Tipología y Cliclo de Vida de los Datos: Práctica 2: Limpieza y análisis de datos

Autor: Jose Manuel Gómez

Junio 2020

Detalles de la actividad

Descripción

En esta actividad se elabora un caso práctico, consistente en el tratamiento de un conjunto de datos orientado a aprender a identificar los datos relevantes para un proyecto analítico y usar las herramientas de integración, limpieza, validación y análisis de las mismas.

Objetivos

Los objetivos que se persiguen mediante el desarrollo de esta actividad práctica son los siguientes:

- Aprender a aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios o multidisciplinares.
- Saber identificar los datos relevantes y los tratamientos necesarios (integración, limpieza y validación) para llevar a cabo un proyecto analítico.
- Aprender a analizar los datos adecuadamente para abordar la información contenida en los datos.
- Identificar la mejor representación de los resultados para aportar conclusiones sobre el problema planteado en el proceso analítico.
- Actuar con los principios éticos y legales relacionados con la manipulación de datos en función del ámbito de aplicación.
- Desarrollar las habilidades de aprendizaje que permita continuar estudiando de un modo que tendrá que ser en gran medida autodirigido o autónomo.
- Desarrollar la capacidad de búsqueda, gestión y uso de información y recursos en el ámbito de la ciencia de datos.

Competencias

En esta práctica se desarrollan las siguientes competencias del Master de Data Science:

- Capacidad de analizar un problema en el nivel de abstracción adecuado a cada situación y aplicar las habilidades y conocimientos adquiridos para abordarlo y resolverlo.
- Capacidad para aplicar las técnicas específicas de tratamiento de datos (integración, transformación, limpieza y validación) para su posterior análisis.

Resolución

Selección de juego de datos

Como juego de datos he escogido un dataset de kaggle llamado Bank Loan modelling que recoge datos bancarios de una entidad americana con el fin de identificar un nuevo conjunto de clientes en el futuro, dependiendo de los atributos disponibles en los datos. https://www.kaggle.com/itsmesunil/bank-loan-modelling

El dataset incluye 5000 observaciones con atributos que contienen información demográfica de los clientes (edad, experiencia, ingresos, código postal, familia, gasto medio en tarjerta de crédito y nivel educativo), información relativa a la relación del cliente con la entidad bancaria (hipoteca, cuenta de seguridad, depósito a plazo, cuenta online y tarjeta de crédito) y la respuesta del cliente a la oferta del banco de firmar un prestamo personal (prestamo) .Tiene catorce variables divididas en cuatro categorías de medición diferentes.

Estos datos se dividen por categorias; la categoría binaria tiene cinco variables, incluyendo crédito personal, cuenta de seguridad, depósito a plazo, cuenta online y tarjeta de crédito. La categoría de intervalos contiene cinco variables: edad, experiencia, ingresos, gasto medio en tarjeta de crédito e hipoteca. La categoría ordinal incluye las variables familia y educación. Y por último la categoría nominal con ID y código postal.

Objetivos del análisis

A partir de este conjunto de datos se plantea determinar qué variables influyen más sobre el hecho de que el cliente acepte o no una oferta de préstamo personal por parte de la entidad bancaria.

Los atribtutos son numéricos o boolenaos, que pueden ser transformados en numéricos, sin cadenas de texto ni categóricos y por lo tanto idóneo para tabajar con métodos de aprendizaje no supervisados como clustering. También se podrían realizar métodos supervisados por ejemplo para predecir la contratación de un prestamo personal por el cliente en el futuro.

Limpieza, acondicionado y análisis exploratorio

Pasamos a importar los datos para visualizar la estructura y resumen de los datos:

```
bd <-read.csv('Bank_personal_Loan_Modelling.csv')
colnames(bd) <- c("ID", "edad", "experiencia", "ingresos", "CP", "familia", "gastoTC", "educacion", "hi
head(bd)
```

```
##   ID edad experiencia ingresos    CP familia gastoTC educacion hipoteca
## 1  1   25           1       49 91107       4     1.6         1        0
## 2  2   45          19       34 90089       3     1.5         1        0
## 3  3   39          15       11 94720       1     1.0         1        0
## 4  4   35           9      100 94112       1     2.7         2        0
## 5  5   35           8       45 91330       4     1.0         2        0
## 6  6   37          13       29 92121       4     0.4         2      155
##   prestamo cuentaseguridad deposito online TC
## 1        0               1        0      0  0
## 2        0               1        0      0  0
## 3        0               0        0      0  0
## 4        0               0        0      0  0
## 5        0               0        0      0  1
## 6        0               0        0      1  0
```

El dataset contiene 5000 entradas y 14 variables:

```r
dim(bd)
```

```
## [1] 5000   14
```

Las variables son las siguientes:

- ID: Identificador del cliente
- edad: edad del cliente en años
- experiencia: años de experiencia profesional
- ingresos: ingreso anual del cliente en miles de dólares
- CP: Código Postal del cliente.
- familia: número de miembros de la unidad familiar
- gastoTC: gasto medio mesual en tarjeta de crédito en miles de dólares
- educacion: nivel de educación. 1: Undergrad; 2: Graduate; 3: Advanced/Professional
- hipoteca: valor de de la hipoteca en miles de dólares (si tiene)
- prestamo: aceptó el cliente el prestamo personal ofrecido en la última campaña del banco?
- cuentaseguridad: tiene el cliente una cuenta de seguridad en el banco?
- depósito: tiene el cliente un depósito a plazos en el banco?
- online: tiene el cliente una cuenta online?
- TC: tiene el cliente una tarjeta de crédito en el banco?

Comprobamos las clases de las variables y mostramos el summary:

```r
str(bd)
```

```
## 'data.frame':    5000 obs. of  14 variables:
##  $ ID             : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ edad           : int  25 45 39 35 35 37 53 50 35 34 ...
##  $ experiencia    : int  1 19 15 9 8 13 27 24 10 9 ...
##  $ ingresos       : int  49 34 11 100 45 29 72 22 81 180 ...
##  $ CP             : int  91107 90089 94720 94112 91330 92121 91711 93943 90089 93023 ...
##  $ familia        : int  4 3 1 1 4 4 2 1 3 1 ...
##  $ gastoTC        : num  1.6 1.5 1 2.7 1 0.4 1.5 0.3 0.6 8.9 ...
##  $ educacion      : int  1 1 1 2 2 2 2 3 2 3 ...
##  $ hipoteca       : int  0 0 0 0 0 155 0 0 104 0 ...
##  $ prestamo       : int  0 0 0 0 0 0 0 0 0 1 ...
##  $ cuentaseguridad: int  1 1 0 0 0 0 0 0 0 0 ...
##  $ deposito       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ online         : int  0 0 0 0 0 1 1 0 1 0 ...
##  $ TC             : int  0 0 0 0 1 0 0 1 0 0 ...
```

```r
summary(bd)
```

```
##        ID            edad        experiencia      ingresos           CP
##  Min.   :   1   Min.   :23.00   Min.   :-3.0   Min.   :  8.00   Min.   : 9307
##  1st Qu.:1251   1st Qu.:35.00   1st Qu.:10.0   1st Qu.: 39.00   1st Qu.:91911
##  Median :2500   Median :45.00   Median :20.0   Median : 64.00   Median :93437
##  Mean   :2500   Mean   :45.34   Mean   :20.1   Mean   : 73.77   Mean   :93153
##  3rd Qu.:3750   3rd Qu.:55.00   3rd Qu.:30.0   3rd Qu.: 98.00   3rd Qu.:94608
##  Max.   :5000   Max.   :67.00   Max.   :43.0   Max.   :224.00   Max.   :96651
##     familia         gastoTC         educacion        hipoteca
```

```
##  Min.   :1.000   Min.   : 0.000   Min.   :1.000   Min.   :  0.0
##  1st Qu.:1.000   1st Qu.: 0.700   1st Qu.:1.000   1st Qu.:  0.0
##  Median :2.000   Median : 1.500   Median :2.000   Median :  0.0
##  Mean   :2.396   Mean   : 1.938   Mean   :1.881   Mean   : 56.5
##  3rd Qu.:3.000   3rd Qu.: 2.500   3rd Qu.:3.000   3rd Qu.:101.0
##  Max.   :4.000   Max.   :10.000   Max.   :3.000   Max.   :635.0
##     prestamo     cuentaseguridad     deposito          online
##  Min.   :0.000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.000   Median :0.0000   Median :0.0000   Median :1.0000
##  Mean   :0.096   Mean   :0.1044   Mean   :0.0604   Mean   :0.5968
##  3rd Qu.:0.000   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:1.0000
##  Max.   :1.000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##        TC
##  Min.   :0.000
##  1st Qu.:0.000
##  Median :0.000
##  Mean   :0.294
##  3rd Qu.:1.000
##  Max.   :1.000
```

