

Coursework

Applied Statistics and Data Visualisation

(Principles of Data Science)

MSc Data Science



University of
Salford
MANCHESTER

ANALYZING WEALTH BASED ON TAX, GDP, EXPENDITURE, AND INFLATION

ODEWALE OLUWASEUN ELIZABETH

SDP060

16/12/2022

TABLE OF CONTENTS

INTERACTIVE DASHBOARD DESIGN	3
1. INTRODUCTION	3
2. BACKGROUND RESEARCH	3
3. EXPLORATION OF THE DATASET	4
4. INVESTIGATION OF DATA WORKFLOWS & PROPOSAL FOR DESIGN OF DASHBOARD	5
5. DISCUSSION	12
6. CONCLUSIONS	12
STATISTICAL ANALYSIS.....	13
1. INTRODUCTION	13
2. BACKGROUND RESEARCH	13
3. EXPLORATION OF THE DATASET	14
4. ANALYSIS.....	17
5. DISCUSSIONS	50
6. CONCLUSIONS	50
REFERENCES AND APPENDICES	51

1. INTRODUCTION

The goal of this assessment or research is to compare the wealth of some selected developed countries regarding Taxes, GDP, and major Expenditures. I believe that you cannot measure an economy's wealth without considering utilities or consumptions hence the need for expenditure. This will enlighten us on the performance of these countries and their capacity. Developed countries are generally known to be performing better than others because of their ability to manage their economy industry-wise, education-wise, importation, and exportation-wise, health-wise amongst others, and Taxes are believed to be the major strength for generating revenue.

Taxes are used to help fund public works and services and build as well as maintain the infrastructure used in a country. A government usually taxes its individual and corporate residents and they are used for the betterment of the economy and all who are living in it (Gorton, Updated November 30, 2022).

Expenditure on the other end is the necessary or important goods and services that the government incurs from the revenue generated from taxes and other sources to make the economy better. This public spending enables the government to produce and purchase goods and services to fulfill their objectives like the provision of public goods or the redistribution of resources (Roser & 2016).

All of these different taxes and expenditures are captured in the GDP of the country which is then used to measure the wealth of the nation.

The visualization below reveals the evolution of government expenditure as a share of national income, for a selection of the 10 countries we are using in this assessment over the last century.

Figure 1.1: Government spending, 1880 to 2011



Note (Mauro, 2015)

2. BACKGROUND RESEARCH

Dashboard designs are now popular in the 21st century which allows people to make use of themselves as against using graphs and metrics that are unexplainable or requiring the expertise of certain professionals to analyze; it encourages self-service analytics. Technology has reformed the world and is always evolving daily which also affects the way we use it to collect and analyze data. Today, a BI dashboard refers to an analytics tool that provides

a consolidated view of relevant business data, such as key performance metrics and operational data which allows business owners to quickly and easily view, explore, and breakdown data from different sources, at a glance and on one screen (Team, 2022).

Dashboards have been very useful in decision-making since the 1970s and this has exposed different domains of businesses to embrace this platform. Although it still requires the expertise of some business analysts to analyze and interpret the data as well as make predictions that will boost the income level of the business. Before now, data collected are just for research analysis but now it is more complex and becoming a key factor in decision-making.

These days, most dashboard platforms allow you to get data from different sources and have various information before finally taking a decision based on the data analysis.

PowerBI which is the main tool for this visualization was originally designed by Ron George and was named Project Crescent in 2010 summer and was released in 2011, July 11th precisely (Wikipedia, 2022b).

This tool was designed to read datasets directly from a database, structured files such as excel spreadsheets, CSV, JSON, XML, or webpage. It is a collection of different software services, connectors, and apps that cooperate or function together and then convert data sources that are not related producing interactive views or insights.

A single-screen dashboard explains all about an imported dataset that has been transformed with a tool in PowerBI called Power Query so that it can be visualized in the chosen format and capture the attention of its audience. There are several visualization tools embedded in PowerBI similar to another BI dashboard known as Tableau which is used in telling stories. They include bar charts (stacked or clustered), column charts (stacked or clustered), line charts, maps, slicer, cards, scatter plots, and many more which have their peculiarity to further explain the dataset.

3. EXPLORATION OF THE DATASET

For this visualization, the dataset used was collected from the World Data Bank database which ranges from 2011 to 2020 for 10 countries. The countries extracted are Finland, France, Greece, Germany, the United Kingdom, the United States, Turkiye, Russian Federation, Canada, and Belgium (Group, 2022). Based on the objective of this assessment, the following indicators were carefully selected;

Table 1.0: World Bank Indicators and its definition

S/N	Indicator Name	Definition / Meaning
1.	Taxes on income, profits, and capital gains (current LCU)	These are charges on the main or net earnings of people/individuals, the interest of corporate organizations including sole proprietors or entities, on principal gains irrespective of achieved status, and also on valuable properties. Local government dues are exempted for coalition purposes.
2.	Tax revenue (% of GDP)	These are necessary payments made to core government wallets for general use with the exemption of transfers like retribution charges or other penalties and mostly public welfare allowances.
3.	Taxes on goods and services (% of revenue)	These are majorly taxes on trades and yields especially on some important commodities with tariffs, levies on valuables, duty payments on abstraction and manufacture of crude products, and benefits of budgetary patents or holdings.
4.	GDP (current US\$)	This is a measure of all the values obtained via the internally made or manufactured goods and services in the country including any levied product without any appropriation. These calculations are made without considering any devaluation of some essential resources in current U.S. dollars.

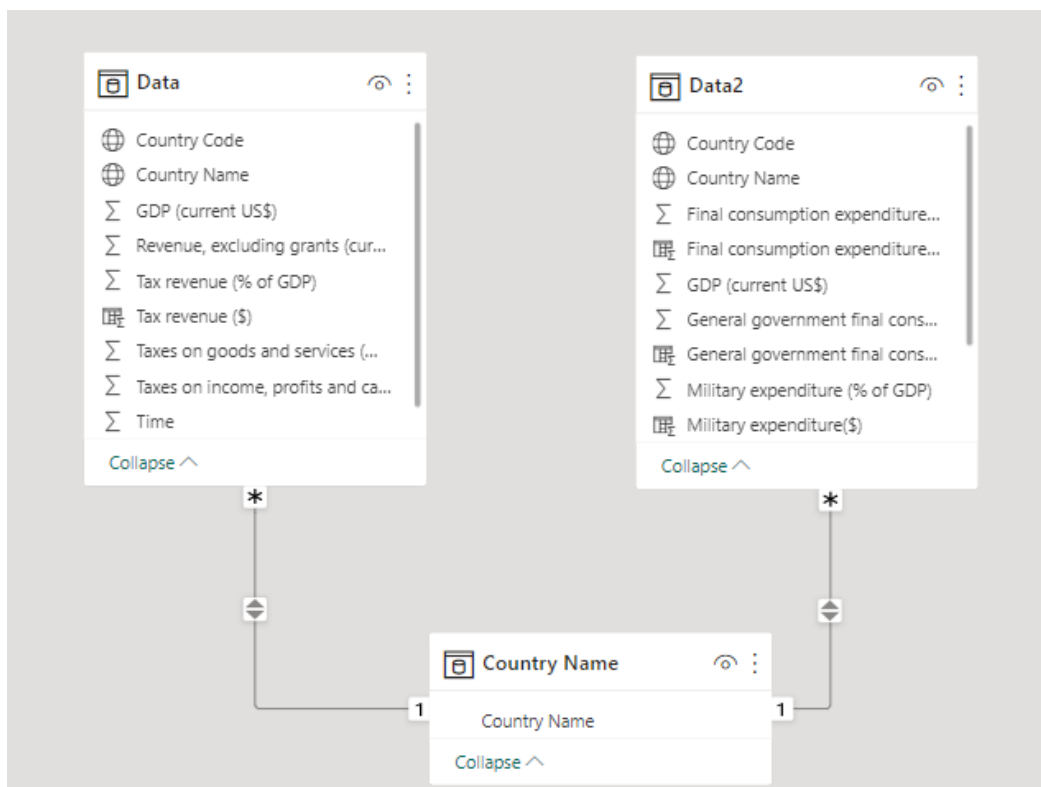
5.	Revenue, excluding grants (current LCU)	These are receivables from levies, general fees derived from sales, and commodities without including grants.
6.	Final consumption expenditure (% of GDP)	This encapsulates all private and government utilization in the economy previously known as total consumption.
7.	General government final consumption expenditure (% of GDP)	Similar to final consumption expenditure, this includes the addition of procurement of the economy's necessities and also workers' remuneration. All expenses made on national defense and security are also captured with exception of military expenditures which are within the team of government.
8.	Military expenditure (% of GDP)	These are all armed forces capital expenditures including all entitlements of military and civil personnel both previous and current.
9.	Research and development expenditure (% of GDP)	These captured all recent expenditures within the four active or important sectors of the government.

The data pre-processing was majorly replacing some missing values with zeros, renaming the columns, changing some columns to whole numbers, and two decimal places, and data type to currency to properly visualize the dataset. To create a relationship, the dataset was downloaded twice; one for income/revenue by tax and the other for expenditure, and were combined having country as their connection. New columns were also created by using the DAX expression to generate some actual figures for some indicators.

4. INVESTIGATION OF DATA WORKFLOWS & PROPOSAL FOR DESIGN OF DASHBOARD

After transforming the dataset on Power Query and loading it to the PowerBI desktop, the tool recognized a relationship between the two datasets and the country data set which was extracted from the main dataset.

Figure 1.2: Autodetected relationships within the dataset



According to the objective of this assessment which is to compare the wealth strength of the selected countries with emphasis on Taxes and Expenditure, the first visualization tool selected and loaded to the canvas to explain the data were the clustered bar chart revealing the Tax revenue(\$) and Taxes on income, profits and capital gains by country.

The Tax revenue(\$) was obtained by using a DAX expression;

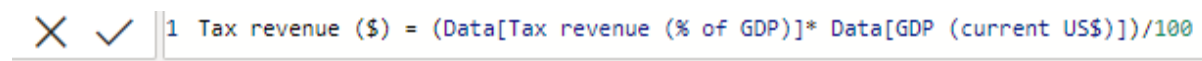
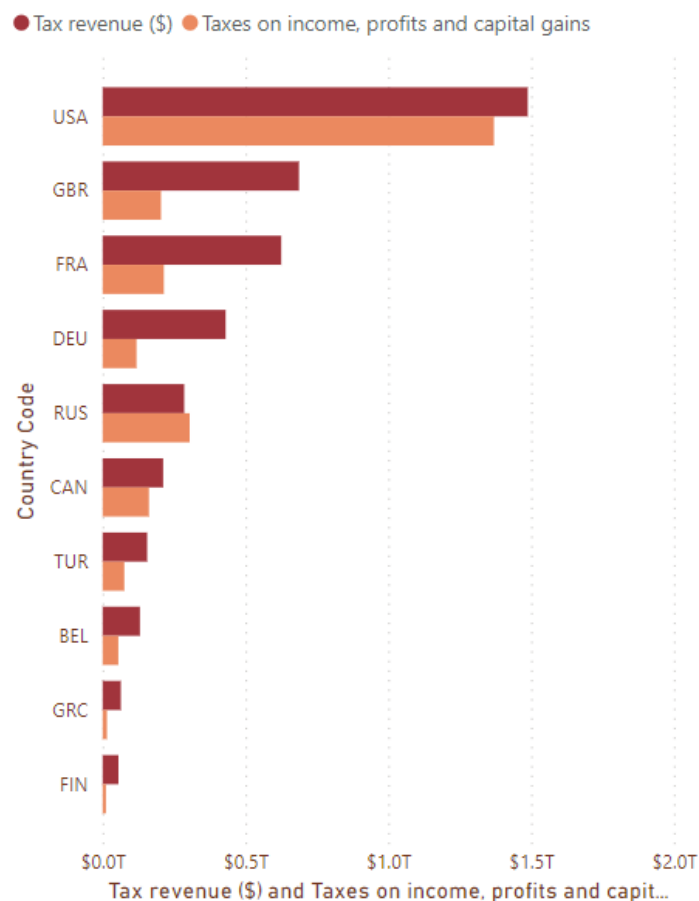


Figure 1.3: Taxes by Country(2011)

Tax revenue (\$) and Taxes on income, profits and capital gains

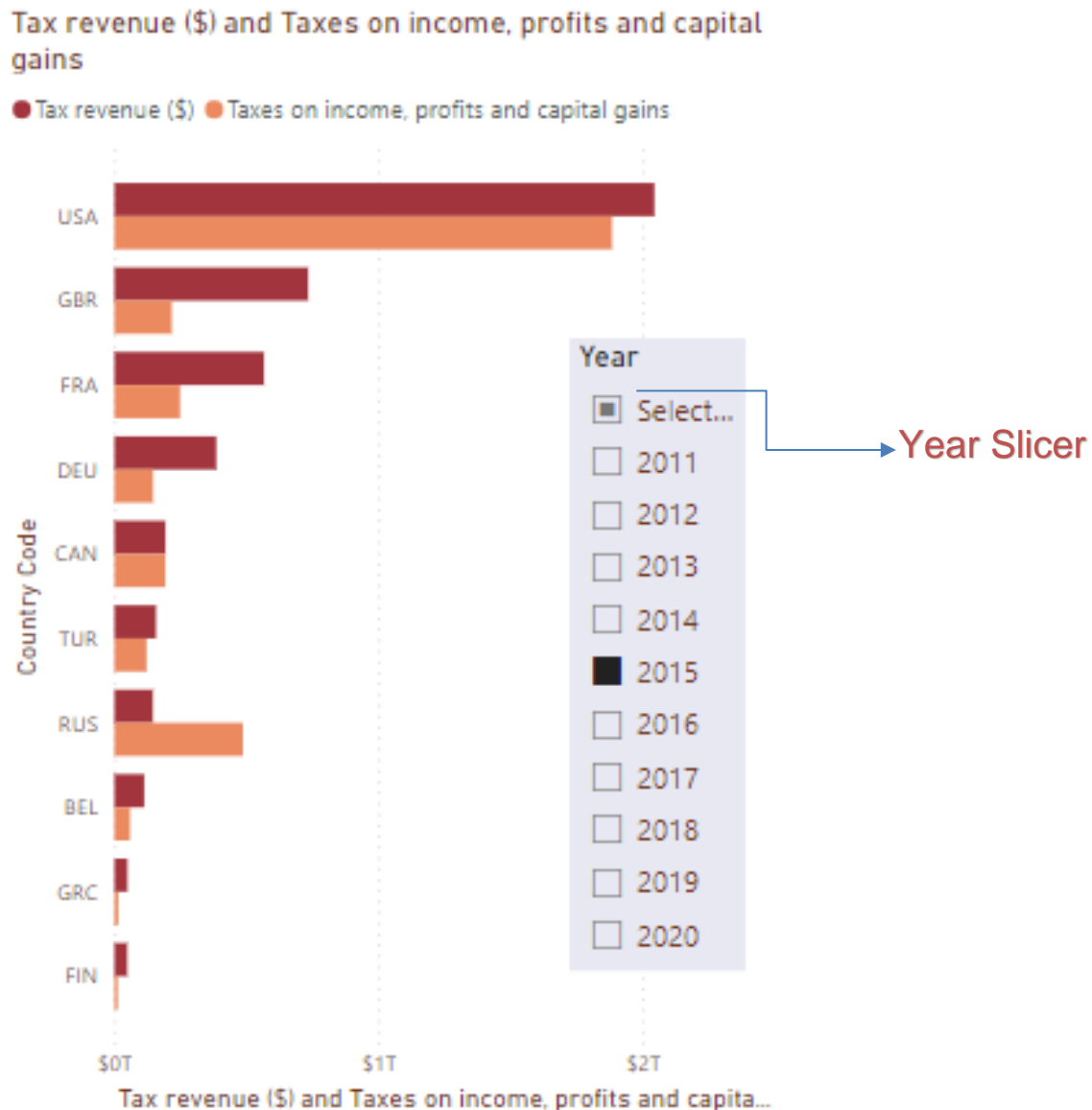


This visualization reveals that the USA has the highest and outstanding income generated from both Tax revenue and Taxes on income, profits, and capital gains compared to other countries. Next to the USA is the United Kingdom which is almost on the same level as France and then Germany, then Russia (having an equal proportion of Tax revenue and Taxes on income, profits, and capital gains), then Canada, Turkey, Belgium, Greece and finally Finland which is the last on the list. We can immediately identify countries that are doing better based on this simple visualization.

Then we created Year slicer to visualize each country's performance every year from 2011 to 2020.

Fig.1.2 visualized the performance for 2011 and each country maintained its position in ascending order till 2015, which revealed that Russia's income on Taxes on income, profits, and capital gains was more than the Tax revenue.

Figure 1.4: Taxes by Country(2015)



In 2016, there was a decline in Taxes on income, profits, and capital gains in Russia as against the performance in 2015 and 2017. What went wrong? Figure 1.5 revealed that in 2016, Russia had Military Expenditure of \$69,267,579,502 against \$66,422,181,076 in 2015, which is a rise in expenditure amongst other expenses that are not captured in this visualization but later reduced in 2017. Fig.1.6 revenue line chart for Russia revealed just a small increase in 2016 from what was achieved in 2015 and then it kicked up again in 2018 but with a forecast that there will be a steady increase in revenue for the next five(5) years.

Figure 1.5: Taxes by Country(2016)

Tax revenue (\$) and Taxes on income, profits and capital gains

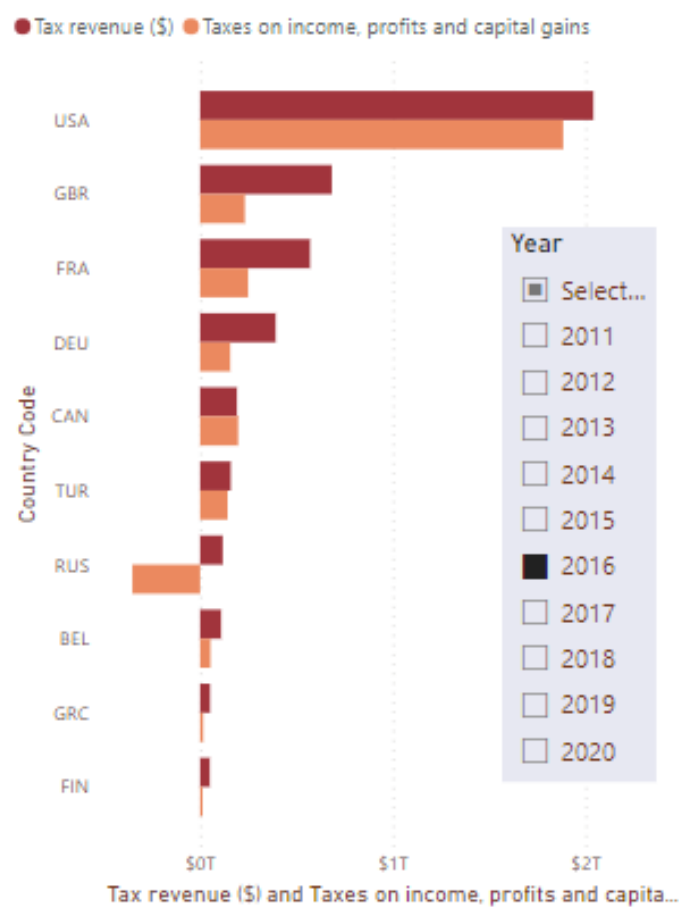
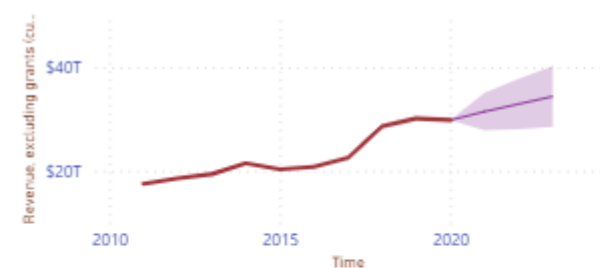


Figure 1.6: Revenue Forecast and Military Expenditure(Russia)

Forecast of Revenue, excluding grants for 5 Years



\$12.41T

Military expenditure for 10 Years



Country

Belgium	Greece
Canada	Russian Federation
Finland	Turkiye
France	United Kingdom
Germany	United States

We noticed that in 2017, Finland overtook Greece with a small difference of about \$1b, then we realized that Greece had a shortfall in their Revenue and increased Military expenditure. The card visualization for other expenditures also captured the total sum of expenditure incurred by Greece as against the previous year which made the country last on the list till 2020, see figures 1.7 to 1.9.

