UNIVERSITAT POLITÈCNICA DE CATALUNYA

Exemple d'aplicació de tècniques d'anàlisi multivariant a l'estudi de l'economia familiar filipina

Data d'entrega: 11/04/2019

Components del grup:
Miquel de Jover Boira (m.dejover@gmail.com)
Laura Julià Melis (laaurajulia@gmail.com)
Ferran Lacasta Roig (ferranlacasta@gmail.com)
Víctor Navarro Garcés (vng1997@gmail.com)
Guiu Puigcercós Vilar (96guiu@gmail.com)
Guillem Querol Llaveria (guillem.querolet@gmail.com)
Carles Requena Sánchez (carlesrequenasanchez@gmail.com)

Índex

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 Expediture.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.

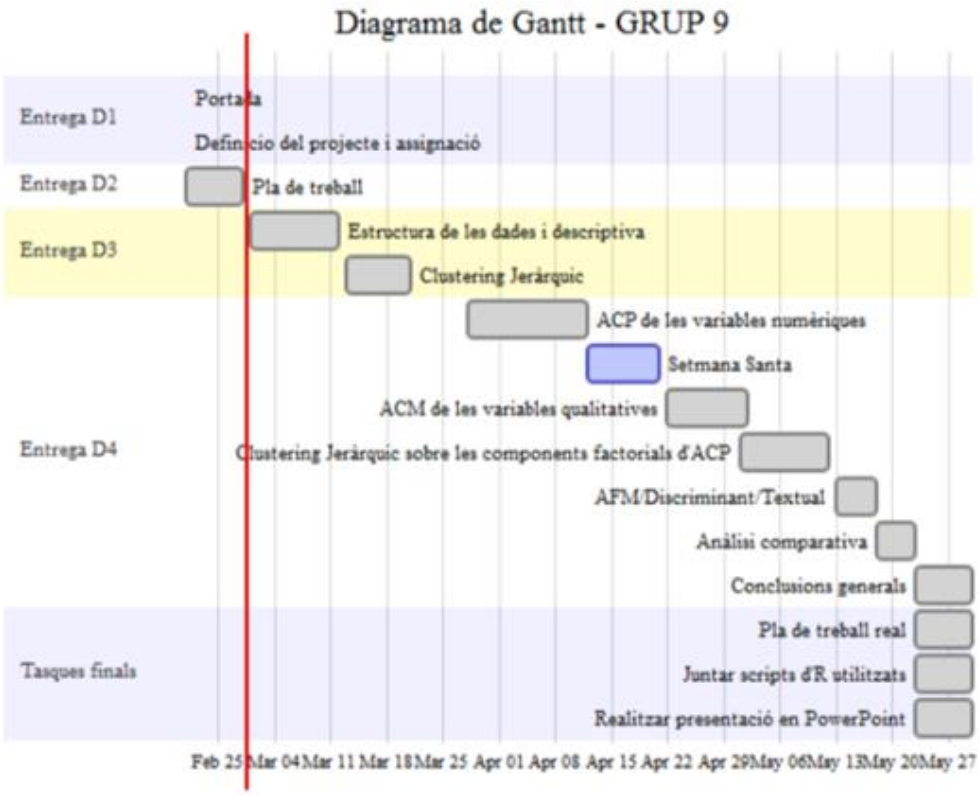


Figura 1: Diagrama de Gantt com a seqüenciació temporal de les tasques.

● 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
Un membre del grup es posa malalt i no pot fer la seva part de la tasca.	Les tasques assignades als membres del grup no seran individuals. Així doncs si un no pot fer-la, l'altre persona a la que se li ha assignat la podrà fer.	S'haurà de reajustar l'assignació de treballs per tal de compensar la pèrdua.

