SESSION 10: Correlations

Assignment 1

Problem Statement

1. Import dataset from the following link:

https://archive.ics.uci.edu/ml/machine-learning-databases/00360/

Perform the below written operations:

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

b. Create Univariate for all the columns.

c. Check for missing values in all columns.

d. Impute the missing values using appropriate methods

e. Create bi-variate analysis for all relationships

f. Test relevant hypothesis for valid relations

g. Create cross tabulations with derived variables

h. check for trends and patterns in time series

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

Miss <- read.csv(file.choose(),na.strings=c(""))

```
summarize(CO(GT), PT08.S1(CO) = PTS1, NMHC(GT) = NMHCGT, PT08.S2(NMHC) =
PTNMHC , NOx(GT) = NxGT , PT08.S3(NOx) = PTS3Nx , NO2(GT) = NGT ,
PT08.S4(NO2) = PTS4 , PT08.S5(O3) = PTS5)
```

```
na.test <-  function (data) {
  if (colSums(!is.na(data) == 0)){
     stop ("The some variable in the dataset has all missing value,
    remove the column to proceed")
     }
     }
na.test (test1)
mice_imputes = mice(nhanes, m=5, maxit = 40)
```

```
ggplot(tdata, aes(x= PT08.S4(NO2), y= PT08.S5(O3))) +
geom_point(shape=1) +
geom_smooth(method=lm)
ggsave(file="scatter-plot.png", dpi=500)
```

```
with(Chlordata, tapply(Chlorophyll, list(Treatment=Treatment,Stage=Stage), mean) )
with(Chlordata, tapply(Chlorophyll, list(Treatment=Treatment,Stage=Stage), sd) )
```

```
ts (inputData, frequency = 4, start = c(1959, 2)) # frequency 4 => Quarterly Data
ts (1:10, frequency = 12, start = 1990) # freq 12 => Monthly data.
ts (inputData, start=c(2009), end=c(2014), frequency=1) # Yearly Data
```