Code

* [Show All Code](#gjdgxs)
* [Hide All Code](#gjdgxs)

# Práctica 2: Limpieza y análisis de datos

David Dávila y Mónica Gómez

21/12/2020

* [1 Limpieza de los datos](#30j0zll)
  + [1.1 Carga del archivo](#1fob9te)
  + [1.2 Análisis del archivo](#3znysh7)
  + [1.3 Selección de los datos de interés](#2et92p0)
  + [1.4 Normalización de los datos cuantitativos y cualitativos](#tyjcwt)
  + [1.5 Ceros y elementos vacíos](#3dy6vkm)
  + [1.6 Valores extremos](#1t3h5sf)
    - [1.6.1 Customer\_Age](#4d34og8)
    - [1.6.2 Months\_on\_book](#2s8eyo1)
    - [1.6.3 Months\_Inactive\_12\_mon](#17dp8vu)
    - [1.6.4 Contacts\_Count\_12\_mon](#3rdcrjn)
    - [1.6.5 Credit\_Limit](#26in1rg)
    - [1.6.6 Avg\_Open\_To\_Buy](#lnxbz9)
    - [1.6.7 Total\_Amt\_Chng\_Q4\_Q1](#35nkun2)
    - [1.6.8 Total\_Trans\_Amt](#1ksv4uv)
    - [1.6.9 Total\_Trans\_Ct](#44sinio)
    - [1.6.10 Total\_Ct\_Chng\_Q4\_Q1](#2jxsxqh)
  + [1.7 Exportación de los datos preprocesados](#z337ya)

getwd()

## [1] "C:/Users/David/Dropbox/Data Science/Máster Universitario en Ciencia de Datos/Materias/Semestre 2/Tipología y ciclo de vida de los datos/Prácticas/PRA2/Entrega"

1 Limpieza de los datos

1.1 Carga del archivo

getwd()

## [1] "C:/Users/David/Dropbox/Data Science/Máster Universitario en Ciencia de Datos/Materias/Semestre 2/Tipología y ciclo de vida de los datos/Prácticas/PRA2/Entrega"

Se procede a abrir el archivo formato .csv y examinar el tipo de datos con los que R ha interpretado cada variable.

mydata <- data.frame(read.csv("C:\\Users\\David\\Dropbox\\Data Science\\Máster Universitario en Ciencia de Datos\\Materias\\Semestre 2\\Tipología y ciclo de vida de los datos\\Prácticas\\PRA2\\Entrega\\data\\BankChurners.csv", header=TRUE))  
df <- data.frame(mydata)  
attach(df)  
head(df)

1.2 Análisis del archivo

Reviso el resumen del dataframe:

summary(df)

## CLIENTNUM Attrition\_Flag Customer\_Age Gender   
## Min. :708082083 Length:10127 Min. :26.00 Length:10127   
## 1st Qu.:713036770 Class :character 1st Qu.:41.00 Class :character   
## Median :717926358 Mode :character Median :46.00 Mode :character   
## Mean :739177606 Mean :46.33   
## 3rd Qu.:773143533 3rd Qu.:52.00   
## Max. :828343083 Max. :73.00   
## Dependent\_count Education\_Level Marital\_Status Income\_Category   
## Min. :0.000 Length:10127 Length:10127 Length:10127   
## 1st Qu.:1.000 Class :character Class :character Class :character   
## Median :2.000 Mode :character Mode :character Mode :character   
## Mean :2.346   
## 3rd Qu.:3.000   
## Max. :5.000   
## Card\_Category Months\_on\_book Total\_Relationship\_Count  
## Length:10127 Min. :13.00 Min. :1.000   
## Class :character 1st Qu.:31.00 1st Qu.:3.000   
## Mode :character Median :36.00 Median :4.000   
## Mean :35.93 Mean :3.813   
## 3rd Qu.:40.00 3rd Qu.:5.000   
## Max. :56.00 Max. :6.000   
## Months\_Inactive\_12\_mon Contacts\_Count\_12\_mon Credit\_Limit   
## Min. :0.000 Min. :0.000 Min. : 1438   
## 1st Qu.:2.000 1st Qu.:2.000 1st Qu.: 2555   
## Median :2.000 Median :2.000 Median : 4549   
## Mean :2.341 Mean :2.455 Mean : 8632   
## 3rd Qu.:3.000 3rd Qu.:3.000 3rd Qu.:11068   
## Max. :6.000 Max. :6.000 Max. :34516   
## Total\_Revolving\_Bal Avg\_Open\_To\_Buy Total\_Amt\_Chng\_Q4\_Q1 Total\_Trans\_Amt  
## Min. : 0 Min. : 3 Min. :0.0000 Min. : 510   
## 1st Qu.: 359 1st Qu.: 1324 1st Qu.:0.6310 1st Qu.: 2156   
## Median :1276 Median : 3474 Median :0.7360 Median : 3899   
## Mean :1163 Mean : 7469 Mean :0.7599 Mean : 4404   
## 3rd Qu.:1784 3rd Qu.: 9859 3rd Qu.:0.8590 3rd Qu.: 4741   
## Max. :2517 Max. :34516 Max. :3.3970 Max. :18484   
## Total\_Trans\_Ct Total\_Ct\_Chng\_Q4\_Q1 Avg\_Utilization\_Ratio  
## Min. : 10.00 Min. :0.0000 Min. :0.0000   
## 1st Qu.: 45.00 1st Qu.:0.5820 1st Qu.:0.0230   
## Median : 67.00 Median :0.7020 Median :0.1760   
## Mean : 64.86 Mean :0.7122 Mean :0.2749   
## 3rd Qu.: 81.00 3rd Qu.:0.8180 3rd Qu.:0.5030   
## Max. :139.00 Max. :3.7140 Max. :0.9990   
## Naive\_Bayes\_Classifier\_Attrition\_Flag\_Card\_Category\_Contacts\_Count\_12\_mon\_Dependent\_count\_Education\_Level\_Months\_Inactive\_12\_mon\_1  
## Min. :0.0000077   
## 1st Qu.:0.0000990   
## Median :0.0001815   
## Mean :0.1599975   
## 3rd Qu.:0.0003373   
## Max. :0.9995800   
## Naive\_Bayes\_Classifier\_Attrition\_Flag\_Card\_Category\_Contacts\_Count\_12\_mon\_Dependent\_count\_Education\_Level\_Months\_Inactive\_12\_mon\_2  
## Min. :0.00042   
## 1st Qu.:0.99966   
## Median :0.99982   
## Mean :0.84000   
## 3rd Qu.:0.99990   
## Max. :0.99999

Reviso la estructura del dataframe:

str(df)

## 'data.frame': 10127 obs. of 23 variables:  
## $ CLIENTNUM : int 768805383 818770008 713982108 769911858 709106358 713061558 810347208 818906208 710930508 719661558 ...  
## $ Attrition\_Flag : chr "Existing Customer" "Existing Customer" "Existing Customer" "Existing Customer" ...  
## $ Customer\_Age : int 45 49 51 40 40 44 51 32 37 48 ...  
## $ Gender : chr "M" "F" "M" "F" ...  
## $ Dependent\_count : int 3 5 3 4 3 2 4 0 3 2 ...  
## $ Education\_Level : chr "High School" "Graduate" "Graduate" "High School" ...  
## $ Marital\_Status : chr "Married" "Single" "Married" "Unknown" ...  
## $ Income\_Category : chr "$60K - $80K" "Less than $40K" "$80K - $120K" "Less than $40K" ...  
## $ Card\_Category : chr "Blue" "Blue" "Blue" "Blue" ...  
## $ Months\_on\_book : int 39 44 36 34 21 36 46 27 36 36 ...  
## $ Total\_Relationship\_Count : int 5 6 4 3 5 3 6 2 5 6 ...  
## $ Months\_Inactive\_12\_mon : int 1 1 1 4 1 1 1 2 2 3 ...  
## $ Contacts\_Count\_12\_mon : int 3 2 0 1 0 2 3 2 0 3 ...  
## $ Credit\_Limit : num 12691 8256 3418 3313 4716 ...  
## $ Total\_Revolving\_Bal : int 777 864 0 2517 0 1247 2264 1396 2517 1677 ...  
## $ Avg\_Open\_To\_Buy : num 11914 7392 3418 796 4716 ...  
## $ Total\_Amt\_Chng\_Q4\_Q1 : num 1.33 1.54 2.59 1.4 2.17 ...  
## $ Total\_Trans\_Amt : int 1144 1291 1887 1171 816 1088 1330 1538 1350 1441 ...  
## $ Total\_Trans\_Ct : int 42 33 20 20 28 24 31 36 24 32 ...  
## $ Total\_Ct\_Chng\_Q4\_Q1 : num 1.62 3.71 2.33 2.33 2.5 ...  
## $ Avg\_Utilization\_Ratio : num 0.061 0.105 0 0.76 0 0.311 0.066 0.048 0.113 0.144 ...  
## $ Naive\_Bayes\_Classifier\_Attrition\_Flag\_Card\_Category\_Contacts\_Count\_12\_mon\_Dependent\_count\_Education\_Level\_Months\_Inactive\_12\_mon\_1: num 9.34e-05 5.69e-05 2.11e-05 1.34e-04 2.17e-05 ...  
## $ Naive\_Bayes\_Classifier\_Attrition\_Flag\_Card\_Category\_Contacts\_Count\_12\_mon\_Dependent\_count\_Education\_Level\_Months\_Inactive\_12\_mon\_2: num 1 1 1 1 1 ...