No entregar a temps les tasques.	Fer els scripts a classe i les entregues amb temps.	Seguir el diagrama de Gantt, en el qual ja prevenim possibles contratemps deixant uns dies de marge.
Pèrdua del treball o parts d'aquest.	Compartir-ho entre tots els membres del grup en un espai virtual tots els avenços del treball (Google drive).	Assegurar-nos de que tots els components pujant els avenços al espai virtual.
Entregues o scripts mal realitzats.	Per a cada tasca o entrega hi haurà un grup de 2 persones que s'encarregaren de revisar la feina realitzada.	En el repartiment de tasques també sortiran en cada entrega el grup que s'encarregarà de la revisió de la feina.
Errades ortogràfiques a l'informe.	Fer servir en tot moment el corrector, assegurant-nos del seu bon funcionament.	Designar un encarregat de revisions ortogràfiques per a cada part del informe.
Discussions causades per la diversitat d'opinions entre alguns o tots els membres de l´equip.	Tenir una bona comunicació i respecte a l´hora de discutir els temes referits a la feina.	S´ha d´arribar a un acord amb el qual tots els membres estiguin totalment o parcialment d´acord.
En el cas que es produeixi una vaga d'estudiants o de transport.	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.
Un membre de l'equip decideix renunciar a l'avaluació contínua.	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.
Algun integrant del grup no faci la seva tasca o no posi intères en el treball.	Donar-li tocs d'atenció.	Al final del treball fer una avaluació justa d´aquest membre de l´equip.
Un membre de l'equip no sap fer la seva part del treball.	Quan un membre de l'equip tingui dubtes ho comunicarà a la resta abans que sigui massa tard.	Si es disposa d'algun integrant capaç de dur a terme aquesta tasca se li demanarà col·laboració amb la persona corresponent.
No es respecta el calendari de coordinadors i algú agafa el control contínuament.	Cadascú ha de ser respectuós amb les pautes marcades i intentar no agafar el control quan no li correspon.	El grup haurà de reunir-se i parlar d'aquest problema fins que es solucioni.
El campus virtual no funciona o en el moment d'entrega està en manteniment.	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'economía familiar a Filipines cumplia tots els requisits alhora que ens oferia un tema que satisfeia 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 familia 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 - SOCCKSARGEN*"}
    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).
    a. Rang de valors: {0, 5}
    b. 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
ClothingFootwearandOther.WearExpenditure	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 lo que sembla haver 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.

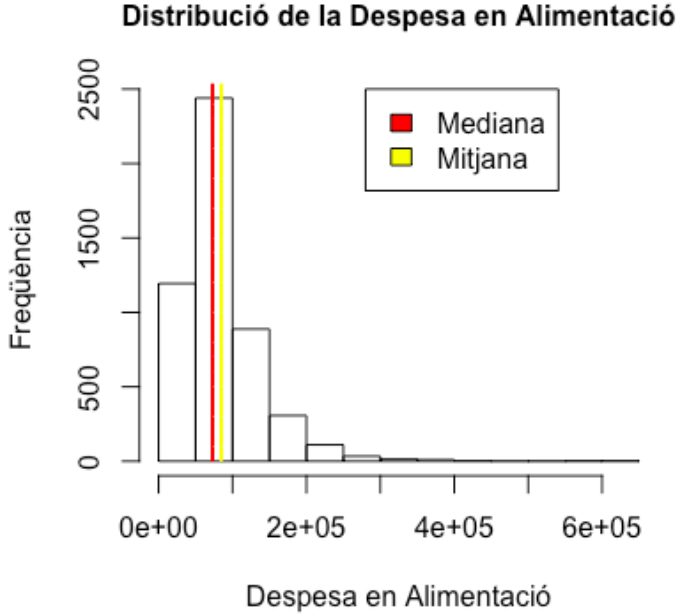
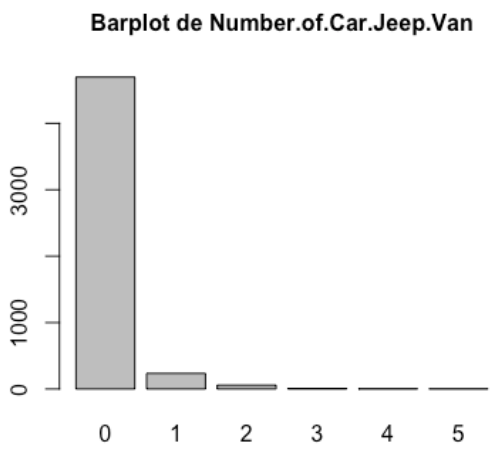
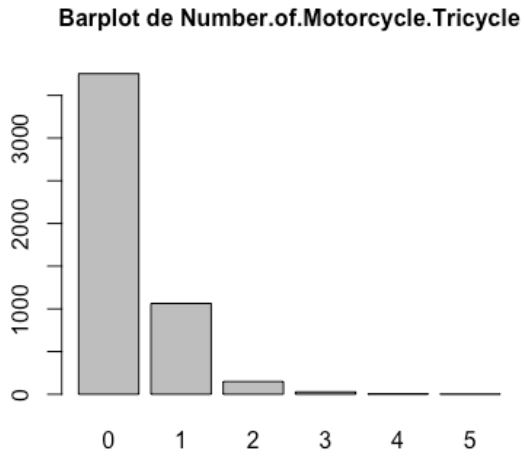


