Projeto Data Mining I

Lucas Parada, Ana Catarina Monteiro, Lucas de Paula

11/16/2019

Data importation, clean-up and pre-processing

Importando os Dados

Inicialmente importamos todos as bibliotecas que utilizaresmos neste trabalho Foi importado o dataset "PRSA_Data_Aotizhongxin_20130301-20170228" - Foi utilizado o DataFrame do R para manipular os dados, pois este tipo de estrutura de dados possui um conjunto de funcionalidades e ferrantas que auxiliam neste processo

```
library(na.tools)
library(naniar)
```

```
##
## Attaching package: 'naniar'
```

```
## The following objects are masked from 'package:na.tools':
##
##     all_na, any_na, is_na, which_na
```

```
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(zoo)
```

```
##
## Attaching package: 'zoo'
```

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

```
library(ggplot2)
library(caret)
```

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

```
library(tidyimpute)
```

```
##
## Attaching package: 'tidyimpute'
```

```
## The following objects are masked from 'package:naniar':
##
```

```
##     impute_mean, impute_mean_all, impute_mean_at, impute_mean_if,
##     impute_median, impute_median_all, impute_median_at,
##     impute_median_if
```

```
library("ggplot2")

df = read.csv('data/PRSA_Data_Aotizhongxin_20130301-20170228.csv')
```

Analisando os dados

Utilizando a função "summary" da linguagem R foi feita uma análise dos dados.

```
summary(df)
```

```
##        No             year          month             day
##  Min.   :    1   Min.   :2013   Min.   : 1.000   Min.   : 1.00
##  1st Qu.: 8767   1st Qu.:2014   1st Qu.: 4.000   1st Qu.: 8.00
##  Median :17532   Median :2015   Median : 7.000   Median :16.00
##  Mean   :17532   Mean   :2015   Mean   : 6.523   Mean   :15.73
##  3rd Qu.:26298   3rd Qu.:2016   3rd Qu.:10.000   3rd Qu.:23.00
##  Max.   :35064   Max.   :2017   Max.   :12.000   Max.   :31.00
##
##       hour           PM2.5             PM10            SO2
##  Min.   : 0.00   Min.   :  3.00   Min.   :  2.0   Min.   :  0.2856
##  1st Qu.: 5.75   1st Qu.: 22.00   1st Qu.: 38.0   1st Qu.:  3.0000
##  Median :11.50   Median : 58.00   Median : 87.0   Median :  9.0000
##  Mean   :11.50   Mean   : 82.77   Mean   :110.1   Mean   : 17.3759
##  3rd Qu.:17.25   3rd Qu.:114.00   3rd Qu.:155.0   3rd Qu.: 21.0000
##  Max.   :23.00   Max.   :898.00   Max.   :984.0   Max.   :341.0000
##                  NA's   :925      NA's   :718     NA's   :935
##       NO2              CO              O3               TEMP
##  Min.   :  2.00   Min.   :  100   Min.   :  0.2142   Min.   :-16.80
##  1st Qu.: 30.00   1st Qu.:  500   1st Qu.:  8.0000   1st Qu.:  3.10
##  Median : 53.00   Median :  900   Median : 42.0000   Median : 14.50
##  Mean   : 59.31   Mean   : 1263   Mean   : 56.3534   Mean   : 13.58
##  3rd Qu.: 82.00   3rd Qu.: 1500   3rd Qu.: 82.0000   3rd Qu.: 23.30
##  Max.   :290.00   Max.   :10000   Max.   :423.0000   Max.   : 40.50
##  NA's   :1023     NA's   :1776    NA's   :1719       NA's   :20
##       PRES             DEWP              RAIN                wd
##  Min.   : 985.9   Min.   :-35.300   Min.   : 0.00000   NE     : 5140
##  1st Qu.:1003.3   1st Qu.: -8.100   1st Qu.: 0.00000   ENE    : 3950
##  Median :1011.4   Median :  3.800   Median : 0.00000   SW     : 3359
##  Mean   :1011.8   Mean   :  3.123   Mean   : 0.06742   E      : 2608
##  3rd Qu.:1020.1   3rd Qu.: 15.600   3rd Qu.: 0.00000   NNE    : 2445
##  Max.   :1042.0   Max.   : 28.500   Max.   :72.50000   (Other):17481
##  NA's   :20       NA's   :20        NA's   :20         NA's   :   81
##       WSPM                station
##  Min.   : 0.000   Aotizhongxin:35064
##  1st Qu.: 0.900
##  Median : 1.400
##  Mean   : 1.708
##  3rd Qu.: 2.200
##  Max.   :11.200
##  NA's   :14
```

Começamos por verificar se existia algum dia em falta no dataframe e vimos que nao. Samendo que o dataset

possui os valores referentes a 4 anos completos especificados por hora então sevem existir $(4365+1)24 = 35064$ rows

```
  #samendo que o dataset possui os valores referentes a 4 anos completos especificados por hora então s
dim(df)
```

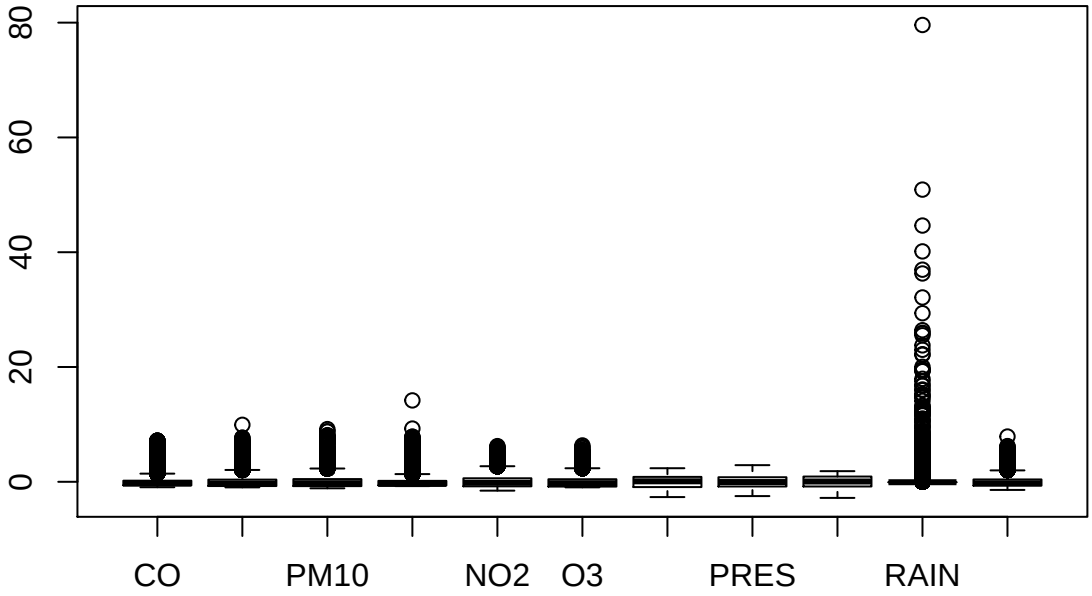
```
## [1] 35064    18
```

Outliers

Seguimos com a analise da existencia de outliers em variáveis numericas. Começando por ver fazer a análise por variavel como um todo, em seguida fizemos a análise por variavel tendo em conta a estação do ano e por último por variavel por mês.

