Activitat 4

Marc Cervera Rosell

21-06-2024

```
setRepositories(ind=2) # Per descarregar paquets de CRAN
```

1 Preprocessament

1.1 Variables Income i Year_Birth

```
tryCatch({
  data <- read.csv("marketing.csv", header = TRUE)
  print("Fitxer llegit correctament")
}, error = function(e){
  cat("ERROR en la lectura del fitxer:",conditionMessage(e),"\n")
})
```

```
## [1] "Fitxer llegit correctament"
```

```
columns <- names(data)
type <- sapply(data, class)
for (i in seq_along(columns)) {
  cat("La columna", columns[i], "es de tipus", type[i], "\n")
}
```

```
## La columna ID es de tipus integer
## La columna Year_Birth es de tipus integer
## La columna Education es de tipus character
## La columna Marital_Status es de tipus character
## La columna Income es de tipus integer
## La columna Kidhome es de tipus integer
## La columna Teenhome es de tipus integer
## La columna Dt_Customer es de tipus character
## La columna Recency es de tipus integer
## La columna MntWines es de tipus integer
## La columna MntFruits es de tipus integer
## La columna MntMeatProducts es de tipus integer
## La columna MntFishProducts es de tipus integer
## La columna MntSweetProducts es de tipus integer
## La columna MntGoldProds es de tipus integer
## La columna NumDealsPurchases es de tipus integer
## La columna NumWebPurchases es de tipus integer
## La columna NumCatalogPurchases es de tipus integer
## La columna NumStorePurchases es de tipus integer
## La columna NumWebVisitsMonth es de tipus integer
## La columna AcceptedCmp3 es de tipus integer
## La columna AcceptedCmp4 es de tipus integer
```

```
## La columna AcceptedCmp5 es de tipus integer
## La columna AcceptedCmp1 es de tipus integer
## La columna AcceptedCmp2 es de tipus integer
## La columna Complain es de tipus integer
## La columna Z_CostContact es de tipus integer
## La columna Z_Revenue es de tipus integer
## La columna Response es de tipus integer
```

Després d'observar els tipus de les variables del conjunt de dades, solament es procedirà a fer un canvi de tipus. Aquest canvi es produirà en la variable *Dt_Customer* que passarà de ser de tipus *character* a tipus *date*.

```
data_transformed <- transform(data,
                              Dt_Customer = as.Date(Dt_Customer))
```

```
columns <- names(data_transformed)
type <- sapply(data_transformed, class)
for (i in seq_along(columns)) {
  cat("La columna", columns[i], "es de tipus", type[i], "\n")
}
```

```
## La columna ID es de tipus integer
## La columna Year_Birth es de tipus integer
## La columna Education es de tipus character
## La columna Marital_Status es de tipus character
## La columna Income es de tipus integer
## La columna Kidhome es de tipus integer
## La columna Teenhome es de tipus integer
## La columna Dt_Customer es de tipus Date
## La columna Recency es de tipus integer
## La columna MntWines es de tipus integer
## La columna MntFruits es de tipus integer
## La columna MntMeatProducts es de tipus integer
## La columna MntFishProducts es de tipus integer
## La columna MntSweetProducts es de tipus integer
## La columna MntGoldProds es de tipus integer
## La columna NumDealsPurchases es de tipus integer
## La columna NumWebPurchases es de tipus integer
## La columna NumCatalogPurchases es de tipus integer
## La columna NumStorePurchases es de tipus integer
## La columna NumWebVisitsMonth es de tipus integer
## La columna AcceptedCmp3 es de tipus integer
## La columna AcceptedCmp4 es de tipus integer
## La columna AcceptedCmp5 es de tipus integer
## La columna AcceptedCmp1 es de tipus integer
## La columna AcceptedCmp2 es de tipus integer
## La columna Complain es de tipus integer
## La columna Z_CostContact es de tipus integer
## La columna Z_Revenue es de tipus integer
## La columna Response es de tipus integer
```

S'observa que després de l'aplicació de la funció *transform()* el tipus de la variable *Dt_Customer* queda modificat.

Finalment, cal excloure les variables *Z_CostContact* i *Z_Revenue*, atès que són variables de control i que així s'indica a l'enunciat de l'activitat.

```
columns_to_exclude <- c("Z_CostContact", "Z_Revenue")
data_with_no_control_variables <- data_transformed[, !(names(data_transformed) %in%
                                                         columns_to_exclude)]
print(names(data_with_no_control_variables))
```

```
##  [1] "ID"                 "Year_Birth"          "Education"
##  [4] "Marital_Status"     "Income"              "Kidhome"
##  [7] "Teenhome"           "Dt_Customer"         "Recency"
## [10] "MntWines"           "MntFruits"           "MntMeatProducts"
## [13] "MntFishProducts"    "MntSweetProducts"    "MntGoldProds"
## [16] "NumDealsPurchases"  "NumWebPurchases"     "NumCatalogPurchases"
## [19] "NumStorePurchases"  "NumWebVisitsMonth"   "AcceptedCmp3"
## [22] "AcceptedCmp4"       "AcceptedCmp5"        "AcceptedCmp1"
## [25] "AcceptedCmp2"       "Complain"            "Response"
```

S'observa que en treure per pantalla les columnes del *dataset data_with_no_control_variables* les variables de control indicades anteriorment ja no hi són. Per tant, aquest conjunt final queda completament operatiu per a poder treballar.

1.2 Valors absents

Abans de res, encara que l'enunciat ja ho diu, és bona pràctica comprovar si realment hi ha valors absents.

```
any(is.na(data_with_no_control_variables$Income))
```

```
## [1] TRUE
```

```
any(is.na(data_with_no_control_variables$Year_Birth))
```

```
## [1] FALSE
```

```
library(VIM)
```

```
## Loading required package: colorspace
## Loading required package: grid
## VIM is ready to use.
## Suggestions and bug-reports can be submitted at: https://github.com/statistikat/VIM/issues
##
## Attaching package: 'VIM'
## The following object is masked from 'package:datasets':
##
##     sleep
```

Primer cal seleccionar aquelles variables que s'usaran per al càlcul de la distància de Gower. Com s'especifica en l'enunciat seran les variables 10 a 15.

```
variables_gower <- c("MntWines", "MntFruits", "MntMeatProducts", "MntFishProducts",
                     "MntSweetProducts", "MntGoldProds")
```

Com s'indica en les instruccions de l'activitat en desenvolupament, per imputar els valors *NA* de la varible *Income* s'aplica la funció *kNN* de la llibreria *VIM*.

```
income_imputed <- kNN(data_with_no_control_variables, variable = "Income",
                      dist_var = variables_gower, k = 5)
```

Un cop aplicada la funció *kNN* s'observa que la variable *Income* ja no té valors *NA*

```
any(is.na(income_imputed$Income))
```

```
## [1] FALSE
```

En primer lloc, abans de calcular la mitjana d'edat cal seleccionar les persones que són vídues.

```
widowed_people <- subset(income_imputed, Marital_Status == "Widow")
```

Un cop seleccionades aquestes persones ja es pot procedir al càlcul de la mitjana d'edat.

```
mean_widow_age <- mean(widowed_people$Year_Birth)
cat("L'any de naixement mitjà de les persones vidues es:"
    , round(mean_widow_age))
```

```
## L'any de naixement mitjà de les persones vidues es: 1959
```

```
income_imputed$Year_Birth <- ifelse(is.na(income_imputed$Year_Birth), mean_widow_age,
                                    income_imputed$Year_Birth)
```

```
any(is.na(income_imputed$Year_Birth))
```

```
## [1] FALSE
```

