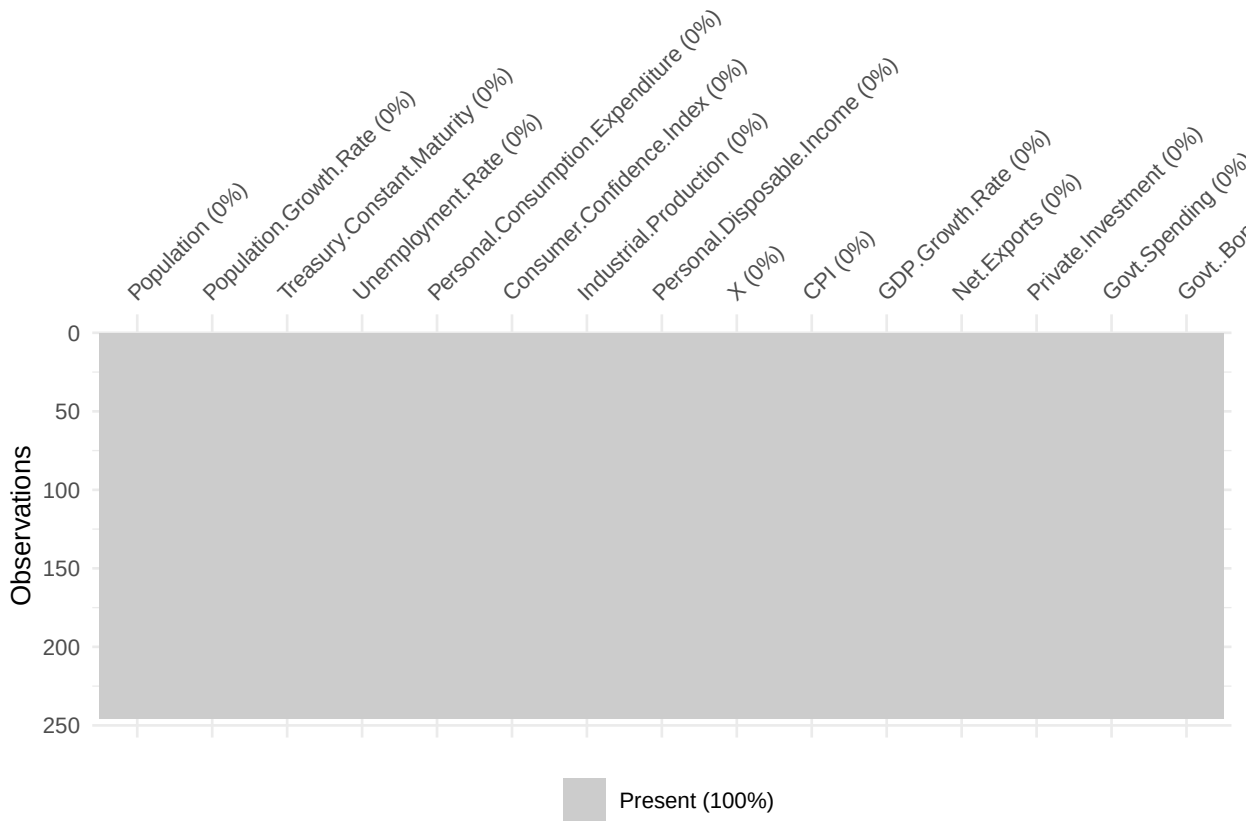
Predicting US GDP Growth Rate using Multiple Linear Regression

Mohammed Inamalhasan Dastagir Faras

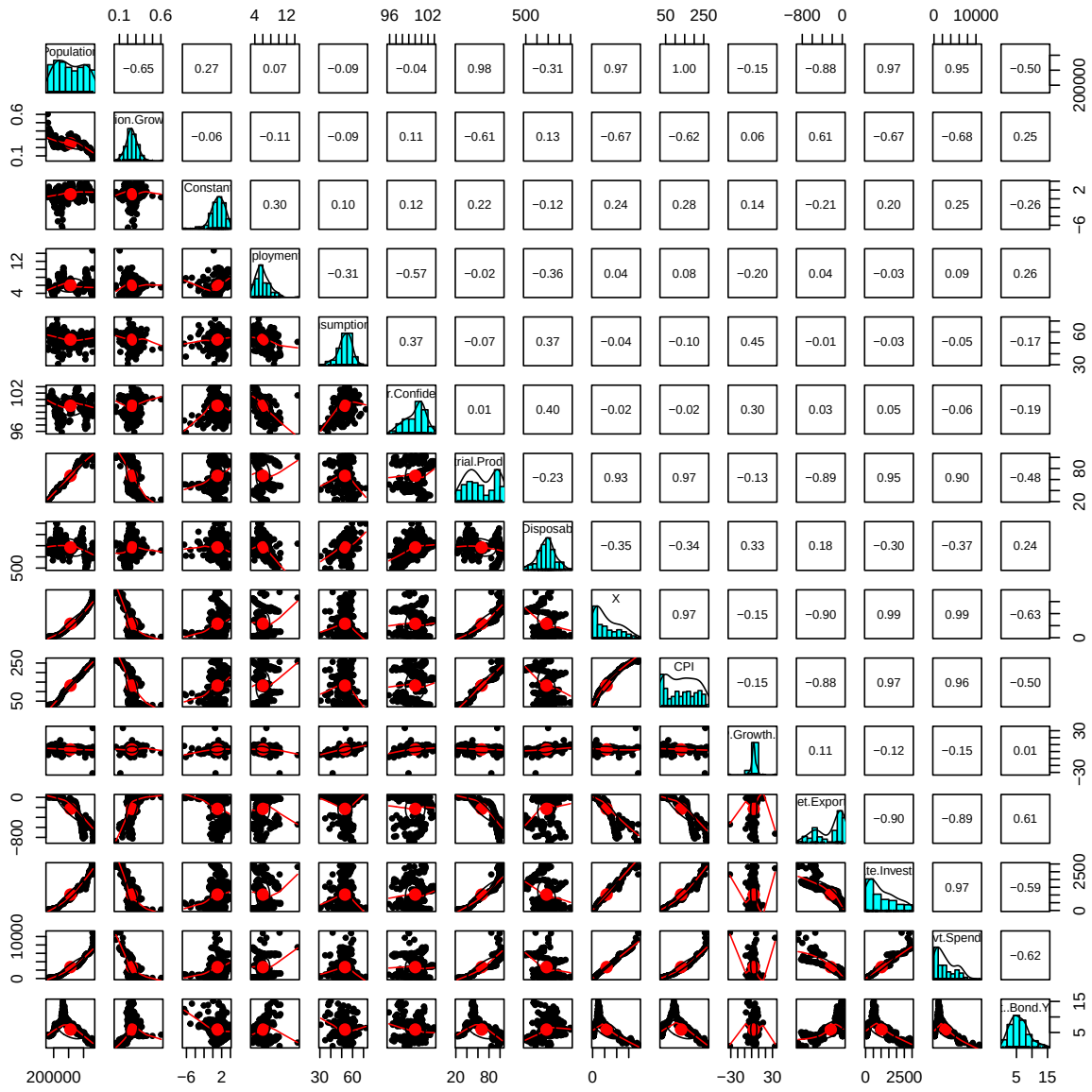
21/02/2022

This project aims at the prediction of GDP Growth rate by using factors that contribute to it.

Exploratory Data Analysis



By looking at the above plot, we see there are no NA's in the data.



