

# Exemple d'aplicació de tècniques de mineria de dades a la indústria hotelera \*

**Hugo Allès Pons** Universitat de Barcelona  
**Carles Blanco Conde** Universitat de Barcelona  
**Aleix Fibla Salgado** Universitat de Barcelona  
**Victor Miranda Hernández** Universitat de Barcelona  
**Pablo Morante López** Universitat de Barcelona  
**Antoni Ramoneda Montoya** Universitat de Barcelona  
**Oriol Rovira Tauler** Universitat de Barcelona  
**Aleix Salvador Barrera** Universitat de Barcelona

---

Aquest document mostra un exemple d'aplicació de tècniques de mineria de dades i anàlisi multivariant dins del marc de la indústria hotelera, amb l'objectiu de millorar les experiències dels viatgers i optimitzar el targeting dels hotels

*Keywords:* clúster, preprocessament, profiling, ACP, ACM

---

## Contents

<b>Introducció</b>	<b>3</b>
<b>Pla de treball</b>	<b>5</b>
Diagrama de Gantt . . . . .	5
Distribució de tasques . . . . .	5
Pla de riscos . . . . .	5
<b>Estructura de les dades i anàlisi descriptiva</b>	<b>7</b>
Motivació del treball . . . . .	7
Descripció formal de l'estructura de les dades . . . . .	7
Anàlisi exploratori inicial de les variables . . . . .	11
Procés de preprocessament . . . . .	24
Anàlisi descriptiva univariant post preprocessament . . . . .	29
<b>Clúster jeràquic</b>	<b>34</b>
<b>Profiling dels clústers</b>	<b>37</b>
<b>ACP de les variables numèriques</b>	<b>54</b>
Gràfics d'individus . . . . .	56
Gràfics de variables . . . . .	58
Biplots mixtes de variables i individus . . . . .	65
<b>ACM de les variables qualitatives</b>	<b>68</b>

\*Tots els arxius per a replicar l'anàlisi es troben al compte de Github <https://github.com/aleixfiblasalgado>

<b>Clustering jeràrquic sobre les components factorials retingudes a l'ACP i a l'ACM</b>	<b>75</b>
<b>Profiling del Clúster Jeràrquic sobre ACP</b>	<b>92</b>
<b>Anàlisi Textual</b>	<b>104</b>
<b>Anàlisi Comparativa dels diferents mètodes i conclusions generals</b>	<b>108</b>
<b>Pla de treball real</b>	<b>109</b>
Diagrama de Gantt . . . . .	109
Distribució de tasques . . . . .	110

## Introducció

Aquest treball s'ha desenvolupat amb l'objectiu de millorar les experiències de viatges en l'àmbit de la indústria hotelera. Per a assolir aquest objectiu, s'empren una sèrie de tècniques de mineria de dades dins el marc de l'anàlisi multivariant.

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 otorgades 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.

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 aleatoriament 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.

Table 1: Taula de valors missing

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 %

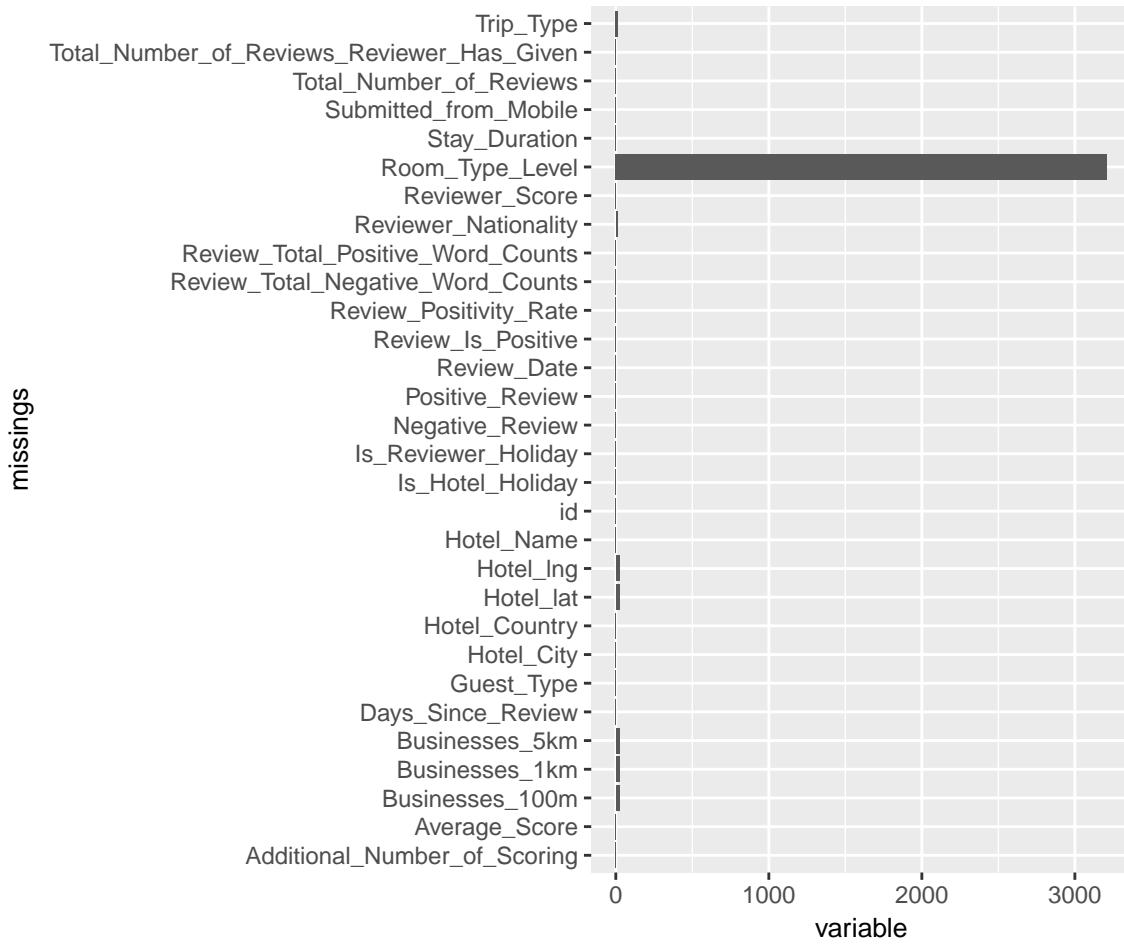


Figure 1: Histograma dels valors missing

## Pla de treball

En aquest capítol es resumeix la organització del treball i les tasques que es persegueix implementar, a priori. Un cop finalitzat el treball es presentarà un appendix on cada component del grup valorarà si el treball s'ha dut a terme d'acord amb el pla de treball establert en aquesta secció.

### Diagrama de Gantt

El Diagrama de Gant és una eina que es fa servir per a descompondre el projecte en tasques “*indivisibles*” i sequenciar-les temporalment. La *Figura 2* mostra el Diagrama de Gant per a aquest projecte.

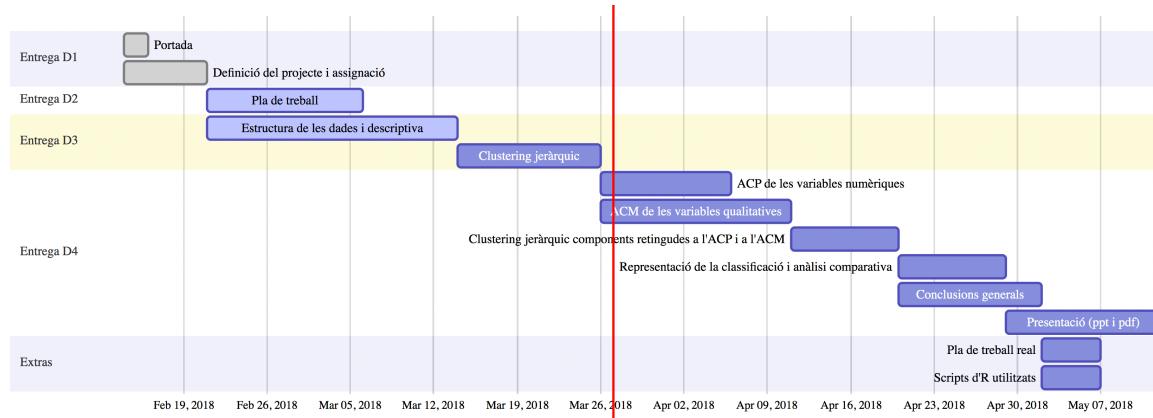


Figure 2: Diagrama de Gantt

### Distribució de tasques

A la *Figura 3* es resumeix la distribució de les diferents tasques presentades anteriorment al Diagrama de Gant.

### Pla de riscos

Finalment, cal elaborar un pla on es consideren riscos potencials que puguin presentar-se al llarg del projecte, així com un llistat de possibles solucions per als mateixos (*Figura 4*).

	Hugo Allès Pons	Carles Blanco Conde	Aleix Fibla Salgado	Victor Miranda Hernández	Pablo Morante López	Antoni Ramoneda Montoya	Oriol Rovira Tauler	Aleix Salvador Barrera
<b>D1</b>	Portada	x	X				x	
	Definició del projecte	x	x	X	x	x	x	x
<b>D2</b>	Pla de treball		x	x	x			X
	Estructura i descriptiva de les dades	x			x		x	x
<b>D3</b>	Clúster jeràrquic			x	x	X		x
	ACP de les variables numèriques		x		x	x	X	
	ACM de les variables qualitatives	X		x			x	x
<b>D4</b>	Clustering jeràrquic sobre les components factorials retingudes a l'ACP i a l'ACM	x		X		x	x	
	Representació de la classificació i anàlisi comparativa	x		x	x		x	x
	Conclusions	x	x	x	x	x	x	X
	Pla de treball REAL	x	x	x	X	x	x	x
	Scripts d'R utilitzats	x	x	x	X	x	x	x

Figure 3: Plantilla de Tasques

	Possible Risc	Com prevenir-lo	Com gestionar-lo
	<i>Un membre del grup es posa malalt i no pot fer la seva part de la tasca</i>	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.
	<i>Pèrdua del treball o parts d'aquest</i>	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.
	<i>Entregues o scripts mal realitzats</i>	Per a cada tasca o entrega hi haurà un grup de 2 persones que s'encarreguen de revisar la feina realitzada.	En el repartiment de tasques també sortiran en cada entrega el grup que s'encarregará de la revisió de la feina.
	<i>No entregar a temps les tasques</i>	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.
	<i>Distribució del treball inadequadament</i>	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.
	<i>Discussions internes per diferències conceptuais</i>	Les decisions finals sempre seran fetes amb un consens del grup.	En cas d'haver una discussió els dos membres podran exposar les seves idees i s'escollirà en majoria mitjançant una votació.
	<i>Desobediència d'algun membre del grup en les indicacions grupals i actitud individualista</i>	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 al llarg del treball. Aquest s'encarregará de tindre la última paraula en termes de repartiment de tasques i si fa falta parlar amb el professor, que normalment serà un dels que li toqui fer la feina de revisió.
	<i>Errades ortogràfiques a l'informe</i>	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.

Figure 4: Pla de riscos

## Estructura de les dades i anàlisi descriptiva

Aquest capítol recull tota la informació necessària per a que el lector es familiaritzi amb les dades. En primer lloc, es planteja la motivació del treball, seguida de les metadades que acompanyen a la base de dades. Finalment, es plantegen tots els procediments que s'han dut a terme en la fase de preprocessament de les dades, així com un anàlisi exploratori inicial de cadascuna de les variables.

### *Motivació del treball*

El debat per escollir el tema del treball no ha estat una discussió difícil, ja que a tots els integrants del grup ens agrada viatjar, observar noves cultures i estils de vida; alhora que, estar en bona companyia i viure experiències enriquidores. Així doncs, vam decidir, aplicar les tècniques d'anàlisi de dades que estem aprenent en el present curs per tal de fer un estudi sobre una indústria amb molt de pes en el nostre territori, com és la indústria hotelera, tot i que no només ens hem centrat en un àmbit local. 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 detectar característiques (tant del client com de l'hotel) que estiguin associades a una puntuació elevada, o per contra, molt baixa. Però, no només volem quedar-nos aquí, sinó que la nostra intenció és la de millorar l'experiència dels clients que utilitzen els portals web especialitzats, a través de la conglomeració d'hotels a partir de les ressenyes, i les preferències dels hostes, per tal d'ofrir un millor servei.

### *Descripció formal de l'estructura de les dades*

Tal i com s'ha mencionat anteriorment, tot l'anàlisi es duu a terme amb una mostra aleatòria de 5.000 observacions  $\{seed(20359724)\}$  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.

---

## Diccionari de metadades

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
6. *Hotel\_lng* (numeric): Longitud de l'hotel.
  - a. **rang:** {-0.3697581 , 16.4219737}
  - 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 , 254}
  - b. **rol:** variable explicativa
  - c. **missing\_code:** NULL
8. *Businesses\_1km* (integer): Nombre de negocis a 1 km a la rodona de l'hotel.
  - a. **rang:** {5 , 5500}
  - b. **rol:** variable explicativa
  - c. **missing\_code:** NULL
9. *Businesses\_5km* (integer): Nombre de negocis a 5 km a la rodona de l'hotel.
  - a. **rang:** {408 , 30300}
  - 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 trip , 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 , 730}
  - 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.
- modalitats:** {"", Abkhazia Georgia , Albania , Andorra , Angola , Argentina , Armenia , Australia , Austria , Azerbaijan , Bahrain , Bangladesh ,...} (123 modalitats)
  - rol:** variable explicativa
  - missing\_code:** ""
22. *Negative\_Review* (text): Ressenya negativa escrita per l'usuari.
- rol:** text
  - missing\_code:** "Na"
23. *Review\_Total\_Negative\_Word\_Counts* (integer): Nombre total de paraules a la ressenya negativa escrita per l'usuari.
- rang:** {0 , 372}
  - rol:** variable explicativa
24. *Positive\_Review* (text): Ressenya positiva escrita per l'usuari.
- rol:** text
  - missing\_code:** "Na"
25. *Review\_Total\_Positive\_Word\_Counts* (integer): Nombre total de paraules a la ressenya positiva escrita per l'usuari.
- range:** {0 , 247}
  - rol:** variable explicativa
26. *Average\_Score* (numeric): Valoració mitjana de l'hotel a la pàgina de Booking a data 31 de desembre de 2016.
- range:** {5.2 , 9.6}
  - rol:** variable explicativa
27. *Reviewer\_Score* (numeric): Valoració global otorgada per l'usuari que ha escrit la ressenya.
- range:** {2.5 , 10}
  - rol:** variable explicativa
28. *Total\_Number\_of\_Reviews\_Reviewer\_Has\_Given* (integer): Nombre total de ressenyes a la web de Booking escriptes per l'usuari.
- range:** {1 , 156}
  - rol:** variable explicativa
29. *Additional\_Number\_of\_Scoring* (integer): Nombre total de valoracions adicionals vàlides sobre diferents aspectes de l'hotel.
- range:** {8 , 2682}
  - rol:** variable explicativa
30. *Submitted\_from\_Mobile* (qualitative): Variable binària que indica si la ressenya s'ha pujat a Booking via telèfon mòvil.
- modalitats:** {0: No , 1: Yes}
  - rol:** variable explicativa
-

## Anàlisi exploratori inicial de les variables

Un cop tenim ben definida la base de dades, així com el diccionari de metadades que l'acompanya, convé conduir 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. Aquesta aproximació inicial, també ens permetrà detectar anomalies, que posteriorment corregirem a la fase de preprocessament.

Tot i que majoritàriament farem servir tècniques d'anàlisi univariant, hi ha un ampli ventall d'eines que es poden fer servir de manera complementària. Més que un procediment formal, l'anàlisi exploratori de dades és una aproximació a l'anàlisi de dades que posposa les suposicions habituals sobre quin tipus de model de dades tenim, amb una aproximació molt més directe a aquestes, que revela la seva estructura i model subjacent. Això va molt més enllà que una mera col·lecció de tècniques; EDA (exploratory data analysis) és una filosofia sobre com diseccionem un conjunt de dades, el que busquem, com les veiem i com les interpretarem. En aquest apartat, pretenem aconseguir aquest objectiu a partir de gràfics estadístics univarians de les nostres variables.

Per a no perdre el fil, proposem seguir l'ordenació que hem especificat al diccionari de dades. 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

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

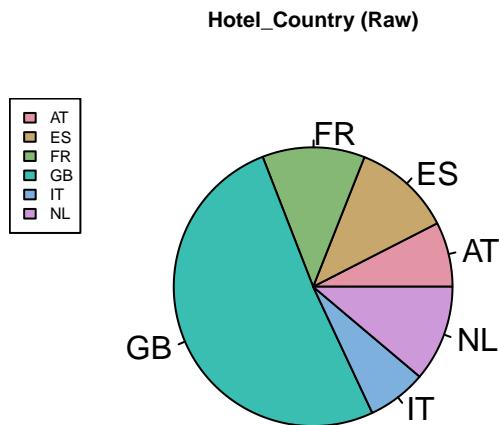


Figure 5: Pie Chart Hotel\_Country (Raw)

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	Barcelona
545	13	570
Boulogne Billancourt	Donauinsel	El Prat de Llobregat

3	6	4
Fitzrovia	London	Milan
23	2378	343
Paddington	Paris	Paris 06
150	587	1
Paris 12	Vienna	Vincennes
1	370	3
Woodford Green		
3		

**Hotel\_City (Raw)**

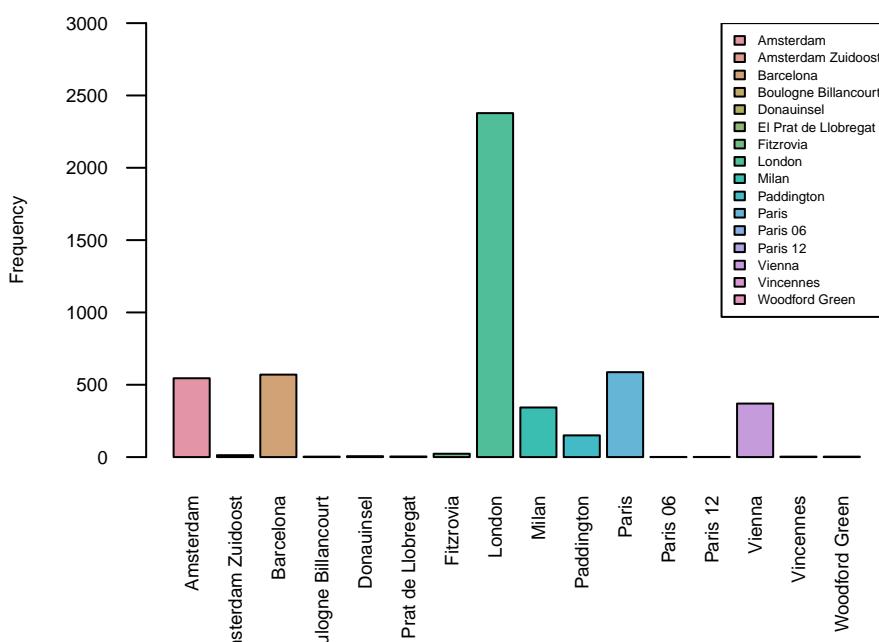


Figure 6: Bar plot Hotel\_City (Raw)

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. Considerarem fer un resum numèric per a totes dues variables, acompanyat de dos boxplots (*Figure 7*).

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

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
-0.36976	-0.14387	-0.00025	2.81000	4.83110	16.42197	23

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

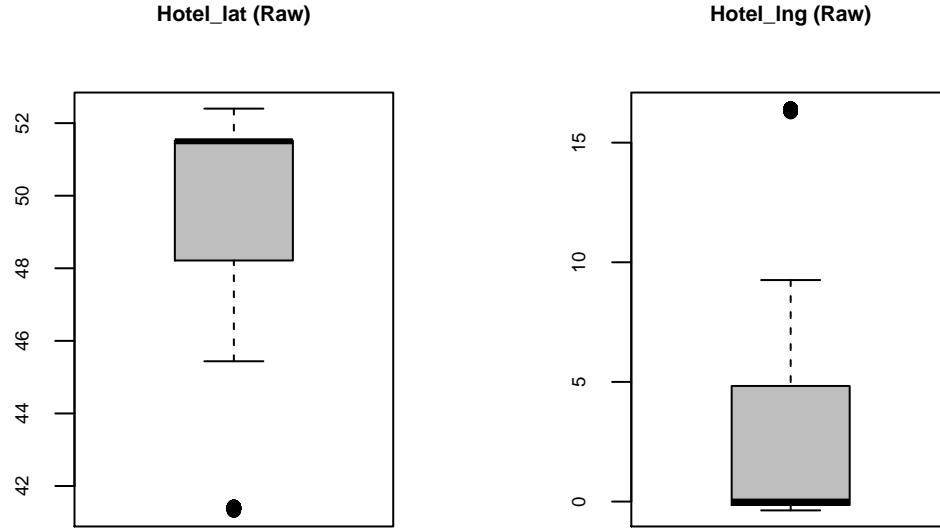


Figure 7: Boxplot Mesures de localització (Raw)

Pel que fa a la latitud on es troben els hotels, observem que la majoria es troben al voltant de  $51^{\circ}\text{N}$ , 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  $41^{\circ}$  i  $52^{\circ}\text{ N}$ , els quals coincideixen, obviament amb les latituds de la majoria dels països europeus.

Complementàriament, per conèixer la localització dels hotels també necessitem la longitud, que és una línia imàginaria que va de pol a pol. Aquí podem observar com el gran gruix dels hotels es troben al voltant del  $0^{\circ}$ , 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 conseqüència, 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 desprenden 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	City
1	18	58	3	9
Classic	Deluxe	Duplex	Executive	Family
343	191	10	83	131
Luxury	NULL	Premium	Privilege	Standard
9	3209	1	8	546
Studio	Suite	Superior		
23	71	286		

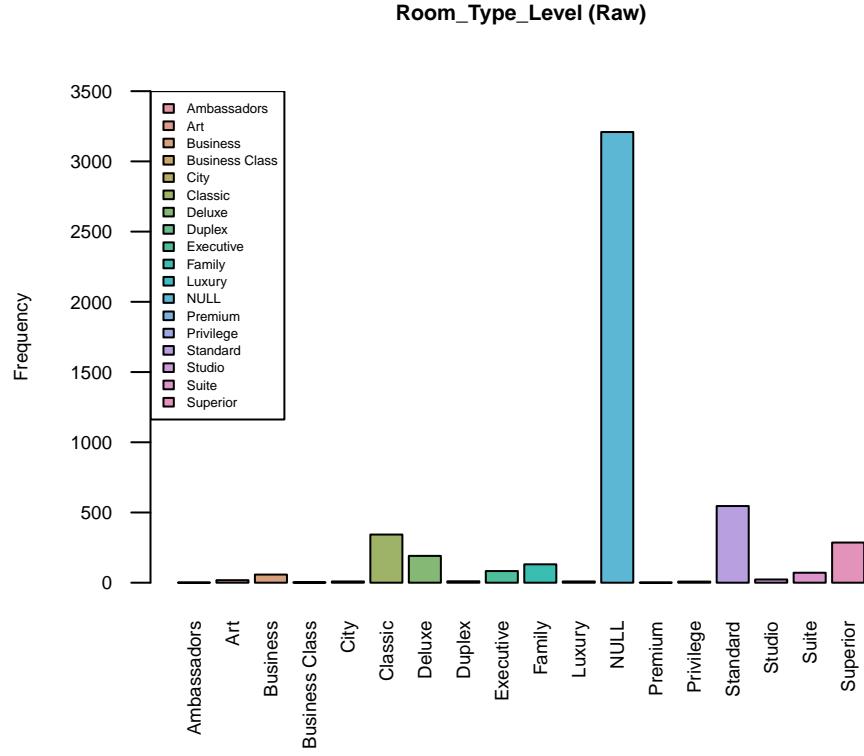


Figure 8: Barplot Room\_Type\_Level (Raw)

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 reduir-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).

Couple Family with older children	
2425	246
Family with young children	Group
569	658
Solo traveler	Travelers with friends
1077	21
With a pet	
4	
Business trip	Couple
803	56
Family with older children	Family with young children
11	31
Leisure trip	NULL

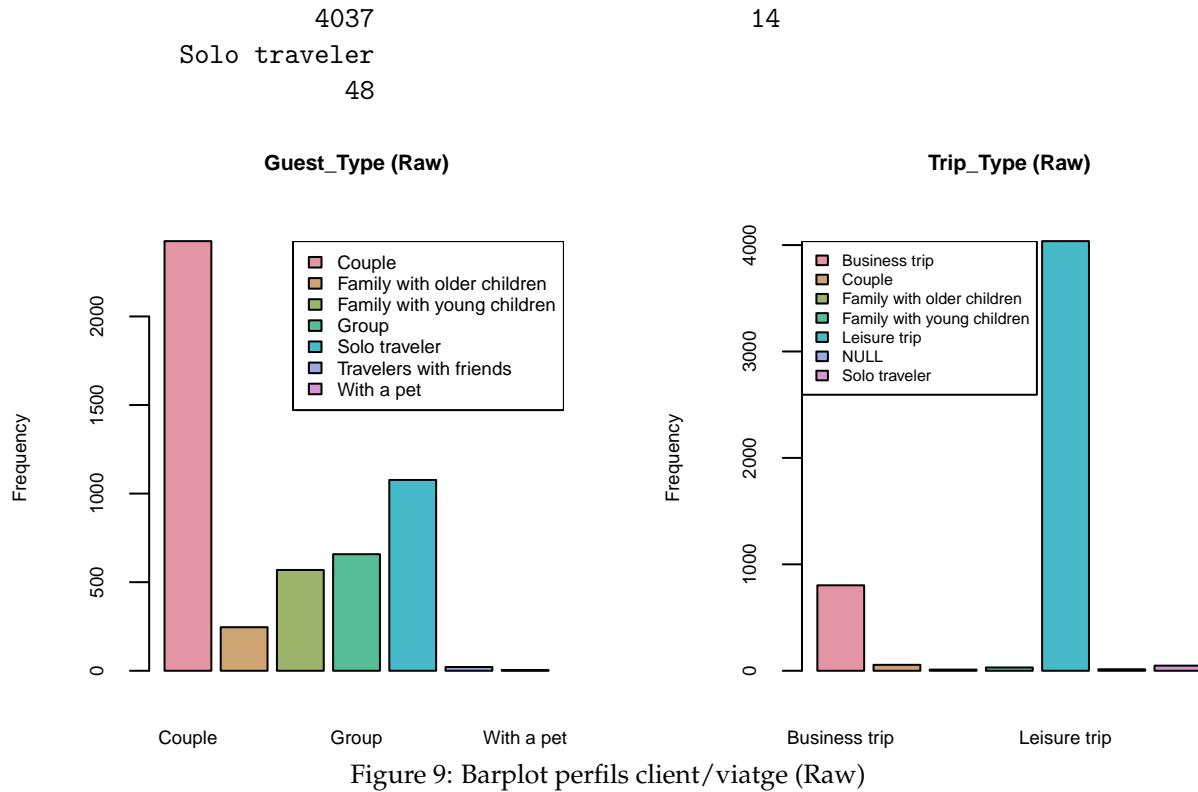


Figure 9: Barplot perfiles client/viatge (Raw)

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

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

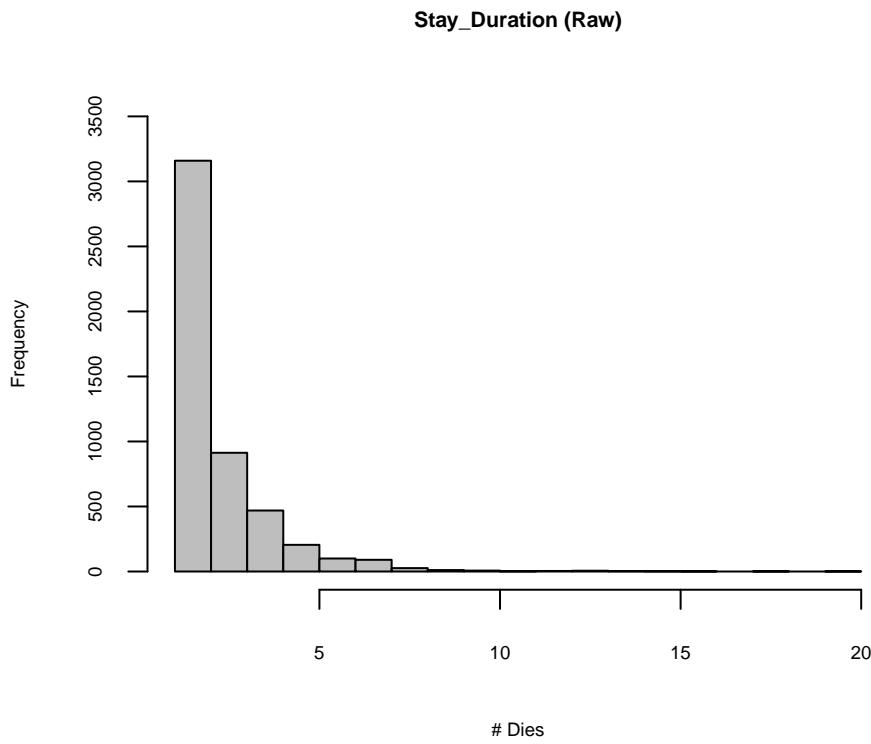


Figure 10: Histograma Stay\_Duration (Raw)

igualment construir la proporció de 1 (Sí) que tenim sobre el total mitjançant un pie chart (*Figure 11*).

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 variable *Is\_Reviewer\_Holiday* que coincideixen.

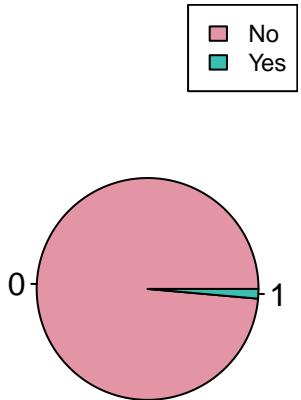
Seguidament, analitzem la variable *Total\_Number\_of\_Reviews*. Aquesta apareix codificada correctament com a numèrica així que construïm un histograma (*Figure 12*) i el complementem amb un resum numèric, com en els casos anteriors.

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

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*).

**Is\_Hotel\_Holiday (Raw)**



**Is\_Reviewer\_Holiday (Raw)**

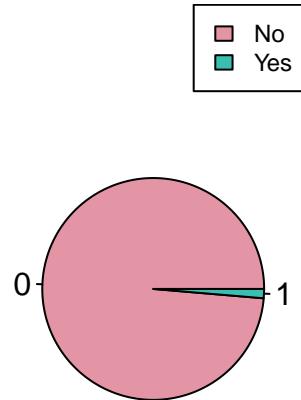


Figure 11: Pie Chart Festivitats (Raw)

**Total\_Number\_of\_Reviews (Raw)**

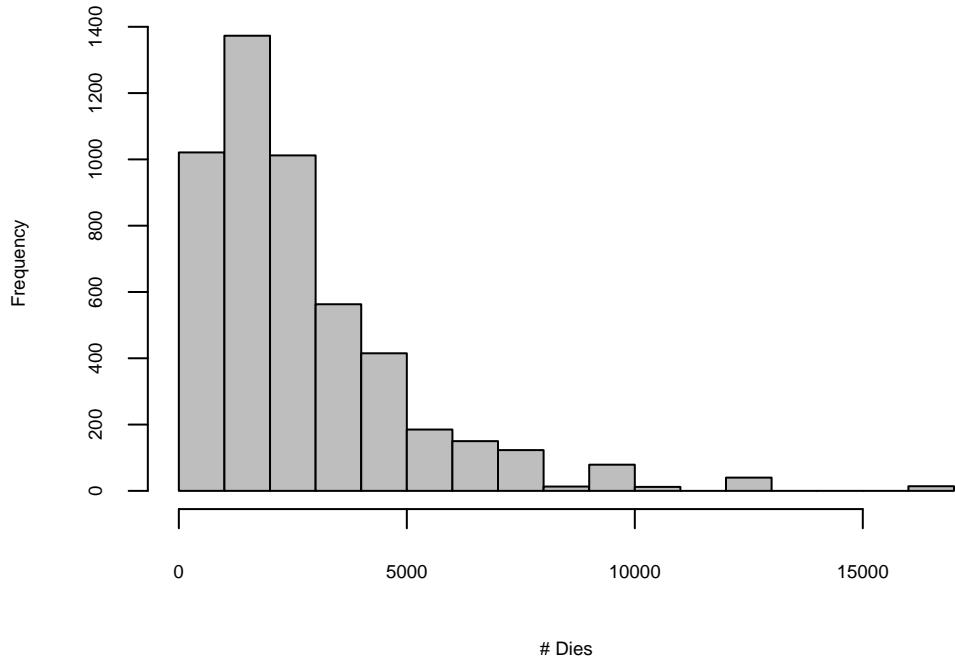


Figure 12: Histograma Total\_Number\_of\_Reviews (Raw)

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

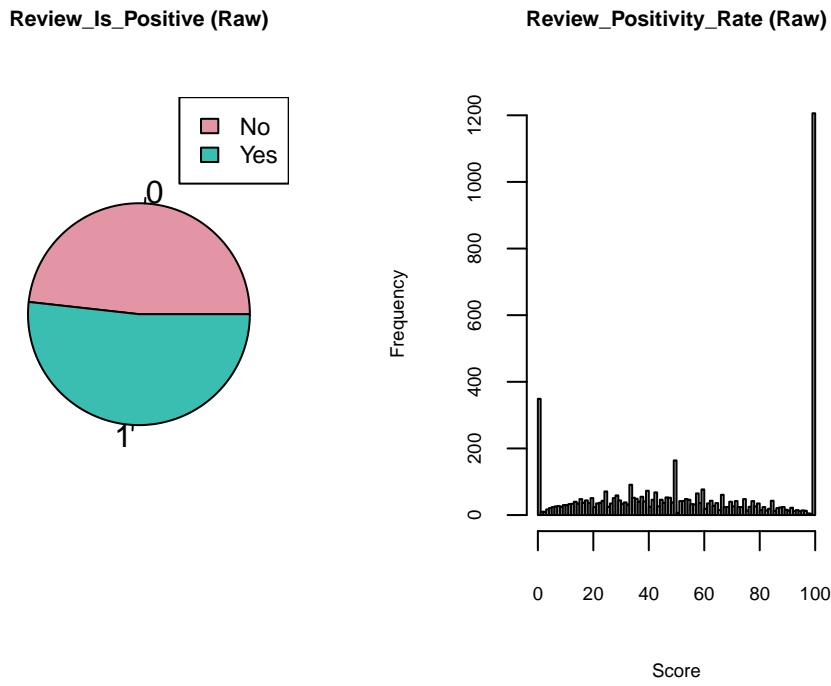


Figure 13: Gràfics variables positivisme de la ressenya (Raw)

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 empata.

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.

United Kingdom	United States of America
2321	365
Australia	Ireland
209	134
United Arab Emirates	Netherlands
123	89
Switzerland	Germany
88	81
Canada	Saudi Arabia
74	74
Belgium	France

	69		66
Israel	64	Italy	59
Kuwait	48	Spain	48
Turkey	45	New Zealand	43
Greece	41	Sweden	41
South Africa	39	Romania	37
Russia	37	Singapore	33
Poland	32	India	30
Qatar	30	Austria	28
China	28	Portugal	26
Egypt	25	Finland	25
Lebanon	24	Czech Republic	21
Norway	21	Denmark	20
Brazil	19	Hong Kong	19
Hungary	19	Cyprus	18
Thailand	18	Bahrain	16
Oman	16	Croatia	15
Malaysia	15	Malta	14
Japan	13	Serbia	13
Luxembourg	12	Bulgaria	11
Iran	11	Slovakia	11
	10	Indonesia	
Pakistan	10	South Korea	10
Iceland	9	Lithuania	8
Nigeria		Latvia	

	8		7
Philippines	7	Slovenia	7
Isle of Man	6	Taiwan	6
Azerbaijan	5	Estonia	5
Guernsey	5	Kenya	5
Sri Lanka	5	Ukraine	5
Gibraltar	4	Jersey	4
Jordan	4	Mauritius	4
Mexico	4	Albania	3
Argentina	3	Bermuda	3
Colombia	3	Peru	3
Zambia	3	Angola	2
Armenia	2	Bangladesh	2
Barbados	2	Chile	2
Iraq	2	Kosovo	2
Macedonia	2	Moldova	2
Monaco	2	Morocco	2
Puerto Rico	2	Tunisia	2
Uganda	2	Abkhazia Georgia	1
Andorra	1	Bosnia and Herzegovina	1
Botswana	1	(Other)	24

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 inter-

essant fer un boxplot conjunt de les dues per a comparar-les (*Figure 14*). A priori, haurien de ser complementàries, aquelles ressenyes amb major nombre de paraules als comentaris positius haurien 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).

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

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

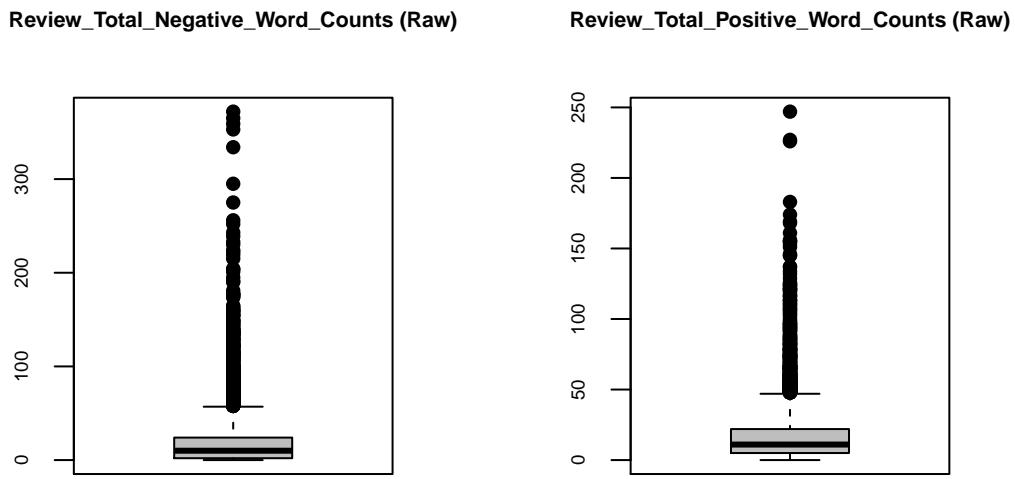


Figure 14: Boxplot Paraules als comentaris negatius/positius (Raw)

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.

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

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*).

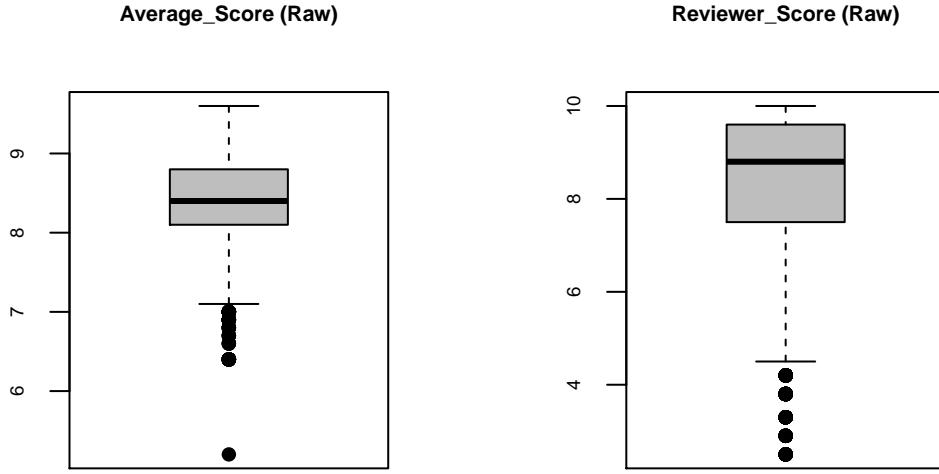


Figure 15: Boxplot Puntuacions (Raw)

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	3.000	7.323	9.000	156.000

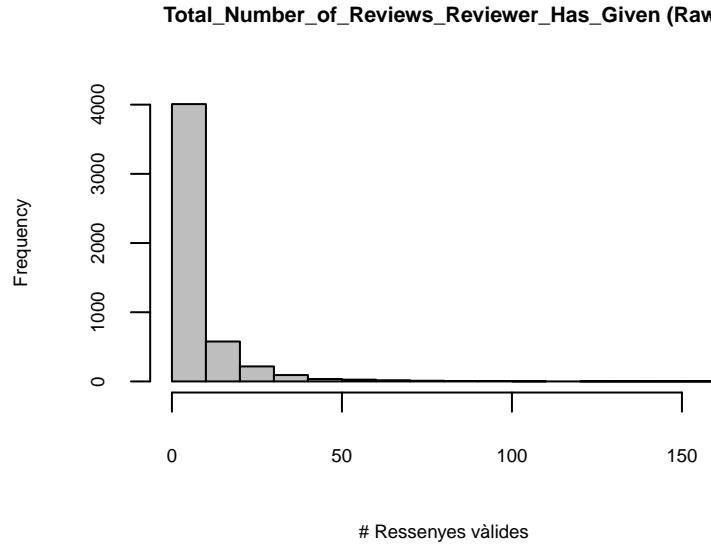


Figure 16: Histograma Total\_Number\_of\_Reviews\_Reviewer\_Has\_Given (Raw)

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 enquestat 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

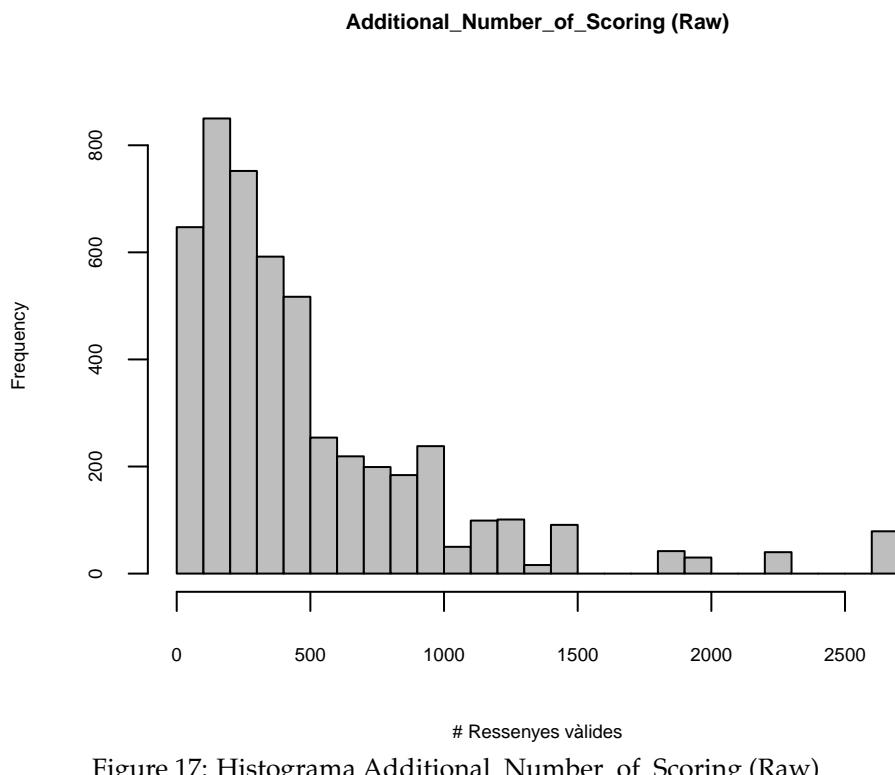


Figure 17: Histograma Additional\_Number\_of\_Scoring (Raw)

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 exploratori 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 variables binàries i construir un pie chart (*Figure 18*).

