

# Aplicació de tècniques d'anàlisi multivariant a l'estudi de l'economia familiar filipina

**Data d'entrega:** 30/05/2019

**Components del grup:**

Miquel de Jover Boira ([m.dejover@gmail.com](mailto:m.dejover@gmail.com))

Laura Julià Melis ([laaurajulia@gmail.com](mailto:laaurajulia@gmail.com))

Ferran Lacasta Roig ([ferranlacasta@gmail.com](mailto:ferranlacasta@gmail.com))

Víctor Navarro Garcés ([vng1997@gmail.com](mailto:vng1997@gmail.com))

Guiu Puigcercós Vilar ([96guiu@gmail.com](mailto:96guiu@gmail.com))

Guillem Querol Llaveria ([guillem.querolet@gmail.com](mailto:guillem.querolet@gmail.com))

Carles Requena Sánchez ([carlesrequenasanchez@gmail.com](mailto:carlesrequenasanchez@gmail.com))

# Índex

1. Definició del projecte i assignació.	2
• Font i informació sobre la base de dades.	2
• Estructura de la base de dades.	2
2. Pla de treball.	2
• Descomposició de la pràctica en tasques.	3
• Seqüenciació temporal en diagrama de Gantt.	3
• Distribució de tasques.	4
• Pla de riscos.	4
3. Estructura de les dades i descriptiva.	6
• Motivació del treball.	6
• Descripció formal de l'estructura de dades.	6
• Anàlisi descriptiva univariant inicial de totes les variables.	9
• Preprocessament de dades.	12
• Anàlisi descriptiva univariant de les dades preprocessades.	13
4. Clustering jeràrquic.	14
5. ACP de les variables numèriques.	18
6. ACM de les variables qualitatives.	20
7. Clustering jeràrquic sobre les components factorials d'ACP.	22
8. Anàlisi Factorial Múltiple.	26
10. Conclusions generals.	29
11. Pla de treball real.	30
12. Codi d'R utilitzat.	32
• Definició del projecte i assignació.	32
• Pla de treball.	32
• Anàlisis descriptives univariants (inicial i final).	32
• Preprocessament de dades.	34
• Clustering jeràrquic.	35
• Profiling.	36
• Anàlisi de Correspondències Múltiples.	37
• Clustering jeràrquic sobre les components factorials d'ACP.	37
• Anàlisi Factorial Múltiple.	38

## 1. Definició del projecte i assignació.

Aquest treball es desenvoluparà amb l'objectiu de conèixer l'activitat econòmica familiar a la República de Filipines a partir dels ingressos i les despeses així com també altres característiques rellevants. Per a assolir-ho, s'utilitzaran una sèrie de tècniques d'anàlisi multivariant.

### ● Font i informació sobre la base de dades.

La base de dades s'ha obtingut de la pàgina web *kaggle*, una comunitat online de científics de dades i *Machine learners* que permet als usuaris obtenir dades públiques de manera gratuïta. L'enllaç on es troba la base de dades és el següent: <https://www.kaggle.com/grosvenpaul/family-income-and-expenditure/home>, malgrat que les dades són propietat de la *Philippine Statistics Authority* i es poden trobar també a la seva pàgina, <http://openstat.psa.gov.ph/search>.

El conjunt de dades va ser recollit realitzant enquestes d'ingressos i despeses familiars (Family Income and Expenditure Survey, FIES) cada 3 anys, amb l'objectiu de predir els ingressos a les llars de Filipines en funció de determinades característiques i conèixer les principals fonts d'ingressos de les famílies.

A la pàgina es pot trobar la base de dades amb el nom de "Family Income and Expenditure.csv".

### ● Estructura de la base de dades.

La base de dades original conté més de 40.000 observacions i 60 variables, 45 de les quals són numèriques (nombres enters i continus, dates i codis identificadors) i 15 categòriques (dicotòmiques i politòmiques).

Per tal d'ajustar-nos a la dimensionalitat de les dades proposada a les bases del treball, s'han seleccionat aleatòriament un subconjunt de 5000 observacions fent servir R i s'han eliminat de la base de dades aquelles variables que es consideren poc rellevants en relació als objectius del treball. A més, s'han agrupat algunes columnes de manera que abans hi havia unes 10 variables en relació a les despeses en menjar (fruites, verdures, carn, etc.) i ara només n'hi ha una. D'aquesta forma, assegurem que tots els procediments requerits podran ser implementats de manera eficient i satisfactòria amb les nostres dades.

Així doncs, finalment la base de dades que utilitzarem d'ara en endavant té 5000 observacions i 34 variables, dues de les quals tenen missings (`Household.Head.Occupation` i `Household.Head.Class.of.Worker`).

## 2. Pla de treball.

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

- **Descomposició de la pràctica en tasques.**

- Entrega D1 (21/02/2019)
  - o Portada
  - o Definició del projecte i assignació
- Entrega D2 (28/02/2019)
  - o Pla de treball
- Entrega D3 (11/04/2019)
  - o Estructura de les dades i descriptiva
  - o Clustering jeràrquic
- Entrega D4 (30/5/2019)
  - o ACP de les variables numèriques
  - o ACM de les variables qualitatives
  - o Clustering jeràrquic sobre les components factorials d'ACP
  - o AFM/Discriminant/Textual
  - o Anàlisi comparativa
  - o Conclusions generals

- **Seqüenciació temporal en diagrama de Gantt.**

Aquest diagrama s'utilitza per planificar i programar tasques al llarg d'un període determinat. Gràcies a una visualització còmoda i senzilla de les accions previstes (tasques, durada, seqüència i calendari general). En funció de les diferents dates de lliuraments i les tasques a realitzar en el període establert s'ha realitzat el següent diagrama de Gantt.

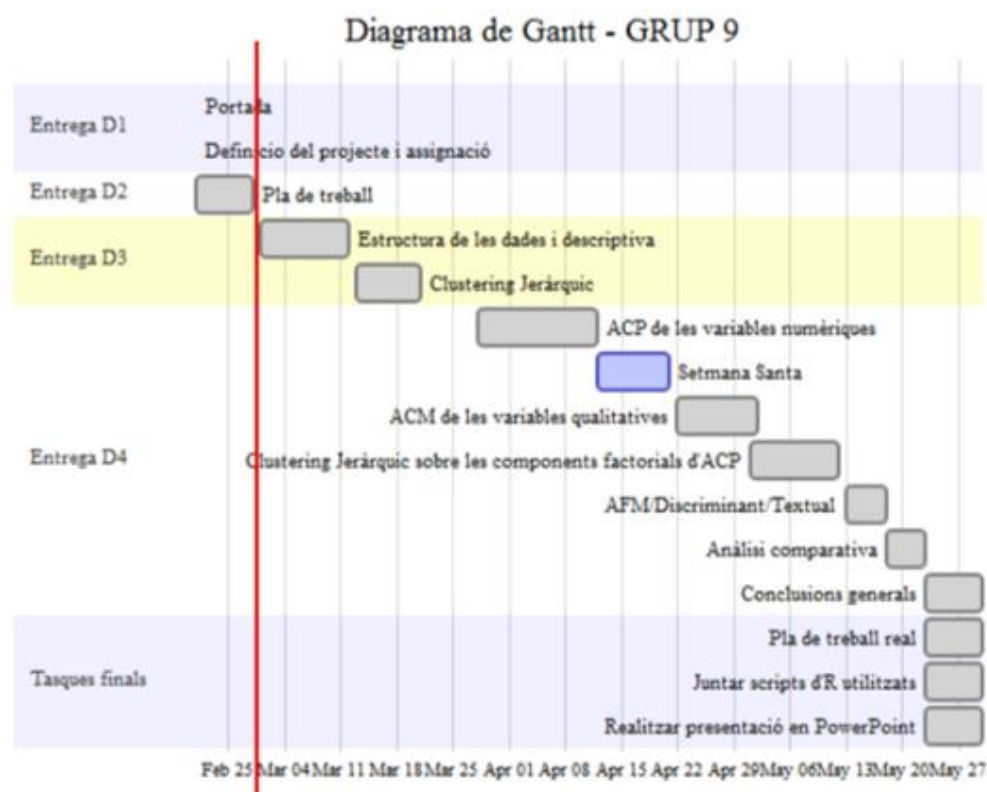


Figura 1: Diagrama de Gantt com a seqüenciació temporal de les tasques.

## ● Distribució de tasques.

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

El següent repartiment pot patir canvis futurs.

		Miquel	Laura	Ferran	Víctor	Guiu	Guillem	Carles
Entrega D1	Portada					X		
	Definició del projecte i assignació		X	X	X			X
Entrega D2	Pla de treball	X				X	X	X
Entrega D3	Estructura de les dades i descriptiva	X	X					
	Clustering Jeràrquic			X	X	X	X	X
Entrega D4	ACP de les variables numèriques	X	X		X		X	
	ACM de les variables qualitatives	X	X					X
	Clustering Jeràrquic sobre les components factorials d'ACP			X	X		X	
	AFM/Discriminant/Textual		X		X	X		X
	Anàlisi comparativa	X		X		X	X	
	Conclusions finals		X	X	X		X	X
Tasques finals	Pla de treball real	X		X		X		
	Juntar scripts d'R utilitzats	X	X	X	X	X	X	X
	Realitzar presentació en PowerPoint	X	X	X	X	X	X	X
		8	8	8	8	8	8	8

Taula 1: Repartiment de les tasques entre els membres del grup.

## ● Pla de riscos.

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

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

Possible risc	Com prevenir-lo	Com gestionar-lo
<b>Un membre del grup es posa malalt i no pot fer la seva part de la tasca.</b>	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.

<b>No entregar a temps les tasques.</b>	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.
<b>Pèrdua del treball o parts d'aquest.</b>	Compartir-ho entre tots els membres del grup en un espai virtual tots els avenços del treball (Google drive).	Assegurar-nos de que tots els components pujant els avenços al espai virtual.
<b>Entregues o scripts mal realitzats.</b>	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.
<b>Errades ortogràfiques a l'informe.</b>	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.
<b>Discussions causades per la diversitat d'opinions entre alguns o tots els membres de l'equip.</b>	Tenir una bona comunicació i respecte a l'hora de discutir els temes referits a la feina.	S'ha d'arribar a un acord amb el qual tots els membres estiguin totalment o parcialment d'acord.
<b>En el cas que es produeixi una vaga d'estudiants o de transport.</b>	Ser previsors, anticipar-nos a aquesta situació i deixar avançada la tasca a realitzar el dia en qüestió.	Haurem de gestionar la feina els dies posteriors, per tal de no quedar endarrerits.
<b>Un membre de l'equip decideix renunciar a l'avaluació contínua.</b>	Totes les tasques tenen almenys dos membres de l'equip assignat.	Haurem de repartir les tasques assignades a la persona en qüestió entre els membres restants.
<b>Algun integrant del grup no faci la seva tasca o no posi interès en el treball.</b>	Donar-li tocs d'atenció.	Al final del treball fer una avaluació justa d'aquest membre de l'equip.
<b>Un membre de l'equip no sap fer la seva part del treball.</b>	Quan un membre de l'equip tingui dubtes ho comunicarà a la resta abans que sigui massa tard.	Si es disposa d'algun integrant capaç de dur a terme aquesta tasca se li demanarà col·laboració amb la persona corresponent.
<b>No es respecta el calendari de coordinadors i algú agafa el control contínuament.</b>	Cadascú ha de ser respectuós amb les pautes marcades i intentar no agafar el control quan no li correspon.	El grup haurà de reunir-se i parlar d'aquest problema fins que es solucioni.
<b>El campus virtual no funciona o en el moment d'entrega està en manteniment.</b>	Entregar les tasques amb els màxim dies d'antelació possibles.	Enviar la tasca a través del correu electrònic del professor.

*Taula 2: Definició, solució i prevenció de possibles riscos.*

### 3. Estructura de les dades i 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 d'una descripció detallada de la base de dades (descripció i anàlisi descriptiva univariant). Finalment, es plantegen tots els procediments que s'han dut a terme en la fase de preprocessament de les dades, així com un anàlisi exploratòria inicial de cadascuna de les variables.

#### ● Motivació del treball.

Escollir el tema del treball no ha estat una tasca fàcil ja que tots els integrants del grup hem anat trobant bases de dades interessants a diferents pàgines web però que per diversos motius hem hagut d'anar descartant: tema semblant a algun treball passat, dificultat per trobar relacions entre les variables disponibles que ens permetessin fer l'estudi, etc. Finalment, el conjunt de dades sobre l'economia familiar a Filipines complia tots els requisits alhora que ens ofería un tema que satisfecia l'interès de tots els components.

Així doncs, vam decidir aplicar les tècniques d'anàlisi multivariant que estem aprenent en la present assignatura per tal de fer un estudi sobre l'activitat econòmica familiar a Filipines amb l'objectiu d'establir un model (basat en el nombre de membres, la regió del país o la feina, entre d'altres variables d'interès) que ens permeti detectar característiques associades a uns ingressos i/o despeses elevades.

#### ● Descripció formal de l'estructura de dades.

Tal i com s'ha mencionat anteriorment, tot l'anàlisi es duu a terme amb una mostra aleatòria de 5.000 observacions  $\{seed(2019)\}$  i una selecció de 34 variables que es consideren rellevants per als objectius perseguits (base de dades original més de 40.000 files i 60 columnes). En aquest sentit, cada observació de la matriu de dades representa una família filipina.

Tot seguit, es presenta el llistat de les variables seleccionades amb la corresponent metainformació:

1. **Id:** Identificació de cada família (quantitativa).
  - a. Rang de valors:  $\{1, 5000\}$
  - b. Rol: variable identificadora (*primary key*)
2. **Total.Household.Income:** ingressos totals de la llar (quantitativa).
  - a. Rang de valors:  $\{11285, 11639365\}$
  - b. Rol: variable resposta.
3. **Region:** Regió de Filipines (qualitativa politòmica).
  - a. Modalitats (17):  $\{“ARMM”, “CAR”, “Caraga”, “NCR”, “I - Ilocos Region”, “II - Cagayan Valley”, “III - Central Luzon”, “IVA - CALABARZON”, “IVB - MIMAROPA”, “V - Bicol Region”, “VI - Western Visayas”, “VII - Central Visayas”, “VIII - Eastern Visayas”, “IX - Zasmboanga Peninsula”, “X - Northern Mindanao”, “XI - Davao Region”, “XII - SOCCSARGEN”\}$
  - b. Rol: variable explicativa
4. **Total.Food.Expenditure:** despesa total en menjar (quantitativa, entera).
  - a. Rang de valors:  $\{2947, 603187\}$
  - b. Rol: variable explicativa.
5. **Main.Source.of.Income:** principal font d'ingressos (qualitativa politòmica).

- a. Modalitats (3): {"Entrepreneurial Activities", "Wage/Salaries", "Other Sources of Income"}
  - b. Rol: variable explicativa
- 6. **Agricultural.Household.indicator:** indicador de si la llar és agrícola (qualitativa dicotòmica).
  - a. Modalitats (3): {"0" = NS/NC, "1" = SI, "2" = NO}
  - b. Rol: variable explicativa
- 7. **Restaurant.and.hotels.Expenditure:** despeses en restaurants i hotels (quantitativa).
  - a. Rang de valors: {0, 383650}
  - b. Rol: variable explicativa.
- 8. **Alcohol.and.tobacco.Expenditure:** despeses en alcohol i tabac (quantitativa).
  - a. Rang de valors: {0, 66730}
  - b. Rol: variable explicativa.
- 9. **Clothing..Footwear.and.Other.Wear.Expenditure:** despeses en roba i sabates (quantitativa).
  - a. Rang de valors: {0, 94635}
  - b. Rol: variable explicativa.
- 10. **Housing.and.water.Expenditure:** despeses de vivenda i d'aigua (quantitativa).
  - a. Rang de valors: {2400, 1458300}
  - b. Rol: variable explicativa.
- 11. **Imputed.House.Rental.Value:** valor imputat del lloguer de la casa (quantitativa).
  - a. Rang de valors: {0, 1200000}
  - b. Rol: variable explicativa.
- 12. **Medical.education.transport.and.communication.Expenditure:** despeses en metges, educació, transports i comunicació (quantitativa).
  - a. Rang de valors: {0, 717039}
  - b. Rol: variable explicativa.
- 13. **Miscellaneous.goods.and.special.occasions.expenditure:** despeses en béns diversos i ocasions especials (quantitativa).
  - a. Rang de valors: {18, 542444}
  - b. Rol: variable explicativa.
- 14. **Crop.Farming.and.Gardening.expenses:** despeses en cultiu i jardineria (quantitativa).
  - a. Rang de valors: {0, 510935}
  - b. Rol: variable explicativa.
- 15. **Total.Income.from.Entrepreneurial.Activities:** Ingressos totals per activitats empresarials (quantitativa).
  - a. Rang de valors: {0, 2281500}
  - b. Rol: variable explicativa.
- 16. **Household.Head.Sex:** gènere del cap de la llar (qualitativa).
  - a. Modalitats (2): {"Female", "Male"}
  - b. Rol: variable explicativa
- 17. **Household.Head.Age:** edat del cap de la llar (quantitativa).
  - a. Rang de valors: {15, 96}
  - b. Rol: variable explicativa.
- 18. **Household.Head.Marital.Status:** estat civil del cap de la llar (qualitativa).
  - a. Modalitats (6): {"Married", "Widowed", "Single", "Divorced/Separated", "Annulled", "Unknow"}
  - b. Rol: variable explicativa
- 19. **Household.Head.Highest.Grade.Completed[1] [2] [3] :** grau més alt completat del cap de la llar (qualitativa).
  - a. Modalitats (4): {"High School Studies", "Elementary Studies", "No Studies", "Program Studies"}
  - b. Rol: variable explicativa



20. **Household.Head.Job.or.Business.Indicator:** indicador de si el cap de la llar treballa (qualitativa).
  - a. Modalitats (2): {"No Job/Business", "With Job/Business"}
  - b. Rol: variable explicativa
21. **Household.Head.Occupation[4]** : ocupació del cap de la llar (qualitativa).
  - a. Modalitats (3): {"Primary Sector", "Secondary Sector", "Tertiary Sector"}
  - b. Rol: variable explicativa
  - c. Imputació de missings: "NA"
22. **Household.Head.Class.of.Worker:** tipus de feina del cap de la llar (qualitativa).
  - a. Modalitats (7): {"Employer in own family-operated farm or business", "Self-employed without any employee", "Worked for government/government corporation", "Worked for private establishment", "Worked for private household", "Worked with pay in own family-operated farm or business", "Worked without pay in own family-operated farm or business"}
  - b. Rol: variable explicativa
  - c. Imputació de missings: "NA"
23. **Total.Number.of.Family.members:** nombre total de membres a la família (quantitativa).
  - a. Rang de valors: {1, 15}
  - b. Rol: variable explicativa.
24. **Members.with.age.less.than.5.year.old:** membres de la família menors de 5 anys (quantitativa).
  - a. Rang de valors: {0, 5}
  - b. Rol: variable explicativa.
25. **Members.with.age.5...17.years.old:** membres de la família entre 5 i 17 anys (quantitativa).
  - a. Rang de valors: {0, 8}
  - b. Rol: variable explicativa.
26. **Total.number.of.family.members.employed:** nombre total de membres de la família que treballen (quantitativa).
  - a. Rang de valors: {0, 8}
  - b. Rol: variable explicativa.
27. **Type.of.Building.House:** tipus de casa (qualitativa politòmica).
  - a. Modalitats (5): {"Commercial/industrial/agricultural building", "Duplex", "Institutional living quarter", "Multi-unit residential", "Single house"}
  - b. Rol: variable explicativa
28. **House.Floor.Area:** superfície de la casa (quantitativa).
  - a. Rang de valors: {5, 720}
  - b. Rol: variable explicativa.
29. **House.Age:** antiguitat de la casa en anys (quantitativa).
  - a. Rang de valors: {0, 115}
  - b. Rol: variable explicativa.
30. **Number.of.bedrooms:** nombre d'habitacions (quantitativa).
  - a. Rang de valors: {0, 9}
  - b. Rol: variable explicativa.
31. **Electricity:** accés a electricitat (qualitativa dicotòmica).
  - a. Modalitats (5): {"1" = SI, "0" = NO}
  - b. Rol: variable explicativa.
32. **Number.of.Car..Jeep..Van:** nombre d'automòbils(quantitativa).
  - a. Rang de valors: {0, 5}
  - b. Rol: variable explicativa.
33. **Number.of.Cellular.phone:** nombre de telèfons mòbils (quantitativa).
  - a. Rang de valors: {0, 10}
  - b. Rol: variable explicativa.

