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 mul- tivariant 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**

**Introducció 3**

Pla de treball 5 DiagramadeGantt . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 5 Distribuciódetasques . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 5 Pladeriscos . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 5

Estructura de les dades i anàlisi descriptiva 7 Motivaciódeltreball . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 7 Descripció formal de l’estructura de les dades . . . . . . . . . . . . . . . . . . . . . . . . . . 7 Anàlisiexploratòriinicialdelesvariables . . . . . . . . . . . . . . . . . . . . . . . . . . . . 11 Procésdepreprocessament. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 24 Anàlisidescriptivaunivariantpostpreprocessament . . . . . . . . . . . . . . . . . . . . . . 29

**Clúster jeràrquic 34**

**Profiling dels clústers 37**

ACP de les variables numèriques 54 Gràficsd’individus . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 56 Gràficsdevariables . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 58 Biplotsmixtesdevariablesiindividus . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 65

**ACM de les variables qualitatives 68**

∗Tots els arxius per a replicar l’anàlisi es troben al compte de Github https://github.com/aleixfiblasalgado

1

**Clustering jeràrquic sobre les components factorials retingudes a l’ACP i a l’ACM 75**

**Profiling del Clúster Jeràrquic sobre ACP 92**

**Anàlisi Textual 104**

**Anàlisi Comparativa dels diferents mètodes i conclusions generals 108**

Pla de treball real 109 DiagramadeGantt . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 109 Distribuciódetasques . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 110

2

**Introducció**

Aquest treball s’ha desenvolupat amb l’objectiu de millorar les experiències de viatges en l’àmbit de la industria 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 materia 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 otor- gades 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 aleatòriament un subconjunt de 5.000 ob- servacions 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 · n). En aquest sentit, presentem la Taula 1 que mostra com es distribuieixen 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 %

3

Trip\_Type Total\_Number\_of\_Reviews\_Reviewer\_Has\_Given Total\_Number\_of\_Reviews Submitted\_from\_Mobile Stay\_Duration Room\_Type\_Level Reviewer\_Score Reviewer\_Nationality Review\_Total\_Positive\_Word\_Counts Review\_Total\_Negative\_Word\_Counts Review\_Positivity\_Rate Review\_Is\_Positive

s g n i s s

Review\_Date Positive\_Review Negative\_Review i m

Is\_Reviewer\_Holiday Is\_Hotel\_Holiday id Hotel\_Name Hotel\_lng Hotel\_lat Hotel\_Country Hotel\_City Guest\_Type Days\_Since\_Review Businesses\_5km Businesses\_1km Businesses\_100m Average\_Score Additional\_Number\_of\_Scoring

0 1000 2000 3000

variable Figure 1: Histograma dels valors missing

4

**Pla de treball**

En aquest capítol es resumeix la organització del treball i les tasques que es persegueix implemen- tar, 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 Gantt 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 Dia- grama de Gantt.

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).

5

*Oriol Victor*

*Pablo*

*Antoni Miranda*

*Morante*

*Ramoneda*

*Rovira Hernández*

*López*

*Montoya*

*Tauler*

*Aleix Salvador Barrera*

***D1***

*Hugo Allès*

*Carles*

*Aleix Fibla Pons*

*Blanco Conde*

*Salgado*

Portada x X x Definició del projecte x x X x x x x x D2 Pla de treball x x x x X

***D3***

Estructura i descriptiva de les dades

x x x X X

Clúster jeràrquic x x X x

***D4***

ACP de les variables numèriques

x x x X x

ACM de les variables qualitatives

X 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 x X x

Conclusions x x x x x x x X Pla de treball REAL x x x X x x x x

Scripts d’R utilitzats x x x X x x x x

Figure 3: Plantilla de Tasques

*Possible Risc Com prevenir-lo Com gestionar-lo*

*Un membre del grup es posa malalt i no pot fer la seva part de la tasca*

Les tasques assignades als membres del grup no seran individuals. Així doncs si un no pot fer-la ,l’altre persona a la que se li ha assignat la podrà fer.

S’haurà de reajustar l’assignació de treballs per tal de compensar la pèrdua.

Compartir-ho entre tots els membres del grup en un Pèrdua del treball o parts d’aquest

espai virtual tots els avenços del treball (Google drive).

Assegurar-nos de que tots els components pujant els avenços al espai virtual.

*Entregues o scripts mal realitzats*

Per a cada tasca o entrega hi haurà un grup de 2 persones que s’encarregaren de revisar la feina realitzada.

En el repartiment de tasques també sortiran en cada entrega el grup que s’encarregarà de la revisió de la feina.

No entregar a temps les tasques Fer els scripts a classe i les entregues amb temps.

Seguir el diagrama de Gantt, en el qual ja prevenim possibles contratemps deixant uns dies de marge.

*Distribució del treball*

inadequadament Revisar amb antelació la plantilla de tasques.

En cas de que algun component del grup tingui molta feina i altres el contrari, haurem de tornar a fer la plantilla de riscos.

*Discussions internes per diferències conceptuals*

En cas d’haver una discussió els dos membres podran exposar les seves idees i s’escollirà en majoria mitjançant una votació.

*Desobediència d’algun membre del grup en les indicacions grupals i actitud individualista*

Les decisions finals sempre seran fetes amb un consens del grup.

S’escollirà un líder per majoria, que pot anar variant al llarg del treball. Aquest s’encarregarà 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.

*Errades ortogràfiques a l’informe*

Intentar que totes les decisions es facin de forma que tot el grup estigui d’acord.

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

6

**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 exploratòri 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 com- panyia 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 inten- ció 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’oferir 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

7

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

8

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 Book-

ing. 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 mit- jana 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 Nega- tive Word Counts). a. rang: {0 , 100} b. rol: variable resposta

9

21. Reviewer Nationality (qualitative): Nacionalitat de l’usuari que ha escrit la ressenya.

a. modalitats: {“” , Abkhazia Georgia , Albania , Andorra , Angola , Argentina , Armenia , Australia , Austria , Azerbaijan , Bahrain , Bangladesh ,. . . } (123 modalitats) b. rol: variable explicativa c. missing\_code: “”

22. Negative\_Review (text): Ressenya negativa escrita per l’usuari.

**a. rol: text b. missing\_code: “Na”**

23. Review\_Total\_Negative\_Word\_Counts (integer): Nombre total de paraules a la ressenya nega-

tiva escrita per l’usuari. a. rang: {0 , 372} b. rol: variable explicativa

24. Positive\_Review (text): Ressenya positiva escrita per l’usuari.

**a. rol: text b. missing\_code: “Na”**

25. Review\_Total\_Postive\_Word\_Counts (integer): Nombre total de paraules a la ressenya positiva

escrita per l’usuari. a. range: {0 , 247} b. rol: variable explicativa

26. Average\_Score (numeric): Valoració mitjana de l’hotel a la pàgina de Booking a data 31 de

desembre de 2016. a. range: {5.2 , 9.6} b. rol: variable explicativa

27. Reviewer\_Score (numeric): Valoració global otorgada per l’usuari que ha escrit la ressenya.

a. range: {2.5 , 10} b. rol: variable explicativa

28. Total\_Number\_of\_Reviews\_Reviewer\_Has\_Given (integer): Nombre total de ressenyes a la web

de Booking escrites per l’usuari. a. range: {1 , 156} b. rol: variable explicativa

29. Additional\_Number\_of\_Scoring (integer): Nombre total de valoracions adicionals vàlides so-

bre diferents aspectes de l’hotel. a. range: {8 , 2682} b. rol: variable explicativa

30. Submited\_from\_Mobile (qualitative): Variable binària que indica si la ressenya s’ha pujat a

Booking via telèfon mòvil. a. modalitats: {0: No , 1: Yes} b. rol: variable explicativa

10

Anàlisi exploratòri 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 preprocessa- ment.

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 interpretem. En aquest apartat, pretenem aconseguir aquest objectiu a partir de gràfics estadístics univariants de les nostres vari- ables.

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.

**Hotel\_Country (Raw)**

AT ES FR GB IT NL

AT

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 vari- able 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

11

FR

ES

GB

IT

NL

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)**

F

3000

Amsterdam Amsterdam Zuidoost Barcelona 2500

Boulogne Billancourt Donauinsel El Prat de Llobregat Fitzrovia 2000

London Milan y c n e u q e r

Paddington Paris 1500

Paris 06 Paris 12 Vienna Vincennes

1000

Woodford Green

500

0

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. Consid- erem 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

12

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

**Hotel\_lat (Raw) Hotel\_lng (Raw)**

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 51oN, el qual passa per països com el Regne Unit, França o Bèlgica. D’altra banda, mencionar que en el nostre estudi només hi trobem hotels entre les latituds 41o i 52o N, els quals coincideixen, òbviament amb les latituds de la majoria dels països europeus.

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

A continuació, passem a analitzar les variables Businesses\_100m, Businesses\_1km, Businesses\_- 5km. A primera vista, ja veiem que aquestes tres variables apareixen mal codificades (les tenim com a factors i haurien de ser numèriques). En conseqüènia, les apartem per analitzar-les un cop haguem fet el preprocessament i les tinguem en el tipus adequat. Tanmateix sabem que tenim 23 valors missings per a aquestes variables que es desprenen de la falta d’informació de mesures de localització i coincideixen en les observacions (caldrà tenir-ho en consideració per al preprocessa- ment).

Seguidament, considerem la variableRoom\_Type\_Leveli observem quins són els tipus d’habitacions més freqüents (Figure 8). De nou, la variable és qualitativa i repliquem el procediment descrit an- teriorment 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

13

**Room\_Type\_Level (Raw)**

F

3500

Ambassadors Art Business 3000

Business Class City Classic Deluxe 2500

Duplex Executive Family

y c n e u q e r

Luxury 2000

NULL Premium Privilege Standard 1500

Studio Suite Superior

1000

500

0

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’estandarització dels noms de les habitacions (cada hotel pot fer servir noms diferents per a referir-se al mateix tipus d’habitació). Tot i això, si ens quedem amb les denominacions de les que disposem, obtenim que les tres més freqüents són: Standard, Superior, Deluxe i Classic. Finalment, dir que tenim un nombre considerable de modalitats per a la variable i, potser, hauríem de considerar reduïr-lo a la fase de preprocessament.

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

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

14

4037 14 Solo traveler 48

**Guest\_Type (Raw)**

F

Couple Group With a pet

**Trip\_Type (Raw)**

Couple

Business trip Family with older children Family with young children Group Solo traveler Travelers with friends With a pet

Couple 0 0 0 2

Family with older children Family with young children Leisure trip NULL

y c n e u q e r

0 0 5 1

0

Solo traveler

0 0 1

0 0 5

0

Business trip Leisure trip Figure 9: Barplot perfils 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 variableDays\_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 codifi- cades com a numèriques, quan interessaria més tenir-les com a factors. En aquest sentit, plantegem

15

**Stay\_Duration (Raw)**

0 0 5 3

0 0 0 3

0 0 5 2

y c n e u q e r

0 0 0 2

F

0 0 5 1

0 0 0 1

0 0 5

0

5 10 15 20

# Dies 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 varible Is\_Reviewer\_Holiday que coincideixen.

Seguidament, analitzem la variable Total\_Number\_of\_Reviews. Aquesta apareix codificada cor- rectament 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\_Pos- itive, 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\_Positiv- ity\_Rate i un gràfic conjunt (histograma) d’aquesta amb Review\_Is\_Positive (Figure 13).

16

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

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

No

No Yes

Yes

0

1

0

1

Figure 11: Pie Chart Festivitats (Raw)

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

0 0 4 1

0 0 2 1

0 0 0 1

y c n e u q

0 0 8

e r F

0 0 6

0 0 4

0 0 2

0

0 5000 10000 15000

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

17

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

**Review\_Is\_Positive (Raw)**

**Review\_Positivity\_Rate (Raw)**

0 No Yes 0

1

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

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

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

A continuació tenim la variable Reviewer\_Nationality qualitativa i amb 123 modalitats. Com a conseqüència d’aquest nombre tan elevat de nivells del factor, els gràfics es tornen il.legibles.

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

18

0 2 1

0 0 0 1

0 0 y c n e

8

u q e r F

0 0 6

0 0 4

0 0 2

0

0 20 40 60 80 100

69 66 Israel Italy

64 59 Kuwait Spain

48 48 Turkey New Zealand

45 43 Greece Sweden

41 41 South Africa Romania

39 37 Russia Singapore

37 33 Poland India

32 30 Qatar Austria

30 28 China Portugal

28 26 Egypt Finland

25 25 Lebanon Czech Republic

24 21 Norway Denmark

21 20 Brazil Hong Kong

19 19 Hungary Cyprus

19 18 Thailand Bahrain

18 16 Oman Croatia

16 15 Malaysia Malta

15 14 Japan Serbia

13 13 Luxembourg Bulgaria

12 11 Iran Slovakia

11 11

Indonesia 10 10 Pakistan South Korea

10 10 Iceland Lithuania

9 8 Nigeria Latvia

19

8 7 Philippines Slovenia

7 7 Isle of Man Taiwan

6 6 Azerbaijan Estonia

5 5 Guernsey Kenya

5 5 Sri Lanka Ukraine

5 5 Gibraltar Jersey

4 4 Jordan Mauritius

4 4 Mexico Albania

4 3 Argentina Bermuda

3 3 Colombia Peru

3 3 Zambia Angola

3 2 Armenia Bangladesh

2 2 Barbados Chile

2 2 Iraq Kosovo

2 2 Macedonia Moldova

2 2 Monaco Morocco

2 2 Puerto Rico Tunisia

2 2 Uganda Abkhazia Georgia

2 1 Andorra Bosnia and Herzegovina

1 1 Botswana (Other) 1 24

Seguidament, comentem de manera conjunta les variables Negative\_Review, Review\_Total\_Neg- ative\_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-

20

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

**Review\_Total\_Negative\_Word\_Counts (Raw) Review\_Total\_Positive\_Word\_Counts (Raw)**

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\_Scorei Reviewer\_Score, ambdues de gran rellevàn- cia 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).

21

**Average\_Score (Raw) Reviewer\_Score (Raw)**

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

**Total\_Number\_of\_Reviews\_Reviewer\_Has\_Given (Raw)**

0 0 0 4

y c n e

0 0 0 3

u q e r F

0 0 0 2

0 0 0 1

0

0 50 100 150

# Ressenyes vàlides 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 en questió al portal de Booking (Figure 17).

22

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

**Additional\_Number\_of\_Scoring (Raw)**

0 0 8

0 0 6

y c n e u q e r F

0 0 4

0 0 2

0

0 500 1000 1500 2000 2500

# Ressenyes vàlides 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 exploratòri de dades considerem la variable Sub- mitted\_from\_Mobile binària. De nou, la variable binària apareix codificada com a numèrica i caldrà tractar-la a la fase de preprocessament per codificar-la correctament. Tanmateix, podem emprar el mateix procediment que per a les altres vairables binàries i construir un pie chart (Figure 18).

Observem com predominen les ressenyes escrites des del telèfon mòbil.