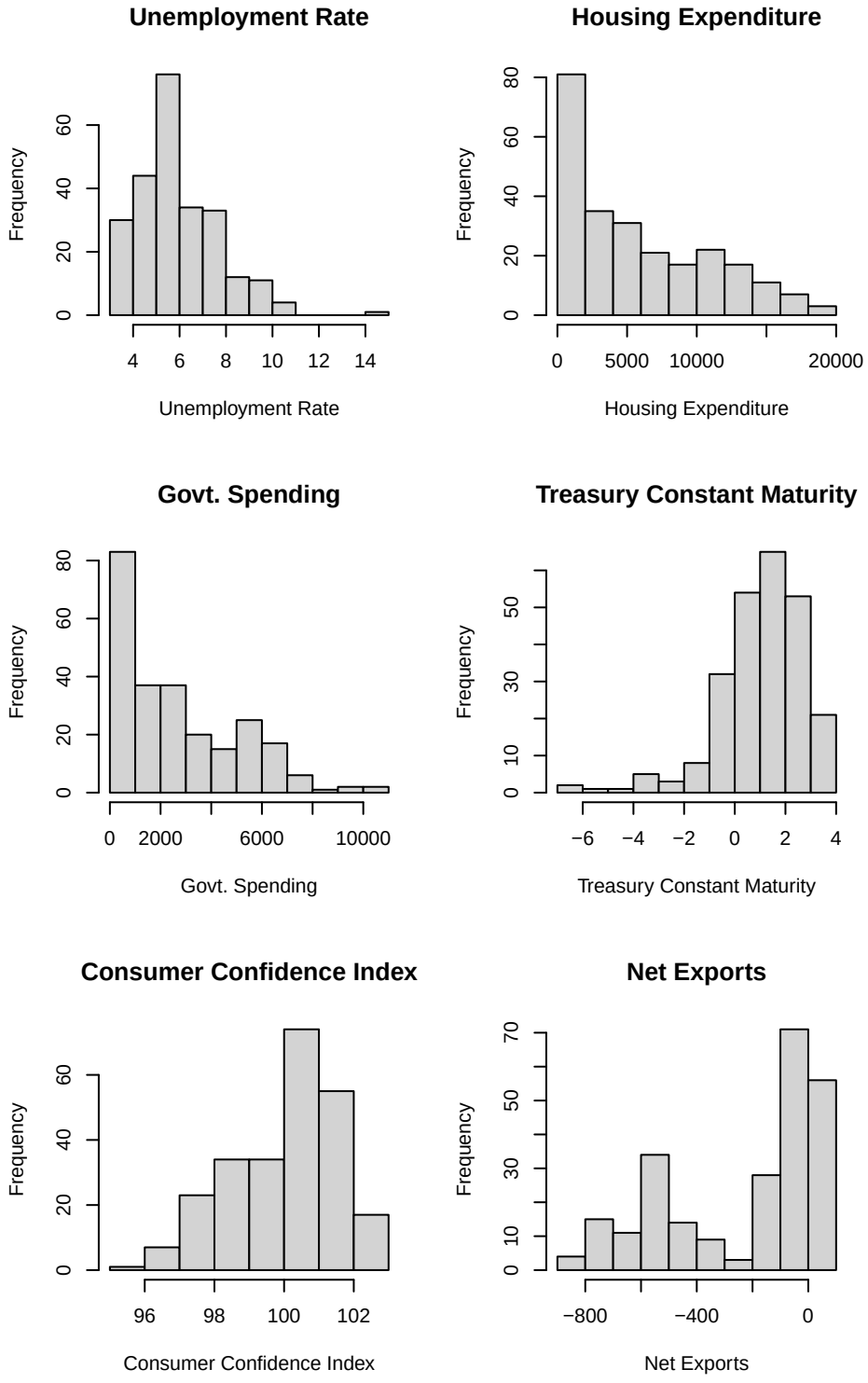
Insights from the panels plot:

1. There appears to be skewness in some variables
2. There is multicollinearity between variables

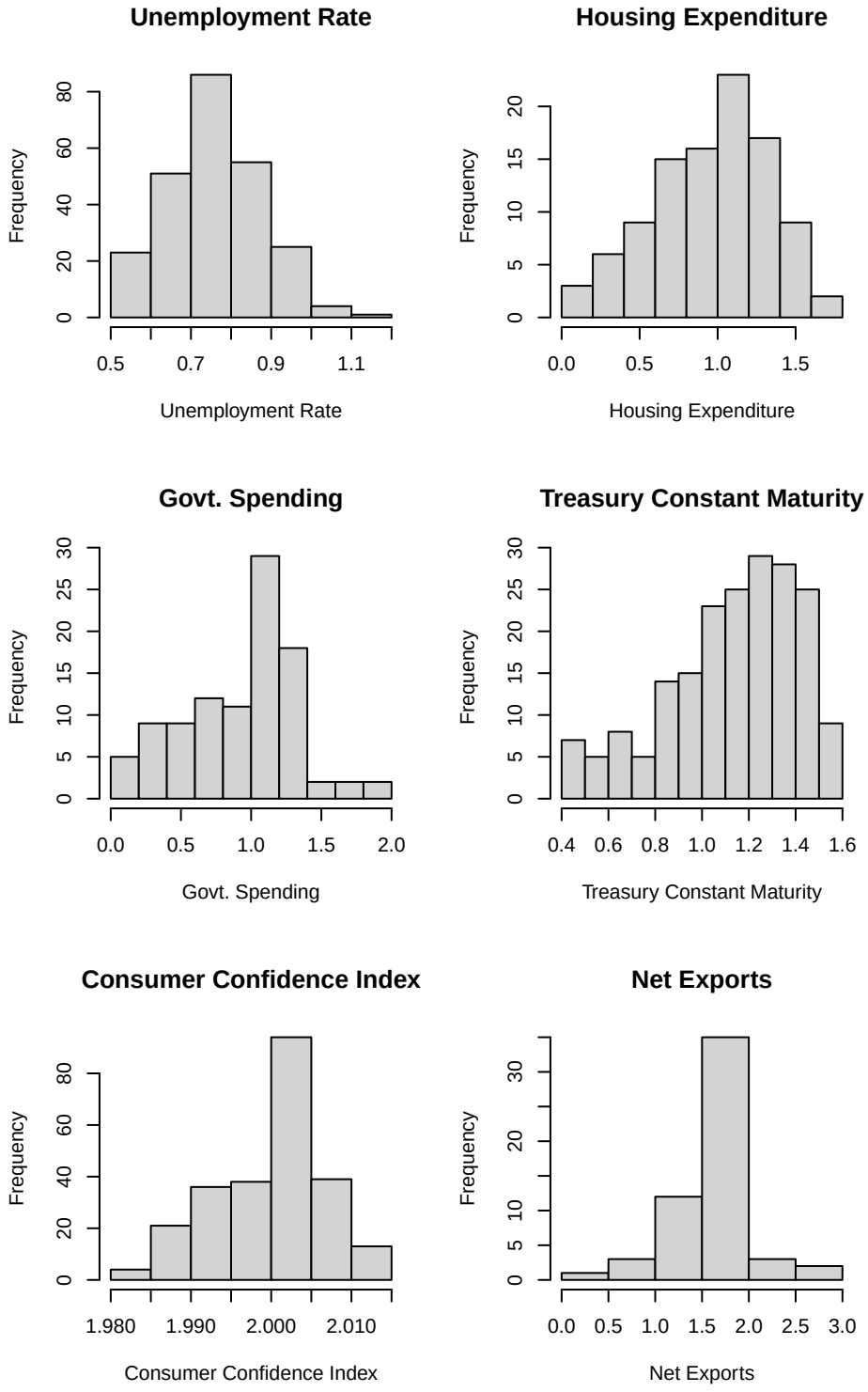
Variables that do not follow a normal distribution are: Treasury Constant Maturity, Unemployment Rate, Consumer Confidence Index, X, CPI, Gross Private Domestic Investment, Govt Expenditure, Private Investment and Govt Spending.

In order to satisfy the assumptions of Linear Regression, these distributions have to be transformed to be normal.

