TP1 - Toyota Corolla

Navarro Matias, Ortiz Fausto - Universidad Tecnológica Nacional - Facultad Regional Tucumán

11/10/2020

Carga de librerías y datos

```
library("dplyr")
```

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

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

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

```
library("gdata")
```

```
## gdata: Unable to locate valid perl interpreter
## gdata:
## gdata: read.xls() will be unable to read Excel XLS and XLSX files
## gdata: unless the 'perl=' argument is used to specify the location of a
## gdata: valid perl intrpreter.
## gdata:
## gdata: (To avoid display of this message in the future, please ensure
## gdata: perl is installed and available on the executable search path.)
```

```
## gdata: Unable to load perl libaries needed by read.xls()
## gdata: to support 'XLX' (Excel 97-2004) files.
```

```
##
```

```
## gdata: Unable to load perl libaries needed by read.xls()
## gdata: to support 'XLSX' (Excel 2007+) files.
```

```
##
```

```
## gdata: Run the function 'installXLSXsupport()'
## gdata: to automatically download and install the perl
## gdata: libaries needed to support Excel XLS and XLSX formats.
```

```
##
## Attaching package: 'gdata'
```

```
## The following objects are masked from 'package:dplyr':
##
##     combine, first, last
```

```
## The following object is masked from 'package:stats':
##
##     nobs
## The following object is masked from 'package:utils':
##
##     object.size
## The following object is masked from 'package:base':
##
##     startsWith
```

```
library("corrplot")
```

```
## corrplot 0.84 loaded
```

```
library("moments")
library("fastDummies")
library("ggplot2")
library("psych")
```

```
##
## Attaching package: 'psych'
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
library("car")
```

```
## Loading required package: carData
##
## Attaching package: 'car'
## The following object is masked from 'package:psych':
##
##     logit
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
library("corrplot")
library("caret")
```

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

Carga de Datos

```
datos = read.csv("ToyotaCorolla.csv")
```

Mostrar los datos del dataset

```
head(datos,20)
```

```
##    Id                                              Model Price Age_08_04
## 1   1          TOYOTA Corolla 2.0 D4D HATCHB TERRA 2/3-Doors 13500        23
## 2   2          TOYOTA Corolla 2.0 D4D HATCHB TERRA 2/3-Doors 13750        23
## 3   3         ?TOYOTA Corolla 2.0 D4D HATCHB TERRA 2/3-Doors 13950        24
## 4   4          TOYOTA Corolla 2.0 D4D HATCHB TERRA 2/3-Doors 14950        26
## 5   5            TOYOTA Corolla 2.0 D4D HATCHB SOL 2/3-Doors 13750        30
```

```
## 6   6            TOYOTA Corolla 2.0 D4D HATCHB SOL 2/3-Doors 12950        32
## 7   7         ?TOYOTA Corolla 2.0 D4D 90 3DR TERRA 2/3-Doors 16900        27
## 8   8          TOYOTA Corolla 2.0 D4D 90 3DR TERRA 2/3-Doors 18600        30
## 9   9           ?TOYOTA Corolla 1800 T SPORT VVT I 2/3-Doors 21500        27
## 10 10           ?TOYOTA Corolla 1.9 D HATCHB TERRA 2/3-Doors 12950        23
## 11 11      TOYOTA Corolla 1.8 VVTL-i T-Sport 3-Drs 2/3-Doors 20950        25
## 12 12 TOYOTA Corolla 1.8 16V VVTLI 3DR T SPORT BNS 2/3-Doors 19950        22
## 13 13     TOYOTA Corolla 1.8 16V VVTLI 3DR T SPORT 2/3-Doors 19600        25
## 14 14     TOYOTA Corolla 1.8 16V VVTLI 3DR T SPORT 2/3-Doors 21500        31
## 15 15     TOYOTA Corolla 1.8 16V VVTLI 3DR T SPORT 2/3-Doors 22500        32
## 16 16     TOYOTA Corolla 1.8 16V VVTLI 3DR T SPORT 2/3-Doors 22000        28
## 17 17    ?TOYOTA Corolla 1.8 16V VVTLI 3DR T SPORT 2/3-Doors 22750        30
## 18 18 ?TOYOTA Corolla 1.6 VVTI Linea Terra Comfort 2/3-Doors 17950        24
## 19 19                 TOYOTA Corolla 1.6 16v L.SOL 2/3-Doors 16750        24
## 20 20       TOYOTA Corolla 1.6 16V VVT I 3DR TERRA 2/3-Doors 16950        30
##    Mfg_Month Mfg_Year    KM Fuel_Type  HP Met_Color Automatic   cc Doors
## 1         10     2002 46986    Diesel  90         1         0 2000     3
## 2         10     2002 72937    Diesel  90         1         0 2000     3
## 3          9     2002 41711    Diesel  90         1         0 2000     3
## 4          7     2002 48000    Diesel  90         0         0 2000     3
## 5          3     2002 38500    Diesel  90         0         0 2000     3
## 6          1     2002 61000    Diesel  90         0         0 2000     3
## 7          6     2002 94612    Diesel  90         1         0 2000     3
## 8          3     2002 75889    Diesel  90         1         0 2000     3
## 9          6     2002 19700    Petrol 192         0         0 1800     3
## 10        10     2002 71138    Diesel  69         0         0 1900     3
## 11         8     2002 31461    Petrol 192         0         0 1800     3
## 12        11     2002 43610    Petrol 192         0         0 1800     3
## 13         8     2002 32189    Petrol 192         0         0 1800     3
## 14         2     2002 23000    Petrol 192         1         0 1800     3
## 15         1     2002 34131    Petrol 192         1         0 1800     3
## 16         5     2002 18739    Petrol 192         0         0 1800     3
## 17         3     2002 34000    Petrol 192         1         0 1800     3
## 18         9     2002 21716    Petrol 110         1         0 1600     3
## 19         9     2002 25563    Petrol 110         0         0 1600     3
## 20         3     2002 64359    Petrol 110         1         0 1600     3
##    Cylinders Gears Quarterly_Tax Weight Mfr_Guarantee BOVAG_Guarantee
## 1          4     5           210   1165             0               1
## 2          4     5           210   1165             0               1
## 3          4     5           210   1165             1               1
## 4          4     5           210   1165             1               1
## 5          4     5           210   1170             1               1
## 6          4     5           210   1170             0               1
## 7          4     5           210   1245             0               1
## 8          4     5           210   1245             1               1
## 9          4     5           100   1185             0               1
## 10         4     5           185   1105             0               1
## 11         4     6           100   1185             1               1
## 12         4     6           100   1185             1               1
## 13         4     6           100   1185             1               1
## 14         4     6           100   1185             1               1
## 15         4     6           100   1185             1               1
## 16         4     6           100   1185             0               1
## 17         4     5           100   1185             0               1
```

```
## 18         4     5            85   1105             0              0
## 19         4     5            19   1065             0              0
## 20         4     5            85   1105             1              1
##    Guarantee_Period ABS Airbag_1 Airbag_2 Airco Automatic_airco Boardcomputer
## 1                 3   1        1        1     0               0             1
## 2                 3   1        1        1     1               0             1
## 3                 3   1        1        1     0               0             1
## 4                 3   1        1        1     0               0             1
## 5                 3   1        1        1     1               0             1
## 6                 3   1        1        1     1               0             1
## 7                 3   1        1        1     1               0             1
## 8                 3   1        1        1     1               0             1
## 9                 3   1        1        0     1               0             0
## 10                3   1        1        1     1               0             1
## 11               12   1        1        1     1               1             0
## 12                3   1        1        1     1               1             1
## 13                3   1        1        1     1               1             1
## 14                3   1        1        1     1               1             1
## 15                3   1        1        1     1               1             1
## 16                3   1        1        1     1               1             1
## 17                3   1        1        1     1               1             1
## 18               18   1        1        0     1               0             0
## 19                3   1        1        1     1               1             1
## 20                3   1        1        1     1               0             1
##    CD_Player Central_Lock Powered_Windows Power_Steering Radio Mistlamps
## 1          0            1              1              1     0         0
## 2          1            1              0              1     0         0
## 3          0            0              0              1     0         0
## 4          0            0              0              1     0         0
## 5          0            1              1              1     0         1
## 6          0            1              1              1     0         1
## 7          0            1              1              1     0         0
## 8          1            1              1              1     0         0
## 9          0            1              1              1     1         0
## 10         0            0              0              1     0         0
## 11         1            1              1              1     0         0
## 12         0            1              1              1     0         1
## 13         0            1              1              1     0         1
## 14         1            1              1              1     0         1
## 15         1            1              1              1     0         1
## 16         0            1              1              1     0         1
## 17         1            1              1              1     0         1
## 18         0            1              1              1     1         0
## 19         1            1              1              1     0         1
## 20         1            1              1              1     0         0
##    Sport_Model Backseat_Divider Metallic_Rim Radio_cassette Tow_Bar
## 1            0                1           0              0       0
## 2            0                1           0              0       0
## 3            0                1           0              0       0
## 4            0                1           0              0       0
## 5            0                1           0              0       0
## 6            0                1           0              0       0
## 7            1                1           0              0       0
## 8            0                1           0              0       0
```

```
## 9            0               0            1              1       0
## 10           0               1            0              0       0
## 11           0               0            1              0       0
## 12           1               1            1              0       0
## 13           1               1            1              0       0
## 14           1               1            1              0       0
## 15           1               1            1              0       0
## 16           1               1            1              0       0
## 17           0               1            1              0       0
## 18           0               0            0              1       1
## 19           0               0            0              0       0
## 20           1               1            0              0       0
```

Estructura de dataset

```
str(datos)
```

```
## 'data.frame':    1436 obs. of  37 variables:
##  $ Id              : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Model           : chr  "TOYOTA Corolla 2.0 D4D HATCHB TERRA 2/3-Doors" "TOYOTA Corolla 2.0 D4D HA
##  $ Price           : int  13500 13750 13950 14950 13750 12950 16900 18600 21500 12950 ...
##  $ Age_08_04       : int  23 23 24 26 30 32 27 30 27 23 ...
##  $ Mfg_Month       : int  10 10 9 7 3 1 6 3 6 10 ...
##  $ Mfg_Year        : int  2002 2002 2002 2002 2002 2002 2002 2002 2002 2002 ...
##  $ KM              : int  46986 72937 41711 48000 38500 61000 94612 75889 19700 71138 ...
##  $ Fuel_Type       : chr  "Diesel" "Diesel" "Diesel" "Diesel" ...
##  $ HP              : int  90 90 90 90 90 90 90 90 192 69 ...
##  $ Met_Color       : int  1 1 1 0 0 0 1 1 0 0 ...
##  $ Automatic       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ cc              : int  2000 2000 2000 2000 2000 2000 2000 2000 1800 1900 ...
##  $ Doors           : int  3 3 3 3 3 3 3 3 3 3 ...
##  $ Cylinders       : int  4 4 4 4 4 4 4 4 4 4 ...
##  $ Gears           : int  5 5 5 5 5 5 5 5 5 5 ...
##  $ Quarterly_Tax   : int  210 210 210 210 210 210 210 210 100 185 ...
##  $ Weight          : int  1165 1165 1165 1165 1170 1170 1245 1245 1185 1105 ...
##  $ Mfr_Guarantee   : int  0 0 1 1 1 0 0 1 0 0 ...
##  $ BOVAG_Guarantee : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Guarantee_Period: int  3 3 3 3 3 3 3 3 3 3 ...
##  $ ABS             : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Airbag_1        : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Airbag_2        : int  1 1 1 1 1 1 1 1 0 1 ...
##  $ Airco           : int  0 1 0 0 1 1 1 1 1 1 ...
##  $ Automatic_airco : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Boardcomputer   : int  1 1 1 1 1 1 1 1 0 1 ...
##  $ CD_Player       : int  0 1 0 0 0 0 0 1 0 0 ...
##  $ Central_Lock    : int  1 1 0 0 1 1 1 1 1 0 ...
##  $ Powered_Windows : int  1 0 0 0 1 1 1 1 1 0 ...
##  $ Power_Steering  : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Radio           : int  0 0 0 0 0 0 0 0 1 0 ...
##  $ Mistlamps       : int  0 0 0 0 1 1 0 0 0 0 ...
##  $ Sport_Model     : int  0 0 0 0 0 0 1 0 0 0 ...
##  $ Backseat_Divider: int  1 1 1 1 1 1 1 1 0 1 ...
##  $ Metallic_Rim    : int  0 0 0 0 0 0 0 0 1 0 ...
##  $ Radio_cassette  : int  0 0 0 0 0 0 0 0 1 0 ...
##  $ Tow_Bar         : int  0 0 0 0 0 0 0 0 0 0 ...
```

Formatear algunas variables para una mejor observación

```
deleteme = datos
deleteme$Fuel_Type = as.factor(datos$Fuel_Type)
deleteme$Mfg_Month =  as.factor(datos$Mfg_Month)
deleteme$Mfg_Year = as.factor(datos$Mfg_Year)
deleteme$Met_Color = as.factor(datos$Met_Color)
deleteme$Automatic = as.factor(datos$Automatic)
deleteme$Doors = as.factor(datos$Doors)
deleteme$Cylinders = as.factor(datos$Cylinders)
deleteme$Gears = as.factor(datos$Gears)
deleteme$Mfr_Guarantee = as.factor(datos$Mfr_Guarantee)
deleteme$BOVAG_Guarantee = as.factor(datos$BOVAG_Guarantee)
deleteme$Guarantee_Period = as.factor(datos$Guarantee_Period)
deleteme$ABS = as.factor(datos$ABS)
deleteme$Airbag_1 = as.factor(datos$Airbag_1)
deleteme$Airbag_2 = as.factor(datos$Airbag_2)
deleteme$Airco = as.factor(datos$Airco)
deleteme$Automatic_airco = as.factor(datos$Automatic_airco)
deleteme$Boardcomputer = as.factor(datos$Boardcomputer)
deleteme$CD_Player = as.factor(datos$CD_Player)
deleteme$Central_Lock = as.factor(datos$Central_Lock)
deleteme$Power_Windows = as.factor(datos$Powered_Windows)
deleteme$Power_Steering = as.factor(datos$Power_Steering)
deleteme$Radio = as.factor(datos$Radio)
deleteme$Mistlamps = as.factor(datos$Mistlamps)
deleteme$Sport_Model = as.factor(datos$Sport_Model)
deleteme$Backseat_Divider = as.factor(datos$Backseat_Divider)
deleteme$Metallic_Rim = as.factor(datos$Metallic_Rim)
deleteme$Radio_cassette = as.factor(datos$Radio_cassette)
deleteme$Tow_Bar = as.factor(datos$Tow_Bar)


str(deleteme)
```

```
## 'data.frame':    1436 obs. of  38 variables:
##  $ Id              : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Model           : chr  "TOYOTA Corolla 2.0 D4D HATCHB TERRA 2/3-Doors" "TOYOTA Corolla 2.0 D4D HA
##  $ Price           : int  13500 13750 13950 14950 13750 12950 16900 18600 21500 12950 ...
##  $ Age_08_04       : int  23 23 24 26 30 32 27 30 27 23 ...
##  $ Mfg_Month       : Factor w/ 12 levels "1","2","3","4",..: 10 10 9 7 3 1 6 3 6 10 ...
##  $ Mfg_Year        : Factor w/ 7 levels "1998","1999",..: 5 5 5 5 5 5 5 5 5 5 ...
##  $ KM              : int  46986 72937 41711 48000 38500 61000 94612 75889 19700 71138 ...
##  $ Fuel_Type       : Factor w/ 3 levels "CNG","Diesel",..: 2 2 2 2 2 2 2 2 3 2 ...
##  $ HP              : int  90 90 90 90 90 90 90 90 192 69 ...
##  $ Met_Color       : Factor w/ 2 levels "0","1": 2 2 2 1 1 1 2 2 1 1 ...
##  $ Automatic       : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ cc              : int  2000 2000 2000 2000 2000 2000 2000 2000 1800 1900 ...
##  $ Doors           : Factor w/ 4 levels "2","3","4","5": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Cylinders       : Factor w/ 1 level "4": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Gears           : Factor w/ 4 levels "3","4","5","6": 3 3 3 3 3 3 3 3 3 3 ...
##  $ Quarterly_Tax   : int  210 210 210 210 210 210 210 210 100 185 ...
##  $ Weight          : int  1165 1165 1165 1165 1170 1170 1245 1245 1185 1105 ...
##  $ Mfr_Guarantee   : Factor w/ 2 levels "0","1": 1 1 2 2 2 1 1 2 1 1 ...
```

```
##  $ BOVAG_Guarantee : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Guarantee_Period: Factor w/ 9 levels "3","6","12","13",..: 1 1 1 1 1 1 1 1 1 ...
##  $ ABS             : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Airbag_1        : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Airbag_2        : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 1 2 ...
##  $ Airco           : Factor w/ 2 levels "0","1": 1 2 1 1 2 2 2 2 2 2 ...
##  $ Automatic_airco : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Boardcomputer   : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 1 2 ...
##  $ CD_Player       : Factor w/ 2 levels "0","1": 1 2 1 1 1 1 1 2 1 1 ...
##  $ Central_Lock    : Factor w/ 2 levels "0","1": 2 2 1 1 2 2 2 2 2 1 ...
##  $ Powered_Windows : int  1 0 0 0 1 1 1 1 1 0 ...
##  $ Power_Steering  : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Radio           : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 2 1 ...
##  $ Mistlamps       : Factor w/ 2 levels "0","1": 1 1 1 1 2 2 1 1 1 1 ...
##  $ Sport_Model     : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 2 1 1 1 ...
##  $ Backseat_Divider: Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 1 2 ...
##  $ Metallic_Rim    : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 2 1 ...
##  $ Radio_cassette  : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 2 1 ...
##  $ Tow_Bar         : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Power_Windows   : Factor w/ 2 levels "0","1": 2 1 1 1 2 2 2 2 2 1 ...
```

Al analizar la estructura de "deleteme" podemos observar que tenemos muchas variables binarias y enumeraciones ya formateados con los tipo de dato que tendria que tener el dataset.

```
summary(deleteme)
```

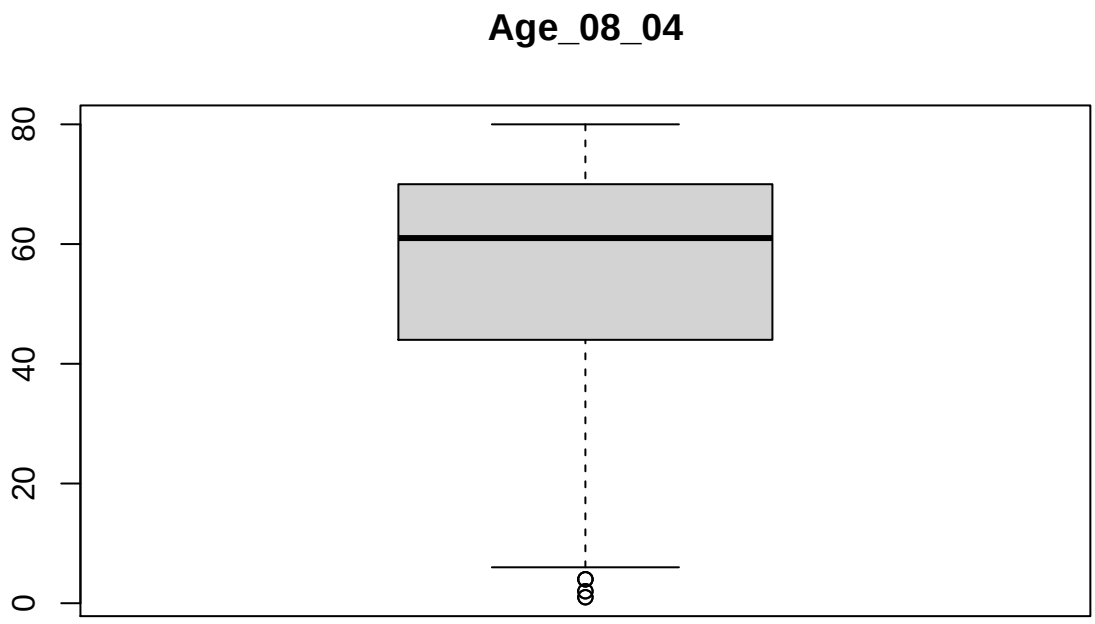
```
##        Id              Model                Price          Age_08_04
##  Min.   :   1.0   Length:1436        Min.   : 4350   Min.   : 1.00
##  1st Qu.: 361.8   Class :character   1st Qu.: 8450   1st Qu.:44.00
##  Median : 721.5   Mode  :character   Median : 9900   Median :61.00
##  Mean   : 721.6                      Mean   :10731   Mean   :55.95
##  3rd Qu.:1081.2                      3rd Qu.:11950   3rd Qu.:70.00
##  Max.   :1442.0                      Max.   :32500   Max.   :80.00
##
##    Mfg_Month   Mfg_Year         KM          Fuel_Type          HP
##  1      :207   1998:392   Min.   :     1   CNG   :  17   Min.   : 69.0
##  4      :154   1999:441   1st Qu.: 43000   Diesel: 155   1st Qu.: 90.0
##  3      :138   2000:225   Median : 63390   Petrol:1264   Median :110.0
##  2      :134   2001:192   Mean   : 68533                 Mean   :101.5
##  7      :133   2002: 87   3rd Qu.: 87021                 3rd Qu.:110.0
##  6      :120   2003: 75   Max.   :243000                 Max.   :192.0
##  (Other):550   2004: 24
##  Met_Color Automatic       cc        Doors   Cylinders Gears
##  0:467     0:1356    Min.   : 1300   2:  2   4:1436    3:   2
##  1:969     1:  80    1st Qu.: 1400   3:622             4:   1
##                      Median : 1600   4:138             5:1390
##                      Mean   : 1577   5:674             6:  43
##                      3rd Qu.: 1600
##                      Max.   :16000
##
##  Quarterly_Tax        Weight     Mfr_Guarantee BOVAG_Guarantee Guarantee_Period
##  Min.   : 19.00   Min.   :1000   0:848         0: 150          3      :1274
##  1st Qu.: 69.00   1st Qu.:1040   1:588         1:1286          6      :  77
##  Median : 85.00   Median :1070                                 12     :  73
##  Mean   : 87.12   Mean   :1072                                 24     :   4
```

```
##  3rd Qu.: 85.00   3rd Qu.:1085                                    36     :   4
##  Max.   :283.00   Max.   :1615                                    13     :   1
##                                                                   (Other):   3
##  ABS      Airbag_1 Airbag_2 Airco   Automatic_airco Boardcomputer CD_Player
##  0: 268   0:  42   0: 398   0:706   0:1355          0:1013        0:1122
##  1:1168   1:1394   1:1038   1:730   1:  81          1: 423        1: 314
##
##
##
##
##
##  Central_Lock Powered_Windows Power_Steering Radio    Mistlamps Sport_Model
##  0:603        Min.   :0.000   0:  32         0:1226   0:1067    0:1005
##  1:833        1st Qu.:0.000   1:1404         1: 210   1: 369    1: 431
##               Median :1.000
##               Mean   :0.562
##               3rd Qu.:1.000
##               Max.   :1.000
##
##  Backseat_Divider Metallic_Rim Radio_cassette Tow_Bar  Power_Windows
##  0: 330           0:1142       0:1227         0:1037   0:629
##  1:1106           1: 294       1: 209         1: 399   1:807
##
##
##
##
##
```

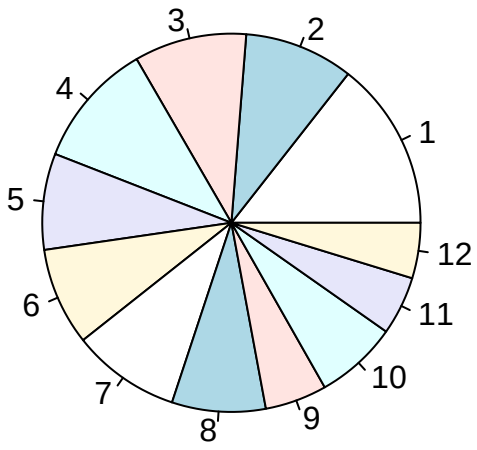
Análisis Exploratorio

Distribución de cada variable del dataset deleteme con boxplot.

