Pràctica 2 (35% nota final)

Presentació

En aquesta pràctica s'elabora un cas pràctic orientat a aprendre a identificar les dades rellevants per un projecte analític i usar les eines d'integració, neteja, validació i anàlisi de les mateixes. Per fer aquesta pràctica haureu de treballar en grups de fins a 3 persones, o si preferiu, també podeu fer-ho de manera individual. Haureu de lliurar un sol fitxer amb l'enllaç Github (https://github.com) on es trobin les solucions incloent els noms dels components de l'equip. Podeu utilitzar la Wiki de Github per descriure el vostre equip i els diferents arxius que corresponen a la vostra entrega. Cada membre de l'equip haurà de contribuir amb el seu usuari Github. Podeu utilitzar aquests exemples com guia:

- Exemple: https://github.com/Bengis/nba-gap-cleaning
- Exemple complex (fitxer adjunt).

Competències

En aquesta pràctica es desenvolupen les següents competències del Màster de Data Science:

- Capacitat d'analitzar un problema en el nivell d'abstracció adequat a cada situació i aplicar les habilitats i coneixements adquirits per abordar-lo i resoldre'l.
- Capacitat per aplicar les tècniques específiques de tractament de dades (integració, transformació, neteja i validació) per al seu posterior anàlisi.

Objectius

Els objectius concrets d'aquesta pràctica són:

- Aprendre a aplicar els coneixements adquirits i la seva capacitat de resolució de problemes en entorns nous o poc coneguts dintre de contextos més amplis o multidisciplinaris.
- Saber identificar les dades rellevants i els tractaments necessaris (integració, neteja i validació) per dur a terme un projecte analític.
- Aprendre a analitzar les dades adequadament per abordar la informació continguda en les dades.
- Identificar la millor representació dels resultats per tal d'aportar conclusions sobre el problema plantejat en el procés analític.
- Actuar amb els principis ètics i legals relacionats amb la manipulació de dades en funció de l'àmbit d'aplicació.
- Desenvolupar les habilitats d'aprenentatge que els permetin continuar estudiant d'una manera que haurà de ser en gran manera autodirigida o autònoma. Tipologia i cicle de vida de les dades
- Desenvolupar la capacitat de cerca, gestió i ús d'informació i recursos en l'àmbit de la ciència de dades.

Descripció de la PAC a realitzar

L'objectiu d'aquesta activitat serà el tractament d'un dataset, que pot ser el creat a la pràctica 1 o bé qualsevol dataset lliure disponible a Kaggle (https://www.kaggle.com). Alguns exemples de dataset amb els que podeu treballar són:

- Red Wine Quality (https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009).
- Titanic: Machine Learning from Disaster (https://www.kaggle.com/c/titanic).

L'últim exemple correspon a una competició activa a *Kaggle* de manera que, opcionalment, podeu aprofitar el treball realitzat durant la pràctica per entrar en aquesta competició. Seguint les principals etapes d'un projecte analític, les diferents tasques a realitzar (i justificar) són les següents:

1. Descripció del set de dades. Perquè és important i quina pregunta/problema pretén respondre?
2. Integració i selecció de les dades d'interès a analitzar.
3. Neteja de les dades.
    1. Les dades contenen zeros o elements buits? Com gestionaries aquests casos?
    2. Identificació i tractament de valors extrems.
4. Anàlisi de les dades.
    1. Selecció dels grups de dades que es volen analitzar/comparar (planificació dels anàlisis a aplicar).
    2. Comprovació de la normalitat i homogeneïtat de la variància.
    3. Aplicació de proves estadístiques per comparar els grups de dades. En funció de les dades i de l'objectiu de l'estudi, aplicar proves de contrast d'hipòtesis, correlacions, regressions, etc.
5. Representació dels resultats a partir de taules i gràfiques.
6. Resolució del problema. A partir dels resultats obtinguts, quines són les conclusions? Els resultats permeten respondre al problema?
7. Codi: Cal adjuntar el codi, preferiblement en R, amb el que s'ha realitzat la neteja, anàlisi i representació de les dades. Si ho preferiu, també podeu treballar en Python.

1. Descripció del set de dades

El set de dades que es vol estudiar prové de l'enllaç web següent:

https://www.kaggle.com/mehdidag/black-friday/home

Tal i com es descriu a la documentació (c.f. paràgraf següent), es tracta d'un set de dades que conté les transaccions realitzades en una botiga durant un període concret: El *Black Friday*. Aquestes dades han estat obtingudes amb l'objectiu d'analitzar els comportaments de compra dels usuaris durant aquest període en un local concret, per tal de poder predir els volums de compra, l'edat del client, el tipus de producte venut o bé simplement per classificar els clients en diferents clústers. Es tracta d'un set de dades pertanyent a un concurs (finalitza el 30 de Desembre, més informació a l'enllaç següent: https://datahack.analyticsvidhya.com/contest/black-friday/#problem_statement).

Description

The dataset here is a sample of the transactions made in a retail store. The store wants to know better the customer purchase behavior against different products. Specifically, here the problem is a regression problem where we are trying to predict the dependent variable (the amount of purchase) with the help of the information contained in the other variables.

Classification problem can also be settled in this dataset since several variables are categorical, and some other approaches could be "Predicting the age of the consumer" or even "Predict the category of goods bought". This dataset is also particularly convenient for clustering and maybe find different clusters of consumers within it.

Acknowledgements

The dataset comes from a competition hosted by Analytics Vidhya.

De manera addicional, es pot extreure molta informació en relació amb el comportament del client per cara a realitzar accions que fomentin el seu consum. Per exemple, es poden identificar quins rangs d'edat consumeixen més, o bé si hi ha diferencies entre gènere per tal d'enfocar les campanyes de màrqueting cap a un públic més concret.

Donat que es tracta de dades de caire real, la identitat i informació sensible dels clients no apareix. Més concretament, el set de dades es composa de 12 columnes que descriuen la informació següent:

Columna	***Descripció***
User_ID	Número de 7 xifres, identificador del client (ex. 1000001).
Product_ID	Número de 8 xifres, identificador del producte comprat. Per marcar que es un codi de producte ve encapçalat per una P (ex. P00069042).
Gender	Gènere del client (M o F per Masculí o Femení, respectivament).
Age	Edat del client per rangs de 10 anys (ex. 26-35).
Occupation	Codi numèric (ex. 20).
City_Category	Lletra que defineix la categoria de la ciutat (ex. A).

Stay_In_Current_City_Years	Nombre d'anys que el client ha residit a la ciutat (ex. 4+)
Marital_Status	Codi numèric que defineix l'estat civil del client (ex. 0).
Product_Category_1	Codi numèric que defineix la categoria del producte comprat (ex. 3).
Product_Category_2	Codi numèric que defineix la categoria del producte comprat (ex. 3).
Product_Category_3	Codi numèric que defineix la categoria del producte comprat (ex. 3).
Purchase	Cost del producte en cèntims de dòlar (ex. 1570).

Un dels principals inconvenients d'aquest format de dades és el fet de no disposar dels diccionaris per les diferents columnes, que ens permetrien arribar a les mateixes conclusions però permetent-nos contextualitzar més els resultat. Així doncs, podrem concloure quins productes son els més venuts, per exemple, però no sabrem de quins productes es tractarà més enllà del seu identificador. El mateix passarà per a la ocupació del client, el seu estat civil, o les categories de producte. D'altra banda, i tot i que sembla que podem identificar si el mateix client ha comprat dues vegades a la botiga (els codis d'usuari es repeteixen a mesura que processem el set de dades), no disposem de la hora de compra i per tant no podem saber si es tracta d'una compra o diverses. Voldrem analitzar doncs la compra diària i no tant, doncs, les vegades que el client ha anat a comprar.

2. Integració i selecció de les dades d'interès a analitzar

Per tal d'integrar les dades, llegirem en un entorn de R el fitxer en format CSV descarregat de *Kaggle*:

```
> vendes_bf <- read.csv("C:/Users/josepconsuegra/Desktop/BlackFriday.csv", he
ader=TRUE, sep=",")
```

Si analitzem el set de dades processat, veurem que disposem de 537577 files (productes comprats) i 12 atributs. En realitat, i per inspecció visual, veurem que cada fila correspon a la compra d'un sol producte, i per tant necessitarem agregar les dades per a trobar l'import total de compra del dia. D'altra banda, les variables categòriques ja apareixen com a tal, i en canvi certes variables numèriques que corresponen a categories no ho són encara. Cadrà doncs factoritzar aquestes columnes.

```
> str(vendes_bf)
'data.frame':  537577 obs. of  12 variables:
 $ User_ID                   : int  1000001 1000001 1000001 1000001 1000002 1
000003 1000004 1000004 1000004 1000005 ...
 $ Product_ID                : Factor w/ 3623 levels "P00000142","P00000242",
..: 671 2375 851 827 2733 1830 1744 3319 3597 2630 ...
 $ Gender                    : Factor w/ 2 levels "F","M": 1 1 1 1 2 2 2 2 2
2 ...
 $ Age                       : Factor w/ 7 levels "0-17","18-25",..: 1 1 1 1
7 3 5 5 5 3 ...
 $ Occupation                : int  10 10 10 10 16 15 7 7 7 20 ...
 $ City_Category             : Factor w/ 3 levels "A","B","C": 1 1 1 1 3 1 2
2 2 1 ...
 $ Stay_In_Current_City_Years: Factor w/ 5 levels "0","1","2","3",..: 3 3 3 3
5 4 3 3 3 2 ...
 $ Marital_Status            : int  0 0 0 0 0 0 1 1 1 1 ...
 $ Product_Category_1        : int  3 1 12 12 8 1 1 1 1 8 ...
 $ Product_Category_2        : int  NA 6 NA 14 NA 2 8 15 16 NA ...
 $ Product_Category_3        : int  NA 14 NA NA NA NA 17 NA NA NA ...
 $ Purchase                  : int  8370 15200 1422 1057 7969 15227 19215 158
54 15686 7871 ...
```