Figura 2: Histograma de la variable ***Total.Food.Expenditure****.*

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) mentres 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.Status	**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.Grade.Completed	**243 nivells**					
Household.Head.Job.or.Business.Indicator	**Sense treball**	874	17.48	**Amb treball**	4126	82.52
Household.Head.Occupation	**46 nivells**					
Household.Head.Class.of.Worker	**Employer in own family-operated farm or business**	305	6.10	**Self-employed wihout 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/agricultural 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 algún nivell molt majoritari que la resta, com "Electricity", que ens indica que un 88% de les famílies disposen d'electricitat:

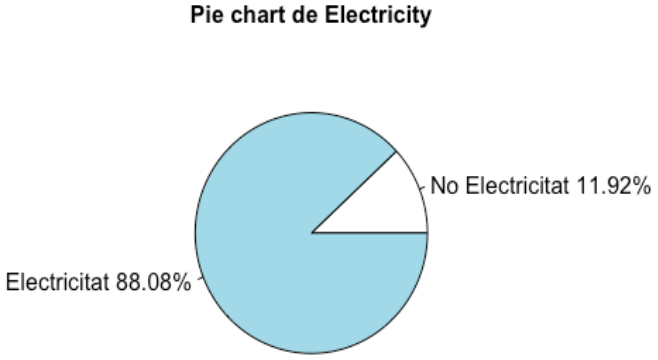


*Figura 5: Diagrama circular de la variable **Elecricity**.*

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 treball.

● 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 243 i 46 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.

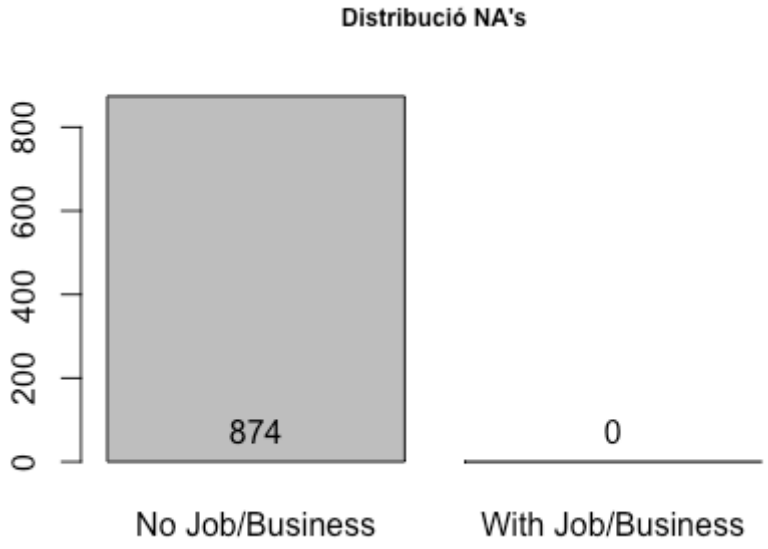


Figura 6: Distribució dels missings en la variable NA's.

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.

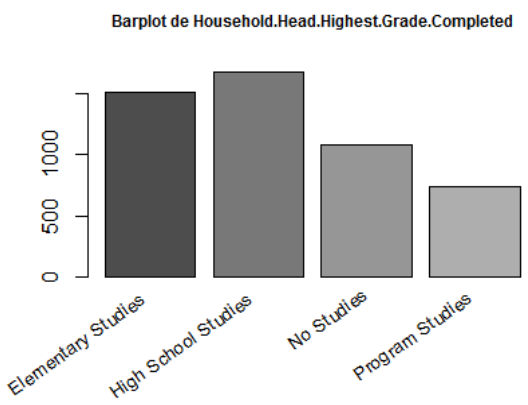
- **Variables Categòriques (ara 11):**

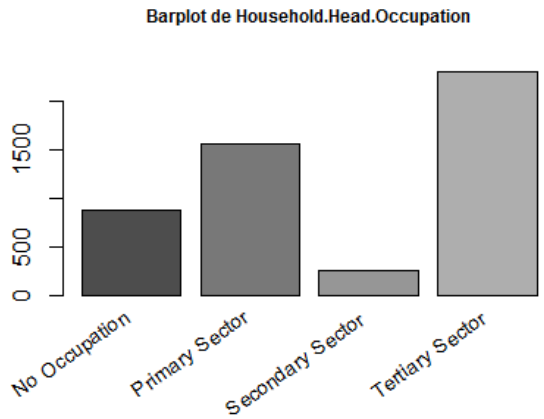
En la taula adjuntada sota aquestes línies es pot veure un resum numèric de les quatre variables que han patit canvis en el *preprocessing*. Ara ja és possible analitzar aquelles variables que abans tenien tants nivells ("Household.Head.Highest. Grade.Completed" i "Household.Head.Occupation").

