Proyecto FMAD

ICAI. Máster en Big Data. Fundamentos Matemáticos del Análisis de Datos (FMAD).

Álvaro Rodríguez González, Ignacio Perez-Cea, Pablo Sanz Caperote

Curso 2021-22. Última actualización: 2021-12-03

Índice

1. Introducción **3**

2. Definición de las variables **4**

3. Preprocesamiento **5**

3.1. Resumen de datos 5

3.2. Análisis de las variables 7

3.3. Visualización de los datos 9

4. Análisis Gráfico **13**

4.1. Relacción nº de hijos, gasto y estudios. 13

5. Análisis Predictivo y Analítica Avanzada **23**

5.1. Ajuste parámetros de control 23

5.2. Análisis sobre variable Complain 23

1. Introducción

El principal objetivo de este proyecto es plasmar los conocimientos adquiridos durante la primera parte del curso en la asignatura de Fundamentos Matemáticos del Análisis de Datos. Para ello trabajaremos sobre un dataset de campañas de marketing en EEUU.

Nuestra idea es primero realizar un breve estudio de nuestras variables, donde realizaremos cambios en caso de considerarlo oportuno (usando el paquete tidyverse), este estudio será tanto gráfico como no gráfico.

Una vez que hemos realizado la limpieza de los datos, realizaremos un análisis gráfico sobre el comportamiento de las variables así como también sobre alguna posible relación que nos resulte interesante.

Para terminar el proyecto aplicaremos técnicas de Machine Learning que hemos aprendido durante la segunda parte del cuatrimestre.

Antes de iniciar el breve estudio, lo que haremos será cargar las diferentes librerias que usaremos para nuestro proyecto. Entre ellas encontraremos librerias públicas como tidyverse y algunas privadas como MLTools:

```
library(tidyverse)
library(lubridate)
library(caret)
library(grid)
library(corrplot)
library(gridExtra)
library(ROCR)
library(MLTools)
library(GGally)
library(rpart)
library(rpart.plot)
library(partykit)
library(kernlab)
library(NeuralNetTools)
library(NeuralSens)
library(nnet)
library(ROSE)
library(randomForest)
```

A su vez tambien leeremos los datos con los que trabajaremos:

```
datos <- read.csv("marketing_campaign.csv", header = TRUE, sep = "")
```

2. Definición de las variables

Antes de comenzar con el preprocesamiento de los datos lo que haremos será listar las variables y lo que representa cada una de ellas:

- **ID**: El ID del cliente.
- **Year_Birth:** Indica el año de nacimiento del cliente.
- **Education:** Indica el nivel de educación del cliente.
- **Marital_Status:** Indica el estado civil del cliente.
- **Income:** Presenta el ingreso familiar anual del cliente.
- **Kidhome:** Indica el número de niños pequeños en casa del cliente.
- **Teenhome:** Indica el número de adolescentes en el hogar del cliente.
- **Dt_Customer:** Muestra la fecha de inscripción del cliente en la empresa.
- **Recency:** El número de días desde la última compra.
- **MntWines:** El gasto en productos vitivinícolas en los últimos 2 años.
- **MntGoldProds:** El gasto en productos premium en los últimos 2 años.
- **NumDealsPurchases:** El número de compras con uso de descuento.
- **NumWebPurchases:** El número de compras a través de la web.
- **NumCatalogPurchases:** El número de compras usando catalogo.
- **NumWebVisitsMonth:** El número de visitas por mes a la web.
- **AcceptedCmp1:** 1 si el cliente acepta la oferta en la 1ra campaña, 0 si no lo acepta.
- **AcceptedCmp2:** 1 si el cliente acepta la oferta en la 2nd campaña, 0 si no lo acepta.
- **Complain:** 1 si el cliente se ha quejado en los dos últimos años.
- **Z_CostContact:** El coste de contactar con cliente.
- **Z_Revenue:** Los ingresos/beneficios después de que el cliente acepte la campaña.
- **Response:** 1 si el cliente acepta la oferta en la última campaña y 0 si no la acepta.

3. Preprocesamiento

3.1. Resumen de datos

Lo primero que haremos será ver como esta estructurado nuestro dataset. Para ello veremos que tamaño tiene, tanto filas como columnas. A su vez también veremos con que tipo de datos estamos trabajando.

```
cat(cat(cat(cat("El conjunto de datos tiene", nrow(datos)), "filas y"),
        ncol(datos)), "columnas")
```

```
## El conjunto de datos tiene 2440 filas y 29 columnas
```

```
str(datos)
```

```
## 'data.frame':    2440 obs. of  29 variables:
##  $ ID                 : int  5524 2174 4141 6182 5324 7446 965 6177 4855 5899 ...
##  $ Year_Birth         : int  1957 1954 1965 1984 1981 1967 1971 1985 1974 1950 ...
##  $ Education          : chr  "Graduation" "Graduation" "Graduation" "Graduation" ...
##  $ Marital_Status     : chr  "Single" "Single" "Together" "Together" ...
##  $ Income             : chr  "58138" "46344" "71613" "26646" ...
##  $ Kidhome            : int  0 1 0 1 1 0 0 1 1 1 ...
##  $ Teenhome           : chr  "0" "1" "0" "0" ...
##  $ Dt_Customer        : chr  "04-09-2012" "08-03-2014" "21-08-2013" "10-02-2014" ...
##  $ Recency            : chr  "58" "38" "26" "26" ...
##  $ MntWines           : int  635 11 426 11 173 520 235 76 14 28 ...
##  $ MntFruits          : int  88 1 49 4 43 42 65 10 0 0 ...
##  $ MntMeatProducts    : int  546 6 127 20 118 98 164 56 24 6 ...
##  $ MntFishProducts    : int  172 2 111 10 46 0 50 3 3 1 ...
##  $ MntSweetProducts   : int  88 1 21 3 27 42 49 1 3 1 ...
##  $ MntGoldProds       : int  88 6 42 5 15 14 27 23 2 13 ...
##  $ NumDealsPurchases  : int  3 2 1 2 5 2 4 2 1 1 ...
##  $ NumWebPurchases    : int  8 1 8 2 5 6 7 4 3 1 ...
##  $ NumCatalogPurchases: int  10 1 2 0 3 4 3 0 0 0 ...
##  $ NumStorePurchases  : int  4 2 10 4 6 10 7 4 2 0 ...
##  $ NumWebVisitsMonth  : int  7 5 4 6 5 6 6 8 9 20 ...
##  $ AcceptedCmp3       : int  0 0 0 0 0 0 0 0 0 1 ...
##  $ AcceptedCmp4       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AcceptedCmp5       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AcceptedCmp1       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AcceptedCmp2       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Complain           : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Z_CostContact      : int  3 3 3 3 3 3 3 3 3 3 ...
##  $ Z_Revenue          : int  11 11 11 11 11 11 11 11 11 11 ...
##  $ Response           : int  1 0 0 0 0 0 0 0 1 0 ...
```

Una vez visto el tipo de variables con las que trabajamos es facilmente observable la necesidad de realizar algunas modificaciones en algunas de ellas.

Ahora veremos un resumen de las variables que tenemos:

```
summary(datos)
```

```
##        ID          Year_Birth    Education         Marital_Status
##  Min.   :    0   Min.   :1893   Length:2440        Length:2440
##  1st Qu.: 2108   1st Qu.:1959   Class :character   Class :character
##  Median : 5048   Median :1970   Mode  :character   Mode  :character
##  Mean   : 5134   Mean   :1969
##  3rd Qu.: 8147   3rd Qu.:1977
##  Max.   :11191   Max.   :1996
##                  NA's   :200
##     Income             Kidhome        Teenhome         Dt_Customer
##  Length:2440        Min.   :    0   Length:2440        Length:2440
##  Class :character   1st Qu.:    0   Class :character   Class :character
##  Mode  :character   Median :    0   Mode  :character   Mode  :character
##                     Mean   : 4253
##                     3rd Qu.:    1
##                     Max.   :96547
##                     NA's   :200
##    Recency             MntWines        MntFruits       MntMeatProducts
##  Length:2440        Min.   :   0.0   Min.   :   0.00   Min.   :   0.0
##  Class :character   1st Qu.:  25.0   1st Qu.:   2.00   1st Qu.:  14.0
##  Mode  :character   Median : 138.0   Median :   9.00   Median :  57.0
##                     Mean   : 288.5   Mean   :  42.39   Mean   : 156.8
##                     3rd Qu.: 476.5   3rd Qu.:  38.00   3rd Qu.: 211.5
##                     Max.   :1493.0   Max.   :1215.00   Max.   :1725.0
##                     NA's   :200      NA's   :200       NA's   :200
##  MntFishProducts  MntSweetProducts  MntGoldProds    NumDealsPurchases
##  Min.   :  0.00   Min.   :  0.00   Min.   :  0.00   Min.   :  0.000
##  1st Qu.:  3.00   1st Qu.:  1.00   1st Qu.:  7.00   1st Qu.:  1.000
##  Median : 13.00   Median :  9.00   Median : 22.00   Median :  2.000
##  Mean   : 45.34   Mean   : 28.44   Mean   : 42.45   Mean   :  6.326
##  3rd Qu.: 55.00   3rd Qu.: 35.00   3rd Qu.: 54.00   3rd Qu.:  4.000
##  Max.   :974.00   Max.   :362.00   Max.   :321.00   Max.   :246.000
##  NA's   :200      NA's   :200      NA's   :200      NA's   :200
##  NumWebPurchases  NumCatalogPurchases NumStorePurchases NumWebVisitsMonth
##  Min.   : 0.000   Min.   : 0.000      Min.   : 0.000    Min.   : 0.000
##  1st Qu.: 2.000   1st Qu.: 1.000      1st Qu.: 3.000    1st Qu.: 3.000
##  Median : 3.000   Median : 2.000      Median : 5.000    Median : 6.000
##  Mean   : 3.929   Mean   : 2.817      Mean   : 5.503    Mean   : 5.278
##  3rd Qu.: 6.000   3rd Qu.: 4.000      3rd Qu.: 8.000    3rd Qu.: 7.000
##  Max.   :27.000   Max.   :28.000      Max.   :13.000    Max.   :20.000
##  NA's   :200      NA's   :200         NA's   :200       NA's   :200
```

```
##   AcceptedCmp3      AcceptedCmp4       AcceptedCmp5       AcceptedCmp1
##  Min.   :0.0000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000
##  Median :0.0000   Median :0.00000   Median :0.00000   Median :0.00000
##  Mean   :0.5545   Mean   :0.07679   Mean   :0.07277   Mean   :0.06161
##  3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :9.0000   Max.   :1.00000   Max.   :1.00000   Max.   :1.00000
##  NA's   :200      NA's   :200       NA's   :200       NA's   :200
##   AcceptedCmp2         Complain       Z_CostContact      Z_Revenue
##  Min.   :0.00000   Min.   :0.00000   Min.   : 0.000   Min.   : 0.00
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.: 3.000   1st Qu.:11.00
##  Median :0.00000   Median :0.00000   Median : 3.000   Median :11.00
##  Mean   :0.01875   Mean   :0.03661   Mean   : 2.809   Mean   :10.18
##  3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.: 3.000   3rd Qu.:11.00
##  Max.   :1.00000   Max.   :3.00000   Max.   :11.000   Max.   :11.00
##  NA's   :200       NA's   :200       NA's   :200      NA's   :200
##     Response
##  Min.   : 0.000
##  1st Qu.: 0.000
##  Median : 0.000
##  Mean   : 1.132
##  3rd Qu.: 0.000
##  Max.   :11.000
##  NA's   :221
```

Observamos que existen numerosos valores nulos en nuestras variables. En el punto siguiente veremos que hacer con estos casos.