Before performing transformations

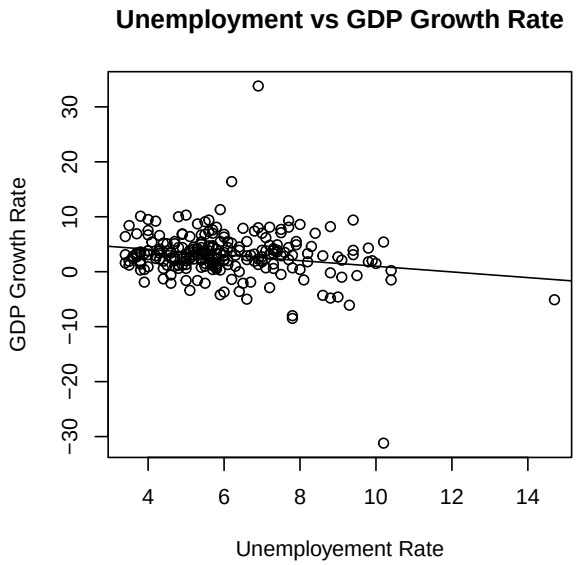


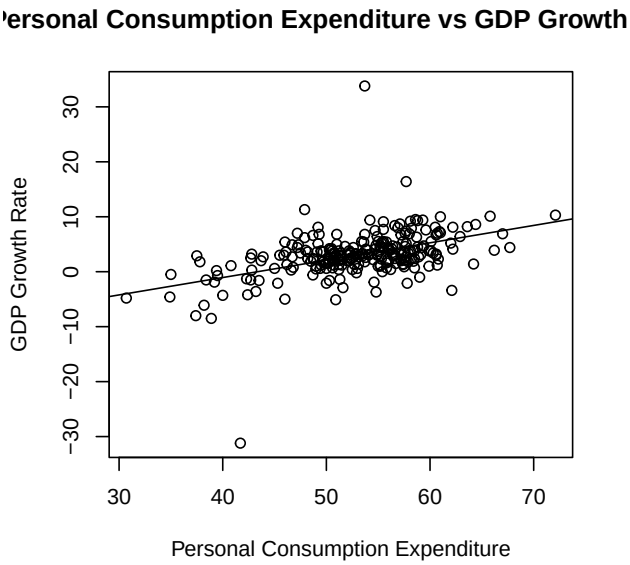
Transfromed Data

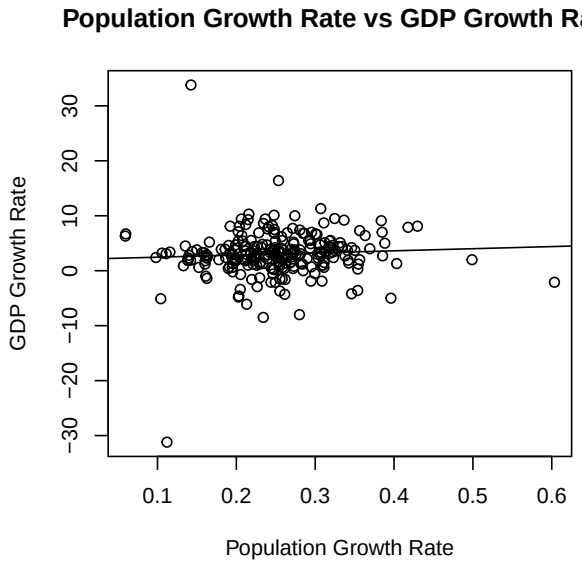


The plot shows distributions of variables after performing transformations. It now looks like the skewness has been dealt with and the variables follow an almost normal distribution.

Variables that have a linear relationship with GDP Growth Rate







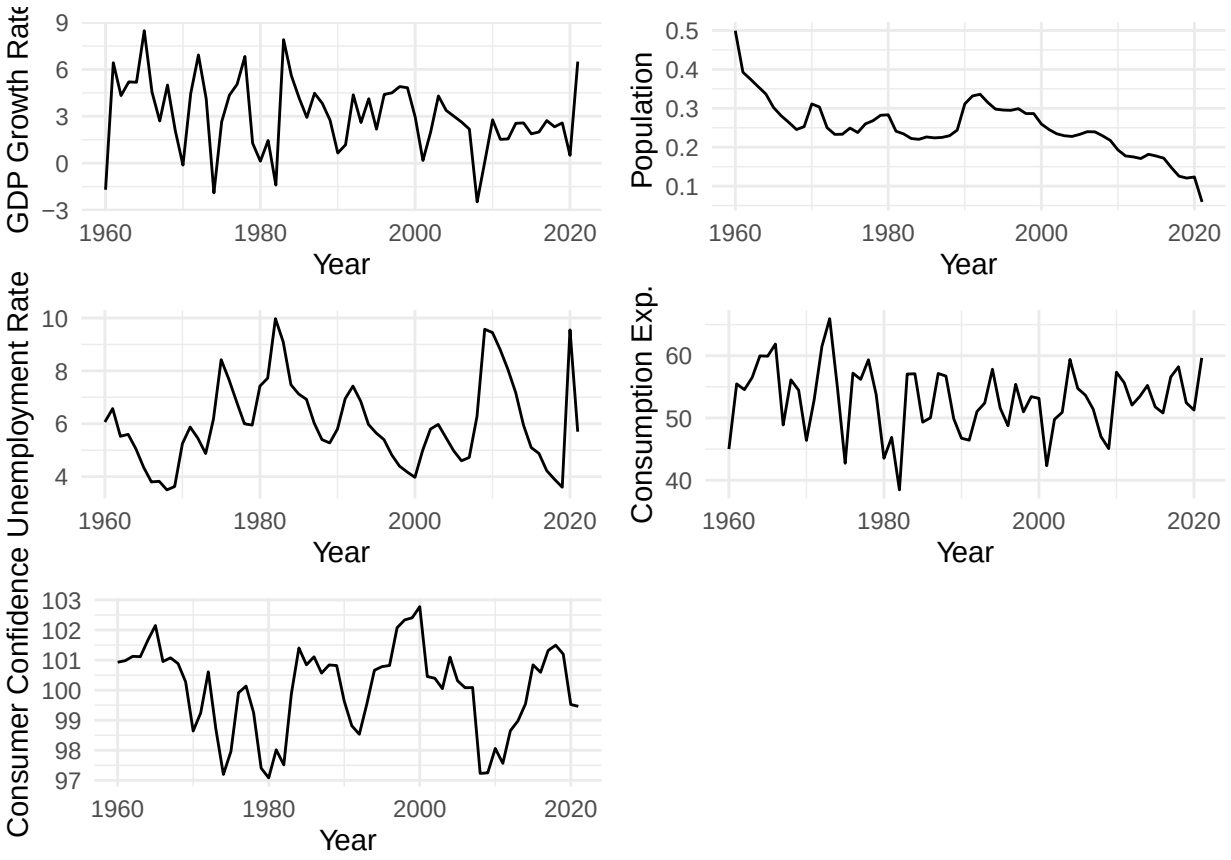
Insights from the plot:

1. There appears to be a linear relationship between the variables unemployment rate, Personal consumption expenditure and Population Growth Rate.

2. There appears to be negative linear relationship between Unemployment rate and GDP Growth rate. We can see, as the unemployment rate goes up, the GDP growth rate drops.