Como um todo

```
col_list = c("CO", "PM2.5", "PM10", "SO2", "NO2", "CO", "O3", "TEMP", "PRES", "DEWP", "RAIN", "WSPM")

df_ALL_boxplot <- df %>% select(col_list)
df_ALL_boxplot <- scale(df_ALL_boxplot)

boxplot(x = df_ALL_boxplot)
```



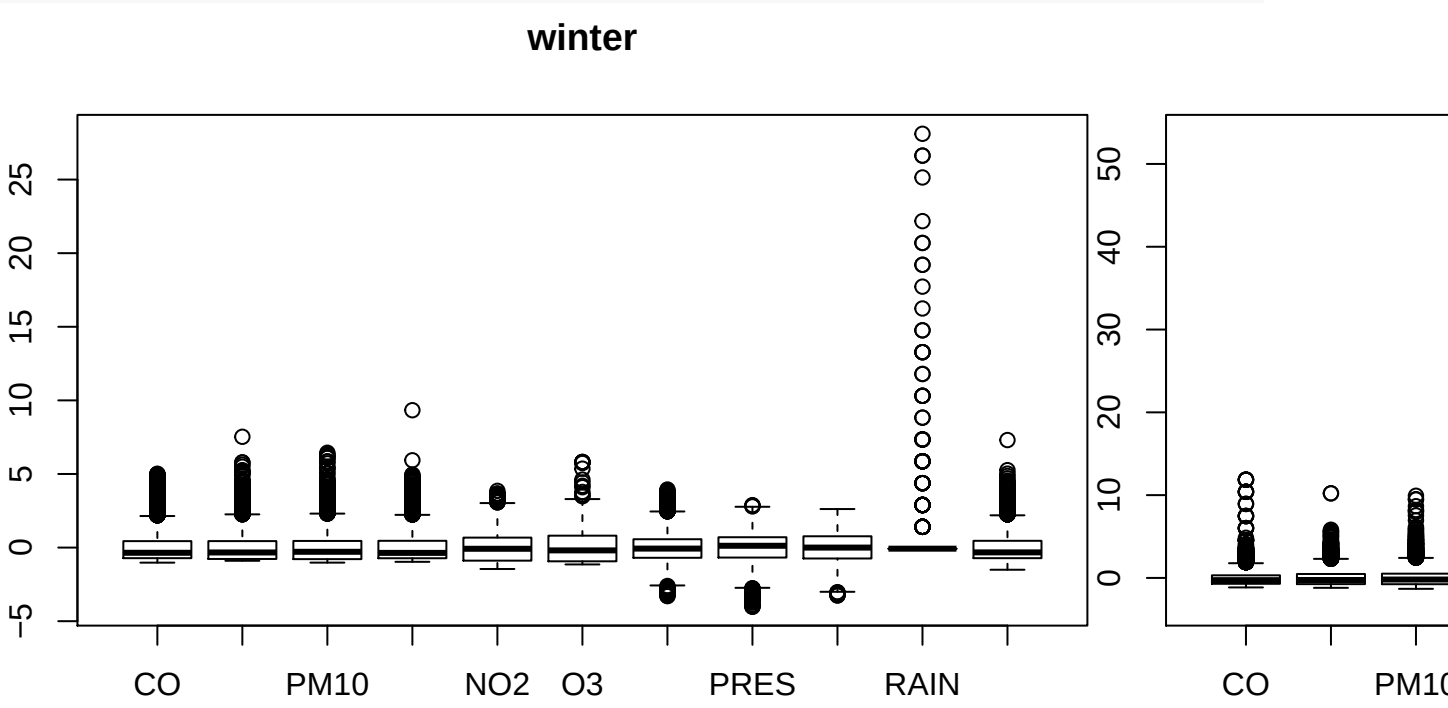

```
#for(i in col_list){
#  plot(df[i], pch=".", cex=2, main=i)  # plot cook's distance
#  abline(h = 4*mean(c(df[i]), na.rm=T), col="red")  # add cutoff line
#}
#text(x=1:length(df$CO)+1, y=df$CO, labels=ifelse(df$CO>4*mean(df$CO, na.rm=T),names(df$CO),""), col="r
```

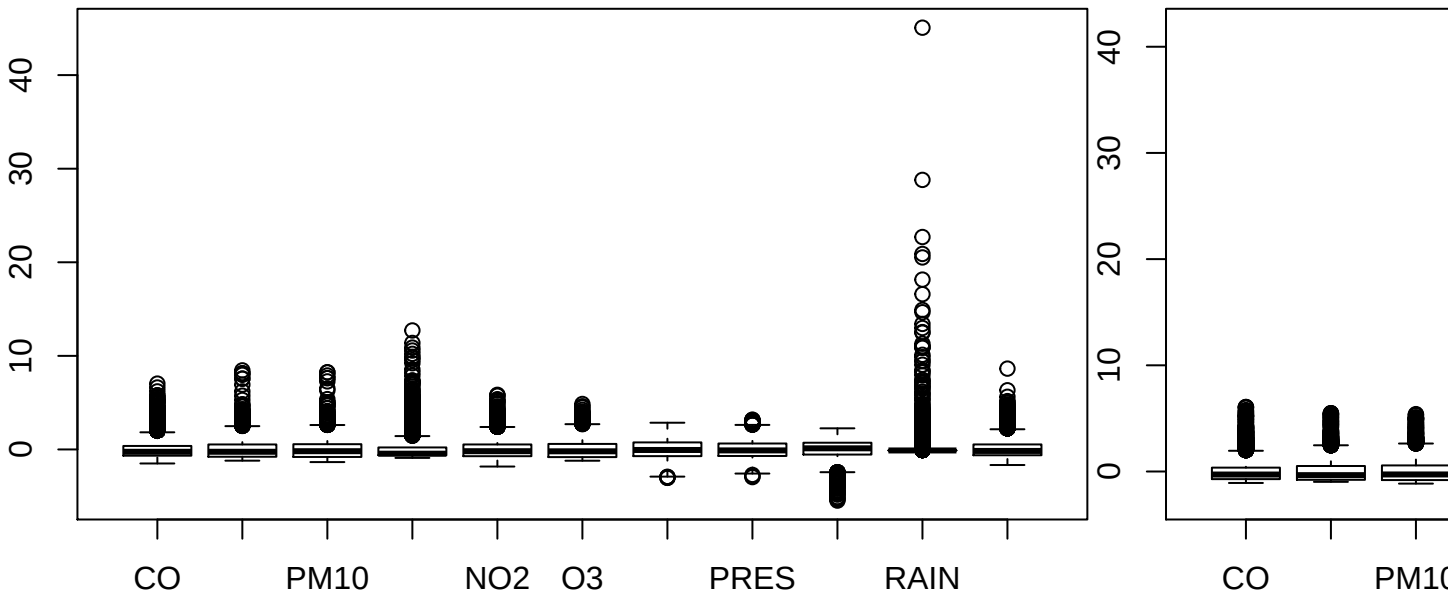
Como

Por estação do ano

```
col_list = c("season","CO", "PM2.5", "PM10", "SO2", "NO2", "CO", "O3", "TEMP", "PRES", "DEWP", "RAIN",
df_seasons_boxplot <- df %>% mutate(season =
                                ifelse(month == 12 & day >= 21, 'winter',
```

