Preprocessing

FERRAN ARMENGOL Y SALOMÓN SABAL

2025-10-06

Preprocessing

```
#install.packages("dplyr")
library(dplyr)
```

```
train <- read.csv("train.csv", header = TRUE, sep = ",")
test  <- read.csv("test.csv", header = TRUE, sep = ",")
datos <- bind_rows(train, test)

dim(datos)
```

```
## [1] 10000    24
```

```
head(datos)
```

```
##   X Tenure Gender EducationLevel  LoanStatus NetPromoterScore TransactionFrequency    Surname Age Ge
## 1 1     NA Female     University        <NA>               10                   34       <NA>  29
## 2 2      1   <NA>     University     No loan                9                   31 Chiagoziem  41
## 3 3      8   Male     University     No loan                8                   26       <NA>  43
## 4 4     NA Female    High School Active loan               NA                   32     Mackay  48   
## 5 5      3   Male    High School Active loan                6                   41       <NA>  34
## 6 6     NA   Male     University        <NA>                9                   33 O'Loughlin  NA
##   ComplaintsCount HasCrCard EstimatedSalary IsActiveMember AvgTransactionAmount CustomerSegment Mari
## 1               0        NA              NA              0                   NA     Mass Market
## 2               0        NA              NA              0                   NA     Mass Market
## 3               0         1       135650.72              0            228.83782     Mass Market
## 4               0         0       102640.52             NA            133.89406        Affluent
## 5              NA         1        83773.02              1             91.57003            <NA>
## 6              NA        NA       116706.00              0             87.96360            <NA>
##   DigitalEngagementScore   ID CreditScore SavingsAccountFlag  Balance NumOfProducts Exited
## 1                     60 6222         832                 1       NA             2      0
## 2                     NA 3217          NA                 1       NA            NA      0
## 3                     NA   NA         577                 1 79757.21             1      1
## 4                     NA   NA         482                 0       NA            NA      1
## 5                     67 5511         635                 0       NA            NA      0
## 6                     43 3575         656                NA     0.00             2      0
```

```
summary(datos)
```

```
##        X            Tenure          Gender          EducationLevel      LoanStatus        NetPromote
##  Min.   :   1   Min.   : 0.000   Length:10000       Length:10000       Length:10000       Min.   : 0
##  1st Qu.:1751   1st Qu.: 3.000   Class :character   Class :character   Class :character   1st Qu.: 4
##  Median :3500   Median : 5.000   Mode  :character   Mode  :character   Mode  :character   Median : 8
##  Mean   :3500   Mean   : 5.005                                                           Mean   : 6
##  3rd Qu.:5250   3rd Qu.: 7.000                                                           3rd Qu.: 9
##  Max.   :7000   Max.   :10.000                                                           Max.   :10
##  NA's   :3000   NA's   :3000                                                             NA's   :30
##  TransactionFrequency   Surname               Age          Geography          ComplaintsCount    HasCr
##  Min.   :13.00        Length:10000       Min.   :18.00   Length:10000       Min.   :0.0000   Min.
##  1st Qu.:26.00        Class :character   1st Qu.:32.00   Class :character   1st Qu.:0.0000   1st Qu.
##  Median :30.00        Mode  :character   Median :37.00   Mode  :character   Median :0.0000   Median
##  Mean   :30.06                           Mean   :39.01                      Mean   :0.3696   Mean
##  3rd Qu.:34.00                           3rd Qu.:44.00                      3rd Qu.:0.0000   3rd Qu.
##  Max.   :58.00                           Max.   :92.00                      Max.   :5.0000   Max.
##  NA's   :3000                            NA's   :3000                       NA's   :3000     NA's
##  EstimatedSalary     IsActiveMember   AvgTransactionAmount CustomerSegment    MaritalStatus      Digi
##  Min.   :    11.58   Min.   :0.0000   Min.   : 19.60       Length:10000       Length:10000       Min
##  1st Qu.: 50755.81   1st Qu.:0.0000   1st Qu.: 70.12       Class :character   Class :character   1st
##  Median : 99796.85   Median :1.0000   Median : 98.61       Mode  :character   Mode  :character   Medi
##  Mean   : 99912.26   Mean   :0.5167   Mean   :111.76                                             Mean
##  3rd Qu.:148823.10   3rd Qu.:1.0000   3rd Qu.:138.02                                             3rd
##  Max.   :199992.48   Max.   :1.0000   Max.   :611.35                                             Max
##  NA's   :3000        NA's   :3000     NA's   :3000                                               NA's
##        ID        CreditScore    SavingsAccountFlag    Balance       NumOfProducts       Exited
##  Min.   :   1   Min.   :350.0   Min.   :0.0000     Min.   :     0   Min.   :1.000   Min.   :0.0000
##  1st Qu.:1167   1st Qu.:584.0   1st Qu.:0.0000     1st Qu.:     0   1st Qu.:1.000   1st Qu.:0.0000
##  Median :2334   Median :652.0   Median :1.0000     Median : 96951   Median :1.000   Median :0.0000
##  Mean   :2750   Mean   :650.7   Mean   :0.6573     Mean   : 76267   Mean   :1.536   Mean   :0.2071
##  3rd Qu.:4210   3rd Qu.:718.0   3rd Qu.:1.0000     3rd Qu.:127686   3rd Qu.:2.000   3rd Qu.:0.0000
##  Max.   :6999   Max.   :850.0   Max.   :1.0000     Max.   :250898   Max.   :4.000   Max.   :1.0000
##  NA's   :2100   NA's   :3000    NA's   :3000       NA's   :3000     NA's   :3000    NA's   :3000
```

```
clases <- sapply(datos, class)
datos$HasCrCard <- as.factor(datos$HasCrCard)
datos$IsActiveMember <- as.factor(datos$IsActiveMember)
datos$SavingsAccountFlag <- as.factor(datos$SavingsAccountFlag)


varNum <- names(datos)[sapply(datos, is.numeric)]
varCat <- names(datos)[!sapply(datos, is.numeric)]
```

```
#install.packages("dlookr")
library(dlookr)
dlookr::diagnose(datos)
```

```
## # A tibble: 24 x 6
##    variables            types     missing_count missing_percent unique_count unique_rate
##    <chr>                <chr>             <int>           <dbl>        <int>       <dbl>
##  1 X                    integer            3000              30         7001      0.700
##  2 Tenure               integer            3000              30           12      0.0012
##  3 Gender               character          3000              30            3      0.0003
##  4 EducationLevel       character          3000              30            5      0.0005
```

```
