# TRABAJO DATA VISUALIZATION

## Limpieza y analisis de la base de datos Pollution

En este trabajo vamos a realizar un analisis y lipieza de la base de datos llamada Pollution que contiene datos sobre la contaminacion de cuatro gases diferentes en cada estado de EEUU desde 2000 hasta 2016

## Limpieza

**Esquema de trabajo:**

- Estraccion de datos a traves de la lectura del csv 'pollution_us_2000_2016.csv'
- Realizacion de una copia de la informacion para mantener la integriedad de los datos originales
- Eliminar columnas que consideramos no relevantes en el analisis
- Renombrar las columnas a nombres intuitivos y en castellano
- Casting necesarios. Fecha como principal cambio
- Creacion de una nueva columna producto de otras de la tabla

**Antes de comenzar con el analisis vamos a importar las librerias necesarias:**

```
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(tidyr)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following object is masked from 'package:base':
##
##     date
```

```
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
## gdata: of a valid perl intrpreter.
## gdata:
## gdata: (To avoid display of this message in the future, please
## gdata: ensure perl is installed and available on the executable
## gdata: 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(reshape2)
```

```
##
## Attaching package: 'reshape2'
```

```
## The following object is masked from 'package:tidyr':
##
##     smiths
```

**Lectura de los datos:**

```
# Read in csv files
pollution <- read.csv("pollution_us_2000_2016.csv", stringsAsFactors = FALSE)
pollution <- read.csv("pollution_us_2000_2016.csv", header = T, sep=",")
```

# Analisis inicial de la base de datos

Aqui veremos los tipos de datos, numero y nombre de filas y columnas, dimensiones, asi como el resumen estadistico de la base de datos

```
#como la base de datos es muy grande y tarda mucho en cargar, cogemos una muestra aleatoria de l
a bbdd
muestramia <- sample(1:nrow(pollution),size=10000,replace=FALSE)

pollutionmuestramia <- pollution[muestramia, ]
head(pollutionmuestramia)
```

| | X | State.Code | County.Code | Site.Num | Address |
|---|---|---|---|---|---|
| | <int> | <int> | <int> | <int> | <fctr> |
| 1520016 | 55603 | 11 | 1 | 43 | 2500 1ST STREET, N.W. WASHINGTON DC |
| 1616485 | 20483 | 6 | 37 | 1103 | 1630 N MAIN ST, LOS ANGELES |
| 628174 | 99707 | 80 | 2 | 12 | UABC, CALZADA BENITO JUAREZ, MEXICALI |
| 802107 | 57423 | 17 | 31 | 4201 | 750 DUNDEE ROAD |
| 305662 | 31002 | 6 | 83 | 1025 | LFC #1-LAS FLORES CANYON |
| 825528 | 80844 | 37 | 119 | 41 | 1130 EASTWAY DRIVE |

6 rows | 1-6 of 30 columns

```
dim(pollutionmuestramia)         # filas x columnas
```

```
## [1] 10000    29
```

```
head(pollutionmuestramia)       #bbdd aleatoria de pollution
```

| | X | State.Code | County.Code | Site.Num | Address |
|---|---|---|---|---|---|
| | <int> | <int> | <int> | <int> | <fctr> |
| 1520016 | 55603 | 11 | 1 | 43 | 2500 1ST STREET, N.W. WASHINGTON DC |
| 1616485 | 20483 | 6 | 37 | 1103 | 1630 N MAIN ST, LOS ANGELES |
| 628174 | 99707 | 80 | 2 | 12 | UABC, CALZADA BENITO JUAREZ, MEXICALI |
| 802107 | 57423 | 17 | 31 | 4201 | 750 DUNDEE ROAD |
| 305662 | 31002 | 6 | 83 | 1025 | LFC #1-LAS FLORES CANYON |
| 825528 | 80844 | 37 | 119 | 41 | 1130 EASTWAY DRIVE |

6 rows | 1-6 of 30 columns

```
summary(pollutionmuestramia)     #resumen de la bbdd pollution
```

```
##        X            State.Code     County.Code        Site.Num
##  Min.   :    19   Min.   : 1.00   Min.   :  1.00   Min.   :   1
##  1st Qu.: 25297   1st Qu.: 6.00   1st Qu.: 17.00   1st Qu.:   9
##  Median : 53017   Median :17.00   Median : 59.00   Median :  60
##  Mean   : 54302   Mean   :22.12   Mean   : 70.51   Mean   :1112
##  3rd Qu.: 79986   3rd Qu.:40.00   3rd Qu.: 95.00   3rd Qu.:1039
##  Max.   :134264   Max.   :80.00   Max.   :650.00   Max.   :9997
##
##                          Address              State
##  PIKE AVE AT RIVER ROAD      : 215   California  :3331
##  5888 MISSION BLVD., RUBIDOUX: 170   Pennsylvania:1039
##  1130 EASTWAY DRIVE          : 157   Texas       : 727
##  1415 Hinton Street          : 156   Arizona     : 419
##  1061-A Leesville Ave        : 149   New York    : 373

##  304 TUOLUMNE ST.            : 145   Illinois    : 307
##  (Other)                     :9008   (Other)     :3804
##            County                     City           Date.Local
##  Los Angeles  : 530   Not in a city    : 796   2001-06-18:   8
##  Contra Costa : 472   New York         : 254   2013-07-21:   8
##  Santa Barbara: 465   Los Angeles      : 235   2002-08-04:   7
##  San Diego    : 322   Phoenix          : 229   2007-06-06:   7
##  Maricopa     : 282   El Paso          : 217   2009-09-11:   7
##  Harris       : 247   North Little Rock: 215   2010-11-01:   7
##  (Other)      :7682   (Other)          :8054   (Other)   :9956
##              NO2.Units        NO2.Mean       NO2.1st.Max.Value
##  Parts per billion:10000   Min.   :-0.6042   Min.   :  0.00
##                            1st Qu.: 5.7083   1st Qu.: 13.00
##                            Median :10.8229   Median : 24.00
##                            Mean   :12.7509   Mean   : 25.35
##                            3rd Qu.:17.7836   3rd Qu.: 36.00
##                            Max.   :76.4500   Max.   :137.00
##
##  NO2.1st.Max.Hour    NO2.AQI                    O3.Units
##  Min.   : 0.0     Min.   :  0.00   Parts per million:10000
##  1st Qu.: 6.0     1st Qu.: 12.00
##  Median : 9.0     Median : 23.00
##  Mean   :11.8     Mean   : 23.85
##  3rd Qu.:20.0     3rd Qu.: 34.00
##  Max.   :23.0     Max.   :108.00
##
##     O3.Mean        O3.1st.Max.Value O3.1st.Max.Hour     O3.AQI
##  Min.   :0.00000   Min.   :0.0000   Min.   : 0.00   Min.   :  0.00
##  1st Qu.:0.01775   1st Qu.:0.0290   1st Qu.: 9.00   1st Qu.: 25.00
##  Median :0.02575   Median :0.0380   Median :10.00   Median : 33.00
##  Mean   :0.02608   Mean   :0.0391   Mean   :10.18   Mean   : 35.89
##  3rd Qu.:0.03389   3rd Qu.:0.0480   3rd Qu.:11.00   3rd Qu.: 42.00
##  Max.   :0.08304   Max.   :0.1110   Max.   :23.00   Max.   :200.00
##
##              SO2.Units        SO2.Mean       SO2.1st.Max.Value
##  Parts per billion:10000   Min.   :-1.4500   Min.   : -0.700
##                            1st Qu.: 0.2609   1st Qu.:  0.900
##                            Median : 1.0000   Median :  2.000
##                            Mean   : 1.8413   Mean   :  4.371
```

```
##                          3rd Qu.: 2.2917   3rd Qu.:  5.000
##                          Max.   :27.5417   Max.   :147.000
##
##  SO2.1st.Max.Hour    SO2.AQI                     CO.Units
##  Min.   : 0.000   Min.   :  0.000   Parts per million:10000
##  1st Qu.: 4.000   1st Qu.:  1.000
##  Median : 8.000   Median :  3.000
##  Mean   : 9.564   Mean   :  6.678
##  3rd Qu.:14.000   3rd Qu.:  7.000
##  Max.   :23.000   Max.   :133.000
##                   NA's   :5017
##     CO.Mean          CO.1st.Max.Value  CO.1st.Max.Hour      CO.AQI
##  Min.   :-0.07083   Min.   : 0.0000   Min.   : 0.000   Min.   : 0.000
##  1st Qu.: 0.18696   1st Qu.: 0.2968   1st Qu.: 0.000   1st Qu.: 2.000
##  Median : 0.29167   Median : 0.4000   Median : 6.000   Median : 5.000
##  Mean   : 0.36778   Mean   : 0.6192   Mean   : 7.888   Mean   : 5.972
##  3rd Qu.: 0.46667   3rd Qu.: 0.8000   3rd Qu.:13.000   3rd Qu.: 8.000
##  Max.   : 4.04583   Max.   :16.5000   Max.   :23.000   Max.   :81.000
##                                                        NA's   :5011
```

```
head(pollutionmuestramia, 10)    # primeras diez filas
```

