Alex Ferrando de las Morenas
Carlos Hurtado Comín
Maria Ribot i Vilà

Model de predicció del consum del cànnabis.

Objectius

L'objectiu últim d'aquest projecte és familiaritzar-nos amb l'estudi d'una base de dades fent servir mètodes d'aprenentatge automàtic. Això inclou el preprocessament de les dades, així com un estudi previ per tal d'entendre les dades de què disposem i després escollir els mètodes correctes en quant a tipus i distribució de les dades.

Per fer això, a les següents pàgines s'estudiarà una base de dades que conté dades demogràfiques de 1885 persones, així com els seus resultats en exàmens de personalitat i caràcter. També es disposa de dades relacionades amb el consum de diverses drogues.

Encara que tenim informació respecte al consum de diverses drogues, hem volgut fer un estudi del potencial d'una persona a consumir cànnabis, una droga ilegal, a partir de dades demogràfiques, de personalitat i del consum de drogues legals i socialment normalitzades com són l'alcohol i la nicotina. És per això que no es tindran en compte l'abús d'altres drogues.

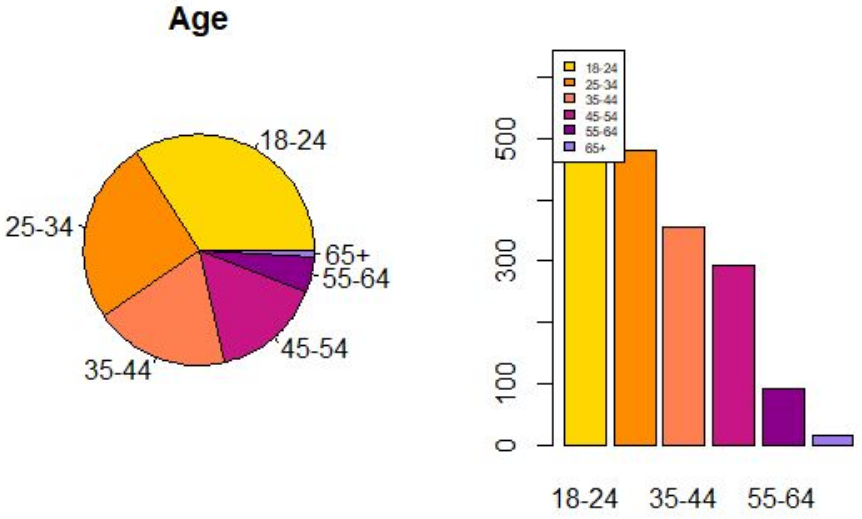
Descripció del conjunt de dades

Hem extret la base de dades del següent enllaç:
http://archive.ics.uci.edu/ml/datasets/Drug+consumption+%28quantified%29
És un repositori de dades per aprenentatge automàtic i a través d'aquí veiem que es tracta d'unes dades que originalment són del departament de Matemàtiques de la universitat de Leicester.

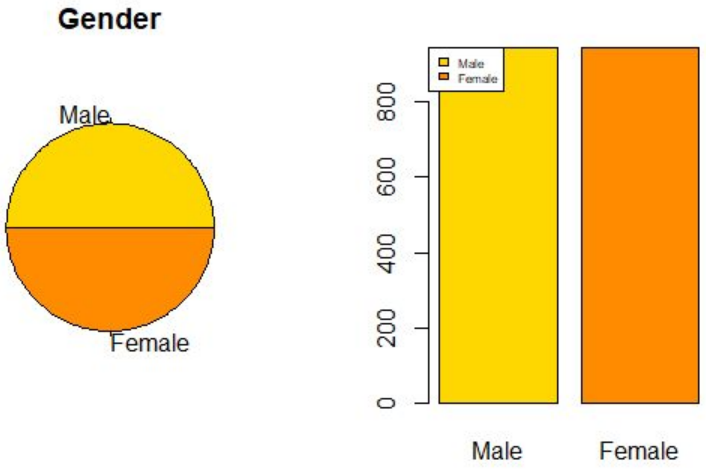
La base de dades va ser recollida per Elaine Fehrman entre el març de 2011 i el març de 2012 i les dades estan disponibles a la següent adreça:
https://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/36536/version/1

Aquest enllaç ens dona més informació sobre l'obtenció d'aquestes dades. L'estudi es va fer amb l'eina d'enquestes *Survey Gizmo* que permet obtenir dades amb anonimitat màxima, especialment rellevant pel tema específic que s'està estudiant. Es van aconseguir 2051 respostes, 166 de les quals no van satisfer els criteris de validació i per tant es van retirar i es va acabar amb 1885 participants .
Un dels criteris es va introduir per detectar afirmacions exagerades sobre el consum de substàncies va ser la introducció d'una droga fictícia per tal d'eliminar aquells qui afirmaven haver-la consumit.
El mètode de mostreig utilitzat s'anomena *Snowball* i es tracta d'una tècnica utilitzada sobretot per fer estudis estadístics de poblacions difícils d'accedir amb altres maneres d'aconseguir mostres poblacionals. La tècnica es basa en garantir l'anonimitat dels participants i que ells mateixos siguin els qui enviïn l'enquesta als seus contactes de manera encadenada. Això pot duu a models amb bastant de biaix ja que les mostres no acaben de ser independents, degut als cercles d'amistats. Això ho veiem amb els diferents grups poblacionals: Països d'origen, ètinies i edat, ja que la comunicació es fa amb grups de contacte i costa sortir-ne.
Aquesta és la mostra restant de 1885 participants amb les següents categories:

1. Edat: Majoritàriament població jove, dividida en els següents intervals:
   18-24 years (643; 34.1%), 25-34 years (481; 25.5%), 35-44 years (356; 18.9%), 45-54 years (294; 15.6%), 55-64 (93; 4.9%), and over 65 (18; 1%).



