TREBALL DE FI DE GRAU

Grau en Estadística

Títol: Clústers amb variables mixtes per a la caracterització de clients

Autora: Marta Carbonell Cabutí

Director/a: Ignasi Puig De Dou i Lourdes Rodero De Lamo

Departament: Estadística i Investigació Operativa

Convocatòria: Juliol 2020

UNIVERSITAT DE BARCELONA

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

RESUM

L'anàlisi de conglomerats és un mètode multivariant que té com a objectiu principal identificar grups d'objectes amb característiques similars dins d'una base de dades numèriques. Actualment però, aquesta branca de l'estadística està desenvolupant mètodes que permetin l'anàlisi de bases de dades mixtes, per tal de poder utilitzar tant les variables descriptives numèriques com les categòriques dels diversos objectes. Aquests criteris d'agrupació es poden classificar en dos grans grups: els mètodes jeràrquics i els no jeràrquics.

En el següent treball es realitza una clusterització de les dades dels clients d'un majorista de ferreteria industrial a fi de poder-los agrupar en varis grups homogenis, mitjançant dos mètodes d'agrupació: el mètode de Ward i el *Partition Around Medoids.* A fi de poder crear aquests grups és necessari calcular un coeficient de similitud per tal de conèixer les distàncies entre els individus. Així doncs, s'utilitzarà el coeficient de Gower, ja que permet tractar amb dades numèriques i categòriques a la vegada. No obstant, també es realitzarà un anàlisi de sensibilitat d'aquesta mesura per tal de comprovar la seva robustesa.

Clúster: grup homogeni d'objectes o observacions.

Dissimilitud: qualitat de ser diferent o desigual a la resta.

Dendrograma: gràfic en forma d'arbre que agrupa les dades en funció de la seva semblança.

Medoid: objecte representatiu dins d'un grup d'observacions, a causa de tenir la menor dissimilaritat mitjana respecte la resta d'objectes del conjunt.

SUMMARY

Traditionally, the cluster analysis has only been used in numerical data bases with the main objective being the identification of object groups with similar characteristics within those databases. This is an on-growing branch of statistics science that tries to incorporate methods that allow us to perform this mix database object categorization.

In this project we will be carrying out a cluster analysis of an industrial hardware wholesaler's client information. By doing so, we can recognize those that follow specific patterns of behaviour, which will allow us to perform a more accurate follow-up and segment the price cut campaigns according to their own interests.

For us to accomplish this objective, we have opted for the use of both a hierarchical aggregation method and a non-hierarchical one: the Ward and the Partition Around Medoids methods, respectively. However, for us to be able to create these groups we need to calculate a similarity coefficient for us to know the distances between the various individuals. In this study, we have opted for the use of the Grower coefficient, seeing as it allows us to handle numerical and categorical data at the same time.

Hereunder is the detailed explanation of the client characterisation process, starting with the creation of the database and following up with the description of the various groups used. A Grower distance sensibility analysis is also included for the validation of the solidness of this measure.

Cluster: a group of similar objects or observations.

Dissimilarity: the quality or state of being dissimilar.

Dendrogram: a cluster tree diagram where the distance of split or merge is recorded.

Medoid: a mathematically representative object in a set of objects; with the smallest average dissimilarity to all other objects in the set.

62H30 Classification and discrimination; cluster analysis

ÍNDEX

I. INTRODUCCIÓ

Una distribuïdora catalana de peces de recanvi està interessada en classificar els seus clients en funció de les seves compres, a fi d'identificar els que tenen un comportament comú per poder-ne fer un seguiment més acurat i segmentar les campanyes promocionals d'acord als seus interessos. La distribuïdora ha facilitat una base de dades mixta, ordenada de forma cronològica, amb les línies de compra de tots els seus clients des del 2016 fins el 2019.

Per tal d'aconseguir la caracterització dels clients de la companyia, s'ha fet ús de la distància de Gower i s'han decidit utilitzar dos mètodes d'agrupament de variables mixtes, el Mètode de Ward i el PAM (*Partition Around Mediods*).

En el següent informe es troba l'explicació detallada del procés de caracterització dels clients, començant per la creació de la base de dades i acabant amb la descripció dels diversos grups realitzats.

II. METODOLOGIA

En el següent estudi s'ha dut a terme un treball experimental. L'objectiu d'aquest ha estat la creació de conglomerats a partir d'una base de dades mixtes (dades numèriques i categòriques). A fi d'abordar aquesta qüestió, s'ha fet ús de diverses tècniques d'anàlisi multivariant de dades.

A partir d'una matriu de dades s'han calcular les distàncies i similituds entre les observacions mitjançant el coeficient de similitud de Gower. Posteriorment, s'han utilitzat dues tècniques de classificació diferents per tal de realitzar els clústers i aconseguir l'objectiu principal d'aquest projecte. El mètode de Ward, pertanyent als mètodes jeràrquics aglomeratius, i el *Partition Around Mediods,* pertanyent als mètodes no jeràrquics, han estat les dues tècniques utilitzades.

La base de dades ha estat proporcionada per un client anònim de Datancia, una empresa innovadora en tecnologia i mètodes d'analítica avançada, especialitzada en anàlisi de dades.

En darrer terme, cal esmentar que tot el procés s'ha realitzat mitjançant les eines estadístiques que ens proporciona el software R.

III. BASE DE DADES

Per tal de realitzar l'anàlisi en qüestió, la distribuïdora va proporcionar dues bases de dades on s'hi trobaven: el detall de les línies de comanda de les compres (*Tiquets*) i la descripció de cadascun d'ells (*Clients*). A partir d'aquestes, s'han creat un total de quatre resums anuals (2016, 2017, 2018 i 2019) on s'hi recopilen les principals variables numèriques i categòriques que han permès realitzar l'agrupació dels clients.

1. Tiquets

Taula 3.1. Estructura de la base de dades "Tiquets"

	customerId	tratamiento	mercado	sectorId	delId	contId	albId	fAlb	sku	familiaId	qty	precio
1	47	G	OEM	10	11	5491239	171982	2016-05-19	58600	333	55	7.9825
2	47	G	OEM	10	11	5510849	180879	2016-06-10	137283	3	5	6.3119
3	47	G	OEM	10	11	5510849	180879	2016-06-10	314765	502	10	6.2279
4	47	G	OEM	10	11	5510849	180879	2016-06-10	335239	5	5	1.1020
5	47	G	OEM	10	11	5510849	180879	2016-06-10	342958	5	5	0.6572
6	47	G	OEM	10	11	5510849	180879	2016-06-10	441949	2	5	3.0423
7	47	G	OEM	10	11	5510849	180879	2016-06-10	441949	2	5	1.2871
8	47	G	OEM	10	11	5510849	180879	2016-06-10	552406	2	5	1.0066
9	47	G	OEM	10	11	5510849	180879	2016-06-10	552406	2	5	1.1815
10	47	G	OEM	10	11	5510849	180879	2016-06-10	570089	5	5	0.0924

dto	importe	up	descUP	digital	tAlb	fCont	envioId	botiga	contOrigenId	descContOrigen	descEnvio
0	439.0375	8	HERRAM+NEUMATIC	N	CA	2016-05-19	3	S	TDA	Tienda	AGENCIA
0	31.5595	1	ESTANQUEIDAD	N	CA	2016-06-08	3	S	TDA	Tienda	AGENCIA
0	62.2790	1	ESTANQUEIDAD	N	CA	2016-06-08	3	S	TDA	Tienda	AGENCIA
0	5.5100	1	ESTANQUEIDAD	N	CA	2016-06-08	3	S	TDA	Tienda	AGENCIA
0	3.2860	1	ESTANQUEIDAD	N	CA	2016-06-08	3	S	TDA	Tienda	AGENCIA
0	15.2115	1	ESTANQUEIDAD	N	CA	2016-06-08	3	S	TDA	Tienda	AGENCIA
0	6.4355	1	ESTANQUEIDAD	N	CA	2016-06-08	3	S	TDA	Tienda	AGENCIA
0	5.0330	1	ESTANQUEIDAD	N	CA	2016-06-08	3	S	TDA	Tienda	AGENCIA
0	5.9075	1	ESTANQUEIDAD	N	CA	2016-06-08	3	S	TDA	Tienda	AGENCIA
0	0.4620	1	ESTANQUEIDAD	N	CA	2016-06-08	3	S	TDA	Tienda	AGENCIA

La taula 3.1 mostra l'estructura original de la base de dades "Tiquets". Cada línia conté la informació d'una línia de comanda, on s'hi especifica l'identificador del client i de l'albarà per tal de poder distingir les diverses compres realitzades. Les variables d'aquesta primera taula són:

- customerId: identificador únic del client.
- tratamiento: variable binària amb valor "G" o "S", en funció de si el client és considerat "Gold" o "Silver".
- mercado: variable categòrica que fa referència al mercat en el qual treballa el client. Té un total de 4 categories:
    - MRO: Maintance, Repair and Operations.
    - OEM: Original Equipment Manufacturing.
    - OTR: Others
    - SI: Industrial Systems.

- sectorId: variable categòrica amb 29 nivells possibles que fan referència al sector d'activitat en el qual treballa el client.
- delId: variable categòrica amb 11 nivells possibles que indica a quina delegació pertany el client.
- contId: identificador de compra. Cada vegada que el client contacta amb l'empresa per realitzar una compra es crea un contenidor nou. Aquest contenidor pot tenir un nombre qualsevol de transaccions.
- albId: Un contenidor, no té perquè enviar-se de forma conjunta sinó que es pot entregar per parts. Cadascuna d'aquestes parts és un albarà i és amb l'identificador que es treballarà al llarg de l'estudi.
- fAlb: data en la qual es va realitzar la compra.
- sku: Tercer i últim nivell d'identificació del producte. Cada peça té el seu propi número d'identificació.
- familiaId: Segon nivell d'identificació del producte. Dins de cada família hi ha una quantitat significativa d'skus. En total hi ha 122 famílies diferents.
- qty: variable numèrica que indica el nombre d'unitats de producte que ha comprat.
- precio: variable numèrica que indica el preu unitari del producte comprat:
- dto: variable numèrica entre zero i u, que determina el tant per cent de descompte que s'ha aplicat en la compra.
- importe: import total de la línia de transacció. És el producte entre la quantitat comprada i el preu de la peça menys el descompte aplicat.
- up: primer nivell d'identificació del producte. Dins de cada grup up, hi ha diferents famílies. Existeixen un total de 12 grups: estanquitat, segells mecànics, rodaments, etc.
- descUP: Descripció de la variable "up".
- digital: variable binària que pren per valor "S" o "N" en funció del mètode de registre que ha emprat el client. El nivell "S" és per aquells que es van registrar per la plataforma digital. El nivell "N" per a la resta.
- tAlb: variable binària que pren per valor "AB" o "CA" en funció de si l'albarà tracta d'un abonament o d'un càrrec.
- fCont: data en la qual es va crear el contenidor.
- envioId: variable categòrica amb 10 nivells diferents en funció del mètode d'enviament utilitzat per la compra.
- botiga: variable binària amb nivells "S" i "N" que mostra si la compra s'ha realitzat a la botiga del distribuïdor o bé per altres mitjans (telèfon, correu electrònic, web, etc.).
- contOrigenId: variable categòrica amb 13 nivells diferents que defineixen a través de quin mitjà s'ha creat el contenidor.
- descContOrigen: descripció de la variable "contOrigenId".
- descEnvio: descripció de la variable "envioId".

2. Clients

Taula 3.2. Estructura de la base de dades “Clients”

	customerId	delId	sectorId	codPos	codPais	loc	prov	alta	NIF	vendedorId
1	13	3	15	20829	34	ITZIAR-DEBA	GUIPUZCOA	2013-01-16	Empresa	502
2	14	17	18	33203	34	GIJON	ASTURIAS	2013-01-23	Empresa	800
3	23	2	28	45311	34	DOS BARRIOS	TOLEDO	2013-01-31	Empresa	808
4	30	12	28	48980	34	SANTURCE	BIZKAIA	2013-02-04	Particular	800
5	32	1	28	08440	34	CARDEDEU	BARCELONA	2013-02-06	Particular	808
6	36	11	18	41410	34	CARMONA	SEVILLA	2013-02-07	Particular	808
7	37	1	27	08960	34	SANT JUST DESVERN	BARCELONA	2013-02-07	Empresa	808
8	41	3	28	20280	34	HONDARRIBI	GUIPUZCOA	2013-02-12	Empresa	808
9	44	1	28	17460	34	CELRA	GIRONA	2013-02-12	Particular	808
10	45	1	28	08295	34	SANT VICEN DE CASTELLET	BARCELONA	2013-02-12	Particular	808

La taula 3.2 mostra l’estructura original de la base de dades “Clients”. Cada línia conté la informació bàsica d’un client. Les tres primeres variables s’han definit en l’apartat anterior ja que també formaven part de la base de dades “Tiquets”. Les resta de variables d’aquesta segona taula són:

- codPos: codi postal del municipi en el qual es localitza el client.
- codPais: codi del país en el qual es localitza el client.
- loc: població en la qual es localitza el client.
- prov: província en la qual es localitza el client.
- alta: data en què l'empresa va fer el registre del client.
- NIF: variable binària que pren per valor “Empresa” o “Particular” en funció del tipus de negoci que té el client.
- vendedorId: identificador del treballador que va registrar el client a la base de dades.

3. Resums anuals

Un cop definides les dues bases de dades proporcionades per la distribuïdora, és necessari explicar el procés de creació dels resums anuals. En primer lloc, es va dividir la base de dades “Tiquets” en quatre conjunts de dades nous en funció a la variable que definia en quin moment s’havia realitzat la compra. Cal recordar que l’objectiu principal d’aquesta tercera base de dades era obtenir una síntesi de les diverses compres realitzades per cada client al llarg de l’any en una sola observació, per tal de poder aplicar els diversos mètodes d’agrupament i classificar els clients segons el seu comportament anual de compra. Així doncs, hi ha variables que són idèntiques a les originals, però n’hi ha d’altres que són estadístics de les diverses compres realitzades per cada client en la distribuïdora.

La distribuïdora parteix d'una classificació binària inicial dels clients en funció de si tenen un tracte preferent o no. Els clients "Gold" són un grup de consumidors reduït que estan fidelitzats amb l'empresa i gasten una quantitat important de diners al llarg de l'any; motiu pel qual reben un tracte preferencial. Com que la distribuïdora creu conèixer les seves tendències a l'hora de realitzar les compres, tant sols vol que realitzin els resums anuals pels clients "Silver".

Així doncs, en tots els resums anuals dels clients "Silver" s'hi troben les següents variables:

- clientID: identificador únic del client.
- Preumig: variable numèrica que indica el preu mitjà de les peces que compra al llarg de l'any.
- Preuvar: variable numèrica que mostra la variància en el preu de les diverses peces comprades al llarg de l'any. Permetrà identificar si sempre compra peces amb un cost similar o no.
- preutotal_compra: variable numèrica que indica l'import total gastat al llarg de l'any en la distribuïdora.
- preuvar_compra: variable numèrica que mostra la variància entre els imports gastats en les diverses compres realitzades al llarg de l'any. Permetrà observar si totes les compres que realitza són del mateix volum econòmic o no.
- total_compra: variable numèrica que mostra el nombre de compres realitzades al llarg de l'any.
- liniesmig_compra: variable numèrica que indica línies mitjanes de facturació en les diverses compres al llarg de l'any.
- liniesvar_compra: variable numèrica que indica variància entre les línies de facturació de les compres realitzades al llarg de l'any.
- mercat: variable categòrica que expressa a quin mercat està assignat el client.
- botiga: proporció de diners gastats al llarg de l'any en la botiga física.
- contenidor: variable categòrica amb 7 nivells que indica per quin mitjà s'ha creat el contenidor. Els nivells són:
    - DOC: el contenidor s'ha creat mitjançant un document.
    - EM: el contenidor s'ha creat mitjançant el correu electrònic.
    - EXT: el contenidor s'ha creat a través d'internet o extranet.
    - POR: el contenidor s'ha creat a través del portal.
    - TDA: el contenidor s'ha creat a la mateixa botiga.
    - TF: el contenidor s'ha creat mitjançant una trucada telefònica.
    - ALT: el contenidor s'ha creat mitjançant el fax, els sistema de reposició d'estoc o a través del propi personal.
- enviament: proporció d'enviaments al client que s'han realitzat a través dels mitjans de la distribuïdora, per contraposició al transport realitzat pels mitjans del propi client.