| X | State.Code | County.Code | Site.Num | Address |
|---|---|---|---|---|
| <int> | <int> | <int> | <int> | <fctr> |
| 1520016 | 55603 | 11 | 1 | 43 | 2500 1ST STREET, N.W. WASHINGTON DC |
| 1616485 | 20483 | 6 | 37 | 1103 | 1630 N MAIN ST, LOS ANGELES |
| 628174 | 99707 | 80 | 2 | 12 | UABC, CALZADA BENITO JUAREZ, MEXICALI |
| 802107 | 57423 | 17 | 31 | 4201 | 750 DUNDEE ROAD |
| 305662 | 31002 | 6 | 83 | 1025 | LFC #1-LAS FLORES CANYON |
| 825528 | 80844 | 37 | 119 | 41 | 1130 EASTWAY DRIVE |
| 796558 | 51874 | 11 | 1 | 41 | 420 34th Street N.E.,Washington, DC 20019 |
| 549759 | 21292 | 6 | 65 | 8001 | 5888 MISSION BLVD., RUBIDOUX |
| 830717 | 86033 | 42 | 3 | 10 | CARNEGIE SCIENCE CENTER - 1 ALLEGHENY RI |
| 515538 | 72316 | 42 | 69 | 2006 | GEORGE ST TROOP AND CITY OF SCRANTON |

1-10 of 10 rows | 1-6 of 30 columns

```
tail(pollutionmuestramia, 10)    #ultimas diez filas
```

| X | State.Code | County.Code | Site.Num | Address |
|---|---|---|---|---|
| <int> | <int> | <int> | <int> | <fctr> |
| 617084 | 88617 | 48 | 201 | 416 | 7421 Park Place Blvd |

| | X | State.Code | County.Code | Site.Num | Address |
|---|---|---|---|---|---|
| | <int> | <int> | <int> | <int> | <fctr> |
| 476940 | 33718 | 6 | 83 | 2004 | 128 S 'H' ST, LOMPOC |
| 1351314 | 21477 | 6 | 37 | 1002 | 228 W. PALM AVE., BURBANK |
| 678286 | 48980 | 11 | 1 | 41 | 420 34th Street N.E.,Washington, DC 20019 |
| 960327 | 96094 | 48 | 113 | 69 | 1415 Hinton Street |
| 1017294 | 48886 | 11 | 1 | 41 | 420 34th Street N.E.,Washington, DC 20019 |
| 1704860 | 108858 | 44 | 7 | 1010 | FRANCIS SCHOOL, 64 BOURNE AVE |
| 951966 | 87733 | 42 | 101 | 4 | 1501 E. LYCOMING AVE. |
| 29070 | 29069 | 6 | 83 | 1025 | LFC #1-LAS FLORES CANYON |
| 1626277 | 30275 | 6 | 71 | 306 | 14306 PARK AVE., VICTORVILLE, CA |

1-10 of 10 rows | 1-6 of 30 columns

```
class(pollutionmuestramia)   #tipo
```

```
## [1] "data.frame"
```

```
nrow(pollutionmuestramia)    # número de filas
```

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

```
ncol(pollutionmuestramia)    # número de columnas
```

```
## [1] 29
```

# Transformacion de los datos a DataFrame

```
#convertimos a dataframe
pollutionmuestramia <- as.data.frame(pollutionmuestramia)
#como vemos con str el tipo de cada variables esta bien no hay que transformar
str(pollutionmuestramia)     #las filas son oobservaciones y columnas las variables
```

```
## 'data.frame':    10000 obs. of  29 variables:
##  $ X                : int  55603 20483 99707 57423 31002 80844 51874 21292 86033 72316 ...
##  $ State.Code       : int  11 6 80 17 6 37 11 6 42 42 ...
##  $ County.Code      : int  1 37 2 31 83 119 1 65 3 69 ...
##  $ Site.Num         : int  43 1103 12 4201 1025 41 41 8001 10 2006 ...
##  $ Address          : Factor w/ 204 levels "          6100 ARLINGTON BLVD MONTG WARD",..: 61
36 199 125 170 19 95 112 145 160 ...
##  $ State            : Factor w/ 47 levels "Alabama","Alaska",..: 10 5 8 15 5 32 10 5 37 37
...
##  $ County           : Factor w/ 133 levels "Ada","Adair",..: 36 73 10 28 111 79 36 101 6 67
...
##  $ City             : Factor w/ 144 levels "Albuquerque",..: 136 79 83 92 26 28 136 112 101 1
20 ...
##  $ Date.Local       : Factor w/ 5996 levels "2000-01-01","2000-01-02",..: 5446 5656 2236 3146
1209 3154 3154 2396 3021 2035 ...
##  $ NO2.Units        : Factor w/ 1 level "Parts per billion": 1 1 1 1 1 1 1 1 1 1 ...
##  $ NO2.Mean         : num  8.57 15 38.5 8.25 1.26 ...
##  $ NO2.1st.Max.Value: num  29.1 28.9 69 18 2 31 43 35 35 16 ...
##  $ NO2.1st.Max.Hour : int  23 11 20 2 8 7 5 3 5 0 ...
##  $ NO2.AQI          : int  27 26 67 17 2 29 41 33 33 15 ...
##  $ O3.Units         : Factor w/ 1 level "Parts per million": 1 1 1 1 1 1 1 1 1 1 ...
##  $ O3.Mean          : num  0.0176 0.0255 0.012 0.0126 0.0484 ...
##  $ O3.1st.Max.Value : num  0.026 0.046 0.032 0.024 0.052 0.077 0.066 0.09 0.048 0.05 ...
##  $ O3.1st.Max.Hour  : int  8 10 9 9 10 10 10 10 12 9 ...
##  $ O3.AQI           : int  24 43 27 20 44 104 71 137 41 42 ...
##  $ SO2.Units        : Factor w/ 1 level "Parts per billion": 1 1 1 1 1 1 1 1 1 1 ...
##  $ SO2.Mean         : num  0.5375 -0.0261 9.9565 0 0 ...
##  $ SO2.1st.Max.Value: num  1.8 0.3 42 0 0 4.6 7.6 1.3 4 11 ...
##  $ SO2.1st.Max.Hour : int  12 11 22 0 5 9 8 14 11 8 ...
##  $ SO2.AQI          : num  1 0 59 0 NA 6 NA NA NA NA ...
##  $ CO.Units         : Factor w/ 1 level "Parts per million": 1 1 1 1 1 1 1 1 1 1 ...
##  $ CO.Mean          : num  0.229 0.4 3.791 0.163 0.3 ...
##  $ CO.1st.Max.Value : num  0.506 0.5 16.5 0.39 0.3 0.6 1 0.5 0.7 0.1 ...
##  $ CO.1st.Max.Hour  : int  23 8 22 4 0 0 6 5 6 7 ...
##  $ CO.AQI           : num  NA 6 NA NA NA 7 11 6 NA 1 ...
```