Observem com predominen les ressenyes escrites des del telèfon mòbil.

**Submitted\_from\_Mobile (Raw)**

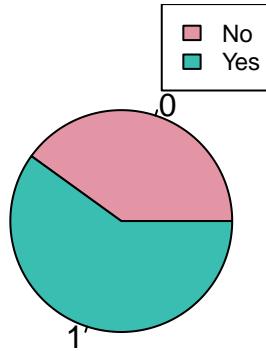


Figure 18: Pie Chart Submitted\_from\_Mobile (Raw)

### 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 textuales *Positive\_Review* i *Negative\_Review* no s'ajusten exactament 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 textuales). Com a filtre adicional, considerem tractar únicament aquells factors amb nombre de nivells inferior a 150. A banda de les variables textuales, 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.

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

Variable
3: Hotel_Country
4: Hotel_City
10: Room_Type_Level
11: Guest_Type
12: Trip_Type
16: Is_Hotel_Holiday
17: Is_Reviewer_Holiday

---

Variable
19: Review_Is_Positive
21: Reviewer_Nationality
30: Submitted_from_Mobile

---

Nivells
3: {AT, ES, FR, GB, IT, NL}
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: {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}
12: {Business trip Couple Family with older children, Family with young children, Leisure trip, NULL, Solo traveler}
16: {0, 1}
17: {0, 1}
19: {0 1}
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}
30: {0 1}

---

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 convenientes. En general, és recomanable recategoritzar factors amb nivells que reflexen el mateix, o aquelles variables amb un gran nombre de categories per a facilitar ànalisis 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, Vienna} , London{Fitzrovia, London, Paddington, Woodford Green}, Milan{Milan}, Paris{Paris, Paris 06, Paris 12}, Vincennes{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, Duplex, Executive, Family, Luxury, NULL, Premium, Privilege, Standard, Studio, Suite, Superior.
- **NOUS NIVELLS:** Deluxe{Ambassadors, Art, Deluxe, Executive, Luxury, Premium, Privilege, 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, Australia, Austria, Azerbaijan, Bahrain, Bangladesh, Barbados, Belgium, Bermuda, Bosnia and Herzegovina, Botswana, Brazil, Brunei, Bulgaria, Cameroon, Canada, Chile, China, Colombia, Comoros, Costa Rica, Croatia, Curaçao, 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.
- **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, Kazakhstan, Kenya, Kosovo, Latvia, Lebanon, Liechtenstein, Lithuania, Luxembourg, Macedonia, 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}, Switzerland{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 aleatori 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:

- $H_0$ : Missings are completely random (MCAR)
- $H_1$ : Missings are not random

this could take a while

```
[1] 0.0008004798
```

El valor del test ens porta a rebutjar la hipòtesi nula 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\_Ing*, *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 establerts i no apareguin anomalies a les dades (*Figure 19*).

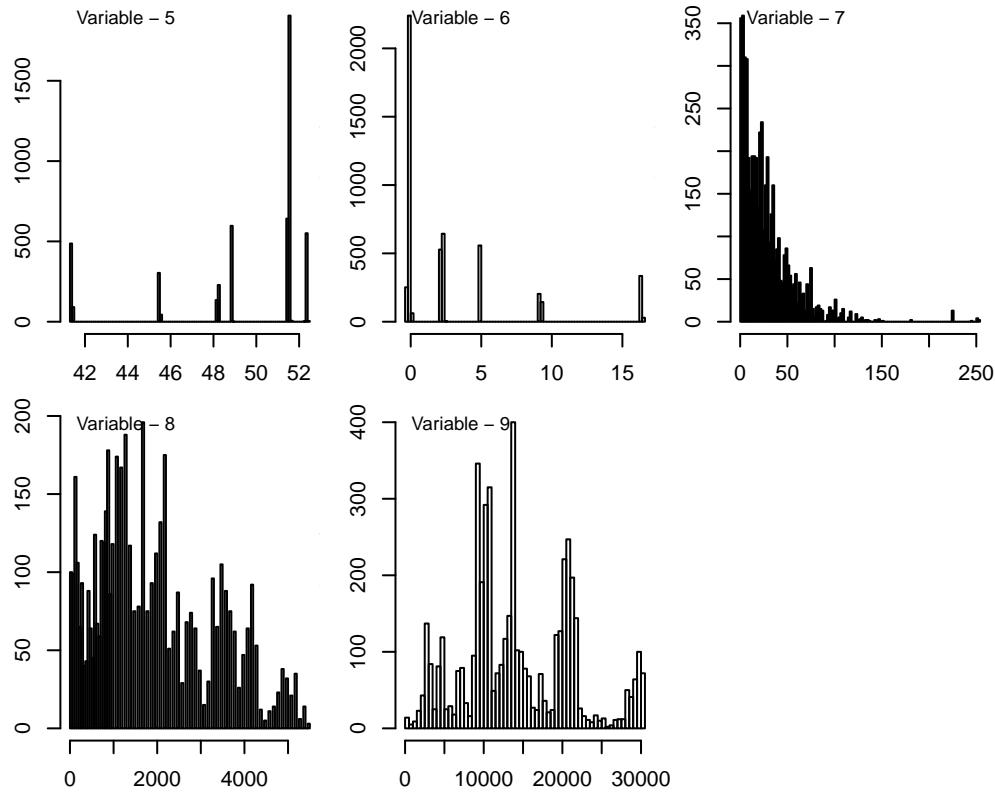


Figure 19: Histogrames sense NA

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 textuales (a la variable de dates no en tenim). En aquest sentit, podem plantear 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 ànalsis).

## Anàlisi descriptiva univariant post preprocessament

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 exploratori 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 especificuem en la fase de preprocessament. En aquest sentit, reconstruïm el gràfic amb les noves modalitats (Figure 20).

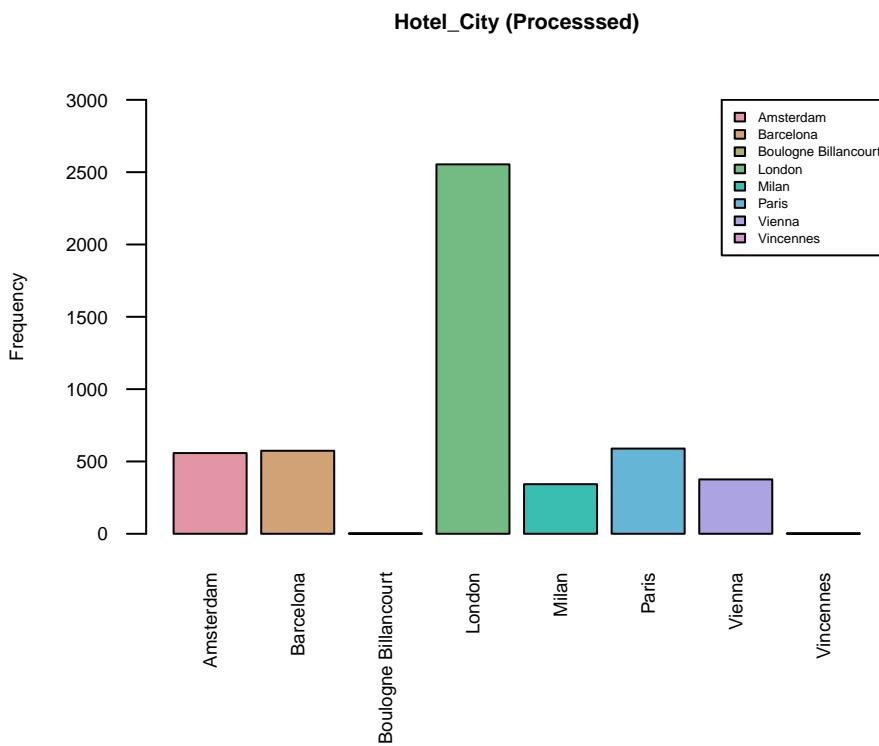


Figure 20: Bar plot Hotel\_City (Processed)

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 conoure 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 especificuem 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 redefinició 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 recodificat 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.

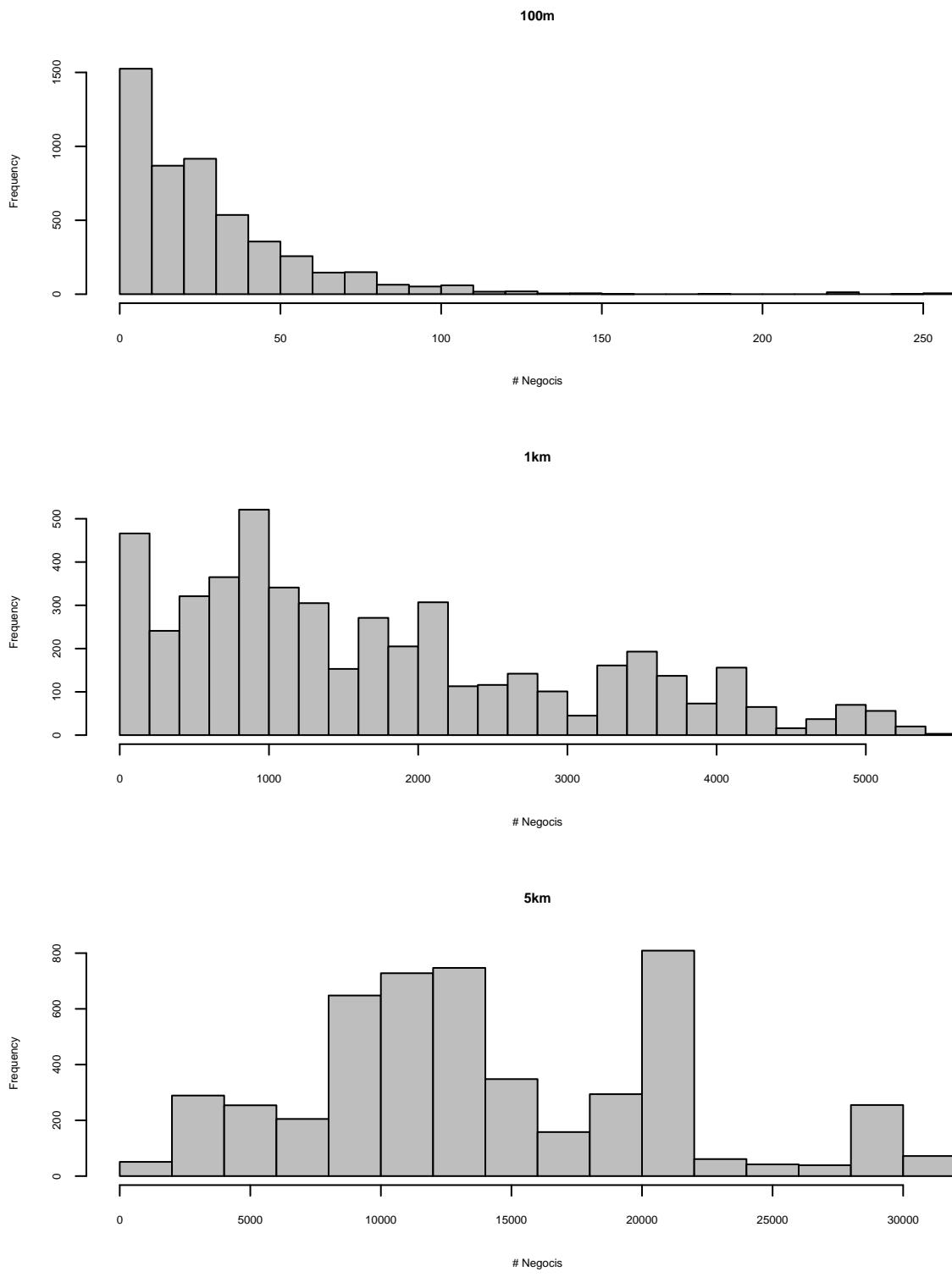


Figure 21: Histogrammes Negocios a la rodona (Processed)

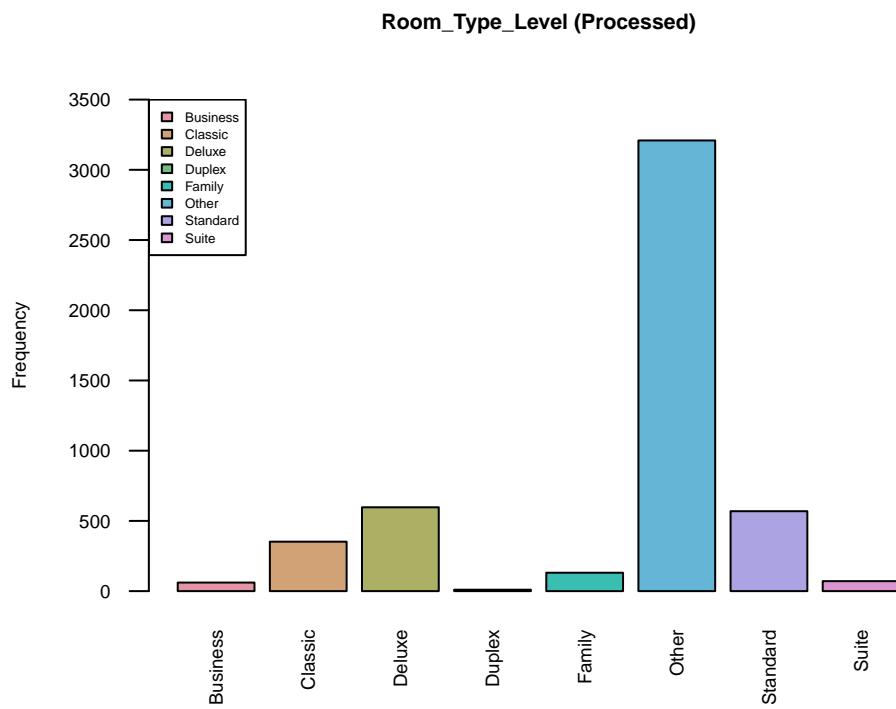


Figure 22: Barplot Room\_Type\_Level (Processed)

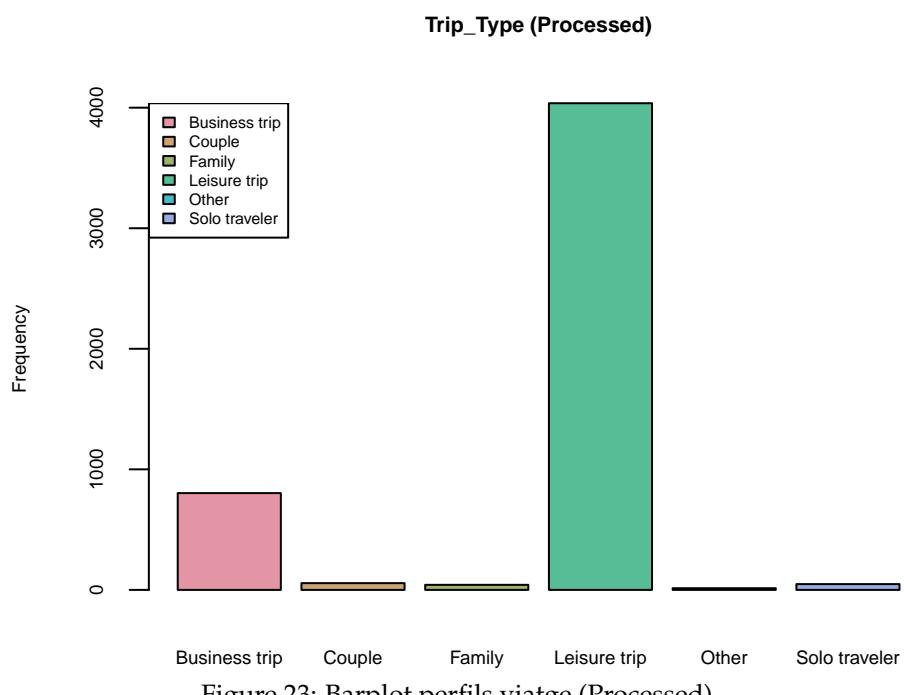


Figure 23: Barplot perfils viatge (Processed)

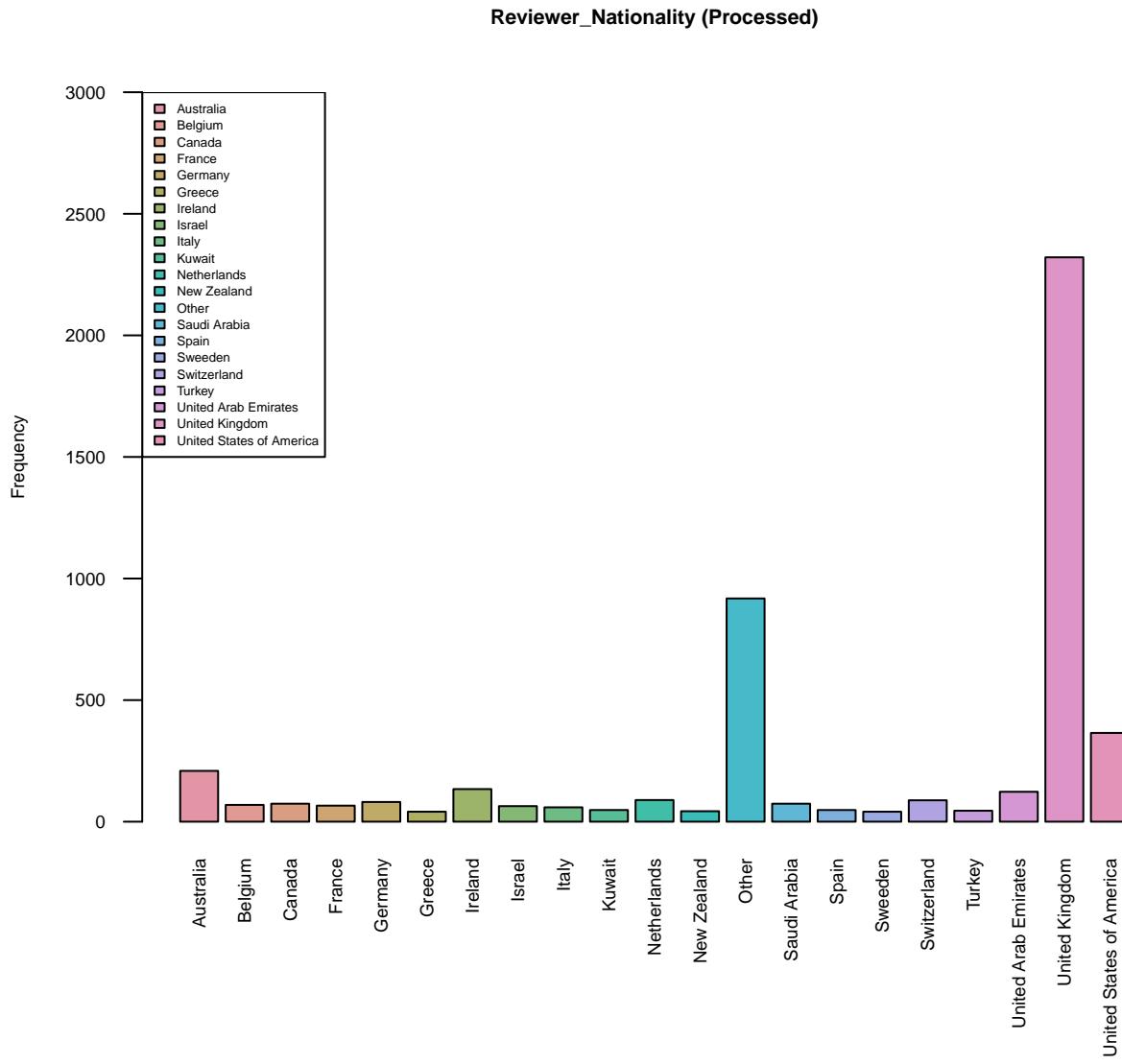


Figure 24: Bar plot Reviewer\_Nationality (Processed)

## Clúster jeràrquic

En l'àmbit de la ciència de dades, es sol treballar amb bases de dades molt grans que tenen un nombre d'observacions molt elevat. Sovint, aquestes observacions són semblants entre sí, de manera que és possible organitzar aquesta gran quantitat de dades en un nombre reduït de *grups* o *clústers*, els quals es componen d'observacions similars. Les diferents tècniques de clustering es fan servir per agrupar dades/observacions en diferents segments de manera que les dades dins de cadascun d'aquests segments són similars, però significativament diferents entre segments. Determinar què vol dir "similars" o "diferents" és la part essencial del *cluster analysis* i està estretament relacionat amb la estadística.

Existeixen diferents tipus de mètodes o tècniques de clústering (particions, jeràrquics etc...) que es basen en diferents metodologies i es serveixen de coneixements teòrics provinents de diferents camps d'estudi. En línia amb els objectius perseguits en aquest projecte, el clústering que es duu a terme en aquest capítol és un *clústering jeràrquic*.

S'empra el mètode de Ward, que consisteix en fer servir la pèrdua d'informació que es produeix al 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.

Es defineix la distància de Gower com  $d_{ij}^2 = 1 - s_{ij}$ , on:

$$s_{ij} = \frac{\sum_{h=1}^{p_1} (1 - |x_{ih} - x_{jh}| / G_h) + a + \alpha}{p_1 + (p_2 - d) + p_3} \quad \text{és el coeficient de similitud de Gower}$$

- $p_1$  és el nombre de variables quantitatives contínues,
- $p_2$  és el nombre de variables binàries,
- $p_3$  és el nombre de variables qualitatives (no binàries),
- $a$  és el nombre de coincidències (1, 1) en les variables binàries,
- $d$  és el nombre de coincidències (0, 0) en les variables binàries,
- $\alpha$  és el nombre de coincidències de les variables qualitatives (no binàries) i
- $G_h$  és el rang (o recorregut) de la  $h$ -èssima variable quantitativa.

Tanmateix, hem considerat exoure del procés de clústering la variable identificadora del registre i les dues variables textuales amb les ressenyes dels clients, ja que la realització d'aquestes variables és única per a cada observació i rebran un tractament diferent. Addicionalment hem considerat oportú no incloure la variable *Hotel\_Name*, ja que també és una característica única per a cada establiment i no té massa sentit pensar en agrupacions en funció d'aquesta variable.

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 dendograma (Figure 25).

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 adicional no compensi

## Dendrograma-WARD



Figure 25: Dendrograma-mètode de Ward amb distàncies de gower

la variabilitat que aconseguim explicar amb aquest tall extra. En aquest treball, proposem un mètode heurístic alternatiu anomenat Elbow Method (*Figure 26*) que es basa en el mateix principi de la suma de quadrats intra clústers. Ens interessarà tenir una suma de quadrats intra clústers petita ja que això voldrà dir que els individus dins d'un mateix conglomerat seran molt similars, però a la vegada realitzar particions només fins al punt que el benefici marginal d'un grup més en superi el cost.

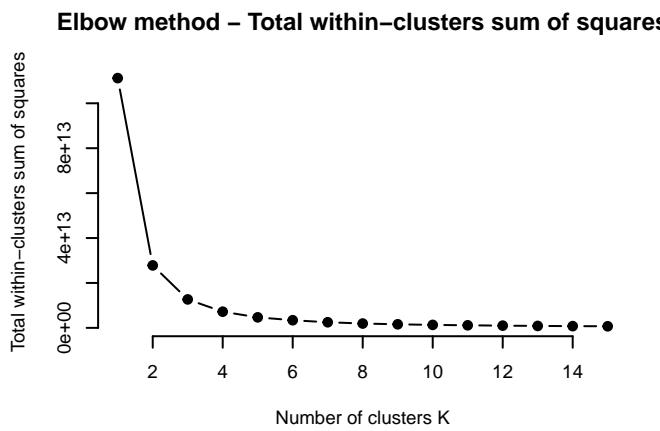


Figure 26: Elbow Method

El Elbow method fa servir només variables numèriques ja que calcula sumes de quadrats. Podría passar que el resultat no es correspongués amb el nostre dendrograma, on hem fet servir totes les variables. Tanmateix, veiem com una partició en 5 clústers sembla força raonable en tots dos casos i el mètode de Elbow pot servir per a reforçar la nostra creença prèvia. Ara bé, caldrà

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

c5

	1	2	3	4	5
1	1268	565	1324	1248	595

c6

c5	1	2	3	4	5	6
1	713	0	555	0	0	0
2	0	565	0	0	0	0
3	0	0	0	1324	0	0
4	0	0	0	0	1248	0
5	0	0	0	0	0	595

c4

c5	1	2	3	4
1	1268	0	0	0
2	0	565	0	0
3	0	0	1324	0
4	0	0	1248	0
5	0	0	0	595

Observant la distribució dels clústers, veiem que els grups 1, 3 i 4 són semblants en mida i més abundants que els grups 2 i 5. A priori, sembla raonable seleccionar aquesta partició. Ara bé, 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:

- **98,17080** amb 4 conglomerats.
- **98,80219** amb 5 conglomerats.
- **99,12840** amb 6 conglomerats.

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

## Profiling dels clústers

Un cop realitzat el procés de clustering, convé pensar en tenir una comprensió més àmplia de com són les unitats d'estudi dins de cada conglomerat. Per a aquest propòsit fem servir les variables de la nostra base de dades que han participat en el clustering per a elaborar panells de classes i estadístiques descriptives per grups<sup>1</sup>. En última instància, l'objectiu és crear un “perfil” per a cada clúster que representi els atributs d'aquest en relació a les variables mencionades.

Les eines descriptives que hem fet servir per a la caracterització dels clústers són:

- Snake plots, diagrames de barres o taules de contingència per a les variables qualitatives.
- Boxplots i diagrames de barres per a les variables numèriques.

A més, per cada variable analitzada, fem servir un seguit de contrastos, tant paramètrics com no paramètrics, per a testar la hipòtesis de diferències significatives 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 llògic pensar que la caracterització dels clústers anirà en la mateixa línia en tots quatre casos.

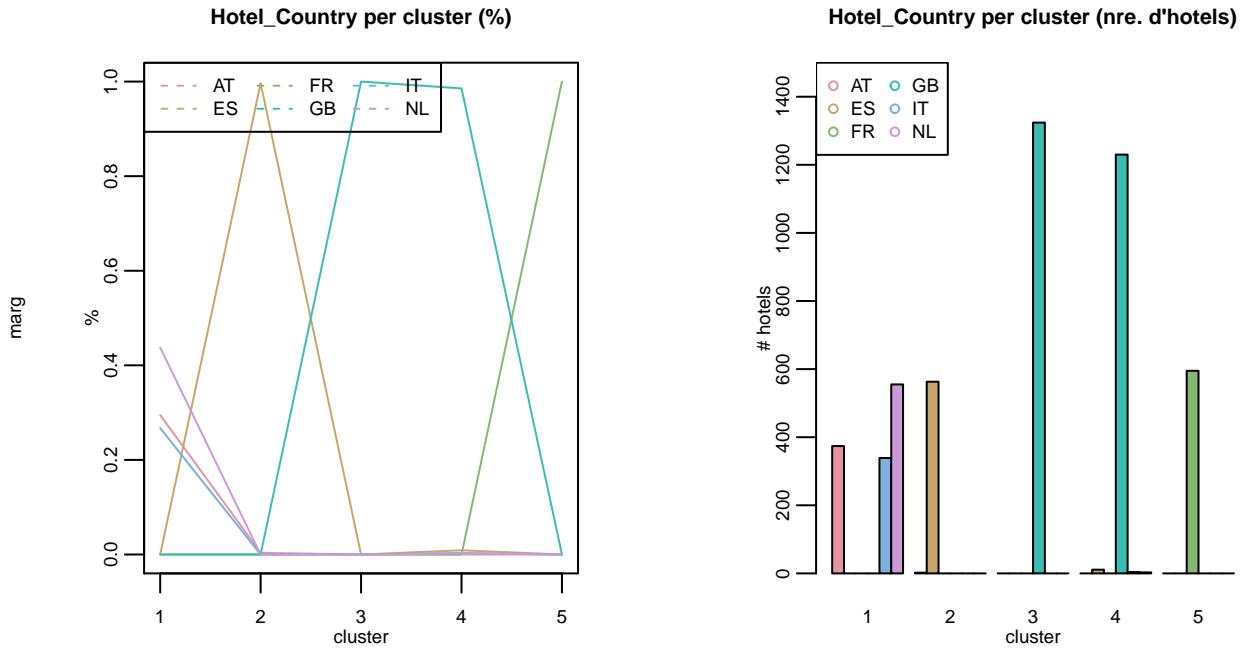


Figure 27: Profiling variable Hotel\_Country

Si construim el snake plot i el diagrama de barres de la variable *Hotel\_Country* estratificant per clúster (Figure 27), 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, Itàlia i Països Baixos 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 del Regne Unit.
- Clúster 4: Predominen hotels del Regne Unit amb lleugera presència d'hotels d'altre països.
- Clúster 5: Predominen hotels de França.

<sup>1</sup>Deixarem fora l'identificador, les dues variables textuals i Hotel\_Name.

D'altra banda si en comptes de considerar la variable *Hotel\_Country* fem exactament el mateix per a la ciutat de l'hotel (*Figure 28*) observem com existeix una correspondència en la relació entre països i ciutats:

- Al clúster 1 hi trobem hotels de Ámsterdam, Vienna i Milà.
- Al clúster 2 hi trobem hotels de Barcelona.
- Als clústers 3 i 4 trobem majoritàriament hotels de Londres.
- Al clúster 5 predominen els hotels situats a París.

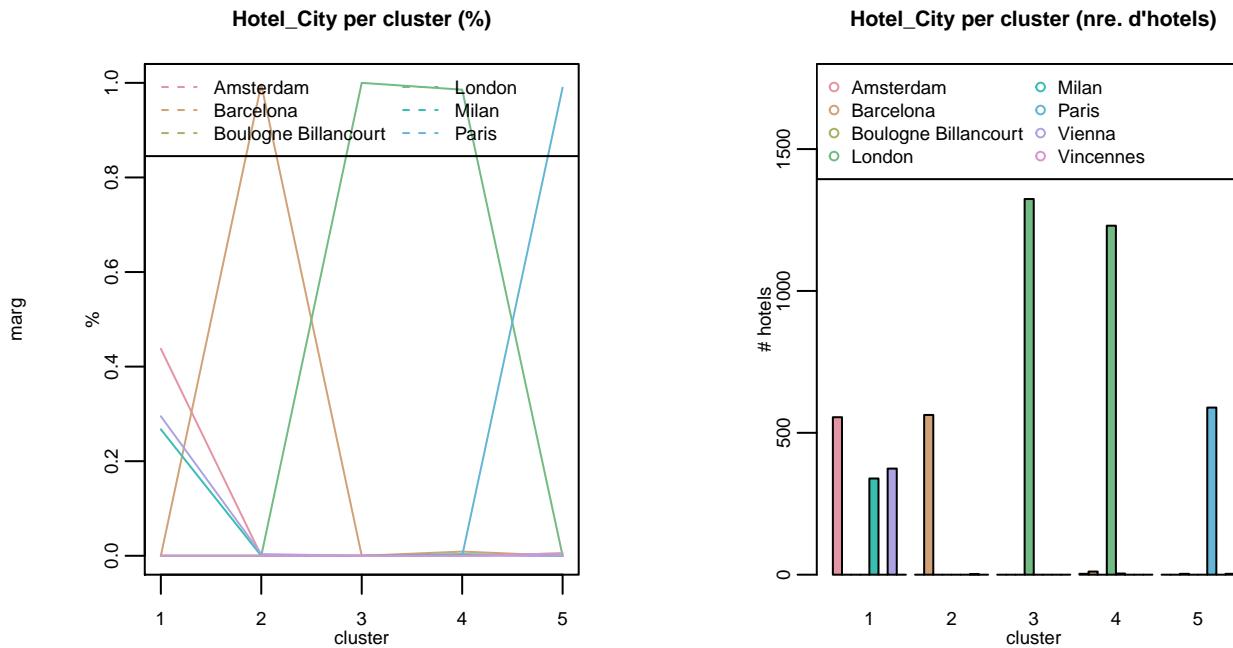


Figure 28: Profiling variable Hotel\_City

Per últim, podem realitzar un test  $\chi^2$  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 = 14818, 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 = 14818, df = 28, p-value < 2.2e-16
```

Tots dos contrastos són significatius.

Per a facilitar la comprensió dels resultats descrits anteriorment, podem pensar en fer servir les variables latitud i longitud per a realitzar una geolocalització de les dades i visualitzar els clústers en un mapa. Als següents gràfics representem al mapa la localització dels hotels, estratificant per clústers, els quals distingim amb colors, i on la mida dels cercles fa referència al nombre total de ressenyes que tenen els hotels (*Figure 29*).

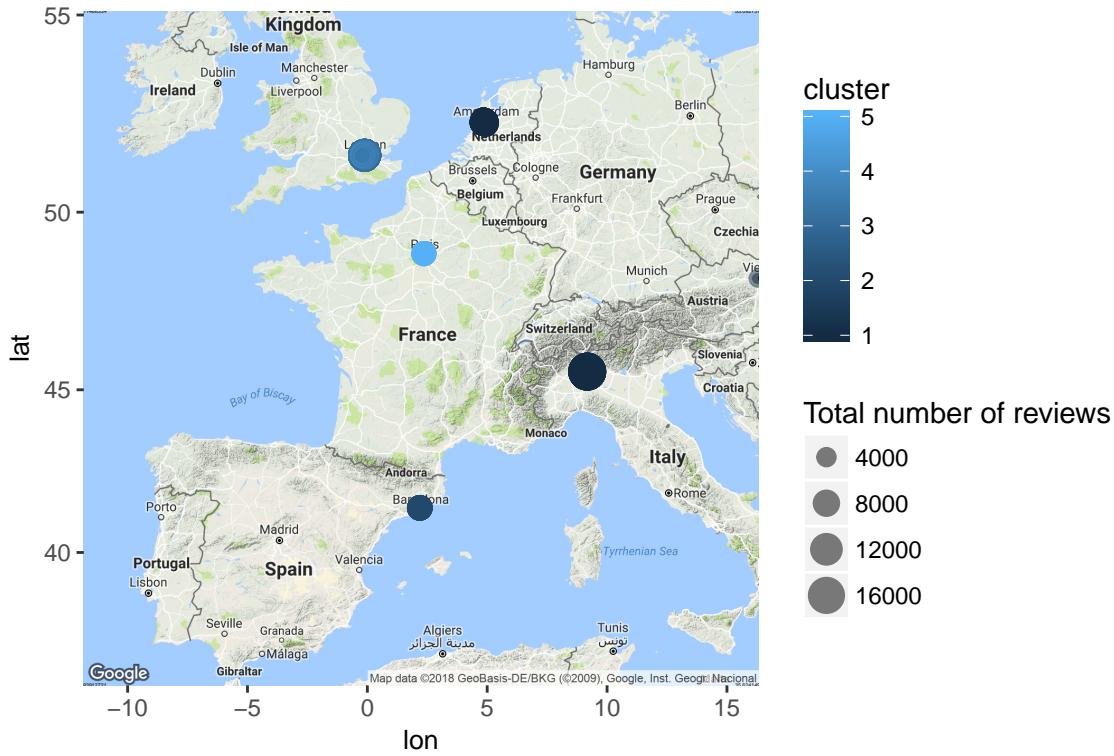


Figure 29: Geolocalització

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. Com que les variables *Business\_100m*, *Business\_1km* i *Business\_5Km* tenen connotacions semblants, hem considerat representar-les juntes mitjançant un plotMeans (*Figure 30*).

Observem com els hotels que trobem als clústers 2 i 5 tendeixen a ser més urbans, en especial els del clúster 5 ja que destaquen en tots els llindars. Els hotels inclosos al clúster 2 destaquen, encara que en menor mesura que els del 5, quan la distància és inferior a 1 km a la rodona (pot ser degut a que Barcelona és menys extensa que les altres ciutats). En contrapartida, els hotels del clúster 1 semblen ser els menys urbans ja que ocupen posicions baixes en tots tres llindars, i es troben força allunyats de la mitjana. Les dues primeres variables comparteixen moltes característiques i és l'última (*Businesses\_5km*) on més evidents es fan les diferències.

Proposem la realització d'una ANOVA i un test de Kruskall-Wallis per a testar si existeixen diferències globals entre clusters. A part, calculem la significació de cada clúster amb la funció *ValorTestXnum*. En aquest sentit, p-valors molt extrems per a un clúster voldrà dir que aquest està molt allunyat de la mitjana.