34. **Number.of.Motorcycle.Tricycle:** nombre de motos (quantitativa).

- Rang de valors: {0, 5}
- Rol: variable explicativa.

## ● Anàlisi descriptiva univariant inicial de totes les variables.

En aquest apartat s'adjunten dues taules que recullen la informació numèrica de cada variable (tant per a les quantitatives com per les qualitatives) així com també alguns gràfics per tal d'observar d'una manera més gràfica les característiques més rellevants de les variables que calguin.

### 1. Variables Numèriques (23):

A partir de la següent taula (*Taula 3*) és possible fer-se una idea de la distribució de cada variable.

Variable	Min	1st Qu.	Median	Mean.	3rd Qu	Max	Desv. Tip.
Total.Household.Income	11289	104100	162800	243800	294299.9	11640000	297478.5
Total.Food.Expenditure	2947	51090	72630	84600	104000	603200	50939.81
Agricultural.Household.indicator	0	0	0	0.44	1	2	0.67
Restaurant.and.hotels.Expenditure	0	1820	7282	15470	19880	383600	23996.29
Alcohol.and.tobacco.expenditure	0	0	1440	3440	4754	66730	5164.20
ClothingFootwear.and.Other.Wear.Expenditure	0	1360	2750	4873	5481	94640	6774.18
Housing.and.water.Expenditure	2400	13090	22800	38720	45460	1458000	59337.49
Imputed.House.Rental.Value	0	5700	10800	21110	24000	1200000	44140.98
Medical.education.transport.and.communication.expenditure	0	5774	13850	30070	34920	717000	49276.02
Miscellaneous.goods.and.special.occasions.expenditure	18	4815	9310	17330	20040	542400	24537.87
Crop.Farming.and.Gardening.expenses	0	0	0	13750	6555	510900	37413.65
Total.Income.from.Entrepreneurial.Acitivites	0	0	18070	50490	65000	2282000	101458.30
Household.Head.Age	15.00	41.00	50.50	51.28	61.00	96.00	14.00
Total.Number.of.Family.members	1.00	3.00	4.00	4.63	6.00	15.00	2.25
Members.with.age.less.than.5.year.old	0	0	0	0.41	1.00	5.00	0.69
Members.with.age.5...17.years.old	0	0	1.00	1.35	2.00	8.00	1.41
Total.number.of.family.members.employed	0	0	1.00	1.30	2.00	8.00	1.15
House.Floor.Area	5.00	25.00	40.00	55.04	70.00	720.00	52.41
House.Age	0	10.00	17.00	20.05	27.00	115.00	14.33
Number.of.bedrooms	0	1.00	2.00	1.79	2.00	9.00	1.10
Number.of.Car..Jeep..Van	0	0	0	0.08	0	5.00	0.34
Number.of.Cellular.phone	0	1.00	2.00	1.92	3.00	10.00	1.54
Number.of.Motorcycle.Tricycle"	0	0	0	0.29	0	5.00	0.56

*Taula 3: Resum numèric per a les variables numèriques.*

En general, s'observen grans diferències entre els valors màxim i mínim de les variables. També desviacions tipus força grans, per el que sembla haver-hi bastanta variabilitat en les característiques estudiades entre les famílies filipines.

Per exemple, “Total.Household.Income” hi ha una serie d’outliers que afecten a la distribució de manera que la mediana és molt més baixa que la mitjana. Això també ocorre en altres variables com a “Total.Food.Expenditure”, de la que es mostra un gràfic de freqüències a continuació, a partir del qual s’observa que la distribució de la despesa total en alimentació és asimètrica cap a l’esquerra: poques famílies gasten una quantitat molt alta en menjar.

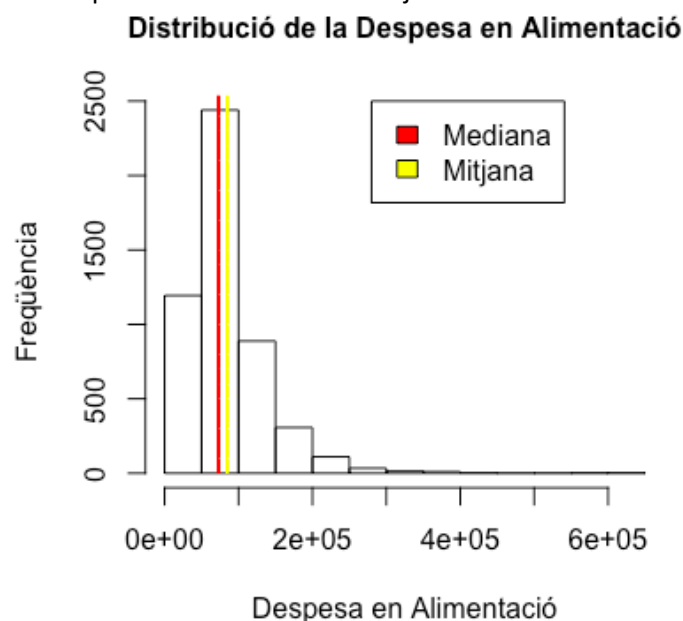
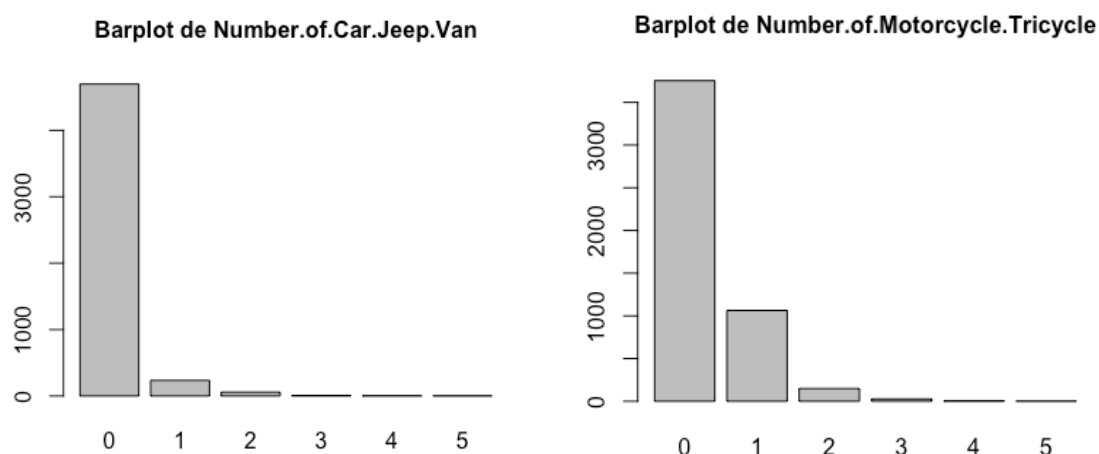


Figura 2: Histograma de la variable **Total.Food.Expenditure**.

També crida l’atenció que més d’un 75 % de les famílies no disposen de cap vehicle motoritzat (de 2 o 4 rodes, sigui del tipus que sigui) mentre que algunes poques n’arriben a tenir fins a un màxim de 5.



Figures 3 i 4: Gràfics de barres de **Number.of.Car.Jeep.Van** (esquerra) i de **Number.of.Motorcycle.Tricycle** (dreta).

## 2. Variables Categòriques(10):

Variable	Nivells	Freq	%	Nivells	Freq	%
Region	ARMM	280	5.60	NCR	497	9.94
	CAR	189	3.78	V - Bicol Region	328	6.56
	CARAGA	218	4.36	VI - Western Visayas	378	7.56
	I - Ilocos Region	267	5.34	VII - Central Visayas	308	6.16
	II - Cagayan Valley	258	5.16	VIII - Eastern Visayas	304	6.08
	III - Central Luzon	375	7.50	X - Northern Mindanao	201	4.02

	IVA - CALABARZON	460	9.20	XI - Davao Region	297	5.94
	IVB - MIMAROPA	163	3.26	XII - SOCCSKSARGEN	263	5.26
	IX - Zasmboanga Peninsula	214	4.28			
Main.Source.of.Incme	Entrepreneurial Activities	1219	23.38	Other sources of Income	1306	26.12
	Wage/Salaries	2475	49.50			
Household.Head.Sex	Male	3933	78.66	Female	1067	21.34
Household.Head.Marital.Stat us	Annulled	2	0.04	Divorced/Separated	159	3.18
	Married	3787	75.74	Single	243	4.86
	Unkown	1	0.02	Widowed	808	16.16
Household.Head.Highest.Gra de.Completed	243 nivells					
Household.Head.Job.or.Busin ess.Indicator	Sense treball	874	17.48	Amb treball	4126	82.52
Household.Head.Occupation	46 nivells					
Household.Head.Class.of.Wo rker	Employer in own family- operated farm or business	305	6.10	Self-employed without any employee	1669	33.38
	Unemployed	874	17.48	Worked for government/government corporation	360	7.20
	Worked for private establishment	1675	33.50	Worked for private household	88	1.76
	Worked with pay in own family- operated farm or business	2	0.04	Worked without pay in own family-operated farm or business	27	0.54
Type.of.Building.House	Commercial/industrial/agricultu ral building	6	0.12	Duplex	139	2.78
	Institutional living quarter	1	0.02	Multi-unit residential	158	3.16
	Single house	4696	93.92			
Electricity	NO (0)	596	11.92	SI (1)	4404	88.08

Taula 4: Resum numèric per a les variables categòriques.

El primer que s'observa és que hi ha dues variables amb massa nivells, les que recullen la ocupació i el nivell d'estudis màxim del cap de família. També es veuen variables amb una distribució prou heterogènia com es el cas de "Region", en la qual observem unes freqüències relatives entre el 3 i el 10% aproximadament per a cadascuna de les regions. D'altres en canvi, tenen algun nivell molt majoritari que la resta, com "Electricity", que ens indica que un 88% de les famílies disposen d'electricitat:

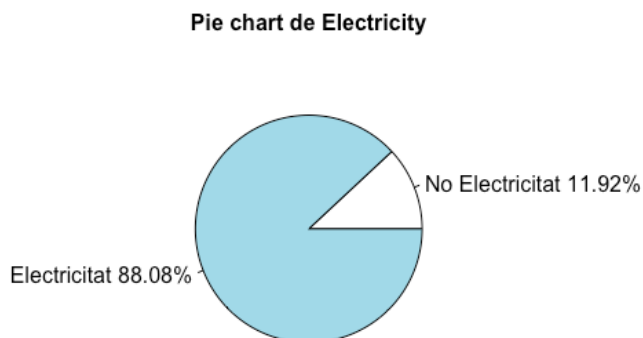


Figura 5: Diagrama circular de la variable **Electricity**.

També és digne de menció que el 79% dels caps de família de la nostra base de dades són homes. I, més d'un 70% de vegades, és una persones casada i amb ocupació.

## ● Preprocessament de dades.

En el pre-processament s'ha tractat la base de dades resultant dels passos de selecció de variables inicial a partir del primer arxiu en brut. La intenció d'aquest procés ha estat arribar a tenir unes dades les quals siguin prou refinades per a poder aplicar-les-hi un tractament d'anàlisi multivariant satisfactòriament, començant pel *clustering* i *profiling*.

S'ha començat per assegurar que totes les variables estaven definides en la classe corresponent per la seva definició. La variable "Agricultural.Household.indicator" sortia com a numèrica i s'ha categoritzat com a factor, ja que és un indicador que pren valors categòrics: {0,1,2}.

A continuació s'han modelitzat les variables categòriques "Household.Head.Highest.Grade.Completed" i "Household.Head.Occupation" ja que tenien 46 i 243 nivells respectivament, fet que les feia impossibles d'analitzar.

- I. La variable "Household.Head.Highest.Grade.Completed" s'ha modelitzat en 4 nous nivells segons la major etapa d'educació que hagin completat: {"High School Studies", "Program Studies", "Elementary School Studies", "No Studies"}.
- II. Pel que fa a "Household.Head.Occupation", s'han modelitzat 3 nivells nous els quals engloben les ocupacions en els 3 sectors de treball: {"Primary Sector", "Secondary Sector", "Tertiary Sector"}. També s'ha assignat "NA" a aquells casos on l'ofici no estigués reportat.

S'ha finalitzat el *pre-processing* amb el tractament de *missings*. S'ha vist anteriorment en aquest document que tots els missings es troben a les variables "Household.Head.Occupation" i "Household.Head.Class.of.Worker" en les mateixes 874 files.

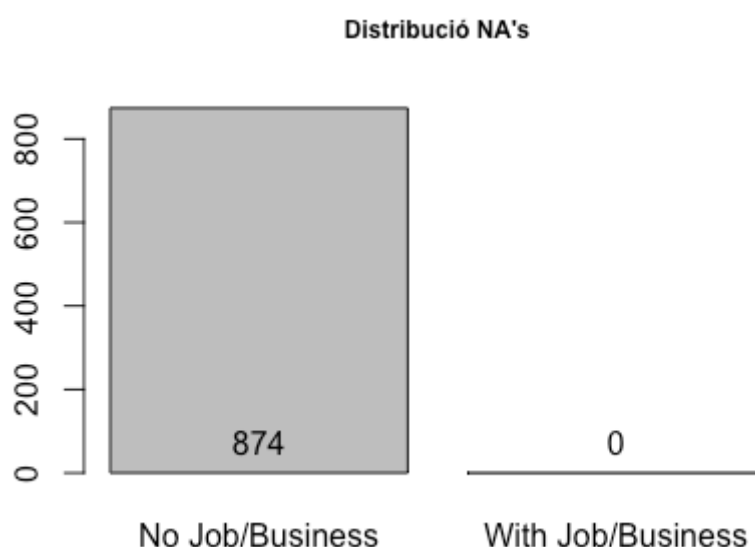


Figura 6: Distribució dels missings en la variable NA's.

Utilitzant comandes per trobar la ubicació dels *missings* s'ha descobert que tots ells es corresponen al nivell de la variable Household.Head.Job.or.Business.Indicator: "No Job / Business" (veure Figura 6) indicant que la totalitat dels missings en les variables sobre el treball són degudes a que la persona es troba a l'atur.

En vista d'això, s'ha decidit definir els *missings* de "Household.Head.Occupation" com a "No Occupation" i els *missings* de "Household.Head.Job.or.Business.Indicator" com a "Unemployed". D'aquesta manera es disposa d'una base de dades pre-processada amb zero missings.

### ● Anàlisi descriptiva univariant de les dades preprocessades.

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

Quant a les variables numèriques, no és necessari repetir res ja que ninguna ha sofert canvis llevat de "Agricultural.Household.indicator", la qual ha passat a ser considerada categòrica i per tant s'analitzarà juntament amb la resta de variables qualitatives.

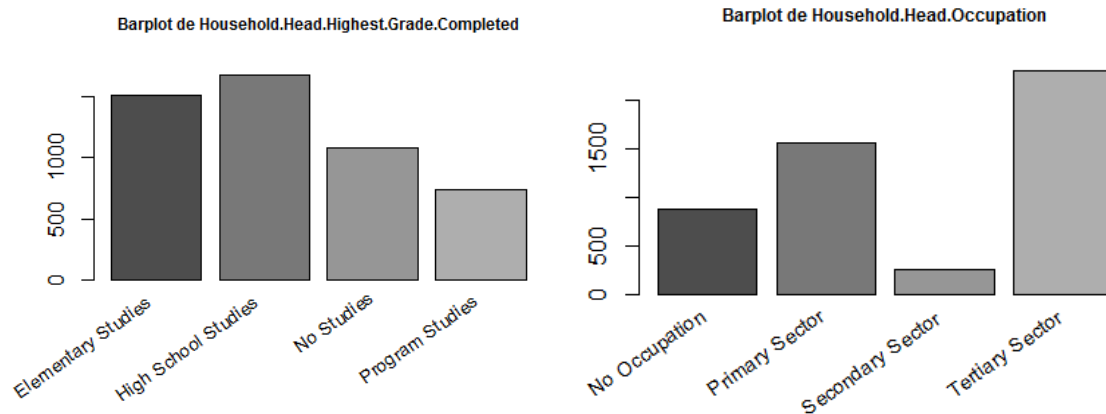
#### - Variables Categòriques (ara 11):

En la taula adjuntada sota aquestes línies es pot veure un resum numèric de les quatre variables que han patit canvis en el *preprocessing*. Ara ja és possible analitzar aquelles variables que abans tenien tants nivells ("Household.Head.Highest. Grade.Completed" i "Household.Head.Occupation").

Variable	Nivells	Freq	%	Nivells	Freq	%
Agricultural.Household.indicator	NS/NC (0)	3328	66.56	SI (1)	1150	23.00
	NO (2)	522	10.44			
Household.Head.Highest. Grade.Completed	Elementary studies	1511	30.22	High School Studies	1673	33.46
	No Studies	1079	21.58	Program Studies	737	14.74
Household.Head.Occupation	No Occupation	874	17.48	Primary Sector	1570	32.40
	Secondary Sector	255	5.10	Tertiary Sector	2301	46.02
Household.Head.Class.of. Worker	Employer in own family-operated farm or business	305	6.10	Self-employed without any employee	1669	33.38
	Unemployed	874	17.48	Worked for government /government corporation	360	7.20
	Worked for private establishment	1675	33.50	Worked for private household	88	1.76
	Worked with pay in own family-operated farm or business	2	0.04	Worked without pay in own family-operated farm or business	27	0.54

Taula 5: Resum numèric de les dades preprocessades per a les variables categòriques.

Amb aquest nou anàlisi, veiem que les persones que han volgut respondre si la seva llar és agrícola és menys d'un 40% del total de famílies. A més, s'observa una minoria dels caps de família amb estudis superiors (*Program Studies*) i que quasi la meitat d'aquests treballen al sector terciari.



Figures 7 i 8: Gràfics de **Household.Head.Highest.Grade.Completed** (esquerra) i de **Household.Head.Occupation** (dreta).

## 4. Clustering jeràrquic.

L'objectiu en aquest capítol és trobar grups homogenis amb individus diferenciats, mirant de trobar observacions semblants entre sí, de manera que poguem organitzar aquesta gran quantitat de dades en un nombre reduït de *clústers*. Existeixen diferents tipus de tècniques de clústering (particions, jeràrquics, etc...) basats en diferents metodologies que utilitzen coneixements teòrics provinents de diferents camps d'estudi. En aquest cas, el clústering que es duu a terme és el *clústering jeràrquic*.

S'empra el **mètode de Ward**, que consisteix en fer servir la pèrdua d'informació que es produeix al integrar els diferents individus en els clústers. Aquesta pèrdua es pot mesurar a través de la suma total dels quadrats de les desviacions de cada individu respecte la mitjana del clúster, de manera que s'aniran agrupant aquells individus que menys incrementin aquesta magnitud al ajuntar-se.

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

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

$$s_{ij} = \frac{\sum_{h=1}^{p_1} (1 + |x_{ih} - x_{jh}|/G_h) + a + \alpha}{p_1 + (p_2 - d) + p_3},$$

amb

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

Així doncs, s'ha calculat la matriu de discrepàncies fent servir la distància de Gower i, amb el mètode de Ward, s'ha realitzat el procés de *clustering*. A continuació es mostra una representació del resultat amb un dendrograma (Figura 9).

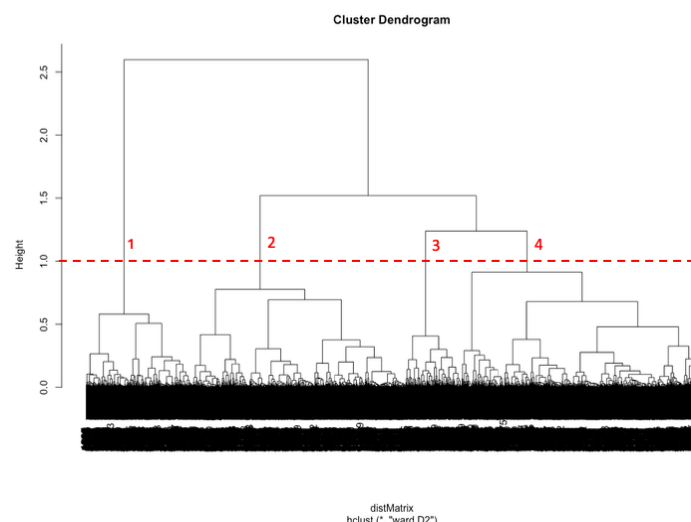


Figura 9: Arbre jeràrquic de les 5000 observacions obtingut amb la funció de R "hclust".