3.2. Análisis de las variables

Lo primero que haremos será eliminar las filas que contienen datos nulos. Esto podemos hacerlo ya que disponemos de una muestra muy grande y eliminar los valores nulos no afectará para nuestro trabajo.

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

Además, también observamos que hay algunos datos erróneos por lo que por el mismo motivo que antes procederemos a eliminarlos.

```
datos <- datos %>%
  filter(ID != 0 & ID != 1 & Education != "2n" & Income > 10 & Income != "2")
```

Además el conjunto de datos tiene muchas columnas las cuales no nos resultan interesantes, por ello vamos a eliminar algunas de ellas: “NumDealsPurchases”, “Receny”, “AcceptedCmp1”, “AcceptedCmp2”, “AcceptedCmp3”, “AcceptedCmp4” y “AcceptedCmp5”, “Z_CostContact” y “Z_Revenue”.

```
datos <- datos %>%
  select(-c(AcceptedCmp3:AcceptedCmp2), -NumDealsPurchases, -Recency,
```

```
         -Z_CostContact, -Z_Revenue)
```

También como vimos cuando hicimos la visión general de las variables y su tipo, nos dimos cuenta de que algunas de ellas estaban mal tipadas. Por ello cambiaremos el tipado de algunas columnas.

```
datos$Teenhome <- as.numeric(datos$Teenhome)
datos$Income <- as.numeric(datos$Income)
datos$Complain <- as.factor(datos$Complain)
datos$Education <- as.factor(datos$Education)
datos$Response <- as.factor(datos$Response)
```

A su vez hemos observado que algunas columnas podrían tener un formato más útil o sencillo, como es el caso del año de nacimiento, donde es mas cómodo trabajar con edades. Por tanto, para un mejor procesamiento y una mayor útilidad realizaremos un mutate para generar una nueva columna formada por la edad de los clientes. A su vez eliminaremos la columna de año de nacimiento.

```
datos <- datos %>%
  mutate(edad = 2021 - Year_Birth) %>%
  select(-Year_Birth)
```

También nos pareció interesante en vez de distinguir entre número de hijos los cuales son pequeños o son adolescentes, tomarlos como una única variable que nos indique el número de hijos que hay en cada hogar. Para ello sumaremos el total de niños de cada cliente agrupando las columnas Kidhome y Teenhome.

```
datos <- datos %>%
  mutate(totalHijos = Kidhome + Teenhome) %>%
  select(-Kidhome, -Teenhome)
```

Como en el caso de los hijos para las compras haremos algo similar, donde cogeremos las columnas "NumWebPurchases", "NumCatalogPurchases" y "NumStorePurchases" que indican el número de compras hechas en cada sitio, en tiendas, por catalogo y por la web y las sumaremos todas en una única columna que indique el total de compras que ha realizado el cliente.

```
datos <- datos %>%
  rowwise(ID) %>%
  mutate(suma_compras = sum(c(NumWebPurchases, NumCatalogPurchases,
                              NumStorePurchases))) %>%
  select(-c(NumWebPurchases, NumCatalogPurchases, NumStorePurchases))
```

A su vez, para el gasto en los diferentes tipos de producto sumaremos las columans: "MntWines", "MntFruits", "MntMeatProducts", "MntFishProducts", "MntSweetProducts", "MntGoldProds" lo cual nos indicara cuánto dinero se ha gastado un cliente en total.

```
datos <- datos %>%
  rowwise(ID) %>%
  mutate(Dinero_Gastado = sum(c(MntWines, MntFruits, MntMeatProducts, MntFishProducts,
                                MntSweetProducts, MntGoldProds)))
```

También existe una variable que nos indica el estado civil del cliente, al existir numerosas situaciones nosotros agruparemos el estado civil de cada cliente y lo simplificamos para ver si vive solo o en pareja. Ya que esto nos podrá resultar interesante para análisis posteriores.

```
datos <- datos %>%
  mutate(Marital_Status = factor(Marital_Status== "Single" | Marital_Status== "Divorced",
                           levels = c(TRUE, FALSE), labels = c('Single','Not single')))
```

Por último cambiaremos la columna Dt_Customer y estableceremos 3 grupos que representan la longevidad del cliente en la empresa.

```
datos$Dt_Customer = as.Date(datos$Dt_Customer, "%d-%m-%Y")
fechaMinima = min(datos$Dt_Customer)
datos$Dt_Customer <-  factor(cut_number(as.duration(datos$Dt_Customer-fechaMinima),
                      n = 3), labels = c("Nuevo", "Con Experiencia", "Muy Antiguo"),
                      ordered = TRUE)
```

3.3. Visualización de los datos

En este apartado lo que realizaremos será un análisis mediante gráficas de las variables según su tipo, viendo si siguen una distribución normal y si presentarían outliers entre otros factores.

Comenzaremos con las variables continuas, que son las siguientes:

```
h1 <- ggplot(datos)+
  geom_histogram(aes(x = Income), color = "black", alpha = 0.35)
h2 <- ggplot(datos)+
  geom_histogram(aes(x = MntWines ), color = "black", alpha = 0.35)
h3 <- ggplot(datos)+
  geom_histogram(aes(x = MntFruits ), color = "black", alpha = 0.35)
h4 <- ggplot(datos)+
  geom_histogram(aes(x = MntMeatProducts ), color = "black", alpha = 0.35)
h5 <- ggplot(datos)+
  geom_histogram(aes(x = MntFishProducts ), color = "black", alpha = 0.35)
h6 <- ggplot(datos)+
  geom_histogram(aes(x = MntSweetProducts ), color = "black", alpha = 0.35)
h7 <- ggplot(datos) +
  geom_histogram(aes(x = MntGoldProds ), color = "black", alpha = 0.35)
h8 <- ggplot(datos) +
  geom_histogram(aes(x = Dinero_Gastado), color = "black", alpha = 0.35)
h9 <- ggplot(datos)+
  geom_histogram(aes(x = suma_compras ), color = "black", alpha = 0.35)
h10 <- ggplot(datos)+
  geom_histogram(aes(x = NumWebVisitsMonth ), color = "black", alpha = 0.35)
h12 <- ggplot(datos)+
  geom_histogram(aes(x = edad ), color = "black", alpha = 0.35)
```

```
grid.arrange(h1,h2,h3,h4,h5,h6,h7,h8,h9,h10,h12)
```

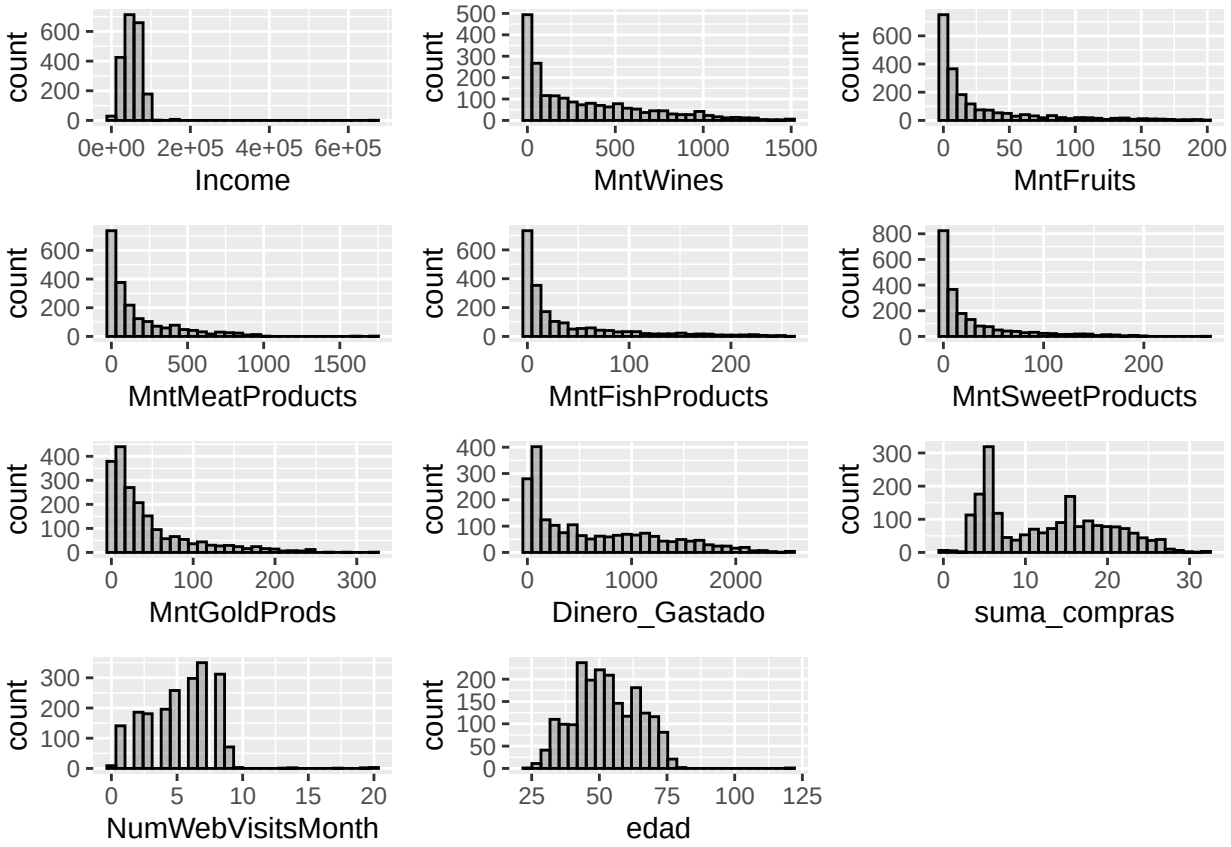

Figura 1: Gráfico variables continuas

Podemos observar en los histogramas que ninguna de las variables sigue una distribución normal. El patrón más común es un gran número de datos con valores pequeños y muchos menos datos a medida que el valor de la variable del eje x aumenta. Por este motivo, aunque no los mostremos, podemos deducir que existen outliers en casi todas las variables. Un caso bastante claro de outlier se puede ver en la variable edad donde vemos que el eje x llega a 125 lo cual nos indica que debe existir alguna observación con valor por encima de 115 y menor de 125.