S'observa que després de l'execució de la cel · la que conté la sentència *ifelse* els valors *NA* de la variable *Year_Birth* ja no són tals.

```
any(is.na(income_imputed))
```

```
## [1] FALSE
```

Un cop eliminats els valors *NA* de les variables *Income* i *Year_Birth* es pot observar (en la cel · la anterior) el conjunt de dades queda completament lliure de valors *NA*, per tant, cal redenominar aquest conjunt a *markclean*.

```
markclean <- income_imputed
```

Finalment, es demana una reflexió sobre el nombre de valors *NA* del fitxer, per tant, cal, en primer lloc, comptar aquests valors.

```
na_per_column <- colSums(is.na(data_with_no_control_variables))
total_na_values <- sum(na_per_column)
cat("El nombre total de valors NA es:", total_na_values)
```

```
## El nombre total de valors NA es: 24
```

Tot i que el nombre de valors *NA* del fitxer no és considerablement elevat, és important abordar aquests valors. El fet de l'existència de valors *NA* pot ser un indicador de problemes en les dades com: problemes en el moment de la recollida de les dades o senzillament que en el moment d'introduir les dades en el fitxer s'ha comès error humà. Com s'acaba d'esmentar, malgrat que el nombre de valors absents no és molt significatiu, és molt important gestionar aquests buits per evitar problemes en les anàlisis que es puguin arribar a fer tant en aquesta activitat com les anàlisis que pugui fer una altra persona completament aliena a la UOC.

2 Estadística descriptiva

2.1 Income

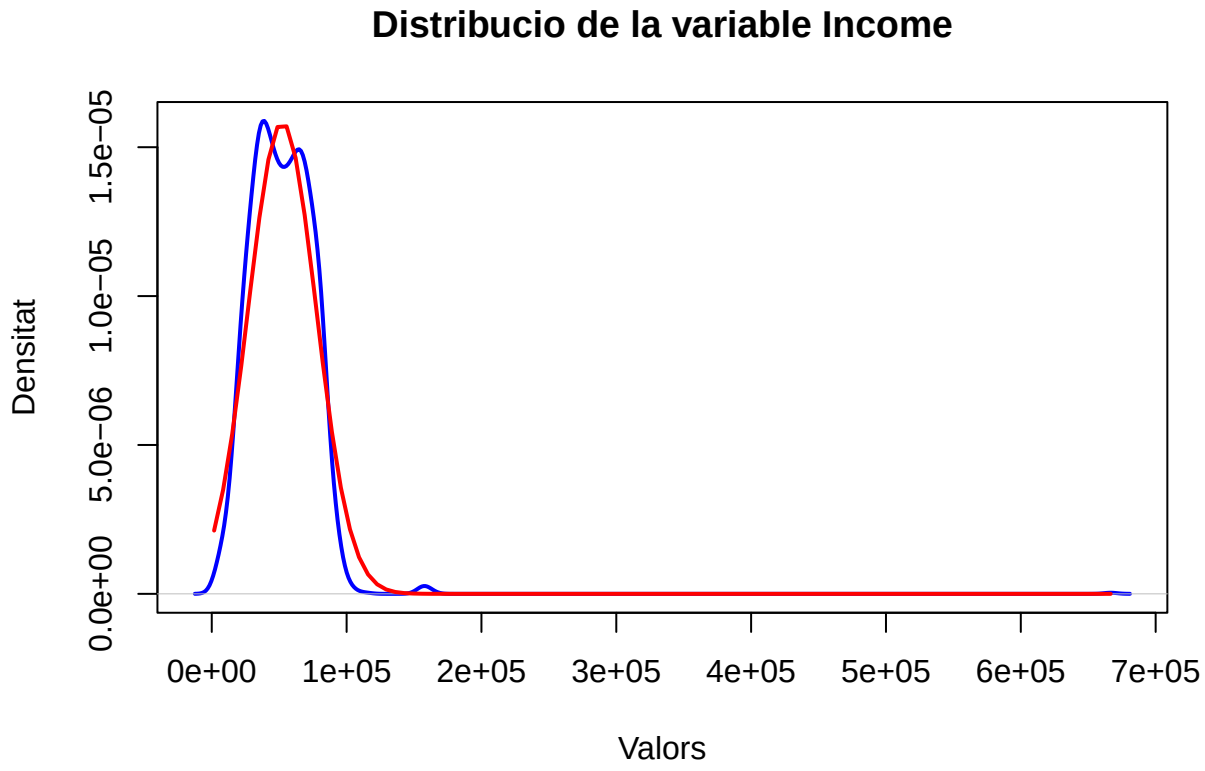
```
mean_abs <- abs(mean(markclean$Income))
standard_deviation <- sd(markclean$Income)
coeffitient_of_variation <- standard_deviation / mean_abs
cat("El coeficient de variacio de la variable Income es:", coeffitient_of_variation)
```

```
## El coeficient de variacio de la variable Income es: 0.4824468
```

Es pot observar un coeficient de variació de 0.4824468. Això indica que, efectivament, existeix una variabilitat en la distribució, però que no és excessiva. Per tant, l'ingrés mitjà sí que és un valor representatiu de la distribució dels ingressos.

Per respondre a la segona pregunta cal, prèviament, utilitzar alguna eina per visualitzar la distribució de les dades de la variable. En aquest cas s'ha escollit veure la distribució en un gràfic de densitat. També s'introduirà una campana de Gauss (densitat normal) de les dades d'estudi per veure com de lluny estan de distribuir-se de manera normal.

```
plot(density(markclean$Income),
     main = "Distribucio de la variable Income",
     xlab = "Valors",
     ylab = "Densitat",
     col = "blue",
     lwd = 2)
values_normal_distribution <- seq(min(markclean$Income), max(markclean$Income),
                                  length = 100)
normal_distribution <- dnorm(values_normal_distribution, mean = mean_abs,
                             sd = standard_deviation) # Campana de Gauss de les dades
lines(values_normal_distribution, normal_distribution, col = "red", lwd = 2)
```



Després d'observar ambdós gràfics, es pot observar que les dades no normalitzades (línia blava) no segueixen una distribució normal atès que la forma de la corba en el gràfic de densitat queda molt allunyada de la forma de la línia vermella que representa les dades normalitzades.