Així doncs, aplicarem la funció *factor* als atributs *Occupation*, *Marital_Status* i *Product_Category_N*:

```
> vendes_bf_net <- vendes_bf
> vendes_bf_net$Occupation <- as.factor(vendes_bf_net$Occupation)
> vendes_bf_net$Marital_Status <- as.factor(vendes_bf_net$Marital_Status)
> vendes_bf_net$Product_Category_1 <- as.factor(vendes_bf_net$Product_Categor
y_1)
> vendes_bf_net$Product_Category_2 <- as.factor(vendes_bf_net$Product_Categor
y_2)
> vendes_bf_net$Product_Category_3 <- as.factor(vendes_bf_net$Product_Categor
y_3)
> str(vendes_bf_net)
'data.frame':  537577 obs. of  12 variables:
 $ User_ID                   : int  1000001 1000001 1000001 1000001 1000002 1
000003 1000004 1000004 1000004 1000005 ...
 $ Product_ID                : Factor w/ 3623 levels "P00000142","P00000242",
..: 671 2375 851 827 2733 1830 1744 3319 3597 2630 ...
```

```
 $ Gender                    : Factor w/ 2 levels "F","M": 1 1 1 1 2 2 2 2 2
2 ...
 $ Age                       : Factor w/ 7 levels "0-17","18-25",..: 1 1 1 1
7 3 5 5 5 3 ...
 $ Occupation                : Factor w/ 21 levels "0","1","2","3",..: 11 11
11 11 17 16 8 8 8 21 ...
 $ City_Category             : Factor w/ 3 levels "A","B","C": 1 1 1 1 3 1 2
2 2 1 ...
 $ Stay_In_Current_City_Years: Factor w/ 5 levels "0","1","2","3",..: 3 3 3 3
5 4 3 3 3 2 ...
 $ Marital_Status            : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 2 2 2
2 ...
 $ Product_Category_1        : Factor w/ 18 levels "1","2","3","4",..: 3 1 12
12 8 1 1 1 1 8 ...
 $ Product_Category_2        : Factor w/ 17 levels "2","3","4","5",..: NA 5 N
A 13 NA 1 7 14 15 NA ...
 $ Product_Category_3        : Factor w/ 15 levels "3","4","5","6",..: NA 11
NA NA NA NA 14 NA NA NA ...
 $ Purchase                  : int  8370 15200 1422 1057 7969 15227 19215 158
54 15686 7871 ...
```

De cara a l'estudi que volem realitzar, ens interessa analitzar tots els camps excepte potser les categories de producte, en concret les dues categories addicionals (*Product_Category_1* i *Product_Category_2*), ja que no aportaran massa informació rellevant. D'altra banda, i enllaçant amb el proper punt, presenten molts valors no disponibles (NA). En qualsevol cas, analitzarem aquesta casuística en el proper apartat.

D'altra banda, la resta de camps ja es troba en un format adequat per a l'estudi i tots els atributs son rellevants respecte a les preguntes inicials que es plantegen. Així doncs, conservarem tots els camps que tenim disponibles.

3. Neteja de les dades

a. Les dades contenen zeros o elements buits? Com gestionaries aquests casos?

De cara a analitzar la presencia de zeros o elements buits, executem la comanda *summary* per a obtenir una descripció del set de dades:

```
> summary(vendes_bf_net)
   User_ID            Product_ID     Gender        Age           Occupation     City_Category Stay_In_Current_City_Years
Min.   :1000001   P00265242:  1858   F:132197   0-17 : 14707   4      : 70862   A:144638      0 : 72725
1st Qu.:1001495   P00110742:  1591   M:405380   18-25: 97634   0      : 68120   B:226493      1 :189192
Median :1003031   P00025442:  1586              26-35:214690   7      : 57806   C:166446      2 : 99459
Mean   :1002992   P00112142:  1539              36-45:107499   1      : 45971                 3 : 93312
3rd Qu.:1004417   P00057642:  1430              46-50: 44526   17     : 39090                 4+: 82889
Max.   :1006040   P00184942:  1424              51-55: 37618   20     : 32910
                  (Other)   :528149             55+  : 20903   (Other):222818
Marital_Status Product_Category_1 Product_Category_2 Product_Category_3    Purchase
0:317817       5      :148592     8      : 63058     16     : 32148     Min.   :  185
1:219760       1      :138353     14     : 54158     15     : 27611     1st Qu.: 5866
               8      :112132     2      : 48481     14     : 18121     Median : 8062
               11     : 23960     16     : 42602     17     : 16449     Mean   : 9334
               2      : 23499     15     : 37317     5      : 16380     3rd Qu.:12073
               6      : 20164     (Other):124975     (Other): 53569     Max.   :23961
               (Other): 70877     NA's   :166986     NA's   :373299
```

Aprofitem l'avinentesa per a descobrir diversos punts que analitzarem un a un:

- Existència de zeros com a valor dels atributs *Occupation*, *Stay_In_Current_City_Years* i *Marital_Status*:

En aquest cas, tots els zeros existents no representen absència de valor sinó que tenen un significat. Per al cas de l'ocupació, el 0 és l'identificador d'una categoria, i si disposéssim del diccionari trobaríem que la classe 0 correspon a "metge", per exemple. En el cas dels anys viscuts a la ciutat actual, el 0 indica que es tracta del primer any que el client resideix al seu domicili actual. En canvi, per a l'estat civil, observem que només existeixen dos valors, i que per tant podem assumir que es tracta d'un flag binari: 1 vol dir que el client està casat, 0 que no ho està.

- Existència d'elements buits en les columnes Product_Category_2 i Product_Category_3:

Segons el que podem veure a les dades, un producte pot tenir fins a 3 categories. D'aquesta manera, un producte amb totes tres categories informades no presentarà cap valor buit, mentre que un producte amb només dues categories tindrà un valor buit per a la columna Product_Category_3. Si fem extensiva aquesta explicació, un producte amb una sola categoria no tindrà informat cap valor per a les columnes Product_Category_2 i Product_Category_3.

b. Identificació i tractament de valors extrems.

Observant els resultats anteriors amb la comanda *summary* i analitzant els valors obtinguts per a la columna *purchase*, única variable numèrica, podem comprovar que el preu màxim dels productes (de manera individual) és 239.61$ i el preu mínim de 1.85$. Aquests valor extrems no semblen en cap cas anòmals, i per tant es consideraran dintre de l'estudi.

Donat que la resta d'atributs són categòrics, simplement comprovarem que no hi hagi cap categoria anòmala per a aquests atributs, més concretament per als camps *Occupation* i *Product_Category_N.*

```
> boxplot.stats(vendes_bf_net$Occupation)$out
factor(0)
Levels: 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Warning message:
In Ops.factor(x[floor(d)], x[ceiling(d)]) : '+' not meaningful for factors

> boxplot.stats(vendes_bf_net$Product_Category_1)$out
factor(0)
Levels: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
Warning message:
In Ops.factor(x[floor(d)], x[ceiling(d)]) : '+' not meaningful for factors

> boxplot.stats(vendes_bf_net$Product_Category_2)$out
factor(0)
Levels: 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
Warning message:
In Ops.factor(x[floor(d)], x[ceiling(d)]) : '+' not meaningful for factors

> boxplot.stats(vendes_bf_net$Product_Category_3)$out
factor(0)
Levels: 3 4 5 6 8 9 10 11 12 13 14 15 16 17 18
Warning message:
In Ops.factor(x[floor(d)], x[ceiling(d)]) : '+' not meaningful for factors
```

Observem que totes les categories són numèriques i que es troben dins un rang acceptable de valors ([1;18] per a *Product_Category_N,* [0;20] per a *Occupation*), per tant entendrem que o bé no existien valors extrems incoherents de base, o be el set de dades ja ha estat tractat prèviament per a realitzar una primera neteja de dades.

Com podrem veure en l'apartat d'Anàlisi de dades, però, ens interessarà obtenir l'import total de compra d'un client. Com a apunt interessant, si tornem a comprovar els valors extrems un cop hem realitzat agregació de les compres, podem veure que ara n'apareixen 424 sobre el total de mesures que estem considerant (5891):

```
> length(boxplot.stats(vendes_bf_agg$Purchase)$out)

[1] 424
```

En qualsevol cas, aquests valors semblen correctes i res ens fa pensar que siguin valors anòmals. Tot el contrari, són dades que expliquen una realitat i ens han d'ajudar a entendre aquesta realitat. Així doncs, aquests valors es consideraran dintre de l'estudi.

4. Anàlisi de les dades i representació dels resultats a partir de taules i gràfiques.

Aquest set de dades permet flexibilitat a l'hora d'estudiar el comportament dels clients de l'establiment, donat que podem analitzar diversos factors que poden resultar molt interessants, i per tant dintre de l'espectre de preguntes que ens agradaria poder respondre trobaríem, de manera no exhaustiva, les següents :

- ¿Un dels dos gèneres compra més que l'altre? ¿Quin?
- ¿Hi ha altres factors que influeixin directament en el volum de compra?
- ¿Podem predir el consum de l'usuari?

a. Selecció dels grups de dades que es volen analitzar/comparar (planificació dels anàlisis a aplicar).

Per tal de respondre a les preguntes esmentades prèviament, cal entendre primerament el format de les dades. En aquest sentit, estem tractant dades transaccionals separades per productes. Així doncs, per un mateix client A disposarem de tots els productes que ha comprat de manera individual, cadascun d'ells corresponent a una fila del nostre set de dades. Per tal d'analitzar l'import i volum de compra necessitarem primer de tot fer un agregat de les dades, de cara a disposar de la següent informació:

- El codi de client
- Els camps *Gender*, *Age*, *Occupation*, *City_Category*, *Stay_In_Current_City_Years* i *Marital_Status* per a analitzar correlacions.
- L'import total de compra
- El nombre de productes comprats
- El preu mig dels productes comprats (import total dividit per productes comprats)

Deixarem de banda, d'entrada, les columnes *Product_Category_N* ja que hi ha altres preguntes més interessants sobre el set de dades. Addicionalment, a l'hora de fer el total agregat per client estaríem perdent informació sobre la categoria dels productes, i per tant s'hauria de plantejar una altra estratègia per a analitzar com afecta la categoria dels productes i com gestionar productes multi-categòrics.

