Reto Etapa 2

Ruben Dario Castro Terrazas

2023-08-21

# Importar librerias y dependencias

library(readxl)

## Warning: package 'readxl' was built under R version 4.2.3

library(Hmisc)

## Warning: package 'Hmisc' was built under R version 4.2.3

##   
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:base':  
##   
## format.pval, units

library(gridExtra)

## Warning: package 'gridExtra' was built under R version 4.2.2

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.2.3

library(corrplot)

## Warning: package 'corrplot' was built under R version 4.2.2

## corrplot 0.92 loaded

# Etapa I

# Etapa II: Comprensión y preparación de los datos

Para dirigir la solución al reto, seguirás la metodología CRISP-DM. En esta primera actividad realizarás el segundo paso Comprensión y Preparación de los datos que permitirá añadir el análisis descriptivo e introductorio de las variables.

## 1) Comprensión de los datos del negocio

### Importar base de datos

sheet\_names = excel\_sheets("datos2022\_2023\_estaciones.xlsx")  
  
for (i in 1:length(sheet\_names)) {  
 variable\_name = paste("M",sheet\_names[i],sep="\_")  
 value = read\_excel("datos2022\_2023\_estaciones.xlsx",sheet = sheet\_names[i])  
 assign(variable\_name,value) #guardamos cada hoja en una variable  
}

Lo que se hizo en el código de arriba fue guardar los datos de cada estación en una variable. Por fines prácticos utilizaremos solamente una de ellas para realizar el análisis descriptivo de esta entrega.

### A Dimensión del dataset

cat("Número de observaciones:",dim(M\_CENTRO)[1])

## Número de observaciones: 14255

cat("\nNúmero de variables:",dim(M\_CENTRO)[2])

##   
## Número de variables: 16

### Describe claramente cada una de las variables, incluyendo su nombre, descripción, tipo (categórico/Numérico) y valores posibles que puede tomar, valores nulos.

cat("Tipo dato de cada variable\n")

## Tipo dato de cada variable

t(sapply(M\_CENTRO,class))

## date CO NO NO2 NOX O3   
## [1,] character,2 "character" "character" "character" "character" "character"  
## PM10 PM2.5 PRS RAINF RH SO2   
## [1,] "character" "character" "character" "character" "character" "character"  
## SR TOUT WSR WDR   
## [1,] "character" "character" "character" "character"

summary(M\_CENTRO)

## date CO NO   
## Min. :2022-01-01 00:00:00.00 Length:14255 Length:14255   
## 1st Qu.:2022-05-29 11:30:00.00 Class :character Class :character   
## Median :2022-10-24 23:00:00.00 Mode :character Mode :character   
## Mean :2022-10-24 23:18:54.42   
## 3rd Qu.:2023-03-22 11:30:00.00   
## Max. :2023-08-17 23:00:00.00   
## NO2 NOX O3 PM10   
## Length:14255 Length:14255 Length:14255 Length:14255   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
## PM2.5 PRS RAINF RH   
## Length:14255 Length:14255 Length:14255 Length:14255   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
## SO2 SR TOUT WSR   
## Length:14255 Length:14255 Length:14255 Length:14255   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
## WDR   
## Length:14255   
## Class :character   
## Mode :character   
##   
##   
##

Si observamos a simple vista la tabla hecha por la función summary(), inicialmente vemos que las variables que deberían catalogarse como cuantitativas se identifican como carateres, lo cual no resulta eficiente para hacer un analisis estadistico y descriptivo. Para solucionar esto, transformaremos los datos de las variables para que sean numéricos.

### Transformación a variables numéricas

variables\_names = colnames(M\_CENTRO)  
M\_CENTRO[,2:length(variables\_names)] = sapply(M\_CENTRO[,2:length(variables\_names)],as.numeric)  
t(sapply(M\_CENTRO,class))

## date CO NO NO2 NOX O3 PM10   
## [1,] character,2 "numeric" "numeric" "numeric" "numeric" "numeric" "numeric"  
## PM2.5 PRS RAINF RH SO2 SR TOUT   
## [1,] "numeric" "numeric" "numeric" "numeric" "numeric" "numeric" "numeric"  
## WSR WDR   
## [1,] "numeric" "numeric"

cat("Valores únicos por cada variable\n")

## Valores únicos por cada variable

for (i in 1:length(variables\_names)) {  
cat("\n-",variables\_names[i],":",length(unique(sapply(M\_CENTRO [,i],as.list)))) #Transformamos las columnas a vectores para poder calcular sus valores únicos  
}

##   
## - date : 14255  
## - CO : 404  
## - NO : 655  
## - NO2 : 617  
## - NOX : 1021  
## - O3 : 133  
## - PM10 : 1090  
## - PM2.5 : 4420  
## - PRS : 240  
## - RAINF : 2  
## - RH : 93  
## - SO2 : 226  
## - SR : 984  
## - TOUT : 3222  
## - WSR : 209  
## - WDR : 360

cat("\nNúmero de valores nulos (NA) por variable")

