Pràctica 2: Neteja i validació de dades

David Martin Tinaquero

4 de gener de 2019

Índex

1. Descripció del *dataset*

El conjunt de dades objecte d'anàlisi s'ha obtingut de l'enllaç https://www.kaggle.com/mehdidag/black-friday i està constituït per 537.577 observacions (files o registres) amb 12 característiques (columnes). A continuació es llisten i descriuen tots els camps continguts a aquest conjunt de dades:

- **User_ID.** Identificador únic de cada usuari.
- **Product_ID.** Identificador únic de cada producte.
- **Gender.** Variable categòrica que indica si el comprador és home (M - *Male*) o dona (F - *Female*).
- **Age.** Rang d'edat a la que pertany l'usuari comprador. Aquests intervals són els següents: 0-17, 18-25, 26-35, 36-45, 46-50, 51-55 i 55+.
- **Occupation.** Codi numèric en l'interval 0-20 que indica la ocupació del comprador.
- **City_Category.** Categoria a la qual pertany la ciutat de residència del comprador, representada per les lletres: A, B i C.
- **Stay_In_Current_City.** Anys que fa que el comprador viu a la ciutat actual. Els anys poden ser: 0, 1, 2, 3 i 4+.
- **Marital_Status.** Codi numèric que indica si el comprador està cassat ("1") o no ("0"). La traducció de la codificació s'ha interpretat a partir de que no hi ha cap comprador a l'interval d'edat 0-17 amb el camp **Marital_Status** a "1".
- **Product_Category_1.** Codi numèric en l'interval 1-18 que indica la categoria a la qual pertany el producte.
- **Product_Category_2.** Codi numèric en l'interval entre 2-18 que indica la categoria a la qual pertany el producte.
- **Product_Category_3.** Codi numèric en l'interval entre 3-18 que indica la categoria a la qual pertany el producte.
- **Purchase.** Número que indica el preu de la compra del producte. Per la magnitud dels imports sembla que inclou els decimals.

2. Importància i objectius de les anàlisis

Aquest conjunt de dades és una mostra de les transaccions realitzades en una tenda minorista. La tenda vol conèixer millor el comportament de compra del client enfront a diferents productes. Específicament, és tracta d'un problema de regressió, al que es vol predir la variable dependent, l'import de la compra, amb ajuda de la informació continguda a la resta de variables.

També es podria realitzar una anàlisi de classificació, ja que les variables són categòriques. Alguns enfocaments podrien ser predir l'edat del consumidor o, inclús, predir la categoria dels béns comprats. El conjunt de dades també és particularment convenient per agrupar i, potser, trobar diferents grups de consumidors dintre d'ell.

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

En primer lloc, comencem amb una lectura del fitxer en format CSV. El resultat tornat per la crida a la funció **read.csv()** és un objecte **data.frame**.

6 primers registres amb les 6 primeres columnes:

```
> # Lectura de dades
> blackFriday <- read.csv("BlackFriday.csv", header = TRUE)
> # 5 primers registres
> head(blackFriday[1:6])
  User_ID Product_ID Gender   Age Occupation City_Category
1 1000001  P00069042      F  0-17         10             A
2 1000001  P00248942      F  0-17         10             A
3 1000001  P00087842      F  0-17         10             A
4 1000001  P00085442      F  0-17         10             A
5 1000002  P00285442      M   55+         16             C
6 1000003  P00193542      M 26-35         15             A
```

6 darrers registres amb les 6 primeres columnes:

```
> # 5 últims registres
> tail(blackFriday[1:6])
       User_ID Product_ID Gender   Age Occupation City_Category
537572 1004737  P00221442      M 36-45         16             C
537573 1004737  P00193542      M 36-45         16             C
537574 1004737  P00111142      M 36-45         16             C
537575 1004737  P00345942      M 36-45         16             C
537576 1004737  P00285842      M 36-45         16             C
537577 1004737  P00118242      M 36-45         16             C
```

A la següent captura es mostren els tipus de dades de les columnes:

```
> # Tipus de les variables
> sapply(blackFriday, function(x) class(x))
                    User_ID                  Product_ID
                  "integer"                    "factor"
                     Gender                         Age
                   "factor"                    "factor"
                 Occupation               City_Category
                  "integer"                    "factor"
Stay_In_Current_City_Years              Marital_Status
                   "factor"                   "integer"
         Product_Category_1         Product_Category_2
                  "integer"                   "integer"
         Product_Category_3                   Purchase
                  "integer"                   "integer"
```

A continuació, es mostren els tipus de dades amb algunes de les dades contingudes a les variables:

```
> # Tipus de variables i continguts
> glimpse(blackFriday)
Observations: 537,577
Variables: 12
$ User_ID                   <int> 1000001, 1000001, 1000001, 1000001, ...
$ Product_ID                <fct> P00069042, P00248942, P00087842, P00...
$ Gender                    <fct> F, F, F, F, M, M, M, M, M, M, M, M, ...
$ Age                       <fct> 0-17, 0-17, 0-17, 0-17, 55+, 26-35, ...
$ Occupation                <int> 10, 10, 10, 10, 16, 15, 7, 7, 7, 20,...
$ City_Category             <fct> A, A, A, A, C, A, B, B, B, A, A, A, ...
$ Stay_In_Current_City_Years <fct> 2, 2, 2, 2, 4+, 3, 2, 2, 2, 1, 1, 1,...
$ Marital_Status            <int> 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, ...
$ Product_Category_1        <int> 3, 1, 12, 12, 8, 1, 1, 1, 1, 8, 5, 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, ...
$ Purchase                  <int> 8370, 15200, 1422, 1057, 7969, 15227...
```

La majoria dels atributs presents al joc de dades es corresponen a categories dels compradors i dels productes comprats, per tant, és important no descartar cap d'aquests camps. No obstant, podem prescindir dels dos primers camps (User_ID i Product_ID), donat que no pertanyen a cap categoria dels usuaris ni dels productes, sinó que són els seus identificadors clau, i el que es pretén amb l'anàlisi és generalitzar pels altres camps categòrics i no concretar per usuaris ni per productes concrets. Amb la següent comanda s'eliminen els camps esmentats.

```
# Selecció de les dades d'interés
blackFriday <- blackFriday[, -(1:2)]
```

Es verifica que els camps eliminats no es troben entre les 10 variables restants al *dataset*:

```
> # Tipus de variables i continguts
> glimpse(blackFriday)
Observations: 537,577
Variables: 10
$ Gender                    <fct> F, F, F, F, M, M, M, M, M, M, M, M, ...
$ Age                       <fct> 0-17, 0-17, 0-17, 0-17, 55+, 26-35, ...
$ Occupation                <int> 10, 10, 10, 10, 16, 15, 7, 7, 7, 20,...
$ City_Category             <fct> A, A, A, A, C, A, B, B, B, A, A, A, ...
$ Stay_In_Current_City_Years <fct> 2, 2, 2, 2, 4+, 3, 2, 2, 2, 1, 1, 1,...
$ Marital_Status            <int> 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, ...
$ Product_Category_1        <int> 3, 1, 12, 12, 8, 1, 1, 1, 1, 8, 5, 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, ...
$ Purchase                  <int> 8370, 15200, 1422, 1057, 7969, 15227...
```

4. Neteja de les dades

4.1. Zeros i elements buits

Les columnes **Product_Category_2** y **Product_Category_3** són les úniques que contenen elements buits, concretament, 166.986 i 373.299 registres respectivament.

```
# Nombres de valors desconeguts per camp
sapply(blackFriday, function(x) sum(is.na(x)))
```

```
> # Nombres de valors desconeguts per camp
> sapply(blackFriday, function(x) sum(is.na(x)))
                      Gender                          Age
                           0                            0
                  Occupation                City_Category
                           0                            0
Stay_In_Current_City_Years                Marital_Status
                           0                            0
            Product_Category_1            Product_Category_2
                           0                       166986
            Product_Category_3                     Purchase
                      373299                            0
```