##  5 LoanStatus           character          3000              30            4      0.0004
##  6 NetPromoterScore     integer            3000              30           12      0.0012
##  7 TransactionFrequency integer            3000              30           42      0.0042
##  8 Surname              character          3000              30         2427      0.243
##  9 Age                  integer            3000              30           70      0.007
## 10 Geography            character          3000              30            4      0.0004
## # i 14 more rows
```

```
head(overview(datos), n = 9)
```

```
##     division              metrics   value
## 1       size         observations   10000
## 2       size            variables      24
## 3       size               values  240000
## 4       size          memory size 1509120
## 5 duplicated duplicate observation       0
## 6    missing  complete observation       0
## 7    missing   missing observation   10000
## 8    missing     missing variables      24
## 9    missing        missing values   71100
```

```
diagnose_numeric(datos)
```

```
##                 variables       min          Q1         mean      median          Q3         max zero
## 1                       X   1.00000  1750.75000 3.500500e+03  3500.50000   5250.2500   7000.0000    0
## 2                  Tenure   0.00000     3.00000 5.005286e+00     5.00000      7.0000     10.0000  299
## 3        NetPromoterScore   0.00000     4.00000 6.440286e+00     8.00000      9.0000     10.0000  423
## 4    TransactionFrequency  13.00000    26.00000 3.005900e+01    30.00000     34.0000     58.0000    0
## 5                     Age  18.00000    32.00000 3.900557e+01    37.00000     44.0000     92.0000    0
## 6         ComplaintsCount   0.00000     0.00000 3.695714e-01     0.00000      0.0000      5.0000 5617
## 7         EstimatedSalary  11.58000 50755.80750 9.991226e+04 99796.84500 148823.0950 199992.4800    0
## 8    AvgTransactionAmount  19.60371    70.11772 1.117624e+02    98.61161    138.0194    611.3528    0
## 9  DigitalEngagementScore   5.00000    50.00000 5.955029e+01    60.00000     70.0000    100.0000    0
## 10                    ID   1.00000  1167.00000 2.749697e+03  2334.00000   4210.2500   6999.0000    0
## 11            CreditScore 350.00000   584.00000 6.506946e+02   652.00000    718.0000    850.0000    0
## 12                Balance   0.00000     0.00000 7.626726e+04 96950.71000 127685.6825 250898.0900 2545
## 13          NumOfProducts   1.00000     1.00000 1.535714e+00     1.00000      2.0000      4.0000    0
## 14                 Exited   0.00000     0.00000 2.071429e-01     0.00000      0.0000      1.0000 5550
```

```
diagnose_category(datos)
```

```
##             variables         levels     N freq ratio rank
## 1              Gender           Male 10000 3813 38.13    1
## 2              Gender         Female 10000 3187 31.87    2
## 3              Gender           <NA> 10000 3000 30.00    3
## 4      EducationLevel     University 10000 3206 32.06    1
## 5      EducationLevel           <NA> 10000 3000 30.00    2
## 6      EducationLevel    High School 10000 2400 24.00    3
## 7      EducationLevel   Postgraduate 10000 1064 10.64    4
## 8      EducationLevel          Other 10000  330  3.30    5
## 9          LoanStatus        No loan 10000 4216 42.16    1
## 10         LoanStatus           <NA> 10000 3000 30.00    2
```