```
[1] "p-valueANOVA Businesses_100m: 2.76654704606887e-37"
```

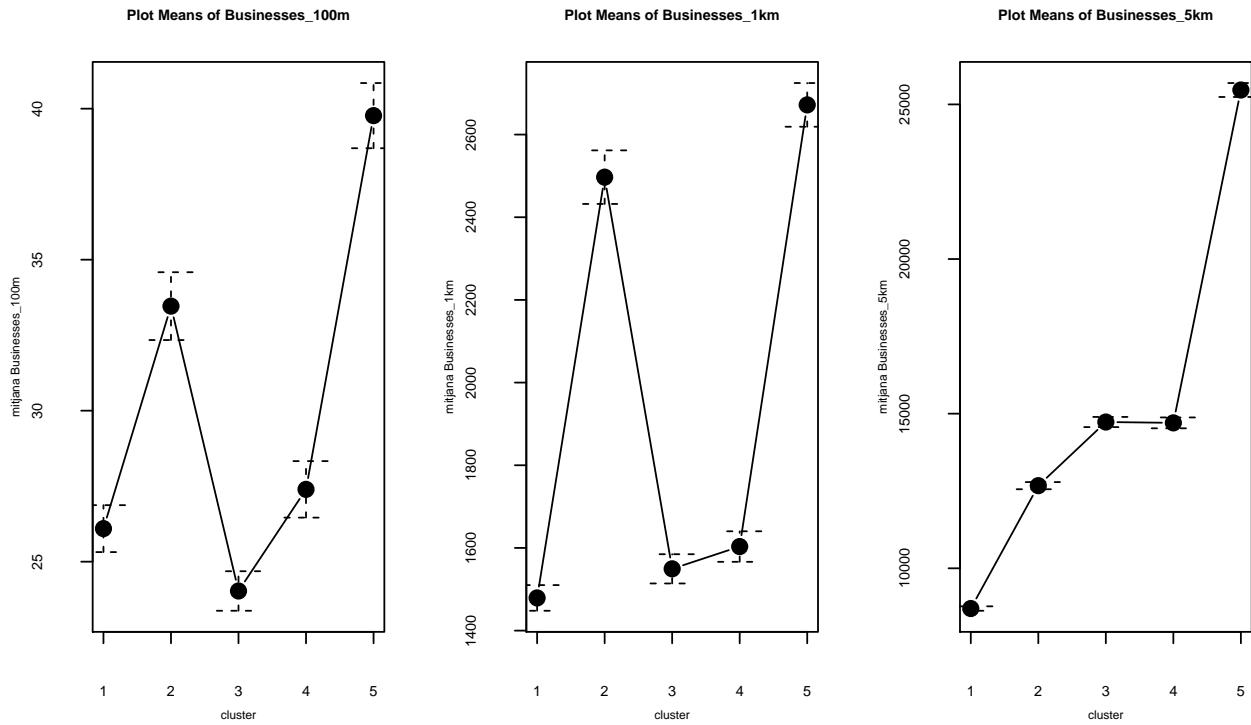


Figure 30: Profiling Nre. Negocis a 100m, 1km i 5km a la rodona

```
[1] "p-value Kruskal-Wallis Businesses_100m: 3.30050126457872e-60"

[1] "p-values ValorsTest Businesses_100m:"
[2] "0.000576639115411615"
[3] "2.54663976133031e-06"
[4] "6.57803811421331e-11"
[5] "0.0887515703701272"
[6] "9.53742443473316e-26"

[1] "p-valueANOVA Businesses_1km: 2.64238824497962e-105"

[1] "p-value Kruskal-Wallis Businesses_1km: 2.94063599839316e-107"

[1] "p-values ValorsTest Businesses_1km:"
[2] "0"
[3] "1.05571553750797e-39"
[4] "8.01581023779363e-14"
[5] "2.13068271914807e-08"
[6] "2.4992576114904e-63"

[1] "p-valueANOVA Businesses_5km: 0"

[1] "p-value Kruskal-Wallis Businesses_5km: 0"

[1] "p-values ValorsTest Businesses_5km:"
[2] "0"
[3] "6.88717216590362e-09"
```

```
[4] "0.00136745579630548"
[5] "0.00326407457008836"
[6] "0"
```

Si ens fixem en els p-valors tots tres contrastos 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.

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 (*Figure 31*)

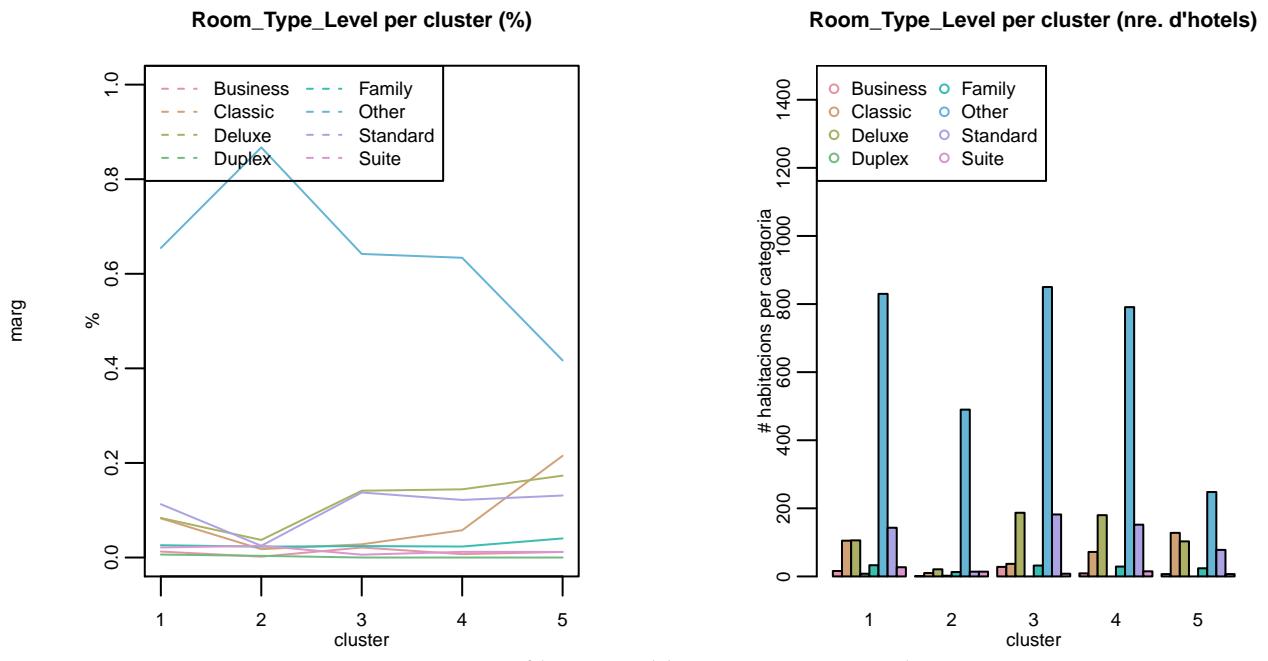


Figure 31: Profiling variable Room\_Type\_Level

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 2 (més del 80%). La resta de modalitats són força homogènies per a tots els clústers amb la salvetat que al clúster 5 hi ha major presència d'habitacions del tipus clàssic i en el clúster 3 les del tipus Standard i Deluxe. El test  $\chi^2$  és significatiu però hem de tenir en compte que els valors centrals poden estar altament influïts per la modalitat "Other".

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

Pearson's Chi-squared test

```
data: dd[, k] and as.factor(P)
X-squared = 507.11, df = 28, p-value < 2.2e-16
```

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

(Figure 32), observem com als clústers 1, 3 i 4 les parelles són més abundants que a la resta. El clúster 3 també conté més viatgers solitaris que la resta, mentre que al clúster 2 augmenten lleugerament els grups. La resta de categories es manté força constant per a tots els conglomerats.

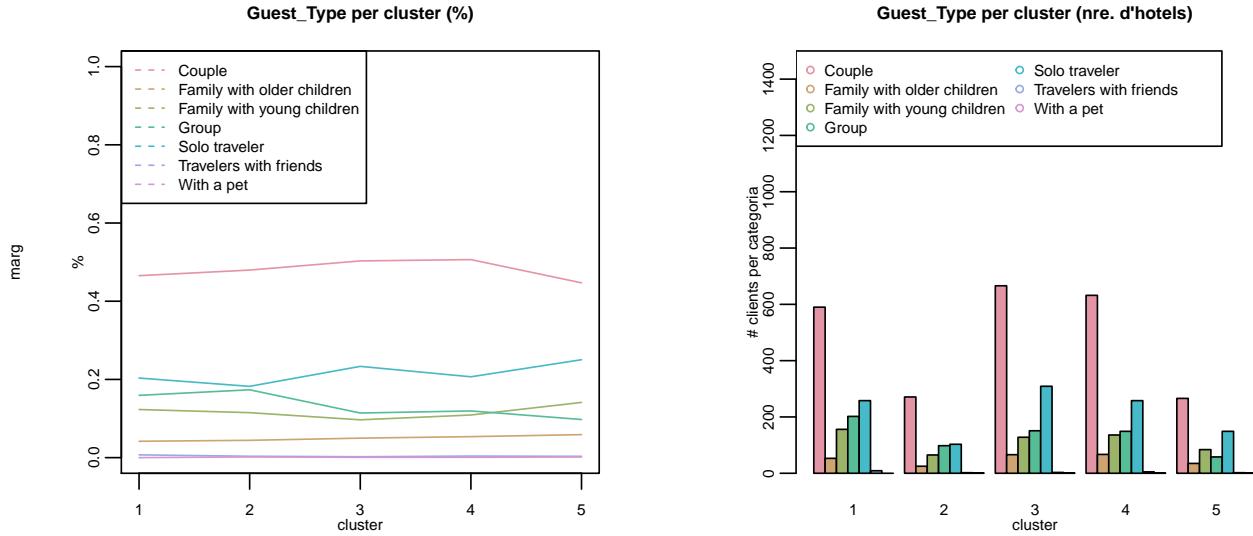


Figure 32: Profiling variable Guest\_Type

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

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

Pearson's Chi-squared test

```
data: dd[, k] and as.factor(P)
X-squared = 57.116, df = 24, p-value = 0.0001612
```

A continuació, passem a la variable *Trip\_Type* que reflecteix el motiu del viatge (Figure 33). En aquest cas, observem com, en tots els conglomerats predominen els viatges amb motiu d'oci. Tanmateix, trobem diferències subtils al clúster 3 on aquesta dominància dels viatges per plaer disminueix en benefici dels viatges per motiu de negocis.

Tot i això, podem concloure que la distribució de la variable és més aviat homogènia per a tots els clústers. Si construim el test, veiem com existeixen diferències significatives entre els conglomerats, però el p-valor no és tan petit com el trobat per a altres variables (menys evidència de diferències entre clústers).

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

Pearson's Chi-squared test

```
data: dd[, k] and as.factor(P)
X-squared = 42.378, df = 20, p-value = 0.002468
```

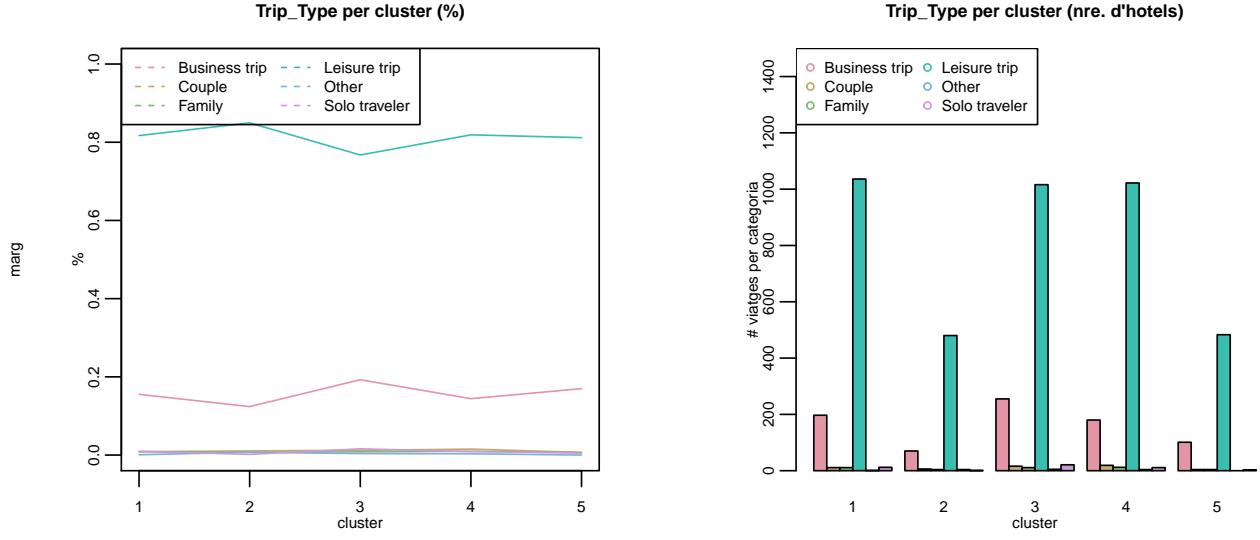


Figure 33: Profiling variable Trip\_Type

La següent variable és *Stay\_Duration*. Al tenir davant una variable numèrica, construim un boxplot i un gràfic de barres (Figure 34). Observem com els clústers 1, 2 i 5 inclouen llargues estades, mentre que els clústers 3 i 4 (recordem, majoria de parelles o viatgers en solitari) tendeixen a incloure estades més curtes.

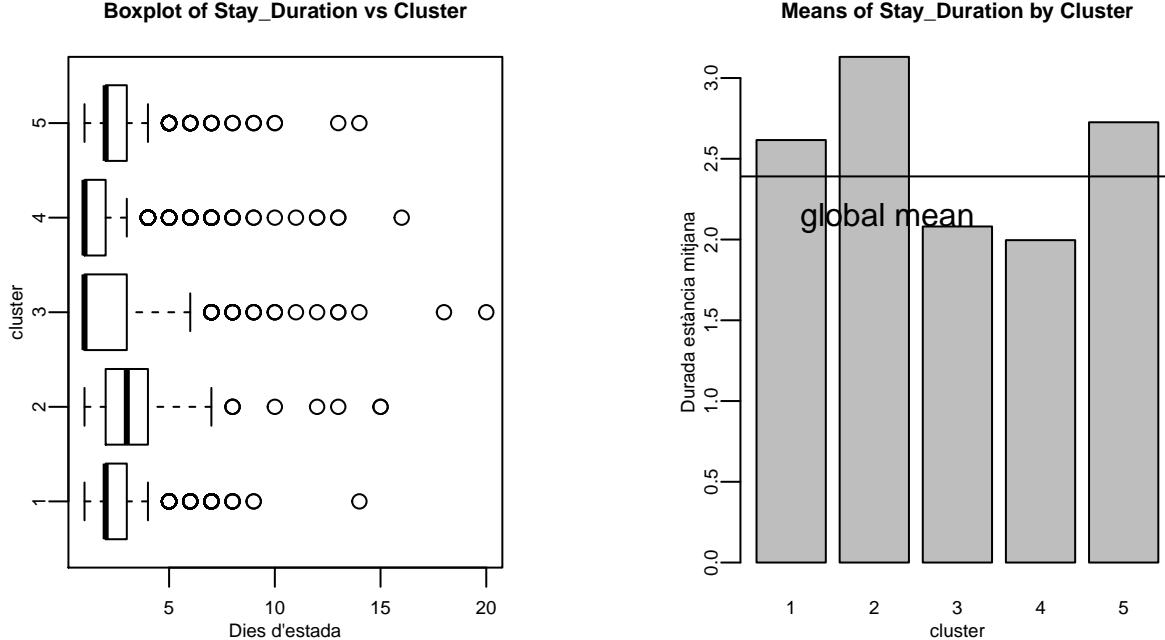


Figure 34: Profiling variable Stay\_Duration

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

[1] "p-valueANOVA Stay\_Duration: 9.59192701931915e-53"

[1] "p-value Kruskal-Wallis Stay\_Duration: 5.80529941960794e-109"

```
[1] "p-values ValorsTest Stay_Duration:"
[1] 2.302153e-08 3.162802e-28 5.107026e-15 0.000000e+00 1.425555e-07
```

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 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 al segon i el tercer.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.0	177.8	351.0	350.8	513.0	730.0	
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	
0.0	164.0	343.0	341.3	487.0	730.0	
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	
0.0	178.0	344.5	354.5	528.2	730.0	
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	
0.0	181.0	356.5	356.1	527.0	730.0	
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	
1.0	192.5	374.0	380.6	577.0	730.0	

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.

```
[1] "p-valueANOVA Days_Since Review: 0.027114559022556"
[1] "p-value Kruskal-Wallis Days_Since Review: 0.0220796017640812"
[1] "p-values ValorsTest Days_Since Review:"
[1] 0.171837020 0.042948950 0.413178419 0.460035852 0.000918868
```

A continuació ens fixem conjuntament en les variables *Is\_Hotel\_Holiday* i *Is\_Reviewer\_Holiday*, dicotòmiques les dues. Per a caracteritzar els clústers, construim dos diagrames de barres apilades de manera que poguem comparar entre els conglomerats si la ciutat de l'hotel, o la de l'usuari es troba en dia festiu (*Figure 35*). L'objectiu és veure si en algun clúster els clients tendeixen a escriure les ressenyes en dies festius.

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 Sí és molt baixa per a les dues variables. Si ens fixem en els p-valors dels contrastos, observem com els dos testos surten significatius.

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

Pearson's Chi-squared test

```
data: dd[, k] and as.factor(P)
X-squared = 74.728, df = 4, p-value = 2.274e-15
```

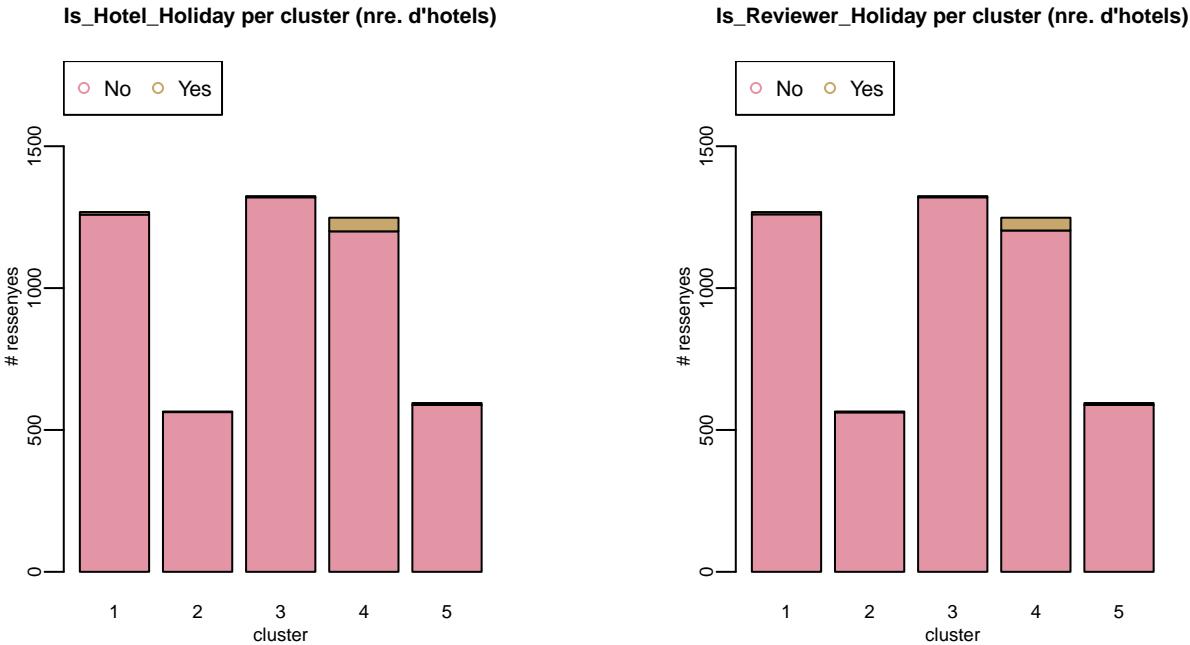


Figure 35: Profiling variables dia Festiu

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

Pearson's Chi-squared test

```
data: dd[, k] and as.factor(P)
X-squared = 66.99, df = 4, p-value = 9.797e-14
```

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 repercutix en la valoració de l'hotel.

Al gràfic (*Figure 36*) 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.

```
[1] "p-valueANOVA Total_Number_of_Reviews: 2.06602500452088e-142"
[1] "p-value Kruskal-Wallis Total_Number_of_Reviews: 1.86632785558636e-127"
[1] "p-values ValorsTest Total_Number_of_Reviews:"
[1] 9.359835e-11 3.682168e-06 3.464789e-12 5.307368e-03 0.000000e+00
```

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 cluster, juntament amb un gràfic de barres o un PlotMeans de la variable *Review\_Positivity\_Rate* (*Figure*

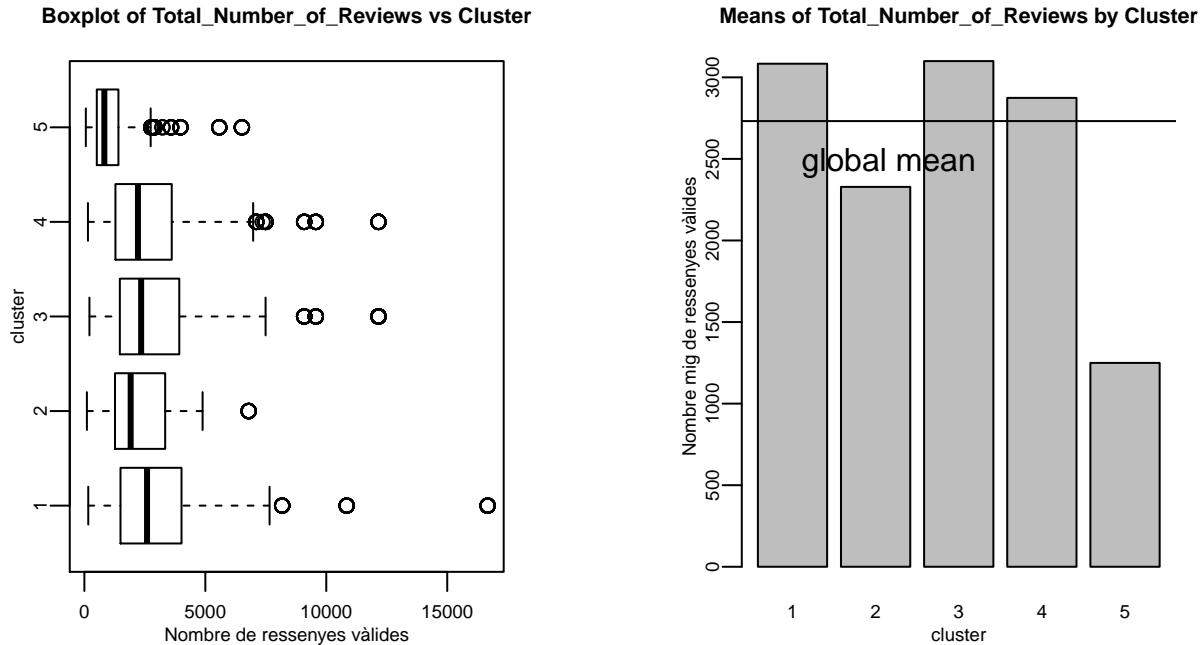


Figure 36: Profiling variable Total\_Number\_of\_Reviews

37), ens pot donar un indici de 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.

Observant el gràfic, podem concloure que:

- El clúster 4 té gairebé totes les ressenyes i comentaris positius.
- Els clústers 1, 2 i 5 tenen un percentatge de ressenyes positives superior al 50% i superior a la mitjana global en tots tres casos.
- El clúster 3 té gairebé totes les ressenyes i comentaris negatius.

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

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

Pearson's Chi-squared test

```
data: dd[, k] and as.factor(P)
X-squared = 2499.4, df = 4, p-value < 2.2e-16
```

```
[1] "p-valueANOVA Review_Positivity_Rate: 0"
```

```
[1] "p-value Kruskal-Wallis Review_Positivity_Rate: 0"
```

```
[1] "p-values ValorsTest Review_Positivity_Rate:"
```

```
[1] 1.841366e-01 8.408602e-05 0.000000e+00 5.113320e-215 1.255479e-03
```

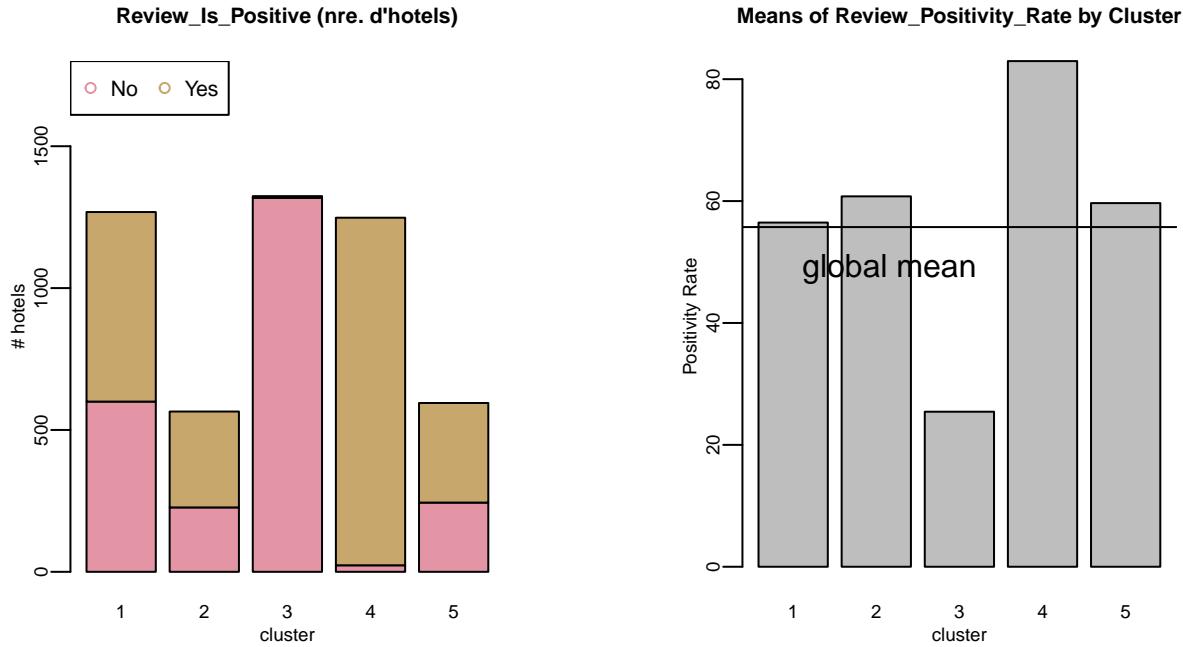


Figure 37: Profiling variables grau de positivisme de la ressenya

La següent variable és *Reviewer Nationality*. Tal i com hem fet anteriorment, construim un snake plot i un gràfic de barres (variable categòrica) (Figure 38). Es pot apreciar com la nacionalitat que predomina està altament correlacionada amb el país de l'hotel, un clar exemple en són els clústers 3 i 4 on predominen els hotels situats a Gran Bretanya i les ressenyes escriptes per britànics. Això pot indicar que el turisme intern és molt més usual a Booking o, al menys, que és més comú escriure ressenyes per a hotels del teu país.

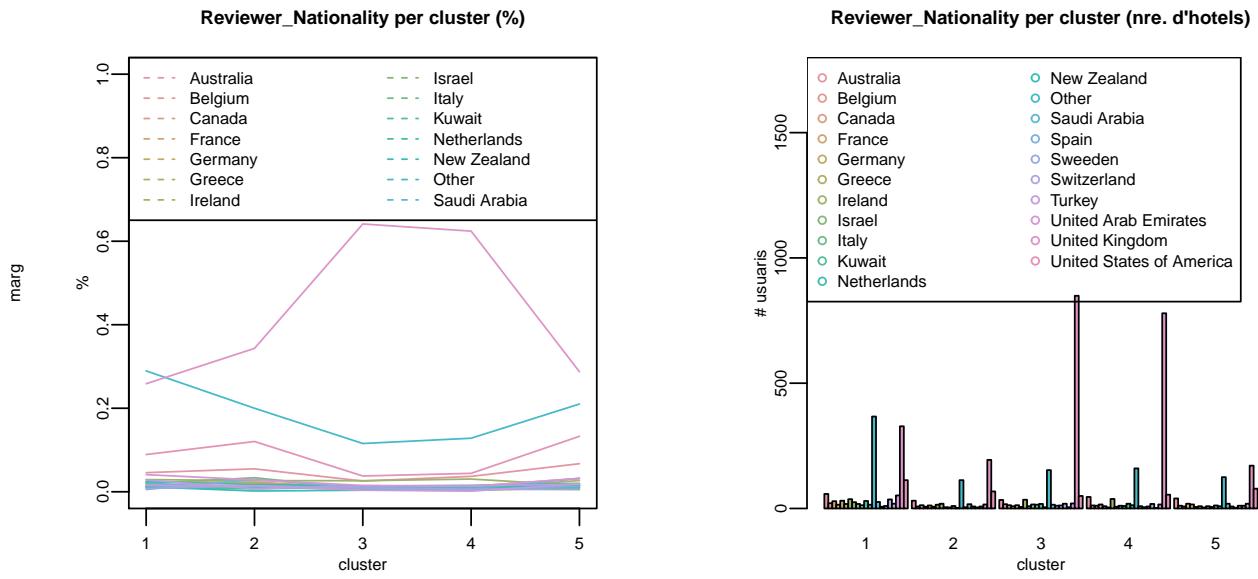


Figure 38: Profiling variable Reviewer\_Nationality

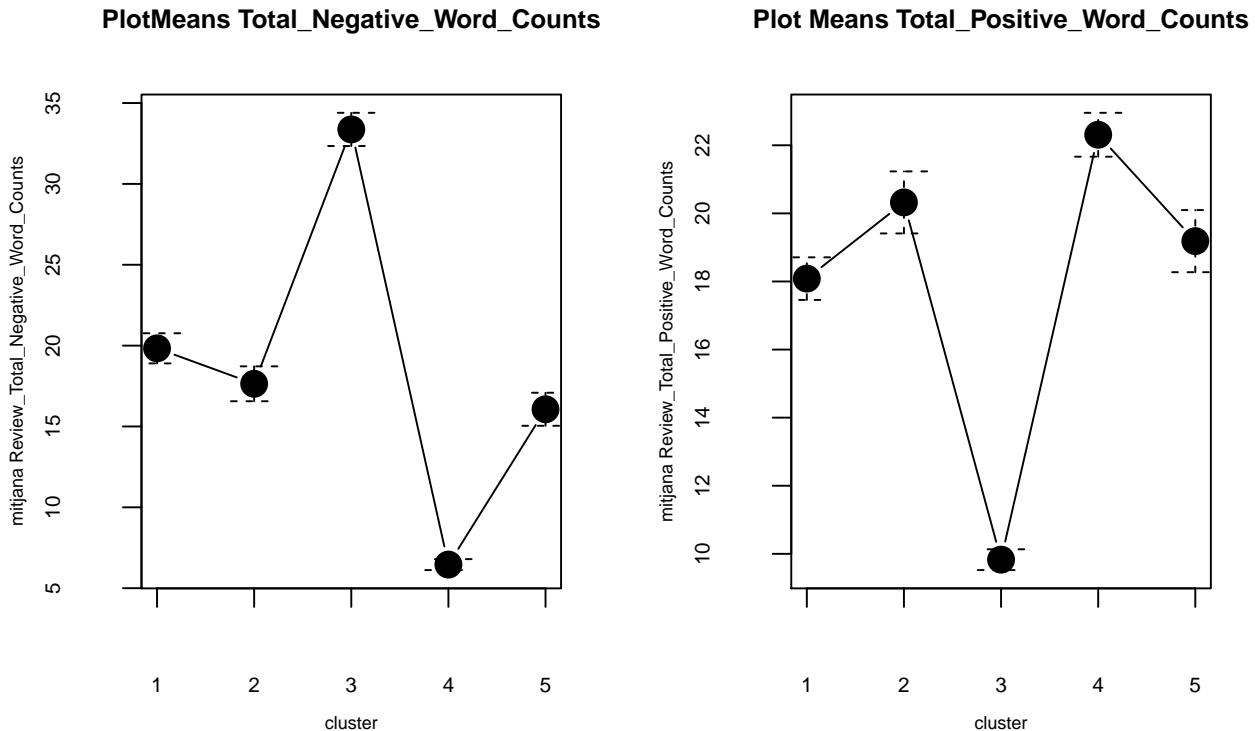
De nou, el test mostra una alta significació estadística.

[1] "Test Chi quadrat Reviewer\_Nationality:"

Pearson's Chi-squared test

```
data: dd[, k] and as.factor(P)
X-squared = 825.77, df = 80, p-value < 2.2e-16
```

Les següents variables, si excloem les textuais, són *Review\_Total\_Negative\_Word\_Counts* i *Review\_Total\_Positive\_Word\_Counts*. Entenem que aquestes variables han de anar en línia amb el resultat anterior on hem valorat el grau de positivisme del comentari. Comparem totes dues juntes mitjançant un plotMeans i veiem com el resultat, llògicament, és complementari. Aquells clústers amb valors als per a *Review\_Total\_Negative\_Word\_Counts* tindràn valors baixos per a *Review\_Total\_Positive\_Word\_Counts* i viceversa.



Les conclusions que obtenim del gràfic són:

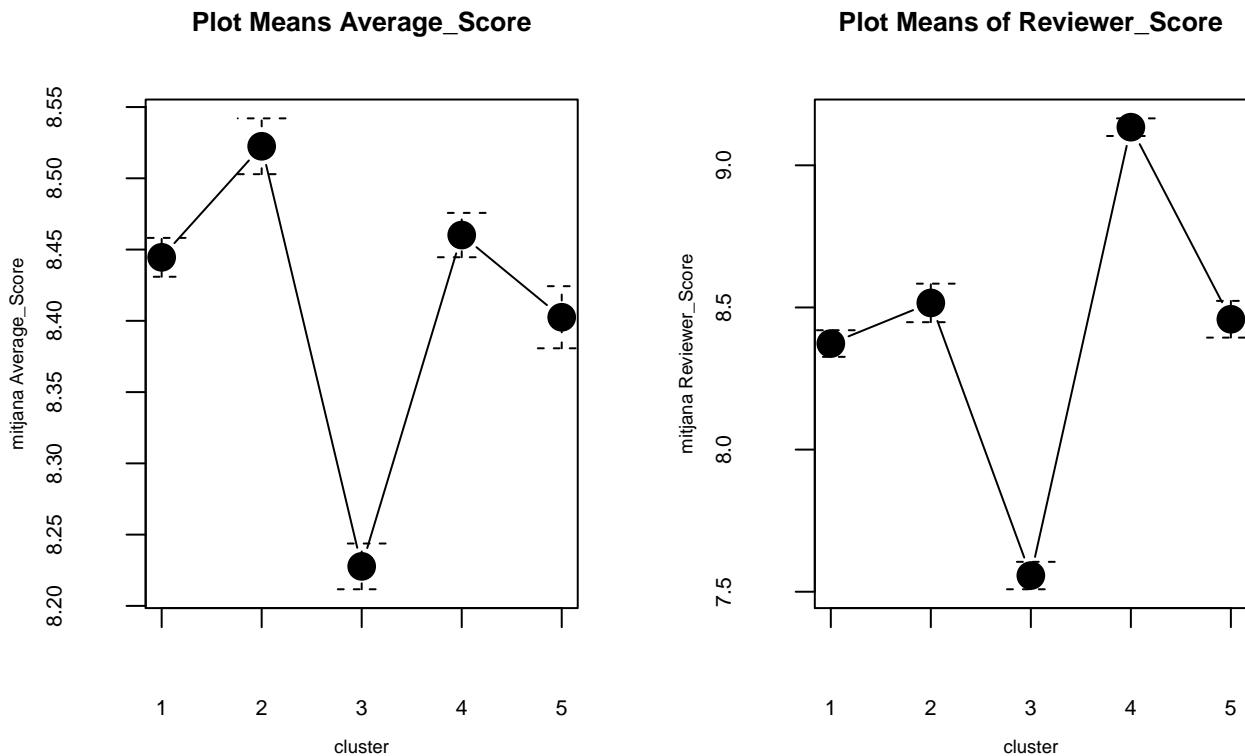
- El clúster 3 és el que conté els hotels on els comentaris són més negatius.
- Els hotels dels clústers 1, 2 i 5 es mantenen prop de la mitjana global encara que les valoracions per a aquest últim tendeixen a ser més negatives.
- El clúster 4 és el que conté els hotels on els comentaris són més positius.

El contrast, reforça la informació proporcionada pel gràfic, mostrant una forta validesa estadística en relació al test de diferències entre clústers.

```
[1] "p-valueANOVA Review_Total_Negative_Word_Counts: 2.63231256048715e-143"
[1] "p-value Kruskal-Wallis Review_Total_Negative_Word_Counts: 3.9141297138178e-247"
[1] "p-values ValorsTest Review_Total_Negative_Word_Counts:"
```

```
[1] 2.713636e-01 7.431889e-02 6.932987e-82 0.000000e+00 2.326173e-03
[1] "p-valueANOVA Review_Total_Positive_Word_Counts: 2.87512630818527e-90"
[1] "p-value Kruskal-Wallis Review_Total_Positive_Word_Counts: 8.78970324320619e-103"
[1] "p-values ValorsTest Review_Total_Positive_Word_Counts:"
[1] 6.703907e-02 1.242412e-04 0.000000e+00 4.643612e-23 9.730768e-03
```

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 otorgant 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 llògic pensar que els resultats haurien d'anar en línia amb els anteriors.



En efecte, les conclusions que podem obtenir d'aquests gràfics són calcades a les que hem obtingut analitzant *Review\_Is\_Positive*, *Review\_Positivity\_Rate*, *Review\_Total\_Negative\_Word\_Counts* i *Review\_Total\_Positive\_Word\_Counts*: clúster 3 pitjors valoracions, clúster 4 millors valoracions i la resta mantenint-se en la mitjana global. Això pot ser un indicí 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, 2 i 5 que, recordem es troben a prop de la mitjana global.

```
[1] "p-valueANOVA Average_Score: 5.28801912822158e-36"
[1] "p-value Kruskal-Wallis Average_Score: 1.35068451835033e-33"
[1] "p-values ValorsTest Average_Score:"
```

```
[1] 8.000340e-05 1.532450e-09 0.000000e+00 4.647395e-07 3.564448e-01
[1] "p-valueANOVA Reviewer_Score: 8.91299361097231e-143"
[1] "p-value Kruskal-Wallis Reviewer_Score: 5.44372655792561e-142"
[1] "p-values ValorsTest Reviewer_Score:"
[1] 4.986789e-01 1.465757e-02 0.000000e+00 2.068419e-76 8.999279e-02
```

La següent variable que analitzem, ja a falta de només dues per conoure el profiling, 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 39*). 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, exepte els del tercer conglomerat, on la mitjana de comentaris totals escrits pels usuaris és més baixa.

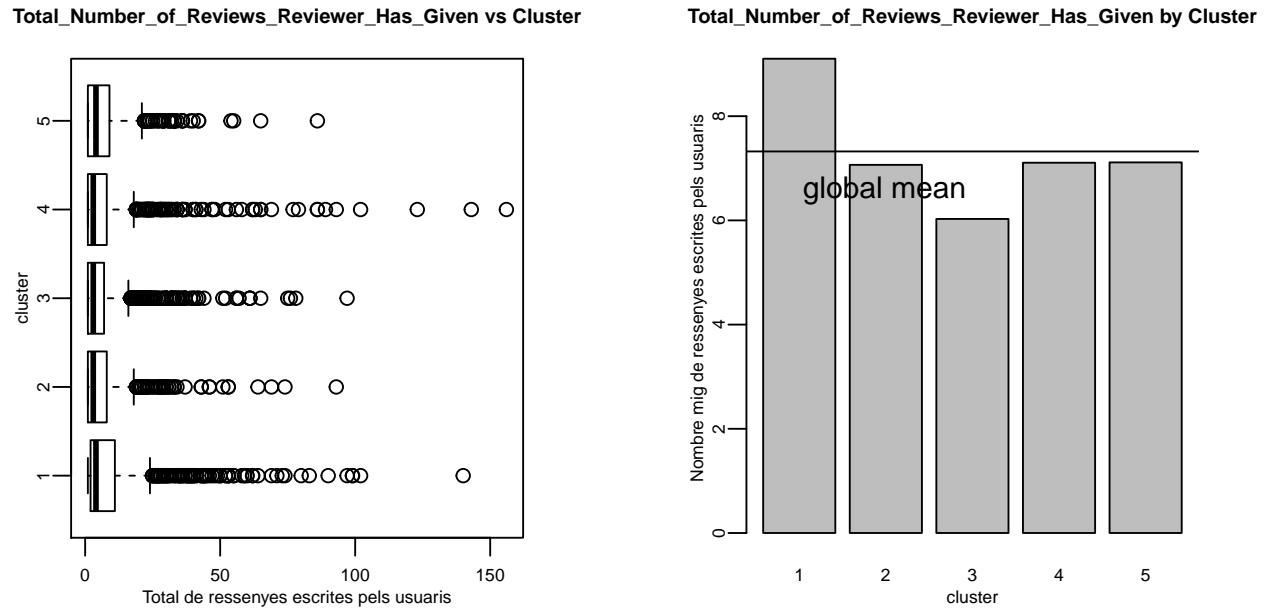


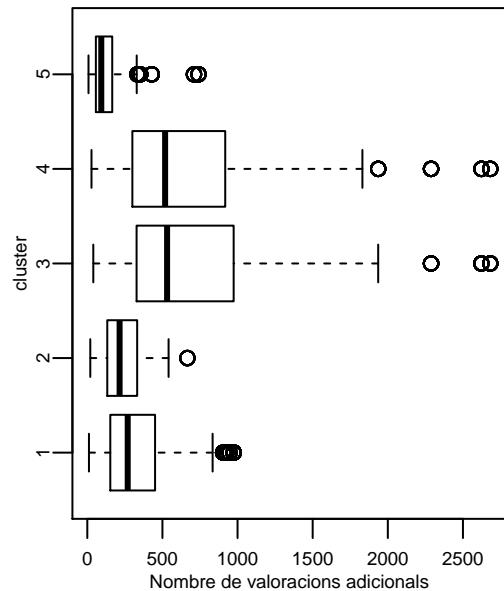
Figure 39: Profiling variable Total\_Reviews\_Given

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

```
[1] "p-valueANOVA Total_Number_of_Reviews_Reviewer_Has_Given: 2.34234297335977e-09"
[1] "p-value Kruskal-Wallis Total_Number_of_Reviews_Reviewer_Has_Given: 2.67414288494901e-16"
[1] "p-values ValorsTest Total_Number_of_Reviews_Reviewer_Has_Given:"
[1] 6.995392e-11 2.844451e-01 7.374909e-07 2.195685e-01 3.158726e-01
```

A continuació ens fixem en la variable *Additional\_Number\_of\_Scoring* relativa al total de valoracions adicionals que rep l'hotel (localització, nejeta, servei...). Seria interessant veure si els hotels amb valoracions més positives també reben un major nombre de comentaris adicionals, o aquest efecte és just al contrari (*Figure 40*).

Boxplot of Additional\_Number\_of\_Scoring vs Cluster



Means of Additional\_Number\_of\_Scoring by Cluster

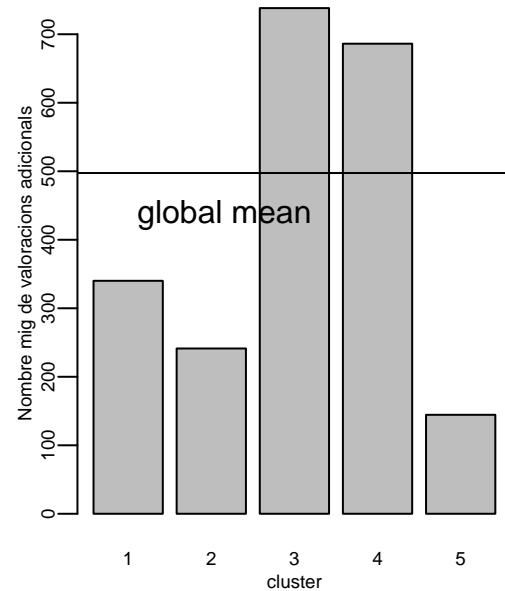


Figure 40: Profiling variable Additional\_Number\_of\_Scoring

Veiem com el resultat és clarament significatiu i, curiosament, els hotels dels clústers on les puntuacions són més extremes (clúster 3 les més negatives i clúster 4 les més positives) 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'pcionals.

Els contrastos mostren una forta validesa estadística per als resultats obtinguts.