- gf: variable categòrica que mostra l'ID del grup família en el qual el client ha gastat la majoria dels diners.
- IVQ: variable numèrica que mostra l'Índex de Variació Qualitativa entre els diversos grups-familia en els quals el client ha comprat:

$$IVQ = \frac{1 - \sum_{i=1}^{k} p_i^2}{(k-1)/k} \qquad (1)$$

On p_i és la proporció de diners gastats en cada grup-familia i, i k és el nombre total de grups-familia en els quals el client ha gastat diners al llarg de l'any.
Expressa el grau en que els diners que s'ha gastat el client estan dispersos en les diferents categories de la variable, agafant el seu màxim (IVQ=1) en el cas de que les freqüències relatives siguin iguals per totes les categories de la variable. Així doncs, si un client ha gastat tots els diners en el mateix grup-familia, obtindrà un índex de variació igual a zero. En canvi, si un client ha comprat en 4 famílies i ha gastat 850€ en una i 50€ en cadascuna de la resta, el seu IVQ seria de 0,36.

- RV: variable numèrica que representa la Raó de la Variació dels diners gastats entre els diversos grups-familia en els quals el client ha comprat:

$$RV = 1 - \frac{q_{max}}{q} \qquad (2)$$

On q és la quantitat total de diners gastats al llarg de l'any i q_{max} és el màxim de diners gastats en un dels grups-familia. Així doncs, aquest estadístic indica la proporció de diners que s'ha gastat el client en un grup-familia diferent en el qual s'ha gastat més diners. Per tant si l'estadístic és igual a zero ens indicarà que no hi ha dispersió entre els grups-familia en els quals ha gastat diners, mentre que si és proper a 1 indicarà una gran dispersió. Per exemple, si un client ha gastat el 20% de l'import total en el seu grup-familia modal, la seva Raó de Variació serà de 0,8.
- gf1: variable dicotòmica que pren valor 1 si el client ha comprat en el grup família Estanquitat al llarg del l'any en qüestió, o 0 en cas contrari.
- gf2: variable dicotòmica que pren valor 1 si el client ha comprat en el grup família Segells mecànics al llarg del l'any en qüestió, o 0 en cas contrari.
- gf3: variable dicotòmica que pren valor 1 si el client ha comprat en el grup família Rodaments i Coixinets al llarg del l'any en qüestió, o 0 en cas contrari.
- gf4: variable dicotòmica que pren valor 1 si el client ha comprat en el grup família Transmissors de potència al llarg del l'any en qüestió, o 0 en cas contrari.
- gf5: variable dicotòmica que pren valor 1 si el client ha comprat en el grup família Control de moviment al llarg del l'any en qüestió, o 0 en cas contrari.
- gf6: variable dicotòmica que pren valor 1 si el client ha comprat en el grup família Fixacions al llarg del l'any en qüestió, o 0 en cas contrari.

- gf7: variable dicotòmica que pren valor 1 si el client ha comprat en el grup família Antivibració al llarg del l'any en qüestió, o 0 en cas contrari.
- gf801: variable dicotòmica que pren valor 1 si el client ha comprat en el grup família Eines al llarg del l'any en qüestió, o 0 en cas contrari.
- gf802: variable dicotòmica que pren valor 1 si el client ha comprat en el grup família Pneumàtica al llarg del l'any en qüestió, o 0 en cas contrari.
- gf9: variable dicotòmica que pren valor 1 si el client ha comprat en el grup família Productes diversos al llarg del l'any en qüestió, o 0 en cas contrari.
- gf1001: variable dicotòmica que pren valor 1 si el client ha comprat en el grup família Adhesius al llarg del l'any en qüestió, o 0 en cas contrari.
- gf1002: variable dicotòmica que pren valor 1 si el client ha comprat en el grup família Lubricació al llarg del l'any en qüestió, o 0 en cas contrari.
- gf1003: variable dicotòmica que pren valor 1 si el client ha comprat en el grup família Prevenció al llarg del l'any en qüestió, o 0 en cas contrari.
- gf11: variable dicotòmica que pren valor 1 si el client ha comprat en el grup família TSP al llarg del l'any en qüestió, o 0 en cas contrari.
- gf12: variable dicotòmica que pren valor 1 si el client ha comprat en el grup família Serveis al llarg del l'any en qüestió, o 0 en cas contrari.
- delId: variable categòrica que fa referència a la delegació a la qual pertany el client.
- sectorId: variable categòrica que indica el sector d'activitat en el qual treballa el client.
- prov: variable categòrica que ens indica a quina província pertany el client.
- NIF: variable binària que mostra si el client és una empresa o un particular.

IV. ANÀLISI DE CONGLOMERATS

L'Anàlisi de Clústers, o de Conglomerats, és una tècnica estadística multivariant que té com a objectiu principal l'agrupació d'elements en grups homogenis. Tanmateix, cada clúster establert ha de diferenciar-se al màxim de la resta.

Aquesta tècnica de classificació automàtica de dades intenta, a partir d'una taula de casos-variables, situar els casos (individus) en grups homogenis no coneguts anteriorment, però suggerits per la pròpia essència de les dades. Així doncs, els individus que puguin ser considerats similars aniran a un mateix clúster, mentre que els individus diferents es localitzaran en clústers diferents.

Abans de realitzar una anàlisi d'aquest estil s'han de prendre un seguit de decisions:

- Selecció de les variables més rellevants per identificar el grup.
- Elecció de la mesura de proximitat entre els individus.
- Selecció del criteri per agrupar els individus en conglomerats.
- Validació de la coherència dels conglomerats creats.

La selecció de variables és decisiva a l'hora d'identificar idòniament els grups; és per aquest motiu que en aquest projecte tant aquesta decisió com les tres restants s'han pres de forma conjunta entre l'autora del treball i els directors d'aquest.

Consegüentment, en aquest capítol s'hi podrà trobar una explicació detallada de la mesura de proximitat entre els individus escollida, així com els criteris d'agrupació triats i els resultats obtinguts gràcies a aquests.

1. Selecció de les variables

Com bé s'ha dit anteriorment, la primera decisió que s'ha de prendre a l'hora de realitzar un anàlisi de conglomerats és la selecció de les variables que permetran crear els grups homogenis d'individus màximament diferenciats de la resta. Aquesta tria s'ha de realitzar envers les variables creades en els resums anuals i tenint en compte quina influència poden tenir a l'hora de crear els conglomerats.

Al llarg de l'estudi s'han realitzat vàries simulacions dels clústers amb diverses variables dels resums anuals, a fi de trobar quines d'aquestes havien de formar part de l'anàlisi. Gràcies a aquesta exploració es van poder observar mancances en el pes que se li donava a les variables quantitatives en la mesura de proximitat entre els individus triada per l'anàlisi. No obstant, aquesta carència serà analitzada en l'apartat "Anàlisi de sensibilitat de la distància de Gower", ja que es considera necessari dedicar-li una recerca particular.

Així doncs, després d'aquest anàlisi fet al llarg de la investigació, les dades seleccionades per a realitzar els clústers són: *preumig, preuvar, preutotal_compra, preuvar_compra, total_compra, liniesmig_compra, liniesvar_compra, mercat, botiga, enviament, gf1, gf2, gf3, gf4, gf5, gf6, gf7, gf801, gf802, gf9, gf1001, gf1002, gf1003, gf11, gf12, IVQ, RV, gf, delId, sectorID* i *NIF*. Resumint, s'utilitzaran totes les variables dels resums anuals, exceptuant l'identificador del client, el contenidor i la província a la qual pertany. Es poden trobar les descripcions d'aquestes variables a la secció 3 de l'apartat III.

2. Coeficient de similitud de Gower

Abans de realitzar l'agrupament dels individus, cal triar la mesura de proximitat entre ells. Com que la base de dades amb la qual s'està treballant hi ha variables numèriques però també categòriques, és necessari utilitzar un coeficient de similitud per a variables mixtes.

El coeficient de similitud de Gower permet el tractament simultani de dades qualitatives i quantitatives en una base de dades. Amb l'aplicació d'aquest coeficient s'aconsegueix trobar la similitud entre individus als quals se'ls hi ha mesurat una sèrie de característiques comunes. Una semblança alta, és a dir propera a l'1, indica una gran homogeneïtat entre els individus; mentre que una semblança baixa, és a dir propera al 0, indica que els individus són diferents.

Així doncs, es defineix la distància de Gower com $d_{ij}^2 = 1 - s_{ij}$. En el numerador del coeficient de similitud de Gower (s_{ij}), es comptabilitzen les coincidències de les variables categòriques i binàries dels diversos individus, i la similitud entre les variables numèriques d'aquests subjectes. En el denominador s'hi troben el nombre de variables totals que s'utilitzen.

En l'estudi realitzat per Guerrero, S, *et al.* (2017) es defineix el coeficient com

$$s_{ij} = \frac{\sum_{h=1}^{p_1}\left(1 - \frac{|x_{ih} - x_{jh}|}{G_h}\right) + a + \alpha}{p_1 + (p_2 - d) + p_3} \qquad (3)$$

on:

p_1 és el nombre de variables quantitatives contínues,

p_2 és el nombre de variables binàries,

p_3 és el nombre de variables qualitatives (no binàries),

x_{ih} és el valor que pren la variable quantitativa *h* en l'individu *i*,

x_{jh} és el valor que pren la variable quantitativa *h* en l'individu *j*,

a és el nombre de coincidències (1,1) en les variables binàries,

d és el nombre de coincidències (0,0) en les variables binàries,

α és el nombre de coincidències en les variables qualitatives (no binàries) i

G_h és el rang (o recorregut) de l'h-èssima variable quantitativa.

Mitjançant aquesta transformació de les variables mixtes amb l'ús de la similitud s_{ij} i la definició de la distància de Gower com d_{ij}^2 es poden agrupar els individus de forma que els conglomerats resultants estiguin integrats per unitats homogènies i els conglomerats siguin molt heterogenis entre ells.

Finalment, cal dir que s'ha triat aquest coeficient ja que és apropiat per calcular similituds quan es té una barreja de dades quantitatives, qualitatives i binàries. No obstant, un dels avantatges que també té aquesta mesura és que ens permet treballar amb bases de dades en les quals hi hagi observacions mancants d'algunes variables, sense recórrer a l'eliminació de tot el vector que representa la unitat mostral ni utilitzar cap mètode d'imputació.

3. Criteris d'agrupació

El penúltim pas per a realitzar l'anàlisi de conglomerats, és decidir quin criteri d'agrupació s'utilitzarà. Es poden classificar les diverses tècniques possibles en dos grans grups: els mètodes jeràrquics i els mètodes no jeràrquics. Els mètodes jeràrquics tenen com a objectiu l'agrupació de clústers per formar-ne de nous, de manera que si successivament es va realitzant aquest procés d'aglomeració, es minimitza alguna distància o bé es maximitza alguna mesura de similitud. D'altra banda, els mètodes no jeràrquics estan dissenyats per la classificació d'individus en k grups. El procediment es basa en seleccionar una partició dels individus en k grups i intercanviar els membres dels clústers per obtenir la millor partició.

En aquest treball es farà ús de dos mètodes d'agrupació: el mètode de Ward i el k-mediods. El primer pertany als mètodes jeràrquics i el segon als mètodes no jeràrquics.

3.1 Mètode de Ward

El mètode de Ward és un criteri d'agrupació que forma part dels mètodes jeràrquics de creació de clústers. En els mètodes jeràrquics, es poden distingir dues tècniques:

- Els mètodes aglomeratius, també coneguts com a ascendents. En aquest cas, es parteix de tants grups com individus hi ha en l'estudi i es van fent grups de forma ascendent, fins a tenir tots els casos en el mateix conglomerat.

- Els mètodes dissociatius, també coneguts com a descendents. En aquest cas es realitza el procés invers a l'anterior mètode. Es comença amb un sol grup que conté tots els casos i a través de successives divisions es formen grups cada cop més petits.

Tots dos mètodes permeten construir un arbre de classificació o dendrograma, en base al qual es podrà decidir en quants conglomerats s'acaben agrupant els individus del conjunt de dades inicial.

El mètode de Ward, també conegut com el mètode la mínima variància, és un mètode aglomeratiu que uneix els individus per tal de minimitzar la variància dins de cada grup. Per a això calcula en primer lloc, la mediana de totes les variables en cada conglomerat. A continuació, calcula la distància entre cada individu i la mediana del conglomerat, sumant després les distàncies entre tots els individus. Finalment, s'agrupen els conglomerats que generen menys augments en la suma de les distàncies dins de cada conglomerat. Aquest procediment permet obtenir grups homogenis i amb tamanys similars.

Així doncs, aquest és un procediment jeràrquic en el qual en cada etapa s'uneixen els dos clústers amb els quals es produeixi el menor increment en el valor total de la suma dels quadrats de les diferències dins de cada clúster. Consegüentment, la notació serà:

- x_{ij}^k : valor de la j-èssima variable sobre l'i-èssim individu del k-èssim conglomerat, suposant que aquest clúster contè n_k individus.
- m_k : centroide del clúster k, amb components m_j^k.
- E_k : suma dels quadrats dels errors del clúster k, és a dir, la distància euclidiana al quadrat entre cada individu del clúster k al seu centroide:

$$E_k = \sum_{i=1}^{n_k}\sum_{j=1}^{n}\left(x_{ij}^k - m_j^k\right)^2 = \sum_{i=1}^{n_k}\sum_{j=1}^{n}\left(x_{ij}^k\right)^2 - n_k\sum_{j=1}^{n}\left(m_j^k\right)^2 \qquad (4)$$

- E : suma dels quadrats dels errors per tots els clústers. Suposant que hi ha h clústers:

$$E = \sum_{k-1}^{h} E_k \qquad (5)$$

El procés comença amb m clústers, cada un dels quals està compost per un sol individu, per tant cada individu coincideix amb el centre del conglomerat i per tant en aquesta primera fase es tindrà $E_k = 0$ per cada clúster i conseqüentment $E = 0$. L'objectiu d'aquest mètode és trobar en cada etapa aquells dos clústers la unió dels quals proporcioni el menor increment en la suma total dels errors, E.

3.2 Partition Around Medoids

Dins dels mètodes no jeràrquics de creació de clústers, un dels algoritmes més populars de partició per bases de dades numèriques, és el k-means. No obstant, en aquest estudi és impossible fer-ne ús ja que la base de dades és mixta i per tant, la mitjana com a centre del clúster és inviable. Quan aquest problema sorgeix, el més comú és aplicar el *Partition Around Medoids* (PAM), que reemplaça els centroides pels *medoids.*

El terme *medoid* es refereix a un objecte dins d'un clúster per al qual la dissimilitud entre ell i tots els altres membres és mínima. Per tant, es pot definir el *medoid* com el punt més cèntric del clúster.

El k-medoids és una alternativa més robusta a l'algoritme k-means. Això vol dir que l'algoritme és menys sensible al soroll i als outliers ja que utilitza els *medoids* com a centres de clúster en lloc de les mitjanes.

Així doncs, l'algoritme PAM s'executa mitjançant tres fases: la selecció dels *medoids*, la permutació dels *medoids* i l'assignació final dels objectes als seus respectius *medoids*.