**Una vez tenemos analizada la vista inicial dela tabla vamos a comenzar con la limpieza. En este caso lo realizaremos sobre columnas. Viendo el nombre de las mismas, eliminando las sobrantes, cambiando el nombre de todas ellas y por ultimo seprando la fecha creando tres diferentes con los datos del dia, el mes y el año**

```
names(pollutionmuestramia)     #nombre de las columnas
```

```
##  [1] "X"                 "State.Code"        "County.Code"
##  [4] "Site.Num"          "Address"           "State"
##  [7] "County"            "City"              "Date.Local"
## [10] "NO2.Units"         "NO2.Mean"          "NO2.1st.Max.Value"
## [13] "NO2.1st.Max.Hour"  "NO2.AQI"           "O3.Units"
## [16] "O3.Mean"           "O3.1st.Max.Value"  "O3.1st.Max.Hour"
## [19] "O3.AQI"            "SO2.Units"         "SO2.Mean"
## [22] "SO2.1st.Max.Value" "SO2.1st.Max.Hour"  "SO2.AQI"
## [25] "CO.Units"          "CO.Mean"           "CO.1st.Max.Value"
## [28] "CO.1st.Max.Hour"   "CO.AQI"
```

```
colnames(pollutionmuestramia)  #nombre de las columnas
```

```
##  [1] "X"                 "State.Code"        "County.Code"
##  [4] "Site.Num"          "Address"           "State"
##  [7] "County"            "City"              "Date.Local"
## [10] "NO2.Units"         "NO2.Mean"          "NO2.1st.Max.Value"
## [13] "NO2.1st.Max.Hour"  "NO2.AQI"           "O3.Units"
## [16] "O3.Mean"           "O3.1st.Max.Value"  "O3.1st.Max.Hour"
## [19] "O3.AQI"            "SO2.Units"         "SO2.Mean"
## [22] "SO2.1st.Max.Value" "SO2.1st.Max.Hour"  "SO2.AQI"
## [25] "CO.Units"          "CO.Mean"           "CO.1st.Max.Value"
## [28] "CO.1st.Max.Hour"   "CO.AQI"
```

```
#pasamos los nombres de las variables a minusculas
names(pollutionmuestramia) <- tolower(names(pollutionmuestramia))

#separamos la fecha en la columna date local, y la divido en 3 columnas distintas: en año, mes y
dia, ahora tenemos 31 variables en vez de 29
pollutionmuestramia <- separate(pollutionmuestramia, date.local, c("year", "month", "day"))

pollutionmuestramia$x <- NULL   #asigno NULL al valor del area, es decir borror la primera colum
na que no es interesante

#cambiamos el nombre de las variables
colnames(pollutionmuestramia) <- c("codigo_estado", "codigo_condado", "numero_sitio", "direccio
n", "estado", "condado", "ciudad",
                           "ano", "mes", "dia","unidades_NO2", "media_NO2", "valor_max_1st_NO2",
"hora_max_1st_NO2",
                           "AQI_NO2", "unidades_O3", "media_O3", "valor_max_1st_O3", "hora_max_1
st_O3",
                           "AQI_O3", "unidades_SO2", "media_SO2", "valor_max_1st_SO2", "hora_max
_1st_SO2",
                           "AQI_SO2", "unidades_CO", "media_CO", "valor_max_st_CO", "hora_max_1s
t_CO",
                           "AQI_CO")

#ponemos en minuscula el contenido de la variable direccion
pollutionmuestramia$direccion <- tolower(pollutionmuestramia$direccion)

#creamos una nueva columna que sea la media de la suma de los 4 gases
pollutionmuestramia$valoresmaximos <- (pollutionmuestramia$valor_max_1st_NO2 + pollutionmuestram
ia$valor_max_1st_O3 +
                                         pollutionmuestramia$valor_max_1st_SO2 + pollutionmuestr
amia$valor_max_st_CO)
```

# Filas

Al tener ya limpias las columnas podemos centrarnos en las filas. Verificaremos si existen valores nulos o NaN y los corregiremos, asi como saber las filas que tienen los datos correctos

```
#check Nas values
head(is.na(pollutionmuestramia), 1)
```

```
##         codigo_estado codigo_condado numero_sitio direccion estado condado
## 1520016         FALSE          FALSE        FALSE     FALSE  FALSE   FALSE
##         ciudad   ano   mes   dia unidades_NO2 media_NO2 valor_max_1st_NO2
## 1520016  FALSE FALSE FALSE FALSE        FALSE     FALSE             FALSE
##         hora_max_1st_NO2 AQI_NO2 unidades_O3 media_O3 valor_max_1st_O3
## 1520016            FALSE   FALSE       FALSE    FALSE            FALSE
##         hora_max_1st_O3 AQI_O3 unidades_SO2 media_SO2 valor_max_1st_SO2
## 1520016           FALSE  FALSE        FALSE     FALSE             FALSE
##         hora_max_1st_SO2 AQI_SO2 unidades_CO media_CO valor_max_st_CO
## 1520016            FALSE   FALSE       FALSE    FALSE           FALSE
##         hora_max_1st_CO AQI_CO valoresmaximos
## 1520016           FALSE   TRUE          FALSE
```

```
#preguntamos si hay alguno
head(any(is.na(pollutionmuestramia)), 1)
```

```
## [1] TRUE
```

```
#contamos el numero de Nas
sum(is.na(pollutionmuestramia))
```

```
## [1] 10028
```

```
#encontrar las filas que no falten datos
head(complete.cases(pollutionmuestramia), 30)
```

```
##  [1] FALSE  TRUE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE  TRUE
## [12] FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE FALSE FALSE
## [23]  TRUE FALSE FALSE  TRUE FALSE  TRUE FALSE  TRUE
```

```
#subset data solo con los casos completos
head(pollutionmuestramia[complete.cases(pollutionmuestramia),])
```

| | codigo_estado <int> | codigo_condado <int> | numero_sitio <int> | direccion <chr> |
|---|---|---|---|---|
| 1616485 | 6 | 37 | 1103 | 1630 n main st, los angeles |
| 825528 | 37 | 119 | 41 | 1130 eastway drive |
| 1522609 | 13 | 89 | 2 | 2390-b wildcat road, decatur ga 30034 |
| 1148310 | 25 | 25 | 42 | harrison ave |
| 786541 | 6 | 83 | 4003 | sts power plant, vandenberg afb |
| 1304264 | 42 | 49 | 3 | 10th and marne streets |

