Caracterització dels compradors del producte *"Alfa"* i els seus competidors
Mineria de Dades

Miquel Perelló Nieto, Marc Bergés Garrido

20 d'abril de 2012

Índex

Índex de figures

Capítol 1

Introducció

Una certa marca de productes d'alt consum té la intenció d'emetre un anunci televisiu d'un dels seus productes. Abans de l'emisió, per veure la reacció dels consumidors davant l'anunci i els dels seus competidors, la companyia decideix dur a terme una enquesta a una mostra de 252 compradors potencials, als que se'ls pregunta sobre la última marca que han comprat i la seva preferència després de veure els anuncis, així com informació sòcio-demogràfica.

Amb aquesta informació se'ns demana caracteritzar els compradors de les diferents marques ALFA, BETA, GAMMA i DELTA.

Capítol 2

Perfils dels consumidors de les marques ALPHA, BETA, GAMMA i DELTA

2.1 Resum inicial

En aquesta secció farem un repas general a les variables amb les què comptem, explicant-ne el significat i intentant treure alguna conclusió.

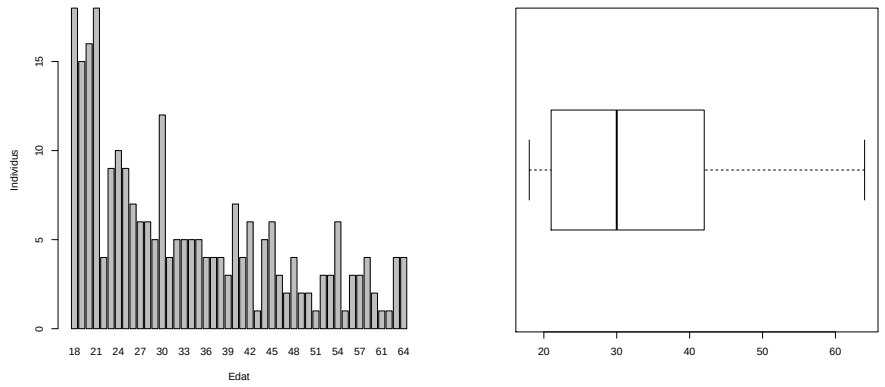
2.1.1 Edat (edat)

L'edat dels participants ens vé donada en el rang dels 18 als 64 anys així que podem deduïr que la mostra s'ha pres només a persones en edat de treballar. És una variable contínua i veiem que no té cap dada mancant.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Variance
18.00	21.00	30.00	33.16	42.00	64.00	174.9894

Taula 2.1: Resum de la variable edat

La mitjana d'edats de la mostra es troba en 33 anys. Observem que el 50% dels individus tenen una edat igual o inferior als 30 anys i d'aquests, la meitat en té 21 o menys. Del 50% restant, la meitat estan per sobre dels 42 anys. Per tant, la densitat d'individus joves és força més elevada que la de gent gran.



(a) Barplot d'edat

(b) Boxplot d'edat

Figura 2.1: Gràfiques d'edat

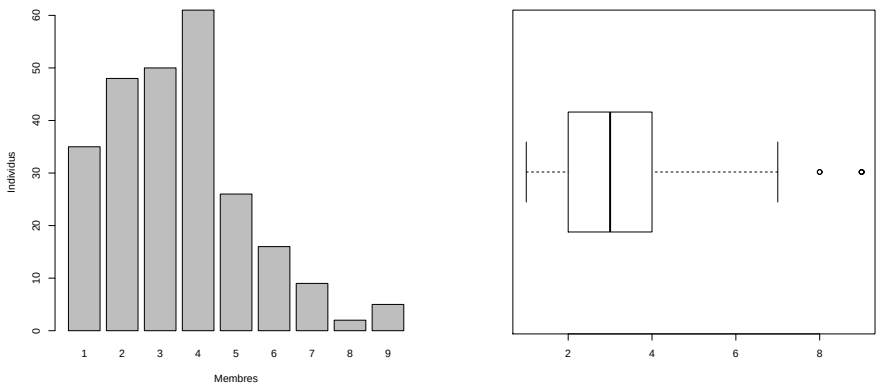
2.1.2 Membres de la família (memb)

El nombre de membres de la família de cada individu està en el rang de 1 (persona soltera o separada o vídua sense fills) fins a 9. No es té coneixement de com es pot interpretar exactament aquest nombre; pero pels valors que es donen podem imaginar que només es tenen en compte pares i fills. És una variable contínua i no hi ha cap dada mancant.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Variance
1.000	2.000	3.000	3.472	4.000	9.000	3.246237

Taula 2.2: Resum de la variable memb

Un 25% de la mostra son famílies amb només un o dos membres (solters i parelles), un altre 25% tenen de 2 a 3 membres (parella i un fill), un 25% més en tenen de 3 a 4 i la resta 4 o més (grup que es podria considerar com a família nombrosa). El grup més abundant és el de famílies amb 4 membres (parella i dos fills). També hi observem algun outlier o cas extrem que donarem per vàlid ja que no és molt freqüent trobar famílies amb 8 o 9 membres però sí és possible.



(a) Plot de memb

(b) Boxplot de memb

Figura 2.2: Gràfiques de memb

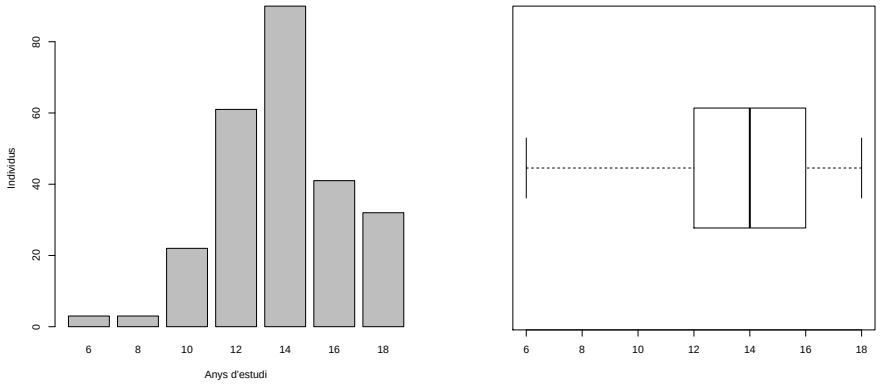
2.1.3 Anys d'estudi (estd)

Una altra variable sòcio-econòmica són els anys d'estudi de cada un dels individus enquestats. Aquests van dels 6 anys (educació primària) fins als 18 anys (carrera superior). És una variable contínua i no té dades mancants.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Variance
6.00	12.00	14.00	13.83	16.00	18.00	6.23506

Taula 2.3: Resum de la variable estd

La mitjana es troba en uns 14 anys que equivalen aproximadament a un batxillerat. De tota la mostra, un 25% de persones tenen un nivell baix d'estudis (de 6 a 12 anys; educació primària o secundària). La meitat té el títol de batxillerat, només un 25% té una carrera universitària i el tercer quartil es troba als 16 anys que es podrien interpretar com un cicle formatiu.



(a) Plot d'estd

(b) Boxplot d'estd

Figura 2.3: Gràfiques d'estd

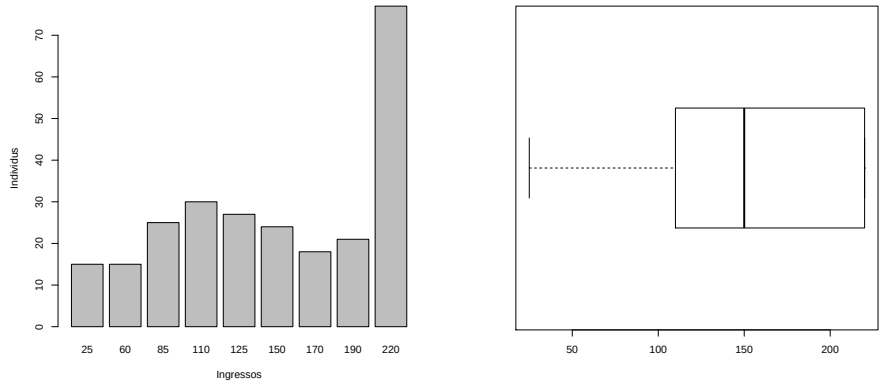
2.1.4 Ingressos (ingr)

Els ingressos són una variable que està en el rang de 25 a 220 encara que s'aprecia que fa salts discrets. No sabem quin tipus de moneda o equivalència té. És una variable contínua i no té cap dada mancant.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Variance
25.0	110.0	150.0	149.5	220.0	220.0	3762.560

Taula 2.4: Resum de la variable ingr

Veiem que la mitjana es situa en 150 aproximadament al igual que la mediana. Hi ha molts individus que tenen uns ingressos de 220 que es suposa que és el màxim que permetia respondre l'enquesta. Un 25% dels enquestats cobra per sota de 110, que es considerarien uns ingressos baixos, un 50% cobra igual o menys de 150 i un altre 25% cobra igual o més de 220.



(a) Plot dels ingressos

(b) Boxplot dels ingressos

Figura 2.4: Gràfiques dels ingressos

2.1.5 Estat civil (eciv)

L'estat civil només contempla dues categories: solters i casats. Aquesta és una variable categòrica i tampoc té dades mancants.

solter	casat
109	143

Taula 2.5: Resum de la variable eciv

Així doncs, veiem que un 43% dels individus estan solters i la resta, casats.