1.3 Selección de los datos de interés

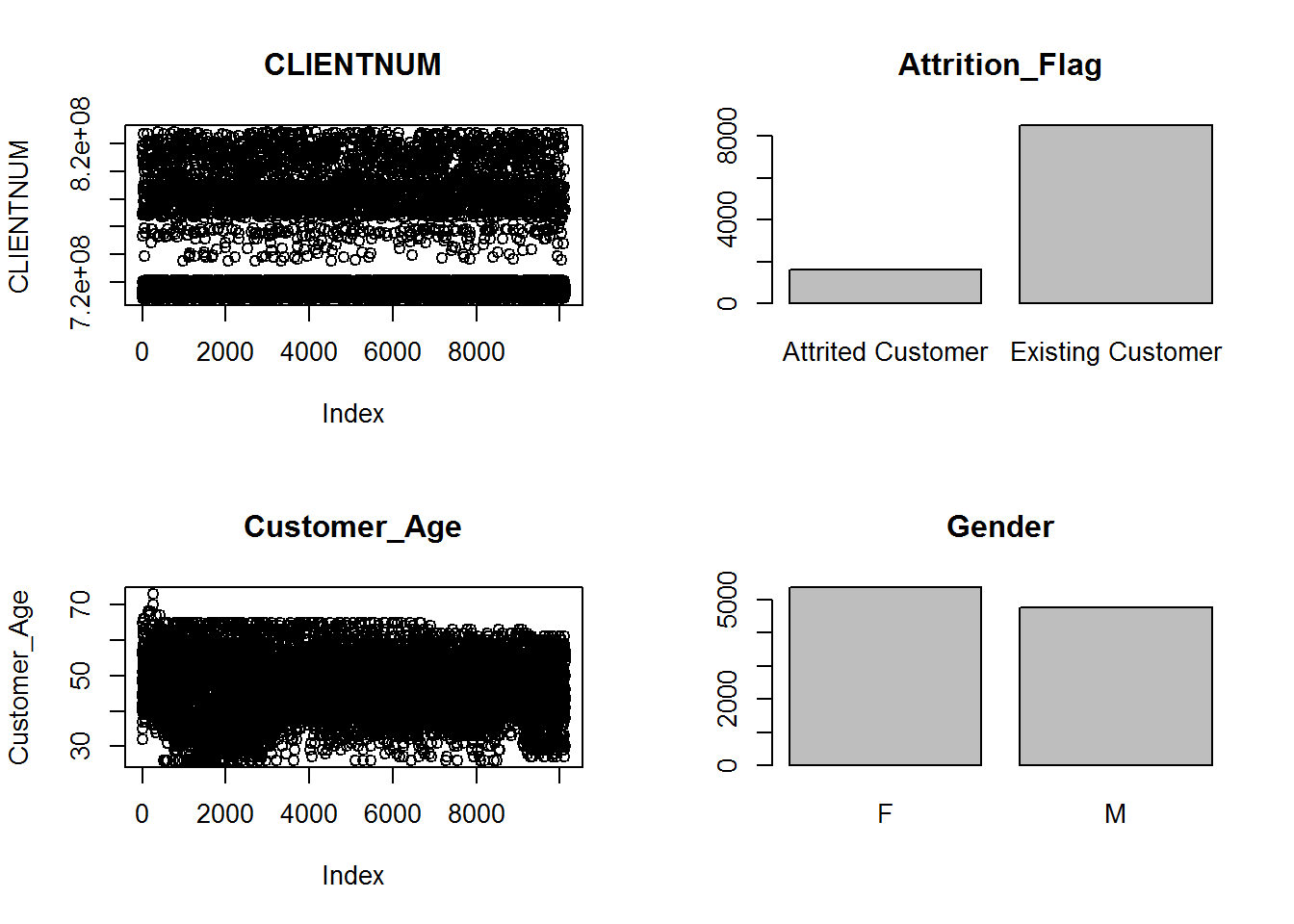
Para el presente proyecto se utilizarán todas las variables presentes en el juego de datos a excepción de las dos últimas que no serán de utilidad en los análisis que se planean hacer a posteriori.

df <- df[,-(22:23)]  
names(df)

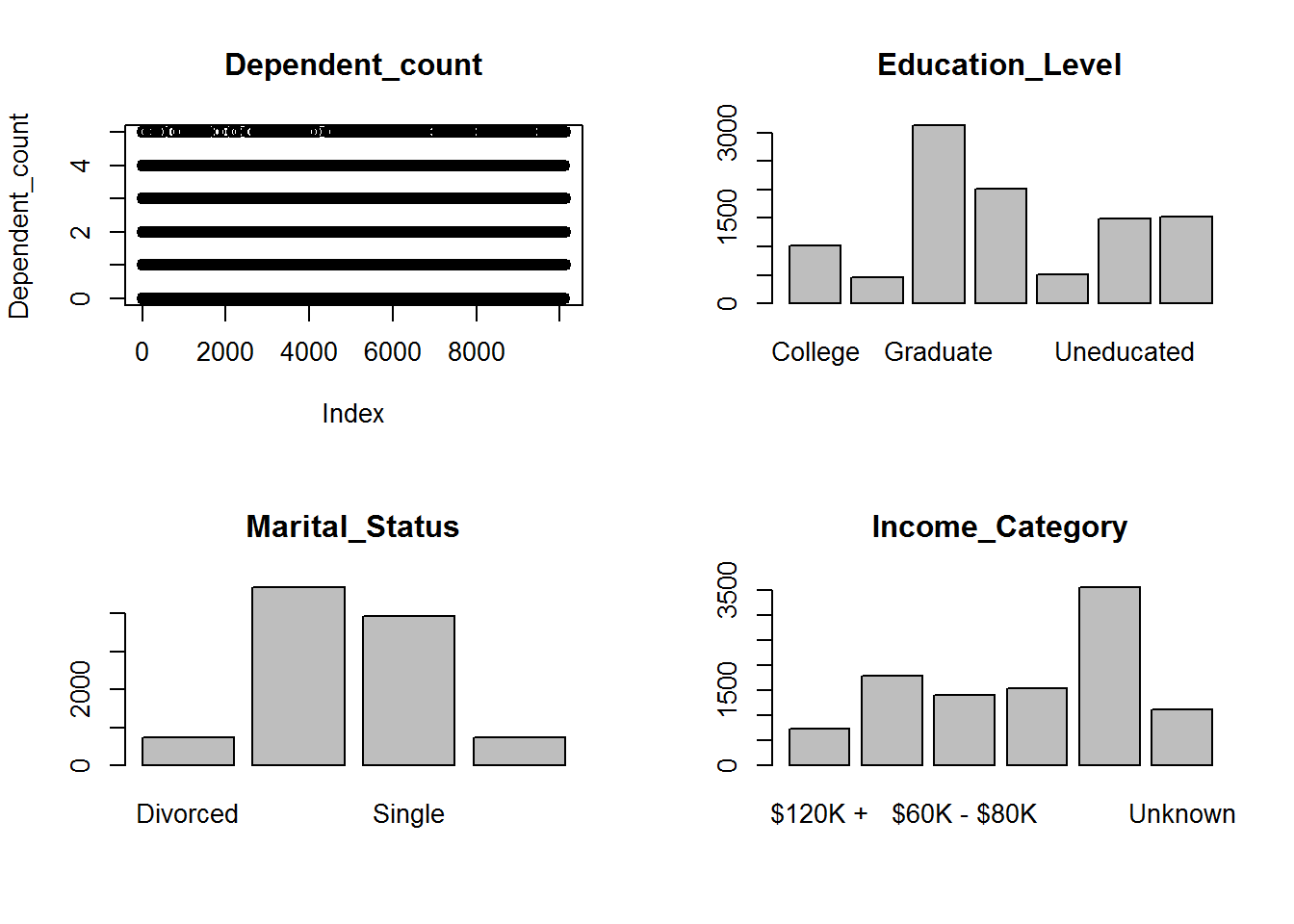
## [1] "CLIENTNUM" "Attrition\_Flag"   
## [3] "Customer\_Age" "Gender"   
## [5] "Dependent\_count" "Education\_Level"   
## [7] "Marital\_Status" "Income\_Category"   
## [9] "Card\_Category" "Months\_on\_book"   
## [11] "Total\_Relationship\_Count" "Months\_Inactive\_12\_mon"   
## [13] "Contacts\_Count\_12\_mon" "Credit\_Limit"   
## [15] "Total\_Revolving\_Bal" "Avg\_Open\_To\_Buy"   
## [17] "Total\_Amt\_Chng\_Q4\_Q1" "Total\_Trans\_Amt"   
## [19] "Total\_Trans\_Ct" "Total\_Ct\_Chng\_Q4\_Q1"   
## [21] "Avg\_Utilization\_Ratio"