En la primera fase, es calcula la distància entre cada parell d'objectes mitjançant la mesura de dissimilaritat triada i seguidament es crea el vector v_j per cada individu *j* on:

$$v_j = \sum_{i=1}^{n} \frac{d_{ij}}{\sum_{l=1}^{n} d_{il}}, \quad j = 1, \dots, n \qquad (6)$$

Un cop generat el vector es trien de forma aleatòria els *k* objectes com a *medoids* inicials. Gràcies a aquesta selecció s'obtenen els clústers inicials mitjançant l'assignació de cada individu al *medoid* més proper. Per acabar amb aquesta fase, es calcula la suma de les distàncies de tots els individus als seus respectius *medoids*.

En la fase de permuta s'intenta millorar la qualitat dels clústers, trobant un nou *medoid* que minimitzi la distància total cap als altres individus, per cada un dels conglomerats creats. Així doncs, s'actualitzen els *medoids* a través del reemplaçament. Aquest procediment es repeteix fins que ja no es pot disminuir la funció objectiu. La finalitat d'aquesta fase és trobar els *k* objectes representatius que minimitzin la suma de les diferències de les observacions al seu objecte representatiu més proper.

Finalment, s'assigna cada objecte al seu *medoid* més proper a fi d'obtenir els clústers finals.

4. Validació dels clústers

L'últim pas abans d'extreure els resultats d'un anàlisi de conglomerats és comprovar que el nombre de grups d'individus que s'han creat és l'òptim.

El mètode de *Silhouette* serà utilitzat com a recurs per interpretar i validar la coherència dins dels clústers de dades. Aquesta mesura quantifica com de similar és un objecte al seu propi clúster en comparació als altres. El rang de valors que pot prendre aquesta mètrica va des del –1 fins al 1, on un valor elevat indica una bona agrupació mentre que un valor negatiu manifesta un mal aparellament.

Un dels avantatges d'aquesta tècnica és que proporciona una representació gràfica fàcilment interpretable. Si la majoria d'objectes tenen un valor elevat, aleshores la creació dels conglomerats es considera que s'ha realitzar adequadament. En canvi, si molts individus obtenen un valor baix o negatiu, el nombre de conglomerats creats es considerà excessiu o bé insuficient.

La mesura es defineix com:

$$s(i) = \begin{cases} 1 - a(i)/b(i) & si\ a(i) < b(i) \\ 0 & si\ a(i) = b(i) \\ b(i)/a(i) - 1 & si\ a(i) > b(i) \end{cases} \qquad (7)$$

on $a(i)$ és la distància mitjana entre i i tots els altres individus dins del mateix clúster, i $b(i)$ és la distància mitjana entre i i tots els punts de qualsevol altre clúster del qual no n'és membre. Així doncs, com bé s'ha indicat anteriorment i observant la definició de *Silhouette* queda clar que $-1 \leq -s(i) \leq 1$.

Per tant, la distància mitjana entre un individu i els seus companys de clúster és:

$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} d(i,j) \qquad (8)$$

essent $d(i,j)$ la distància entre els individus i i j en el clúster C_i. El sumatori de les distàncies es divideix entre $|C_i| - 1$ ja que no s'inclou la distància $d(i,i)$ en la suma. Així doncs, com més petit sigui el valor d' $a(i)$ millor serà l'assignació.

D'altra banda, la distància mitjana des de cada individu i fins a tots els punts de qualsevol altre clúster esdevé:

$$b(i) = \min_{k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} d(i,j) \qquad (9)$$

En aquest cas es diu que el clúster amb la diferència mitjana més petita és el "clúster veí" perquè és el següent punt que millor s'ajusta a l'individu i. Al contrari d' $a(i)$, com més gran sigui el valor de $b(i)$ més ben assignada estarà l'observació.

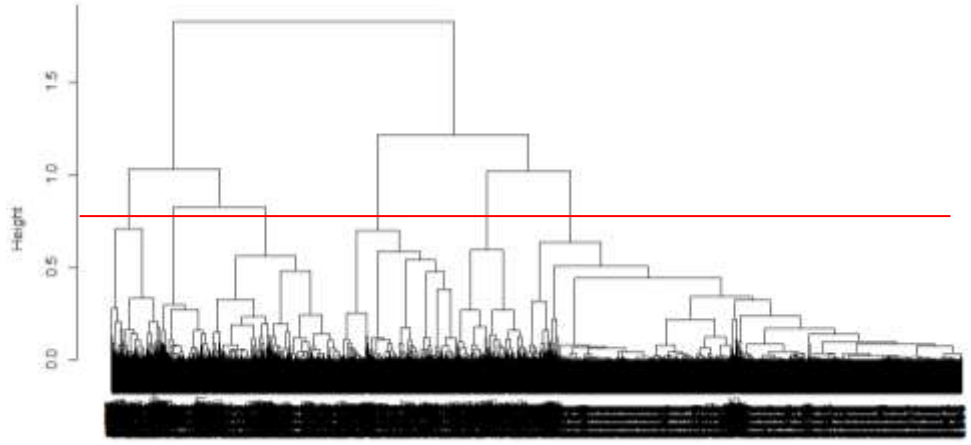
5. Resultats

En aquest darrer apartat de l'anàlisi de conglomerats, es pot trobar una descripció detallada dels clústers obtinguts mitjançant els diversos criteris d'agrupació esmentats anteriorment, així com un petit resum del procés que s'ha dut a terme per tal de realitzar-los. En ambdós casos s'ha utilitzat la distància de Gower per calcular les similituds entre els diversos individus.

5.1Mètode de Ward

En primer lloc, fent ús de la funció *daisy* del software R s'ha calculat la matriu amb els coeficients de similitud de Gower. Amb aquesta matriu de similituds, el mètode de Ward permet fer un dendrograma per tal de determinar el nombre òptim de clústers a realitzar.

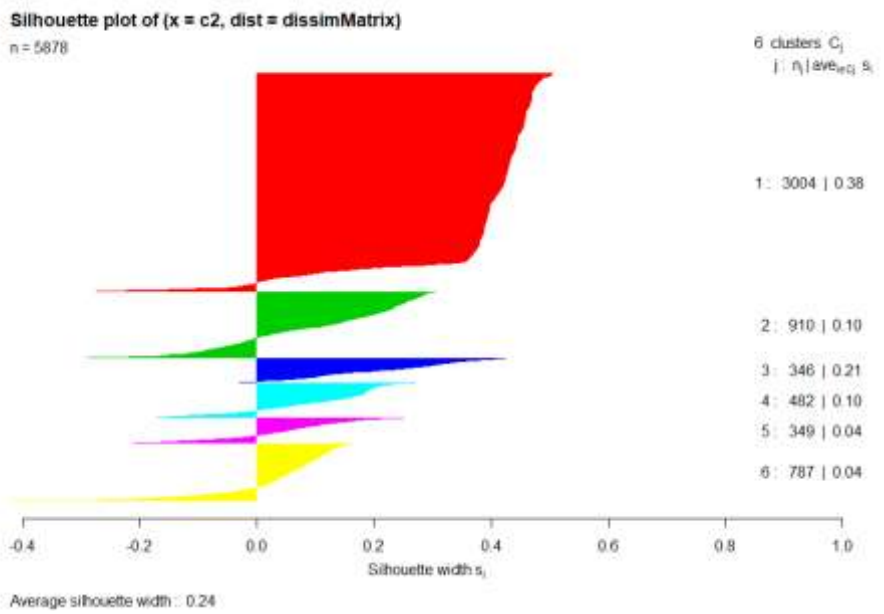
Gràfic 4.5.1.1 Dendrograma Mètode de Ward



Tenint en compte les demandes del client i fent ús del gràfic adjunt, es considera convenient realitzar un total de 6 grups. D'aquesta manera es podrà obtenir una quantitat considerable de conjunts amb característiques pròpies que ajudin a segmentar el mercat de clients.

Per tal de validar la decisió presa, mitjançant el mètode de *Silhouette* i la seva respectiva funció d'R, s'ha creat un gràfic on es poden observar els valor mitjans de la mesura en qüestió. Aquesta mètrica quantifica com de similar és un individu respecte el seu propi conglomerat en comparació amb els altres.

Gràfic 4.5.1.2 Trama de les mesures de *Silhouette*



En el gràfic adjunt es pot veure quina proporció d'individus tenen una valor mitjà superior al zero i quins no. Aquells objectes que tenen un valor negatiu, es considera que no han estat ben classificats i que per tant encaixarien millor en un altre conglomerat. Així doncs, tot i que en el segon i sisè clúster sembla que hi hagi una quantitat considerable d'individus mal col·locats, s'accepta aquesta agrupació com a correcta ja que comparant-la amb les altres possibilitats que sembla haver-hi en el dendrograma (3 o 5 clústers), l'última realitzada és la que obté uns valors mitjans més elevats.

La funció *hclust* del R ha estat la que ha permès crear, mitjançant el mètode de Ward, un total de 6 clústers. Una vegada efectuats tots els càlculs, s'ha realitzat un anàlisi descriptiu de cadascun dels grups. L'objectiu d'aquest anàlisi és obtenir les característiques principals de cada cúmul, tant les numèriques com categòriques.

A continuació, es troben adjunts un seguit de gràfics i taules per tal de trobar quins trets característics dels individus han esdevingut importants a l'hora de crear els diversos conjunts de dades.

En primer lloc, per tal de conèixer la naturalesa de cada un dels clústers es considera necessari presentar-ne les seves dimensions.

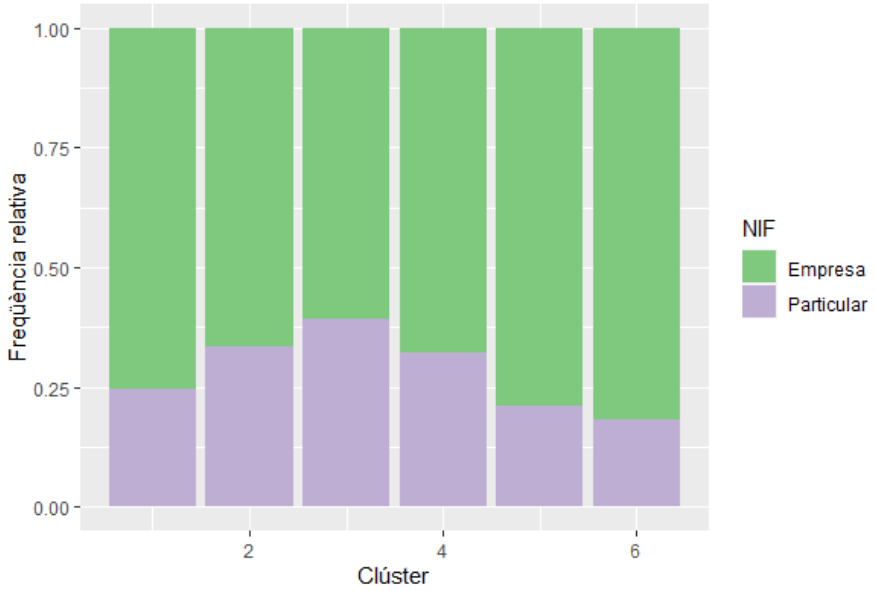
Taula 4.5.1.1 Grandària dels clústers

	Freqüència absoluta	Freqüència relativa
Clúster 1	3004	0.5111
Clúster 2	910	0.1548
Clúster 3	346	0.0589
Clúster 4	482	0.0820
Clúster 5	349	0.0594
Clúster 6	787	0.1339

Com es pot veure, s'ha creat un cúmul molt nombrós respecte als altres que agrupa el 51,11% dels individus, amb un total de 3004 clients. En canvi, dels cinc clústers restants cap d'ells arriba al miler de clients.

La següent característica que es considera necessària analitzar és el perfil del client (particular o empresa).

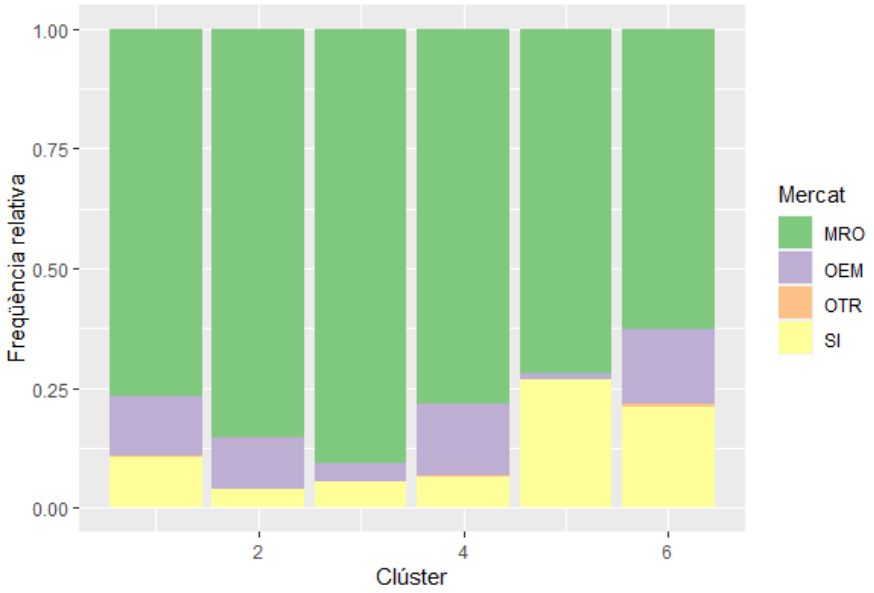
Gràfic 4.5.1.3 Proporció de clients a cada clúster en funció del seu NIF



Tot i semblar que la variable NIF no ha estat important a l'hora de crear els clústers, s'ha realitzat un test binomial exacte per determinar si la proporció de clients particulars i d'empresa en cada un dels clústers és la mateixa que en el total de la població. Per tal de realitzar aquesta prova s'ha utilitzat la comanda *binom.test* del software R. Amb un nivell de significació del 95% es pot afirmar que el primer, el cinquè i el sisè clúster tenen un proporció de clients particulars significativament inferior que al total, mentre que en el segon, el tercer i el quart clúster, aquesta proporció és significativament superior a la del total.

Continuant amb les variables categòriques, la següent característica a estudiar és el mercat del qual formen part els diversos individus en funció dels clúster que se'ls hi ha assignat.

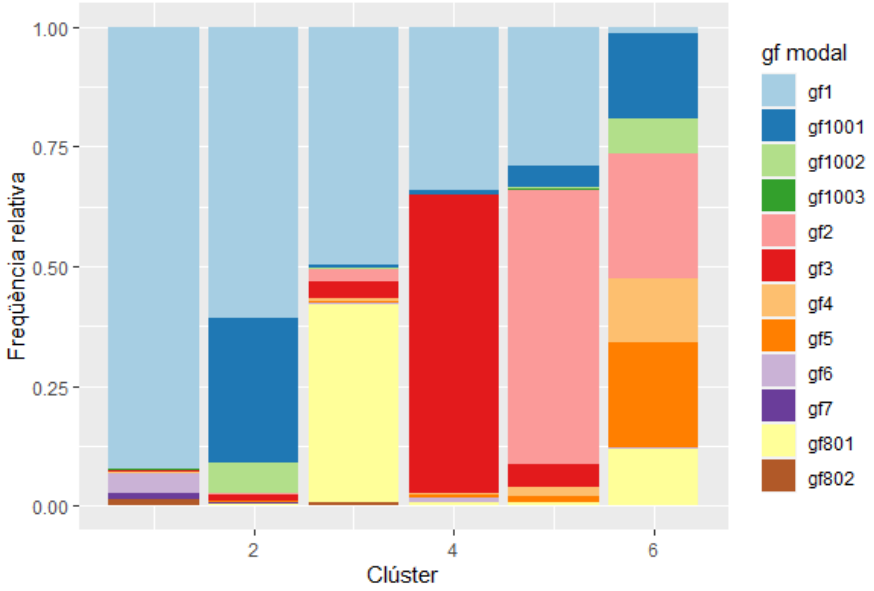
Gràfic 4.5.1.4 Proporció de clients a cada clúster en funció del mercat



Els quatre primers cúmuls presenten unes proporcions força similars pel que fa a la distribució de la variable mercat. No obstant, els dos darrers clústers destaquen per l'elevada presència de individus pertanyents al mercat de sistemes industrials. Un altre aspecte a valorar és la poca presència de clients del mercat OEM en els clústers 3 i 5.