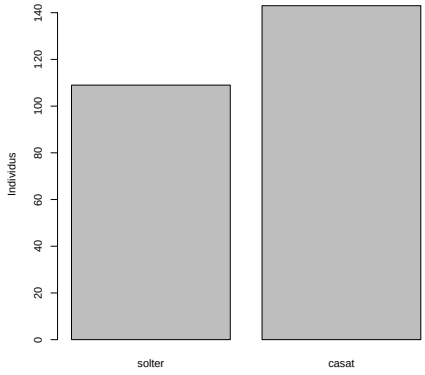


Figura 2.5: Gràfica de la distribució de l'estat civil

2.1.6 Professió (prof)

La professió divideix els individus en tres grups que són: obrers, administratius o directius i altres. És una variable categòrica i no té dades mancants.

altres	obrer	admin/dir
42	98	112

Taula 2.6: Resum de la variable prof

Tenim un 44% d'administratius/directius, un 39% d'obrers i un 17% de persones que composen la mostra i es dediquen a d'altres professions no contemplades en els dos grups anteriors.

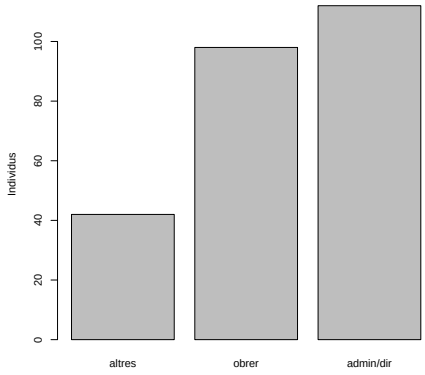


Figura 2.6: Gràfica de la distribució de la professió

2.1.7 Marca 1 (mar1) i marca 2 (mar2)

Aquestes dues variables de tipus categòric ens indiquen quina és la última marca comprada per cada individu abans de veure els anuncis i quina és la seva preferència després. Són variables de tipus categòric i aquestes sí que tenen dades mancants, cosa que farà necessària la seva imputació.

beta1	gamma1	delta1	alfa1	altra1	estr1	NA's
26	8	61	48	80	10	19

Taula 2.7: Resum de l'últim producte comprat abans de veure els anuncis

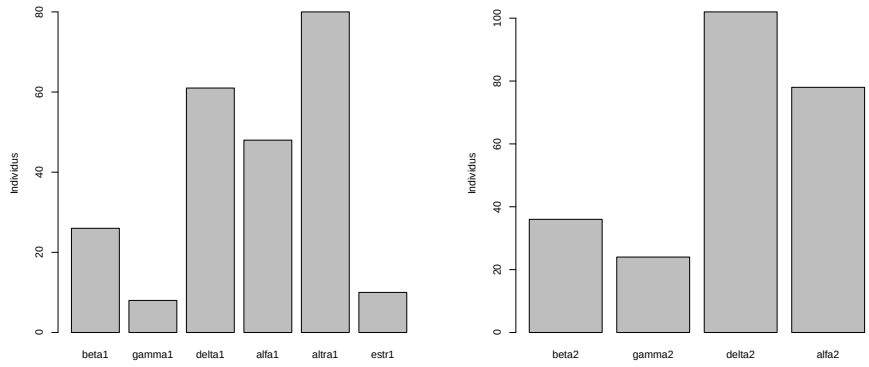
beta2	gamma2	delta2	alfa2	NA's
36	24	102	78	12

Taula 2.8: Resum de la preferència de compra un cop vista la publicitat

Un primer cop d'ull ens permet veure ràpidament que hi ha dues categories a mar1 que no apareixen a mar2 (altra1 i estr1) així com que hi ha 19 dades mancants a la primera i 12 a la segona. També podem veure quin percentatge del total de persones compra cada marca:

beta1	gamma1	delta1	alfa1	altra1	estr1	NA's
10.32	3.17	24.2	19.05	31.75	3.97	7.54

Taula 2.9: Percentatge de la mostra que ha comprat cada marca abans de l'anunci



(a) Plot abans

(b) Plot després

Figura 2.7: Gràfiques de marca preferida abans i després de veure els anuncis

beta2	gamma2	delta2	alfa2	NA's
14.28	9.52	40.48	30.95	4.76

Taula 2.10: Percentatge de la mostra que té intenció de comprar cada marca després de l'anunci

2.2 Imputació de dades mancants

En aquesta secció imputarem les dades mancants que hem descobert abans (de les variables mar1 i mar2) buscant la distància mínima entre les mostres amb totes les dades i les mostres amb dades desconegudes. D'aquesta manera, els assignarem el valor més pròxim.

2.2.1 Marca 1 (mar1) i marca 2 (mar2)

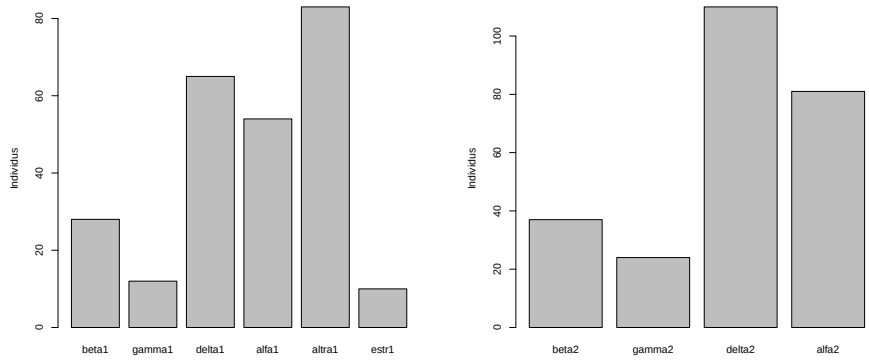
Existeixen individus que no ens han indicat quina és l'última marca que han comprat (mar1) i, per tant, primer farem aquesta imputació. Posteriorment, farem la imputació de la variable de preferència de compra després d'haver vist els anuncis televisius (mar2).

beta1	gamma1	delta1	alfa1	altra1	estr1
28	12	65	54	83	10

Taula 2.11: Resum de l'últim producte comprat

beta2	gamma2	delta2	alfa2
37	24	110	81

Taula 2.12: Resum de la preferència de compra un cop vista la publicitat



(a) Plot abans

(b) Plot després

Figura 2.8: Gràfiques de marca preferida abans i després de veure els anuncis

Ara ja no trobem a faltar cap dada. Totes les que necessitàvem les hem assignat tenint en compte els individus que estaven més a prop i prenent el seu valor per la imputació. D'aquesta manera aconseguim no distorsionar gaire el resultat ja que suposem que dues persones de semblants característiques també

tindran el mateix comportament. Els percentatges després de les imputacions queden de la següent manera:

beta1	gamma1	delta1	alfa1	altra1	estr1
11.11	4.76	25.79	21.43	32.94	3.97

Taula 2.13: Percentatge de la mostra que ha comprat cada marca abans de l'anunci

beta2	gamma2	delta2	alfa2
14.68	9.52	43.65	32.14

Taula 2.14: Percentatge de la mostra que té intenció de comprar cada marca després de l'anunci

Podem comprovar com els percentatges han pujat lleugerament ja que s'ha assignat les dades mancants a les diverses categories però mantenen la distribució.

2.3 Categorització de variables contínues

Per poder treballar necessitem recodificar les variables contínues a variables categòriques. Això ho farem delimitant uns rangs per cada una d'elles i intentant que la distribució sigui el més homogènia possible, sempre mirant que els rangs tinguin un significat aplicable al món real.

2.3.1 Edat (redat)

A la variable edat li assignarem els rangs d'1 a 21 anys (persones que definirem com adolescents), de 22 a 26 anys (joves), de 27 a 34 (adults-joves), de 35 a 45 (adults) i de 45 en endavant (gent gran).

ED (0-21]	ED (21-26]	ED (26-34]	ED (34-45]	ED (45-99]
67	39	48	49	49

Taula 2.15: Resum de l'edat categoritzada

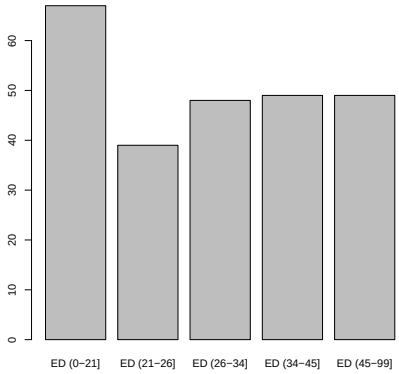


Figura 2.9: Gràfica de la distribució de l'edat

2.3.2 Anys d'estudi (restd)

Els anys d'estudi els dividirem en rangs segons el que creiem correspon als nivells d'estudis actuals (encara que no sabem on està presa l'enquesta). Així doncs prendrem, de manera orientativa, els intervals d'1 a 10 anys (estudis primaris i secundaris), d'11 a 12 (batxillerat), de 13 a 14 (formació professional), de 15 a 16 (carrera tècnica) i més de 16 (carrera superior).

ES (0-10]	ES (10-12]	ES (12-14]	ES (14-16]	ES (16-20]
28	61	90	41	32

Taula 2.16: Resum dels anys d'estudi categoritzats

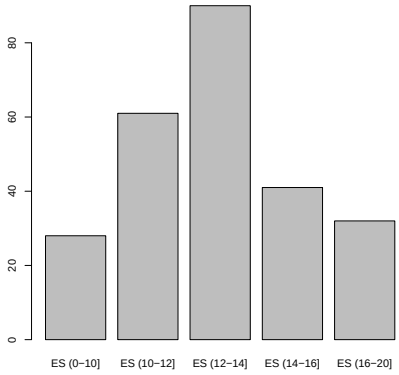


Figura 2.10: Gràfica de la distribució dels anys d'estudi

2.3.3 Ingressos (ringr)

Com no sabem de quin tipus de moneda o equivalència té aquesta variable, decidim crear els rangs d'1 a 50 (ingressos baixos), de 51 a 100 (baixos-mitjans), de 101 a 150 (mitjans), de 151 a 200 (mitjans-alts) i de 201 a 250 (ingressos alts).