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 wihout 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 com una minoria dels caps de familia te estudis de grau (*Program Studies*) i com quasi la meitat treballen en el 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 diferenciables, 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 es serveixen dels 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 juntar-se.

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

Es defineix la distància de Gower com $d_{ij}^2 = 1 - s_{ij}$, on s_{ij} 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.
- α = 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 dendograma (Figura 9).

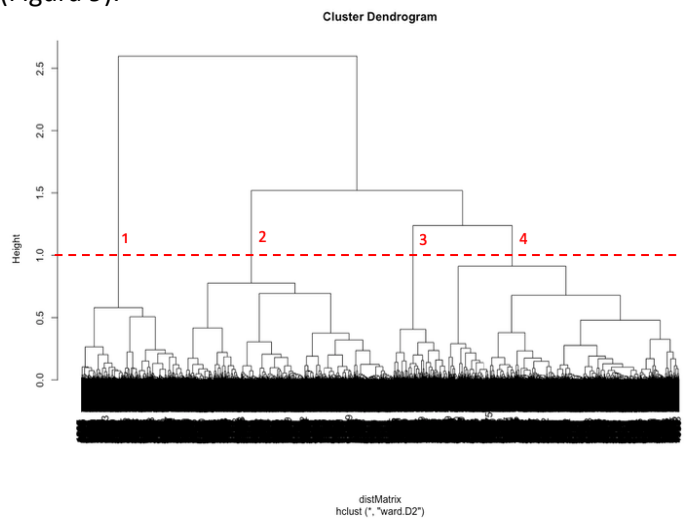


Figura 9: Arbre jeràrquic de les 5000 observacions obtingut amb la funció de R "hclust".

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 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.

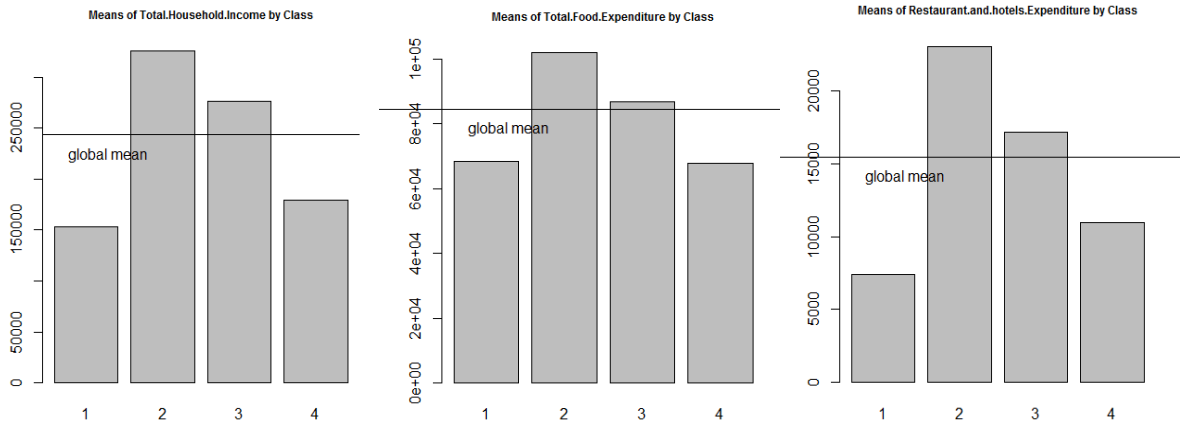
En primer lloc, s'ha realitzat un test *Chi-quadrat* per a comprovar quines diferències en les variables qualitatives són estadísticament significatives entre clúster i, per a les diferències entre variables numèriques, els tests ANOVA o de *Kruskal-Wallis*. Llevat de "id", com és normal ja que és la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

