Exemple d'aplicació de tècniques de mineria de dades a la industria hotelera

MINERIA DE DADES, FACULTAT DE MATEMÀTIQUES I ESTADÍSTICA

DATA D'ENTREGA: 18.12.2018

AUTORS:

Divisió 1

Guillem Querol Llaveria

Pablo Morante López

Víctor Miranda Hernández

Carles Requena Sánchez

Aleix Salvador Barrera

Divisió 2

Laura Julià Melis

Marta Piñol Palau

Antoni Ramoneda Montoya

Sofía Touceda Suárez

ÍNDEX

1. Definició del projecte i assignació.

Aquest treball es desenvoluparà amb l'objectiu de millorar les experiències de viatges en l'àmbit de la industria hotelera. Per a assolir aquest objectiu, s'utilitzaran una sèrie de tècniques de mineria de dades.

i. Font i informació sobre la base de dades.

Com a matèria prima per a l'anàlisi, s'han utilitzat dades importades a través d'una API des de 'Booking'. Aquestes dades són propietat de *Booking* però un usuari de kaggle les ha fet públiques i en permet l'ús amb finalitats acadèmiques. A la web trobem dues versions de les dades: un primer 'dataset' amb la informació obtinguda de *Booking* (https://www.kaggle.com/jiashenliu/ 515k- hotel-reviews-data-in-europe), i un de segon, en el qual la informació del primer ha estat enriquida amb noves variables. Aquest últim és el que s'utilitza en aquest treball https://www. kaggle.com/ycalisar/hotel-reviews-dataset-enriched.

La base de dades conté, aproximadament, 515.000 opinions de clients, i les puntuacions atorgades pels mateixos, que han contractat estades a un total de 1.493 hotels d'Europa. *Booking* també proporciona les coordenades de cada hotel per a realitzar geolocalitzacions.

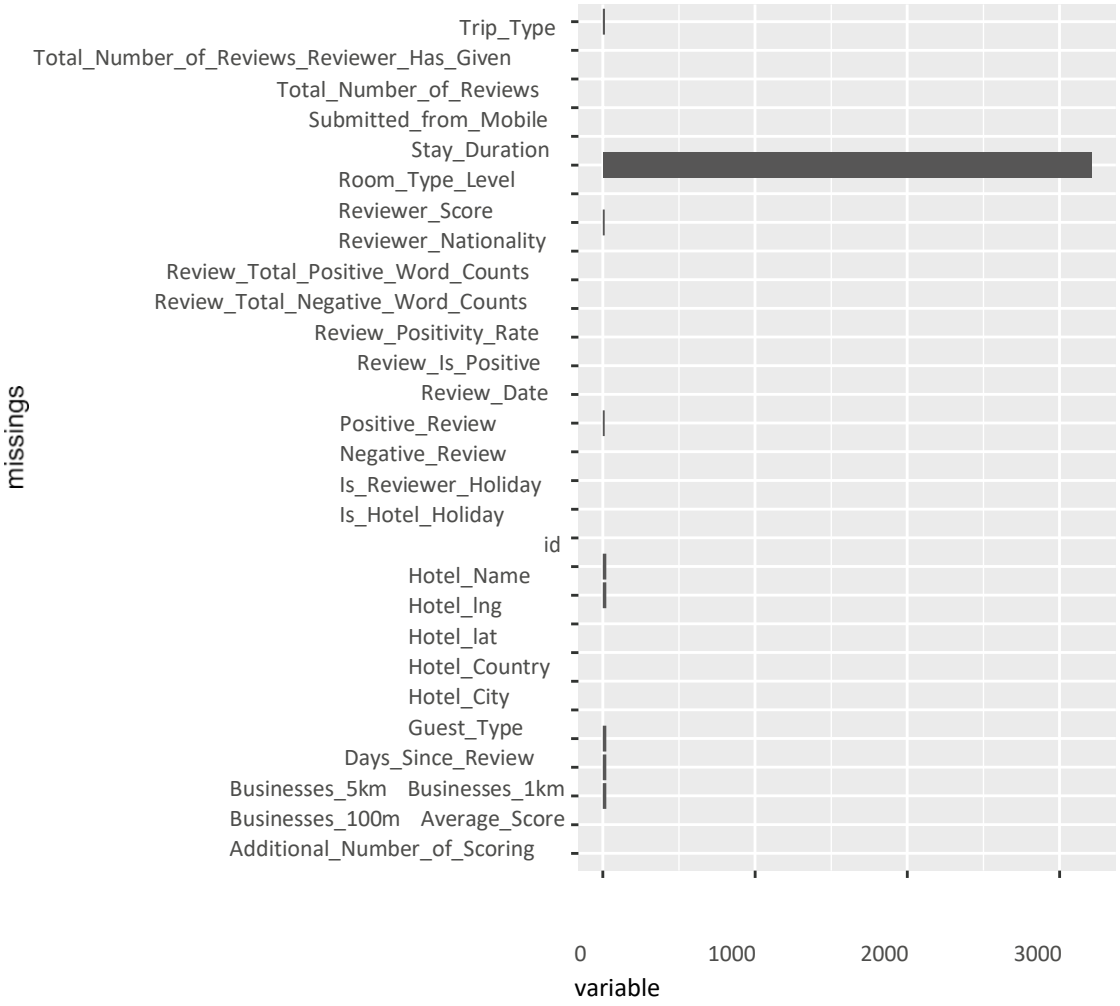
ii. Estructura de la base de dades.

A grans trets, la matriu de dades resultant conté 41 variables, 21 de les quals són numèriques i 20 categòriques, i 515.738 observacions. Tanmateix, per a ajustar-nos a la dimensionalitat de les dades proposada a les bases del treball, s'ha seleccionat aleatòriament un subconjunt de 5.000 observacions i s'han eliminat de la base de dades les variables [*Hotel_Address, Hotel_State, Room_Type, Tags, Day_of_Week, Day_of_Year, Bed_Type, Week_of_Month, Week_of_Year, Quarter_of_Year,Reviewer_Country*], que es consideren poc rellevants en relació als objectius del treball. D'aquesta manera, assegurem que tots els procediments requerits podran ser implementats de manera eficient i satisfactòria amb les nostres dades.

Respecte als valors *missing*, la base de dades revela un total de 3.350, que representen un 2, 23% de la matriu de dades completa ($m \cdot n$). En aquest sentit, presentem la Taula 1 que mostra com es distribueixen els valors *missing* entre les diferents variables, així com la Figura 1.1 on es representa un histograma que resumeix la taula anterior.

variable	nre.Missings	freq.Missings
Hotel_lat	23	0.0153 %
Hotel_lng	23	0.0153 %
Businesses_100m	23	0.0153 %
Businesses_1km	23	0.0153 %
Businesses_5km	23	0.0153 %
Room_Type_Level	3209	2.1393 %
Trip_Type	14	0.0093 %
Reviewer_Nationality	10	0.0067 %
Negative_Review	2	0.0013 %

Taula 1: Taula de valors missing



<u>Figura 1</u>: Histograma dels valors missing

2. Pla de treball

Un pla de treball és una eina que permet ordenar i sistematitzar informació rellevant per a realitzar un treball, a més permet organitzar a un grup per dur a terme les diferents tasques que intervenen durant el procés i recollir els objectius d'aquestes.

i. Tasques.

Part comuna inicial:

- Portada.
- Definició del projecte i assignació
- Pla de treball.
- Estructura de les dades i descriptiva.
    - Descripció formal de l'estructura de les dades.
    - Anàlisi descriptiva univariant inicial de totes les variables.
    - Descripció detallada del procés de preprocessament de dades seguit i justificació de totes les decisions preses.
    - Anàlisi descriptiva univariant de les dades preprocessades i discussió sobre l'aleatorietat de les dades mancants si n'hi ha.
- Disseny dels dos processos de mineria de dades a seguir.

Divisió (treball paral·lel):

- Procés de mineria de dades.
    - Aplicar almenys un mètode de cada un dels següents: profiling, associatiu, discriminant i predictiu.
    - Comentar resultats obtinguts.
    - Realitzar proves de validació quan sigui necessari.

Part comuna final:

- Anàlisi comparativa entre els processos seguits per les dues divisions.
- Conclusions generals.
- Pla de treball real.
- Scripts d'R utilitzats.

ii. Seqüenciació temporal en diagrama de Gantt.

Aquest diagrama s'utilitza per planificar i programar tasques al llarg d'un període determinat. Gràcies a una visualització còmoda i senzilla de les accions previstes (tasques, durada, seqüència i calendari general).

En funció de les diferents dates de lliuraments i les tasques a realitzar en el període establert s'ha realitzat el següent diagrama de Gantt:

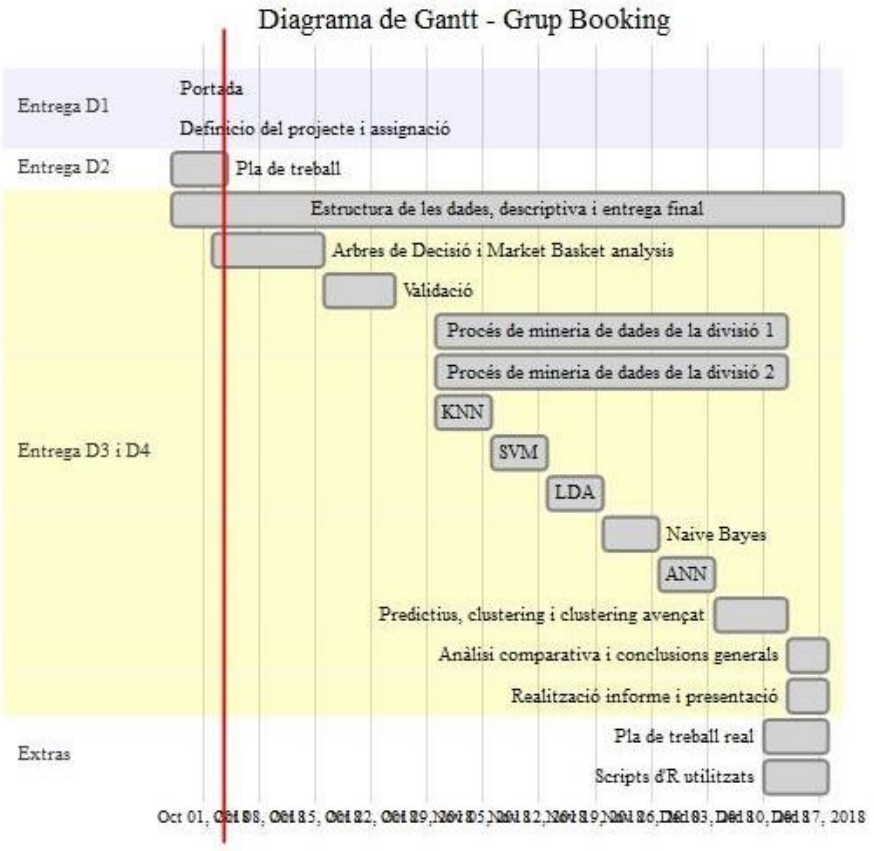


Figura 2: Diagrama de Gantt

iii. Distribució de les tasques.

S'ha realitzat un repartiment de les tasques entre els membres del grup el més equitatiu possible. El repartiment s'ha fet de forma aleatòria entre els 7 membres del grup amb l'objectiu d'aconseguir portar a terme aquest treball d'una manera ordenada i evitant possibles descompensacions de treball entre els membres. D'aquesta manera tots els integrants del grup han de ser capaços de explicar a la resta del grup les tasques que ha realitzat de manera que tots el puguin entendre i explicar.

A més, s'ha repartit de la mateixa manera la funció de coordinador perquè això no recaigui només sobre una persona sinó que tots els participants es facin càrrec almenys una vegada durant aquest període de temps.

El següent repartiment pot ser patir canvis futurs.

				GRUP 1					GRUP 2			
				Guillem	Pablo	Antoni	Carles	Aleix	Laura	Víctor	Marta	Sofía
Entrega D1			Portada								X	X
			Definició del projecte i assignació		X	X	X					
Entrega D2	Estructura de les dades, descriptiva i entrega final		Pla de treball	X				X	X	X		
Entrega D3 iD4			Estructura de les dades, descriptiva i entrega final	X	X						X	X
			Arbres de decisió			X	X	X	X	X		
			Validació		X		X				X	X
		Procés de mineria de dades	KNN	X	X					X	X	
			SVM			X	X		X			X
			LDA		X		X	X		X	X	
			Naive Bayes	X		X		X	X			X
			ANN		X	X	X		X	X		
			Prectius, clustering i clustering avençat	X		X		X			X	X
			Anàlisi comparativa i conclusions generals	X	X	X	X	X	X	X	X	X
			Realització informe i presentació	X	X	X	X	X	X	X	X	X
Extres			Pla de treball real	X				X	X	X		
			Scripts d'R utilitzats	X	X	X	X	X	X	X	X	X

Figura 3:Plantilla de Tasques

iv. Pla de riscos.

S'ha de tenir en compte que en un grup format per diversos integrants possibles fets futurs no planejats que perjudiquin el grup de una manera o altra.

Per tant, s'ha realitzat una taula que recull alguns dels possibles riscos que pot patir el grup a l'hora de fer la feina, així com una manera d'evitar-ho o prevenir-lo i la manera d'assumir-ho.

Possible risc	Com prevenir-lo	Com gestionar-lo
Els dos grups no treballen igual.	Anar revisant la feina dels dos grups periòdicament.	En el moment que hi hagin diferències considerables entre els dos grups, fer un canvi de components i reorganitzar el repartiment.
Un membre del grup es posa malalt i no pot fer la seva part de la tasca.	Les tasques assignades als membres del grup no seran individuals. Així doncs si un no pot fer-la, l'altre persona a la que se li ha assignat la podrà fer.	S'haurà de reajustar l'assignació de treballs per tal de compensar la pèrdua.
Distribució del treball inadequadament.	Revisar amb antelació la plantilla de tasques.	En cas de que algun component del grup tingui molta feina i altres el contrari, haurem de tornar a fer la plantilla de riscos.
No entregar a temps les tasques	Fer els scripts a classe i les entregues amb temps.	Seguir el diagrama de Gantt, en el qual ja prevenim possibles contratemps deixant uns dies de marge.
Pèrdua del treball o parts d'aquest	Compartir-ho entre tots els membres del grup en un espai virtual tots els avenços del treball (Google drive).	Assegurar-nos de que tots els components pujant els avenços al espai virtual.
Entregues o scripts mal realitzats	Per a cada tasca o entrega hi haurà un grup de 2 persones que s'encarregaren de revisar la feina realitzada.	En el repartiment de tasques també sortiran en cada entrega el grup que s'encarregarà de la revisió de la feina.

Errades ortogràfiques a l'informe	Fer servir en tot moment el corrector, assegurant-nos del seu bon funcionament.	Designar un encarregat de revisions ortogràfiques per a cada part del informe.
Discussions causades per la diversitat d'opinions entre alguns o tots els membres del´equip	Tenir una bona comunicació i respecte a l´hora de discutir els temes referits a la feina.	S´ha d´arribar a un acord amb el qual tots els membres estiguin totalment o parcialment d´acord.
Desobediència d'algun membre del grup en les indicacions grupals i actitud individualista	Intentar que totes les decisions es facin de forma que tot el grup estigui d'acord.	S'escollirà un líder per majoria, que pot anar variant. Aquest s'encarregarà de tenir la última paraula en termes de repartiment de tasques i si fa falta parlar amb el professor.
Un membre de l'equip decideix renunciar a l'avaluació contínua	Totes les tasques tenen almenys dos membres del´equip assignat.	Haurem de repartir les tasques assignades a la persona en qüestió entre els membres restants.
Algun integrant del grup no faci la seva tasca o no posi interès en el treball	Donar-li tocs d'atenció.	Al final del treball fer una avaluació justa d´aquest membre de l´equip.
Un membre de l'equip no sap fer la seva part del treball	Quan un membre de l'equip tingui dubtes ho comunicarà a la resta abans que sigui massa tard.	Si es disposa d'algun integrant capaç de dur a terme aquesta tasca se li demanarà col·laboració amb la persona corresponent.
Dos dels integrants han realitzat la mateixa part del treball per separat i els dos pensen que la seva és millor	Els membres a càrrec de la mateixa tasca hauran de repartir el treball de manera que no hi hagi duplicats.	Es debatrà entre tot el grup quina part s´ha d´escollir.
No obtenir tot el que es desitja de la base de dades i per tant, deixar la feina incompleta	Aconseguir una base de dades en la qual totes les variables siguin conegudes i totalment enteses pels integrants de l´equip.	Demanar ajuda al professor i intentar buscar alguna solució entre tots.

No es respecta el calendari de coordinadors i algú agafa el control contínuament	Cadascú ha de ser respectuós amb les pautes marcades i intentar no agafar el control quan no li correspon.	El grup haurà de reunir-se i parlar d'aquest problema fins que es solucioni.
No hi ha una bona comunicació i ningú puja el treball o el lliurament que toca pensant que ho feia l´altre	Abans que acabi el temps de lliurament tots s´han d´assegurar que la feina o lliurament està pujat.	Explicar el problema i intentar que no torni a passar millorant la comunicació i fent que tots s´assegurin que el treball estigui pujat.

Figura 4: Pla de Riscos

3. Estructura de dades i descriptiva.

i. Motivació del treball.

La raó per la qual es va decidir realitzar el treball sobre aquesta base de dades, relativa a la indústria hotelera, ha sigut l'interés de comparar els resultats obtinguts l'any passat al treball realitzat a l'assignatura d'Anàlisi Multivariant amb els que podriem obtenir utilitzant tècniques de mineria de dades.

L'objectiu principal que perseguim amb la realització del present projecte, és tractar d'establir un model, basat en les valoracions i puntuacions que han realitzat els hostes, per tal de detectar característiques (tant del client com de l'hotel) que estiguin associades a una puntuació elevada, o per contra, molt baixa.

ii. Descripició formal de l'estructura de dades.

Tal i com s'ha mencionat anteriorment, tot l'anàlisi es duu a terme amb una mostra aleatòria de 5.000 observacions i una selecció de 30 variables que es consideren rellevants per als objectius perseguits (*Base de dades original 515.738 files per 41 columnes*). En aquest sentit, cada observació de la matriu de dades representa una ressenya escrita a Booking per un client que ha visitat cert hotel registrat al portal.

Tot seguit, es presenta el llistat de les variables seleccionades, així com les metadades requerides en cada cas. Cal destacar que tota la informació que apareix al diccionari de dades es basa en la base de dades un cop fet la selecció dels registres, i no sempre serà aplicable a les dades originals. Adicionalment, de cara a mencions posteriors, hem considerat oportú indexar les variables, de manera que, en endavant, podem referenciar una variable pel seu índex.

El valor NULL a les variables *Businesses_100m, Businesses_1km, Businesses_5km, Room_Type_Level, Trip_Type* indica dada faltant, mentre que a les variables *Hotel_lat, Hotel_lng* i *Negative_Review* tenim NA. Per últim la variable *Reviewer_Nationality* codifica les dades faltants mitjançant espais buits.

(Els nivells de les variables categòriques apareixen ordenats)

1. *id* (integer): Identifica cada ressenya.
   a. **rang** : {170 , 515.740}
   b. **rol**: variable índex
2. *Hotel_Name* (qualitative): Nom de l'hotel.
   a. **modalitats**: {11 Cadogan Gardens , 1K Hotel , 25hours Hotel beim MuseumsQuartier , 88 Studios , 9Hotel Republique , Abba Sants , AC Hotel Barcelona Forum a Marriott Lifestyle
   Hotel , AC Hotel Diagonal L Illa a Marriott Lifestyle Hotel , AC Hotel Milano a Marriott Lifestyle Hotel , AC Hotel Sants a Marriott Lifestyle Hotel, ...} *(1135 modalitats)* b. **rol**: variable explicativa
3. *Hotel_Country* (qualitative): País de l'hotel.
   a. **modalitats**: {AT, ES , FR , GB , IT , NL}
   b. **rol**: variable explicativa

4. *Hotel_City* (qualitative): Ciutat de l'hotel.
   a. **modalitats**: {Amsterdam , Amsterdam Zuidoost , Barcelona , Boulogne Billancourt , Donauinsel , El Prat de Llobregat , Fitzrovia , London , Milan , Paddington , Paris , Paris 06 , Paris 12 , Vienna , Vincennes , Woodford Green}
   b. **rol**: variable explicativa

5. *Hotel_lat* (numeric): Latitud de l'hotel.
   a. **rang**: {41.32838, 52.40018}
   b. **rol**: geolocalització
   c. **missing_code**: NA

7. *Businesses_100m* (integer): Nombre de negocis a 100 metres a la rodona de l'hotel.
   a. **rang**: {1, 124}
   b. **rol**: variable explicative
   c. **missing_code**: NULL.

8. *Businesses_1km* (integer): Nombre de negocis entre els 100 metres fora de la rodona del hotel fins a 1km a la rodona de l'hotel.
   a. **rang**: {1, 336}
   b. **rol**: variable explicativa
   c. **missing_code**: NULL

9. *Businesses_5km* (integer): Nombre de negocis entre 1km fora de la rodona del hotel fins a 5km a la rodona de l'hotel.
   a. **rang**: {1, 242}
   b. **rol**: variable explicativa
   c. **missing_code**: NULL

10. *Room_Type_Level* (qualitative): Tipus d'habitació contractada per l'usuari que ha escrit la ressenya.
    a. **modalitats**: {Ambassadors, Art, Business, Business Class, City, Classic, Deluxe, Duplex, Executive, Family, Luxury, NULL, Premium, Privilege, Standard, Studio, Suite, Superior}
    b. **rol**: variable explicativa
    c. **missing_code**: NULL

11. *Guest_Type* (qualitative): Perfil del client que ha escrit la ressenya, obtingut a partir de tags.
    a. **modalitats**: {Couple, Family with older children, Family with young children, Group, Solo traveler, Travelers with friends, With a pet}
    b. **rol**: variable explicativa

12. *Trip_Type* (qualitative): Tipus de viatge realitzat pel client que ha escrit la ressenya, obtingut a partir de tags.
    a. **modalitats**: {Business trip, Couple, Family with older children, Family with young children, Leisure tri , NULL, Solo traveler}
b. **rol**: variable explicativa

13. *Stay_Duration* (integer): Total de nits d'estada.
    a. **rang**: {1, 20}
    b. **rol**: variable explicativa

14. *Review_Date* (data: yyyy-mm-dd): Data en la qual l'usuari ha escrit la ressenya a Booking.
    a. **rang**: {2015-08-04, 2017-08-03}
    b. **rol**: variable explicativa

15. *Days_Since_Review* (integer): Diferència de dies entre la data en la qual l'usuari ha escrit la ressenya a Booking i la data de *checkout*.
    a. **rang**: {0, 720}
    b. **rol**: variable explicativa

16. *Is_Hotel_Holiday* (qualitative): Variable binària que indica si va ser festiu a la ciutat on es troba l'hotel, a la Review date.
    a. **modalitats**: {0: No, 1: Yes}
    b. **rol**: variable explicativa

17. *Is_Reviewer_Holiday* (qualitative): Variable binària que indica si va ser festiu al pais del client, a la Review date.
    a. **modalitats**: {0: No, 1: Yes}
    b. **rol**: variable explicativa

18. *Total_Number_of_Reviews* (integer): Nombre total de ressenyes vàlides que té l'hotel a Booking.
    a. **rang**: {60, 16670}
    b. **rol**: variable explicativa

19. *Review_Is_Positive* (qualitative): Variable binària que indica si el nombre de paraules a la variable *Review Total Positive Word Counts* és major que a la variable *Review Total Negative Word Counts*.
    a. **modalitats**: {0: No, 1: Yes}
    b. **rol**: variable explicativa

20. *Review_Positivity_Rate* (numeric): Mesura el grau de positivisme de la ressenya fent la mitjana ponderada del total de paraules a la ressenya positiva (*Review Total Positive Word Counts*) sobre la suma del total de paraules tant en la positiva com en la negativa (*Review Total Negative Word Counts*).
    a. **rang**: {0, 100}
    b. **rol**: variable resposta

21. *Reviewer Nationality* (qualitative): Nacionalitat de l'usuari que ha escrit la ressenya.
    a. **modalitats**: {"" , Abkhazia Georgia, Albania, Andorra, Angola, Argentina, Armenia, Australia, Austria, Azerbaijan, Bahrain, Bangladesh,...} *(123 modalitats)*
    b. **rol:** variable explicativa
    c. **missing_code:** ""

22. *Negative_Review* (text): Ressenya negativa escrita per l'usuari.
    a. **rol**: text
    b. **missing_code:** "Na"

23. *Review_Total_Negative_Word_Counts* (integer): Nombre total de paraules a la ressenya negativa escrita per l'usuari.
    a. **rang**: {0, 372}
    b. **rol**: variable explicativa

24. *Positive_Review* (text): Ressenya positiva escrita per l'usuari.
    a. **rol**: text
    b. missing_code: "Na"

25. *Review_Total_Postive_Word_Counts* (integer): Nombre total de paraules a la ressenya positiva escrita per l'usuari.
    a. **range**: {0, 247}
    b. **rol**: variable explicativa

26. *Average_Score* (numeric): Valoració mitjana de l'hotel a la pàgina de Booking a data 31 de desembre de 2016.
    a. **range**: {5.2, 9.6}
    b. **rol**: variable explicativa

27. *Reviewer_Score* (numeric): Valoració global otorgada per l'usuari que ha escrit la ressenya.
    a. **range**: {2.5, 10}
    b. **rol**: variable explicativa

28. *Total_Number_of_Reviews_Reviewer_Has_Given* (integer): Nombre total de ressenyes a la web de Booking escrites per l'usuari.
    a. **range**: {1, 156}
    b. **rol**: variable explicativa

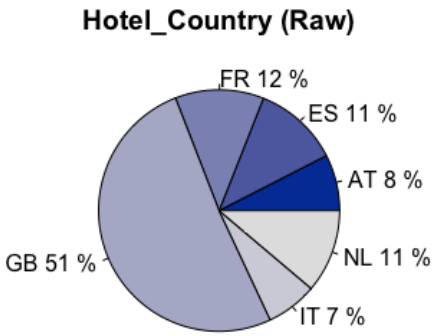
29. *Additional_Number_of_Scoring* (integer): Nombre total de valoracions adicionals vàlides sobre diferents aspectes de l'hotel.
    a. **range**: {8, 2682}
    b. **rol**: variable explicativa

30. *Submited_from_Mobile* (qualitative): Variable binària que indica si la ressenya s'ha pujat a Booking via telèfon mòvil.
    a. **modalitats**: {0: No, 1: Yes}
    b. **rol**: variable explicativa

iii. Anàlisi descriptiva univariant inicial.

Un cop definida la base de dades, convé fer un anàlisi exploratori inicial de les variables amb l'objectiu de millorar la percepció que tenim sobre aquestes, i descobrir més a fons la seva estructura. Això s'aconseguirà a partir de gràfics i mesures estadístiques univariants de les nostres variables. A més, també ens permetrà detectar anomalies, que posteriorment corregirem a la fase de preprocessament.

Les primeres variables *id* i *Hotel_Name* no les representem a cap gràfic, ja que són identificadors dels registres. Així doncs, començem amb la variable *Hotel_Country (Figure 5)*.




```
 AT  ES  FR   GB  IT  NL
376 574 595 2554 343 558
```

Figura 5: Pie Chart Hotel_Country

Aquesta, apareix codificada com una variable qualitativa de 6 modalitats amb nombre d'individus a cada modalitat igual al resultat anterior.

Es pot observar que la meitat de la mostra de l'estudi es troba al Regne Unit, seguit de França, Espanya i els Països Baixos, entre els quals componen un 1/3 del total.

La següent variable a estudiar és *Hotel_City*. Aquesta també apareix codificada com una variable qualitativa amb 16 modalitats, de manera que el procediment a aplicar és idèntic a l'anterior *(Figure 6)*.

```
  Amsterdam   Amsterdam Zuidoost
        545                  13
  Barcelona Boulogne Billancourt
        570                   3
Donauinsel El Prat de Llobregat
          6                   4
  Fitzrovia              London
         23                2378
      Milan           Paddington
        343                 150
      Paris            Paris 06
        587                   1
   Paris 12              Vienna
          1                 370
  Vincennes       Woodford Green
          3                   3
```

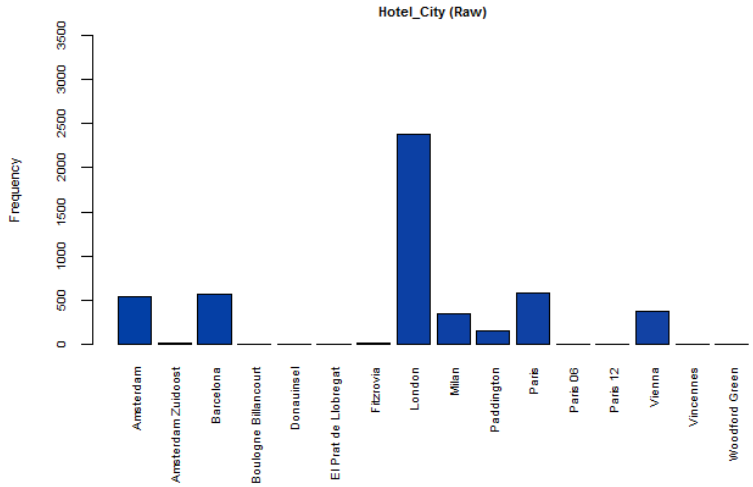


Figura 6: Bar plot Hotel_City

Al igual que hem vist per al cas univariant per països, el Regne Unit, és qui mes hotels aporta a l'estudi, localitzant-se la majoria a la capital, Londres. També, al igual que en el cas anterior, en una segona escala trobem les capitals dels Països Baixos i França. La totalitat dels hotels espanyols analitzats es troben a Barcelona. Adicionalment, tenim un ampli ventall de ciutats i districtes que, de cara el preprocessing, en reduirem el nombre per tal de sintetitzar la informació i analitzar les ciutats més rellevants.

A continuació, tractem conjuntament les variables de localització latitud i longitud. Considerem fer un resum numèric per a totes dues variables, acompanyat de dos boxplots *(Figure 7)*.

Hotel_lat

Min.	*1st Qu.*	*Median*	*Mean*	*3rd Qu.*	*Max.*	*NA's*
41.33	*48.21*	*51.50*	*49.47*	*51.52*	*52.40*	*23*

Hotel_lng

Min.	*1st Qu.*	*Median*	*Mean*	*3rd Qu.*	*Max.*	*NA's*
-0.36980	*-0.14390*	*-0.00025*	*2.81000*	*4.83100*	*16.42000*	*23*

Aquí tenim les primeres dades mancants (NA). Cal pendre nota d'aquesta anomalia per a corregir-la a la fase de preprocessament.

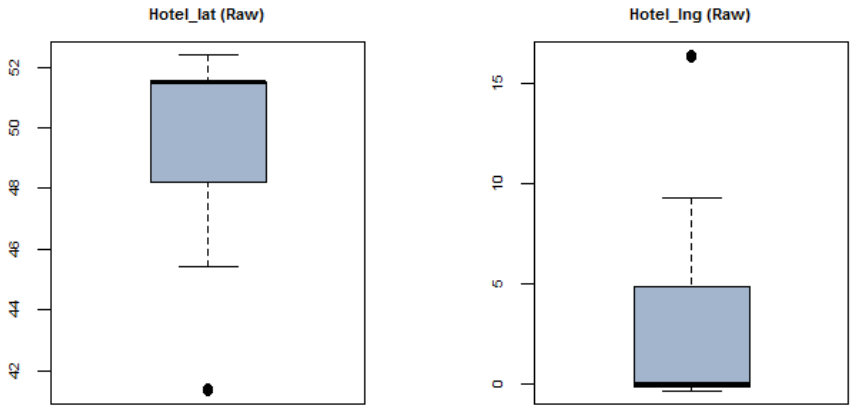


Figura 7: Boxplot Mesures de localització

Pel que fa a la latitud on es troben els hotels, observem que la majoria es troben al voltant de 51ºN, el qual passa per països com el Regne Unit, França o Bèlgica. D'altra banda, mencionar que en el nostre estudi només hi trobem hotels entre les latituds 41º i 52º N, els quals coincideixen, òbviament amb les latituds de la majoria dels països europeus.

Complementàriament, per conèixer la localització dels hotels també necessitem la longitud, que és una linia imàginaria que va de pol a pol. Aquí podem observar com el gran gruix dels hotels es troben al voltant del 0º, és a dir, al voltant del Meridià de Greenwich, el qual passa per Espanya, França i el Regne Unit.

A continuació, passem a analitzar les variables *Businesses_100m, Businesses_1km, Businesses_5km*. A primera vista, ja veiem que aquestes tres variables apareixen mal codificades (les tenim com a factors i haurien de ser numèriques). En conseqüènia, les apartem per analitzar-les un cop haguem fet el preprocessament i les tinguem en el tipus adequat. Tanmateix sabem que tenim 23 valors missings per a aquestes variables que es desprenen de la falta d'informació de mesures de localització i coincideixen en les observacions (caldrà tenir-ho en consideració per al preprocessament).

Seguidament, considerem la variable *Room_Type_Level* i observem quins són els tipus d'habitacions més freqüents *(Figure 8)*. De nou, la variable és qualitativa i repliquem el procediment descrit anteriorment per a variables qualitatives.

Ambassadors	Art	Business	Business Class
1	18	58	3
City	Classic	Deluxe	Duplex
9	343	191	10
Executive	Family	Luxury	NULL
83	131	9	3209
Premium	Privilege	Standard	Studio
1	8	546	23
Suite	Superior		
71	286		

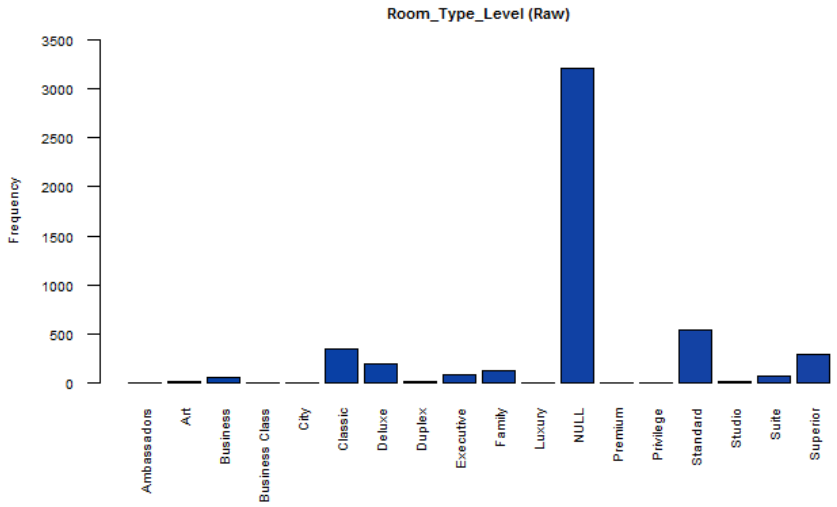


Figura 8: Barplot Room_Type_Level

En aquest cas, el tipus d'habitacions que més abunden tendeixen a ser diferents a les habituals, doncs la més repetida és null. Aquesta variable vol dir que no es refereix a cap de les altres categories, potser degut a una falta d'estandardització dels noms de les habitacions (cada hotel pot fer servir noms diferents per a referir-se al mateix tipus d'habitació).
Tot i això, si ens quedem amb les denominacions de les que disposem, obtenim que les tres més freqüents són: Standard, Superior, Deluxe i Classic. Finalment, dir que tenim un nombre considerable de modalitats per a la variable i, potser, hauríem de considerar reduïr-lo a la fase de preprocessament.

