

Tipologia i cicle de vida de les dades - Pràctica 2

Autor: Isabel Barrera Benavent i Maria Font Sánchez

Desembre 2021

Descripció del dataset.

Perquè és important i quina pregunta/problema pretén respondre?

El dataset que hem decidit fer servir per aquesta pràctica és el següent:

<https://www.kaggle.com/imakash3011/customer-personality-analysis>

És un conjunt de dades obtingut de la web Kaggle i construït per 29 atributs, un dels quals és l'identificador únic del registre, 25 atributs numèrics enters i 3 atributs de tipus text.

Es tracta d'una enquesta feta a 2.240 clients d'una empresa per a poder valorar la seva personalitat a l'hora de comprar els seus productes i amb la finalitat de comprendre i analitzar els gustos i necessitats que tenen segons el tipus de clients que siguin.

D'aquesta manera l'empresa pot prendre decisions respecte als productes que ven o fer campanyes de màrqueting més específiques segons el tipus de clients i producte.

Els atributs que conté el dataset són els següents:

- **ID:** Identificador únic del client.
- **Year_Birth:** Any de naixement del client.
- **Education:** Nivell d'educació del client.
- **Marital_Status:** Estat marital del client.
- **Income:** Ingress familiar del client.
- **Kidhome:** Número de fills amb qui viu el client.
- **Teenhome:** Número d'adolescents amb qui viu el client.
- **Dt_Customer:** Data en que s'inscriu com client de l'empresa.
- **Recency:** Número de dies que han passat des de l'última compra.
- **MntWines:** Quantitat gastada en vi durant els últims dos anys.
- **MntFruits:** Quantitat gastada en vi durant els últims dos anys.
- **MntMeatProducts:** Quantitat gastada en carn durant els últims dos anys.
- **MntFishProducts:** Quantitat gastada en peix durant els últims dos anys.
- **MntSweetProducts:** Quantitat gastada en dolços durant els últims dos anys.
- **MntGoldProds:** Quantitat gastada en or durant els últims dos anys.
- **NumDealsPurchases:** Número de compres realitzades amb descomptes.
- **NumWebPurchases:** Número de compres realitzades des de la web de l'empresa.
- **NumCatalogPurchases:** Número de compres realitzades des del catàleg de l'empresa.
- **NumStorePurchases:** Número de compres realitzades directament en una tenda de l'empresa.

- **NumWebVisitsMonth:** Número de visites realitzades durant l'últim mes a la web de l'empresa.
- **AcceptedCmp3:** el client va acceptar l'oferta en la 3a campanya (1), o no (0).
- **AcceptedCmp4:** el client va acceptar l'oferta en la 4a campanya (1), o no (0).
- **AcceptedCmp5:** el client va acceptar l'oferta en la 5a campanya (1), o no (0).
- **AcceptedCmp1:** el client va acceptar l'oferta en la 1a campanya (1), o no (0).
- **AcceptedCmp2:** el client va acceptar l'oferta en la 2a campanya (1), o no (0).
- **Complain:** si hi ha hagut alguna queixa del client en els últims dos anys (1), o no (0).
- **Z_CostContact:** Cost Contacte. Té valor 3 per defecte.
- **Z_Revenue:** Ingressos del client: Té valor 11 per defecte.
- **Response:** 1 si el client va acceptar l'oferta en la última campanya (1), o no (0).

A partir del conjunt de dades d'aquest dataset, es pretén estudiar com influeix el perfil del client amb la quantitat gastada en les compres de productes de carn. Per això, s'aplicaran mètodes de regressió per a veure la relació entre les diferents variables, i mètodes de classificació per a crear el perfil del client. Aquest estudi és important per a les empreses de ventes de productes (tan online com físiques) per a poder identificar el tipus de client que compra de carn i crear així campanyes de màrqueting específiques.

Integració i selecció de les dades d'interès a analitzar

Per tal de poder realitzar la selecció de les dades d'interès a analitzar, el primer pas és fer la lectura del fitxer CSV anomenat 'marketing_campaign.csv':

```
data_clients<-read.csv("../data/marketing_campaign.csv", header=T, sep="\t")
```

D'aquestes 29 variables, ens interessa quedar-nos amb les següents:

- ID: identificador únic

Perfil del client:

- Year_Birth, Education, Marital_status, Income, Kidhome, Teenhome.

Gasto per productes alimentaris:

- MntMeatProducts, NumWebPurchases, NumCatalogPurchases, NumStorePurchases.

La resta de variables, degut a que no es faran servir, s'eliminen del conjunt de dades:

```
data_clients <- data_clients[, c('ID', 'Year_Birth', 'Education', 'Marital_Status', 'Income', 'Kidhome', 'Teenhome', 'MntMeatProducts', 'NumWebPurchases', 'NumCatalogPurchases', 'NumStorePurchases')]
```

Neteja de dades

En les dades del dataset, es pot observar que les variables 'Education' i 'Marital Status' són les úniques expressades com a text. La resta de variables són numèriques.