En cuanto a las variables discretas tenemos lo siguiente:

```
hb1 <- ggplot(datos)+
  geom_bar(aes(x = Education), color = "black", alpha = 0.35)
hb2 <- ggplot(datos) +
  geom_bar(aes(x = Marital_Status), color = "black", alpha = 0.35)
hb3 <- ggplot(datos) +
  geom_bar(aes(x = Dt_Customer), color = "black", alpha = 0.35)
hb4 <- ggplot(datos) +
  geom_bar(aes(x = Complain), color = "black", alpha = 0.35)
grid.arrange(hb1, hb2, hb3, hb4, ncol = 2)
```

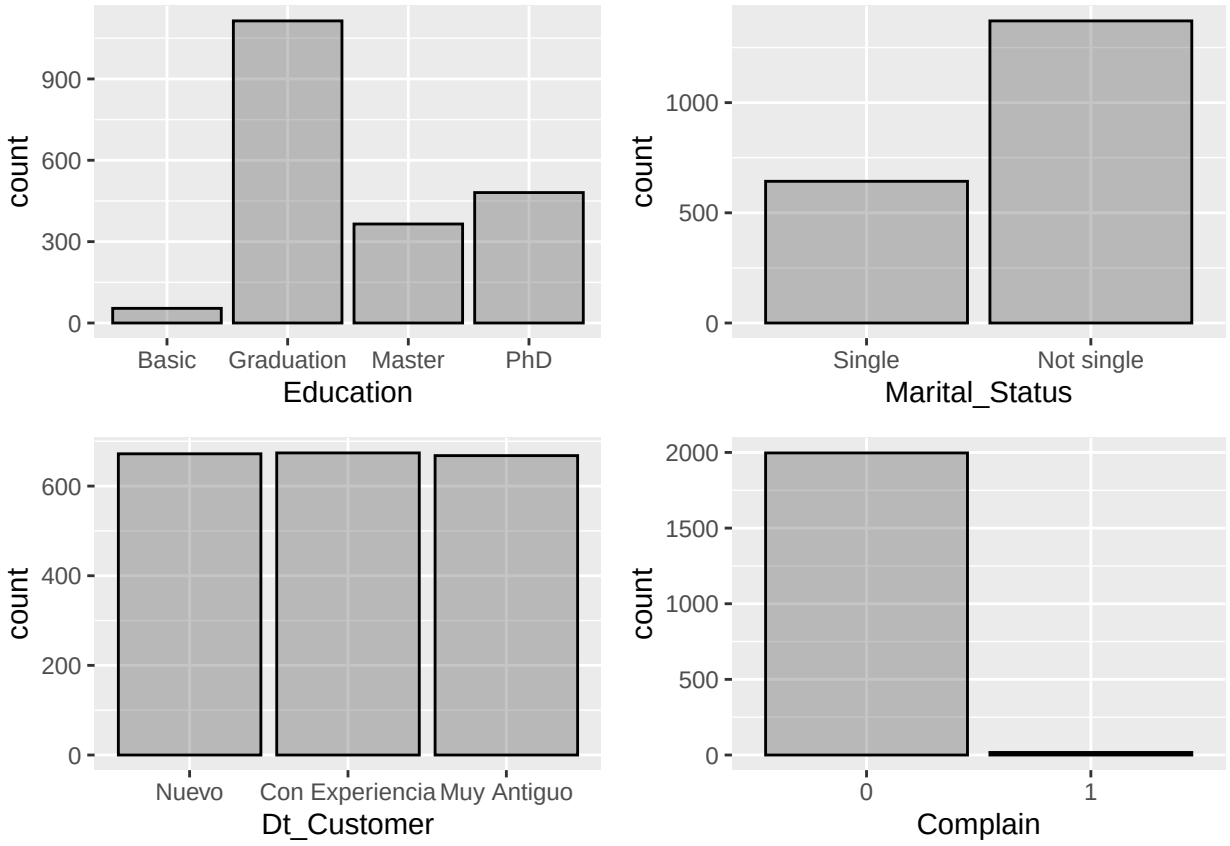

Figura 2: Gráfico variables discretas

En estas variables podemos observar diferentes patrones. Si miramos la educación de los clientes nos damos cuenta que hay mucha diferencia entre el número de clientes que tienen una educación básica y el resto de tipos, sobre todo los clientes graduados. De la misma forma vemos que hay más del doble de clientes not single que single. Por otro lado tenemos que la longividad de los clientes es uniforme. En cuanto a la variable complain observamos que unicamente un porcentaje muy pequeño de los clientes se han quejado durante los dos últimos años.

4. Análisis Gráfico

En esta sección nuestro objetivo será sacar conclusiones acerca de diversas variables así como de posibles relaciones entre estas. Nuestros principales apoyos para sacar dichas conclusiones serán los elementos gráficos (histogramas, boxplot, tablas, etc).

4.1. Relacción nº de hijos, gasto y estudios.

Para ello crearemos un nuevo conjunto de datos sacados de los datos que hemos refinado en la parte anterior. A su vez en un primer momento hemos analizado la relación entre tener hijos y los gastos en compras de cada tipo.

```
datos_ML<-datos %>%
  mutate(sonPadres = factor(totalHijos>0, levels = c(FALSE, TRUE), labels = c(0,1)))
borrar <- c("MntWines", "MntFruits", "MntMeatProducts",
            "MntFishProducts", "MntSweetProducts",
            "MntGoldProds","Dinero_Gastado", "sonPadres", "totalHijos")
datos_ML <- datos_ML[(names(datos_ML) %in% borrar)]
head(datos_ML)
```

```
## # A tibble: 6 x 9
## # Rowwise:
##   MntWines MntFruits MntMeatProducts MntFishProducts MntSweetProducts
##      <int>     <int>           <int>           <int>             <int>
## 1      635        88             546             172                88
## 2       11         1               6               2                 1
## 3      426        49             127             111                21
## 4       11         4              20              10                 3
## 5      173        43             118              46                27
## 6      520        42              98               0                42
## # ... with 4 more variables: MntGoldProds <int>, totalHijos <dbl>,
## #   Dinero_Gastado <int>, sonPadres <fct>
```

Lo primero que haremos sera ver como se relacionan las diferentes variables con las variables de output, empezando por si tienen hijos.

```
g1 <-ggplot(datos_ML) +
  geom_boxplot(aes(x = sonPadres, y = MntWines))
g2<-ggplot(datos_ML) +
  geom_boxplot(aes(x = sonPadres, y = MntFruits))
g3<-ggplot(datos_ML) +
  geom_boxplot(aes(x = sonPadres, y = MntMeatProducts))
g4<-ggplot(datos_ML) +
  geom_boxplot(aes(x = sonPadres, y = MntFishProducts))
g5<-ggplot(datos_ML) +
  geom_boxplot(aes(x = sonPadres, y = MntSweetProducts))
```

```
g6<-ggplot(datos_ML) +
  geom_boxplot(aes(x = sonPadres, y = MntGoldProds))
g7<-ggplot(datos_ML) +
  geom_boxplot(aes(x = sonPadres, y = Dinero_Gastado))
gridExtra::grid.arrange(g1,g2,g3,g4, g5, g6, g7)
```

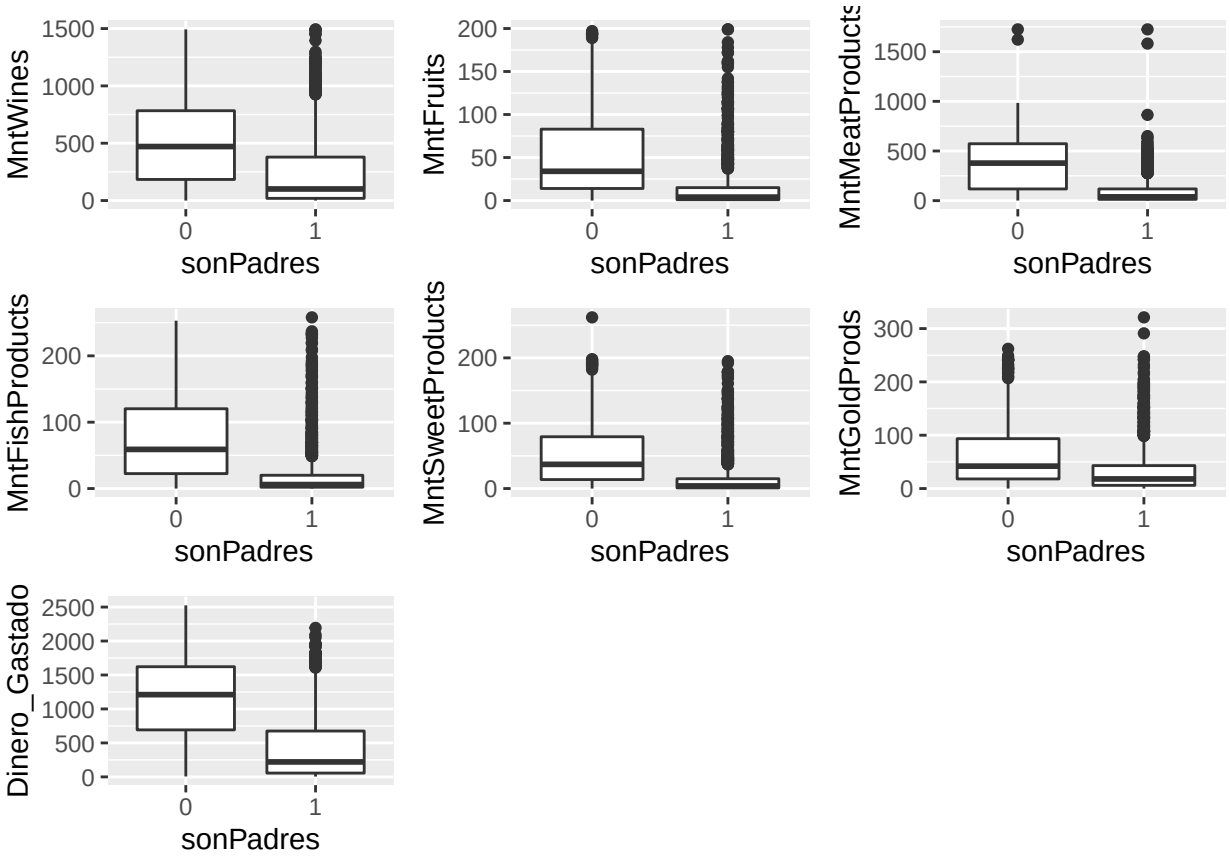

Figura 3: Gráfico relaciones variables

Vemos que si los clientes tienen hijos, el gasto en todos los productos se reduce. Pero ya que hemos llegado hasta aquí, queremos además observar la relación entre el número de hijos y dichos gastos. Veamos los siguientes gráficos.

```
gb1 <-ggplot(datos_ML) +
  geom_boxplot(aes(x = factor(totalHijos), y = MntWines))
gb2<-ggplot(datos_ML) +
  geom_boxplot(aes(x = factor(totalHijos), y = MntFruits))
gb3<-ggplot(datos_ML) +
  geom_boxplot(aes(x = factor(totalHijos), y = MntMeatProducts))
gb4<-ggplot(datos_ML) +
  geom_boxplot(aes(x = factor(totalHijos), y = MntFishProducts))
gb5<-ggplot(datos_ML) +
  geom_boxplot(aes(x = factor(totalHijos), y = MntSweetProducts))
```

```
gb6<-ggplot(datos_ML) +
  geom_boxplot(aes(x = factor(totalHijos), y = MntGoldProds))
gb7<-ggplot(datos_ML) +
  geom_boxplot(aes(x = factor(totalHijos), y = Dinero_Gastado))
gridExtra::grid.arrange(gb1,gb2,gb3,gb4, gb5, gb6, gb7)
```