23

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

No Yes

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 vari- ables numèriques enteres. D’altra banda les variables Is\_Hotel\_Holiday, Is\_Reviewer\_holiday, Sub- mitted\_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 textu- als Positive\_Review i Negative\_Review no s’ajusten exactement a un factor amb diferents nivells així que les podríem codificar com cadenes de caràcters. Tanmateix, com que només les farem servir en un àmbit reduït de l’anàlisi i, el seu estat actual no representa un problema, hem considerat mantenir-les com a factor per convertir-les a caràcter quan realitzem l’anàlisi textual.

Un cop tenim totes les variables en el tipus d’R convenient, ens centrem en les modalitats de les variables qualitatives. Observem els diferents nivells que presenten, a partir de definir una funció que selecciona les variables de la base de dades segons de la seva classe, per tal d’extreure només els factors i mirar els seus nivells (recordem que cal obviar les variables textuals). Com a filtre adicional, considerem tractar únicament aquells factors amb nombre de nivells inferior a 150. A banda de les variables textuals, la variable Hotel\_Name tampoc la considerem ja que cada nom és únic i serveix per a identificar completament un registre, no té sentit aplicar cap tipus d’agrupació de modalitats en aquest cas.

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

24

0

1

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 re- dundàncies entre aquests, si podem eliminar algun, o si simplement les etiquetes no són conve- nients. En general, és recomanable recategoritzar factors amb nivells que reflexen el mateix, o aquelles variables amb un gran nombre de categories per a facilitar anàlisis posteriors.

En primer lloc, assignem etiquetes (No, Sí) a les variables dicotòmiques. Seguidament, plantegem redefinir les categories de la resta de variables qualitatives de la taula anterior. Un cop arribats a aquest punt, és convenient mencionar que en aquest procediment es realitzarà el tractament dels valors missing de les variables categòriques. Mentre que, per a les variables numèriques necessitarem realitzar una imputació de valors, aquí podem simplement

25

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, Donauin- sel, El Prat de Llobregat, Fitzrovia, London, Milan, Paddington, Paris, Paris 06, Paris 12, Vienna, Vincennes, Woodford Green.

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

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

• ANTICS NIVELLS: Ambassadors, Art, Business, Business Class, City, Classic, Deluxe, Du- plex, Executive, Family, Luxury, NULL, Premium, Privilege, Standard, Studio, Suite, Supe- rior.

• NOUS NIVELLS: Deluxe{Ambassadors, Art, Deluxe, Executive, Luxury, Premium, Privi- lege, Superior} , Business{Business, Business Class} , Classic{City, Classic} , Duplex{Duplex} , Family{Family} , Other{NULL} , Standard{Standard, Studio}, Suite{Suite}.

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

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

La següent variable és Trip\_Type. Aquest cas és similar a l’anterior, caldrà recodificar els valors missing (els tenim com a NULL) com “Other”. La correspondència entre els nivells antics i els nous és:

• ANTICS NIVELLS: Business trip, Couple, Family with older children, Family with young children, Leisure trip, NULL, Solo traveler.

• NOUS NIVELLS: Business trip{Business trip}, Couple{Couple}, Family {Family with older children, Family with young children}, Leisure trip{Leisure trip}, Other{NULL}, Solo trav- eler{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:

26

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

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

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

Ara bé, és necessàri realitzar la imputació dels NA de les variables numèriques. En aquest sentit, plantegem, en primer lloc, la pregunta de si la presència de valors missing a la nostra base de dades és, o no, aleatòria. En el cas que tinguem random missing, podem pensar que es deu a un fenòmen aleatòri casual i que tots ells segueixen la mateixa distribució amb valor esperat 0, de manera que coneixent informació adicional no hauríem de tenir problemes per a realitzar la imputació. Tanmateix, la qüestió es torna més complexe en cas que tinguem missings no aleatoris ja respondrien a algun tipus de sistemàtica.

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

27

• 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 nul.la de valors missing aleatòris. Si ens parem a pensar, aquest resultat sembla raonable, ja que les úniques 23 observacions per a les que tenim dades mancants, en relació a les variables numèriques Hotel\_lat, Hotel\_lng, Businesses\_100m, Busi- nesses\_1km, Businesses\_5km, són conseqüència d’una absència de coneixement de la situació ge- ogràfica de l’hotel. Així mateix, és llògic pensar que això es deu en un error en la mesura, o la manca de dades geogràfiques d’una zona particular (totes deuen ser en un espai força proper).

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

dd[, i]

0

Variable − 7

F

Variable − 5

0 0 5 1

0 0 0 2

Variable − 6

5 3

y c n e u q

0 0 0 1

y c n e u q

0 0 5 1 0 0

y c n e u q

0 5 2

e r F

0 0 5

e r F

0 1

e r F

0 5 1

0 0 5

0 5

0

0

0

42 44 46 48 50 52

0 5 10 15

0 50 150 250 0 0 2

Variable − 8

0 0 4

Variable − dd[, 9

i]

dd[, i]

0 5 1 y c n e u q e r

0 0 1

y c n e u q e r F

0 0 3

0 0 2

0 5

0 0 1

0

0

0 2000 4000

0 10000 20000 30000 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 miss- ing a les variables textuals (a la variable de dates no en tenim). En aquest sentit, podem plantejar un procés similar al de les variables qualitatives, creant un “text” que digui “Ressenya no vàlida” per a aquelles observacions amb NAs.

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

28

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 plante- jat en l’apartat d’anàlisi exploratòri inicial, concretant quines variables han patit alguna modifi- cació i quines romanen igual.

Les dues primeres variables id i Hotel\_Name segueixent sent identificadors per a cada obser- vació de la base de dades.

La variable Hotel\_Country no pateix cap tipus de modificació.

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

**Hotel\_City (Processsed)**

F

3000

Amsterdam Barcelona Boulogne Billancourt 2500

London Milan Paris Vienna 2000

Vincennes

y c n e u q e r

1500

1000

500

0

Figure 20: Bar plot Hotel\_City (Processed)

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

A les variables Hotel\_lat, Hotel\_lng, Businesses\_100m, Businesses\_1km, Businesses\_5km hem re- alitzat 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 Ho- tel\_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.

29

41.33 48.21 51.50 49.46 51.52 52.40

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

En aquest sentit, observem com ens movem en el mateix interval de valors i, al no tenir cap ciutat en una localització molt diferent a la resta, podem concloure que la imputació ha estat exitosa al nivell de precisió geogràfica al que treballem. Tot seguit, construïm histogrames per a les variables Businesses\_100m, Businesses\_1km, Businesses\_5km, ja que ara les tenim codificades com a variables numèriques (Figure 21).

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

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

La variable Guest\_Type no pateix cap tipus de modificació.

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

La variable Stay\_Duration no pateix cap tipus de modificació.

La següent variable és Review\_Date. Aquesta variable la hem recodifcat com a data i podem considerar determinar l’interval de temps corresponent a les nostres dades.

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

Observem com aquest interval comprèn 2 anys.

Les variables Is\_Hotel\_Holiday, Is\_Reviewer\_Holiday, Review\_Is\_Positive i Submitted\_from\_Mobile les hem recodificat com a factors però els gràfics construïts anteriorment són perfectament extrap- olables, 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, Re- view\_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.

30

**100m**

0 0 5 1

y c n e u q e r F

0 0 0 1

0 0 5

0

0 50 100 150 200 250

# Negocis

**1km**

0 0 5

0 0 4

y c n e u q e r

0 0 3

F

0 0 2

0 0 1

0

0 1000 2000 3000 4000 5000

# Negocis

**5km**

0 0 8

0 0 6

y c n e u q e r F

# Negocis 0 0 4

0 0 2

0

0 5000 10000 15000 20000 25000 30000

Figure 21: Histogrames Negocis a la rodona (Processed)

31

**Room\_Type\_Level (Processed)**

F

3500

Business Classic Deluxe 3000

Duplex Family Other Standard 2500

Suite

y c n e u q e r

2000

1500

1000

500

0

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

**Trip\_Type (Processed)**

Business trip Couple Family Leisure trip Other Solo traveler

Business trip Couple Family Leisure trip Other Solo traveler

Figure 23: Barplot perfils viatge (Processed)

32

**Reviewer\_Nationality (Processed)**

F

3000

Australia Belgium Canada France Germany Greece 2500

Ireland Israel Italy Kuwait Netherlands New Zealand Other

2000

Saudi Arabia Spain Sweeden Switzerland Turkey y c n e u q e r

United Arab Emirates United Kingdom United States of America 1500

1000

500

0

Figure 24: Bar plot Reviewer\_Nationality (Processed)

33

**Clúster jeràrquic**

En l’àmbit de la ciència de dades, es sol treballar amb bases de dades molt grans que tenen un nom- bre 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 difer- ents 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 pro- dueix 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 d2 ij

= 1 − s

ij

, on:

s

ij

=

C

h=1 p

1

(1 − |x

ih − p

1

x

jh

|/G

h

) + a + α

és el coeficient de similitud de Gower

• p

1

+ (p

2

− d) + p

3

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,

• α 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 excloure del procés de clústering la variable identificadora del reg- istre i les dues variables textuals 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 par- ticions, 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

34

**Dendrograma−WARD**

s e i c n à t s i D

observació Figure 25: Dendograma-mètode de Ward amb distàncies de gower

la variabilitat que aconseguim explicar amb aquest tall extra. En aquest treball, proposem un mè- tode 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.

**Elbow method − Total within−clusters sum of squares**

s e r a u q s f o m u s s r e t

3 1 + e 8

s u l c − n i h t i w

3 1 + e 4

l a t o T

0 0 + e 0

2 4 6 8 10 12 14

Number of clusters K 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 dendograma, 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

35

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 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.

36

**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 1. 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.

0

**Hotel\_Country per cluster (%)**

cluster

**Hotel\_Country per cluster (nre. d'hotels)**

0 . 1

AT

FR

IT ES

GB

NL

0 0 4 1

AT

GB ES

IT

0 0

FR

NL

8 . 0

2 1

0 0 0 6 . 0

m

cluster 1

g r a

%

s l e t o

0 0 8

4 . 0

1 2 3 4 5

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.

1

Deixarem fora l’identificador, les dues variables textuals i Hotel\_Name.

37

h #

0 0 6

2 . 0

0 0 4

0 0 2

0 .

0

1 2 3 4 5

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.

0

**Hotel\_City per cluster (%)**

cluster

**Hotel\_City per cluster (nre. d'hotels)**

0 . 1

Amsterdam

London

Amsterdam

Milan Barcelona

Milan

Barcelona

Paris Boulogne Billancourt

Paris

0 0 5 1

Boulogne Billancourt

Vienna

8 . 0

London

Vincennes

6 . 0 g r a m

s l e t

0 0 0 1

1 2 3 4 5

cluster Figure 28: Profiling variable Hotel\_City

Per últim, podem realitzar un test χ2 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

38 %

o

4 . 0

h #

2 . 0

0 0 5

0 .

0

1 2 3 4 5

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).

55

55.0927511466334 55.0927511466334

40

−10 −5 0 5 10 15 lon

cluster

50

t a l

45

Total number of reviews

4000

8000

12000

16000

35.6241493912721 35.6241493912721

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 caracterís- tiques i és l’ultima (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"

39

5

4

3

2

1

**Plot Plot Means of Businesses\_100m**

**Plot Means of Businesses\_1km**

**Means of Businesses\_5km**

0

0 0 4

0 0 6 2

1 2 3 4 5

1 2 3 4 5

cluster

cluster

cluster 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"

40

0 5 2

0 0 4 2

5 3

1 2 3 4 5

0 0 0 0 m 0 0 1 \_ s e s s e n i s u B a n a j t i

2

2

m k 1 \_ s e s s e n i s u B a n a j t i

0 0 2 2

0

0 0 0 2 m

3

m

1

m k 5 \_ s e s s e n i s u B a n a j t i m

0 0 0 5 1 0 0 8 1

5

0 0 6 1

0 0 0 0 1

0 0 4

[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)

0

**Room\_Type\_Level per cluster (%)**

cluster

**Room\_Type\_Level per cluster (nre. d'hotels)**

0 . 1

Business

Family Classic

Other

0 0 4 1

Business

Family Classic

Other Deluxe

Standard Duplex

Suite

0 0

Deluxe Duplex

Standard

8 .

2 1

Suite 0

a

m

cluster Figure 31: Profiling variable Room\_Type\_Level

Aquest gràfic presenta la limitació que tenim molts valors missing que hem inclòs en la cat- egoria “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 χ2 és significatiu però hem de tenir en compte que els valors centrats poden estar altament influits 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

41

i r o g 6 . 0

e t a c r

0 0 0 1

g r a

%

e p

0 0 8

4 . 0

1 2 3 4 5

s n o i c a t i b a h #

0 0 6

2 . 0

0 0 4

0 0 2

0 .

0

1 2 3 4 5

(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 lleuger- ament els grups. La resta de categories es manté força constant per a tots els conglomerats.

0

**Guest\_Type per cluster (%)**

cluster

**Guest\_Type per cluster (nre. d'hotels)**

0 . 1

Couple Family with older children

0 0 4 1

Couple

Solo traveler Family with older children

Travelers with friends Family with young children Group

0 0

Family with young children

With a pet

8 .

Solo traveler

2 1

Group 0

Travelers with friends With a pet

a

0 0

6 . 0

m

cluster 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 dominànça dels viatges per plaer disminueix en benefici dels viatges per motiu de negoci.

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

42

i r o g e t a

0 1

g r a

%

c r e

0 0 8

4 . 0

1 2 3 4 5

p s t n e i l c #

0 0 6

2 . 0

0 0 4

0 0 2

0 .

0

1 2 3 4 5

**Trip\_Type per cluster (%)**

cluster

**Trip\_Type per cluster (nre. d'hotels)**

cluster

0

m

.

0 Business trip

Business trip Couple

Couple Family

Family Leisure trip Other Solo traveler

Leisure trip Other Solo traveler

1

0 4 1

0 0

8

2

.

1

0

0 a

0

6 .

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.

1

i r o g

0 1 0 g r a

%

e t a c r e

0 0 8

4 . 0

1 2 3 4 5

p s e g t a i v #

0 0 6

2 . 0

0 0 4

0 0 2

0 . 0

0

1 2 3 4 5

**Boxplot of Stay\_Duration vs Cluster**

Dies d'estada

**Means of Stay\_Duration by Cluster**

0 . 3

5

5 . 2

4

a n a j t i

global mean

5 10 15 20

cluster 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"

43

0 . 2

r e t s u l c

3

m a i c n à t s e a d a r

5 . 1

2

u D

0 . 1

5 . 0

0 . 0

1 2 3 4 5

[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. Tanmteix, evidencia com el clúster 5 presenta una forta desviació significativa respecte la tendència general de tots els con- glomerats. 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

44

**Is\_Hotel\_Holiday per cluster (nre. d'hotels)**

cluster

**Is\_Reviewer\_Holiday per cluster (nre. d'hotels)**

No Yes

No Yes

0 0 5 1

0 0 5 1

s e y n e s s e r #

0 0 0 1

s

1 2 3 4 5 cluster 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 repercuteix 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

45

e y n e s s e r #

0 0 0 1

0 0 5

0 0 5

0

0

1 2 3 4 5

**Boxplot of Total\_Number\_of\_Reviews vs Cluster**

Nombre de ressenyes vàlides

**Means of Total\_Number\_of\_Reviews by Cluster**

0 0 0 3

5

0

global mean

0 5000 10000 15000

cluster 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

46

0 s e

5 2

4

d i l à v s e

0 0 0 2

r e t s u l c

3

y n e s s e r e d g i m

0 0 5 1

2

e r b m o N

0 0 0 1

1

0 0 5

0

1 2 3 4 5

**Review\_Is\_Positive (nre. d'hotels)**

cluster

**Means of Review\_Positivity\_Rate by Cluster**

No Yes

0 8

cluster

0 0 5 1

0 6

e global mean

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 escrites 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.