Figure 1.7: Taxes by Country(2017)

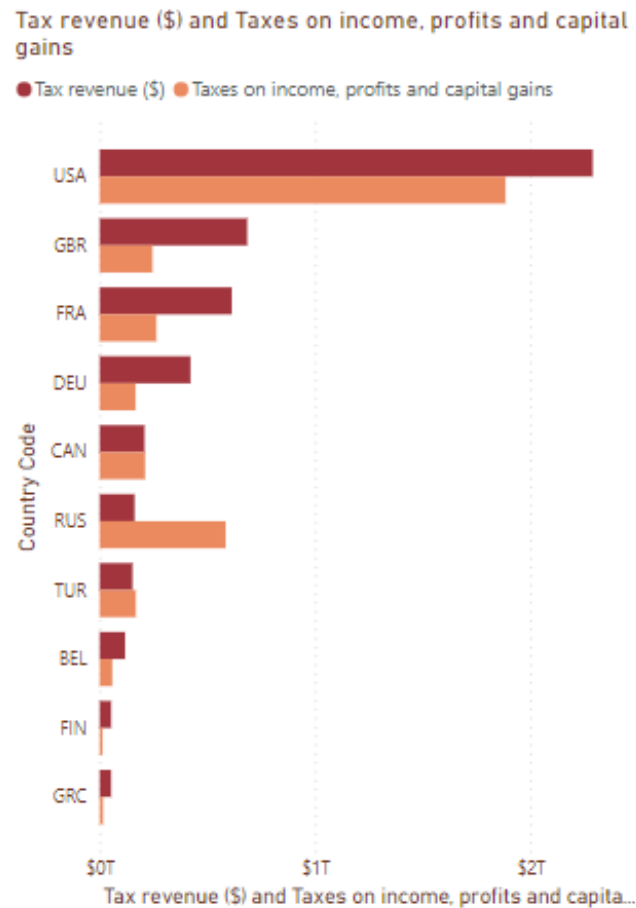


Figure 1.8: Different Expenditure in value(Greece)

General Government Final
Consumption Expenditure

\$461.6bn

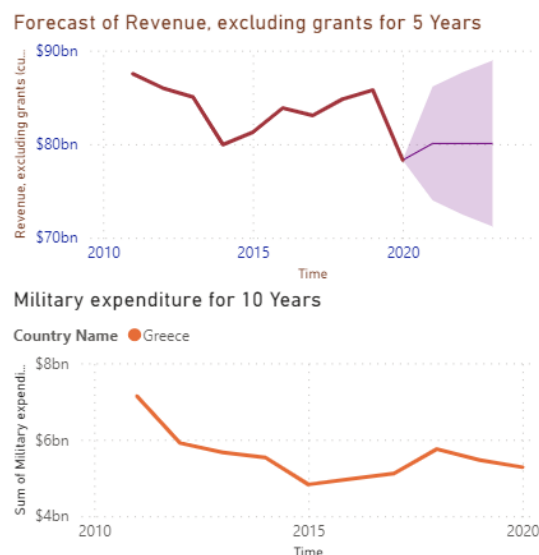
Total Research and Development
Expenditure

\$21.73bn

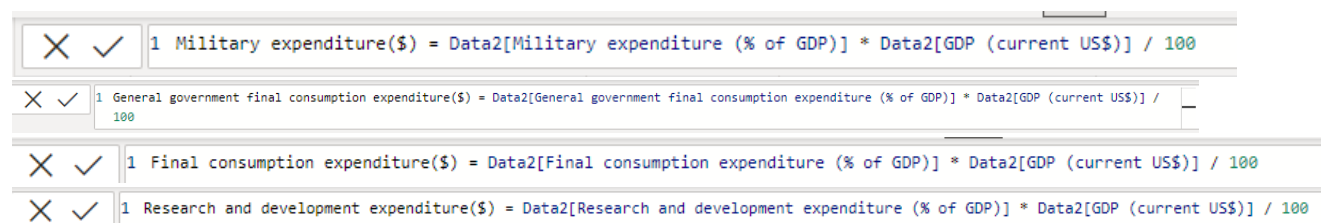
Final Consumption Expenditure

\$1.98T

Figure 1.9: Revenue Forecast and Military Expenditure(Greece)



To generate actual figures for the expenditures created with the card visualization on the canvas, we had to create new columns for each expenditure using the DAX expression as well.



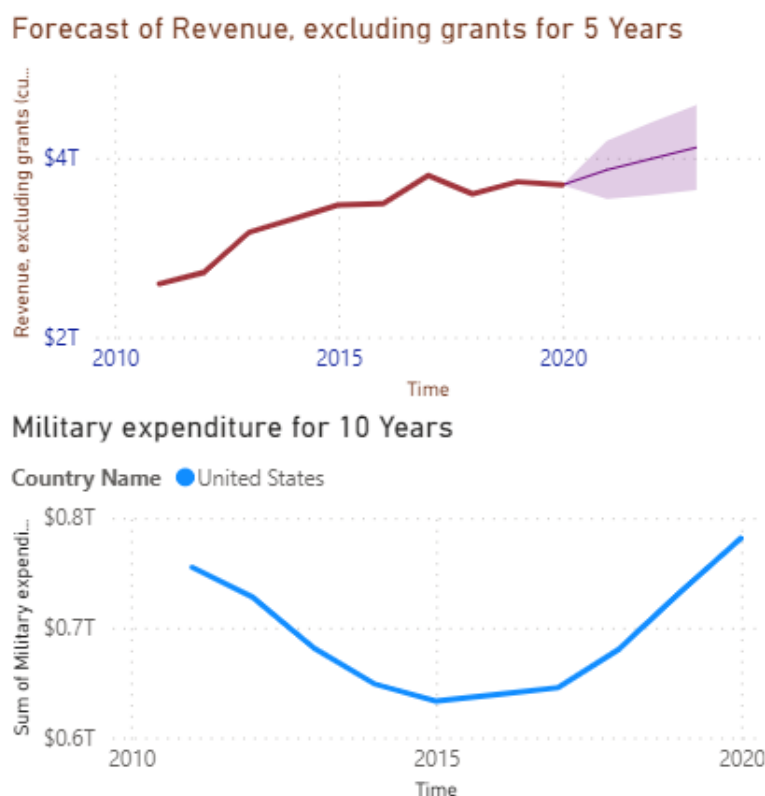
All these interactions were easy because we decided to edit the slicer interactions; the year slicer only affects the clustered bar chart while the country slicer affects the expenditure cards, the revenue forecast, and the military expenditure line charts.

The map tool was used to visualize the GDP of each country in different regions. The USA has the highest in North America and even among all countries and then some other European countries. See Fig.1.10 below;

Figure 1.10: GDP by Country

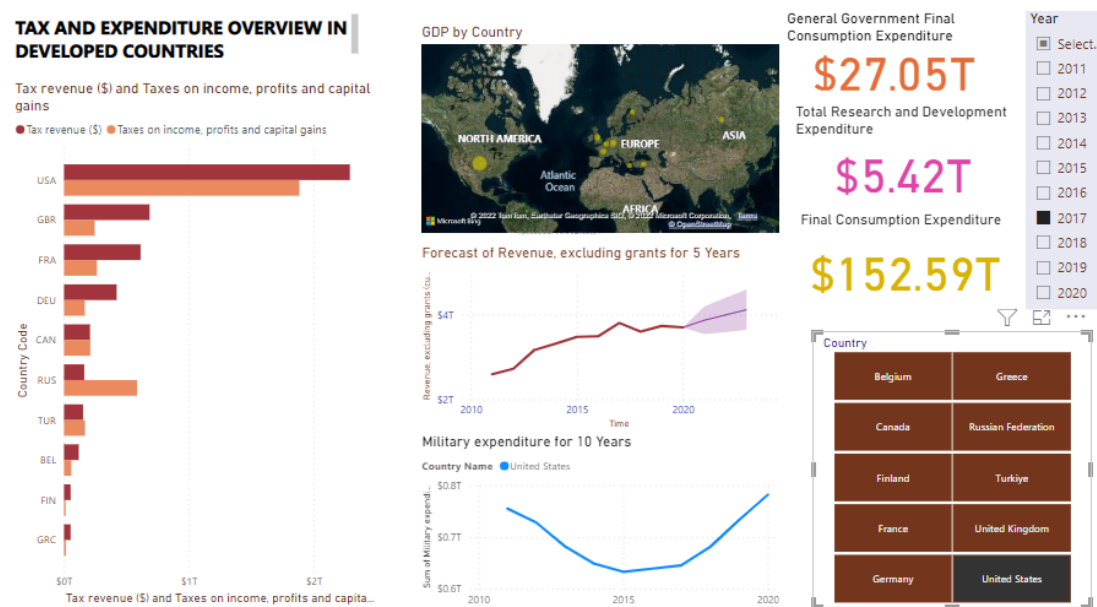


Figure 1.11: Revenue Forecast and Military Expenditure(USA)



In as much as the USA has the highest income compared to other countries, it also has the highest expenditure. Fig. 1.11 reveals that the USA has high military expenditure in 2011 but there was a drop in 2015 then started increasing at an increasing rate from 2017 to 2020. The US economy is one of the largest in the world and this visualization also confirmed this.

Figure 1.12: Single Screen Dashboard of Tax and Expenditure of Selected Developed Countries



5. DISCUSSION

In a nutshell, we have been able to identify and analyze the taxes generated by each country and the expenses incurred for a period of 10 years and we realized that the results gotten depend on their population. Although this is not visualized on the dashboard screen, we know that the US is highly populated hence, their manpower and manufacturing capacity is broad compared to other countries. Another factor deduced from this dashboard points to the USA as the largest showing that it is an English-speaking country and people migrate there legally or illegally helping the economy to grow.

This research will therefore propose to other countries to intensify their efforts to make their country attain a level like the USA in terms of building more in the economy. The more the expenditure on relevant needs that will boost the economy, the more it generates revenue from products and services.

6. CONCLUSIONS

Aside from taxes and expenditure, other relevant aspects of the economy can also be focused on like the health sector, education sector, agricultural sector, manufacturing, and construction sector to ease the residents of each country. Most people complain because of the high increase in taxes on most products and services so the country should channel its energy on other factors that can generate revenue for the economy. Countries like Finland, Belgium, and Greece which are the least performing on the list should also build tourist attraction centers that will increase the country's population. United Kingdom, France, and Germany should also strive to keep building the economy and also divert their focus on other revenue-generating aspects which will keep the economy in a healthy state.

Finally, dashboards can help you get insight, but to find the answers to the unexpected questions they raise, you must be able to look beyond the dashboards' most granular inquiries (Wexler, 2022).

STATISTICAL ANALYSIS

1. INTRODUCTION

The objective of this analysis is to identify key factors that determine the GDP of a country with a particular interest in Inflation. The rate at which prices increase over a specific period is known as inflation. Inflation is often measured in broad terms, such as the general rise in prices or the rise in a nation's cost of living.

By removing government-set prices and the more volatile costs of goods like food and energy that are most impacted by seasonal variables or transient supply problems, core consumer inflation concentrates on the underlying and persistent trends in inflation. Policymakers also keep a careful eye on core inflation. An index with greater coverage, like the GDP deflator, is needed to calculate the overall inflation rate for a nation, rather than just for consumers, for example (Oner, 2022). So in the real world before using any models to check if GDP and Inflation are correlated, both indicators affect each other or are determinant factors.

The market worth of all the final goods and services produced and sold (not resold) in a certain period by countries is measured in dollars using the term "gross domestic product" (GDP). This measure is frequently amended before being regarded as a trustworthy indication due to its subjective and complex character. The per capita GDP of a region is determined by the GDP to total population ratio (also called the Mean Standard of Living) but we will be using the GDP(current US\$) in this analysis(Wikipedia, 2022a).

2. BACKGROUND RESEARCH

For this statistical analysis, we will be using different models to analyze the dataset. We will check the correlation of the indicators to one another, and use regression models, time series, and hypothesis testing to achieve our objectives.

Correlation and linear regression are the two most frequently used methods for examining the relationship between two quantitative variables. Regression expresses the relationship as an equation, whereas correlation assesses the strength of the linear link between two variables. In a scatter plot, the linear relationship between two variables is greater the closer the points are to a straight line. We may compute the correlation coefficient to put a number on the strength of the link. In algebraic notation, the correlation coefficient is represented by the following equation if two variables, x and y, are present and the data are presented as n pairs, such as [x1, y1], [x2, y2], [x3, y3],...[xn,yn].

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

where \bar{x} is the mean of the x values, and \bar{y} is the mean of the y values.(Bewick V, 2003)

The equation of this line which is the linear relationship can be discovered via regression. The regression line is the term most often used to describe this line. The response variable is always displayed on the vertical (y) axis in a scatter diagram.

$Y = a + bx$, where the coefficients a and b represent the line's intercept on the y-axis and the gradient, respectively, is the equation for a straight line (Bewick V, 2003).

A particular method of examining a set of data points gathered over a period of time is called a "time series analysis." Instead of just capturing the data points intermittently or arbitrarily, time series analyzers record the data points at regular intervals over a predetermined length of time. But this kind of study involves more than just gathering data over time (Group, 2003-2022).

A survey or experiment's results can be tested using the statistical procedure known as hypothesis testing to see whether the results are significant.

$$z = \frac{\hat{p} - p}{\sqrt{pq/n}}$$

By calculating the likelihood that your results are the result of chance, you are essentially verifying their validity. If it's possible that your results were a coincidence, the experiment won't be repeatable and won't be of much help. One of the most difficult concepts for students to understand is hypothesis testing, mainly because you have to know your null hypothesis before you can even start the test. You may frequently find it challenging to understand those challenging word problems. But it's simpler than you might imagine. All you have to do is:

- Determine your null hypothesis,
- Describe your null hypothesis,
- Select the type of test you need to run—either one that supports or rejects the null hypothesis.

All these statistical tools will be used to analyze the objective of this assessment and reach a conclusion.

3. EXPLORATION OF THE DATASET

The dataset was extracted from the World Data Bank database which ranges from 2011 to 2020 for 10 countries. The following countries were carefully selected based on their development level; Brazil, Canada, China, Australia, Finland, Greece, Germany, India, Mexico, the United Kingdom, the United States, Nigeria, Spain, Turkiye, and Russian Federation (Group, 2022). See below table revealing all the downloaded variables or indicators for this analysis;

Table 2.0: World Bank Indicators and its definition

S/N	Indicator Name	Time/Year	Data Source	Definition / Meaning
1.	Inflation, consumer prices (annual %)	2011 - 2020	World Data Bank	The annual percentage change in the average consumer's cost of acquiring a basket of goods and services, which may be set or modified at predetermined intervals, such as annually, is reflected in inflation as measured by the consumer price index.
2.	GDP (current LCU)	2011 - 2020	World Data Bank	The gross value contributed by all resident producers in the economy, along with any applicable product taxes and any subsidies not reflected in the value of the goods themselves, is what is referred to as the GDP at purchaser's prices. It is estimated without taking into account natural resource deterioration or the depreciation of manufactured assets.
3.	GDP growth (annual %)	2011 - 2020	World Data Bank	GDP growth rate, expressed as a percentage, based on market pricing and constant local currency. The prices used to calculate aggregates are constant 2015 values, given in US dollars.
4.	Unemployment, total (% of total labor)	2011 - 2020	World Data Bank	The percentage of the labour force that is neither employed nor actively seeking one is referred to as unemployment.

	force) (modeled ILO estimate)			
5.	Wage and salaried workers, total (% of total employment)	2011 - 2020	World Bank	Data Workers who hold jobs classified as "paid employment jobs," where the incumbents hold explicit (written or oral) or implicit employment contracts that provide them with a basic salary that is not directly based on the unit's revenue, are referred to as wage and salaried workers (also known as employees).
6.	Self-employed, total (% of total employment)	2011 - 2020	World Bank	Data Employees who hold the kind of positions classified as "self-employment jobs" on their own, with one or more partners, or in a cooperative are considered self-employed workers. i.e., occupations where pay directly reflects profits from the items and services produced.
7.	Imports of goods and services (current LCU)	2011 - 2020	World Bank	Data All products and other market services that are brought into the country from other countries are represented by imports. Included in them are the costs associated with goods, shipping, insurance, travel, royalties, licence fees, and other services like government, financial, informational, communication, construction, and building-related services. They do not include transfer payments, investment income, or what was once referred to as factor services.
8.	Insurance and financial services (% of commercial service imports)	2011 - 2020	World Bank	Data Freight insurance on imported goods, other direct insurance like life insurance, financial intermediation services like commissions, foreign exchange trades, and brokerage services, and ancillary services like financial market operational and regulatory services are all included in insurance and financial services.
9.	Exports of goods and services (current US\$)	2011 - 2020	World Bank	Data The value of all products and other market services that are exported to other countries is represented by their exports of goods and services. The value of goods, freight, insurance, transport, travel, royalties, licence fees, and other services, such as government, financial, informational, communication, construction, and commercial services, are among them. They do not include employee remuneration, investment income or transfer payments.

Note(Group, 2022)

The data pre-processing was carried out on R-studio; after setting the working directory, the CSV file was read into the R-script.

The library(readr) was used to load and change the dataset to the required class in place of using the inbuilt library as.numeric()

```
library(readr)
WDI <- read_csv("world_Development_Indicator.csv",
  col_types = cols(Time = col_character(),
    `Wage and salaried workers, total (% of total employment) (modeled ILO estimate) [SL.EMP.WORK.ZS]` = col_number(),
    `Self-employed, total (% of total employment) (modeled ILO estimate) [SL.EMP.SELF.ZS]` = col_number(),
    `Imports of goods and services (current LCU) [NE.IMP.GNFS.CN]` = col_number(),
    `Insurance and financial services (% of commercial service exports) [TX.VAL.INSF.ZS.WT]` = col_number(),
    `Exports of goods and services (current US$) [NE.EXP.GNFS.CD]` = col_number()))
```

Rows displaying the citation of the dataset were also removed with the below syntax(Webster, 2015);

```
###Need to eliminate unwanted rows with NA, which are rows 151 to 155
WDI <- WDI[-c(151:155),]
```

The column names were changed by using the %in% function with concatenation.

```
###View column names and rename
colnames(WDI)
colnames(WDI)[colnames(WDI) %in%
  c("Country Name","Country Code","Time","Time Code",
    "Inflation, consumer prices (annual %) [FP.CPI.TOTL.ZG]",
    "GDP (current LCU) [NY.GDP.MKTP.CN]","GDP growth (annual %) [NY.GDP.MKTP.KD.ZG]",
    "Unemployment, total (% of total labor force) (modeled ILO estimate) [SL.UEM.TOTL.ZS]",
    "Wage and salaried workers, total (% of total employment) (modeled ILO estimate) [SL.EMP.WORK.ZS]",
    "Self-employed, total (% of total employment) (modeled ILO estimate) [SL.EMP.SELF.ZS]",
    "Imports of goods and services (current LCU) [NE.IMP.GNFS.CN]",
    "Insurance and financial services (% of commercial service exports) [TX.VAL.INSF.ZS.WT]",
    "Exports of goods and services (current US$) [NE.EXP.GNFS.CD]")] <-
  c("Country","C-Code","Year","Y-Code","Inflation","GDP", "GDP growth","Unemployment","Wage and salaried workers",
    "Self-employed","Imports","Insurance and financial services","Export")
```

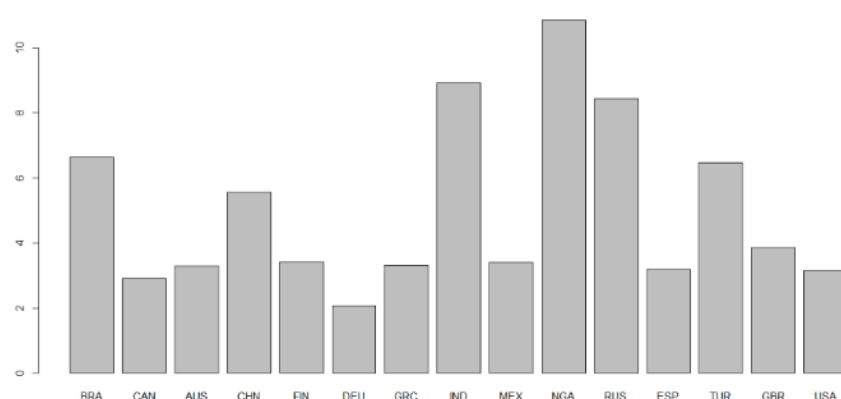
Initially, the number of countries selected from the World Bank Data online were 15 but since we are only working on 10, we decided to choose a year (2011) and plot a graph based on Inflation to select the 10 countries for this analysis.

See the below code and barplot;

```
WDI_2011 <- WDI[WDI$Year==2011,]
WDI_2011

###Plotting WDI_2011 using names.arg for Inflation to select 10 countries###
barplot(WDI_2011$Inflation,names.arg = WDI_2011$`C-Code`)
```

Figure 2.1: Histogram Plot of Inflation by Country(2011)



Based on the plot, we decided to remove Germany, Canada, Spain, Australia, and Mexico from the dataset and we were left with 100 observations of 13 variables.

We also checked for missing values which were replaced with their Mean and detected outliers with boxplot. All these were further explained in the descriptive analysis section of this assessment.

4. ANALYSIS

4.1a. Descriptive Analysis

The dataset was explored and assessed using different techniques and methods. By using the `summary()` function, it describes the dataset by displaying the class and mode of character vectors and also the minimum value, 1st quartile, median, mean, 3rd quartile, maximum value, and missing values of numeric vectors. Other ways of viewing the dataset are by using `str()`, `head()`, `tail()`, etc functions.