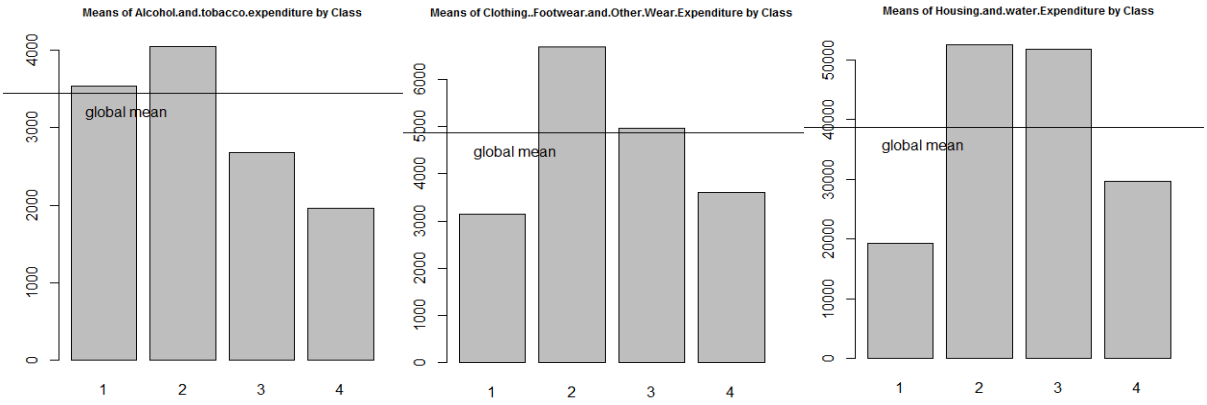
Clúster	Nom	Descripció
1	Homes pagesos de classe baixa	Famílies agrícoles en les que el cap de família és un home sense estudis o amb estudis bàsics. També són famílies amb molts nens petits i uns ingressos baixos (150000 de mitjana). Tenen despeses baixes respecte els altres clúster llevat de la despesa 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.

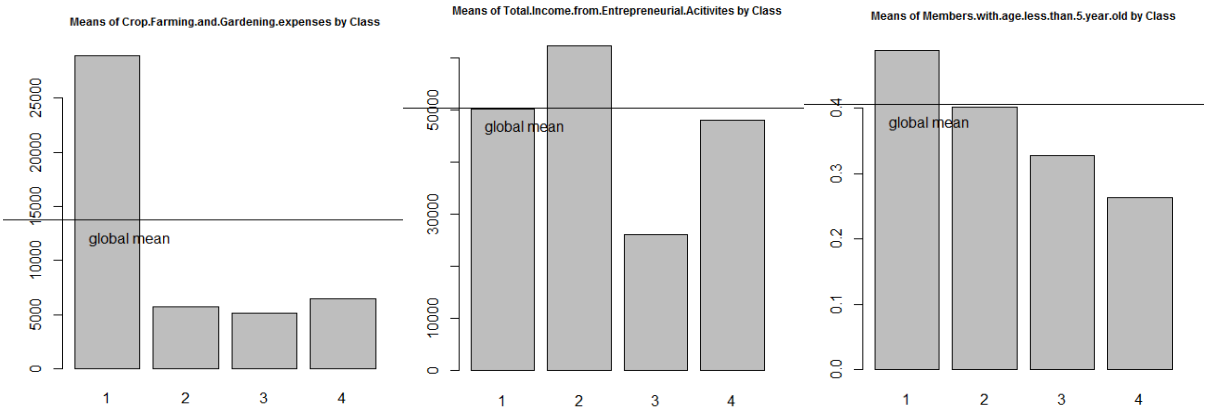
Per acabar, s'adjunten les representacions més rellevants per a la realització dels perfils dels 4 grups descrits en la Taula 7.



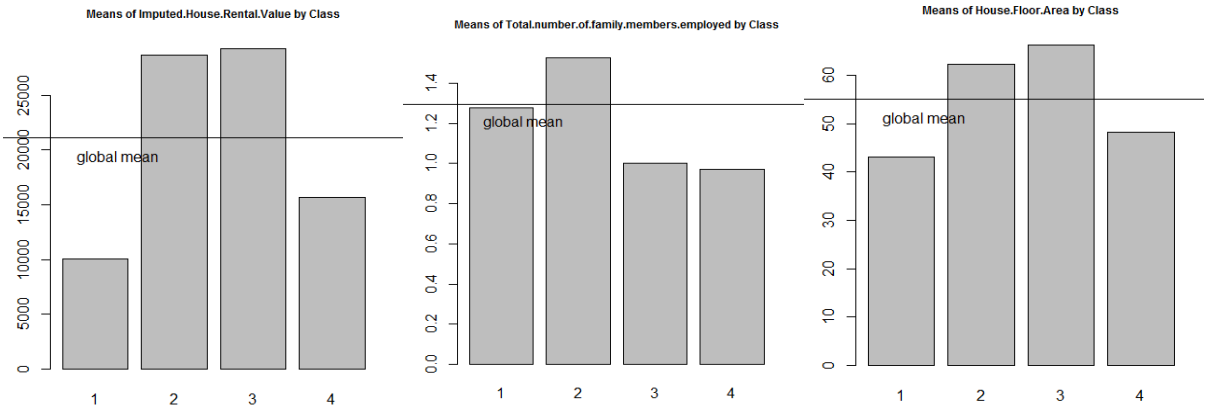
Figures 10-12: Barplots de "Total.Household.Income", "Total.Food.Expenditure" i "Restaurants.and.hotels.Expenditure" per clústers.