Per acabar amb la descriptiva de les variables categòriques, s'adjunta un gràfic en el qual es pot veure en quin grup família (*gf*) s'ha gastat la majoria dels diners cada un dels clients.

Gràfic 4.5.1.4 Proporció de clients a cada clúster en funció del seu gf modal



Abans de descriure les proporcions en cada un dels clústers creats, cal destacar l'elevada presència de compres en estanquitat dels 5 primers cúmuls, sobretot en els tres inicials. En el clúster 1 més del 90% dels clients ha comprat majoritàriament en aquest grup família. En el segon grup, tot i que es redueix aquesta proporció, més del 60% dels clients continua tenint com a grup família modal l'estanquitat. No obstant, quasi el 30% dels individus han gastat la majoria dels seus diners en adhesius. Pel que fa al tercer cúmul, es redueix al 50% els individus que tenen el gf1 com a grup família modal ja que gairebé la meitat de clients restants, tenen com a grup família modal les eines.

El clúster 4 destaca per tenir com a grup família modal majoritari els rodaments i coixinets. Malgrat això la presència de clients que gasten la major part dels seus diners en estanquitat continua estant al voltant del 30%. En el cinquè cúmul més de la meitat dels clients destaquen per les seves compres de segells mecànics. No obstant, una quarta part dels clients continuen tenint com a grup família modal l'estanquitat.

Finalment, el sisè clúster és el més diferenciat de la resta ja que els clients no destaquen en cap grup família en concret, sinó que la distribució en funció de la variable en qüestió es veu repartida entre 6 grups famílies diferents. En ordre descendent, els grups famílies són els següents: segells mecànics, control del moviment, adhesius, transmissors de potència, eines i finalment, lubricació. El primer de tots acull un 25% dels clients, mentre que el darrer es troba al voltant del 10% dels clients.

D'altra banda per descriure els clústers en funció de les variables numèriques, s'ha considerat més adequat presentar els valors que han pres aquestes en una taula. En la primera fila, es presenta el valor mitjà que pren cada una de les diverses variables sense tenir en compte els clústers creats. D'aquesta manera esdevindrà més senzill comparar els grups i saber en què destaquen cada un d'ells.

Taula 4.5.1.2 Resum de les principals variables numèriques

	Preu mig del total de compres	Mitjana de compres realitzades	Percentatge de compres realitzades a la botiga	Índex de Variació Qualitativa	Raó de Variació	Percentatge d'enviaments realitzats per la distribuïdora
Clúster 1	222.41	2.91	61.91	0.05	0.02	14.27
Clúster 2	308.80	8.44	83.61	0.68	0.28	3.25
Clúster 3	289.07	7.70	90.63	0.66	0.28	2.40
Clúster 4	371.93	5.57	61.75	0.45	0.19	28.86
Clúster 5	635.14	10.96	66.59	0.69	0.33	2.54
Clúster 6	403.11	2.06	37.20	0.08	0.03	22.23
Valors mitjans	300.67	4.63	63.91	0.26	0.11	13.44

Pel que fa al preu mitjà de les compres realitzades al llarg de l'any, el cinquè clúster destaca per la gran quantitat de diners gastats (635,14€/any). El sisè clúster tot i no prendre un valor tant extrem, també gasta una quantitat de diners considerable (403.11€/any). D'altra banda el primer grup, que com bé s'ha dit anteriorment comprèn més de la meitat dels clients, és el que menys diners gasta al llarg de l'any, amb una mitjana de 222,41€.

Amb relació al total de compres realitzades al llarg de l'any, torna a destacar per l'elevat valor que pren, el clúster número cinc amb quasi onze compres anuals. No obstant, el sisè clúster que també gasta forces diners durant l'any, destaca per tant sols realitzar dues compres anuals. Com en la variable anterior, el clúster majoritari torna a ressaltar per les poques compres que realitza.

Si s'analitza el percentatge de compres realitzades a la botiga hi ha tres grups que cal destacar. D'una banda, el sisè clúster tant sols fa un 37,20% de les seves compres a la botiga. D'altra banda, el segon i tercer cúmul ressalten per fer més del 80% de les compres presencialment. Finalment, el primer, quart i cinquè clúster realitzen al voltant d'un 60% de compres a la botiga.

Respecte les variables que quantifiquen la variabilitat entre les compres que s'han fet en funció del grup família en el qual s'ha comprat (IVQ i RV), destaquen el primer i el sisè grup pels seus baixos valors. Aquest fet indica que els individus pertanyents a aquests clústers habituen a comprar sempre productes del mateix grup família, és a dir, que estan especialitzats en algun sector molt concret.

Finalment, si s'analitza el percentatge d'enviaments que s'han realitzat a través dels mitjans de la distribuïdora, sembla ser que hi hagi dues tendències. Els clients pertanyents als clúster dos, tres i cinc no utilitzen quasi mai el servei d'enviament de la distribuïdora. D'altra banda, els individus dels grups u, quatre i sis en fan un ús mitjà del 14, 29 i 22% respectivament.

Un cop descrites les variables rellevants de l'anàlisi es prossegueix a realitzar una breu descripció per cada un dels clústers creats:

- **Clúster 1:** és el cúmul més nombrós ja que està format pel 51,11% del total de clients, una quarta part dels quals són particulars. Pertanyen majoritàriament al mercat de MRO, tot i tenir clients dels altres dos mercats. Habituen a gastar la majoria dels seus diners en estanquitat i no varien de grup família en les diverses compres que realitzen al llarg de l'any. El preu mitjà en el total de compres anuals destaca per ser el més baix de tots els clústers i tant sols realitzen al voltant de 3 compres anuals. Acostuma a comprar a la botiga i fer ús dels seus propis mitjans per endur-se la compra.
- **Clúster 2:** és el segon grup més nombrós, amb un 15,48% del total de clients, tant empreses com particulars. La majoria d'aquests formen part del mercat MRO i habituen a comprar adhesius i productes d'estanquitat. La seva despesa anual en la

distribuïdora és similar a la mitjana però en canvi realitza més compres del que és habitual. Aquestes adquisicions les du a terme a la botiga física i es pot considerar que mai utilitza el servei d'enviament.

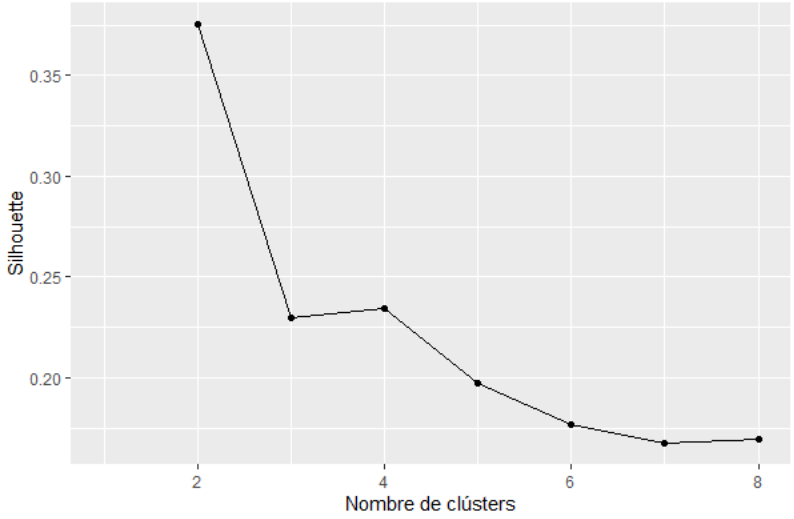
- **Clúster 3:** està format per tant sols el 5,89% dels clients, amb una proporció aproximada del 40-60 en particulars-empreses. Formen part del mercat MRO i habituen a comprar eines i productes d'estanquitat. Gasten aproximadament els mateixos diners que la mitjana, el 90% de les compres les realitza de forma presencial i per tant, no fa ús del servei d'enviament.
- **Clúster 4:** hi pertanyen el 8,20% dels clients, tant empreses com particulars. Formen part del mercat MRO majoritàriament, tot i comprar als tres mercats principals. Destaca per gastar la majoria dels diners en rodaments i coixinets. Es deixa aproximadament 370€ anuals a la distribuïdora i és el grup que utilitza més el servei d'enviament tot i realitzar més del 60% de les compres presencialment.
- **Clúster 5:** està format pel 5,94% dels clients. Es caracteritza per tenir més d'un 25% dels individus en el mercat de SI i per gastar la majoria dels seus diners en segells mecànics. És el cúmul de clients que realitza més compres al llarg de l'any i també el que es gasta més diners en la distribuïdora. Així doncs, destaca tant pel volum com per la freqüència de compres.
- **Clúster 6:** hi pertanyen tant empreses com particulars, representant el 13,35% del total d'individus. Té clients de tots els mercats i destaca per comprar en molts grups família diferents. Principalment compren adhesius, segells mecànics o bé productes de control de moviment. Només realitza dues compres anuals, habitualment telemàtiques, però es gasten més de 400€ en la distribuïdora. Per tant, aquest últim grup es caracteritza per gastar-se molts diners tot i realitzar poques compres llarg de l'any.

5.2 Partition Around Medoids

A diferència del Mètode de Ward, amb el PAM és necessari decidir el nombre total de clústers (*k*) des d'un primer moment. Existeixen diverses mesures que ajuden a triar aquesta *k,* no obstant, com ja ha fet anteriorment es farà ús Silhouette. Cal recordar que aquesta mètrica de validació interna proporciona una mesura de similitud d'una observació amb el seu propi clúster en comparació amb el seu clúster veí més proper.

La mesura de Silhouette es troba en el rang [-1, 1], on els valors més elevats indiquen una similitud més elevada. Així doncs, s'ha calculat la mesura per als clústers, que van des de 2 fins a 8, mitjançant l'algoritme PAM i s'ha realitzat un gràfic per observar el nombre òptim de grups a realitzar.

Gràfic 4.5.2.1 Nombre de clústers òptims en funció de *Silhouette*



La separació dels clients en dos únics grups, no permet l'assoliment de l'objectiu principal d'aquest estudi. Així doncs, tot i que sigui la *k* òptima, no serà la utilitzada. Si es continua analitzant el gràfic fent cas de la mesura de *Silhouette* esdevindria raonable seleccionar un total de 4 clústers, però la distribuïdora demana segmentar encara més el mercat. Per tant, acceptant les seves sol·licituds i necessitats es decideixen realitzar un total de 6 clústers. L'avantatge d'aquesta decisió, serà poder comparar els resultats d'aquest apartat amb els obtinguts amb el mètode de Ward.

En aquest cas, la creació dels clústers es farà mitjançant la matriu de coeficients de similitud de Gower, calculada anteriorment, i la funció *pam* del software R.

A diferència del mètode de Ward, mitjançant els *k-medoids* s'obtenen 6 clústers de tamanys molt més similars. Aquests valors es troben descrits en la següent taula:

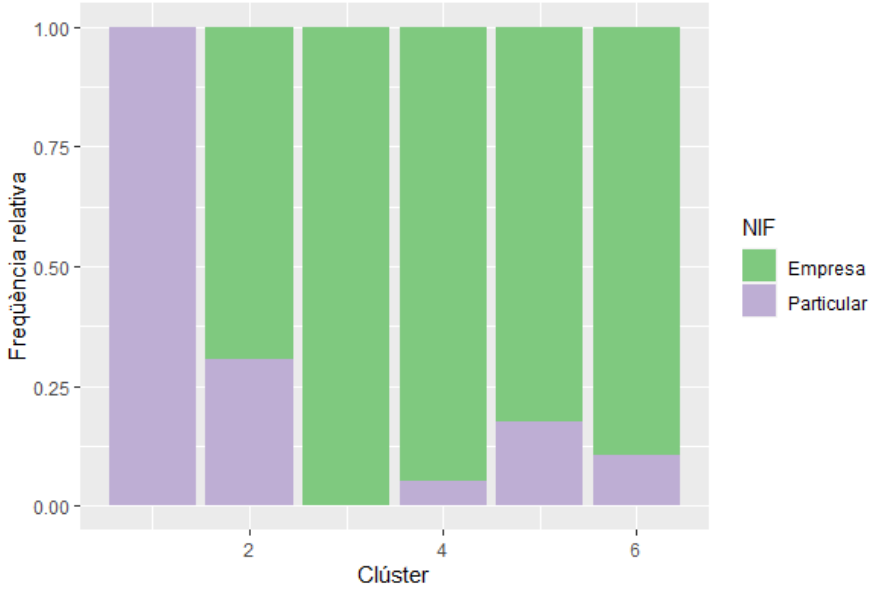
Taula 4.5.2.1 Tamany dels clústers

	Freqüència absoluta	Freqüència relativa
Clúster 1	1042	0.1773
Clúster 2	1065	0.1812
Clúster 3	1734	0.2950
Clúster 4	1226	0.2086
Clúster 5	424	0.0721
Clúster 6	387	0.0658

El clúster que aglutina més clients és el tercer, amb gairebé un 30% dels individus. Deixant de banda aquest cúmul, es poden separar la resta de clústers en dos grups. D'una banda, el primer, el segon i el quart clúster conten al voltant d'un 20% de clients cada un d'ells, mentre que el cinquè i sisè clúster no arriben al 10%.

Un cop conegudes les dimensions dels diversos grups, es prossegueix a fer l'anàlisi descriptiu de les variables categòriques en funció del clústers. En primer lloc, es considera necessari analitzar el perfil del client (particular o empresa).

Gràfic 4.5.2.2 Proporció de clients a cada clúster en funció del seu NIF



En aquest cas no és necessari realitzar un test binomial exacte, ja que el gràfic ens permet observar que la variable NIF ha esdevingut significativa a l'hora de crear els diversos grups. El primer clúster tant sols conté clients particulars, mentre que el tercer tant sols conté empreses. Els cúmuls restants integren tot tipus de clients, tot i que en el clúster número quatre la presència de particulars és tant sols del 5,06%.

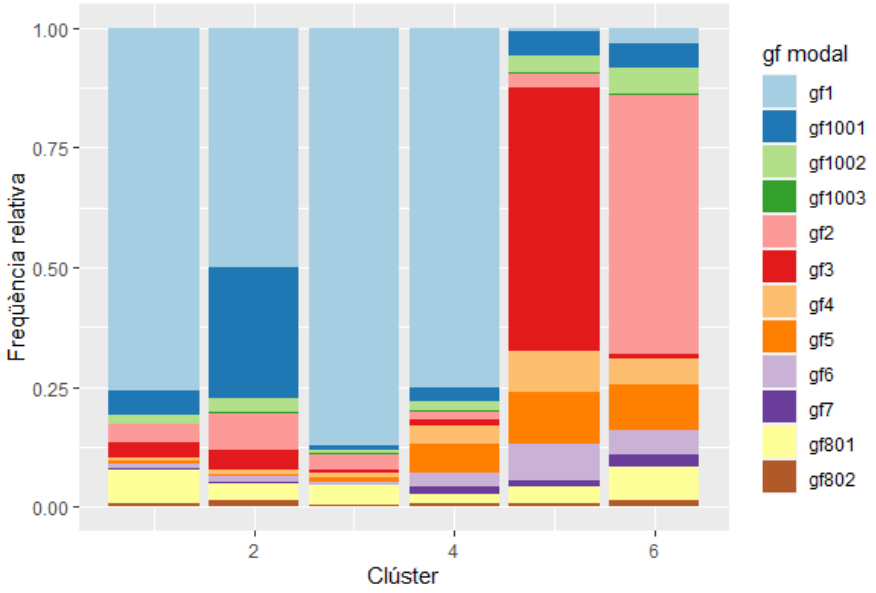
La següent variable qualitativa a analitzar és el mercat al qual pertanyen els clients de cada clúster.