Elaboro gráficas por variable para tener un primer acercamiento sobre su comportamiento

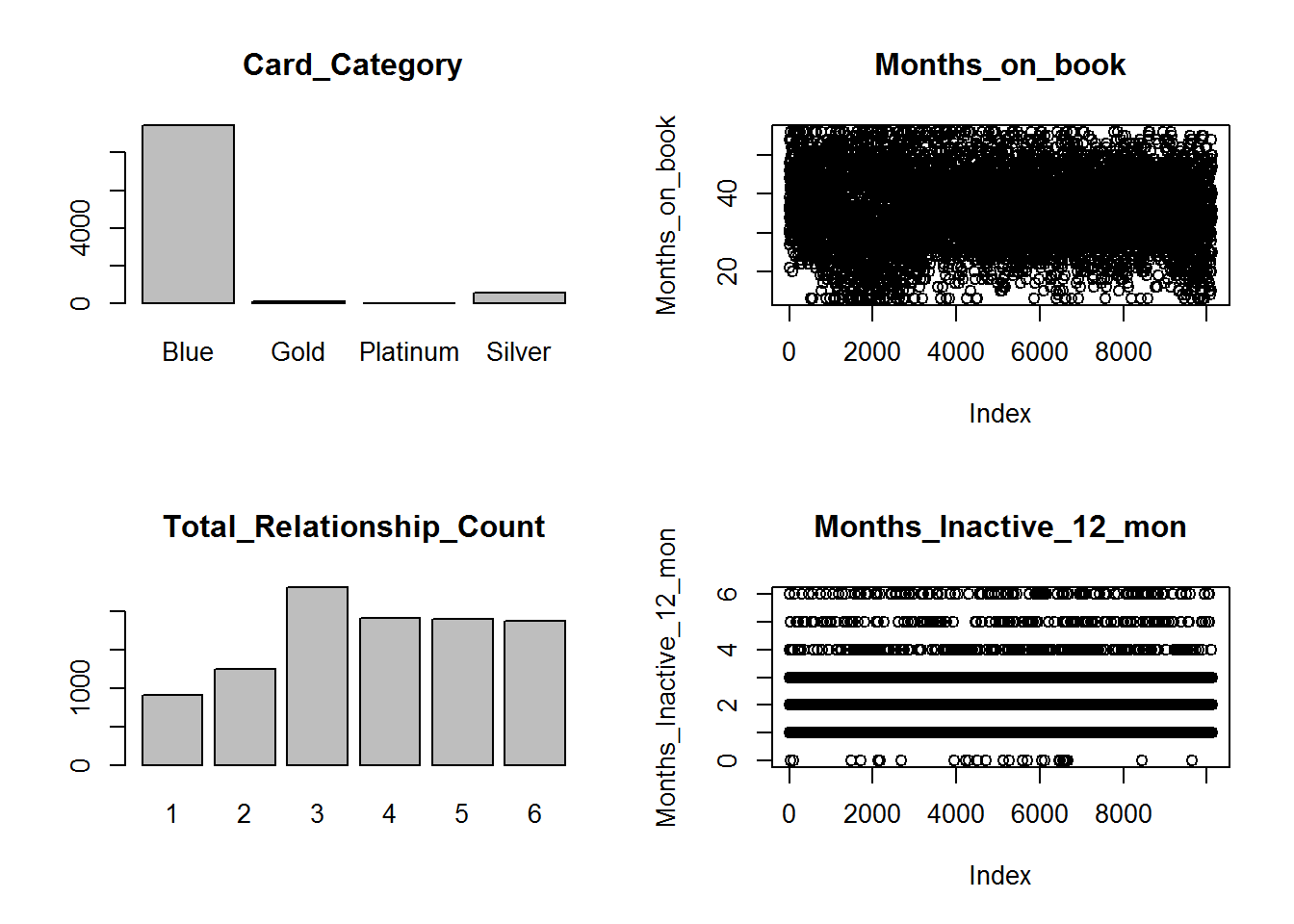
library(ggplot2)  
  
par(mfrow=c(2,2))  
  
p1 <- plot(CLIENTNUM, main="CLIENTNUM")  
p2 <- plot(as.factor(Attrition\_Flag), main="Attrition\_Flag")  
p3 <- plot(Customer\_Age, main="Customer\_Age")  
p4 <- plot(as.factor(Gender), main="Gender")



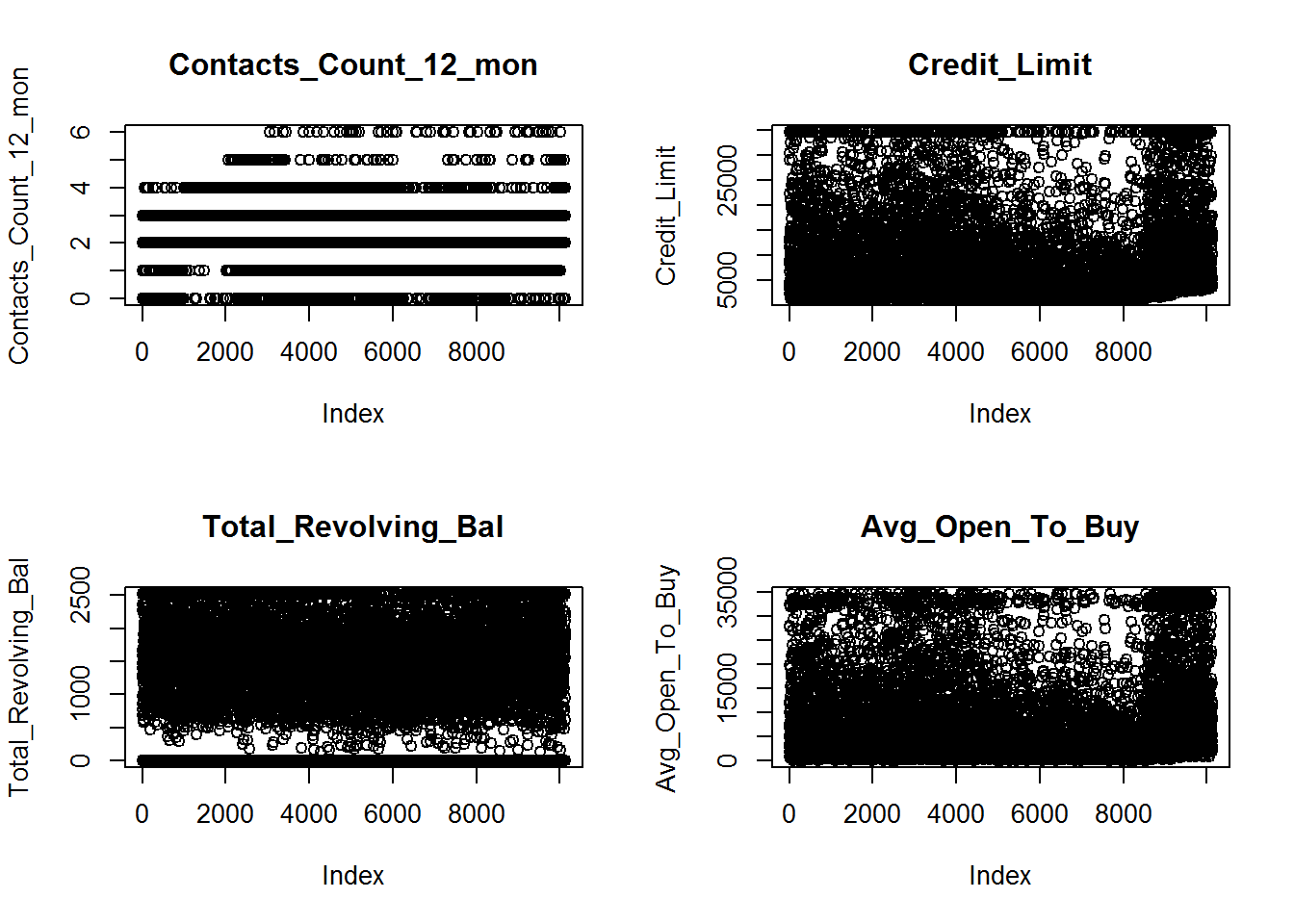
p5 <- plot(Dependent\_count, main="Dependent\_count")  
p6 <- plot(as.factor(Education\_Level), main="Education\_Level")  
p7 <- plot(as.factor(Marital\_Status), main="Marital\_Status")   
p8 <- plot(as.factor(Income\_Category), main="Income\_Category")



