

TÈCNIQUES DE MINERIA DE DADES: APLICACIÓ A LA INDUSTRIA HOTELERA

Allès Pons, Hugo
Blanco Conde, Carles
Fibla Salgado, Aleix
Miranda Hernández, Victor

Morante López, Pablo
Ramoneda Montoya, Antoni
Rovira Tauler, Oriol
Salvador Barrera, Aleix

3r Curs

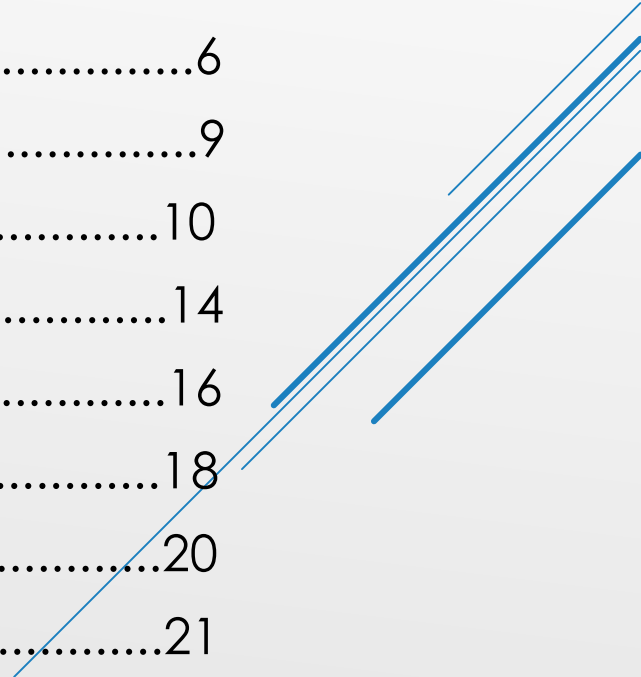
LES NOSTRES DADES

- ▶ Són propietat de Booking.com. Nosaltres estem autoritzats per ús acadèmic.
- ▶ Publicades inicialment a *kaggle* per Jason Liu i enriquides posteriorment per Yanir, un altre usuari.
- ▶ Es pot obtenir del següent enllaç: <https://www.kaggle.com/ycalisar/hotel-reviews-dataset-enriched>.
- ▶ Conté aproximadament 515,000 ressenyes de clients de 1,493 hotels de luxe a Europa.

Aquest treball s'ha desenvolupat amb l'objectiu de millorar les experiències de viatges en l'àmbit de la indústria hotelera i generar valor tant per als establiments com per als clients.

























ÍNDEX




▶ Descripció formal de l'estructura de les dades.....	4
▶ Clústering jeràrquic.....	5
▶ Profiling.....	6
▶ Perfil de les classes	9
▶ Anàlisi de components principals.....	10
▶ Anàlisi de components múltiples.....	14
▶ Clustering jeràrquic sobre ACP	16
▶ Anàlisi textual	18
▶ Conclusions	20
▶ Annex: Planificació original i planificació final	21



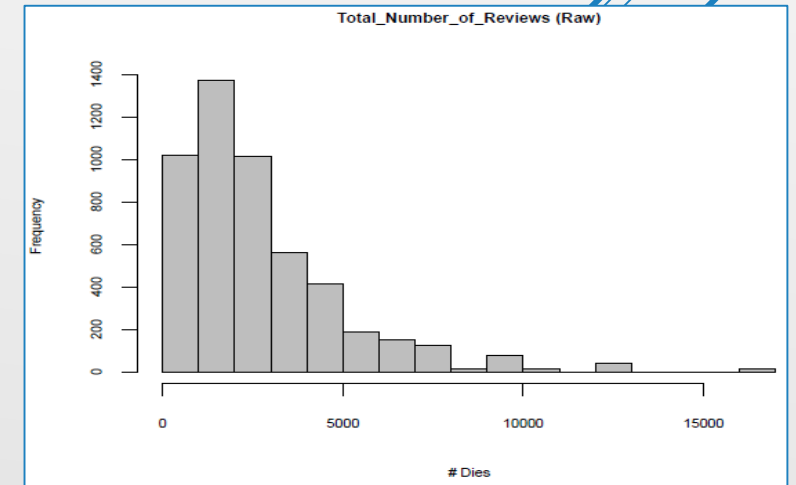
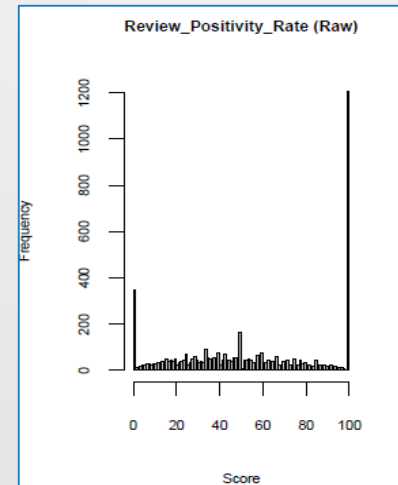
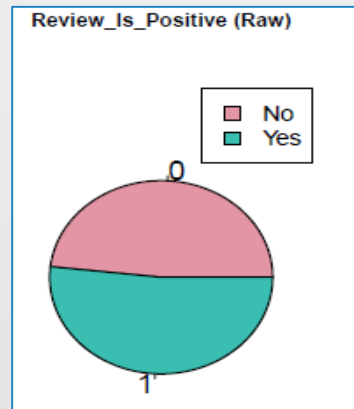
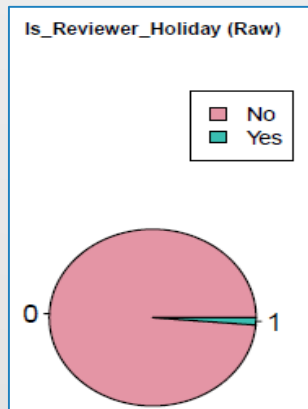
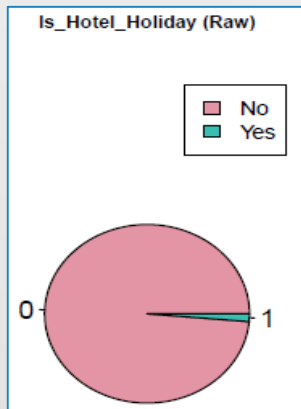
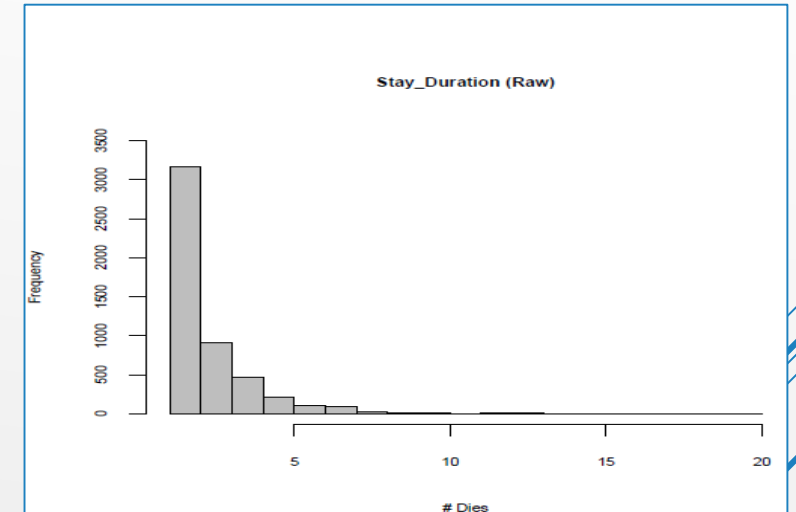
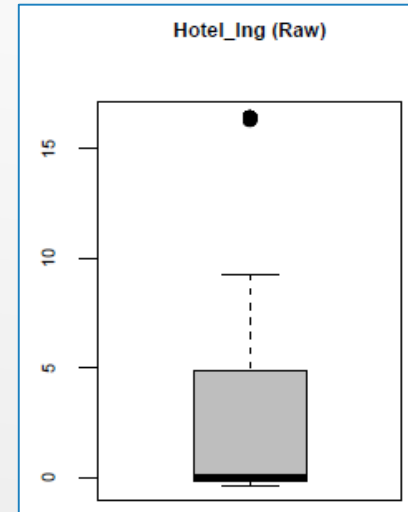
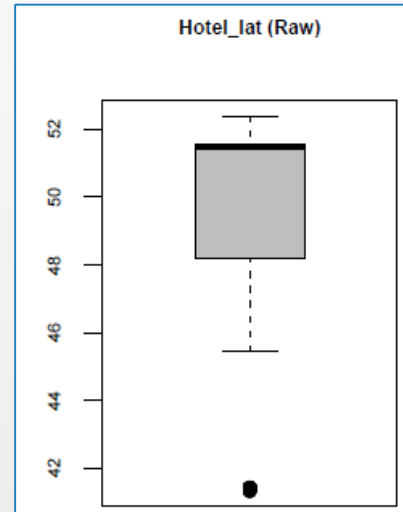
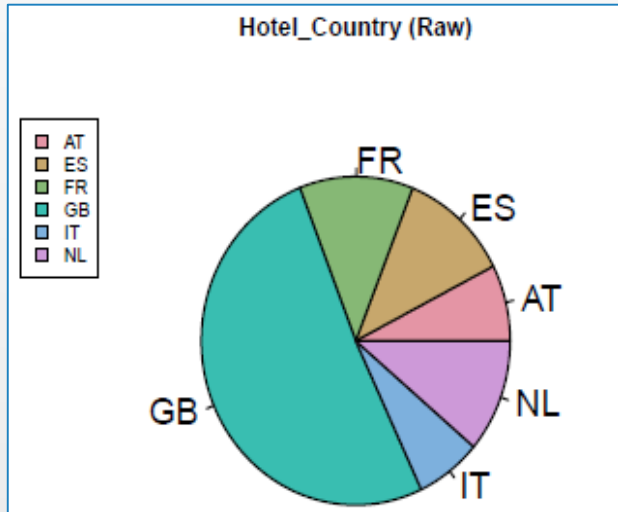
DESCRIPCIÓ FORMAL DE L'ESTRUCTURA DE LES DADES