There is a package called “DataExplorer” which when installed can be used to create a report of the dataset using the `create_report()` function. The report shows the correlation plot, histogram, and `qplot` / scatterplot and also reveals rows and columns with missing values. This report summarizes all about the dataset giving an insight into what the dataset truly entails.

To get the mean, median, standard deviation, and mode of the dataset, we need to install a package called “skimr” or “mosaic as these two packages will help us to generate all at once.

Using “skimr”

```
install.packages("skimr")
library(skimr)
skim(WDI_New)
```

--- Data Summary ---

Name	Values
Number of rows	WDI_New 100
Number of columns	13

Column type frequency:

character	4
numeric	9

Group variables: None

--- Variable type: character ---

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
1 Country	0	1	5	18	0	10	0
2 C-Code	0	1	3	3	0	10	0
3 Year	0	1	4	4	0	10	0
4 Y-Code	0	1	6	6	0	10	0

--- Variable type: numeric ---

skim_variable	n_missing	complete_rate	mean	sd	p0
1 Inflation	0	1	4.78e 0	4.35e 0	-1.74e+ 0
2 GDP	0	1	4.43e13	5.45e13	1.65e+11
3 GDP growth	0	1	2.22e 0	3.97e 0	-1.01e+ 1
4 Unemployment	0	1	8.39e 0	5.28e 0	3.67e+ 0
5 wage and salaried workers	10	0.9	6.50e 1	2.60e 1	1.74e+ 1
6 Self-employed	10	0.9	3.50e 1	2.60e 1	6.09e+ 0
7 Imports	1	0.99	8.62e12	1.14e13	5.71e+10
8 Insurance and financial services	0	1	8.17e 0	9.20e 0	6.30e- 1
9 Export	1	0.99	7.29e11	8.56e11	3.73e+10

p25 p50 p75 p100 hist

1	1.50e+ 0	3.36e 0	7.70e 0	1.65e 1	
2	1.82e+12	1.15e13	8.38e13	2.01e14	
3	7.88e- 1	2.20e 0	4.83e 0	1.12e 1	
4	4.97e+ 0	6.95e 0	8.85e 0	2.75e 1	
5	5.27e+ 1	6.78e 1	8.65e 1	9.39e 1	
6	1.35e+ 1	3.22e 1	4.73e 1	8.26e 1	

Using the “mosaic” package, some missing values(NA) will be generated for character vectors.

```
install.packages("mosaic")
library(mosaic)
dfapply(WDI_New,favstats)
```

```

$Country
  min Q1 median Q3 max mean sd n missing
NA NA      NA NA  NA  NA  NA 0      100

$`C-Code`
  min Q1 median Q3 max mean sd n missing
NA NA      NA NA  NA  NA  NA 0      100

$Year
  min Q1 median Q3 max mean sd n missing
2011 2013 2015.5 2018 2020 2015.5 2.886751 100 0

$`Y-Code`
  min Q1 median Q3 max mean sd n missing
NA NA      NA NA  NA  NA  NA 0      100

$Inflation
  min Q1 median Q3 max mean sd n missing
-1.735888 1.495717 3.355756 7.696924 16.52354 4.780747 4.353843 100 0

$GDP
  min Q1 median Q3 max mean sd n missing
1.65326e+11 1.815863e+12 1.153366e+13 8.380673e+13 2.00749e+14 4.427112e+13 5.454826e+13 100 0

$`GDP growth`
  min Q1 median Q3 max mean sd n missing
-10.14931 0.7884818 2.203253 4.826298 11.20011 2.217762 3.966225 100 0

$Unemployment
  min Q1 median Q3 max mean sd n missing
3.67 4.9675 6.95 8.8525 27.47 8.38977 5.277659 100 0

$`wage and salaried workers`
  min Q1 median Q3 max mean sd n missing
17.41 52.7275 67.845 86.51 93.91 64.97489 26.03204 90 10

```

We need to check for missing values in the dataset, and we can identify the rows and columns with NAs and replace them with either the mean or median of the dataset. I chose to replace with mean with the assumption that the dataset is normally distributed (Learning, 2022).

```

> colnames(WDI_New)[colSums(is.na(WDI_New)) > 0]
[1] "wage and salaried workers" "Self-employed" "Imports"
[4] "Export"

```

Four columns had missing values and are replaced with their mean; after performing data cleansing, the character columns were dropped from the dataset to create the report using the “DataExplorer” library.

We could not determine the mode of the dataset because they are continuous variables and are not discrete.

The skewness and kurtosis of the dataset can be determined by installing the package “moments” and loading its library.

```

###Checking skewness and kurtosis
install.packages("moments")
library(moments)

skewness(WDI_New)
      Inflation      GDP      GDP growth
      0.9373690      1.0820825      -0.6691382
Unemployment wage and salaried workers Self-employed
      2.0013208      -0.7495344      0.7495503
Imports Insurance and financial services Export
      1.3751306      1.5005409      1.2921545

kurtosis(WDI_New)
      Inflation      GDP      GDP growth
      3.173629      3.175222      4.104017
Unemployment wage and salaried workers Self-employed
      6.667478      2.478915      2.478931
Imports Insurance and financial services Export
      4.034127      3.897953      3.045969

```

The result indicates a substantially skewed distribution, both negatively and positively; and kurtosis indicates that some variables in the distribution are too peaked which is popularly known as a leptokurtic distribution.

4.1b. Graphical Exploration

Graphical methods are used for the graphical representation of the dataset.

The graphs below are representations of Inflation and GDP on the dataset.

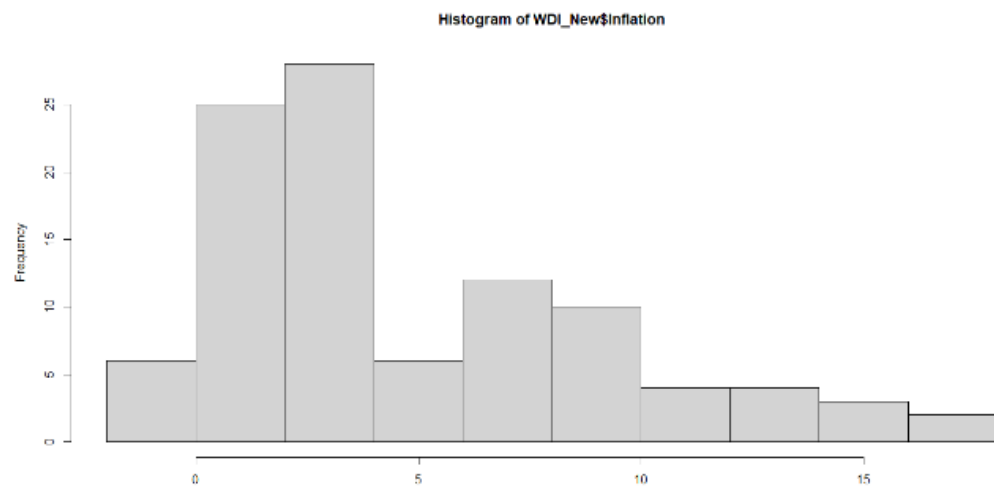
```
plot(WDI_New)
```

Figure 2.2: Plot of the whole dataset



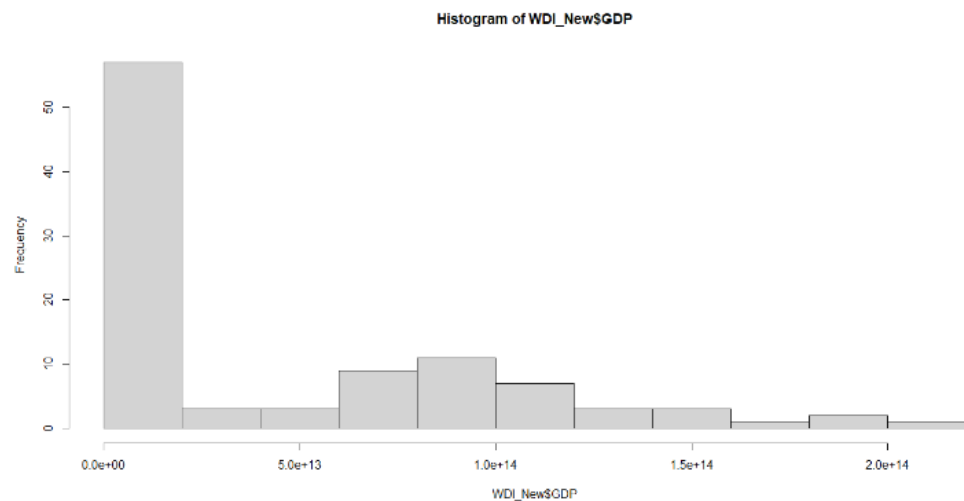
```
hist(WDI_New$Inflation)
```

Figure 2.3: Histogram of Inflation Indicator



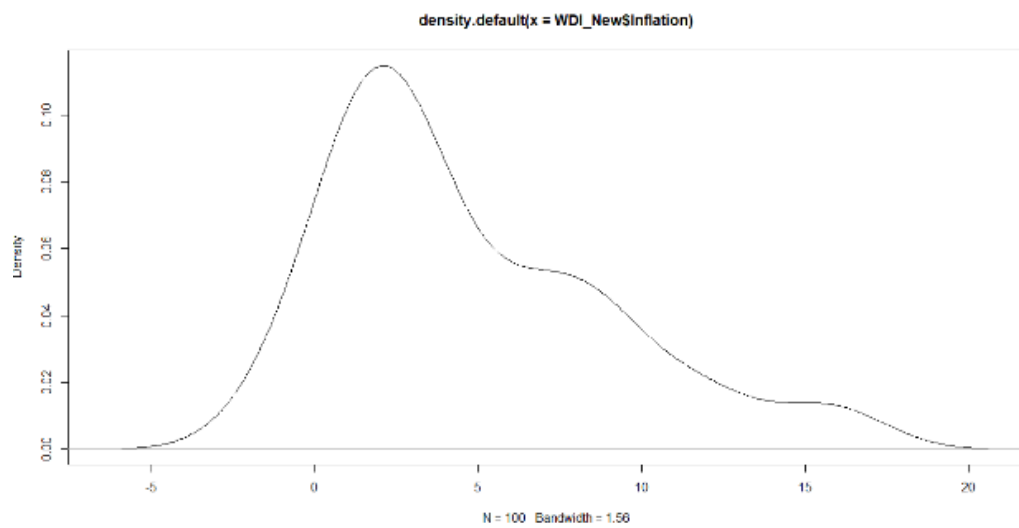
```
hist(WDI_New$GDP)
```

Figure 2.4: Histogram of GDP Indicator



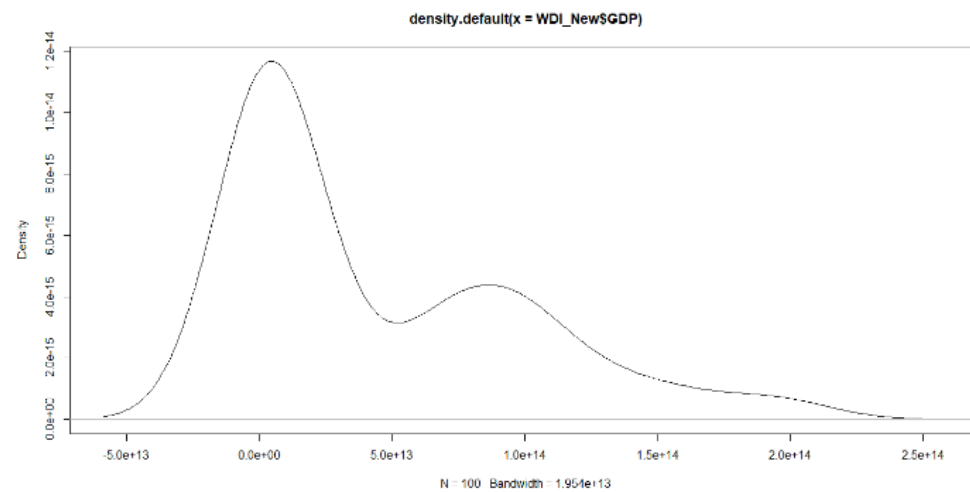
```
plot(density(WDI_New$Inflation))
```

Figure 2.5: Density plot of Inflation



```
plot(density(WDI_New$GDP))
```

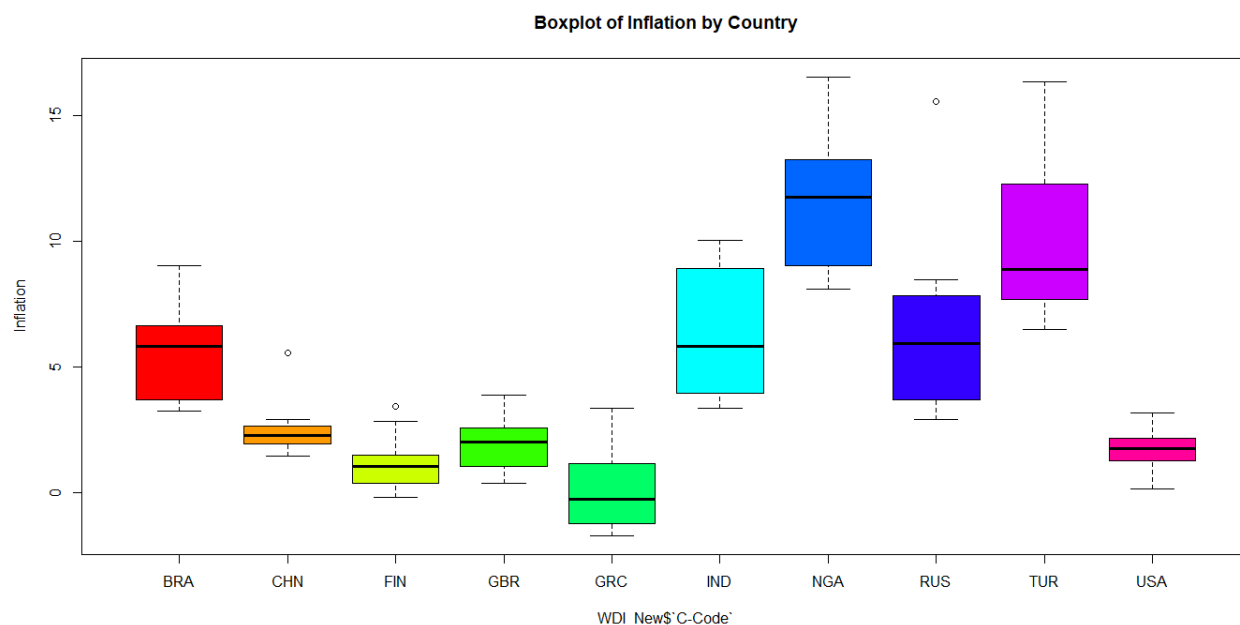
Figure 2.6: Density plot of GDP



Using boxplot to visualize the dataset and check for outliers

```
boxplot(WDI_New$Inflation~WDI_New$`C-Code`,
        main='Boxplot of Inflation by Country',ylab='Inflation',
        col= rainbow(10))->b_Inflation_Country
```

Figure 2.7: Boxplot of Inflation by Country



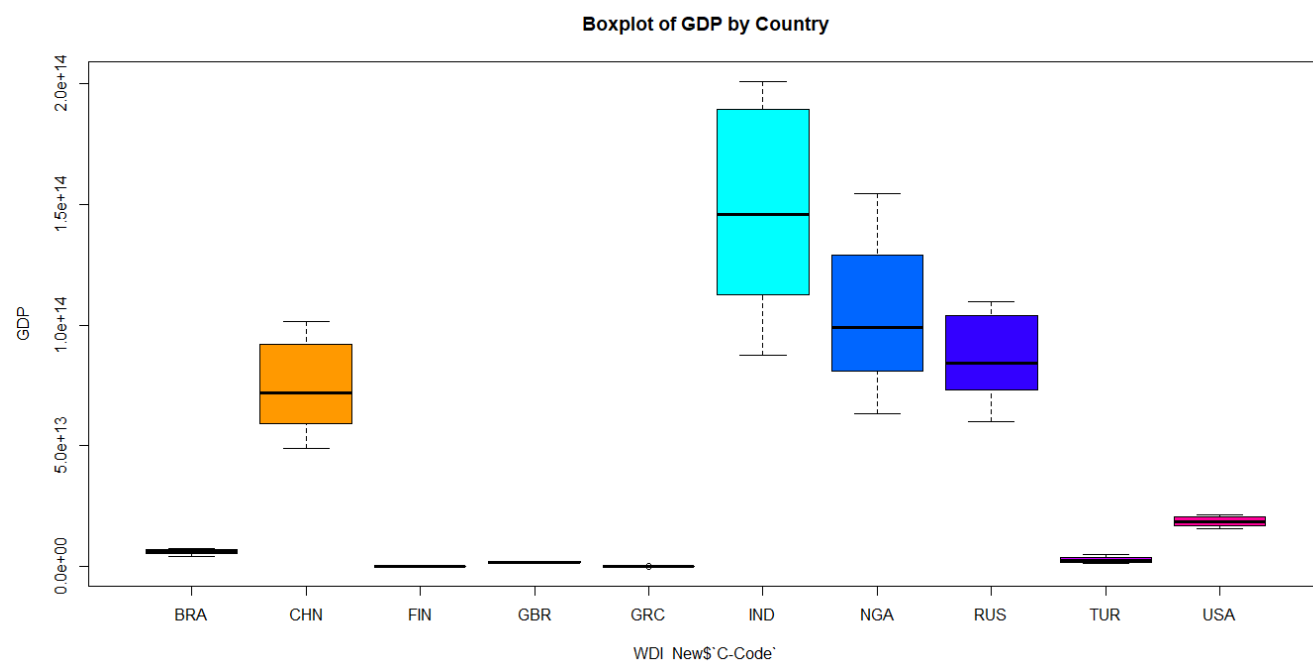
The boxplot reveals the median, and interquartile range and also shows outliers within the plots. From the boxplot, we can see that there are a few points on the high end that appear to be outliers which are in China, Finland, and Russia.

To identify the points of outliers, we can use the below syntax;

```
> ###To identify the outliers
> b_inflation_Country$out
[1] 5.553899 3.416808 15.534405

boxplot(wDI_New$GDP~wDI_New$`C-Code`,
        main='Boxplot of GDP by Country',ylab='GDP'
        col= rainbow(10))>b_GDP_Country
```

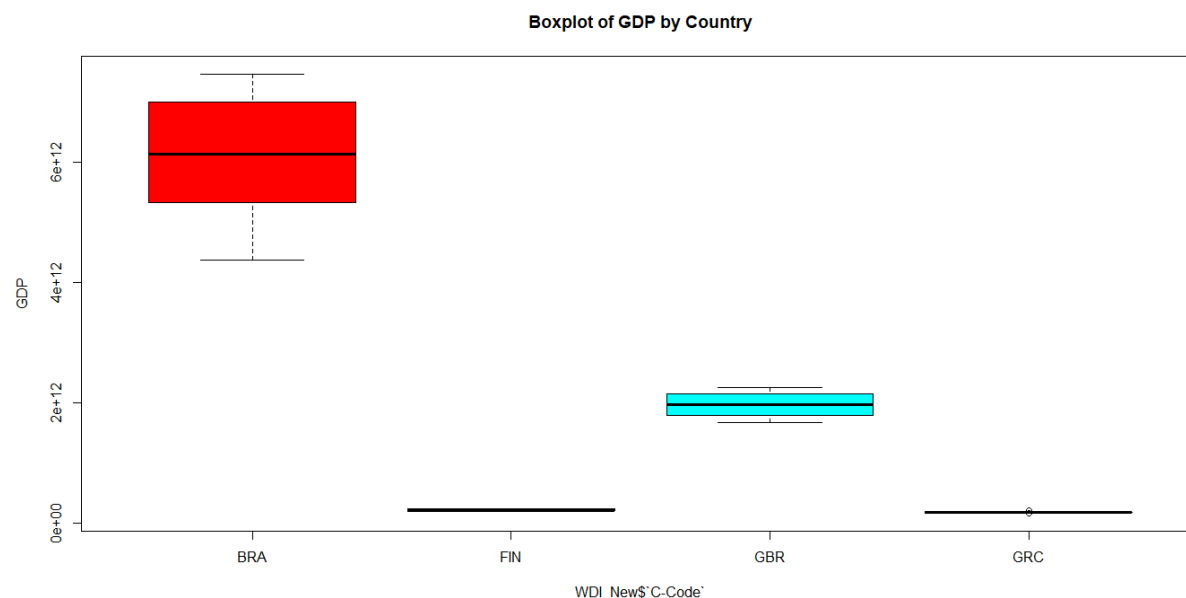
Figure 2.8a: Boxplot of GDP by Country



The plot could only clearly review that of China, India, Nigeria, Russia, partially Turkiye, and the USA. We need to run the codes again by defining ing subset to review the countries not visible and check for outliers.

