PR2 IMPLEMENTATION REPORT : CLEANING AND VALIDATION OF DATA

Daura Hernández Díaz ; Xiaowei Cai

2019/6/10

Load R libraries

```r
library(RcmdrMisc)
```

```
## Loading required package: car

## Loading required package: carData

## Loading required package: sandwich
```

```r
library(stringr)
library(Hmisc)
```

```
## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

## Loading required package: ggplot2

##
## Attaching package: 'Hmisc'

## The following object is masked from 'package:RcmdrMisc':
##
##     Dotplot

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

```r
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```r
library(MVN)
```

```
## sROC 0.1-2 loaded
```

```r
library(gvlma)
library(stargazer)
```

```
##
## Please cite as:
```

```
##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.
```

```
##  R package version 5.2.2. https://CRAN.R-project.org/package=stargazer
```

```r
library(easyGgplot2)
library(psych)
```

```
##
## Attaching package: 'psych'
```

```
## The following object is masked from 'package:Hmisc':
##
##     describe
```

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

```
## The following object is masked from 'package:car':
##
##     logit
```

```r
library(car)
library(caret)
```

```
## Warning: package 'caret' was built under R version 3.5.3
```

```
##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:survival':
##
##     cluster
```

```r
library(mgcv)
```

```
## Loading required package: nlme
```

```
## This is mgcv 1.8-25. For overview type 'help("mgcv-package")'.

##
## Attaching package: 'mgcv'

## The following object is masked from 'package:MVN':
##
##     mvn
```

```
library(lmtest)
```

```
## Warning: package 'lmtest' was built under R version 3.5.3

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
library(rpart)
```

```
##
## Attaching package: 'rpart'

## The following object is masked from 'package:survival':
##
##     solder
```

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

Load data and clean data

```
#Load the datasets Casa Palacio-Puerto del Rosario2011.
Fuerteventura_2011_1 <- read.csv("Casa Palacio-Puerto del Rosario20111.csv",encoding="UTF-8")
Fuerteventura_2011_2 <- read.csv("Casa Palacio-Puerto del Rosario20112.csv",encoding="UTF-8")
Fuerteventura_2011_3 <- read.csv("Casa Palacio-Puerto del Rosario20113.csv",encoding="UTF-8")
Fuerteventura_2011_4 <- read.csv("Casa Palacio-Puerto del Rosario20114.csv",encoding="UTF-8")
#Combine the datasets.
Fuerteventura_2011 <- cbind(cbind(Fuerteventura_2011_1,Fuerteventura_2011_2),cbind(Fuerteventura_2011_3
#Create a new variable to assign the location of weather station.
Fuerteventura_2011$YEAR <- "2011"

#Load the datasets Casa Palacio-Puerto del Rosario2012.
Fuerteventura_2012_1 <- read.csv("Casa Palacio-Puerto del Rosario20121.csv",encoding="UTF-8")
```

```
Fuerteventura_2012_2 <- read.csv("Casa Palacio-Puerto del Rosario20122.csv",encoding="UTF-8")
Fuerteventura_2012_3 <- read.csv("Casa Palacio-Puerto del Rosario20123.csv",encoding="UTF-8")
Fuerteventura_2012_4 <- read.csv("Casa Palacio-Puerto del Rosario20124.csv",encoding="UTF-8")
#Combine the datasets.
Fuerteventura_2012 <- cbind(cbind(Fuerteventura_2012_1,Fuerteventura_2012_2),cbind(Fuerteventura_2012_3
#Create a new variable to assign the location of weather station.
Fuerteventura_2012$YEAR <- "2012"


#Load the datasets Casa Palacio-Puerto del Rosario2013.
Fuerteventura_2013_1 <- read.csv("Casa Palacio-Puerto del Rosario20131.csv",encoding="UTF-8")
Fuerteventura_2013_2 <- read.csv("Casa Palacio-Puerto del Rosario20132.csv",encoding="UTF-8")
Fuerteventura_2013_3 <- read.csv("Casa Palacio-Puerto del Rosario20133.csv",encoding="UTF-8")
Fuerteventura_2013_4 <- read.csv("Casa Palacio-Puerto del Rosario20134.csv",encoding="UTF-8")
#Combine the datasets.
Fuerteventura_2013 <- cbind(cbind(Fuerteventura_2013_1,Fuerteventura_2013_2),cbind(Fuerteventura_2013_3
#Create a new variable to assign the location of weather station.
Fuerteventura_2013$YEAR <- "2013"


#Load the datasets Casa Palacio-Puerto del Rosario2014.
Fuerteventura_2014_1 <- read.csv("Casa Palacio-Puerto del Rosario20141.csv",encoding="UTF-8")
Fuerteventura_2014_2 <- read.csv("Casa Palacio-Puerto del Rosario20142.csv",encoding="UTF-8")
Fuerteventura_2014_3 <- read.csv("Casa Palacio-Puerto del Rosario20143.csv",encoding="UTF-8")
Fuerteventura_2014_4 <- read.csv("Casa Palacio-Puerto del Rosario20144.csv",encoding="UTF-8")
#Combine the datasets.
Fuerteventura_2014 <- cbind(cbind(Fuerteventura_2014_1,Fuerteventura_2014_2),cbind(Fuerteventura_2014_3
#Create a new variable to assign the location of weather station.
Fuerteventura_2014$YEAR <- "2014"


#Load the datasets Casa Palacio-Puerto del Rosario2015.
Fuerteventura_2015_1 <- read.csv("Casa Palacio-Puerto del Rosario20151.csv",encoding="UTF-8")
Fuerteventura_2015_2 <- read.csv("Casa Palacio-Puerto del Rosario20152.csv",encoding="UTF-8")
Fuerteventura_2015_3 <- read.csv("Casa Palacio-Puerto del Rosario20153.csv",encoding="UTF-8")
Fuerteventura_2015_4 <- read.csv("Casa Palacio-Puerto del Rosario20154.csv",encoding="UTF-8")
#Combine the datasets.
Fuerteventura_2015 <- cbind(cbind(Fuerteventura_2015_1,Fuerteventura_2015_2),cbind(Fuerteventura_2015_3
#Create a new variable to assign the location of weather station.
Fuerteventura_2015$YEAR <- "2015"


#Load the datasets Casa Palacio-Puerto del Rosario2016.
Fuerteventura_2016_1 <- read.csv("Casa Palacio-Puerto del Rosario20161.csv",encoding="UTF-8")
Fuerteventura_2016_2 <- read.csv("Casa Palacio-Puerto del Rosario20162.csv",encoding="UTF-8")
Fuerteventura_2016_3 <- read.csv("Casa Palacio-Puerto del Rosario20163.csv",encoding="UTF-8")
Fuerteventura_2016_4 <- read.csv("Casa Palacio-Puerto del Rosario20164.csv",encoding="UTF-8")
#Combine the datasets.
Fuerteventura_2016 <- cbind(cbind(Fuerteventura_2016_1,Fuerteventura_2016_2),cbind(Fuerteventura_2016_3
#Create a new variable to assign the location of weather station.
Fuerteventura_2016$YEAR <- "2016"
```

```r
#Load the datasets Casa Palacio-Puerto del Rosario2017.
Fuerteventura_2017_1 <- read.csv("Casa Palacio-Puerto del Rosario20171.csv",encoding="UTF-8")
Fuerteventura_2017_2 <- read.csv("Casa Palacio-Puerto del Rosario20172.csv",encoding="UTF-8")
Fuerteventura_2017_3 <- read.csv("Casa Palacio-Puerto del Rosario20173.csv",encoding="UTF-8")
Fuerteventura_2017_4 <- read.csv("Casa Palacio-Puerto del Rosario20174.csv",encoding="UTF-8")
#Combine the datasets.
Fuerteventura_2017 <- cbind(cbind(Fuerteventura_2017_1,Fuerteventura_2017_2),cbind(Fuerteventura_2017_3
#Create a new variable to assign the location of weather station.
Fuerteventura_2017$YEAR <- "2017"


#Load the datasets Casa Palacio-Puerto del Rosario2018.
Fuerteventura_2018_1 <- read.csv("Casa Palacio-Puerto del Rosario20181.csv",encoding="UTF-8")
Fuerteventura_2018_2 <- read.csv("Casa Palacio-Puerto del Rosario20182.csv",encoding="UTF-8")
Fuerteventura_2018_3 <- read.csv("Casa Palacio-Puerto del Rosario20183.csv",encoding="UTF-8")
Fuerteventura_2018_4 <- read.csv("Casa Palacio-Puerto del Rosario20184.csv",encoding="UTF-8")
#Combine the datasets.
Fuerteventura_2018 <- cbind(cbind(Fuerteventura_2018_1,Fuerteventura_2018_2),cbind(Fuerteventura_2018_3
#Create a new variable to assign the location of weather station.
Fuerteventura_2018$YEAR <- "2018"



clean_data <- function(dataset){


  #Rename some variables
  dataset$DATE <- dataset$Fecha
  dataset$SO2 <- dataset[ , which(str_detect(names(dataset), pattern = "SO2"))]
  dataset$NO <- dataset[ , which(str_detect(names(dataset), pattern = "NO.. g.m3."))]
  dataset$NO2 <- dataset[ , which(str_detect(names(dataset), pattern = "NO2.. g.m3."))]
  dataset$NOX <- dataset[ , which(str_detect(names(dataset), pattern = "NOX.. g.m3."))]
  dataset$PM10 <- dataset[ , which(str_detect(names(dataset), pattern = "PM10"))]
  dataset$CO <- dataset[ , which(str_detect(names(dataset), pattern = "CO"))]
  dataset$PM2.5 <- dataset[ , which(str_detect(names(dataset), pattern = "PM2.5"))]
  dataset$O3 <- dataset[ , which(str_detect(names(dataset), pattern = "O3"))]
  dataset$VV <- dataset[ , which(str_detect(names(dataset), pattern = "VV"))]
  dataset$DD <- dataset[ , which(str_detect(names(dataset), pattern = "DD"))]
  dataset$TMP <- dataset[ , which(str_detect(names(dataset), pattern = "TMP"))]
  dataset$HR <- dataset[ , which(str_detect(names(dataset), pattern = "HR"))]
  dataset$PRB <- dataset[ , which(str_detect(names(dataset), pattern = "PRB"))]

  #Rename some columns
  dataset$DATE <- dataset$Fecha
  dataset$HOUR <- dataset$Hora

  #Include the necessary variables into the dataset.
  dataset <-  dataset[,c("DATE","HOUR","YEAR","SO2","NO","NO2","NOX","O3","CO","PM10","PM2.5","VV","DD"



  return(dataset)
}
```

```r
#Conduct the data cleaning process for each of the datasets.
Fuerteventura_2011 <- clean_data(Fuerteventura_2011)
Fuerteventura_2012 <- clean_data(Fuerteventura_2012)
Fuerteventura_2013 <- clean_data(Fuerteventura_2013)
Fuerteventura_2014 <- clean_data(Fuerteventura_2014)
Fuerteventura_2015 <- clean_data(Fuerteventura_2015)
Fuerteventura_2016 <- clean_data(Fuerteventura_2016)
Fuerteventura_2017 <- clean_data(Fuerteventura_2017)
Fuerteventura_2018 <- clean_data(Fuerteventura_2018)


#Bind the observations from different datasets in one unique dataset.
provisional1 <- rbind(Fuerteventura_2011,Fuerteventura_2012)
provisional2 <- rbind(Fuerteventura_2013,Fuerteventura_2014)
provisional3 <- rbind(Fuerteventura_2015,Fuerteventura_2016)
provisional4 <- rbind(Fuerteventura_2017,Fuerteventura_2018)
provisional5 <- rbind(provisional1,provisional2)
provisional6 <- rbind(provisional3,provisional4)
Canarydataset <-rbind(provisional5,provisional6)

#Configure levels for the variable "YEAR".
Canarydataset$YEAR <- factor(Canarydataset$YEAR,levels = c("2011","2012","2013","2014","2015","2016","20

#Create a new variable "MONTH".
Canarydataset$MONTH <- ifelse(str_detect(Canarydataset$DATE, pattern = "-01-"),"January",
                              ifelse(str_detect(Canarydataset$DATE, pattern = "-02-"),"February",
                                     ifelse(str_detect(Canarydataset$DATE, pattern = "-03-"),"March",
                                            ifelse(str_detect(Canarydataset$DATE, pattern = "-04-"),"Ap
                                                   ifelse(str_detect(Canarydataset$DATE, pattern = "-05
                                                          ifelse(str_detect(Canarydataset$DATE, pattern
                                                                 ifelse(str_detect(Canarydataset$DATE,
                                                                        ifelse(str_detect(Canarydataset
                                                                               ifelse(str_detect(Canary
                                                                                      ifelse(str_detect
                                                                                             ifelse(str


#Configure levels for the variable "MONTH".
Canarydataset$MONTH <- factor(Canarydataset$MONTH,levels = c("January","February","March","April","May"

#Detect outliers and replace them by NA.
outlier_SO2 <- which(Canarydataset$SO2 %in% boxplot.stats(Canarydataset$SO2)$out)
Canarydataset[outlier_SO2,"SO2"] <- NA

outlier_NO <- which(Canarydataset$NO %in% boxplot.stats(Canarydataset$NO)$out)
Canarydataset[outlier_NO,"NO"] <- NA

outlier_NO2 <- which(Canarydataset$NO2 %in% boxplot.stats(Canarydataset$NO2)$out)
Canarydataset[outlier_NO2,"NO2"] <- NA

outlier_PM10 <- which(Canarydataset$PM10 %in% boxplot.stats(Canarydataset$PM10)$out)
Canarydataset[outlier_PM10,"PM10"] <- NA
```

```r
outlier_NOX <- which(Canarydataset$NOX %in% boxplot.stats(Canarydataset$NOX)$out)
Canarydataset[outlier_NOX,"NOX"] <- NA

outlier_CO <- which(Canarydataset$CO %in% boxplot.stats(Canarydataset$CO)$out)
Canarydataset[outlier_CO,"CO"] <- NA

outlier_PM2.5 <- which(Canarydataset$PM2.5 %in% boxplot.stats(Canarydataset$PM2.5)$out)
Canarydataset[outlier_PM2.5,"PM2.5"] <- NA

outlier_O3 <- which(Canarydataset$O3 %in% boxplot.stats(Canarydataset$O3)$out)
Canarydataset[outlier_O3,"O3"] <- NA

outlier_VV <- which(Canarydataset$VV %in% boxplot.stats(Canarydataset$VV)$out)
Canarydataset[outlier_VV,"VV"] <- NA

outlier_DD <- which(Canarydataset$DD %in% boxplot.stats(Canarydataset$DD)$out)
Canarydataset[outlier_DD,"DD"] <- NA

outlier_TMP <- which(Canarydataset$TMP %in% boxplot.stats(Canarydataset$TMP)$out)
Canarydataset[outlier_TMP,"TMP"] <- NA

outlier_HR <- which(Canarydataset$HR %in% boxplot.stats(Canarydataset$HR)$out)
Canarydataset[outlier_HR,"HR"] <- NA

outlier_PRB <- which(Canarydataset$PRB %in% boxplot.stats(Canarydataset$PRB)$out)
Canarydataset[outlier_PRB,"PRB"] <- NA

#Replace the missing values with the mean values in each of the numerical variables.
Canarydataset$SO2[is.na(Canarydataset$SO2)] <- mean(Canarydataset$SO2,na.rm=T)
Canarydataset$NO[is.na(Canarydataset$NO)] <- mean(Canarydataset$NO,na.rm=T)
Canarydataset$NO2[is.na(Canarydataset$NO2)] <- mean(Canarydataset$NO2,na.rm=T)
Canarydataset$PM10[is.na(Canarydataset$PM10)] <- mean(Canarydataset$PM10,na.rm=T)
Canarydataset$NOX[is.na(Canarydataset$NOX)] <- mean(Canarydataset$NOX,na.rm=T)
Canarydataset$CO[is.na(Canarydataset$CO)] <- mean(Canarydataset$CO,na.rm=T)
Canarydataset$PM2.5[is.na(Canarydataset$PM2.5)] <- mean(Canarydataset$PM2.5,na.rm=T)
Canarydataset$O3[is.na(Canarydataset$O3)] <- mean(Canarydataset$O3,na.rm=T)
Canarydataset$VV[is.na(Canarydataset$VV)] <- mean(Canarydataset$VV,na.rm=T)
Canarydataset$DD[is.na(Canarydataset$DD)] <- mean(Canarydataset$DD,na.rm=T)
Canarydataset$TMP[is.na(Canarydataset$TMP)] <- mean(Canarydataset$TMP,na.rm=T)
Canarydataset$HR[is.na(Canarydataset$HR)] <- mean(Canarydataset$HR,na.rm=T)
Canarydataset$PRB[is.na(Canarydataset$PRB)] <- mean(Canarydataset$PRB,na.rm=T)
```

Descriptive data analysis

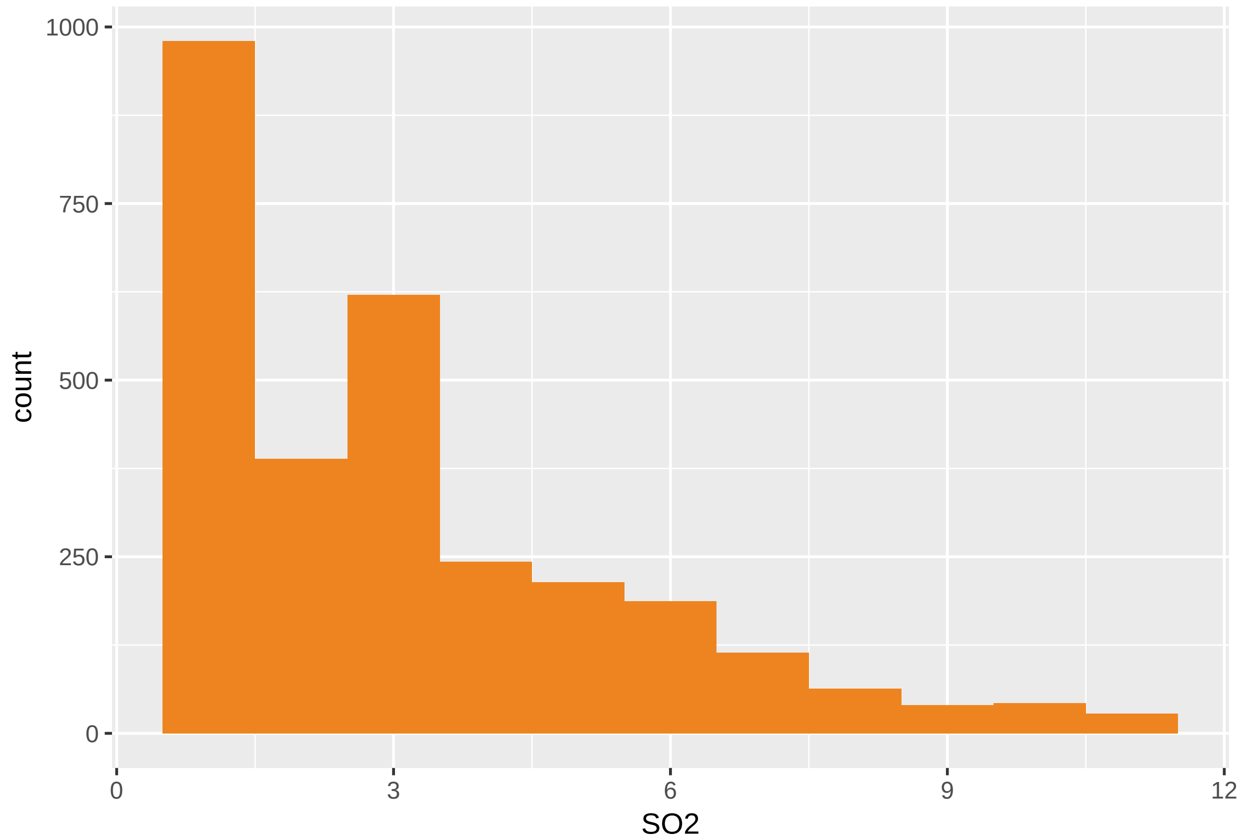
SO2

```r
#Concentration of SO2 -  g / m³
summary(Canarydataset$SO2)
```

```
##     Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
##    1.000    1.000    3.000    3.161    4.000   11.000
```

```
ggplot(Canarydataset, aes(x=SO2))+geom_histogram(binwidth=1,fill="#EE8420")
```

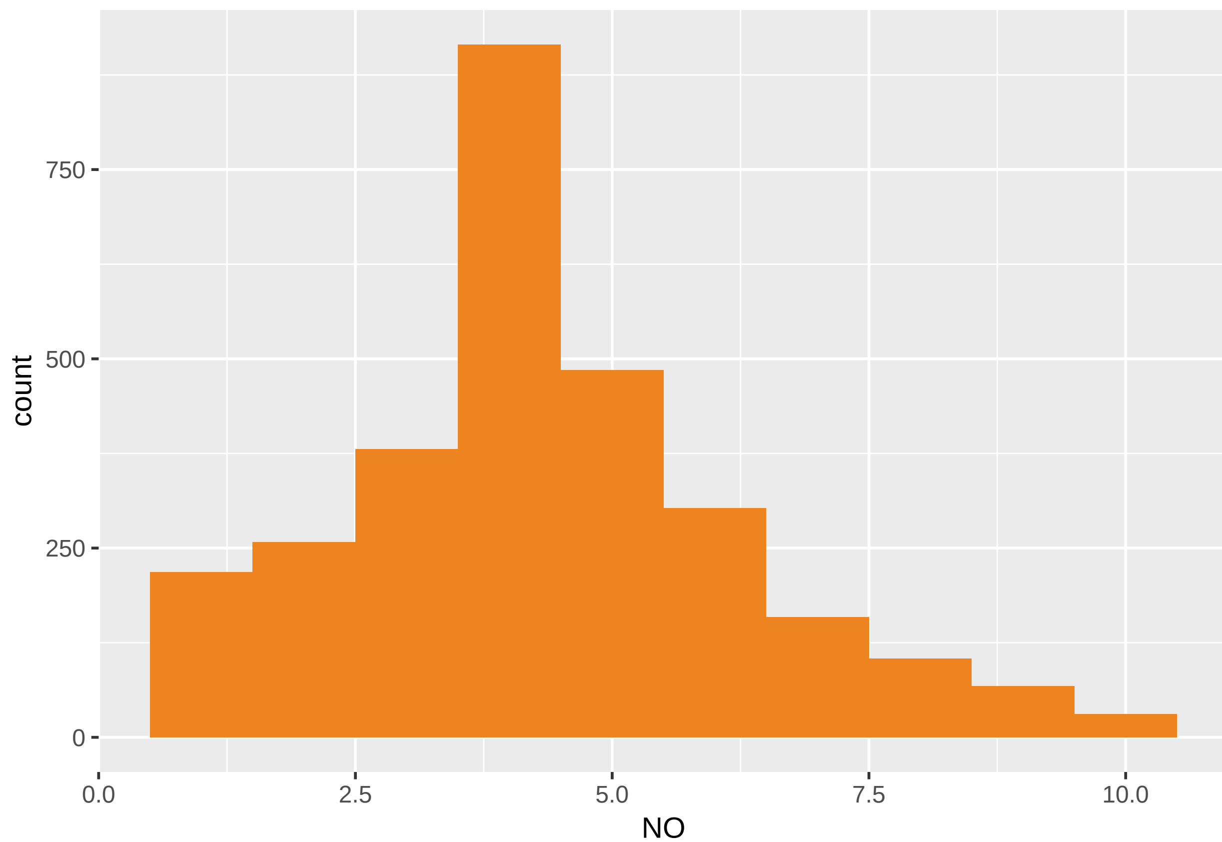


NO

```
#Concentration of NO - g / m³
summary(Canarydataset$NO)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    3.00    4.37    4.37    5.00   10.00
```

```r
ggplot(Canarydataset, aes(x=NO))+geom_histogram(binwidth=1,fill="#EE8420")
```

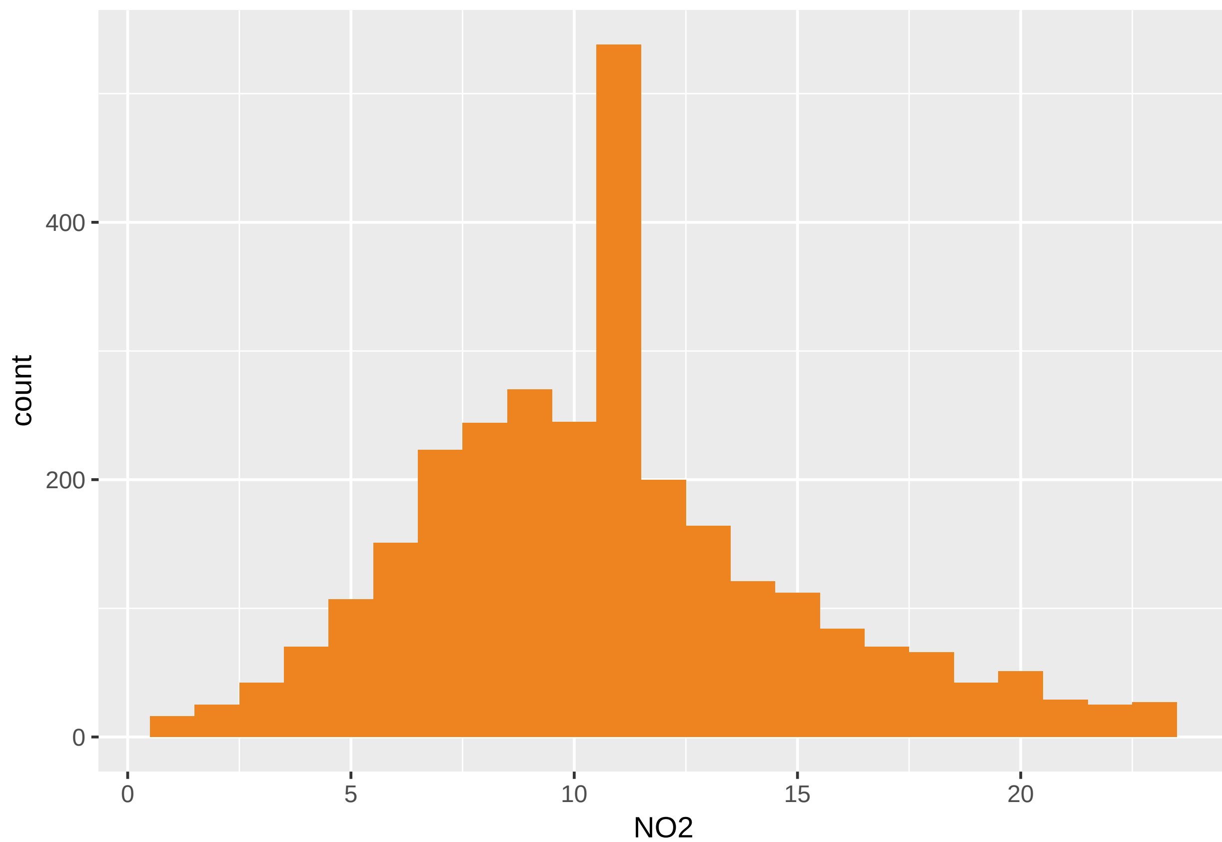


NO2

```r
#Concentration of NO2 - g / m³
summary(Canarydataset$NO2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    8.00   10.73   10.73   13.00   23.00
```

```r
ggplot(Canarydataset, aes(x=NO2))+geom_histogram(binwidth=1,fill="#EE8420")
```

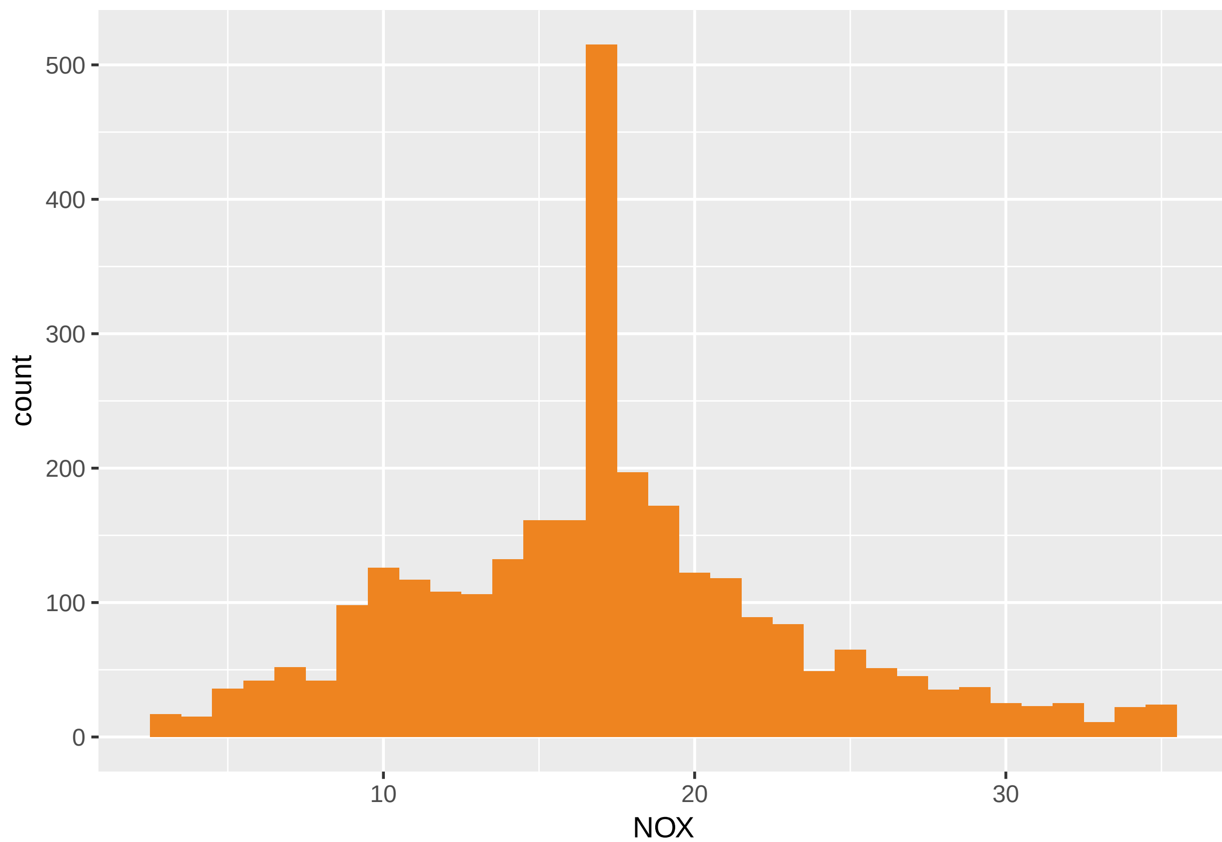


NOX

```
#Concentration of NOX -  g / m³
summary(Canarydataset$NOX)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3.00   13.00   17.19   17.19   20.00   35.00
```