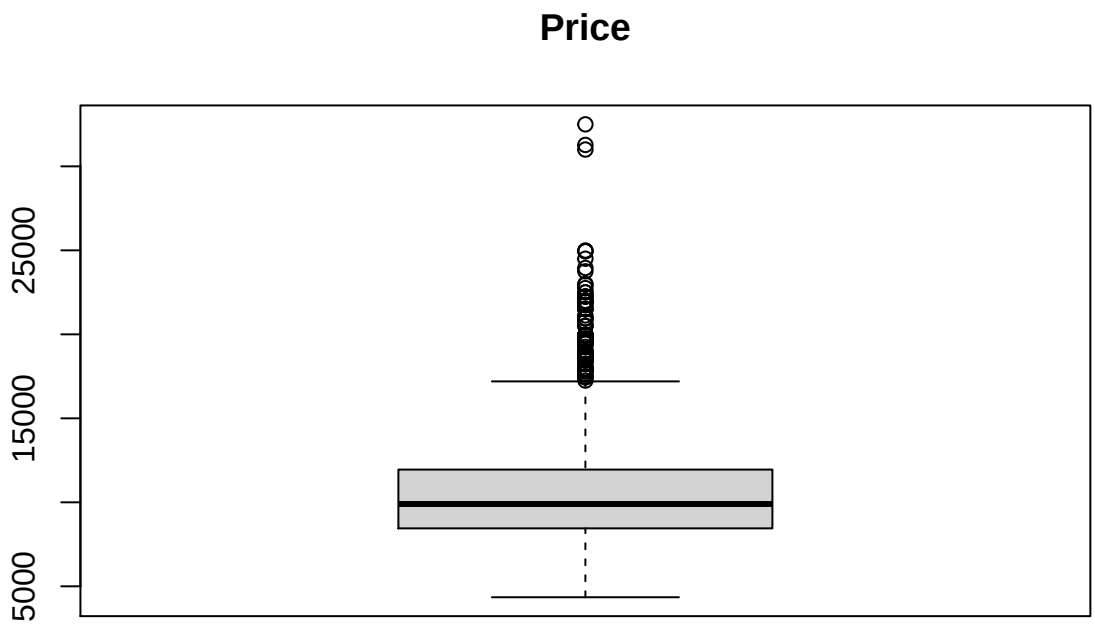
```
boxplot(deleteme$Age_08_04, main="Age_08_04")
```



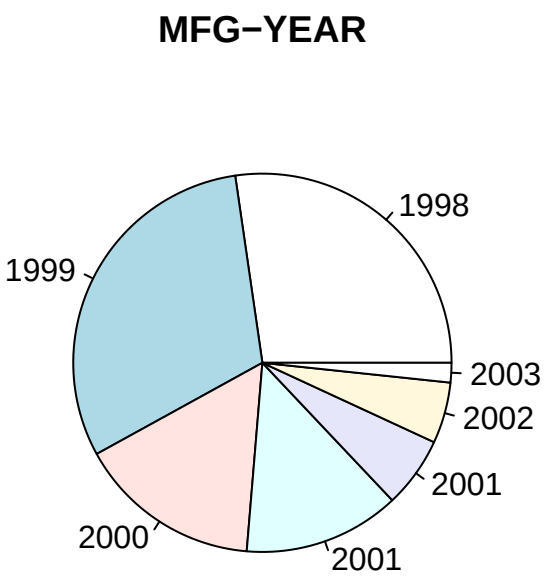

```
pie(summary(deleteme$Mfg_Month), main = "MFG-MONTH"  )
```



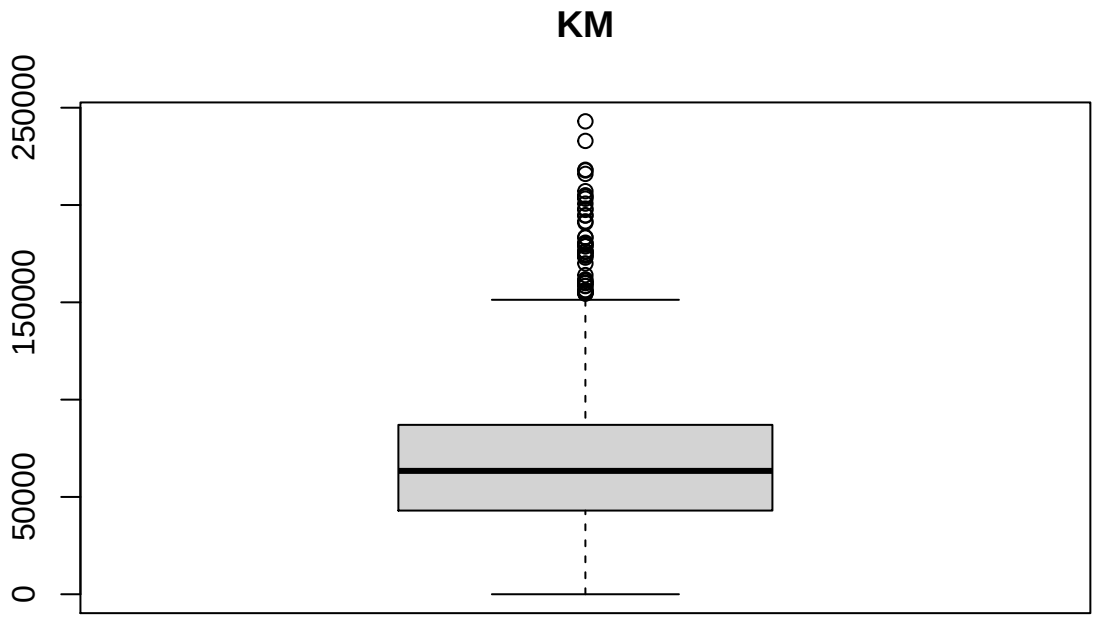

```
boxplot(deleteme$Price, main = "Price")
```



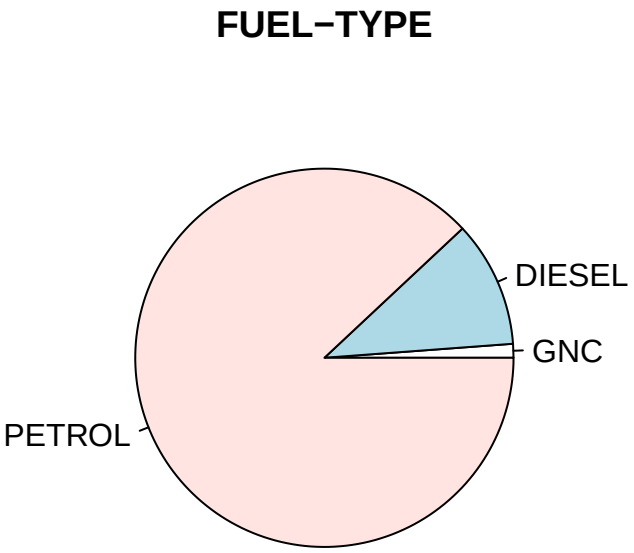

```
pie(summary(deleteme$Mfg_Year), labels = c("1998","1999","2000","2001", "2001", "2002", "2003", "2004",
```



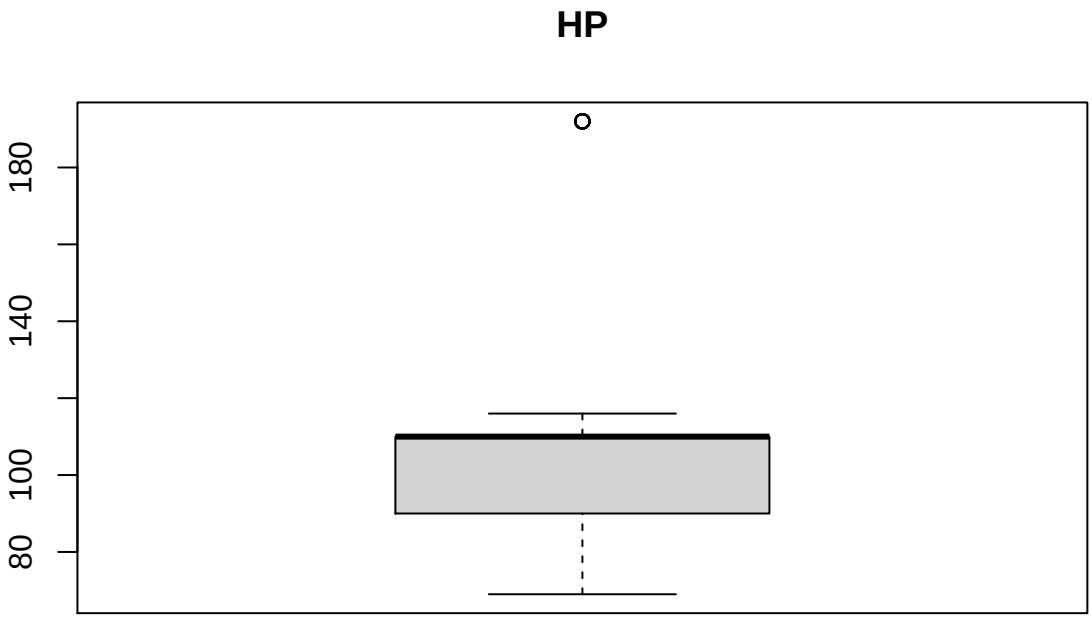

```
boxplot(deleteme$KM, main = "KM")
```



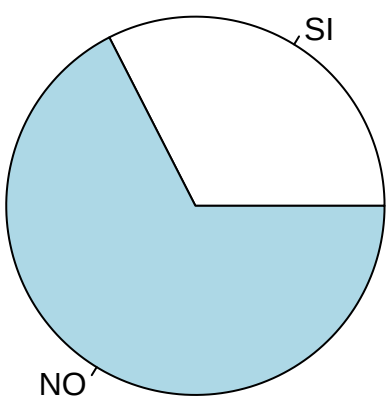

```
pie(summary(deleteme$Fuel_Type), labels = c("GNC", "DIESEL", "PETROL"), main ="FUEL-TYPE")
```



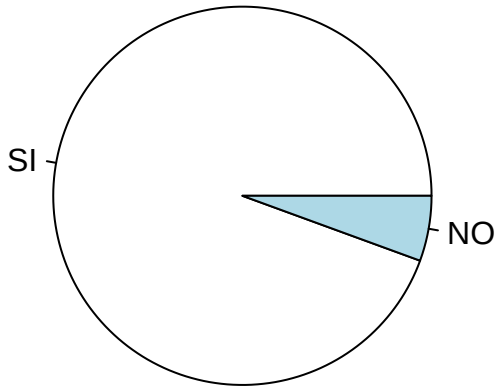

```
boxplot(deleteme$HP, main = "HP")
```



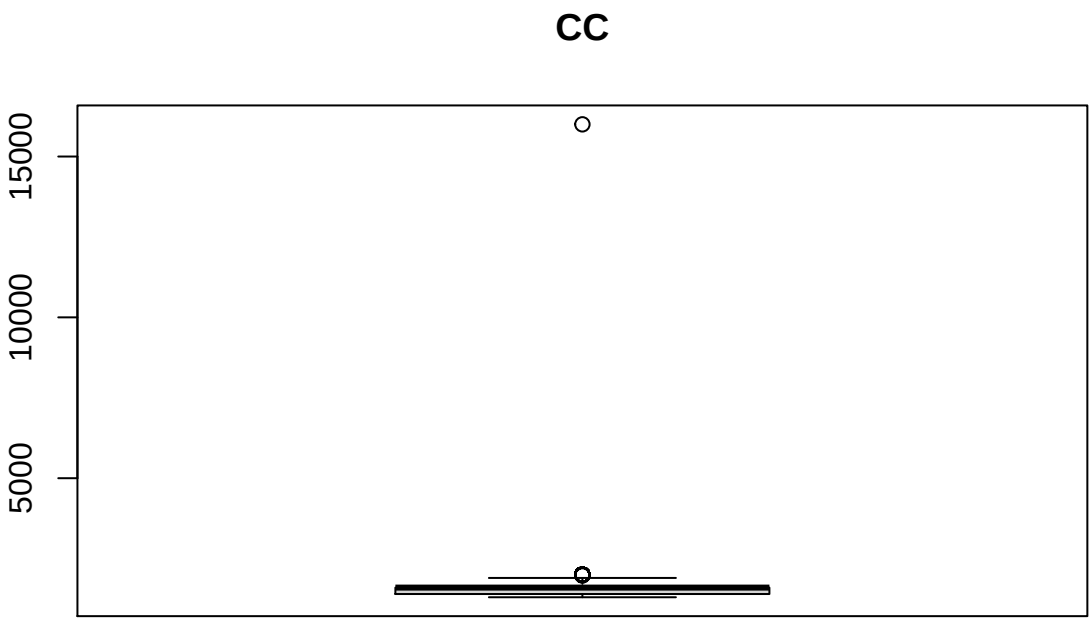

```
pie(summary(deleteme$Met_Color), labels = c("SI", "NO"), main = "MET-COLOR")
```



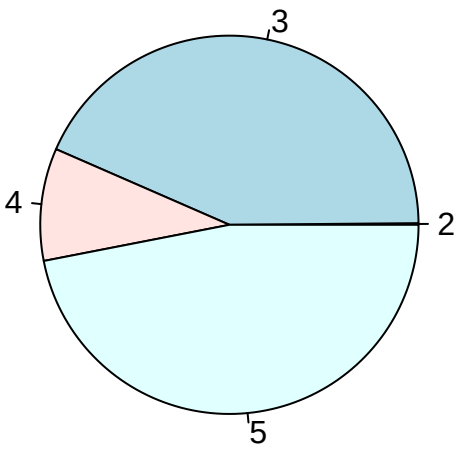

```
pie(summary(deleteme$Automatic), labels = c("SI", "NO"), main = "AUTOMATIC")
```




```
boxplot(deleteme$cc, main="CC")
```

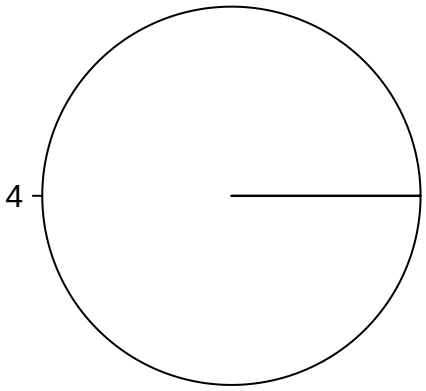



```
pie(summary(deleteme$Doors), labels = c("2", "3", "4", "5"), main = "DOORS")
```

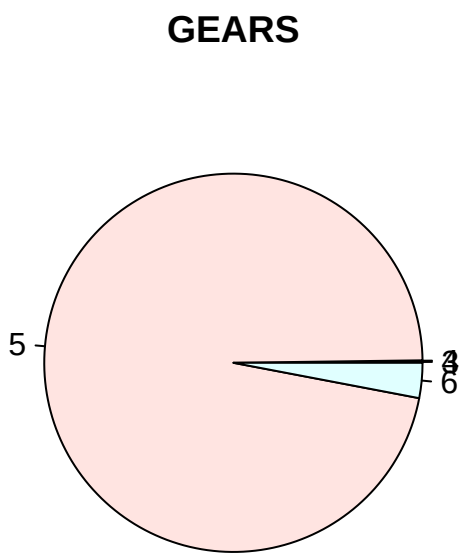



```
pie(summary(deleteme$Cylinders), labels =c("4", "otro"), main =  "CYLINDERS")
```

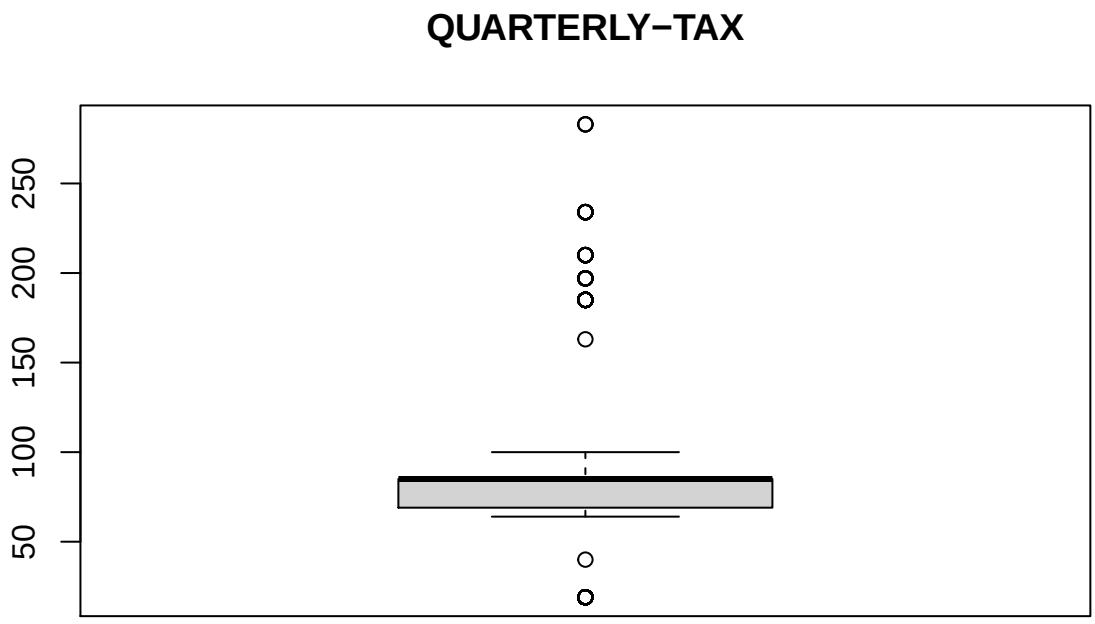
CYLINDERS



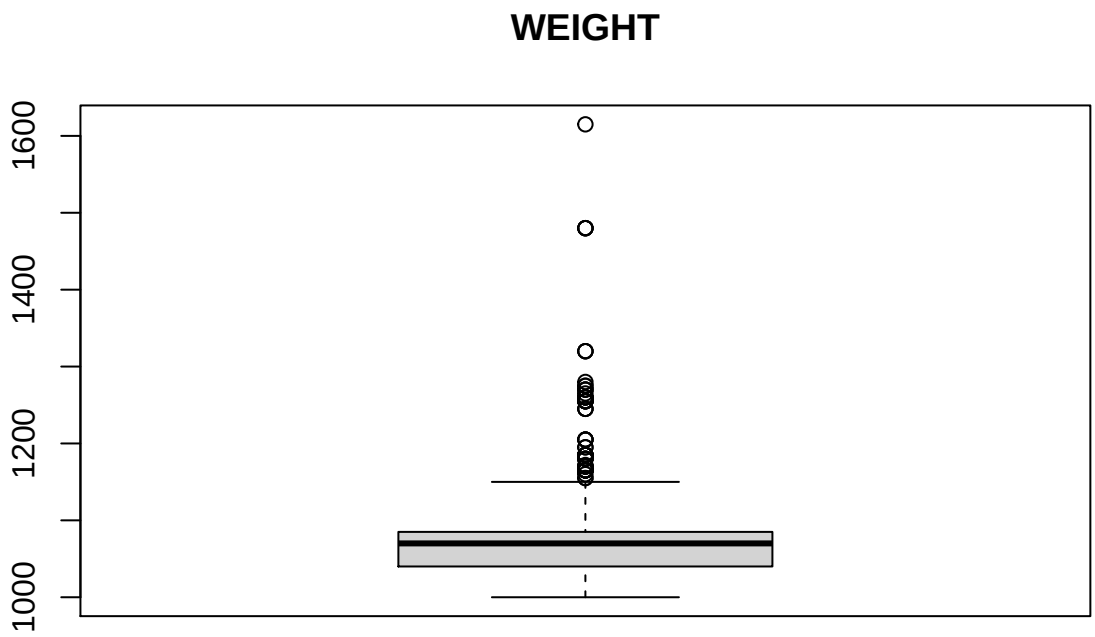

```
pie(summary(deleteme$Gears), labels = c("3", "4", "5", "6"), main = "GEARS")
```



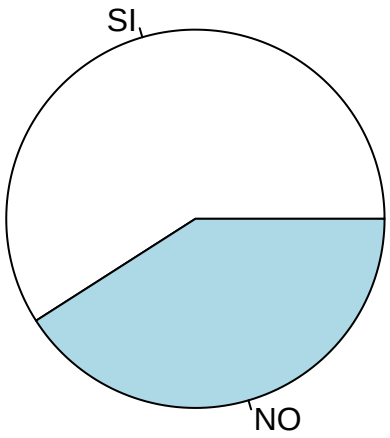

```
boxplot(deleteme$Quarterly_Tax, main = "QUARTERLY-TAX")
```



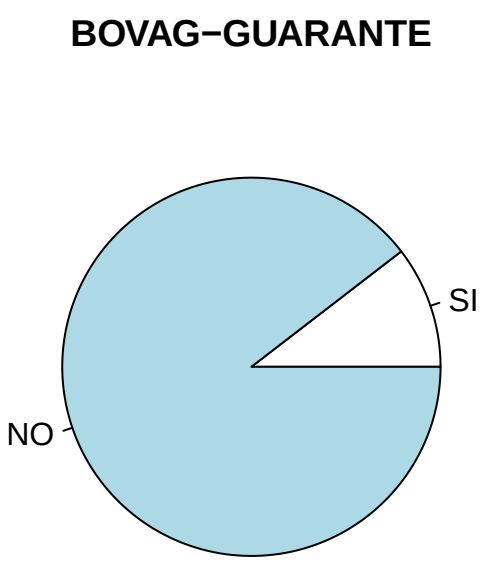

```
boxplot(deleteme$Weight, main = "WEIGHT")
```



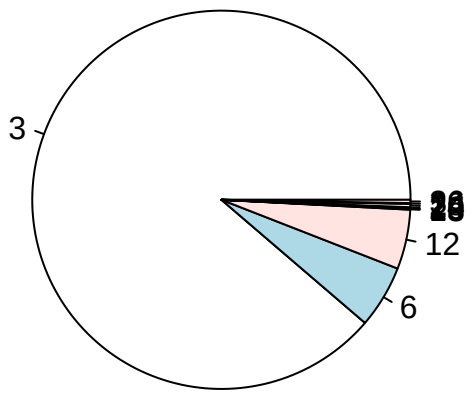

```
pie(summary(deleteme$Mfr_Guarantee), labels = c("SI", "NO"), main = "MFR-GUARANTE")
```



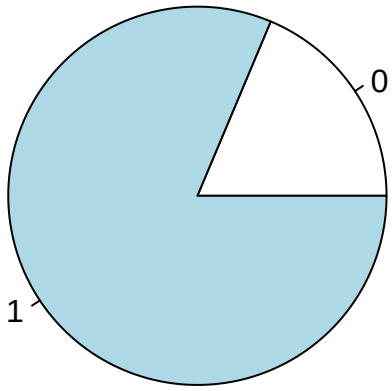

```
pie(summary(deleteme$BOVAG_Guarantee), labels = c("SI", "NO"), main = "BOVAG-GUARANTE")
```



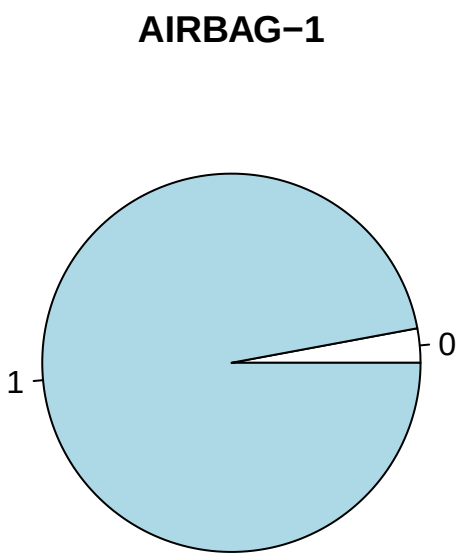

```
pie(summary(deleteme$Guarantee_Period), main="GUARANTE-PERIOD")
```



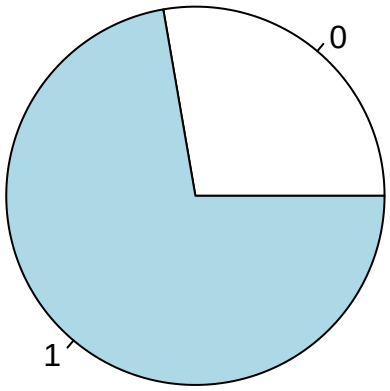

```
pie(summary(deleteme$ABS), main="ABS")
```



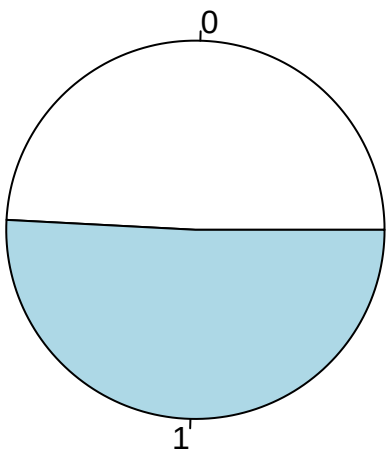

```
pie(summary(deleteme$Airbag_1), main = "AIRBAG-1")
```



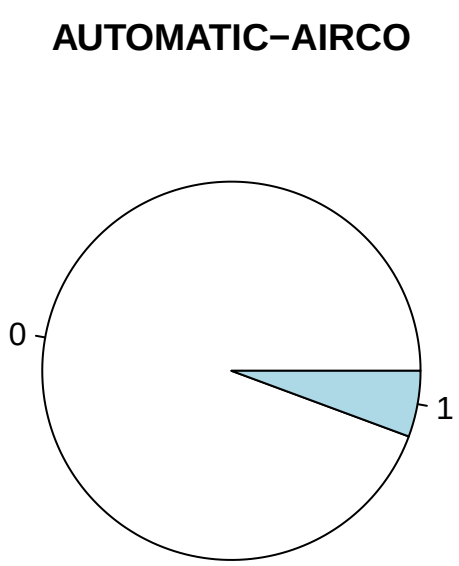

```
pie(summary(deleteme$Airbag_2), main = "AIRBAG-2")
```



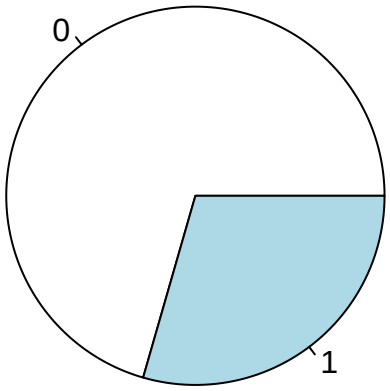

```
pie(summary(deleteme$Airco), main = "AIRCO")
```



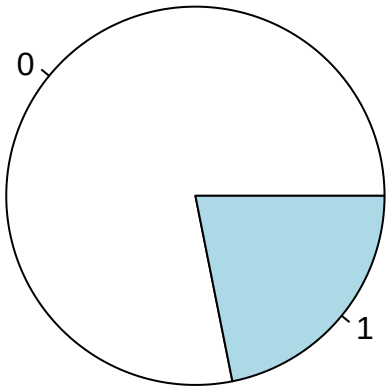

```
pie(summary(deleteme$Automatic_airco), main = "AUTOMATIC-AIRCO")
```



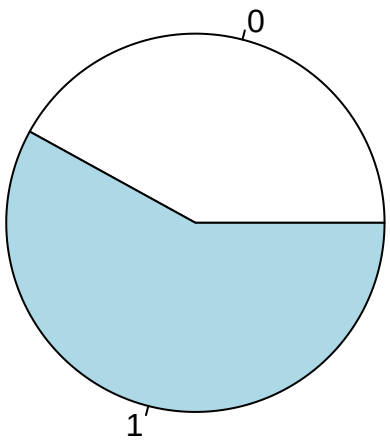

```
pie(summary(deleteme$Boardcomputer), main = "BOARDCOMPUTER")
```



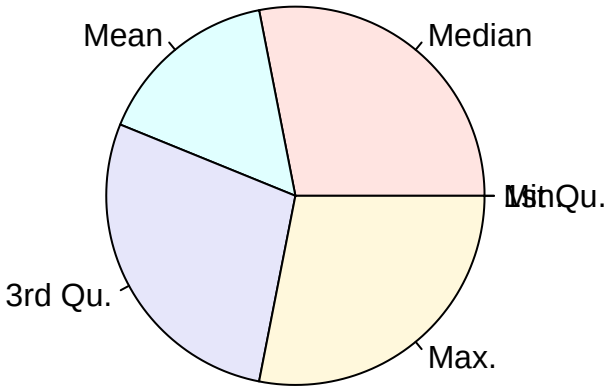

```
pie(summary(deleteme$CD_Player), main = "CD-PLAYER")
```



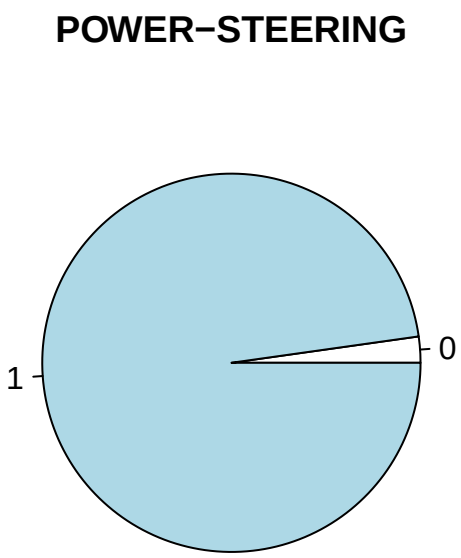

```
pie(summary(deleteme$Central_Lock), main = "CENTRAL-LOCK")
```