Primer, es validarà que aquestes variables de text es puguin categoritzar:

```
table(data_clients$Education)

##
##   2n Cycle   Basic Graduation   Master   PhD
##       203       54       1127       370       486

table(data_clients$Marital_Status)

##
##   Absurd   Alone Divorced   Married   Single Together   Widow   YOLO
##       2       3       232       864       480       580       77       2
```

Totes dues variables es poden categoritzar fent ús de la funció 'as.factor':

```
data_clients$Education <- as.factor(data_clients$Education)
data_clients$Marital_Status <- as.factor(data_clients$Marital_Status)
```

Zeros o elements buits

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

Les variables numèriques utilitzen el valor 0 per indicar que no es té informació. I les de tipus text ho indiquen mitjançant els elements buits. El primer pas per a fer aquesta comprovació, es validar si existeixen valors buits en les variables seleccionades:

```
colSums(is.na(data_clients))

##           ID           Year_Birth           Education           Marital_Stat
us
##           0           0           0
0
##           Income           Kidhome           Teenhome           MntMeatProduc
ts
##           24           0           0
0
##   NumWebPurchases NumCatalogPurchases NumStorePurchases
##           0           0           0
```

En aquest cas, la variable 'Income' en té 24. Per tal de gestionar aquests valors, hi ha diverses tècniques que es poden aplicar:

- Eliminar aquests registres.
- Aplicar la mitjana dels valors més propers.

Degut a que es tracta d'1% dels registres totals, es preferible eliminar-los ja que l'impacte i la perduda de dades és mínima:

```
data_clients <- na.omit(data_clients)
```

Identificació i tractament de valors extrems

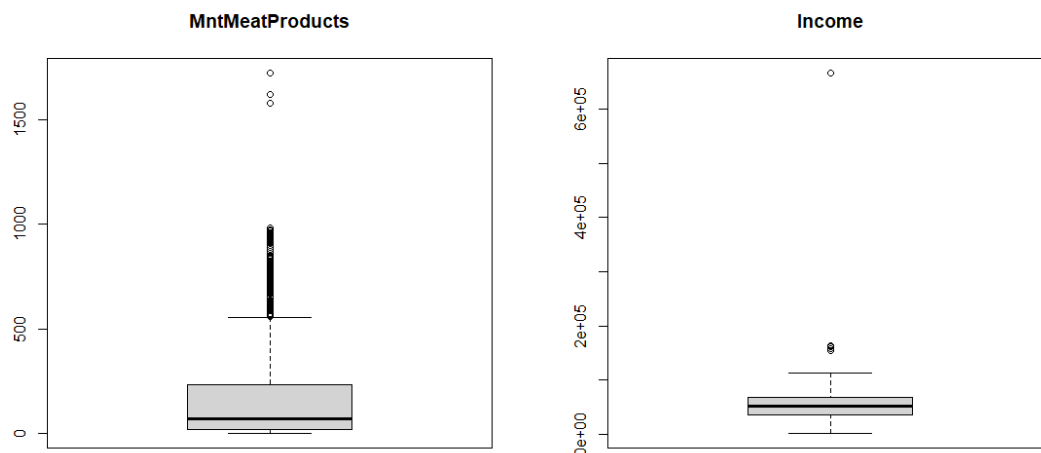
El següent pas és identificar els valors extrems, que són les dades que es troben allunyades de la distribució normal.

Utilitzarem la funció `boxplots.stats()` de R per anar revisant variable a variable si trobem registres amb valors extrems que es puguin considerar incorrectes o que ens pugui implicar una desviació important.

Es visualitzen la gràfica `boxPlot` d'algunes de les variables, i posteriorment es fa l'anàlisi de cadascuna d'elles:

```
par(mfrow=c(1,2))

extrems_mntMeat <- boxplot(data_clients$MntMeatProducts, main = "MntMeatProducts")
extrems_income <- boxplot(data_clients$Income, main = "Income")
```



VARIABLE YEAR_BIRTH

```
boxplot.stats(data_clients$Year_Birth)$out
```

```
## [1] 1900 1893 1899
```

```
summary(data_clients$Year_Birth)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      1893   1959   1970   1969   1977   1996
```

En aquest cas considerem que aquests 3 valors són incorrectes, ja que implicaria que actualment aquests clients tenen més de 130 anys.

Decidim canviar-los per el valor mitjà.

```
median_year_birth <- median(data_clients$Year_Birth)
```

```
data_clients$Year_Birth[(data_clients$Year_Birth==1893)] <- median_year_birth
data_clients$Year_Birth[(data_clients$Year_Birth==1900)] <- median_year_birth
data_clients$Year_Birth[(data_clients$Year_Birth==1899)] <- median_year_birth
```

VARIABLES EDUCATION, MARITAL STATUS, KIDHOME, TEENHOME I NUMSTOREPURCHASES:

```
boxplot.stats(data_clients$Education)$out
```

```

## factor(0)
## Levels: 2n Cycle Basic Graduation Master PhD

boxplot.stats(data_clients$Marital_Status)$out

## factor(0)
## Levels: Absurd Alone Divorced Married Single Together Widow YOLO

boxplot.stats(data_clients$Kidhome)$out

## integer(0)

boxplot.stats(data_clients$Teenhome)$out

## integer(0)

boxplot.stats(data_clients$NumStorePurchases)$out

## integer(0)

## No sembla que hi hagin valors extrems.
## Revisem els seus valor amb la funció table
table(data_clients$Education)

##
##      2n Cycle      Basic Graduation      Master      PhD
##      200         54      1116         365      481

table(data_clients$Marital_Status)

##
##      Absurd      Alone Divorced      Married      Single Together      Widow      YOLO
##      2         3      232      857      471      573      76      2

table(data_clients$Kidhome)

##
##      0      1      2
## 1283  887  46

table(data_clients$Teenhome)

##
##      0      1      2
## 1147 1018  51

table(data_clients$NumStorePurchases)

##
##      0      1      2      3      4      5      6      7      8      9     10     11     12     13
##     14      6  220  484  319  211  177  141  147  106  124   80  104   83

```

VARIABLE NUMWEBPURCHASES:

```

boxplot.stats(data_clients$NumWebPurchases)$out

## [1] 23 27 25

```

```
summary(data_clients$NumWebPurchases)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.000   2.000   4.000   4.085   6.000   27.000
```

Els valors extrems són valors que es poden donar. Els donem per bons.