2.2 Education

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

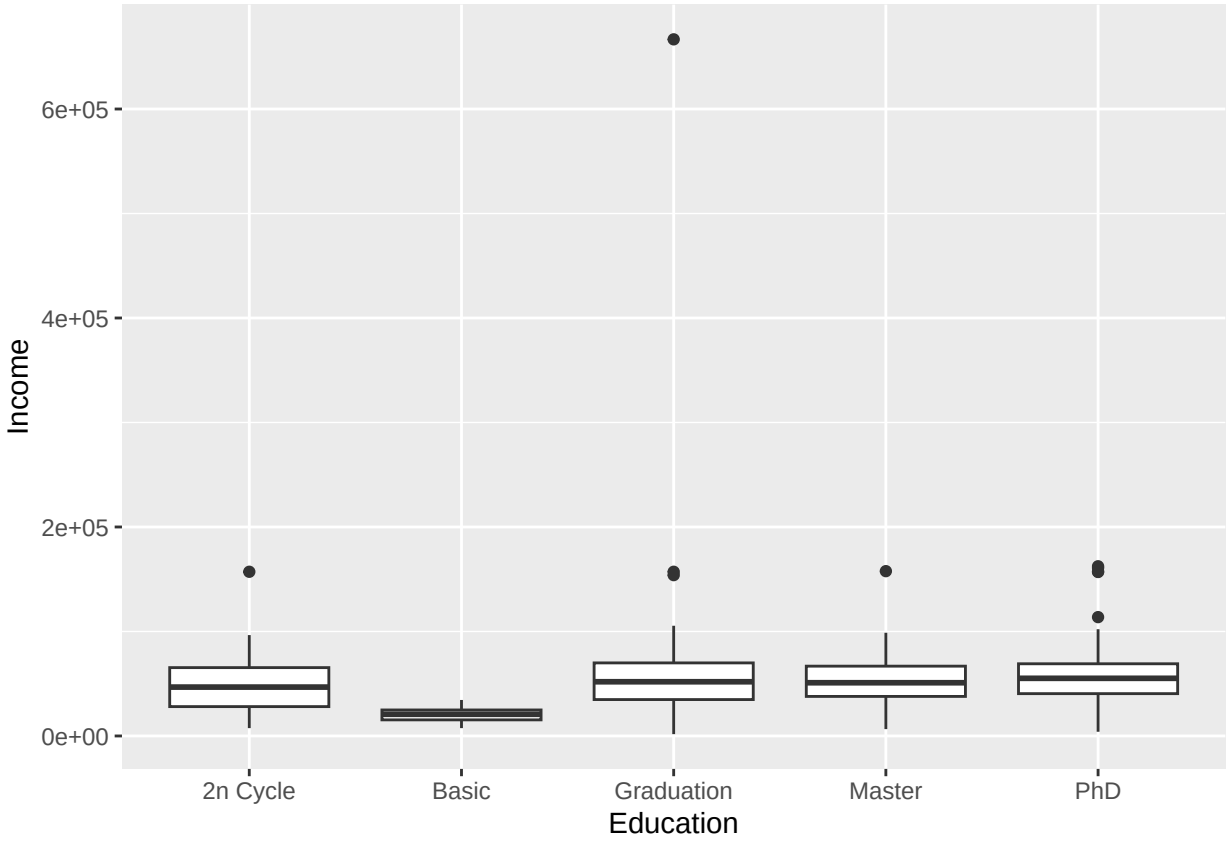
```
library(ggplot2)
library(magrittr) # Per l'operador %>%
```

L'operador "%>%" és un operador que crea un *pipe* i permet que el resultat d'una funció passi com a primer argument de la següent funció.

```
markclean %>%
  group_by(Education) %>%
  summarize(
    mean_value = mean(Income),
    observations = n(),
    deviation = sd(Income)
  ) %>%
  arrange(desc(Education))
```

```
## # A tibble: 5 x 4
##   Education  mean_value observations deviation
##   <chr>           <dbl>        <int>     <dbl>
## 1 PhD            56061.          486    20544.
## 2 Master         53014.          370    20082.
## 3 Graduation     52616.         1127    28125.
## 4 Basic          20306.           54     6235.
## 5 2n Cycle       48051.          203    23298.
```

```
markclean %>%
  arrange(desc(Education)) %>%
  ggplot(aes(x = Education, y = Income)) +
  geom_boxplot() +
  labs(x = "Education", y = "Income", tittle = "Ingressos segons nivell educatiu")
```



En primer lloc, es pot observar que la posició de la mediana (línia interna de la caixa), es troba just al centre en els nivells educatius "Graduation", "Master" i "PhD", per tant, en aquests tres casos es pot concloure que el 50% dels valors estan per sota de la mediana i l'altre 50% per sobre. Per als nivells educatius "2n Cycle" i "Basic", es pot observar que en el cas del primer nivell educatiu comentat, la línia de la mediana es troba lleugerament desplaçada a la part superior de la caixa i en el cas del nivell "Basic" la línia de la mediana està a la part superior de la caixa, per tant, en aquests dos casos les dades presenten una asimetria positiva cosa que indica que la majoria de les dades es troben a l'esquerra de la mediana. El fet de tenir una asimetria positiva és indicatiu, en aquest cas, que una petita proporció de persones tenen més ingressos que la majoria de les persones del mateix nivell educatiu.

Observant els bigotis de les caixes, es pot veure que n'hi ha de més curt i de més llargs. Segons la longitud dels bigotis de cada caixa, es podrà veure com d'agrupades estan les dades, és a dir, es podrà observar com són de dispersos els valors extrems. Per interpretar els bigotis de les caixes cal mirar la seva longitud. Com més llargs siguin els bigotis més dispersos estaran els valors extrems, és a dir, els valors extrems estaran més lluny de la resta de valors. Per contra, com més curts siguin els bigotis de les caixes, menys dispersos estaran els valors extrems i, per tant, més propers estaran a la resta de valors.

Finalment, s'observa la presència de valors atípics. Els valors atípics són els punts que estan situats fora dels bigotis, però això no significa que no siguin valors importants.

3. Estadística inferencial

3.1 Contrast d'hipòtesi per a la diferència de les mitjanes

3.1.1 Escriviu la hipòtesi nul · la i l'alternativa

Hipòtesi nul · la: Els ingressos mitjans de les persones sense estudis universitaris són iguals als de les persones amb estudis universitaris.

Hipòtesi alternativa: Els ingressos mitjans de les persones sense estudis són inferiors als de les persones amb estudis universitaris.

3.1.2 Justificació del test a aplicar

Per a poder aplicar un test de diferència de mitjanes, és condició necessària que les mostres a examinar siguin independents.

Per comprovar la igualtat, o diferència, de les variàncies s'aplicarà un test de Fisher.

Hipòtesi nul · la: Les variàncies són iguals.

Hipòtesi nul · la: Les variàncies són diferents.

```
alpha <- 0.01
# Ingressos no universitaris
income_no_uni <- markclean$Income[markclean$Education %in% c("2n Cycle", "Basic")]
# Ingressos universitaris
income_uni <- markclean$Income[markclean$Education %in% c("Graduation",
                                                          "Master", "PhD")]
mean_no_uni <- mean(income_no_uni)
mean_uni <- mean(income_uni)
n_no_uni <- length(income_no_uni)
n_uni <- length(income_uni)
s_no_uni <- sd(income_no_uni)
s_uni <- sd(income_uni)
c(mean_no_uni, mean_uni, s_no_uni, s_uni, n_no_uni, n_uni)
```

```
## [1] 42221.33 53534.81 23761.62 25096.78   257.00  1983.00
```

```
fobs <- s_no_uni^2 / s_uni^2
fcritL <- qf(alpha/2, df1 = n_no_uni - 1, df2 = n_uni - 2)
fcritU <- qf(1 - alpha/2, df1 = n_no_uni - 1, df2 = n_uni - 2)
pvalue <- min(pf(fobs, df1 = n_no_uni - 1, df2 = n_uni - 2, lower.tail = FALSE),
              pf(fobs, df1 = n_no_uni - 1, df2 = n_uni - 2)) * 2
c(fobs, fcritL, fcritU, pvalue)
```

```
## [1] 0.8964297 0.7769809 1.2620087 0.2613326
```

Atès que el valor observat es troba dins dels límits L i U i que el valor P és major al valor alfa es pot concloure que no hi ha evidència suficient per a rebutjar la hipòtesi nul · la i que amb un 99% de confiança les variàncies d'ambdues poblacions són iguals.

