TREBALL DE FI DE GRAU

Grau en Estadística

Títol: Clustering d'estratègies en el joc del pòquer

Autor: Mariona Martín Llaveria

Director: Ferran Reverter Comes

Departament: Genètica, Microbiologia i Estadística.
Secció Estadística

Convocatòria: Juny 2020

UNIVERSITAT DE BARCELONA

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

RESUM

Aquest treball pretén estudiar quines són les estratègies seguides pels nostres oponents en un nivell NL5 i NL10 del joc *Texas Hold'em Poker* per tal de poder aplicar el nostre joc adaptat a cada jugador. Per dur a terme la classificació de possibles estratègies es realitzarà un estudi específic prèvi per aquelles variables més significatives i es descriuran aquells rangs de percentatges guanyadors i perdedors.
Posteriorment farem un estudi de possibles agrupacions de clústers entre jugadors per més tard, intentar predir un model que ens classifiqui aquests mitjançant arbres de classificació.

Paraules clau: pòquer, estratègies, *clúster*, PAM clustering, *Random Forest, CART*, arbres de classificació.

Classificació:

- 62H30 *Classification and discrimination; cluster analysis [See also* 68T10, 91C20]
- 91C20 *Clustering [See also* 62H30]
- 05C05 *Trees*

ABSTRACT

This work aims to study the strategies followed by our opponents at an NL5 and NL10 level of Texas Hold'em Poker game in order to be able to apply our game adapted to each player. To carry out the classification of possible strategies, a specific prior study will be carried out for those most significant variables and those ranges of winning and losing percentages will be described.
Later we will make a study of possible groupings of clusters between players for later, to try to predict a model that classifies these to us by means of classification trees.

Key words: poker, strategies, *clúster*, PAM clustering, *Random Forest, CART*, classification tress.

Classification:

- 62H30 *Classification and discrimination; cluster analysis [See also* 68T10, 91C20]
- 91C20 *Clustering [See also* 62H30]
- 05C05 *Trees*

ÍNDEX DE CONTINGUTS

ÍNDEX D'IL·LUSTRACIONS

I. INTRODUCCIÓ

Des de fa 3 anys que vaig començar a jugar al pòquer de manera amateur. A mesura que anava jugant mans em sorgien diverses preguntes de caire estadístic. Poc a poc, vaig anar comprovant que no només es tracta d'un joc d'atzar, com assegura gran part de la població, si no que està compost per moltes variables que amb ajut de diferents softwares es poden semi-controlar. És a dir, que aquests softwares t'avisen de com hauries d'apostar una mà envers un *flop* concret o un adversari en concret. No obstant que aquests softwares estadístics et faciliten els valors de les diferents variables, cal saber interpretar-les i aplicar-les de tal manera que siguis capaç de construir una pròpia estratègia guanyadora.

És per aquest motiu que he volgut dedicar-me, en aquest treball, a cercar possibles estratègies mitjançant l'estudi d'aquelles variables més representatives realitzant *clusterings* jeràrquics on provarem de trobar grups d'individus que siguin el més homogenis possible entre ells i que, alhora, es diferenciïn al màxim amb els individus d'altres grups. Clarament ens serà una tasca complicada cercar i definir aquestes possibles estratègies, ja que hem de tenir en compte molts altres factor com pot ser l'atzar i la probabilitat que caiguin certes cartes o bé, la banca de la que disposem per jugar, el nivell on juguem, el *rake* de la sala de joc, etc. Un cop trobat els clústers més representatius per a la nostre base de dades, mirarem de prediure un model mitjançant arbres de classificació.

En aquest treball implementarem diferents tècniques de Clustering i s'aplicaran en les diferents variables del pòquer online. Més concretament, analitzarem mans des de nivells de NL10 a NL50 (*No Limit*) de *Texas Hold'em* de 6*max* (taules de 6 persones). Per fer aquest anàlisis, s'extraurà la base de dades del software estadístic *Holdem Manager 2* que recopila informació dels jugadors contra el que s'ha jugat. Més concretament ens resumeix totes les mans o jugades que hem tingut amb el jugador.

L'objectiu principal, per tant, consisteix en trobar quines i com són les estratègies guanyadores i perdedores de la modalitat *Texas Hold'em*, per tal d'enfocar el nostre joc a aquelles que siguin guanyadores. Realitzarem, per tant un estudi de validació creuada i ajustarem els paràmetres i variables per tal d'obtenir el millor resultat. És a dir, les estratègies que produeixin més benefici per al jugador/a.

Vull agrair al meu tutor Ferran Reverter, l'ajuda prestada durant aquesta època tant complicada i la dedicació que ha mostrat per a que aquest treball sortís endavant. També agrair-li que hagi acceptat el tema del treball, un tema que ell desconeixia però que jo tenia ganes d'investigar.

II. CONCEPTES SOBRE EL TEXAS HOLD'EM POKER

ORIGEN DEL TEXAS HOLD'EM, EVOLUCIÓ I SITUACIÓ ACTUAL:

L'origen del pòquer és un tant incert. La creença més popular és que els xinesos el van inventar vora l'any 99 *d.C*. com una modalitat del dominó. Altres afirmen que el pòquer va néixer a partir del joc persa "*as nas"*, un joc de cinc jugador on es necessita un mall especial de vint-i-cinc cartes i 5 pals diferents.
No obstant, la teoria més acceptada el relaciona amb el joc francès "*poque"*. Els francesos van portar aquest joc a Nova Orleans vora el 1480 i allà es va popularitzar gràcies als vaixells de vapor que navegaven pel riu Mississipí. El "*poque"* es va començar a denominar *poker* (pòquer en català), i durant la Guerra Civil Americana (1861-1865) es van fer populars algunes modificacions com el *stud poker,* el *draw* i el *straight.* Posteriorment arribaria una altre modalitat anomenada *Texas Hold'em Poker,* nascuda a Robstown (Texas) sobre l'any 1900. Aquesta es va convertir en la modalitat triada per disputar les Sèries Mundials de Pòquer (*WSOP*) i actualment és la variant més jugada i popular arreu del món.

La majoria dels jugadors de pòquer actuals procedeixen d'internet. Cada vegada més les cases d'apostes ofereixen més possibilitats de jugar *online* i, és aquest un dels motius de la gran difusió que ha tingut i segueix tenint aquest joc.
Segons publica la revista *Forbes*, el president i director executiu de nombrosos casinos a Las Vegas (Sheldon Adelson) ha sigut, a nivell mundial, un dels empresaris que més ha incrementat la seva fortuna des del 2013.
Un altre factor decisiu en el *boom* del pòquer està associat al nom Chris Moneymaker. Aquest jugador jove i sense gaire experiència, va aconseguir classificar-se per les WSOP al 2003 després d'haver guanyat un torneig classificatori (torneig satèl·lit *online*) de 39 dòlars d'inscripció a la sala *PokerStars*. Si es té en compte que, aleshores, la entrada ordinària al WSOP costava uns deu mil dòlars, va ser sorprenent com aquest jugador va guanyar el torneig i es va fer amb els dos milions i mig de dòlars de premi.

Chris Moneymaker guanya les WSOP al 2003

Aquesta forma, aparentment fàcil, de guanyar diners va augmentar l'interès pel pòquer, sobretot pel pòquer online. Aquesta fita va causar un gran punt d'inflexió pel pòquer *online*. A més a més, cada vegada amb més freqüència els mitjans de comunicació retransmeten partides i tornejos de *Texas Hold'em*.

REGLES DEL TEXAS HOLD'EM POKER:

Com s'ha comentat a l'apartat anterior, aquesta variant del pòquer és la més pràctica a nivell mundial. Té una mecànica d'aprenentatge fàcil que pot semblar que es denomina el joc un cop jugades poques mans. No obstant, es tracta d'un joc amb infinitats de detalls on es requereix molta habilitat i perspicàcia.

Per jugar es necessita un mall de 52 cartes franceses sense el comodí. La baralla està organitzada en 4 pals (piques, diamants, cors i trèvols) amb tretze cartes cada un, organitzades de major a menor com: A, K, Q, J, 10, 9, 8, 7, 6, 5, 4, 3, 2. També són necessàries un joc de fixes (*chips*) o bé, diners en metàl·lic per realitzar les apostes. Poden jugar de dos a deu jugadors en una mateixa taula i el jugador quedarà eliminat quan es quedi sense fixes per apostar.

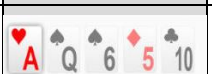
Per començar, un dels jugadors ha d'exercir de repartidor (*dealer*), càrrec que anirà rotant mà rere mà en el sentit de les agulles del rellotge. Es diu que aquest jugador ocupa la posició del botó a la taula. Els dos jugadors situats a la seva esquerra han d'apostar forçosament una certa quantitat predeterminada (cegues *o blinds*). El primer jugador aporta la cega petita (*small blind, SB*) i el segon aporta la cega gran (*big blind, BB*) que acostuma a ser el doble de l'anterior. A continuació, es reparteixen dues cartes boca avall , que són privades, a cada jugador. A mesura que avança la mà, es col·loquen cinc cartes, compartides, boca amunt al centre de la taula. Aquestes 5 cartes s'ensenyen en 3 etapes distintes: *flop* (tres cartes), *turn* (1 carta) i *river* (1 carta).

Tot aquest procés dona lloc a quatre rondes d'apostes. Si durant alguna ronda tots els jugadors menys un abandona, es dona per finalitzada la mà i el jugador que resta viu rep totes les fixes que s'han apostat fins aquell moment. Aquest conjunt de fixes rep el nom de pot comú (*pot*).

Cal saber que cada jugador, quan arriba el seu torn per "parlar" pot realitzar una d'aquestes accions:

- Acceptar l'aposta actual (***call***).
- Abandonar la mà (***fold***).
- Continuar jugant però sense apostar (***check***).
- Realitzar una aposta (***raise***).
- Pujar l'aposta realitzada per un altre jugador. "*Resubir"* (***re-raise***).
- Apostar totes les seves fixes restants (***all in*** *o* ***push***).

Finalment, i no menys important, cal tenir clara quina és la jerarquia o *ranking* de les diverses mans possibles. Aquesta jerarquia, en forma descendent, és:

Ranking	Mà	Denominació	Explicació
1º		**Escala real** (*Royal flush*)	És la millor mà. Cinc cartes més altes possibles, consecutives, del mateix pal.
2º		**Escala de color** (*Straight flush*)	Cinc cartes consecutives del mateix pal
3º		**Poker** (*Poker, Four of a kind*)	Quatre cartes d'igual numeració.
4º		**Full** (*Full house*)	Dos grups de cartes: un de tres i l'altre de 2. Ambdós amb igual numeració.
5º		**Color** (*Flush*)	Cinc cartes del mateix pal.
6º		**Escala** (*Straight*)	Cinc cartes amb numeració consecutiva.
7º		**Trio** (*Three of a kind*)	Tres cartes d'igual numeració.
8º		**Doble parella** (*Two pairs*)	Dos grups de dos cartes d'igual numeració.
9º		**Parella** (*Pair*)	Dos cartes d'igual numeració.
10º		**Carta Alta** *(High card)*	Cap de les combinacions anteriors.

Jerarquia de mans en el Texas Hold'em pòquer

Es pot donar el cas que dos mans del mateix rang, per exemple, dos jugadors que tinguin *full,* escala, dobles parelles, etc. En aquest casos ens regim per la carta més alta. Si es donés el cas que dues mans fossin exactament iguals (tenir la mateixa parella, per exemple) es repartiria el pot comú en parts iguals pels dos jugadors.

DIFERENTS SALES DE JOC:

En aquest treball ens centrem en la modalitat del *Texas Hold'em* pòquer *online* i és important saber quines sales o plataformes *online* hi han, per tal de comparar-les i triar jugar a la que més ens convingui.

Abans cal saber que no va ser fins al 2011 que el pòquer es va regularitzar a molts països (sobretot europeus). Aquest va ser el cas d'Espanya, on es va començar a regularitzar el mercat, és a dir a cobrar impostos als jugadors. Altres països com l'Anglaterra, van decidir fer una regularització oberta on els seus jugadors seguien jugant contra rivals d'arreu del món sense cobrar-los impostos, no obstant se'ls afegia un cobrament extra abans de la inscripció en cada torneig.

En el cas d'Espanya, es va decidir realitzar un mercat tancat del joc on només jugaven entre espanyols, però aplicant els cobraments que es portaven a terme a l'Anglaterra. A més, es va afegir un extra a l'hora de fer la declaració de la renta. En resum, es va dur a terme una doble tributació, fet que es va considerar un tant abusiu.
Tots aquests canvis en la regularització van fer que molts professionals del pòquer decidissin emigrar a terres més pròsperes, fet que va causar que Espanya passes de tenir una gran fluència de jugadors a tenir-ne molt poca. Va ser doncs al 2017 que la situació va donar un gir molt favorable gràcies a que Espanya, Portugal, Itàlia i França es posessin d'acord i ajuntessin els seus mercats. Avui dia la situació és molt més sostenible i la professionalitat a Espanya torna a ser viable tot i que amb una regulació menys favorable que en altres països.

Des del 2017, les quatre principals sales on es pot jugar des d'Espanya i que tenen un tràfic òptim són ***PokerStars, 888 Poker, WinaMax*** **i** ***PartyPoker***.
A continuació les compararem segons el seu tràfic de jugadors (a dia 20 de maig del 2020, a les 4.49PM) i el *rake o* comissió (diners que es queda la sala, on s'està jugant, de cada bot que es juga) aplicat:

- **PokerStars:**

  L'afluència de jugadors en aquesta sala és:

Poker Site	Online	Cash	24 H Peak	7 Day avg	Last Week
PokerStars	55540	8731	19789	12000	

Afluència jugadors en sala PokerStars

  Es poden observar unes 55.540 persones jugant *online* des de la plataforma **PokerStars** i en la modalitat de *cash* (no tornejos) unes 8.731 . A més a més, de mitjana durant els set dies de la setmana es connecten unes 12.000 persones.

I el seu *rake* en els diferents nivells:

Sin límite y pot limit (salvo en Zoom)

Nivel de apuestas	% de comisión	Máx. para 2-4 jugadores	Máx. para 5 jugadores	Máx. para 6 o más jugadores
0,01 €/0,02 €	6 %	0,10 €	0,15 €	0,25 €
0,02 €/0,05 €	6 %	0,40 €	0,50 €	0,70 €
0,05 €/0,10 €	6 %	0,75 €	0,95 €	1,25 €
0,10 €/0,25 €	5,75 %	1,60 €	2,05 €	2,85 €
0,25 €/0,50 €	5,75 %	1,85 €	2,35 €	3,00 €
0,50 €/1 €	5,75 %	2,50 €	2,85 €	3,00 €
1 €/2 €	5,75 %	3,00 €	3,00 €	3,00 €
2,50 €/5 €	5,75 %	3,00 €	3,00 €	3,00 €
5 €/10 €	5,75 %	3,00 €	3,00 €	3,00 €

Rake de la sala PokerStars

Pel nivell que tractem en aquest treball (NL5,NL10) , la sala *PokerStars* ens cobraria un *rake* o comissió del 6% del *stake* del qual partim en cada taula.

- **888Poker:**

L'afluència de jugadors en aquesta sala és:

Poker Site	Online	Cash	24 H Peak	7 Day avg	Last Week
888poker.es	383	104	575	325	

Afluència jugadors en sala 888Poker

Veiem unes 383 persones jugant online des de la plataforma **888Poker** i en la modalitat de *cash* (no tornejos) unes 104. De mitjana, durant els set dies de la setmana, es connecten unes 325 persones a aquesta hora.

I el seu *rake* en el nivell NL5/NL10:

Límites de las mesas: **5¢/10¢-10¢/20¢**

Nº de jugadores	Comisión por bote	MÁX
2	1¢ por cada 18¢ del bote	1€
3+	1¢ por cada 18¢ del bote	4€

Rake de la sala 888Poker

La comissió (*rake*) cobrada per la sala 888, que serà la nostra sala d'estudi, és d'1 cèntim per cada 18 cèntims del pot comú arribant a cobrar.

- **WinaMax:**

L'afluència de jugadors d'aquesta sala és:

Poker Site	Online	Cash	24 H Peak	7 Day avg	Last Week
Winamax.fr	5878	1062	5065	2800	

Afluència jugadors en sala WinaMax

Es poden observar unes 5.878 persones jugant online des de la plataforma **Winamax** i en la modalitat de *cash* (no tornejos) unes 1.062 . A més a més, de mitjana durant els set dies de la setmana es connecten unes 2.800 persones.

I el seu *rake* en els diferents nivells:

	Número de jugadores en la mesa			
Ciegas	2 jugadores	3 jugadores	4 jugadores	5 jugadores o más
0,01/0,02 €	0,25 €	0,30 €	0,35 €	0,40 €
0,02/0,05 €	0,50 €	0,65 €	0,80 €	1 €
0,05/0,10 €	0,75 €	1 €	1,25 €	1,50 €
0,10/0,20 €	1,5 €	1,75 €	2 €	2,5 €
0,15/0,30 €	1,5 €	2 €	2,5 €	3 €
0,25/0,50 €	1,5 €	2 €	2,5 €	3 €
0,50/1 €	1,5 €	2 €	2,5 €	3 €

Rake de la sala WinaMax

Observem pel nivell NL5 i NL10 quin és l'import de la comissió (*rake*) cobrat per la sala *Winamax* depenent dels jugadors presents a la taula. A més jugadors més import i sabem que la majoria de vegades hi ha 5 jugadors o més presents a la taula. Així que la comissió cobrada en el nivell NL5 serà de 1,50 euros; i en canvi, pel nivell NL10 serà de 2,50 euros.

- **PartyPoker:**

L'afluència de jugadors d'aquesta sala és:

Poker Site	Online	Cash	24 H Peak	7 Day avg	Last Week
PartyPoker	1568	2870	3455	2000	

Afluència jugadors en sala PartyPoker

Es poden observar unes 1.568 persones jugant *online* des de la plataforma **PartyPoker** i en la modalitat de *cash* (no tornejos) unes 2.870 . A més a més, de mitjana durant els set dies de la setmana es connecten unes 2.000 persones.

I el seu *rake* en els diferents nivells:

Límites	Número de jugadores	Rake	Máximo
5/10 € y 10/20 €	2	0,50 por 7,50 €	2 €
	3 - 4	0,50 por 7,50 €	3 €
	5 - 10	0,50 por 7,50 €	4 €
3/6 €	2	0,50 por 7,50 €	2 €
	3 - 4	0,50 por 7,50 €	3 €
	5 - 10	0,50 por 7,50 €	4 €
2/4 €	2	0,25 € por 3,75 €	2 €
	3 - 4	0,25 € por 3,75 €	3 €
	5 - 10	0,25 € por 3,75 €	4 €
1/2 €	2 - 10	0,05 € por 0,75 €	2 €
0,25/0,50 € y 0,50/1€	2 - 10	0,05 € por 0,75 €	2 €
0,10/0,20 € y 0,15/0,30 €	2 - 10	0,01 € por 0,15 €	1 €
0,02/0,04 € y 0,05/0,10 €	2 - 10	0,01 € por 0,15 €	1 €

Rake de la sala PartyPoker

A la sala *PartyPoker* la comissió (*rake*) cobrada en el nivell NL5 i NL10 és de 1 cèntim per 15 cèntims presents en el pot comú.

En resum doncs podríem dir que la sala *888Poker* és la que menys comissió (*rake*) cobra, seguit per la sala *Party Poker*. I, referent a l'afluència de jugadors jugant *cash* s'ha pogut observar com la sala líder en aquest aspecte és *PokerStars*.

A més a més, d'aquestes 4 sales principals explicades hi han, aproximadament, unes 10 sales més petites que tenen un tràfic reduït però que encara així estan en funcionament. Algunes d'aquestes són: *GGPoker*, *Bwin Poker*, etc.

**rake:* comissió que el jugador/a paga a la sala per cada pot jugat.

**rakeback*: és la devolució d'una part del *rake* que generes en la sala gràcies al "acord" fet o estipulat per la sala on s'està jugant.

GESTIÓ DEL BANKROLL:

El més important en el pòquer i el primer que ha d'aprendre un jugador si vol ser guanyador i tenir una vida longeva en el joc, és el fet de saber gestionar el seu *bankroll*. Aquells jugadors recreacionals (*fishes*) no tenen en compte aquest fet, ja que juguen per pura diversió.

Què és el *bankroll*?

Són els diners dels que disposa un jugador de pòquer per jugar. La quantitat la tria el propi jugador i no s'ha de barrejar mai els diners destinats al pòquer amb els diners per viure o d'altres usos (del dia a dia). Sense una bona gestió de la banca (*bankroll*), el jugador de pòquer està destinat a perdre-ho tot.

Imaginem que som uns jugadors de pòquer excel·lents i només disposem de 100 dòlars per jugar. Se'ns ofereixen dues alternatives: seure a una taula de NL5 (la qual té una entrada de €5, amb cegues de €0.05 i €0.1) o seure en una altre de NL100 (€100 d'entrada amb cegues de €0.5 i €1). Sabem que el nivell dels rivals és similar en ambdues taules però molt inferior al nostre nivell, així que arribarem a ser guanyadors, a llarg termini, en qualsevol de les dues.
Les diferències fonamentals entre una partida i l'altre són, en primer lloc, que en la taula de NL5 arriscarem un 5% dels nostres diners mentre que en la taula de NL100 n'arriscarem el 100%. En segon lloc, a la taula de NL100 tindrem la possibilitat de guanyar molts més diners en comparació amb una de NL5.

Coneixent aquestes dades: quina partida hauria de jugar una persona que només disposa de 100 dòlars?

En aquest cas extrem, la resposta pot semblar òbvia. És una temeritat arriscar tots els nostres diners o *bankroll* en una sola taula, per molts diners que puguem arribar a guanyar en ella ja que qualsevol mà traïdora ens pot portar a la banca fallida, és a dir a perdre-ho tot. Tot i que els nostres guanys mitjos siguin menors en una taula de NL5, sembla l'opció més sensata, doncs el risc de fallida és molt menor.
Per arribar a aquesta conclusió ens hem vist obligats a realitzar un petit anàlisis de gestió de risc, que ens ha portat a prendre la decisió que més "fora de perill" posa la nostra banca. Aquests mateixos principis són els que utilitzarem per dissenyar la nostra estratègia de gestió de banca.

En l'exemple anterior, no ens ha quedat cap altre remei que decantar-nos per la taula de NL5, però això no significa que posar en joc el 5% del nostre *bankroll* sigui el correcte. De fet segueix sent una gran imprudència, pel que ens porta a fer-nos la següent pregunta: quina quantitat és l'adequada?

Com veurem a continuació, no existeix una única estratègia de gestió de banca; i per dissenyar-la correctament s'ha de tenir en compte la modalitat de joc a la que ens dedicarem (*cash games*, tornejos, *sit & go*, etc.).
Tota estratègia de gestió de banca que estigui ben dissenyada ha de respondre a un mateix propòsit: eliminar les opcions de fallida durant una ratxa negativa.

Abans d'entrar en matèria és convenient explicar els perills de no respectar les normes de la nostra estratègia. A més del ja anomenat risc de fallida, el fet de jugar fora de banca també pot arribar a afectar al nostre estil de joc, per exemple en situacions on la jugada correcta sigui realitzar una gran aposta i, per por a perdre molts diners en ella, prendre una decisió més aviat conservadora on ens veiem obligats a realitzar un joc excessivament temorós i passiu.

Estratègia de gestió de la banca en taules de *cash*:

Es parteix de la base que tot jugador de pòquer hauria de començar jugant les taules més baixes de la sala NL2 (amb cegues €0,01- 0.02).

Nivell	Comencem amb:	Meta per pujar de nivell:	Baixem al nivell anterior si retrocedim a:
NL2	20 €	400 €	-
NL5	400 €	800 €	100 € (pèrdues de 300 €)
NL10	800 €	2000 €	250 € (pèrdues de 550 €)
NL25	2000 €	4000 €	500 € (pèrdues de 1500 €)
NL50	4000 €	8000 €	1000 € (pèrdues de 3000 €)
NL100	8000 €	16000 €	2500 € (pèrdues de 5500 €)

Gestió bankroll per nivells en taules de cash

Quan s'aconsegueix un nivell?

De la taula anterior es poden extreure vàries conclusions. La primera és que, a excepció del NL2, un nivell queda batut quan al menys hem duplicat el *bankroll* amb el que hem accedit a ell. La segona, i més important, és la necessitat de baixar al nivell anterior quan les pèrdues són les estipulades en la quarta columna. Això pot resultar dur psicològicament però és fonamental realitzar-ho correctament per tal de disminuir el nostre risc de fallida. D'aquesta manera podrem recuperar el nostre *bankroll* una vegada tornem al nivell on ja havíem demostrat ser guanyadors amb anterioritat.

Per últim, és necessari indicar que el número de mans jugades en un nivell és tan important com els guanys obtinguts en ell, per tal de deixar patent que aquest ha sigut superat. De no ser així, podria donar-se el cas que degut a una sèrie de sessions excepcionals, arribéssim a la fita sense haver jugat masses mans.

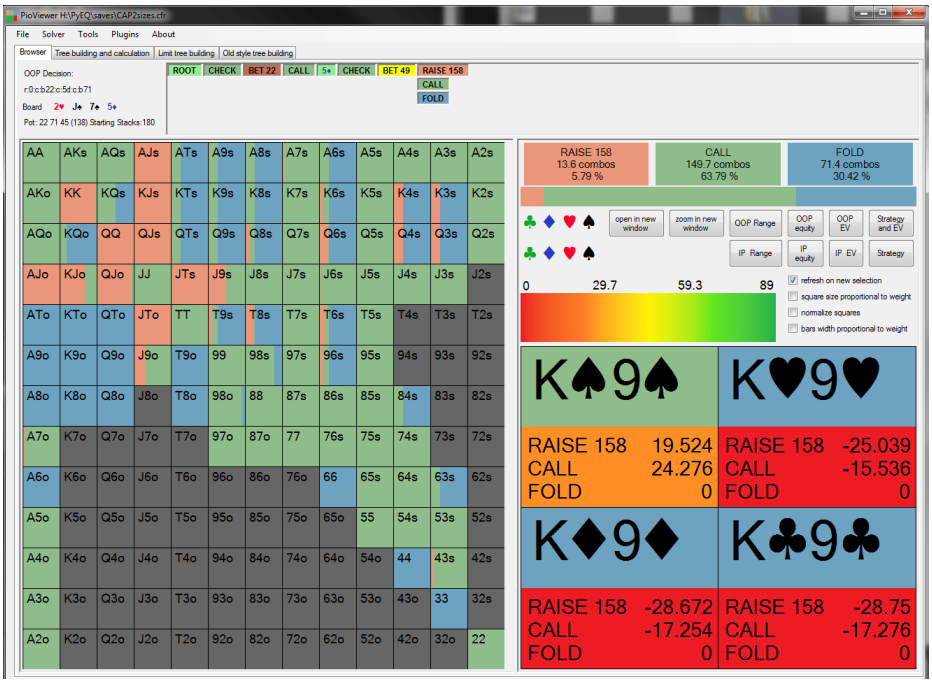
En aquest cas, els nostres números no serien gaire fiables, ja que els nostres guanys podrien haver estat provocats per la bona sort o l'atzar.
És necessari, per tant, contrastar mitjançant una mostra de mans grans (mai inferior a cinquanta mil mans) que estem preparats per donar el salt al següent nivell. En aquest aspecte és important no impacientar-se ja que un salt prematur fomentat per una bona ratxa a curt termini sense haver jugat aquest número de mans mínimes, pot ser contraproduent i frustrant pel nostre futur com a jugadors de pòquer.

JOC GTO (EQUILIBRI EN EL JOC):

L'abreviatura GTO prové de l'anglès i significa *Game Theory Optimal.* És un terme que fa referència a l'estratègia òptima que cal seguir un jugador de pòquer. Aquesta utilitza diversos postulats de la Teoria de Jocs de les matemàtiques aplicades.
El GTO ha estat creat per tal que el jugador de pòquer pugui construir un equilibri en el seu joc. Es tracta de modelar el joc dels oponents i les diferents situacions en terme de rangs i probabilitats, en comptes d'estar estrictament orientats o enfocats als propis resultats.

Com arribem, doncs, al joc GTO en el pòquer?

Al 2010 el joc GTO no existia, la majoria d'estratègies eren creades de forma intuïtiva. Si els millors professionals de pòquer innovaven moviments, la resta es limitava a imitar-los. Han existit, doncs, corrents de tot tipus en el progrés dels *cash games*, però va ser al 2015 quan van aparèixer els *solvers.* Els *solvers* són programes informàtics (d'alt cost) que saben jugar al pòquer de la manera més òptima, seguint el joc GTO.



Solució òptima per la mà K9 en el board 2J75 (PioSolver)

Antigament es discutia entre jugadors per saber quina era la manera més òptima de jugar una mà en concret. Ara això ja no és necessari perquè si introduïm els *solvers* a la jugada, aquests ens indiquen quina és la línia més rentable de la mà o bé amb quina línia guanyarem més diners a llarg termini.

S'ha de jugar sempre un joc GTO?

La resposta és "NO", tot i que pugui semblar el contrari, perquè suposadament s'ha de jugar sempre de la forma més perfecta i equilibrada possible i; es donen casos especials on seguir un joc GTO pot ser no ho és.

En el pòquer, a la mateixa vegada que hi han infinites línies de joc també hi ha infinits i diversos tipus de jugadors. Aquesta infinitat de possibilitats de jugadors els *solvers* no les saben identificar ja que no coneixen a "X" jugador com podem arribar a conèixer-lo nosaltres una vegada hem jugat milers de mans envers.

Per exemple, si es dona el cas que juguem envers un jugador recreacional (*fish*) , la forma més rentable de jugar-li una mà igual no és fer-li una estratègia GTO, sinó una estratègia explotativa o el que és el mateix, intentar treure el màxim partit de la mà jugant d'una forma excessivament agressiva o excessivament passiva. D'aquesta manera li estem parant una "trampa" al rival que té menys experiència.

Aleshores, seguim l'estratègia GTO o bé, l'estratègia explotativa?

El GTO és el joc perfecte per excel·lència, un joc balancejat sense fissures. Però en ocasions per guanyar diners extres a rivals precisos necessitarem desbalancejar-nos per treure-hi, encara, més partit a la mà.
Pel que la millor estratègia seria una estratègia mixta, combinació de les dues, sempre i quan sapiguem com i quan aplicar-la.

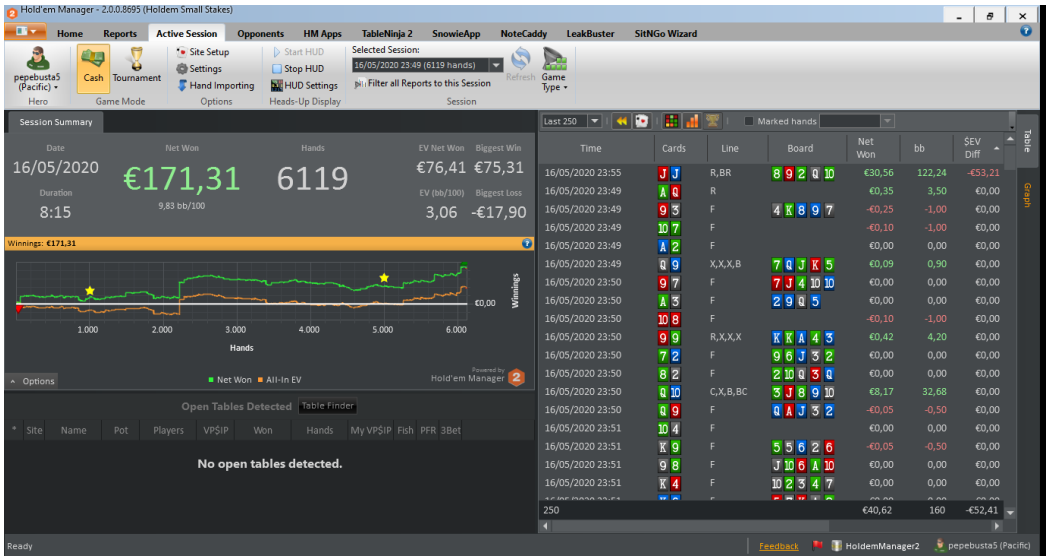
Resumint, la teoria GTO es tractaria en definir des d'un punt de vista matemàtic el joc perfecte contra cada rival i en cada acció en el pòquer. Vindria a ser la "pedra filosofal" del pòquer.

SOFTWARE HOLDEM MANAGER 2:

El software *Holdem Manager 2* és una eina professional dissenyada per facilitar informació en viu a qualsevol jugador de pòquer. Aporta molta informació sobre el joc dels oponents. Amb l'ajuda d'aquesta eina es podrà visualitzar els valors dels paràmetres de cada jugador en temps real, a la taula on s'estigui jugant.

Recopila totes les mans jugades, totes les accions dels oponents i crea estadístiques detallades per a cada jugador. Els jugadors que disposin d'aquesta eina tindran un cert avantatge sobre la resta ja que podran visualitzar les targetes (*HUD)* de tots els jugadors, de forma intuïtiva, utilitzant informació sobre les seves mans anteriors enregistrades.

A més a més d'estadístiques en viu sobre cada jugador, facilita informació d'un mateix. Per exemple, quins guanys nets s'han obtingut durant tot el mes o bé, quantes vegades s'ha perdut AA, etc.



Captura del resum de la sessió en el software Hold'em Manager 2

A la part superior-esquerra de la imatge podem visualitzar les nostres mans jugades de la sessió (6119), els nostres guanys nets (171,31€), el dia de la sessió (16/05/2020),la duració de la sessió (8 hores i 15 minuts),etc. A la part dreta es visualitzen totes les 6119 mans jugades, una per una, amb les seves accions i característiques corresponents.

No totes les sales de pòquer permeten utilitzar aquest software. Algunes sales es centren en un joc tradicional i prohibeixen l'ús d'aquest tipus de programes (per exemple, la sala P*okerDom*). Sales com *PokerStars, 888 Poker, Winamax, Party Poker* són algunes de les que admeten l'ús i aplicació d'aquesta eina.

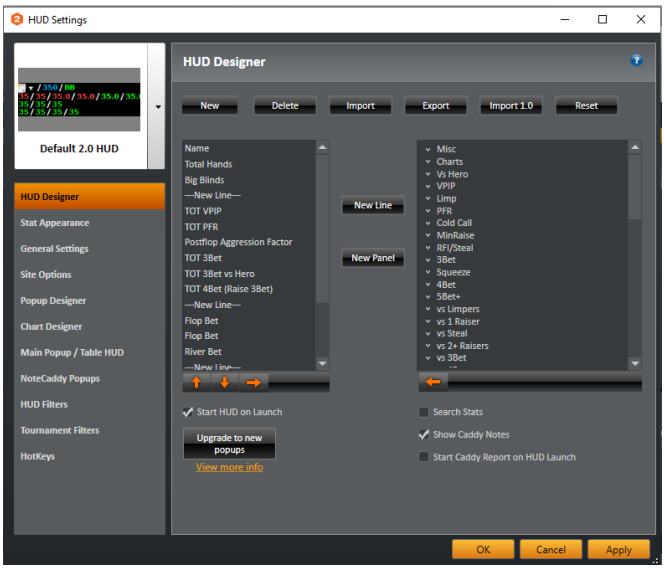
Què és l'HUD?

L'HUD (*heads-up display*) són estadístiques d'altres jugadors que es mostren directament en la pantalla durant la sessió de joc.

Captura taula de NL10 a la sala 888Poker amb els HUD dels jugadors presents

Les estadístiques es formen durant el joc. Quantes més mans s'hagin jugat amb l'oponent, més estadístiques es tindran sobre les seves accions i més precisa serà, aquesta. Amb l'ajuda dels diferents indicadors, es pot predir les accions i la mà d'un jugador amb major precisió i, també , es podrà exprimir el major valor possible a la nostre mà.

Les principals estadístiques dels oponents es mostren vora el seu nom, a la posició on estan asseguts. Quan es senyalen amb el ratolí, apareix una finestra emergent amb les estadístiques més detallades. Tots poden personalitzar l'HUD per a sí mateixos, seleccionant les estadístiques que consideren més necessàries a tenir en compte i; així, poder visualitzar-les en viu mentre la mà s'està jugant (veure imatge). Gràcies a l'HUD es poden prendre decisions ràpides i amb sentit.

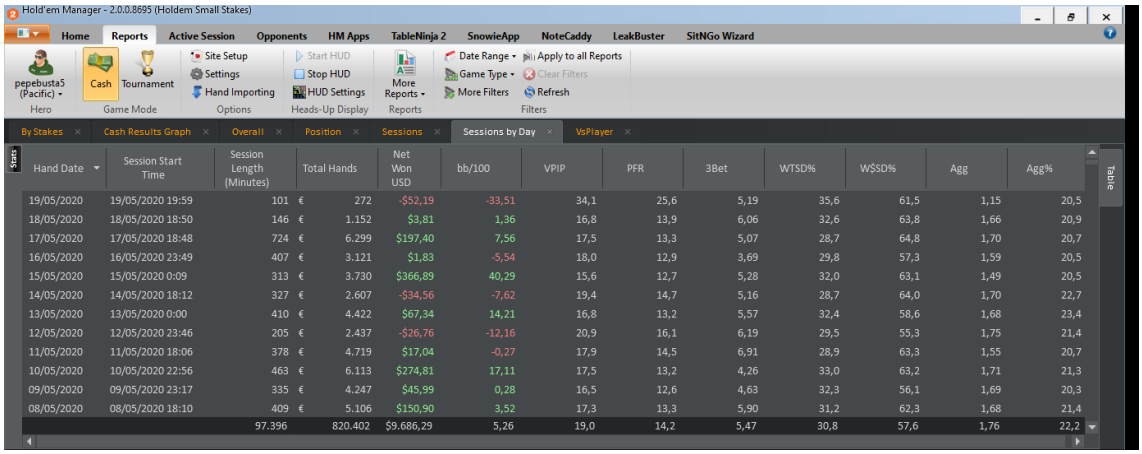


Captura del disseny propi dels HUD

III. METODOLOGIA

TRACTAMENT BASE DE DADES D'ESTUDI

Es disposa d'una base de dades, anomenada ***holdem***, extreta del software ***Holdem Manager 2*** en format ".csv" . La base de dades es caracteritza per tenir 8.054 registres que són els diversos jugadors/es contra els que he jugat alguna mà i jo mateixa, en una mateixa taula, i 29 columnes que són les diferents variables de que es composa la base de dades *holdem*. Les files són jugadors no repetits i cada fila resumeix les característiques o valors d'aquell jugador quan he estat jo present a la taula. És a dir, resumeix totes les seves mans.



Hand Date	Session Start Time	Session Length (Minutes)		Total Hands	Net Won USD	bb/100	VPIP	PFR	3Bet	WTSD%	W$SD%	Agg	Agg%
19/05/2020	19/05/2020 19:59	101	€	272	-$52,19	-33,51	34,1	25,6	5,19	35,6	61,5	1,15	20,5
18/05/2020	18/05/2020 18:50	146	€	1.152	$3,81	1,36	16,8	13,9	6,06	32,6	63,8	1,66	20,9
17/05/2020	17/05/2020 18:48	724	€	6.299	$197,40	7,56	17,5	13,3	5,07	28,7	64,8	1,70	20,7
16/05/2020	16/05/2020 23:49	407	€	3.121	$1,83	-5,54	18,0	12,9	3,69	29,8	57,3	1,59	20,5
15/05/2020	15/05/2020 0:09	313	€	3.730	$366,89	40,29	15,6	12,7	5,28	32,0	63,1	1,49	20,5
14/05/2020	14/05/2020 18:12	327	€	2.607	-$34,56	-7,62	19,4	14,7	5,16	28,7	64,0	1,70	22,7
13/05/2020	13/05/2020 0:00	410	€	4.422	$67,34	14,21	16,8	13,2	5,57	32,4	58,6	1,68	23,4
12/05/2020	12/05/2020 23:46	205	€	2.437	-$26,76	-12,16	20,9	16,1	6,19	29,5	55,3	1,75	21,4
11/05/2020	11/05/2020 18:06	378	€	4.719	$17,04	-0,27	17,9	14,5	6,91	28,9	63,3	1,55	20,7
10/05/2020	10/05/2020 22:56	463	€	6.113	$274,81	17,11	17,5	13,2	4,26	33,0	63,2	1,71	21,3
09/05/2020	09/05/2020 23:17	335	€	4.247	$45,99	0,28	16,5	12,6	4,63	32,3	56,1	1,69	20,3
08/05/2020	08/05/2020 18:10	409	€	5.106	$150,90	3,52	17,3	13,3	5,90	31,2	62,3	1,68	21,4
		97.396		820.402	$9.686,29	5,26	19,0	14,2	5,47	30,8	57,6	1,76	22,2

Base de dades d'estudi en el software Holdem Manager 2

Fent un primer anàlisis es pot veure com no hi ha cap *missing* (*NA*) present en la base de dades. Per tant, en aquest aspecte no caldrà modificar cap variable. No obstant, es va considerar necessari fer alguna modificació en la base de dades.