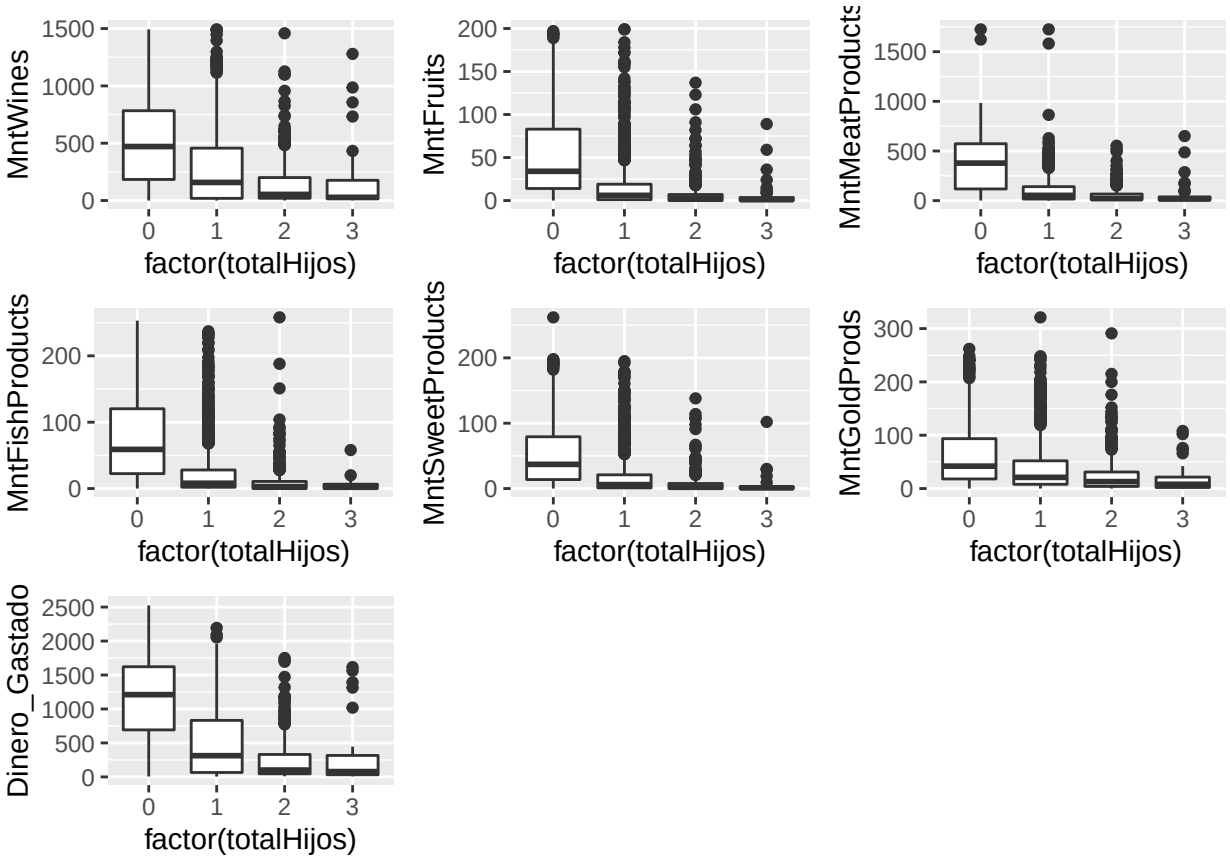

Figura 4: Gráfico relación nº hijos y gasto

Siguiendo con la relación anterior, observamos que a más hijos menor es el gasto en todos los productos.

Veamos la relación entre la edad y los gastos de compras. Primero vamos a ver la distribución de la edad.

```
ggplot(datos) +
  geom_histogram(aes(x = edad, y =stat(density)), bins = 15, fill = "darkgreen",
                 color = "black") +
  geom_density(aes(x = edad), color="red", size=1.5)
```

Como hemos comentado anteriormente existe un valor atípico dentro de nuestra variable, por lo que al unicamente ser uno lo eliminaremos (ya que no afectará a la información) y volvemos a examinar los datos.

```
datos <- datos %>%
  filter(edad < 100)
ggplot(datos) +
```

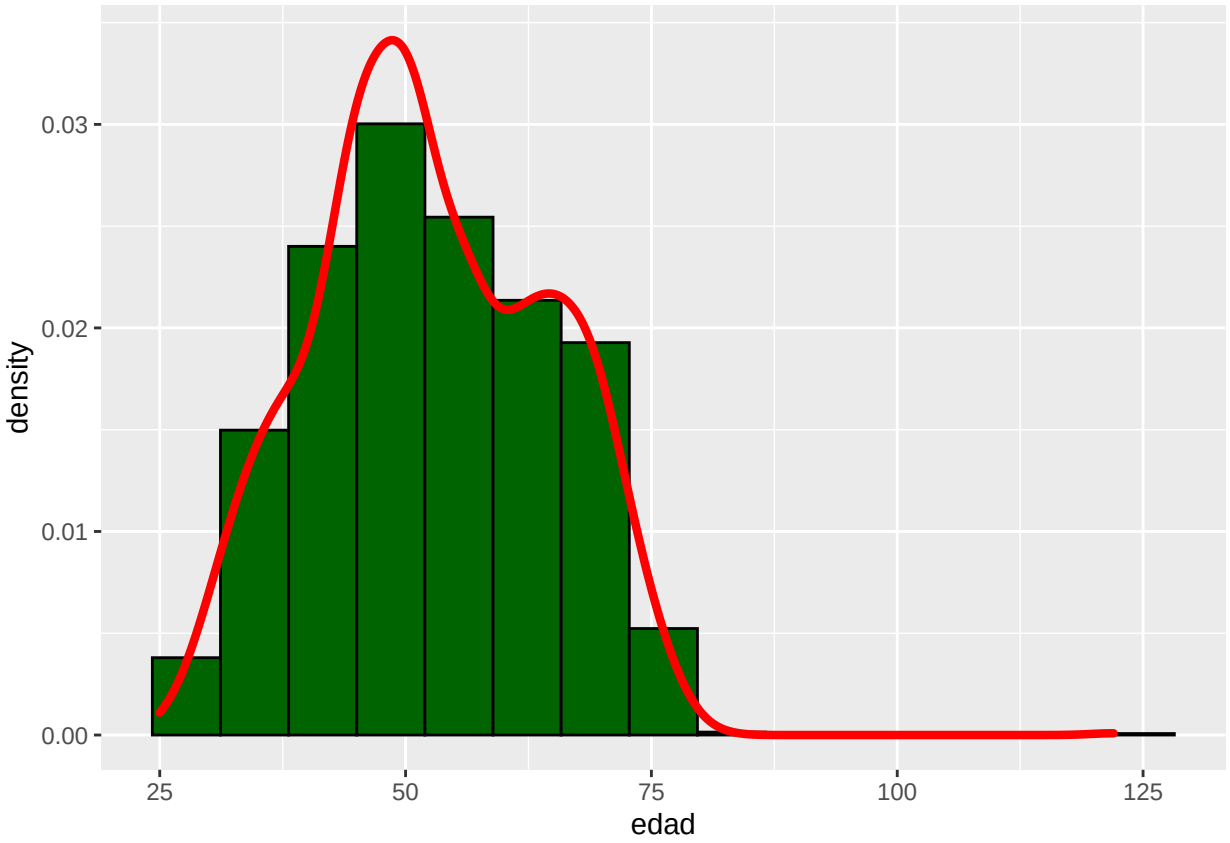

Figura 5: Distribución edad

```
  geom_histogram(aes(x = edad, y =stat(density)), bins = 15, fill = "darkgreen",
                 color ='black') +
  geom_density(aes(x = edad), color="red", size=1.5)
```

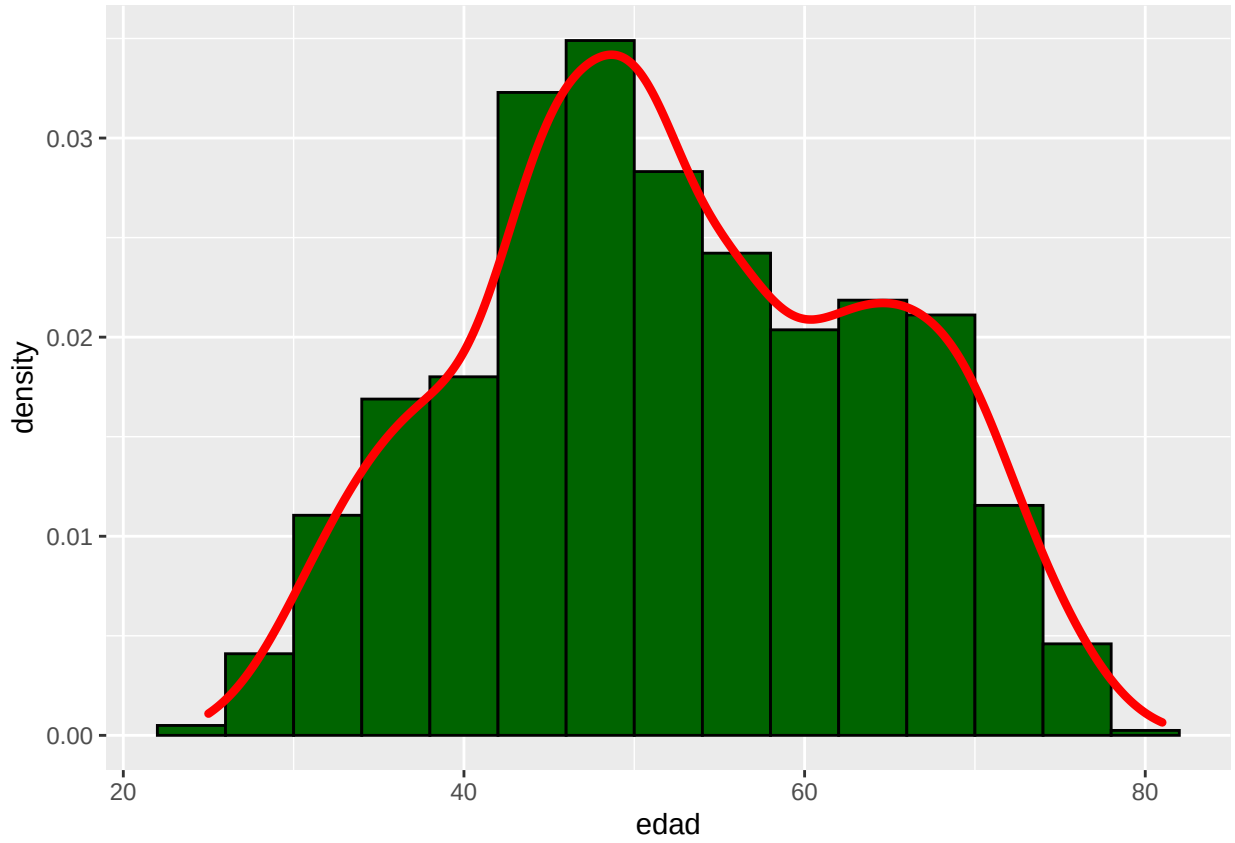

Figura 6: Distribución edad retocado

Observamos que nuestros datos ya tienen una forma que es plausible, por lo que podemos iniciar el análisis de los datos.

```
g1 <-ggplot(datos) +
  geom_jitter(aes(x = edad, y = MntWines))
g2<-ggplot(datos) +
  geom_jitter(aes(x = edad, y = MntFruits))
g3<-ggplot(datos) +
  geom_jitter(aes(x = edad, y = MntMeatProducts))
g4<-ggplot(datos) +
  geom_jitter(aes(x = edad, y = MntFishProducts))
g5<-ggplot(datos) +
  geom_jitter(aes(x = edad, y = MntSweetProducts))
g6<-ggplot(datos) +
  geom_jitter(aes(x = edad, y = MntGoldProds))
g7<-ggplot(datos) +
```

```
  geom_jitter(aes(x = edad, y = Dinero_Gastado))
gridExtra::grid.arrange(g1,g2,g3,g4, g5, g6, g7)
```