Gràfic 4.5.2.3 Proporció de clients a cada clúster en funció del mercat

Els tres primers cúmuls són majoritàriament clients del mercat MRO, tot i tenir una petita proporció d'individus pertanyents a altres mercats. Tant el clúster quatre com el cinc, continuen tenint un percentatge molt elevat de clients en MRO, però al voltant d'un 20% dels individus compren al mercat OEM. Finalment, el sisè clúster és el més rellevant ja que més de la meitat dels seus clients formen part del mercat de sistemes industrials.

L'última variable categòrica a analitzar és el grup família modal (*gf*) en cada clúster creat.

Gràfic 4.5.2.4 Proporció de clients a cada clúster en funció del seu gf modal



La descripció dels clústers en funció del grup família modal la podem dividir en 4 conjunts. El primer d'ells compren els clústers número u, tres i quatre, i es caracteritza per tenir més d'un 75% dels clients els quals han gastat la majoria dels seus diners en el grup família estanquitat. En segon lloc, ens trobem el clúster número dos, que tot i tenir una proporció del 50% dels seus individus que tenen estanquitat com a grup família modal, destaca per tenir un 25% dels clients en el grup família adhesius.

D'altra banda el cinquè i el sisè cúmul es diferencien de la resta per tenir un percentatge molt disminuït de clients que tinguin com a grup família modal l'estanquitat. El cinquè clúster destaca per tenir el 50% de clients amb el grup família modal a rodaments i coixinets. En canvi, el sisè grup té la meitat d'individus amb segells mecànics com a grup modal.

A fi de descriure els clústers en funció de les variables numèriques, com bé s'ha fet en l'apartat anterior, es presenten els valors que han pres aquestes en una taula. En la primera fila, es presenta el valor mitjà que pren cada una de les diverses variables sense tenir en compte els clústers creats. D'aquesta manera es facilita la comparació dels clústers i les seves característiques principals.

Taula 4.5.2.2 Resum de les principals variables numèriques

	Preu mig del total de compres	Mitjana de compres realitzades	Percentatge de compres realitzades a la botiga	Índex de Variació Qualitativa	Raó de Variació	Percentatge d'enviaments realitzats per la distribuïdora
Clúster 1	103.73	2.61	90.32	0.20	0.08	4.56
Clúster 2	242.32	7.06	87.73	0.76	0.34	31.07
Clúster 3	199.41	4.97	87.66	0.12	0.04	13.06
Clúster 4	554.53	4.63	8.78	0.12	0.05	29.50
Clúster 5	413.32	2.37	32.25	0.18	0.08	25.87
Clúster 6	517.56	4.38	30.24	0.17	0.07	0.68
Valors mitjans	300.67	4.63	63.91	0.26	0.11	13.44

Pel que fa al preu mig de les compres realitzades al llarg de l'any, sembla haver-hi dos tipus de clústers: els de clients que gasten menys diners que la mitjana i els que en gasten més. El primer clúster conté clients que gasten al voltant de 100€ en compres al llarg de l'any, mentre que el segon i el tercer cúmul estan al voltant dels 200€. D'altra banda els tres últims clústers són els que gasten més diners. El clúster número 5 gasta al voltant de 400€ en compres anuals, mentre que el quart i el sisè clúster es deixen més de 500€ anuals en productes de la distribuïdora.

Amb relació al total de compres realitzades al llarg de l'any, sembla haver-hi 3 tendències diferents. El primer i el cinquè clúster es caracteritzen per comprar al voltant de 2,5 vegades a l'any. El tercer, quart i sisè cúmul prenen aproximadament el mateix valor que la mitjana, és a dir entre 4 i 5 compres anuals. Finalment, el segon cúmul destaca per l'elevada quantitat de compres realitzades al llarg de l'any, que es troben prop de les set.

Si s'estudia el percentatge de compres realitzades a la botiga hi ha tres grups que cal destacar. En primer lloc, el quart clúster destaca per no arribar al 10% de compres presencials. D'altra banda, els tres primers clústers realitzen la majoria de compres a la botiga física. Finalment el cinquè i el sisè clúster tant sols fan una tercera part de les seves compres presencialment.

Pel que fa a les variables que quantifiquen la variabilitat entre les compres que s'han fet en funció del grup família en el qual s'ha comprat (IVQ i RV), cal destacar el segon cúmul pels seus elevats valors.

Finalment, si s'analitza el percentatge d'enviaments que s'han realitzat a través dels mitjans de la distribuïdora, sembla ser que hi hagi tres tendències. Els clients pertanyents als clúster u i 6 no utilitzen quasi mai el servei d'enviament de la distribuïdora. D'altra banda, els individus dels grups tres l'acostumen a utilitzar en el 13% de les seves compres. Finalment, els cúmuls dos, quatre i cinc l'utilitzen aproximadament entre el 25 i el 30% de les vegades.

Per acabar aquest apartat, es realitzarà una breu descripció per cada un dels clústers creats:

- **Clúster 1:** està format únicament per particulars, exactament per 1042 clients. Pertanyen majoritàriament al mercat MRO i tenen tendència a comprar productes d'estanquitat. La seva despesa en la distribuïdora és reduïda i es produeix entre dues i tres vegades a l'any de manera presencial. Aquest últim fet provoca que quasi mai facin ús del servei d'enviament.
- **Clúster 2:** hi pertanyen clients dels dos perfils, tant empreses com particulars. Predominen individus del mercat MRO i habituen a comprar adhesius i productes d'estanquitat. Són el cúmul que realitza més compres al llarg de l'any, les realitzen a la botiga física i aquestes no són homogènies, és a dir que habituen a comprar en més d'un grup família al llarg de l'any. La seva despesa anual és inferior a la mitjana.
- **Clúster 3:** està format únicament per empreses i representen el 29,50% dels clients totals, sent així el cúmul més nombrós. Predominen els clients del mercat manteniment, reparacions i operacions que compres productes d'estanquitat. Són el segon grup amb una despesa més baixa, tot i realitzar quasi 5 compres presencials al llarg de l'any.

- **Clúster 4:** hi pertanyen majoritàriament empreses ja que el percentatge de particulars dins del clúster no arriba al 10%. Inclou clients de tots tres mercats, però continuen comprant majoritàriament productes d'estanquitat. Realitza al voltant de 5 compres a l'any i és el cúmul que es gasta més diners. Les compres quasi sempre les realitza de manera telemàtica. Per les seves característiques descrites, es poden considerar clients potencials.
- **Clúster 5:** grup reduït format per clients particulars i empreses. Pertanyen principalment al mercat MRO i OEM, però destaquen per comprar rodaments i coixinets. Gasten una quantitat elevada de diners però per contra, realitzen poques compres. Aquest fet ens indica que les compres són grans despeses. Habituen a adquirir els productes de manera telemàtica tot i que també compren a la botiga física.
- **Clúster 6:** cúmul reduït de clients particulars i empreses. Destaquen la meitat dels clients per pertànyer al mercat de sistemes industrials i comprar majoritàriament al grup família de segells mecànics. Com en els dos clústers anteriors, la despesa que realitzen a la distribuïdora és elevada. Fan un total d'aproximadament quatre compres a l'any i es pot considerar que mai utilitzen el servei d'enviament tot i realitzar quasi un 70% de les seves compres de forma no presencial.

V. ANÀLISI DE SENSIBILITAT DE LA DISTÀNCIA DE GOWER

Anteriorment, en l'apartat de selecció de variables, s'ha esmentat el procediment que s'ha dut a terme per triar quines variables formaven part de l'estudi i quines no. Durant el procés de simulacions de clústers amb les diverses variables dels resums anuals, es va poder observar que quan s'incloïa la variable categòrica que feia referència a la província a la qual pertanyia el client, els clústers perdien tota la seva estructura per passar a agrupar els clients en funció d'aquest ítem.

Com que es va considerar alarmant aquest canvi en l'agrupació dels individus a causa d'una variable categòrica, es va creure convenient realitzar aquest apartat i observar quin pes se li dóna a les variables qualitatives en la distància de Gower.

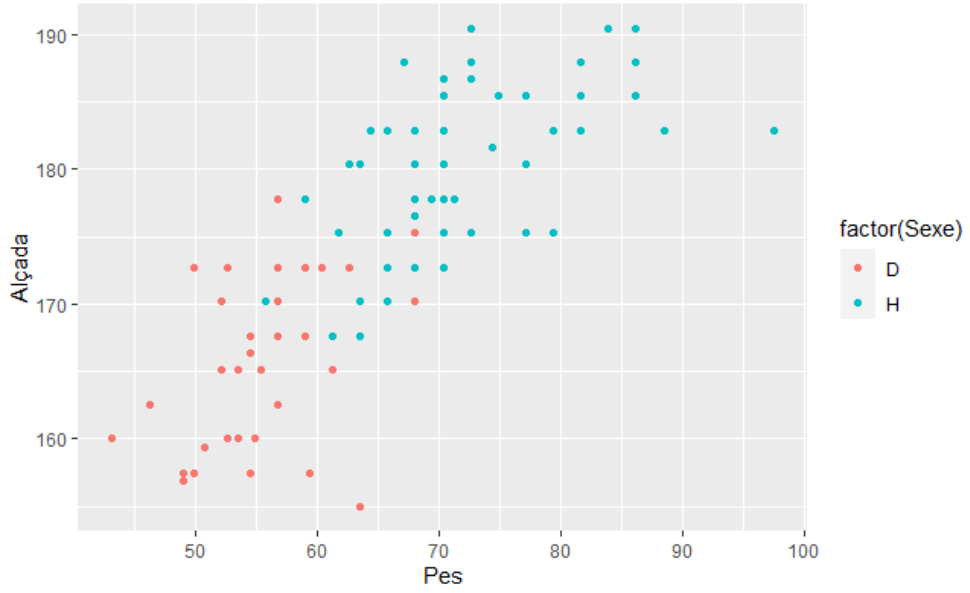
Per tal de dur a terme aquest anàlisi de sensibilitat s'ha fet ús d'una base de dades amb 92 observacions. Cadascuna d'aquestes fa referència a un individu diferent i inclou la seva informació referent al sexe, l'alçada en centímetres i el pes en quilograms. No obstant, s'han afegit dues variables categòriques creades de forma aleatòria, una amb dos nivells i una altra amb sis. A continuació es pot veure les primeres observacions d'aquesta base de dades:

Taula 5.1. Estructura de la base de dades "Simulacions"

	Sexe	Alçada	Pes	Var_2	Var_6
1	H	167.640	63.50288	1	3
2	H	182.880	65.77084	2	4
3	H	186.690	72.57472	1	2
4	H	185.420	86.18248	2	2
5	H	175.260	70.30676	2	5
6	H	185.420	74.84268	1	1
7	H	182.880	68.03880	2	3
8	H	187.960	86.18248	2	4
9	H	182.880	88.45044	2	4
10	H	180.340	62.59570	1	2

A fi de conèixer una mica millor la naturalesa de la base de dades inicial, seguidament s'adjunta un diagrama bivariant en el qual es representen els diversos individus en base al seu pes (eix d'abscisses), la seva alçada (eix d'ordenades) i al seu sexe (color).

Gràfic 5.1. Gràfic de dispersió dels inidivuds de la base de dades “Simulacions”

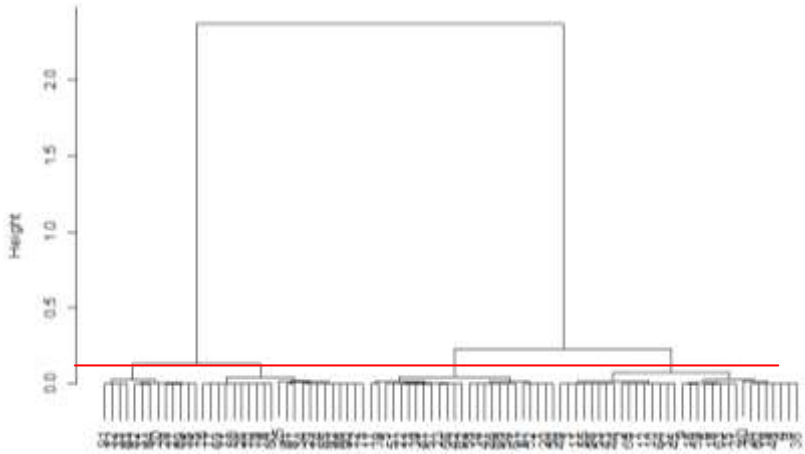


Si la distància de Gower funcionés de manera adequada, els individus sempre serien agrupats en funció de les seves condicions físiques. Per contra, si se li donés un pes massa important a les variables categòriques els individus podrien ser agrupats en funció d’aquestes, tot i no contenir informació sobre l’essència real dels subjectes.

A fi d’observar les diferències entre la clusterització mitjançant les 3 variables reals i la clusterització quan se li afegeixen les variables categòriques aleatòries, es realitzaran un total de 3 simulacions. La primera de totes serà per conèixer l’essència real de les dades, és a dir, només s’utilitzaran les tres variables reals i es crearan els conglomerats mitjançant el mètode de Ward, decidint el nombre total d’aquests en funció del seu dendrograma. Les dues simulacions restants seran per dur a terme les comparacions entre els clústers originals i aquells als quals se li han afegit variables aleatòries.

En primer lloc, amb la base de dades original, s’ha calculat la distància de Gower i s’ha realitzat el procés de clusterització de Ward fent ús de les mateixes comandes del R que s’havien utilitzat anteriorment. El dendrograma resultant ha estat el següent:

Gràfic 5.2. Dendrograma amb les variables originals

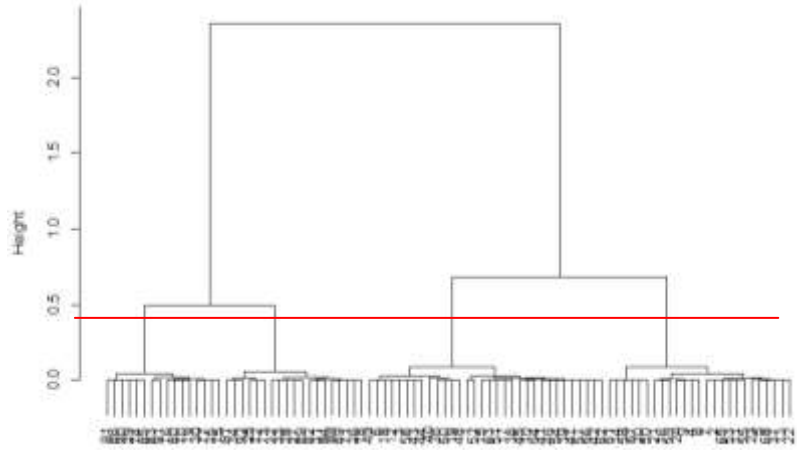


Tot i que analitzant el gràfic el més coherent és fer tan sols dos clústers, s'ha considerat oportú realitzar-ne un total de 4. Com bé s'havia pronosticat, els grups d'individus resultants s'agrupen en funció de la seva condició física:

- **Clúster 1:** Individus de sexe masculí, de baixa alçada i poc pes.
- **Clúster 2:** Individus de sexe masculí, amb alçada i pes superiors.
- **Clúster 3:** Individus de sexe femení, de baixa alçada i poc pes.
- **Clúster 4:** Individus de sexe femení, amb alçada i pes superiors.