Primerament, es va reduir el nombre de registres (files/jugadors-es) a aquells que almenys haguessin jugat 300 mans (variable "*Hands*") en contra nostra, per poder fer un anàlisi amb consistència. Per tant, amb la condició que mínim hi hagin 300 mans jugades amb aquell jugador/a (*Hands* > 300), es passarà de 8054 registres a 2222, sabent que no estan repetits els jugadors en cap cas. Seguidament, vaig decidir suprimir-me a mi mateixa de la base de dades ja que l'estudi està enfocat als meus oponents i no són rellevants les meves dades. Es va prendre aquesta decisió ja que a la base de dades es guarden totes les meves mans però, en canvi, per la resta de jugadors/es només es guarden aquelles mans jugades quan he estat jo present a la taula. Per això la observació del primer jugador/a és un punt extrem per a totes les variables i és convenient suprimir-lo de la base de dades a estudiar. Per tant de 2222 observacions (files) que es tenien a la base de dades pre-processada, es passarà a 2221 observacions / individus / jugadors-es. A més a més, s'agafaran aquelles columnes/variables més rellevants, és a dir les més importants i fàcils de modificar pel joc per a l'estudi que es farà a posteriori. Doncs, ens quedarem amb 12 de les 29 variables inicials. Per tant es tindrà la base de dades d'estudi pre-processada (*holdem*) amb un total de 2221 files i 12 columnes.

D'aquestes 12 variables triades, n'hi hauran de 3 tipus diferents:

- Primerament, trobem la variable "*Player Name*" que és l'*id* de la nostre base de dades i fa referència als *nicks* dels diferents jugadors/es amb els que s'ha jugat, a taula i en contra.
  És de **tipus caràcter** i són una simple etiqueta tan informativa com distintiva.

- Seguidament, hi ha la variable "*Site*" que és l'única **variable categòrica** de la base de dades. Fa referència a les diferents plataformes de joc de pòquer que es recullen en el software *Holdem Manager 2*.

- Finalment, la resta de **variables són de caràcter numèric**. Les variables "*Net Won*" i "*Hands*" són variables numèriques discretes i la resta de variables estan expressades en tant per cent (%), per tant són variables numèriques contínues.

Aleshores, a partir d'ara, es procedirà a realitzar una descriptiva univariant de les variables més rellevant per l'estudi, una descriptiva univariant segregada per la variable *Net Won* de les mateixes variables i, finalment, una descriptiva bivariant.

DESCRIPCIÓ VARIABLES D'ESTUDI

Un cop es tenen les variables de la base de dades triades per a l'estudi, cal descriure-les per tal de saber què tracta cada una de les variables.

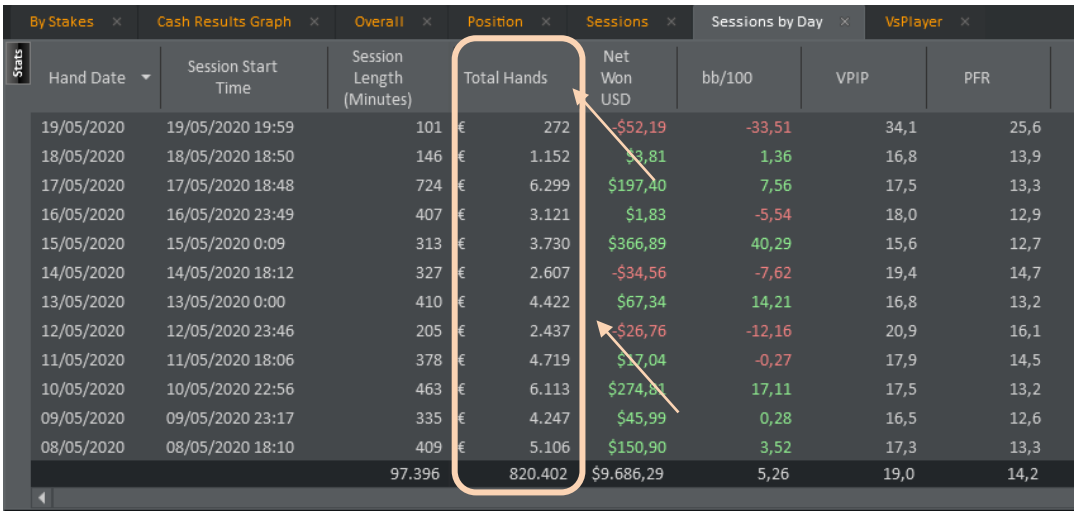
- ***SITE :***

La variable *Site* fa referència a la sala on s'ha jugat contra aquell jugador/a. En la nostra base de dades només tenim 3 sales registrades, que són en les que hem jugat mans i aquestes són: *PokerStars, 888Poker* i *Winamax*.

3 sales de pòquer analitzades (PokerStars, 88Poker i WinaMax

- ***HANDS:***

La variable *Hands* fa referència a la quantitat de mans jugades contra aquell jugador en una mateixa taula. És la variable que recull el número de mans acumulades i jugades contra aquell jugador en concret.



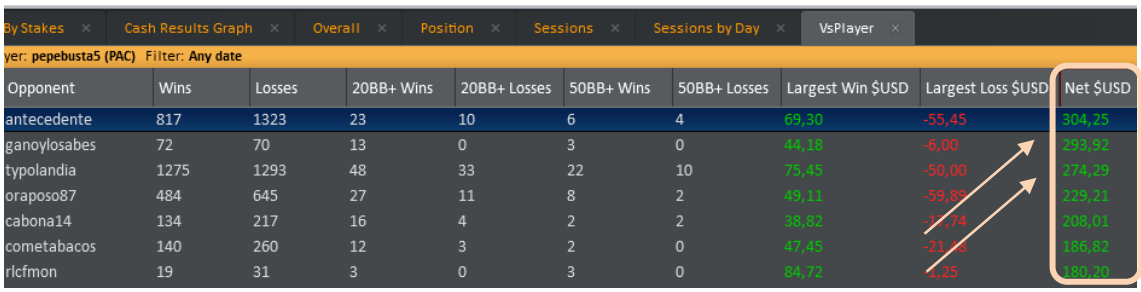
Hand Date	Session Start Time	Session Length (Minutes)	Total Hands	Net Won USD	bb/100	VPIP	PFR
19/05/2020	19/05/2020 19:59	101	€ 272	-$52,19	-33,51	34,1	25,6
18/05/2020	18/05/2020 18:50	146	€ 1.152	$3,81	1,36	16,8	13,9
17/05/2020	17/05/2020 18:48	724	€ 6.299	$197,40	7,56	17,5	13,3
16/05/2020	16/05/2020 23:49	407	€ 3.121	$1,83	-5,54	18,0	12,9
15/05/2020	15/05/2020 0:09	313	€ 3.730	$366,89	40,29	15,6	12,7
14/05/2020	14/05/2020 18:12	327	€ 2.607	-$34,56	-7,62	19,4	14,7
13/05/2020	13/05/2020 0:00	410	€ 4.422	$67,34	14,21	16,8	13,2
12/05/2020	12/05/2020 23:46	205	€ 2.437	-$26,76	-12,16	20,9	16,1
11/05/2020	11/05/2020 18:06	378	€ 4.719	$17,04	-0,27	17,9	14,5
10/05/2020	10/05/2020 22:56	463	€ 6.113	$274,81	17,11	17,5	13,2
09/05/2020	09/05/2020 23:17	335	€ 4.247	$45,99	0,28	16,5	12,6
08/05/2020	08/05/2020 18:10	409	€ 5.106	$150,90	3,52	17,3	13,3
		97.396	820.402	$9.686,29	5,26	19,0	14,2

Columna referent a les mans jugades per rival

En la imatge podem visualitzar la pantalla principal del software Holdem Manager 2 on apareix la informació del jugador, i senyalat per les fletxes taronges apareix la variable Hands, en aquest cas concret les mans jugades per día. No obstant a la nostra base de dades apareixen les mans totals jugades d'aquell jugador quan he estat jo present, no només d'un dia en concret.

- ***NET WON:***

Aquesta variable conté els guanys nets o pèrdues netes acumulades del jugador quan he estat present a la taula.



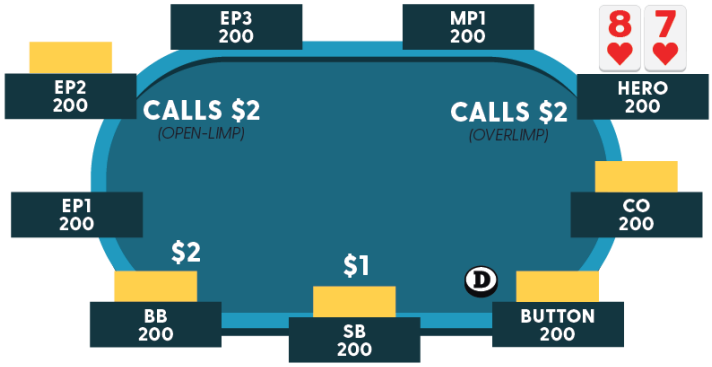
Opponent	Wins	Losses	20BB+ Wins	20BB+ Losses	50BB+ Wins	50BB+ Losses	Largest Win $USD	Largest Loss $USD	Net $USD
antecedente	817	1323	23	10	6	4	69,30	-55,45	304,25
ganoylosabes	72	70	13	0	3	0	44,18	-6,00	293,92
typolandia	1275	1293	48	33	22	10	75,45	-50,00	274,29
oraposo87	484	645	27	11	8	2	49,11	-59,89	229,21
cabona14	134	217	16	4	2	2	38,82	-17,74	208,01
cometabacos	140	260	12	3	2	0	47,45	-21,48	186,82
rlcfmon	19	31	3	0	3	0	84,72	-4,25	180,20

Columna referent als guanys nets per rival

En el requadre taronja podem observar la variable Net Won dels oponents contra els que he jugat, entre d'altres.

- ***VP$IP (Voluntary Put In the Pot) :***

  Indica el percentatge (%) de vegades que els diners són voluntàriament introduïts en el pot comú, i senyala en quin percentatge de casos un jugador/a posa diners en el pot abans del *flop*. Si un jugador/a que està situat a la cega veu el *flop* sense cap inversió prèvia addicional (perquè ningú ha decidit pujar el valor de la cega) aleshores no està inclòs en el *VP$IP*. És a dir, si estàs situat en la cega gran i t'arriba l'opció de veure el flop sense cap inversió no formaria part del *VP$IP*.



Jugador EP2 i HERO fan limp (VP$IP)

- ***PFR (Preflop Raise):***

  És la pujada de la cega abans de veure el *flop*. Indica el percentatge (%) de vegades que un jugador/a puja la cega, en la primera ronda d'apostes. Un "PFR" del 10%, per exemple, indicaria que un jugador/a ha pujat el 10% de les seves mans en la primera ronda. Un PFR per tant és un *VP$IP*.

Jugador EP realitza un PFR

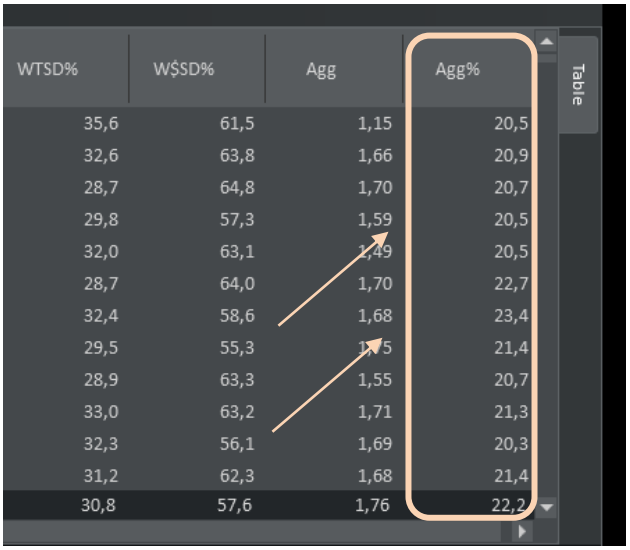
- ***3Bet:***

  És el percentatge (%) de vegades que un jugador/a "resube" una aposta, és a dir puja una aposta anterior d'un jugador/a abans de veure el *flop*. Es pot realitzar per tal de robar el pot o per treure rendibilitat a una mà potent.

Jugador HERO realitza un 3 Bet

- ***Postflop Agg%:***

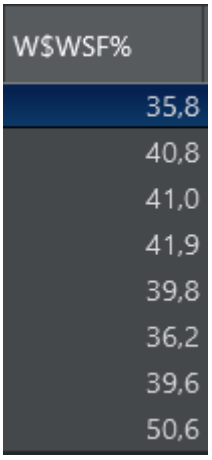
  És el percentatge (%) general de carrils agressius en tot el joc *postflop* d'un jugador/a a la vegada. L'*Agg*% de cada carril és el percentatge de veus agressives en un carril determinat. És a dir, senzillament ens dona la quantitat de vegades que un jugador/a aposta o “resube”, però no dona cap informació sobre quina és la seva freqüència de *call* o *check*.



Columna variable Postflop Agg%

- ***W$WSF%:***

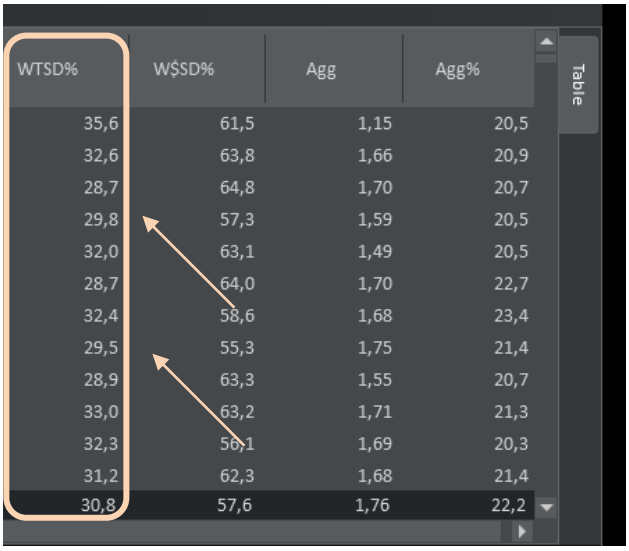
  És el percentatge (%) de vegades que un jugador/a guanya una mà una vegada has vist el *flop*, sense tenir en compte "com" has guanyat. Ens ajuda a saber amb quina freqüència el jugador/a roba el pot en el *postflop*.



Columna variable W$WSF%

- ***WTSD%:***

  És el percentatge (%) de vegades que un jugador/a arriba al *river (*5èna i última carta compartida) una vegada vist el *flop*. Podem saber doncs si un jugador/a és propens a abandonar ràpid la mà o persistir i confiar en la seva mà fins el final.



Columna variable WTSD%

- ***Won $ at SD:***

  És el percentatge (%) de vegades que un jugador/a guanya la mà quan s'arriba al *showdown **.

  **showdown:* part final de la mà, en la que els jugadors/es que queden, ensenyen les seves cartes i les comparen per veure qui té la millor mà i, per tant, guanya el pot acumulat. El *showdown* té lloc després de l'última ronda d'apostes.

Jugador en botó (delaer) guanya al showdown amb A8

- ***Squeeze:***

  És el percentatge (%) de vegades que un jugador/a fa una "resubida" després de que hagi hagut una pujada anterior i al menys un altre jugador/a hagi pagat a aquesta.

Jugador en UTG realitza moviment Squeeze

IV. ANÀLISI DESCRIPTIVA UNIVARIANT

En aquest apartat es pretén classificar, presentar, descriure, resumir i analitzar les dades relatives a cada variable en concret (una a una) per tal de tenir una idea prèvia sobre com estaran distribuïdes i quines característiques tindran les observacions de la nostra base de dades.

Variable SITE:

Ens indica a quina sala o plataforma de joc s'està jugant. En el nostre cas, només tenim 3 possibles sales de joc (*PokerStars, 888Poker,Winamax*) , que són les que enregistra el software *Holdem Manager 2*. A continuació, s'observa la taula de freqüències de la variable *Site*:

Site	Freq	FreqRel	%
2	99	0.0446	4.46%
12	2122	0.9554	95.54%
22	0	0.0000	0,00%

Taula 1. Taula de freqüències Variable SITE

Gràfic 1. Diagrama de sectors Variable SITE

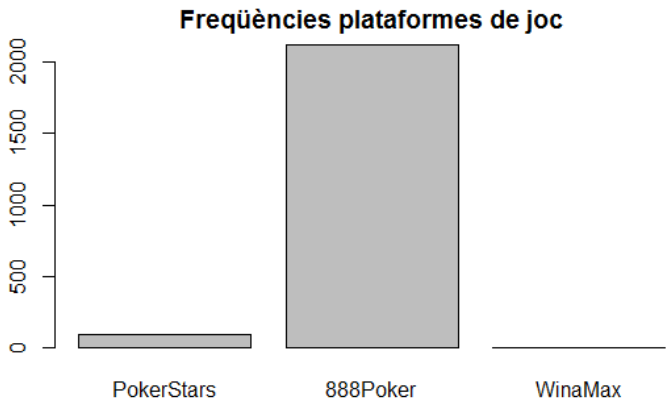
La variable *Site* presenta 3 nivells:

- **2:** Plataforma de joc és ***PokerStars***

- **12:** Plataforma de joc és ***888Poker***

- **22:** Plataforma de joc és ***WinaMax***

Es pot observar en el gràfic que el 95.54% (2122) dels jugadors/es contra els que s'ha jugat, s'ha produït en la plataforma *888Poker* ; un 4.46% (99) en la plataforma *PokerStars* i un 0,00% (0) a *WinaMax*. Això és degut a que quasi sempre s'ha jugat a la plataforma *888Poker* i per això hi han més oponents registrats en allà.



Gràfic 2. Diagrama de barres Variable SITE

Si es jugués més a qualsevol de les altres dues plataformes, aquestes dades canviarien.
Una vegada fet l'anàlisi descriptiu de la variable *Site*, es creu necessari eliminar aquelles dues plataformes on es tenen menys mans enregistrades per tal de fer un estudi més concret. Aquestes dues plataformes eliminades són *PokerStars* i *WinaMax* i; per tant, la base de dades quedarà reduïda a aquelles observacions que s'hagin enregistrat únicament en la plataforma *888Poker*.
Pel que ara, la base de dades d'estudi passarà de tenir 2221 observacions/jugadors a tenir-ne 2122. Haurem eliminat de la base de dades d'estudi 99 jugadors/es.

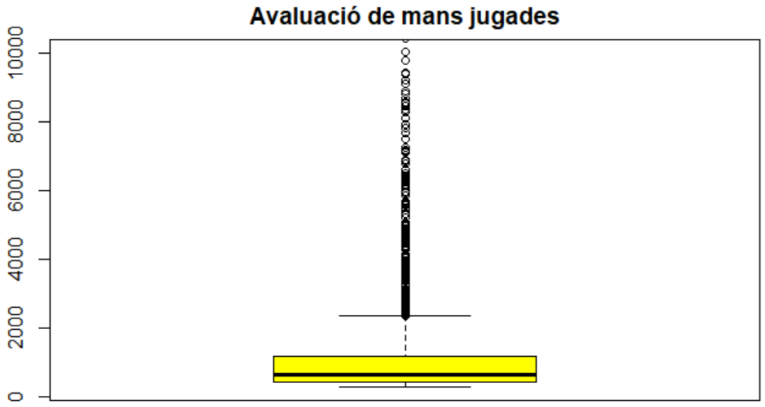
Variable HANDS:

Aquesta variable ens indica quantes mans hem jugat amb cadascú dels 2122 jugadors presents a la nostra base de dades. Observem el resum de la variable *Hands*:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
301.0	419.0	654.5	1262.4	1188.8	38643.0

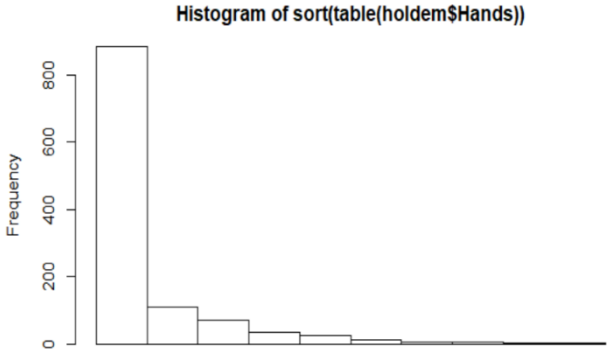
Taula 2. Summary Variable Hands

Té **desviació típica** (*sd*) = 2264.442



Gràfic 3. Boxplot Variable Hands

La variable *Hands* té una mitjana de 1262.4 mans jugades per jugador/a, aproximadament, però és molt distant del jugador/a màxim amb el que s'han compartit més mans en la taula (38643 mans).
Es pot observar en el *boxplot* que les mans jugades estan distribuïdes amb molta desviació, i molts valors atípics e *outliers*.



Gràfic 4. Histograma Variable Hands

Els calculem de la manera següent:

IQR = Q3-Q1 = 1188.8 - 419.0 = 769.8 (càlcul del rang interquartílic)

- Els **valors atípics lleus** estaran per sobre de:
  Q3 + 1.5 * IQR = 1188.8 + 1.5 * 769.8 = **2343.5**
- Els **valors atípics extrems** estaran per sobre de:
  Q3 + 3 * IQR = 1188.8 + 3 * 769.8 = **3498.2**
- Els ***outliers*** estaran per sobre de:
  Q1 + 1.5 * IQR = 419.0 + 1.5 * 769.8 = **1573.7**

Per tant, per la variable *Hands* hi hauran:

Valors atípics lleus	236
Valors atípics extrems	127
Outliers	372

No obstant la presència d'*outliers*, no es creu convenient eliminar-los o modificar-los ja que és una variable molt important que pot determinar la diferència entre aquells jugadors/es bons i els no tant bons.
Cal deixar clar, que quantes més mans haguem jugat contra un jugador, més fiables seran els valors de les seves variables i més precises les decisions a prendre.

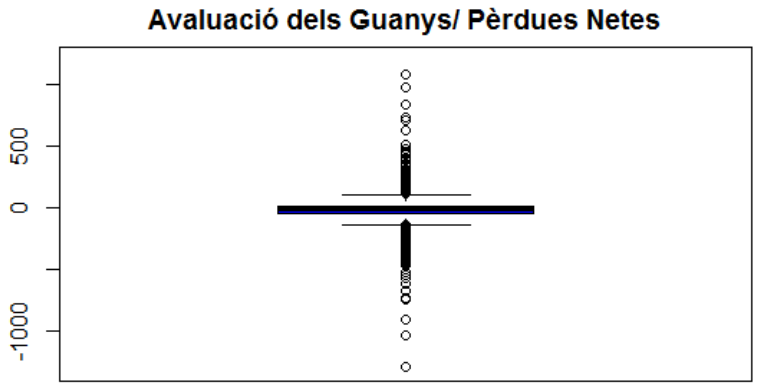
Variable NET WON:

Ens indica quins són els guanys nets o pèrdues netes dels jugadors de la nostre base de dades. Els valors estan expressats en euros (€).
Observem el resum de la variable *Net Won*:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-1293.03	-47.61	-10.36	-20,60	15.95	1080.77

Taula 3. Summary Variable Net Won

Té **desviació típica** (*sd*) = 119.1389

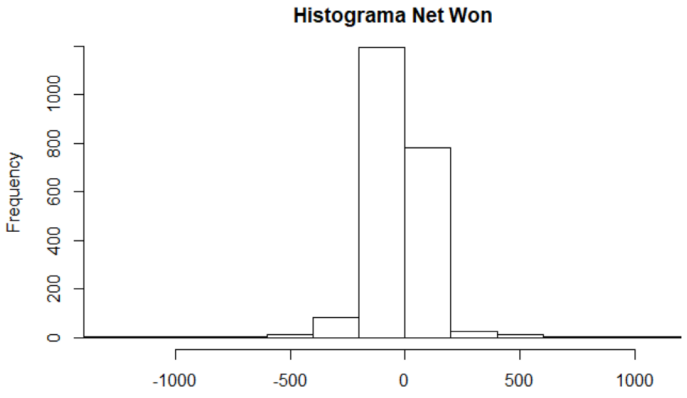


Gràfic 5. Boxplot Variable Net Won

D'entrada ja veiem com aquesta variable té valors positius i valors negatius.

Al parlar sobre la variable de guanys nets (*Net Won*) es pot observar com la mediana es situa vora el 0, distribuint-se les observacions de manera quasi simètrica entre els valors negatius (possibles pèrdues) i els valors positius (possibles guanys), però mirant l'histograma es veuen més pèrdues que guanys.

Veiem , en l'histograma, com les pèrdues són més abundats que no pas els guanys. De mitjana si nosaltres som guanyadors, gran part dels nostres adversaris seran perdedors ja que els hi estarem treien diners. Recordem que estem analitzant les dades de les accions que han fet aquells jugadors envers nosaltres, no envers els altres jugadors... Això és degut a dos motius principals:



Gràfic 6. Histograma Variable Net Won

- Tots els jugadors/es han de pagar obligatòriament i per cada mà un "*rake*", és a dir una comissió, a la sala on s'està jugant. Fet que decanta a un jugador a tenir pèrdues involuntàriament.

- La majoria de jugadors/es són perdedors/es per falta d'habilitat i coneixement.

Variable VP$IP:

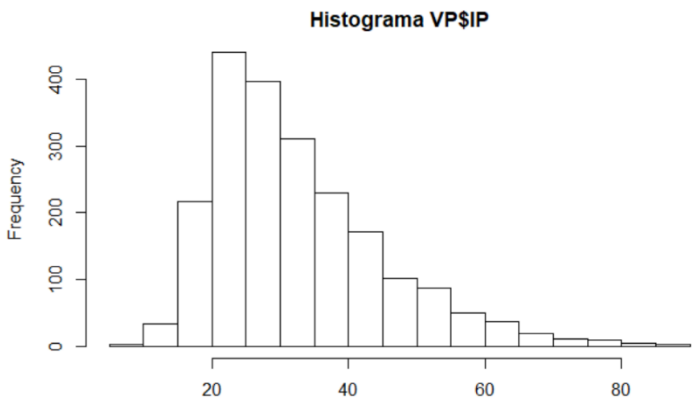
Indica quin percentatge de vegades els jugadors, introdueixen diners voluntàriament en el pot comú. Observem el resum de la variable *VP$IP*:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
5.044	23.481	29.645	32.523	39.191	89.474

Taula 4. Summary Variable VP$IP

Té **desviació típica** (*sd*) = 12.55827

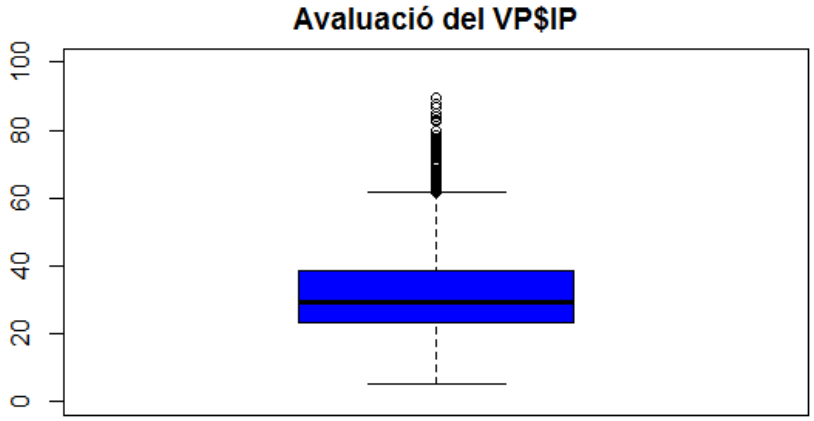
És observable com els jugadors/es registrats en la base de dades tenen un *VP$IP* mitjà de 32,523 %. Es pot dir que de mitjana el 32,523% dels jugadors/es introdueixen diners voluntàriament en el pot comú abans de veure el *flop*.



Gràfic 7. Histograma Variable VP$IP

Un percentatge bastant alt, tenint en compte que és una acció totalment voluntària, no són diners "obligats" a apostar com ho és el *rake* o comissió.

En l'histograma anterior s'observa com la majoria de jugadors/es tenen un *VP$IP* d'entre el 20% i el 40%.



Gràfic 8. Boxplot Variable VP$IP

En aquest *boxplot* es pot visualitzar que la variable *VP$IP* es distribueix de manera força simètrica, amb alguns valors atípics a la part superior.

Variable PFR:

Indicarà el percentatge de vegades que els jugadors/es han pujat la cega en la primera ronda d'apostes. Observem el resum de la variable *PFR*:

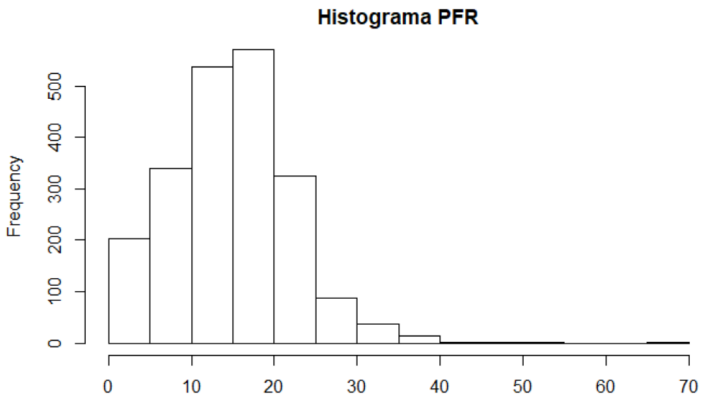
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.000	9.847	14.837	14.897	19.366	65.546

Amb **desviació típica** (*sd*) = 7.385453

Taula 5. Summary Variable PFR

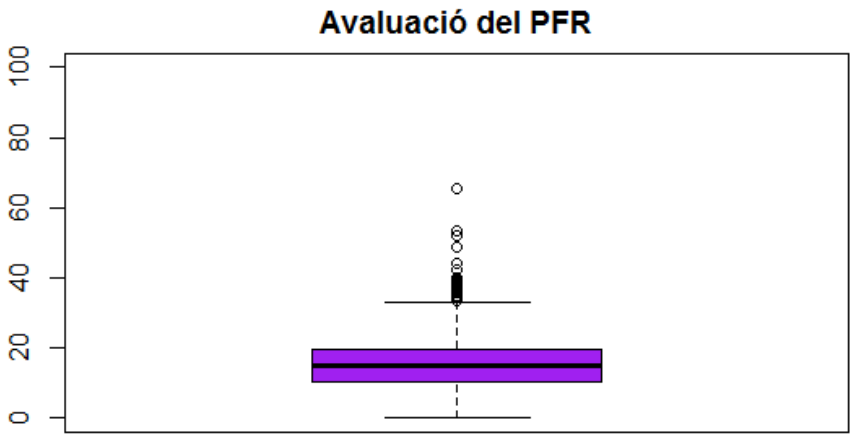
S'observa que els jugadors/es tenen un *PFR* mitjà de 14.90%, aproximadament.

Es pot observar, en l'histograma, que la freqüència de la majoria de jugadors/es es situa entre el 10% i el 20%. Més enllà del 40% de *PFR* es visualitzen molts pocs jugadors/es.



Gràfic 9. Histograma Variable PFR

En l'anàlisi descriptiu segregant, que es realitzarà més endavant, veurem quin serà el percentatge teòric d'aquells jugadors/es guanyadors pel *PFR*.



Gràfic 10. Boxplot Variable PFR

Es visualitza en el *boxplot* que la variable *PFR* té poca dispersió. Tot i que té algun valor extrem observable en la part superior, la majoria d'ells estan concentrats entre el 40% i 70% del *PFR*. Aquests jugadors/es extrems seran perdedors/es, en l'àmbit del pòquer se'ls anomena **recreacionals o *fish*** i són molt comuns.

Variable 3BET:

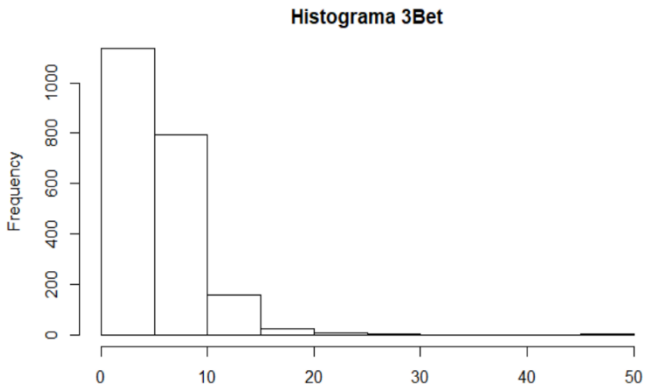
Indica quin percentatge de vegades els jugadors pugen una aposta anterior (*resubir*) feta per un jugador, abans de veure el *flop*. Observem el resum:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.000	2.586	4.684	5.260	7.160	45.631

Taula 6. Summary Variable 3Bet

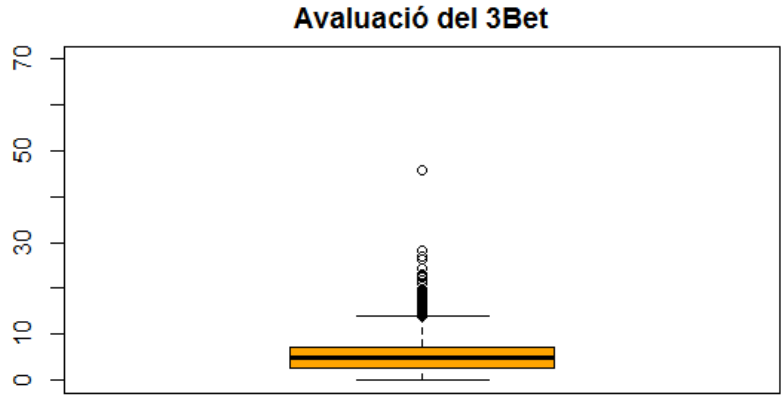
Amb **desviació típica** (*sd*) = 3.74928

No és sorprenent que els jugadors/es només tornin a pujar una aposta (*resubir*) per sobre de la pujada d'un jugador/a anterior abans del *flop* un 5,26% de mitjana. Es sol entendre que una pujada d'un jugador/a significa que porta una bona mà, però qui sap si està jugant com a *farol*...



Gràfic 11. Histograma Variable 3Bet

Els jugadors/es guanyadors/es fan *3Bet* amb bones mans però també juguen amb un rang de mans de *farol* per tal que la seva mà o la seva manera de jugar no sigui previsible pels oponents; i així poder, també, robar els diners acumulats en el pot comú abans de veure el *flop.* Es pot intuir que per aquesta variable (*3Bet*) els percentatges teòrics guanyadors no seran superiors al 20%. Ho comprovarem més endavant.



Gràfic 12. Boxplot Variable 3Bet

S'observa que la caixa del *boxplot* és força simètrica, amb valors petits però amb algun valor extrem en la part superior. Hi ha un en concret (el valor màxim = 45.63%) que és excessivament alt per la variable *3Bet*. Aquest jugador/a molt probablement sigui perdedor/a.

Variable POSTFLOP AGG%:

Indica el percentatge de carrils agressius en tot el joc *postflop* que tenen els jugadors/es. Observem el resum de la variable *Postflop Agg%:*

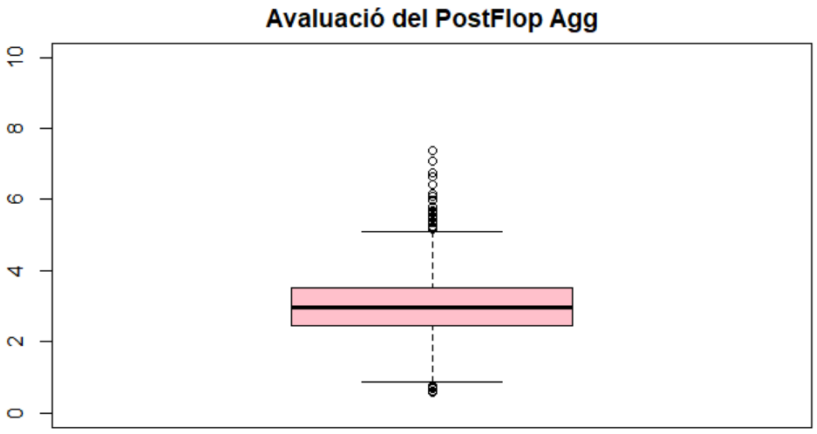
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.569	2.446	2.959	3.010	3.539	7.386

Taula 7. Summary Variable Postflop Agg%

Amb **desviació típica** (*sd*) = 0.8579544

S'obté que els jugadors/es tenen de mitjana un 3.01% de vegades que són agressius en els carrils *postflop*. Sabent que el *Postflop Agg%* ens dona la quantitat de vegades que un jugador/a aposta o puja una posta anterior (*resubir)* i, que no dona cap informació sobre la freqüència de *call* o *check*, es pot dir que no és estrany tenir una mitjana del 3,01%.

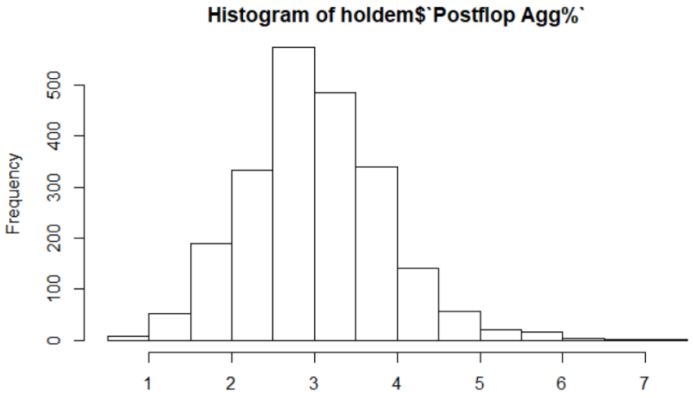
Sol estar entre el 1% i el 3%, així que la base de dades d'estudi ens ha donat un *Postflop Agg%* més aviat alt, però no preocupant perquè hi ha jugadors/es molt agressiu en el nivell NL10.



Gràfic 13. Boxplot Variable Postflop Agg%

Es pot observar, en el *boxplot*, una variable força simètrica i homogènia entre el 1% i 5%. Hi ha un valor màxim d'un 7,386% i, per tant, es té poca dispersió entre les dades de la variable *Postflop Agg%*. Es tracta d'una variable amb poca desviació típica, i conseqüentment poca variància.

En l'histograma es torna a observar la simetria comentada anteriorment, aparentant una distribució semblant a la de la distribució Normal. La majoria de jugadors tenen un *Posflop Agg%* entre el 2.5% i el 3.5%.



Gràfic 14. Histograma Variable Postflop Agg%

Variable W$WSF%:

Fa referència al percentatge de vegades que els jugadors guanyen la mà una vegada han vist el *flop*. El resum d'aquesta variable és:

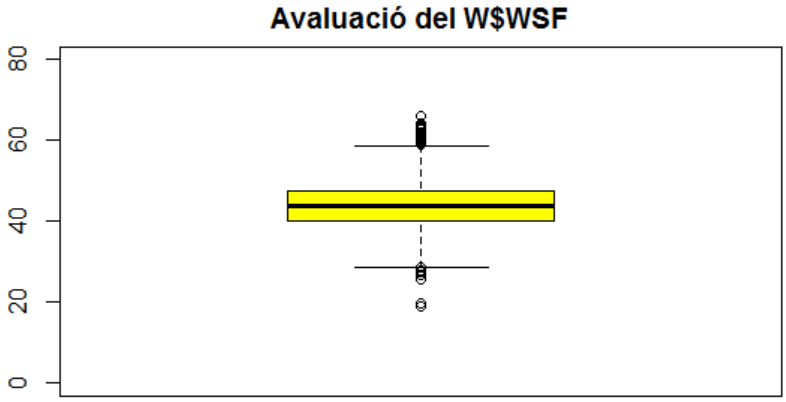
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
19.51	40.00	43.75	43.74	47.42	66.14

Taula 8. Summary Variable W$WSF%

Amb **desviació típica** (*sd*) = 5.843965

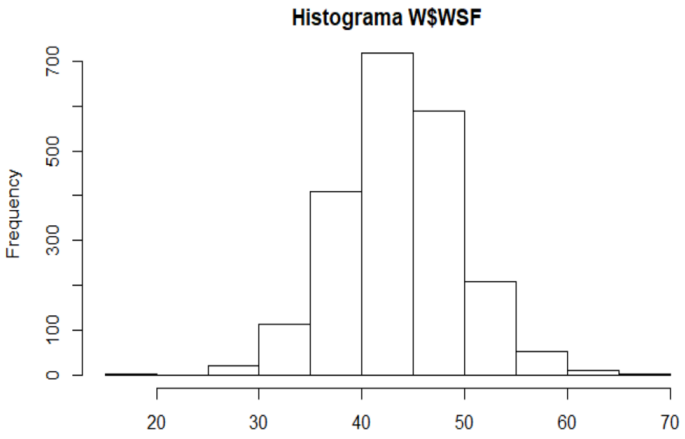
Els jugadors/es guanyen la mà quan han vist el *flop* un 43.74% de mitjana.
Es pot dir que és un percentatge raonable, perquè significa que una vegada vist el *flop* segueixen "confiant" en la seva mà fins el final, on aconsegueixen guanyar el pot comú amb les seves cartes.

No obstant, no és bo treure conclusions a priori sobre la diferència entre jugadors/es guanyadors/es i perdedors/es per aquesta variable, ja que dintre d'aquest 43.74% també es poden trobar jugadors/es perdedors/es, que a simple vista no són identificables.



Gràfic 15. Boxplot Variable W$WSF%

En ambdós plots s'observa una variable força simètrica en quan a com es distribueixen les seves dades.
Presència d'alguns valors extrems en els seus extrems, tant superior com inferior, però es tracta d'una variable amb observacions bastant homogènies.