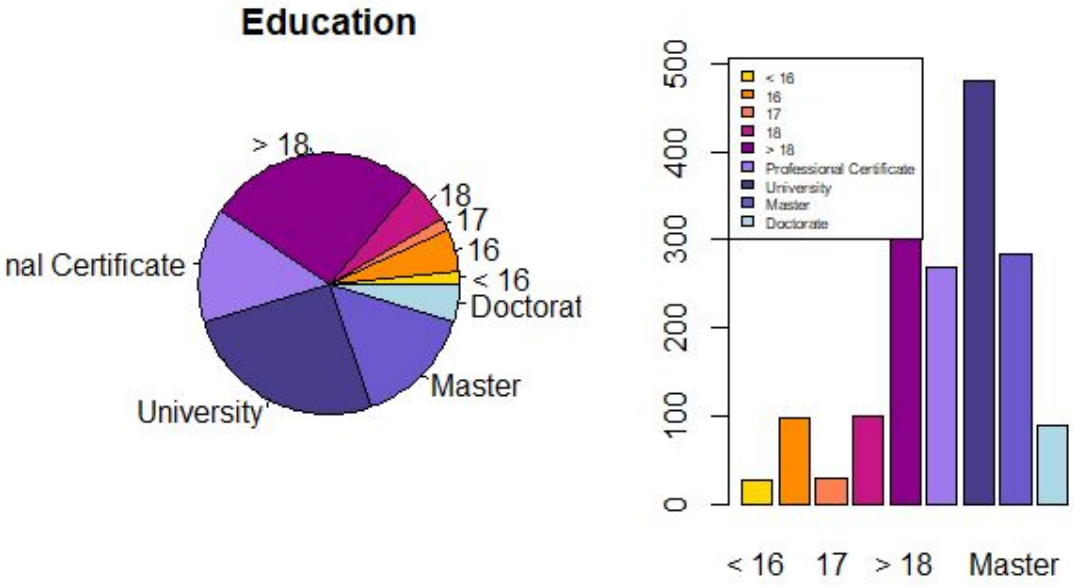
2. Gènere: male/female = 943/942.



3. Educació: Població amb alts nivells d'educació:

   14.4% (271) professional certificate or diploma, 25.5% (481) an undergraduate degree, 15% (284) a master's degree, and 4.7% (89) a doctorate.

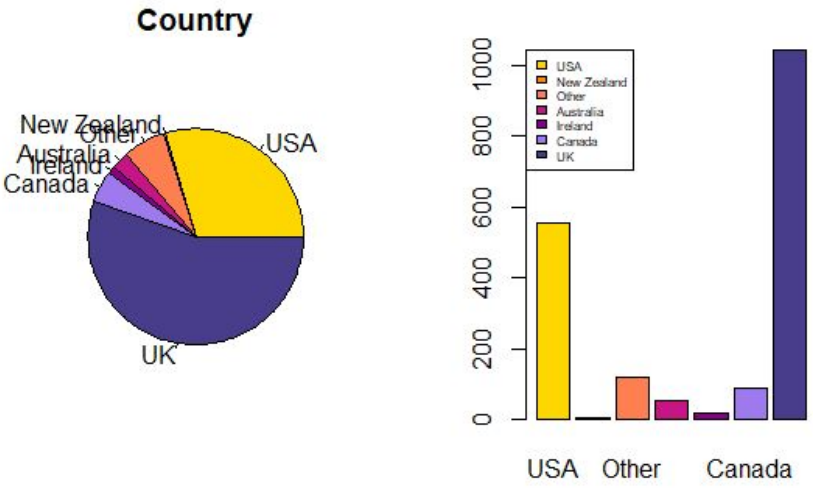
   26.8% (506) college without diploma; 13.6% (257) left school at the age of 18 or younger.



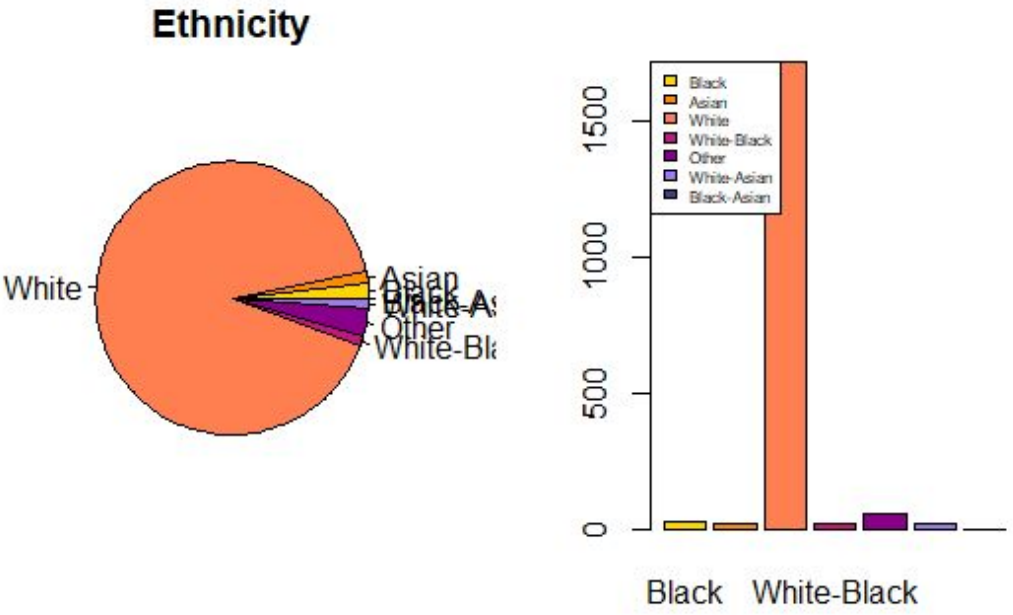
4. País d'origen: Majoritàriament països anglosaxons (93.7%), hi ha informació significativa dels següents:

   the UK (1044; 55.4%), the USA (557; 29.5%), Canada (87; 4.6%), Australia (54; 2.9%), New Zealand (5; 0.3%) and Ireland (20; 1.1%).

   Other(118 ,6.3%)



5. Ètnia o entorn cultural: majoritàriament gent blanca:
   (91.2%; 1720) reported being white, (1.8%; 33) stated they were Black, and (1.4%; 26) Asian. (5.6%; 106) described themselves as 'Other' or 'Mixed' categories.