##   
## Número de valores nulos (NA) por variable

for (i in 1:length(variables\_names)){  
 cat(cat("\n-",variables\_names[i],":",length(which(is.na(M\_CENTRO[,i]),arr.ind=TRUE))))  
}

##   
## - date : 0  
## - CO : 544  
## - NO : 1546  
## - NO2 : 446  
## - NOX : 444  
## - O3 : 908  
## - PM10 : 586  
## - PM2.5 : 2506  
## - PRS : 452  
## - RAINF : 424  
## - RH : 584  
## - SO2 : 2258  
## - SR : 266  
## - TOUT : 424  
## - WSR : 426  
## - WDR : 458

### Exploración de los datos

1. Variables cuantitativas
   * Medidas de posición no-central: cuartiles, outlier (valores atípicos), boxplots

summary(M\_CENTRO)

## date CO NO   
## Min. :2022-01-01 00:00:00.00 Min. :0.070 Min. : 0.600   
## 1st Qu.:2022-05-29 11:30:00.00 1st Qu.:1.150 1st Qu.: 2.500   
## Median :2022-10-24 23:00:00.00 Median :1.510 Median : 3.600   
## Mean :2022-10-24 23:18:54.42 Mean :1.675 Mean : 8.444   
## 3rd Qu.:2023-03-22 11:30:00.00 3rd Qu.:2.030 3rd Qu.: 8.700   
## Max. :2023-08-17 23:00:00.00 Max. :4.940 Max. :166.600   
## NA's :272 NA's :773   
## NO2 NOX O3 PM10   
## Min. : 0.40 Min. : 2.70 Min. : 1.00 Min. : 3.00   
## 1st Qu.: 7.90 1st Qu.: 10.90 1st Qu.: 12.00 1st Qu.: 41.00   
## Median :14.20 Median : 18.70 Median : 23.00 Median : 57.00   
## Mean :16.76 Mean : 25.17 Mean : 26.65 Mean : 64.19   
## 3rd Qu.:22.60 3rd Qu.: 31.90 3rd Qu.: 37.00 3rd Qu.: 78.00   
## Max. :89.10 Max. :186.80 Max. :164.00 Max. :735.00   
## NA's :223 NA's :222 NA's :454 NA's :293   
## PM2.5 PRS RAINF RH   
## Min. : 0.00 Min. :698.7 Min. :0 Min. : 1.00   
## 1st Qu.: 14.89 1st Qu.:709.0 1st Qu.:0 1st Qu.:36.00   
## Median : 21.68 Median :711.0 Median :0 Median :53.00   
## Mean : 24.86 Mean :711.4 Mean :0 Mean :51.99   
## 3rd Qu.: 31.42 3rd Qu.:713.3 3rd Qu.:0 3rd Qu.:69.00   
## Max. :120.10 Max. :724.9 Max. :0 Max. :92.00   
## NA's :1253 NA's :226 NA's :212 NA's :292   
## SO2 SR TOUT WSR   
## Min. : 0.500 Min. :0.0000 Min. :-1.89 Min. : 0.700   
## 1st Qu.: 3.200 1st Qu.:0.0010 1st Qu.:18.27 1st Qu.: 5.100   
## Median : 4.800 Median :0.0080 Median :23.79 Median : 7.800   
## Mean : 5.175 Mean :0.1918 Mean :22.91 Mean : 8.213   
## 3rd Qu.: 6.100 3rd Qu.:0.3190 3rd Qu.:28.04 3rd Qu.:10.800   
## Max. :73.900 Max. :1.0570 Max. :41.47 Max. :24.800   
## NA's :1129 NA's :133 NA's :212 NA's :213   
## WDR   
## Min. : 1.0   
## 1st Qu.: 47.0   
## Median : 72.0   
## Mean :102.1   
## 3rd Qu.:117.0   
## Max. :360.0   
## NA's :229

#Desviacion estandar por variable  
cat("Desviacion estandar por variable")

## Desviacion estandar por variable

for (i in 1:length(variables\_names)) {  
cat(cat("\n-",variables\_names[i],":",sapply(M\_CENTRO[,i],var)))  
}

##   
## - date : 2.195167e+14  
## - CO : NA  
## - NO : NA  
## - NO2 : NA  
## - NOX : NA  
## - O3 : NA  
## - PM10 : NA  
## - PM2.5 : NA  
## - PRS : NA  
## - RAINF : NA  
## - RH : NA  
## - SO2 : NA  
## - SR : NA  
## - TOUT : NA  
## - WSR : NA  
## - WDR : NA

Vemos que obtenemos “NA”s o “valores faltantes” cuando tratamos de obtener la desviación estándar de cada variable-a excepción de la variable de fecha ya que esa es categórica y no aporta valor al análisis. Esto se debe principalmente a que R no acepta valores faltantes al calcular la desviación estandar o la varianza por cada columna. Mucho del análisis requiere que no haya valores faltantes, por lo que haremos una limpieza de ellos para seguir con el análisis de exploración.