```
[1] "p-valueANOVA Additional_Number_of_Scoring: 7.89067302316596e-319"
[1] "p-value Kruskal-Wallis Additional_Number_of_Scoring: 0"
[1] "p-values ValorsTest Additional_Number_of_Scoring:"
[1] 0.000000e+00 0.000000e+00 3.767604e-91 1.470819e-53 0.000000e+00
```

Per últim, considerem la variable *Submitted\_from\_Mobile*. Aquesta variable ens mostrerà si en algun clúster són més freqüents les ressenyes i valoracions escrites des del mòvil. Aquesta variable pot ser indicativa de si els usuaris estan satisfets, o no, amb el funcionament de la aplicació de Booking.

No existeixen grans diferències entre els clústers pel que fa a la proporció de ressenyes escrites des del telèfon mòbil (Figure 41). La diferència més rellevant apareix en els usuaris compressos al clúster 3, els quals tendeixen a fer servir més el telèfon mòvil per a escriure les seves ressenyes. Recordem que els hotels dins aquest conglomerat tendeixen a rebre valoracions més negatives i, en conseqüència, és possible inferir que aquestes s'escriuen en major grau des de telèfons mòvil. Pel que fa al contrast, veiem com la significació global no és massa forta, segurament degut a que, excloent el clúster 3, tots els altres es troben força a prop de la mitjana global.

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

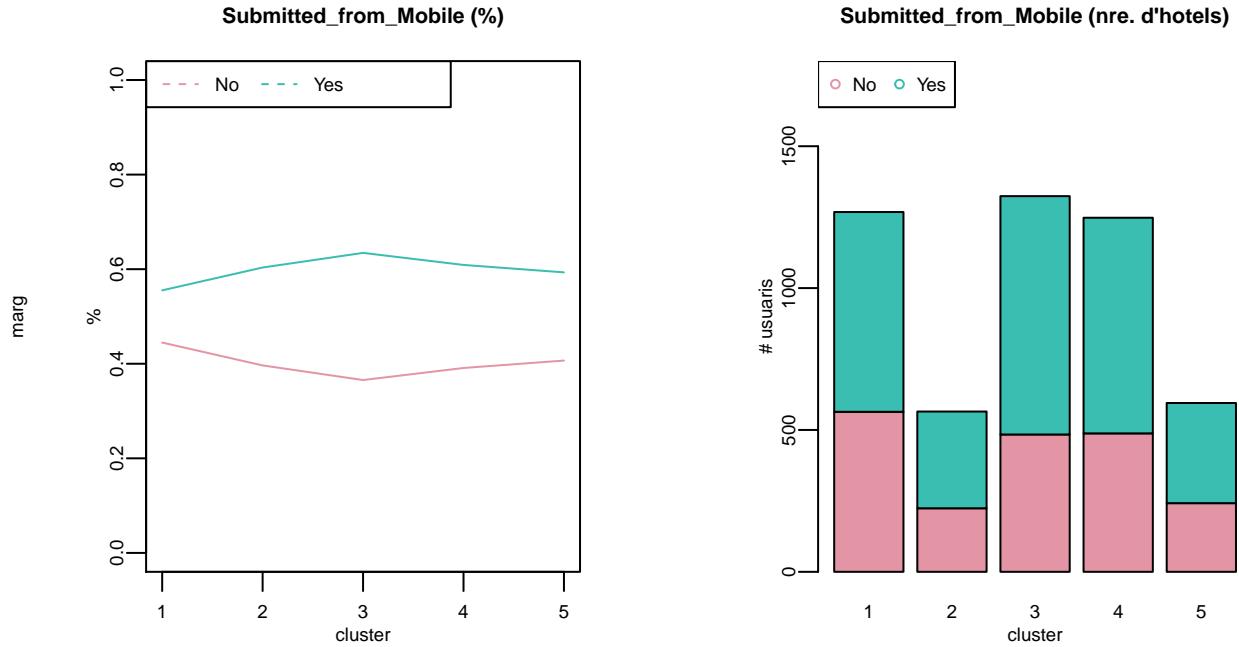


Figure 41: Profiling variable Submitted\_from\_Mobile

#### Pearson's Chi-squared test

```
data: dd[, k] and as.factor(P)
X-squared = 17.696, df = 4, p-value = 0.001415
```

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	Utilitaris a la periferia	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 seràn els més econòmics). La 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.
2	Anglesos a Barcelona	Hotels de Barcelona, majoritariament reservats per persones de nacionalitat anglesa amb presència més abundant de grups de viatgers i reservats per a llargues estades. Les valoracions són lleugerament superiors a la mitjana general però és el grup del que menys informació disposem sobre el perfil i característiques de l'hotel. El nombre de revisions i puntuacions adicionals també es baix.

Clúster	Nom	Descripció
3	Mal puntuats	Són hotels majoritàriament situats a Londres, no massa cèntrics, i els que han obtingut les valoracions més baixes (els que menys agraden als clients). La majoria de ressenyes corresponen a un turisme local i són freqüents per parelles o viatgers solitaris. Mencionar també que són els que reben més ressenyes i valoracions, tot i que els usuaris que han visitat aquests hotels són els menys actius. Estades no massa llargues
4	Ben puntuats	Són hotels majoritàriament situats a Londres, no massa cèntrics, i els que han obtingut les valoracions més altes (els que més agraden als clients). De nou reben molts comentaris tot i que els usuaris que han escrit les ressenyes d'aquest clúster no són gaire actius a Booking. En aquest grup són una mica més freqüents les famílies i les parelles, en detriment del percentatge de grups grans de viatgers. Estades no massa llargues
5	Experiència Urbana	Hotels amb valoracions mitjanes i poques valoracions, que es caracteritzen per estar situats al cor de París. Predominen famílies amb nens petits i grups (les parelles no són tan habituals).

## ACP de les variables numèriques

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

És a dir, el output que busquem és la projecció de la nostra base de dades original ( $n \times d$ ) 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 el "soroll" i 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 aquests. 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 capture 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).

Per a evidenciar la explicació anterior, proposem el següent gràfic on veiem com, sobre l'espai original es projecten els dos components principals que constitueixen els nous eixos de coordenades sobre els que rotaran les dades. La idea és aplicar aquest principi sobre les nostres dades i anar analitzant les components en plans bidimensionals.

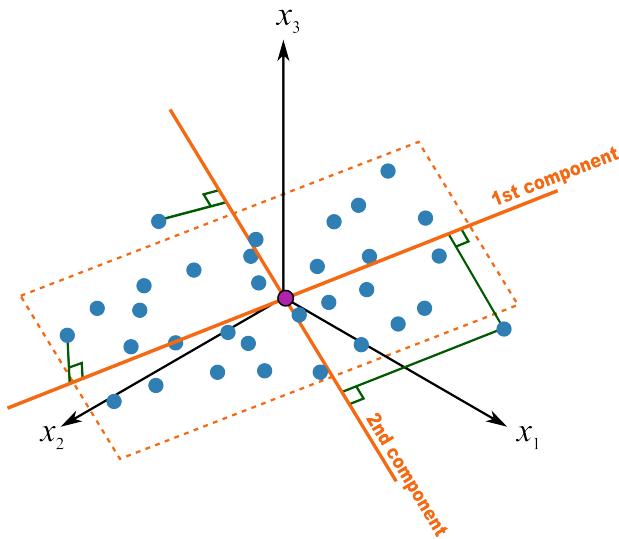


Figure 42: Exemple: Projecció de les dues primeres components principals

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'estimaràn 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'estimaràn 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 (*Figure 43*) per a veure el percentatge de variabilitat total de les dades que capturen cada dimensió (component principal).

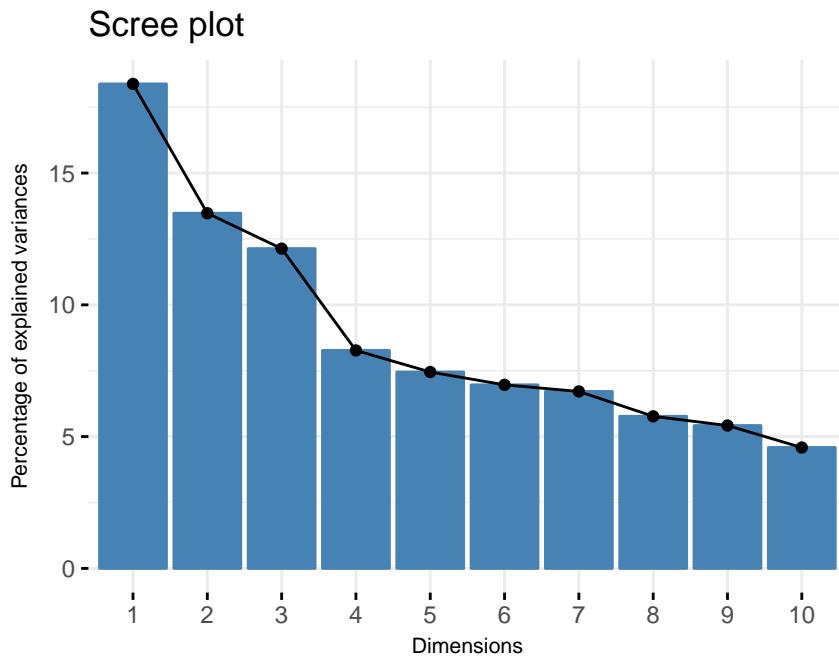


Figure 43: Percentatge de variància capturat a cada dimensió

Repetim el càlcul numèricament i observem com, per a assolir el llindar del 80% de la variància total de les dades, necessitem 8 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. El resultat és molt semblant, tant si fem servir un criteri com l'altre (8 vs. 7).

	eigenvalue	variance.percent	cumulative.variance.percent
Dim. 1	2.7574235	18.3828235	18.38282

Dim.2	2.0214369	13.4762459	31.85907
Dim.3	1.8200343	12.1335620	43.99263
Dim.4	1.2404610	8.2697403	52.26237
Dim.5	1.1176526	7.4510173	59.71339
Dim.6	1.0446726	6.9644841	66.67787
Dim.7	1.0070475	6.7136502	73.39152
Dim.8	0.8654886	5.7699242	79.16145
Dim.9	0.8130658	5.4204388	84.58189
Dim.10	0.6879154	4.5861029	89.16799
Dim.11	0.4870421	3.2469471	92.41494
Dim.12	0.4544410	3.0296066	95.44454
Dim.13	0.3100775	2.0671836	97.51173
Dim.14	0.2723073	1.8153821	99.32711
Dim.15	0.1009337	0.6728912	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)
```

- `res.var$coord` # Coordenades
- `res.var$contrib` # Contribucions als components principals
- `res.var$cos2` # Qualitat de la representació

*Resultats per als individus*

```
res.ind <- get_pca_ind(res.pca)
```

- `res.ind$coord` # Coordenades
- `res.ind$contrib` # Contribucions als components principals
- `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 `cos2` 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) (*Figure 44*). 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 homogèni, veiem com els individus amb menys representativitat prenen valors propers al centre de coordenades. A continuació construïm gràfics adicionals per a la resta de dimensions 3, 4, 5, 6 (*Figure 45*) (per abreujar l'anàlisi, tractem només les 6 primeres dimensions, on es concentra la majoria de la variabilitat).

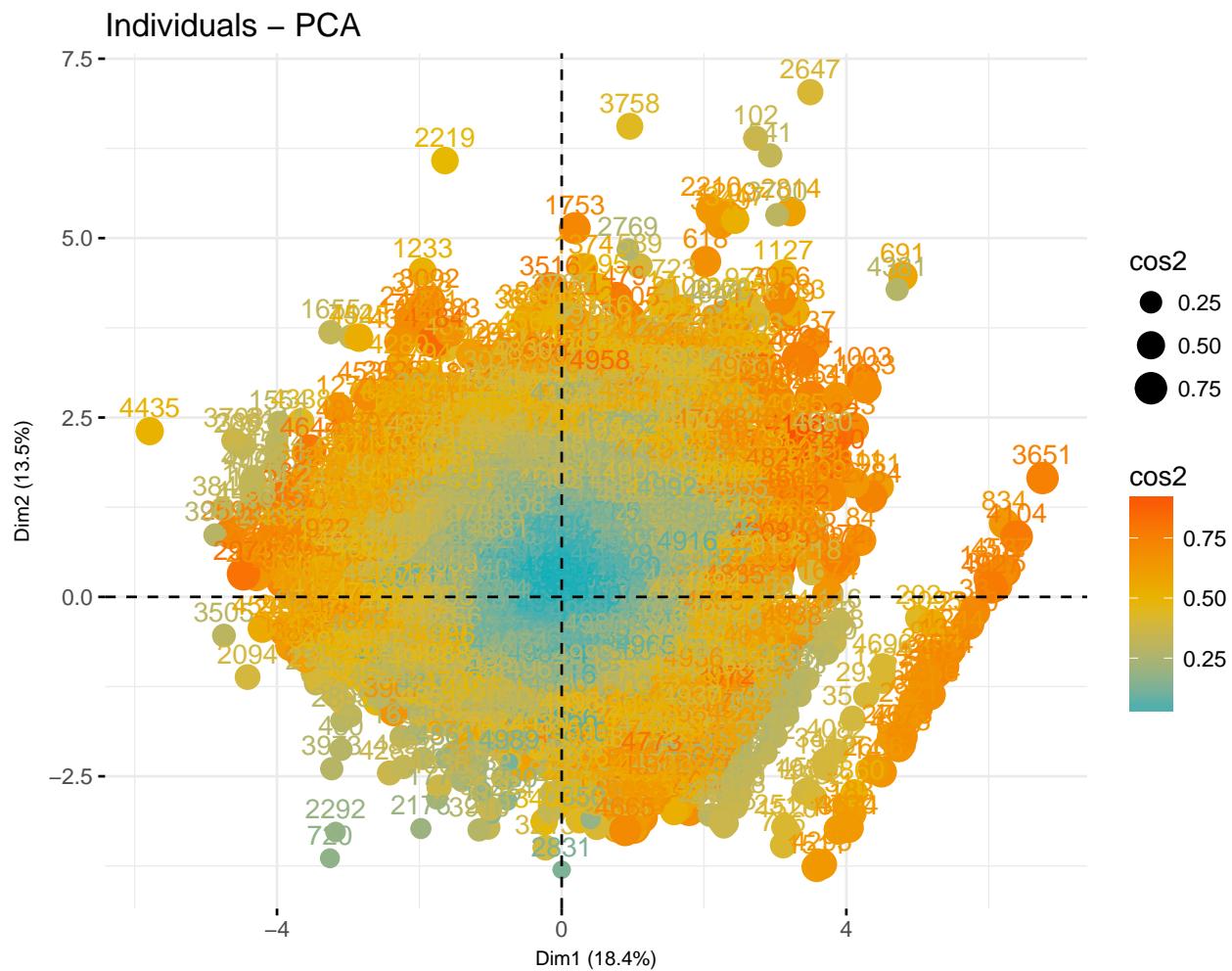


Figure 44: Representació dels individus sobre els PC (1er pla)

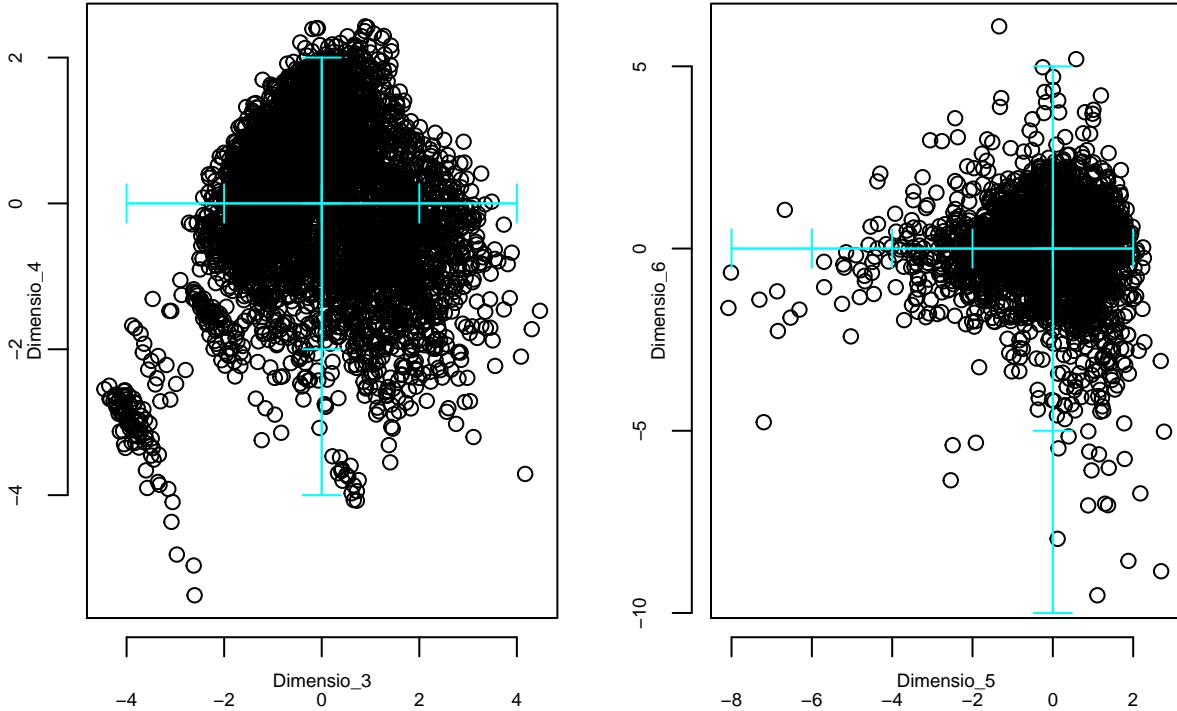


Figure 45: Representació dels individus sobre els PC

Veiem com, a mesura que ens acosetem 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 capturen menys variància). Adicionalment podem testar si aquests nous eixos construïts a partir dels components principals ens separen bé els clústers obtinguts anteriorment (*Figure 46*).

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 (*Figure 47*). 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.

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 poguem veure ràpidament les que capturen més variància de l'eix. Construïm el gràfic per al primer pla factorial (components principals 1 i 2) (*Figure 48*). 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.

En aquest primer pla veiem com les variables més rellevants són la puntuació, el grau de

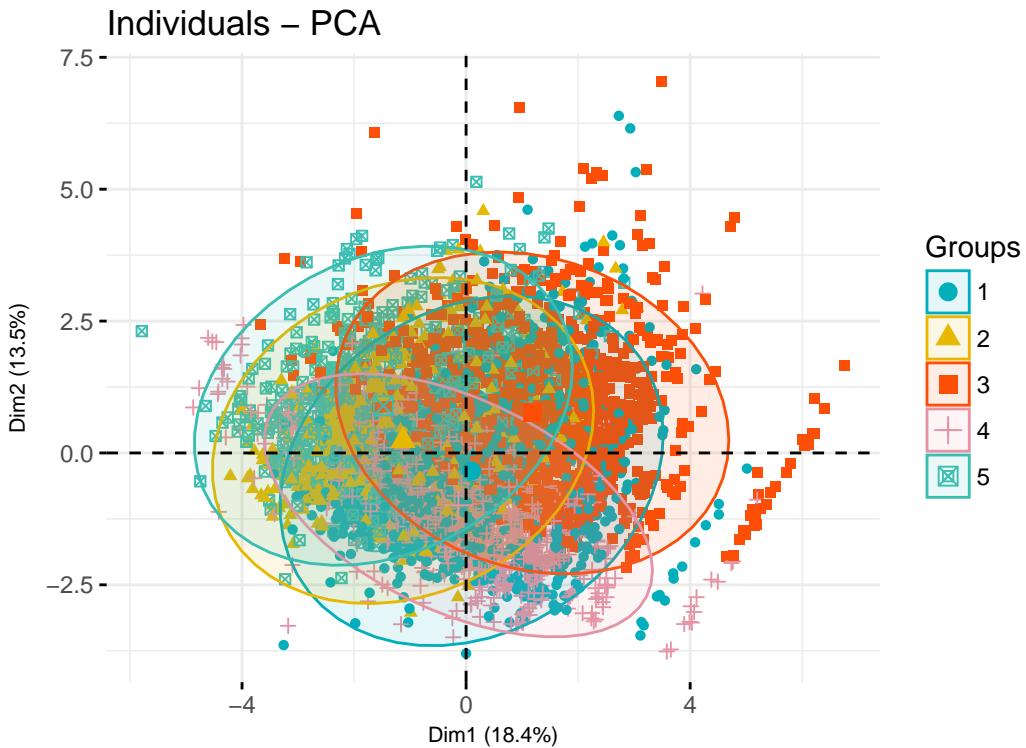


Figure 46: Clustering jeràrquic sobre el primer pla factorial

positivisme del comentari i els negocis a 1km a la rodona, tot i que no totes en la mateixa direcció. En general, les variables implicades en la valoració del hotel (Reviewer\_Score, Review\_Positivity\_Rate, Review\_Total\_Negative\_Word\_Counts...) tenen major representativitat a l'eix 1, mentre que les variables referides als negocis a la rodona i els comentaris adicionals es projecten millor a l'eix 2. En aquest sentit, podem pensar que els individus situats al tercer sector del mapa de coordenades (valors negatius per a PC1 i PC2) són els que han obtingut millors valoracions. De la mateixa manera, és lògic que aquesta direcció sigui contrària a la que mostra la variable *Review\_Total\_Negative\_Word\_Counts*.

A continuació, i com en el cas anterior, construim gràfics adicionals per a les dimensions 3, 4, 5 i 6 (Figure 49).

En general, les variables ben representades a la Dimensió 2, també ho estan a la dimensió 3, en especials els comptatges de comentaris adicionals. D'aquest primer gràfic també es desprèn una alta representativitat de les latituds i longituds 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 *Total\_Number\_of\_Reviews\_Reviewer\_Has\_Given* i *Stay\_Duration*.

Adicionalment, el software ens permet crear una espècie de clúster de les variables numèriques fent servir kmeans (Figure 50). 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).

L'agrupació sembla llògica 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

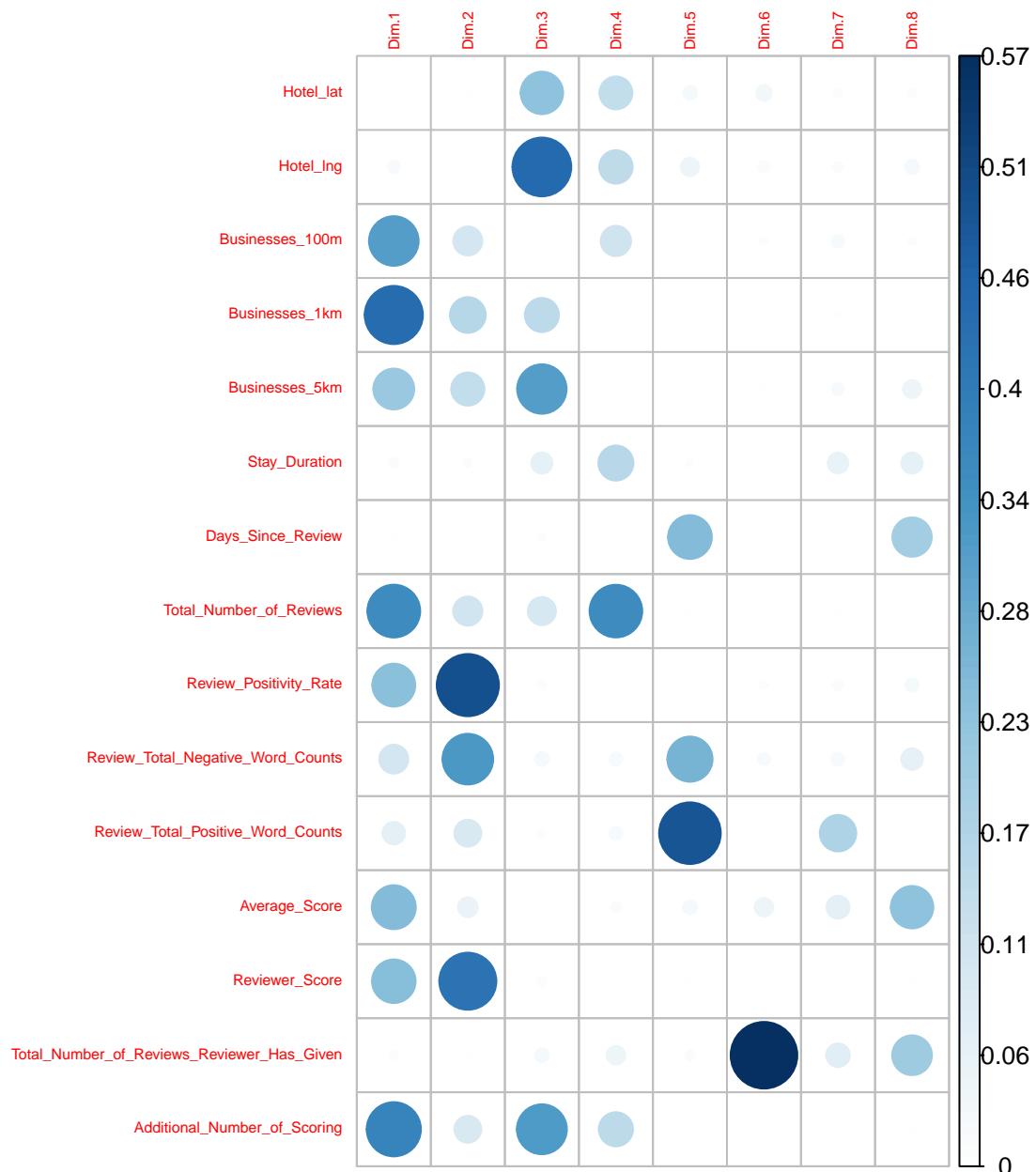


Figure 47: Representativitat de les variables a cada dimensió

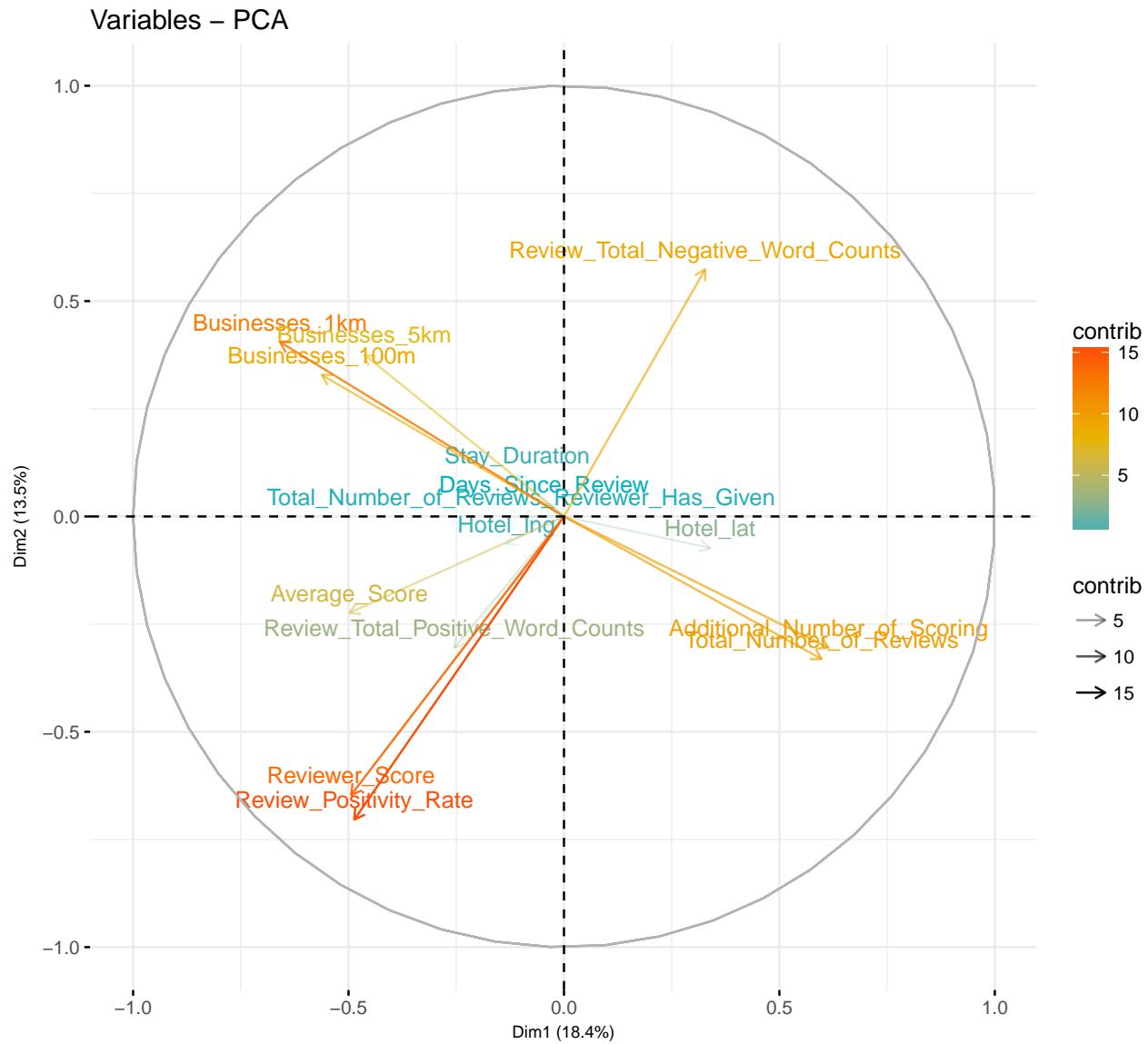


Figure 48: Representació de les variables sobre els PC(1er pla)

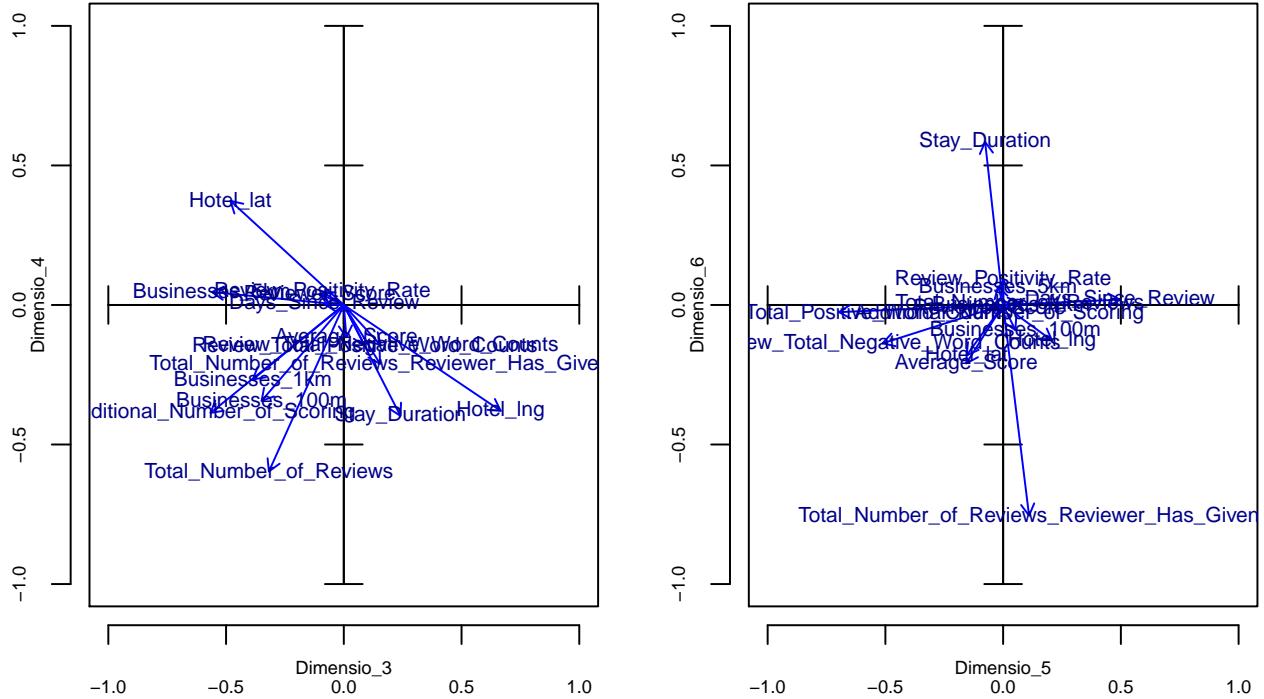


Figure 49: Representació de les variables sobre els PC

es desplacen conjuntament en aquest nou subespai.

Un cop tractades les variables numèriques actives a l'anàlisi, no podem olvidar 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. Començem fent una representació senzilla de les categories de la variable *Hotel\_Country* en el primer pla factorial (*Figure 51*). Més endavant proposarem construir un gràfic amb tots els nivells de totes les variables categòriques, de manera que poguem identificar intuitivament relacions entre variables als plans factorials.

Veiem com, en general, les categories es troben força a prop del centre de coordenades i no podem inferir gaire cosa. Tanmateix, la modalitat “França” es troba en la direcció de *Businesses\_1km*. Aquest resultat concorda amb el que hem vist en l’etapa de profiling, ja que els hotels més urbans (cluster 5) és troben majoritàriament a França.

A continuació generem un gràfic amb totes les modalitats de totes les variables categòriques (*Figure 52*). Si tenim molts nivells en total, com és el nostre cas, podem desdoblar el gràfic per a poder apreciar millor els nivells i evitar solapaments. D'aquesta manera, la interpretació serà més fàcil, i de més qualitat.

La interpretació és idèntica al cas de una sola variable, veiem com les modalitats de França i París es troben al segon sector de la circumferència. En general, el gràfic ens confirma les relacions que hem establert al profiling anterior ja que, per exemple la nacionalitat britànica i els hotels de Londres ocupen espais molt propers, els valors positius per a la variable *Review\_Is\_Positive* es troben al tercer sector que concorda amb la direcció de creixement de les variables *Reviewer\_Score* i *Review\_Positivity\_Rate*, etc. A continuació, repetim el procediment per a les dimensions tercera i quarta (*Figure 53*).

Aquest segon gràfic mostra bona representació per a modalitats com “Grecia”, “Israel”, “Barcelona” o “Milan” que abans consideràvem inertes. En aquest cas hem projectat les modalitats únicament sobre les 4 primeres dimensions que, com sabem, són les capturen major variabilitat de les dades.

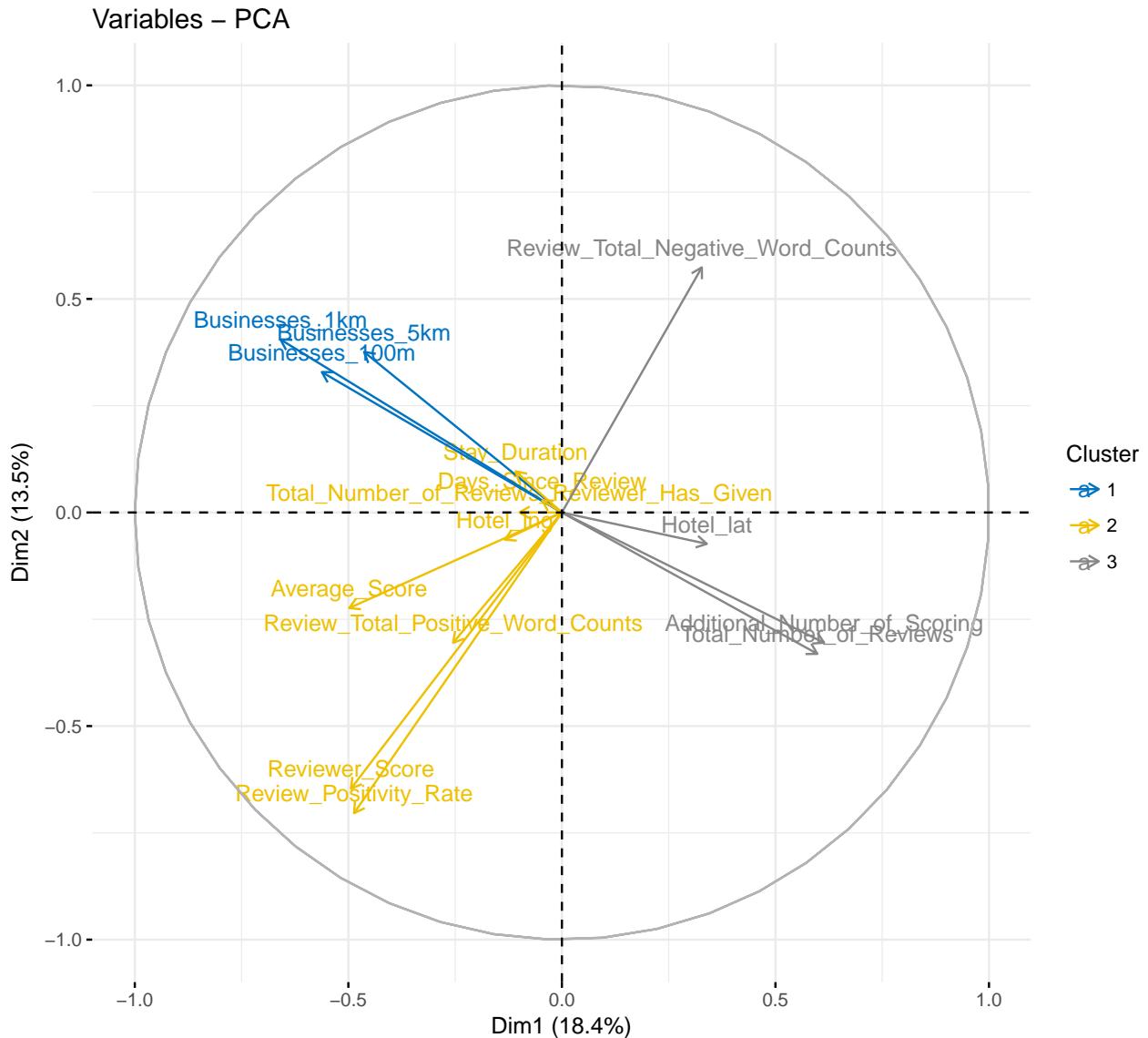


Figure 50: Representació de les variables sobre els PC(1er pla)

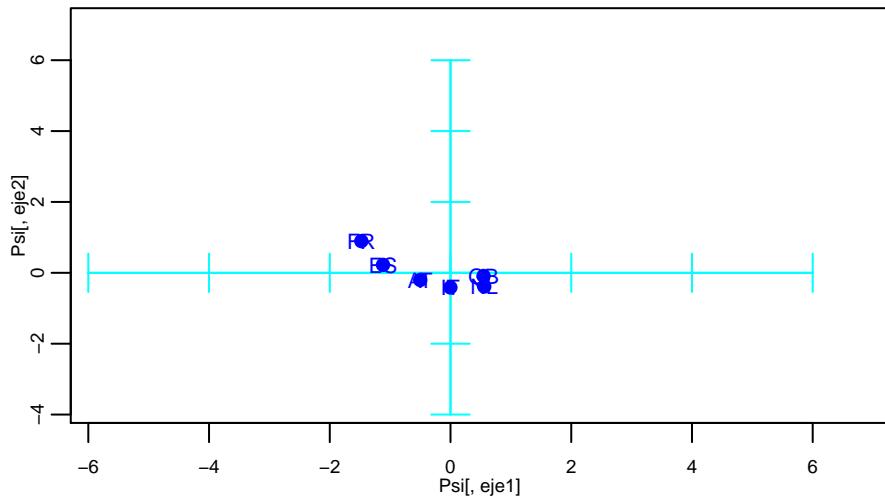


Figure 51: Hotel\_Country al primer pla factorial

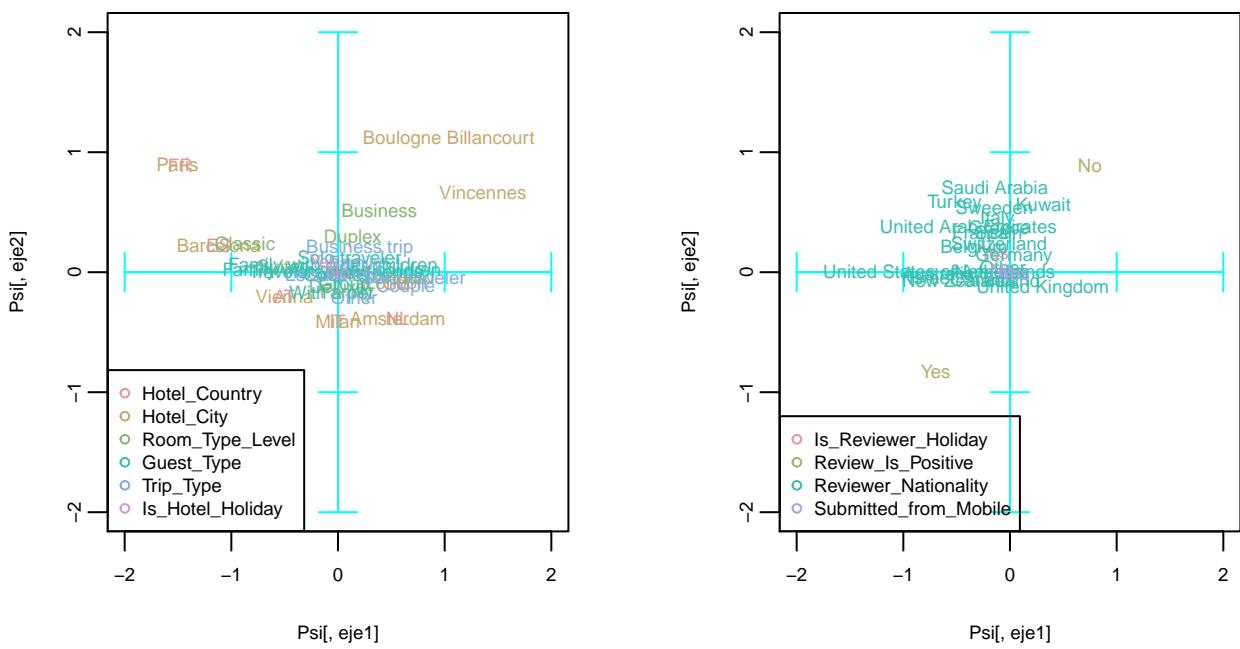


Figure 52: Projeccions de totes les modalitats sobre Dim.1 i Dim.2

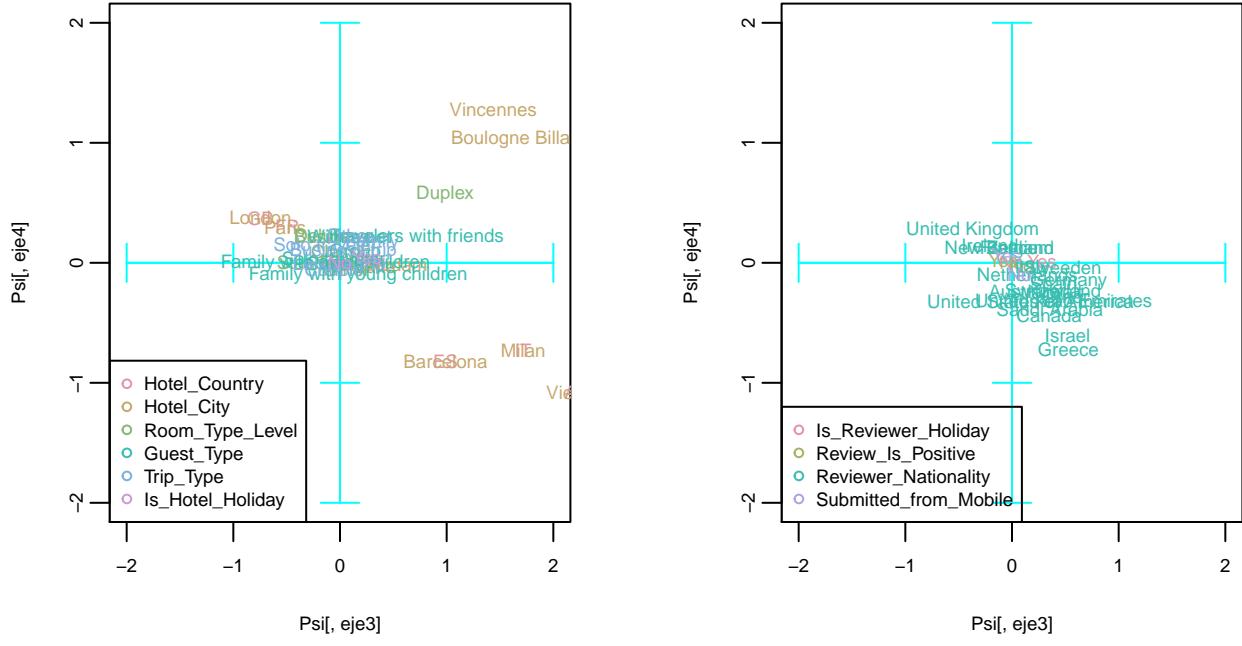


Figure 53: Projeccions de totes les modalitats sobre Dim.3 i Dim.4

Tanmateix, de cara a estudis futurs podria ser interessant representar les modalitats en un major nombre de dimensions. Si anem una mica més enllà, el més interessant tenint l'input dels gràfics anteriors, és identificar possibles correlacions entre modalitats de les variables categòriques i les projeccions de les variables numèriques sobre els plans factorials (l'espai sobre el que projectem és el mateix i per tant són perfectament comparables). 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 la 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 (*Figure 54*). Generalment, es comú fer servir un factor d'escala per a poder tenir una bona visualització dels nivells de les variables qualitatives i les direccions de les numèriques. En aquest sentit, per prova i error hem detectat que un factor corrector de 3 unitats s'adapta bé a la dimensionalitat de les nostres dades. En aquesta línia d'aconseguir representacions clares i fàcils d'interpretar, hem dividit de nou les variables categòriques en dos conjunts.