6. Mesures de personalitat:
   a. Es va utilitzar el qüestionari *Revised NEO Five-Factor Inventory* (NEO-FFI-R) que és una mesura significant de els dominis bàsics de la personalitat
      versió online: https://www.truity.com/test/big-five-personality-test

   Aquest test consisteix en cinc factors:

i. Neurotisme (N): tendència a llarg termini a experimentar emocions negatives: nervis, tensió, ansietat o depressió.
ii. Extraversió (E): es manifesta en ser extrovertit, proper, actiu, assertiu, alegre i la cerca de característiques estimuladores.
iii. Obertura a noves experiències (O): apreciació general per l'art, idees poc convencionals o usuals i interessos eclèctics, imaginatius i creatius.
iv. Agrat (A) dimensió de relacions interpersonals: altruisme, confiança, modèstia, amabilitat, compassió i cooperativitat.
v. Consciència (C) tendència a ser ordenat i algú de qui es depèn, conviccions fortes, persistent, eficient.

b. També s'utilitza *Barratt Impulsiveness Scale* (BIS-11) que és un qüestionari de 30 preguntes que mesura la construcció de comportament impulsiu amb tres sub escales:
i. Impulsivitat motor: Actuar sense pensar
ii. Impulsivitat d'atenció: Poca concentració i intrusions al pensament.
iii. Poca planificació: falta de consideració de les conseqüències

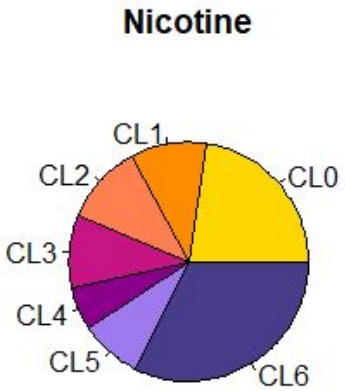
Podem realitzar el test i veure'n un estudi a la pàgina : https://scielo.conicyt.cl/pdf/rchnp/v51n4/art03.pdf

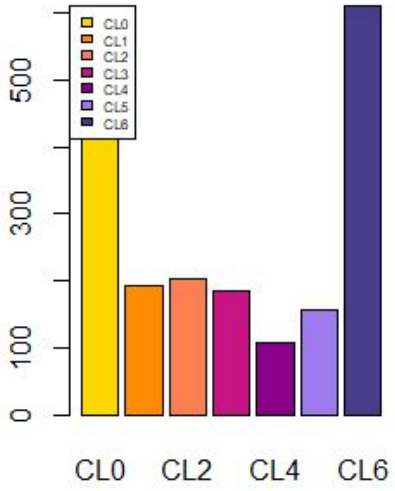
c. I per acabar també s'usa *Impulsiveness Sensation-Seeking* (ImpSS) com a eina de mesura del comportament, sobretot de la tendència a explorar noves sensacions

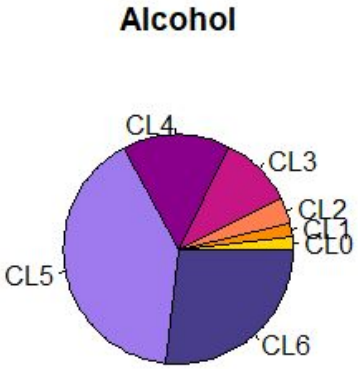
7. Consum de drogues:

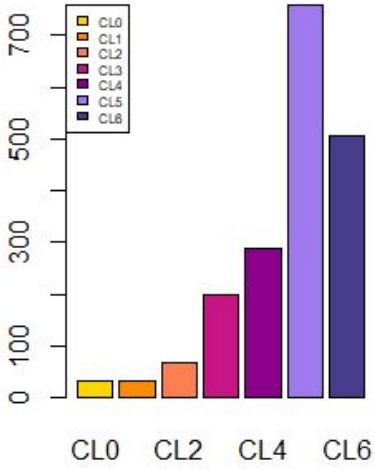
La base de dades conté 18 factors que descriuen el consum de diferents drogues en els perfils estudiats, degut a la quantitat de informació donada per l'enquesta i l'objectiu de la nostra pràctica hem reduït el nombre de categòriques que ho defineixen:
Obtenim de cada factor 7 nivells : 0 no ha consumit mai - 6 consumit avui i per la variable objectiu separem entre USER(0-2), NON-USER(3-6).

a. Variable objectiu: Consum de Cànnabis.
b. Variables categòriques escollides:









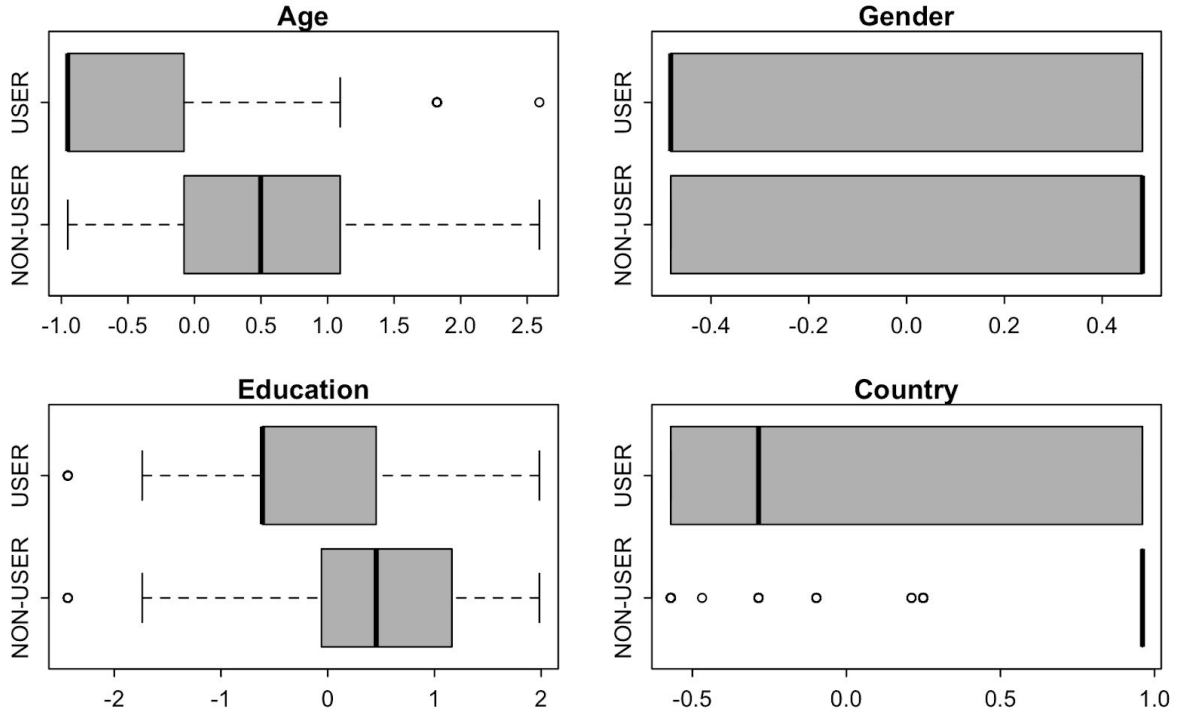
Al descarregar la base de dades vam observar que les dades no eren tal com pensavem, ja que en comptes de nivells descrits de manera intel·ligible vam observar que cada nivells era descrit per un nombre real sense significat aparent. Després de preprocessar les dades vam decidir mantenir aquests nivells per fer-ne els models ja que el resultat que obtenim és més ràpid i aparentment surt menys error.