```
ggplot(Canarydataset, aes(x=NOX))+geom_histogram(binwidth=1,fill="#EE8420")
```

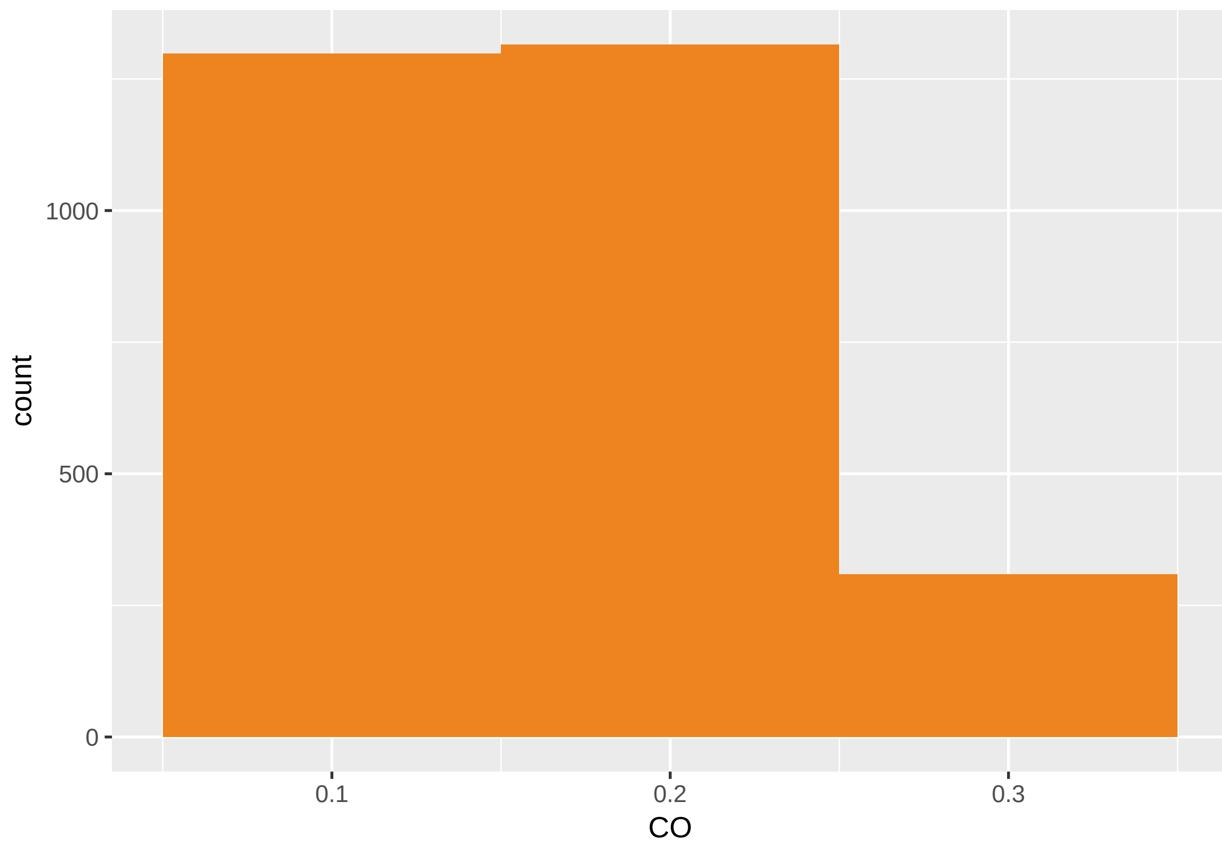


CO

```r
#Concentration of CO - mg / m³
summary(Canarydataset$CO)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1000  0.1000  0.1602  0.1602  0.2000  0.3000
```

```r
ggplot(Canarydataset, aes(x=CO))+geom_histogram(binwidth=0.1,fill="#EE8420")
```

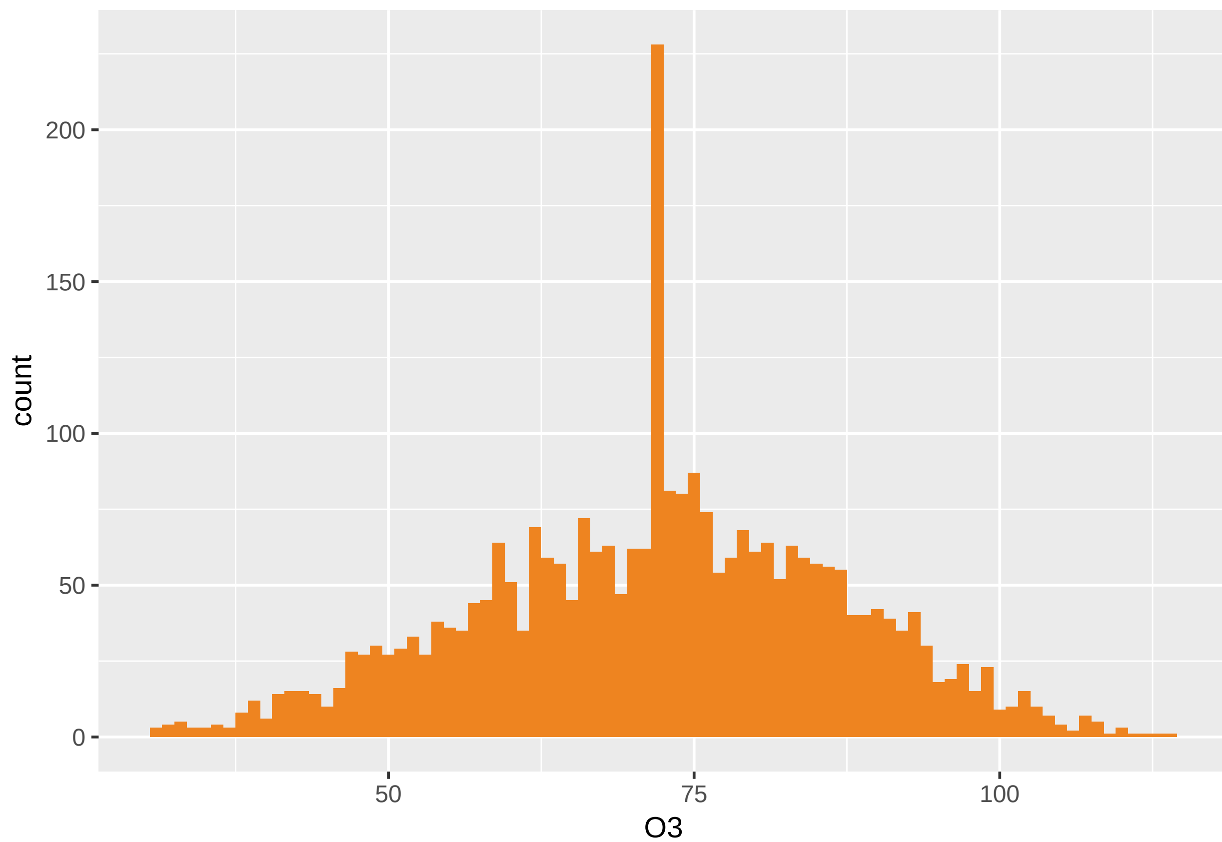


O3

```
#Concentration of O3 - g / m³
summary(Canarydataset$O3)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   31.00   62.00   72.11   72.11   83.00  114.00
```

```
ggplot(Canarydataset, aes(x=O3))+geom_histogram(binwidth=1,fill="#EE8420")
```

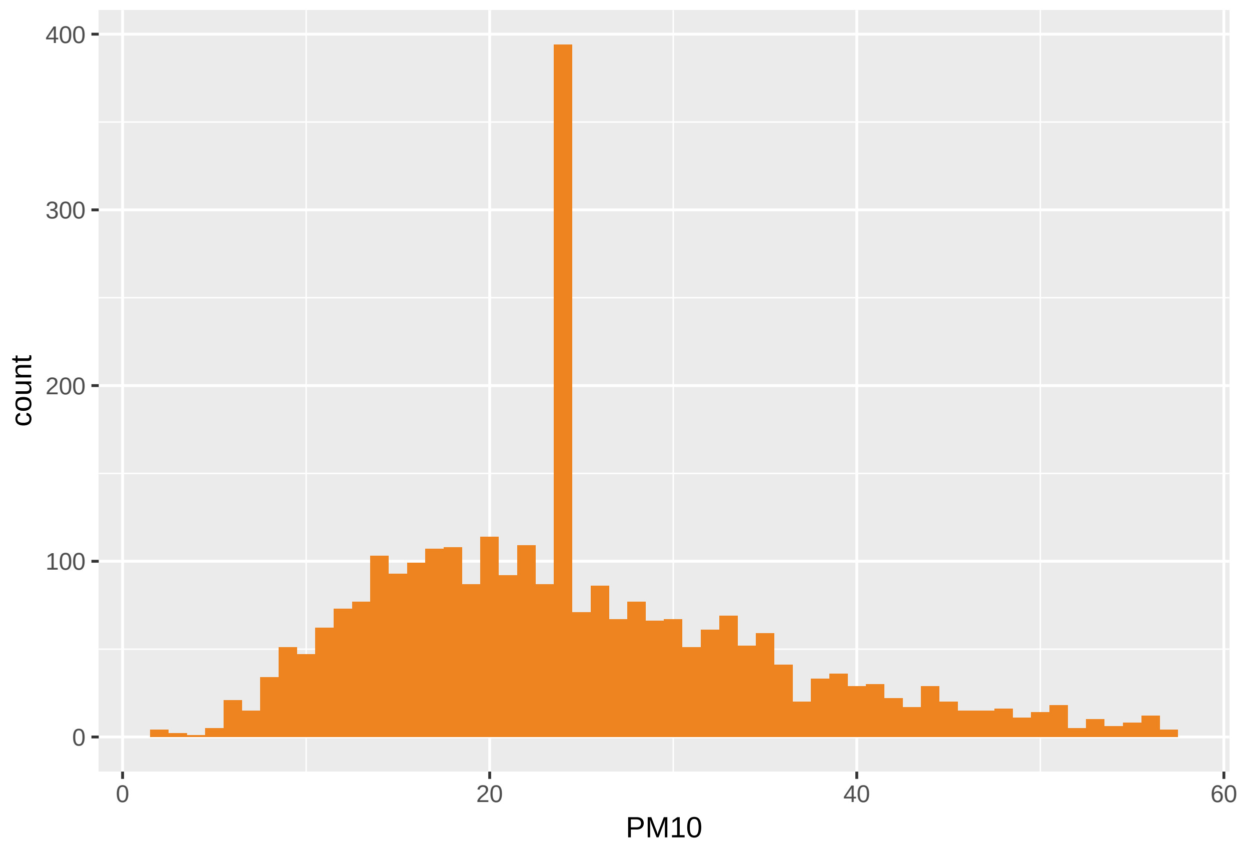


PM10

```r
#Particulate matter (PM10) - g / m³
summary(Canarydataset$PM10)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00   17.00   24.00   24.47   30.00   57.00
```

```r
ggplot(Canarydataset, aes(x=PM10))+geom_histogram(binwidth=1,fill="#EE8420")
```

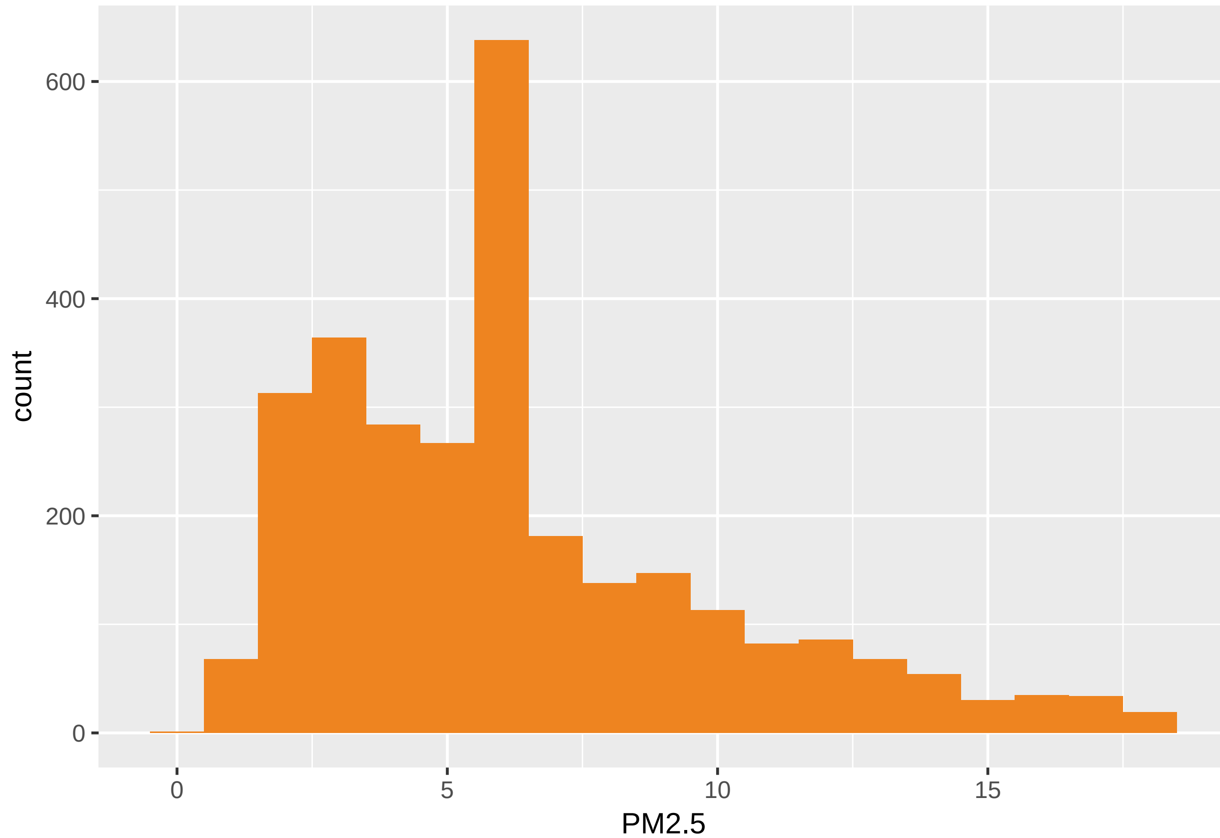


PM2.5

```
#Particulate matter (PM2.5) - g / m³
summary(Canarydataset$PM2.5)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   3.000   6.000   6.347   8.000  18.000
```

```
ggplot(Canarydataset, aes(x=PM2.5))+geom_histogram(binwidth=1,fill="#EE8420")
```

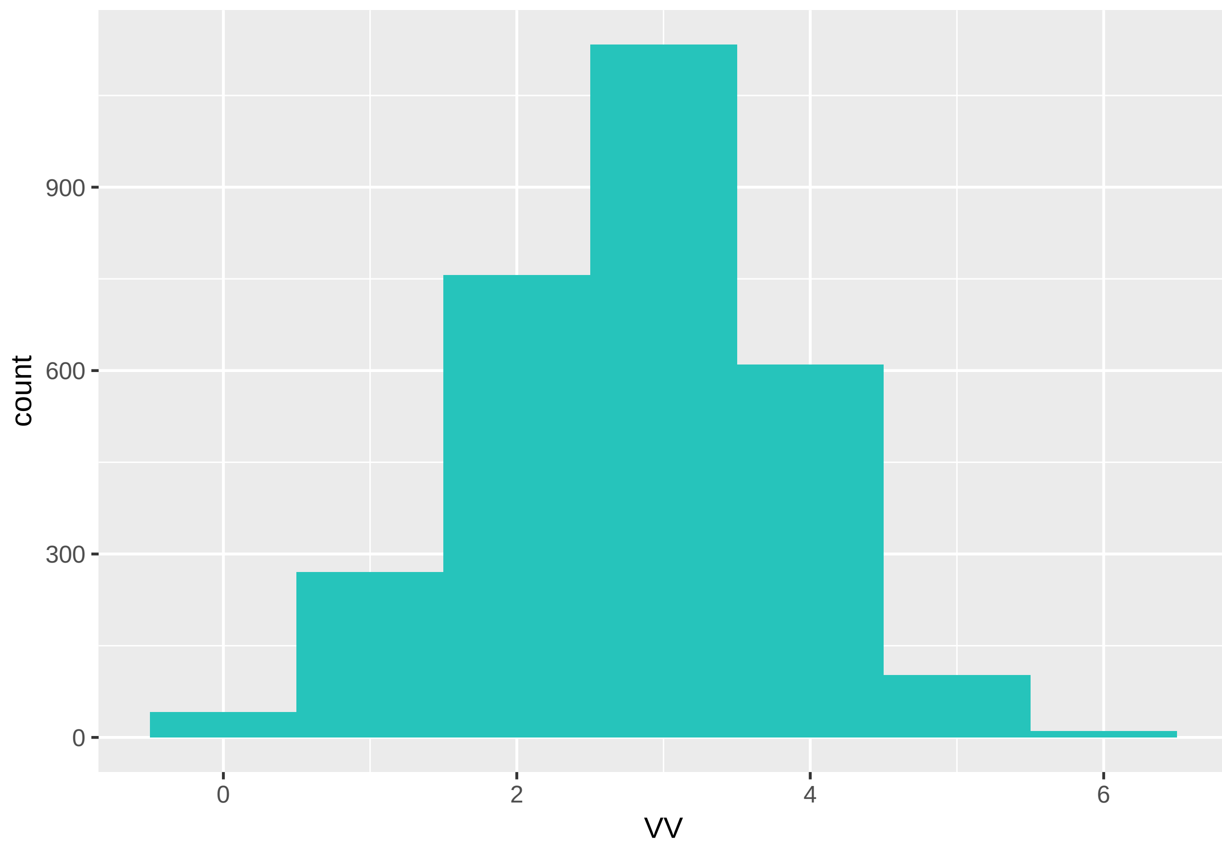


Wind speed

```r
#Wind speed - m / s
summary(Canarydataset$VV)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.100   2.125   2.900   2.847   3.500   5.800
```

```r
ggplot(Canarydataset, aes(x=VV))+geom_histogram(binwidth=1,fill="#26C4BB")
```

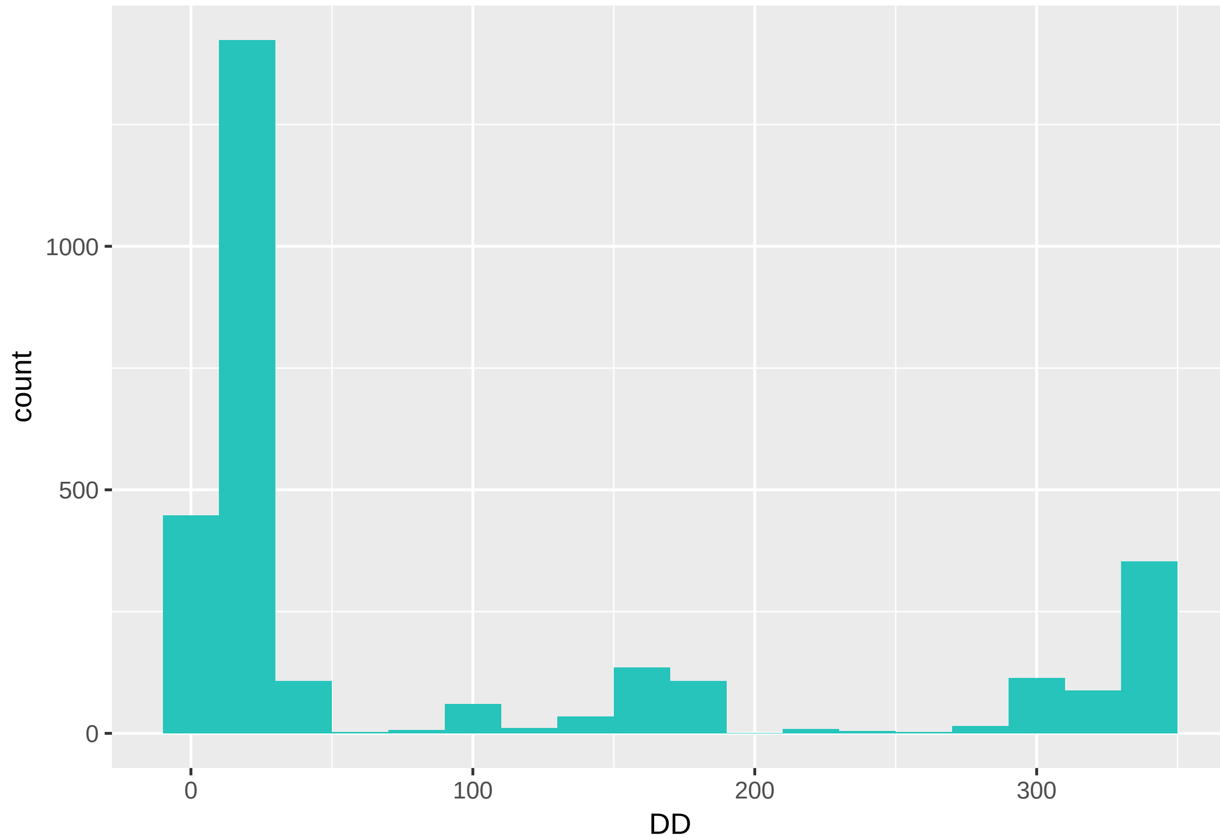


Wind direction

```r
#Wind direction - Grd
summary(Canarydataset$DD)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   21.00   23.00   95.61  158.00  337.00
```

```r
ggplot(Canarydataset, aes(x=DD))+geom_histogram(binwidth=20,fill="#26C4BB")
```

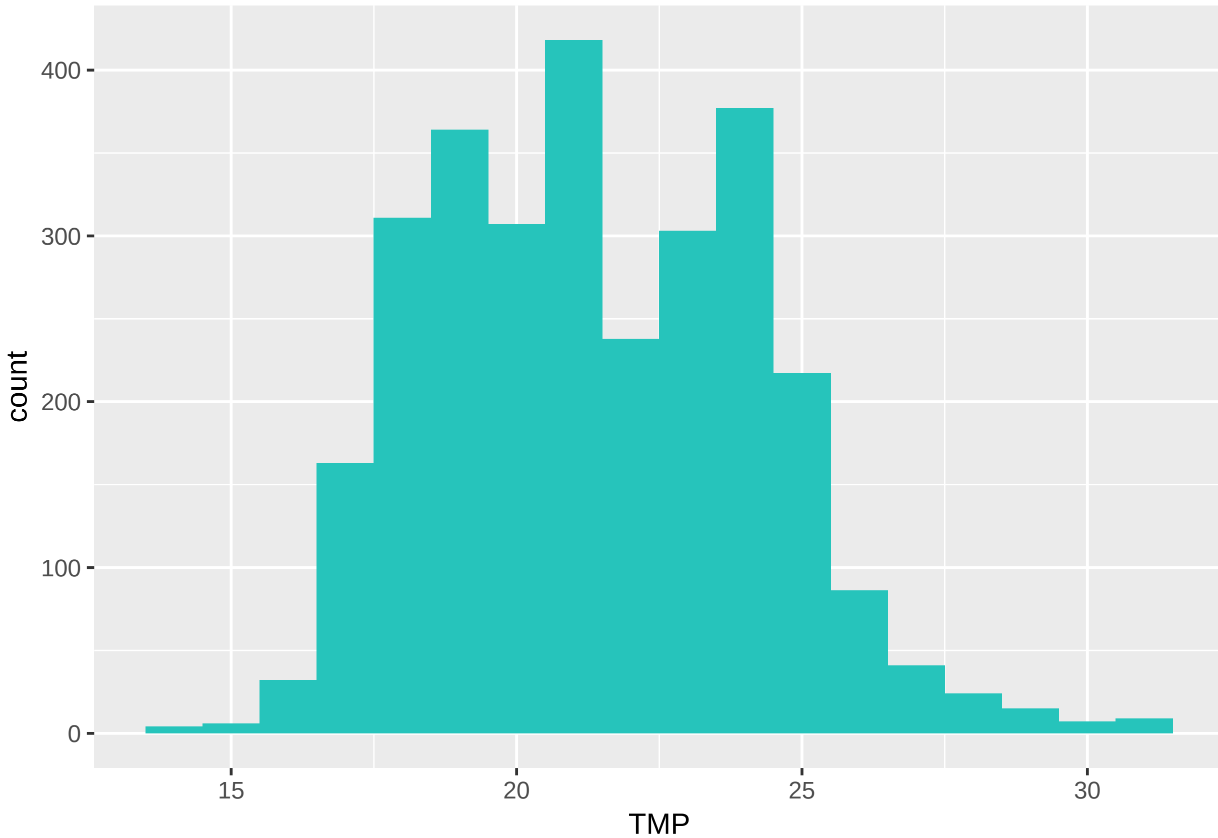


Average temperature

```r
#Average temperature - ºC
summary(Canarydataset$TMP)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   13.80   19.10   21.45   21.45   23.70   31.00
```

```r
ggplot(Canarydataset, aes(x=TMP))+geom_histogram(binwidth=1,fill="#26C4BB")
```

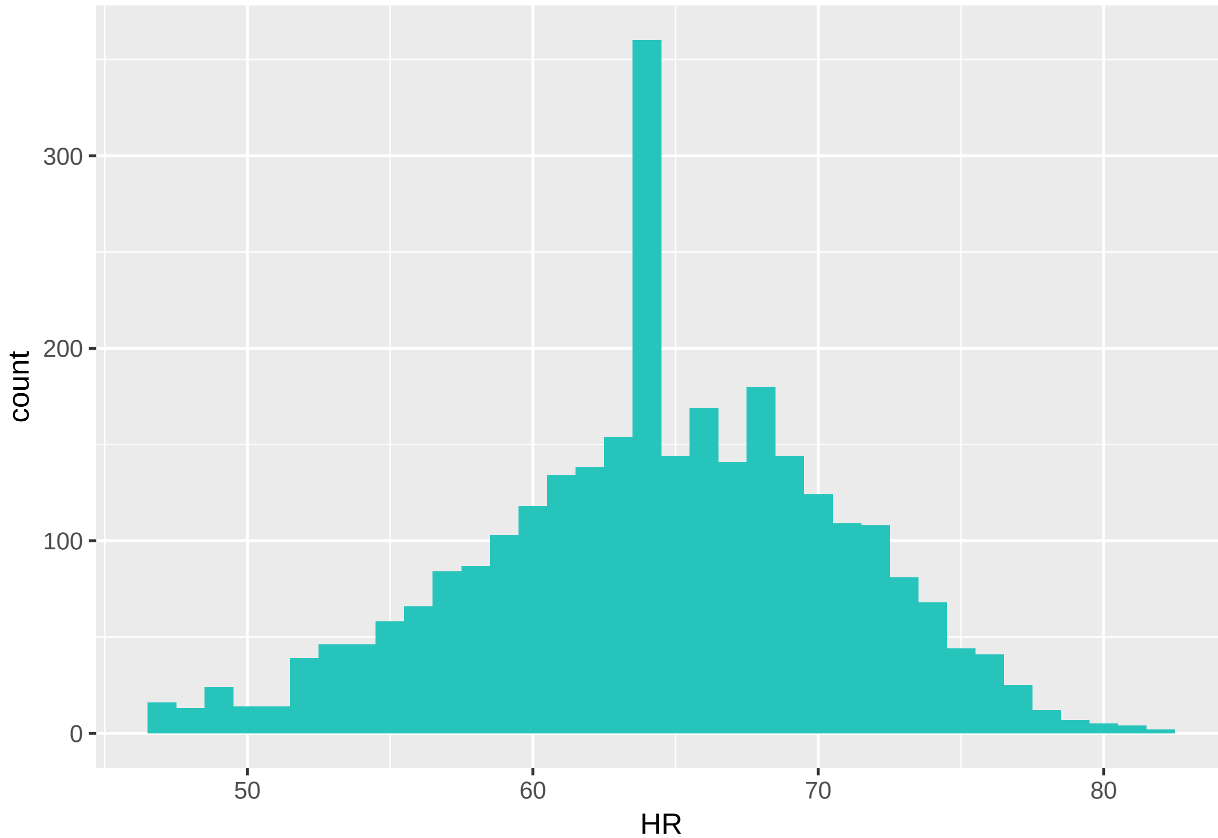


Relative humidity

```
#Relative humidity -%
summary(Canarydataset$HR)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   47.00   61.00   64.46   64.46   69.00   82.00
```

```
ggplot(Canarydataset, aes(x=HR))+geom_histogram(binwidth=1,fill="#26C4BB")
```



Barometric pressure

```
#Barometric pressure - mb
summary(Canarydataset$PRB)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     999    1012    1016    1016    1020    1034
```

```
ggplot(Canarydataset, aes(x=PRB))+geom_histogram(binwidth=1,fill="#26C4BB")
```

Summary of descriptive statistics