#### *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 dels 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 (*Figure 55*).

Observem com, en general, les categories es troben força sobreposades. La distinció més clara la trobem en la variable *Review\_Is\_Positive*, on clarament veiem com, valors negatius per a les components principals 1 i 2, estan altament correlacionats amb valoracions positives (i valors alts per a les variables *Review\_Positivity\_Rate*, *Reviewer\_Score*, *Review\_Total\_Positive\_Word\_Counts*...). Podem fer distincions lleus, en relació al primer gràfic on es pot apreciar subtilment que les modal-

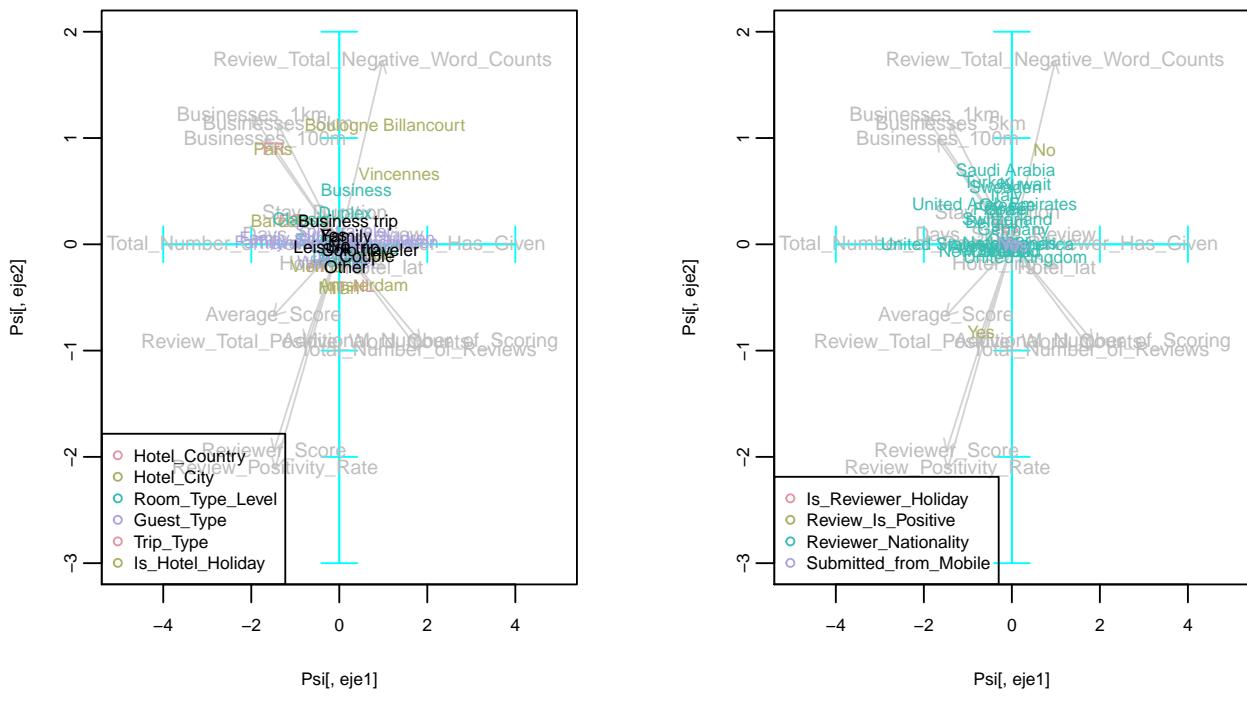
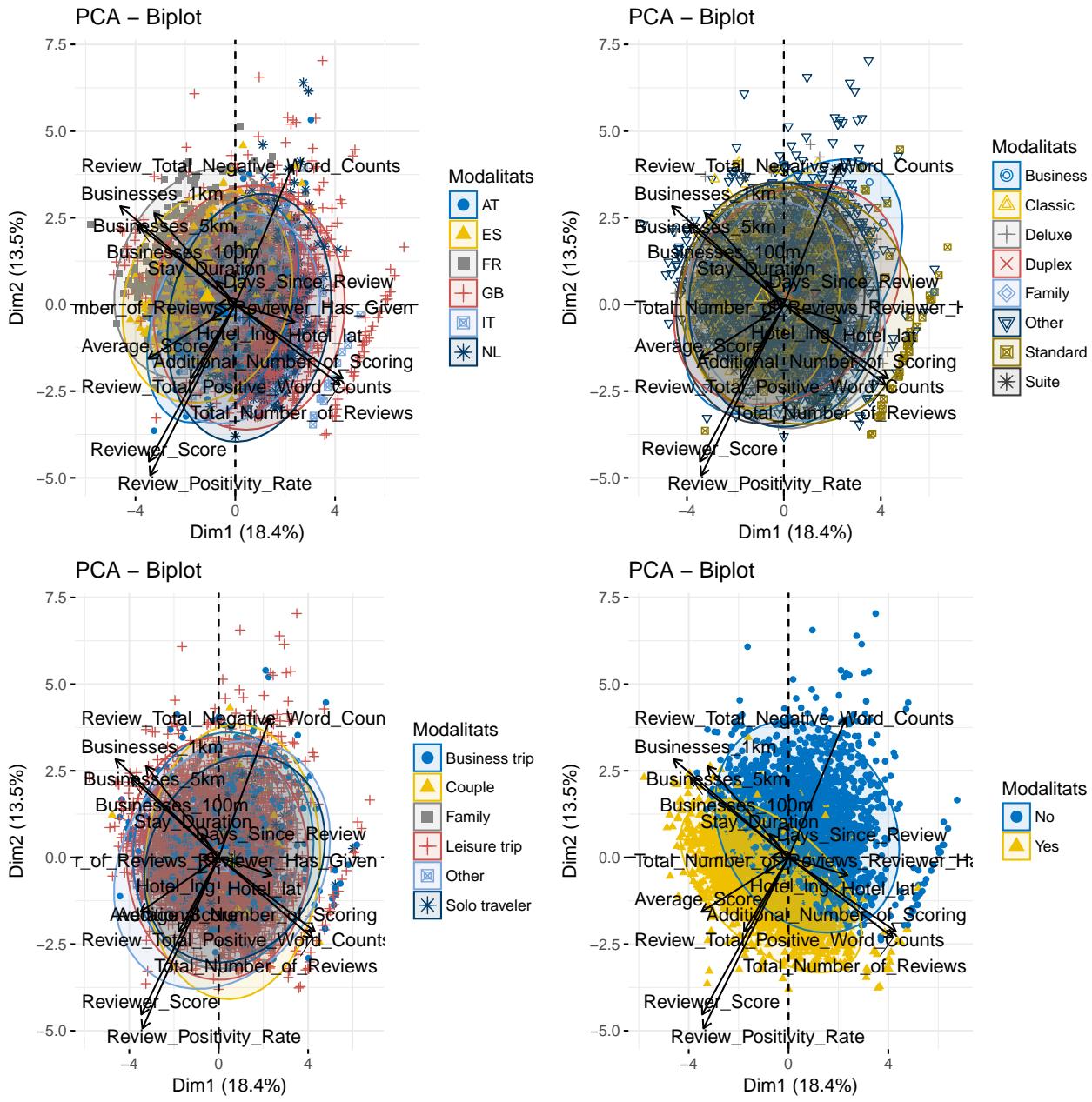


Figure 54: Projeccions de les modalitats amb variables numèriques

itats “França” i “Gran Bretanya” es corresponen amb el segon i quart sector del gràfic. En aquest sentit, podem veure com hotels ubicats a França seran més urbans (valors alts per a variables Businesses), mentre que els hotels anglesos tindran més ressenyes i valoracions addicionals.



## ACM de les variables qualitatives

L'anàlisi de correspondència múltiple (ACM) és una extensió de l'anàlisi de correspondència simple per tal de resumir i visualitzar una taula de dades que contingui més de dues variables categòriques. L'ACM també s'usa com una generalització de l'anàlisi de components principals quan les variables a analitzar són categòriques en lloc de quantitatives.

L'ACM s'ha utilitzat generalment per analitzar un conjunt de dades d'una enquesta.

L'objectiu de l'ACM és identificar:

- Grups d'individus amb un perfil semblant respecte a la resposta a les variables categòriques.
- Associacions entre categories de les variables.

La funció que usarem per calcular l'ACM s'anomena MCA i en aquesta funció indiquem:

- dd: la nostra base de dades.
- quanti.sup: Les variables quantitatives de la nostra base de dades.

Per tal de mirar la proporció de variància explicada per les diferents dimensions utilitzarem els valors propis.

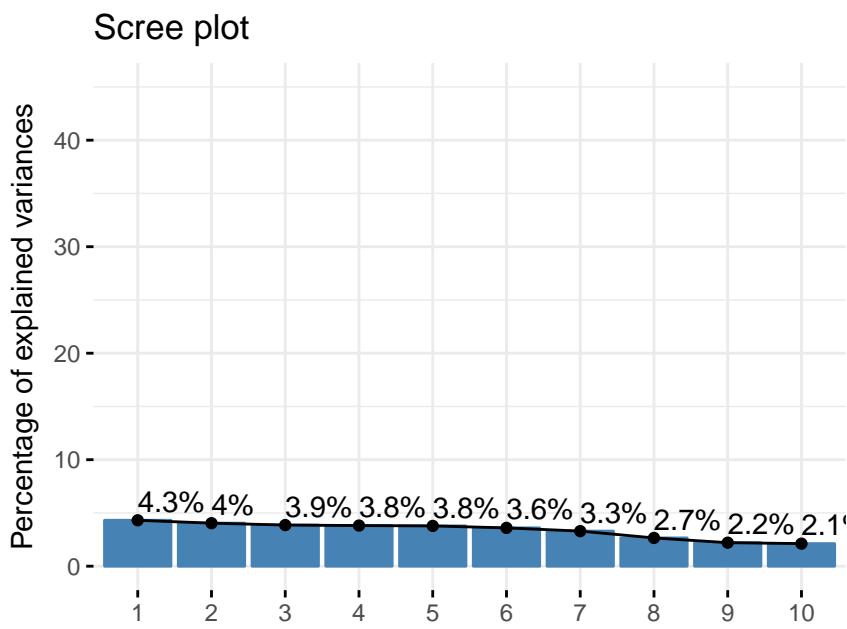


Figure 56: Scree Plot ACM

Mitjançant un Scree Plot (Figure 56) podem veure la variància acumulada per a cada una de les dimensions. Com podem apreciar a la cinquena dimensió arribem a una variància acumulada del 19.81%, per tal de seguir amb l'anàlisi ens quedarem amb aquestes 5 dimensions.

	eigenvalue	variance.percent	cumulative.variance.percent
Dim. 1	0.2327627	4.310421	4.310421
Dim. 2	0.2182383	4.041450	8.351871
Dim. 3	0.2085977	3.862921	12.214792
Dim. 4	0.2061175	3.816991	16.031783
Dim. 5	0.2042524	3.782453	19.814235

Seguidament podem veure la contribució tant de les 5000 observacions com de les categories de cada variable qualitativa a les 5 dimensions que hem definit per a analitzar.

	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5
1	0.041320772	0.0106688161	0.1044179096	0.050221085	3.526090e-02
2	0.006131659	0.1096921970	0.0104976734	0.001646055	2.369746e-03
3	0.001443370	0.0057330754	0.0008151055	0.107302782	2.902581e-02
4	0.025214746	0.0005758638	0.0000299075	0.002257324	2.415743e-04
5	0.008517406	0.0674088975	0.0075838167	0.004965617	1.414784e-06

	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5
AT	7.2389412	0.3778540	2.903169	2.5482064	28.7019965
ES	0.6670875	25.3473991	4.457692	0.9876952	0.5430264
FR	8.5436786	10.2056219	16.484967	0.8212250	1.9947729
GB	15.7270586	1.9097145	0.294194	1.1674542	0.3918830
IT	5.7452202	0.5992421	16.483260	10.2924863	6.5521884

Per tal de poder treure millors conclusions sobre les contribucions de cada modalitat a les definicions de cada dimensió realitzarem un plot on poder interpretar els resultats anteriors. A continuació, fem un petit recordatori on es mostra el nombre de modalitats de cada variable qualitat que ha intervengut en l'anàlisi.

Hotel_Country	Hotel_City	Room_Type_Level
6	8	8
Guest_Type	Trip_Type	Is_Hotel_Holiday
7	6	2
Is_Reviewer_Holiday	Review_Is_Positive	Reviewer_Nationality
2	2	21
Submitted_from_Mobile		
2		

Al següent gràfic (*Figure 57*) podem veure representades cada modalitat i, com hem dit, la seva contribució a les dues dimensions proposades. Mitjançant aquest plot podem identificar les variables que estan més correlacionades amb les dues dimensions.

Es pot observar com les modalitats de Londres i Regne Unit són les més correlacionades amb la dimensió 1 (aporten més a la dimensió 1) i Barcelona i Espanya les més correlacionades amb la dimensió 2. Les altres modalitats aporten menys, en comparació amb aquestes, i hauríem de mirar-ho a la taula de contribucions per saber quant aporten exactament.

Per últim, també podem fixar-nos en les coordenades dels individus i de les modalitats sobre el pla factorial.

	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5
1	0.6934672	0.34119967	-1.04358373	-0.7194249	0.600088496
2	0.2671349	1.09405298	0.33089205	-0.1302461	-0.155567725
3	-0.1296076	0.25011763	-0.09220335	1.0515936	0.544453483
4	-0.5417127	-0.07927028	-0.01766159	-0.1525244	-0.049669985
5	0.3148440	0.85764803	0.28124429	-0.2262190	-0.003801139

## ACM: Contribució modalitats a les dimensions

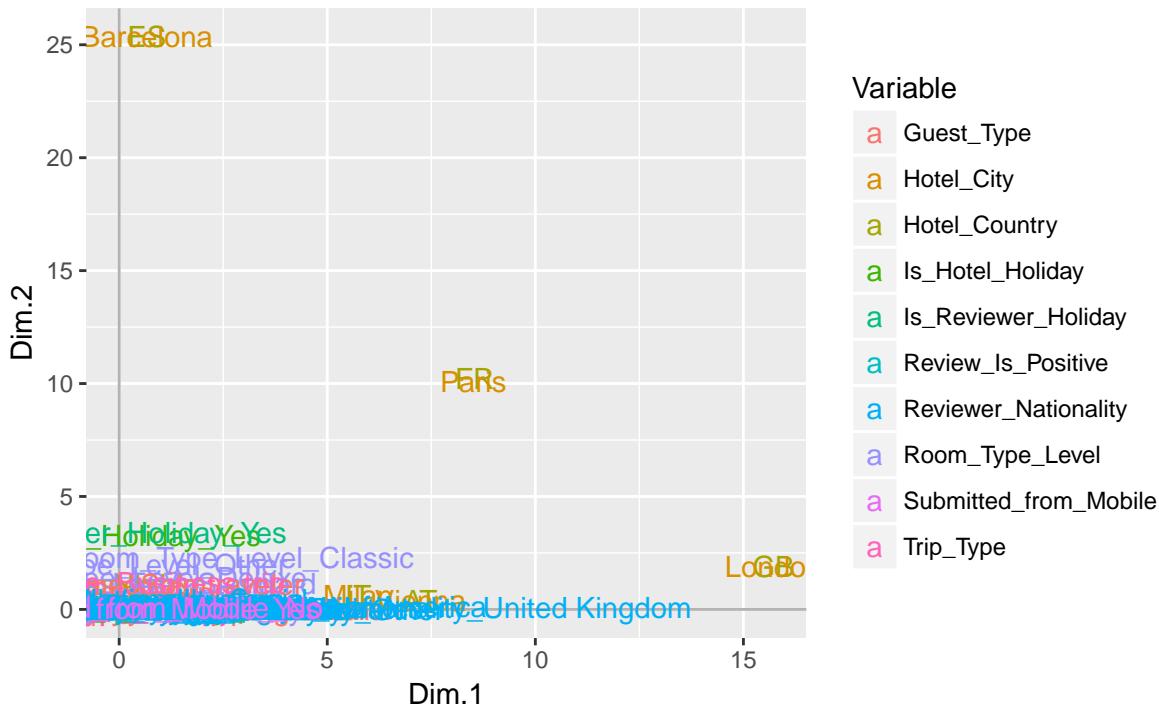


Figure 57: Contribucions de les modalitats sobre els PC(1er pla)

	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5
AT	1.4968742	-0.3311453	-0.8973917	0.8357297	-2.7920987
ES	0.3677707	2.1951350	0.8999925	-0.4211124	-0.3108302
FR	1.2927236	-1.3680808	1.6999078	-0.3771508	0.5851361
GB	-0.8465547	-0.2856434	-0.1096090	-0.2170461	-0.1251804
IT	1.3962011	0.4366209	-2.2387949	-1.7585533	1.3967374

Mitjançant aquestes taules, podem observar les coordenades per a cada registre i modalitat sobre el pla factorial.

Mitjançant el següent gràfic (*Figure 58*), igual que anteriorment amb la contribució, podem millorar la taula prèvia. Per tal de poder interpretar aquest plot hem de tenir en consideració que:

- Les modalitats de les variables amb un perfil similar s’agrupen.
- Les modalitats de les variables negativament correlacionades se situen els quadrants opositats.
- La distància entre els punts de la modalitat i l’origen mesura la qualitat de la categoria variable en el mapa de factors. Els punts de modalitat que estan allunyats de l’origen estan ben representats en el mapa de factors.

Com podem veure hi ha modalitats ben allunyades de l’origen són entre altres Room\_type\_level\_duplex, trip\_type\_other, Vicennes, Paris, Billancourt, Vienna, i per tant són algunes de les modalitats millor representades en el mapa de factors.

Per altra banda podem observar que Vicennes, Paris, Billancourt, es troben agrupades, això vol dir que tenen característiques similars (cosa que té sentit, ja que totes formen part de la regió de París).

## ACM coordenades modalitats

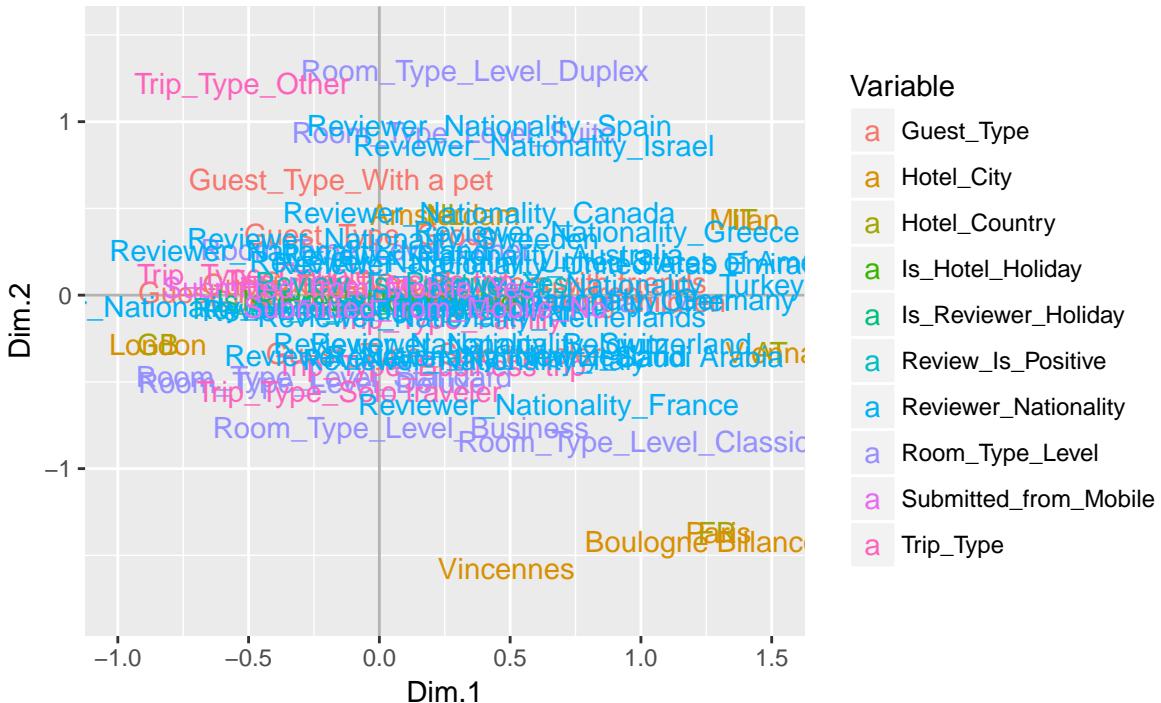


Figure 58: Representació de les modalitats sobre els PC(1er pla)

Per últim en el següent gràfic, a més de les modalitats també podem observar les coordenades de les observacions (*Figure 59*). D'aquest mode podem observar diferents núvols de punts representats en el plot. Per tal de relacionar les observacions amb les modalitats, podem interpretar que els núvols de punts situats juntament amb les modalitats tenen les característiques d'aquelles mateixes categories.

Seguidament analitzarem l'ACM amb la combinació d'altres dimensions que no siguin la 1 i la 2. En primer lloc realitzarem l'anàlisi entre la dimensió 1 i 3, mitjançant els gràfics anteriors podrem extreure a les mateixes conclusions que abans sobre aquestes noves dimensions, i saber quines modalitats i/o observacions contribueixen més a l'ACM (*Figure 60*).

Veiem com, lògicament, les modalitats que més aporten a la dimensió 1 són les mateixes que en el primer gràfic. Pel que fa a la dimensió 3 veiem que les que estan millor representades són Milà (Itàlia) i París (França). També podem destacar Barcelona i que la Review\_date sigui festiu a la ciutat de l'allotjament. Per tal d'entrar a analitzar més modalitats hauríem de mirar a la taula de contribucions.

Mitjançant la taula de coordenades, podem observar aquestes contribucions per a cada registre i, de manera anàloga al estudi de les dimensions 1 i 2, podem representar cada modalitat sobre el pla factorial (*Figure 61*).

Tal com hem vist abans les modalitats que es troben agrupades tenen un perfil similar, també podem interpretar que les modalitats situades en els quadrants opositos estan correlacionades negativament i una de les conclusions que podem extreure per exemple, és que els hostes suïssos solen viatjar sense mascotes. Per últim podem apreciar que les categories més allunyades de l'origen de coordenades seran les que estiguin millor representades en el mapa de factors (*Figure 62*).

El gràfic que compara les modalitats amb les coordenades de les observacions, no té sentit repetir-lo, ja que en aquests no intervenen les dimensions i per tant la seva representació, inde-

## AMC coordenades modalitats i registres

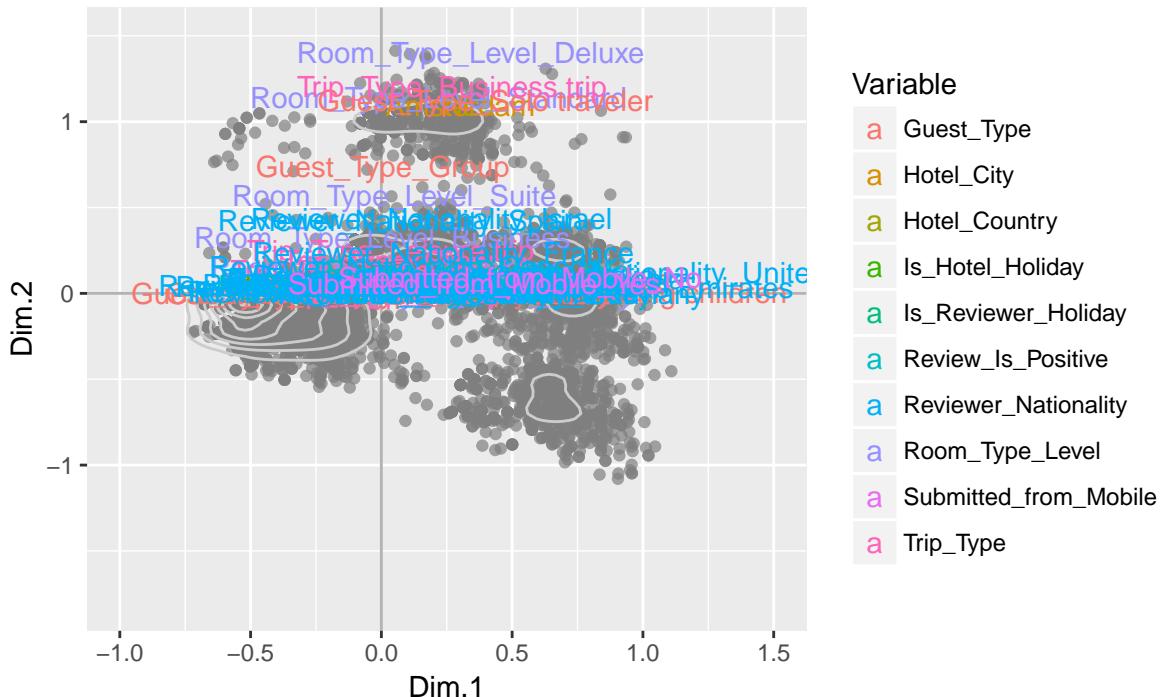


Figure 59: Representació de les modalitats i el núvol d'individus sobre els PC(1er pla)

## AMC Contribució modalitats a les dimensions

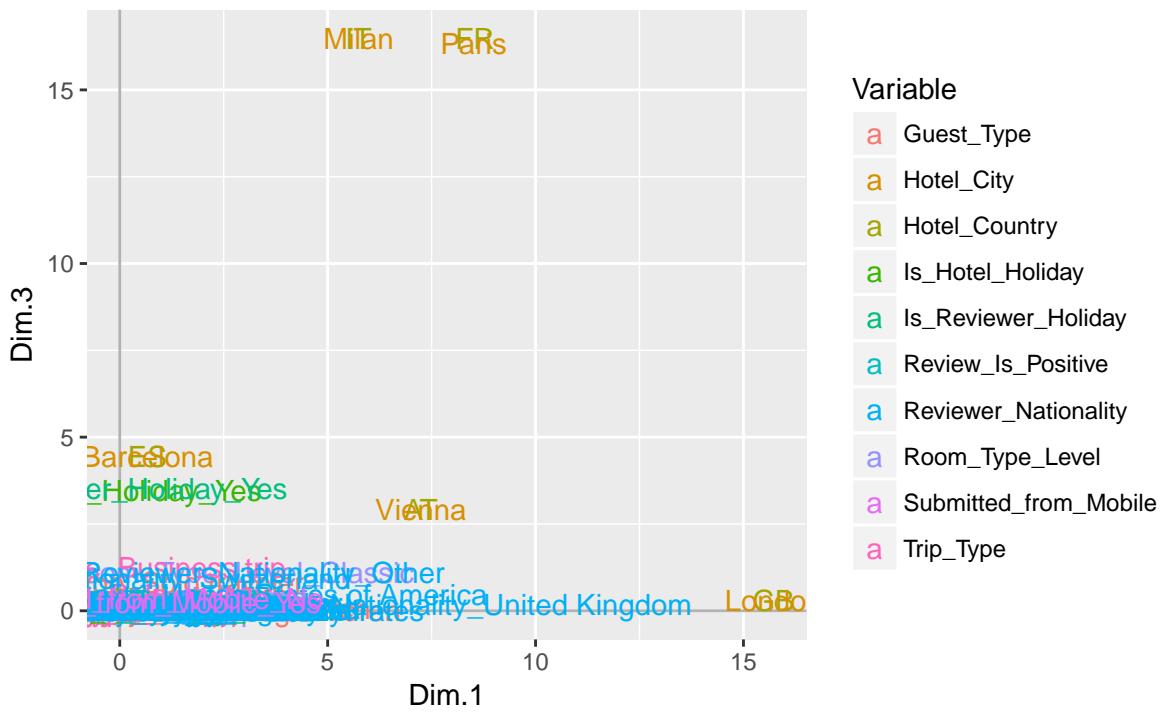


Figure 60: Contribucions de les modalitats sobre els PC(Dim.1 vs Dim.3)

### AMC coordenades modalitats

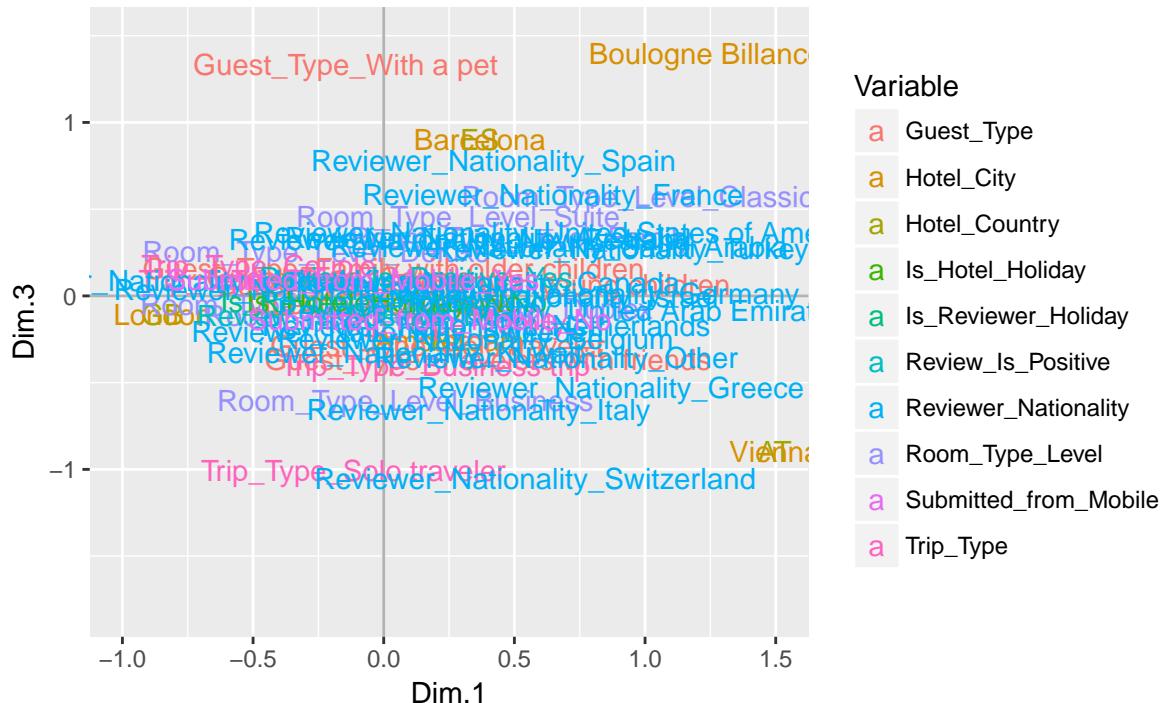


Figure 61: Representació de les modalitats sobre els PC(Dim.1 vs Dim.3)

### AMC coordenades modalitats i registres

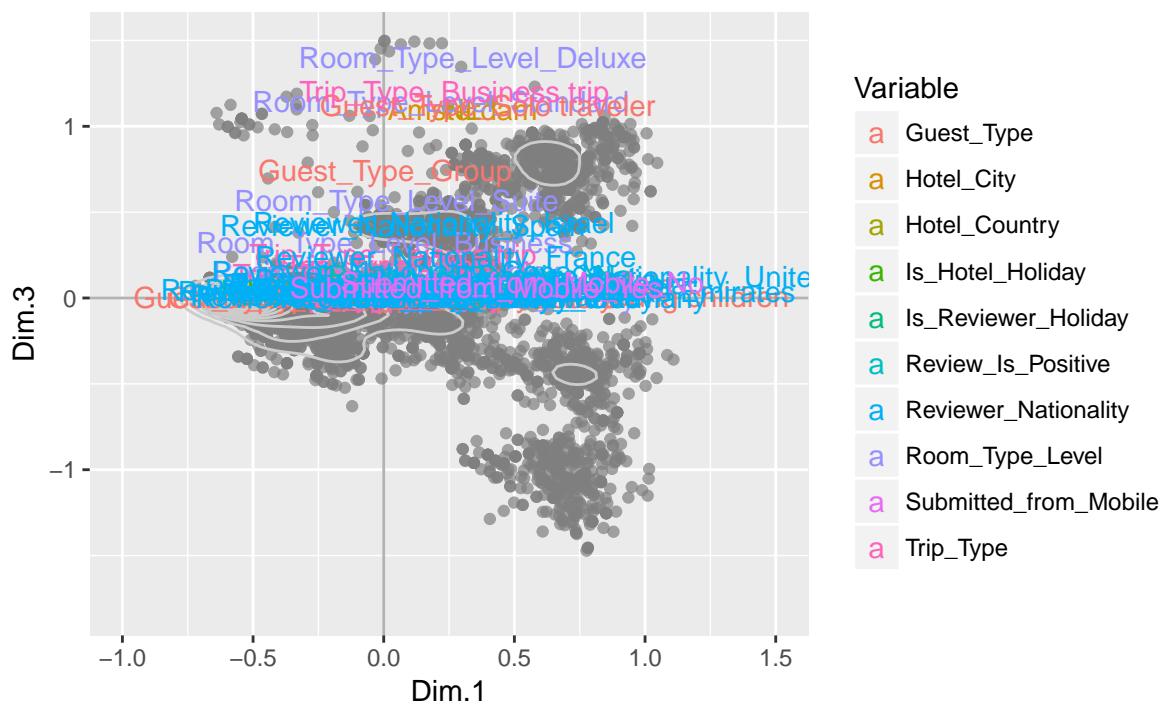


Figure 62: Representació de les modalitats i el núvol d'individus sobre els PC(Dim.1 vs Dim.3)

pendentment de les dimensions escollides, serà la mateixa.

Tal com hem pogut observar en l'anàlisi de les dimensions 1,2 i 3, sigui quina sigui la combinació entre elles, s'acaben treien les mateixes conclusions sobre quina modalitat està millor representada a cada dimensió (ja que les representacions sobre els eixos són les mateixes). Si a simple vista no es veu quina modalitat contribueix més a cada dimensió, podem fixar-nos en les taules de coordenades i contribucions per esclarir els dubtes. Per a seguir avançant, ara ens centrarem les dues dimensions que ens queden per analitzar.

Repetim el gràfic de contribucions, per a cada modalitat (*Figure 63*)

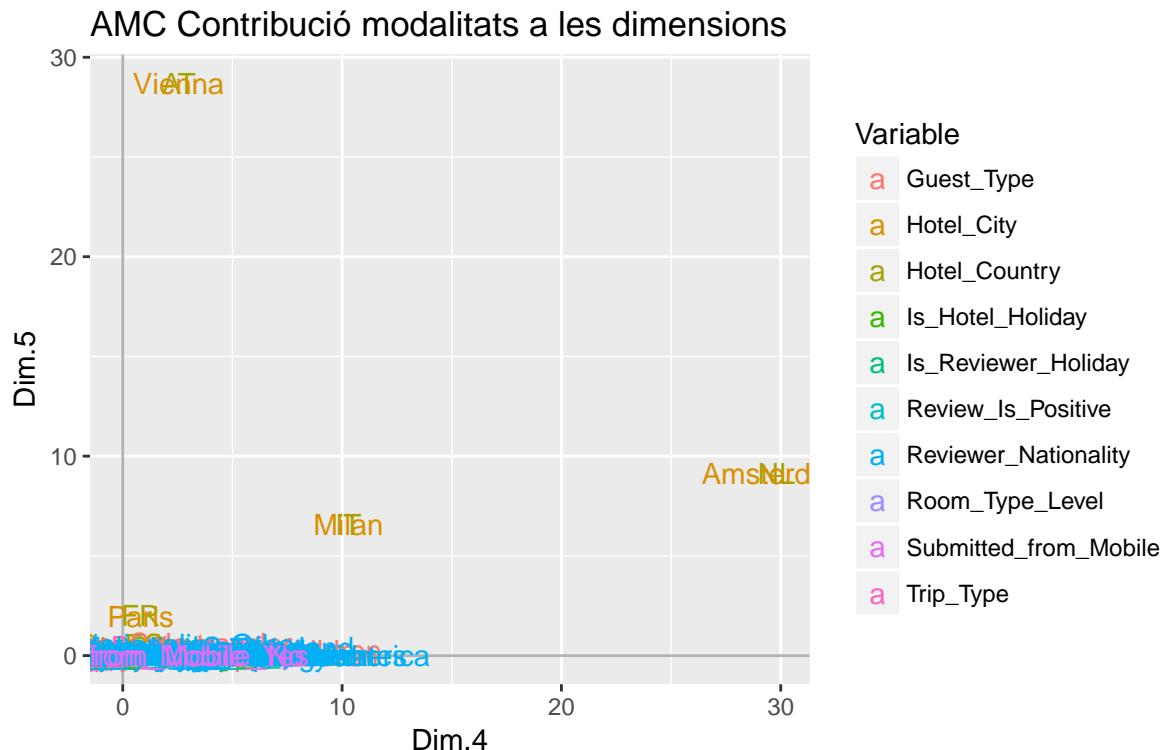


Figure 63: Contribucions de les modalitats sobre els PC(Dim.4 vs Dim.5)

Tal com es pot veure en aquest gràfic, les modalitats que més aporten a la dimensió 4 són Amsterdam (Holanda) i Milà (Itàlia). Entre les altres categories a simple vista no es pot diferenciar quina té una major contribució. En canvi, a la dimensió 5 és Viena (Aòstria) és la millor representada.

Per a fer aquest resultat més visual i, com en els casos anteriors, dibuixem les modalitats sobre els PC (*Figure 64*)

Mencionar que, de nou, en aquest gràfic trobem les modalitats que tenen un perfil similar agrupades. Podem observar com, per exemple, la modalitat Nationality\_Israel està correlacionada negativament amb ciutat de l'hotel Billancourt, ja que aquestes es troben en quadrants opositius i les modalitats més allunyades de l'origen de coordenades seran les que estiguin millor representades en el mapa de factors.

Com a conclusió de l'anàlisi de correspondències múltiples, podem extreure que les variables Hotel\_city i Hotel\_country són les que tenen les modalitats millor representades a les 5 dimensions estudiades i, en conseqüència, les més rellevants a l'hora de diferenciar entre individus.