Gràfic 16. Histograma Variable W$WSF%

Variable WTSD%:

Fa referència al percentatge de vegades que els jugadors/es arriben fins el *river*, una vegada han vist el *flop*. Observem el resum d'aquesta variable:

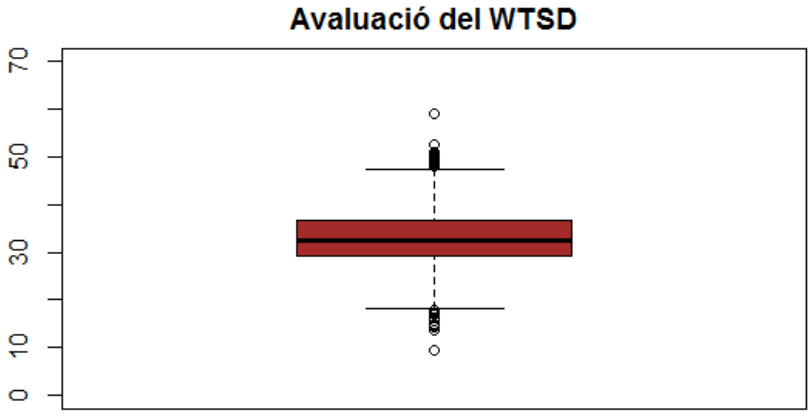
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
9.524	29.333	32.573	33.004	36.645	59.041

Taula 9. Summary Variable WTSD%

Amb **desviació típica** (*sd*) = 5.845999

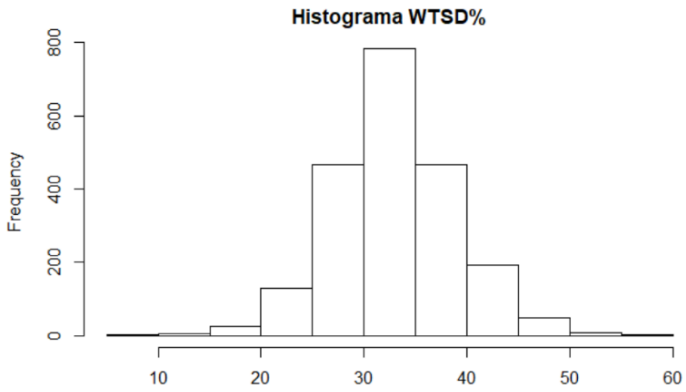
S'obté un 33% de mitjana del *WTSD%.* És a dir, un 33% de vegades que els jugadors/es de la base de dades arriben al *river* una vegada vist el *flop*. És un percentatge una mica alt pel que acostuma a ser (vora el 26%), però no és preocupant ja que està dintre del % guanyador, com comprovarem més endavant en la segregació. Hi ha jugadors/es que no els hi agrada abandonar, però els jugadors/es guanyadors/es acostumen a saber abandonar a temps

quan veuen que podria ser que no acabessin guanyant la mà. En canvi, hi ha molts jugadors/es (majoritàriament perdedors/es) que segueixen i segueixen apostant fins al final sense parar-se a pensar si val la pena seguir pagant o és millor abandonar.



Gràfic 17. Boxplot Variable WTSD%

Les dades es distribueixen de forma simètrica vora la mediana, com es pot observar en ambdós gràfics. La freqüència més alta de les nostres dades per aquesta variable es situa entre el 30% i el 35%, seguint aparentment una distribució Normal.



Gràfic 18. Histograma Variable WTSD%

Variable WON $ AT SD:

Fa referència al percentatge de vegades que els jugadors/es guanyen la seva mà quan s'arriba al *showdown*. Observem el resum d'aquesta variable:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
19.35	42.65	48.21	48.30	53.49	89.47

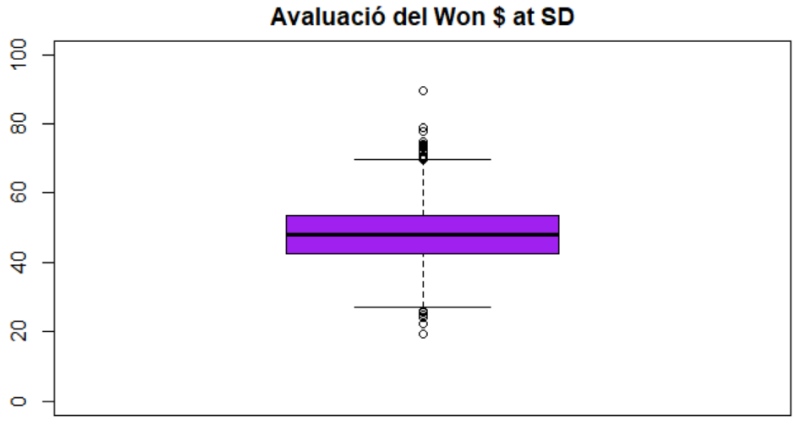
Taula 10. Summary Variable Won $ at SD

Amb **desviació típica** (*sd*) = 8.473874

Un 48.30% de mitjana els jugadors/es guanyen la mà quan arriben al *showdown*.

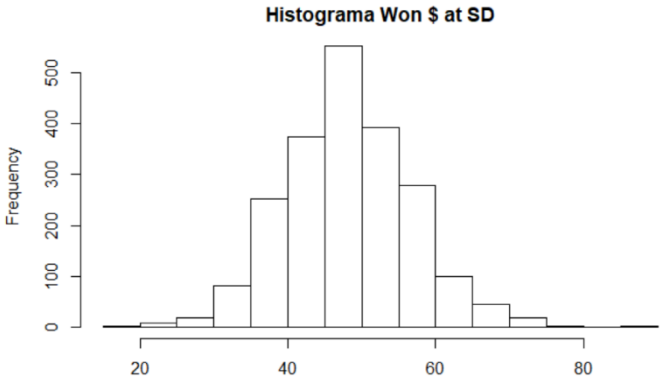
Quan es parla de *showdown,* ens referim a la part final de la mà, quan tots els jugadors/es implicats en la jugada ensenyen les seves cartes per veure quina ha estat la mà guanyadora.

És un percentatge bastant pobre, ja que s'entén que quan arribes al *showdown* és perquè hauries de tenir una mà molt propera a guanyar el pot comú. Però, com tenim molt jugadors/es perdedors/es d'aquí que el percentatge de *Won $ at SD* sigui més baix del que hauria de ser. Per tant, es deduïble que com més baix sigui el % de la variable *Won $ at SD*, aquest percentatge farà referència a jugadors/es perdedors, en la majoria d'ocasions.
Ho comprovarem, en l'apartat de segregació que realitzarem seguidament.



Gràfic 19. Boxplot Variable Won $ at SD

Podem observar ens aquests gràfics, com les observacions es distribueixen de forma simètrica i amb poca dispersió entre el 0% i 100%.
Tendeix a seguir una distribució Normal.



Gràfic 20. Histograma Variable Won $ at SD

Variable SQUEEZE:

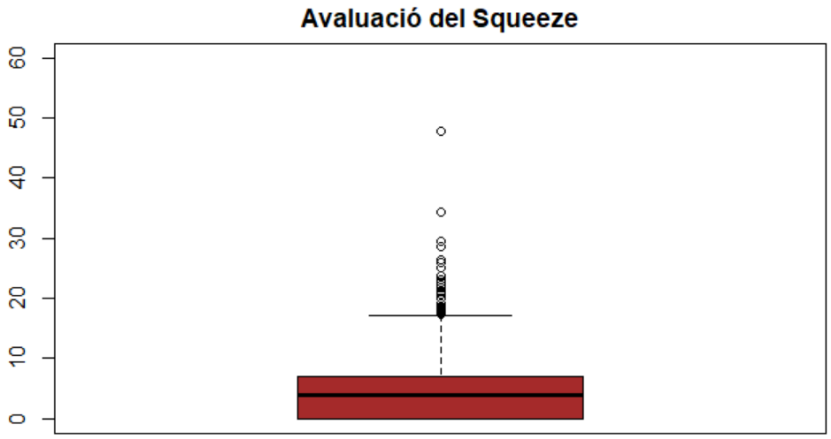
Indica el percentatge de vegades que els jugadors/es pugen una aposta feta prèviament, després de que hagi hagut una pujada anterior i al menys un altre jugador/a hagi pagat a aquesta. Observem el resum d'aquesta variable:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.000	0.000	4.000	4.792	6.891	47.826

Taula 11. Summary Variable Squeeze

Amb **desviació típica** (*sd*) = 4.79677

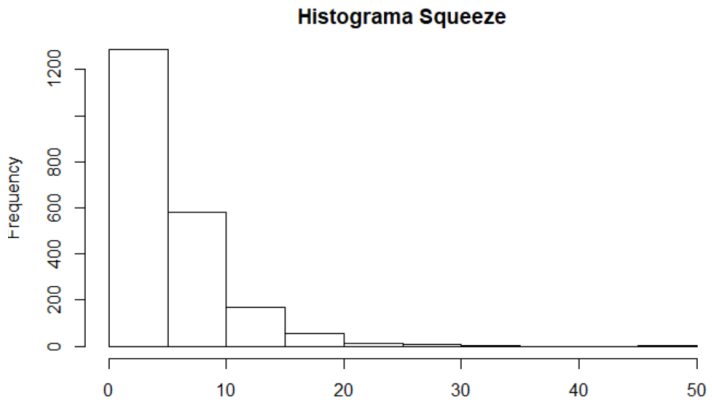
Només un 4.79% dels jugadors/es fan una aposta per sobre una aposta anterior (*resubida*) pagada per algun altre jugador/a. És comprensible, doncs quan ja hi han hagut 2 jugadors/es que han pagat una pujada i ve un altre que puja encara més el preu del pot comú, és evident que la gent es tiri de la mà pensant que aquell jugador/a porta possiblement una mà millor que la seva.
Aquest moviment, solen fer-ho aquells jugadors/es bons/es, que tenen experiència en el joc. Un excés en el percentatge d' *Squeeze* el solen tenir aquells jugadors més agressius.



Gràfic 21. Boxplot Variable Squeeze

Podem veure en el *boxplot*, que és una variable amb valors baixos, majoritàriament, però que té valors atípics i extrems en la part superior, però no sobrepassa el 48%.

Hi ha molts 0% i és degut a que l'*Squeeze* és el moviment més tècnic del joc, i molts jugadors no el contemplen.



Gràfic 22. Histograma Variable Squeeze

V. ANÀLISI DESCRIPTIVA UNIVARIANT SEGREGADA

En aquest apartat es farà una descriptiva univariant segregada, segregant cada variable pel ***Net Won***, i així visualitzar cada una de les variables de la nostra base de dades amb aquells jugadors/es guanyadors/es i, aquells jugadors/es perdedors/es.

Amb els *boxplots* i histogrames comparatius podrem anar treien conclusions sobre quines variables contribuiran més a crear estratègies guanyadores i quines a contribuir-ne de perdedores.
D'entrada, partim que totes les variables tindran dos grups amb mostres de:

- **n = 822** pels jugadors/es **guanyadors/es** (*Net Won* >= 0)
- **n = 1300** pels jugadors/es **perdedors/es** (*Net Won* < 0)

Abans de comparar ambdós grups, s'ha fet el test de normalitat *Shapiro-Wilk* per veure si les dades segueixen una distribució normal. Ens ha donat un *p-valor* inferior al 0.05 pel que hem rebutjat que les dades de la base de dades segueixin una distribució normal. Per tant, si més endavant es comparessin mitjanes faríem servir un test d'hipòtesis no paramètric, ja que no compleix el supòsit de normalitat de les dades.

Procedim doncs a analitzar cada variable per separat, segons la segregació feta:

Variable SITE:

Per aquesta variable no té sentit fer segregació ja que és una variable quantitativa amb un únic factor possible (*888Poker*), així que la diferenciació entre jugadors/es guanyadors/es i perdedors/es serà la mateixa que la de la variable *Net Won*.

Variable NET WON:

No té sentit segregar aquesta variable per ella mateixa, ja que és la variable que utilitzem per segregar la resta de variables de la nostra base de dades.

Variable HANDS:

Segregant la variable mans jugades, s'obtenen les següents dades:

- Jugadors/es guanyadors/es:

  sd= 2724.052

- Jugadors/es perdedors/es:

  sd= 1911.031

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
301.0	420.0	681.5	1434.3	1269.8	34819.0

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
301.0	416.8	637.5	1153.7	1117.0	38643.0

Taula 12. Summaries guanyadors i perdedors Variable Hands

Per començar, es pot dir que la mitjana de mans jugades per aquells jugadors que tenen un *Net Won* positiu és superior a la mitjana de mans jugades per aquells jugadors que tenen pèrdues. Aproximadament, els jugadors amb guanys juguen 281 mans més de mitjana que aquells que tenen pèrdues. És una xifra bastant significativa. Té sentit, si més no, que aquells jugadors/es amb guanys hagin jugat més mans que no pas els que tenen pèrdues. En el pòquer, quantes més mans es juguin millor s'entendrà el joc i millors decisions es prendran, així com més possibilitats de guanyar.

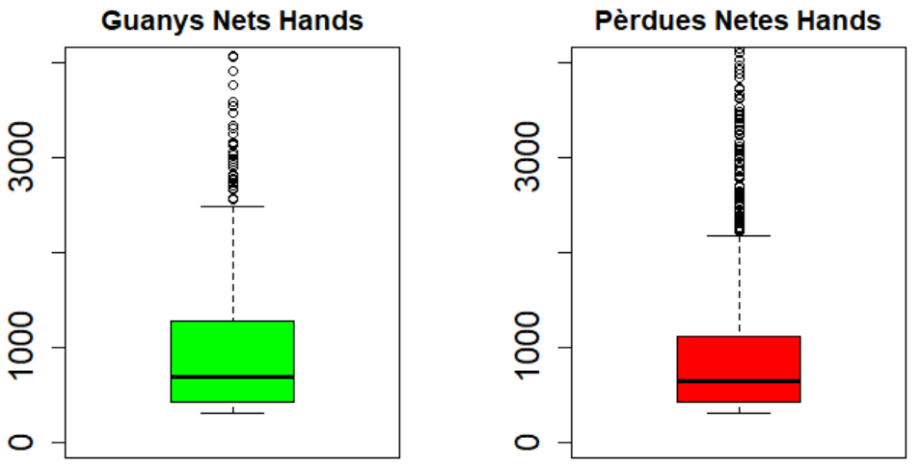
Per comparar aquestes dues mitjanes independents (perdedors/es i guanyadors/es) s'ha fet un test d'hipòtesis no paramètric *U Mann-Whitney-Wilcoxon*, ja que les dades no compleixen el supòsit de normalitat. Aquest test contrastará si dues mostres procedeixen de poblacions equidistribuïdes o no.

```
        Wilcoxon rank sum test with continuity correction

data:  g$Hands and p$Hands
W = 556370, p-value = 0.1085
alternative hypothesis: true location shift is not equal to 0
```

Taula 13. Test Wilcoxon Variable Hands

Com el p-valor del test és superior al 0.05 , ens indica que no hi ha evidències per considerar que la localització de les poblacions és diferent i conclou que les medianes d'ambdues poblacions són iguals. Per tant, la variable *Hands* no és significativa.

Seguidament es fa un *boxplot* per cada grup segregat:



Gràfic 23. Boxplots segregats Variable Hands

Dels *boxplots* anteriors se'n pot treure una interpretació similar, doncs les seves observacions estan distribuïdes de manera força similar però es pot observar com el *boxplot* dels guanys tindria la caixa superior més ample que la dels perdedors/es, per això també tenen una mitjana superior a la dels perdedors i perdedores.

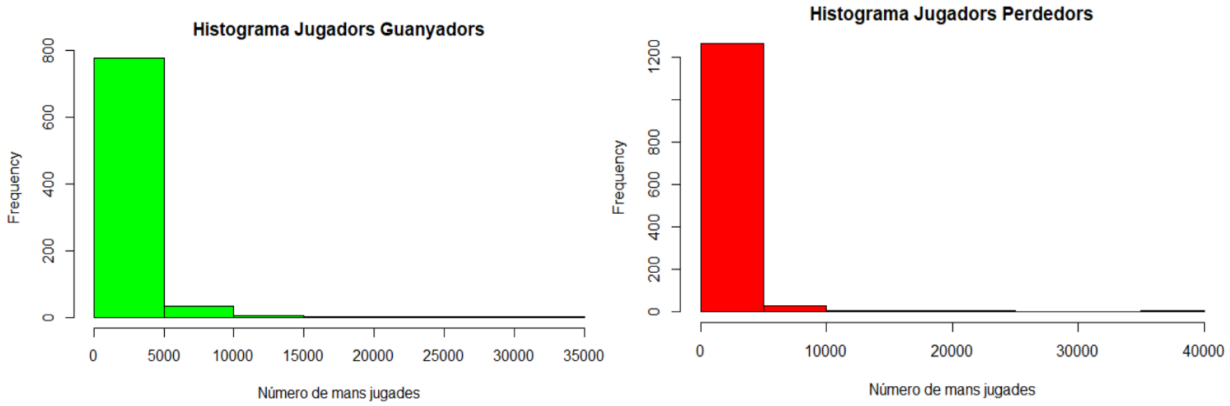
Com es sap que el valor màxim d'ambdós grups no supera les 40.000, es fa una taula i un histograma per grup, per visualitzar quines són les freqüències de mans jugades per cada grup:

	[0,5000)		[5000,10000)		[10000,20000)		[20000,30000)		[30000,40000)	
	F.AB	**%**	**F.AB**	**%**	**F.AB**	**%**	**F.AB**	**%**	**F.AB**	**%**
G	778	94.65%	34	4.14%	6	0.73%	2	0.24%	2	0.24%
P	1266	97.38%	26	2%	6	0.46%	1	0.08%	1	0.08%

Taula 14. Freqüències segregades de la Variable Hands

És visible com el 94.65% dels jugadors/es guanyadors/es i el 97.38% dels jugadors/es perdedors/es juguen menys de 5.000 mans. Només un 4.14% dels guanyadors/es i un 2% dels perdedors/es en juguen d'entre 5.000 i 10.000 mans. Un 0.73% dels guanyadors/es i un 0.46% en juguen d'entre 10.000 i 20.000 mans; i el 0.48% dels guanyadors/es i el 0.16% dels perdedors/es, en juguen més de 2.000.

No són sorprenent els resultats, doncs es suposa que un jugador bo /guanyador amb poques mans jugades ja hauria de tenir guanys i tenir molt marge d'avantatge en quan a guanys ($); i pel contrari un jugador-a dolent-a/perdedor-a amb poques mans que jugui ja tindria suficient per tenir pèrdues.



Gràfic 24. Histogrames segregats Variable Hands

En l'histograma es fa visible com la majoria de jugadors/es no superen les 10.000 mans. Aquells jugadors/es que superen les 10000 mans són jugadors/es que juguen amb molta freqüència i durant molta estona.

Variable VP$IP:

Segregant la variable *VP$IP*, s'obtenen els següents resums:

- Jugadors/es guanyadors/es:

  sd= 11.04516

- Jugadors/es perdedors/es:

  sd=13.26311

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Guanyadors	8.015	22.557	28.126	30.437	35.688	86.829
Perdedors	5.044	24.055	30.932	33.842	41.345	89.474

Taula 15. Summaries guanyadors i perdedors Variable VP$IP

Es pot veure com la mitjana dels jugadors/es amb pèrdues és superior a la dels jugadors/es amb guanys. D'entrada podem dir que té sentit, ja que aquells jugadors que no en saben gaire apostaran més diners voluntàriament al pot comú que no pas aquells que en saben més. Els jugadors/es "bons" només n'afegiran quan portin una bona mà o quan coneguin tant bé al seu adversari que sàpiguen que afegint diners voluntàriament el jugador/a s'acabarà tirant (*fold*).

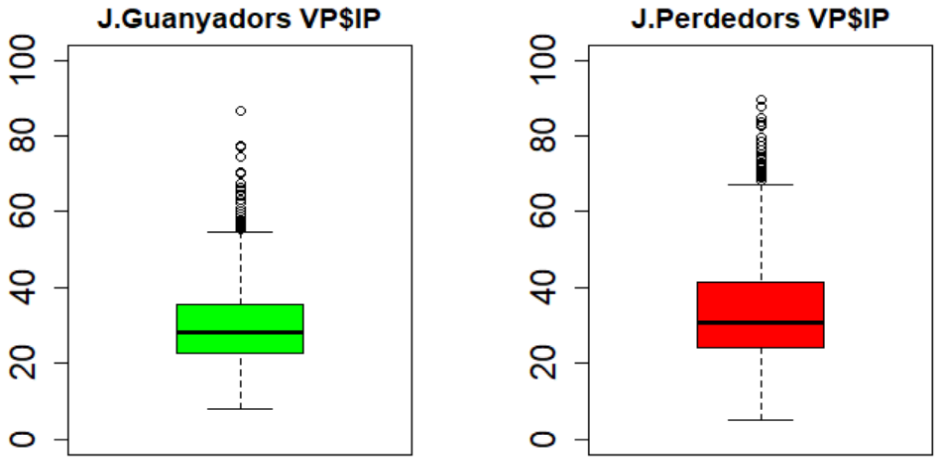
Realitzem el test d'hipòtesis no paramètric *U Mann-Whitney-Wilcoxon* per tal de comparar ambdues mitjanes:

```
        Wilcoxon rank sum test with continuity correction

data:  g$`VP$IP` and p$`VP$IP`
W = 455812, p-value = 1.141e-08
alternative hypothesis: true location shift is not equal to 0
```

Taula 16. Test Wilcoxon Variable VP$IP

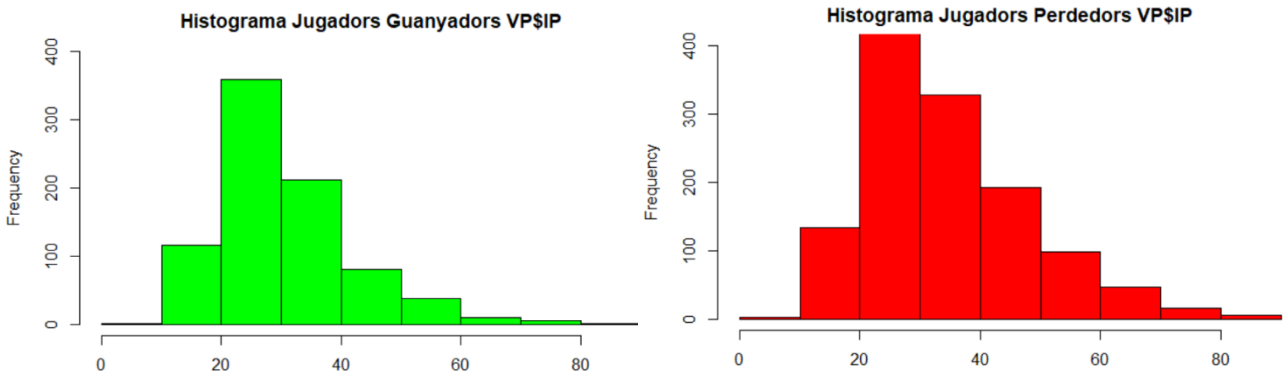
Com el *p-valor* del test és inferior al 0.05 , indica que sí hi ha evidències per considerar que la localització de les poblacions és diferent i conclou que les medianes d'ambdues poblacions són diferents. Per tant, la variable *VP$IP* és significativa.

Es sap que entre el 10% i el 35% del *VP$IP*, aproximadament, es trobarien els jugadors/es guanyadors/es. Per sota del 23% es trobaran els jugadors/es més guanyadors/es ja que els altres, tot i acabar resultant guanyadors/es o no, estaran jugant masses mans dèbils. Els jugadors/es "guanyadors/es" acostumen a jugar poques mans i, aquestes, bones.



Gràfic 25. Boxplots segregats Variable VP$IP

El *boxplot* d'aquells jugadors/es guanyadors té una aparença molt simètrica en quan a amplada de la caixa. No tant, simètrica és la dels perdedors/es.
Tenim una mitjana del 30.44% de l' *VP$IP* en aquells jugadors/es que guanyen, i una mitjana de 33.84% en aquells jugadors/es que perden. És a dir, que en mitjana els perdedors/es introdueixen, més vegades, diners voluntàriament en el pot comú.
Això es pot entendre com que els jugadors/es guanyadors, només introdueixen diners voluntàriament en el pot quan porten mans realment potents; en canvi els perdedors/es ho fan amb un rang més ampli de mans.



Gràfic 26. Histogrames segregats Variable VP$IP

En l'histograma es fa visible la gran diferència entre guanyadors/es i perdedors/es comentada anteriorment. Entre el 20% i el 50% de l'*VP$IP* d'ambdós gràfics resideix la diferència. Hi han moltíssims més jugadors per sobre del 30% en el grup dels perdedors/es. Aquests, amb menys experiència acostumen a afegir diners voluntàriament quan no porten mans gaire bones o contra adversaris que són bons i porten mans millors que les seves.

Variable PFR:

Segregant la variable *PFR*, s'obtenen les dades següents*:*

- Jugadors/es guanyadors/es:

  sd= 6.826438

- Jugadors/es perdedors/es:

  sd=7.692081

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.00	11.23	15.50	15.55	19.74	48.74

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.000	8.997	14.411	14.484	19.065	65.546

Taula 17. Summaries guanyadors i perdedors Variable PFR

Veiem com la mitjana de *PFR* és superior pel grup de guanyadors/es, però només un 1% més. Intuïm que tindran una distribució semblant tot i que el grup perdedor tindrà més dispersió ja que té jugadors amb valors màxims més elevats.
No obstant, aquesta petita diferència entre mitjanes, és normal veure que aquells jugadors amb guanys tenen un percentatge més alt de vegades que pugen la cega en la primera ronda d'apostes. Aquests, no només pugen la cega quan porten bones mans si no que també la pugen quan a la taula es situen jugadors que es tiren ràpid de la mà quan veuen que un altre jugador puja l'aposta feta anteriorment. I més, abans de visualitzar el *flop*. És una manera de jugar de "farol" fent PFR contra jugadors concrets.

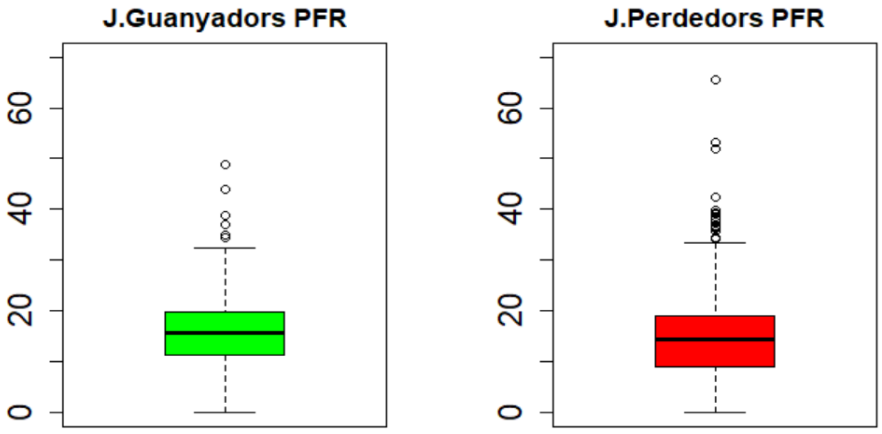
Es comparen les mitjanes utilitzant el test *U Mann-Whitney-Wilcoxon*:

```
        Wilcoxon rank sum test with continuity correction

data:  g$PFR and p$PFR
W = 586372, p-value = 0.0001524
alternative hypothesis: true location shift is not equal to 0
```

Taula 18. Test Wilcoxon Variable PFR

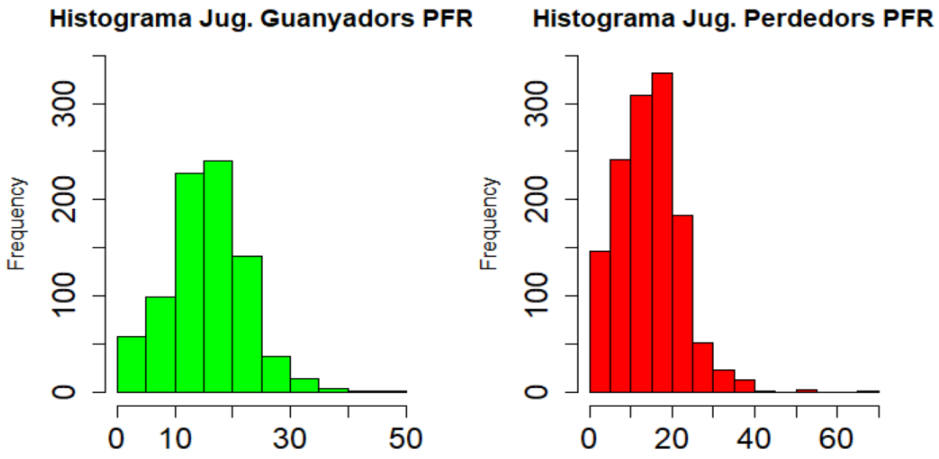
Com el *p-valor* del test és inferior al 0.05 , indica que sí hi ha evidències per considerar que la localització de les poblacions és diferent i conclou que les medianes d'ambdues poblacions són diferents. Per tant, el *PFR* és una variable significativa.

Fem els *boxplots* per cada grup en concret:



Gràfic 27. Boxplots segregats Variable PFR

Es pot corrobora les hipòtesis fetes anteriorment, tot i tenir una aparença simètrica, es veu com l'amplada de la caixa en el grup perdedor és superior i com, també, té valors extrems que fan que hi hagi més dispersió per aquest grup que no pas pel guanyador.



Gràfic 28. Histogrames segregats Variable PFR

Es fa visible la gran diferència en quant a freqüències absolutes. Entre el 15% i el 25% de *PFR* tenen moltes més observacions el grup perdedors que el guanyador.
Tenint en compte que parlem sempre del nivell *NL10*, els jugadors/es que tenen un *PFR* superior al 18% no solen ser guanyadors/es, degut a una excessiva agressivitat; i, els que el tenen inferior al 10% resulten ser perdedors/es a causa de tenir poca agressivitat i falta d'iniciativa. No obstant, hi ha jugadors/es amb un percentatge superior al 18% que acabarien sent guanyadors/es.

Per entendre com es classificarien els jugadors/es, segons el seu *PFR*, sense tenir en compte el nivell en el que es juga, obtindríem la següent classificació:

- **[0% , 18%) :** jugadors/es conservadors/es o que igualen la cega sense pujar amb massa freqüència.
- **[18%, 25%) :** jugadors/es normals i sòlids/es que juguen un número de mans apropiades.
- **[25% , 30%) :** jugadors/es agressius/ves que no eviten el conflicte i que intenten pressionar als seus rivals intentant robar molts pots i fent pujades a apostes anteriors amb mans marginals.
- **[30% , 100%] :** jugadors/es que juguen masses mans.

Veiem com per nivells superiors el % guanyador podria augmentar fins el 25%, però pel nivell que tractem en el nostre cas (*NL10*) el percentatge guanyador quedaria delimitat pel 18%.

Variable 3BET:

Segregant la variable *3Bet*, s'obtenen els següents resultats*:*

- Jugadors/es guanyadors/es:

  sd= 3.302908

- Jugadors/es perdedors/es:

  sd=4.006458

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.000   3.035   4.965   5.335   7.299  22.656
```

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.000   2.376   4.460   5.213   7.051  45.631
```

Taula 19. Summaries guanyadors i perdedors Variable 3Bet

Fixant-nos en la mitjana d'ambdós grups, s'obtenen mitjanes semblants vora el 5%, que és un percentatge bastant raonable per aquesta variable. Però intuïm, mirant el valor màxim de cada grup, que els perdedors/es tindran una dispersió major.

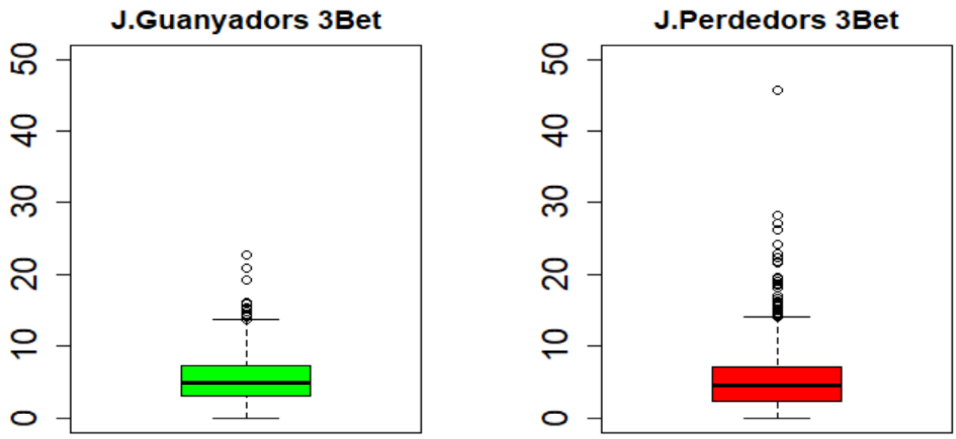
Comparem, doncs, ambdues mitjanes:

```
        Wilcoxon rank sum test with continuity correction

data:  g$`3Bet` and p$`3Bet`
W = 569283, p-value = 0.01095
alternative hypothesis: true location shift is not equal to 0
```

Taula 20. Test Wilcoxon Variable 3Bet

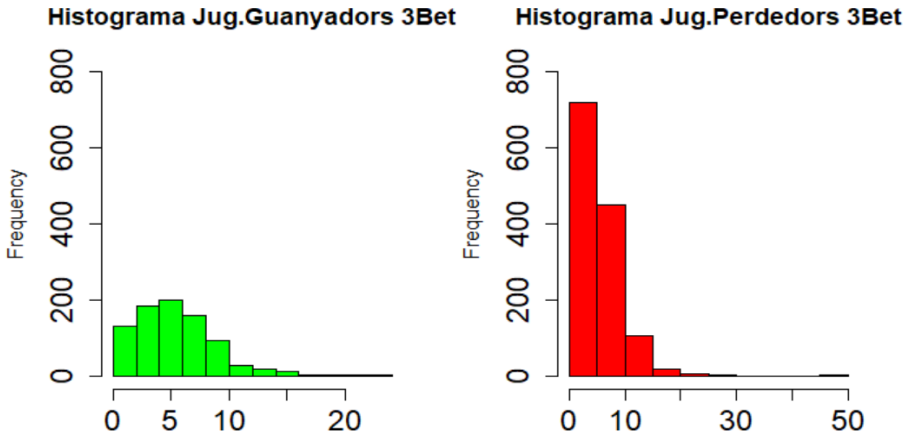
Com el p-valor del test és inferior al 0.05 , indica que sí hi ha evidències per considerar que la localització de les poblacions és diferent i conclou que les medianes d'ambdues poblacions són diferents. Per tant, la variable *3Bet* és significativa.

Sabem que el percentatge de *3Bet* d'aquells jugadors/es guanyadors/es hauria d'estar entre el 3% - 10%. Per sobre del 10% significaria que el jugador/a està pujant una aposta anterior amb masses mans i per sota del 3% no està fent un *3Bet* suficient.



Gràfic 29. Boxplots segregats Variable 3Bet

Mirant els *boxplots* anteriors és fa visible com el grup de jugadors/es perdedors/es és superior ja que té jugadors amb valors massa elevats.



Gràfic 30. Histogrames segregats Variable 3Bet

A partir dels histogrames, arribem a la conclusió de que la majoria de jugadors/es perdedors/es fan un 3Bet d'entre 0 i 15%. En canvi, el percentatge de *3Bet* en el grup guanyador està bastant equilibrat en quan a freqüències entre el 0% i el 10%.

Variable POSTFLOP AGG%:

Segregant la variable *Postflop Agg%*, s'obtenen els següents resultats*:*

- Jugadors/es guanyadors/es:
  sd= 0.7915806
- Jugadors/es perdedors/es:
  sd=0.8935469

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1.125	2.581	3.047	3.095	3.609	6.108

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.569	2.363	2.901	2.956	3.486	7.386

Taula 21. Summaries guanyadors i perdedors Variable Postflop Agg%

El percentatge guanyador per aquesta variable sol estar entre l' 1% i el 3%, així que la base de dades d'estudi ens ha donat un *Postflop Agg%* més aviat alt, però no preocupant perquè hi ha jugadors/es molt agressius en el nivell *NL10*. D'aquí que es tinguin més jugadors/es perdedors que guanyadors en la base de dades (*holdem*).
En aquest cas, la mitjana és superior pel grup guanyador. És a dir, que en mitjana, els jugadors/es guanyadors són més agressius després de veure el *flop* que no pas els perdedors. Això es pot interpretar com que, una vegada han vist el *flop* corroboren que porten bona mà (o no) i aposten d'una manera agressiva per tal de que els altres jugadors involucrats en la jugada s'acabin retirant.

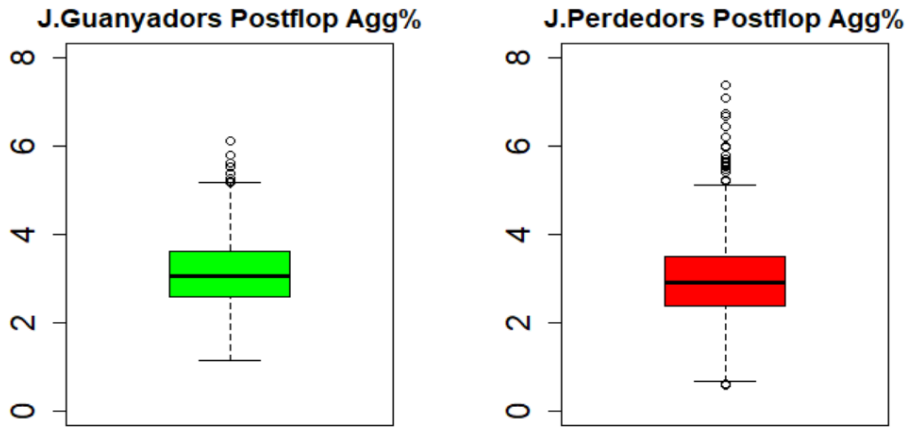
Compararem les mitjanes dels jugadors/es perdedors/es i la dels guanyadors/es:

```
        Wilcoxon rank sum test with continuity correction

data:  g$`Postflop Agg%` and p$`Postflop Agg%`
W = 592360, p-value = 2.415e-05
alternative hypothesis: true location shift is not equal to 0
```

Taula 22. Test Wilcoxon Variable Postflop Agg%

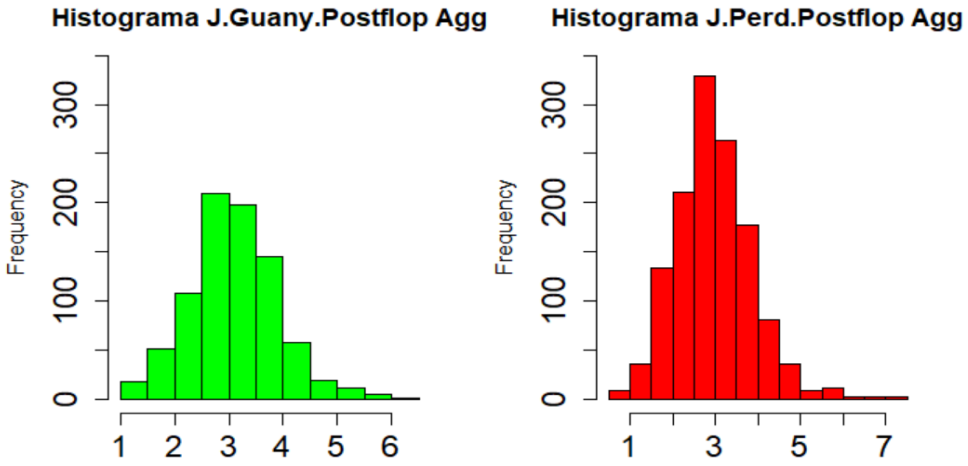
Com el p-valor del test és inferior al 0.05 , indica que sí hi ha evidències per considerar que la localització de les poblacions és diferent i conclou que les medianes d'ambdues poblacions són diferents. Per tant, aquesta variable és significativa.

Seguidament, obtenim els dos *bloxplots* segregats:



Gràfic 31. Boxplots segregats Variable Postflop Agg%

Dels *boxplots* es pot comentar que tenen una distribució molt semblant, tot i que el grup perdedor té valors extrems molt superiors però sense sobrepassar el 8%.



Gràfic 32. Histogrames segregats Variable Postflop Agg%

En l'histograma es torna a fer visible com el repartiment dels jugadors, dintre els límits teòrics de la variable, és més equitatiu en el grup guanyador que no pas en el perdedor.