Observant el gràfic, es considera oportú realitzar una partició en 4 clústers, cadascun amb el següent nombre de individus:

Classe	1	2	3	4
Nre. Individus	1738	1932	874	456

Taula 6: Taula de contingència del nombre d'individus i les classes.

Un cop realitzat el procés de *clustering*, convé tenir una idea més àmplia de com són les unitats d'estudi dins de cada grup. Així doncs, s'han elaborat estadístiques descriptives per grups amb l'objectiu de crear un perfil per a cada clúster que representi les seves característiques.

En primer lloc, s'ha realitzat un test *Chi-quadrat* per a comprovar quines diferències en les variables qualitatives són estadísticament significatives entre clúster i, per a les diferències entre variables numèriques, els tests ANOVA o de *Kruskal-Wallis*. Llevat de "id", com és normal ja que és la variable identificadora, tots els contrastos han sigut significatius.

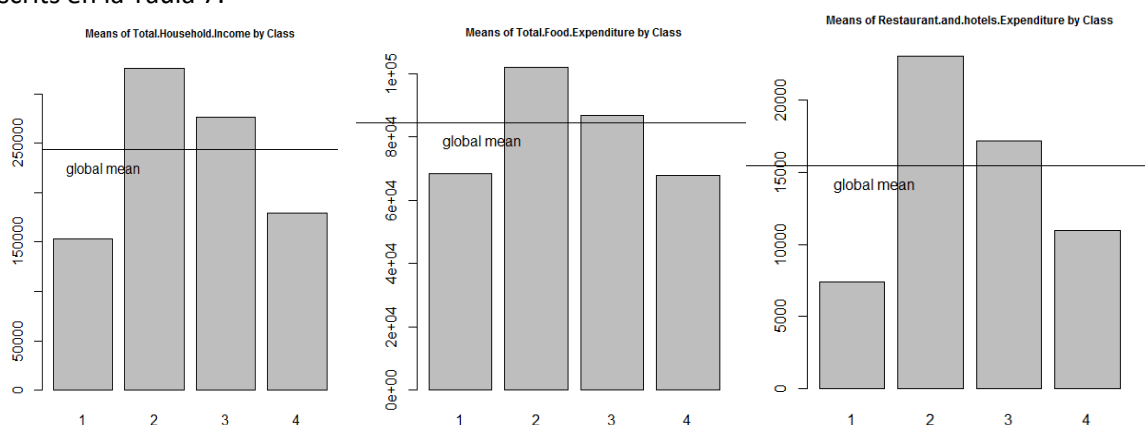
Després, s'han realitzat barplots (variables numèriques) i snakeplots (categòriques) per clústers, per tal de reconèixer característiques clau de cada grup (veure Figures 10-30). A continuació s'ha agregat una taula en la que es recopila tota la informació descriptiva de cada clúster (obtinguda a partir de les representacions gràfiques) mitjançant una petita descripció que inclou els seus atributs principals.

Clúster	Nom	Descripció
1	Homes pagesos de classe baixa	Famílies agrícoles en les que el cap de família és un home sense estudis o amb estudis bàsics. També són famílies amb molts nens petits i uns ingressos baixos (150000 de mitjana). Tenen despeses baixes respecte els altres clúster llevat de la despesa en tabac i alcohol (segon grup amb més despeses en aquests productes) i, òbviament, en jardineria i granja. A més, són famílies amb pocs béns (cotxes, mòbils...)
2	Homes de classe alta que treballen en empreses	Famílies amb els ingressos i les despeses més altes així com també disposen de molts béns. El cap de la família és un home amb estudis superiors que treballa en el sector terciari (majoritàriament en empreses privades). Són el grup amb més membres de la família treballant.
3	Dones de classe alta no treballadores	Famílies dirigides per una dona vídua o soltera que no treballa, amb ingressos i despeses altes (no les que més) excepte en els productes de tabac i alcohol. Són les famílies amb el millor tipus d'habitatge: més metres quadrats, més habitacions, etc. Tenen pocs nens petits.
4	Dones de classe baixa treballadores	Famílies amb el menor nombre de nens i les que menys béns posseeixen. Tenen ingressos i despeses força baixos. El cap de família és una dona soltera (o vídua en alguns casos) treballadora amb estudis baixos.

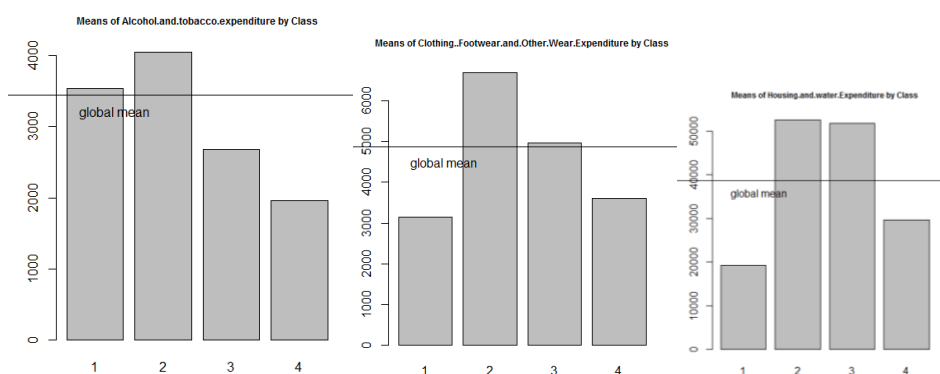
Taula 7: Taula resum profiling del grups obtinguts amb el clústering jeràrquic.



Per acabar, s'adjunten les representacions més rellevants per a la realització dels perfils dels 4 grups descrits en la Taula 7.



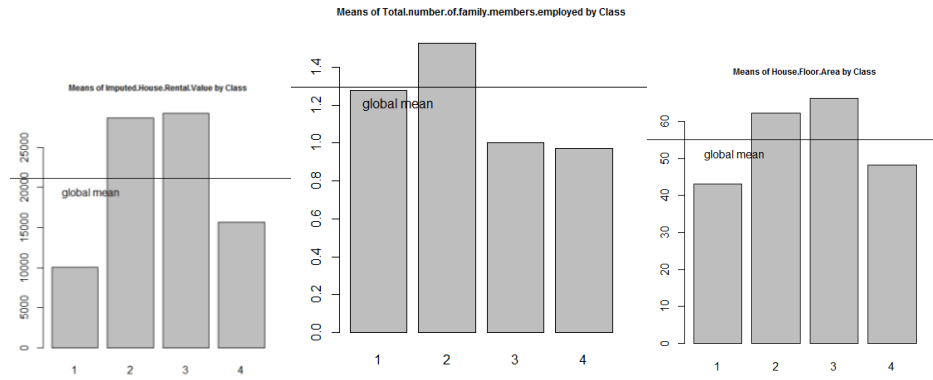
Figures 10-12: Barplots de "Total.Household.Income", "Total.Food.Expenditure" i "Restaurants.and.hotels.Expenditure" per clústers.



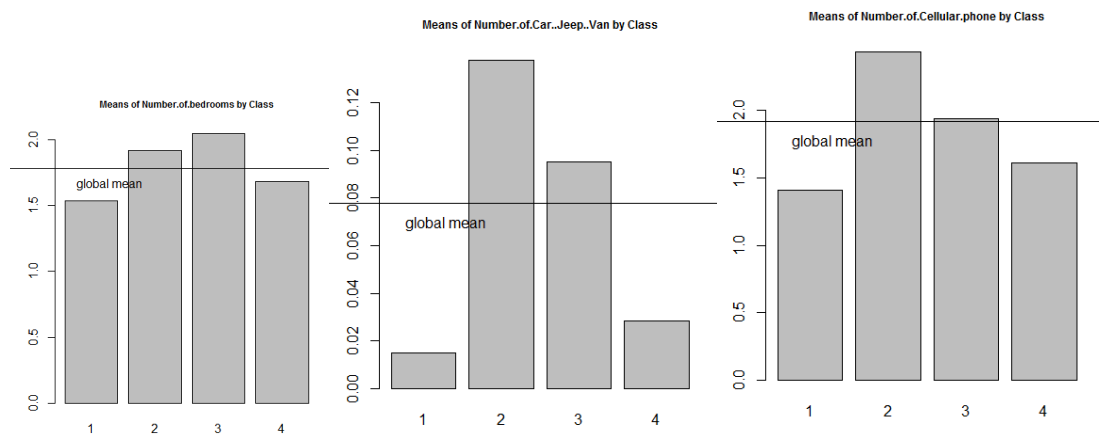
Figures 13-15: Barplots de "Alcohol.and.tobacco.expenditure", "Clothing\_Footwear.and.Other.Wear.Expenditure" i "Housing.and.water.Expenditure" per clústers.



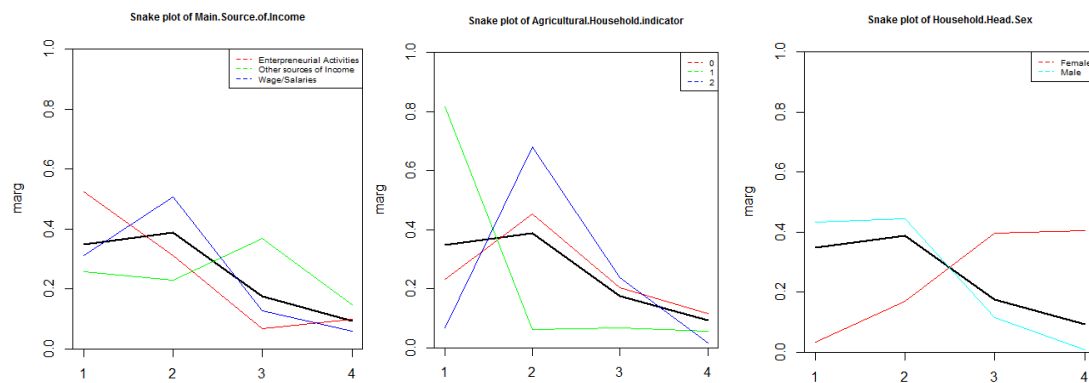
Figures 16-18: Barplots de "Crop.Farming.and.Gardening.expenses", "Total.Income.from.Entrepreneurial.Activities" i "Members.with.age.less.than.5.year.old" per clústers.



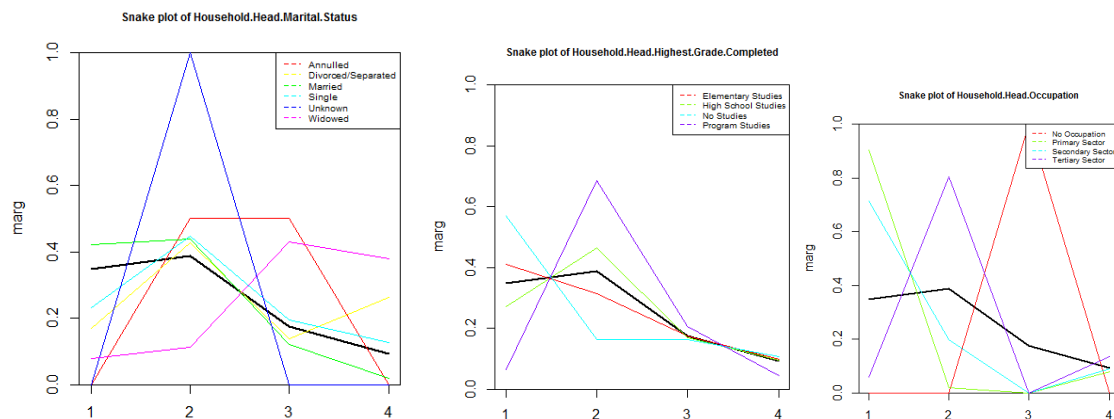
Figures 19-21: Barplots de “Imputed.House.Rental.Value”, “Total.number.of.family.members.employed” i “House.Floor.Area” per clústers.



Figures 22-24: Barplots de “Number.of.bedrooms”, “Number.of.Car..Jeep..Van” i “Number.of.Cellular.phone” per clústers.



Figures 25-27: Snakeplots de “Main.Source.of.Income”, “Agricultural.Household.indicator” i “Household.Head.Sex” per clústers.



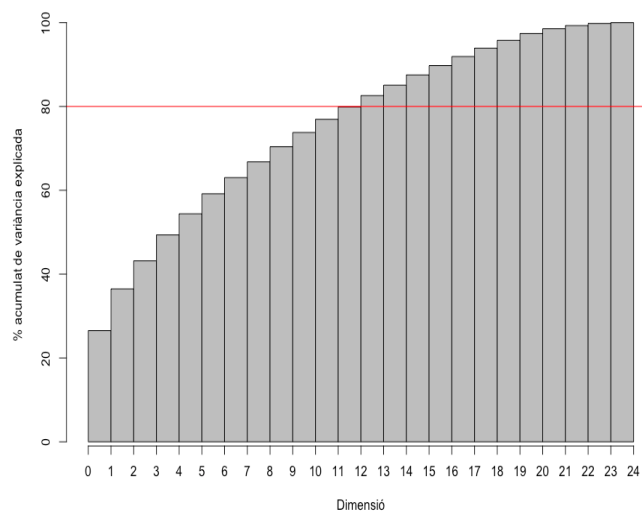
Figures 28-30: Snakeplots de "Household.Head.Marital-Status", "Household.Head.Highest.Grade.Completed" i "Household.Head.Ocupation" per clústers.

## 5. ACP de les variables numèriques.

L'Anàlisi de Components Principals (ACP) és una tècnica estadística de síntesi de la informació, o reducció de la dimensió (nombre de variables) d'un conjunt de dades. L'objectiu d'aquesta anàlisi serà reduir les variables a un menor nombre perdent el mínim d'informació possible. Els nous components principals o factors seran una combinació lineal de les variables originals, i a més seran independents entre si.

Un aspecte clau de l'ACP és la interpretació dels factors, ja que aquesta no ve donada a priori, sinó que serà deduïda després d'observar la relació dels factors amb les variables inicials (s'haurà doncs d'estudiar tant el signe com la magnitud de les correlacions).

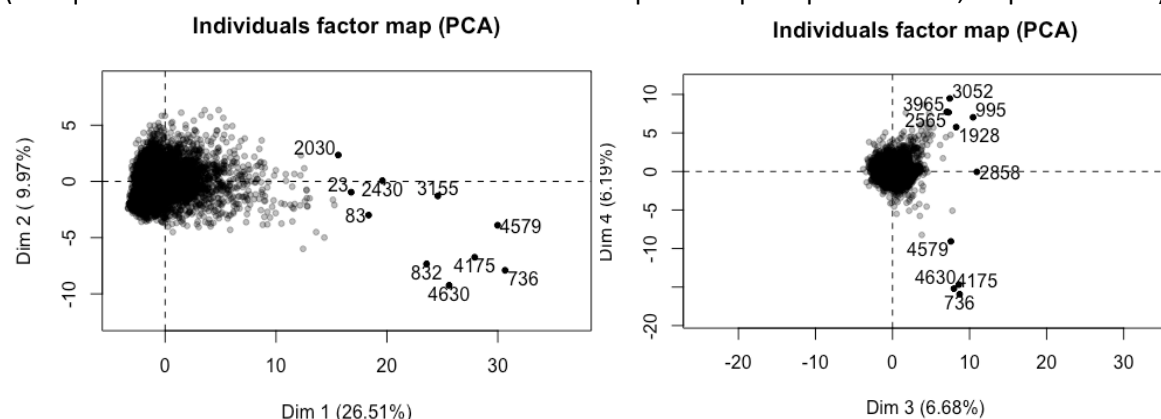
En primer lloc, s'ha fet un *scree plot* del percentatge de variància així com també una taula amb els valors propis i percentatges de cada dimensió per tal de seleccionar el nombre de dimensions necessàries per assolir un llindar del 80% de la variància total de les dades. Amb el % acumulat de la taula veiem que si considerem 12 dimensions ja és suficient.



Dimensió	Valor propi	% de variància	% acumulat
1	6.36317859	26.5132441	26.51324
2	2.39348746	9.9728644	36.48611
3	1.60230597	6.6762749	43.16238
4	1.48472021	6.1863342	49.34872
5	1.21324356	5.0551815	54.40390
6	1.14517718	4.7715716	59.17547
7	0.92521958	3.8550816	63.03055
8	0.90575634	3.7739847	66.80454
9	0.86046582	3.5852742	70.38981
10	0.81734454	3.4056022	73.79541
11	0.75358861	3.1399526	76.93537
12	0.70919917	2.9549966	79.89036
13	0.64888104	2.7036710	82.59403
...	...	...	...
24	0.04785091	0.1993788	100.00000

Figura 31: Barplot del % acumulat de variància capturat. Taula 8: Valors propis i percentatge de cada dimensió.

Per començar, s'han representat en un núvol de punts els individus sobre els diferents plans factorials (en aquest informe només es mostraran els dels components principals 1-2 i 3-4, respectivament).



Figures 32 i 33: Representacions dels individus sobre els components principals 1-2 (esquerra) i 3-4 (dreta).

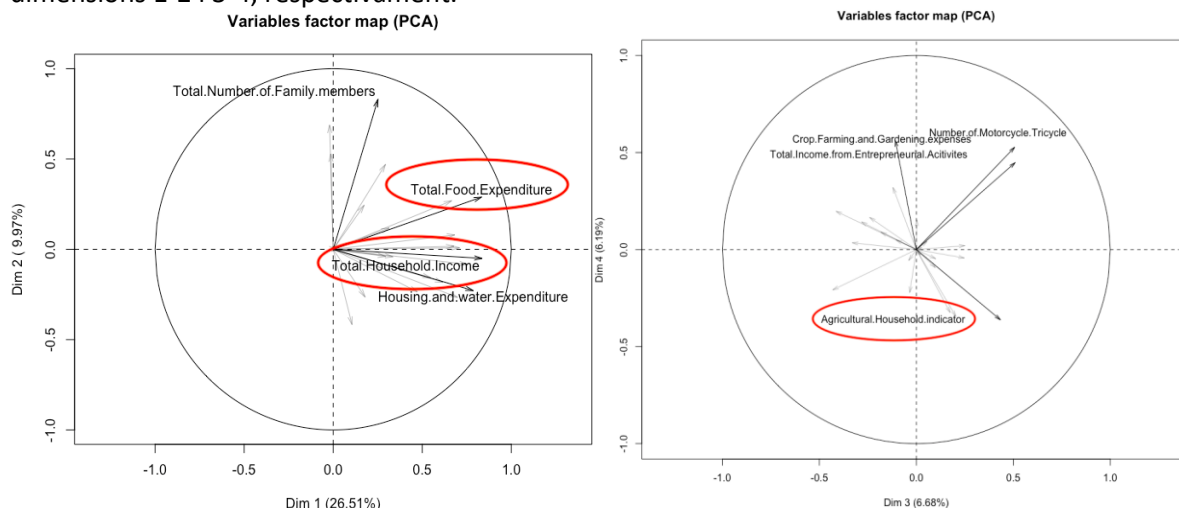
Amb aquests gràfics es pot observar la manera com els individus queden projectats sobre l'espai generat per les dues components principals. Els individus marcats són els 10 que més contribueixen, és a dir, aquells que més variabilitat aporten (són els més allunyats al centre). És important comentar que en el gràfic d'individus de les dimensions majors (3 i 4) el núvol de punts és més compacte ja que hi ha menys dispersió entre els punts (és captura menys variància).