Seguidament, s'ha realitzat el mateix procés però afegint a la base de dades original la variable categòrica de dos nivells. El dendrograma obtingut ha estat el següent:

Gràfic 5.3. Dendrograma amb les variables originals i la variable categòrica amb 2 nivells

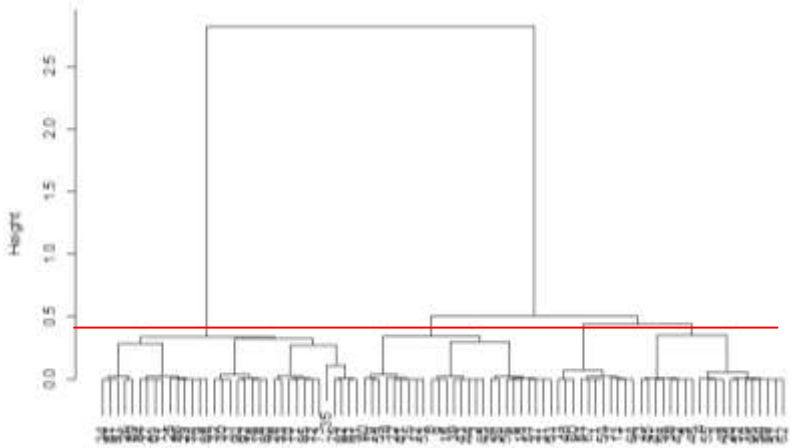


En aquest gràfic encara és més evident la necessitat de crear 4 grups. Per contra, els clústers resultants aquest cop han deixat de tenir relació amb les condicions físiques de l'individu, i s'han creat fent ús d'aquesta variable aleatòria:

- **Clúster 1:** Homes en el nivell 1 de la variable qualitativa aleatòria.
- **Clúster 2:** Homes en el nivell 2 de la variable qualitativa aleatòria.
- **Clúster 3:** Dones en el nivell 2 de la variable qualitativa aleatòria.
- **Clúster 4:** Dones en el nivell 1 de la variable qualitativa aleatòria.

Finalment, seguint amb el procediment realitzat fins ara, s'ha afegit la variable aleatòria de 6 nivells a la base de dades original. Calculant la distància de Gower i fent ús del mètode de Ward, el dendrograma resultant ha estat el següent:

Gràfic 5.4. Dendrograma amb les variables originals i la variable categòrica amb 6 nivells



En aquest últim gràfic la decisió ha esdevingut més complicada ja que un cop feta la primera partició en dos grups, les branques del dendrograma pateixen diverses bifurcacions. No obstant, s'ha considerat convenient fer 4 grups ja que es poden observar clarament al dendrograma i permeten comparar el resultat actual amb els calculats anteriorment. Una vegada més, els grups d'individus no s'han dut a terme tenint en compte les seves condicions físiques sinó que s'han centrat en les variables categòriques:

- **Clúster 1:** Homes en el nivell 3 de la variable categòrica aleatòria.
- **Clúster 2:** Homes en el nivell 1, 4 i 6 de la variable categòrica aleatòria.
- **Clúster 3:** Homes en el nivell 2 i 5 de la variable categòrica aleatòria.
- **Clúster 4:** Dones en tots els nivells de la variable categòrica aleatòria.

Un cop realitzades aquestes tres simulacions, es pot assegurar que hi ha una manca de consistència en la distància de Gower quan s'afegeixen variables categòriques en l'anàlisi. Per tal de reafirmar aquesta idea s'ha fet una cerca bibliogràfica per veure quines explicacions hi ha darrere d'aquest inconvenient i les possibles alternatives que s'haurien de dur a terme a fi d'erradicar-lo.

Si analitzem detalladament la fórmula de Gower (3) podem observar que aquesta tendeix a donar pesos molt més grans a les variables categòriques que a les contínues. Aquest fet és fruit de que quan es tracten variables categòriques la distància entre elles només pot prendre valor 0 o 1. Així doncs, dos individus que comparteixin el mateix valor en totes les variables quantitatives obtindran un valor molt elevat en el seu coeficient de similitud de Gower. Contràriament, quan la distància està analitzant variables numèriques el seu coeficient no pot ser igual a la unitat, a no sé que les variables siguin exactament iguals.

Tanmateix, una altra de les mancances que s'observa en aquest coeficient és el tractament que reben les variables quantitatives. Si s'aïlla la part de la funció en la qual es calcula la similitud entre dos individus respecte la seva variable numèrica, s'obté la següent expressió:

$$1 - \frac{|x_{ih} - x_{jh}|}{G_h} \qquad (10)$$

Mentre que el numerador és la diferència en termes absoluts d'ambdues variables numèriques, en el denominador s'hi troba el rang de la variable en qüestió. Per tant, si existeix un individu amb un valor atípic en una d'aquestes variables el rang esdevindrà elevat fet que impedirà la detecció de diferències reals entre la resta d'observacions.

Recollint tot el que s'ha dit al llarg d'aquest apartat, el fet més preocupant és el tractament que reben les variables categòriques en la distància de Gower ja que és el que influeix més en l'anàlisi de conglomerats realitzat. Consegüentment, s'ha considerat necessari buscar alteratives possibles per evitar que aquesta mancança afecti al procés de clusterització.

La primera possibilitat, té en compte que no totes les variables tenen la mateixa importància en la determinació de les característiques principals que permeten descriure els individus. Així doncs, considera que una bona solució pel problema descrit és crear un vector amb els pesos (w_k) que se li assignen a cada variable (k) a l'hora de calcular la distància entre les diverses observacions. D'aquesta manera, es podria desinflar el pes que la pròpia distància de Gower dóna a les variables quantitatives assignant-les-hi un pes inferior al de la resta. Tot i que aquesta alternativa sembla una solució òptima i senzilla té un problema que s'ha de considerar. L'assignació dels pesos es pot establir amb els coneixements que té l'estadístic envers les dades, però habitualment és necessari una cerca més profunda i acurada per tal d'establir els pesos òptims. Així doncs, si es tria per aquesta alternativa és necessari fer una recerca d'informació sobre l'optimització dels pesos en aquesta distància.

Deixant de banda la distància de Gower, una altra opció per estudiar conjuntament variables qualitatives i quantitatives ha estat extreta de Sakar *et al.* (2015). En aquest estudi, es proposen un total de 6 mesures per a calcular la distància entre individus en bases de dades mixtes. Aquestes sis possibilitats són fruit d'una barreja de tres mesures per a les variables quantitatives i dues mesures per a les variables qualitatives.

Abans de presentar els resultats d'aquest estudi, cal remarcar que la combinació que va ser considerada com la millor opció, estava fent-se servir en un estudi de genètica en plantacions d'arròs. Així doncs, caldria veure quina de les combinacions proposades s'adapten millor al nostre propòsit.

Per les variables quantitatives es proposaven les següents tres distàncies:

- La mitjana de la diferència absoluta estandarditzada per rang:

$$A_1 = \frac{1}{p}\sum_{k=1}^{p}\frac{|x_{ik} - x_{jk}|}{r_k} \qquad (11)$$

  És la distància utilitzada en el coeficient de similitud de Gower. On x_{ik} i x_{jk} són els valors que prenen les k variables quantitatives en el i-èssim i el j-èssim individu; r_k és el rang de la k-èssima variable; i p és el nombre total de nombre de variables categòriques.

- El coeficient de correlació de Pearson:

$$A_2 = \left(1 - r_{ij}^2\right) \qquad (12)$$

  On r_{ij} és el producte de la correlació entre el i-èssim i el j-èssim individu.

- La puntuació estàndard reescalada:

$$A_3 = \sum_{k=1}^{p}\frac{\left[\frac{x_{ik} - x_{jk}}{\sigma_k}\right]^2}{\max\left(d_{ij}^*\right)} \qquad (13)$$

  On σ_k és la desviació estàndard de la k-èssima variable i $\max\left(d_{ij}^*\right)$ és la distància màxima entre dues observacions en la base de dades.

En canvi, per les variables qualitatives, tan sols es proposaven dues distàncies:

- El desajustament mitjà:

$$B_1 = \frac{1}{m}\sum_{k=1}^{m} d_k \qquad (14)$$

  Sent m el total de variables qualitatives, $d_k = 0$ si $y_{ik} = y_{jk}$, i $d_k = 1$ en el cas contrari.

- La mitjana de la diferència absoluta:

$$B_2 = \frac{3}{2} * \frac{\frac{1}{m}\sum_{k=1}^{m}|y_{ik} - y_{jk}|}{1 + \frac{1}{m}\sum_{k=1}^{m}|y_{ik} - y_{jk}|} \qquad (15)$$

  on y_{ik} i y_{jk} són la i-èssima i j-èssima accessió de la k-èssim variable qualitativa, i m el nombre total de variables categòriques

Els rangs que prenen les diverses distàncies enumerades sempre estan entre zero i u. Així doncs, per obtenir la matriu de distàncies total per a bases de dades mixtes, tant sols cal sumar la matriu de distàncies quantitatives amb la matriu de distàncies qualitatives. De totes les combinacions possibles, amb la que es van obtenir millors resultats fou l'A_1B_2.

Recordant tot el que s'ha dit, després de les diverses simulacions realitzades i els resultats obtinguts, es pot afirmar que el coeficient de similitud de Gower no és òptim a l'hora de calcular distàncies en bases de dades mixtes. Realitzant la cerca d'alternatives possibles a aquest problema, s'han obtingut dos resultats possibles que caldria estudiar en propers estudis d'anàlisi multivariant de dades. Sembla ser que l'estudi que tracta d'optimitzar la distància de Gower mitjançant pesos, està prenent importància entre els especialistes en el tema, ja que el nombre de projectes al voltant d'aquest tema està creixent de forma important.

VI. CONCLUSIONS

Un cop finalitzat l'estudi i el corresponent anàlisi de conglomerats, és necessari ressaltar els problemes i els resultats obtinguts. En primer lloc cal recordar que l'objectiu inicial del projecte era obtenir una segmentació dels clients de la distribuïdora, per tal de localitzar quines eren els individus potencials i quina estratègia de compra utilitzaven. No obstant, el treball ha acabat abordant un segon objectiu, de finalitat purament estadística: analitzar si el coeficient de similitud de Gower és robust a la inclusió de variables categòriques en l'anàlisi.

Si s'analitza l'assoliment del primer objectiu, cal tractar els dos mètodes d'agrupació utilitzats per separat:

- El mètode de Ward ha creat un grup amb més de la meitat dels individus, on el comportament d'aquests ha resultat ser força homogeni i estable. Els cinc grups restants es diferencien entre ells a través de les variables numèriques, especialment per la quantitat de diners gastats, el nombre de compres realitzades i l'ús del servei d'enviament. No obstant, la variable categòrica que també ha influït ha estat el grup família modal.

  Així doncs, aquest criteri ha acabat proporcionant dos grups de tamany reduït de possibles clients potencials. El primer gasta més diners, però compra de forma més variable; mentre que el segon grup té una despesa anual més petita però les compres són menys variants.
- El mètode de PAM ha creat grups de mida més similar, tot i que dos d'ells són de tamany significativament més reduït. La variable binària NIF, ha estat decisiva per crear els clústers ja que en un d'ells tant sols hi ha clients particulars i en un altre només hi consten empreses. El mercat al qual pertanyen els clients i el percentatge de compres realitzades a la botiga han estat clau per crear els diversos cúmuls.

  Amb l'ús d'aquest criteri d'agrupació s'han acabat creant un total de 3 cúmuls de possibles clients potencials. El primer d'ells és el que gasta més diners, mai compra a la botiga i els clients que hi pertanyen compren productes d'estanquitat tot i ser dels tres mercats. El segon cúmul de clients potencials, gasta menys però també efectua menys compres al llarg de l'any. Habitua a comprar rodaments i coixinets, tant presencialment com de forma telemàtica. Finalment, l'últim cúmul destaca per tenir una gran presència de clients del mercat de serveis industrials i per comprar segells mecànics. Compra tant a la botiga com de manera telemàtica però mai utilitza el servei d'enviament.

Per finalitzar amb aquest primer bloc, s'ha pogut observar que els dos mètodes d'agrupació han proporcionat una correcta segmentació del mercat. Queda en mans de la distribuïdora decidir quina d'aquestes prefereix per tal d'aplicar les seves tècniques de venda o màrqueting.

No obstant, cal remarcar que els mètodes d'agrupació no han pogut fer ús de totes les variables que es consideraven rellevants, ja que la inclusió d'aquestes en l'estudi comportava grans canvis en la creació dels clústers. Les variables que han quedat fora han estat la província a la qual pertany el client i la manera com realitzava la compra, és a dir el seu contenidor. Quan s'afegien a l'estudi ambdues variables, els clústers creats fins aleshores perdien tota l'estructura i es passaven a crear en funció únicament de les variables categòriques. Aquest fet va ser el que va comportar la creació d'un segon objectiu en l'estudi.

L'apartat d'anàlisi de sensibilitat de la distància de Gower ha permès corroborar la hipòtesis de que el coeficient no és robust a la inclusió de variables categòriques. Tot i ser un mètode fàcil d'utilitzar, ja que permet el tractament de bases de dades mixtes i amb observacions mancants, no es pot considerar que calculi les distàncies de manera adequada.

Gràcies a les diverses simulacions realitzades s'ha pogut comprovar que la inclusió de variables categòriques a una base de dades comporta que l'agrupació de les observacions passi a realitzar-se únicament en funció d'aquestes. És per aquest motiu, que per possibles propers estudis caldria assignar un vector de pesos a les diverses variables per tal de millorar el càlcul de la distància. No obstant, aquesta alternativa suposaria la realització d'un altre projecte ja que hauria d'incloure un estudi del repartiment de pesos òptims per la distància en qüestió.

VII. BIBLIOGRAFIA

Budiaji, W. (2019). *Distance-Based K-Medoids*. Recuperat de: https://cran.r-project.org/web/packages/kmed/vignettes/kmed.html

Budiaji, W; Leisch, F. (2019). *Simple K-Medoids Partitioning Algorithm for Mixed Variable Data*. Recuperat de: https://www.mdpi.com/1999-4893/12/9/177/pdf

Cuadras, M; Salvo-Garrido, S. (2018). *Predicción multivariante basada en distancias.* Recuperat de: http://www.ub.edu/stat/personal/cuadras/Capdbufror70sin.pdf

De la Fuente, S. (2011). *Análisi de conglomerados.* Recuperat de: http://www.estadistica.net/Master-Econometria/Analisis_Cluster.pdf

Gibert, K; Cortés, U. (1997). *Weighting Quantitative and Qualitative Variables in Clustering Methods*. Recuperat de: https://upcommons.upc.edu/bitstream/handle/2099/3494/Gibert%20-%20Cort%C3%A9s.pdf

Gomes, P; *et al* . (2019) *Combinations of distance measures and clustering algorithms in pepper germplasm characterization*. Recuperat de: https://www.researchgate.net/publication/334529932_Combinations_of_distance_measures_and_clustering_algorithms_in_pepper_germplasm_characterization

Guerrero, S; Melo, O. (2017). *Una metodología para el tratamiento de la multicolinealidad a través del escalamiento multidimensional*. Recuperat de: http://www.scielo.org.co/pdf/cide/v8n2/0121-7488-cide-8-02-00009.pdf

Hae.Sang, P; Chi-Hyuck, J. (2009). *A simple and fast algorithm for K-medoids clustering*. Recuperat de: https://isiarticles.com/bundles/Article/pre/pdf/79087.pdf

Hoven, J. (2015) Clustering with optimised weights for gowers metric. Recuperat de: https://beta.vu.nl/nl/Images/stageverslag-hoven_tcm235-777817.pdf

Kassambra, A. (2019). *K-Medoids in R: Algorithm and Practical Examples*. Recuperat de: https://www.datanovia.com/en/lessons/k-medoids-in-r-algorithm-and-practical-examples/

Kaufman, L; Rousseeuw, P. (1987). *Clustering by means of medoids.* Recuperat de: https://www.researchgate.net/profile/Peter_Rousseeuw/publication/243777819_Clustering_by_Means_of_Medoids/links/00b7d531493fad342c000000/Clustering-by-Means-of-Medoids.pdf

Métodos Jerárquicos de Análisis Cluster (s.d). Recuperat de: https://www.ugr.es/~gallardo/pdf/cluster-3.pdf

S. Pavoine, J. Vallet, Anne-Béatrice Dufour, S. Gachet, H. Daniel. (2009) *On the challenge of treating various types of variables: application for improving the measurement of functional diversity*. Recuperat de: http://www.cef-cfr.ca/uploads/Membres/PavoineEtal2009.pdf

Wang, S; Yabes, J; Chang, C. (2019). *Hybrid Density- and Partition-based Clustering Algorithm for Data with Mixed-type Variables*. Recuperat de: https://arxiv.org/pdf/1905.02257.pdf

Wikipedia. *Silhouette (Clustering)*. Recuperat de: https://en.wikipedia.org/wiki/Silhouette_(clustering)

Referències

Sakar, Rk; Meher, Pk; Wahi, Sd; Mohapatra, T; Rao, Ar. (2015). An approach to the development of a core set of gerplasm using a mixture of qualitative and quantitative data. *Plant Genetic Resources: Characterization and Utilization* 13: 96-103.