3.1.3 Càlculs

```
combined_s <- sqrt((((n_no_uni - 1) * s_no_uni^2) + ((n_uni - 1) * s_uni^2))/
                   (n_no_uni + n_uni - 2))
t_obs <- (mean_no_uni - mean_uni) / (combined_s * sqrt((1 / n_no_uni) + (1 / n_uni)))
tcritL <- qt(alpha / 2, n_no_uni + n_uni - 2)
tcritU <- qt(1 - alpha / 2, n_no_uni + n_uni - 2)
```

```r
pvalue_test <- pt(abs(t_obs), df = n_no_uni + n_uni - 2, lower.tail = FALSE) * 2
c(t_obs, tcritL, tcritU, pvalue_test)
```

```
## [1] -6.840222e+00 -2.578028e+00  2.578028e+00  1.016752e-11
```

3.1.4 Interpretació del test

Vist que el valor observat (t_obs) es troba fora del rang de valors crítics i que el valor p (pvalue_test) és inferior al valor alfa (0.01) es rebutja la hipòtesi nul · la. Per tant, es pot afirmar que amb un 99% de confiança els ingressos mitjans de les persones sense estudis són inferiors als de les persones amb estudis universitaris.

4. Model de regressió

4.1 Regressió lineal múltiple

```r
markclean_bkp <- markclean
# Backup del dataset
```

```r
markclean$Education <- factor(markclean$Education,
                              levels = c("Basic","2n Cycle","Graduation", "Master", "PhD"))

set.seed(123)
training_index <- sample(seq_len(nrow(markclean)),
                         size = 0.8 * nrow(markclean))
set_training <- markclean[training_index,]
set_ttesting <- markclean[-training_index,]

model <- lm(Income ~ Year_Birth + Kidhome + Teenhome + Education + MntWines + MntFruits +
              MntMeatProducts + MntFishProducts + MntSweetProducts + MntGoldProds +
              NumDealsPurchases + NumCatalogPurchases + NumStorePurchases +
              NumWebVisitsMonth, data = set_training)
summary_model <- summary(model)
cat("Valor R-squared = ",summary_model$r.squared, "\n")
```

```
## Valor R-squared =  0.5112118
```

```r
cat("Valor R-sqaured ajustat = ",summary_model$adj.r.squared)
```

```
## Valor R-sqaured ajustat =  0.5065278
```

```r
predictions <- predict(model, newdata = set_ttesting)
sum_for_RMSE <- 0
for (i in 1:length(predictions)) {
  sum_for_RMSE <- sum_for_RMSE + ((set_ttesting$Income[i] - predictions[[i]])^2)
}
rmse <- sqrt(sum_for_RMSE / nrow(set_ttesting))
cat("RMSE = ",rmse)
```

```
## RMSE =  12305.76
```

Per comprovar la qualitat de l'ajust, s'observaran i interpretaran tres valors que atorga el *summary* del model. El primer valor és el *Multiple R-squared*. Aquest valor és de 0.5112 (51.12%), això indica que les variables incloses en el model poden explicar fins al 51.12% de les variacions dels ingressos. El fet que aquest valor superi el 50% és indicatiu de què el model és relativament bo. Per considerar el model bo i no relativament bo, caldria que aquest valor superés el 70%.

El segon valor a interpretar és el *Adjusted R-squared*. Aquest valor és una versió modificada de l'anterior que ha estat ajustat al nombre de predictors del model. Aquest valor és de 0.5065 (50.65%), per tant, el model encara pot explicar el 50.65% de les variacions dels ingressos. Vist que continua superant el 50%, es pot concloure que l'ajust continua sent relativament bo.

Finalment, l'últim valor a considerar per la qualitat de l'ajust és el *Residual standard error* (RSE). Aquest valor indica la dispersió dels punts al voltant de la línia del millor ajustament.

Com el RMSE està calculat en els ingressos (variable depenent), es pot interpretar el resultat del RMSE comparant amb el valor mitjà dels ingressos.

```
mean_income <- mean(markclean$Income)
cat("Ingressos mitjans:", mean_income)
```

```
## Ingressos mitjans: 52236.79
```

S'observa que el RMSE és bastant menor que els ingressos mitjans. En vista d'aquest fet, es conclou que el model realitza relativament bé les prediccions.

4.1.1 Multicolinealitat

```
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
vif_vals <- vif(model)
```

```
print(vif_vals)
```

```
##                       GVIF Df GVIF^(1/(2*Df))
## Year_Birth        1.248919  1        1.117550
## Kidhome           1.839880  1        1.356422
## Teenhome          1.571542  1        1.253612
## Education         1.240658  4        1.027322
## MntWines          2.621946  1        1.619242
## MntFruits         1.905105  1        1.380256
## MntMeatProducts   2.744147  1        1.656547
## MntFishProducts   2.095572  1        1.447609
## MntSweetProducts  1.887620  1        1.373907
## MntGoldProds      1.516210  1        1.231345
## NumDealsPurchases 1.616485  1        1.271411
## NumCatalogPurchases 3.033996  1        1.741837
## NumStorePurchases 2.257205  1        1.502400
## NumWebVisitsMonth 1.978885  1        1.406729
```

Després d'observar els resultats de la funció *vif()* és pot observar que la multicol · linealitat del model no és un problema (no significa que no hi hagi multicol · linealitat) atès que cap valor GVIF és superior a 5. Pel que fa als resultats de la tercera columna, els valors tampoc són alarmants tot i que cal estar atent.

4.2 Regressió logística

4.2.1 Model predictiu

```
markclean <- markclean_bkp
# Desfeta del factor a la columna Education
```

```
set.seed(123)
training_index_log <- sample(seq_len(nrow(markclean)),
                         size = 0.8 * nrow(markclean))
set_training_log <- markclean[training_index,]
set_testing_log <- markclean[-training_index,]
model_log <- glm(Response ~ NumDealsPurchases + NumWebVisitsMonth + AcceptedCmp1 +
                  AcceptedCmp2 + AcceptedCmp3 + AcceptedCmp4 + AcceptedCmp5,
                data = set_training_log, family = binomial)
summary(model_log)
```

```
##
## Call:
## glm(formula = Response ~ NumDealsPurchases + NumWebVisitsMonth +
##     AcceptedCmp1 + AcceptedCmp2 + AcceptedCmp3 + AcceptedCmp4 +
##     AcceptedCmp5, family = binomial, data = set_training_log)
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -3.10032    0.21498 -14.422  < 2e-16 ***
## NumDealsPurchases  0.10078    0.03780   2.666  0.00767 **
## NumWebVisitsMonth  0.07825    0.03523   2.221  0.02634 *
## AcceptedCmp1       1.58969    0.25265   6.292 3.13e-10 ***
## AcceptedCmp2       1.00192    0.57979   1.728  0.08397 .
## AcceptedCmp3       1.69404    0.21816   7.765 8.15e-15 ***
## AcceptedCmp4       0.67325    0.25775   2.612  0.00900 **
## AcceptedCmp5       2.22936    0.25760   8.654  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1512.2  on 1791  degrees of freedom
## Residual deviance: 1217.5  on 1784  degrees of freedom
## AIC: 1233.5
##
## Number of Fisher Scoring iterations: 5
```

Per saber quines variables són significatives cal, primer de tot, establir un nivell sé significança de referència. Per tal de continuar treballant amb valors anteriorment vists en aquesta PAC es considerarà un nivell de significança de 0.01. Per tant, amb aquest valor establert es pot afirmar que totes aquelles variables que tinguin un valor p associat major a 0.01 no seran significatives. En conseqüència, en el cas del *summary* del model de regressió logística que es presenta en l'anterior cel · la, les variables significatives seran aquelles que tinguin dos o tres asteriscs, on dos asteriscs és sinònim de "variable significativa" ($p < 0.01$) i tres asteriscs és senyal de "variable molt significativa" ($p < 0.001$). Un cop donada aquesta explicació, s'observa que totes les variables menys *NumWebVisitsMonth* i *AcceptedCmp2* són significatives.