Degut a que són camps que indiquen la categoria a la qual pertanyen els productes, és recomanable convertir els camps a tipus **factor** en comptes de mantenir-los com a tipus numèric (*integer*). Per convertir-los es pot fer mitjançant la funció **factor()** de R però, si prèviament hem reemplaçat els valors buits per zeros, al convertir el camp al tipus factor els zeros passen a ser NA de nou, cosa que no interessa. Per això, es pren la decisió d'afegir el caràcter "C" com a prefix a cadascun dels codis numèrics que representen les diferents categories (els valors nuls queden amb el codi "C0"), després es converteixen els camps a tipus **factor**.

```
# Substitució dels valors nuls del camp Product_Category_2 per zeros ("0")
blackFriday$Product_Category_2[is.na(blackFriday$Product_Category_2)] <- "0"
# Substitució dels valors nuls del camp Product_Category_3 per zeros ("0")
blackFriday$Product_Category_3[is.na(blackFriday$Product_Category_3)] <- "0"
# Adició del prefix "C" als codis dels camps de les categories dels productes
blackFriday$Product_Category_1 <- paste("C",blackFriday$Product_Category_1,sep="")
blackFriday$Product_Category_2 <- paste("C",blackFriday$Product_Category_2,sep="")
blackFriday$Product_Category_3 <- paste("C",blackFriday$Product_Category_3,sep="")
# Conversió dels camps de categories dels productes a tipus factor
blackFriday$Product_Category_1 <- factor(blackFriday$Product_Category_1)
blackFriday$Product_Category_2 <- factor(blackFriday$Product_Category_2)
blackFriday$Product_Category_3 <- factor(blackFriday$Product_Category_3)
```

Es verifica que ja no queden valors nuls als camps **Product_Category_2** y **Product_Category_3**.

```
> # Nombres de valors desconeguts per camp
> sapply(blackFriday, function(x) sum(is.na(x)))
                    User_ID                  Product_ID
                          0                           0
                     Gender                         Age
                          0                           0
                 Occupation               City_Category
                          0                           0
Stay_In_Current_City_Years              Marital_Status
                          0                           0
           Product_Category_1          Product_Category_2
                          0                           0
           Product_Category_3                    Purchase
                          0                           0
```

Es segueixen procediments semblants pels camps **Occupation** i **Marital_Status**, al primer s'afegeix el prefix "O" als codis que indiquen les ocupacions, i al segon es reemplacen els "0" per "NO" i els "1" per "SI". Posteriorment es converteixen els dos camps a tipus factor.

```
# Adició del prefix "O" als codis del camp Occupation
blackFriday$Occupation <- paste("O",blackFriday$Occupation,sep="")
# Conversió del camp Occupation a tipus factor
blackFriday$Occupation <- factor(blackFriday$Occupation)
# Substitució dels valors 0 per "NO" y 1 per "SI" al camp Marital_Status
blackFriday$Marital_Status[blackFriday$Marital_Status == 0] <- "NO"
blackFriday$Marital_Status[blackFriday$Marital_Status == 1] <- "SI"
# Conversió del camp Marital_Status a tipus factor
blackFriday$Marital_Status <- factor(blackFriday$Marital_Status)
```

Amb l'objectiu de que les dades siguin el més realistes possibles es divideix el camp **Purchase** per 100, ja que sembla que els imports del *dataset* contenen els decimals.

```
# Conversió de l'import de la compra a preu real
blackFriday$Purchase <- blackFriday$Purchase/100
```

També es crea la nova columna **Categories** que inclou la concatenació dels valors de les tres columnes de categories dels productes: **Product_Category1**, **Product_Category2** i **Product_Category3**.

```
# Creació de la variable Categories amb el conjunt de les categories
categories <- paste(blackFriday$Product_Category_1,
                    blackFriday$Product_Category_2,
                    blackFriday$Product_Category_3)
blackFriday$Categories <- categories
blackFriday$Categories <- factor(blackFriday$Categories)
```

Es verifica que tots els camps, excepte **Purchase**, son del tipus factor:

```
> # Tipus de les variables i continguts
> glimpse(blackFriday)
Observations: 534,912
Variables: 11
$ Gender                    <fct> F, F, F, F, M, M, M, M, M, M, M, M...
$ Age                       <fct> 0-17, 0-17, 0-17, 0-17, 55+, 26-35...
$ Occupation                <fct> O10, O10, O10, O10, O16, O15, O7, ...
$ City_Category             <fct> A, A, A, A, C, A, B, B, B, A, A, A...
$ Stay_In_Current_City_Years <fct> 2, 2, 2, 2, 4+, 3, 2, 2, 2, 1, 1, ...
$ Marital_Status            <fct> NO, NO, NO, NO, NO, NO, SI, SI, SI...
$ Product_Category_1        <fct> C3, C1, C12, C12, C8, C1, C1, C1, ...
$ Product_Category_2        <fct> C0, C6, C0, C14, C0, C2, C8, C15, ...
$ Product_Category_3        <fct> C0, C14, C0, C0, C0, C0, C17, C0, ...
$ Purchase                  <dbl> 83.70, 152.00, 14.22, 10.57, 79.69...
$ Categories                <fct> C3 C0 C0, C1 C6 C14, C12 C0 C0, C1...
```

4.2. Identificació i tractament de valors extrems

Els valors extrems o *outliers* són els que semblen no ser congruents si es comparen amb la resta de dades. Per identificar-los, utilitzem la funció **boxplot.stats()** de R. Es pot veure que al camp **Purchase** s'han detectat 2.665 valors extrems amb quantitats entre el 213,84 i el 239,61, que són el ~0,50% respecte del total.

```
> outliers <- boxplot.stats(blackFriday$Purchase)$out
> # Numero de registres amb valors extrems
> paste("Número de registres que contenen valors extrems al dataset: ",length(outliers))
[1] "Número de registres que contenen valors extrems al dataset:  2665"
> # Valor més baix
> paste("Valor extrem més baix: ",min(outliers))
[1] "Valor extrem més baix:  213.84"
> # Vaor mes alt
> paste("Valor extrem més alt: ",max(outliers))
[1] "Valor extrem més alt:  239.61"
> # Percentatge de registres que representen els valors extrems
> outliersVsTotal <- (length(outliers)/nrow(blackFriday))*100
> paste("Percentatge de registres que representen els valors extrems al dataset: ",round(outliersVsTotal,3),
 "%")
[1] "Percentatge de registres que representen els valors extrems al dataset:  0.496 %"
```

Degut a que representen un percentatge molt baix respecte al total, podem prescindir d'aquests registres, per tant, els eliminem mitjançant la comanda següent:

```
# Eliminació dels outliers
blackFriday <- filter(blackFriday,blackFriday$Purchase < min(outliers))
```

