Import dataset from the following link: AirQuality Data Set

Perform the following written operations:

1. Read the file in Zip format and get it into R.

2. Create Univariate for all the columns.

**3. Check for missing values in all columns.**

require(ggplot2)

require(moments)

aq <- data.frame(airquality)

colSums(is.na(aq))

4. Impute the missing values using appropriate methods.

**5. Create bi-variate analysis for all relationships.**

# Only numerical variables can be used for correlation

columns <- c('Ozone.sqrt', 'Solar.R', 'Wind', 'Temp')

# We identify non-na rows as rows with na values cause problems

rows <- rowSums(is.na(aq)) == 0

# Compute correlation of variables and round to 2 digits

round(cor(aq[rows, columns]), 2)

6. Test relevant hypothesis for valid relations.

7. Create cross tabulations with derived variables.

8. Check for trends and patterns in time series.

9. Find out the most polluted time of the day and the name of the chemical compound.

require("datasets")

data("airquality")

str(airquality)

col1<- mapply(anyNA,airquality) # apply function anyNA() on all columns of airquality dataset

col1

## Ozone Solar.R Wind Temp Month Day

## TRUE TRUE FALSE FALSE FALSE FALSE

# Impute monthly mean in Ozone

for (i in 1:nrow(airquality)){

if(is.na(airquality[i,"Ozone"])){

airquality[i,"Ozone"]<- mean(airquality[which(airquality[,"Month"]==airquality[i,"Month"]),"Ozone"],na.rm = TRUE)

}

# Impute monthly mean in Solar.R

if(is.na(airquality[i,"Solar.R"])){

airquality[i,"Solar.R"]<- mean(airquality[which(airquality[,"Month"]==airquality[i,"Month"]),"Solar.R"],na.rm = TRUE)

}