TREBALL DE FI DE GRAU

Grau en Estadística

Títol: Estudi de classificació a partir de tècniques discriminants

Autor: Silvia Mena Guix

Director: Mireia Besalú Mayol

Departament: Genètica, microbiologia i estadística

Convocatòria: Juny de 2020

UNIVERSITAT DE BARCELONA

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

AGRAÏMENTS

En primer lloc, vull agrair a totes aquelles persones que han col·laborat en la realització d'aquest projecte, ja m'hagin ajudat a adquirir capacitats i coneixements, com ajudant-me donant-me suport.

M'agradaria començar agraint el paper imprescindible de la Mireia Besalú com a directora d'aquest estudi. Per la seva paciència, orientació, implicació i suport durant el desenvolupament de l'estudi.
També vull agrair a la Manuela Alcañiz, per la seva implicació i els ànims que m'ha donat sempre.

En l'àmbit personal, vull agrair als meus pares tot el suport que m'han donat durant els anys de carrera. Sense la seva ajuda, i sobretot confiança no hauria arribat fins aquí. Moltes gràcies per tot l'amor i suport incondicional. I també la interminable paciència.
També als meus amics, per estar sempre disposats a ajudar-me i treure'm un somriure en els moments més difícils.

Sense la vostra ajuda aquest projecte no hauria estat possible.
Moltes gràcies!

RESUM

La necessitat d'identificar característiques que ens permetin diferenciar dos o més grups cada cop és més freqüent. L'anàlisi discriminant consisteix a analitzar aquestes característiques que ajuden a la classificació de dos o més grups. L'objectiu d'aquest estudi és arribar a una bona classificació d'una base de dades que té com a observacions les reserves d'un complex hoteler diferenciat en tres seccions diferents. Aquestes tres seccions de l'hotel seran els nostres grups de la variable dependent. Els mètodes de classificació amb els quals volem aconseguir una assignació correcta de reserves noves són l'anàlisi discriminant lineal, l'anàlisi discriminant quadràtic i l'anàlisi k-nn veïns més pròxims. Seguidament validarem aquests resultats estudiant la capacitat predictiva. D'aquesta forma arribarem a veure amb quin mètode obtenim la taxa d'encerts de classificació més elevada.

Paraules Clau

Classificació, discriminant lineal, Discriminant quadràtic, k-nn veïns més pròxims, reserves, capacitat predictiva, taxa d'encerts.

ABSTRACT

The necessity of identifying characteristics that allow us to distinguish two or more groups of individuals is becoming more common. The discriminant analysis consists in analyze these characteristics that you can use at the time of classifying in two or more groups. The goal of this study is to get a good classification of a data base that contain observations of reservations of a hotel complex differentiated in three different sections. These sections will be the groups of our dependent variable. The classification methods with which we want to achieve the classification are linear discriminant analysis, quadratic discriminant analysis and k-nn nearest neighboring analysis. Then, we will validate these results by studying the predictive ability of those methods. In this way, we will be able to see with which method we obtain the highest success rate of classification.

Key words

Classification, linear discriminant, quadratic discriminant, k-nn nearest neighboring, reservations, predictive ability, success rate.

Classificació AMS

62-07 Data analysis

62P20 Applications to economics

62H30 Classification and discrimination; cluster analysis

97K40 Descriptive statistics

62J10 Analysis of variance and covariance

ÍNDEX

I. INTRODUCCIÓ

El volum d'informació creix dia a dia de forma exponencial, i amb aquest creixement cada cop es genera més la necessitat de poder segmentar dades per poder treballar-les de la manera més personalitzada possible, d'aquesta manera cada cop és més important conèixer quines són les característiques que diferencien grups de dades, per poder realitzar futures prediccions.

En aquest estudi es fa un l'anàlisi d'una base de dades que té com a observacions les reserves de tot un any d'un *Resort.* Les variables que trobem en aquesta base de dades ens donen informació tant sobre la pròpia reserva com sobre el client que ha realitzat la reserva. A més, les reserves poden ser per una de les tres seccions diferents de l'hotel. En aquest estudi es buscarà la millor classificació de les reserves en cada secció a partir de les altres variables associades a la reserva, com per exemple, l'edat dels hostes, l'antelació de compra, el dia de compra o el dia d'entrada a l'hotel.

Conèixer aquesta informació té un gran valor, ja que actualment el comerç digital va prenent més presència, i el màrqueting digital acompanya directament al creixement d'aquest. Un bon màrqueting digital contempla l'impacte als usuaris amb publicitat on-line buscant trobar un públic més afí i un impacte més personalitzat per assegurar-ne la compra. D'aquesta manera, a partir d'aquesta classificació podríem conèixer millor algunes característiques dels clients de cada hotel i altres sobre el seu comportament. Donat un tipus de client potencial, a partir de les seves característiques, podríem associar-li publicitat de la secció d'hotel que millor se li ajusti.

Aquest estudi té l'objectiu principal d'aconseguir un model predictiu que permeti classificar nous individus al grup de pertinència. D'aquesta manera, i coneixent dades associades a clients i les seves respectives reserves, podríem determinar quin hotel serà més afí a cada hoste mitjançant una regla de decisió que assigni una reserva nova a un dels hotels.

El mètode estadístic que utilitzarem per classificar les reserves en cada hotel és l'anàlisi discriminant. A partir d'aquest mètode podrem assignar noves observacions a la secció d'hotel més afí, calculada anteriorment amb reserves de les quals ja en coneixem l'hotel.

Posteriorment procedirem a avaluar l'exactitud de la classificació mitjançant taules creuades on es compararà la classificació real de cada individu a cada hotel al grup amb pronòstic que es plantejarà amb cada mètode, per conèixer quin és el grau d'encert de classificació de les reserves als hotels.

Utilitzarem diferents mètodes per dur a terme aquesta classificació, així que un altre objectiu de l'anàlisi serà la comparació entre diferents tipus de mètodes multivariants utilitzats i la valoració sobre quin mètode és millor aplicar a les nostres dades.

II. METODOLOGIA

L'anàlisi multivariant és un conjunt de mètodes estadístics utilitzats per a determinar la contribució de varis factors en un sol esdeveniment o resultat. Aquestes anàlisis estudien, analitzen, representen i interpreten dades que resulten de l'observació de més d'una variable sobre una mostra d'individus.

Segons Maurice Kendall, estadístic britànic molt influent en l'estadística durant els anys 70, l'anàlisi multivariant es classifica segons tres blocs:

- Si les variables poden ser classificades o no en dependents o independents, és a dir, si s'assumeix existència o no de relacions causals entre variables. Aquest es poden anomenar mètodes de dependència i alguns exemples en serien: anàlisis de regressió, MANOVA, anàlisi discriminant...
- Si hi ha un model de dependència, la determinació de quantes variables dependents han sigut incloses a l'anàlisi. S'anomenen mètodes d'interdependència i dos exemples en són l'anàlisi de components principals i l'anàlisi de correspondències.
- Com s'han mesurat les variables, distingint l'efecte entre variables quantitatives i variables qualitatives i com estan relacionades les variables dels grups entre si. Aquests són els models estructurals. Un tipus de model estructural seria l'anàlisi factorial.

A partir d'aquesta classificació, a cada problema li correspondria un determinat tipus d'anàlisi, complint tota una sèrie de fases necessàries per tal d'aconseguir una realització correcte d'un anàlisi multivariant

- Primera fase: definició del problema de la investigació, definició dels objectius i elecció d'una tècnica multivariant convenient.
- Segona fase: desenvolupament del projecte de l'anàlisi posant a la pràctica la tècnica multivariant seleccionada.
- Tercera fase: recollida de dades i revisió del compliment de condicions bàsiques com són la normalitat de les dades, la linealitat, correlació... depenent de la tècnica que s'utilitzarà.
- Quarta fase: estimació del model i capacitat predictiva del model (amb la significació dels paràmetres utilitzats).
- Cinquena fase: interpretació dels valors obtinguts en el model estimat. En el cas que la interpretació no sigui coherent, es realitzarà una segona estimació del model.
- Sisena fase: valoració del model mitjançant la comprovació dels resultats amb la major fiabilitat possible a través dels contrastos específics.

En el cas del nostre estudi, ens centrarem en una de les principals tècniques multivariants: l'anàlisi discriminant. Aquesta tècnica és utilitzada per classificar diferents individus en grups a partir d'un conjunt de variables estudiades, tenint en compte que cada individu només pot pertànyer a un grup.
Aquesta anàlisi es pot considerar com una anàlisi de regressió on la variable dependent és categòrica i té com a valors les etiquetes de cadascun dels grups. Les variables independents són quantitatives i són les que determinen a quin grup pertanyen els individus.

L'objectiu de l'anàlisi és construir una regla de decisió que pugui assignar un individu nou, del qual no sabem a quina categoria pertany ja que no s'ha pogut classificar anteriorment, a un dels grups prefixats. Assumint que un conjunt de casos d'estudi ja estan classificats en una sèrie de grups, és a dir, se sap a quin grup pertanyen, l'anàlisi discriminant té l'objectiu de trobar una combinació de les variables independents que permeti diferenciar els grups. Un cop trobada aquesta combinació, mitjançant una funció discriminant, s'utilitzaria per classificar nous casos de l'estudi. Per tant, l'anàlisi discriminant ens permetrà estudiar, doncs, diferències entre dos o més grups definits a priori.

Per a la realització d'aquesta tècnica és necessària la consideració d'una sèrie de restriccions:

- Es té una variable categòrica que actua com a variable dependent i la resta de variables són independents respecte d'ella.
- És necessari que existeixin almenys dos grups per la variable categòrica.
- Per cada grup de la variable categòrica es necessiten dos o més casos.
- El nombre de variables discriminants ha de ser menor que el nombre d'individus menys 2: $x_1, ..., x_p$, on $p < (n - 2)$ i n és el número d'individus.
- Cap variable discriminant pot ser combinació lineal de les altres variables discriminants.
- En el cas de més de dos grups, el nombre màxim de funcions discriminants és igual al mínim entre el nombre de variables i el nombre de grups menys 1 (amb q grups i n variables tindrem, mín(n;q-1) funcions discriminants).
- Les matrius de variàncies i covariàncies de les variables independents dins de cada grup han de ser aproximadament iguals
- Les variables contínues han de seguir una distribució normal multivariant.

A continuació, i en primer lloc, plantejarem els diferents discriminants de classificació de dues poblacions. Posteriorment, plantejarem el raonament pel cas de $k \geq 3$ grups.

1. Classificació en dues poblacions

Siguin Ω_1 i Ω_2 dues poblacions i $X_1,\ldots, X_p$ variables observables, i indicant $x = (x_1,\ldots. x_p)$ les observacions sobre un individu ω. Considerant que es coneix com a regla discriminant un criteri que permet assignar ω a una de les poblacions considerades Ω_1 i Ω_2 en funció de les observacions $(x_1,\ldots,x_p)$, s'expressa aquest criteri mitjançant una funció discriminant $D(x_1,\ldots, x_p)$ amb la regla de classificació següent:

$$\begin{cases} \text{Si } D(x_1,\ldots, x_p) > 0 & \text{s'assigna } \omega \text{ a } \Omega_1 \\ \text{Altrament} & \text{s'assigna } \omega \text{ a } \Omega_2 \end{cases}$$

Així, una regla discriminant equival a fer una partició de l'espai mostral en dues regions diferenciades:

$$R_1 = \{ x \mid D(x) > 0 \} \text{ i } R_2 = \{ x \mid D(x) \leq 0 \}$$

Per poder discriminar les observacions mitjançant una funció discriminant, existeixen diferents funcions matemàtiques basades en diferents criteris.

A continuació, presentarem el discriminador lineal, la regla de la màxima versemblança, la regla de Bayes i el discriminant quadràtic. En un primer pas explicarem aquests criteris pel cas de classificació de dues poblacions.

1.1 Discriminador lineal

Siguin μ_1 i μ_2 els vectors de mitjanes de les variables $X_1,\ldots,X_p$ en Ω_1 i Ω_2, i suposant que la matriu de covariàncies Σ és la mateixa en les dues poblacions. Les distàncies de Mahalanobis de les observacions $x = (x_1,\ldots, x_p)$ d'un individu ω a les poblacions són:

$$M^2(x, \mu_i) = (x - \mu_i)' \Sigma^{-1} (x - \mu_i), \qquad i = 1, 2$$

El criteri d'assignació que veiem a continuació consisteix a assignar l'individu ω a la població més pròxima:

$$\begin{cases} M^2(x, \mu_1) < M^2(x, \mu_2) & \text{s'assigna } \omega \text{ a } \Omega_1, \\ \text{Altrament} & \text{s'assigna } \omega \text{ a } \Omega_2. \end{cases}$$

A partir de la diferència entre $M^2(x, \mu_2) - M^2(x, \mu_1)$ es construeix la funció discriminant:

$$L(x) = \left[x - \frac{1}{2}(\mu_1 - \mu_2)\right]' \Sigma^{-1}(\mu_1 - \mu_2)$$

Aquesta funció discriminant és anomenada **discriminador lineal de Fisher** i el criteri és el següent:

$$\begin{cases} L(x) > 0 & \text{s'assigna } \omega \text{ a } \Omega_1, \\ \text{Altrament} & \text{s'assigna } \omega \text{ a } \Omega_2. \end{cases}$$

1.2 Regla de la màxima versemblança

Suposem que coneixem $f_1(x)$, $f_2(x)$ que són les densitats del vector d'interès X en Ω_1 i Ω_2 respectivament.
El criteri d'assignació de ω serà a la població on la versemblança de les observacions x sigui major:

$$\begin{cases} f_1(x) > f_2(x) & \text{s'assigna } \omega \text{ a } \Omega_1, \\ \text{Altrament} & \text{s'assigna } \omega \text{ a } \Omega_2. \end{cases}$$

La funció discriminant es defineix com:

$$V(x) = log\, f_1(x) - log\, f_2(x)$$

Aquesta funció discriminant és anomenada **discriminador de la màxima versemblança** i el criteri és el següent:

$$\begin{cases} V(x) > 0 & \text{s'assigna } \omega \text{ a } \Omega_1, \\ \text{Altrament} & \text{s'assigna } \omega \text{ a } \Omega_2. \end{cases}$$

1.3 Regla de Bayes

El teorema de Bayes és útil per obtenir probabilitats condicionals, ja que expressa la probabilitat condicional d'un esdeveniment aleatori A_j donat que un altre esdeveniment B ja és un fet.

$$P(A_j \mid B) = \frac{P(A_j \cap B)}{P(B)} = \frac{P(A_j) \cdot P(B \mid A_j)}{\sum_{i=1}^{n} P(A_i) \cdot P(B \mid A_i)}$$

On: **$P(A_j \mid B)$** = probabilitat a posteriori (*posterior probability*)

$P(A_j)$ = probabilitat a priori (*prior probability*)

Aplicat a l'anàlisi discriminant s'estaria calculant la probabilitat que la variable resposta Y pertanyi a cada un dels possibles nivells, donats uns determinats valors de les variables predictores.

Suposem que coneixem les probabilitats a priori[1], $q_1 = P(\Omega_1)$ i $q_2 = P(\Omega_2)$. Aquestes probabilitats representen la probabilitat que una observació aleatòria ω pertanyi a la classe 1 o 2 de la variable resposta. Buscarem conèixer les probabilitats a posteriori de que es produeixi cada succés:

$$P(\Omega_i \mid x) = \frac{q_i f_i(x)}{q_1 f_1(x) + q_2 f_2(x)}, \qquad i = 1,2$$

La classificació de cada individu ω seguirà la condició de classificar-se dins del nivell que tingui la probabilitat $P(\Omega_i \mid x)$ més elevada.

El criteri d'assignació de ω serà a la població on amb una probabilitat a posteriori més gran de les observacions x:

$$\begin{cases} P(\Omega_1 \mid x) > P(\Omega_2 \mid x) & \text{s'assigna } \omega \text{ a } \Omega_1, \\ \text{Altrament} & \text{s'assigna } \omega \text{ a } \Omega_2. \end{cases}$$

La funció discriminant de Bayes es defineix com:

$$B(x) = log\, f_1(x) - log\, f_2(x) + log\,(q_1 / q_2)$$

Podem indicar-ho mitjançant la següent **regla de classificació de Bayes**:

$$\begin{cases} B(x) > 0 & \text{s'assigna } \omega \text{ a } \Omega_1, \\ \text{Altrament} & \text{s'assigna } \omega \text{ a } \Omega_2. \end{cases}$$

[1] Són les probabilitats que una observació pertanyi a la classe k i es correspon amb nombre d'observacions d'aquella classe entre el nombre total d'observacions $q_k = n_k / N$.

Recordant la funció discriminant de la màxima versemblança, observem que l'única diferència amb la funció discriminant de Bayes és la constant $log\ (q_1/\ q_2\ )$, per tant podem dir que

$$B\ (x)\ =\ V\ (x)\ +\ log\ (q_1/\ q_2\ )$$

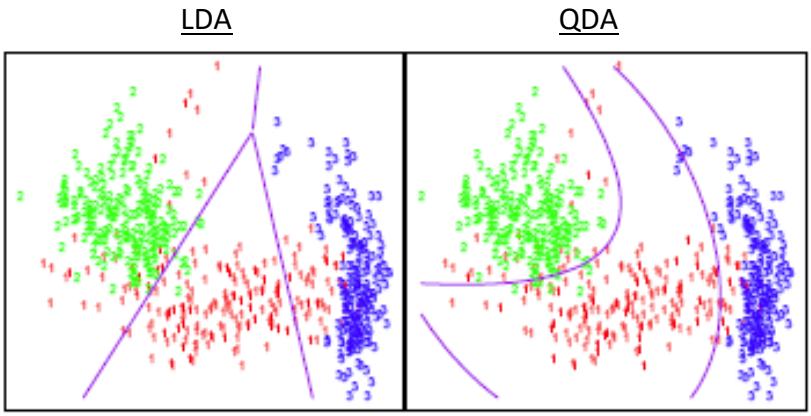
En el cas particular que $q_1\ =\ q_2\ =$ ½, llavors $B\ (x)\ =\ V\ (x)$.

1.4 Discriminador quadràtic

L'anàlisi discriminant lineal (LDA) assumeix que la covariància de les variables de predicció són comunes als dos grups de la variable de resposta. L'anàlisi discriminant quadràtic (QDA) proporciona un enfocament alternatiu que assumeix que cada grup té la seva pròpia matriu de covariància. És a dir, no suposa que les variables de predicció tinguin una variació comuna entre els dos grups en la variable dependent.

Considerem la imatge següent. Al intentar classificar les observacions en els tres grups (marcades en diferents colors), LDA proporciona límits de decisió lineals basats ens els supòsits que les observacions varien de forma constant entre tots els grups i quan observem les dades veiem que la variabilitat de les observacions dins de cada grup és diferent. En canvi, QDA és capaç de capturar les covariàncies diferents i proporcionar límits de decisió de classificació no lineals més precisos.

Gràfic 2.1. Comparació entre LDA i QDA



An Introducion to Statistical Learning, Gareth James

2. Classificació de poblacions normals

Fins ara hem observat com es desenvolupen els diferents criteris discriminants sense fixar la distribució que segueixen les dades. En aquest apartat observarem com queden aquests mateixos criteris en el cas de que les dades segueixin una distribució normal. A més, coneixem que la normalitat multivariant és un suposat bàsic de l'anàlisi discriminant que ens ajudarà a aconseguir un resultat més òptim.

Per a l'estimació de $f_i(x)$ haurem d'assumir que es compleixen algunes hipòtesis, en aquest apartat tractarem la normalitat i l'homoscedasticitat.

Suposem a partir d'ara que les poblacions són normals, és a dir, que la distribució de $X=(X_1,\dots, X_p)$ és en Ω_1 és $N_p(\mu_1,\sum_1)$ i en Ω_2 és $N_p(\mu_2,\sum_2)$ i, per tant la seva funció de densitat és per i=1,2

$$f_i(x) = (2\pi)^{-p/2} \left|\Sigma_i^{-1}\right|^{1/2} exp\left\{-\frac{1}{2}(x - \mu_i)'\Sigma_i^{-1}(x - \mu_i)\right\}$$

A continuació, observem com queden els criteris de discriminació suposant que les poblacions són normals.

2.1 Discriminador lineal i Màxima versemblança

Si suposem que $\mu_1 \neq \mu_2$ i $\sum_1 = \sum_2 = \sum$. Si les dues poblacions són normals el discriminador lineal i el de màxima versemblança coincideixen:

$$V(x) = -\frac{1}{2}(x - \mu_1)'\Sigma^{-1}(x - \mu_1) + -\frac{1}{2}(x - \mu_2)'\Sigma^{-1}(x - \mu_2) = L(x)$$

A continuació, comprovem la igualtat $V(x) = L(x)$.
Sabem que:

$$L(x) = \left[x - \tfrac{1}{2}(\mu_1 - \mu_2)\right]'\Sigma^{-1}(\mu_1 - \mu_2) \quad \text{i} \quad V(x) = log\, f_1(x) - log\, f_2(x)$$

Agafem $V(x) = log\left(\frac{f_1(x)}{f_2(x)}\right)$

Sabent que $f_i(x) = (2\pi)^{-p/2} \left|\Sigma_i^{-1}\right|^{1/2} exp\left\{-\frac{1}{2}(x - \mu_i)'\Sigma_i^{-1}(x - \mu_i)\right\}$ per a i=1,2 i $\sum_1 = \sum_2 = \sum$.

Calculem:

$$\frac{f_1(x)}{f_2(x)} = \frac{(2\pi)^{-p/2}|\Sigma^{-1}|^{1/2}\, exp\left\{-\frac{1}{2}(x-\mu_1)'\Sigma^{-1}(x-\mu_1)\right\}}{(2\pi)^{-p/2}|\Sigma^{-1}|^{1/2}\, exp\left\{-\frac{1}{2}(x-\mu_2)'\Sigma^{-1}(x-\mu_2)\right\}} =$$

$$exp\left\{-\frac{1}{2}(x-\mu_1)'\,\Sigma^{-1}(x-\mu_1) + \frac{1}{2}(x-\mu_2)'\,\Sigma^{-1}(x-\mu_2)\right\}$$

Per tant,

$V(X) = \frac{1}{2}\{(x-\mu_2)'\,\Sigma^{-1}(x-\mu_2)$ - $(x-\mu_1)'\,\Sigma^{-1}(x-\mu_1)\}$ =

$\frac{1}{2}(x'\,\Sigma^{-1}x - x'\,\Sigma^{-1}\mu_2' - \mu_2'\,\Sigma^{-1}x + \mu_2'\,\Sigma^{-1}\mu_2 - x'\,\Sigma^{-1}x - x'\,\Sigma^{-1}\mu_1' - \mu_1'\,\Sigma^{-1}x +$

$\mu_1'\,\Sigma^{-1}\mu_1) = \frac{1}{2}(-x'\,\Sigma^{-1}\mu_2' - \mu_2'\,\Sigma^{-1}x + \mu_2'\,\Sigma^{-1}\mu_2 - x'\,\Sigma^{-1}\mu_1' - \mu_1'\,\Sigma^{-1}x +$

$\mu_1'\,\Sigma^{-1}\mu_1) =$

$\frac{1}{2}\,(\,2x - (\mu_2-\mu_1))'\,\Sigma^{-1}(\mu_1-\mu_2)) = (x-\frac{1}{2}(\mu_1-\mu_2))'\,\Sigma^{-1}(\mu_1-\mu_2)$

Per tant, podem veure que el discriminador lineal coincideix amb els discriminadors de màxima versemblança.

2.2 Regla de Bayes

Seguim suposant que:

- Les observacions de cada grup segueixen una distribució normal multivariant
- $\mu_1 \neq \mu_2$
- $\sum_1 = \sum_2 = \sum$
- Coneixem les probabilitats a priori: q_1= P(Ω_1), q_2= P(Ω_2), tals que $q_1 + q_2 = 1$.

Aleshores podem dir que la funció discriminant de Bayes, igual que en el cas general, coincideix amb la funció discriminant de la regla de màxima versemblança ($V(x)$) més la constant $log\,(q_1/\,q_2)$. Però a més, en aquest cas com que $V(x)$ coincideix amb el discriminador lineal tenim que

$$B(x) = V(x) + log\,(q_1/\,q_2) = L(x) + log\,(q_1/\,q_2)$$

2.3 Discriminador quadràtic

Al igual que la resta de criteris, en aquest cas, el discriminador quadràtic també assumeix que les observacions de cada grup segueixen una distribució normal multivariant però ara suposem que $\mu_1 \neq \mu_2$ i $\sum_1 \neq \sum_2$.
Aleshores, el criteri de màxima versemblança proporciona el **discriminador quadràtic**

$$Q(x) = \frac{1}{2}x'(\Sigma_2^{-1} - \Sigma_1^{-1})x + x'(\Sigma_1^{-1}\mu_1 - \Sigma_2^{-1}\mu_2) + \frac{1}{2}\mu_2{}'\Sigma_2^{-1}\mu_2 - \frac{1}{2}\mu_1{}'\Sigma_1^{-1}\mu_1 + \frac{1}{2}log|\Sigma_2| - \frac{1}{2}log|\Sigma_1|$$

3. Classificació en el cas de més de 2 poblacions

En aquest apartat suposem que l'individu ω pot provenir de k poblacions $\Omega_1,\ldots,\Omega_k$ on $k \geq 3$. Buscarem establir una regla que ens permeti assignar ω a una de les k poblacions possibles a partir de les observacions $x = (x_1,\ldots, x_p)'$ de *p* variables $X = (X_1,\ldots, X_p)$.