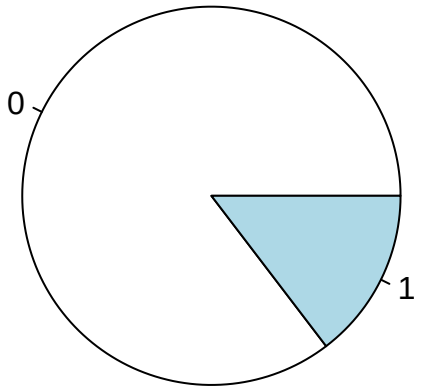

```
pie(summary(deleteme$Powered_Windows), main = "POWERED-WINDOWS")
```



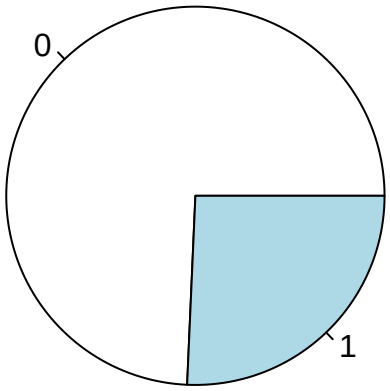

```
pie(summary(deleteme$Power_Steering), main = "POWER-STEERING")
```



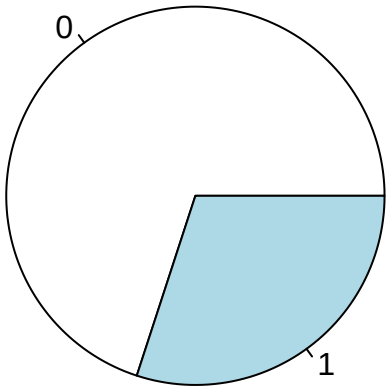

```
pie(summary(deleteme$Radio), main = "RADIO")
```



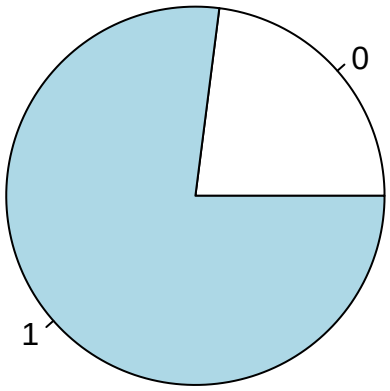

```
pie(summary(deleteme$Mistlamps), main = "MITSLAMPS")
```



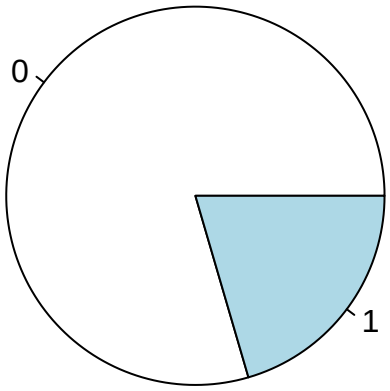

```
pie(summary(deleteme$Sport_Model), main = "SPORT-MODEL")
```



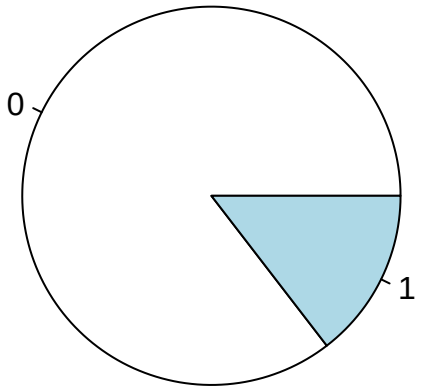

```
pie(summary(deleteme$Backseat_Divider), main = "BACKSEAT-DIVIDER")
```



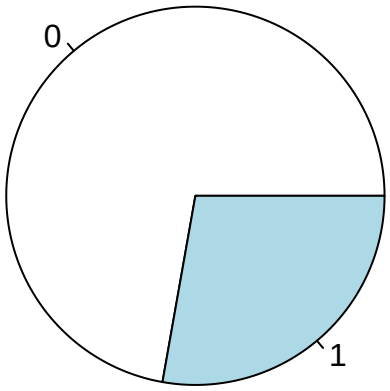

```
pie(summary(deleteme$Metallic_Rim), main = "METALIC-RIM")
```



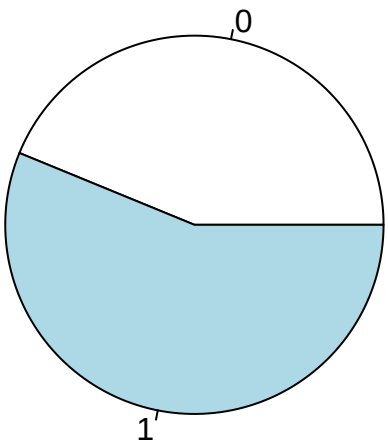

```
pie(summary(deleteme$Radio_cassette), main = "RADIO-CASSETTE")
```




```
pie(summary(deleteme$Tow_Bar), main = "TOW-BAR")
```

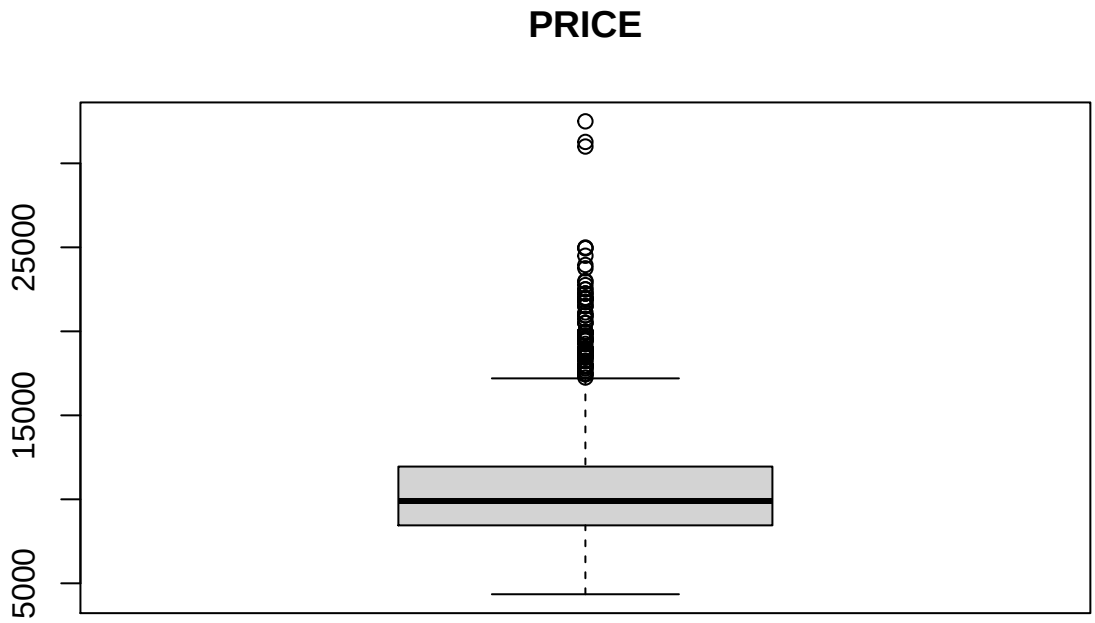



```
pie(summary(deleteme$Power_Windows), main = "POWER-WINDOWS")
```

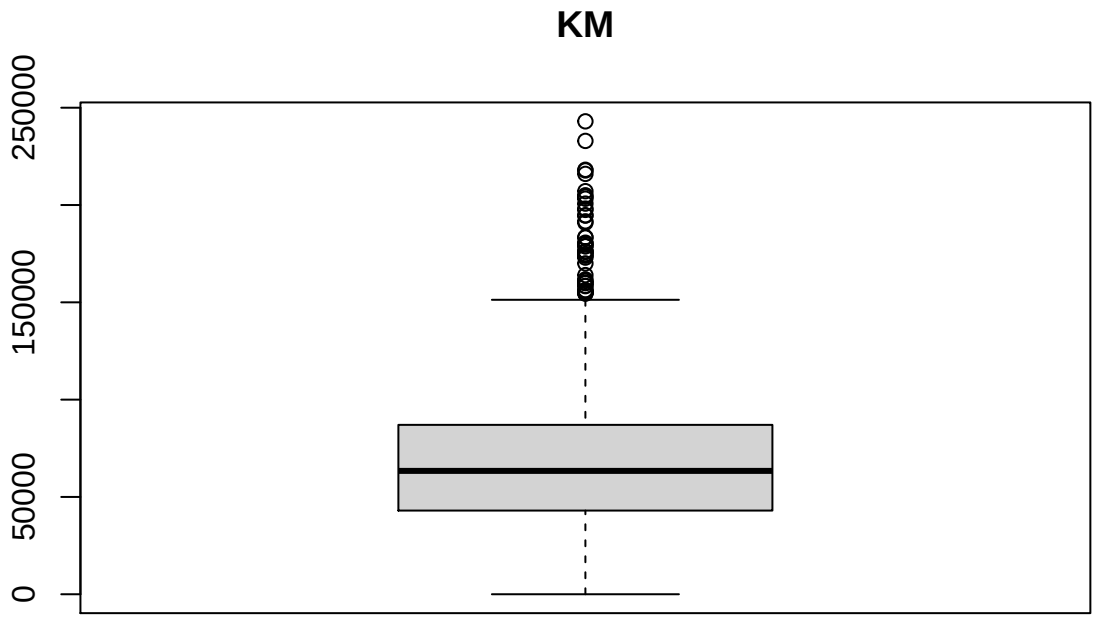


Distribución de las variables del dataset datos.

```
boxplot(datos$Price, main="PRICE")
```

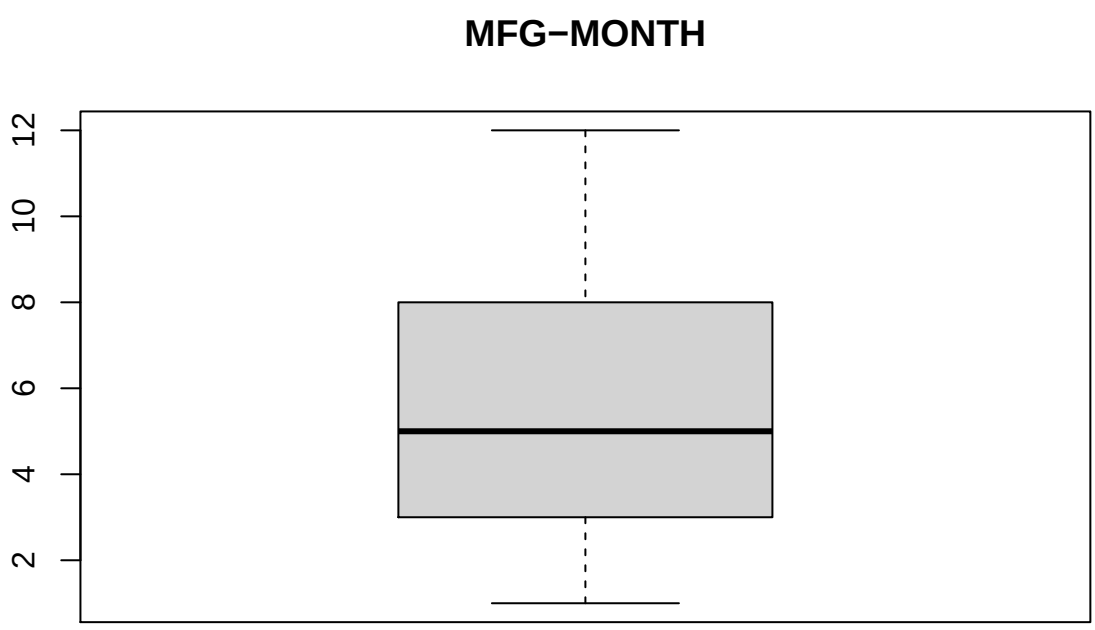



```
boxplot(datos$KM, main="KM")
```

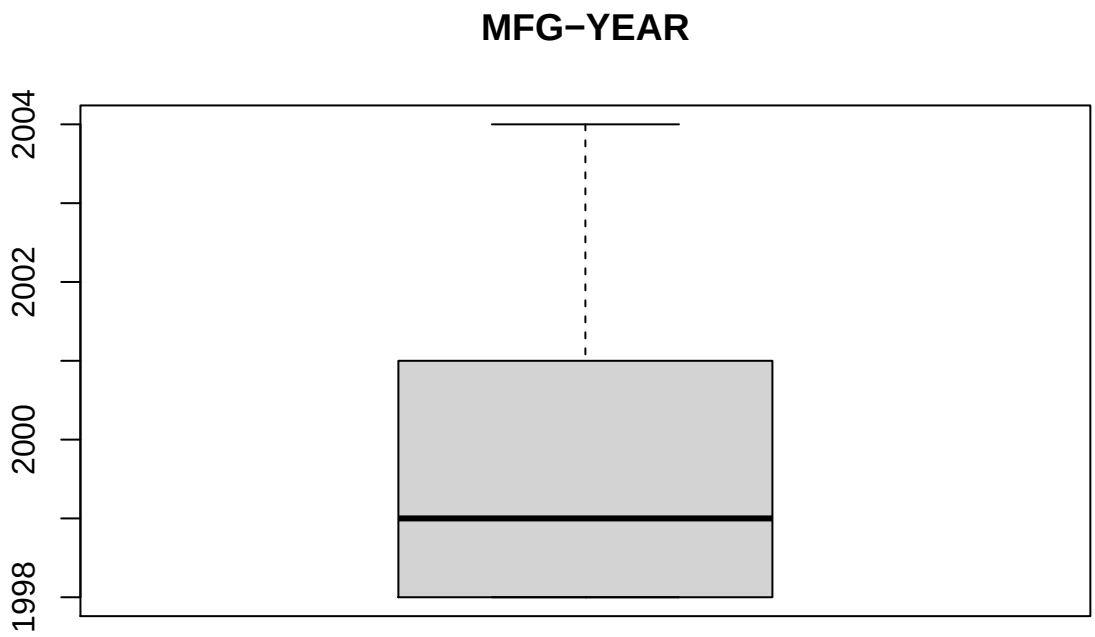



```
boxplot(datos$Age_08_04, main="AGE-08-04")
```

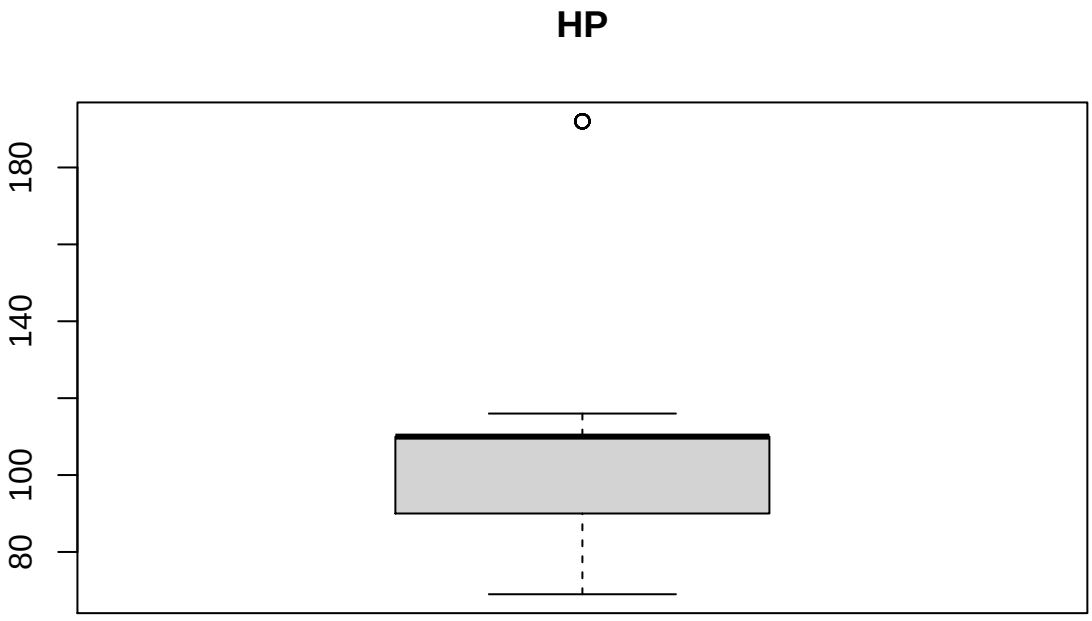

```
boxplot(datos$Mfg_Month, main="MFG-MONTH")
```



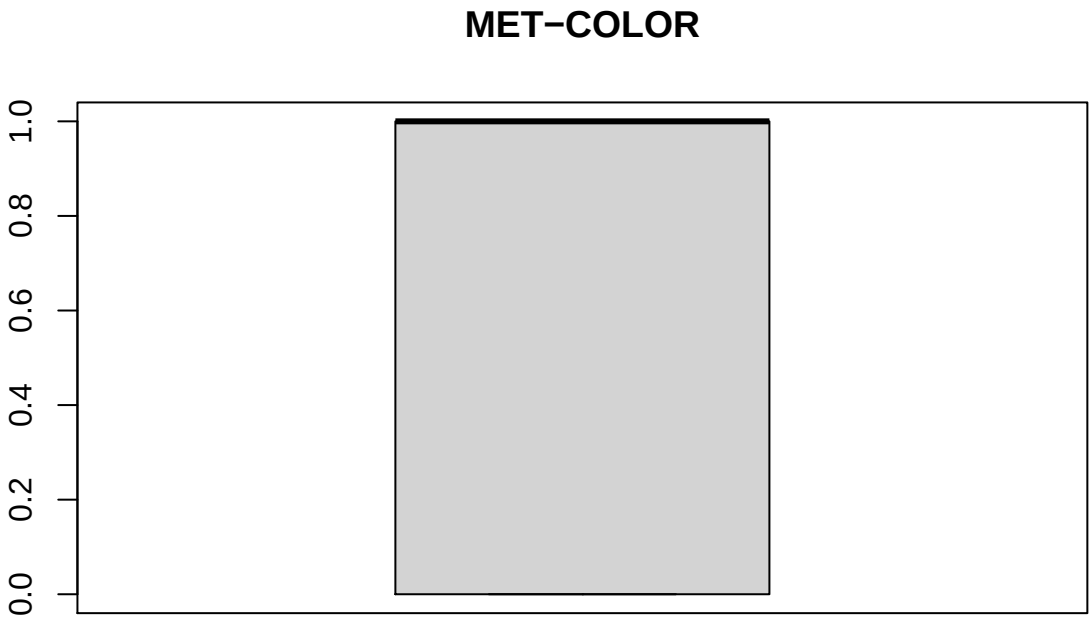

```
boxplot(datos$Mfg_Year, main="MFG-YEAR")
```



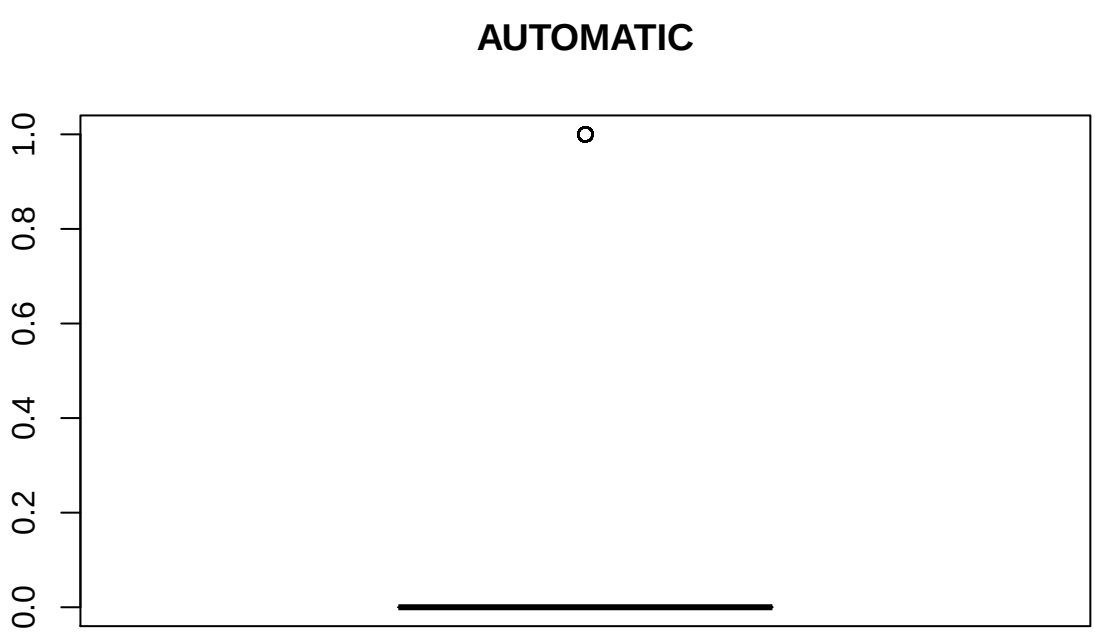

```
#boxplot(datos$Fuel_Type, main="FUEL-TYPE")
boxplot(datos$HP, main="HP")
```



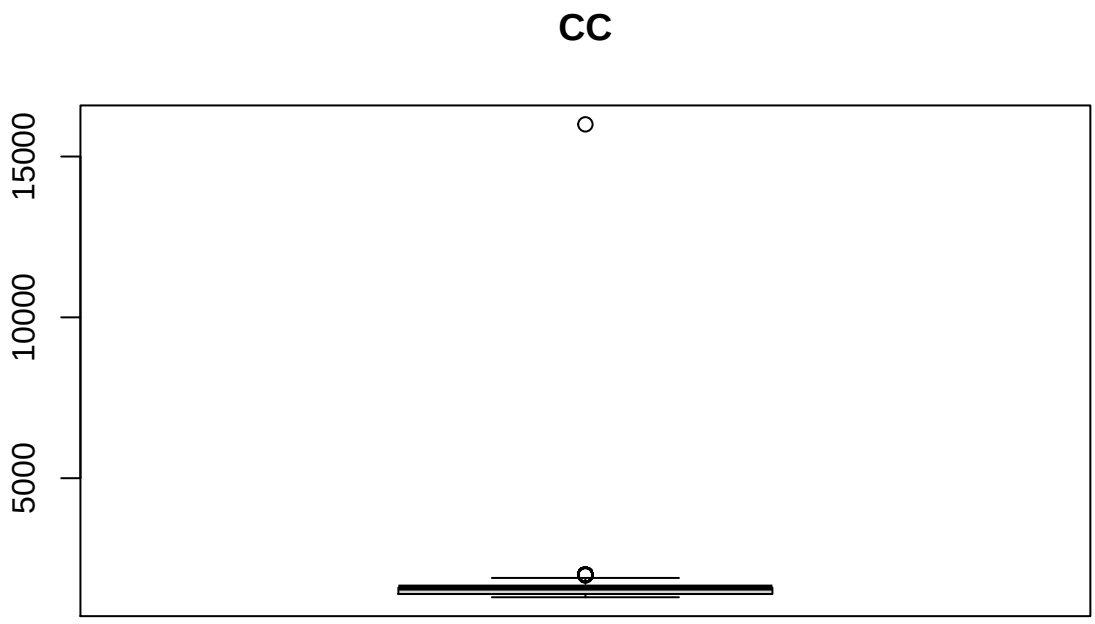

```
boxplot(datos$Met_Color, main="MET-COLOR")
```



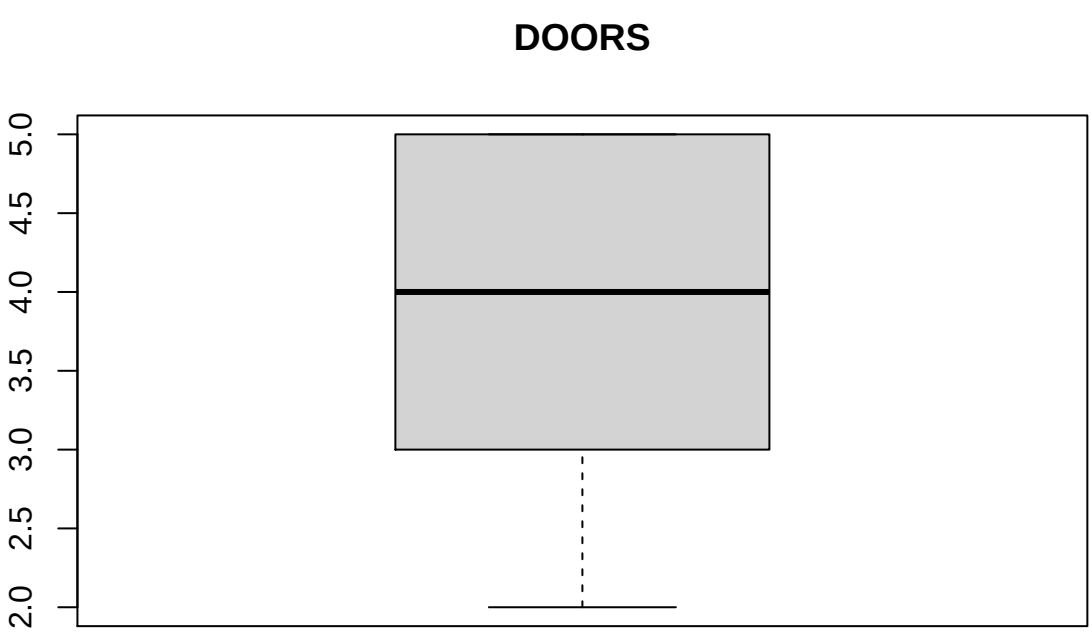

```
boxplot(datos$Automatic, main="AUTOMATIC")
```