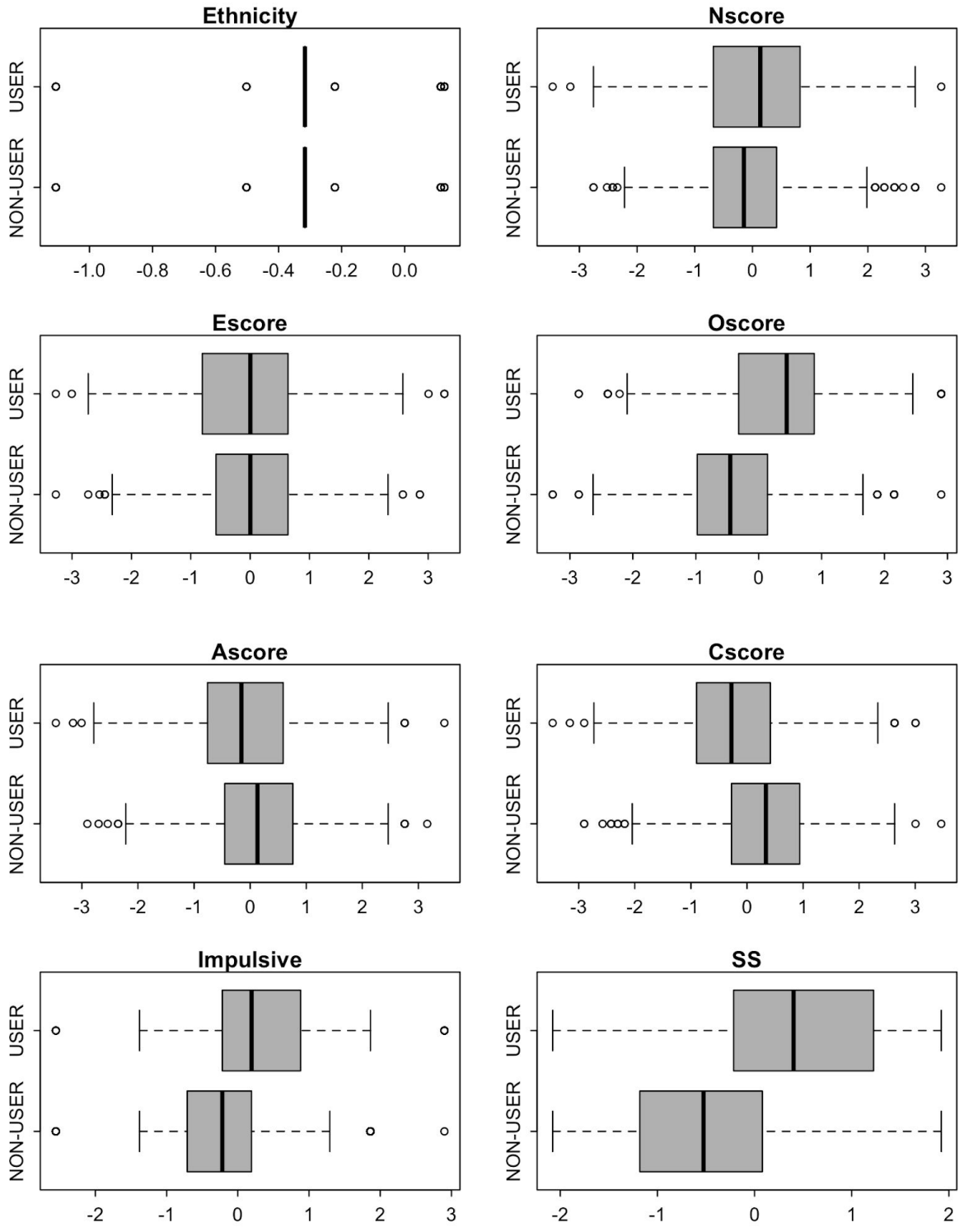
Per tal de visualitzar les diferents variables amb el conjunt de dades hem canviat els noms dels nivells per fer-ne les gràfiques i estudi previ, ja que l'objectiu de la pràctica és també extreure'n conclusions i aquest fet ho simplifica.

Anàlisis

Abans d'entrenar cap model per a la predicció del consum de cànnabis en persones adultes hem de fer una anàlisi exploratori de la base de dades per garantir el bon funcionament dels models predictors. En primer lloc, analitzarem aquells aspectes que podem distorsionar tant l'anàlisi previ com l'exploració del model.

Aplicant que no tenim cap mena de missings (ens ho diuen al mateix repositori de dades del qual hem descarregat les dades) només hem de buscar outliers importants que puguin modificar el comportament dels models lineals generalitzats que utilitzarem més endavant. Per a la cerca d'aquests outliers, graficarem y analitzarem els boxplots de la base de dades.