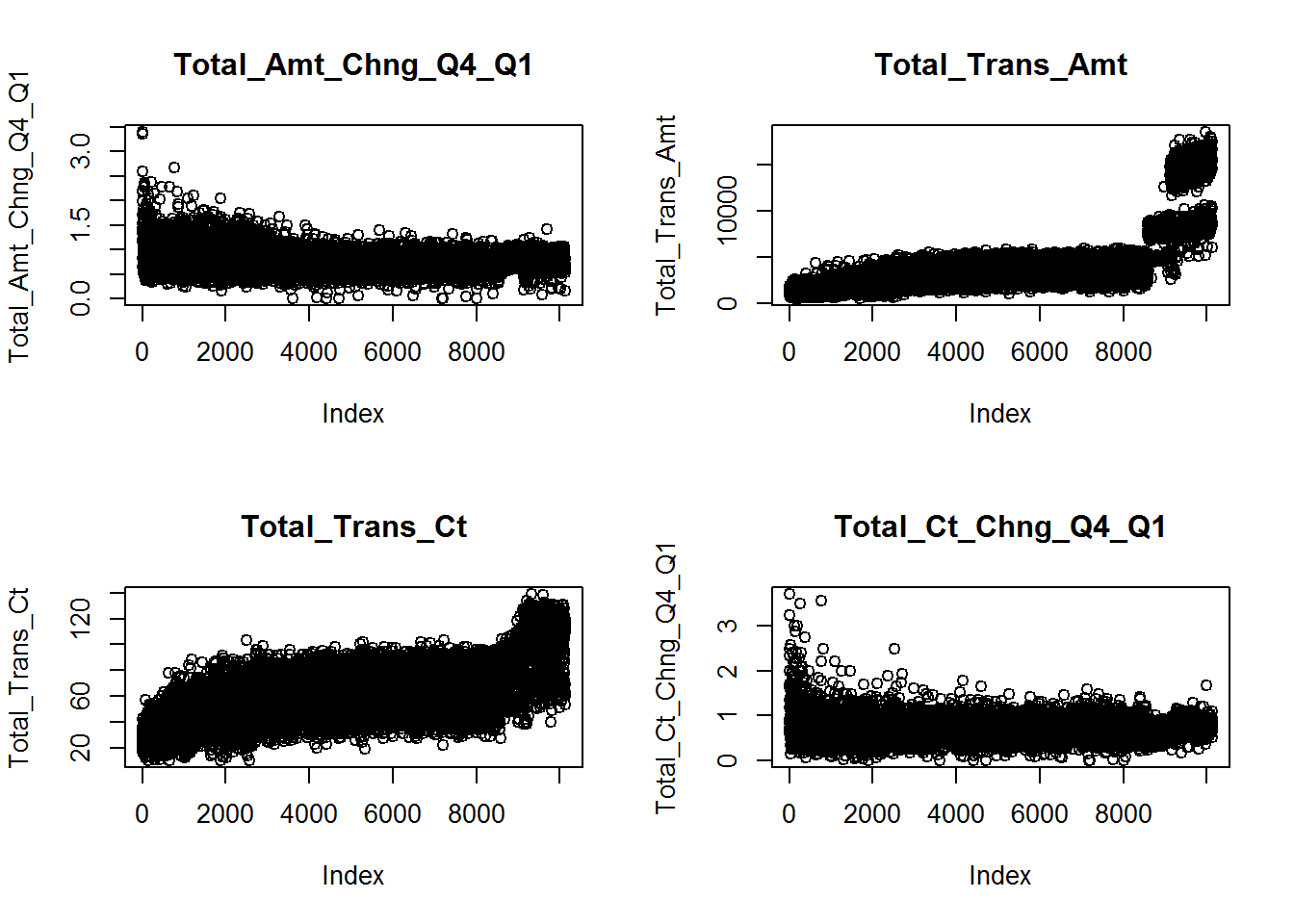
p9 <- plot(as.factor(Card\_Category), main="Card\_Category")   
p10 <- plot(Months\_on\_book, main="Months\_on\_book")  
p11 <- plot(as.factor(Total\_Relationship\_Count), main="Total\_Relationship\_Count")  
p12 <- plot(Months\_Inactive\_12\_mon, main="Months\_Inactive\_12\_mon")



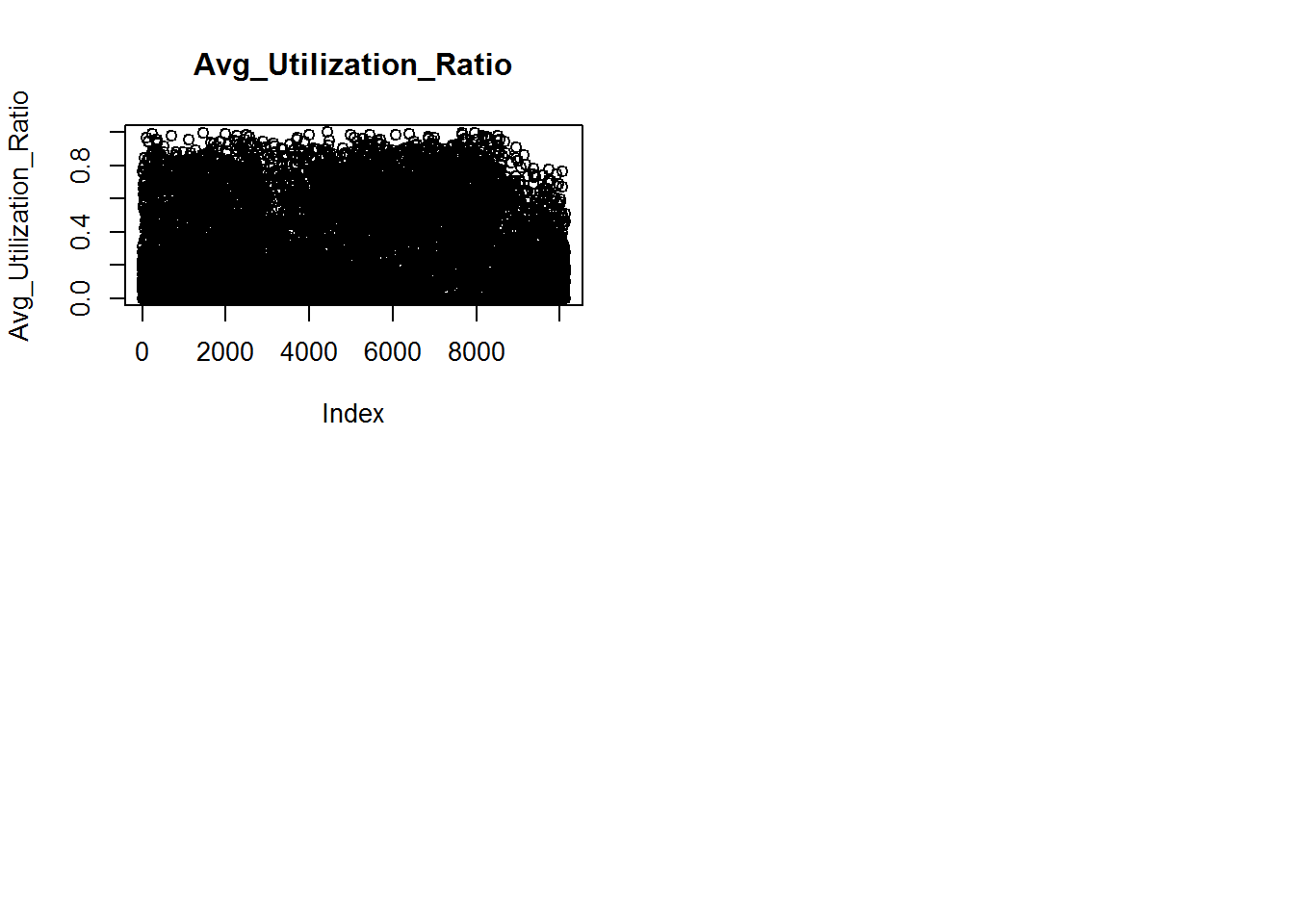
p13 <- plot(Contacts\_Count\_12\_mon, main="Contacts\_Count\_12\_mon")   
p14 <- plot(Credit\_Limit, main="Credit\_Limit")   
p15 <- plot(Total\_Revolving\_Bal, main="Total\_Revolving\_Bal")   
p16 <- plot(Avg\_Open\_To\_Buy, main="Avg\_Open\_To\_Buy")



p17 <- plot(Total\_Amt\_Chng\_Q4\_Q1, main="Total\_Amt\_Chng\_Q4\_Q1")   
p18 <- plot(Total\_Trans\_Amt, main="Total\_Trans\_Amt")   
p19 <- plot(Total\_Trans\_Ct, main="Total\_Trans\_Ct")   
p20 <- plot(Total\_Ct\_Chng\_Q4\_Q1, main="Total\_Ct\_Chng\_Q4\_Q1")



p21 <- plot(Avg\_Utilization\_Ratio, main="Avg\_Utilization\_Ratio")   
#p22 <- plot(Naive\_Bayes\_Classifier\_Attrition\_Flag\_Card\_Category\_Contacts\_Count\_12\_mon\_Dependent\_count\_Education\_Level\_Months\_Inactive\_12\_mon\_1, main="Naive\_Bayes\_Classifier\_1")   
#p23 <- plot(Naive\_Bayes\_Classifier\_Attrition\_Flag\_Card\_Category\_Contacts\_Count\_12\_mon\_Dependent\_count\_Education\_Level\_Months\_Inactive\_12\_mon\_2, main="Naive\_Bayes\_Classifier\_2")



Los gráficos generados nos permiten explorar los datos de forma preliminar. Algunos de los insights que se pueden extraer son:

Variables cuantitativas: presentan una alta dispersión en sus datos.