VARIABLE NUMCATALOGPURCHASES:

```
boxplot.stats(data_clients$NumCatalogPurchases)$out
```

```
## [1] 28 11 22 11 11 11 11 11 28 11 11 11 11 11 11 11 28 11 11 11 11 11
```

```
summary(data_clients$NumCatalogPurchases)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.000   0.000   2.000   2.671   4.000   28.000
```

Els valors extrems són valors que es poden donar. Els donem per bons.

VARIABLE MNTMEATPRODUCTS;

Al tenir tants valors diferents, representem el seu boxplot

```
summary(data_clients$MntMeatProducts)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      0.0    16.0    68.0   167.0   232.2   1725.0
```

Sembla que hi han uns 4 valors superiors dels 1500 i diferenciats dels altres

```
table(data_clients$MntMeatProducts[(data_clients$MntMeatProducts>1500)] )
```

```
##
```

```
## 1582 1622 1725
```

```
##    1    1    2
```

Els canviem per la seva mitjana:

```
data_clients$MntMeatProducts[(data_clients$MntMeatProducts>1500)] <- median(data_clients$MntMeatProducts)
```

VARIABLE INCOME:

Al tenir tants valors diferents, representem el seu boxplot

```
summary(data_clients$Income)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1730   35303   51382   52247   68522   666666
```

Hi ha un valor molt superior al altres.

```
table(data_clients$Income[(data_clients$Income>300000)] )
```

```
##
```

```
## 666666
```

```
##    1
```

Els canviem per la seva mitjana:

```
data_clients$Income[(data_clients$Income>300000)] <- median(data_clients$Income)
```

Anàlisi de les dades

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

La selecció dels grups de dades ha de permetre realitzar les proves estadístiques en funció dels grups a comparar o analitzar.

Per això, es seleccionaran les dades en funció de les variables 'Education' i 'Marital Status', aquestes són les que s'han categoritzat anteriorment.

#Agrupació per estudis:

```
data_clients.basic <- data_clients[data_clients$Education == "Basic",]  
data_clients.graduation <- data_clients[data_clients$Education == "Graduation",]  
data_clients.cycle <- data_clients[data_clients$Education == "2n Cycle",]  
data_clients.master <- data_clients[data_clients$Education == "Master",]  
data_clients.phd <- data_clients[data_clients$Education == "PhD",]
```

#Agrupació per estat civil:

```
data_clients.divorced <- data_clients[data_clients$Marital_Status == "Divorced",]  
data_clients.married <- data_clients[data_clients$Marital_Status == "Married",]  
data_clients.single <- data_clients[data_clients$Marital_Status == "Single" || data_clients$Education == "Alone",]  
data_clients.together <- data_clients[data_clients$Marital_Status == "Together",]  
data_clients.widow <- data_clients[data_clients$Marital_Status == "Widow",]
```

Comprovació de la normalitat i homogeneïtat de la variància

Per a poder realitzar les proves estadístiques, cal comprovar la normalitat i la homogeneïtat de la variància.

En aquest cas utilitzarem el test de Shapiro-Wilk, on diu que si el p-valor és més petit que $\alpha=0,05$, es conclou que les dades no estan seguint una distribució normal. Per contra, si p-valor és major que $\alpha=0,05$, les dades sí que estan seguint una distribució normal.

Les variables amb les que es mesurarà la normalitat són les del tipus numèric, i que no siguin la variable ID (corresponen a l'identificador), ni Kidhome ni Teenhome ja que corresponen a valors booleans del tipus 0 i 1.

```
options(scipen = 100)  
par(mfrow=c(2,3))  
p_value <- 0.05  
columnes <- colnames(data_clients)  
colnames_not_check <- c("ID", "Kidhome", "Teenhome")  
for (i in 1:ncol(data_clients)) {  
  col <- data_clients[,i]  
  if(is.numeric(col) && !is.element(colnames(data_clients)[i], colnames_not_check)){  
    qqnorm(col, main = paste("Normal Q-Q plot: ", colnames(data_clients)[i]));  
    qqline(col, col = 2)  
    normality_pvalue <- shapiro.test(col)$p.value  
    if(normality_pvalue < p_value){  
      cat("Distribució NO normalitzada de la columna: ", colnames(data_clients)[i])  
    }else{  

```

```

    cat("Distribució normal de la columna: ", colnames(data_clients)[i])
  }
  cat("\n")
}