#Tratamiento de valores faltantes  
M\_CENTRO = na.exclude(M\_CENTRO)

#Desviacion estandar por variable  
cat("Desviacion estandar por variable")

## Desviacion estandar por variable

for (i in 1:length(variables\_names)) {  
cat(cat("\n-",variables\_names[i],":",sapply(M\_CENTRO[,i],var)))  
}

##   
## - date : 2.31351e+14  
## - CO : 0.5602519  
## - NO : 124.1917  
## - NO2 : 132.7635  
## - NOX : 406.2848  
## - O3 : 394.929  
## - PM10 : 1265.622  
## - PM2.5 : 202.5249  
## - PRS : 13.17473  
## - RAINF : 0  
## - RH : 407.0543  
## - SO2 : 8.267503  
## - SR : 0.08622075  
## - TOUT : 55.94773  
## - WSR : 15.21981  
## - WDR : 7534.048

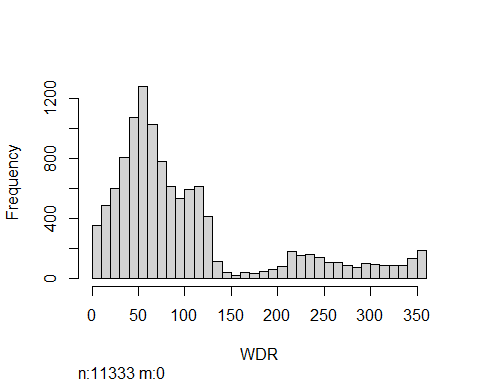