3.1 Discriminador lineal

Sigui μ_i la mitjana de les variables en Ω_i i suposem que la matriu de covariàncies $\sum$ és comuna per a totes les poblacions.
Considerem les distàncies de Mahalanobis de cada individu ω a les poblacions

$$M^2(x, \mu_i) = (x - \mu_i)'\Sigma^{-1}(x - \mu_i), \quad i = 1, \ldots, k$$

Seguint un criteri de classificació que assignarà cada individu ω a la població més pròxima:

$$\text{Si } M^2(x, \mu_i) = min\{M^2(x, \mu_1), \ldots, M^2(x, \mu_k)\} \qquad \text{assignarem } \omega \text{ a } \Omega_i$$

En el cas de dues poblacions havíem definit la funció discriminant de la següent manera:

$$L(x) = \left[x - \frac{1}{2}(\mu_1 - \mu_2)\right]' \Sigma^{-1}(\mu_1 - \mu_2)$$

I pel cas de k poblacions les **funcions discriminants lineals** seran per a i, j=1,...,k i≠j

$$L_{ij}(x) = (\mu_i - \mu_j)'\,\Sigma^{-1}x - \tfrac{1}{2}(\mu_i - \mu_j)'\Sigma^{-1}(\mu_i - \mu_j)$$

3.2 Regla de la màxima versemblança

Coneixent la funció de densitat de x, $f_i(x)$, en la població Ω_i, obtindrem la regla de classificació assignant ω a la població on la versemblança sigui major:

$$\text{si } f_i(x) = màx\{f_1(x), \ldots, f_k(x)\} \qquad \text{assignarem } \omega \text{ a } \Omega_i$$

A més, si la funció discriminant en el cas de dues poblacions era:

$$V(x) = log\, f_1(x) - log\, f_2(x)$$

per k poblacions les funcions discriminants seran per a i, j=1,...,k i≠j

$$V_{ij}(x) = log\, f_i(x) - log\, f_j(x)$$

Tal com havíem comentat en el cas de poblacions normals amb matriu de covariàncies comuna, també en el cas de més de dues poblacions es verificarà la igualtat entre els discriminants de la regla de versemblança i els discriminants lineals, és a dir,

$$V_{ij}(x) = L_{ij}(x)$$

3.3 Regla de Bayes

Suposant que coneixem la funció de densitat $f_i(x)$ i les probabilitats a priori, $q_i, \ldots, q_k$, amb la regla de Bayes assignarem ω a la població que la probabilitat a posteriori és màxima

$$\text{si } q_i f_i(x) = màx\{q_1 f_1(x), \ldots, q_k f_k(x)\} \qquad \text{assignarem } \omega \text{ a } \Omega_i$$

i estarà associada a les següents funcions discriminants per a i, j=1,...,k i≠j

$$B_{ij}(x) = log\, f_i(x) - log\, f_j(x) + log\,(q_i / q_j)$$

3.4 Discriminador quadràtic

En el cas que les matrius de covariàncies siguin diferents, utilitzarem aquest criteri que es correspondrà als discriminants quadràtics per a i, j=1,...,k i≠j

$$Q_{ij}(x) = \frac{1}{2}x'\left(\Sigma_j^{-1} - \Sigma_i^{-1}\right)x + x'\left(\Sigma_i^{-1}\mu_i - \Sigma_j^{-1}\mu_j\right) + \frac{1}{2}\mu_j{}'\Sigma_j^{-1}\mu_j - \frac{1}{2}\mu_i{}'\Sigma_i^{-1}\mu_i + \frac{1}{2}log|\Sigma_j| - \frac{1}{2}log|\Sigma_i|$$

El QDA genera límits de decisió corbats, no lineals, pel que es aplicable pels casos en que la separació entre grups no és lineal. Per aquest motiu, QDA també és molt més flexible que LDA.

4. Classificació K-NN poblacions veïnes

L'algoritme K-NN poblacions veïnes, *k-nearest neighbors algorithm*, classifica cada observació nova al grup que pertanyi segons tingui k observacions veïnes més a prop d'un grup o d'un altre. Els passos que segueix aquesta classificació són els següents:

- Càlcul de les distàncies de l'observació a cada un dels existents.
- Ordre de les distàncies de menor a major.
- Selecció de les k distàncies menors.
- Tria del grup amb major freqüència dels k seleccionats amb menor distàncies.

Es tracta d'un algoritme d'aprenentatge supervisat, és a dir, que a partir d'una base de dades inicial d'estudi el seu objectiu serà el de classificar correctament totes les observacions noves. És denominat també com a *lazy learning*, ja que la classificació s'endarrereix tot el possible sense construir-se cap model. En aquest cas, el model serà la base de dades d'estudi, i es treballa quan arriba una nova observació a classificar.

Durant la fase d'entrenament de l'algoritme s'emmagatzemen els vectors de les variables d'interès i les etiquetes de les classes de les observacions d'entrenament. En la fase de classificació, l'avaluació de l'observació, de la qual no en coneixem la classe, és representada per un vector de les variables d'interès. Posteriorment, es calcula la distància entre els vectors emmagatzemats i el nou vector. Generalment s'utilitza la distància euclidiana:

$$d(x_i, x_j) = \sqrt{\sum_{i=1}^{k}\left(x_i - x_j\right)}$$

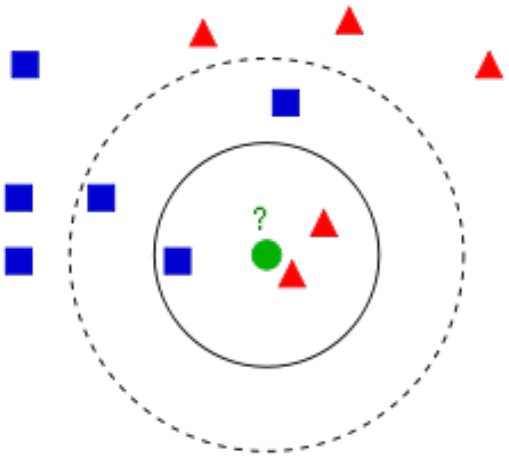
Es seleccionen les k observacions més pròximes. La nova observació es classificarà a la classe que més es repeteix en els vectors seleccionats.

Per aquesta classificació és necessària la definició del valor de k, que determinarà a quin grup pertany cada variable. Aquesta elecció depèn fonamentalment de les dades. La versió més simple del classificador k-NN és 1-NN. Per k=1 només s'obté un veí més pròxim en el conjunt de la base de dades d'estudi i assigna l'etiqueta de classe a la nova observació que es va a classificar.

Veiem en el següent gràfic un exemple per k=3 i k=5:

El punt verd representa la mostra de prova que volem classificar o bé en els quadrats blaus o els triangles vermells. Com que a priori hem definit que k=3, representat per la línia contínua, el punt verd s'assignaria als triangles vermells, ja que hi ha 2 triangles i només un quadrat dins del cercle interior. Canviem ara per k=5, representat per la línia discontínua, el punt verd s'assignaria als quadrats blaus, ja que dins d'aquest cercle hi ha més quadrats blaus que triangles vermells.

Gràfic 2.2. Exemple classificació 3-NN i 5-NN poblacions veïnes



Antti Ajanki

5. Probabilitat de classificació errònia

Una vegada que hem establert el mètode de classificació i hem classificat les nostres dades hem d'avaluar com de bona és la classificació resultant. Per tant, hem d'avaluar el percentatge d'encerts en les classificacions.
Com a mesura per conèixer el percentatge d'error de cada anàlisi s'utilitza la probabilitat de classificació errònia, és a dir, el nombre d'individus mal classificats dividit pel nombre total d'individus.

Les matrius de confusió són una de les millors maneres d'avaluar la capacitat d'encert que tenen els models discriminants. Aquestes matrius ens mostren el nombre de verdaders positius, verdaders negatius, falsos positius i falsos negatius.

Taula 2.1. Matriu de confusió

		Classe predita		
		-	+	TOTAL
Classe verdadera	-	Verdaders Negatius	Falsos Positius	N
	+	Falsos Negatius	Verdaders Positius	P
	TOTAL	N*	P*	N + P

La probabilitat de classificació errònia queda sobreestimada quan es realitza sobre el mateix conjunt d'individus que s'utilitzen per estimar la funció discriminant. Per evitar aquesta sobreestimació s'acostumen a utilitzar dos conjunts d'individus, un per estimar la funció i l'altre per valorar la classificació.

Per a l'obtenció d'una classificació correcte de les observacions haurem de tenir una bona estimació de les probabilitats a priori i la funció de densitat. Com més pròximes siguin al valor real, més s'aproximarà el classificador LDA al classificador de Bayes, i podem dir que serà millor ja que la regla de Bayes minimitzarà la probabilitat de classificació errònia.

III. DADES DE L'ESTUDI

Per la realització d'aquest estudi utilitzarem dades de la base de dades de l'empresa on actualment treballo, Selenta Group. Selenta Group és una empresa hotelera que té un total de 7 hotels distribuïts a diferents ciutats espanyoles, centralitzant alguns dels seus serveis a Barcelona.

En concret, s'utilitzarà la base de dades d'uns aquests hotels, situat a Tenerife: Mare Nostrum *Resort*. Aquest gran *resort* té un total de 1.036 habitacions i es troba segmentat en 3 seccions d'hotel diferents, cadascun d'ells amb els seus propis serveis i instal·lacions: Cleopatra Palace (CL), Mediterranean Palace (MP) i Sir Anthony (SA).

La base de dades té un total de 40.042 registres i 16 columnes i corresponen a un total de 40.042 estades que es van gaudir durant tot l'any 2018. Cada registre correspon a les dades d'un client que ha estat, necessàriament, uns dies allotjat en alguna d'aquestes tres seccions de Mare Nostrum *Resort*.

A continuació trobem una taula resum on hi trobem una breu descripció de cada variable de la base de dades que volem analitzar:

Taula 3.1. Descripció de les variables

VARIABLE	DESCRIPCIÓ	TIPUS DE VARIABLE
Hotel	Hotel on s'allotja l'hoste	Categòrica
Hoste	Identificador de l'hoste	Categòrica
Edat	Edat de l'hoste	Numèrica
Sexe	Gènere de l'hoste	Categòrica (Binària)
Canal	Canal de compra	Categòrica
Dia setmana alta	Dia de la setmana en què s'ha realitzat la reserva	Categòrica
Mes alta	Mes en què s'ha realitzat la reserva	Categòrica
Antelació	Antelació de compra (dia de check-in – dia realització de la reserva)	Numèrica

Dia setmana entrada	Dia de la setmana que l'hoste realitzarà el check-in a l'hotel	Categòrica
Mes entrada	Mes en què l'hoste realitzarà el check-in a l'hotel	Categòrica
Règim	Pensió contractada per reserva	Categòrica
Hostes per reserva	Número de persones per reserva	Numèrica
RN	Total de nits que els hostes s'allotjaran a l'hotel	Numèrica
RR	Import en € pel cost de l'habitació	Numèrica
AR	Import en € corresponent als costos extres per reserva	Numèrica
TR	Import total en € per reserva	Numèrica

A continuació realitzarem una anàlisi descriptiu de les dades, mostrant en cada cas les categories de cada variable, possibles diferències entre grups, relacions entre variables i punts destacables de cada una d'elles.

1. Anàlisi descriptiva

L'estadística descriptiva és una metodologia que ens proporciona informació sobre les dades d'una mostra: obtenint, organitzant, presentant i descrivint el conjunt de dades.

Aquesta anàlisi ens permetrà fer un primer nivell de selecció de les variables, que posteriorment haurem de decidir si les considerem valides per ser incloses a l'estudi o no, i valorar els suposats bàsics que s'han de tenir en compte.

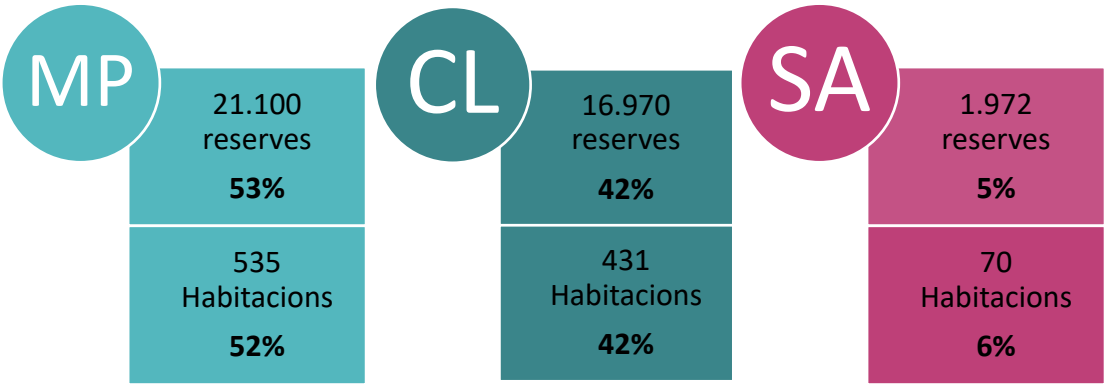
A través d'aquesta anàlisi exploratori de les dades, analitzarem la base de dades per trobar patrons, relacions entre variables, diferències significatives entre grups...

- **Hotel**: En aquesta variable hi trobem diferenciades les reserves per 3 grups diferents, corresponents a la secció de l'hotel on es va allotjar cada hoste:

    - Cleopatra Palace (CL)
    - Mediterranean Palace (MP)
    - Sir Anthony (SA)

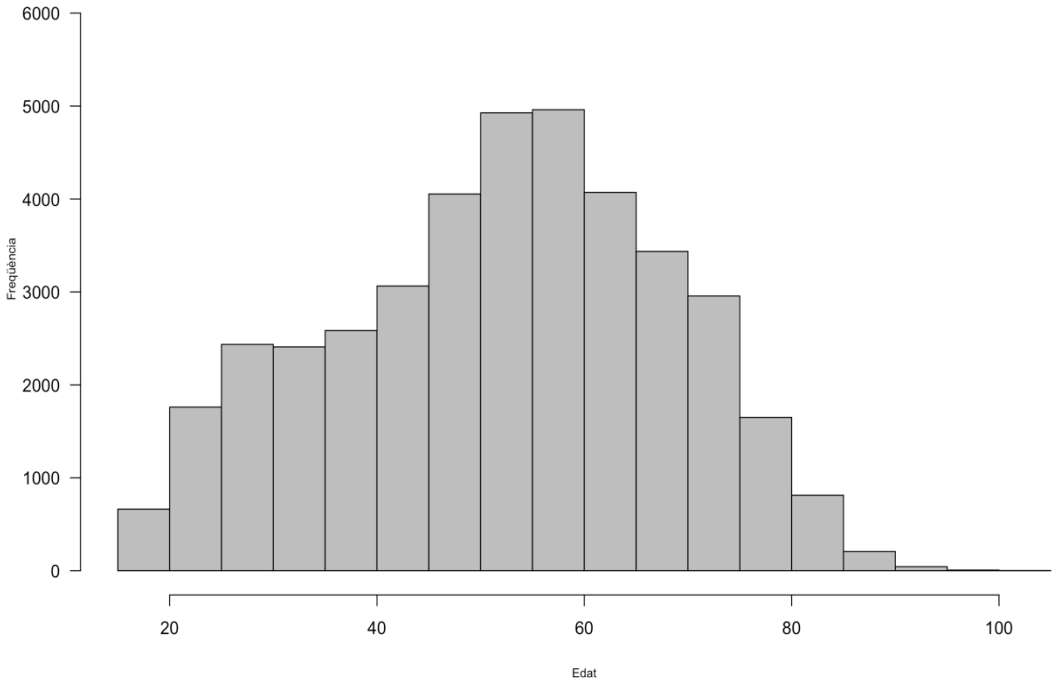
En el següent gràfic podem veure com es distribueixen els pesos de les seccions per reserva. Veiem que un 95% de les reserves corresponen als hotels MP i CL. El 5% restant pertanyen a SA. A més, podem observar que els percentatges de reserves i habitacions són pràcticament iguals en els tres hotels, això ens indica que cada hotel es reserva en la mateixa proporció, és a dir, els nivells d'ocupació dels tres hotels són quasi iguals.

Gràfic 3.1. Distribució de l'ocupació del hotels



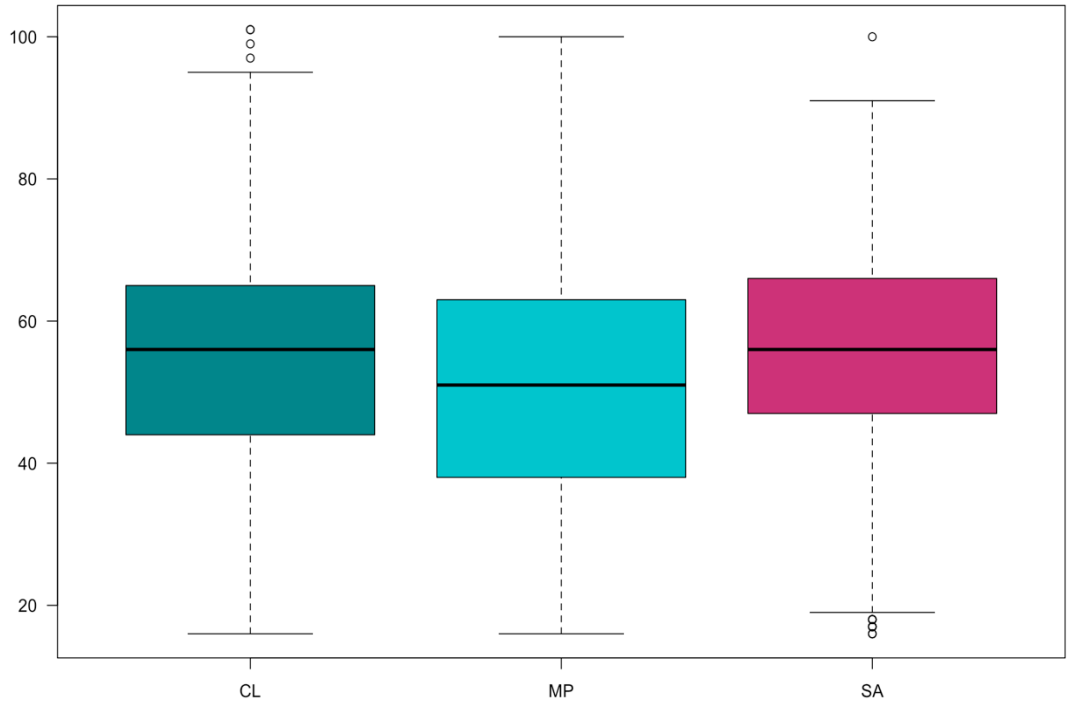
- **Edats**: L'edat que trobem a les reserves correspon a l'edat de l'hoste principal, que és qui ha realitzat la reserva. Després de veure'n l'histograma veiem que un 50% dels hostes tenen una edat d'entre 41 i 64 anys, situant-se la mitjana als 52 anys. A més, sabem que l'edat mínima és de 16 anys mentre que l'edat màxima és de 101 anys.

Gràfic 3.2. Histograma de la variable edat



A continuació veiem diferenciades les edats dels hostes de cada hotel: els hostes més joves es trobarien a l'hotel MP i els de més edat a SA.

Gràfic 3.3. Boxplot de la variable edat segmentat per hotel



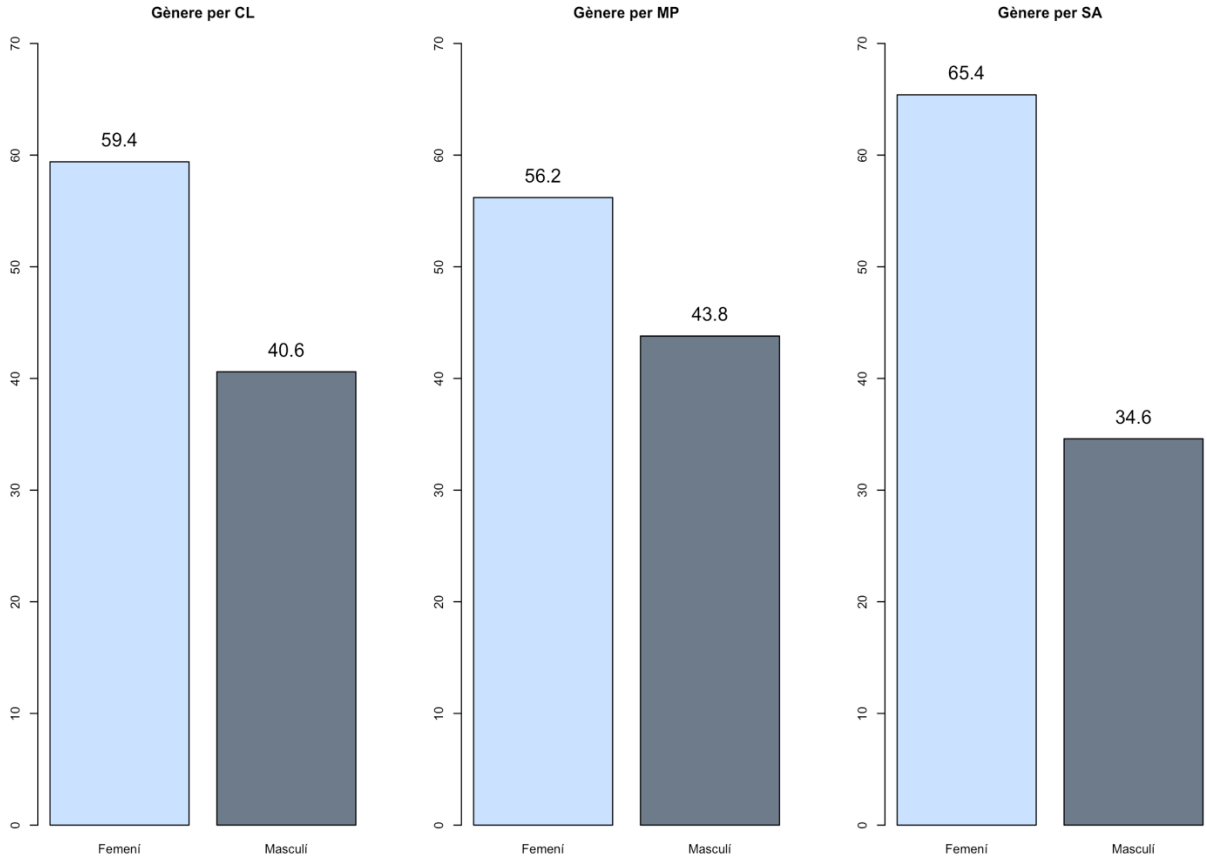
- **Gènere:** En aquesta variable queda recollit el sexe de l'hoste que ha realitzat la reserva. Un 58% corresponen al gènere femení, mentre que el 42% restant correspon al gènere masculí.

Taula 3.2. Distribució de la variable sexe

Sexe	*N*	*%*
Femení	23230	58%
Masculí	16812	42%

Veiem en la separació per grups que en els 3 hotels el gènere que segueix predominant, igual que estudiant les dades de forma univariada, és el femení. Els percentatges de gènere entre MP i CL són força semblants, una mica més diferenciats en el cas de SA, ja que trobem més pes pel gènere femení.

Gràfic 3.4. Gràfics de barres per gènere i hotel

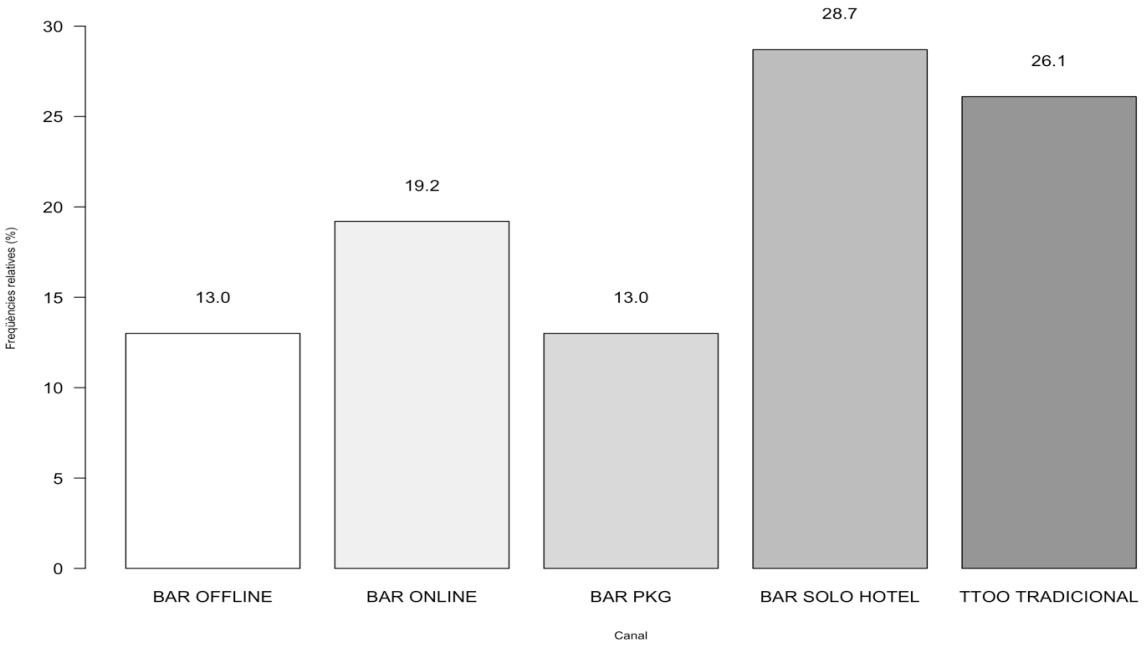


- **Canal de compra:** Aquesta variable ens indica per quin mitjà els hostes han realitzat la reserva. A continuació trobem les opcions que tenen els hostes a l'hora de fer una reserva:

    - BAR OFFLINE: aquest canal de compra és el telèfon, els hostes poden reservar trucant al mateix hotel.
    - BAR ONLINE: aquest canal de compra és la web pròpia de l'hotel on hi poden veure les diferents seccions de l'hotel i tota la resta de serveis.
    - BAR PKG: aquest canal de compra són els anomenats paquets, hotel + vols d'avió fins a l'illa. Els hostes que reserven per BAR PKG ho fan des d'una pàgina web externa a l'empresa que els busca una oferta d'allotjament i vols inclosos.
    - BAR SOLO HOTEL: aquest canal de compra són les OTA's, les agències de viatge online *(Online Travel Agency)*. Els exemples més coneguts d'OTA són Booking i Expedia.
    - TTOO: Tour operador, són les empreses que ofereixen serveis turístics, generalment integren diferents serveis: transport, allotjament, trasllats, excursions... Un exemple conegut és Thomas Cook, tour operador britànic.