IN (0-50]	IN (50-100]	IN (100-150]	IN (150-200]	IN (200-250]
15	40	81	39	77

Taula 2.17: Resum dels ingressos categoritzats

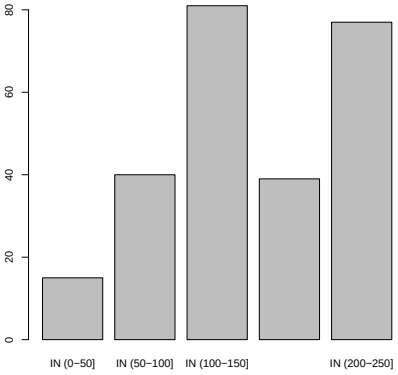


Figura 2.11: Gràfica de la distribució dels ingressos

2.3.4 Membres de la família (rmemb)

Per acabar, els intervals de la variable memb seran 1 membre (solter), 2 (parella), 3 (parella amb un fill), 4 (parella amb dos fills) i de 5 en endavant (família nombrosa.

ME (0-1]	ME (1-2]	ME (2-3]	ME (3-4]	ME (4-9]
35	48	50	61	58

Taula 2.18: Resum dels membres de la família categoritzats

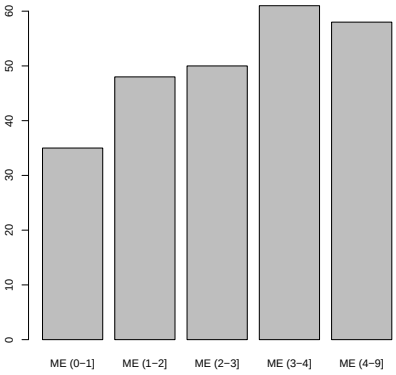
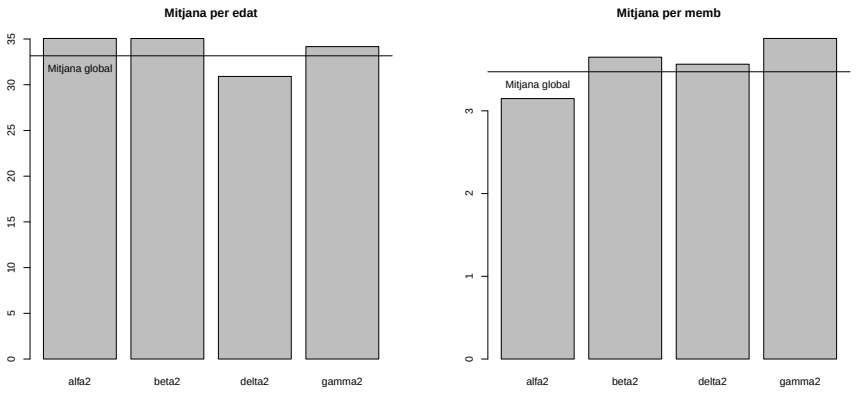


Figura 2.12: Gràfica de la distribució dels membres de la família

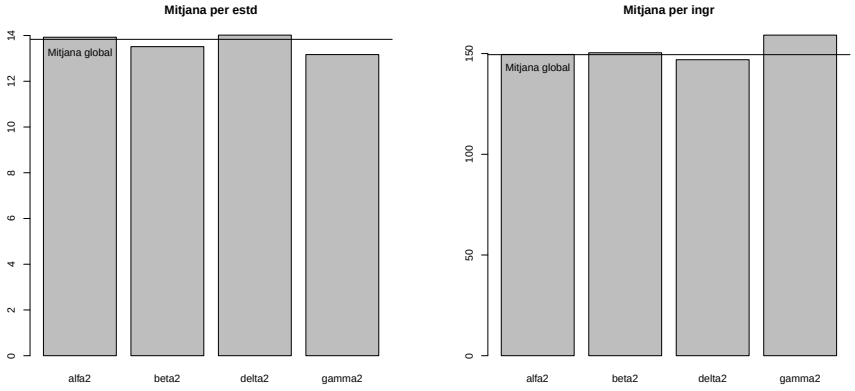
2.4 Visualització de les dades

En aquesta secció representarem les dades visualment. Per començar, representem els valors de les mitjanes de cada marca, contrastats amb el valor de la mitjana general.



(a) Edat (b) Membres

Figura 2.13: Mitjanes de les variables contínues per cada marca



(a) Professió (b) Ingressos

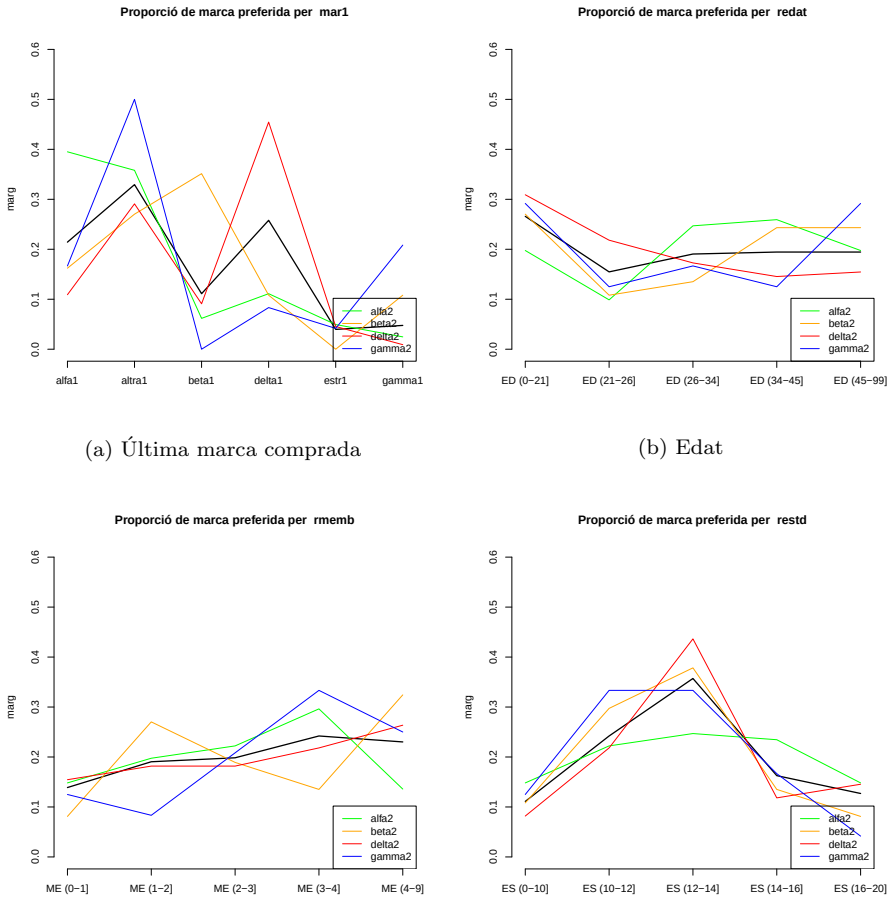
Figura 2.14: Mitjanes de les variables contínues per cada marca

Observem que existeixen dependències segons les marques ja que no totes les mitjanes de cada una d'elles es corresponen a la mitjana global.

Així doncs, podem destacar que pel que sembla, la marca gamma serà comprada per individus que tenen una mitjana d'estudis inferior a la normal i que tenen més ingressos però també famílies amb més membres. Delta serà comprada per gent més jove i alfa, per gent amb unitats familiars més reduïdes.

En aquestes gràfiques, la línia negra representa el valor marginal mig. Quan més s'assemblin les altres línies a la negra, menys representatives seran.

Per tant, un cop revisades, podem dir que les més significatives són la marca comprada (mar1), l'estat civil i la edat.



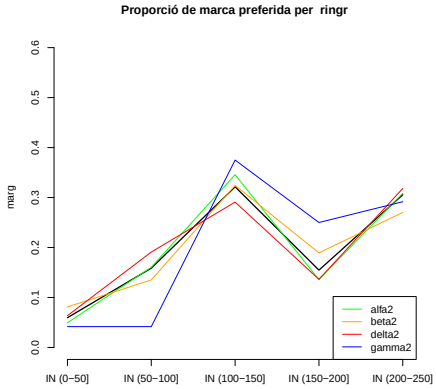
(a) Última marca comprada

(b) Edat

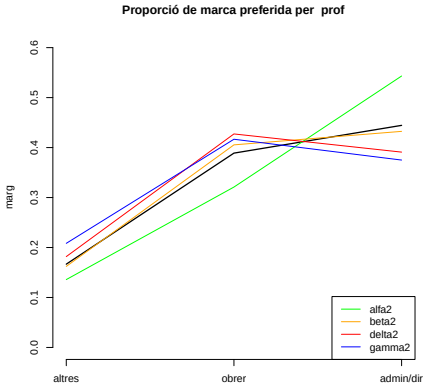
(c) Membres de la família

(d) Anys d'estudi

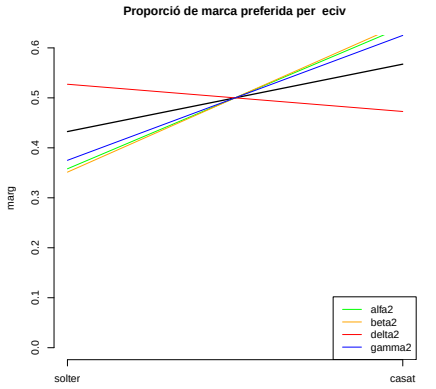
Figura 2.15: Proporció de preferències per diferents variables 1



(a) Ingressos



(b) Professió



(c) Estat civil

Figura 2.16: Proporció de preferències per diferents variables 2

2.5 Selecció de característiques

En un estudi on el nombre de variables explicatives és molt gran, és necessari realitzar un anàlisi de quines d'elles són rellevants i quines només afegeixen soroll per tal d'eliminar-les.