Variable W$WSF%:

Segregant la variable *W$WSF%*, s'obtenen els següents resultats*:*

- Jugadors/es guanyadors/es:

  sd= 5.670259

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
28.39	42.05	45.45	45.64	48.85	65.96

- Jugadors/es perdedors/es:

  sd=5.630293

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
19.51	38.84	42.41	42.54	46.24	66.14

Taula 23. Summaries guanyadors i perdedors Variable W$WSF%

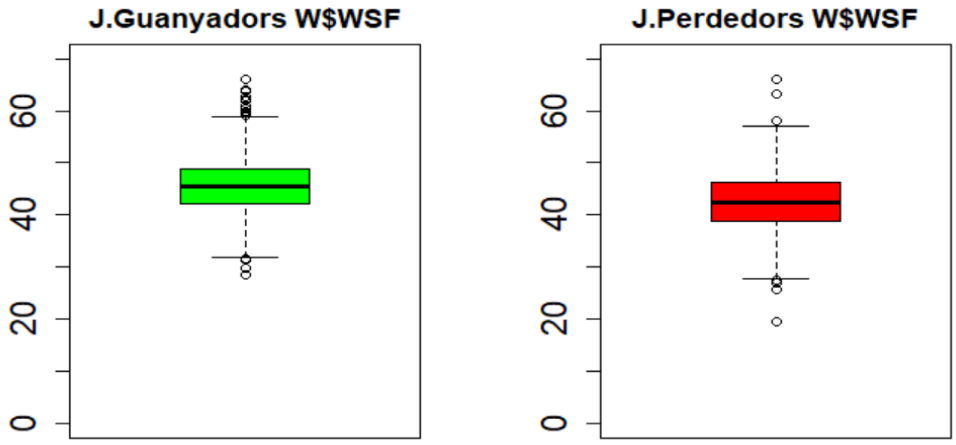
Aquesta variable és bastant curiosa ja que ens indica en quin percentatge els jugadors guanyen la mà, una vegada han vist el *flop*. Sense tenir molta idea, el grup guanyador hauria de tenir un percentatge superior i; veiem com és així. De mitjana els jugadors/es amb guanys nets, guanyen la mà un 46% de les vegades, aproximadament, una vegada vist el *flop* i els perdedors ho fan en un 43% de mitjana, aproximadament. Tot i ser, el percentatge mitjà superior en el grup guanyador és curiós veure que els perdedors tenen un percentatge decent per aquesta variable.

Compararem les mitjanes dels dos grups segregats:

```
        Wilcoxon rank sum test with continuity correction

data:  g$`W$WSF%` and p$`W$WSF%`
W = 696003, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

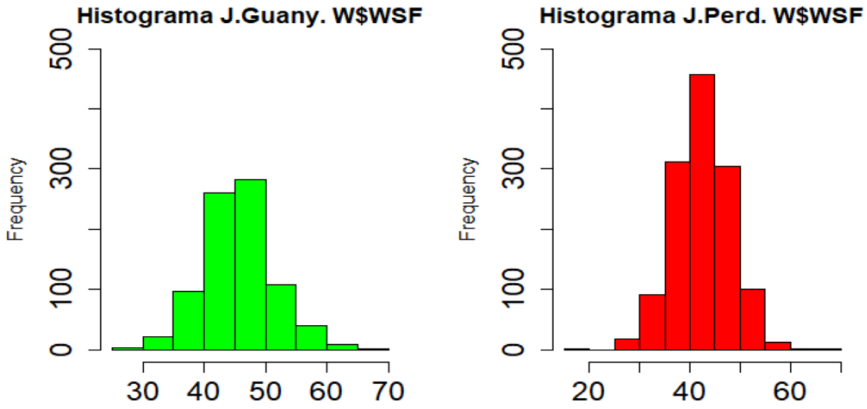
Taula 24. Test Wilcoxon Variable W$WSF%

Com el *p-valor* del test és inferior al 0.05 , indica que sí hi ha evidències per considerar que la localització de les poblacions és diferent i conclou que les medianes d'ambdues poblacions són diferents. Per tant, es tracta d'una variable significativa.
Els límits en percentatge del *W$WSF%* per aquells jugadors/es guanyadors hauria d'estar entre el 30% i el 45%. Realitzem els dos *boxplots* segregats, per tal de visualitzar possibles diferències entre grups:



Gràfic 33. Boxplots segregats Variable W$WSF%

En el *boxplot* dels jugadors/es amb guanys es pot veure com la caixa està situada per sobre del 40%, en canvi pels jugadors/es amb pèrdues està entre el 30% i el 50%. Es pot comentar que, en aquest cas, el grup amb més valors extrems és el guanyador. I no és estrany ja que dintre dels guanyadors es trobaran jugadors que guanyin més del 60% de les seves mans una vegada han vist el *flop*.
Aquests jugadors acostumen a ser els anomenats "piedra", que es caracteritzen per només pagar o apostar quan porten bones mans sense fer farols gairebé mai.



Gràfic 34. Histogrames segregats Variable W$WSF%

Veiem com entre el 40% i el 45% es troben la majoria d'observacions dels jugadors/es perdedors, en canvi entre el 45% i el 50% es troben la dels guanyadors/es.

Variable WTSD%:

Segregant la variable *WTSD*, s'obtenen els següents resultats*:*

- Jugadors/es guanyadors/es:

  sd= 5.456267

- Jugadors/esperdedors/es:

  sd=6.037897

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
13.64	28.77	32.00	32.29	35.74	50.00

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
9.524	29.653	33.005	33.457	37.382	59.041

Taula 25. Summaries guanyadors i perdedors Variable WTSD%

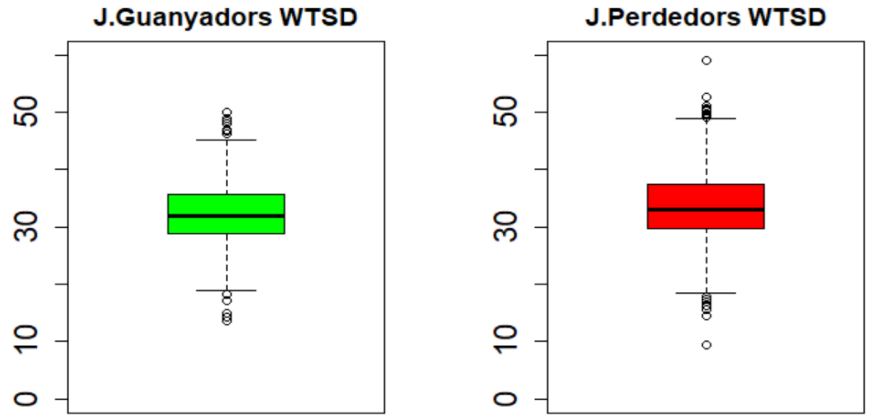
Els jugadors/es guanyadors solen estar entre el 25% i el 40%, pel que d'entrada sembla que ambdós grups tenen, de mitjana, valors dintre dels límits. El grup guanyador té una mitjana inferior (un 1.23% menys) que el perdedor.
Fem comparació d'ambdues mitjanes:

```
        Wilcoxon rank sum test with continuity correction

data:  g$`WTSD%` and p$`WTSD%`
W = 477103, p-value = 3.184e-05
alternative hypothesis: true location shift is not equal to 0
```

Taula 26. Test Wilcoxon Variable WTSD%

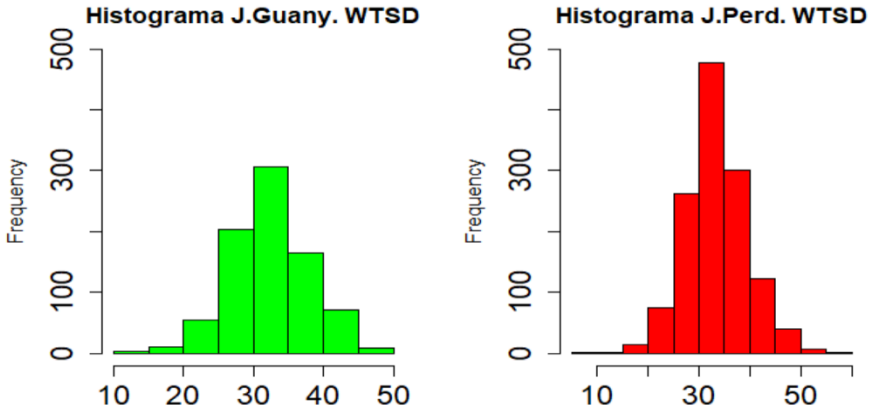
Com el *p-valor* del test és inferior al 0.05 , indica que sí hi ha evidències per considerar que la localització de les poblacions és diferent i conclou que les medianes d'ambdues poblacions són diferents. Per tant, aquesta variable és significativa.

Hi ha jugadors/es que no els hi agrada abandonar, però els jugadors/es guanyadors acostumen a saber abandonar a temps quan veuen que podria ser que no acabessin guanyant la mà. En canvi, hi ha molts jugadors/es (majoritàriament perdedors/es) que segueixen i segueixen apostant fins al final sense parar-se a pensar si val la pena seguir pagant o és millor abandonar.



Gràfic 35. Boxplots segregats Variable WTSD%

Veiem, en el *boxplot*, que tenen una distribució semblant tot i que la caixa dels perdedors/es és més ample i també són visibles punts extrems. Hi ha un en concret , pel grup perdedor, que és massa alt ja que supera el 50%.



Gràfic 36. Histogrames segregats Variable WTSD%

En ambdós histogrames es veu com la majoria d'observacions/jugadors-es es situen entre el 30% i el 35%.

Variable WON $ AT SD:

Segregant la variable *Won $ at SD*, s'obtenen els següents resultats*:*

- Jugadors/es guanyadors/es:

  sd= 7.935198

- Jugadors/es perdedors/es:

  sd=7.76892

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
30.53	47.15	52.08	52.36	57.14	89.47

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
19.35	40.44	45.83	45.72	50.73	78.95

Taula 27. Summaries guanyadors i perdedors Variable Won $ at SD

El percentatge guanyador per aquesta variable hauria d'estar per sobre del 55% aproximadament. Entre el 55% i el 70% per ser més exactes. El grup guanyador té un percentatge mitjà més elevat que el perdedor tot i que no arriba al 55%.

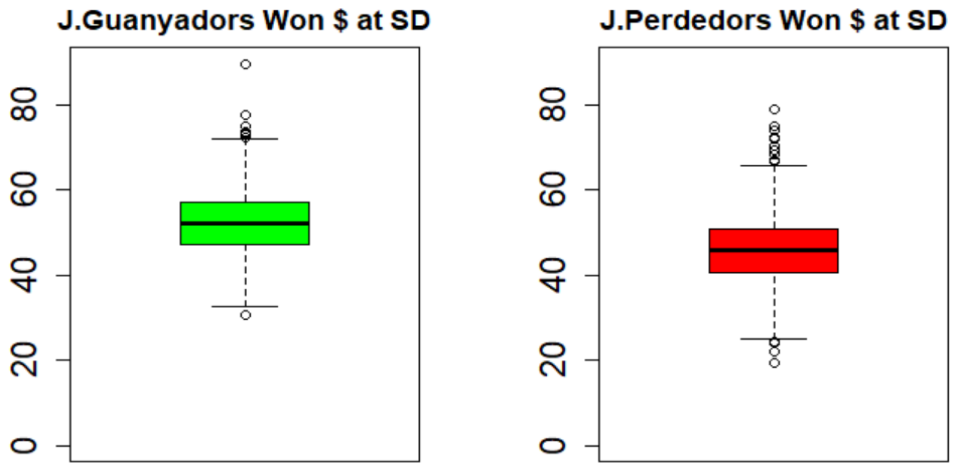
Comparem ambdues mitjanes:

```
        Wilcoxon rank sum test with continuity correction

data:  g$`Won $ at SD` and p$`Won $ at SD`
W = 777327, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

Taula 28. Test Wilcoxon Variable Won $ at SD

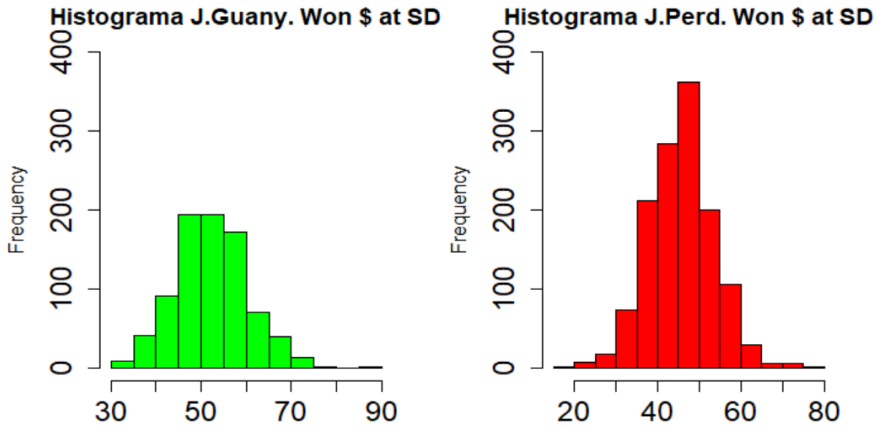
Com el *p-valor* del test és inferior al 0.05 , indica que sí hi ha evidències per considerar que la localització de les poblacions és diferent i conclou que les medianes d'ambdues poblacions són diferents. Per tant, aquesta variable és significativa.

Visualitzem els dos *boxplots* segregats:



Gràfic 37. Boxplots segregats Variable Won $ at SD

Es veu com la caixa dels guanyadors/es està una mica més elevada, tot i ser ambdós *boxplots* molt semblants. I, es pot observar un punt extrem en el grup guanyador per sobre del 80%. Aquesta observació pertany a un jugador/a que guanya més del 80% de les vegades que arriba al *showdown*, és a dir la gran majoria de vegades.



Gràfic 38. Histogrames segregats Variable Won $ at SD

En l'histograma es pot visualitzar que la majoria de jugadors/es guanyadors/es es troben entre el 45% i el 60% de les vegades que guanyen una vegada arriben al *showdown*. En canvi, en el grup guanyador la majoria d'observacions es troben entre el 40% i el 50%, un percentatge bo però no lo suficient bo que hauria de ser.

Variable SQUEEZE:

Segregant la variable *Squeeze*, s'obtenen els següents resultats*:*

- Jugadors/es guanyadors/es:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.000	1.498	4.271	4.827	6.742	28.571

sd= 4.354445

- Jugadors/es perdedors/es:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.000	0.000	3.774	4.770	6.977	47.826

sd=5.058041

Taula 29. Summaries guanyadors i perdedors Variable Squeeze

El percentatge d'aquells jugadors/es que resultarien guanyadors/es, hauria d'estar entre el 3% i el 8%, aproximadament, per aquesta variable. L'*Squeeze* és un moviment molt tècnic, el qual no tots els jugadors el contemplen o el saben efectuar quan cal.

Tots dos grups tenen una mitjana dintre del límit percentual guanyador (aproximadament un 5%) . No obstant, podem veure com el grup guanyador té un valor màxim molt més superior que el valor màxim del grup guanyador.

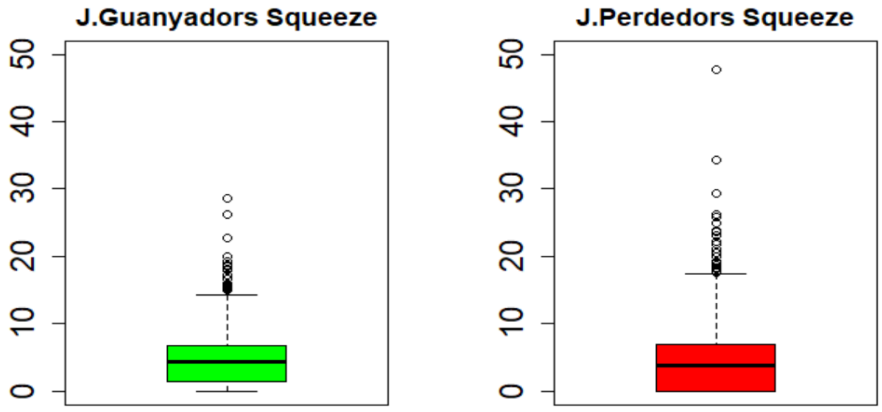
Comparem, doncs, ambdues mitjanes:

```
        Wilcoxon rank sum test with continuity correction

data:  g$Squeeze and p$Squeeze
W = 557653, p-value = 0.08662
alternative hypothesis: true location shift is not equal to 0
```

Taula 30. Test Wilcoxon Variable Squeeze

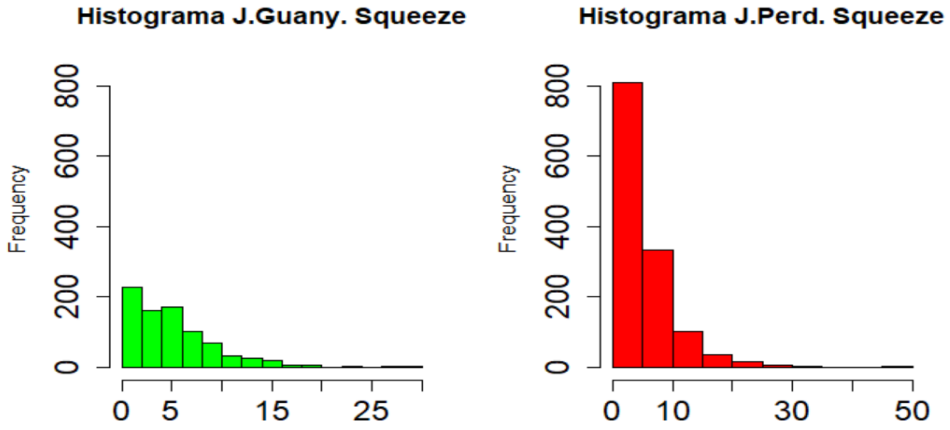
Com el *p-valor* del test és superior al 0.05 , indica que no hi ha evidències per considerar que la localització de les poblacions és diferent i conclou que les medianes d'ambdues poblacions són iguals. Per tant, aquesta variable és no significativa.

Observem els dos *boxplots* segregats:



Gràfic 39. Boxplots segregats Variable Squeeze

En aquests *boxplots* sí que s'observen més diferències. La caixa dels perdedors/es és més ample que la dels guanyadors/es i es veu com el valor mínim dels perdedors/es coincideix amb el primer quartil (0%), pel que intuïm que tindran més jugadors/es amb freqüència 0%. També són visibles molts més punts extrems amb percentatges superiors, en els jugadors/es perdedors/es.



Gràfic 40. Histogrames segregats Variable Squeeze

Ens reafirmem, a través de l'histograma perdedor, en el fet de que la majoria d'observacions tenen un 0% d'*Squeeze* fen evident la gran diferència entre un grup i l'altre.

RESUM VARIABLES SEGREGADES. SIGNIFICATIVES O NO SIGNIFICATIVES?

Per concloure l'apartat descriptiu segregat de les nostres variables d'estudi, fem un petit resum on es poden visualitzar quines variables ens han aparegut com a significatives i quines com a no significatives en quan a la seva segregació per la variable *Net Won*, que és la nostra variable resposta de l'estudi ...

VARIABLE	p-value		SIGNIFICATIVA?
Hands	0.1085	> 0.05	NO
VP$IP	1.141e-08	< 0.05	**SÍ**
PFR	0.0001524	< 0.05	**SÍ**
3Bet	0.01095	< 0.05	**SÍ**
Postflop Agg%	2.415e-05	< 0.05	**SÍ**
W$WSF%	2.2e-16	< 0.05	**SÍ**
WTSD%	3.184e-05	< 0.05	**SÍ**
Won $ at SD	2.2e-16	< 0.05	**SÍ**
Squeeze	0.08662	> 0.05	NO

Taula 31. Resum variables significatives

Per tant, tenim que totes són significatives, per a la segregació feta, excepte la variable número de mans jugades (*Hands*) i la variable *Squeeze*. Podem intuir doncs que aquestes dues variables no tindran gaire importància per a la determinació d'estratègies, tot i que òbviament com millor les controlem millor serà el nostre joc i majors seran els beneficis obtinguts.

VI. ANÀLISI DESCRIPTIVA BIVARIANT

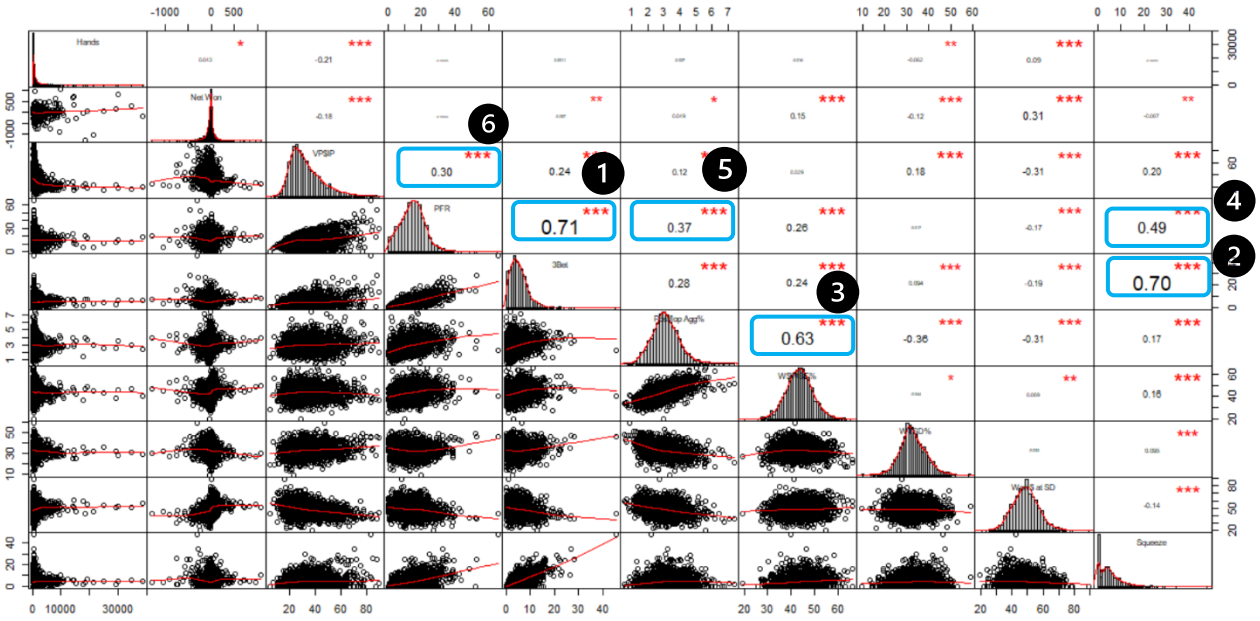
En aquest apartat, es realitzarà un anàlisi bivariant entre aquelles variables més rellevants per tal de comentar quina tendència lineal segueixen i veure si estan correlacionades o no. Només hem analitzat aquelles de caràcter numèric ja que la variable *Player* és l'*id* de la nostra base de dades i la variable *Site*, que és categòrica només ens hem quedat amb 1 factor dels 3 inicials.
Primerament es realitzarà un matriu de correlacions entre les variables, i aquelles que estiguin més correlacionades n'estudiarem el seu plot bivariant on distingirem quines observacions quedarien dintre d'aquells jugadors que resultarien guanyadors i quines resultarien perdedors, juntament amb les seves respectives rectes de regressió.
Comentant, també, on s'haurien de trobar els guanyadors per cada variable tractada si partíssim dels seus límits teòrics "guanyadors". Aquests, ja anomenats anteriorment per aquelles variables percentuals, són (aproximadament) :

- ***VP$IP :*** [10% , 35%]
- ***PFR:*** [10% , 18%]
- ***3Bet:*** [3% , 10%]
- ***Postflop Agg%:*** [1% , 3%]
- ***W$WSF:*** [30% , 45%]
- ***WTSD%:*** [25% , 40%]
- **Won $ at SD:** [55% , 70%]
- ***Squeeze:*** [3% , 8%]

Per fer aquesta diferenciació cal saber que no hi ha només una estratègia guanyadora. El joc evoluciona amb els anys i, conseqüentment, els jugadors també ho fan.
Les estratègies que hi havia en l'any 2010 avui dia resultarien estratègies perdedores i estratègies que avui dia són guanyadores, segurament en uns 10 anys deixaran de ser-ho per passar a ser perdedores.
El pòquer és un joc que requereix una adaptació continua on els jugadors/es pretenen acostar-se a l'estratègia GTO (*Game Theory Optimal*) amb els diferents programes informàtics disponibles.

Sabent això, s'ha decidit crear una nova columna (*jug*) a la base de dades amb els caràcters "G" i "P" que ens diferenciarà aquells jugadors amb *Net Won* > 0 (G) i aquells amb *Net Won* < 0 (P) directament. Amb aquesta columna ens serà més fàcil fer gràfics segregats i poder visualitzar possibles diferències en l'anàlisi bivariant corresponent.

Partim doncs de la següent matriu de correlacions entre les variables d'estudi (variables numèriques):



Gràfic 41. Matriu de correlacions entre variable numèriques

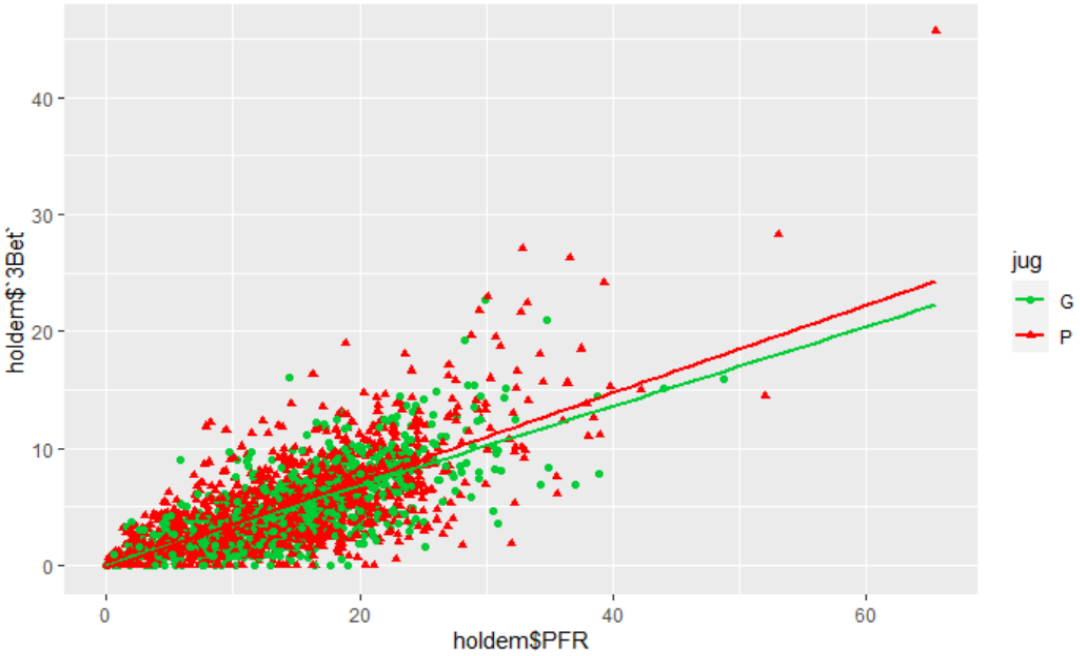
A la matriu de correlacions entre variables podem observar els valors de la correlació entre cada parell de variables (així com el seu nivell de significació representat per les estrelles) a la seva a la part superior dreta; els plots bivariants a la part inferior esquerra; i, els histogrames representatius de cada variable en la diagonal de la matriu.
Aquests histogrames fan visible com és la distribució de les observacions per a cada variable.
Els valors encerclats amb el requadre blau, són aquells coeficients de correlació més elevats entre variables. Aquestes parelles de variables correlacionades les estudiarem en detall a continuació, on procedirem a fer el seu anàlisi descriptiu bivariant segregat per la variable *Net Won*.

Tot i que ara les tractarem ja podem comentar que la variable *PFR* i *Postflop Agg%,* esdevindran variables importants pel nostre posterior estudi de clústers, ja que tenen més d'una variable amb correlació positiva i elevada.

Seguidament detallem i estudiem com és aquesta correlació lineal entre variables:

❶ PFR segons 3Bet:

S'estudiarà el *PFR* dels jugadors analitzats segons el seu *3Bet* mitjançant un diagrama de dispersió segregat per la variable *Net Won*:



Gràfic 42. Gràfic de dispersió segregat PFR ~ 3Bet

Els punts verds fan referència a aquells jugadors/es que tenen guanys i, els vermells, a aquells que tenen pèrdues.

Es pot veure que la majoria d'observacions es troben entre el 0% i el 20% del *PFR* i, entre el 0% i el 10% de *3Bet*. A mesura que augmenta el *3Bet* i el *PFR*, es van reduint notablement el número de jugadors/es. Podem comentar que diversos jugadors/es perdedors i guanyadors se situen en una mateixa zona (entre 0-10% del 3Bet i 0-30% del PFR). Però és evident com hi ha diversos jugadors/es perdedors que sobresurten i resulten ser extrems, amb valors massa elevats per ambdues variables.

Si es tenen en compte els límits citats prèviament, es sap que el *PFR* dels jugadors/es guanyadors/es hauria d'estar entre el 10% - 18% (ambdós inclosos) i que el *3Bet* dels guanyadors/es sol estar entre el 3%-10% (ambdós inclosos). La majoria de les nostres observacions es troben dins aquests marges.
No obstant, cal remarcar que poden haver observacions fora d'aquests límits que resultin guanyadors/es degut a que tindran altres variables amb millors percentatges que faran que es compensin amb altres factors del joc.

Si ens fixem en les rectes de regressió, veiem com la dels jugadors/es amb pèrdues (P) té una pendent superior a la d'aquells amb guanys (G). Aquest fet es degut a que hi han punts extrems amb pèrdues que fan que aquesta recta tingui més pendent. Però totes dues pendents són positives, fet que fa visible una certa relació lineal d'ambdues variables analitzades (*PFR i 3Bet*).
Finalment, s'estudia la correlació entre les dues variables analitzades realitzant un test de correlació *Pearson*:

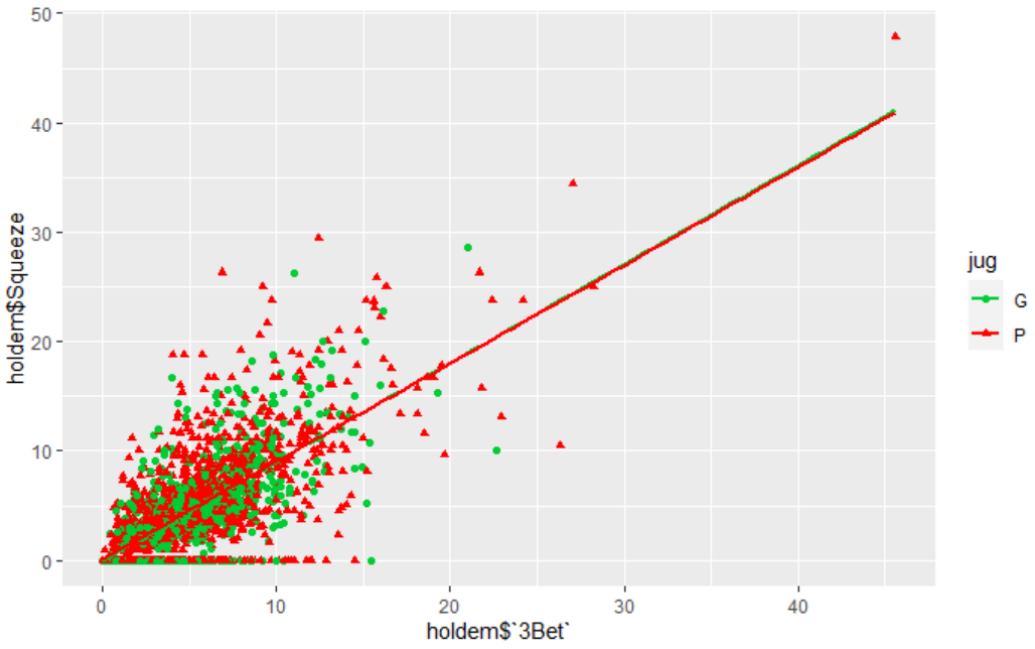
```
        Pearson's product-moment correlation

data:  holdem$PFR and holdem$`3Bet`
t = 46.521, df = 2120, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6890288 0.7311807
sample estimates:
      cor
0.7107421
```

Taula 32. Correlació de Pearson PFR ~ 3Bet

Com el *p-valor* del test és més petit que el 0.05, rebutgem la hipòtesis nul·la que diu: "la correlació entre aquestes dues variables serà igual a 0". A més, tenim un coeficient de correlació de 0.7107421. Per tant diem que un 71,07 % del *PFR* és explicat per la variable *3Bet* (correlació força alta i positiva).

❷ 3Bet segons l'Squeeze:

En aquest cas, s'estudiarà el *3Bet* dels jugadors segons el seu *Squeeze*:



Gràfic 43. Gràfic de dispersió segregat 3Bet ~ Squeeze

S'intueix una certa relació i tendència lineal entre aquestes dues variables. La majoria de *3Bet* i d' *Squeeze* es concentra entre el 0% i 10%; i a mesura que augmenta l'*Squeeze* també ho fa el *3Bet*. S'observa com aquells jugadors/es amb pèrdues tenen més dispersió, ja que hi han diversos punts extrems.

Un fet curiós és la quantitat d'observacions situades en el 0% d'*Squeeze*. Aquest fet ja l'hem intuït en l'apartat d'anàlisis descriptiu univariant on vèiem que la freqüència del 0% de la variable *Squeeze* era molt elevada tant per jugadors/es amb pèrdues com per jugadors/es amb guanys. Tot i que els jugadors/es amb pèrdues (situats en el 0% de l'*Squeeze*) eren gairebé el doble dels jugadors/es amb guanys, amb aquest tant per cent.
Concloem, per tant, que els jugadors contemplats en la nostra base de dades, i per tant en el nostre nivell de joc, no saben fer aquesta aposta concreta i no la contemplen ni executen com a conseqüència.
Pot ser un gran focus d'atenció per a la diferenciació de possibles estratègies.

Les dues rectes de regressió es solapen des d'un inici fent evident una certa similitud en la seva dispersió i, mostrant relació lineal.
Finalment s'estudia la correlació entre les dues variables amb un test de correlació *Pearson*:

```
        Pearson's product-moment correlation

data:  holdem$`3Bet` and holdem$Squeeze
t = 45.304, df = 2120, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6790778 0.7223568
sample estimates:
      cor
0.7013631
```

Taula 33. Correlació de Pearson 3Bet ~ Squeeze

Com el *p-valor* del test és més petit que el 0.05, rebutgem la hipòtesis nul·la que diu: "la correlació entre aquestes dues variables serà igual a 0".
A més, tenim un coeficient de correlació igual a 0.7013631 pel que podem dir que un 70,14% del *3Bet* és explicat per l'*Squeeze*. Concloem que aquestes dues variables tenen correlació positiva i alta (més propera al 1 que al 0).

❸ Postflop Agg% segons el W$WSF%:

Estudiarem el *Postflop Agg%* dels jugadors/es segons el seu *W$WSF%*:

Gràfic 44. Gràfic de dispersió segregat Postflop Agg% ~ W$WSF%

Podem observar una certa relació i tendència lineal entre aquestes dues variables. La majoria d'observacions es situen vora el 2% - 4% de *Postflop Agg%* i vora el 35% - 50% de *W$WSF%.* A mesura que augmenta una variable també ho fa l'altre.
Les dues rectes de regressió tenen pendent positiu i no arriben mai a solapar-se. La recta verda (jugadors amb guanys) té més pendent que la vermella(jugadors amb pèrdues) i; a mesura que augmenten ambdues variables, la recta dels guanys es va distanciant notablement de la recta de les pèrdues.

Finalment s'estudia la correlació entre les dues variables amb un test de correlació *Pearson*:

```
        Pearson's product-moment correlation

data:  holdem$`Postflop Agg%` and holdem$`W$WSF%`
t = 37.167, df = 2120, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.601642 0.653207
sample estimates:
      cor
0.6281136
```

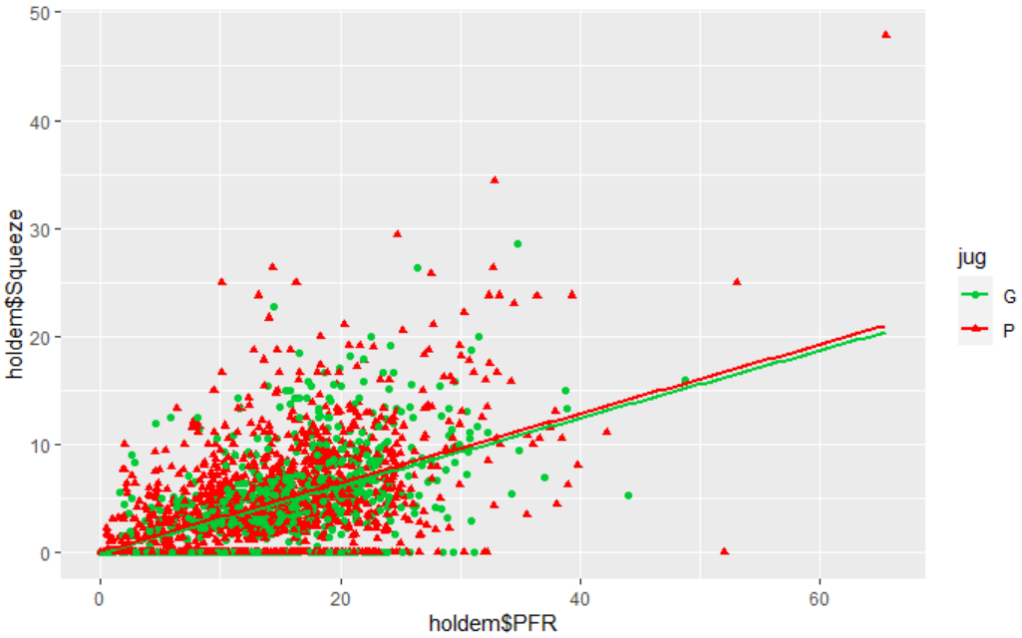
Taula 34. Correlació de Pearson Postflop Agg% ~ W$WSF%

Com el *p-valor* del test és més petit que el 0.05, rebutgem la hipòtesis nul·la que diu: "la correlació entre aquestes dues variables serà igual a 0".

A més, tenim un coeficient de correlació igual a 0.6281136 pel que podem dir que un 62,81% del *W$WSF%* és explicat per el *Postflop Agg%.* Concloem que aquestes dues variables tenen correlació positiva i alta (més propera al 1 que al 0).

4 PFR segons l' Squeeze:

Analitzem doncs el *PFR* dels jugadors/es segons el seu *Squeeze*:



Gràfic 45. Gràfic de dispersió segregat PFR ~ Squeeze

És visible una certa relació i tendència lineal entre aquestes dues variables. La majoria d'observacions es troben entre el 0% i el 20% de *PFR*, i entre el 0% i 10% de l'*Squeeze*. A mesura que augmenta una variable també ho fa l'altre.
Les dues rectes de regressió tenen pendent positiu i són paral·leles en tot el seu recorregut. La recta verda (jugadors amb guanys) té menys pendent que la vermella(jugadors amb pèrdues), però sense gaire diferència.

Es poden observar diverses observacions extremes, sobretot comentar una en concret (jugador amb pèrdues) que té valors molt alts per ambdues variables.

Finalment s'estudia la correlació entre les dues variables amb un test de correlació *Pearson*:

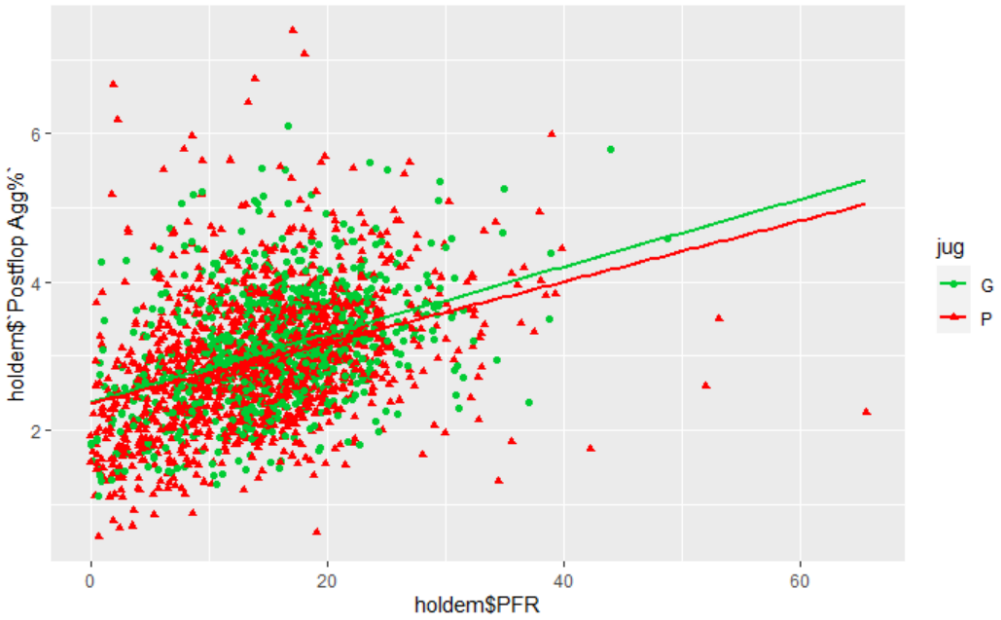
```
        Pearson's product-moment correlation

data:  holdem$PFR and holdem$Squeeze
t = 25.561, df = 2120, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.4521537 0.5172366
sample estimates:
      cor
0.4853673
```

Taula 35. Correlació de PFR ~ Squeeze

Com el *p-valor* del test és més petit que el 0.05, rebutgem la hipòtesis nul·la que diu: "la correlació entre aquestes dues variables serà igual a 0".
A més, tenim un coeficient de correlació igual a 0.4853673 pel que podem dir que un 48,54% de l'*Squeeze* és explicat per el *PFR.* Concloem que aquestes dues variables tenen correlació positiva i alta (més propera al 1 que al 0).

5 PFR segons Postflop Agg%:

Estudiem, a continuació, el *PFR* dels jugadors/es segons el seu *Postflop Agg%*:



Gràfic 46. Gràfic de dispersió segregat PFR ~ Postflop Agg%

En el *scatter plot* entre aquestes dues variables podem observar que la majoria d'observacions es situen entre el 0% i 25% del *PFR,* i vora el 2% i 4% del *Postflop Agg%.* Hi ha no obstant, molts jugadors/es extrems amb pèrdues tan per sobre del núvol de punts com per sota, explicant per tant la separació entre rectes de regressió a mesura que augmentem ambdues variables. Aquestes rectes, amb pendent positiu totes dues.