- ▶ Mostra aleatòria de 5.000 observacions a partir de les dades originals.
- ▶ Selecció de 30 variables considerades rellevants per als objectius de l'estudi

▶ ID 	▶ Stay_Duration 	rd_Counts 
▶ Hotel_Name 	▶ Review_Date 	▶ Positive_Review 
▶ Hotel_Country 	▶ Days_Since_Review 	▶ Review_Total_Positive_Word_Counts 
▶ Hotel_City 	▶ Is_Hotel_Holiday 	▶ Average_Score 
▶ Hotel_Lat 	▶ Is_Reviewer_Holiday 	▶ Reviewer_Score 
▶ Hotel_Ing 	▶ Total_Number_of_Reviews 	▶ Total_Number_of_Reviews_Has_Given 
▶ Businesses_100m 	▶ Review_Is_Positive 	▶ Additional_Number_of_Scoring 
▶ Businesses_1km 	▶ Review_Positivity_Rate 	▶ Submitted_from_Mobile 
▶ Businesses_5km 	▶ Reviewer_Nationality 	
▶ Room_Type_Level 	▶ Negative_Review 	
▶ Guest_Type 	▶ Review_Total_Negative_Word_Counts 	
▶ Trip_type 		

 Alfanumérica
 Numérica
 Data
 Binària

DESCRIPTIVA UNIVARIANT



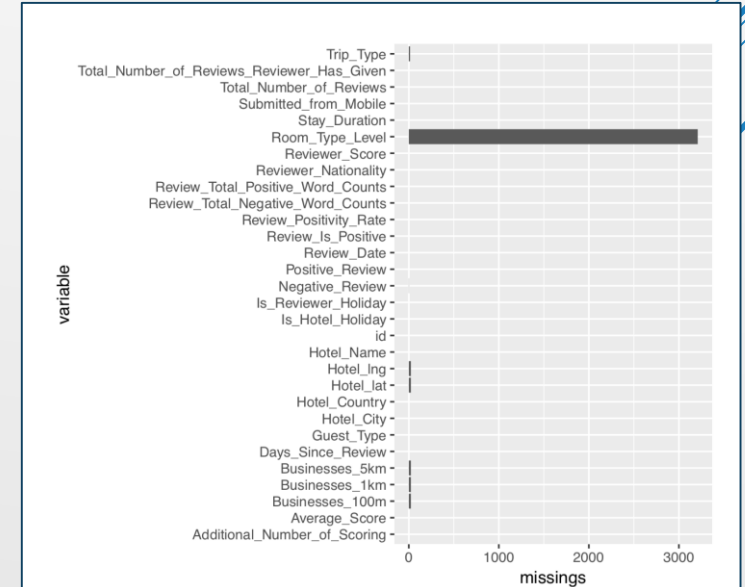
PREPROCESSAMENT DE LES DADES

- ▶ Fem un resum inicial per controlar els màxims, mínims, errors i valors anòmals.
- ▶ Corregir la classe de cada variable (entre d'altres):
 - ▶ Character → Numèric.
 - ▶ Character → Factor.
 - ▶ Integer → Date.
- ▶ Modalitats variables categòriques → recategorització dels nivells.
- ▶ Control de Valors Missing (NA):
 - ▶ Variables numèriques → Imputació KNN.
 - ▶ Variables categòriques → Recategoritzem com "Altres".

Room Type Level

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		

Ambassadors	Art	Business	City	Classic	Deluxe
1	18	61	9	343	191
Duplex	Executive	Family	Luxury	Other	Premium
10	83	131	9	3209	1
Privilege	Standard	Studio	Suite	Superior	
8	546	23	71	286	