5. Anàlisi de dades

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

A continuació, es seleccionen els grups dintre del nostre conjunt de dades que poden resultar interesants per analitzar i/o comparar, encara que, com es veurà als apartats posteriors amb les proves estadístiques, no tots els utilitzarem.

```
# Agrupació per gènere
blackFriday.gender_male <- blackFriday[blackFriday$Gender == "M",]
blackFriday.gender_female <- blackFriday[blackFriday$Gender == "F",]

# Agrupació per rangs d'edad
blackFriday.age_0_17 <- blackFriday[blackFriday$Age == "0-17",]
blackFriday.age_18_25 <- blackFriday[blackFriday$Age == "18-25",]
blackFriday.age_26_35 <- blackFriday[blackFriday$Age == "26-35",]
blackFriday.age_36_45 <- blackFriday[blackFriday$Age == "36-45",]
blackFriday.age_46_50 <- blackFriday[blackFriday$Age == "46-50",]
blackFriday.age_51_55 <- blackFriday[blackFriday$Age == "51-55",]
blackFriday.age_55_ <- blackFriday[blackFriday$Age == "55+",]
```

```r
# Agrupació per categoria de ciutat
blackFriday.city_cat_A <- blackFriday[blackFriday$City_Category == "A",]
blackFriday.city_cat_B <- blackFriday[blackFriday$City_Category == "B",]
blackFriday.City_cat_C <- blackFriday[blackFriday$City_Category == "C",]

# Agrupació per anys vivint a la ciutat actual
blackFriday.years_0 <- blackFriday[blackFriday$Stay_In_Current_City_Years == "0",]
blackFriday.years_1 <- blackFriday[blackFriday$Stay_In_Current_City_Years == "1",]
blackFriday.years_2 <- blackFriday[blackFriday$Stay_In_Current_City_Years == "2",]
blackFriday.years_3 <- blackFriday[blackFriday$Stay_In_Current_City_Years == "3",]
blackFriday.years_4_ <- blackFriday[blackFriday$Stay_In_Current_City_Years == "4+",]

# Agrupació per estat civil
blackFriday.marital_SI <- blackFriday[blackFriday$Marital_Status == "SI",]
blackFriday.marital_NO <- blackFriday[blackFriday$Marital_Status == "NO",]

# Agrupació per ocupació
blackFriday.ocu_O0 <- blackFriday[blackFriday$Occupation == "O0",]
blackFriday.ocu_O1 <- blackFriday[blackFriday$Occupation == "O1",]
blackFriday.ocu_O2 <- blackFriday[blackFriday$Occupation == "O2",]
blackFriday.ocu_O3 <- blackFriday[blackFriday$Occupation == "O3",]
blackFriday.ocu_O4 <- blackFriday[blackFriday$Occupation == "O4",]
blackFriday.ocu_O5 <- blackFriday[blackFriday$Occupation == "O5",]
blackFriday.ocu_O6 <- blackFriday[blackFriday$Occupation == "O6",]
blackFriday.ocu_O7 <- blackFriday[blackFriday$Occupation == "O7",]
blackFriday.ocu_O8 <- blackFriday[blackFriday$Occupation == "O8",]
blackFriday.ocu_O9 <- blackFriday[blackFriday$Occupation == "O9",]
blackFriday.ocu_O10 <- blackFriday[blackFriday$Occupation == "O10",]
blackFriday.ocu_O11 <- blackFriday[blackFriday$Occupation == "O11",]
blackFriday.ocu_O12 <- blackFriday[blackFriday$Occupation == "O12",]
blackFriday.ocu_O13 <- blackFriday[blackFriday$Occupation == "O13",]
blackFriday.ocu_O14 <- blackFriday[blackFriday$Occupation == "O14",]
blackFriday.ocu_O15 <- blackFriday[blackFriday$Occupation == "O15",]
blackFriday.ocu_O16 <- blackFriday[blackFriday$Occupation == "O16",]
blackFriday.ocu_O17 <- blackFriday[blackFriday$Occupation == "O17",]
blackFriday.ocu_O18 <- blackFriday[blackFriday$Occupation == "O18",]
blackFriday.ocu_O19 <- blackFriday[blackFriday$Occupation == "O19",]
blackFriday.ocu_O20 <- blackFriday[blackFriday$Occupation == "O20",]

# Agrupació per categories
blackFriday.prod_cat_C0 <- blackFriday[blackFriday$Product_Category_1 == "C0" |
                                       blackFriday$Product_Category_2 == "C0" |
                                       blackFriday$Product_Category_3 == "C0",]
blackFriday.prod_cat_C1 <- blackFriday[blackFriday$Product_Category_1 == "C1" |
                                       blackFriday$Product_Category_2 == "C1" |
                                       blackFriday$Product_Category_3 == "C1",]
blackFriday.prod_cat_C2 <- blackFriday[blackFriday$Product_Category_1 == "C2" |
                                       blackFriday$Product_Category_2 == "C2" |
                                       blackFriday$Product_Category_3 == "C2",]
blackFriday.prod_cat_C3 <- blackFriday[blackFriday$Product_Category_1 == "C3" |
                                       blackFriday$Product_Category_2 == "C3" |
                                       blackFriday$Product_Category_3 == "C3",]
blackFriday.prod_cat_C4 <- blackFriday[blackFriday$Product_Category_1 == "C4" |
                                       blackFriday$Product_Category_2 == "C4" |
                                       blackFriday$Product_Category_3 == "C4",]
blackFriday.prod_cat_C5 <- blackFriday[blackFriday$Product_Category_1 == "C5" |
                                       blackFriday$Product_Category_2 == "C5" |
                                       blackFriday$Product_Category_3 == "C5",]
```

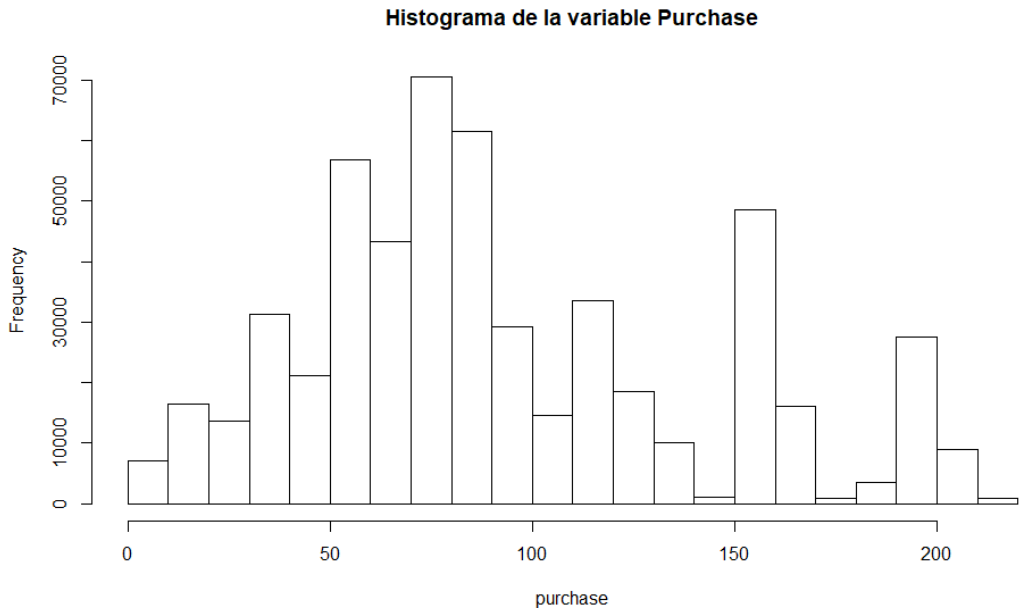
```
blackFriday.prod_cat_C6 <- blackFriday[blackFriday$Product_Category_1 == "C6" |
                                       blackFriday$Product_Category_2 == "C6" |
                                       blackFriday$Product_Category_3 == "C6",]
blackFriday.prod_cat_C7 <- blackFriday[blackFriday$Product_Category_1 == "C7" |
                                       blackFriday$Product_Category_2 == "C7" |
                                       blackFriday$Product_Category_3 == "C7",]
blackFriday.prod_cat_C8 <- blackFriday[blackFriday$Product_Category_1 == "C8" |
                                       blackFriday$Product_Category_2 == "C8" |
                                       blackFriday$Product_Category_3 == "C8",]
blackFriday.prod_cat_C9 <- blackFriday[blackFriday$Product_Category_1 == "C9" |
                                       blackFriday$Product_Category_2 == "C9" |
                                       blackFriday$Product_Category_3 == "C9",]
blackFriday.prod_cat_C10 <- blackFriday[blackFriday$Product_Category_1 == "C10" |
                                        blackFriday$Product_Category_2 == "C10" |
                                        blackFriday$Product_Category_3 == "C10",]
blackFriday.prod_cat_C11 <- blackFriday[blackFriday$Product_Category_1 == "C11" |
                                       blackFriday$Product_Category_2 == "C11" |
                                       blackFriday$Product_Category_3 == "C11",]
blackFriday.prod_cat_C12 <- blackFriday[blackFriday$Product_Category_1 == "C12" |
                                       blackFriday$Product_Category_2 == "C12" |
                                       blackFriday$Product_Category_3 == "C12",]
blackFriday.prod_cat_C13 <- blackFriday[blackFriday$Product_Category_1 == "C13" |
                                       blackFriday$Product_Category_2 == "C13" |
                                       blackFriday$Product_Category_3 == "C13",]
blackFriday.prod_cat_C14 <- blackFriday[blackFriday$Product_Category_1 == "C14" |
                                       blackFriday$Product_Category_2 == "C14" |
                                       blackFriday$Product_Category_3 == "C14",]
blackFriday.prod_cat_C15 <- blackFriday[blackFriday$Product_Category_1 == "C15" |
                                       blackFriday$Product_Category_2 == "C15" |
                                       blackFriday$Product_Category_3 == "C15",]
blackFriday.prod_cat_C16 <- blackFriday[blackFriday$Product_Category_1 == "C16" |
                                       blackFriday$Product_Category_2 == "C16" |
                                       blackFriday$Product_Category_3 == "C16",]
blackFriday.prod_cat_C17 <- blackFriday[blackFriday$Product_Category_1 == "C17" |
                                       blackFriday$Product_Category_2 == "C17" |
                                       blackFriday$Product_Category_3 == "C17",]
blackFriday.prod_cat_C18 <- blackFriday[blackFriday$Product_Category_1 == "C18" |
                                       blackFriday$Product_Category_2 == "C18" |
                                       blackFriday$Product_Category_3 == "C18",]
```