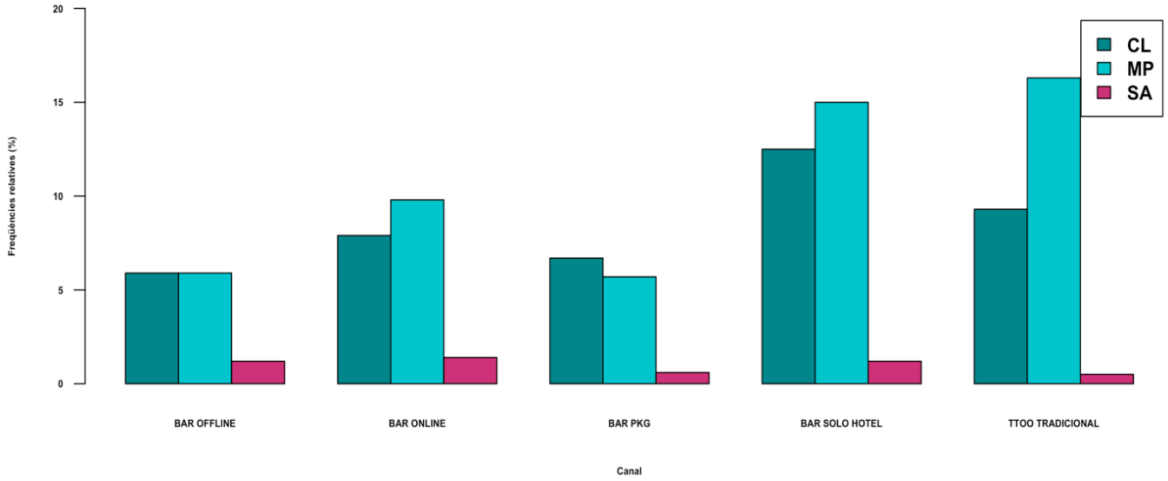
En el següent gràfic veiem com es troba repartit el pes en els diferents canals de compra pels quals es poden realitzar reserves. La major part de les reserves, un 29% ho van fer a través de BAR SOLO HOTEL, seguit del TTOO amb un 26%. Els canals propis de l'hotel, BAR ONLINE i BAR OFFLINE es reparteixen el 19% i 13% respectivament. El 13% restant van reservar a partir de BAR PKG.

Gràfic 3.5. Freqüències relatives de la variable canal de compra



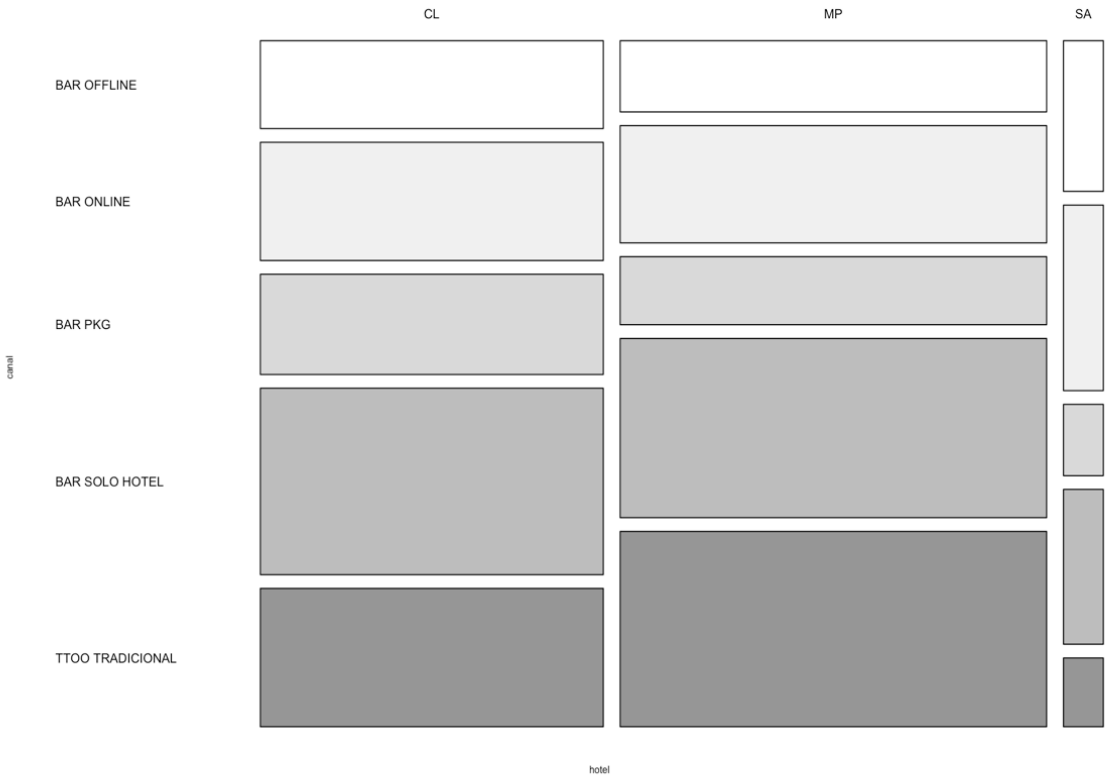
Ara ens interessa veure segmentat el canal de compra de cada hotel. Ho veiem a través del següent diagrama de barres i en destaquem que pel cas de SA la majoria d'hostes van realitzar les reserves pels canals ONLINE i BAR SOLO HOTEL. A més, veiem que la BAR PKG va aconseguir més reserves per l'hotel CL, mentre que un 16% de les reserves van realitzar-se per TTOO i van fer-se per l'hotel MP.

Gràfic 3.6. Freqüències relatives de la variable canal de compra segmentada per hotel



Però per poder observar de manera més clara les diferències entre hotels, observem un gràfic on podrem veure el pes que representa cada canal de venda per cada hotel i en destaquem que el comportament entre CL i MP és bastant similar, en canvi per SA en veiem algunes diferències, començant per tenir uns majors pesos de venda BAR OFFLINE i BAR ONLINE i menors en TTOO, en comparació amb els altres dos hotels.

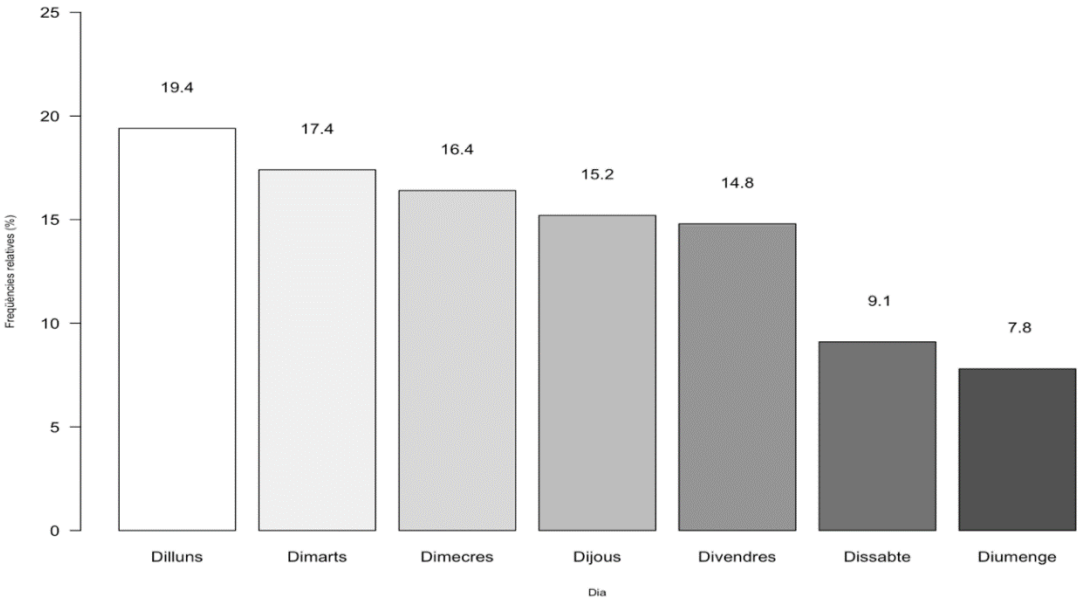
Gràfic 3.7. Gràfic de mosaic del canal de reserva segmentat per hotel



- **Dia de la setmana de la reserva:** A partir d'aquesta variable podem veure quin és el dia preferit de compra pels hostes.

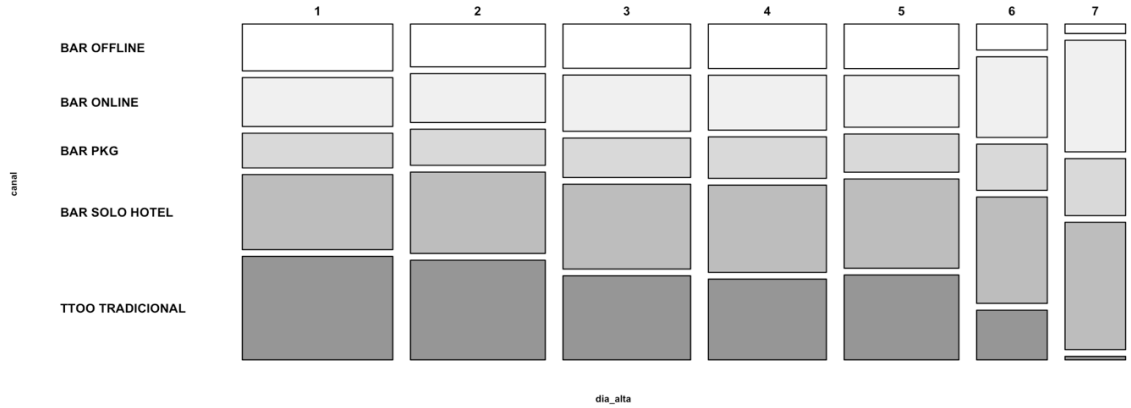
Observem el dia de la setmana de compra i veiem que el 83% dels hostes van preferir reservar entre setmana, mentre que el 9,1% i el 7,8% van preferir-ho fer dissabte i diumenge, respectivament. Veiem que el dia en el qual es van realitzar més reserves és dilluns, seguit de dimarts i en ordre descendent fins a final de setmana.

Gràfic 3.8. Freqüències relatives de la variable dia de la reserva



N'estudiem la possible relació entre el dia de la setmana de reserva i el canal de compra i a simple vista podríem pensar que aquestes variables són dependents: observem que durant la setmana les proporcions de la venda per canal es mantenen constants, els dissabtes incrementa la venda del canal BAR ONLINE i BAR SOLO HOTEL i el diumenge encara incrementen més aquests dos canals mentre que els canals BAR OFFLINE i TTOO es veuen reduïts quasi a 0, aquest fet era molt previsible, ja que aquests dos canals són els presencials i el diumenge és el dia de la setmana festiu, per tant no hi hauria cap persona per atendre i anotar les reserves en diumenge. Així doncs observem en el gràfic el pes que representen els canals BAR ONLINE, BAR PKG i BAR SOLO HOTEL, suposant que aquests segurament es mantenen, però en tenir els canals de venda presencials a 0, la freqüència relativa dels altres canals augmenta.

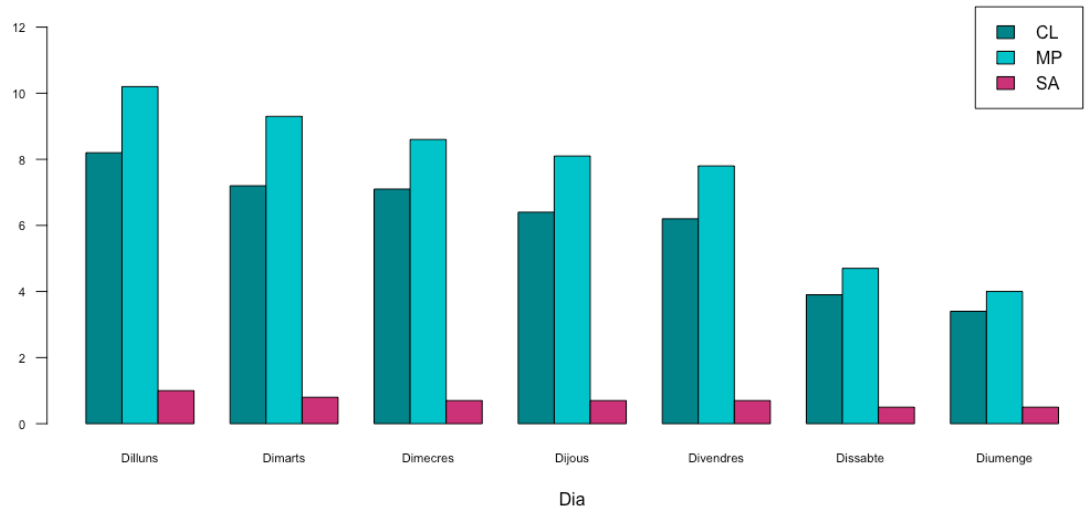
Gràfic 3.9. Gràfic de mosaic del dia de al reserva segmentat pel canal de compra



Per comprovar-ho realitzem el test de la Chi-quadrada de Pearson i el resultat és un p-valor molt inferior a 0.05 (X-squared = 2578.7, p-value < 2.2e-16), per tant es pot rebutjar la hipòtesi d'independència i per tant, aquestes dues variables seran dependents.

Segmentem aquesta variable per cada hotel, per veure si se segueixen la mateixa tendència de decreixement de reserva a mesura que va passant la setmana, de dilluns a diumenge, i podem comprovar que els hotels MP i CL si que segueixen aquesta tendència, en canvi SA segueix un comportament més lineal durant tots els dies de la setmana.

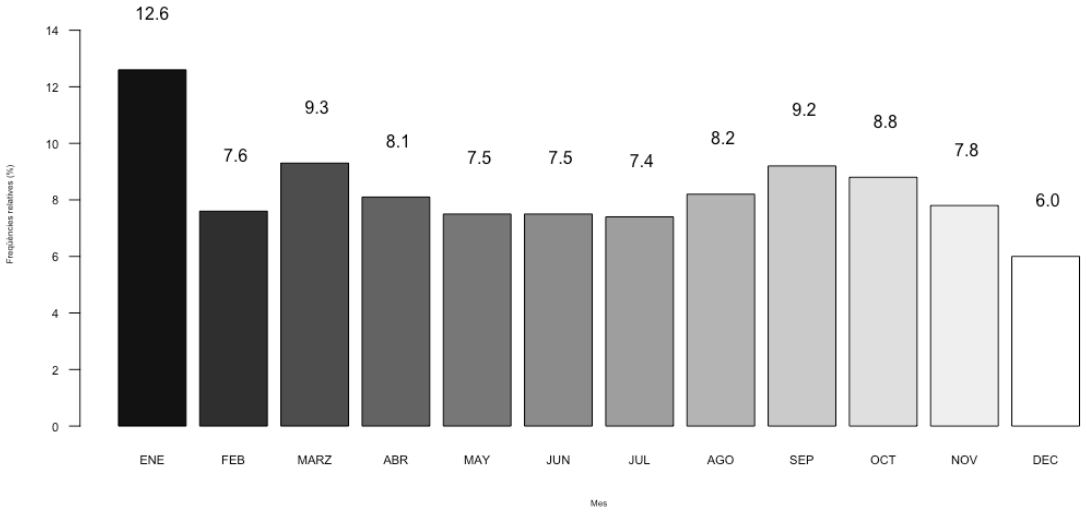
Gràfic 3.10. Gràfic de barres del dia de la reserva segmentada per hotel



- **Mes en què es va realitzar la reserva:** De la mateixa manera que coneixem a quin dia de la setmana es va realitzar cada reserva, també en coneixem el mes, on també hi veiem tendències de major o menor proporció de reserves.

A simple vista diferenciem el mes de Gener. Aquest seria el mes preferit de reserva dels hostes de Mare Nostrum *Resort*. D'altra manera també destaquem el mes de Desembre, que és quan menys reserves es van realitzar.

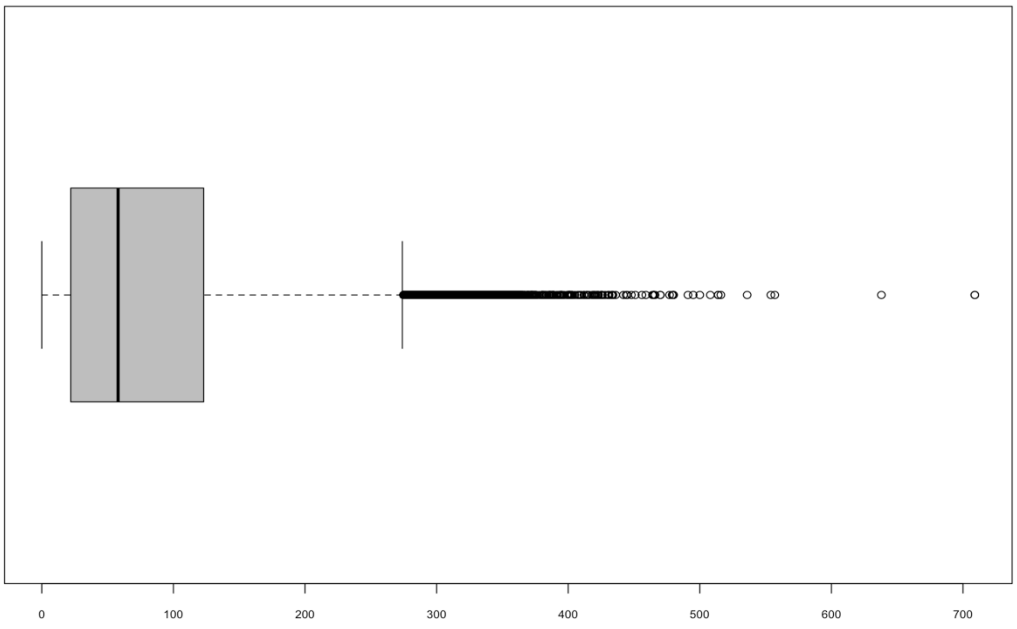
Gràfic 3.11. Freqüències relatives del mes de la reserva



- **Antelació de compra:** L'antelació amb què reserven els hostes es construeix a partir del dia d'entrada a l'hotel menys el dia de compra. Així doncs, es pot saber amb quants dies d'antelació ha realitzat la reserva cada hoste.

L'antelació mitjana és de 86 dies, mentre que el 50% dels hostes han reservat entre els 22 i 123 dies abans de l'entrada a l'hotel.

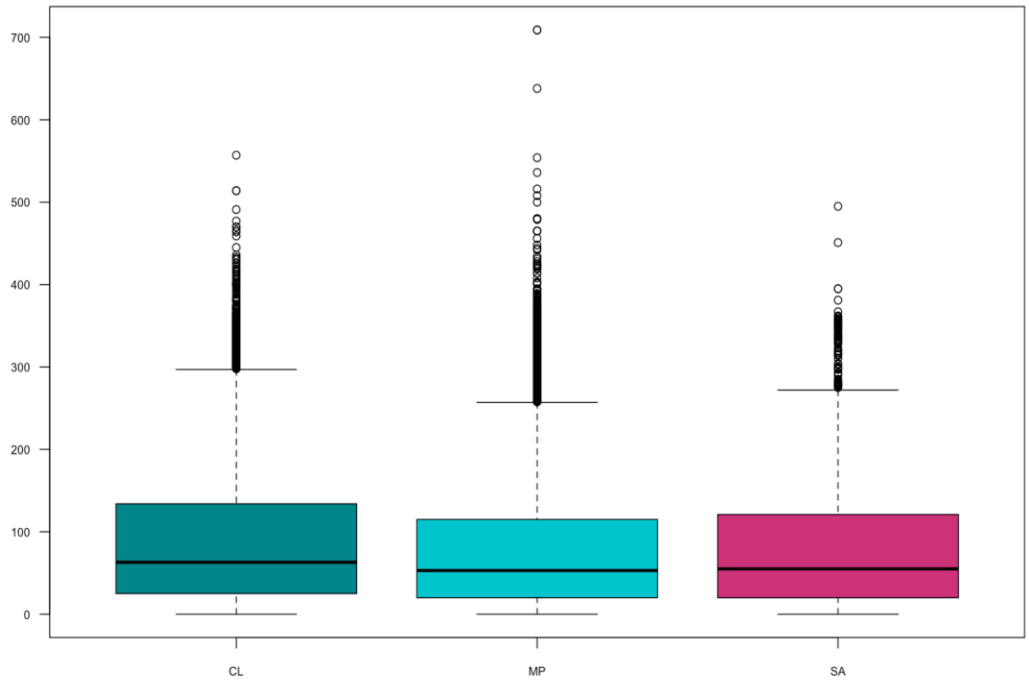
Gràfic 3.12. Boxplot de la variable antelació de compra



Busquem si hi ha diferencies per aquesta variable entre les diferents seccions d'hotel i veiem que els hostes de l'hotel CL són els que van reservar amb més antelació tot i ser les reserves de més antelació per MP (outliers). Majoritàriament, i sense contemplar les

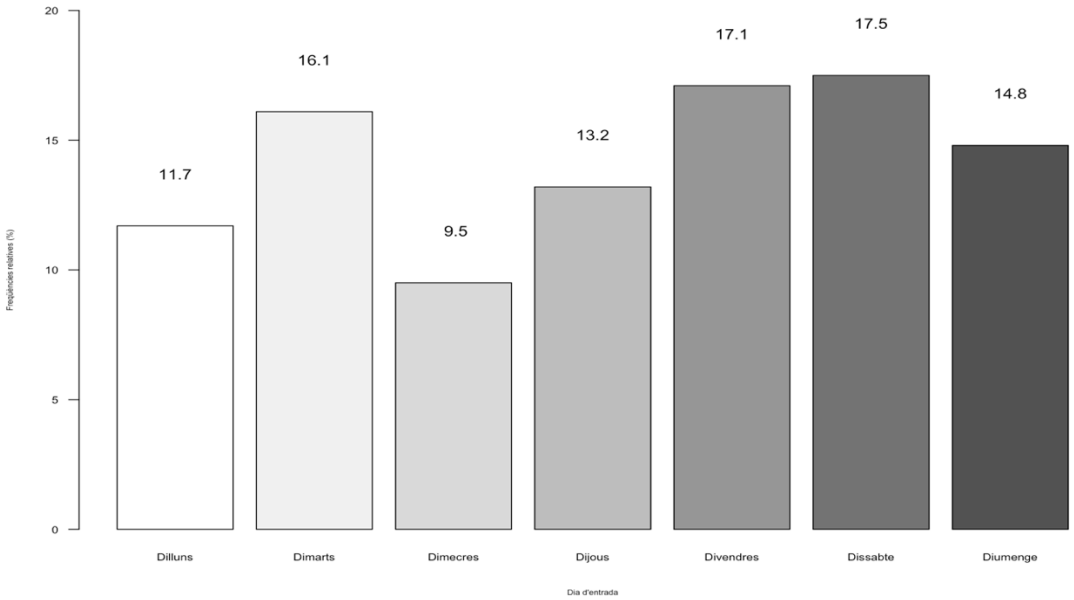
reserves atípiques, els hostes de MP i SA van tenir un comportament similar en quant a la antelació de reserva.

Gràfic 3.13. Boxplots de la variable antelació de compra per hotel



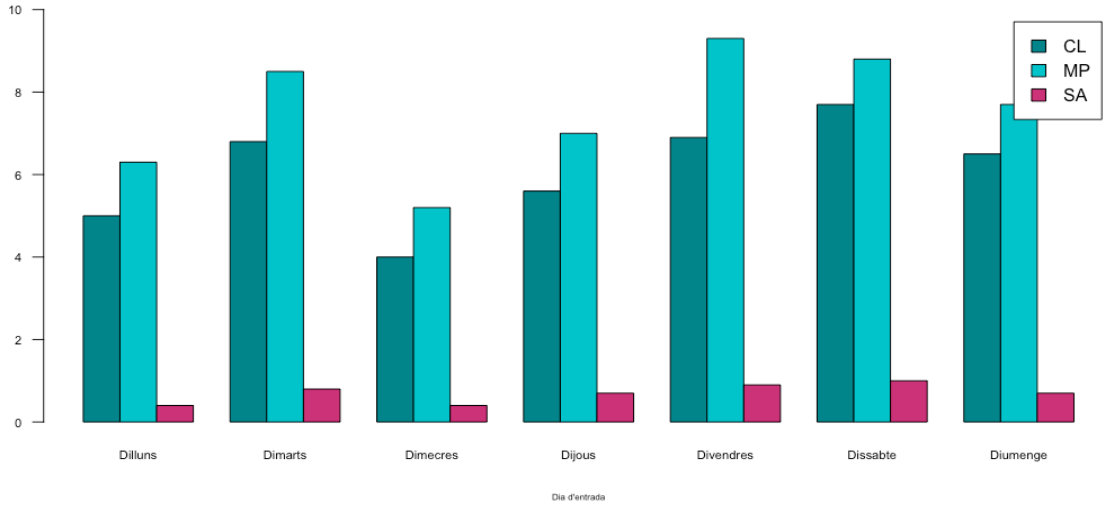
- **Dia d'entrada:** Aquesta variable ens indica el dia de la setmana en el que els hostes realitzen el check-in i que, per tant, dia en el qual es començaran a allotjar a l'hotel. Observem que el dia de la setmana preferit per començar la seva estança a l'hotel va ser el divendres i el dissabte.

Gràfic 3.14. Freqüències relatives del dia d'entrada



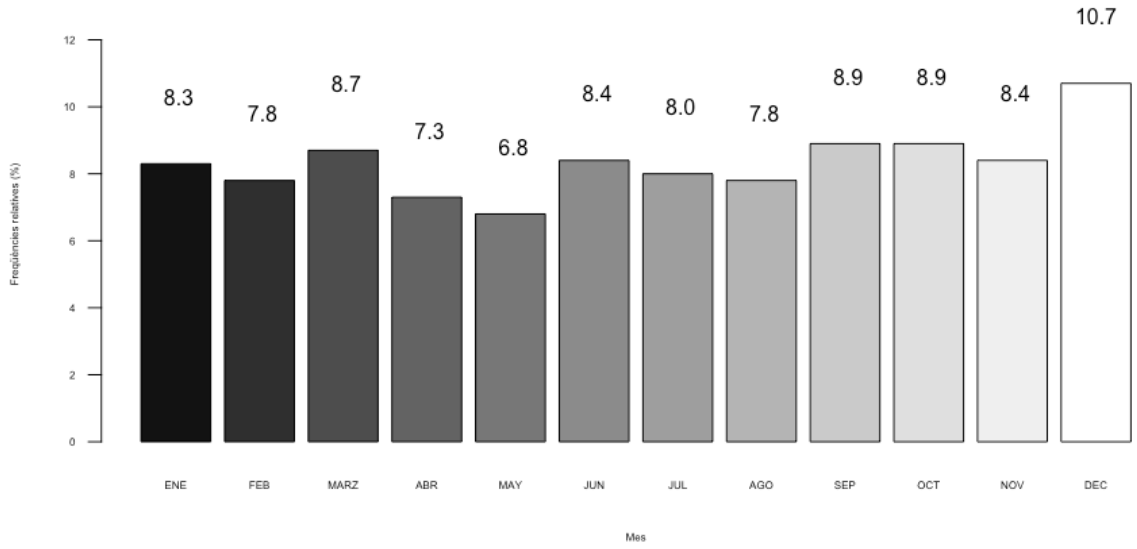
Observem si a simple vista hi podem diferenciar pesos entre cada grup: El divendres seria el dia preferit d'entrada a l'hotel MP, en canvi per l'hotel CL seria el dissabte. En el cas de SA els dies preferits serien divendres i dissabte.