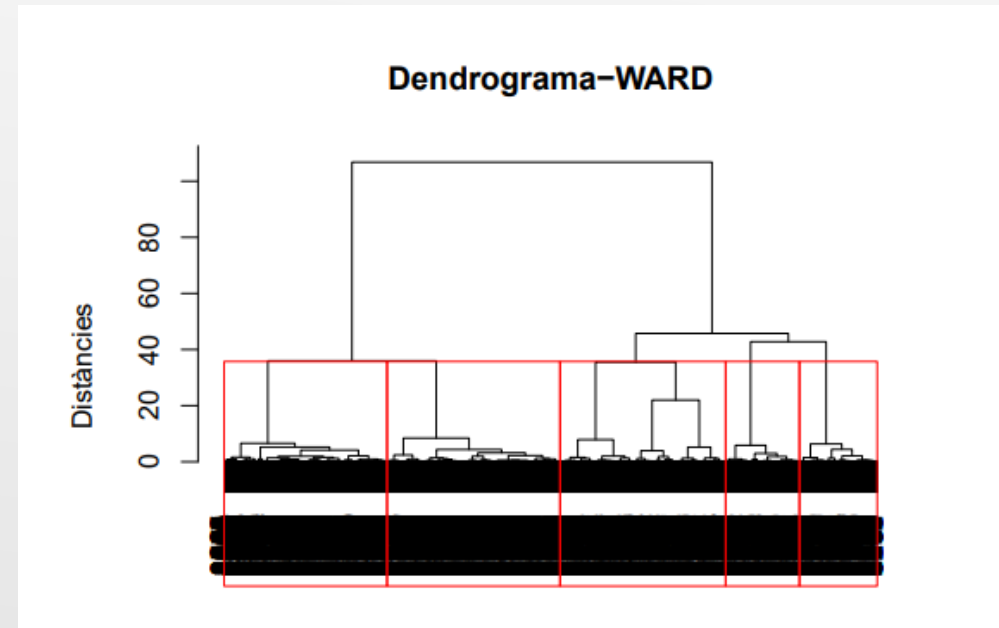


Test de
Little:
0,0587

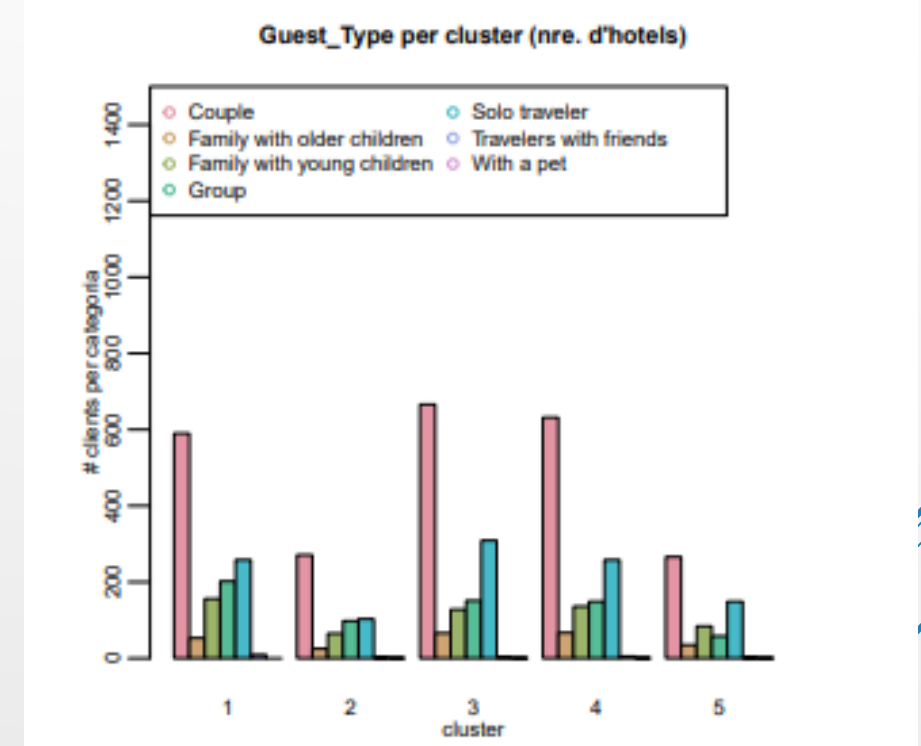
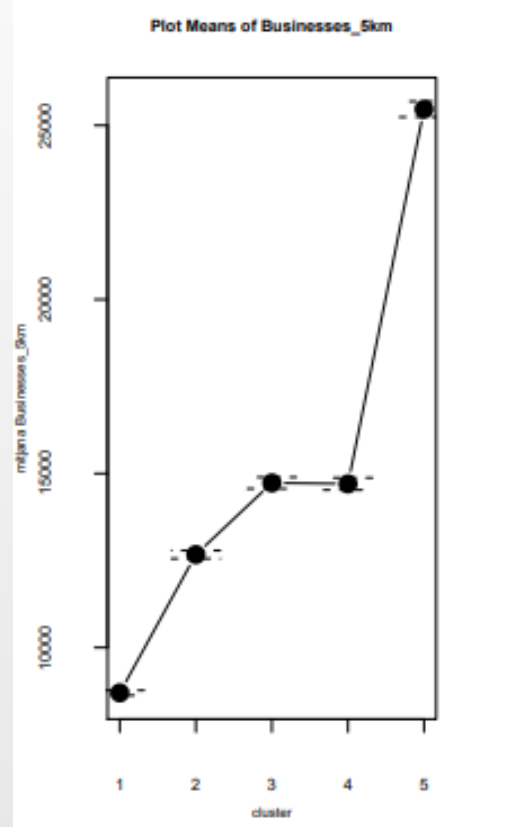
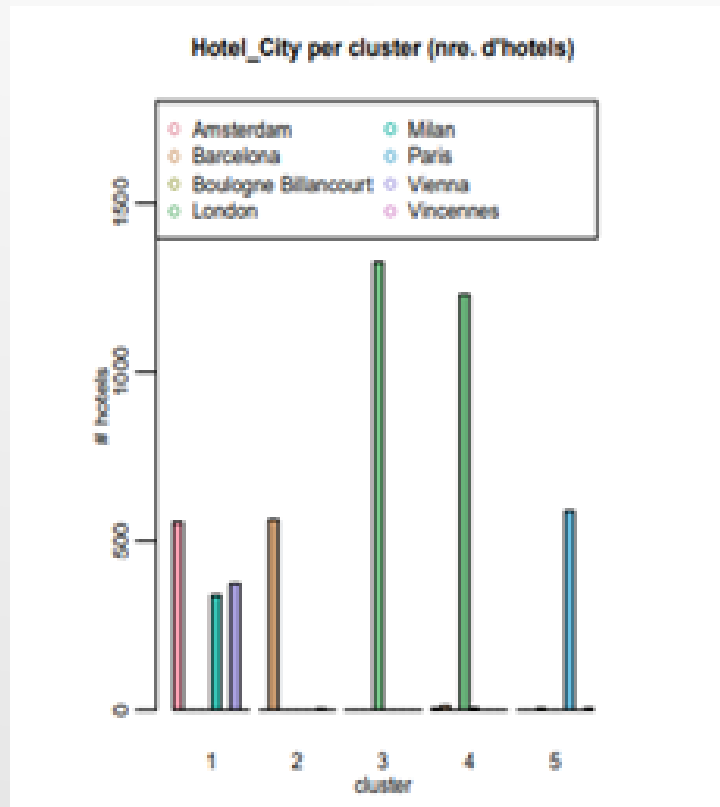
CLÚSTERING JERÀRQUIC

- ▶ Una vegada fet el preprocessing, es tracta d'agrupar conjunts d'objectes de manera que els objectes dins un grup siguin similars entre ells, i diferents en conjunt respecte als altres grups
- Panell de classes
- Dendrograma de les observacions

Clúster	Nre Individus
1	1268
2	565
3	1324
4	1248
5	595



PROFILING



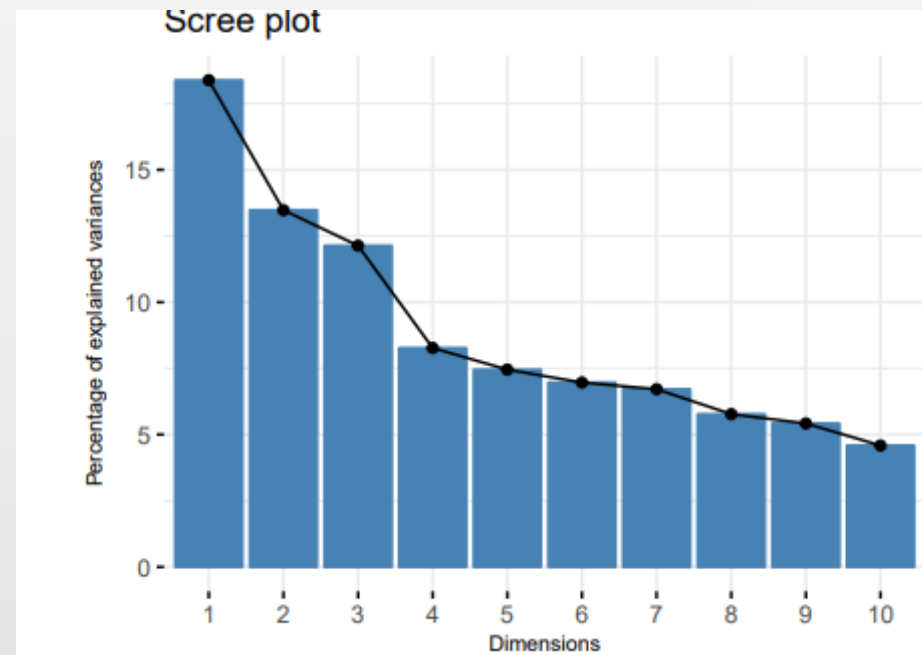
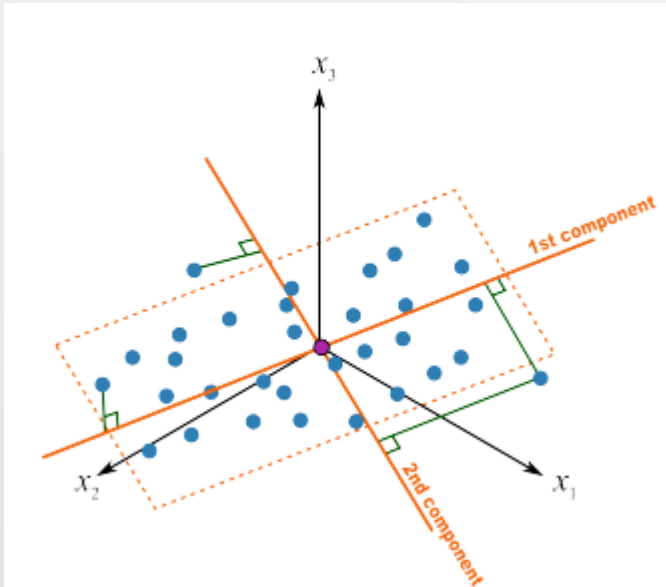
PERFILS DE LES CLASSES