5.2. Comprovació de la normalitat i homogeneïtat de la variància

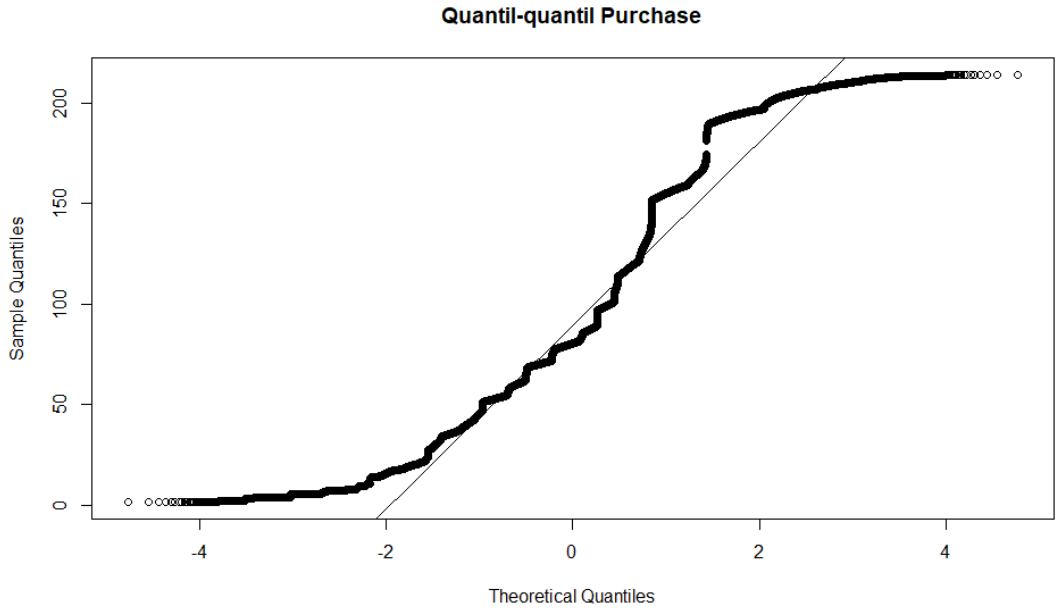
Abans de realitzar els contrasts d'hipòtesi, necessitem conèixer si la variable **Purchase** té un tipus de distribució normal o no. Conèixer el tipus de distribució de les dades és fonamental, ja que si és normal sabem que la mitjana és representativa i, en cas contrari, ho serà la mediana. Amb la mitjana i la desviació estàndard es descriu una distribució normal.

El contrast d'hipòtesi consisteix en comparar dos elements. La hipòtesi nul·la és aquella que afirma que entre els dos valors comparats no hi ha diferència. Qualsevol contrast d'hipòtesi es basa en que s'ha d'acceptar o rebutjar aquesta hipòtesi. Si l'acceptem, sabem que no hi ha diferència entre els valors, si la rebutgem sabem que si que hi ha diferència.

En primer lloc, fem una ullada mitjançant un histograma. Encara que no es el mètode més fiable per comprovar la normalitat, es pot apreciar a la gràfica següent que la variable Purchase no té una distribució de Gauss, per tant, sembla que no té distribució normal.



La segona comprovació que hem fet, ha sigut mitjançant una gràfica quantil-quantil, que és més informativa. Com es pot apreciar a la gràfica, els punts que representen els valors de Purchase no es distribueixen al llarg de la línia que representa la distribució normal:



L'últim mètode que s'ha emprat per comprovar la normalitat ha sigut el mètode numèric, que evita errors deguts a les possibles interpretacions de les gràfiques. Concretament s'ha emprat el mètode d'asimetria (*skewness*) d'Agostino i el mètode d'Anderson-Darling. Ambdues proves han retornat un p-valor inferior a 0.05, per tant, es rebutja la hipòtesi nul·la i es confirma que no existeix una distribució normal a la variable Purchase.

```
> skewness(sample(purchase))
[1] 0.5886613
> agostino.test(sample(purchase,46340))

        D'Agostino skewness test

data:  sample(purchase, 46340)
skew = 0.58961, z = 48.13500, p-value < 2.2e-16
alternative hypothesis: data have a skewness

> # Mètode de Anderson-Darling
> ad.test(purchase)

        Anderson-Darling normality test

data:  purchase
A = 9376.6, p-value < 2.2e-16
```

Ara que sabem que la variable Purchase no té una distribució normal, toca comprovar la homogeneïtat de la variància. Per avaluar la distribució de la variància es considera com a hipòtesi nul·la que la variància és igual entre els grups, i com hipòtesi alternativa que no ho és. En aquest cas, degut a que no existeix una distribució normal, haurem d'utilitzar un test que utilitzi la mediana en comptes de la mitjana.