```
numSummary(Canarydataset[,c("SO2","NO","NO2","NOX","O3","CO","PM10","PM2.5","VV","DD","TMP","HR","PRB"
), drop=FALSE], statistics=c("mean", "sd", "quantiles", "skewness", "kurtosis"),
quantiles=c(0,.25,.5,.75,1), type="2")
```

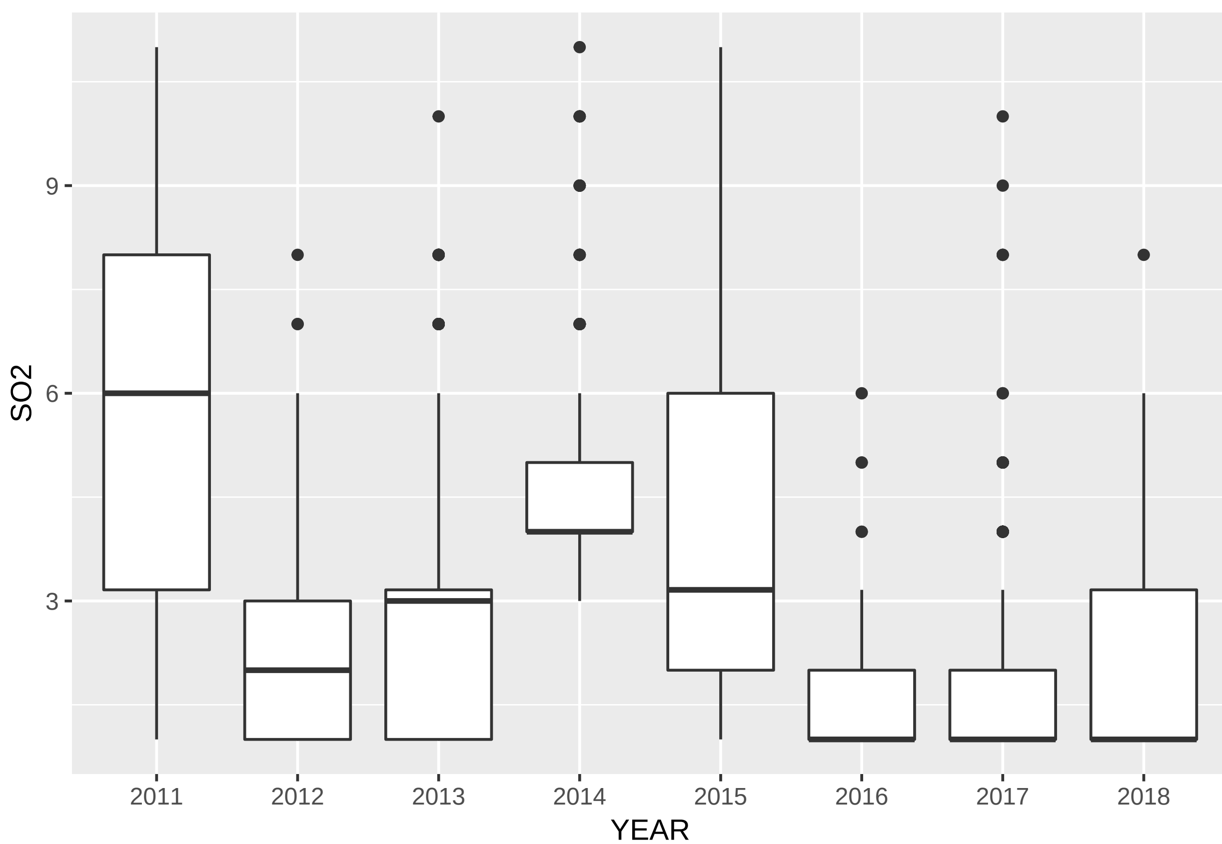
```
##              mean           sd   skewness   kurtosis    0%      25%
## SO2     3.1605979   2.32292870  1.2043834  1.03181218   1.0    1.000
## NO      4.3696993   1.90752114  0.4345443  0.31080297   1.0    3.000
## NO2    10.7337437   4.23940170  0.5611819  0.34734162   1.0    8.000
## NOX    17.1933100   6.21614900  0.4139438  0.33971332   3.0   13.000
## O3     72.1123596  14.78001612 -0.1178599 -0.25784218  31.0   62.000
## CO      0.1601531   0.06443907  0.7911818 -0.26963883   0.1    0.100
## PM10   24.4685583  10.37049745  0.7039339  0.31241089   2.0   17.000
## PM2.5   6.3471237   3.68640572  0.9893938  0.63328987   0.0    3.000
## VV      2.8469067   1.00996492 -0.1167381 -0.12239138   0.1    2.125
## DD     95.6127006 121.06461461  1.1604696 -0.37285103   0.0   21.000
## TMP    21.4533479   2.81879417  0.2748527 -0.33420350  13.8   19.100
## HR     64.4642195   6.33750073 -0.2627982 -0.09194337  47.0   61.000
## PRB  1016.1312478   6.82598030  0.1799721 -0.24083163 999.0 1012.000
##               50%    75%   100%    n
## SO2     3.0000000    4.0   11.0 2922
## NO      4.3696993    5.0   10.0 2922
```

```
## NO2      10.7337437   13.0   23.0 2922
## NOX      17.1933100   20.0   35.0 2922
## O3       72.1123596   83.0  114.0 2922
## CO        0.1601531    0.2    0.3 2922
## PM10     24.0000000   30.0   57.0 2922
## PM2.5     6.0000000    8.0   18.0 2922
## VV        2.9000000    3.5    5.8 2922
## DD       23.0000000  158.0  337.0 2922
## TMP      21.4533479   23.7   31.0 2922
## HR       64.4642195   69.0   82.0 2922
## PRB    1016.0000000 1020.0 1034.0 2922
```

Boxplot by year

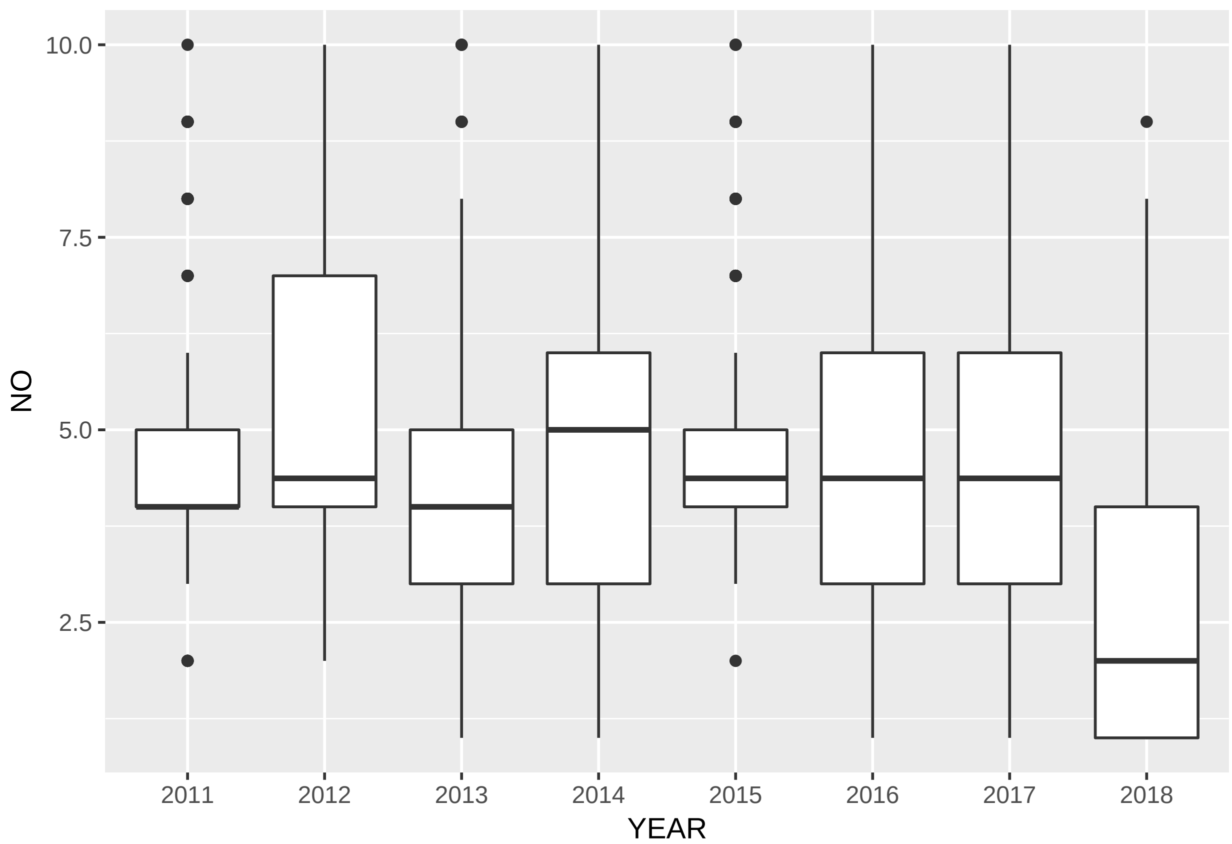
```
#Boxplots for each year
ggplot(Canarydataset, aes(x=YEAR, y=SO2)) + geom_boxplot()+geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



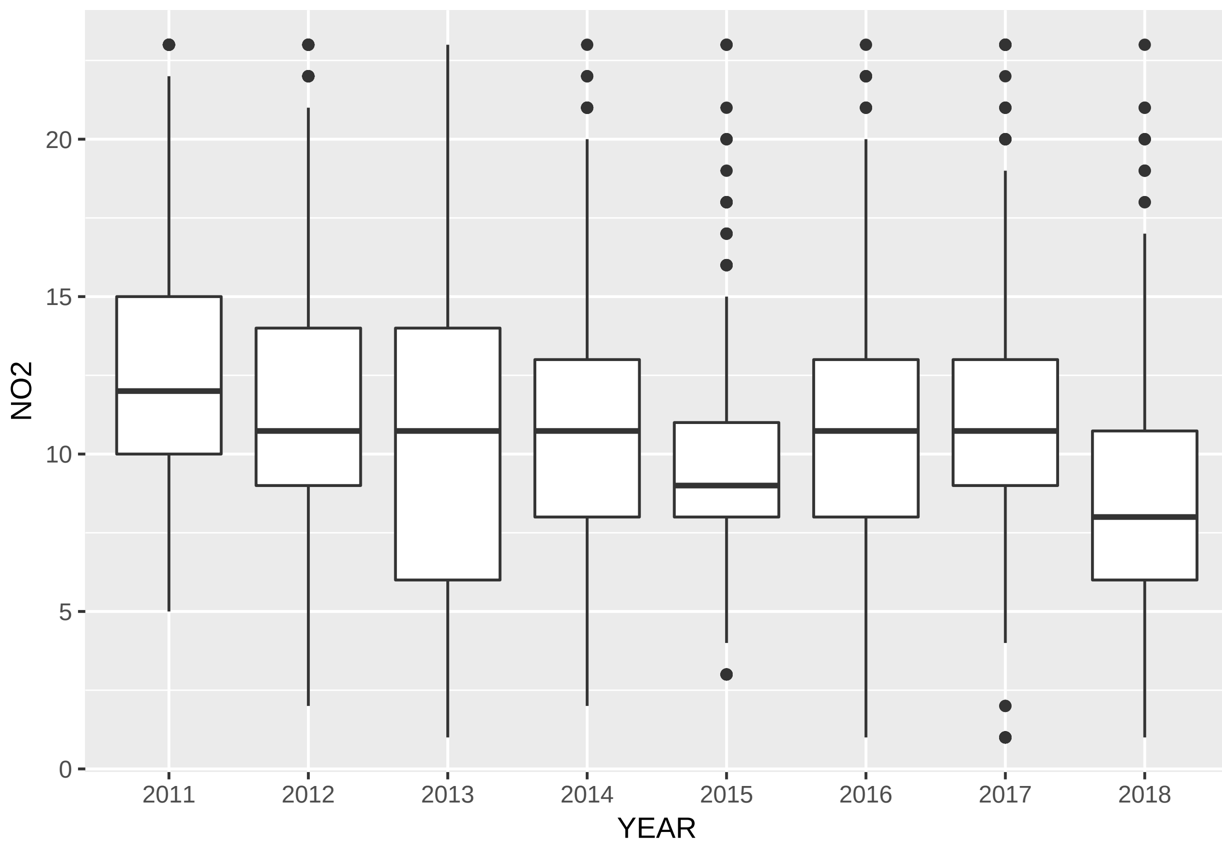

```
ggplot(Canarydataset, aes(x=YEAR, y=NO)) + geom_boxplot()+geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



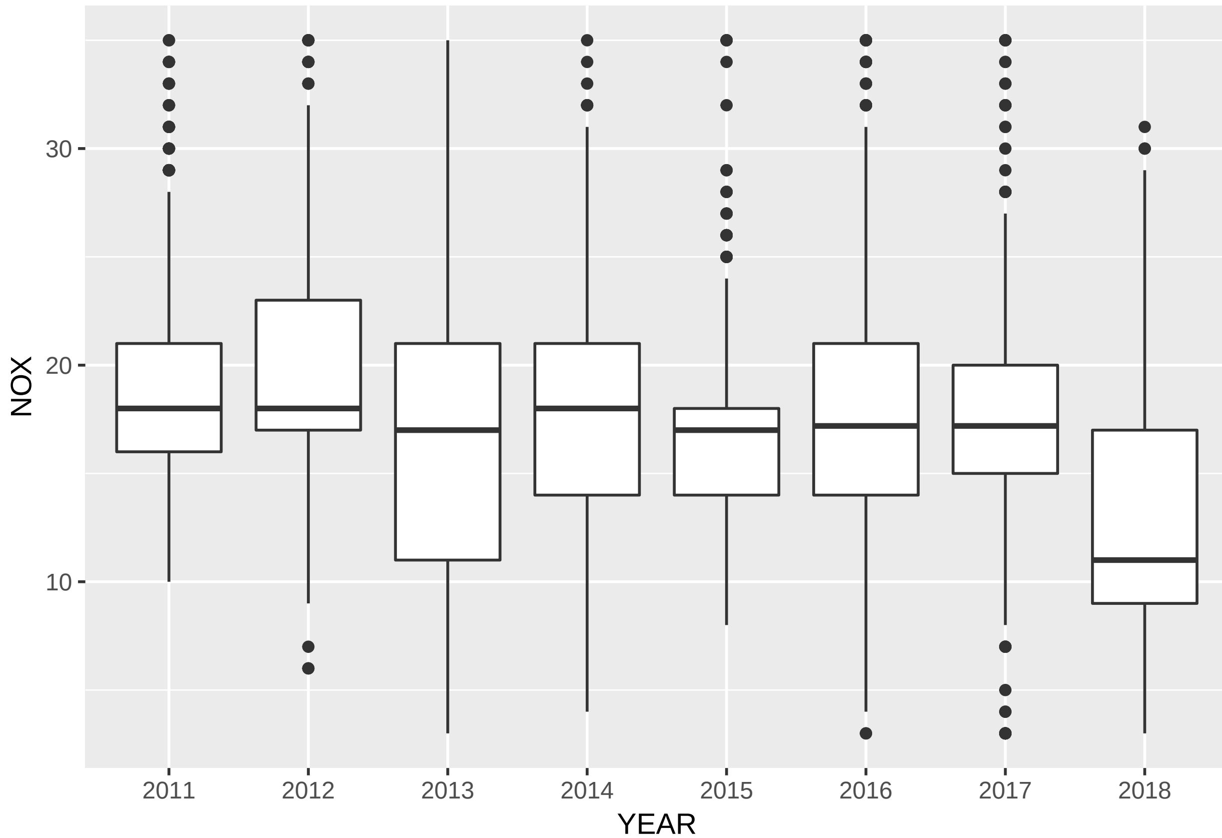

```r
ggplot(Canarydataset, aes(x=YEAR, y=NO2)) + geom_boxplot()+geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



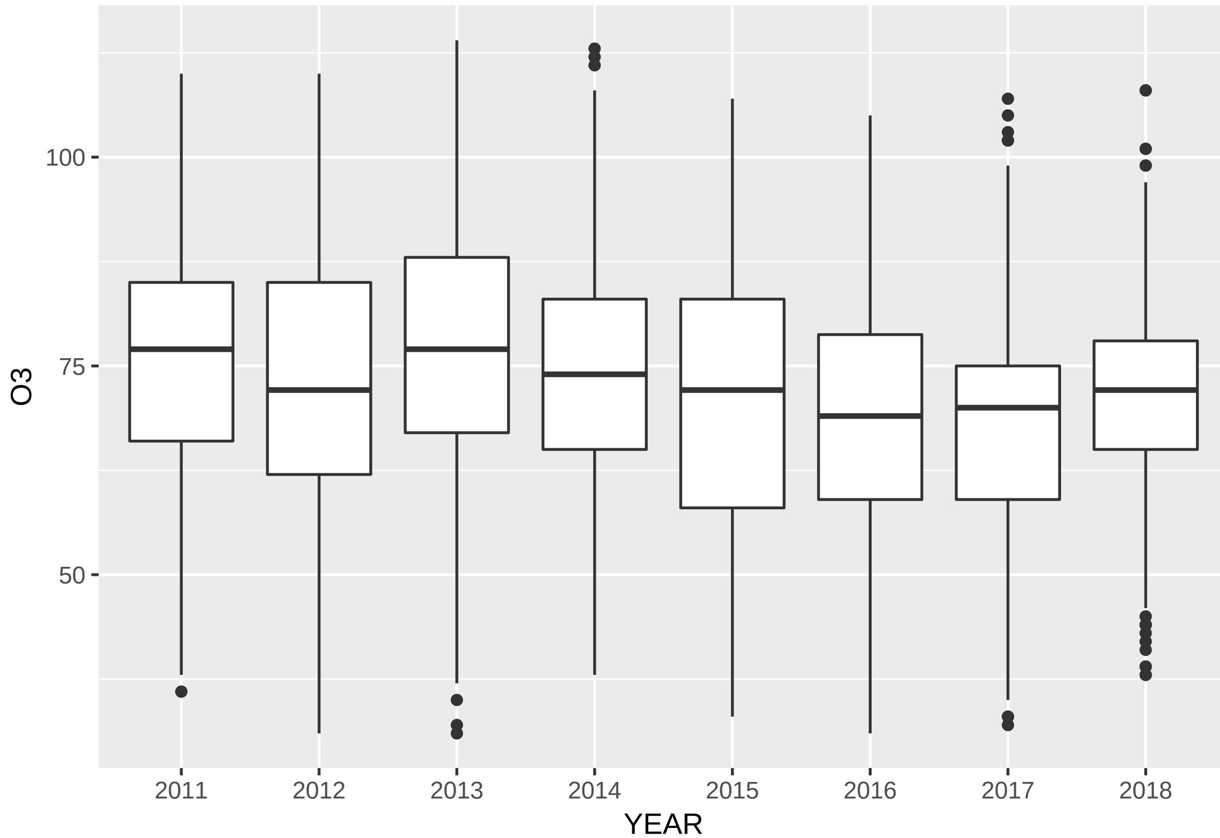

```
ggplot(Canarydataset, aes(x=YEAR, y=NOX)) + geom_boxplot()+geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



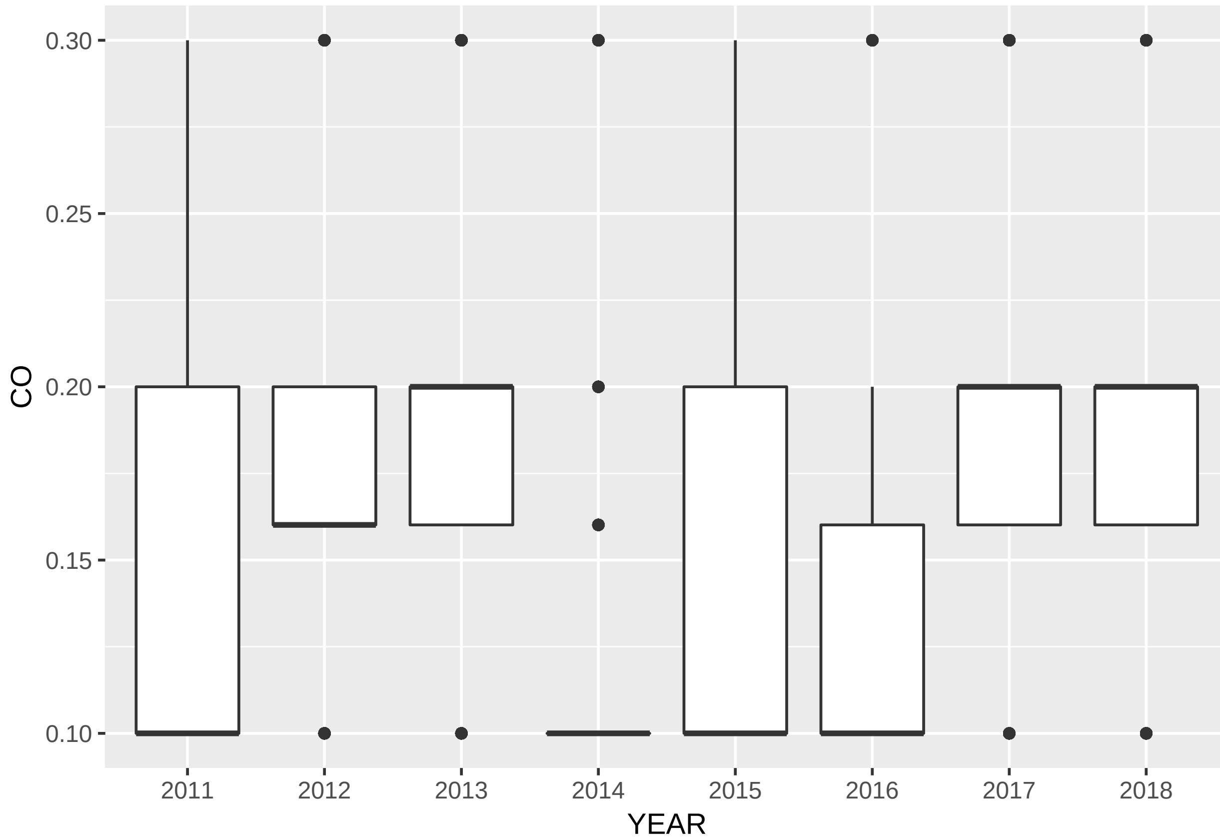

```r
ggplot(Canarydataset, aes(x=YEAR, y=O3)) + geom_boxplot()+geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



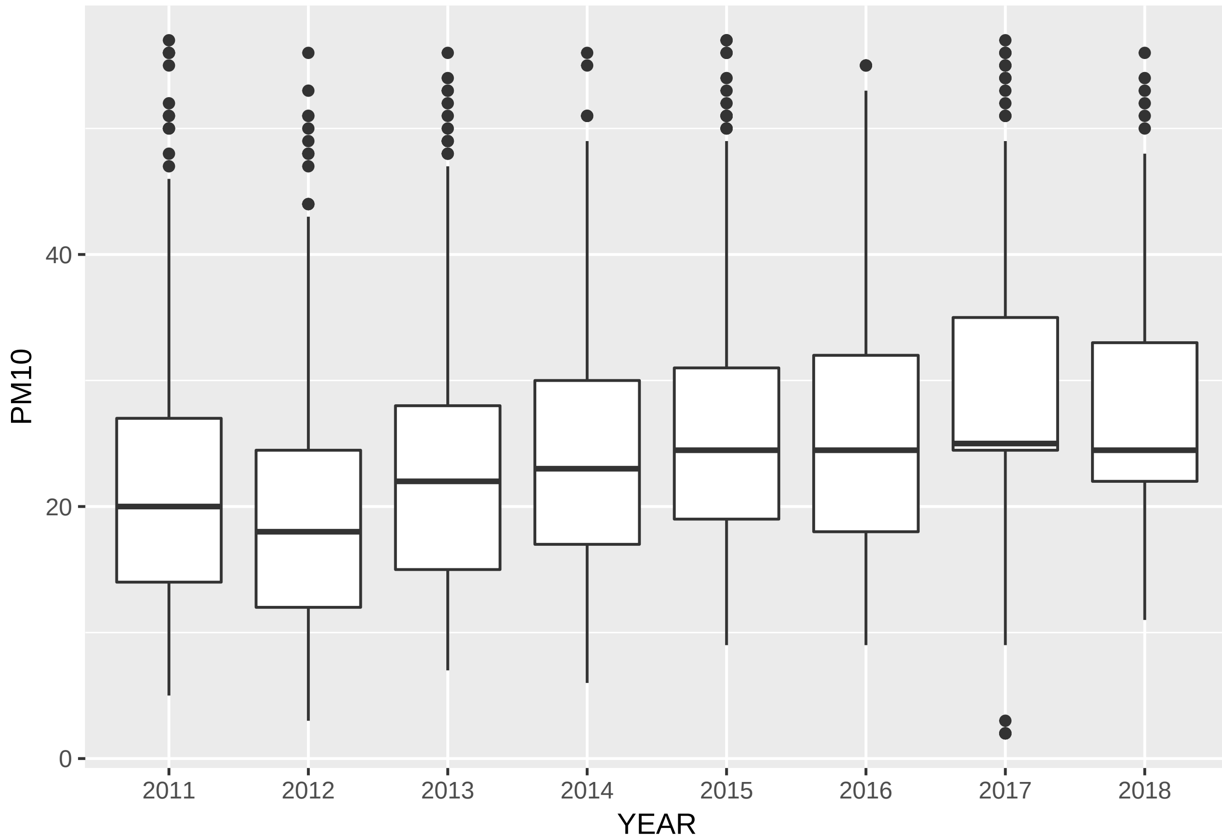

```
ggplot(Canarydataset, aes(x=YEAR, y=CO)) + geom_boxplot()+geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



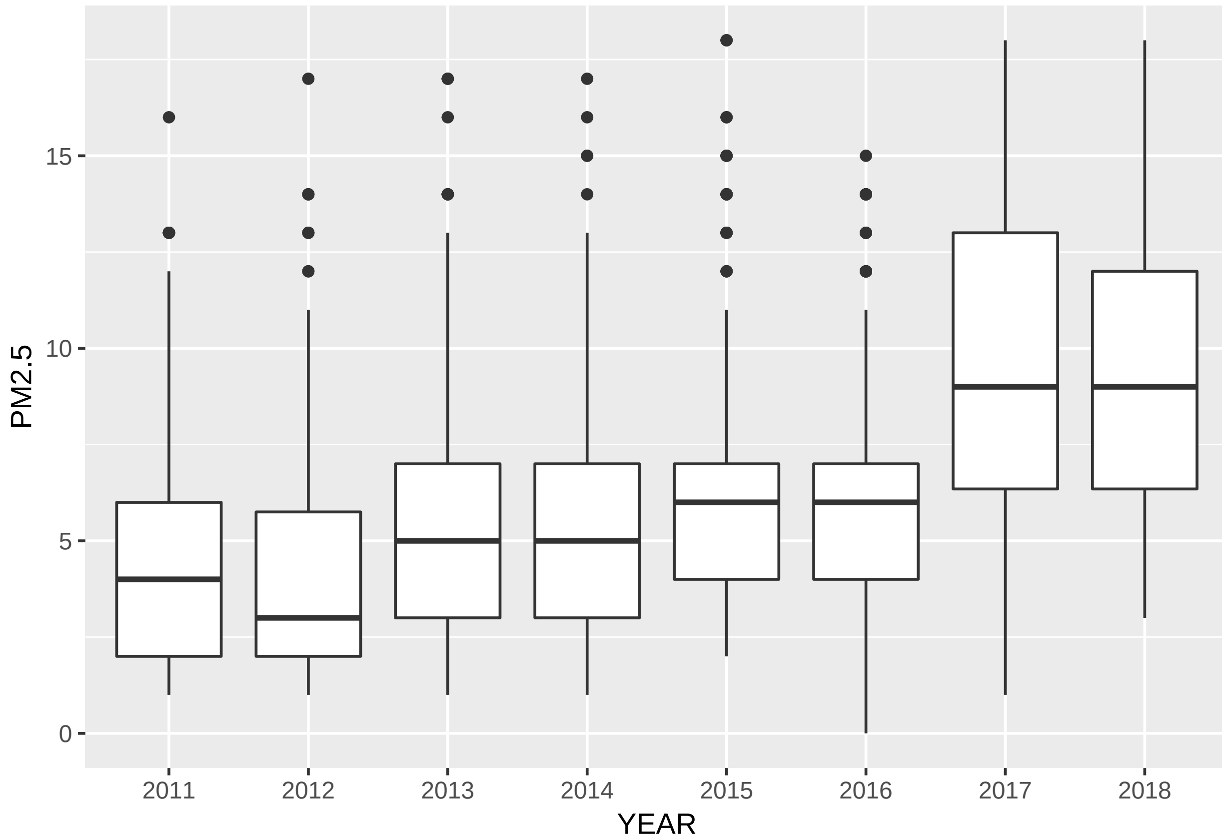

```
ggplot(Canarydataset, aes(x=YEAR, y=PM10)) + geom_boxplot()+geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



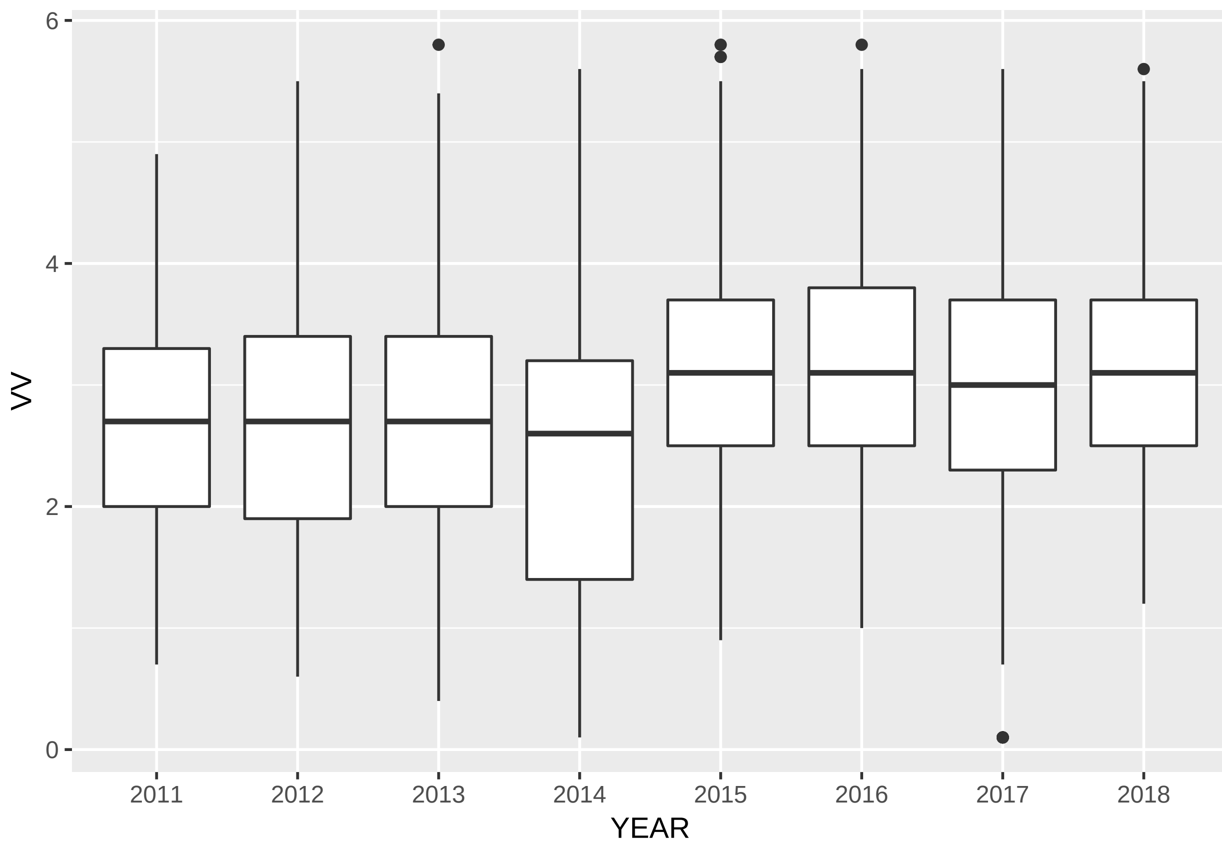

```
ggplot(Canarydataset, aes(x=YEAR, y=PM2.5)) + geom_boxplot()+geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



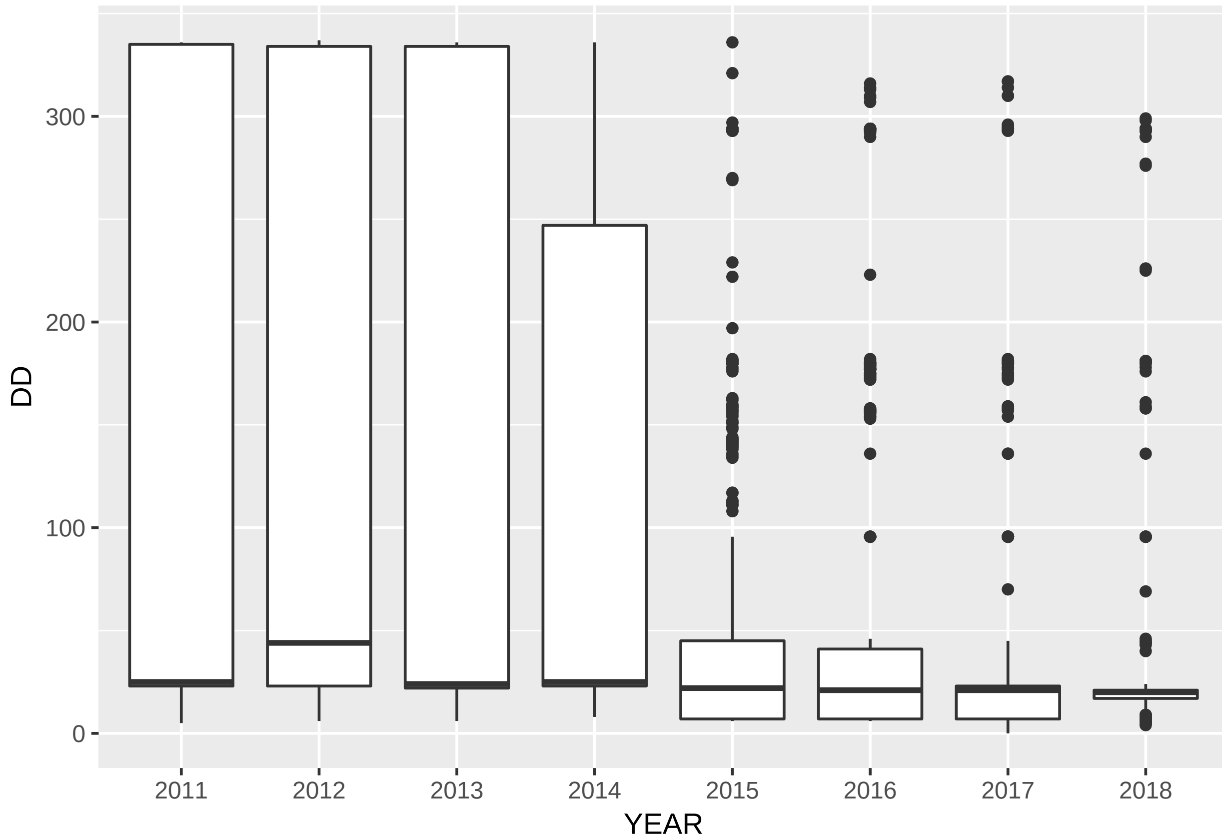