0

t s l e t o h

0 0 0 1

#

a R y t i v i t i

0 4 s o P

0 0 5

0 2

0

0

1 2 3 4 5

1 2 3 4 5

**Reviewer\_Nationality per cluster (%)**

cluster

**Reviewer\_Nationality per cluster (nre. d'hotels)**

0 . 1

Australia

Israel

S

Australia

New Zealand Belgium

Italy

S

Belgium

Other Canada France Germany

Kuwait Netherlands New Zealand

S T U

0 0 5 1

Canada France Germany Greece

Other

U Ireland

Saudi Arabia

U

m

cluster

Saudi Arabia

8 . 0

Spain Sweeden Greece

Switzerland Ireland

Turkey

6 . 0 g r a

Israel

United Arab Emirates Italy

United Kingdom Kuwait

United States of America Netherlands %

s i r a u

0 0 0 1

4 . 0

1 2 3 4 5

Figure 38: Profiling variable Reviewer\_Nationality

De nou, el test mostra una alta significació estadística.

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

47

s u #

2 . 0

0 0 5

0 .

0

1 2 3 4 5

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 textuals, són Review\_Total\_Negative\_Word\_Counts i Re- view\_Total\_Positive\_Word\_Counts. Entenem que aquestes variables han de anar en línia amb el re- sultat 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\_To- tal\_Positive\_Word\_Counts i viceversa.

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

cluster

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

5 3

2 s t n u o C \_ d r o W \_ e v i t a g e N

0 3

1 2 3 4 5

1 2 3 4 5

cluster Les conslusions 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 valo- racions 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 estadís- tica 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:"

48

2

5

s t n u o C \_ d

0 2

5 2

0

r o W \_ e v i t i s o P \_

8 1

\_ l a t

2

l a t

6 1 o T \_ w e i v

5 1

o T \_ w e i v e

4 1 e R a n a j t i m

0 1

R a n a j t i m

2 1

0 1

[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 otor- gant els usuaris de Booking que han escrit les ressenyes. Teòricament aquells hotels on les valo- racions són més positves haurien de rebre millors valoracions així que, sembla llògic pensar que els resultats hauríen d’anar en línia amb els anteriors.

**Plot Means Average\_Score**

cluster

**Plot Means of Reviewer\_Score**

5 5 . 8

0 5 . 8

0 . 9

e r o c S \_ e g a r e v A a n a j t i m

5 4 . 8

0 4 . 8

1 2 3 4 5

1 2 3 4 5

cluster 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 indici d’alta correlació entre aquestes variables. Tots dos testos mantenen la significació global amb valors més propers a la no significació per als clústers 1, 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:"

49

e r o c S \_ r e w e i

5 . 8

5 3 . 8

0 3 . 8

v e R a n a j t i m

0 . 8

5 2 . 8

8 0 2 .

5 . 7

[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 concloure el profiling, és Total\_- Number\_of\_Reviews\_Reviewer\_Has\_Given. Aquesta ens dona informació sobre com d’actiu a Book- ing 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.

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

1

Total de ressenyes escrites pels usuaris

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

5

s

8

global mean

0 50 100 150

cluster Figure 39: Profiling variable Total\_Reviews\_Given

Els contrastos reforcen 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 val- oracions 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).

50

i r a u s u s

4

l e p s e t i r c

6

r e t s u l c

s

3

e s e y n e s s e r e d

4

2

g i m e r b m o N

2

0

1 2 3 4 5

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

Nombre de valoracions adicionals

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

0 0 7 5

s

0 0

global mean

0

0 500 1000 1500 2000 2500

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 (cluster 3 les més negatives i clúster 4 les més positives) és on els usuaris s’entreten més a valorar aspectes extra de l’hotel. En contrapartida, aquells establiments hotelers amb valoracions mitjanes no solen rebre d’adicionals.

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 mostrarà si en algún clúster són més freqüents les ressenyes i valoracions escrites des del mòvil. Aquesta variable pot ser indicativa de si els usuaris estàn 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 mòvil (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:"

51

l a n o

6

4

i c i d a s

0 0 5

r e t s

3

n o i c a r o l

0 0 4 u l c

a v e d g i m

0 0 3

2

e r b m o N

0 0 2

1

0 0 1

1 2 3 4 5

**Submitted\_from\_Mobile (%)**

cluster

**Submitted\_from\_Mobile (nre. d'hotels)**

cluster

0

m

.

No Yes

No Yes

1

0 0 5 1

8 . 0

6 . 0 g r a

%

s i r a u

0 0 0 1

4 . 0

1 2 3 4 5

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

s u #

2 . 0

0 0 5

0 . 0

0

1 2 3 4 5

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 tendèndia 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.

52

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 freqüentats 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).

53

**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 reduïnt 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 (nxd) en un subespai més petit, però que mantingui una bona representació de les dades i les descrigui correc- tament. 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 redundàn- cies.

• Identificar correlacions entre variables.

L’ACP aconsegueix aquests objectius transformant les variables inicials en nou (i més petit) conjunt de variables sense que perdem la informació més rellevant que ens aporten aquestes. Les noves variables les anomenem components principals, i no són més que combinacions lineals de les variables originals. La metodologia assumeix que les direccions principals amb major variàn- cia són les més importants. Per exemple, el primer component principal és una combinació lineal de les variables originals que captura la variància màxima de la base dades, determinant la di- recció 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 coorde- nades sobre els que rotaran les dades. La idea és aplicar aquest principi sobre les nostres dades i anar analizant les components en plans bidimensionals.

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

54

En aquest sentit, és important mencionar des del començament que, com que hem de calcu- lar 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 combina- cions lineals. En terminologia PCA, tenim:

• Individus actius: Totes aquelles observacions que intervenen en l’ACP. En el nostre cas totes les observacions (5000).

• Individus suplementaris: Les coordenades d’aquests individus s’estimaran fent servir la informació obtinguda de l’anàlisi de components principals en base als individus actius.

• Variables actives: Totes les variables que utilitzem en l’ACP. En el nostre cas totes les numèriques i numèriques enteres.

• Variables suplementaries: Com en el cas dels individus suplementaris, les coordenades d’aquestes variables s’estimaran amb els resultats de l’anàlisi. En aquest cas, no hem espci- ficat 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 captura cada dimensió (component principal).

0

Scree plot

s e c n a i r a v d e n i

15

a l p x e f o e g a t n

10

e c r e P

5

1 2 3 4 5 6 7 8 9 10 Dimensions 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 variàn- cia 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

55

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 cos 2 indicaran baixa representativitat, poca correlació dels valors de l’individu/variable amb la component principal (poca rellevància en l’anàlisi).

Gràfics d’individus

Per a començar l’anàlisi podem representar, en primer lloc, el nùvol 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 construim 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).

56

) % 5 . 3 1 ( 2 m i D

Individuals − PCA 7.5

−4 0 4

Dim1 (18.4%) Figure 44: Representació dels individus sobre els PC (1er pla)

57

2647

2219

3758

102

541

5.0

1753

2769

103

3759

542

2220

2648

2221

104

543

2210

3149

1209

3417

3700

2814

618

3701 2.5

3150

3151 0.0

1210 2816

2770

2815 1754

3418

1211

2817

2211

3702 −2.5

4435

3971

3840

3505

2592 3798

2978 2053

4498 388

1408

2094 4972 1040

2709 3732 4099

2926 1894

4205

4596

1056 3813 4721 4524

1460 3932 3111 2187

3898 1554

112 4260 2107 989

2673 4740 1368 2426 2227 4386

1843 472

647

3864 582 440 4603 2440 2676 365

3070 258 528 4338

3895 4763 4039

4845 3642

2081 1152 1882 1741 4644

2275 3456 3944 2888

3033 1486 1391 2153 974 1875

4852

2126 2762

1326 1335

1950

2580 4368 935 4922

297 4171 3858 3761

786 1640

372 80 4670 1137 2418 3067 2829

330 4361

2535 3995 1655

1134

3403 3512 2933

3966 4921 740

3973 4322 3901 4147 2946

4281

3752

2110 414

99

2292 4902 2843

533 228 242

2590

720 1165 227 4844 3906 1268 1836 3578 1835 4724 2253 1990

1939 4380

3229 1232

2895 2923

1715 3955 3623

4538 2242 2403

1846 4374 4926 127

3314 652 4874 804

2595 335 1009 857

2097

2586 4995 497

3458 3390 1181

480 765 680

3897 4328

4084 34 2177 662 600 4552 580

2113 1033 293 575 1052 3555 963 4160 1371

2070 4294 926 3488 4854

4468 952

3292 4678 4931

1184 4044 742 4545

3976 3011 361 3936 4043

4893 1685 2917

2957 1466 4637 2186 2017 4760 4686 23142315 790 3711

3036 1269 1537 3814 3393

1686

270 4267

1267 2672 4535

907 3617 171 1555

4774 3446 702

2959 3180 1239 957 998 2277 2289 2791

822 1070 2607 435 3030

4134 49764977 1045 2932 3320

2925 322

1313

2195 4947 701 4269

4568 3938 4796 1831 73

353

549 2192 3017 2165 1221 764

4491 4017

3558 2664

4682 4620 1660

4144 4505 3214

2830 469 3091 1208

4624 1201 38

411 3206 2308 1717 109 1110

2691 1698 3063

4063 1871

4276 1878

4667 3593 3124 2252 2636 453 1364

2822 2556 4849 2296 3401 1415 2766 2077 4315

2417 4174

1912 694 2263

4199 4558 4930

375 1901

4285 3416 213 3920 1744 573 4778 3438 2516 1036 3377

348

4924 2998

2149 2649

1261 2026

1958

117 1973

2169 3369 782

4275 4226 3780

4705 68 3174 1661

728 4757

2696 1418

3907 1501 1225 1705 2860 4699 2702 2956 4307 1953 2390

4030

1704 4602 3889

4255 4019 1761

2074 4397 3981 813

2164 1787 784 4196 1339

523 4484 2127 2996 3850 2501 3172 4280

1430

3261

4784

4495

3534 4345

2739

1079 313 705

3675 2824

2112 4726

4067 4712 4800 676 315

1638 2835

2072 2526

416 3626 1245 532 2233

4897 3484 3856 4329 1348 2788 1988 3974 2181 1096 433 4451 4317 4163

3718

1877 3664

3634 2411 2368 840

1713 1586 2351 1726 291

1292 3360 2058 2745

2686 4033 451

4621 1468

4767

1858

139 4041 406 3249 3533 845 1582

3286

4890

4170 1442 917 2732 1156

3677 882 289 1818 1324

1480

769 2317

455

2269 3051

1285

767 2908 620 3844

460 4365 2424 2549

1327

1101

2205 3948

2475 2863 3684

3886 3670 62 1773 3949

3474 3862

1606 3540 1983

1757 2789 2088 928 4089 4394

2986 4185 4050 2111

2893 4387 2717

4056 1043 1898 4817 3284 4441

3146 683

3343

441 1915 41534154

3477 2751

4650 4543 992 1671 4872

1332 2478 1139

1428 3205 3986 746 4898 2085

389 356 1760 3027 157 4913 2320 3090 2197

832 2616

424 1464 3476 3705 4366 1833 92 3630 985 2356 4282 77 1065 4923 2301 1226 3405 2506 2389 3374

