S

Métodologías supervisadas

Realización de Regresión lineal múltiple, SVM y Árbol de decisión.

Análisis exploratorio de datos

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.1     v readr     2.1.4
v forcats   1.0.0     v stringr   1.5.0
v ggplot2   3.4.2     v tibble    3.2.1
v lubridate 1.9.2     v tidyr     1.3.0
v purrr     1.0.1
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to beco
```

```
library(caret)
```

```
Loading required package: lattice

Attaching package: 'caret'

The following object is masked from 'package:purrr':

    lift
```

```r
library(ggplot2)
library(e1071)
library(DataExplorer)
library(caTools)
library(plotly)
```

```
Attaching package: 'plotly'

The following object is masked from 'package:ggplot2':

    last_plot

The following object is masked from 'package:stats':

    filter

The following object is masked from 'package:graphics':

    layout
```

```r
library(readr)
library(pROC)
```

```
Type 'citation("pROC")' for a citation.

Attaching package: 'pROC'

The following objects are masked from 'package:stats':

    cov, smooth, var
```

```r
library(MASS)
```

```
Attaching package: 'MASS'

The following object is masked from 'package:plotly':
```

```
    select

The following object is masked from 'package:dplyr':

    select
```

```
library(dplyr)
library(factoextra)
```

```
Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

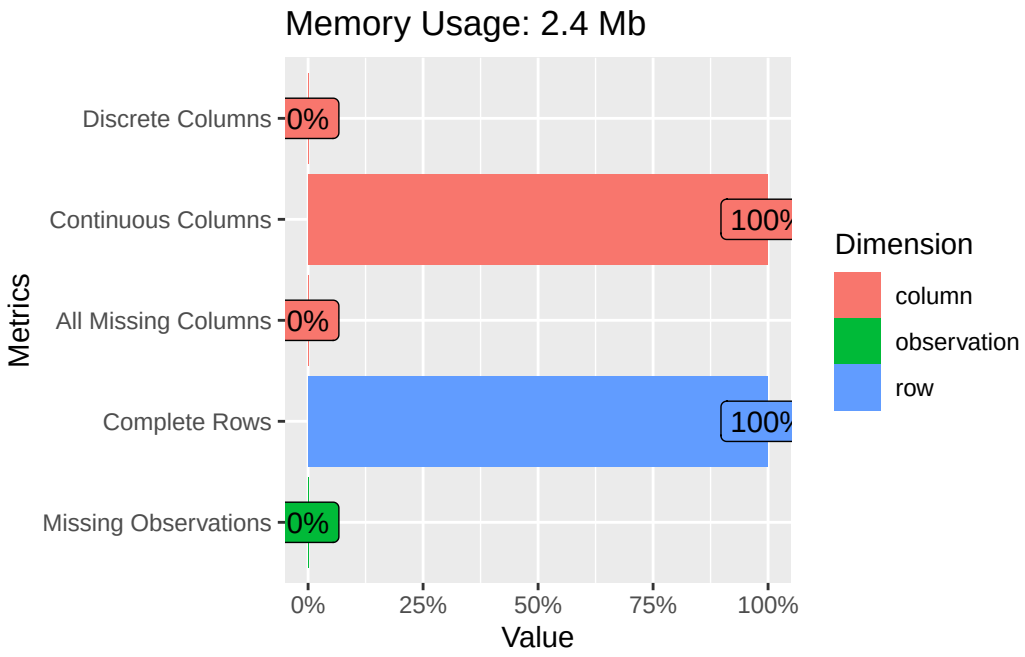
```
library(ggpubr)
library(caret)

# Importación de los datos


data <- read_csv("Steel_industry_data.csv",
                 col_types = cols(date = col_datetime(format = "%d/%m/%Y %H:%M")))

View(data)

data<- as.data.frame(data)

data <- data[, -c(1,10)]

View(data)

data <- data %>%
  mutate(Load_Type = as.character(Load_Type)) %>%
  mutate(Load_Type = case_when(
    Load_Type == "Light_Load" ~ as.numeric("1"),
    Load_Type == "Medium_Load" ~ as.numeric("2"),
    Load_Type == "Maximum_Load" ~ as.numeric("3"),
    TRUE ~ as.numeric(Load_Type)
  ))
```

```
Warning: There was 1 warning in `mutate()`.
i In argument: `Load_Type = case_when(...)`.
Caused by warning:
! NAs introducidos por coerción
```

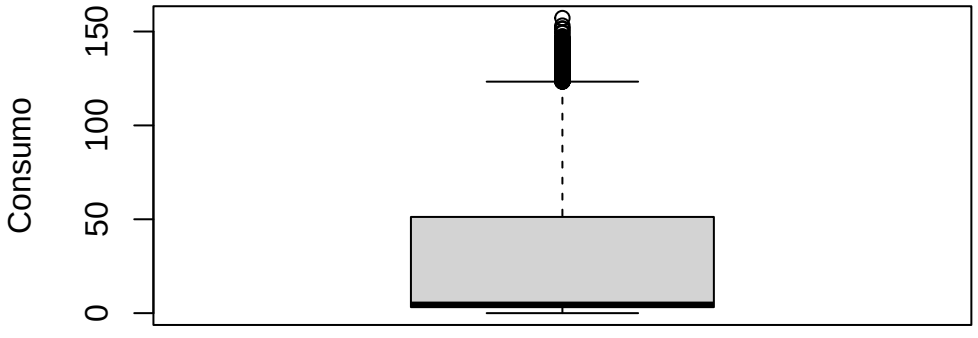
```r
data <- data %>%
  mutate(WeekStatus = as.character(WeekStatus)) %>%
  mutate( WeekStatus = case_when(
    WeekStatus == "Weekday" ~ as.numeric("1"),
    WeekStatus == "Weekend" ~ as.numeric("0"),
    TRUE ~ as.numeric(WeekStatus)
  ))
```

```
Warning: There was 1 warning in `mutate()`.
i In argument: `WeekStatus = case_when(...)`.
Caused by warning:
! NAs introducidos por coerción
```