A continuació es mostra una taula amb les variables que més contribueixen en cada dimensió (en vermell):

Variable	Dim 1	Dim 2	Dim 3	Dim 4
Total.Household.Income	10.998543	0.106425	0.621078	0.149995
Total.Food.Expenditure	10.916569	3.492815	0.097981	0.213410
Agricultural.Household.indicator	0.005952	0.000508	11.72594	8.83443
Restaurant.and.hotels.Expenditure	7.321359	0.266079	0.072986	3.303973
Alcohol.and.tobacco.expenditure	0.4738356	2.486619	2.08414	0.549147
Clothes.Footwear.Other.Wear.Expenditure	7.9032331	0.003174	0.046484	0.006925

Taula 9: Resum de les variables més contributors a les dimensions 1-4.

En els gràfics que s'adjunten sota aquestes línies s'han projectat les variables quantitatives sobre les dimensions 1-2 i 3-4, respectivament.

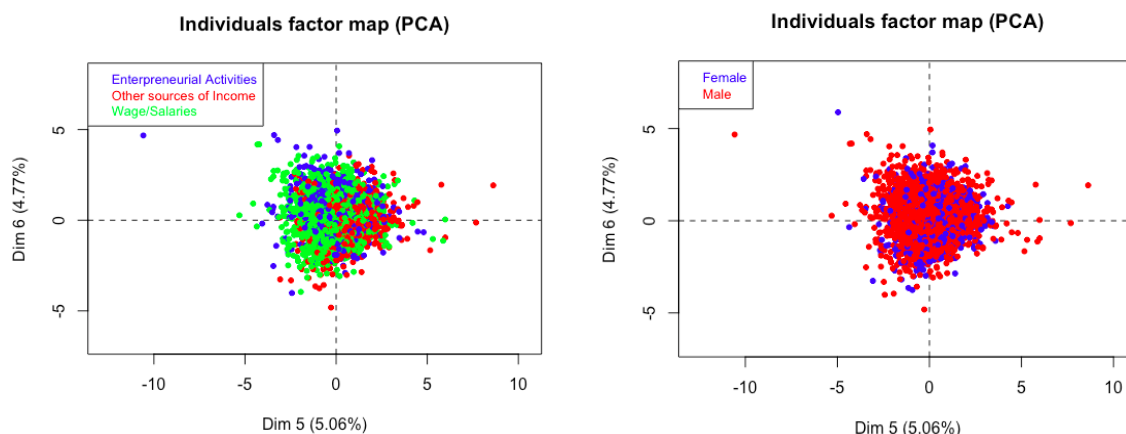


Figures 34 i 35: Projeccions de les variables sobre els components principals 1-2 (esquerra) i 3-4 (dreta).

Pel que fa a la interpretació dels gràfics cal comentar que la llargada de les fletxes correspon a la correlació de la variable amb la dimensió. Si estan dintre del quadrant positiu (de 0 a +1) significa que

les variables estan correlacionades positivament. Així doncs, es veu que en el primer gràfic *Total.Food.Expenditure* és una variable correlacionada positivament amb ambdues dimensions (més amb la dimensió 1 que amb la 2, ja que la fletxa arriba molt a la dreta però no tant adalt) i en canvi en el segon, *Agricultural.Household.indicator* està correlacionada positivament amb la dimensió 3 i negativament amb la dimensió 4.

Per acabar, s'han representat els individus sobre els diferents plans factorials agrupats per variables qualitatives, però en la majoria de casos no s'ha vist cap relació entre la variable i la dimensió:



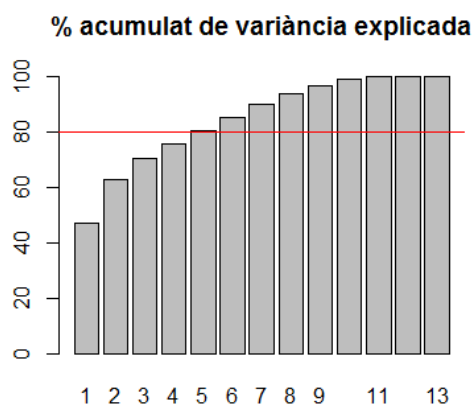
Figures 36 i 37: Núvol dels individus sobre les dimensions 5 i 6, agrupades segons treball (esquerra) i sexe (dreta)

## 6. ACM de les variables qualitatives.

L'Anàlisi de correspondències múltiples (ACM) és una tècnica estadística per a dades categòriques utilitzada per detectar i representar estructures subjacents en un conjunt de dades. L'objectiu d'aquest anàlisi és identificar grups d'individus amb un perfil semblant respecte a la resposta a les variables categòriques i reconèixer associacions entre categories de les variables. El mètode utilitzat per realitzar aquesta anàlisi és la **matriu de Burt** ja que és el més adequat per a dades de grans dimensions, com és el nostre cas.

La funció d'R que s'ha fet servir per calcular l'ACM és la *MCA*, en la qual s'ha indicat quina és la base de dades (*X*) i el vector d'índexs de les variables quantitatives suplementàries (*quanti.sup*). En el nostre cas hem considerat com a variables actives aquelles variables qualitatives que conceptualment fan referència al treball del cap de família (ocupació, tipus de treball, etc) i com a suplementàries, totes les quantitatives i la resta de qualitatives.

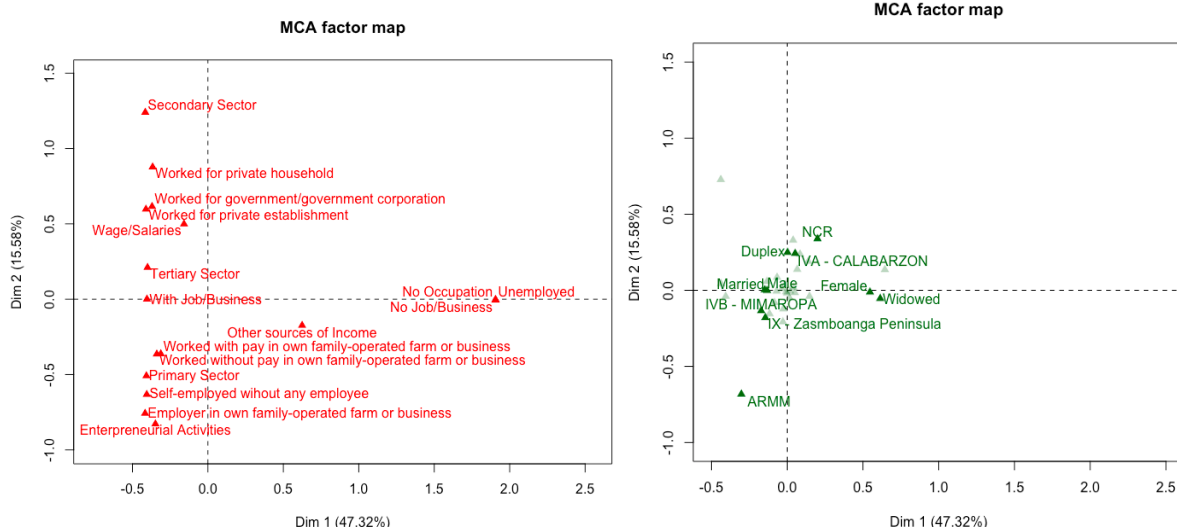
Primerament, per tal de mirar la proporció de variància explicada per les diferents dimensions utilitzarem els valors propis. S'ha fet un *barplot* del percentatge de variància acumulada així com també una taula a partir de la qual s'observa que considerant 5 dimensions s'assoleix el llindar del 80% de la variància total de les dades.



Dimensió	Valors propis	% de variància	% acumulada
1	6.133124e-01	4.731667e+01	47.31667
2	2.019817e-01	1.558277e+01	62.89944
3	9.789680e-02	7.552677e+00	70.45211
4	6.792992e-02	5.240751e+00	75.69287
5	6.264319e-02	4.832883e+00	80.52575
6	6.250000e-02	4.821836e+00	85.34759
7	6.250000e-02	4.821836e+00	90.16942
8	4.641433e-02	3.580837e+00	93.75026
9	3.903649e-02	3.011641e+00	96.76190
10	3.110375e-02	2.399635e+00	99.16153
11	1.086808e-02	8.384653e-01	100.00000
12	1.033605e-55	7.974198e-54	100.00000
13	4.590644e-56	3.541653e-54	100.00000

Figura 38: Barplot del % acumulat de variància capturada. Taula 10: Valors propis i percentatge de cada dimensió.

A continuació es mostren les projeccions sobre el primer pla factorial (dimensió 1 a l'eix de les x i 2 a l'eix vertical) de les modalitats de les variables qualitatives actives, en vermell, i il·lustratives, en verd.



Figures 39 i 40: Projecció de les modalitats de les variables actives (esquerra) i complementàries (dreta) sobre el primer pla factorial, dimensions 1 i 2.

S'observa com, d'entre les variables actives, les categories que més s'associen amb la dimensió 1 són "No Occupation", "Unemployed" i "No Job/Business", les tres amb un 25.89% de contribució. Quant a les modalitats més relacionades amb la dimensió 2, aquestes són "Entrepreneurial Activities", "Employer in own family-operated farm or business" i "Secondary Sector", les dues primeres correlacionades negativament. Aquestes sis variables, les variables que més es correlacionen amb components principals 1 i 2, són les més importants per explicar la variabilitat del conjunt de dades.

També s'ha realitzat un MCA plot de les modalitats de les variables actives per tal d'estudiar relacions i oposicions entre les variables que conceptualment fan referència al treball del cap de família.

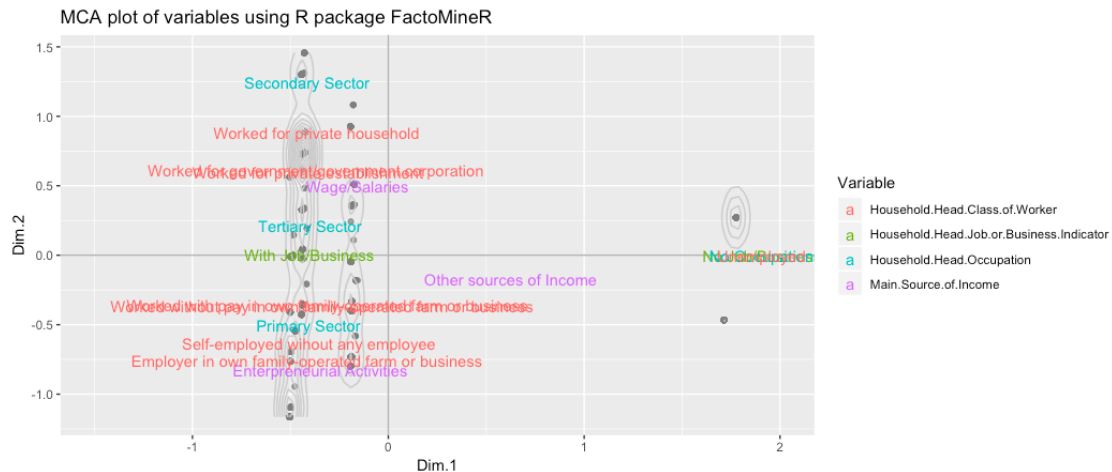


Figura 41: MCA plot de les modalitats de les variables actives.

A partir de la Figura 41 es pot veure com aquelles persones que treballen en el sector primari són també famílies en les que la principal font d'ingressos és d'activitats empresarials i en les que el cap de família és un treballador autònom sense treballadors o treballador de la seva pròpia granja. També es pot veure que els que treballen en el sector terciari obtenen la majoria dels seus ingressos dels seus salaris. Per acabar, com és obvi i com ja s'ha dit amb anterioritat, les modalitats de "No Occupation", "Unemployed" i "No Job/Business" estan molt relacionades.

Finalment s'ha realitzat el gràfic *plotellipses*, en el que es dibuixen el·lipses de confiança al voltant de les categories de diverses variables categòriques, en aquest cas, de les variables il·lustratives.

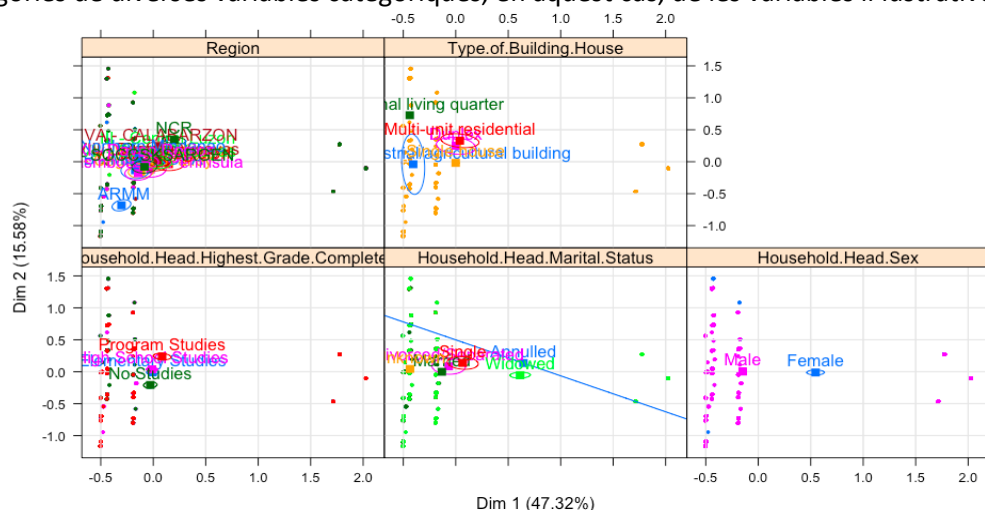


Figura 42: Plotellipses de les modalitats de les variables complementàries sobre el primer pla factorial.

Les categories de gènere són significativament diferents entre si. Per a la variable "Region", només la modalitat "ARMM" és significativament diferent a la resta. Pel que fa al grau més alt completat del cap de la llar, les categories de estudis elementals i secundària no són significativament diferents de la resta i per acabar, les opcions "Anulled" i "Widowed" són les diferents a la resta de l'estat civil.

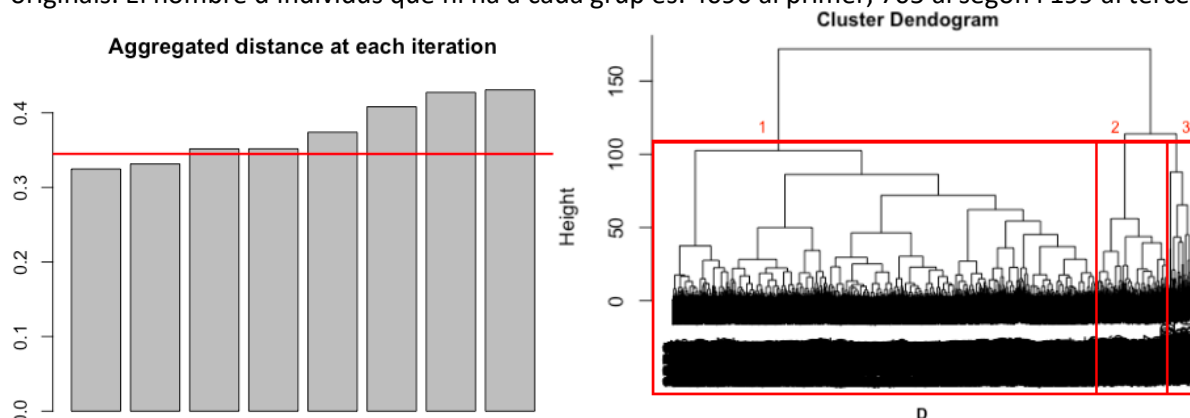
## 7. Clustering jeràrquic sobre les components factorials d'ACP.

En aquest apartat es replicarà el procés de *clustering* jeràrquic del capítol 4 fent servir com a *input* el resultat de l'anàlisi de components principals i no les dades originals.

Primerament, s'han extret les coordenades de tots els individus de la base de dades sobre les 12 primeres dimensions (en el capítol 5 ja s'ha vist que eren suficients perquè capturen aproximadament el 80% de la variabilitat de les dades) per després poder calcular la matriu de **distàncies euclidianes** (D)

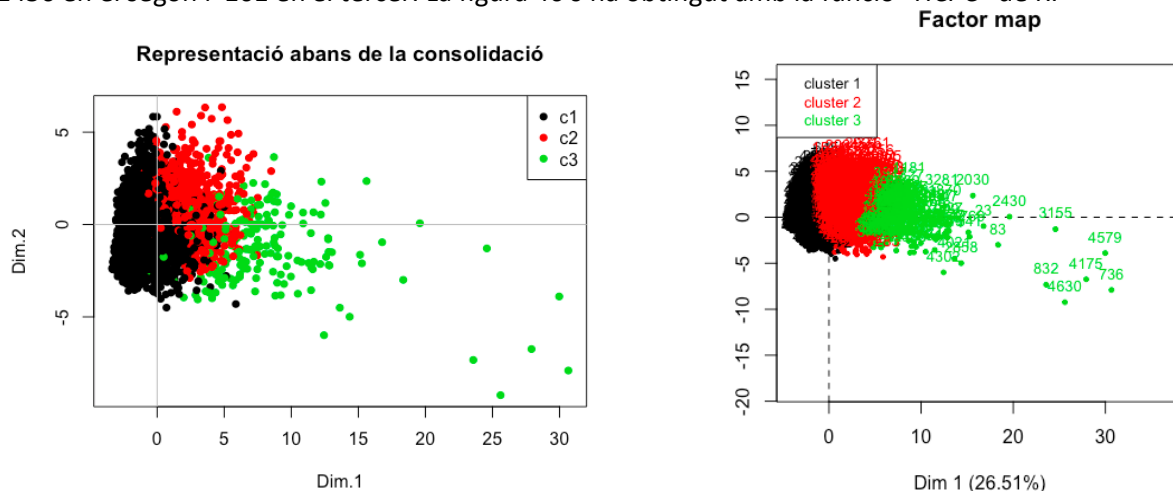
que hi ha entre els individus, que serà l'*input* del procés de *clustering*. Un altre cop, s'ha fet servir el **mètode d'agregació de Ward**.

En vista del gràfic amb les distàncies agregades a cada iteració (*Figura 43*), amb el que s'observa que quasi bé no hi ha diferències entre la tercera i quarta barra, s'ha considerat oportú tallar l'arbre jeràrquic resultant (*Figura 44*) en 3 *clusters*, un grup menys que en el clustering realitzat sobre les dades originals. El nombre d'individus que hi ha a cada grup és: 4096 al primer, 705 al segon i 199 al tercer.



Figures 43 i 44: Distància acumulada a cada tall (esquerra) i dendograma (dreta).

Per avaluar si la qualitat de les particions és bona, mirem quin percentatge de la variància total correspon a la variància entre classes de manera que si aquest percentatge és alt, voldrà dir que la qualitat és bona (volem que hi hagi molta variància entre classes i poca intra classes, és a dir, entre els individus d'un mateix *cluster*). En un primer moment, aquest percentatge ha sigut de poc més del 22% (veure *Figura 45*), molt baix, per la qual cosa s'ha assignat cada individu al grup amb el centre de gravetat més proper a ell. Així, els *clusters* s'han consolidat i ara el 26.5% de la variància total és variància entre classes. El tamany de cada grup ha variat, i ara és: 3383 individus en el primer *cluster*, 1456 en el segon i 161 en el tercer. La figura 46 s'ha obtingut amb la funció "*HCPC*" de R.



Figures 45 i 46: Representació per clusters abans (esquerra) i després (dreta) de la consolidació de les particions.

A continuació cal crear un perfil per a cada *clúster* que representi les seves característiques per tal de tenir una idea més clara de com és cada grup.

En primer lloc s'ha mirat el test Chi-quadrat de la sortida de la funció "*HCPC*" per veure si hi ha independència o no entre les variables categòriques i la variable de classe, la qual indica a quin grup pertany cada observació. També, per mirar si la relació entre les numèriques i la variable de classe és significativa, s'han mirat els valors de  $\eta^2$  (percentatge de variància explicada obtingut a partir de la regressió lineal entre ambdues) i el p valor del test de l'ANOVA. Els resultats s'adjunten a la taula següent i es pot observar que, com que tots els p-valors de  $\chi^2$  són inferiors a 0.05, totes les relacions



del *cluster* amb les categòriques són estadísticament significatives. Pel que fa a les numèriques, només s'han considerat significatives aquelles variables amb un  $\eta^2$  superior al 30% ja que, un p valor inferior a 0.05 pot ser degut al gran tamany de la mostra (n=5000).

