Data preparation using R

```
#Import
   packages-----------------------------------------------------
library(tidyverse)
library(readxl)
library(writexl)

#Data Preparation----------------------------------------------------

#Dataset period
Start_year = 2000
End_year = 2020


Main_dataframe <- data.frame()

#Creating the indexing year vaector
Years <- 2000:2020
Main_dataframe <- data.frame(Years=2000:2020)

#Import GDP data---------------------------------------------------
GDP_raw <- read_excel("GDPNEW.xls", sheet = "Data")
View(GDP_raw)

#columindex vector
C_index <- c()
for (year in Years){
     C_index
        <-append(C_index,which(colnames(GDP_raw)==as.character(year)))
}

#Select the data
SL_GDP <- GDP_raw[GDP_raw$`Country Name`=="Sri Lanka",C_index]
```

```r
SL_GDP <-unlist(SL_GDP , use.names = F)
#Normalize the data and add to the main data frame
Main_dataframe$GDP <- (SL_GDP - mean(SL_GDP))/sd(SL_GDP)
hist((SL_GDP - mean(SL_GDP))/sd(SL_GDP))


Data_bundle <- read_csv("Bundle.csv")
View(Data_bundle)

#Import the CAR
   data---------------------------------------------------------------------------
SL_CAR <- unlist(Data_bundle[2,],use.names =
   F)[4:length(colnames(Data_bundle))]
Main_dataframe$`Capital adequacy ratio` <- as.double(SL_CAR)/100

#Import Return on
   equity-------------------------------------------------------------------------
SL_ROE <- unlist(Data_bundle[3,],use.names =
   F)[4:length(colnames(Data_bundle))]
Main_dataframe$`Return On Equity (ROE)` <- as.double(SL_ROE)/100

#Import Return on
   assets-------------------------------------------------------------------------
SL_ROA <- unlist(Data_bundle[4,],use.names =
   F)[4:length(colnames(Data_bundle))]
Main_dataframe$`Return On asset (ROA)` <- as.double(SL_ROA)/100

#Import Net Interest
   Margin-------------------------------------------------------------------------
SL_Net_interest_margin <- unlist(Data_bundle[5,],use.names =
   F)[4:length(colnames(Data_bundle))]
Main_dataframe$`Net Interest Margin` <-
   as.double(SL_Net_interest_margin)/100

#Import the inflation
   rates--------------------------------------------------------------------------
RAW_inflation_data <- read_csv("inflation_data.csv")
View(RAW_inflation_data)

#columindex vector
C_index <- c()
for (year in Years){
  C_index
     <-append(C_index,which(colnames(RAW_inflation_data)==as.character(year)))
}

#Select the data
```

```
SL_Inflation <- RAW_inflation_data[RAW_inflation_data$`Country
   Name`=="Sri Lanka",C_index]
SL_Inflation <-unlist(SL_Inflation , use.names = F)

Main_dataframe$Inflation <- SL_Inflation/100


#-----------------------------------------------------------------------------
#Outliar detection

#GDP-----------------------------------
#Discriptive statistics
summary(Main_dataframe$GDP)

boxplot(SL_GDP,
      ylab = "hwy"
)

#The boxplot shows there is no any out liars
#CAR-----------------------------------
boxplot(Main_dataframe$`Capital adequacy ratio`,
      ylab = "hwy"
)
#ROE-----------------------------------
boxplot(Main_dataframe$`Return On Equity (ROE)`,
      ylab = "hwy"
)
#ROA----------------------------------
boxplot(Main_dataframe$`Return On asset (ROA)`,
      ylab = "hwy"
)
#Net interest margin
boxplot(Main_dataframe$`Net Interest Margin`,
      ylab = "hwy"
)
#Inflation
boxplot(Main_dataframe$Inflation,
      ylab = "hwy"
)
#In box plot we can see a outliar so we must handle this situation
ggplot(data = Main_dataframe) + geom_point(aes(x=Years ,
   y=Inflation))

quartiles <- quantile(Main_dataframe$Inflation, probs=c(.25, .75),
   na.rm = FALSE)
IQR <- IQR(Main_dataframe$Inflation)

Lower <- quartiles[1] - 1.5*IQR
```

```
Upper <- quartiles[2] + 1.5*IQR

for (x in 1:21){
  if (Main_dataframe[x,7]>Upper){
    print(x)
  }
}
#It's the 9 th row value, so we remove the data and predict a value
    for that using
#linearregreesion model on year and inflation rate

train <- data.frame(yr=Years[-9],Infla=Main_dataframe$Inflation[-9])

Inflation_model <- lm(Infla ~ yr , data = train)

summary(Inflation_model)

pred<-data.frame(yr=2008)
new_value <- predict(Inflation_model,pred)

Main_dataframe[9,7]=new_value
boxplot(Main_dataframe$Inflation,
        ylab = "hwy"
)

#There are some outliars still there so we must remove them

quartiles <- quantile(Main_dataframe$Inflation, probs=c(.25, .75),
    na.rm = FALSE)
IQR <- IQR(Main_dataframe$Inflation)

Lower <- quartiles[1] - 1.5*IQR
Upper <- quartiles[2] + 1.5*IQR

for (x in 1:21){
  if (Main_dataframe[x,7]>Upper){
    print(x)
  }
}

pred<-data.frame(yr=Years[c(2,8)])
new_value <- predict(Inflation_model,pred)

Main_dataframe[c(2,8),7]=new_value
boxplot(Main_dataframe$Inflation,
        ylab = "hwy"
)
```

```
#Now the data are ok!