A continuació, tractem conjuntament les variables *Guest_Type* i *Trip_Type*, ja que estan força relacionades entre sí en referència al perfil del client i el viatge que busca. Fem servir la mateixa metodologia, tots dos són factors amb 7 modalitats (*Figure 9*).

Guest_type

```
      Couple Family with older children Family with young children
        2425                       246                        569
       Group              Solo traveler       Travelers with friends
         658                      1077                         21
With a pet
           4
```

Trip_type

```
      Couple Family with older children Family with young children
        2425                       246                        569
       Group              Solo traveler       Travelers with friends
         658                      1077                         21
With a pet
           4
```

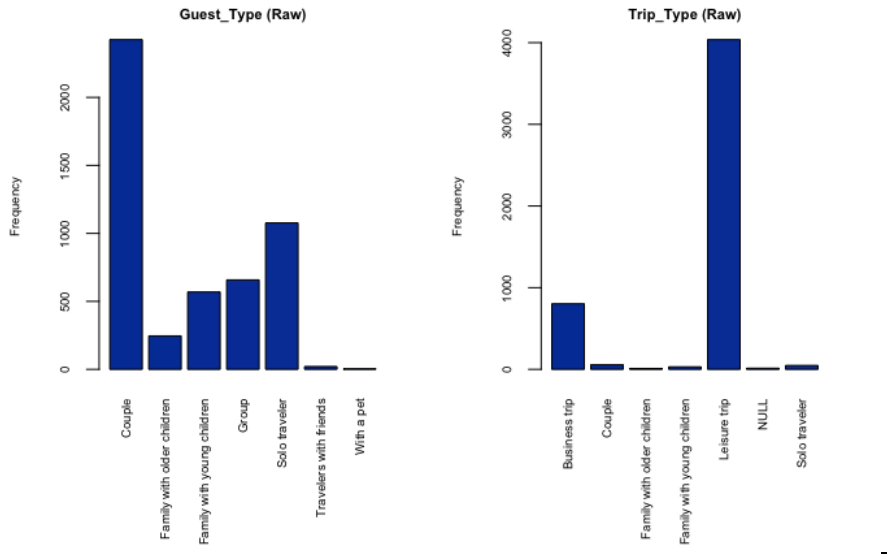


Figura 9: Barplot perfils client/viatge

Realitzant els gràfics i els recomptes per variable, podem concloure que les parelles són el tipus d'hostes més freqüents pel que fa a la variable *Guest_Type*, mentre que en segona posició tindriem els que realitzen el viatge en solitari (ja sigui per motius laborals o pròpiament d'oci). En referència a la variable *Trip_Type* estudiada, s'observa clarament com el tipus de viatge més freqüent és el d'oci, seguit a molta distància pel viatge amb motius de negocis, éssent la resta valors (inclosa la categoria NULL) residuals. De cara el preprocessing ajuntarem les variables family en una sola.

La següent variable a estudiar és la que informa sobre la duració de l'estància *Stay_Duration.* Aquesta, la tenim codificada com a numèrica i podem explorar-la a través d'un resum numèric i representar-la gràficament amb un histograma (*Figure 10*).

Min.	*1st Qu.*	*Median*	*Mean*	*3rd Qu.*	*Max.*
1.000	*1.000*	*2.000*	*2.391*	*3.000*	*20.000*

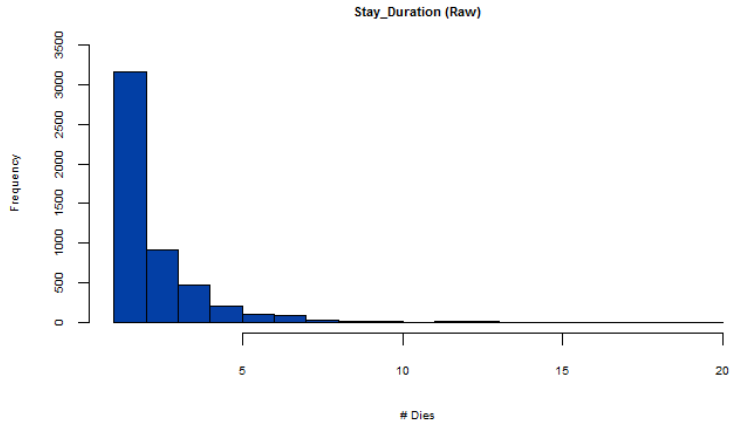


Figure 10: Histograma Stay_Duration

Veiem com els dies que un hoste passa a l'hotel com a màxim arriba a 20, però si ens quedem amb els més freqüents diríem que entre 1 i 3, donat que aquest interval comprén, aproximadament, el 75% de les observacions.
La següent variable *"Review_Date"* fa referència a la data en la que es va escriure la ressenya a Booking. En aquest cas, apareix codificada com a numèrica. Per a poder treballar amb ella caldrà que la transformem en data del tipus *yyyy/mm/dd* (pendent de preprocessament).

Passa una cosa semblant amb la variable *"Days_Since_Review",* actualment factor amb 720 nivells. D'aquesta variable ens interessarà tenir, codificat com a variable numèrica, el nombre de dies que han transcorregut (pendent de preprocessament).

Seguidament, les variables binàries *"Is_Hotel_Holiday"* i *"Is_Reviewer_Holiday"* apareixen codificades com a numèriques, quan interessaria més tenir-les com a factors. En aquest sentit, plantegem igualment construir la proporció de 1 (Sí) que tenim sobre el total mitjançant un pie chart (*Figure 11*).

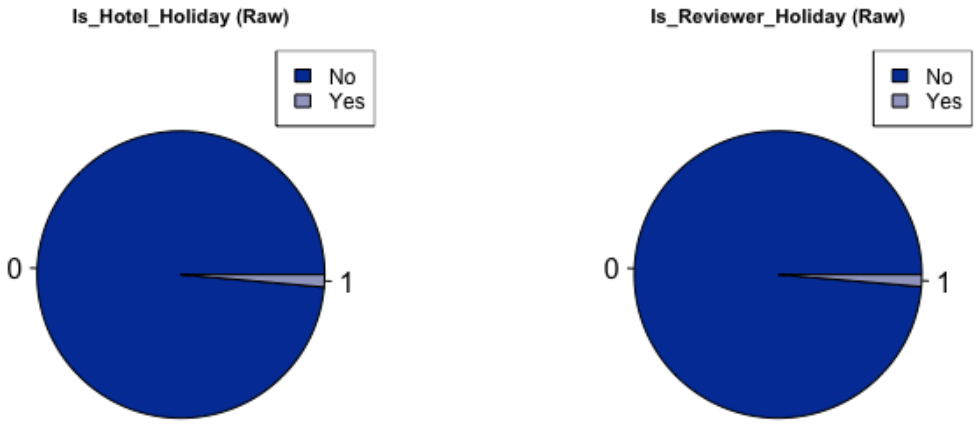


Figura 11: Pie Chart Festivitats

En primer lloc, s'observa que, majoritàriament, l'hotel no es troba en dia festiu en la data que es va escriure la ressenya. En segon lloc, al igual que per a l'hotel, també és pràcticament total la resposta no, en referència a la festivitat al país d'origen del client. Totes dues freqüències són semblants a les dues variables. Per depurar més l'anàlisi podriem considerar analitzar si els valors 1 per a una de les dues es corresponen amb els valors 1 de l'altra. Veiem que hi ha 65 casos dels 67 de la varible *"Is_Reviewer_Holiday"* que coincideixen.

Seguidament, analitzem la variable "*Total_Number_of_Reviews*". Aquesta apareix codificada correctament com a numèrica així que construïm un histograma *(Figure 12)* i el complementem amb un resum numèric, com en els casos anteriors.

Min.	*1st Qu.*	*Median*	*Mean*	*3rd Qu.*	*Max.*
60	*1179*	*2135*	*2732*	*3611*	*16670*

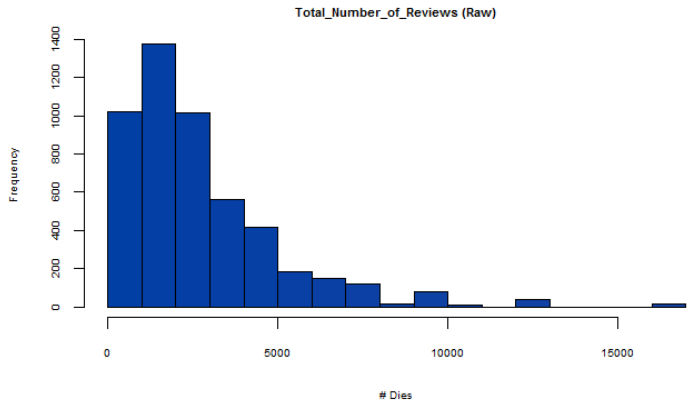


Figure 12: Histograma Total_Number_of_Reviews

El nombre total de ressenyes, oscil.la entre 60 i 16.670, existint la concentració màxima entre els 60 i 5000. Veiem com la variable mostra una distribució semblant a una Chi_Quadrat, igual que la variable "Stay_Duration". Aquest fet, sembla llògic ja que les dues tenen un zero natural.

A continuació tenim dues variables que reflecteixen el grau de positivisme del comentari. En aquest sentit, hem considerat tractar-les conjuntament tot i que la primera d'elles *"Review_Is_Positive",* apareix mal codificada (és una variable binària i per tant hauria de ser un factor, no una variable numèrica). En aquest cas, considerem un resum numèric per a la variable *"Review_Positivity_Rate"* i un gràfic conjunt (histograma) d'aquesta amb *"Review_Is_Positive"* (Figure 13).

Min.	*1st Qu.*	*Median*	*Mean*	*3rd Qu.*	*Max.*
0.00	*28.10*	*52.94*	*55.73*	*94.30*	*100.00*

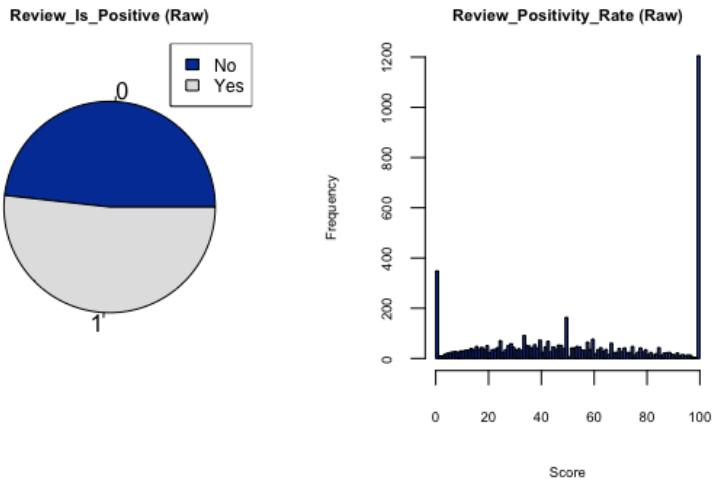


Figure 13: Gràfics variables positivisme de la ressenya

En primer lloc, dir que aquestes variables tenen molta rellevància en l'anàlisi ja que representen, o estan molt relacionades, amb la resposta que volem estudiar. Es pot observar que les ressenyes positives són lleugerament més elevades però pràcticament observem un empat.

En segon lloc, en referència al ratio de positivitat, al gràfic es veu com la variable presenta molta variabilitat. Ara bé, si ens centrem en els valors que es situen una mica per sobre de la línia que marca les variables menys freqüents, observem com aquestes són 0, 25, 33.3, 50, 60, 66.6 i 100, els quals podriem dividir en dos subgrups, un primer per 0 i 10, i un segon per la resta. Això vol dir que les valoracions són molt polaritzades i caldria considerar si definitivament seria bo fer servir aquesta variable com a resposta.

A continuació tenim la variable "*Reviewer_Nationality*" qualitativa i amb 123 modalitats. Com a conseqüència d'aquest nombre tan elevat de nivells del factor, els gràfics es tornen il·legibles.

Seguidament, comentem de manera conjunta les variables *"Negative_Review", "Review_Total_Negative_Word_Counts", "Positive_Review i Review_Total_Positive_Word_Counts".* La primera i la tercera contenen, per a cada registre, una cadena de caràcters que fa referència a la ressenya completa que l'usuari ha escrit a la pàgina web de Booking. Considerem, de moment, deixar-les de banda per a l'anàlisi textual que realitzarem a l'últim capítol del treball. Ara, si ens centrem en les dues restants, observem com apareixen degudament codificades (variables numèriques), i pot ser interessant fer un boxplot conjunt de les dues per a comparar-les *(Figure 14).* A priori, hauríen de ser complementàries, aquelles ressenyes amb major nombre de paraules als comentaris positius hauren de tenir un nombre molt petit de paraules als comentaris negatius, i viceversa. Acompanyem els boxplots amb resums numèrics per a les dues variables (negative/positve).

Review_Total_Negative_Word_Counts

Min.	*1st Qu.*	*Median*	*Mean*	*3rd Qu.*	*Max.*
0.00	*2.00*	*10.00*	*19.38*	*24.00*	*372.00*

Review_Total_Positive_Word_Counts

Min.	*1st Qu.*	*Median*	*Mean*	*3rd Qu.*	*Max.*
0.00	*5.00*	*11.00*	*17.34*	*22.00*	*247.00*

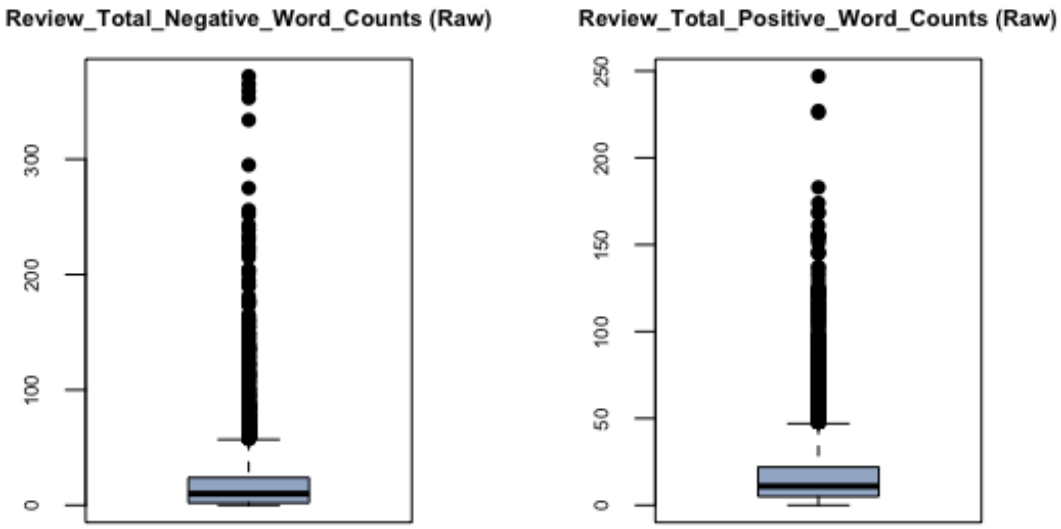


Figure 14: Boxplot Paraules als comentaris negatius/positius

Pel que fa al nombre de paraules negatives per ressenya, destacar que la freqüència modal és entre 0 i 20, trobant-se la mitjana en 19,38. Si ara considerem els comentaris positius, obtenim que podríem reduir més aquest interval i situar-lo entre 0 i 10, tot i trobar-se la mitjana en 17. Això és indicatiu que quan la experiència no ha estat bona l'usuari tendeix a escriure ressenyes més llargues.

A continuació, tractarem les variables *"Average_Score"* i *"Reviewer_Score",* ambdues de gran rellevància per a l'anàlisi. Apareixen codificades correctament, de manera que podem construir resums numèrics i boxplots per a les dues.

Average score

Min.	*1st Qu.*	*Median*	*Mean*	*3rd Qu.*	*Max.*
5.200	*8.100*	*8.400*	*8.395*	*8.800*	*9.600*

Reviewer score

Min.	*1st Qu.*	*Median*	*Mean*	*3rd Qu.*	*Max.*
2.500	*7.500*	*8.800*	*8.373*	*9.600*	*10.000*

Observem com les puntuacions que han anat otorgant els usuaris són, en promig, similars a la valoració mitjana acumulada pels hotels a finals de l'any 2016 *(Figure 15)*. Tanmateix, la variabilitat en la valoració dels usuaris és més alta, situació llògica ja que la variable "*Average_Score*" és un promig. En general, una valoració agregada de tots els hotels estaria al voltant de 8,3 i seria interessant en seccions posteriors veure com evoluciona la puntuació que otorguen els clients al llarg del temps (en funció de la variable "*Review_Date*").

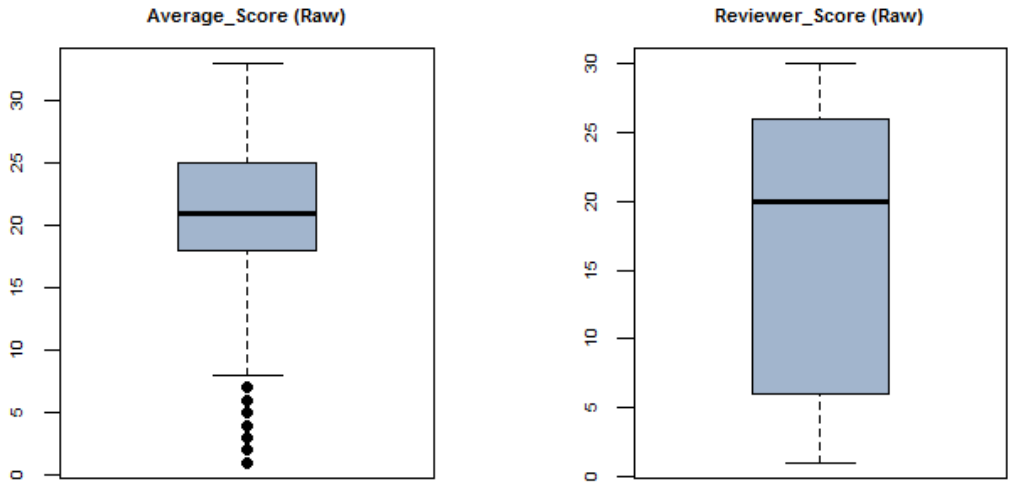


Figura 15: Boxplot Puntuacions

Seguidament, passem a la variable "*Total_Number_of_Reviews_Reviewer_Has_Given*", indicador de si l'usuari és molt actiu, o no, al portal web. La variable és numèrica i la tenim codificada correctament, així que, com en els casos anteriors, elaborem un resum numèric i un histograma (*Figure 16*).

Min.	*1st Qu.*	*Median*	*Mean*	*3rd Qu.*	*Max.*
1.000	*1.000*	*3.400*	*7.323*	*9.000*	*156.000*

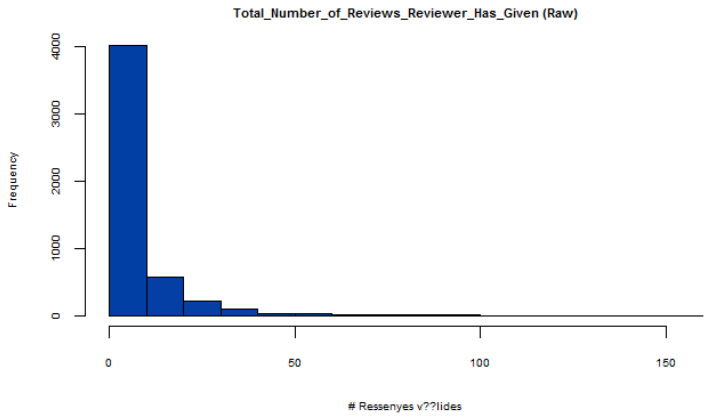


Figure 16: Histograma Total_Number_of_Reviews_Reviewer_Has_Given

Veiem com, per més d'un 25% dels usuaris, és la primera ressenya que escriuen. La mitjana es situa en 7,3, i el 75% dels valors es troben compresos entre 1 i 9. No sembla un nombre de ressenyes massa elevat (poca participació).

Seguidament analitzem la variable "*Additional_Number_of_Scoring*", cas idèntic al de la variable anterior, però ara considerem totes les valoracions adicionals (serveis, localització etc..) vàlides que té l'hotel en questió al portal de Booking (*Figure 17*).

Min.	*1st Qu.*	*Median*	*Mean*	*3rd Qu.*	*Max.*
8.0	*170.0*	*342.5*	*497.4*	*666.0*	*2682.0*

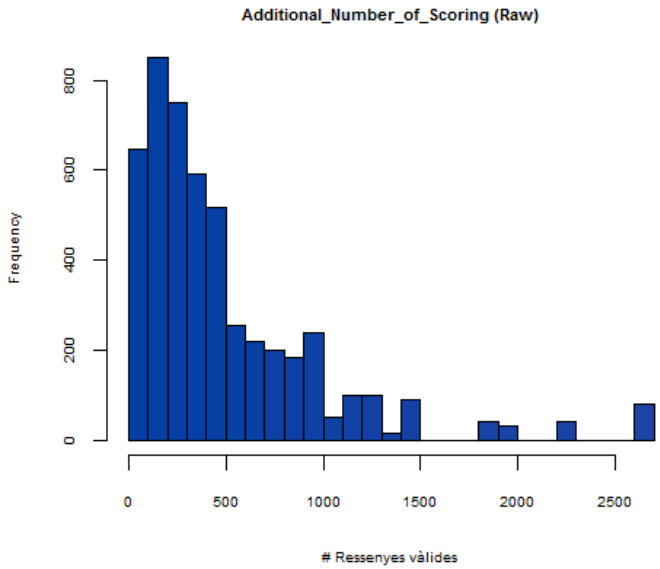


Figure 17: Histograma Additional_Number_of_Scoring

Veiem com el nombre de puntuacions adicionals va des de 8 (mínim) fins a 2682 (màxim). Cal destacar que els dos quartils són més propers, entre 170 i 666 unitats, i per tant podríem considerar que tenim valors anòmals a la variable (caldria estudiar en detall aquestes observacions amb valors extrems). Tanmateix, observem com, en general, les puntuacions de serveis són més elevades que les de ressenyes escrites, ja que el mètode és més còmode per a l'usuari (marcar estrelles vs. escriure un comentari).

Per finalitzar aquesta primera part d'anàlisi exploratòri de dades considerem la variable "*Submitted_from_Mobile*" binària. De nou, la variable binària apareix codificada com a numèrica i caldrà tractar-la a la fase de preprocessament per codificar-la correctament. Tanmateix, podem emprar el mateix procediment que per a les altres vairables binàries i construir un pie chart (*Figure 18*). Observem com predominen les ressenyes escrites des del telèfon mòbil.

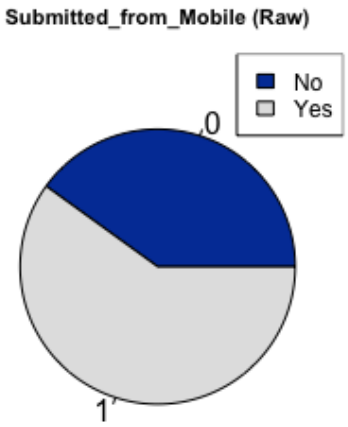


Figure 18: Pie Chart Submitted_from_Mobile

iv. Descripció detallada del procés de preprocessament.

Un cop feta l'aproximació incial a les dades que hem platejat a l'apartat anterior, en la fase de preprocessament el que es busca és corregir els errors o problemes detectats (outliers, NAs ...) que perjudicaràn la qualitat dels anàlisis posteriors.

En primer lloc, transformem el tipus/classe de les variables mal codificades. Les variables *Businesses_100m*, *Businesses_1km*, *Businesses_5km* i *Days_Since_Review*, les hem transformat en variables numèriques enteres. D'altra banda les variables *Is_Hotel_Holiday*, *Is_Reviewer_holiday*, *Submitted_from_Mobile* i *Review_Is_Positive*, les hem transformat en factor. Totes elles són variables dicotòmiques i, en conseqüència, hem declarat els nivells (1: Sí i 0:No). Per últim, codifiquem com a data la variable *Review_Date*. En aquest apartat podem meniconar també que les variables textuals *Positive_Review* i *Negative_Review* no s'ajusten exactement a un factor amb diferents nivells així que les podríem codificar com cadenes de caràcters. Tanmateix, com que només les farem servir en un àmbit reduït de l'anàlisi i, el seu estat actual no representa un problema, hem considerat mantenir-les com a factor per convertir-les a caràcter quan realitzem l'anàlisi textual.

Un cop tenim totes les variables en el tipus d'R convenient, ens centrem en les modalitats de les variables qualitatives. Observem els diferents nivells que presenten, a partir de definir una funció que selecciona les variables de la base de dades segons de la seva classe, per tal d'extreure només els factors i mirar els seus nivells (recordem que cal obviar les variables textuals). Com a filtre adicional,

considerem tractar únicament aquells factors amb nombre de nivells inferior a 150. A banda de les variables textuals, la variable *Hotel_Name* tampoc la considerem ja que cada nom és únic i serveix per a identificar completament un registre, no té sentit aplicar cap tipus d'agrupació de modalitats en aquest cas.

Variables	**Nivells**
3: Hotel_Country	3: {AT, ES, FR, GB, IT, NL}
4: Hotel_City	5: {Amsterdam, Amsterdam Zuidoost, Barcelona, Boulogne Billancourt, Donauinsel, El Prat de Llobregat, Fitzrovia, London, Milan, Paddington, Paris, Paris 06, Paris 12, Vienna, Vincennes, Woodford Green}
10: Room_Type_Level	10: {Ambassadors, Art, Business, Business Class, City, Classic, Deluxe, Duplex, Executive, Family, Luxury, NULL, Premium, Privilege, Standard, Studio, Suite, Superior} 11: {Couple, Family with older children, Family with young children, Group, Solo traveler, Travelers with friends, With a pet}
11: Guest_Type	12: {Business trip Couple Family with older children, Family with young children, Leisure trip, NULL, Solo traveler}
12: Trip_Type	16: {0, 1}
16: Is_Hotel_Holiday	17: {0, 1}
17: Is_Reviewer_Holiday	19: {0 1}
19: Review_Is_Positive	21: {Abkhazia Georgia, Albania, Andorra, Angola, Argentina, Armenia, Australia, Austria, Azerbaijan, Bahrain, Bangladesh, Barbados, Belgium, Bermuda, Bosnia and Herzegovina, Botswana, Brazil, Brunei, Bulgaria, Cameroon, Canada, Chile, China, Colombia, Comoros, Costa Rica, Croatia, Cura ao, Cyprus, Czech Republic, Denmark, Dominican Republic, Egypt, Estonia, Ethiopia, Finland, France, Gabon, Georgia, Germany, Gibraltar, Greece, Guernsey, Honduras, Hong Kong, Hungary, Iceland, India, Indonesia, Iran, Iraq, Ireland, Isle of Man, Israel, Italy, Ivory Coast, Jamaica, Japan, Jersey, Jordan, Kazakhstan, Kenya, Kosovo, Kuwait, Latvia, Lebanon, Liechtenstein, Lithuania, Luxembourg, Macedonia, Malaysia, Malta, Mauritius, Mexico, Moldova, Monaco, Montenegro, Morocco, Namibia, Nepal, Netherlands, New Caledonia, New Zealand, Nigeria, Norway, Oman, Pakistan, Panama, Peru, Philippines, Poland, Portugal, Puerto Rico, Qatar, Romania, Russia, Saint Barts, Saint Lucia, Saudi Arabia, Serbia, Singapore, Slovakia, Slovenia, South Africa, South Korea, Spain, Sri Lanka, Sweden, Switzerland, Taiwan, Thailand, Tunisia, Turkey, Uganda, Ukraine, United Arab Emirates, United Kingdom, United States, Minor Outlying Islands, United States of America, Venezuela, Vietnam, Zambia}
21: Reviewer_Nationality	30: {0 1}
30: Submitted_from_Mobile	

Taula 2: Modalitats de les variables qualitatives (continued below)

En aquest sentit, volem analitzar els nivells de cada variable a fi d'esbrinar si existeixen redundàncies entre aquests, si podem eliminar algun, o si simplement les etiquetes no són convenients. En general, és recomanable recategoritzar factors amb nivells que reflexen el mateix, o aquelles variables amb un gran nombre de categories per a facilitar anàlisis posteriors.

En primer lloc, assignem etiquetes (No, Sí) a les variables dicotòmiques. Seguidament, plantegem redefinir les categories de la resta de variables qualitatives de la taula anterior. Un cop arribats a aquest punt, és convenient mencionar que en aquest procediment es realitzarà el tractament dels valors missing de les variables categòriques. Mentre que, per a les variables numèriques necessitarem realitzar una imputació de valors, aquí podem simplement agrupar les dades mancants en una nova categoria (p.e. "Altres") i declarar-les com un nou nivell del factor.

La primera variable és *Hotel_Country*. En aquest cas, veiem com no tenim valors missing i els nivells de la variable són únicament 6 (AT, ES, FR, GB, IT, NL). En aquest sentit, es conclou que aquesta variable no necessita cap tipus de tractament.

En segon lloc, considerem la variable *Hotel_City* i repetim la operació (en aquest cas la variable no té valors missing).

- *ANTICS NIVELLS*: Amsterdam, Amsterdam Zuidoost, Barcelona, Boulogne Billancourt, Donauinsel, El Prat de Llobregat, Fitzrovia, London, Milan, Paddington, Paris, Paris 06, Paris 12, Vienna, Vincennes, Woodford Green.
- *NOUS NIVELLS*: Amsterdam{Amsterdam, Amsterdam Zuidoost}, Barcelona{Barcelona, El Prat de Llobregat} , Boulogne Billancourt{Boulogne Billancourt}, Vienna{Donauinsel, Vi- enna} , London{Fitzrovia, London, Paddington, Woodford Green}, Milan{Milan}, Paris{Paris, Paris 06, Paris 12}, Vicennes{Vincennes}.

Seguim amb el tractament de la variable *Room_Type_Level*. Recordem que, per a aquesta variable, tenim els missings codificats com a NULL. Definim la correspondència entre els antics nivells de la variable i els nous:

- *ANTICS NIVELLS*: Ambassadors, Art, Business, Business Class, City, Classic, Deluxe, Du- plex, Executive, Family, Luxury, NULL, Premium, Privilege, Standard, Studio, Suite, Supe- rior.
- *NOUS NIVELLS*: Deluxe{Ambassadors, Art, Deluxe, Executive, Luxury, Premium, Privi- lege, Superior} , Business{Business, Business Class} , Classic{City, Classic} , Duplex{Duplex} , Family{Family} , Other{NULL} , Standard{Standard, Studio}, Suite{Suite}.

Observeu que, els NULL que teníem al principi els hem recodificat com a "Other".

En relació a la variable Guest_Type, mencionar que aquesta no rep cap tipus de modificació ja que la seva codificació inicial és adequada.

La següent variable és Trip_Type. Aquest cas és similar a l'anterior, caldrà recodificar els valors missing (els tenim com a NULL) com "Other". La correspondència entre els nivells antics i els nous és:

- *ANTICS NIVELLS*: Business trip, Couple, Family with older children, Family with young children, Leisure trip, NULL, Solo traveler.
- *NOUS NIVELLS*: Business trip{Business trip}, Couple{Couple}, Family {Family with older children, Family with young children}, Leisure trip{Leisure trip}, Other{NULL}, Solo traveler{Solo traveler}.

Per últim, cal recodificar la variable *Reviewer_Nationality*. Observem com, en aquest cas tenim un gran nombre de nivells (moltes nacionalitats diferents). Com a solució, hem proposat escollir les 20

nacionalitats més freqüents i agrupar la resta en la categoria “Altres” (incloent en aquest grup els valors missings codificats com “”). Les correspondències són:

- *ANTICS NIVELLS*: Abkhazia Georgia, Albania, Andorra, Angola, Argentina, Armenia, Aus- tralia, Austria, Azerbaijan, Bahrain, Bangladesh, Barbados, Belgium, Bermuda, Bosnia and Herzegovina, Botswana, Brazil, Brunei, Bulgaria, Cameroon, Canada, Chile, China, Colom- bia, Comoros, Costa Rica, Croatia, Curaçao, Cyprus, Czech Republic, Denmark, Domini- can Republic, Egypt, Estonia, Ethiopia, Finland, France, Gabon, Georgia, Germany, Gibral- tar, Greece, Guernsey, Honduras, Hong Kong, Hungary, Iceland, India, Indonesia, Iran, Iraq, Ireland, Isle of Man, Israel, Italy, Ivory Coast, Jamaica, Japan, Jersey, Jordan, Kaza- khstan, Kenya, Kosovo, Kuwait, Latvia, Lebanon, Liechtenstein, Lithuania, Luxembourg, Macedonia, Malaysia, Malta, Mauritius, Mexico, Moldova, Monaco, Montenegro, Morocco, Namibia, Nepal, Netherlands, New Caledonia, New Zealand, Nigeria, Norway, Oman, Pak- istan, Panama, Peru, Philippines, Poland, Portugal, Puerto Rico, Qatar, Romania, Russia, Saint Barts, Saint Lucia, Saudi Arabia, Serbia, Singapore, Slovakia, Slovenia, South Africa, South Korea, Spain, Sri Lanka, Sweden, Switzerland, Taiwan, Thailand, Tunisia, Turkey, Uganda, Ukraine, United Arab Emirates, United Kingdom, United States Minor Outlying Islands, United States of America, Venezuela, Vietnam, Zambia.
- *NOUS NIVELLS*: Other{Abkhazia Georgia, Albania, Andorra, Angola, Argentina, Armenia, Azerbaijan, Bahrain, Bangladesh, Barbados, Bermuda, Bosnia and Herzegovina, Botswana, Brazil, Brunei, Bulgaria, Cameroon, Chile, China, Colombia, Comoros, Costa Rica, Croatia, Curaçao, Cyprus, Czech Republic, Denmark, Dominican Republic, Egypt, Estonia, Ethiopia, Finland, Gabon, Georgia, Gibraltar, Guernsey, Honduras, Hong Kong, Hungary, Iceland, India, Indonesia, Iran, Iraq, Isle of Man, Ivory Coast, Jamaica, Japan, Jersey, Jordan, Kaza- khstan, Kenya, Kosovo, Latvia, Lebanon, Liechtenstein, Lithuania, Luxembourg, Macedo- nia, Malaysia, Malta, Mauritius, Mexico, Moldova, Monaco, Montenegro, Morocco, Namibia, Nepal, New Caledonia, Nigeria, Norway, Oman, Pakistan, Panama, Peru, Philippines, Poland, Portugal, Puerto Rico, Qatar, Romania, Russia, Saint Barts, Saint Lucia, Serbia, Singapore, Slovakia, Slovenia, South Africa, South Korea, Sri Lanka, Taiwan, Thailand, Tunisia, Uganda, Ukraine,United States Minor Outlying Islands, Venezuela, Vietnam, Zambia}, Australia{Australia}, Belgium{Belgium}, Canada{Canada}, France{France}, Germany{Germany}, Greece{Greece}, Ireland{Ireland}, Israel{Israel}, Italy{Italy}, Kuwait{Kuwait}, Netherlands{Netherlands}, New Zealand{New Zealand}, Saudi Arabia{Saudi Arabia}, Spain{Spain}, Sweeden{Sweeden}, Switzer- land{Switzerland}, Turkey{Turkey}, United Arab Emirates{United Arab Emirates}, United Kingdom{United Kingdom}, United States of America{United States of America}.