```

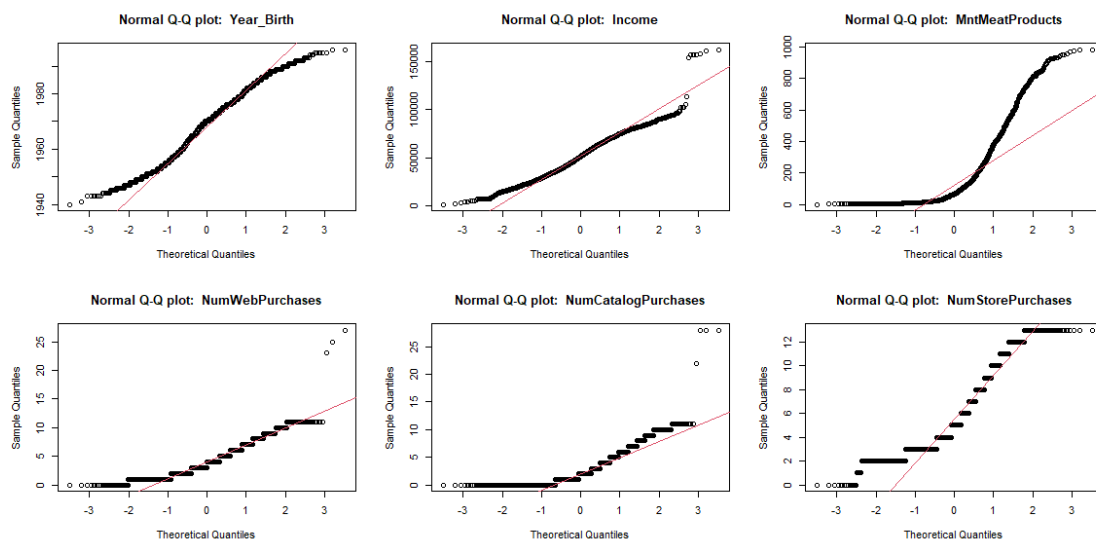
```
## Distribució NO normalitzada de la columna: Year_Birth
```

```
## Distribució NO normalitzada de la columna: Income
```

```
## Distribució NO normalitzada de la columna: MntMeatProducts
```

```
## Distribució NO normalitzada de la columna: NumWebPurchases
```

```
## Distribució NO normalitzada de la columna: NumCatalogPurchases
```



```
## Distribució NO normalitzada de la columna: NumStorePurchases
```

Es pot observar en el diagrama Q-Q i fent el test Shapiro-Wilk, que cap de les variables està normalitzada.

Es proposa normalitzar les variables següents, ja que la resta, es bo que es vegin els valors reals: MntMeatProducts, NumWebPurchases, NumCatalogPurchases, NumStorePurchases. En aquest cas s'aplica la transformació Box-Cox de la llibreria forecast:

```

if (!require('ggplot2')) install.packages('ggplot2'); library('ggplot2')
if (!require('forecast')) install.packages('forecast'); library('forecast')
if (!require('gridExtra')) install.packages('gridExtra'); library('gridExtra')
data_clients_normalitzat <- data_clients
data_clients_normalitzat$MntMeatProducts <- BoxCox(data_clients[,c('MntMeatProducts')], 1/2)
data_clients_normalitzat$NumWebPurchases <- BoxCox(data_clients[,c('NumWebPurchases')], 1/2)
data_clients_normalitzat$NumCatalogPurchases <- BoxCox(data_clients[,c('NumCatalogPurchases')], 1/2)
data_clients_normalitzat$NumStorePurchases <- BoxCox(data_clients[,c('NumStorePurchases')], 1/2)

```


Un cop es tenen les dades normalitzades, es fa l'estudi de l'homoscedasticitat. S'utilitzarà el test de Fligner-Killeen ja que es volen comparar les dades normalitzades amb les que no ho han estat.

Tal i com passava amb el test Shapiro-Wilk, si el p-value és inferior a 0.05, es conclou que les dades no són homogènies:

```
fligner.test(Year_Birth ~ MntMeatProducts, data = data_clients_normalitzat)

##
## Fligner-Killeen test of homogeneity of variances
##
## data: Year_Birth by MntMeatProducts
## Fligner-Killeen:med chi-squared = 659.96, df = 550, p-value = 0.0008539

fligner.test(Year_Birth ~ NumWebPurchases, data = data_clients_normalitzat)

##
## Fligner-Killeen test of homogeneity of variances
##
## data: Year_Birth by NumWebPurchases
## Fligner-Killeen:med chi-squared = 19.778, df = 14, p-value = 0.1373

fligner.test(Year_Birth ~ NumCatalogPurchases, data = data_clients_normalitzat)

##
## Fligner-Killeen test of homogeneity of variances
##
## data: Year_Birth by NumCatalogPurchases
## Fligner-Killeen:med chi-squared = 52.732, df = 13, p-value =
## 0.00001006

fligner.test(Year_Birth ~ NumStorePurchases, data = data_clients_normalitzat)

##
## Fligner-Killeen test of homogeneity of variances
##
## data: Year_Birth by NumStorePurchases
## Fligner-Killeen:med chi-squared = 33.543, df = 13, p-value = 0.001412
```

Les variables següents són homogènies ja que el p-value és superior a 0.05: Year_Birth by NumWebPurchases. La resta no ho són.

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.

Anàlisi de correlació:

Procedim a realitzar un anàlisi de correlació amb la variable MntMeatProducts i així verificar si hi ha alguna variable que tingui una dependència prou gran per creure que pot influir amb el valor de la compra de Carn.

Per fer l'estudi utilitzarem les dades normalitzades (data_clients_normalitzat):

```

corr_matrix <- matrix(nc = 2, nr = 0)
colnames(corr_matrix) <- c("estimate", "p-value")

for (i in 1:(ncol(data_clients_normalitzat) - 1)) {
  if (is.integer(data_clients_normalitzat[,i]) | is.numeric(data_clients_normali
tztat[,i])) {
    spearman_test = cor.test(data_clients_normalitzat[,i],
    data_clients_normalitzat[,10],
    method = "spearman")
    corr_coef = spearman_test$estimate
    p_val = spearman_test$p.value
    pair = matrix(ncol = 2, nrow = 1)
    pair[1][1] = corr_coef
    pair[2][1] = p_val
    corr_matrix <- rbind(corr_matrix, pair)
    rownames(corr_matrix)[nrow(corr_matrix)] <- colnames(data_clients_normalitza
t)[i]
  }
}
print(corr_matrix)

##              estimate      p-value
## ID                -0.01133448 5.938398e-01
## Year_Birth        -0.17932270 1.803758e-17
## Income             0.79277637 0.000000e+00
## Kidhome           -0.59896565 6.659968e-216
## Teenhome          -0.04604096 3.021372e-02
## MntMeatProducts    0.85072506 0.000000e+00
## NumWebPurchases    0.62100133 1.651990e-236
## NumCatalogPurchases 1.00000000 0.000000e+00

```

Les variables amb més correlació amb la quantitat de carn comprada són els que estan més a prop dels valors -1 i 1. En aquest cas la variable amb una correlació més alta és NumCatalogPurchases i Income per tant podem concloure que l'ingrés familiar i el número de compres realitzades per catàleg pot està relacionat amb les compres de carn d'un client.