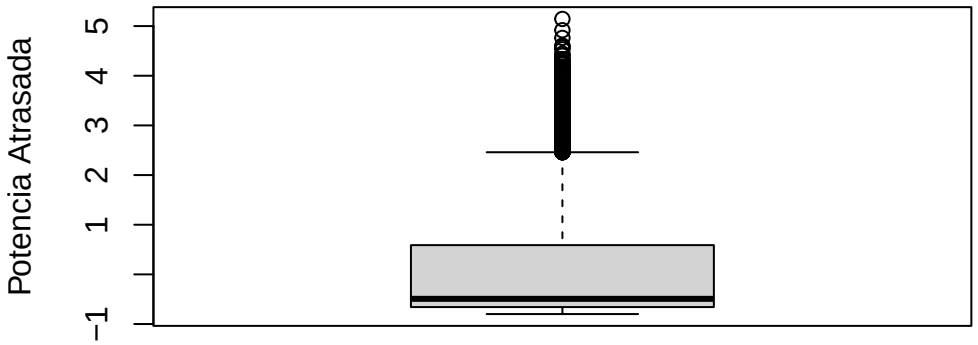
```r
data_sc <- as.data.frame(cbind(scale(data[ ,c(2:9)]), data$Usage_kWh))

View(data_sc)


colnames(data_sc) <- c("Potencia_atrasada",
                       "Potencia_principal",
                       "tCO2",
                       "Factor_potencia_atrasada",
                       "Factor_potencia_principal",
                       "Segundos_desde_medianoche",
                       "Estado_semana" ,
                       "Tipo_carga",
                       "Consumo")


View(data_sc)

plot_intro(data)
```



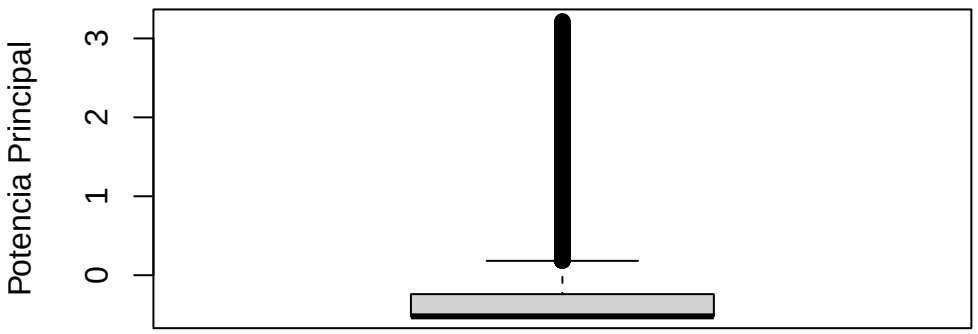

```
attach(data_sc)
boxplot(Consumo,ylab="Consumo")
```



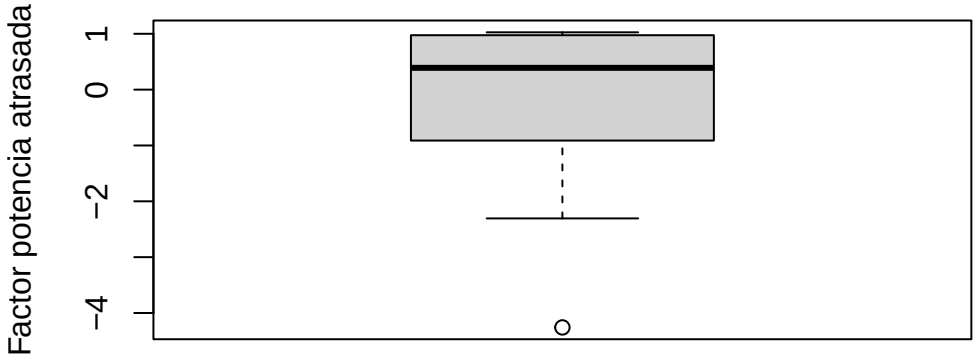
```r
boxplot(Potencia_atrasada,ylab="Potencia Atrasada")
```



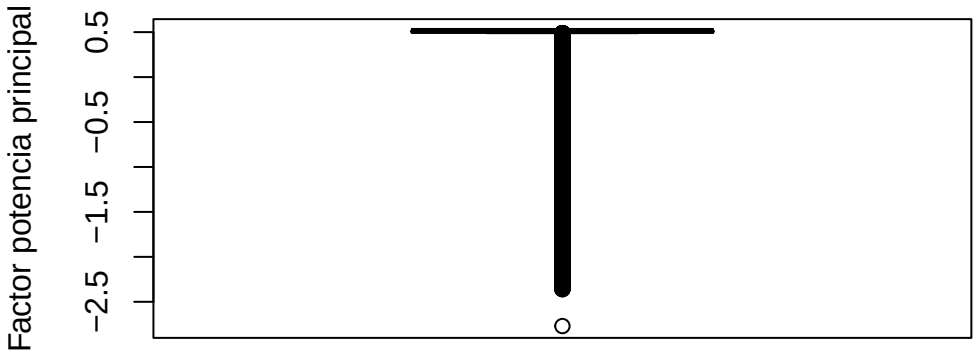

```r
boxplot(Potencia_principal,ylab="Potencia Principal")
```