Un cop realitzat cop definides les modalitats correctament, la següent passa és la imputació dels valors missing de la base de dades, ja que d’ara endavant treballarem sense dades mancants. Com ja hem mencionat anteriorment, aquest procés ja s’ha iniciat en el pas anterior agrupant els valors missing de les variables categòriques en un nou nivell del factor que anomenem “Altres” o “Other”.

Ara bé, és necessàri realitzar la imputació dels NA de les variables numèriques. En aquest sentit, plantegem, en primer lloc, la pregunta de si la presència de valors missing a la nostra base de dades és, o no, aleatòria. En el cas que tinguem random missing, podem pensar que es deu a un fenòmen aleatòri casual i que tots ells segueixen la mateixa distribució amb valor esperat 0, de manera que

coneixent informació adicional no hauríem de tenir problemes per a realitzar la imputació. Tanmateix, la qüestió es torna més complexe en cas que tinguem missings no aleatoris ja respondrien a algun tipus de sistemàtica.

Per a verificar la naturalesa dels nostres NA (recordem que estem parlant de variables numèriques) fem servir el *Little's MCAR* test on:

- H0: Missings are completely random (MCAR)
- H1: Missings are not random

this could take a while

[1] 0.0008004798

El valor del test ens porta a rebutjar la hipòtesi nul.la de valors missing aleatòris. Si ens parem a pensar, aquest resultat sembla raonable, ja que les úniques 23 observacions per a les que tenim dades mancants, en relació a les variables numèriques *Hotel_lat*, *Hotel_lng*, *Businesses_100m*, *Businesses_1km*, *Businesses_5km*, són conseqüència d'una absència de coneixement de la situació geogràfica de l'hotel. Així mateix, és llògic pensar que això es deu en un error en la mesura, o la manca de dades geogràfiques d'una zona particular (totes deuen ser en un espai força proper).

Com que el nombre d'observacions afectades per aquests valors missing és molt reduït (23) podem fer la imputació igualment, controlant que els valors que obtinguem es mantinguin dins dels rangs establert i no apareguin anomalies a les dades (Figure 19).

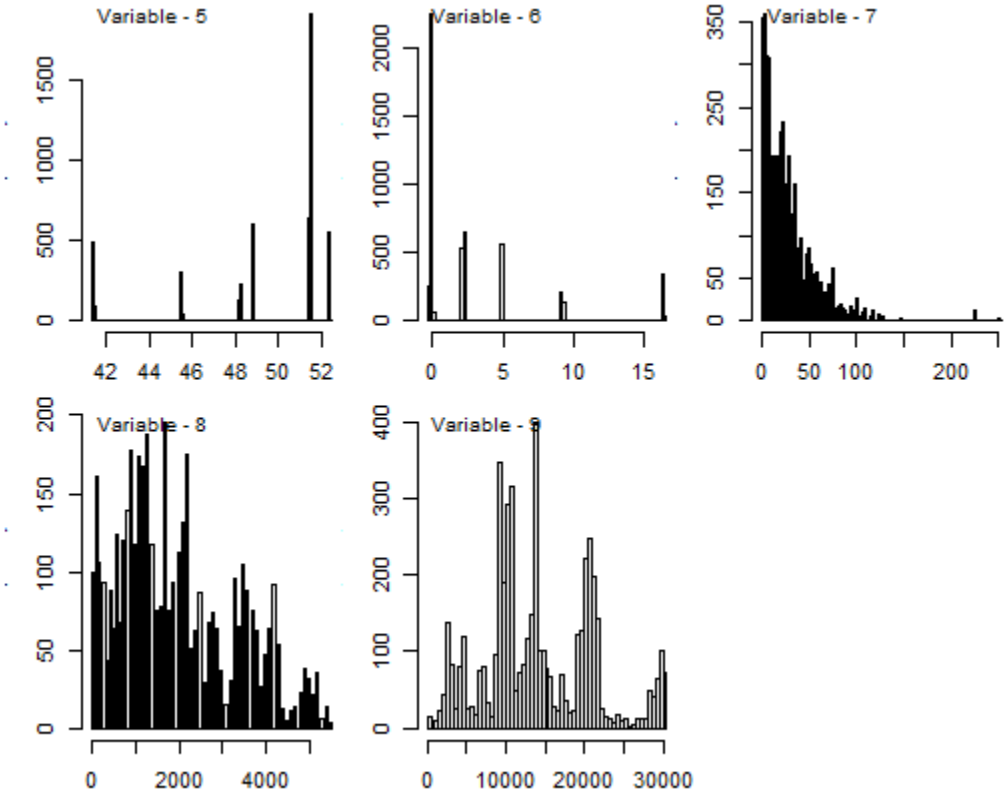


Figure 19: Histogrames sense NA

Apel.lant al diccionari de dades i a l'anàlisi univariant previ, sembla que la imputació s'ha realitzat correctament.

Finalment, respecte a la imputació dels valors missing, només quedarà tractar els valors missing a les variables textuals (a la variable de dates no en tenim). En aquest sentit, podem plantejar un procés similar al de les variables qualitatives, creant un "text" que digui "Ressenya no vàlida" per a aquelles observacions amb NAs.

Finalment, arribats a aquest punt, guardem la base de dades un cop processada de manera adient (en endavant farem servir aquestes noves dades processades en tots els anàlisis).

v. Anàlisi descriptiva univariant de les dades preprocessades.

Un cop completada la fase de preprocessament, és interessant repetir l'anàlisi univariant, per a les variables que anteriorment presentaven algun problema de codificació, o aquelles que han patit alguna modificació (redefinició de categories, imputació de valors missing etc. . .)

En aquest sentit, repetim el procediment d'anàlisi gràfic i numèric univariant que hem plantejat en l'apartat d'anàlisi exploratòri inicial, concretant quines variables han patit alguna modificació i quines romanen igual.

Les dues primeres variables id i *Hotel_Name* segueixent sent identificadors per a cada observació de la base de dades.

La variable *Hotel_Country* no pateix cap tipus de modificació.

Per a la variable *Hotel_City*, hem redefinit les modalitats tal i com especifiquem en la fase de preprocessament. En aquest sentit, reconstruïm el gràfic amb les noves modalitats (Figure 20).

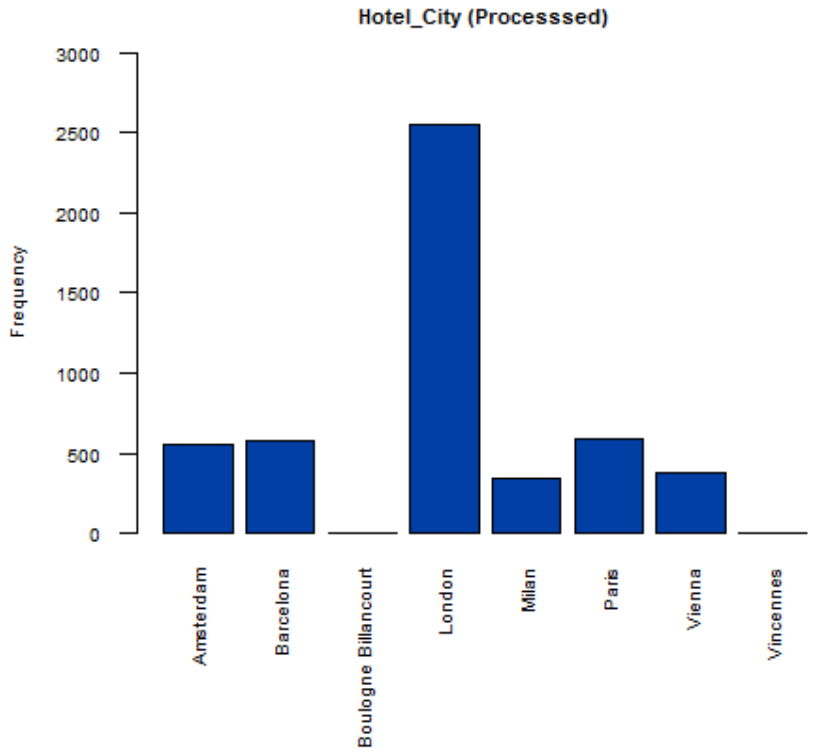


Figure 20: Bar plot Hotel_City (Processada)

Observem com les proporcions es mantenen, bàsicament el que hem fet és agrupar les categories més marginals amb la ciutat gran més propera.

A les variables *Hotel_lat, Hotel_lng, Businesses_100m, Businesses_1km, Businesses_5km* hem realitzat la imputació dels valors missing (a part de la recodificació a numèriques de les variables "Businesses"). En aquest cas, el que volem verificar és que els nous valors que hem introduït no es troben fora del rang o la tendència general de les dades. Per a les dues primeres variables *Hotel_lat*, i *Hotel_lng* fem el resum numèric i ens fixem especialment en els extrems per a detectar si algun valor ha caigut fora dels límits anteriors a la imputació o es corresponen amb localitzacions molt allunyades de les ciutats amb les que estem tractant.

Min.	*1st Qu.*	*Median*	*Mean*	*3rd Qu.*	*Max.*
41.33	*48.21*	*51.50*	*49.46*	*51.52*	*52.40*

Min.	*1st Qu.*	*Median*	*Mean*	*3rd Qu.*	*Max.*
-0.36976	*-0.14366*	*0.01989*	*2.83413*	*4.83110*	*16.42197*

En aquest sentit, observem com ens movem en el mateix interval de valors i, al no tenir cap ciutat en una localització molt diferent a la resta, podem concloure que la imputació ha estat exitosa al nivell de precisió geogràfica al que treballem. Tot seguit, construïm histogrames per a les variables *Businesses_100m*, *Businesses_1km, Businesses_5km*, ja que ara les tenim codificades com a variables numèriques (Figure 21).

Observem com, a mesura que augmentem el radi d'inclusió, les dades tendeixen a una distribució més semblant a la normal. Respecte a la imputació que hem realitzat, no tenim prou evidència per a considerar que els 23 valors imputats puguin afectar a les conclusions globals de l'anàlisi.

En relació a la variable *Room_Type_Level* hem redefinit les modalitats tal i com especifiquem en la fase de preprocessament. En aquest sentit, reconstruïm el gràfic amb les noves modalitats (Figure 22).

La variable *Guest_Type* no pateix cap tipus de modificació.

La variable *Trip_Type* ha patit una redefinció de les modalitats tal i com hem especificat en l'apartat anterior. Repetim el barplot considerant les dades preprocessades (Figure 23).

La variable *Stay_Duration* no pateix cap tipus de modificació.

La següent variable és *Review_Date*. Aquesta variable la hem recodifcat com a data i podem considerar determinar l'interval de temps corresponent a les nostres dades.

[1] "2015-08-04" "2017-08-03"

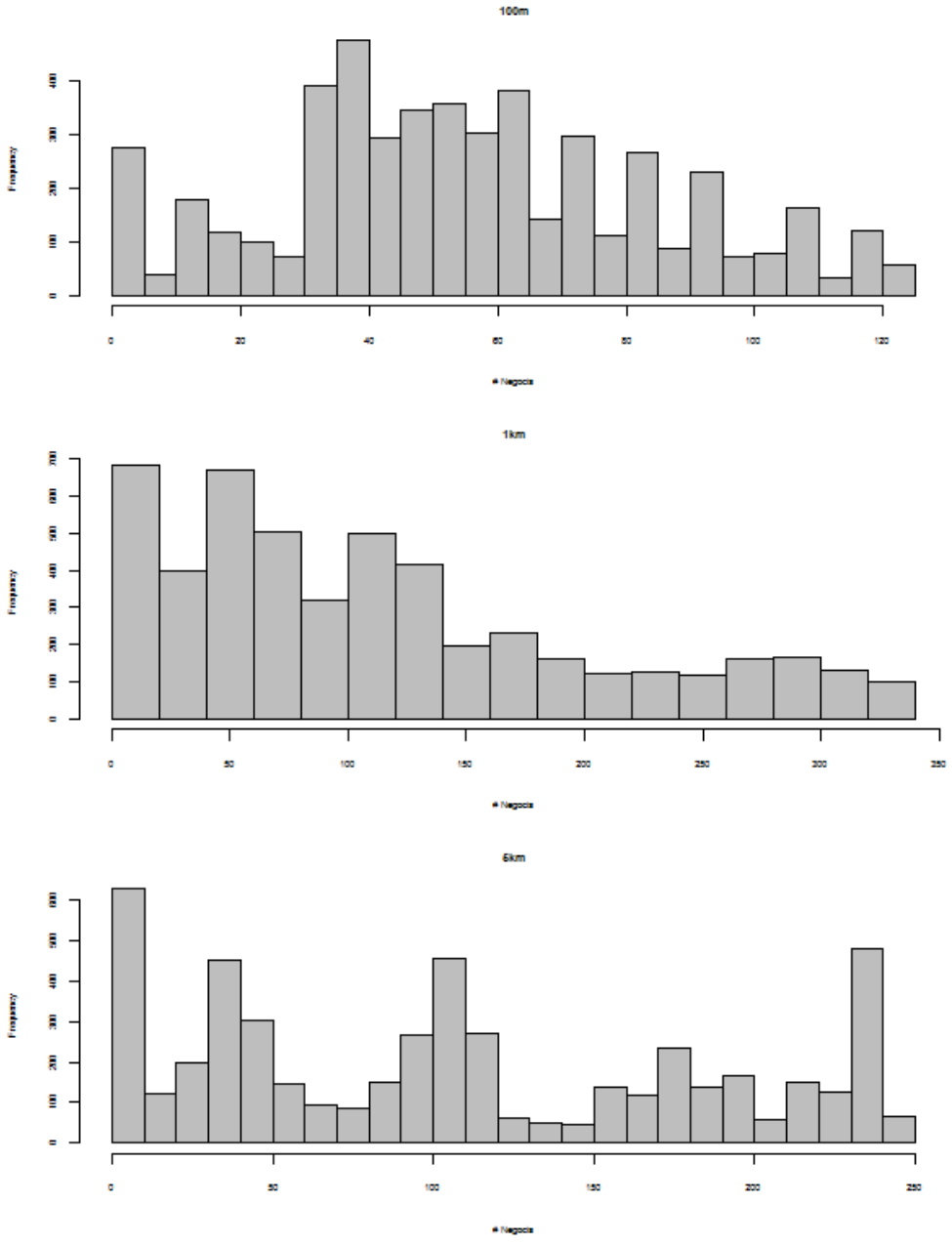
Observem com aquest interval comprèn 2 anys.

Les variables *Is_Hotel_Holiday, Is_Reviewer_Holiday, Review_Is_Positive* i Submitted_from_Mobile les hem recodificat com a factors però els gràfics construïts anteriorment són perfectament extrapolables, ja que simplement canviem 1 per Sí i 0 per No.

Les variables Total_Number_of_Reviews i Review_Positivity_Rate romanen igual.

La següent variable és *Reviewer_Nationality* que, recordem, tenia 123 modalitats. Treballar amb aquest nombre tan alt de nivells pot resultar difícil per a etapes posteriors de l'anàlisi així que, tal i com hem descrit a la fase de preprocessament, redefinim les modalits. En conseqüència, la variable queda finalment definida del següent mode (Figure 24).

La resta de variables *Negative_Review, Review_Total_Negative_Word_Counts, Positive_Review, Review_Total_Positive_Word_Counts, Average_Score, Reviewer_Score, Total_Number_of_Reviews_Reviewer_Has_Given, Additional_Number_of_Scoring* no han patit cap tipus de modificació.



<u>Figura 21:</u> Histogrames Negocis a la rodona (Processada)

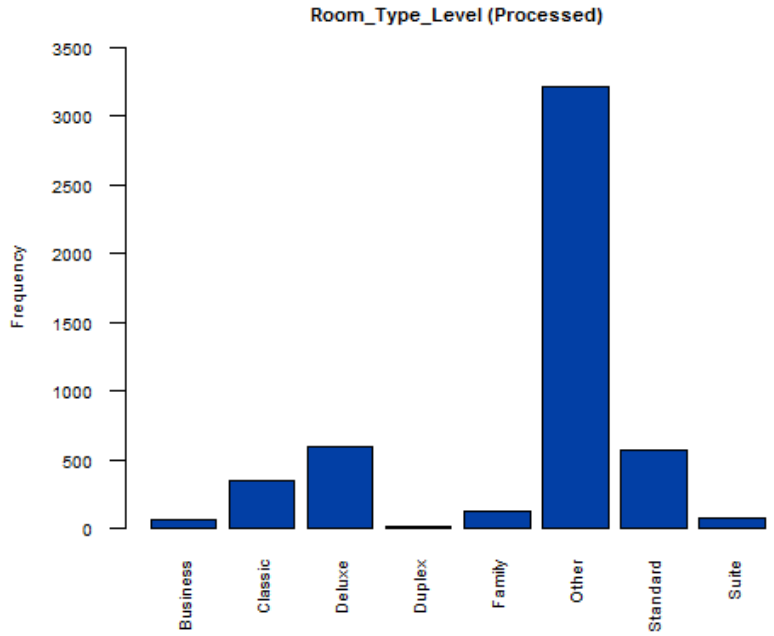


Figura 22: Barplot Room_Type_Level (Processada)

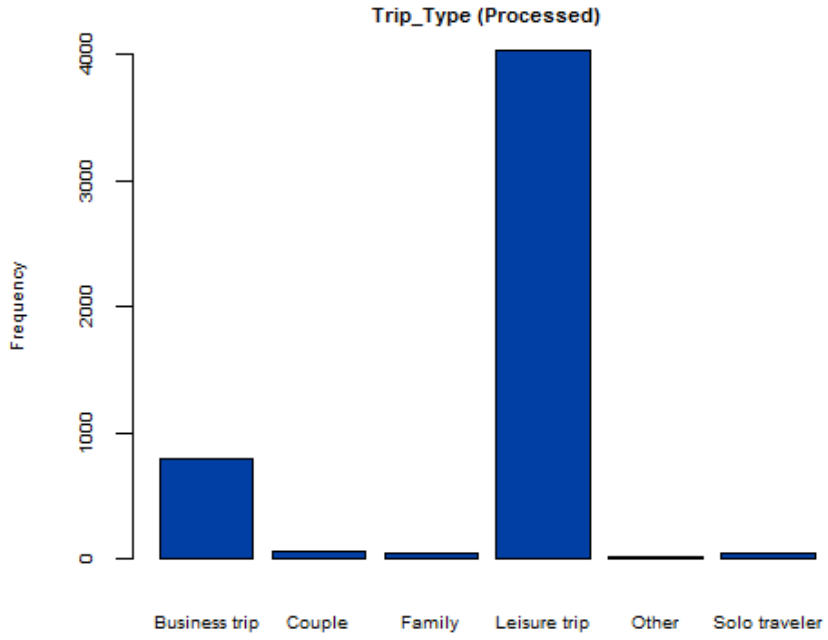


Figura 23: Barplot perfils viatge (Processada)

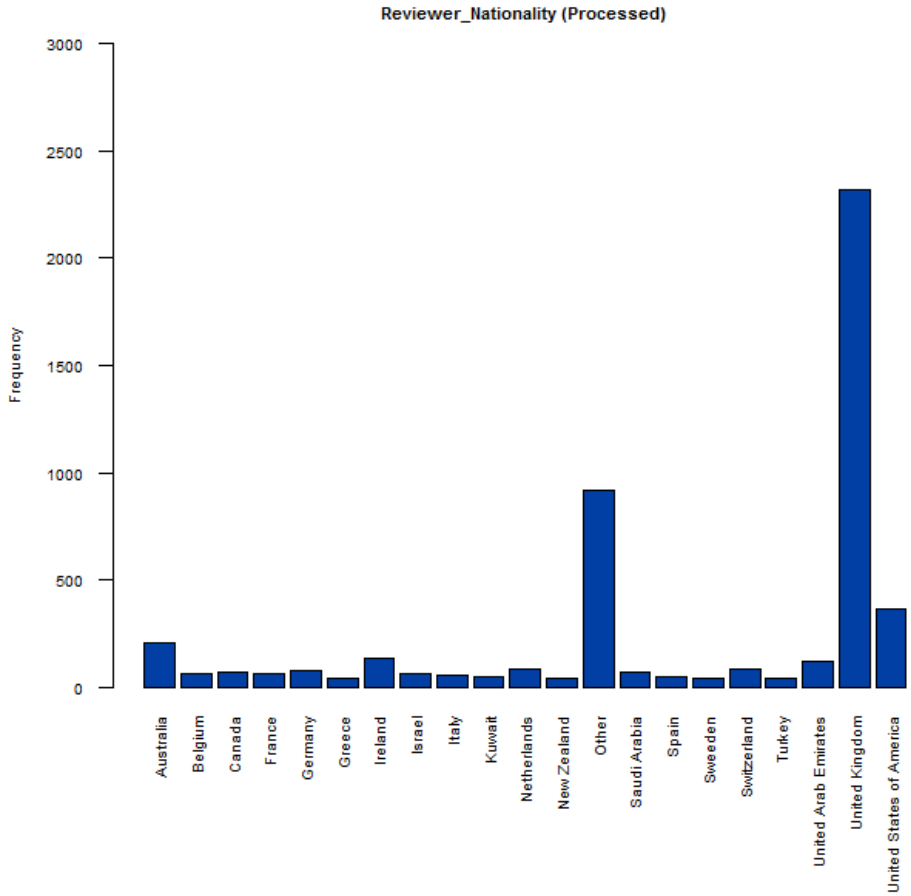


Figura 24: Bar plot Reviewer_Nationality (Processada)

4. Disseny dels processos de mineria de dades.

A partir d'aquest punt només fem servir 25 variables. Hem prescindit de les variables textuals (*Negative_Review i Positive_Review*) ja que en cap dels mètodes de mineria de dades utilitzats les fem servir. També hem prescindit de les variables *Review_Total_Negative_Word_Counts Review_Total_Postive_Word_Counts* que estan relacionades amb les anteriors.

4.1. Disseny dels processos de mineria de dades de la divisió 1.

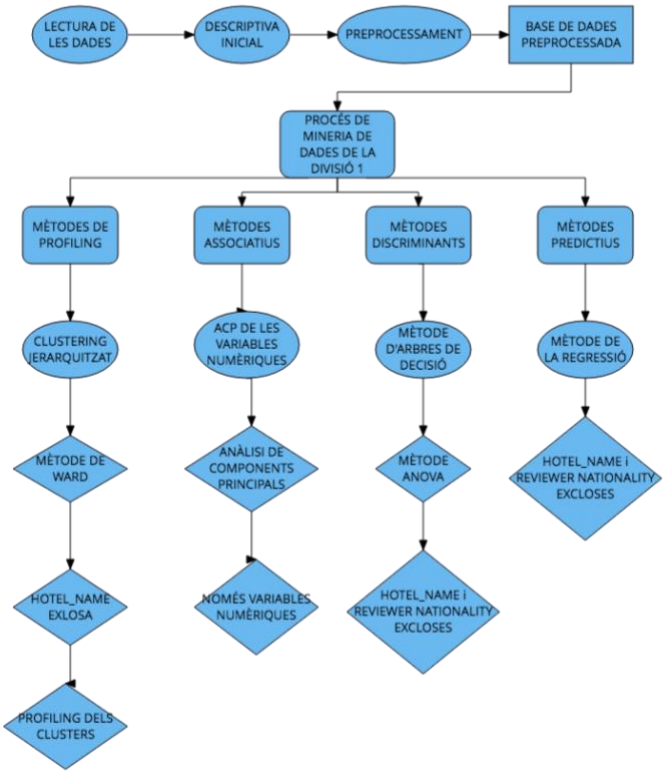


Figura 25: Diagrama de flux de la subdivisió 1

En primer lloc volem destacar que fer realitzar el clústering jerarquitzat s'ha fet servir el mètode de Ward on s'utilitza la distància de Gower. A part, s'ha exclòs la variable HOTEL_NAME perquè per cada observació el valor d'aquesta variable canviarà ja que aquest s'utilitza com un identificador.

Seguidament, a l'ACP només s'han utilitzat les variables numèriques ja que en aquest mètode són aquestes les que es fan servir.

Per altra banda, per construir l'arbre de decisió s'ha fet servir el mètode "anova" ja que tenim com a variable resposta una variable numèrica. En aquest mètode, també s'ha exclòs la variable HOTEL_NAME i REVIEWER_NATIONALITY perquè són factors amb molts nivells.

Finalment, amb els mètodes predictius, s'ha dut a terme el mètode de la regressió en el qual , s'ha creat un model lineal per predir la variable resposta a partir de les explicatives, exepcte HOTEL_NAME i REVIEWER_NATIONALITY, per el mateix motiu que en el mètode discriminant.

4.2. Disseny dels processos de mineria de dades de la divisió 2.

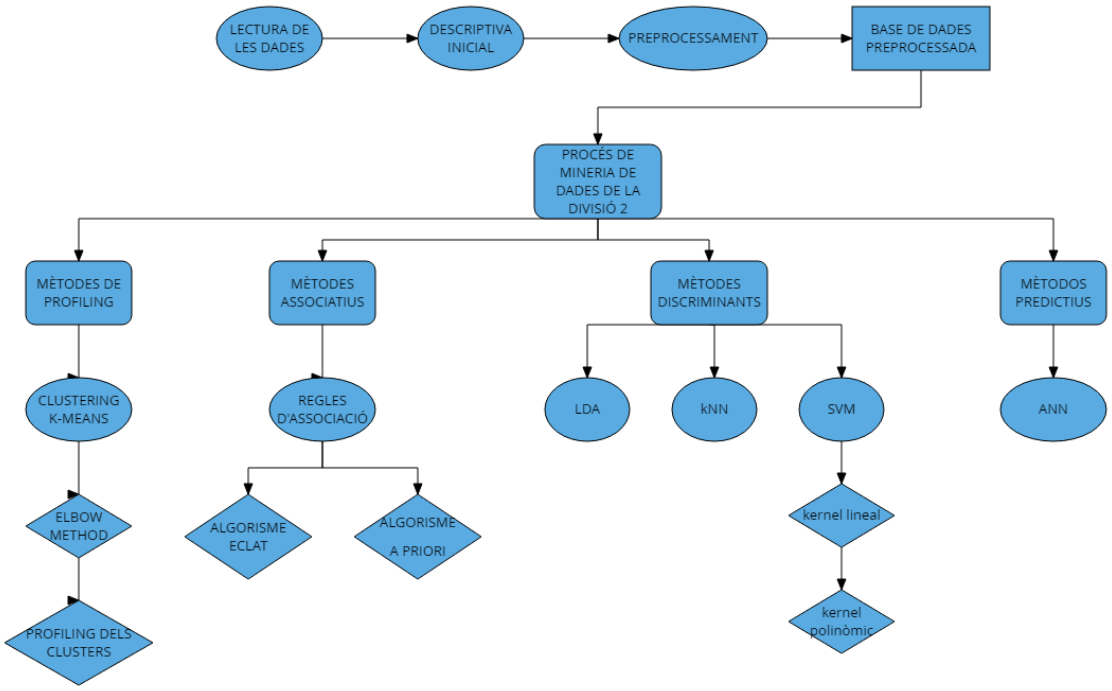


Figura 26: Diagrama de flux de la subdivisió 2

S'han fet servir processos de mineria de dades on no es pren cap variable com a resposta (mètodes de profiling i associatius) i mètodes on es pren la variable *review_positivity_rate* com a resposta (mètodes discriminants i predictius).

En primer lloc, el mètode de profiling utilitzat ha estat el Clustering K-means que és un clustering amb criteri només estadístic. L'Elbow Method ha estat el criteri seguit per escollir el nombre de conglomerats

En segon lloc, s'ha aplicat un mètode associatiu en relació a l'inteligència artificial: les regles d'associació. Per una banda, s'han utilitzat representacions gràfiques i l'algorisme Eclat per trobar itemsets freqüents i per l'altra, l'algorisme a priori (tenint en compte la confiança i el lift) per trobar regles d'associació.

En tercer lloc, s'han aplicat diversos mètodes discrimininats. Un d'ells ha estat el Linear Discriminant Analisis, purament estadístic. Un altre ha estat el k-Nearest Neighbors, mètode no paramètric d'inteligència artifcial. I, per últim, s'ha aplicat el Suport Vector Machine que és un mètode basat tant en l'estadística com en l'inteligència artificial.

Finalment, s'ha escollit el ANN com ha mètode predictiu a fer servir.

5. Procés de mineria de dades de la divisió 1

i. Mètodes de profiling.

En aquest procés es durà a terme un clústering jeràrquic en el qual s'emprarà el mètode de Ward, que consisteix en fer servir la pèrdua d'informació que es produeix a l'integrar els diferents individus en els clústers. Aquesta pèrdua es pot mesurar a través de la suma total dels quadrats de les desviacions de cada individu respecte la mitjana del clúster, de manera que s'aniran agrupant aquells individus que menys incrementin aquesta magnitud al juntar-se.

A més, es pretén que totes les variables intervinguin en el procés de creació dels conglomerats. En aquest sentit, proposem fer servir la distància de Gower, per a conjunts de dades mixtes. És a dir, farem servir aquesta distància quan tinguem un conjunt de registres/individus sobre els quals haguem observat tant variables quantitatives com qualitatives, com és el cas.

Cal destacar que s'ha considerat excloure del procés de clústering la variable Hotel_Name perquè és un camp en què cada registre té un valor diferent i no tindria sentit pensar en possibles agrupacions en funció d'aquesta variable.

Per començar amb el clústering, calculem la matriu de discrepàncies fent servir la distància de gower i, amb el mètode de Ward realitzem el procés de clústering. Podem representar el resultat amb un dendrograma:

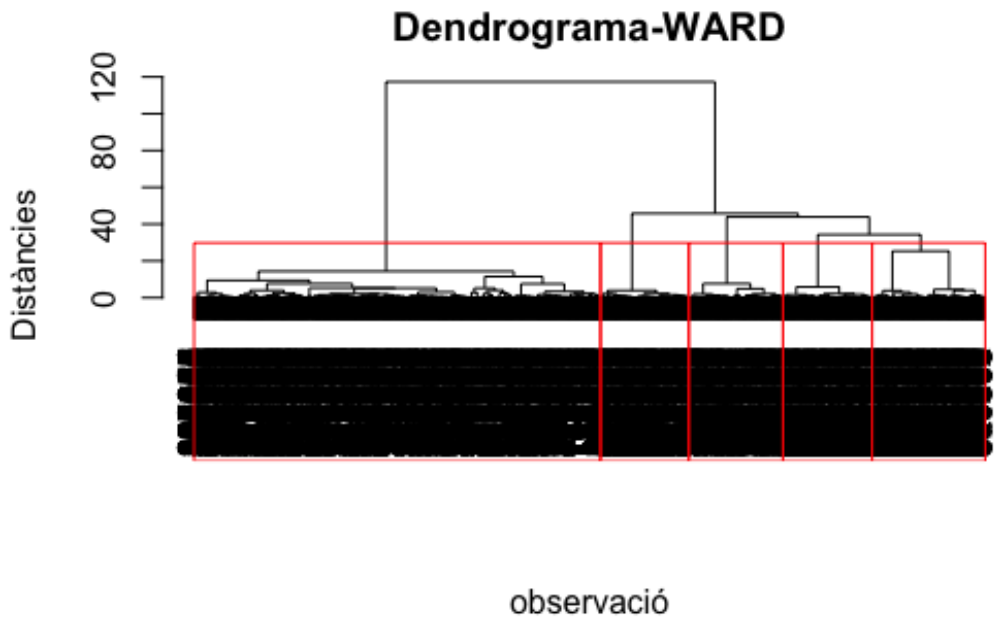


Figura 27: Dendograma del mètode de Ward

En aquest cas, hem considerat oportú realitzar una partició en 5 clústers un cop observat el dendograma. No existeix un procediment formal únic establert per a decidir el nombre de particions, és més comú fer servir tècniques heurístiques com per exemple tallar per aquell salt on guanyem menys inèrcia entre grups de manera que l'esforç d'una partició addicional no compensi la variabilitat que aconseguim explicar amb aquest tall extra.

Per analitzar la qualitat de la nostra partició la comparem amb d'altres. Per exemple, si prenem una hipotètica partició de 4 i una altra de 6 tenim que:

6 PARTICIONS

Cúster	*Nº individos*
1	*343*
2	*563*
3	*558*
4	*2569*
5	*595*
6	*372*

Taula 3: Repartició C. Jeràrquic 6 particions

5 PARTICIONS

Cúster	*Nº individos*
1	*715*
2	*563*
3	*558*
4	*2569*
5	*595*

Taula 4: Repartició C. Jeràrquic 5 particions

4 PARTICIONS

Cúster	*Nº individos*
1	*1278*
2	*558*
3	*2569*
4	*595*

Taula 5: Repartició C. Jeràrquic 4 particions

Podem comparar el percentatge de variabilitat entre grups explicada respecte el total per a les particions immediatament anterior i posterior (k = 4 i k = 6). Fent-ho obtenim:

- 90,82877% amb 4 conglomerats.
- 94.35592% amb 5 conglomerats.
- 95.25504% amb 6 conglomerats.