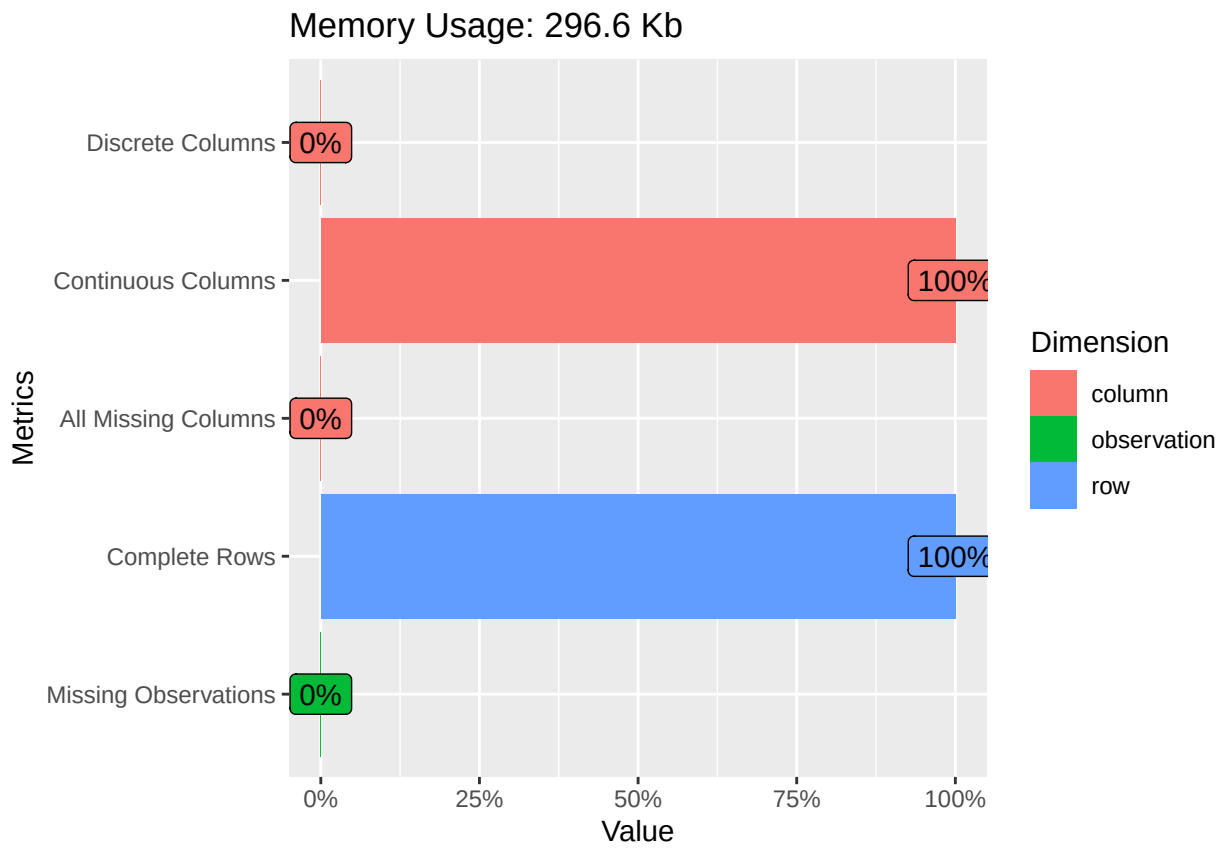
La estadística descriptiva de las variables numéricas sería la siguiente:

- edad: el rango de Q1 a Q3 esta entre 35 y 55. Debido a que la media es muy similar a la mediana, podemos decir que la edad sigue una distribución normal.
- experiencia: el rango de Q1 a Q3 esta entre 20 y 30. Debido a que la media es muy similar a la mediana, podemos decir que la edad sigue una distribución normal. Sin embargo, veremso mas adelante que el campo experiencia contiene algunos valores negativos y que imputaremos nuevos valores.
- ingresos: el rango de Q1 a Q3 esta entre 39 y 98. Ya que la mediana es mayor que la media, podemos decir que la distribución está inclinada a la derecha.
- gastoTC: el rango de Q1 a Q3 esta entre 0.7 y 2.5. Ya que la mediana es mayor que la media, podemos decir que la distribución está inclinada a la derecha.
- hipoteca: el Q3 está en 101 mil dolares y el máximo es de 635 mil dolares. Podemos decir que que el campo hipoteca sigue una distribución que está altamente inclinada a la derecha.

Veamos un resumen gráfico del banco de datos:

```
library(DataExplorer)

plot_intro(bd)
```



Para comprobar si tenemos valores perdidos podemos hacer uso de la función is.na(). Más concretamente, para saber que campos tienen valores perdidos podemos hacer lo siguiente:

```
colSums(is.na(bd))
```

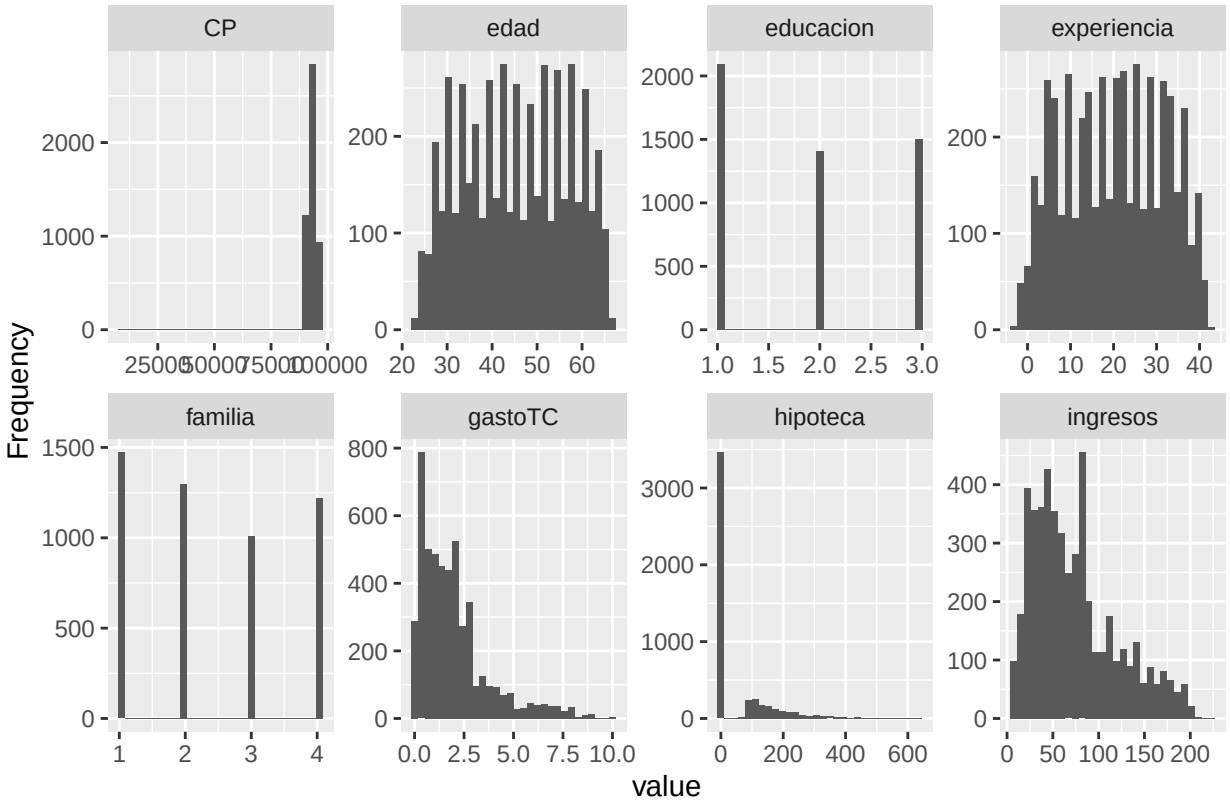
```
##              ID            edad     experiencia        ingresos              CP
##               0               0               0               0               0
##         familia        gastoTC       educacion        hipoteca        prestamo
##               0               0               0               0               0
## cuentaseguridad        deposito          online              TC
##               0               0               0               0
```