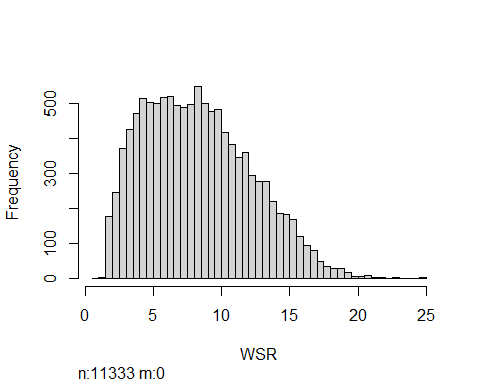
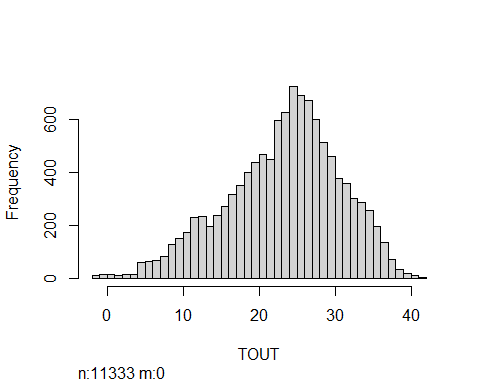
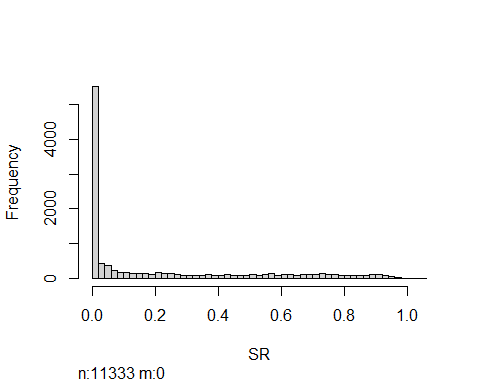
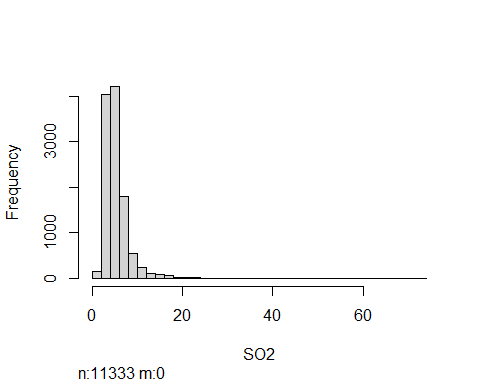
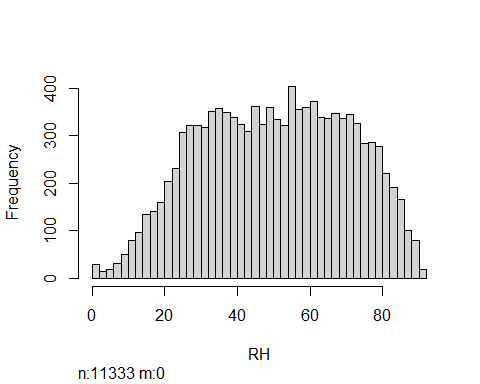
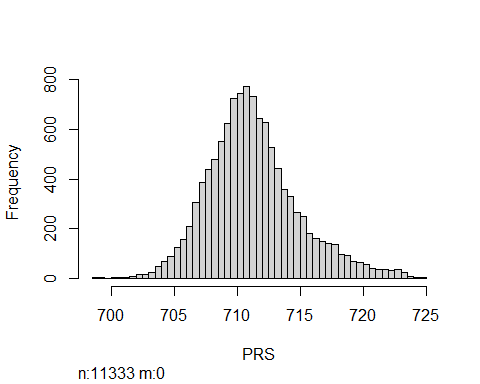
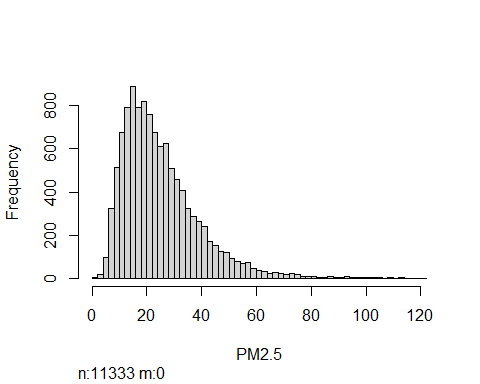
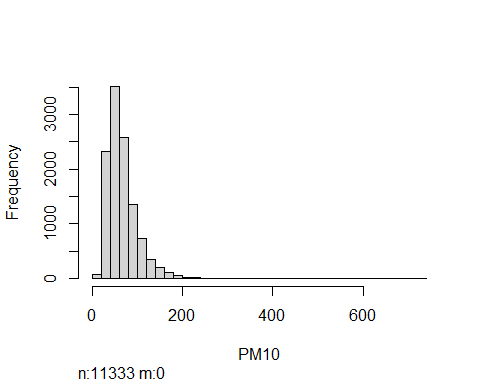
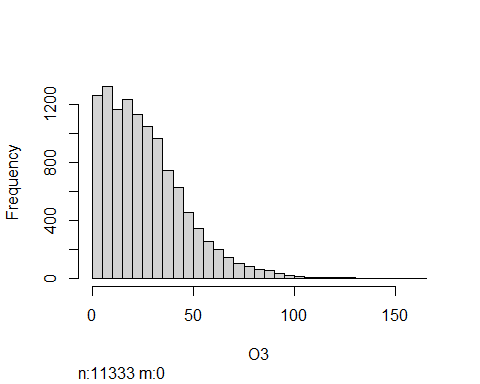
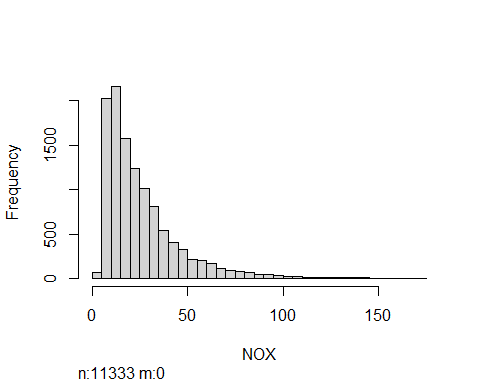
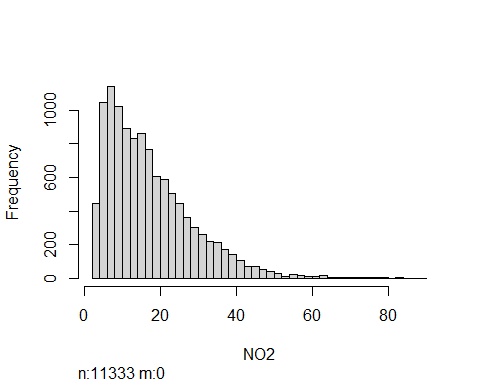
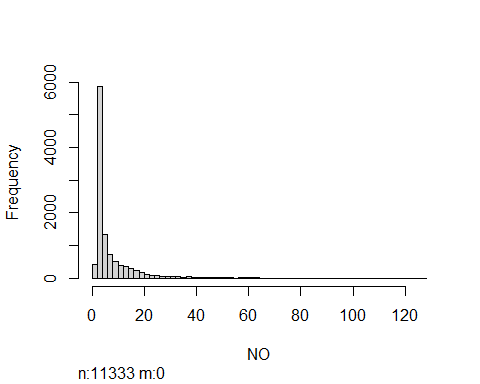
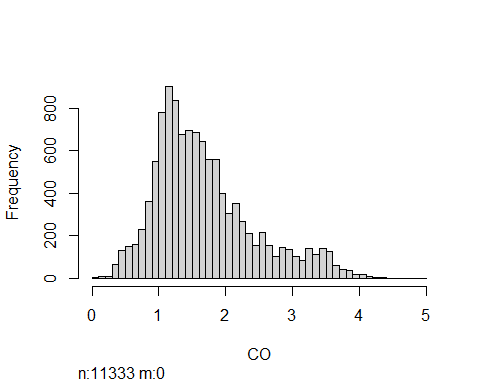
Ahora, trataremos de detectar los outliers o valores atípicos dentro de los datos de la estación central (M\_CENTRO) calculando los rangos intercuartílicos (IQR) de cada variable.