VIII. ANNEX

En aquest darrer apartat, es poden consultar les diverses comandes realitzades per tal de dur a terme l'estudi. Es troben separades amb la mateixa enumeració que en el treball en qüestió.

```
# III. BASES DE DADES
        ## 3.3 Resums anuals
        load("C:/Users/Usuari/Desktop/TFG/Treball/R/data_20191220.RData") #Imporació de les
        dades
        tiquets2018<- data[data$fAlb >= "2018-01-01" & data$fAlb < "2019-01-01",] #Extracció
        de les dades referents al 2018
        tiquets2018<- tiquets2018[order(tiquets2018$customerId),]
        ### Preu mitjà de les peces i variancia entre els seus preus
        preumig<- c()
        preuvar<- c()
        preu<- c(tiquets2018$precio[1])
        clientID<- c(tiquets2018$customerId[1])
        for (i in 2:nrow(tiquets2018)){
          if(tiquets2018$customerId[i]==tiquets2018$customerId[i-1]){
            preu<- c(preu, tiquets2018$precio[i])
          }else{
            clientID<- c(clientID, tiquets2018$customerId[i])
            preumig<- c(preumig, mean(preu))
            preuvar<- c(preuvar, var(preu))
            preu<- c(tiquets2018$precio[i])
          }}
        preumig<- c(preumig, mean(preu))
        preuvar<- c(preuvar, var(preu))
        ### Preu mitjà de les compres, variància entre el preu d'aquestes
        ### Nombre total de compres, línies de les compres i variància entre línies de
        compra.
        preu_compra<- c(tiquets2018$importe[1])
        preumig_compra<- c()
        preuvar_compra<- c()
        total_compra<- c()
        preutotal_compra<- c()
        linies_compra<- c()
        liniesmig_compra<- c()
        liniesvar_compra<- c()
        preus<- c()
        compra<- 1
        for (i in 2:nrow(tiquets2018)){
          if(tiquets2018$customerId[i]==tiquets2018$customerId[i-1]){
            if(tiquets2018$albId[i]==tiquets2018$albId[i-1]){
              preu_compra<- c(preu_compra, tiquets2018$importe[i])
            }else{
```

```
      preumig_compra<- c(preumig_compra, mean(preu_compra))
      linies_compra<- c(linies_compra, length(preu_compra))
      preus<- c(preus, preu_compra)
      compra<- compra+1
      preu_compra<- c(tiquets2018$importe[i])}
  }else{
    preumig_compra<- c(preumig_compra, mean(preu_compra))
    linies_compra<- c(linies_compra, length(preu_compra))
    preus<- c(preus, preu_compra)
    preu_compra<- c(tiquets2018$importe[i])
    total_compra<- c(total_compra, compra)
    compra<- 1
    preutotal_compra<- c(preutotal_compra, sum(preus))
    preus<- c()
    preuvar_compra<- c(preuvar_compra, var(preumig_compra))
    liniesmig_compra<- c(liniesmig_compra, mean(linies_compra))
    liniesvar_compra<- c(liniesvar_compra, var(linies_compra))
    preumig_compra<- c()
    linies_compra<- c()
  }}
preumig_compra<- c(preumig_compra, mean(preu_compra))
linies_compra<- c(linies_compra, length(preu_compra))
preus<- c(preus, preu_compra)
compra<- compra+1
preu_compra<- c(tiquets2018$importe[i])
total_compra<- c(total_compra, compra)
preutotal_compra<- c(preutotal_compra, sum(preus))
preuvar_compra<- c(preuvar_compra, var(preumig_compra))
liniesmig_compra<- c(liniesmig_compra, mean(linies_compra))
liniesvar_compra<- c(liniesvar_compra, var(linies_compra))
### Creació base de dades resum anual
resum2018<- cbind(clientID, preumig, preuvar, preutotal_compra,
                  preuvar_compra, total_compra, liniesmig_compra,
                  liniesvar_compra)
resum2018<- as.data.frame(resum2018)
resum2018$liniesvar_compra[which(is.na(liniesvar_compra))]<- 0 #Si no existeix
variància és igual a zero
resum2018$preuvar_compra[which(is.na(preuvar_compra))]<- 0
resum2018$preuvar[which(is.na(preuvar))]<- 0
### Tractament
trac<- c(tiquets2018$tratamiento[1])
tractament<- c()
for (i in 2:nrow(tiquets2018)){
  if(tiquets2018$customerId[i]==tiquets2018$customerId[i-1]){
    trac<- c(trac, tiquets2018$tratamiento[i])
  }else{
```

```
    tractament<- c(tractament, max(trac))
    trac<- c(tiquets2018$tratamiento[i])
  }}
tractament<- c(tractament, max(trac))
resum2018<- cbind(resum2018, tractament)
### Mercat
mer<- c(tiquets2018$mercado[1])
mercat<- c()
for (i in 2:nrow(tiquets2018)){
  if(tiquets2018$customerId[i]==tiquets2018$customerId[i-1]){
    mer<- c(mer, tiquets2018$mercado[i])
  }else{
    mercat<- c(mercat, max(mer))
    mer<- c(tiquets2018$mercado[i])
  }}
mercat<- c(mercat, max(mer))
resum2018<- cbind(resum2018, mercat)
### Botiga
diners<- 0
if (tiquets2018$botiga[1]=='S'){
  diners<- tiquets2018$importe[1]
}
diners_totals<- tiquets2018$importe[1]
botiga<- c()
for (i in 2:nrow(tiquets2018)){
  if(tiquets2018$customerId[i]==tiquets2018$customerId[i-1]){
    if (tiquets2018$botiga[i]=='S' && tiquets2018$importe[i]>0){
      diners<- diners + tiquets2018$importe[i]
      diners_totals<- diners_totals + tiquets2018$importe[i]
    }else{
      if (tiquets2018$importe[i]>0){
        diners_totals<- diners_totals + tiquets2018$importe[i]
      }
    }
  }else{
    botiga<- c(botiga, diners/diners_totals)
    if (tiquets2018$botiga[i]=='S' && tiquets2018$importe[i]>0){
      diners<- tiquets2018$importe[i]
      diners_totals<- tiquets2018$importe[i]
    }else{
      if (tiquets2018$importe[i]>0){
        diners<- 0
        diners_totals<- tiquets2018$importe[i]
} } } }
botiga<- c(botiga, diners/diners_totals)
resum2018<- cbind(resum2018, botiga)
```

```
### Enviament
env<- c(tiquets2018$envioId[1])
enviament<- c()
for (i in 2:nrow(tiquets2018)){
  if(tiquets2018$customerId[i]==tiquets2018$customerId[i-1]){
    env<- c(env, tiquets2018$envioId[i])
  } else{
    mitjans<- sum(env==1)
    total<- length(env)
    enviament<- c(enviament, mitjans/total)
    env<- c(tiquets2018$envioId[i])
  }
}
mitjans<- sum(env==1)
total<- length(env)
enviament<- c(enviament, mitjans/total)
resum2018<- cbind(resum2018, enviament)
### Famílies
families2018<- read.csv2("C:/Users/Usuari/Desktop/TFG/Dades/Dades
bones/familia2018.csv")
zeros<-  c(families2018$X1[1],
families2018$X2[1],families2018$X3[1],families2018$X4[1],

families2018$X5[1],families2018$X6[1],families2018$X7[1],families2018$X801[1],

families2018$X802[1],families2018$X9[1],families2018$X1001[1],families2018$X1002[1],
          families2018$X1003[1],families2018$X11[1],families2018$X12[1])
matriu <- c(rep(0,15))
for (i in 2:nrow(families2018)){
  if(tiquets2018$customerId[i]==tiquets2018$customerId[i-1]){
    fam <- c(families2018$X1[i],
families2018$X2[i],families2018$X3[i],families2018$X4[i],

families2018$X5[i],families2018$X6[i],families2018$X7[i],families2018$X801[i],

families2018$X802[i],families2018$X9[i],families2018$X1001[i],families2018$X1002[i],
             families2018$X1003[i],families2018$X11[i],families2018$X12[i])
    zeros<- zeros+fam
  }else{
    matriu<- rbind(matriu, zeros)
    zeros <- c(families2018$X1[i],
families2018$X2[i],families2018$X3[i],families2018$X4[i],

families2018$X5[i],families2018$X6[i],families2018$X7[i],families2018$X801[i],

families2018$X802[i],families2018$X9[i],families2018$X1001[i],families2018$X1002[i],
```

```
              families2018$X1003[i],families2018$X11[i],families2018$X12[i])
}}
matriu<- rbind(matriu,zeros)
matriu<- as.data.frame(matriu)
matriu<- matriu[-1,]
colnames(matriu)<- c("gf1","gf2","gf3","gf4","gf5","gf6","gf7","gf801",
                     "gf802","gf9","gf1001","gf1002","gf1003","gf11","gf12")
gf<- c()
for(i in 1:nrow(matriu)){
  valors<- c(matriu$gf1[i], matriu$gf2[i],matriu$gf3[i],matriu$gf4[i],
             matriu$gf5[i],matriu$gf6[i],matriu$gf7[i],matriu$gf801[i],
             matriu$gf802[i],matriu$gf9[i],matriu$gf1001[i],matriu$gf1002[i],
             matriu$gf1003[i],matriu$gf11[i],matriu$gf12[i])
  gf<- c(gf, colnames(matriu[which.max(valors)]))
}
gf<- as.factor(gf)
resum2018<- cbind(resum2018, gf)
problemes<- c()
gfs<- c(rep(0,15))
for (i in 1:nrow(matriu)){
  valors<- c(matriu$gf1[i], matriu$gf2[i],matriu$gf3[i],matriu$gf4[i],
             matriu$gf5[i],matriu$gf6[i],matriu$gf7[i],matriu$gf801[i],
             matriu$gf802[i],matriu$gf9[i],matriu$gf1001[i],matriu$gf1002[i],
             matriu$gf1003[i],matriu$gf11[i],matriu$gf12[i])
  valors[which(valors<0)]<-0
  gfs<- rbind(gfs,valors)
  if (all(valors==0)==TRUE){
    problemes<- c(problemes, i)
  }
}
gfs<- as.data.frame(gfs)
gfs<- gfs[-1,]
resum2018<- resum2018[-problemes,]
gfs<- gfs[-problemes,]
### IVQ: Índex de Variació Qualitativa
ns<- c()
IVQ<- c()
for (i in 1:nrow(gfs)){
  tot<- c(rep(sum(gfs[i,1:15]),15))
  ns<- gfs[i,1:15]
  p<- sum((ns/tot)^2)
  den<- length(which(gfs[i,]!=0))
  den<- (den-1)/den
  IVQ<- c(IVQ, (1-p)/den)
}
IVQ[which(is.na(IVQ))]<- 0
```

```
resum2018<- cbind(resum2018, IVQ)
### RV: Raó de Variació
RV<- c()
for (i in 1:nrow(gfs)){
  nmoda<- max(gfs[i,1:15])
  valor<- 1-(nmoda/sum(gfs[i,1:15]))
  RV<- c(RV, valor)
}
resum2018<- cbind(resum2018, RV)
### GF's
colnames(gfs)<- c("gf1","gf2","gf3","gf4","gf5","gf6","gf7","gf801",
                  "gf802","gf9","gf1001","gf1002","gf1003","gf11","gf12")
gfs$gf1<- ifelse(gfs$gf1 == 0, 0, 1)
gfs$gf2<- ifelse(gfs$gf2 == 0, 0, 1)
gfs$gf3<- ifelse(gfs$gf3 == 0, 0, 1)
gfs$gf4<- ifelse(gfs$gf4 == 0, 0, 1)
gfs$gf5<- ifelse(gfs$gf5 == 0, 0, 1)
gfs$gf6<- ifelse(gfs$gf6 == 0, 0, 1)
gfs$gf7<- ifelse(gfs$gf7 == 0, 0, 1)
gfs$gf801<- ifelse(gfs$gf801 == 0, 0, 1)
gfs$gf802<- ifelse(gfs$gf802 == 0, 0, 1)
gfs$gf9<- ifelse(gfs$gf9 == 0, 0, 1)
gfs$gf1001<- ifelse(gfs$gf1001 == 0, 0, 1)
gfs$gf1002<- ifelse(gfs$gf1002 == 0, 0, 1)
gfs$gf1003<- ifelse(gfs$gf1003 == 0, 0, 1)
gfs$gf11<- ifelse(gfs$gf11 == 0, 0, 1)
gfs$gf12<- ifelse(gfs$gf12 == 0, 0, 1)
resum2018<- cbind(resum2018, gfs)
### Afegim variables a la base de dades definitiva
#install.packages("dplyr")
library(dplyr)
clients<- select(clientes, -codPos, -vendedorId, -loc, -alta, -customerName, -
codPais)
colnames(clients)[1]<- "clientID"
resum2018<- resum2018%>%
  left_join(clients, by="clientID")
### Definim les variables correctament
resum2018$clientID<- as.character(resum2018$clientID)
resum2018$gf1<- as.factor(resum2018$gf1)
resum2018$gf2<- as.factor(resum2018$gf2)
resum2018$gf3<- as.factor(resum2018$gf3)
resum2018$gf4<- as.factor(resum2018$gf4)
resum2018$gf5<- as.factor(resum2018$gf5)
resum2018$gf6<- as.factor(resum2018$gf6)
resum2018$gf7<- as.factor(resum2018$gf7)
resum2018$gf801<- as.factor(resum2018$gf801)
```

```r
resum2018$gf802<- as.factor(resum2018$gf802)
resum2018$gf9<- as.factor(resum2018$gf9)
resum2018$gf1001<- as.factor(resum2018$gf1001)
resum2018$gf1002<- as.factor(resum2018$gf1002)
resum2018$gf1003<- as.factor(resum2018$gf1003)
resum2018$gf11<- as.factor(resum2018$gf11)
resum2018$gf12<- as.factor(resum2018$gf12)
resum2018$mercat<- as.factor(resum2018$mercat)
resum2018$gf<- as.factor(resum2018$gf)
resum2018$delId<- as.factor(resum2018$delId)
resum2018$sectorId<- as.factor(resum2018$sectorId)
resum2018$prov<- as.factor(resum2018$prov)
resum2018$NIF<- as.factor(resum2018$NIF)
resum2018<- resum2018[which(resum2018$tractament == "S"),] # Ens quedem només amb
clients Silver
resum2018$tractament<- NULL
library("readr")
write_csv(resum2018, path='C:/Users/Usuari/Desktop/TFG/Treball/R/resum_2018.csv')
#resum2018<-read_csv('C:/Users/Usuari/Desktop/TFG/Treball/R/resum_2018.csv')
# IV. ANÀLISI DE CONGLOMERATS
## 4.2 Coeficient de similitud de Gower
library(cluster)
actives<-c(2:31,33) # definim les variables
n <- dim(resum2018)[1] # nombre d'individus
filtro <- c(1:n) # selecció d'individus
dissimMatrix <- daisy(resum2018[filtro, actives], metric = "gower", stand = TRUE)
distMatrix<-dissimMatrix^2
## 4.5 Resultats
### 4.5.1 Mètode de Ward
h1 <- hclust(distMatrix, method = "ward.D2") # classificació