```
## 11         LoanStatus    Active loan 10000 2065 20.65    3
## 12         LoanStatus   Default risk 10000  719  7.19    4
## 13            Surname           <NA> 10000 3000 30.00    1
## 14            Surname         Walker 10000   25  0.25    2
## 15            Surname          Smith 10000   24  0.24    3
## 16            Surname         Martin 10000   22  0.22    4
## 17            Surname          Scott 10000   21  0.21    5
## 18            Surname            Yeh 10000   21  0.21    5
## 19            Surname        Maclean 10000   20  0.20    7
## 20            Surname          Brown 10000   19  0.19    8
## 21            Surname           Shih 10000   18  0.18    9
## 22            Surname            Sun 10000   18  0.18    9
## 23          Geography         France 10000 3522 35.22    1
## 24          Geography           <NA> 10000 3000 30.00    2
## 25          Geography        Germany 10000 1755 17.55    3
## 26          Geography          Spain 10000 1723 17.23    4
## 27          HasCrCard              1 10000 4918 49.18    1
## 28          HasCrCard           <NA> 10000 3000 30.00    2
## 29          HasCrCard              0 10000 2082 20.82    3
## 30     IsActiveMember              1 10000 3617 36.17    1
## 31     IsActiveMember              0 10000 3383 33.83    2
## 32     IsActiveMember           <NA> 10000 3000 30.00    3
## 33    CustomerSegment    Mass Market 10000 3566 35.66    1
## 34    CustomerSegment           <NA> 10000 3000 30.00    2
## 35    CustomerSegment       Affluent 10000 2069 20.69    3
## 36    CustomerSegment High Net Worth 10000 1365 13.65    4
## 37      MaritalStatus        Married 10000 3549 35.49    1
## 38      MaritalStatus           <NA> 10000 3000 30.00    2
## 39      MaritalStatus         Single 10000 2074 20.74    3
## 40      MaritalStatus       Divorced 10000 1029 10.29    4
## 41      MaritalStatus        Widowed 10000  348  3.48    5
## 42 SavingsAccountFlag              1 10000 4601 46.01    1
## 43 SavingsAccountFlag           <NA> 10000 3000 30.00    2
## 44 SavingsAccountFlag              0 10000 2399 23.99    3
```

```
mapply(function(x, name) {
  if (is.numeric(x)) {
    cat("var. ", name, ": \n\t min: ", min(x, na.rm = TRUE), "\n\t max: ", max(x, na.rm = TRUE), "\n")
  }
  invisible(NULL)
}, datos[, varNum], colnames(datos[, varNum]))
```

```
## var.  X :
##   min:  1
##   max:  7000
## var.  Tenure :
##   min:  0
##   max:  10
## var.  NetPromoterScore :
##   min:  0
##   max:  10
## var.  TransactionFrequency :
##   min:  13
##   max:  58
```

```
## var.  Age :
##   min:  18
##   max:  92
## var.  ComplaintsCount :
##   min:  0
##   max:  5
## var.  EstimatedSalary :
##   min:  11.58
##   max:  199992.5
## var.  AvgTransactionAmount :
##   min:  19.60371
##   max:  611.3528
## var.  DigitalEngagementScore :
##   min:  5
##   max:  100
## var.  ID :
##   min:  1
##   max:  6999
## var.  CreditScore :
##   min:  350
##   max:  850
## var.  Balance :
##   min:  0
##   max:  250898.1
## var.  NumOfProducts :
##   min:  1
##   max:  4
## var.  Exited :
##   min:  0
##   max:  1
```

```
## $X
## NULL
##
## $Tenure
## NULL
##
## $NetPromoterScore
## NULL
##
## $TransactionFrequency
## NULL
##
## $Age
## NULL
##
## $ComplaintsCount
## NULL
##
## $EstimatedSalary
## NULL
##
## $AvgTransactionAmount
## NULL
```

```