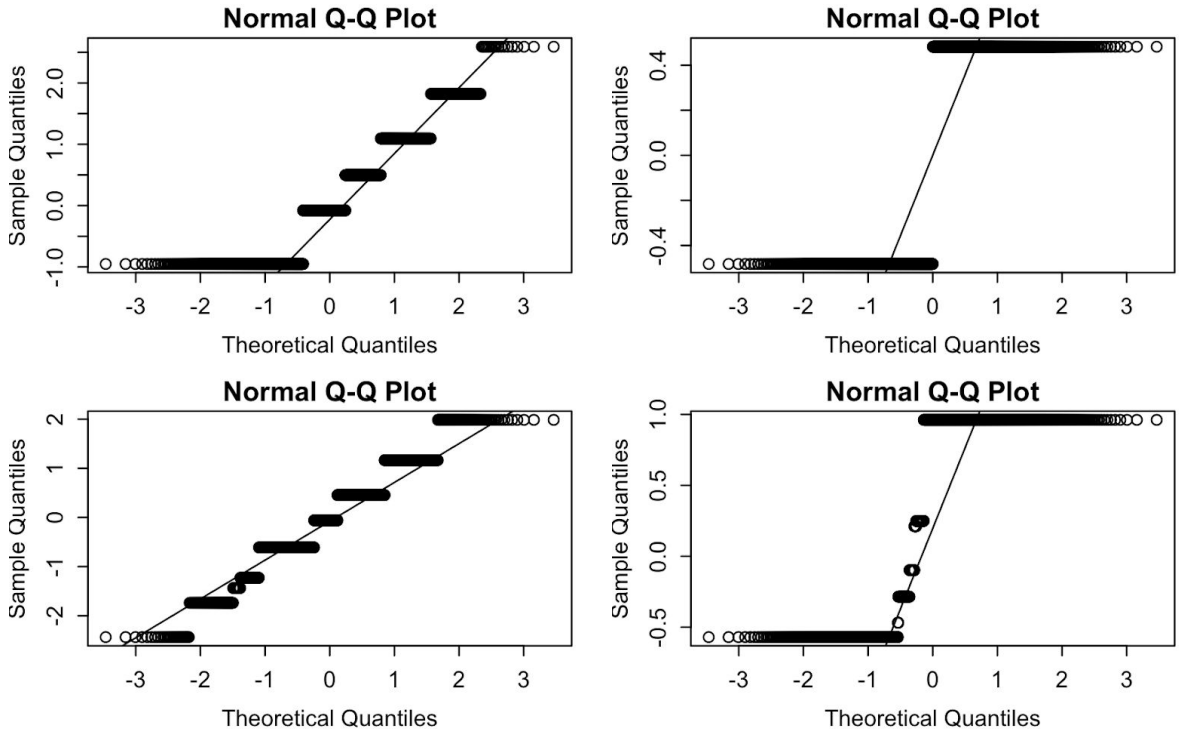


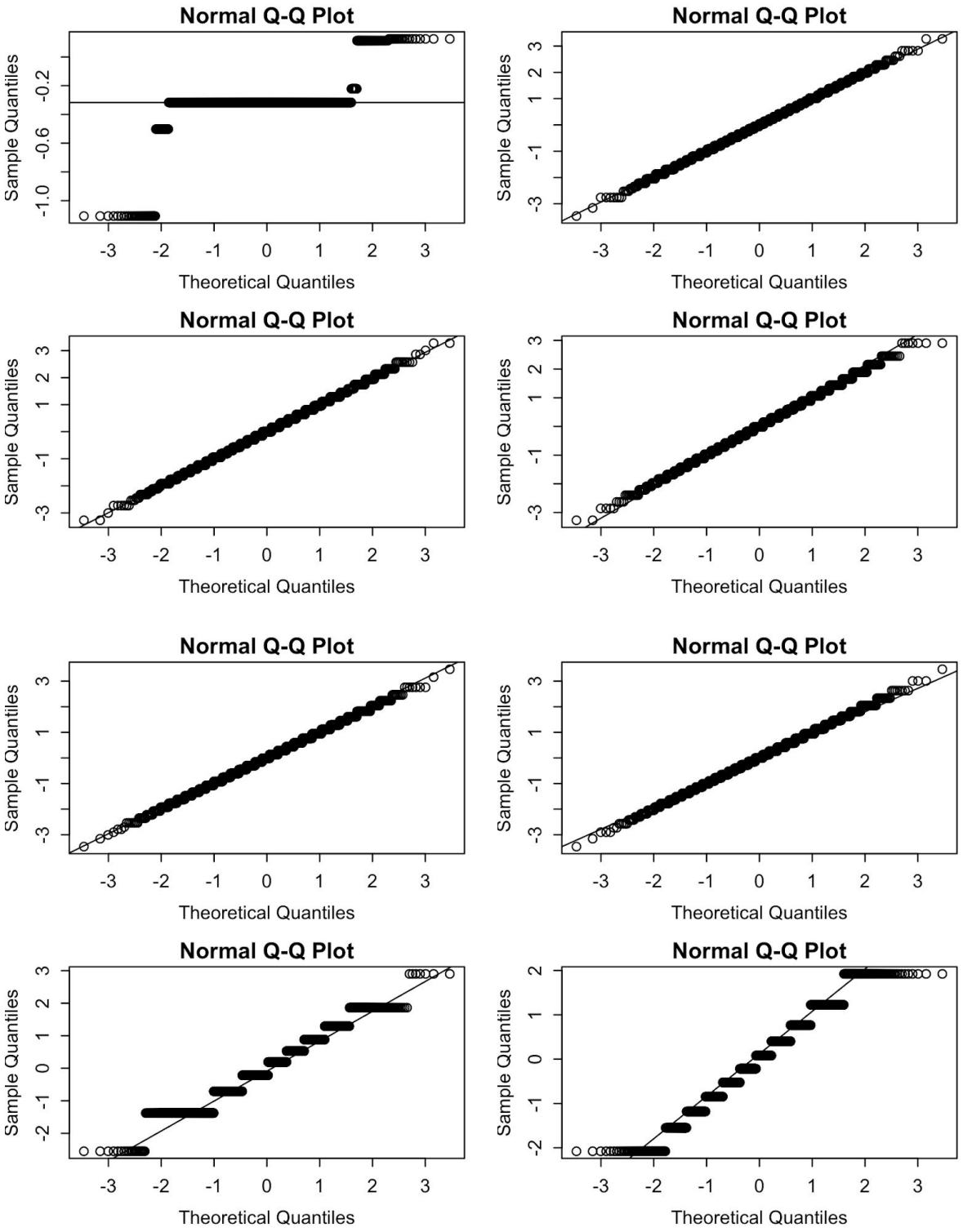


En els boxplots podem veure que la correcció aplicada a les variables funciona correctament i aquestes, tot i ser categòriques, es troben prou bé distribuïdes (boxplots no desequilibrats). Tot i això, observem que aquesta distribució no és prou bona per aquelles variables amb molt pocs nivells o que tenen moltes desigualtats poblacionals (com ara ètnia, on la presència de població blanca és molt major a les altres).