Variables cuanlitativas: se puede observar que el banco en cuestión presenta una mayor cantidad de clientes con cuentas activas que inactivas, existen más mujeres que hombres, la mayoría están solteros o casados con estudios terminados y que usan la tarjeta de crédito categoría “Blue”.

1.4 Normalización de los datos cuantitativos y cualitativos

Estas normalizaciones tienen como objetivo uniformizar los formatos. En este caso no es necesario normalizarlos. Sin embargo, los valores perdidos o valores extremos (de haberlos), se tratarán más adelante.

Se detallan los tipos de datos por variable:

sapply(df, function(x) class(x))

## CLIENTNUM Attrition\_Flag Customer\_Age   
## "integer" "character" "integer"   
## Gender Dependent\_count Education\_Level   
## "character" "integer" "character"   
## Marital\_Status Income\_Category Card\_Category   
## "character" "character" "character"   
## Months\_on\_book Total\_Relationship\_Count Months\_Inactive\_12\_mon   
## "integer" "integer" "integer"   
## Contacts\_Count\_12\_mon Credit\_Limit Total\_Revolving\_Bal   
## "integer" "numeric" "integer"   
## Avg\_Open\_To\_Buy Total\_Amt\_Chng\_Q4\_Q1 Total\_Trans\_Amt   
## "numeric" "numeric" "integer"   
## Total\_Trans\_Ct Total\_Ct\_Chng\_Q4\_Q1 Avg\_Utilization\_Ratio   
## "integer" "numeric" "numeric"

1.5 Ceros y elementos vacíos

Este dataset en particular presenta una completitud del 100% como se demuestra a continuación:

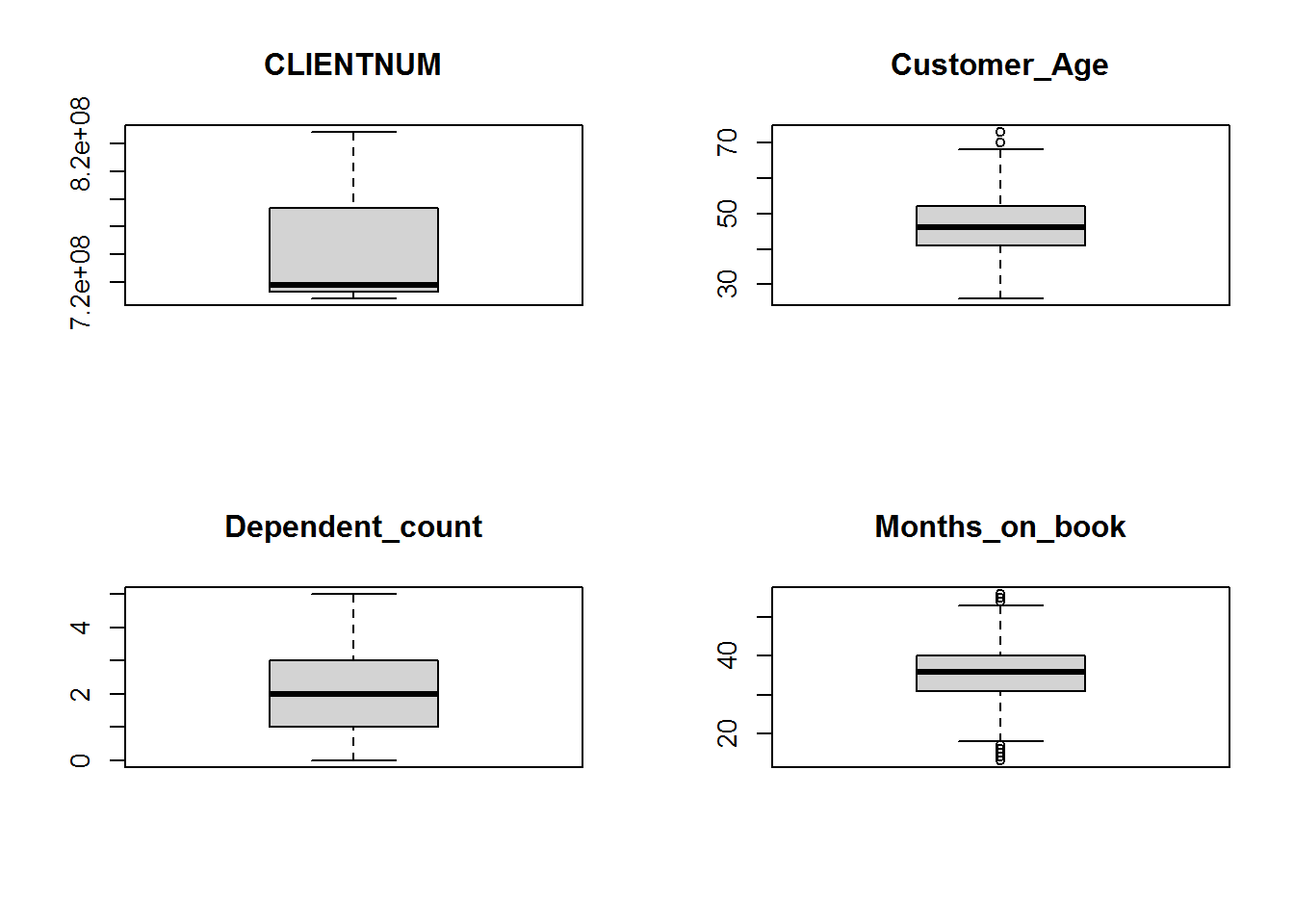
sapply(df, function(x) sum(is.na(x)))

## CLIENTNUM Attrition\_Flag Customer\_Age   
## 0 0 0   
## Gender Dependent\_count Education\_Level   
## 0 0 0   
## Marital\_Status Income\_Category Card\_Category   
## 0 0 0   
## Months\_on\_book Total\_Relationship\_Count Months\_Inactive\_12\_mon   
## 0 0 0   
## Contacts\_Count\_12\_mon Credit\_Limit Total\_Revolving\_Bal   
## 0 0 0   
## Avg\_Open\_To\_Buy Total\_Amt\_Chng\_Q4\_Q1 Total\_Trans\_Amt   
## 0 0 0   
## Total\_Trans\_Ct Total\_Ct\_Chng\_Q4\_Q1 Avg\_Utilization\_Ratio   
## 0 0 0