6 rows | 1-5 of 32 columns

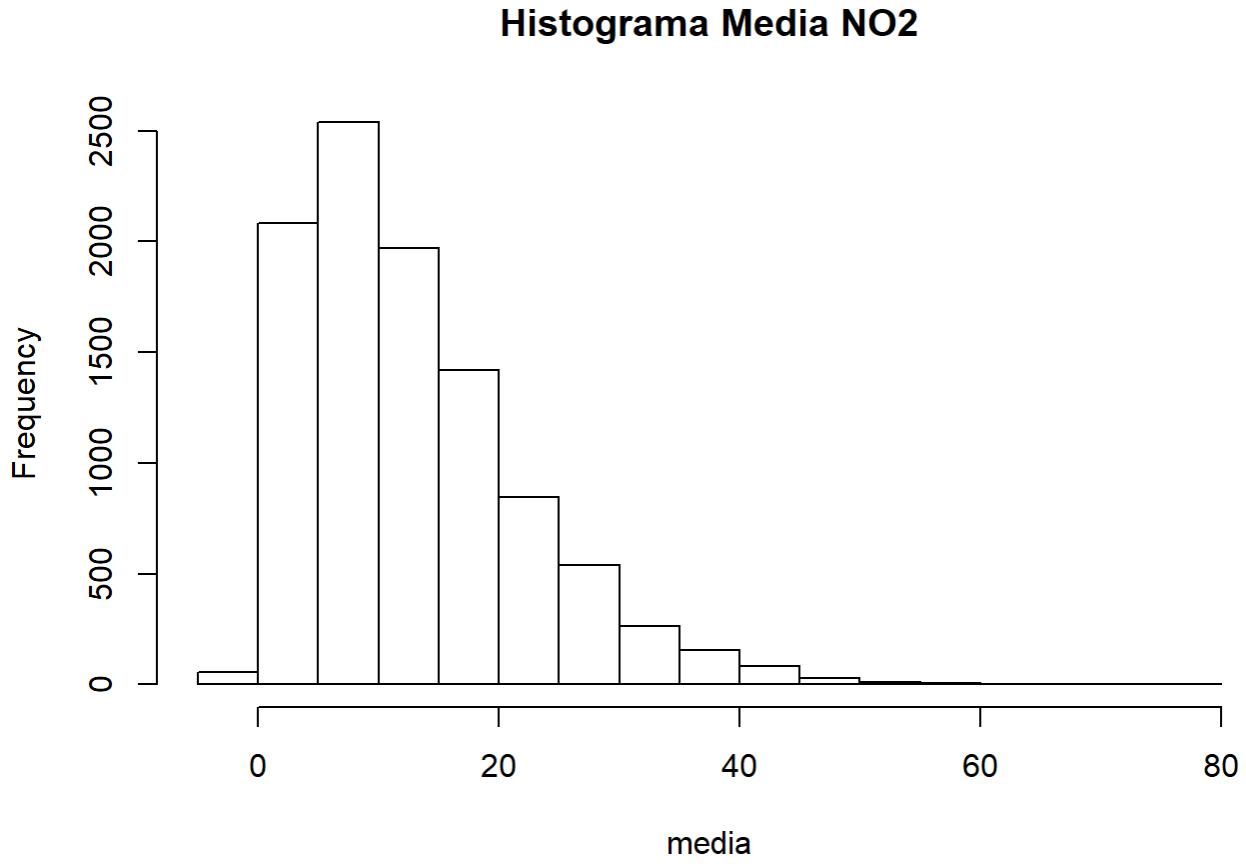
# Realizacion de los graficos

## Histogramas

Para poder ver las distribuciones de cada variable

### Histograma N02

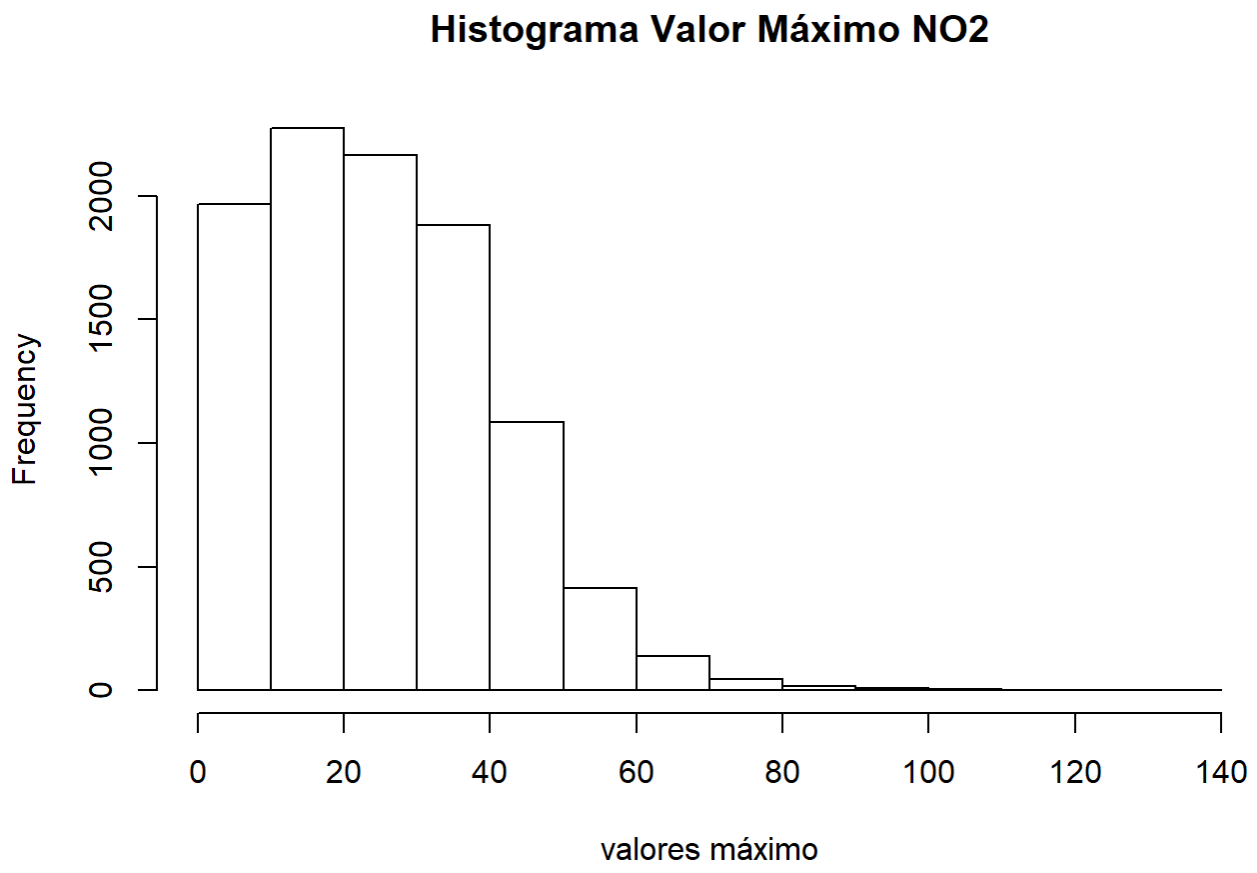
```
hist(pollutionmuestramia$media_NO2, main = "Histograma Media NO2", xlab = "media")
```