Hem utilitzat el test de Fligner-Killen, que es caracteritza perquè compara les variàncies basant-se en la mediana. Hem realitzat els tests entre el camp Purchase i els grups definits en l'apartat 5.1.

```
> # Purchase amb gènere
> fligner.test(x = list(blackFriday.gender_male$Purchase,
+                       blackFriday.gender_female$Purchase))

        Fligner-Killeen test of homogeneity of variances

data:  list(blackFriday.gender_male$Purchase, blackFriday.gender_female$Purchase)
Fligner-Killeen:med chi-squared = 1569.8, df = 1, p-value < 2.2e-16

> # Purchase amb rangs d'edad
> fligner.test(x = list(blackFriday.age_0_17$Purchase,
+                       blackFriday.age_18_25$Purchase,
+                       blackFriday.age_26_35$Purchase,
+                       blackFriday.age_36_45$Purchase,
+                       blackFriday.age_36_45$Purchase,
+                       blackFriday.age_46_50$Purchase,
+                       blackFriday.age_51_55$Purchase,
+                       blackFriday.age_55_$Purchase))

        Fligner-Killeen test of homogeneity of variances

data:  list(blackFriday.age_0_17$Purchase, blackFriday.age_18_25$Purchase,     blackFriday.
age_26_35$Purchase, blackFriday.age_36_45$Purchase,     blackFriday.age_36_45$Purchase, bla
ckFriday.age_46_50$Purchase,     blackFriday.age_51_55$Purchase, blackFriday.age_55_$Purcha
se)
Fligner-Killeen:med chi-squared = 136.98, df = 7, p-value < 2.2e-16

> # Purchase amb categoria de ciutat
> fligner.test(x = list(blackFriday.city_cat_A$Purchase,
+                       blackFriday.city_cat_B$Purchase,
+                       blackFriday.City_cat_C$Purchase))

        Fligner-Killeen test of homogeneity of variances

data:  list(blackFriday.city_cat_A$Purchase, blackFriday.city_cat_B$Purchase,     blackFrid
ay.City_cat_C$Purchase)
Fligner-Killeen:med chi-squared = 222.78, df = 2, p-value < 2.2e-16
```

```
> # Purchase amb anys vivint a la ciutat actual
> fligner.test(x = list(blackFriday.years_0$Purchase,
+                       blackFriday.years_1$Purchase,
+                       blackFriday.years_2$Purchase,
+                       blackFriday.years_3$Purchase,
+                       blackFriday.years_4_$Purchase))

        Fligner-Killeen test of homogeneity of variances

data:  list(blackFriday.years_0$Purchase, blackFriday.years_1$Purchase,     blackFriday.yea
rs_2$Purchase, blackFriday.years_3$Purchase,     blackFriday.years_4_$Purchase)
Fligner-Killeen:med chi-squared = 18.294, df = 4, p-value = 0.001081

> # Purchase amb estat civil
> fligner.test(x = list(blackFriday.marital_NO$Purchase,
+                       blackFriday.marital_SI$Purchase))

        Fligner-Killeen test of homogeneity of variances

data:  list(blackFriday.marital_NO$Purchase, blackFriday.marital_SI$Purchase)
Fligner-Killeen:med chi-squared = 12.886, df = 1, p-value = 0.0003311

> #Purchase amb ocpacions
> fligner.test(x = list(blackFriday.ocu_00$Purchase,
+                       blackFriday.ocu_01$Purchase,
+                       blackFriday.ocu_02$Purchase,
+                       blackFriday.ocu_03$Purchase,
+                       blackFriday.ocu_04$Purchase,
+                       blackFriday.ocu_05$Purchase,
+                       blackFriday.ocu_06$Purchase,
+                       blackFriday.ocu_07$Purchase,
+                       blackFriday.ocu_08$Purchase,
+                       blackFriday.ocu_09$Purchase,
+                       blackFriday.ocu_010$Purchase,
+                       blackFriday.ocu_011$Purchase,
+                       blackFriday.ocu_012$Purchase,
+                       blackFriday.ocu_013$Purchase,
+                       blackFriday.ocu_014$Purchase,
+                       blackFriday.ocu_015$Purchase,
+                       blackFriday.ocu_016$Purchase,
+                       blackFriday.ocu_017$Purchase,
+                       blackFriday.ocu_018$Purchase,
+                       blackFriday.ocu_019$Purchase,
+                       blackFriday.ocu_020$Purchase))

        Fligner-Killeen test of homogeneity of variances

data:  list(blackFriday.ocu_00$Purchase, blackFriday.ocu_01$Purchase,     blackFriday.ocu_0
2$Purchase, blackFriday.ocu_03$Purchase,     blackFriday.ocu_04$Purchase, blackFriday.ocu_0
5$Purchase,     blackFriday.ocu_06$Purchase, blackFriday.ocu_07$Purchase,     blackFriday.o
cu_08$Purchase, blackFriday.ocu_09$Purchase,     blackFriday.ocu_010$Purchase, blackFriday.
ocu_011$Purchase,     blackFriday.ocu_012$Purchase, blackFriday.ocu_013$Purchase,     black
Friday.ocu_014$Purchase, blackFriday.ocu_015$Purchase,     blackFriday.ocu_016$Purchase, bl
ackFriday.ocu_017$Purchase,     blackFriday.ocu_018$Purchase, blackFriday.ocu_019$Purchase,
     blackFriday.ocu_020$Purchase)
Fligner-Killeen:med chi-squared = 681.68, df = 20, p-value < 2.2e-16
```

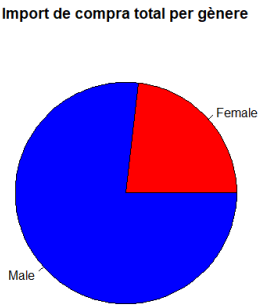
```
> # Purchase amb categories de productes
> fligner.test(x = list(blackFriday.prod_cat_C0$Purchase,
+                       blackFriday.prod_cat_C1$Purchase,
+                       blackFriday.prod_cat_C2$Purchase,
+                       blackFriday.prod_cat_C3$Purchase,
+                       blackFriday.prod_cat_C4$Purchase,
+                       blackFriday.prod_cat_C5$Purchase,
+                       blackFriday.prod_cat_C6$Purchase,
+                       blackFriday.prod_cat_C7$Purchase,
+                       blackFriday.prod_cat_C8$Purchase,
+                       blackFriday.prod_cat_C9$Purchase,
+                       blackFriday.prod_cat_C10$Purchase,
+                       blackFriday.prod_cat_C11$Purchase,
+                       blackFriday.prod_cat_C12$Purchase,
+                       blackFriday.prod_cat_C13$Purchase,
+                       blackFriday.prod_cat_C14$Purchase,
+                       blackFriday.prod_cat_C15$Purchase,
+                       blackFriday.prod_cat_C16$Purchase,
+                       blackFriday.prod_cat_C17$Purchase,
+                       blackFriday.prod_cat_C18$Purchase))

        Fligner-Killeen test of homogeneity of variances

data:  list(blackFriday.prod_cat_C0$Purchase, blackFriday.prod_cat_C1$Purchase,     blackFr
iday.prod_cat_C2$Purchase, blackFriday.prod_cat_C3$Purchase,     blackFriday.prod_cat_C4$Pu
rchase, blackFriday.prod_cat_C5$Purchase,     blackFriday.prod_cat_C6$Purchase, blackFriday
.prod_cat_C7$Purchase,     blackFriday.prod_cat_C8$Purchase, blackFriday.prod_cat_C9$Purcha
se,     blackFriday.prod_cat_C10$Purchase, blackFriday.prod_cat_C11$Purchase,     blackFrid
ay.prod_cat_C12$Purchase, blackFriday.prod_cat_C13$Purchase,     blackFriday.prod_cat_C14$P
urchase, blackFriday.prod_cat_C15$Purchase,     blackFriday.prod_cat_C16$Purchase, blackFri
day.prod_cat_C17$Purchase,     blackFriday.prod_cat_C18$Purchase)
Fligner-Killeen:med chi-squared = 74685, df = 18, p-value < 2.2e-16
```

Com que a tots els tests s'ha obtingut un p-valor inferior a 0.05, es rebutja la hipòtesi nul·la, per tant, les variàncies de totes les mostres no són homogènies.

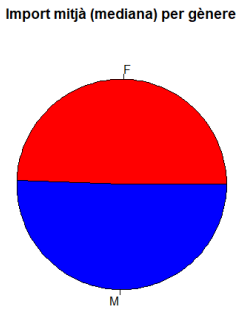
5.3. Aplicació de proves estadístiques per comparar els grups de dades

Degut a que totes les variables, excepte **Purchase**, són categòriques s'analitzarà si existeixen relacions entre les categories relatives als compradors i als productes que facin variar l'import de la compra (*purchase*).