Gràfic 3.15. Gràfic de barres del dia d'entrada segmentada per hotel



- **Mes d'entrada:** Tal com es coneix, en el món hoteler hi ha temporades en què la demanda és superior. En el cas d'un hotel de Tenerife, on les temperatures són elevades i constants durant tot l'any, és possible que no existeixi una tendència d'altes ocupacions durant uns mesos en concret. El que veuríem en hotels de la península, és una tendència d'alta ocupació durant els mesos d'estiu, però en aquest cas i tal com podem comprovar en el següent gràfic, cap mes destaca significativament respecte als altres. En podríem destacar una major ocupació durant el mes de desembre.

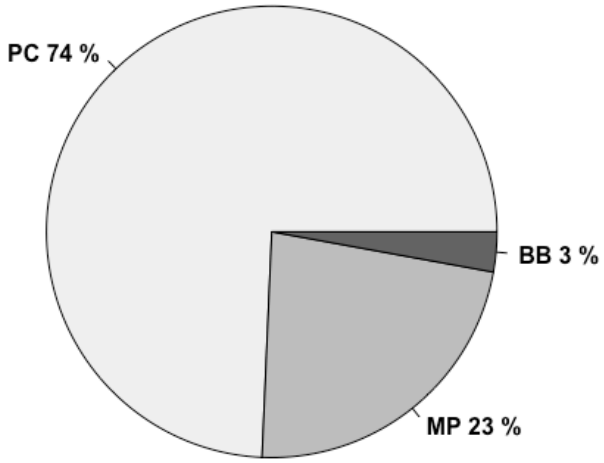
Gràfic 3.16. Freqüències relatives del mes d'entrada



- **Règim de la reserva (pensió):** En realitzar la reserva d'una habitació a l'hotel, hi ha la possibilitat de reservar una sèrie de serveis extres que ofereix l'hotel. L'extra més reservat és el de les pensions, trobant les següents opcions:

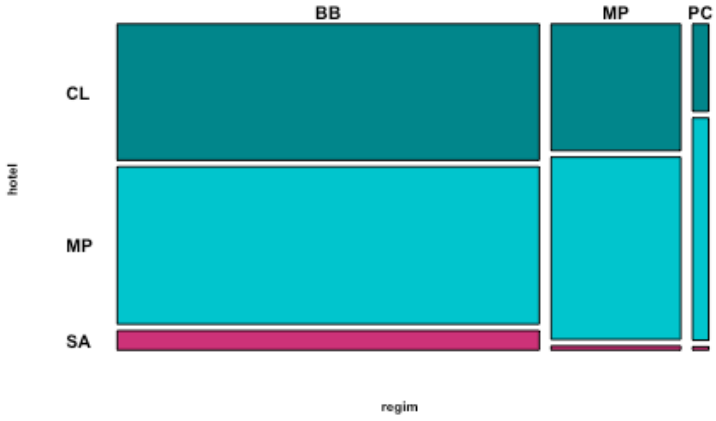
    - BB: *Bed and breakfast,* és l'opció bàsica, en reservar l'habitació s'inclou en el preu l'esmorzar.
    - MP: Mitja pensió. Els hostes que escullen aquesta opció tindran inclòs, a més a més de l'esmorzar, el sopar.
    - PC: Pensió completa. Aquesta és l'opció més completa, i per tant també serà la més cara. Inclou l'esmorzar, el dinar i el sopar.

Gràfic 3.17. Gràfic de sectors de la variable Règim



En el següent gràfic observem la variable règim segmentada per hotel, i en traiem algunes conclusions: el pes de les pensions MP i PC pel cas de Sir Anthony són mínimes. Pel cas de Mediterranean Palace les pensions MP i PC són superiors a Cleopatra Palace.

Gràfic 3.18. Gràfic de mosaic del règim segmentat per hotel

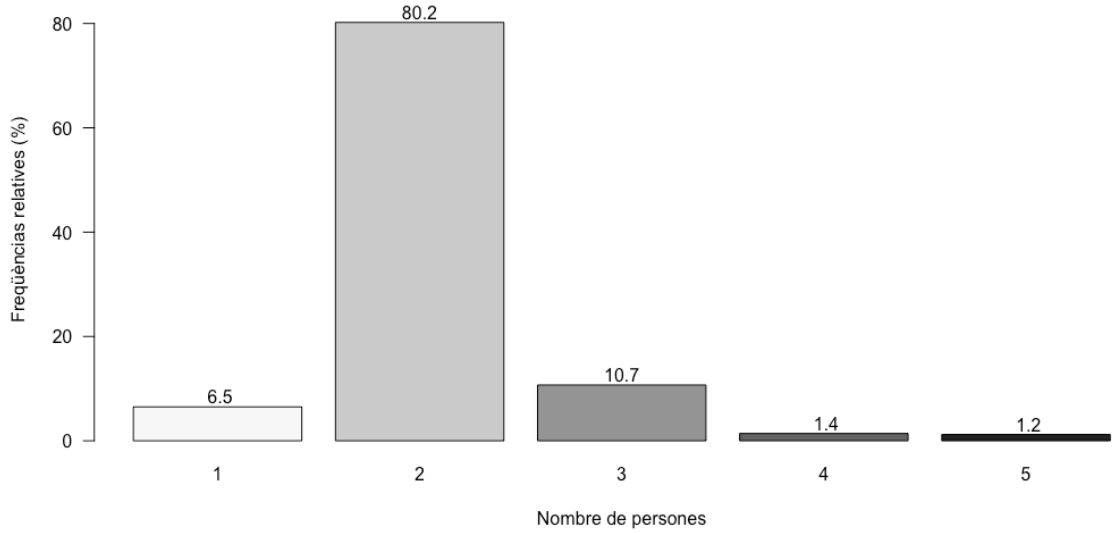


Després de realitzar el test de la chi-quadrat (X-squared = 606.6, p-value < 2.2e-16), podem rebutjar H_0 que ens indica que les variables no són independents, i per tant el règim de les reserves seria dependent de l'hotel.

- **Hostes per reserva:** En aquesta variable hi trobem el nombre total de persones que per reserva, és a dir l'hoste principal de la reserva més els seus possibles acompanyants. El valor mínim per reserva és d'una persona mentre que el màxim de persones en una reserva és de 5.

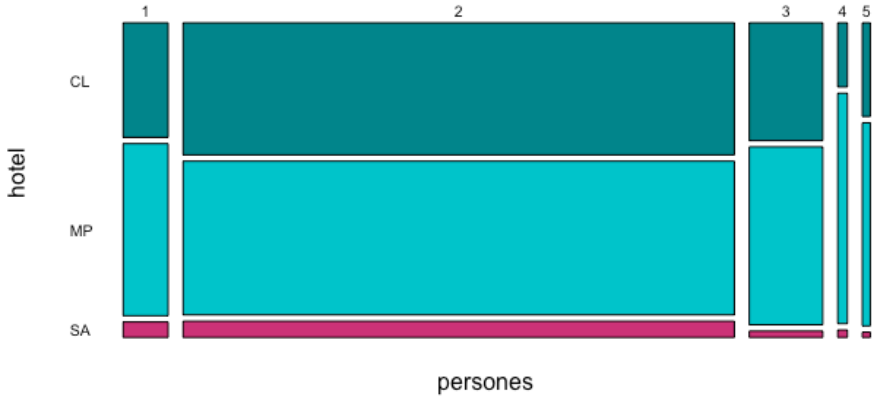
  Un 80% de les reserves són de 2 persones, mentre que el total de reserves de 4 o 5 persones no arriba al 3%.

Gràfic 3.19. Freqüències relatives del nombre de persones per reserva



Observem el nombre de persones per reserva en cada hotel i veiem que la majoria de reserves de 4 o 5 persones s'allotgen a MP.

Gràfic 3.20. Gràfic de mosaic del nombre de persones segmentat per hotel



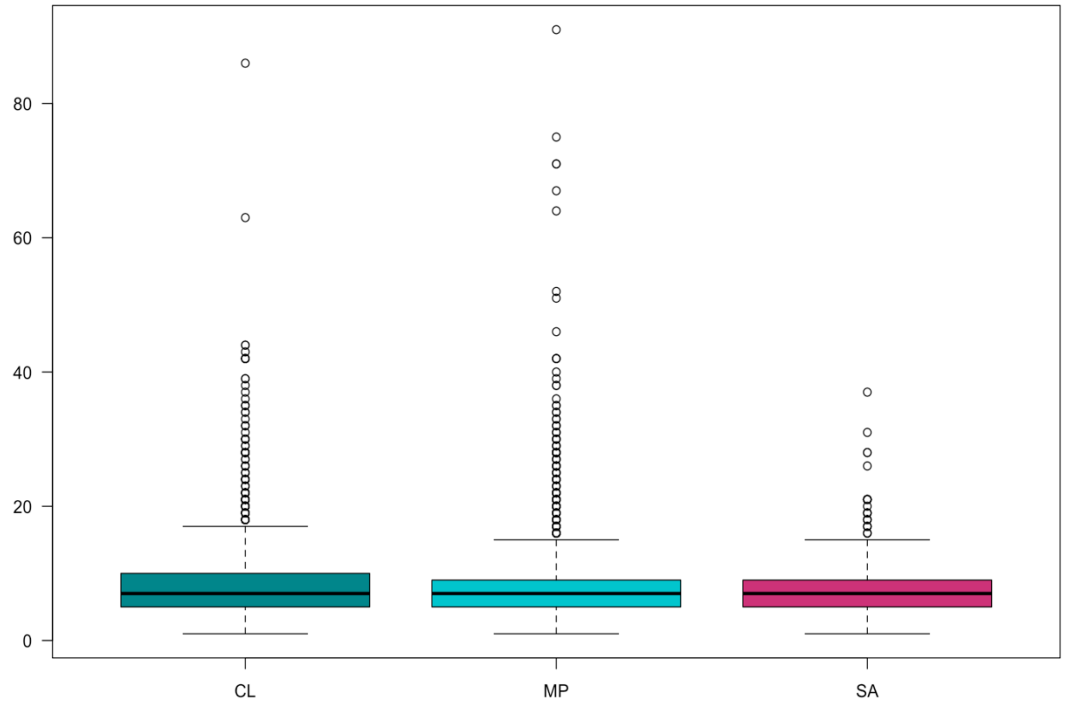
- **Nits dormides a l'hotel:** Aquesta variable està construïda a partir del càlcul entre dues variables: Data de sortida – Data d'entrada. D'aquesta manera coneixem el nombre total de nits que han estat els hostes allotjats a l'hotel.
  La mitjana de nits per reserva del 2018 va ser de 7,5 nits/reserva i un 50% dels hostes van reservar entre 5 i 9 nits.

Taula 3.3. *Resum estadístic* de la variable RN

Min.	1st Qu	Media	Mean	3rd Qu	Max.
1.000	5.000	7.000	7.486	9.000	91.000

A més a més, volem veure els valors d'aquesta variable segmentats per cada hotel. Observant el boxplot segmentat per seccions de l'hotel veiem que la mitjana de nits és pràcticament igual per les 3 seccions. Això ho podem veure a través de les caixes, ja que són molt semblants. En canvi, podríem observar algunes diferències en els outliers, hi veiem que on més estades de llarga durada és a MP i CL.

Gràfic 3.21. Boxplots de la variable total de nits segmentada per hotel



- **Import per habitació (€):** El Room revenue és l'import associat al preu de l'habitació, expressat en €, que ha pagat per tota l'estada de la reserva cada hoste.

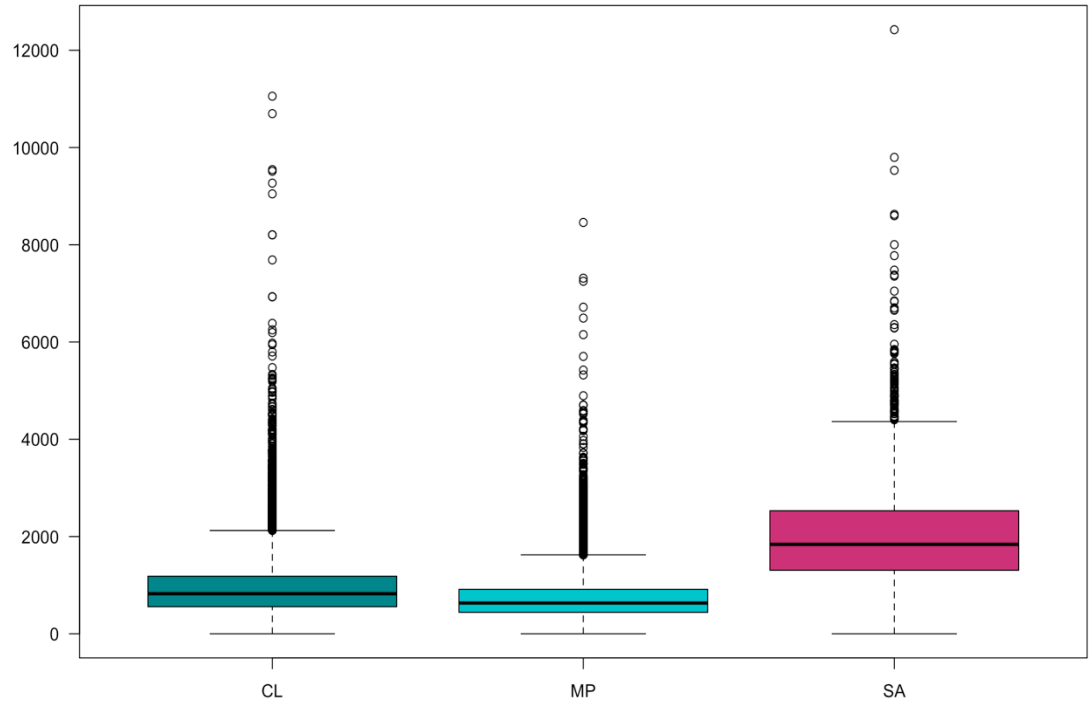
Observem que el 50% del room revenue de les reserves es troba entre els 483€ i els 1102€. Veiem que hi ha reserves que tenen un import de RR de 0, aquestes són mínimes, no representen ni el 0.01% de les reserves, i es tractaria d'unes reserves que han sigut gaudides sense cost, o bé treballadors desplaçats a l'hotel per feina o hostes guanyadors d'un concurs fruit de la publicitat digital.

Taula 3.4. *Resum estadístic d*e la variable RR

Min.	1st Qu	Media	Mean	3rd Qu	Max.
0	483.1	730.5	892	1102.5	12424.1

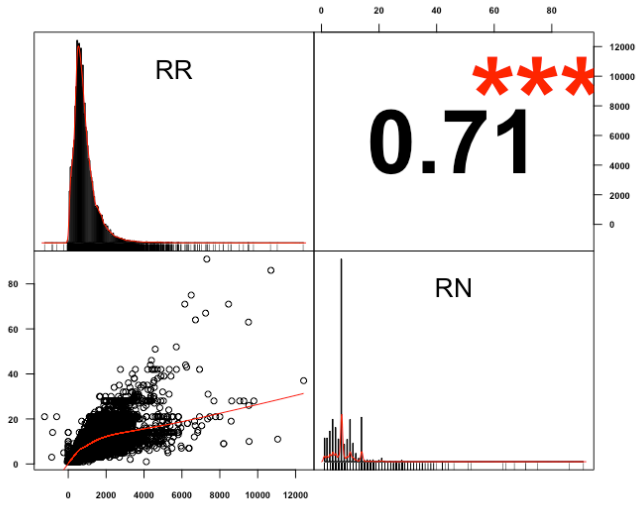
A continuació observem la segmentació d'aquesta variable per hotel, i tal com havíem comentat anteriorment, s'hi pot destacar a simple vista la diferència entre CL i MP amb SA. SA és l'hotel on més cares són les estances, ja que es tracta d'un hotel de luxe.

Gràfic 3.22. Diagrama de caixes de la variable RR segmentada per hotel



En el cas d'aquesta variable, podríem suposar que està fortament correlacionada amb la variable de les nits dormides a l'hotel, degut que l'import en euros del Room Revenue s'assigna en funció de les nits dormides. Visualitzem el resultat d'aquesta hipòtesi a partir d'un paquet de R anomenat `PerformanceAnalytics` i veiem que les suposicions són certes, les variables RN i RR tenen una correlació positiva.

Gràfic 3.23. Gràfic de correlació entre les variables RR i RN



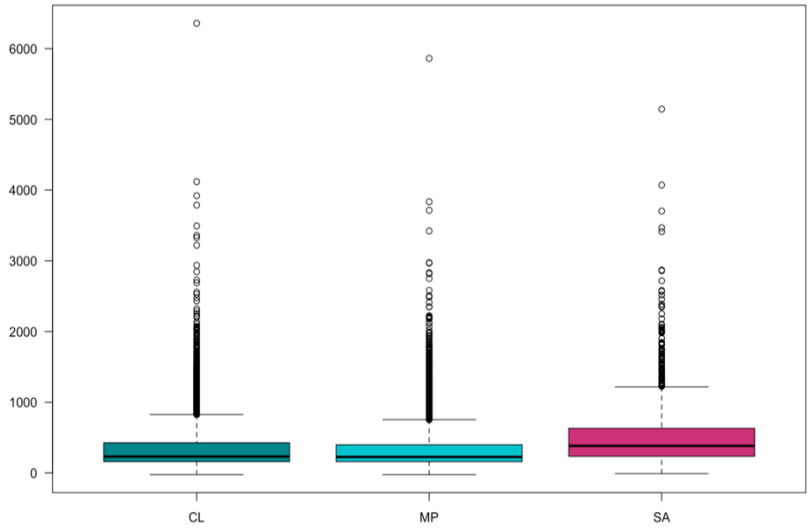
- **Imports extres (€):** L'extra revenue representa els imports generats en € pels hostes amb serveis que no són el room revenue. Alguns d'aquests serveis poden ser: les pensions, el transport, un servei de spa, etc.

Taula 3.5. *Resum estadístic* de la variable AR

Min.	1st Qu	Media	Mean	3rd Qu	Max.
-23.78	161	230.15	324.94	425.67	6358.58

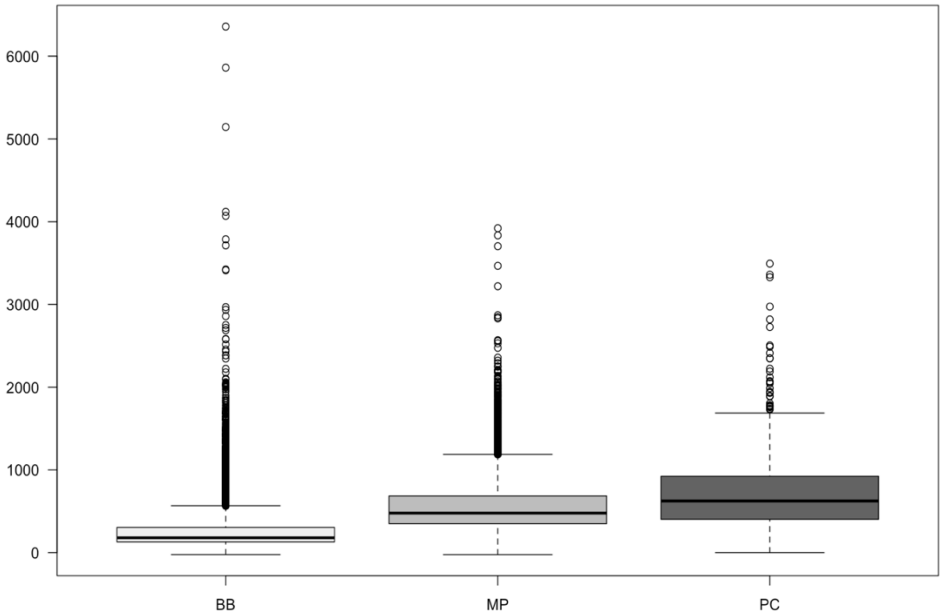
A continuació, trobem un boxplot que ens mostra l'extra revenue generat per hotel. N'observem les mitjanes i veiem que pot haver-hi diferències significatives: el preu mitjà generat per reserva d'extra revenue en el cas de CL és 26€ superior al de MP i el de SA ho és 200€ més.

Gràfic 3.24. Diagrama de caixes de la variable AR segmentada per hotel



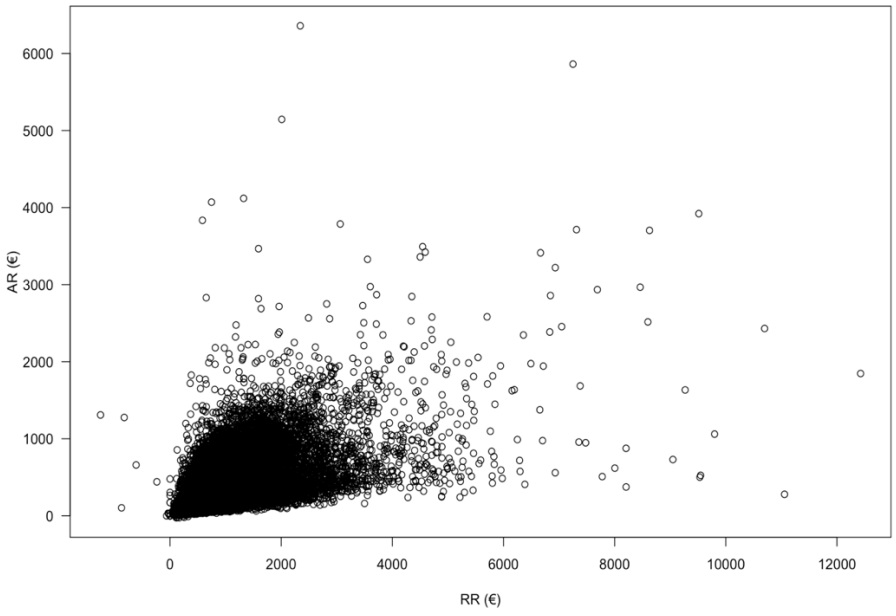
Pel cas d'aquesta variable, podríem considerar dues possibles relacions: pensions i room revenue. Suposem una possible correlació entre les variables pensions i extra revenue, ja que coneixem que les pensions són els serveis extres que més contracten els hostes.
En l'anàlisi bivariant de l'extra revenue i el règim, podem veure que on hi ha un extra revenue major és en el cas de les pensions completes, contemplant una major variància en la contractació del règim bàsic.

Gràfic 3.25. Diagrama de caixes de la variable AR segmentada per règim



Entre les variables contínues extra revenue i room revenue hi observem una correlació positiva: a més room revenue més extra revenue, amb un coeficient de correlació r=0,58.

Gràfic 3.26. Gràfic de correlació entre AR i RR



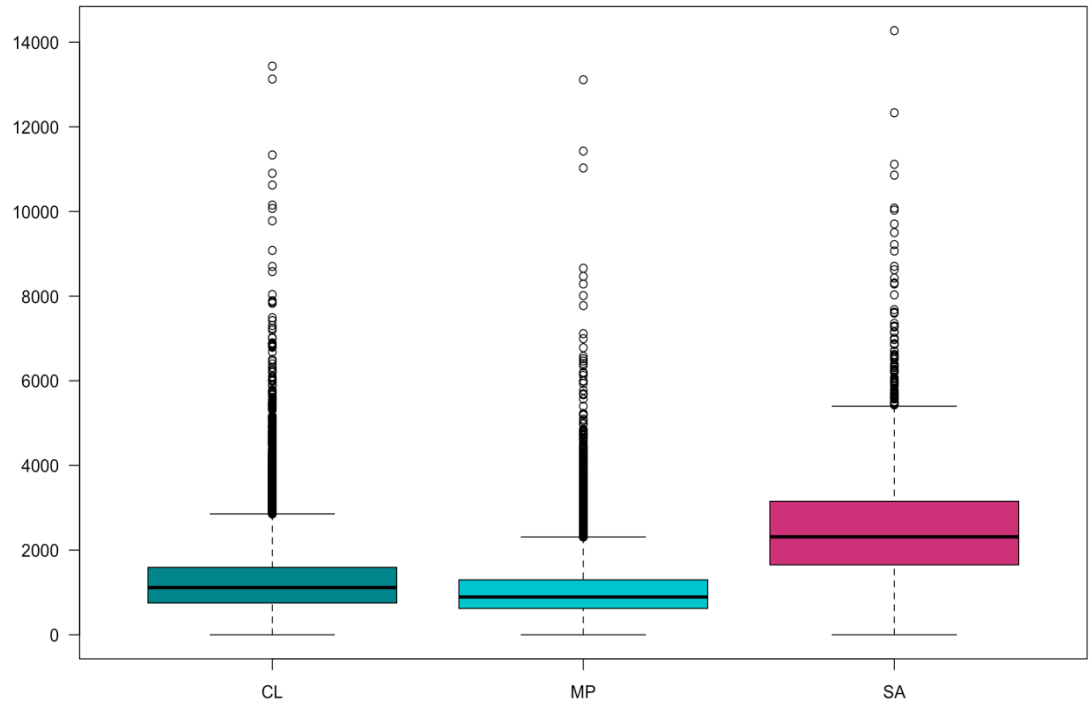
- **Total revenue (€):** Aquesta variable és la suma entre el Room Revenue i l'Extra Revenue, per tant, el preu total per reserva que paga cada hoste per cada reserva. El preu mitjà per reserva és 1.217€, mentre que hi hauria reserves que han costat fins a 14.270€.

Taula 3.6. *Resum estadístic* de la variable TR

Min.	1st Qu	Media	Mean	3rd Qu	Max.
0	677.8	1014.2	1217.2	1517.8	14270.6

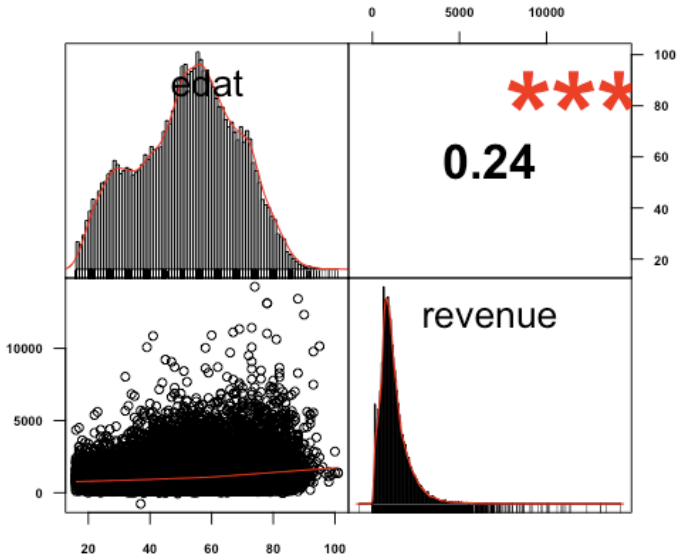
Observem també, aquesta variable segmentada per hotel per veure'n les principals diferències. Tal com hem comentat anteriorment, a partir de les mitjanes de total revenue de cada secció, podem afirmar que els hostes que s'allotgen a SA es gasten en mitjana molts més diners, i a més a més, tal com hem vist anteriorment, podem afirmar que no tenen les estades més llargues.