La media de NO2 parece seguir una distribución similar a la de Poisson con una cola larga a la derecha. La mayoría de los estados tienen una media similar que se encuentra entre 5 y 20, sin embargo, hay ciertos estados que salen de esa

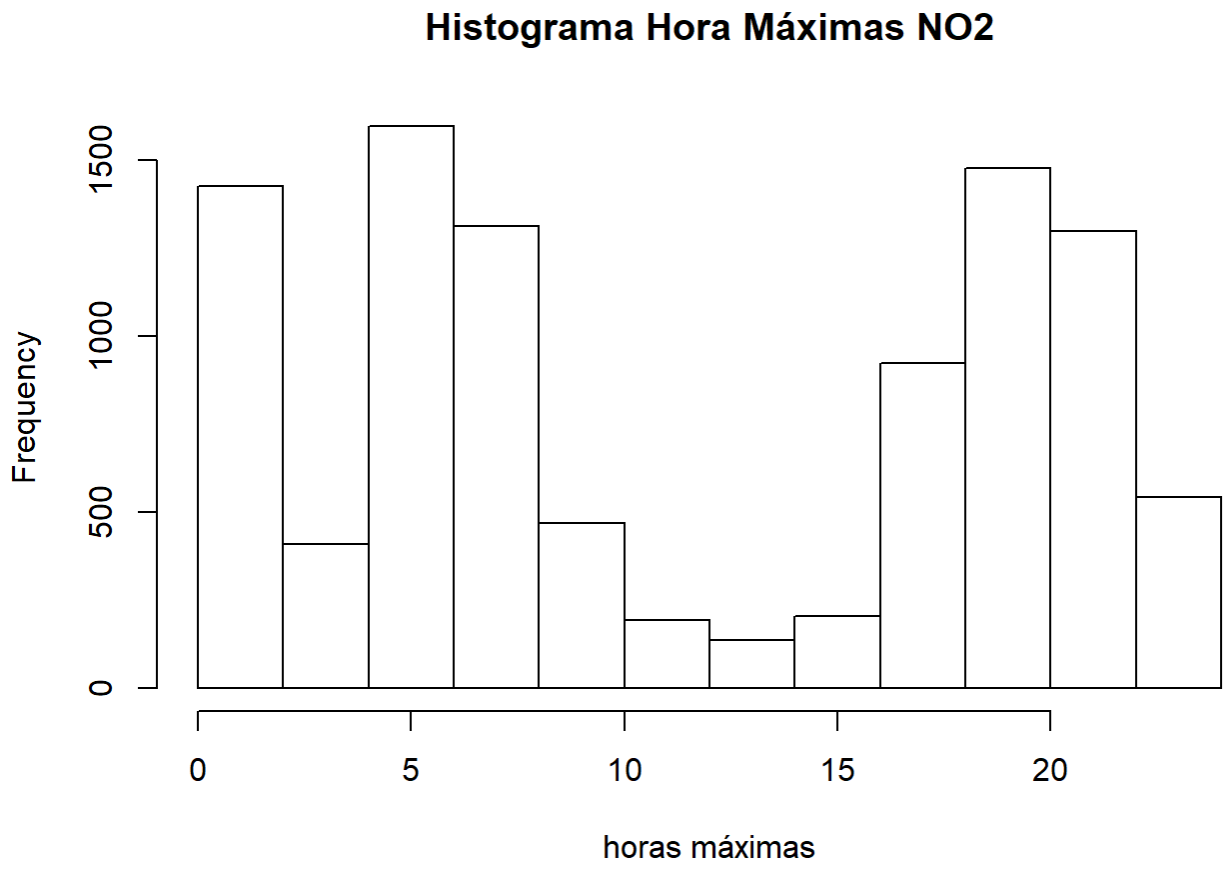
media. California o Pennsylvania son los más representativos y son los que generan, en parte, esa cola lateral alargada.

```
hist(pollutionmuestramia$valor_max_1st_NO2, main = "Histograma Valor Máximo NO2", xlab = "valore
s máximo")
```