Variables categòriques	P valor	Graus de llibertat	Variables numèriques	eta2	P valor
Household.Head.Highest.Grade.Completed	3.440742e-199	6	Total.Household.Income	0.485983578	0
Household.Head.Occupation	3.785828e-81	6	Total.Food.Expenditure	0.539966029	0
Region	2.108654e-80	32	Restaurant.and.hotels.Expenditure	0.351974822	0
Household.Head.Class.of.Worker	2.702482e-62	14	Clothing..Footwear.and.Other.Wear.Expenditure	0.405531359	0
Main.Source.of.Income	4.818270e-28	4	Housing.and.water.Expenditure	0.382160461	0
Household.Head.Job.or.Business.Indicator	2.224944e-06	2	Medical.education.transport.and.communication.expenditure	0.509352499	0
Household.Head.Marital.Status	3.872770e-04	10	Miscellaneous.goods.and.special.occasions.expenditure	0.378482755	0
Type.of.Building.House	9.533519e-03	8	Number.of.Car..Jeep..Van	0.390867651	0
			Number.of.Cellular.phone	0.415096835	0

Taula 11: Taula de resultats del test Chi-quadrat d'independència entre les variables categòriques i la classe.

Un cop definides les variables rellevants per caracteritzar els grups, s'ha realitzat un resum numèric de les variables numèriques comparant les mitjanes globals amb les de cada *clúster* (Taula 12), informació molt pràctica per descriure cadascun dels grups.

Variable	Mean global	Mean 1	Mean 2	Mean 3
Total.Household.Income	243812.38	136036.40	389342.10	1192353
Total.Food.Expenditure	84600.03	61255.21	123999.90	218819.80
Restaurant.and.hotels.Expenditure	15474.50	7408.18	27581.77	75475.81
Clothing..Footwear.and.Other.Wear.Expenditure	4873.08	2531.79	8209.21	23899.11
Housing.and.water.Expenditure	38716.19	21433.20	59180.54	216804.80
Medical.education.transport..communication.expenditure	30074.35	12693.75	52060.15	196454.80
Miscellaneous.goods.and.special.occasions.expenditure	17327.39	8395.87	31448.47	77296.60
Number.of.Car..Jeep..Van	0.08	0.005	0.12	1.19
Number.of.Cellular.phone	1.92	1.24	3.23	4.23

Taula 12: Taula resum de variables numèriques per cada cluster.

Seguidament s'ha mirat el percentatge d'individus que pertanyen a cadascun dels nivells que tenen les variables categòriques respecte al total del grup. D'aquesta manera, s'ha pogut estudiar si hi havia nivells que fossin característics d'algun dels tres *clusters*. La taula amb les dades és massa gran per la qual cosa en aquest document només es mostrarà com a exemple els valors de la variable "*Household.Head.Occupation*", adjuntats a la Taula 13.

Els valors dins dels parèntesis fan referència al percentatge d'individus d'aquell nivell sobre el total d'individus de la base de dades. Aquests valors són molt més alts al *clúster* 1 degut al gran tamany respecte a la resta. La distribució que s'observa dintre dels 3 *clusters*, és que mentre en el primer, els diferents sectors estan ben repartits, en els *clusters* 2 i 3 tenen un percentatge molt més alt al sector terciari que a la resta.

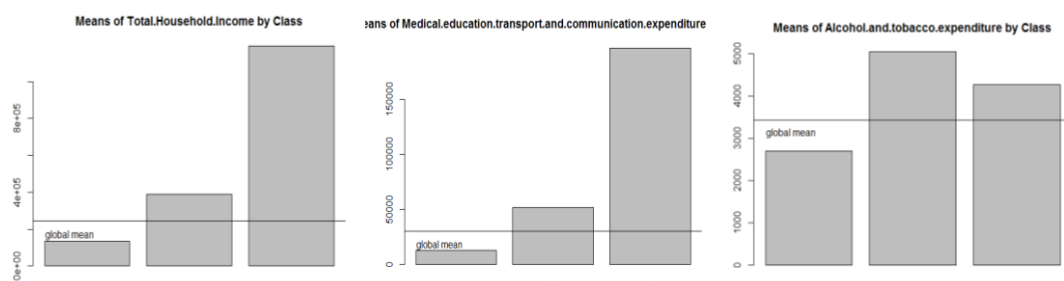
Nivell	Cluster 1	Cluster 2	Cluster 3
Primary Sector	39.88% (85.92%)	14.75% (13.70%)	3.73% (0.38%)
Secondary Sector	5.53% (73.33%)	4.55% (25.89%)	1.24% (0.78%)
Tertiary Sector	38.99% (57.32%)	59.55% (37.68%)	71.43% (5%)
No Occupation	15.60% (60.41%)	21.15% (35.24%)	23.60% (4.35%)

Taula 13: Percentatge d'individus per cadascun dels nivells de la variable *Household.Head.Occupation* segons el cluster.

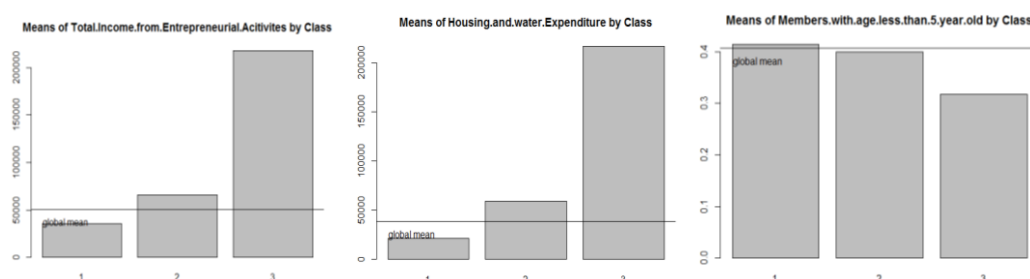
Finalment, s'han realitzat barplots i snakeplots per les variables numèriques i categòriques, respectivament (veure Figures 47-61) de manera que s'ha pogut definir més concretament com és cada grup. Aquestes descripcions s'adjunten a la taula sota aquestes línies (*Taula XX*).

Clúster	Nom	Descripció
1	Famílies amb pocs recursos	Famílies amb pocs ingressos i, en general, poques despeses. La principal font d'ingressos ve d'activitats empresarials ja que el cap de família és majoritàriament un autònom sense treballadors. Les cases estan situades en barris de protecció oficial, tenen poques habitacions i pocs metres quadrats. No disposen de cotxes i són el grup amb més nens menors de 5 anys. De manera predominant, el responsable de la família no té estudis i és solter, divorciat o vidu. Són la major part de la mostra.
2	Famílies no destacables	Grup que es caracteritza per no destacar en cap variable a excepció de que són les famílies que més gasten en alcohol i tabac i que més motocicletes tenen. El cap de família és normalment una persona casada, amb estudis superiors.
3	Famílies adinerades	Famílies minoritàries en la mostra, que tenen molts ingressos i despeses, alt nombre de cotxes i telèfons mòbils, poc nombre de fills. Viuen en cases molt ben valorades, amb bastants metres quadrats i dormitoris. El cap de família normalment treballa pel govern o té una empresa o granja pròpia; té estudis superiors, sol treballar en el sector terciari i hi ha casats i divorciats per igual.

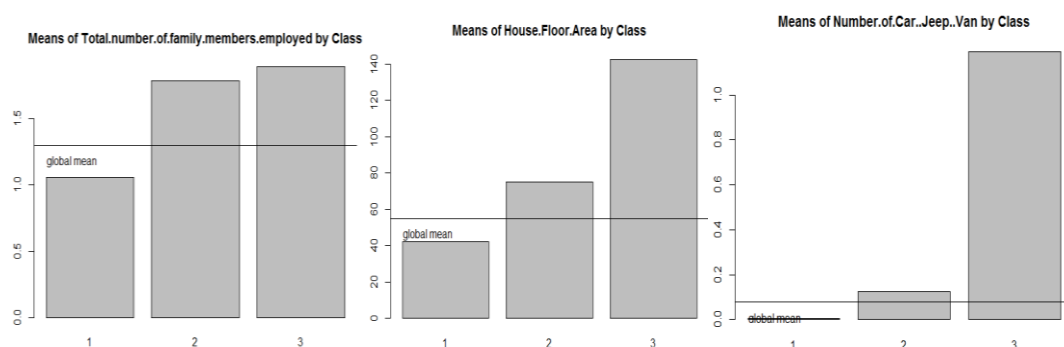
*Taula 14: Descripció dels clústers obtinguts amb el clústering jeràrquic sobre els components principals.*



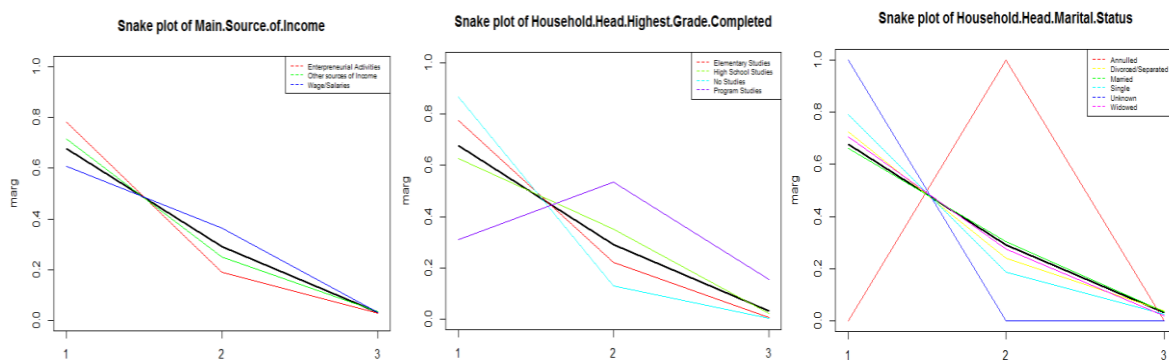
*Figures 47-49: Barplots de "Total.Household.Income", "Medical.education.transport.and.communication.expenditure" i "Alcohol.and.tobacco.expenditure" per clústers.*



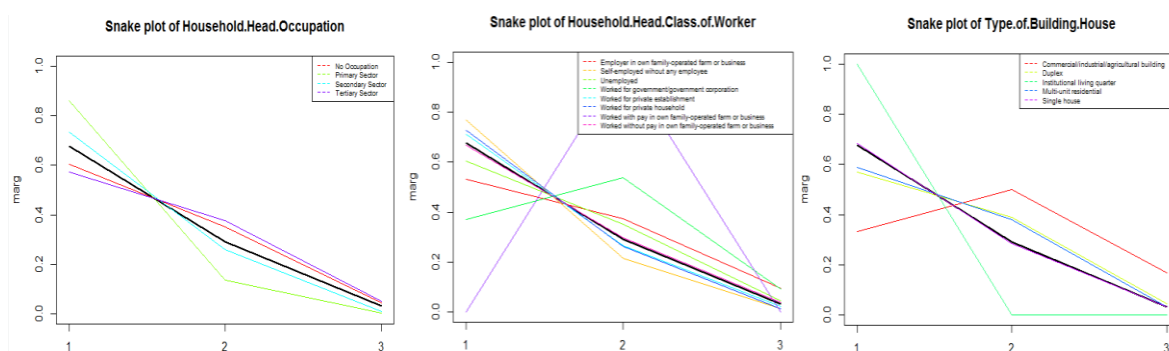
*Figures 50-52: Barplots de "Total.Income.from.Entrepreneurial.Activities", "Housing.and.water.Expenditure" i "Members.with.age.less.than.5.year.old" per clústers.*



*Figures 53-55: Barplots de "Number.of.family.members.employed", "House.Floor.Area" i "Number.of.Car.Jeep.Van" per clústers.*



Figures 56-58: Snakeplots de "Main.Source.of.Income", "Household.Head.Highest.Grade.Completed" i "Household.Head.Marital.Status" per clústers.



Figures 59-61: Snakeplots de "Household.Head.Occupation", "Household.Head.Class.of.Worker" i "Type.of.Building.House" per clústers.

## 8. Anàlisi Factorial Múltiple.

L'Anàlisi factorial múltiple (AFM) és una tècnica estadística per a dades tant numèriques com categòriques. Aquest procediment és utilitzat per estudiar la similitud entre individus respecte el conjunt de variables i la relació entre aquestes. Per aplicar aquesta tècnica, s'han de dividir les variables de la base de dades en grups per similituds entre significat de variables, aquests grups han d'estar formats per variables numèriques o categòriques, no poden ser mixtos. Els objectius d'aquest anàlisi són: estudiar les similituds i diferències entre grups i comparar les característiques dels individus des d'un anàlisi per separat.

La funció d'R que s'ha fet servir per calcular l'AFM és la *MFA*, en la qual s'ha indicat quina és la base de dades (*X*), un vector del tamany de cada grup (*group*), un vector del tipus de variables que hi ha en un grup (*type*), (es posarà "s" per numèriques i "n" per categòriques), un vector dels índexs dels grups suplementaris (*num.group.sup*), el numero de dimensions a mantenir en els resultats (*ncp*) i un vector amb els noms de cada grup (*name.group*). En el nostre cas s'han format tres grups de variables numèriques. Aquests fan referència a les variables que tenen a veure amb ingressos i despeses (*Incomes and Expenditures*, *n*= 11), objectes i informació de la casa (*Objects and home info*, *n*= 8) i per últim informació de la família (*Family info*, *n*= 5). També s'han creat dos grups de variables qualitatives, aquests han sigut en referencia a les variables d'ocupacions (*Occupations*, *n*= 5) i un de les variables qualitatives restants (*Other qualitative*, *n*= 4). Les agrupacions que s'han considerat actives han sigut *Income and Expenditures*, *Objects and Home info* i *Occupations*.

Primerament, per tal de mirar la proporció de variància explicada per les diferents dimensions s'han utilitzat els valors propis. S'ha vist com 17 dimensions expliquen aproximadament un 80% de la variància acumulada per tant s'ha fet l'anàlisi amb l'argument *ncp=17*.

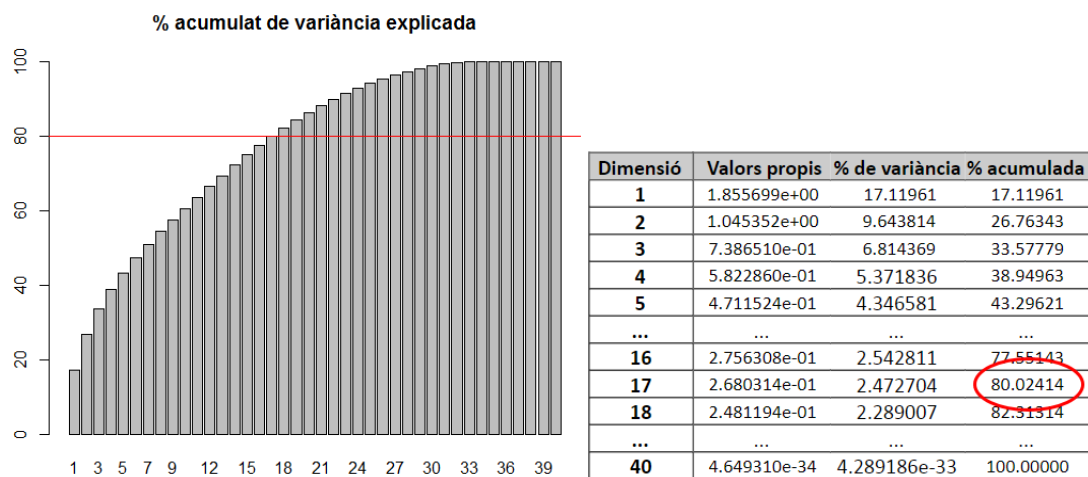


Figura 62: Barplot del % acumulat de variància capturat. Taula 15: Valors propis i percentatge de cada dimensió.

A continuació es mostren les projeccions sobre el primer pla factorial (dimensió 1 a l'eix de les x i 2 a l'eix vertical) de tots els grups.

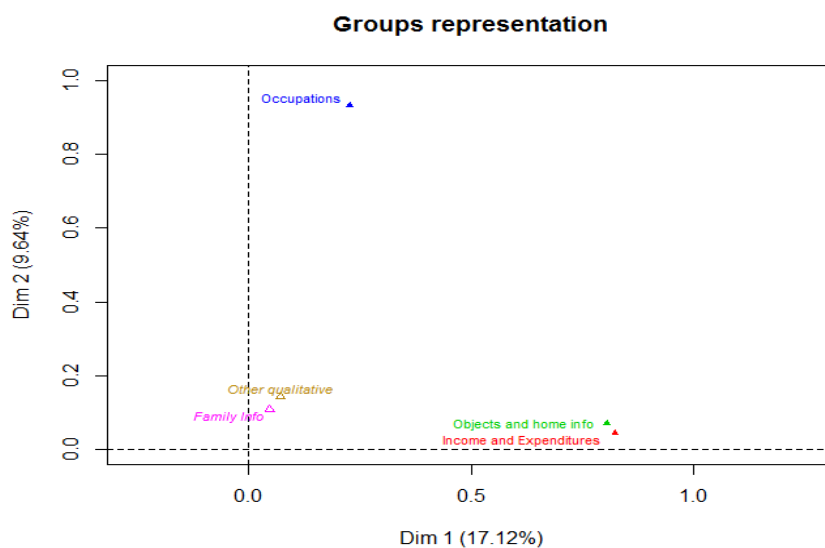
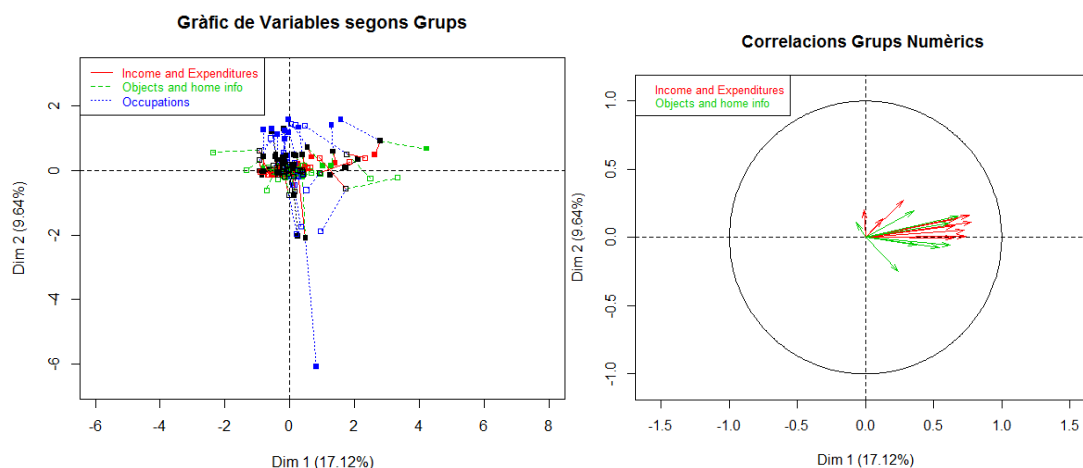


Figura 63: Gràfic de la representació dels grups sobre primer pla factorial.

En aquest gràfic es pot observar com els grups actius, són els que aporten més informació. El grups *Objects and home info* i *Income and Expenditures* són molt semblants i els dos aporten molta informació de la dimensió 1, pel que fa a *Objects and home info* dona un 43.383% de la informació i el grup *Income and Expenditures* un 44.368%. Per altra banda, el grup de *Occupations* dona molta informació de la dimensió 2, en concret dona un 89.153% d'informació.