Pel que fa a la qualitat del model, es pot observar que la null deviance és menor a la residual deviance i això és indicatiu que les variables significatives contribueixen a millorar les prediccions del model amb variables si es compara amb un model sense variables.

4.2.2 Matriu de confusió

```
predictions_log <- predict(model_log, newdata = set_testing_log, type = "response")
converted_predictions <- ifelse(predictions_log >= 0.5, 1, 0)
```

```
chooseCRANmirror(ind = 1)
install.packages("caret")
```

```
## Installing package into 'C:/Users/mcr99/AppData/Local/R/win-library/4.3'
## (as 'lib' is unspecified)
```

```
## package 'caret' successfully unpacked and MD5 sums checked
```

```
## Warning: cannot remove prior installation of package 'caret'
```

```
## Warning in file.copy(savedcopy, lib, recursive = TRUE): problem copying
## C:\Users\mcr99\AppData\Local\R\win-library\4.3\00LOCK\caret\libs\x64\caret.dll
## to C:\Users\mcr99\AppData\Local\R\win-library\4.3\caret\libs\x64\caret.dll:
## Permission denied
```

```
## Warning: restored 'caret'
```

```
##
## The downloaded binary packages are in
##  C:\Users\mcr99\AppData\Local\Temp\RtmpQXGcrT\downloaded_packages
```

S'ha col · lapsat el *output* de la cel · la anterior atesa la gran quantitat de *logs* que s'han generat, però es pot observar en la següent cel · la que el paquet s'ha instal · lat i importat correctament.

```
library(caret)
```

```
## Loading required package: lattice
```

```
confusion_matrix <- confusionMatrix(data = factor(converted_predictions,
                                                  levels = c(0, 1)),
                                    reference = factor(set_testing_log$Response, levels
                                                       = c(0, 1)))
print(confusion_matrix)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 367  53
##          1  15  13
##
##                Accuracy : 0.8482
##                  95% CI : (0.8116, 0.8802)
##     No Information Rate : 0.8527
##     P-Value [Acc > NIR] : 0.6357
##
##                   Kappa : 0.207
##
##  Mcnemar's Test P-Value : 7.226e-06
##
##             Sensitivity : 0.9607
##             Specificity : 0.1970
##          Pos Pred Value : 0.8738
##          Neg Pred Value : 0.4643
```

```
##              Prevalence : 0.8527
##          Detection Rate : 0.8192
##    Detection Prevalence : 0.9375
##       Balanced Accuracy : 0.5789
##
##        'Positive' Class : 0
##
```

Interpretació de la matriu de confusió:

- Quadrant superior esquerra -> Vertaders positius -> Valor predit = valor real = 0
- Quadrant superior dret -> Falsos negatius -> Valor predit = 0 i valor real = 1
- Quadrant inferior esquerra -> Falsos positius -> Valor predit = 1 i valor real = 0
- Quadrant inferior dret -> Vertaders positius -> Valor predit = valor real = 1

S'observa que hi ha una quantitat substancialment superior de valors ben predits que de valors mal predits i això és indicatiu de què el model funciona bé.

4.2.3 Predicció

```r
model_coefs <- coef(model_log)
print(model_coefs)
```

```
##       (Intercept) NumDealsPurchases NumWebVisitsMonth      AcceptedCmp1
##       -3.10031817        0.10078493        0.07824664        1.58968778
##      AcceptedCmp2      AcceptedCmp3      AcceptedCmp4      AcceptedCmp5
##        1.00191939        1.69403799        0.67324875        2.22936326
```

```r
data_calc_prediction <- data.frame(model_coefs[1] +
                                   model_coefs[2] * 5 +
                                   model_coefs[3] * 10 +
                                   model_coefs[4] +
                                   model_coefs[5] +
                                   model_coefs[6] +
                                   model_coefs[7] +
                                   model_coefs[8])
prob_pred <- 1 / (1 + exp(-data_calc_prediction))
cat("Predicció amb càlculs manuals: "
,prob_pred$model_coefs.1....model_coefs.2....5...model_coefs.3....10...model_coefs.4....)
```

```
## Predicció amb càlculs manuals:  0.9953874
```

```r
new_data <- data.frame(NumDealsPurchases = 5, NumWebVisitsMonth = 10,
                       AcceptedCmp1 = 1, AcceptedCmp2 = 1, AcceptedCmp3 = 1,
                       AcceptedCmp4 = 1, AcceptedCmp5 = 1)
probability_predict <- predict(model_log, newdata = new_data, type = "response")
cat("Probabilitat amb predict():",probability_predict)
```

```
## Probabilitat amb predict(): 0.9953874
```