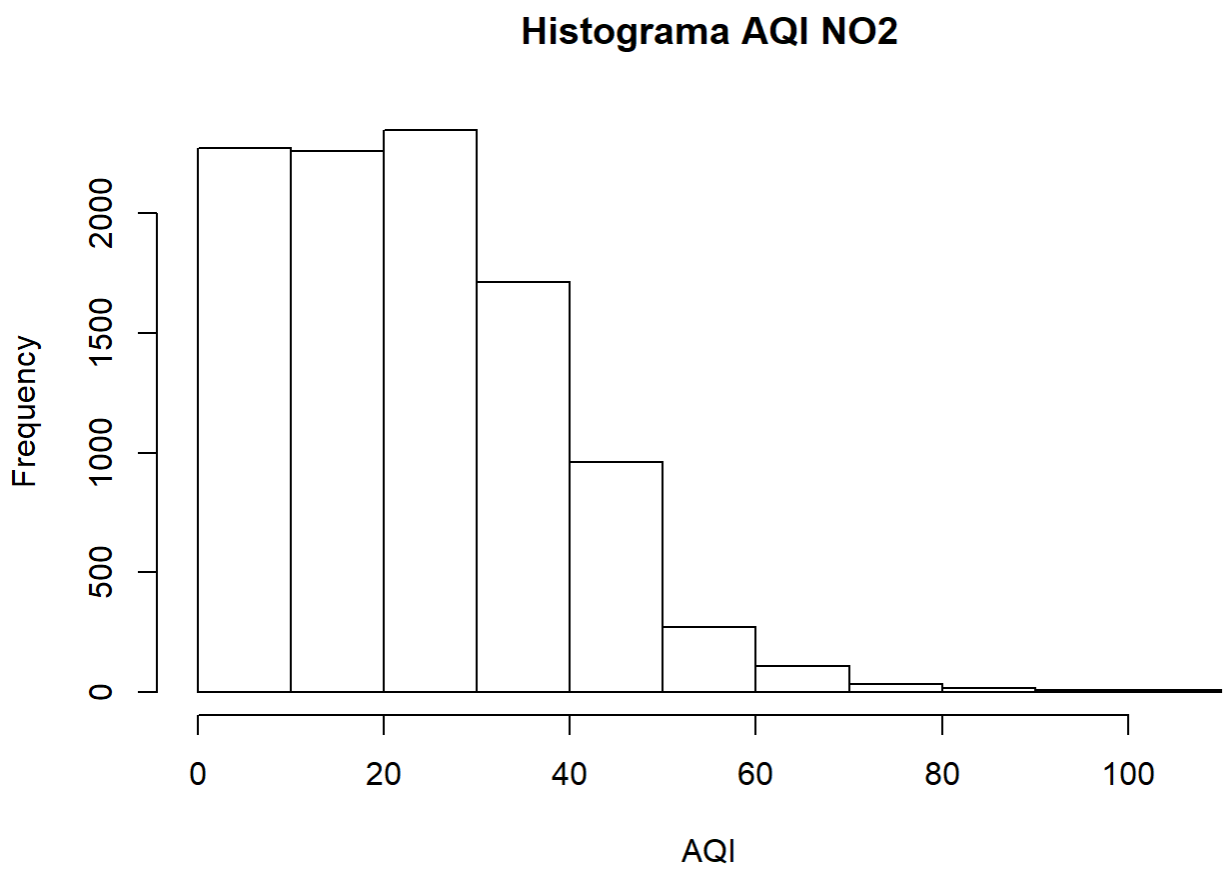

Los valores máximos se concentran en valores más pequeños en su mayoría, generando una distribución, de nuevo, similar a Poisson, aunque con una cola derecha más corta, ya que hay pocos datos que se vayan lejos de la media.

```
hist(pollutionmuestramia$hora_max_1st_NO2, main = "Histograma Hora Máximas NO2", xlab = "horas m
áximas")
```




Parece observarse una curva en forma de U en el histograma en el que hay valores altos durante la noche y tarde y muy bajos durante la mañana, hasta el mediodía. Ligera excepción de madrugada, con orígenes desconocidos. Quizás incluso una mala medición.

```
hist(pollutionmuestramia$AQI_NO2, main = "Histograma AQI NO2", xlab = "AQI")
```

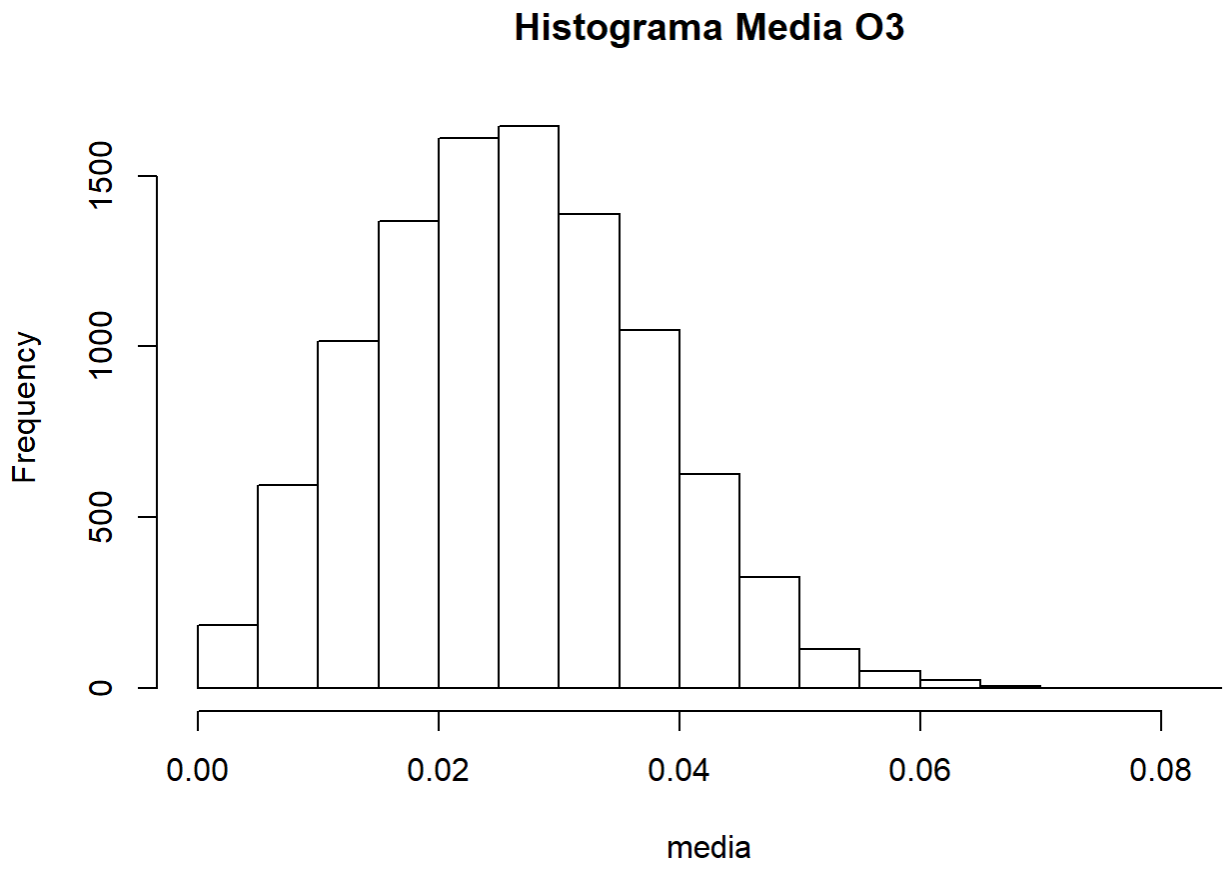



Al igual que ocurre con la media de NO2, el AQI de NO2 vuelve a parecer una distribución de Poisson muy pegada a la izquierda, con la mayoría de valores muy bajos y, por lo tanto, la media también.

En cambio, la cola derecha, a pesar de ser pequeña, es larga. Lo que significa que hay ciertos estados que tienen una gran diferencia, incluso entre ellos mismos.