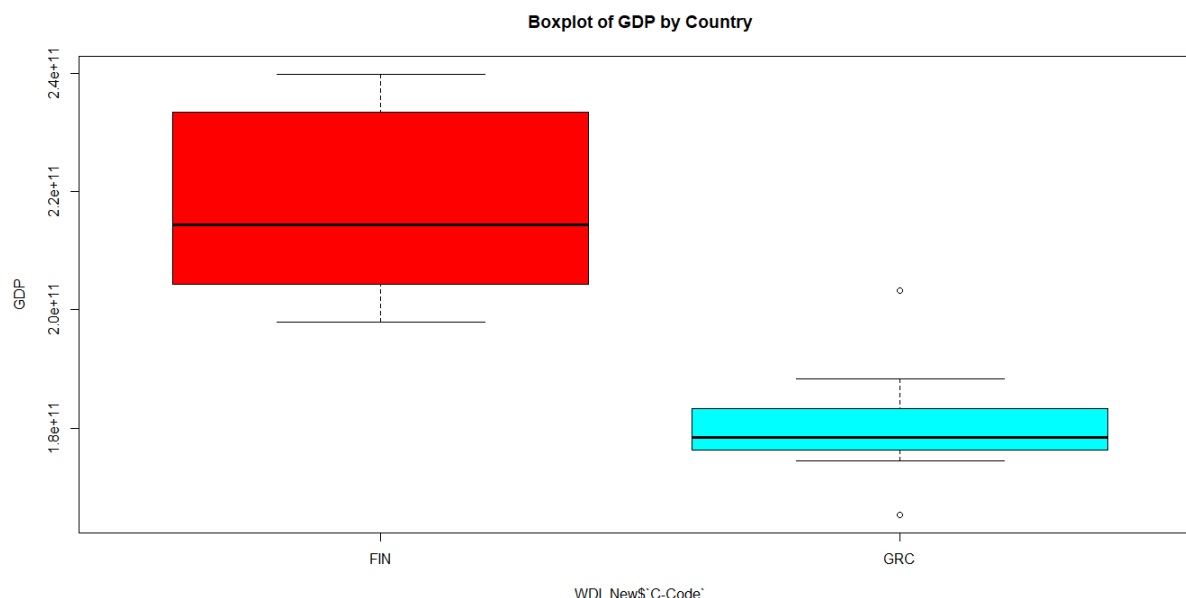
```
###We need to create a subset
boxplot(WDI_New$GDP~WDI_New$`C-Code`,
        main='Boxplot of GDP by Country',ylab='GDP',
        col= rainbow(4),
        subset= WDI_New$`C-Code` %in% c("BRA","FIN","GRC","GBR"))
```

Figure 2.8b: Boxplot of GDP by Country



Still, we need to check for Finland and Greece

Figure 2.8c: Boxplot of GDP by Country



The plot shows outliers on the both high and low end in the boxplot for Greece.

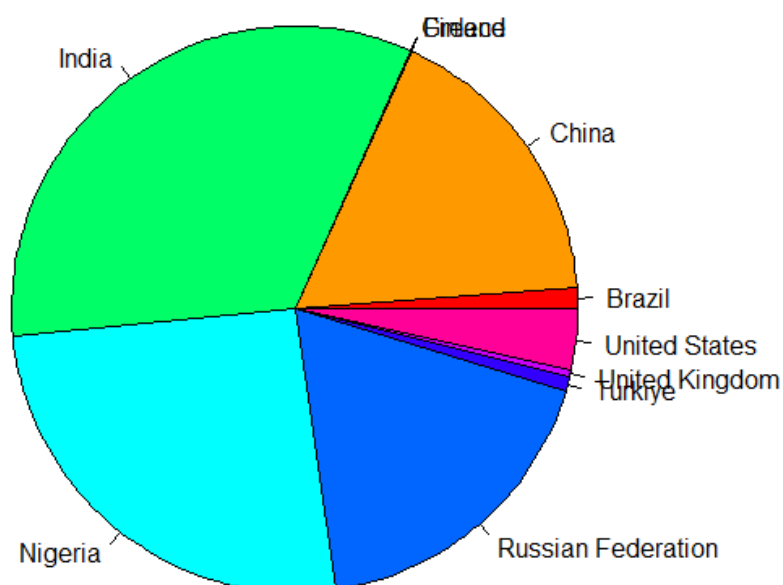
```
> b_GDP_Country$out
[1] 2.03308e+11 1.65326e+11
```

Using the Pie function to evaluate each country in 2020 the recent year in the dataset.

```
###Using Pie Function to evaluate each country in 2020
WDI_2020 <- WDI_New[WDI_New$Year==2020,]
WDI_2020

pie(WDI_2020$GDP, labels=WDI_2020$Country,col=rainbow(10))
```

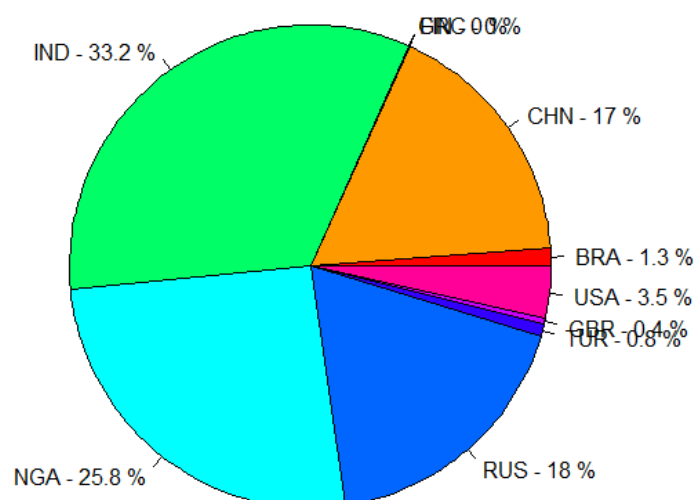
Figure 2.9a: Pie Chart of GDP by Country



The chart showed that India has the highest GDP in 2020 while Finland and Greece are the lowest with the same GDP. Presenting the chart to show their percentage;

```
percent <- round(100*WDI_2020$GDP/sum(WDI_2020$GDP),1)
percent <- paste(WDI_2020$c-Code`, "-",percent,"%")
pie(WDI_2020$GDP, labels=percent,col=rainbow(10))
```

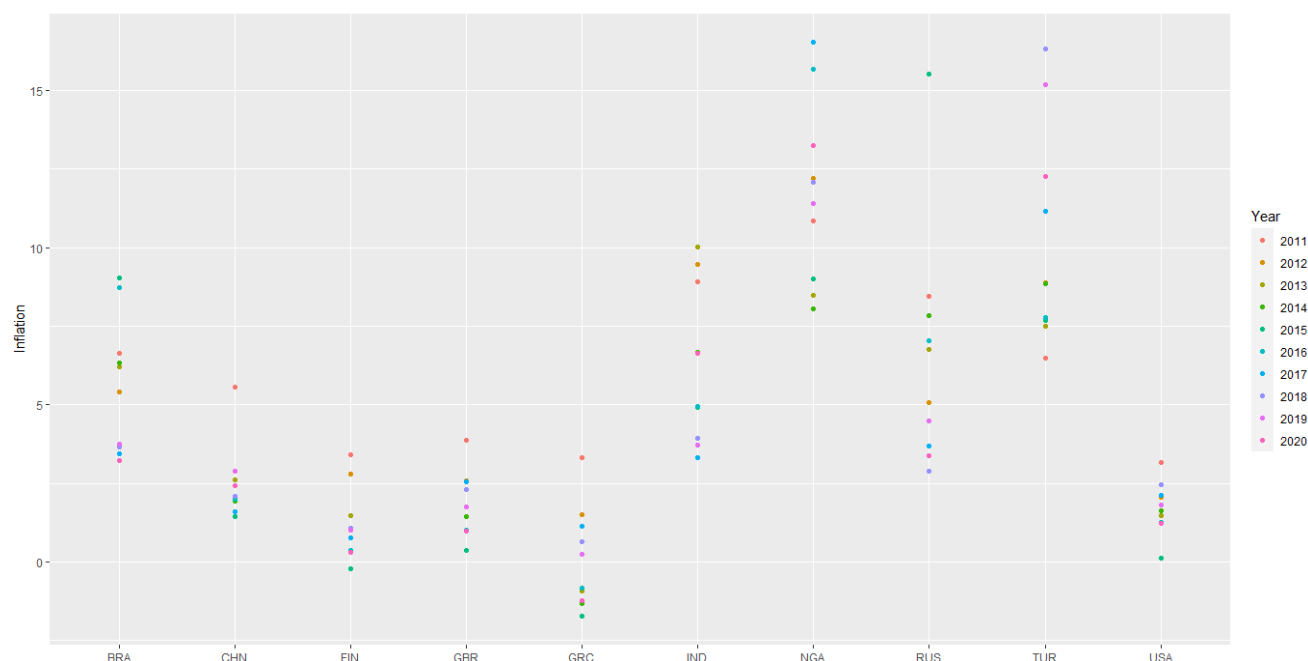
Figure 2.9b: Pie Chart of GDP by Country



Making scatterplot with the qplot function

```
qplot(`c-Code`,Inflation,data=WDI_New,color = Year)
```

Figure 2.10: Scatter plot of Inflation by Country for 10 years



All these graphical representations and more are mostly used to visualize the dataset and have a better understanding of it even without running any model.

4.2. Correlation Analysis

Correlation analysis simply means the statistical method that defines the strength of the relationship between two or more variables showing the magnitude in numerical figures usually between -1 and +1.

Since our focus is on inflation and GDP, we will run a correlation between the two variables using Pearson and Spearman correlation method after which we will run for the whole dataset.

```
> ##Pearson correlation
> cor(WDI_New$Inflation,WDI_New$GDP)
[1] 0.3438165
> ##Spearman correlation
> cor(WDI_New$Inflation,WDI_New$GDP, method = "spearman")
[1] 0.5313531
```

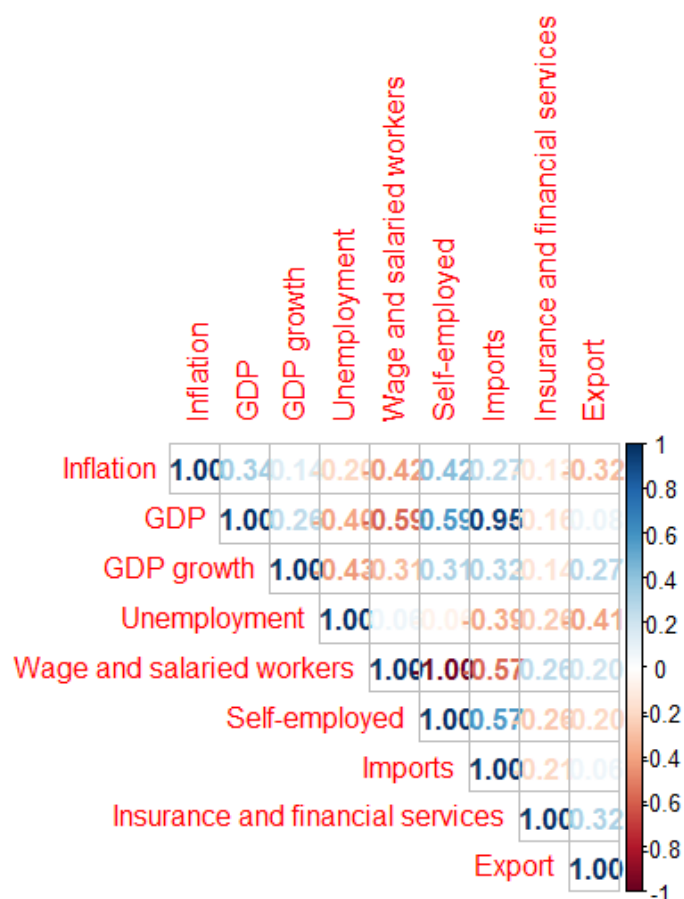
While Pearson shows a 34% positive correlation between Inflation and GDP, Spearman showed a 53% positive correlation.

We need to load the library("corrplot") to run a correlation analysis for all variables in the dataset.

The syntax can be in two forms; to display the figures and without figures

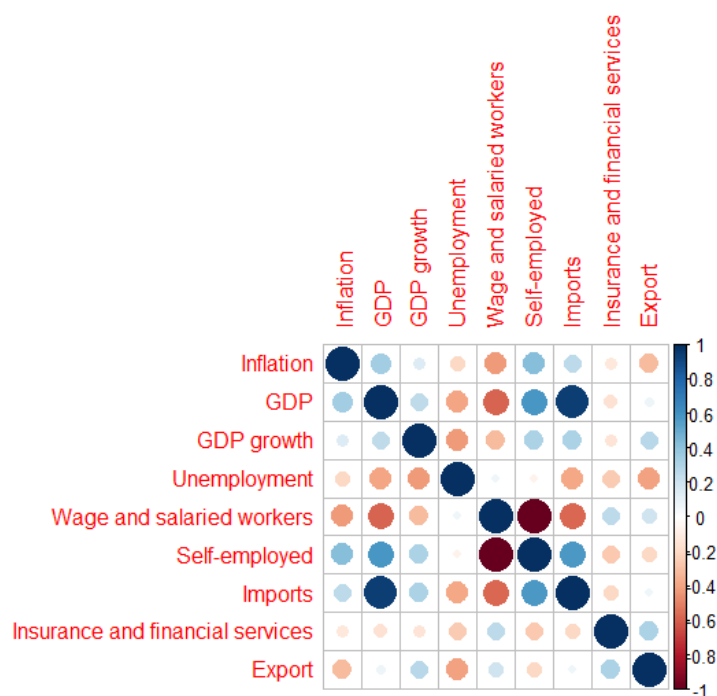
```
library("corrplot")
corrplot(cor(WDI.New), method = "number", type = "upper")
```


Figure 2.11a: Correlation plot of all indicators



####To display the full correlation matrix without figures
`corrplot(cor(WDI.New))`

Figure 2.11b: Correlation plot of all indicators



According to the plots and with emphasis on GDP, the following are the correlation matrix figures;

- Positive correlation of 0.34 between GDP and Inflation
- Positive correlation of 0.59 between GDP and Self-employed
- Positive correlation of 0.95 between GDP and imports
- Negative correlation of 0.20 between GDP and unemployment
- Negative correlation of 0.42 between GDP and Wage and salaried workers
- Negative correlation of 0.16 between GDP and Insurance and financial services
- Positive correlation of 0.08 between GDP and Exports

While other variables have little or insignificant correlation with GDP, other variables are highly correlated with each other. For example, Self-employed and Wage and the salaried worker has a very high negative correlation (-1) between each other which means variability in either of them does not have any effect on the other.

Furthermore, the result shows that there is a 95% correlation between GDP and Imports which could mean that importation has a major effect or makes a lump sum of the GDP in this dataset. Other variables like Inflation, Self-employed, Unemployment, Wage and salaried workers, and Insurance and financial services have a lower correlation compared to Imports.

4.3. Regression Analysis

This is a model built to mathematically sort out variables that have a high impact on a chosen target or dependent variable.

The objective of this regression analysis is to determine the possible linear relation between GDP(target variable) and other variables with an emphasis on Inflation. We will be running two regression models; Simple Linear Regression and Multiple Linear Regression. This model is selected because our predictor and target variables are continuous.

- Simple Linear Regression:** This is to determine a linear relationship between GDP and Inflation, GDP and Self-employed which are the most correlated. Although, it is either use the forward stepwise method or the backward stepwise method we will be running this analysis with Inflation and Self-employed as the only independent variables with 34% and 59% correlation respectively.

```
#####REGRESSION ANALYSIS#####
install.packages("car")
install.packages("caret")
library("car")
library("caret")

##Based on the correlation, will run simple linear regression on the most
##correlated variables to GDP
#Y = GDP
#X = Inflation

model_1 <- lm(GDP~Inflation,WDI.New)
summary.lm(model_1)

Call:
lm(formula = GDP ~ Inflation, data = WDI.New)

Residuals:
    Min       1Q   Median       3Q      Max
-9.027e+13 -3.180e+13 -1.709e+13  3.163e+13  1.610e+14

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.368e+13  7.667e+12   3.088  0.002620 **
Inflation    4.308e+12  1.188e+12   3.625  0.000461 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.148e+13 on 98 degrees of freedom
Multiple R-squared:  0.1182,    Adjusted R-squared:  0.1092
F-statistic: 13.14 on 1 and 98 DF,  p-value: 0.0004612
```

The result shows that both coefficients are significant and the R^2 is approximately 0.11. So the SLR equation will be:

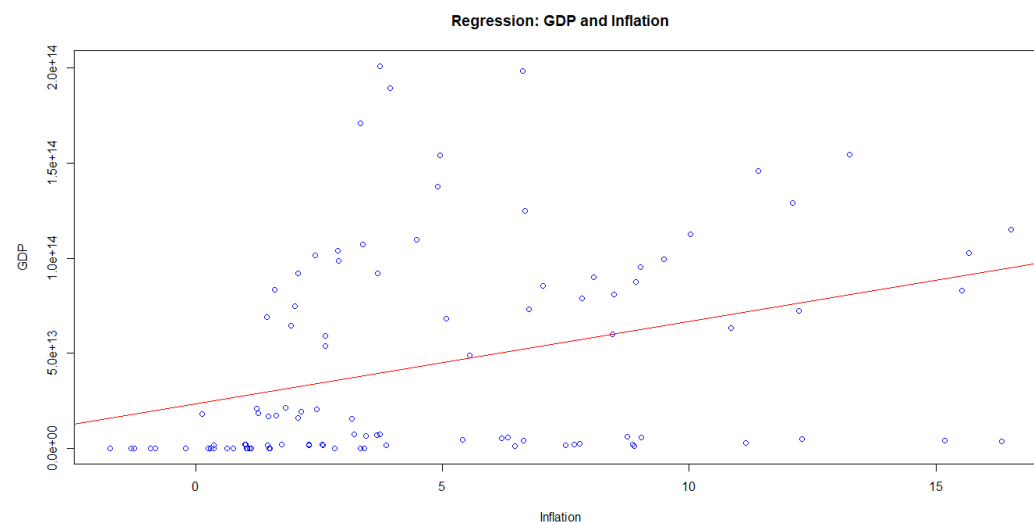
$$\text{GDP} = 2.368\text{e}+13 + 4.308\text{e}+12 \times \text{Inflation}$$

This means that with this equation, Inflation can predict an 11% variability of the entire GDP.

Let's draw the scatter plot and the regression line:

```
plot(GDP~Inflation,WDI.New, col="blue",
     main = "Regression: GDP and Inflation",
     xlab = "Inflation",
     ylab = "GDP")
abline(model_1,col="red")
```

Figure 2.12: Scatter plot and Regression line of GDP by Inflation



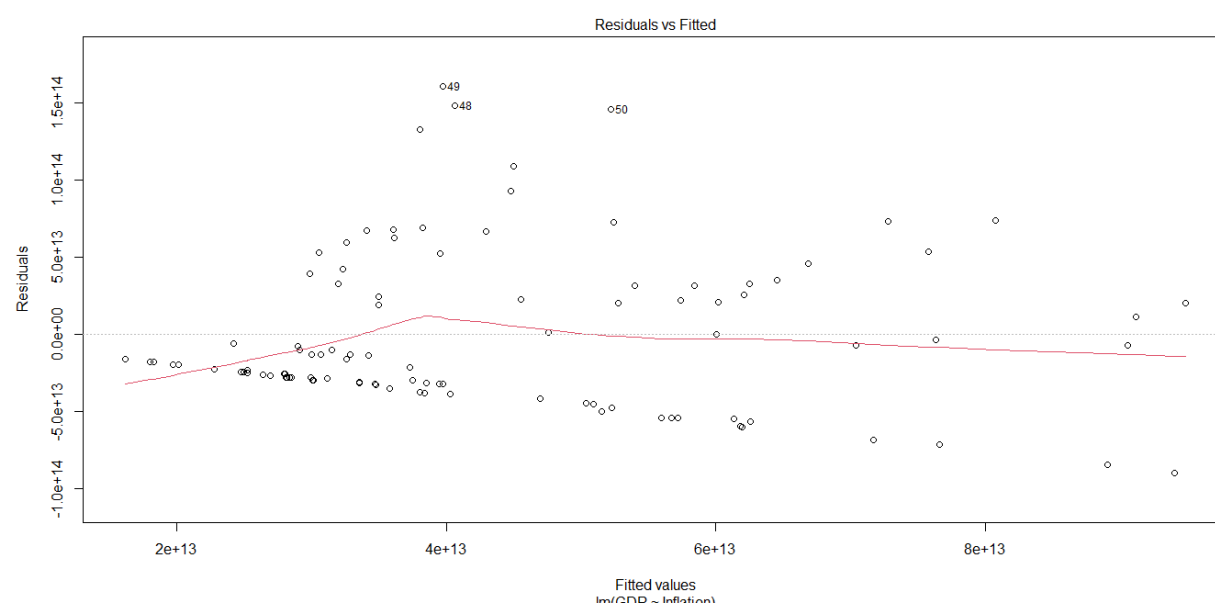
Now, let's check if the fitted model meets the assumptions of SLR.