5. ANOVA unifactorial

5.1 Visualització gràfica

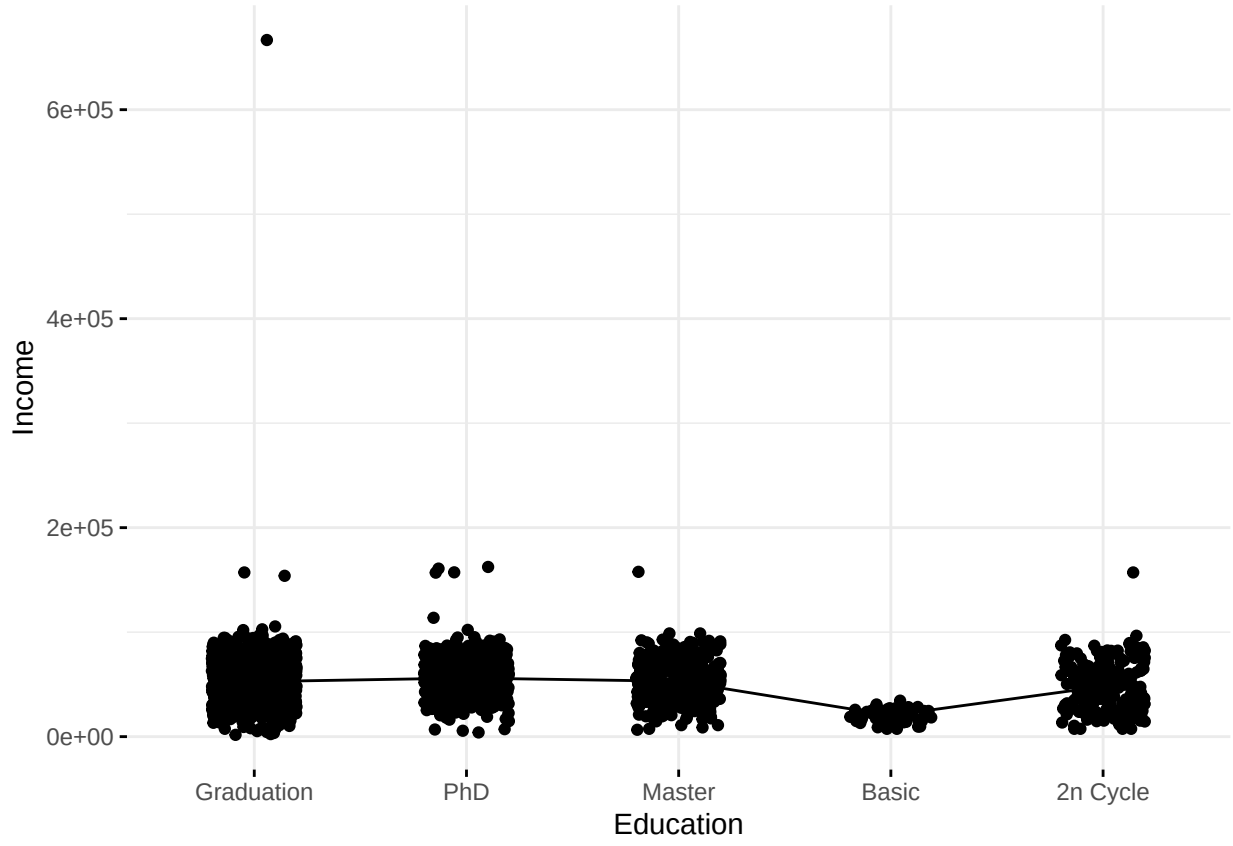
```
install.packages("ggpubr")
```

```
## Installing package into 'C:/Users/mcr99/AppData/Local/R/win-library/4.3'
## (as 'lib' is unspecified)
```

```
## package 'ggpubr' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\mcr99\AppData\Local\Temp\RtmpQXGcrT\downloaded_packages
```

```
library(ggpubr)
```

```
# Gràfic ordenat
ggline(markclean, x = "Education", y = "Income",
     add = c("mean_se", "jitter"),
     ylab = "Income", xlab = "Education",
     ggtheme = theme_minimal())
```



5.2 Hipòtesi nul · la i alternativa

Hipòtesi nul · la: Els ingressos entre els diferents nivells educatius de la població estudiada no tenen diferències significatives.

Hipòtesi alternativa: Els ingressos entre els diferents nivells educatius de la població estudiada sí que tenen

diferències significatives.

5.3 Model

```
anova_model <- aov(Income ~ Education, data = markclean)
summary(anova_model)
```

```
##               Df    Sum Sq   Mean Sq F value Pr(>F)
## Education      4 6.611e+10 1.653e+10   27.24 <2e-16 ***
## Residuals   2235 1.356e+12 6.067e+08
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Per interpretar els resultats del model ANOVA, es continuarà usant un nivell de significança del 0.01. Com s'observa al *summary* del model la variable *Education* té impacte en els ingressos de les persones a causa del fet que el valor p associat a la variable explicativa és menor al nivell de significança (0.01), per tant, es rebutja la hipòtesi nul · la.

5.4 Efectes dels nivells del factor i força de relació

```
anova_components <- anova(anova_model)
ssb <- anova_components$`Sum Sq`[1]
sse <- sum(anova_components$`Sum Sq`[2:length(anova_components$`Sum Sq`)])
sst <- ssb + sse
relation_force <- ssb / sst
cat("SSB: ",ssb,"\n")
```

```
## SSB:  66107929149
```

```
cat("SSE: ",sse,"\n")
```

```
## SSE:  1.355913e+12
```

```
cat("SST: ",sst,"\n")
```

```
## SST:  1.422021e+12
```

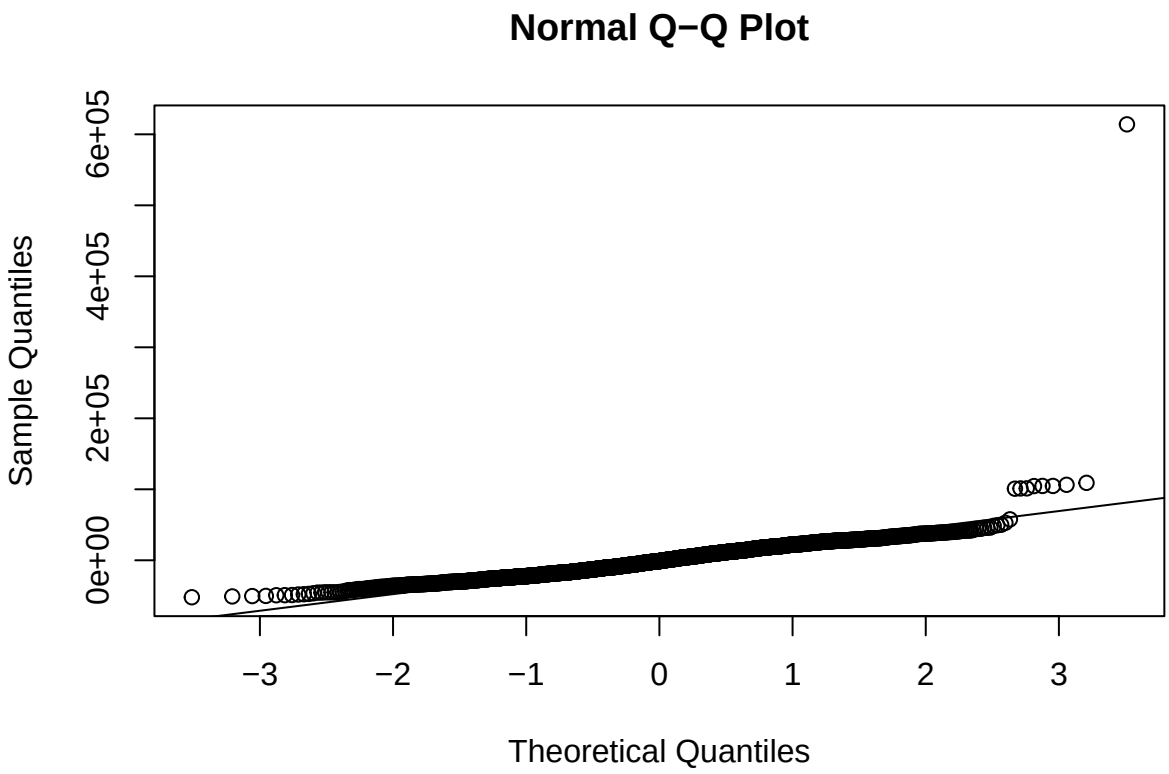
```
cat("Força de relacio: ",relation_force*100,"%")
```

```
## Força de relacio:  4.648871 %
```

S'observa que al voltant del 4.65% de la variabilitat dels ingressos de la població pot ser explicada mitjançant el nivell d'educació. En altres paraules, el nivell d'educació causa un efecte petit a l'hora d'establir el sou d'una persona.

5.5 Normalitat dels residus