##
## $DigitalEngagementScore
## NULL
##
## $ID
## NULL
##
## $CreditScore
## NULL
##
## $Balance
## NULL
##
## $NumOfProducts
## NULL
##
## $Exited
## NULL
```

```
library(EnvStats)

IQROutlier <- function(variable, rmnas = TRUE) {
  variable <- variable[!is.na(variable)]  # eliminar NAs
  IQ <- iqr(variable, na.rm = rmnas)
  intInf <- quantile(variable, probs = 0.25, na.rm = TRUE) - 1.5 * IQ
  intSup <- quantile(variable, probs = 0.75, na.rm = TRUE) + 1.5 * IQ
  posicions <- which(variable >= intSup | variable <= intInf)
  if (length(posicions) > 0) {
    cat("Existeixen outliers en les posicions:", paste0(posicions, collapse = ", "))
  } else {
    cat("No existeixen outliers")
  }
  return(posicions)
}

outliers_todos <- lapply(varNum, function(v) {
  cat("\n\nVariable:", v, "\n")
  IQROutlier(datos[[v]])
})
```

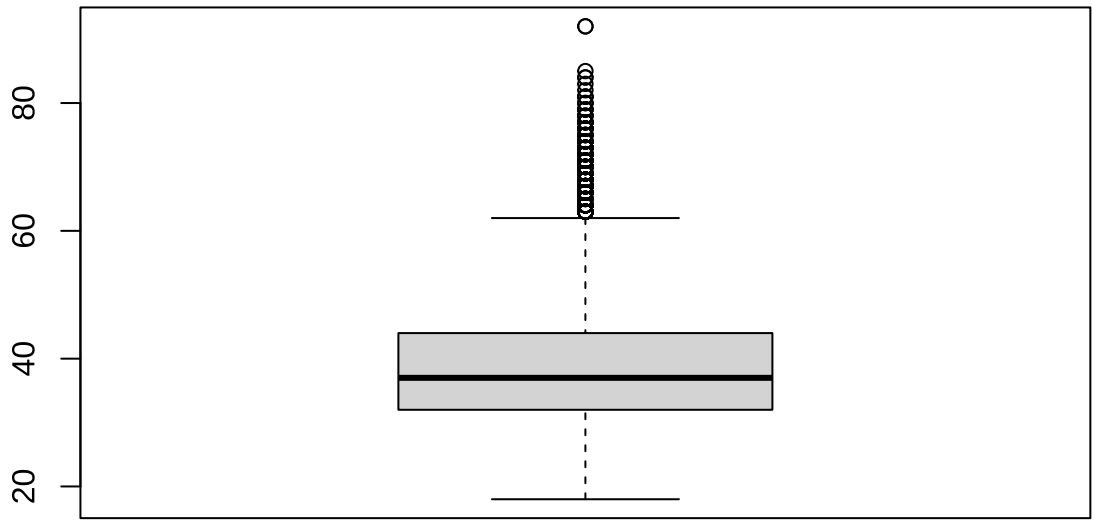
```
##
##
## Variable: X
## No existeixen outliers
##
## Variable: Tenure
## No existeixen outliers
##
## Variable: NetPromoterScore
## No existeixen outliers
##
## Variable: TransactionFrequency
## Existeixen outliers en les posicions: 110, 183, 359, 643, 1099, 1464, 1468, 1511, 1584, 1827, 2060, 1
##
```

```
## Variable: Age
## Existeixen outliers en les posicions: 31, 33, 87, 116, 152, 197, 200, 216, 243, 274, 287, 308, 328, 3
##
## Variable: ComplaintsCount
## Existeixen outliers en les posicions: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18,
##
## Variable: EstimatedSalary
## No existeixen outliers
##
## Variable: AvgTransactionAmount
## Existeixen outliers en les posicions: 65, 72, 94, 101, 150, 175, 185, 188, 216, 218, 231, 332, 333, 4
##
## Variable: DigitalEngagementScore
## Existeixen outliers en les posicions: 7, 92, 385, 606, 1386, 1428, 1769, 2175, 2196, 2400, 2464, 248
##
## Variable: ID
## No existeixen outliers
##
## Variable: CreditScore
## Existeixen outliers en les posicions: 524, 878, 995, 1072, 2221, 2245, 3005, 3196, 5240, 6991
##
## Variable: Balance
## No existeixen outliers
##
## Variable: NumOfProducts
## Existeixen outliers en les posicions: 456, 1087, 1152, 1265, 1438, 1720, 1846, 2145, 2154, 2170, 226
##
## Variable: Exited
## Existeixen outliers en les posicions: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18,
```

```r
library(ggplot2)

variable <- "Age"

boxplot(datos[, variable])
```

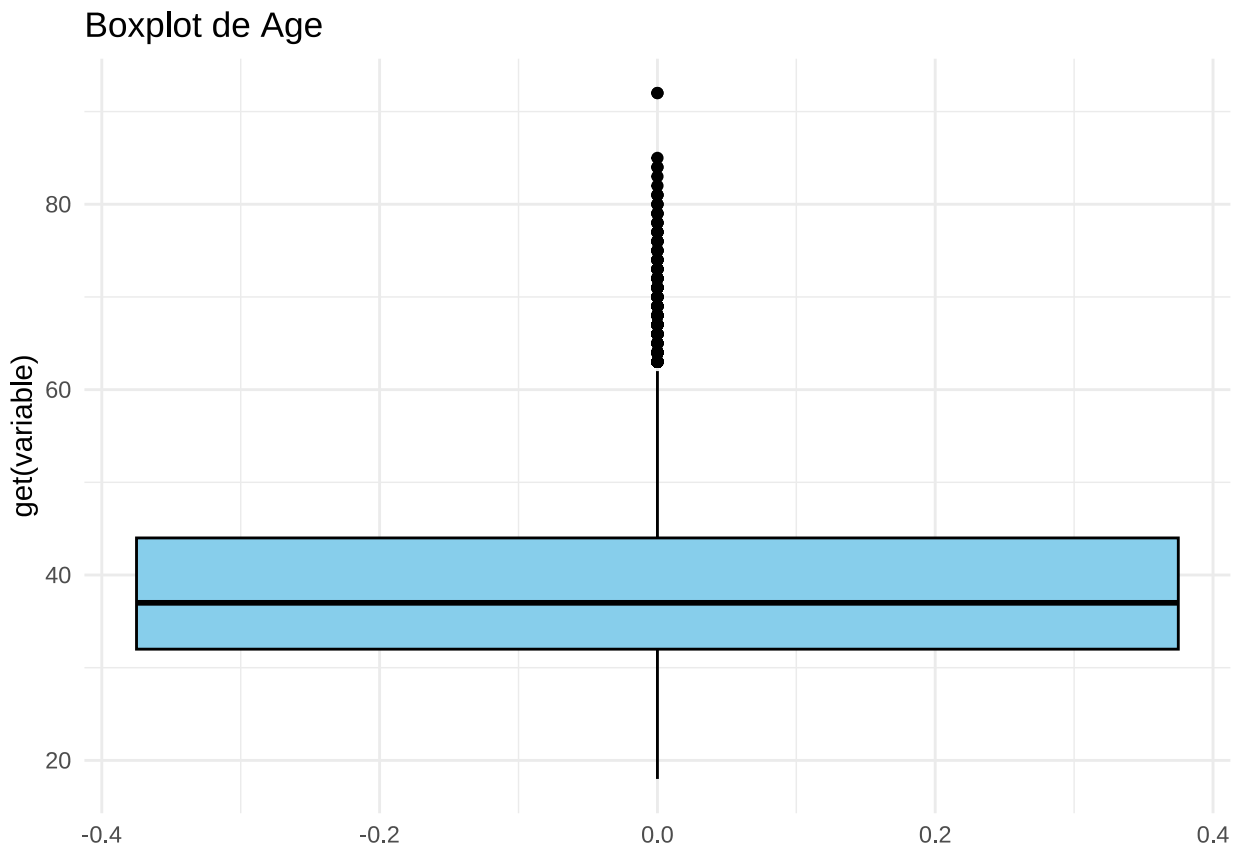



```
boxplot.stats(datos[, variable])$out
```

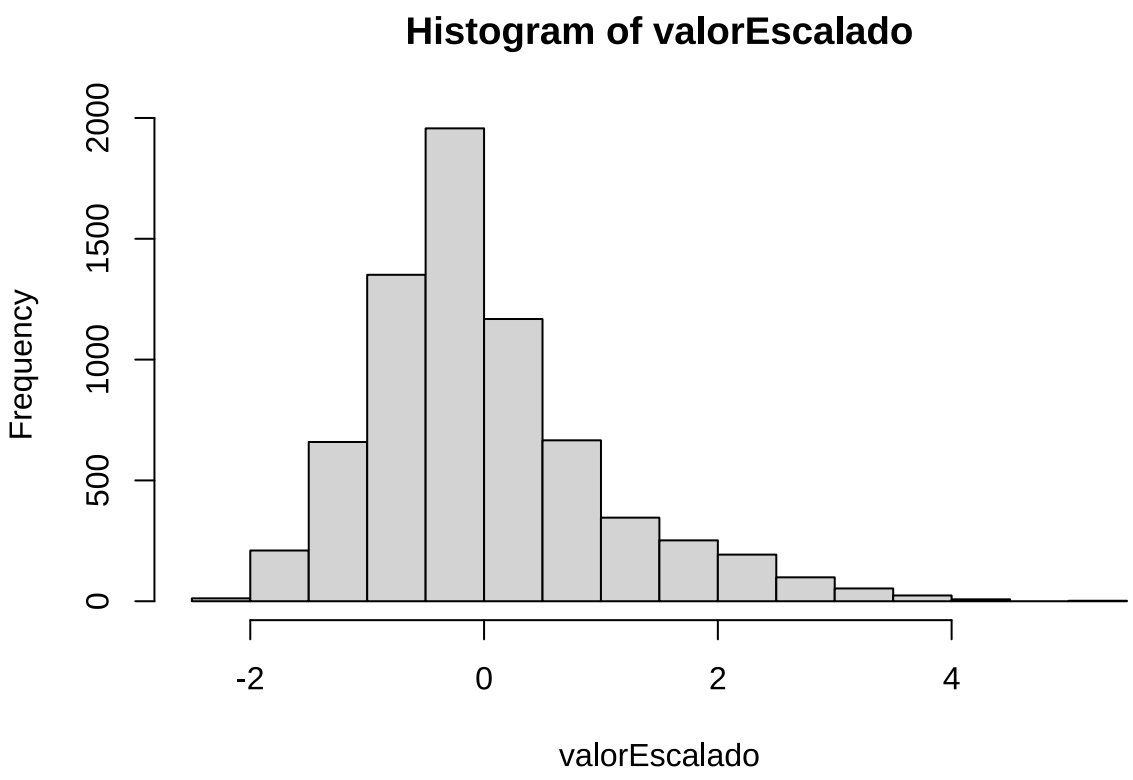
```
##   [1] 74 67 68 64 67 63 77 72 66 75 66 64 70 67 63 68 66 75 69 78 77 67 76 72 71 76 69 67 63 63 85 64
##  [39] 63 66 63 64 66 65 71 76 73 92 67 67 75 71 64 79 67 75 64 80 66 66 70 63 68 82 63 67 73 66 72 64
##  [77] 68 63 72 68 66 70 63 64 63 76 69 79 84 66 63 78 69 66 72 63 76 63 75 66 67 67 68 63 74 64 68 70
## [115] 66 73 64 66 68 66 76 69 63 76 73 67 63 66 80 73 63 72 67 67 66 63 64 74 67 66 70 70 67 66 70 64
## [153] 77 77 74 63 69 73 73 63 71 67 64 71 63 92 74 70 69 67 81 63 67 66 69 68 70 74 75 70 65 74 63 63
## [191] 69 70 74 63 72 84 67 65 71 66 68 67 71 64 80 64 77 63 69 73 66 63 64 71 65 83 64 66 67 63 64 71
## [229] 69 66 73 71 67 77 71 65 68 73 63 63 70 64 70 68 64 71 78 71 68 69 67
```

```
# Crear un boxplot
ggplot(datos, aes(y = get(variable))) +
  geom_boxplot(fill = "skyblue", color = "black") +
  labs(title = paste0("Boxplot de ", variable)) +
  theme_minimal()
```