2826

2176 778

4341

657 937 1222 4804 910 4405 3188

329 3704

3571 704 1233

2741 4641 4128

2284 1014

641 1359 108 1432 2225 4313 199

3918 2188 1771

1575 3005

1149 4169 892 3510 4437

2232 1369 3054 636

2642 4349 621

177 961

495 1271

4963

4335 458 3463

4419 1231

4660 1273 3788

4581

4013 3803 1076 4189

1545 3778 3964 1814 3034

2889 2021 4103 996 1330 4577 4851 3092

2488 875

4555 2020 248 1910 3809 2079

1017 4218 544

465 3724 145

323 1080 8 965 2729 1090 4133 3059 630 3098

3530

519 1706 896 1776 1727 2116

4048 4658 4807 1652 2037 1855 1038 2437 2481

696 4184

4346 3327 1645

224 4449 1320 4692 4443 4083 2905 666

428 3835 967 3291 978 3246 6

1410

561 717 4635 4248 2309 915 4015

1476

1005 3498 4711 1722

1270

1011

2573 794 3604

3303 4012

2385

1772 604

3747 316 2422 174

526

4689 1784

239 3461

665 1204 3263 3159 4206 1816

4836 339 355 909 3086

2378 4343 3339 1670 4502 18

4298 1197

2597 4988 4268 409

3608

663 2366 279 370 3441

3348 3407 4858

1929 4130 3409 2445 22992300

3404 2316 673

162 3148

1405 4747 509 4722

4424 214 4254

3919 4320 2446

3762 2641 4903 4040 812 3783 1598 2621 4986

4861 671

4113

404 4966 4352 2651 2212 3868 4038 486 3525

1461

1803 4736 4290 2706 463 2631 1031 4554 2640

3853

237 2539 2001 2331

2287 1250

1478 4251 2845 1441 2090 4162

1081 3532

1830 3880 3849

4653

861 4506 3105 4080 2502 4490 1344

1955 3473 2339 96 466 4398 3244 2206 3297 1812 1157 229 2031 246 4783 3083

2015 4257

4220 1566 3196 2941 259

230 4396 3504 4871 1975 744 3216

2524

818 3460

3487 445

36 4756 1279 3351 2873 635 4733 2804

2381 3079

3495

1032 3454 4157 3792

24

2547

3309 4970

2716

1919 689 1235 484

545

4357 1448

222 184

263 4167 3176 3014 2147 4036 4211 3950 2167 412

1969 4066

700

1416 3931

3352

4363 2774

3212 605 2497

1388 3629

3045

4634 1735

1807

3720 734

4469 4300 2397 4725 713 4702 898 4101 4788 2118 1523

2882 2451 1321

4250 2087 1956 3410 1509

1297 1980 4436 2698

2180 3024 1845

1543 4488 3828

2453 447 1584 87

574 2288

4253 1020 4331 311 1844 4727 4879 955 1290 2878 89

1475

3587 1889 1026

2141 3285 3696

4839 634

2827 3160 2671 1404 3713

793 675 4178 2625

2434 3293 3887 4207 3230

4330 2185

3271 3715 4110 25 3192

1765 3 2290 4074 2633 1342

3765 207 3655 4090 2306 2452

608 2223 3064 4791

3008

3782 478 1601 4088

160 2558 2594 1926 1203 3822

4519 1337

749 3613 49

3050 670 3265 940 3233 2461 2198 1218 2062 1176 1500

4264 1317 3859 1129

2350 45694570 4123

3344

4466

2009 4598 2235 3697

4869 3637 2404 4576 592 1617 4004 2687 2929

2945

1826 2973 4422 4856 4615 2214 427

2682

2752 3706 1155 3541 1556 619 3225

3288 4508 4417 2942 4855 2796 4842

868

2091 3560

4482 3308 4707 1223

869 23542355 4527 11771178 4072

2868 595 3722

11671168 3480 3528

2332 2226 2619

4310 2060 3273 1002 4486 3992 3324 4738 2265 165

3638

1801 2384 2968 3099 4109 1793 3653 1691 839

4693 887 3672 3871

1206 512 477 816 3020

4873

16 238 3022 3321

1097 454

3620 1420 3917 518 178 3969 1436 1171

1790 1349 2545 1266

1198 1576

1531

1341 2132 83 3594 571

3916 4640 231 2043 28382839 4688 2911 491 351 3445

2084 3221 4782

1672 319 1160 189 1618 3001 1529 3280 1989

1728 4578 1015 4886

1792 397 2435 4129 2179 4351 2735

2694 1259 4046 449

4828 4256 588 4948 1911 1100

3385 4515

3940 1641 1128 2553 2083 872

835 2432

3867 2125 4350 110

3341 711 3106 3019 1829 4270 3281 394

3915 2465 4378 3648 3237 1276 4531 1191

2727 2173 4107 3744 2023 2421 4798

1251 3057 4542 3128 2458 3432 2281 4811 4095

1518 2869 1116 625 129 4195 3041 569

1971 1538 1435 4321

2204

860 2448 4117 2712 1220

3431

3791 4112

3247 1679

1587 1255 4590 1024 430 4166

736

2582

3688 1142 2433 2213 626 4021 4016 4945

3043 2264

1257 1729 540

4407

2326 3436 785

3529 4677

4429

3582 4115 3526 2160 3186

825

1469 4121 2635 4480 4029

45

320 3073 3647 3012 754 4216 2529 1075 2374 468 1967 2951

1646 4312

4769 1852

4291 1470 609 1553 3547 2884 1446 66 4622 1346

2610

159 730731 2508 3025

2515 616 1536 741 1934 3993

2131 42014202 627 3119 3490 3452

2990 4795 303

2375 326 1884 3725 715 3579 2352

1720 215

763 3746 1588 420

2130 3426

2172 4586 1560

3296 3439 1493 2720

930 4217 1692

200 2069 1053 629 697 2877 1307 1312

138 2850 3816

2194 4393 3667 2588 1936 2479 1308 1665 496 3842 3122 2379

4301 1840 3557 2562

4204 190 1353 3312

4065 2910 4483 3197 4964 576 3464 4245

2014

4881

4786

3358 3493 337 4891

1063 181 479

2875 3690 824 4000 554

1386 269 3596 2027

977 4454 4573 4344 4797 2532 1298 444

1653 2662

1750 1370 4541 2108 151 4518 2617 4509 2773 4120 4100 195

2555 2067

4571 4297 49174918 4920 4055 4950 2019 3836 844

2474

1632 1540 1985 1697 1805 1406 198 3433 4671 236 4708 1484 2216 2358 1740

4091 2778

3330 594

3527 1025 128

3262 923 815 4042 4723 2581 3326 3719

4214

3394 4941 562 1795 2035

2914 4382 28662867 4124 1886 1719 1278 3298 292 1505 2401

1925 2154

654 1099 1590 3795 3047 4539 4031 1533

945 4776

1966 687

2520 2915 2129 1316 402

686

1711 1196 3738 3937

1280 4463 1502 3926 1403 4442 1633 1462

481 1817 4145

4210 3035 3453 2362 2245 3158 2093 1028 3266 4140

3240

3537 2008 1105

4814

4989 3016 745

4496

2271 807 2460 3590 4007 115 3539 3479 2431 2487

3328 546 439 2668

2322 106 3733

601 347 265

4714 4983 1230

4452

3945

1421 1283 1490

1329

4741 781 3564

814

3923 219 3210 4829 1786 953 2202 3440 2046 2189

1589 4456 11231124 4831 4183 1109 1192 725 39463947 2802 1702 3565 4647 1173 1135 1114

2183 4266

4969 4574

515 1492 4440 733 1861 4181 1010 3797 2050 88 858 3260 1457 1643 3175 2819 79 3000 1237 1121 4808

41384139 3652

1984 3998 2979 3881 3522 1948 3710 3353

4172

284 2036

890 63 3553

105

4940 226 3378 3902 2921 1528 615

3818 1515 4892

3679 3741 4504

452 4984 4974 2250 363 2057 3238 3839 2402 4434

4752 85 2105 1539 1272 1635 4252 3888

3384 1673

415 1994 2969 2891 1477 3257 4137

4190 3419

2798 1597

3299 462 3619 3928

2480 2099

1252 2000

2984

1498 4718 4645 357

2764 2473

4481 2754 2441 1385 2387 2244 4830

3598 3147 4146 4233 223 762 2701 1402

1450 2786 4731 3347 2230

4446 962 2967 3748 852 2303 3187 1182 2175 314 3800 722 4585 4691 4423

4061 4908 2080 23 4401 4453 3774 3635

1991

4114 1086 70 856

2049 396

1413 5 3114 333334 1340 1188

1027 4925 86 3694 502 1767

4032

2781 2953 1409

3776

3215

3595 1696 2981 716 4418 2660 1284 1938 4575 2209 3396 1857 3766 850 2665

4135 1578

1001

1356 572

245 3588 4564

2954

1115 4662 3773 2425 2162 2565

2333 2353 4241 1243 718 2388 2419 1636 1064 1570 2161 3640 4823 4059 4697 4759 1732 1357 4990 187

1147 712 3865 2450 4528 1472 1039

3269 885 2444 4583 1049 3039 4263 4077

133 1904

4122 504 3457 374 2159 3687 3357 1658 2618 2670

3790 4305 3201 4956 386 1997 801 2228 3356 4385

2171 1733 197 2683 2844 4550 122 3678 3144 4002 1850

3767 2138 475 735 2898 726 3817 39293930 2561 1599 3519

2484 944 2100 1042 2509 1254 4561 4011 299

4676 1779 3402 4324

3305 134 25702571 2609

3693 1832 1066 3395 2260 4556

1016

2006 2089 3970 2298 3042

929 4127

3576 4838 2991 4223 3999 1922 3799 1202 4919 4152 467 4993 367 623

1624 4197 4809 2872 4944 1838 3965 373 4526 380 4546

2987 901 2948 4141 1094 2731 1609 505 4230 3346 3481

880 559

3029

2536 612 19 811 4961 4700 4592 4764

3520

4416 3581

4379 2557 2482 3847 1228 4600

3287 3069

1721 3574 2808 2135 954 2294 983 2150 2466 1881 4464 4982 3345 780 1644 1141 3843

2799 2410 1397 3671

867 1433

964 4097 2469 3491

2040

4734 4710 1880 4362 3193 3563

4939 3535 3764

4900

2551 1676 3469 4485 1823 167 4771

1585

1300 1205 1908 2412

624 3414

1579

4832 1169 2380 1240 3325 622 1331 2985

4497

3170 864 3575 936

3387 1111 186

3211 2937 4069 4187 1806 2623

1906 52 3837 81 4111 3097

4503 603 1842

779 755 4889 4835 3413 4027

210 2148

3562 4410 2567 2377

3408 2408

4119 4070

4824 1630 564 4878 760 4229 4572 4105 1104 485

1146 1120 4687 2710 2583 301 1534 1688 706 149 11 2569 295 4632 2086 4530 350

3213 3760

1227 2286 3591 4656 2051 2039

1054 2274 4062 2901 488 598 4915 4580 381 3366 1748 788 1392

4610

2340 1158 2028 4934 4445

1924 2498

3190 4337

359360 4265 1482 1487

4616 991 1508 2746

4743 275

1449 1593

2861 2585

4336

2722

2522 1916 2793 3516

4754 362

1443 2768 4408 1791 2396 1625 681

3167 1782

4787 4690

4052 1229

2348 419 298 1377

4286 3668 4910 2239 3109 3368

1452 2139 393

4522 3483 1463 4935 3807

1190 871 2128 2310 4395 3908

3669

2892 3631 3911 1532 848 422 1048 3552 4627 3676 3586

3486 980 173 1296 1863 4746 795 2327 1897 3424

1977 327 2639

4642 461 1059 1707 2096 1920

2420 4663

797 1108 2262 3489 166 566 4098 2634

407 866

2614

3685 3117 986 401 2504 4661 3939 4014 2742 3278 2243 1022 2939 3471 3707 3183 3354 1940 1804 4748

877

3392 4272 3422 4425 4753 3499 1107 4646

1862 423 1780 3983 4822 830 2190 216 1249 2468 4477 3921

4271

2436

999 3222 1749 513 1360 4092 710 1677

969 1291 4247 4215 3804 931 286 1928 1389

232

58 4367

971 819

2749 4475 2715 4880 3002 2927 2439

3232

994 1896 3245 4794 3126 2500

904 2367 2115

1314 4960

4236

3485 2960

4799 520 2851

172 3740 3040 1558 4302

3554 2047 2767 1131 4529 1547 2842 1567 960 4319 1837 2486 3031 903

1431 3075 3449 2980 651 1473 3848 3430 4709 3709

4364 60

2792

3615

4058 3545 3823 3913 3492 3751 3730 2760 613 552 4073

2831 958 1869 2427 2222 3639 2400 3383 4399 4106 1067 2218 1992 3644 3130

2755 1521 1682 1777 4565 1013 257

3335 2832 4400 349 4413 1294 682 3138

1062 3682 107 4359

3203 3466 4238 2003 1752 776 2790 3322

2266 1649 993 921 4309

2970 3658 4224

900 3234 1412 2011 2459

4428 777

2966

1281 2207 2061 4155

3777 1891 1008 4716 2063

959 4284

883 1398 2542 1180 792 2577 2 2495 942

4951 2644 2168 2144

1546 1215 2579 4411 204

3721 1183 4516 3683 354 4866 753 429 4406

1873

1960 906

4194 29 2285

249 4649

4086 300 1944

1927 4018

2627

2772 4981 3892 201

2231 61 4334 1189

4186 4579

1766 2371 2638 3820 4126 2600 3524 3478 1802 421 3227 3632 3013 2523 2916 277 4457 3769 3373 3943 2975 661

2703 4998 2229 3152

3779 3768 1058 3665 1253

796 2955 664 3503

482 2123 1872 3866

65 1145 2156 4068 660

118 3735 2606 2319 2928 2347 4659 1968 2707

413 346

3656

633 4857 3182 645 3549 3121 3451

558 1738

4907 1163 3726 2200 1769 1293 3329 2659 3224 1185 4478 2675 4674

2134

1440 37 94 3482 4643 3423 2559 4639 2537 1361 1678 4698 1241

98 4587

4450 3391

358 1902 3896 3592 1684

3135 4358 808 22582259 1467 3447

3435 4005 538 4813 2307

378 1651 194 2823 4008 709

2208 878 3082 3162

3538 2382 3912

638 2282 3231 2848 4513

2728

4232 2414 3421 2988 2906 2879

1375 4544 3979 3350 225 4860 1295 3602 3890

1611 4510 3301

4311 1879 4906 2919 3811 3662 1758 49674968 1367

2248 271

1637

2965 143

4611 1978 4173 3023 1976 2697 4548 2795 1569 3264 2514 3235 20642065 2837 193 2013 838 3636 3101 3465

1827 748 3536 1825 44324433 3157 2961 4991 2854 3112 4549 648 873 4750 4512 19131914 2145 1061 2238 522 2554 4638

1113 2876 4306

3548 3078 1874 156 3861 3984 3107 585 771 3692 1366

3885 3143 390 3787 1739

4604 3204 310 4177 1374

264 1774 4744 276 3198 1495 4999

3388 2736 3962 774 132

147 459

2343 3815 4975

884

3058 131 4045 2657 1119 503 1238 3927 3048

4001 1414 2178 2834 1856 274 3178

2994

1809 1604 3985 1499 4259 4026

1854 4501 579 4720

4487 39

3219 2881 101 3933 4766 3007 3411

4904

2713 3987 3878 1289 1451 1724 32 3267 1799 837

1242

4222 3585

4629 3827

2462 4812 1841 3763

1770

3185 3338 4159 1848 4373 4914

3543 4273 1893 1680 649 787 3884

2311 4028

2098 2626 1930 646

3429 3806 4149 120 3935 614 2999 1986

1504 3365 766 984 3120

141 1668 578

2552 1700

3077 2493 100 1909

708 2038

3502 2336

1311 1834 876 2151 3323 2587 770 2971 4840 4859 446 2995 3805

3333 2779

91

2152 43254326 2761 1631 826

2883

123 3279 4240 119 4278 4079 3621 2661 950 4597 4093

3169 4198 67 4231 3362

188 521

2056 1859 2629

2608

15 3708

739 2416 2699 1794 1788 3627 4514 4739 3869 3277 179 4980 3349 2158 2681 3389 48 3334 4768 2124 1917