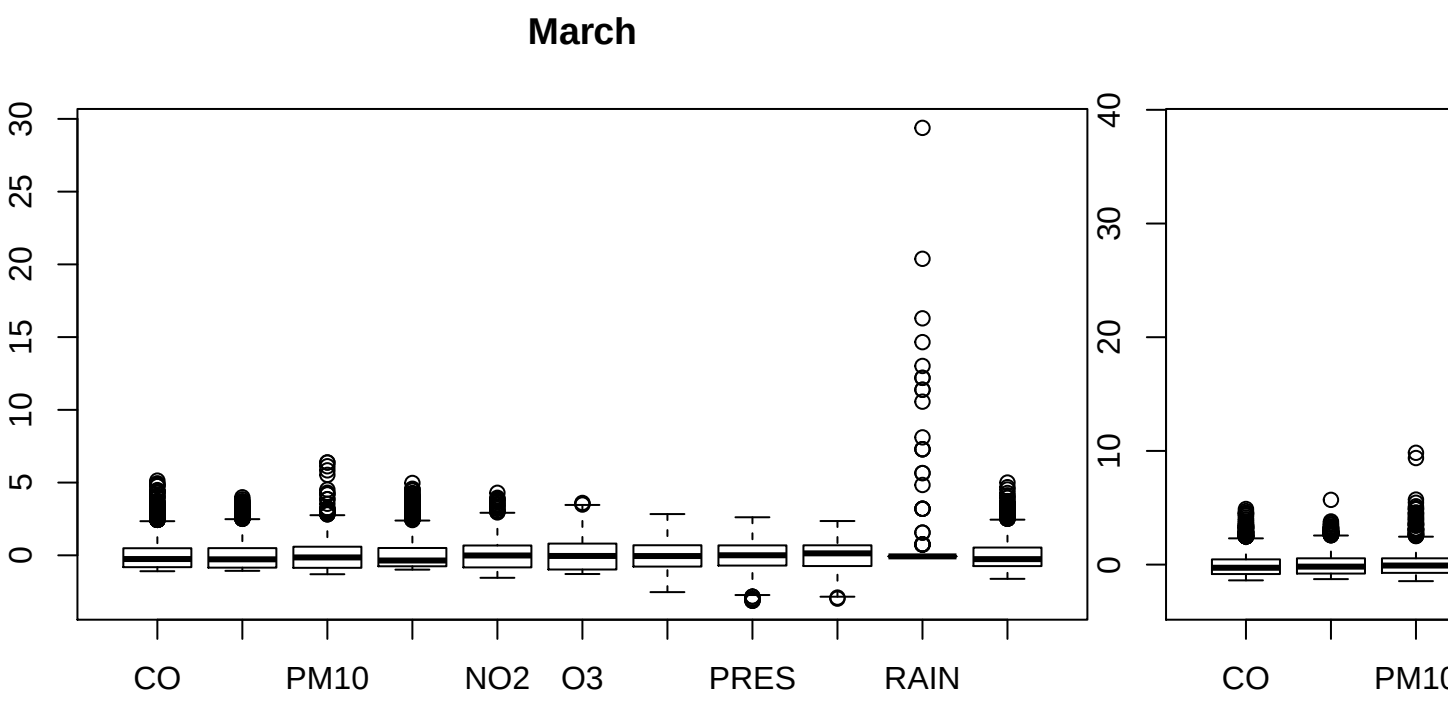
```
                                ifelse(month == 1 | month == 2, 'winter',
                                ifelse(month == 3 & day < 20, 'winter',

                                ifelse(month == 3 & day >= 20, 'spring',
                                ifelse(month == 4 | month == 5, 'spring',
                                ifelse(month == 6 & day < 21, 'spring',

                                ifelse(month == 6 & day >= 21, 'summer',
                                ifelse(month == 7 | month == 8, 'summer',
                                ifelse(month == 9 & day < 21, 'summer',

                                ifelse(month == 9 & day >= 21, 'autumn',
                                ifelse(month == 10 | month == 11, 'autumn',
                                ifelse(month == 12 & day < 21, 'autumn',

                                       0)))))))))))) %>%
                    select(col_list)

for(s in df_seasons_boxplot$season %>% unique()){
   df_seasons_boxplot %>% filter(season == s) %>% select(-season) %>% scale() %>% boxplot(main=s)
}
```