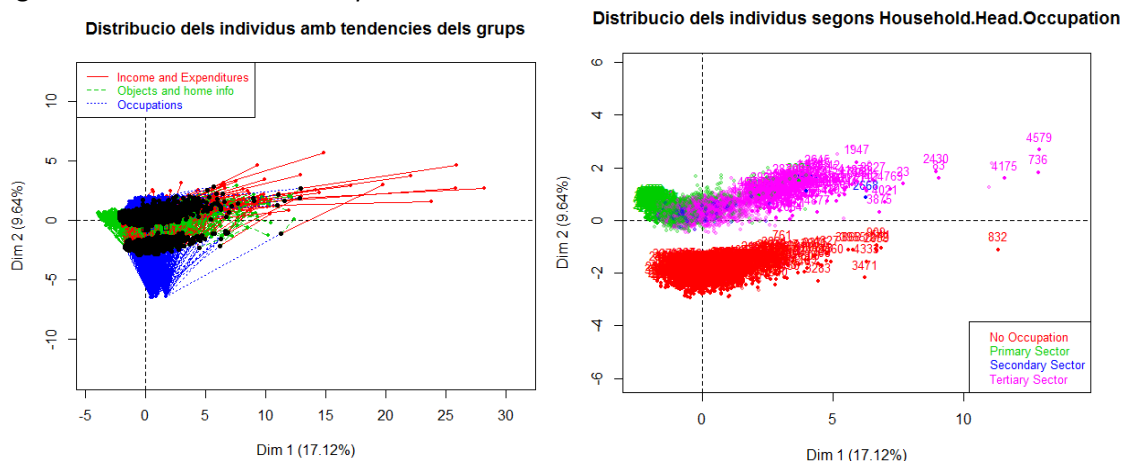
També s'han realitzat dos gràfics per veure la distribució de les variables segons el grup al que pertanyen.



Figures 64 i 65: Gràfic de representació de variables segons grup actiu i gràfic de correlacions dels dos grups actius de variables numèriques.

En el gràfic anterior situat a l'esquerra, es pot veure la distribució de les variables, segons el grup al que corresponen són d'un color diferent. Pel que fa a les variables del grup *Occupations* es pot veure com varien principalment en la dimensió 2. Les variables dels grups *Income and expenditures* i *Objects and home info*, varien en la dimensió 1, per això s'ha afegit el gràfic de la dreta, per veure amb més detall les variacions de les variables d'aquests dos grups. Són variacions principalment positives i molt semblants entre els dos.

Finalment, s'han fet dos gràfics per observar la distribució dels individus, un ens mostrarà la tendència de cada grup en els individus i un altre la distribució dels individus separats per factors de la variable categòrica *Household.Head.Occupation*.



Figures 66 i 67: Gràfic de representació d'individus segons tendència de grup actiu i gràfic d'individus segons *Household.Head.Occupation*.

En el primer gràfic (figura 66), es pot veure quina tendència tenen els grups en els individus. El grup *Occupations* tendeix a valors baixos de la dimensió 2, el grup *Income and Expenditures* tendeix a grups alts de la dimensió 1 i el grup *Objects and home info* tendeix a valors positius i negatius de la dimensió 1.

En la figura 67, es veu com hi ha una gran diferència entre els factors *Primary Sector*, *Secondary Sector* i *Tertiary Sector* amb el factor *No Occupation*, aquesta diferència afecta a la dimensió 2. Pel que fa a les diferències que es poden veure a la dimensió 1, es troben entre els factors *Primary Sector*, *Tertiary Sector* i *Secondary Sector*, es pot veure com els valors més baixos de la dimensió 2 els tenen els individus de *Primary Sector*, els més propers a 0 *Secondary Sector* i els més alts *Tertiary Sector*.

## 9. Anàlisi comparativa.

En aquest apartat s'exposaran les semblances i diferències que s'han obtingut en els diferents mètodes d'anàlisi que s'han fet servir en aquest estudi. Més concretament, es compararan l'ACP i l'ACM i, d'altra banda, els dos processos de *clustering* jeràrquic.

### ACP-ACM

Per començar, l'Anàlisi de Components Principals fa servir com a variables actives només variables numèriques i en canvi l'Anàlisi de Correspondències Múltiples únicament usa les qualitatives, de manera que en el primer cas s'ha treballat directament amb variables i en el segon, amb les categories o nivells de les variables (en ambdós casos s'ha pogut treballar a més amb els individus, tant numèrica com gràficament). A més, en l'ACP s'han fixat com a variables actives totes les numèriques i en l'ACM, només algunes de les categòriques, deixant la resta com a variables il·lustratives.

Pel que fa a la selecció de components factorials a retenir per arribar al límit del 80% de variància explicada, s'han hagut de considerar 12 i 5 dimensions, en l'ACP i l'ACM, respectivament. Això es deu a que el nombre de variables actives en el primer ha sigut més alt.

### Clusterings jeràrquics

La divergència principal entre el *clustering* jeràrquic sobre el conjunt de dades original i el fet sobre les components factorials de l'ACP és que l'*input* del primer ha sigut la matriu de discrepàncies fent servir la distància de Gower i en el segon, la matriu amb les distàncies euclidianes entre els individus de les primeres 12 dimensions. A més, en el primer clustering s'han fet 4 particions d'un tamany relativament semblant mentres que en el segon se n'han fet 3 de tamany molt diferents. També s'ha detectat que fent servir les components de l'ACP, les diferències entre els *clústers* s'accentuen molt per alguns grups en particular: en el primer clustering hi ha punts de trobada entre les particions i totes tenen característiques concretes del grup mentre que en el segon, el grup 2 no destaca en res i les disparitats entre els grups 1 i 3 són molt grans.

Pel que fa a les coincidències entre els dos processos cal comentar tres aspectes: en ambdós s'ha fet servir el mètode d'agregació de Ward, totes les variables de la base de dades han intervingut en el procés i totes han sigut estadísticament significatives entre els *clústers*.

Finalment, algunes associacions entre variables obtingudes en aquests els mètodes d'ACP i ACM ja s'havien observat abans amb el primer *clustering*. Per exemple, en l'ACM s'ha observat que la categoria d'ocupació "Sector Terciari" està relacionada amb la categoria "With Job/Business" de la variable indicadora de feina del cap de família, fet que en el *profiling* de les classes ja s'havia comprovat també perquè en el grup 2 (Homes de classe alta que treballen en empreses) la majoria treballen en el sector terciari i tenen feina.

## 10. Conclusions generals.

Per finalitzar, es presentaran les conclusions a les que s'ha arribat fent servir aquesta petita mostra de 5000 dades, les quals poden ajudar a definir millor com són les famílies filipines i la seva economia familiar, objectiu principal d'aquest estudi.

- Com resulta obvi, les famílies amb més ingressos són les que més gasten, de més béns disposen i millors cases tenen.

- Les famílies en les que el cap de família no té parella (ja sigui divorciat, vidu, etc) sol tenir més problemes de parella. Aquest fet s'accentua molt si el cap de la llar és una dona.
- Quan la persona treballa en el sector terciari, aquesta no sol tenir ingressos baixa. Al contrari, la pràctica totalitat de treballadors del sector primari són agrícoles i tenen pocs ingressos.
- Les famílies amb ingressos alts però no màxims, són les que més despeses fan en tabac i alcohol (fa pensar que per aquest motiu potser no pertanyen al grup de famílies amb ingressos superiors ja que les famílies de classe alta són les que menys despeses tenen en aquest sector).
- Les famílies més pobres són aquelles amb més joves i infants.
- El més comú en les famílies filipines és que el cap de família sigui un home.

## 11. Pla de treball real.

### • Distribució de tasques.

En el transcurs de la realització del treball s'han produït alguns canvis en la realització de tasques respecte a l'assignació distribuïda inicialment. Això és degut a la disponibilitat de temps dels membres en alguns moments del semestre. Amb la intenció de no aturar la feina, s'ha pres l'acció que es va definir al pla de riscos. És a dir, s'ha reassignat la tasca en qüestió a un altre component del grup. Després, en una tasca posterior, s'ha fet el mateix per tal de repartir-nos la feina entre tots els membres de manera equitativa.

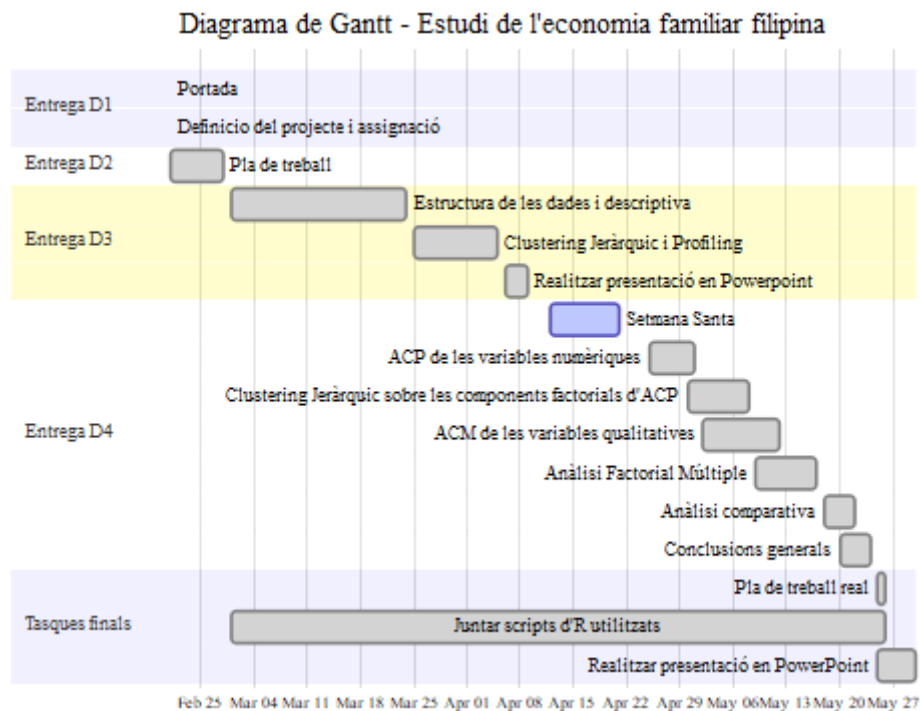
En la taula següent, es pot observar quines són les tasques que ha realitzat cada component del grup:

		Miquel	Laura	Ferran	Victor	Guiu	Guillem	Carles
Entrega D1	Portada					x		
	Definició del projecte i assignació		x	x	x			x
Entrega D2	Pla de treball	x				x	x	x
Entrega D3	Estructura de les dades i descriptiva	x	x	x	x		x	
	Clustering Jeràrquic	x		x	x	x	x	x
Entrega D4	ACP de les variables numèriques	x	x		x		x	
	ACM de les variables qualitatives	x	x	x		x		x
	Clustering Jeràrquic sobre les components factorials d'ACP	x	x	x	x		x	x
	AFM/Discriminant/Textual		x	x	x	x	x	x
	Anàlisi comparativa	x	x	x	x	x	x	
	Conclusions finals		x	x	x		x	x
Tasques finals	Pla de treball real	x		x		x		
	Juntar Scripts d'R utilitzats	x	x	x	x	x	x	x
	Realitzar presentació en PowerPoint	x	x	x	x	x	x	x
		8	8	8	8	8	8	8

Taula 16: Repartiment de les tasques entre els membres del grup.

Les tasques marcades de color vermell, representen aquelles que van ser assignades inicialment a un membre però que finalment no han sigut realitzades per aquest. I en color verd, apareixen les tasques que no havien sigut assignades en la planificació del treball, però que han acabat sent realitzades per aquest membre.

- Diagrama de Gantt



*Figura 68: Diagrama de Gantt com a seqüenciació temporal de les tasques.*

Finalment, no s'ha seguit completament el diagrama de Gantt inicial. Això és degut a que el calendari del qual disposàvem a principi de curs ha patit canvis. Com es pot observar en els gràfics anteriors, La part de l'ACP es va fer després de Setmana Santa degut a que la part teòrica de l'assignatura va ser donada més tard del previst. Un altre canvi important que hem tingut ha sigut la tasca d'ajuntar els diferents scripts utilitzats al llarg del treball. Aquesta tasca estava prevista fer-la tota de cop, un cop fet el treball, però, després hem vist més oportú anar-ho ajuntant cada cop que afegiem nova informació.



## 12. Codi d'R utilitzat.

### ● Definició del projecte i assignació.

```
#1.1 FILTRATGE DADES
bdd_filipines<- read.csv("C:/Users/96gui/Documents/BBDD_filip.csv", header = TRUE, sep = ";")
View(bdd_filipines)
nrow(bdd_filipines)
set.seed(2019)
p <- sample(x = nrow(bdd_filipines), size = 5000)
bdd_final <- bdd_filipines[p,]
summary(bdd_final)

# 1.2 ESTRUCTURA DE LES DADES
summary(bdd_final)
str(bdd_final)
dim(bdd_final)
# reparam el nom de la primera variable
colnames(bdd_final)[1] <- c("Total.Household.Income")
# esborrem els rownames del dataframe
rownames(bdd_final) <- NULL
# escrivim la bdd
write.csv(bdd_final, file="bdd_final.csv")
# missings
sum(is.na(bdd_final))
colSums(is.na(bdd_final))
colSums(is.na(bdd_final))/5000
```

### ● Pla de treball.

```
install.packages("viridisLite")
install.packages("DiagrammeR")
library(viridisLite)
library(DiagrammeR)

m <- mermaid("
  graph TD
    dateFormat YYYY-MM-DD
    title Diagrama de Gantt - Grup 9
    section Entrega D1
      Portada: done, des1, 2019-02-21, 2019-02-21
      Definició del projecte i assignació : done, des2, 2019-02-21, 2019-02-21
    section Entrega D2
      Pla de treball : done, import_1, after des2, 2019-02-28
    section Entrega D3
      Estructura de les dades i descriptiva : done, import_2, 2019-03-01, 2019-03-12
      Clustering Jeràrquic : done, import_3, 2019-03-13, 2019-03-21
    section Entrega D4
      ACP de les variables numèriques : done, import_4, 2019-03-28, 2019-04-12
      Setmana Santa : active, import_5, 2019-04-12, 2019-04-21
      ACM de les variables qualitatives : done, import_6, 2019-04-22, 2019-05-02
      Clustering Jeràrquic sobre les components factorials d'ACP : done, import_7, 2019-05-01, 2019-05-12
      ACM/Discriminant/Textual : done, import_8, 2019-05-13, 2019-05-18
      Anàlisi comparativa : done, import_9, 2019-05-18, 2019-05-23
      Conclusions generals : done, import_10, 2019-05-23, 2019-05-30
    section Tasques finals
      Pla de treball real : done, ex_1, 2019-05-23, 2019-05-30
      Juntar scripts d'R utilitzats: done, ex_2, 2019-05-23, 2019-05-30
      Realitzar presentació en PowerPoint : done, ex_3, 2019-05-23, 2019-05-30
  )
m$x$config = list(ganttConfig = list(axisFormatter = list(list("%b %d",htmlwidgets::JS('function(d){ return d.getDay() == 1 }')))))
m
```

