Data Cleaning-02 Duplicates & Outliers

Created by H. M. Samadhi Chathuranga Rathnayake

```r
setwd("D:\\Workshops\\R Programming for Data Science Workshop\\Part 02 - Data Manipulation & Cleaning\\Datasets")

#Dealing with Duplicated Rows
data=read.csv("iris - Duplicate.CSV")
head(data)

#Base R
duplicated(data)#Detecting duplicates

df_dup=data.frame(table(data$Id))
df_dup[df_dup$Freq>1,]

new_data1=data[!duplicated(data),] #Removing duplicated rows
head(new_data1)

new_data2=unique(data)
head(new_data2)

#dplyr library
library(dplyr)

new_data=distinct(data,Id,.keep_all = TRUE)
head(new_data)

data%>%distinct(Id,.keep_all=TRUE)->new_data1
head(new_data1)

data%>%distinct(Sepal.Length,Sepal.Width,.keep_all=TRUE)->new_data2
head(new_data2)

#Dealing with Outliers
data=read.csv("iris - Outliers.CSV")
head(data)

num_data=data[,c("Sepal.Length","Sepal.Width","Petal.Length","Petal.Width")]
head(num_data)

boxplot(num_data)

boxplot(num_data$Sepal.Width)$out

Q=quantile(num_data$Sepal.Width,probs = c(0.25,0.75))
Q
```

```r
iqr=IQR(num_data$Sepal.Width)

upper=Q[2]+1.5*iqr # Upper Range
lower=Q[1]-1.5*iqr # Lower Range

#Remove the most significant outliers
num_data_new=num_data[num_data$Sepal.Width > lower & num_data$Sepal.Width <
upper,]
head(num_data_new)

boxplot(num_data_new)
```

```r
data_new=data[data$Sepal.Width > lower & data$Sepal.Width < upper,]
head(data_new)
```