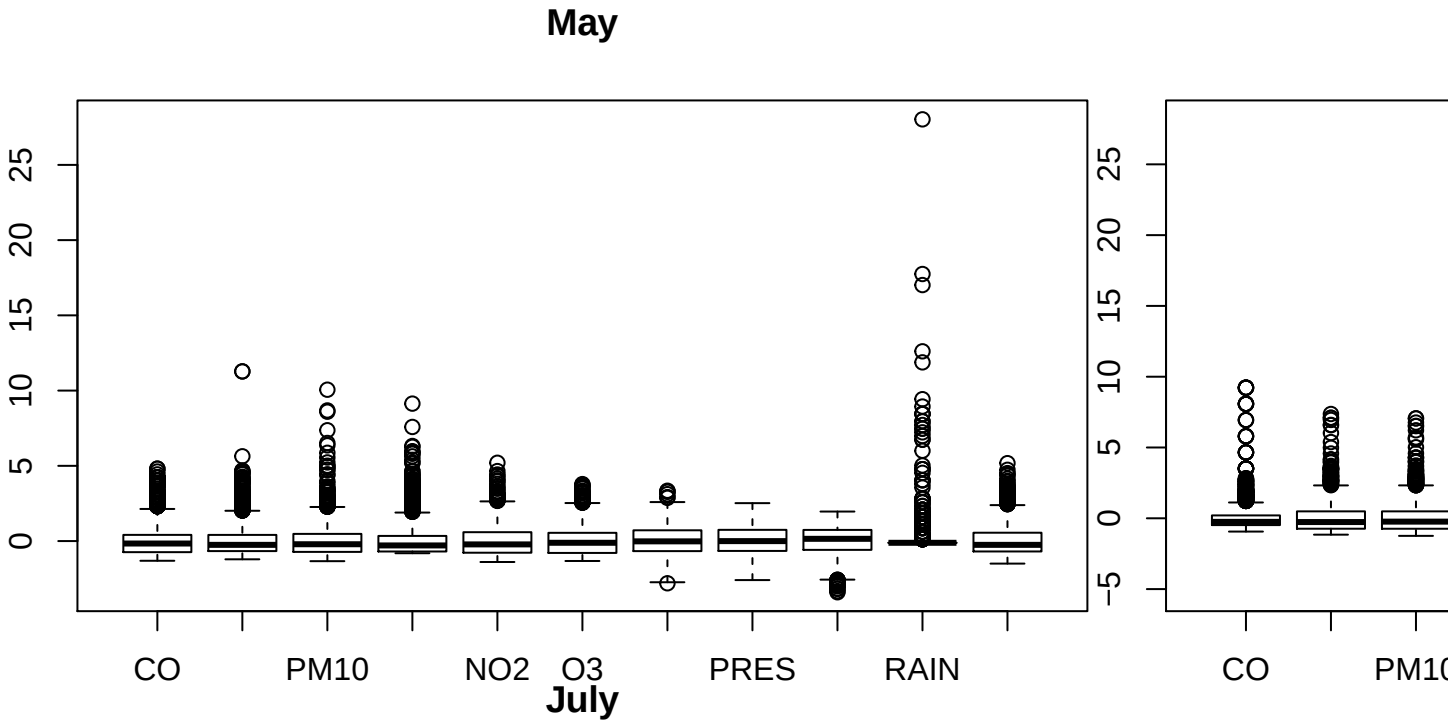


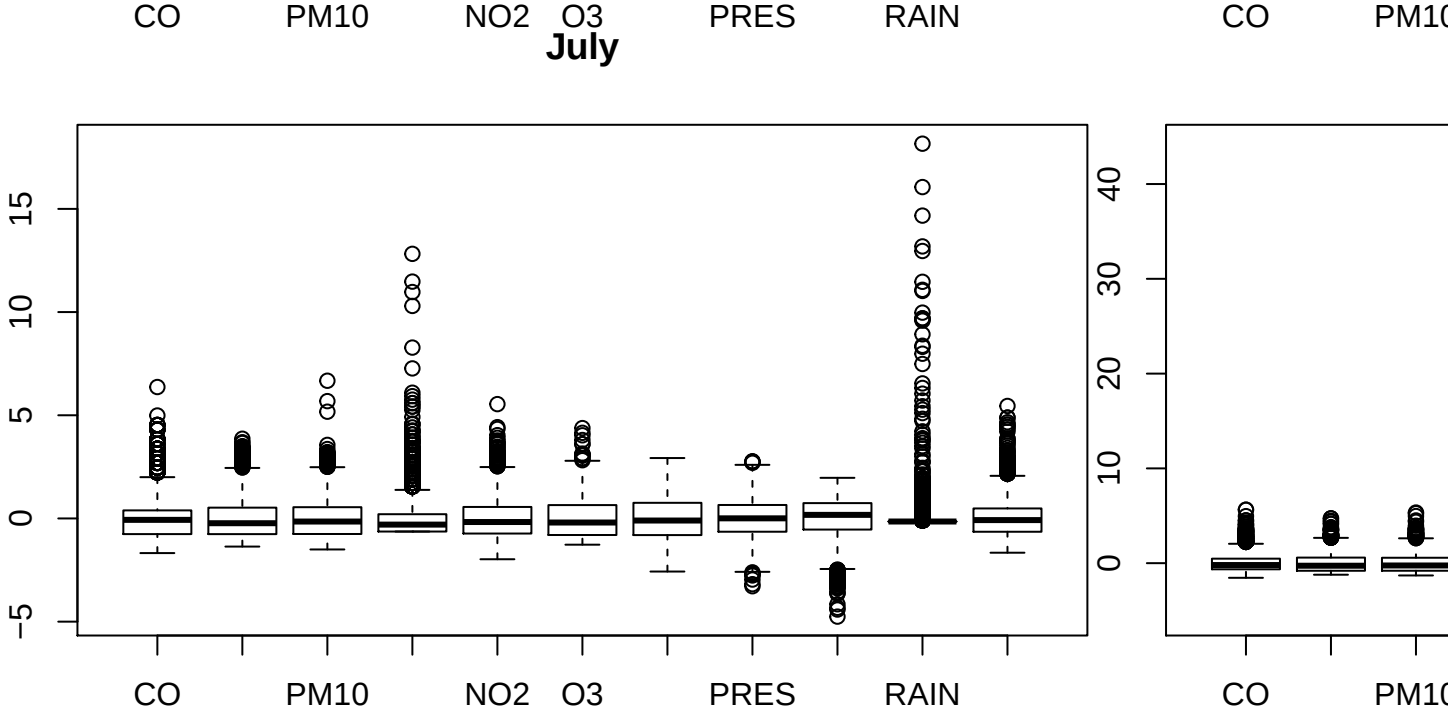


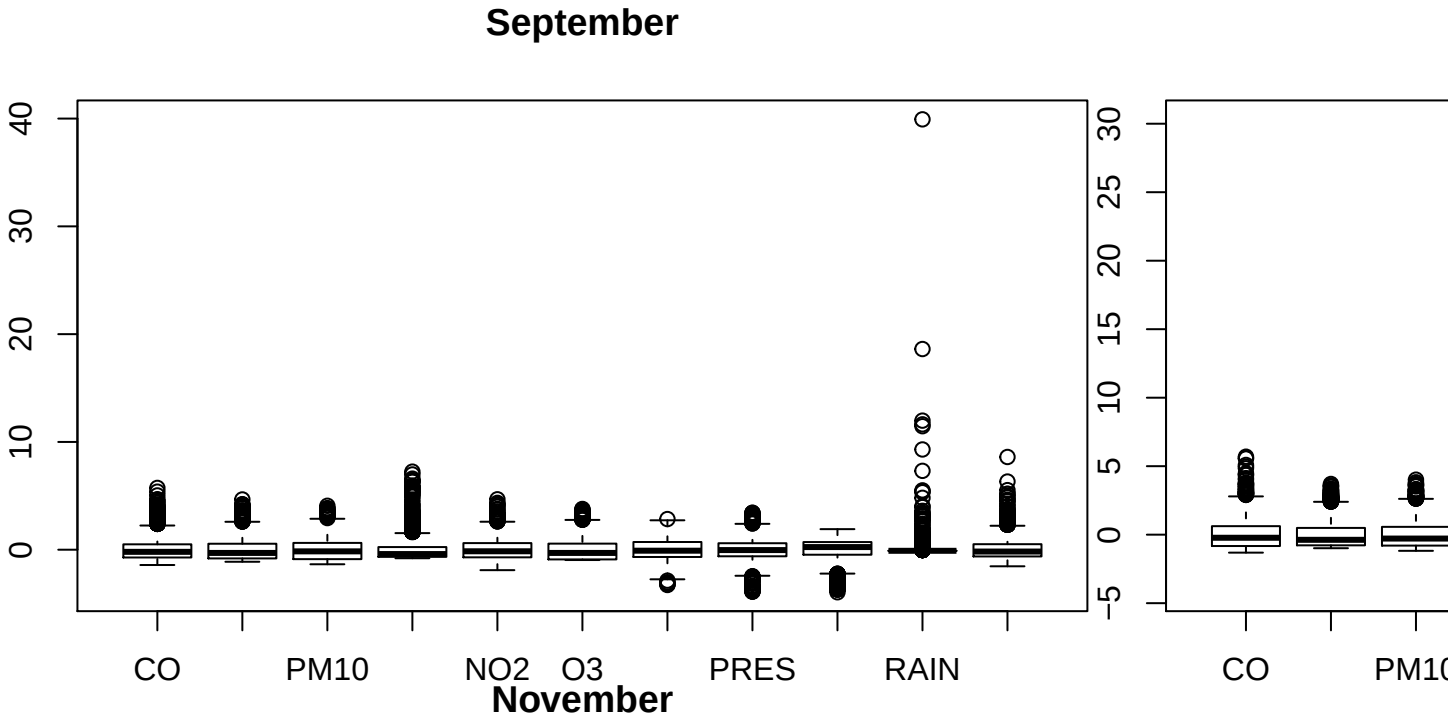
Por mês

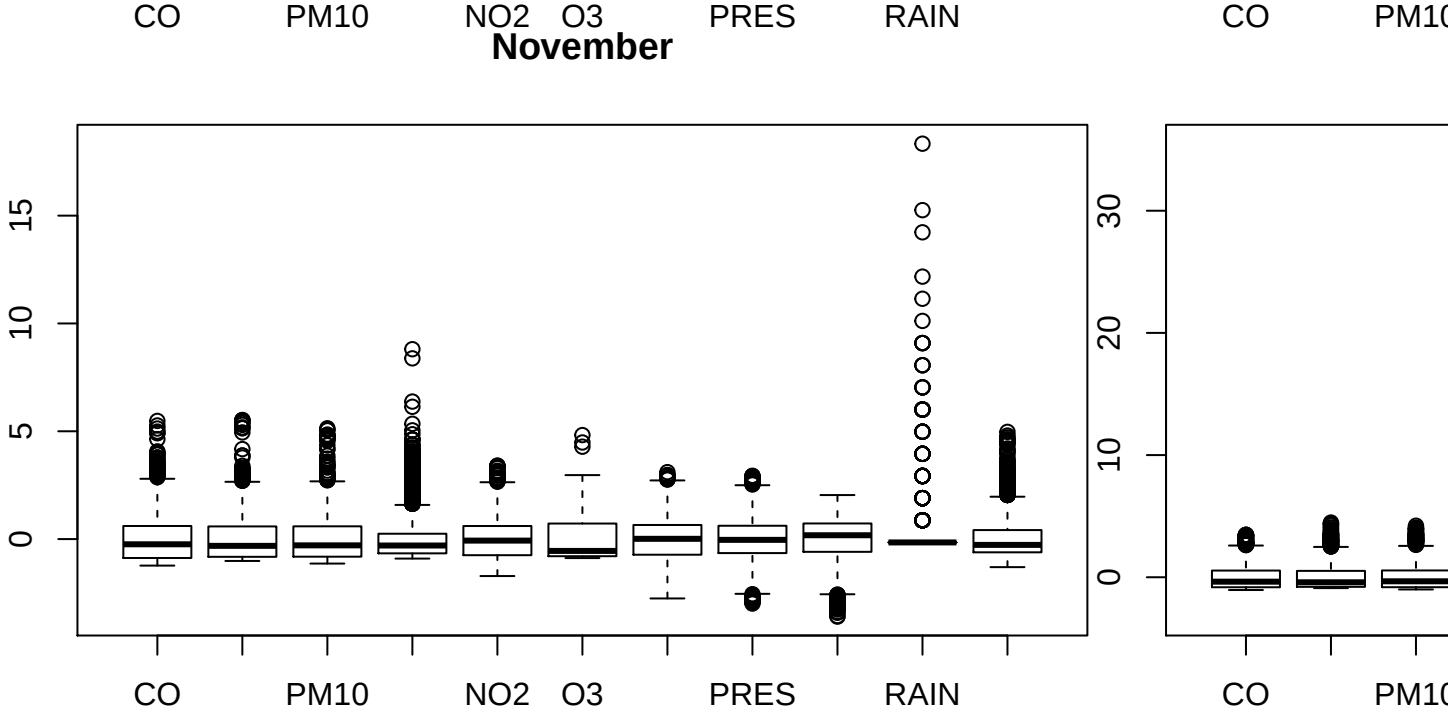
```
col_list = c("month", "CO", "PM2.5", "PM10", "SO2", "NO2", "CO", "O3", "TEMP", "PRES", "DEWP", "RAIN",


df_month_boxplot <- df %>% select(col_list)
month_names <- c("January", "February", "March", "April", "May", "June", "July", "August", "September",
for(m in df_month_boxplot$month %>% unique()){
   df_month_boxplot %>% filter(month == m) %>% select(-month) %>% scale() %>% boxplot(main=month_names[
}
```