#Detección de datos atípicos al calcular los límites de los valores atípicos utilizando IQR  
  
outliers\_table = data.frame()  
  
  
q1 <- quantile(M\_CENTRO$CO, 0.25)  
q3 = quantile(M\_CENTRO$CO,0.75)  
iqr = q3 - q1  
lowerLimit = q1 - 1.5 \* iqr  
upperLimit = q3 + 1.5 \* iqr  
variable\_name = paste("outliers",variables\_names[i],sep="\_")  
outliers <- M\_CENTRO[M\_CENTRO$CO < lowerLimit | M\_CENTRO$CO > upperLimit,]  
outliers

## # A tibble: 486 × 16  
## date CO NO NO2 NOX O3 PM10 PM2.5 PRS RAINF  
## <dttm> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 2022-01-09 00:00:00 3.44 44.6 66.7 111. 3 139 82.9 713. 0  
## 2 2022-01-09 01:00:00 3.52 60.4 67.3 128. 4 172 107. 713. 0  
## 3 2022-03-10 11:00:00 3.6 57.1 61.7 119 18 126 82.0 710. 0  
## 4 2022-04-05 08:00:00 3.53 58.7 27.4 86.4 3 114 55.3 706. 0  
## 5 2022-04-05 09:00:00 3.53 46.1 35.4 81.7 7 132 54.1 706. 0  
## 6 2022-05-16 09:00:00 3.49 41.5 39 80.8 11 78 34.5 711. 0  
## 7 2022-05-16 10:00:00 3.61 42.8 55.6 98.5 19 93 48.9 711. 0  
## 8 2022-05-17 08:00:00 3.5 19.3 26.7 46.1 10 80 27.5 710. 0  
## 9 2022-05-17 09:00:00 3.69 21.2 31.9 53.3 16 82 30.2 711. 0  
## 10 2022-05-18 11:00:00 3.45 15.3 41.9 57.5 36 106 46.8 710. 0  
## # … with 476 more rows, and 6 more variables: RH <dbl>, SO2 <dbl>, SR <dbl>,  
## # TOUT <dbl>, WSR <dbl>, WDR <dbl>