2821

3625 2800 2847 547 3681 2806

3597 1098 2776

2918 4742

4758

1887

1138 1093 2054 3131 2949

1608 1034 3607 3601 2757 758 4348 2825

4821 1288 1132

1301 3300 3295

1725 3857 2952

1117

2103 3914 3217

568 2840 90 3307 2010 1851 114 3742 2415 1485

2859 889 3074

3037 3163 2174 2398 4296 1615

196 2082 4958

3319 1264

1742 4426 4191

3137 4816 1921 2812 3793 3359

3544

1091 2688 4946 3475 3834 3132 72 1444

3191 4116

3123

4607 1179 9 251 450

4932 2605

4051 1797 1000 894 3894 1199 534 1933 695 4494 282 40 253

2267

1695 3650

2550

4672

2903 268

1345 4094 1506 3317

508 1055 2184 1148 650 567 3006 54

2602 1030 218

2803

113 990 371 4511 1212 1522 3891 1963

2345 2809

1172 2041 4654 4388 3641 1951 3556 1778

4391 2325

4540 2643 3494

1084 809

116 281 71

1074 3380 3115 2612 2477 2476 4666 136

3468 2546 3282 76 1849 3961

1564

4415 631 1785 4132 3925

148 2894

4820 946 4996 4770 43 536 4459 2615 1572 3021 3727 3236 925 1494

1380

3673 4772 2963 2510

2777 1417 679 2849 3942 3363

506 1731 587

1483 1354 2748 847 2611 1819 2693 3415 1642 4593

3580 548

294 2030 806 4785

1519 4243 2137 642 3496

1150 1683 1959 1592 1868 1006 2853 2022 2528

934 2887 4246 3990

2763

2133 4020 247

4287 4370 4717 4793 1358 3975 4473 2756 2201 1073

3444 4258

1174 2563 820

2337 1864 1323 3165 4952 287 4347 4803 2004 1175 331 3829 2203 2215

1166 724 1214 4745 2738 2143

997 27232724 1164

737 2771 1867 1957 2363 3603 4655 3513 4375

1106 1937 699 3851 4179 1217 851 240 4076 2924

2052 2291 22 4049

4156

2989

476 158 2359

2820 639 1479

3136 37363737 3068 2593 318

4959 4356

69 3255 1946 2780

1459 4293

2940

1306

254

2678 244 308

2747

278 557 2782

4355

2653 1603

4148 2575 4867 3523

3646 340 2655 4006 438 1808

4884

208

2902 2846 4599

399

3428 290 2018 4887 943 3566

342 2540 4078 3202 1559 2690 1325 4234

810 2430

3437 1046 2395 836 4617

1895 437 250 4703 3226 2170 3018 4087

2818 31533154

1890 3501 3371 443 3952 855 3096 410

684 1372 3164 4462 306 4500 4997 672 3714

3497 895

2101 1513

2622 3134 1514 4979 2024

821

751 2836 4262 550

4681 3102 1151 4203 4339 2525 4876 4142 2120 4805 4458 3274 1310 2342 3924

3145 2852 4182 2256 3108 902

3716 3997 4237 2076

2283 1445 2272 2454

2572 1510 4612

1876

1648 2730 391 2247 859 3991 1775

823

4665 4765 2305 3624 4636 205 2318

4261 1102 1153 2700 4292

913 4862

593 169 30 3583 4521 4158

1216 4228 3517

2584 4212 2246 535 3600 1947 309 418 1982

2794 886 2075 951

4213

2871

3729 4525 2007

1699 2598 1544 1438 4402 3283 4749 384

1574 1087 4943 2496

4353 2797 4761 385

93 3470 1664 2725 916 3951 833 2758

3900 693 2765

1060

4161 2005

2313 3854 4762

596 3755 344 1659 1454 2654 2667 1798 2885

21 2066 4551

3386 4409 4806 4054 4929 4414 3785

2857 2913 1600 1943

3978 1395

1594 2078

2257

1103

3577 3139 2604 1407

3749 759 2360 4905 1343 1552

677 1714 471

3812 891 3304 4949 44 560

1605 586

2870 241

3500

2862 152 1071

3228 2104 4472

432 644 3254

995 3093

2566 775 3584 1429 2055

2304 212

4460

4626 570 1447

3877 3689

302 2092 2544 653 4438

3833 1136 2122

4560 1144

2449 55 1607 1035

1304 3166 1143 3382

2947 1993 3208 4439

2630

3628 2338

4023 510

3242 2373

3611 4594 3756 3775 1892

1213

4953 1542 1639 2958 4047 3968 3010

4249

1866

3572 2759 1821

1813 1763

4605

970

1737

3253 155 843 2890

3771 1820

3988

2365 3618 537

3306 881 368

3038 1287

3832 2983 3616 3173

24052406 3567 176

2386 4719 4360 1303

2241 203

3376 2136 4035 1516 1362

2489 59

1050 3996 4022 260

41504151 4493 1525 1194 3066

2157 517

3367 4633 4729 3379 191 4588 2897 2708

3189 4683 3161

1037 2068

304 4489

514 4447

870 1047 1305 829 4037 732 656 1654 3794

2499 1972 768

1122 2012 2095 4404 124 1581

3826 4619 911 3004 2726 1507 493 1092

939

4965 1954

678 2620 1811 4606 4790 1275

4304

956 1263 26

2669 2865 2394 1822 3910

1095

3838 1853 3569

3723 589

2864

3956 1595 487 2637 2251

1596 4780 1422

1379

1161 142 3967 4474

4102

4471 352

1619 2472 667 4342

4277 283 4461 1474 1657

2357 3610 637 4928 3904 2182 3129 2578 668 4695 2470 3256 2775

632 4327 4024 2810 4371 4164 789

4175 4060 426 3181 4455

1583 692 13 2935 2750

1899 474 3272 4773 553 2695 3819 3863 772 3168

3605

4520

3743 1497

583 2992 2391 150 4499

4010 2685 3400 4675 4901

841 2140 1888 3551 2530 1286 3514 2442

4657 2733 1437 3609 4899 3071

206 752 1488 827 4421

933 3177 2413 1622

966 1620

395 4751 1689 3903

927 2297 2568 3808

4534 267 4870 2527 3258 3570 1550 1258 1931 1455 2858 1935 2787 799 4631 2034 1078 3959

387 516

2904 3294 95 1419

1439 1577 908 919 121

1768 1551 4883

1565

1623 28 473

1207 22542255

3318

1666 3142 3442

1885 1730 1974 4648 2589 3960 2392 436 727 1186 262 3568 3789 2677 3195

3251 4737 1302

4912 1125

3899 2236

2324

987 4308 140 4384

1662 4403 403 4372

738 2534 2155 2457 1426

1282 4563 1424 4034

4476

1573 1527

3954

1130 4582 4685

757

1274 4180 1260 976

905 345 3375 3802 4668 3589 1351

366 1425 2833 483 2734 3989 3110 1248 161

1041

607 1378 3315 4732 4289 1634

175 1355

3825 2428

1427 3663

3003 4819 3381 565 1723

2543

2273

3087 2538 3905 3127 3336 4992

4825 2974

1800 918

4895 4219 539

1693 3420 3561 4430 3052

2721 4383 364 4985 2121

3531 610

4283 685 3893 3654

2718

4479 3116 941

707 1072

2109 3103 4227

3081 2029 4595

1401 1453

51 1561 4755 1018

1077 1399 4096

3757 3275

2443 4566

1383 3612

723 2370 1322 4882 2576 1610 3699

3259 2217 4057 1394 2646

2117 4340

4868

4853

1900 1942

2302 3521

4239 3506 3546 4987 4299

3179 4517 3550 1687 2447

3055

4536 2503 2048 3171 3425

3542 3450

2312 2880 2531 1883 1762

747

2663 4630

4244 17

3061

2376

1247 153

4792

3745 1159 3772

1964

164

4833

1481 3957 252

3882

2456

4728

831 2507 2455 3511 4735 719

4618 531 2920 3754 3810 3876 4448

2993 4589

3364 2719 332 2492 4978

4651 3874 690

4875

234 4909

4843 1663

3781 1591 3241 3872 4323 4242

192

376 2711 312 1824 2106

888

4136 2119

1423 1112

3659 2624 3830 4314 2323 4025 2714 2249 1669 1847

2807 863 2964

2740

1021

1562 1265 3397 4303

3786 2813 1903 3614 4802

2943 846 1517

1315 529530 3220

255 4864

2279 2464 3870 4235

3982

3248 130 4834 800 1381

2934

1088 125

1012 1580 2044 2485 1083

3831 4916 2828 4553 2280

714 1905 2705

2025 431 4954 2982

1815 3080 2163 492 3252 879 802 2512 2490

2591 1365 2505 3331 47 3717 1262 4075 1471 405

4613

1756 674

4818 4064 1434

235 3342 854

221

3801 3643

4492

14

170 4420 4937 4684

1628 1690

698 3032 2329 2896 2656

1694 1530

4609

4936

4994 2278

448

4333 1051 3209 1019 4125 597

57 288 2045

3113 2599 4777 1718

1945

3268

3412 581 3270

3909 1118

12

972 1396

1376 527 4837 2753 377

2059 3994 3599

1044 4427

2548

3645 3852 3100 75 1999

1068

551 3606

2407 1568

4971 922

1200

3980 1352

2650 3728

3958 949 4332

20 3633

4704 2936 3841 2483 2628

4715

1627 135 947

2224 3883 4775 1318 1126 3977 3372 1140 3796 1839 2102

1319

3125 31 2560 4628

750 4706

1489 979 1069 27432744 3155

1336

1961 628 3784

1350 555 4815

3370 4377 3104

3140 1085 3739 10

3223

1347

4143 2349

3009 343 392 1746 499

2977 1007

3313 968 220 773

1675 111

2658 272 4003

4847

3622 2912

4431 3824

4389 1736 2674

4894 4846

791 2931 3049 53 1029 3332

1981 4221 4131 4933 1747 1548

328 211

2146 3302 498 4810

2652 2533 1616

4523 3459 3922

3770 1526

2680 382

2423

1557 3141 470

874 2704

4942 4444 4557 2944

163

3199 2166

4465 1865 3089 4390 4694

1023 2409

434 74 3427

1907 1932

1004 4108 305 1709

4009 4841 3559 400

4081 1962

1996

2632

50 3398 2372

1503

501 577 3953

408

4188 154 3399

1755

3518 2276

2364 4591

2361 3467 2344 4877

4274 64

4559 2596

2467 2471 4200

341 4826 1629

4392 2907 4288 3731 1373

2938

1987 2295

325 3855

3972 805 1219 4888 3076 1458

2689 2335 379 3666 5000 1870 1949

3455 2909 3873 500

464 563 3846 3355

2841 2234 2645 3660

507 1995

4547 144 7 3448

4863 2805 525

2341

2521 3661

2196

4053

3934 862

3845 4652 3316 3053 1382

4376 4789 4192

3712 261

1781 490 2692

938 4 456 1563

893

324 4369 3046 1170

2684 2603

1923 2033 2191 3085

4885 3472 1681 217 669 3026 4567

3060 35

1621 1246

988

3028 3695

643 369 1309

3361 1154 4955

3340

591 4957 1751

3691

3508

1393 524 982

1277

2328 1828

853

3406

2737 1082 2268

3184 2900

1759 2541 897 617

1535 1193

4608 4176 1602

1626

842 817

1057

4467 4713

3734 1256 4209 611

1703 721 457 2511 599 4801

442 1512

2399 3434

4354 4848

2334

1743 296

2016

1998 1613 4973 1524

2193 4679 168

2261 3084 1390 3290

3243 4701 1712 233

2042 3276

920 27

743 4533 1979

4911 1571 1796 975

398 4071 2429

4601 2002

209

41 1941 2142 4625 425 729 2199 4225 2922 1965 1541 1549

4208

56 3753

4614 1162

3310 1612 3750 3156

1334 3218

3686

4669 3649 606

688 3963 307

3680 2811 2369

2972 3879 1710

1745 1224 1187 803 273

828

2270 912

1328 2950

3875 2513 1195

1708 590

1491

1465 1614

3698

78

3250 2613 924 4827

494 46 3094

2976 1716 2601 2393 1650 2783

3860 3044 3289 4532 1384 4562 3056

3821 4507

1400 3239 1127

3200

4470

417 82

2114

2517 1299

4865 3573

783 3657

1701

1387 3207 3941

640 914

4781

3194

1520 182

2886 2073 321 4664 3062

4779

756 3515

2491

4938 489 338 4085

602 1783

4168

126

1363 849 3133 3337 2801

2032 2899

3462 280 42 1734

3088

1952 2997

336 3065

180

584

4584 1236 1764

3118 2383

4962 659

556 2071 3509

2237

1511 798 1656 137

266

4295 4680 658

703 4279

3443 761

1918 4412 4850 1674 4896

4118 2240 185

4082 3072 4316

1411 256 1234 932

4730 146

3674 3507 3311

1860 511 183 243 1003 1333

2930 2784

2666 84

1496 4696 1133 981

2785

4381

4193 4673 655 2321 691

899 383 4318 202

1456 2494 3095

2463 97 948 2574

1338

1244

2330

1810 33 4927 3015

4537 834

317 4104

3651

2438 2874 1

2564 285

2519 4165 3703

2962 2346

1789 2679 1089 865 1647 973 2855 1667

2293 4623

2856 1970

2518

cos2

0.25

0.50

0.75

cos2

0.75

0.50

0.25

2

Dimensio\_3

5

Dimensio\_5

0

0

4 \_ o i s n

6

−8 −6 −4 −2 0 2 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 captura menys variància). Adicionalment podem testar si aquests nous eixos construits 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 estandaritzada). 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 circunferència 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 com- ponents 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. Construim 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

58

\_ o i s n

2 −

e m i D

e m i D

5 −

4 −

0 1 −

−4 −2 0 2 4

Individuals − PCA 7.5

5.0

−4 0 4

Dim1 (18.4%) 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 llogic 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, 5i6 (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 pri- ori, 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 exempleReview\_Positivity\_Rate, Reviewer\_Scorei Review\_Total\_Positive\_Word\_Counts, o les tres variables corresponents als negocis a la rodona. És natural pensar que aquestes variables

59

Groups

)

2.5

1 % 5 . 3 1 ( 2 m i D

2

3

0.0

4

5

−2.5

1 . im

2 . im D

3 . im D

4 . im D

5 . im D

6 . im D

7 . im D D

8 . im D

Hotel\_lat

Hotel\_lng

Businesses\_100m

Businesses\_1km

Businesses\_5km

Stay\_Duration

Days\_Since\_Review

Total\_Number\_of\_Reviews

Review\_Positivity\_Rate

Review\_Total\_Negative\_Word\_Counts

Review\_Total\_Positive\_Word\_Counts

Average\_Score

Reviewer\_Score

Total\_Number\_of\_Reviews\_Reviewer\_Has\_Given

Additional\_Number\_of\_Scoring

Figure 47: Representativitat de les variables a cada dimensió

60

0.57

0.51

0.46

0.4

0.34

0.28

0.23

0.17

0.11

0.06

0

) % 5 . 3 1 ( 2 m i D

Variables − PCA

1.0

Review\_Total\_Negative\_Word\_Counts

0.5

Businesses\_1km

Businesses\_100m

Businesses\_5km

Stay\_Duration

0.0

Total\_Number\_of\_Reviews\_Reviewer\_Has\_Given

Days\_Since\_Review

Hotel\_lng Hotel\_lat Average\_Score Review\_Total\_Positive\_Word\_Counts

Additional\_Number\_of\_Scoring

Total\_Number\_of\_Reviews

−0.5

Reviewer\_Score Review\_Positivity\_Rate

−1.0

−1.0 −0.5 0.0 0.5 1.0 Dim1 (18.4%) Figure 48: Representació de les variables sobre els PC(1er pla)

61

contrib 15

10

5

contrib

5

10

15

0 . 1

0 . 1

5 . 0

5 . 0

Stay\_Duration

Hotel\_lat

0 . 0

4 \_ o i s n e m i

Businesses\_5km

Review\_Positivity\_Rate Days\_Since\_Review

Reviewer\_Score

D

Total\_Number\_of\_Reviews\_Reviewer\_Has\_Given

Businesses\_1km

Review\_Total\_Positive\_Word\_Counts

Review\_Total\_Negative\_Word\_Counts Average\_Score

Review\_Total\_Positive\_Word\_Counts

0 . 0

6 \_ o i s n e m i D Review\_Total\_Negative\_Word\_Counts

Additional\_Number\_of\_Scoring Review\_Positivity\_Rate

Average\_Score Total\_Number\_of\_Reviews

Reviewer\_Score

Businesses\_5km

Hotel\_lat

Businesses\_100m Businesses\_1km

Hotel\_lng Days\_Since\_Review 5 . 0 −

Additional\_Number\_of\_Scoring

Businesses\_100m

Stay\_Duration

Hotel\_lng

Total\_Number\_of\_Reviews

Total\_Number\_of\_Reviews\_Reviewer\_Has\_Given

Dimensio\_3

Dimensio\_5 −1.0 −0.5 0.0 0.5 1.0 Figure 49: Representació de les variables sobre els PC

es desplacin 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 intuïtivament 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 catègoriques (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 circunferència. En general, el gràfic ens confirma les relacions que hem establert al profiling anterior ja que, per exemple la nationalitat 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.