### ● Anàlisis descriptives univariants (inicial i final).

```
# 2. Total.Household.Income:
c(summary(bbdd$Total.Household.Income), Desv.Tip.=sqrt(var(bbdd$Total.Household.Income)))
plot(bbdd$Total.Household.Income, main="Resum Gràfic ", xlab="Observacions", ylab="Ingressos Totals")
lines(1:5000, rep(median(bbdd$Total.Household.Income), 5000), col=c("red"))
lines(1:5000, rep(mean(bbdd$Total.Household.Income), 5000), col=c("yellow"))

# 3. Region
table(bbdd$Region)
prop.table(table(bbdd$Region))*100
regbarplot <- barplot(table(bbdd$Region), names.arg="")
text(regbarplot, par("usr")[3], labels = names(table(bbdd$Region)), srt = 35, adj = c(1.1,1.1), xpd = TRUE, cex=0.6)

# 4. Total.Food.Expenditure
c(summary(bbdd$Total.Food.Expenditure), Desv.Tip.=sqrt(var(bbdd$Total.Food.Expenditure)))
hist(bbdd$Total.Food.Expenditure, xlab = "Despesa en Alimentació", ylab="Frequència", main="Distribució de la Despesa en Alimentació")
lines(rep(median(bbdd$Total.Food.Expenditure), 5000), 1:5000, col=c("red"))
lines(rep(mean(bbdd$Total.Food.Expenditure), 5000), 1:5000, col=c("yellow"))
legend(280000, 2500, c("Mediana", "Mitjana"), c("red", "yellow"))
boxplot(bbdd$Total.Food.Expenditure)

# 5. Main.Source.of.Income
table(bbdd$Main.Source.of.Income)
prop.table(table(bbdd$Main.Source.of.Income))*100
pie(table(bbdd$Main.Source.of.Income))

# 6. Agricultural.Household.indicator
table(bbdd$Agricultural.Household.indicator)
prop.table(table(bbdd$Agricultural.Household.indicator))*100
pie(table(bbdd$Agricultural.Household.indicator))

# 7. Restaurant.and.hotels.Expenditure
c(summary(bbdd$Restaurant.and.hotels.Expenditure), Desv.Tip.=sqrt(var(bbdd$Restaurant.and.hotels.Expenditure)))
hist(bbdd$Restaurant.and.hotels.Expenditure, xlab = "Despesa en Restaurants i Hotels", ylab="Frequència", main="Distribució de la Despesa en Restaurants i Hotels")
lines(rep(median(bbdd$Restaurant.and.hotels.Expenditure), 5000), 1:5000, col=c("red"))
lines(rep(mean(bbdd$Restaurant.and.hotels.Expenditure), 5000), 1:5000, col=c("yellow"))
legend(150000, 3000, c("Mediana", "Mitjana"), c("red", "yellow"))
```

```

boxplot(hbdd$Restaurant.and.hotels.Expenditure)

# 8. Alcohol.and.tobacco.expenditure
c(summary(hbdd$Alcohol.and.tobacco.expenditure),Desv.Tip.=sqrt(var(hbdd$Alcohol.and.tobacco.expenditure)))
hist(hbdd$Alcohol.and.tobacco.expenditure,xlab = "Despesa en Alcohol i tabac",ylab="Frequència",main="Distribució de la Despesa en Alcohol i tabac")
lines(rep(median(hbdd$Alcohol.and.tobacco.expenditure),5000),1:5000,col=c("red"))
lines(rep(mean(hbdd$Alcohol.and.tobacco.expenditure),5000),1:5000,col=c("yellow"))
legend(30000,3000,c("Mediana","Mitjana"),c("red","yellow"))
boxplot(hbdd$Alcohol.and.tobacco.expenditure)

# 9. Clothing,.Footwear.and.Other.Wear.Expenditure
c(summary(hbdd$Clothing,.Footwear.and.Other.Wear.Expenditure),Desv.Tip.=sqrt(var(hbdd$Clothing,.Footwear.and.Other.Wear.Expenditure)))
hist(hbdd$Clothing,.Footwear.and.Other.Wear.Expenditure,xlab = "Despesa en Roba",ylab="Frequència",main="Distribució de la Despesa en Roba")
lines(rep(median(hbdd$Clothing,.Footwear.and.Other.Wear.Expenditure),5000),1:5000,col=c("red"))
lines(rep(mean(hbdd$Clothing,.Footwear.and.Other.Wear.Expenditure),5000),1:5000,col=c("yellow"))
legend(30000,3000,c("Mediana","Mitjana"),c("red","yellow"))
boxplot(hbdd$Clothing,.Footwear.and.Other.Wear.Expenditure)

### 10. Housing.and.water.Expenditure
c(summary(hbdd$Housing.and.water.Expenditure),Desv.Tip.=sqrt(var(hbdd$Housing.and.water.Expenditure)))
hist(hbdd$Housing.and.water.Expenditure,xlab = "Despesa en llar i aigua",ylab="Frequència",main="Distribució de la Despesa en llar i aigua")
lines(rep(median(hbdd$Housing.and.water.Expenditure),5000),1:5000,col=c("red"))
lines(rep(mean(hbdd$Housing.and.water.Expenditure),5000),1:5000,col=c("yellow"))
legend(200000,3000,c("Mediana","Mitjana"),c("red","yellow"))
boxplot(hbdd$Housing.and.water.Expenditure)

# 11. Imputed.House.Rental.Value
c(summary(hbdd$Housing.and.water.Expenditure),Desv.Tip.=sqrt(var(hbdd$Housing.and.water.Expenditure)))
hist(hbdd$Housing.and.water.Expenditure,xlab = "Despesa en llar i aigua",ylab="Frequència",main="Distribució de la Despesa en llar i aigua")
lines(rep(median(hbdd$Housing.and.water.Expenditure),5000),1:5000,col=c("red"))
lines(rep(mean(hbdd$Housing.and.water.Expenditure),5000),1:5000,col=c("yellow"))
legend(200000,3000,c("Mediana","Mitjana"),c("red","yellow"))
boxplot(hbdd$Housing.and.water.Expenditure)

# 12. Medical.education.transport.and.communication.expenditure
c(summary(hbdd$Medical.education.transport.and.communication.expenditure),Desv.Tip.=sqrt(var(hbdd$Medical.education.transport.and.communication.expenditure)))
hist(hbdd$Medical.education.transport.and.communication.expenditure,xlab = "Despesa en Educació, transport i comunicació",ylab="Frequència",main="Distribució de la Despesa en Educació, transport i comunicació")
lines(rep(median(hbdd$Medical.education.transport.and.communication.expenditure),5000),1:5000,col=c("red"))
lines(rep(mean(hbdd$Medical.education.transport.and.communication.expenditure),5000),1:5000,col=c("yellow"))
legend(400000,3000,c("Mediana","Mitjana"),c("red","yellow"))
boxplot(hbdd$Medical.education.transport.and.communication.expenditure)

# 13. Miscellaneous.good.and.special.occasions.expenditure
c(summary(hbdd$Miscellaneous.good.and.special.occasions.expenditure),Desv.Tip.=sqrt(var(hbdd$Miscellaneous.good.and.special.occasions.expenditure)))
hist(hbdd$Miscellaneous.good.and.special.occasions.expenditure,xlab = "Despesa en béns diversos",ylab="Frequència",main="Distribució de béns diversos")
lines(rep(median(hbdd$Miscellaneous.good.and.special.occasions.expenditure),5000),1:5000,col=c("red"))
lines(rep(mean(hbdd$Miscellaneous.good.and.special.occasions.expenditure),5000),1:5000,col=c("yellow"))
legend(400000,3000,c("Mediana","Mitjana"),c("red","yellow"))
boxplot(hbdd$Miscellaneous.good.and.special.occasions.expenditure)

# 14. Crop.Farming.and.Gardening.expenses
c(summary(hbdd$Crop.Farming.and.Gardening.expenses),Desv.Tip.=sqrt(var(hbdd$Crop.Farming.and.Gardening.expenses)))
hist(hbdd$Crop.Farming.and.Gardening.expenses,xlab = "Despesa en productes de granja i jardineria",ylab="Frequència",main="Distribució de productes de granja i jardineria")
lines(rep(median(hbdd$Crop.Farming.and.Gardening.expenses),5000),1:5000,col=c("red"))
lines(rep(mean(hbdd$Crop.Farming.and.Gardening.expenses),5000),1:5000,col=c("yellow"))
legend(300000,3000,c("Mediana","Mitjana"),c("red","yellow"))
boxplot(hbdd$Crop.Farming.and.Gardening.expenses)

# 15. Total.Income.from.Entrepreneurial.Activities
c(summary(hbdd$Total.Income.from.Entrepreneurial.Activities),Desv.Tip.=sqrt(var(hbdd$Total.Income.from.Entrepreneurial.Activities)))
hist(hbdd$Total.Income.from.Entrepreneurial.Activities,xlab = "Ingresos d'activitats laborals",ylab="Frequència",main="Distribució dels ingressos d'activitats laborals")
lines(rep(median(hbdd$Total.Income.from.Entrepreneurial.Activities),5000),1:5000,col=c("red"))
lines(rep(mean(hbdd$Total.Income.from.Entrepreneurial.Activities),5000),1:5000,col=c("yellow"))
legend(300000,3000,c("Mediana","Mitjana"),c("red","yellow"))
boxplot(hbdd$Total.Income.from.Entrepreneurial.Activities)

# 16. Household.Head.Sex
table(hbdd$Household.Head.Sex)
prop.table(table(hbdd$Household.Head.Sex))*100
lbls<-paste(c("Dones","Homes"),prop.table(table(hbdd$Household.Head.Sex))*100)
lbls <- paste(lbls,"%",sep="")
pie(table(hbdd$Household.Head.Sex),labels =lbls)

# 17. Household.Head.Age
c(summary(hbdd$Household.Head.Age),Desv.Tip.=sqrt(var(hbdd$Household.Head.Age)))
hist(hbdd$Household.Head.Age,xlab = "Edat del cap de família",ylab="Frequència",main="Distribució de les edats dels caps de família")
#lines(rep(median(hbdd$Household.Head.Age),5000),1:5000,col=c("red"))
#lines(rep(mean(hbdd$Household.Head.Age),5000),1:5000,col=c("yellow"))
#legend("topright",c("Mediana","Mitjana"),c("red","yellow"))
boxplot(hbdd$Household.Head.Age) ## Millor!

# 18. Household.Head.Marital.Status
table(hbdd$Household.Head.Marital.Status)
prop.table(table(hbdd$Household.Head.Marital.Status))*100
lbls<-paste(c("Null","Divorciat/da","Casat/da","Solter/a","Desconegut/da","Vidu/a"),prop.table(table(hbdd$Household.Head.Marital.Status))*100)
lbls <- paste(lbls,"%",sep="")
pie(table(hbdd$Household.Head.Marital.Status),labels =lbls)
barplot18 <- barplot(table(hbdd$Household.Head.Marital.Status), names.arg="")
text(barplot18, par("usr")[3], labels = lbls, srt = 35, adj = c(1.1,1.1), xpd = TRUE, cex=0.9)

# 19. Household.Head.Highest.Grade.Completed
table(hbdd$Household.Head.Highest.Grade.Completed)
prop.table(table(hbdd$Household.Head.Highest.Grade.Completed))*100
pie(table(hbdd$Household.Head.Highest.Grade.Completed))
barplot19 <- barplot(table(hbdd$Household.Head.Highest.Grade.Completed), names.arg="")
text(barplot19, par("usr")[3], labels = names(table(hbdd$Household.Head.Highest.Grade.Completed)), srt = 35, adj = c(1.1,1.1), xpd = TRUE, cex=0.9)
## Seria necessaria l'agrupacio de variables categoriques, massa nivells.

# 20. Household.Head.Job.or.Business.Indicator
table(hbdd$Household.Head.Job.or.Business.Indicator)
prop.table(table(hbdd$Household.Head.Job.or.Business.Indicator))*100
lbls<- paste(c("Sense treball/empresa","Amb treball/empresa"),prop.table(table(hbdd$Household.Head.Job.or.Business.Indicator))*100)
lbls <- paste(lbls,"%",sep="")
pie(table(hbdd$Household.Head.Job.or.Business.Indicator),labels =lbls)

# 21. Household.Head.Occupation
table(hbdd$Household.Head.Occupation)
prop.table(table(hbdd$Household.Head.Occupation))*100
pie(table(hbdd$Household.Head.Occupation))
barplot21 <- barplot(table(hbdd$Household.Head.Occupation), names.arg="")
text(barplot21, par("usr")[3], labels = names(table(hbdd$Household.Head.Occupation)), srt = 35, adj = c(1.1,1.1), xpd = TRUE, cex=0.9)
## Seria necessaria l'agrupacio de variables categoriques, massa nivells.

# 22. Household.Head.Class.of.Worker
table(hbdd$Household.Head.Class.of.Worker)
prop.table(table(hbdd$Household.Head.Class.of.Worker))*100
lbls<- paste(c("Empleat/da empresa familiar","Autònom/a","Funcionari/a","Empleat/da","Empleat/da de la llar","Assalariat/da en empresa familiar","Voluntari/a en empresa familiar"),round(prop.table(table(hbdd$Household.Head.Class.of.Worker))*100,2))

```

```

lbls <- paste(lbls,"%",sep="")
pie(table(bbdd$Household.Head.Class.of.Worker),labels =lbls)
barplot22 <- barplot(table(bbdd$Household.Head.Class.of.Worker), names.arg="")
text(barplot22, par("usr")[3], labels = lbls, srt = 35, adj = c(1.1,1.1), xpd = TRUE, cex=0.9)

# 23. Total.Number.of.Family.members
## Com a categoria:
table(bbdd$Total.Number.of.Family.members)
prop.table(table(bbdd$Total.Number.of.Family.members))*100
lbls<-paste(c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15),round(prop.table(table(bbdd$Total.Number.of.Family.members))*100,2))
lbls <- paste(lbls,"%",sep="")
pie(table(bbdd$Total.Number.of.Family.members),labels =lbls)
barplot23 <- barplot(table(bbdd$Total.Number.of.Family.members), names.arg="")
text(barplot23, par("usr")[3], labels = lbls, srt = 35, adj = c(1.1,1.1), xpd = TRUE, cex=0.9)
## Com a numérica:
c(summary(bbdd$Total.Number.of.Family.members),Desv.Tip.=sqrt(var(bbdd$Total.Number.of.Family.members)))
hist(bbdd$Total.Number.of.Family.members,xlab = "Edat del cap de familia",ylab="Frequència",main="Distribució de les edats dels caps de família")
lines(rep(median(bbdd$Total.Number.of.Family.members),5000),1:5000,col=c("red"))
lines(rep(mean(bbdd$Total.Number.of.Family.members),5000),1:5000,col=c("yellow"))
legend("topright",c("Mediana","Mitjana"),c("red","yellow"))
boxplot(bbdd$Total.Number.of.Family.members)

## 24. Members.with.age.less.than.5.year.old
table(bbdd$Members.with.age.less.than.5.year.old)
prop.table(table(bbdd$Members.with.age.less.than.5.year.old))*100
lbls<-paste(c("1","2","3","4","5"),round(prop.table(table(bbdd$Total.Number.of.Family.members))*100,2))
lbls <- paste(lbls,"%",sep="")
barplot24 <- barplot(table(bbdd$Members.with.age.less.than.5.year.old), names.arg="")
text(barplot24, par("usr")[3], labels = lbls, srt = 35, adj = c(1.1,1.1), xpd = TRUE, cex=0.9)

# 25. Members.with.age.5...17.years.old
table(bbdd$Members.with.age.5...17.years.old)
prop.table(table(bbdd$Members.with.age.5...17.years.old))*100
lbls<-paste(c("0","1","2","3","4","5","6","7","8"),round(prop.table(table(bbdd$Members.with.age.5...17.years.old))*100,2))
lbls <- paste(lbls,"%",sep="")
barplot25 <- barplot(table(bbdd$Members.with.age.5...17.years.old), names.arg="")
text(barplot25, par("usr")[3], labels = lbls, srt = 35, adj = c(1.1,1.1), xpd = TRUE, cex=0.9)

# 26. Total.number.of.family.members.employed
table(bbdd$Total.number.of.family.members.employed)
prop.table(table(bbdd$Total.number.of.family.members.employed))*100
lbls<-paste(c("0","1","2","3","4","5","6","7","8"),round(prop.table(table(bbdd$Total.number.of.family.members.employed))*100,2))
lbls <- paste(lbls,"%",sep="")
barplot26 <- barplot(table(bbdd$Total.number.of.family.members.employed), names.arg="")
text(barplot26, par("usr")[3], labels = lbls, srt = 35, adj = c(1.1,1.1), xpd = TRUE, cex=0.9)

# 27. Type.of.Building.House
table(bbdd$Type.of.Building.House)
prop.table(table(bbdd$Type.of.Building.House))*100
lbls<- paste(c("Edifici Industrial/Agrícola","Duplex","Barri d'empresa","Residencia","Casa Unifamiliar",round(prop.table(table(bbdd$Type.of.Building.House))*100,2))
lbls <- paste(lbls,"%",sep="")
barplot27 <- barplot(table(bbdd$Type.of.Building.House), names.arg="")
text(barplot27, par("usr")[3], labels = lbls, srt = 35, adj = c(1.1,1.1), xpd = TRUE, cex=0.5)

# 28. House.Floor.Area
c(summary(bbdd$House.Floor.Area,Desv.Tip.=sqrt(var(bbdd$House.Floor.Area))))
hist(bbdd$House.Floor.Area,xlab = "Superfície del habitatge",ylab="Frequència",main="Distribució de la superfície dels habitatges")
lines(rep(median(bbdd$House.Floor.Area),5000),1:5000,col=c("red"))
lines(rep(mean(bbdd$House.Floor.Area),5000),1:5000,col=c("yellow"))
legend(300,3000,c("Mediana","Mitjana"),c("red","yellow"))

# 29. House.Age
c(summary(bbdd$House.Age),Desv.Tip.=sqrt(var(bbdd$House.Age)))
hist(bbdd$House.Age,xlab = "Anys de la Casa",ylab="Frequència",main="Distribució dels anys de les cases")
lines(rep(median(bbdd$House.Age),5000),1:5000,col=c("red"))
lines(rep(mean(bbdd$House.Age),5000),1:5000,col=c("yellow"))
legend(60,2000,c("Mediana","Mitjana"),c("red","yellow"))

# 30. Number.of.bedrooms
table(bbdd$Number.of.bedrooms)
prop.table(table(bbdd$Number.of.bedrooms))*100
barplot(table(bbdd$Number.of.bedrooms))

# 31. Electricity
table(bbdd$Electricity)
prop.table(table(bbdd$Electricity))*100
lbls<- paste(c("No Electricitat","Electricitat"),prop.table(table(bbdd$Electricity))*100)
lbls <- paste(lbls,"%",sep="")
pie(table(bbdd$Electricity),labels =lbls)

# 32. Number.of.Car..Jeep..Van
table(bbdd$Number.of.Car..Jeep..Van)
prop.table(table(bbdd$Number.of.Car..Jeep..Van))*100
barplot(table(bbdd$Number.of.Car..Jeep..Van))

# 33. Number.of.Cellular.phone
table(bbdd$Number.of.Cellular.phone)
prop.table(table(bbdd$Number.of.Cellular.phone))*100
barplot(table(bbdd$Number.of.Cellular.phone))

# 34. Number.of.Motorcycle.or.Tricycle
table(bbdd$Number.of.Motorcycle.Tricycle)
prop.table(table(bbdd$Number.of.Motorcycle.Tricycle))*100
barplot(table(bbdd$Number.of.Motorcycle.Tricycle))

```

## ● Preprocessament de dades.

```

#1. Importació de Dades
bbdd <- read.csv("bdd_final_revisada.csv",sep=",")

#2. Dimensions BDD
class(bbdd)
dim(bbdd)
colnames(bbdd)
str(bbdd)

#3. Primer Control
summary(bbdd)
summary(bbdd$Total.Household.Income) # Variable resposta
hist(bbdd$Total.Household.Income)
# Només les variables Household.Head.Class.of.Worker i #Household.Head.Occupation donen NA's, i ho fan en les mateixes #entries.

#4. Categoritzar
sapply(bbdd, class)
# Agricultural.Household.indicator surt com a integer però ha de ser #categòrica.La resta de variables són ja la classe que els #correspon.
bbdd$Agricultural.Household.indicator <- as.factor(bbdd$Agricultural.Household.indicator)
levels(bbdd$Agricultural.Household.indicator)

```

```
#5. Modalitats
# Amb la funció str() s'ha vist que les variables #Household.Head.Occupation i Household.Head.Highest.Grade.Completed #tenen 243 i 46 levels
respectivament. Es procedeix a remodelar #aquestes dues variables.

#5.1. Variable Household.Head.Highest.Grade.Completed
for (i in 1:nrow(bbdd)){
  if (bbdd$Household.Head.Highest.Grade.Completed[i] == "High School Graduate" || bbdd$Household.Head.Highest.Grade.Completed[i] == "No Grade Completed"
  || bbdd$Household.Head.Highest.Grade.Completed[i] == "First Year College" || bbdd$Household.Head.Highest.Grade.Completed[i] == "Second Year
College"
  || bbdd$Household.Head.Highest.Grade.Completed[i] == "Third Year College" || bbdd$Household.Head.Highest.Grade.Completed[i] == "Fourth Year
College"){
    bbdd$Household.Head.Highest.Grade.Completed.new[i] = "High School Studies"
  }else if (bbdd$Household.Head.Highest.Grade.Completed[i] == "Elementary Graduate" || bbdd$Household.Head.Highest.Grade.Completed[i] == "First Year
High School"
  || bbdd$Household.Head.Highest.Grade.Completed[i] == "Second Year High School"
  || bbdd$Household.Head.Highest.Grade.Completed[i] == "Third Year High School"
  || bbdd$Household.Head.Highest.Grade.Completed[i] == "Fourth Year High School") {

    bbdd$Household.Head.Highest.Grade.Completed.new[i] = "Elementary Studies"
  }else if (bbdd$Household.Head.Highest.Grade.Completed[i] == "Agriculture" || bbdd$Household.Head.Highest.Grade.Completed[i] == "Preschool" ||
bbdd$Household.Head.Highest.Grade.Completed[i] == "Grade 1" || bbdd$Household.Head.Highest.Grade.Completed[i] == "Grade 2"
  || bbdd$Household.Head.Highest.Grade.Completed[i] == "Grade 3" || bbdd$Household.Head.Highest.Grade.Completed[i] == "Grade 4"
  || bbdd$Household.Head.Highest.Grade.Completed[i] == "Grade 5" || bbdd$Household.Head.Highest.Grade.Completed[i] == "Grade 6") {
    bbdd$Household.Head.Highest.Grade.Completed.new[i] = "No Studies"
  } else {
    bbdd$Household.Head.Highest.Grade.Completed.new[i] = "Program Studies"
  }
}
table(bbdd$Household.Head.Highest.Grade.Completed.new)

#5.2. Variable Household.Head.Occupation
for (i in 1:nrow(bbdd)){
  if (is.na(bbdd$Household.Head.Occupation[i]) || bbdd$Household.Head.Occupation[i] == " "){
    bbdd$Household.Head.Occupation.new[i] <- NA
  } else if (bbdd$Household.Head.Occupation[i] == "Rice farmers" || bbdd$Household.Head.Occupation[i] == "Root crops farmers"
  || bbdd$Household.Head.Occupation[i] == "Forestry laborers" || bbdd$Household.Head.Occupation[i] == "Fishery laborers and helpers"
  || bbdd$Household.Head.Occupation[i] == "Corn farmers" || bbdd$Household.Head.Occupation[i] == "Coffee and cacao farmers"
  || bbdd$Household.Head.Occupation[i] == "Fruit" || bbdd$Household.Head.Occupation[i] == "Garbage collectors"
  || bbdd$Household.Head.Occupation[i] == "Mining and quarrying laborers"
  || bbdd$Household.Head.Occupation[i] == "Motorized farm and forestry plant operators"
  || bbdd$Household.Head.Occupation[i] == "Other plant growers" || bbdd$Household.Head.Occupation[i] == "Other livestock farmers"
  || bbdd$Household.Head.Occupation[i] == "Other field crop farmers" || bbdd$Household.Head.Occupation[i] == "Cattle and dairy farmers"
  || bbdd$Household.Head.Occupation[i] == "Chicken farmers" || bdd$Household.Head.Occupation[i] == "Coconut farmers"
  || bbdd$Household.Head.Occupation[i] == "Cotton and fiber crops farmers" || bdd$Household.Head.Occupation[i] == "Deep-sea fishermen"
  || bdd$Household.Head.Occupation[i] == "Field legumes farmers" || bdd$Household.Head.Occupation[i] == "Forest tree planters"
  || bdd$Household.Head.Occupation[i] == "Fruit tree farmers" || bdd$Household.Head.Occupation[i] == "Hog raising farmers"
  || bdd$Household.Head.Occupation[i] == "Inland and coastal waters fishermen" || bdd$Household.Head.Occupation[i] == "Miners and quarry
workers"
  || bdd$Household.Head.Occupation[i] == "Minor forest products gatherers" || bdd$Household.Head.Occupation[i] == "Ornamental plant
growers"
  || bdd$Household.Head.Occupation[i] == "Other aqua products cultivators" || bdd$Household.Head.Occupation[i] == "Other orchard farmers"
  || bdd$Household.Head.Occupation[i] == "Other poultry farmers" || bdd$Household.Head.Occupation[i] == "Production and operations managers
in agriculture"
  || bdd$Household.Head.Occupation[i] == "Seaweeds cultivators" || bdd$Household.Head.Occupation[i] == "Sugarcane farmers"
  || bdd$Household.Head.Occupation[i] == "Vegetable farmers" || bdd$Household.Head.Occupation[i] == "Wood processing plant operators"
  || bdd$Household.Head.Occupation[i] == "Farmhands and laborers" || bdd$Household.Head.Occupation[i] == "Tree nut farmers"){
    bdd$Household.Head.Occupation.new[i] = "Primary Sector"
  }else if (bbdd$Household.Head.Occupation[i] == "Brewers and wine and other beverage machine operators"
  || bdd$Household.Head.Occupation[i] == "Building and related electricians"
  || bdd$Household.Head.Occupation[i] == "Building construction laborers"
  || bdd$Household.Head.Occupation[i] == "Cement and other mineral products machine operators"
  || bdd$Household.Head.Occupation[i] == "Fishery laborers and helpers"
  || bdd$Household.Head.Occupation[i] == "Industrial robot operators" || bdd$Household.Head.Occupation[i] == "Metal finishing"
  || bdd$Household.Head.Occupation[i] == "Metal finishing" || bdd$Household.Head.Occupation[i] == "Wood and related products assemblers"
  || bdd$Household.Head.Occupation[i] == "Wood products machine operators"
  || bdd$Household.Head.Occupation[i] == "Earth-moving and related plant operators"
  || bdd$Household.Head.Occupation[i] == "Freight handlers" || bdd$Household.Head.Occupation[i] == "Hand launderers and pressers"
  || bdd$Household.Head.Occupation[i] == "Marine craft mechanics" || bdd$Household.Head.Occupation[i] == "Masons and related concrete
finishers"
  || bdd$Household.Head.Occupation[i] == "Metal" || bdd$Household.Head.Occupation[i] == "Production and operations managers in construction"
  || bdd$Household.Head.Occupation[i] == "Sheet-metal workers" || bdd$Household.Head.Occupation[i] == "Textile"
  || bdd$Household.Head.Occupation[i] == "Sewers" || bdd$Household.Head.Occupation[i] == "Weavers") {
    bdd$Household.Head.Occupation.new[i] = "Secondary Sector"
  }else {
    bdd$Household.Head.Occupation.new[i] = "Tertiary Sector"
  }
}
table(bbdd$Household.Head.Occupation.new)

# Posem els nous valors a les variables corresponents i s'eliminen #les variables provisionals que s'havien fet amb anterioritat.
bbdd$Household.Head.Highest.Grade.Completed <- as.factor(bbdd$Household.Head.Highest.Grade.Completed.new)
bbdd$Household.Head.Occupation <- as.factor(bbdd$Household.Head.Occupation.new)
bbdd$Household.Head.Highest.Grade.Completed.new <- NULL
bbdd$Household.Head.Occupation.new <- NULL
str(bbdd)

### 6. Tractament de Missings
bd_NA <- bdd[NAS,] # ens quedem només amb les files que tenen missings
head(bd_NA) # A simple vista sembla que totes les files amb missings corresponen a individus que no treballen

# Ho confirmem amb el següent gràfic:
tab<- table(bd_NA$Household.Head.Job.or.Business.Indicator)

par(oma=c(1,1,1,1),cex.main=0.75)
bp<-barplot(tab, main="Distribució NA's")
text(bp, 0, round(tab, 1), cex=1, pos=3)
# Procedim a #codificar-les com a un level de més: "No Occupation" per a #Household.Head.Occupation i "Unemployed" per a Household.Head.Class.#.of.Worker.
levels(bbdd$Household.Head.Class.of.Worker)<-c(levels(bdd$Household.Head.Class.of.Worker),"Unemployed")
bbdd$Household.Head.Class.of.Worker[is.na(bdd$Household.Head.Class.of.Worker)] <- "Unemployed"
levels(bdd$Household.Head.Occupation)<-c(levels(bdd$Household.Head.Occupation),c("No Occupation"))
bbdd$Household.Head.Occupation[is.na(bdd$Household.Head.Occupation)] <- c("No Occupation")

# 7. Escrivim la nova base de dades ja processada.
write.csv(bbdd, file= "bdd_preprocessed.csv", row.names = FALSE)
```