Gràfic 3.27. Diagrama de caixes de la variable Total revenue segmentada per hotel



Ara que sabem que la mitjana d'edat més elevada, és a dir, gent més gran, pertany a SA, i que es tracta de l'hotel més de luxe, plantegem una hipòtesi de correlació entre l'edat i el total revenue. En el següent gràfic, del paquet `PerformanceAnalytics`, podem veure que la variable revenue està poc correlacionada amb l'edat.

Gràfic 3.28. Gràfic de correlació entre les variables edat i revenue



IV. TRACTAMENT DE LES DADES

Abans de l'aplicació de tècniques estadístiques multivariants, les variables de la base de dades han de complir una sèrie de suposats bàsics per tal de la correcta realització i interpretació dels resultats.

En el cas de l'anàlisi discriminant és recomanable el compliment d'alguns suposats bàsics que garantiran uns resultats fiables. Es resumeix principalment en la comprovació de dues principals hipòtesis:

- Les variables independents o predictives han de seguir una distribució normal multivariant
- Les matrius de covariàncies han de ser igual en tots els grups.

Tot i que cal tenir en compte que l'anàlisi discriminant és considerat una tècnica robusta[2], i per tant, els resultats no es veurien altament afectats si no es complís alguna de les condicions.

Abans d'analitzar els suposats bàsics, estudiarem la multicol·linealitat de les dades, per evitar realitzar el tractament de les dades amb multicol·linealitats exactes.

1. Multicol·linealitat

La multicol·linealitat és una condició que succeeix quan algunes variables predictores estan correlacionades amb altres variables predictores.
Per mesurar la multicol·linealitat, examinem l'estructura de correlació de les variables predictores. En aquest cas, ho visualitzarem a través de la matriu de correlacions del paquet `PerformanceAnalytics`.

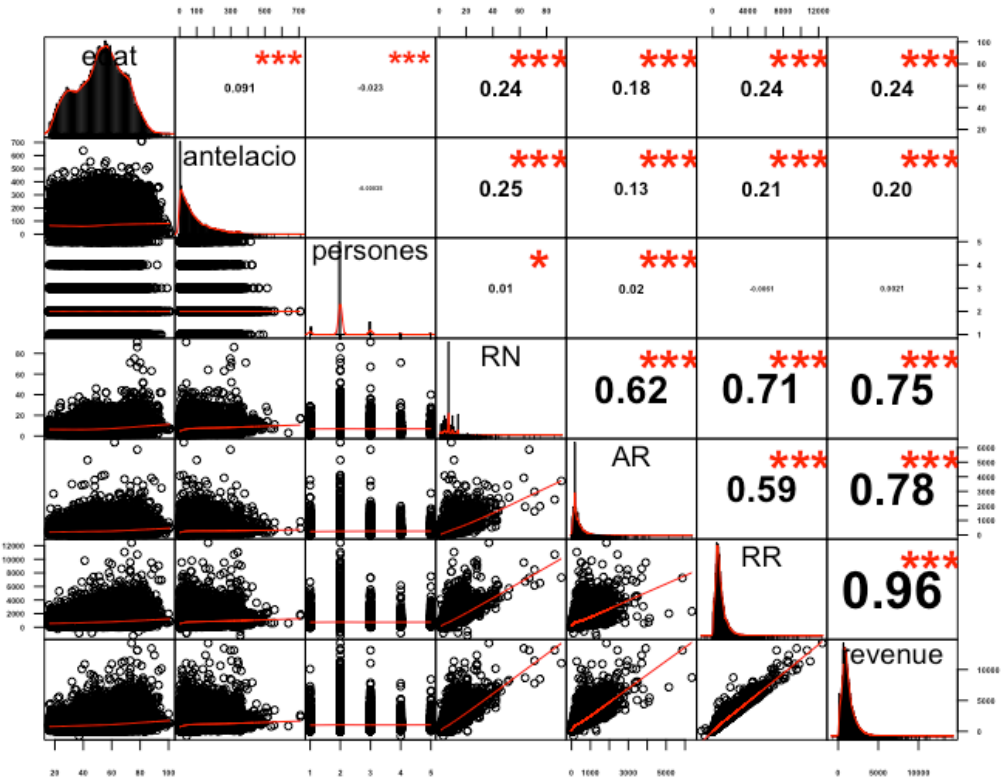
Observem la presència d'una multicol·linealitat exacte entre revenue, AR i RR. Es considera que una multicol·linealitat és exacte quan el coeficient de correlació és 1.
A priori sabíem que això passaria en el nostre estudi, ja que revenue és una variable que pot ser tractada de manera desglossada, per conèixer quina part del preu total de la reserva correspon a l'habitació i quina als serveis gaudits. En la nostra base de dades havíem decidit analitzar les dades amb aquesta variable de manera desglossada i de manera entera. Observem al gràfic de correlacions l'alta correlació entre revenue - RR i revenue - AR, amb valors molt pròxims a 1 en el cas de revenue - RR. Per aquest motiu, considerem una millor

[2] Tècnica que no es veurà alterada per conseqüència de la violació d'alguns dels suposats bàsics.

opció l'eliminació de la variable revenue i la realització de l'anàlisi de les dades amb les variables AR i RR.

Seguim observant els coeficients de correlació i podem suposar que l'estudi no es veurà afectat per greus problemes de correlació entre variables. La variable més correlacionada és RN, i ho està amb les variables AR i RR. Aquest fet és totalment lògic, ja que el preu que paguen els hostes per reserva va en funció de les RN: com més RN més elevat serà el preu. El mateix passa amb la correlació entre RN i AR: com més dies està l'hoste allotjat a l'hotel més costos tindrà.

Gràfic 4.1. Gràfic de correlacions



2. Normalitat

Com s'ha assenyalat anteriorment, un dels suposats bàsics que s'ha de complir és que les variables observades segueixin una distribució normal multivariant, ja que donat que el cas contrari, ni els estimadors plantejats ni els ajustos globals serien els òptims.

Primer de tot, farem referència a alguns procediments que ens permetran estudiar la normalitat univariant i posteriorment utilitzarem tècniques per contrastar la hipòtesi de normalitat multivariada.

2.1 Normalitat univariant

Per tal de contrastar si el conjunt de dades s'ajusten o no a una distribució normal hem decidit utilitzar el test de Kolmogorov-Smirnov. Kolmogorov-Smirnov és una prova no paramètrica[3] que determina la bondat d'ajust[4] entre dues distribucions de probabilitat entre si. Aquesta tècnica és similar al test de Shapiro Wilk, amb la principal diferència de la grandària de la mostra. Mentre que Shapiro Wilk es pot utilitzar fins a una grandària mostral de 50 observacions, el test de Kolmogorov-Smirnov és recomenable a partir de més de 50 observacions.

Abans de l'aplicació d'aquesta prova a R, es necessari el plantejament del contrast d'hipòtesis que realitzarem:

H_0: les dades provenen d'una distribució normal.
H_1: les dades no provenen d'una distribució normal.

Una vegada hem realitzat el test de kolmogorov-Smirnov, hem pogut comprovar que per totes les variables els p-valors són menors a 0.05 i per tant es rebutja H_0.

Taula 4.1. Resultats test de Kolmogorov-Smirnov

Hotel	***D***	***p-valor***	***Rebuig H_0***
Edat	0.044	< 2.2e-16	✓
Antelació	0.157	< 2.2e-16	✓
RN	0.232	< 2.2e-16	✓
AR	0.144	< 2.2e-16	✓
RR	0.138	< 2.2e-16	✓

Hem de tenir en compte que estem treballant amb un nombre molt gran d'observacions i en tractar-se de p-valors, com major sigui la grandària de la mostra més poder estadístic tenen i més fàcil és trobar evidencies en contra de H_0. D'altra banda, com major sigui la

[3] Tècniques que tenen en comú l'absència d'assumpcions sobre la llei de probabilitat que segueix la població de les dades.
[4] Descriu com de bo és l'ajust d'un conjunt d'observacions. Resumeix la discrepància entre els valors observats i els esperats en el model d'estudi.

mida de la mostra, menys sensibles són els mètodes paramètrics a la falta de normalitat. Per aquestes raons, observarem també algunes representacions gràfiques.

Analitzarem dues representacions gràfiques que ens ajudaran a conèixer la normalitat o no normalitat de les dades:

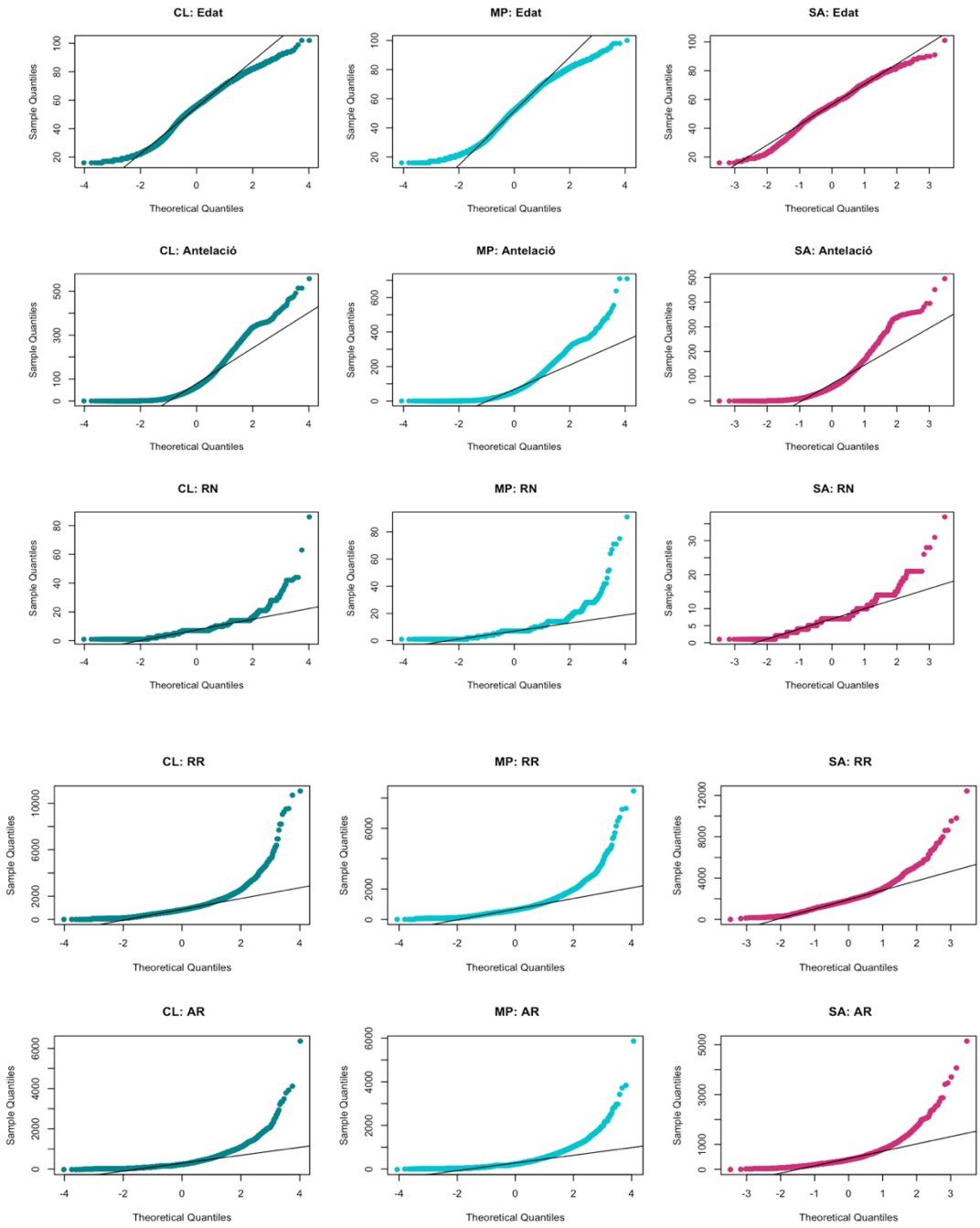
- Histograma: per tal de veure si la distribució que segueixen les dades és similar a la campana de Gauss (unimodal, campaniforme, simètrica...).
- Gràfic de la probabilitat normal: en el que les dades haurien d'aproximar-se a la línia recta per poder acceptar que són normals, tot i que haurem de tenir en compte que sempre es tendirà a veure una major desviació als extrems.

A partir d'aquests gràfics podem saber la causa de la desviació, si els punts es disposen en forma de U o amb alguna corba, la distribució és asimètrica[5], mentre que si es presenten en forma de S, significarà que la distribució no és mesocúrtica[6].

Analitzem cada variable separadament per cada grup d'interès, i a simple vista observem que les variables segueixen la mateixa distribució pels 3 grups. Podríem acceptar únicament la normalitat de la variable Edat, tant per CL com per MP i SA. La resta de variables, seguirem suposant que no segueixen una distribució normal, podríem suposar que segueixen una distribució asimètrica positiva.

Gràfic 4.2. Gràfics QQ de les variables contínues segmentades per hotel

[5] Mesura que indica la simetria de la distribució d'una variable respecte la mitjana aritmètica
[6] Concentració normal de les dades al voltant de la seva mitjana.



Tindrem en compte que si la distribució és normal multivariant, cada una de les variables segueix una distribució normal univariant, però no a la inversa. Tot i haver descartat normalitat univariant, a continuació comprovarem la normalitat multivariant.

2.2 Normalitat multivariant

Fins ara hem pogut veure que, excepte per la variable edat, no podem suposar que les altres variables segueixin una distribució normal univariant, així que tal com hem

comentat, serà difícil que les variables conjuntament de cada grup segueixin una distribució normal multivariant.

La comprovació de la normalitat multivariant la realitzarem a partir del paquet de R "`mvn`". S'utilitzarà el test de Henze-Zirkler utilitzant `mvnTest = "hz"`. Aquesta prova multivariant de R permet un màxim de 5000 observacions així que s'ha escollit una mostra aleatòria de cada grup de 1000 observacions, i s'ha realitzat una iteració del test fins l'última observació de la base de dades de cada hotel.

L'última columna de la sortida de R d'aquest test ens indica si les dades segueixen o no una normalitat multivariada a un nivell de significació del 5%. Els valors crítics de HZ que trobem a la taula representen la mitjana dels valors crítics de totes les iteracions de cada grup:

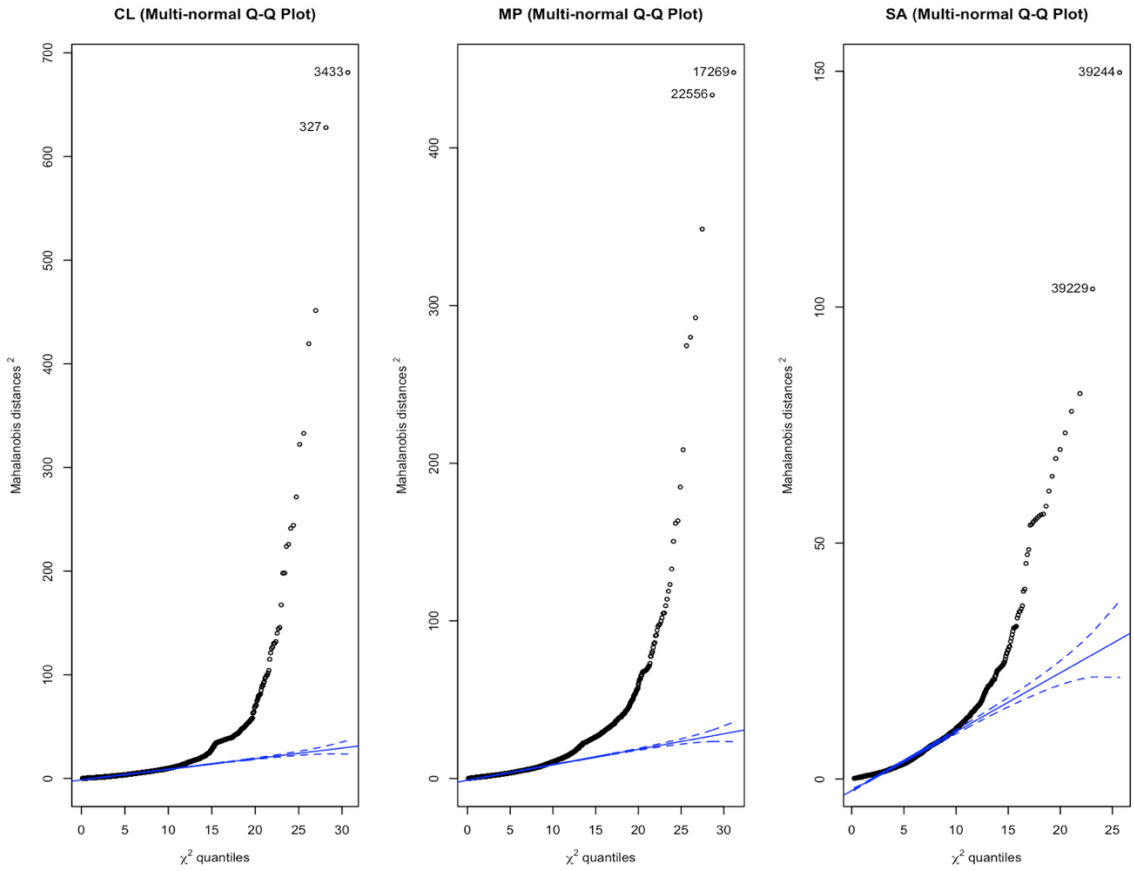
Taula 4.2. Resultats test de Henze-Zirkler

Hotel	***HZ***	***p-valor***	***MVN***
CL	14.65	0	✖
MP	11.81	0	✖
SA	18.50	0	✖

En cap de les iteracions el test de Henze-Zirkler ens indica que les dades segueixin una distribució normal multivariant, per tant, de moment rebutgem que les variables segueixin una distribució normal multivariant.

A més, utilitzarem el paquet de R `RVAideMemoire` amb la funció `mqqnorm()` per acabar d'afirmar que les dades no seguiran una distribució normal multivariant per cap dels grups.

Gràfic 4.3. Gràfics QQ de les variables contínues segmentats per hotel



Tot i l'incompliment d'aquest suposat bàsic, l'anàlisi discriminant és considerada una tècnica robusta i no es veurà greument afectada per la falta de normalitat, tot i que tindrem en compte aquests incompliments a l'hora d'interpretar els resultats.

3. Homoscedasticitat

L'homogeneïtat de variàncies és una considerada una condició molt important per a la realització d'anàlisis de variàncies, com ANOVA, tècniques de regressió i anàlisis discriminants. A continuació analitzarem l'homoscedasticitat univariant i multivariant com a contrast previ abans de la realització de l'anàlisi discriminant.

3.1 Homoscedasticitat univariant

L'estadística ens diu que un model predictiu presenta homoscedasticitat quan la variància de l'error de les variables explicatives és constant al llarg de les observacions o dit d'una altra manera, si l'error comès pel model té sempre la mateixa variància.

En bases de dades on la grandària de la mostra no està equilibrada, la falta d'homoscedasticitat té major impacte; si els grups de menor grandària són els que presenten una desviació estàndard major, augmentarà el nombre de falsos positius, si pel contrari els grups d'una grandària major, tenen major desviació estàndard, augmentaran els falsos negatius.

El supòsit d'homogeneïtat de variàncies, supòsit d'homoscedasticitat, considera que la variància és constant entre diferents grups. Per tal de veure si les dades del nostre estudi compleixen aquest supòsit utilitzarem el test de Levene. Es tracta d'un test que s'utilitza per comprovar que les diferències són iguals per a totes les mostres. Aquest test ens permet poder comparar 2 o més poblacions i poder escollir entre diferents estadístics de centralitat (mediana, mitjana), fet important a l'hora de contrastar aquesta hipòtesi quan els grups es distribueixen de manera normal o no normal.
En el cas que tinguéssim indicis que les dades provenen d'una distribució normal hauríem d'utilitzar el test de Barlett, però tal com hem observat a l'apartat anterior no podem suposar que les variables segueixin una distribució normal i per tant utilitzarem el test de Levene.

Plantegem dues hipòtesis:

$$\begin{cases} H_0: \sigma_{CL}^2 = \sigma_{MP}^2 = \sigma_{SA}^2 \\ H_1: \text{la variància no és constant per tots els grups.} \end{cases}$$

Una vegada hem realitzat el test de Levene per totes les variables a R mitjançant la funció `leveneTest()`, podem rebutjar totes les hipòtesis nul·les, i per tant acceptarem que hi ha diferències significatives entre les variàncies de les poblacions. En veiem un breu resum en format de taula per observar els resultats:

Taula 4.3. Resultats test de Levene

Hotel	***F value***	***p-valor***	***Rebuig H_0***
Edat	93.34	< 2.2e-16	✓
Antelació	49.991	< 2.2e-16	✓
Persones	140.08	< 2.2e-16	✓
RN	2.0297	< 2.2e-16	✓
AR	144.8	< 2.2e-16	✓
RR	1201.8	< 2.2e-16	✓

3.2 Homoscedasticitat multivariant

Després de realitzar el test per acceptar homoscedasticitat univariant, assumim heteroscedasticitat. Igualment, comprovarem l'homoscedasticitat multivariant, ja que en el nostre estudi, aquesta condició ens ajudarà a saber si és millor utilitzar LDA o QDA, per tant realitzarem la hipòtesi d'homoscedasticitat multivariant que ens diu que, la variància de la variable dependent és constant en els grups definits pels factors.

La variabilitat de les dades i les relacions lineals entre les variables es poden resumir a la matriu de variàncies i covariàncies. Aquesta matriu és quadrada i simètrica d'ordre k (en el nostre cas k=6, que és el nombre de variables considerades), on els termes en diagonal són les variàncies i les no diagonals, les covariàncies entre les variables. Anomenant la matriu com a S, tindrem per definició:

$$S = \begin{bmatrix} s_1^2 & s_{12} \cdots s_{16} \\ \vdots & \vdots \ddots \vdots \\ s_{61} & s_{62} \cdots s_6^2 \end{bmatrix}$$

sent

$$S_{jj'} = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x_j})(x_{ij'} - \bar{x_{j'}})$$

la covariància mostral entre les variables x_j , $x_{j'}$. Tenint en compte que x_{ij} = x_j (ω_i) l'observació de la variable x_j sobre l'individu ω_i i $\bar{x_j} = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$.

Aplicant les nostres dades obtindrem les següents matrius de variàncies – covariàncies de cada un dels hotels mitjançant la funció `cov()` d'R :

Taula 4.4. Matriu de variàncies – covariàncies de l'hotel CL

CL	*edat*	*antelacio*	*persones*	*RN*	*AR*	*RR*
edat	243.79	160.70	-0.24	16.11	847.34	2466.16
antelacio	160.70	7817.52	-0.11	90.19	3054.84	13312.79
persones	-0.24	-0.11	0.25	0.00	3.38	7.09
RN	16.11	90.19	0.00	16.75	739.65	2097.31
AR	847.34	3054.84	3.38	739.65	85340.42	109656.70
RR	2466.16	13312.79	7.09	2097.31	109656.70	420067.57

Taula 4.5. Matriu de variàncies – covariàncies de l'hotel MP

MP	*edat*	*antelacio*	*persones*	*RN*	*AR*	*RR*
edat	270.46	85.17	-0.10	15.90	735.07	1768.02
antelacio	85.17	6705.16	0.37	88.51	3406.56	9026.04
persones	-0.10	0.37	0.39	0.05	5.76	5.88
RN	15.90	88.51	0.05	18.35	776.11	1749.00
AR	735.07	3406.56	5.76	776.11	71826.71	86060.44
RR	1768.02	9026.04	5.88	1749.00	86060.44	230751.81

Taula 4.6. Matriu de variàncies – covariàncies de l'hotel SA

SA	*edat*	*antelacio*	*persones*	*RN*	*AR*	*RR*
edat	215.12	53.95	0.16	15.38	1032.33	4576.83
antelacio	53.95	7561.21	-0.31	70.17	4153.11	17755.65
persones	0.16	-0.31	0.18	0.02	1.73	9.50
RN	15.38	70.17	0.02	14.11	795.15	3891.54
AR	1032.33	4153.11	1.73	795.15	186658.90	241223.40
RR	4576.83	17755.65	9.50	3891.54	241223.40	1462632.31

Per comprovar aquest fet en les nostres dades, i tenir una primera idea de com podem obtenir millors resultats, si amb discriminants lineals o discriminants quadràtics, ens centrarem a comprovar, mitjançant la prova M de box del paquet `biotools`, la igualtat de les matrius de variància-covariància múltiple.

La prova M de box és un test paramètric que es fa servir per comparar la variació en mostres multivariades. Hem de tenir en compte que l'incompliment de la normalitat multivariant afecta aquest test, amb el que comprovarem l'homoscedasticitat de les matrius de variàncies-covariàncies. A més, l'acceptació d'homoscedasticitat és difícil, amb menor importància en estudis de mostres grans, com és el nostre cas, i superable amb l'ús del model en forma quadràtica a l'hora d'estimar-lo.

Plategem dues hipòtesis:

H_0: la matriu de covariància és constant per tots els grups.
H_1: la matriu de covariància no és constant per tots els grups.

Després de realitzar el test M de box, podem rebutjar la hipòtesi nul·la d'homogeneïtat de les matrius de covariàncies. N'observem el resultat a partir de la sortida de R:

Taula 4.7. Resultat del test M de box

	Chi-Sq	***p-valor***	***Rebuig H_0***
Box's M-test for homogeneity of Covariances matrix	3757.9	< 2.2e-16	✓

Per tant, serà més adequada la utilització de la discriminació quadràtica que la discriminació lineal.

4. Variables fictícies

En la majoria dels casos, en una base de dades hi trobem presència de variables quantitatives i qualitatives. En el cas que sigui una tècnica per variables mètriques, com és el cas de l'anàlisi discriminant, i hi hagi presència de variables qualitatives, per tal de no prescindir d'informació rellevant, es representaran amb l'ajuda de variables fictícies[7].