62

5 . 0 −

0 . 1 −

0 . 1 −

−1.0 −0.5 0.0 0.5 1.0

Variables − PCA

1.0

Review\_Total\_Negative\_Word\_Counts

0.5

Businesses\_1km

Businesses\_100m

Businesses\_5km

) % 5 .

Stay\_Duration Days\_Since\_Review

Review\_Total\_Positive\_Word\_Counts

Total\_Number\_of\_Reviews

Review\_Positivity\_Rate

−1.0 −0.5 0.0 0.5 1.0 Dim1 (18.4%) Figure 50: Representació de les variables sobre els PC(1er pla)

63

Cluster

3 1 ( 2 m i D

0.0

Total\_Number\_of\_Reviews\_Reviewer\_Has\_Given

a

1

a

2

a

3

Additional\_Number\_of\_Scoring

−1.0

Hotel\_lng Hotel\_lat Average\_Score

−0.5

Reviewer\_Score

6

4

] 2 e j e , [ i s P

2

0

FR

ES

AT

IT GB NL

2 −

4 −

−6 −4 −2 0 2 4 6

Psi[, eje1] Figure 51: Hotel\_Country al primer pla factorial

2

−2 −1 0 1 2

Psi[, eje1]

2

1

Paris FR

Boulogne Billancourt

1

Vienna

London GB

Milan

IT Amsterdam

NL Psi[, eje1]

No Vincennes Business

Turkey

Saudi Arabia

] 2 e j e , [ i s P

0

Barcelona ES

Classic

Family Family Travelers Suite

AT

with Leisure with With Solo Couple

Business Deluxe Yes

Other

Group young Family Duplex older No Family with Other a traveler pet

trip

Standard Solo friends

children children Couple

trip

traveler

] 2 e j e , [ i s P

0

United United States New Israel

Arab Belgium of Zealand Canada Switzerland France

Netherlands Sweeden

AmericaNoYes Greece Spain Germany United Yes Other Italy Ireland Emirates

No

Kuwait

Kingdom

Yes Hotel\_Country Hotel\_City Room\_Type\_Level

Is\_Reviewer\_Holiday Guest\_Type

Review\_Is\_Positive Trip\_Type

Reviewer\_Nationality Is\_Hotel\_Holiday

Submitted\_from\_Mobile

−2 −1 0 1 2

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

64

Australia

1 −

1 −

2 −

2 −

2

Psi[, eje3]

2

Vincennes

1

Boulogne Billancourt

1

Duplex ] 4 e j e , [ i s P

0

Family London GB Family Paris Solo FR StandardSuite with Solo Business Leisure Business

Deluxe

Couple Couple With with Classic traveler older Family Other traveler Other No Group Travelers NLAmsterdam Family young a Yes trip

children pet

trip

children

with friends ] 4 e j e , [ i s P

0

United United New Ireland

Kingdom

United Netherlands France

Belgium States Zealand Yes

Australia Yes Saudi No Switzerland

No YesNo Turkey Italy

Canada

Sweeden Kuwait Spain Germany Arab Other

of Arabia

America

Emirates

1 −

Hotel\_Country Hotel\_City

Barcelona ES

Milan IT

Vienna AT

1 −

Greece Israel

Room\_Type\_Level

Is\_Reviewer\_Holiday Guest\_Type

Review\_Is\_Positive Trip\_Type

Reviewer\_Nationality 2 −

Is\_Hotel\_Holiday

2 −

Submitted\_from\_Mobile

−2 −1 0 1 2

−2 −1 0 1 2

Psi[, eje3] Figure 53: Projeccions de totes les modalitats sobre Dim.3 i Dim.4

Tanmteix, de cara a estudis futurs podria ser interessent 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 qual- itatives 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 línea 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 ex- posem juntament el núvol de punts del les projeccions dels individus sobre els eixos de compo- nents 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-

65

2

Psi[, eje1]

2 Review\_Total\_Negative\_Word\_Counts

Review\_Total\_Negative\_Word\_Counts

1

Businesses\_1km

Businesses\_100m

Businesses\_5km

Paris FR

Boulogne Billancourt

1

Businesses\_1km

Businesses\_100m

Businesses\_5km

No

Business

Vincennes

Days\_Since\_Review

Barcelona

Stay\_Duration

Classic

ES

Solo Business Duplex

trip

] 2 e j e , [ i

0

Total\_Number\_of\_Reviews\_Reviewer\_Has\_Given

Family Family Travelers Hotel\_lng

with Vienna Leisure with Suite With AT

Couple Deluxe Other Group Family Yes Family Solo Other young No Standard older with London traveler a Couple

Hotel\_lat GB

pet

trip

traveler friends

children

children

Milan NLAmsterdam IT Average\_Score

]

0

Total\_Number\_of\_Reviews\_Reviewer\_Has\_Given

United United Days\_Since\_Review

Stay\_Duration

States New Australia Saudi Israel Switzerland Netherlands Turkey

United Belgium Arab Sweeden

Canada France

Germany Zealand Greece Spain

Ireland Other Italy

of Yes

NoYes Kuwait

No

Arabia

Emirates

America Kingdom

Review\_Total\_Positive\_Word\_Counts

Additional\_Number\_of\_Scoring Total\_Number\_of\_Reviews

Review\_Total\_Positive\_Word\_Counts

Additional\_Number\_of\_Scoring

Yes

Total\_Number\_of\_Reviews

Hotel\_Country Hotel\_City

Review\_Positivity\_Rate

Review\_Positivity\_Rate

Room\_Type\_Level

Is\_Reviewer\_Holiday Guest\_Type

Review\_Is\_Positive Trip\_Type

Reviewer\_Nationality Is\_Hotel\_Holiday

Submitted\_from\_Mobile

−4 −2 0 2 4

Psi[, eje1] Figure 54: Projeccions deles 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.

66

2 e j e , [ i

Hotel\_lng Hotel\_lat s P

1 −

Reviewer\_Score

s P

1 −

Average\_Score

2 −

2 −

Reviewer\_Score

3 −

3 −

−4 −2 0 2 4

PCA PCA − Biplot

− Biplot

7.5

7.5

5.0

5.0

Review\_Total\_Negative\_Word\_Counts

Modalitats

Business Businesses\_1km

Classic Businesses\_5km

Businesses\_100m

Stay\_Duration

Days\_Since\_Review

Hotel\_lng Hotel\_lat −4 0 4

−4 0 4 Dim1 (18.4%)

Dim1 (18.4%)

Hotel\_lng

Hotel\_lat Deluxe Duplex

Family

Other

Standard

Review\_Total\_Positive\_Word\_Counts

Suite

Total\_Number\_of\_Reviews

Review\_Positivity\_Rate

Review\_Total\_Negative\_Word\_Counts

Total\_Number\_of\_Reviews\_Reviewer\_Has\_Given

) % 5 . 3 1 ( 2 m i D

2.5

0.0

Modalitats

AT

ES

FR

GB

IT

) % 5 . 3 1 ( 2 m i D

2.5

0.0

Businesses\_1km

Businesses\_5km

Businesses\_100m

Stay\_Duration

Days\_Since\_Review Total\_Number\_of\_Reviews\_Reviewer\_Has\_Given

Average\_Score

Additional\_Number\_of\_Scoring

Hotel\_lng

Hotel\_lat NL

Average\_Score

Additional\_Number\_of\_Scoring

−2.5

−2.5

Review\_Total\_Positive\_Word\_Counts

Total\_Number\_of\_Reviews

Reviewer\_Score

Reviewer\_Score

−5.0

−5.0

Review\_Positivity\_Rate

PCA − Biplot

PCA − Biplot 7.5

7.5

5.0

5.0

Review\_Total\_Negative\_Word\_Counts Businesses\_1km

Businesses\_5km

Businesses\_100m

Stay\_Duration

Days\_Since\_Review

Modalitats

Business trip

Couple

Review\_Total\_Negative\_Word\_Counts

tal\_Number\_of\_Reviews\_Reviewer\_Has\_Given

) % 5 . 3 1 ( 2 m i D

)

Businesses\_1km 2.5

Businesses\_5km

Family

Leisure trip 2.5

Modalitats

0.0

Stay\_Duration

Days\_Since\_Review

No

Yes

Other

Hotel\_lng Hotel\_lat Additional\_Number\_of\_Scoring

Solo traveler

Review\_Total\_Positive\_Word\_Counts

Review\_Total\_Positive\_Word\_Counts Total\_Number\_of\_Reviews

Total\_Number\_of\_Reviews

−5.0

Review\_Positivity\_Rate

Review\_Positivity\_Rate −4 0 4

−4 0 4 Dim1 (18.4%)

Dim1 (18.4%) Figure 55: Representació de les variables sobre els PC (1er pla)

67

% 5 . 3 1 ( 2 m i

Businesses\_100m

D

0.0

Total\_Number\_of\_Reviews\_Reviewer\_Has\_Given

Average\_Score

Average\_Score

Additional\_Number\_of\_Scoring

−2.5

−2.5

Reviewer\_Score

Reviewer\_Score

−5.0

**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 analizar 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.

4.3% 4% 3.9% 3.8% 3.8% 3.6% 3.3%

2.7% 2.2% 2.1% 0

Scree plot

s e c n a i r a v d

40

e n i a l p x e

30

f o e g a t

20

n e c r e P

10

1 2 3 4 5 6 7 8 9 10 Dimensions 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

68

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 qual- itat que ha intervingut 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

69

ACM: Contribució modalitats a les dimensions

25

Barcelona ES

FR

GB Boulogne Billancourt

IT

0 5 10 15

Dim.1 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 oposats.

• La distància entre els punts de la modalitat i l’origen mesura la qualitat de la categoria vari- able 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).

70

Variable

a a a a a a a a a a

Guest\_Type 20

Hotel\_City

Hotel\_Country

2 . m i D

15

10

Paris

Is\_Hotel\_Holiday

Is\_Reviewer\_Holiday

Review\_Is\_Positive

Reviewer\_Nationality

Room\_Type\_Level

5

Submitted\_from\_Mobile

Reviewer\_Nationality\_France viewer\_Nationality\_Saudi Room\_Type\_Level\_Standard Reviewer\_Nationality\_Spain st\_Type\_Family iewer\_Nationality\_New Reviewer\_Nationality\_Turkey Room\_Type\_Level\_Duplex wer\_Nationality\_United eviewer\_Nationality\_Belgium viewer\_Nationality\_Switzerland Reviewer\_Nationality\_Israel \_Type\_Family Submitted\_from\_Mobile\_Yes eviewer\_Nationality\_Canada Reviewer\_Nationality\_Italy viewer\_Nationality\_Sweeden er\_Nationality\_United eviewer\_Nationality\_Australia Is\_Reviewer\_Holiday\_Yes iewer\_Nationality\_Netherlands eviewer\_Nationality\_Kuwait oom\_Type\_Level\_Business st\_Type\_Travelers eviewer\_Nationality\_Germany eviewer\_Nationality\_Ireland Is\_Reviewer\_Holiday\_No Room\_Type\_Level\_Other eviewer\_Nationality\_Greece Room\_Type\_Level\_Suite Room\_Type\_Level\_Deluxe Submitted\_from\_Mobile\_No Room\_Type\_Level\_Family Guest\_Type\_With Trip\_Type\_Solo Guest\_Type\_Solo Trip\_Type\_Business Trip\_Type\_Leisure Review\_Is\_Positive\_Yes Is\_Hotel\_Holiday\_Yes Review\_Is\_Positive\_No Is\_Hotel\_Holiday\_No

Guest\_Type\_Group

Trip\_Type\_Couple Trip\_Type\_Family Guest\_Type\_Couple Trip\_Type\_Other Vincennes NLAmsterdam

0 Room\_Type\_Level\_Classic

Reviewer\_Nationality\_Other with with traveler Reviewer\_Nationality\_United older with a traveler

young trip pet

Arab trip Zealand States friends children Arabia

children

Emirates Milan

of Vienna

America AT

Trip\_Type London Kingdom

ACM coordenades modalitats

Trip\_Type\_Other

Room\_Type\_Level\_Duplex

Room\_Type\_Level\_Family

Trip\_Type\_Solo traveler

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 rep- resentats 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 oposats estan correlacionades neg- ativament 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-

71

Variable 1

Room\_Type\_Level\_Suite

Reviewer\_Nationality\_Spain

Reviewer\_Nationality\_Israel

a

Guest\_Type

Guest\_Type\_With a pet

a

Hotel\_City

Reviewer\_Nationality\_Sweeden

a

Hotel\_Country

Reviewer\_Nationality\_United 2

. m i

0

Guest\_Type\_Family Guest\_Type\_Couple

Guest\_Type\_Family with older with young children

children

a a D

Reviewer\_Nationality\_Switzerland

a a a a a

−1.0 −0.5 0.0 0.5 1.0 1.5 Dim.1 Reviewer\_Nationality\_Ireland

London

Room\_Type\_Level\_Standard Trip\_Type\_Couple

GB Room\_Type\_Level\_Deluxe

Submitted\_from\_Mobile\_Yes Review\_Is\_Positive\_No

Room\_Type\_Level\_Other

Reviewer\_Nationality\_Kuwait

Room\_Type\_Level\_Business Is\_Reviewer\_Holiday\_No Reviewer\_Nationality\_New Trip\_Type\_Leisure Is\_Hotel\_Holiday\_No Guest\_Type\_Group

Submitted\_from\_Mobile\_No Reviewer\_Nationality\_United Review\_Is\_Positive\_Yes

Reviewer\_Nationality\_Netherlands Guest\_Type\_Travelers Guest\_Type\_Solo Reviewer\_Nationality\_United Reviewer\_Nationality\_Belgium Trip\_Type\_Business Reviewer\_Nationality\_Australia Reviewer\_Nationality\_Canada

Reviewer\_Nationality\_Italy

Reviewer\_Nationality\_Saudi Trip\_Type\_Family

Kingdom

Reviewer\_Nationality\_France

NLAmsterdam

Reviewer\_Nationality\_Other

Reviewer\_Nationality\_Germany Reviewer\_Nationality\_Greece

Reviewer\_Nationality\_Turkey

Room\_Type\_Level\_Classic

trip

traveler

trip with Zealand

friends States Arab Arabia Milan of Vienna Emirates IT AT

America

−1

Is\_Hotel\_Holiday

Is\_Reviewer\_Holiday

Review\_Is\_Positive

Reviewer\_Nationality

Room\_Type\_Level

Submitted\_from\_Mobile

Vincennes

Boulogne Paris

FR

Billancourt

Trip\_Type

AMC coordenades modalitats i registres

Room\_Type\_Level\_Deluxe

1

Room\_Type\_Level\_Standard

Trip\_Type\_Business Guest\_Type\_Solo NLAmsterdam

traveler

trip

Variable

a

Room\_Type\_Level\_Suite

a a Guest\_Type\_Family Guest\_Type\_Travelers Room\_Type\_Level\_Duplex

Guest\_Type\_With Boulogne Guest\_Type\_Family Trip\_Type\_Family

Room\_Type\_Level\_Family Trip\_Type\_Couple Vincennes

Billancourt with Guest\_Type\_Couple

older with a pet

friends children with young children

a a a a a a a

−1.0 −0.5 0.0 0.5 1.0 1.5 Dim.1 Figure 59: Representació de les modalitats i el núvol d’individus sobre els PC(1er pla)