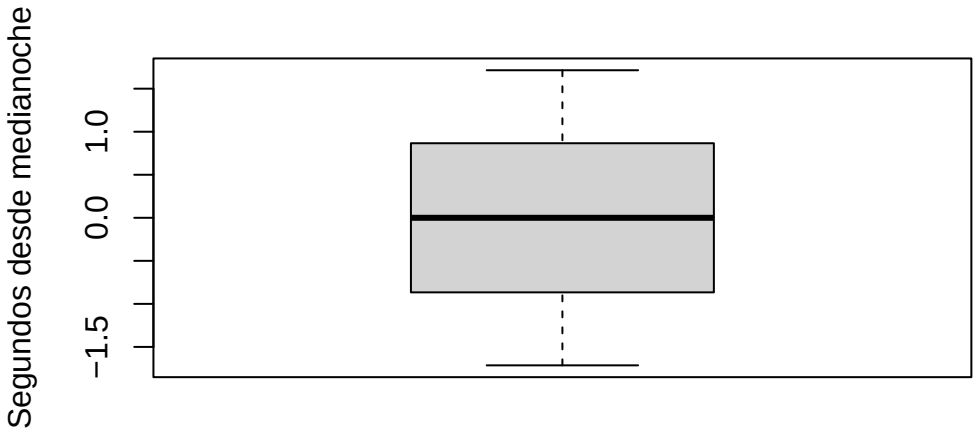

```
boxplot(Factor_potencia_atrasada,ylab="Factor potencia atrasada")
```



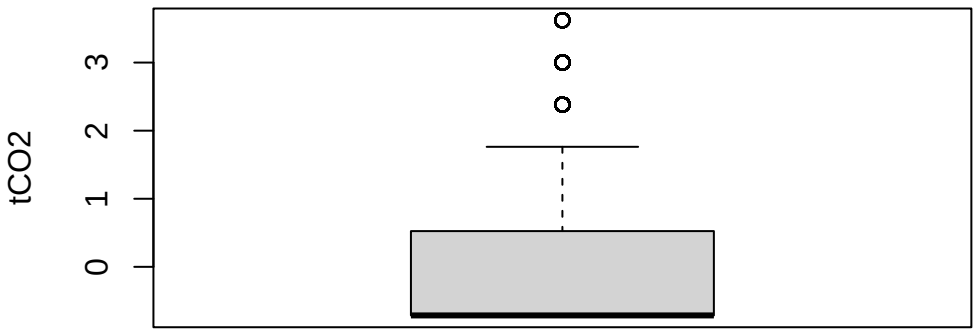
```r
boxplot(Factor_potencia_principal,ylab="Factor potencia principal")
```




```r
boxplot(Segundos_desde_medianoche,ylab="Segundos desde medianoche")
```



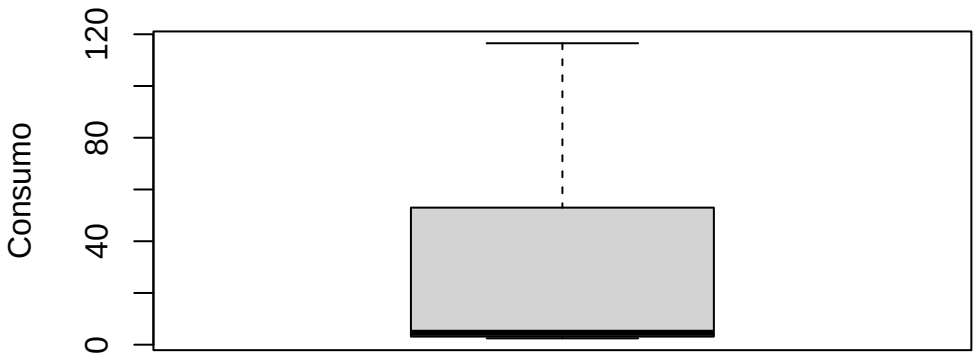

```
boxplot(tCO2,ylab="tCO2")
```



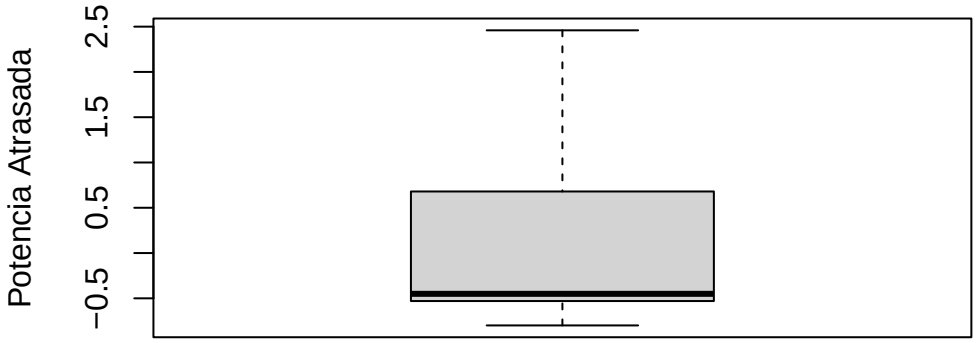
```r
data_sc <- filter_if(data_sc, is.numeric , all_vars(!is.na(.)))

view(data_sc)

for (i in c("Consumo","Potencia_atrasada", "Potencia_principal", "tCO2", "Factor_potencia_
{
  outliers <- boxplot.stats(data_sc[[i]])$out
  data_sc[[i]][data_sc[[i]] %in% outliers] <- NA
}

data_sc <- filter_if(data_sc, is.numeric , all_vars(!is.na(.)))
View(data_sc)

attach(data_sc)
```

```
The following objects are masked from data_sc (pos = 3):

    Consumo, Estado_semana, Factor_potencia_atrasada,
    Factor_potencia_principal, Potencia_atrasada, Potencia_principal,
    Segundos_desde_medianoche, tCO2, Tipo_carga
```