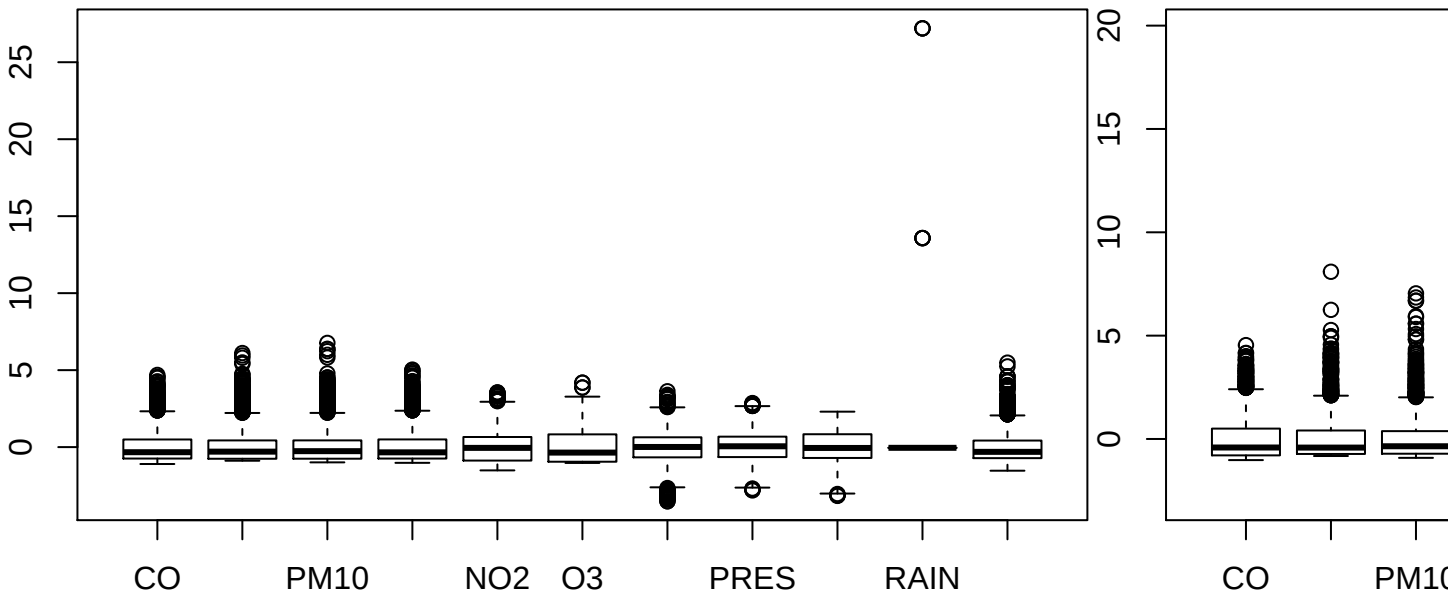



May
0
5
10
15
20
25
CO
PM10
NO2
O3
PRES
RAIN
−5
CO
PM10


July
−5
0
5
10
15
0
10
20
30
40
CO
PM10
NO2
O3
PRES
RAIN
CO
PM10

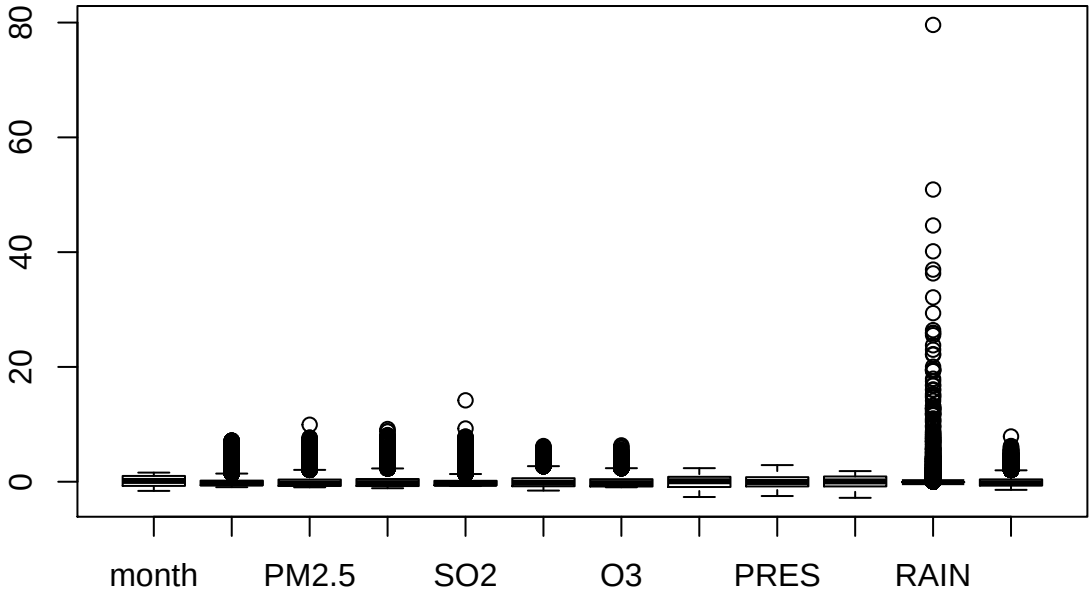

September
40
30
20
10
0
CO
PM10
NO2
O3
PRES
RAIN
30
25
20
15
10
5
0
−5
CO
PM10


November
15
10
5
0
CO
PM10
NO2
O3
PRES
RAIN
30
20
10
0
CO
PM10




```
df_month_boxplot %>% group_by(month) %>% select(-month) %>% scale() %>% boxplot()
```

```
## Adding missing grouping variables: `month`
```



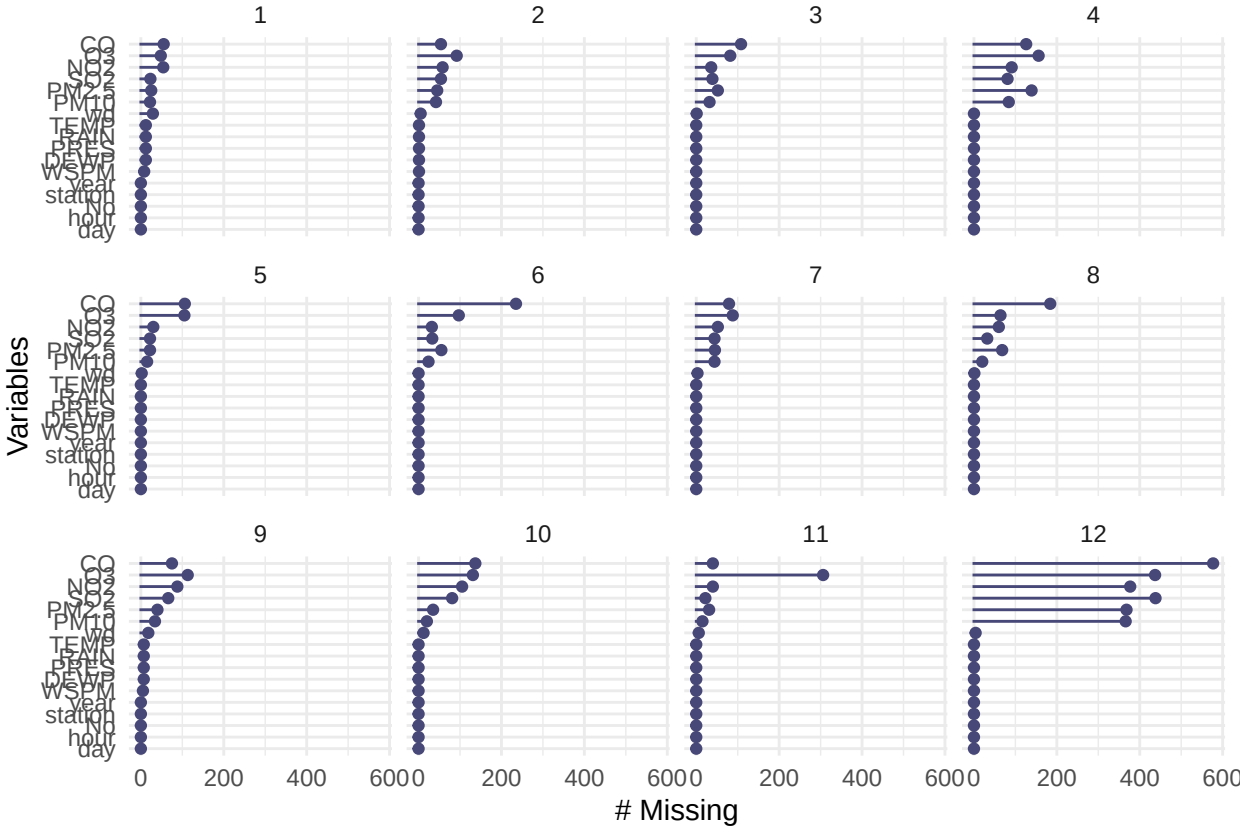
Apartir da análise mencionada anteriormente foi identificada que as varáveis “PM2.5”, “PM10”, “SO2”, “NO2”, “CO”, “O3”, “TEMP”, “PRES”, “DEWP”, “RAIN”, “wd”, “WSPM” possuem missing values.