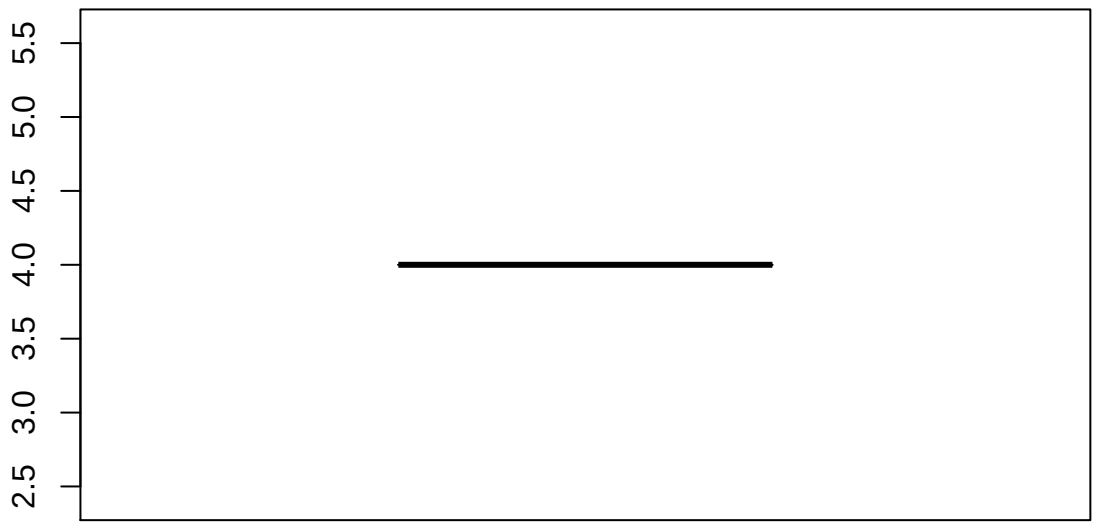

```
boxplot(datos$cc, main="CC")
```



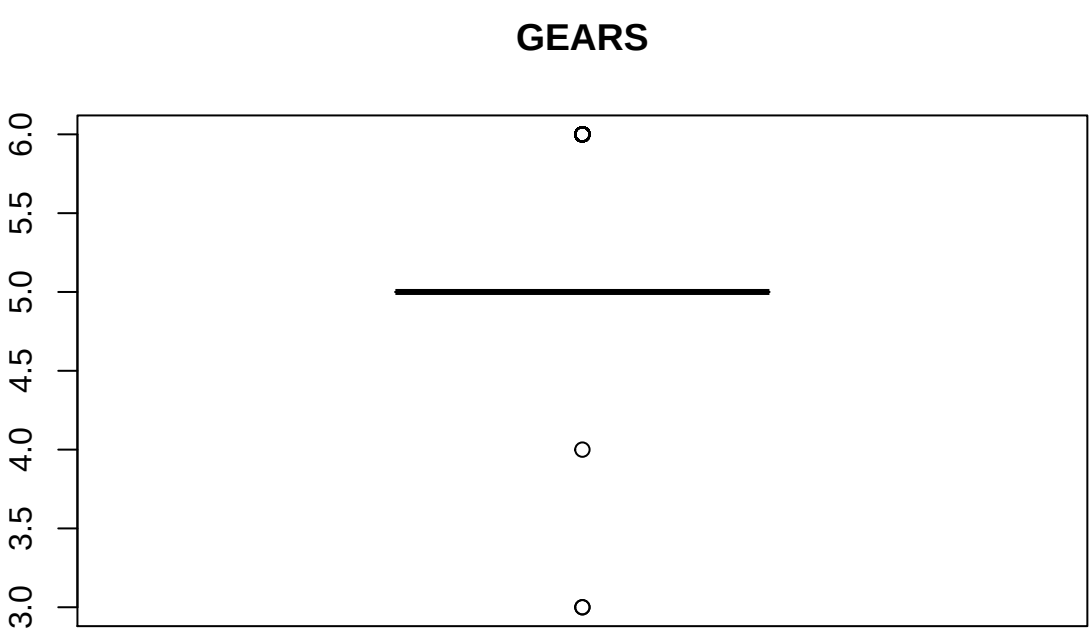

```
boxplot(datos$Doors, main="DOORS")
```



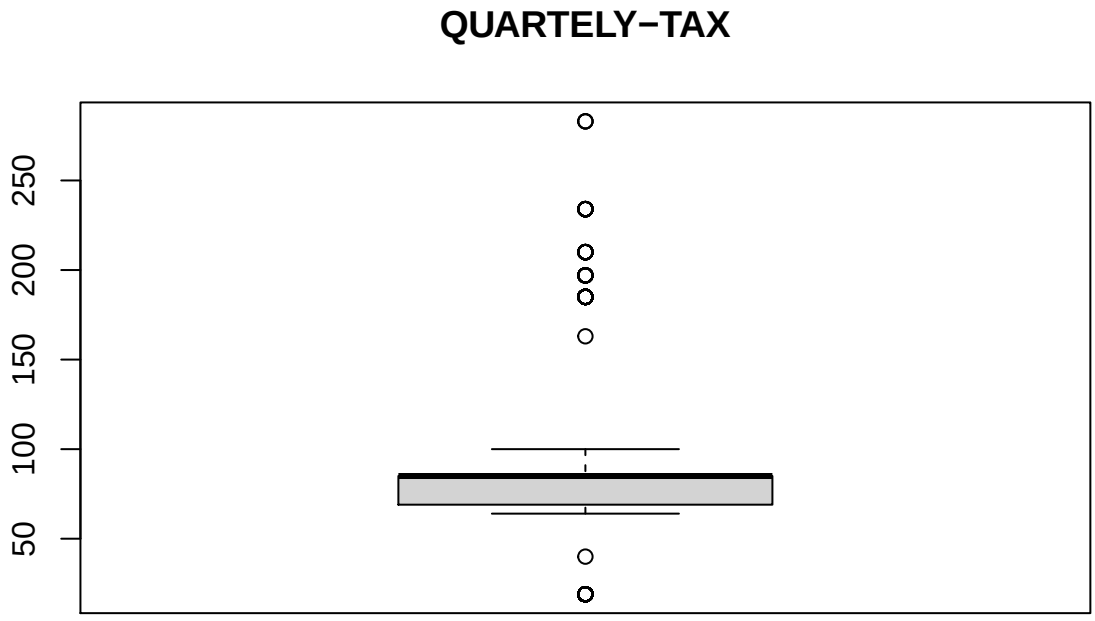

```
boxplot(datos$Cylinders, main="CYLINDERS")
```



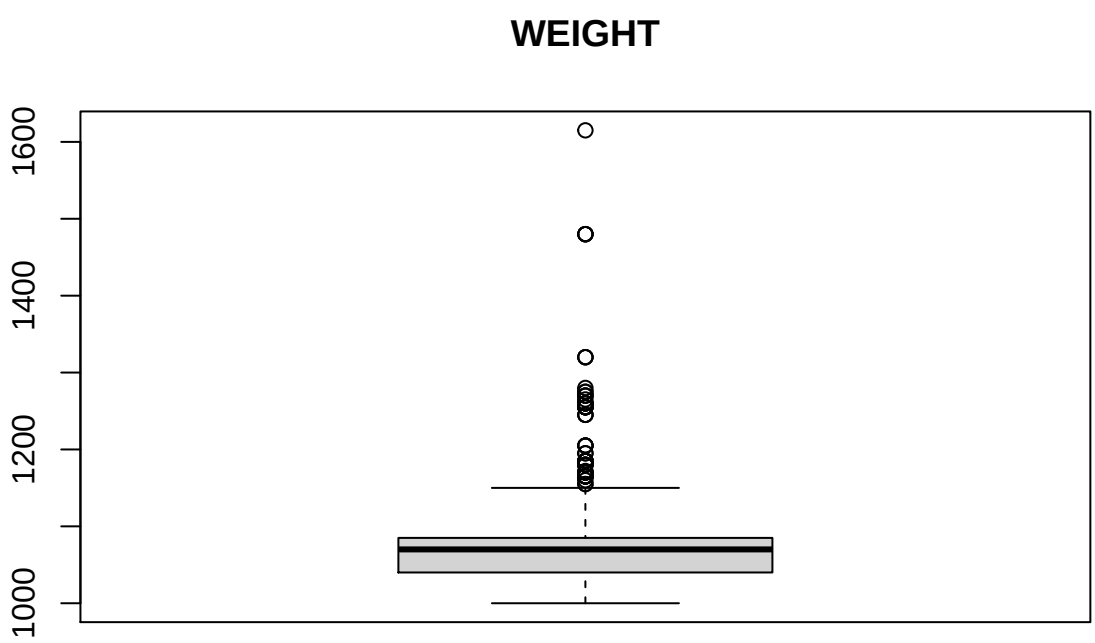

```
boxplot(datos$Gears, main="GEARS")
```



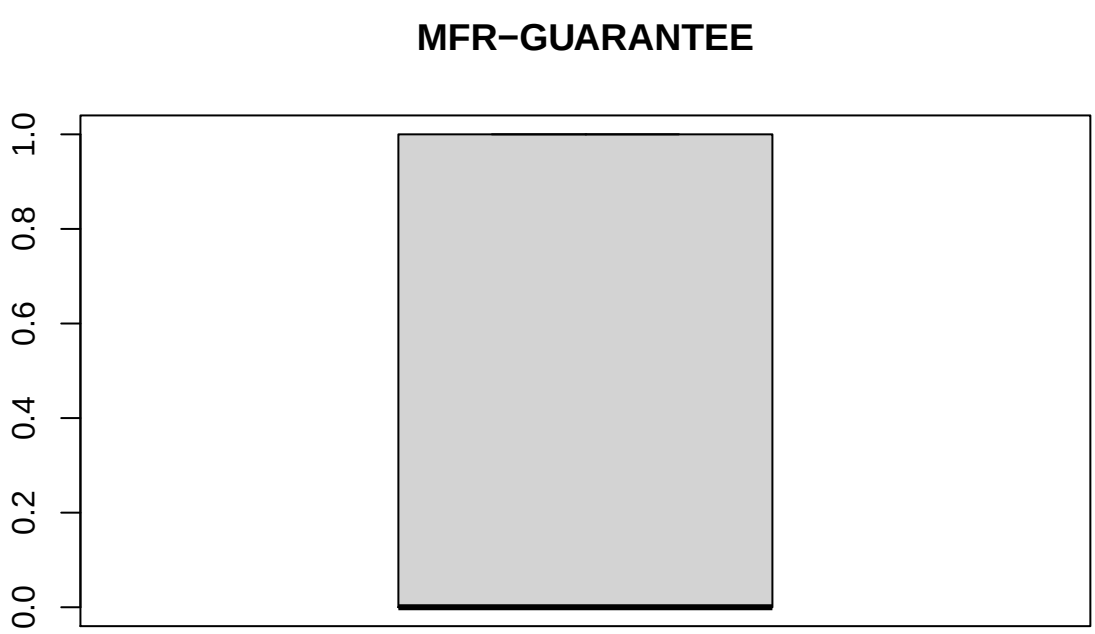

```
boxplot(datos$Quarterly_Tax, main="QUARTELY-TAX")
```



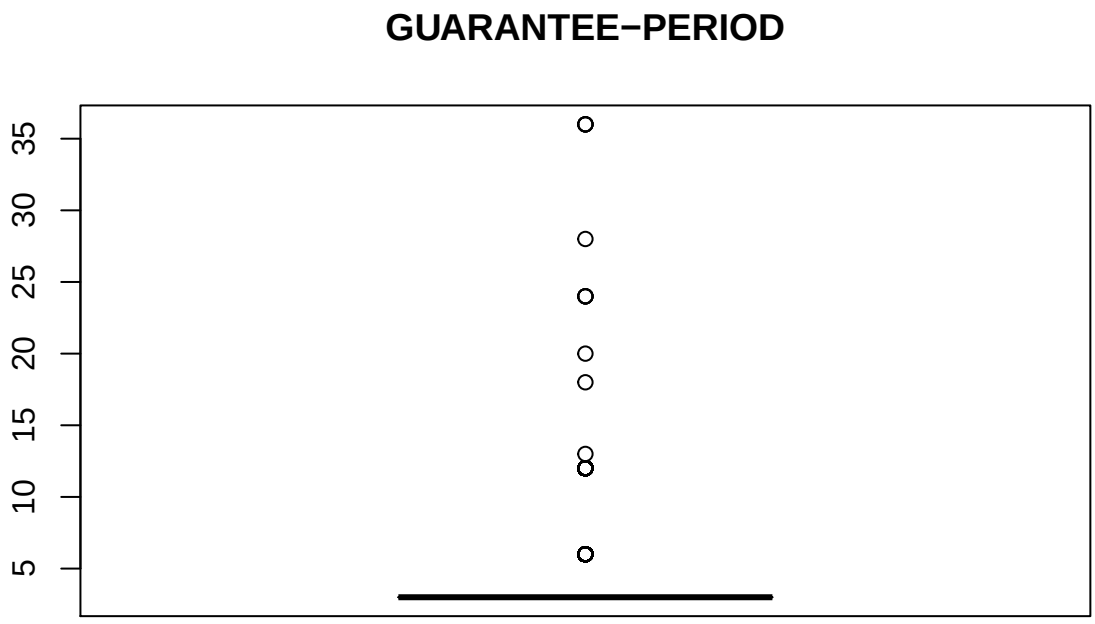

```
boxplot(datos$Weight, main="WEIGHT")
```



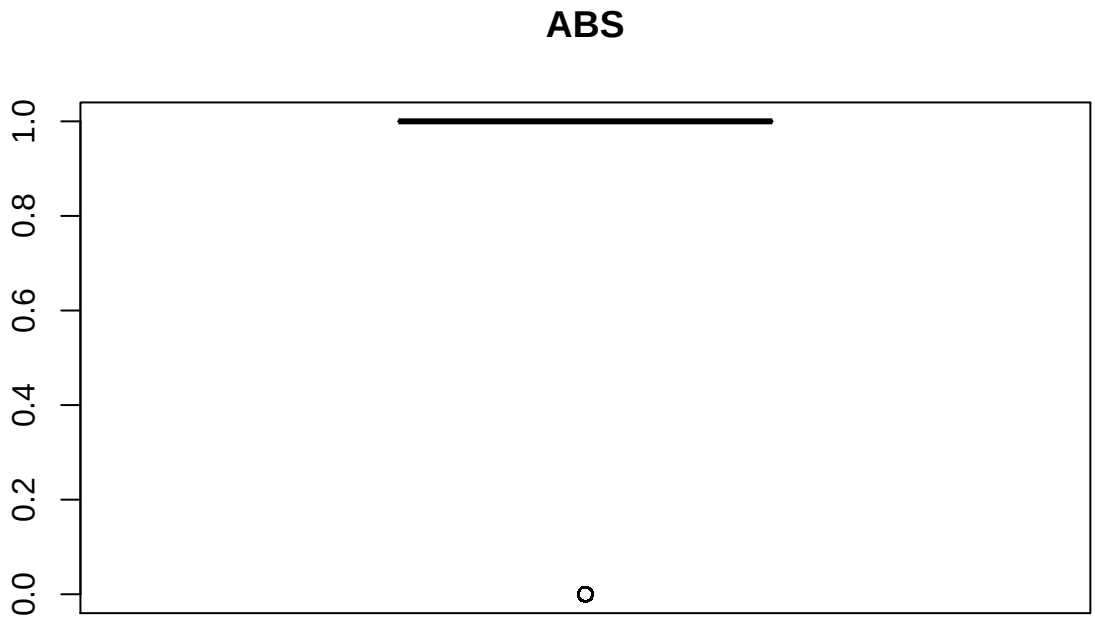

```
boxplot(datos$Mfr_Guarantee, main="MFR-GUARANTEE")
```



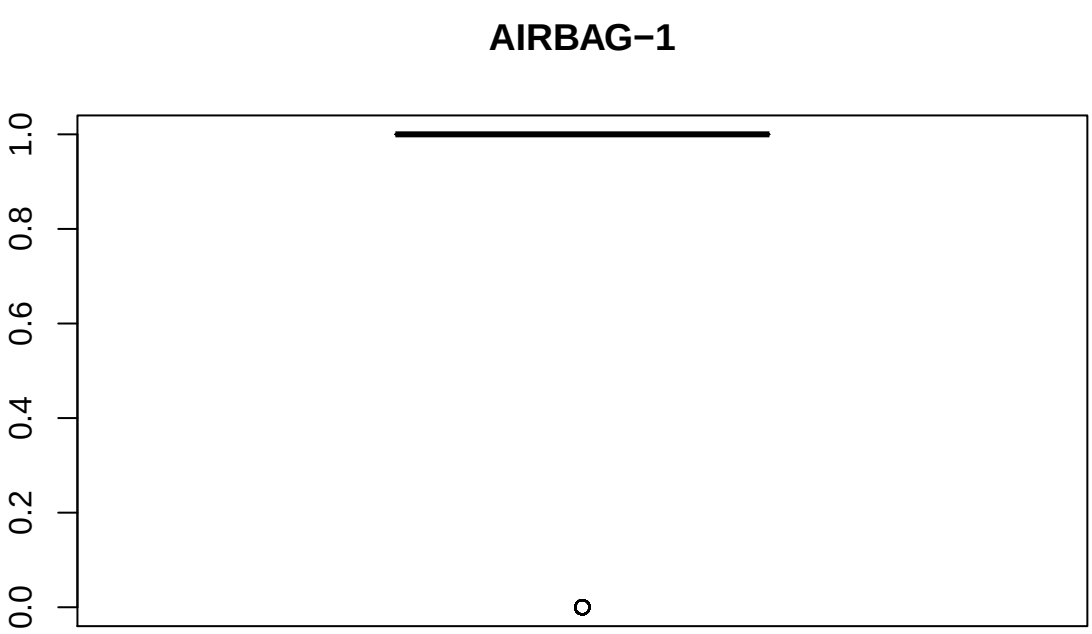

```
boxplot(datos$Guarantee_Period, main="GUARANTEE-PERIOD")
```



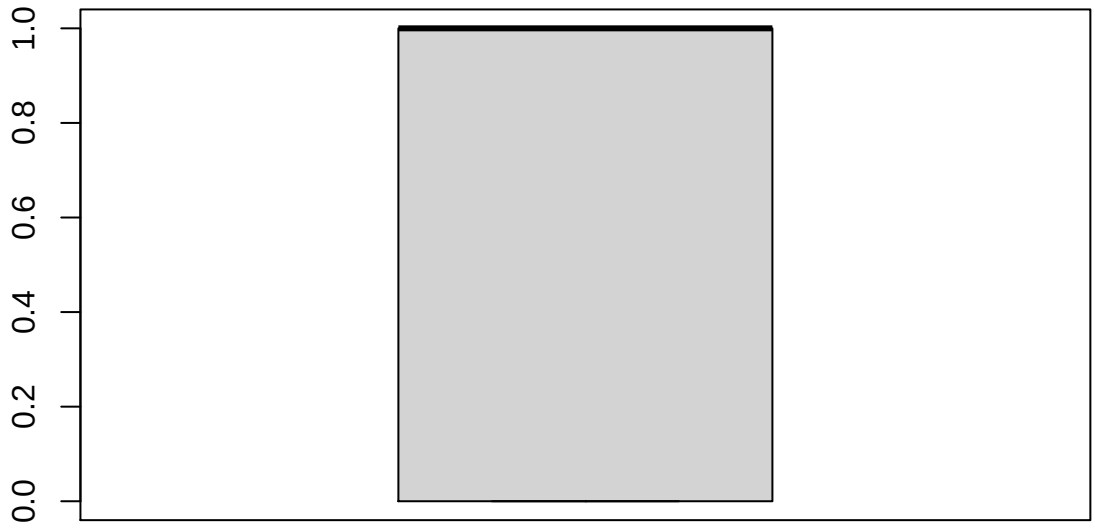

```
boxplot(datos$ABS, main="ABS")
```



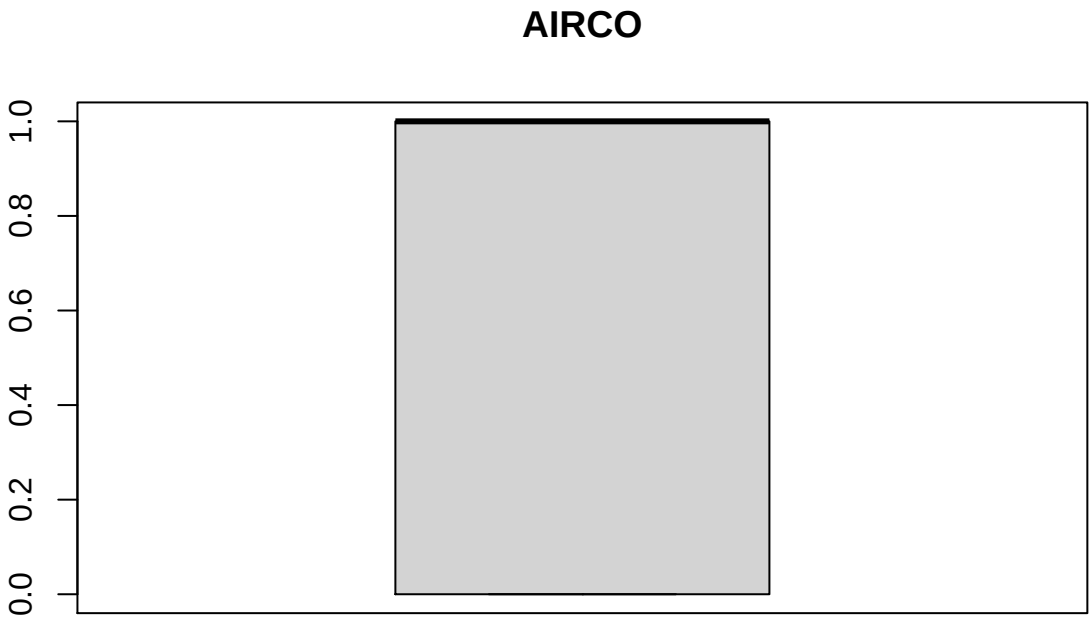

```
boxplot(datos$Airbag_1, main="AIRBAG-1")
```



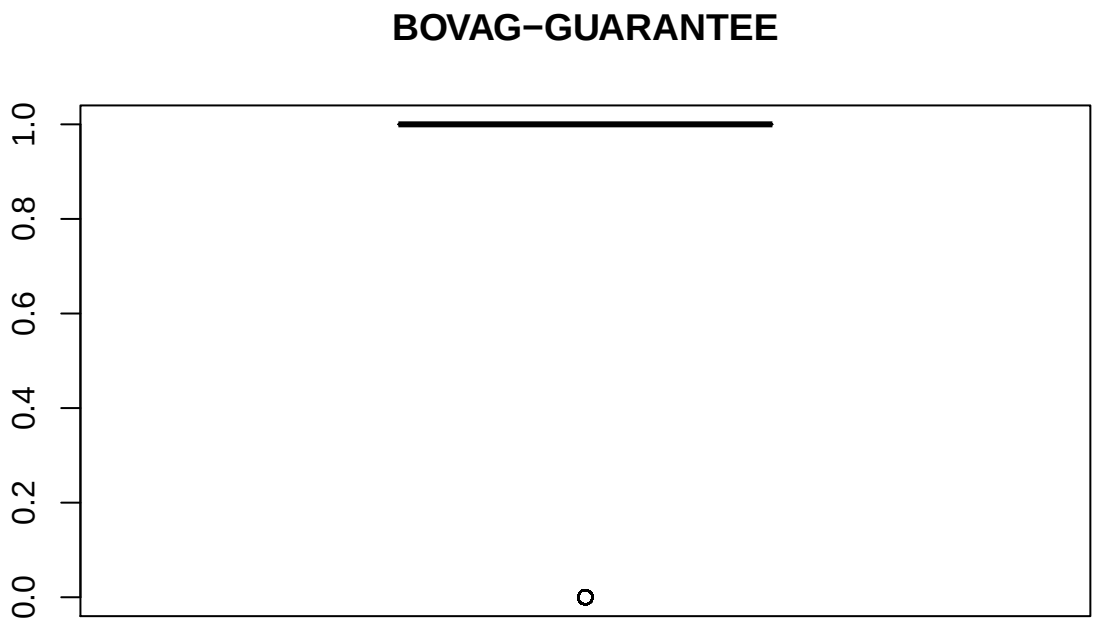

```
boxplot(datos$Airbag_2, main="AIRBAG-2")
```



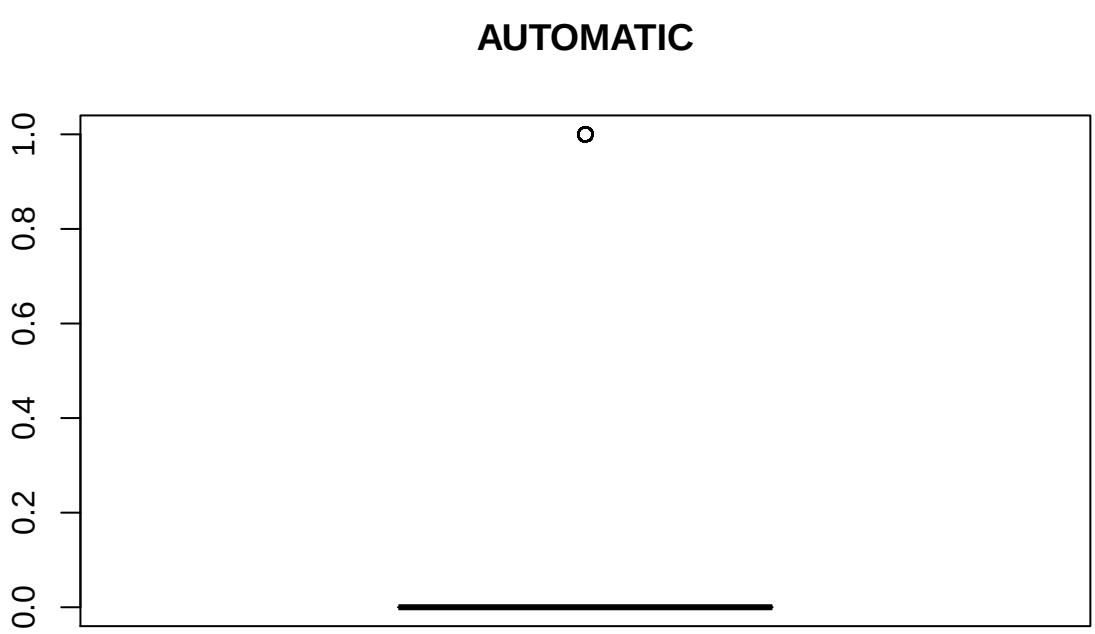

```
boxplot(datos$Airco, main="AIRCO")
```



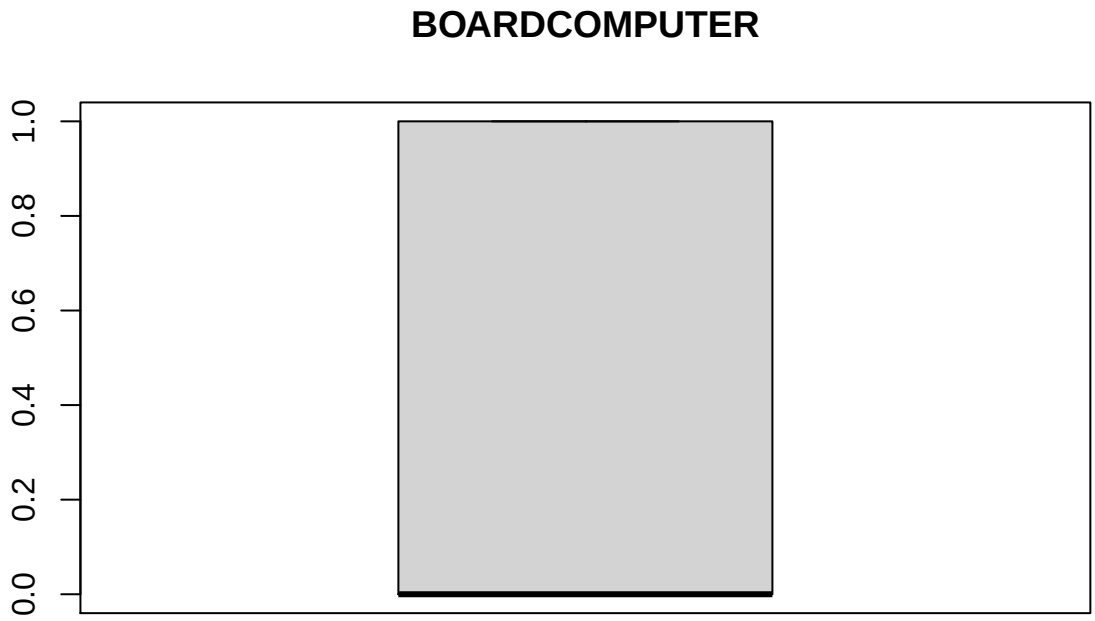

```
boxplot(datos$BOVAG_Guarantee, main="BOVAG-GUARANTEE")
```