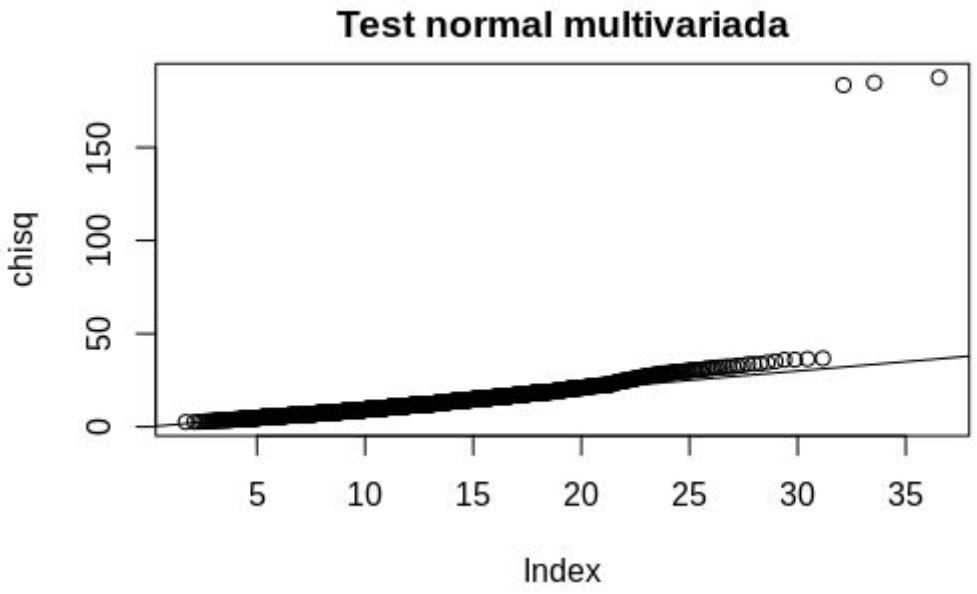
Ja en aquest anàlisi, podem veure el que semblen ser diferències significatives en les mitjanes de les variables la qual cosa acabarem de corroborar amb proves wilcox sobre les variables individualment. Així mateix, veiem certs valor anormals per a cada variable per separat, la qual cosa era d'esperar pel gran nombre de dades que tenim. En l'anàlisi multivariant que farem a continuació podrem detectar aquells outliers més importants i eliminar-los de la base de dades.

Per poder aplicar correctament models d'anàlisi discriminant (QDA i LDA), hem d'assegurar que la distribució conjunta de les variables explicatives aproximi una distribució normal en el gros de les dades que disposem. Per fer aquest anàlisi, grafiquem primerament els *qqplots* de cada variable per separat i veiem que, tot i ser variables categòriques, gràcies a la transformació numèrica que s'ha fet a la base de dades, les dades intenten aproximar una distribució normal. Aquesta observació es fa més òbvia a mesura que el nombre de nivells augmenta com veiem en els *qqplots* dels scores.