#In box plot we can see a outliar so we must handle this situation
ggplot(data = Main_dataframe) + geom_point(aes(x=Years ,
   y=Inflation))

#---------------------------------------------------------------------------
write_csv(Main_dataframe,file="Main_data_frame.csv")
```

Exploratory Data Analysis (EDA)

```
  #Import relevent libraries
library(tidyverse)
library(corrplot)
#Load the data set
df = read_csv("Main_data_frame.csv")
View(df)

#Discriptive statistics
summary(df)

#GDP----------------------------------------------------------------------
plt1 <- df %>% select(GDP) %>%
  ggplot(aes(x="", y = GDP)) +
  geom_boxplot(fill = "lightblue", color = "black") +
  coord_flip() +
  theme_classic() +
  xlab("") +
  theme(axis.text.y=element_blank(),
        axis.ticks.y=element_blank())

plt2 <- df %>% select(GDP) %>%
  ggplot() +
  geom_histogram(aes(x = GDP, y = (..count..)),
                 position = "identity", binwidth = 0.29,
                 fill = "lightblue", color = "black") +
  ylab("Frequency") +
  theme_classic()

egg::ggarrange(plt2, plt1, heights = 2:1)
cowplot::plot_grid(plt2, plt1,
                   ncol = 1, rel_heights = c(2, 1),
                   align = 'v', axis = 'lr')
library(patchwork)
plt2 + plt1 + plot_layout(nrow = 2, heights = c(2, 1))
```

```
#Capital adequacy ratio---------------------------------------------

plt1 <- df %>% select(`Capital adequacy ratio`) %>%
  ggplot(aes(x="", y = `Capital adequacy ratio`)) +
  geom_boxplot(fill = "lightblue", color = "black") +
  coord_flip() +
  theme_classic() +
  xlab("") +
  theme(axis.text.y=element_blank(),
        axis.ticks.y=element_blank())

plt2 <- df %>% select(`Capital adequacy ratio`) %>%
  ggplot() +
  geom_histogram(aes(x = `Capital adequacy ratio`, y = (..count..)),
                 position = "identity", binwidth = 0.025,
                 fill = "lightblue", color = "black") +
  ylab("Frequency") +
  theme_classic()

egg::ggarrange(plt2, plt1, heights = 2:1)
cowplot::plot_grid(plt2, plt1,
                   ncol = 1, rel_heights = c(2, 1),
                   align = 'v', axis = 'lr')
library(patchwork)
plt2 + plt1 + plot_layout(nrow = 2, heights = c(2, 1))

#Return On Equity (ROE)------------------------------------------

plt1 <- df %>% select(`Return On Equity (ROE)`) %>%
  ggplot(aes(x="", y = `Return On Equity (ROE)`)) +
  geom_boxplot(fill = "lightblue", color = "black") +
  coord_flip() +
  theme_classic() +
  xlab("") +
  theme(axis.text.y=element_blank(),
        axis.ticks.y=element_blank())

plt2 <- df %>% select(`Return On Equity (ROE)`) %>%
  ggplot() +
  geom_histogram(aes(x = `Return On Equity (ROE)`, y = (..count..)),
                 position = "identity", binwidth = 0.025,
                 fill = "lightblue", color = "black") +
  ylab("Frequency") +
  theme_classic()
```

```
egg::ggarrange(plt2, plt1, heights = 2:1)
cowplot::plot_grid(plt2, plt1,
                   ncol = 1, rel_heights = c(2, 1),
                   align = 'v', axis = 'lr')
library(patchwork)
plt2 + plt1 + plot_layout(nrow = 2, heights = c(2, 1))

#Return On asset (ROA)-------------------------------------------

plt1 <- df %>% select(`Return On asset (ROA)`) %>%
  ggplot(aes(x="", y = `Return On asset (ROA)`)) +
  geom_boxplot(fill = "lightblue", color = "black") +
  coord_flip() +
  theme_classic() +
  xlab("") +
  theme(axis.text.y=element_blank(),
        axis.ticks.y=element_blank())

plt2 <- df %>% select(`Return On asset (ROA)`) %>%
  ggplot() +
  geom_histogram(aes(x = `Return On asset (ROA)`, y = (..count..)),
                 position = "identity", binwidth = 0.004,
                 fill = "lightblue", color = "black") +
  ylab("Frequency") +
  theme_classic()

egg::ggarrange(plt2, plt1, heights = 2:1)
cowplot::plot_grid(plt2, plt1,