Estudiem, seguidament, la correlació observada a priori mitjançant el test de correlació *Pearson:*

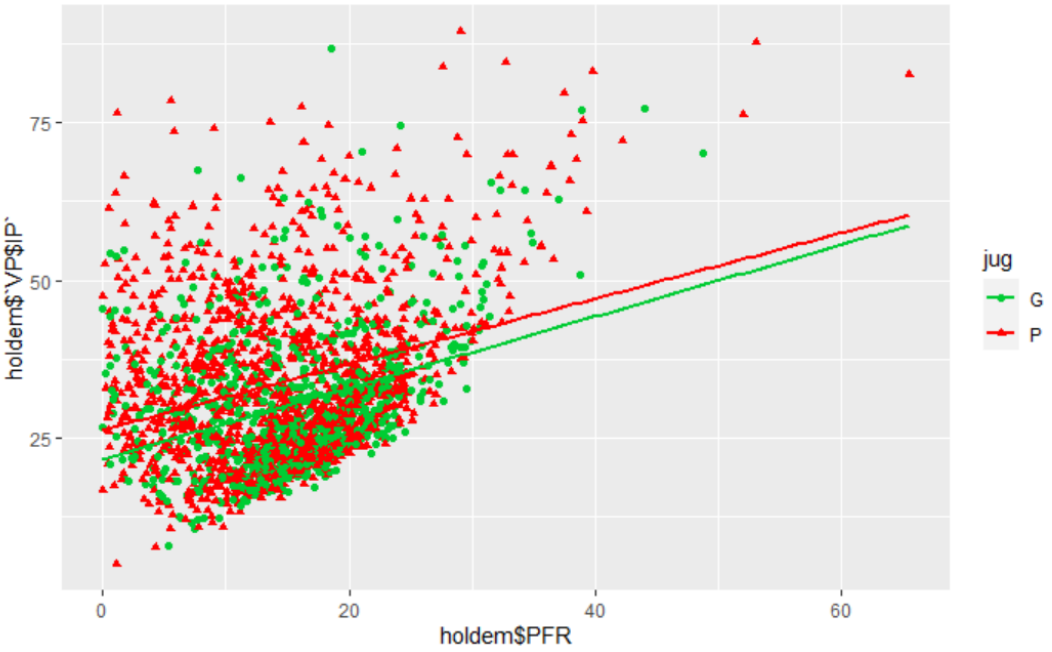
```
        Pearson's product-moment correlation

data:  holdem$PFR and holdem$`Postflop Agg%`
t = 18.333, df = 2120, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3326013 0.4060780
sample estimates:
      cor
0.369918
```

Taula 36. Correlació de PFR ~ Postflop Agg%

Com el *p-valor* del test és més petit que el 0.05, rebutgem la hipòtesis nul·la que diu: "la correlació entre aquestes dues variables serà igual a 0".
A més, tenim un coeficient de correlació igual a 0.369918 pel que podem dir que un 37% (aproximadament) del *Postflop Agg%* és explicat per el *PFR.* Concloem que aquestes dues variables tenen correlació positiva i alta (més propera al 1 que al 0).

6 PFR segons VP$IP:

S'estudiarà el *PFR* dels jugadors/es segons, aquesta vegada, el seu *VP$IP*:



Gràfic 47. Gràfic de dispersió segregat PFR ~ VP$IP

És visible com la majoria de *PFR* es troba vora el 10% i el 25% i que, a mesura que augmenta el *VP$IP* alhora que el *PFR*, es van reduint molt el número d'observacions. Destacar que aquells jugadors amb un *VP$IP* elevat, majoritàriament, són jugadors que perden a un ritme més alt.

Sembla que tinguin distribució semblant, però els perdedors/es tenen moltes més observacions amb un *PFR* i un *VP$IP* superior, i més dispersió.
Les dues rectes de regressió són bastant paral·leles, sent superior la recta dels jugadors/es perdedors/es ja que tenen valors superiors més extrems. Intuïm per el gràfic de dispersió que aquestes dues variables tindran una certa relació lineal.

Per finalitzar, s'estudia la correlació entre les dues variables analitzades a través del test de correlació *Pearson*:

```
        Pearson's product-moment correlation

data:  holdem$PFR and holdem$`VP$IP`
t = 14.704, df = 2120, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.2650912 0.3423322
sample estimates:
      cor
0.3042116
```

Taula 37. Correlació de Pearson PFR ~ VP$IP

Com el *p-valor* del test és inferior al 0.05, rebutgem la hipòtesis nul·la que diu: "la correlació entre aquestes dues variables serà igual a 0".
A més, tenim un coeficient de correlació del 0.3042116. Per tant, diem que un 30,42% del *PFR* és explicat pel *VP$IP* (correlació positiva, però més aviat baixa).

VII. CLUSTERING

La tècnica d'anàlisi clúster o anàlisi de conglomerats consisteix en classificar als individus d'estudi formant grups o conglomerats (clústers) d'elements, tals que els individus dins de cada conglomerat presentin un cert grau d'homogeneïtat en base als valors adoptats sobre un conjunt de variables. És a dir, que els individus dins de cada clúster comparteixin característiques en comú i alhora estiguin diferenciats dels altres clústers.

Aquesta tècnica desconeix aquests conglomerats i, per tant, consisteix en formar-los d'una manera òptima. Els possibles grups que es formin vindran determinats per les diferents variables de la base de dades d'estudi. Un cop fet aquest agrupament caldrà trobar respostes a aquestes agrupacions creades.

De cara a realitzar el *clustering* de la nostre base de dades, tenim en compte que la variable qualitativa *Player Name* és l' *id* de la base de dades i que, la variable *Site* que tenia 3 factors s'ha reduït a un únic factor. Per tant, tindrem només variables quantitatives / numèriques per la realització del *Clustering*.
Dintre d'aquestes variables numèriques es diferencien dos tipus:

- **Variables numèriques discretes**
  - *Hands*
  - *Net Won*
- **Variables numèriques en percentatges (contínues)**
  - *VP$IP*
  - *PFR*
  - *3Bet*
  - *Postflop Agg%*
  - *W$WSF%*
  - *WTSD%*
  - *Won $ at SD*
  - *Squeeze*

Per tant, al fer aquesta diferenciació de variables farem 3 clusterings:

1- Clustering PAM amb **distància de gower** (totes variables juntes escalant variable Hands i Net Won).

2- Clustering PAM amb **distància Bray-Curtis** (només variables percentuals).

3- Clustering PAM ajuntant clustering 1 i 2 combinant dues distàncies (**D=D1+D2**). D1 amb la distància de *Gower* entre *Hands i Net Won* i, D2 amb la distància de *Bray-Curtis* entre variables percentuals.

1- CLUSTERING VARIABLES NUMÈRIQUES

Realitzarem un *clustering* amb totes les variables numèriques de la nostre base de dades ja processada, escalant la variable *Hands* i *Net Won*.
Per a que un algorisme que encara està per triar pugui agrupar observacions, primer hem de definir alguna noció de similitud o dissimilitud entre les observacions. En el nostre cas volem definir nocions de dissimilitud entre observacions. Una opció popular per l'agrupació és la distància euclidiana. No obstant, la distància euclidiana només és vàlida per aquelles variables contínues i, per tant, no es pot aplicar en el nostre cas. Per a que un algoritme d'agrupament produeixi resultats sensibles, hem d'utilitzar una mètrica de distància que pugui manejar tipus de dades mixtes. Com aquest és el nostre cas, utilitzarem doncs la **distància de *Gower*** . Aquesta es calcula:

$$d_{ijk} = \frac{|x_{ik} - x_{jk}|}{R_k}$$

Com que aquests dos tipus de variables numèriques tenen alguns *outliers* i l'ús del mètode *k-means* és molt sensible als *outliers* , podria afectar greument en l'assignació de les observacions al clúster. Per tant, s'ha decidit utilitzar l'algorisme PAM que és més robust. L'algorisme PAM és un mètode que es basa en la cerca de *medoids* o objectes representatius entre les observacions de les variables triades de la base de dades d'estudi. L'objectiu serà trobar k *medoids* que minimitzin la suma de les diferències de les seves observacions (individus) al seu *medoid* més proper. Per estimar el nombre òptim de clústers hi ha moltes tècniques. Nosaltres farem servir les 3 següents:

- L'**amplada de la silueta** que és una mètrica de validació interna que consisteix en una mesura agregada de la similitud d'una observació amb el seu propi clúster en comparació amb el seu clúster veí més proper.

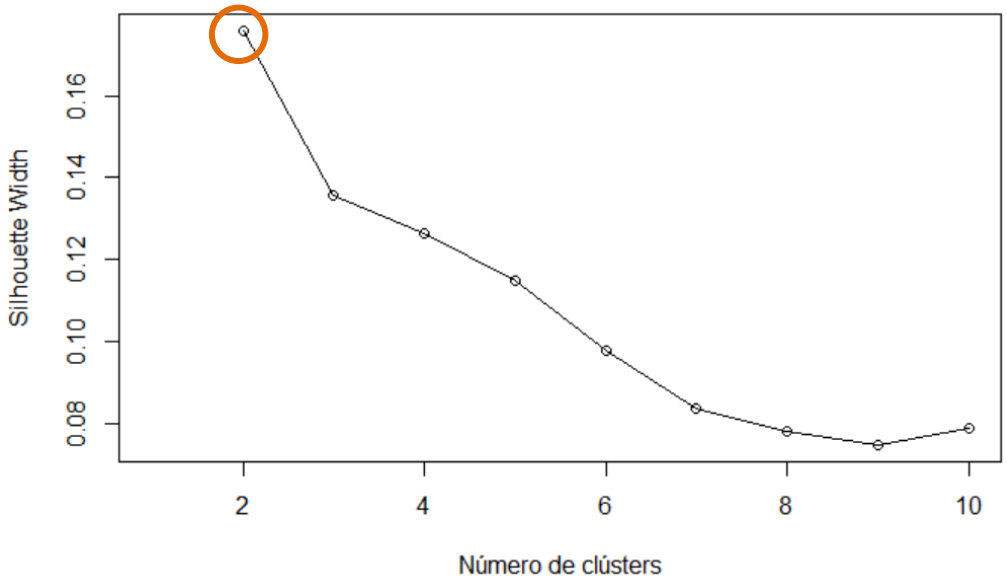
- El **criteri Calinski-Harabasz** que és un tipus de criteri intern (intra-grups) que s'utilitza per comparar diverses possibilitats de particions pròximes entre sí, mitjançant *k-means*. Ens permetrà validar l'elecció adoptada i proporcionarà informació sobre el grau de cohesió, relació o similitud entre els elements dins de cada grup. En definitiva, es basa en la relació entre la variància entre grups i la variància dins dels grups, de manera que un valor major del quocient indica una millor partició.

- **Dendrograma** amb mètode de Ward.

Aquests 3 mètodes de validació interna per tal de triar el nombre de clústers òptims els farem servir en els dos clustering següents, on es seguirà el mateix procediment que realitzarem a continuació.

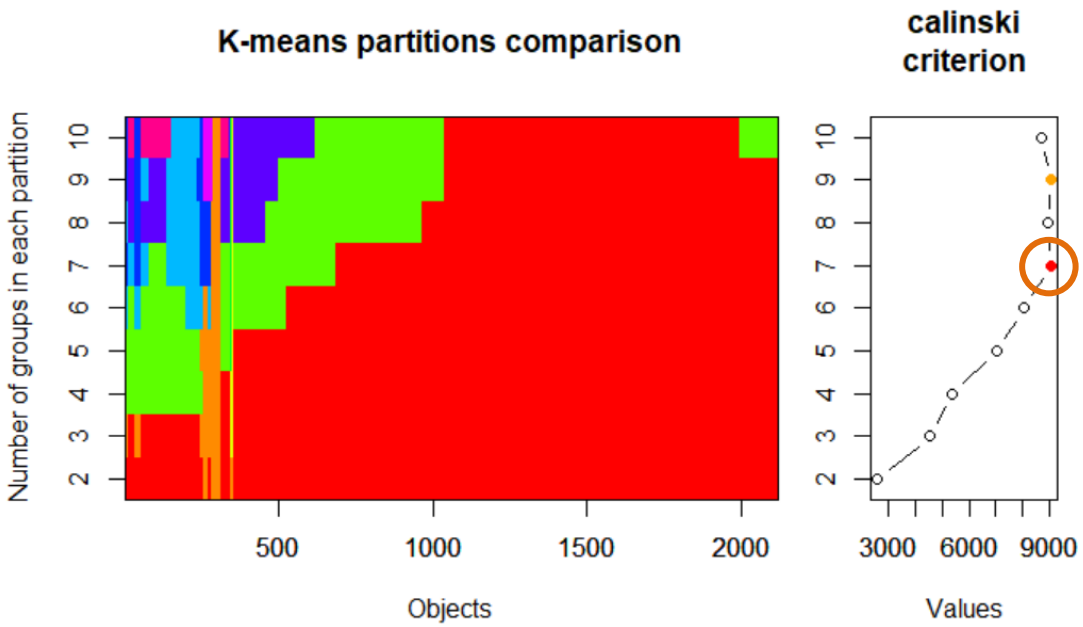
Quan hem calculat la distància de gower pel nostre *dataframe* amb aquelles variables numèriques, realitzarem diversos mètodes de validació interna per tal de triar el número de clústers òptims.

- Mitjançant el gràfic de **l'amplada de la silueta** obtenim k=2 com a número de clústers òptims:



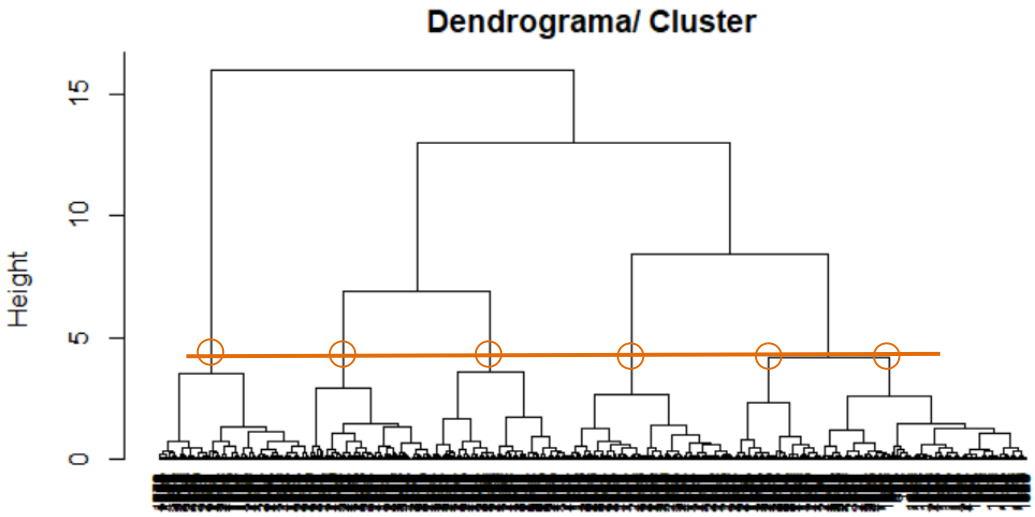
Gràfic 48. Amplada de la silueta clustering 1 variables numèriques

- Pel **criteri de Calinski**, en canvi, obtenim k=7 número de clústers òptims:



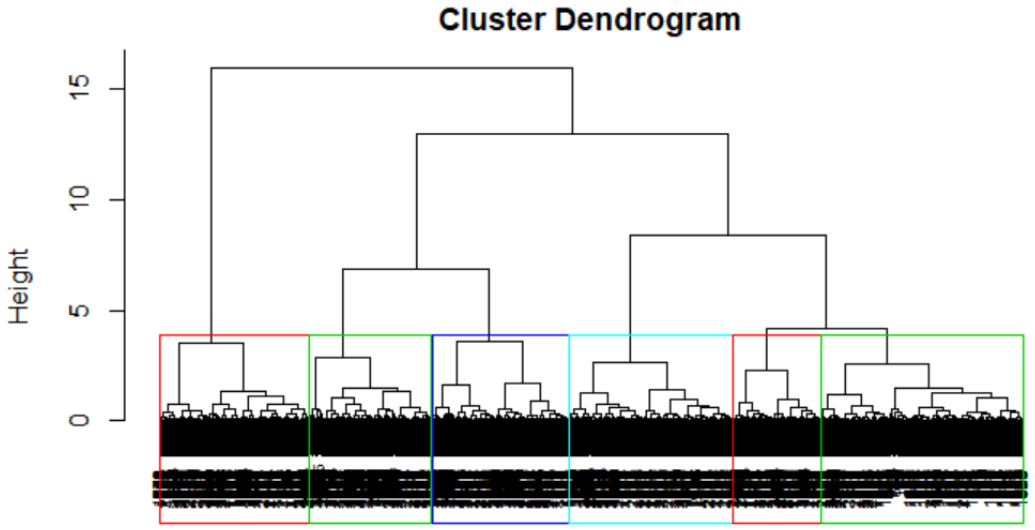
Gràfic 49. Criteri Calinski clustering 1 variables numèriques

- I, finalment, visualitzem el **dendrograma** que hem realitzat utilitzant el mètode de Ward on podem intuir uns k=6 clústers òptims:



Gràfic 50. Dendrograma clustering 1 variables numèriques

On aquests 6 clústers quedarien partits de la següent manera:



Gràfic 51. Dendrograma clustering 1 variables numèriques, amb 6 clústers marcats

Per tant, comparant aquest 3 mètodes de validació interna per triar els clústers òptims veiem que tots diuen una cosa diferent. Agafar k=2 clústers seria obtenir poca informació de les possibles estratègies presents i; agafar-ne k=7, pot ser seria agafar-ne masses. Per tant he cregut convenient quedar-me amb k=6 clústers òptims.

Un cop triat el número de clústers amb els quals treballarem, realitzem **l'algorisme PAM** on trobarem quins són aquests 6 medoids, les mides de cada clúster i els seus respectius valors per cada variable d'estudi.

- La mida dels 6 clústers calculats per l'algorisme PAM són:

Clústers	1	2	3	4	5	6
n	477	486	399	298	248	214

Taula 38. Mida dels 6 clústers mitjançant PAM (1)

Observem que les mides estan bastant equilibrades entre els 6 grups. El grup 2 és el grup que conté més jugadors; i, en canvi, el grup 6 el que menys.

- Els medoids d'aquests 6 clústers són:

Clústers	1	2	3	4	5	6
medoids	480	1036	715	785	1761	1872

Taula 39. Medoids dels 6 clústers mitjançant PAM (1)

- La màxima i mitjana dissimilaritat entre les observacions del clúster i el medoide del clúster, el diàmetre del clúster (màxima dissimilaritat entre dues observacions del clúster) i la separació del clúster (mínima dissimilaritat entre una observació del clúster i una observació d'un altre clúster) :

Clústers	max_diss	av_diss	diameter	separation
1	0.1433289	0.05292036	0.2587680	0.01605065
2	0.1259166	0.05344748	0.2290530	0.01717285
3	0.1454778	0.06027452	0.2701438	0.01700522
4	0.1223848	0.05764086	0.2104315	0.01871844
5	0.3113426	0.07863699	0.3865471	0.01605065
6	0.1265905	0.06407563	0.2306920	0.01935506

Taula 40. Valors dissimilaritat dels 6 clústers mitjançant PAM (1)

- La mitjana dels 5 clúster per cada una de les 8 variables analitzades són:

Clústers	Hands	Net Won	VP$IP	PFR	3Bet
1	2546.9859	2.012213	24.82629	14.241722	4.820378
2	1057.6865	-36.123973	30.97498	8.103737	2.618746
3	993.6119	22.313582	28.68456	18.170837	6.522247
4	880.3267	-62.763067	42.16877	23.290559	10.170128
5	640.6310	-44.756458	46.15374	14.242251	3.782885
6	657.9171	-29.684562	25.46058	11.325324	3.929583

Clústers	Postflop Agg%	W$WSF%	WTSD%	Won $ at SD	Squeeze
1	2.786997	43.59165	33.36237	51.96693	4.225713
2	1.965686	38.46495	37.42216	52.14414	2.500929
3	3.675213	49.53877	30.18867	51.20958	5.445231
4	3.366671	45.65002	35.14571	43.10036	10.960428
5	3.429574	44.18065	32.02304	43.27280	2.937299
6	2.925915	39.02292	28.42513	42.88406	3.128366

Taula 41. Mitjanes dels 6 clústers mitjançant PAM (1)

Centrant-nos en els valors de les mitjanes dins de cada clúster per a cada variable, podem extreure característiques rellevants dels diversos grups. Els valors de color verd són els valors més baixos dels 5 clústers per aquella variable; i els de color blau, els més elevats. No significa que els més alts siguin bons ni a la inversa, dependrà dels valors per variable. Per tant, podem comentar que en el:

- **Clúster 1:** ens trobem davant un grup de jugadors guanyadors ja que el seu *VP$IP* és bastant bo i no té un *gap* molt gran amb el seu *PFR*. Són un tipus de jugadors sòlids com indica el seu *3Bet* i sobretot el seu *Postflop Agg%*. Aquests jugadors compensen els seus valors i, les seves jugades per valor estan vora el 2.8%, valor una mica alt el qual ens indica que afegeixen algun farol de més i que haurien de pagar-ne menys. Aquests valors també ens indiquen que juguen les seves mans per valor fins el *river*, en el que guanyen un 52% (aproximadament) de les ocasions. Aquesta xifra podría millorar-se. És un grup guanyador, com indica el seu *Net Won*, que juga bastantes mans de mitjana en comparació a la resta de grups.

- **Clúster 2:** trobem jugadors amb certa experiència però que resulta impossible que esdevinguin guanyadors ja que entre el seu *VP$IP* i el seu *PFR* hi ha un *gap* molt gran. Aquest fet ens indica que són jugadors que fan molts *limps* (igualen la cega) i que paguen de més el *flop.* Un *VP$IP* del 31% (aproximadament) com tenen aquests jugadors de mitjana només seria sostenible amb un *PFR* del 25% però no és el cas ja que tenen un valor del 14%, aproximadament. Observem que en el *river* guanyen (grup que més guanya) però el seu problema és la quantitat de diners que es deixen abans d'arribar al *river* per ser massa porucs. Podriem definir-los com jugadors que exploten de menys les seves mans bones pel fet de ser passius i, quan porten una mà potent *preflop* tipus AA-KK-QQ, no saben abandonar-les.

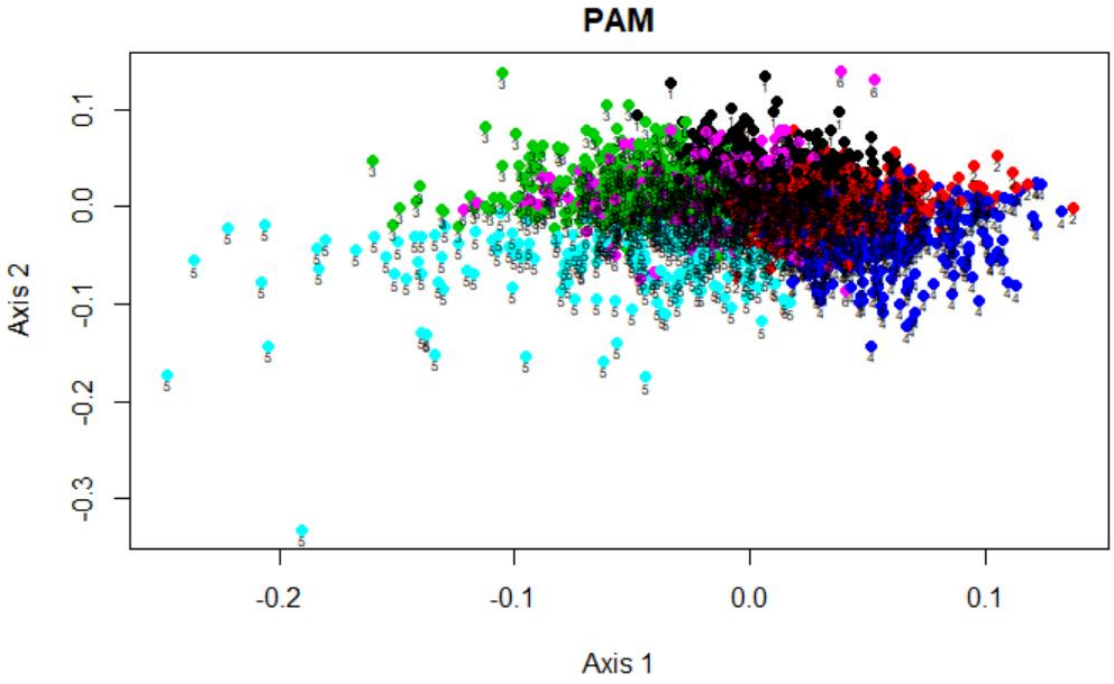
- **Clúster 3:** es tracta d'un grup de jugadors molt guanyador, com indica el seu *Net Won*. Tenen un estil agressiu de joc com demostra el seu *3Bet* i el seu *Squeeze*, indicant que juguen mans per valor i mans de farol. Solen apostar molt postflop i així consegueixen que els seus rivals abandonin la seva mà. Realitzen molts farols però no guanyen més d'un 51% (aproximadament) al *showdown*. Tot i així, aquest grup aconsegueix guanyar fet que ens indica que segueixen una estratègia tremendament guanyadora i efectiva.

- **Clúster 4:** Podem observar com aquest grup està contingut per aquells jugadors més perdedors. Podriem dir que són excessivament agressius *preflop* i que juguen un 42% de la baralla de cartes, la qual és una xifra insostenible. L'Squeeze del quasi 11% és una altra xifra que ens fa veure que són jugadors que els hi agrada ficar-se en els pots de moltes maneres, però sobretot sent agressius i amb iniciativa. Esdevenen fàcil detectables pels jugadors competents de la taula i que simplement jugant un 20%-25% de les mans contra ells i tenint paciència acabaran enxampant-lis un dels seus múltiples farols per quedar-se així amb tot el seu *stack*.

- **Clúster 5:** També es tracta d'un grup perdedor, però el seu estil de joc és totalment diferent al del clúster 4, ja que es tracta de jugadors molt passius que entren fent *limp* als pots deixant-se així milers de cegues abandonant quan el flop no els hi sembla apropiat. Aquest tipus de jugadors perden diners de forma més lenta e inclús podrien arribar a guanyar algún dia esporàdic, fent que tinguin una sensació enganyosa de que no són tan dolents com realment sí són. Pot semblar, a simple vista, que el seu *Postflop Agg%* sigui alt i ens faci pensar que són jugadors agressius però és pel fet que, a vegades, en pots de 20 cegues aposten 1 i fan que aquest tipus de jugades pugin el valor d'aquesta variable. En realitat, es tracta d'apostes que no fan cap mal als rivals. Es podria dir que són jugadors inexperts que juguen per diversió.

- **Clúster 6:** Aquest grup no té mals números de partida, ja que el seu *VP$IP* vora el 25% és bo i el seu *PFR* tot i ser una mica baix (11% aproximadament) és mig sostenible. Aquest hauria d'estar entorn al 15%. Aquest grup també té un 3Bet sòlid amb poc rang de farol. Pel que ens porta a pensar que el seu problema es troba en els guanys al *showdown*. Són jugadors pagadors que tenen el concepte del pòquer i els rangs establerts però que es neguen a abandonar les seves mans, fet que els deriva a ser perdedors. Podriem dir que són jugadors que han dedicat temps a estudiar i quan perden no ho entenen i juguen sota *tilt*. Si aquests jugadors realitzessin lleugers canvis a nivell estratègic i, sobretot, mental podrien arribar a ser guanyadors en poc temps.

Per finalitzar aquest clustering, mirem de corroborar les explicacions fetes mitjançant la representació dels clústers trobats amb el **mètode MDS** (*Multidimensional Scaling*) e interpretarem els seus dos eixos.



Gràfic 52. Representació 6 clústers mètode MDS (1)

- Axis 1:

Hands	Net Won	VP$IP	PFR	3Bet	Postflop Agg%	W$WSF%	WTSD%	Won $ at SD	Squeeze
0.106	0.081	-0.49	-0.776	-0.756	-0.677	-0.539	0.025	0.429	-0.624

Sembla que el Axis 1 està altament i positivament correlacionat amb la variable *Won $ at SD* i, negativament correlacionat (valors molt alts) amb les variables *VP$IP, PFR, 3Bet, Postflop Agg%, W$WSF% i Squeeze*.

- Axis 2:

Hands	Net Won	VP$IP	PFR	3Bet	Postflop Agg%	W$WSF%	WTSD%	Won $ at SD	Squeeze
0.225	0.341	-0.628	-0.024	-0.101	0.493	0.524	-0.59	0.335	-0.131

L'Axis 2 del plot, segueix la mateixa tendència que l'Axis 1 a diferència que el Postflop Agg% i el W$WSF% passen a estar positivament correlacionats amb l'axis. La variable Hands i Net Won prenen valors més alts i positius de correlació que en l'Axis 1.

Per tant podem identificar els 5 clústers com:
Clúster 1, Clúster 2, Clúster 3, Clúster 4, Clúster 5 i **Clúster 6.**

2- CLUSTERING VARIABLES NUMÈRIQUES EN FORMAT PERCENTATGE:

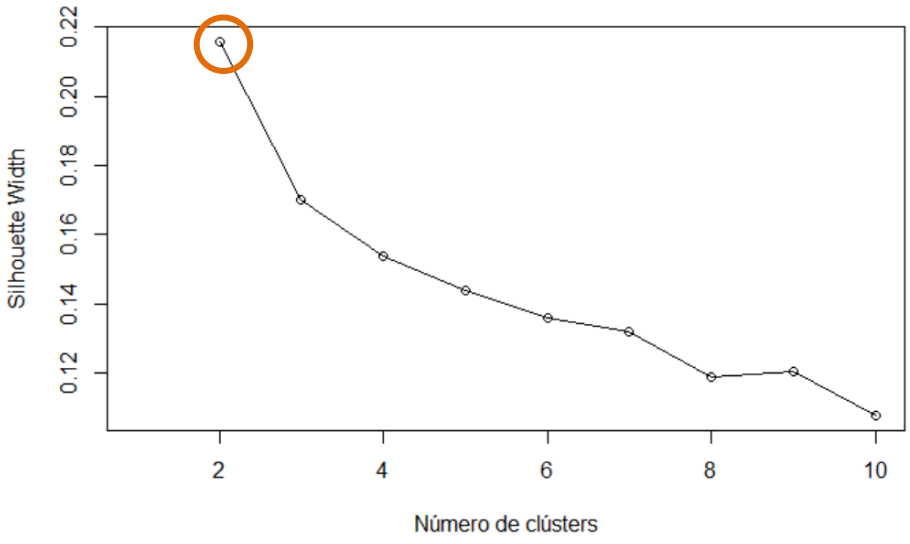
En aquest clustering es treballarà només amb aquelles variables de la base de dades que siguin numèriques i estiguin en format percentatge (%).
Aquestes variables són:
{ *VP$IP, PFR, 3Bet, Postflop Agg%, W$WSF%, WTSD%, Won $ at SD i Squeeze* }

A l'hora de tractar aquest tipus de variables caldrà treballar amb una distància de ***Bray-Curtis***. Aquesta es basa en l'abundància e intensitat de les dades utilitzant la dissimilitud entre dos llocs en funció dels recomptes dels elements en cada lloc. Per tant, si la classe té dissimilitud de 0 seran similars, en canvi si té dissimilitud igual a 1 les classes no tindran res en comú. La distància de *Bray-Curtis* es calcula de la següent forma:

$$d_{Bray}(\bar{p}, \bar{q}) = \frac{\sum_{i=1}^{n} |p_i - q_i|}{\sum_{i=1}^{n} p_i + \sum_{i=1}^{n} q_i}$$

Per tant, un cop calculada la distància pel nostre *dataframe* amb aquelles variables en format percentatge, realitzarem diversos mètodes de validació interna per tal de triar el número de clústers òptims.

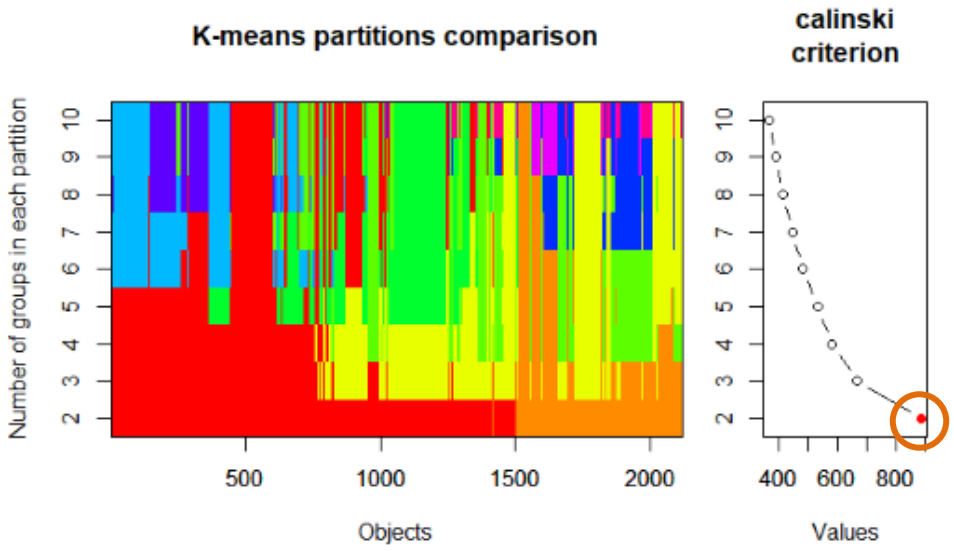
- Mitjançant el gràfic de **l'amplada de la silueta** obtenim k=2 com a número de clústers òptims:



Gràfic 53. Amplada de la silueta clustering 2 variables numèriques continues (en percentatge)

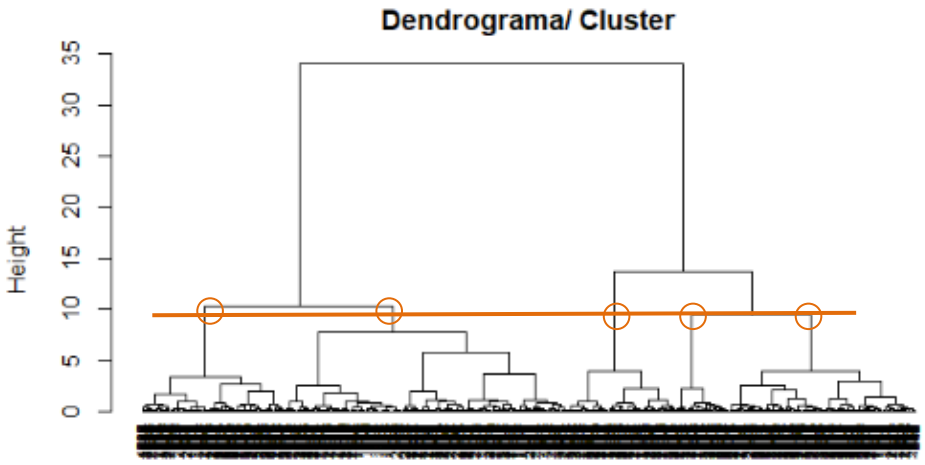
Com més proper sigui el punt al valor 1, aquell nombre serà el número de clústers més òptim trobat.

- Pel **criteri de Calinski** obtenim k=2 número de clústers òptims, també:



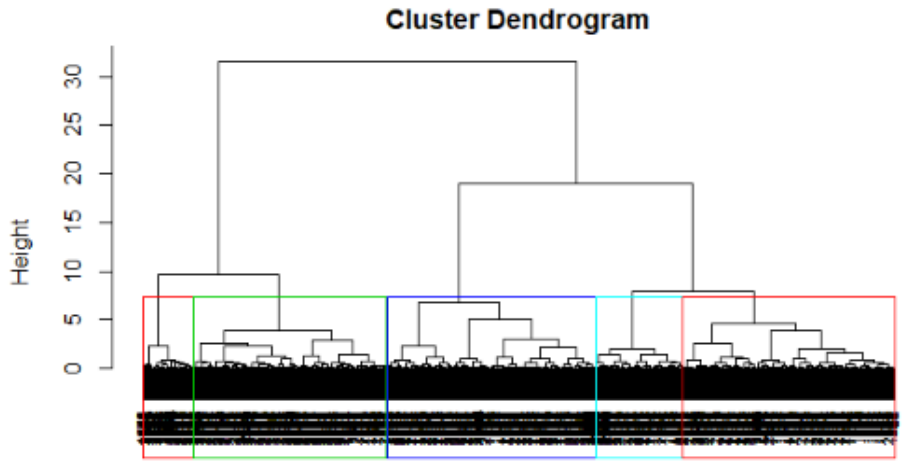
Gràfic 54. Criteri Calinski clustering 2 variables numèriques continues (en percentatge)

- I, finalment, visualitzem el **dendrograma** que hem realitzat utilitzant el mètode de Ward on, a simple vista, s'observen uns k=5 clústers òptims:



Gràfic 55. Dendrograma clustering 2 variables numèriques continues (en percentatge)

On aquests 5 clústers quedarien partits de la següent manera:



Gràfic 56. Dendrograma clustering 2 variables numèriques continues (en percentatge), amb 5 clústers marcats

Per tant, com el que ens interesa és trobar el màxim d'estratègies possibles presents en el nivell on juguem NL5 i NL10, triarem **k=5** com a número de clústers. A més a més si en triessim només dues d'estratègies (k=2), aquestes vindrien caracteritzades simplement per la variable *Net Won* i tot i que és important, no volem quedar-nos només amb aquesta diferència entre estratègies.

Un cop hem triat el número de clústers amb el qual treballarem, realitzem **l'algorisme PAM** on trobarem quins són aquests 5 medoids, les mides de cada clúster i els seus respectius valors per variable d'estudi.

- La mida dels 5 clústers calculats per l'algorisme PAM són:

Clústers	1	2	3	4	5
n	681	521	360	351	210

Taula 42. Mida dels 5 clústers mitjançant PAM (2)

El clúster número 1 és el que té més observacions/jugadors i el clúster 5 és el que menys en té. No obstant, estan bastant equilibrats en quan a mida es tracta, els 5 clústers.

- Els medoids d'aquests 5 clústers són:

Clústers	1	2	3	4	5
medoids	772	223	546	943	611

Taula 43. Medoids dels 5 clústers mitjançant PAM (2)

- La màxima i mitjana dissimilaritat entre les observacions del clúster i el medoide del clúster, el diàmetre del clúster (màxima dissimilaritat entre dues observacions del clúster) i la separació del clúster (mínima dissimilaritat entre una observació del clúster i una observació d'un altre clúster) :

Clústers	max_diss	av_diss	diameter	separation
1	0.2569090	0.07351957	0.3662090	0.01503290
2	0.1927610	0.07050823	0.3167920	0.01656398
3	0.2740998	0.08132250	0.3933982	0.01804416
4	0.1939107	0.08284273	0.3071822	0.01503290
5	0.2824992	0.08805639	0.3804711	0.02260725

Taula 44. Valors dissimilaritat dels 5 clústers mitjançant PAM (2)

- La mitjana dels 5 clúster per cada una de les 8 variables analitzades són:

Clústers	VP$IP	PFR	3Bet	Postflop Agg%
1	23.07475	12.222599	4.076983	2.737288
2	26.97992	13.950271	4.772842	3.026528
3	37.02383	20.661761	7.899343	3.526652
4	42.56362	9.046308	2.633211	2.675708
5	61.83948	23.439792	8.983931	3.286975

Clústers	W$WSF%	WTSD%	Won $ at SD	Squeeze
1	43.15689	32.66297	54.40700	3.643432
2	41.58787	31.83677	42.40533	3.826013
3	46.98072	32.74753	44.61287	7.669089
4	42.67343	34.51263	49.47807	2.191384
5	43.39509	35.81239	41.90725	8.726416

Taula 45. Mitjanes dels 5 clústers mitjançant PAM (2)

Centrant-nos en els valors de les mitjanes dins de cada clúster per a cada variable, podem extreure característiques rellevants dels diversos grups. Els valors de color verd són els valors més baixos dels 5 clústers per aquella variable; i els de color blau, els més elevats. No significa que els més alts siguin bons ni a la inversa.
En aquest agrupament no es tenen en compte el nombre de mans jugades ni els seus guanys o pèrdues netes, simplement tenim en compte els moviments del joc de *cash*. Per tant, podem comentar que en el:

- **Clúster 1:** Trobem valors que ens fan pensar que estem parlant de jugadors amb un joc bastant estructurat ja que tenen un percentatge de *VP$IP* bo, vóra el 23%, i també podriem dir que són jugadors sòlids *preflop* ja que pugen una aposta anterir amb les seves millors mans un 4%. Aquests valors defineixen un joc sòlid que a la llarga podria ser guanyadora, molt probablement.