Procedim doncs a realitzar l'agregació de les dades. En primer lloc calcularem l'import total, i, tot seguit, el nombre de productes comprats per l'usuari, que ajuntarem a una mateixa taula:

```
> library(plyr)
> library(data.table)
> vendes_bf_count <- count(vendes_bf_net, c("User_ID"))
> vendes_bf_sum <- aggregate(Purchase ~ User_ID + Gender + Age + Occupation +
City_Category + Stay_In_Current_City_Years + Marital_Status, vendes_bf_net, s
um)
> vendes_bf_agg <- join(vendes_bf_sum, vendes_bf_count, by="User_ID", type="l
eft", match="all")
> View(vendes_bf_agg)
> vendes_bf_agg$mean_price <- vendes_bf_agg$Purchase/vendes_bf_agg$freq
> setnames(vendes_bf_agg, old=c("freq"), new=c("Items"))
> setnames(vendes_bf_agg, old=c("mean_price"), new=c("Average_Price"))
```

	User_ID	Gender	Age	Occupation	City_Category	Stay_In_Current_City_Years	Marital_Status	Purchase	Items	Average_Price
1	1000034	F	18-25	0	A	0	0	807747	99	8159.061
2	1000524	M	18-25	0	A	0	0	5212846	558	9342.018
3	1003389	M	18-25	0	A	0	0	2580955	321	8040.358
4	1003789	F	26-35	0	A	0	0	844854	76	11116.500
5	1001579	M	26-35	0	A	0	0	3977702	444	8958.788
6	1003678	M	26-35	0	A	0	0	773414	71	10893.155
7	1004643	M	26-35	0	A	0	0	885018	116	7629.466
8	1005720	M	26-35	0	A	0	0	1463389	178	8221.287
9	1003217	F	36-45	0	A	0	0	1882898	201	9367.652

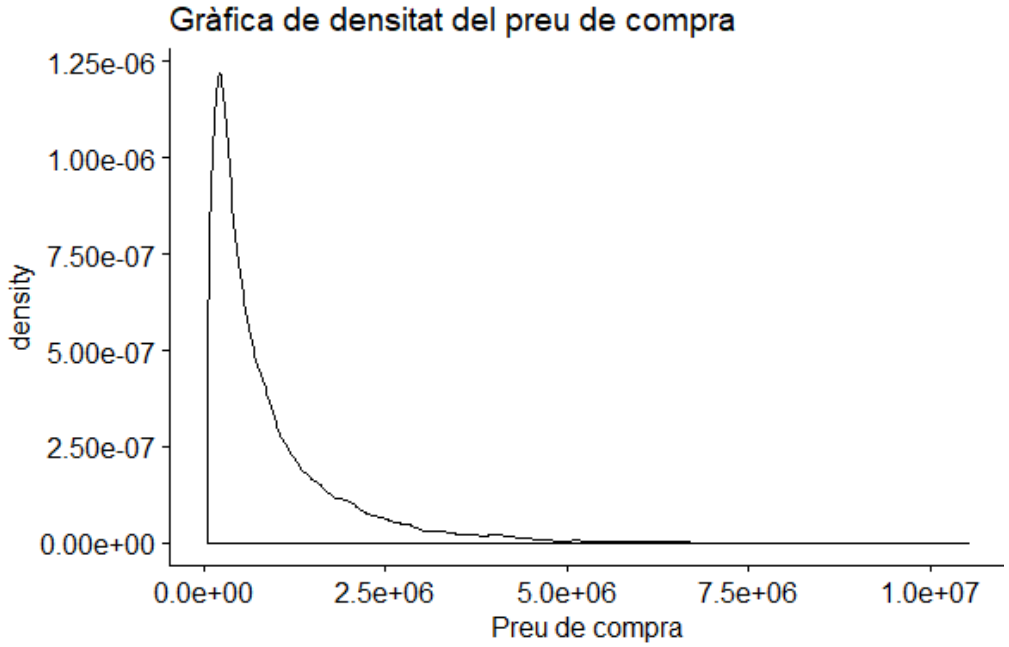
b. Comprovació de la normalitat i homogeneïtat de la variància.

Per tal de comprovar la normalitat de la mostra, comprovarem per l'import de compra si es segueix una distribució normal, tal i com es realitza en l'exemple del mòdul teòric:

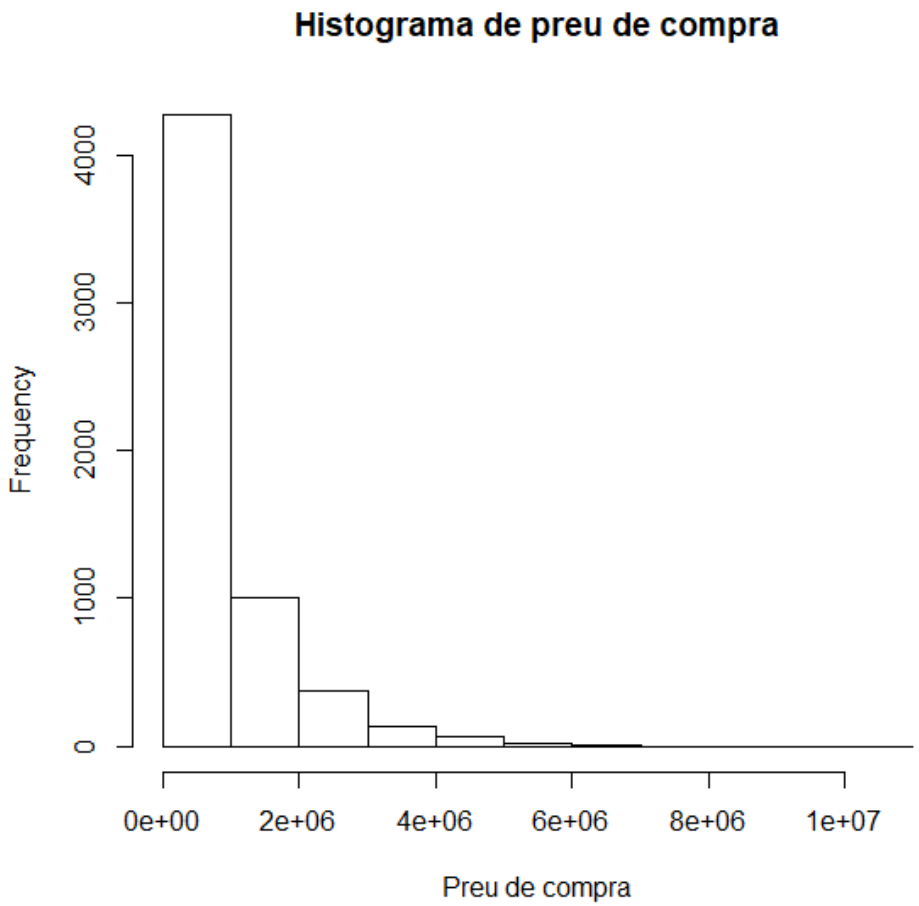
```
> library(nortest)
> p_Val=ad.test(vendes_bf_agg$Purchase)$p.value
> p_Val
[1] 3.7e-24
```

Vistos els resultats, sembla que el preu de compra segueix una distribució normal, donat que el p-valor obtingut és clarament inferior a 0.05. Si intentem reproduir gràficament aquest fet, observem el següent:

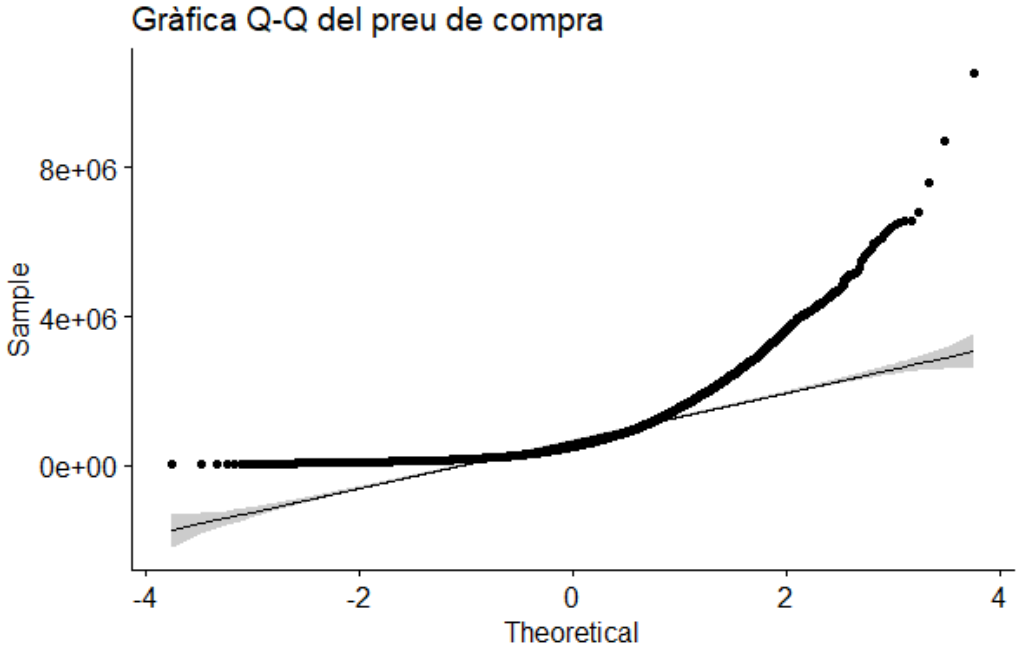
```
> library(ggpubr)
> ggdensity(vendes_bf_agg$Purchase, main = "Gràfica de densitat del preu de c
ompra", xlab = "Preu de compra")
```




```
> hist(vendes_bf_agg$Purchase, main="Histograma de preu de compra", xlab="Pre
u de compra")
```




```
> ggqqplot(vendes_bf_agg$Purchase, main="Gràfica Q-Q del preu de compra")
```

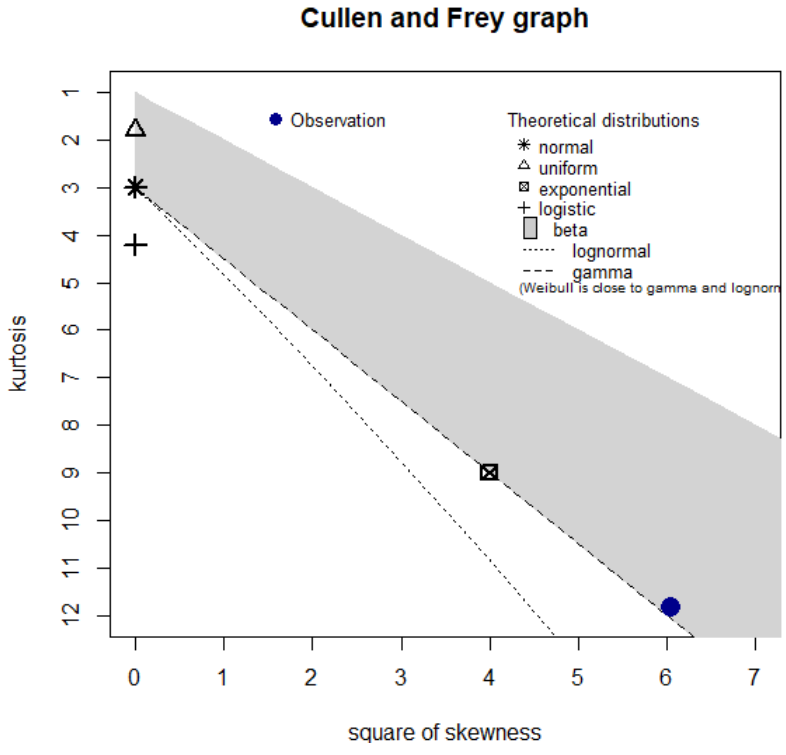


Podem observar com clarament la mostra no segueix una distribució normal (els valors de la gràfica Q-Q no es troben dintre de l'àrea grisa), fet que veiem demostrat també per el test de Kolmogorov-Smirnov, on el p-valor obtingut és molt inferior a 0.05:

```
> library(vcd)
> library(MASS)
> ks.test(vendes_bf_agg$Purchase, "pnorm", mean=mean(vendes_bf_agg$Purchase),
sd=sd(vendes_bf_agg$Purchase))

        One-sample Kolmogorov-Smirnov test

data:  vendes_bf_agg$Purchase
D = 0.19554, p-value < 2.2e-16
alternative hypothesis: two-sided
```