Clúster	Caracterització
1	Hotels allunyats del centre de les ciutats, freqüentat per parelles i grups (suposadament els més econòmics) i amb estància de cap de setmana .
2	Hotels de Barcelona, per persones de nacionalitat anglesa, grups grans i reservats per llargues estades.
3	Hotels situats a Londres, amb valoracions baixes, referents a turisme local i són freqüentats per parelles o viatger solitaris.
4	Hotels situats a Londres, amb valoracions altes, són freqüentats per famílies i reben molts comentaris
5	Hotels amb valoracions mitjanes i poques valoracions, que es caracteritzen per estar al centre de París. Predominen famílies amb nens petits i grups (les parelles no són tan habituals)

ANÀlisi DE COMPONENTS PRINCIPALS

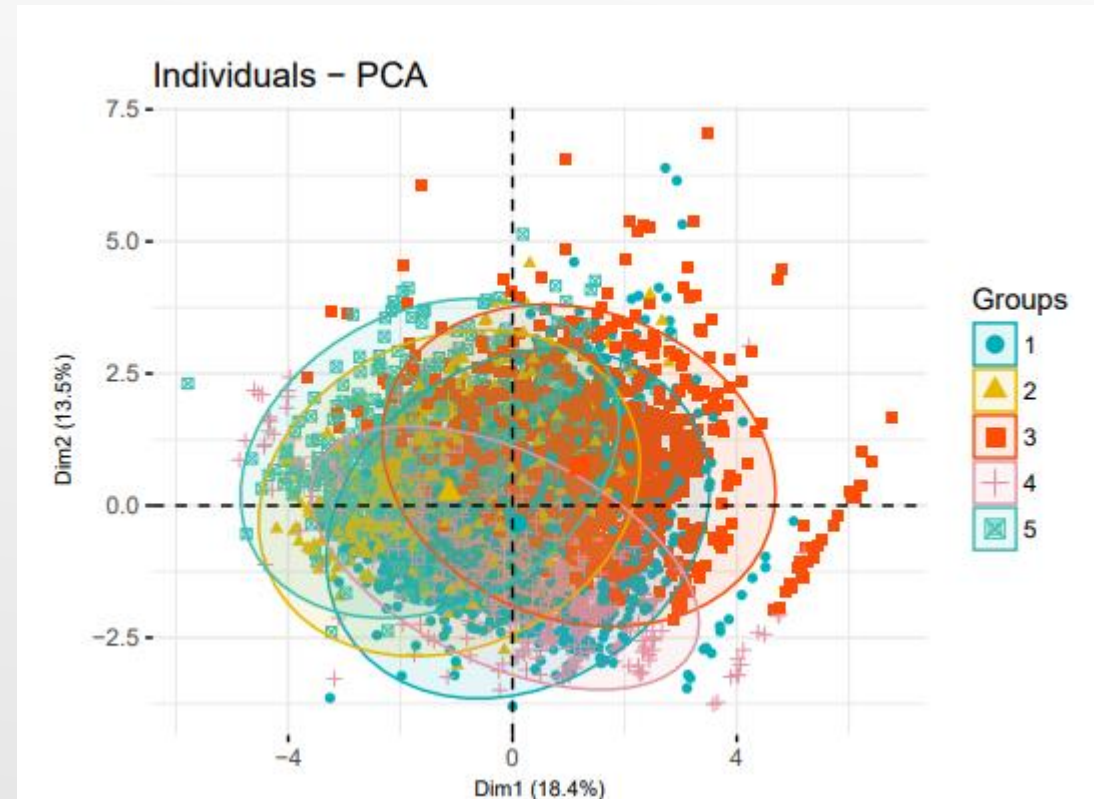
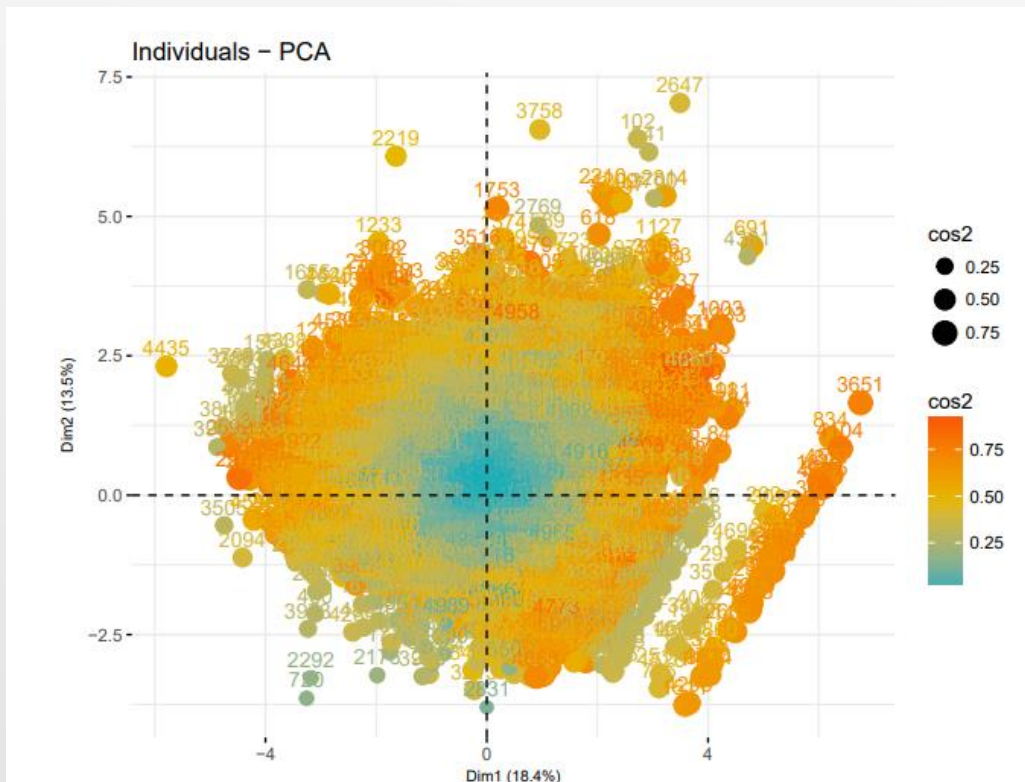
OBJECTIUS:

- ▶ Identificar patrons a les dades.
- ▶ Reduir la dimensionalitat de la base de dades original eliminant el “soroll” i les redundàncies.
- ▶ Identificar correlacions entre variables.