3. As the value of the goods and services purchased by the US residents increases, so does the GDP growth rate.

Looking at the data with respect to time



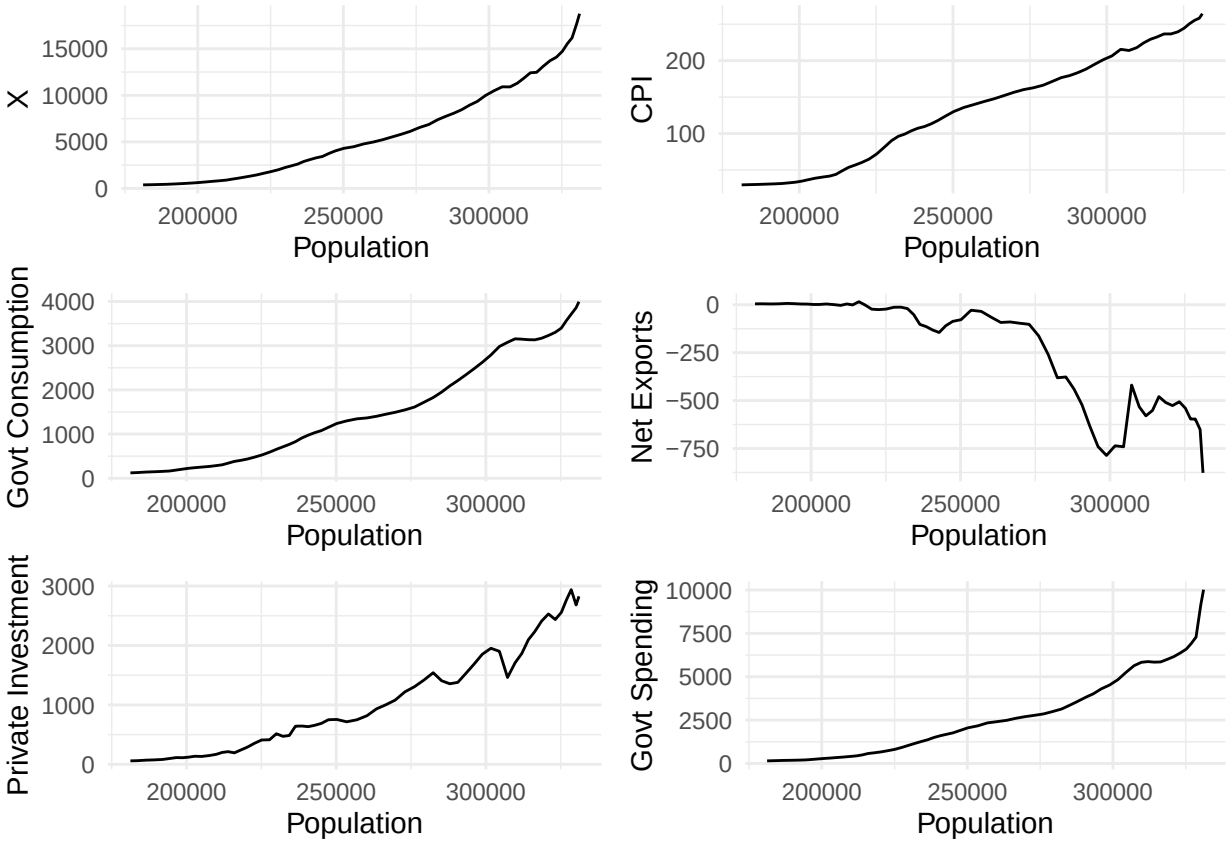
Insights:

1. By looking at the plots for unemployment rate and GDP Growth rate against Year we can draw some insights. During the year 2020 when Covid-19 pandemic hit, the unemployment rate soared which resulted in GDP growth rate to drop for the same time period.

2. It also can be seen that during the Great Recession which lasted from December 2007 to June 2019, a similar pattern can be seen. The unemployment rate went high which contributed to a plunge in GDP Growth rate.

3. Population growth rate is on a decline as years pass.

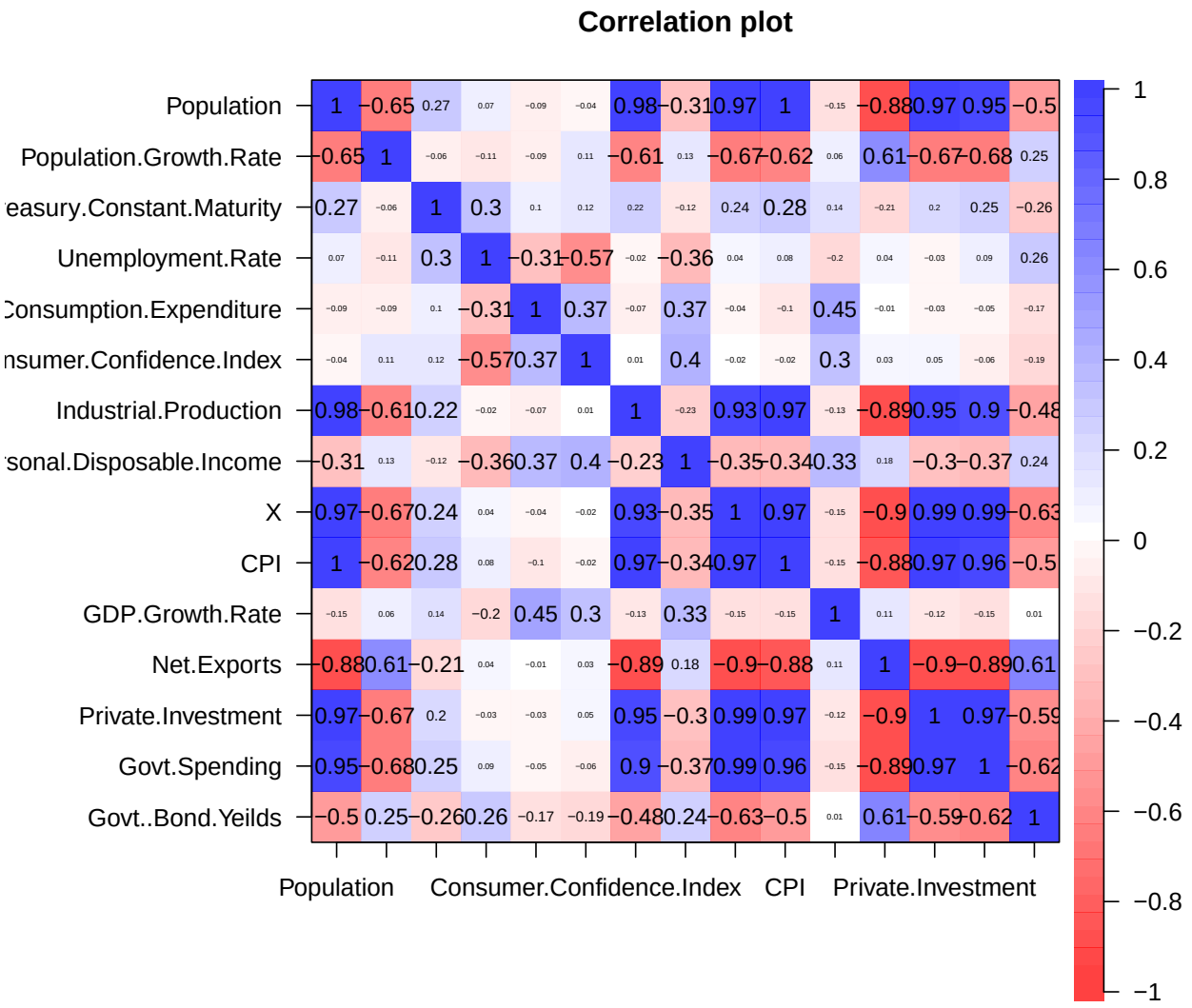
Variables that have a linear relationship with GDP Growth Rate



This plot shows variables that appear to have a linear relationship with population.

Insights:
It can be seen that personal consumption, X, Private Investment, CPI and government spending have increased with increase in population whereas net exports dropped. It is understood that when the population rises, it means there can be increase in need for imports too.