Observem com el guany obtingut en passar de 5 conglomerats a 6 és mínim, així doncs, ens quedem amb l'opció escollida anteriorment de 5 clústers.

Seguidament, després del clústering es realitzarà el profiling per tal de comprendre quines característiques tenen els individus que formen cada clúster. Per fer això s'utilitzaran Snake plots, diagrames de barres o taules de contingència per a les variables qualitatives així com boxplots i diagrames de barres per a les variables numèriques. A part d'això, per assegurar-nos que hi ha diferències significatives entre els diferents clústers es realitzaran un seguit de contrasts tant paramètrics com no paramètrics en els quals es testaran les diferències entre clústers.

Les primeres variables de la base de dades Hotel_Country, Hotel_City, Hotel_lat, Hotel_lng fan referència a l'àmbit geogràfic i és lògic pensar que la caracterització dels clústers anirà en la mateixa línia en tots quatre casos.

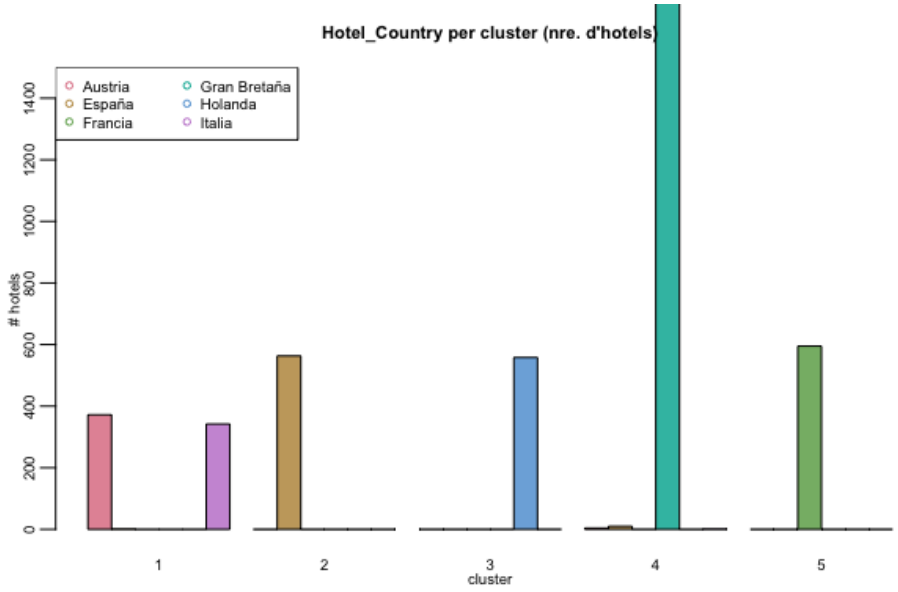


Figura 28: Profiling Hotel_Coutry

Si construim el diagrama de barres de la variable Hotel_Country estratificant per clúster, observem com el clúster 1 és el més divers pel que fa al país on es troba l'hotel en qüestió. En aquest conglomerat hi trobem hotels d'Àustria i Itàlia en proporcions força semblants. Ara bé, la resta de clústers té una forta caracterització pel que fa al país de l'hotel:

- Clúster 2: Predominen hotels d'Espanya.
- Clúster 3: Predominen hotels d'Holanda.
- Clúster 4: Predominen hotels del Regne Unit amb lleugera presència d'hotels d'altre països.
- Clúster 5: Predominen hotels de França.

D'altra banda si en comptes de considerar la variable Hotel_Country fem exactament el mateix per a la ciutat de l'hotel observem com existeix una correspondència en la relació entre països i ciutats:

- Al clúster 1 hi trobem hotels de Vienna i Milà.
- Al clúster 2 hi trobem hotels de Barcelona.
- Al clúster 3 hi trobem hotels d'Amsterdam.
- Al clúster 4 trobem majoritàriament hotels de Londres.
- Al clúster 5 predominen els hotels situats a París.

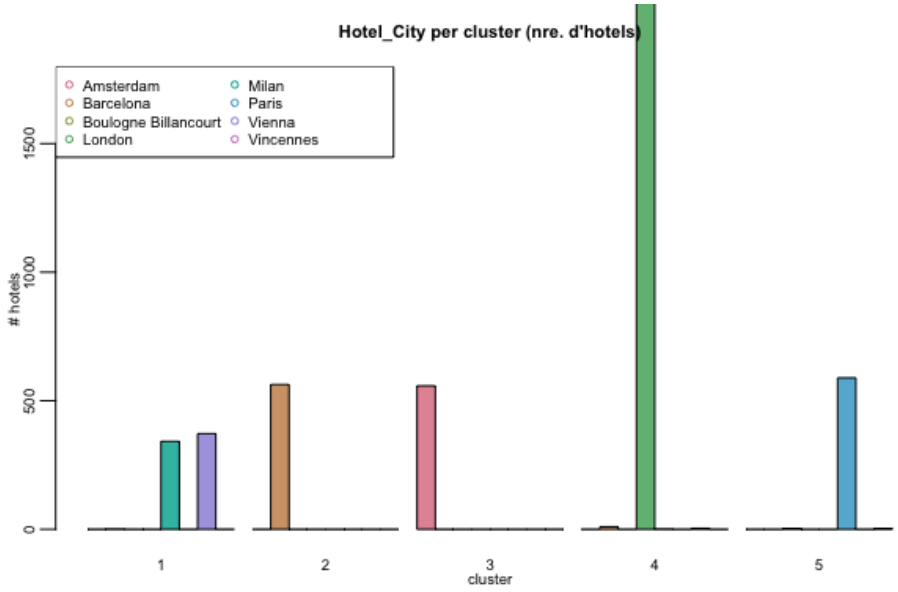


Figura 29: Profiling Hotel_CIty

Per últim, podem realitzar un test Qhi-Quadrat per a validar estadísticament que les diferències entre aquestes variables són significatives entre clústers:

```
[1] "Test Chi quadrat Hotel_Country:"

        Pearson's Chi-squared test

data:  dd[, k] and as.factor(P)
X-squared = 19834, df = 20, p-value < 2.2e-16
```

```
[1] "Test Chi quadrat Hotel_City:"

        Pearson's Chi-squared test

data:  dd[, k] and as.factor(P)
X-squared = 19834, df = 28, p-value < 2.2e-16
```

Tots dos contrastos són significatius.

Tenint en compte que els resultats de les variables Hotel_Country i Hotel_City han coincidit (a cada clúster estan les ciutats pertanyents a cada país), si es realitza el profiling per les variables Hotel_lat (latitud) i Hotel_lng (longitud) s'obtenen uns valors per a aquestes que coincideixen amb els resultats anteriors.

Si seguim amb la resta de variables de la base de dades, les següents tres fan referència al nombre de negocis a la rodona considerant diferents distàncies. L'objectiu cercat amb aquestes variables és fer una distinció entre hotels urbans i hotels més allunyats del centre de les ciutats.
Proposem en primer lloc la realització d'una ANOVA i un test de Kruskall-Wallis per a testar si existeixen diferències globals entre clústers. Els p-valors dels contrastos de totes tres variables són significatius. Tanmateix, cal mencionar que la significació va augmentant a mesura que augmentem el llindar de km a la rodona. És a dir, com més àmplia és la zona que considerem, més evidents es fan les diferències entre els grups. Per aquest motiu, a l'hora d'elaborar els perfils, donem més pes als resultats obtinguts per a la variable Businesses_5km, ja que ens permet detectar millor les diferències entre hotels més o menys allunyats del centre de les ciutats. Ara s'adjunta el plot means de la variable Businesses_5km que és la que ha resultat més significativa:

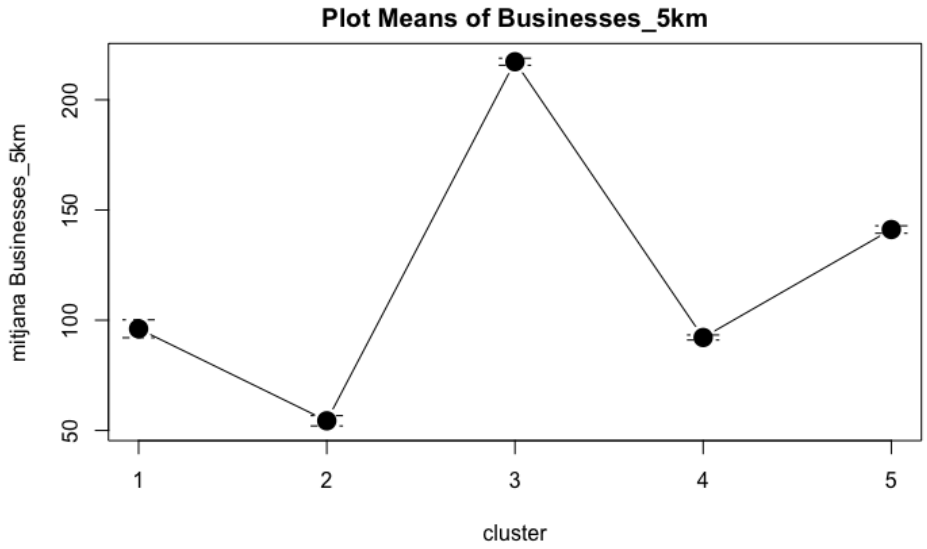


Figura 30: Profiling Businesses_5km

Observem com els hotels que trobem als clústers 3 i 5 tendeixen a ser més urbans, en especial els del clúster 3. En contrapartida, els hotels del clúster 2 semblen ser els menys urbans ja que ocupen posicions baixes, i es troben força allunyats de la mitjana, això pot ser degut a que Barcelona és menys extensa que altres ciutats europees. Aquest fet pot provocar que hi hagi un menor nombre de negocis al voltant de la ciutat.

A continuació ens fixem en el tipus d'habitació. Construïm un snake plot i un diagrama de barres per a comparar les modalitats dins de cada clúster:

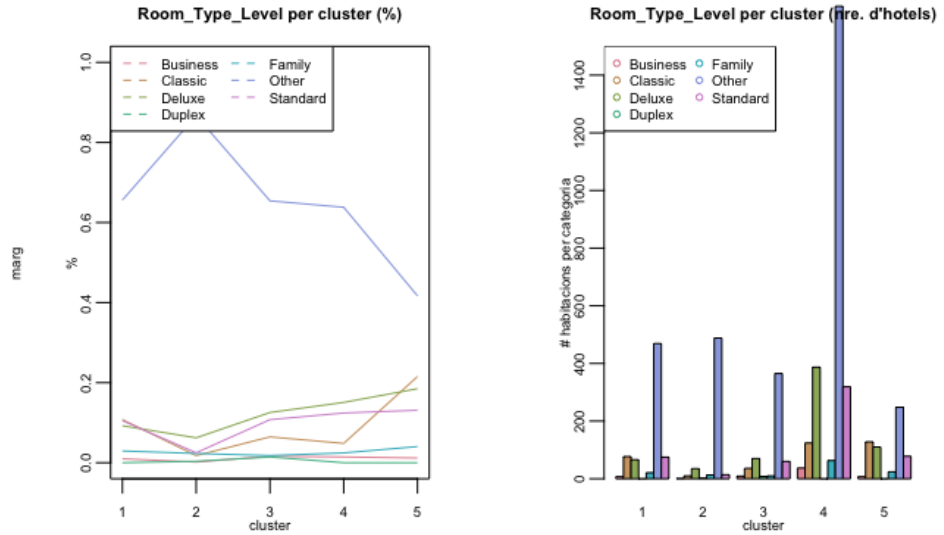


Figura 31: Profiling Room_Type_Level

Aquest gràfic presenta la limitació que tenim molts valors missing que hem inclòs en la categoria "Others". Aquesta categoria és la més freqüent en tots els clústers, en especial al clúster 4 (més del 80%). La resta de modalitats són força homogènies per a tots els clústers amb la distinció que al clúster

4 hi ha major presència d'habitacions del tipus Deluxe i Standard. El test de khi-quadrat és significatiu però hem de tenir en compte que els valors centrats poden estar altament influits per la modalitat "Other".

La següent variable és Guest_Type. Aquesta fa referència al perfil del client que ha escrit la ressenya. Si analitzem com es distribueixen les modalitats d'aquesta variable entre els 5 clústers, observem com al clúster 4 les parelles són més abundants que a la resta de clústers (en tots ells les parelles són els més abundants). El clúster 4 a més també conté més viatgers solitaris que la resta. La resta de categories es manté força constant per a tots els conglomerats.

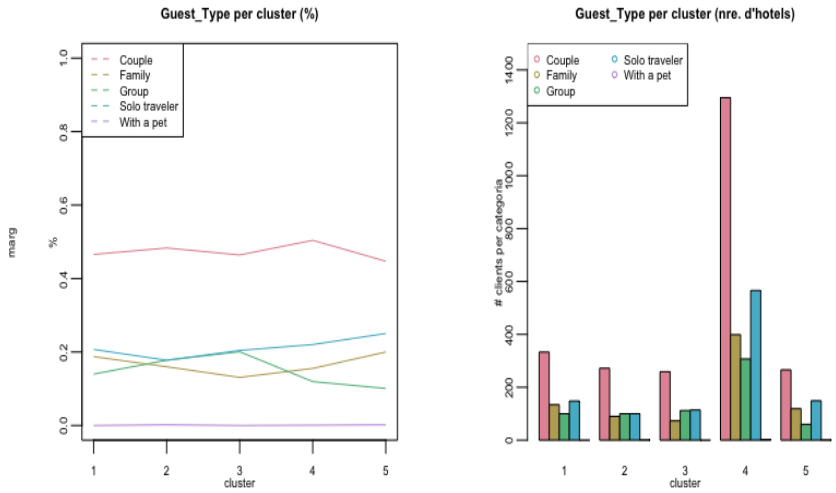
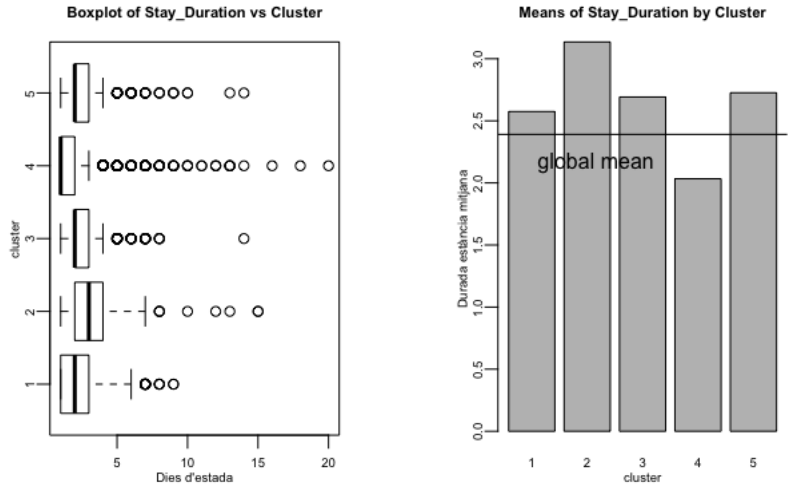


Figura 32: Profiling Guest_Type

Si construïm el test, veiem com existeixen diferències significatives entre els clústers, tot i que la significació global és menys forta que en els casos anteriors.

A continuació, passem a la variable Trip_Type que reflecteix el motiu del viatge. En aquest cas, observem com, en tots els conglomerats predominen els viatges amb motiu d'oci. Tanmateix, trobem que al clúster 4 predominen secundàriament els viatges per motius de negoci.
Podem concloure que la distribució de la variable és més aviat homogènia per a tots els clústers. Si construïm el test, veiem com no existeixen diferències significatives entre els conglomerats ja que el p-valor és superior al nivell de significació del 5%.

La següent variable és Stay_Duration. Al tenir davant una variable numèrica, construïm un boxplot i un gràfic de barres. Observem com els clústers 1, 2, 3 i 5 inclouen llargues estades, mentre que el clúster 4 (recordem, majoria de parelles o viatgers en solitari) tendeixen a incloure estades més curtes.



<u>Figura 33:</u> Profiling Stay_Duration

El test revela diferències significatives entre els conglomerats, tant a nivell global com per a cada clúster.

La següent variable és Days_Since Review. En aquest cas, fem un resum numèric per a cada segment (clústers del 1 al 5). Observem que la mitjana del clúster 5 (380 dies) és considerablement superior a la resta. Els altres conglomerats semblen situar-se força a prop de la mitjana global, tot i que els clústers 1 i 4 semblen tenir valors lleugerament superiors (356 i 355 respectivament) al segon i el tercer (341 i 344 respectivament).

Si realitzem el test no podem parlar de significació global forta. Tanmateix, evidencia com el clúster 5 presenta una forta desviació significativa respecte la tendència general de tots els conglomerats. Per tant, la conclusió a la que arribem és que les ressenyes incloses en el clúster 5 s'han publicat amb més retard respecte la norma general.

A continuació ens fixem conjuntament en les variables Is_Hotel_Holiday i Is_Reviewer_Holiday, dicotòmiques les dues. Per a caracteritzar els clústers, construïm dos diagrames de barres apilades de manera que puguem comparar entre els conglomerats si la ciutat de l'hotel, o la de l'usuari es troba en dia festiu. L'objectiu és veure si en algun clúster els clients tendeixen a escriure les ressenyes en dies festius.

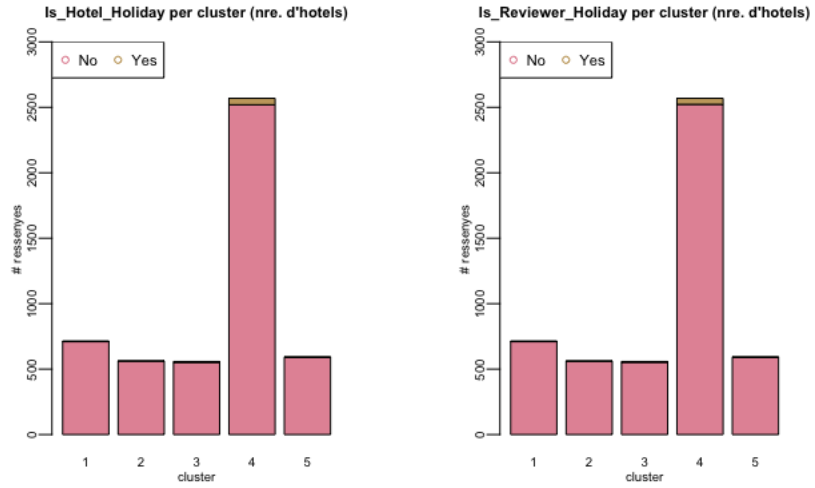


Figura 34: Profiling Is_Hotel_Holiday

Veiem com, al clúster 4 és on la presència de dies festius en el moment d'escriure la ressenya és més alt. A la resta de clústers la proporció de SI és molt baixa per a les dues variables. Si ens fixem en els p-valors dels contrastos, observem com els dos testos surten significatius.

A continuació, ens fixem en la variable Total_Number_of_Reviews. Aquesta variable, recordem, representa el nombre total de ressenyes vàlides que té l'hotel en qüestió. Un estudi interessant seria observar si existeix algun clúster on els hotels tinguin major nombre de ressenyes, i veure com això repercuteix en la valoració de l'hotel.

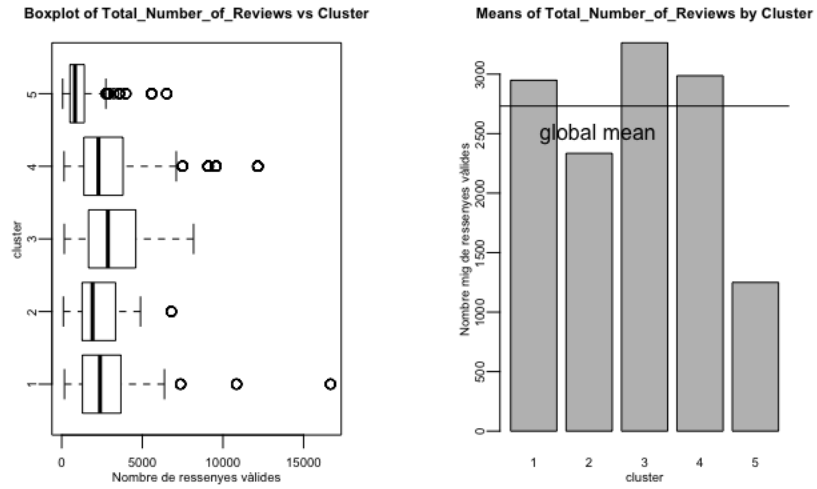


Figura 35: Profiling Total_Number_of_Reviews

Al gràfic veiem com, en nombre de ressenyes vàlides destaquen els clústers 1, 3 i 4. Els clústers 2 i 5 es troben considerablement per sota la mitjana global, en especial el clúster 5 amb un valor mig de ressenyes vàlides al voltant de 1000. Tal i com evidencia la figura anterior, el test revela diferències

significatives entre els conglomerats.

Les següents dues variables, estan relacionades amb el grau de positivitat de la ressenya de Booking. Review_Is_Positive és una variable binària que pren valor 1 (Sí) si el nombre de paraules a la ressenya positiva és major que a la negativa. La proporció de ressenyes positives per clúster, juntament amb un gràfic de barres o un PlotMeans de la variable Review_Positivity_Rate ens pot donar un indici d'en quin dels clústers els comentaris són més positius. Aquest coneixement combinat amb anàlisis posteriors ens revelarà quin tipus d'hotel és el més preferit pels clients, també en funció de les seves característiques.

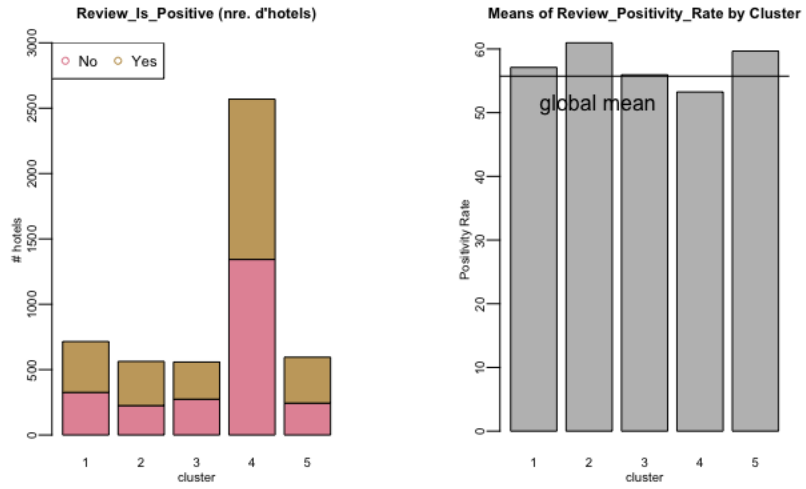


Figura 36: Profiling Reveiw_Is_Positive

Observant els gràfics, podem concloure que:

- Els clústers 1, 2, 3 i 5 tenen un percentatge de ressenyes positives superior al 50% i superior a la mitjana global en tots tres casos.

- El clúster 4 té més o menys la meitat de comentaris positius i l'altra meitat de negatius.

Tal i com sembla indicar la figura, els contrastos evidencien una forta validesa estadística de les conclusions a les que hem arribat.

Tot seguit entrem a analitzar variables relacionades amb les puntuacions dels hotels. Prenem conjuntament la puntuació mitjana que presentava l'hotel a finals de 2016 i la que han anat atorgant els usuaris de Booking que han escrit les ressenyes. Teòricament aquells hotels on les valoracions són més positves haurien de rebre millors valoracions així que, sembla lògic pensar que els resultats haurien d'anar en línia amb els anteriors.

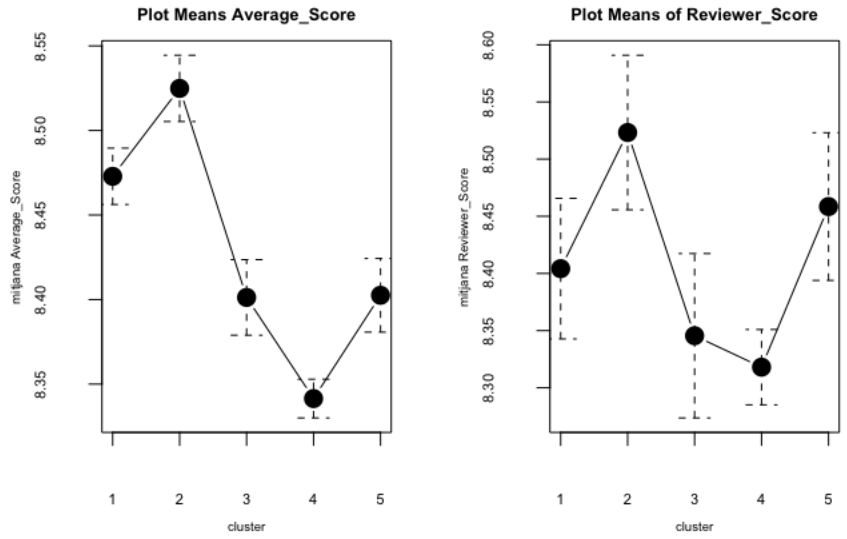


Figura 37: Profiling Average_Score i Reviewer_Score

En efecte, les conclusions que podem obtenir d'aquests gràfics són calcades a les que hem obtingut analitzant Review_Is_Positive, Review_Positivity_Rate: clúster 4 pitjors valoracions, clúster 2 millors valoracions i la resta mantenint-se en la mitjana global. Això pot ser un indici d'alta correlació entre aquestes variables.

Tots dos testos mantenen la significació global amb valors més propers a la no significació per als clústers 1, 3 i 5 que, recordem es troben a prop de la mitjana global.

La següent variable que analitzem, ja a falta de només dues per concloure el profiling, és Total_Number_of_Reviews_Reviewer_Has_Given. Aquesta ens dóna informació sobre com d'actiu a Booking és l'usuari que ha escrit la ressenya. En aquest sentit, observem com els usuaris relatius al clúster 1 són els més actius i la resta es manté a prop de la mitjana global, excepte els del tercer conglomerat, on la mitjana de comentaris totals escrits pels usuaris és més baixa.

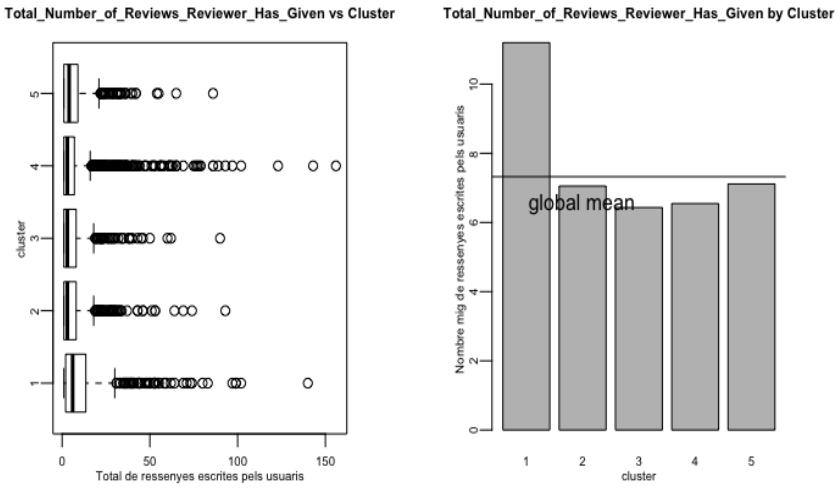


Figura 38: Profiling Total_Number_of_Reviews_Reviewer_Has_Given

Els contrastos reforcen la conclusió obtinguda observant el gràfic, ja que la significació més forta apareix als clústers 1 i 3.

A continuació ens fixem en la variable Additional_Number_of_Scoring relativa al total de valoracions addicionals que rep l'hotel (localització, neteja, servei...). Seria interessant veure si els hotels amb valoracions més positives també reben un major nombre de comentaris addicionals, o aquest efecte és just al contrari.

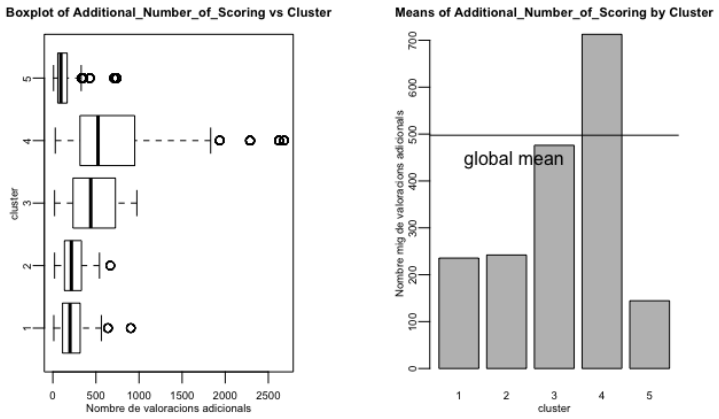


Figura 39: Profiling Additional_Number_of_Scoring

Veiem com el resultat és clarament significatiu i, curiosament, els hotels dels clústers on les puntuacions són més extremes (clúster 4 les més negatives) és on els usuaris s'entretenen més a valorar aspectes extra de l'hotel. En contrapartida, aquells establiments hotelers amb valoracions mitjanes no solen rebre d'addicionals.

Els contrastos, en aquest cas mostren una forta validesa estadística per als resultats obtinguts.

Per últim, considerem la variable Submitted_from_Mobile. Aquesta variable ens mostrarà si en algun clúster són més freqüents les ressenyes i valoracions escrites des del mòbil. Aquesta variable pot ser indicativa de si els usuaris estan satisfets, o no, amb el funcionament de l'aplicació de Booking.

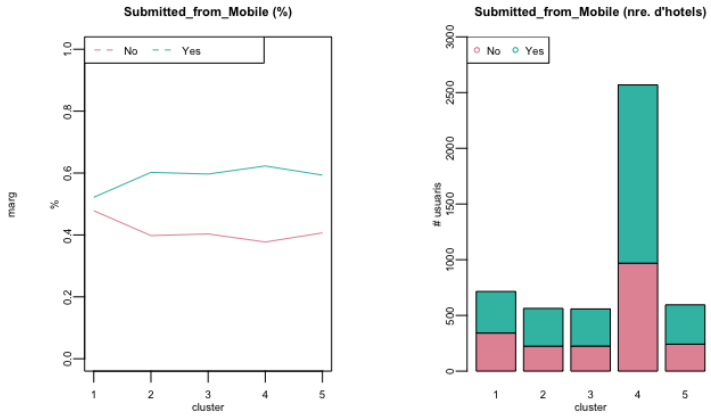


Figura 40: Profiling Submitted_from_Mobile

No existeixen grans diferències entre els clústers pel que fa a la proporció de ressenyes escrites des del telèfon 48nfor. La 48nformació més rellevant apareix en els usuaris compresos al clúster 4, els quals tendeixen a fer servir més el telèfon 48nfor per a escriure les seves ressenyes. Recordem que els hotels dins aquest conglomerat tendeixen a rebre valoracions més negatives i, en conseqüència, és 48nforma inferir que aquestes s'escriuen en major grau des de telèfons mòbil. Pel que fa al contrast, veiem com la significació global no és massa forta, segurament degut a que, excloent el clúster 4, tots els altres es troben força a prop de la mitjana global.

Un cop tenim recopilada tota la informació descriptiva de cada clúster, podem resumir-la en una taula, donant un nom al segment en qüestió i elaborant una petita descripció que inclogui els seus atributs principals.

Clúster	Nom	Descripció
1	Hotels d'Àustria i Itàlia	Hotels un tant allunyats del centre de les ciutats amb valoracions mitjanes (ni massa bones ni massa dolentes), generalment freqüentats per parelles, però també són abundants els grups (suposem que seran els més econòmics). L'estància mitjana està al voltant de la tendència general de 2,5 dies (cap de setmana llarg). També són els hotels que reben més comentaris i on els usuaris són més actius.
2	Millor puntuats	Hotels a Barcelona que tenen menys establiments que els altres al voltant, amb la majoria d'habitacions Deluxe i amb estades dels clients més llargues que la mitja i més comentaris positius.
3	Hotels cèntrics	Hotels amb molts negocis als voltants freqüentats per parelles majoritàriament per motius de vacances. Usuaris poc actius.
4	Mal puntuats	Hotels situats a Londres, relativament cèntrics que han obtingut les valoracions més baixes. Freqüentats majorment per parelles o per viatgers solitaris. Són els que reben més valoracions addicionals (podria ser degut a les insatisfaccions dels clients).
5	Viatges de poca durada	Hotels amb poques valoracions situats a París. Predominen les parelles que viatgen pocs dies (podria ser degut als alts preus d'aquesta ciutat).

Taula 6:Taula resum profiling Clustering Jeràrquic

ii. Mètodes associatius.

En aquest apartat es durà a terme l'ACP de les variables numèriques.

L'objectiu general de l'anàlisi de components principals és identificar patrons a les nostres dades, de manera que puguem analitzar-les reduint la dimensió de la base de dades original amb una pèrdua d'informació mínima.

És a dir, l'output que busquem és la projecció de la nostra base de dades original (nxd) en un subespai més petit, però que mantingui una bona representació de les dades i les descrigui correctament. En aquest sentit, més endavant veurem com aconseguir aquesta "bona" representació de les dades mitjançant els valors propis i vectors propis de la nostra matriu de dades.

En general, podríem dir que els objectius perseguits amb l'ACP són:

- Identificar patrons a les dades
- Reduir la dimensionalitat de la base de dades original eliminant les redundàncies
- Identificar correlacions entre variables
-

L'ACP aconsegueix aquests objectius transformant les variables inicials en nou (i més petit) conjunt de variables sense que perdem la informació més rellevant que ens aporten aquestes. Les noves variables les anomenem components principals, i no són més que combinacions lineals de les variables originals. La metodologia assumeix que les direccions principals amb major variància són les més importants.

Per exemple, el primer component principal és una combinació lineal de les variables originals que captura la variància màxima de la base dades, determinant la direcció de més variabilitat en les nostres dades n-dimensionals (cap component pot capturar més variabilitat que el primer).