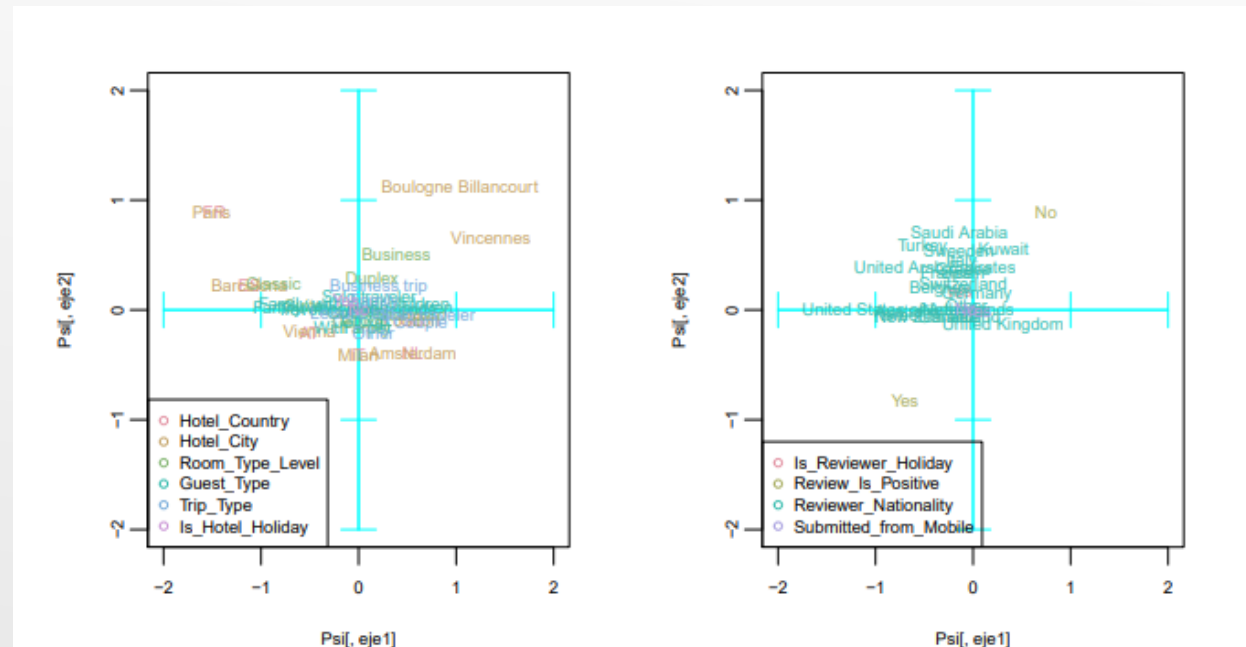
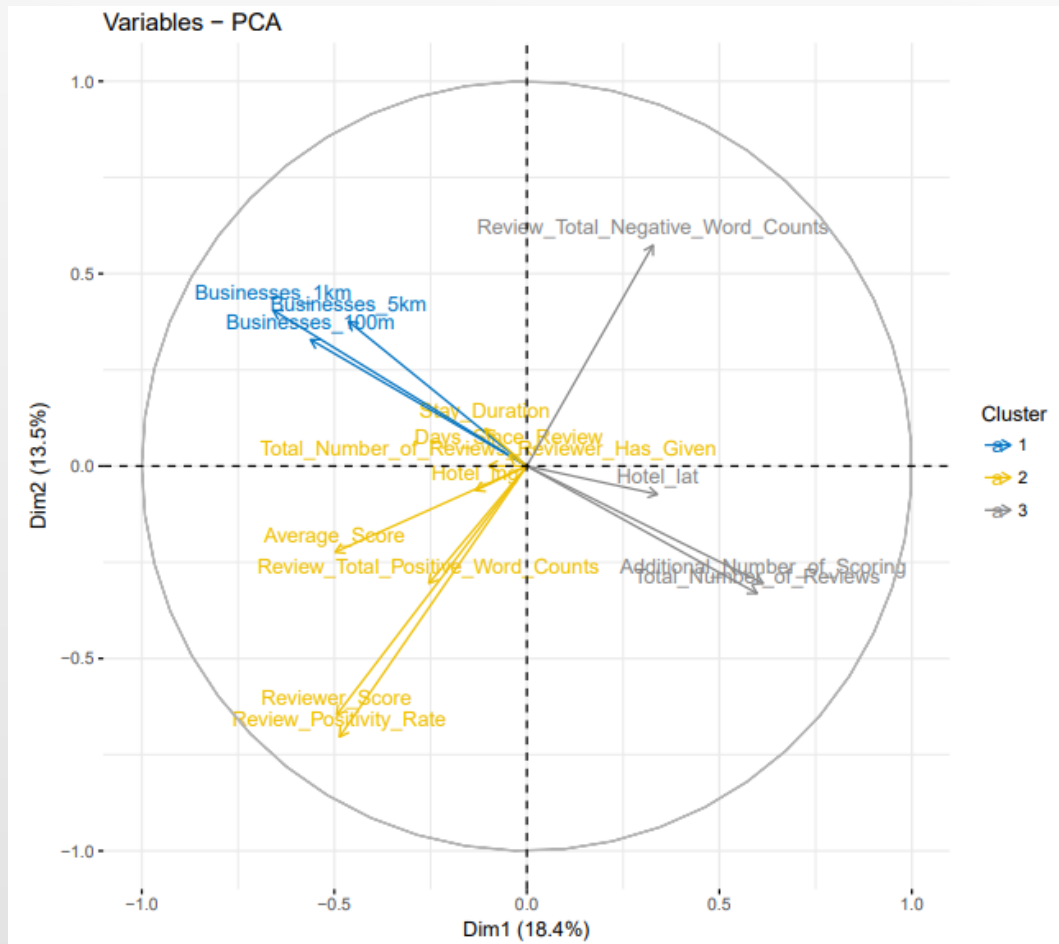
ANÀLISI DE COMPONENTS PRINCIPALS

ACP-INDIVIDUS:



ANÀLISI DE COMPONENTS PRINCIPALS

ACP-VARIABLES:



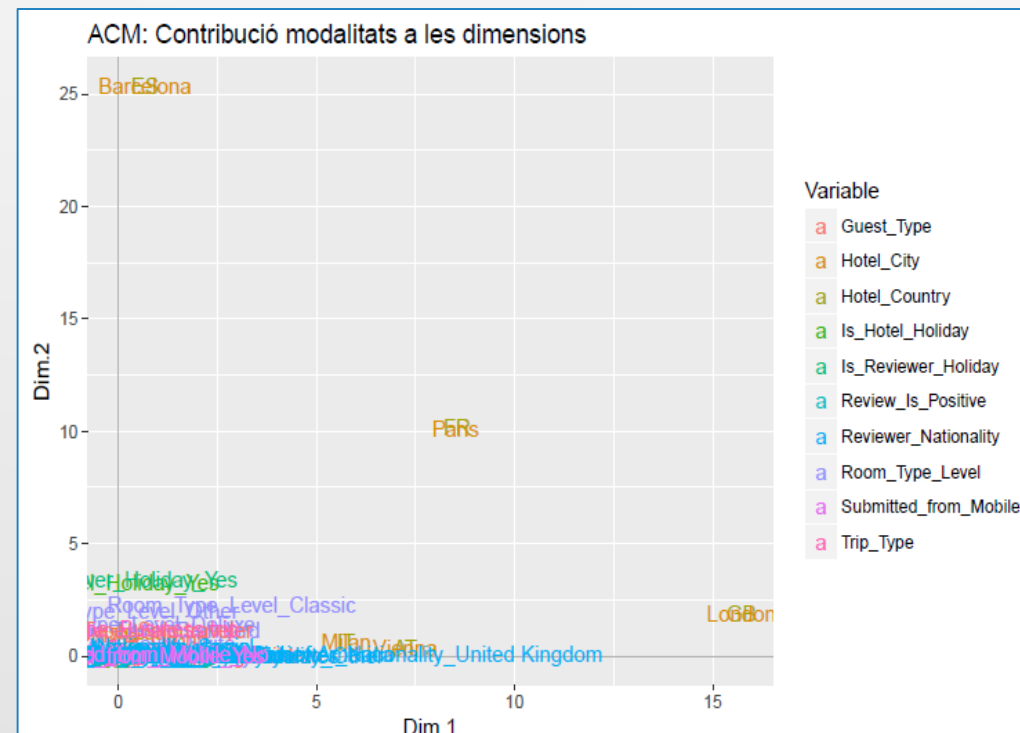
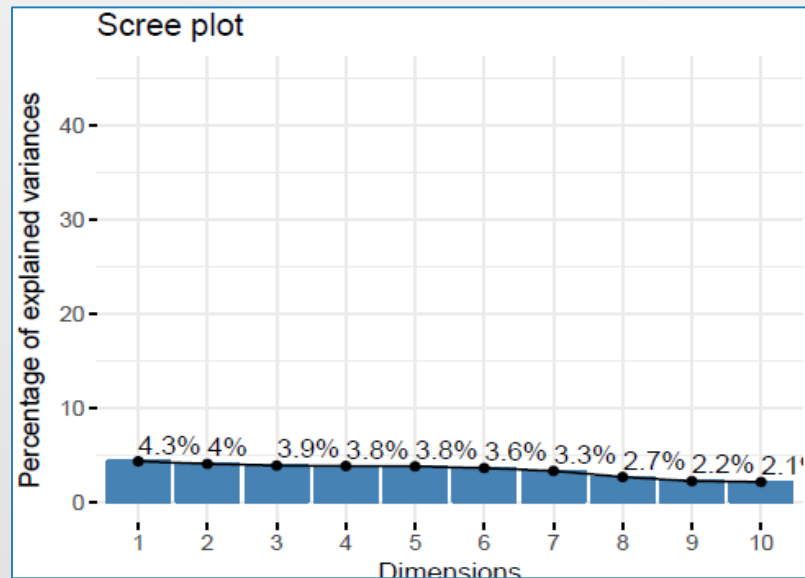
ANÀLISI DE COMPONENTS PRINCIPALS