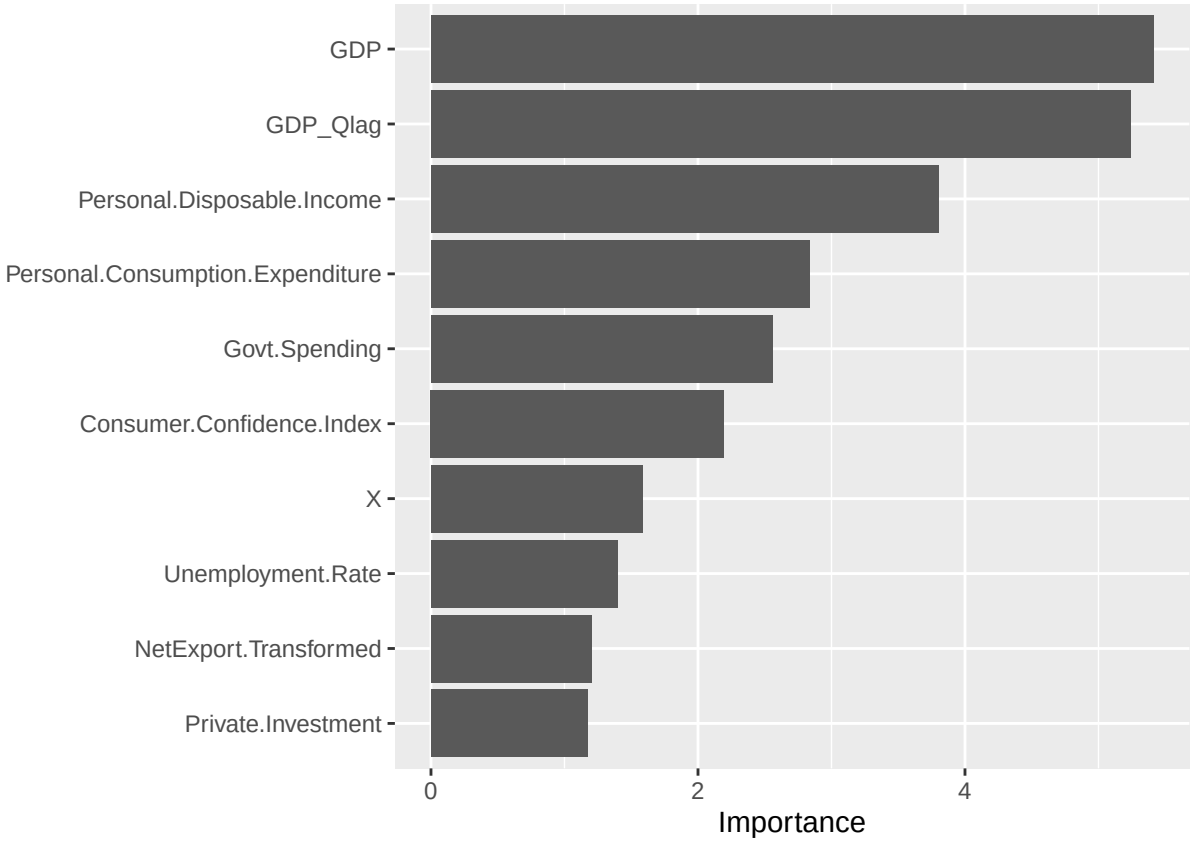
Correlation Plot



By looking at the correlation plot, there appears to be multicollinearity in the data. Variables are highly correlated with each other.

Insights:
1. Population has high correlation over 0.95 with variables - Industrial Production, Housing Expenditure, CPI, Government Spending and Private Investment. This seems to be obvious as population has a big role to play for these factors.
2. Industrial Production is highly correlated with variables - Housing Expenditure, CPI, Govt Spending and Private Investment
3. Similarly, other variables have high correlation among them.
Multicollinearity can be dealt with using PCA and removing some variables that have very high correlation with others.

Importance of variables in model



The above plot shows the relative importance of variables if all predictors were to be included in the model.

Building regression model with 1 quarter Lag

Observations	183
Dependent variable	gdp_train$GDP.Growth.Rate
Type	OLS linear regression

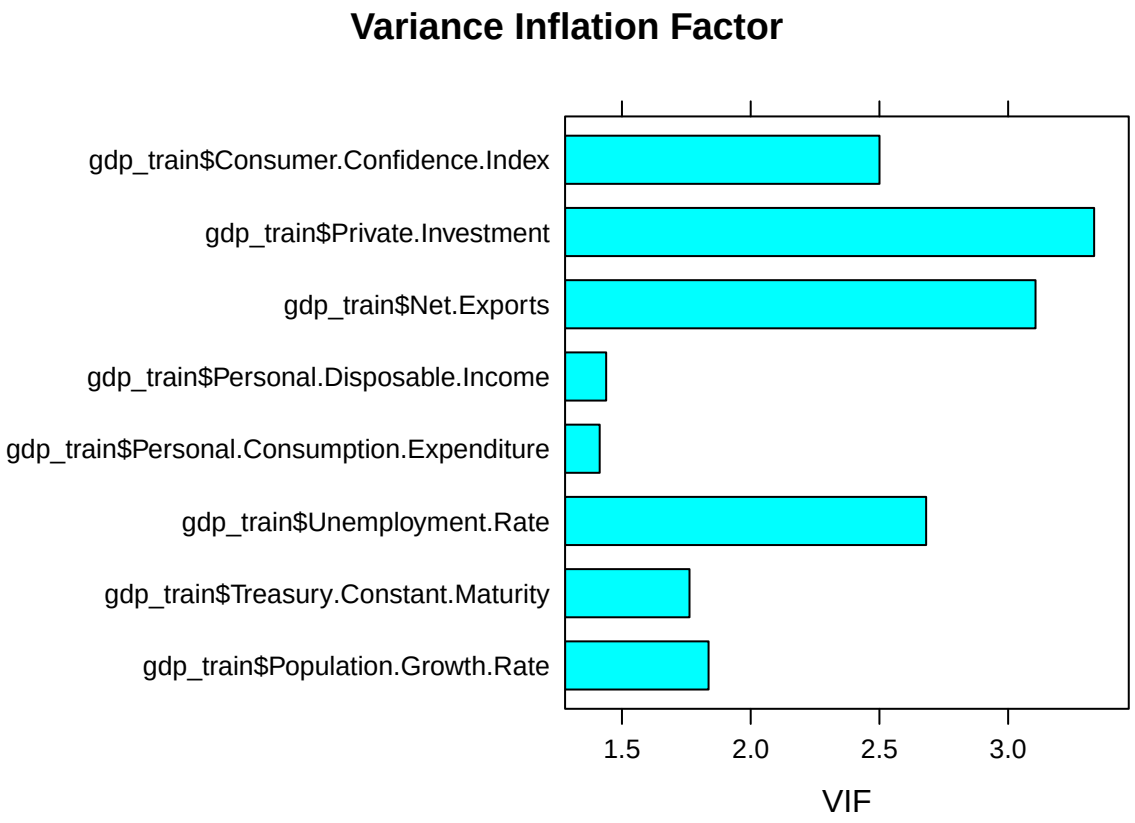
F(8,174)	8.19
R^2	0.27
Adj. R^2	0.24

Insights from the summary of model:

1. The model explains 66% of the variability in the data
2. The small p-value of model indicates the model is statistically significant.
3. The significance of the selected variables is not very high.
4. The residulas look to be normally distributed

	Est.	S.E.	t val.	p
(Intercept)	-56.11	65.60	-0.86	0.39
gdp_train$Population.Growth.Rate	0.86	5.92	0.15	0.88
gdp_train$Treasury.Constant.Maturity	0.51	0.25	2.04	0.04
gdp_train$Unemployment.Rate	0.23	4.47	0.05	0.96
gdp_train$Personal.Consumption.Expenditure	0.26	0.06	4.43	0.00
gdp_train$Personal.Disposable.Income	0.00	0.00	2.54	0.01
gdp_train$Net.Exports	0.00	0.00	0.44	0.66
gdp_train$Private.Investment	0.09	1.13	0.08	0.93
gdp_train$Consumer.Confidence.Index	4.12	6.46	0.64	0.52

Standard errors: OLS



The above plot shows the variance inflation factor of variables in the model. It shows the measure of how much the behavior of an independent variable is influenced by its interaction with other variables. Consumer Confidence Index has the highest VIF.

Building regression model with 2 quarters Lag

Observations	183
Dependent variable	gdp_train$GDP_Quarter_lag
Type	OLS linear regression

Insights from the summary of model:

F(8,174)	5.79
R^2	0.21
Adj. R^2	0.17

	Est.	S.E.	t val.	p
(Intercept)	-184.81	56.67	-3.26	0.00
gdp_train$Population.Growth.Rate	-18.43	5.11	-3.61	0.00
gdp_train$Treasury.Constant.Maturity	0.31	0.22	1.44	0.15
gdp_train$Unemployment.Rate	10.57	3.86	2.74	0.01
gdp_train$Personal.Consumption.Expenditure	-0.03	0.05	-0.59	0.56
gdp_train$Consumer.Confidence.Index	19.09	5.58	3.42	0.00
gdp_train$Personal.Disposable.Income	0.00	0.00	1.70	0.09
gdp_train$Net.Exports	0.00	0.00	0.04	0.97
gdp_train$Private.Investment	-2.50	0.97	-2.57	0.01

Standard errors: OLS

1. The model explains only 28% of the variability in the data
2. The p-value of model indicates the model is not very statistically significant.
3. The relatively large p-values indicate that none of the variables have high significance.
4. The residuals look to be normally distributed