En aquest sentit, és important mencionar des del començament que, com que hem de calcular distàncies per a maximitzar aquestes variàncies, considerarem com a actives únicament les variables numèriques ja que si incloem variables qualitatives en aquests càlculs, podem introduir errors considerables ja que R les tractarà com a dummies per a la construcció d'aquestes combinacions lineals. En terminologia PCA, tenim:

- Individus actius: Totes aquelles observacions que intervenen en l'ACP. En el nostre cas totes les observacions (5000).
- Individus suplementaris: Les coordenades d'aquests individus s'estimaran fent servir la informació obtinguda de l'anàlisi de components principals en base als individus actius.
- Variables actives: Totes les variables que utilitzem en l'ACP. En el nostre cas totes les numèriques i numèriques enteres.
- Variables suplementaries: Com en el cas dels individus suplementaris, les coordenades d'aquestes variables s'estimaran amb els resultats de l'anàlisi. En aquest cas, no hem especificat cap variable categòrica suplementaria (les projectarem sobre els nous eixos).

Podem fer servir prcomp per a que R calculi els valors propis de la matriu de dades un cop seleccionades les variables numèriques actives en l'anàlisi. A continuació, els representem en un gràfic per a veure el percentatge de variabilitat total de les dades que captura cada dimensió (component principal).

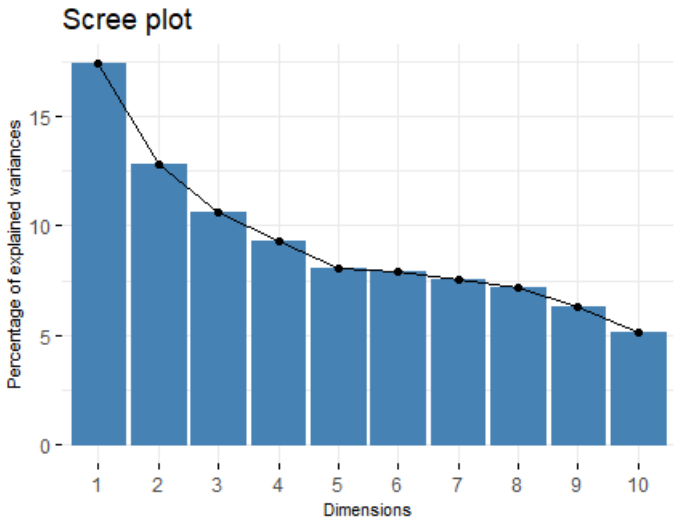


Figura 41: Percentatge de variància capturat a cada dimensió

Repetim el càlcul numèricament i observem com, per a assolir el llindar del 80% de la variància total de les dades, necessitem 10 dimensions. Un altre llindar que normalment es fa servir és seleccionar dimensions fins que trobem un valor propi inferior a 1, ja que voldrà dir que aquell valor propi capturarà menys variabilitat que alguna de les variables originals.

	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	2.26373003	17.4133079	17.41331
Dim.2	1.66236040	12.7873877	30.20070
Dim.3	1.37520376	10.5784905	40.77919
Dim.4	1.20625810	9.2789085	50.05809
Dim.5	1.04763755	8.0587504	58.11684
Dim.6	1.03016390	7.9243377	66.04118
Dim.7	0.97534579	7.5026599	73.54384
Dim.8	0.93476037	7.1904644	80.73431
Dim.9	0.82222193	6.3247841	87.05909
Dim.10	0.66380156	5.1061659	92.16526
Dim.11	0.51427479	3.9559599	96.12122
Dim.12	0.40531620	3.1178170	99.23903
Dim.13	0.09892561	0.7609663	100.00000

Un cop hem seleccionat les dimensions, podem guardar els resultats del PCA. Si partim de l'output res.pca de la comanda prcomp:

- Resultats per a les variables: res.var <- get_pca_var(res.pca):

    1. res.var$coord: Coordenades.
    2. res.var$contrib: Contribucions als components principals.
    3. res.var$cos2: Qualitat de la representació.

- Resultats per als individus : res.ind <- get_pca_ind(res.pca):

    1. res.ind$coord: Coordenades.
    2. res.ind$contrib: Contribucions als components principals.
    3. res.ind$cos2: Qualitat de la representació.

La qualitat de la representació (cos2 a R) fa referència a com de "bona" és la representació de l'individu o la variable al component principal. Valors alts indicaran una alta representativitat, és a dir, més rellevant serà aquell individu/variable per a la interpretació dels resultats (major pes en l'anàlisi). En contrapartida, valors baixos per a cos 2 indicaran baixa representativitat, poca correlació dels valors de l'individu/variable amb la component principal (poca rellevància en l'anàlisi).

Gràfics d'individus

Per a començar l'anàlisi podem representar, en primer lloc, el núvol de punts dels individus sobre el primer pla factorial (components principals 1 i 2). S'ha considerat incloure en el gràfic la representativitat de cada individu, de manera que punts més grans indicaran individus que mostren correlacions més elevades amb la component principal.

El núvol de punts és molt homogeni, veiem com els individus amb menys representativitat prenen valors propers al centre de coordenades. A continuació construïm gràfics addicionals per a la resta de dimensions 3, 4, 5, 6 (per abreujar l'anàlisi, tractem només les 6 primeres dimensions, on es concentra la majoria de la variabilitat).

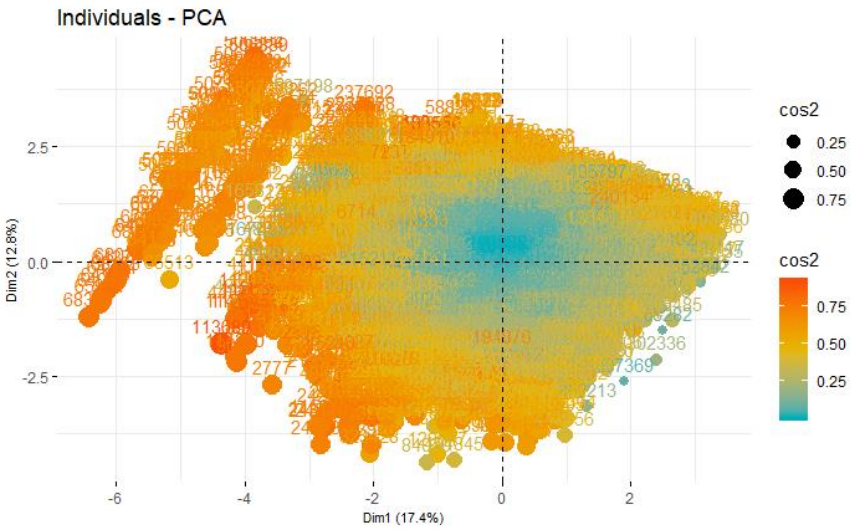


Figura 42: Representació dels individus sobre els PC (1er pla)

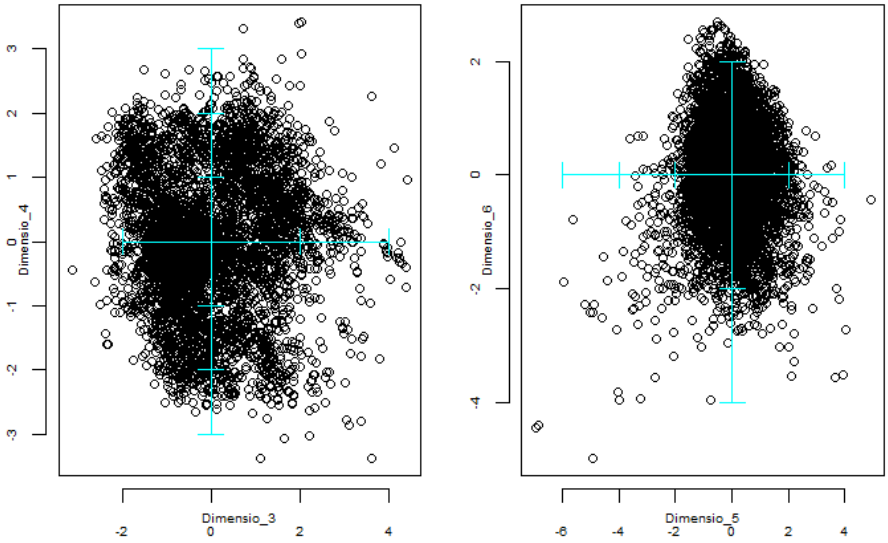


Figura 43: Representació dels individus sobre els PC

Veiem com, a mesura que ens acostem a dimensions d'ordre major, aquesta massa de punts es fa tornant més compacte i ocupa menys a l'espai (menys dispersió entre els punts ja que es captura menys variància). Addicionalment podem testar si aquests nous eixos construïts a partir dels components principals ens separen bé els clústers obtinguts anteriorment.

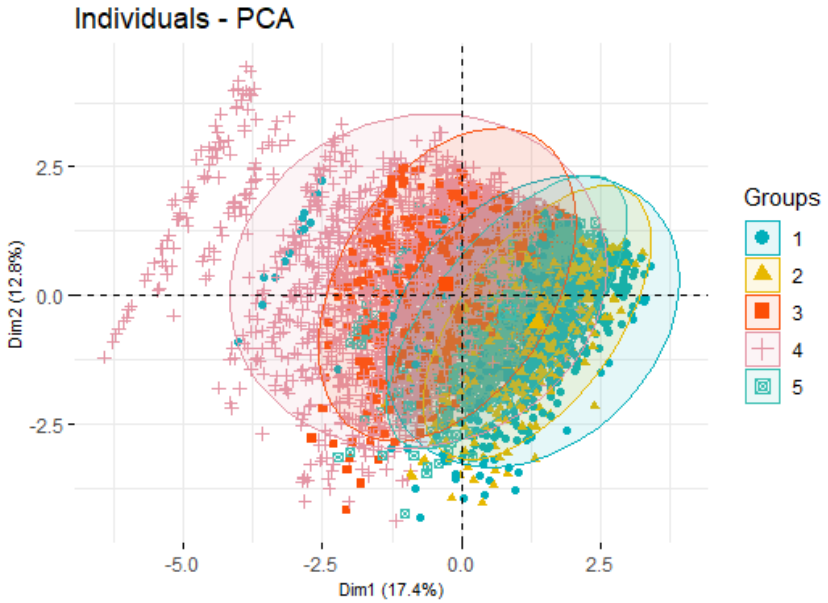


Figura 43: Clustering jeràrquic sobre el primer pla factorial

Veiem com, sobre els eixos de dimensions, el clúster anterior no separa massa bé els grups, en especial l'1 i el 2 (a simple vista, sembla que potser aniria millor amb 3 clústers). Deixant de banda aquest últim apartat, un cop hem representat els individus sobre aquests eixos de components principals, el següent pas serà tractar amb les variables i representar-les en els eixos d'aquest nou

subespai. D'aquesta manera que podrem establir una relació entre els individus i les variables, a partir de les posicions que ocupen.

Gràfics de variables

El primer pas, per a obtenir les projeccions de les variables és obtenir la correlació entre els valors originals i les seves projeccions (fem servir les correlacions per a tenir una mesura estandarditzada). Un cop tinguem aquest resultat, la primera aproximació que proposem és la creació d'un gràfic per a veure la representativitat de cada variable en cada component principal. La interpretació és la mateixa que en el cas dels individus, a major representativitat, més a prop de la circumferència del cercle de correlacions i més rellevància en l'anàlisi.

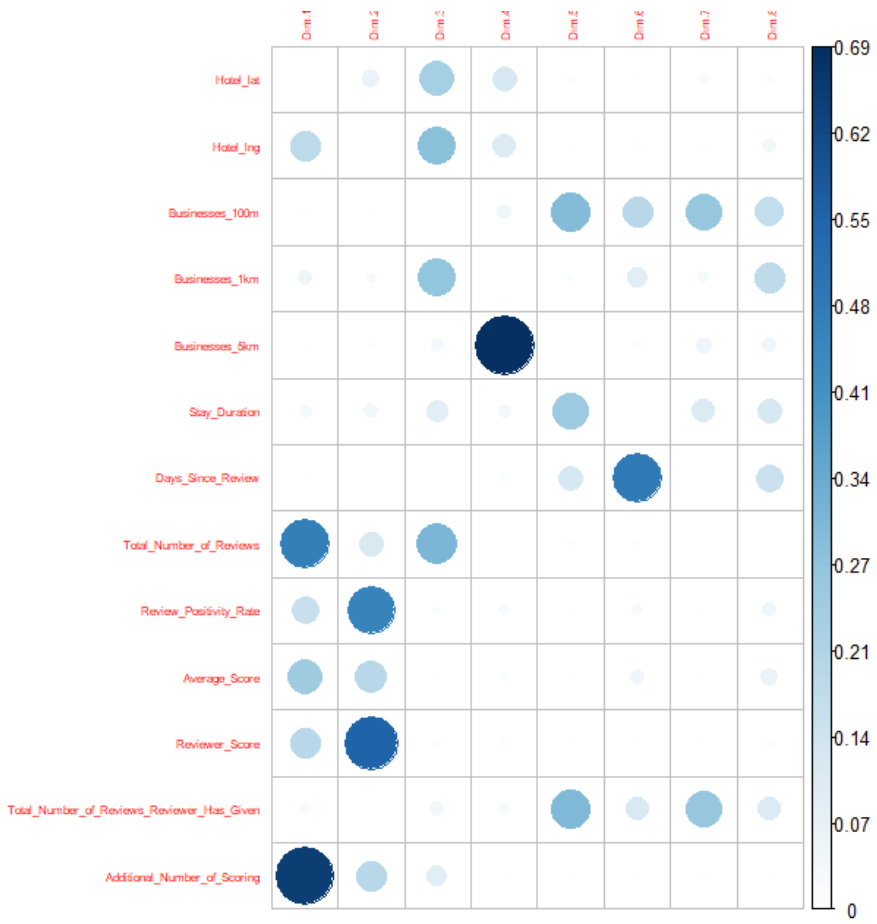
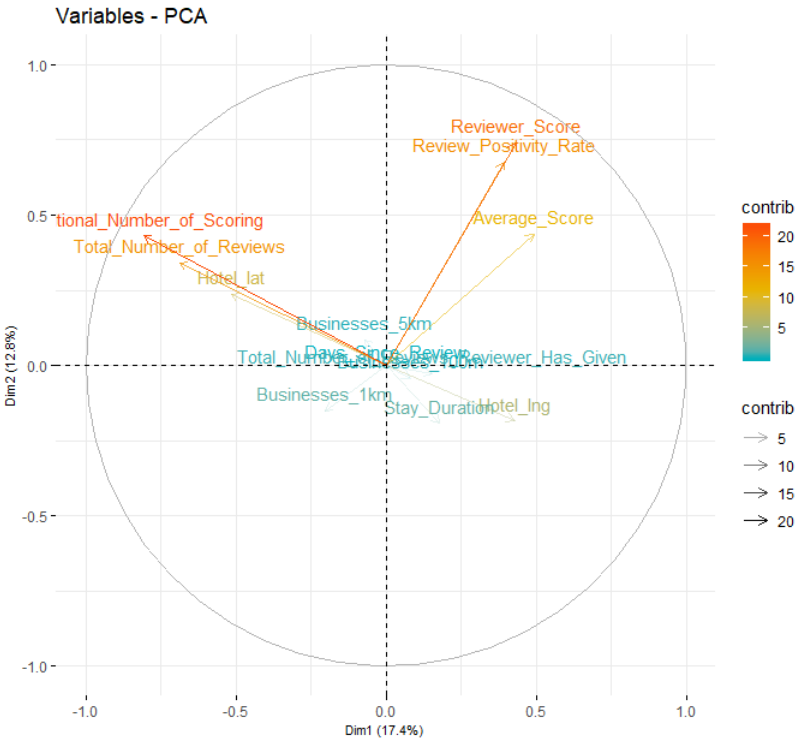


Figura 44: Representativitat de les variables a cada dimensió

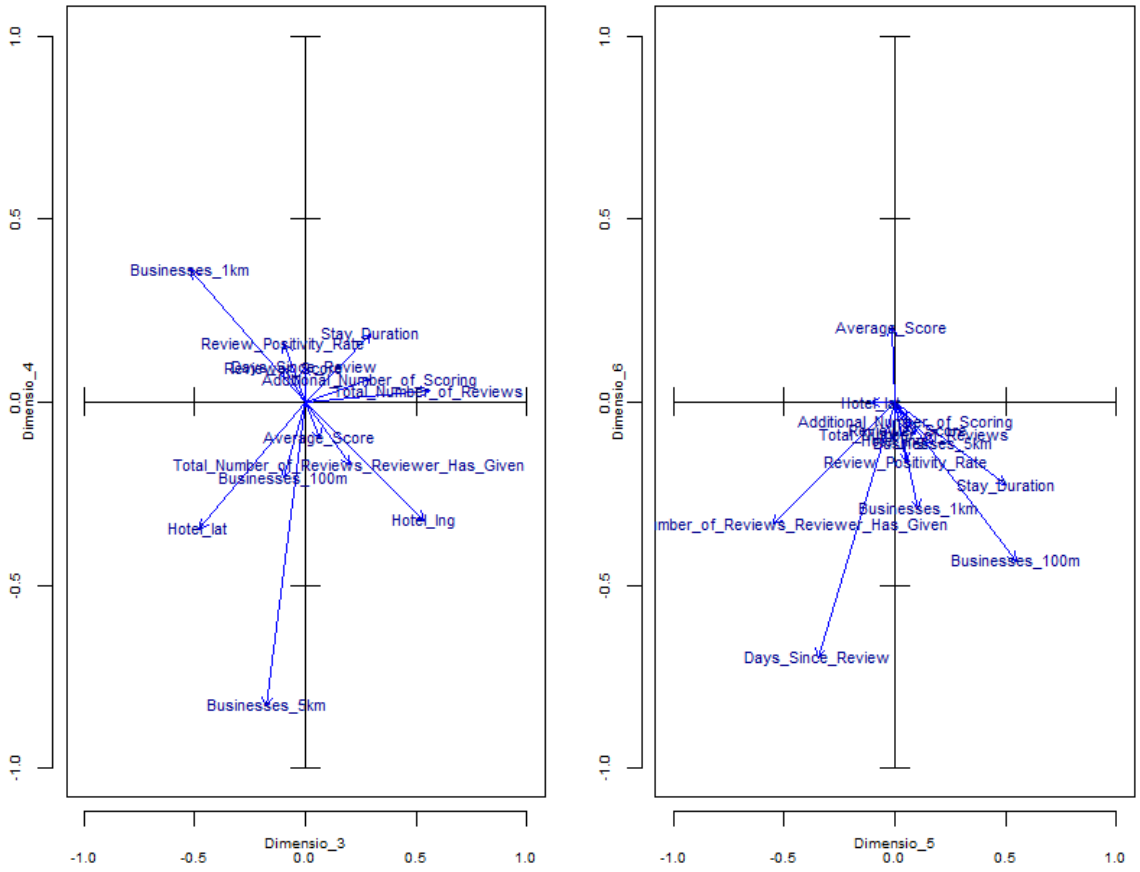
Un cop calculades les projeccions de les variables, caldrà representar-les sobre els eixos de components principals. Podem pensar en un gràfic on representem aquests valors en funció de la seva contribució, de manera que puguem veure ràpidament les que capturen més variància de l'eix. Construïm el gràfic per al primer pla factorial (components principals 1 i 2). Recordem que aquestes variables són les actives en l'anàlisi, ja que hem calculat les seves projeccions a partir de la matriu de correlacions descrita anteriorment.



<u>Figura 45:</u> Representació de les variables sobre els PC(1er pla)

En aquest primer pla veiem com les variables més rellevants són la puntuació, el grau de positivisme del comentari, valoració mitjana de l'hotel, nombre total de ressenyes i nombre total de valoracions addicionals vàlides tot i que no totes en la mateixa direcció. En general, les variables implicades en la valoració de l'hotel (Reviewer_Score, Review_Positivity, Average_Score) tenen major representativitat a l'eix 1, mentre que les variables referides als comentaris addicionals i valoracions addicionals es projecten millor a l'eix 2. En aquest sentit, podem pensar que els individus situats al primer sector del mapa de coordenades són els que han obtingut millors valoracions.

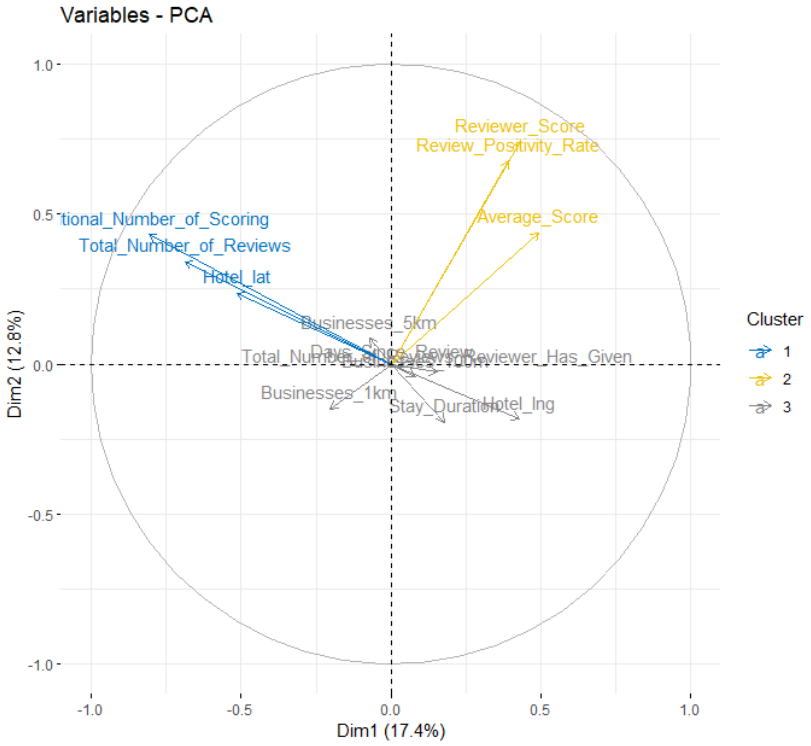
A continuació, i com en el cas anterior, construïm gràfics addicionals per a les dimensions 3, 4, 5 i 6:



<u>Figura 46:</u> Representació de les variables sobre els PC

En general, les variables ben representades a la Dimensió 2, també ho estan a la dimensió 4, la majoria variables relacionades amb la puntuació d'hotel i el nombre de ressenyes. D'aquest primer gràfic també es desprèn una alta representativitat de les latituds, longituds i les variables relacionades amb el nombre de negocis a prop dels hotels a les dimensions 3 i 4. En contrapartida, a les dimensions 5 i 6 tenim, en general, una pitjor representació global destacant únicament, però en gran mesura, les variables Days_Since_Review i Number_of_Reviews_Reviewer_Has_Given.

Addicionalment, el software ens permet crear una espècie de clúster de les variables numèriques fent servir kmeans. D'aquesta manera, creem conglomerats de variables que, a priori, assumim es relacionen del mateix mode amb les components principals. Recordem que l'algoritme kmeans requereix especificar el nombre de categories que desitgem (en aquest cas hem considerat 3).



<u>Figura 47:</u> Representació de les variables sobre els PC(1er pla)

L'agrupació sembla lògica ja que les variables que s'agrupen estan altament correlacionades entre elles. Per exemple Review_Positivity_Rate, Reviewer_Score i Review_Total_Positive_Word_Counts, o les tres variables corresponents als negocis a la rodona. És natural pensar que aquestes variables es desplacin conjuntament en aquest nou subespai.

Un cop tractades les variables numèriques actives a l'anàlisi, no podem oblidar les variables qualitatives que hem deixat de banda inicialment. Podem fer un càlcul de la variància (en relació a les modalitats) de les variables qualitatives de la nostra base de dades i projectar-les als eixos juntament amb les numèriques. Comencem fent una representació senzilla de les categories de la variable Hotel_Country en el primer pla factorial. Més endavant proposarem construir un gràfic amb tots els nivells de totes les variables categòriques, de manera que puguem identificar intuïtivament relacions entre variables als plans factorials.

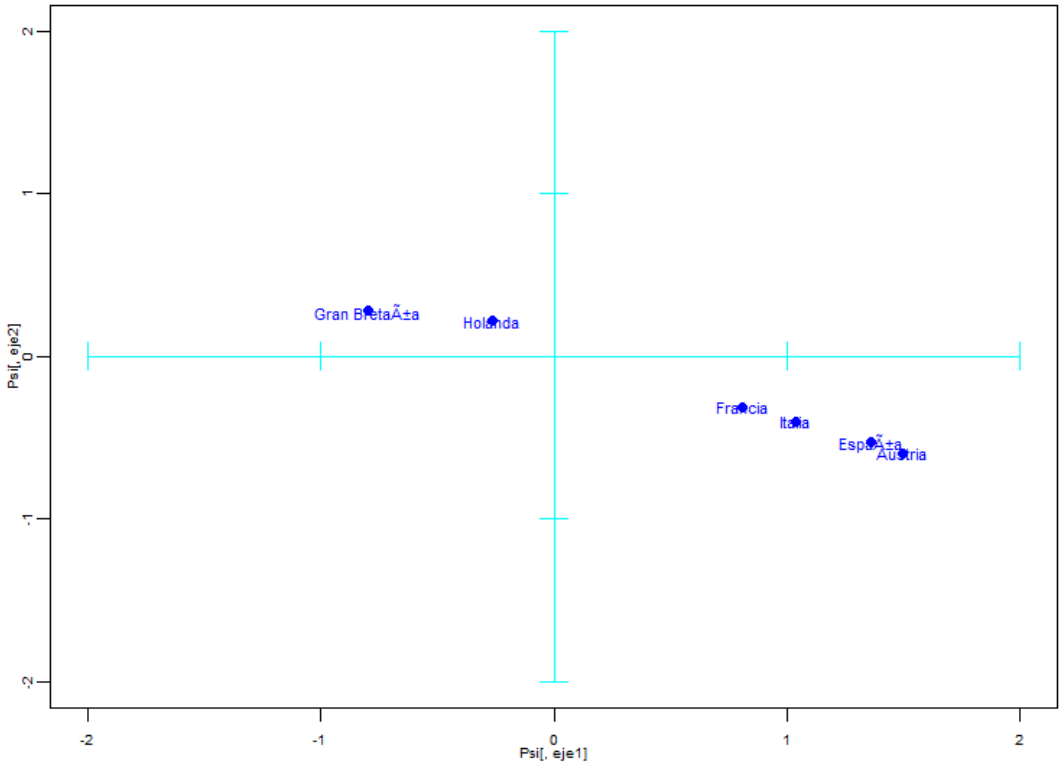


Figura 48: Hotel_Country al primer pla factorial

Veiem com, en general, les categories es troben força a prop del centre de coordenades i no podem inferir gaire cosa. Tanmateix, la modalitat "Gran Bretanya" es troba en la direcció de Additional_number_of_scoring. Aquest resultat concorda amb el que hem vist en l'etapa de profiling, ja que els hotels amb un nombre de valoracions addicionals (clúster 4) es troben majoritàriament a Gran Bretanya.

A continuació generem un gràfic amb totes les modalitats de totes les variables categòriques. La interpretació és la mateixa que anteriorment:

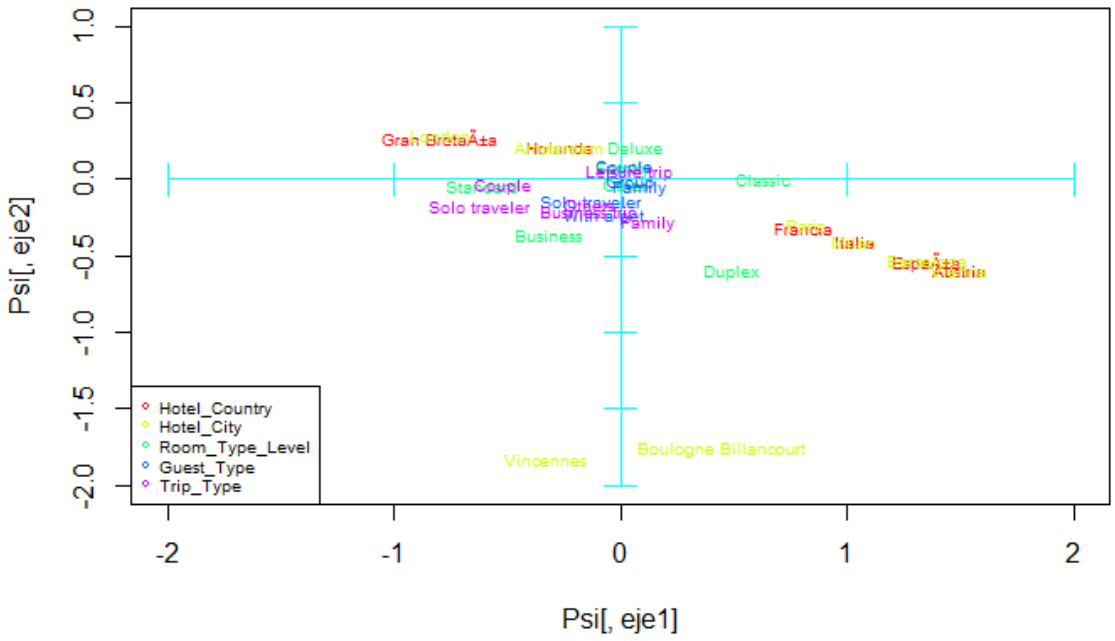


Figura 49: Projeccions de totes les modalitats sobre Dim.1 i Dim.2

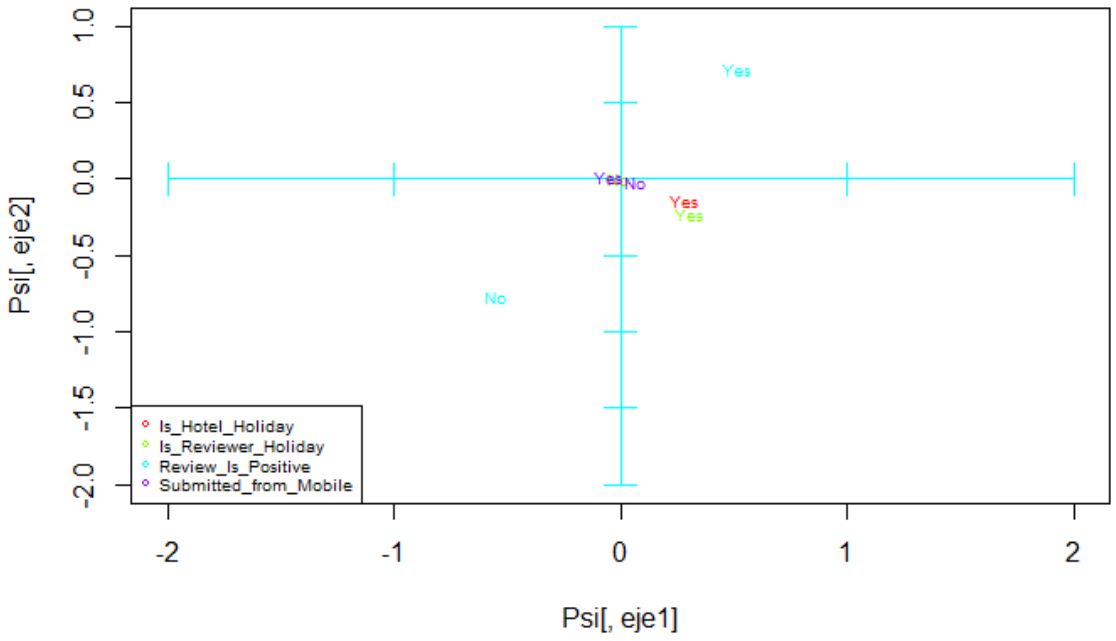
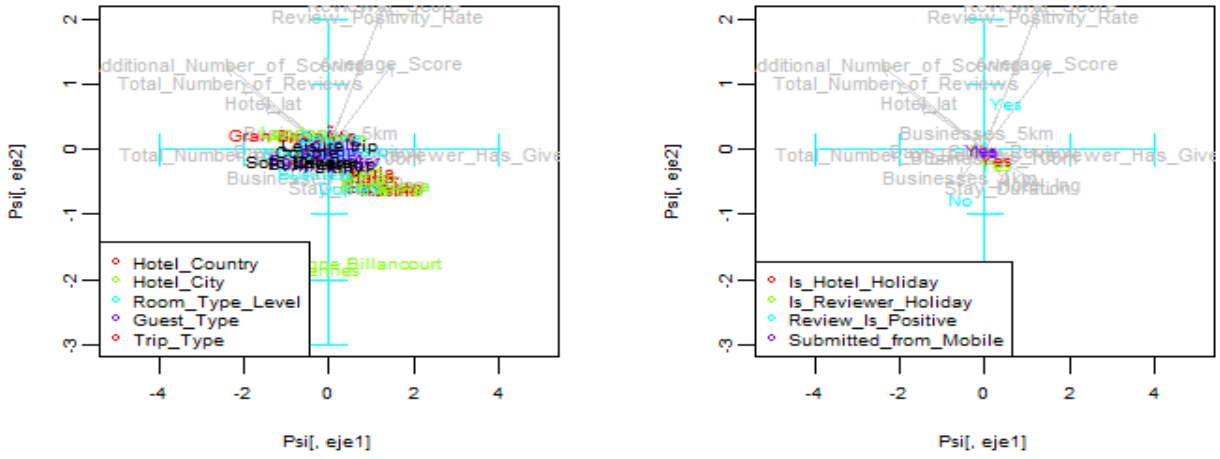


Figura 50: Projeccions de totes les modalitats sobre Dim.1 i Dim.2

Tanmateix, de cara a estudis futurs podria ser interessant representar les modalitats en un major nombre de dimensions. Mitjançant aquests gràfics podem saber quines modalitats de les variables categòriques tenen una major aportació a les dues primeres dimensions del nostre estudi. S'ha realitzat una separació de les variables categòriques en dos grups tal com es pot observar en els dos gràfics.



<u>Figura 51:</u> Projeccions deles modalitats amb variables numèriques

D'aquesta manera, podrem establir relacions entre variables determinant quines modalitats són indicatives de valors alts/baixos per qualsevol variable numèrica de la base de dades. La metodologia és idèntica a l'emprada per a construir els gràfics anteriors, únicament cal representar un fons amb les variables numèriques indicant les seves direccions de creixement respecte als plans factorials.

Biplots mixtes de variables i individus

Per últim, i com a ampliació, hem considerat la representació gràfica de diferents biplots on exposem juntament el núvol de punts de les projeccions dels individus sobre els eixos de components principals i les projeccions de les variables actives (numèriques). Hem inclòs les variables qualitatives en aquests gràfics creant agrupacions d'individus en base a les modalitats, de manera que podem identificar, al mateix temps, la posició en l'espai dels individus corresponents a una modalitat d'una variable qualitativa determinada.