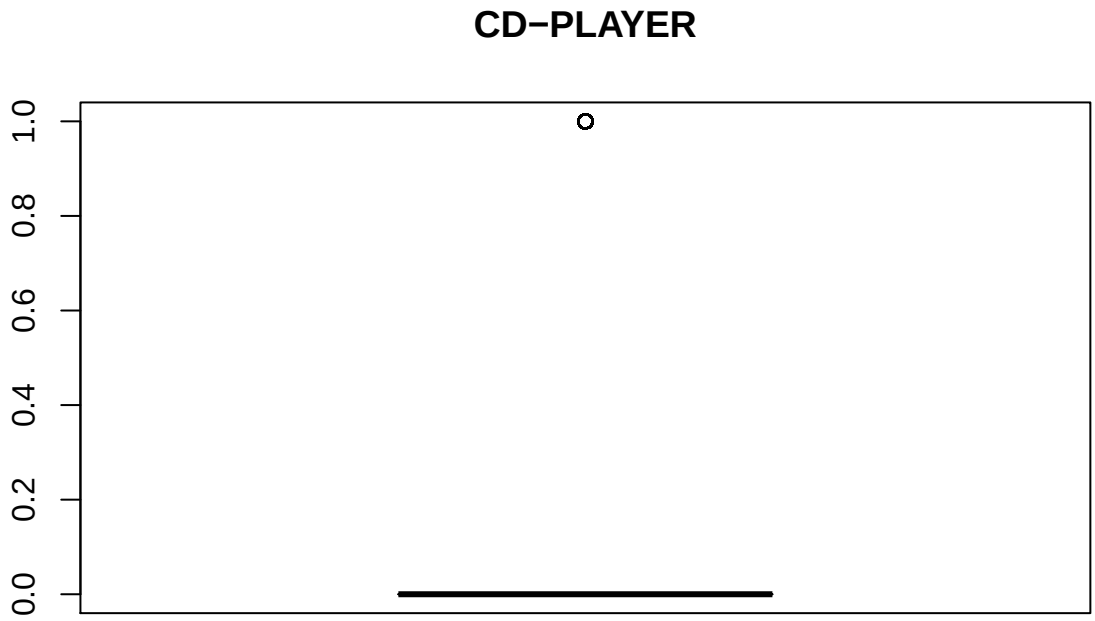

```
boxplot(datos$Automatic_airco, main="AUTOMATIC")
```



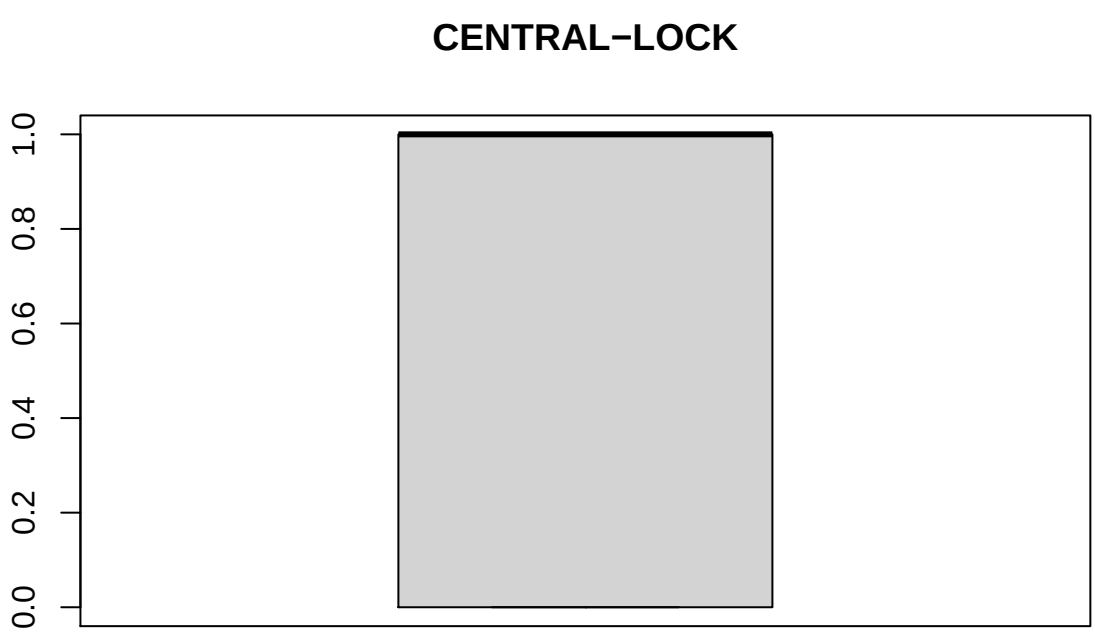

```
boxplot(datos$Boardcomputer, main="BOARDCOMPUTER")
```



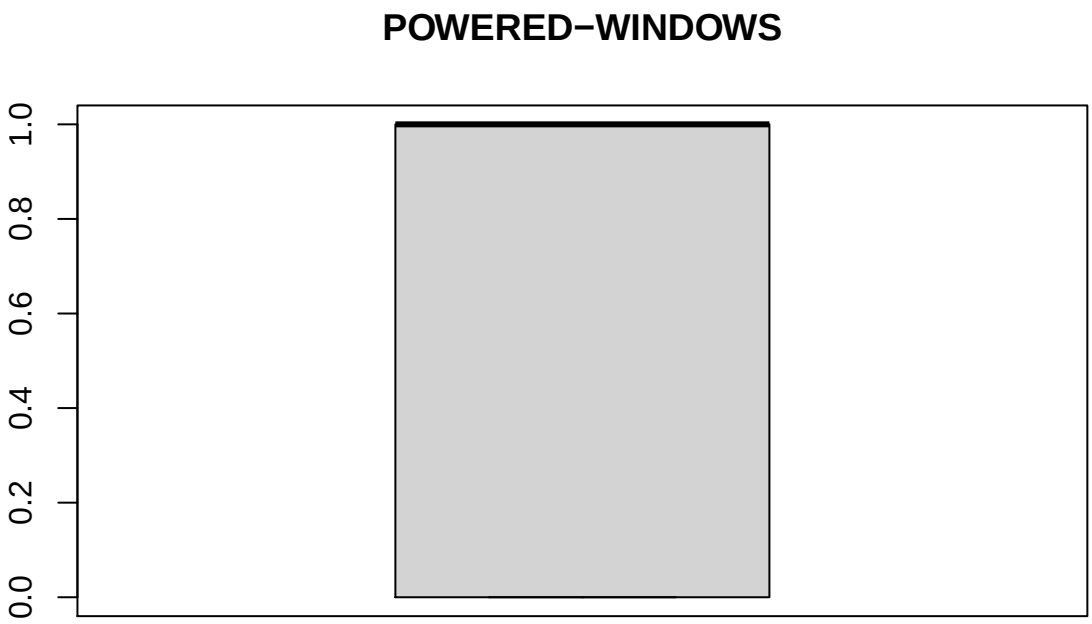

```
boxplot(datos$CD_Player, main="CD-PLAYER")
```



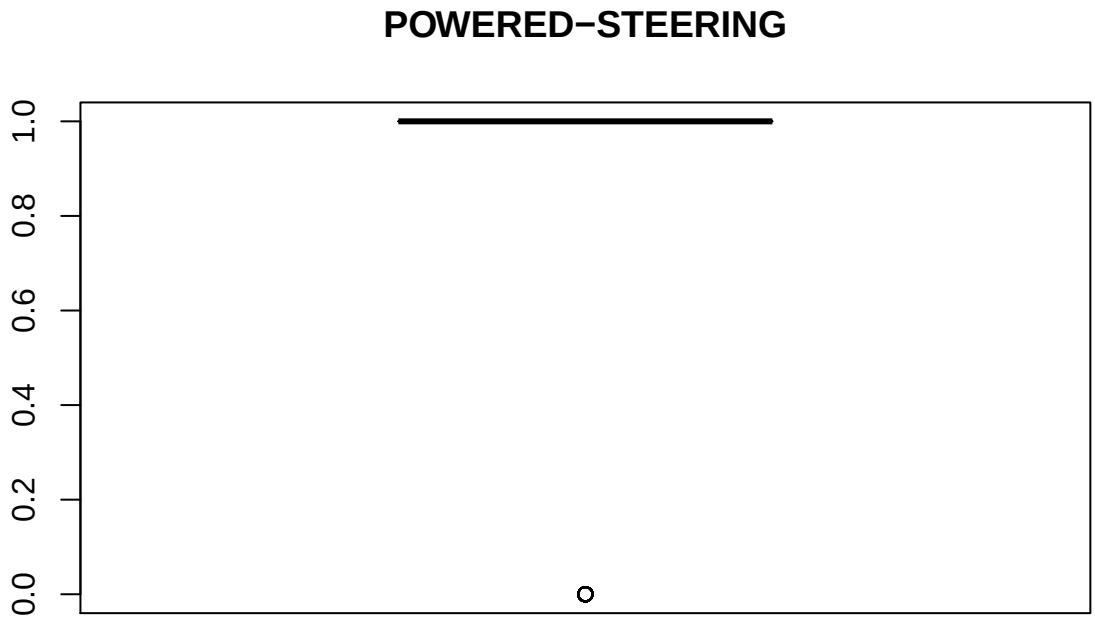

```
boxplot(datos$Central_Lock, main="CENTRAL-LOCK")
```



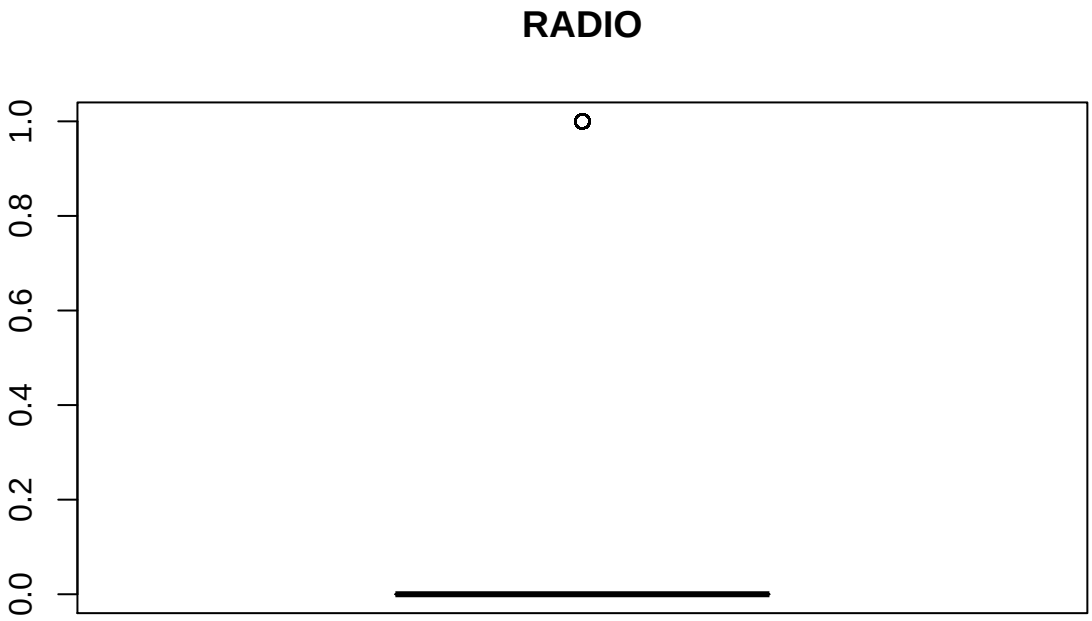

```
boxplot(datos$Powered_Windows, main="POWERED-WINDOWS")
```



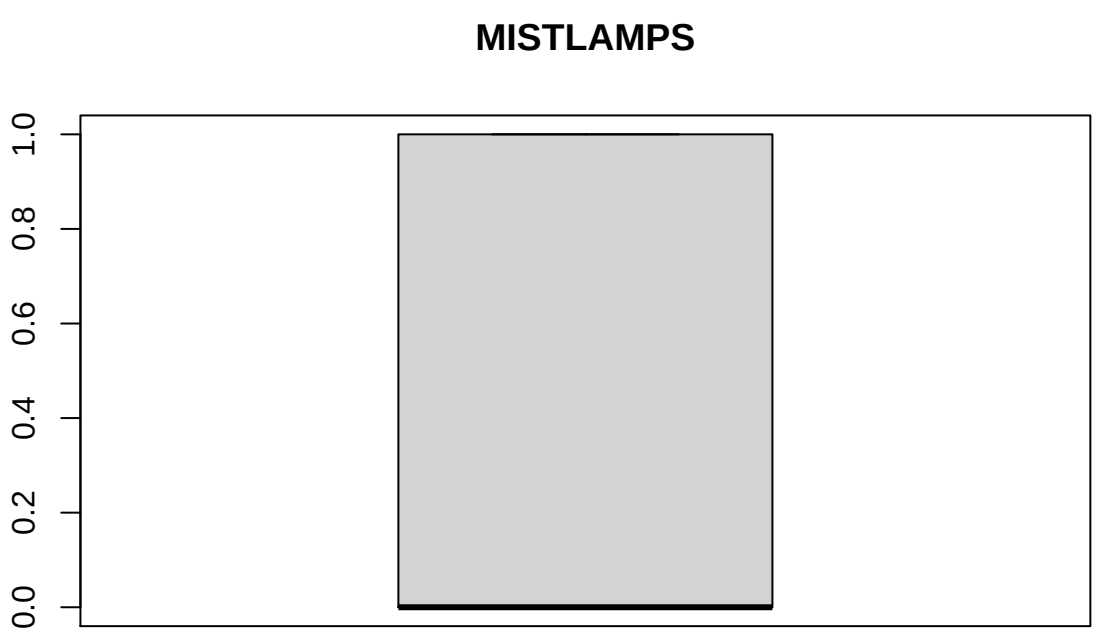

```
boxplot(datos$Power_Steering, main="POWERED-STEERING")
```



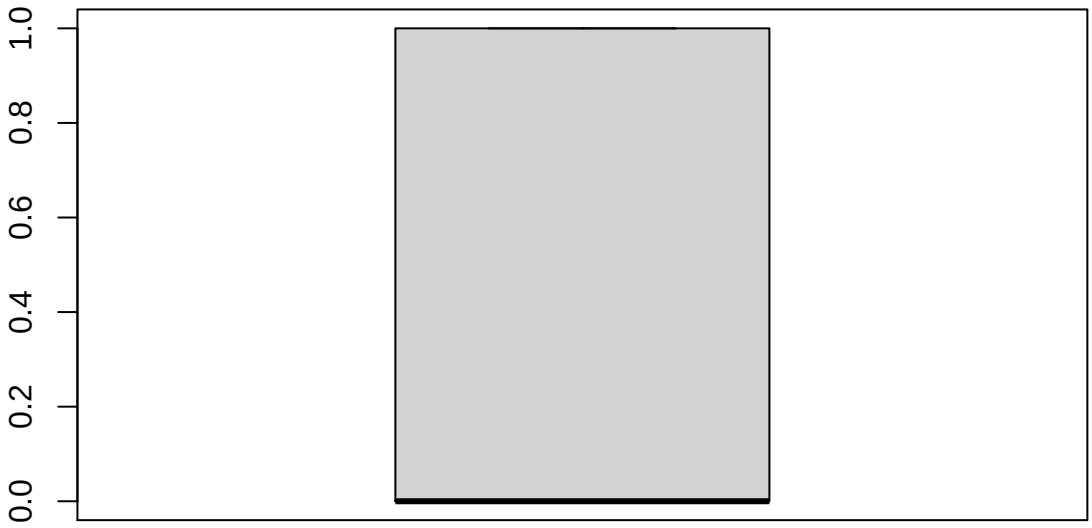

```
boxplot(datos$Radio, main="RADIO")
```



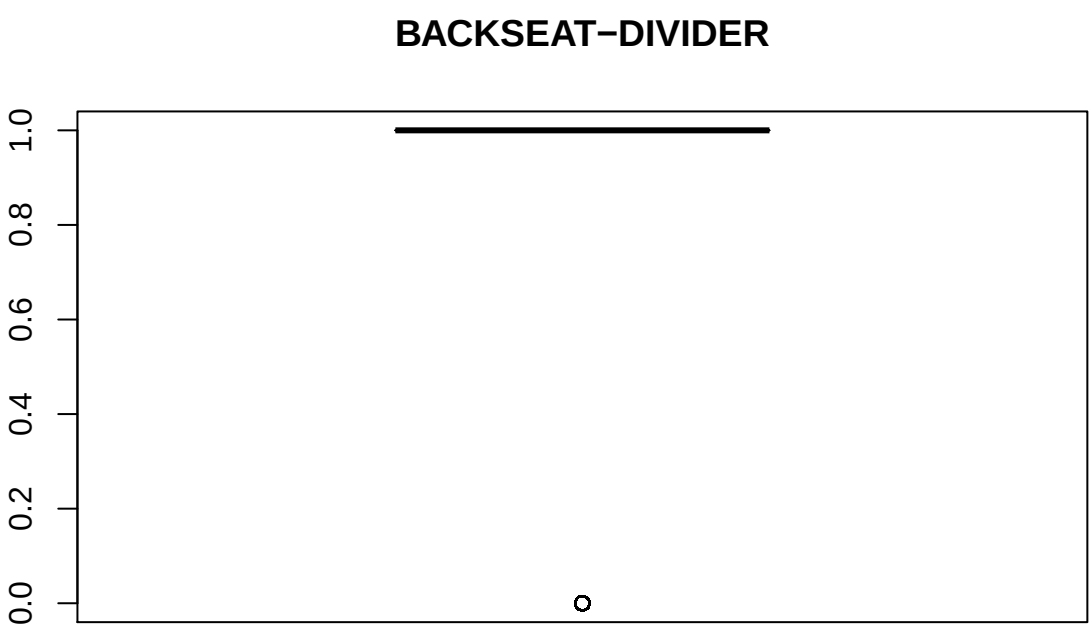

```
boxplot(datos$Mistlamps, main="MISTLAMPS")
```



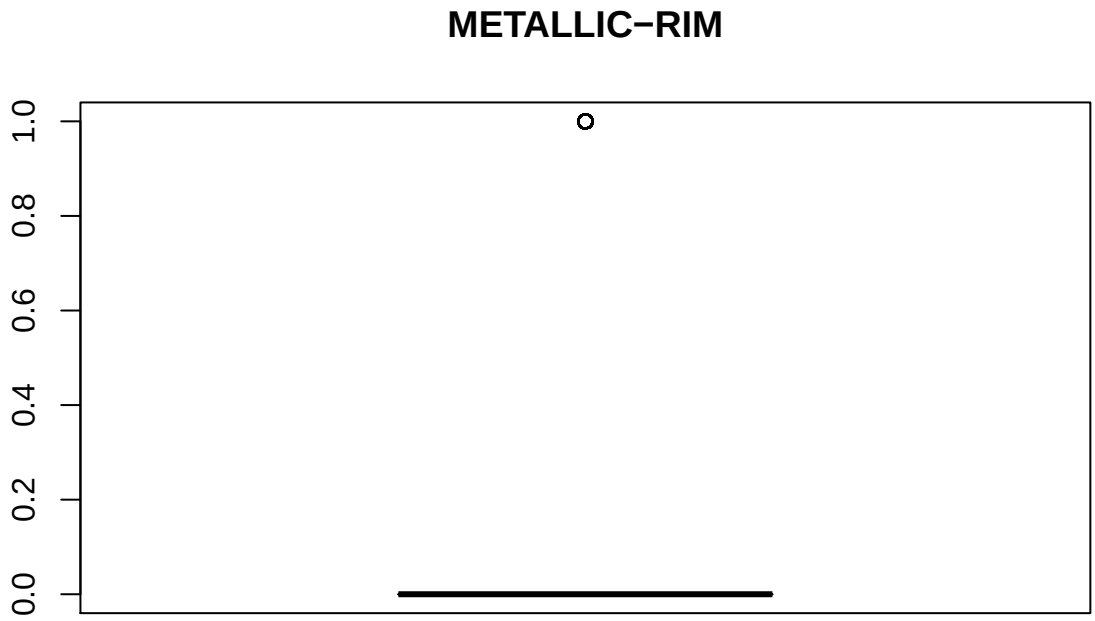

```
boxplot(datos$Sport_Model, main="SPORT-MODEL")
```



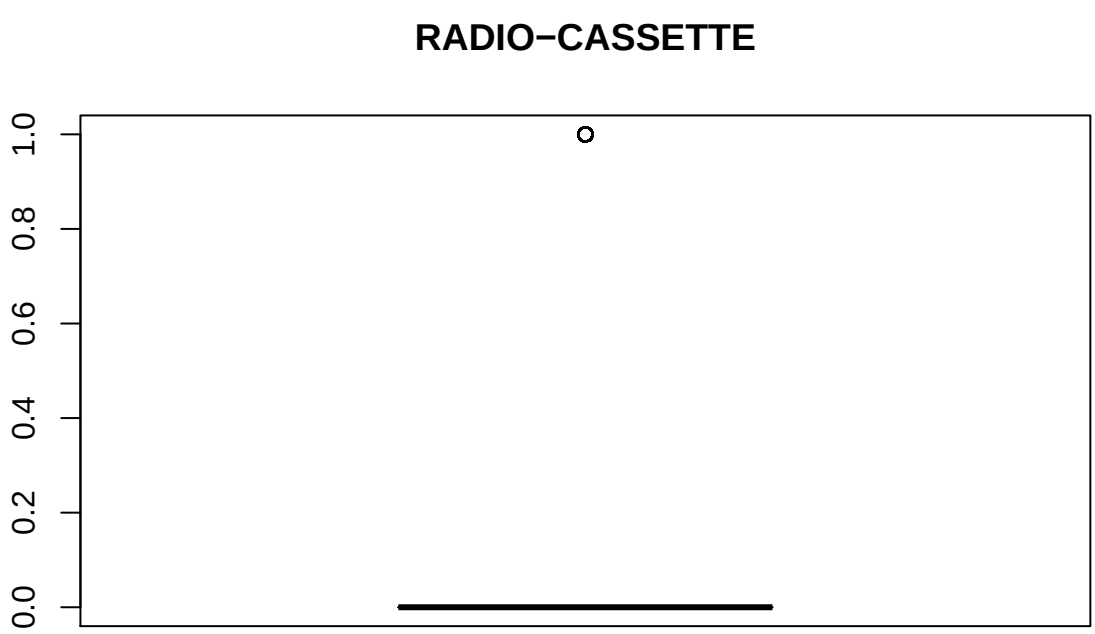

```
boxplot(datos$Backseat_Divider, main="BACKSEAT-DIVIDER")
```



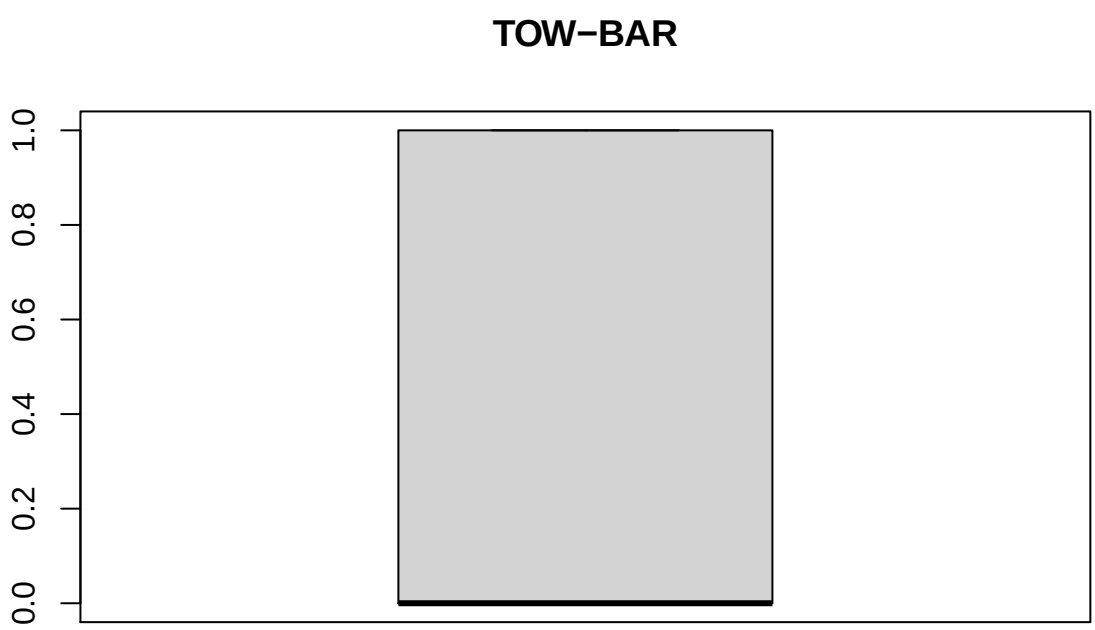

```
boxplot(datos$Metallic_Rim, main="METALLIC-RIM")
```




```
boxplot(datos$Radio_cassette, main="RADIO-CASSETTE")
```




```
boxplot(datos$Tow_Bar, main="TOW-BAR")
```



Notamos en las distribuciones que hay muchas variables binarias y que las variables que tienen datos continuos presentan muchos problemas.Un ejemplo de esto es el boxplot de precio donde notamos que la mayor distribución se concentra en un aproximado a los $10.000 y despues de $15.000 pueden ser un conjunto de posibles outliers. Ahora vamos a elegir a nuestro criterio un conjunto de variables para estudiarlas más a fondo.