```
residuals_model <- residuals(anova_model)
qqnorm(residuals_model)
qqline(residuals_model)
```



En el gràfic QQ es pot observar que els residus del model ANOVA no segueixen una distribució normal atès que les dades no s'ajusten a la línia recta. Es pot observar que a la part dreta del gràfic hi ha alguns valors per sobre (bastant) de la línia igual que a l'esquerra. La part que contra major quantitat de dades, la central, es pot observar que no és una línia recta sinó que té forma de S.

Per dur a terme el test shapiro es plantegen les següents hipòtesis:

Hipòtesi nul · la: Els residus segueixen una distribució normal.

Hipòtesi alternativa: Els residus no segueixen una distribució normal.

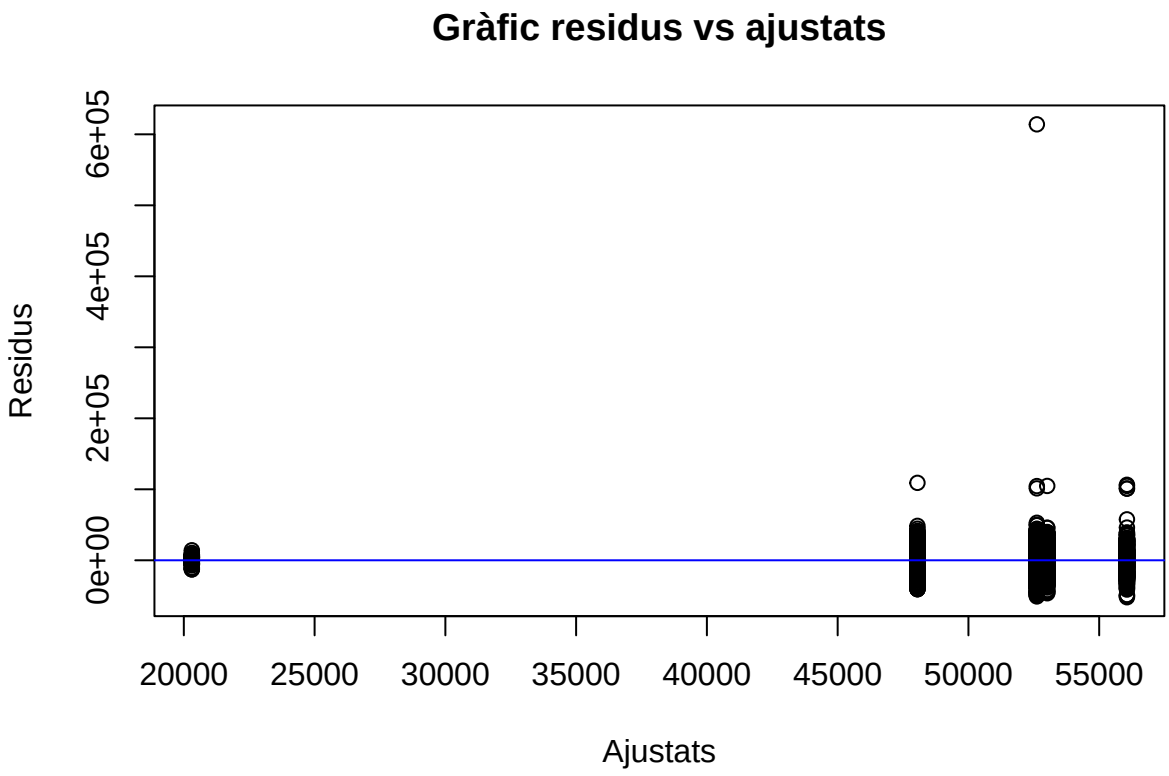
```
shapiro.test(residuals_model)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals_model
## W = 0.76853, p-value < 2.2e-16
```

Prenent un nivell de significança de 0.01, s'observa en el test de Shapiro-Wilk que el valor p està molt sota del nivell de significança. Per tant, es conclou rebutjar la hipòtesi nul · la.

Un cop vist el gràfic QQ i els resultats del test Shapiro-Wilk, es conclou que els residus no segueixen una distribució normal.

```
plot(fitted(anova_model), residuals_model,
     xlab = "Ajustats",
     ylab = "Residus",
     main = "Gràfic residus vs ajustats")
abline(h = 0, col = "blue")
```



Observant el gràfic de residus vs. valors ajustats, s'observa que per a valors petits valors ajustats petits, els residus s'apropen a 0 i per a valors grans els residus estan més dispersos. Aquest augmenta no constant a mesura que els valors dels ajustats creixen es coneixen com a efecte *fanning* i indica la presència de heteroscedasticitat en els residus.

6. Comparacions múltiples

```
bonferroni <- pairwise.t.test(markclean$Income, markclean$Education,
                              p.adjust.method = "bonferroni")
print(bonferroni)
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  markclean$Income and markclean$Education
##
##            2n Cycle Basic   Graduation Master
## Basic      2.6e-12  -       -          -
## Graduation 0.151    < 2e-16 -          -
## Master     0.212    < 2e-16 1.000      -
## PhD        0.001    < 2e-16 0.100      0.730
##
## P value adjustment method: bonferroni
```

Per a fer aquest test de comparació s'estableixen les següents hipòtesis:

Hipòtesi nul · la: No hi ha diferència en els ingressos mitjans quan es comparen ambdós nivells educatius.

Hipòtesi alternativa: Hi ha diferència en els ingressos mitjans quan es comparen ambdós nivells educatius.

Prenent com a nivell de significança 0.01 s'observa que per les comparacions: *Basic - 2n Cycle*, *PhD - 2n Cycle*, *Graduation - Basic*, *Master - Basic* i *PhD - Basic* es rebutja la hipòtesi nul · la i, per tant, es conclou que sí que hi ha diferència en la comparació de sous dels nivells educatius.

En les comparacions: *Graduation - 2n Cycle*, *Master - 2n Cycle*, *Master - Graduation*, *PhD - Graduation* i *PhD - Master* no es rebutja la hipòtesi nul · la i, en conseqüència, es conclou que no hi ha diferència en la comparació de sous dels nivells educatius.

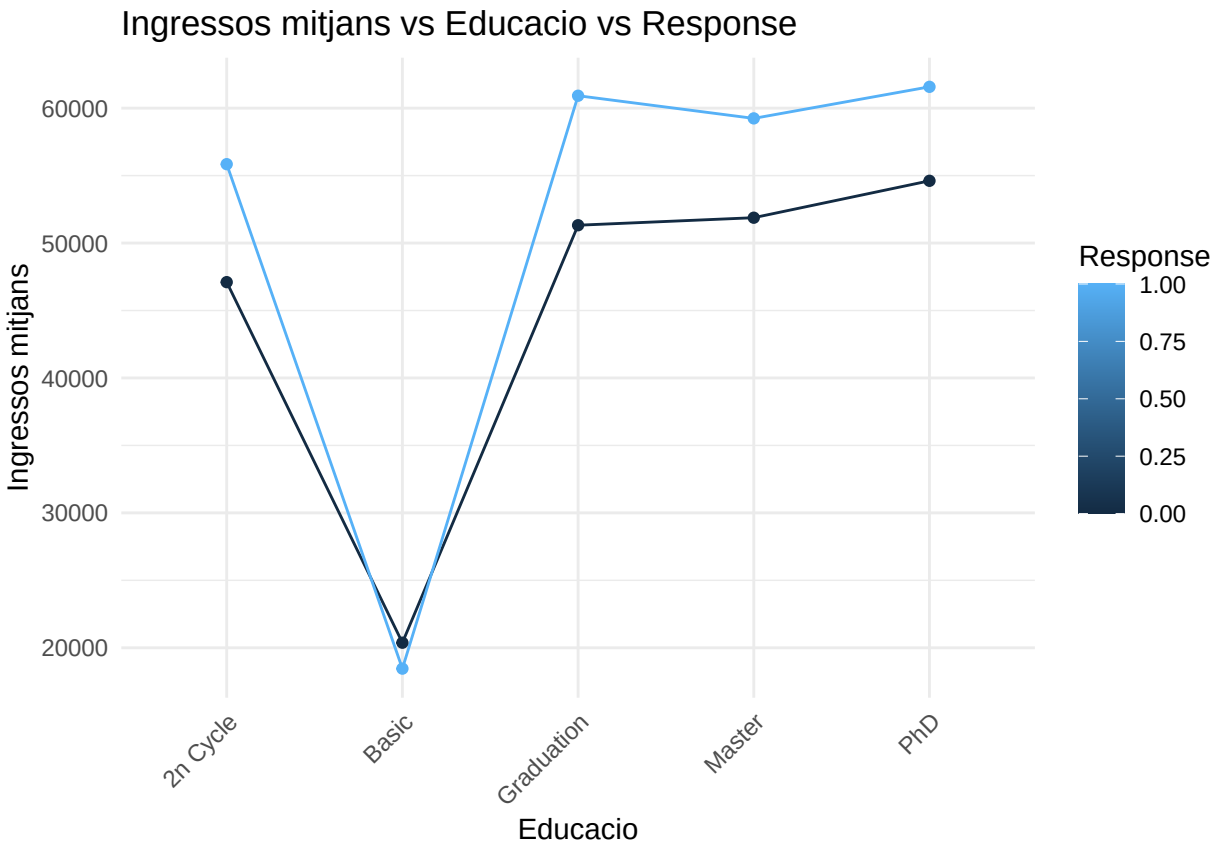
7. ANOVA multifactorial

7.1 Anàlisi visual dels efectes principals i possibles interaccions