                   ncol = 1, rel_heights = c(2, 1),
                   align = 'v', axis = 'lr')
library(patchwork)
plt2 + plt1 + plot_layout(nrow = 2, heights = c(2, 1))

#Net Interest Margin-------------------------------------------

plt1 <- df %>% select(`Net Interest Margin`) %>%
  ggplot(aes(x="", y = `Net Interest Margin`)) +
  geom_boxplot(fill = "lightblue", color = "black") +
  coord_flip() +
  theme_classic() +
  xlab("") +
  theme(axis.text.y=element_blank(),
        axis.ticks.y=element_blank())

plt2 <- df %>% select(`Net Interest Margin`) %>%
  ggplot() +
  geom_histogram(aes(x = `Net Interest Margin`, y = (..count..)),
                 position = "identity", binwidth = 0.002,
```

```
             fill = "lightblue", color = "black") +
  ylab("Frequency") +
  theme_classic()

egg::ggarrange(plt2, plt1, heights = 2:1)
cowplot::plot_grid(plt2, plt1,
                  ncol = 1, rel_heights = c(2, 1),
                  align = 'v', axis = 'lr')
library(patchwork)
plt2 + plt1 + plot_layout(nrow = 2, heights = c(2, 1))

#Inflation-------------------------------------------------------

plt1 <- df %>% select(`Inflation`) %>%
  ggplot(aes(x="", y = `Inflation`)) +
  geom_boxplot(fill = "lightblue", color = "black") +
  coord_flip() +
  theme_classic() +
  xlab("") +
  theme(axis.text.y=element_blank(),
        axis.ticks.y=element_blank())

plt2 <- df %>% select(`Inflation`) %>%
  ggplot() +
  geom_histogram(aes(x = `Inflation`, y = (..count..)),
             position = "identity", binwidth = 0.02,
             fill = "lightblue", color = "black") +
  ylab("Frequency") +
  theme_classic()

egg::ggarrange(plt2, plt1, heights = 2:1)
cowplot::plot_grid(plt2, plt1,
                  ncol = 1, rel_heights = c(2, 1),
                  align = 'v', axis = 'lr')
library(patchwork)
plt2 + plt1 + plot_layout(nrow = 2, heights = c(2, 1))

#Discriptive statistics

summary(df$Inflation)

#Correlation
M<-cor(df)
head(round(M,2))

corrplot(M, method="color",tl.col = "black",addCoef.col = "black")

#Scatter plot
```

```r
library(GGally)
# Loading
library(ggcorrplot)

pairs(df[,-1],
     col = "blue",
     pch = 19,
     cex = 1,
     labels = c("GDP","Capital adequacy ratio","Return On Equity
        (ROE)",
              "Return On asset (ROA)","Net Interest
                 Margin","Inflation"),
     gap = 0.3,
     upper.panel = NULL,       # Turn off the upper panel above the
        diagonal
     main = "Scatterplot Matrix of dataset",
     cex.labels = 1.5)
corr <- round(cor(df[,-1]), 4)
ggcorrplot(corr,
         hc.order = TRUE,
         type = "lower",
         outline.color = "white",
         lab = TRUE)
#-----------------------------------------------------------------------------
Model1 <- lm(data = df,formula = `Return On asset (ROA)` ~ `Capital
   adequacy ratio` +
           `Inflation` + `Return On Equity (ROE)` )

moel_f <- data.frame(residuals=Model1$residuals)
plt1 <- moel_f %>% select(residuals) %>%
  ggplot(aes(x="", y = residuals)) +
  geom_boxplot(fill = "lightblue", color = "black") +
  coord_flip() +
  theme_classic() +
  xlab("") +
  theme(axis.text.y=element_blank(),
      axis.ticks.y=element_blank())

plt2 <- moel_f %>% select(residuals) %>%
  ggplot() +
  geom_histogram(aes(x = residuals, y = (..count..)),
             position = "identity", binwidth = 0.0009,
             fill = "lightblue", color = "black") +
  ylab("Frequency") +
  theme_classic()

egg::ggarrange(plt2, plt1, heights = 2:1)
cowplot::plot_grid(plt2, plt1,
```

```
               ncol = 1, rel_heights = c(2, 1),
               align = 'v', axis = 'lr')
library(patchwork)
plt2 + plt1 + plot_layout(nrow = 2, heights = c(2, 1))
```

Quantitative Analysis

```r
#Previously We discuss the characteristic of the indipendent
    variables and now
#We are going to Build Regression model.