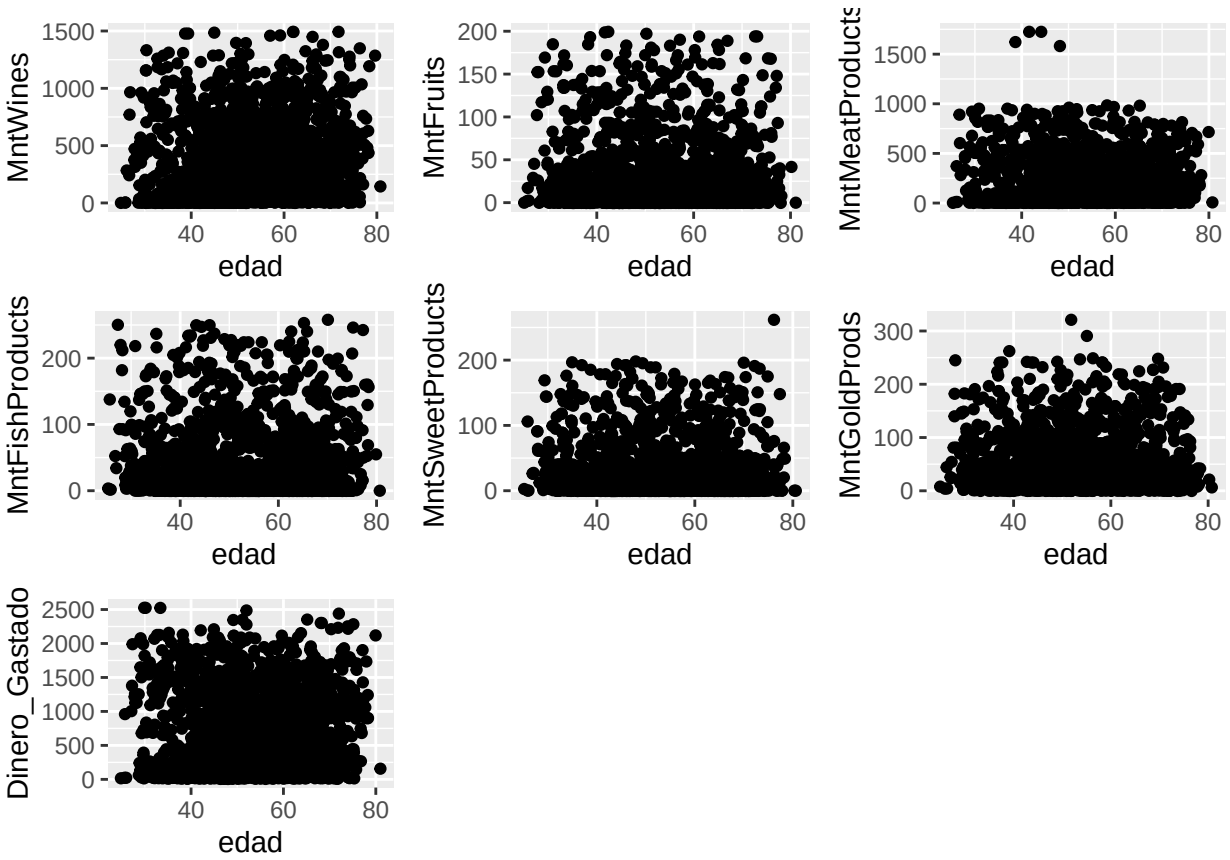

Figura 7: Gráfico edad y gasto por productos

En una primera observación podemos ver que el gasto en vino es mucho mayor que en cualquiera de las otras secciones y para cualquier edad.

Por otra parte, no parece que existan patrones muy claros en ninguna de las variables. Sin embargo, podríamos decir que las personas de mayor edad gastan más dinero en vino y eso probablemente repercute en que gasten más dinero en general.

Realizamos un último estudio en función del nivel de estudios y el estado sentimental. Pero primero vamos a ver cuantos datos hay de cada tipo.

```
table(datos$Education, datos$Marital_Status)
```

```
##
##              Single Not single
##   Basic          19         35
##   Graduation    364        750
##   Master        112        253
##   PhD           148        332
```

```
ggplot(datos) +
  geom_bar(aes(x = Education, fill = Marital_Status), position = "dodge")
```

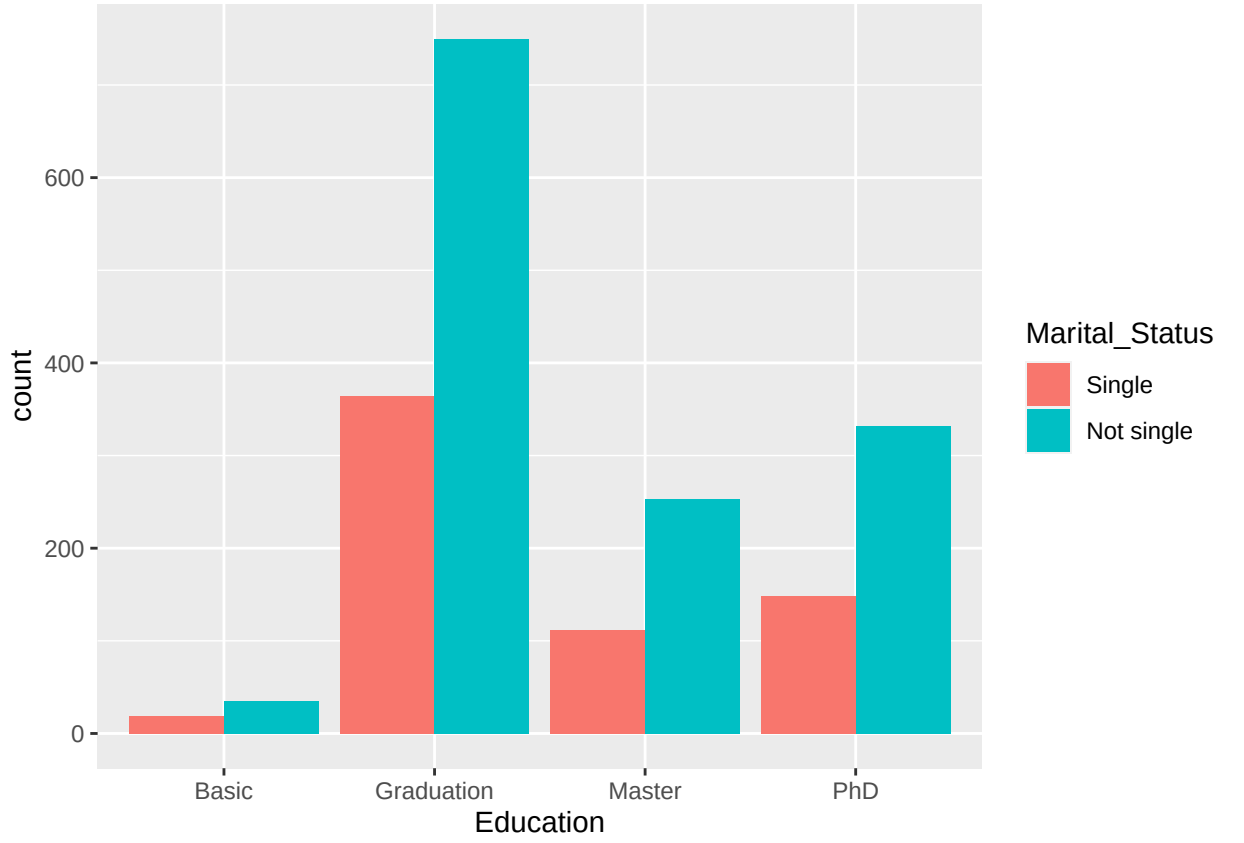

Figura 8: Relación educación y estado civil

Anteriormente hemos visto de forma general que el número de clientes con pareja son el doble de los sin pareja, pero ahora vemos esta relación se cumple además, para todos los grupos de niveles de estudio

Veamos la relación entre el salario y el nivel de estudios:

```
ggplot(datos) +
  geom_boxplot(aes(x = factor(Education), y = Income))
```

Vemos que hay un outlier que no nos permite ver correctamente los gráficos. Por este motivo lo quitamos.

```
aux <- datos %>%
  filter(Income < 500000)
ggplot(aux) +
  geom_boxplot(aes(x = factor(Education), y = Income))
```

Se puede observar cláramente que las personas con un nivel de estudio “Basic” ganan mucho menos que cualquiera de los otros 3 grupos. Otra cosa que hay que tener en cuenta es que entre los 3 grupos restantes no existen deiferencias significativas.

Por último vemos la relación entre el gasto y el nivel de estudios y la situación sentimental.