Linearity: Checking the linearity between GDP and Inflation is the plotted regression line on the scatter plot. The relationship is linear.

Residuals' Independence: This is done by examining the scatterplots of “the residuals versus fits”. The correlation must be approximately zero meaning that it should not look like it has any relationship.

```
plot(model_1,1)
```

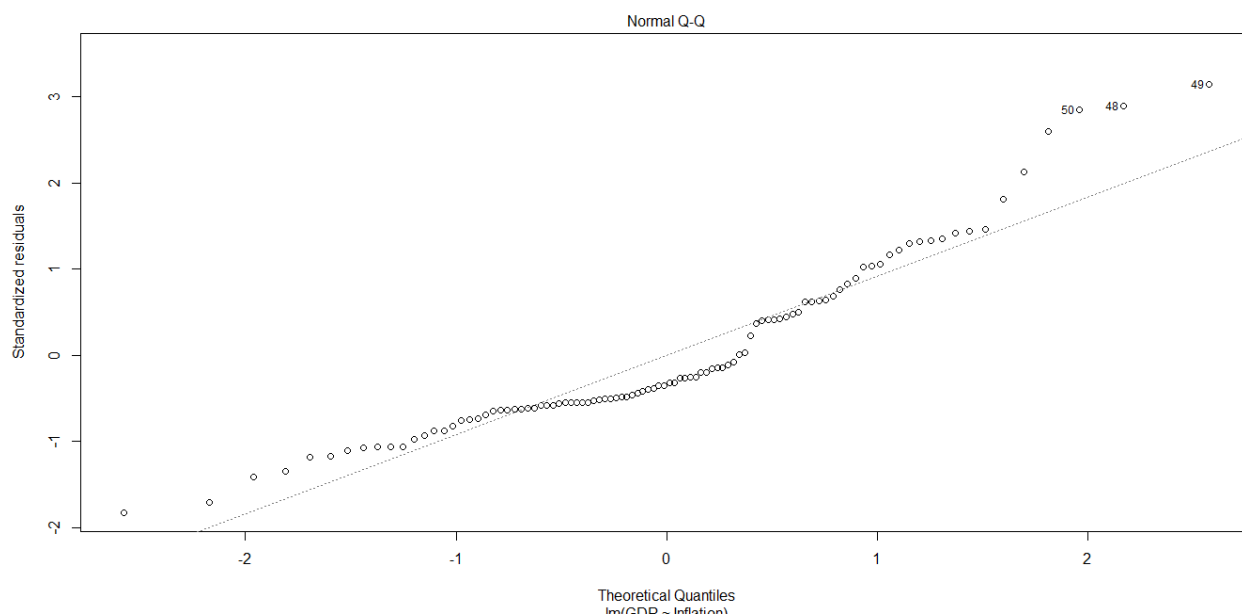
Figure 2.13: Scatter plot of the residuals and fitted



Normality of Residuals: The residuals when plotted should be approximately normally distributed. This is examined with a plot that would usually fall within a straight line if they are normally distributed. It can also be examined with a histogram plot.

```
plot(model_1,2)
```

Figure 2.14: Normality of residuals plot

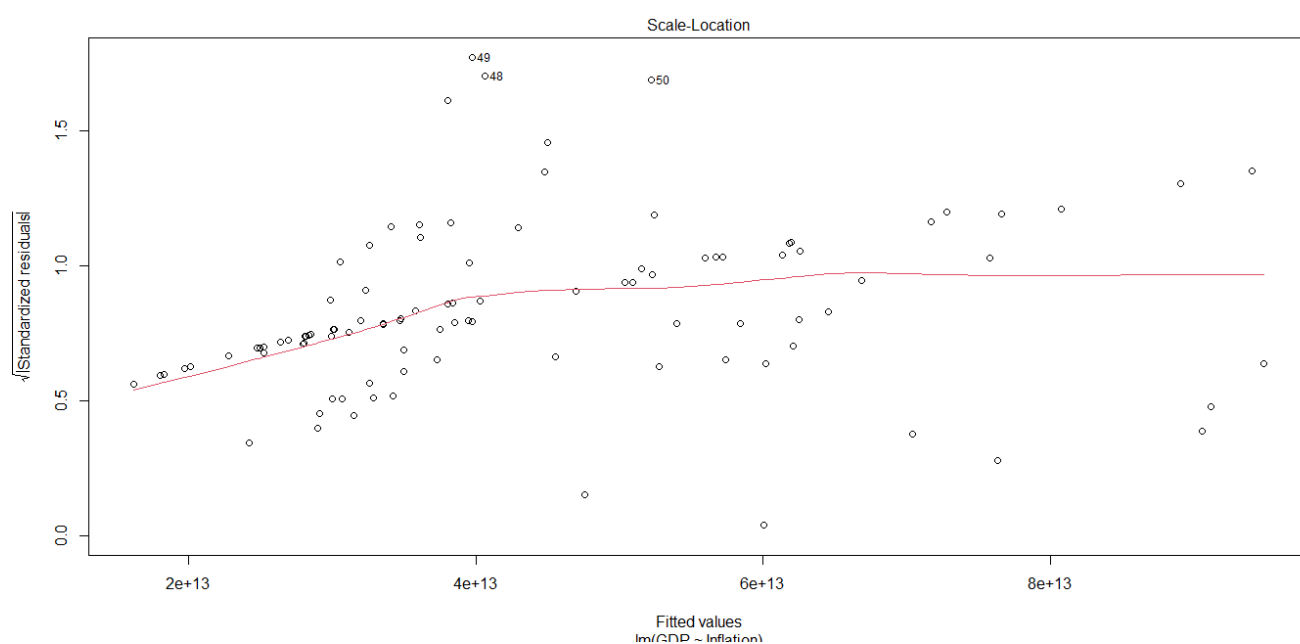


The plot shows that they are not normally distributed as most of the plots are far away from the line. So this assumption is not met meaning this model does not fit.

Equal Variance of Residuals(Homoscedasticity): This assumption says that the variance of the residuals is constant and not related to the fitted value.

```
plot(model_1,3)
```

Figure 2.15: Homoscedasticity plot



Although the residuals are randomly scattered around the red line we cannot approve the model because one of the assumptions is not met.

Now running SLR for GDP and Self-employed.

```
##Correlation of GDP and Self-employed
#Y = GDP
#X = Self-employed

model_1a <- lm(GDP~`Self-employed`,WDI.New)
summary.lm(model_1a)

Call:
lm(formula = GDP ~ `Self-employed`, data = WDI.New)

Residuals:
    Min       1Q   Median       3Q      Max
-4.709e+13 -3.461e+13 -1.610e+13  1.462e+13  1.537e+14

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.025e+12  7.744e+12  -0.132    0.895
`Self-employed` 1.293e+12  1.810e+11   7.143 1.62e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.446e+13 on 98 degrees of freedom
Multiple R-squared:  0.3424,    Adjusted R-squared:  0.3357
F-statistic: 51.03 on 1 and 98 DF,  p-value: 1.621e-10
```

The result shows that the coefficient of intercept is not significant while that of Self-employed is significant. The Adjusted R² is approximately 0.34. So the SLR equation will be:

$$\text{GDP} = 1.293\text{e}+12 \times \text{Self-employed}$$

This means that with this equation, the Self-employed can predict a 34% variability of the entire GDP.

After checking the assumptions for this model on Self-employed, the residuals are randomly scattered but not around the line; so this model does not meet all the assumptions. Hence, the SLR model is not a good fit.

b. Multiple Linear Regression

Here we will be determining a linear relationship between the target variable and two or more predictive or independent variables; the focus will be on variables that are **positively** correlated to GDP.

We will start with two variables, Inflation and Self-employed since both have a higher correlation than others.

```
###PERFORMING MULTIPLE LINEAR REGRESSION MODEL###
## Y = GDP
## x1 = Inflation
## x2 = Self-employed

model_2 = lm(GDP~Inflation + `Self-employed`,WDI.New)
summary.lm(model_2)

Call:
lm(formula = GDP ~ Inflation + `Self-employed`, data = WDI.New)

Residuals:
    Min       1Q   Median       3Q      Max
-5.398e+13 -3.493e+13 -1.366e+13  1.686e+13  1.510e+14

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.228e+12  8.096e+12  -0.522    0.603
Inflation    1.475e+12  1.128e+12   1.307    0.194
`Self-employed` 1.183e+12  1.990e+11   5.947 4.32e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.43e+13 on 97 degrees of freedom
Multiple R-squared:  0.3538,    Adjusted R-squared:  0.3405
F-statistic: 26.55 on 2 and 97 DF,  p-value: 6.352e-10
```

The result showed that only Self-employed is significant, so it makes no difference adding Inflation to the equation.

Now let's add another variable;

```
##The next correlated variable is Imports
## Y = GDP
## X1 = Inflation
## X2 = Self-employed
## X3 = Imports

model_3 = lm(GDP~Inflation + `Self-employed` + Imports ,WDI.New)
summary.lm(model_3)

Call:
lm(formula = GDP ~ Inflation + `Self-employed` + Imports, data = WDI.New)

Residuals:
    Min       1Q   Median       3Q      Max
-5.068e+13 -5.081e+12 -1.458e+12  4.518e+12  1.004e+14

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -6.946e+11  3.106e+12  -0.224   0.8235
Inflation      1.127e+12  4.325e+11   2.605   0.0107 *
`Self-employed` 5.747e+10  8.977e+10   0.640   0.5236
Imports        4.359e+00  1.834e-01  23.764 <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.697e+13 on 96 degrees of freedom
Multiple R-squared:  0.9061,    Adjusted R-squared:  0.9032
F-statistic: 308.8 on 3 and 96 DF,  p-value: < 2.2e-16
```

The result showed that Imports and Inflation are significant and the adjusted R^2 is approximately 0.90. So the MLR equation will be:

$$\text{GDP} = 1.126\text{e}+12 \times \text{Inflation} + 4.359\text{e}+00 \times \text{Imports}$$

This means that with this equation, Inflation and Imports can predict a 90% variability of the entire GDP. So adding these two variables makes a huge difference in the adjusted R^2 .

Let's add another variable;

```
##The next correlated variable is Exports
## Y = GDP
## X1 = Inflation
## X2 = Self-employed
## X3 = Imports
## X4 = Export

model_4 = lm(GDP~Inflation + `Self-employed` + Imports + Export,WDI.New)
summary.lm(model_4)

Call:
lm(formula = GDP ~ Inflation + `Self-employed` + Imports + Export,
    data = WDI.New)

Residuals:
    Min       1Q   Median       3Q      Max
-5.024e+13 -5.020e+12 -3.591e+11  3.827e+12  9.838e+13

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -5.477e+12  3.880e+12  -1.412   0.16133
Inflation      1.371e+12  4.430e+11   3.094   0.00259 **
`Self-employed` 9.293e+10  9.015e+10   1.031   0.30525
Imports        4.270e+00  1.860e-01  22.959 < 2e-16 ***
Export         4.305e+00  2.150e+00   2.003   0.04806 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.671e+13 on 95 degrees of freedom
Multiple R-squared:  0.9099,    Adjusted R-squared:  0.9061
F-statistic: 239.9 on 4 and 95 DF,  p-value: < 2.2e-16
```

The result showed that only Imports, Exports, and Inflation are significant and the adjusted R^2 is approximately 0.91. So the MLR equation will be:

$$\text{GDP} = 1.371\text{e}+12 \times \text{Inflation} + 4.270\text{e}+00 \times \text{Imports} + 4.305\text{e}+00 \times \text{Exports}$$

This means that with this equation, Inflation, Imports, and Exports can predict 91% variability of the entire GDP. So adding the Export variable to the model makes little or no difference compared with the previous model in the adjusted R^2 while Self-employed is still insignificant.

So, we decided to remove Self-employed from the equation and run the fifth model on Inflation, Imports, and Exports only.

```
model_5 = lm(GDP~Inflation + Imports + Export,WDI.New)
summary.lm(model_5)

Call:
lm(formula = GDP ~ Inflation + Imports + Export, data = WDI.New)

Residuals:
    Min       1Q   Median       3Q      Max
-4.906e+13 -5.117e+12 -9.076e+11  4.466e+12  9.735e+13

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.396e+12  3.314e+12  -1.025  0.308148
Inflation    1.493e+12  4.270e+11   3.495  0.000719 ***
Imports      4.375e+00  1.555e-01  28.146 < 2e-16 ***
Export       3.870e+00  2.108e+00   1.835  0.069541 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.672e+13 on 96 degrees of freedom
Multiple R-squared:  0.9089,    Adjusted R-squared:  0.9061
F-statistic: 319.3 on 3 and 96 DF,  p-value: < 2.2e-16
```

There is a negative intercept that is not significant, Imports and Exports are highly significant while Export has little significance. The adjusted R^2 is approximately 0.91.

Since we realize that Export is not so relevant, we decided to run the sixth model with only Inflation and Imports but the R^2 was approximately 0.90, then we decided that the 5th model is the best fit with variables Inflation, Imports, and Exports.

The MLR equation will be:

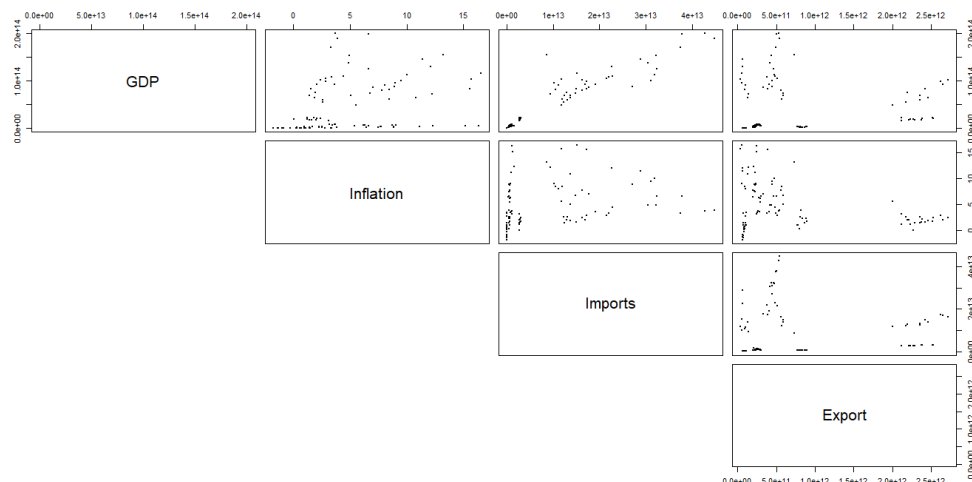
$$\text{GDP} = 1.493\text{e}+12 \times \text{Inflation} + 4.375\text{e}+00 \times \text{Imports} + 3.870\text{e}+00 \times \text{Exports}$$

Checking the five assumptions of the model;

Linearity: The relationship between the target and independent variables must be linear and so we need to find the index of each variable and examine it with the scatterplot.

```
##Linearity:
colnames(WDI.New)
##To draw scatterplot, we need to input the indices of the variables
##Putting GDP as the first variable (index number 2)
pairs(WDI.New[,c(2,1,7,9)], lower.panel = NULL, pch = 19, cex = 0.2)
```

Figure 2.16: Scatter plot for linearity of GDP, Inflation, Imports and Export

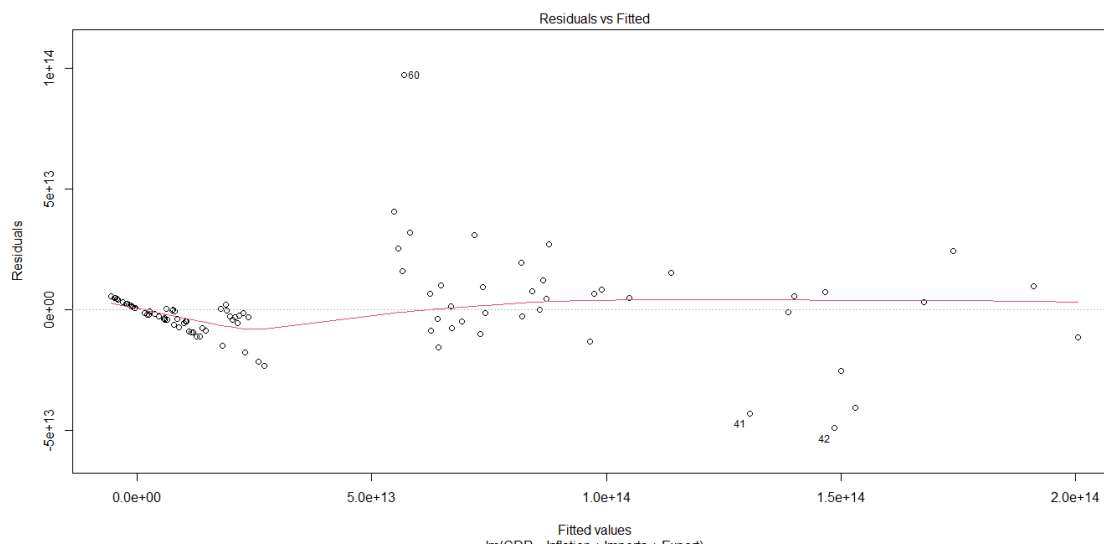


The first row signifies that all variables have a linear relation with GDP.

Residuals' Independence: This plot would not have a pattern where the red line is approximately horizontal at zero.

```
plot(model_5,1)
```

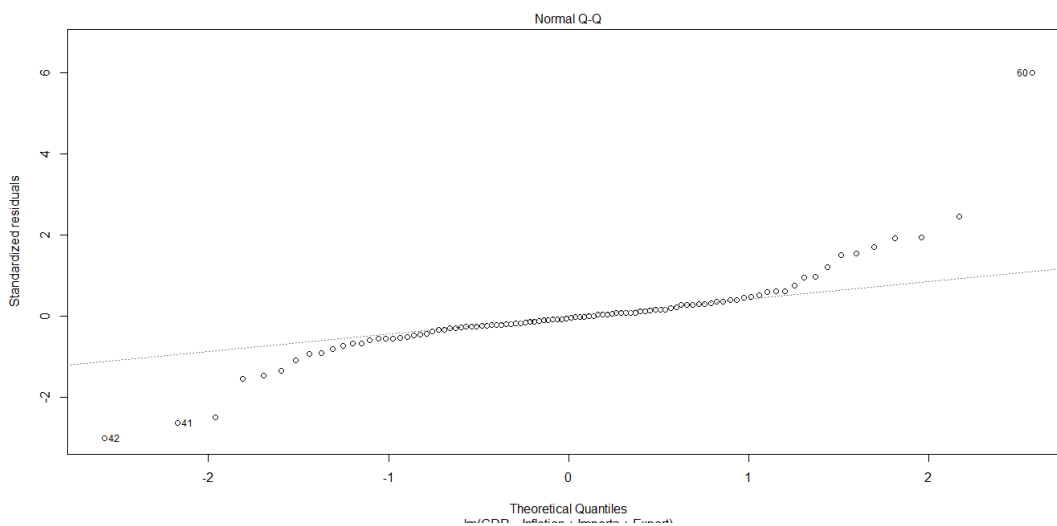
Figure 2.17: Scatterplot of Residuals and Fitted



Normality of residuals: The residuals must be approximately normally distributed so the observations should be near the line.

```
plot(model_5,2)
```

Figure 2.18: Normality of residuals plot

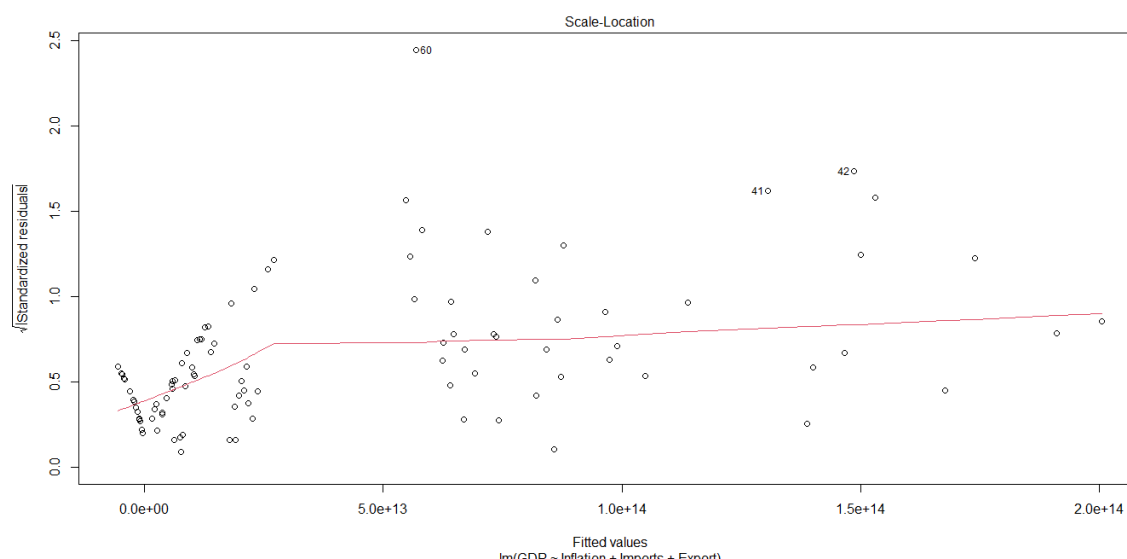


Most of the residuals are near the line which shows that this assumption has been met.