Per veure quines d'aquestes variables són realment importants, es realitza una comparació entre els valors d'aquestes i els valors de la variable de resposta que permet veure si aquesta última està linealment condicionada per les variables explicatives.

2.5.1 Variables contínues

Per poder saber quines variables contínues són significants, realitzem un test de F de Fisher amb la variable de resposta i ens fixem el el p-value que retorna. Si aquest és inferior a 0.1, es considera que hi ha una relació lineal entre les dues variables i aleshores agafem la variable explicativa com a representativa.

	p-value
edat	0.13
memb	0.17
estd	0.36
ingr	0.83

Cap dels p-values és inferior a 0.1 encara que la edat s'hi acosta i, per tant, no considerem les variables com a rellevants.

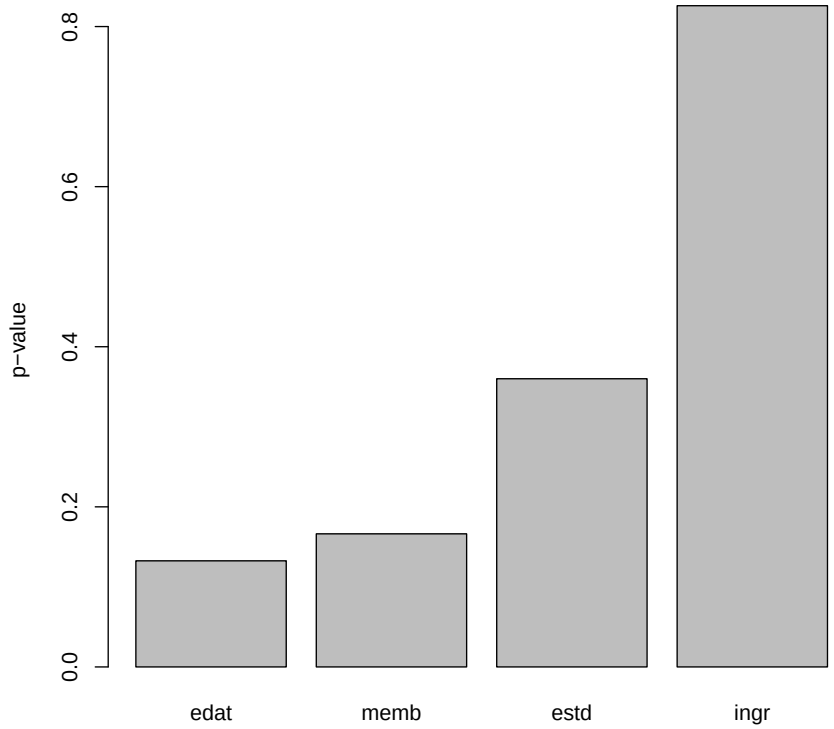


Figura 2.17: Variables explicatives contínues, per ordre creixent de p-value

2.5.2 Variables categòriques

Per les variables categòriques realitzarem una prova semblant però ara amb un test de Chi-Quadrat.

Ara hem trobat dues variables que són rellevants, mar1 i eciv o estat civil ja que tenen p-values inferiors a 0.1.

Tot i això, en el nostre cas no es realitzarà la eliminació de cap variable ja que el nombre de variables explicatives és prou reduït com per poder treballar còmodament.

	p-value
mar1	0.00
prof	0.53
eciv	0.07
redat	0.18
restd	0.24
ringr	0.89
rmemb	0.35

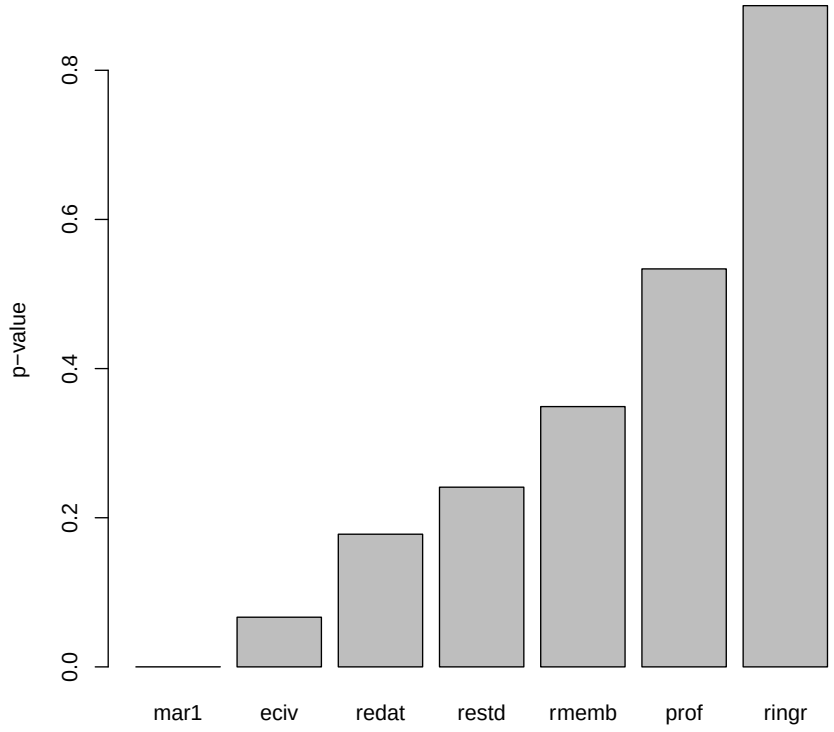


Figura 2.18: Variables explicatives categòriques, per ordre creixent de p-value

2.6 Perfil de cada marca

Ens interessa trobar el perfil dels compradors de cada una de les marques i, per fer això, seguirem un procediment semblant a l'anterior però aplicant dues funcions (p.xk i p.zkj) depenent del tipus de variables explicatives i per cada un dels valors de la variable de resposta.

2.6.1 Variables contínues

Fent servir la funció p.xk amb les variables contínues i mar2 obtenim:

	edat	memb	estd	ingr
alfa2	0.06	0.97	0.34	0.49
beta2	0.17	0.26	0.80	0.46
delta2	0.99	0.24	0.15	0.72
gamma2	0.35	0.13	0.91	0.21

Seguint el criteri d'abans on dèiem que una variable és representativa si el seu p-value és inferior a 0.1, veiem que només ens resulta significativa la edat per la marca alfa.

2.6.2 Variables categòriques

Ara farem servir la funció p.zkj amb les variables contínues i mar2. Desglossarem els resultats en cada una de les variables explicatives per fer-ho més entenedor.

Última marca comprada

	alfa1	altra1	beta1	delta1	estr1	gamma1
alfa2	0.00	0.25	0.96	1.00	0.29	0.88
beta2	0.80	0.80	0.00	0.99	0.91	0.03
delta2	1.00	0.87	0.82	0.00	0.34	0.99
gamma2	0.72	0.03	0.97	0.98	0.48	0.00

Queda ben clar que el fet d'haver comprat una marca anteriorment és molt representatiu del fet de voler comprar-la després del visionat dels anuncis. També observem com el fet d'haver comprat primerament el producte gamma és representatiu de comprar després beta i el fet d'haver comprat una altra marca, de gamma.

Edat

Ara veiem que els individus de les edats de (26-45] anys són clars compradors d'alfa mentre que els de delta estan compresos en l'interval (0-26] i els compradors de gamma són els més grans (45-99] anys.

	ED (0-21]	ED (21-26]	ED (26-34]	ED (34-45]	ED (45-99]
alfa2	0.95	0.95	0.06	0.04	0.47
beta2	0.47	0.80	0.82	0.21	0.21
delta2	0.09	0.01	0.74	0.96	0.92
gamma2	0.38	0.66	0.62	0.82	0.10

Membres de la família

	ME (0-1]	ME (1-2]	ME (2-3]	ME (3-4]	ME (4-9]
alfa2	0.38	0.42	0.26	0.08	0.99
beta2	0.86	0.09	0.56	0.95	0.07
delta2	0.26	0.62	0.72	0.78	0.13
gamma2	0.58	0.92	0.45	0.14	0.40

En quant a nombre de membres, els compradors d'alfa formen part de famílies amb pares i dos fills i els de beta són o bé parelles o famílies nombroses.

Anys d'estudi

	ES (0-10]	ES (10-12]	ES (12-14]	ES (14-16]	ES (16-20]
alfa2	0.10	0.69	0.99	0.02	0.24
beta2	0.53	0.20	0.39	0.69	0.82
delta2	0.90	0.78	0.01	0.95	0.22
gamma2	0.41	0.14	0.60	0.48	0.91

La gent amb una carrera tècnica o sense gaires estudis es decanta per la marca alfa mentre que la gent amb formació professional prefereix delta.

Ingressos

Els ingressos no són gaire significatius tret, potser, de la gent amb uns ingressos mitjans-alts que prefereix la marca gamma.

	IN (0-50]	IN (50-100]	IN (100-150]	IN (150-200]	IN (200-250]
alfa2	0.68	0.48	0.29	0.72	0.47
beta2	0.27	0.66	0.48	0.27	0.69
delta2	0.40	0.11	0.82	0.76	0.35
gamma2	0.65	0.95	0.28	0.09	0.56

Estat civil

	solter	casat
alfa2	0.95	0.05
beta2	0.86	0.14
delta2	0.00	1.00
gamma2	0.73	0.27

Amb l'estat civil sí podem veure a primera vista que els individus casats compren alfa mentre que els solters compren delta.