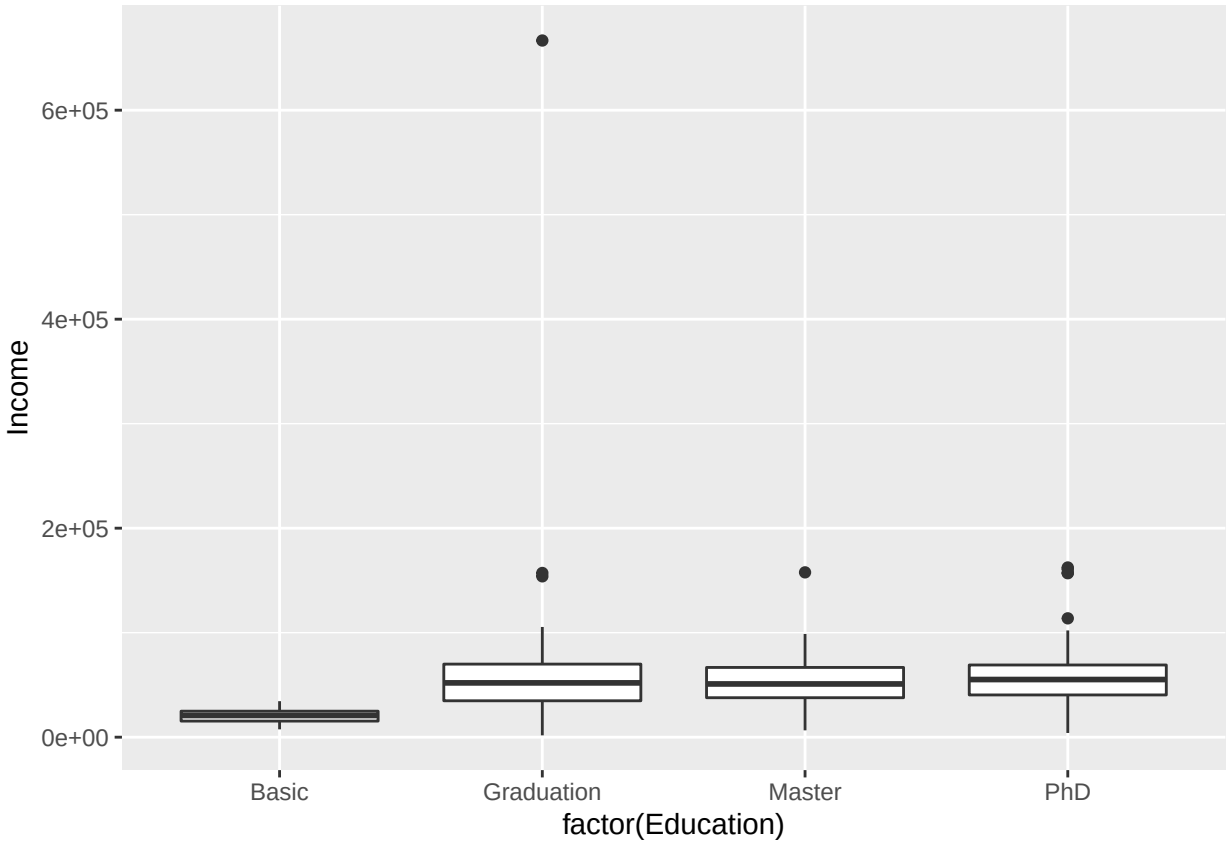

Figura 9: Relación salario y nivel de estudios

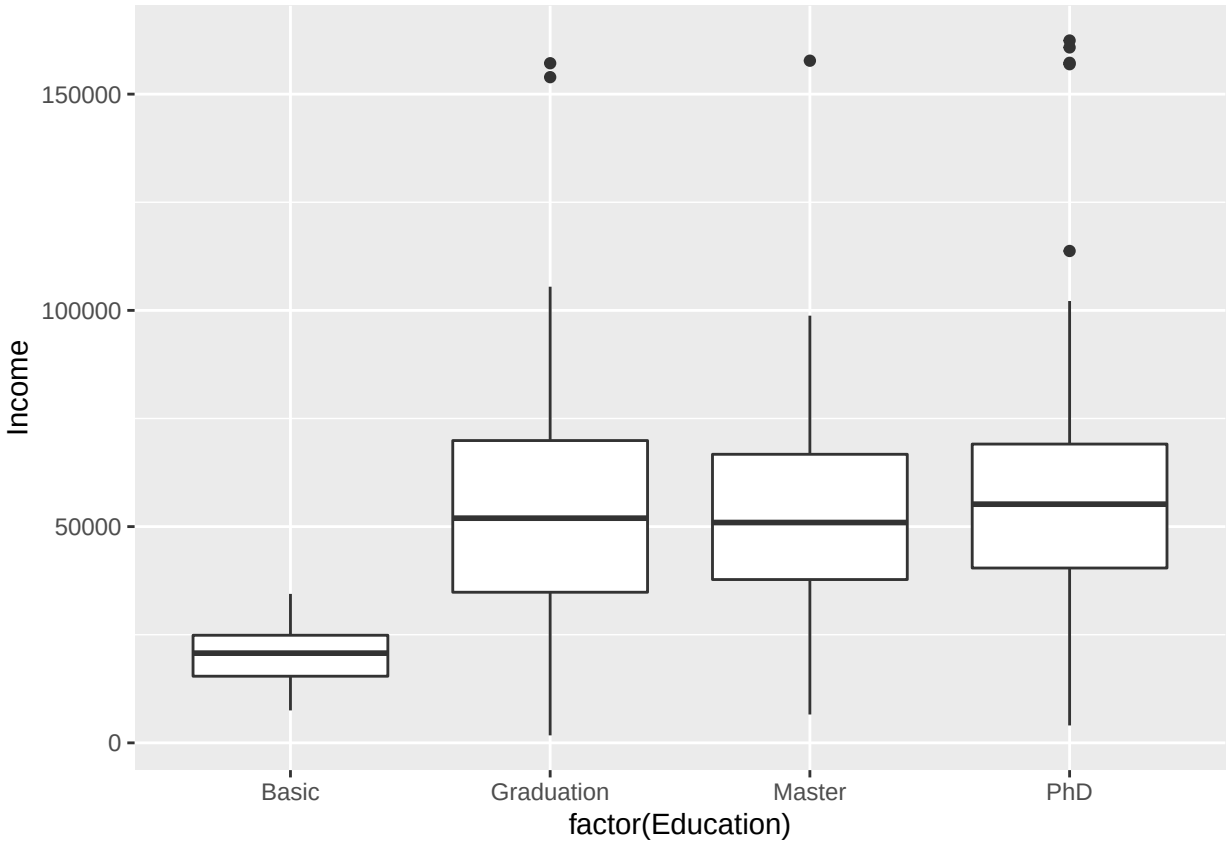

Figura 10: Modificación outlier

```
g1 <- ggplot(datos) +
  geom_col(aes(x = Education, y = MntWines, fill = Marital_Status),position = "dodge")
g2<-ggplot(datos) +
  geom_col(aes(x = Education, y = MntFruits, fill = Marital_Status),position = "dodge")
g3<-ggplot(datos) +
  geom_col(aes(x = Education, y = MntMeatProducts, fill = Marital_Status),position = "dodge")
g4<-ggplot(datos) +
  geom_col(aes(x = Education, y = MntFishProducts, fill = Marital_Status),position = "dodge")
g5<-ggplot(datos) +
  geom_col(aes(x = Education, y = MntSweetProducts, fill = Marital_Status),position = "dodge")
g6<-ggplot(datos) +
  geom_col(aes(x = Education, y = MntGoldProds, fill = Marital_Status),position = "dodge")
g7<-ggplot(datos) +
  geom_col(aes(x = Education, y = Dinero_Gastado, fill = Marital_Status),position = "dodge")
gridExtra::grid.arrange(g1,g2,g3,g4, g5, g6, g7)
```

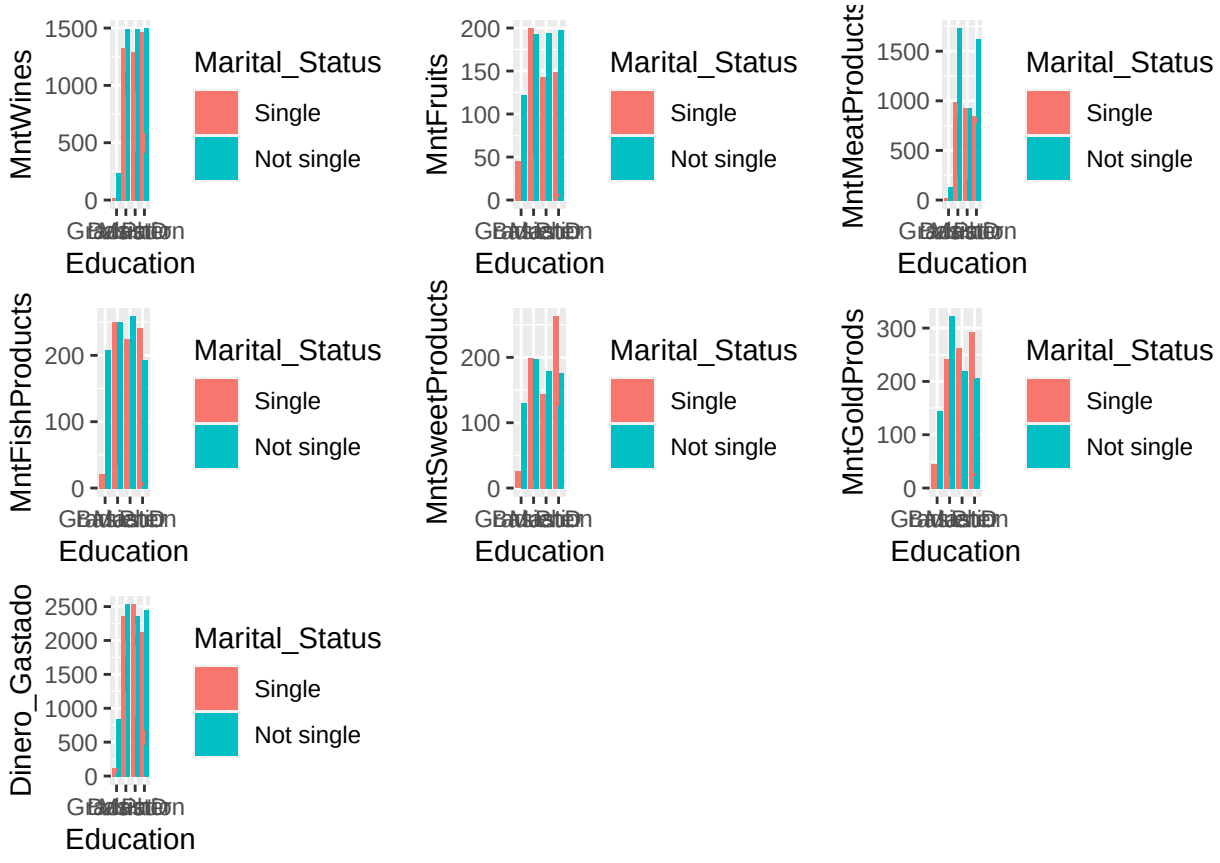

Figura 11: Relación gasto y nivel de estudios

Estos gráficos reflejan lo visto anteriormente, ya que en todas las secciones, los clientes con un nivel de estudio "Basic" gastan mucho menos dinero que cualquiera de los otros 3 grupos, lo que concuerda con que su salario sea menor.

5. Análisis Predictivo y Analítica Avanzada

En esta sección nuestro objetivo es usar técnicas de Machine Learning para predecir tendencias y comportamientos sobre diferentes variables que nos puedan resultar interesantes. Antes de iniciar cualquier estudio lo que haremos será definir nuestros parámetros de control.

5.1. Ajuste parámetros de control

Lo que haremos para para tener siempre un mismo resultado es usar una semilla, en nuestro caso será 2021. Por otro lado para el tema de control usaremos el método de cross-validation con un fold de 10.

```
ctrl <- trainControl(method = "cv",number = 10,summaryFunction = defaultSummary,
                     classProbs = TRUE)
```

5.2. Análisis sobre variable Complain

Otro análisis que se puede realizar mediante técnicas de machine learning es encontrar aquellas variables que son claves a la hora de detectar de forma anticipada que clientes se pueden quejar. Para ello trabajaremos y realizaremos diferentes modelos de clasificación.

Lo primero antes de iniciar cualquier modelo, será ver como está distribuida la variable complain y también cambiarla para poder trabajar con ella:

```
datos$Complain <- ifelse(datos$Complain == 1, 'Yes','No')
table(datos$Complain)
```

```
##
##   No  Yes
## 1996   17
```

Tenemos que el dataset está totalmente desbalanceado, para evitar este problema lo que haremos será rebalancear nuestros datos. Esto nos puede generar alguna anomalía ya que estamos tratando los datos iniciales.

```
set.seed(2021)
datos_comp_rebal <- ovun.sample(Complain ~ ., data = datos, method = 'both',
                                N = table(datos$Complain)[1]*2)$data
table(datos_comp_rebal$Complain)
```

```
##
##   No  Yes
## 2027 1965
```

```
datos_comp_rebal <- datos_comp_rebal %>%
  select(-ID)
datos_comp_rebal$Complain <- as.factor(datos_comp_rebal$Complain)
```

Una vez que tenemos los datos rebalanceados podemos pasar al siguiente paso, que consistirá en mirar si tenemos variables que tengan una alta correlación:

```
catvars <- sapply(datos_comp_rebal, class) %in% c("character","factpr")
numvars <- sapply(datos_comp_rebal, class) %in% c("integer","numeric")
C <- cor(datos_comp_rebal[,numvars])
corrplot(C, method = "circle")
```

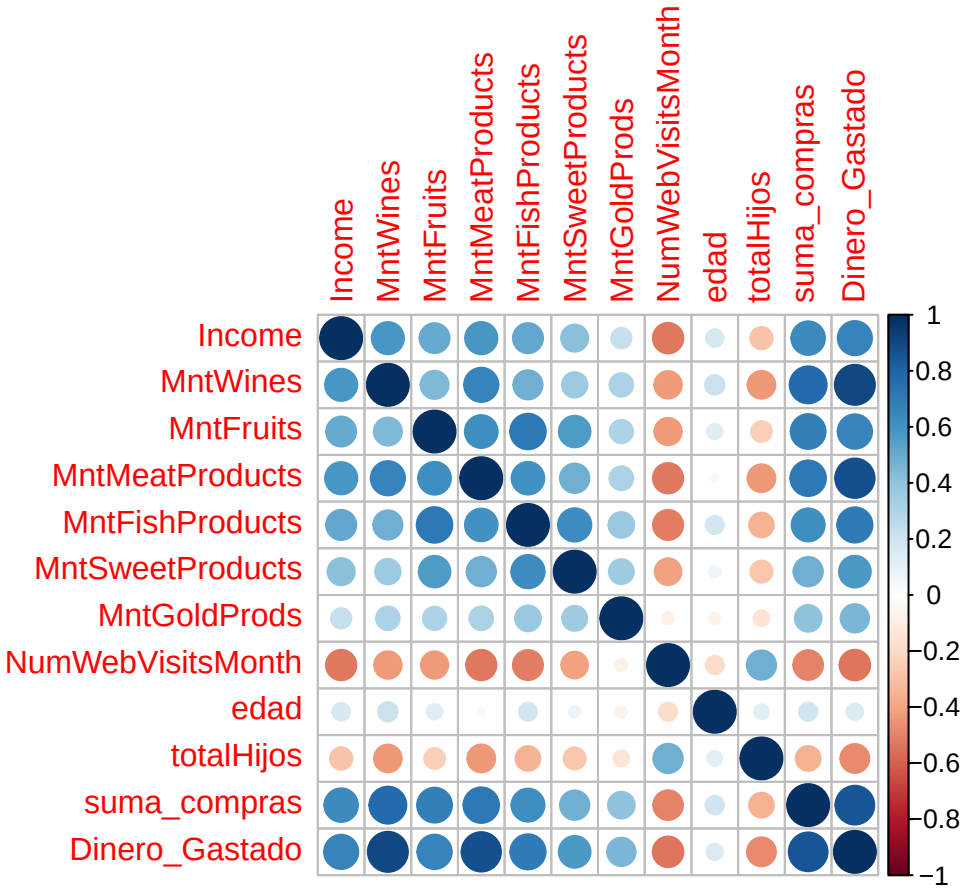

Figura 12: Gráfico de correlación

Vemos que muchas variables tienen una alta correlación entre si, por ahora no haremos nada pero posteriormente veremos si es necesario eliminar alguna o no.