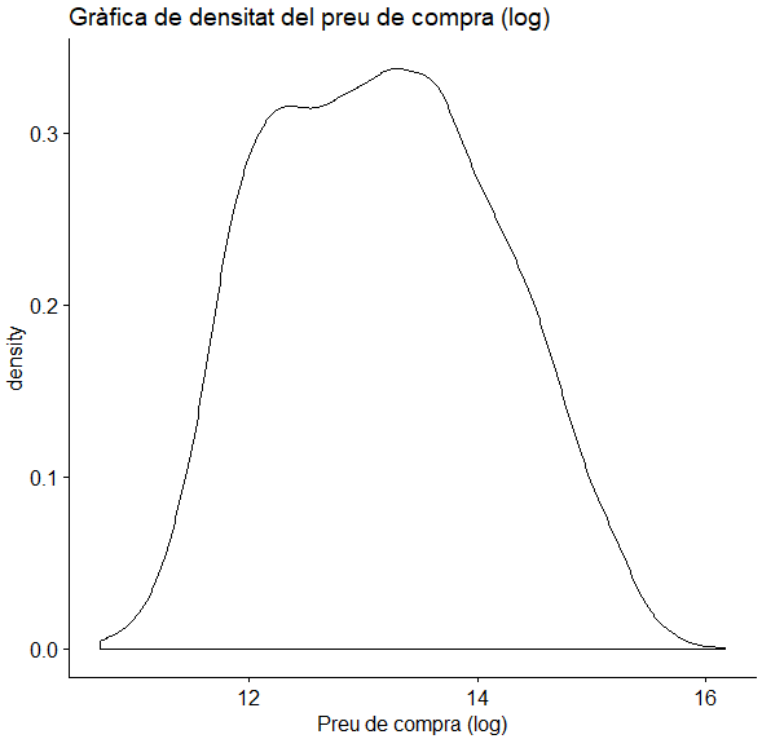
De cara a intentar endevinar quina distribució pot representar el nostre set de dades, podem generar la corba de Cullen i Frey i estimar a *grosso modo* la nostra distribució:

```
> library(fitdistrplus)
> library(logspline)
> descdist(vendes_bf_agg$Purchase, discrete = FALSE)

summary statistics
------
min:  44108   max:  10536783
median:  512612
mean:  851751.5
estimated sd:  932997.8
estimated skewness:  2.459439
estimated kurtosis:  11.83409
```

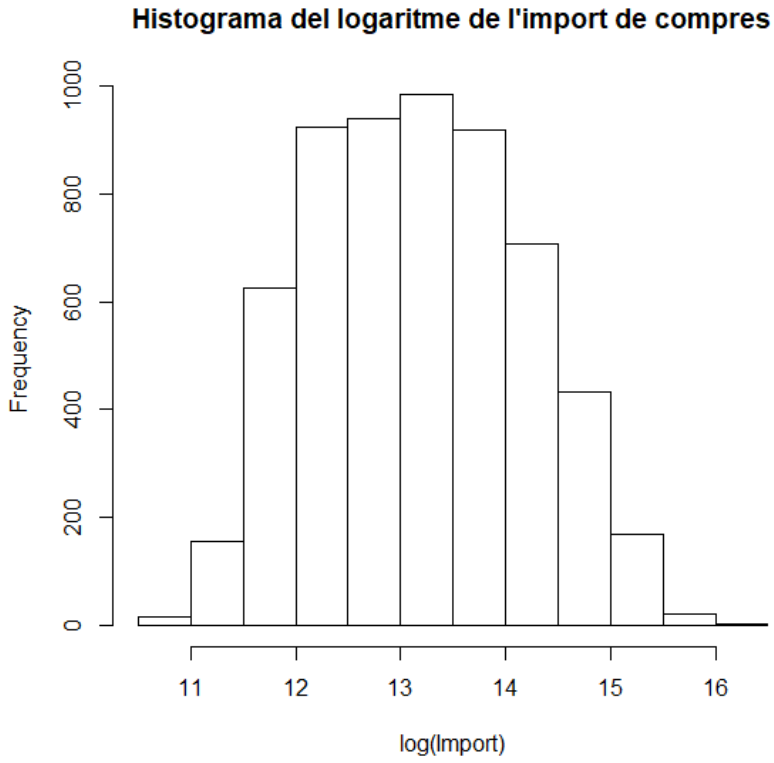


Seguint la gràfica obtinguda, sembla que la nostra mostra segueix o bé una distribució gamma o bé una distribució beta. De cara a intentar obtenir una distribució normal, podríem intentar encara realitzar una transformació de les dades a escala logarítmica. En aquest cas:

```
> vendes_bf_log = log(vendes_bf_agg$Purchase)
> ggdensity(vendes_bf_log, main = "Gràfica de densitat del preu de compra (lo
g)", xlab = "Preu de compra (log)")
```

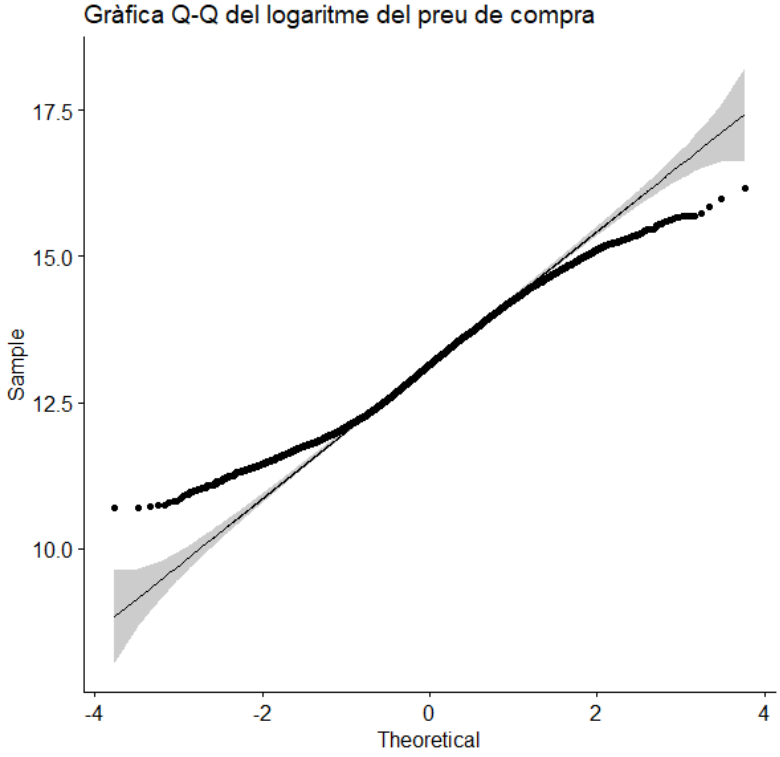



```
> hist(vendes_bf_log, main="Histograma del logaritme de l'import de compres",
xlab = "log(Import)")
```



Un cop aplicada la transformació, aquest set de dades sembla més proper a una distribució normal, com es pot apreciar en el gràfic Q-Q. De totes maneres, sembla que aquesta distribució de les dades tampoc es pot considerar normal.

```
> ggqqplot(vendes_bf_log, main="Gràfica Q-Q del logaritme del preu de compra")
```



Per evitar interpretacions incorrectes, podem avaluar la normalitat de la distribució com en el cas anterior. Estimarem en primer lloc quins paràmetres podrien correspondre al set de dades si seguís una distribució normal, i sobre aquests paràmetres avaluarem la normalitat de la distribució amb el test de Kolmogov-Smirnov.

```
> fitdistr(vendes_bf_log, "normal")
       mean             sd
  13.170003263    0.992650310
 ( 0.012933074) ( 0.009145065)

> ks.test(vendes_bf_log, "pnorm", mean=13.17, sd=0.99265)

        One-sample Kolmogorov-Smirnov test

data:  vendes_bf_log
D = 0.043241, p-value = 5.416e-10
alternative hypothesis: two-sided
```

Veiem que el p-valor, tot i ser més alt, no compleix amb el llindar de 0.05 imposat per a afirmar la hipòtesi de que les dades segueixen una distribució normal.

D'altra banda, si analitzem la homogeneïtat de la variança, tal i com s'explica en l'exemple del mòdul teòric, veiem que no obtenim en cap cas un p-valor que permeti acceptar la hipòtesi d'homogeneïtat de variances:

```
> fligner.test(Purchase ~ Gender, data = vendes_bf_agg)

        Fligner-Killeen test of homogeneity of variances

data:  Purchase by Gender
Fligner-Killeen:med chi-squared = 121.01, df = 1, p-value < 2.2e-16

> fligner.test(Purchase ~ Occupation, data = vendes_bf_agg)

        Fligner-Killeen test of homogeneity of variances

data:  Purchase by Occupation
Fligner-Killeen:med chi-squared = 139.93, df = 20, p-value < 2.2e-16

> fligner.test(Purchase ~ Marital_Status, data = vendes_bf_agg)

        Fligner-Killeen test of homogeneity of variances

data:  Purchase by Marital_Status
Fligner-Killeen:med chi-squared = 10.11, df = 1, p-value = 0.001474
```

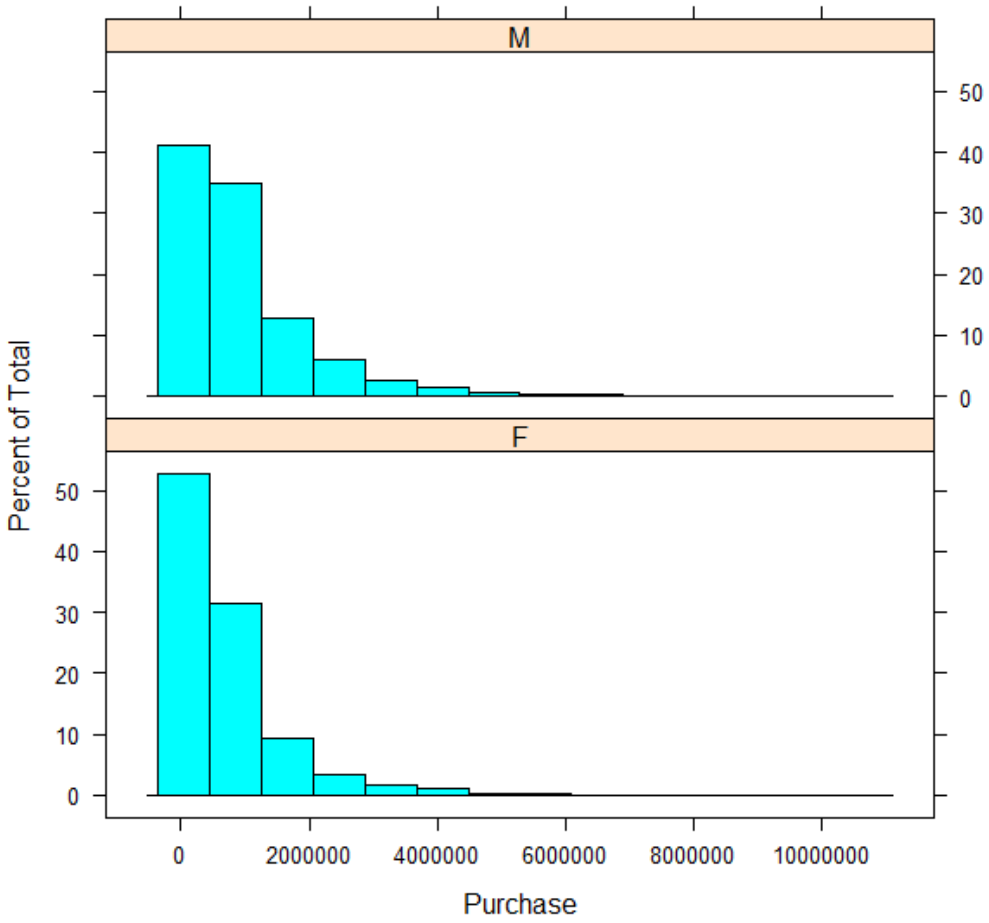
c. Aplicació de proves estadístiques per comparar els grups de dades. En funció de les dades i de l'objectiu de l'estudi, aplicar proves de contrast d'hipòtesis, correlacions, regressions, etc.