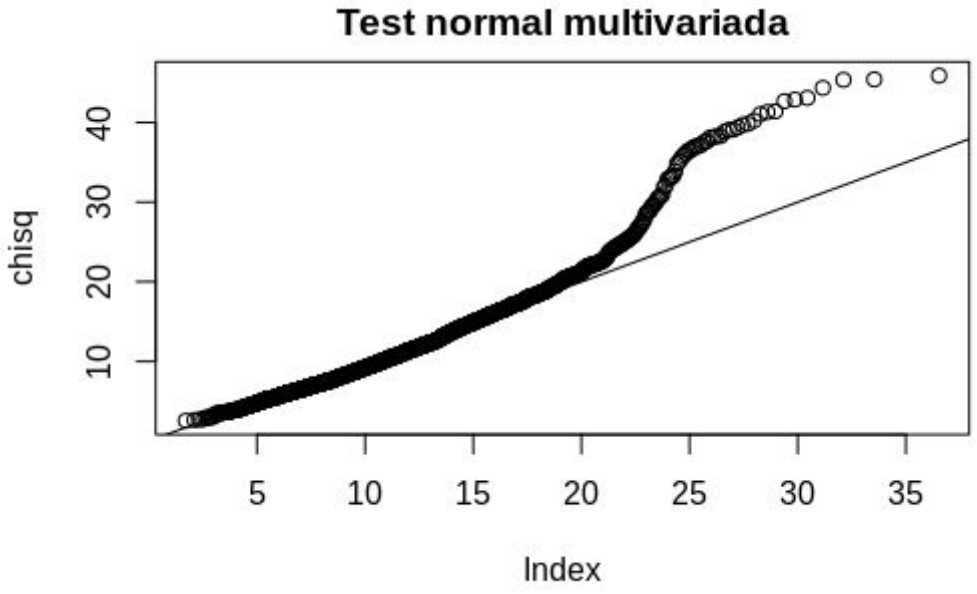




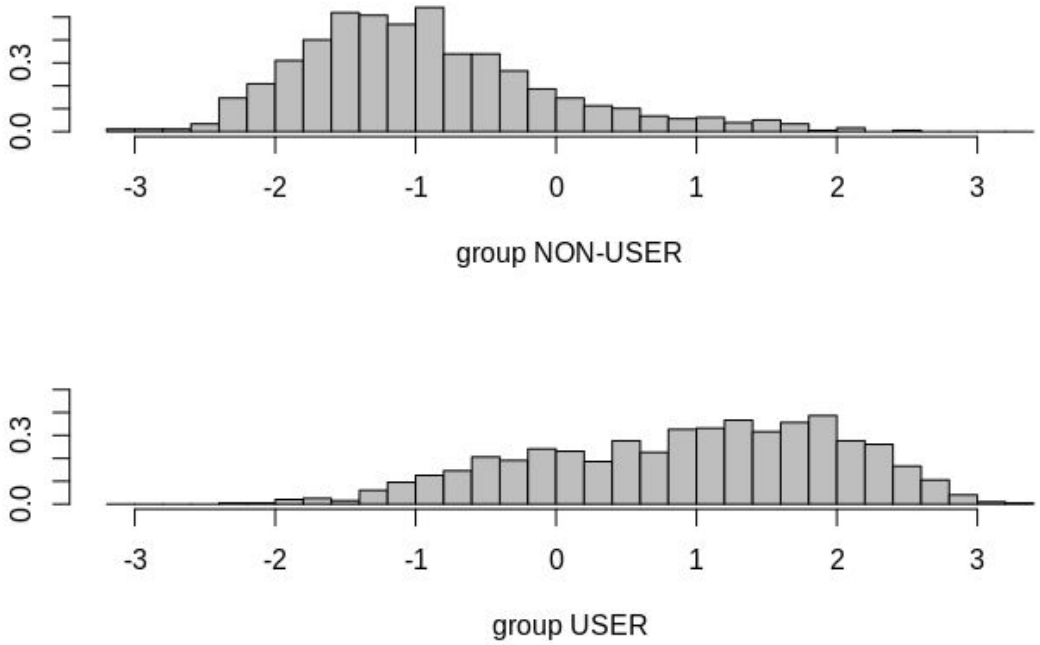
Observem ara, si la distribució conjunta de les dades aproximen prou bé una normal multivariada. Si fem l'anàlisi chisq amb la base de dades sencera obtenim:



Veiem tres outliers molt importants que poden afectar el comportament final dels model que entrenarem. Si eliminem aquestes tres mostres i tornem a repetir l'anàlisi observem:



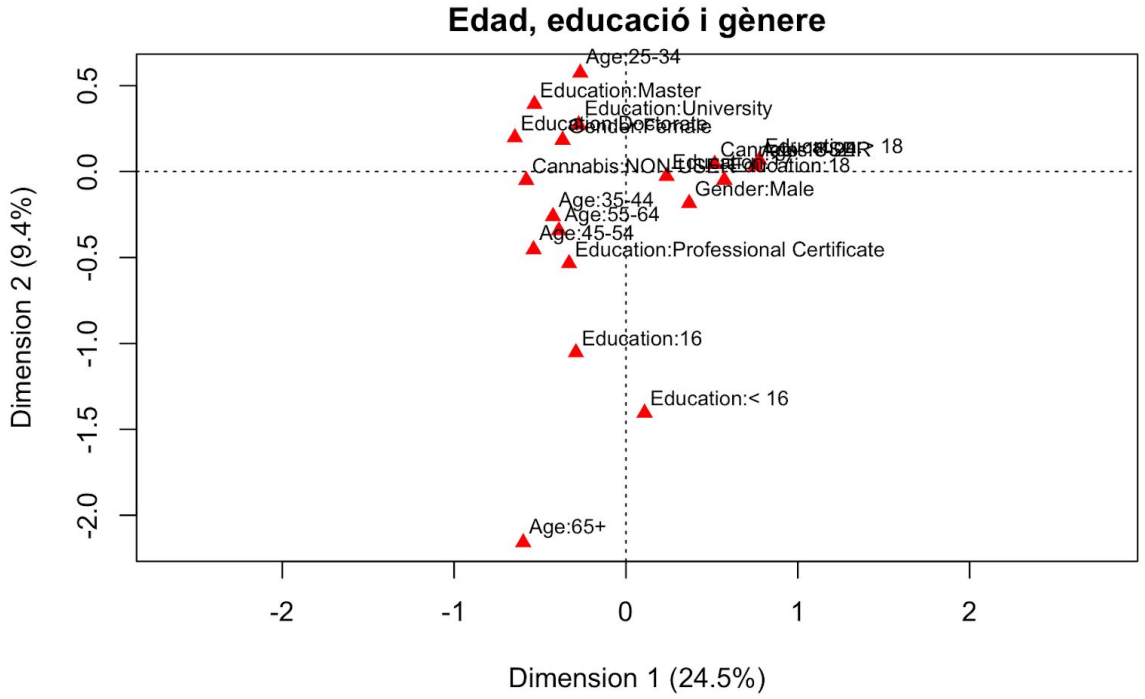
Ara, podem veure de més a prop la gràfica que abans i veure de fet, que la distribució conjunta de la base de dades s'allunya prou per a valors alts. Tanmateix, veiem que en conjunt podem suposar una aproximació a una normal multivariada. Si fem un plot del resultat de LDA sobre tota la base de dades veiem que els grups d'usuaris i no usuaris de cànnabis si que presenten diferències prou significatives. Considerant els resultats obtinguts podem considerar com a models vàlids QDA i LDA.



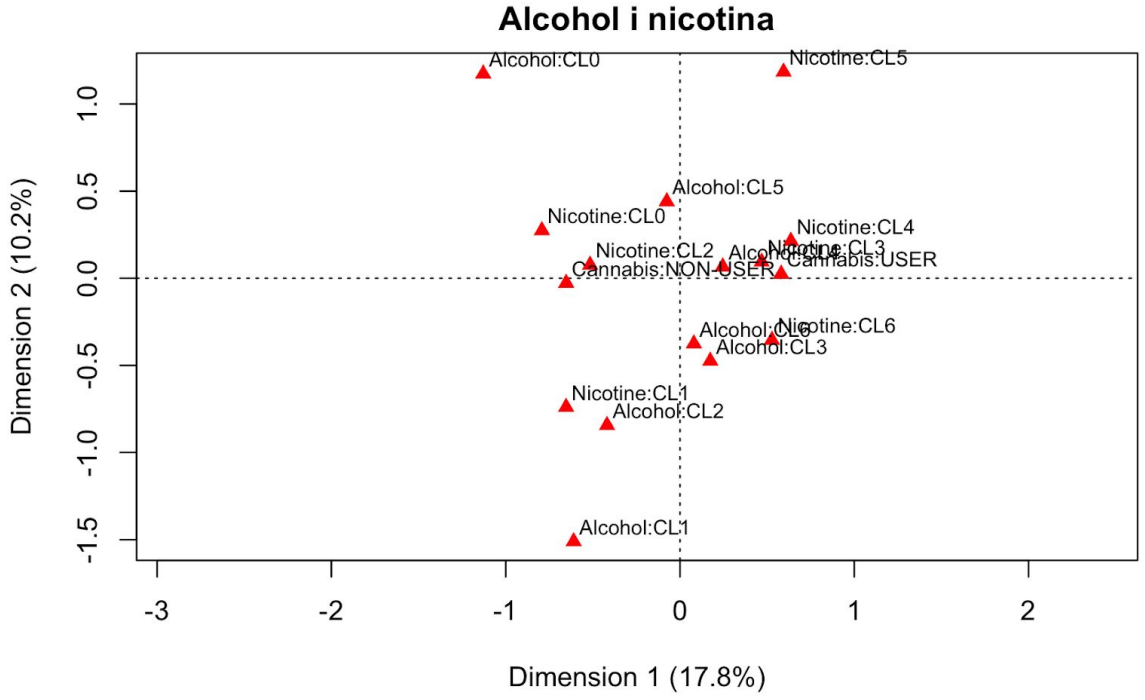
Com podem observar, disposem d'uns boxplots prou ben distribuïts gràcies a les transformacions de factors a variables numèriques comentades a l'apartat anterior sense la presència de outliers importants que ens pugin