```
ggplot(Canarydataset, aes(x=YEAR, y=VV)) + geom_boxplot()+geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



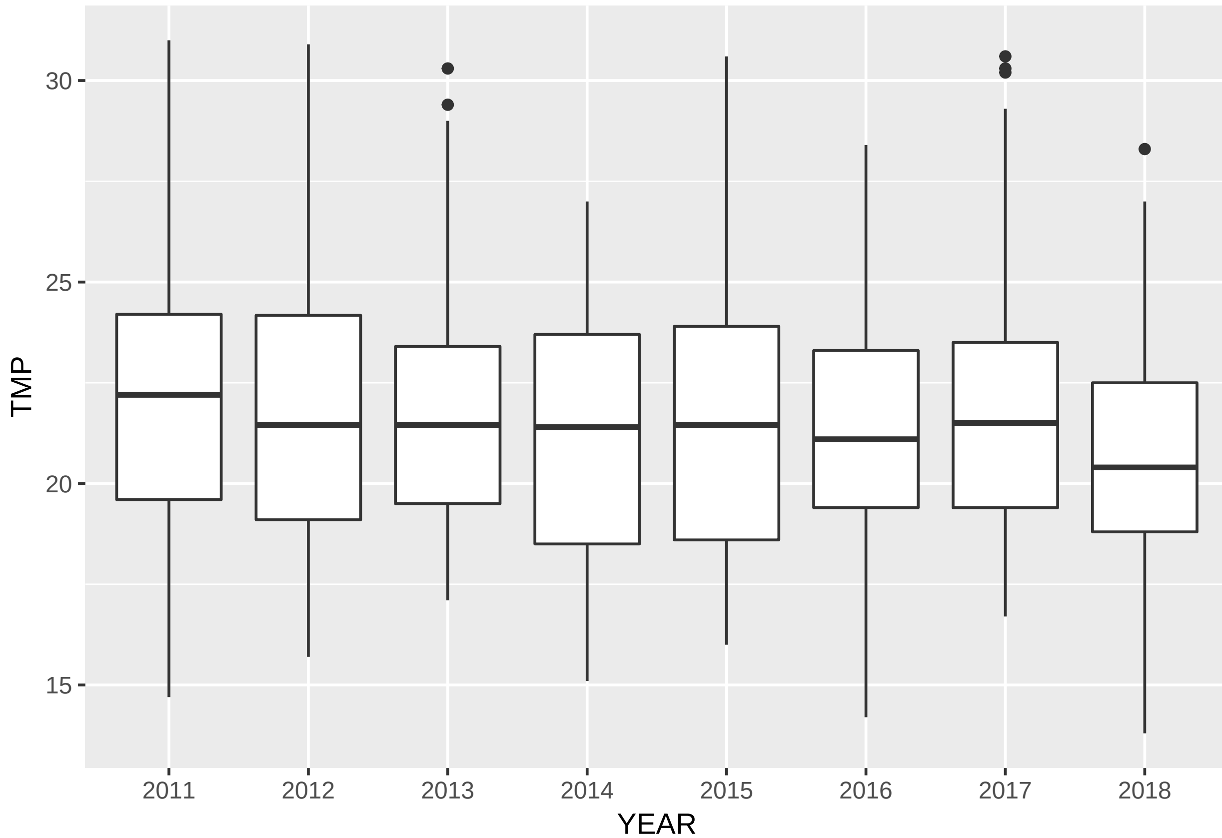

```r
ggplot(Canarydataset, aes(x=YEAR, y=DD)) + geom_boxplot()+geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



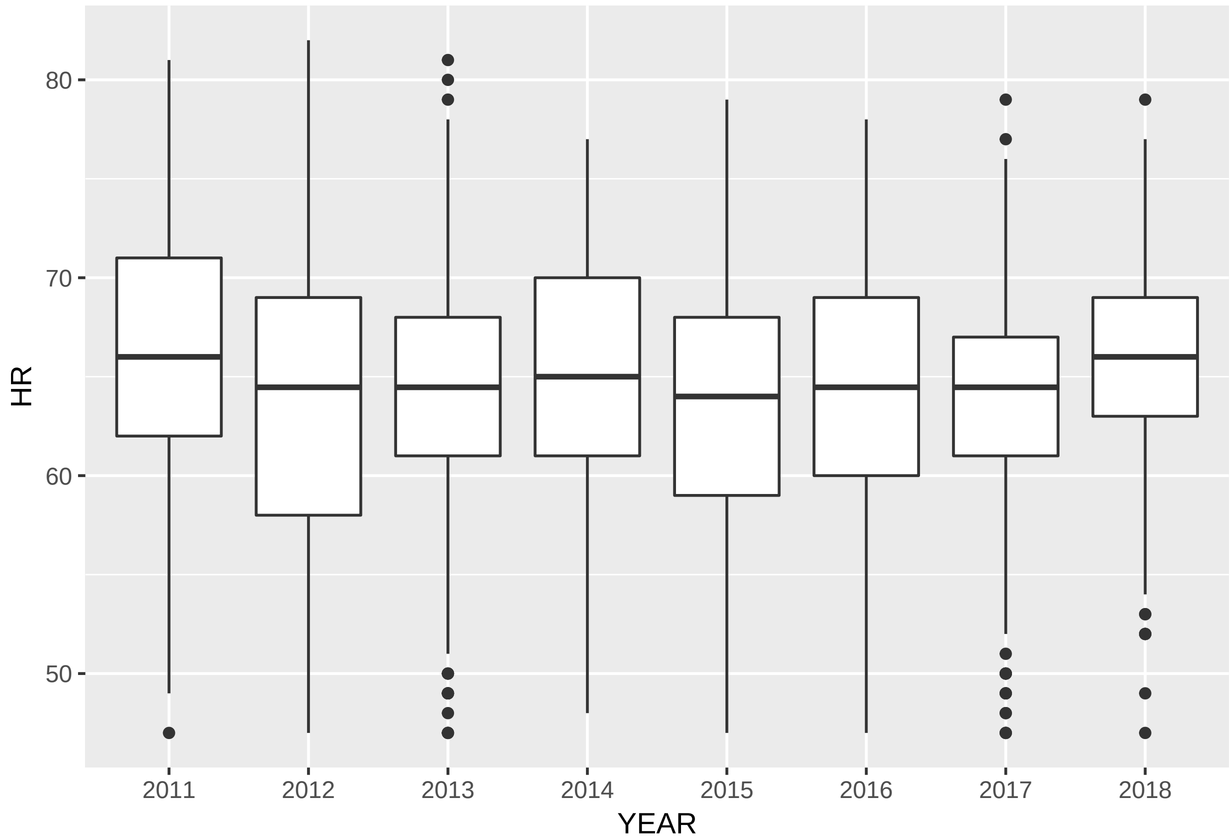

```r
ggplot(Canarydataset, aes(x=YEAR, y=TMP)) + geom_boxplot()+geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



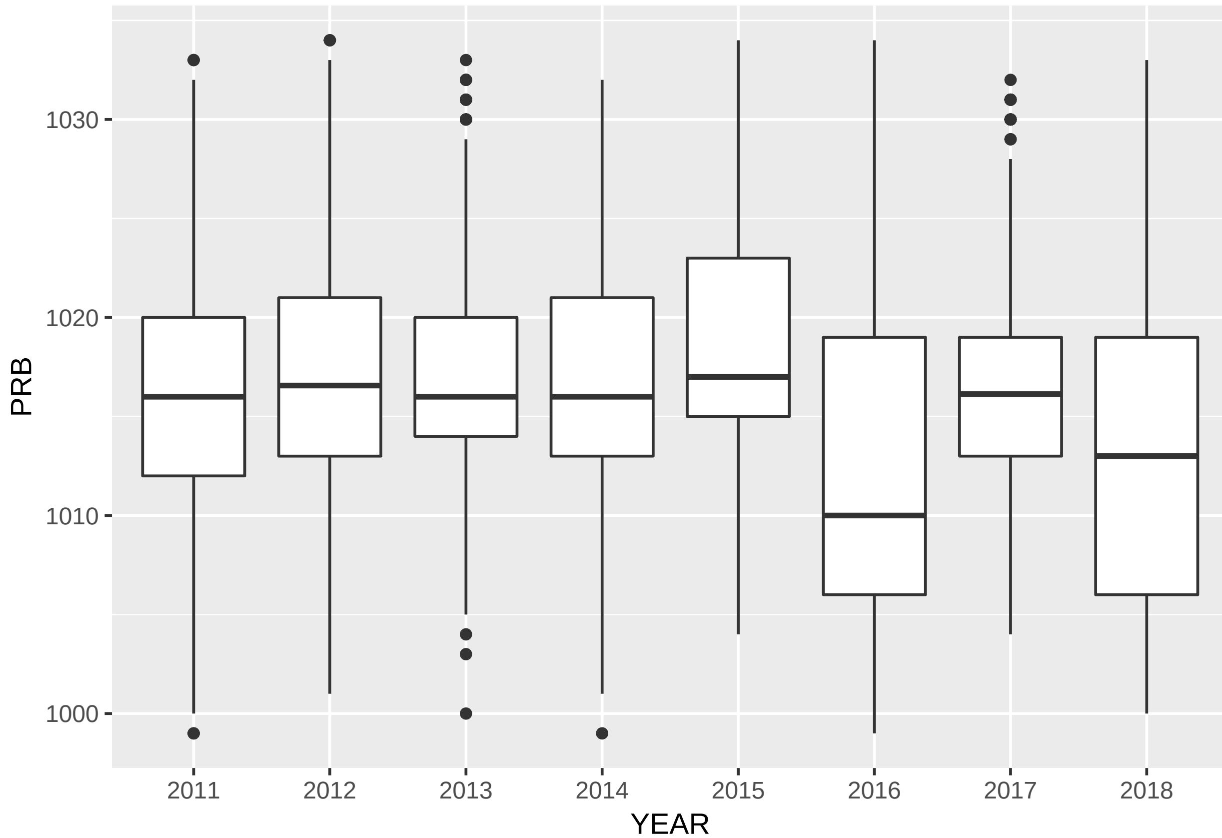

```
ggplot(Canarydataset, aes(x=YEAR, y=HR)) + geom_boxplot()+geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



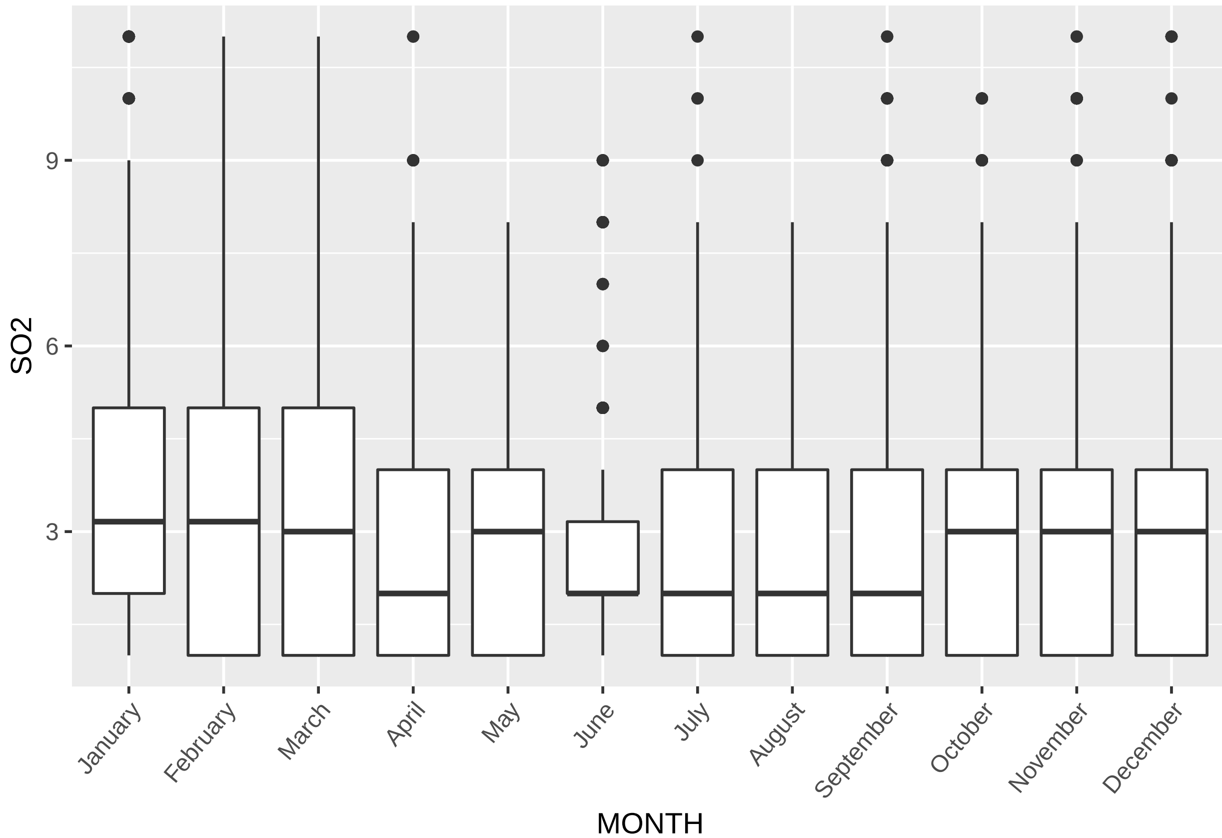

```
ggplot(Canarydataset, aes(x=YEAR, y=PRB)) + geom_boxplot()+geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

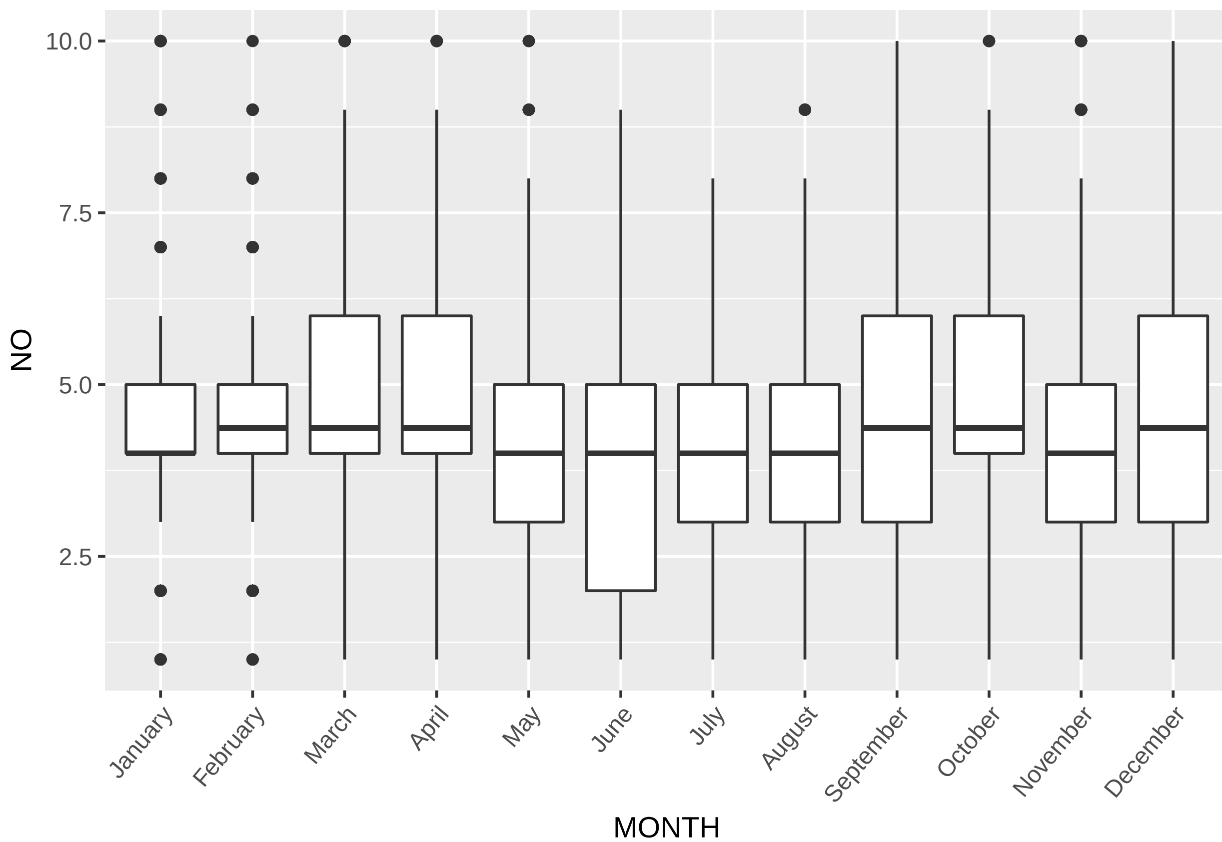


Boxplot by month

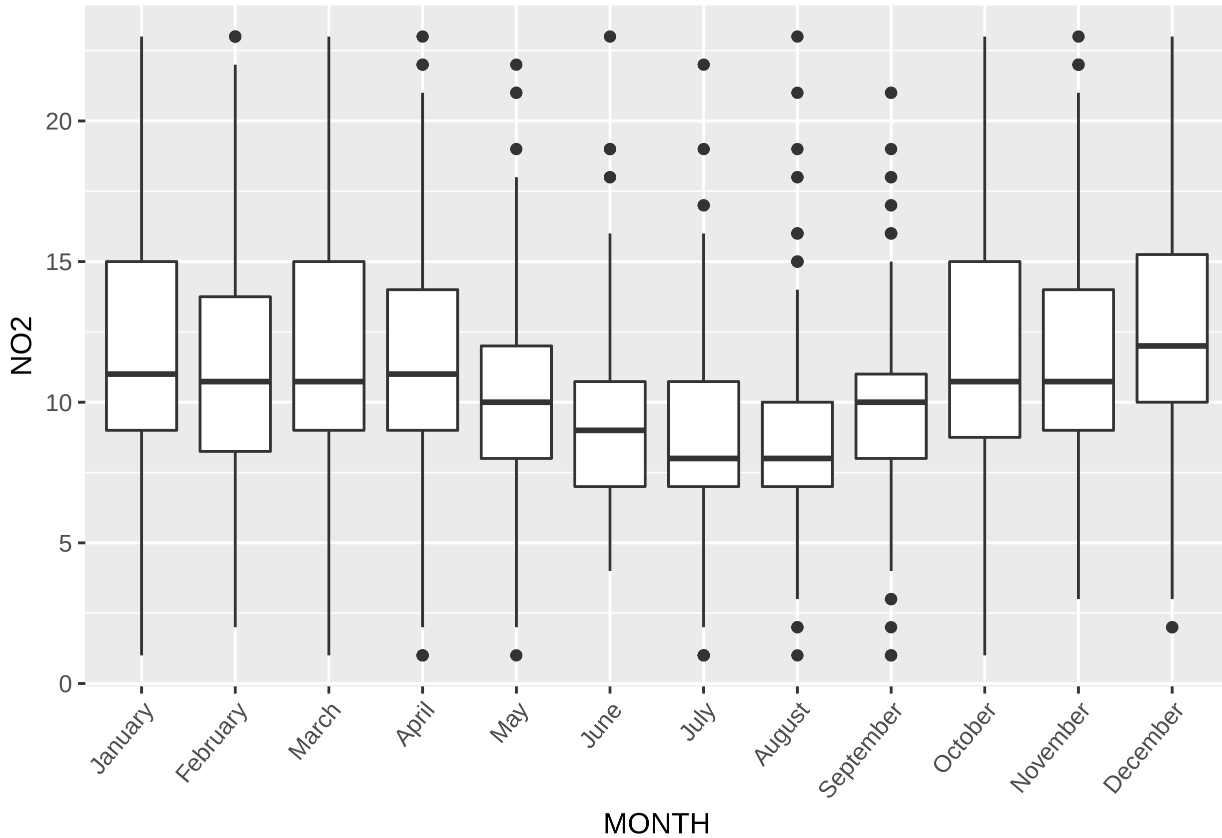
```
#Boxplots for each month
ggplot(Canarydataset, aes(x=MONTH, y=SO2)) + geom_boxplot()+theme(axis.text.x = element_text(angle = 50
```



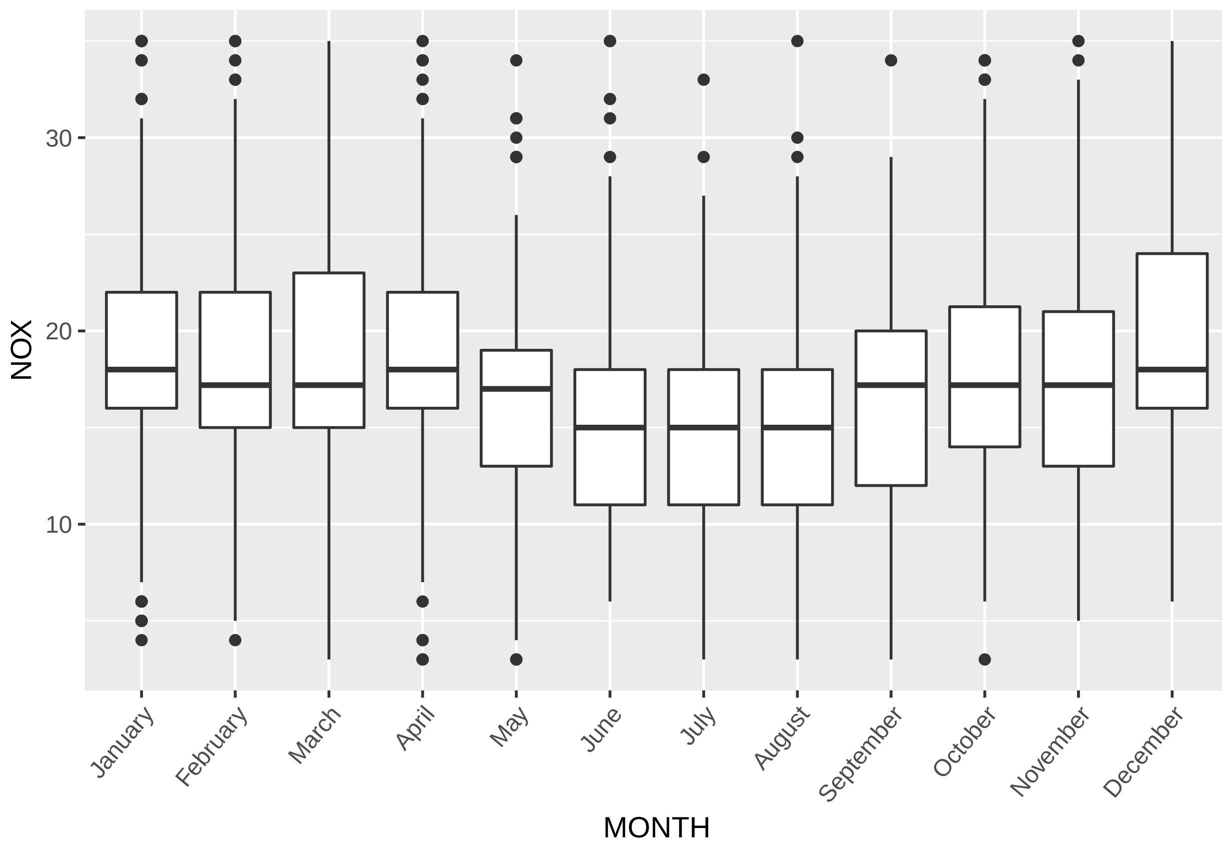

```
ggplot(Canarydataset, aes(x=MONTH, y=NO)) + geom_boxplot()+theme(axis.text.x = element_text(angle = 50,
```



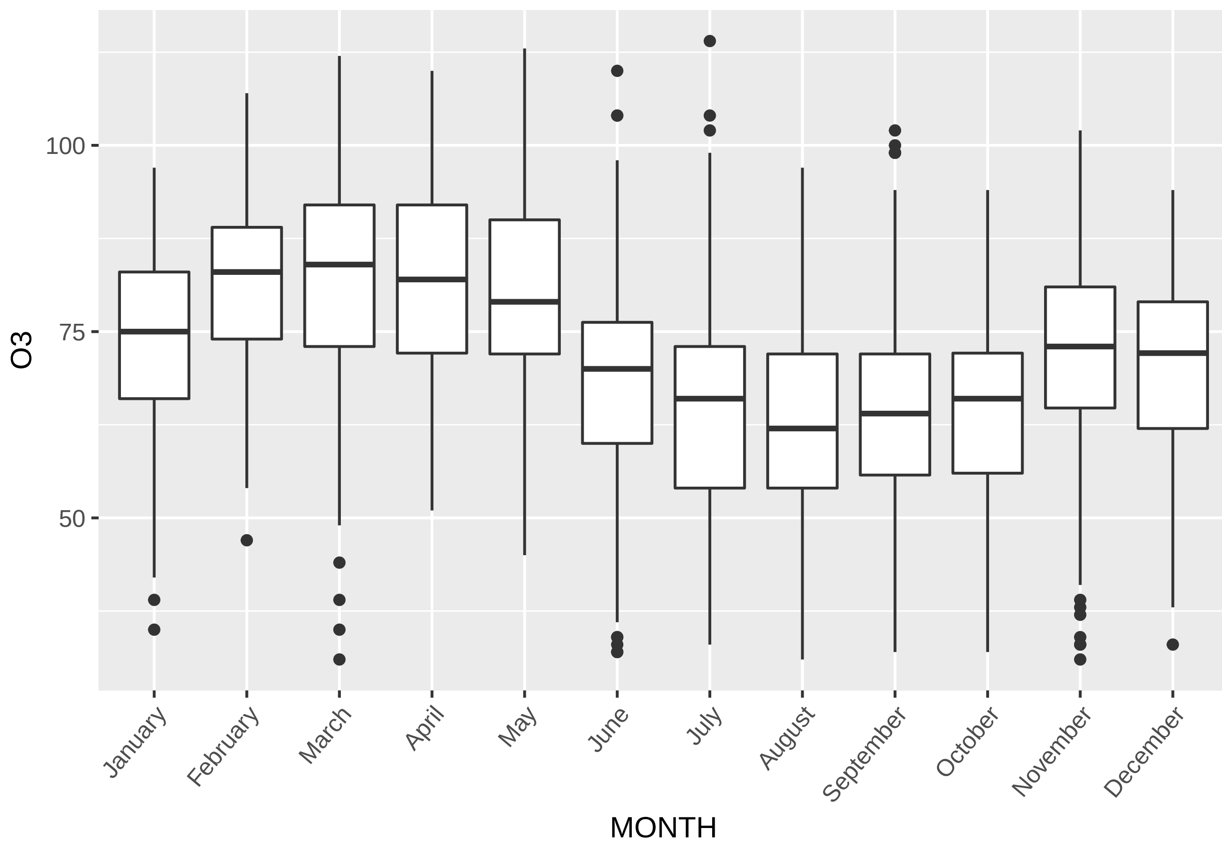

```
ggplot(Canarydataset, aes(x=MONTH, y=NO2)) + geom_boxplot()+theme(axis.text.x = element_text(angle = 50
```