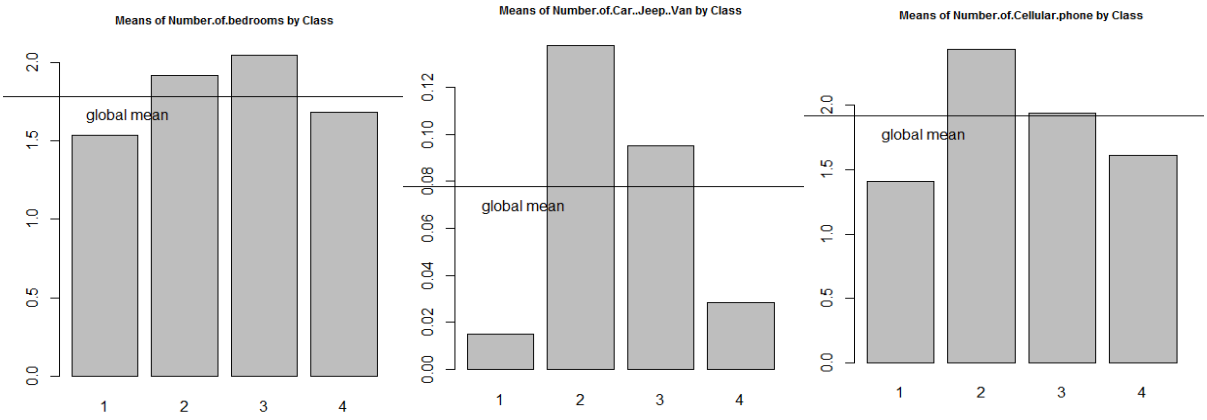
Figures 13-15: Barplots de "Alcohol.and.tobacco.expenditure", "Clothing_Footwear.and.Other.Wear.Expenditure" i "Housing.and.water.Expenditure" per clústers.



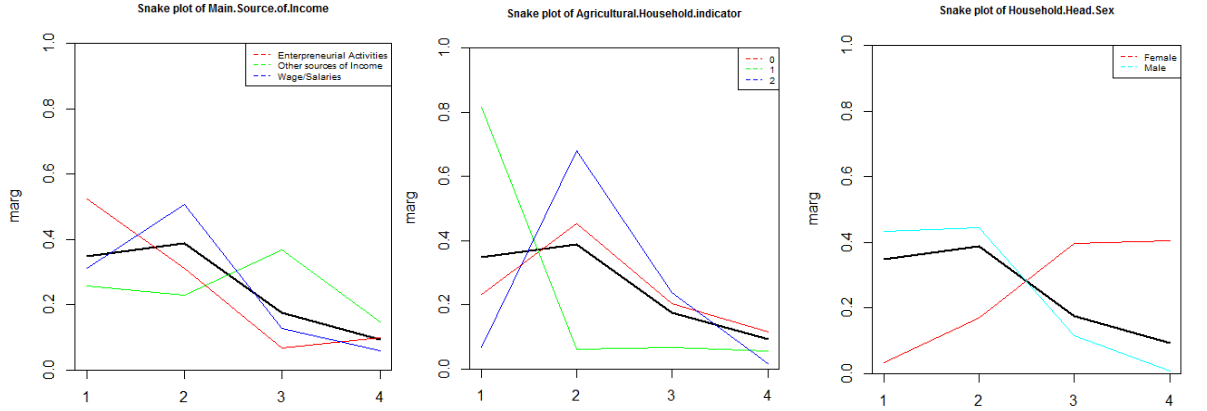
Figures 16-18: Barplots de "Crop.Farming.and.Gardening.expenses", "Total.Income.from.Entrepeneurial.Activities" i "Members.with.age.less.than.5.year.old" per clústers.