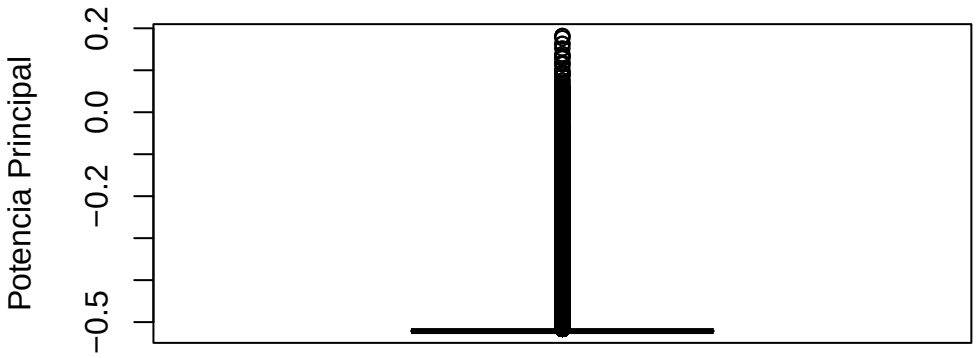
```r
boxplot(Consumo,ylab="Consumo")
```



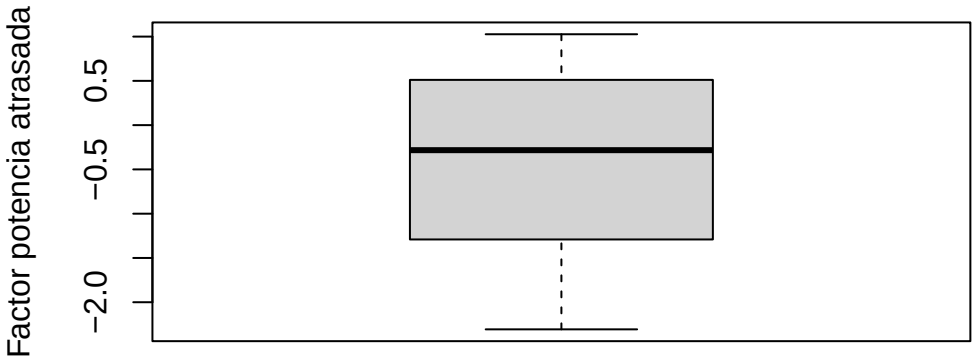

```
boxplot(Potencia_atrasada,ylab="Potencia Atrasada")
```



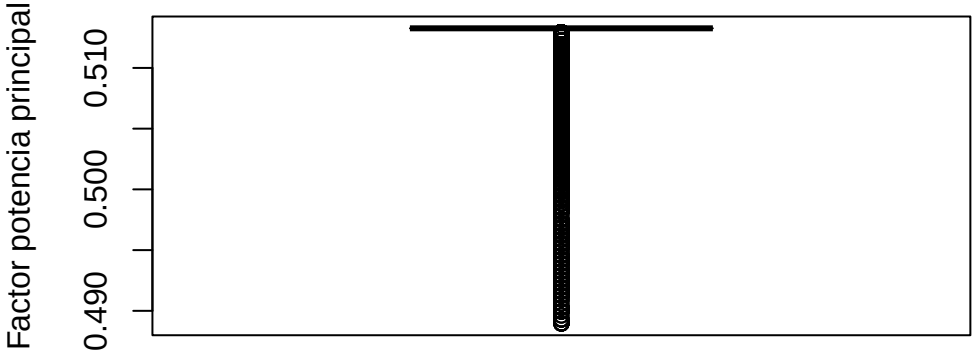
```r
boxplot(Potencia_principal,ylab="Potencia Principal")
```



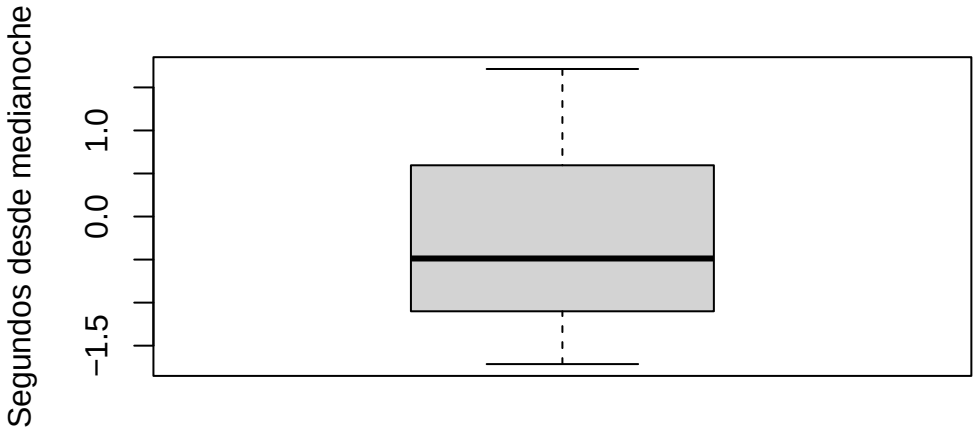

```r
boxplot(Factor_potencia_atrasada,ylab="Factor potencia atrasada")
```



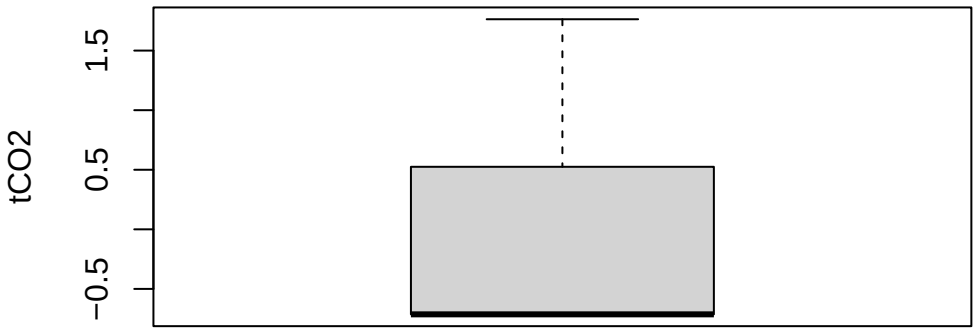

```
boxplot(Factor_potencia_principal,ylab="Factor potencia principal")
```



```r
boxplot(Segundos_desde_medianoche,ylab="Segundos desde medianoche")
```




```r
boxplot(tCO2,ylab="tCO2")
```