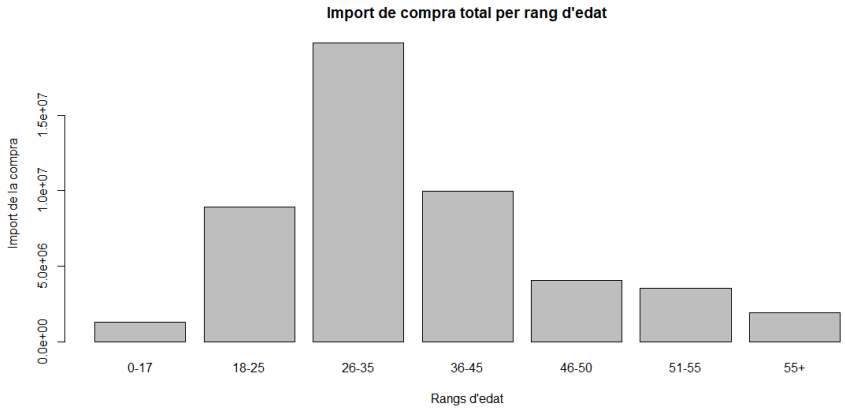
Com es pot observar a la gràfica següent, més del 75% de l'import de les compres totals han estat realitzades per homes:



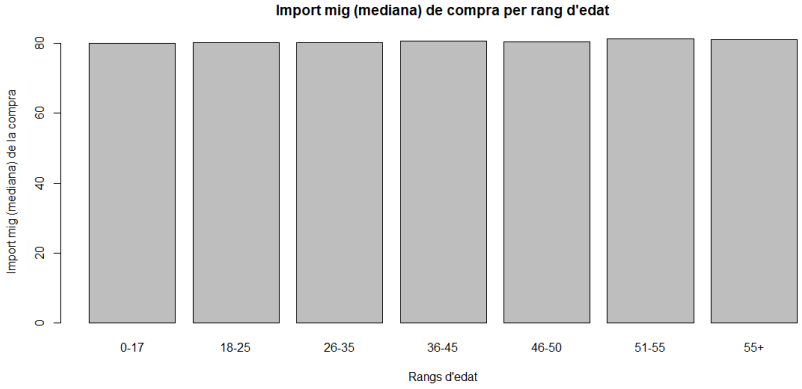
No obstant, l'import mig dels productes comprats és quasi igual entre gèneres, encara que els homes han comprat productes amb preu sensiblement superior:



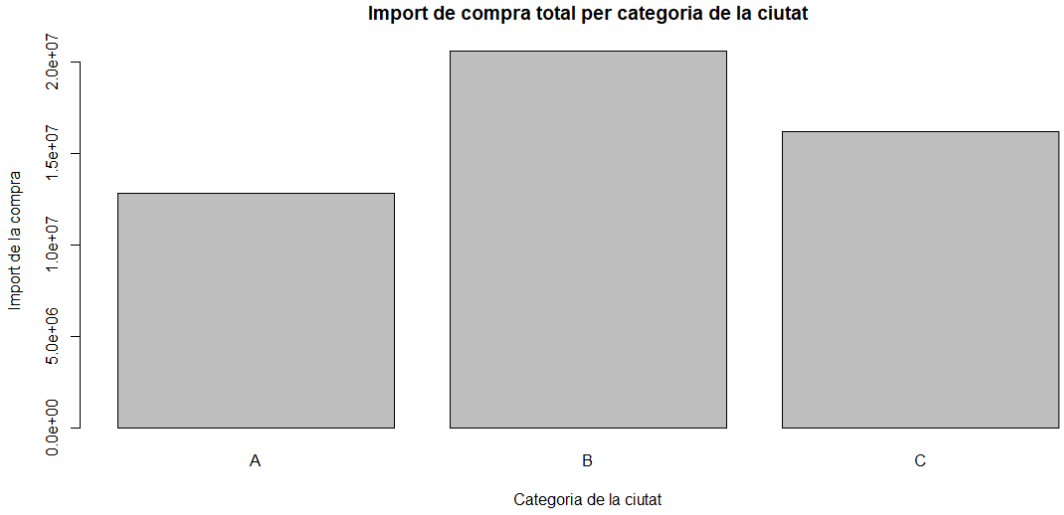
Amb els rangs d'edat passa una cosa similar, els que més import de compra han acumulat són els d'edat compresa entre 26 i 35 anys, i els que sumen menys import de compra són els que estan entre 0 i 17 anys:



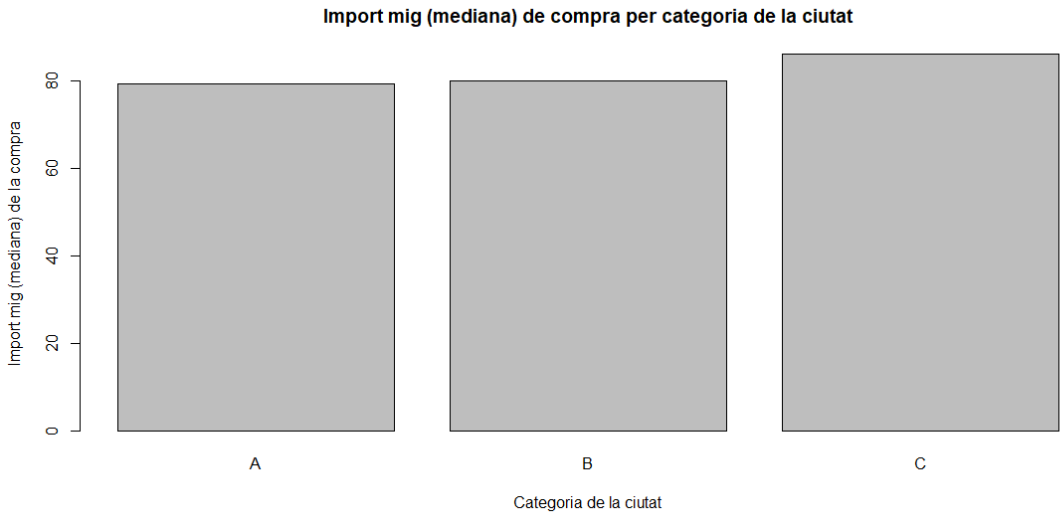
Però l'import mig dels productes comprats és quasi idèntic a tots els rangs d'edat, encara que el rang d'edat amb l'import mig més alt és el de entre 51 i 55 anys:



A continuació, es procedeix a realitzar la mateixa anàlisi amb la categoria de la ciutat de residència del comprador. L'import total de les compres realitzades més alt l'han realitzat els usuaris que viuen a ciutats de tipus B.



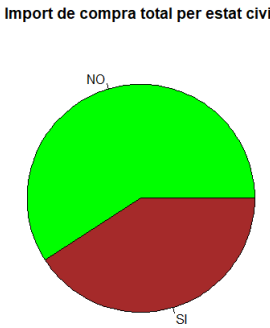
Mentre que l'import mig dels productes més elevat el tenen les compres realitzades per persones de ciutats de categoria C.



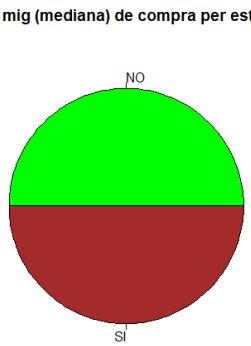
Quan s'analitza l'import de les compres totals agrupades per anys de permanència a la ciutat actual, l'import més elevat de les compres és el de les persones que porten un any vivint a la ciutat actual.

En canvi, quan es realitza l'anàlisi per l'import mig del producte, el resultat és molt semblant, però els que compren productes sensiblement més cars són els que porten 2 anys a la ciutat actual.