Homoscedasticity: Here the residuals must be scattered around the regression line.

```
plot(model_5,3)
```


Figure 2.19: Homoscedasticity plot



No Multicollinearity: MLR assumes that none of the independent variables otherwise known as predictor variables are highly correlated with each other. We obtain this by calculating the Variance Inflation Factor(VIF) which measures the correlation and strength between the predictor variables.

```
vif(model_5)
```

```
Inflation Imports Export
1.224193 1.105208 1.140826
```

The result shows that there is no correlation between the predictor variables.

Now we can say that our regression line is:

$$\text{GDP} = 1.493\text{e}+12 \times \text{Inflation} + 4.375\text{e}+00 \times \text{Imports} + 3.870\text{e}+00 \times \text{Exports}$$

According to the MLR model, Inflation, Imports, and Exports can predict 91% variability in the GDP of our dataset.

4.4. Time Series Analysis

We need to install the packages we need for this analysis; “TTR” and “forecast” and run the analysis on the Inflation rate of the dataset setting the frequency at 12 i.e. every month.

```
> Inf_series<- ts(WDI_New$Inflation,frequency=12,start=(2011),end=(2020))
> Inf_series
```

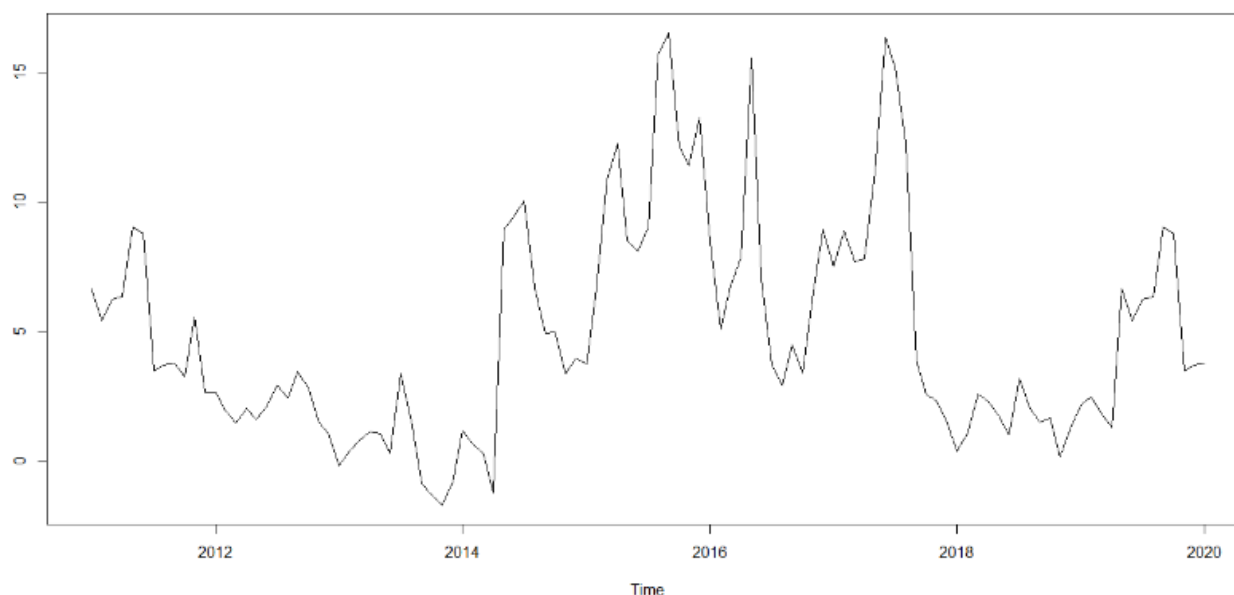
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep
2011	6.6364496	5.4034991	6.2043107	6.3290402	9.0299010	8.7391435	3.4463734	3.6648503	3.7329762
2012	2.6210500	1.9216416	1.4370238	2.0000018	1.5931360	2.0747904	2.8992342	2.4194219	3.4168075
2013	-0.2079288	0.3566845	0.7540150	1.0838210	1.0240939	0.2905546	3.3298532	1.5015270	-0.9212695
2014	1.1212545	0.6256214	0.2530075	-1.2479836	8.9117934	9.4789969	10.0178785	6.6656567	4.9069734
2015	3.7295057	6.6234368	10.8400275	12.2177817	8.4758273	8.0624858	9.0093872	15.6753405	16.5235400
2016	8.4404649	5.0747430	6.7537103	7.8234118	15.5344051	7.0424476	3.6833294	2.8782972	4.4703666
2017	7.4930903	8.8545727	7.6708536	7.7751342	11.1443111	16.3324639	15.1768216	12.2789575	3.8561124
2018	0.3680468	1.0084174	2.5577558	2.2928399	1.7381046	0.9894867	3.1568416	2.0693373	1.4648327
2019	2.1301100	2.4425833	1.8122101	1.2335844	6.6364496	5.4034991	6.2043107	6.3290402	9.0299010
2020	3.7329762								
	Oct	Nov	Dec						
2011	3.2117680	5.5538989	2.6195243						
2012	2.8083362	1.4782862	1.0411962						
2013	-1.3122610	-1.7358880	-0.8256540						
2014	4.9482163	3.3281734	3.9388265						
2015	12.0947316	11.3967950	13.2460234						
2016	3.3816594	6.4718797	8.8915700						
2017	2.5732348	2.2916667	1.4511202						
2018	1.6222230	0.1186271	1.2615832						
2019	8.7391435	3.4463734	3.6648503						
2020									

The result shows a time series set for inflation every month from 2011 to 2020.

Plotting Time Series: This is to visualize the time series before running any model

```
plot.ts(Inf_series)
```

Figure 2.20: Inflation Time Series Plot



An additive model can be used to explain this plot, being that the random fluctuations in the data are roughly constant in size over time. We can say it is seasonal time series data.

Decomposing Time Series(Seasonal): This is done to separate the time series into its separate components, which are mostly trend, seasonal, and irregular.

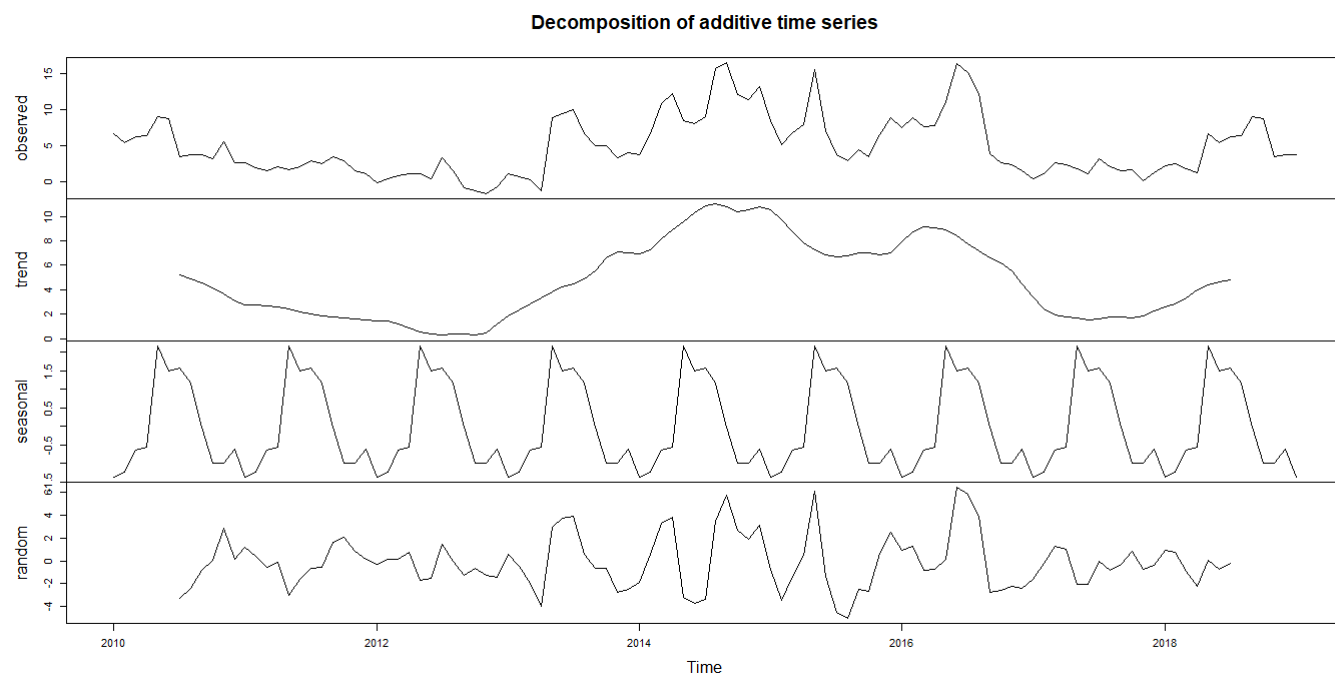
```
> Inf_seriescomp <- decompose(Inf_series)
> Inf_seriescomp$seasonal
```

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug
2011	-1.36048650	-1.23721334	-0.63231461	-0.55119272	2.16593677	1.49854457	1.56003662	1.16217389
2012	-1.36048650	-1.23721334	-0.63231461	-0.55119272	2.16593677	1.49854457	1.56003662	1.16217389
2013	-1.36048650	-1.23721334	-0.63231461	-0.55119272	2.16593677	1.49854457	1.56003662	1.16217389
2014	-1.36048650	-1.23721334	-0.63231461	-0.55119272	2.16593677	1.49854457	1.56003662	1.16217389
2015	-1.36048650	-1.23721334	-0.63231461	-0.55119272	2.16593677	1.49854457	1.56003662	1.16217389
2016	-1.36048650	-1.23721334	-0.63231461	-0.55119272	2.16593677	1.49854457	1.56003662	1.16217389
2017	-1.36048650	-1.23721334	-0.63231461	-0.55119272	2.16593677	1.49854457	1.56003662	1.16217389
2018	-1.36048650	-1.23721334	-0.63231461	-0.55119272	2.16593677	1.49854457	1.56003662	1.16217389
2019	-1.36048650	-1.23721334	-0.63231461	-0.55119272	2.16593677	1.49854457	1.56003662	1.16217389
2020	-1.36048650							

	Sep	Oct	Nov	Dec
2011	-0.01241027	-0.97829980	-0.99235376	-0.62242085
2012	-0.01241027	-0.97829980	-0.99235376	-0.62242085
2013	-0.01241027	-0.97829980	-0.99235376	-0.62242085
2014	-0.01241027	-0.97829980	-0.99235376	-0.62242085
2015	-0.01241027	-0.97829980	-0.99235376	-0.62242085
2016	-0.01241027	-0.97829980	-0.99235376	-0.62242085
2017	-0.01241027	-0.97829980	-0.99235376	-0.62242085
2018	-0.01241027	-0.97829980	-0.99235376	-0.62242085
2019	-0.01241027	-0.97829980	-0.99235376	-0.62242085
2020				

```
> plot(Inf_seriescomp)
```

Figure 2.21: Decomposition of Inflation Time series plot

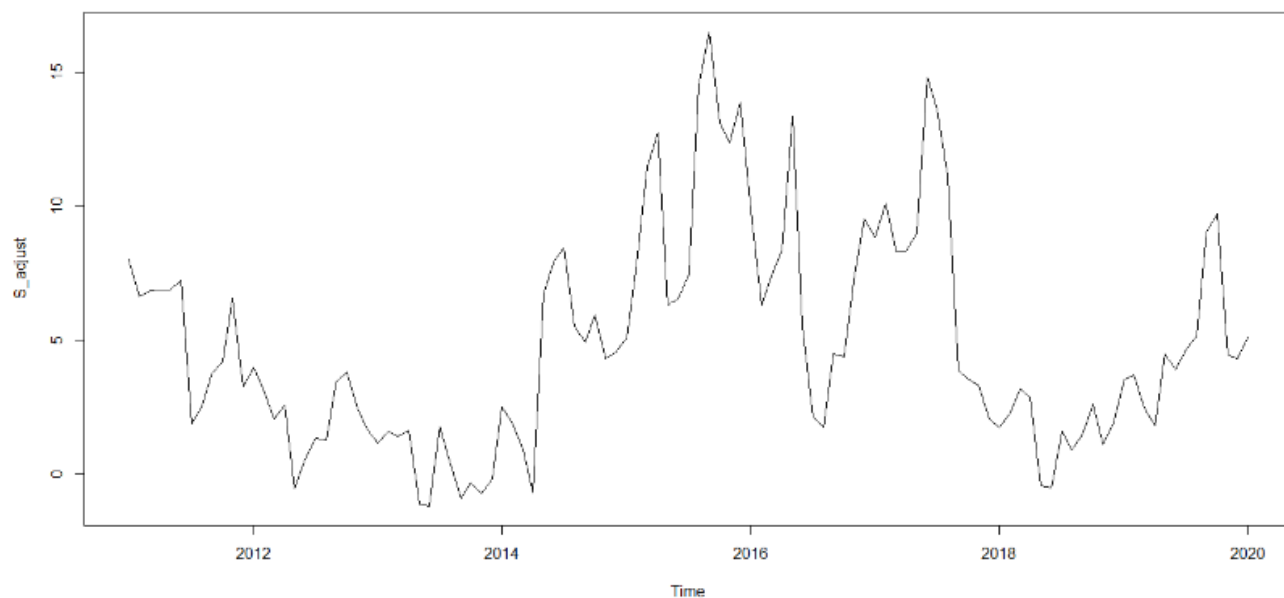


The above plot is sectioned into four reveals from the top the main time series, the trend component, the seasonal component, and the irregular component. We see that the trend component shows a minute downfall from about 6 in 2011 to about 0.5 in 2013, after which it gradually increased from then on to about 11 in 2015.

Seasonally Adjusting: This is done to eliminate the seasonality in the time series

```
#Seasonally adjusting
s_adjust <- Inf_series - Inf_seriescomp$seasonal
plot(s_adjust)
```

Figure 2.22: Seasonally Adjusted Inflation Time series plot



Forecasting: We will be running two forecasting models, Holt-winters and Arima. First, we need to run the “forecast” library and define the function.

```

#FORECASTING#
library("forecast")
#we need to define the function
plotForecastErrors <- function(forecasterrors)
{
  # make a histogram of the forecast errors:
  mybinsize <- IQR(forecasterrors)/4
  mysd <- sd(forecasterrors)
  mymin <- min(forecasterrors) - mysd*5
  mymax <- max(forecasterrors) + mysd*3
  # generate normally distributed data with mean 0 and standard deviation mysd
  mynorm <- rnorm(10000, mean=0, sd=mysd)
  mymin2 <- min(mynorm)
  mymax2 <- max(mynorm)
  if (mymin2 < mymin) {mymin <- mymin2}
  if (mymax2 > mymax) {mymax <- mymax2}
  #make a red histogram of the forecast errors, with the normally distributed data overlaid:
  mybins <- seq(mymin, mymax, mybinsize)
  hist(forecasterrors, col="red", freq=FALSE, breaks=mybins)
  # freq=FALSE ensures the area under the histogram = 1
  # generate normally distributed data with mean 0 and standard deviation mysd
  myhist <- hist(mynorm, plot=FALSE, breaks=mybins)
  # plot the normal curve as a blue line on top of the histogram of forecast errors:
  points(myhist$mids, myhist$density, type="l", col="blue", lwd=2)
}

```

Holt-Winters Smoothing(with trend and seasonality): This is done to make a brief forecast on an additive model with trend and seasonality.

```

> Inf_seriesforecasts <- Holtwinters(Inf_series)
> Inf_seriesforecasts
Holt-winters exponential smoothing with trend and additive seasonal component.

```

```

Call:
Holtwinters(x = Inf_series)

```

```

Smoothing parameters:
  alpha: 0.9018409
  beta : 0
  gamma: 1

```

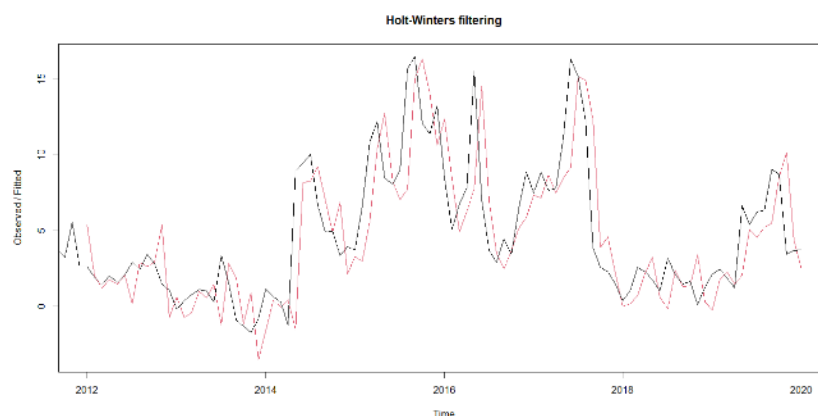
```

Coefficients:
      [,1]
a    3.3109418
b   -0.2777719
s1    0.2685919
s2    0.2847546
s3    0.2031173
s4    1.7460087
s5    0.5084580
s6    0.0855799
s7   -0.5243832
s8   -0.7514998
s9   -0.9106940
s10   0.1377433
s11   1.2539252
s12   0.4220345
> plot(Inf_seriesforecasts)

> Inf_seriesforecasts$SSE
[1] 845.3859

```

Figure 2.23: Holt-Winters Filtering of Inflation Time series plot

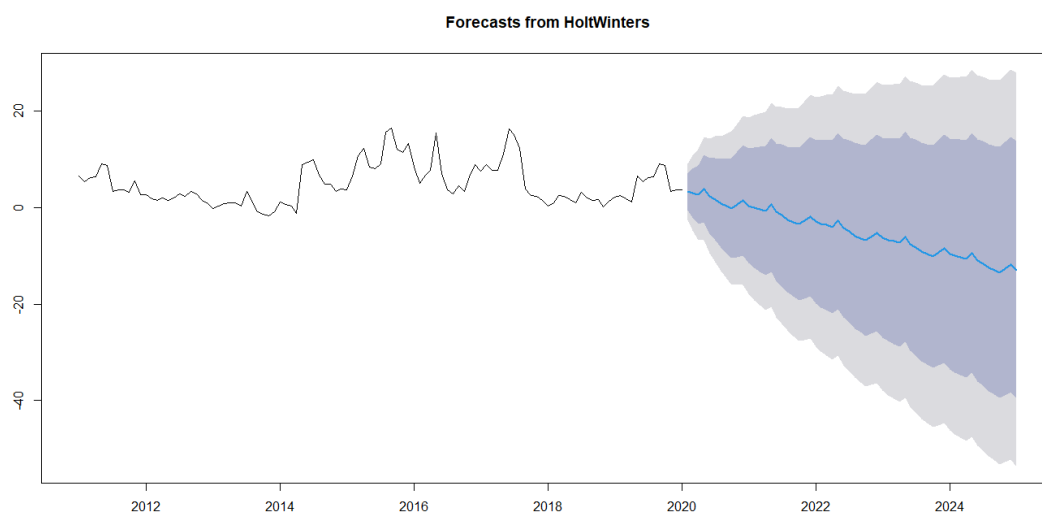


The values estimated reveal alpha as 0.90, beta as 0.00, and gamma as 1.00. Alpha value seems to be high meaning it is based on recent findings or observations while Beta at zero means there is no update on the trend component and left at its previous value. Lastly, Gamma has a very high value meaning it is also based on recent observations.