Una vez hecho esto podemos pasar a la parte probabilidad y de modelos, lo primero que haremos será calcular la probabilidad de que se queje si tiene hijos y luego, pasaremos a dividir el conjunto de datos entre entrenamiento y test.

Al realizar el análisis de probabilidades de ser padre y quejarse, obsevamos que la probabilidad de quejarse sin tener hijos es muy pequeña, por lo que se puede afirmar que el hecho de tener hijos puede afectar al caracter de los padres, haciéndolos más irritables.

```
datos_prob<-datos %>%
  mutate(sonPadres = factor(totalHijos>0, levels = c(FALSE, TRUE), labels = c(0,1)))
tabla <- table(datos_prob$Complain, datos_prob$sonPadres)
prop.table(tabla)
```

```
##
##                 0           1
##   No  0.280675609 0.710879285
##   Yes 0.000993542 0.007451565
```

Tras realizar un análisis probabilístico básico, pasamod a la división del dataset.

Una vez hecho esto podemos pasar a la parte de modelos, lo primero que haremos será dividir el conjunto de datos entre entrenamiento y test.

```
set.seed(2021)
trainIndex2 <- createDataPartition(datos_comp_rebal$Complain, p = 0.8, list = FALSE, times = 1)
fTR2 <- datos_comp_rebal[trainIndex2,]
fTS2 <- datos_comp_rebal[-trainIndex2,]
fTR2_eval <- fTR2
fTS2_eval <- fTS2
```

Una vez definido tanto el conjunto de entrenamiento como el de test realizaremos un modelo sencillo para ver cómo se comporta todo, este será una regresión logística. A su vez también usaremos como método de control un cross-validation con un fold de 10.

```
set.seed(2021)
LogReg.fit <- train(form = Complain ~ . , data = fTR2, method = "glm",
                    trControl = ctrl, metric = "Accuracy")
LogReg.fit
```

```
## Generalized Linear Model
##
## 3194 samples
##   16 predictor
##    2 classes: 'No', 'Yes'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 2875, 2874, 2875, 2875, 2875, 2874, ...
## Resampling results:
##
##   Accuracy   Kappa
##   0.7069357  0.4140293
```

Podemos observar que nuestro primer modelo, donde tenemos todas las variables, nos devuelve un accuracy de 0.695. Lo cual no esta mal para empezar, pero vayamos a lo que realmente nos interesa, que variables son importantes.

```
summary(LogReg.fit)
```

```
##
## Call:
```

```
## NULL
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.79747  -0.83460  -0.00053   0.83918   1.34729
##
## Coefficients: (1 not defined because of singularities)
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -1.495e+01  1.980e+02  -0.075 0.939825
## EducationGraduation         1.637e+01  1.980e+02   0.083 0.934103
## EducationMaster             1.567e+01  1.980e+02   0.079 0.936912
## EducationPhD                1.380e+01  1.980e+02   0.070 0.944443
## `Marital_StatusNot single` -7.128e-01  1.124e-01  -6.344 2.24e-10 ***
## Income                     -1.780e-06  2.161e-06  -0.824 0.410176
## Dt_Customer.L              -1.058e+00  1.006e-01 -10.518  < 2e-16 ***
## Dt_Customer.Q               3.357e-01  8.705e-02   3.857 0.000115 ***
## MntWines                   -2.759e-03  3.899e-04  -7.077 1.47e-12 ***
## MntFruits                   1.757e-02  2.448e-03   7.178 7.06e-13 ***
## MntMeatProducts            -2.869e-03  5.185e-04  -5.533 3.15e-08 ***
## MntFishProducts            -1.932e-03  1.858e-03  -1.040 0.298366
## MntSweetProducts           -2.358e-02  2.563e-03  -9.199  < 2e-16 ***
## MntGoldProds               -1.435e-02  1.409e-03 -10.183  < 2e-16 ***
## NumWebVisitsMonth          -2.455e-01  3.163e-02  -7.761 8.45e-15 ***
## Response1                   1.110e+00  1.802e-01   6.162 7.18e-10 ***
## edad                        2.517e-02  4.269e-03   5.895 3.75e-09 ***
## totalHijos                  1.233e-01  7.844e-02   1.572 0.116008
## suma_compras                6.074e-02  1.505e-02   4.036 5.43e-05 ***
## Dinero_Gastado                     NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4427.0  on 3193  degrees of freedom
## Residual deviance: 3266.7  on 3175  degrees of freedom
## AIC: 3304.7
##
## Number of Fisher Scoring iterations: 14
```

Vemos que las variables importantes en nuestro modelo son el estado civil, la antigüedad del cliente, el consumo en diferentes productos, las veces que visitan la web, si respondes a la ofertas, la edad y el total de compras. Cabe destacar que este modelo solo detecta importancia de variables lineales con el output.

Evaluamos nuestro modelo para obtener datos más claros.

```
set.seed(2021)
fTR2_eval$LRprob <- predict(LogReg.fit, type="prob", newdata = fTR2)
fTR2_eval$LRpred <- predict(LogReg.fit, type="raw", newdata = fTR2)
fTS2_eval$LRprob <- predict(LogReg.fit, type="prob", newdata = fTS2)
fTS2_eval$LRpred <- predict(LogReg.fit, type="raw", newdata = fTS2)
```

Ahora implementaremos otro modelo, que será el de arbol de decisiones. Este tiene la ventaja de que si nos dará las variables más importantes incluso si tienen relación no lineal con el output.

```
set.seed(2021)
tree2.fit <- train(x = fTR2[,c(seq(1,11),seq(13,17))],y = fTR2$Complain,method ="rpart",
            control=rpart.control(minsplit=20,minbucket = 20), parms = list(split = "gini"),
            tuneGrid = data.frame(cp = seq(0,0.1,0.001)), trControl = ctrl,metric = "Accuracy")
ggplot(tree2.fit)
```

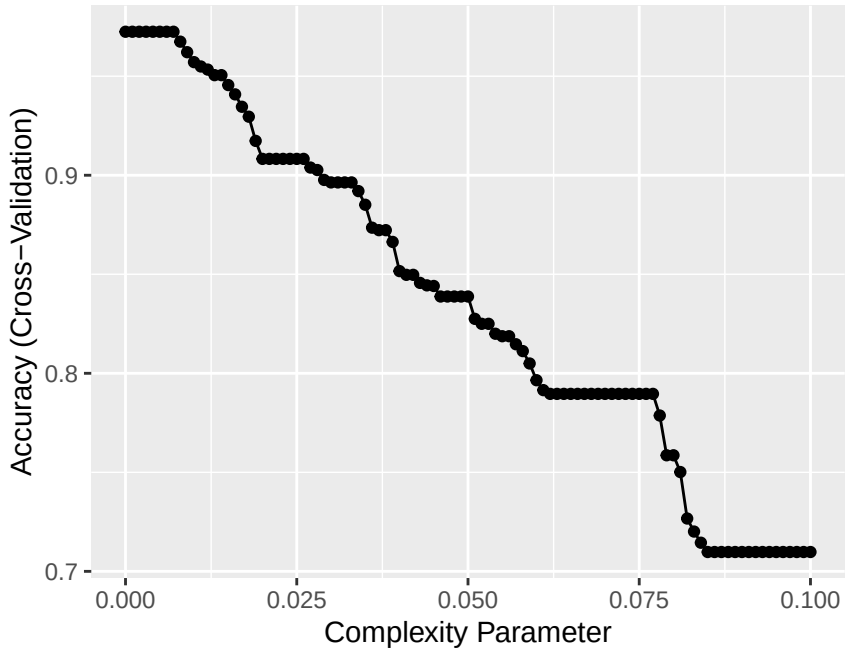

Figura 13: Gráfico de selección hiperparámetro

Podemos observar que este modelo tiene un accuracy aproximado de 0.97 con un hiperparámetro c igual a 0. Elegimos $c = 0$ ya que es el valor que maximiza el accuracy.

A continuación veremos que variables son importantes, donde entrarán también en juego aquellas variables las cuales tengan relaciones no lineales con nuestro output.

```
plot(varImp(tree2.fit,scale = FALSE))
```

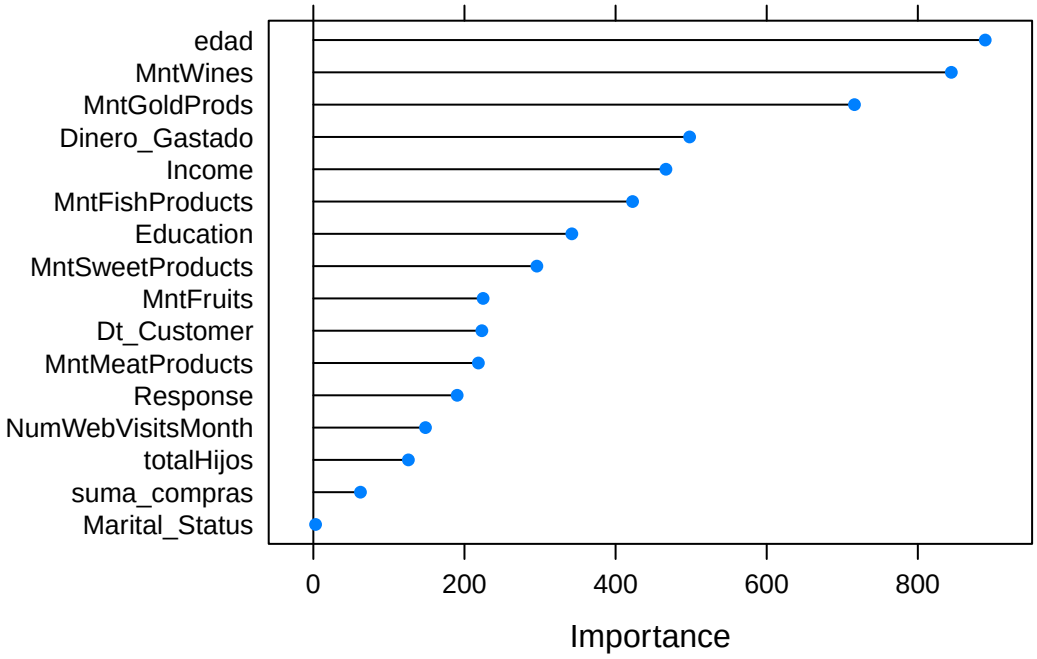

Figura 14: Importancia de variables

Obtenemos que las variables más importantes son el consumo de productos "gourmet", carne y vino así como también la edad, los ingresos y el dinero gastado.