```
## Warning: Removed 3000 rows containing non-finite outside the scale range (`stat_boxplot()`).
```

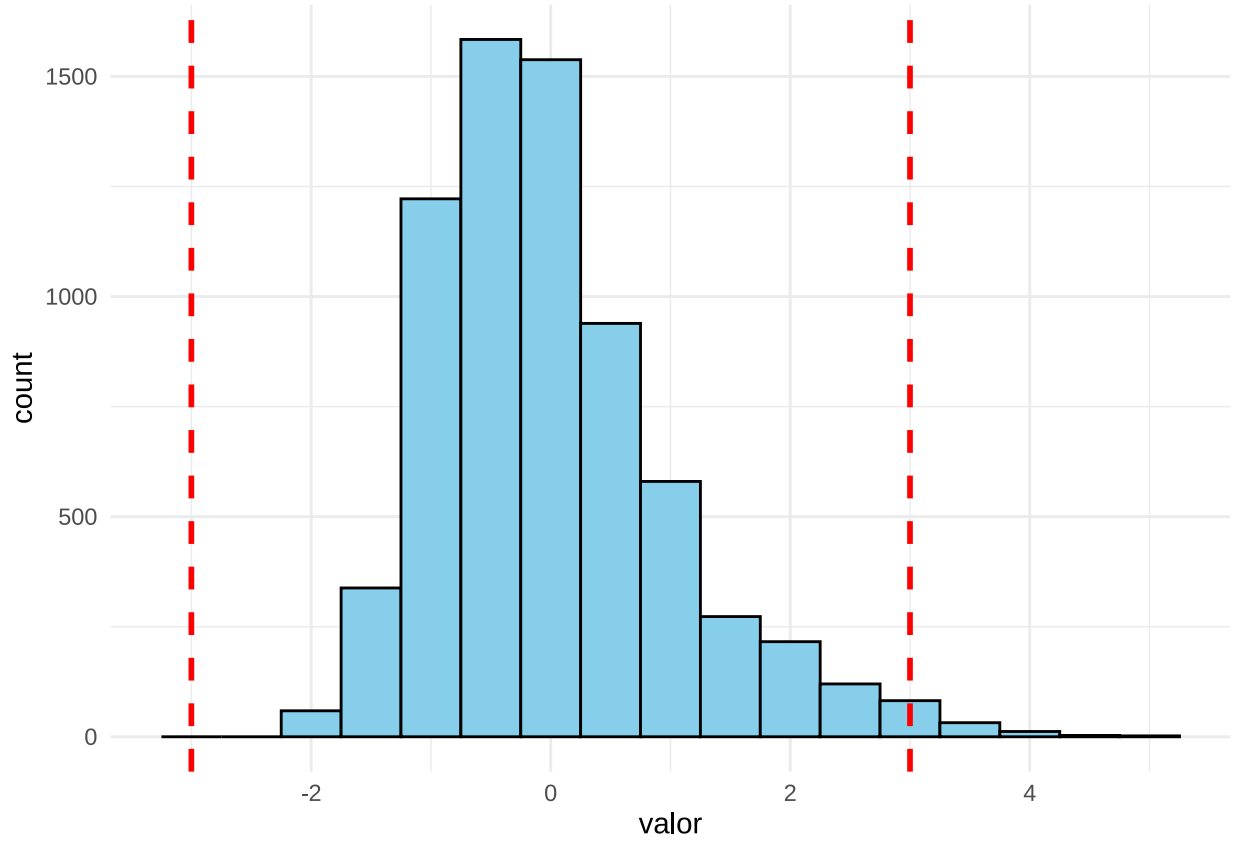



```
variable <- "Age"
valorEscalado <- scale(datos[, variable])
hist(valorEscalado)
```




```
ggplot(data.frame(valor = valorEscalado), aes(x = valor)) +
  geom_histogram(binwidth = 0.5, fill = "skyblue", color = "black") +  # Histograma
  geom_vline(xintercept = c(3, -3), linetype = "dashed", color = "red", size = 1) + # Líneas horizontal
  theme_minimal()
```

```
## Warning: Removed 3000 rows containing non-finite outside the scale range (‘stat_bin()‘).
```




```
#library(dplyr)

train <- read.csv("train.csv", header = TRUE, sep = ",")
test  <- read.csv("test.csv", header = TRUE, sep = ",")
datos <- bind_rows(train, test)

varNum <- names(datos)[sapply(datos, is.numeric)]
datos_num <- datos[, varNum]
summary(datos_num)
```

```
##        X            Tenure       NetPromoterScore TransactionFrequency      Age        ComplaintsCou
##  Min.   :   1   Min.   : 0.000   Min.   : 0.00    Min.   :13.00        Min.   :18.00   Min.   :0.0000
##  1st Qu.:1751   1st Qu.: 3.000   1st Qu.: 4.00    1st Qu.:26.00        1st Qu.:32.00   1st Qu.:0.0000
##  Median :3500   Median : 5.000   Median : 8.00    Median :30.00        Median :37.00   Median :0.0000
##  Mean   :3500   Mean   : 5.005   Mean   : 6.44    Mean   :30.06        Mean   :39.01   Mean   :0.3690
##  3rd Qu.:5250   3rd Qu.: 7.000   3rd Qu.: 9.00    3rd Qu.:34.00        3rd Qu.:44.00   3rd Qu.:0.0000
##  Max.   :7000   Max.   :10.000   Max.   :10.00    Max.   :58.00        Max.   :92.00   Max.   :5.0000
##  NA's   :3000   NA's   :3000     NA's   :3000     NA's   :3000         NA's   :3000    NA's   :3000
##  EstimatedSalary     IsActiveMember   AvgTransactionAmount DigitalEngagementScore       ID          Cr
##  Min.   :    11.58   Min.   :0.0000   Min.   : 19.60       Min.   :  5.00          Min.   :   1   Min
##  1st Qu.: 50755.81   1st Qu.:0.0000   1st Qu.: 70.12       1st Qu.: 50.00          1st Qu.:1167   1st
##  Median : 99796.85   Median :1.0000   Median : 98.61       Median : 60.00          Median :2334   Medi
##  Mean   : 99912.26   Mean   :0.5167   Mean   :111.76       Mean   : 59.55          Mean   :2750   Mean
##  3rd Qu.:148823.10   3rd Qu.:1.0000   3rd Qu.:138.02       3rd Qu.: 70.00          3rd Qu.:4210   3rd
##  Max.   :199992.48   Max.   :1.0000   Max.   :611.35       Max.   :100.00          Max.   :6999   Max
```

```
##  NA's   :3000        NA's   :3000     NA's   :3000         NA's   :3000          NA's   :2100   NA's
##  SavingsAccountFlag    Balance       NumOfProducts       Exited
##  Min.   :0.0000     Min.   :     0   Min.   :1.000   Min.   :0.0000
##  1st Qu.:0.0000     1st Qu.:     0   1st Qu.:1.000   1st Qu.:0.0000
##  Median :1.0000     Median : 96951   Median :1.000   Median :0.0000
##  Mean   :0.6573     Mean   : 76267   Mean   :1.536   Mean   :0.2071
##  3rd Qu.:1.0000     3rd Qu.:127686   3rd Qu.:2.000   3rd Qu.:0.0000
##  Max.   :1.0000     Max.   :250898   Max.   :4.000   Max.   :1.0000
##  NA's   :3000       NA's   :3000     NA's   :3000    NA's   :3000
```

```
center <- colMeans(datos_num, na.rm = TRUE)
cov_matrix <- cov(datos_num, use = "complete.obs")


dist_mahal <- mahalanobis(datos_num, center, cov_matrix)

umbral <- qchisq(0.999, df = ncol(datos_num))


outliers_mahal <- which(dist_mahal > umbral)
cat("Número de outliers multivariantes detectados:", length(outliers_mahal), "\n")