Professió

	altres	obrer	admin/dir
alfa2	0.82	0.94	0.01
beta2	0.53	0.41	0.56
delta2	0.29	0.14	0.93
gamma2	0.28	0.38	0.76

Per acabar, en quant a professió, només podem dir que alfa és comprada per administratius o directius.

Així doncs, hem vist que les característiques que tenen els compradors de cada marca són:

- Alfa: Antics compradors d'alfa; edat entre 27 i 45 anys; famílies amb 4 membres (pares i dos fills); de 15 a 16 anys d'estudi (carrera tècnica) o 10 o menys (estudis primaris-secundaris); casats; administratius o directius
- Beta: Antics compradors de beta; parelles o famílies nombroses
- Gamma: Antics compradors de gamma o altres marques; gent gran; ingressos mitjans-alts
- Delta: Antics compradors de delta; gent jove de 18 a 26 anys; 13-14 anys d'estudi (formació professional); solters

A partir d'això extreiem les següents conclusions dels compradors de cada marca:

- Alfa: Gent adulta, casada, que viu en família i té una bona feina (directiu o administratiu).

- Beta: Parelles i famílies nombroses.
- Gamma: Gent gran amb ingressos mitjans o alts que ja té la vida solucionada.
- Delta: Gent jove i que viu sola.

Capítol 3

Posicionament multidimensional

En aquesta secció mirarem de representar de la millor manera possible les variables socio-econòmiques amb les que comptem. Per fer això primer farem un anàlisis multivariable. Com a resposta utilitzarem la última marca comprada i la marca preferida un cop s'han visualitzat els anuncis. Un cop realitzat aquest anàlisi podrem reduïr les dimensions a visualitzar, de manera que puguin ser representades en una gràfica de dues dimensions.

3.1 Anàlisi multivariant

Per realitzar l'anàlisis multivariant ens hem ajudat d'un anàlisi de correspondències múltiples també anomenat MCA (de l'anglès Multiple Correspondece Analysis). Un cop generades totes les dimensions del nou espai vectorial, podem generar una gràfica de barres amb la inèrcia de cada un dels nous eixos.

En aquesta gràfica generada podem observar tots els eixos ordenats decreixentment per inèrcia. En aquest cas veiem que existeix un eix molt diferenciat que utilitzarem per fer la visualització. El segon eix de referències serà el segon eix de la llista; que ens dóna una mica més d'inèrcia que el tercer. El significat real d'aquests eixos en aquest punt ens és totalment desconegut, ja que només hem mirat de maximitzar la seva inèrcia, posteriorment als següents apartats intentarem donar un caràcter representatiu a aquests, gràcies a la visualització de les diferents categories en aquest espai vectorial.

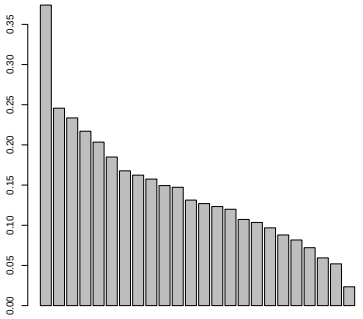


Figura 3.1: **Valors propis dels diferents eixos**

3.2 Projecció de variables il·lustratives

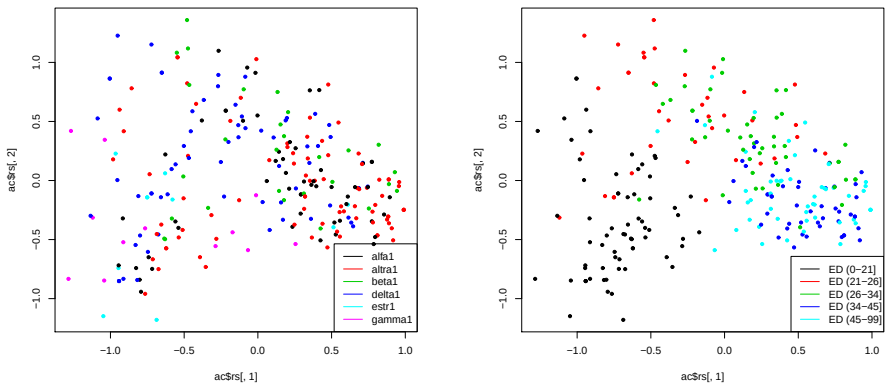
Un cop decidits els eixos més representatius de la inèrcia total, anem a representar les diferents variables explicatives en vàries gràfiques. El que podrem observar és la seva distribució en aquest nou espai, i li donarem explicacions que ajudin a entendre el seu sentit.

Per això hem fet gràfiques de punts, on cada punt és un individu pintat amb la categoria a la qual pertany. En aquesta gràfica podem veure la seva distribució a l'espai, però per poder definir en quin punt es troba el seu centre de gravetat hem creat una gràfica on apareixen els centres de gravetat de cada una d'aquestes variables.

Tot seguit expliquem les conclusions que podem extreure d'aquestes visualitzacions.

3.2a **Última marca comprada:** En aquesta categoria podem veure, en els centres de gravetat, que les marques *"gamma1"* i *"estr1"* estan molt separades del punt central de la gràfica, en concret en el tercer quadrant (part inferior esquerra). A la gràfica de punts veiem que la seva dispersió no és molt gran. La resta de marques no s'allunyen molt del centre però es veuen tant la marca *"altra1"* com la *"alfa1"* bastant juntes a la part dreta, i *"beta1"* que tendeix a la part superior.

3.2b **Edat:** Sembla seguir una diagonal perfectament descendent amb l'increment d'edat. Encara que el cas concret d'individus d'edat inferior a 21 semblen un punt separat a la part inferior esquerra de la gràfica. A més, no presenta una distribució gran respecte altres variables, ja que es poden veure els centres de gravetat a simple vista a la gràfica de punts.



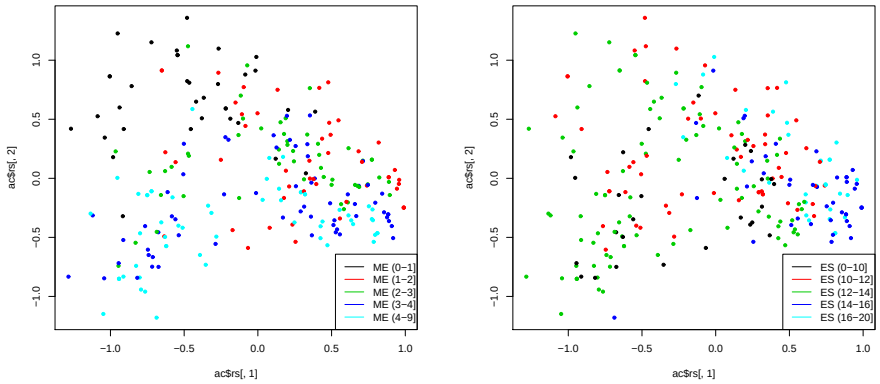
(a) Última marca comprada (b) Edat

Figura 3.2: Projecció del núvol de punts segons els dos primers eixos 1

3.3a **Membres de la família:** Observant la gràfica 3.5 sembla que segueixen l'eix vertical en nombre de membres, però existeix una tendència a la

dreta en els de dos a tres membres. A l'extrem oposat hi tenim els d'un membre, que es troben totalment a la part superior esquerra. Encara que a la gràfica de punts sembla que la distribució d'aquests està bastant estesa, i no es centra en els seus centres de gravetat.

3.3b **Anys d'estudi:** Amb aquesta variable no es pot veure una tendència a les abscisses positives pels individus amb més de 16 anys d'estudis; en canvi, pels de menys edat es veuen distribuïts caòticament però amb el centre de gravetat al centre.



(a) Membres de la família (b) Anys d'estudi

Figura 3.3: Projecció del núvol de punts segons els dos primers eixos 2

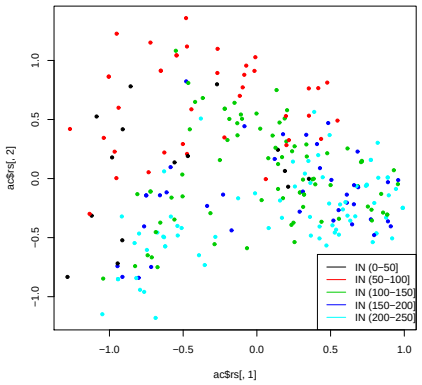
3.4a **Ingressos:** Aquesta variable sembla seguir l'eix de coordenades éssent els ingressos més baixos els valors positius de l'eix, i els més alts els negatius. Peró, concretament amb els ingressos d'entre 0 i 50 es veu que no segueixen aquesta tendència lineal, i es queden a la part esquerra separats dels altres.

3.4b **Professió:** Aquesta variable sembla seguir l'eix d'abscisses. Es veu clarament que les altres professions queden a l'esquerra i part inferior, els obrers semblen estar distribuïts per tota la gràfica i, per tant, creen el centre de gravetat molt proper al punt zero i, per últim, els administradors tenen una tendència a la dreta de la gràfica bastant centrat en la verticalitat.

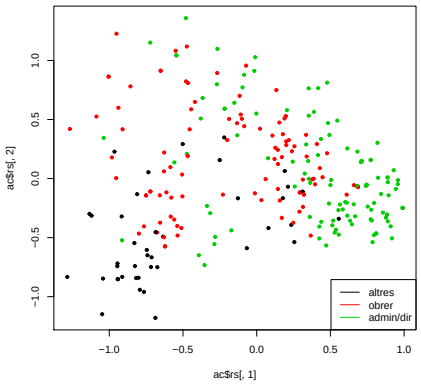
3.4c **Estat civil:** Igual que en el cas de la professió, es pot veure clarament que un dels criteris que defineixen l'eix d'abscisses és l'estat civil, en el qual, en el sentit negatiu, estan els solters i en el positiu, els casats.