Dataset elegidos.

```
dataset = datos[c("Price", "KM", "Age_08_04", "HP", "cc", "Doors", "Gears", "Weight",
                  "Fuel_Type", "Central_Lock", "Powered_Windows","Automatic_airco")]

#dataset1 = deleteme[c("Price", "KM", "Age_08_04", "HP", "cc", "Doors", "Gears", "Weight", "Fuel_Type",
```

Una vez conformado el dataset con las variables que elegimos a nuestro criterio, procedemos a realizar la regresión lineal.

```
mlr <- lm(formula = Price ~ ., data = dataset)
summary(mlr)
```

```
##
## Call:
## lm(formula = Price ~ ., data = dataset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8073.9  -689.8   -12.9   731.4  5660.9
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)     -3.098e+03  1.457e+03  -2.125  0.03373 *
## KM              -1.750e-02  1.222e-03 -14.319  < 2e-16 ***
## Age_08_04       -1.136e+02  2.470e+00 -45.970  < 2e-16 ***
## HP               1.869e+01  3.255e+00   5.741 1.15e-08 ***
## cc              -1.410e-01  8.400e-02  -1.679  0.09334 .
## Doors            2.988e+01  3.758e+01   0.795  0.42681
## Gears            4.060e+02  1.813e+02   2.240  0.02525 *
## Weight           1.530e+01  1.144e+00  13.371  < 2e-16 ***
## Fuel_TypeDiesel  6.237e+02  3.483e+02   1.791  0.07353 .
## Fuel_TypePetrol  7.761e+02  3.108e+02   2.497  0.01262 *
## Central_Lock     2.579e+01  1.368e+02   0.188  0.85054
## Powered_Windows  3.928e+02  1.366e+02   2.876  0.00409 **
## Automatic_airco  2.637e+03  1.684e+02  15.665  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1224 on 1423 degrees of freedom
## Multiple R-squared:  0.887,  Adjusted R-squared:  0.8861
## F-statistic: 930.9 on 12 and 1423 DF,  p-value: < 2.2e-16
```

En este caso en los residuales hay una variación entre los extremos lo que denota que no es simétrico entre el 1Q y 3Q los valores se acercan por lo tanto esta dentro de todo bien. Al mirar las variables vemos que hay muchas que presentan t value cercanos a ceros lo que deriva en un pr alto quitandole significancia a dichas variables para nuestro modelo.para la siguientes regresiones buscaremos excluir las variables que no sean significantes para nuestro modelo.

Nueva selección de variables

```
dataset1 <- dataset[c("Price", "KM", "Age_08_04", "HP", "cc", "Doors", "Gears", "Weight",
                      "Powered_Windows","Automatic_airco")]
```

```
mlr2 <- lm(formula = Price ~ ., data =  dataset1)

summary(mlr2)
```

```
##
## Call:
## lm(formula = Price ~ ., data = dataset1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7711.1  -689.7   -16.8   740.4  5716.8
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -1.908e+03  1.239e+03  -1.541   0.1237
## KM              -1.834e-02  1.112e-03 -16.485  < 2e-16 ***
## Age_08_04       -1.129e+02  2.453e+00 -46.038  < 2e-16 ***
## HP               1.932e+01  2.463e+00   7.847 8.28e-15 ***
## cc              -1.477e-01  8.175e-02  -1.807   0.0710 .
## Doors            3.868e+01  3.667e+01   1.055   0.2917
## Gears            4.381e+02  1.806e+02   2.425   0.0154 *
## Weight           1.468e+01  8.258e-01  17.772  < 2e-16 ***
## Powered_Windows  4.247e+02  7.065e+01   6.011 2.34e-09 ***
## Automatic_airco  2.668e+03  1.673e+02  15.950  < 2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1226 on 1426 degrees of freedom
## Multiple R-squared:  0.8865, Adjusted R-squared:  0.8858
## F-statistic:  1238 on 9 and 1426 DF,  p-value: < 2.2e-16
```

En esta nueva regresión podemos notar que la asimetría de los residuales disminuyó de forma leve en comparación con la anterior regresión. el modelo se ajusta a la primera regresion ya que al sacar variables insignificantes. pero notamos que siguen estando variables que para nuestro modelo no tiene relevancia. para un próximo análisis iremos excluyendo dichas variables.

Nueva selección de variables para nuestro dataset.

```
dataset3 <- dataset1[c("Price", "KM", "Age_08_04", "HP", "cc", "Gears", "Weight",
                       "Powered_Windows","Automatic_airco")]
```

```
mlr4 <- lm(formula = Price ~ ., data =  dataset3)

summary(mlr4)
```

```
##
## Call:
## lm(formula = Price ~ ., data = dataset3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7808.0  -697.2    -9.1   722.2  5668.3
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -1.846e+03  1.237e+03  -1.492   0.1360
## KM              -1.834e-02  1.112e-03 -16.485  < 2e-16 ***
## Age_08_04       -1.130e+02  2.453e+00 -46.045  < 2e-16 ***
## HP               1.961e+01  2.448e+00   8.010 2.36e-15 ***
## cc              -1.494e-01  8.174e-02  -1.828   0.0677 .
## Gears            4.011e+02  1.772e+02   2.264   0.0237 *
## Weight           1.491e+01  7.950e-01  18.758  < 2e-16 ***
## Powered_Windows  4.285e+02  7.056e+01   6.073 1.61e-09 ***
## Automatic_airco  2.650e+03  1.664e+02  15.926  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1226 on 1427 degrees of freedom
## Multiple R-squared:  0.8864, Adjusted R-squared:  0.8858
## F-statistic:  1392 on 8 and 1427 DF,  p-value: < 2.2e-16
```

En cuanto a los valores residuales 1Q y 3Q a pesar no estar simétrico mantiene un buen balance, la mediana se acerca a cero, pero en los extremos siguen dispersos lo que lleva a tener residuales que no son simétricos. En general los valores de la mayoria de las variables tiene un buen t value y pr salvo algunas variables que tendremos que tener en cuenta para su próxima depuración como por ejemplo Gears y cc, posterior análisis deberemos tomar una decisión de ver si nos quedamos con la misma o la eliminamos del dataset.

nuevo dataset

```
dataset4 <- dataset3[c("Price", "KM", "Age_08_04", "HP","Gears", "Weight",
                       "Powered_Windows","Automatic_airco")]
```

```r
mlr5 <- lm(formula = Price ~ ., data = dataset4)

summary(mlr5)
```

```
##
## Call:
## lm(formula = Price ~ ., data = dataset4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7705.0  -701.0   -11.5   724.5  5761.2
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -1.642e+03  1.233e+03  -1.331   0.1832
## KM              -1.856e-02  1.106e-03 -16.776  < 2e-16 ***
## Age_08_04       -1.130e+02  2.455e+00 -46.027  < 2e-16 ***
## HP               1.942e+01  2.448e+00   7.935 4.23e-15 ***
## Gears            4.032e+02  1.773e+02   2.273   0.0231 *
## Weight           1.453e+01  7.672e-01  18.935  < 2e-16 ***
## Powered_Windows  4.299e+02  7.061e+01   6.088 1.47e-09 ***
## Automatic_airco  2.634e+03  1.663e+02  15.840  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1227 on 1428 degrees of freedom
## Multiple R-squared:  0.8861, Adjusted R-squared:  0.8856
## F-statistic:  1588 on 7 and 1428 DF,  p-value: < 2.2e-16
```

En esta última regresión podemos observar en los residuales que estan dando unos valores bastantes simétricos pero tienden a dispersarse en los extremos lo cual el problema de la simetria continua. en cuanto a los 1Q y 3Q estan bastante bien y la mediana esta cerca a cero. Las variables tienen un buen t value y pr value notamos que gears entro pero habra que realizarle un nuevo análisis sobre esta variable para ver si continuamos con la misma.

Validación del modelo

Análisis sobre los residuales.

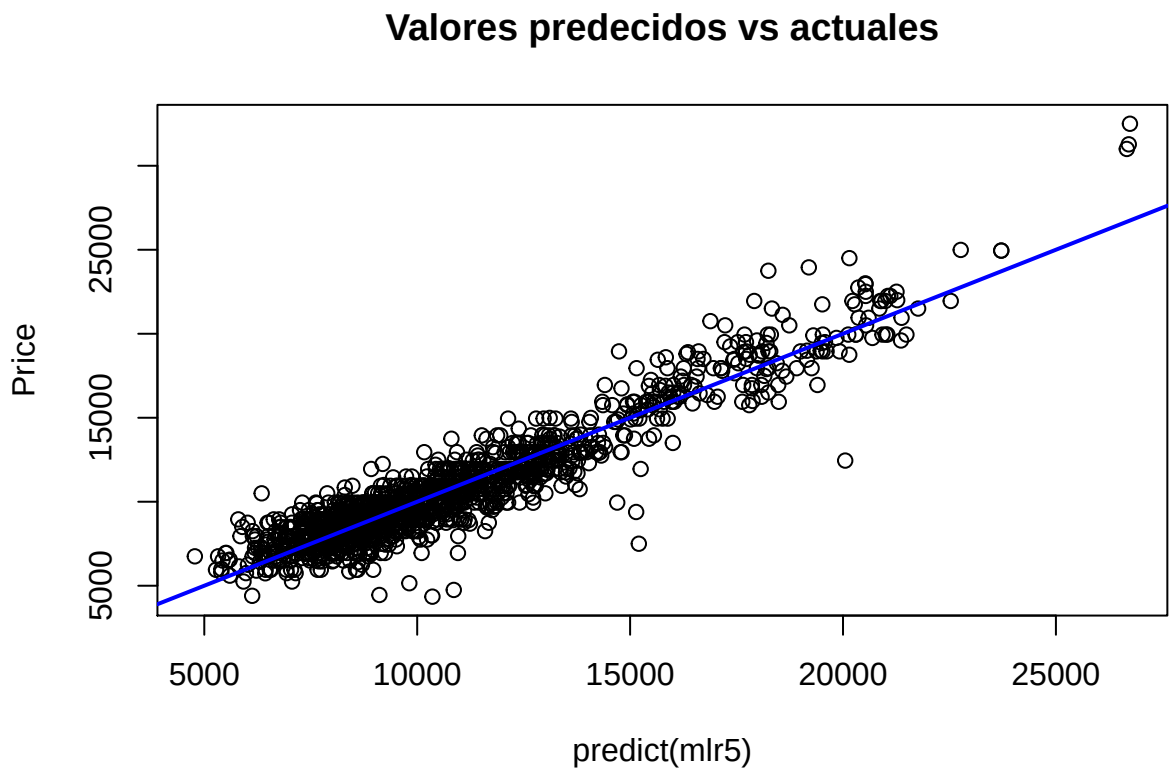
```r
stem(mlr5$residuals)
```

```
##
##   The decimal point is 3 digit(s) to the right of the |
##
##   -7 | 76
##   -6 | 10
##   -5 | 8
##   -4 | 7770
##   -3 | 332100
##   -2 | 98776655555555443333332222221111111000000
##   -1 | 99999999999888888888888888888877777777777777776666666666665555555+119
##   -0 | 99999999999999999999999999999999999999998888888888888888888888+388
##    0 | 00000000000000000000000011111111111111111111111111111111111111111+355
##    1 | 00000000000000000000000000000000000000011111111111111111111111111+128
```

```
##     2 | 00000000000001111112222223333333344455555556667888889
##     3 | 012239
##     4 | 0223468
##     5 | 58
```

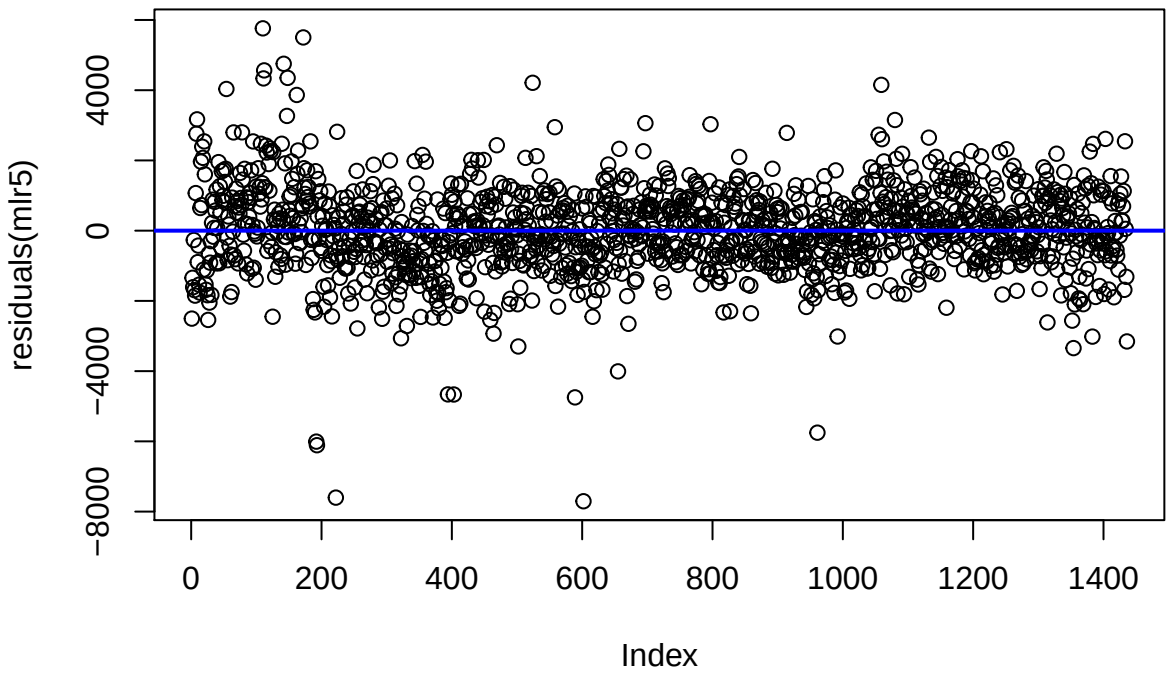
Acá notamos que al aplicar stem sobre los residuales de la regresión mlr5,confirmamos que no son simétricos en los extremos.

```
plot(predict(mlr5), datos$Price, ylab = "Price" , main = "Valores predecidos vs actuales")
abline(a=0,b=1, col="blue", lwd=2)
```



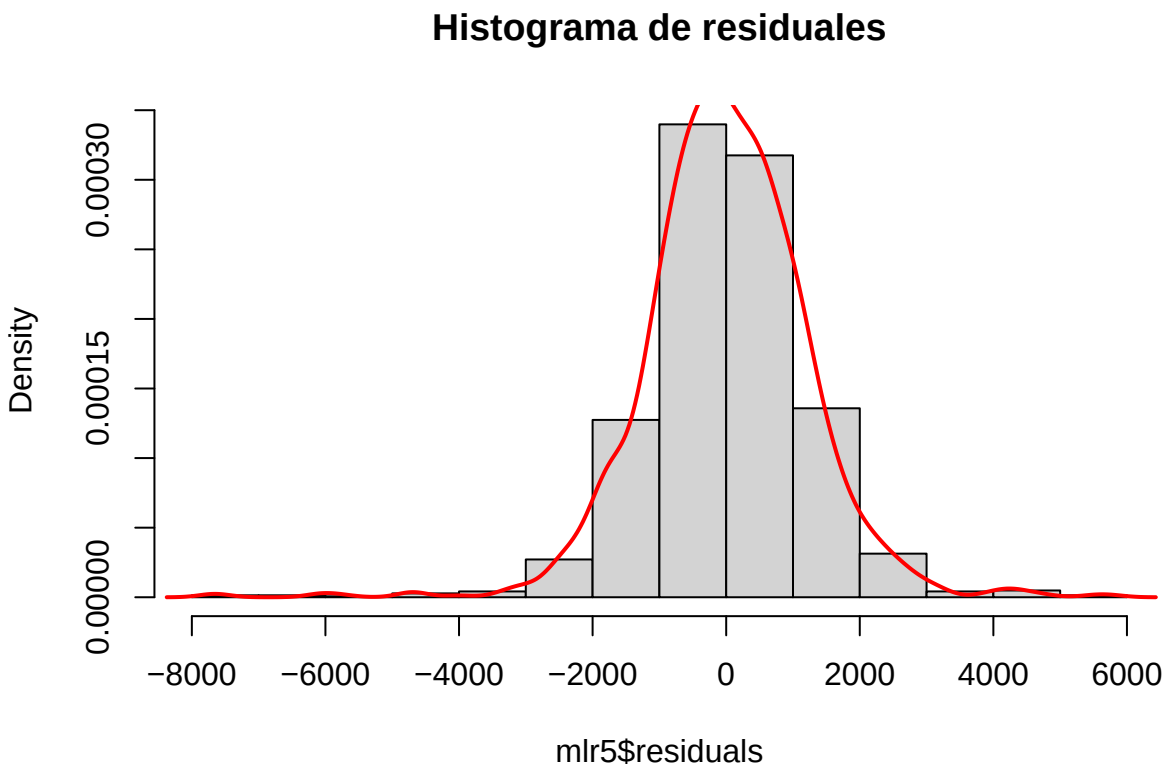
Con esta gráfica notamos que se concentran las observaciones entre 5000 y 15000 produciendo un área de mayor densidad comprendido esto podemos decir también despues de esos valores hay 2 grupo de datos que tendremos que analizar a posterior

```
plot(residuals(mlr5))
abline(a=0,b=0, col="blue", lwd =2)
```



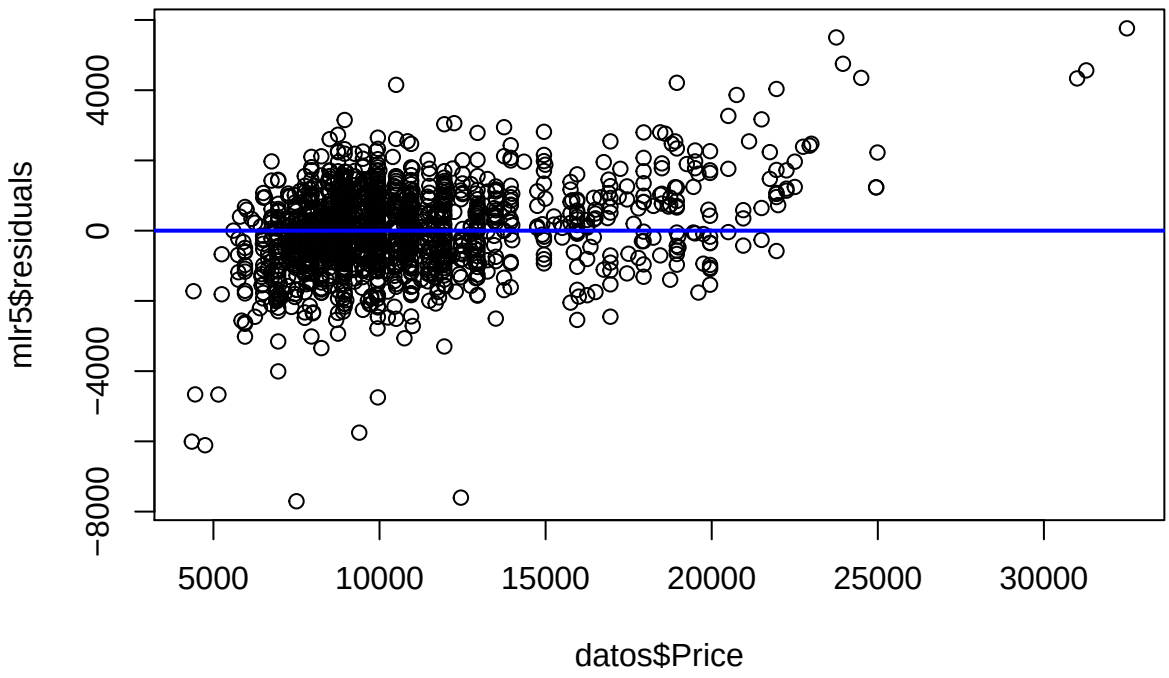
La gráfica aquí en este caso se ve con bastantes problemas entre 0 a 200 los datos tienden a estar por encima de la recta pasa lo mismo en el siguiente rango por lo tanto decimos que no tiene una distribución aleatoria.

```
hist(mlr5$residuals , main = "Histograma de residuales", freq = F)
lines(density(mlr5$residuals), col="red", lwd=2)
```



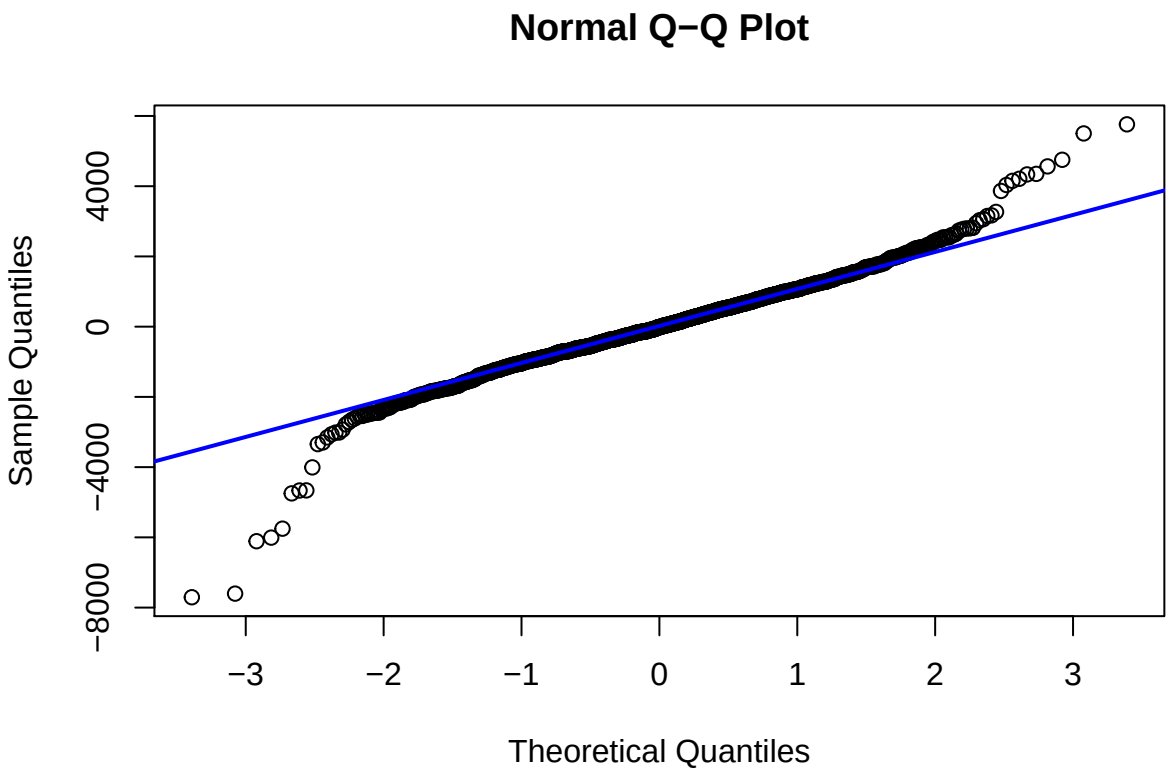
En el histograma con una tendencia hacia la derecha lo que seguimos confirmando que los residuales no son simétricos.

```
plot(mlr5$residuals ~ datos$Price)
abline(a=0,b=0, col = "blue", lwd=2)
```



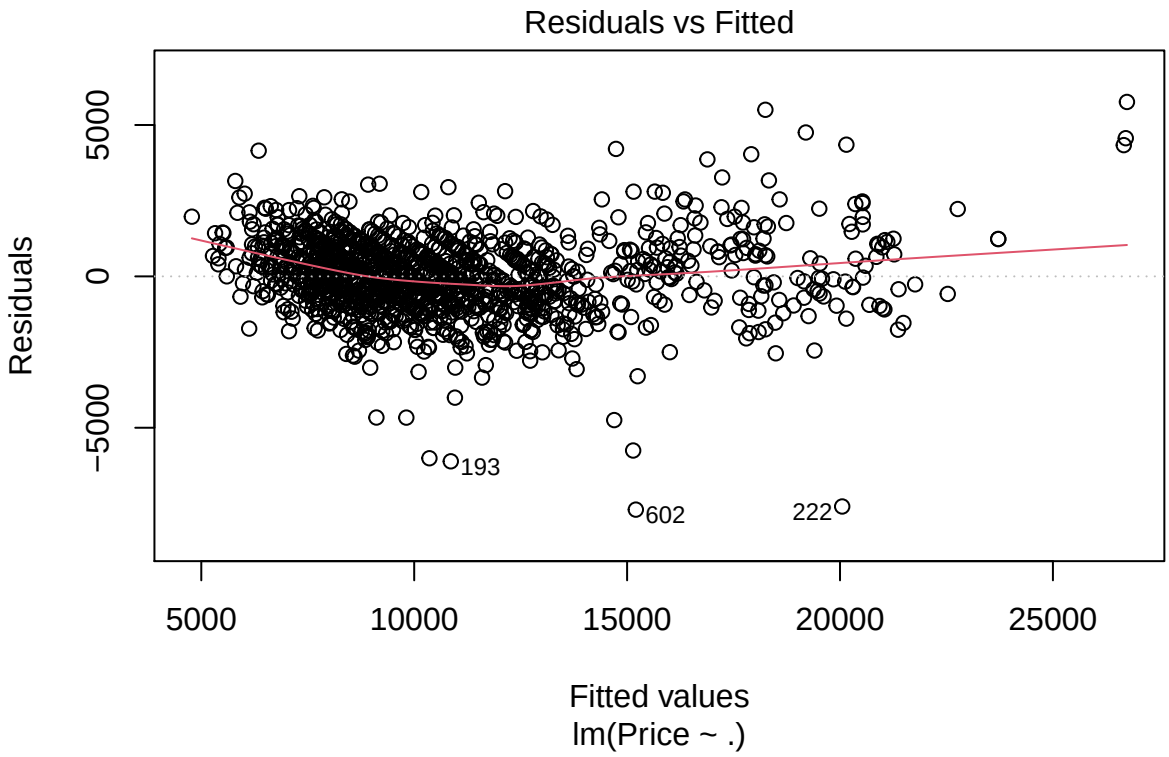
Se puede observar 3 grupos definidos lo que pùeden llegar a ser un conjunto de posibles outliers. desde el 5000 a 15000 es el grupo con mayor densidad, y consideramos que despues de 15000 se podria decir que estamos en presencia de un posible conjunto de outliers.