CONCLUSIONS ACP

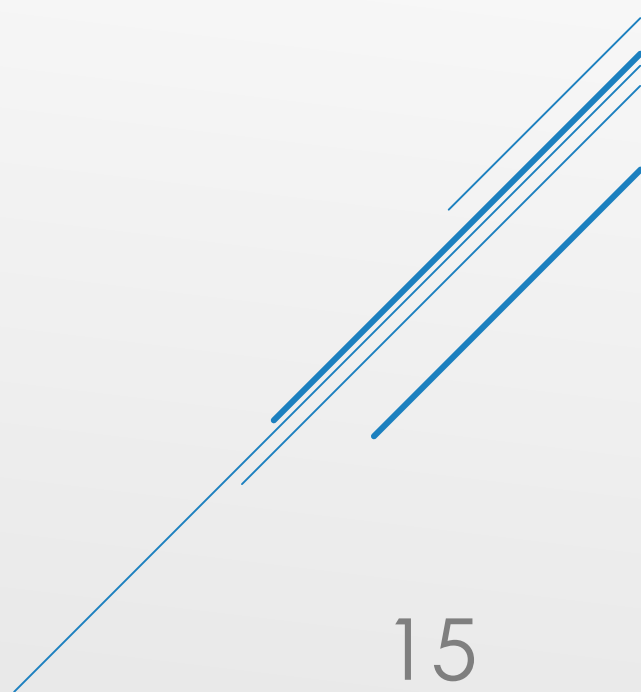
- ▶ Els hotels de França estan relacionats amb valors alts per a nombre de negocis a la rodona.
- ▶ Els clients que viatgen per negocis acostumen a puntuar pitjor els hotels i agafar hotels allunyats del centre de les ciutats.
- ▶ Barcelona i Milà solen tenir temps d'estada més llarg.

ANÀlisi CORRESPONDÈNCIES MÚLTIPLES

- ▶ Objectiu : Grups d'individus amb un perfil semblant respecte a la resposta a les variables categòriques. Associacions entre categories de les variables.
- ▶ Usem els valors propis per mirar la proporció de variància explicada