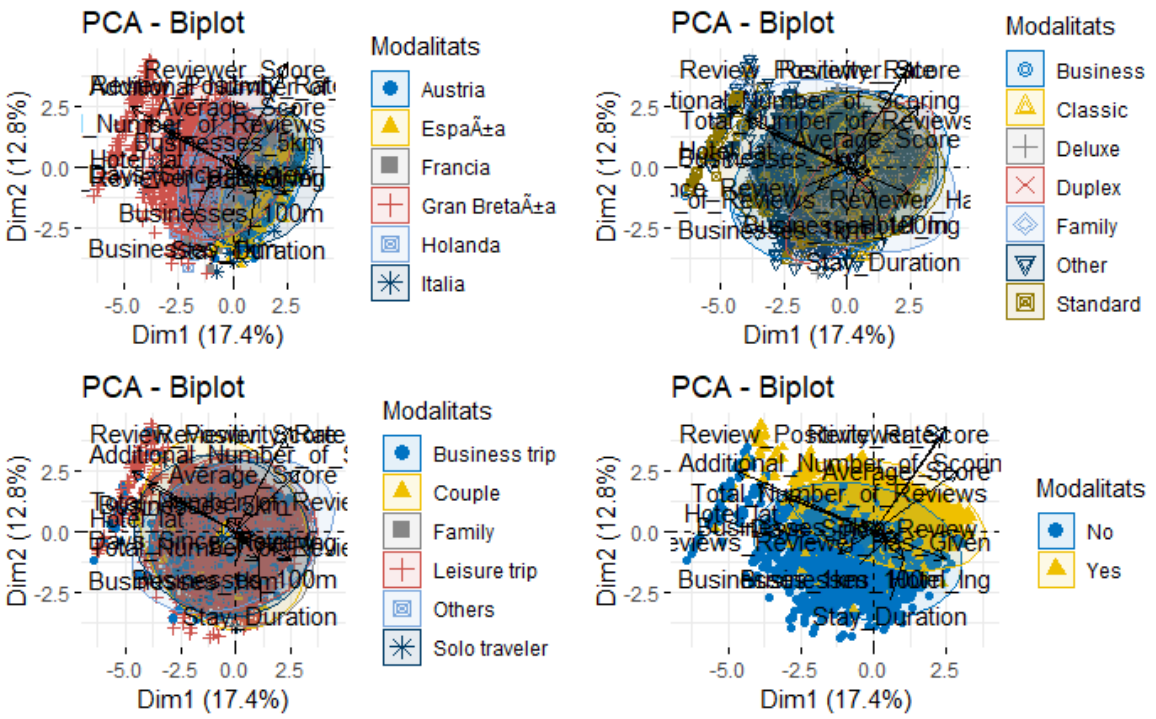


Figura 52: Representació de les variables sobre els PC (1er pla)

Observem com, en general, les categories es troben força sobreposades. Mitjançant aquests biplots podríem trobar relacions entre les diferents modalitats de les variables categòriques, les components principals i les variables numèriques.

iii. Mètodes discriminants.

En aquest apartat es durà a terme el mètode d'arbres de decisió. Aquest mètode analític facilita la presa de millors decisions (especialment quan existeixen riscs, costs, beneficis i opcions múltiples) a través d'una representació esquemàtica de les alternatives.

Cal destacar que l'arbre de decisió es durà a terme mitjançant el mètode "anova" que és l'apropiat en bases de dades en que la variable resposta és numèrica. A part d'això també cal dir que aquesta tècnica es durà a terme sense les variables Hotel_Name ni Reviewer_Nationality perquè són variables categòriques amb molts nivells i si es considera la seva inclusió a la tècnica, l'arbre no surt bé.

Cal destacar que per realitzar aquest mètode discriminant es farà servir la funció rpart de la llibreria

rpart per crear l'arbre i la funció rpart.plot de la llibreria rpart.plot per visualitzar-lo gràficament. Un cop realitzat el mètode, l'arbre resultant és el següent:

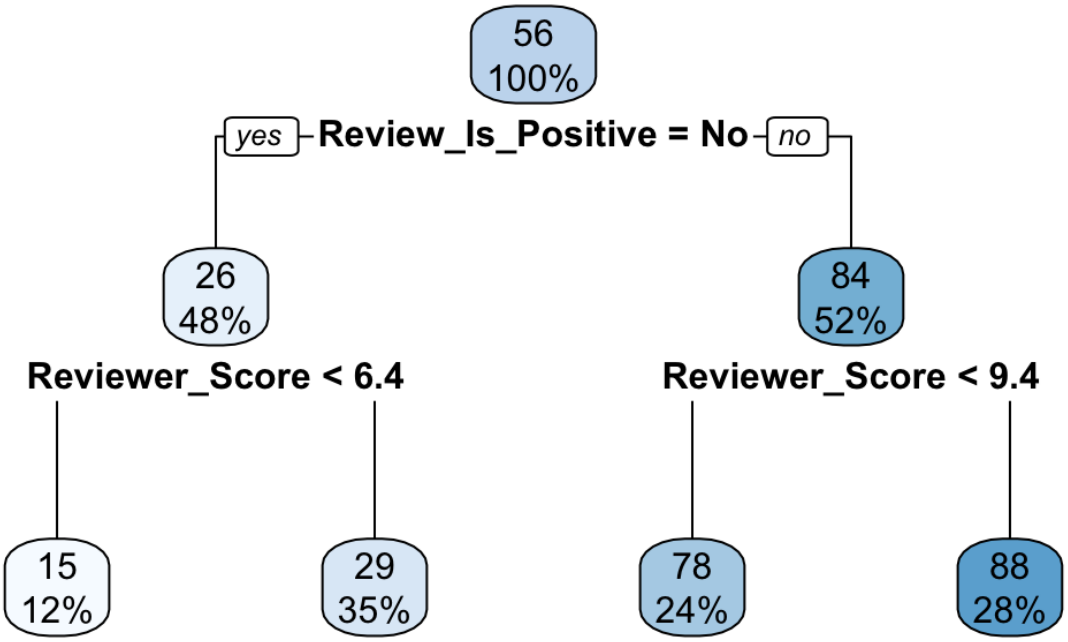


Figura 53: Arbre de decisió

Volem destacar que després de podar l'arbre utilitzant com a paràmetre de complexitat 'cp' el menor CP possible, s'ha obtingut com a resultat el mateix arbre que abans de podar-lo. Per tant, això significa que des d'un primer moment ja hem obtingut l'arbre més complex.

A partir de l'arbre obtingut podem arribar a les següents conclusions:

Les variables més importants a l'hora de determinar el fet de tenir una millor nota per a cada hotel són Reviewer_Score i Review_Is_Positive. Per tant, podem veure que les característiques que més influiran en la nota de l'hotel seran les valoracions dels usuaris (tal i com era d'esperar).

Per una banda tindrem que a major valoració global atorgada per l'usuari (Reviewer_Score), major serà la puntuació de l'hotel.

En general, quan la ressenya positiva de l'usuari conté més paraules que la ressenya negativa (Review_Is_Positive = YES) la nota de l'hotel serà major que quan la ressenya negativa és la que té més paraules (Review_Is_Positive = No). Per altra banda, en els casos on la ressenya positiva té més paraules, si el valor de Reviewer_Score és superior a 9.4, la nota de l'hotel serà major mentre que si aquesta valoració és menor a 9.4 s'obtindrà una nota de l'hotel pitjor. Mentre que en els casos on la ressenya negativa té més paraules, el punt de tall de la variable Reviewer_Score és 6.4, sent els hotels pitjor valorats els que es troben amb valors de Reviewer_Score per sota de 6.4 i una mica millor els que estan per sobre.

iv. Mètodes predictius.

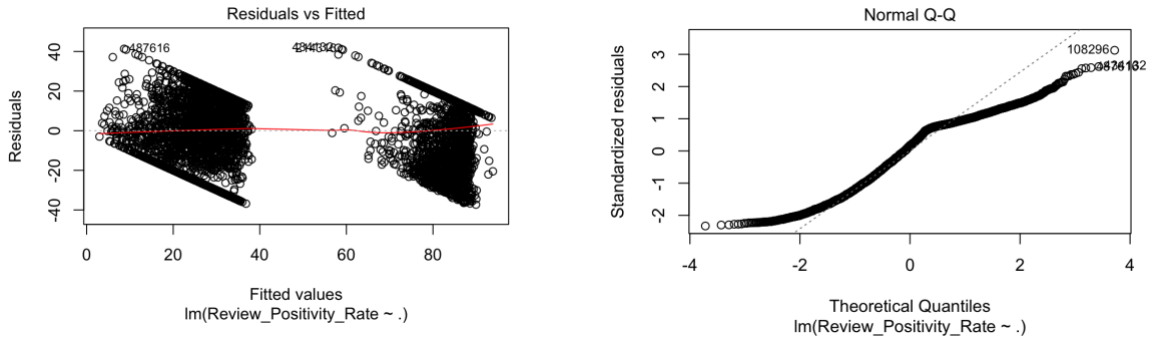
Aquest mètode consisteix en realitzar un model lineal amb la funció d'R "lm" la variable depenent del qual serà la nostra variable resposta (Reviewer_Positivity_Rate) i les explicatives seran la resta de variables de la nostra base de dades excepte Hotel_Name i Reviewer_Nationality ja que aquestes dos últimes són categòriques i tenen molts nivells, cosa que afavorirà a l'augment de la dificultat en la interpretació del model.

Un cop realitzat el model es pot obtindre una taula informativa d'aquest amb la funció "summary" d'R. Aquesta taula no serà adjuntada a l'informe per la seva voluminositat causada pel gran nombre de variables que hi intervenen però si que s'explicaran els trets informatius més importants que ens dóna aquesta funció.

En primer lloc ens fixarem en la significació global de les variables explicatives mitjançant el darrer contrast que dóna el summary. Aquest contrast consisteix en un test que utilitza una distribució F de Fisher amb 38 i 4961 graus de llibertat i té un estadístic de contrast igual a 453, per tant, el p-valor associat al test és molt inferior al nivell de significació del 5% i es pot dir que tenim indicis per rebutjar H0 (no significació de les variables explicatives) i dir que el model és útil per explicar la variable resposta. A partir d'aquest punt, si anem mirant variable per variable la seva significació individual ens adonem que si que hi ha bastantes variables que són rellevants a l'hora d'explicar la variable resposta ja que el p-valor associat al test de significació de la t-student és inferior al nivell de significació (també n'hi ha que resulten no significatives).

Seguidament, un cop validada la significació global del model es procedirà a comprovar la bondat de l'ajust d'aquest. Això es farà amb el coeficient de determinació R2, valor que ens dóna també el summary d'R. L'R2 d'aquest model ha resultat ser igual a 0.7763, valor una mica per sota dels punts desitjables ja que quan més proper a 1 es troba aquest coeficient, millor ajust tenim a les dades. El fet de tenir un R2 proper al 75% significa que amb el model només s'aconsegueix explicar un 75% de la variabilitat total de Reviewer_Positivity_Rate mentre que el 25% restant és deguda a altres factors que no es poden recollir amb el model lineal. Per tant, parlant de bondat de l'ajust, no tenim un model del tot desitjable però tampoc tenim un model molt dolent.

En tercer lloc farem servir els gràfics que ens dona el software R per tal de validar el model (linealitat, homoscedasticitat, normalitat…):



<u>Figura 54:</u> Gràfic dels residus vs ajustats i QQ -Plot del model de regressió

Mitjançant aquest dos gràfics podem veure si hi haurà o no linealitat, homoscedasticitat i normalitat. En el primer dels dos, el qual mostra els residus en front dels valors predits, es pot comprovar que els residus estan situats aleatòriament al voltant del zero i que la variabilitat d'aquests no varia a mesura que s'augmenta el valor de les prediccions però que es tenen dos grups clarament diferenciats, els que tenen un valor predit per sobre de 50 i els que el tenen per sota, per tant, hi ha dos grups que podríem qualificar com aprovats i suspesos. Per tant amb aquest primer gràfic es pot afirmar que el model assoleix la linealitat i l'homoscedasticitat. El segon gràfic testa la normalitat mitjançant un QQ-Plot la qual no sembla ser assolida ja que pateix molt als extrems. A part es troben alguns valors atípics.

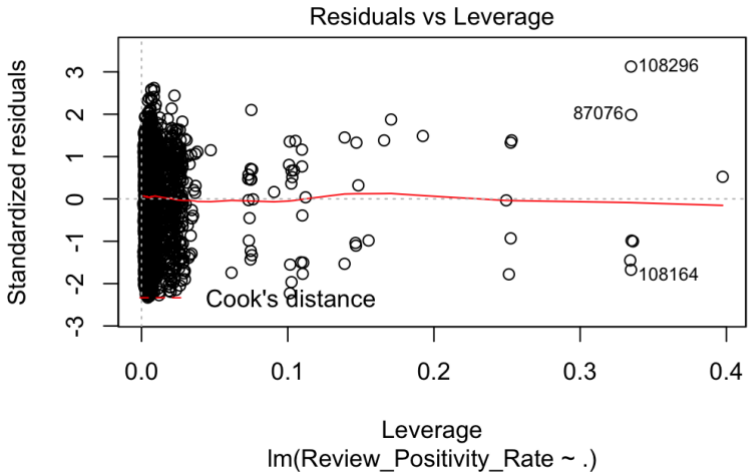


Figura 55: Gràfic dels residus vs Leverage del model de regressió

Finalment, amb aquest gràfic podem observar la més que possible presència de valors atípics així com els que tenen influència real i/o potencial. Hi ha en concret tres valors que es pot observar a simple vista que són influents, com per exemple el 108296 que té un residu estandaritzat al voltant del 4 (outlier) o el 87076 que té un leverage al voltant de 0.35 (potencialment influent).

6. Procés de mineria de dades de la divisió 2

i. Mètodes de profiling.

Clúster K-means

El Clúster K-means és un algoritme de machine learning que serveix per dividir un conjunt de dades en un conjunt de k grups (és a dir, k clústers), on k representa el nombre de grups predefinits. Es classifiquen els objectes en conglomerats de manera que dins de cadascún, els objectes siguin el més similars possible, mentre que els objectes de diferents clústers siguin tan diferents com sigui possible. És important tenir en compte que és un mètode no supervisat, el que implica que busca trobar relacions entre les n observacions sense ser entrenades per una variable resposta.

El mètode d'agrupació k-means només pren les variables numèriques per a que intervinguin en el procés de creació dels conglomerats . A més, les dades s'estandarditzen amb l'objectiu que les variables siguin comparables. L'estandardització consisteix en transformar les variables numèriques de manera que la seva mitjana sigui 0 i la desviació estàndard 1.

La idea bàsica del algorisme K-means consisteix en la definició de clústers de manera que es minimitzi la variació total intra clúster. Hi ha diversos algoritmes de k-means disponibles. La variació total intra clúster pot ser definida com la suma de les distàncies quadrades de les distàncies euclidianes entre els ítems i el centroide corresponent:

$$W(C_k) = \sum_{x_i \in C_k} (x_i - \mu_k)^2$$

On x_i és una observació pertanyent al cluster C_k i μ_k és el valor mitjà dels punts assignats al clúster. Cada observació (x_i) s'assigna a un clúster determinat de manera que la suma de quadrats (SS) de distància de l'observació als seus centroides (μ_k) es minimitza.

Definim la suma de la variació total dins del clúster de la manera següent:

$$\text{tot.withiness} = \sum_{k=1}^{k} W(C_k) = \sum_{k=1}^{k} \sum_{x_i \in C_k} (x_i - \mu_k)^2$$

Aquesta mesura la compacitat (i la bondat) del clúster i volem que sigui tan petita com sigui possible.

L'algoritme K-means pot resumir-se de la manera següent:

1. Especificar el nombre de clústers (K).

2. Se selecciona aleatòriament k objectes del conjunt de dades com a centroides dels clústers inicials.

3. S'assigna a cada observació al seu centroide més proper, sobre la base de la distància euclidiana entre l'objecte i el centroide.

4. Per a cada un dels k clústers, s'actualitza el centroide calculant els nous valors mitjans de tots els punts de dades del clúster.

5. Iterativament es minimitza la variació total dins del clúster. És a dir, es repeteixen els passos 3 i 4 fins que les assignacions de grups no deixin de canviar o el nombre màxim d'iteracions s'aconsegueixi. Per defecte, el programari R utilitza 10 com a valor predeterminat per al nombre màxim d'iteracions.

Per determinar de forma òptima el nombre de conglomerats (k) s'ha emprat el mètode del colze (Elbow Method) que es basa en el mateix principi de la suma de quadrats intra clústers. Els resultats suggereixen que 4 és el nombre òptim de clústers, ja que sembla ser la corba del colze:

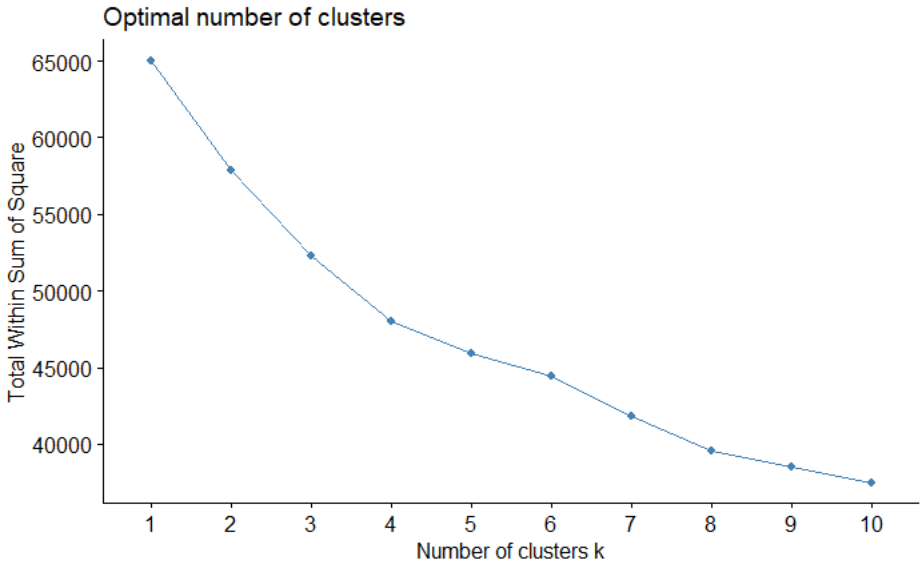


Figure 56: Elbow Method

Podem visualitzar els resultats de la divisió de les dades estandarditzades en el següent gràfic (Figura x) on les observacions estan representades per punts en el primer pla de les components principals. Veiem que els 4 grups estan perfectament diferenciats.

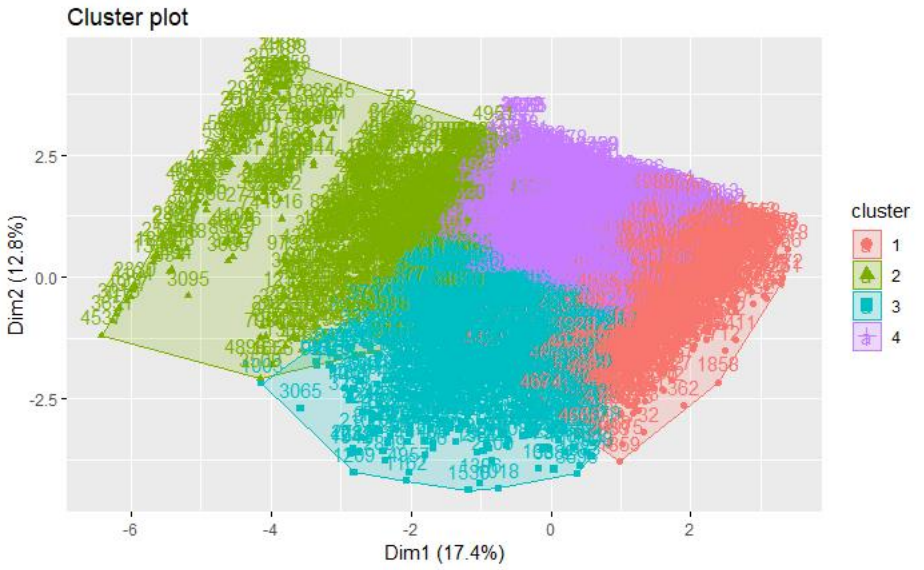


Figure 57: Bar Cluster plot

En aquesta taula (Taula x) veiem el número d'individus que han caigut en cada grup:

Cúster	*Nº individos*
1	*1175*
2	*583*
3	*1221*
4	*2021*

Taula 7: Repartició Clustering K-means

Profiling dels clústers

Ara que ja es tenen els clústers creats convé descobrir quines particularitats té cadascun d'ells i comprendre com són les unitats d'estudi dins de cada conglomerat. Per a aquest propòsit fem servir les variables que han participat en la creació dels grups, és a dir, les variables numèriques, que són les que apareixen a la següent taula (Taula x). En aquesta taula, obtinguda amb la funció `catdes()` d'R trobem el grau de relació entre cada variable numérica i la variable clúster. Veiem que la variable més relacionada amb la variable clúster és *Average_Score* ja ques és la que te un p-valor associat més petit, seguit de *Businesses_5km* i *Business_1km,* aquestes són les que han contribuit més en la creació dels clústers.

	Eta2	*P-value*
Hotel_lat	*0.668942*	*0.00E+00*
Hotel_lng	*0.38996*	*0.00E+00*
Total_Number_of_Reviews	*0.556044*	*0.00E+00*
Review_Positivity_Rate	*0.300284*	*0.00E+00*
Reviewer_Score	*0.413507*	*0.00E+00*
Additional_Number_of_Scoring	*0.611458*	*0.00E+00*
Average_Score	*0.236572*	*3.92E-292*
Businesses_5km	*0.102841*	*3.36E-117*
Businesses_1km	*0.056018*	*3.86E-62*
Stay_Duration	*0.027163*	*1.24E-29*
Total_Number_of_Reviews_Reviewer_Has_Given	*0.016128*	*1.66E-17*
Businesses_100m	*0.009454*	*2.77E-10*

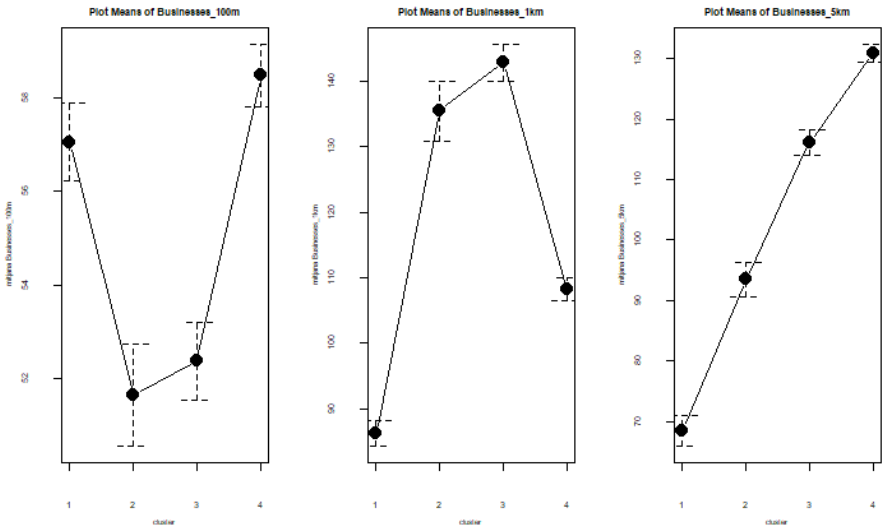
Taula 8: Relació entre la variable de clúster i les variables quantitatives

- ***Hotel_lat, Hotel_lng***

Les variables *Hotel_lat* i *Hotel_lng* fan referència a la latitud i longitud dels hotels, per aquest motiu és complicat utilitzar-les per fer una diferenciació entre els quatre clústers.

- ***Business_100m, Business_1km, Business_5Km***

Les variables *Business_100m*, *Business_1km* i *Business_5Km* fan referència al nombre de negocis a la rodona considerant diferents distàncies, per això les representarem juntes mitjançant un plotMean. L'objectiu cercat amb aquestes variables és fer una distinció entre hotels urbans i hotels més allunyats del centre de les ciutats.



<u>Figura 58:</u> Profiling Nre. Negocis a 100m, 1km i 5km a la rodona

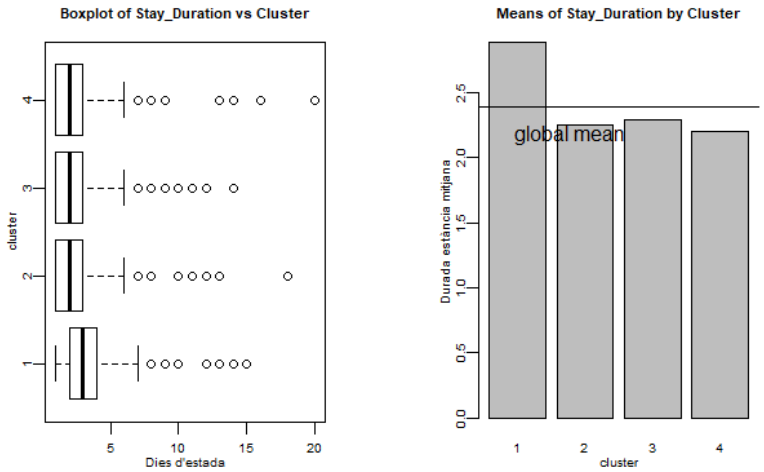
S'observa que els hotels de les ressenyes tenen...

- Clúster 1: molts establiments a 100m, pocs a 1km i a 5km.
- Clúster 2: pocs establiments a 100m, molts establiments a 1km.
- Clúster 3: pocs establiments a 100m, molts a 1km i 5km.
- Clúster 4: molts establiments a 100m i 5km. → hotels cèntrics

- ***Stay_Duration***

La següent variable és *Stay_Duration*. Observem com els clústers 2, 3 i 4 són molt similars en el fet que tendeixen a incloure estades més curtes, mentre que el clúster 1 inclou estades més llargues.

Clúster	*Min.*	*1st Qu.*	*Median*	*Mean*	*3rd Qu.*	*Max.*
1	*1.000*	*2.000*	*3.000*	*2.981*	*4.000*	*15.000*
2	*1.00*	*1.00*	*2.00*	*2.25*	*3.00*	*18.00*
3	*1.000*	*1.000*	*2.000*	*2.2288*	*3.000*	*14.000*
4	*1.000*	*1.000*	*2.000*	*2.202*	*3.000*	*20.000*



<u>Figura 59:</u> Profiling Durada de l'estada

- **Days_Since_Review**

La següent variable és *Days_Since Review*. Observem que la distribució d'aquesta variable és molt similar en cadascun dels 4 conglomerats com s'observa al gràfic (Figura).

Clúster	*Min.*	*1st Qu.*	*Median*	*Mean*	*3rd Qu.*	*Max.*
1	*0.0*	*191.0*	*353.0*	*355.9*	*513.0*	*730.0*
2	*0.0*	*164.0*	*347.0*	*355.1*	*534.0*	*730.0*
3	*0.0*	*157.0*	*335.0*	*345.5*	*522.0*	*730.0*
4	*0.0*	*193.0*	*362.0*	*361.6*	*530.0*	*730.0*

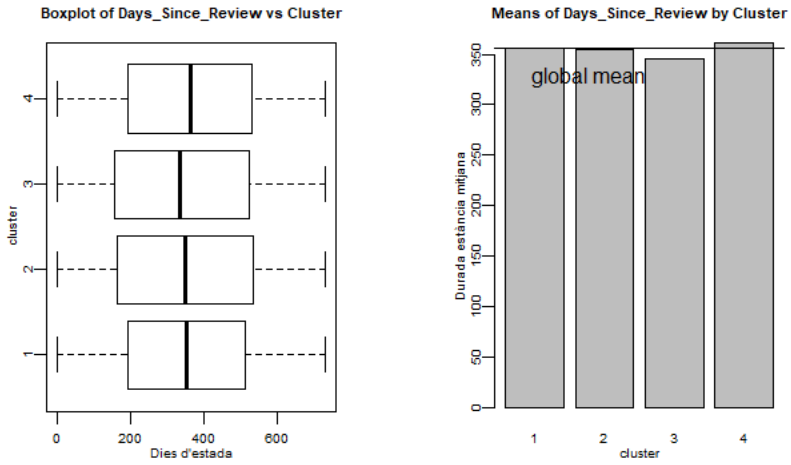


Figura 60: Profiling Nre. Dies desde l'última Ressenya

- ***Total_Number_of_Reviews***

A continuació, ens fixem en la variable *Total_Number_of_Reviews*. Aquesta variable, recordem, representa el nombre total de ressenyes vàlides que té l'hotel en qüestió.
Al gràfic (Figura) veiem com, en nombre de ressenyes vàlides destaca el clúster 1. Aquest es troba considerablement per sobre la mitjana global. Un estudi interessant seria observar com això repercuteix en la valoració dels hotels d'aquest grup.

Clúster	*Min.*	*1st Qu.*	*Median*	*Mean*	*3rd Qu.*	*Max.*
1	*172*	*1264*	*2011*	*2421*	*3461*	*7371*
2	*867*	*5726*	*6608*	*7357*	*8177*	*16670*
3	*74*	*1179*	*1945*	*2166*	*2898*	*6792*
4	*60*	*941*	*1686*	*1920*	*2692*	*7108*

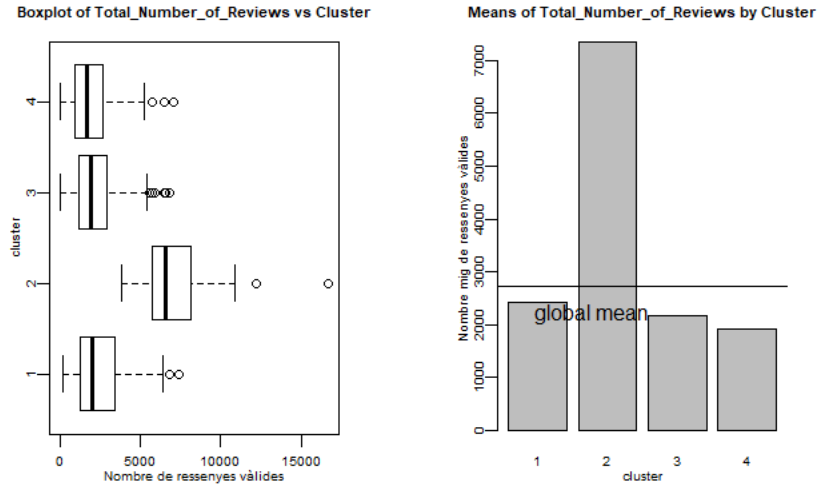


Figura 61: Profiling Nre. Total de Ressenyes

- ***Review_Positivity_Rate***

La següent variable està relacionada amb el grau de positivitat de la ressenya de Booking i ens pot donar un indici de quin dels clústers els comentaris són més positius o, al contrari, negatius.
S'observa que en els conglomerats 1 i 4 domina el grau de positivat ja que en mitjana aquest és del 62.08% i 71.73, respectivament. En el clúster 3 predominen els comentaris negatius ja que el grau de positivat és del 25.20%. Per altra banda, en el clúster 2 hi ha el mateix grau de positivat que de negativitat ja que el grau de positivitat en mitjana és del 51.43%.

Clúster	*Min.*	*1st Qu.*	*Median*	*Mean*	*3rd Qu.*	*Max.*
1	*0.00*	*35.71*	*60.87*	*62.08*	*100.00*	*100.00*
2	*0.00*	*25.32*	*49.51*	*51.43*	*77.78*	*100.00*
3	*0.000*	*6.667*	*21.429*	*25.202*	*38.000*	*100.000*
4	*0.00*	*50.00*	*76.19*	*71.73*	*100.00*	*100.00*

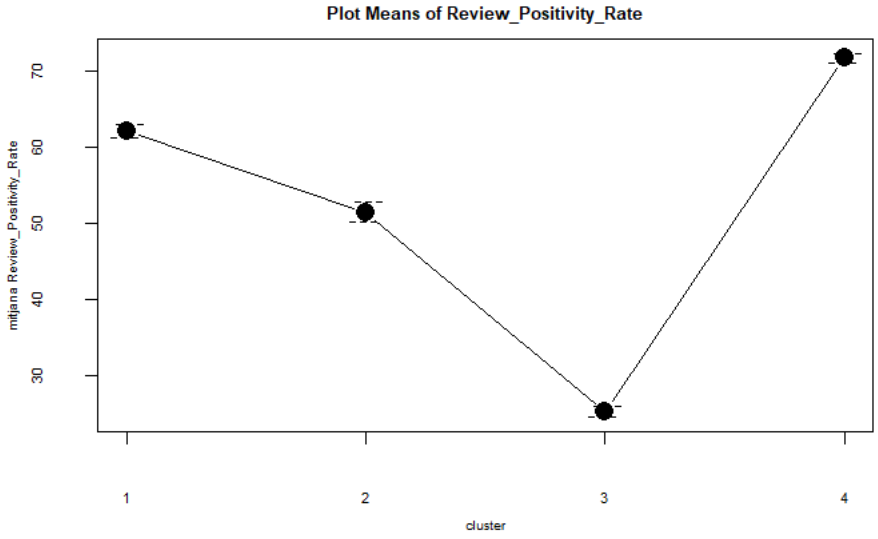


Figura 62: Profiling Grau de Positivitat de la Ressenya

- ***Average_Score, Reviewer_Score***

Tot seguit entrem a analitzar variables relacionades amb les puntuacions dels hotels. Prenem conjuntament la puntuació mitjana que presentava l'hotel a finals de 2016 i la que han anat atorgant els usuaris de Booking que han escrit les ressenyes. Teòricament aquells hotels on les valoracions són més positives haurien de rebre millors valoracions així que, sembla lògic pensar que els resultats haurien d'anar en línia amb l'anterior anàlisi de la variable *Review_Positivity_Rate*.

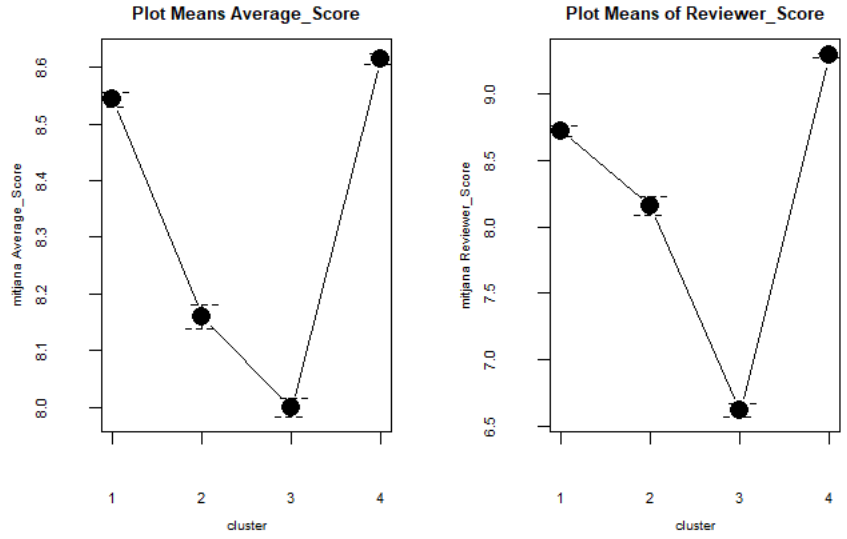


Figura 63: Profiling Puntuació Mitjana i Puntuació del Usuari

Efectivament, els clústers 4 és el que té les puntuacions més altes seguit del clúster 1. La classe 3 és la que té les puntuacions més baixes, mentre que la classe 2 es manté a la mitjana.