Els predictors categòrics que volem introduir a l'anàlisi discriminant són: sexe, canal, règim, dia d'alta, mes d'alta, dia d'entrada i mes d'entrada.

Per tal d'incloure aquestes variables a l'anàlisi discriminant les convertirem en factors mitjançant R i la funció "`as.factor()`". Una vegada tinguem les variables categòriques com a factors, tant la funció "`lda()`" com "`qda()`" que utilitzarem per realitzar l'anàlisi discriminant lineal i quadràtic, respectivament, ja processaran de manera automàtica els predictors categòrics i aquestes variables seran tractades com a fictícies.

Per tant, les variables predictores que finalment estudiarem seran les següents: edat, sexe, canal, dia alta, mes alta, antelació, dia entrada, mes entrada, règim, hostes per reserva, RN, RR i AR.

[7] Una variable fictícia és una variable utilitzada per explicar valors qualitatius en models de regressió.

5. Base de dades en submostres

Des d'un primer moment s'ha especificat la selecció de la variable dependent i de les variables independents. La variable dependent és la variable hotel, i té un total de 3 grups.
Coneixem que la mostra mínima per aquest tipus d'anàlisis és de 5 observacions per la variable dependent, tot i que es recomana tenir-ne més de 20. La mostra ha de ser representativa de la població i la grandària del grup més petit ha de ser major que el nombre de variables independents. En el nostre cas d'estudi tenim 21.100 observacions per MP, 16.970 observacions per CL i 1.972 per SA.

A més, per tal d'evitar errors de classificació, separarem la mostra en dues submostres: un 80% de les dades s'utilitzaran com a mostra d'anàlisis i el 20% restant es deixarà per una segona etapa de validació del model. Es recomana mantenir les proporcions de la població per mantenir-ne la representativitat.

En el cas que utilitzéssim la mateixa base de dades per estudiar l'anàlisi discriminant i també per comprovar-ne els resultats, obtindríem l'anomenat *training error*. Per tant, serà més adequada la divisió d'aquestes dades de manera aleatòria per tal de poder comprovar els resultats de l'estudi amb observacions que no s'hagin estudiat anteriorment.

V. APLICACIÓ EN R

Una vegada hem detallat la metodologia de l'anàlisi que volem aplicar i hem analitzat les dades de l'estudi, aprofundirem l'anàlisi discriminant lineal i l'anàlisi discriminant quadràtic. Coneixem que les dades no segueixen una distribució normal multivariant i que la matriu de variàncies – covariàncies no és constant per a tots els grups, així que segons la metodologia d'aquestes dues anàlisis el més adequat seria l'anàlisi discriminant quadràtic. Tot i això, aplicarem les dues anàlisis per veure quin dels dos mètodes és el més òptim per a les nostres dades. A més, també en comprovarem les seves capacitats predictives.

Tal com s'ha treballat en la resta de l'estudi, el software escollit per analitzar aquestes dades i realitzar l'anàlisi discriminant és R.

En l'apartat anterior, havíem dividit la mostra en dues submostres: 80% de les observacions per l'entrenament de l'algoritme i el 20% restant com a submostra de test per comprovar-ne la capacitat predictiva del model i així evitar el *training error*. A més a més, també tenim convertides les variables qualitatives en factors per tal de poder ser tractades com a variables fictícies.

Recordem les variables del model:

- Variable dependent: hotel
- Variables predictores: edat, antelació, persones, RN, AR, RR, sexe, canal, regim, dia d'alta, mes d'alta, dia d'entrada, mes d'entrada.

1. Anàlisi discriminant lineal

Tot i la dificultat que podríem considerar que té la definició de l'anàlisi discriminant, realitzar aquesta anàlisi a R és més senzill gràcies al paquet `MASS` a partir de la funció `lda()`. Aquesta funció segueix el criteri de la discriminació lineal de Fisher.

Les funcions discriminants obtingudes com a combinacions lineals de les variables explicatives de l'anàlisi discriminant, permeten la classificació dels individus de la mostra en els grups definits per la variable dependent, a través d'establir un punt de tall per les puntuacions calculades a partir de la funció corresponent.

Tenim que la funció `dades_estudi` escull de manera aleatòria un 80% de les dades de la base de dades. Per tant, aquesta és la nova base de dades que utilitzarem per a aquesta part de l'estudi.

```
funciolda <- lda(hotel~ ., data=dades_estudi)
funciolda
```

La sortida de R ens indica que les probabilitats a priori són:

Taula 5.1. Probabilitats a priori

Probabilitats a priori		
$P(\Omega_{CL})$	$P(\Omega_{MP})$	$P(\Omega_{SA})$
0.42	0.53	0.05

També ens proporciona les mitjanes dels grups de cada variable i els coeficients discriminants lineals de la combinació entre les variables que s'utilitzen per crear la regla de dedició de LDA.

L'objecte `funciolda` conté tota la informació rellevant per a l'anàlisi. Els coeficients de la funció discriminant lineal de Fisher els extraurem mitjançant `funciolda$scaling`. A partir d'aquests coeficients, que són les ponderacions assignades a les variables, trobem les funcions discriminants. Això significa que el límit entre les classes diferents s'especificarà mitjançant la següent fórmula:

Y1= -0.0069 * edat + 0.0004 * antelacio + 0.1570 * persones + ... + 0.0741 * mes_alta12
Y2 = 0.0291 * edat + 0.0053 * antelacio - 0.2982 * persones + ... + 0.1259 * mes_alta12

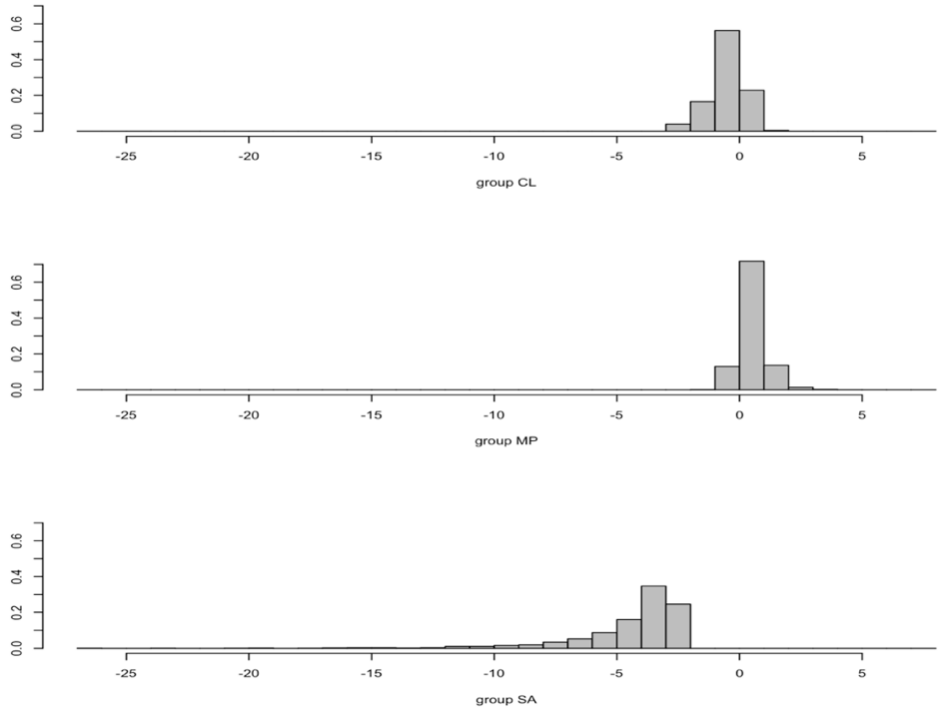
Una bona manera de mostrar els resultats de l'anàlisi discriminant lineal és fer un histograma apilat dels valors de la funció discriminant per les observacions dels diferents grups. Podem fer-ho mitjançant la funció `ldahist()`. El valor de cada funció discriminant s'escala de manera que el seu valor mitjà és 0. Però per la realització d'aquest histograma necessitarem executar la funció `predict()`.

```
pred<- predict(funciolda, data=dades_estudi)
```

Observem primer l'histograma apilat dels valors de la primera funció discriminant per a les observacions dels tres hotels, escrivim:

```
ldahist(pred$x[,1], g= pred$class, col=c("grey"))
```

Gràfic 5.1. Histograma apilat dels valors de LDA1



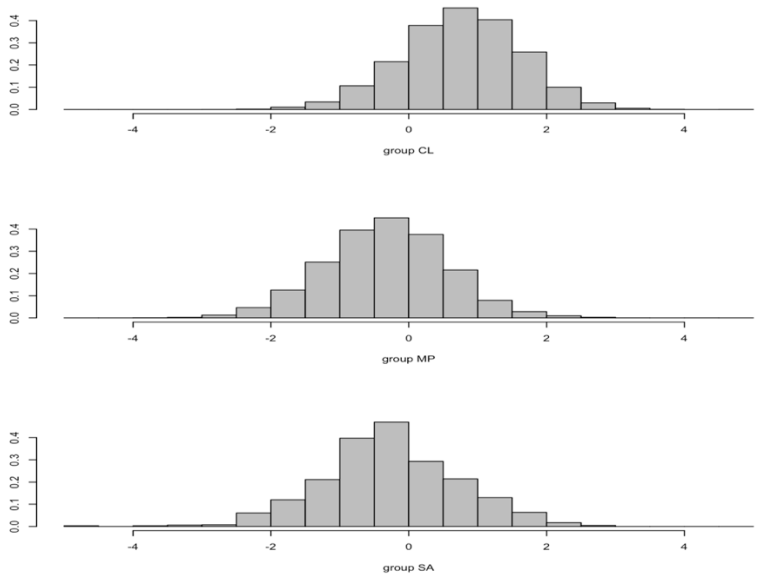
Podem veure que els valors de la funció discriminant de les observacions de CL se superposen majoritàriament amb els de MP, en aquestes zones hi podrem veure un augment de la taxa de l'error. En canvi, els valors de les observacions de SA quasi no se superposen amb els altres dos grups.

En molts casos els valors de la funció discriminant de les observacions separen correctament els grups. Hi ha una zona en què els tres grups se superposen, en aquesta zona s'incrementarà el valor de les taxes d'error.

Observem també els valors de la segona funció discriminant de les observacions de cada grup mitjançant

```
ldahist(pred$x[,2], g= pred$class, col=c("grey"))
```

Gràfic 5.2. Histograma apilat dels valors de LDA2



Aquesta segona funció discriminant separa millor les observacions de CL. Observem que els valors d'aquesta funció discriminant de les observacions de MP i SA estan superposades.

Ara utilitzarem la funció `predict()`, que utilitza els resultats de `lda()` per assignar les observacions als grups per la base de dades de test. És a dir, ja que `lda()` va derivar de la funció lineal que hauria de classificar els grups, `predict()` ens permet aplicar aquesta funció a unes noves dades de test per veure l'èxit de la classificació.

```
pred_lda <- predict(object = funcio_lda, dades_prediccio)
```

La sortida comença amb les classificacions assignades a cadascuna de les observacions. A continuació, enumera les probabilitats posteriors de cada observació a cada hotel, i les probabilitats per a cada observació, que sumen 1. Aquestes probabilitats mesuren la força de la classificació. Si una d'aquestes probabilitats per a una observació és molt superior a les altres, aquesta mostra s'assignarà a un grup amb un alt grau de certesa. Si les tres probabilitats són gairebé iguals, l'assignació serà menys segura.

N'observem les 5 primeres observacions:

Taula 5.2. Probabilitats a posteriori LDA

	CL	MP	SA
5	0.2258	0.7742	0.0000
7	0.4439	0.5560	0.0001
10	0.5801	0.4178	0.0021
14	0.6247	0.3751	0.0001
21	0.4929	0.5068	0.0003

1.1 Capacitat predictiva

Gràcies a la base de dades amb la que hem treballat, coneixem a quin grup pertany cada individu. Anteriorment, havíem segmentat la base de dades en dues submostres per tal de tenir una base de dades d'estudi, que conté un 80% de les dades, i una altra base de dades de test que conté el 20% restant. Per conèixer la capacitat predictiva de l'anàlisi discriminant s'utilitza la base de dades de test.

La capacitat predictiva de les funcions discriminants l'avaluarem a partir de la matriu de classificació de la mostra, en la que es recullen els valors estimats per la variable dependent i els estimats pel model.

Les matrius de confusió són una de les millors tècniques per avaluar la capacitat predictiva que té el model LDA. Aquestes matrius ens mostren el número de positius verdaders, negatius verdaders, falsos positius i falsos negatius. LDA busca aconseguir el mínim % d'error però no diferencia entre falsos positius o falsos negatius.

A partir de la funció `CrossTable()` i mitjançant una regla que ens calculi el percentatge d'observacions mal classificades podrem saber l'error de classificació, i per tant podrem conèixer quin percentatge d'observacions s'ha classificat correctament.

Taula 5.3. Matriu de confusió LDA

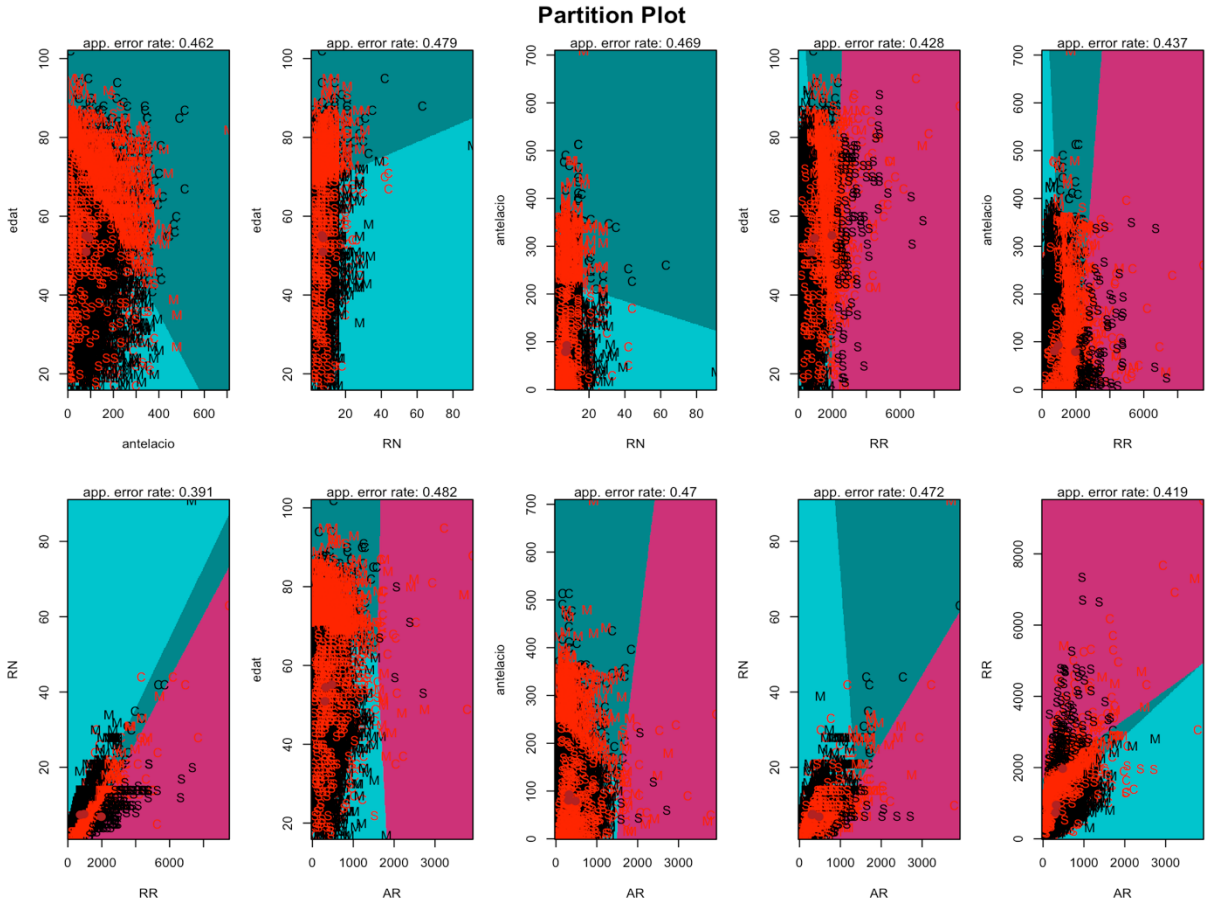
		Classe verdadera		
Classe predita	**CL**	**MP**	**SA**	**Total**
CL	1476	777	122	2375
	62.15%	32.71%	5.14%	29.65%
MP	1829	3453	15	5297
	34.53%	65.19%	0.28%	66.15%
SA	85	8	243	336
	25.30%	2.38%	72.32%	4.2%
Total	3390	4238	380	8008

En el cas de CL podem veure que el nombre d'encerts és superior al nombre d'errors, assignant-se correctament al seu grup un 62.15% de les observacions. El grup amb el qual més confusió hi ha és amb observacions que s'assignen a MP.

Les observacions de MP tenen un grau d'encert més elevat que les de CL, amb un 65.19%, tot i que la confusió també és més elevada amb observacions que s'assignen a CL quan realment són de MP.
Pel que fa a les observacions de SA, són les que millor s'han classificat al grup que els hi pertoca. Un 72.32% de les observacions s'han classificat correctament. Un 25.30% de les observacions de SA s'han assignat errròniament a CL.

A més, podem observar mitjançant les funcions discriminants mitjançant el paquet `klaR` i la funció `partimat()`. Les regions pintades de color representen cada àrea de classificació. A més hi podem veure la taxa d'error aparent de les observacions, mentre que cada observació està indicada amb C, M, o S, segons la classe de la qual prové (CL, MP i SA, respectivament). Les observacions pintades en vermell representen una classificació errònia, mentre que les observacions en negre s'han classificat correctament al seu grup.
Mostrem només els gràfics entre les variables quantitatives contínues.

Gràfic 5.3. Gràfics de classificació LDA de les observacions



Pel model corresponent a la mostra dels individus d'estudi hem aconseguit un percentatge total de classificació correcta del 64.6%. Per tant, mitjançant aquesta tècnica un 35.4% dels individus serien classificats de manera incorrecta.

A continuació, mitjançant una transformació de les variables contínues es buscarà reduir la taxa d'error de l'anàlisi discriminant lineal.

1.2 Transformació de les variables contínues

Tal com hem comentat al llarg de l'estudi, per tal d'aconseguir un resultat més òptim s'hauria de complir el suposat bàsic de la normalitat multivariant de les variables contínues.

L'asimetria estadística és una mesura de simetria per una distribució. El valor pot ser negatiu, positiu o pot no estar definit. En una distribució esbiaixada, les mesures de tendència central (mitjana, mediana...) no seran iguals. La direcció d'asimetria és definida pel signe del coeficient d'asimetria: un zero significaria que no hi ha asimetria (distribució normal), un valor negatiu significaria que la distribució està esbiaixada negativament i un valor positiu que la distribució està esbiaixada positivament.
Ja havíem observat que les variables del nostre estudi no seguien una distribució normal. Per tant, volem observar si amb unes variables contínues seguint una normalitat multivariant milloraríem la taxa d'encert. Transformem aquestes variables amb logaritmes per comprovar-ho.

Utilitzarem la funció `skewness()` del paquet `moments` per avaluar el signe de l'asimetria. Segons el signe, realitzem les següents transformacions:

- Log10(x) per les variables esbiaixades positivament: edat
- Log10(max(x+1)-x) per les esbiaixades negativament: antelacio, RN, RR i AR.

Observem els valors de l'asimetria abans i després de fer la transformació logarítmica a cada variable:

Taula 5.4. Valors d'asimetria

	VARIABLE	**VARIABLE TRANSFORMADA**
EDAT	-0.16	-0.74
ANTELACIÓ	1.41	-0.97
RN	2.32	-1.03

RR	3.10	-0.86
AR	3.05	-0.63

Observem que els valors d'asimetria de les variables transformades són més pròxims a zero. Això significa que les variables seran més properes a seguir una distribució normal. Pel cas de la variable edat, el valor d'asimetria s'allunya més de 0, per tant conservarem aquesta variable sense la transformació.

Ara, tornem a executar `lda()` a la base de dades d'estudi però substituint les variables contínues, excepte la variable edat, per les seves transformacions i comprovem mitjançant la matriu de confusió amb les dades de test que els resultats obtinguts són millors.

Taula 5.5. Matriu de confusió LDA amb les variables contínues transformades

	Classe verdadera			
Classe predita	**CL**	**MP**	**SA**	**Total**
CL	1773	878	55	2706
	65.52%	32.45%	2.03%	33.80%
MP	1542	3353	3	4898
	31.48%	68.46%	0.06%	61.16%
SA	75	7	322	404
	18.56%	1.73%	79.70%	5.04%
Total	3390	4238	380	8008

Sent la taxa d'encerts 68.03%. Recordem que la taxa d'encert de l'anàlisi discriminant amb les variables antelació, RN, RR i AR sense transformació era de 62.15%, per tant, amb la transformació de les variables hauríem reduït en un 5.88% la taxa d'error.

Després de comprovar que amb les variables contínues transformades aconseguim un millor resultat de l'anàlisi discriminant lineal, i considerant que l'anàlisi quadràtic és més robust a la falta de normalitat, en un primer lloc realitzarem l'anàlisi discriminant quadràtic i n'avaluarem la capacitat predictiva. Posteriorment, realitzarem la mateixa anàlisi amb les variables contínues transformades, per tal de veure si s'aconseguirà una millor taxa d'encerts.

2. Anàlisi discriminant quadràtic

Per obtenir les funcions discriminants hem utilitzat la funció `qda()` del paquet `MASS`. La sintaxi i l'output que obtenim amb aquesta funció és molt similar que la funció LDA, amb l'única diferència que no obtindrem els coeficients dels discriminadors lineals, ja que el classificador QDA, tal com hem comentat anteriorment, implica una funció quadràtica dels predictors, en lloc de lineal.

El que ens indica la sortida del model de QDA són les probabilitats a priori i les mitjanes de cada variable en cada grup, i podem observar que coincideixen amb les de LDA.

```
funcioqda <- qda(hotel~ ., data=dades_estudi)
pred_qda <- predict(object = funcio_qda, dades_prediccio)
```

La funció `predict()` aplicada al model conté informació de la classificació predita pel model. Podrem observar les probabilitats a posteriori de que les observacions pertanyin a cada classe. S'assigna l'observació al grup amb major probabilitat a posteriori. N'observem les 5 primeres:

Taula 5.6. Probabilitats a posteriori QDA

	CL	MP	SA
5	0.0318	0.9682	0.0000
7	0.4038	0.5961	0.0001
10	0.9647	0.0233	0.0121
14	0.8557	0.1440	0.0003
21	0.8236	0.1764	0.0000

2.1 Capacitat predictiva

Utilitzem la mateixa tècnica utilitzada en el cas de LDA: la matriu de classificació. Tal com hem comentat al subapartat de capacitat predictiva de LDA, una forma de valorar la bondat de classificació de cada observació és aplicar el procediment als casos pels quals coneixem el seu grup, i comprovar si coincideix el grup observat amb el grup predit. El percentatge de casos correctament classificats indicarà la correcció del procediment.

La matriu de classificació, o matriu de confusió permet presentar, per cada grup, quants d'ells s'esperaven en aquell grup i quants en la resta de grups. D'aquesta manera, podrem saber quin tipus d'error de classificació es produeixen.

Utilitzant la mateixa funció que per LDA, visualitzem la matriu de confusió:

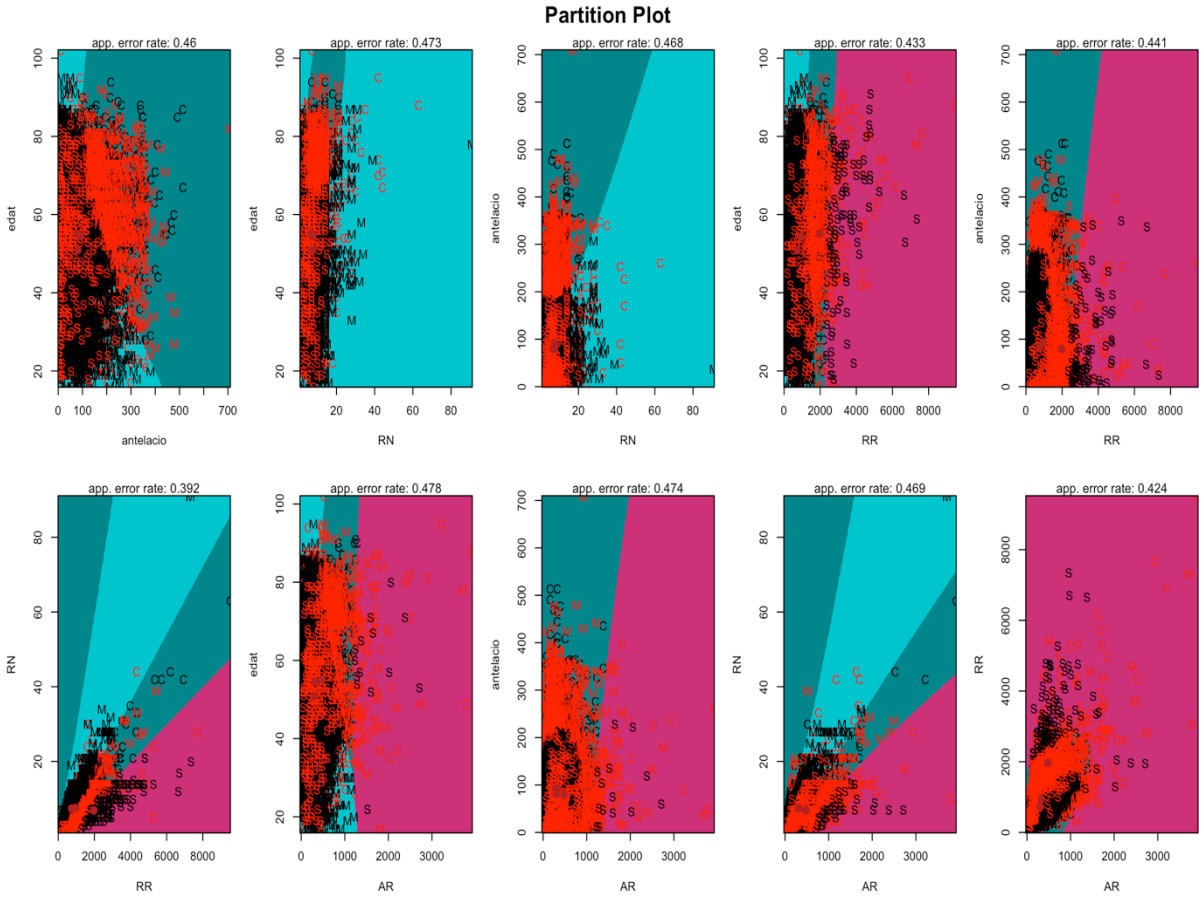
Taula 5.7. Matriu de confusió QDA

		Classe verdadera		
Classe predita	**CL**	**MP**	**SA**	**Total**
CL	1844	1221	98	3163
	58.30%	38.60%	3.10%	39.5%
MP	1389	2919	12	4320
	32.15%	67.57%	0.28%	53.95%
SA	157	98	270	525
	30.00%	18.67%	51.43%	6.55%
Total	3390	4238	380	8008

Observem que amb QDA, igual que per LDA, el nombre d'encerts continua sent major al nombre d'errors en cada grup. Disminueix el nombre d'encerts en CL i SA però augmenta en MP.