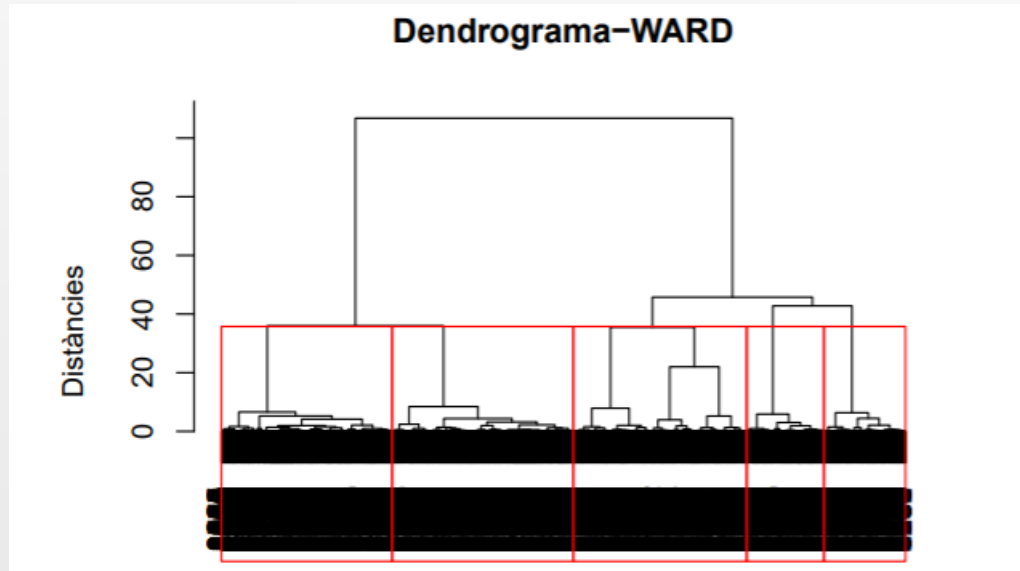


15



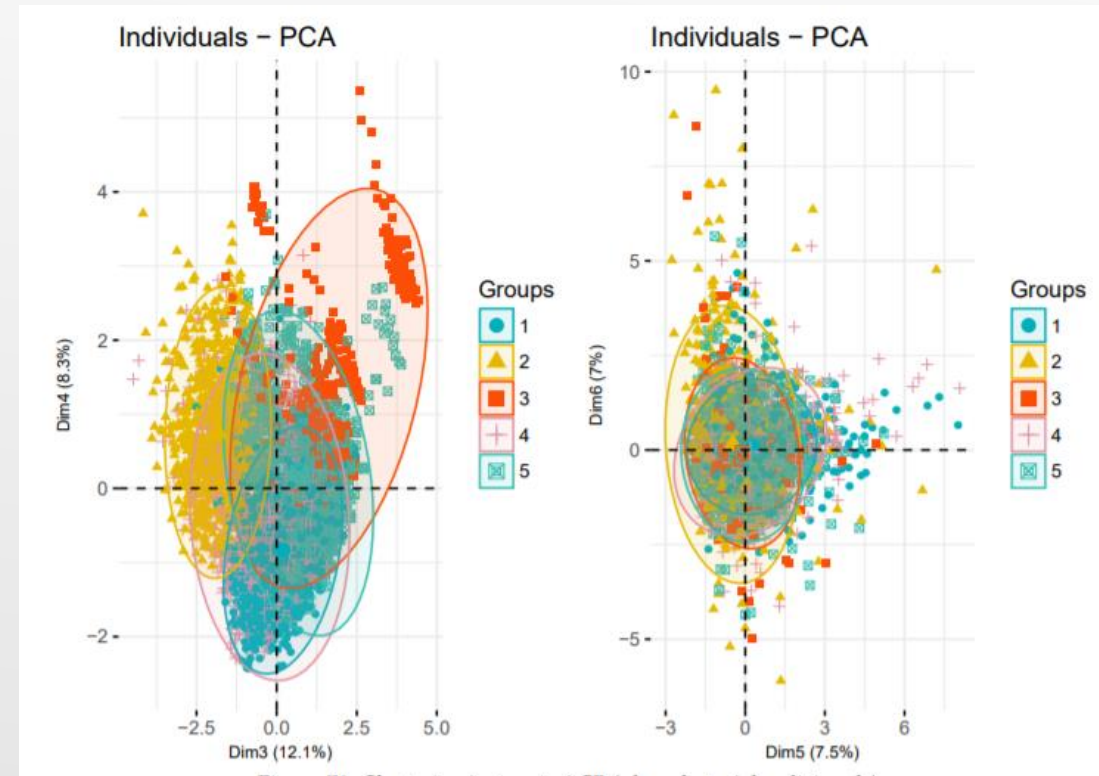
CLUSTERING JERÀRQUIC SOBRE LES COMPONENTS FACTORIALS RETINGUDES A L'ACP

Comparació de la quarta fins la sisena dimensió



Tests Chi-quadrats

	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



PROFILING DEL CLÚSTER JERÀRQUIC SOBRE ACP

Variables qualitatives

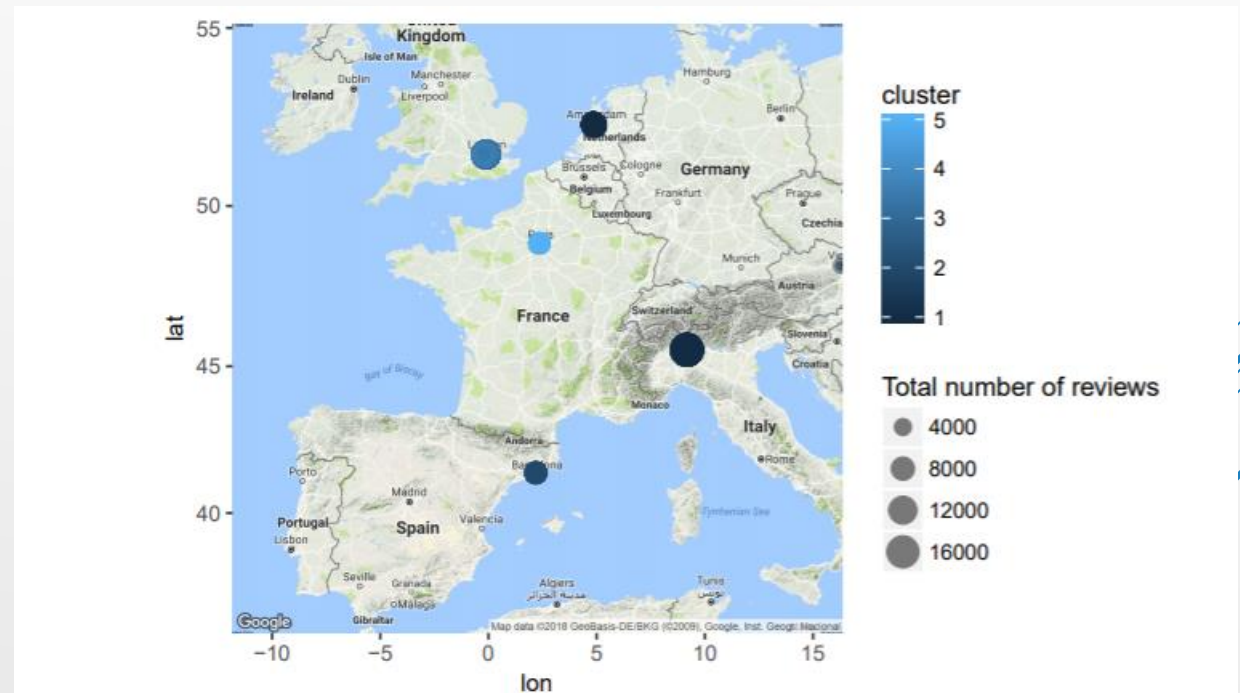


Snake plots, diagrames de
barres o taules de contingència.

Variables numèriques



Boxplots i diagrames de
barres variables.



ANÀLISI TEXTUAL

OBJECTIUS:

- ▶ Preguntes obertes, buscar motivacions o opinions dels entrevistats sobre algun assumpte.

MÈTODE:

- ▶ Amb l'ajuda de l'anàlisi de correspondències, ens permet descobrir tendències, desviacions i associacions entre individus i paraules.

ANÀLISI TEXTUAL

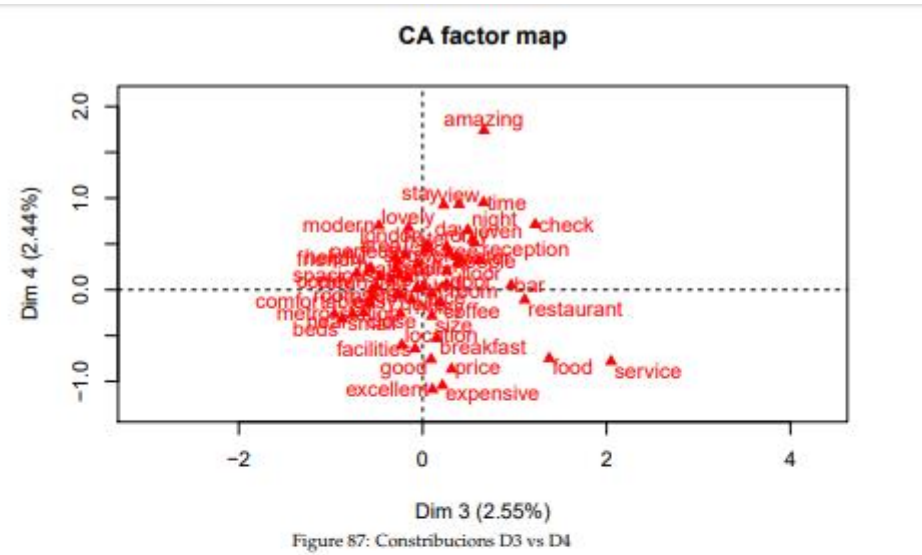
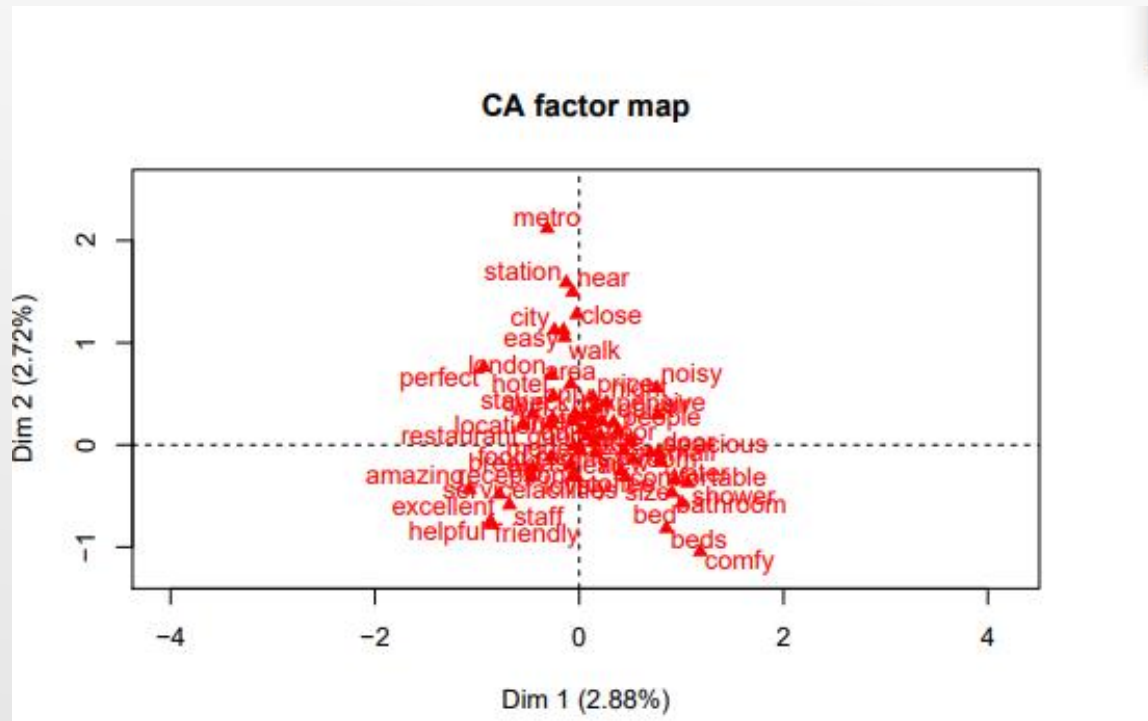