Vemos que este dataset no tiene ningún valor perdido.

Mostramos histograma de los valores numéricos:

```
library(DataExplorer)

plot_histogram(bd[2:14.])
```



Los atributos ID y Código Postal no aportan información relevante para el estudio y por lo tanto, podemos eliminarlos de la muestra

```
#Eliminamos ID
bd <- bd[-1]

#Eliminados CP
bd <- bd[-4]
```

Analizamos el campo experencia y observamos que contiene valores negativos. Esto carece de sentido ya que un individuo no pueden tener -x años de experiencia:

```
table(bd$experiencia)
```

```
##
##  -3  -2  -1   0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16
##   4  15  33  66  74  85 129 113 146 119 121 119 147 118 116 102 117 127 119 127
##  17  18  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36
## 125 137 135 148 113 124 144 131 142 134 125 138 124 126 104 154 117 125 143 114
##  37  38  39  40  41  42  43
## 116  88  85  57  43   8   3
```

Esto puede tener afectación en el estudio, y por lo tanto tenemos que sustituir dichos valores. Hay diversos métodos de sustitución. Podemos hacerlo por la media, la mediana o incluso hacer uso de algoritmos como K-Nearest Neighbours Imputation (kNN) para popularlos con los valores cercanos en base a la distancia Gower

```r
library(VIM)

bd[bd$experiencia < 0,'experiencia']
```

```
##  [1] -1 -1 -2 -2 -1 -1 -1 -1 -1 -2 -1 -1 -1 -2 -2 -1 -1 -1 -1 -1 -1 -1 -2 -1 -3
## [26] -2 -1 -2 -2 -1 -1 -2 -1 -1 -1 -1 -1 -1 -3 -2 -1 -2 -1 -1 -1 -2 -3 -2 -2 -3
## [51] -1 -1
```

```r
summary(bd$experiencia)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    -3.0    10.0    20.0    20.1    30.0    43.0
```

La primera opción sería sustituir los valores negativos por la media o la mediana calculada arriba. Vamos a ver que valores imputaría el algoritmo knn:

```r
#Para usar knn necesitamos convertir los valores negativos en NA

id_experiencia_negativa <- which(bd$experiencia < 0)

for (i in id_experiencia_negativa){
  bd$experiencia[i] <- NA
}

output <- kNN(bd,k=3)

#Comprobamos los valores introducidos por knn
output[output$experiencia_imp == TRUE,"experiencia"]
```

```
##  [1]  1  0  0  1  1  2  7  2  8 15  4  1  3  2  0  9  0  1  1  3  0  2  1  2  2
## [26]  1  1  2  1  1 13  8  1  3  1  1  2  1  0 13  6  5  2  5  1  5  6 11  2  2
## [51]  0 12
```

```r
#Y su impacto en la media y mediana
summary(output$experiencia)
```

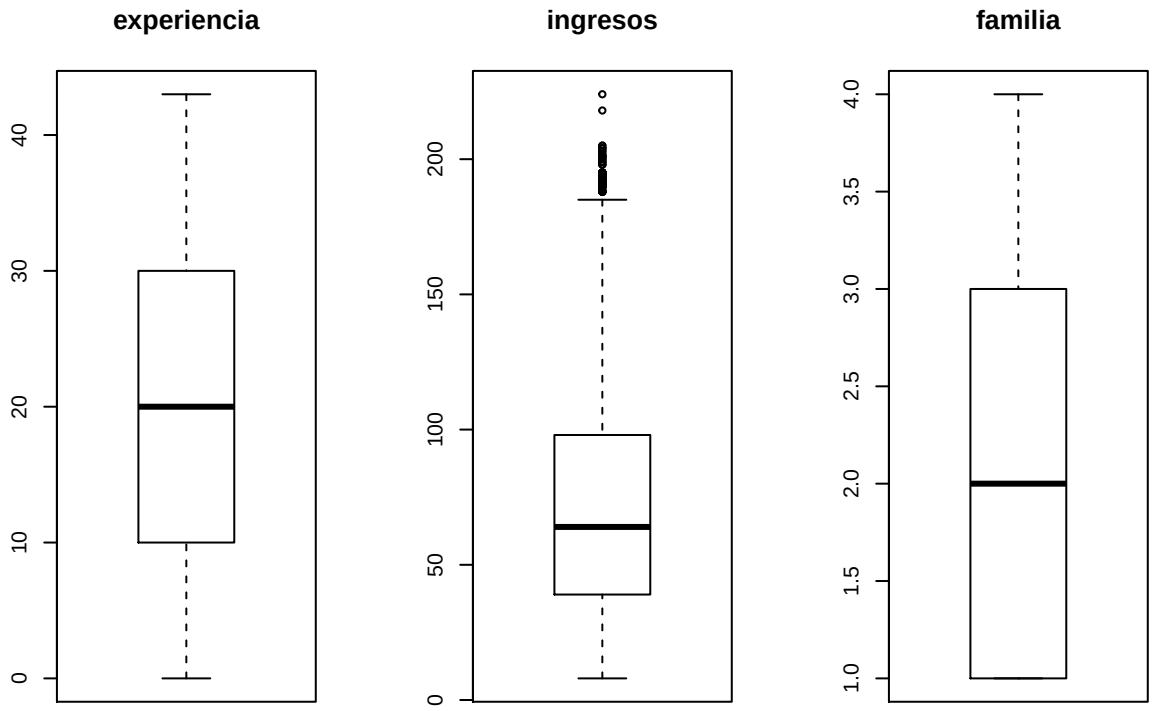
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   10.00   20.00   20.15   30.00   43.00
```

Vemos que el impacto en la media y mediana usando knn es mínimo, por lo tanto nos quedamos con esta opción en lugar de rellenar dichos campos con la media o mediana ya que es mas equitativa. Imputamos los valores al dataset:

```r
bd$experiencia <- output$experiencia
```

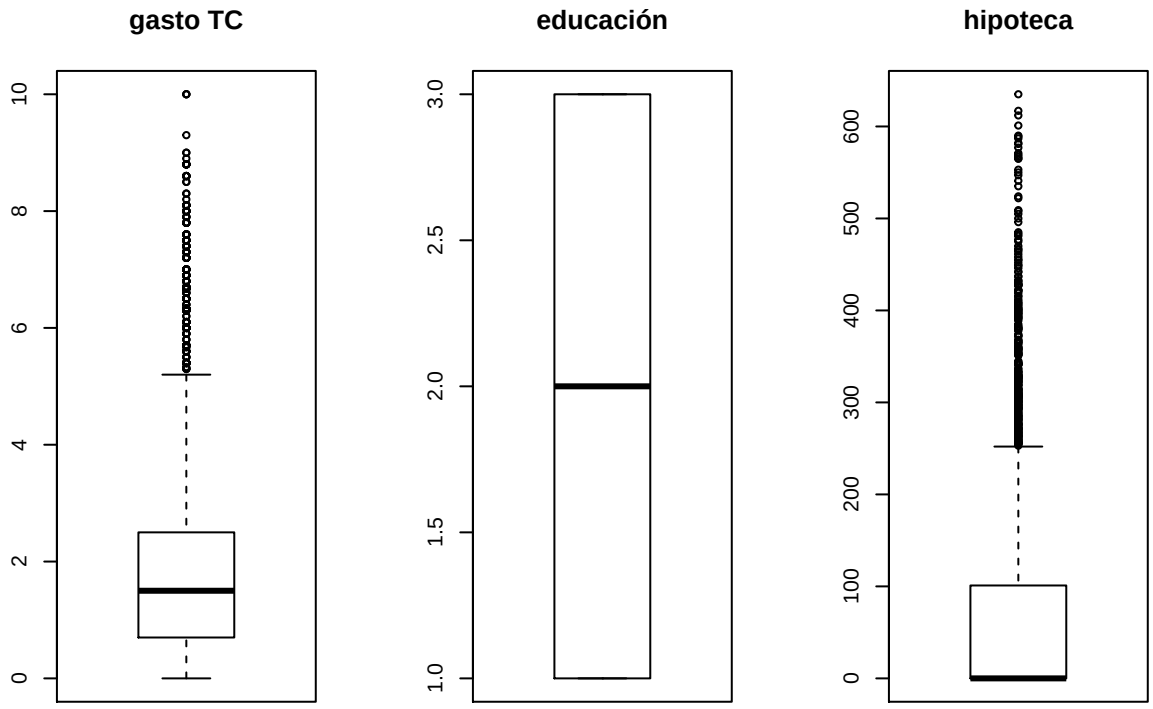
Echemos un vistazo a los posibles outliers de la muestra:

```r
par(mfrow=c(1,3))
boxplot( bd$experiencia, main="experiencia" )
boxplot( bd$ingresos, main="ingresos" )
boxplot( bd$familia, main="familia" )
```



Vemos algunos outliers en ingresos, pero tambien comprobamos que pueden ser valores posibles (ingresos de hasta 224k)

```
par(mfrow=c(1,3))
boxplot (bd$gastoTC, main="gasto TC")
boxplot (bd$educacion, main="educación")
boxplot (bd$hipoteca, main="hipoteca")
```



También podemos visualizar outliers en gasto TC e hipoteca. Aunque los valores pueden estar dentro de la posibilidad real, podemos comprobar si hay correlación entre los gastos de la tarjeta de crédito y los montantes de la hipoteca con la capacidad adquisitiva del cliente. Si vemos que hay correlación, nos podría confirmar que los valores estan dentro de lo razonable y no son outliers a tratar. (Comprobación realizada en apartado 2.6)

Comprobación de la normalidad y homogeneidad de la varianza

Para la comprobación de que los valores que toman nuestras variables cuantitativas provienen de una población distribuida normalmente, utilizaremos la prueba de normalidad de Anderson-Darling. Así, se comprueba que para que cada prueba se obtiene un p-valor superior al nivel de significación prefijado = 0, 05. Si esto se cumple, entonces se considera que variable en cuestión sigue una distribución normal.

```
library(nortest)

alpha = 0.05
col.names = colnames(bd)
for (i in 1:ncol(bd)) {
  if (i == 1) cat("Variables que no siguen una distribución normal:\n")
  if (is.integer(bd[,i]) | is.numeric(bd[,i])) {
    p_val = ad.test(bd[,i])$p.value
    if (p_val < alpha) {
      cat(col.names[i])
      # Format output
      if (i < ncol(bd) - 1) cat(", ")
```

```
      if (i %% 3 == 0) cat("\n")}
  }
}
```

```
## Variables que no siguen una distribución normal:
## edad, experiencia, ingresos,
## familia, gastoTC, educacion,
## hipoteca, prestamo, cuentaseguridad,
## deposito, onlineTC
```

Seguidamente, pasamos a estudiar la homogeneidad de varianzas mediante la aplicación de un test de Fligner-Killeen. En este caso, estudiaremos esta homogeneidad en cuanto a los grupos conformados por los clientes que aceptan o no la hipoteca y su gasto en tarjetas de crédito. En el siguiente test, la hipótesis nula consiste en que ambas varianzas son iguales.