## AMC coordenades modalitats

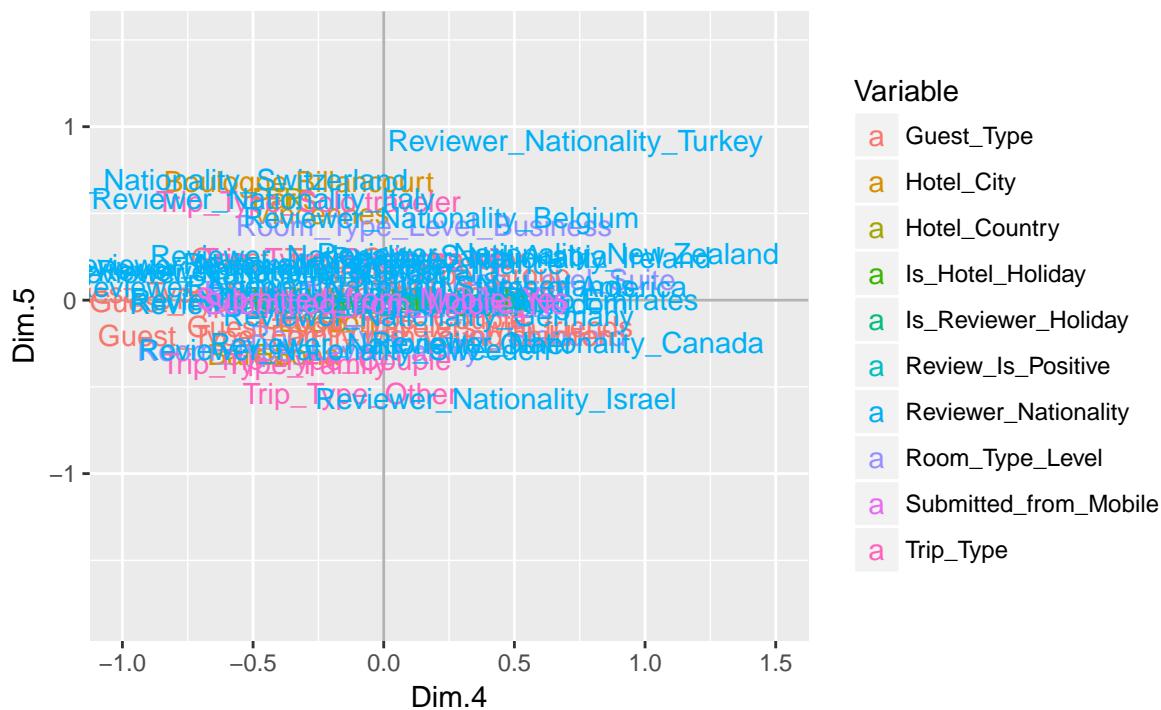


Figure 64: Representació de les modalitats sobre els PC(Dim.4 vs Dim.5)

## Clustering jeràrquic sobre les components factorials retingudes a l'ACP i a l'ACM

Aquesta secció té com a objectiu replicar el procés de clustering jeràrquic de la secció 4 però, en comptes d'executar l'algoritme de clustering sobre les dades originals, fer servir com a input el resultat de l'anàlisi de components principals.

En aquest sentit, comencem replicant el procediment de l'anàlisi de components principals, ja que necessitarem la matriu de coordenades dels individus sobre els components principals per a implementar el càlcul de les distàncies. En aquest cas, hem fet servir la llibreria FactoMineR i hem tallat en les 8 dimensions que, ja hem vist, capturen aproximadament el 80% de la variabilitat de la matriu de les nostres dades.

Com a recordatori, proposem un la realització d'un petit resum gràfic per individus i per variables dels resultats de l'ACP (*Figure 65*).

A continuació, procedim amb el clúster en si. En primer lloc, hem d'extreure les coordenades dels individus per tots els components principals i calcular les distàncies entre ells amb la funció `dist` (per defecte calcula distàncies euclidianes).

Un cop tenim la matriu de distàncies  $D$ , la fem servir com a input per al càlcul del clúster jeràrquic i realitzem els mateixos passos que en la secció 4 que, recordem empra el mètode de Ward (*Figure 66*).

Obsevem com, en vista de la distància agregada en cada iteració, podriem considerar un nombre de clústers entre 5 i 10. Per a concretar una mica més, podem representar el dendograma (*Figure 67*) i veure quin d'aquests talls s'adapta més al resultat del clustering, d'un mode més heurístic.

A partir del dendograma, hem considerat tallar en 5 clústers, igual que en el clustering que hem fet anteriorment sobre les dades originals. Si recordem el procediment anterior, un cop hem fet efectiva la partició, procedim a analitzar la qualitat d'aquesta, calculant la suma de quadrats

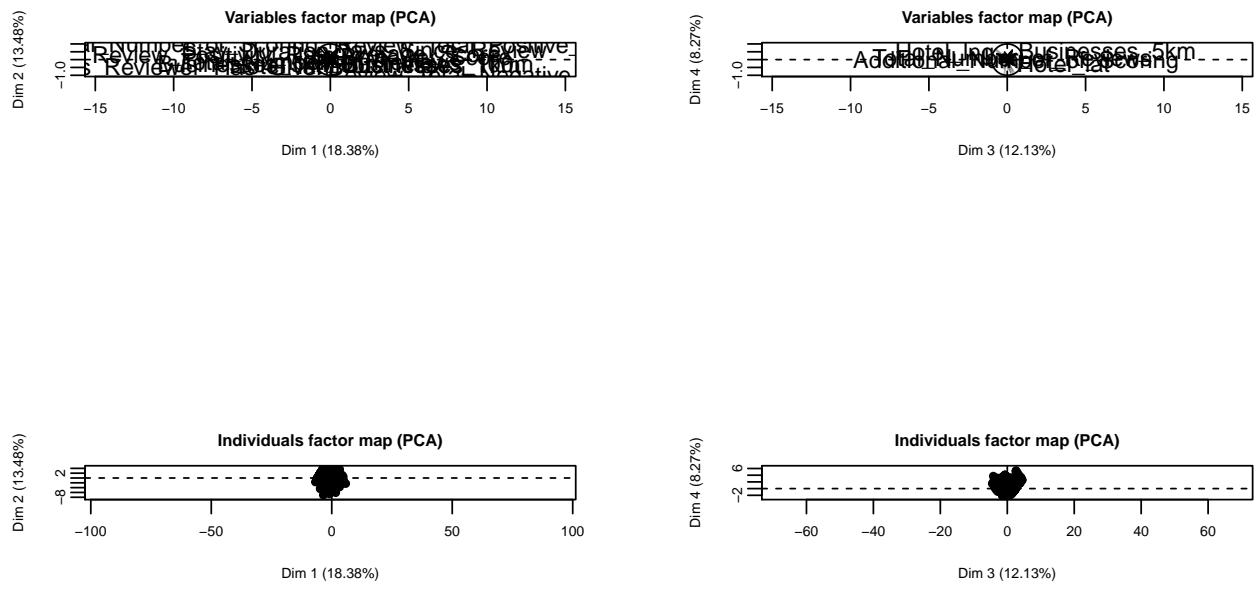


Figure 65: Resum ACP

## Aggregated distance at each iteration

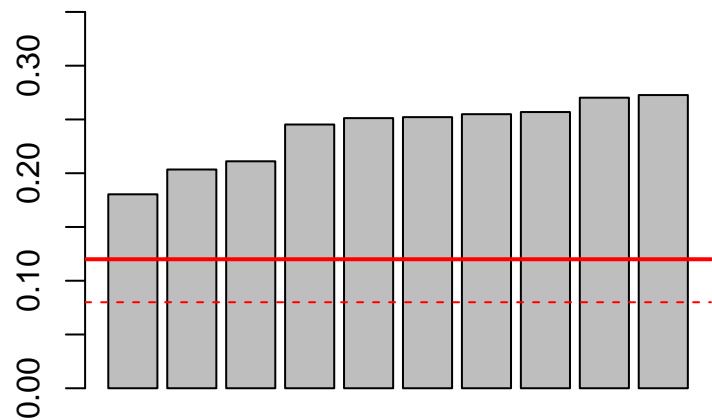


Figure 66: Distància acumulada a cada tall (1:10)

## Dendrograma-WARD

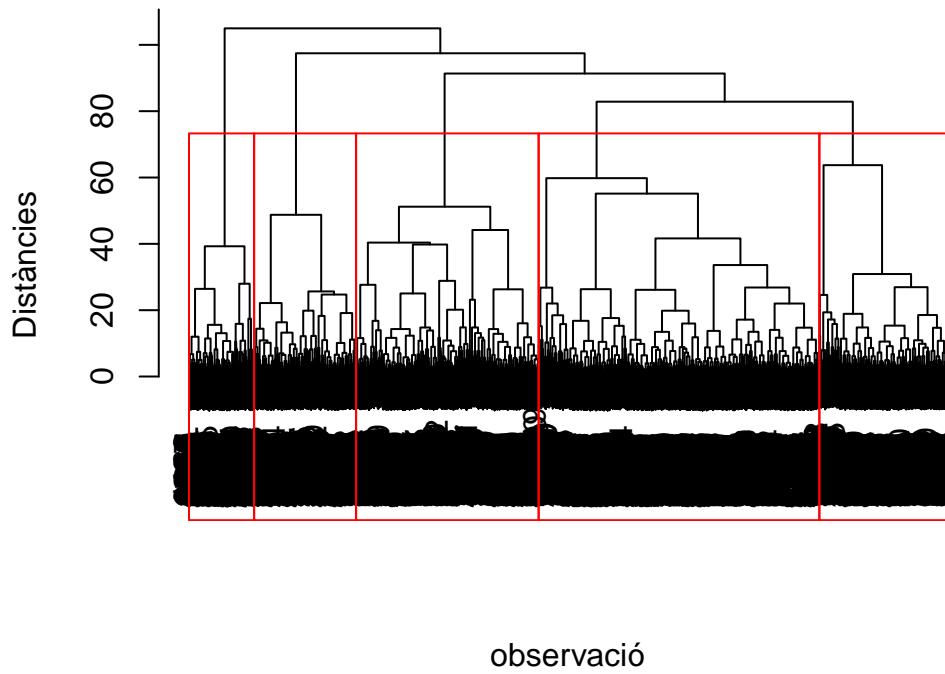


Figure 67: Dendrograma-mètode de Ward amb distàncies euclidianes

entre clústers sobre la total. Recordem, és interessant minimizar la variància intra-clusters (que a la vegada maximitzarà la entre-clusters) ja que volem tenir els individus agrupats en conglomerats homogènis internament però heterogènis entre ells.

Veiem com, en aquest cas, la partició no és tan bona (inclus la consolidada) ja que amb l'ACP reduïm la dimensionalitat de les dades a un conjunt reduït de components principals que capturen una major proporció de la variància. A continuació, podem representar les dades als eixos de components principals un cop consolidada la partició. Comencem fent la representació sobre el primer pla factorial (components principals 1 i 2) (*Figure 68*).

Observem com el resultat és força semblant a l'obtingut en el clustering anterior, una forta distinció entre els conglomerats 3 i 5 que, recordem, es corresponen amb els perfils d'hotels ubicats, respectivament, a França i el Regne Unit (tot i que potser ara han canviat aquestes etiquetes).

Tot seguit, fem servir la funció *HCPC* de FactoMineR per a obtenir informació adicional sobre la partició. En primer lloc, mencionar que el output de la funció ens torna una list amb diferents components relacionats amb característiques de la partició. Comencem mostrant el test Chi-quadrat que especifica la rellevància de cada variable categòrica en la caracterització dels clústers.

	p.value	df
Hotel_Country	0.000000e+00	20
Hotel_City	0.000000e+00	28
Review_Is_Positive	4.715591e-232	4
Reviewer_Nationality	2.061743e-79	80
Room_Type_Level	8.882218e-25	28
Trip_Type	1.566886e-04	20
Submitted_from_Mobile	4.994905e-04	4
Guest_Type	2.479092e-03	24

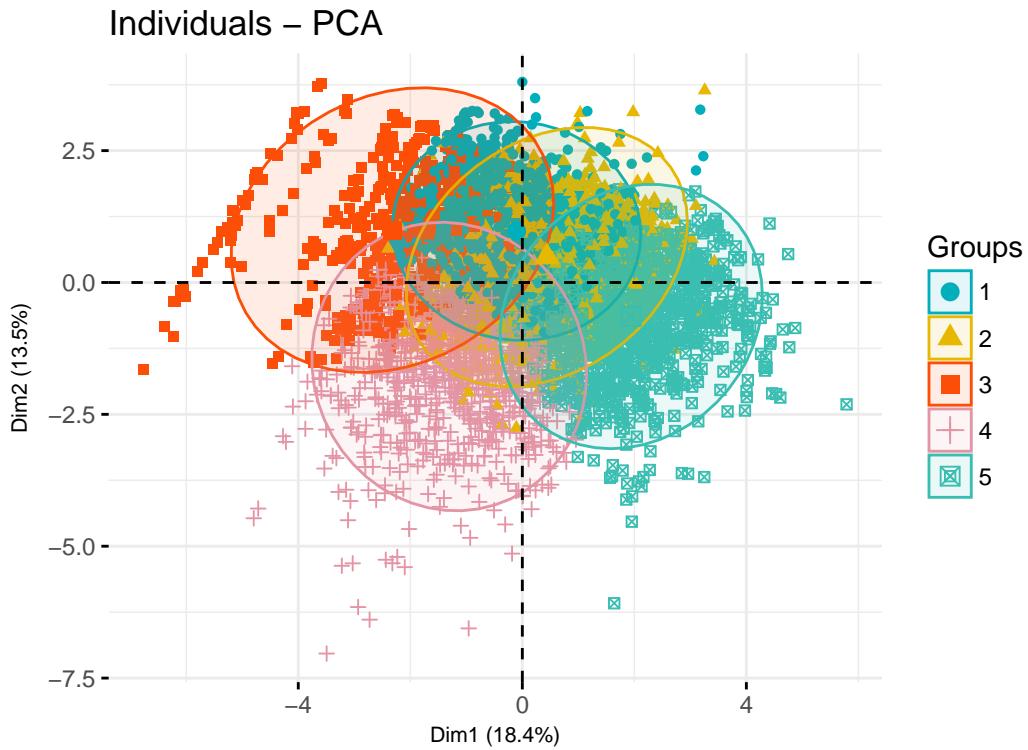


Figure 68: Clustering jeràrquic ACP sobre el primer pla factorial

Observem com les variables geogràfiques són les que millor discriminen en relació als nostres conglomerats. FactoMineR també ens permet realitzar una descripció completa de cada modalitat per variable, mostrant la representativitat de cada cluster en cada modalitat i de cada modalitat en cada clúster, juntament amb un test Chi-quadrat per a cada modalitat.

\$‘1‘

	Cla/Mod	Mod/Cla	Global
Hotel_City=London	16.0140955	95.1162791	51.08
Hotel_Country=GB	16.0140955	95.1162791	51.08
Reviewer_Nationality=United Kingdom	11.8483412	63.9534884	46.42
Room_Type_Level=Standard	16.6959578	22.0930233	11.38
Review_Is_Positive>No	10.6965174	60.0000000	48.24
Trip_Type=Couple	19.6428571	2.5581395	1.12
Is_Hotel_Holiday>No	8.7018256	99.7674419	98.60
Is_Reviewer_Holiday>No	8.6965335	99.7674419	98.66
Reviewer_Nationality=Israel	1.5625000	0.2325581	1.28
Is_Reviewer_Holiday>Yes	1.4925373	0.2325581	1.34
Reviewer_Nationality=Turkey	0.0000000	0.0000000	0.90
Reviewer_Nationality=Other	6.6448802	14.1860465	18.36
Reviewer_Nationality=Belgium	1.4492754	0.2325581	1.38
Reviewer_Nationality=Australia	4.3062201	2.0930233	4.18
Is_Hotel_Holiday>Yes	1.4285714	0.2325581	1.40
Guest_Type=Group	5.9270517	9.0697674	13.16
Reviewer_Nationality=France	0.0000000	0.0000000	1.32
Hotel_City=Milan	4.0816327	3.2558140	6.86

Hotel_Country=IT	4.0816327	3.2558140	6.86
Review_Is_Positive=Yes	6.6460587	40.0000000	51.76
Room_Type_Level=Classic	0.5681818	0.4651163	7.04
Hotel_City=Amsterdam	1.2544803	1.6279070	11.16
Hotel_Country=NL	1.2544803	1.6279070	11.16
Hotel_City=Vienna	0.0000000	0.0000000	7.52
Hotel_Country=AT	0.0000000	0.0000000	7.52
Hotel_City=Barcelona	0.0000000	0.0000000	11.48
Hotel_Country=ES	0.0000000	0.0000000	11.48
Hotel_City=Paris	0.0000000	0.0000000	11.78
Hotel_Country=FR	0.0000000	0.0000000	11.90
	p.value	v.test	
Hotel_City=London	5.950867e-98	21.004619	
Hotel_Country=GB	5.950867e-98	21.004619	
Reviewer_Nationality=United Kingdom	2.346941e-14	7.630033	
Room_Type_Level=Standard	2.619545e-11	6.666512	
Review_Is_Positive>No	3.277967e-07	5.106721	
Trip_Type=Couple	9.514186e-03	2.593003	
Is_Hotel_Holiday>No	1.531768e-02	2.424780	
Is_Reviewer_Holiday>No	1.946507e-02	2.336503	
Reviewer_Nationality=Israel	2.470259e-02	-2.246022	
Is_Reviewer_Holiday>Yes	1.946507e-02	-2.336503	
Reviewer_Nationality=Turkey	1.715684e-02	-2.383329	
Reviewer_Nationality=Other	1.682671e-02	-2.390472	
Reviewer_Nationality=Belgium	1.659367e-02	-2.395589	
Reviewer_Nationality=Australia	1.589857e-02	-2.411236	
Is_Hotel_Holiday>Yes	1.531768e-02	-2.424780	
Guest_Type=Group	6.505776e-03	-2.721137	
Reviewer_Nationality=France	2.539548e-03	-3.018589	
Hotel_City=Milan	8.260369e-04	-3.343920	
Hotel_Country=IT	8.260369e-04	-3.343920	
Review_Is_Positive>Yes	3.277967e-07	-5.106721	
Room_Type_Level=Classic	3.763045e-12	-6.945806	
Hotel_City=Amsterdam	4.230522e-15	-7.847903	
Hotel_Country=NL	4.230522e-15	-7.847903	
Hotel_City=Vienna	5.099010e-16	-8.109114	
Hotel_Country=AT	5.099010e-16	-8.109114	
Hotel_City=Barcelona	1.320569e-24	-10.239397	
Hotel_Country=ES	1.320569e-24	-10.239397	
Hotel_City=Paris	2.843550e-25	-10.386931	
Hotel_Country=FR	1.536144e-25	-10.445518	

\$‘2‘

	Cla/Mod	Mod/Cla	Global
Review_Is_Positive>No	43.076285	86.2240664	48.24
Hotel_City=London	27.682067	58.6721992	51.08
Hotel_Country=GB	27.682067	58.6721992	51.08
Hotel_Country=FR	30.588235	15.1037344	11.90

Hotel_City=Paris	30.390492	14.8547718	11.78
Reviewer_Nationality=Saudi Arabia	43.243243	2.6556017	1.48
Trip_Type=Business trip	27.895392	18.5892116	16.06
Room_Type_Level=Business	37.704918	1.9087137	1.22
Guest_Type=Solo traveler	26.555246	23.7344398	21.54
Room_Type_Level=Suite	14.084507	0.8298755	1.42
Reviewer_Nationality=Other	21.350763	16.2655602	18.36
Room_Type_Level=Family	16.030534	1.7427386	2.62
Reviewer_Nationality=Israel	12.500000	0.6639004	1.28
Trip_Type=Leisure trip	23.334159	78.1742739	80.74
Hotel_City=Milan	8.454810	2.4066390	6.86
Hotel_Country=IT	8.454810	2.4066390	6.86
Hotel_City=Vienna	7.978723	2.4896266	7.52
Hotel_Country=AT	7.978723	2.4896266	7.52
Review_Is_Positive=Yes	6.414219	13.7759336	51.76
	p.value	v.test	
Review_Is_Positive>No	1.901735e-217	31.475139	
Hotel_City=London	1.346953e-09	6.061694	
Hotel_Country=GB	1.346953e-09	6.061694	
Hotel_Country=FR	1.167744e-04	3.852801	
Hotel_City=Paris	2.020963e-04	3.716382	
Reviewer_Nationality=Saudi Arabia	2.856457e-04	3.627982	
Trip_Type=Business trip	6.717684e-03	2.710525	
Room_Type_Level=Business	1.748792e-02	2.376286	
Guest_Type=Solo traveler	3.467195e-02	2.112169	
Room_Type_Level=Suite	4.003005e-02	-2.053439	
Reviewer_Nationality=Other	2.987626e-02	-2.171727	
Room_Type_Level=Family	2.424486e-02	-2.253227	
Reviewer_Nationality=Israel	2.260617e-02	-2.280026	
Trip_Type=Leisure trip	1.026676e-02	-2.566713	
Hotel_City=Milan	1.106280e-14	-7.726403	
Hotel_Country=IT	1.106280e-14	-7.726403	
Hotel_City=Vienna	4.360807e-17	-8.402767	
Hotel_Country=AT	4.360807e-17	-8.402767	
Review_Is_Positive=Yes	1.901735e-217	-31.475139	

\$‘3‘

	Cla/Mod	Mod/Cla	Global
Review_Is_Positive=Yes	49.536321	69.1851052	51.76
Hotel_City=London	48.316366	66.5947113	51.08
Hotel_Country=GB	48.316366	66.5947113	51.08
Hotel_City=Amsterdam	68.100358	20.5072855	11.16
Hotel_Country=NL	68.100358	20.5072855	11.16
Reviewer_Nationality=United Kingdom	45.454545	56.9347005	46.42
Room_Type_Level=Deluxe	47.738693	15.3804641	11.94
Reviewer_Nationality=Ireland	50.000000	3.6157582	2.68
Reviewer_Nationality=Spain	22.916667	0.5936319	0.96
Reviewer_Nationality=Kuwait	22.916667	0.5936319	0.96

Reviewer_Nationality=Switzerland	26.136364	1.2412304	1.76
Guest_Type=Family with young children	32.864675	10.0917431	11.38
Guest_Type=Solo traveler	33.983287	19.7517539	21.54
Reviewer_Nationality=Italy	22.033898	0.7015650	1.18
Reviewer_Nationality=Australia	28.708134	3.2379924	4.18
Reviewer_Nationality=United Arab Emirates	26.016260	1.7269293	2.46
Reviewer_Nationality=Saudi Arabia	20.270270	0.8094981	1.48
Room_Type_Level=Other	35.119975	60.8202914	64.18
Reviewer_Nationality=Other	27.995643	13.8694010	18.36
Hotel_Country=FR	25.042017	8.0410146	11.90
Hotel_City=Paris	24.957555	7.9330815	11.78
Hotel_City=Milan	10.787172	1.9967620	6.86
Hotel_Country=IT	10.787172	1.9967620	6.86
Hotel_City=Vienna	5.319149	1.0793308	7.52
Hotel_Country=AT	5.319149	1.0793308	7.52
Hotel_City=Barcelona	5.749129	1.7808958	11.48
Hotel_Country=ES	5.749129	1.7808958	11.48
Review_Is_Positive>No	23.673300	30.8148948	48.24
	p.value	v.test	
Review_Is_Positive=Yes	2.538956e-81	19.099710	
Hotel_City=London	1.759644e-64	16.955266	
Hotel_Country=GB	1.759644e-64	16.955266	
Hotel_City=Amsterdam	3.429392e-56	15.793860	
Hotel_Country=NL	3.429392e-56	15.793860	
Reviewer_Nationality=United Kingdom	2.533749e-30	11.443526	
Room_Type_Level=Deluxe	1.354245e-08	5.679076	
Reviewer_Nationality=Ireland	2.037398e-03	3.084726	
Reviewer_Nationality=Spain	3.850149e-02	-2.069475	
Reviewer_Nationality=Kuwait	3.850149e-02	-2.069475	
Reviewer_Nationality=Switzerland	3.004507e-02	-2.169496	
Guest_Type=Family with young children	2.697998e-02	-2.211807	
Guest_Type=Solo traveler	1.791394e-02	-2.367392	
Reviewer_Nationality=Italy	1.399421e-02	-2.457412	
Reviewer_Nationality=Australia	9.716393e-03	-2.585762	
Reviewer_Nationality=United Arab Emirates	9.004448e-03	-2.611885	
Reviewer_Nationality=Saudi Arabia	1.868529e-03	-3.110370	
Room_Type_Level=Other	1.496556e-04	-3.791640	
Reviewer_Nationality=Other	1.667502e-10	-6.389205	
Hotel_Country=FR	3.527283e-11	-6.622687	
Hotel_City=Paris	3.295708e-11	-6.632714	
Hotel_City=Milan	7.527801e-30	-11.348695	
Hotel_Country=IT	7.527801e-30	-11.348695	
Hotel_City=Vienna	8.584102e-51	-14.989621	
Hotel_Country=AT	8.584102e-51	-14.989621	
Hotel_City=Barcelona	1.212302e-76	-18.528677	
Hotel_Country=ES	1.212302e-76	-18.528677	
Review_Is_Positive>No	2.538956e-81	-19.099710	

\$‘4‘

	Cla/Mod	Mod/Cla	Global
Hotel_City=Vienna	72.340426	32.419547	7.52
Hotel_Country=AT	72.340426	32.419547	7.52
Hotel_City=Milan	74.344023	30.393325	6.86
Hotel_Country=IT	74.344023	30.393325	6.86
Reviewer_Nationality=Other	31.154684	34.088200	18.36
Hotel_City=Barcelona	35.714286	24.433850	11.48
Hotel_Country=ES	35.714286	24.433850	11.48
Room_Type_Level=Other	18.572764	71.036949	64.18
Submitted_from_Mobile=No	19.130869	45.649583	40.04
Reviewer_Nationality=Italy	35.593220	2.502980	1.18
Review_Is_Positive=Yes	18.353941	56.615018	51.76
Reviewer_Nationality=Switzerland	29.545455	3.098927	1.76
Reviewer_Nationality=Israel	29.687500	2.264601	1.28
Reviewer_Nationality=Canada	27.027027	2.383790	1.48
Reviewer_Nationality=United Arab Emirates	24.390244	3.575685	2.46
Reviewer_Nationality=Australia	22.488038	5.601907	4.18
Reviewer_Nationality=Kuwait	29.166667	1.668653	0.96
Reviewer_Nationality=Germany	25.925926	2.502980	1.62
Trip_Type=Business trip	19.302615	18.474374	16.06
Guest_Type=Couple	15.422680	44.576877	48.50
Trip_Type=Leisure trip	16.125836	77.592372	80.74
Reviewer_Nationality=Ireland	8.208955	1.311085	2.68
Review_Is_Positive=No	15.091211	43.384982	48.24
Room_Type_Level=Standard	12.126538	8.224076	11.38
Submitted_from_Mobile=Yes	15.210140	54.350417	59.96
Room_Type_Level=Deluxe	9.882747	7.032181	11.94
Hotel_City=Amsterdam	4.659498	3.098927	11.16
Hotel_Country=NL	4.659498	3.098927	11.16
Hotel_Country=FR	2.857143	2.026222	11.90
Hotel_City=Paris	2.716469	1.907032	11.78
Reviewer_Nationality=United Kingdom	7.970702	22.050060	46.42
Hotel_City=London	2.505873	7.628129	51.08
Hotel_Country=GB	2.505873	7.628129	51.08
	p.value	v.test	
Hotel_City=Vienna	1.046257e-140	25.253134	
Hotel_Country=AT	1.046257e-140	25.253134	
Hotel_City=Milan	7.927513e-136	24.804950	
Hotel_Country=IT	7.927513e-136	24.804950	
Reviewer_Nationality=Other	1.058310e-33	12.099829	
Hotel_City=Barcelona	4.536821e-32	11.787333	
Hotel_Country=ES	4.536821e-32	11.787333	
Room_Type_Level=Other	4.336447e-06	4.594569	
Submitted_from_Mobile=No	3.002620e-04	3.615074	
Reviewer_Nationality=Italy	4.675103e-04	3.498713	
Review_Is_Positive=Yes	2.020826e-03	3.087154	
Reviewer_Nationality=Switzerland	2.742486e-03	2.995218	

Reviewer_Nationality=Israel	1.003438e-02	2.574642
Reviewer_Nationality=Canada	2.544799e-02	2.234534
Reviewer_Nationality=United Arab Emirates	2.891706e-02	2.184615
Reviewer_Nationality=Australia	2.922482e-02	2.180441
Reviewer_Nationality=Kuwait	3.207486e-02	2.143476
Reviewer_Nationality=Germany	3.550357e-02	2.102568
Trip_Type=Business trip	3.927739e-02	2.061269
Guest_Type=Couple	1.267159e-02	-2.492868
Trip_Type=Leisure trip	1.244911e-02	-2.499152
Reviewer_Nationality=Ireland	4.072768e-03	-2.872470
Review_Is_Positive>No	2.020826e-03	-3.087154
Room_Type_Level=Standard	1.133909e-03	-3.255004
Submitted_from_Mobile=Yes	3.002620e-04	-3.615074
Room_Type_Level=Deluxe	3.991482e-07	-5.069364
Hotel_City=Amsterdam	3.907774e-20	-9.190567
Hotel_Country=NL	3.907774e-20	-9.190567
Hotel_Country=FR	9.870132e-30	-11.324980
Hotel_City=Paris	3.651029e-30	-11.411798
Reviewer_Nationality=United Kingdom	2.415358e-57	-15.960308
Hotel_City=London	5.638146e-191	-29.477233
Hotel_Country=GB	5.638146e-191	-29.477233

\$‘5‘

	Cla/Mod	Mod/Cla	Global
Hotel_City=Paris	41.935484	36.7013373	11.78
Hotel_Country=FR	41.512605	36.7013373	11.90
Hotel_City=Barcelona	35.888502	30.6092125	11.48
Hotel_Country=ES	35.888502	30.6092125	11.48
Review_Is_Positive>Yes	19.049459	73.2540862	51.76
Reviewer_Nationality=United States of America	23.561644	12.7786033	7.30
Room_Type_Level=Classic	22.443182	11.7384844	7.04
Reviewer_Nationality=Australia	25.358852	7.8751857	4.18
Trip_Type=Leisure trip	14.515730	87.0728083	80.74
Reviewer_Nationality=Israel	29.687500	2.8231798	1.28
Reviewer_Nationality=United Arab Emirates	21.951220	4.0118871	2.46
Room_Type_Level=Suite	22.535211	2.3774146	1.42
Reviewer_Nationality=France	22.727273	2.2288262	1.32
Guest_Type=Couple	14.474227	52.1545319	48.50
Trip_Type=Solo traveler	4.166667	0.2971768	0.96
Trip_Type=Couple	3.571429	0.2971768	1.12
Room_Type_Level=Deluxe	10.385260	9.2124814	11.94
Guest_Type=Group	10.486322	10.2526003	13.16
Room_Type_Level=Business	3.278689	0.2971768	1.22
Reviewer_Nationality=Kuwait	2.083333	0.1485884	0.96
Trip_Type=Business trip	9.838107	11.7384844	16.06
Hotel_City=Milan	2.332362	1.1887073	6.86
Hotel_Country=IT	2.332362	1.1887073	6.86
Reviewer_Nationality=United Kingdom	9.521758	32.8380386	46.42

Hotel_City=Amsterdam	3.225806	2.6745914	11.16
Hotel_Country=NL	3.225806	2.6745914	11.16
Review_Is_Positive=No	7.462687	26.7459138	48.24
Hotel_City=London	5.481597	20.8023774	51.08
Hotel_Country=GB	5.481597	20.8023774	51.08
	p.value	v.test	
Hotel_City=Paris	1.283305e-77	18.649138	
Hotel_Country=FR	1.784316e-76	18.507866	
Hotel_City=Barcelona	3.490596e-49	14.741478	
Hotel_Country=ES	3.490596e-49	14.741478	
Review_Is_Positive=Yes	2.640899e-34	12.213260	
Reviewer_Nationality=United States of America	5.263888e-08	5.442158	
Room_Type_Level=Classic	1.760807e-06	4.779101	
Reviewer_Nationality=Australia	2.552404e-06	4.703898	
Trip_Type=Leisure trip	3.294175e-06	4.651570	
Reviewer_Nationality=Israel	6.793330e-04	3.397787	
Reviewer_Nationality=United Arab Emirates	8.992099e-03	2.612354	
Room_Type_Level=Suite	3.537225e-02	2.104071	
Reviewer_Nationality=France	3.864947e-02	2.067900	
Guest_Type=Couple	4.164218e-02	2.037078	
Trip_Type=Solo traveler	4.204137e-02	-2.033110	
Trip_Type=Couple	1.679146e-02	-2.391242	
Room_Type_Level=Deluxe	1.636626e-02	-2.400644	
Guest_Type=Group	1.419853e-02	-2.452201	
Room_Type_Level=Business	9.318243e-03	-2.600152	
Reviewer_Nationality=Kuwait	8.932319e-03	-2.614634	
Trip_Type=Business trip	7.201839e-04	-3.381778	
Hotel_City=Milan	1.565182e-13	-7.381498	
Hotel_Country=IT	1.565182e-13	-7.381498	
Reviewer_Nationality=United Kingdom	1.685733e-14	-7.672578	
Hotel_City=Amsterdam	5.954268e-18	-8.633423	
Hotel_Country=NL	5.954268e-18	-8.633423	
Review_Is_Positive=No	2.640899e-34	-12.213260	
Hotel_City=London	3.423151e-67	-17.318301	
Hotel_Country=GB	3.423151e-67	-17.318301	

A continuació, repetim el procediment de trobar la variable que millor defineix els conglomerats, aquest cop per a les variables numèriques. Podem fer un ajust lineal amb els clusters com a covariables per a veure la relació que existeix entre les nostres variables numèriques i les components del clúster. En aquest cas, com a repistes testem les valoracions dels hotels, variables *Reviewer\_Score* i *Average\_Score*.

	Eta2	P-value
Hotel_lat	0.34178533	0.000000e+00
Hotel_lng	0.35166900	0.000000e+00
Businesses_100m	0.30013932	0.000000e+00
Businesses_1km	0.35703173	0.000000e+00
Total_Number_of_Reviews	0.47016499	0.000000e+00

Review_Positivity_Rate	0.26377625	0.000000e+00
Additional_Number_of_Scoring	0.58967032	0.000000e+00
Reviewer_Score	0.23468498	4.567920e-288
Businesses_5km	0.17787674	1.598067e-210
Review_Total_Negative_Word_Counts	0.16511388	7.592318e-194
Average_Score	0.16001530	2.954659e-187
Total_Number_of_Reviews_Reviewer_Has_Given	0.09700051	5.195372e-109
Review_Total_Positive_Word_Counts	0.06175739	1.117095e-67
Days_Since_Review	0.02879422	1.488267e-30
Stay_Duration	0.02692301	1.704847e-28

Call:

```
lm(formula = res.hcpc$data.clust$Reviewer_Score ~ res.hcpc$data.clust$clust)
```

Residuals:

Min	1Q	Median	3Q	Max
-6.5440	-0.7440	0.4186	0.9909	2.9221

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	7.88140	0.06972	113.036	< 2e-16 ***
res.hcpc\$data.clust\$clust2	-0.80347	0.08122	-9.893	< 2e-16 ***
res.hcpc\$data.clust\$clust3	1.12767	0.07739	14.571	< 2e-16 ***
res.hcpc\$data.clust\$clust4	0.66199	0.08575	7.720	1.4e-14 ***
res.hcpc\$data.clust\$clust5	1.16259	0.08926	13.024	< 2e-16 ***
---				
Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'
	0.1 , , 1			

Residual standard error: 1.446 on 4995 degrees of freedom

Multiple R-squared: 0.2347, Adjusted R-squared: 0.2341

F-statistic: 382.9 on 4 and 4995 DF, p-value: < 2.2e-16

Call:

```
lm(formula = res.hcpc$data.clust$Average_Score ~ res.hcpc$data.clust$clust)
```

Residuals:

Min	1Q	Median	3Q	Max
-2.95585	-0.31015	0.04415	0.36805	1.44415

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	8.00558	0.02398	333.792	< 2e-16 ***
res.hcpc\$data.clust\$clust2	0.15027	0.02794	5.379	7.84e-08 ***
res.hcpc\$data.clust\$clust3	0.50456	0.02662	18.953	< 2e-16 ***
res.hcpc\$data.clust\$clust4	0.40693	0.02950	13.796	< 2e-16 ***
res.hcpc\$data.clust\$clust5	0.72637	0.03070	23.657	< 2e-16 ***

```

---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4973 on 4995 degrees of freedom
Multiple R-squared: 0.16, Adjusted R-squared: 0.1593
F-statistic: 237.9 on 4 and 4995 DF, p-value: < 2.2e-16

```

En consonància amb el resultat previ, les variables geogràfiques, són les que millor caracteritzen els nostres conglomerats (latitud i longitud, Negocis a la rodona...). A més, tots dos models són significatius i les puntuacions (respostes) semblen ser màximes al clúster 5.

A continuació, formulem la descripció dels clústers en termes de variables quanitatives. Un cop realitzat el nou profiling d'aquests clústers, farem més intuïtiva aquesta caracterització. Tanmateix, a continuació mostrem una pinzellada del que ens ofereix FactoMineR per a realitzar aquesta descripció.

FactoMineR ens permet calcular estadístics numèrics per variable i clúster. En aquest cas, proposem calcular les mitjanes i desviacions tipus per al clúster 1 i el global, prenent com a variable numèrica *Total\_Number\_of\_Reviews*.

```

[1] 7595.6
[1] 2731.855
[1] 2729.571
[1] 2268.892

```

Noteu que, podem replicar el procediment anterior per a qualsevol combinació de variable numèrica i clúster.