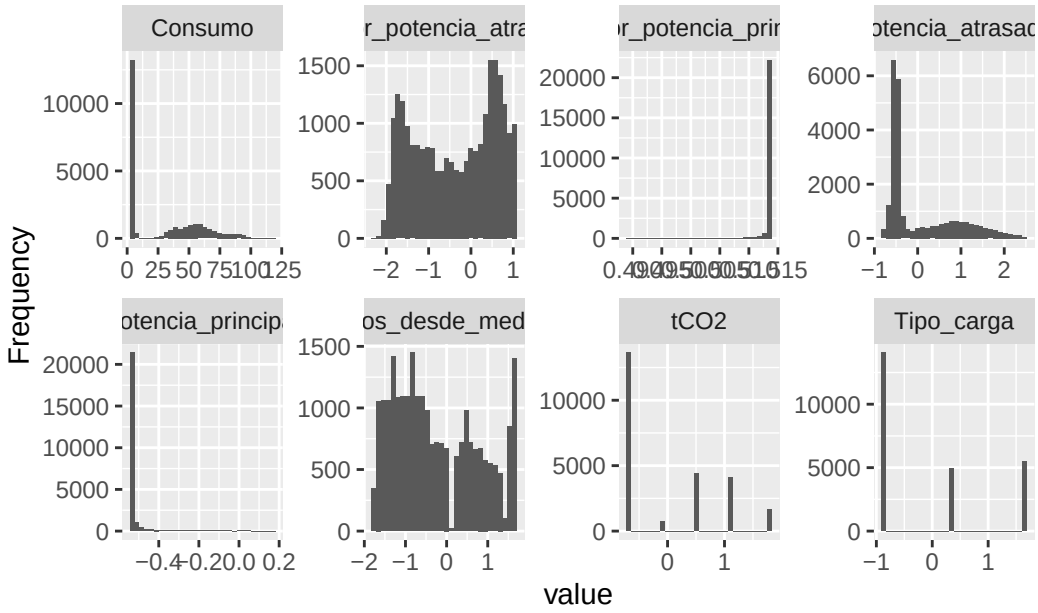



```
summary(data_sc)
```

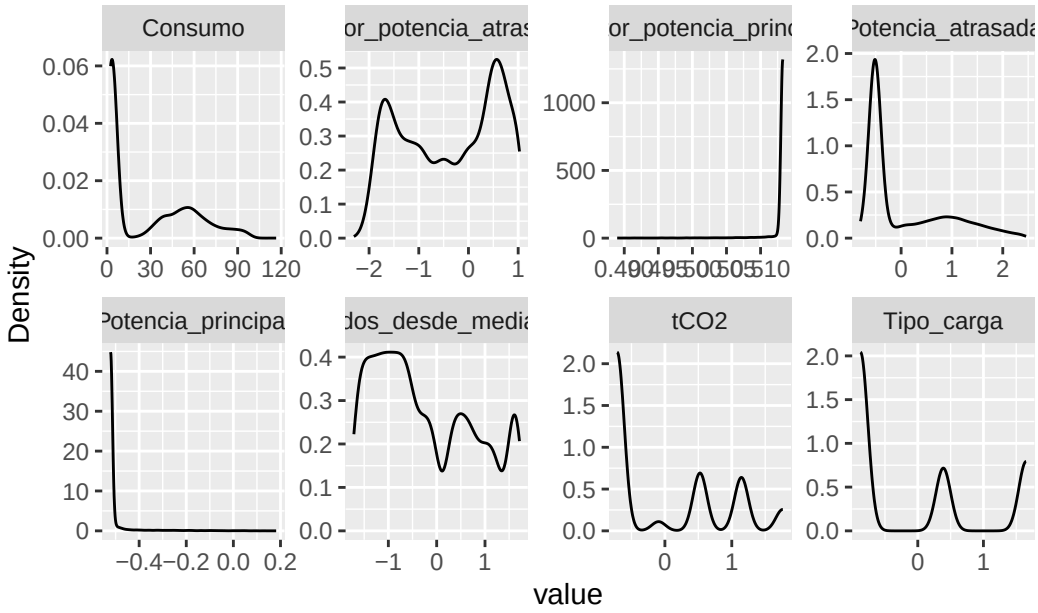
```
Potencia_atrasada  Potencia_principal      tCO2
Min.   :-0.79942   Min.   :-0.5214    Min.   :-0.713540
1st Qu.:-0.53020   1st Qu.:-0.5214    1st Qu.:-0.713540
Median :-0.45047   Median :-0.5214    Median :-0.713540
Mean   : 0.05262   Mean   :-0.5061    Mean   : 0.002744
3rd Qu.: 0.67979   3rd Qu.:-0.5214    3rd Qu.: 0.524787
Max.   : 2.45950   Max.   : 0.1817    Max.   : 1.763114
Factor_potencia_atrasada Factor_potencia_principal Segundos_desde_medianoche
Min.   :-2.3063          Min.   :0.4890            Min.   :-1.7141
1st Qu.:-1.2910          1st Qu.:0.5133            1st Qu.:-1.1006
Median :-0.2826          Median :0.5133            Median :-0.4872
Mean   :-0.3795          Mean   :0.5128            Mean   :-0.2428
3rd Qu.: 0.5112          3rd Qu.:0.5133            3rd Qu.: 0.5954
Max.   : 1.0265          Max.   :0.5133            Max.   : 1.7141
Estado_semana       Tipo_carga          Consumo
Min.   :-1.5842   Min.   :-0.87274   Min.   :  2.45
1st Qu.: 0.6312   1st Qu.:-0.87274   1st Qu.:  3.13
Median : 0.6312   Median :-0.87274   Median :  4.61
Mean   : 0.1334   Mean   :-0.05487   Mean   : 27.25
3rd Qu.: 0.6312   3rd Qu.: 0.38884   3rd Qu.: 52.99
```

```
Max.   : 0.6312    Max.   : 1.65042    Max.   :116.53
```