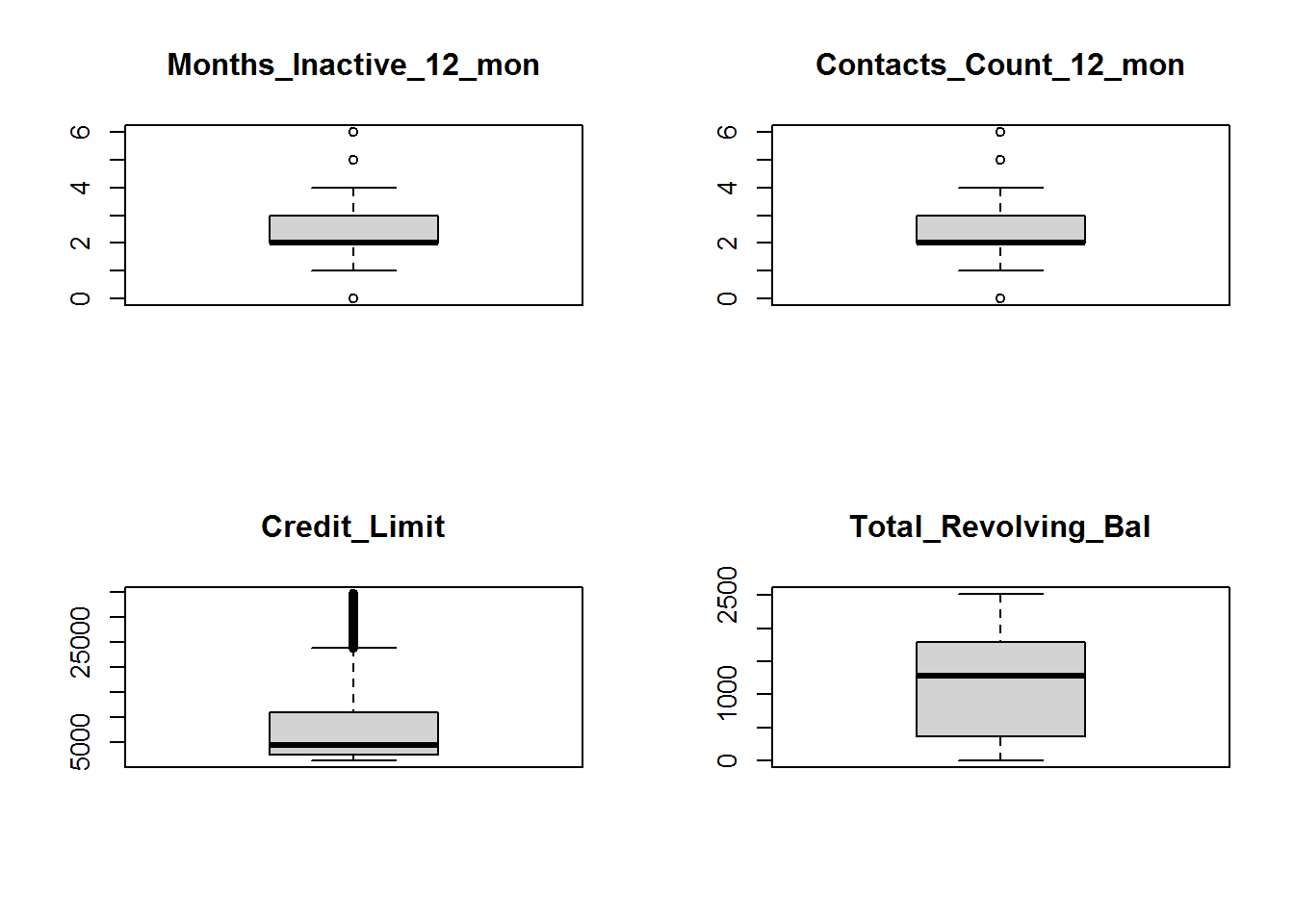
1.6 Valores extremos

Se procede a hacer la exploración de outliers con el diagráma de caja y bigote, para posteriormente de acuerdo a un análisis por cada variable extraer e imputar los valores con la mediana de aquellas variables donde se haya considerado meritorio hacerlo.

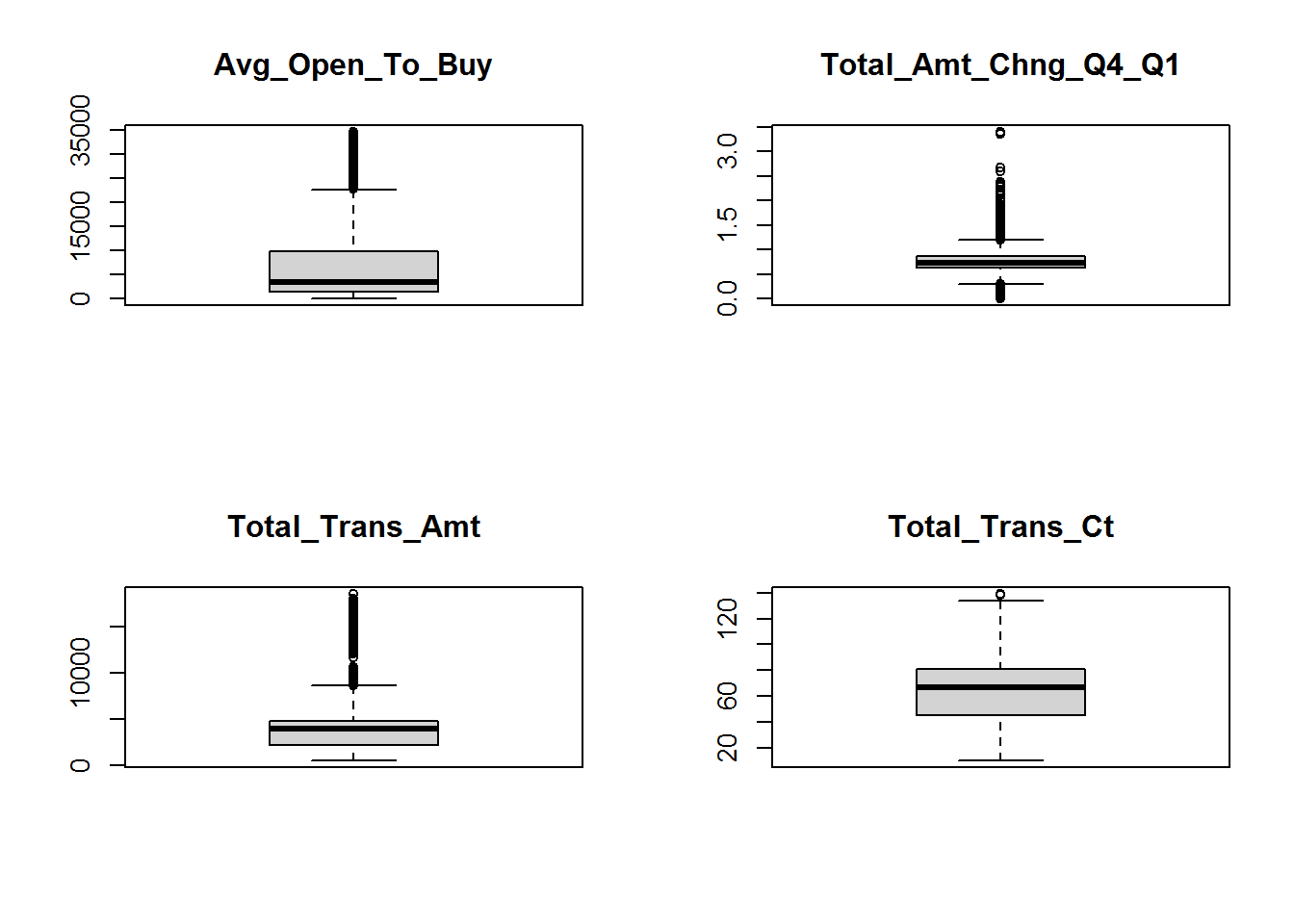
par(mfrow=c(2,2))  
  
p1 <- boxplot(CLIENTNUM, main="CLIENTNUM")  
p3 <- boxplot(Customer\_Age, main="Customer\_Age")  
p5 <- boxplot(Dependent\_count, main="Dependent\_count")  
p10 <- boxplot(Months\_on\_book, main="Months\_on\_book")



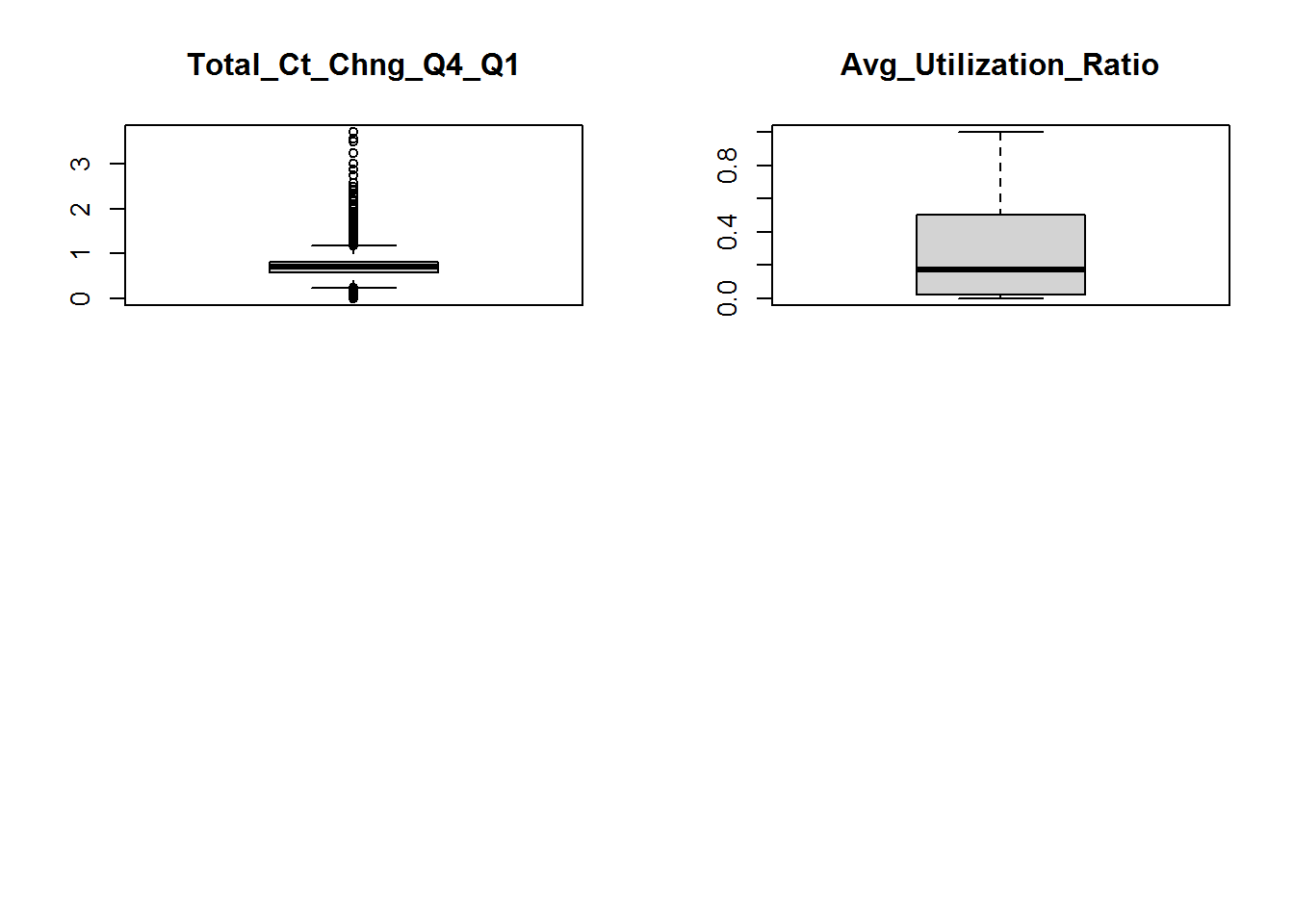
p12 <- boxplot(Months\_Inactive\_12\_mon, main="Months\_Inactive\_12\_mon")  
p13 <- boxplot(Contacts\_Count\_12\_mon, main="Contacts\_Count\_12\_mon")   
p14 <- boxplot(Credit\_Limit, main="Credit\_Limit")   
p15 <- boxplot(Total\_Revolving\_Bal, main="Total\_Revolving\_Bal")