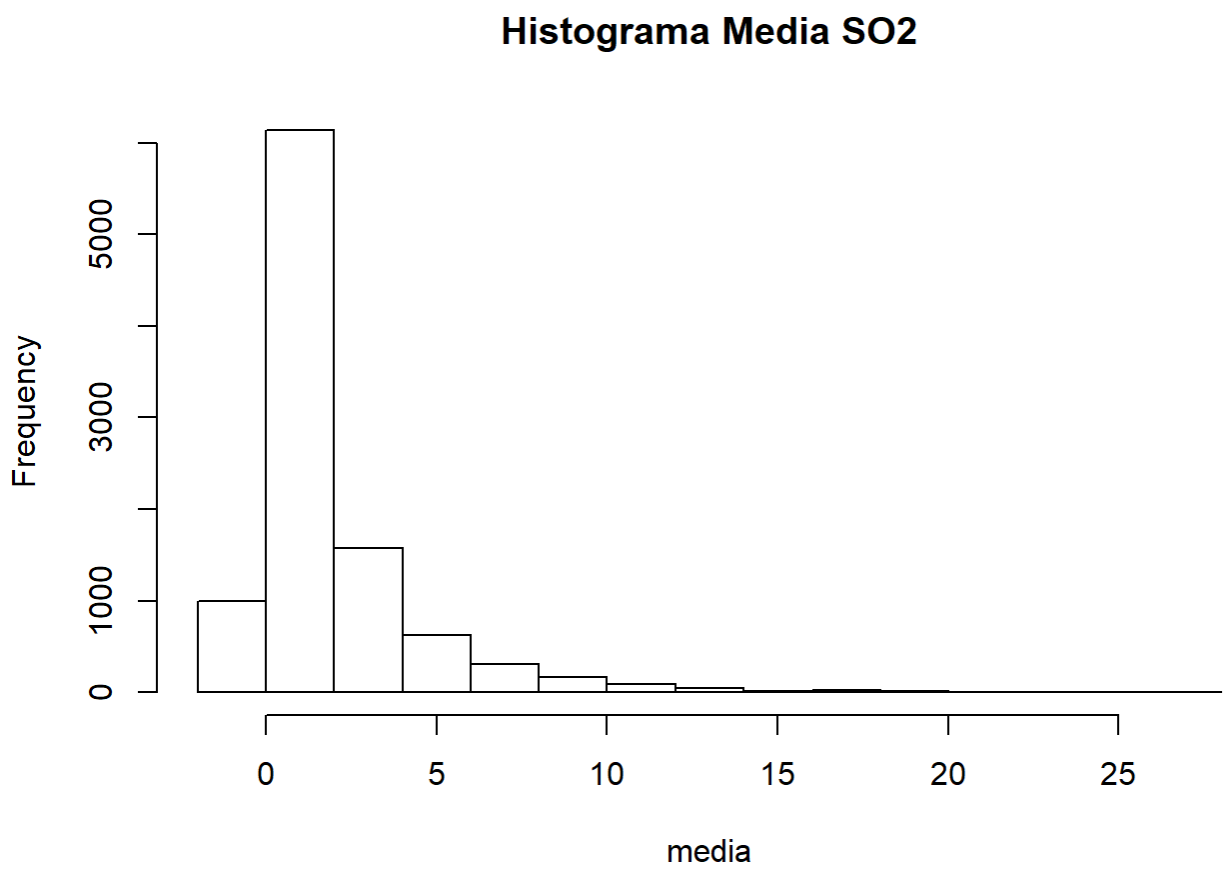
Histogramas de los datos de la media de los demas gases

```
hist(pollutionmuestramia$media_O3, main = "Histograma Media O3", xlab = "media")
```

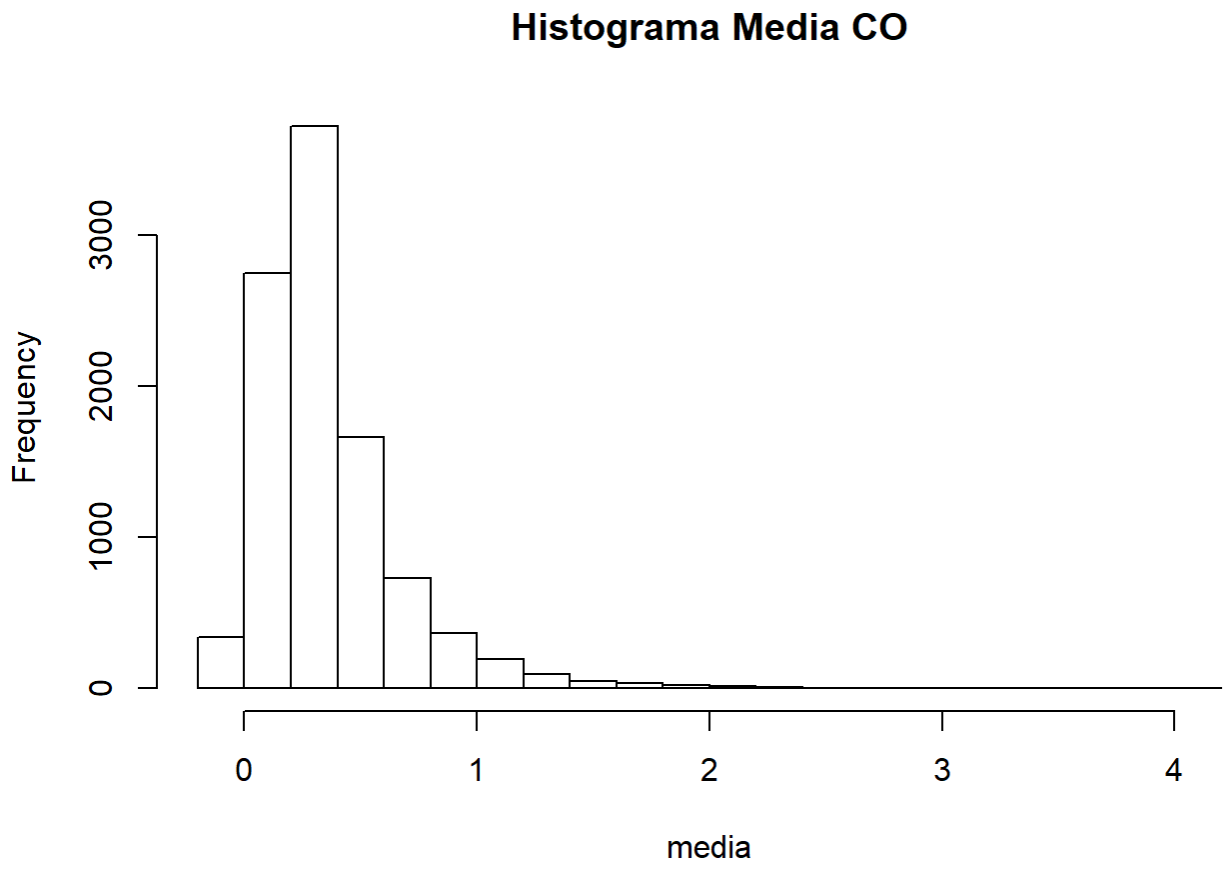



En el caso del ozono, sin embargo, la distribución se parece mas a una normal, sin embargo, existen ligeras asimetrías. Puede ser debido, sobre todo, al amplio número de estados a medir.

```
hist(pollutionmuestramia$media_SO2, main = "Histograma Media SO2", xlab = "media")
```