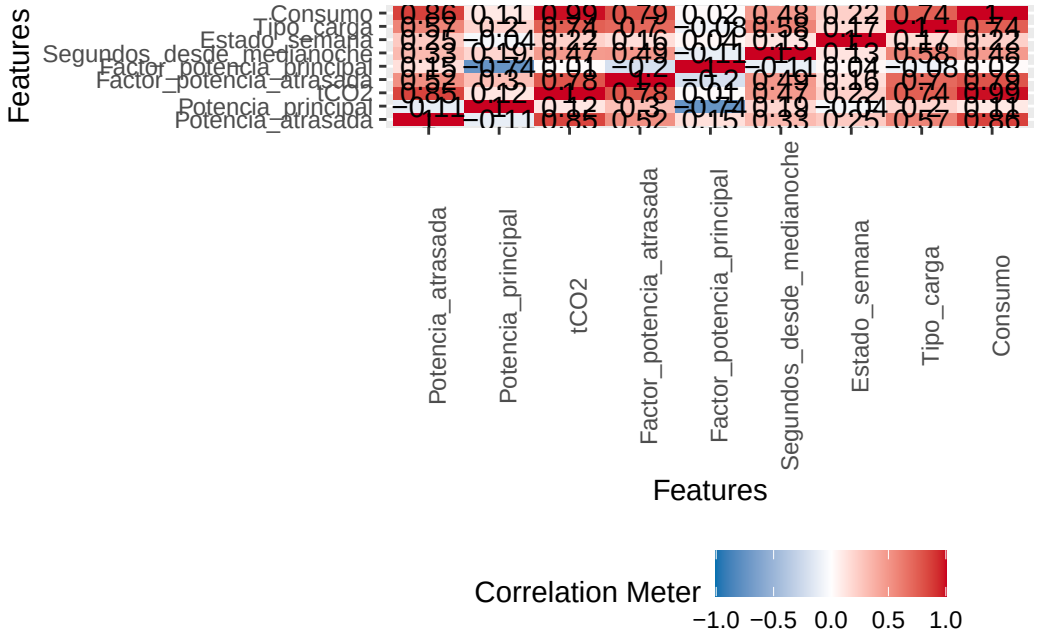
```
plot_histogram(data_sc)
```




```
plot_density(data_sc)
```




```
plot_correlation(data_sc)
```



Modelo de regresión lineal múltiple

```r
set.seed(163)

# Separación entre datos de train y test

sample <- sample.split(data_sc$Consumo, SplitRatio = 0.75)

train <- subset(data_sc, sample == TRUE)
test <- subset(data_sc, sample == FALSE)

## Modelo de regresión lineal
lm_fit <- lm(Consumo ~ Potencia_principal + tCO2 + Factor_potencia_atrasada + Factor_poten

summary(lm_fit)
```

```
Call:
lm(formula = Consumo ~ Potencia_principal + tCO2 + Factor_potencia_atrasada +
    Factor_potencia_principal + Segundos_desde_medianoche + Tipo_carga +
    Potencia_atrasada + Estado_semana, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-14.139  -1.379   0.133   1.125  94.719

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               -77.28332   10.74116  -7.195 6.48e-13 ***
Potencia_principal          4.10784    0.80540   5.100 3.42e-07 ***
tCO2                       26.40006    0.11792 223.880  < 2e-16 ***
Factor_potencia_atrasada    2.54924    0.06614  38.544  < 2e-16 ***
Factor_potencia_principal 209.35842   21.51239   9.732  < 2e-16 ***
Segundos_desde_medianoche   0.13258    0.04069   3.258  0.00112 **
Tipo_carga                  0.62916    0.05352  11.755  < 2e-16 ***
Potencia_atrasada           4.50152    0.09151  49.192  < 2e-16 ***
Estado_semana               0.09416    0.03734   2.522  0.01169 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.531 on 18478 degrees of freedom
```

```
Multiple R-squared:  0.9759,    Adjusted R-squared:  0.9759
F-statistic: 9.372e+04 on 8 and 18478 DF,  p-value: < 2.2e-16
```

```r
predictions <- predict(lm_fit, newdata = test)

rmse <- sqrt(mean((test$Consumo - predictions)^2))
rmse
```

```
[1] 4.095178
```

```r
mae <- mean(abs(test$Consumo - predictions))
mae
```

```
[1] 2.582004
```

```r
rsquared <- 1 - sum((test$Consumo - predictions)^2) / sum((test$Consumo - mean(test$Consum
rsquared
```

```
[1] 0.9792471
```