p16 <- boxplot(Avg\_Open\_To\_Buy, main="Avg\_Open\_To\_Buy")   
p17 <- boxplot(Total\_Amt\_Chng\_Q4\_Q1, main="Total\_Amt\_Chng\_Q4\_Q1")   
p18 <- boxplot(Total\_Trans\_Amt, main="Total\_Trans\_Amt")   
p19 <- boxplot(Total\_Trans\_Ct, main="Total\_Trans\_Ct")



p20 <- boxplot(Total\_Ct\_Chng\_Q4\_Q1, main="Total\_Ct\_Chng\_Q4\_Q1")   
p21 <- boxplot(Avg\_Utilization\_Ratio, main="Avg\_Utilization\_Ratio")



De los diagramas se puede extraer la siguiente información:

* Simetría de la distribución de los datos.
* Detectar la presencia de valores atípicos o outliers.
* Ver cómo es la dispersión de los puntos con la mediana, los percentiles 25 y 75 y los valores máximos y mínimos.

Al revisar los diagramas generados se encuentra que:

La distribución no es simétrica para la mayoría de variables, y es simétrica en los casos de: “Customer\_Age”, “Dependent\_count” y “Months\_on\_book”.

1.6.1 Customer\_Age

x<-boxplot.stats(Customer\_Age)$out  
idx <- which(Customer\_Age %in% x)  
ca <- df$Customer\_Age[idx] #Valores atípicos  
min(ca)

## [1] 70

max(ca)

## [1] 73

length(ca)

## [1] 2

ca

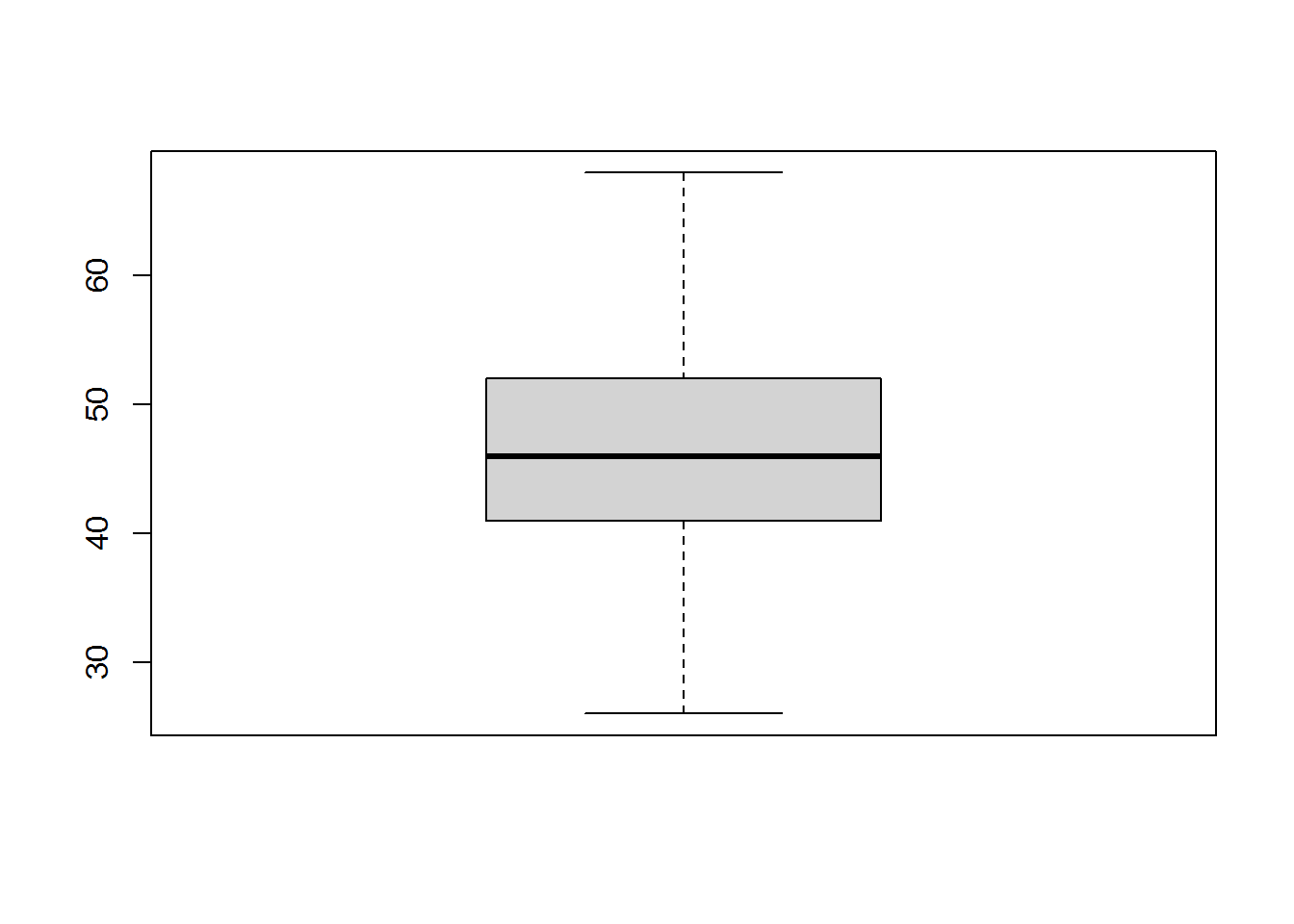
## [1] 73 70

En este caso, es un poco extraño que únicamente 2 de 10 mil clientes superen los 70 años y por ello se puede considerar que se trata de valores que probablemente se ingresaron o calcularon erróneamente. Si se tuviera la fecha de nacimiento sería fácil comprobar la edad real de ese par de clientes, sin embargo al desconocer ese dato se procede a imputar.

df$Customer\_Age[df$Customer\_Age>69] <- NA #Dejamos en NA solo los más extremos  
df$Customer\_Age[idx]

## [1] NA NA

boxplot(df$Customer\_Age)



idx <- which(is.na(df$Customer\_Age))  
length(idx) #número de valores perdidos

## [1] 2

for (i in 1:length(idx)){  
index <- idx[i]  
df[index,]$Customer\_Age <- median(df$Customer\_Age, na.rm=TRUE ) #imputación  
}  
df$Customer\_Age[idx] #mostramos el resultado

## [1] 46 46

1.6.2 Months\_on\_book

x<-boxplot.stats(Months\_on\_book)$out  
idx <- which(Months\_on\_book %in% x)  
mob <- df$Months\_on\_book[idx] #Valores atípicos  
min(mob)

## [1] 13

max(mob)

## [1] 56

length(mob)

## [1] 386

Como se observa, la variable ‘Months\_on\_book’ tiene 386 outliers. Sin embargo, dado que representa el período de pertenencia con la entidad financiera en meses va a ser normal el encontrar clientes muy antiguos (56 meses) o muy recientes (13 meses) por lo que no se los eliminará del dataset.

1.6.3 Months\_Inactive\_12\_mon

x<-boxplot.stats(Months\_Inactive\_12\_mon)$out  
idx <- which(Months\_Inactive\_12\_mon %in% x)  
mi <- df$Months\_Inactive\_12\_mon[idx] #Valores atípicos  
min(mi)

## [1] 0

max(mi)

## [1] 6

length(mi)

## [1] 331

La variable ‘Months\_Inactive\_12\_mon’ tiene 331 outliers. Sin embargo, dado que pueden haber clientes que han permanecido inactivos entre 0-6 meses, este podría ser un parámetro a considerar para saber si el cliente está por suspender o no algún servicio bancario. Por lo tanto, no se los eliminarán del dataset.