Ens centrarem en aquest apartat en analitzar les següents problemàtiques i decidir quins tests hauríem d'aplicar per poder respondre a aquestes problemàtiques:

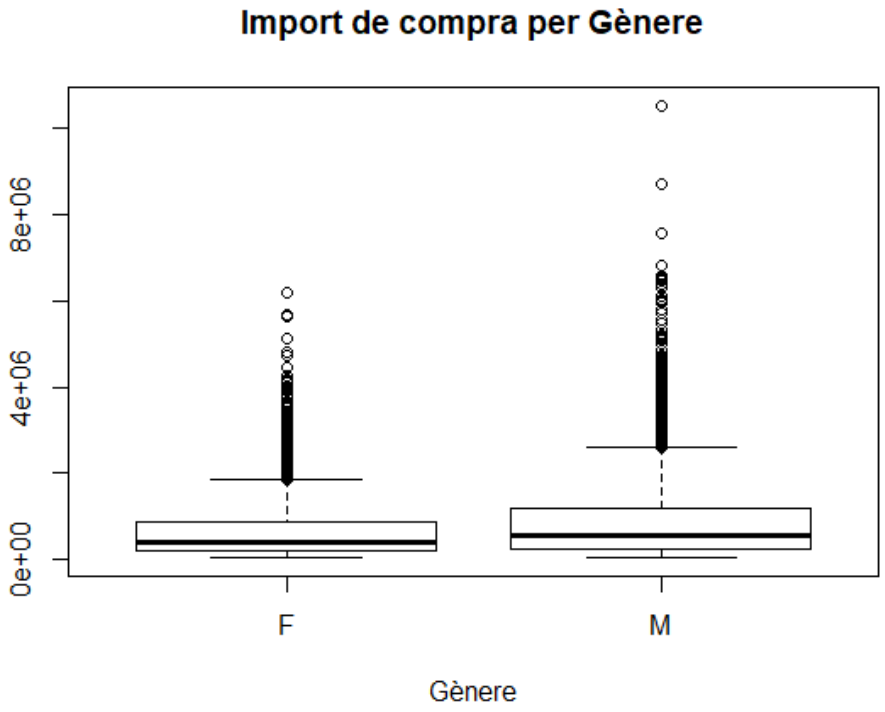
- ¿Un dels dos gèneres compra més que l'altre? ¿Quin?

Volem comparar aquí una variable numèrica amb una variable categòrica, tot i que no podem aplicar el t.test ja que la nostra distribució no compleix amb els requisits de proves paramètriques: normalitat de la distribució i homogeneïtat de les variances. Així doncs, cal realitzar proves no paramètriques, com poden ser Mann-Whitney o Kruskal-Wallis. Mirem doncs, primer de tot, quina pinta fan les dades:

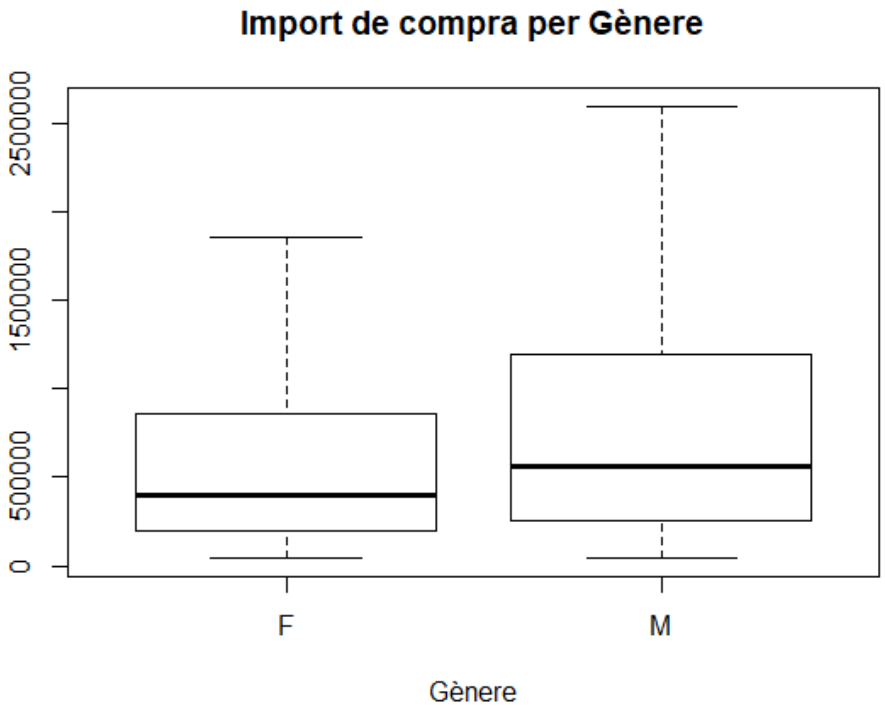
```
> library(lattice)
> histogram(~ Purchase | Gender, data=vendes_bf_agg, layout=c(1,2))
```



```
> boxplot(Purchase ~ Gender, data = vendes_bf_agg, xlab="Gènere", main="Impor
t de compra per Gènere")
```




```
> boxplot(Purchase ~ Gender, data = vendes_bf_agg, xlab="Gènere", main="Impor
t de compra per Gènere", outline = FALSE)
```



En aquest cas, i donat que volem estudiar l'import de compra en funció del gènere, ens podem limitar al U test de Mann-Whitney (una sola variable amb dues categories):

```
> wilcox.test(Purchase ~ Gender, data = vendes_bf_agg)

        Wilcoxon rank sum test with continuity correction

data:  Purchase by Gender
W = 2961600, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

Per tal de validar que els valors extrems no afecten a aquest càlcul, farem el mateix procés tot eliminant els outliers amb la llibreria *boxplot*, i veurem que obtenim els mateixos resultats:

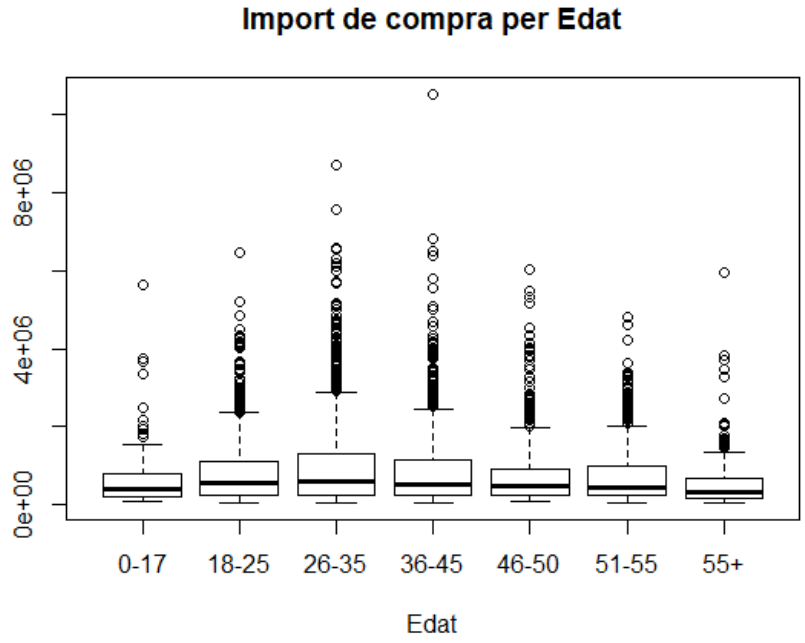
```
> vendes_bf_in <- vendes_bf_agg[!vendes_bf_agg$Purchase %in% boxplot.stats(ve
ndes_bf_agg$Purchase)$out,]
> with(vendes_bf_in, wilcox.test(Purchase ~ Gender))

        Wilcoxon rank sum test with continuity correction

data:  Purchase by Gender
W = 2625500, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

Donat que la hipòtesi nul·la que volem validar per al U test de Mann-Whitney és que les dues mostres (F i M) pertanyen a la mateixa població, i veient que el p-valor obtingut està per sota del llindar d'acceptabilitat, podem assumir que es tracta efectivament de dues poblacions diferents, i que per tant **el gènere es un factor diferencial en l'import de compra, sent el gènere masculí el que a priori més gasta.**

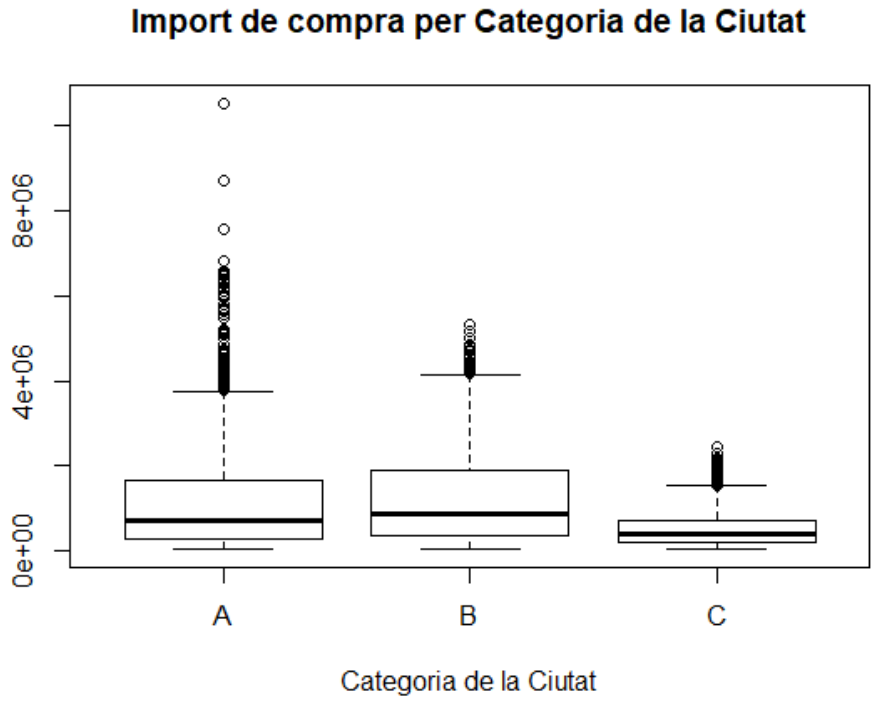
- ¿Hi ha altres factors que influeixin directament en el volum de compra?