AT

Guest\_Type Guest\_Type\_Group

Hotel\_City Reviewer\_Nationality\_Israel

Hotel\_Country

2 . m i D

0

Reviewer\_Nationality\_New Reviewer\_Nationality\_Sweeden Reviewer\_Nationality\_Netherlands

Reviewer\_Nationality\_Kuwait Room\_Type\_Level\_Business Reviewer\_Nationality\_Belgium Reviewer\_Nationality\_Canada Reviewer\_Nationality\_Spain

Is\_Reviewer\_Holiday\_No Reviewer\_Nationality\_Ireland Reviewer\_Nationality\_United Reviewer\_Nationality\_Italy Reviewer\_Nationality\_Switzerland Reviewer\_Nationality\_Saudi Trip\_Type\_Solo Is\_Hotel\_Holiday\_No Trip\_Type\_Leisure Reviewer\_Nationality\_Australia Reviewer\_Nationality\_France Reviewer\_Nationality\_Greece Reviewer\_Nationality\_Turkey Trip\_Type\_Other Submitted\_from\_Mobile\_Yes Reviewer\_Nationality\_Germany Review\_Is\_Positive\_Yes Review\_Is\_Positive\_No Submitted\_from\_Mobile\_No traveler

Reviewer\_Nationality\_United trip

Zealand

Arabia

Arab Emirates

States Is\_Hotel\_Holiday

of Is\_Reviewer\_Holiday

Review\_Is\_Positive

Reviewer\_Nationality

−1

Room\_Type\_Level

Submitted\_from\_Mobile

Trip\_Type

AMC Contribució modalitats a les dimensions

Milan IT

Paris FR

15

Variable

a a a a a a a a

Vienna

a a

oom\_Type\_Level\_Business Vincennes NLAmsterdam

GB

0 5 10 15

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

72

Guest\_Type

Hotel\_City

Hotel\_Country 10 3 . m i D

Is\_Hotel\_Holiday

Is\_Reviewer\_Holiday

Review\_Is\_Positive

Reviewer\_Nationality 5

Barcelona ES

Room\_Type\_Level Is\_Reviewer\_Holiday\_Yes Is\_Hotel\_Holiday\_Yes

Submitted\_from\_Mobile

viewer\_Nationality\_Saudi Room\_Type\_Level\_Standard st\_Type\_Family Reviewer\_Nationality\_Turkey viewer\_Nationality\_Switzerland wer\_Nationality\_United eviewer\_Nationality\_Belgium Room\_Type\_Level\_Duplex \_Type\_Family Submitted\_from\_Mobile\_Yes eviewer\_Nationality\_Canada eviewer\_Nationality\_Australia er\_Nationality\_United st\_Type\_Travelers Is\_Reviewer\_Holiday\_No Room\_Type\_Level\_Other Room\_Type\_Level\_Suite Room\_Type\_Level\_Deluxe Submitted\_from\_Mobile\_No Room\_Type\_Level\_Family Trip\_Type\_Solo Guest\_Type\_With Guest\_Type\_Solo Trip\_Type\_Business Trip\_Type\_Leisure Review\_Is\_Positive\_Yes Review\_Is\_Positive\_No Is\_Hotel\_Holiday\_No

Boulogne Guest\_Type\_Group Trip\_Type\_Couple Trip\_Type\_Family Guest\_Type\_Couple Trip\_Type\_Other

0 Room\_Type\_Level\_Classic

Reviewer\_Nationality\_Other

Billancourt

with with traveler

Reviewer\_Nationality\_United older with a traveler

young trip pet Arab trip

States friends children Arabia

children Emirates of America Kingdom

Trip\_Type

Reviewer\_Nationality\_France Reviewer\_Nationality\_Spain iewer\_Nationality\_New Reviewer\_Nationality\_Israel Reviewer\_Nationality\_Italy viewer\_Nationality\_Sweeden iewer\_Nationality\_Netherlands eviewer\_Nationality\_Kuwait eviewer\_Nationality\_Germany eviewer\_Nationality\_Ireland eviewer\_Nationality\_Greece Zealand

London

AMC coordenades modalitats

Guest\_Type\_With a pet

Boulogne Billancourt

London

Trip\_Type\_Business trip

Vienna

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

NLAmsterdam

Boulogne Vincennes

Billancourt Variable 1

a a a

Reviewer\_Nationality\_United 3

.

0

Reviewer\_Nationality\_Spain

Barcelona ES

Reviewer\_Nationality\_Ireland

Trip\_Type\_Other Trip\_Type\_Couple

GB Guest\_Type\_Family Room\_Type\_Level\_Deluxe

Submitted\_from\_Mobile\_Yes

Reviewer\_Nationality\_Sweeden

Room\_Type\_Level\_Other Guest\_Type\_Couple Guest\_Type\_Family Reviewer\_Nationality\_New Trip\_Type\_Leisure Submitted\_from\_Mobile\_No

Guest\_Type\_Group Reviewer\_Nationality\_United Review\_Is\_Positive\_Yes Guest\_Type\_Travelers Guest\_Type\_Solo Reviewer\_Nationality\_United Reviewer\_Nationality\_Belgium

Reviewer\_Nationality\_Australia Reviewer\_Nationality\_Canada

Room\_Type\_Level\_Suite

Room\_Type\_Level\_Duplex

Reviewer\_Nationality\_Italy

Reviewer\_Nationality\_Saudi Trip\_Type\_Family Room\_Type\_Level\_Family Kingdom Reviewer\_Nationality\_Israel Reviewer\_Nationality\_France

NLAmsterdam Reviewer\_Nationality\_Other

Reviewer\_Nationality\_Germany with Reviewer\_Nationality\_Greece Reviewer\_Nationality\_Turkey Room\_Type\_Level\_Classic

older trip with traveler

young children

with Zealand

friends children States Arab Arabia

of Emirates

America

a a a a a a a

−1.0 −0.5 0.0 0.5 1.0 1.5 Dim.1

Guest\_Type

Hotel\_City

Hotel\_Country

Is\_Hotel\_Holiday

m i D

Room\_Type\_Level\_Standard

Review\_Is\_Positive\_No

Reviewer\_Nationality\_Kuwait

Is\_Reviewer\_Holiday\_No

Is\_Hotel\_Holiday\_No Reviewer\_Nationality\_Netherlands

Is\_Reviewer\_Holiday

Review\_Is\_Positive Room\_Type\_Level\_Business

Reviewer\_Nationality

−1

Trip\_Type\_Solo Reviewer\_Nationality\_Switzerland

traveler

AT

Room\_Type\_Level

Submitted\_from\_Mobile

Trip\_Type

AMC coordenades modalitats i registres

Room\_Type\_Level\_Deluxe

1

Room\_Type\_Level\_Standard

Trip\_Type\_Business Guest\_Type\_Solo traveler

trip

Variable

a

Room\_Type\_Level\_Suite

a a Guest\_Type\_Family Guest\_Type\_Travelers Room\_Type\_Level\_Duplex

Guest\_Type\_With Guest\_Type\_Family Trip\_Type\_Family

Room\_Type\_Level\_Family Trip\_Type\_Couple with Guest\_Type\_Couple

older with a pet

friends children with young children

a a a a a a a

−1.0 −0.5 0.0 0.5 1.0 1.5 Dim.1 Figure 62: Representació de les modalitats i el núvol d’individus sobre els PC(Dim.1 vs Dim.3)

73

Guest\_Type Guest\_Type\_Group

Hotel\_City Reviewer\_Nationality\_Israel

Hotel\_Country

3 . m i D

0

Reviewer\_Nationality\_New Reviewer\_Nationality\_Sweeden Reviewer\_Nationality\_Netherlands Reviewer\_Nationality\_Kuwait Room\_Type\_Level\_Business Reviewer\_Nationality\_Belgium Reviewer\_Nationality\_Canada Reviewer\_Nationality\_Spain

Is\_Reviewer\_Holiday\_No Reviewer\_Nationality\_Ireland Reviewer\_Nationality\_United Reviewer\_Nationality\_Italy Reviewer\_Nationality\_Switzerland Reviewer\_Nationality\_Saudi Trip\_Type\_Solo Is\_Hotel\_Holiday\_No Trip\_Type\_Leisure Reviewer\_Nationality\_Australia Reviewer\_Nationality\_France Reviewer\_Nationality\_Greece Reviewer\_Nationality\_Turkey Trip\_Type\_Other Submitted\_from\_Mobile\_Yes

Reviewer\_Nationality\_Germany Review\_Is\_Positive\_Yes Review\_Is\_Positive\_No Submitted\_from\_Mobile\_No traveler

Reviewer\_Nationality\_United trip

Zealand

Arabia

Arab Emirates

States Is\_Hotel\_Holiday

of Is\_Reviewer\_Holiday

Review\_Is\_Positive

Reviewer\_Nationality

−1

Room\_Type\_Level

Submitted\_from\_Mobile

Trip\_Type

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 combi- nació entre elles, s’acaben treien les mateixes conclusions sobre quina modalitat esta millor repre- sentada 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)

AMC Contribució modalitats a les dimensions 30

Vienna AT

Room\_Type\_Level\_Classic Room\_Type\_Level\_Deluxe

oom\_Type\_Level\_Business Vincennes ES

0 10 20 30

Dim.4 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 difer- enciar quina té una major contribució. En canvi, a la dimensió 5 és Viena (Aùstria) 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 Nacionality\_Israel està correlacionada negativament amb ciutat de l’hotel Billancourt, ja que aquestes es troben en quadrants oposats 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 diferènciar entre individus.

74

Variable

a a a a a a a a a a

Guest\_Type

Hotel\_City 20

Hotel\_Country

5 . m i D

Is\_Hotel\_Holiday

Is\_Reviewer\_Holiday

Review\_Is\_Positive

10

NLAmsterdam

Reviewer\_Nationality

Room\_Type\_Level

Submitted\_from\_Mobile

Trip\_Type

Boulogne Barcelona

Billancourt

Milan IT

Reviewer\_Nationality\_Spain \_Type\_Family Reviewer\_Nationality\_Israel iewer\_Nationality\_Saudi eviewer\_Nationality\_Belgium iewer\_Nationality\_New t\_Type\_Family eviewer\_Nationality\_Australia wer\_Nationality\_United viewer\_Nationality\_Switzerland iewer\_Nationality\_Netherlands r\_Nationality\_United Nationality\_United eviewer\_Nationality\_Greece Reviewer\_Nationality\_Italy Room\_Type\_Level\_Family viewer\_Nationality\_Sweeden eviewer\_Nationality\_France viewer\_Nationality\_Germany st\_Type\_Travelers eviewer\_Nationality\_Canada ubmitted\_from\_Mobile\_Yes Guest\_Type\_Solo eviewer\_Nationality\_Kuwait Room\_Type\_Level\_Other Is\_Reviewer\_Holiday\_No oom\_Type\_Level\_Standard ubmitted\_from\_Mobile\_No eviewer\_Nationality\_Turkey eviewer\_Nationality\_Ireland eviewer\_Nationality\_Other Room\_Type\_Level\_Suite Review\_Is\_Positive\_Yes Trip\_Type\_Business Trip\_Type\_Solo Review\_Is\_Positive\_No Guest\_Type\_With Room\_Type\_Level\_Duplex Trip\_Type\_Leisure Is\_Hotel\_Holiday\_No Guest\_Type\_Couple Is\_Reviewer\_Holiday\_Yes Guest\_Type\_Group

Trip\_Type\_Couple Trip\_Type\_Family Trip\_Type\_Other Is\_Hotel\_Holiday\_Yes 0 London

Paris FR GB with with traveler young traveler

with States older a Arab trip pet trip

Zealand Kingdom friends Arabia children children of Emirates America

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 recordartori, 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 nom- bre 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

75

Variable 1

a a a a a a a a a a

−1.0 −0.5 0.0 0.5 1.0 1.5 Dim.4 Reviewer\_Nationality\_Turkey

Guest\_Type Reviewer\_Nationality\_Switzerland

5 . m i D

Guest\_Type\_Family Reviewer\_Nationality\_Spain Reviewer\_Nationality\_Greece 0 Reviewer\_Nationality\_Kuwait Reviewer\_Nationality\_United Guest\_Type\_With Reviewer\_Nationality\_Italy

Reviewer\_Nationality\_Australia Guest\_Type\_Family Reviewer\_Nationality\_United Reviewer\_Nationality\_United Room\_Type\_Level\_Family

Reviewer\_Nationality\_Sweeden Reviewer\_Nationality\_Saudi Trip\_Type\_Solo Trip\_Type\_Family Boulogne Reviewer\_Nationality\_France

Reviewer\_Nationality\_Netherlands Guest\_Type\_Travelers Guest\_Type\_Solo Submitted\_from\_Mobile\_Yes Trip\_Type\_Business Submitted\_from\_Mobile\_No Room\_Type\_Level\_Classic Room\_Type\_Level\_Deluxe Room\_Type\_Level\_Other

Barcelona Reviewer\_Nationality\_Other Trip\_Type\_Couple Review\_Is\_Positive\_Yes Reviewer\_Nationality\_Germany Is\_Reviewer\_Holiday\_No Room\_Type\_Level\_Business Review\_Is\_Positive\_No Reviewer\_Nationality\_Belgium Trip\_Type\_Leisure Room\_Type\_Level\_Standard Paris Is\_Hotel\_Holiday\_No Vincennes

ES

Guest\_Type\_Couple FR

London

Billancourt

GB

Reviewer\_Nationality\_New Guest\_Type\_Group

Reviewer\_Nationality\_Ireland

Room\_Type\_Level\_Suite

traveler

Reviewer\_Nationality\_Canada

traveler

States trip

with Kingdom Arab Arabia

friends of Zealand

America Emirates Hotel\_City

Hotel\_Country

Is\_Hotel\_Holiday

Is\_Reviewer\_Holiday

Review\_Is\_Positive Reviewer\_Nationality\_Israel

Reviewer\_Nationality

−1

Room\_Type\_Level

Submitted\_from\_Mobile

Trip\_Type with a pet

with older young children

trip

children

Trip\_Type\_Other

)

of\_Reviews\_Reviewer\_Has\_Given

Additional\_Number\_of\_Scoring

% 8 4 .

**Variables factor map (PCA)**

Businesses\_5km

Stay\_Duration Total\_Number\_of\_Reviews

Hotel\_lat Days\_Since\_Review

Hotel\_lng Review\_Total\_Positive\_Word\_Coun Average\_Score Reviewer\_Score Businesses\_100m Businesses\_1km )

% 7 2

Hotel\_lat

Dim 1 (18.38%)

**Variables factor map (PCA)**

3 1 ( 2

0 . 1 −

Review\_Positivity\_Rate

Review\_Total\_Negative\_Word\_Cou

. 8 ( 4 im

0 . 1 −

Additional\_Number\_of\_Scoring

Total\_Number\_of\_Reviews Hotel\_lng Businesses\_5km

m i D

−15 −10 −5 0 5 10 15

D

−15 −10 −5 0 5 10 15

Dim 3 (12.13%)

) % 8 4 .