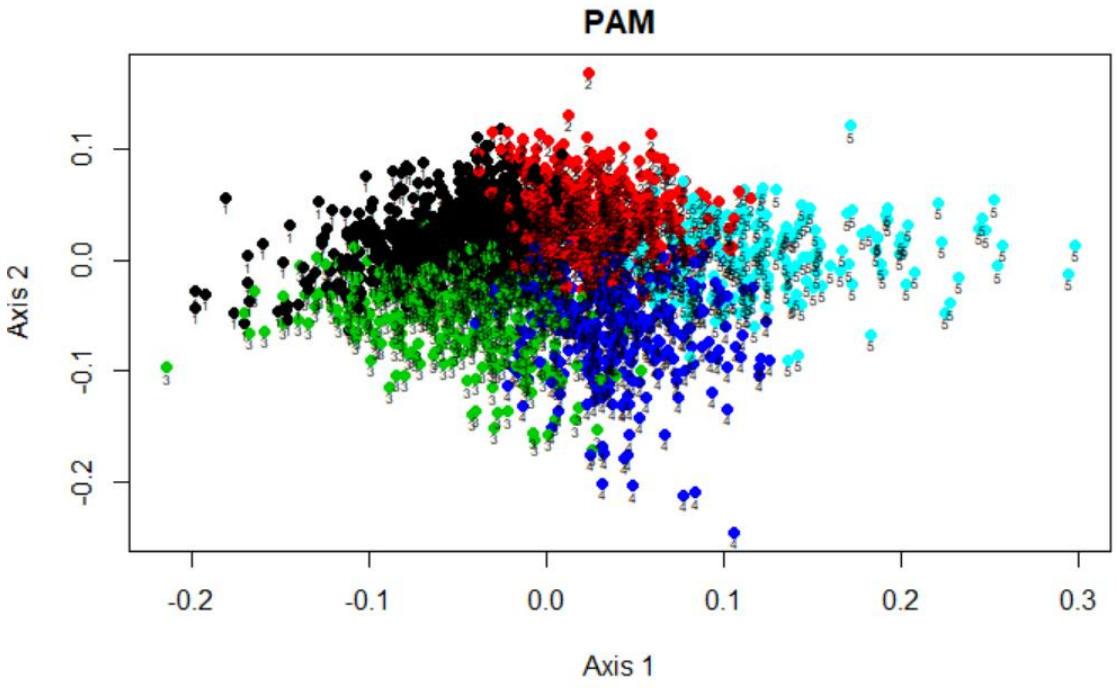
- **Clúster 2:** Aquest grup és potencialment guanyador ja que tenen un *VP$IP* mínimament alt, és a dir, juguen alguna mà que no haurien, però ho compensen amb un valor molt bo de *Postflop Agg%.*
  Ens trobem doncs davant jugadors que són molt agressius *postflop*, però sense pasar-se. Jugadors que saben fer farols i moltes vegades s'acaben enduent el pot abans d'arribar al *showdown*. Les vegades que arriben al *showdown* tenen problemes ja que guanyen un 42% de les vegades, és a dir, arrosseguen les seves mans de farol. El seu valor d'*Squeeze* és òptim, molt semblant al valor del clúster 1. Quan executen aquest moviment solen portar mans fortes.

- **Clúster 3:** Ens trobem davant un grup de jugadors/es potencialment perdedors a la llarga, però que poden arribar a guanyar (dies esporàdics) ja que tenen un *VP$IP* alt (37%) sense ser exagerat, un *PFR* alt i un *Squeeze* altíssim. Són, per tant, jugadors molt agressius que quan connecten mans acaben treient-li molt valor ja que els seus oponents saben que són molt agressius i fan molts farols, aleshores acostumen a pagar-li de més pensant-se que els pretén enganyar. El valor del seu *Postflop Agg%* és massa elevat fent evident que són jugadors que pretenen que els rivals s'acabin retirant mitjançant apostes fortes de gran tamany.
  Pretenen espantar al rival per tal d'obtenir el pot comú. Segurament aquests jugadors guanyin dies puntuals que connectin mans, però a la llarga seran perdedors/es.

- **Clúster 4:** Es poden observar valors característics d'aquells jugadors anomenats recreacionals. Tenen un *VP$IP* molt alt i un *PFR* molt baix, és a dir que tenen un *gap* enorme. Són jugadors que *limpean* molt (igualen la cega gran) per, finalment, acabar abandonant la mà. Quasi mai pugen la cega anterior i no porten mai la iniciativa de la mà. Aquests jugadors van perdent de forma lenta degut a que no porten la iniciativa i resulta molt poc probable que s'acabin enduent els pots. Tenen valors molt baixos en lo que iniciativa i agressió ens referim: *PFR, 3Bet, Postflop Agg%* i *Squeeze*. Es podria dir que juguen amb tanta por que, fins i tot, quan connecten mans amb el flop els espanta que els seus oponents es retirin. Fet que els porta a realitzar apostes molt insignificants o esperan a que sigui el rival qui aposti amb el seu farol per, després, pagar. El fet que el valor del *Won $ at SD* sigui dels més alts és normal degut a que quan porten cartes bones no es tiren i si acaben arribant al *showdown* és perquè han connectat molt. Resumint, es podria dir que aquest grup està format per jugadors que veuen molt el *flop* però, que si no connecten, acaben abandonant amb facilitat.

- **<u>Clúster 5:</u>** Aquest grup és el més perdedor amb diferència. Tenen un *VP$IP* del 61% que ens indica que juguen tot tipus de mans dèbils i a més a més, tenen un *PFR* del 23% que és insostenible. Juguen masses mans a diferència del clúster 1 que només jugaven les mans bones. Aleshores, sempre estaràn vençuts quan portin mans de rang dèbil. El seu *Squeeze* del 8% també és insostenible.
  Aquest tipus de jugadors/es se'ls coneix com a maníacs, que es refereix al fet que són excessivament agressius, juguen masses mans i tenen tendència al *tilt* (jugar cabrejat). Solen perdre grans quantitats de diners en poc temps. Són els jugadors més rentables pel joc i els més fugaços. És a dir que, quan ells estan presents a una taula, són fàcilment detectables pels jugadors bons causant que aquests abandonin el seu model de joc *GTO* per pasar al model explotatiu e intentar treure'ls-hi els màxims diners possibles abans que ho faci algú altre.

Per acabar i corroborar les explicacions fetes, representarem els clústers trobats amb el **mètode MDS** (*Multidimensional Scaling*) e interpretarem els seus dos eixos.



Gràfic 57. Representació 5 clústers mètode MDS (2)

Per interpretar correctament els Axis del plot, hem extret les correlacions entre variables per cada un dels Axis:

- Axis 1:

```
VP$IP   PFR  3Bet Postflop Agg% W$WSF% WTSD% Won $ at SD Squeeze
0.818 0.655 0.613         0.367  0.246 0.199      -0.526    0.51
```

Sembla que el Axis 1 està altament i positivament correlacionat amb les variables *VP$IP*, *PFR* i *3Bet*; i, negativament correlacionat amb la variable *Won $ at SD*, únicament.

- Axis 2:

```
VP$IP    PFR   3Bet Postflop Agg% W$WSF% WTSD% Won $ at SD Squeeze
).444 -0.565 -0.527        -0.422 -0.487 0.287      -0.168  -0.442
```

L'Axis 2 del plot anterior, per contra, està altament correlacionat de forma negativa amb les variables *PFR, 3Bet* i *W$WSF%.* També està força correlacionat negativament amb les variables *Postflop Agg%* i l' *Squeeze*.

Per tant podem identificar els 5 clústers com:

Clúster 1, Clúster 2, Clúster 3, Clúster 4 i **Clúster 5.**

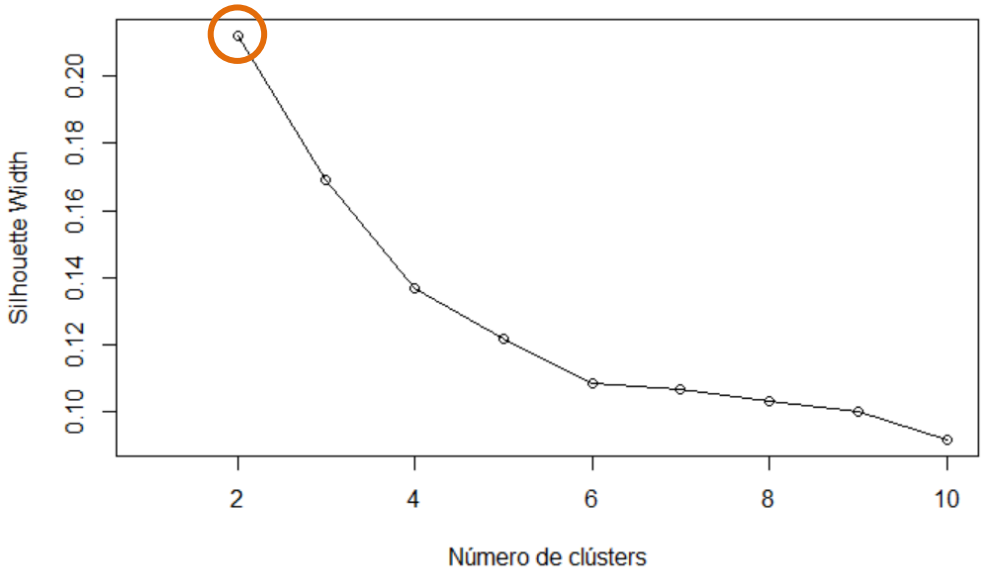
3- CLUSTERING VARIABLES NUMÈRIQUES (SUMA DISTÀNCIES CLUSTERING 1 I CLUSTERING 2) :

En aquest últim clustering pretenem sumar les dues distàncies realitzades als darrers 2 apartats de clustering fent servir: **D = D1+D2**, on D1 és la distància de *Gower* del paquet *StatMatch* només amb les variables discretes, i D2 és la distància de *Bray-Curtis* només amb les variables continues (en format percentatge).

Com la D1 és la matriu de distàncies de les discretes, aquestes s'haurien d'escalar per tal de poder obtenir la D. Una manera de fer-ho seria usant la funció *Gower* de *StatMatch* però només usant aquestes dues variables (*Hands* i *NetWon*), no totes les altres. La matriu de distàncies resultant seria la D1 que hauríem de sumar a la D2 que ja la tenim del cas 2, no cal aplicar-ne canvis. No obstant, si fem D = D1+D2 estem donant el mateix pes a les dues matrius. Per tant, una possibilitat és ponderar cada matriu pel nombre de variables que participen en la matriu dividit respecte el total de variables.

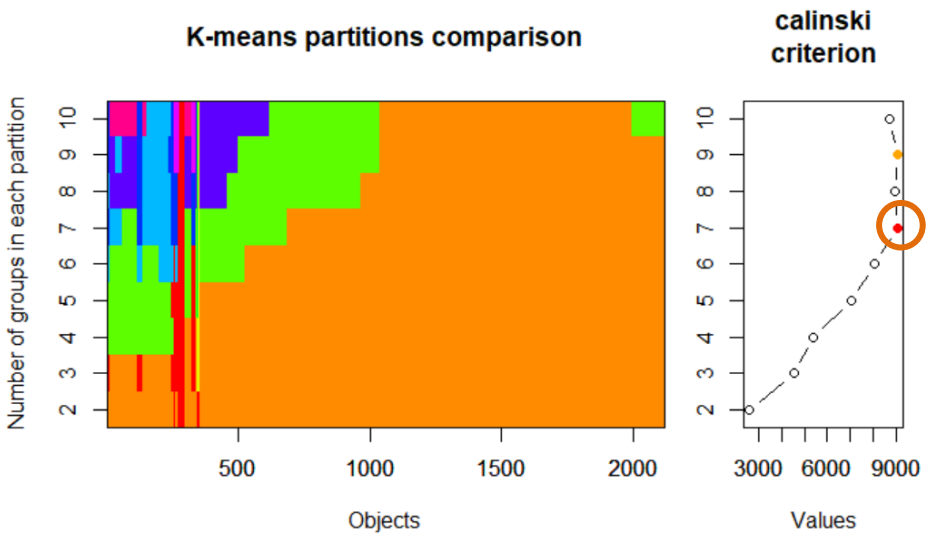
Per tant, un cop sumades i ponderades les dues matrius de distància (D1 i D2), realitzarem diversos mètodes de validació interna per tal de triar el número de clústers òptims.

- Mitjançant el gràfic de **l'amplada de la silueta** obtenim k=2 com a número de clústers òptims:



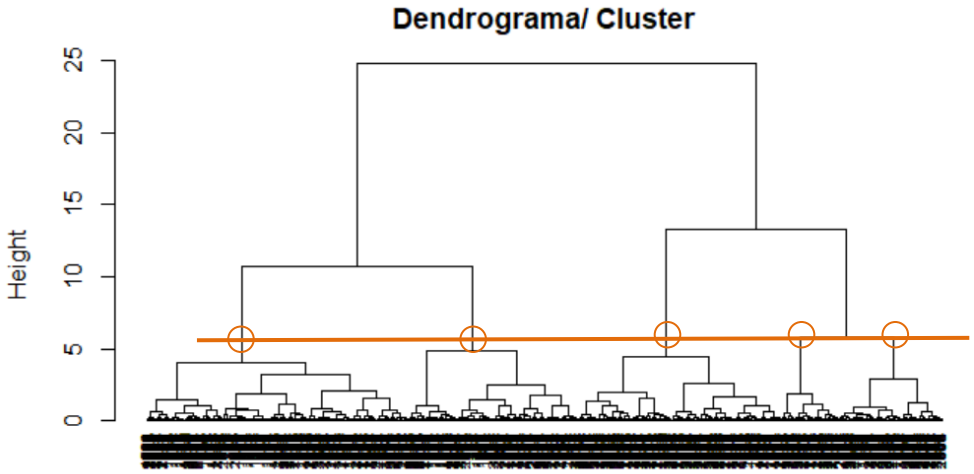
Gràfic 58. Amplada de la silueta clustering 3 suma distàncies dels clusterings 1 i 2

- Pel **criteri de Calinski** obtenim k=7 número de clústers òptims:



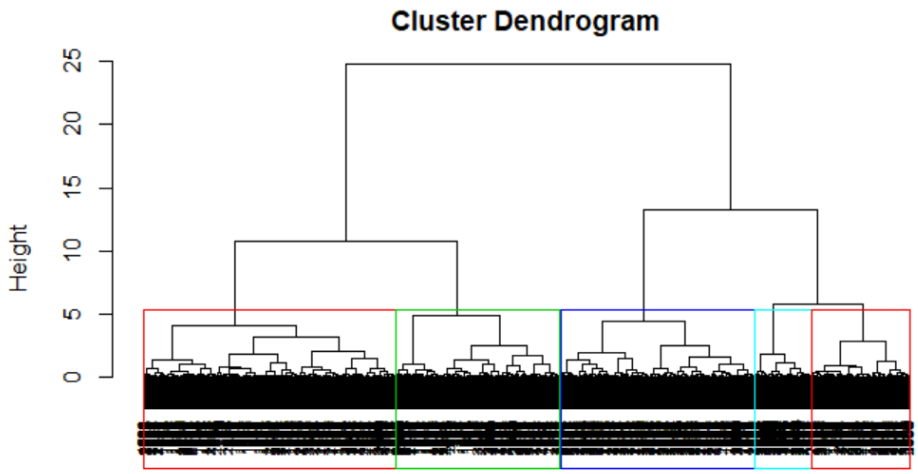
Gràfic 59. Criteri Calinski clustering 3 suma de distàncies dels clusterings 1 i 2

- I, finalment, visualitzem el **dendrograma** que hem realitzat utilitzant el mètode de Ward on, a simple vista, s'observen uns k=5 clústers òptims:



Gràfic 60. Dendrograma clustering 3 suma de distàncies dels clusterings 1 i 2

On aquests 5 clústers quedarien partits de la següent manera:



Gràfic 61. Dendrograma clustering 3 suma de distàncies del clusterings 1 i 2, amb 5 clústers marcats

Per tant, com el que ens interesa és trobar el màxim d'estratègies possibles presents en el nivell on juguem NL5 i NL10 i k=2 són massa poques estratègies i k=7 pot ser en són masses, triarem **k=5** com a número de clústers.

Un cop hem triat el número de clústers realitzem **l'algorisme PAM** on trobarem quins són aquests 5 medoids, les mides de cada clúster i els seus respectius valors per variable d'estudi.

- La mida dels 5 clústers calculats per l'algorisme PAM són:

Clústers	1	2	3	4	5
n	571	386	370	429	366

Taula 46. Mida dels 5 clústers mitjançant PAM (3)

El clúster número 1 és el que té més observacions/jugadors i el clúster 5 és el que menys en té. No obstant, estan bastant equilibrats en quan a mida es tracta, els 5 clústers.

- Els medoids d'aquests 5 clústers són:

Clústers	1	2	3	4	5
medoids	1844	471	558	669	974

Taula 47. Medoids dels 5 clústers mitjançant PAM (3)

- La màxima i mitjana dissimilaritat entre les observacions del clúster i el medoide del clúster, el diàmetre del clúster (màxima dissimilaritat entre dues observacions del clúster) i la separació del clúster (mínima dissimilaritat entre una observació del clúster i una observació d'un altre clúster) :

Clústers	max_diss	av_diss	diameter	separation
1	0.1555488	0.06136085	0.2905647	0.01450877
2	0.2034399	0.06479840	0.3068212	0.01450877
3	0.1386463	0.06806077	0.2383315	0.01527336
4	0.2572742	0.07415715	0.3340406	0.01736379
5	0.1575606	0.07052348	0.2628115	0.01603316

Taula 48. Valors dissimilaritat dels 5 clústers mitjançant PAM (3)

- La mitjana dels 5 clúster per cada una de les 8 variables analitzades són:

Clústers	Hands	Net Won	VP$IP	PFR	3Bet
1	1885.7817	12.12441	26.17774	17.469054	6.318491
2	1339.6586	-16.64694	21.69203	9.909508	3.008828
3	752.8942	-32.60909	39.21730	9.987032	2.968521
4	899.7253	-56.16886	37.70029	21.648771	9.092044
5	626.9226	-75.71277	60.54677	23.054879	8.283762

Clústers	Postflop Agg%	W$WSF%	WTSD%	Won $ at SD	Squeeze
1	3.183776	45.69197	31.95150	51.17403	5.793467
2	2.558838	40.21133	32.45351	50.48233	2.275077
3	2.780437	42.48769	33.97492	47.55001	2.399667
4	3.537449	46.98214	33.53769	42.66165	9.481579
5	3.414322	43.91185	35.06247	41.38515	7.693068

Taula 49. Mitjanes dels 5 clústers mitjançant PAM (3)

Centrant-nos en els valors de les mitjanes dins de cada clúster per a cada variable, podem extreure característiques rellevants dels diversos grups.
En aquest agrupament es tenen en compte totes les variables numèriques (discretes i continues). Per tant, podem comentar que en el:

- **Clúster 1:** És visible com aquest grup de jugadors és l'unic amb *Net Won* positiu, per tant estem davant un grup guanyador amb poc *gap* entre els seus valors d'*VP$IP* i *PFR*, que són amdós bons. Tot i ser el grup que guanya, en percentatge, més vegades el pot comú una vegada arribat al *showdown* aquest valor (51%) hauria de ser més alt. El seu *Postflop Agg%* del 3.18% demostra que són jugadors agressius que fan bastants farols, però ho compensen perquè aconsegueixen guanyar al *river*. Al fer tants farols i ser agressius obliguen al rival a abandonar la seva mà. Tenen el seu joc compensat.

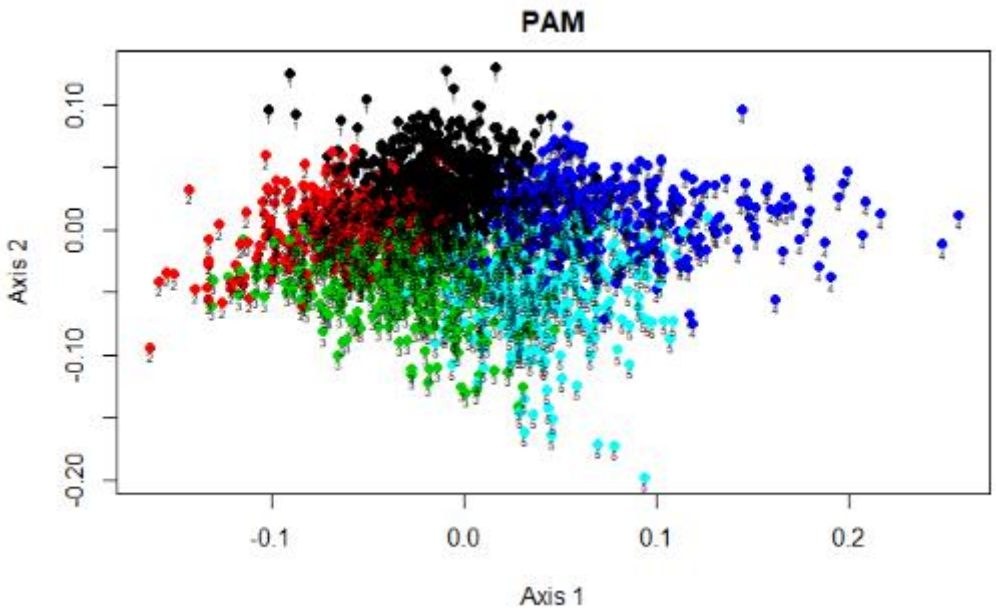
- **<u>Clúster 2:</u>** Veiem un *VP$IP* bo del 21.6% el que ens indica que són jugadors que seleccionen i tenen un rang de mans establerts en el seu joc, però tenen un *PFR* i un *3Bet* massa baix per arribar a ser guanyadors.
  Són jugadors massa sòlids i previsibles en el seu joc d'aquí que resultin ser perdedors, perquè al no tenir iniciativa, no roben pots. Si modifiquessin el seu *PFR* a l'alça vora un 15-16% i el seu 3Bet vora al 5-6%, sense dubte, esdevindrien guanyadors.

- **<u>Clúster 3:</u>** Estem davant uns jugadors recreacionals clàssics del pòquer, amb un 39.21% de *VP$IP* i un 10% de *PFR*, aproximadament, és impossible guanyar a aquest joc (*gap* enorme). Aquests jugadors igualen la cega amb moltes mans i paguen molt preflop per, finalment, acabar abandonant la mà. Per tant, són jugadors previsibles i una font d'ingressos per els bons jugadors de la taula.
  És un perfil recreacional que perd diners lentament ja que tampoc realitza bogeries, més aviat són jugadors passius i que entren als pots amb por i jugant de manera *random* (aleatòria).

- **<u>Clúster 4:</u>** Aquest tipus de jugadors perden a major ritme que els del clúster 3 ja que realitzen molts farols en el seu joc agressiu. Són el grup que més fan 3Bet (9.09%). Aquesta xifra seria sostenible però si es tingués un *VP$IP* menor). També tenen un *Postflop Agg%* exagerat (3.53%) i un *Squeeze* massa elevat (9.48%) que la fa insostenible.
  Són jugadors molt interessant pels bons jugadors ja que només cal esperar-los amb una bona mà i deixar que siguin ells qui prenguin iniciativa per finalment acabar guanyant-los.

- **<u>Clúster 5:</u>** En aquest grup es troben els jugadors maníacs que són jugadors que no tenen cap valor mitjanament ajustat. El seu *Won $ at SD* del 41.39% és una xifra baixíssima i, el *VP$IP* del 60.55% és el que definiriem com "regalar diners al adversari".
  Aquest jugadors perden el seu stack ràpidament i poden durar molt poca estona a la taula ja que juguen quasi tota la baralla de cartes i de forma gressiva.

Per acabar i corroborar les explicacions fetes de cada clúster, representarem els clústers trobats amb el **mètode MDS** (*Multidimensional Scaling*) e interpretarem els seus dos eixos:



Gràfic 62. Representació 5 clústers mètode MDS (3)

Per interpretar correctament els Axis del plot, hem extret les correlacions entre variables per cada un dels Axis:

- Axis 1:

Hands	Net Won	VP$IP	PFR	3Bet	Postflop	Agg%	W$WSF%	WTSD%	Won $ at SD	Squeeze
-0.219	-0.222	0.83	0.632	0.594		0.35	0.225	0.208	-0.538	0.496

Sembla que el Axis 1 està altament i positivament correlacionat amb totes les variables excepte amb les dues variables discretes (*Hands* i *Net Won*) i el *Won $ at SD* que ho estan negativament.

- Axis 2:

Hands	Net Won	VP$IP	PFR	3Bet	Postflop	Agg%	W$WSF%	WTSD%	Won $ at SD	Squeeze
0.226	0.218	-0.401	0.591	0.545		0.432	0.503	-0.276	0.165	0.454

L'Axis 2 del plot anterior, per contra, està altament correlacionat de forma negativa amb les variables *VP$IP* i *WTSD%.* Amb la resta de variables està positivament correlacionat.

Per tant podem identificar els 5 clústers com:
Clúster 1, Clúster 2, Clúster 3, Clúster 4 i **Clúster 5.**

VIII. ARBRES DE CLASSIFICACIÓ

Un arbre de classificació és una eina en forma gràfica i analítica (mapa o esquema) que s'utilitza per representar tots els possibles sucessos que poden sorgir a partir de la presa de decisió.
Ajuden a prendre decisions des d'un punt de vista probabilístic o bé, a traçar un algoritme que predigui matemàticament quina és la opció més acertada a triar en cada cas.
Aquest serà el nostre cas on utilitzarem aquests arbres de classificació per crear models predictius automàtics on es tenen en compte les observacions sobre una variable en concret per predir el valor d'aquella variable resposta.

L'arbre de classificació acostuma a començar per un node que es va bifurcant en possibles opcions/decisions. Cada bifurcació condueix a altres nodes addicionals que es ramifiquen en altres possibles opcions/decisions. Per això, visualment, pren l'estètica d'un arbre.

RANDOM FOREST:

Els *Random Forest* són boscos aleatoris formats per un conjunt d'arbres de classificació o regressió. Aquest algoritme, *Random Forest,* funciona agregant les prediccions fetes per diversos arbres de decisió de diferent profunditat que promedien les prediccions individuals de cada arbre. Tots els arbres de decisió del bosc s'entrenen en un subconjunt del conjunt de dades anomenat conjunt de dades d'arrancada (*bootstrapped dataset*).

El *Random Forest* ens proporcionarà idees (regles) de com intervenen les variables en l'assignació dels clústers trobats en apartats anteriors.

L'aleatorietat d'aquest algoritme s'introdueix en el model amb l'objectiu de reduir la variància mitjançant la reducció de la correlació entre els arbres.

Aquest algoritme és crea seguint els pasos esmentats a continuació:

1- Per cada un dels arbres es trien de forma aleatòria N dades de la mostra amb reemplaçament (*bootstrapping*).

2- En cada node de cada un dels arbres, es trien de forma aleatòria $m < P$ variables candidates per la partició (P són les variables explicatives del model). El número de variables m triat serà constant durant tot el procès de formació de l'arbre.

3- Deixarem crèixer cada arbre, sense podar, fins la màxima extensió possible.

Cal saber que els *Random Forest* tenen dos paràmetres fonamentals de disseny:

- **Ntree:** número d'arbres individuals que formen el *Forest* com a tal.
- **Mtry:** número de variables m triades en cada una de les particions modals.

El valor recomenat pel valor *mtry* si el que busquem són arbres de classificació, com és el nostre cas, serà igual a l'arrel quadrada de les *P* variables candidates per la partició. Important tenir en compte que quants més arbres individuals diferents es construeixin millor serà el caràcter d'anàlisis del *Forest* i millor seran les seves prediccions. No obstant, existeix un cert valor de ntree en el qual l'error de predicció s'estabilitza.

RANDOM FOREST SEGONS *NET WON*:

Per tant, partim del nostre *dataframe* amb totes les variables numèriques (*num*) i en modifiquem la variable *Net Won* passant-la a factor en una nova variable anomenada ***Class***. Si el *Net Won* > 0, *Class* serà *TRUE* i, si *Net Won* < = 0, *Class* serà *FALSE*. Aquesta nova variable l'hem anomenat *Class* i serà la variable resposta del nostre model predictiu.

A continuació, utilitzem el mètode *repeatedcv* per dividir el nostre conjunt de dades en 10 *folds* (plecs) de validació creuada i ho repetim 3 vegades per tal de tenir replicat el procediment i explicar, per tant, millor la variabilitat de les dades.
Mantindrem la validació establerta per fer proves a posteriori.

El paquet *caret* de l'R ens proporcionarà el nombre de variables aleatòries (*mtry*) a seleccionar. Li diem que, de les 9 variables predictores possibles, n'agafi de 1 a 6 a l'hora de triar els nodes dels arbres, i la mètrica *Accuracy* ens seleccionarà el model més òptim utilitzant el valor més alt. Observem, doncs, que el valor final utilitzat pel model serà de ***mtry* = 4**:

```
Random Forest

2122 samples
   9 predictor
   2 classes: 'FALSE', 'TRUE'

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 1910, 1910, 1910, 1909, 1910, 1910, ...
Resampling results across tuning parameters:

  mtry  Accuracy   Kappa
  1     0.7137951  0.3633498
  2     0.7148854  0.3739566
  3     0.7164548  0.3787007
  4     0.7183401  0.3850433
  5     0.7162909  0.3810965
  6     0.7183350  0.3866163

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was mtry = 4.
```

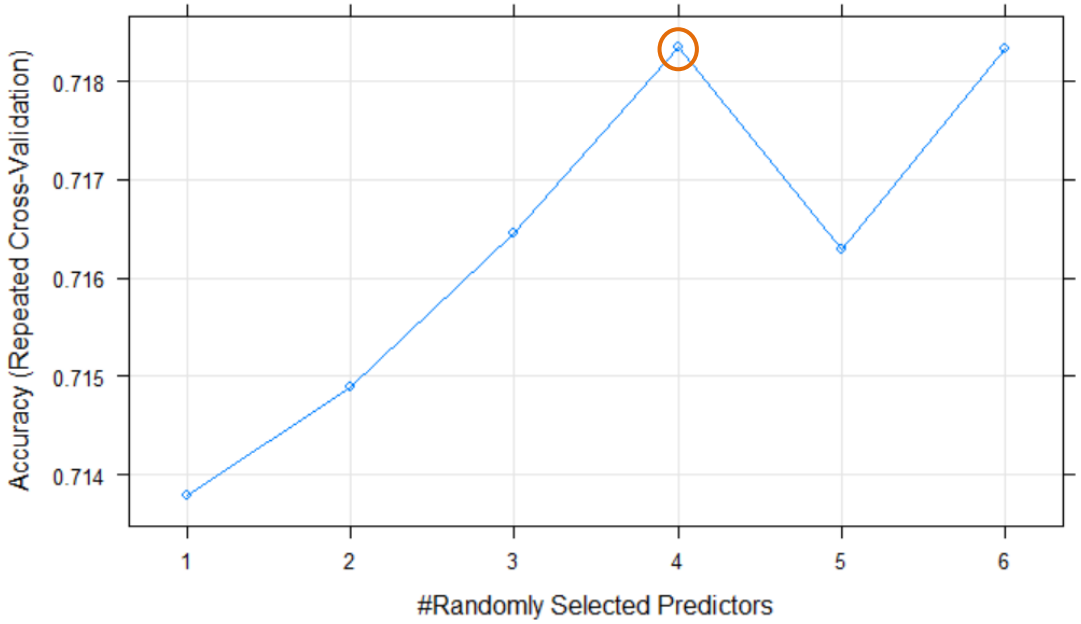
Taula 50. Elecció nombre mtry mitjançant Accuracy

Cal tenir clar que indiquen els valors d'Accuracy i Kappa:

- ***Accuracy:*** és la precisió. Percentatge de classificacions correctes de totes les observacions.

- ***Kappa:*** és la precisió de la classificació. Ajusta l'efecte de l'atzar en la proporció de la concordància observada de les observacions.

En el següent gràfic podem visualitzar-ho millor:



Gràfic 63. Nombre aleatori de predictors a seleccionar

Veiem com la major presició (accuracy) és igual al 72% aproximadament quan mtry = 4.

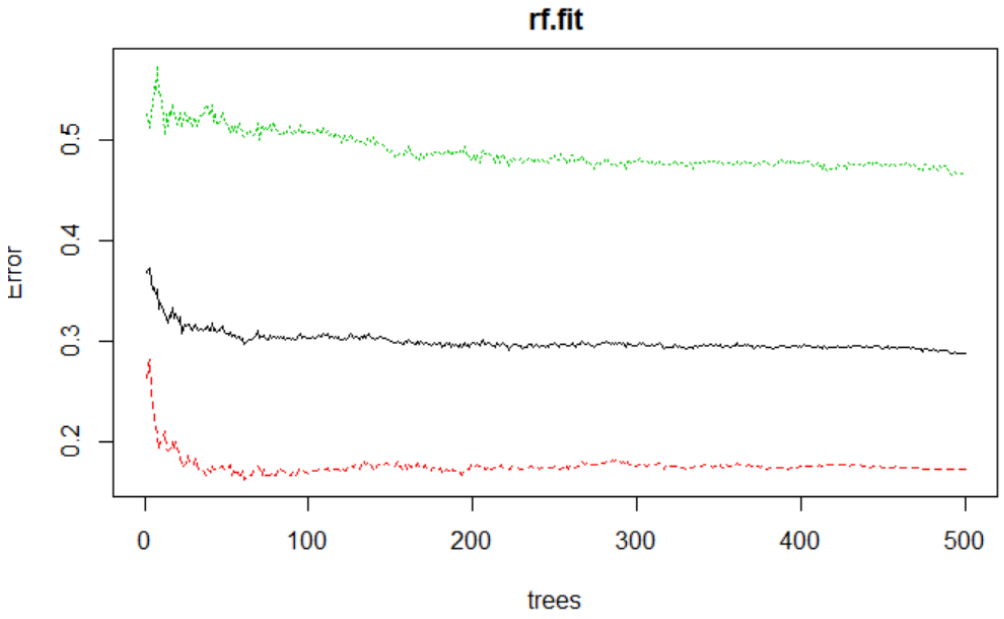
Una vegada establert el valor de mtry executem la funció randomForest amb mtry=4 i tenint la variable Class com la variable objectiu del nostre model i obtenim:

```
 randomForest(formula = Class ~ ., data = train, mtry = 4, importance = TRUE,       nodesize = 30)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 4

        OOB estimate of  error rate: 28.71%
Confusion matrix:
      FALSE TRUE class.error
FALSE   701  145   0.1713948
TRUE    257  297   0.4638989
```

Taula 51. Solució algoritme Random Forest amb mtry = 4

Com l'algoritme *randomForest* selecciona una mostra amb reemplaçament per crear un arbre en una iteració, algunes observacions es queden fora (*out of bag*) i no són utilitzades per crear l'arbre. Per aquestes observacions que s'han quedat fora de l'arbre, es fa una predicció i es calcula quin és l'error de la predicció. Veiem que el nostre error estimat (*OOB error*) és del 28.71%.

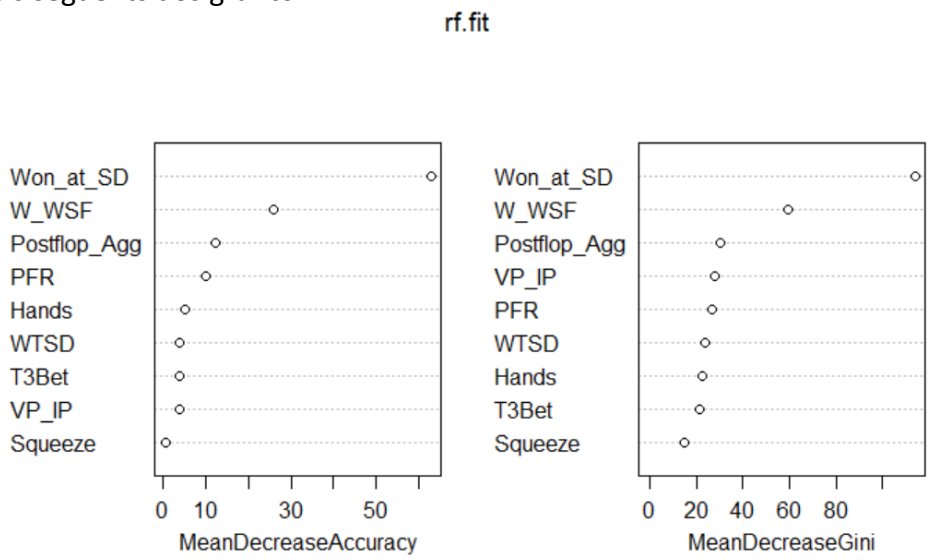
Si volem saber el promig de l'error a mesura que s'agreguen més arbres podem graficar aquest error de la següent manera:



Gràfic 64. Evolució error estimat segons número de trees

On la línea negra representa l' *OOB*, la línea vermella representa l'error al intentar predir *Net Won* > 0 igual a *FALSE* i la verda representa l'error al intentar predir *Net Won* > 0 igual a *TRUE*. Cal comentar que, pot ser, no caldrien tant arbres com 500 ja que vora els 100 arbres veiem que s'estabilitza l'error.

Volem, per tant, saber quines variables són les més significatives o importants per al model predictiu segons el nostre algoritme. Utilitzem la funció *varImpPlot* per visualitzar-les i obtenim els següents dos gràfics:



Gràfic 65. Mesures de la importància de cada variable pel model predictiu

El gràfic de l'esquerra mostra el fet que si a una variable se li assignen valors per permutació aleatòria quan augmentarà el *MSE*. Aleshores, en el nostre cas si permutem el *Won $ at SD* aleatòriament, el *MSE* augmentarà en un 65% en un promig. Té sentit que els guanys nets

(*Net Won* positiu o negatiu) tinguin relació amb si es guanya o es perd en el *showdown.* D'altre banda, el gràfic de la dreta mesura la puresa del node que es medeix amb l'índex de Gini (diferència entre *RSS* d'abans i de després de la divisió per aquella variable).

Veiem que per tots dos gràfics, les 3 primeres variables més significatives són les mateixes però l'ordre d'importància de las següents variables no ho és. Com el concepte del criteri d'importància de les variables és diferent per aquests dos casos, té diferents classificacions per diferents variables. No existeix un criteri fix per seleccionar la "millor" mesura d'importància de les variables, dependrà del problema que volguem mesurar.

A simple vista, i pel que em vist en els test de significació bivariant, la variable *VP$IP* és important per a l'estudi de possibles estratègies, així que seguiré el gràfic dret.

Seguidament obtenim les prediccions per a tots els arbres amb la funció *predict* i obtenim la següent matriu de confusió a partir de les dades *test*:

	pred	
	FALSE	TRUE
FALSE	383	71
TRUE	120	148

Taula 52. Matriu de confusió de les prediccions pels arbres

I obtenim una precisió de: (383 + 148) / 722 = 0.73545. És a dir, hem aconseguit una precisió, aproximada, del **73.55%** amb aquest model predictiu.

Per acabar, utilitzem la funció *getTree* per tal de mostrar-nos com serien les "regles" d'un arbre específic del nostre randomForest mesurat.

	left daughter <dbl>	right daughter <dbl>	split var <fctr>	split point <dbl>	status <dbl>	prediction <chr>
1	2	3	VP_IP	26.9918780	1	*NA*
2	4	5	W_WSF	47.7975576	1	*NA*
3	6	7	WTSD	38.1070247	1	*NA*
4	8	9	Postflop_Agg	3.2680858	1	*NA*
5	10	11	Postflop_Agg	3.9806716	1	*NA*
6	12	13	Won_at_SD	47.4401531	1	*NA*
7	14	15	Won_at_SD	53.8461524	1	*NA*
8	16	17	Won_at_SD	50.5088485	1	*NA*
9	18	19	VP_IP	13.5232022	1	*NA*
10	20	21	PFR	10.4597701	1	*NA*

1-10 of 123 rows Previous 1 2 3 4 5 6 ... 13 Next

Taula 53. Informació arbre de classificació RF

Les files que posa que hi ha 123 són el nombre total de nodes de l'arbre. Les columnes següents indiquen:

- *Left daughter:* fila on el node "fill" esquerre està.
- *Right daughter:* fila on el node "fill" dret està.
- *Split var:* quina variable s'ha utilitzat per separar el node.

- *Split point:* on està la millor partició (expressat en % com les nostres dades).
- *Status:* si és (-1) o no (1) terminal el node.
- *Prediction:* la predicció pel node. Si és 0 el node no és terminal.

En el cas dels *Random Forest* els arbres són molt profunds per això hi ha tantes pàgines... És complex descriure tot un arbre en concret del *Random Forest*, per tant hem decidit calcular-ne un de més senzill mitjançant un model CART, a continuació.

CART:

El model de classificació i regressió d'arbres de decisió *CART* ens permetrà ajustar un sol arbre que podrem representar visualment.

Aquest model CART es pot esquematitzar en 4 fases:

1- Construcció (*building*) de l'arbre

2- Parada (*stopping*) del procés de creixement de l'arbre. Es construeix un arbre màxim que sobre ajusti la informació obtinguda a la nostre base de dades.

3- Podat (*pruning*) de l'arbre, fent-lo així el més senzill(menor mida) possible i deixant només aquells nodes més importants. Té la finalitat de reduir el sobre ajust (*overfiting*) a les dades de *train* i així disminuir el *test error*.

4- Selecció *(selection)* de l'arbre òptim.

La construcció d'aquest arbre, per tant, començarà amb el node *root* (arrel) que inclourà tots els registres de la nostra base de dades. A partir d'aquest node l'algoritme buscarà la variable més adequada per partir-lo en dos subnodes. Per triar aquesta variable es farà servir una mesura de puresa (*purity*). Una de les funcions més utilitzades per fer-ho és la *Gini*. L'índex de Gini ens ajudarà a mesurar quines dades de cada camp estan prenent menys part en la presa de decisions del procés.

Es calcula així:

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

On p_i és la probabilitat de que una observació sigui classificada en una classe en concret.