La conclusió que podem extreure en aquest cas és que els ingressos i l'ús del catàleg com a forma de comprar pot està relacionat amb els clients que compren Carn. Hauríem de fer un anàlisi més extens per poder veure com.

Anàlisi de contrast:

Mirem ara de fer una prova de contrast d'hipòtesi:

Com que quasi totes les relacions entre les variables del dataset original no són homogènies, utilitzarem el tests de kruskal entre els diferents variables i les variables d'estudi, revisant quin valor de p-value dona. Si és menor a 0,05 conclourem que els valors de la variable d'estudi és variant segons els diferents valors de l'altra variable.

Test de Kruskal

```

cont_matrix <- matrix(nc = 2, nr = 0)
colnames(cont_matrix) <- c("statistic", "p-value")

for (i in 1:(ncol(data_clients) - 1)) {
  if (is.integer(data_clients[,i]) | is.numeric(data_clients[,i])) {

```

```

    kruskal_test = kruskal.test(MntMeatProducts ~ data_clients[,i] , data = data
_clients)
    statistic = kruskal_test$statistic
    p_val = kruskal_test$p.value
    pair = matrix(ncol = 2, nrow = 1)
    pair[1][1] = statistic
    pair[2][1] = p_val
    cont_matrix <- rbind(cont_matrix, pair)
    rownames(cont_matrix)[nrow(cont_matrix)] <- colnames(data_clients)[i]
  }
}
print(cont_matrix)

##              statistic      p-value
## ID              2215.0000 4.960041e-01
## Year_Birth      124.6843 2.510831e-07
## Income          2203.3756 1.969276e-04
## Kidhome         676.6741 1.153689e-147
## Teenhome        39.4317 2.738511e-09
## MntMeatProducts 2215.0000 5.336403e-198
## NumWebPurchases 1244.1462 5.584708e-257
## NumCatalogPurchases 1638.5785 0.000000e+00

```

En aquest cas numCatalogPurchases té un valor superior a 0,05. Un resultat coherent amb el resultat que ens ha donat anteriorment revisant la correlació de les variables.

Revisem el test entre els diferents grups d'estudis i estat civil per veure si hi ha alguna relació: Revisem el número d'observacions per dataset:

```

print(paste("Número de Registres clients.basic: ", nrow(data_clients.basic)))
## [1] "Número de Registres clients.basic: 54"

print(paste("Número de Registres clients.graduation: ",nrow(data_clients.graduation)))
## [1] "Número de Registres clients.graduation: 1116"

print(paste("Número de Registres clients.cycle: ",nrow(data_clients.cycle)))
## [1] "Número de Registres clients.cycle: 200"

print(paste("Número de Registres clients.master: ",nrow(data_clients.master)))
## [1] "Número de Registres clients.master: 365"

print(paste("Número de Registres clients.phd: ",nrow(data_clients.phd)))
## [1] "Número de Registres clients.phd: 481"

print(paste("Número de Registres clients.divorced: ",nrow(data_clients.divorced)))
## [1] "Número de Registres clients.divorced: 232"

print(paste("Número de Registres clients.married: ",nrow(data_clients.married)))

```



```

## [1] "t-students graduation vs married: 0.988193901006762"

print(paste("t-students graduation vs single:",t.test(data_clients.graduation$MntMeatProducts, data_clients.single$MntMeatProducts,alternative = "less")$p.value))

## [1] "t-students graduation vs single: 0.950841784310524"

print(paste("t-students graduation vs widow:",t.test(data_clients.graduation$MntMeatProducts, data_clients.widow$MntMeatProducts,alternative = "less")$p.value))

## [1] "t-students graduation vs widow: 0.378878727878976"

print(paste("t-students cycle vs divorced:",t.test(data_clients.cycle$MntMeatProducts, data_clients.divorced$MntMeatProducts,alternative = "less")$p.value))

## [1] "t-students cycle vs divorced: 0.206130641913088"

print(paste("t-students cycle vs married:",t.test(data_clients.cycle$MntMeatProducts, data_clients.married$MntMeatProducts,alternative = "less")$p.value))

## [1] "t-students cycle vs married: 0.0953053949320644"

print(paste("t-students cycle vs single:",t.test(data_clients.cycle$MntMeatProducts, data_clients.single$MntMeatProducts,alternative = "less")$p.value))

## [1] "t-students cycle vs single: 0.0230222844863549"

print(paste("t-students cycle vs together:",t.test(data_clients.cycle$MntMeatProducts, data_clients.together$MntMeatProducts,alternative = "less")$p.value))

## [1] "t-students cycle vs together: 0.0405216750043374"

print(paste("t-students cycle vs widow:",t.test(data_clients.cycle$MntMeatProducts, data_clients.widow$MntMeatProducts,alternative = "less")$p.value))

## [1] "t-students cycle vs widow: 0.0388312320769009"

print(paste("t-students master vs divorced:",t.test(data_clients.master$MntMeatProducts, data_clients.divorced$MntMeatProducts,alternative = "less")$p.value))

## [1] "t-students master vs divorced: 0.77355834021978"

print(paste("t-students master vs married:",t.test(data_clients.master$MntMeatProducts, data_clients.married$MntMeatProducts,alternative = "less")$p.value))

## [1] "t-students master vs married: 0.709468196727142"

print(paste("t-students master vs single:",t.test(data_clients.master$MntMeatProducts, data_clients.single$MntMeatProducts,alternative = "less")$p.value))

## [1] "t-students master vs single: 0.461952541364877"

print(paste("t-students master vs together:",t.test(data_clients.master$MntMeatProducts, data_clients.together$MntMeatProducts,alternative = "less")$p.value))

## [1] "t-students master vs together: 0.482616837621924"