library(tidyverse)
library(olsrr)

#Import the Data
DF <- read_csv("Main_data_frame.csv")

#In previously we dicided to remove the GDP variable from the model
   because it's
#not fullfill the assumptions of MLP.

New_DF <- DF[,-2]
glimpse(New_DF)
colnames(New_DF)<-c("Years","CAR","ROE","ROA","NIM","INF")
View(New_DF)

#------------------------------------------------#
# 01 - Find the best regression model
#-----------------------------------------------#

###########################
#ALL POSSIBLE REGRESSIONS
##########################

#Creating the model using all Variables
model <- lm(data=New_DF,ROE~CAR+ROA+NIM+INF)
model$coefficients
summary(model)
anova(model)
```

```
ols_plot_resid_fit(model) + labs(subtitle = "Model 01")


Sum_1 <-data.frame(confint(model,
   c('(Intercept)','CAR','ROA','NIM','INF'), level=0.95))
confINT <-
   data.frame(Variables=c('(Intercept)','CAR','ROA','NIM','INF'),
mean=c(0.28265661 ,-1.32473051 ,11.17604361, -3.42957033 ,
   0.03307611 ))
confINT$low = Sum_1$X2.5..
confINT$high = Sum_1$X97.5..

ggplot(confINT,          ### The data frame to use.
      aes(x = Variables,
         y = mean)) +
  geom_errorbar(aes(ymin = low,
                 ymax = high),
              width = 0.05,
              size = 0.5) +
  geom_point(shape = 15,
            size = 1) +
  theme_bw() +
  theme(axis.title = element_text(face = "bold")) +

  ylab("Mean steps")

Models <-
   select(ols_step_all_possible(model)$result,n,predictors,rsquare,adjr)
Models$rsquare <-round(Models$rsquare*100,4)
Models$adjr <-round(Models$adjr*100,4)
View(Models)
confint(model)
model1_summary<-summary(model)
model1_summary$r.squared
#Using the R square values of all possible models, we can say the
   best model is
#that the model use "CAR"(Capital adequacy ratio),"ROA"(Return On
   asset (ROA)) and
#"NIM"(Net Interest Margin)

Best_model_1 <- lm(data = New_DF,ROE~CAR+ROA+NIM)
summary(Best_model_1)
Anova_result <- anova(Best_model_1)
Anova_result
ols_plot_resid_fit(Best_model_1) + labs(subtitle = "Model 02")

ols_test_breusch_pagan(Best_model_1)
```

```
###################################
Sum_2 <-data.frame(confint(Best_model_1,
   c('(Intercept)','CAR','ROA','NIM'), level=0.95))
confINT2 <-
   data.frame(Variables=c('(Intercept)','CAR','ROA','NIM'),mean=c(0.2879
   , -1.3520 , 11.2488 , -3.4387 ))
confINT2$low = Sum_2$X2.5..
confINT2$high = Sum_2$X97.5..

ggplot(confINT2,         ### The data frame to use.
      aes(x = Variables,
         y = mean)) +
  geom_errorbar(aes(ymin = low,
                ymax = high),
             width = 0.05,
             size = 0.5) +
  geom_point(shape = 15,
           size = 1) +
  theme_bw() +
  theme(axis.title = element_text(face = "bold")) +

  ylab("Mean steps")

############################
#Partial F test
################################

anova(model,Best_model_1)

#The p value is 0.8823 so it's greater than our significant level
   0.05 so
#we cant reject the null hypothesis. So the reduced model is more
   suitable.

################################
#plot the residuals
###################################
# create a new plotting window and set the plotting area into a 1*2
   array
qqnorm(Best_model_1$residuals,main = "Model 02 residual plot")
qqline(Best_model_1$residuals,main = "Model 02 residual plot")

qqnorm(model$residuals,main = "Model 01 residual plot")
qqline(model$residuals,main = "Model 01 residual plot")

hist(Best_model_1$residuals,main = "Model 02 ",xlab = 'residual')
hist(model$residuals,main = "Model 01 ",xlab = 'residual')
```

```
######################################################################
library(plyr)
library(ggplot2)
library(scales)
df<- data.frame(x = c("Capital Adequacy Ratio (CAR)", "eturn on
   Asset(ROA)", "Net Interest Margin (NIM) "),
              y = c(-1.3520 , 11.2488 , -3.4387))
ggplot(df,aes(x=x,fill=y,y=y))+
  # Plotting the bar plot using geom_bar()
  geom_bar(stat='identity',width=0.5)+coord_flip()+
  # Adding text labels to bar in the bar plot using geom_text()
  geom_text(aes(label=round(y,3))) + labs(xlab='value of
     coefficent',ylab='Variable',title = 'Coefficent values of Model
     02')
```