- ***Total_Number_of_Reviews_Reviewer_Has_Given***

La següent variable que analitzem és *Total_Number_of_Reviews_Reviewer_Has_Given*. Aquesta ens dona informació sobre com d'actiu a Booking és l'usuari que ha escrit la ressenya (Figure). En aquest sentit, observem com els usuaris relatius al clúster 1 són molt més actius que la mitjana, al clúster 4 es mantenen a prop de la mitjana global, i els del segon i tercer conglomerat, on la mitjana de comentaris totals escrits pels usuaris és més baixa.

Clúster	*Min.*	*1st Qu.*	*Median*	*Mean*	*3rd Qu.*	*Max.*
1	*1.000*	*1.000*	*5.000*	*9.699*	*12.000*	*140.000*
2	*1.000*	*1.000*	*2.000*	*5.859*	*6.000*	*143.000*
3	*1.000*	*1.000*	*3.000*	*5.864*	*7.000*	*78.000*
4	*1.000*	*1.000*	*3.000*	*7.245*	*9.000*	*156.000*

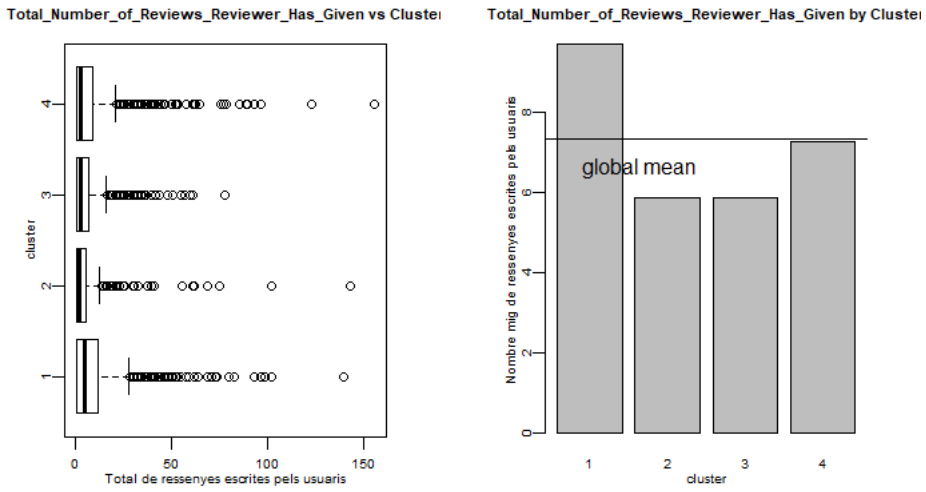


Figura 64: Profiling Nre. De Ressenyes escrites per l'Usuari

- ***Additional_Number_of_Scoring***

Finalment ens fixem en la variable *Additional_Number_of_Scoring* relativa al total de valoracions addicionals que rep l'hotel (localització, neteja, servei. . .). Veiem com els usuaris del clúster 2 tendeixen molt més a valorar aspectes extra de l'hotel.

Clúster	*Min.*	*1st Qu.*	*Median*	*Mean*	*3rd Qu.*	*Max.*
1	*11.0*	*122.0*	*206.0*	*229.1*	*315.0*	*666.0*
2	*563*	*1172*	*1322*	*1543*	*1936*	*2682*
3	*8.0*	*198.0*	*364.0*	*408.9*	*556.0*	*1299.0*
4	*11.0*	*161.0*	*355.0*	*405.2*	*602.0*	*1258.0*

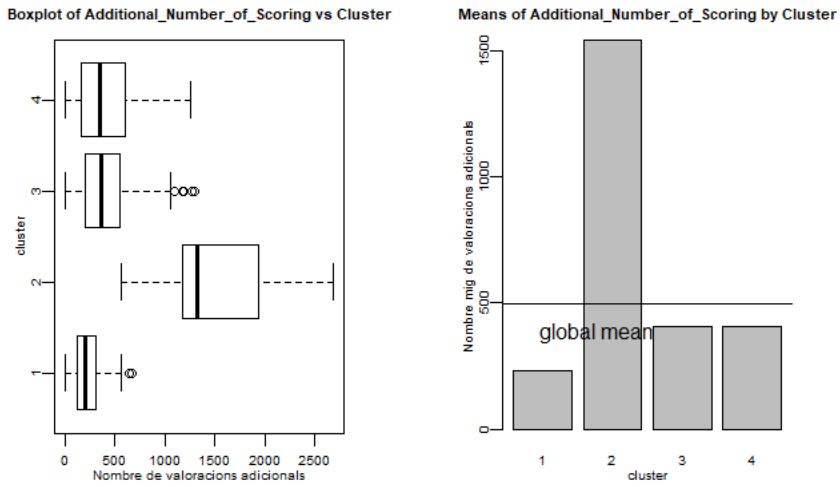


Figura 65: Profiling Nre. De Valoracions Addicionals

Un cop tenim recopilada tota la informació descriptiva de cada clúster, podem resumir-la en una taula, donant un nom al segment en qüestió i elaborant una petita descripció que inclogui els seus atributs principals.

Clúster	Nom	Descripció
1	Viatger atent	Hotels amb puntuació alta que reben usuaris molt actius a Booking i que hi passen llargues estades.
2	Viatger conformista	Hotels amb puntuació modesta tirant a alta que reben usuaris poc actius a Booking. Tenen moltes ressenyes vàlides i valoracions extres.
3	Viatger despistat	Hotels amb puntuació baixa que reben usuaris poc actius a Booking.
4	Viatger que busca luxe	Hotels cèntrics amb puntuació molt alta.

Taula 9: Taula resum profiling Clustering K-means

ii. Mètodes associatius.

Els mètodes associatius permeten conèixer patrons freqüents, associacions, correlacions o estructures causals d'entre els elements d'una base de dades de transaccions. Així doncs, el primer que fem per a realitzar aquests mètodes és seleccionar únicament aquelles variables que siguin categòriques. Després, hem transformat la nostra de base de dades en una base de dades *transaccional*.

Estudi dels ítems més freqüents.

En total hi ha 2018 nivells diferents entre les 12 variables categòriques que tenim a la base de dades. I més concretament, en cada variable hi ha el següent nombre de nivells:

Hotel_name	1135
Hotel_country	6
Hotel_City	8
Room_Type_Level	7
Guest_Type	5
Trip_Type	6
Review_Date	720
Is_Hotel_Holiday	2
Is_Reviewer_Holiday	2
Review_Is_Positive	2
Reviewer_Nationality	123
Submitted_from_Mobile	2
TOTAL	2018

Si realitzem un resum numèric del arxiu transaccional, obtenim que els ítems o elements més freqüents són:

- Is_Reviewer_Holiday = No
- Is_Hotel_Holiday = No
- Trip_Type = Leisure trip
- Room_Type_Level = Other
- Submitted_from_Mobile = Yes

Analitzant-ho gràficament, obtenim la següent representació:

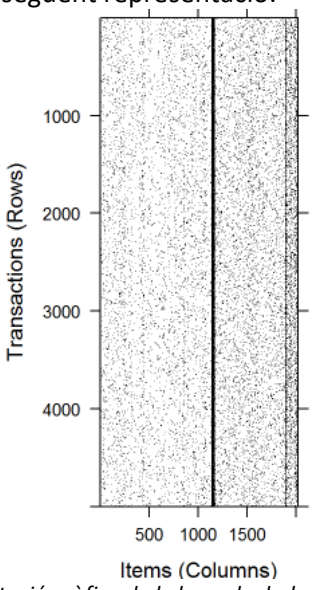


Figura 66: Representació gràfica de la base de dades transaccional completa.

Com ja s'ha comentat anteriorment, a la base de dades hi ha 5000 observacions (transaccions o files, observables a l'eix y del gràfic) y 2018 nivells o classes (representats a l'eix x). Les línies amb menys espais en blanc (les més negres) són les categories que més cops s'han donat.

Com que hi ha massa nivells, a la imatge no podem treure conclusions perquè no es veu res clar i per tant, es faran alguns canvis a la base de dades:

1. La variable `Hotel_name` té molts nivells tot ja que es tracta d'un "identificador", el nom dels hotels no ens és útil com a variable explicativa.
2. La variable `Review_Date`, que té tants nivells com dies diferents hagin fet la ressenya els clients, s'agruparà en anys.

El nou gràfic obtingut és el següent:

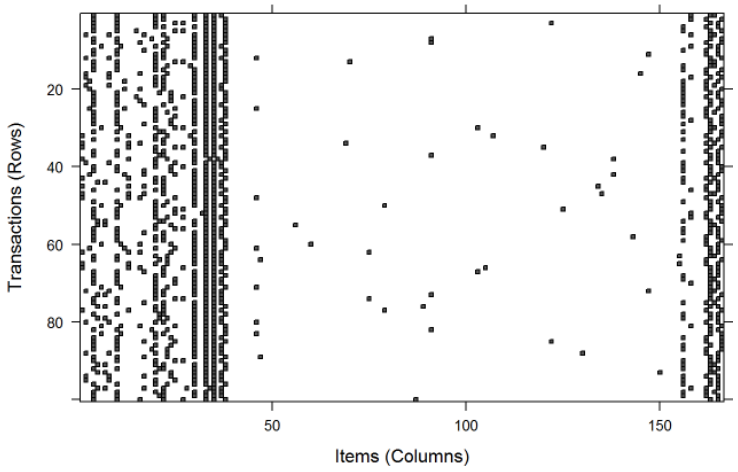


Figura 67: Representació gràfica de la base de dades transaccional modificada.

S'observa com els items compresos entre el nombre 50 o 150, més o menys, són poc freqüents (es veuen pocs punts negres, poques repeticions). Sabent que el nombre de nivells per categoria que tenim en aquest nova base de dades és:

		Acumulat
Hotel_country	6	6
Hotel_City	8	14
Room_Type_Level	7	21
Guest_Type	5	26
Trip_Type	6	32
Is_Hotel_Holiday	2	34
Is_Reviewer_Holiday	2	36
Review_Is_Positive	2	38
Reviewer_Nationality	123	161
Submitted_from_Mobile	2	163
Review_date_year	3	166
TOTAL	166	-

Podem accedir a quines són les categories més freqüents fent

```
levels(dcat$Factor[Nº de la categoria que es vol mirar])
```

Per exemple, veiem que les linies 4 i 10 són molt negres, mirem quines són:

- Com que el primer factor (Hotel_country) té 6 nivells, el nivell 4 serà la tercera categoria de la variable Hotel_country, que és "Gran Bretanya".
- Per a mirar la categoria 10, veiem amb la taula de les categories acumulades que la 10 es correspon a una de les categories de la variable Hotel_City i, més concretament, a la 10-6=4. És la categoría "London".

Això ho veiem també amb el summary i un itemFrequencyPlot (gràfic dels ítems més freqüents):

```
## transactions as itemMatrix in sparse format with
##  5000 rows (elements/itemsets/transactions) and
##  166 columns (items) and a density of 0.06626506
##
## most frequent items:
##    Is_Reviewer_Holiday=No       Is_Hotel_Holiday=No
##                      4933                      4930
##    Trip_Type=Leisure trip     Room_Type_Level=Other
##                      4037                      3209
## Submitted_from_Mobile=Yes                   (Other)
```

Figura x: Summary de la base de dades transaccional nova.

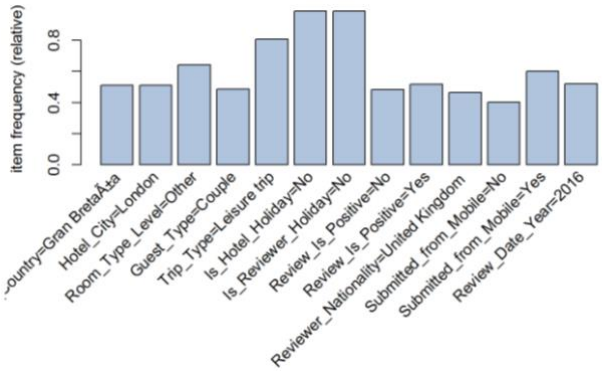


Figura 68: Histograma dels elements més freqüents.

Ara, s'utilitzarà un algorisme per a generar itemsets (grups d'elements) freqüents, el **Eclat** (*Equivalence Class Transformation*). Es mirarà el *support* de cada itemset, i s'agafaran els més alts. Cal recordar que el *support* d'un itemset és la fracció de transaccions que el contenen:

$$Support(X) = \frac{Frequency \quad occurrence \quad of \quad X}{|\tau|}$$

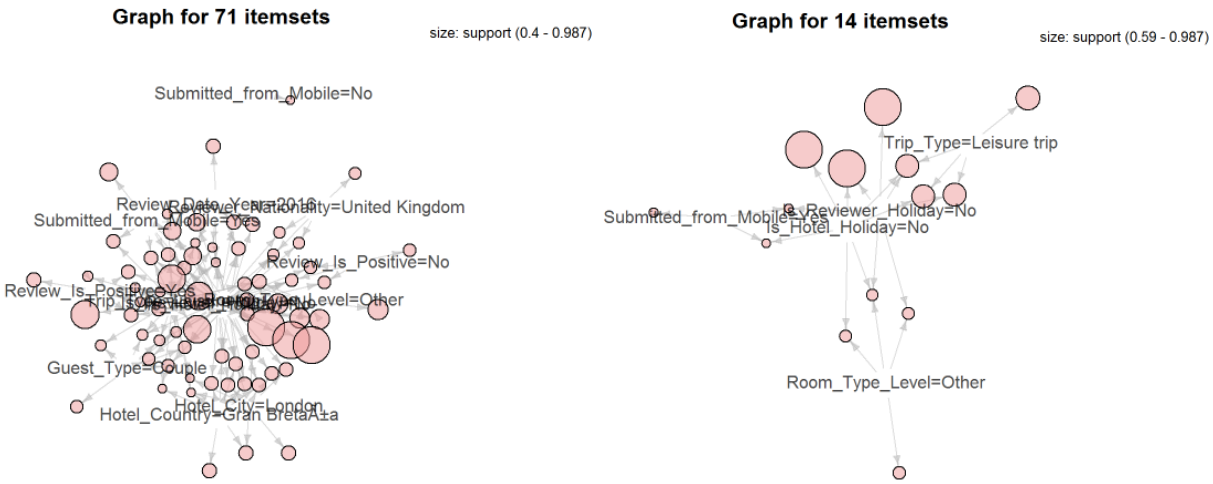
Si apliquem la funció inpect amb l'argument by "support", s'obtenen com a items molt freqüents alguns dels que s'han trobat abans:

- Is_Reviewer_Holiday=No
- Is_Hotel_Holiday=No
- Trip_Type=Leisure trip
- Room_Type_Level=Other
- Submitted_from_Mobile=Yes

Però ara, a més, trobem itemsets de dos elements que abans no haviem obtingut:

- {Is_Hotel_Holiday = No, Is_Reviewer_Holiday = No}
- {Trip_Type = Leisure trip, Is_Reviewer_Holiday = No}
- {Trip_Type = Leisure trip, Is_Hotel_Holiday = No}

Ho veiem gràficament:



<u>Figures 69 i 70:</u> Plots dels 71 itemsets més freqüents (esquerra) i dels 14 amb el support més alt (dreta.

<u>Regles d'associació.</u>

A continuació s'utilitzarà l'algorisme Apriori (tenint en compte la confiança i el lift) per tal de trobar regles d'associació en la nostra base de dades.

- **Confidence.**

La confiança ens indica quants cops la regla (X→Y)ha resultat ser certa. Així doncs, una confiança alta indica una alta proporció de transaccions que contenen el item X que també conté Y.

$$Confidence(X \rightarrow Y) = \frac{Support(X \cup Y)}{Support(X)}$$

Ho apliquem a la nostra base de dades:

```
     lhs                                                rhs                          support confidence      lift count
[1]  {Hotel_City=Milan}                              => {Hotel_Country=Italia}        0.0686          1 14.577259   343
[2]  {Hotel_Country=Italia}                          => {Hotel_City=Milan}            0.0686          1 14.577259   343
[3]  {Reviewer_Nationality=United States of America } => {Is_Reviewer_Holiday=No}     0.0730          1  1.013582   365
[4]  {Hotel_City=Vienna}                             => {Hotel_Country=Austria}       0.0752          1 13.297872   376
[5]  {Hotel_Country=Austria}                         => {Hotel_City=Vienna}           0.0752          1 13.297872   376
[6]  {Hotel_City=Amsterdam}                          => {Hotel_Country=Holanda}       0.1116          1  8.960573   558
[7]  {Hotel_Country=Holanda}                         => {Hotel_City=Amsterdam}        0.1116          1  8.960573   558
[8]  {Hotel_City=Barcelona}                          => {Hotel_Country=EspaÃ±a}       0.1148          1  8.710801   574
[9]  {Hotel_Country=EspaÃ±a}                         => {Hotel_City=Barcelona}        0.1148          1  8.710801   574
[10] {Hotel_City=Paris}                              => {Hotel_Country=Francia}       0.1178          1  8.403361   589
[11] {Hotel_City=London}                             => {Hotel_Country=Gran BretaÃ±a} 0.5108          1  1.957713  2554
[12] {Hotel_Country=Gran BretaÃ±a}                   => {Hotel_City=London}           0.5108          1  1.957713  2554
[13] {Is_Hotel_Holiday=No,
      Reviewer_Nationality=United Arab Emirates }    => {Is_Reviewer_Holiday=No}      0.0238          1  1.013582   119
[14] {Is_Reviewer_Holiday=No,
      Reviewer_Nationality=United Arab Emirates }    => {Is_Hotel_Holiday=No}         0.0238          1  1.014199   119
[15] {Room_Type_Level=Family,
      Is_Hotel_Holiday=No}                           => {Is_Reviewer_Holiday=No}      0.0254          1  1.013582   127
```

Observem que hi ha regles que són molt òbvies, aquelles en que es ja que relaciona una ciutat amb un país, ja que aquestes prediccions es poden fer molt fàcilment: si sabem la ciutat del hotel (Hotel_City) sempre podrem saber el país del Hotel (Hotel_Country).

Però, a més, es mostren altres regles interessants a comentar:

- {Reviewer_Nationality = United States of America } => {Is_Reviewer_Holiday = No}
- {Room_Type_Level = Family, Is_Reviewer_Holiday = No} => {Is_Hotel_Holiday = No}

Com que la confiança per a aquestes regles és 1, això significa que sempre que una transacció conté Reviewer_Nationality=United States of America, la variable Is_Reviewer_Holiday ha estat **No**.

- **Lift**

L'elevació (*Lift*) mesura la relació entre el suport observat i esperat si els itemsets X i Y fossin independendts. Tres possibles resultats:

- **Lift = 1:** la probabilitat d'ocurrència de X i Y es independent (no hi ha regla d'associació)
- **Lift > 1:** hi ha cert grau de dependencia, per tant existeix una regla útil per a realitzar prediccions.
- **Lift < 1:** els ítems es substitueixen entre si, és a dir, quan apareix X en una transació no apareix Y i a l'inversa.

$$Lift(X \rightarrow Y) = \frac{Support(X \cup Y)}{Support(X) \cdot Support(Y)}$$

Així doncs, buscarem regles a priori ordenant-les segons el valor del lift per obtenir aquelles regles amb els lifts més alts.

```
     lhs                                rhs                          support confidence     lift count
[1]  {Trip_Type=Leisure trip,
      Is_Reviewer_Holiday=Yes}       => {Is_Hotel_Holiday=Yes}       0.0108  0.9818182 70.12987    54
[2]  {Trip_Type=Leisure trip,
      Is_Hotel_Holiday=Yes}          => {Is_Reviewer_Holiday=Yes}    0.0108  0.9310345 69.48019    54
[3]  {Is_Reviewer_Holiday=Yes}      => {Is_Hotel_Holiday=Yes}       0.0130  0.9701493 69.29638    65
[4]  {Is_Hotel_Holiday=Yes}         => {Is_Reviewer_Holiday=Yes}    0.0130  0.9285714 69.29638    65
[5]  {Hotel_Country=Italia}         => {Hotel_City=Milan}           0.0686  1.0000000 14.57726   343
[6]  {Hotel_City=Milan}             => {Hotel_Country=Italia}       0.0686  1.0000000 14.57726   343
[7]  {Hotel_Country=Italia,
      Guest_Type=Group}             => {Hotel_City=Milan}           0.0110  1.0000000 14.57726    55
[8]  {Hotel_City=Milan,
      Guest_Type=Group}             => {Hotel_Country=Italia}       0.0110  1.0000000 14.57726    55
[9]  {Hotel_Country=Italia,
      Trip_Type=Business trip}      => {Hotel_City=Milan}           0.0122  1.0000000 14.57726    61
[10] {Hotel_City=Milan,
      Trip_Type=Business trip}      => {Hotel_Country=Italia}       0.0122  1.0000000 14.57726    61
[11] {Hotel_Country=Italia,
      Guest_Type=Family}            => {Hotel_City=Milan}           0.0134  1.0000000 14.57726    67
[12] {Hotel_City=Milan,
      Guest_Type=Family}            => {Hotel_Country=Italia}       0.0134  1.0000000 14.57726    67
[13] {Hotel_Country=Italia,
      Review_Date_Year=2015}        => {Hotel_City=Milan}           0.0110  1.0000000 14.57726    55
[14] {Hotel_City=Milan,
      Review_Date_Year=2015}        => {Hotel_Country=Italia}       0.0110  1.0000000 14.57726    55
[15] {Hotel_Country=Italia,
      Guest_Type=Solo traveler}     => {Hotel_City=Milan}           0.0154  1.0000000 14.57726    77
```

Totes les regles obtingudes mostren una forta associació entre els itemsets de l'esquerra i els de la dreta.

iii. Mètodes discriminants.

Per realitzar els diferents mètodes d'anàlisi discriminant realitzarem una mineria de dades, dividint en una mostra d'entrenament i una altra de test, la mostra d'entrenament inclourà el 70% de les observacions de la base de dades, i el test, el 30 % restant.

L'objectiu de la classificació és trobar un model (una funció) per predir la classe a la qual pertanyeria cada registre, aquesta assignació a una classe s'ha de fer amb la major precisió possible.

a. Mètode LDA

La idea de la tècnica "linear discriminant analysis" consisteix bàsicament en buscar una adreça de projecció de les dades (en aquest cas de dimensió 4), de manera que les dades projectades sobre la recta corresponent siguin el més separables possibles. Això ens permetrà addicionalment dibuixar les dades sobre els dos primers eixos de projecció per fer-nos una idea més exacta de com són els grups que estem determinant.

```
Prior probabilities of groups:
      Bueno Excepcional        Malo      Normal      pesimo
  0.1255587   0.2958147   0.1917920   0.1930110   0.1938236

Group means:
           Hotel_lat Hotel_lng Businesses_100m Businesses_1km Businesses_5km Stay_Duration Days_Since_Review
Bueno       49.45567  2.657771        53.78641       110.6731       105.7799      2.368932          359.2557
Excepcional 49.57914  3.054311        55.55495       109.3173       110.9794      2.342033          344.1607
Malo        49.21653  2.772865        54.12712       119.1038       102.0614      2.451271          346.0953
Normal      49.89564  2.636879        56.90105       111.6526       115.1474      2.261053          359.5474
pesimo      49.24240  2.792977        56.85535       111.8029       100.3166      2.477987          361.6771
           Total_Number_of_Reviews_Reviewer_Has_Given Average_Score Reviewer_Score
Bueno                                        7.184466      8.395793       8.305502
Excepcional                                  7.135989      8.379670       8.308654
Malo                                         7.574153      8.376483       8.285381
Normal                                       7.431579      8.406316       8.304842
pesimo                                       7.488470      8.413208       8.433333

Coefficients of linear discriminants:
                                                  LD1          LD2          LD3          LD4
Hotel_lat                                1.727272e-01 -0.048825504 -0.036588613  0.024161303
Hotel_lng                                3.677899e-02  0.077283073 -0.137381485  0.042173634
Businesses_100m                          5.415128e-03 -0.016753105 -0.013898179  0.020087605
Businesses_1km                          -2.845780e-03  0.001684058  0.005117082  0.007311932
Businesses_5km                           6.127511e-03  0.002712973  0.003516571  0.001010018
Stay_Duration                           -2.147316e-01 -0.054407356 -0.059306029  0.046686044
Days_Since_Review                        5.989859e-05 -0.002810609  0.000867231 -0.001451674
Total_Number_of_Reviews_Reviewer_Has_Given -4.596534e-03 -0.012754430  0.017808244  0.029564185
Average_Score                            1.312001e-01 -0.656624334  0.553752025  0.027466148
Reviewer_Score                          -9.922352e-02 -0.134203860 -0.262204640 -0.002250814

Proportion of trace:
   LD1    LD2    LD3    LD4
0.6069 0.2216 0.1232 0.0483
```

Figura 71: Resultado mètode LDA.

Com veiem, la funció ens retorna les probabilitats a priori dels grups, que les calcula utilitzant la proporció d'elements de cada classe. En segon lloc, calcula la mitjana de cada grup, que donaria la descripció mitjana de la resposta típica de cada grup. Els coeficients dels discriminants lineals s'usen per la funció per predir a quina classe pertany cadascuna de les categories.

Les últimes mesures donen una idea de la importància de cada eix discriminant. LD1 és el més gran, però no està molt proper a l'1, de manera que tot i ser aquest eix el més discriminant, les respostes no es poden classificar utilitzant només aquest eix.

La Matriu de Confusió conté informació sobre les prediccions realitzades per un Mètode o Sistema de Classificació, comparant per al conjunt d'individus de la taula d'aprenentatge o de testing, la predicció donada versus la classe a la qual aquests realment pertanyen. Per veure com de correctes

van ser les prediccions, fem una taula creuant les veritables classes amb les prediccions.

	Bueno	Excepcional	Malo	Normal	pesimo
Bueno	0	2	1	2	1
Excepcional	193	451	259	303	244
Malo	0	0	0	0	0
Normal	2	15	11	6	10
pesimo	0	0	0	0	0

Taula 10: Taula creuada

.

Podem veure que el classificador ha comès molts errors. Per a les classes dolent i pèssim no ha assignat cap dada, mentre que a la classe excepcional és on més dades ha assignat, molts de manera errònia. Per a aquest mètode el percentatge de dades ben classificats és només del 30%.

b. Mètode K veïns més propers

El mètode k-nn suposa que els veïns més propers ens donen la millor classificació i això es fa utilitzant tots els atributs; el problema d'aquesta suposició és que és possible que es tinguin molts atributs irrellevants que dominin sobre la classificació, de manera que els atributs rellevants perdrien pes entre d'altres irrellevants.

La millor elecció de k depèn fonamentalment de les dades; generalment, valors grans de k redueixen l'efecte de soroll en la classificació, però creen límits entre classes semblants. A través d'un procés d'optimització vam triar una K = 11 per realitzar aquest mètode.

La següent taula mostra la matriu de confusió per a un classificador de cinc classes:

knn.cross	Bueno	Excepcional	Malo	Normal	pesimo
Bueno	45	83	49	44	51
Excepcional	225	660	209	319	146
Malo	49	85	104	99	127
Normal	77	177	155	147	157
pesimo	39	66	132	87	168

Taula 11: Matriu de confusió

La categoria pitjor classificada és "Bueno" amb 45 prediccions ben classificades que representa un 16.54%, mentre que 227 no per al 83.45% restants.
La categoria millor classificada és "Excepcional", amb 660 prediccions ben classificades per a un 42.33%.
Per a les categories "Malo" i "Normal" les prediccions van ser ben classificades per al voltant d'un 20%, mentre que no per al 80% aproximadament.
Finalment, per a la categoria "Pésimo" 168 prediccions estan ben classificades per a un 34.14%.

El percentatge de ben classificades per aquest mètode és del 34.6%, més gran que amb el mètode anterior.

c. Màquines de vector de suport

La SVM, intuïtivament, és un model que, partint d'un conjunt d'exemples d'entrenament, podem etiquetar en diferents classes i representar aquestes mostres en punts en l'espai per tractar de separar les diferents classes mitjançant un espai el més ampli possible, perquè quan les noves mostres dels casos de test es posin en correspondència amb aquest model puguin ser classificades correctament en funció de la seva proximitat.

En aquest concepte de "separació òptima" és on resideix la característica fonamental de les SVM: aquest tipus d'algoritmes busquen el hiperplà que tingui la màxima distància (marge) amb els punts que estiguin més a prop d'ell mateix. Per això també a vegades se'ls coneix a les SVM com classificadors de marge màxim. D'aquesta manera, els punts del vector que són etiquetats amb una categoria estaran a un costat del hiperplà i els casos que es trobin en l'altra categoria estaran a l'altre costat.

```
Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  linear
       cost:  10
      gamma:  0.07692308

Number of Support Vectors:  119

 ( 7 31 36 27 18 )


Number of Classes:  5

Levels:
 Bueno Excepcional Malo Normal pesimo

2-fold cross-validation on training data:

Total Accuracy: 98.32
Single Accuracies:
 98.64 98
```

Figura 72: Summary del model lineal.

Hi ha 119 vectors suport. Usant validació creuada amb la mostra dividida en dues parts s'estima una probabilitat d'encert en la classificació d'aproximadament el 98.32%. Pel que el paràmetre de penalització indicat és molt bo ja que fa aquest valor molt proper al 100%.

Vam provar ara amb un SVM quadràtic en comptes de lineal:

```
Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  polynomial
       cost:  10
     degree:  2
      gamma:  1
     coef.0:  1

Number of Support Vectors:  331

 ( 50 79 85 28 89 )


Number of Classes:  5

Levels:
 Bueno Excepcional Malo Normal pesimo

10-fold cross-validation on training data:

Total Accuracy: 81.72
Single Accuracies:
 83 79.2 75.6 84 83 87.6 84.4 76 82.6 81.8
```

Figura 73: Summary del model quadràtic.

Veiem que per aquesta regla de classificació quadràtica tenim 331 vectors suport. Usant validació creuada en aquest cas obtenim una probabilitat d'encert en la classificació d'un 81.72%, notablement menor a l'obtinguda amb la lineal.

A causa de la dificultat per interpretar els resultats en haver 5 classes en la variable resposta, s'agrupen en 3 grups; Malo, Normal i Bueno. A partir d'aquí farem l'anàlisi tenint en compte això.

	Bueno	Malo	Normal
total	1984.0	1546.00	1470.00
%	0.4	0.31	0.29

Taula 12: Taula amb el percentatje per classe.

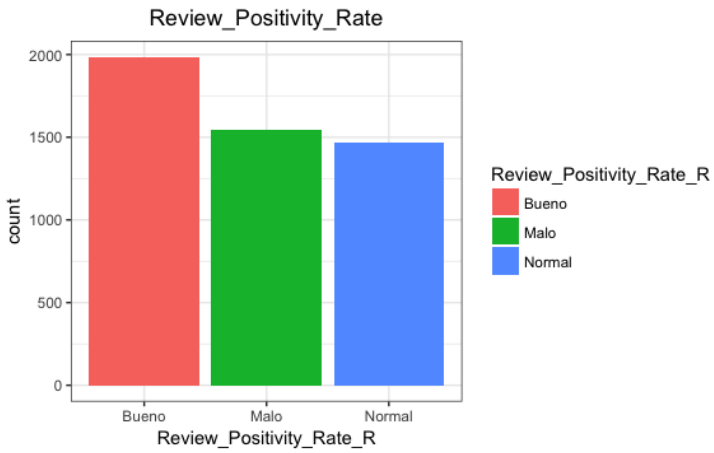


Figura 74: Histograma amb el percentatje per classes.

Perquè els models generats siguin útils, el percentatge d'encerts a que la classificació de les observacions ha de superar un nivell mínim, en aquest cas, el que s'obtindrà si la predicció de totes les observacions es correspongués amb la classe majoritària. La classe majoritària (moda) en aquest

cas és la classe "Bé" amb el 40% de les puntuacions. Aquest serà el nivell base a superar pel model (Aquest és el percentatge mínim d'encerts si sempre es va predir Bé). Abans de procedir a generar els models, dividim el data Set a un grup d'entrenament (per a l'ajust dels models) i un altre de test (per a l'avaluació dels mateixos). Aquesta divisió com especifiquem al començament d'aquest anàlisi serà d'un 70% -30%.

Necessitem fixar un marge de separació entre observacions a priori. Pel que avaluarem diferents valors del mateix mitjançant la validació creuada per escollir l'òptim.

```
Parameter tuning of 'svm':

- sampling method: 10-fold cross validation

- best parameters:
 cost
   50

- best performance: 0.003999186

- Detailed performance results:
   cost       error  dispersion
1 1e-03 0.108821327 0.019223826
2 1e-02 0.051694750 0.012125473
3 1e-01 0.018848189 0.008088238
4 1e+00 0.011992674 0.006270865
5 5e+00 0.010281644 0.003609046
6 1e+01 0.005711030 0.002686074
7 5e+01 0.003999186 0.001998250
```

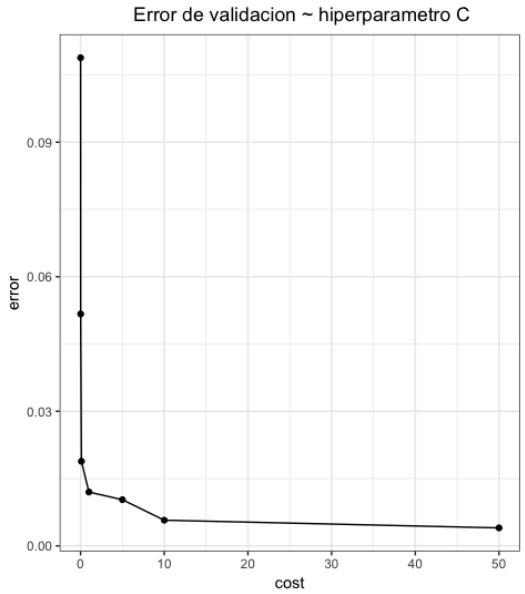


Figura 75: Validació creuada.