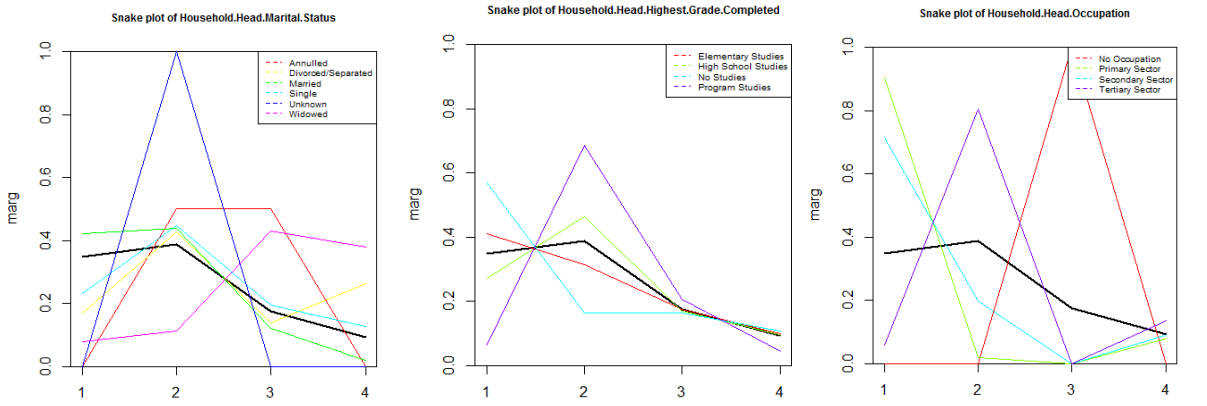
Figures 19-21: Barplots de "Imputed.House.Rental.Value", "Total.number.of.family.members.employed" i "House.Floor.Area" per clústers.



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. 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)
# reparem 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("
  gantt
  dateFormat  YYYY-MM-DD
  title Diagrama de Gantt - Grup 9
 section Entrega D1
  Portada:                          done, des1,    2019-02-21,   2019-02-21
  Definicio 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$Total.Food.Expenditure,xlab = "Despesa en Alimentació",ylab="Freqüència",main="Distribució de la Despesa en Alimentació")
lines(rep(median(bbdd$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$Restaurant.and.hotels.Expenditure,xlab = "Despesa en Restaurants i Hotels",ylab="Freqüència",main="Distribució de la Despesa en Restaurants
i Hotels")
lines(rep(median(bbdd$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(bbdd$Restaurant.and.hotels.Expenditure)

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

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

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

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

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

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

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

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

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

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

# 18. Household.Head.Marital.Status
table(bbdd$Household.Head.Marital.Status)
prop.table(table(bbdd$Household.Head.Marital.Status))*100
lbls<-paste(c("Nul","Divorciat/da","Casat/da","Solter/a","Desconegut/da","Vidu/a"),prop.table(table(bbdd$Household.Head.Marital.Status))*100)
lbls <- paste(lbls,"%",sep="")
pie(table(bbdd$Household.Head.Marital.Status),labels =lbls)
barplot18 <- barplot(table(bbdd$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(bbdd$Household.Head.Highest.Grade.Completed)
prop.table(table(bbdd$Household.Head.Highest.Grade.Completed))*100
pie(table(bbdd$Household.Head.Highest.Grade.Completed))
barplot19 <- barplot(table(bbdd$Household.Head.Highest.Grade.Completed), names.arg="")
text(barplot19, par("usr")[3], labels = names(table(bbdd$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(bbdd$Household.Head.Job.or.Business.Indicator)
prop.table(table(bbdd$Household.Head.Job.or.Business.Indicator))*100
lbls<- paste(c("Sense treball/empresa","Amb treball/empresa"),prop.table(table(bbdd$Household.Head.Job.or.Business.Indicator))*100)
lbls <- paste(lbls,"%",sep="")
pie(table(bbdd$Household.Head.Job.or.Business.Indicator),labels =lbls)