1.6.4 Contacts\_Count\_12\_mon

x<-boxplot.stats(Contacts\_Count\_12\_mon)$out  
idx <- which(Contacts\_Count\_12\_mon %in% x)  
cc <- df$Contacts\_Count\_12\_mon[idx] #Valores atípicos  
min(cc)

## [1] 0

max(cc)

## [1] 6

length(cc)

## [1] 629

La variable ‘Contacts\_Count\_12\_mon’ presenta 629 outliers pero dado el contexto de que pueden haber clientes con los cuales no se ha contactado entre 0-6 meses, este podría ser un parámetro a considerar para saber si el cliente está por suspender o no algún servicio bancario. Por lo tanto, no se los eliminarán del dataset.

1.6.5 Credit\_Limit

x<-boxplot.stats(Credit\_Limit)$out  
idx <- which(Credit\_Limit %in% x)  
cl <- df$Credit\_Limit[idx] #Valores atípicos  
min(cl)

## [1] 23848

max(cl)

## [1] 34516

length(cl)

## [1] 984

Dado que el límite de crédito puede variar dependiendo de muchos factores para a un cliente en específico se opta por mantenerlos.

1.6.6 Avg\_Open\_To\_Buy

x<-boxplot.stats(Avg\_Open\_To\_Buy)$out  
idx <- which(Avg\_Open\_To\_Buy %in% x)  
avg <- df$Avg\_Open\_To\_Buy[idx] #Valores atípicos  
min(avg)

## [1] 22664

max(avg)

## [1] 34516

length(avg)

## [1] 963

Dado que la variable ‘Avg\_Open\_To\_Buy’ presenta 963 outliers y representa línea de crédito abierta para compra se mantienen los valores.

1.6.7 Total\_Amt\_Chng\_Q4\_Q1

x<-boxplot.stats(Total\_Amt\_Chng\_Q4\_Q1)$out  
idx <- which(Total\_Amt\_Chng\_Q4\_Q1 %in% x)  
tac <- df$Total\_Amt\_Chng\_Q4\_Q1[idx] #Valores atípicos  
min(tac)

## [1] 0

max(tac)

## [1] 3.397

length(tac)

## [1] 396

La variable ‘Total\_Amt\_Chng\_Q4\_Q1’ tiene 396 outliers y muestra los cambios transaccionales que pueden ser muy variables dependiendo del perfil de cada cliente, así que se opta por mantenerlos.

1.6.8 Total\_Trans\_Amt

x<-boxplot.stats(Total\_Trans\_Amt)$out  
idx <- which(Total\_Trans\_Amt %in% x)  
tta <- df$Total\_Trans\_Amt[idx] #Valores atípicos  
min(tta)

## [1] 8620

max(tta)

## [1] 18484

length(tta)

## [1] 896

La variable ‘Total\_Trans\_Amt’ tiene 896 outliers y muestra la totalidad de cantidad de transacciones que pueden ser muy variables de persona a persona, así que se opta por mantenerlos.

1.6.9 Total\_Trans\_Ct

x<-boxplot.stats(Total\_Trans\_Ct)$out  
idx <- which(Total\_Trans\_Ct %in% x)  
ttc <- df$Total\_Trans\_Ct[idx] #Valores atípicos  
min(ttc)

## [1] 138

max(ttc)

## [1] 139

length(ttc)

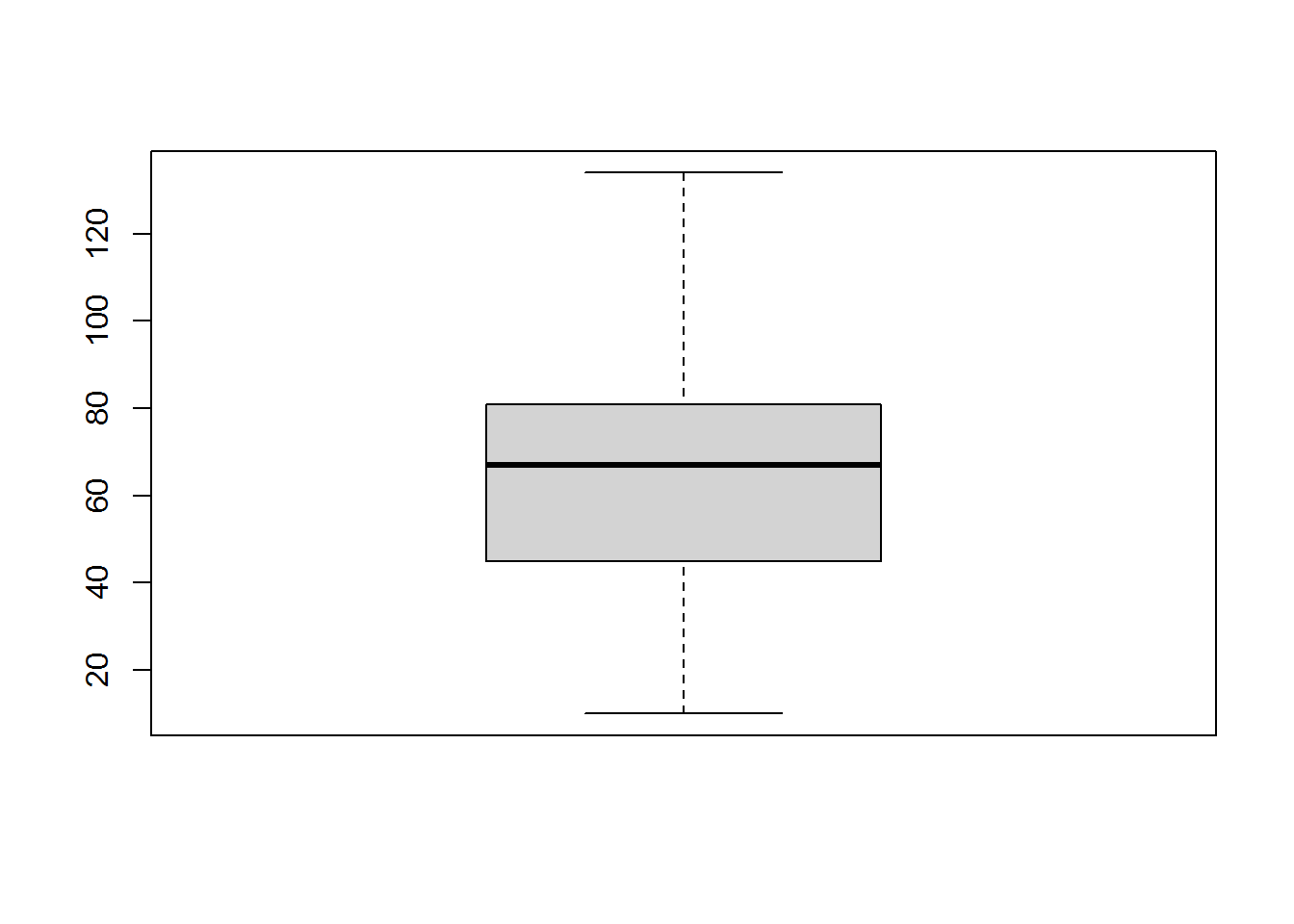
## [1] 2

La variable ‘Total\_Trans\_Ct’ tiene 2 outliers. Es extraño que se existan únicamente dos clientes que durante los últimos 12 meses hayan realizado casi 140 transacciones. Por ello se opta por imputar esos datos.

df$Total\_Trans\_Ct[df$Total\_Trans\_Ct>137] <- NA #Dejamos en NA solo los más extremos  
df$Total\_Trans\_Ct[idx]

## [1] NA NA

boxplot(df$Total\_Trans\_Ct)



idx <- which(is.na(df$Total\_Trans\_Ct))  
length(idx) #número de valores perdidos

## [1] 2

for (i in 1:length(idx)){  
index <- idx[i]  
df[index,]$Total\_Trans\_Ct <- median(df$Total\_Trans\_Ct, na.rm=TRUE ) #imputación  
}  
df$Total\_Trans\_Ct[idx] #mostramos el resultado

## [1] 67 67

1.6.10 Total\_Ct\_Chng\_Q4\_Q1

x<-boxplot.stats(Total\_Ct\_Chng\_Q4\_Q1)$out  
idx <- which(Total\_Ct\_Chng\_Q4\_Q1 %in% x)  
tcc <- df$Total\_Ct\_Chng\_Q4\_Q1[idx] #Valores atípicos  
min(tcc)

## [1] 0

max(tcc)

## [1] 3.714

length(tcc)

## [1] 394

La variable ‘Total\_Ct\_Chng\_Q4\_Q1’ tiene 394 outliers y muestra el cambio en el conteo de las transacciones que pueden ser muy variables de cliente a cliente, así que se opta por mantenerlos.

1.7 Exportación de los datos preprocesados

# Exportación de los datos limpios en formato .csv  
write.csv(df, "BankChurners\_clean.csv")