Em seguida foi uma análise e foi verificado que as variáveis “TEMP”, “PRES”, “DEWP”, “RAIN” esta a faltar no mesmo dia atravez do gráfico:

```
#check missing values
#df %>% select(RAIN, TEMP, PRES, wd, WSPM) %>% filter_any_na()
gg_miss_var(df, facet = month)
```

```
## Warning: `cols` is now required.
```

```
## Please use `cols = c(data)`
```




```
#for(col_name in c('TEMP', 'PRES', 'DEWP', 'RAIN')){
#  (df %>% mutate(test_col = ifelse(is_na(TEMP), 1, 0)))$test_col %>% plot(pch=".", cex=2, main=col_nam
#  abline(h = 4*mean(c(df[col_name]), na.rm=T), col="red")  # add cutoff line
#}
```

Data Analisys

```
head(df, 20)
```

```
##    No year month day hour PM2.5 PM10 SO2 NO2  CO O3 TEMP   PRES  DEWP RAIN
## 1   1 2013     3   1    0     4    4   4   7 300 77 -0.7 1023.0 -18.8    0
## 2   2 2013     3   1    1     8    8   4   7 300 77 -1.1 1023.2 -18.2    0
## 3   3 2013     3   1    2     7    7   5  10 300 73 -1.1 1023.5 -18.2    0
## 4   4 2013     3   1    3     6    6  11  11 300 72 -1.4 1024.5 -19.4    0
## 5   5 2013     3   1    4     3    3  12  12 300 72 -2.0 1025.2 -19.5    0
## 6   6 2013     3   1    5     5    5  18  18 400 66 -2.2 1025.6 -19.6    0
## 7   7 2013     3   1    6     3    3  18  32 500 50 -2.6 1026.5 -19.1    0
## 8   8 2013     3   1    7     3    6  19  41 500 43 -1.6 1027.4 -19.1    0
## 9   9 2013     3   1    8     3    6  16  43 500 45  0.1 1028.3 -19.2    0
## 10 10 2013     3   1    9     3    8  12  28 400 59  1.2 1028.5 -19.3    0
## 11 11 2013     3   1   10     3    6   9  12 400 72  1.9 1028.2 -19.4    0
## 12 12 2013     3   1   11     3    6   9  14 400 71  2.9 1028.2 -20.5    0
## 13 13 2013     3   1   12     3    6   7  13 300 74  3.9 1027.3 -19.7    0
## 14 14 2013     3   1   13     3    6   7  12 400 76  5.3 1026.2 -19.3    0
## 15 15 2013     3   1   14     6    9   7  11 400 77  6.0 1025.9 -19.6    0
## 16 16 2013     3   1   15     8   15   7  14 400 76  6.2 1025.7 -18.6    0
```

```
## 17 17 2013     3   1   16     9   19   9  13 400 76  5.9 1025.6 -18.1    0
## 18 18 2013     3   1   17    10   23  11  15 400 74  4.3 1026.3 -18.7    0
## 19 19 2013     3   1   18    11   20   8  20 500 70  3.1 1027.4 -18.4    0
## 20 20 2013     3   1   19     8   14  12  30 500 60  2.3 1028.3 -18.4    0
##     wd WSPM      station
## 1  NNW  4.4 Aotizhongxin
## 2    N  4.7 Aotizhongxin
## 3  NNW  5.6 Aotizhongxin
## 4   NW  3.1 Aotizhongxin
## 5    N  2.0 Aotizhongxin
## 6    N  3.7 Aotizhongxin
## 7  NNE  2.5 Aotizhongxin
## 8  NNW  3.8 Aotizhongxin
## 9  NNW  4.1 Aotizhongxin
## 10   N  2.6 Aotizhongxin
## 11 NNW  3.6 Aotizhongxin
## 12   N  3.7 Aotizhongxin
## 13 NNW  5.1 Aotizhongxin
## 14  NW  4.3 Aotizhongxin
## 15  NW  4.4 Aotizhongxin
## 16 NNE  2.8 Aotizhongxin
## 17 NNW  3.9 Aotizhongxin
## 18 NNE  2.8 Aotizhongxin
## 19 NNE  2.1 Aotizhongxin
## 20   N  2.8 Aotizhongxin
```

```
summary(df)
```

```
##        No             year          month             day
##  Min.   :    1   Min.   :2013   Min.   : 1.000   Min.   : 1.00
##  1st Qu.: 8767   1st Qu.:2014   1st Qu.: 4.000   1st Qu.: 8.00
##  Median :17532   Median :2015   Median : 7.000   Median :16.00
##  Mean   :17532   Mean   :2015   Mean   : 6.523   Mean   :15.73
##  3rd Qu.:26298   3rd Qu.:2016   3rd Qu.:10.000   3rd Qu.:23.00
##  Max.   :35064   Max.   :2017   Max.   :12.000   Max.   :31.00
##
##       hour           PM2.5             PM10            SO2
##  Min.   : 0.00   Min.   :  3.00   Min.   :  2.0   Min.   :  0.2856
##  1st Qu.: 5.75   1st Qu.: 22.00   1st Qu.: 38.0   1st Qu.:  3.0000
##  Median :11.50   Median : 58.00   Median : 87.0   Median :  9.0000
##  Mean   :11.50   Mean   : 82.77   Mean   :110.1   Mean   : 17.3759
##  3rd Qu.:17.25   3rd Qu.:114.00   3rd Qu.:155.0   3rd Qu.: 21.0000
##  Max.   :23.00   Max.   :898.00   Max.   :984.0   Max.   :341.0000
##                  NA's   :925      NA's   :718     NA's   :935
##       NO2              CO              O3               TEMP
##  Min.   :  2.00   Min.   :  100   Min.   :  0.2142   Min.   :-16.80
##  1st Qu.: 30.00   1st Qu.:  500   1st Qu.:  8.0000   1st Qu.:  3.10
##  Median : 53.00   Median :  900   Median : 42.0000   Median : 14.50
##  Mean   : 59.31   Mean   : 1263   Mean   : 56.3534   Mean   : 13.58
##  3rd Qu.: 82.00   3rd Qu.: 1500   3rd Qu.: 82.0000   3rd Qu.: 23.30
##  Max.   :290.00   Max.   :10000   Max.   :423.0000   Max.   : 40.50
##  NA's   :1023     NA's   :1776    NA's   :1719       NA's   :20
##       PRES             DEWP              RAIN                wd
##  Min.   : 985.9   Min.   :-35.300   Min.   : 0.00000   NE     : 5140
##  1st Qu.:1003.3   1st Qu.: -8.100   1st Qu.: 0.00000   ENE    : 3950
```

```
##  Median :1011.4   Median :  3.800   Median : 0.00000   SW     : 3359
##  Mean   :1011.8   Mean   :  3.123   Mean   : 0.06742   E      : 2608
##  3rd Qu.:1020.1   3rd Qu.: 15.600   3rd Qu.: 0.00000   NNE    : 2445
##  Max.   :1042.0   Max.   : 28.500   Max.   :72.50000   (Other):17481
##  NA's   :20       NA's   :20        NA's   :20         NA's   :   81
##       WSPM               station
##  Min.   : 0.000   Aotizhongxin:35064
##  1st Qu.: 0.900
##  Median : 1.400
##  Mean   : 1.708
##  3rd Qu.: 2.200
##  Max.   :11.200
##  NA's   :14
```

```
dim(df)
```

```
## [1] 35064    18
```