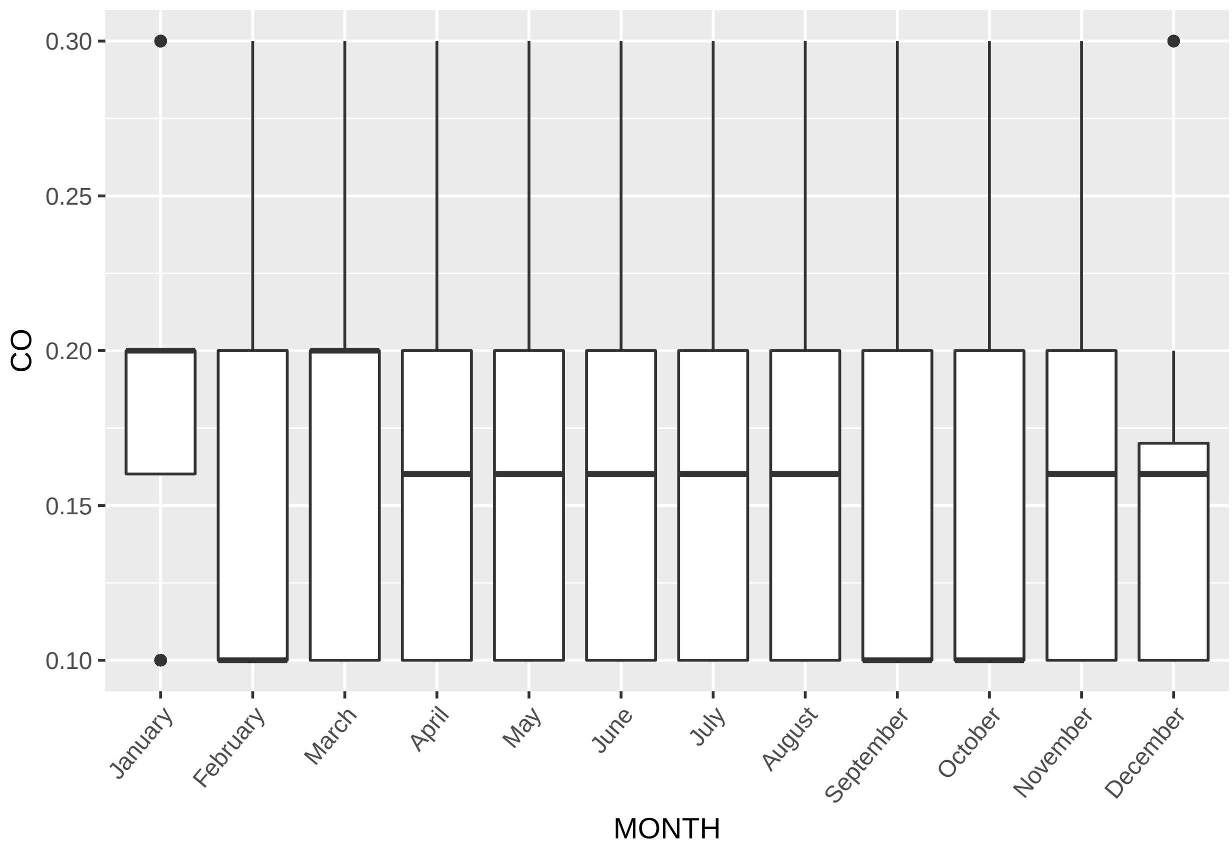

```
ggplot(Canarydataset, aes(x=MONTH, y=NOX)) + geom_boxplot()+theme(axis.text.x = element_text(angle = 50
```



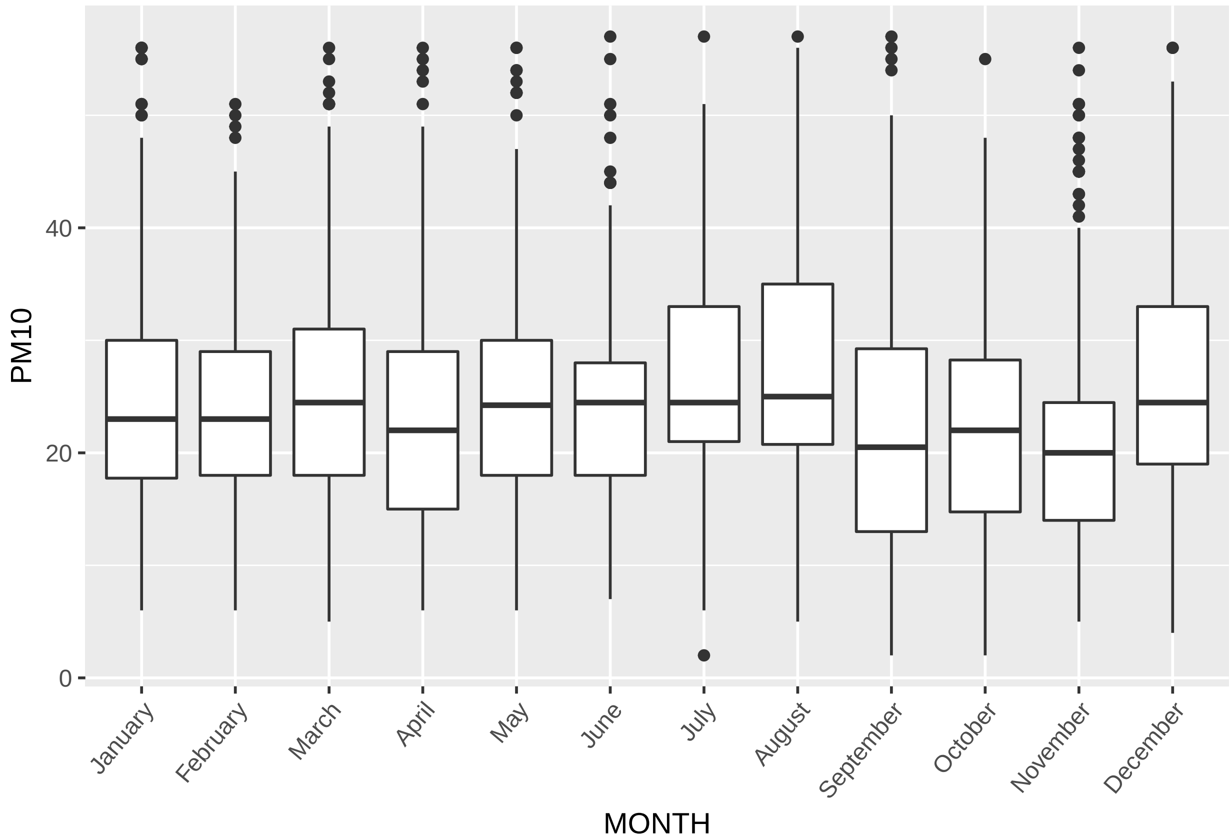

```
ggplot(Canarydataset, aes(x=MONTH, y=O3)) + geom_boxplot()+theme(axis.text.x = element_text(angle = 50,
```



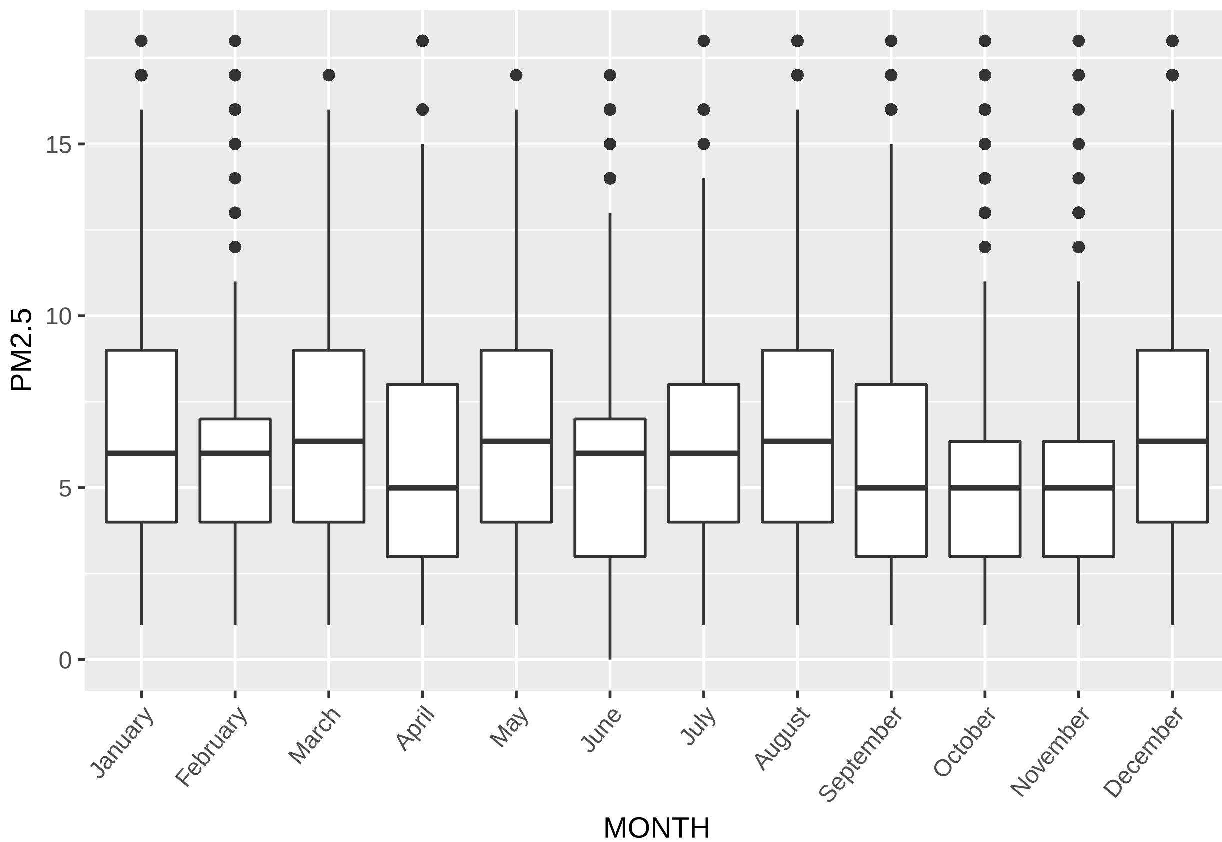

```
ggplot(Canarydataset, aes(x=MONTH, y=CO)) + geom_boxplot()+theme(axis.text.x = element_text(angle = 50,
```



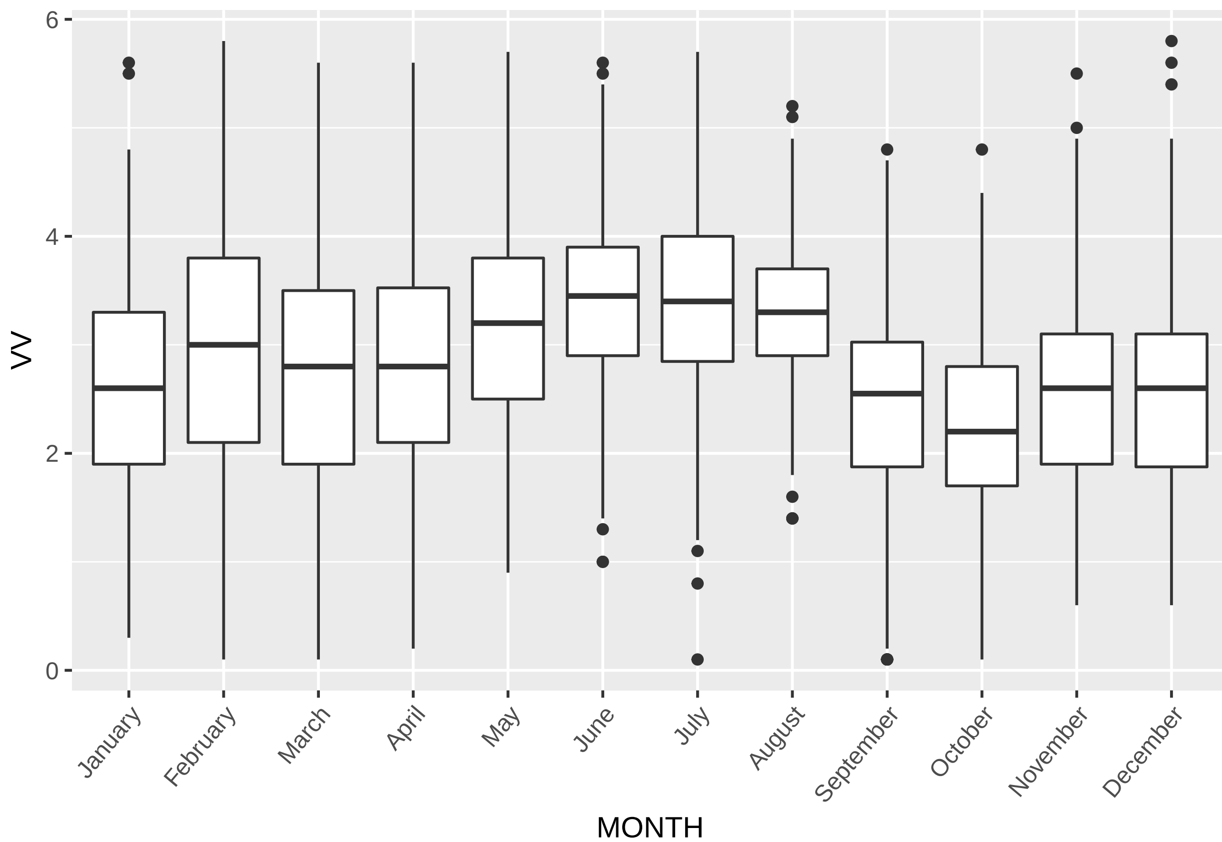

```
ggplot(Canarydataset, aes(x=MONTH, y=PM10)) + geom_boxplot()+theme(axis.text.x = element_text(angle = 50
```



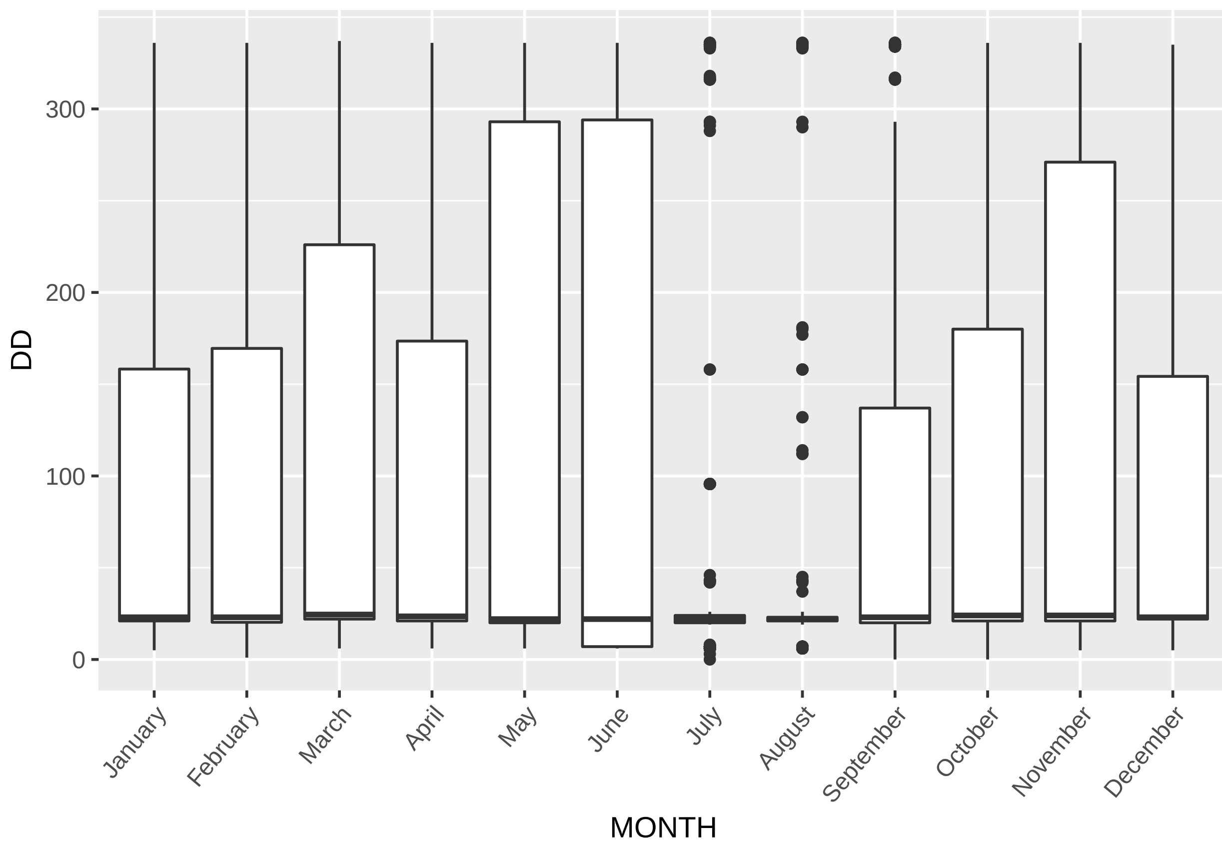

```
ggplot(Canarydataset, aes(x=MONTH, y=PM2.5)) + geom_boxplot()+theme(axis.text.x = element_text(angle = 5
```



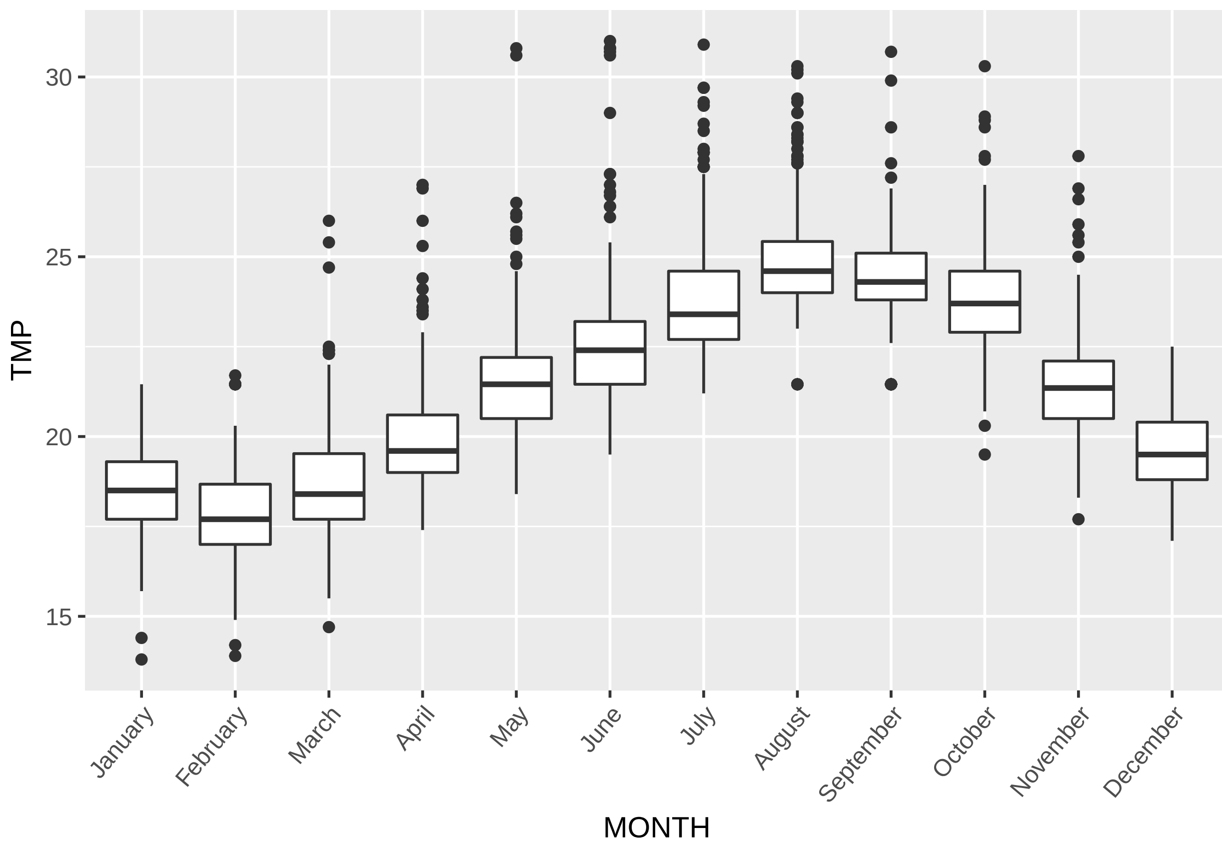

```
ggplot(Canarydataset, aes(x=MONTH, y=VV)) + geom_boxplot()+theme(axis.text.x = element_text(angle = 50,
```



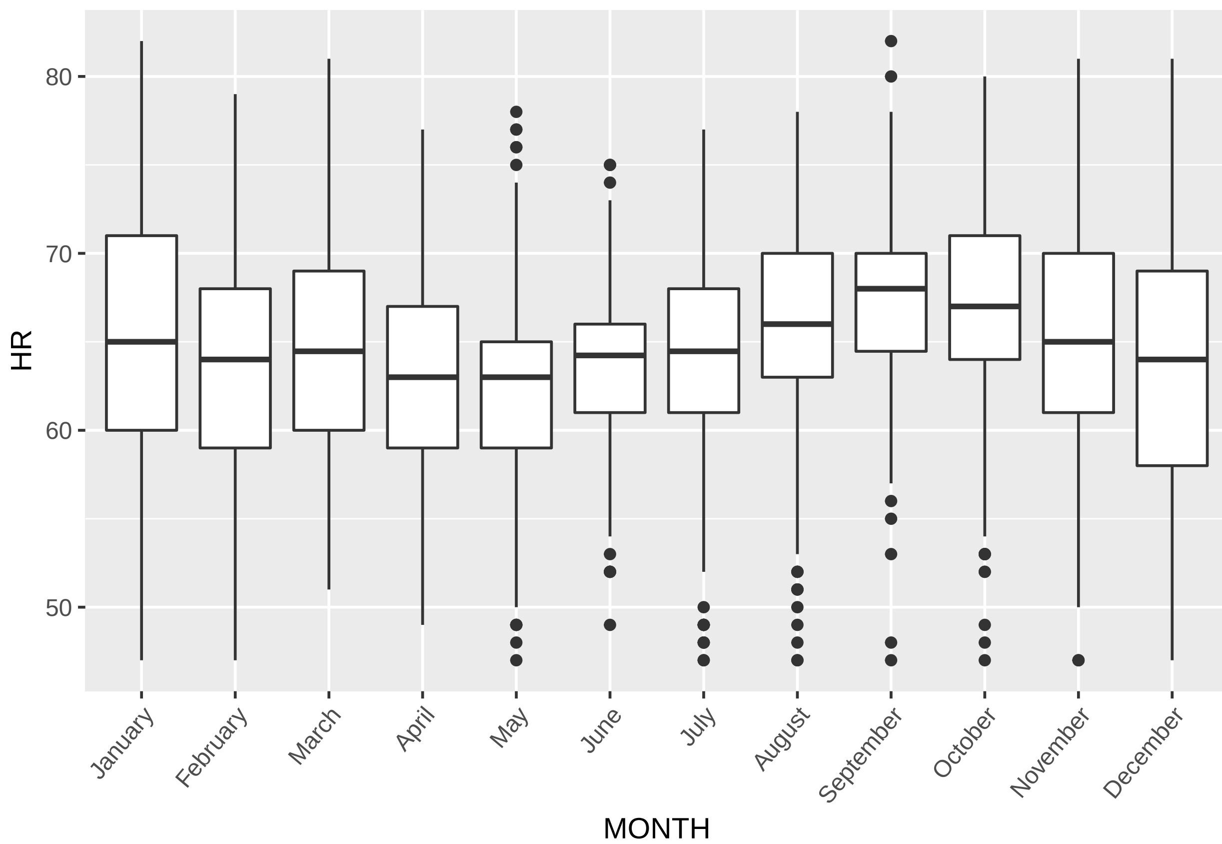

```
ggplot(Canarydataset, aes(x=MONTH, y=DD)) + geom_boxplot()+theme(axis.text.x = element_text(angle = 50,
```



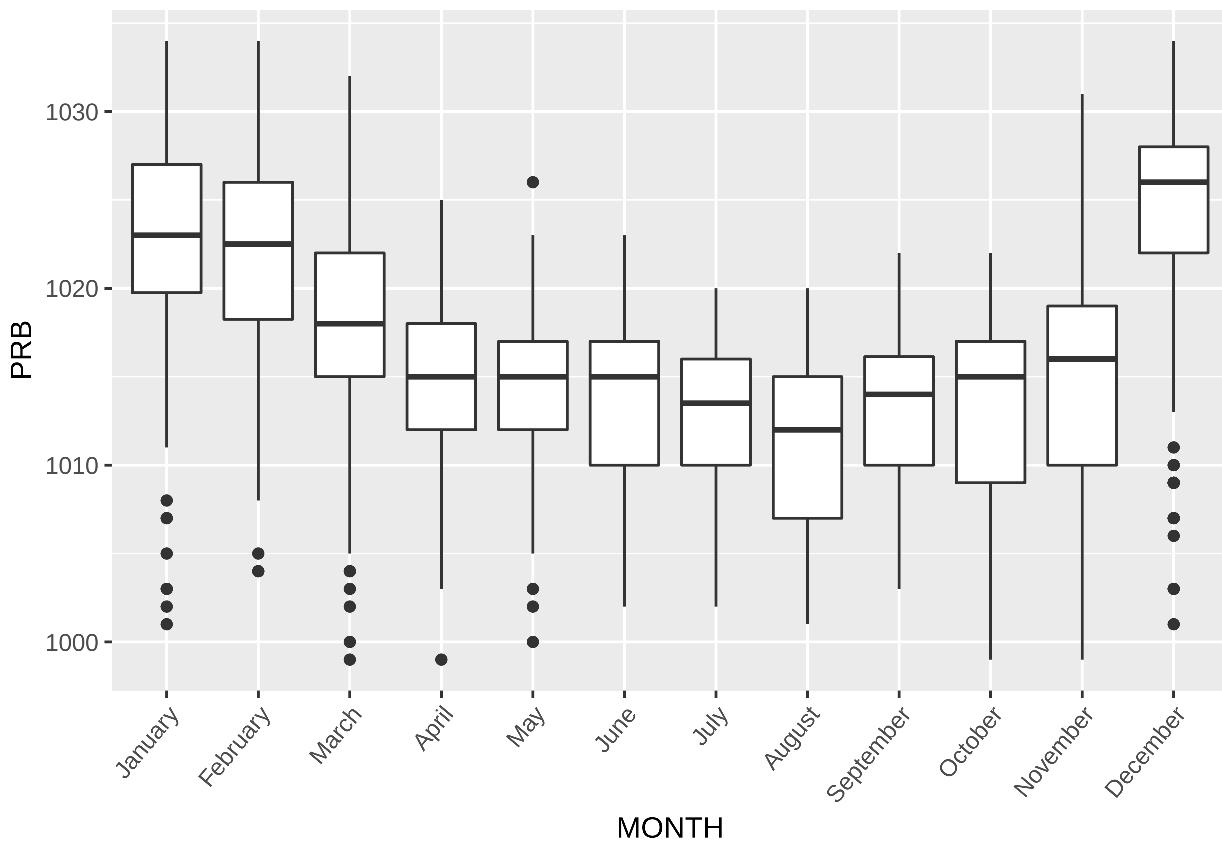

```
ggplot(Canarydataset, aes(x=MONTH, y=TMP)) + geom_boxplot()+theme(axis.text.x = element_text(angle = 50
```




```
ggplot(Canarydataset, aes(x=MONTH, y=HR)) + geom_boxplot()+theme(axis.text.x = element_text(angle = 50,
```




```
ggplot(Canarydataset, aes(x=MONTH, y=PRB)) + geom_boxplot()+theme(axis.text.x = element_text(angle = 50
```



Normality test

```
Canarydataset_NUME <- Canarydataset[,4:16]
MVN::mvn(Canarydataset_NUME)
```

```
## $multivariateNormality
##              Test        Statistic p value Result
## 1 Mardia Skewness 10466.8611916701       0     NO
## 2 Mardia Kurtosis 71.7347131098657       0     NO
## 3             MVN             <NA>    <NA>     NO
##
## $univariateNormality
##            Test  Variable Statistic   p value Normality
## 1  Shapiro-Wilk    SO2       0.8458  <0.001      NO
## 2  Shapiro-Wilk    NO        0.9556  <0.001      NO
## 3  Shapiro-Wilk    NO2       0.9690  <0.001      NO
## 4  Shapiro-Wilk    NOX       0.9754  <0.001      NO
## 5  Shapiro-Wilk    O3        0.9959  <0.001      NO
## 6  Shapiro-Wilk    CO        0.7961  <0.001      NO
## 7  Shapiro-Wilk   PM10       0.9629  <0.001      NO
## 8  Shapiro-Wilk   PM2.5      0.9174  <0.001      NO
## 9  Shapiro-Wilk    VV        0.9965  <0.001      NO
## 10 Shapiro-Wilk    DD        0.6785  <0.001      NO
```

```
## 11 Shapiro-Wilk    TMP       0.9842  <0.001      NO
## 12 Shapiro-Wilk    HR        0.9907  <0.001      NO
## 13 Shapiro-Wilk    PRB       0.9882  <0.001      NO
##
## $Descriptives
##          n         Mean      Std.Dev       Median   Min    Max     25th
## SO2   2922    3.1605979   2.32292870    3.0000000   1.0   11.0    1.000
## NO    2922    4.3696993   1.90752114    4.3696993   1.0   10.0    3.000
## NO2   2922   10.7337437   4.23940170   10.7337437   1.0   23.0    8.000
## NOX   2922   17.1933100   6.21614900   17.1933100   3.0   35.0   13.000
## O3    2922   72.1123596  14.78001612   72.1123596  31.0  114.0   62.000
## CO    2922    0.1601531   0.06443907    0.1601531   0.1    0.3    0.100
## PM10  2922   24.4685583  10.37049745   24.0000000   2.0   57.0   17.000
## PM2.5 2922    6.3471237   3.68640572    6.0000000   0.0   18.0    3.000
## VV    2922    2.8469067   1.00996492    2.9000000   0.1    5.8    2.125
## DD    2922   95.6127006 121.06461461   23.0000000   0.0  337.0   21.000
## TMP   2922   21.4533479   2.81879417   21.4533479  13.8   31.0   19.100
## HR    2922   64.4642195   6.33750073   64.4642195  47.0   82.0   61.000
## PRB   2922 1016.1312478   6.82598030 1016.0000000 999.0 1034.0 1012.000
##         75th       Skew    Kurtosis
## SO2      4.0  1.2031471  1.02523820
## NO       5.0  0.4340983  0.30595474
## NO2     13.0  0.5606058  0.34240594
## NOX     20.0  0.4135189  0.33479589
## O3      83.0 -0.1177389 -0.26132935
## CO       0.2  0.7903697 -0.27309776
## PM10    30.0  0.7032114  0.30755882
## PM2.5    8.0  0.9883782  0.62766976
## VV       3.5 -0.1166183 -0.12620275
## DD     158.0  1.1592784 -0.37606293
## TMP     23.7  0.2745705 -0.33750789
## HR      69.0 -0.2625285 -0.09582762
## PRB   1020.0  0.1797873 -0.24435952
```