**Individuals factor map (PCA)**

)

−100 −50 0 50 100

−60 −40 −20 0 20 40 60

Dim 1 (18.38%)

Figure 65: Resum ACP

**Aggregated distance at each iteration**

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

76

% 7 2

6

**Individuals factor map (PCA)**

3 1 ( 2

2 8 −

. 8 ( 4 m i

2 − m i D

D

Dim 3 (12.13%)

**Dendrograma−WARD**

s e i c n à t s i D

observació Figure 67: Dendograma-mètode de Ward amb distàncies euclidianes

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

Veiem com, en aquest cas, la partició no és tan bona (inclús 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 dis- tinció entre els conglomerats 3 i 5 que, recordem, es corresponien 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 difer- ents 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

77

Individuals − PCA

2.5

−4 0 4

Dim1 (18.4%) 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

78

Groups

) % 5 . 3 1 ( 2 m i

0.0

1

2

3

D

−2.5

4

5

−5.0

−7.5

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

79

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

80

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

81

$‘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

82

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

83

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 defineixi els conglom- erats, 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 com- ponents del clúster. En aquest cas, com a repostes 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

84

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 \*\*\*

85

--- 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 caracter- itzen els nostres conglomerats (latitud i longitud, Negocis a la rodona. . . ). A més, tots dos models són significatius i les puntuacions (respostes) semblen ser máximes 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. Tan- mateix, a continuació mostrem una pinzellada del que ens ofereix FactoMineR per a realitzar aque- sta descripció.

FactoMineR ens permet calcular estadístics numèrics per variable i clúster. En aquest cas, pro- posem 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 anteior 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.hcpcdesc.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

86

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

87

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

88

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

89

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

90

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 factorials. En aquest cas, hem considerat representacions fins a la sisena dimensió (Figure 69).

−2

Individuals − PCA

−5

Individuals − PCA

Figure 69: Clustering jeràrquic ACP (plans factorials adicionals)

Veiem com, a mesura que avancem en les dimensions, la variabilitat dismineix 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.

91

10

4

5

Groups

1

2

3

4

5

−2.5 0.0 2.5 5.0

−3 0 3 6 Dim3 (12.1%)

Dim5 (7.5%)

Groups

) %

2

3 .

1

8 (

) % 7 ( 6

2

4 m i

m i D

3 D

4 0

5

0

**Profiling del Clúster Jeràrquic sobre ACP**

En aquesta secció, repetim el profiling, ara considerant el clúster obtingut a partir de l’ACP. Recor- dem 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 constrastos, 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\_lng 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).

0

**Hotel\_Country per cluster (%)**

cluster

**Hotel\_Country per cluster (nre. d'hotels)**

0 . 1

AT

FR

IT ES

GB

NL

0 0 4 1

AT

GB ES

IT

0 0

FR

NL

8 . 0

2 1

0 0 0 6 . 0

m

cluster 1

g r a

%

s l e t o

0 0 8

4 . 0

1 2 3 4 5

Figure 70: Profiling2 variable Hotel\_Country

Observem com la situació ha canviat totalment, ara els hotels de Gran Bretanya es trobem 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 forta presència Francesa.

Si considerem la variable Hotel\_City, de nou observem com les ciutats concorden amb els re- sultat 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 estrat- ificant 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

92

h #

0 0 6

2 . 0

0 0 4

0 0 2

0 .

0

1 2 3 4 5

**Hotel\_City per cluster (%)**

cluster

**Hotel\_City per cluster (nre. d'hotels)**

cluster

0

m

.

Amsterdam Barcelona Boulogne Billancourt

London Milan Paris

Amsterdam Barcelona Boulogne Billancourt London

Milan Paris Vienna Vincennes

1

0 0 5 1

8 . 0

6 . 0 g r a

s l e t %

o

0 0 0 1

4 . 0

1 2 3 4 5

Figure 71: Profiling2 variable Hotel\_City

h #

2 . 0

0 0 5

0 . 0

0

1 2 3 4 5

t a l

55

55.0927511466334 55.0927511466334

50

45

40

35.6241493912721 35.6241493912721 −10 −5 0 5 10 15 lon

cluster

Total number of reviews

4000

8000

12000

16000

Figure 72: Geolocalització

93

5

4

3

2

1

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).

**Plot Means of Businesses\_100m**

cluster

**Plot Means of Businesses\_1km**

cluster

**Plot Means of Businesses\_5km**

0 0 5 3

0 0 0 0 2

0 5

0 0 0 3

0 0 0 8 1

m 0 0 1 \_ s

0 4 e s s

m k 1 \_ s e s s

0 0 5 2 e

m k 5 \_ s e s s e

0 0 0 6 1 e n i s u B a n a j t i m

n i s u B a n a j t i m

0 0 0

n i s u B a n a j t i m 0 3

2

0 0 0 4 1

0 0 5 1

2

1 1

1 2 3 4 5

1 2 3 4 5

1 2 3 4 5

cluster 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 salvetat 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

94

0 0 0 2 1

0

0 0 0

0 0 0 0

**Room\_Type\_Level per cluster (%)**

cluster

**Room\_Type\_Level per cluster (nre. d'hotels)**

cluster

0

m

.

0 Business

Business Classic

Classic Deluxe

Deluxe Duplex

Duplex Family Other Standard Suite

Family Other Standard Suite

1

0 4 1

0 0

8

2

.

1

0

a

Figure 74: Profiling2 variable Room\_Type\_Level

0

6 .

i r o g e t

0 0 0 1 0 g r a

%

a c r e p

0 0 8

4 . 0

1 2 3 4 5

s n o i c a t i b a h #

0 0 6

2 . 0

0 0 4

0 0 2

0 . 0

0

1 2 3 4 5

**Guest\_Type per cluster (%)**

cluster

**Guest\_Type per cluster (nre. d'hotels)**

0 . 1

Couple Family with older children

0 0 4 1

Couple

Solo traveler Family with older children

Travelers with friends Family with young children Group

0 0

Family with young children

With a pet

8 .

Solo traveler

2 1

Group 0

Travelers with friends With a pet

a

0 0

6

m

cluster Figure 75: Profiling2 variable Guest\_Type

95

i r o g

0 1 . 0 g r a

%

e t a c r e

0 0 8

4 . 0

1 2 3 4 5

p s t n e i l c #

0 0 6

2 . 0

0 0 4

0 0 2

0 .

0

1 2 3 4 5

**Trip\_Type per cluster (%)**

cluster

**Trip\_Type per cluster (nre. d'hotels)**

cluster

0

m

.

0 Business trip

Business trip Couple

Couple Family

Family Leisure trip Other Solo traveler

Leisure trip Other Solo traveler

1

0 4 1

0 0

8

2

.

1

0

0 a

0

6 .

Figure 76: Profiling2 variable Trip\_Type

1

i r o g

0 1 0 g r a

%

e t a c r e

0 0 8

4 . 0

1 2 3 4 5

p s e g t a i v #

0 0 6

2 . 0

0 0 4

0 0 2

0 . 0

0

1 2 3 4 5

**Boxplot of Stay\_Duration vs Cluster**

Dies d'estada

**Means of Stay\_Duration by Cluster**

5

5 . 2

global mean

4

a n a j t i

5 10 15 20

cluster Figure 77: Profiling2 variable Stay\_Duration

96

0 . 2

r e t s u

3

m a i c n à t

5 . 1 l c

2

s e a d a r u D

0 . 1

5 . 0

0 . 0

1 2 3 4 5

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.

**Is\_Hotel\_Holiday per cluster (nre. d'hotels)**

1 2 3 4 5 cluster

**Is\_Reviewer\_Holiday per cluster (nre. d'hotels)**

No Yes

No Yes

0 0 5 1

0 0 5 1

s e y n

0 0 0 1

cluster s

e s s e r #

1 2 3 4 5

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 extremada- ment 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 ocòrre 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 accen- tuat respecte al profiling anterior.

97

e y n e s s e r #

0 0 0 1

0 0 5

0 0 5

0

0

**Boxplot of Total\_Number\_of\_Reviews vs Cluster**

Nombre de ressenyes vàlides

**Means of Total\_Number\_of\_Reviews by Cluster**

5

0 0 0 7

s

0 0 0

global mean

0 5000 10000 15000

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

1 2 3 4 5

e

6

4

d i l à v s e

0 0 0 5

r e t s u l

3

y n e s s e r e

0 0 0 4 c

d g i m e r

0 0 0 3 2

b m o N

0 0 0 2

1

0 0 0 1

0

1 2 3 4 5

**Review\_Is\_Positive (nre. d'hotels)**

cluster

**Means of Review\_Positivity\_Rate by Cluster**

No Yes

0 7

0 0 5 1

0 6

global mean

cluster Figure 80: Profiling2 variables grau de positivisme de la ressenya

98

0 5

e t s l e t

0 0 0 1

a R y

0 4 o h #

t i v i t i s o P

0 3

0 0 5

0 2

0 1

0

0

1 2 3 4 5

Observant el gràfic, podem concloure 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.

0

**Reviewer\_Nationality per cluster (%)**

cluster

**Reviewer\_Nationality per cluster (nre. d'hotels)**

0 . 1

Australia

Israel

S

Australia

New Zealand Belgium

Italy

S

Belgium

Other Canada France Germany

Kuwait Netherlands New Zealand

S T U

0 0 5 1

Canada France Germany Greece

Other

U Ireland

Saudi Arabia

U

m

cluster

Saudi Arabia

8 . 0

Spain Sweeden Greece

Switzerland Ireland

Turkey

6 . 0 g r a

Israel

United Arab Emirates Italy

United Kingdom Kuwait

United States of America Netherlands %

is r a u

0 0 0 1

4 . 0

1 2 3 4 5

Figure 81: Profiling2 variable Reviewer\_Nationality

Les següents variables, si excloem les textuals, són Review\_Total\_Negative\_Word\_Counts i Re- view\_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, ab- sorbeix 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 analitzantReview\_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 exepció 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.

99

s u #

2 . 0

0 0 5

0 .

0

1 2 3 4 5

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

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

0 5

2 2

s t n u o C

0

1 2 3 4 5

cluster

cluster Figure 82: Profiling2 Nre.paraules comentari Negatiu/Postiu

**Plot Means Average\_Score**

cluster

s t

0

1 1

1 2 3 4 5

n u o C \_ d r

2 \_ d r o W \_ e v i t a g

4

o W \_ e v i t i

8 1

e N \_ l a t o T \_ w e i v

0 3

s o P \_ l a t o

6 1

T \_ w e i v e

4 1

e R a n a j t i m

0 2

R a n a j t i m

2 1

0 0

**Plot Means of Reviewer\_Score**

6 . 8

0 . 9

e r o c

5 . 8

8

1 2 3 4 5

1 2 3 4 5

cluster Figure 83: Profiling2 variable Valoracions

100

5 e r o c S \_ r e w e i v e R a n a j t i m

. 8

S \_ e g a r e

4 . 8

0 . 8

v A a n a j t i

3 . 8

5 . 7

m

2 . 8

0 . 7 1 . 8

5 . 6 0 .

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

Total de ressenyes escrites pels usuaris

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

2 1 s 5

global mean

cluster Figure 84: Profiling2 variable Total\_Reviews\_Given

1

i r a

0 50 100 150

u s u s l

0 1

4

e p s e t i r c

8

r e t s u l c

s

3

e s e y n e s s e

6

2

r e d g i m e r

4

b m

2 1

o N

0

1 2 3 4 5

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

Nombre de valoracions adicionals

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

5

0 0 5 1 s

global mean

0

0 500 1000 1500 2000 2500

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

101

l a n o i 4

c i d a s

0 0

r

n o i c a

0 1 e t s u l c

3

2

r o l a v e d g i m e r b m o N

0 0 5

1 2 3 4 5

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 per- centatge de ressenyes escrites des del mòvil al clúster 3.

0

**Submitted\_from\_Mobile (%)**

cluster

**Submitted\_from\_Mobile (nre. d'hotels)**

0 . 1

No Yes

No Yes

8 . 0

0 0 5 1

g r a m

6 . 0

%

s

0

cluster 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

i r a u

0 0 1

4 . 0

1 2 3 4 5

s u #

2 . 0

0 0 5

0 .

0

1 2 3 4 5

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, Vienna i Milan no excessivament cèntrics pero

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 families 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

102

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 altes puntuacions

103

**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 pa- raules 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 pre- viament) 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 textuals que disposem, a partir de la funció Textdata.

Mitjançant aquesta funció podem saber quantes vegades apareix cada paraula de les que for- men les nostres variables textuals. 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

104

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 anàlisi 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 estudi- ades.

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 mostrarà aquestes contribucions d’una manera més gràfica:

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

105

**CA factor map**

2 ) % 2 7 . 2 ( 2

1

metro

m i D

0

amazing

excellent restaurant perfect helpful

service

reception location

breakfast london

facilities station

food staytime hotel

staff check

friendly close

bar city

wifi

nice quite rooms bathroom night

lovely day

free clean even

easy walk area

water coffeecomfortable near

quiet

expensive good price

little only modern

comfy

people view

floor room

small noisy air door spacious

beds bed shower

Dim 1 (2.88%) Figure 87: Constribucions 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 textuals 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 textuals. 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 con- tribuiran a aquesta dimensió. Com podem veure aire, sorollós i confortable són les paraules que més aporten a la cinquena dimensió.

106

size

1 −

−4 −2 0 2 4

**CA factor map**

0 . 2

amazing

) % 4 4 . 2 ( 4

0 . 1

m i

0 . 0 D

0 . 1 −

comfortable metro

friendly

beds comfy modern

station helpful bathroombed

rooms clean

location perfect

facilities expensive

close easy london area

near

quite staff lovely

quiet city

noisy spacious stay little hotel walk

good sizesmall

free

shower air only

nice view

wifi

excellent breakfast coffee door room day people

price water floor

night even reception time bar

restaurant

check

food

service

−2 0 2 4

Dim 3 (2.55%) Figure 88: Contribucions D3 vs D4

**CA factor map**

2

1 ) % 4 3 . 2 ( 5

−5 0 5 0

comfortable bathroom excellent friendly comfy

helpful spacious breakfast facilities shower service beds

coffee staff

rooms bed

water clean

small size

quite lovely nice

day door view wifi bar little amazing

free time stay food

reception hotel room even floor modern people good location restaurant quiet expensive

check only night price area london perfect walk city easy

close near station

metro

m i D

2 −

noisy

4 −

air

Dim 2 (2.72%) Figure 89: Dimensio vs nre. paraules

107

**Anàlisi Comparativa dels diferents mètodes i conclusions generals**

Per finalitzar, exposem les conclusions que hem obtingut a partir dels diferents mètodes i tèc- niques 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 re- forçades 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 breuement mostrar com el nostre anàlisi por 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 aventatge per als hotels amb una pitjor performance ja que son 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 prior 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 sen- tit, aquests hotels més allunyats del centre que a priori ofereixen un preu més econòmic

108

tenen un gran marge de millorar per a intentar contentar aquest col.lectiu. Una possible es- tratègia 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 col.lectiu 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

109

Distribució de tasques

Figure 91: Distribució de tasques

110