Building regression model with 4 quarters lag

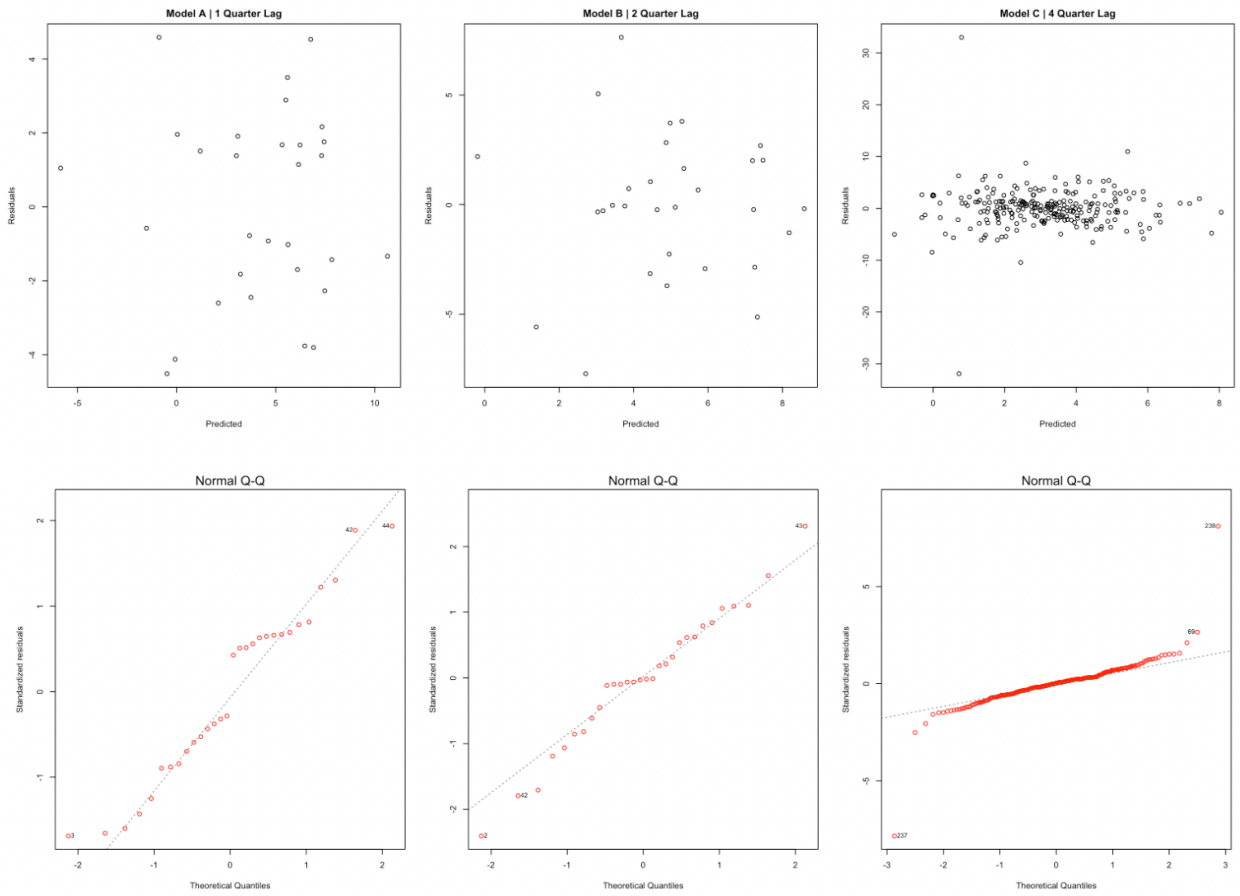
Observations	190 (55 missing obs. deleted)
Dependent variable	gdp_year_lag$GDP_Lag_Year
Type	OLS linear regression

F(7,182)	6.79
R^2	0.21
Adj. R^2	0.18

	Est.	S.E.	t val.	p
(Intercept)	5.37	4.47	1.20	0.23
gdp_year_lag$Personal.Consumption.Expenditure	-0.05	0.04	-1.08	0.28
gdp_year_lag$Government.Consumption.Expenditure.and.Gross.Investment	-0.00	0.00	-2.36	0.02
gdp_year_lag$Real	-0.00	0.00	-0.92	0.36
gdp_year_lag$Govt..Bond.Yeilds	-0.48	0.17	-2.76	0.01
gdp_year_lag$t	0.05	0.03	1.67	0.10
gdp_year_lag$Treasury.Constant.Maturity	0.68	0.71	0.95	0.34
gdp_year_lag$Unemployment.Rate	8.23	2.98	2.76	0.01

Standard errors: OLS

Residuals



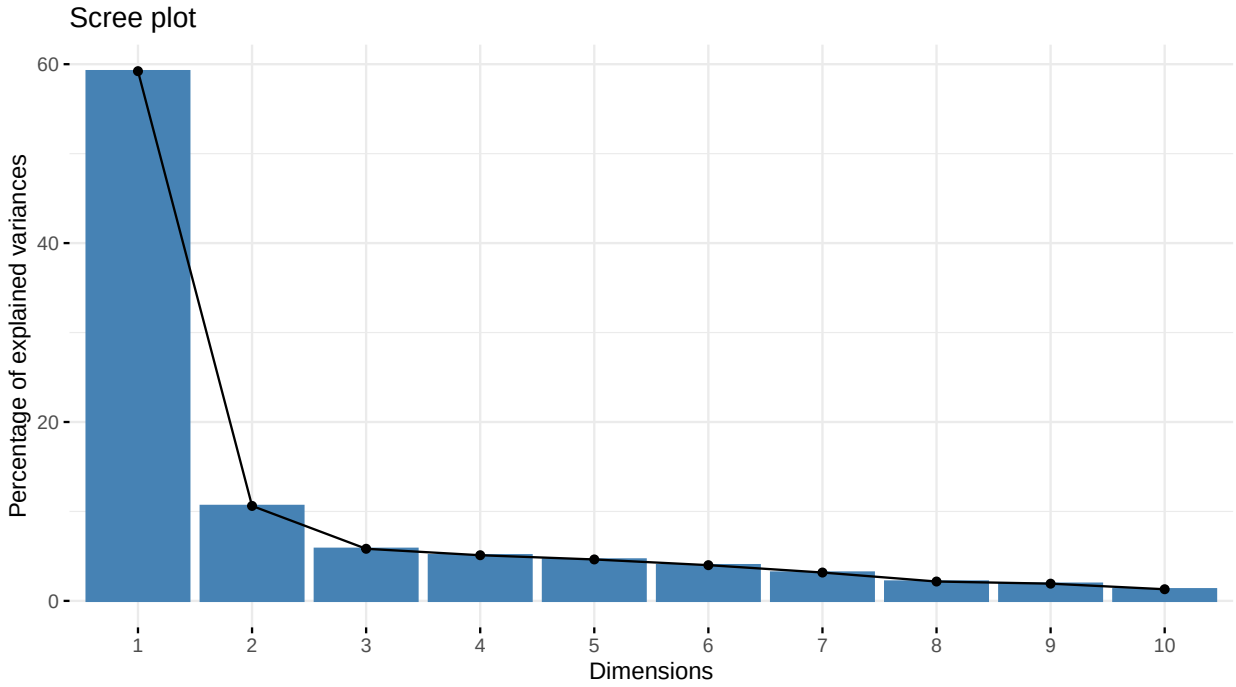
Model A: The residuals do not show any significant relationship and are scattered around zero which indicate that model A is performing well.

Model B: The residuals do not show any significant relationship and are scattered around zero which indicate that the regression model is performing well.

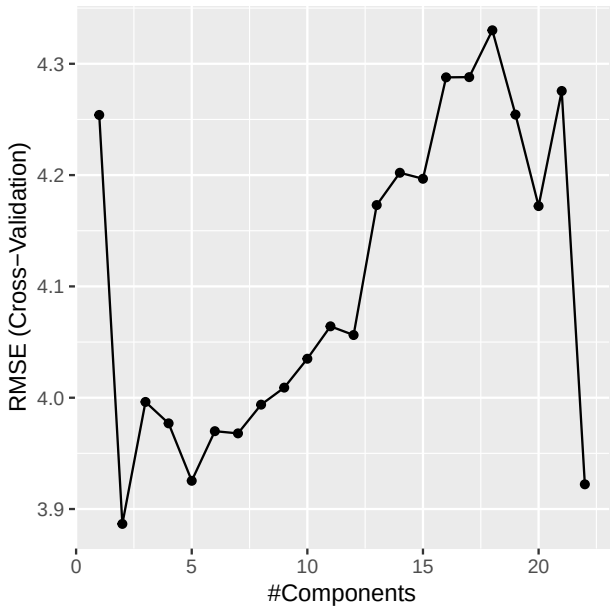
Model C: The residuals show high correlation between residuals and predicted values, one of the reasons for that might be slight skew in the data.