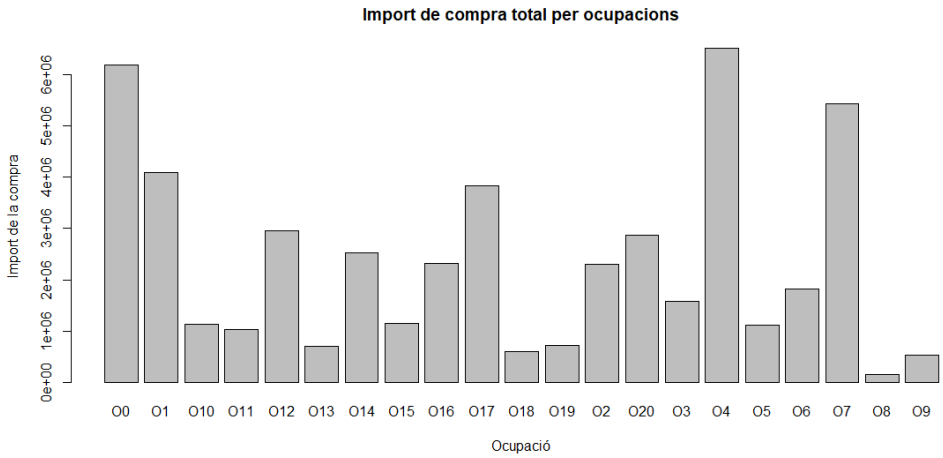
Continuant amb l'anàlisi per l'estat civil, podem observar com la suma de l'import més gran del total de les compres l'han realitzat persones que no estan casades.



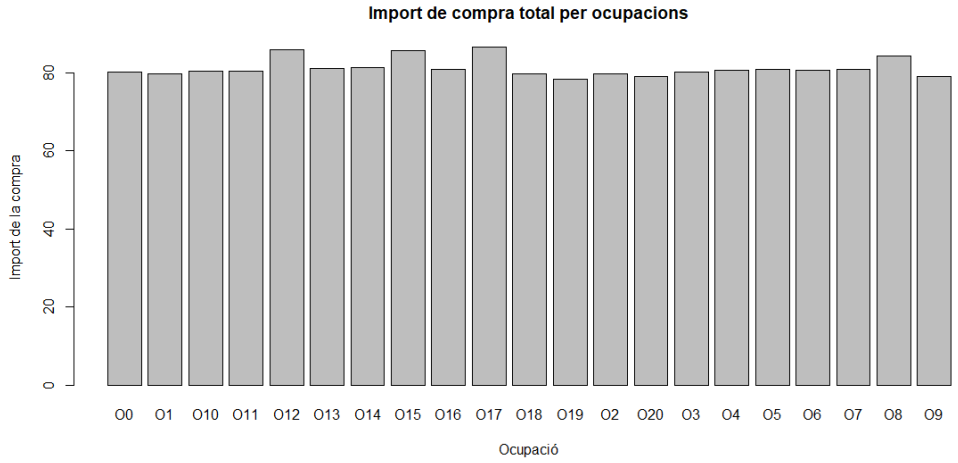
Però l'import mig del producte comprat és quasi idèntic entre cassats i la resta:



Les persones amb l'ocupació O4 són les que sumen un import de compra total més gran. Els segueixen els que l'ocupació és desconeguda (O0) i els de la O7. L'import més baix el tenen els de l'ocupació O8.



Quan ho analitzem per import mig del producte comprat, les persones amb les ocupacions O17, O15, O12 i O8 són les que tenen un import de compra més alt.



Per últim, s'han realitzat les mateixes anàlisis amb les categories dels productes comprats. A les següents captures es mostren les 10 combinacions de categories que més i menys import de compra sumen:

```
> # Top 10 categories que més import de compra sumen
> sumByProdCatMax
      Category       x
210   C8 C0 C0 4494372
155   C5 C0 C0 3831265
16   C1 C16 C0 1981476
24   C1 C2 C15 1919904
182   C5 C8 C0 1490897
92   C16 C0 C0 1431680
166  C5 C14 C0 1417265
13  C1 C15 C16 1284346
221  C8 C14 C0 1127433
30    C1 C2 C5 1098983

> # Top 10 categories que menys import de compra sumen
> sumByProdCatMin
     Category       x
156 C5 C10 C16   72.18
234   C9 C0 C0  184.56
61  C10 C11 C0  192.06
152   C4 C9 C0  291.49
225 C8 C14 C18 1174.48
167 C5 C14 C15 1461.20
97  C2 C12 C14 1563.12
81  C13 C15 C0 1651.86
159 C5 C11 C15 1700.77
153  C4 C9 C12 1906.92
```

A les captures següents es mostren les 10 combinacions de productes que tenen més alt i més baix l'import mig per producte:

```
> # Top 10 categories amb import de compra més alt
> medianByProdCatMax
       Category       x
91   C15 C17 C0 207.935
61   C10 C11 C0 192.060
62   C10 C13 C0 187.745
64  C10 C14 C16 187.200
67   C10 C16 C0 186.380
65   C10 C15 C0 186.190
60    C10 C0 C0 185.740
63  C10 C13 C16 184.875
234    C9 C0 C0 184.560
66  C10 C15 C16 184.370

> # Top 10 categories amb import de compra més baix
> medianByProdCatMin
       Category      x
81   C13 C15 C0  7.410
82  C13 C15 C16  7.425
80  C13 C14 C16  7.500
79    C13 C0 C0  7.580
83   C13 C16 C0  7.580
77  C12 C14 C17 13.780
75    C12 C0 C0 13.960
76   C12 C14 C0 14.070
78   C12 C17 C0 14.100
146   C4 C5 C15 15.010
```

5.4. Regressió lineal múltiple per predir l'import de la compra

Un cop vistes quines són les persones amb les categories que més compres fan i les que compren els productes amb preu mig més elevat, estem en disposició de crear un model de regressió lineal que permeti predir l'import de compra d'una persona en funció de les categories a les qual pertany: sexe, edat, estat civil, categoria i anys de permanència a la ciutat de residència, ocupació; i les categories dels productes comprats.

Al model que crearem només utilitzarem regressors qualitatius, degut a que totes les variables del *dataset* pertanyen a aquest tipus. Amb l'objectiu de trobar el model més eficient, crearem varis models de regressió utilitzant les variables corresponents als compradors i les que indiquen les categories dels productes, correlacionades amb el preu de la compra, i escollirem aquell que presenti major coeficient de determinació (R^2).

```
# ALGORITME DE REGRESSIÓ PREDICITU
# Regresors qualitatius dels compradors
gender = blackFriday$Gender
age = blackFriday$Age
cityCat = blackFriday$City_Category
cityYears = blackFriday$Stay_In_Current_City_Years
maritalStatus = blackFriday$Marital_Status
occupation = blackFriday$Occupation
prodCat = blackFriday$Categories

# Variable a predir
purchase = blackFriday$Purchase

# Models
model1 <- lm(purchase ~ gender, data = blackFriday)
model2 <- lm(purchase ~ age, data = blackFriday)
model3 <- lm(purchase ~ cityCat, data = blackFriday)
model4 <- lm(purchase ~ cityYears, data = blackFriday)
model5 <- lm(purchase ~ maritalStatus, data = blackFriday)
model6 <- lm(purchase ~ occupation, data = blackFriday)
model7 <- lm(purchase ~ prodCat, data = blackFriday)
model8 <- lm(purchase ~ prodCat + cityCat, data = blackFriday)
model9 <- lm(purchase ~ prodCat + cityCat + occupation, data = blackFriday)
model10 <- lm(purchase ~ prodCat + cityCat + occupation +
                gender, data = blackFriday)
model11 <- lm(purchase ~ prodCat + cityCat + occupation +
                gender + age, data = blackFriday)
model11 <- lm(purchase ~ prodCat + cityCat + occupation +
                gender + age + cityYears, data = blackFriday)
model12 <- lm(purchase ~ prodCat + cityCat + occupation + gender +
                age + cityYears + maritalStatus, data = blackFriday)


# Taula amb els coeficients de determinació de cada model
taula.coeficients <- matrix(c(1, summary(model1)$r.squared,
                              2, summary(model2)$r.squared,
                              3, summary(model3)$r.squared,
                              4, summary(model4)$r.squared,
                              5, summary(model5)$r.squared,
                              6, summary(model6)$r.squared,
                              7, summary(model7)$r.squared,
                              8, summary(model8)$r.squared,
                              9, summary(model9)$r.squared,
                              10, summary(model10)$r.squared,
                              11, summary(model11)$r.squared,
                              12, summary(model12)$r.squared),
                            ncol = 2, byrow = TRUE)
colnames(taula.coeficients) <- c("Model", "R^2")
taula.coeficients
```