# 21. Household.Head.Occupation
table(bbdd$Household.Head.Occupation)
prop.table(table(bbdd$Household.Head.Occupation))*100
pie(table(bbdd$Household.Head.Occupation))
barplot21 <- barplot(table(bbdd$Household.Head.Occupation), names.arg="")
text(barplot21, par("usr")[3], labels = names(table(bbdd$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(bbdd$Household.Head.Class.of.Worker)
prop.table(table(bbdd$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(bbdd$Household.Head.Class.of.Worker))*100,2))
lbls <- paste(lbls,"%",sep="")
pie(table(bbdd$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 categorica:
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$Total.Number.of.Family.members), names.arg="")
text(barplot23, par("usr")[3], labels = lbls, srt = 35, adj = c(1.1,1.1), xpd = TRUE, cex=0.9)
## Com a numèrica:
c(summary(bbdd$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 família",ylab="Freqüència",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/Agricola","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$House.Floor.Area,xlab = "Superfície del habitatge",ylab="Freqüència",main="Distribució de la superficie dels habitatges")
lines(rep(median(bbdd$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$House.Age,xlab = "Anys de la Casa",ylab="Freqüència",main="Distribució dels anys de les cases")
lines(rep(median(bbdd$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 dónen 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$Household.Head.Occupation[i]== "Root croops farmers"
             || bbdd$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" || bbdd$Household.Head.Occupation[i]== "Coconut farmers"
             || bbdd$Household.Head.Occupation[i]== "Cotton and fiber crops farmers" || bbdd$Household.Head.Occupation[i]== "Deep-sea fishermen"
             || bbdd$Household.Head.Occupation[i]== "Field legumes farmers" || bbdd$Household.Head.Occupation[i]== "Forest tree planters"
             || bbdd$Household.Head.Occupation[i]== "Fruit tree farmers" || bbdd$Household.Head.Occupation[i]== "Hog raising farmers"
             || bbdd$Household.Head.Occupation[i]== "Inland and coastal waters fishermen" || bbdd$Household.Head.Occupation[i]== "Miners and quarry
workers"
             || bbdd$Household.Head.Occupation[i]== "Minor forest products gatherers" || bbdd$Household.Head.Occupation[i]== "Ornamental plant
growers"
             || bbdd$Household.Head.Occupation[i]== "Other aqua products cultivators" || bbdd$Household.Head.Occupation[i]== "Other orchard farmers"
             || bbdd$Household.Head.Occupation[i]== "Other poultry farmers" || bbdd$Household.Head.Occupation[i]== "Production and operations managers
in agriculture"
             || bbdd$Household.Head.Occupation[i]== "Seaweeds cultivators" || bbdd$Household.Head.Occupation[i]== "Sugarcane farmers"
             || bbdd$Household.Head.Occupation[i]== "Vegetable farmers" || bbdd$Household.Head.Occupation[i]== "Wood processing plant operators"
             || bbdd$Household.Head.Occupation[i]== "Farmhands and laborers" || bbdd$Household.Head.Occupation[i]== "Tree nut farmers"){
    bbdd$Household.Head.Occupation.new[i]= "Primary Sector"
  }else if (bbdd$Household.Head.Occupation[i]== "Brewers and wine and other beverage machine operators"
            ||  bbdd$Household.Head.Occupation[i]== "Building and related electricians"
            || bbdd$Household.Head.Occupation[i]== "Building construction laborers"
            || bbdd$Household.Head.Occupation[i]== "Cement and other mineral products machine operators"
            || bbdd$Household.Head.Occupation[i]== "Fishery laborers and helpers"
            || bbdd$Household.Head.Occupation[i]== "Industrial robot operators" || bbdd$Household.Head.Occupation[i]== "Metal finishing"
            || bbdd$Household.Head.Occupation[i]== "Metal finishing" || bbdd$Household.Head.Occupation[i]== "Wood and related products assemblers"
            || bbdd$Household.Head.Occupation[i]== "Wood products machine operators"
            || bbdd$Household.Head.Occupation[i]== "Earth-moving and related plant operators"
            || bbdd$Household.Head.Occupation[i]== "Freight handlers" || bbdd$Household.Head.Occupation[i]== "Hand launderers and pressers"
            || bbdd$Household.Head.Occupation[i]== "Marine craft mechanics" || bbdd$Household.Head.Occupation[i]== "Masons and related concrete
finishers"
            || bbdd$Household.Head.Occupation[i]== "Metal" || bbdd$Household.Head.Occupation[i]== "Production and operations managers in construction"
            || bbdd$Household.Head.Occupation[i]== "Sheet-metal workers" || bbdd$Household.Head.Occupation[i]== "Textile"
            || bbdd$Household.Head.Occupation[i]== "Sewers" || bbdd$Household.Head.Occupation[i]== "Weavers") {
    bbdd$Household.Head.Occupation.new[i]="Secondary Sector"
  }else {
            bbdd$Household.Head.Occupation.new[i]="Tertiary Sector"
  }
}
table(bbdd$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 <-bbdd[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$Household.Head.Class.of.Worker)<-c(levels(bbdd$Household.Head.Class.of.Worker),"Unemployed")
bbdd$Household.Head.Class.of.Worker[is.na(bbdd$Household.Head.Class.of.Worker)] <- "Unemployed"
levels(bbdd$Household.Head.Occupation)<-c(levels(bbdd$Household.Head.Occupation),c("No Occupation"))
bbdd$Household.Head.Occupation[is.na(bbdd$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 (sapply(dd,class)[k] == "integer"){
    print(paste("Anàlisi per classes de la Variable:", names(dades)[k]))
    o<-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)
  }
}
```