Check Outliers

Verificamos para cada uma das colunas se existem valores outliers Para isso plotamos os dados utilizando o boxplot ....

```
#df %>% group_by(month) %>% ggplot(aes(group = month, y = TEMP)) + geom_boxplot()
#temp_out <- boxplot(TEMP~month+year , data=df)$out
#temp_out <- boxplot(df$TEMP~month)$out

#ed_exp1 <- df[c(10:21),c(2,6:7)]

#df_new <- df[-which(df$TEMP %in% temp_out),]
#boxplot(TEMP~month+year , data=df_new)

#boxplot(df.NEW_TEMP$TEMP)
#boxplot(df$TEMP)
```

Missing Values

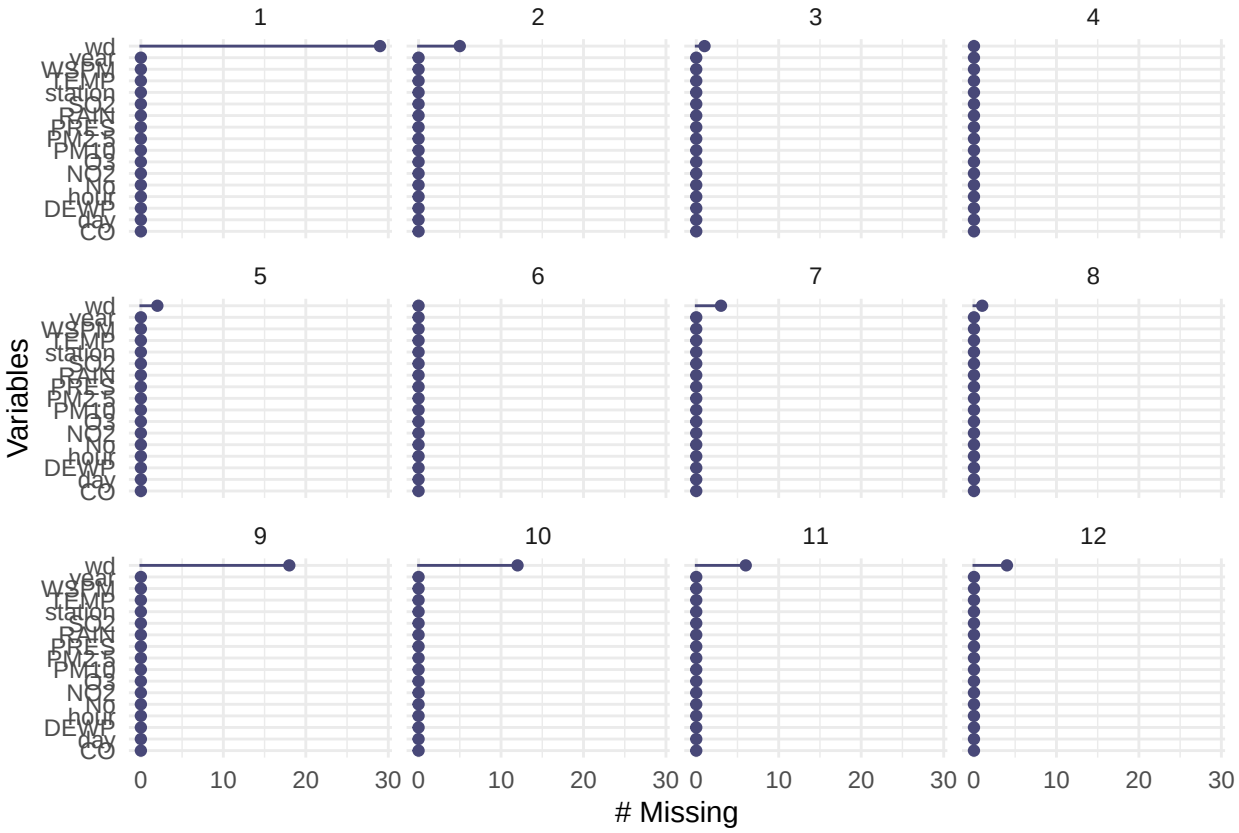
```
#function interpolation
interpolation_df <- function(df, col_names ){
  for(col in col_names){
    df[col] <- na.approx(df[col], rule=2)
  }
  return(df)
}

head(df)
```

```
##   No year month day hour PM2.5 PM10 SO2 NO2  CO O3 TEMP   PRES  DEWP RAIN
## 1  1 2013     3   1    0     4    4   4   7 300 77 -0.7 1023.0 -18.8    0
## 2  2 2013     3   1    1     8    8   4   7 300 77 -1.1 1023.2 -18.2    0
## 3  3 2013     3   1    2     7    7   5  10 300 73 -1.1 1023.5 -18.2    0
## 4  4 2013     3   1    3     6    6  11  11 300 72 -1.4 1024.5 -19.4    0
## 5  5 2013     3   1    4     3    3  12  12 300 72 -2.0 1025.2 -19.5    0
## 6  6 2013     3   1    5     5    5  18  18 400 66 -2.2 1025.6 -19.6    0
##    wd WSPM      station
```

```
## 1 NNW  4.4 Aotizhongxin
## 2   N  4.7 Aotizhongxin
## 3 NNW  5.6 Aotizhongxin
## 4  NW  3.1 Aotizhongxin
## 5   N  2.0 Aotizhongxin
## 6   N  3.7 Aotizhongxin
```

```
col_names = c("PM2.5", "PM10", "SO2", "NO2", "CO", "O3", "TEMP", "PRES", "DEWP", "RAIN", "WSPM")
df <- interpolation_df(df, col_names)
gg_miss_var(df, facet = month)
```

```
## Warning: `cols` is now required.
## Please use `cols = c(data)`
```




```
#falta remover os missing values da coluna wd
#sum(is.na(df$wd))
```

Calc a class variable

```
df$aqi<-NA
df[, "aqi"] <- apply(df[, 6:11], 1, max)

head(df)
```

```
##   No year month day hour PM2.5 PM10 SO2 NO2  CO O3 TEMP   PRES  DEWP RAIN
## 1  1 2013     3   1    0     4    4   4   7 300 77 -0.7 1023.0 -18.8    0
## 2  2 2013     3   1    1     8    8   4   7 300 77 -1.1 1023.2 -18.2    0
## 3  3 2013     3   1    2     7    7   5  10 300 73 -1.1 1023.5 -18.2    0
## 4  4 2013     3   1    3     6    6  11  11 300 72 -1.4 1024.5 -19.4    0
```

```
## 5  5 2013     3   1    4     3    3  12  12 300 72 -2.0 1025.2 -19.5    0
## 6  6 2013     3   1    5     5    5  18  18 400 66 -2.2 1025.6 -19.6    0
##    wd WSPM      station aqi
## 1 NNW  4.4 Aotizhongxin 300
## 2   N  4.7 Aotizhongxin 300
## 3 NNW  5.6 Aotizhongxin 300
## 4  NW  3.1 Aotizhongxin 300
## 5   N  2.0 Aotizhongxin 300
## 6   N  3.7 Aotizhongxin 400
```

Data exploratory analysis

Predictive modelling