Abans d'implementar els mètodes de classificació hem fet un estudi de la relació entre algunes variables per tal d'entendre el seu comportament així com la seva interacció. D'aquesta manera s'han fet algunes hipòtesis sobre el pes d'algunes variables sobre el consum de cànnabis. Aquest estudi s'ha fet aplicant anàlisi de correspondències sobre les variables que esperàvem que tinguéssin (o no) efectes significatius. Aquesta prova només es pot fer sobre variables categòriques, per això les variables que es volen estudiar s'han recodificat com a categòriques nomñes per l'anàlisi.

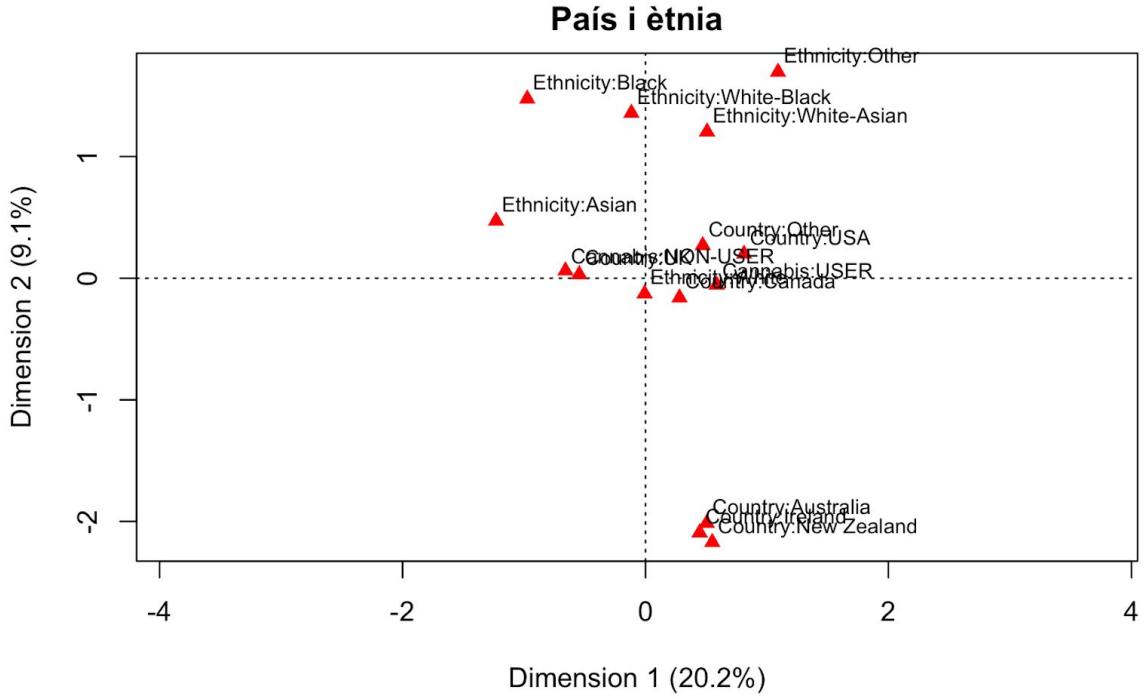
El primer que hem volgut visualitzar és l'ús de cànnabis en funció de l'edat, l'educació i el gènere de les persones consultades. De l'anàlisi següent caldria destacar que hi ha una forta associació entre els homes i l'ús de cànnabis. De la mateixa manera hi ha una associació de consum amb la gent jove i en especial que abandona els estudis universitaris. D'altra banda les dones i la gent més gran, així com aquelles persones amb formació superior estan associades a no utilitzar cànnabis.



També s'ha volgut estudiar la influència de drogues com l'alcohol i la nicotina sobre el consum de cànnabis. És clar que el consum de nicotina té una forta relació amb el consum de cànnabis, de la mateixa manera que no consumir alcohol ni nicotina son actituds associades al no ús. D'altra banda, el consum d'alcohol en diferents mesures no sembla estar relacionat amb el cànnabis.



Per últim, s'ha volgut contrastar els prejudicis en quant a país d'origen, ètnia i consum de drogues. A l'anàlisi següent queda clar que no hi ha cap relació entre l'etnia i el consum de cànnabis. L'única conclusió que podem treure és que a la base de dades hi ha poca gent d'Austràlia, Irlanda i Nova Zelanda, així com persones d'etnia negre-asiàtica.



Un cop estudiades les dades demogràfiques, falta comentar la influència dels tests comentats prèviament sobre l'ús de cànnabis. Per fer això es durà a terme un Wilcoxon test per a cada un dels tests per tal d'estudiar la diferència de mitjanes entre usuaris i no usuaris de cànnabis. L'hipòtesi nula del test és doncs la igualtat de mitjanes d'ambdós grups.

Per veure si els scores són significatius a l'hora de configurar un model de predicció comprovem si les mitjanes dels scores són diferents o no considerant si s'és usuari o no. Com podem comprovar en l'annex secció 1.6 observem que les dues categories sí que presenten diferències significatives. L'única categoria en la qual no tenim evidència suficient per rebutjar l'hipòtesi d'igualtat de variàncies és en la variable categòrica extraversió.