Por último haremos un modelo solo con las variables más importantes, tanto aquellas que tienen una relación lineal como aquellas que son no lineales. Para una mayor facilidad de comprensión del modelo y viendo como ha salido el último, realizaremos un arbol de decisión.

```
set.seed(2021)
tree2_1.fit <- train(x = fTR2[,c(1,3,5,8,9,10,14,17)],y = fTR2$Complain,method ="rpart",
            control=rpart.control(minsplit=20,minbucket = 20), parms = list(split = "gini"),
            tuneGrid = data.frame(cp = seq(0,0.1,0.001)), trControl = ctrl,metric = "Accuracy")
ggplot(tree2_1.fit)
```

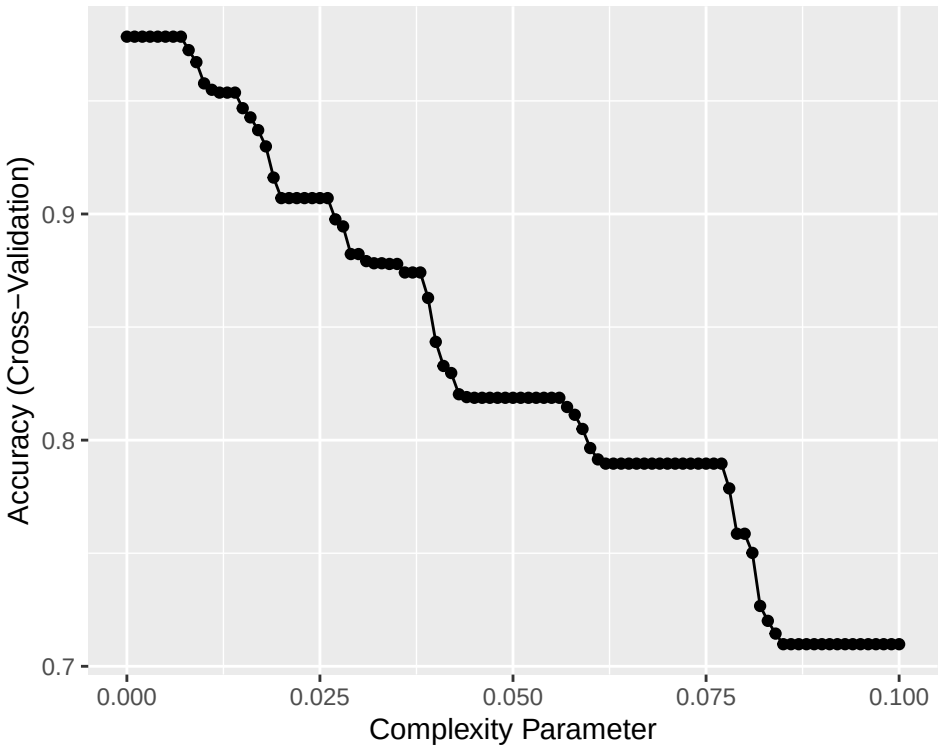

Figura 15: Gráfico de selección hiperparámetro

En este caso vemos que el accuracy no varía y que se vuelve a coger el mismo valor de hiperparámetro.

Veamos como esta construido nuestro modelo de arbol de decisión:

```
rpart.plot(tree2_1.fit$finalModel, type = 2, fallen.leaves = FALSE, box.palette = "Oranges")
```

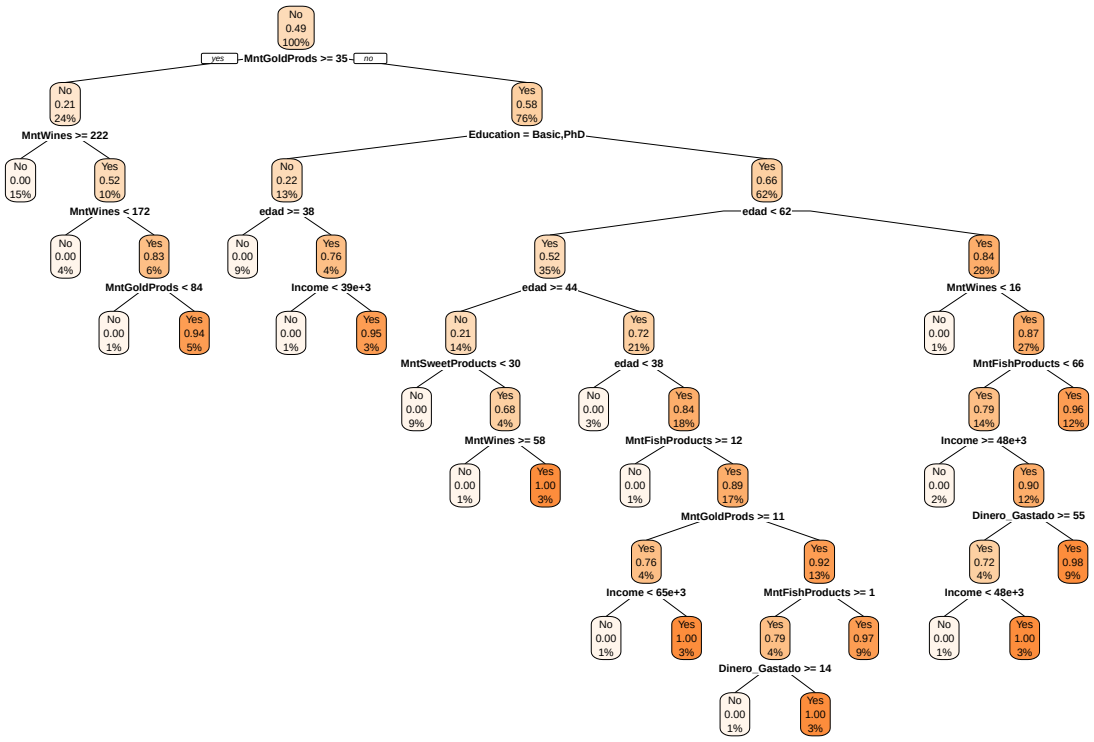

Figura 16: Estructura Arbol de Decisión

Por último haremos como en el caso de la regresión logística, donde representamos graficamente la predicción frente a los valores reales de cada observación.

```
set.seed(2021)
fTR2_eval$tree2_1_prob <- predict(tree2_1.fit, type="prob", newdata = fTR2)
fTR2_eval$tree2_1_pred <- predict(tree2_1.fit, type="raw", newdata = fTR2)
fTS2_eval$tree2_1_prob <- predict(tree2_1.fit, type="prob", newdata = fTS2)
fTS2_eval$tree2_1_pred <- predict(tree2_1.fit, type="raw", newdata = fTS2)
```

```
Plot2DClass(fTR2[,c(1,3,5,8,9,10,14,17)],fTR2$Complain,tree2_1.fit,var1 = "edad",
            var2 = "Dinero_Gastado", selClass = "Yes")
```

Figura 17: Gráfico de predicción

A su vez, para ver que tal trabaja tanto en training como en test sacaremos ambas matrices de confusión:

```
confusionMatrix(data = fTR2_eval$tree2_1_pred,reference = fTR2_eval$Complain,positive = "Yes")
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   No  Yes
##        No  1580    0
##        Yes   42 1572
##
##                Accuracy : 0.9869
##                  95% CI : (0.9823, 0.9905)
##     No Information Rate : 0.5078
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.9737
##
##  Mcnemar's Test P-Value : 2.509e-10
##
##             Sensitivity : 1.0000
##             Specificity : 0.9741
##          Pos Pred Value : 0.9740
##          Neg Pred Value : 1.0000
##              Prevalence : 0.4922
##          Detection Rate : 0.4922
##    Detection Prevalence : 0.5053
##       Balanced Accuracy : 0.9871
##
##        'Positive' Class : Yes
##
```

```
set.seed(2021)
confusionMatrix(data = fTS2_eval$tree2_1_pred,reference = fTS2_eval$Complain,positive = "Yes")
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  No Yes
##        No  395   0
##        Yes  10 393
##
##                Accuracy : 0.9875
##                  95% CI : (0.9771, 0.994)
##     No Information Rate : 0.5075
```

```
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.9749
##
##  Mcnemar's Test P-Value : 0.004427
##
##             Sensitivity : 1.0000
##             Specificity : 0.9753
##          Pos Pred Value : 0.9752
##          Neg Pred Value : 1.0000
##              Prevalence : 0.4925
##          Detection Rate : 0.4925
##    Detection Prevalence : 0.5050
##       Balanced Accuracy : 0.9877
##
##        'Positive' Class : Yes
##
```

Por tanto, podemos afirmar que las variables que más importancia tienen para detectar que un cliente se va a quejar en un futuro son el gasto en productos "gourmet" y vino así como la edad, los ingresos y el dinero gastado. Esto se puede ver en la siguiente imagen:

```
plot(varImp(tree2_1.fit,scale = FALSE))
```

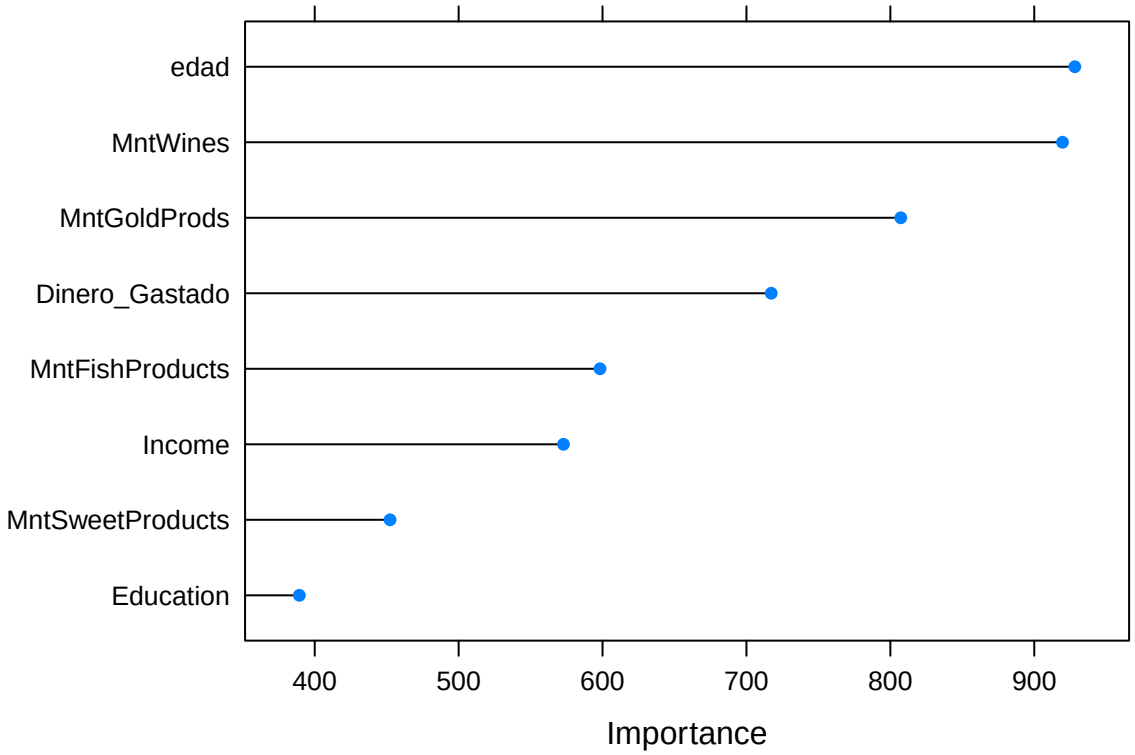

Figura 18: Gráfico importancia varibales