## ● Clustering jeràrquic.

```
setwd("C:/Users/laura.julia/Desktop")
dd <- read.csv("bdd_preprocessed.csv")
dim(dd)
summary(dd)
attach(dd)

library(cluster) # CLUSTERING JERÀRQUIC

# Dissimilarity matrix
actives<-c(1:34) # variables que volem utilitzar
n <- 5000 # nombre d'observacions
filtro<- c(1:n) # totes
```

```
dissimMatrix <- daisy(dd[filtro,actives], metric = "gower", stand=TRUE) # calculem matriu de distàncies utilitzant mètode de gower
distMatrix<-dissimMatrix^2 # matriu de distàncies nova

# Mètode de ward "ward.D2" important!!!!
h1 <- hclust(distMatrix,method="ward.D2") # NOTICE THE COST
plot(h1) # Dendograma

k <- 4 #mirar el gràfic per decidir-ho
c2 <- cutree(h1,k) #cutree fa talls a l'arbre d'hclust i genera una columna
dd[,35]<-c2 #afegim la columna identificadora del cluster a la base de dades

table(c2) #class sizes, podem veure si les classes estan equilibrades
```

## ● Profiling.

```
dades<-dd #dades contain the dataset
K<-dim(dades)[2] # nombre de variables
par(ask=TRUE, cex.main=0.75) # per a que en le bucle de després vagi fent les coses poc a poc
P<-dd[,35] # la última variable (creada en el clustering) és la variable de classe, ara P.
nc<-length(levels(factor(P)))
nameP<-"Class"

## 1. Tests per veure la significació de les variables ente clústers.
for(k in 1:K){
  if (is.numeric(dades[,k]) == "integer"){
    print(paste("Anàlisi per classes de la Variable:", names(dades)[k]))
    c<-oneway.test(dades[,k]~P)
    print(paste("p-valueANOVA:", o$p.value))
    kw<-kruskal.test(dades[,k]~P)
    print(paste("p-value Kruskal-Wallis:", kw$p.value))
  }else{
    #qualitatives
    print(paste("Variable qualitativa", names(dades)[k]))

    print("Test Chi quadrat: ")
    print(chisq.test(dades[,k], as.factor(P))$p.value)
  }
}

## 2. Mètodes gràfics per al profiling
for(k in 2:34){
  if (is.numeric(dades[,k])){
    print(paste("Anàlisi per classes de la Variable:", names(dades)[k]))
    boxplot(dades[,k]~P, main=paste("Boxplot of", names(dades)[k], "vs", nameP ), horizontal=TRUE)

    barplot(tapply(dades[,k], P, mean),main=paste("Means of", names(dades)[k], "by", nameP ))
    abline(h=mean(dades[,k]))
    legend(0,mean(dades[,k]),"global mean",bty="n")
  }else{
    print(paste("Variable", names(dades)[k])) # qualitatives
    table<-table(P,dades[,k])
    rowperc<-prop.table(table,1)
    colperc<-prop.table(table,2)
    dades[,k]<-as.factor(dades[,k])

    marg <- table(as.factor(P))/n
    print(append("Categories=",levels(as.factor(dades[,k]))))

    # Snake plot
    plot(marg,type="l",ylim=c(0,1),main=paste("Snake plot of",names(dades)[k]))
    paleta<-rainbow(length(levels(dades[,k])))
    for(c in 1:length(levels(dades[,k]))){lines(colperc[,c],col=paleta[c]) }
    legend("topright", levels(dades[,k]), col=paleta, lty=2, cex=0.6)

    #diagrames de barres apilades
    paleta<-rainbow(length(levels(dades[,k])))
    barplot(table(dades[,k], as.factor(P)), beside=FALSE,col=paleta,main=paste("Means of", names(dades)[k]))
    legend("topright",levels(as.factor(dades[,k])),pch=1,cex=0.5, col=paleta)
  }
}
}
```

## ● Anàlisi de Components Principals.

```
#### Anàlisi de Components principals ####
library(FactoMineR)
# Base de dades processada:
dd <- read.table("bdd_preprocessed.csv",header=T, sep=",");
dd <- dd[, -1] # traïem la variable identificadora

# 1. Triem només les variables numèriques:
num <- which(sapply(dd, class) == "numeric"| sapply(dd, class) == "integer")
dcon<- dd[, num]

# 2. Selecció del nombre de dimensions significatives.
pcl <- prcomp(dcon, scale=TRUE)
barplot(100*cumsum(pcl$sdev[1:dim(dcon)[2]]^2)/dim(dcon)[2],space=0, xlab="Dimensió", ylab="% acumulat de variància explicada")
abline(h=80, v=NULL, col = "red1")
axis(1, at = 0:24)

# 3. Fem l'ACP amb la funció PCA de R.
res.pca<-PCA(dd, ncp = 12, quali.sup = c(2,4,15,17,18,19,20,21,26),graph=FALSE)

# 4. Valors propis i percentatge de variància captat a cada dimensió:
res.pca$eig # per quasi bé assolir el llindar del 80% de la variància total de les dades necessitem 10 dimensions.

# 5. Gràfics ACP
# 5.1. Plot dels individus.
plot(res.pca,choix="ind",invisible="quali",axes=c(1,2),select = "contrib 10")
plot(res.pca,choix="ind",invisible="quali",axes=c(3,4),select = "contrib 10")
plot(res.pca,choix="ind",invisible="quali",axes=c(5,6))

# 5.2. Plots de les variables
plot(res.pca,choix="var",axes=c(1,2),select = "cos2 0.6")
plot(res.pca,choix="var",axes=c(3,4),select = "cos2 0.1")
plot(res.pca,choix="var",axes=c(5,6),select = "cos2 0.2")
plot(res.pca,choix="var",axes=c(7,8),select = "cos2 0.1")
plot(res.pca,choix="var",axes=c(9,10),select = "cos2 0.1")
plot(res.pca,choix="var",axes=c(11,12),select = "cos2 0.1")

head(res.pca$var$contrib) # variables que més contribueixen en la inèrcia

# 5.3. Plots dels individus segons algunes variables qualitatives.
```

```

plot(res.pca,habillage=4,col.hab=c("blue","red","green"),invisible="quali",label="none",axes=c(1,2))
plot(res.pca,habillage=4,col.hab=c("blue","red","green"),invisible="quali",label="none",axes=c(3,4))
plot(res.pca,habillage=4,col.hab=c("blue","red","green"),invisible="quali",label="none",axes=c(5,6))
plot(res.pca,habillage=15,col.hab=c("blue","red"),invisible="quali",label="none",axes=c(1,2))
plot(res.pca,habillage=15,col.hab=c("blue","red"),invisible="quali",label="none",axes=c(3,4))
plot(res.pca,habillage=15,col.hab=c("blue","red"),invisible="quali",label="none",axes=c(5,6))
plot(res.pca,habillage=18,invisible="quali",label="none",axes=c(1,2))
plot(res.pca,habillage=18,invisible="quali",label="none",axes=c(3,4))
plot(res.pca,habillage=18,invisible="quali",label="none",axes=c(5,6))

```

## ● Anàlisi de Correspondències Múltiples.

```

#### ANÀLISI DE CORRESPONDÈNCIES MÚLTIPLES ####
# Carreguem els paquets que necessitarem
library(FactoMineR)
library(Matrix)

# 1. Lectura de la base de dades.
dd <- read.table("bdd_preprocessed.csv",header=T, sep=",")
dd <- dd[, -1] # traïem la variable id identificadora
cat <- which(sapply(dd, class) == "factor")
num <- which(sapply(dd, class) == "numeric"| sapply(dd, class) == "integer")

# 2. Fem l'ACM amb la funció MCA de R
res.mca<-MCA(dd, ncp = 5, quanti.sup=c(1,3,5,6,7,8,9,10,11,12,13,14,16,22,23,24,25,27,28,29,30,31,32,33), quali.sup = c(2,15,17,18,26),
method="Burt",,graph=FALSE)
res.mca$eig # considerem 5 dimensions.
barplot(res.mca$eig[,3],main="% acumulat de variància explicada",names.arg=1:nrow(res.mca$eig))
abline(h=80, col="red")

# 4. Anàlisi numèric.
# Contribucions
res.mca$ind$contrib
res.mca$var$contrib # Per veure les categories més contribuents en cada dimensió
round(res.mca$var$contrib,4)

# 5. Anàlisi gràfic.
# Plots de les modalitats de variables actives i complementàries per separat.
plot(res.mca,invisible=c("ind","quali.sup"))
plot(res.mca,invisible=c("ind","var"), xlim=c(-0.5,2.5), ylim=c(-1, 1.5), selectMod = "cos2 10")

# El.lipses de confiança
plotellipses(res.mca,keepvar=c("quali"))
plotellipses(res.mca,keepvar=c("quali.sup"))

# 6. Millorem els grafics
noudd <- dd[, c(4,19,20,21)]
cats = apply(noudd, 2, function(x) nlevels(as.factor(x)))
cats
mca_vars_df = data.frame(res.mca$var$coord, Variable = rep(names(cats), cats))
mca_obs_df = data.frame(res.mca$ind$coord)

library(ggplot2)
ggplot(data = mca_obs_df, aes(x = Dim.1, y = Dim.2)) +
  geom_hline(yintercept = 0, colour = "gray70") +
  geom_vline(xintercept = 0, colour = "gray70") +
  geom_point(colour = "gray50", alpha = 0.7) +
  geom_density2d(colour = "gray80") +
  geom_text(data = mca_vars_df,
    aes(x = Dim.1, y = Dim.2,
      label = rownames(mca_vars_df), colour = Variable)) +
  ggtitle("MCA plot of variables using R package FactoMineR") +
  scale_colour_discrete(name = "Variable")

```

## ● Clustering jeràrquic sobre les components factorials d'ACP.

```

#### CLUSTERING JERÀRQUIC SOBRE ELS COMPONENTS PRINCIPALS ####
# Paquets necessaris:
library(FactoMineR)

# Base de dades processada:
dd <- read.table("bdd_preprocessed.csv",header=T, sep=",");
dd <- dd[, -1] # traïem la variable id identificadora

# 1. Fem l'ACP amb la funció PCA de R (ja sabem que seleccionar 12 dimensions és suficient)
res.pca<-PCA(dd, ncp = 12, quali.sup = c(2,4,15,17,18,19,20,21,26),graph=FALSE)

# 2. Càlcul de distàncies
Psi<-res.pca$ind$coord # coordenades de tots els individus sobre les 12 primeres dimensions
D = dist(Psi) # matriu de distàncies euclídees que hi ha entre els individus.

# 3. Clustering jeràrquic.
hc <- hclust(D, method = "ward.D2")
plot(hc)
barplot(hc$height[1:8],main="Aggregated distance at each iteration")
abline(h=0.345,col="red",lwd=2)
nc = 3 # Nombre de clusters
clus1<- cutree(hc,nc)
table(clus1) # Nombre d'individus a cada clúster

# 4. Mirem si la qualitat de les particions és bona.
# 4.1. Representació dels clusters en el pla principal (dimensions 1 i 2)
plot(Psi, col=as.numeric(clus1),pch=16, main="Representació abans de la consolidació")
abline(h=0,v=0,col="gray")
legend("topright",c("c1","c2","c3"),pch=16,col=c(1:3))

# 4.2. Determinem quina quantitat de la variància correspon a la var entre classes.
cdg <- aggregate(as.data.frame(Psi),list(clus1),mean)[,2:13]
Bss <- sum(rowSums(cdg^2)*as.numeric(table(clus1)))
Tss <- sum(rowSums(Psi^2))

```

```

100*Bss/Tss # 22% del total és de la variabilitat entreclasses, no és molt

# 5. Consolidem les particions
clus2<-kmeans(Psi,centers=cdg) # assignem a cada individu el centre de gravetat més proper que té
clus2$size # Diferent que abans, aquestes seran les classes definitives.
Bss <- sum(rowSums(clus2$centers^2)*clus2$size)
Wss <- sum(clus2$withinss)
100*Bss/(Bss+Wss) # 26.5%, millor!
# 5.1. Individus representats segons el cluster en el primer pla factorial després de la consolidació
plot(Psi, col=as.numeric(clus2$cluster),pch=16,main="Representació després de la consolidació")
abline(h=0,v=0,col="gray")
legend("topright",c("c1","c2","c3"),pch=16,col=c(1:3))

#### PROFILING DE LES CLASSES ####
# 1. Clustering jeràrquic amb la funció HCPC de R (paquet FactoMineR)
res.hcpc<-HCPC(res.pca, consol=TRUE)
summary(res.hcpc$data.clust$clust) # nombre d'individus final a cada classe

# 2. Descripció de les particions.
## 2.1. Descripció per VARIABLES ($desc.var).
### 2.1.1. Variables categòriques
res.hcpc$desc.var$test.chi2 # test d'independència chi-quadrat
res.hcpc$desc.var$category # Descripció de cada cluster

### 2.1.2. Variables numèriques
res.hcpc$desc.var$quanti.var# Significació
res.hcpc$desc.var$quanti # Descripció de cada cluster

# 3. Mètodes gràfics per al profiling.
n <- 5000
dades <- res.hcpc$data.clust
K<-dim(dades)[2] # nombre de variables
par(ask=TRUE, cex.main=0.75) # per a que en le bucle de després vagi fent les coses poc a poc
P<-dades[,34] # la última variable (creada en el clustering) és la variable de classe, ara P.
nc<-length(levels(factor(P)))
nameP<-"Class"

for(k in 1:33){
  if (is.numeric(dades[,k])){
    print(paste("Anàlisi per classes de la Variable:", names(dades)[k]))
    boxplot(dades[,k]~P, main=paste("Boxplot of", names(dades)[k], "vs", nameP ), horizontal=TRUE)

    barplot(tapply(dades[[k]], P, mean),main=paste("Means of", names(dades)[k], "by", nameP ))
    abline(h=mean(dades[[k]]))
    legend(0,mean(dades[[k]]),"global mean",bty="n")
  }else{
    print(paste("Variable", names(dades)[k])) # qualitatives
    table<-table(P,dades[,k])
    rowperc<-prop.table(table,1)
    colperc<-prop.table(table,2)
    dades[,k]<-as.factor(dades[,k])

    marg <- table(as.factor(P))/n
    print(append("Categories=",levels(as.factor(dades[,k]))))

    # Snake plot
    plot(marg,type="l",ylim=c(0,1),main=paste("Snake plot of",names(dades)[k]))
    paleta<-rainbow(length(levels(dades[,k])))
    for(c in 1:length(levels(dades[,k]))){lines(colperc[,c],col=paleta[c]) }
    legend("topright", levels(dades[,k]), col=paleta, lty=2, cex=0.6)

    #diagrames de barres apilades
    paleta<-rainbow(length(levels(dades[,k])))
    barplot(table(dades[,k], as.factor(P)), beside=FALSE,col=paleta,main=paste("Means of", names(dades)[k]))
    legend("topright",levels(as.factor(dades[,k])),pch=1,cex=0.5, col=paleta)
  }
}

```

## ● Anàlisi Factorial Múltiple.

```

# 0. Carguem el paquet FactoMineR
library(FactoMineR)

# 1. Llegim i ordenem la base de dades segons les variables que hi haurà a cada grup
dd <- read.csv("bdd_preprocessed.csv",header= T)

ddafm <- dd[,c(2,4,7:15)]
ddafm <- cbind(ddafm,dd[, c(6,28:34)])
ddafm <- cbind(ddafm,dd[, c(5,19:22)])
ddafm <- cbind(ddafm,dd[, c(17,23:26)])
ddafm <- cbind(ddafm,dd[, c(3,16,18,27)])

# 2. Fem el MFA, els grups suplementaris seran Family info i Other qualitative.
res2 <- MFA(ddafm[,1:33], group=c(11,8,5,5,4), type=c("s","s","n","s","n"),num.group.sup = c(4,5),ncp=17,
            name.group=c("Income and Expenditures", "Objects and home info", "Occupations", "Family Info", "Other qualitative"),graph=F)
summary(res2)
dimdesc(res2)

# 3. Gràfics
# 3.1 Gràfics de grups
plot(res2,choix="group",cex=0.7)

# 3.2 Gràfics de Variables
plot(res2, cex=0.8, invisible="ind", lab.var=F,lab.ind = F,partial="all", title="Graph of the individuals")
plot(res2,choix="var",lab.var=F, invisible="quanti.sup", cex=0.7)

# 3.3 Gràfics de Individus
# 3.3.1 Amb inèrcia de grup
plot(res2,partial="all",lab.var=F,lab.ind = F)

# 3.3.2 Per variable categorica (Household.Head.Occupation) del grup Occupations
plot(res2, invisible="quali", habillage=23, cex=0.8,select="cos2 0.4",legend=list(bty = "y", x = "bottomright"))

```