Homogeneity test

YEAR

```
fligner.test(Canarydataset$SO2,Canarydataset$YEAR)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Canarydataset$SO2 and Canarydataset$YEAR
## Fligner-Killeen:med chi-squared = 484.83, df = 7, p-value <
## 2.2e-16
```

```
fligner.test(Canarydataset$NO,Canarydataset$YEAR)
```

```
##
```

```
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Canarydataset$NO and Canarydataset$YEAR
## Fligner-Killeen:med chi-squared = 165.62, df = 7, p-value <
## 2.2e-16
```

```r
fligner.test(Canarydataset$NO2,Canarydataset$YEAR)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Canarydataset$NO2 and Canarydataset$YEAR
## Fligner-Killeen:med chi-squared = 87.335, df = 7, p-value =
## 4.357e-16
```

```r
fligner.test(Canarydataset$NOX,Canarydataset$YEAR)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Canarydataset$NOX and Canarydataset$YEAR
## Fligner-Killeen:med chi-squared = 70.553, df = 7, p-value =
## 1.142e-12
```

```r
fligner.test(Canarydataset$O3,Canarydataset$YEAR)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Canarydataset$O3 and Canarydataset$YEAR
## Fligner-Killeen:med chi-squared = 58.16, df = 7, p-value =
## 3.513e-10
```

```r
fligner.test(Canarydataset$CO,Canarydataset$YEAR)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Canarydataset$CO and Canarydataset$YEAR
## Fligner-Killeen:med chi-squared = 88.829, df = 7, p-value <
## 2.2e-16
```

```r
fligner.test(Canarydataset$PM10,Canarydataset$YEAR)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Canarydataset$PM10 and Canarydataset$YEAR
## Fligner-Killeen:med chi-squared = 17.873, df = 7, p-value =
## 0.01256
```

```r
fligner.test(Canarydataset$PM2.5,Canarydataset$YEAR)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Canarydataset$PM2.5 and Canarydataset$YEAR
## Fligner-Killeen:med chi-squared = 191.29, df = 7, p-value <
## 2.2e-16
```

```r
fligner.test(Canarydataset$VV,Canarydataset$YEAR)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Canarydataset$VV and Canarydataset$YEAR
## Fligner-Killeen:med chi-squared = 63.402, df = 7, p-value =
## 3.148e-11
```

```r
fligner.test(Canarydataset$DD,Canarydataset$YEAR)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Canarydataset$DD and Canarydataset$YEAR
## Fligner-Killeen:med chi-squared = 548.27, df = 7, p-value <
## 2.2e-16
```

```r
fligner.test(Canarydataset$TMP,Canarydataset$YEAR)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Canarydataset$TMP and Canarydataset$YEAR
## Fligner-Killeen:med chi-squared = 82.17, df = 7, p-value =
## 4.968e-15
```

```r
fligner.test(Canarydataset$HR,Canarydataset$YEAR)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Canarydataset$HR and Canarydataset$YEAR
## Fligner-Killeen:med chi-squared = 51.948, df = 7, p-value =
## 5.977e-09
```

```r
fligner.test(Canarydataset$PRB,Canarydataset$YEAR)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Canarydataset$PRB and Canarydataset$YEAR
## Fligner-Killeen:med chi-squared = 196.83, df = 7, p-value <
## 2.2e-16
```

MONTH

```r
fligner.test(Canarydataset$SO2,Canarydataset$MONTH)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Canarydataset$SO2 and Canarydataset$MONTH
## Fligner-Killeen:med chi-squared = 46.776, df = 11, p-value =
## 2.357e-06
```

```r
fligner.test(Canarydataset$NO,Canarydataset$MONTH)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Canarydataset$NO and Canarydataset$MONTH
## Fligner-Killeen:med chi-squared = 33.019, df = 11, p-value =
## 0.0005224
```

```r
fligner.test(Canarydataset$NO2,Canarydataset$MONTH)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Canarydataset$NO2 and Canarydataset$MONTH
## Fligner-Killeen:med chi-squared = 99.477, df = 11, p-value =
## 2.266e-16
```

```r
fligner.test(Canarydataset$NOX,Canarydataset$MONTH)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Canarydataset$NOX and Canarydataset$MONTH
## Fligner-Killeen:med chi-squared = 36.073, df = 11, p-value =
## 0.0001646
```

```r
fligner.test(Canarydataset$O3,Canarydataset$MONTH)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Canarydataset$O3 and Canarydataset$MONTH
## Fligner-Killeen:med chi-squared = 26.017, df = 11, p-value =
## 0.006451
```

```r
fligner.test(Canarydataset$CO,Canarydataset$MONTH)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Canarydataset$CO and Canarydataset$MONTH
## Fligner-Killeen:med chi-squared = 69.012, df = 11, p-value =
## 1.881e-10
```

```r
fligner.test(Canarydataset$PM10,Canarydataset$MONTH)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Canarydataset$PM10 and Canarydataset$MONTH
## Fligner-Killeen:med chi-squared = 43.01, df = 11, p-value =
## 1.082e-05
```

```r
fligner.test(Canarydataset$PM2.5,Canarydataset$MONTH)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Canarydataset$PM2.5 and Canarydataset$MONTH
## Fligner-Killeen:med chi-squared = 24.026, df = 11, p-value =
## 0.01262
```

```r
fligner.test(Canarydataset$VV,Canarydataset$MONTH)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Canarydataset$VV and Canarydataset$MONTH
## Fligner-Killeen:med chi-squared = 129.82, df = 11, p-value <
## 2.2e-16
```

```r
fligner.test(Canarydataset$DD,Canarydataset$MONTH)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Canarydataset$DD and Canarydataset$MONTH
## Fligner-Killeen:med chi-squared = 87.189, df = 11, p-value =
## 5.915e-14
```

```r
fligner.test(Canarydataset$TMP,Canarydataset$MONTH)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
```

```
##
## data:  Canarydataset$TMP and Canarydataset$MONTH
## Fligner-Killeen:med chi-squared = 37.608, df = 11, p-value =
## 9.11e-05
```

```r
fligner.test(Canarydataset$HR,Canarydataset$MONTH)
```

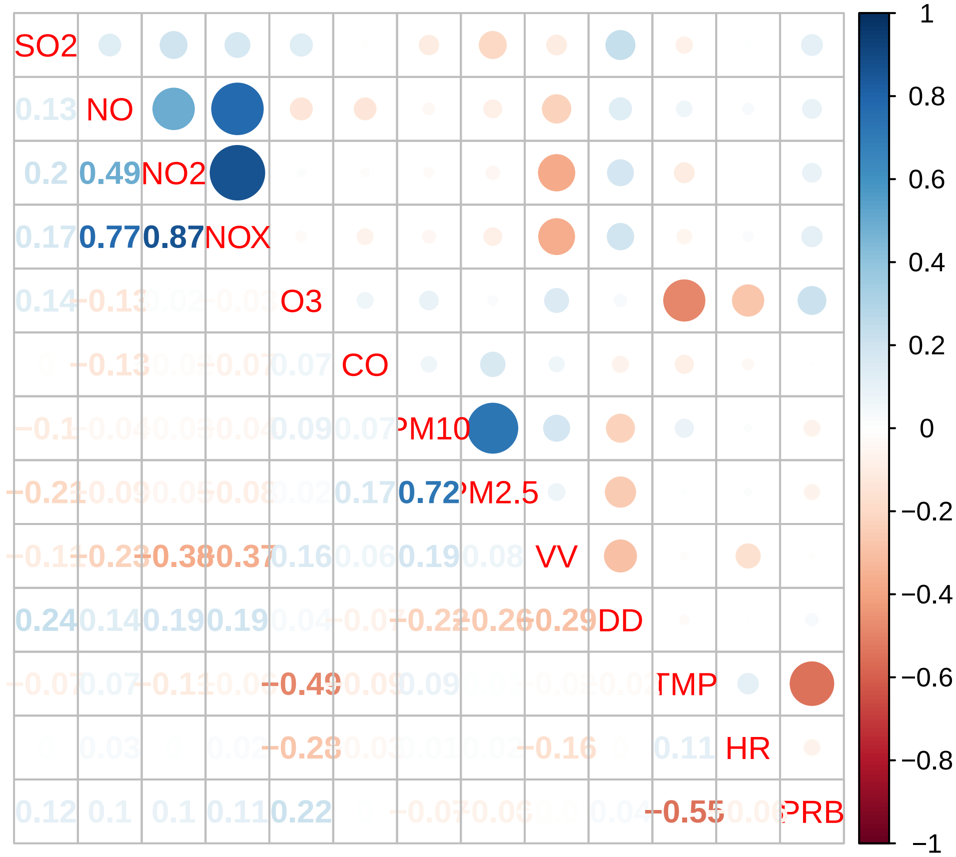
```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Canarydataset$HR and Canarydataset$MONTH
## Fligner-Killeen:med chi-squared = 129.98, df = 11, p-value <
## 2.2e-16
```

```r
fligner.test(Canarydataset$PRB,Canarydataset$MONTH)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Canarydataset$PRB and Canarydataset$MONTH
## Fligner-Killeen:med chi-squared = 65.106, df = 11, p-value =
## 1.029e-09
```

Correlation analysis

```r
numerical_items <- Canarydataset[,c("SO2","NO","NO2","NOX","O3","CO","PM10","PM2.5","VV","DD","TMP","HR
correlation_numerical_items <- cor(numerical_items,method =  "spearman")
corrplot.mixed(correlation_numerical_items, order = "original")
```



Kruskal-Wallis test

YEAR

```r
kruskal.test(SO2 ~ YEAR,data= Canarydataset)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  SO2 by YEAR
## Kruskal-Wallis chi-squared = 1176.6, df = 7, p-value < 2.2e-16
```

```r
kruskal.test(NO ~ YEAR,data= Canarydataset)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  NO by YEAR
## Kruskal-Wallis chi-squared = 478, df = 7, p-value < 2.2e-16
```

```r
kruskal.test(NO2 ~ YEAR,data= Canarydataset)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  NO2 by YEAR
## Kruskal-Wallis chi-squared = 229.44, df = 7, p-value < 2.2e-16
```

```r
kruskal.test(NOX ~ YEAR,data= Canarydataset)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  NOX by YEAR
## Kruskal-Wallis chi-squared = 369.87, df = 7, p-value < 2.2e-16
```

```r
kruskal.test(CO ~ YEAR,data= Canarydataset)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  CO by YEAR
## Kruskal-Wallis chi-squared = 631.74, df = 7, p-value < 2.2e-16
```

```r
kruskal.test(O3 ~ YEAR,data= Canarydataset)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  O3 by YEAR
## Kruskal-Wallis chi-squared = 116.38, df = 7, p-value < 2.2e-16
```

```r
kruskal.test(PM10 ~ YEAR,data= Canarydataset)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  PM10 by YEAR
## Kruskal-Wallis chi-squared = 288.07, df = 7, p-value < 2.2e-16
```

```r
kruskal.test(PM2.5 ~ YEAR,data= Canarydataset)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  PM2.5 by YEAR
## Kruskal-Wallis chi-squared = 875.5, df = 7, p-value < 2.2e-16
```

```r
kruskal.test(VV ~ YEAR,data= Canarydataset)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  VV by YEAR
## Kruskal-Wallis chi-squared = 151.66, df = 7, p-value < 2.2e-16
```

```r
kruskal.test(DD ~ YEAR,data= Canarydataset)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  DD by YEAR
## Kruskal-Wallis chi-squared = 737.22, df = 7, p-value < 2.2e-16
```

```r
kruskal.test(TMP ~ YEAR,data= Canarydataset)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  TMP by YEAR
## Kruskal-Wallis chi-squared = 49.284, df = 7, p-value = 1.996e-08
```

```r
kruskal.test(HR ~ YEAR,data= Canarydataset)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  HR by YEAR
## Kruskal-Wallis chi-squared = 53.625, df = 7, p-value = 2.79e-09
```

```r
kruskal.test(PRB ~ YEAR,data= Canarydataset)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  PRB by YEAR
## Kruskal-Wallis chi-squared = 195.25, df = 7, p-value < 2.2e-16
```

MONTH

```r
kruskal.test(SO2 ~ MONTH,data= Canarydataset)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  SO2 by MONTH
## Kruskal-Wallis chi-squared = 44.339, df = 11, p-value = 6.339e-06
```

```r
kruskal.test(NO ~ MONTH,data= Canarydataset)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  NO by MONTH
## Kruskal-Wallis chi-squared = 73.909, df = 11, p-value = 2.193e-11
```

```r
kruskal.test(NO2 ~ MONTH,data= Canarydataset)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  NO2 by MONTH
## Kruskal-Wallis chi-squared = 302.1, df = 11, p-value < 2.2e-16
```

```r
kruskal.test(NOX ~ MONTH,data= Canarydataset)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  NOX by MONTH
## Kruskal-Wallis chi-squared = 210.69, df = 11, p-value < 2.2e-16
```

```r
kruskal.test(CO ~ MONTH,data= Canarydataset)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  CO by MONTH
## Kruskal-Wallis chi-squared = 100.35, df = 11, p-value < 2.2e-16
```

```r
kruskal.test(O3 ~ MONTH,data= Canarydataset)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  O3 by MONTH
## Kruskal-Wallis chi-squared = 765.75, df = 11, p-value < 2.2e-16
```

```r
kruskal.test(PM10 ~ MONTH,data= Canarydataset)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  PM10 by MONTH
## Kruskal-Wallis chi-squared = 107.31, df = 11, p-value < 2.2e-16
```

```r
kruskal.test(PM2.5 ~ MONTH,data= Canarydataset)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  PM2.5 by MONTH
## Kruskal-Wallis chi-squared = 77.617, df = 11, p-value = 4.256e-12
```

```r
kruskal.test(VV ~ MONTH,data= Canarydataset)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  VV by MONTH
## Kruskal-Wallis chi-squared = 476.66, df = 11, p-value < 2.2e-16
```

```r
kruskal.test(DD ~ MONTH,data= Canarydataset)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  DD by MONTH
## Kruskal-Wallis chi-squared = 75.879, df = 11, p-value = 9.192e-12
```

```r
kruskal.test(TMP ~ MONTH,data= Canarydataset)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  TMP by MONTH
## Kruskal-Wallis chi-squared = 2244.7, df = 11, p-value < 2.2e-16
```

```r
kruskal.test(HR ~ MONTH,data= Canarydataset)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  HR by MONTH
## Kruskal-Wallis chi-squared = 171.35, df = 11, p-value < 2.2e-16
```

```r
kruskal.test(PRB ~ MONTH,data= Canarydataset)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  PRB by MONTH
## Kruskal-Wallis chi-squared = 1120.2, df = 11, p-value < 2.2e-16
```

Spliting dataset

```r
#Establish a training set and a verification set.
set.seed(1)
sample <- sample.int(n = nrow(Canarydataset), size = floor(0.50*nrow(Canarydataset)), replace = F)
Canarydataset_training <- Canarydataset[sample, ]
Canarydataset_verification  <- Canarydataset[-sample, ]
```

Linal regression models

Model1(Dependent variable: SO2)

```r
lineal_model_SO2 <- lm(log(SO2+1) ~ NO+NO2+NOX+O3+PM2.5+VV,data=Canarydataset_training)
summary(gvlma(lineal_model_SO2))
```

```
##
## Call:
## lm(formula = log(SO2 + 1) ~ NO + NO2 + NOX + O3 + PM2.5 + VV,
##     data = Canarydataset_training)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.07015 -0.41442 -0.03956  0.35739  1.35229
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.8851181  0.0892710   9.915  < 2e-16 ***
## NO           0.0168947  0.0111501   1.515    0.130
## NO2          0.0295485  0.0062532   4.725 2.52e-06 ***
## NOX         -0.0075556  0.0055913  -1.351    0.177
## O3           0.0049183  0.0008621   5.705 1.41e-08 ***
## PM2.5       -0.0301385  0.0035325  -8.532  < 2e-16 ***
## VV          -0.0040727  0.0139775  -0.291    0.771
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4921 on 1454 degrees of freedom
## Multiple R-squared:  0.1068, Adjusted R-squared:  0.1031
## F-statistic: 28.97 on 6 and 1454 DF,  p-value: < 2.2e-16
##
##
## ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
## USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
## Level of Significance =  0.05
##
## Call:
##  gvlma(x = lineal_model_SO2)
##
##                        Value   p-value                   Decision
```

```
## Global Stat        64.1341 3.916e-13 Assumptions NOT satisfied!
## Skewness           20.0849 7.408e-06 Assumptions NOT satisfied!
## Kurtosis           36.3671 1.634e-09 Assumptions NOT satisfied!
## Link Function       7.5235 6.090e-03 Assumptions NOT satisfied!
## Heteroscedasticity  0.1587 6.904e-01    Assumptions acceptable.
```

Model2(Dependent variable: NO)

```
lineal_model_NO <- lm(log(NO+1) ~ NO2+NOX+O3+HR+PRB,data=Canarydataset_training)
summary(gvlma(lineal_model_NO))
```

```
##
## Call:
## lm(formula = log(NO + 1) ~ NO2 + NOX + O3 + HR + PRB, data = Canarydataset_training)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.26189 -0.08958  0.03640  0.12498  1.21646
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.9053088  0.9463421  -3.070  0.00218 **
## NO2         -0.0432431  0.0027279 -15.852  < 2e-16 ***
## NOX          0.0714522  0.0018848  37.910  < 2e-16 ***
## O3          -0.0019214  0.0004251  -4.520 6.69e-06 ***
## HR          -0.0027705  0.0009942  -2.787  0.00539 **
## PRB          0.0040041  0.0009329   4.292 1.89e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2356 on 1455 degrees of freedom
## Multiple R-squared:  0.6321, Adjusted R-squared:  0.6309
## F-statistic:   500 on 5 and 1455 DF,  p-value: < 2.2e-16
##
##
## ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
## USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
## Level of Significance =  0.05
##
## Call:
##  gvlma(x = lineal_model_NO)
##
##                      Value p-value                   Decision
## Global Stat        2088.37  0.0000 Assumptions NOT satisfied!
## Skewness            133.37  0.0000 Assumptions NOT satisfied!
## Kurtosis           1410.27  0.0000 Assumptions NOT satisfied!
## Link Function       543.15  0.0000 Assumptions NOT satisfied!
## Heteroscedasticity    1.59  0.2073    Assumptions acceptable.
```

Model3(Dependent variable: NO2)

```
lineal_model_NO2 <- lm(log(NO2+1) ~ SO2+NO+NOX+O3+HR+PRB,data=Canarydataset_training)
summary(gvlma(lineal_model_NO2))
```

```
##
## Call:
## lm(formula = log(NO2 + 1) ~ SO2 + NO + NOX + O3 + HR + PRB, data = Canarydataset_training)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.05732 -0.06609  0.02324  0.08944  1.02393
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.7549985  0.8099158   3.402 0.000688 ***
## SO2          0.0123548  0.0023169   5.333 1.12e-07 ***
## NO          -0.0525854  0.0041458 -12.684  < 2e-16 ***
## NOX          0.0660137  0.0012636  52.241  < 2e-16 ***
## O3          -0.0003627  0.0003698  -0.981 0.326768
## HR          -0.0021504  0.0008482  -2.535 0.011340 *
## PRB         -0.0011274  0.0007987  -1.412 0.158277
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2008 on 1454 degrees of freedom
## Multiple R-squared:  0.7492, Adjusted R-squared:  0.7482
## F-statistic:   724 on 6 and 1454 DF,  p-value: < 2.2e-16
##
##
## ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
## USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
## Level of Significance =  0.05
##
## Call:
##  gvlma(x = lineal_model_NO2)
##
##                       Value p-value                   Decision
## Global Stat        5503.0739  0.0000 Assumptions NOT satisfied!
## Skewness            303.5701  0.0000 Assumptions NOT satisfied!
## Kurtosis           4467.4699  0.0000 Assumptions NOT satisfied!
## Link Function       731.0475  0.0000 Assumptions NOT satisfied!
## Heteroscedasticity    0.9864  0.3206    Assumptions acceptable.
```

Model4(Dependent variable: NOX)

```
lineal_model_NOX <- lm(log(NOX+1) ~ SO2+NO+NO2+O3+HR+PRB,data=Canarydataset_training)
summary(gvlma(lineal_model_NOX))
```

```
##
```

```
## Call:
## lm(formula = log(NOX + 1) ~ SO2 + NO + NO2 + O3 + HR + PRB, data = Canarydataset_training)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.02745 -0.02975  0.03083  0.06433  0.77306
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.3779839  0.5960215   3.990 6.94e-05 ***
## SO2          0.0038516  0.0017133   2.248  0.02473 *
## NO           0.0919798  0.0023619  38.943  < 2e-16 ***
## NO2          0.0549963  0.0010462  52.566  < 2e-16 ***
## O3           0.0007439  0.0002719   2.736  0.00629 **
## HR           0.0013757  0.0006237   2.205  0.02758 *
## PRB         -0.0006771  0.0005879  -1.152  0.24964
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1478 on 1454 degrees of freedom
## Multiple R-squared:  0.8501, Adjusted R-squared:  0.8495
## F-statistic:  1375 on 6 and 1454 DF,  p-value: < 2.2e-16
##
##
## ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
## USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
## Level of Significance =  0.05
##
## Call:
##  gvlma(x = lineal_model_NOX)
##
##                       Value p-value                   Decision
## Global Stat        8373.416 0.00000 Assumptions NOT satisfied!
## Skewness            612.061 0.00000 Assumptions NOT satisfied!
## Kurtosis           6937.773 0.00000 Assumptions NOT satisfied!
## Link Function       818.484 0.00000 Assumptions NOT satisfied!
## Heteroscedasticity    5.099 0.02394 Assumptions NOT satisfied!
```

Model5(Dependent variable: O3)

```
lineal_model_O3 <- lm(log(O3+1) ~ SO2+NO+NO2+NOX+HR+PRB,data=Canarydataset_training)
summary(gvlma(lineal_model_O3))
```

```
##
## Call:
## lm(formula = log(O3 + 1) ~ SO2 + NO + NO2 + NOX + HR + PRB, data = Canarydataset_training)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.78739 -0.11344  0.02508  0.13546  0.49205
##
## Coefficients:
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.2494489  0.8428801  -0.296  0.76731
## SO2          0.0138578  0.0024084   5.754 1.06e-08 ***
## NO          -0.0290192  0.0047109  -6.160 9.40e-10 ***
## NO2         -0.0071992  0.0026779  -2.688  0.00726 **
## NOX          0.0074973  0.0023784   3.152  0.00165 **
## HR          -0.0083958  0.0008577  -9.788  < 2e-16 ***
## PRB          0.0050082  0.0008257   6.065 1.68e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2101 on 1454 degrees of freedom
## Multiple R-squared:  0.1293, Adjusted R-squared:  0.1257
## F-statistic: 35.99 on 6 and 1454 DF,  p-value: < 2.2e-16
##
##
## ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
## USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
## Level of Significance =  0.05
##
## Call:
##  gvlma(x = lineal_model_O3)
##
##                      Value   p-value                   Decision
## Global Stat        198.254 0.000e+00 Assumptions NOT satisfied!
## Skewness           145.352 0.000e+00 Assumptions NOT satisfied!
## Kurtosis            49.080 2.457e-12 Assumptions NOT satisfied!
## Link Function        3.701 5.438e-02    Assumptions acceptable.
## Heteroscedasticity   0.121 7.280e-01    Assumptions acceptable.
```

Model6(Dependent variable: CO)

```r
lineal_model_CO <- lm(log(CO+1) ~ NO+NO2+NOX+PM2.5+DD+HR,data=Canarydataset_training)
summary(gvlma(lineal_model_CO))
```

```
##
## Call:
## lm(formula = log(CO + 1) ~ NO + NO2 + NOX + PM2.5 + DD + HR,
##     data = Canarydataset_training)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.086842 -0.046737  0.000505  0.034351  0.143194
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.580e-01  1.514e-02  10.435   <2e-16 ***
## NO          -2.423e-03  1.208e-03  -2.006    0.045 *
## NO2          1.070e-03  6.823e-04   1.568    0.117
## NOX         -4.150e-04  6.089e-04  -0.682    0.496
## PM2.5        2.272e-03  3.941e-04   5.764    1e-08 ***
## DD           9.657e-06  1.195e-05   0.808    0.419
```

```
## HR          -3.274e-04  2.200e-04  -1.488    0.137
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05377 on 1454 degrees of freedom
## Multiple R-squared:  0.03607,    Adjusted R-squared:  0.03209
## F-statistic: 9.068 on 6 and 1454 DF,  p-value: 9.352e-10
##
##
## ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
## USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
## Level of Significance =  0.05
##
## Call:
##  gvlma(x = lineal_model_CO)
##
##                       Value p-value                   Decision
## Global Stat        138.6251 0.00000 Assumptions NOT satisfied!
## Skewness           130.5753 0.00000 Assumptions NOT satisfied!
## Kurtosis             5.3040 0.02128 Assumptions NOT satisfied!
## Link Function        0.1715 0.67878    Assumptions acceptable.
## Heteroscedasticity   2.5742 0.10862    Assumptions acceptable.
```

Model7(Dependent variable: PM10)

```
lineal_model_PM10 <- lm(log(PM10+1) ~ NO+PM2.5+DD+TMP+HR+PRB,data=Canarydataset_training)
summary(gvlma(lineal_model_PM10))
```

```
##
## Call:
## lm(formula = log(PM10 + 1) ~ NO + PM2.5 + DD + TMP + HR + PRB,
##     data = Canarydataset_training)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.67770 -0.20347  0.03429  0.21679  0.91571
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.3885002  1.6058915   1.487 0.137143
## NO           0.0090540  0.0047053   1.924 0.054521 .
## PM2.5        0.0711554  0.0024218  29.381  < 2e-16 ***
## DD          -0.0003418  0.0000735  -4.651  3.6e-06 ***
## TMP          0.0137128  0.0036332   3.774 0.000167 ***
## HR          -0.0018089  0.0013538  -1.336 0.181709
## PRB          0.0001280  0.0015328   0.083 0.933482
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3303 on 1454 degrees of freedom
## Multiple R-squared:  0.416,  Adjusted R-squared:  0.4135
## F-statistic: 172.6 on 6 and 1454 DF,  p-value: < 2.2e-16
```

```
##
##
## ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
## USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
## Level of Significance =  0.05
##
## Call:
##  gvlma(x = lineal_model_PM10)
##
##                         Value   p-value                   Decision
## Global Stat        472.6976 0.000e+00 Assumptions NOT satisfied!
## Skewness            61.5222 4.330e-15 Assumptions NOT satisfied!
## Kurtosis           115.0167 0.000e+00 Assumptions NOT satisfied!
## Link Function      295.6639 0.000e+00 Assumptions NOT satisfied!
## Heteroscedasticity   0.4948 4.818e-01    Assumptions acceptable.
```

Model8(Dependent variable: PM2.5)

```r
lineal_model_PM2.5 <- lm(log(PM2.5+1)~ SO2+PM10+VV+DD+TMP,data=Canarydataset_training)
summary(gvlma(lineal_model_PM2.5))
```

```
##
## Call:
## lm(formula = log(PM2.5 + 1) ~ SO2 + PM10 + VV + DD + TMP, data = Canarydataset_training)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.47588 -0.24591 -0.01433  0.19908  1.12132
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.516e+00  8.191e-02  18.510  < 2e-16 ***
## SO2         -2.488e-02  4.080e-03  -6.100 1.36e-09 ***
## PM10         3.160e-02  9.486e-04  33.316  < 2e-16 ***
## VV          -2.068e-02  9.863e-03  -2.097  0.03617 *
## DD          -4.210e-04  8.198e-05  -5.135 3.20e-07 ***
## TMP         -1.074e-02  3.370e-03  -3.186  0.00147 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.359 on 1455 degrees of freedom
## Multiple R-squared:  0.4829, Adjusted R-squared:  0.4811
## F-statistic: 271.7 on 5 and 1455 DF,  p-value: < 2.2e-16
##
##
## ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
## USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
## Level of Significance =  0.05
##
## Call:
##  gvlma(x = lineal_model_PM2.5)
##
```

```
##                      Value   p-value                   Decision
## Global Stat        239.744 0.000e+00 Assumptions NOT satisfied!
## Skewness            16.053 6.158e-05 Assumptions NOT satisfied!
## Kurtosis             3.846 4.987e-02 Assumptions NOT satisfied!
## Link Function      214.911 0.000e+00 Assumptions NOT satisfied!
## Heteroscedasticity   4.934 2.633e-02 Assumptions NOT satisfied!
```

Generalized additive model(GAM)

Model1(Dependent variable: SO2)

```r
GAM_model_SO2 <- gam(SO2 ~ s(NO)+s(NO2)+s(NOX)+s(O3)+s(PM2.5)+s(VV),data=Canarydataset_training)
summary(GAM_model_SO2)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## SO2 ~ s(NO) + s(NO2) + s(NOX) + s(O3) + s(PM2.5) + s(VV)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.18996    0.05571   57.26   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##            edf Ref.df      F  p-value
## s(NO)    3.573  4.550 10.131 1.13e-08 ***
## s(NO2)   4.173  5.288  7.623 2.82e-07 ***
## s(NOX)   7.604  8.463  5.109 4.61e-06 ***
## s(O3)    5.661  6.783  6.886 1.20e-07 ***
## s(PM2.5) 1.000  1.000 39.800 3.70e-10 ***
## s(VV)    2.371  3.033  3.073   0.0266 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.169   Deviance explained = 18.3%
## GCV = 4.6149  Scale est. = 4.5348    n = 1461
```

Model2(Dependent variable: NO)

```r
GAM_model_NO <- gam(NO ~ s(NO2)+s(NOX)+s(O3)+s(HR)+s(PRB),data=Canarydataset_training)
summary(GAM_model_NO)
```

```
##
## Family: gaussian
```

```
## Link function: identity
##
## Formula:
## NO ~ s(NO2) + s(NOX) + s(O3) + s(HR) + s(PRB)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.3537     0.0238   182.9   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##          edf Ref.df       F  p-value
## s(NO2) 4.753  5.912 167.835  < 2e-16 ***
## s(NOX) 8.289  8.829 317.684  < 2e-16 ***
## s(O3)  1.907  2.415  14.843 7.66e-08 ***
## s(HR)  5.462  6.558   2.689 0.009673 **
## s(PRB) 5.737  6.872   4.268 0.000121 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.767   Deviance explained = 77.1%
## GCV = 0.84327  Scale est. = 0.8276    n = 1461
```

Model3(Dependent variable: NO2)

```
GAM_model_NO2 <- gam(NO2 ~ s(SO2)+s(NO)+s(NOX)+s(O3)+s(HR)+s(PRB),data=Canarydataset_training)
summary(GAM_model_NO2)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## NO2 ~ s(SO2) + s(NO) + s(NOX) + s(O3) + s(HR) + s(PRB)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.70595    0.04606   232.4   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##          edf Ref.df       F  p-value
## s(SO2) 7.608  8.459   4.319 4.48e-05 ***
## s(NO)  8.126  8.771  73.431  < 2e-16 ***
## s(NOX) 7.604  8.486 520.077  < 2e-16 ***
## s(O3)  3.357  4.215   6.418 3.01e-05 ***
## s(HR)  2.102  2.667   4.512 0.004758 **
## s(PRB) 6.977  8.013   3.414 0.000688 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## R-sq.(adj) =  0.828   Deviance explained = 83.3%
## GCV = 3.1794  Scale est. = 3.0994    n = 1461
```