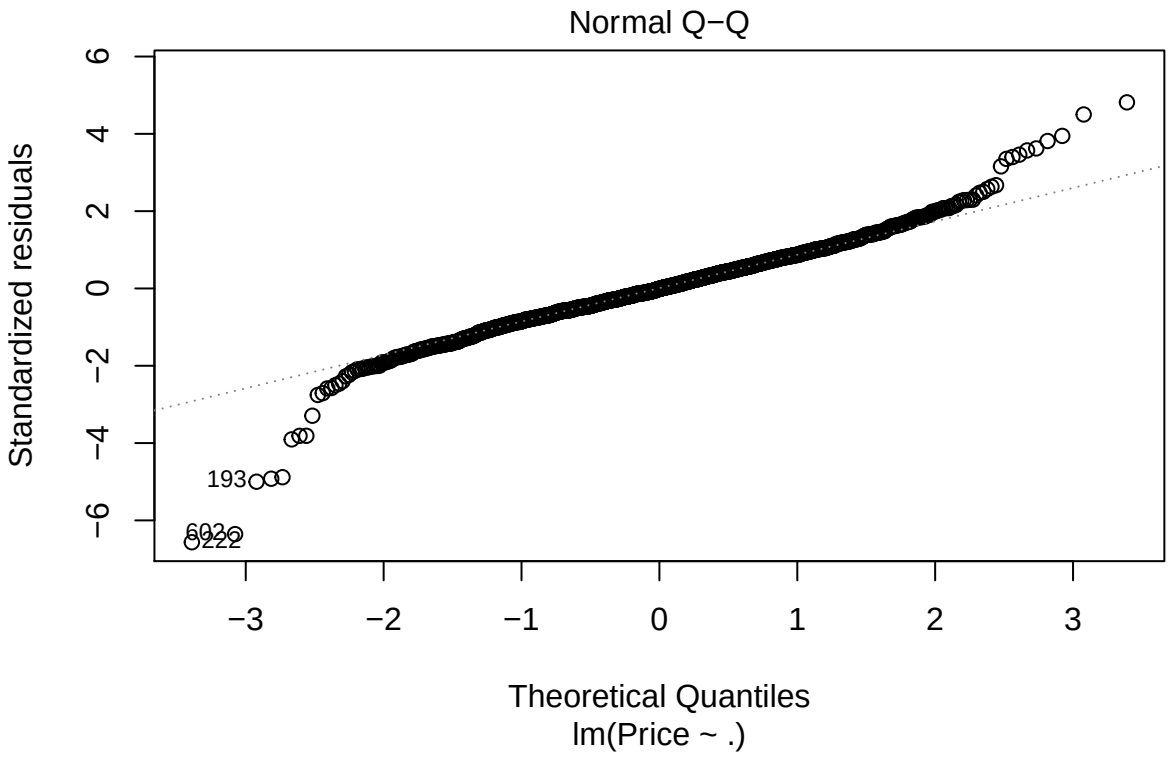
```
qqnorm(mlr5$residuals)

qqline(mlr$residuals, col = "blue ", lwd=2)
```

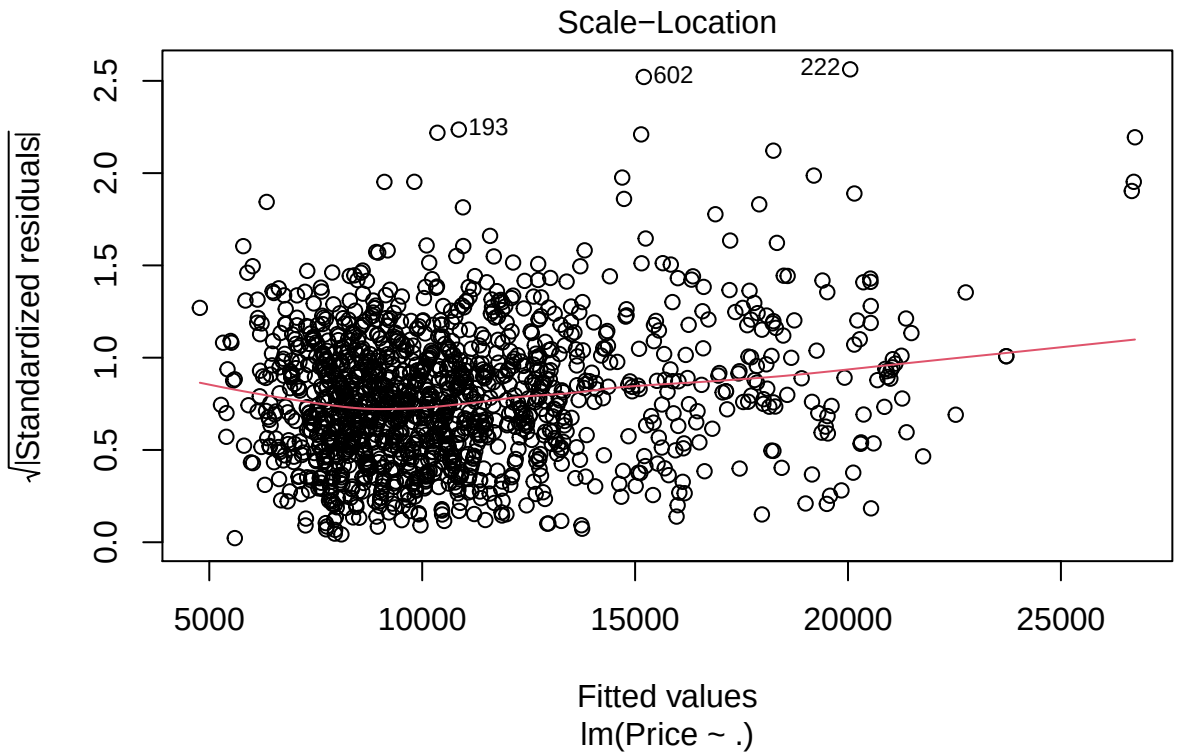


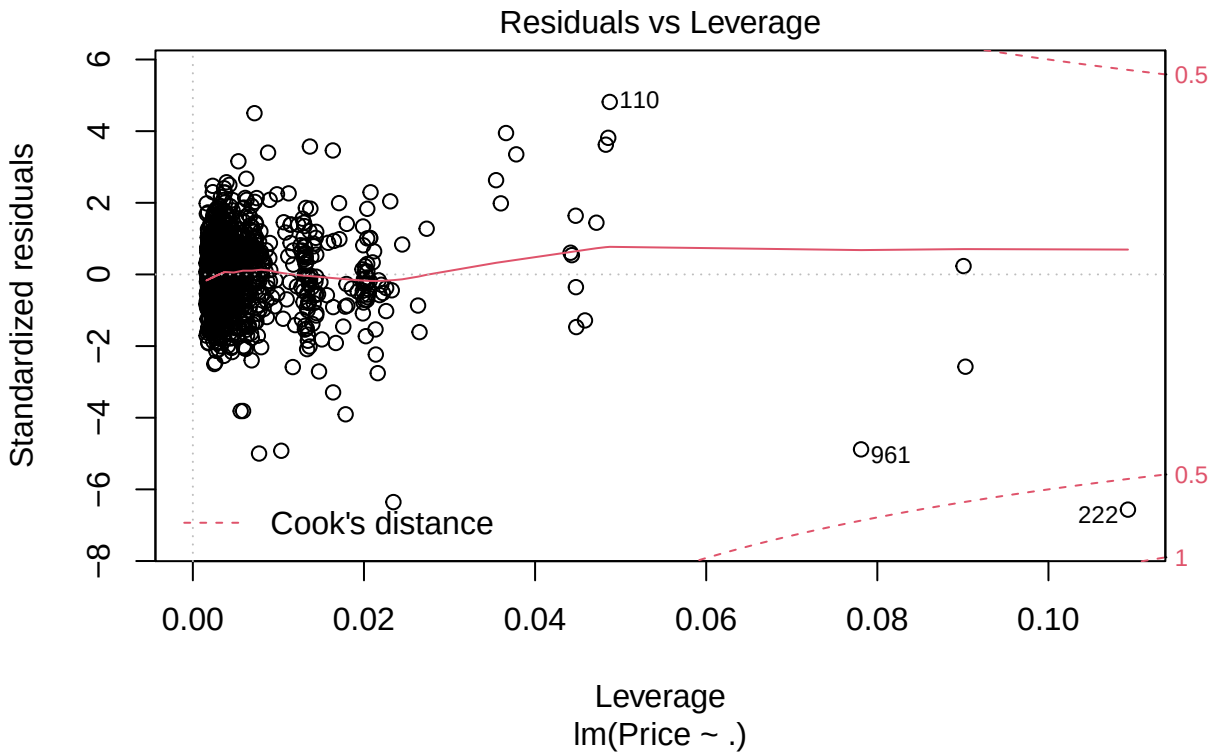
En esta gráfica se observa alteraciones respecto al patron dominante (puntos sobre la recta) por fuera del intervalo de -2 y 2, deberiamos analizar mas a fondo estos puntos ya que pueden ser posibles outliers.

```
plot(mlr5)
```


Residuals vs Fitted
Residuals
5000
0
−5000
193
602
222
5000
10000
15000
20000
25000
Fitted values
lm(Price ~ .)


Normal Q−Q
Standardized residuals
6
4
2
0
−2
−4
−6
193
602
222
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(Price ~ .)

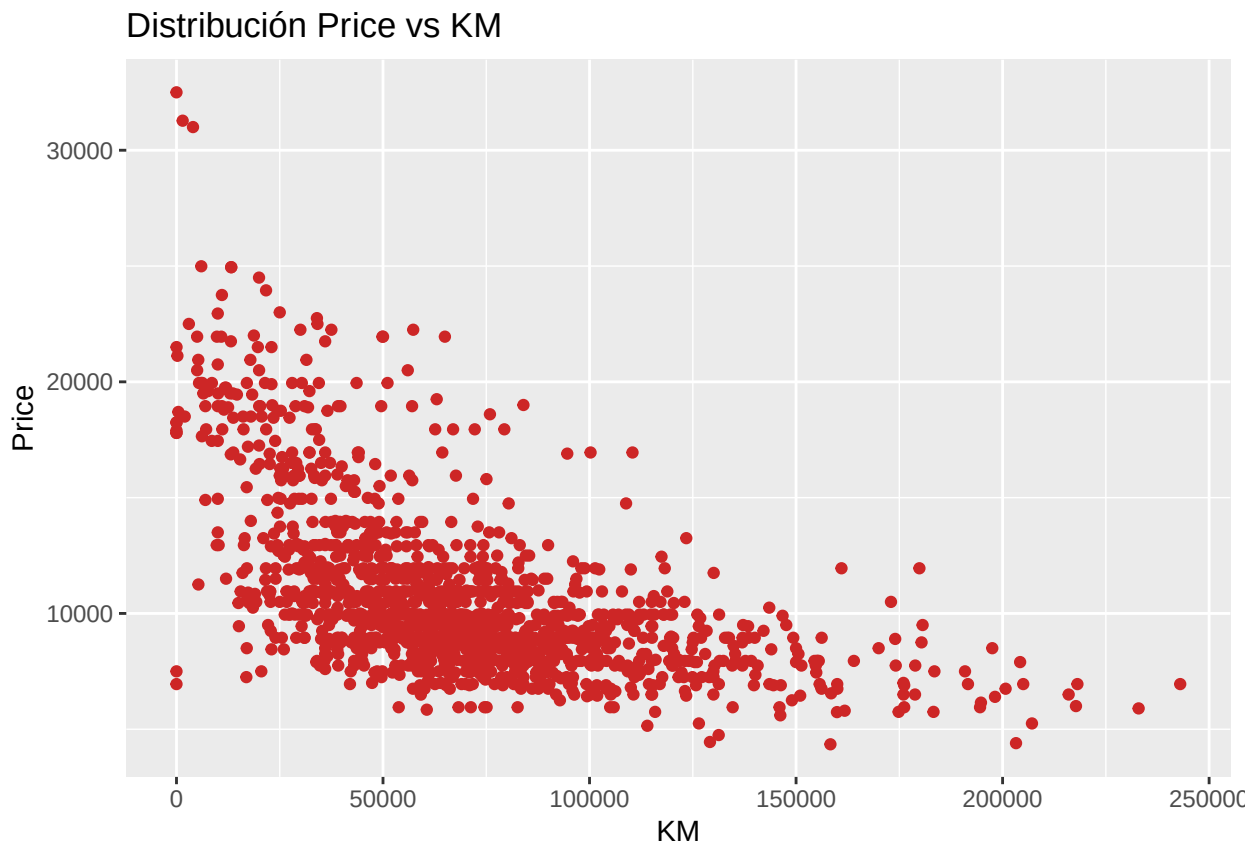

Scale−Location
602
222
193
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
5000
10000
15000
20000
25000
Fitted values
lm(Price ~ .)



Aplicando plot a la regresion mlr5, podemos representar todas las gráficas que veniamos ejecuando pero con los puntos (observaciones)donde se encontraria los posibles outliers. en esta última gráfica se ven los puntos muy dispersos y lo que nos lleva a confirmar que estamos en presencia de outliers los cuales tendran que ser tratados en posterioridad.

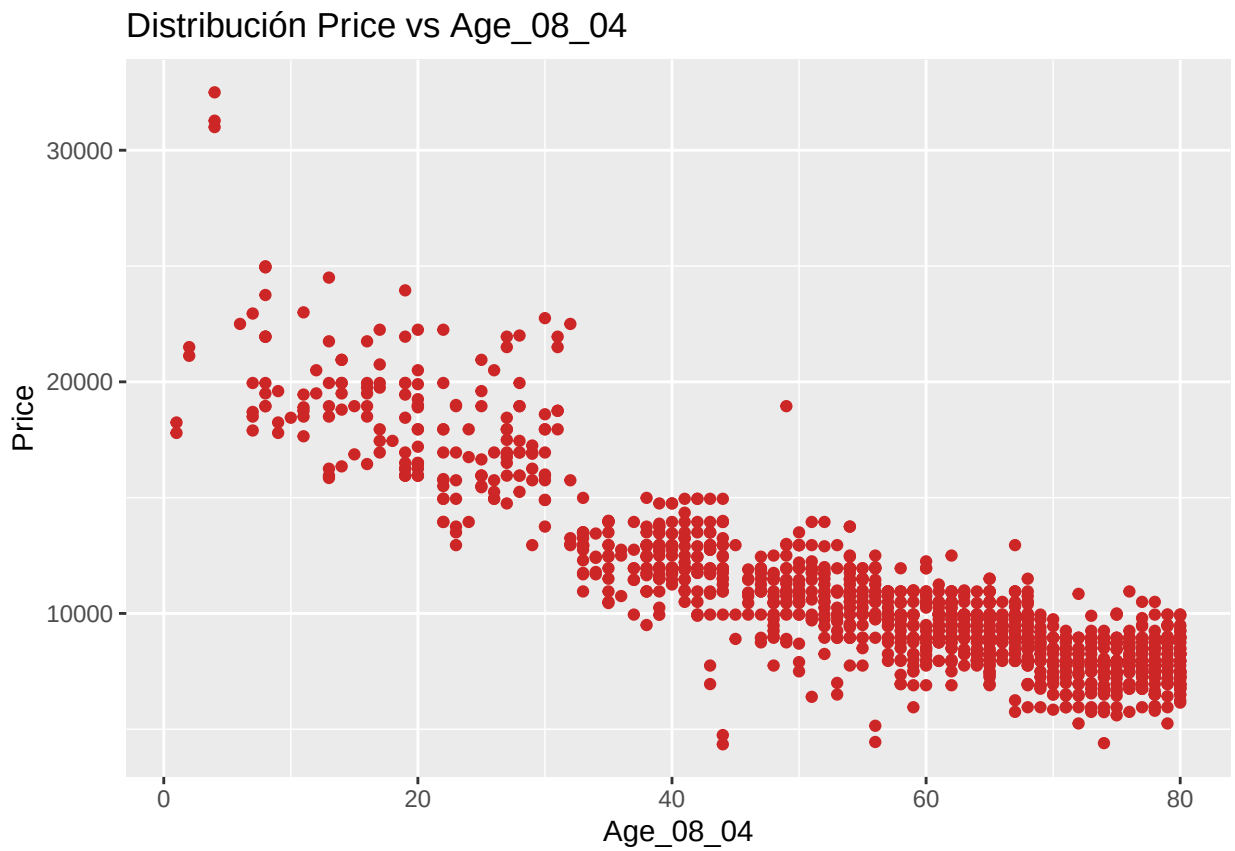
Distribución de las distintas variables frente al precio

```
ggplot(dataset4, aes(x=KM, y=Price)) + geom_point(colour = "firebrick3") +
        ggtitle("Distribución Price vs KM")
```



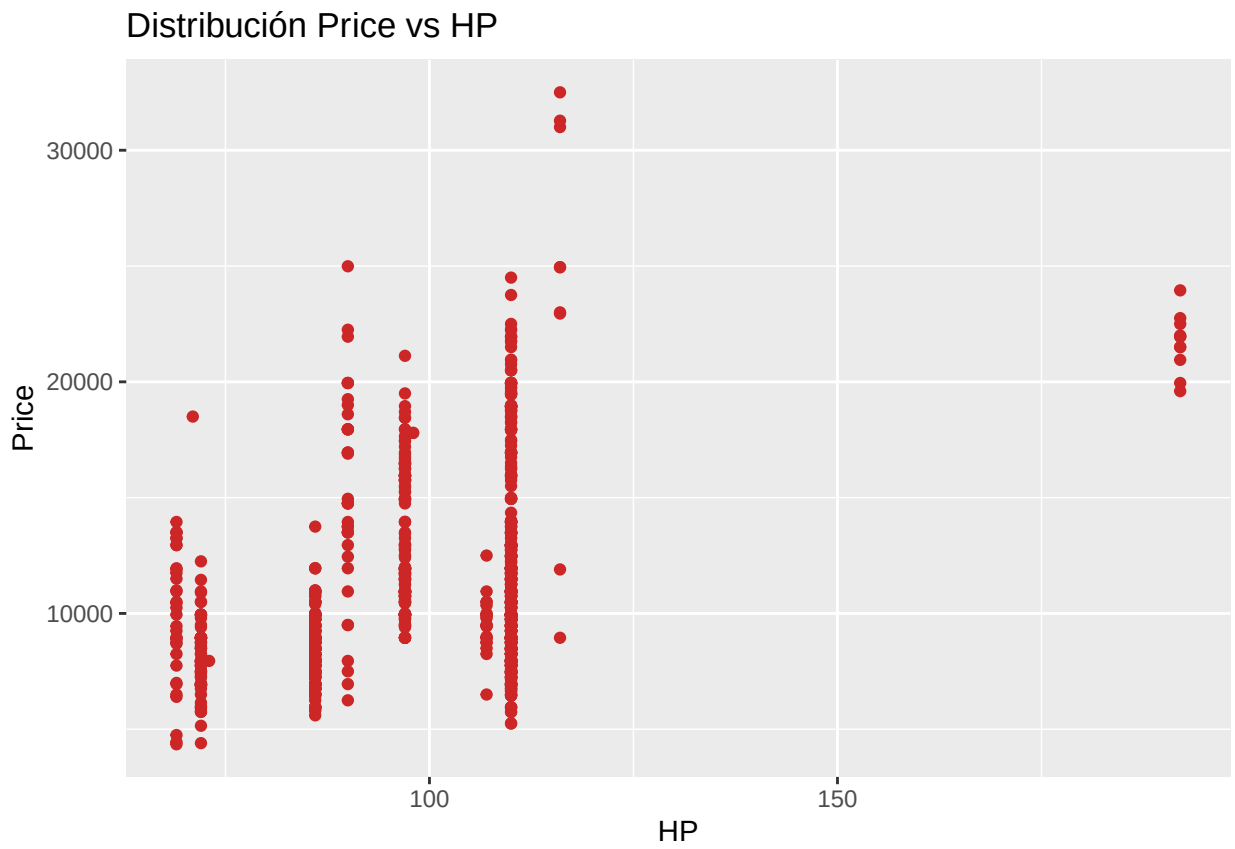
Con esta gráfica observamos la distribución de los KM frente al precio, y a nuestro criterio observamos que los datos mayores a 150.000km los exlcuiremos del modelo porque consideramos que son autos demasiados viejos, al igual que los datos por debajo de los 15km al cuales consideramos autos practicamente nuevos.

```
ggplot(dataset4, aes(x=Age_08_04, y=Price)) + geom_point(colour = "firebrick3") +
          ggtitle("Distribución Price vs Age_08_04")
```



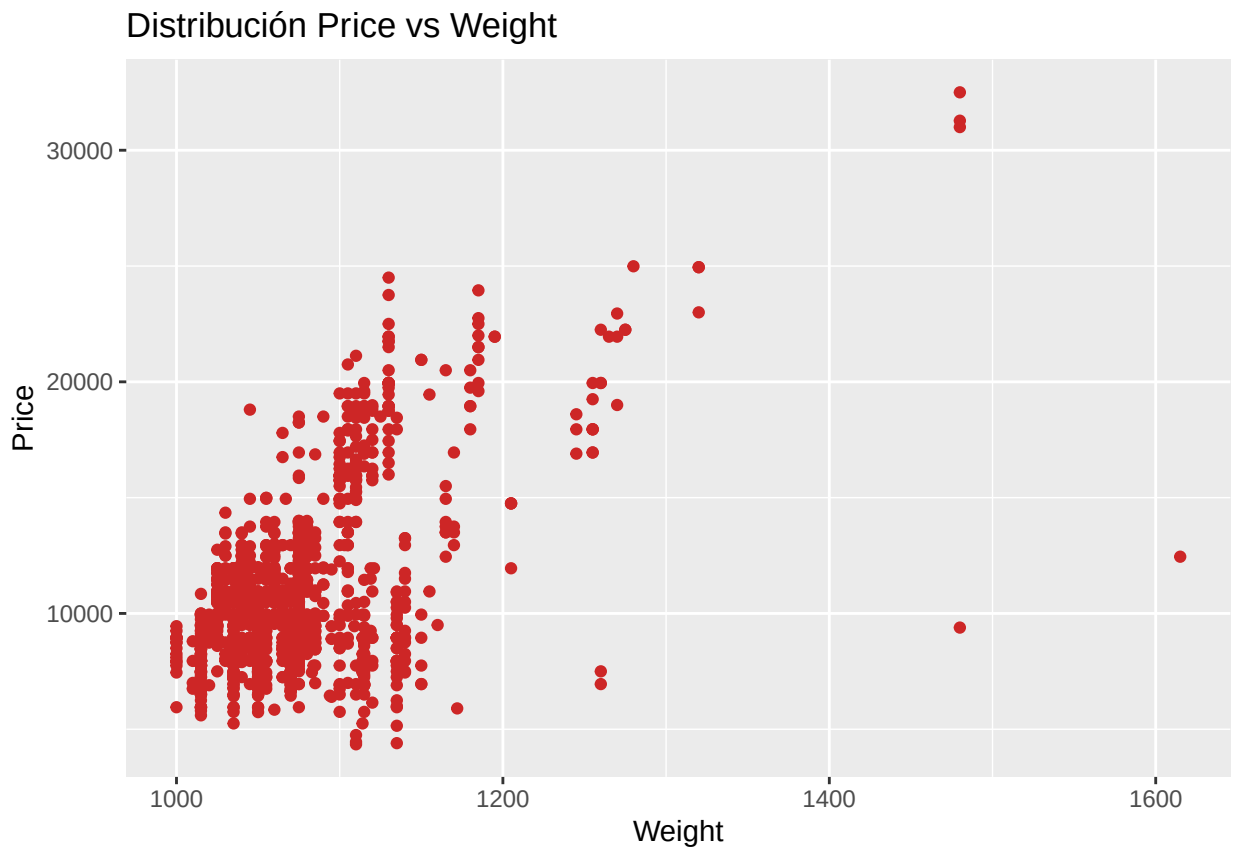
A partir de esta gráfica aplicando nuestro criterio optamos por excluir del modelos a los autos menores a 20 por ser considerados autos demasiados nuevos.

```
ggplot(dataset4, aes(x=HP, y=Price)) + geom_point(colour = "firebrick3") +
        ggtitle("Distribución Price vs HP")
```



En este caso notamos que en los mayores a 150 esta muy separado del resto, lo que a nuestro criterio decimos que son outliers y debemos excluirlos del modelo.

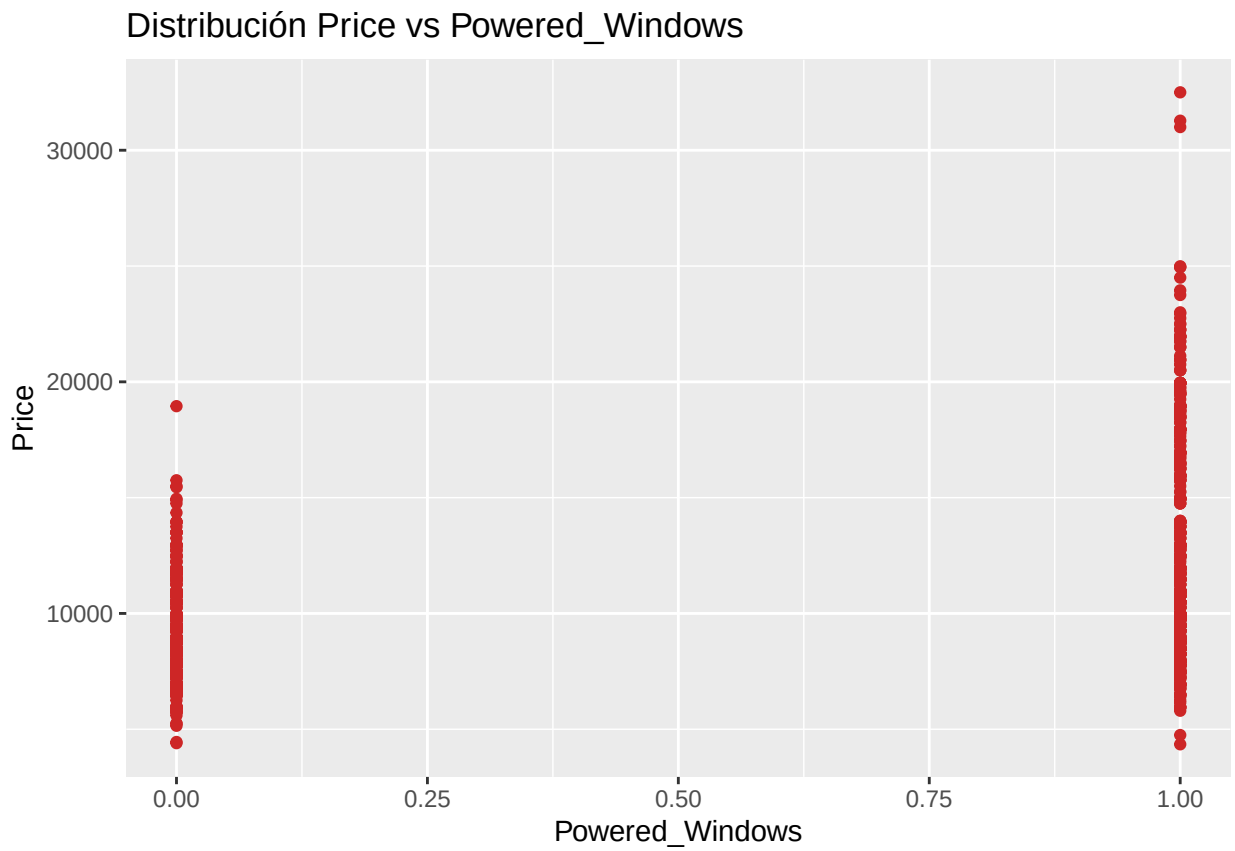
```
ggplot(dataset4, aes(x=Weight, y=Price)) + geom_point(colour = "firebrick3") +
        ggtitle("Distribución Price vs Weight")
```



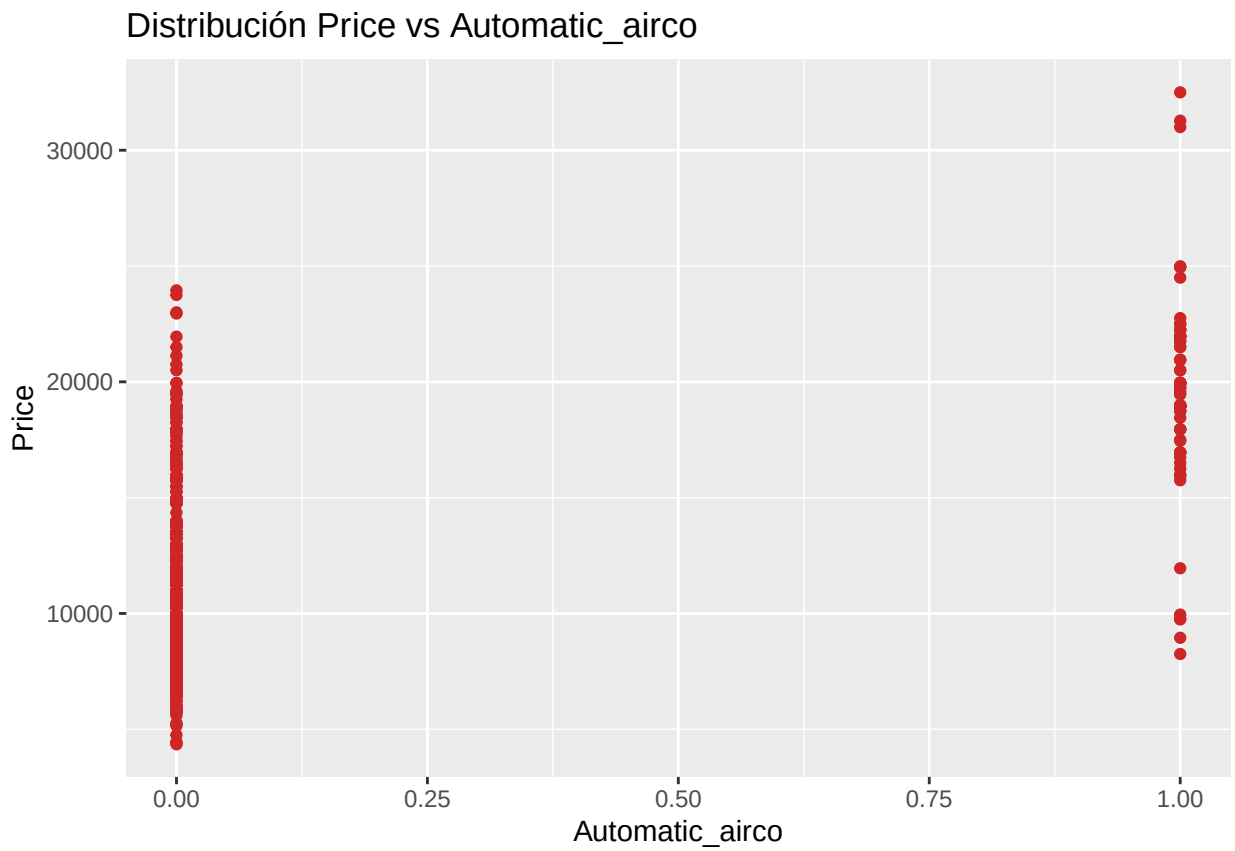
De esta gráfica rescatamos que todos aquellos autos cuyo peso supere los 1200kg debera ser excluido ya que consideramos que son vehiculos que son caros de mantener en cuanto a consumo de combustible.