Si volem un resum numèric complet (totes les variables i tots els clústers), cridem la comanda *res.hpcdesc.varquanti*.

```

$ '1'
v.test Mean in category
Additional_Number_of_Scoring      51.569866    1674.5534884
Total_Number_of_Reviews           46.496278    7595.6000000
Hotel_lat                          11.747007    51.3254612
Businesses_100m                   -3.036092    24.3627907
Stay_Duration                      -3.870004    2.0883721
Total_Number_of_Reviews_Reviewer_Has_Given -4.898827    4.7465116
Review_Total_Positive_Word_Counts   -5.450150    12.1720930
Review_Positivity_Rate             -5.504199    47.1600531
Reviewer_Score                     -6.457465    7.8813953
Businesses_1km                     -7.795937    1298.6906977
Hotel_lng                           -12.200198    0.2426063
Average_Score                      -15.565289    8.0055814
Overall mean sd in category
Additional_Number_of_Scoring      497.437800    597.3201788
Total_Number_of_Reviews           2731.855000    2726.3954717
Hotel_lat                          49.458637     1.0682937

```

Businesses_100m	28.332000	19.5378824
Stay_Duration	2.390800	1.5039807
Total_Number_of_Reviews_Reviewer_Has_Given	7.323000	6.4550989
Review_Total_Positive_Word_Counts	17.336400	14.6038594
Review_Positivity_Rate	55.733893	33.6096699
Reviewer_Score	8.373240	1.8069061
Businesses_1km	1785.679200	832.5099173
Hotel_lng	2.834132	1.7583892
Average_Score	8.394840	0.4832704
	Overall sd	p.value
Additional_Number_of_Scoring	495.0406296	0.000000e+00
Total_Number_of_Reviews	2268.6646545	0.000000e+00
Hotel_lat	3.4466240	7.316473e-32
Businesses_100m	28.3535073	2.396663e-03
Stay_Duration	1.6948379	1.088336e-04
Total_Number_of_Reviews_Reviewer_Has_Given	11.4065363	9.641053e-07
Review_Total_Positive_Word_Counts	20.5504461	5.032722e-08
Review_Positivity_Rate	33.7830579	3.708504e-08
Reviewer_Score	1.6518995	1.064717e-10
Businesses_1km	1354.7778449	6.393231e-15
Hotel_lng	4.6068739	3.100698e-34
Average_Score	0.5423738	1.253028e-54

\$‘2‘

	v.test	Mean in category
Review_Total_Negative_Word_Counts	28.544174	41.200830
Businesses_5km	7.304958	15513.447303
Stay_Duration	4.976232	2.602490
Hotel_lat	4.906470	49.883094
Businesses_1km	3.583898	1907.548548
Total_Number_of_Reviews_Reviewer_Has_Given	-6.858712	5.359336
Days_Since_Review	-9.352497	306.529461
Hotel_lng	-10.491474	1.620984
Additional_Number_of_Scoring	-10.583066	365.938589
Review_Total_Positive_Word_Counts	-12.192598	11.047303
Total_Number_of_Reviews	-12.674242	2010.144398
Average_Score	-17.555333	8.155851
Reviewer_Score	-31.240709	7.077925
Review_Positivity_Rate	-33.581854	27.258192
	Overall mean	sd in category
Review_Total_Negative_Word_Counts	19.382200	47.440475
Businesses_5km	14238.567200	6540.614123
Stay_Duration	2.390800	2.158567
Hotel_lat	49.458637	3.268240
Businesses_1km	1785.679200	1349.140701
Total_Number_of_Reviews_Reviewer_Has_Given	7.323000	6.383290
Days_Since_Review	355.551000	205.763915
Hotel_lng	2.834132	3.082142

Additional_Number_of_Scoring	497.437800	273.036251
Review_Total_Positive_Word_Counts	17.336400	12.179035
Total_Number_of_Reviews	2731.855000	1336.128143
Average_Score	8.394840	0.598207
Reviewer_Score	8.373240	1.812761
Review_Positivity_Rate	55.733893	23.439399
	Overall sd	p.value
Review_Total_Negative_Word_Counts	30.4536389	3.317874e-179
Businesses_5km	6953.1378727	2.773540e-13
Stay_Duration	1.6948379	6.483392e-07
Hotel_lat	3.4466240	9.273010e-07
Businesses_1km	1354.7778449	3.385047e-04
Total_Number_of_Reviews_Reviewer_Has_Given	11.4065363	6.948414e-12
Days_Since_Review	208.8279239	8.560251e-21
Hotel_lng	4.6068739	9.454198e-26
Additional_Number_of_Scoring	495.0406296	3.570786e-26
Review_Total_Positive_Word_Counts	20.5504461	3.403945e-34
Total_Number_of_Reviews	2268.6646545	8.214936e-37
Average_Score	0.5423738	5.414851e-69
Reviewer_Score	1.6518995	2.985419e-214
Review_Positivity_Rate	33.7830579	3.087589e-247

\$‘3‘

	v.test	Mean in category
Hotel_lat	26.421871	51.137157
Review_Positivity_Rate	21.913843	69.379303
Reviewer_Score	20.882661	9.009066
Review_Total_Positive_Word_Counts	13.114192	22.303832
Average_Score	11.534081	8.510146
Days_Since_Review	10.790682	397.085267
Additional_Number_of_Scoring	2.686804	521.953589
Total_Number_of_Reviews	-3.724878	2576.096600
Total_Number_of_Reviews_Reviewer_Has_Given	-8.469394	5.542364
Stay_Duration	-9.244955	2.101997
Businesses_5km	-10.155644	12937.027523
Review_Total_Negative_Word_Counts	-13.877445	11.592553
Hotel_lng	-15.859938	1.487413
Businesses_100m	-16.563739	19.675661
Businesses_1km	-17.925039	1338.071775
	Overall mean	sd in category
Hotel_lat	49.458637	1.7994299
Review_Positivity_Rate	55.733893	29.4979719
Reviewer_Score	8.373240	1.1486560
Review_Total_Positive_Word_Counts	17.336400	26.2747115
Average_Score	8.394840	0.4720289
Days_Since_Review	355.551000	201.3134134
Additional_Number_of_Scoring	497.437800	332.9382496
Total_Number_of_Reviews	2731.855000	1721.7211249

Total_Number_of_Reviews_Reviewer_Has_Given	7.323000	6.8006033
Stay_Duration	2.390800	1.3891342
Businesses_5km	14238.567200	6881.0049328
Review_Total_Negative_Word_Counts	19.382200	17.8271877
Hotel_lng	2.834132	2.7665142
Businesses_100m	28.332000	18.5024622
Businesses_1km	1785.679200	1077.6607653
Overall sd	p.value	
Hotel_lat	3.4466240	7.683310e-154
Review_Positivity_Rate	33.7830579	1.917055e-106
Reviewer_Score	1.6518995	7.698346e-97
Review_Total_Positive_Word_Counts	20.5504461	2.730757e-39
Average_Score	0.5423738	8.883089e-31
Days_Since_Review	208.8279239	3.809526e-27
Additional_Number_of_Scoring	495.0406296	7.213934e-03
Total_Number_of_Reviews	2268.6646545	1.954095e-04
Total_Number_of_Reviews_Reviewer_Has_Given	11.4065363	2.466731e-17
Stay_Duration	1.6948379	2.353394e-20
Businesses_5km	6953.1378727	3.127379e-24
Review_Total_Negative_Word_Counts	30.4536389	8.678275e-44
Hotel_lng	4.6068739	1.200126e-56
Businesses_100m	28.3535073	1.274243e-61
Businesses_1km	1354.7778449	7.519818e-72

\$‘4‘

	v.test	Mean in category
Hotel_lng	40.709166	8.741239
Total_Number_of_Reviews_Reviewer_Has_Given	21.394420	15.009535
Review_Positivity_Rate	3.332628	59.280084
Reviewer_Score	3.270086	8.543385
Stay_Duration	2.815875	2.541120
Review_Total_Positive_Word_Counts	-5.054995	14.064362
Review_Total_Negative_Word_Counts	-6.859840	12.802145
Businesses_100m	-9.106787	20.199046
Businesses_1km	-11.831036	1280.823600
Additional_Number_of_Scoring	-14.310516	274.300358
Businesses_5km	-17.717005	10358.426698
Hotel_lat	-31.288364	46.061969
Overall mean	sd	in category
Hotel_lng	2.834132	6.033975
Total_Number_of_Reviews_Reviewer_Has_Given	7.323000	21.035809
Review_Positivity_Rate	55.733893	32.175252
Reviewer_Score	8.373240	1.425939
Stay_Duration	2.390800	1.576168
Review_Total_Positive_Word_Counts	17.336400	11.859855
Review_Total_Negative_Word_Counts	19.382200	15.636788
Businesses_100m	28.332000	18.471502
Businesses_1km	1785.679200	867.936958

Additional_Number_of_Scoring	497.437800	223.318248
Businesses_5km	14238.567200	3533.361654
Hotel_lat	49.458637	3.256278
	Overall sd	p.value
Hotel_lng	4.606874	0.000000e+00
Total_Number_of_Reviews_Reviewer_Has_Given	11.406536	1.505903e-101
Review_Positivity_Rate	33.783058	8.603002e-04
Reviewer_Score	1.651899	1.075147e-03
Stay_Duration	1.694838	4.864456e-03
Review_Total_Positive_Word_Counts	20.550446	4.304020e-07
Review_Total_Negative_Word_Counts	30.453639	6.893766e-12
Businesses_100m	28.353507	8.485755e-20
Businesses_1km	1354.777845	2.697905e-32
Additional_Number_of_Scoring	495.040630	1.881120e-46
Businesses_5km	6953.137873	3.099841e-70
Hotel_lat	3.446624	6.718602e-215

\$‘5‘

	v.test	Mean in category	Overall mean
Businesses_1km	40.232546	3740.423477	1785.67920
Businesses_100m	37.978062	66.949480	28.33200
Businesses_5km	24.573216	20366.121842	14238.56720
Average_Score	17.331018	8.731947	8.39484
Review_Positivity_Rate	11.945118	70.206067	55.73389
Reviewer_Score	11.322117	9.043982	8.37324
Stay_Duration	6.942533	2.812779	2.39080
Review_Total_Positive_Word_Counts	6.733086	22.298663	17.33640
Review_Total_Negative_Word_Counts	-7.083110	11.646360	19.38220
Total_Number_of_Reviews	-16.780072	1366.616642	2731.85500
Additional_Number_of_Scoring	-17.234370	191.466568	497.43780
Hotel_lat	-18.929722	47.118817	49.45864
	sd in category	Overall sd	
Businesses_1km	963.8621145	1354.7778449	
Businesses_100m	42.1430078	28.3535073	
Businesses_5km	7368.9771365	6953.1378727	
Average_Score	0.4596063	0.5423738	
Review_Positivity_Rate	29.1675755	33.7830579	
Reviewer_Score	1.1738724	1.6518995	
Stay_Duration	1.6025658	1.6948379	
Review_Total_Positive_Word_Counts	21.5831427	20.5504461	
Review_Total_Negative_Word_Counts	17.4267234	30.4536389	
Total_Number_of_Reviews	917.3027107	2268.6646545	
Additional_Number_of_Scoring	135.1525664	495.0406296	
Hotel_lat	4.0128756	3.4466240	
	p.value		
Businesses_1km	0.000000e+00		
Businesses_100m	0.000000e+00		
Businesses_5km	2.443028e-133		

Average_Score	2.744299e-67
Review_Positivity_Rate	6.885218e-33
Reviewer_Score	1.019788e-29
Stay_Duration	3.851321e-12
Review_Total_Positive_Word_Counts	1.661020e-11
Review_Total_Negative_Word_Counts	1.409551e-12
Total_Number_of_Reviews	3.414197e-63
Additional_Number_of_Scoring	1.466431e-66
Hotel_lat	6.490427e-80

Per últim, podem considerar la representació dels clústers sobre altres plans factorialis. En aquest cas, hem considerat representacions fins a la sisena dimensió (*Figure 69*).

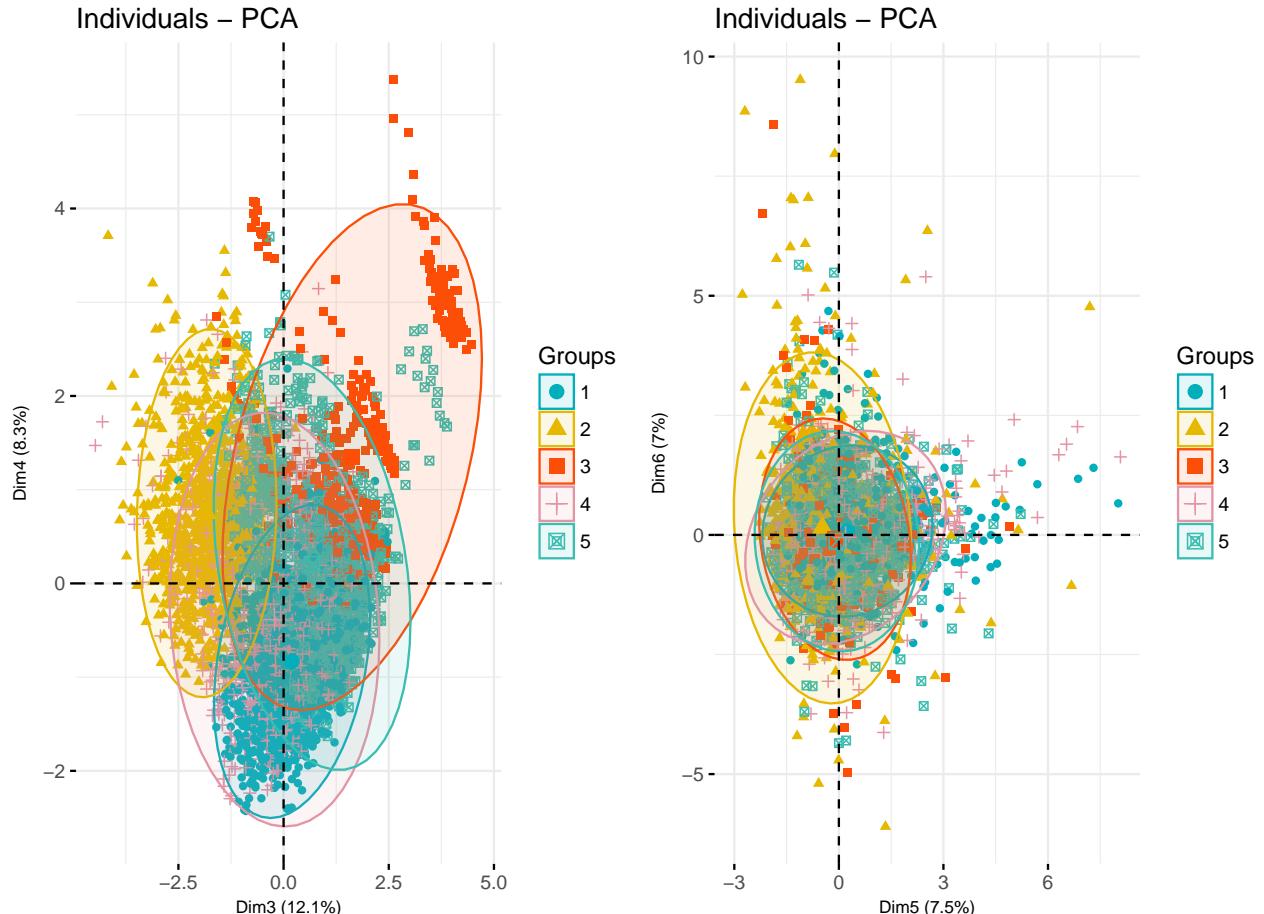


Figure 69: Clustering jeràrquic ACP (plans factorialis adicionals)

Veiem com, a mesura que avancem en les dimensions, la variabilitat disminaix ja que aquesta és capturada per les dimensions superiors. En general, a major nombre de dimensió més residual es torna aquesta, en termes de contribució a la variabilitat de les dades.

## Profiling del Clúster Jeràrquic sobre ACP

En aquesta secció, repetim el profiling, ara considerant el clúster obtingut a partir de l'ACP. Recordem les eines descriptives que fem servir per a la caracterització dels conglomerats:

- Snake plots, diagrames de barres o taules de contingència per a les variables qualitatives.
- Boxplots i diagrames de barres per a les variables numèriques.

En aquest cas, obviem els contrastos, ja que aquests apareixen resumits en les taules de l'apartat anterior. L'objectiu d'aquesta secció passa més per aportar una visió més conceptual i fàcil d'entendre de les caracteritzacions de cada clúster, i no pas endinsar-nos en càlculs numèrics (les taule de l'apartat anterior contenen tota la informació resumida en aquest apartat).

Les primeres variables de la base de dades *Hotel\_Country*, *Hotel\_City*, *Hotel\_lat*, *Hotel\_Ing* fan referència a l'àmbit geogràfic i és llògic pensar que la caracterització dels clústers continuarà en la mateixa línia en tots quatre casos (*Figure 70*).

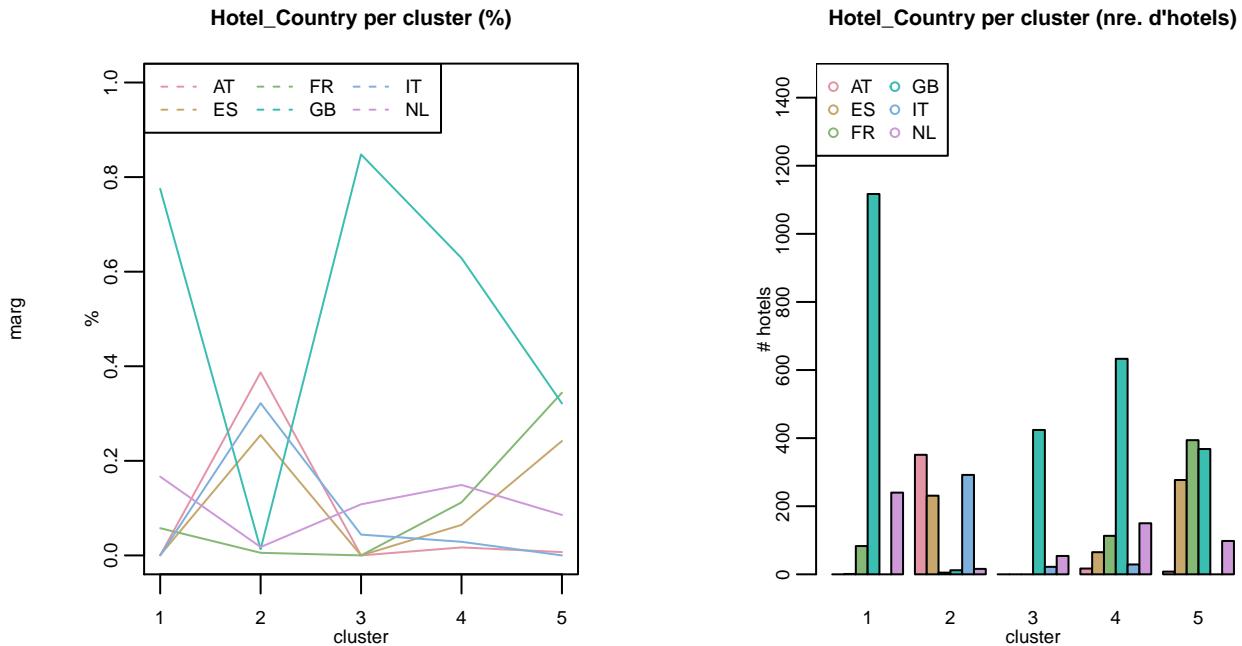


Figure 70: Profiling2 variable Hotel\_Country

Observem com la situació ha canviat totalment, ara els hotels de Gran Bretanya es troben majoritàriament concentrats als clusters 1 3 i 4. Al clúster 2 predominen els hotels d'Itàlia, Espanya i Àustria, mentre que al 5 seguim tenint fortia presència Francesa.

Si considerem la variable *Hotel\_City*, de nou observem com les ciutats concorden amb els resultat obtingut en el profiling prèvi (*Figure 71*).

De nou, Per a facilitar la comprensió dels resultats geogràfics anteriors, podem pensar en fer servir les variables latitud i longitud per a realitzar una geolocalització de les dades i visualitzar els clústers en un mapa. Als següents gràfics representem al mapa la localització dels hotels estratificant per clústers, els quals distingim amb colors, i on la mida dels cercles fa referència al nombre total de ressenyes que tenen els hotels (*Figure 72*).

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

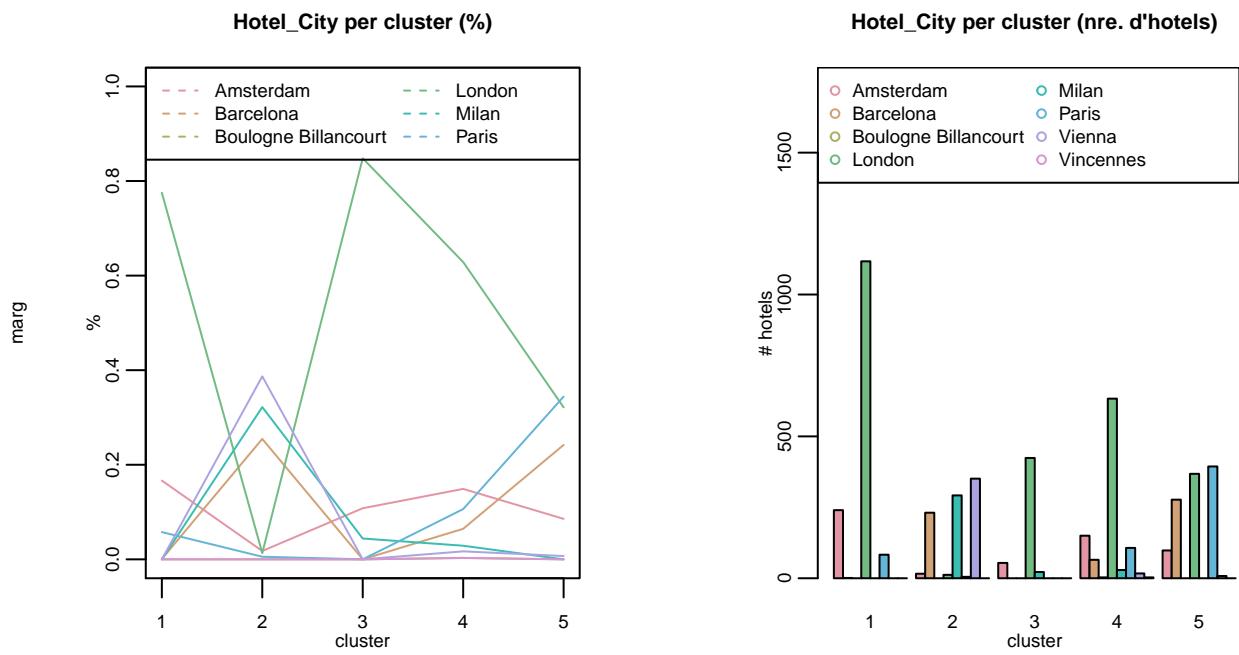


Figure 71: Profiling2 variable Hotel\_City

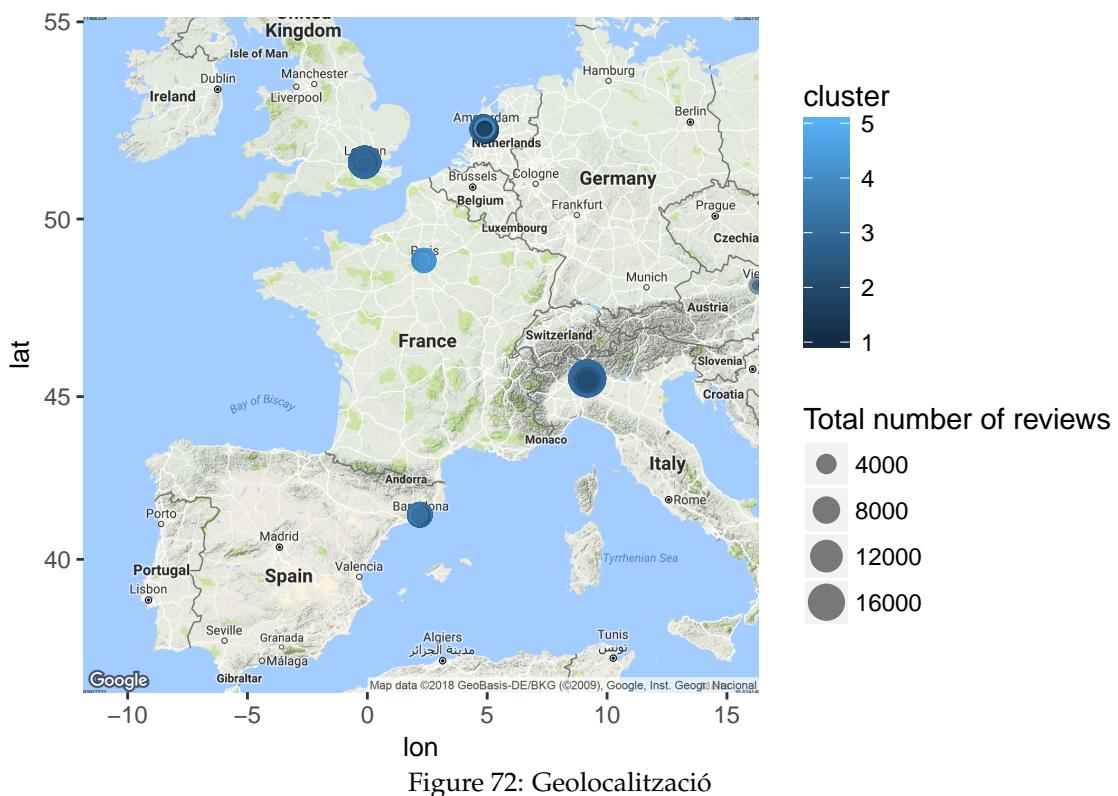


Figure 72: Geolocalització

variables és fer una distinció entre hotels urbans i hotels més allunyats del centre de les ciutats. Com que les variables *Business\_100m*, *Business\_1km* i *Business\_5Km* tenen connotacions semblants, hem considerat novament representar-les juntes mitjançant un plotMeans (*Figure 73*).

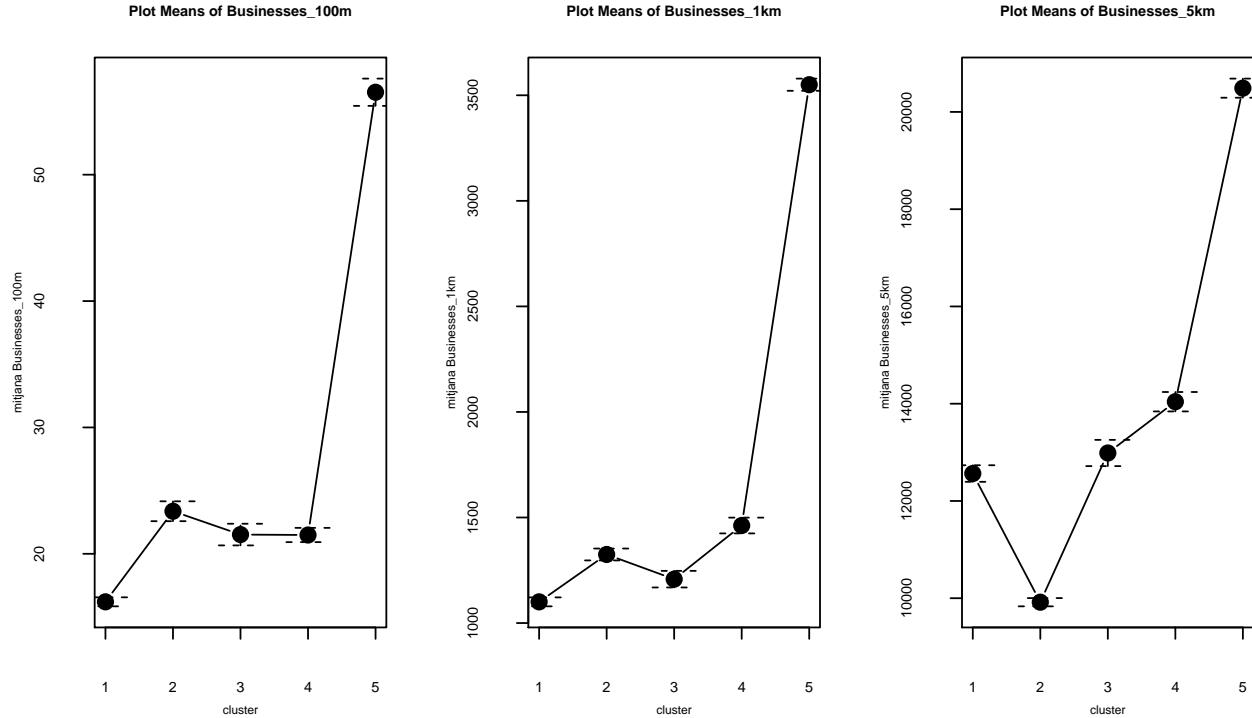


Figure 73: Profiling2 Nre. Negocis a 100m, 1km i 5km a la rodona

En aquest cas, observem com el clúster 5 absorbeix la gran majoria dels valors elevats per a variables de negocis a la rodona, està molt per sobre dels altres en tots els llindars

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 (*Figure 74*)

De nou, 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 2. La resta de modalitats són força homogènies per a tots els clústers, amb la salutat que al clúster 3 hi ha major presència d'habitacions de tipus Standard (concorda amb el profiling anterior).

La següent variable és *Guest\_Type*. En aquest cas, les parelles són més abundants als clústers 1 i 5 (*Figure 75*). Les persones que viatgen soles, en comptes de concentrar-se en el clúster 3, ara són més freqüents al 4.

A continuació, passem a la variable *Trip\_Type* que reflecteix el motiu del viatge. Veiem com es manté força constant, predominen els viatges amb motiu d'oci per a tots els clústers, amb lleugeres diferències al clúster 4 on aquests baixen una mica i pugen els viatges per negocis (concorda amb els que viatgen sols) (*Figure 76*).

La següent variable és *Stay\_Duration*. Al tenir davant una variable numèrica, recordem, cal construir un boxplot i un gràfic de barres (*Figure 77*). En aquest cas, els clústers 2 i 5 inclouen llargues estades, mentre que el primer, que en l'anterior profiling presentava un valor força alt, ara conté els hotels on les estades són les més curtes.

La següent variable és *Days\_Since\_Review*. En aquest cas, fem un resum numèric per a cada segment (clústers del 1 al 5). Sembla que la mitjana del clúster 5 és considerablement superior a la resta (igual que el profiling prèvi). Tanmateix, recordem que aquesta variable no està massa ben

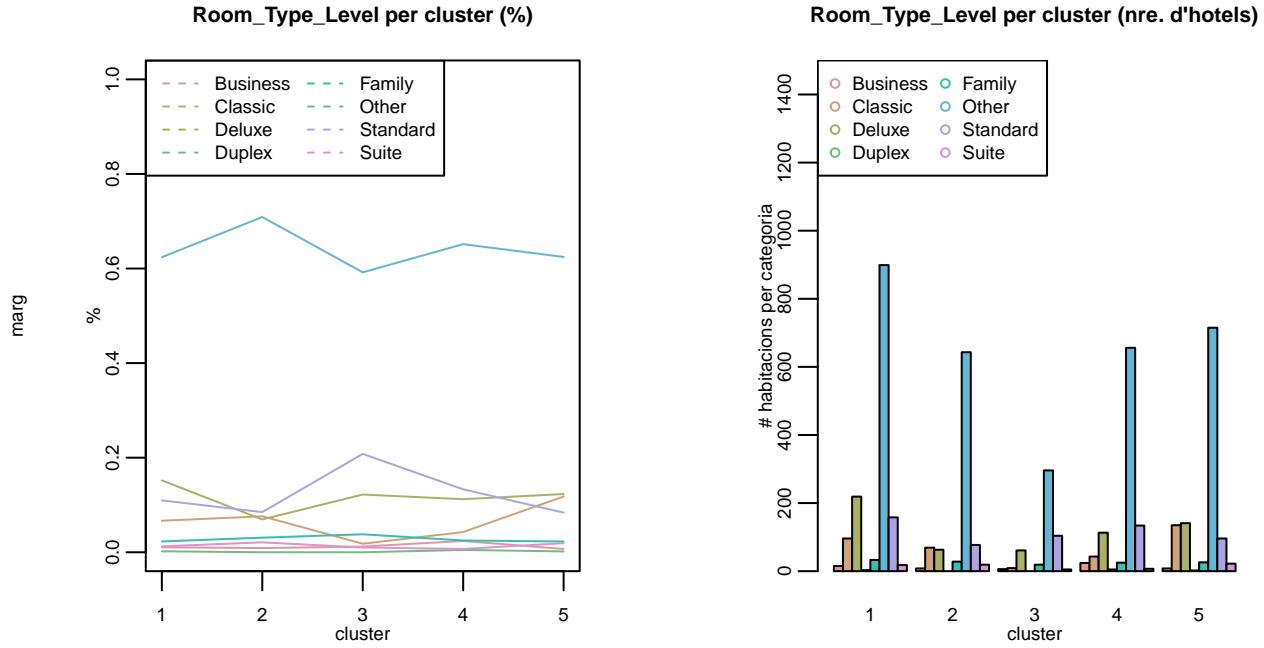


Figure 74: Profiling2 variable Room\_Type\_Level

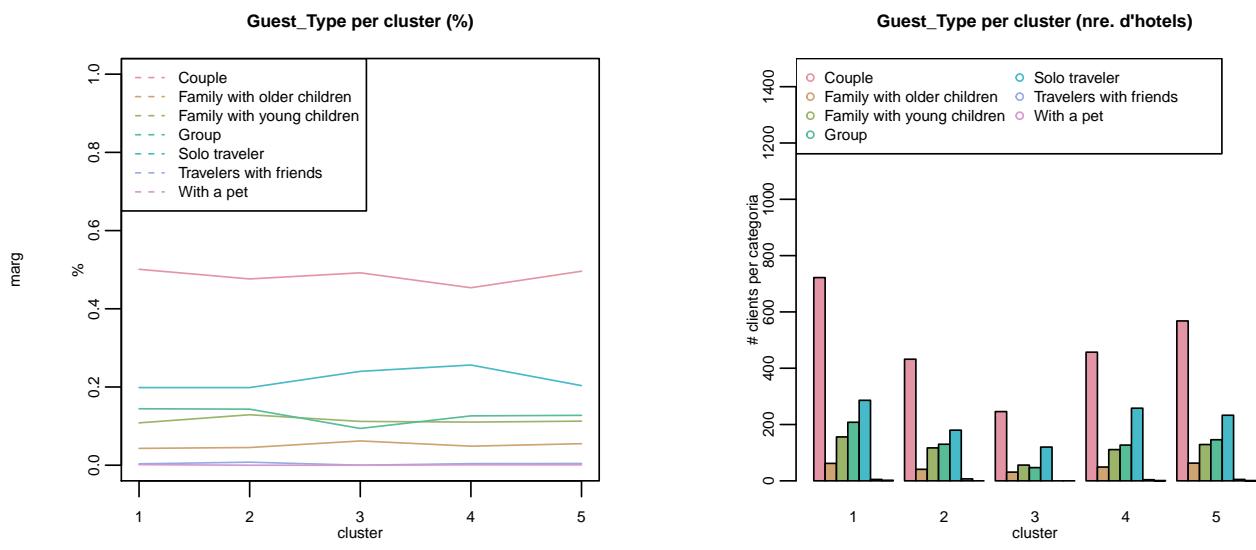


Figure 75: Profiling2 variable Guest\_Type

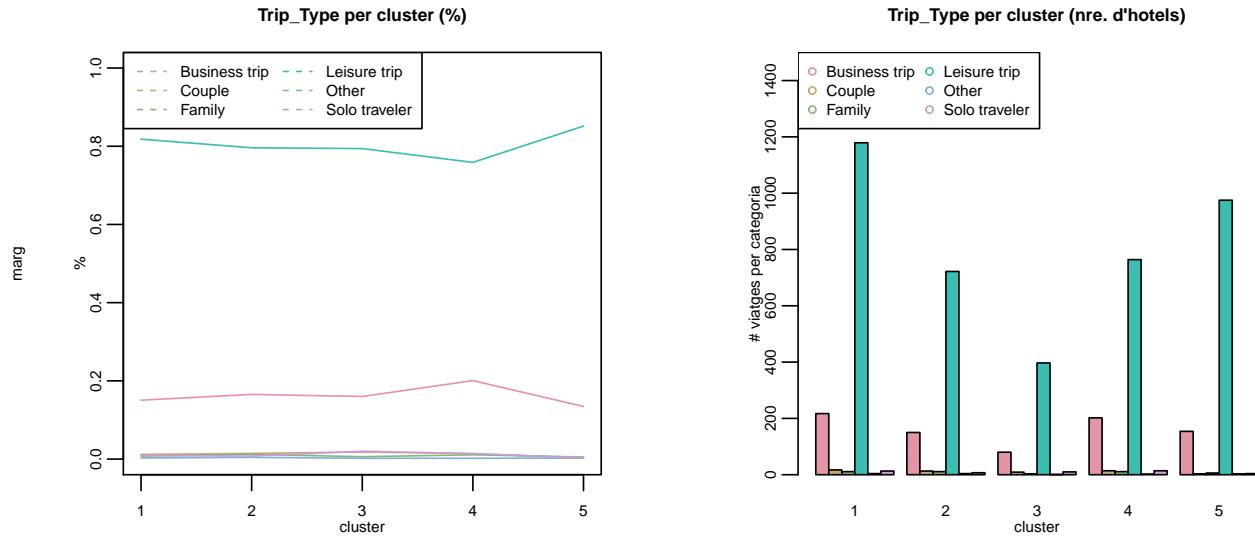


Figure 76: Profiling2 variable Trip\_Type

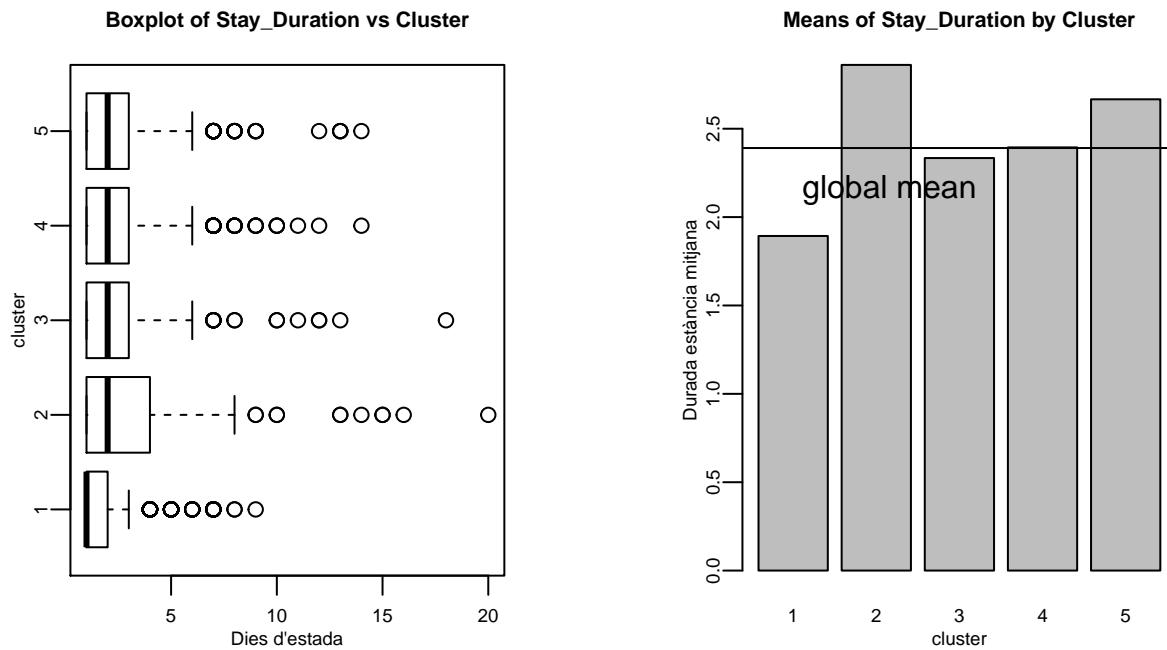


Figure 77: Profiling2 variable Stay\_Duration

representada a les dimensions superiors.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
	0.0	171.0	354.0	351.5	522.0	730.0
	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
	0.0	199.5	348.0	354.7	513.0	730.0
	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
	0.0	164.0	340.0	350.7	534.0	730.0
	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
	0.0	157.0	338.0	343.5	509.0	730.0
	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1	207	372	374	546	730	

A continuació ens fixem conjuntament en les variables *Is\_Hotel\_Holiday* i *Is\_Reviewer\_Holiday*, dicotòmiques les dues. Per a caracteritzar els clústers, construim dos diagrames de barres apilades (*Figure 78*) de manera que poguem 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.

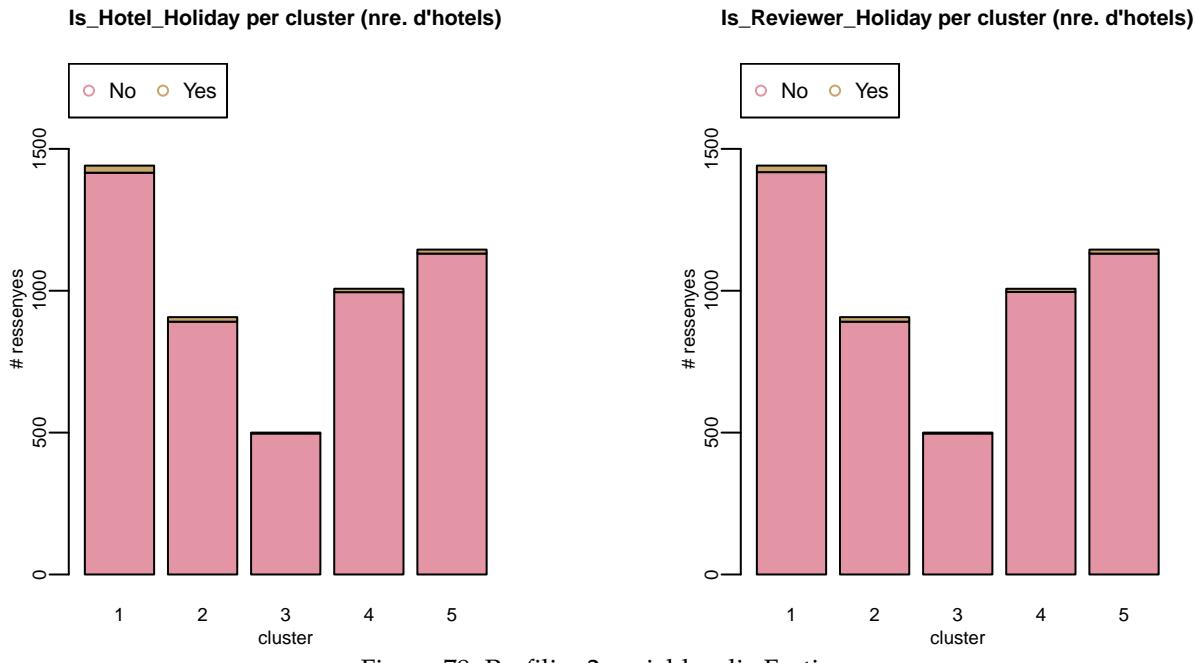


Figure 78: Profiling2 variables dia Festiu

Veiem com les proporcions es mantenen constants per a tots els clústers. Per a aquestes dues variables s'han reduït les diferències entre clústers, en comparació amb el profiling anterior.

A continuació, ens fixem en la variable *Total\_Number\_of\_Reviews* (*Figure 79*). En aquest cas, el resultat es veu altament modificat i observem com, el clúster 3 presenta un valor extremadament alt mentre que tots els altres conglomerats es mantenen per sota la mitjana (concentració de registres amb valors alts al clúster 3). En aquest cas ocurre el fenòmen contràri a l'anterior, les diferències entre clústers s'han fet més evidents.

Les següents dues variables, són *Review\_Is\_Positive* i *Review\_Positivity\_Rate*, relacionades amb el grau de positivisme de la ressenya (*Figure 80*). Veiem com, de nou, les diferències s'han accentuat respecte al profiling anterior.