plot(h1, main = "Dendrograma")
k <- 6
c2 <- cutree(h1,k) # Assignem a cada individu la seva classe
table(c2) #class sizes
resum2018[,34] <- c2 #Guardem la classificació a la bdd
colnames(resum2018)[34] <- "Cluster"
silInter = silhouette(c2, dist=dissimMatrix) # Silhouette
print(summary(silInter))
plot(silInter,col=2:7, border=NA)
#### Descriptiva per grups
aggregate(resum2018$preutotal_compra,by=list(resum2018$Cluster),FUN=summary)
#Preutotal_compra
aggregate(resum2018$total_compra,by=list(resum2018$Cluster),FUN=summary) #Total
compra
aggregate(resum2018$botiga,by=list(resum2018$Cluster),FUN=summary) #Percentatge
compra botiga
```

```
aggregate(resum2018$IVQ,by=list(resum2018$Cluster),FUN=summary) #IVQ
aggregate(resum2018$RV,by=list(resum2018$Cluster),FUN=summary) #RV
aggregate(resum2018$enviament,by=list(resum2018$Cluster),FUN=summary) #Enviament
# Mercat
round(prop.table(x=table(resum2018$mercat, resum2018$Cluster), margin=2), 4)
round(prop.table(x=table(resum2018$mercat)), 4)
ggplot(data=resum2018, aes(x = Cluster, fill = mercat) ) +
  geom_bar(aes(fill = as.factor(mercat)), position = "fill") +
  xlab("Clúster") +
  ylab("Freqüència relativa") +
  scale_fill_manual("Mercat", values = alpha( c("#7fc97f","#beaed4",
"#fdc086","#ffff99"),1))
# Grupfam
round(prop.table(x=table(resum2018$gf, resum2018$Cluster), margin=2), 4)
round(prop.table(x=table(resum2018$gf)), 4)
ggplot(data=resum2018, aes(x = Cluster, fill = gf) ) +
  geom_bar(aes(fill = as.factor(gf)), position = "fill") +
  xlab("Clúster") +
  ylab("Freqüència relativa") +
  scale_fill_manual("gf modal", values = alpha(
c("#a6cee3","#1f78b4","#b2df8a","#33a02c","#fb9a99","#e31a1c","#fdbf6f","#ff7f00",

"#cab2d6","#6a3d9a","#ffff99","#b15928"),1))
round(prop.table(x=table(resum2018$gf1, resum2018$Cluster), margin=2), 4) #GF1
round(prop.table(x=table(resum2018$gf1)), 4) #GF1
round(prop.table(x=table(resum2018$gf2, resum2018$Cluster), margin=2), 4) #GF2
round(prop.table(x=table(resum2018$gf2)), 4) #GF2
round(prop.table(x=table(resum2018$gf3, resum2018$Cluster), margin=2), 4) #GF3
round(prop.table(x=table(resum2018$gf3)), 4) #GF3
round(prop.table(x=table(resum2018$gf4, resum2018$Cluster), margin=2), 4) #GF4
round(prop.table(x=table(resum2018$gf4)), 4) #GF4
round(prop.table(x=table(resum2018$gf5, resum2018$Cluster), margin=2), 4) #GF5
round(prop.table(x=table(resum2018$gf5)), 4) #GF5
round(prop.table(x=table(resum2018$gf6, resum2018$Cluster), margin=2), 4) #GF6
round(prop.table(x=table(resum2018$gf6)), 4) #GF6
round(prop.table(x=table(resum2018$gf7, resum2018$Cluster), margin=2), 4) #GF7
round(prop.table(x=table(resum2018$gf7)), 4) #GF7
round(prop.table(x=table(resum2018$gf801, resum2018$Cluster), margin=2), 4) #GF801
round(prop.table(x=table(resum2018$gf801)), 4) #GF801
round(prop.table(x=table(resum2018$gf802, resum2018$Cluster), margin=2), 4) #GF802
round(prop.table(x=table(resum2018$gf802)), 4) #GF802
round(prop.table(x=table(resum2018$gf9, resum2018$Cluster), margin=2), 4) #GF9
round(prop.table(x=table(resum2018$gf9)), 4) #GF9
round(prop.table(x=table(resum2018$gf1001, resum2018$Cluster), margin=2), 4) #GF1001
round(prop.table(x=table(resum2018$gf1001)), 4) #GF1001
round(prop.table(x=table(resum2018$gf1002, resum2018$Cluster), margin=2), 4) #GF1002
```

```r
round(prop.table(x=table(resum2018$gf1002)), 4) #GF1002
round(prop.table(x=table(resum2018$gf1003, resum2018$Cluster), margin=2), 4) #GF1003
round(prop.table(x=table(resum2018$gf1003)), 4) #GF1003
round(prop.table(x=table(resum2018$gf11, resum2018$Cluster), margin=2), 4) #GF11
round(prop.table(x=table(resum2018$gf11)), 4) #GF11
round(prop.table(x=table(resum2018$gf12, resum2018$Cluster), margin=2), 4) #GF12
round(prop.table(x=table(resum2018$gf12)), 4) #GF12
round(prop.table(x=table(resum2018$delId, resum2018$Cluster), margin=2),4) #DelId
round(prop.table(x=table(resum2018$delId)), 4) #DelId
# NIF
round(prop.table(x=table(resum2018$NIF, resum2018$Cluster), margin = 2), 4)
round(prop.table(x=table(resum2018$NIF)), 4)
ggplot(data=resum2018, aes(x = Cluster, fill = NIF) ) +
  geom_bar(aes(fill = as.factor(NIF)), position = "fill") +
  xlab("Clúster") +
  ylab("Freqüència relativa") +
  scale_fill_manual("NIF", values = alpha( c("#7fc97f","#beaed4"),1))
 binom.test(x=2270, n=2270+734, p=0.737)
 binom.test(x=606, n=606+304, p=0.737)
 binom.test(x=211, n=211+135, p=0.737)
 binom.test(x=327, n=327+155, p=0.737)
 binom.test(x=275, n=275+74, p=0.737)
 binom.test(x=643, n=643+144, p=0.737)
### 4.5.2 Mètode PAM
#### Càlcul de pesos Sihouette en funció de les possibles k's
sil_width <- c(NA)
for(i in 2:14){
  pam_fit <- pam(distMatrix,
                 diss = TRUE,
                 k = i)
  sil_width[i] <- pam_fit$silinfo$avg.width
}
plot(1:14, sil_width,
     xlab = "Nombre de clústers",
     ylab = "Silhouette")
lines(1:12, sil_width, col="red")
#### Creació clústers
PAM <- pam(x=distMatrix,k=6,diss=T)
resum2018$Cluster<- PAM$clustering
print(table(PAM$clustering))
#### Descriptiva per grups
aggregate(resum2018$preutotal_compra,by=list(resum2018$Cluster),FUN=summary)
#Preutotal_compra
aggregate(resum2018$total_compra,by=list(resum2018$Cluster),FUN=summary) #Total
compra
```

```r
aggregate(resum2018$botiga,by=list(resum2018$Cluster),FUN=summary) #Percentatge
compra botiga
aggregate(resum2018$IVQ,by=list(resum2018$Cluster),FUN=summary) #IVQ
aggregate(resum2018$RV,by=list(resum2018$Cluster),FUN=summary) #RV
aggregate(resum2018$enviament,by=list(resum2018$Cluster),FUN=summary) #Enviament
# Mercat
round(prop.table(x=table(resum2018$mercat, resum2018$Cluster), margin=2), 4)
round(prop.table(x=table(resum2018$mercat)), 4)
ggplot(data=resum2018, aes(x = Cluster, fill = mercat) ) +
  geom_bar(aes(fill = as.factor(mercat)), position = "fill") +
  xlab("Clúster") +
  ylab("Freqüència relativa") +
  scale_fill_manual("Mercat", values = alpha( c("#7fc97f","#beaed4",
"#fdc086","#ffff99"),1))
# Grupfam
round(prop.table(x=table(resum2018$gf, resum2018$Cluster), margin=2), 4)
round(prop.table(x=table(resum2018$gf)), 4)
ggplot(data=resum2018, aes(x = Cluster, fill = gf) ) +
  geom_bar(aes(fill = as.factor(gf)), position = "fill") +
  xlab("Clúster") +
  ylab("Freqüència relativa") +
  scale_fill_manual("gf modal", values = alpha(
c("#a6cee3","#1f78b4","#b2df8a","#33a02c","#fb9a99","#e31a1c","#fdbf6f","#ff7f00",

"#cab2d6","#6a3d9a","#ffff99","#b15928"),1))
round(prop.table(x=table(resum2018$gf1, resum2018$Cluster), margin=2), 4) #GF1
round(prop.table(x=table(resum2018$gf1)), 4) #GF1
round(prop.table(x=table(resum2018$gf2, resum2018$Cluster), margin=2), 4) #GF2
round(prop.table(x=table(resum2018$gf2)), 4) #GF2
round(prop.table(x=table(resum2018$gf3, resum2018$Cluster), margin=2), 4) #GF3
round(prop.table(x=table(resum2018$gf3)), 4) #GF3
round(prop.table(x=table(resum2018$gf4, resum2018$Cluster), margin=2), 4) #GF4
round(prop.table(x=table(resum2018$gf4)), 4) #GF4
round(prop.table(x=table(resum2018$gf5, resum2018$Cluster), margin=2), 4) #GF5
round(prop.table(x=table(resum2018$gf5)), 4) #GF5
round(prop.table(x=table(resum2018$gf6, resum2018$Cluster), margin=2), 4) #GF6
round(prop.table(x=table(resum2018$gf6)), 4) #GF6
round(prop.table(x=table(resum2018$gf7, resum2018$Cluster), margin=2), 4) #GF7
round(prop.table(x=table(resum2018$gf7)), 4) #GF7
round(prop.table(x=table(resum2018$gf801, resum2018$Cluster), margin=2), 4) #GF801
round(prop.table(x=table(resum2018$gf801)), 4) #GF801
round(prop.table(x=table(resum2018$gf802, resum2018$Cluster), margin=2), 4) #GF802
round(prop.table(x=table(resum2018$gf802)), 4) #GF802
round(prop.table(x=table(resum2018$gf9, resum2018$Cluster), margin=2), 4) #GF9
round(prop.table(x=table(resum2018$gf9)), 4) #GF9
round(prop.table(x=table(resum2018$gf1001, resum2018$Cluster), margin=2), 4) #GF1001
```

```
round(prop.table(x=table(resum2018$gf1001)), 4) #GF1001
round(prop.table(x=table(resum2018$gf1002, resum2018$Cluster), margin=2), 4) #GF1002
round(prop.table(x=table(resum2018$gf1002)), 4) #GF1002
round(prop.table(x=table(resum2018$gf1003, resum2018$Cluster), margin=2), 4) #GF1003
round(prop.table(x=table(resum2018$gf1003)), 4) #GF1003
round(prop.table(x=table(resum2018$gf11, resum2018$Cluster), margin=2), 4) #GF11
round(prop.table(x=table(resum2018$gf11)), 4) #GF11
round(prop.table(x=table(resum2018$gf12, resum2018$Cluster), margin=2), 4) #GF12
round(prop.table(x=table(resum2018$gf12)), 4) #GF12
round(prop.table(x=table(resum2018$delId, resum2018$Cluster), margin=2),4) #DelId
round(prop.table(x=table(resum2018$delId)), 4) #DelId
# NIF
round(prop.table(x=table(resum2018$NIF, resum2018$Cluster), margin = 2), 4)
round(prop.table(x=table(resum2018$NIF)), 4)
ggplot(data=resum2018, aes(x = Cluster, fill = NIF) ) +
  geom_bar(aes(fill = as.factor(NIF)), position = "fill") +
  xlab("Clúster") +
  ylab("Freqüència relativa") +
  scale_fill_manual("NIF", values = alpha( c("#7fc97f","#beaed4"),1))
```

```
# V. ANÀLISI DE SENSIBILITAT DE LA DISTÀNCIA DE GOWER
```

```
Simulacions<-read.csv2('C:/Users/Usuari/Desktop/TFG/Treball/R/simulacions.csv')
set.seed(123)
Var_2<- sample(1:2, 92, replace=T)
Var_6<-sample(1:6, 92, replace=T)
Simulacions<- cbind(Simulacions[,1:3], Var_2, Var_6)
colnames(Simulacions)<- c("Sexe", "Alçada", "Pes", "Var_2","Var_6")
Simulacions[,4]<- as.factor(Simulacions[,4])
Simulacions[,5]<- as.factor(Simulacions[,5])
library(ggplot2)
ggplot(Simulacions, aes(Pes, Alçada)) + geom_point()+ geom_point(aes(colour =
factor(Sexe)))
### Simulacions possibles
library(cluster)
actives<-c(1:3)
actives_2<- c(1:4)
actives_6<- c(1:3,5)
n <- dim(Simulacions)[1] # nombre d'individus
filtro <- c(1:n) # selecció d'individus
### Creació dels clústers
dissimMatrix1 <- daisy(Simulacions[filtro, actives], metric = "gower", stand = TRUE)
```

```
dissimMatrix2 <- daisy(Simulacions[filtro, actives_2], metric = "gower", stand =
TRUE)
dissimMatrix6 <- daisy(Simulacions[filtro, actives_6], metric = "gower", stand =
TRUE)
distMatrix1<-dissimMatrix1^2
distMatrix2<-dissimMatrix2^2
distMatrix6<-dissimMatrix6^2
h1 <- hclust(distMatrix1, method = "ward.D2") # classificació
h2 <- hclust(distMatrix2, method = "ward.D2") # classificació
h6 <- hclust(distMatrix6, method = "ward.D2") # classificació
plot(h1, main = "Dendrograma amb les variables originals")
plot(h2, main = "Dendrograma amb la variable aleatòria a dos nivells")
plot(h6, main = "Dendrograma amb la variable aleatòria a sis nivells")
## Assignació de clústers
k<-4
c <- cutree(h1,k) # Assignem a cada individu la seva classe
c2<- cutree(h2,k)
c6<- cutree(h6,k)
table(c6)
Simulacions[,6] <- c
Simulacions[,7] <- c2
Simulacions[,8] <- c6
colnames(Simulacions)[6] <- "Cluster"
colnames(Simulacions)[7] <- "Cluster_2"
colnames(Simulacions)[8] <- "Cluster_6"
barplot(table(Simulacions$Sexe, Simulacions$Cluster))
round(prop.table(x=table(Simulacions$Sexe, Simulacions$Cluster_6)), 4)
aggregate(Simulacions$Alçada,by=list(Simulacions$Cluster),FUN=mean)
aggregate(Simulacions$Pes,by=list(Simulacions$Cluster),FUN=mean)
barplot(table(Simulacions$Sexe, Simulacions$Cluster_2))
barplot(table(Simulacions$Sexe, Simulacions$Cluster_6))
barplot(table(Simulacions$Var_2, Simulacions$Cluster_2))
round(prop.table(x=table(Simulacions$Var_6, Simulacions$Cluster_6)), 4)
aggregate(Simulacions$Alçada,by=list(Simulacions$Cluster_2),FUN=mean)
aggregate(Simulacions$Pes,by=list(Simulacions$Cluster_2),FUN=mean)
barplot(table(Simulacions$Var_6, Simulacions$Cluster_6))
aggregate(Simulacions$Alçada,by=list(Simulacions$Cluster_6),FUN=mean)
aggregate(Simulacions$Pes,by=list(Simulacions$Cluster_6),FUN=mean)
```