A partir de la gràfica dels centres de gravetat de totes les modalitats es poden veure certs aparellaments en les variables categòriques.

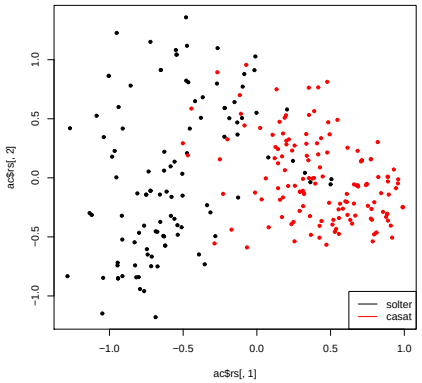
Per exemple, els individus que a la última compra van comprar la marca gamma o l'estrangera semblen tenir menys de 21 anys i una professió no definida.



(a) Ingressos



(b) Professió



(c) Estat civil

Figura 3.4: Projecció del núvol de punts segons els dos primers eixos 3

També sembla que hi ha una relació entre els individus amb edat compresa entre els 21 i 26 anys, amb uns ingressos tirant a baixos i que viuen sols.

Pel centre de la gràfica es pot veure que els individus amb professió obrera van comprar per últim cop la marca delta, i no tenen molt lluny la beta. A més sembla que tinguin entre 10 i 12 anys d'estudis.

Una altra dada curiosa és que els administratius solen estar casats, i la edat a la que s'apropen més és de més de 34 anys, i estan al voltant dels 14 als 20 anys d'estudis.

Per últim tenim la gràfica de les modalitats suplementàries, que ens mostra la última marca que havien comprat, i la seva marca preferida un cop visualitzats els anuncis. Per la distància entre elles es pot dir que els individus que compraven alfa, segueixen tenint a alfa com a preferit. Una certa similitud té el producte delta. Els altres tenen unes distancies considerables per suposar alguna relació de grup.

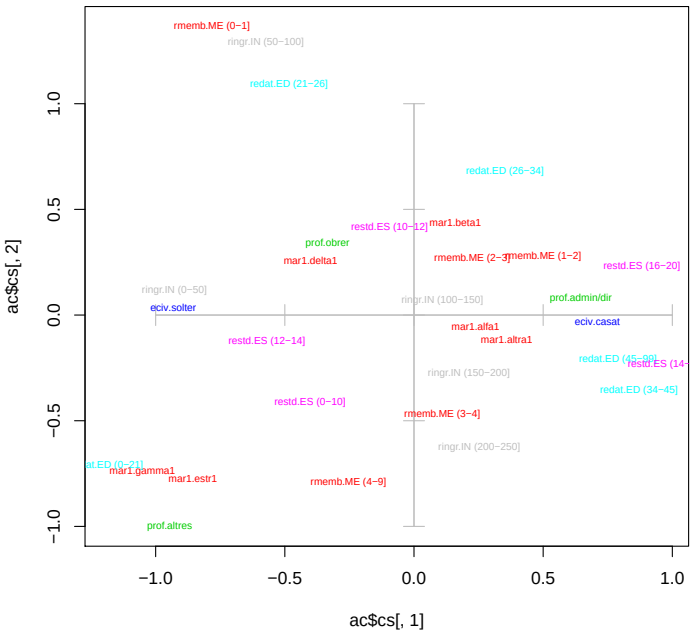


Figura 3.5: Projecció de les 24 modalitats

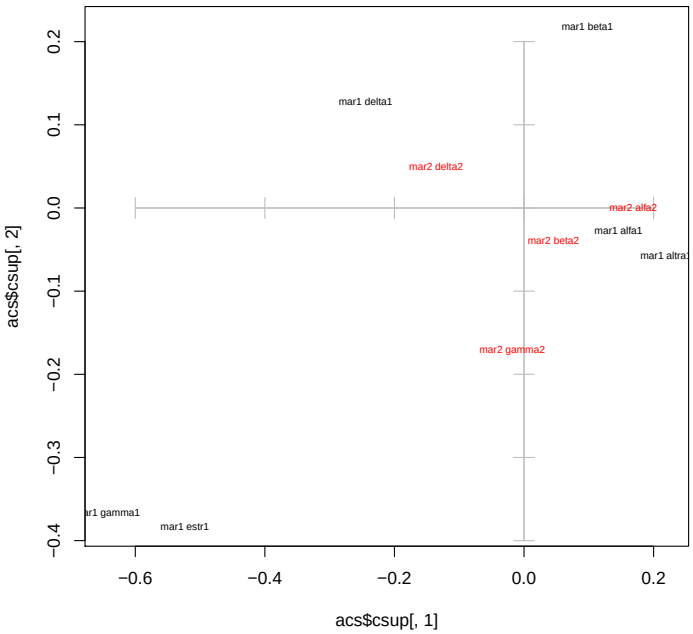


Figura 3.6: Projecció de les modalitats suplementàries

Capítol 4

Clústering

En aquest capítol intentarem agrupar els individus de la mostra. Per fer això, primer realitzarem un clústering jeràrquic que ens permetrà trobar el nombre de grups i el seu centre de gravetat per poder, posteriorment, realitzar clústering amb kmeans i fer-ne la consolidació.

4.1 Clústering jeràrquic i kmeans

Realitzem el clústering jeràrquic que dóna com a resultat el següent arbre. Cada una de les agregacions que s'hi produeixen mostra la diferència entre els grups que s'uneixen depenent de l'alçada a la que es trobi.

Seguidament veiem les inèrcies de les últimes agregacions.

Després de decidir quedar-nos amb sis clústers, utilitzem kmeans calculant els centres de gravetat de cada un dels grups per tal de fer la consolidació i els individus ens queden agrupats de la següent manera:

	Individus
c1	55
c2	43
c3	36
c4	51
c5	46
c6	21

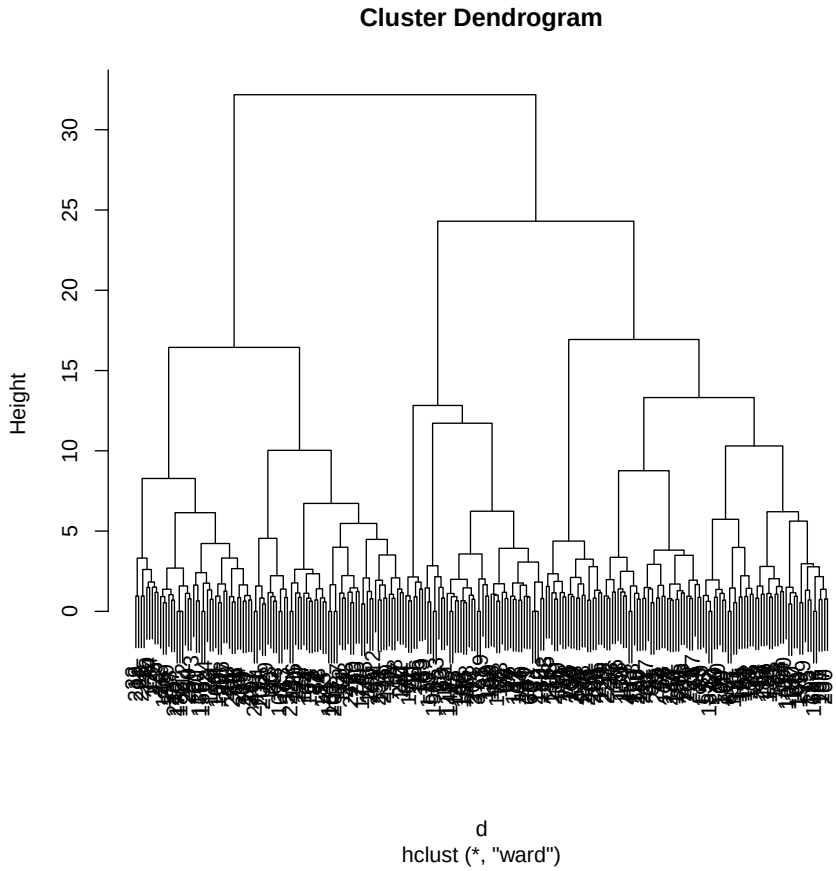


Figura 4.1: Arbre de totes les agregacions

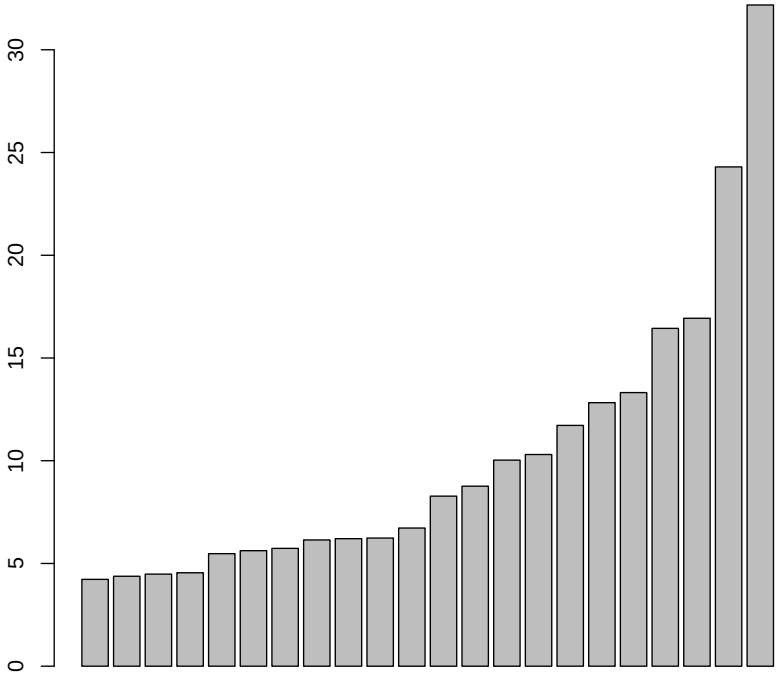


Figura 4.2: Inèrcia de cada una de les últimes agregacions

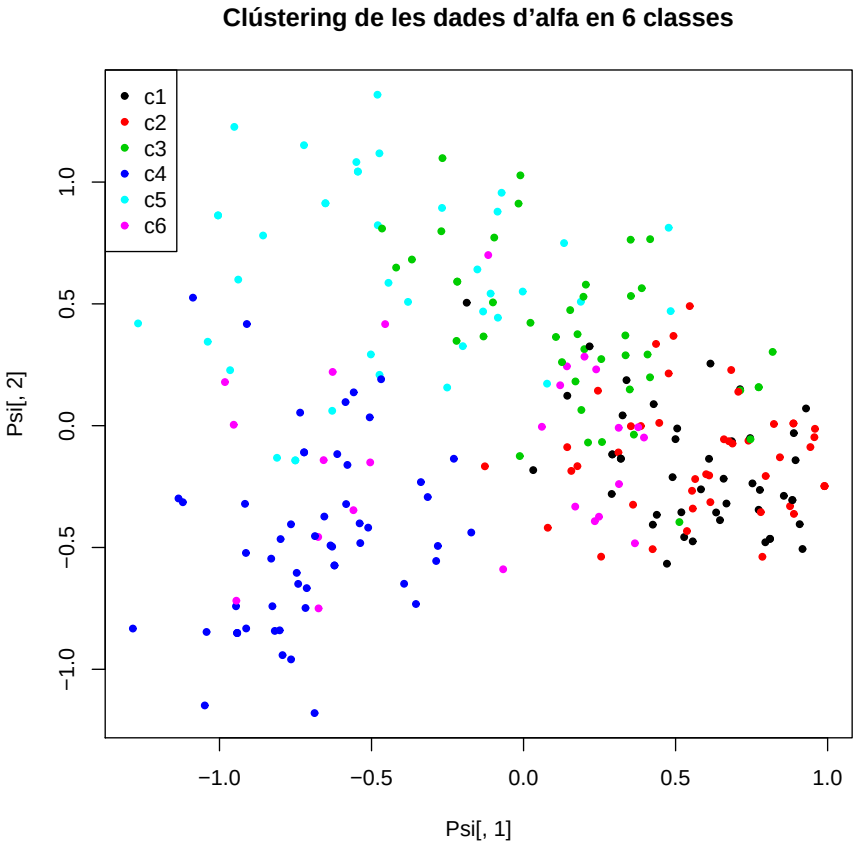


Figura 4.3: Projecció dels individus per clústers

4.2 Interpretació de les particions

Per poder interpretar els diferents clústers tornarem a fer servir la funció p.zkj per tal de trobar els p-values de cada categoria de les variables per cada clúster. Després els ordenarem de menor a major per veure quines d'aquestes variables explicatives són significatives. Aquest és el resultat:

```
[1] "P-values del clÃºster  1 "
  ED (34-45]         casat     admin/dir    ES (14-16]      ME (4-9]    ES (16-20]
2.822221e-46 1.418052e-07 8.421893e-06 1.459298e-03 1.372972e-02 1.712467e-02
     ME (3-4] IN (150-200] IN (100-150]         beta2         alfa2      ME (2-3]
2.493209e-02 3.323043e-02 5.508795e-02 1.330684e-01 2.943045e-01 3.292852e-01
IN (200-250]    ES (10-12]        gamma2        delta2         obrer     IN (0-50]
4.644353e-01 4.835911e-01 7.643643e-01 8.096157e-01 9.682491e-01 9.687024e-01
     ME (0-1]    ES (12-14]     ES (0-10]    ED (21-26]  IN (50-100]      ME (1-2]
9.783765e-01 9.814251e-01 9.955625e-01 9.966596e-01 9.971531e-01 9.992415e-01
  ED (45-99]         altres    ED (26-34]     ED (0-21]         solter
9.993609e-01 9.995336e-01 9.998348e-01 9.999958e-01 9.999999e-01
[1] "P-values del clÃºster  2 "
  ED (45-99]         casat      ME (1-2]    ES (14-16]     admin/dir IN (200-250]
3.956985e-32 3.355327e-09 9.288654e-09 9.734185e-08 3.562442e-05 1.261583e-04
       beta2         gamma2         alfa2 IN (150-200]    ES (16-20]      ME (2-3]
1.330684e-01 1.683895e-01 2.943045e-01 3.188179e-01 4.438790e-01 4.882492e-01
  ES (10-12]         altres     IN (0-50]      ME (4-9]    ED (21-26]      ME (3-4]
6.341460e-01 7.460185e-01 8.783723e-01 9.052045e-01 9.111795e-01 9.325226e-01
IN (100-150]         delta2    ED (26-34]    ES (12-14]   IN (50-100]     ES (0-10]
9.420534e-01 9.693556e-01 9.722416e-01 9.814251e-01 9.896884e-01 9.955625e-01
     ME (0-1]         obrer    ED (34-45]     ED (0-21]         solter
9.985232e-01 9.998019e-01 9.998618e-01 9.999958e-01 1.000000e+00
[1] "P-values del clÃºster  3 "
  ED (26-34]    ES (16-20]      ME (0-1]     admin/dir      ME (2-3]         casat
2.001557e-47 5.027214e-04 7.456670e-03 2.359939e-02 7.265819e-02 1.118936e-01
IN (100-150]      ME (3-4]   IN (50-100]         alfa2        delta2     IN (0-50]
1.272088e-01 1.555217e-01 1.595046e-01 3.363009e-01 3.389429e-01 3.776181e-01
  ES (10-12]    ES (12-14]         obrer         beta2        gamma2 IN (200-250]
4.085984e-01 5.496638e-01 7.229321e-01 7.328457e-01 7.339431e-01 7.815770e-01
  ES (14-16]      ME (1-2]        solter IN (150-200]        altres     ES (0-10]
8.174214e-01 8.248755e-01 8.881064e-01 9.546882e-01 9.694092e-01 9.945479e-01
  ED (21-26]    ED (45-99]    ED (34-45]      ME (4-9]     ED (0-21]
9.989688e-01 9.990788e-01 9.997981e-01 9.997994e-01 9.999927e-01
[1] "P-values del clÃºster  4 "
    ED (0-21]         solter        altres      ME (4-9]    ES (12-14]     IN (0-50]
2.122846e-39 7.530366e-17 2.183094e-11 3.715710e-07 8.727562e-05 7.016812e-03
IN (200-250]    ES (10-12]        delta2      ME (3-4]        gamma2 IN (150-200]
1.514381e-02 3.865697e-02 8.520863e-02 1.469389e-01 1.657049e-01 2.427928e-01
       obrer         beta2     ES (0-10] IN (100-150]         alfa2      ME (2-3]
6.235882e-01 6.564445e-01 8.357175e-01 9.608824e-01 9.608824e-01 9.652220e-01
     ME (0-1]   IN (50-100]      ME (1-2]    ED (21-26]    ES (16-20]    ES (14-16]
9.771228e-01 9.971169e-01 9.978033e-01 9.991046e-01 9.992159e-01 9.993991e-01
  ED (26-34]    ED (34-45]    ED (45-99]     admin/dir         casat
9.998585e-01 9.999768e-01 9.999768e-01 9.999982e-01 1.000000e+00
[1] "P-values del clÃºster  5 "
  ED (21-26]   IN (50-100]      ME (0-1]        solter    ES (12-14]         obrer
4.441170e-35 2.399447e-12 9.635172e-08 9.860900e-05 2.371180e-04 4.238450e-03
       delta2      ME (1-2]    ES (10-12]    ES (16-20] IN (150-200] IN (100-150]
4.387618e-03 6.887168e-02 4.491676e-01 7.118634e-01 7.147842e-01 7.534859e-01
     ME (2-3]         beta2     IN (0-50]        gamma2         altres         alfa2
7.987232e-01 8.191872e-01 8.428278e-01 8.560318e-01 8.913044e-01 9.227379e-01
    admin/dir      ME (4-9]     ED (0-21]    ES (14-16]     ES (0-10]    ED (45-99]
9.512917e-01 9.575228e-01 9.647367e-01 9.823745e-01 9.926137e-01 9.940757e-01
     ME (3-4]    ED (26-34]    ED (34-45]         casat IN (200-250]
9.990056e-01 9.995883e-01 9.996477e-01 9.999014e-01 9.999347e-01
[1] "P-values del clÃºster  6 "
    ES (0-10]         obrer         alfa2 IN (100-150]    ED (45-99]     IN (0-50]
2.982955e-46 5.442467e-07 1.254779e-02 3.682999e-02 4.733374e-02 8.906259e-02
    ED (0-21]      ME (2-3]    ED (34-45]   IN (50-100]      ME (3-4]        gamma2
1.308456e-01 1.405717e-01 2.721274e-01 2.759328e-01 3.203973e-01 3.283786e-01
       casat      ME (1-2]        altres      ME (0-1]        solter         beta2
3.646672e-01 4.491650e-01 5.375405e-01 6.132272e-01 6.353328e-01 6.551587e-01
IN (150-200]      ME (4-9]    ED (21-26]    ES (16-20]    ED (26-34]        delta2
8.619166e-01 9.160025e-01 9.526988e-01 9.777415e-01 9.782452e-01 9.815722e-01
IN (200-250]    ES (14-16]    ES (10-12]    ES (12-14]     admin/dir
```

```
66 9.830857e-01 9.898901e-01 9.935264e-01 9.999569e-01 9.999988e-01
```