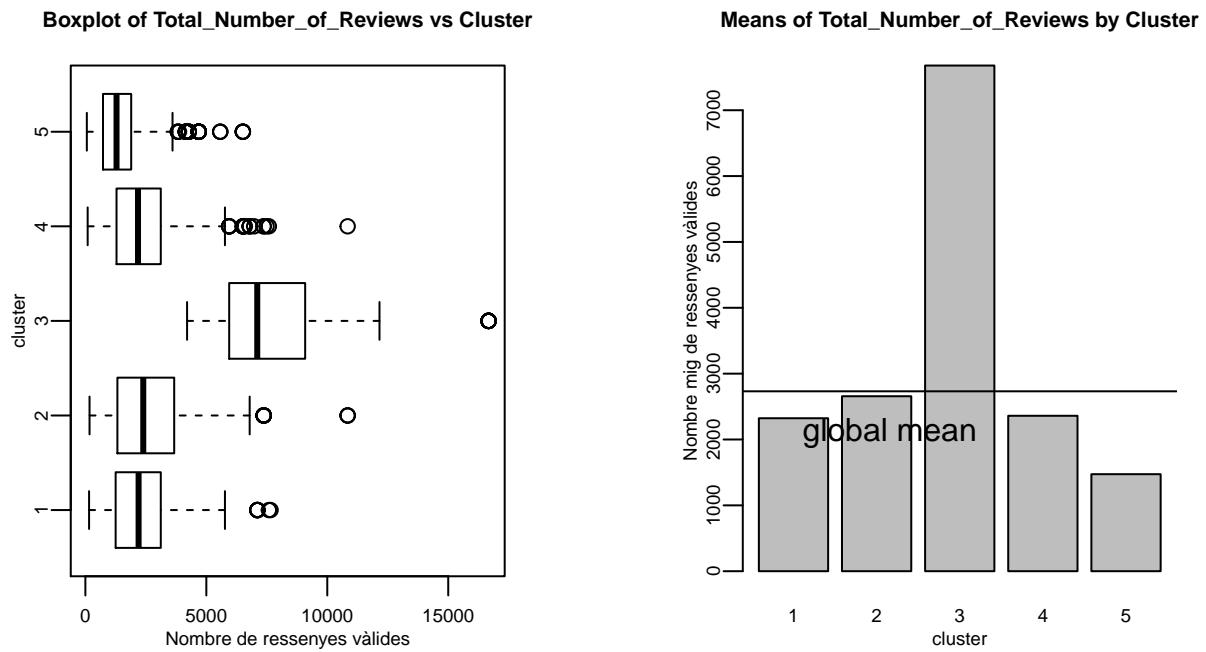


Figure 79: Profiling2 variable Total\_Number\_of\_Reviews

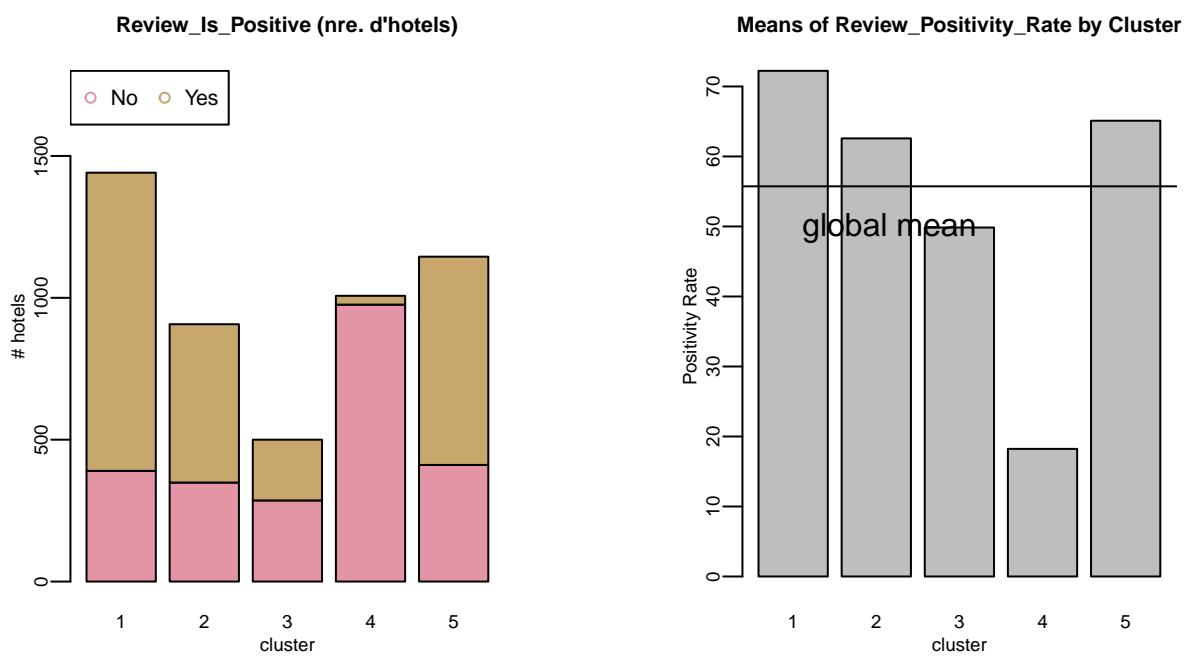


Figure 80: Profiling2 variables grau de positivisme de la ressenya

Observant el gràfic, podem concluir que:

- Els comentaris positius s'han concentrat als clústers 1, 2 i 5 (alt percentatge de ressenyes positives)
- El clúster 4 té gairebé totes les ressenyes i comentaris negatius (al profiling anterior aquest paper l'adoptava el clúster 3).

La següent variable és *Reviewer Nationality*. Els resultats són consistents amb la hipòtesi prèvia que el turisme predominant és intern (*Figure 81*), ja que als clústers 1 i 3, on la majoria d'hotels es troben a Gran Bretanya, també es concentren la majoria de clients de nacionalitat anglesa.

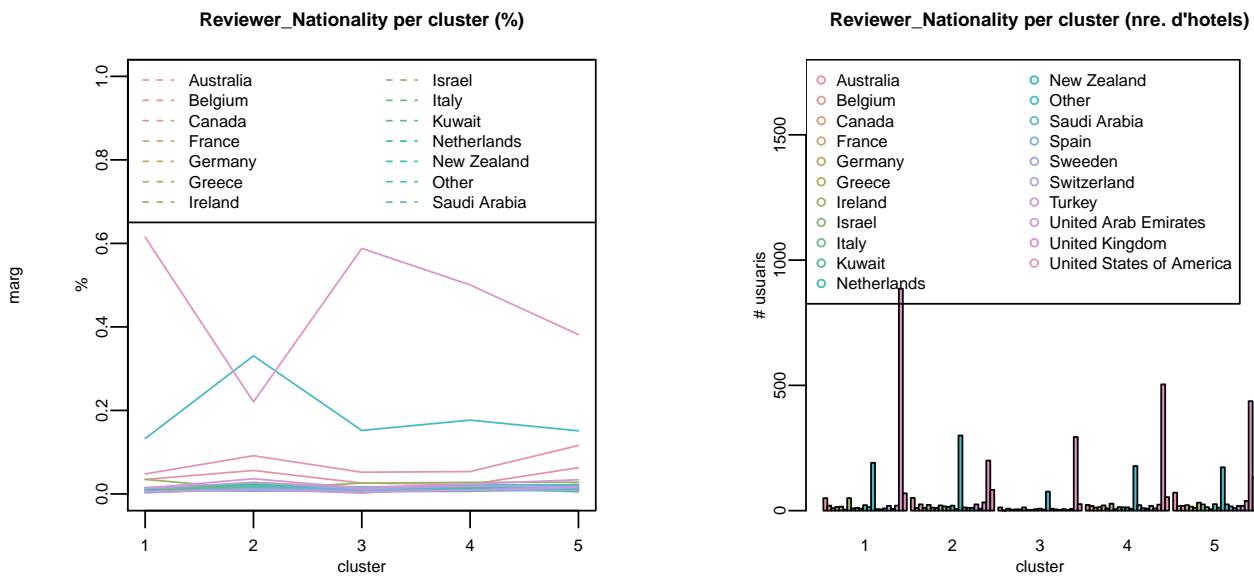


Figure 81: Profiling2 variable Reviewer\_Nationality

Les següents variables, si excloem les textuals, són *Review\_Total\_Negative\_Word\_Counts* i *Review\_Total\_Positive\_Word\_Counts*. Els resultats concordem amb les conclusions ofertes pels gràfics que reflectien el grau de positivisme (*Figure 82*). En aquest nou clustering, el conglomerat 4, absorbeix la majoria de ressenyes negatives.

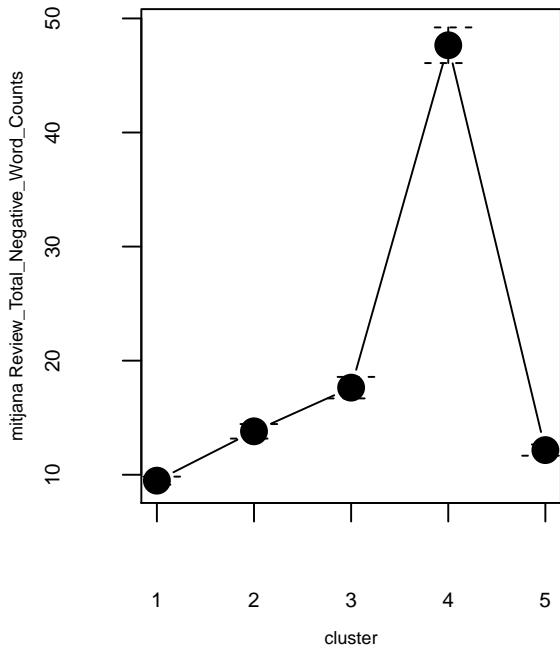
Tot seguit entrem a analitzar variables relacionades amb les puntuacions dels hotels (tot i ja tenir indicis del que obtindrem) (*Figure 83*). De nou, prenem conjuntament la puntuació mitjana que presentava l'hotel a finals de 2016, i la que han anat otorgant els usuaris de Booking que han escrit les ressenyes.

En efecte, les conclusions que podem obtenir d'aquests gràfics són calcades a les que hem obtingut analitzant *Review\_Is\_Positive*, *Review\_Positivity\_Rate*, *Review\_Total\_Negative\_Word\_Counts* i *Review\_Total\_Positive\_Word\_Counts*: clúster 4 pitjors valoracions, seguit d'aprop pel 3, i la resta amb valoracions elevades.

La següent variable que analitzem, és *Total\_Number\_of\_Reviews\_Reviewer\_Has\_Given*. Veiem com, en aquest clustering els individus del clúster 2 són els més actius a Booking, mentre que en l'anterior ho eren els del clúster 3 (*Figure 84*). La resta, a excepció del 5, es manté considerablement per sota la mitjana.

A continuació ens fixem en la variable *Additional\_Number\_of\_Scoring* (*Figure 85*). En aquest cas, la majoria de valors alts per a aquesta variable es concentren al clúster 3 mentre que, en el primer profiling, aquesta posició era compartida pel 3 i el 4.

**PlotMeans Total\_Negative\_Word\_Counts**



**Plot Means Total\_Positive\_Word\_Counts**

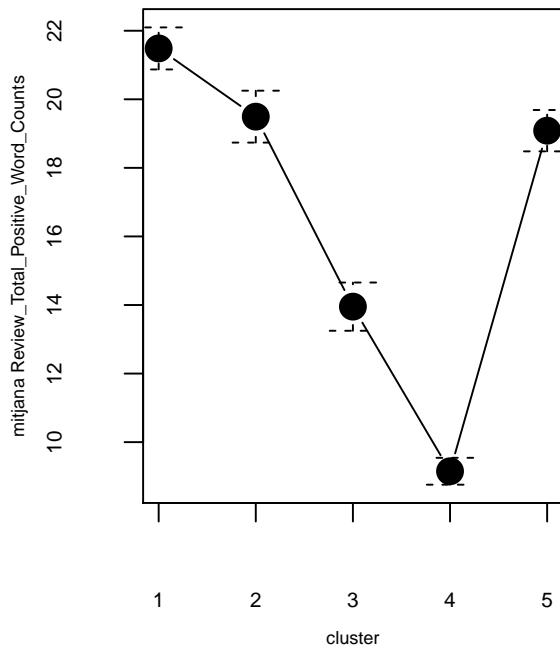
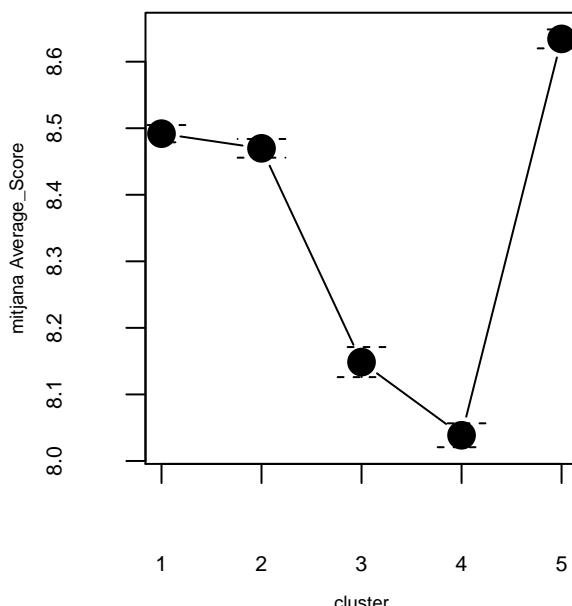


Figure 82: Profiling2 Nre.paraules comentari Negatiu/Postiu

**Plot Means Average\_Score**



**Plot Means of Reviewer\_Score**

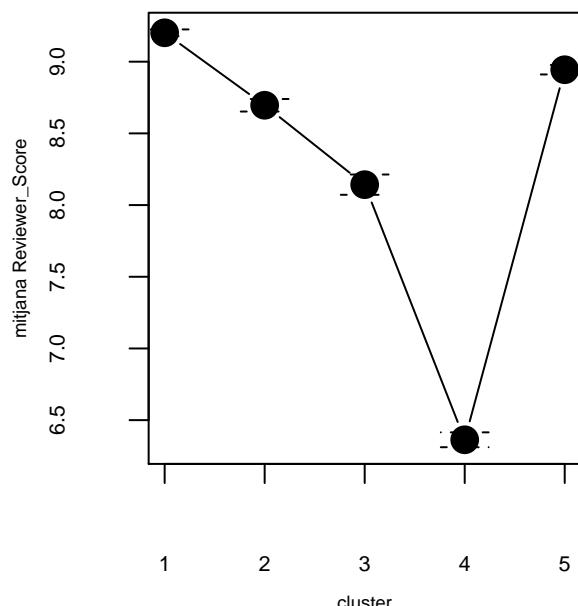
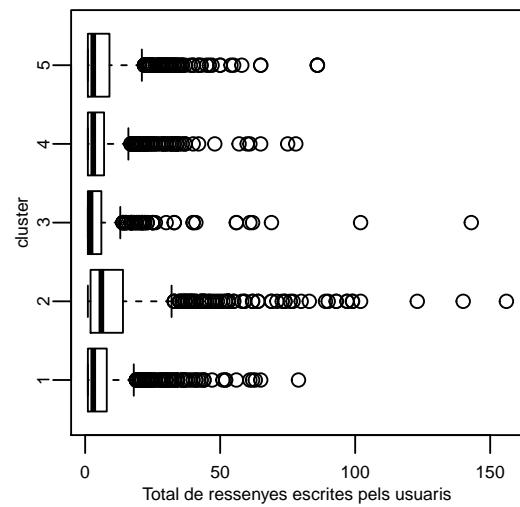


Figure 83: Profiling2 variable Valoracions

**Total\_Number\_of\_Reviews\_Reviewer\_Has\_Given vs Cluster**



**Total\_Number\_of\_Reviews\_Reviewer\_Has\_Given by Cluster**

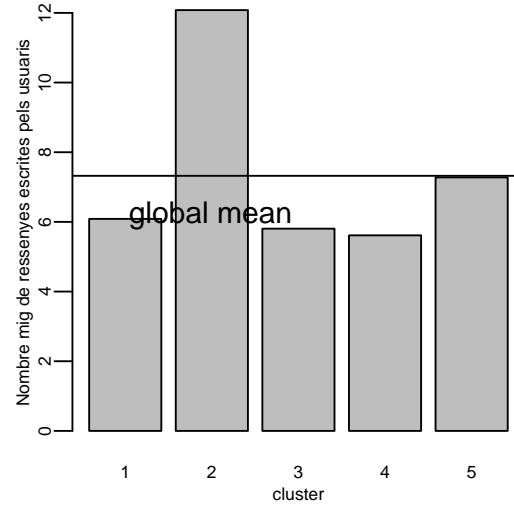
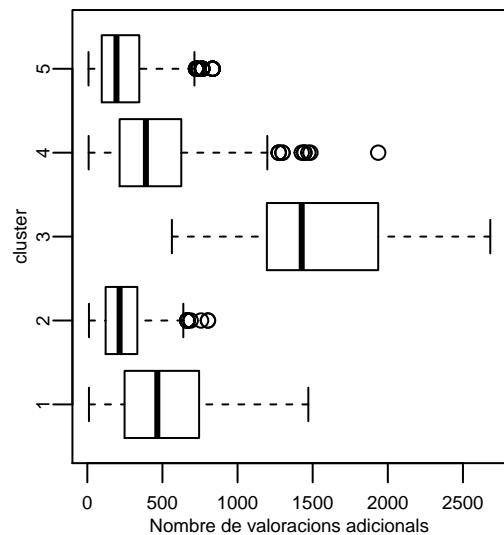


Figure 84: Profiling2 variable Total\_Reviews\_Given

**Boxplot of Additional\_Number\_of\_Scoring vs Cluster**



**Means of Additional\_Number\_of\_Scoring by Cluster**

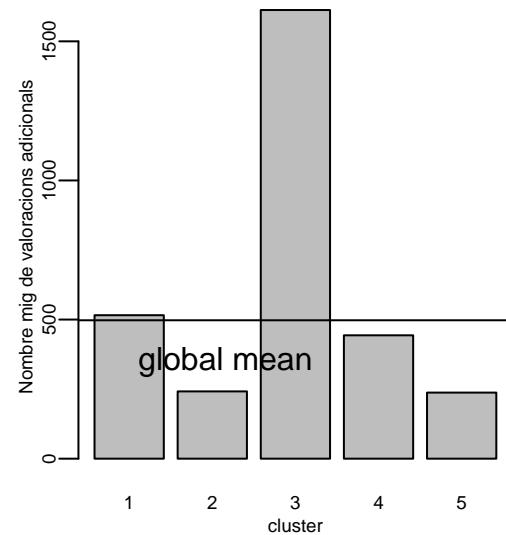


Figure 85: Profiling2 variable Additional\_Number\_of\_Scoring

Per últim, considerem la variable *Submitted\_from\_Mobile* (Figure 86). Observem molt poca variació entre clusters pel que fa a aquesta variable, a excepció d'un lleuger augment en el percentatge de ressenyes escrites des del mòvil al clúster 3.

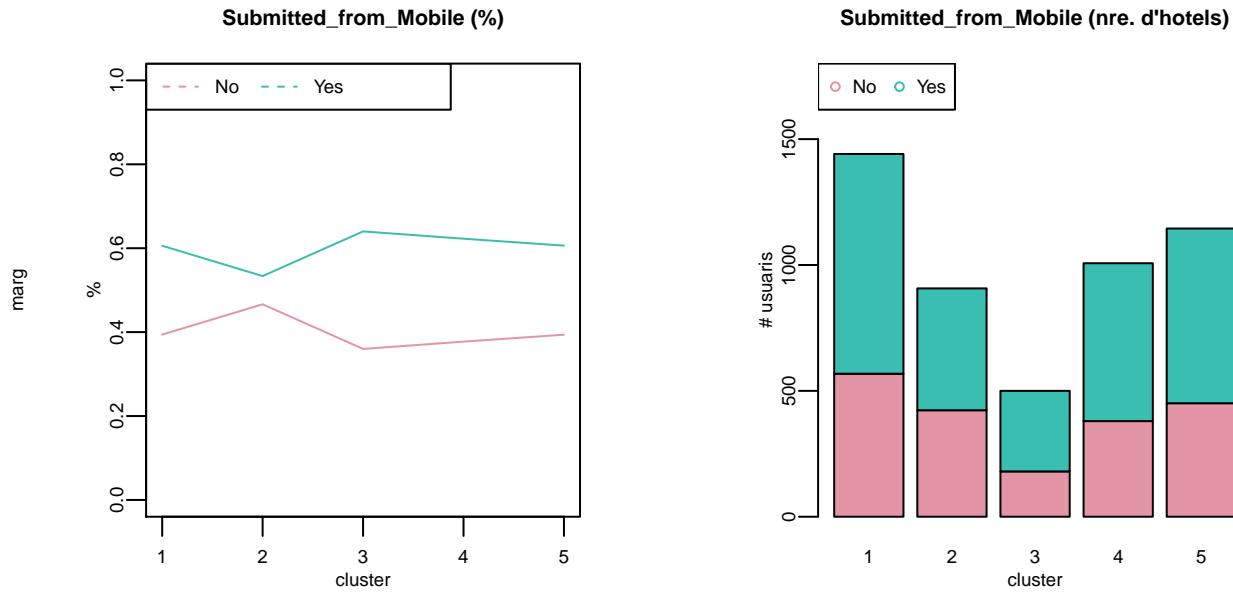


Figure 86: Profiling2 variable Submitted\_from\_Mobile

De nou, 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	Elegància Anglesa	Anàlegs als ben putuats del profiling anterior. Són ressenyes sobre hotels ubicats majoritàriament a Londres lluny del centre urbà. Aquests hotels són els que reben millors valoracions i el tipus de client més habitual són parelles que viatgen per motius d'oci. També són hotels que reben poques ressenyes i tampoc massa valoracions adicionals, cosa llògica ja que els seus clients són poc actius a la plataforma. Les estades també solen ser curtes en aquests hotels
2	En Familia	Hotels a Barcelona, Vienn i Milan no excessivament cèntrics però amb valoracions acceptables. Els més semblants del profiling anterior són els utilitaris a la perifèria. Aquests hotels normalment són per a llargues estades i hi abunden lleugerament més que en els anteriors les famílies amb nens petits/joves. En aquest grup predominen els clients de diferents nacionalitats (grup others) i, tot i que tampoc destaquen per rebre moltes valoracions/comentaris, els seus clients són els més actius a la plataforma
3	Londres mal valorats	Les ressenyes d'aquest segment estan associades amb hotels ubicats a Londres que no han rebut massa bones puntuacions
4	Els que no agraden	Anàlegs al grup 3 anterior

Clúster	Nom	Descripció
5	Experiència Urbana	Hotels amb valoracions mitjanes i poques valoracions, que es caracteritzen per estar situats al cor de París. Predominen les parelles en viatges d'oci. Aquests hotels també han rebut altres puntuacions

## Anàlisi Textual

Per tal de poder aprofitar les nostres dues variables Positive\_Review i Negative\_Review, que es basen en els comentaris tant positius com negatius que escriuen els clients dels allotjaments a Booking, realitzarem l'anàlisi textual amb l'objectiu de poder determinar quines són aquelles paraules més utilitzades i que més contribueixen a la nostra base de dades.

Per tal de poder realitzar aquest apartat necessitarem els paquets FactoMiner (ja utilitzat previament) i Xplortext.

Ja que no disposem de variables que ens indiqui quantes vegades apareix una paraula en cada registre, haurem de fer un recompte general per a cada paraula que apareix en les dues variables textuales que disposem, a partir de la funció **Textdata**.

Mitjançant aquesta funció podem saber quantes vegades apareix cada paraula de les que formen les nostres variables textuales. Hem escollit per tal d'acotar la recerca, que aquestes paraules hi han d'aparèixer un mínim de 150 vegades en els 5000 registres que disposem en la base dades. Un cop executada, ens hem trobat que la funció ens ha proporcionat algunes paraules poc rellevants i que per tant no podrien aportar informació en el posterior anàlisi. És per això que d'aquesta llista final hem tret algunes paraules com preposicions, articles o verbs sense relació amb les variables estudiades.

Un cop hem escollit les paraules a estudiar realitzem l'anàlisi de correspondència:

	eigenvalue	percentage of variance	cumulative percentage of variance
dim 1	0.2681131	2.880596	2.880596
dim 2	0.2532007	2.720378	5.600974
dim 3	0.2369859	2.546167	8.147141
dim 4	0.2267313	2.435992	10.583133
dim 5	0.2176028	2.337915	12.921048
dim 6	0.2132300	2.290935	15.211983
dim 7	0.2069069	2.222999	17.434982
dim 8	0.2024480	2.175093	19.610075
dim 9	0.1985697	2.133425	21.743500
dim 10	0.1952723	2.097998	23.841497
dim 11	0.1869476	2.008557	25.850054
dim 12	0.1868051	2.007026	27.857081
dim 13	0.1823228	1.958869	29.815950
dim 14	0.1812076	1.946887	31.762837
dim 15	0.1773792	1.905755	33.668591
dim 16	0.1762069	1.893160	35.561751
dim 17	0.1732263	1.861136	37.422888
dim 18	0.1718947	1.846830	39.269717
dim 19	0.1708756	1.835881	41.105598
dim 20	0.1675805	1.800478	42.906076
dim 21	0.1641178	1.763275	44.669352
dim 22	0.1632735	1.754204	46.423556
dim 23	0.1620243	1.740782	48.164338
dim 24	0.1584583	1.702469	49.866807
dim 25	0.1570214	1.687032	51.553839
dim 26	0.1567849	1.684491	53.238330
dim 27	0.1553964	1.669573	54.907903

dim 28	0.1519781	1.632847	56.540750
dim 29	0.1507381	1.619524	58.160273
dim 30	0.1487645	1.598320	59.758593
dim 31	0.1464464	1.573414	61.332007
dim 32	0.1453736	1.561888	62.893895
dim 33	0.1437013	1.543921	64.437816
dim 34	0.1422021	1.527814	65.965630
dim 35	0.1417711	1.523183	67.488814
dim 36	0.1399029	1.503111	68.991924
dim 37	0.1376111	1.478488	70.470412
dim 38	0.1374300	1.476542	71.946955
dim 39	0.1344081	1.444075	73.391030
dim 40	0.1318594	1.416692	74.807722
dim 41	0.1306342	1.403529	76.211251
dim 42	0.1293032	1.389229	77.600479
dim 43	0.1266698	1.360936	78.961415
dim 44	0.1247174	1.339959	80.301374

Un cop fet podem observar mitjançant els valors propis, com tenim moltes dimensions i per tal d'arribar a una variància del 80% necessitaríem 44 dimensions.

Nosaltres en el nostre analisi textual ens centrarem en unes poques dimensions, i en aquestes ens focalitzarem en veure les contribucions de les paraules a cada una de les dimensions estudiades.

	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5
air	1.38618700	0.2155324	0.007987311	0.001384074	38.0038741
service	1.24416605	0.5839723	27.006640107	4.112665853	0.8409517
staff	12.97221259	10.1602254	2.055653125	0.812924064	1.6494086
friendly	8.36164875	7.1889046	4.260257454	0.668613597	1.7205027
bed	9.51586918	3.2307428	1.279509769	0.127504475	7.7630039
excellent	4.86540571	1.9791362	0.103293580	11.178046279	0.2339528
amazing	2.92568386	0.5072217	1.276735623	9.048876491	4.0441400
room	10.57117976	0.8540372	2.899327379	0.170498352	3.2053598
helpful	6.82087249	5.3407621	3.244361667	0.607762264	1.4967588
comfy	3.81243807	3.1392387	1.573922095	0.104090160	8.5677728
location	2.03435776	1.0801985	1.418238671	10.501508600	0.2978584
station	0.06505078	10.9442113	1.531788762	0.104946405	0.7371295
hotel	1.54421039	5.8246768	0.048317618	5.598791545	0.2659308
metro	0.20938424	10.1764517	2.231118181	0.176551735	0.3163968
noisy	1.16465608	0.6669934	0.032368090	0.030189153	10.4792115

Mitjançant aquesta taula podem observar les 15 variables que més aporten o contribueixen, en les 5 primeres dimensions. Que una paraula aporti més a una dimensió indica que aquesta paraula tindrà molta influència en aquesta dimensió i que serà aquella paraula la que separarà més als individus o registres entre ells o més els diferenciarà.

Per tal de veure aquesta taula d'una manera més clara i entenedora podem fer un seguit de plots (*Figures 87, 88 & 89*) que ens mostraran aquestes contribucions d'una manera més gràfica:

En aquest gràfic es pot observar com les paraules en els comentaris dels clients que més contribueixen a la dimensió 1, són aquelles que estan més allunyades de l'eix Y, és a dir que es situen

## CA factor map

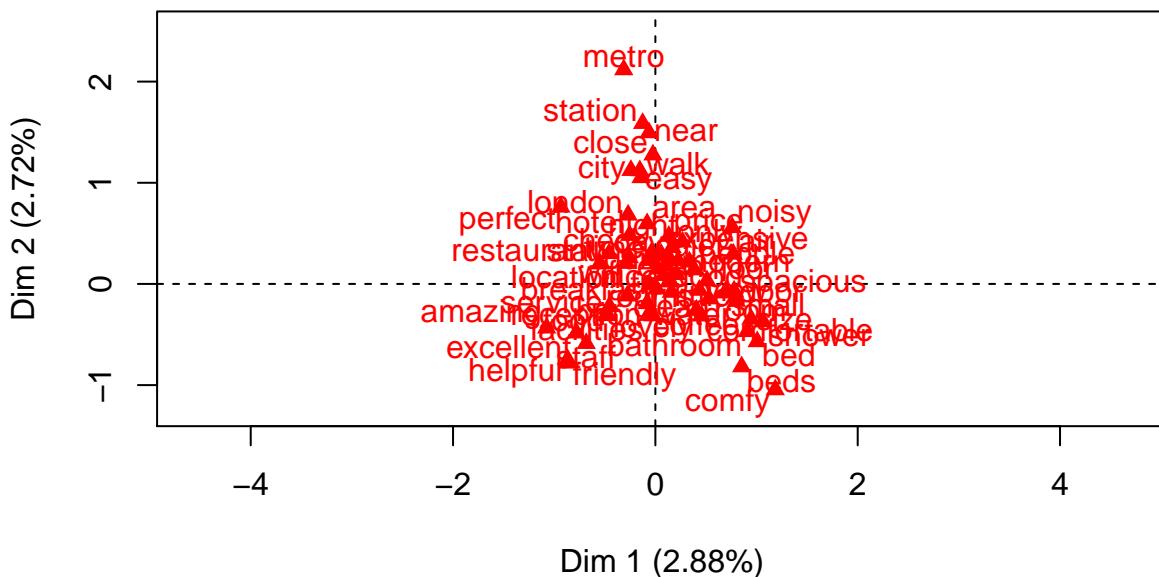


Figure 87: Contribucions D1 vs D2

més a prop dels extrems horitzontals. Per altra banda aquelles que estan més allunyades de l'eix X, com pot ser metro, estació, proper (totes elles relacionades) o confortable són algunes de les que més aporten a la dimensió 2.

Podem seguir analitzant el pes de les paraules de les nostres variables textuales en diferents dimensions, analitzarem les dimensions 3, 4 i 5.

D'igual forma que en el gràfic anterior, interpretarem aquest plot. En la dimensió 3 aquelles paraules que més aporten són menjar, servei i restaurant (relacionades amb la gastronomia i els àpats que prenen els clients als seus respectius viatges). Pel que fa a la dimensió 4 són les paraules excel·lent, car, o espectacular.

Per últim com hem dit analitzarem la dimensió vs les paraules de les variables textuais.

Per tal de fer el plot relacionarem la 5 amb qualsevol de les dimensions anteriors i ens fixarem en com de separades estan les paraules respecte a l'eix X, com més allunyades estiguin més contribuiran a aquesta dimensió. Com podem veure aire, sorollós i confortable són les paraules que més aporten a la cinquena dimensió.

### CA factor map

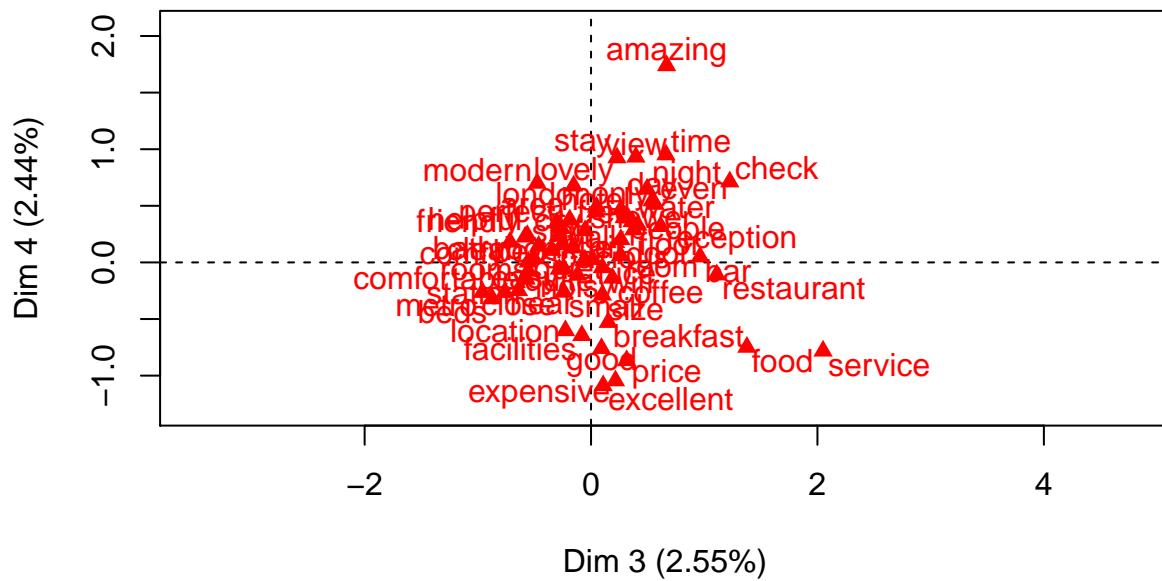


Figure 88: Contribucions D3 vs D4

### CA factor map

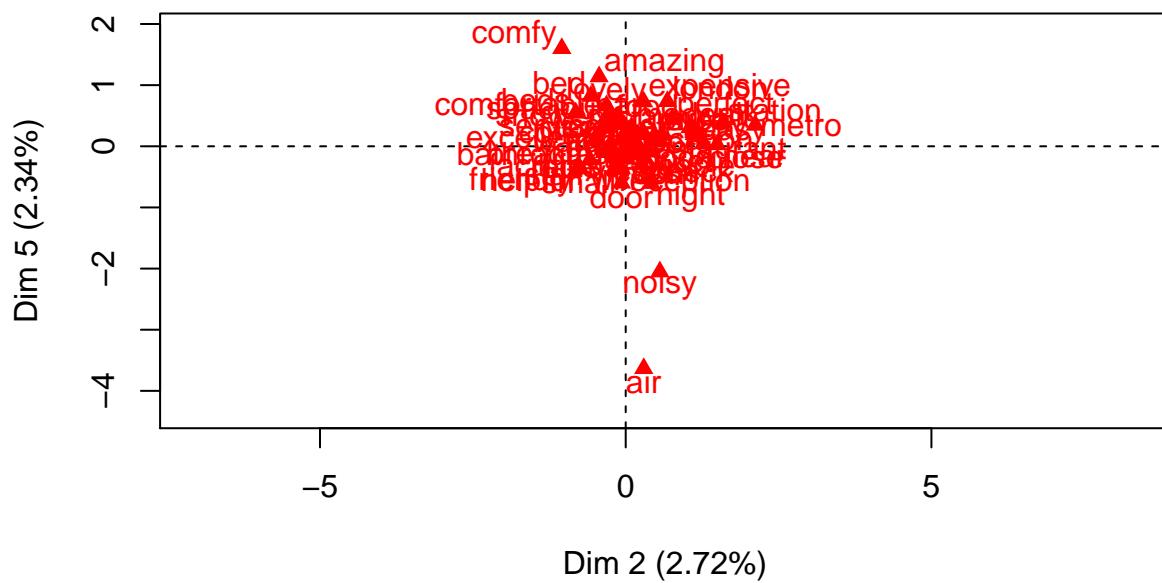


Figure 89: Dimensio vs nre. paraules

## Anàlisi Comparativa dels diferents mètodes i 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. En primer lloc, mencionar que els resultats obtinguts en cada apartat han seguit una llògica general de concordança al llarg de tot l'estudi. Totes les petites conclusions que hem anat elaborant en cada fase, s'han vist posteriorment reforçades per altres resultats. En relació a les dues branques principals de l'estudi que són els dos clusterings (amb el posterior profiling) es pot apreciar com molts dels perfils tenen un alt grau de similitud. La diferència més important que hem detectat es que, si fem servir les components de l'ACP al clústering, hi ha variables que s'estabilitzen (les diferències entre clusters es redueixen fins a nivells mínim), mentre que d'altres s'accentúen molt per a un clúster particular. D'aquesta manera el procés de profiling es fa més evident i les característiques amb les que etiquetem els segments són més rellevants.

Tot seguit, intentem resoldre les qüestions plantejades a l'inici en relació a les experiències dels usuaris de Booking. Fent servir aquesta petita mostra de dades hem pogut extreure les següents conclusions que, creiem, poden ajudar a millorar els viatges dels usuaris:

- En primer lloc, mencionar que per a viatges en parella, basant-nos en experiències prèvies d'altres usuaris, París és destinació ideal. Les parelles han valorat molt positivament la seva experiència en aquesta ciutat i recomanem estades llargues, de més de tres dies.
- Barcelona és un destí molt recomanable per a famílies amb nens, que busquen estades més aviat llargues.
- En general, els destins Gran Bretanya i Països Baixos obtenen valoracions més baixes en relació a la resta. En especial, per a estades curtes el servei als hotels ha rebut valoracions força dolentes.
- Tot i no estar segurs de tenir prou evidències, un nombre elevat de negocis a la rodona, en especial si el radi es àmpli, és indicatiu, en promig, de valoracions elevades. Una possible explicació és que, per als usuaris, la ubicació del hotel juga un paper fonamental en la seva experiència.

Un cop cobertes les conclusions que hem considerat, poden ser útils per als usuaris de Booking, també volem breument mostrar com el nostre anàlisi pot ajudar als hotels a l'hora de millorar el servei que ofereixen i satisfer els clients d'una manera més eficient.

- En primer lloc, a nivell general s'ha observat que les experiències negatives són les que més comentaris i valoracions generen. En aquest sentit, tot i que les valoracions no siguin bones, això representa una gran avantatge per als hotels amb una pitjor performance ja que són els que disposen de més informació sobre el client i perquè no ha quedat satisfet. És molt important que els hotels prenguin nota de tots els comentaris, valoracions, opinions dels seus clients, especialment aquells que no han quedat contents ja que són els que aporten informació més valuosa per a millor (quin o quins serveis han puntuat pitjor, paraules clau que han deixat a la ressenya...). D'aquesta manera l'hotel pot centrar esforços en millorar el que, a priori fa pitjor.
- Seguidament, hem detectat un gran mercat de usuaris descontents amb els hotels d'Amsterdam i Londres que escullen per motius de viatges de negocis o per estades curtes. En aquest sentit, aquests hotels més allunyats del centre que a priori ofereixen un preu més econòmic

tenen un gran marge de millorar per a intentar contentar aquest collectiu. Una possible estratègia seria centrar-se en les demandes d'aquest perfil de client (molts viatgen en solitari o en parella i escullen habitacions de la categoria Classic o Deluxe) i intentar corregir les males actuacions ja que, al tenir un collectiu en general descontent, el fet de poguer satisfer les seves demandes, representarà una gran avantatge competitiva.

- Per últim, mencionar que en relació als hotels de Barcelona hem pogut veure com, les estades que es contracten generalment són força llargues i el perfil de client és molt internacional. Caldrà que els hotels estiguin preparats per adaptar-se a les demandes del tipus de client que sol contractar aquestes estades, predominen les parelles però aquest tipus de viatges el soLEN fer també famílies amb nens petits. D'aquesta manera aspectes com, oferir serveis per als infants, pensar en que el client hi estarà hospedat bastants dies i no volem que s'aburreixi o tenir personal que domini moltes llengües, poden ser aspectes clau per a tenir èxit.

## Pla de treball real

Per tancar, annexem el pla de treball real que hem seguit durant aquest semestre per a la realització del treball de l'assignatura.

### Diagrama de Gantt

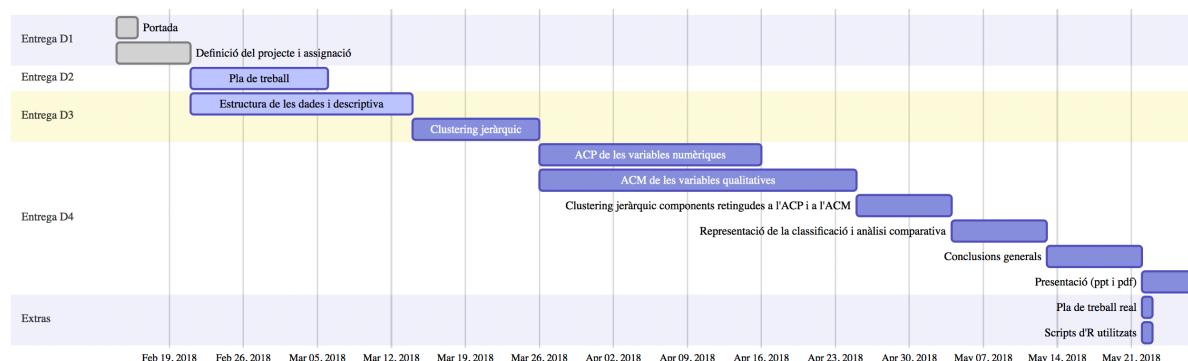


Figure 90: Diagrama de Gantt

## Distribució de tasques

		Hugo Allès Pons	Carles Blanco Conde	Aleix Fibla Salgado	Víctor Miranda Hernández	Pablo Morante López	Antoni Ramoneda Montoya	Oriol Rovira Tauler	Aleix Salvador Barrera
<b>D1</b>	Portada	x	X						
	Definició del projecte	x	x	x	X	x	x	x	x
<b>D2</b>	Pla de treball			x		X	x		x
<b>D3</b>	Estructura i descriptiva de les dades	x						X	x
<b>D4</b>	Cluster jeràrquic				x	X	x		
	ACP de les variables numèriques		x				X	x	
	ACM de les variables qualitatives	X		x	x				
	Clustering jeràrquic sobre les components factorials retingudes a l'ACP i a l'ACM		x	X	x		x		
	Representació de la classificació i anàlisi comparativa	x						X	x
	Conclusions					x			x
	Pla de treball REAL			x		X	x		
	Scripts d'R utilitzats	x	x	X	x	x	x	x	x

Figure 91: Distribució de tasques