SVM

```r
## SVM

library(caTools)

svm_fit <- svm(formula = Consumo ~ Potencia_principal + tCO2 + Factor_potencia_atrasada +

summary(svm_fit)
```

```
Call:
svm(formula = Consumo ~ Potencia_principal + tCO2 + Factor_potencia_atrasada +
    Factor_potencia_principal + Segundos_desde_medianoche + Tipo_carga +
    Potencia_atrasada + Estado_semana, data = train, kernel = "linear")
```

```
Parameters:
   SVM-Type:  eps-regression
 SVM-Kernel:  linear
       cost:  1
      gamma:  0.125
    epsilon:  0.1


Number of Support Vectors:  5883
```

```r
svm_fit2 <- svm(formula = Consumo ~ Potencia_principal + tCO2 + Factor_potencia_atrasada +

summary(svm_fit2)
```

```
Call:
svm(formula = Consumo ~ Potencia_principal + tCO2 + Factor_potencia_atrasada +
    Factor_potencia_principal + Segundos_desde_medianoche + Tipo_carga +
    Potencia_atrasada + Estado_semana, data = train, kernel = "radial")


Parameters:
   SVM-Type:  eps-regression
 SVM-Kernel:  radial
       cost:  1
      gamma:  0.125
    epsilon:  0.1


Number of Support Vectors:  1159
```

```r
predictions_linear <- predict(svm_fit, newdata = test)
predictions_radial <- predict(svm_fit2, newdata = test)


linear_metrics <- caret::postResample(predictions_linear, test$Consumo)
radial_metrics <- caret::postResample(predictions_radial, test$Consumo)


rmse_linear <- sqrt(mean(linear_metrics^2))
```

```r
rmse_radial <- sqrt(mean(radial_metrics^2))

rmse_linear
```

```
[1] 2.979357
```

```r
rmse_radial
```

```
[1] 1.34891
```

```r
mae_linear <- mean(abs(linear_metrics))
mae_radial <- mean(abs(radial_metrics))

mae_linear
```

```
[1] 2.676437
```

```r
mae_radial
```

```
[1] 1.319071
```

```r
rsquared_linear <- 1 - sum((test$Consumo - predictions_linear)^2) / sum((test$Consumo - me
rsquared_radial <- 1 - sum((test$Consumo - predictions_radial)^2) / sum((test$Consumo - me

rsquared_linear
```

```
[1] 0.9785404
```

```r
rsquared_radial
```

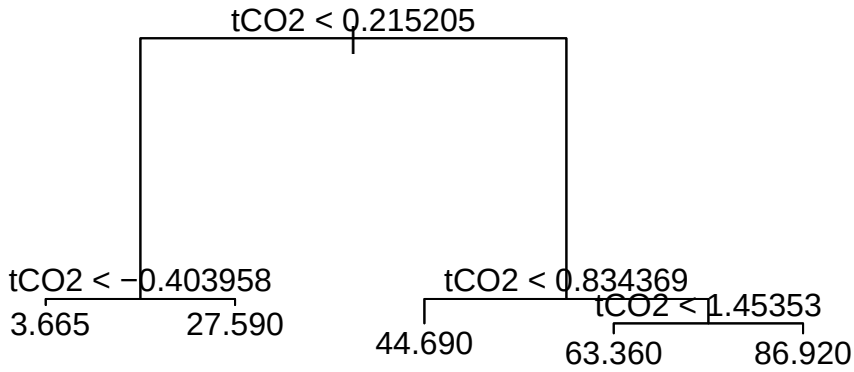
```
[1] 0.9964915
```

Árbol de decisión

```r
## Árbol de decisión

library(tree)


tree.fit = tree(Consumo ~ Potencia_principal + tCO2 + Factor_potencia_atrasada + Factor_po
summary(tree.fit)
```

```
Regression tree:
tree(formula = Consumo ~ Potencia_principal + tCO2 + Factor_potencia_atrasada +
    Factor_potencia_principal + Segundos_desde_medianoche + Tipo_carga +
    Potencia_atrasada + Estado_semana, data = train)
Variables actually used in tree construction:
[1] "tCO2"
Number of terminal nodes:  5
Residual mean deviance:  23.46 = 433600 / 18480
Distribution of residuals:
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-15.8200  -0.9654  -0.3994   0.0000   1.0150 112.9000
```

```r
plot(tree.fit)
text(tree.fit, pretty=0)
```