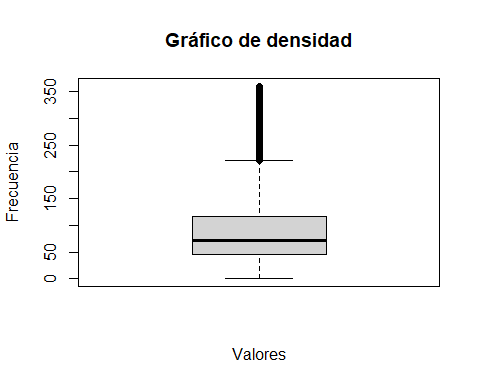
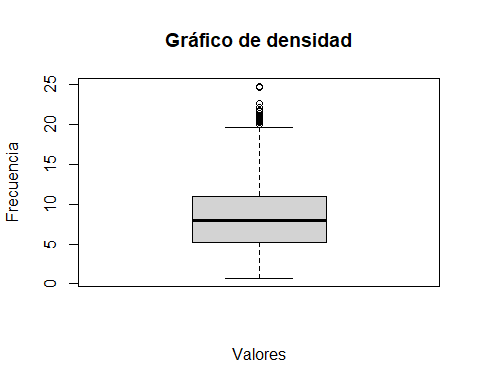
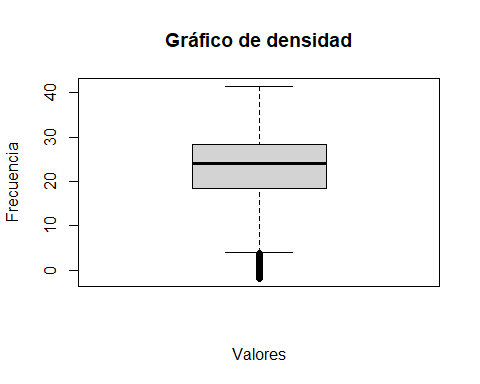
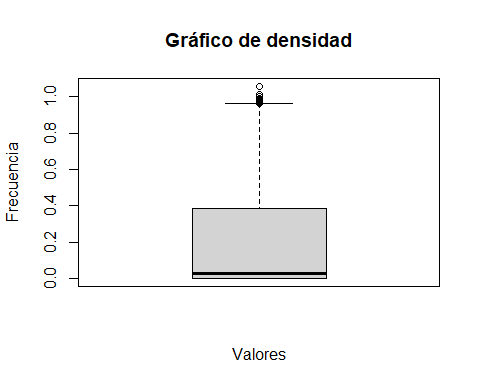
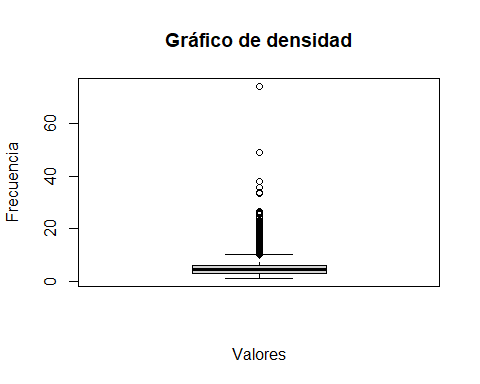
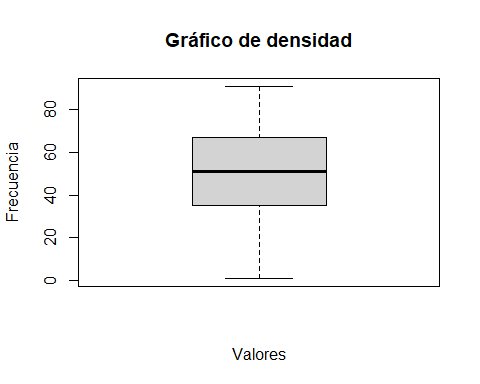
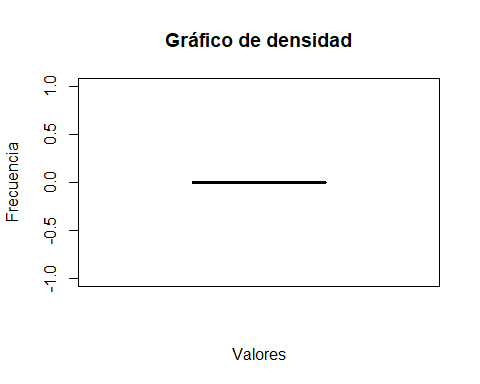
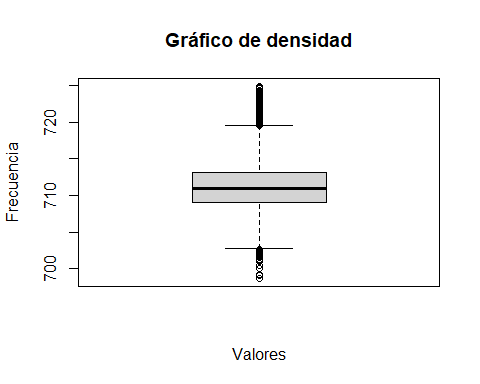
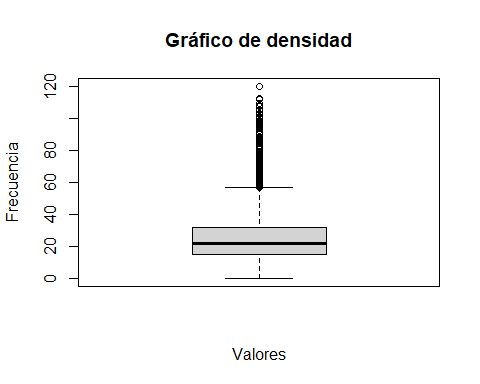
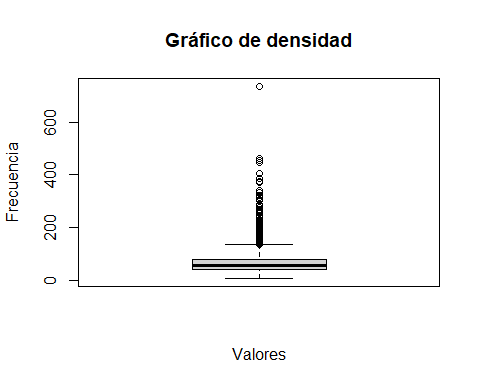
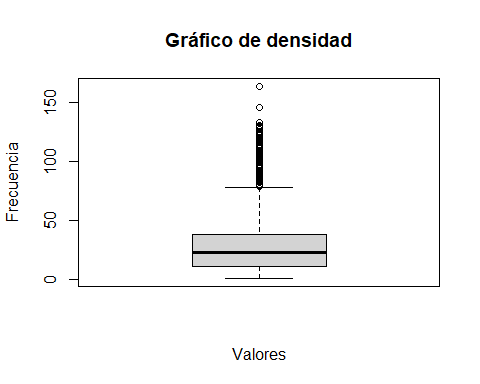
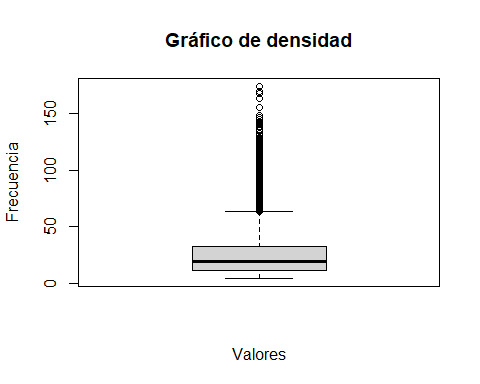
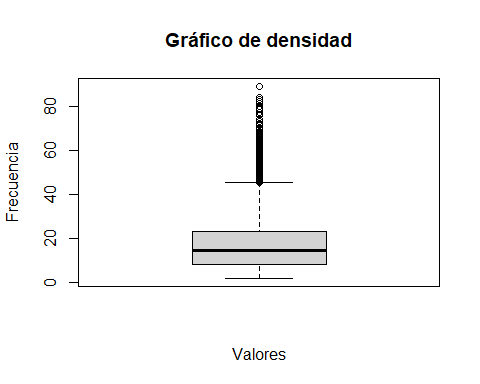
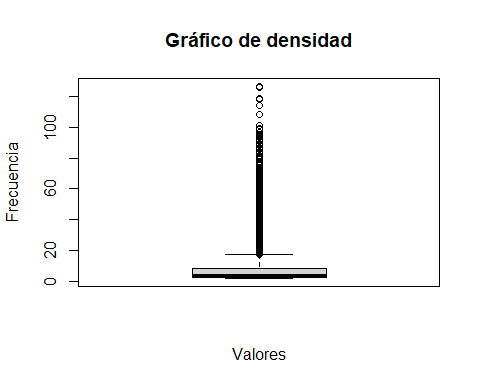
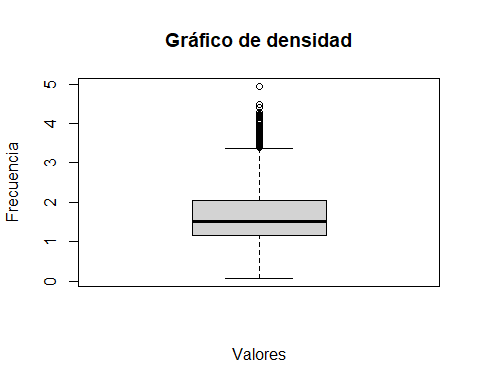
* Análisis de distribución de los datos (Histogramas). Identificar si tiene forma simétrica o asimétrica