```

```

print(paste("t-students master vs widow:",t.test(data_clients.master$MntMeatProducts, data_clients.widow$MntMeatProducts,alternative = "less")$p.value))

## [1] "t-students master vs widow: 0.206626746542938"

print(paste("t-students phd vs divorced:",t.test(data_clients.phd$MntMeatProducts, data_clients.divorced$MntMeatProducts,alternative = "less")$p.value))

## [1] "t-students phd vs divorced: 0.80160757513977"

print(paste("t-students phd vs married:",t.test(data_clients.phd$MntMeatProducts, data_clients.married$MntMeatProducts,alternative = "less")$p.value))

## [1] "t-students phd vs married: 0.749489704805712"

print(paste("t-students phd vs single:",t.test(data_clients.phd$MntMeatProducts, data_clients.single$MntMeatProducts,alternative = "less")$p.value))

## [1] "t-students phd vs single: 0.471333066464326"

print(paste("t-students phd vs together:",t.test(data_clients.phd$MntMeatProducts, data_clients.together$MntMeatProducts,alternative = "less")$p.value))

## [1] "t-students phd vs together: 0.49394976105415"

print(paste("t-students phd vs widow:",t.test(data_clients.phd$MntMeatProducts, data_clients.widow$MntMeatProducts,alternative = "less")$p.value))

## [1] "t-students phd vs widow: 0.203924333993742"

```

Anàlisi de regressió:

Per obtenir un model de regressió lineal eficient ho farem a partir de les variables amb més correlació que haguem trobat de les variables en estudi.

En aquest cas, i segons els estudis anteriors, ens centrarem amb la variable MntMeatProducts i Income, NumCatalogPurchase, NumStorePurchases i NumWebPurchases.

```

model1 <- lm(data_clients$MntMeatProducts ~ data_clients$Income , data = data_clients)
model2 <- lm(data_clients$MntMeatProducts ~ data_clients$NumCatalogPurchase , data = data_clients)
model3 <- lm(data_clients$MntMeatProducts ~ data_clients$Income + data_clients$NumCatalogPurchase , data = data_clients)
model4 <- lm(data_clients$MntMeatProducts ~ data_clients$Income + data_clients$NumCatalogPurchase + data_clients$NumStorePurchases , data = data_clients)
model5 <- lm(data_clients$MntMeatProducts ~ data_clients$Income + data_clients$NumCatalogPurchase + data_clients$NumStorePurchases + data_clients$NumWebPurchases , data = data_clients)

tabla.quoficients <- 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),
ncol = 2, byrow = TRUE)

```

```
colnames(tabla.quoficients) <- c("Model", "R^2")
tabla.quoficients
```

```
##      Model      R^2
## [1,]      1 0.4637420
## [2,]      2 0.4304784
## [3,]      3 0.5279354
## [4,]      4 0.5368668
## [5,]      5 0.5375996
```

En aquest cas veiem que tots tenen una coeficient de determinació baix, per tant ens quedem l'últim model per fer la predicció del número de diners gastats per un client amb carn.

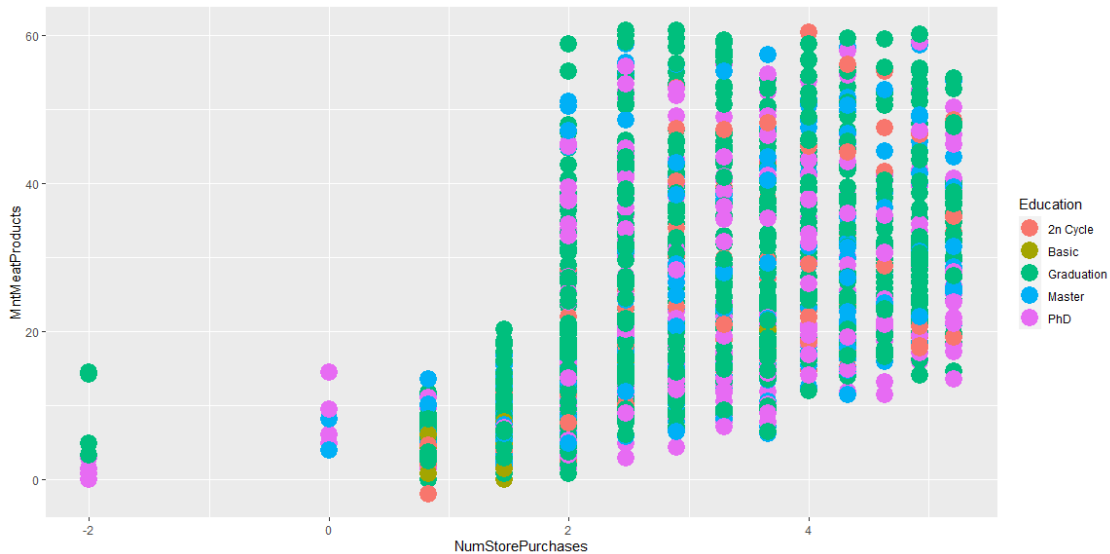
Representació dels resultats a partir de taules gràfiques

Les taules gràfiques han d'ajudar a representar els resultats mostrats anteriorment.