Si visualitzem l'import de compra en funció a altres variables trobarem els següents grafs:

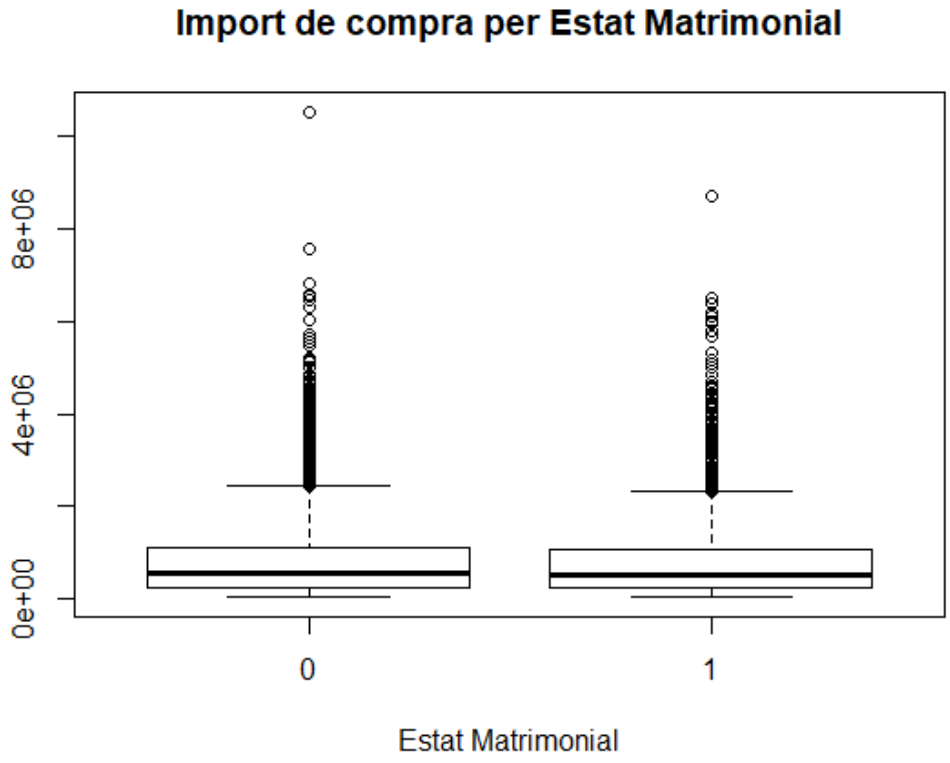
```
> boxplot(Purchase ~ Age, data = vendes_bf_agg, xlab="Edat", main="Import de compra per Edat")
```



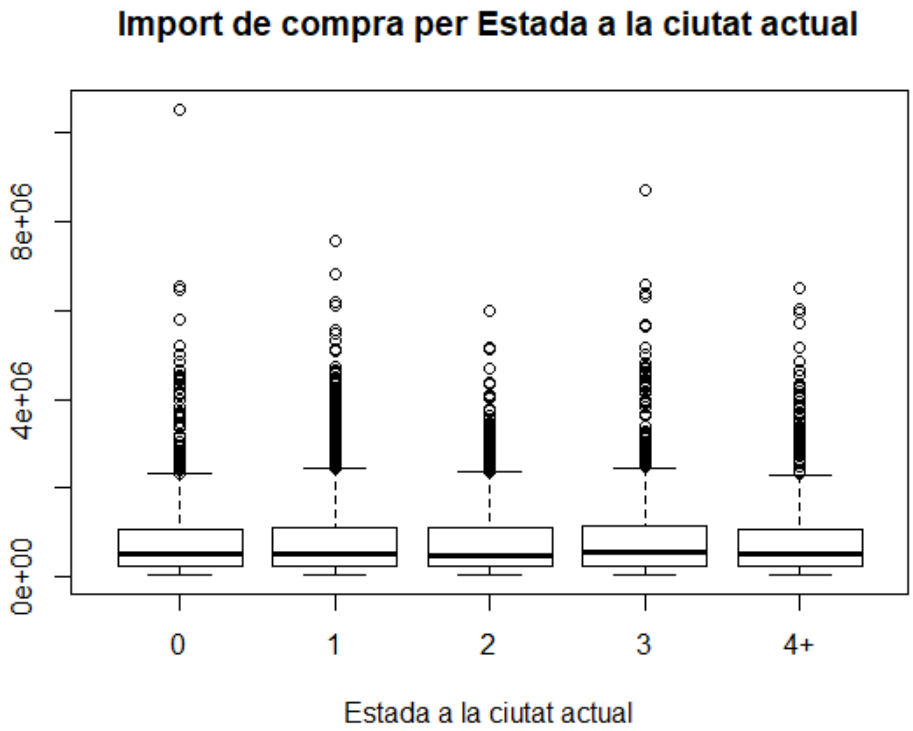

```
> boxplot(Purchase ~ City_Category, data = vendes_bf_agg, xlab="Categoria de la Ciutat", main="Import de compra per Categoria de la Ciutat")
```



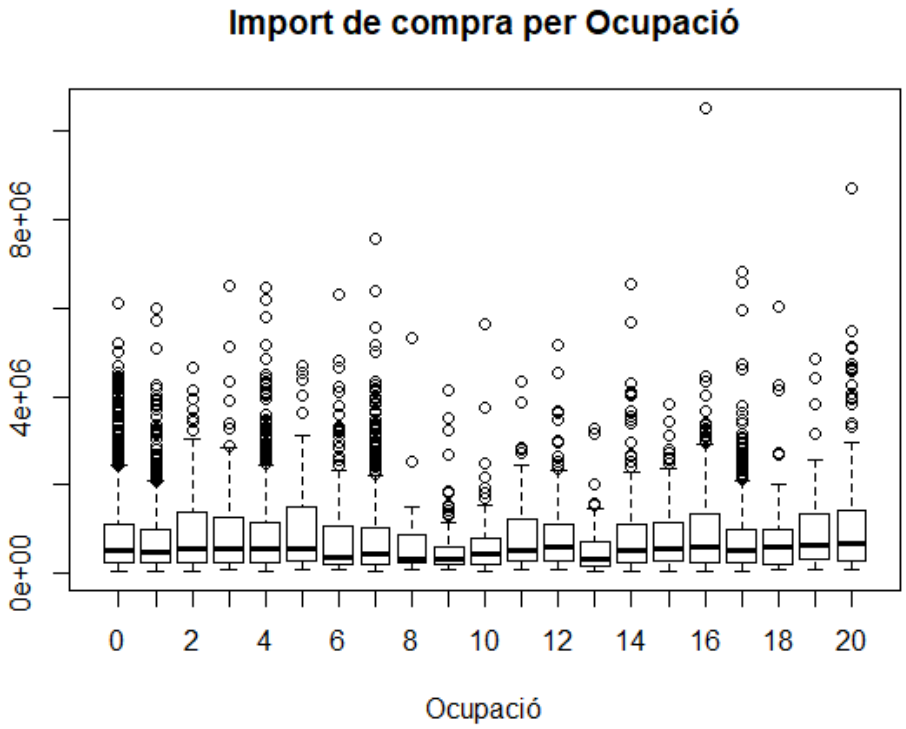
```
> boxplot(Purchase ~ Marital_Status, data = vendes_bf_agg, xlab="Estat Matrim
onial", main="Import de compra per Estat Matrimonial")
```




```
> boxplot(Purchase ~ Stay_In_Current_City_Years, data = vendes_bf_agg, xlab="
Estada a la ciutat actual", main="Import de compra per Estada a la ciutat act
ual")
```

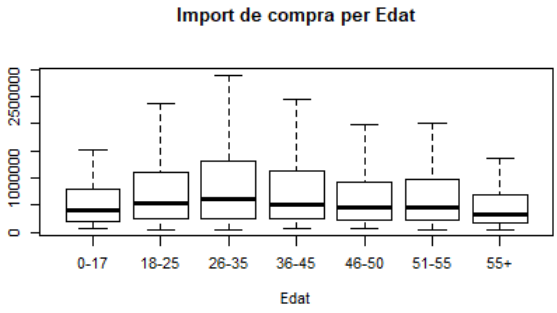


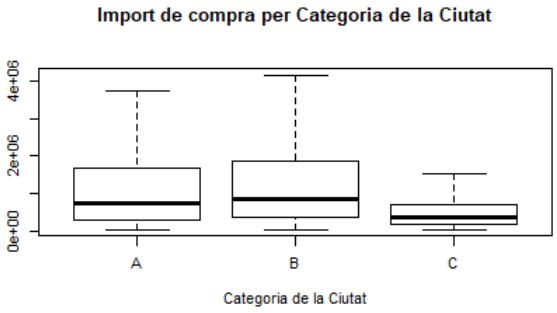
```
> boxplot(Purchase ~ Occupation, data = vendes_bf_agg, xlab="Ocupació", main=
"Import de compra per Ocupació")
```

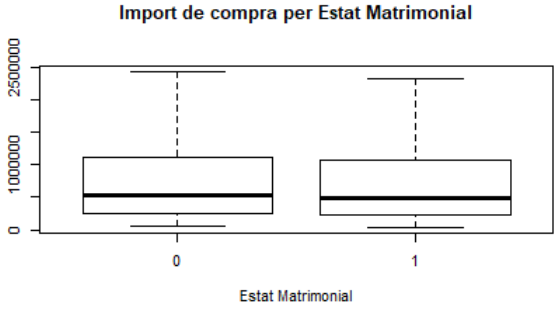


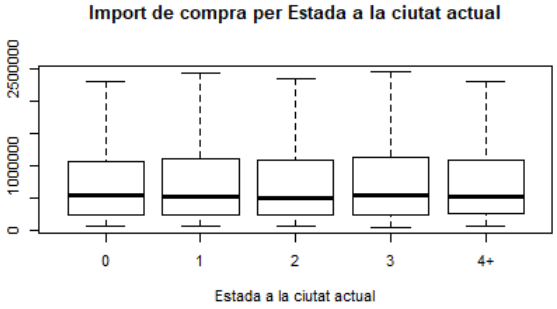
De cara a mostrar aquests resultats de manera més entenedora, podem eliminar els valors extrems de la gràfica i ajuntar-los tots en un mateix plot:

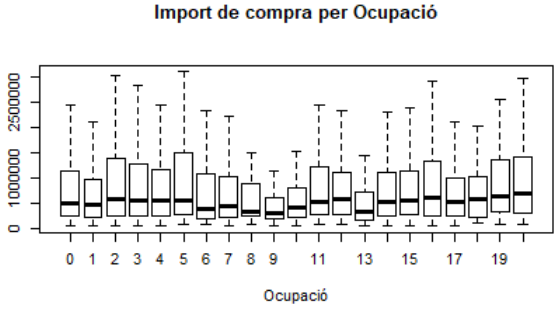
```
> attach(vendes_bf_agg)
> par(mfrow=c(3,2))
> boxplot(Purchase ~ Age, xlab="Edat", main="Import de compra per Edat", outl
ine=FALSE)
> boxplot(Purchase ~ City_Category, xlab="Categoria de la Ciutat", main="Impo
rt de compra per Categoria de la Ciutat", outline = FALSE)
> boxplot(Purchase ~ Marital_Status, xlab="Estat Matrimonial", main="Import d
e compra per Estat Matrimonial", outline = FALSE)
> boxplot(Purchase ~ Stay_In_Current_City_Years, xlab="Estada a la ciutat act
ual", main="Import de compra per Estada a la ciutat actual", outline = FALSE)
> boxplot(Purchase ~ Occupation, xlab="Ocupació", main="Import de compra per
Ocupació", outline = FALSE)
> boxplot(Purchase ~ Gender, xlab="Gènere", main="Import de compra per Gènere
", outline = FALSE)
```

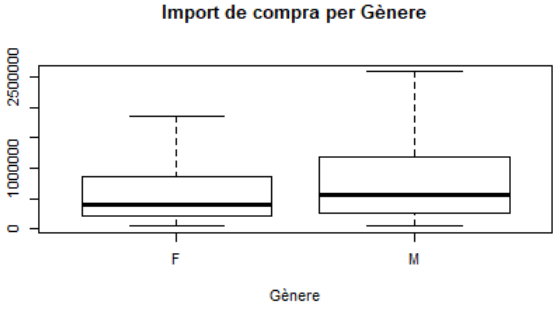












Visualment ja podem apreciar diferencies entre les variables: Per una banda, sembla que l'estat matrimonial no afecta a l'import de les compres, i de mateixa manera per a l'estada en la ciutat actual. En canvi, la categoria de ciutat, l'edat i l'ocupació semblen tenir impacte en l'import de compra, a més del gènere, que ja hem estudiat prèviament. Validarem aquestes hipòtesis amb les proves no-paramètriques necessàries, per variables categòriques de dos nivells amb Mann-Whitney, per variables categòriques de més de dos nivells amb Kruskal-Wallis. Cal tenir present que independentment de la prova realitzada, la hipòtesi nul·la contempla que les mostres pertanyen a la mateixa població, i per tant:

- Si el p-valor es inferior a 0.05, es pot afirmar que les mostres presenten diferències significatives.
- Si el p-valor es superior a 0.05, es pot afirmar que les mostres no presenten diferències significatives, i que per tant la variable independent avaluada no té incidència en la variable dependent (import de compra).

Realitzem doncs el test de Kruskal-Wallis per a la resta de variables dependents, excepte per a l'estat matrimonial, donat que es tracta de una variable binaria.

```
> kruskal.test(Purchase ~ Age, data = vendes_bf_agg)

        Kruskal-Wallis rank sum test

data:  Purchase by Age
Kruskal-Wallis chi-squared = 119.69, df = 6, p-value < 2.2e-16

> kruskal.test(Purchase ~ City_Category, data = vendes_bf_agg)

        Kruskal-Wallis rank sum test

data:  Purchase by City_Category
Kruskal-Wallis chi-squared = 657.34, df = 2, p-value < 2.2e-16

> kruskal.test(Purchase ~ Occupation, data = vendes_bf_agg)

        Kruskal-Wallis rank sum test

data:  Purchase by Occupation
Kruskal-Wallis chi-squared = 90.473, df = 20, p-value = 6.126e-11

> kruskal.test(Purchase ~ Stay_In_Current_City_Years, data = vendes_bf_agg)

        Kruskal-Wallis rank sum test

data:  Purchase by Stay_In_Current_City_Years
Kruskal-Wallis chi-squared = 1.4066, df = 4, p-value = 0.843

> wilcox.test(Purchase ~ Marital_Status, data = vendes_bf_agg)

        Wilcoxon rank sum test with continuity correction

data:  Purchase by Marital_Status
W = 4348200, p-value = 0.05967
alternative hypothesis: true location shift is not equal to 0
```

Observem, com ja havíem pogut inspeccionar visualment, que les variables *estat matrimonial* i *estada a la ciutat actual* no tenen afectació en l'import de compra del client (p-valors de 0.05967 i 0.843 respectivament), però la resta ens pot permetre predir el consum del client.

- ¿Podem predir el consum del client?

Volem analitzar si hi ha variables que permeten predir l'import de les compres de l'usuari, i per tant cal analitzar la multi-col·linealitat del model de regressió. De cara a poder avaluar la qualitat del model obtingut, separarem el set de dades en dos, per tal d'entrenar el model d'una banda (70% de les dades, és a dir 4123 registres) i avaluar-lo d'una altra (30% de les dades, és a dir 1768 registres). Aleatoritzem les mostres i definim el model de regressió lineal amb les dades d'entrenament:

```
> sample <- sample.int(n = nrow(vendes_bf_agg), size = floor(.70*nrow(vendes_
bf_agg)), replace = F)
> train <- vendes_bf_agg[sample, 2:8]
> test <- vendes_bf_agg[-sample, 2:8]
```

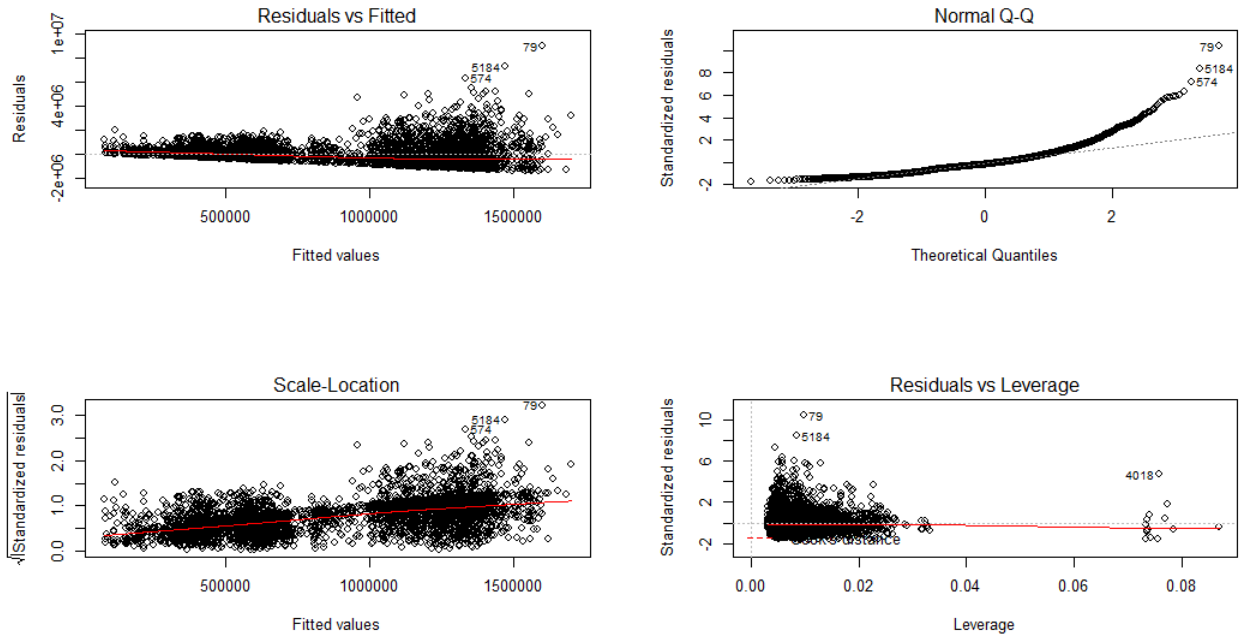
```
> library(car)
> lmfit = lm(Purchase ~ Age + Occupation + Gender + City_Category, data = tra
in)
> summary(lmfit)

Call:
lm(formula = Purchase ~ Age + Occupation + Gender + City_Category,
    data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-1461525  -493000  -162006   320435  8969221

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     1036419     118591   8.739  < 2e-16 ***
Age18-25          60890     112563   0.541   0.5886
Age26-35         162742     112581   1.446   0.1484
Age36-45         143976     114848   1.254   0.2100
Age46-50         131871     119785   1.101   0.2710
Age51-55         -24179     121695  -0.199   0.8425
Age55+           -77212     125358  -0.616   0.5380
Occupation1      -14597      61302  -0.238   0.8118
Occupation2      -66686      77099  -0.865   0.3871
Occupation3      161202      90539   1.780   0.0751 .
Occupation4        3458      60672   0.057   0.9546
Occupation5      150139     110040   1.364   0.1725
Occupation6      -60507      80893  -0.748   0.4545
Occupation7      -88164      57555  -1.532   0.1256
Occupation8       99534     235241   0.423   0.6722
Occupation9      -66025     114876  -0.575   0.5655
Occupation10    -104242     120110  -0.868   0.3855
Occupation11    -225705     102619  -2.199   0.0279 *
Occupation12    -137666      67716  -2.033   0.0421 *
Occupation13    -107105     104535  -1.025   0.3056
Occupation14     -18923      72746  -0.260   0.7948
Occupation15    -130877      95870  -1.365   0.1723
Occupation16     160389      78247   2.050   0.0404 *
Occupation17     -49506      62072  -0.798   0.4252
Occupation18     -52501     133302  -0.394   0.6937
Occupation19     243377     127168   1.914   0.0557 .
Occupation20      14716      74453   0.198   0.8433
GenderM          226777      31533   7.192 7.57e-13 ***
City_CategoryB   -58454      40943  -1.428   0.1535
City_CategoryC  -743095      37842 -19.637  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 866300 on 4093 degrees of freedom
Multiple R-squared:  0.1695,   Adjusted R-squared:  0.1636
F-statistic: 28.81 on 29 and 4093 DF,  p-value: < 2.2e-16
```

Podem observar com els valors del p-valor propi de cada interacció de categoria amb l'import de compra (darrera columna) indiquen si la variable es significativa en el càlcul. Addicionalment, volem obtenir el valor de *F-Statistic* més alt possible. Si analitzem que passa en considerar totes les variables en el model veurem que el *F-statistic* serà més baix:

```
> lmfit = lm(Purchase ~ Age + Occupation + Marital_Status + Gender + City_Cat
egory + Stay_In_Current_City_Years, data = train)
> summary(lmfit)

Call:
lm(formula = Purchase ~ Age + Occupation + Marital_Status + Gender +
    City_Category + Stay_In_Current_City_Years, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-1491340  -496634  -163037   315495  8937347

[...]

F-statistic: 24.65 on 34 and 4088 DF,  p-value: < 2.2e-16
```