#Histograma  
#hist\_list = list()  
#par(mfrow = c(4,4))  
#for (i in 2:length(variables\_names)) {  
 #hist\_plot = ggplot(M\_CENTRO, aes(x = M\_CENTRO[,i])) +  
 #geom\_histogram(binwidth = 0.5, fill = "blue", color = "black", alpha = 0.7) +  
 #labs(title = paste("Histogram for", variables\_names[i]))  
   
 # Agregamos los histogramas a la lista  
 #hist\_list[[variables\_names[i]]] <- hist\_plot  
#}  
  
#grid.arrange(grobs = hist\_list, ncol = 15)  
for (i in 2:length(variables\_names)) {  
hist(M\_CENTRO[,i], main="Distribución de datos", xlab="Valores", ylab="Frecuencia")  
}



A simple vista, podemos deducir que la mayoría de las variables cauntitativas tienen un sesgo hacia la derecha, a excepción de variables como “PRS”,“RH”,y “TOUT”.

#Diagramas de caja y bigotes  
  
for (i in 2:length(variables\_names)) {  
boxplot(M\_CENTRO[,i], main="Gráfico de densidad", xlab="Valores", ylab="Frecuencia")  
}



#qqnorm(M\_CENTRO[,2:length(variables\_names)])

* Análisis de correlación de los datos, mapa de calor

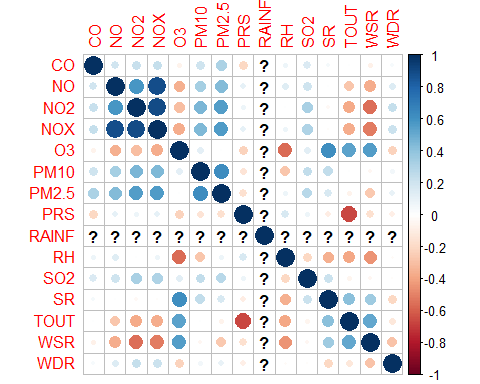
cor\_matrix = cor(M\_CENTRO[,2:length(variables\_names)])