Primerament, es comprova que les compres realitzades la carn i fetes des del catàleg en relació als estudis que tenen (s'ha comprovat en les proves estadístiques que els estudis i les compres de carn tenen relació, s'ha afegit en aquest gràfic les compres de carn per a poder contrastar el diferents productes):

```
if (!require('ggplot2')) install.packages('ggplot2'); library(ggplot2)

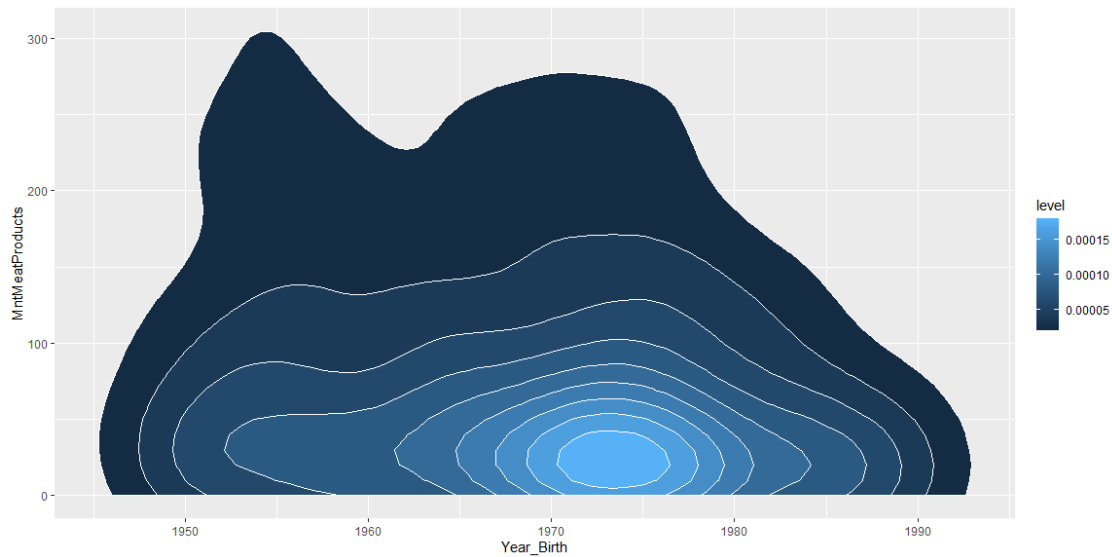
ggplot(data_clients_normalitzat, aes(x=NumStorePurchases, y=MntMeatProducts, col=
  or=Education)) +
  geom_point(size=6)
```



Es pot observar que hi ha una gran diferència de les persones amb estudis bàsics, que gasten menys diners en la compra de carn. Tot i així, les persones amb la resta d'estudis, compren per sobre de la mitjana.

Degut a que l'educació no és un factor rellevant, es decideix fer ús de l'edat:

```
ggplot(data_clients, aes(x=Year_Birth, y=MntMeatProducts) ) +
  stat_density_2d(aes(fill = ..level..), geom = "polygon", colour="white")
```



En aquest cas s'observa que l'edat (any de naixement) és una bona variable que influeix en la compra de productes. Per això es discretitza l'edat en 5 segments, i es visualitza per a cada segment d'edat i educació, el nombre de productes comprats.

```
data_clients["segment_edat"] <- cut(data_clients$Year_Birth, breaks = c(0, 1950,
1960, 1970, 1980, Inf),
labels = c("<1950", "1950-1959", "1960-1969", "1970-1979", ">1
980"))

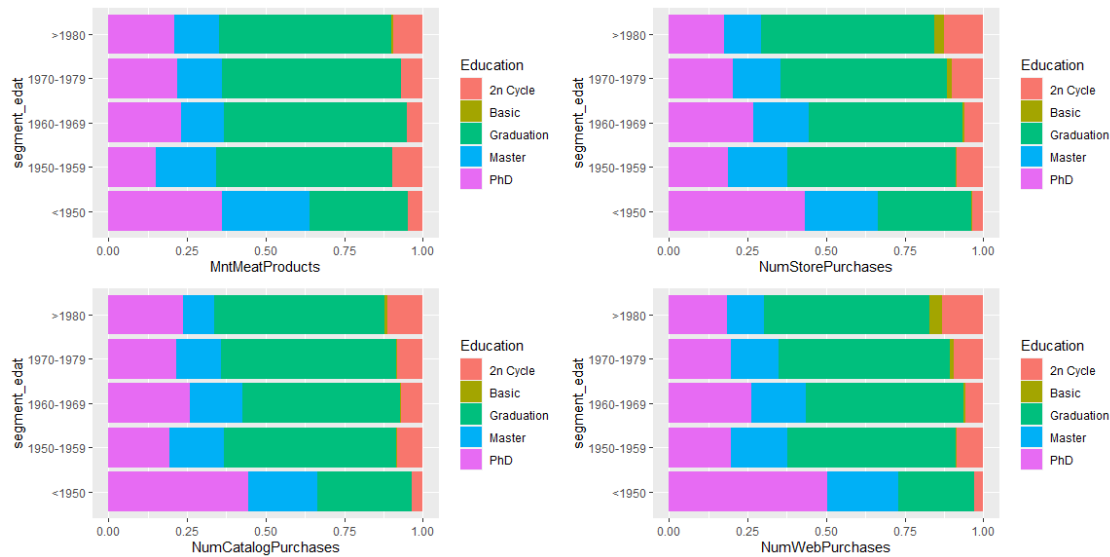
plotMeat <- ggplot(data_clients, aes(fill=Education, y=segment_edat, x=MntMeatPr
oducts)) +
  geom_bar(position="fill", stat="identity")

plotStore <- ggplot(data_clients, aes(fill=Education, y=segment_edat, x=NumStore
Purchases)) +
  geom_bar(position="fill", stat="identity")

plotCatalog <- ggplot(data_clients, aes(fill=Education, y=segment_edat, x=NumCat
alogPurchases)) +
  geom_bar(position="fill", stat="identity")

plotWeb <- ggplot(data_clients, aes(fill=Education, y=segment_edat, x=NumWebPurc
hases)) +
  geom_bar(position="fill", stat="identity")

grid.arrange(plotMeat, plotStore, plotCatalog, plotWeb)
```