Figure 87: Contributions D3 vs D4

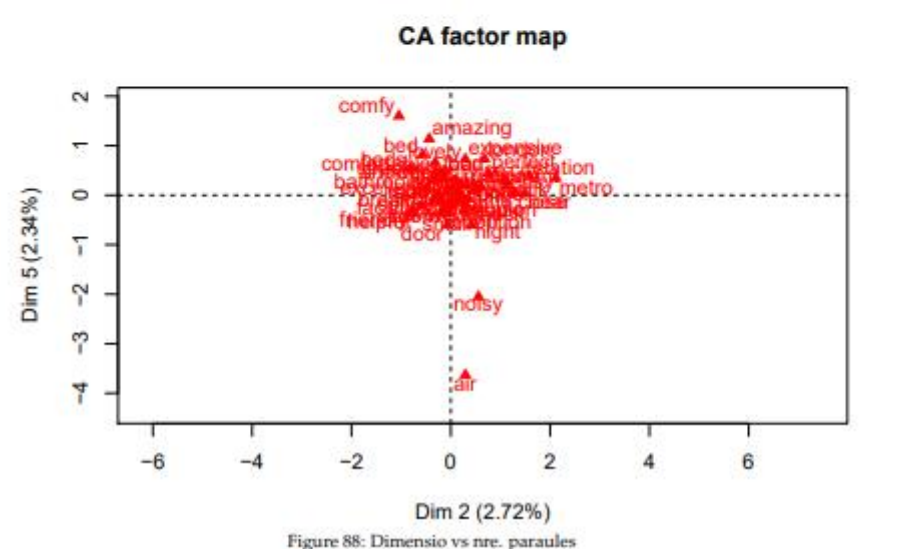


Figure 88: Dimensio vs nre. paraules

ANÀlisi COMPARATIVA I CONCLUSIONS GENERALS

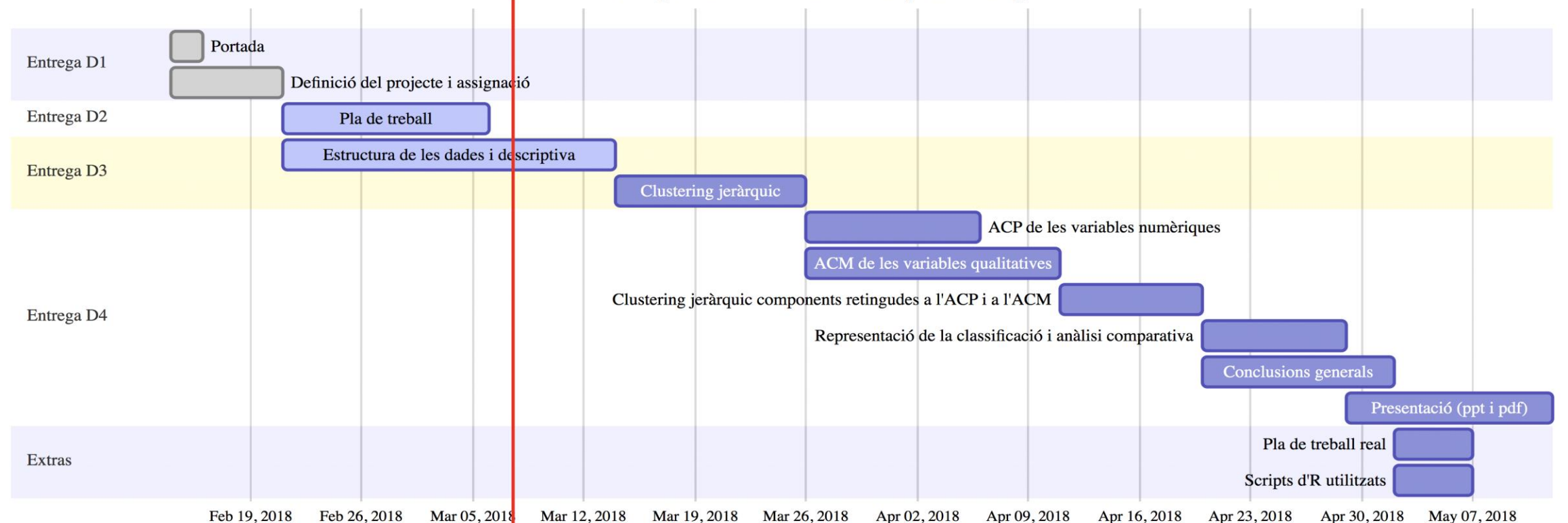
- ▶ Concordança general en els resultats al llarg de l'estudi.
- ▶ Diferències en els clústers (4 ← 3).
- ▶ Recomanacions pels usuaris de la plataforma.
 - ▶ Per a viatges en parella París és la destinació ideal.
 - ▶ Barcelona és un destí molt recomanable per a famílies amb nens.
 - ▶ Gran Bretanya i Països Baixos obtenen valoracions més baixes en relació a la resta.
- ▶ Recomanacions per als empresaris hotelers.
 - ▶ S'ha observat que les experiències negatives són les que més comentaris i valoracions generen.
 - ▶ S'ha detectat un gran mercat de usuaris descontents amb els hotels d'Amsterdam i Londres.
 - ▶ Hotels de Barcelona estades força llargues i el perfil de client és molt internacional.

ANNEX: PLANIFICACIÓ ORIGINAL

		Hugo Allès Pons	Carles Blanco Conde	Aleix Fibla Salgado	Victor Miranda Hernández	Pablo Morante López	Antoni Ramonedà Montoya	Oriol Rovira Tauler	Aleix Salvador Barrera
D1	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
	Cluster 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

ANNEX: PLANIFICACIÓ ORIGINAL

Diagrama de Gantt - Grup Booking



ANNEX: PLANIFICACIÓ FINAL

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

ANNEX: PLANIFICACIÓ FINAL

Diagrama de Gantt

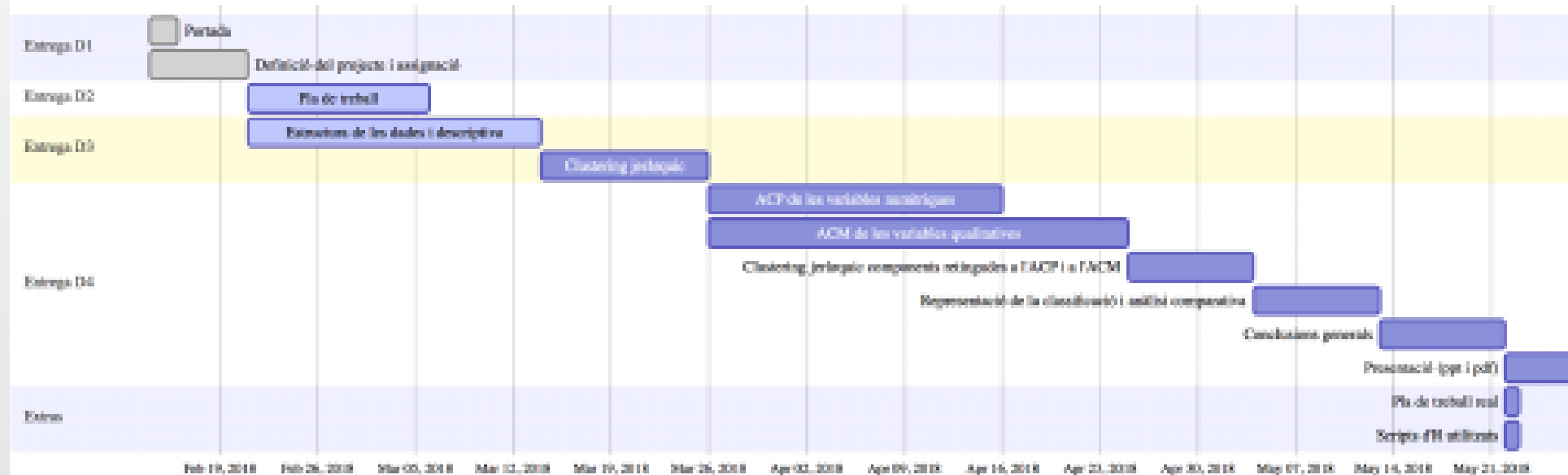


Figure 90: Diagrama de Gantt