## Warning in cor(M\_CENTRO[, 2:length(variables\_names)]): the standard deviation is  
## zero

cor\_matrix

## CO NO NO2 NOX O3  
## CO 1.000000000 0.19353971 0.221381461 0.23363881 -0.060324305  
## NO 0.193539705 1.00000000 0.580690148 0.88537926 -0.351398080  
## NO2 0.221381461 0.58069015 1.000000000 0.89258016 -0.307119487  
## NOX 0.233638815 0.88537926 0.892580156 1.00000000 -0.369923780  
## O3 -0.060324305 -0.35139808 -0.307119487 -0.36992378 1.000000000  
## PM10 0.207063641 0.33136889 0.456616861 0.44431124 0.116993654  
## PM2.5 0.317259624 0.43788977 0.558567133 0.56146053 0.002511103  
## PRS -0.204045029 0.10875113 0.076754243 0.10401368 -0.225746458  
## RAINF NA NA NA NA NA  
## RH 0.070084617 0.13053453 0.006596992 0.07609906 -0.561263150  
## SO2 0.157169515 0.20563440 0.327395845 0.30085697 0.120416545  
## SR 0.044555682 0.00986736 -0.038940546 -0.01643090 0.613314553  
## TOUT 0.004046107 -0.27839827 -0.371887216 -0.36652437 0.530126241  
## WSR -0.074799619 -0.36505226 -0.556834892 -0.51991603 0.557319641  
## WDR 0.034489860 0.14405507 0.239185499 0.21629718 -0.228699299  
## PM10 PM2.5 PRS RAINF RH SO2  
## CO 0.207063641 0.317259624 -0.20404503 NA 0.070084617 0.15716952  
## NO 0.331368888 0.437889773 0.10875113 NA 0.130534533 0.20563440  
## NO2 0.456616861 0.558567133 0.07675424 NA 0.006596992 0.32739585  
## NOX 0.444311239 0.561460534 0.10401368 NA 0.076099058 0.30085697  
## O3 0.116993654 0.002511103 -0.22574646 NA -0.561263150 0.12041655  
## PM10 1.000000000 0.621287860 -0.14847902 NA -0.275698268 0.24318405  
## PM2.5 0.621287860 1.000000000 -0.16403977 NA 0.079819492 0.27154944  
## PRS -0.148479020 -0.164039773 1.00000000 NA 0.162557657 0.07308559  
## RAINF NA NA NA 1 NA NA  
## RH -0.275698268 0.079819492 0.16255766 NA 1.000000000 -0.20790529  
## SO2 0.243184050 0.271549440 0.07308559 NA -0.207905288 1.00000000  
## SR 0.248180155 0.161793362 -0.09233537 NA -0.359562359 0.21501328  
## TOUT 0.003881171 -0.061258671 -0.66167650 NA -0.388931959 -0.05850601  
## WSR -0.046744827 -0.268046001 -0.16906494 NA -0.442403622 -0.04950906  
## WDR 0.058628751 0.099633611 -0.08722834 NA -0.022897182 -0.01038542  
## SR TOUT WSR WDR  
## CO 0.04455568 0.004046107 -0.07479962 0.03448986  
## NO 0.00986736 -0.278398274 -0.36505226 0.14405507  
## NO2 -0.03894055 -0.371887216 -0.55683489 0.23918550  
## NOX -0.01643090 -0.366524371 -0.51991603 0.21629718  
## O3 0.61331455 0.530126241 0.55731964 -0.22869930  
## PM10 0.24818016 0.003881171 -0.04674483 0.05862875  
## PM2.5 0.16179336 -0.061258671 -0.26804600 0.09963361  
## PRS -0.09233537 -0.661676498 -0.16906494 -0.08722834  
## RAINF NA NA NA NA  
## RH -0.35956236 -0.388931959 -0.44240362 -0.02289718  
## SO2 0.21501328 -0.058506014 -0.04950906 -0.01038542  
## SR 1.00000000 0.412479613 0.35248587 -0.20694581  
## TOUT 0.41247961 1.000000000 0.51221300 -0.12470607  
## WSR 0.35248587 0.512212995 1.00000000 -0.28785509  
## WDR -0.20694581 -0.124706072 -0.28785509 1.00000000

# Visualiza la matriz de correlación utilizando corrplot  
corrplot(cor\_matrix, method = "circle")

 Observamos que

1. Variables categóricas
   * Distribución de los datos (diagramas de barras, diagramas de pastel)

* En este caso nadamas tenemos una variable categórica, la cual el la variable “date”, por lo que no tiene mucho caso analizar esta misma

### Verificación de datos

1. Valores faltantes

Como ya se vió anteriormente

1. Valores de los datos
2. Ortografía

## Preparación de los datos

### 1) Selecciona el conjunto de datos a utilizar

1. Decide qué conjunto de datos se utilizará. Explica por qué se incluyeron o excluyeron ciertos datos.
2. Identifica columna objetivo

### 2) Limpieza de datos

1. Elimina duplicados
2. Corrige valores érroneos
3. Maneja valores faltantes
4. Maneja datos categóricos: Transforma a datos numéricos si es necesario.
5. Maneja adecuadamente los valores atípicos (outliers) que encuentres en el dataset.

### Transformación de Datos

1. Revisa si es necesario discretizar los datos (binning)
2. Si es necesario escala y normaliza los datos.
3. Construye atributos si es conveniente (atributos derivados).