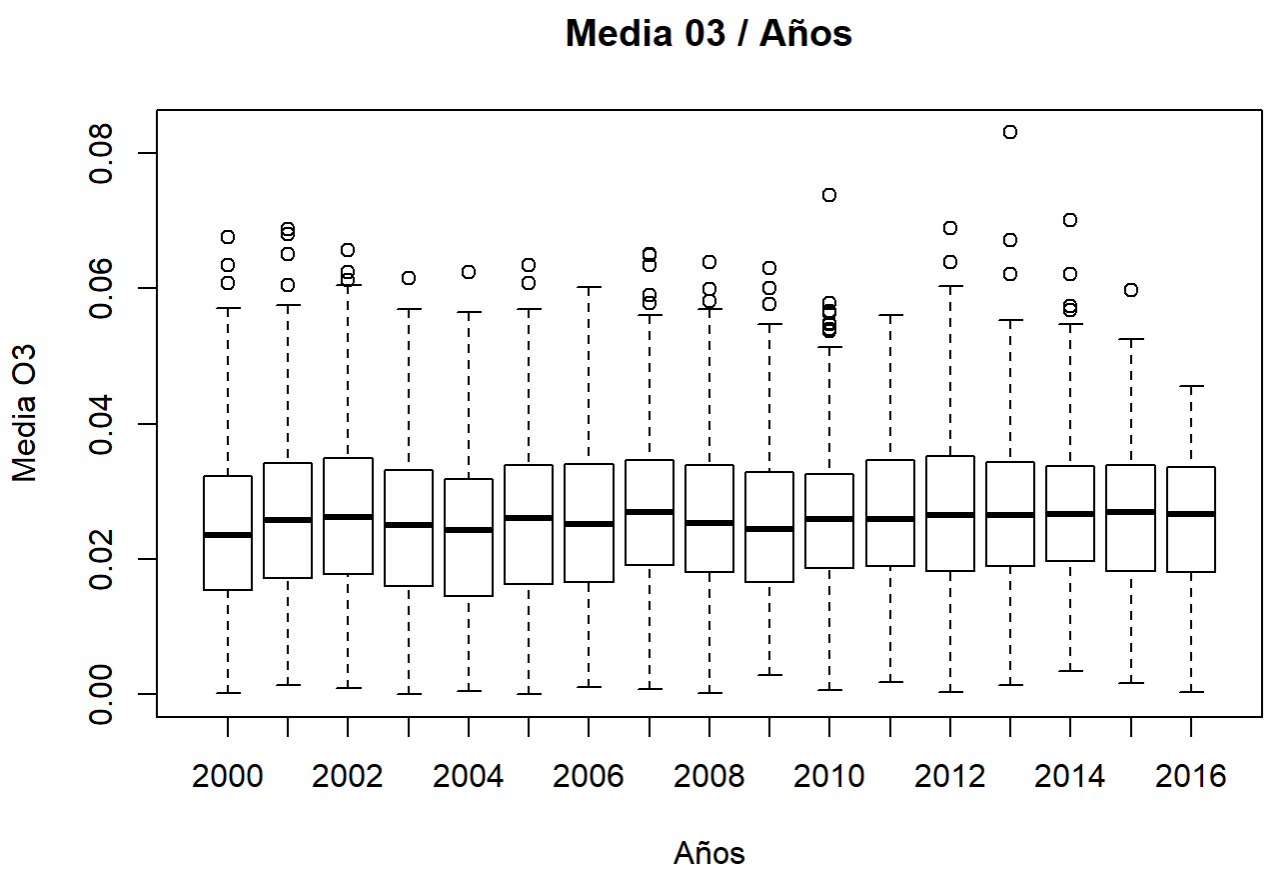

```
hist(pollutionmuestramia$media_CO, main = "Histograma Media CO", xlab = "media")
```




## En el caso del SO2 y el CO, la distribución es más pronunciada y semejante a Poisson. En este caso, la mayoría vuelven a encontrarse de media en la zona más baja, sin embargo, con un pico muy alto. Se debe probablemente a la existencia de numerosos puntos atípicos que desvirtúan en parte la distribución.

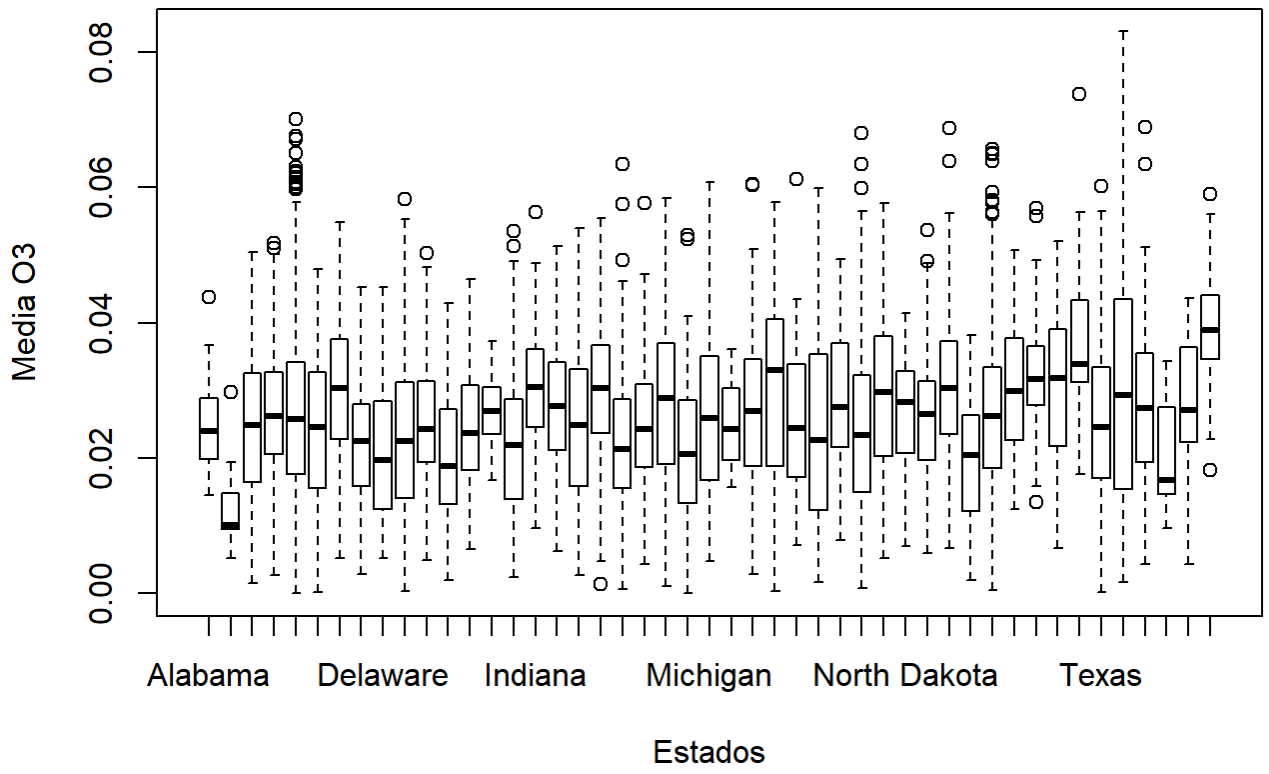
# Boxplots

```
#boxolot de los datos de Media 03 a lo largo de los años
boxplot(media_O3 ~ ano, data=pollutionmuestramia, main="Media 03 / Años", xlab="Años", ylab="Med
ia O3")
```




No parece haber muchas variaciones a lo largo de los años, al menos, respecto a la media. Sin embargo, hay años en los que determinados estados salen fuera de la misma, situándose como puntos atípicos, lo que provoca un desplazamiento de la media.

```
#boxolot de los datos de Media 03 en los diferentes estados
boxplot(media_O3 ~ estado, data=pollutionmuestramia, main="Media 03 / Estados", xlab="Estados",
 ylab="Media O3")
```




Sin embargo, sí se pueden observar diferencias significativas entre las medias de los diferentes estados, viendo cómo algunos tienen una media muy baja, media o muy alta. Aunque sí que existe una tendencia de medias a estar entre dos valores próximos (Entre 0,2 y 0,3).