```
fligner.test(gastoTC ~ prestamo, data = bd)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  gastoTC by prestamo
## Fligner-Killeen:med chi-squared = 144.01, df = 1, p-value < 2.2e-16
```

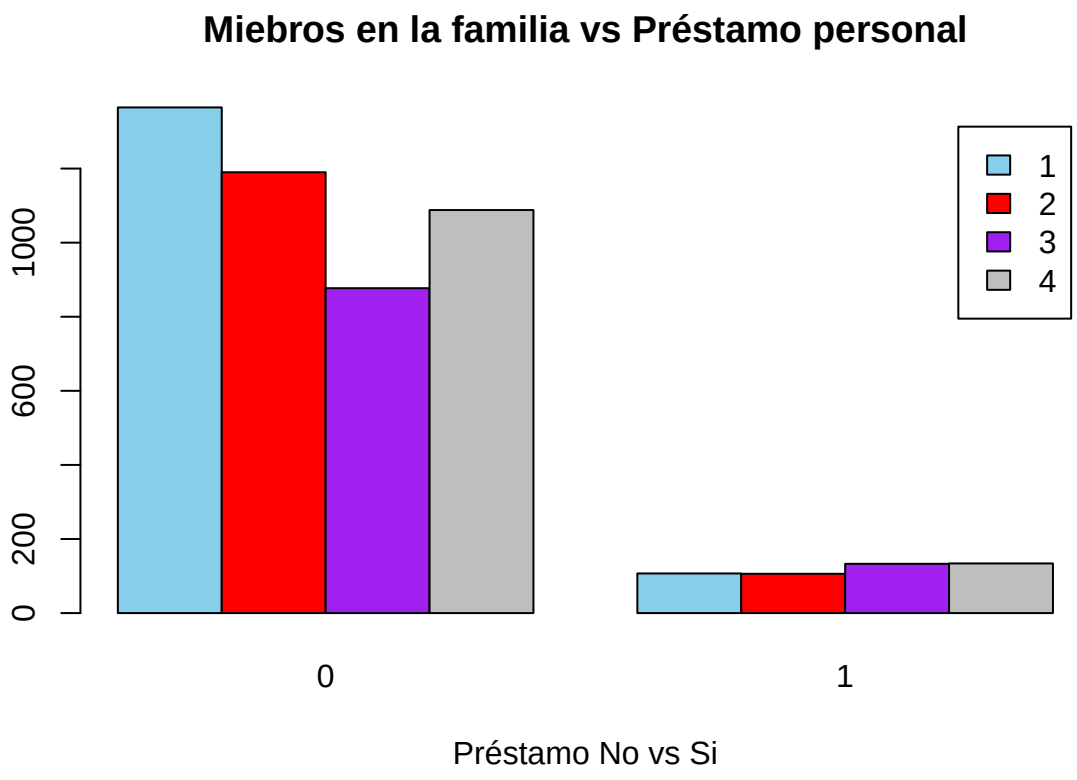
Puesto que obtenemos un p-valor superior a 0,05, aceptamos la hipótesis de que las varianzas de ambas muestras son homogéneas.

Análisis estadísticos

Correlaciones

Veamos las correlaciones entre las diferentes variables del dataset.

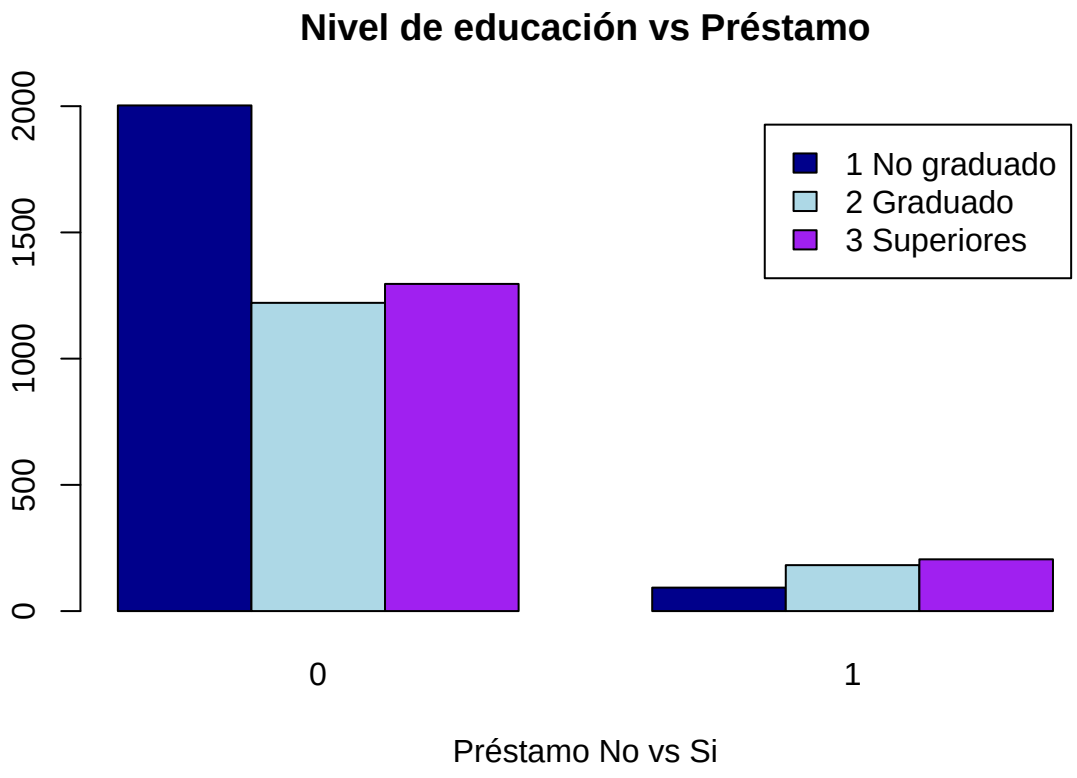
```
barplot(table(bd$familia,bd$prestamo), main="Miebros en la familia vs Préstamo personal", xlab="Préstam
```



Vemos que hay más posibilidad de que acepten un prestamo las familias con mayor número de miembros.

Representamos ahora la correlación entre el nivel de educación y la aceptación del préstamo

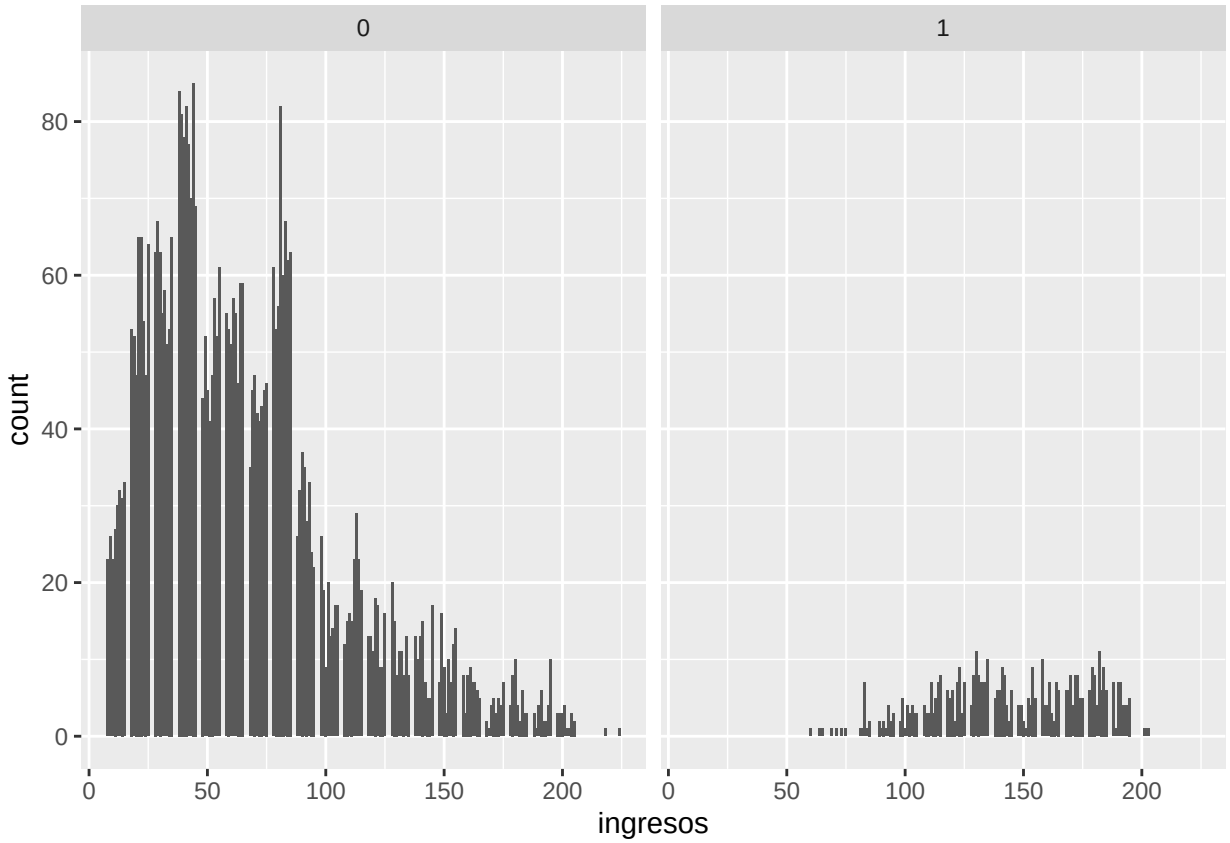
```
barplot(table(bd$educacion, bd$prestamo), main="Nivel de educación vs Préstamo",
xlab="Préstamo No vs Si", col=c("darkblue","lightblue","purple"),
legend = c("1 No graduado", "2 Graduado","3 Superiores"), beside=TRUE)
```



Parece ser que los que aceptan el préstamo tienen mayor grado de educación. Siendo la muestra de Estados Unidos podría estar relacionado con la petición de prestamos de estudio.

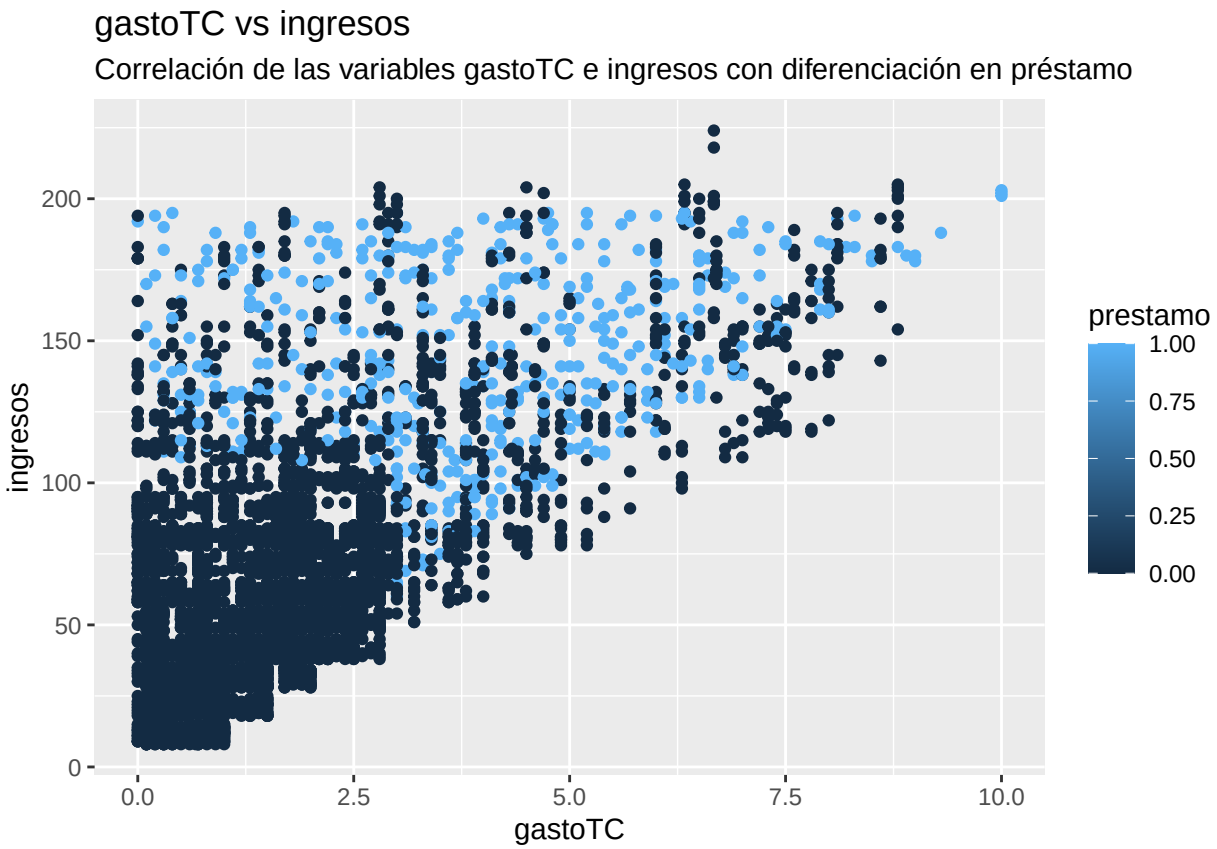
Representamos ahora la relación entre ingresos y aceptación del préstamo:

```
library(ggplot2)
ggplot(data = bd,aes(x=ingresos))+geom_bar()+facet_wrap(~prestamo)
```



Los clientes que han aceptado el préstamo personal tienen una distribución de ingresos muy diferente a aquellos que no han aceptado el préstamo. Añadamos el gasto en tarjetas de crédito a la ecuación para ver si podemos sacar alguna conclusión al respecto:

```
ggplot(bd, aes(x = gastoTC, y = ingresos)) + geom_point(aes(color = prestamo)) + labs(title = "gastoTC
```



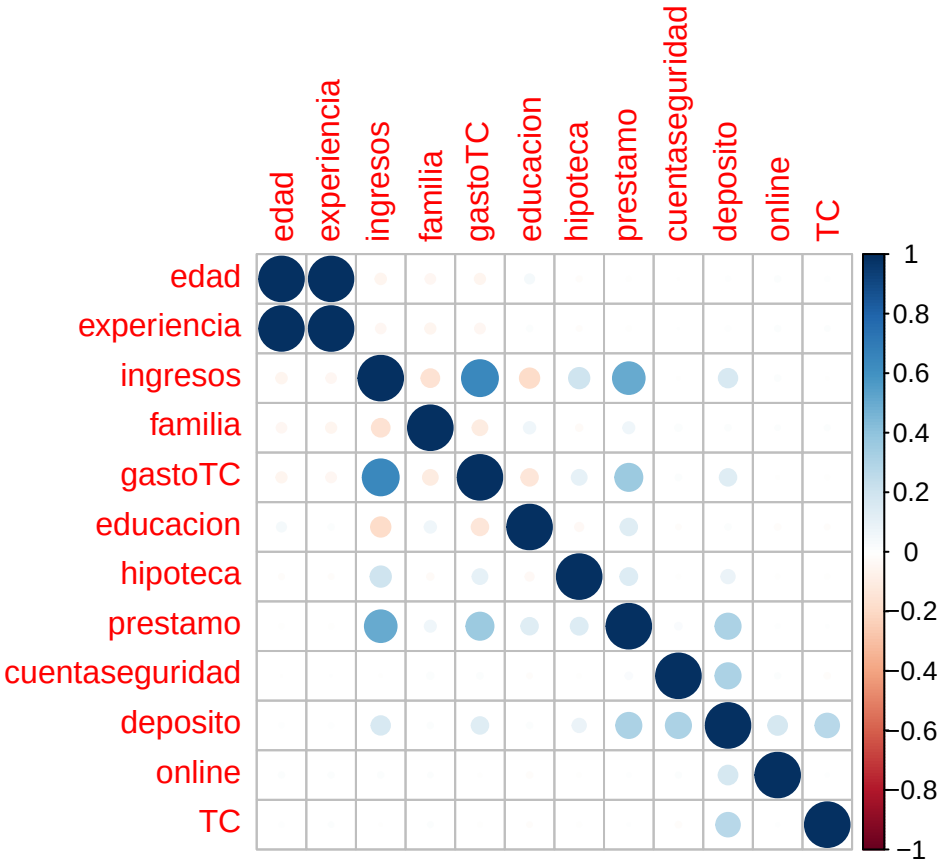
Vemos que hay correlación entre los ingresos y los gastos mensuales en la tarjeta de crédito. También que a mayor ingresos y gatos, mayor posibilidad de aceptación del préstamo. Esto es bastante sorprendente a mi juicio.

Podemos corroborar este análisis realizando el cálculo del coeficiente de correlación de Pearson para las variables de este dataset. Para ello podemos usar la función cor y la librería gráfica corrplot para visualizar los resultados gráficamente:

```
library(corrplot)

correlaciones <- cor(bd)

corrplot(correlaciones, method="circle")
```



Comprobamos que, aparte de la correlación alta entre edad y experiencia (esto parece evidente), las variables más correlacionadas son ingresos y gastoTC, préstamo e ingresos, y préstamo y gastoTC.

Regresión logística

¿Podemos considerar que tener un alto gasto en tarjeta de crédito implica la aceptación del préstamo? Estimamos el modelo de regresión logística tomando como variable dependiente, aceptar el préstamo o no y siendo la variable explicativa, tener un alto gasto en tarjeta de crédito o no.

A traves de la funcion glm() creamos el modelo logistico:

```
rlog <- glm(formula = prestamo ~ gastoTC, family = binomial(), data = bd)
summary(rlog)
```

```
##
## Call:
## glm(formula = prestamo ~ gastoTC, family = binomial(), data = bd)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5832  -0.4050  -0.3241  -0.2587   2.6879
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.58499    0.09129  -39.27   <2e-16 ***
## gastoTC      0.51158    0.02322   22.03   <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: 3162.0  on 4999  degrees of freedom
## Residual deviance: 2653.1  on 4998  degrees of freedom
## AIC: 2657.1
##
## Number of Fisher Scoring iterations: 5
```

Como podemos ver el regresor gastoTCalto tiene una influencia significativa en el modelo, a mas del 99.9% (***). Podemos considerar que el hecho tener un alto gato en tarjeta de crédito, impacta en la aceptación del préstamo.

Añadimos al modelo anterior las variable continua ingresos. ¿Se observa una mejora del modelo?

A traves del modelo de la funcion glm() creamos el modelo logistico añadiendo la nueva variable ingresos:

```
rlog_b <- glm(formula = prestamo ~ gastoTC + ingresos, family = binomial(), data = bd)
summary(rlog_b)
```

```
##
## Call:
## glm(formula = prestamo ~ gastoTC + ingresos, family = binomial(),
##     data = bd)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1652  -0.3013  -0.1763  -0.1134   2.7683
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -6.129701   0.186192 -32.921   <2e-16 ***
## gastoTC      0.063435   0.028414   2.233   0.0256 *
## ingresos     0.035490   0.001553  22.855   <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: 3162.0  on 4999  degrees of freedom
## Residual deviance: 2011.2  on 4997  degrees of freedom
## AIC: 2017.2
##
## Number of Fisher Scoring iterations: 6
```

Para comprobar si hemos mejorado el modelo, además de mirar la influencia significativa de los regresores, podemos comprobar el valor de AIC (Akaike Information Criteria): Este valor es el equivalente a R2 para regresiones logísticas. Mide el ajuste. Mientras menor sea el valor, mas cercano estará el modelo de la verdad. Comprobamos que con la adición de ingresos, hemos bajado de 2657.1 a 2017.2 y por lo tanto hemos mejorado el modelo.

Contraste de hipótesis

La última prueba estadística consistirá en un contraste de hipótesis sobre dos muestras para determinar si el cliente tiene gasta mas en la tarjeta de crédito dependiendo del nivel de ingresos. Para ello, tendremos dos muestras: la primera de ellas se corresponderá con los clientes con un nivel de ingresos de mas de 100K dolares y, por otro lado, los clientes con un nivel de ingresos inferiror o igual a 100K dolares.

Se debe destacar que un test paramétrico como el que a continuación se utiliza necesita que los datos sean normales, si la muestra es de tamaño inferior a 30. Como en nuestro caso, n $>$ 30, el contraste de hipótesis siguiente es válido:

```
bd_altosingresos <- bd[bd$ingresos > 100,]$gastoTC

bd_noaltosingresos <- bd[bd$ingresos <= 100,]$gastoTC
```

Así, se plantea el siguiente contraste de hipótesis de dos muestras sobre la diferencia de medias, el cual es unilateral atendiendo a la formulación de la hipótesis alternativa:

H0 : $\ 1 - \ 2 = 0$ H1 : $\ 1 - \ 2 < 0$

donde $\ 1$ es la media de la población de la que se extrae la primera muestra y $\ 2$ es la media de la población de la que extrae la segunda. Así, tomaremos alpha $= 0,05$.

```
t.test(bd_altosingresos, bd_noaltosingresos,alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  bd_altosingresos and bd_noaltosingresos
## t = 32.923, df = 1356.6, p-value = 1
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##      -Inf 2.395288
## sample estimates:
## mean of x mean of y
##  3.666205  1.384966
```

Puesto que obtenemos un p-valor menor que el valor de significación fijado, rechazamos la hipótesis nula. Por tanto, podemos concluir que, efectivamente, el nivel de ingresos influye en el gasto de la tarjeta de crédito.

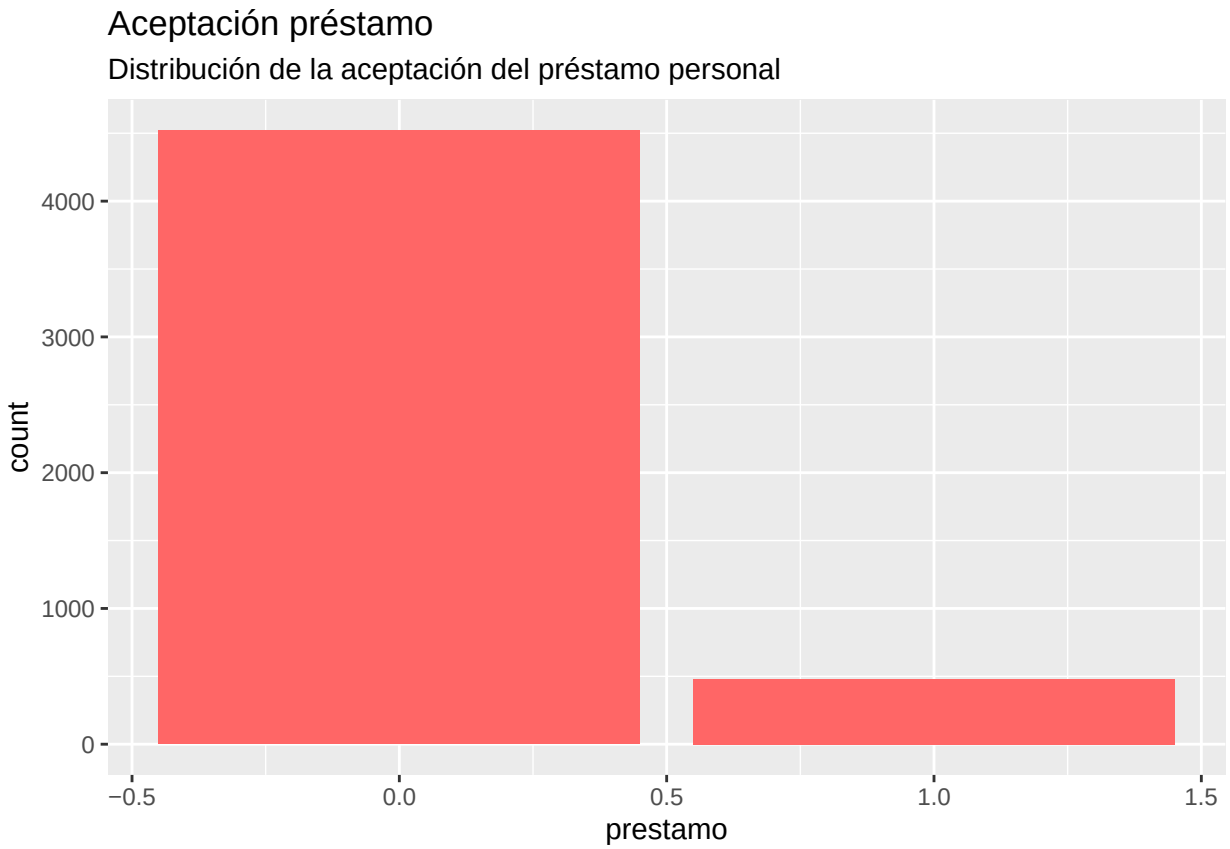
Representación de los resultados a partir de tablas y gráficas

Para empezar veamos cuantos clientes han aceptado el préstamo personal:

```
table(bd$prestamo)
```

```
##
##    0    1
## 4520  480
```

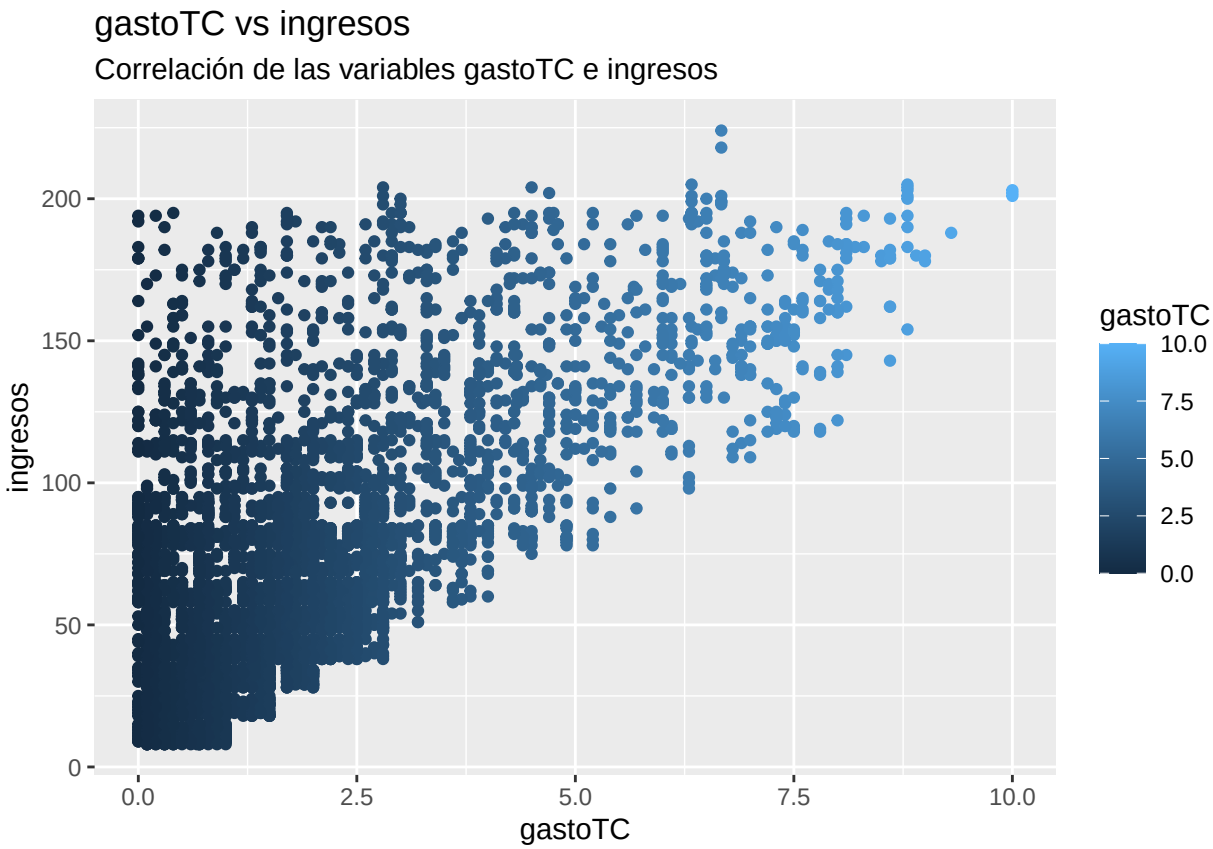
```
ggplot(bd, aes(x = prestamo)) + geom_bar(fill = "#FF6666") + labs(title = "Aceptación préstamo", subtit
```



Vemos que de los 5000 clientes, solo han aceptado el préstamo personal 480. Aproximadamente un 9.6%.

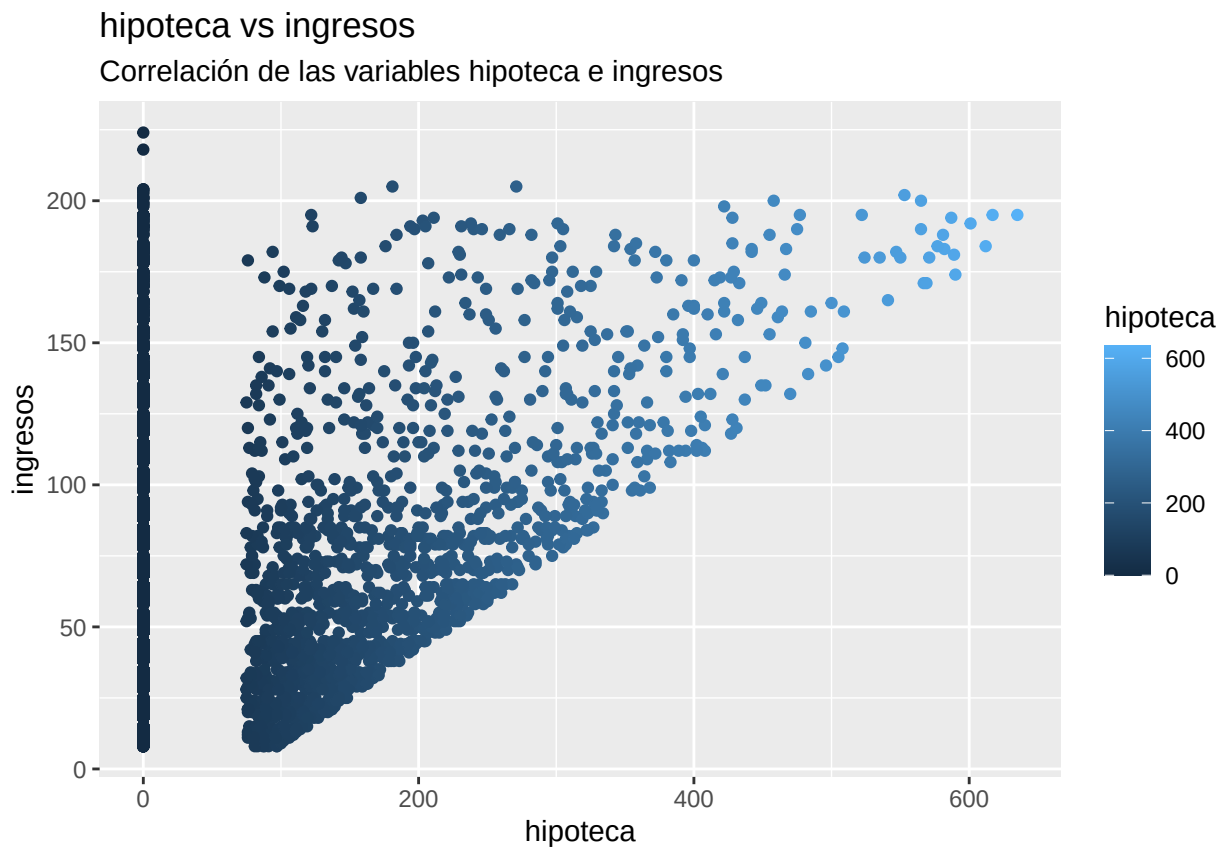
Visualizamos el gasto TC vs ingresos:

```
ggplot(bd, aes(x = gastoTC, y = ingresos)) + geom_point(aes(color = gastoTC)) + labs(title = "gastoTC vs
```



A continuación visualizamos el gasto hipotecario vs ingresos:

```
ggplot(bd, aes(x = hipoteca, y = ingresos)) + geom_point(aes(color = hipoteca)) + labs(title = "hipoteca
```



En ambas gráficas podemos ver que a medida que suben los gatos de tarjeta de crédito y de hipoteca, también lo hacen los ingresos anuales de los clientes, lo cual nos corrobora el pensamiento inicial de que no debían tratarse dichos outliers ya que serían perfectamente posibles.

Conclusiones

Podemos concluir que los clientes que han aceptado el préstamo personal tienen una distribución de ingresos muy diferente a aquellos que no han aceptado el préstamo. También que a mayor ingresos y gatos en tarjeta de crédito, mayor posibilidad de aceptación del préstamo.

Hemos estimado el modelo de regresión logística tomando como variable dependiente, aceptar el préstamo o no y siendo la variable explicativa, tener un alto gasto o no en tarjeta de crédito y la variable continua ingresos.

Por último haciendo uso de un contraste de hipótesis, podemos concluir que el nivel de ingresos influye en el gasto de la tarjeta de crédito.

Exportación del código en R y de los datos producidos

El código en R esta incluido en este fichero con extensión rmd.

Los datos de salida se exportan mediante el siguiente comando y pueden ser descargados desde en GitHub desde la siguiente dirección:

https://github.com/josem-gomez/LoanModeling--Rstudio/blob/master/data/Bank_Personal_Loan_Modelling_output.csv

```r
write.csv(bd, file = "Bank_personal_Loan_Modelling_output.csv")
```