```
data_prep <- markclean %>%
  group_by(Education, Response) %>%
  summarise(mean_income = mean(Income), .groups = 'drop')
# .groups = 'drop' -> Per evitar que el dataset resultat estigui agrupat
# .groups s'ha de posar sinó tira warning
print(data_prep)
```

```
## # A tibble: 10 x 3
##    Education  Response mean_income
##    <chr>         <int>       <dbl>
##  1 2n Cycle          0      47103.
##  2 2n Cycle          1      55849.
##  3 Basic             0      20377.
##  4 Basic             1      18456
##  5 Graduation        0      51322.
##  6 Graduation        1      60920.
##  7 Master            0      51880.
##  8 Master            1      59241.
##  9 PhD               0      54614.
## 10 PhD               1      61581.
```

```
ggplot(data_prep, aes(x = Education, y = mean_income,
                      color = Response, group = Response)) +
  geom_line() + geom_point() + theme_minimal() +
  labs(title = "Ingressos mitjans vs Educacio vs Response",
       x = "Educacio", y = "Ingressos mitjans", color = "Response") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Es pot observar que entre els nivells educatius *Master* i *PhD* les línies són paral · leles. Aquest fet indica que no hi ha una interacció significativa entre l'educació i *Response*. Per la resta de línies (entre nivells educatius), s'observa que, o bé es creuen, o bé es creuaran si les allarguem (si allarguem les línies entre *Graduation* i *Master* s'acabaran creuant). El fet que les línies no siguin paral · leles, indica que, al contrari que en el cas anterior, sí que hi ha una interacció significativa.

7.2 Càlcul del model

```
anova_model_2 <- aov(Income ~ Education * Response, data = markclean)
```

7.3 Interpretació de resultats

```
summary(anova_model_2)
```

```
##                      Df    Sum Sq   Mean Sq F value  Pr(>F)
## Education             4 6.611e+10 1.653e+10  27.590 < 2e-16 ***
## Response              1 1.951e+10 1.951e+10  32.568 1.3e-08 ***
## Education:Response    4 6.113e+08 1.528e+08   0.255   0.907
## Residuals          2230 1.336e+12 5.990e+08
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

En el *summary* del model s'observa que, si es pren un nivell de significança de 0.01, les variables *Education* i *Response* tenen un efecte sobre els ingressos. És a dir, el fet que aquestes variables canvin de valor, afectarà els ingressos. Per contra, el p-value entre l'educació i la *Response* és elevat (> 0.01), per tant, es pot concloure que l'impacte de l'educació sobre els ingressos dels individus és el mateix tant si s'ha acceptat, com si no,

l'última campanya.

8. Resum executiu

Després d'analitzar ek conjunt de dades s'observa que el nivell mitjà d'ingressos de la població, de la qual s'han recollit dades, és un valor que representa bastant bé com es distribueixen els ingressos.

També s'ha pogut observar i concloure amb un 99% de confiança en els resultats que les persones que no estàn en possessió d'un títol universitari ingressen, de mitjana, menys diners que aquelles persones que sí que tenen un títol universitari.

Per comprendre quins factors poden explicar el nivell d'ingressos de les persones s'ha creat una eina que treballa cercant patrons (model de regressió lineal). Aquest model, ara per ara, és capaç d'explicar una mica més del 50% de les variacions dels ingressos en funció de l'any de naixement, el nombre de nens que viuen amb el client, el nombre d'adolescents que viuen amb el client, el nivell d'estudis del client, la quantitat de diners que s'ha gastat el client (en els últims dos anys) en vi, fruita, carn, peix, dolços i productes gold. També es tenen en compte el nombre de compres fetes ambdescompte, el nombre de compres festes usant el catàleg i el nombre de compres fetes a les botigues físiques. Finalment, també es té en compte el nombre de visites que el client ha fet a la web durant l'últim mes. Un cop creada aquesta eina, s'ha pogut veure que tot el que es té en compte per explicar la variació dels ingressos, realment sí que té un impacte en el nivell d'ingressos.

Seguidament, s'ha creat una segona eina per a predir la probabilitat (model de regressió logística) que un client accepti l'oferta en la sisena campanya. Per calcular aquesta probabilitat, l'eina té en compte el nombre de compres efectuades amb descompte els últims 2 anys, el nombre de visites a la web l'últim mes i el acceptar, o no, alguna campanya anterior. Després de la realització de diversos càlculs s'ha pogut observar que aquesta segona eina prediu probabilitats amb una precisió del 84%. En termes numèrics, d'un total de 448 dades que es disposaven per fer proves, l'eina n'ha predit correctament al voltant de 380.

Mitjançant un test anomenat ANOVA s'han comparat els ingressos per als diferents nivells educatius i s'ha pogut observar que el nivell educatiu del client té un impacte en els seus ingressos. Concretament, l'educació explica al voltant del 4.65% de la variabilitat dels ingressos.

Finalment, s'ha creat una eina per avaluar l'efecte que tenen l'educació i acceptar, o no, l'última oferta sobre els ingressos i s'ha pogut observar que l'educació i acceptar, o no, l'última oferta afecta als ingressos, però que l'efecte que causa el nivell d'educació del client sobre els ingressos es similar tant si s'accepta com si no l'oferta.