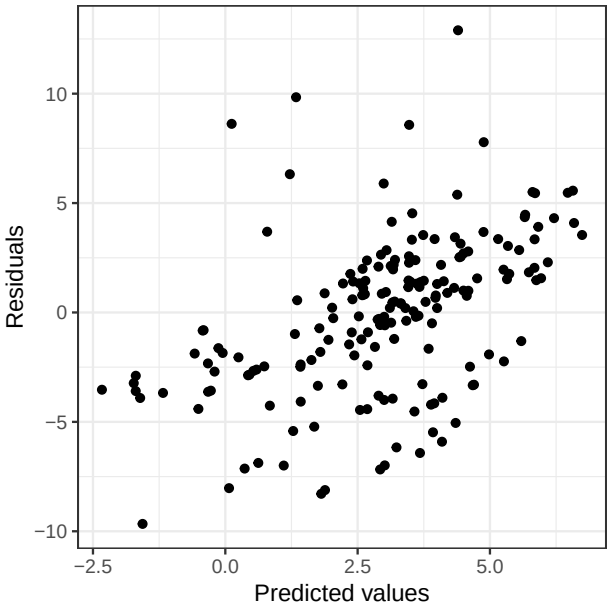
The first two QQ-Plots for Model A & B show almost normally distributed residuals whereas the residuals aren't normal for Model C. Hence, Model A & B prove to be better than Model C.
It can be observed that the residuals tend to show a relationship as the lag intervals increase.

Using PCA



It can be seen from the screeplot that 2 principle components capture about 70% of variance in the data.





It can be observed from the scree plot that the about 2 principal components can capture maximum varianace without making the regression computationally heavy.
Also there appears to be a slight linear relationship between predicted values and residuals.

Data Appendix

This data is taken from Federal Reserve Economic Data [FRED]. T The variables were in seperate excel files which was combined into a single file. All cleaning was done in Excel

Variables:
1. Year: Years from 1960 to 2020
2. Quarter: Quarters 1,2,3,4
3. Population - Population of USA from 1960 to 2020
4. Ppulation Growth Rate - Population quarterly growth rate
5. Treasury Constant Maturity - Value of US treasury
6. Unemployment Rate - Quarterly unemployment rate of USA labor work force
7. Personal Consumption Expenditure - Value of the goods and services purchased by, or on the behalf of, U.S residents.
8. Consumer Confidence Index - Measure of how optimistic consumers are regarding their expected financial situation
9. Industrial Production - Output of industrial sector of US
10. Personal Disposable Income - Amount of money left with US residents after paying off taxes
11. CPI - Consumer Price Index
12. GDP Growth Rate - USA GDP quarterly growth rate 1960 to 2020
13. Net Exports - Balance of trade | Export minus Import
14. Private Investment - Investments by private organisations and Individuals in US
15. Government Spending - Government consumption and expenditure
16. Government Bond Yields - Interest rate at which the government borrows

Transformations:
Log transformation has been applied to the following variables:
Unemployment rate
CPI
Private Investments

Square root transformation was applied to the following variables:
Treasury Constant Maturity
Consumer Confidence Index
Net Exports
Housing Expenditure
Government Spending