Així doncs, hem vist que les característiques que tenen els individus de cada clúster són:

- Clúster 1: ED (34-45]; casat; admin/dir; ES (14-16]; ME (4-9]; ES (16-20]; ME (3-4]; IN (150-200]; IN (100-150]
- Clúster 2: ED (45-99]; casat; ME (1-2]; ES (14-16]; admin/dir; IN (200-250]
- Clúster 3: ED (26-34]; ES (16-20]; ME (0-1]; admin/dir; ME (2-3]
- Clúster 4: ED (0-21]; solter; altres; ME (4-9]; ES (12-14]; IN (0-50]; IN (200-250]; ES (10-12]; delta2
- Clúster 5: ED (21-26]; IN (50-100]; ME (0-1]; solter; ES (12-14]; obrer; delta2; ME (1-2]
- Clúster 6: ES (0-10]; obrer; alfa2; IN (100-150]; ED (45-99]; IN (0-50]

I a partir d'això extreiem les següents conclusions dels individus de cada clúster:

- Clúster 1: Gent adulta casada que viu en família (de 4 o més membres), té molts estudis, una bona feina i uns ingressos elevats.
- Clúster 2: Gent gran casada que viu en parella, té molts estudis, una bona feina i uns ingressos molt elevats.
- Clúster 3: Gent adulta que viu en parella amb molts estudis i una bona feina.
- Clúster 4: Adolescents solters que pertanyen a famílies nombroses i estan estudiant.
- Clúster 5: Joves solters i obrers amb ingressos baixos.
- Clúster 6: Gent gran amb pocs estudis, obrers i amb uns ingressos baixos.

Capítol 5

Regles d'associació

Les regles d'associació treballen sobre un conjunt de dades anomenades transaccions. Cada fila de la taula de dades representa una transacció, mentre que cada columna representa un ítem i un ítem pot aparèixer a una transacció o no, però només un cop.
A continuació podem veure les freqüències d'aparició dels diferents ítems (en aquest cas, les categories de les variables) a les transaccions.

I aquest és el resultat un cop executat el codi per obtenir les regles:

```
> inspect(myalfarules)
  lhs                      rhs             support confidence     lift
1 {ringr=IN (200-250],
   rmemb=ME (0-1]}     => {mar2=alfa2} 0.01587302          1 3.111111
2 {redat=ED (34-45],
   ringr=IN (50-100]}  => {mar2=alfa2} 0.01190476          1 3.111111
3 {prof=obrer,
   restd=ES (0-10],
   rmemb=ME (1-2]}     => {mar2=alfa2} 0.01190476          1 3.111111
> inspect(mybetarules)
  lhs                      rhs             support confidence     lift
1 {mar1=beta1,
   redat=ED (45-99],
   rmemb=ME (1-2]}     => {mar2=beta2} 0.01190476          1 6.810811
2 {mar1=beta1,
   ringr=IN (100-150],
   rmemb=ME (1-2]}     => {mar2=beta2} 0.01190476          1 6.810811
3 {mar1=beta1,
   prof=admin/dir,
   rmemb=ME (1-2]}     => {mar2=beta2} 0.01587302          1 6.810811
> inspect(mygammarules)
  lhs                      rhs              support confidence   lift
1 {mar1=altra1,
   eciv=casat,
   restd=ES (12-14],
   rmemb=ME (2-3]}     => {mar2=gamma2} 0.01190476       1.00 10.500
2 {mar1=gamma1,
   ringr=IN (100-150]} => {mar2=gamma2} 0.01190476       0.75  7.875
3 {mar1=gamma1,
   eciv=casat}         => {mar2=gamma2} 0.01190476       0.60  6.300
> inspect(mydeltarules)
  lhs                   rhs              support confidence     lift
1 {ringr=IN (0-50],
   rmemb=ME (1-2]}   => {mar2=delta2} 0.01587302          1 2.290909
2 {redat=ED (26-34],
   ringr=IN (0-50]}  => {mar2=delta2} 0.01190476          1 2.290909
3 {mar1=beta1,
   rmemb=ME (0-1]}   => {mar2=delta2} 0.01587302          1 2.290909
```

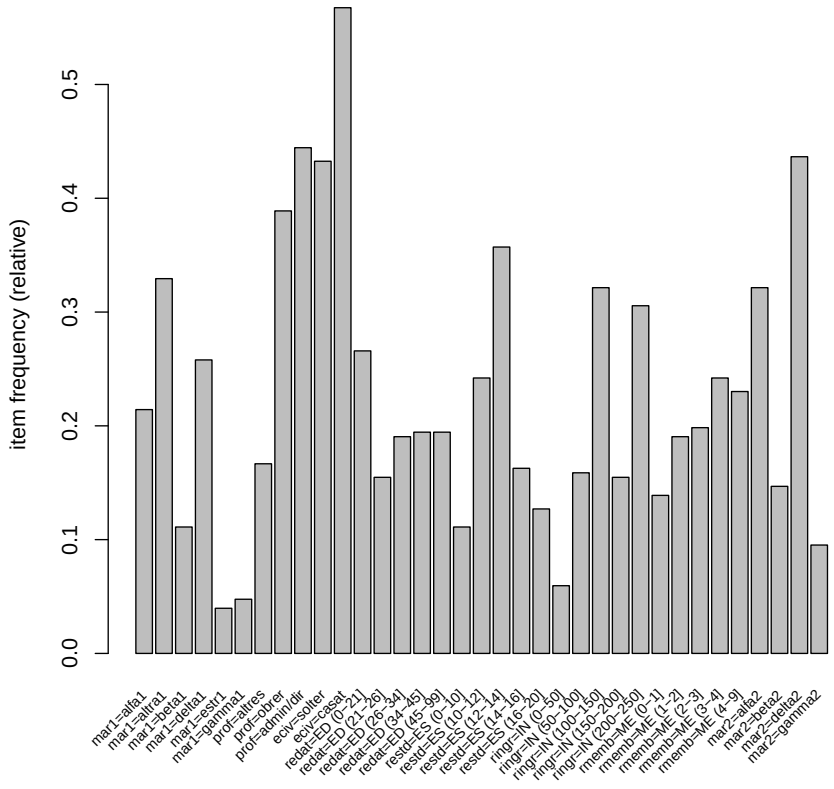


Figura 5.1: Freqüència dels items a les transaccions

5.1 Alfa

Les tres millors regles d'alfa ens indiquen que compren aquesta marca:

- Persones solteres amb molts ingressos
- Gent adulta amb ingressos mitjans-baixos
- Gent amb uns estudis molt bàsics i que viu en parella

5.2 Beta

Les tres millors regles de beta ens indiquen que compren aquesta marca:

- Gent gran que viu en parella i ja ha comprat beta anteriorment
- Persones que viuen en parella amb uns ingressos mitjans-baixos i ja han comprat beta anteriorment

- Administratius o directius que viuen en parella i ja han comprat beta anteriorment

5.3 Gamma

Les tres millors regles de gamma ens indiquen que compren aquesta marca:

- Persones casades amb un fill que tenen estudis universitaris i han comprat una altra marca anteriorment
- Gent amb uns ingressos mitjans-baixos que ja han comprat gamma anteriorment
- Persones casades que ja han comprat gamma anteriorment

5.4 Delta

Les tres millors regles de delta ens indiquen que compren aquesta marca:

- Gent amb uns ingressos baixos que viu en parella
- Persones de 27 a 34 anys que tenen uns ingressos baixos
- Gent que viu sola i anteriorment ha comprat beta

Capítol 6

Conclusions

En aquesta pràctica hem pogut veure una metodologia de Mineria de Dades per poder extreure coneixement d'un conjunt de dades que no teníem analitzat. Si bé la base de dades de partida contenia pocs individus per realitzar una bona explicació dels resultats, tots els passos seguits durant la pràctica han estat els adequats per realitzar en un cas més proper a la realitat, on el nombre d'individus és, com a mínim, de milers.

Per tant, primerament, a la secció 2.1 s'ha realitzat un anàlisi de les variables de partida, per comprovar si els individus seguien algun perfil o realment hi havia moltes dades ben distribuïdes. Com hem detectat que hi havia certs individus que tenien dades mancants n'hem hagut de fer la imputació a 2.2. Posteriorment hem categoritzat les variables contínues a 2.3 i hem visualitzat les dades a 2.4. Per acabar, hem aconseguit triar quines són les variables significatives a 2.5 encara que com en tenim un nombre força reduït hem seguit treballant amb totes elles i també hem esbrinat quins eren els perfils dels compradors de cada marca a 2.6.

A 3.1 hem analitzat quins eixos ens aportaven més informació i a 3.2 hem projectat el núvol de punts dels individus sobre els dos eixos més importants.

Amb el clústering hem aconseguit dividir la mostra en sis grups a 4.1 i a 4.2 n'hem explicat el significat.

Per acabar, hem generat regles d'associació que ens permeten predir el comportament dels individus depenent de les seves característiques.

Bibliografia

[1] "The R Project for Statistical Computing", http://www.r-project.org/, Institute for Statistics and Mathematics of the WU Wien

[2] "Mineria de Dades", http://www.lsi.upc.edu/~belanche/Docencia/mineria/mineria.html, Facultat d'Informàtica de Barcelona, UPC

[3] "R for beginners", http://www.lsi.upc.edu/~belanche/Docencia/mineria/Practiques/R/begin_R.pdf, E. Paradis

[4] "The R Guide", http://www.lsi.upc.edu/~belanche/Docencia/mineria/Practiques/R/Owen-TheRGuide.pdf, W. J. Owen