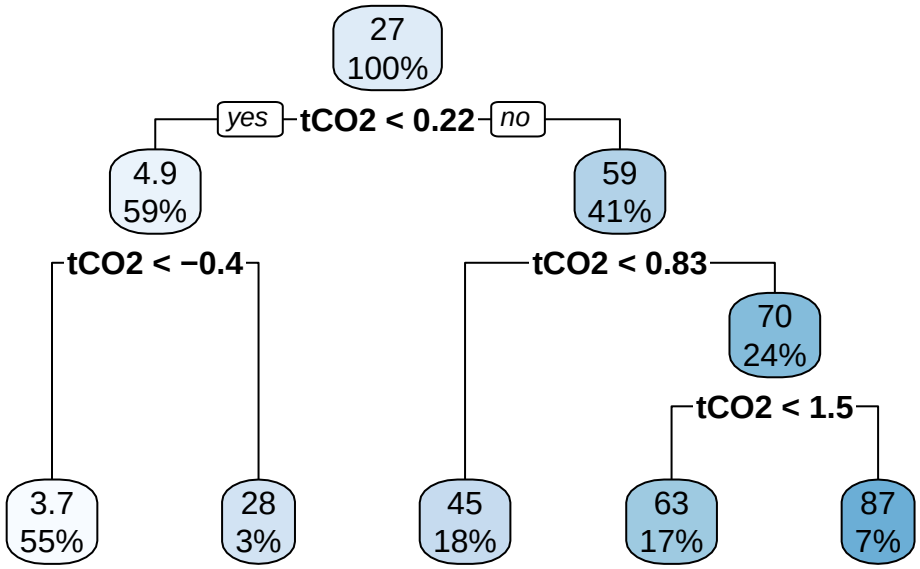

```
library(rpart)
library(rpart.plot)


tree.fit2 = rpart(formula=Consumo ~ Potencia_principal + tCO2 + Factor_potencia_atrasada +
summary(tree.fit2)
```

```
Call:
rpart(formula = Consumo ~ Potencia_principal + tCO2 + Factor_potencia_atrasada +
    Factor_potencia_principal + Segundos_desde_medianoche + Tipo_carga +
    Potencia_atrasada + Estado_semana, data = train)
  n= 18487

          CP nsplit  rel error     xerror        xstd
1 0.84179357      0 1.00000000 1.00007107 0.007535427
2 0.07881587      1 0.15820643 0.15822944 0.002898017
3 0.03219313      2 0.07939057 0.07941838 0.002219527
4 0.01969933      3 0.04719744 0.04722770 0.002069263
5 0.01000000      4 0.02749811 0.02752356 0.001957094

Variable importance
                    tCO2         Potencia_atrasada  Factor_potencia_atrasada
```

```
                    26                          20                          19
          Tipo_carga Segundos_desde_medianoche        Potencia_principal
                    17                          15                           2

Node number 1: 18487 observations,    complexity param=0.8417936
  mean=27.48368, MSE=852.9793
  left son=2 (10822 obs) right son=3 (7665 obs)
  Primary splits:
      tCO2                      < 0.2152053   to the left,  improve=0.8417936, (0 missing)
      Potencia_atrasada         < -0.2992999  to the left,  improve=0.7898259, (0 missing)
      Factor_potencia_atrasada  < 0.0886272   to the left,  improve=0.7020153, (0 missing)
      Tipo_carga                < -0.2419465  to the left,  improve=0.6168224, (0 missing)
      Segundos_desde_medianoche < -0.5412875  to the left,  improve=0.5533223, (0 missing)
  Surrogate splits:
      Potencia_atrasada         < -0.2992999  to the left,  agree=0.946, adj=0.869, (0 split)
      Factor_potencia_atrasada  < -0.03187178 to the left,  agree=0.924, adj=0.817, (0 split)
      Tipo_carga                < -0.2419465  to the left,  agree=0.902, adj=0.764, (0 split)
      Segundos_desde_medianoche < -0.5052017  to the left,  agree=0.871, adj=0.690, (0 split)
      Potencia_principal        < -0.4701955  to the left,  agree=0.614, adj=0.068, (0 split)

Node number 2: 10822 observations,    complexity param=0.01969933
  mean=4.932233, MSE=42.11627
  left son=4 (10249 obs) right son=5 (573 obs)
  Primary splits:
      tCO2                     < -0.4039583  to the left,  improve=0.6815519, (0 missing)
      Factor_potencia_atrasada < 0.491876    to the left,  improve=0.5517775, (0 missing)
      Potencia_atrasada        < -0.291634   to the left,  improve=0.3616348, (0 missing)
      Potencia_principal       < -0.4607672  to the left,  improve=0.3590909, (0 missing)
      Tipo_carga               < -0.2419465  to the left,  improve=0.2060862, (0 missing)
  Surrogate splits:
      Factor_potencia_atrasada  < 0.6454065   to the left,  agree=0.973, adj=0.494, (0 split)
      Potencia_principal        < -0.456053   to the left,  agree=0.970, adj=0.424, (0 split)
      Potencia_atrasada         < -0.2683297  to the left,  agree=0.969, adj=0.422, (0 split)
      Factor_potencia_principal < 0.4901126   to the right, agree=0.947, adj=0.003, (0 split)

Node number 3: 7665 observations,    complexity param=0.07881587
  mean=59.32343, MSE=266.0117
  left son=6 (3303 obs) right son=7 (4362 obs)
  Primary splits:
      tCO2                      < 0.8343688   to the left,  improve=0.60954490, (0 missing)
      Potencia_atrasada         < 1.025366    to the left,  improve=0.22617250, (0 missing)
      Factor_potencia_atrasada  < 0.9950121   to the right, improve=0.07984928, (0 missing)
      Factor_potencia_principal < 0.5130961   to the left,  improve=0.07542308, (0 missing)
```

```
      Potencia_principal        < -0.5186838  to the right, improve=0.07235809, (0 missing)
  Surrogate splits:
      Potencia_atrasada         < 0.6322591   to the left,  agree=0.743, adj=0.404, (0 split)
      Potencia_principal        < -0.5186838  to the right, agree=0.644, adj=0.175, (0 split)
      Factor_potencia_atrasada  < 0.8179631   to the right, agree=0.639, adj=0.163, (0 split)
      Factor_potencia_principal < 0.5130961   to the left,  agree=0.626, adj=0.133, (0 split)
      Estado_semana             < -0.4764601  to the left,  agree=0.601, adj=0.073, (0 split)

Node number 4: 10249 observations
  mean=3.665425, MSE=12.92414

Node number 5: 573 observations
  mean=27.59108, MSE=22.1353

Node number 6: 3303 observations
  mean=44.69015, MSE=38.14039

Node number 7: 4362 observations,    complexity param=0.03219313
  mean=70.40407, MSE=153.6342
  left son=14 (3058 obs) right son=15 (1304 obs)
  Primary splits:
      tCO2                      < 1.453532    to the left,  improve=0.757521100, (0 missing)
      Potencia_atrasada         < 1.531008    to the left,  improve=0.075954400, (0 missing)
      Factor_potencia_atrasada  < 0.3449518   to the left,  improve=0.072276350, (0 missing)
      Estado_semana             < -0.4764601  to the right, improve=0.009387452, (0 missing)
      Segundos_desde_medianoche < 1.29909     to the left,  improve=0.009295822, (0 missing)
  Surrogate splits:
      Segundos_desde_medianoche < 1.335176    to the left,  agree=0.704, adj=0.012, (0 split)
      Potencia_atrasada         < 1.84439     to the left,  agree=0.704, adj=0.008, (0 split)

Node number 14: 3058 observations
  mean=63.35939, MSE=36.41319

Node number 15: 1304 observations
  mean=86.92449, MSE=39.22257
```

```
ree_plot = rpart.plot(tree.fit2)
```




```
predictions <- predict(tree.fit2, newdata = test)

rmse <- sqrt(mean((test$Consumo - predictions)^2))
rmse
```

```
[1] 4.269308
```

```
mae <- mean(abs(test$Consumo - predictions))
mae
```

```
[1] 2.646204
```

```
rsquared <- 1 - sum((test$Consumo - predictions)^2) / sum((test$Consumo - mean(test$Consum
rsquared
```

```
[1] 0.9774447
```