Let's plot for more months aside from the original data, from January 2021 to December 2025 i.e. 60 months.

```
Inf_seriesforecasts2 <- forecast(Inf_seriesforecasts, h=60)
plot(Inf_seriesforecasts2)
```

Figure 2.24: Forecasting from Holt-Winters of Inflation Time series



The gray and purple shaded areas show 95% and 80% prediction intervals respectively while the forecasts are displayed as a blue line. We will make a correlogram and carry out the Ljung-Box test to further investigate the improvement in the predictive model.

```
> #Carrying out Ljung-Box test and making a correlogram
> length(Inf_seriesforecasts2) #check the length of the time series and set as maximum lag
[1] 10
> acf(Inf_seriesforecasts2$residuals, lag.max = 10, na.action = na.pass)
> Box.test(Inf_seriesforecasts2$residuals, lag = 10, type="Ljung-Box")

Box-Ljung test

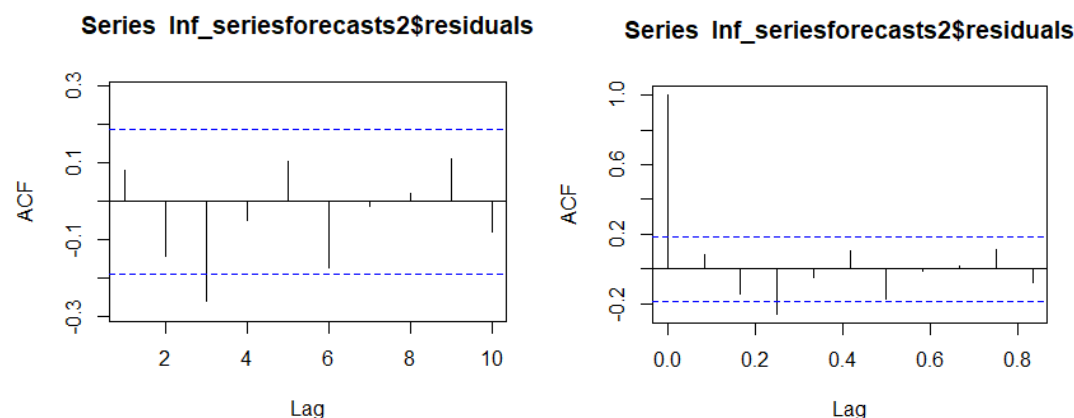
data: Inf_seriesforecasts2$residuals
X-squared = 16.016, df = 10, p-value = 0.09917

> #This was giving decimal result, lets try the capital Acf(capital A) in the package "forecast"
> Acf(Inf_seriesforecasts2$residuals, lag.max = 10, na.action = na.pass)
> Box.test(Inf_seriesforecasts2$residuals, lag = 10, type="Ljung-Box")

Box-Ljung test

data: Inf_seriesforecasts2$residuals
X-squared = 16.016, df = 10, p-value = 0.09917
```

Figure 2.25: Correlogram of Inflation Time series Forecasting and Residuals



The correlogram showed that the forecast error exceeded the significant bounds at lag 3 (plot without decimal) and lag 0.25. The p-value is 0.09 which is proof of non-zero autocorrelations at lags 1 -10.

Let's plot the time series of the residuals and also plot a histogram.

```
plot.ts(Inf_seriesforecasts2$residuals)
Inf_seriesforecasts2$residuals <- Inf_seriesforecasts2$residuals[
!is.na(Inf_seriesforecasts2$residuals)]
#plot the histogram
plotForecastErrors(Inf_seriesforecasts2$residuals)
```

Figure 2.26: Inflation Time series Residuals plot

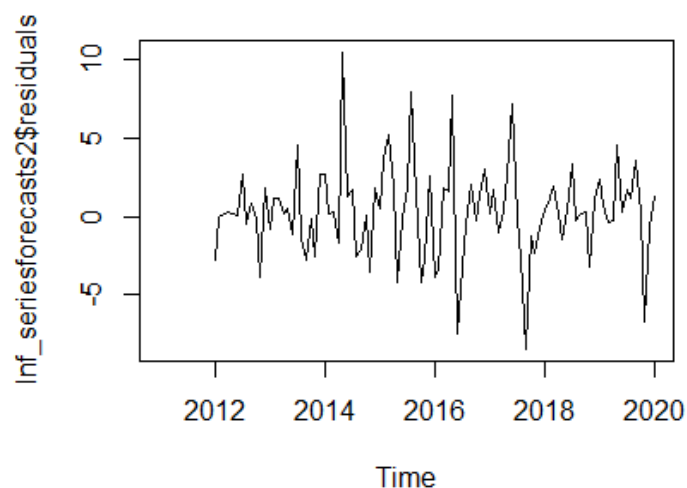
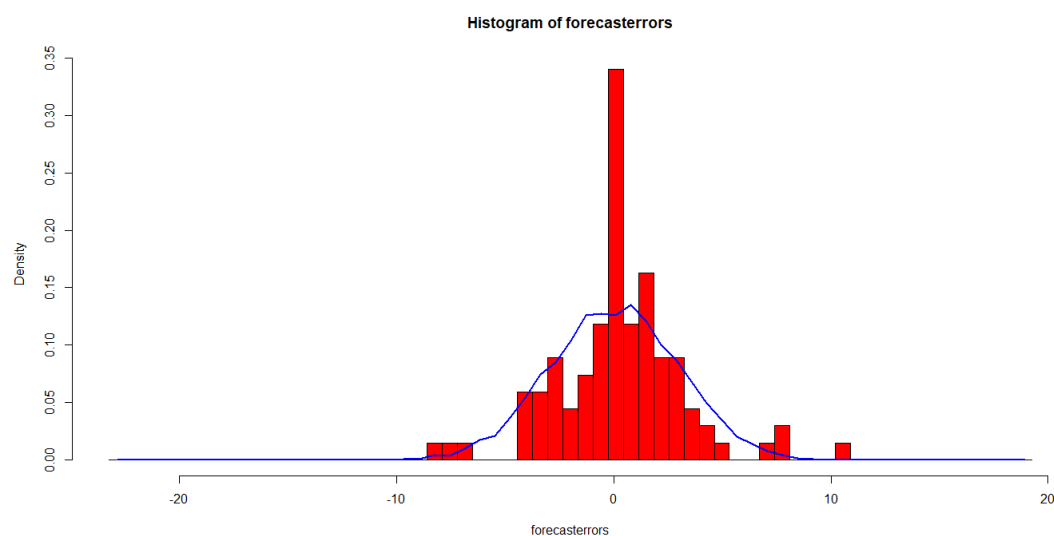


Figure 2.27: Histogram of Inflation Time series Forecast errors

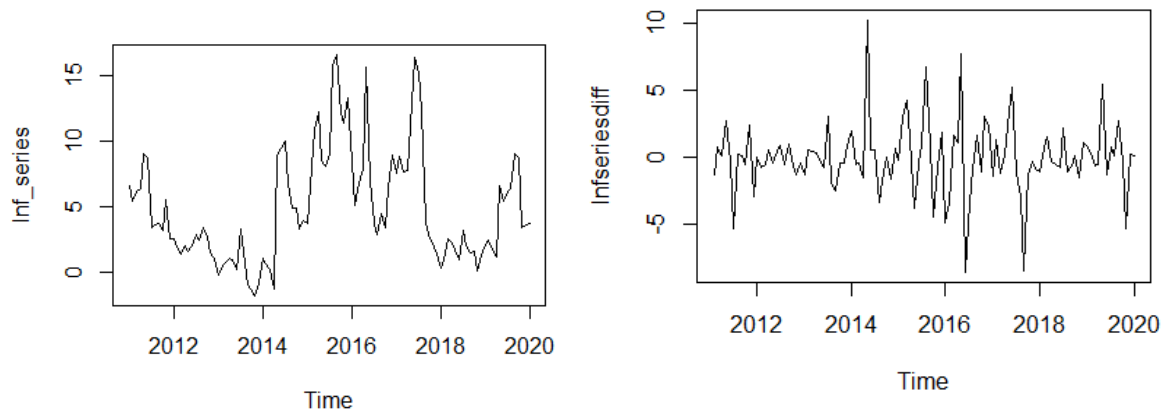


The time series plot shows that the forecast errors have constant variance over time while the histogram indicates that the distribution is normal with a mean value of zero(0).

Arima Model: To perform the Arima model we need to have a stationary time series else, we will perform differencing to obtain one. For Arima(p,d,q), the d is the value for the number of differencing performed before finding values for p and q.

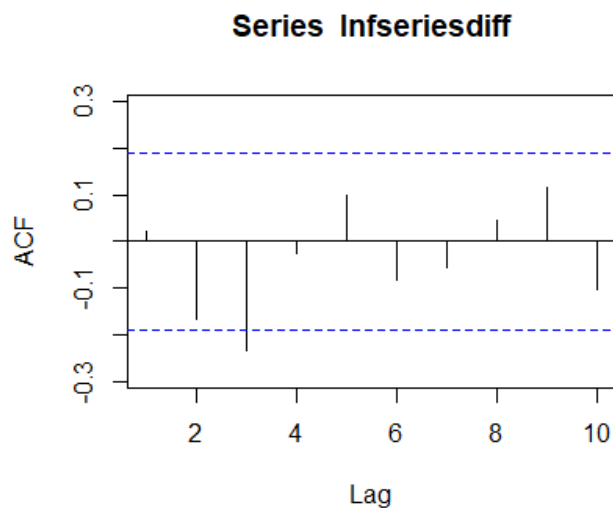
```
##ARIMA MODELS##
#Plot the time series to check if it is stationary
plot.ts(Inf_series)
Infseriesdiff <- diff(Inf_series, differences=1)
plot.ts(Infseriesdiff)
#plot a correlogram
Acf(Infseriesdiff, lag.max = 10)
```

Figure 2.28: Inflation Time series plot and Inflation time series differencing plot



The time series is stationary at the first differencing, so our model is ARIMA(p,1,q) left with getting values for p and q.

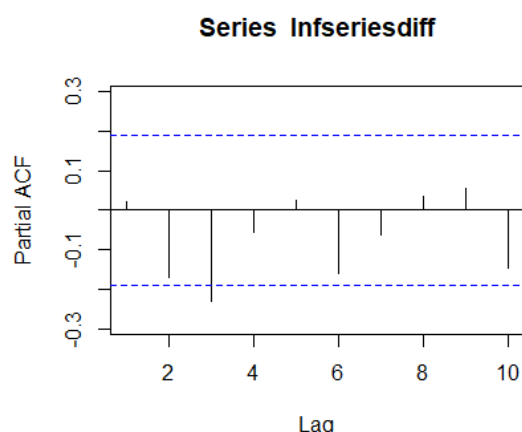
Figure 2.29: Correlogram of Inflation Time series Differencing



Let's get the autocorrelation values for the correlogram and let's plot for the partial correlogram as well.

```
> Acf(Infseriesdiff, lag.max = 10, plot=FALSE)
Autocorrelations of series 'Infseriesdiff', by lag
    0      1      2      3      4      5      6      7      8      9     10
1.000 0.022 -0.167 -0.232 -0.026 0.100 -0.081 -0.054 0.046 0.116 -0.104
> #Plot the partialcorrelation
> Pacf(Infseriesdiff, lag.max = 10)
> Pacf(Infseriesdiff, lag.max = 10, plot=FALSE)
Partial autocorrelations of series 'Infseriesdiff', by lag
    1      2      3      4      5      6      7      8      9     10
0.022 -0.168 -0.230 -0.055 0.024 -0.159 -0.064 0.035 0.056 -0.144
```

Figure 2.30: Partial Correlogram of Inflation Time series Forecasting Residuals



We can see from the correlogram that the autocorrelation at lag 3(-0.232) exceeds significance bounds and tends to shift towards zero after lag 4 while other autocorrelations between 0-3 and 3-10 did not exceed the bounds. The partial correlogram also exceeds the significance bounds at lag 3(-0.230) and tails off to zero(0) after lag 4 as well. We can say that our $p=1$ and $q=1$, the model for differencing will be ARIMA(1,1) while the ARIMA(p,d,q) model of the time series will be ARIMA(1,1,1) but let's find the appropriate model by using the `auto.Arima()` function (Coghlan, 2010).

```
> auto.arima(WDI_New$Inflation)
Series: WDI_New$Inflation
ARIMA(1,0,0) with non-zero mean

Coefficients:
      ar1      mean
      0.8180  4.7109
s.e.    0.0557  1.2970

sigma^2 = 6.195: log likelihood = -232.63
AIC=471.25  AICC=471.5  BIC=479.07
>
> auto.arima(Infseriesdiff)
Series: Infseriesdiff
ARIMA(0,0,0) with zero mean

sigma^2 = 6.761: log likelihood = -256.45
AIC=514.89  AICC=514.93  BIC=517.58
```

For the original data with non-zero mean, the model is ARIMA(1,0,0) while with differencing of one shows ARIMA(0,0,0).

Let's fit our ARIMA(1,1,1) into the time series and forecast.

```
> Inf_seriesArima <- arima(Inf_series, order=c(1,1,1))
> Inf_seriesArima

call:
arima(x = Inf_series, order = c(1, 1, 1))

Coefficients:
      ar1      ma1
      -0.8751  0.9253
s.e.    0.1018  0.0744

sigma^2 estimated as 6.685: log likelihood = -255.87, aic = 517.74
> Inf_seriesforecasts3 <- forecast(Inf_seriesArima, h=12)
> Inf_seriesforecasts3
      Point Forecast      Lo 80      Hi 80      Lo 95      Hi 95
Feb 2020    3.791319    0.4778816    7.104757   -1.276146    8.858785
Mar 2020    3.740263   -1.0647819    8.545308   -3.608419   11.088945
Apr 2020    3.784943   -2.0636057    9.633491   -5.159640   12.729525
May 2020    3.745843   -3.0503928   10.542079   -6.648103   14.139789
Jun 2020    3.780059   -3.7970311   11.357150   -7.808100   15.368219
Jul 2020    3.750117   -4.5745063   12.074740   -8.981296   16.481529
Aug 2020    3.776320   -5.2016900   12.754329   -9.954361   17.507001
Sep 2020    3.753389   -5.8600520   13.366830  -10.949100   18.455879
Oct 2020    3.773456   -6.4141069   13.961018  -11.807077   19.353988
Nov 2020    3.755895   -6.9932409   14.505032  -12.683490   20.195281
Dec 2020    3.771263   -7.4964070   15.038932  -13.461151   21.003676
Jan 2021    3.757815   -8.0182205   15.533850  -14.252077   21.767706
> #Plot the forecasts
> plot(Inf_seriesforecasts3)
> Acf(Inf_seriesforecasts3$residuals, lag.max = 10)
> Box.test(Inf_seriesforecasts3$residuals, lag=10, type="Ljung-Box")
```



```
Box-Ljung test

data: Inf_seriesforecasts3$residuals
X-squared = 13.533, df = 10, p-value = 0.1954

> #Time plot forecast error
> plot.ts(Inf_seriesforecasts3$residuals)
> #Make a histogram
> plotForecastErrors(Inf_seriesforecasts3$residuals)
```

Figure 2.31: Arima forecast error of Inflation time series forecast residuals

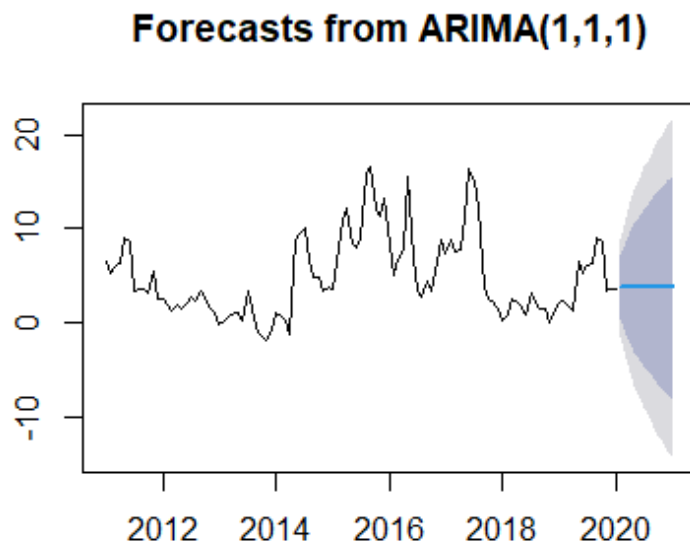
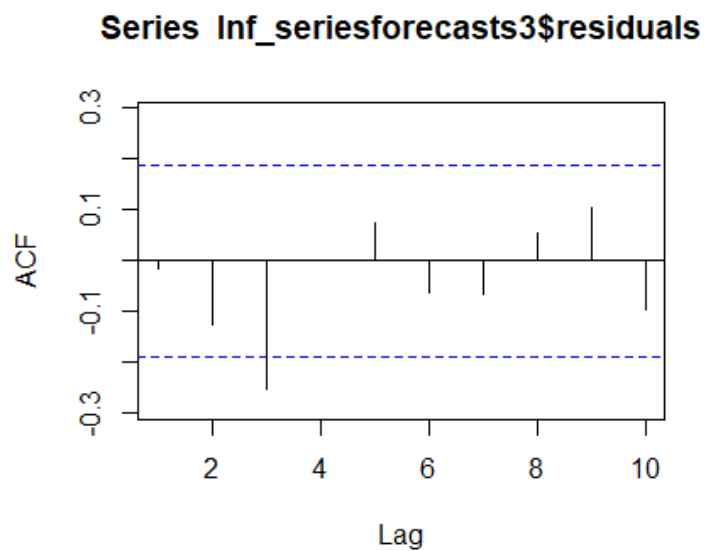


Figure 2.32: Correlogram of Inflation Time series Forecasting Residuals



The autocorrelation moves towards zero(0) after lag 3 which exceeds the significance bounds.

Figure 2.33: Inflation Time series Forecasting Residuals plot

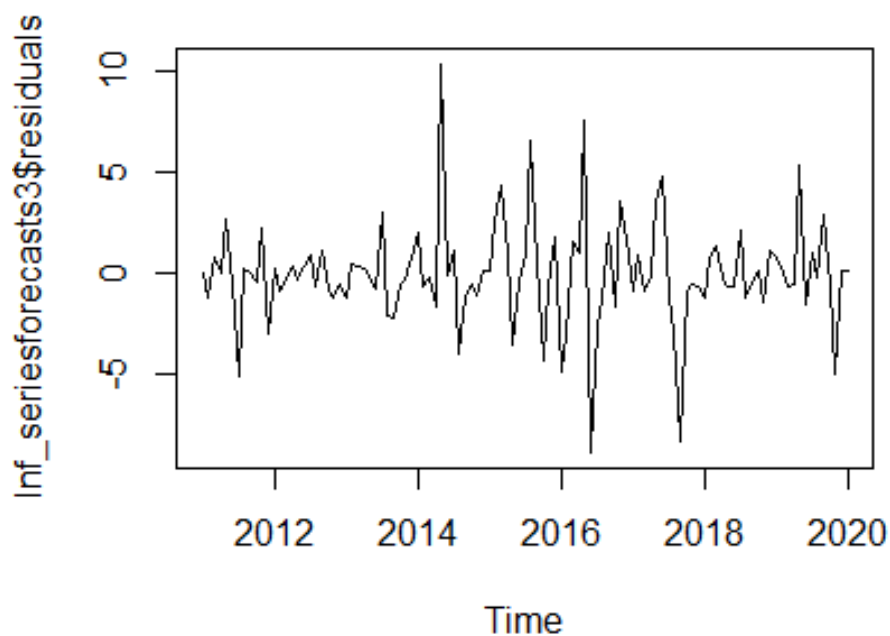
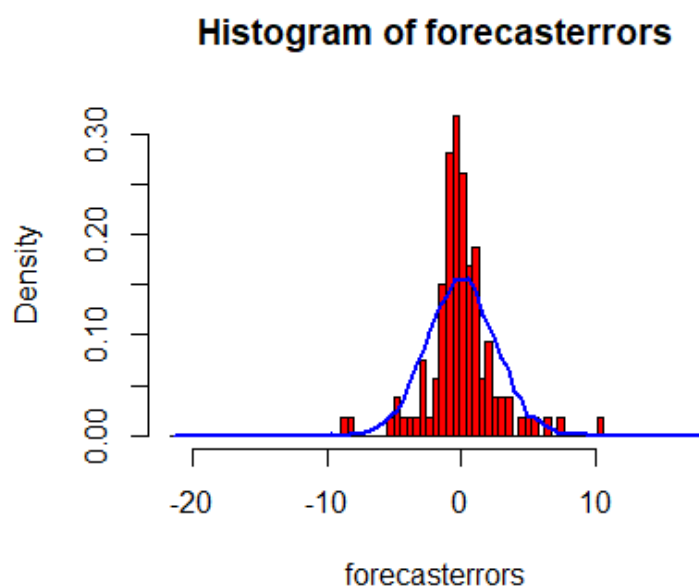


Figure 2.34: Histogram of Inflation Time series Forecast errors



4.5. Hypothesis Testing

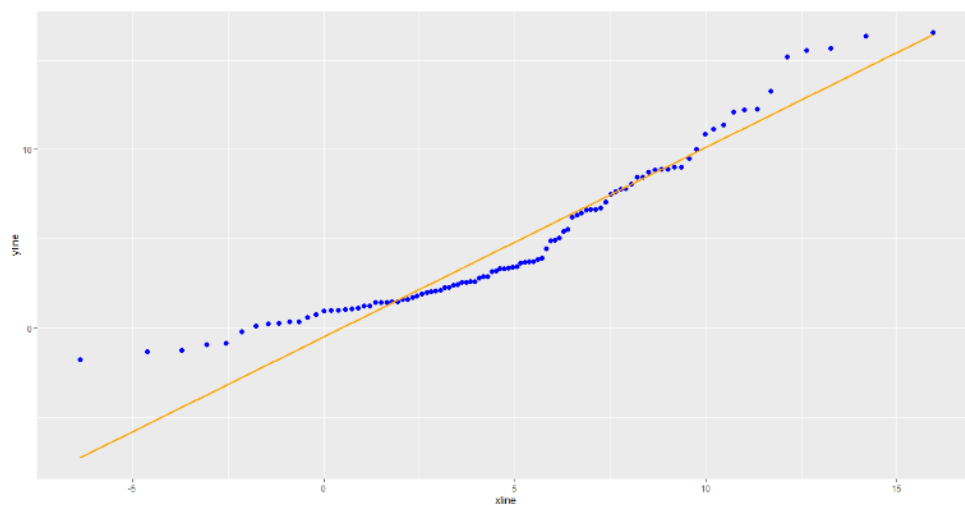
For us to carry out hypothesis testing, we will be making use of the following packages; “datarium”, “qqplotr” and “ggplot2” which will be used in assessing the normality of the dataset.

```
###ASSESSING NORMALITY OF THE DATASET
install.packages("datarium")
install.packages("qqplotr")
install.packages("ggplot2")
library("datarium")
library("qqplotr")
library("ggplot2")
```

Visualizing with Q-Q plot

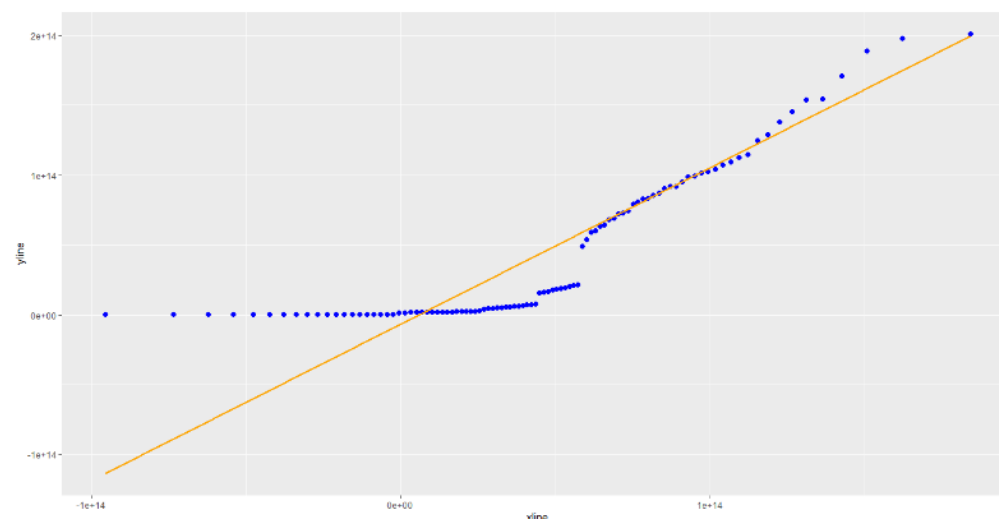
```
ggplot(mapping=aes(sample=WDI_New$Inflation)) +  
  stat_qq_point(size=2,color="blue") +  
  stat_qq_line(color="orange")
```

Figure 2.35a: Q-plot of Inflation



```
ggplot(mapping=aes(sample=WDI_New$GDP)) +  
  stat_qq_point(size=2,color="blue") +  
  stat_qq_line(color="orange")
```

Figure 2.35b: Q-plot of GDP



The plots reveal that the curves are not normally distributed.