Representem gràficament el model obtingut:

```
> par(mfrow=c(2,2))
> plot(lmfit)
```



Mantenim per tant les variables inicials per a la definició del model regressiu. Addicionalment, cal analitzar si les variables són col·lineals, i per tant variables predictives estan correlacionades entre elles. Aquest factor apareix si l'arrel quadrada del factor d'inflació de variança es superior a 2, tot i que en aquest cas no tenim aquest problema:

```
> vif(lmfit)
                    GVIF Df GVIF^(1/(2*Df))
Age             4.877847  6        1.141175
Occupation      5.359915 20        1.042867
Gender          1.110208  1        1.053664
City_Category   1.050949  2        1.012501

> sqrt(vif(lmfit)) > 2
                 GVIF    Df GVIF^(1/(2*Df))
Age              TRUE  TRUE           FALSE
Occupation       TRUE  TRUE           FALSE
Gender          FALSE FALSE           FALSE
City_Category   FALSE FALSE           FALSE
```

Definim doncs el nostre model de regressió lineal:

```
> predict <- predict(lmfit, test, interval="confidence", level=0.95)
> test$IC_prediction <- predict
> test$Error<-(test$Purchase-test$IC_prediction)*100/test$Purchase
> summary(test$Error)
```

```
      fit                lwr                upr
 Min.   :-2457.45   Min.   :-2258.59   Min.   :-2656.31
 1st Qu.: -192.65   1st Qu.: -134.22   1st Qu.: -252.92
 Median :  -44.14   Median :  -10.53   Median :  -73.36
 Mean   : -127.98   Mean   :  -86.91   Mean   : -169.05
 3rd Qu.:   31.62   3rd Qu.:   45.89   3rd Qu.:   15.87
 Max.   :  102.96   Max.   :  180.12   Max.   :   81.24
```

```
> mean(test$Error)
[1] -127.979
> sd(test$Error)
[1] 252.8133
```

L'error relatiu calculat representa un percentatge de l'import total de compra. Podem veure com la mitjana de l'error relatiu comès amb els valors predits es situa en un -127% del valor real de l'import, amb una desviació estàndard de 252, el que indica que el model obtingut no es de gaire qualitat. Existeix però una segona alternativa, que es realitzar un model de regressió per quantils:

```
> library(quantreg)
> quantile <- rq(Purchase ~ Age + Occupation + Gender + City_Category, data =
train, tau = 0.5)
Warning message:
In rq.fit.br(x, y, tau = tau, ...) : Solution may be nonunique
> summary(quantile)

Call: rq(formula = Purchase ~ Age + Occupation + Gender + City_Category,
    tau = 0.5, data = train)

tau: [1] 0.5
```

```
Coefficients:
                Value          Std. Error     t value        Pr(>|t|)
(Intercept)      495077.00000    65940.08196        7.50798        0.00000
Age18-25         155495.00000    62121.70656        2.50307        0.01235
Age26-35         183661.00000    55371.86960        3.31686        0.00092
Age36-45         143399.00000    57536.18082        2.49233        0.01273
Age46-50         134647.00000    62192.24312        2.16501        0.03044
Age51-55          72534.00000    54257.95110        1.33684        0.18135
Age55+            19736.00000    54982.09388        0.35895        0.71965
Occupation1       26474.00000    45249.38184        0.58507        0.55853
Occupation2       -1312.00000    65755.62729       -0.01995        0.98408
Occupation3       51282.00000   119329.06034        0.42975        0.66740
Occupation4        4724.00000    49077.03926        0.09626        0.92332
Occupation5      187980.00000   141553.77856        1.32798        0.18426
Occupation6      -29030.00000    46349.19938       -0.62633        0.53113
Occupation7      -65795.00000    45206.32949       -1.45544        0.14562
Occupation8        3390.00000   156027.30180        0.02173        0.98267
Occupation9      -52352.00000    51223.32252       -1.02203        0.30683
Occupation10       4837.00000    79320.67352        0.06098        0.95138
Occupation11     -92585.00000    73419.99936       -1.26103        0.20737
Occupation12      33491.00000    69599.95023        0.48119        0.63041
Occupation13     -18227.00000    59809.96868       -0.30475        0.76057
Occupation14      -1054.00000    64704.54820       -0.01629        0.98700
Occupation15      -1138.00000    85793.24758       -0.01326        0.98942
Occupation16     169316.00000    90537.66243        1.87012        0.06154
Occupation17     -23613.00000    50589.40257       -0.46676        0.64070
Occupation18      12942.00000   168236.89992        0.07693        0.93869
Occupation19     332462.00000   155189.71971        2.14229        0.03223
Occupation20      45224.00000    79331.40523        0.57006        0.56867
GenderM          155858.00000    21003.62878        7.42053        0.00000
City_CategoryB   111661.00000    57973.93230        1.92606        0.05417
City_CategoryC -350235.00000    39518.36619       -8.86259        0.00000
```

```
> test$QT_prediction <- predict.rq(quantile,test,interval = "confidence", lev
el=0.95)
Warning message:
In summary.rq(object, cov = TRUE, ...) : 15 non-positive fis

> test$QT_Error<-(test$Purchase-test$QT_prediction)*100/test$Purchase
> mean(test$QT_Error)
[1] -61.13645
> sd(test$QT_Error)
[1] 172.7307
```

Per a aquest model hem aconseguit millorar molt les prediccions, reduint el valor mig dels errors comesos a -61% (menys de la meitat respecte al primer model) i la desviació estàndard a 172%, és a dir tres quartes parts de la desviació estàndard del primer model). Tot i presentar clares millores a nivell de qualitat predictiva, el segon model tampoc és prou precís com per a poder-se utilitzar com a eina de predicció.

5. Resolució del problema. A partir dels resultats obtinguts, quines són les conclusions? Els resultats permeten respondre al problema?

Resumim primerament el raonament que s'ha seguit per a arribar fins aquest punt:

- Un cop les dades han estat tractades i netejades, s'ha volgut analitzar si les variables *Ocupació*, *Gènere*, *Categoria de la ciutat*, *Estada en anys a la ciutat actual*, *Edat* i *Estat matrimonial* tenien un impacte significatiu en l'import de la compra del client.
- Donada la natura de les dades, on la variable dependent (*Import de compra*) no segueix una distribució normal i no es manté la homogeneïtat de les variances, s'han realitzat proves no paramètriques per a analitzar l'efecte de les diferents variables sobre l'import de compra del client.
- S'ha pogut observar mitjançant aquestes proves estadístiques que les variables *Estat matrimonial* i *Estada en anys a la ciutat actual* no tenen efecte sobre l'import de compra, mentre que les altres variables si que son significatives.
- Amb les variables obtingudes s'ha volgut realitzar un model per tal de predir l'import de compra dels clients, tot i que cap dels models obtinguts ha estat de qualitat suficient com per a ser aplicable.

De cara a poder explicar el perquè d'aquests resultats, cal revisar les dades i tenir present que no disposem en cap cas de dades numèriques continues, sinó que totes les variables corresponen son categòriques. A la imatge següent podem observar el principal problema de tractar només dades categòriques:

	Gender	Age	Occupation	City_Category	Stay_In_Current_City_Years	Marital_Status	Purchase	IC_prediction
4202	M	46-50	1	C	1	1	191020	637375.2
4203	M	46-50	1	C	1	1	195414	637375.2
4204	M	46-50	1	C	1	1	1019083	637375.2
4206	F	51-55	1	C	1	1	314041	254548.5
4207	F	51-55	1	C	1	1	182399	254548.5
4208	F	51-55	1	C	1	1	321617	254548.5
4209	F	51-55	1	C	1	1	323018	254548.5
4211	F	51-55	1	C	1	1	641090	254548.5
4214	M	51-55	1	C	1	1	690598	481325.8
4236	F	51-55	2	C	1	1	1069855	202459.7
4237	M	51-55	2	C	1	1	991285	429236.9
4241	F	26-35	3	C	1	1	471388	617267.8
4246	F	26-35	3	C	1	1	88360	617267.8
4247	F	26-35	3	C	1	1	558092	617267.8

Si ens fixem en les files 4202, 4203 i 4204, veiem que per als mateixos valors de les variables independents s'obtenen imports de compra dispars, respectivament 191020, 195414 i 1019083, que generen imprecisions en el model. Probablement disposar de variables numèriques continues (Edat exacta, sou del client) ens hagués permès obtenir un model més precís. Tot i així, amb la informació obtinguda podem saber quines variables tenen més impacte en l'import de compra del client, podent així buscar quines categories són les que generen més ingressos i permetent enfocar les campanyes de marketing o promocions de productes a aquells segments que generin més ingressos.

Recursos

Els següents recursos són d'utilitat per la realització de la PAC:

- Megan Squire (2015). Clean Data . Packt Publishing Ltd. Capítols 1 i 2.
- Jiawei Han, Micheine Kamber, Jian Pei (2012). Data mining: concepts and techniques . Morgan Kaufmann. Capítol 3.
- Jason W. Osborne (2010). Data Cleaning Basics: Best Practices in Dealing with Extreme Scores. Newborn and Infant Nursing Reviews; 10 (1): pp. 1527-3369.

Referències

[1] https://en.wikipedia.org/wiki/Type_I_and_type_II_errors

[2] Megan Squire (2015). Clean Data . Packt Publishing Ltd. Capítols 1 i 2.

[3] Jiawei Han, Micheine Kamber, Jian Pei (2012). Data mining: concepts and techniques . Morgan Kaufmann. Capítol 3.

[4] Jason W. Osborne (2010). Data Cleaning Basics: Best Practices in Dealing with Extreme Scores. Newborn and Infant Nursing Reviews; 10 (1): pp. 1527-3369.

[5] http://rcompanion.org/handbook/I_12.html

[6] https://anestesiar.org/2015/no-todo-es-normal-manejo-de-datos-no-normales/

[7] https://stats.stackexchange.com/questions/132652/how-to-determine-which-distribution-fits-my-data-best

[8] http://www.sthda.com/english/wiki/normality-test-in-r

[9] https://flowingdata.com/2012/05/15/how-to-visualize-and-compare-distributions/

[10] http://rcompanion.org/handbook/F_04.html

[11] https://stackoverflow.com/questions/44089894/identifying-the-outliers-in-a-data-set-in-r

[12] http://www.sthda.com/english/wiki/kruskal-wallis-test-in-r

[13] https://www.polyglotdeveloper.com/r-projects/2016-09-30-Predicting-salaries-using-linear-regression/

[14] https://stackoverflow.com/questions/17200114/how-to-split-data-into-training-testing-sets-using-sample-function

[15] https://www.rdocumentation.org/packages/car/versions/3.0-2/topics/vif

[16] https://www.rdocumentation.org/packages/quantreg/versions/5.36/topics/predict.rq

[17] http://rcompanion.org/handbook/F_12.html