Model4(Dependent variable: NOX)

```
GAM_model_NOX <- gam(NOX ~ s(SO2)+s(NO)+s(NO2)+s(O3)+s(HR)+s(PRB),data=Canarydataset_training)
summary(GAM_model_NOX)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## NOX ~ s(SO2) + s(NO) + s(NO2) + s(O3) + s(HR) + s(PRB)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  17.1587     0.0557   308.1   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##          edf Ref.df       F p-value
## s(SO2) 1.478  1.809   2.186  0.0810 .
## s(NO)  5.523  6.694 235.218  <2e-16 ***
## s(NO2) 7.507  8.421 445.337  <2e-16 ***
## s(O3)  3.865  4.815   2.454  0.0320 *
## s(HR)  1.000  1.000   3.323  0.0685 .
## s(PRB) 1.211  1.393   0.215  0.8013
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.881   Deviance explained = 88.3%
## GCV = 4.6006  Scale est. = 4.5326    n = 1461
```

Model5(Dependent variable: O3)

```
GAM_model_O3 <- gam(O3 ~ s(SO2)+s(NO)+s(NO2)+s(NOX)+s(HR)+s(PRB),data=Canarydataset_training)
summary(GAM_model_O3)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## O3 ~ s(SO2) + s(NO) + s(NO2) + s(NOX) + s(HR) + s(PRB)
##
## Parametric coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  72.0152     0.3596   200.2   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##           edf Ref.df      F  p-value
## s(SO2) 7.965  8.678  6.894 1.85e-09 ***
## s(NO)  4.165  5.223 10.627 2.97e-10 ***
## s(NO2) 1.000  1.000 14.477 0.000148 ***
## s(NOX) 8.192  8.809  6.678 2.52e-09 ***
## s(HR)  4.583  5.620 27.319  < 2e-16 ***
## s(PRB) 2.907  3.657  6.940 5.45e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =   0.19   Deviance explained = 20.6%
## GCV = 192.91  Scale est. = 188.97    n = 1461
```

Model6(Dependent variable: CO)

```
GAM_model_CO <- gam(CO ~ s(NO)+s(NO2)+s(NOX)+s(PM2.5)+s(DD)+s(HR),data=Canarydataset_training)
summary(GAM_model_CO)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## CO ~ s(NO) + s(NO2) + s(NOX) + s(PM2.5) + s(DD) + s(HR)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.159229   0.001635   97.42   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##             edf Ref.df     F  p-value
## s(NO)     5.388  6.573 1.016   0.3590
## s(NO2)    6.826  7.901 2.253   0.0251 *
## s(NOX)    4.852  6.044 2.012   0.0603 .
## s(PM2.5) 6.792  7.855 6.018 2.26e-07 ***
## s(DD)     1.000  1.000 0.219   0.6398
## s(HR)     1.457  1.793 1.996   0.1025
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.0659   Deviance explained = 8.27%
## GCV = 0.0039777  Scale est. = 0.0039033  n = 1461
```

Model7(Dependent variable: PM10)

```
GAM_model_PM10 <- gam(PM10 ~ s(NO)+s(PM2.5)+s(DD)+s(TMP)+s(HR)+s(PRB),data=Canarydataset_training)
summary(GAM_model_PM10)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## PM10 ~ s(NO) + s(PM2.5) + s(DD) + s(TMP) + s(HR) + s(PRB)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  24.7508     0.1888   131.1   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##            edf Ref.df       F  p-value
## s(NO)    3.652  4.558   1.304 0.258927
## s(PM2.5) 4.252  5.225 223.109  < 2e-16 ***
## s(DD)    2.093  2.512   4.218 0.007854 **
## s(TMP)   4.271  5.302   9.461 4.06e-09 ***
## s(HR)    6.489  7.548   2.709 0.007574 **
## s(PRB)   3.010  3.786   5.284 0.000532 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.517   Deviance explained = 52.5%
## GCV = 52.967  Scale est. = 52.069    n = 1461
```

Model8(Dependent variable: PM2.5)

```
GAM_model_PM2.5 <- gam(PM2.5 ~  s(SO2)+s(PM10)+s(VV)+s(DD)+s(TMP),data=Canarydataset_training)
summary(GAM_model_PM2.5)
```

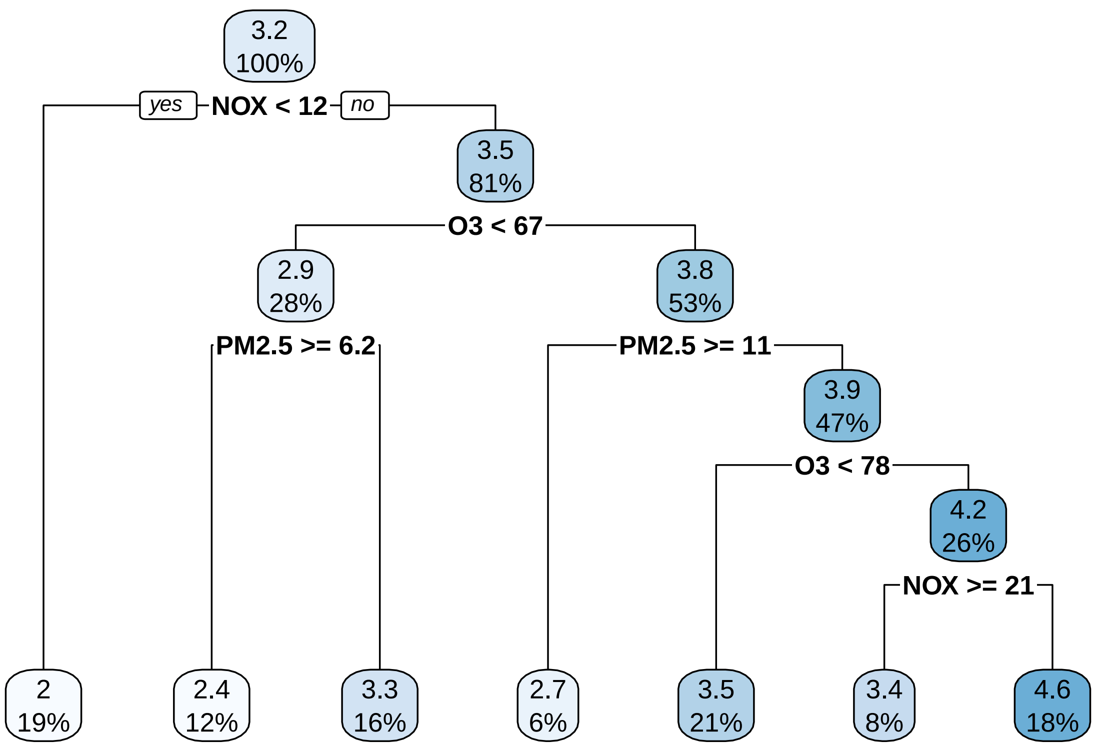
```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## PM2.5 ~ s(SO2) + s(PM10) + s(VV) + s(DD) + s(TMP)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.4535     0.0696   92.73   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Approximate significance of smooth terms:
##            edf Ref.df       F  p-value
## s(SO2)  5.480  6.582  16.803  < 2e-16 ***
## s(PM10) 5.900  7.032 138.921  < 2e-16 ***
## s(VV)   1.000  1.000  14.970 0.000114 ***
## s(DD)   7.228  8.210   4.271 3.68e-05 ***
## s(TMP)  4.105  5.109   2.220 0.049926 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =   0.48   Deviance explained = 48.8%
## GCV = 7.1981  Scale est. = 7.0763    n = 1461
```

Regression tree

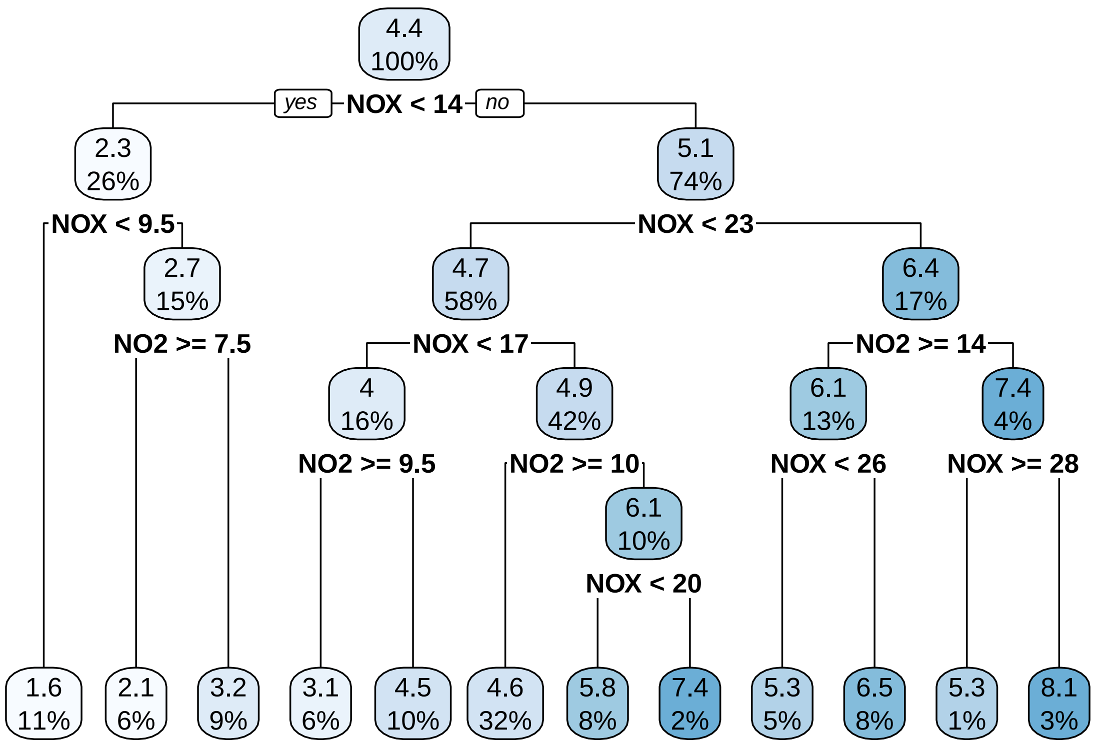
Model1(Dependent variable: SO2

```
tree_model_SO2 <- rpart(SO2 ~NO+NO2+NOX+O3+PM2.5+VV,data=Canarydataset_training)
rpart.plot(tree_model_SO2)
```



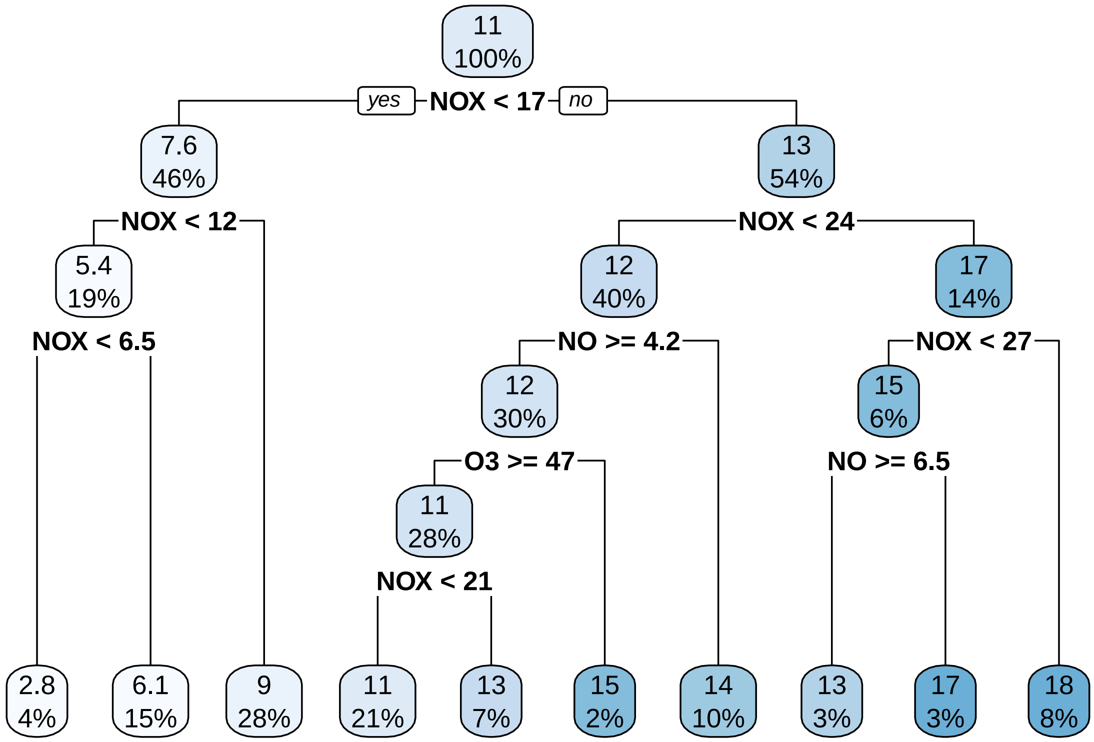
Model2(Dependent variable: NO)

```
tree_model_NO <- rpart(NO ~NO2+NOX+O3+HR+PRB,data=Canarydataset_training)
rpart.plot(tree_model_NO)
```



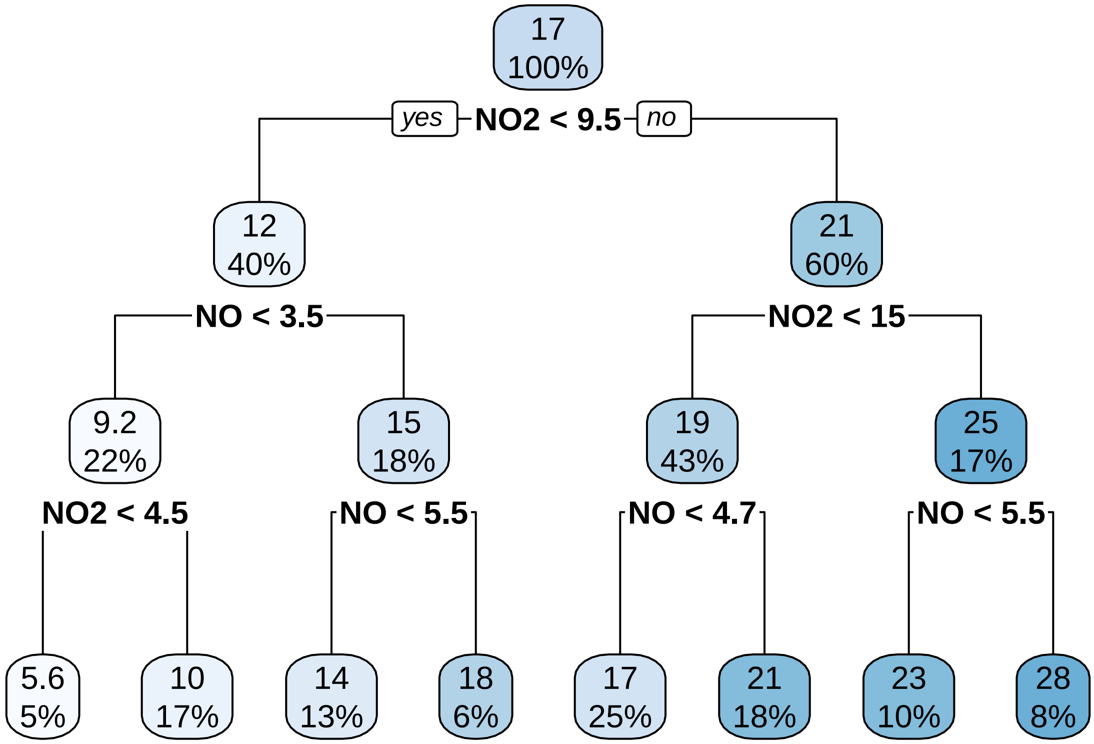
Model3(Dependent variable: NO2)

```
tree_model_NO2 <- rpart(NO2 ~SO2+NO+NOX+O3+HR+PRB,data=Canarydataset_training)
rpart.plot(tree_model_NO2)
```



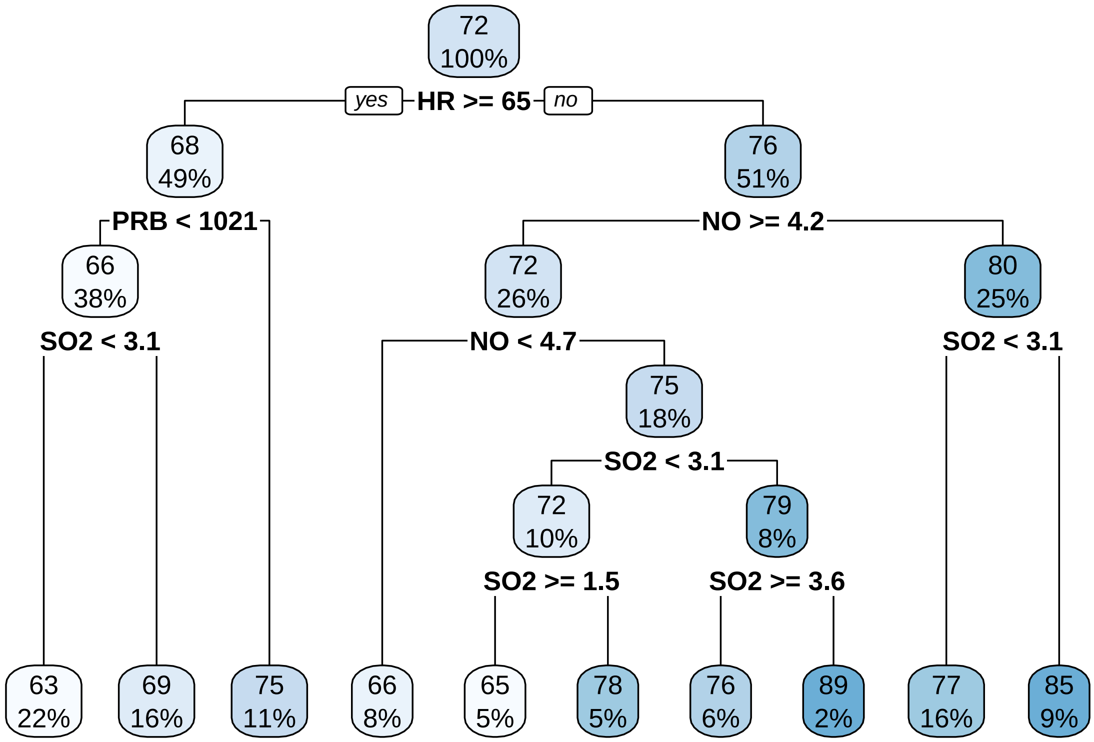
Model4(Dependent variable: NOX)

```
tree_model_NOX <- rpart(NOX ~SO2+NO+NO2+O3+HR+PRB,data=Canarydataset_training)
rpart.plot(tree_model_NOX)
```



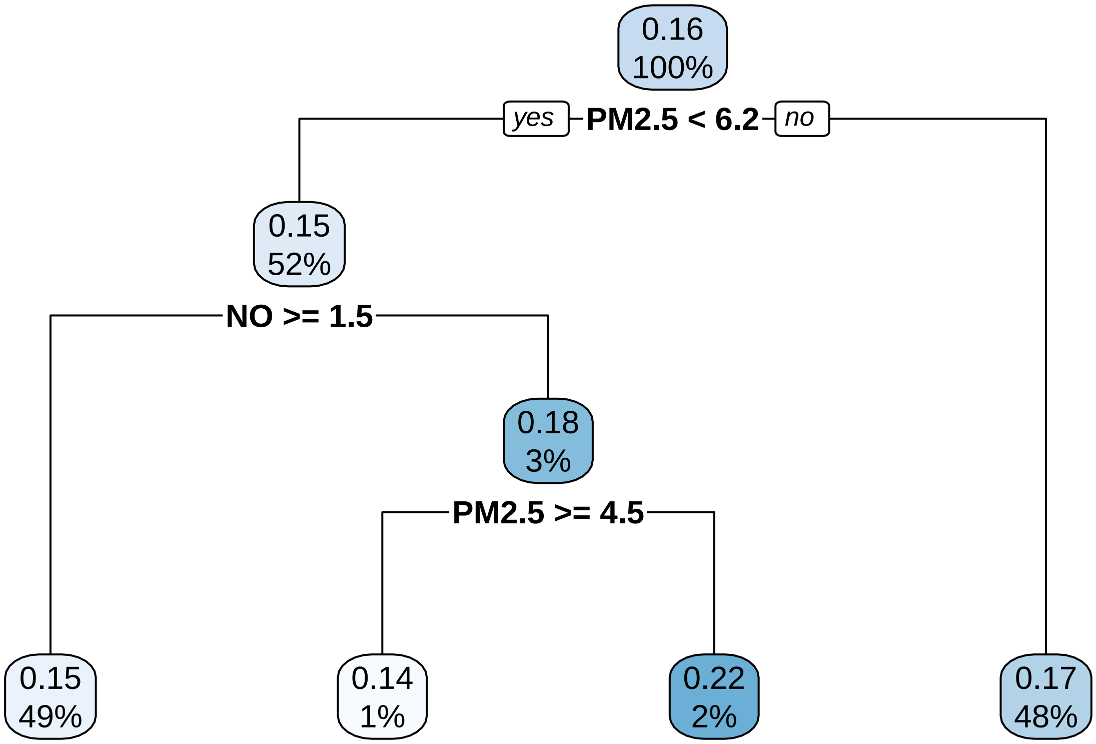
Model5(Dependent variable: O3)

```
tree_model_O3 <- rpart(O3 ~SO2+NO+NO2+NOX+HR+PRB,data=Canarydataset_training)
rpart.plot(tree_model_O3)
```



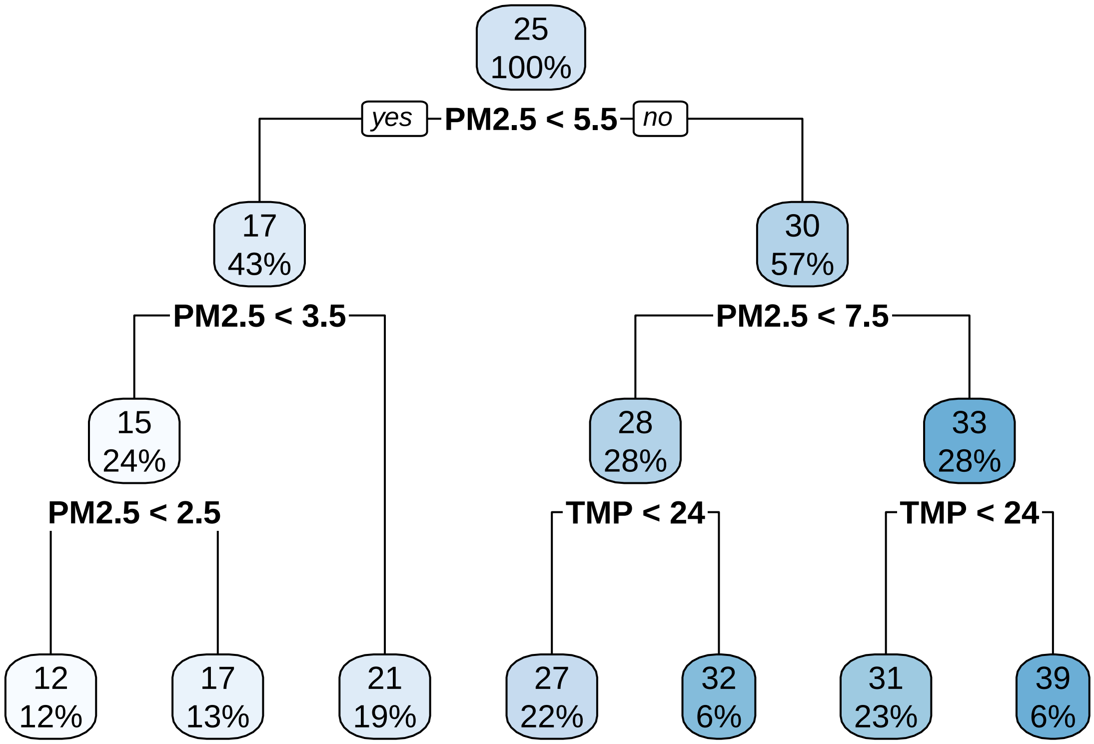
Model6(Dependent variable: CO)

```
tree_model_CO <- rpart(CO ~NO+NO2+NOX+PM2.5+DD+HR,data=Canarydataset_training)
rpart.plot(tree_model_CO)
```



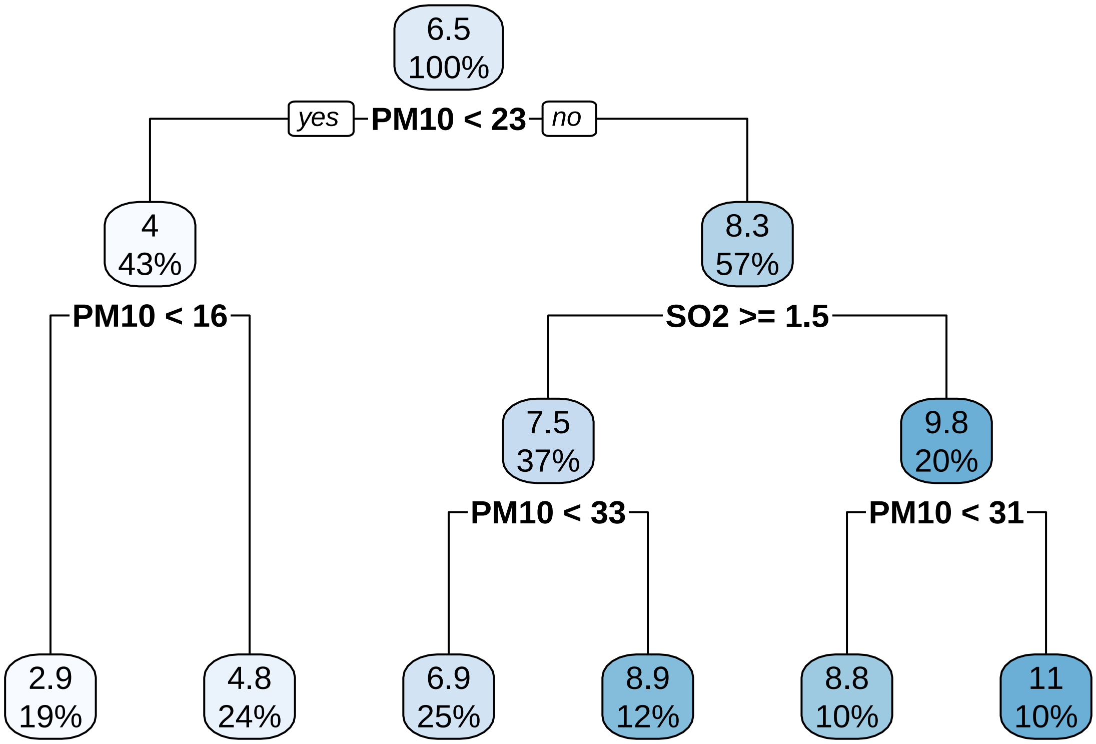
Model7(Dependent variable: PM10)

```
tree_model_PM10 <- rpart(PM10 ~NO+PM2.5+DD+TMP+HR+PRB,data=Canarydataset_training)
rpart.plot(tree_model_PM10)
```



Model8(Dependent variable: PM2.5)

```
tree_model_PM2.5 <- rpart(PM2.5 ~ SO2+PM10+VV+DD+TMP,data=Canarydataset_training)
rpart.plot(tree_model_PM2.5)
```



Model evaluation

Model1(Dependent variable: SO2)

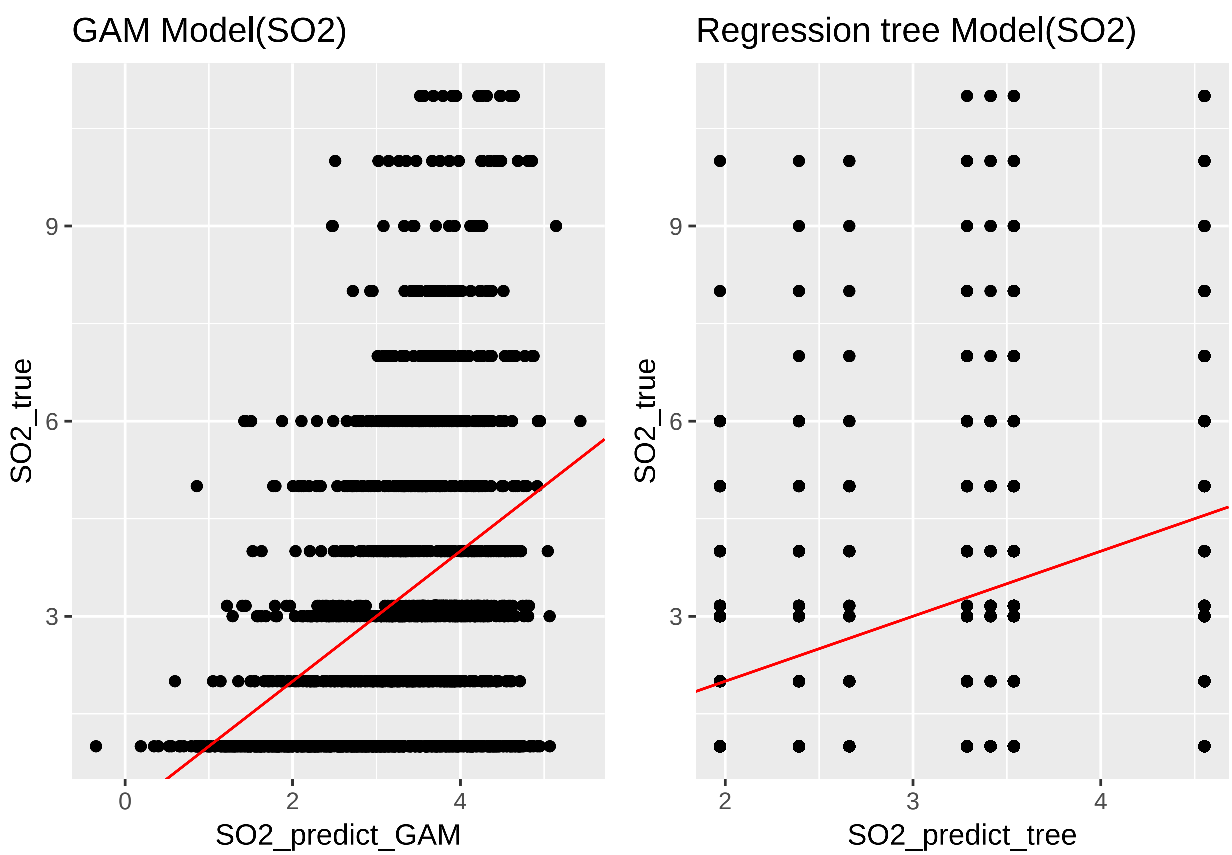
```
#Predict values using verification set.
SO2_predict_GAM <- predict(GAM_model_SO2,newdata = Canarydataset_verification)
SO2_predict_tree <- predict(tree_model_SO2,newdata = Canarydataset_verification)
SO2_true <- Canarydataset_verification$SO2

#Using NMSE to evaluate the model performance.
nmse_SO2_GAM <- mean((SO2_predict_GAM - SO2_true)^2)/mean((mean(SO2_true)-SO2_true)^2)
nmse_SO2_GAM
```

```
## [1] 0.8669353
```

```
nmse_SO2_tree <-  mean((SO2_predict_tree - SO2_true)^2)/mean((mean(SO2_true)-SO2_true)^2)
nmse_SO2_tree
```

```
## [1] 0.9242515
```

```
#Visualisation of the predictive results of these two models.
SO2_data <- data.frame(SO2_predict_GAM = SO2_predict_GAM,
                      SO2_predict_tree = SO2_predict_tree,
                      SO2_true = SO2_true)
plot_SO2_GAM <- ggplot(SO2_data,aes(SO2_predict_GAM,SO2_true))+geom_point()+geom_abline(slope=1, interc
plot_SO2_tree <- ggplot(SO2_data,aes(SO2_predict_tree,SO2_true))+geom_point()+geom_abline(slope=1, inte
ggplot2.multiplot(plot_SO2_GAM, plot_SO2_tree, cols=2)
```