Els 7 primers models només utilitzen un dels 7 regressors cadascun. Del model 7 al 12 s'han anat combinant regressors, afegint un regressor addicional en cada model, començant pels que han presentat major coeficient de determinació. A continuació es llisten els regressors en ordre de major a menor coeficient de determinació (R^2):

- prodCat: combinació de categories de productes.
- cityCat: categoria de ciutat.
- occupation: ocupació.

- **gender:** gènere.
- **age:** rang d'edat.
- **cityYears:** anys vivint a la ciutat actual.
- **maritalSatus:** estat civil.

El model més eficient ha estat el 12, que és el que utilitza les 7 variables juntes:

```
> taula.coeficients
      Model          R^2
 [1,]     1 3.666552e-03
 [2,]     2 3.764299e-04
 [3,]     3 5.158210e-03
 [4,]     4 8.759040e-05
 [5,]     5 7.746095e-07
 [6,]     6 3.972381e-03
 [7,]     7 6.415147e-01
 [8,]     8 6.435588e-01
 [9,]     9 6.441810e-01
[10,]    10 6.441937e-01
[11,]    11 6.445313e-01
[12,]    12 6.445570e-01
```

Hem realitzat 2 prediccions amb el model 12, una amb el perfil d'usuari que compra més i els productes amb les categories més venudes, i una altra amb el perfil d'usuari que menys compra i els productes amb les categories que menys es venen:

```
> # Dades per predir l'import de compra en productes
> # del perfil de comprador que més compra i productes més venuts
> newdata <- data.frame(
+   gender = "M",
+   age = "26-35",
+   cityCat = "B",
+   cityYears = "1",
+   maritalStatus = "NO",
+   occupation = "O4",
+   prodCat = "C8 C0 C0"
+ )
> # Predicció
> predict(model12, newdata)
       1
74.37719

> # Dades per predir l'import de compra en productes
> # del perfil de comprador que menys compra i productes menys venuts
> newdata <- data.frame(
+   gender = "F",
+   age = "0-17",
+   cityCat = "A",
+   cityYears = "0",
+   maritalStatus = "SI",
+   occupation = "O8",
+   prodCat = "C5 C10 C16"
+ )
> predict(model12, newdata)
       1
33.26164
```

5.5. Classificació de categories de productes amb arbre de decisió

S'ha creat un arbre de decisió per veure si és possible predir els tipus d'usuaris que compren determinats productes a partir només de la seva categoria. S'utilitza una mostra aleatòria del 70% dels registres totals per entrenar l'algoritme, i el 30% restant per provar el percentatge d'encert.

```
# Classificació de categories amb arbre de decisió
set.seed(231678)
mida.total <- nrow(blackFriday)
mida.entrenament <- round(mida.total*0.7)
dades.index <- sample(1:mida.total, size = mida.entrenament)
dades.entrenament <- blackFriday[dades.index, ]
dades.test <- blackFriday[-dades.index, ]
```

S'han creat 2 models, un que inclou la variable **Purchase** i l'altre que no. Igual que ha passat amb els models de la regressió de l'apartat anterior, només s'ha aconseguit un percentatge d'encert acceptable quan s'ha utilitzat la variable **Purchase**, degut a la correlació que té amb les categories dels productes.

```
# Model amb totes les variables
model1 <- C5.0(Product_Category_1 ~ Gender + Age + Occupation + City_Category +
               Stay_In_Current_City_Years + Marital_Status + Purchase,
               data = dades.entrenament)
# Model sense la variable Purchase
model2 <- C5.0(Product_Category_1 ~ Gender + Age + Occupation + City_Category +
                Stay_In_Current_City_Years + Marital_Status,
              data = dades.entrenament)
```

Com es pot veure, el model 1 utilitza el 100% la variable **Purchase**, mentre que la resta de variables no arriben ni al 15%.

```
Attribute usage:

100.00% Purchase
 13.62% Gender
 12.90% Age
 12.03% Occupation
  6.30% City_Category
  4.48% Stay_In_Current_City_Years
  4.00% Marital_Status
```

Al model 2, les variables que més s'utilitzen són el gènere, rang d'edat i ocupació. La variable menys utilitzada és la d'estat civil.

```
Attribute usage:

100.00% Gender
100.00% Age
100.00% Occupation
 89.01% Stay_In_Current_City_Years
 84.83% City_Category
 48.93% Marital_Status
```

Es realitzen prediccions amb els dos models:

```
# Prediccions
predict1 <- predict(model1, newdata = dades.test)
taula1 <- table(predict1, dades.test$Product_Category_1)
taula1

predict2 <- predict(model2, newdata = dades.test)
taula2 <- table(predict2, dades.test$Product_Category_1)
taula2
```

El model 1 té un percentatge d'encert molt superior al del model 2, gràcies a la correlació entre les variables Purchase i Product_Category_1.

```
> paste("% Precisió model 1: ", round(100 * sum(diag(taula1)) / sum(taula1),3))
[1] "% Precisió model 1:  86.995"

> paste("% Precisió model 2: ", round(100 * sum(diag(taula2)) / sum(taula2),3))
[1] "% Precisió model 2:  33.469"
```

6. Conclusions

Als apartats anteriors s'han realitzat diferents proves estadístiques sobre un conjunt de dades que es correspon amb les transaccions de les compres d'una tenda minorista, amb diferents variables categòriques que fan referència a l'usuari (comprador) i les variables que indiquen les categories a les quals pertanyen els productes de les compres.

L'anàlisi de la normalitat ha retornat que la variable dependent (*Purchase*) no té una distribució normal. Els contrasts d'hipòtesis també han mostrat que no existeix cap camp que tingui una relació forta amb el preu de la compra, i que el que més influència té són els que indiquen les categories dels productes comprats.

Degut a la poca correlació que existeix entre les variables categòriques i la variable dependent, els models de regressió lineals que s'han provat per realitzar prediccions, han resultat bastant dolents. Només tenen un percentatge d'encert mig acceptable (~64%) si s'utilitza com a regressor la variable que conté les combinacions de les categories dels productes.

D'altra banda, també s'ha intentat trobar el perfil dels usuaris a partir de les categories a les que pertanyen els productes comprats. Tal i com ha passat amb els models de regressió, els resultats obtinguts no sigut gaire prometedors, ja que només s'ha aconseguit un percentatge d'encert acceptable quan s'ha utilitzat la variable *Purchase*.

7. Recursos

- Megan Squire (2015). Clean Data. Packt Publishing Ltd.
- Jiawei Han, Micheine Kamber, Jian Pei (2012). Data mining: concepts and techniques. Morgan Kaufmann.
- Jason W. Osborne (2010). Data Cleaning Basics: Best Practices in Dealing with Extreme Scores. Newborn and Infant Nursing Reviews; 10 (1): pp. 1527-3369.
- Peter Dalgaard (2008). Introductory statistics with R. Springer Science & Business Media.
- Wes McKinney (2012). Python for Data Analysis. O'Reilley Media, Inc.