Data Analytics Lab 5

09/10/19

Step 1. Data Imputation using Mean and median

```
rm(list=ls())
```

```
setwd("your data directory")
```

Read the data and check the content.

```
Data <- read.csv("DI.csv", header=TRUE, sep= ";")
```

```
Data
```

```
str(Data)
```

```
attach(Data)
```

Try to get Mean and Median!

```
mean(Post_BP)
```

```
median(Post_BP)
```

See the missing data:

```
Data$Post_BP[is.na(Data$Post_BP)]
```

Find the Mean and Median of non-missing Data:

```
mean(Data$Post_BP[!is.na(Data$Post_BP)])
```

```
median(Data$Post_BP[!is.na(Data$Post_BP)])
```

Make a copy of your Data:

```
Data1 <- Data
```

Use Mean of your data as replacement for the missing values.

```
Data1$Post_BP[is.na(Data1$Post_BP)] <- mean(Data1$Post_BP[!is.na(Data1$Post_BP)])
```

See what happened:

```
Data1$Post_BP
```

Make a copy of your Data:

```
Data2 <- Data
```

Use Median of your data as replacement for the missing values.

```
Data2$Post_BP[is.na(Data2$Post_BP)] <- median(Data2$Post_BP[!is.na(Data2$Post_BP)])
```

See what happened:

```
Data1$Post_BP
```

Step 2. Data Imputation using Regression

Remove column ID from your data:

```
Data <- Data[,-1]
```

Check the correlation matrix in your data:

```
cor(Data)
```

Remove the missing data and check the correlation matrix again:

```
cor(Data, use = "complete.obs")
```

Let's use some symbols to see the correlations better.

```
symnum(cor(Data, use = "complete.obs"))
```

Define a new column such that its values are 0 when data for column u is missing, and equals 1 otherwise.

```
Ind_Function <- function(u)
{
  x <- dim(length(u))
  x[which(is.na(u))] <- 0
  x[which(!is.na(u))] <- 1
  return(x)
}
```

Generate the column using the function above for the variable Post_BP:

```
Data$I <- Ind_Function(Data$Post_BP)
Data
```

Use a regression model for the Post_BP using the variable Pre_BP as the independent variable.

```
Model<- lm(Post_BP ~ Pre_BP)
```

Identify the intersection and slope:

```
summary(Model)
```

Use the regression model that you got to fill all missing values:

```
for(i in 1:nrow(Data))
{
  if (Data$I[i] == 0)

 { Data$Post_BP[i] = “intercept” + “slope” *Pre_BP[i]
}
}
```

Check your data for Post_BP.

```
Data$Post_BP
```