Model2(Dependent variable: NO)

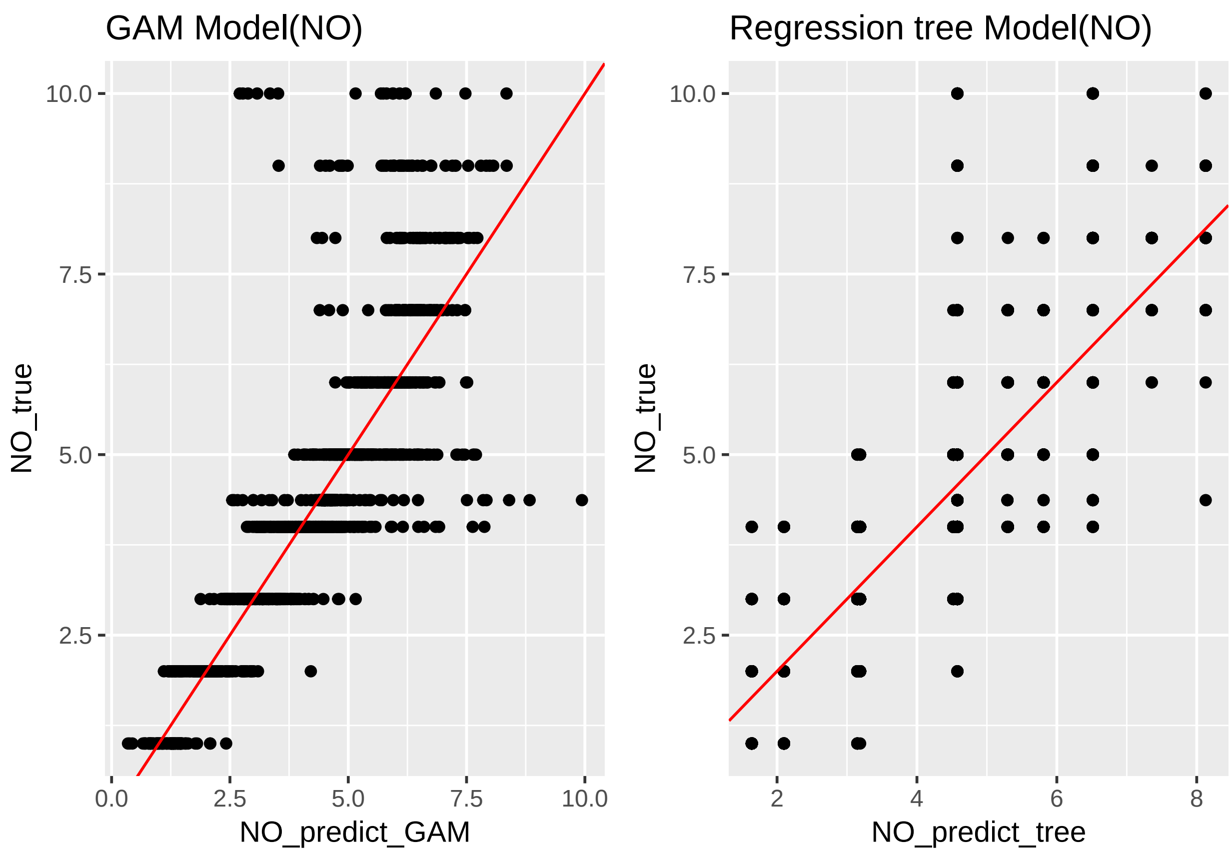
```
#Predict values using verification set.
NO_predict_GAM <- predict(GAM_model_NO,newdata = Canarydataset_verification)
NO_predict_tree <- predict(tree_model_NO,newdata = Canarydataset_verification)
NO_true <- Canarydataset_verification$NO

#Using NMSE to evaluate the model performance.
nmse_NO_GAM <- mean((NO_predict_GAM - NO_true)^2)/mean((mean(NO_true)-NO_true)^2)
nmse_NO_GAM
```

```
## [1] 0.2903619
```

```
nmse_NO_tree <-  mean((NO_predict_tree - NO_true)^2)/mean((mean(NO_true)-NO_true)^2)
nmse_NO_tree
```

```
## [1] 0.32765
```

```
#Visualisation of the predictive results of these two models.
NO_data <- data.frame(NO_predict_GAM = NO_predict_GAM,
                      NO_predict_tree = NO_predict_tree,
                      NO_true = NO_true)
plot_NO_GAM <- ggplot(NO_data,aes(NO_predict_GAM,NO_true))+geom_point()+geom_abline(slope=1, intercept =
plot_NO_tree <- ggplot(NO_data,aes(NO_predict_tree,NO_true))+geom_point()+geom_abline(slope=1, intercept
ggplot2.multiplot(plot_NO_GAM, plot_NO_tree, cols=2)
```



Model3(Dependent variable: NO2)

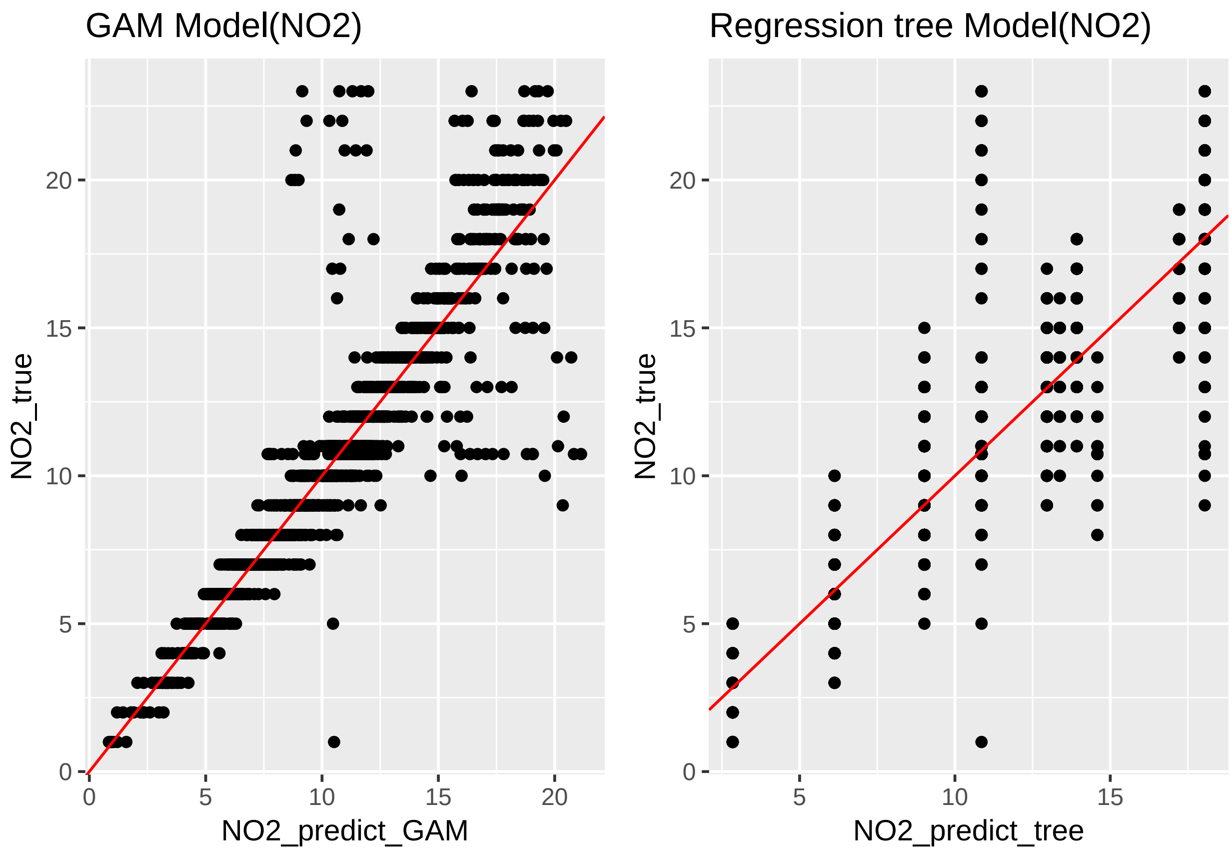
```
#Predict values using verification set.
NO2_predict_GAM <- predict(GAM_model_NO2,newdata = Canarydataset_verification)
NO2_predict_tree <- predict(tree_model_NO2,newdata = Canarydataset_verification)
NO2_true <- Canarydataset_verification$NO2

#Using NMSE to evaluate the model performance.
nmse_NO2_GAM <- mean((NO2_predict_GAM - NO2_true)^2)/mean((mean(NO2_true)-NO2_true)^2)
nmse_NO2_GAM
```

```
## [1] 0.1996935
```

```
nmse_NO2_tree <-  mean((NO2_predict_tree - NO2_true)^2)/mean((mean(NO2_true)-NO2_true)^2)
nmse_NO2_tree
```

```
## [1] 0.270828
```

```
#Visualisation of the predictive results of these two models.
NO2_data <- data.frame(NO2_predict_GAM = NO2_predict_GAM,
                       NO2_predict_tree = NO2_predict_tree,
                       NO2_true = NO2_true)
plot_NO2_GAM <- ggplot(NO2_data,aes(NO2_predict_GAM,NO2_true))+geom_point()+geom_abline(slope=1, interc
plot_NO2_tree <- ggplot(NO2_data,aes(NO2_predict_tree,NO2_true))+geom_point()+geom_abline(slope=1, inte
ggplot2.multiplot(plot_NO2_GAM, plot_NO2_tree, cols=2)
```



Model4(Dependent variable: NOX)

```
#Predict values using verification set.
NOX_predict_GAM <- predict(GAM_model_NOX,newdata = Canarydataset_verification)
NOX_predict_tree <- predict(tree_model_NOX,newdata = Canarydataset_verification)
NOX_true <- Canarydataset_verification$NOX

#Using NMSE to evaluate the model performance.
```

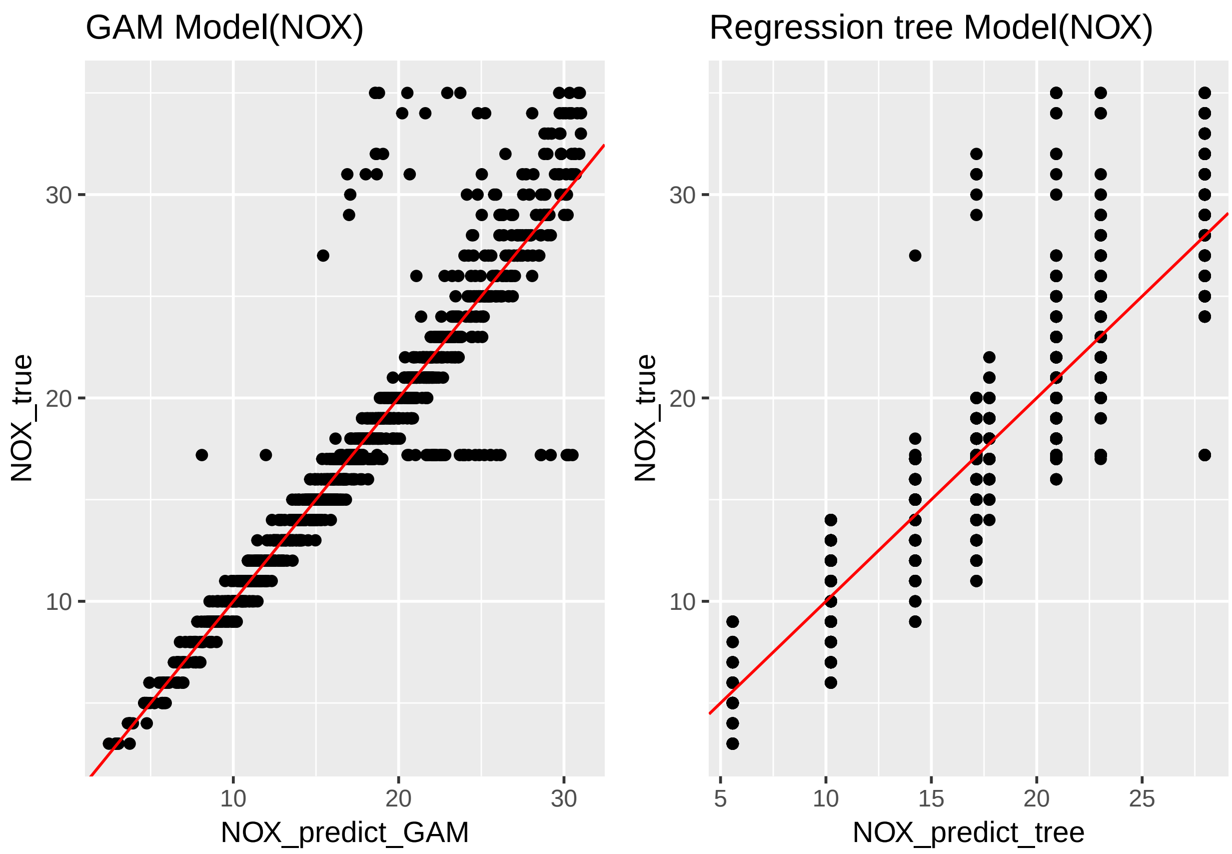
```r
nmse_NOX_GAM <- mean((NOX_predict_GAM - NOX_true)^2)/mean((mean(NOX_true)-NOX_true)^2)
nmse_NOX_GAM
```

```
## [1] 0.124064
```

```r
nmse_NOX_tree <-  mean((NOX_predict_tree - NOX_true)^2)/mean((mean(NOX_true)-NOX_true)^2)
nmse_NOX_tree
```

```
## [1] 0.2002506
```

```r
#Visualisation of the predictive results of these two models.
NOX_data <- data.frame(NOX_predict_GAM = NOX_predict_GAM,
                       NOX_predict_tree = NOX_predict_tree,
                       NOX_true = NOX_true)
plot_NOX_GAM <- ggplot(NOX_data,aes(NOX_predict_GAM,NOX_true))+geom_point()+geom_abline(slope=1, interc
plot_NOX_tree <- ggplot(NOX_data,aes(NOX_predict_tree,NOX_true))+geom_point()+geom_abline(slope=1, inte
ggplot2.multiplot(plot_NOX_GAM, plot_NOX_tree, cols=2)
```



Model5(Dependent variable: O3)

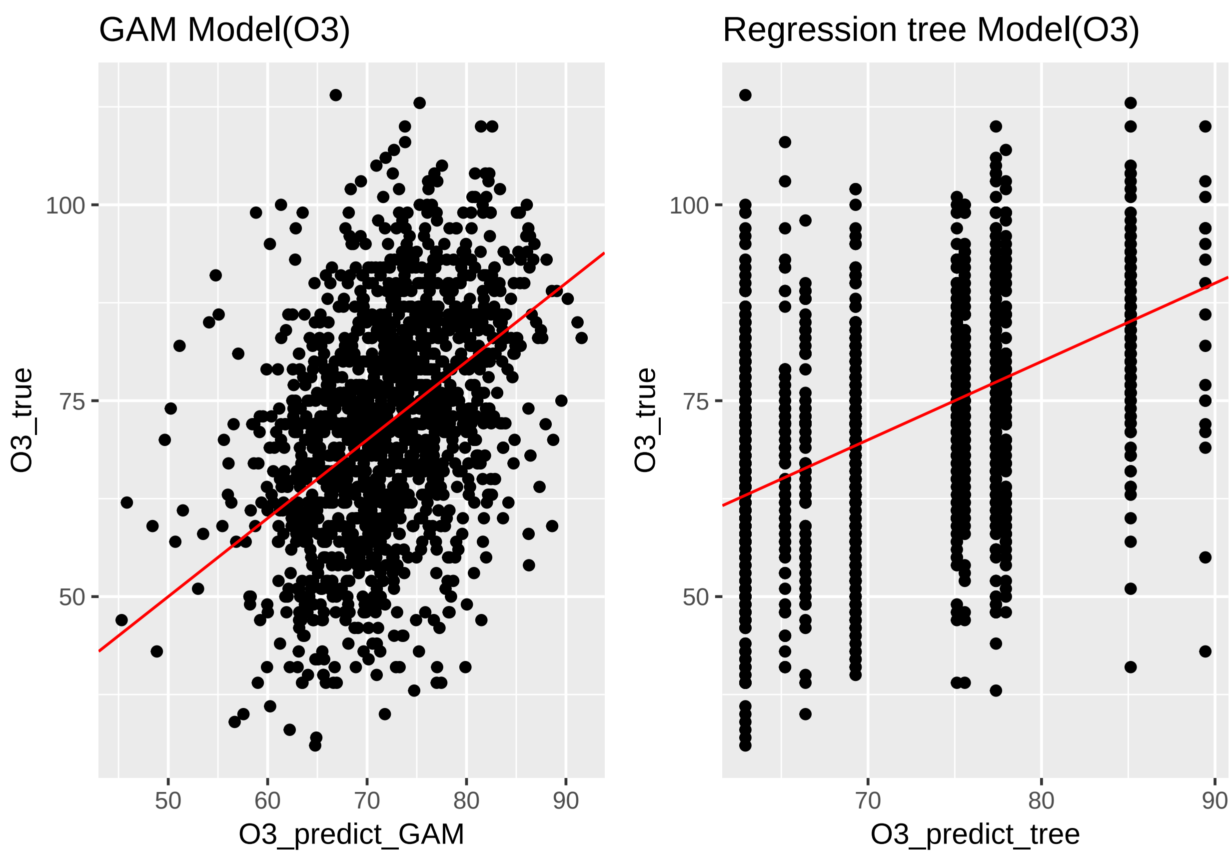
```
#Predict values using verification set.
O3_predict_GAM <- predict(GAM_model_O3,newdata = Canarydataset_verification)
O3_predict_tree <- predict(tree_model_O3,newdata = Canarydataset_verification)
O3_true <- Canarydataset_verification$O3

#Using NMSE to evaluate the model performance.
nmse_O3_GAM <- mean((O3_predict_GAM - O3_true)^2)/mean((mean(O3_true)-O3_true)^2)
nmse_O3_GAM
```

```
## [1] 0.8522244
```

```
nmse_O3_tree <-  mean((O3_predict_tree - O3_true)^2)/mean((mean(O3_true)-O3_true)^2)
nmse_O3_tree
```

```
## [1] 0.8485102
```

```
#Visualisation of the predictive results of these two models.
O3_data <- data.frame(O3_predict_GAM = O3_predict_GAM,
                      O3_predict_tree = O3_predict_tree,
                      O3_true = O3_true)
plot_O3_GAM <- ggplot(O3_data,aes(O3_predict_GAM,O3_true))+geom_point()+geom_abline(slope=1, intercept =
plot_O3_tree <- ggplot(O3_data,aes(O3_predict_tree,O3_true))+geom_point()+geom_abline(slope=1, intercept
ggplot2.multiplot(plot_O3_GAM, plot_O3_tree, cols=2)
```



Model6(Dependent variable: CO)

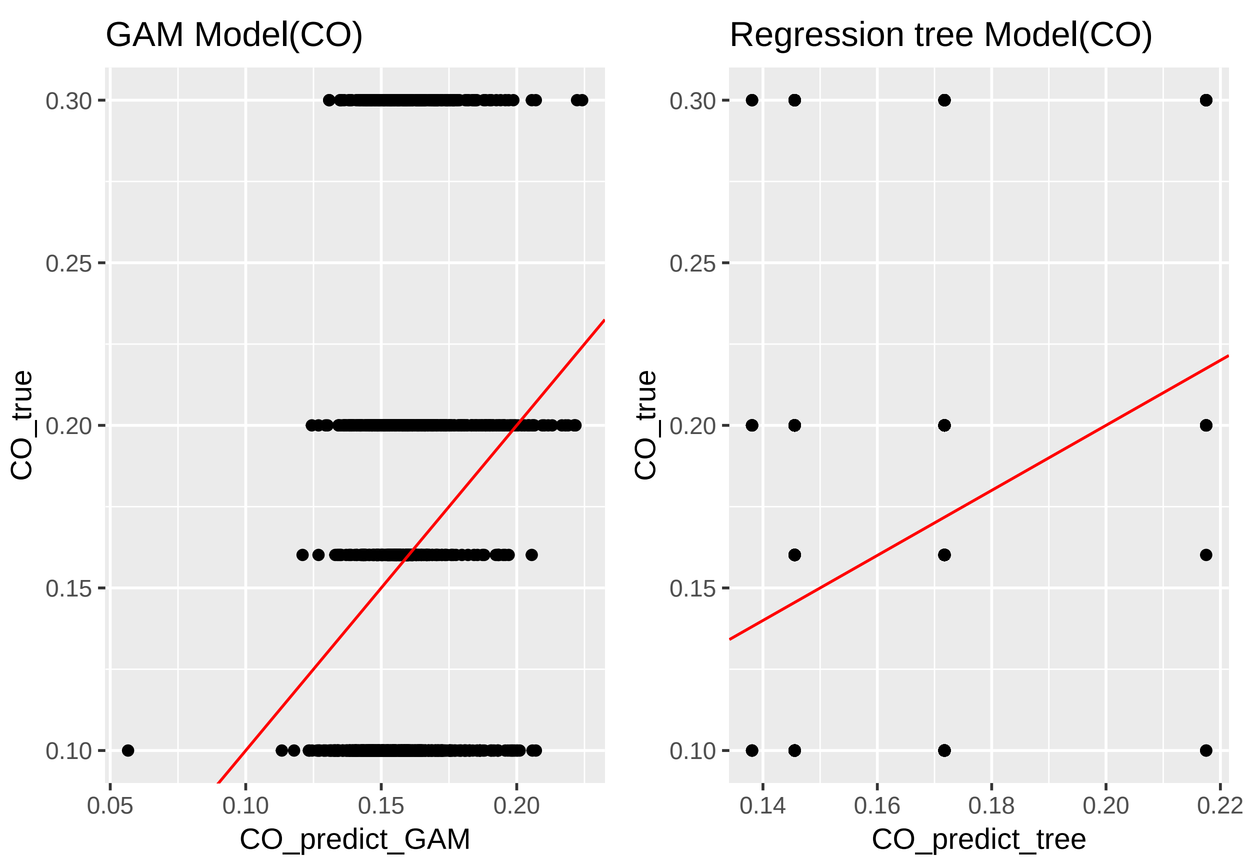
```
#Predict values using verification set.
CO_predict_GAM <- predict(GAM_model_CO,newdata = Canarydataset_verification)
CO_predict_tree <- predict(tree_model_CO,newdata = Canarydataset_verification)
CO_true <- Canarydataset_verification$CO

#Using NMSE to evaluate the model performance.
nmse_CO_GAM <- mean((CO_predict_GAM - CO_true)^2)/mean((mean(CO_true)-CO_true)^2)
nmse_CO_GAM
```

```
## [1] 0.9556282
```

```
nmse_CO_tree <-  mean((CO_predict_tree - CO_true)^2)/mean((mean(CO_true)-CO_true)^2)
nmse_CO_tree
```

```
## [1] 0.945722
```

```
#Visualisation of the predictive results of these two models.
CO_data <- data.frame(CO_predict_GAM = CO_predict_GAM,
                      CO_predict_tree = CO_predict_tree,
                      CO_true = CO_true)
plot_CO_GAM <- ggplot(CO_data,aes(CO_predict_GAM,CO_true))+geom_point()+geom_abline(slope=1, intercept =
plot_CO_tree <- ggplot(CO_data,aes(CO_predict_tree,CO_true))+geom_point()+geom_abline(slope=1, intercep
ggplot2.multiplot(plot_CO_GAM, plot_CO_tree, cols=2)
```



Model7(Dependent variable: PM10)

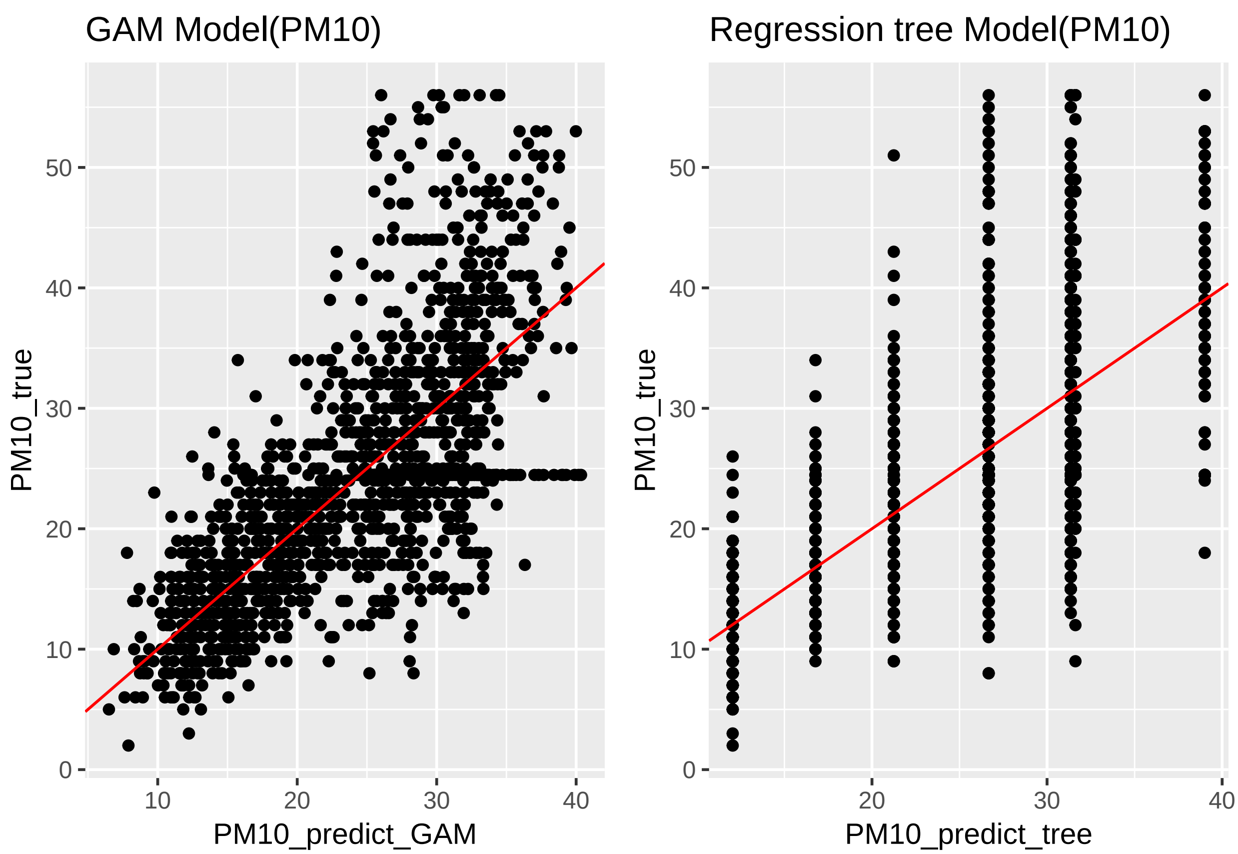
```
#Predict values using verification set.
PM10_predict_GAM <- predict(GAM_model_PM10,newdata = Canarydataset_verification)
PM10_predict_tree <- predict(tree_model_PM10,newdata = Canarydataset_verification)
PM10_true <- Canarydataset_verification$PM10

#Using NMSE to evaluate the model performance.
nmse_PM10_GAM <- mean((PM10_predict_GAM - PM10_true)^2)/mean((mean(PM10_true)-PM10_true)^2)
nmse_PM10_GAM
```

```
## [1] 0.497423
```

```
nmse_PM10_tree <-  mean((PM10_predict_tree - PM10_true)^2)/mean((mean(PM10_true)-PM10_true)^2)
nmse_PM10_tree
```

```
## [1] 0.5187397
```

```
#Visualisation of the predictive results of these two models.
PM10_data <- data.frame(PM10_predict_GAM = PM10_predict_GAM,
                        PM10_predict_tree = PM10_predict_tree,
                        PM10_true = PM10_true)
plot_PM10_GAM <- ggplot(PM10_data,aes(PM10_predict_GAM,PM10_true))+geom_point()+geom_abline(slope=1, int
plot_PM10_tree <- ggplot(PM10_data,aes(PM10_predict_tree,PM10_true))+geom_point()+geom_abline(slope=1, i
ggplot2.multiplot(plot_PM10_GAM, plot_PM10_tree, cols=2)
```



Model8(Dependent variable: PM2.5)

```
#Predict values using verification set.
PM2.5_predict_GAM <- predict(GAM_model_PM2.5,newdata = Canarydataset_verification)
PM2.5_predict_tree <- predict(tree_model_PM2.5,newdata = Canarydataset_verification)
PM2.5_true <- Canarydataset_verification$PM2.5

#Using NMSE to evaluate the model performance.
nmse_PM2.5_GAM <- mean((PM2.5_predict_GAM - PM2.5_true)^2)/mean((mean(PM2.5_true)-PM2.5_true)^2)
nmse_PM2.5_GAM
```

```
## [1] 0.5230972
```

```
nmse_PM2.5_tree <-  mean((PM2.5_predict_tree - PM2.5_true)^2)/mean((mean(PM2.5_true)-PM2.5_true)^2)
nmse_PM2.5_tree
```

```
## [1] 0.5680897
```

```
#Visualisation of the predictive results of these two models.
PM2.5_data <- data.frame(PM2.5_predict_GAM = PM2.5_predict_GAM,
                         PM2.5_predict_tree = PM2.5_predict_tree,
                         PM2.5_true = PM2.5_true)
plot_PM2.5_GAM <- ggplot(PM2.5_data,aes(PM2.5_predict_GAM,PM2.5_true))+geom_point()+geom_abline(slope=1
plot_PM2.5_tree <- ggplot(PM2.5_data,aes(PM2.5_predict_tree,PM2.5_true))+geom_point()+geom_abline(slope=
ggplot2.multiplot(plot_PM2.5_GAM, plot_PM2.5_tree, cols=2)
```