```

```
## Número de outliers multivariantes detectados: 0
```

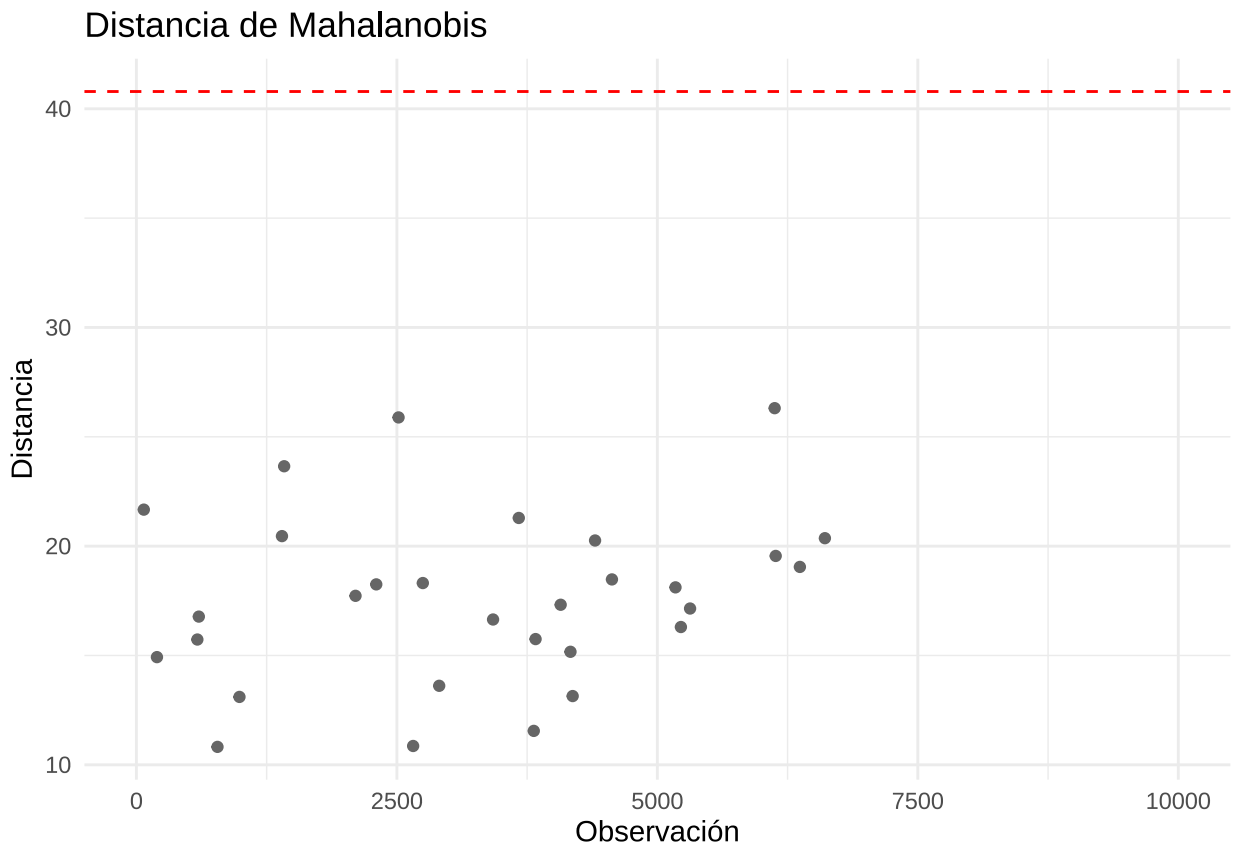
```
summary(dist_mahal)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   10.82   15.31   17.52   17.61   20.08   26.31    9970
```

```
#library(ggplot2)

ggplot(data.frame(dist_mahal), aes(x = 1:length(dist_mahal), y = dist_mahal)) +
  geom_point(color = "grey40") +
  geom_hline(yintercept = umbral, color = "red", linetype = "dashed") +
  labs(title = "Distancia de Mahalanobis", x = "Observación", y = "Distancia") +
  theme_minimal()
```

```
## Warning: Removed 9970 rows containing missing values or values outside the scale range (‘geom_point()
```



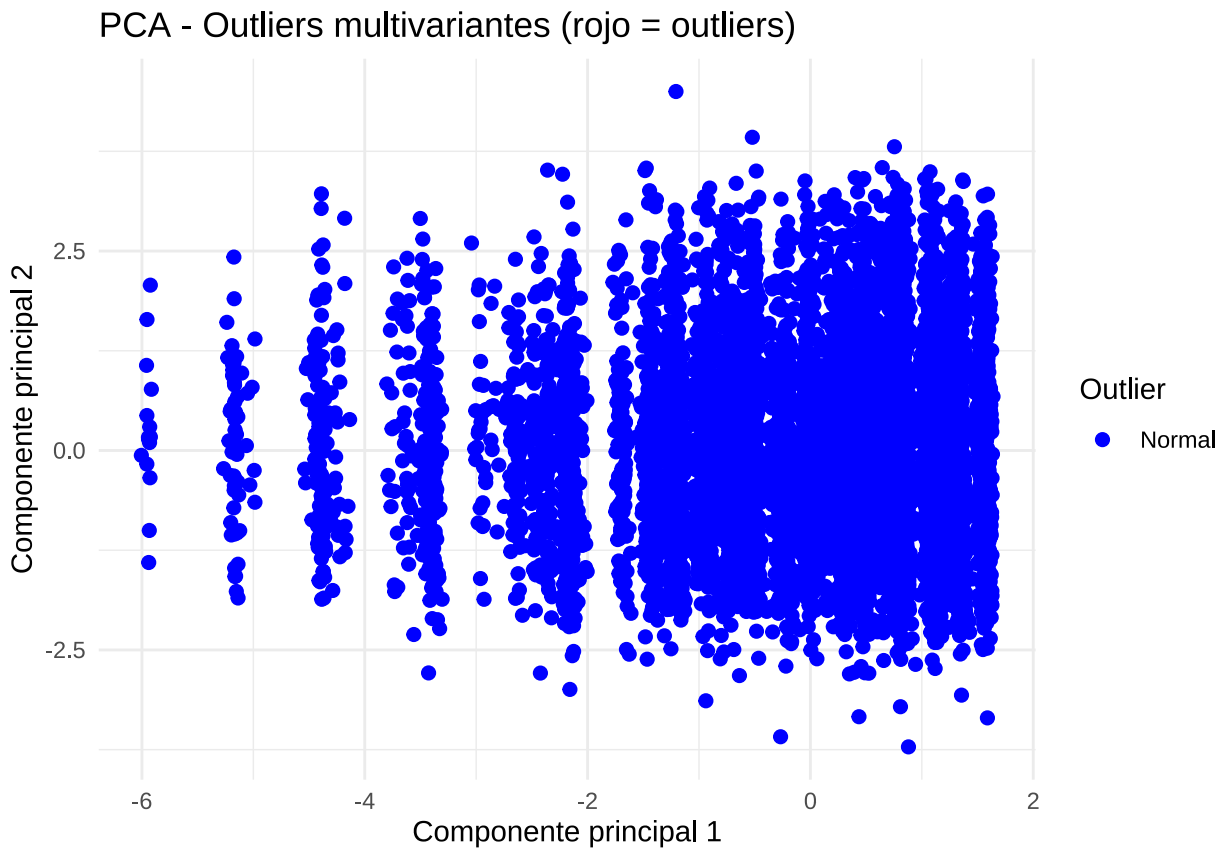

```
library(FactoMineR)
library(factoextra)
library(ggplot2)

# PCA
pca <- PCA(datos_num, scale.unit = TRUE, graph = FALSE)
```

```
## Warning in PCA(datos_num, scale.unit = TRUE, graph = FALSE): Missing values are imputed by the mean o
## should use the imputePCA function of the missMDA package
```

```
# Crear dataframe con coordenadas
coords <- as.data.frame(pca$ind$coord)
coords$Outlier <- factor(ifelse(1:nrow(datos_num) %in% outliers_mahal, "Outlier", "Normal"))

# Gráfico manual con ggplot2 (sin errores de color)
ggplot(coords, aes(x = Dim.1, y = Dim.2, color = Outlier)) +
  geom_point(size = 2) +
  scale_color_manual(values = c("blue", "red")) +
  labs(title = "PCA - Outliers multivariantes (rojo = outliers)",
       x = "Componente principal 1",
       y = "Componente principal 2") +
  theme_minimal()
```




```
datos$Outlier_Mahalanobis <- FALSE
datos$Outlier_Mahalanobis[outliers_mahal] <- TRUE

write.csv(datos, "datos_outliers_multivariantes.csv", row.names = FALSE)
```