Cube root:
Treasury Constant Maturity
Net Exports

Code Appendix

```
library("emulator")
library("corrplot")
library("psych")
library("regclass")
library("car")
library("rsample")
library("ggplot2")
library(gridExtra)
library(tidyverse)
library(vip)
library(cowplot)
library(caret)
library(visdat)
library(naniar)
library(leaps)
library(factoextra)

#Importing Data
gdpdata <- read.csv("Backup2GDP_Dataset.csv")
gdp_1 <- read.csv("Backup2GDP_Dataset.csv")
gdp_year_lag <- read.csv("TestGDP.csv")

cor_data <- subset(gdpdata,select = c("Population","Population.Growth.Rate","Treasury.Constant.Maturity

#-------------------------------------------------------------------------------------------------------
#Data Exploration

vis_miss(cor_data)

pairs.panels(cor_data) # Before Transformations

#Plots before transformations
par(mfrow=c(3,3))

pp1 <- hist(gdp_1$Unemployment.Rate, xlab  = "Unemployment Rate", main = "Unemployment Rate")
pp2 <- hist(gdp_1$X, xlab = "Housing Expenditure", main = "Housing Expenditure")
pp3 <- hist(gdp_1$CPI, xlab = "CPI", main = "CPI")
pp4 <- hist(gdp_1$Govt.Spending, xlab = "Govt. Spending", main = "Govt. Spending")
pp5 <- hist(gdp_1$Treasury.Constant.Maturity, xlab = "Treasury Constant Maturity", main = "Treasury Cons
pp6 <- hist(gdp_1$Consumer.Confidence.Index, xlab = "Consumer Confidence Index", main = "Consumer Confi
pp7 <- hist(gdp_1$Net.Exports,, xlab = "Net Exports", main = "Net Exports")

par(mfrow=c(3,3))

x1 <- log10(gdp_1$Unemployment.Rate)
x2 <- sqrt(scale(gdp_1$X))
x3 <- log10(gdp_1$CPI)
x4 <- log10(gdp_1$Private.Investment)
x5 <- sqrt(scale(gdp_1$Govt.Spending))
x6 <- (gdp_1$Treasury.Constant.Maturity^(1/3))
x7 <- log10(gdp_1$Consumer.Confidence.Index)
```

```r
x8 <- (gdp_1$Net.Exports^(1/3))

pt1 <- hist(x1, xlab  = "Unemployment Rate", main = "Unemployment Rate")
pt2 <- hist(x2, xlab = "Housing Expenditure", main = "Housing Expenditure")
pt3 <- hist(x3, xlab = "CPI", main = "CPI")
pt4 <- hist(x5, xlab = "Govt. Spending", main = "Govt. Spending")
pt5 <- hist(x6, xlab = "Treasury Constant Maturity", main = "Treasury Constant Maturity")
pt6 <- hist(x7, xlab = "Consumer Confidence Index", main = "Consumer Confidence Index")
pt7 <- hist(x8, xlab = "Net Exports", main = "Net Exports")

par(mfrow=c(2,2))

#Unemployment vs GDP Growth Rate
r2 <- lm(GDP.Growth.Rate~Unemployment.Rate, gdpdata)
p2  <- plot(gdpdata$GDP.Growth.Rate~gdpdata$Unemployment.Rate, xlab = "Unemployement Rate", ylab = "GDP
abline(r2)

#Personal Consumption Expenditure vs GDP Growth Rate
r3 <- lm(GDP.Growth.Rate~Personal.Consumption.Expenditure, gdpdata)
p3 <- plot(gdpdata$GDP.Growth.Rate~gdpdata$Personal.Consumption.Expenditure, xlab = "Personal Consumptio
abline(r3)

r4 <- lm(gdpdata$GDP.Growth.Rate~gdpdata$Population.Growth.Rate, gdpdata)
p4 <- plot(gdpdata$GDP.Growth.Rate~gdpdata$Population.Growth.Rate, xlab = "Population Growth Rate", ylab
abline(r4)

# Annualizing variables
gdp_annual <-
  gdp_1%>%
  group_by(Year)%>%
  summarise(
    GDP.Growth.Rate = mean(GDP.Growth.Rate),
    Unemployment.Rate = mean(Unemployment.Rate),
    Personal.Consumption.Expenditure = mean(Personal.Consumption.Expenditure),
    Consumer.Confidence.Index = mean(Consumer.Confidence.Index),
    X=mean(X),
    CPI=mean(CPI),
    Government.Consumption.Expenditure.and.Gross.Investment=mean(Government.Consumption.Expenditure.and
    Net.Exports=mean(Net.Exports),
    Private.Investment=mean(Private.Investment),
    Govt.Spending=mean(Govt.Spending),
    Population = mean(Population),
    Population.Growth.Rate = mean(Population.Growth.Rate)
  )
#-------------------------------------------------------------------------------------------------------

# Creating new dataframe for exploration with respect to Year
population_df <- gdp_annual[c(1,6:13)]

#EDA by year

plot1 <- ggplot(gdp_annual, aes(x=Year, y=gdp_annual$GDP.Growth.Rate)) +
  geom_line()+theme(axis.text.x = element_text(angle = 70)) +
```

```
  labs(x = "Year", y = "GDP Growth Rate")+
  theme_minimal()

plot2 <-ggplot(gdp_annual, aes(x=Year, y=gdp_annual$Population.Growth.Rate)) +
  geom_line()+
  labs(x = "Year", y = "Population ")+
  theme_minimal()

plot3 <-ggplot(gdp_annual, aes(x=Year, y=Unemployment.Rate)) +
  geom_line()+
  labs(x = "Year", y = "Unemployment Rate")+
  theme_minimal()

plot4 <-ggplot(gdp_annual, aes(x=Year, y=gdp_annual$Personal.Consumption.Expenditure)) +
  geom_line() +
  labs(x = "Year", y = "Consumption Exp.")+
  theme_minimal()

plot5 <-ggplot(gdp_annual, aes(x=Year, y=gdp_annual$Consumer.Confidence.Index)) +
  geom_line() +
  labs(x = "Year", y = "Consumer Confidence")+
  theme_minimal()

grid.arrange(plot1, plot2,plot3, plot4,plot5, ncol=2)

#Plots for Variables that have a linear relationship with Population

Plot7 <- ggplot(gdp_annual, aes(x=gdp_annual$Population, y=gdp_annual$X)) +
  geom_line() +
  labs(x = "Population", y = "X") +
  theme_minimal()

Plot8 <- ggplot(gdp_annual, aes(x=gdp_annual$Population, y=gdp_annual$CPI)) +
  geom_line() +
  labs(x = "Population", y = "CPI")+
  theme_minimal()

Plot10 <- ggplot(gdp_annual, aes(x=gdp_annual$Population, y=gdp_annual$Government.Consumption.Expenditu
  geom_line() +
  labs(x = "Population", y = "Govt Consumption")+
  theme_minimal()

Plot11 <- ggplot(gdp_annual, aes(x=gdp_annual$Population, y=gdp_annual$Net.Exports)) +
  geom_line() +
  labs(x = "Population", y = "Net Exports")+
  theme_minimal()

Plot12 <- ggplot(gdp_annual, aes(x=gdp_annual$Population, y=gdp_annual$Private.Investment)) +
  geom_line() +
  labs(x = "Population", y = "Private Investment")+
  theme_minimal()

Plot13 <- ggplot(gdp_annual, aes(x=gdp_annual$Population, y=gdp_annual$Govt.Spending)) +
```

```
  geom_line() +
  labs(x = "Population", y = "Govt Spending")+
  theme_minimal()


grid.arrange(Plot7, Plot8, Plot10,Plot11, Plot12, Plot13, ncol=2)

corPlot(cor_data)

#-----------------------------------------------------------------------------------------------------------

#Transforming variables

gdpdata$Unemployment.Rate <- log10(gdpdata$Unemployment.Rate)
gdpdata$X <- log10(gdpdata$X)
gdpdata$CPI <- log10(gdpdata$CPI)
gdpdata$Private.Investment <- log10(gdpdata$Private.Investment)
gdpdata$Govt.Spending <- log10(gdpdata$Govt.Spending)
#gdpdata$Treasury.Constant.Maturity <- sqrt(gdpdata$Treasury.Constant.Maturity)
gdpdata$Consumer.Confidence.Index <- sqrt(gdpdata$Consumer.Confidence.Index)
#gdpdata$Net.Exports <- sqrt(gdpdata$Net.Exports)

#-----------------------------------------------------------------------------------------------------------

#Scaling the Data
gdp_scaled <- (read.csv("2GDP_Dataset.csv"))
gdp_scaled <- as.data.frame(gdp_scaled)

#-----------------------------------------------------------------------------------------------------------

#Splitting
set.seed(007)
split <- initial_split(gdpdata, prop = 0.75)
gdp_train <- training(split)
gdp_test <- testing(split)

# First Model

lmxx <- lm(GDP.Growth.Rate ~ . , gdp_train)
vip(lmxx)

#Final Model
lm_f <- lm(gdp_train$GDP.Growth.Rate ~  gdp_train$Population.Growth.Rate + gdp_train$Treasury.Constant.
summary(lm_f)

#VIF
vif_vals <- VIF(lm_f)
barchart(vif_vals, main = "Variance Inflation Factor", xlab = "VIF")

#Lag Interval - 2 Quarters

lm_quarter_lag2 <- lm(gdp_train$GDP_Quarter_lag ~  gdp_train$Population.Growth.Rate + gdp_train$Treasur
summary(lm_quarter_lag2)
```

```r
#Year Lag


gdp_year_lag$Unemployment.Rate <- log10(gdp_year_lag$Unemployment.Rate)
gdp_year_lag$X <- log10(gdp_year_lag$X)
gdp_year_lag$CPI <- log10(gdp_year_lag$CPI)
gdp_year_lag$Private.Investment <- log10(gdp_year_lag$Private.Investment)
gdp_year_lag$Govt.Spending <- log10(gdp_year_lag$Govt.Spending)
gdp_year_lag$Treasury.Constant.Maturity <- sqrt(gdp_year_lag$Treasury.Constant.Maturity)
gdp_year_lag$Consumer.Confidence.Index <- sqrt(gdp_year_lag$Consumer.Confidence.Index)
gdp_year_lag$Net.Exports <- sqrt(gdp_year_lag$Net.Exports)

lmylag <- lm(gdp_year_lag$GDP_Lag_Year ~ gdp_year_lag$Personal.Consumption.Expenditure + gdp_year_lag$G
summary(lmylag)

#-------------------------------------------------------------------------------------------------------

#Plotting Residuals

par(mfrow=c(1,3))
pred <- predict(lm_f, data = gdp_test)   # Save the predicted values
res <- residuals(lm_f) # Save the residual values

res_plot1 <- plot(pred, res, xlab = "Predicted", ylab = "Residuals", main = "Model A | 1 Quarter Lag")
abline(pred~res)

pred2 <- predict(lm_quarter_lag2, data = gdp_test)
res2 <- residuals(lm_quarter_lag2, data = gdp_test)

res_plot2 <- plot(pred2, res2, xlab = "Predicted", ylab = "Residuals", main = "Model B | 2 Quarter Lag"


pred3 <- predict(lmylag, data = gdp_test)
res3 <- residuals(lmylag, data = gdp_test)

res_plot3 <- plot(pred3, res3, xlab = "Predicted", ylab = "Residuals", main = "Model C | 4 Quarter Lag"

qq1 <- plot(lm_f, which=2, col=c("red"))
qq2 <- plot(lm_quarter_lag2, which=2,col=c("red"))
qq3 <- plot(lmylag, which=2,col=c("red"))

#PCA

my_pca <- gdpdata %>% prcomp()

scaledpca <- gdpdata %>% prcomp(scale = T, center = T)

fviz_eig(scaledpca) #Screeplot

pca_df <- gdpdata

set.seed(787)
split_pca <- initial_split(pca_df, prop = .8)
```

```
gdp_train_pca <- training(split_pca)
gdp_test_pca <- testing(split_pca)


set.seed(786)
cv_model_pcr <- train(
  GDP.Growth.Rate ~ .,
  data = gdp_train_pca,
  method = 'pcr',
  trControl = trainControl(method = 'cv', number = 10),
  preProcess = c('zv', 'center', 'scale'),
  tuneLength = 50,
  na.rm = T
)

summary(cv_model_pcr)

ggplot(cv_model_pcr, cex.main=2)

predicted<-predict(
  cv_model_pcr,
  gdp_train,
  ncomp = 1:10,
  comps,
  type = c("raw"),
  na.action = na.pass
)


plot20 <- ggplot() +
  geom_point(aes(predicted, predicted-gdp_test$GDP.Growth.Rate)) +
  xlab('Predicted values') +
  ylab('Residuals') +
  theme_bw()
```