CART SEGONS *NET WON:*

Aleshores, a partir dels *dataframe train* i *test* calculats en l'apartat anterior (*Random Forest*) construirem el codi per la predicció del model CART.

Construïm, doncs el model i n'observem els resultats següent:

```
Classification tree:
rpart(formula = Class ~ ., data = train, method = "class")

Variables actually used in tree construction:
[1] Postflop_Agg W_WSF       Won_at_SD

Root node error: 554/1400 = 0.39571

n= 1400

        CP nsplit rel error  xerror     xstd
1 0.182310      0   1.00000 1.00000 0.033027
2 0.064982      1   0.81769 0.84477 0.031861
3 0.021661      2   0.75271 0.80325 0.031449
4 0.014440      4   0.70939 0.76895 0.031075
5 0.010000      5   0.69495 0.76895 0.031075
Call:
rpart(formula = Class ~ ., data = train, method = "class")
  n= 1400

          CP nsplit rel error    xerror       xstd
1 0.18231047      0 1.0000000 1.0000000 0.03302678
2 0.06498195      1 0.8176895 0.8447653 0.03186083
3 0.02166065      2 0.7527076 0.8032491 0.03144908
4 0.01444043      4 0.7093863 0.7689531 0.03107495
5 0.01000000      5 0.6949458 0.7689531 0.03107495

Variable importance
   Won_at_SD        W_WSF Postflop_Agg          PFR        T3Bet        VP_IP         WTSD
          51           16           15            5            4            3            3
       Hands      Squeeze
           2            1
```

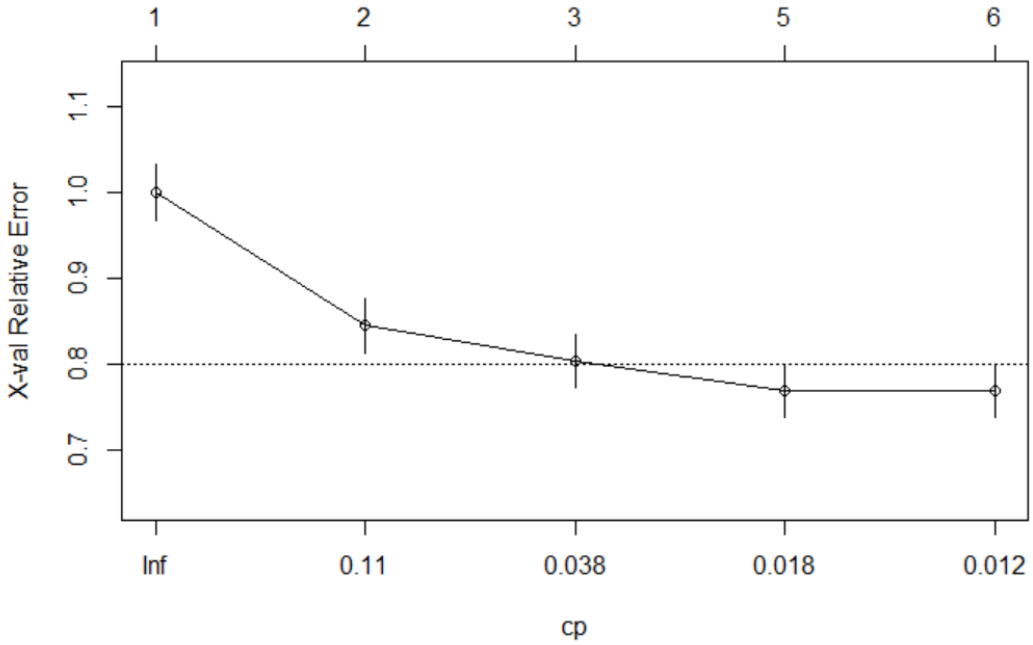
Taula 54. Elecció nsplit mitjançant xerror i CP

Podem observar com les variables seleccionades per a l'explicació del model i la construcció de l'arbre són les variables: *Postflop Agg%, W$WSF%* i *Won $ at SD*.
Els paràmetres de complexitat (*CP)* ens ajuden a controlar la mida de l'arbre. Quan major és el valor del paràmetre *CP* menys decisions conté l'arbre (*nsplit*). Cal tenir clar que si un arbre té molts nodes tindrà més complexitat tot i que cometrà menys errors però, no obstant, es sobreajustarà al *train dataset*. És preferible doncs tenir un arbre menys complexe (més senzill) però que funcioni millor amb el *test dataset*.

Després, el valor *rel error* ens indica quina és la desviació mitjana de l'arbre dividida per la desviació mitjana de l'arbre nul, és a dir *nsplit*=0. El valor de *xerror* és el valor mitjà estimat mitjançant el procediment *cross validation* i el valor de *xstd* és l'error estàndard de l'error relatiu.

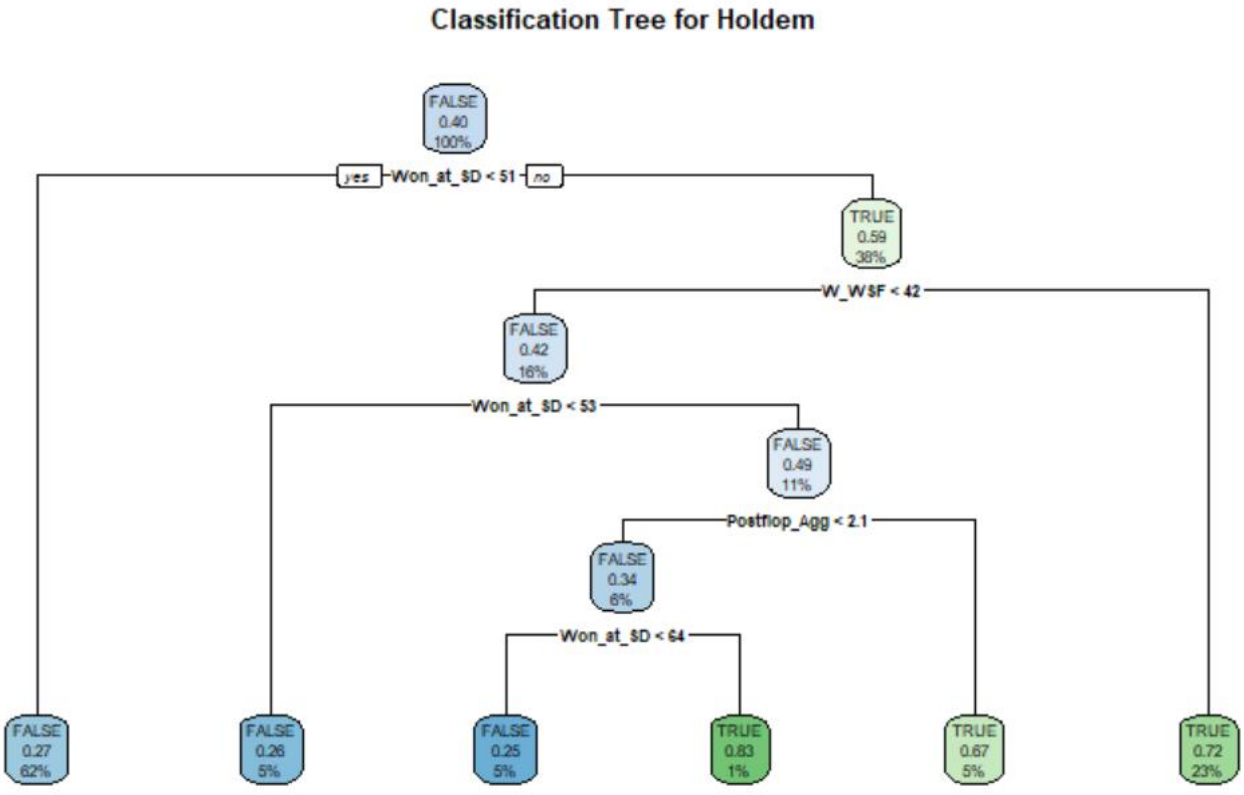
De tots aquests valors, ens hem de fixar en el *xerror*, per tal de prendre la decisió de la poda. Triarem aquell que tingui un *xerror* menor. Veiem que en el nostre cas hem triat *xerror* = 0.7689531, que correspón a un *CP* = 0.01 i un número de divisions *nsplit* = 5.

En el gràfic següent podem observar com evoluciona el *xerror* a mesura que augmenta el valor *CP*:



Gràfic 66. Evolució xerror segons CP

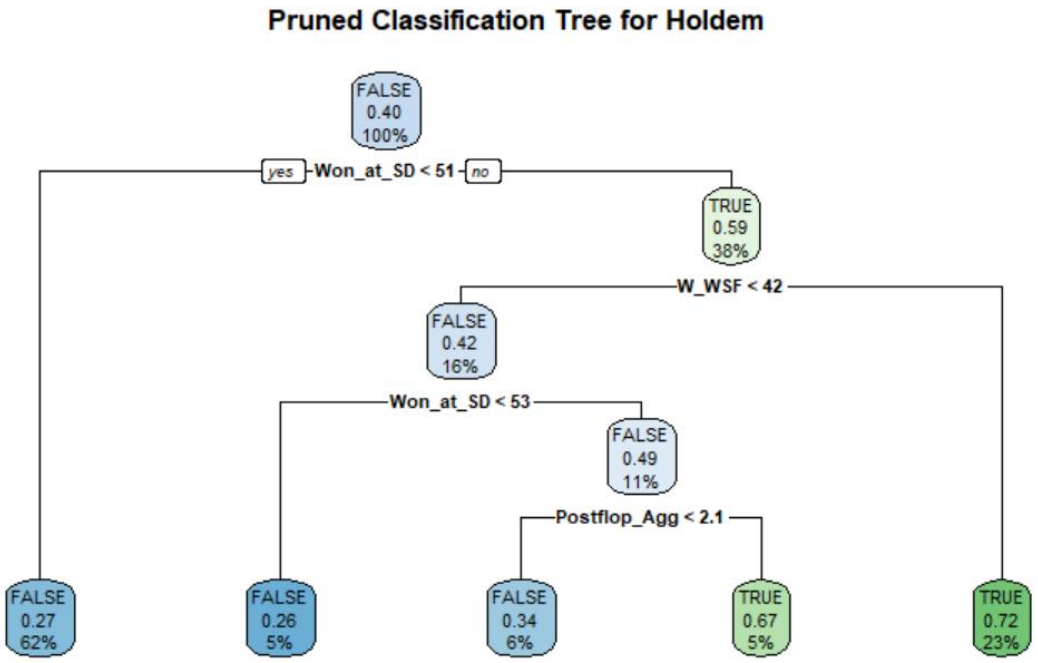
El *xerror* disminueix molt notablement a mesura que el valor de *CP* es menor.
Podem visualitzar la distribució d'aquest arbre abans de podar:



Gràfic 67. Arbre de classificació de la base de dades Holdem

Cada node està marcat per un requadre amb la seva regla de classificació especificada dins seu. Veiem que els números dins seu representa el total de casos representats pel node en concret. Els colors identifiquen quina és la categoría més abundant per aquell node en concret. Observem com el node arrel només té un subnode si la variable *Won $ at SD* és més gran que el 51%.

Si li apliquem la "poda" a el model anterior seleccionant el que té el *xerror* menor, obtenim l'arbre de decisió podat següent:



Gràfic 68. Arbre de classificació podat de la base de dades Holdem

Un cop vist l'arbre de classificació podat representat, podem estudiar si té o no sentit la classificació creada en quan si guanyen o perden. D'entrada i mirant el primer node veiem que té sentit ja que aquells jugadors/es que guanyen menys d'un 51% de les vegades que arriben al *showdown* els classifica directament com a perdedors o com a estratègia perdedora. Seguidament, si d'aquells que guanyen més del 51% de les vegades que arriben al showdown també es dóna el cas que guanyen més del 42% de les vegades que veuen el flop diretament podem dir que són guanyadors/es o qué segueixen una estratègia guanyadora. Que guanyi un percentatge tant alt quan encara queden 2 cartes per sortir (*turn* i *river*) és significatiu que porten una bona mà o que han sapigut jugar contra el rival. Aquells jugadors que no són capaços de guanyar més del 42% de les vegades que veuen el flop però que tenen un *Postflop Agg%* superior al 2.1% també esdevindran guanyadors.

Per tant, estudiades les regles de classificació de l'arbre predit podem afirmar que tenen sentit a l'hora d'identificar possibles estratègies guanyadores i perdedores.

Finalment calculem les prediccions i obtenim la següent matriu de confusió, a partir de les dades *test*:

```
        0   1
FALSE 385  69
TRUE  126 142
```

```
accuracy<-sum(diag(cm))/sum(cm)
accuracy
```

```
[1] 0.7299169
```

Taula 55. Matriu de confusió de les prediccions pels arbres CART i Accuracy

S'observa com el model prediu moltes més pèrdues (*Net Won* > 0 == *FALSE*) que no pas guanys, amb una precisió del 73% aproximadament. És a dir, prediu més estratègies perdedores que no pas guanyadores. Observem també que els *TRUE* (Guanyadors/es) són difícils de classificar ja que n'hi ha 126 mal classificats.

Tot i que aquests arbres de decisió amb la variable *Class* **<-** *Net Won* > 0 ens han donat la informació pertinent per saber com es podria prediure aquesta variable volem veure, a partir dels clústers trobats en l'apartat del clustering amb la ponderació de les dues matrius de distàncies de les variables, quines variables explicarien millor el seu model predictiu.

RANDOM FOREST SEGONS CLÚSTERS IDENTIFICATS:

A partir del nostre *dataframe* amb totes les variables numèriques (*num*), afegim una variable (columna) nova anomenada *Class* que farà referència als clústers trobats anteriorment en l'apartat clustering on vam ponderar les dues matrius de distàncies de les variables numèriques discretes i contínues. Aquesta variable tindrà 5 factors (1,2,3,4,5) depenent del clúster al que pertanyi aquell individu o jugador (fila).

En aquest cas, utilitzem també el mètode *repeatedcv* per dividir el nostre conjunt de dades, dividir-lo en 10 *folds* (plecs) de validació creuada i ho repetim 3 vegades.

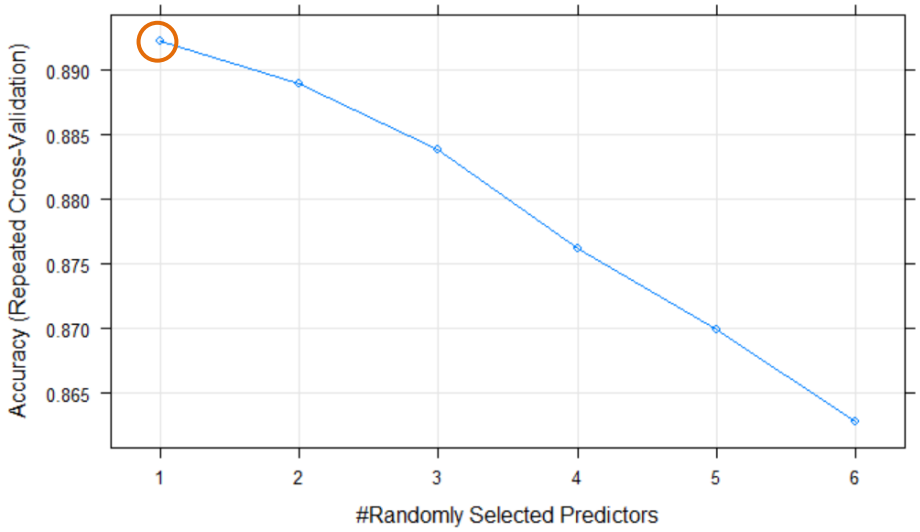
El paquet *caret* de l'R ens proporcionarà el nombre de variables aleatòries (*mtry*) a seleccionar. Li diem que, de les 9 variables predictores possibles, n'agafi de 1 a 6 i la mètrica *Accuracy* ens seleccionarà el model més òptim utilitzant el valor més alt.

Observem, doncs, que el valor final utilitzat pel model serà de ***mtry* = 1**, en aquest cas:

```
Random Forest

2122 samples
   9 predictor
   5 classes: '1', '2', '3', '4', '5'

No pre-processing
Resampling: Cross-Validated (5 fold, repeated 3 times)
Summary of sample sizes: 1697, 1698, 1698, 1697, 1698, 1698, ...
Resampling results across tuning parameters:

  mtry  Accuracy   Kappa
  1     0.8922338  0.8622265
  2     0.8889356  0.8582613
  3     0.8839079  0.8518951
  4     0.8762138  0.8420922
  5     0.8699293  0.8341551
  6     0.8628624  0.8250769

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was mtry = 1.
```

Taula 56. Elecció nombre mtry mitjançant Accuracy

De forma més visual, observem:



Gràfic 69. Nombre aleatori de predictors a seleccionar

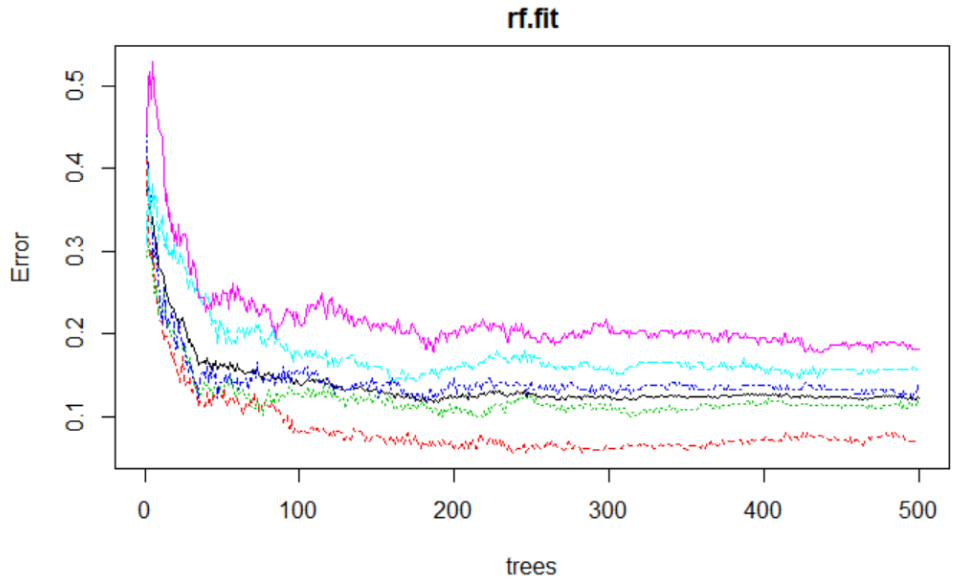
En aquest plot es visualitza millor el mtry triat, és el punt més alt amb major *Accuracy*. Aquest igual al 89.22% aproximadament quan mtry = 1.

Seguidament, executem la funció randomForest amb mtry = 1 i tenint la variable *Class* dels 5 clústers trobats anteriorment com la variable resposta del nostre model obtenim:

```
 randomForest(formula = Class ~ ., data = train, mtry = 1, importance = TRUE)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 1

        OOB estimate of  error rate: 12.21%
Confusion matrix:
    1   2   3   4   5 class.error
1 359   9  11   0   7  0.06994819
2  25 319   0   7  10  0.11634349
3  22   0 184   1   4  0.12796209
4   6  18   1 183   9  0.15668203
5  13  11  10   7 184  0.18222222
```

Taula 57. Solució algoritme Random Forest amb mtry = 1

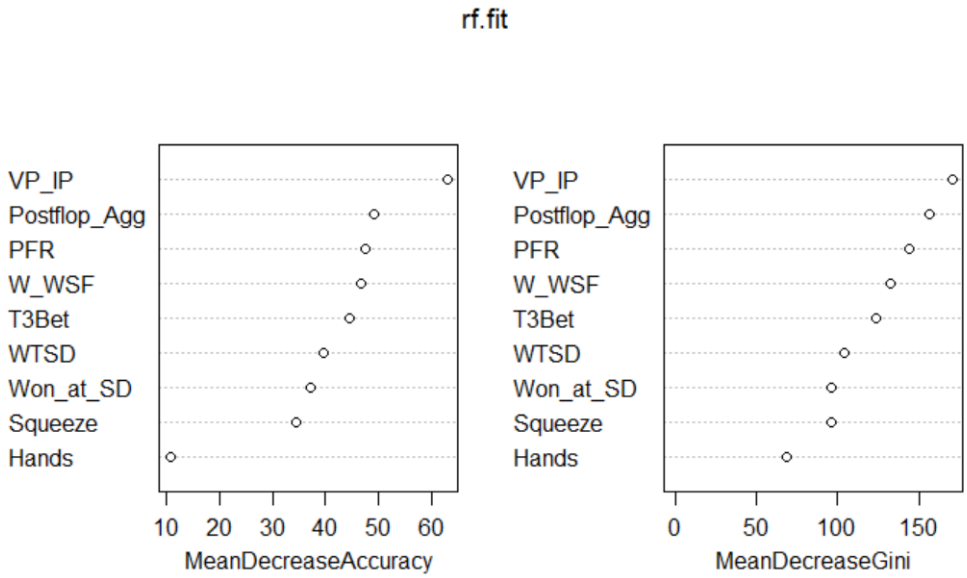
Per les observacions que s'han quedat fora (*out of bag*) i no han estat utilitzades per la creació de l'arbre, es fa una predicció i es calcula quin és l'error de la predicció. Veiem que el nostre error estimat (*OOB error*) és del 12.21%, molt menor que en el *RF* amb la variable resposat *Class <- Net Won > 0*.
Si volem saber el promig de l'error a mesura que s'agreguen més arbres podem graficar aquest error de la següent manera:



Gràfic 70. Evolució error estimat segons número de trees

On la línea negra representa l' *OOB*, i la resta de línies representen l'error de predicció per cada un dels 5 clústers. Els colors de les línies estan marcats en els requadres de cada un dels clústers en el output de la pàgina anterior. El clúster número 5 és el que té més error de predicció a mesura que augmentem el número d'arbres. La resta de clústers tenen un error de predicció bastant proper entre tots ells, sent el clúster número 1 el que en té menys.

Si volem saber quines variables són les més significatives o importants per al model predictiu segons el nostre algoritme, cal utilitzar la funció *varImpPlot* per visualitzar-les i obtenir els següents dos *plots*:



Gràfic 71. Mesures de la importància de cada variable pel model predictiu

El gràfic de l'esquerra mostra el fet que si a una variable se li assignen valors per permutació aleatòria quan augmentarà el *MSE*. D'altre banda, el gràfic de la dreta mesura la puresa del node que es medeix amb l'índex de Gini (diferència entre *RSS* d'abans i de després de la divisió per aquella variable). Veiem que per tots dos gràfics, l'ordre d'importància de les variables és idènticament el mateix, tot i que amb diferents mètriques cada gràfic. Les 3 més importants o significatives veiem que són, en ordre: *VP$IP, Postflop Agg% i PFR*. No existeix un criteri fix per seleccionar la "millor" mesura d'importància de les variables, dependrà del problema que volguem mesurar.

Seguidament obtenim les prediccions per a tots els arbres amb la funció *predict* i obtenim la següent matriu de confusió:

```
 pred
    1   2   3   4   5
1 189   5   2   0   3
2  10 159   0   7   3
3  17   1  72   2   3
4   3   9   0 127   4
5   2   1   4   3  96
```

```
966  accuracy<-sum(diag(cm))/sum(cm)
967  accuracy
968  ```
```

```
[1] 0.8905817
```

Taula 58. Matriu de confusió de les prediccions pels arbres i Accuracy

D'aquesta matriu de confusió podem comentar com el clúster 2 i 3 tenen més dificultat de classificar. L'Accuracy d'aquestes prediccions és del 89.06%.

A continuació analitzem quina és la confusió per cada clúster en concret:

Clústers	Confusió	% confusió
1	(5+2+3) / (189+5+2+3)	5.03%
2	(10+7+3) / (159+10+7+3)	11.17%
3	(17+1+2+3) / (72+17+1+2+3)	24.21%
4	(3+9+4) / (127+3+9+4)	11.19%
5	(2+1+4+3) / (96+2+1+4+3)	9.43%

Taula 59. Taula de confusió per cada clúster

Veim com el clúster 3 és el que té més confusió. No obstant, podem concloure que els clústers que hem trobat són bastant coherents i determinen una tipologia clara de maneres de jugar distintes.

Finalment, utilitzem la funció *getTree* per tal de mostrar-nos com serien les "regles" d'un arbre específic del nostre *randomForest* mesurat:

	left daughter <dbl>	right daughter <dbl>	split var <fctr>	split point <dbl>	status <dbl>	prediction <chr>
1	2	3	W_WSF	43.9999997	1	*NA*
2	4	5	Hands	855.0000000	1	*NA*
3	6	7	VP_IP	32.3358649	1	*NA*
4	8	9	Won_at_SD	45.1027383	1	*NA*
5	10	11	Hands	3803.5000000	1	*NA*
6	12	13	Postflop_Agg	3.1120251	1	*NA*
7	14	15	W_WSF	50.9397063	1	*NA*
8	16	17	W_WSF	39.5993341	1	*NA*
9	18	19	WTSD	33.4733890	1	*NA*
10	20	21	Squeeze	6.4866760	1	*NA*

1-10 of 553 rows Previous 1 2 3 4 5 6 ... 56 Next

Taula 60. Informació arbre de classificació RF

CART SEGONS CLÚSTERS TROBATS:

Aleshores, a partir dels *dataframe train* i *test* calculats en l'apartat anterior (*Random Forest*) construirem el codi per la predicció del model CART.

Construïm, doncs el model i n'observem els resultats següent:

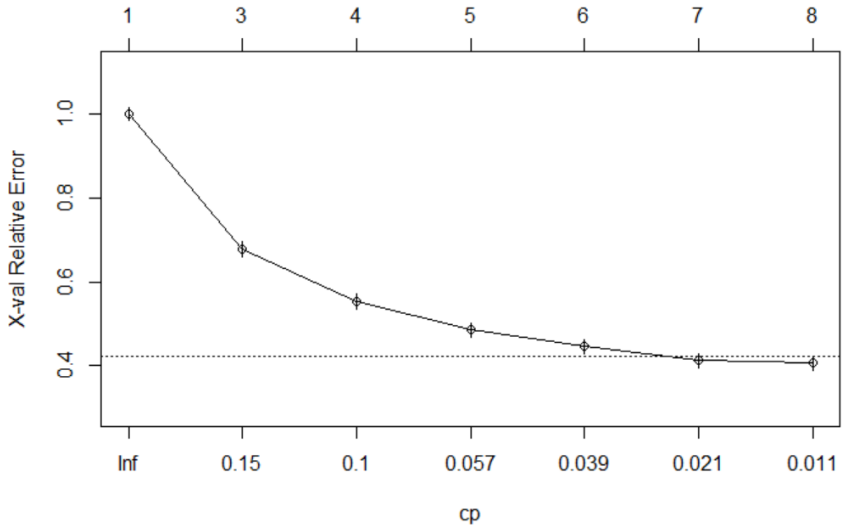
```
Classification tree:
rpart(formula = Class ~ ., data = train, method = "class")

Variables actually used in tree construction:
[1] PFR          Postflop_Agg T3Bet        VP_IP        W_WSF

Root node error: 1014/1400 = 0.72429

n= 1400

        CP nsplit rel error  xerror     xstd
1 0.163708      0   1.00000 1.00000 0.016490
2 0.139053      2   0.67258 0.67850 0.018447
3 0.077909      3   0.53353 0.55227 0.018077
4 0.042406      4   0.45562 0.48521 0.017617
5 0.035503      5   0.41321 0.44576 0.017253
6 0.012821      6   0.37771 0.41223 0.016887
7 0.010000      7   0.36489 0.40631 0.016816

Variable importance
Postflop_Agg          PFR        VP_IP        T3Bet        W_WSF       Squeeze
WTSD
          19           17           16           16           15            8
3
   Won_at_SD        Hands
           3            2
```

Taula 61. Elecció nsplit mitjançant xerror i CP

Podem observar com les variables seleccionades per a l'explicació del model i la construcció de l'arbre són les variables: *PFR*, *Postflop Agg%, VP$IP* i *W$WSF*.

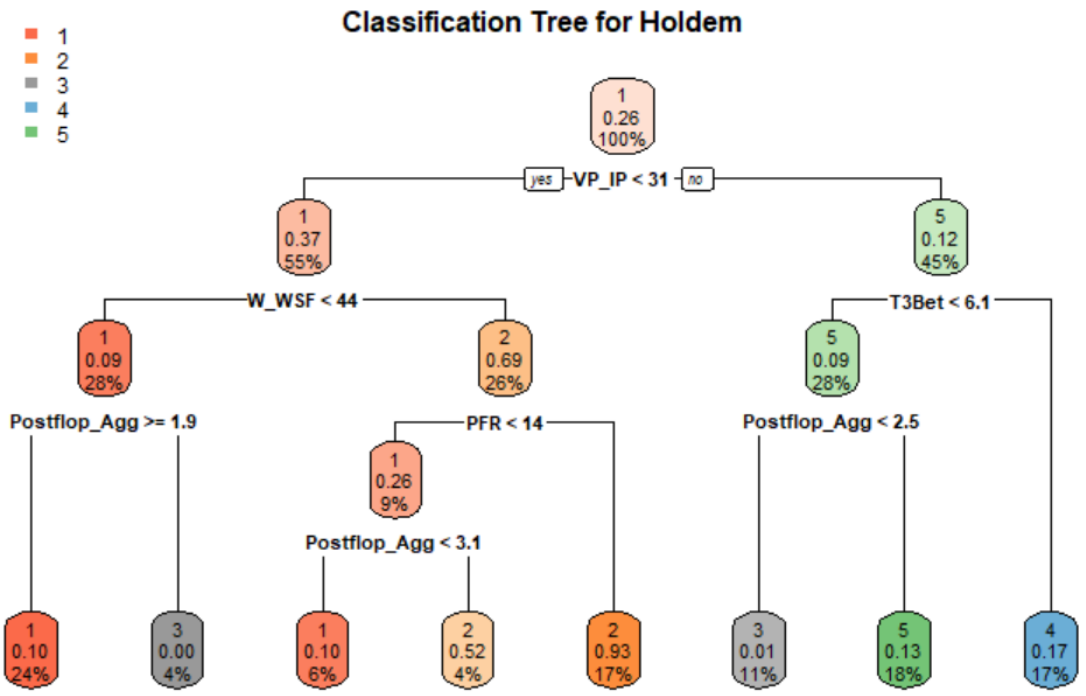
Si ens fixem en el *xerror*, per tal de prendre la decisió de la poda, triarem aquell que tingui un *xerror* menor. Veiem que en el nostre cas hem triat *xerror* = 0.40631, que correspón a un *CP* = 0.01 i un número de divisions *nsplit* = 7.

En el gràfic següent podem observar com evoluciona el *xerror* a mesura que augmenta el valor *CP*:



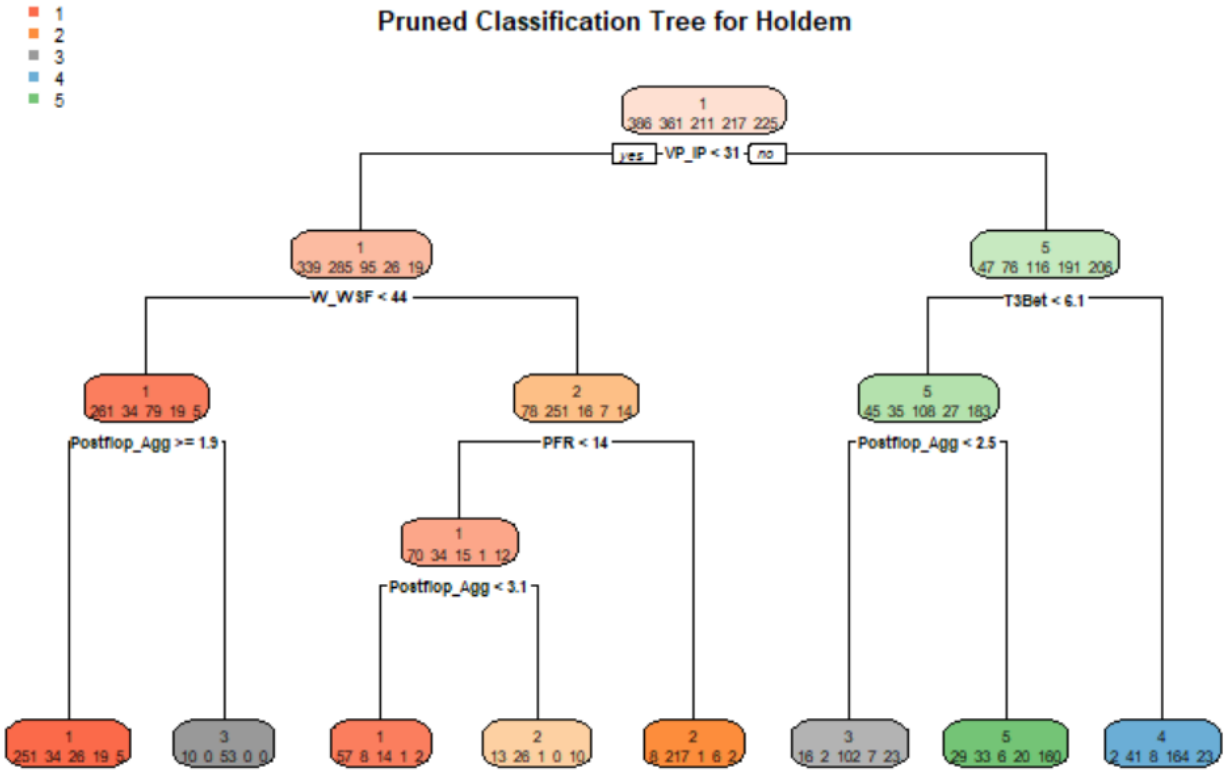
Gràfic 72. Evolució xerror segons CP

El *xerror* disminueix molt notablement a mesura que el valor de *CP* es menor.
Podem visualitzar la distribució d'aquest arbre abans de podar:



Gràfic 73. Arbre de classificació de la base de dades Holdem

Si li apliquem la "poda" a el model anterior seleccionant el que té el *xerror* menor, obtenim l'arbre de decisió podat següent:



Gràfic 74. Arbre de classificació podat de la base de dades Holdem

Observem com el node arrel en el qual s'origina l'arbre de decisió ja podat parteix de les observacions del clúster 1 i; es bifurca en dos camins depenent de si el *VP$IP* és inferior al 31% o no. Com vam observar en l'apartat del clustering, el clúster número 1 era l'únic amb *Net Won* positiu, així que aquest arbre parteix d'una estratègia amb guanys on el joc d'entrada és un joc compensat entre les diferents variables.
Aleshores, tenim que un 55% tenen un valor inferior al 31% *d'VP$IP* i un 45% superior. Aquells amb un *VP$IP* alt si tenen, també, un *3Bet* superior al 6.1% passen a formar part de l'estratègia del clúster 4 on segueixen una estratègia plena de farols amb un *3Bet* i *Postflop Agg%* elevadíssim. Una estratègia massa agressiva. En canvi, si el valor del 3Bet és inferior al 6.1% i a més a més resulta que tenen un *Postflop Agg%* superior al 2.5% formaríen part de l'estratègia seguida pel clúster 5 que seguia una estratègia molt perdedora on només guanyàven un 41% de les vegades que arribàven al *showdown*. Si el valor del *Postflop Agg%* passa a ser inferior al 2.5% formaríen part del clúster 3 que era una estratègia passiva però que es juga de manera més random (aleatòria) que no pas la del clúster 2 on tenien rangs establerts.
Si tornem al node *root* on teníem que un 55% dels jugadors sí que tenien un valor *d'VP$IP* decent, inferior al 31% veiem com depenent si guanyen o no més del 44% de les vegades que veuen el *flop* formaran part d'un clúster o un altre. Per aquells que guanyin menys del 44% de les vegades que veuen el *flop* però tinguin una agressivitat *postflop* superior o igual al 1.9% formaran part del clúster 1 on segueixen una estratègia guanyadora ja que tenen tot el seu joc compensat entre agressivitat i rangs concrets. En canvi si la seva agressivitat *postflop* és poca formaran part del clúster 3 que era una estratègia perdedora passiva però jugada de manera aleatòria.
Per aquells jugadors que no són capaços de guanyar més del 44% de les vegades que veuen el *flop* i a sobre tenen un *PFR* superior al 14% formaran part del clúster número 2 on seguien una estratègia passiva, mediosa però amb poques pèrdues. Per aquells que tenen un *PFR* inferior al 14% i, a més a més, tenen una agressivitat *postflop* superior al 3.1% també estaran dins el clúster 2, si no formaran part del clúster 1.

Per tant, aquest arbre tindria dues seccions on a la dreta de l'arbre tindríem aquelles estratègies més perdedores (clúster 4 i 5) i a l'esquerra tindríem aquelles estratègies més guanyadores (clúster 1 i 2). El clúster número 3, al tenir una estratègia tan random forma part de tots dos grups o tipus d'estratègies.

Quedaríen per tant, definides les estratègies, de forma conscisa, com:

Estratègies Guanyadores	
Clúster 1	**Clúster 2**
Agressivitat necessària per robar pots i obligar rival a abandonar la mà. Joc compensat entre variables.	Poca iniciativa però juguen mans dintre un rang.

Estratègies Perdedores	
Clúster 4	**Clúster 5**
Molta agressivitat en els seus farols. Valors exagerats. Els agrada prendre la iniciativa.	Jugadors maníacs, molt agressius que perden stack ràpidament i guanyen molt poc al river.

Estratègia guanyadora i perdedora
Clúster 3
Poca iniciativa però juguen moltes mans i de manera random (aleatòria).

Taula 62. Característiques de les estratègies segons clusters

Finalment calculem les prediccions i obtenim la següent matriu de confusió, a partir de les dades *test*:

```
     1   2   3   4   5
1  158  10  12   0  19
2   19 129   0  20  11
3   28   1  55   6   5
4    8   7   5 107  16
5    1   2  11  13  79
```

```
accuracy<-sum(diag(cm))/sum(cm)
accuracy
```

```
[1] 0.7313019
```

Taula 63. Matriu de confusió de les prediccions pels arbres CART i Accuracy

I la confusió per cada clúster igual a :

Clústers	Confusió	% confusió
1	(10+12+19) / (158+10+12+19)	20.60%
2	(19+20+11) / (129+19+20+11)	24.12%
3	(28+1+6+5) / (55+28+1+6+5)	42.11%
4	(8+7+5+16) / (107+8+7+5+16)	25.17%
5	(1+2+11+13) / (79+1+2+11+13)	25.47%

Taula 64. Taula de confusió per cada clúster

Veiem com el model prediu més estratègies pertanyents al clúster 1,2 i 4 que no pas als clústers 3 i 5. No obstant, el clúster número 3 és el que té més confusió de tots cinc.

IX. CONCLUSIONS

Com ja prevèiem en un inici, trobar la predicció d'un model que ens permeti identificar estratègies pel pòquer ha estat complicada ja que existeixen molts factos incontrolables que intervenen molt directament en el joc. Alguns d'aquests factors són l'atzar, el lloc on es seu a la taula, l'*stack* del que es disposa, etc. No obstant, si deixem tots aquests factors de banda i ens centrem en aquells que sí podem controlar o millorar per tal de conseguir un joc compensat que ens aporti beneficis, es poden trobar diverses estratègies.

L'objectiu principal era trobar aquestes estratègies i definir-les. Ho hem aconseguit fer mitjançant els diversos clusterings i presentant els diversos arbres de classificació.
Tant l'estudi realitzant mitjançant la variable *Net Won* com l'estudi realitzat mitjançant els 5 clústers trobats em visualitzat que les dues variables més importants i més crucials per a la determinació de possibles estratègies han estat: *Postflop Agg% i W$WSF%.*
Totes dues eren significatives en l'anàlisi segregat i totes dues han esdevingut variables significactives per al clustering i els arbres de classificació. Si es consegueix arribar a veure el *flop* i després (*postflop*) tenir una agressivitat vora el 2.5% - 3%, arribarem a tenir un *W$WSF%* superior al 43% i per tant, aconseguirem seguir una estratègia guanyadora.
Per tant, és molt important tenir un rang de mans molt determinat pel qual seguir pagant un cop vist el *flop* i també, coneixer bé les estratègies seguides pels rivals per tal de saber quan poder fer un farol o no.

Hem sigut capaços, a més a més, de trobar quines són aquelles variables "secundàries" però força importants per a la classificació d'estratègies. Aquestes variables són: *VP$IP, 3Bet i PFR.* Per tal d'aconseguir seguir una estratègia guanyadora en un nivell NL5-NL10 de *Texas Hold'em Poker* cal tenir un *gap* petit entre *VP$IP i PFR*, és a dir que *l'VP$IP* estigui vora el 28% i el *PFR* estigui vora el 13%, si una d'aquestes variables és massa alta i l'altre massa petita el joc estarà molt descompensat i serà molt difícil que el jugador acabi guanyant.

En quan a la variable *3Bet* és important saber quan s'ha de realitzar, aquesta variable potser ha sortit menys significativa del que esperàvem ja que és una variable molt important pel joc però que té molta relació amb el seient que ocupa el jugador a taula. Un jugador que està assegut en primera posició per norma hauria de fer poc *3Bet* en canvi, un que està assegut en el *dealer* (botó) i coincideix amb que té una bona mà, hauria de realitzar aquest moviment. No obstant, en la classificació dels arbres segons els clústers trobats ens ha sortit força significativa per a la discriminació de possibles estratègies. Un *3Bet* vora el 6% compensat amb un bon *gap* entre *VP$IP i PFR* esdevindrà en una estratègia guanyadora.