Las siguientes gráficas son de variables binarias.

```
ggplot(dataset4, aes(x=Powered_Windows, y=Price)) + geom_point(colour = "firebrick3") +
        ggtitle("Distribución Price vs Powered_Windows")
```




```
ggplot(dataset4, aes(x=Automatic_airco, y=Price)) + geom_point(colour = "firebrick3") +
        ggtitle("Distribución Price vs Automatic_airco")
```




```
#dataset2 <- dataset[c("Price", "KM", "Age_08_04", "HP","cc","Gears", "Weight","Powered_Windows","Autom
```

```
#mlr3 <- lm(formula = Price ~ ., data =  dataset2)

#summary(mlr3)
```

```
#plot(mlr3)
```

Aplicación de un modelo de análisis de datos: Regresión lineal múltiple

En base al análisis de los ggplot optamos por limpiar algunas variables para nuestro próximo análisis.

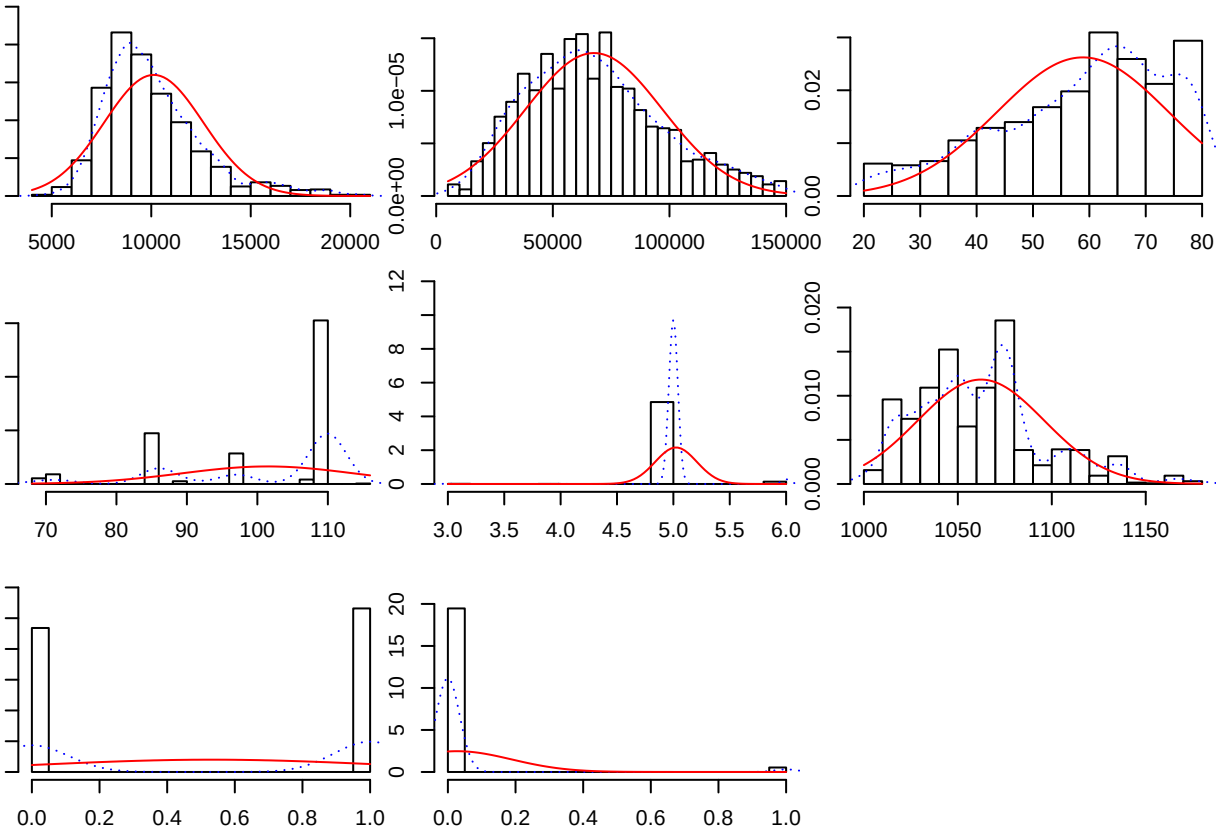
```
dataset4 <- filter(dataset4, !(Weight>1200))
dataset4 <- filter(dataset4,  !(KM>150000))
dataset4 <- filter(dataset4, !(KM<15))
dataset4 <- filter(dataset4, !(HP>150))
dataset4 <- filter(dataset4, !(Age_08_04<20))
```

Visualización del sesgo en la distriibución de las variables elegidas

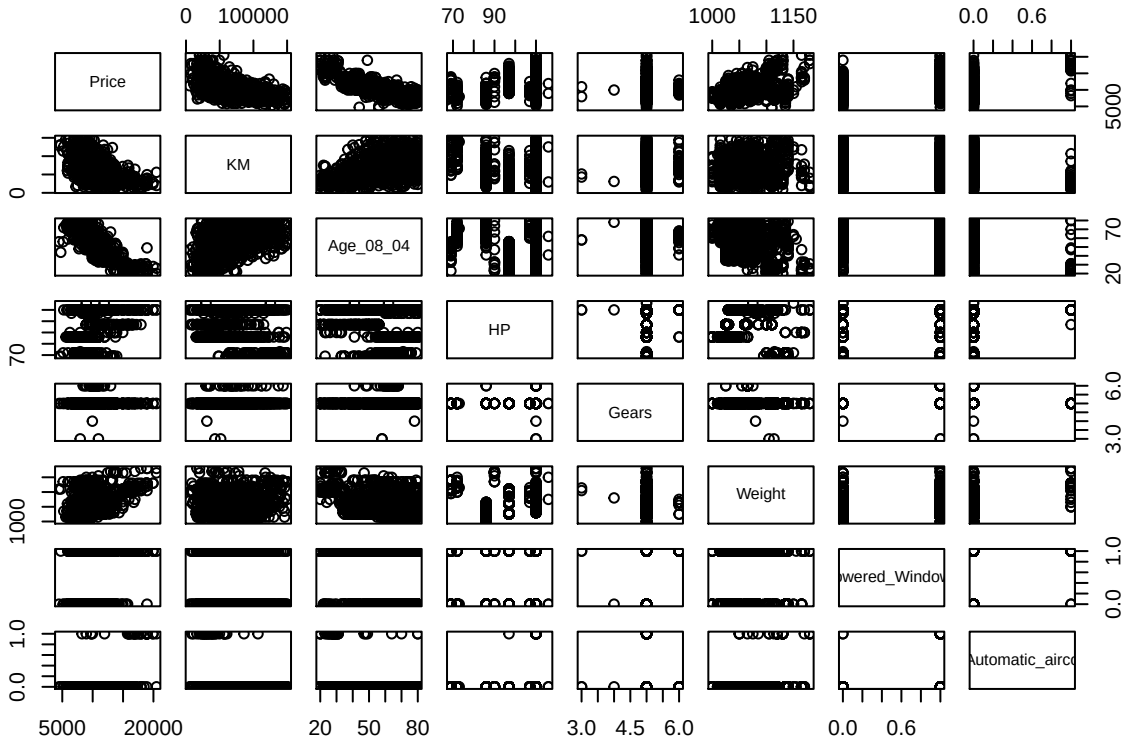
```
skewness(dataset4)
```

```
##           Price              KM        Age_08_04              HP           Gears
##       1.1830649       0.4895120      -0.6244366      -1.1070427       1.8690929
##          Weight Powered_Windows Automatic_airco
##       0.6814009      -0.1275168       5.8709980
```

```
multi.hist(dataset4, dcol = c("blue ", "red"), dlty = c("dotted", "solid"), main = "")
```




```
pairs(dataset4)
```

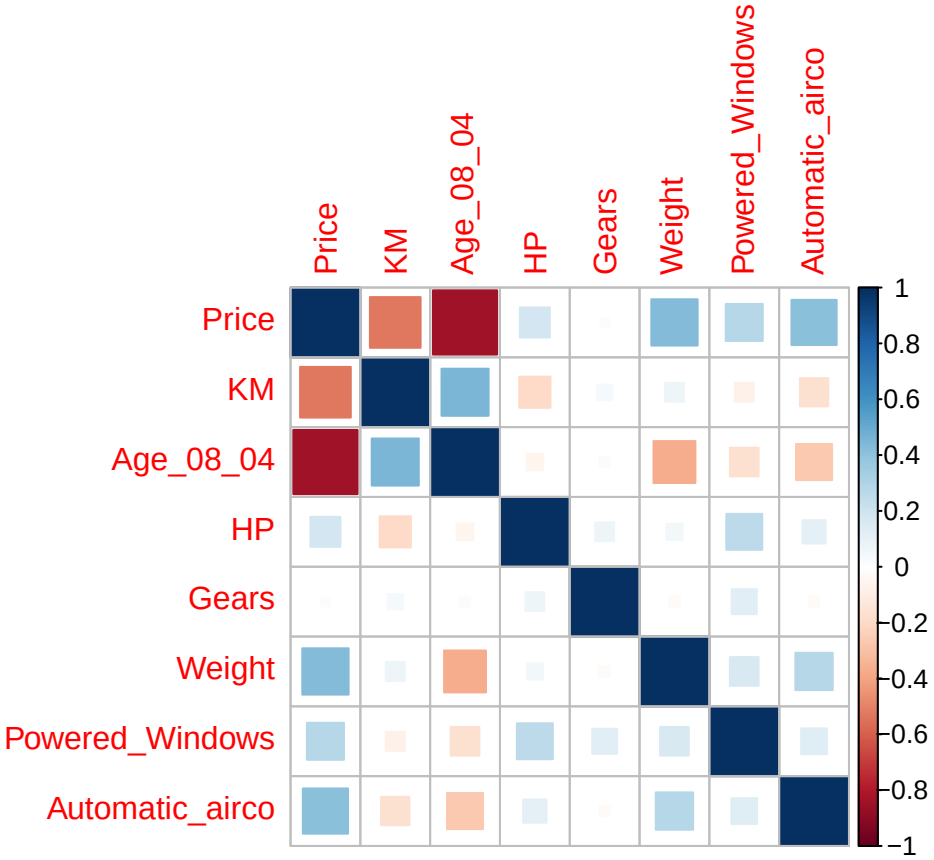


Correlación de las variables.

```
data_correlation <- cor(dataset4)
data_correlation
```

```
##                       Price          KM   Age_08_04          HP       Gears
## Price            1.00000000 -0.52727289 -0.84722824  0.18485980  0.01574659
## KM              -0.52727289  1.00000000  0.45493054 -0.19944808  0.04788529
## Age_08_04       -0.84722824  0.45493054  1.00000000 -0.05656074  0.02118414
## HP               0.18485980 -0.19944808 -0.05656074  1.00000000  0.07276841
## Gears            0.01574659  0.04788529  0.02118414  0.07276841  1.00000000
## Weight           0.43968917  0.07415679 -0.36021243  0.05270309 -0.02439631
## Powered_Windows  0.28177926 -0.07443810 -0.16956527  0.26502879  0.12843881
## Automatic_airco  0.41932986 -0.16460740 -0.26716009  0.11087911 -0.02116908
##                      Weight Powered_Windows Automatic_airco
## Price            0.43968917       0.2817793      0.41932986
## KM               0.07415679      -0.0744381     -0.16460740
## Age_08_04       -0.36021243      -0.1695653     -0.26716009
## HP               0.05270309       0.2650288      0.11087911
## Gears           -0.02439631       0.1284388     -0.02116908
## Weight           1.00000000       0.1687236      0.28627633
## Powered_Windows  0.16872359       1.0000000      0.13590083
## Automatic_airco  0.28627633       0.1359008      1.00000000
```

```r
corrplot(data_correlation, method="square")
```



Las variables con mayor correlación con respecto al precio son KM, Age,hp, weight, powered_windows, automatic_airco.y notamos que gears no tiene corralación frente al precio lo que decimos sacarla de nuestro modelo ya que no nos aporta nada valor.

Nuevo dataset sin Gears.

```r
dataset5 <- dataset4[c("Price", "KM", "Age_08_04", "Weight","HP",
                       "Powered_Windows", "Automatic_airco")]

mlr6 <- lm(formula = Price ~ ., data = dataset5)

summary(mlr6)
```

```
##
## Call:
## lm(formula = Price ~ ., data = dataset5)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6158.4  -618.6   -24.9   680.7  4869.6
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      2.971e+03  1.141e+03   2.604  0.00932 **
## KM              -1.769e-02  1.225e-03 -14.437  < 2e-16 ***
## Age_08_04       -1.039e+02  2.489e+00 -41.763  < 2e-16 ***
```

```
## Weight           1.221e+01  1.025e+00  11.913  < 2e-16 ***
## HP               1.165e+01  2.599e+00   4.481 8.11e-06 ***
## Powered_Windows  4.778e+02  6.324e+01   7.556 7.93e-14 ***
## Automatic_airco  2.308e+03  1.983e+02  11.634  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1063 on 1266 degrees of freedom
## Multiple R-squared:  0.8193, Adjusted R-squared:  0.8184
## F-statistic: 956.7 on 6 and 1266 DF,  p-value: < 2.2e-16
```

Los residuales siguen sin simetría en cuento a los extremos pero con mejores valores achicando mas la brecha, los 1Q y 3Q los valores bastantes simétricos y la mediana tiende a cero. Las variables presenta un buen t value y pr dentro de lo que se estima. Entre el r-squared y su adjustado estamos en presencia de un buen modelo.

Para seleccionar la mejor combinación dentro de la regresión utilizamos step.

```
step(mlr6, direction = "both", trace = 1)
```

```
## Start:  AIC=17749.5
## Price ~ KM + Age_08_04 + Weight + HP + Powered_Windows + Automatic_airco
##
##                   Df  Sum of Sq        RSS   AIC
## <none>                           1430344713 17750
## - HP               1   22684467 1453029180 17768
## - Powered_Windows  1   64505413 1494850126 17804
## - Automatic_airco  1  152929450 1583274162 17877
## - Weight           1  160343951 1590688664 17883
## - KM               1  235487365 1665832078 17942
## - Age_08_04        1 1970595291 3400940004 18850
```

```
##
## Call:
## lm(formula = Price ~ KM + Age_08_04 + Weight + HP + Powered_Windows +
##     Automatic_airco, data = dataset5)
##
## Coefficients:
##     (Intercept)               KM        Age_08_04           Weight
##      2971.12844         -0.01769       -103.94489         12.20895
##              HP  Powered_Windows  Automatic_airco
##        11.64541        477.82279       2307.64388
```

Con esta sentencia podemos decir que considera a todas las varibales de nuestro dataset influyentes para el modelo.

Primera Valadación del modelo.

```
split_data <- createDataPartition(y= dataset5$Price, p=0.7, list= FALSE)

train_data <- dataset5[split_data,]
test_data <- dataset5[-split_data,]

lmfit1 <- train(Price ~ ., data = train_data, method="lm")

summary(lmfit1)
```

```
##
## Call:
## lm(formula = .outcome ~ ., data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4740.1  -670.1   -41.8   680.0  4985.1
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      2.989e+03  1.399e+03   2.137   0.0329 *
## KM              -1.682e-02  1.453e-03 -11.574  < 2e-16 ***
## Age_08_04       -1.063e+02  2.992e+00 -35.537  < 2e-16 ***
## Weight           1.207e+01  1.253e+00   9.629  < 2e-16 ***
## HP               1.356e+01  3.193e+00   4.248 2.38e-05 ***
## Powered_Windows  4.927e+02  7.683e+01   6.412 2.33e-10 ***
## Automatic_airco  2.196e+03  2.141e+02  10.257  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1063 on 886 degrees of freedom
## Multiple R-squared:  0.8275, Adjusted R-squared:  0.8264
## F-statistic: 708.5 on 6 and 886 DF,  p-value: < 2.2e-16
```

Para esta primera validación separamos el dataset en 70% en datos de entrenamiento y 30% en datos de prueba de nuestro último dataset.

```
predict_test <- predict(lmfit1, test_data)

model_test_1 <- data.frame(obs= test_data$Price, pred = predict_test)

defaultSummary(model_test_1)
```

```
##         RMSE     Rsquared          MAE
## 1066.1329432    0.7964694  806.6493423
```

Con esta primera validación podemos decir que el Rsquared de test no hay tanta diferencia entre r squared de los datos de entrenamiento lo cual indica que nuestro modelo predice bien.

Segunda Validación del modelo - Cross Validation

```
control1 <- trainControl(method="cv", number=10)

lmfit2 <- train(Price ~ ., data= dataset5, method="lm", trControl= control1, metric = "Rsquared")

summary(lmfit2)
```

```
##
## Call:
## lm(formula = .outcome ~ ., data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -6158.4  -618.6   -24.9   680.7  4869.6
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      2.971e+03  1.141e+03   2.604  0.00932 **
## KM              -1.769e-02  1.225e-03 -14.437  < 2e-16 ***
## Age_08_04       -1.039e+02  2.489e+00 -41.763  < 2e-16 ***
## Weight           1.221e+01  1.025e+00  11.913  < 2e-16 ***
## HP               1.165e+01  2.599e+00   4.481 8.11e-06 ***
## Powered_Windows  4.778e+02  6.324e+01   7.556 7.93e-14 ***
## Automatic_airco  2.308e+03  1.983e+02  11.634  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1063 on 1266 degrees of freedom
## Multiple R-squared:  0.8193, Adjusted R-squared:  0.8184
## F-statistic: 956.7 on 6 and 1266 DF,  p-value: < 2.2e-16
```

```
predict_test2 <- predict(lmfit2, dataset5)

model_test_2 <- data.frame(obs=dataset5$Price, pred = predict_test2)

defaultSummary(model_test_2)
```

```
##         RMSE     Rsquared          MAE
## 1060.0007087    0.8193014  810.5922107
```

Aplicando Cross Validation también podemos llegar a la misma conclusión: los valores de Rsquared dan los mismos resultados. El modelo predice bien.

Tercera validación del modelo - Leave One Out Cross Validation

```
control2 <- trainControl(method= "LOOCV")

lmfit3 <- train(Price ~ ., data = dataset5, method="lm", trControl=control2)

summary(lmfit3)
```

```
##
## Call:
## lm(formula = .outcome ~ ., data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6158.4  -618.6   -24.9   680.7  4869.6
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      2.971e+03  1.141e+03   2.604  0.00932 **
## KM              -1.769e-02  1.225e-03 -14.437  < 2e-16 ***
## Age_08_04       -1.039e+02  2.489e+00 -41.763  < 2e-16 ***
## Weight           1.221e+01  1.025e+00  11.913  < 2e-16 ***
## HP               1.165e+01  2.599e+00   4.481 8.11e-06 ***
## Powered_Windows  4.778e+02  6.324e+01   7.556 7.93e-14 ***
```

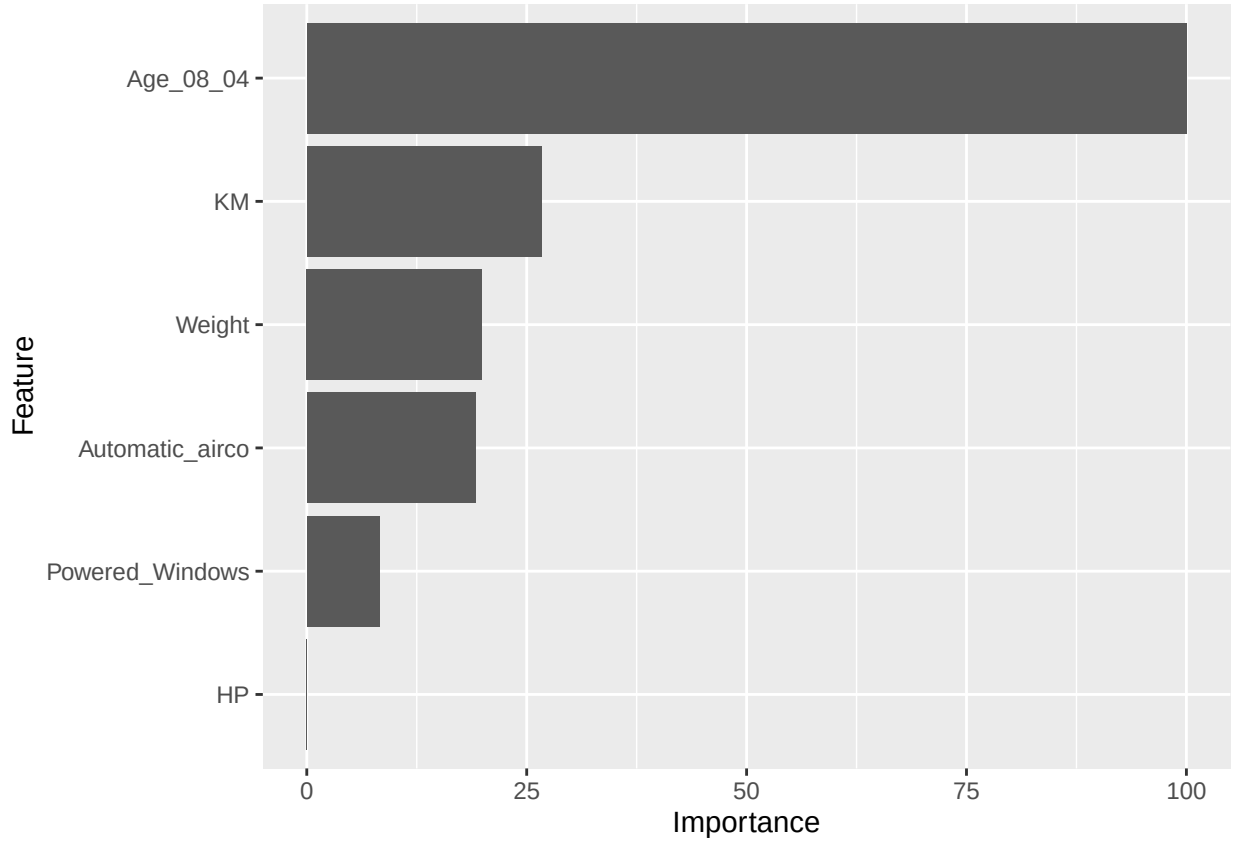
```
## Automatic_airco  2.308e+03  1.983e+02  11.634  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1063 on 1266 degrees of freedom
## Multiple R-squared:  0.8193, Adjusted R-squared:  0.8184
## F-statistic: 956.7 on 6 and 1266 DF,  p-value: < 2.2e-16
```

```
predict_test3 <- predict(lmfit3, dataset5)

model_test_3 <- data.frame(obs= dataset5$Price, pred= predict_test3)

defaultSummary(model_test_3)
```

```
##         RMSE     Rsquared          MAE
## 1060.0007087    0.8193014  810.5922107
```

Con esta última validación confirmamos que nuestro modelo predice bien ya que el rsquared dan los mismos valores.

```
ggplot(varImp(lmfit3))
```



Con este ggplot podemos ver las variables que tienen importancia para nuestro modelo. y aunque predice bien esto nos esta avisando que a la variable HP tranquilamente la podemos descartar del mismo.

Conclusión

Teniendo en cuenta que a nuestro modelo le interesa poder predecir un precio a partir de un conjunto de variables, podemos decir que el mismo es bastante acertado para dicho problema y, segun nuestro criterio, siguiendo este pensamiento a la hora de querer vender o comprar un auto usado podremos aplicarlo y ver los parámetros que tendremos que tener en cuenta para poder conseguir la forma mas optima tanto como para vender como para comprar el vehículo en cuestión. Para este problema las variables a tener en cuenta serían, Age_08_04, KM. Weight, Automatic_airco, Powered_Windows. Para realizar este informe aplicamos todo lo aprendido desde la interpretación de la estructura de los datos hasta la interpretación de los diferentes gráficos y regresiones; los cuales nos fueron guiando para tomar una decisión basada en la información que cada nuevo concepto y técnica aplicada nos brindo.