Finalment, es comprova les compres realitzades segons els diners gastats en les compres dels productes:

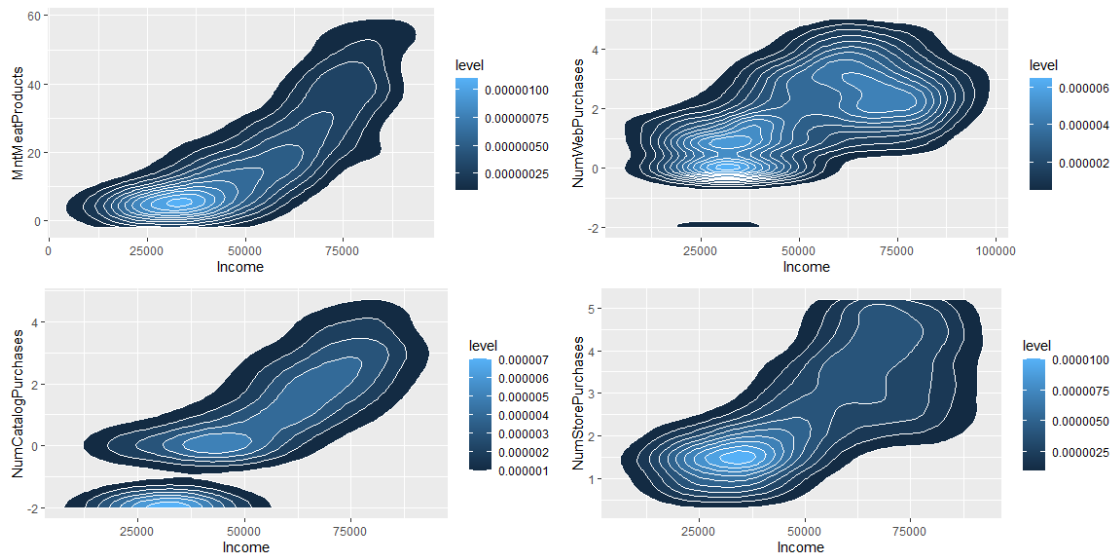
```
plotMeat <- ggplot(data_clients_normalitzat, aes(x=Income, y=MntMeatProducts) )
+
  stat_density_2d(aes(fill = ..level..), geom = "polygon", colour="white")

plotWeb <- ggplot(data_clients_normalitzat, aes(x=Income, y=NumWebPurchases) ) +
  stat_density_2d(aes(fill = ..level..), geom = "polygon", colour="white")

plotCatalog <- ggplot(data_clients_normalitzat, aes(x=Income, y=NumCatalogPurchases) ) +
  stat_density_2d(aes(fill = ..level..), geom = "polygon", colour="white")

plotStore <- ggplot(data_clients_normalitzat, aes(x=Income, y=NumStorePurchases) ) +
  stat_density_2d(aes(fill = ..level..), geom = "polygon", colour="white")

grid.arrange(plotMeat, plotWeb, plotCatalog, plotStore)
```



Resolució del problema

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

L'objectiu d'aquest estudi és identificar el tipus de client que compra de carn i crear així campanyes de màrqueting específiques. Per això, s'han seleccionat les variables adients per a resoldre el problema i s'han netejat i tractat per a poder analitzar les dades i representar-les

Finalment, amb la representació dels resultats es conclou el següent:

- Els aspectes que més influeixen en la compra són els ingressos anuals i l'edat del client. L'educació ens indica que els clients amb estudis bàsics, compren menys, però això es degut a que tenen feines menys qualificades i amb menys ingressos.
- Els clients més habituals són els nascuts entre 1970 i 1975, però el gasto per al producte és baix.
- Hi ha una relació directa entre els ingressos mensuals i el gasto fet en les compres dels productes. Tot i així, és important saber que la majoria dels clients tenen un ingrés de 35,000 dollars anuals. I el gasto de productes augmenta a partir dels 45,000 i 50,000 dollars.
- Pel que fa als estudis, els clients més habituals són els graduats, a excepció dels nascuts abans del 1950 que tenen un PhD.

Es guarden les dades utilitzades en un altre CSV:

```
write.csv(data_clients_normalitzat, '../data/PRA2_Marketing_campaign.csv')
```

Contribucions	Signatura
Investigació prèvia	Isabel Barrera Benavent, Maria Font Sánchez
Redacció de les respostes	Isabel Barrera Benavent, Maria Font Sánchez
Desenvolupament del codi	Isabel Barrera Benavent, Maria Font Sánchez