Per l'apartat final de validació creuada hem partit de les dades de 1400 jugadors (dades *train*) i em predit les estratègies de 722 jugadors (dades *test*).

Un aspecte que valdria la pena tenir en compte seria la coincidència entre les agrupacions que s'obtenen amb els diversos mètodes de clúster que hem aplicat. Per exemple, cada observació quines són les seves observacions veïnes més properes (posem les 10 més properes) en cada Clustering i mesurar si hi ha molta coincidència. Aquest punt no hem tingut temps d'abordar-lo en aquest treball però, potser, es podria posar com a tasca a desenvolupar en un futur pròxim.

Per concloure dir que ara per ara estic seguint una estratègia guanyadora ja que 4 dels 5 clústers trobats, estudiant els meus oponents, han esdevingut estratègies perdedores i això és significatiu que estic anant pel bon camí o que estic sabent jugar envers ells.

X. BIBLIOGRAFIA

LLIBRES:

- Peña, D. (2013). Análisis de datos multivariantes. McGraw-Hill España.
- Izenman, A. J. (2008). *Modern multivariate statistical techniques. Regression, classification and manifold learning*, 10, 978-0.

PÀGINES WEBS:

- https://www.holdem.es/historia/origenes-poker-evolucion-situacion-actual/
- http://www.dpye.iimas.unam.mx/curso_puma/Salvador/PUMA%20no%20parametrica.pdf
- http://www.ics-aragon.com/cursos/salud-publica/2014/pdf/M2T08.pdf
- http://www.cookbook-r.com/Graphs/Colors_(ggplot2)/
- https://ciespinosa.github.io/AnalisisMultivariante/analisis-multivariado-de-la-composicion-de-la-comunidad.html
- https://dpmartin42.github.io/posts/r/cluster-mixed-types
- https://en.cardmates.net/holdem_manager_2_review_where_and_how_to_download_the_program
- https://towardsdatascience.com/random-forest-in-r-f66adf80ec9
- https://uc-r.github.io/random_forests
- https://rpubs.com/phamdinhkhanh/389752
- https://towardsdatascience.com/how-to-visualize-a-decision-tree-from-a-random-forest-in-python-using-scikit-learn-38ad2d75f21c
- http://wwwae.ciemat.es/~cardenas/docs/lessons/RandomForest.pdf

XI. ANNEX

CODI R:

```
setwd("C:/Users/mariona.martin/Desktop/TFG MARIONA")holdem<-
read.csv(file="Summaries.csv",head=FALSE,sep=",",nrows=8054)
#holdem

names(holdem) = c("Player Name","Site","Hands","Net Won","VP$IP","PFR","3Bet","Postflop
Agg%","W$WSF%","WTSD%","Won $ at SD","Flop CBet%","Turn CBet%","River CBet%","Fold to Flop
Cbet","Fold to Turn CBet","Fold to River CBet","Raise Flop Cbet","Raise Turn CBet","Raise River
CBet","Squeeze","Raise Two Raisers","Call Two Raisers","vs 3Bet Fold","vs 3Bet Call","vs 3Bet
Raise","vs 4Bet Fold","vs 4Bet Call","vs 4Bet Raise")

names(holdem)

#Passem bdd a data.frame

holdem<-as.data.frame(holdem)
class(holdem)
dim(holdem)
str(holdem) #per conèixer l'estructura de la bdd
names(holdem)
head(holdem)
nrow(holdem)
summary(holdem)

holdem$Site<-as.factor(holdem$Site)
holdem$`Player Name`<-as.factor(holdem$`Player Name`)
holdem$Hands<-as.numeric(holdem$Hands)
holdem$`Net Won`<-as.numeric(holdem$`Net Won`)
holdem$`VP$IP`<-as.numeric(holdem$`VP$IP`)
holdem$PFR<-as.numeric(holdem$PFR)
holdem$`3Bet`<-as.numeric(holdem$`3Bet`)
holdem$`Postflop Agg%`<-as.numeric(holdem$`Postflop Agg%`)
holdem$`W$WSF%`<-as.numeric(holdem$`W$WSF%`)
holdem$`WTSD%`<-as.numeric(holdem$`WTSD%`)
holdem$`Won $ at SD`<-as.numeric(holdem$`Won $ at SD`)
holdem$Squeeze<-as.numeric(holdem$Squeeze)
#summary(holdem)

                ##################################

                ####### MODIFICACIONS BDD ######

                ##################################
```

```
holdem$`VP$IP`<-(holdem$`VP$IP`)*100
holdem$PFR<-(holdem$PFR)*100
holdem$`3Bet`<-(holdem$`3Bet`)*100
holdem$`Postflop Agg%`<-(holdem$`Postflop Agg%`)*10
holdem$`W$WSF%`<-(holdem$`W$WSF%`)*100
holdem$`WTSD%`<-(holdem$`WTSD%`)*100
holdem$`Won $ at SD`<-(holdem$`Won $ at SD`)*100
holdem$Squeeze<-(holdem$Squeeze)*100

#Mirar si hi han NA'S
#is.na(holdem)

#Com hem vist que no hi ha presència de missings (NA) en la bdd, subdividim la bdd a 14 variables en comptes de 29 i, agafem files on el número de mans jugades sigui > 300

#També eliminem el primer individu (que sóc jo) per tal de no sortir coma punt extrem en tots els gràfics, ja que jo tinc totes les meves mans guardades , en canvi tinc altres jugadors però només les mans que han estat vs mi.

col<-c(holdem$`Player Name`,holdem$Site,holdem$Hands,holdem$`Net
Won`,holdem$`VP$IP`,holdem$PFR,holdem$`3Bet`,holdem$`Postflop
Agg%`,holdem$`W$WSF%`,holdem$`WTSD%`,holdem$`Won $ at
SD`,holdem$Squeeze,holdem$`Raise Two Raisers`,holdem$`Call Two Raisers`)

holdem <- holdem [(holdem$Hands>300),]
holdem<- holdem [c(2:2222),]
holdem<-holdem[,c(1,2,3,4,5,6,7,8,9,10,11,21)]
holdem
#Per tant, tindrem 2221 observacions

#CANVIAR NOMS VARIABLE SITE

levels(holdem$Site)<-c("PokerStars","888Poker","WinaMax")
levels(holdem$Site)
summary(holdem$Site)

    ########################################
    #####ANÀLISI DESCRIPTIVA VARIABLES#####
    ########################################

    ##DESCRIPTIVA UNIVARIANT

#Variable SITE:

s<-table(holdem$Site)
s<-as.data.frame(s)
levels(holdem$Site)
colnames(s)<-c("Site","Freq")
```

```
FreqRel<-round(s$Freq/sum(s$Freq),4)
Percentatge<-FreqRel*100
s<-cbind(s,FreqRel,Percentatge)
a<-c("PokerStars","888Poker","WinaMax")
a<-paste(a,"(",Percentatge,")","%",sep="") #afegirar els % als labels
pie(s$Freq,labels=a,col=rainbow(length(a)),main="Percentatge (%) de jugadors amb els que jugo en una Plataforma o altre")
x<-sort(holdem$Site)
barplot(table(x), main= "Freqüències plataformes de joc")

#Ja que la majoria d'observacions pertanyen a la plataforma 888Poker, ens quedarem només amb aquestes

holdem<-holdem[holdem$Site=="888Poker",]  #de 2221obs passarem a 2122 observacions en la bdd
summary(holdem)
View(holdem)

#Variable HANDS:

summary(holdem$Hands)
sd(holdem$Hands, na.rm = T)
    #install.packages("descr")
    #library(descr)
    #install.packages("RColorBrewer")
    #library(RColorBrewer)

hist(sort(table(holdem$Hands)),main="Histograma Hands")
boxplot(holdem$Hands, col = "yellow", main = "Avaluació de mans jugades", xlab = "Hands", ylim = c(300,10000))

##################
#detecció d'outliers#
##################

iqr<-1188.8 - 419.0
iqr #ens dóna igual a 769.8
#mirem valors atípics lleus
val<-1188.8 + 1.5*769.8
val #2343.5,tots els valors que superin aquest número són valors atípics lleus
c<-(holdem[holdem$Hands>2343.5,])
c #236 jugadors tenen valors atípics lleus

#mirem valors atípics extrems

vae<-1188.8 + 3*769.8
vae #3498.2,tots els valors que superin aquest número són valors atípics extrems
c<-(holdem[holdem$Hands>3498.2,])
c #127 jugadors tenen valors atípics extrems
```

```
#outliers

o<-419.0 + 1.5*769.8
o #1573.7, tots els valors que superin aquest número són outliers
c<-(holdem[holdem$Hands>1573.7,])
c #372 jugadors són outliers

#Variable Net Won:

summary(holdem$`Net Won`)
sd(holdem$`Net Won`, na.rm = T)
hist(holdem$`Net Won`,xlim=c(-1300,1200),main="Histograma Net Won")
boxplot(holdem$`Net Won`, col = "blue", main = "Avaluació dels Guanys/ Pèrdues Netes", xlab = "$"
, ylim = c(-1300,1200))

#Variable VP$IP:

summary(holdem$`VP$IP`)
sd(holdem$`VP$IP`, na.rm = T)
hist(holdem$`VP$IP`, main="Histograma VP$IP")
boxplot(holdem$`VP$IP`, col = "blue", main = "Avaluació del VP$IP", xlab = "en %", ylim = c(0,100))

#Variable PFR:

summary(holdem$PFR)
sd(holdem$PFR, na.rm = T)
hist(holdem$PFR,main="Histograma PFR")
boxplot(holdem$PFR, col = "purple", main = "Avaluació del PFR", xlab = "en %", ylim = c(0,100))

#Variable 3BET:

summary(holdem$`3Bet`)
sd(holdem$`3Bet`, na.rm = T)
hist(holdem$`3Bet`,main="Histograma 3Bet")
boxplot(holdem$`3Bet`, col = "orange", main = "Avaluació del 3Bet", xlab = "en %", ylim = c(0,70))

#Variable POSTFLOP AGG%:

summary(holdem$`Postflop Agg%`)
sd(holdem$`Postflop Agg%`, na.rm = T)
hist(holdem$`Postflop Agg%`,main="Histograma Postflop Agg%")
boxplot(holdem$`Postflop Agg%`, col = "pink", main = "Avaluació del PostFlop Agg", xlab = "en %",
ylim = c(0,10))

#Variable W$WSF:

summary(holdem$`W$WSF%`)
sd(holdem$`W$WSF%`, na.rm = T)
hist(holdem$`W$WSF%`,main="Histograma W$WSF")
```

```
boxplot(holdem$`W$WSF%`, col = "yellow", main = "Avaluació del W$WSF", xlab = "en %", ylim =
c(0,80))

#Variable WTSD%:

summary(holdem$`WTSD%`)
sd(holdem$`WTSD%`, na.rm = T)
hist(holdem$`WTSD%`,main="Histograma WTSD%")
boxplot(holdem$`WTSD%`, col = "brown", main = "Avaluació del WTSD", xlab = "en %", ylim =
c(0,70))

#Variable WON $ AT SD:

summary(holdem$`Won $ at SD`)
sd(holdem$`Won $ at SD`, na.rm = T)
hist(holdem$`Won $ at SD`,main="Histograma Won $ at SD")
boxplot(holdem$`Won $ at SD`, col = "purple", main = "Avaluació del Won $ at SD", xlab = "$", ylim =
c(0,100))

#Variable SQUEEZE:

summary(holdem$Squeeze)
sd(holdem$Squeeze, na.rm = T)
hist(holdem$Squeeze,main="Histograma Squeeze")
boxplot(holdem$Squeeze, col = "brown", main = "Avaluació del Squeeze", xlab = "en %", ylim =
c(0,60))

#######################################
#####ANÀLISI DESCRIPTIVA VARIABLES#####
#######################################

##DESCRIPTIVA SEGREGADA UNIVARIANT

g<-holdem[holdem$`Net Won`>=0,]
p<-holdem[holdem$`Net Won`< 0,]
nrow(g) #822
nrow(p) #1300
summary(g)
summary(p)


#Variable HANDS:

par(mfrow=c(1,2),cex.axis= 1.5)
boxplot(g$Hands,col="green",main= "Guanys Nets Hands",xlab="$",ylim=c(0,4000))
boxplot(p$Hands,col="red",main= "Pèrdues Netes Hands",xlab="$",ylim=c(0,4000))
summary(g$Hands)
sd(g$Hands)
```

```
summary(p$Hands)
sd(p$Hands)

#Podem veure quantes mans juguen en cada grup

length(which(g$Hands<5000)) #778
length(which(g$Hands<10000)) #812 - 778 = 34
length(which(g$Hands<20000)) #818 - 812 = 6
length(which(g$Hands<30000)) #820 - 818 = 2
length(which(g$Hands<40000)) #822 - 820 = 2
hist(g$Hands,col="green",main="Histograma Jugadors Guanyadors",xlab="Número de mans
jugades")
length(which(p$Hands<5000)) #1266
length(which(p$Hands<10000)) #1292 - 1266 = 26
length(which(p$Hands<20000)) #1298 - 1292 = 6
length(which(p$Hands<30000)) #1299 - 1298 = 1
length(which(p$Hands<40000)) #1300 - 1299 = 1
hist(p$Hands,col="red",main="Histograma Jugadors Perdedors",xlab="Número de mans jugades")

#Test comparació de mitjanes
wilcox.test(g$Hands ,p$Hands)

#Variable Net Won:

boxplot(g$`Net Won`,col="green",main= "Avaluació Guanys Nets",xlab="$",ylim=c(0,1090))
boxplot(p$`Net Won`,col="red",main= "Avaluació Pèrdues Netes",xlab="$",ylim=c(-1295,0))
summary(g$`Net Won`)
sd(g$`Net Won`)
summary(p$`Net Won`)
sd(p$`Net Won`)
hist(g$`Net Won`,col="green",main="Histograma Jugadors Guanyadors",xlab="Dòlars ($)")
hist(p$`Net Won`,col="red",main="Histograma Jugadors Perdedors",xlab="Dòlars ($)")

#Test comparació de mitjanes
wilcox.test(g$`Net Won` ,p$`Net Won`)

#Variable VP$IP:

par(mfrow=c(1,2),cex.axis= 1.5)
boxplot(g$`VP$IP`,col="green",main= "J.Guanyadors VP$IP",xlab="$",ylim=c(0,100))
boxplot(p$`VP$IP`,col="red",main= "J.Perdedors VP$IP",xlab="$",ylim=c(0,100))
summary(g$`VP$IP`)
sd(g$`VP$IP`)
summary(p$`VP$IP`)
sd(p$`VP$IP`)
par(mfrow=c(1,2),cex.axis= 1.5)
hist(g$`VP$IP`,col="green",main="Histograma Jugadors Guanyadors VP$IP",xlab="%", ylim=c(0,400))
hist(p$`VP$IP`,col="red",main="Histograma Jugadors Perdedors VP$IP",xlab="%", ylim=c(0,400))
```

```
#Test comparació de mitjanes
wilcox.test(g$`VP$IP` ,p$`VP$IP`)

#Variable PFR:

par(mfrow=c(1,2),cex.axis= 1.5)
boxplot(g$PFR,col="green",main= "J.Guanyadors PFR",xlab="$",ylim=c(0,70))
boxplot(p$PFR,col="red",main= "J.Perdedors PFR",xlab="$",ylim=c(0,70))
summary(g$PFR)
sd(g$PFR)
summary(p$PFR)
sd(p$PFR)
par(mfrow=c(1,2),cex.axis= 1.5)
hist(g$PFR,col="green",main="Histograma Jug. Guanyadors PFR",xlab="%", ylim=c(0,350))
hist(p$PFR,col="red",main="Histograma Jug. Perdedors PFR",xlab="%", ylim=c(0,350))

#Test comparació de mitjanes
wilcox.test(g$PFR ,p$PFR)

#Variable 3BET:

par(mfrow=c(1,2),cex.axis= 1.5)
boxplot(g$`3Bet`,col="green",main= "J.Guanyadors 3Bet",xlab="$",ylim=c(0,50))
boxplot(p$`3Bet`,col="red",main= "J.Perdedors 3Bet",xlab="$",ylim=c(0,50))
summary(g$`3Bet`)
sd(g$`3Bet`)
summary(p$`3Bet`)
sd(p$`3Bet`)
par(mfrow=c(1,2),cex.axis= 1.5)
hist(g$`3Bet`,col="green",main="Histograma Jug.Guanyadors 3Bet",xlab="%", ylim=c(0,850))
hist(p$`3Bet`,col="red",main="Histograma Jug.Perdedors 3Bet",xlab="%", ylim=c(0,850))

#Test comparació de mitjanes
wilcox.test(g$`3Bet` ,p$`3Bet`)

#Variable POSTFLOP AGG%:

par(mfrow=c(1,2),cex.axis= 1.5)
boxplot(g$`Postflop Agg%`,col="green",main= "J.Guanyadors Postflop Agg%",ylim=c(0,8))
boxplot(p$`Postflop Agg%`,col="red",main= "J.Perdedors Postflop Agg%",ylim=c(0,8))
summary(g$`Postflop Agg%`)
sd(g$`Postflop Agg%`)
summary(p$`Postflop Agg%`)
sd(p$`Postflop Agg%`)
par(mfrow=c(1,2),cex.axis= 1.5)
hist(g$`Postflop Agg%`,col="green",main="Histograma J.Guany.Postflop Agg",xlab="%",
ylim=c(0,350))
hist(p$`Postflop Agg%`,col="red",main="Histograma J.Perd.Postflop Agg",xlab="%", ylim=c(0,350))
```

```
#Test comparació de mitjanes
wilcox.test(g$`Postflop Agg%` ,p$`Postflop Agg%`)

#Variable W$WSF:

par(mfrow=c(1,2),cex.axis= 1.5)
boxplot(g$`W$WSF%`,col="green",main= "J.Guanyadors W$WSF",ylim=c(0,70))
boxplot(p$`W$WSF%`,col="red",main= "J.Perdedors W$WSF",ylim=c(0,70))
summary(g$`W$WSF%`)
sd(g$`W$WSF%`)
summary(p$`W$WSF%`)
sd(p$`W$WSF%`)
par(mfrow=c(1,2),cex.axis= 1.5)
hist(g$`W$WSF%`,col="green",main="Histograma J.Guany. W$WSF",xlab="%", ylim=c(0,500))
hist(p$`W$WSF%`,col="red",main="Histograma J.Perd. W$WSF",xlab="%", ylim=c(0,500))

#Test comparació de mitjanes
wilcox.test(g$`W$WSF%` ,p$`W$WSF%`)

#Variable WTSD%:

par(mfrow=c(1,2),cex.axis= 1.5)
boxplot(g$`WTSD%`,col="green",main= "J.Guanyadors WTSD",ylim=c(0,60))
boxplot(p$`WTSD%`,col="red",main= "J.Perdedors WTSD",ylim=c(0,60))
summary(g$`WTSD%`)
sd(g$`WTSD%`)
summary(p$`WTSD%`)
sd(p$`WTSD%`)
par(mfrow=c(1,2),cex.axis= 1.5)
hist(g$`WTSD%`,col="green",main="Histograma J.Guany. WTSD",xlab="%", ylim=c(0,500))
hist(p$`WTSD%`,col="red",main="Histograma J.Perd. WTSD",xlab="%", ylim=c(0,500))

#Test comparació de mitjanes
wilcox.test(g$`WTSD%` ,p$`WTSD%`)

#Variable WON $ AT SD:

par(mfrow=c(1,2),cex.axis= 1.5)
boxplot(g$`Won $ at SD`,col="green",main= "J.Guanyadors Won $ at SD",ylim=c(0,90))
boxplot(p$`Won $ at SD`,col="red",main= "J.Perdedors Won $ at SD",ylim=c(0,90))
summary(g$`Won $ at SD`)
sd(g$`Won $ at SD`)
summary(p$`Won $ at SD`)
sd(p$`Won $ at SD`)
par(mfrow=c(1,2),cex.axis= 1.5)
hist(g$`Won $ at SD`,col="green",main="Histograma J.Guany. Won $ at SD",xlab="%", ylim=c(0,400))
hist(p$`Won $ at SD`,col="red",main="Histograma J.Perd. Won $ at SD",xlab="%", ylim=c(0,400))
```

```
#Test comparació de mitjanes
wilcox.test(g$`Won $ at SD` ,p$`Won $ at SD`)

#Variable SQUEEZE:

par(mfrow=c(1,2),cex.axis= 1.5)
boxplot(g$Squeeze,col="green",main= "J.Guanyadors Squeeze",ylim=c(0,50))
boxplot(p$Squeeze,col="red",main= "J.Perdedors Squeeze",ylim=c(0,50))
summary(g$Squeeze)
sd(g$Squeeze)
summary(p$Squeeze)
sd(p$Squeeze)
par(mfrow=c(1,2),cex.axis= 1.5)
hist(g$Squeeze,col="green",main="Histograma J.Guany. Squeeze",xlab="%", ylim=c(0,900))
hist(p$Squeeze,col="red",main="Histograma J.Perd. Squeeze",xlab="%", ylim=c(0,900))

#Test comparació de mitjanes
wilcox.test(g$Squeeze ,p$Squeeze)



#Primerament, realitzem una matriu de correlacions entre les variables, per visualitzar quines són aquelles variables que estan més correlacionades per així fer el seu anàlisi descriptiu bivariant corresponent.
#install.packages("PerformanceAnalytics")
library("PerformanceAnalytics")
my_data <- holdem[, c(3,4,5,6,7,8,9,10,11,12)]
chart.Correlation(my_data, histogram=TRUE, pch=19)  #AMB HISTOGRAMA
my_data <- holdem[, c(3,4,5,6,7,8,9,10,11,12)]
chart.Correlation(my_data, histogram=FALSE, pch=19) #SENSE HISTOGRAMA
my_data2 <- holdem[, c(5,6,7,8,9,10,11,12)]
chart.Correlation(my_data2, histogram=TRUE, pch=19) #DESCARTEM VARIABLES DISCRETES

#Creem nova columna amb valors "G" o "P" depenent si tenen net won positiu (G) o negatiu(P)
nueva.col<-c(seq(1:2122))
holdem$jug<-nueva.col
holdem$jug<-as.character(holdem$jug)
holdem$jug[holdem$`Net Won` >= 0] <- "G"
holdem$jug[holdem$`Net Won` < 0] <- "P"
View(holdem)                                ##DESCRIPTIVA BIVARIANT

# PFR - 3Bet

par(cex.main=1,font = 1, font.lab = 1, font.axis = 1, las = 1,  cex.lab=1, cex.axis= 0.7)
#boxplot(holdem$PFR~holdem$`3Bet`, main = "PFR según el 3Bet")
plot(`3Bet`~ PFR, holdem, main = "PFR segons 3Bet")
aggregate(holdem$PFR,by=list(holdem$`3Bet`),mean)
with(holdem, cor(`3Bet`, PFR))
```

```
library(ggplot2) # preparar un data.frame
jug<-holdem$jug
df<-data.frame(holdem$PFR,holdem$`3Bet`,jug)
head(df)
sp<-ggplot(df, aes(x=holdem$PFR, y=holdem$`3Bet`, color= jug, shape=jug)) +  geom_point() +
geom_smooth(method=lm, se=FALSE, fullrange=TRUE)
sp + scale_color_manual(values=c("#00CC33", "#FF0000"))

#library(MASS)
cor(x=holdem$PFR, y=holdem$`3Bet`) #coeficient de correlació lineal #0.7107421
cor.test (holdem$PFR,holdem$`3Bet` )

# PFR – VPIP
par(cex.main=1,font = 1, font.lab = 1, font.axis = 1, las = 1,  cex.lab=1, cex.axis= 0.7)
aggregate(holdem$PFR,by=list(holdem$`VP$IP`),mean)
with(holdem, cor(`VP$IP`, PFR))
plot(`VP$IP`~ PFR, holdem, main = "PFR segons el VP$IP")

library(ggplot2) # preparar un data.frame
jug<-holdem$jug
df<-data.frame(holdem$PFR,holdem$`VP$IP`,jug)
head(df)
sp<-ggplot(df, aes(x=holdem$PFR, y=holdem$`VP$IP`, color= jug, shape=jug)) +  geom_point() +
geom_smooth(method=lm, se=FALSE, fullrange=TRUE)
sp + scale_color_manual(values=c("#00CC33", "#FF0000"))

cor(x=holdem$PFR, y=holdem$`VP$IP`) #0.3042116
cor.test (holdem$PFR,holdem$`VP$IP`)



# 3Bet – Squeeze
par(cex.main=1,font = 1, font.lab = 1, font.axis = 1, las = 1,  cex.lab=1, cex.axis= 0.7)
aggregate(holdem$`3Bet`,by=list(holdem$Squeeze),mean)
with(holdem, cor(Squeeze, `3Bet`))
plot(Squeeze ~ `3Bet`, holdem, main = "3Bet segons l'Squeeze")

jug<-holdem$jug
df<-data.frame(holdem$`3Bet`,holdem$Squeeze,jug)
head(df)
sp<-ggplot(df, aes(x=holdem$`3Bet`, y=holdem$Squeeze, color= jug, shape=jug)) +  geom_point() +
geom_smooth(method=lm, se=FALSE, fullrange=TRUE)
sp + scale_color_manual(values=c("#00CC33", "#FF0000"))

cor(x=holdem$`3Bet`, y=holdem$Squeeze) #0.7013631
cor.test (holdem$`3Bet`,holdem$Squeeze)
```

```
# Postflop Agg% - W$WSF%
par(cex.main=1,font = 1, font.lab = 1, font.axis = 1, las = 1,  cex.lab=1, cex.axis= 0.7)
aggregate(holdem$`Postflop Agg%`,by=list(holdem$`W$WSF%`),mean)
with(holdem, cor(`W$WSF%`, `Postflop Agg%`))
plot(`W$WSF%`~ `Postflop Agg%`, holdem, main = "Postflop Agg% segons W$WSF%")
jug<-holdem$jug
df<-data.frame(holdem$`Postflop Agg%`,holdem$`W$WSF%`,jug)
head(df)
sp<-ggplot(df, aes(x=holdem$`Postflop Agg%`, y=holdem$`W$WSF%`, color= jug, shape=jug)) +
geom_point() + geom_smooth(method=lm, se=FALSE, fullrange=TRUE)
sp + scale_color_manual(values=c("#00CC33", "#FF0000"))

cor(x=holdem$`Postflop Agg%`, y=holdem$`W$WSF%`) #0.7013631
cor.test (holdem$`Postflop Agg%`,holdem$`W$WSF%`)

# PFR- Squeeze
par(cex.main=1,font = 1, font.lab = 1, font.axis = 1, las = 1,  cex.lab=1, cex.axis= 0.7)
aggregate(holdem$PFR,by=list(holdem$Squeeze),mean)
with(holdem, cor(Squeeze, PFR))
plot(Squeeze~ PFR, holdem, main = "PFR segons Squeeze")
jug<-holdem$jug
df<-data.frame(holdem$PFR,holdem$Squeeze,jug)
head(df)
sp<-ggplot(df, aes(x=holdem$PFR, y=holdem$Squeeze, color= jug, shape=jug)) +  geom_point() +
geom_smooth(method=lm, se=FALSE, fullrange=TRUE)
sp + scale_color_manual(values=c("#00CC33", "#FF0000"))

cor(x=holdem$PFR, y=holdem$Squeeze) #0.7013631
cor.test (holdem$PFR,holdem$Squeeze)

# PFR - Postflop Agg%
par(cex.main=1,font = 1, font.lab = 1, font.axis = 1, las = 1,  cex.lab=1, cex.axis= 0.7)
aggregate(holdem$PFR,by=list(holdem$`Postflop Agg%`),mean)
with(holdem, cor(`Postflop Agg%`, PFR))
plot(`Postflop Agg%`~ PFR, holdem, main = "PFR segons Postflop Agg%")
jug<-holdem$jug
df<-data.frame(holdem$PFR,holdem$`Postflop Agg%`,jug)
head(df)
sp<-ggplot(df, aes(x=holdem$PFR, y=holdem$`Postflop Agg%`, color= jug, shape=jug)) +
geom_point() + geom_smooth(method=lm, se=FALSE, fullrange=TRUE)
sp + scale_color_manual(values=c("#00CC33", "#FF0000"))

cor(x=holdem$PFR, y=holdem$`Postflop Agg%`) #0.7013631
cor.test (holdem$PFR,holdem$`Postflop Agg%`)
```

```
library(dplyr) # for data cleaning
library(ISLR) # for college dataset
library(cluster) # for gower similarity and pam
library(Rtsne) # for t-SNE plot
library(ggplot2) # for visualization

#############################################################################################
#########################    CLUSTERING VARIABLES NUMERIQUES   ##########################
#############################################################################################

##Utilitzarem distancia de Gower

numericas <- holdem[,c(3:12)] #10 variables numèriques en totalnum <-
as.data.frame(na.omit(numericas))
colnames(num)
m <- apply(num,2,mean)
s <- apply(num,2,sd)
#num <- scale(num,m,s)
rm (m, s)

library(StatMatch)

# matriu de distàncies entre observacions variables numèriques
dx <- as.dist(gower.dist(num))

#Calculem l'amplada de la silueta

sil_width <- c(NA)
for(i in 2:10){
  pam_fit <- pam(dx,diss = TRUE,k = i)
  sil_width[i] <- pam_fit$silinfo$avg.width
}

#Dibuixem el Plot de l'amplada de la silueta (sihouette width) com més alt millor

plot(1:10, sil_width,xlab = "Número de clústers",ylab = "Silhouette Width")
lines(1:10, sil_width)

#Utilitzem mètode calinski per veure quin número de clústers és òptim

library(vegan)
num.cascade<-cascadeKM(num,inf.gr = 1,sup.gr = 10,iter=100,criterion = "calinski")
plot(num.cascade,sortg=TRUE)

#Dendrograma  #is.na(dx)
res<- hclust(dx,method="ward.D")
plot(res,main = "Dendrograma/ Cluster",cex = 0.6, hang = -1)
```

```
library(dendextend)
plot(res, cex = 0.6) # plot tree
rect.hclust(res, k = 6, border = 2:5) # add rectangle

# --------------
# PAM algorisme
# --------------

set.seed(123)
pam.res <- pam(dx, diss = TRUE,6)
#pam.res
pam.res$medoids
pam.res$id.med
pam.res$clustering
pam.res$objective
pam.res$isolation
pam.res$clusinfo
pam.res$silinfo
pam.res$diss
pam.res$call

#Representación clustering PAM

library(cluster)
# Partim llavors, en 5 CLUSTERS
c2 <- cutree(res,k=6)
table(c2)
cdg <- aggregate(as.data.frame(num),list(c2),mean)


# -----
# MDS
# -----
library(MASS)
mds.pam<- cmdscale(dx, eig=TRUE)
plot(mds.pam$points[,1], mds.pam$points[,2], main="PAM", xlab="Axis 1", ylab="Axis 2",
col=as.factor(pam.res$cluster), pch=19)
 text(mds.pam$points[,1], mds.pam$points[,2], labels= pam.res$cluster, pos=1, cex=0.5, offset=0.15)

#interpreting the axes
Axis1<-round(cor(mds.pam$points[,1], num),3)
Axis2<-round(cor(mds.pam$points[,2], num),3)

#########################################################################
###### CLUSTERING VARIABLES NUMERIQUES EN PERCENTATGE #############
#########################################################################
```

```
percentatge<- holdem[,c(5:12)]
per <- as.data.frame(na.omit(percentatge))
summary(per)

## Utilitzarem distancia  de Bray Curtis
library(vegan)
bray<-vegdist(per, "bray")

#Calculem l'amplada de la silueta per veure quin número de clústers és òptim
sil_width <- c(NA)
for(i in 2:10){
  pam_fit <- pam(bray,diss = TRUE,k = i)
  sil_width[i] <- pam_fit$silinfo$avg.width
}
plot(1:10, sil_width,xlab = "Número de clústers",ylab = "Silhouette Width")
lines(1:10, sil_width)

#Utilitzem mètode calinski per veure quin número de clústers és òptim

num.cascade<-cascadeKM(per,inf.gr = 1,sup.gr = 10,iter=100,criterion = "calinski")
plot(num.cascade,sortg=TRUE)

#Dendrograma #is.na(bray)
h1 <- hclust(bray,method="ward.D")
h2<-hclust(bray,method="complete")
plot(h1,main = "Dendrograma/ Cluster",cex = 0.6, hang = -1)
plot(h2,main = "Dendrograma/ Cluster",cex = 0.6, hang = -1)

library(dendextend)
dend.a <- color_branches(h1, k = 5)
plot(dend.a,main = "Dendrograma/ Cluster",cex = 0.6)
dend.b <- color_branches(h2, k = 5)
plot(dend.b,main = "Dendrograma/ Cluster",cex = 0.6)

# ----------
# PAM algorisme
# ----------
set.seed(123)
pam.res <- pam(bray, diss = TRUE,5)
#pam.res
pam.res$medoids
pam.res$id.med
pam.res$clustering
pam.res$objective
pam.res$isolation
pam.res$clusinfo
pam.res$silinfo
```

```
pam.res$diss
pam.res$call

#Representación clustering PAM

c2 <- cutree(h1,k=5)
table(c2)
plot(h1, cex = 0.6) # plot tree
rect.hclust(h1, k = 5, border = 2:5) # add rectangle
cdg <- aggregate(as.data.frame(per),list(c2),mean)
# -----
# MDS
# -----
mds.pam<- cmdscale(bray, eig=TRUE)
plot(mds.pam$points[,1], mds.pam$points[,2], main="PAM", xlab="Axis 1", ylab="Axis 2",
 col=as.factor(pam.res$cluster), pch=19)
text(mds.pam$points[,1], mds.pam$points[,2], labels= pam.res$cluster, pos=1, cex=0.5, offset=0.15)
#interpreting the axes
Axis1<-round(cor(mds.pam$points[,1], per),3)
Axis2<-round(cor(mds.pam$points[,2], per),3)

################################################################
################## CLUSTERING D=D1 + D2  #######################
################################################################

#Definim la D1
discretes<- holdem[,c(3:4)]
dis<-as.data.frame(na.omit(discretes))
summary(dis)
d1 <- as.dist(gower.dist(dis))
#Definim la D2
percentatge<- holdem[,c(5:12)]
per <- as.data.frame(na.omit(percentatge))
summary(per)
d2<-vegdist(per, "bray")

#Ponderem cada matriu de distàncies
d1<-d1*(2/10)
d2<-d2*(8/10)

d<-d1+d2
d<-as.dist(d)

#Calculem l'amplada de la silueta
sil_width <- c(NA)
```

```
for(i in 2:10){
  pam_fit <- pam(d,diss = TRUE,k = i)
  sil_width[i] <- pam_fit$silinfo$avg.width
}

#Dibuixem el Plot de l'amplada de la silueta (sihouette width) com més alt millor
plot(1:10, sil_width,xlab = "Número de clústers",ylab = "Silhouette Width")
lines(1:10, sil_width)

#Utilitzem mètode calinski per veure quin número de clústers és òptim
num.cascade<-cascadeKM(num,inf.gr = 1,sup.gr = 10,iter=100,criterion = "calinski")
plot(num.cascade,sortg=TRUE)

#Dendrograma  #is.na(dx)
fin<- hclust(d,method="ward.D")
plot(fin,main = "Dendrograma/ Cluster",cex = 0.6, hang = -1)
c<- color_branches(fin, k = 5)
plot(c,main = "Dendrograma/ Cluster",cex = 0.6)
plot(fin, cex = 0.6) # plot tree
rect.hclust(fin, k = 5, border = 2:5) # add rectangle

# --------------
# PAM algorisme
# --------------
set.seed(123)
pam.res <- pam(d, diss = TRUE,5)
pam.res$medoids
pam.res$id.med
pam.res$clustering
pam.res$objective
pam.res$isolation
pam.res$clusinfo
pam.res$silinfo
pam.res$diss
pam.res$call

#Representación clustering PAM
c2 <- cutree(fin,k=5)
table(c2)
cdg <- aggregate(as.data.frame(num),list(c2),mean)

#observem els scatterplots
plot(holdem$`Postflop Agg%`~ holdem$`W$WSF%`,holdem,col=pam.res$cluster)

# -----
# MDS
# -----
```

```
mds.pam<- cmdscale(d, eig=TRUE)
plot(mds.pam$points[,1], mds.pam$points[,2], main="PAM", xlab="Axis 1", ylab="Axis 2",
 col=as.factor(pam.res$cluster), pch=19)
text(mds.pam$points[,1], mds.pam$points[,2], labels= pam.res$cluster, pos=1, cex=0.5, offset=0.15)
#interpreting the axes
Axis1<-round(cor(mds.pam$points[,1], num),3)
Axis2<-round(cor(mds.pam$points[,2], num),3)

#########################################################
################  RANDOM FOREST  ####################
#########################################################

########## VARIABLE RESPOSTA NET WON > 0?

#install.packages("randomForest")
library(randomForest)
require(caTools)
library(rpart)
library(rpart.plot)
library(caret)
library(e1071)

colnames(num)<-c("Hands","Net_Won","VP_IP","PFR","T3Bet","Postflop_Agg","W_WSF","WTSD",
"Won_at_SD","Squeeze")
num$Class<-as.factor(num$Net_Won>0)
str(num)
num<-num[,-2]

set.seed(123)
training<-sample(1:2122,1400,replace=F)
train<-num[training,]
test<-num[-training,]
str(train)
str(test)

#10 folds repeat 3 times
control <- trainControl(method='repeatedcv', number=10, repeats=3,search='grid')
tunegrid <- expand.grid(.mtry = (1:6))

#Metric compare model is Accuracy
metric <- "Accuracy" #mtry <- sqrt(ncol(num))   #tunegrid <- expand.grid(.mtry=mtry)

rf <- train(Class~., data=num, method='rf', metric='Accuracy', tuneGrid=tunegrid,
trControl=control)
print(rf)
summary(rf)
plot(rf)
```

```
rf.fit<-randomForest(Class ~ .,data=train,mtry=4,importance=TRUE,nodesize=30)
print(rf.fit)
plot(rf.fit)
varImpPlot(rf.fit)

pred = predict(rf.fit, newdata=test[-10])
head(pred)
cm = table(test[,10], pred)

#test the accuracy
a<-(383+148)/nrow(test)

getTree(rf.fit, k=1, labelVar=TRUE)

############################
####### CART ##############
############################
library(rpart)
fit <- rpart(Class ~ ., method="class", data=train)
printcp(fit) # display the results
plotcp(fit) # visualize cross-validation results
summary(fit)

# plot tree
plot(fit, uniform=TRUE,main="Classification Tree for Holdem")
text(fit, use.n=TRUE, all=TRUE, cex=.9)

library(rpart.plot)
rpart.plot(fit,uniform=TRUE,main=" Classification Tree for Holdem")

pfit<- prune(fit, cp=   fit$cptable[which.min(fit$cptable[,"xerror"]),"CP"])
summary(pfit)
# plot the pruned tree
plot(pfit, uniform=TRUE,main="Pruned Classification Tree for Holdem")
text(pfit, use.n=TRUE, all=TRUE, cex=.8)

library(rpart.plot)
rpart.plot(pfit,uniform=TRUE, main="Pruned Classification Tree for Holdem")
pred = predict(pfit, newdata=test[-10])
pred<-as.data.frame(pred)
pred<-cbind(pred,test$Class)
colnames(pred)<-c("Loss","Win","Class")
pred$hat<-ifelse(pred$Loss<pred$Win,1,0)
cm = table(test[,10], pred$hat)
accuracy<-sum(diag(cm))/sum(cm)

########## VARIABLE RESPOSTA ELS 5 CLÚSTERS DEL CLUSTERING 1
```

```
#10 folds repeat 3 times
control <- trainControl(method='repeatedcv',number=5,repeats=3,search='grid')
tunegrid <- expand.grid(.mtry = (1:6))

#Metric compare model is Accuracy
metric <- "Accuracy"  #mtry <- sqrt(ncol(num))  #tunegrid <- expand.grid(.mtry=mtry)

rf <- train(Class~., data=num, method='rf', metric='Accuracy', tuneGrid=tunegrid,
trControl=control)
print(rf)
plot(rf)

rf.fit<-randomForest(Class ~ .,data=train,mtry=1,importance=TRUE)
plot(rf.fit)
print(rf.fit)
varImpPlot(rf.fit)

pred = predict(rf.fit, newdata=test[-10])
cm = table(test[,10], pred)
accuracy<-sum(diag(cm))/sum(cm)

tr<-getTree(rf.fit, k=1, labelVar=TRUE)

fit <- rpart(Class ~ ., method="class", data=train)
printcp(fit) # display the results
plotcp(fit) # visualize cross-validation results
summary(fit)

# plot tree
plot(fit, uniform=TRUE,main="Classification Tree for Holdem")
text(fit, use.n=TRUE, all=TRUE, cex=.8)
rpart.plot(fit,uniform=TRUE,main=" Classification Tree for Holdem",extra=106)

pfit<- prune(fit, cp=   fit$cptable[which.min(fit$cptable[,"xerror"]),"CP"])
# plot the pruned tree
plot(pfit, uniform=TRUE,main="Pruned Classification Tree for Holdem")
text(pfit, use.n=TRUE, all=TRUE, cex=.8)
rpart.plot(pfit,uniform=TRUE,main=" Pruned Classification Tree for Holdem",extra=1)

pred = predict(pfit, newdata=test[,-10],type="class")
pred<-as.data.frame(pred)
cm = table(test[,10], pred$pred)
accuracy<-sum(diag(cm))/sum(cm)
```