Podem observar visualment les funcions discriminants mitjançant el paquet `klaR` i la funció `partimat()` de les variables quantitatives continues visualitzades anteriorment per LDA.

Gràfic 5.4. Gràfics de classificació QDA de les observacions



Després de veure les taxes d'error, veiem a la taxa d'error total que no s'aconsegueix millora amb l'anàlisi discriminant quadràtic, ja que és capaç de predir correctament un total de 62.85% d'observacions, i per tant s'estaran classificant malament un 37.15% de les observacions.

2.2 Transformació de les variables contínues

De moment, podem afirmar que per la nostra base de dades de test l'anàlisi discriminant quadràtic aconseguirà una taxa d'encert similar a l'anàlisi discriminant lineal. En canvi, si comparem el resultat de QDA amb LDA amb les variables contínues transformades, la taxa d'errors de classificació de QDA serà més elevada.

Realitzem l'anàlisi discriminant quadràtic amb les variables edat, antelació, RN, AR i RR transformades a la base de dades d'estudi i n'avaluem la matriu de confusió de la base de dades de test amb les variables contínues transformades també:

Taula 5.8. Matriu de confusió QDA amb les variables contínues transformades

	Classe verdadera			
Classe predita	**CL**	**MP**	**SA**	**Total**
CL	2113	1499	92	3704
	57.05%	40.47%	2.48%	46.25%
MP	1120	2678	3	3801
	29.47%	70.46%	0.07%	47.46%
SA	157	61	285	503
	31.21%	12.13%	56.66%	6.28%
Total	3390	4238	380	8008

Recordem que la taxa d'encert de QDA sense la transformació de les variables contínues era de 63.39%, i veiem que la taxa d'encert amb les variables transformades és de 63.12%. Per tant, en el cas de l'anàlisi discriminant quadràtic amb la transformació de les variables antelació, RN, RR i AR només haurem aconseguit una reducció de 0.26% de la taxa d'error de classificació.

3. K-NN veïns més pròxims

La funció per aplicar el mètode de k veïns més pròxims és `knn()` del paquet `class`. Primer carregarem el paquet mitjançant `library(class)`. Aquest mètode de classificació també serà realitzat a partir de les submostres: 80% de les dades d'estudi i el 20% restant per avaluar el model.

Volem seguir incloent totes les variables, qualitatives i quantitatives, en aquesta anàlisi, igual que havíem fet en l'anàlisi discriminant. Incloíem les variables qualitatives en LDA i QDA mitjançant aquestes mateixes funcions, que tractaven de manera automàtica aquestes variables com a fictícies. Això no passa amb la funció `knn()`. Per tal d'incloure aquestes variables en aquesta anàlisi o fem mitjançant la funció `dummy_cols()` del paquet `fastdummies`, tant per la base de dades d'estudi, com per la base de dades de predicció.

Un cop incloses les variables qualitatives com a fictícies en el model, realitzarem l'anàlisi escollint diferents valors de K, i guardarem els resultats al següent objecte, per tal de posteriorment poder comprovar amb quin valor de K aconseguim una millor predicció de les dades.

```
KnnTest_k1 <- knn(dades_estudi[,2:55],
dades_prediccio[,2:55], dades_estudi$hotel ,
k = 1, prob = TRUE)
```

El resultat és un vector que ens dóna en quin grup s'inclou cada observació. Per calcular la taxa d'error, tornem a comparar aquests resultats amb les classificacions verdaderes, igual que hem fet en LDA i QDA.

Per tal de seleccionar amb quin valor de k, seguim la següent condició:

$$K = n^{1/2} = 8008^{1/2} = 89$$

A més es recomana considerar els següents punts:

- El valor de K ha de ser imparell.
- El valor de K no pot ser múltiple del nombre de classes.
- No pot ser un nombre ni molt petit ni molt gran.

Un cop hem vist que el valor trobat a partir de la condició també segueix les recomanacions, apliquem l'algoritme k-NN seleccionant uns quants valors senars pròxims de k.

Taula 5.9. Taxes d'encert per k-NN

VALORS DE K	TAXES D'ENCERT
79	59.65%
83	59.49%
89	59.51%
95	59.60%
97	59.59%

Observem que les taxes d'encert pels diferents valors de k són molt semblants, tot i això ens quedarem amb k=79, ja que és el valor de k on aconseguim una millor taxa d'encert.

3.1 Capacitat predictiva

Tal com hem vist en l'anàlisi discriminant, la tècnica més utilitzada per resumir el rendiment dels algoritmes de classificació és la matriu de confusió.

Taula 5.10. Matriu de confusió K-NN amb k = 79

		Classe verdadera		
Classe predita	**CL**	**MP**	**SA**	**Total**
CL	1454	1003	212	2669
	54.48%	37.57%	7.95%	33.32%
MP	1873	3216	71	5160
	36.30%	62.32%	1.38%	64.43%
SA	63	19	97	179
	35.20%	10.61%	54.19%	4.15%
Total	3390	4238	380	8008

Observem, que per cada grup la taxa d'encert és més elevada que la taxa d'error de classificació. És a dir, per les classificacions de CL, el 54.48% de les observacions hauran estat ben classificades, per les de MP, el 62.32% i per SA el 54.19%.

3.2 Transformació de les variables contínues

Tal com hem fet en les anàlisis discriminants lineal i quadràtic, realitzem l'anàlisi k-NN veïns més pròxims amb les variables antelació, RN, RR i AR transformades. Seguim treballant amb k=79. Observem la matriu de confusió i la taxa d'encert obtinguda:

Taula 5.11. Matriu de confusió K-NN amb k = 79

		Classe verdadera		
Classe predita	**CL**	**MP**	**SA**	**Total**
CL	1360	924	247	2531
	53.74%	36.50%	9.76%	31.60%
MP	2030	3314	132	5476

	37.07%	60.52%	2.41%	68.38%
SA	0	0	1	1
	0%	0%	100%	0.02%
Total	3390	4238	380	8008

I obtenim una taxa d'encert de 58.37%. Recordem que la taxa d'encert obtinguda per k-79 veïns més pròxims sense la transformació de les variables nomenades anteriorment era de 59.65%. Per tant, estarem obtenint una taxa d'encert menor, és a dir un resultat menys favorable.

Un cop hem realitzat les anàlisis, procedim a fer una comparativa entre tots els per poder-ne destacar amb quin hem obtingut un millor resultat.

4. Comparativa LDA, QDA i K-NN

De tots els mètodes utilitzats amb la nostra base de dades, l'anàlisi discriminant lineal és el que mostra un test error menor. A més, si observem el resultat de l'anàlisi discriminant lineal amb la transformació de les variables antelació, RN, RR i AR, veiem que encara és millor, obtenint una taxa d'encerts del 68.03% d'observacions ben classificades.

A la següent taula hi observem tots els percentatges d'encert dels diferents mètodes utilitzats:

Taula 5.12. Taxes d'encert pels diferents mètodes discriminants

	PERCENTATGE D'ENCERT
LDA	64.60%
LDA -TRANSFORMACIÓ	68.03%
QDA	62.85%
QDA – TRANSFORMACIÓ	63.39%
K-NN (K=79)	59.65%
K-NN – TRANSFORMACIÓ	58.37%

Destaquem que pels casos de les anàlisis discriminant lineal i quadràtic és millor el mètode amb la transformació de les variables antelació, RN, RR i AR. No passa el mateix pel cas de

l'anàlisi K-NN veïns més pròxims, que tot i que les taxes d'encert són molt semblants, obtenim una millor classificació amb les variables sense transformacions.
Les pitjors taxes d'encert les obtenim amb l'anàlisi de K-NN veïns més pròxims.

Analitzem els resultats per taxes d'encert de classificació pels diferents mètodes i pels diferents grups, és a dir, comprovem les taxes d'encert que hem obtingut per CL, MP i SA per LDA, LDA-transformació, QDA, QDA-transformació, K-NN i K-NN-transformació.

Taula 5.13. Comparativa per mètodes i grups

	CL	**MP**	**SA**
LDA	62.15%	65.19%	72.32%
LDA -TRANSFORMACIÓ	65.52%	68.48%	79.70%
QDA	58.30%	67.57%	51.43%
QDA – TRANSFORMACIÓ	57.05%	70.46%	56.66%
K-NN (K=79)	54.48%	62.32%	54.19%
K-NN – TRANSFORMACIÓ	53.74%	60.52%	100%

Observem que les taxes d'encert més elevades han sigut pel cas de SA i per l'anàlisi discriminant lineal, amb i sense la transformació de les variables. Per la resta de taxes de SA s'obtenen percentatges baixos en comparació la resta d'anàlisis i de grups. En el cas de k-NN amb les variables transformades observem un 100% de classificació correcte, això és perquè la resta d'observacions que eren de SA s'han assignat erròniament a CL o MP.

Les observacions de MP es classifiquen més correctament que les de CL. A més, també tenen unes taxes d'encert més estables, trobant-se per tots les anàlisis dins de l'interval (60.52, 70.46).

Les taxes d'encert de CL també han sigut força similars per les diferents anàlisis, però amb uns valors menors que pels casos de MP i CL. Les trobem compreses entre (53.74, 65.52).

VI. CONCLUSIONS

A partir del que hem vist a l'apartat anterior, podem concloure que el millor anàlisi discriminant per a les nostres és l'anàlisi discriminant lineal.

Tal com hem comentat anteriorment, l'anàlisi discriminant és un anàlisi robust, encara més en el nostre cas que estàvem treballant amb una gran base de dades. Igualment, havíem vist que per l'anàlisi discriminant lineal és recomanable que les dades segueixin una normal multivariant i que la matriu de variàncies-covariàncies sigui igual per cada grup. A l'apartat de tractament de les variables, havíem comprovat que no podíem suposar normalitat multivariant de les nostres dades i podíem suposar que la matriu de variàncies-covariàncies és diferent per cada grup. Per tant, tot indicava a què la millor anàlisi per les nostres dades seria l'anàlisi discriminant quadràtic, un anàlisi discriminant encara més robust a l'incompliment de normalitat multivariant i més adequat per a casos en què la matriu de variàncies-covariàncies era diferent per cada grup. Tot i això, els resultats entre l'anàlisi discriminant lineal i l'anàlisi discriminant quadràtic són molt similars.

Després de realitzar les transformacions de les variables contínues per aproximar-nos a una normalitat multivariant, hem vist una millora més amplia en el cas de l'anàlisi discriminant, menys robusta que l'anàlisi discriminant quadràtic a la normalitat, aconseguint la taxa d'encerts més elevada de tot l'estudi.

Una de les hipòtesis a priori era que les diferències haurien de ser menors entre MP i CL, ja que es tracta de dos hotels amb un perfil bastant similar, en canvi, les diferències clares, i per tant una discriminació més diferenciada, s'haurien de veure amb SA respecta els altres dos.

Quedant-nos amb l'anàlisi discriminant lineal, que és amb el que obtenim una millor taxa d'encert, observem que les observacions amb millors classificacions són les de SA. Per tant, podríem afirmar, que tal com pensàvem en un principi, les reserves de SA es diferencien clarament a les de MP i CL. A més, també podem afirmar que les que seran més difícils de discriminar, ja que tenen les taxes d'encert més baixes, seran les observacions de CL, que tenen un alt percentatge de confusió, del 32.71%, amb MP. El mateix passa amb les observacions de MP que es confonen principalment amb CL.

Per tant, pel nostre cas podem afirmar que el compliment del suposat bàsic de normalitat multivariant de les variables ens ajudarà a aconseguir una millor taxa d'encert i per tant, podem dir que discriminarà millor entre les nostres observacions.

Encara que considero que els objectius principals del treball han estat resolts, m'hagués agradat treballar amb més mètodes de classificació per poder-ne comparar els resultats obtinguts. Com també m'agradaria inclouré més variables en l'estudi de referència als clients que realitzen les reserves. Aquestes no van poder ser incloses, ja que vam trobar-hi alguns problemes a l'hora de tractar i processar que no ens permetien avançar més.

Així, com a possible treball, a més d'incloure l'aplicació de nous mètodes i la incorporació de noves variables sobre els clients, també aplicaria aquests mètodes discriminants amb altres variables dependents, com podria ser molt interessant en aquest sector de l'hostaleria la cancel·lació o no cancel·lació d'aquestes reserves abans de ser gaudides.

VII. BIBLIOGRAFIA

Francisco Parra, 2017. Estadística y Machina Learning con R <https://rstudio-pubs-static.s3.amazonaws.com/293405_4029f1f23f834b7195189d5504a436b2.html>

NCSS Statistical Software. Discriminant Analysis <https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Discriminant_Analysis.pdf>

Luis Saucedo, 2019. Métodos de clasificación <https://rstudio-pubs-static.s3.amazonaws.com/504720_957df439bed24bba9ca7d3f2c5d30a91.html>

Yolanda Larriba González, 2014. Ánalisis discriminante <http://www.eio.uva.es/~valentin/am4g/2014/trabajos_alumnos_13-14/4%20yolanda%20larriba/HATCO.pdf>

Joaquin Aldas Manzano, Ezequiel Uriel Jimenez. Análisis multivariante aplicado con R

Santiago de la Fuente Fernández, 2011. Análisis discriminante <http://www.fuenterrebollo.com/Economicas/ECONOMETRIA/SEGMENTACION/DISCRIMINANTE/analisis-discriminante.pdf>

Jmmarin, 2006. Análisis discriminante lineal <http://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/DM/tema1dm.pdf>

Laura de la Fuente Crespo <http://www.estadistica.net/Master-Econometria/Analisis_Discriminante.pdf>

Joaquín Amat Rodrigo, 2016. Correlación lineal y regresión lineal simple <https://rpubs.com/Joaquin_AR/223351>

Salvador Figueras, M, 2000. Análisis discriminante <https://ciberconta.unizar.es/leccion/discri/100.HTM>

Sergio Ruiz Sánchez, 2016 <https://rstudio-pubs-static.s3.amazonaws.com/237547_0171c04b6d2e4550aea58853c056d29d.html#proceso-de-clasificacion-mediante-k-nn>

Gabriel Martos. *Research Techniques* <https://rstudio-pubs-static.s3.amazonaws.com/35817_2552e05f1d4e4db8ba87b334101a43da.html>