En vista dels resultats, tant gràfics com analítics vam concloure que el millor valor per a l'hiperparametre C és igual a 50 ja que resulta en el menor error de validació (0,00399).

```
Call:
best.tune(method = svm, train.x = Review_Positivity_Rate_R ~ Hotel_lat +
    Hotel_lng + Businesses_100m + Businesses_1km + Businesses_5km +
    Stay_Duration + Days_Since_Review + Total_Number_of_Reviews +
    Review_Positivity_Rate + Average_Score + Reviewer_Score + Additional_Number_of_Scoring +
    Total_Number_of_Reviews_Reviewer_Has_Given, data = datos.dd1.train,
    ranges = list(cost = c(0.001, 0.01, 0.1, 1, 5, 10, 50)), kernel = "linear",
    scale = TRUE)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  linear
       cost:  50
      gamma:  0.07692308

Number of Support Vectors:  105

 ( 21 50 34 )


Number of Classes:  3

Levels:
 Bueno Malo Normal
```

El nombre de vectors suport és de 105: 21 de la classe Bé, 50 de la classe Malo i 34 de la classe Normal. Podem obtenir els índexs de les observacions que es corresponen amb els vectors suport:

```
 375  755  880  978 1351 1578 1660 1932 1940 2040 2043 2102 2196 2482 2609 2688 2848
2872 3009 3387 3485  149  211  311  397  466  672  695  807  967 1227 1243 1299 1368
1442 1499 1511 1523 1534 1535 1619 1711 1749 1971 2003 2011 2041 2046 2123 2170 2197
2237 2283 2351 2384 2508 2532 2605 2633 2753 2757 2826 3068 3079 3097 3101 3150 3195
3210 3253 3419   40  299  476  594  740  846  937  942 1021 1027 1070 1071 1202 1209
1231 1298 1609 1742 2081 2091 2144 2150 2296 2363 2554 2566 2591 2642 2676
```

Figura 76: índex de les observacions que es corresponen amb els vectors suport.

Passarem ara, un cop acabat el millor model, a avaluar-lo. Per això en primer lloc veiem l'error d'entrenament i posteriorment el del test.

Error d'entrenament:

	real		
prediccion	Bueno	Malo	Normal
Bueno	1389	0	1
Malo	0	1081	1
Normal	0	2	1027

En aquest cas veiem que tan sols 4 observacions han estat mal classificades, dues de la classe Malo es van classificar com Normal i altres dues de la classe Normal es van classificar l'una com Bé i una altra com Malo. Això suposa tan sols un error del 0,11%.

Error de test:

	real		
prediccion	Bueno	Malo	Normal
Bueno	595	0	2
Malo	0	463	3
Normal	0	0	436

En aquest cas veiem que tan sols 5 observacions han estat mal classificades, les cinc que pertanyen a la classe Normal les quals 2 es van classificar erròniament com Bé i les tres restants com Malo. Això suposa tan sols un error de test del 0.33% .

Per tant, amb un valor de Cost = 50, el 0.33% de les observacions són incorrectament classificades. Tot i que el model és millor predient les observacions amb les que ha estat entrenat no es pot dir que l'error d'entrenament subestima l'error de test ja que la diferència és mínima.

Per dur a terme l'avaluació final es fa primer un entrenament del model amb un Kernel lineal i es busca optimitzar, en cas de ser necessari, el hiperparametre C. Aquest últim s'escull amb ajuda de la precisió utilitzant el valor més gran. Pel que resulta que el valor final d'aquest segueix sent 50.

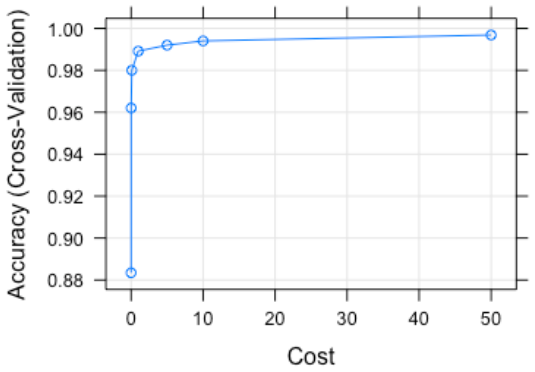
```
Support Vector Machines with Linear Kernel

3501 samples
  13 predictor
   3 classes: 'Bueno', 'Malo', 'Normal'

Pre-processing: centered (13), scaled (13)
Resampling: Cross-Validated (50 fold)
Summary of sample sizes: 3432, 3431, 3431, 3431, 3430, 3430, ...
Resampling results across tuning parameters:

  C      Accuracy   Kappa
  1e-03  0.8834387  0.8246693
  1e-02  0.9620483  0.9425344
  1e-01  0.9800344  0.9697560
  1e+00  0.9891658  0.9835985
  5e+00  0.9920195  0.9879119
  1e+01  0.9940279  0.9909545
  5e+01  0.9968527  0.9952322

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was C = 50.
```



Veiem gràcies a la matriu de confusió que la precisió és gairebé 1 pel que el model lineal és molt bo.

```
          Reference
Prediction Bueno Malo Normal
    Bueno   1389    0      1
    Malo       0 1081      1
    Normal     0    2   1027

Overall Statistics

               Accuracy : 0.9989
                 95% CI : (0.9971, 0.9997)
    No Information Rate : 0.3967
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9983
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Bueno Class: Malo Class: Normal
Sensitivity                1.0000      0.9982        0.9981
Specificity                0.9995      0.9996        0.9992
Pos Pred Value             0.9993      0.9991        0.9981
Neg Pred Value             1.0000      0.9992        0.9992
Prevalence                 0.3967      0.3093        0.2939
Detection Rate             0.3967      0.3088        0.2933
Detection Prevalence       0.3970      0.3091        0.2939
Balanced Accuracy          0.9998      0.9989        0.9986
```

Figura 77: Matriu de confusió.

De la mateixa manera que abans anem a especificar en aquest cas el grau del polinomi, per dur a terme ara el nostre propi nucli polinòmic i veure si millora o no davant del lineal.

```
Parameter tuning of 'svm':

- sampling method: 10-fold cross validation

- best parameters:
 cost degree
   10      3
```

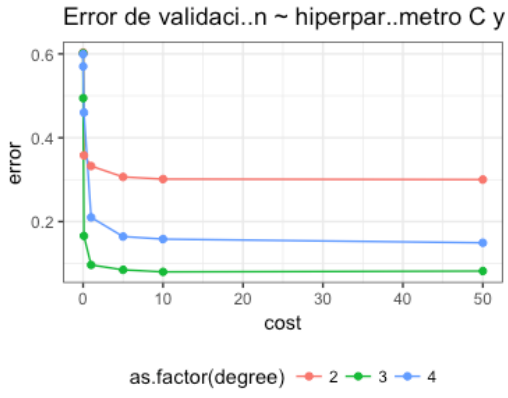


Figura 78: Validació creuada.

Obtenim així el grau igual a 3 i veiem com el valor del cost ha disminuït de 50 a 10.

El nombre de vectors suport és de 943: 201 de la classe “Bueno”, 487 de la classe “Malo” i 255 de la classe “Normal”. Podem obtenir els índexs de les observacions que es corresponen amb els vectors suport:

```
Call:
svm(formula = Review_Positivity_Rate_R ~ Hotel_lat + Hotel_lng +
    Businesses_100m + Businesses_1km + Businesses_5km + Stay_Duration +
    Days_Since_Review + Total_Number_of_Reviews + Review_Positivity_Rate +
    Average_Score + Reviewer_Score + Additional_Number_of_Scoring +
    Total_Number_of_Reviews_Reviewer_Has_Given, data = datos.dd1.train,
    kernel = "polynomial", cost = 10, degree = 3, scale = TRUE)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  polynomial
       cost:  10
     degree:  3
      gamma:  0.07692308
     coef.0:  0

Number of Support Vectors:  943

 ( 201 487 255 )


Number of Classes:  3

Levels:
 Bueno Malo Normal
```

Figura 79: Summary del model.

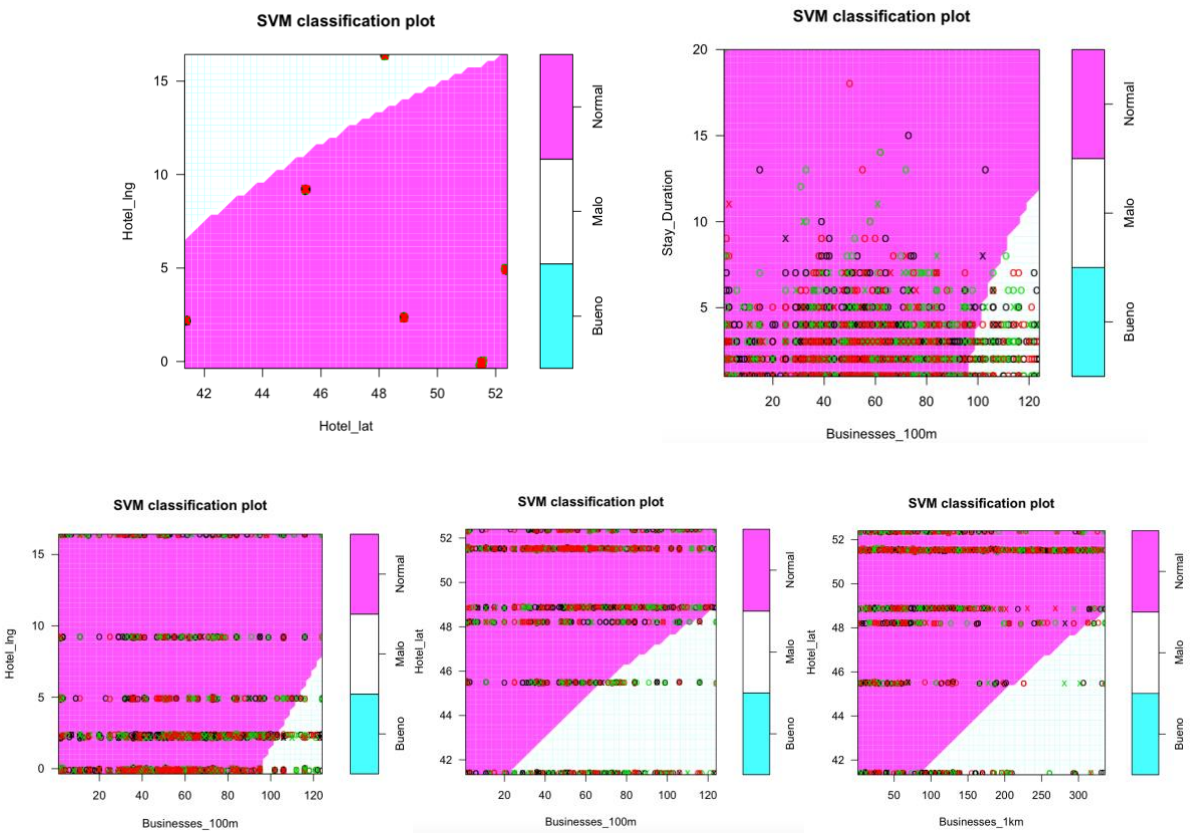
Passarem ara, un cop acabat el millor model, a avaluar-lo. Per això en primer lloc veiem l'error de test.

	real		
prediccion	Bueno	Malo	Normal
Bueno	1366	0	2
Malo	0	1040	12
Normal	23	43	1015

Figura 80: Taula amb l’error de test.

En aquest cas veiem que 80 observacions han estat mal classificades, suposant un error de test del 2.29%, més que per al nucli lineal.

No obstant això, El model ajustat amb un nucli polinòmic de grau 3 ha augmentat els vectors suport de 105 (nucli lineal) a 943.

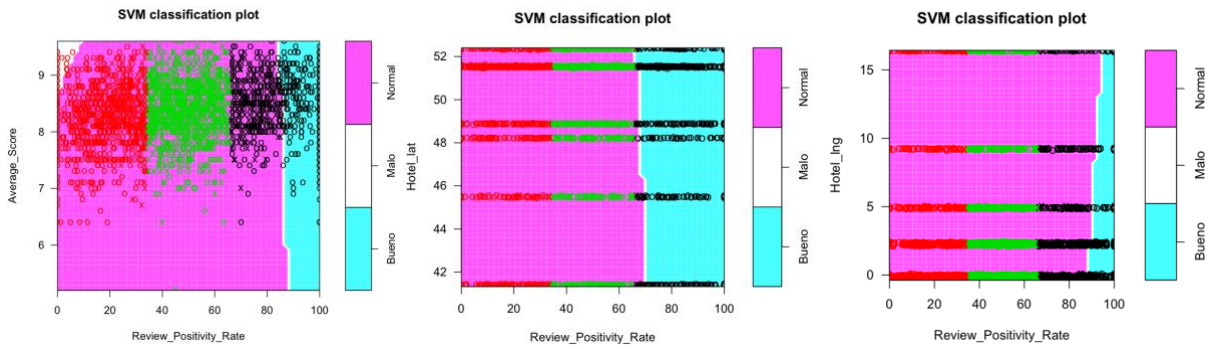


<u>Figura 81:</u> plots amb el mètode de vectors de suport.

Aquests plots representen els vectors de suport de dades. Els vectors de suport estan representats per les creus i també es poden veure els límits de decisió per a les diferents categories. Les regions de classe pronosticades són proporcionades pel color en el fons.

Vam tenir 943 vectors de suport, 201 pertanyen a la classe "Bueno", 487 a la classe “Malo” i, finalment, els 255 restants a la classe "Normal". Les creus vermelles, verdes i negres corresponen a les categories "Normal", "Malo" i "Bueno" respectivament.

Utilitzant les variables sobre la puntuació dels hotels com a eix en la gràfica obtenim:



<u>Figura 82:</u> plots amb el mètode de vectors de suport.

On veiem que la classe “Bueno”, que no apareix als gràfics anteriors, apareix quan Review_Positivity_Rate és elevada. En el segon gràfic veiem que quan la latitud de l'hotel és alta la categoria “Bueno” creix un poc. De forma contrària, veiem que a mesura que la longitud creix es torni més petita. Pel que es treu la conclusió que els hotels amb una latitud major i una longitud menor tenen una major puntuació. En els tres predomina la classe "Normal".

iv. Mètodes predictius.

Mètode ANN

El mètode ANN o Artificial Neural Network (d’aquí endavant ANN) consisteix en crear a partir de la base de dades un graf composat per nodes o neurones y diverses capes. En cada capa es realitzen uns càlculs i es calculen uns pèsos. El graf toma com a *input* les variables del model que es vol calcular i dona com a *output* el valor de la variable que es vol predir.

Per a realitzar la nostra ANN, degut als recursos informàtics disponibles hem realitzat una ANN amb dos capes ocultes amb 6 i 3 nodes o neurones, respectivament. També hem hagut de suprimir les variables categòriques per normalitzar el model i obtindré una millor predicció.

El graf resultant de la nostra ANN es el següent:

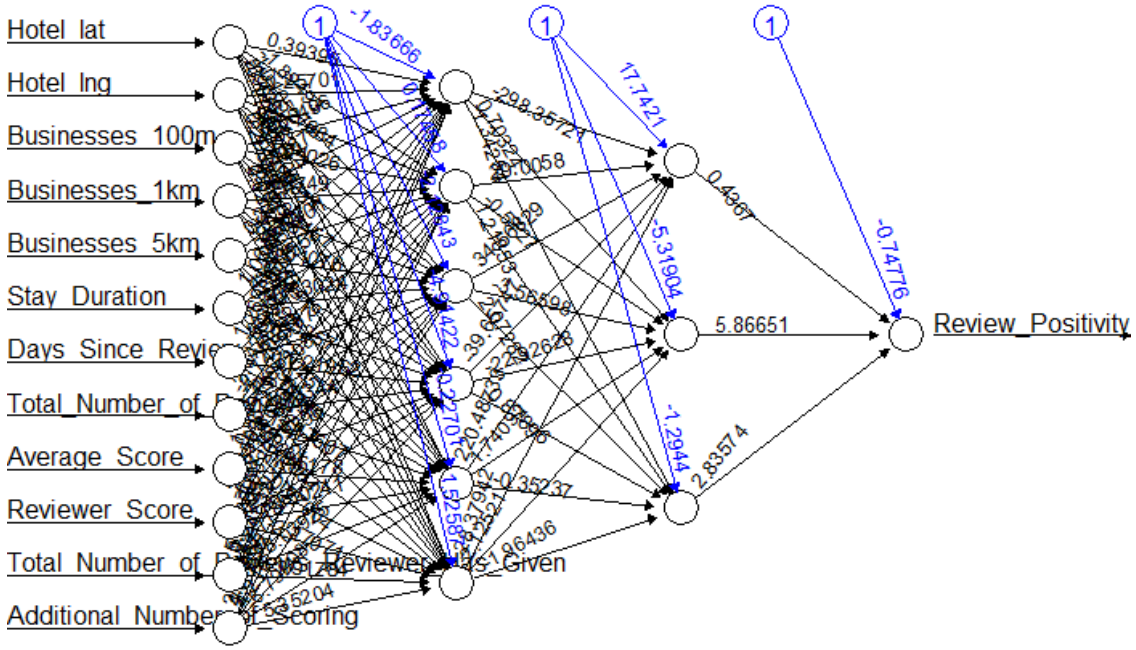


Figura 83: graf ANN

Com podem apreciar les 12 variables que entren al model es situen a l’esquerra. Formen l’*input* del model. Es calculen uns pesos y pasen a la següent capa de 6 nodes y continuen amb la 2na capa oculta (3nodes) y dona una predicció(4ta capa). Aquesta predicció es compara amb el valor de la variable resposta, en el nostre cas Review_Positivity_Rate, de la base de dades de entrenament i es recalculen els pesos per tal de millorar l’ajust. Aquest procés es fá de forma iterativa fins que s’arriba als pesos finals.

Una vegada calculat el model fem una sèrie de gràfics per validar el model. En primer lloc, calculem la suma dels errors al quadrat i el resultat es 798.1, veiem que es molt elevada, per tant, evidencia que a priori el model podria no ser vàlid. En segon lloc, pintem un gràfic on les variables d'entrada son el Review_positivity_rate.real, obtingut directament de les dades de la base de dades i el Review_Positivity_Rate.pred, obtingut de la predicció de la nostra ANN.

El gràfic es el següent.

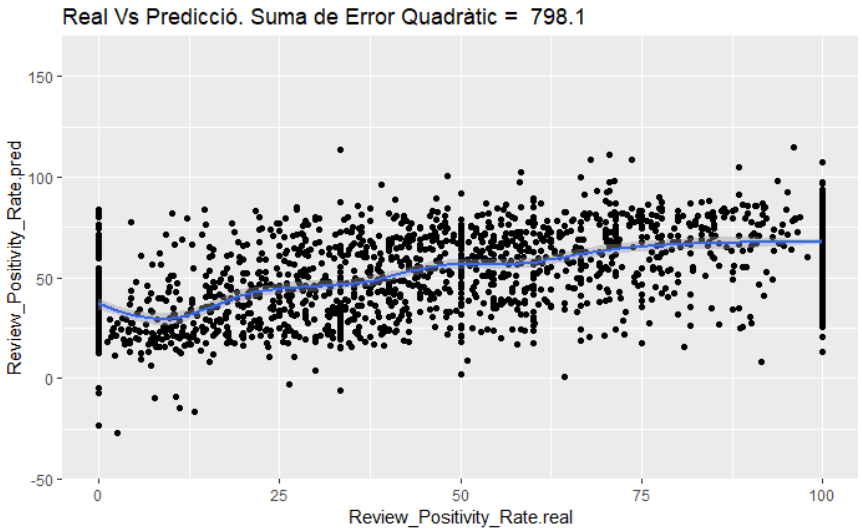


Figura 84: ANN Real vs Predicció

Com podem veure al gràfic hi ha una sèrie de valors que es repiteixen constantment, com per exemple el 0, el 100 o el 50. Això es degut a que la nostra variable resposta es una qualificació, llavors es normal trobar valors repetits als extrems i als valors més freqüents com podria ser el 50. D'altra banda veiem que el model s'ajusta el millor possible tenint en compte la variància real de la variable resposta, aproximadament deixa el 50% dels valors per sobre i per sota per a cada valor predit. Per veure-ho amb més detall grafiquem els errors.

El gràfic següent fà referència als errors estandarditzats del model ANN, com veiem estan centrats en 0 i té forma de recta horitzontal, el qual indica que no hi ha cap patró en els errors que es pugui deure a la omissió d'alguna variable o que, directament, pugui invalidar el model. El errors estan distribuïts aleatòriament respecte a l'eix vertical, el qual es símptoma de que el model es vàlid. Malgrat tot, hi ha erros que semblen bantant elevats, el qual no es cap sospresa tenint en compte que la suma d'errors quadràtic es elevada.

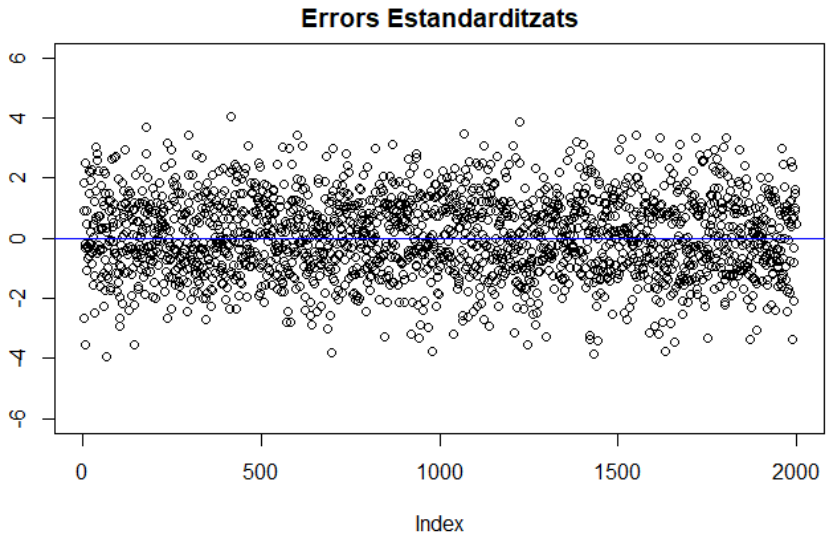


Figura 85: *ANN Errors Estandaritzats*

7. Anàlisi comparativa.

v. Mètodes de profiling.

S'han aplicat dos mètodes de clustering de la branca dels que tenen principis estadístics però que són totalment diferents. Aquests mètodes són el Jeràrquic i el K-means.

La primera cosa a tenir en compte a l'hora de comparar aquests dos mètodes és que en el Clustering Jeràrquic intervenen en el procés de creació de conglomerats tant les variables numèriques com les categòriques, en canvi, en el Clúster K-means només intervenen les variables numèriques. Tenim una base de dades amb 25 variables on 13 de les quals són numèriques i 12 categòriques. Per tant, a priori els resultats obtinguts amb el Clustering Jeràrquic seran millors que els obtinguts amb el Clúster K-means. Aquesta suposició clarament es compleix ja que si comparem les taules resum amb la informació descriptiva de cada clúster, per cada un dels mètodes veiem com els perfils de cada clúster estan molt millor descrits en el Clúster Jeràrquic que en el Clúster K-means.

En quant al criteri seguit, el Clúster Jeràrquic empra el **Mètode de Ward**, que consisteix en fer servir la pèrdua d'informació, mesurada per la **distància de Gower**, que es produeix al integrar els diferents individus en els clústers. Per altra banda, el criteri del K-means és la **minimització de la variació total intra clúster** definida com la suma de quadrats de les **distàncies euclidianes** entre els objectes i el centroide de cada clúster.

Per altra banda, en el Clúster K-means en primer lloc s'ha determinat el nombre de conglomerats i això s'ha fet amb el **Mètode del colze** (Elbow Method). En canvi, en el Clúster Jeràrquic s'escull el nombre de conglomerats posteriorment a l'execució d'aquest a partir del **dendrograma** i del percentatge de variabilitat entre grups explicada respecte el total.

Tanmateix, amb el Clúster K-means com només es treballa amb variables numèriques s'estandarditzen les dades perquè les diferents variables siguin comparables. Això no es fa amb el Clúster Jeràrquic per que usa tant variables numèriques com categòriques.

Finalment, fixant-se en els resultats, es té que amb el Clúster K-means s'han obtingut 4 conglomerats que es diferencien entre sí principalment per la puntuació, la situació i el nombre de ressenyes del hotel i de l'activitat a Booking del usuari que ha fet la ressenya. Mentre que amb el Clúster Jeràrquic s'han obtingut 5 conglomerats i es diferencien entre sí per les mateixes característiques que en l'anterior cas però també segons el tipus de viatger (sol, en parella o família, etc.).

vi. Mètodes associatius.

Els mètodes associatius que s'han dut a terme en els dos processos de mineria de dades han sigut l'Anàlisi de Components Principals (ACP, procés 1) i l'anàlisi per regles d'associació (Support, Confidence i Lift, procés 2). En primer lloc cal destacar que l'ACP té com a objectiu la identificació de patrons en les dades analitzant-les reduint la seva dimensionalitat i perdent la mínima informació possible. Per altra banda, els mètodes associatius fan una cosa semblant ja que permeten observar o descobrir patrons freqüents o correlacions entre els elements d'una base de dades.

Per tant es pot apreciar que els objectius de tots dos processos de mineria de dades són molt semblants. Tanmateix, però, aquests dos mètodes tenen diferències ja que l'ACP es duu a terme sobre les variables numèriques de la base de dades i els mètodes associatius, sobre les categòriques i els seus nivells. Un altre fet que els diferencia és que la base de dades que s'utilitza per als mètodes associatius és transaccional, per tant abans de començar amb aquests mètodes els integrants del subgrup que han realitzat el procés de mineria 2 han transformat la nostra base de dades en una base de dades transaccional.

Per finalitzar destacarem els resultats obtinguts d'ambdós mètodes. Mitjançant l'ACP, s'han extret aquelles variables que tenen més pes en cada una de les dimensions estudiades. Les variables relacionades amb la valoració de l'hotel tenen més influència en la dimensió 1, les que fan referència al nombre de ressenyes a les dimensions 2 i 4 i per últim es pot destacar que les variables longitud, latitud i nombre de negocis al voltant tenen un major pes a la tercera dimensió. Per altra banda, amb els mètodes associatius hem vist quines són els nivells de les variables categòriques més freqüents (hotels a Londres, Regne Unit) així com associacions freqüents de nivells de diferents variables que òbviament a simple vista no es poden observar.

Un cop analitzats tots dos mètodes no es podria dir quin és millor que l'altre perque un utilitza les variables numèriques (ACP) i l'altre les categòriques, i per tant amb un es trobaran patrons i associacions interessants que amb l'altre no es trobarien i el mateix a l'inrevés.

vii. Mètodes discriminants.

Els mètodes discriminants utilitzats han estat el mètode d'arbres de decisió, i el mètode de màquines de vectors de suport. Cal destacar que l'arbre de decisió es du a terme mitjançant el mètode "anova" que és l'apropiat en bases de dades en què la resposta variable és numèrica. A part d'això també cal dir que aquesta tècnica es du a terme sense les variables Hotel_Name ni Reviewer_Nationality perquè són variables categòriques amb molts nivells i si es considera la seva inclusió a la tècnica, l'arbre no surt bé. D'aquesta mateixa manera, en les màquines de suport de vector es pot destacar que hem realitzat amb un nucleó polinòmic de grau 3.

En el cas d'arbres de decisió, les conclusions finals són que a una major valoració global otorgada per l'usuari major serà la puntuació de l'hotel, així com que les característiques que més influeixen en la nota de l'hotel seran les valoracions dels usuaris.

Per al cas de les màquines de vectors de suport es treu la conclusió que els hotels amb una latitud major i una longitud menor tenen una major puntuació. De la mateixa manera es treu la conclusió que una major valoració flobal otorgada per l'usuari major serà la puntuació de l'hotel, igual que amb l'anterior mètode.

viii. Mètodes predictius.

Els mètodes predictius que s'ha dut a terme en aquest treball són el mètode de la regressió (subdivisió 1) i el mètode ANN (subdivisió 2).

El fet més determinant a l'hora de diferenciar aquests dos mètodes és que el mètode de la regressió utilitza tant les variables numèriques com les categòriques mentre que al mètode ANN només s'utilitzen les numèriques. Per tant, a priori es podria dir que amb la regressió es prediran millor les dades ja que s'aporta més informació que amb el mètode ANN a l'incloure també les variables categòriques.

Tanmateix, però, un cop construït el model de regressió ens trobem en una situació on és difícil validar-lo ja que a la vista dels gràfics dels residus les dades no segueixen una distribució normal i difícilment es podrien considerar lineals. Per altra banda, els gràfics dels residus del model implementat mitjançant el mètode ANN resulten ser millors que els anteriors ja que els residus d'aquest semblen estar distribuïts aleatòriament al voltant del zero i tenir una variància constant.

Cal destacar que per realitzar el mètode ANN s'han realitzat dues capes ocultes amb 6 i 3 nodes respectivament i que, encara que s'ha obtingut una suma de quadrats residual força alta, aquest mètode ens donarà unes millors prediccions que el de la regressió.

8. Conclusions generals.

Per finalitzar, exposem les conclusions que hem obtingut a partir dels diferents mètodes i tècniques d'anàlisi que hem fet servir al llarg de l'estudi. Cal recordar primer que el nostre objectiu principal era detectar característiques (tant del client com de l'hotel) associades a puntuacions elevades i que l'anàlisi ens ha aportat resultats interessants que ens permeten concloure que l'objectiu ha sigut assolit.

Pel que fa als mètodes de profiling, els resultats obtinguts amb el Clustering Jeràrquic han sigut millors. S'han pogut caracteritzar 5 perfils d'hotels diferents:

- En els hotels d'Àustria i Itàlia els viatgers solen ser parelles, passen una estància mitjana de dos dies i mitg i és habitual que deixin comentaris.
- Els hotels amb les millor puntuacions són els de Barcelona, i solen tenir pocs edificis al voltant, moltes habitacions luxoses i força comentaris positius.
- Els hotels amb les pitjors puntuacions estan situats a la ciutat de Londres, són hotels cèntrics i són els que més valoracions reben.
- Els hotels cèntrics són aquells que més freqüentat per parelles amb motius de vacances.
- Els viatges més curts es solen realitzar a París, on hi predominen els viatges de parelles.

Els mètodes associatius ens han permès establir relacions entre les diferents variables que disposem. Amb el mètode ACP s'ha pogut concloure que els hotels ubicats a França seran més urbans o cèntrics, mentre que en els hotels ubicats a Gran Bretanya els clients solen deixar més ressenyes o comentaris addicionals. D'altra banda, els algorismes de regles d'associació han indicat que les característiques més habituals dels viatgers són que viatgen sense que hi hagi vacances en el seu país ni en la ciutat en la que viatge, que viatgen per motius d'oci i que pugen la ressenya des del telèfon mòbil.

Quant als mètodes discriminants, aquests ens han fet veure que quan la ressenya positiva de l'usuari conté més paraules que la ressenya negativa, la nota de l'hotel serà més bona, així com també que els hotels amb una latitud alta i una longitud baixa tenen una major puntuació.

En darrer lloc, els mètodes predictius ens han ofert un model basat en les variables més rellevants d'entre totes les que hem estat treballant, és a dir, les variables que poden explicar millor el comportament de la variable resposta (grau de positivisme en les ressenyes).

9. Pla de treball real.

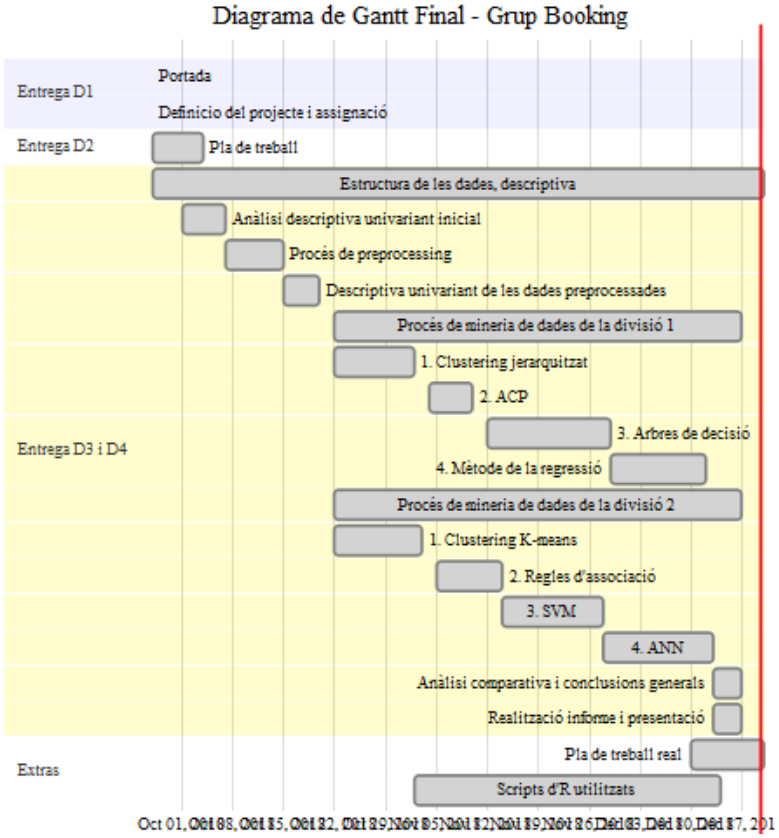


Figura 86: Diagrama de Gantt Final

				GRUP 1					GRUP 2			
				Guillem	Pablo	Antoni	Carles	Aleix	Laura	Víctor	Marta	Sofía
Entrega D1			Portada								X	X
			Definició del projecte i assignació		X	X	X					
Entrega D2	Estructura de les dades, descriptiva i entrega final		Pla de treball	X				X	X	X		
Entrega D3 iD4			Estructura de les dades, descriptiva i entrega final	X	X						X	X
		Procés de mineria de dades	Clustering jerarquitzat	X		X		X				
			ACP				X					
			Arbres de decisió		X			X				
			Mètode de la regressió			X		X				
			Clustering K-means							X	X	
			Regles d'associació						X			X
			SVM								X	X
			ANN						X	X		
			Anàlisi comparativa i conclusions generals	X	X	X	X	X	X	X	X	X
			Realització informe i presentació	X	X	X	X	X	X	X	X	X
Extres			Pla de treball real	X				X	X	X		
			Scripts d'R utilitzats	X	X	X	X	X	X	X	X	X

Figura 87: Plantilla de Tasques Final.