VIII. ANEXOS

```
setwd("~/Desktop")

library(readxl)
library(plotrix)
library(Epi)
library(descr)
library(catspec)
library(survival)
library(Hmisc)
library(psych)


bbdd_client <- read_excel("bbdd_client.xlsx")
bbdd <- bbdd_client


#missings
sum(is.na(bbdd))


hoste <- as.factor(bbdd$Huésped)
sexe <- as.factor(bbdd$Sexo)
edat <- as.numeric(bbdd$Edad)
alta <- as.Date(bbdd$`Fecha alta`, "%d/%m/%y")
entrada <- as.Date(bbdd$`Fecha de entrada pre`, "%d/%m/%y")
antelacio <- as.numeric(bbdd$Antelació)
hotel <- as.factor(bbdd$Hotel)
canal <- as.factor(bbdd$Subsegmento)
dia_alta <- as.numeric(bbdd$`Dia semana alta`)
dia_entrada <- as.numeric(bbdd$`Día semana de entrad`)
mes_alta <- as.numeric(bbdd$`Mes alta`)
mes_entrada <- as.numeric(bbdd$`Mes de entrada previ`)
persones <- as.numeric(bbdd$`Huéspedes Alojados`)
regim <- as.factor(bbdd$Régimen)
RN <- as.numeric(bbdd$RN)
AR <- as.numeric(bbdd$Ancillary)
RR <- as.numeric(bbdd$`Prod Alojamiento Financiera`)
revenue <- as.numeric(bbdd$`Prod Total Financiera`)


bbdd <- data.frame(hoste, sexe, edat, hotel, canal, dia_alta,
mes_alta, dia_entrada,antelacio, mes_entrada,regim, persones,
RN, AR, RR, revenue)
summary(bbdd)

#COLORS
# CL --> turquoise4
# MP --> turquoise3
```

```
# SA --> violetred3

dfCL <- bbdd[bbdd$hotel == "CL",]
dfMP <- bbdd[bbdd$hotel == "MP",]
dfSA <- bbdd[bbdd$hotel == "SA",]

summary( bbdd[ which( bbdd$hotel == "CL" ), ] )
summary( bbdd[ which( bbdd$hotel == "MP" ), ] )
summary( bbdd[ which( bbdd$hotel == "SA" ), ] )

stat.table(list(hotel = hotel), list(N = count(),"%" =
percent(hotel)),data = bbdd, margins = T)

pie(table(bbdd$hotel), labels = paste(c("CL", "MP",
"SA"),round(freq(bbdd$hotel, plot=F)[1:3,2]), "%"),
    main = "Hotel", col =
c("turquoise4","turquoise3","violetred3"))


#EDAT
summary(bbdd$edat)

par(mfrow=c(1,1))
par(cex.axis= 1, cex.lab= 1, cex.sub= 1, las=1)
hist(bbdd$edat,main = "Edats", xlab="Edat",ylab="Freqüència",
cex.main=1, col="grey", ylim=c(0,6500))

with(bbdd,tapply(edat,hotel,summary))

par(mfrow=c(1,3))
hist(dfCL$edat,main = " histograma de les edats de CL",
xlab="Edat",ylab="Freq??ncia"
     ,col="turquoise4")
hist(dfMP$edat,main = " histograma de les edats de MP",
xlab="Edat",ylab="Freq??ncia"
     ,col="turquoise3")
hist(dfSA$edat,main = " histograma de les edats de SA",
xlab="Edat",ylab="Freq??ncia"
     ,col="violetred3")

par(mfrow=c(1,3))
boxplot(dfCL$edat, border=1, col="turquoise4", horizontal=F,
main="Edats CL")
boxplot(dfMP$edat, border=1, col="turquoise3", horizontal=F,
main="Edats MP")
boxplot(dfSA$edat, border=1, col="violetred3", horizontal=F,
main="Edats SA")

par(mfrow=c(1,1))
```

```
plot(bbdd$hotel,bbdd$edat, col=
c("turquoise4","turquoise3","violetred3"), horitzontal=T,
main="Edats per hotel")

install.packages("PerformanceAnalytics")
library(PerformanceAnalytics)
dat1 <- data.frame(edat,revenue)
chart.Correlation(dat1, main="Correlació Edat- Revenue")

cor.test(edat, revenue)
plot(revenue, edat, main="Correlació Edat-Revenue")

dfSA <- bbdd[bbdd$hotel == "SA",]
dat2 <- data.frame(dfSA$edat,dfSA$revenue)
chart.Correlation(dat2, main="Correlació Edat- Revenue")

cor.test(dfSA$edat, dfSA$revenue)


#SEXE
stat.table(list(sexe = sexe), list(N = count(),"%" =
percent(sexe)),data = bbdd, margins = T)
par(mfrow=c(1,1))
pie(table(bbdd$sexe), labels = paste(c("Femení",
"Masculí"),round(freq(bbdd$sexe, plot=F)[1:2,2]), "%"),
    main = "Gènere de l'hoste ", col =
c("lightsteelblue1","lightsteelblue4"))

counts <- table(bbdd$hotel, bbdd$sexe)
counts <- prop.table(counts)*100
counts <- round(counts, 1)
pl<-barplot(counts, main="Gènere per hotel",ylab =
"Freqüències relatives (%)",
            xlab="Gènere",
col=c("turquoise4","turquoise3","violetred3"),
            cex.lab=0.5,
ylim=c(0,30),names.arg=c("Femení","Masculí"), beside=TRUE,
legend = rownames(counts))


count <- table(bbdd$sexe, bbdd$hotel)
count <- prop.table(count)*100
count <- round(count, 1)
barplot(count,col = c("lightsteelblue1","lightsteelblue4"),
        main="Gènere per hotel", ylim=c(0,60), legend =
c("Femení","Masculí"))

par(mfrow=c(1,3))
count <- table(dfCL$sexe)
count <- prop.table(count)*100
count <- round(count, 1)
```

```
sexeCL <- barplot(count,col =
c("lightsteelblue1","lightsteelblue4"),beside=FALSE,
        main="Gènere per CL", ylim=c(0,70),
names.arg=c("Femení","Masculí"))
text(sexeCL, count+2,format(count), cex=1.5)

count <- table(dfMP$sexe)
count <- prop.table(count)*100
count <- round(count, 1)
sexeMP <- barplot(count,col =
c("lightsteelblue1","lightsteelblue4"),beside=FALSE,
                  main="Gènere per MP", ylim=c(0,70),
names.arg=c("Femení","Masculí"))
text(sexeMP, count+2,format(count), cex=1.5)

count <- table(dfSA$sexe)
count <- prop.table(count)*100
count <- round(count, 1)
sexeSA <- barplot(count,col =
c("lightsteelblue1","lightsteelblue4"),beside=FALSE,
                  names.arg=c("Femení","Masculí"),
ylim=c(0,70),  main="Gènere per SA")
text(sexeSA, count+2,format(count), cex=1.5)


#CANAL
library("RColorBrewer",
lib.loc="/Library/Frameworks/R.framework/Versions/3.5/Resource
s/library")

stat.table(list(canal=canal), list(N = count(),"%" =
percent(canal)),data = bbdd, margins = T)

par(mfrow=c(1,1))
CN <- table(bbdd$canal)
CN <- prop.table(CN)*100
CN<- round(CN,1)

par(cex.axis= 1, cex.lab= 0.7, cex.sub= 1, las=1)
grf<- barplot(CN, main="Canal de compra",
col=brewer.pal(9,"Greys"), ylim=c(0,31), cex.main=1,
ylab="Freqüències relatives (%)", xlab="Canal")
text(grf, CN+2,format(CN), cex=1)

plot(x = bbdd$canal, main = "Canal de compra",
     xlab = "Canal", ylab = "Freqüència", ylim=c(0,12000),
     col=brewer.pal(9,"Greys"))

counts <- table(bbdd$hotel, bbdd$canal)
counts <- prop.table(counts)*100
counts <- round(counts, 1)
```

```
pl<-barplot(counts, main="Canal de compra per hotel",ylab =
"Freqüències relatives (%)",
        xlab="Canal",
col=c("turquoise4","turquoise3","violetred3"),
        cex.lab=0.5, ylim=c(0,20), beside=TRUE, legend =
rownames(counts))
text(pl, counts+2,format(counts), cex=0.7)

par(cex.axis= 0.5, cex.lab= 0.5, cex.sub= 0.7, las=1)
taula1<- table(hotel,canal)
plot(taula1, col=c(brewer.pal(9,"Greys")), main="Canal segons
hotel")
chisq.test(taula1)


#DIA SEMANA ALTA
stat.table(list(dia_alta=dia_alta), list(N = count(),"%" =
percent(dia_alta)),data = bbdd, margins = T)

CN <- table(bbdd$dia_alta)
CN <- prop.table(CN)*100
CN<- round(CN,1)

par(cex.axis= 1, cex.lab= 0.7, cex.sub= 1, las=1)
grf<- barplot(CN, main="Dia de la reserva",
names.arg=c("Dilluns","Dimarts","Dimecres","Dijous","Divendres
","Dissabte","Diumenge"),col=brewer.pal(9,"Greys"),
ylim=c(0,25), cex.main=1, ylab="Freqüències relatives (%)",
xlab="Dia")
text(grf, CN+2,format(CN), cex=1)

counts <- table(bbdd$hotel, bbdd$dia_alta)
counts <- prop.table(counts)*100
counts<- round(counts,1)

grf <- barplot(counts, main="Dia de la reserva per hotel",
        xlab="Dia",
col=c("turquoise4","turquoise3","violetred3"),

names.arg=c("Dilluns","Dimarts","Dimecres","Dijous","Divendres
","Dissabte","Diumenge"),
        cex.lab=1,  beside=TRUE, legend = rownames(counts),
ylim=c(0,13))
text(grf, counts+2,format(counts), cex=0.7)


taula1<- table(dia_alta, canal)
plot(taula1, col=brewer.pal(9,"Greys"), main="Dia de reserva
segons el canal")
chisq.test(taula1)
```

```
#MES ALTA
stat.table(list(mes_alta=mes_alta), list(N = count(),"%" =
percent(mes_alta)),data = bbdd, margins = T)

CN <- table(bbdd$mes_alta)
CN <- prop.table(CN)*100
CN<- round(CN,1)

par(cex.axis= 0.7, cex.lab= 0.5, cex.sub= 0.7, las=1)
grf<- barplot(CN, main="Mes de la reserva",
names.arg=c("ENE","FEB","MARZ","ABR","MAY","JUN","JUL","AGO","
SEP","OCT","NOV","DEC"),col=c("grey8","grey20","grey30",
"grey40", "grey48", "grey55", "grey62", "grey72",
"grey80","grey88","grey95", "grey100"), ylim=c(0,15),
cex.main=1, ylab="Freqüències relatives (%)", xlab="Mes")
text(grf, CN+2,format(CN), cex=1)


#ANTELACIO
summary(bbdd$antelacio)
par(mfrow=c(1,1))
boxplot(bbdd$antelacio, border=1, col="grey", horizontal=T,
main="Antelació de compra (dies)", range= 1.5)

plot(bbdd$hotel,bbdd$antelacio, col=
c("turquoise4","turquoise3","violetred3"), horitzontal=T,
main="Antelació de compra per hotel (dies)")

with(bbdd,tapply(antelacio,hotel,summary))

plot(bbdd$antelacio,bbdd$mes_alta, col=
c("turquoise4","turquoise3","violetred3"), horitzontal=T,
main="Mes alta i antelació")
chisq.test(bbdd$antelacio, bbdd$mes_alta)

plot(bbdd$antelacio,bbdd$mes_entrada, col=
c("turquoise4","turquoise3","violetred3"), horitzontal=T,
main="Mes alta i antelació")
chisq.test(bbdd$antelacio, bbdd$mes_entrada)

#REGIM
stat.table(list(regim = regim), list(N = count(),"%" =
percent(regim)),data = bbdd, margins = T)
par(mfrow=c(1,1))
pie(table(bbdd$regim), labels =
paste(c("PC","MP","BB"),round(freq(bbdd$regim,
plot=F)[1:3,2]), "%"),
    main = "Règim de la reserva ", col =
c(brewer.pal(3,"Greys")))
```

```
barplot(table(bbdd$regim))
taula1<- table(regim, hotel)
plot(taula1, col=c("turquoise4","turquoise3","violetred3"),
main="Règim per hotel")
chisq.test(bbdd$regim, bbdd$hotel)

#DIA SEMANA ENTRADA
stat.table(list(dia_entrada=dia_entrada), list(N = count(),"%"
= percent(dia_entrada)),data = bbdd, margins = T)

CN <- table(bbdd$dia_entrada)
CN <- prop.table(CN)*100
CN<- round(CN,1)

grf<- barplot(CN, main="Dia d'entrada a l'hotel",
names.arg=c("Dilluns","Dimarts","Dimecres","Dijous","Divendres
","Dissabte","Diumenge"),col=brewer.pal(9,"Greys"),
ylim=c(0,20), cex.main=1, ylab="Freqüències relatives (%)",
xlab="Dia d'entrada")
text(grf, CN+2,format(CN), cex=1)

counts <- table(bbdd$hotel, bbdd$dia_entrada)
counts <- prop.table(counts)*100
counts<- round(counts,1)
barplot(counts, main="Dia d'entrada a cada hotel",
        xlab="Dia d'entrada",
col=c("turquoise4","turquoise3","violetred3"),

names.arg=c("Dilluns","Dimarts","Dimecres","Dijous","Divendres
","Dissabte","Diumenge"),
        cex.lab=0.5,  beside=TRUE, legend = rownames(counts),
ylim=c(0,10))


#MES ENTRADA
stat.table(list(mes_entrada=mes_entrada), list(N = count(),"%"
= percent(mes_entrada)),data = bbdd, margins = T)

CN <- table(bbdd$mes_entrada)
CN <- prop.table(CN)*100
CN<- round(CN,1)

par(cex.axis= 0.5, cex.lab= 0.5, cex.sub= 0.7, las=1)
grf<- barplot(CN, main="Mes d'entrada a l'hotel",
names.arg=c("ENE","FEB","MARZ","ABR","MAY","JUN","JUL","AGO","
SEP","OCT","NOV","DEC"),col=c("grey8","grey20","grey30",
"grey40", "grey48", "grey55", "grey62", "grey72",
"grey80","grey88","grey95", "grey100"), ylim=c(0,13),
cex.main=1, ylab="Freqüències relatives (%)", xlab="Mes")
text(grf, CN+2,format(CN), cex=1)
```

```
#PERSONES
summary(bbdd$persones)
par(cex.axis= 1, cex.lab= 1, cex.sub= 0.7, las=1)

par(mfrow=c(1,1))
CN <- table(bbdd$persones)
CN <- prop.table(CN)*100
CN<- round(CN,1)
barplot(CN,
        ylab = "Freqüències relatives (%)", xlab = "Nombre de
persones", ylim=c(0,85),
        main = "Nombre de persones per reserva",
        col = c(brewer.pal(5,"Greys")))
text(grf, CN+2,format(CN), cex=1)

taula1<- table(persones, hotel)
plot(taula1, col=c("turquoise4","turquoise3","violetred3"),
main="Règim per hotel")


#RN
summary(bbdd$RN)

par(mfrow=c(1,1))
boxplot(bbdd$RN, border=1, col="grey", horizontal=F,
main="Total de nits de la reserva", range= 1.5)

plot(bbdd$hotel,bbdd$RN, col=
c("turquoise4","turquoise3","violetred3"), horitzontal=T,
main="Total de nits de la reserva i hotel")

par(mfrow=c(3,1))
boxplot(dfCL$RN, border=1, col="turquoise4", horizontal=T,
main="Total de nits de la reserva CL")
boxplot(dfMP$RN, border=1, col="turquoise3", horizontal=T,
main="Total de nits de la reserva MP")
boxplot(dfSA$RN, border=1, col="violetred3", horizontal=T,
main="Total de nits de la reserva SA")

library(PerformanceAnalytics)
dat1 <- data.frame(RR,RN)
chart.Correlation(dat1, main="Correlació Edat- Revenue")


#RR
summary(bbdd$RR)
par(mfrow=c(1,1))
boxplot(bbdd$RR, border=1, col="grey", horizontal=F,
main="Room revenue (€)", range= 1.5)
```

```
plot(bbdd$hotel,bbdd$RR, col=
c("turquoise4","turquoise3","violetred3"), horitzontal=T,main=
"Room revenue per hotel (€)")

with(bbdd,tapply(RR,hotel,summary))


#AR
summary(bbdd$AR)
par(mfrow=c(1,1))
boxplot(bbdd$AR, border=1, col="grey", horizontal=F,
main="Extra revenue (€)", range= 1.5)
plot(bbdd$hotel,bbdd$AR, col=
c("turquoise4","turquoise3","violetred3"), horitzontal=T,main=
"Extra revenue per hotel (€)")

with(bbdd,tapply(AR,hotel,summary))

plot(bbdd$regim,bbdd$AR, col= c(brewer.pal(3,"Greys")),
horitzontal=T,main= "Extra revenue per règim (€)")
plot(bbdd$RR,bbdd$AR, horitzontal=T,main= "Extra revenue i
Room revenue (€)", xlab= "RR (€)", ylab= "AR (€)")
cor.test(bbdd$RR, bbdd$AR)


#REVENUE
summary(bbdd$revenue)
par(mfrow=c(1,1))
boxplot(bbdd$revenue, border=1, col="grey", horizontal=F,
main="Total revenue (€)", range= 1.5)
plot(bbdd$hotel,bbdd$revenue, col=
c("turquoise4","turquoise3","violetred3"), horitzontal=T,main=
"Total revenue per hotel (€)")

with(bbdd,tapply(revenue,hotel,summary))


#TRACTAMENT
library(datasets)

m <- data.frame(hotel, sexe, edat, dia_alta, dia_entrada,
mes_alta, mes_entrada, antelacio, canal, regim, persones, RN,
AR, RR)
dd <- bbdd

dfCL <- bbdd[bbdd$hotel ==
"CL",c("edat","antelacio","persones","RN","AR","RR")]
dfMP <- bbdd[bbdd$hotel ==
"MP",c("edat","antelacio","persones","RN","AR","RR")]
dfSA <- bbdd[bbdd$hotel ==
"SA",c("edat","antelacio","persones","RN","AR","RR")]
```

```
#MULTICOLINEALITAT
install.packages("PerformanceAnalytics")
library(PerformanceAnalytics)
dat1 <- data.frame(edat, antelacio, persones, RN, AR, RR,
revenue)
chart.Correlation(dat1, main="Correlació Edat- Revenue")


#NORMALITAT
num <- data.frame(hotel, edat, antelacio, RN, AR, RR)
par(mfrow=c(3,2))
for(i in 2:ncol(num)){
  plot(density(num[,i]),ylab=names(num)[i] )}

dfCL <- bbdd[bbdd$hotel ==
"CL",c("edat","antelacio","RN","AR","RR")]
dfMP <- bbdd[bbdd$hotel ==
"MP",c("edat","antelacio","RN","AR","RR")]
dfSA <- bbdd[bbdd$hotel ==
"SA",c("edat","antelacio","RN","AR","RR")]

par(mfrow=c(3,3))
qqnorm(dfCL$edat, pch = 19, col = "turquoise4", main=" CL:
Edat")
qqline(dfCL$edat)
qqnorm(dfMP$edat, pch = 19, col = "turquoise3", main=" MP:
Edat")
qqline(dfMP$edat)
qqnorm(dfSA$edat, pch = 19, col = "violetred3", main=" SA:
Edat")
qqline(dfSA$edat)

qqnorm(dfCL$antelacio, pch = 19, col = "turquoise4", main="
CL: Antelació")
qqline(dfCL$antelacio)
qqnorm(dfMP$antelacio, pch = 19, col = "turquoise3", main="
MP: Antelació")
qqline(dfMP$antelacio)
qqnorm(dfSA$antelacio, pch = 19, col = "violetred3", main="
SA: Antelació")
qqline(dfSA$antelacio)

qqnorm(dfCL$RN, pch = 19, col = "turquoise4", main=" CL: RN")
qqline(dfCL$RN)
qqnorm(dfMP$RN, pch = 19, col = "turquoise3", main=" MP: RN")
qqline(dfMP$RN)
qqnorm(dfSA$RN, pch = 19, col = "violetred3", main=" SA: RN")
qqline(dfSA$RN)

qqnorm(dfCL$RR, pch = 19, col = "turquoise4", main=" CL: RR")
qqline(dfCL$RR)
```

```
qqnorm(dfMP$RR, pch = 19, col = "turquoise3", main=" MP: RR")
qqline(dfMP$RR)
qqnorm(dfSA$RR, pch = 19, col = "violetred3", main=" SA: RR")
qqline(dfSA$RR)

qqnorm(dfCL$AR, pch = 19, col = "turquoise4", main=" CL: AR")
qqline(dfCL$AR)
qqnorm(dfMP$AR, pch = 19, col = "turquoise3", main=" MP: AR")
qqline(dfMP$AR)
qqnorm(dfSA$AR, pch = 19, col = "violetred3", main=" SA: AR")
qqline(dfSA$AR)

library("nortest")
lillie.test(x = dd$edat)
lillie.test(x = dd$antelacio)
lillie.test(x = dd$RN)
lillie.test(x = dd$RR)
lillie.test(x = dd$AR)

hist(edat, probability = TRUE, main = "Edat", xlab =
"Població", ylab = "Densitat")
x   <- seq(min(edat), max(edat), length = 1000)
y <- dnorm(x, mean(edat), sd(edat))
lines(x, y, col = "blue")

hist(antelacio, probability = TRUE, main = "Antelació", xlab =
"Població", ylab = "Densitat")
x   <- seq(min(antelacio), max(antelacio), length = 1000)
y <- dnorm(x, mean(antelacio), sd(antelacio))
lines(x, y, col = "blue")

hist(persones, probability = TRUE, main = "Persones", xlab =
"Població", ylab = "Densitat")
x   <- seq(min(persones), max(persones), length = 1000)
y <- dnorm(x, mean(persones), sd(persones))
lines(x, y, col = "blue")

hist(RN, probability = TRUE, main = "RN", xlab = "Població",
ylab = "Densitat")
x   <- seq(min(RN), max(RN), length = 1000)
y <- dnorm(x, mean(RN), sd(RN))
lines(x, y, col = "blue")

hist(RR, probability = TRUE, main = "RR", xlab = "Població",
ylab = "Densitat")
x   <- seq(min(RR), max(RR), length = 1000)
y <- dnorm(x, mean(RR), sd(RR))
lines(x, y, col = "blue")

hist(AR, probability = TRUE, main = "AR", xlab = "Població",
ylab = "Densitat")
```

```
x   <- seq(min(AR), max(AR), length = 1000)
y <- dnorm(x, mean(AR), sd(AR))
lines(x, y, col = "blue")

#QQnorm
par(mfrow=c(3,2))

qqnorm(dd$edat, pch = 19, col = "gray50", main="Edat")
qqline(dd$edat)

qqnorm(dd$antelacio, pch = 19, col = "gray50",
main="Antelació")
qqline(dd$antelacio)

qqnorm(dd$persones, pch=19, col="gray50", main="Persones")
qqline(dd$persones)

qqnorm(dd$RN, pch = 19, col = "gray50", main="RN")
qqline(dd$RN)

qqnorm(dd$RR, pch = 19, col = "gray50", main="RR")
qqline(dd$RR)

qqnorm(dd$AR, pch = 19, col = "gray50", main="AR")
qqline(dd$AR)


#MULTIVARIADA
install.packages("MVN")
library(MVN)
dd<- data.frame(hotel, edat, antelacio, persones, RN, AR, RR)

dfCL <- bbdd[bbdd$hotel ==
"CL",c("edat","antelacio","persones","RN","AR","RR")]
dfMP <- bbdd[bbdd$hotel ==
"MP",c("edat","antelacio","persones","RN","AR","RR")]
dfSA <- bbdd[bbdd$hotel ==
"SA",c("edat","antelacio","persones","RN","AR","RR")]


#NORM MULTI CL

valores <- NULL
i <- 1
while ( i <= nrow(dfCL)) {
  result<-mvn(data = dfCL[sample(1:nrow(dfCL),1000,replace=F)
,], mvnTest = "hz")
  valores <- c(valores,result$multivariateNormality[,2])
  print(valores)
  i <- i + 1000 }
```

```
#NORM MULTI MP
valores <- NULL
i <- 1
while ( i <= nrow(dfMP)) {
  result<-mvn(data = dfMP[sample(1:nrow(dfMP),1000,replace=F)
,], mvnTest = "hz")
  valores <- c(valores,result$multivariateNormality[,2])
  print(valores)
  i <- i + 1000 }


#NORM MULTI SA
install.packages("RVAideMemoire")
library("RVAideMemoire")

valores <- NULL
i <- 1
while ( i <= nrow(dfSA)) {
  result<-mvn(data = dfSA[sample(1:nrow(dfSA),1000,replace=F)
,], mvnTest = "hz")
  valores <- c(valores, result$multivariateNormality[,2])
  print(result)
  print(valores)
  i <- i + 1000 }


#HOMOCEDASTICITAT
install.packages("biotools")
library(biotools)

factor(dd[,7])
boxM(data = dd[, -1], grouping = dd[, 1])

library(car)
leveneTest(edat,hotel)
leveneTest(antelacio,hotel)
leveneTest(persones,hotel)
leveneTest(RN,hotel)
leveneTest(RR,hotel)
leveneTest(AR,hotel)

CLcov <- cov(dfCL[,])
round(CLcov,2)

MPcov <- cov(dfMP[,])
round(MPcov,2)

SAcov <- cov(dfSA[,])
round(SAcov,2)
plotCov(cov)
```

```
res <- boxM(dd[, 2:6], dd[, "hotel"])
boxM(data = num[, 1:7], grouping = num[, 1])


#APLICACIÓ
dades <- data.frame(hotel, edat , antelacio, persones, RN, AR,
RR, sexe, canal, regim, dia_entrada, mes_entrada, dia_alta,
mes_alta)

set.seed(101)
tamany <- nrow(dades)
estudi <- round((nrow(dades))*0.8)
index_dades <- sample(1:tamany , size=estudi)
dades_estudi <- dades[index_dades,]
dades_prediccio <- dades[-index_dades,]


#LDA
library(MASS)
funciolda <- lda(hotel~ ., data=dades_estudi)
funciolda
round(funciolda$scaling,3)
lda.pred <- predict(funciolda)$class

table(dades_estudi$hotel, lda.pred, dnn = c('Actual
Group','Predicted Group'))
mean(lda.pred != dades_estudi$hotel)

pred<- predict(funciolda, data=dades_estudi)
ldahist(pred$x[,1], g= pred$class, col=c("grey"))
ldahist(pred$x[,2], g= pred$class, col=c("grey"))


#preddicio
pred_lda <- predict(object = funciolda, dades_prediccio)
ldahist(pred_lda$x[,1], g= pred$class, col=c("grey"))
ldahist(pred_lda$x[,2], g= pred$class, col=c("grey"))

table(clase_predicha = pred_lda$class,
      clase_real =  dades_prediccio$hotel)

mean(pred_lda$class != dades_prediccio$hotel)
mean(pred_lda$class == dades_prediccio$hotel)

round(head(pred_lda$posterior, n = 5),4)

library(klaR)
par(mfrow=c(1,1))
partimat(formula = hotel ~ edat + antelacio + RN + RR + AR,
data = dades_prediccio, method = "lda",
```

```
         prec = 400, image.colors =
c("turquoise4","turquoise3" ,"violetred3"),col.mean =
"firebrick", nplots.vert = 2)



#QDA
funcioqda <- qda(hotel~ ., data=dades_estudi)
funcioqda

predicciones_qda <- predict(object = funcioqda,
dades_prediccio)
head(predicciones_qda$class)

round(head(predicciones_qda$posterior, n=5),4)
table(clase_predicha = predicciones_qda$class,
      clase_real =  dades_prediccio$hotel)

mean(predicciones_qda$class != dades_prediccio$hotel)
mean(predicciones_qda$class == dades_prediccio$hotel)


library(klaR)
par(mfrow=c(1,1))
partimat(formula = hotel ~ edat + antelacio + RN + RR + AR,
data = dades_prediccio, method = "qda",
         prec = 400, image.colors =
c("turquoise4","turquoise3" ,"violetred3"),col.mean =
"firebrick", nplots.vert = 2)

#K-NN
library("class")
library("fastDummies")

df<- dummy_cols(dades_estudi, select_columns = c("sexe",
"canal","regim",
"dia_entrada","mes_entrada","dia_alta","mes_alta"),
remove_first_dummy = FALSE, remove_selected_columns=TRUE)
new_df<-dummy_cols(dades_prediccio, select_columns = c("sexe",
"canal","regim",
"dia_entrada","mes_entrada","dia_alta","mes_alta"),
remove_first_dummy = FALSE, remove_selected_columns=TRUE)

dades_estudi <- df
dades_prediccio <- new_df


# K = 79
KnnTest_k79 <- knn(dades_estudi[,2:55],dades_prediccio[,2:55],
dades_estudi$hotel , k = 79, prob = TRUE )
table ( KnnTest_k79, dades_prediccio$hotel )
```

```
c1<-sum(KnnTest_k79 == dades_prediccio$hotel)/
length(dades_prediccio$hotel)*100

# K = 83
KnnTest_k83 <- knn(dades_estudi[,2:55],dades_prediccio[,2:55],
dades_estudi$hotel , k = 83, prob = TRUE )
c2<-sum(KnnTest_k83 == dades_prediccio$hotel)/
length(dades_prediccio$hotel)*100

# K = 89
KnnTest_k89 <- knn(dades_estudi[,2:55],dades_prediccio[,2:55],
dades_estudi$hotel , k = 89, prob = TRUE )
c3<-sum(KnnTest_k89 == dades_prediccio$hotel)/
length(dades_prediccio$hotel)*100

# K = 95
KnnTest_k95 <- knn(dades_estudi[,2:55],dades_prediccio[,2:55],
dades_estudi$hotel , k = 95, prob = TRUE )
c4<-sum(KnnTest_k95 == dades_prediccio$hotel)/
length(dades_prediccio$hotel)*100

# K = 97
KnnTest_k97 <- knn(dades_estudi[,2:55],dades_prediccio[,2:55],
dades_estudi$hotel , k = 97, prob = TRUE )
c5<-sum(KnnTest_k97 == dades_prediccio$hotel)/
length(dades_prediccio$hotel)*100


#Transformaciones logaritmiques
par(mfrow=c(1,2))
library(moments)

skewness(edat, na.rm = TRUE)

logedat <- log10(max(edat+1)-edat)
hist(logedat, probability = TRUE, main = "", xlab = "log
Edat", ylab = "Densitat")
x   <- seq(min(logedat), max(logedat), length = 1000)
y <- dnorm(x, mean(logedat), sd(logedat))
lines(x, y, col = "red")
qqnorm(logedat, main = "LogEdat")
qqline(logedat)

skewness(edat, na.rm = TRUE)
skewness(logedat, na.rm = TRUE)

ks.test(logedat, pnorm, mean(edat), sd(edat))


skewness(antelacio, na.rm = TRUE)
logantelacio <- log10(antelacio)
```

```
i <- 1
while ( i <= length(logantelacio)) {
  if (logantelacio[i] == "-Inf"){
    logantelacio[i] <- 0}
  i <- i + 1 }

hist(logantelacio, probability = TRUE, main = "", xlab = "log
Antelación", ylab = "Densitat")
x   <- seq(min(logantelacio), max(logantelacio))
y <- dnorm(x, mean(logantelacio), sd(logantelacio))
lines(x, y, col = "red")
qqnorm(logantelacio, main = "LogAntelació", ylim=c(0,8))
qqline(logantelacio)

skewness(antelacio, na.rm = TRUE)
skewness(logantelacio, na.rm = TRUE)


skewness(RN, na.rm = TRUE)
logRN <- log10(RN)
hist(logRN, probability = TRUE, main = "", xlab = "log RN",
ylab = "Densitat")
x   <- seq(min(logRN), max(logRN))
y <- dnorm(x, mean(logRN), sd(logRN))
lines(x, y, col = "red")
qqnorm(logRN, main = "LogRN")
qqline(logRN)

skewness(RN, na.rm = TRUE)
skewness(logRN, na.rm = TRUE)


skewness(RR, na.rm = TRUE)
logRR <- log(RR)

i <- 1
while ( i <= length(logRR)) {
  if (logRR[i] == "-Inf"){
    logRR[i] <- 0}
  i <- i + 1 }

hist(logRR, probability = TRUE, main = "", xlab = "log RR",
ylab = "Densitat")
x   <- seq(min(logRR), max(logRR))
y <- dnorm(x, mean(logRR), sd(logRR))
lines(x, y, col = "red")
qqnorm(logRR, main = "LogRR", ylim=c(0,10))
qqline(logRR)
```

```
skewness(RR, na.rm = TRUE)
skewness(logRR, na.rm = TRUE)


skewness(AR, na.rm = TRUE)
logAR <- log(AR)
is.na(logAR) <- 0
logAR <- replace(logAR, is.na(logAR), 0)
i <- 1
while ( i <= length(logAR)) {
  if (logAR[i] == "-Inf" ){
    logAR[i] <- 0}
  i <- i + 1 }

hist(logAR, probability = TRUE, main = "", xlab = "log AR",
ylab = "Densitat")
x   <- seq(min(logAR), max(logAR))
y <- dnorm(x, mean(logAR), sd(logAR))
lines(x, y, col = "red")
qqnorm(logAR, main = "LogAR", ylim=c(0,10))
qqline(logAR)

skewness(AR, na.rm = TRUE)
skewness(logAR, na.rm = TRUE)


dades <- data.frame(hotel, edat , logantelacio, persones,
logRN, logAR, logRR, sexe, canal, regim, dia_entrada,
mes_entrada, dia_alta, mes_alta)

set.seed(101)
tamany <- nrow(dades)
estudi <- round((nrow(dades))*0.8)
index_dades <- sample(1:tamany , size=estudi)
dades_estudi <- dades[index_dades,]
dades_prediccio <- dades[-index_dades,]

#LDA
library(MASS)
funciolda <- lda(hotel~ ., data=dades_estudi)
funciolda
round(funciolda$scaling,3)
lda.pred <- predict(funciolda)$class

pred<- predict(funciolda, data=dades_estudi)
ldahist(pred$x[,1], g= pred$class, col=c("grey"))
ldahist(pred$x[,2], g= pred$class, col=c("grey"))


#preddicio
pred_lda <- predict(object = funciolda, dades_prediccio)
```

```
ldahist(pred_lda$x[,1], g= pred$class, col=c("grey"))
ldahist(pred_lda$x[,2], g= pred$class, col=c("grey"))

table(clase_predicha = pred_lda$class,
      clase_real =  dades_prediccio$hotel)

mean(pred_lda$class != dades_prediccio$hotel)
mean(pred_lda$class == dades_prediccio$hotel)

round(head(pred_lda$posterior, n = 5),4)

library(klaR)
par(mfrow=c(1,1))
partimat(formula = hotel ~ edat + antelacio + RN + RR + AR,
data = dades_prediccio, method = "lda",
         prec = 400, image.colors =
c("turquoise4","turquoise3" ,"violetred3"),col.mean =
"firebrick", nplots.vert = 2)


#QDA
modelo_qda <- qda(hotel~ ., data=dades_estudi)
modelo_qda

#prediccio
predicciones_qda <- predict(object = modelo_qda,
dades_prediccio)
head(predicciones_qda$class)

round(head(predicciones_qda$posterior, n=5),4)
table(clase_predicha = predicciones_qda$class,
      clase_real =  dades_prediccio$hotel)

mean(predicciones_qda$class != dades_prediccio$hotel)
mean(predicciones_qda$class == dades_prediccio$hotel)

library(klaR)
par(mfrow=c(1,1))
partimat(formula = hotel ~ edat + antelacio + RN + RR + AR,
data = dades_prediccio, method = "qda",
         prec = 400, image.colors =
c("turquoise4","turquoise3" ,"violetred3"),col.mean =
"firebrick", nplots.vert = 2)
```

```
#K-NN
library("class")
library("fastDummies")

df<- dummy_cols(dades_estudi, select_columns = c("sexe",
"canal","regim",
"dia_entrada","mes_entrada","dia_alta","mes_alta"),
remove_first_dummy = FALSE, remove_selected_columns=TRUE)
new_df<-dummy_cols(dades_prediccio, select_columns = c("sexe",
"canal","regim",
"dia_entrada","mes_entrada","dia_alta","mes_alta"),
remove_first_dummy = FALSE, remove_selected_columns=TRUE)

dades_estudi <- df
dades_prediccio <- new_df

# K = 79
KnnTest_k79 <- knn(dades_estudi[,2:55],dades_prediccio[,2:55],
dades_estudi$hotel , k = 79, prob = TRUE )
table ( KnnTest_k79, dades_prediccio$hotel )
c1<-sum(KnnTest_k79 == dades_prediccio$hotel)/
length(dades_prediccio$hotel)*100

# K = 83
KnnTest_k83 <- knn(dades_estudi[,2:55],dades_prediccio[,2:55],
dades_estudi$hotel , k = 83, prob = TRUE )
c2<-sum(KnnTest_k83 == dades_prediccio$hotel)/
length(dades_prediccio$hotel)*100

# K = 89
KnnTest_k89 <- knn(dades_estudi[,2:55],dades_prediccio[,2:55],
dades_estudi$hotel , k = 89, prob = TRUE )
c3<-sum(KnnTest_k89 == dades_prediccio$hotel)/
length(dades_prediccio$hotel)*100

# K = 95
KnnTest_k95 <- knn(dades_estudi[,2:55],dades_prediccio[,2:55],
dades_estudi$hotel , k = 95, prob = TRUE )
c4<-sum(KnnTest_k95 == dades_prediccio$hotel)/
length(dades_prediccio$hotel)*100

# K = 97
KnnTest_k97 <- knn(dades_estudi[,2:55],dades_prediccio[,2:55],
dades_estudi$hotel , k = 97, prob = TRUE )
c5<-sum(KnnTest_k97 == dades_prediccio$hotel)/
length(dades_prediccio$hotel)*100
```