Shapiro-Wilk and Kolmogorov-Smirnov Test

```
> ##Using Shapiro-wilk test and Kolmogorov-Smirnov test to access the dataset  
> shapiro.test(WDI_New$Inflation)
```

shapiro-wilk normality test

```
data: WDI_New$Inflation  
W = 0.91256, p-value = 5.795e-06
```

```

> shapiro.test(WDI_New$GDP)

      Shapiro-Wilk normality test

data:  WDI_New$GDP
W = 0.79319, p-value = 1.538e-10

>
> ##Using kolmogorov-Smirnov test for the distribution
> ks.test(WDI_New$Inflation,'pnorm')

      Asymptotic one-sample Kolmogorov-Smirnov test

data:  WDI_New$Inflation
D = 0.71464, p-value < 2.2e-16
alternative hypothesis: two-sided

> ks.test(WDI_New$GDP,'pnorm')

      Asymptotic one-sample Kolmogorov-Smirnov test

data:  WDI_New$GDP
D = 1, p-value < 2.2e-16
alternative hypothesis: two-sided

```

The P-value of both tests is less than 0.05 which indicates that the data is not normally distributed.

Normalizing the Dataset: There are three different methods to normalize our dataset, the log transformation, the square root transformation, and the cube root transformation (Bobbitt, 2020).

```

> ##NORMALIZING THE DATASET
> ##1. Log Transformation
> log_Inflation <- log10(WDI_New$Inflation)
warning message:
NaNs produced
> #histogram for both distribution
> ##original distribution
> hist(WDI_New$Inflation, col='steelblue', main='original')
>
> ##log-transformed distribution
> hist(log_Inflation, col='coral2', main='Log Transformed')
> shapiro.test(log_Inflation)

      Shapiro-Wilk normality test

data:  log_Inflation
W = 0.95474, p-value = 0.00255

```

After running the code for log transformation, NaNs were produced but we still need to visualize the distribution with a histogram and run either Shapiro-Wilk or Kolmogorov normality test.

Figure 2.36a: Histogram of Original Inflation distribution

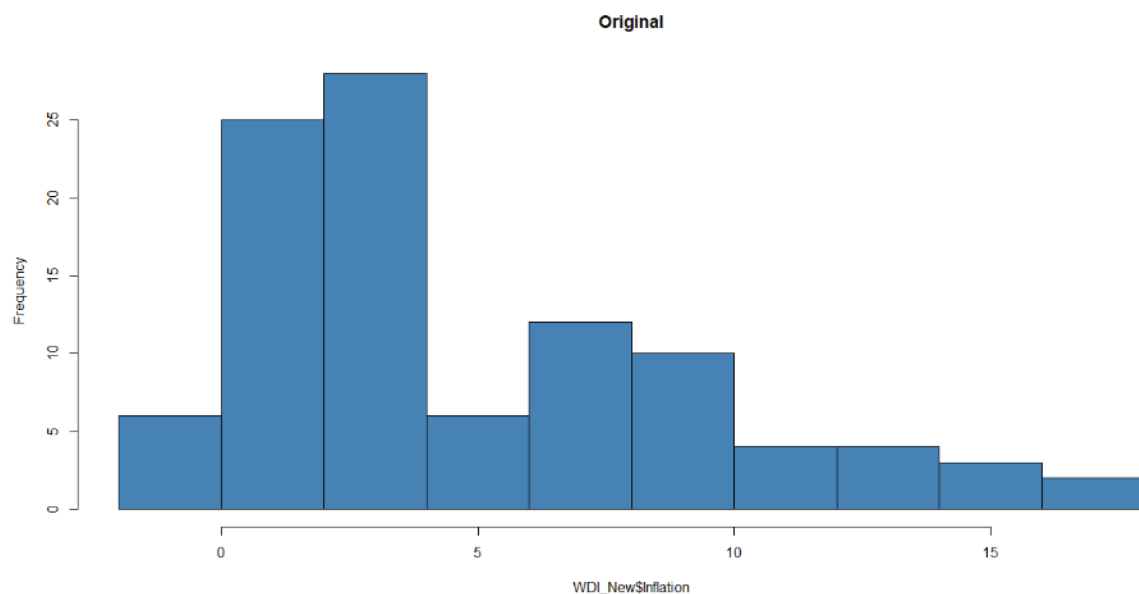
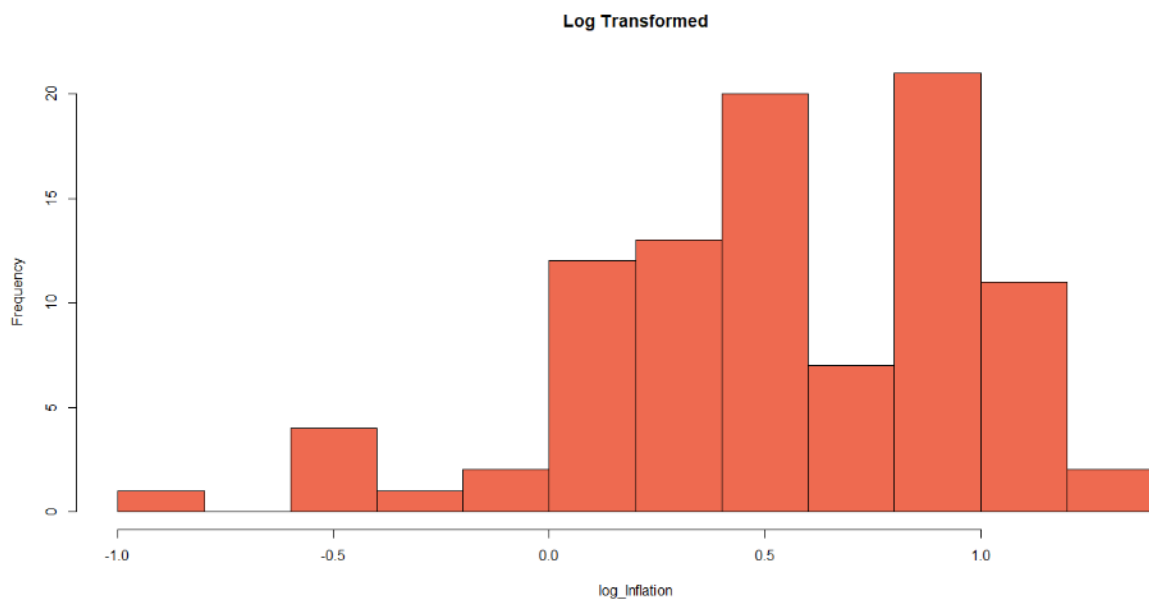


Figure 2.36b: Histogram of Log transformed Inflation distribution



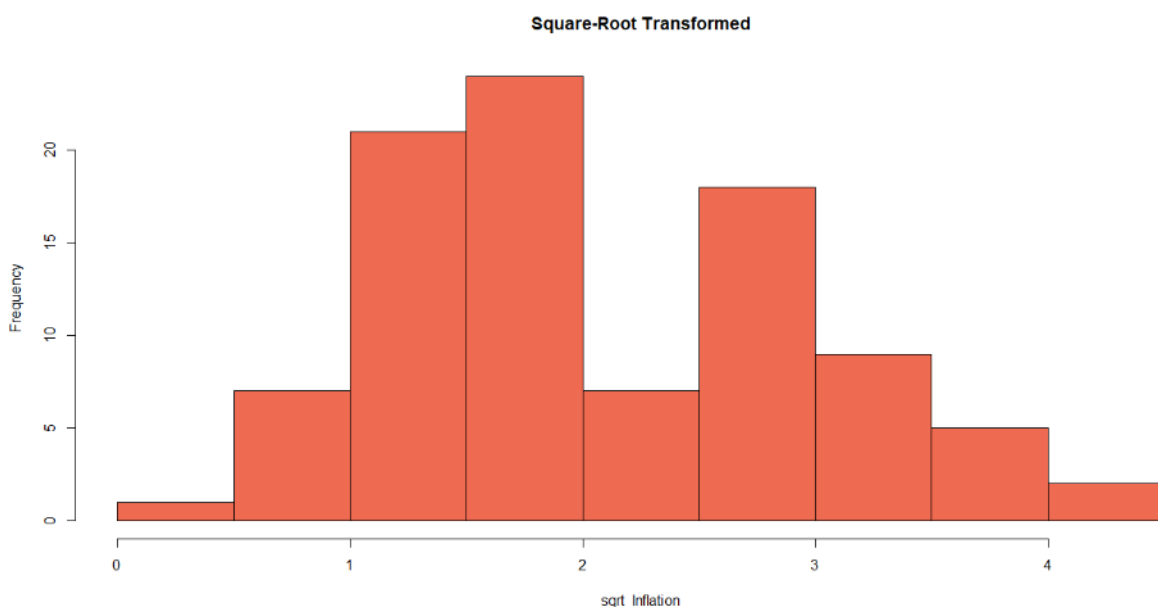
The p-value is less than 0.05, so the distribution is not normalized yet.

```
> ##2. Square Root Transformation
> sqrt_Inflation <- sqrt(WDI_New$Inflation)
Warning message:
In sqrt(WDI_New$Inflation) : NaNs produced
> #histogram for both distribution
> ##original distribution
> hist(WDI_New$Inflation, col='steelblue', main='Original')
> 
> ##Square-root transformed distribution
> hist(sqrt_Inflation, col='coral2', main='Square-Root Transformed')
> shapiro.test(sqrt_Inflation)
```

shapiro-wilk normality test

data: sqrt_Inflation
W = 0.96767, p-value = 0.01987

Figure 2.36c: Histogram of Square-root transformed Inflation distribution



NANs were produced as well after running the codes for square-root transformation. The p-value for the Shapiro-Wilk test is 0.01987 which is also less than 0.05; the distribution is not normalized yet.

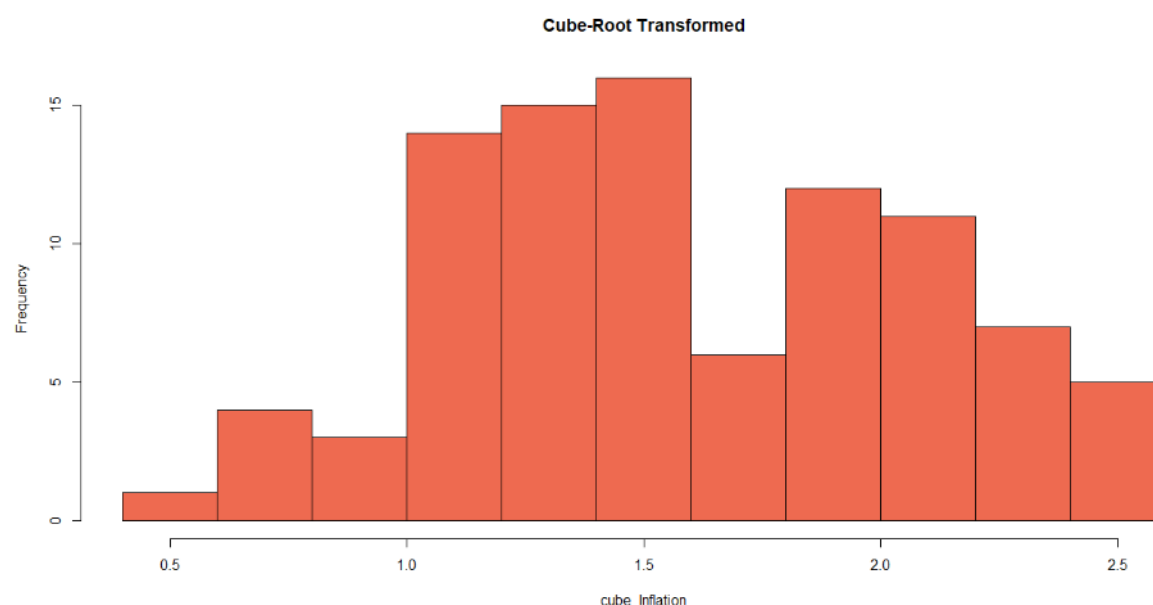
```
> ##3. Cube Root Transformation
> cube_Inflation <- WDI_New$Inflation^(1/3)
> hist(WDI_New$Inflation, col='steelblue', main='Original')
> hist(cube_Inflation, col='coral2', main='Cube-Root Transformed')
> shapiro.test(cube_Inflation)
```

shapiro-wilk normality test

data: cube_Inflation
W = 0.98033, p-value = 0.169

The cube root transformation method eventually normalized the distribution with a p-value of 0.169 which is more than 0.05 and there was no NANs produced.

Figure 2.36d: Histogram of Cube-root transformed Inflation distribution



Now let's insert our normalized variable(Inflation) into the dataset and carry out our hypothesis testing.

One Sample T-test

The Null hypothesis(H_0): Less developed countries have the highest inflation rate

The Alternative hypothesis(H_1): Less developed countries do not have the highest inflation rate

Common p-value: $p < 0.05$

```
> ##ONE SAMPLE T-TEST
> t.test(WDI_New2$Inflation)
```

One sample t-test

data: WDI_New2\$Inflation
t = 31.409, df = 93, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
1.488553 1.689480
sample estimates:
mean of x
1.589016

The p-value is less than 0.05 hence we will reject the null hypothesis

Independent Two Sample T-test

To carry out this test, we need to have a predictor variable that must be categorical hence, we will create a new column for two levels (more developed and less developed countries).

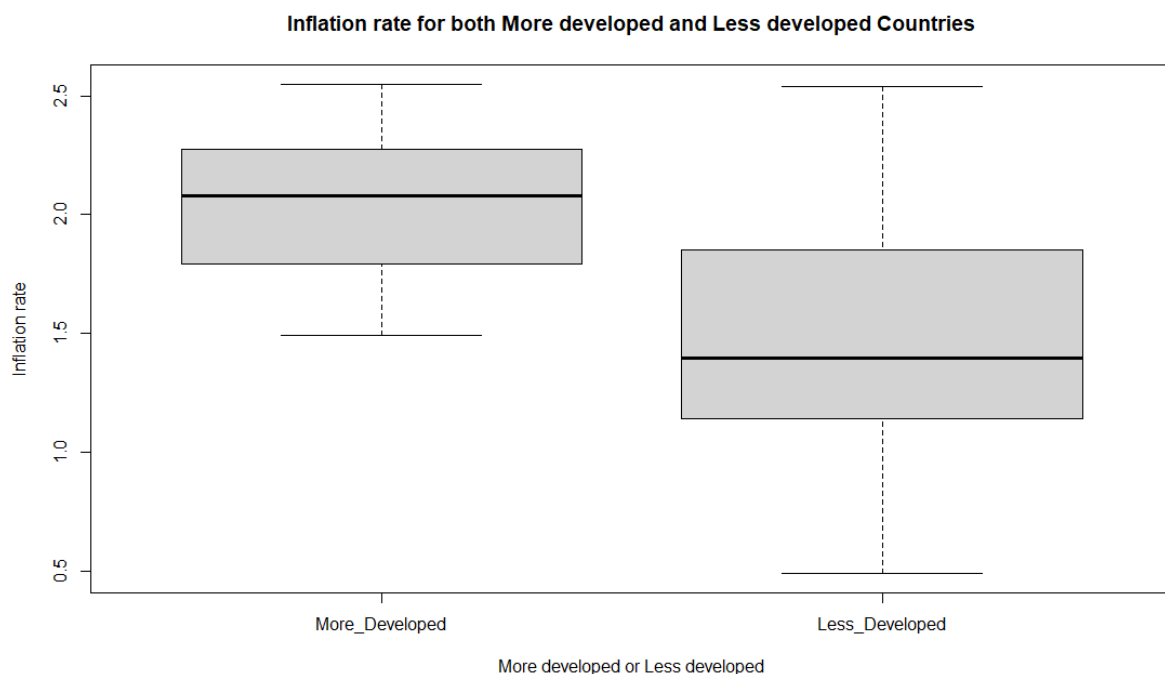
```

> ###convert country to two levels of more developed and less developed countries
> ###Then create a new column with the country code column using "sapply" and "switch"
  function
> WDI_New2$Country_Status <- sapply(WDI_New2$`C-Code`, switch, "GBR" = 'More_Develope
d',
+                               "USA" = 'More_Developed', "BRA" = 'More_Developed',
+                               "CHN" = 'More_Developed', "GRC" = 'More_Developed',
+                               "FIN" = 'More_Developed', "RUS" = 'More_Developed',
+                               "TUR" = 'More_Developed', "NGA" = 'Less_Developed',
+                               "IND" = 'Less_Developed')
> str(WDI_New2$Country_Status)
  Named chr [1:100] "More_Developed" "More_Developed" "More_Developed" ...
  - attr(*, "names")= chr [1:100] "BRA" "BRA" "BRA" "BRA" ...

> WDI_New2$Country_Status <- as.factor(WDI_New2$Country_Status)
> ##Using box plot to compare More developed and Less developed
> boxplot(Inflation~`Country_Status`,data=WDI_New2, names=c("More_Developed", "Less_Deve
loped"),
+         xlab = "More developed or Less developed", ylab= "Inflation rate",
+         main = "Inflation rate for both More developed and Less developed Countries")
~ |

```

Figure 2.37: Boxplot of Inflation rate for More developed and Less developed Countries



The boxplot revealed that more developed countries are more than less developed countries in the distribution.

Let's perform the test using the same function by passing this argument 'Inflation~Country_Status'.

```

##Using two sample t-test using the same function
t.test(Inflation~`Country_Status`,WDI_New2)

      welch Two sample t-test

data:  Inflation by Country_Status
t = 6.4807, df = 44.056, p-value = 6.588e-08
alternative hypothesis: true difference in means between group Less_Developed and group
More_Developed is not equal to 0
95 percent confidence interval:
 0.3918633 0.7455720
sample estimates:
mean in group Less_Developed mean in group More_Developed
      2.036730              1.468013

```

The p-value is less than 0.05, so we will reject the null hypothesis.

Let's perform a one-tail test on the distribution

```

> ##Using a one tail test
> t.test(Inflation~Country_Status`,WDI_New2, alternative="less")

Welch Two Sample t-test

data: Inflation by Country_Status
t = 6.4807, df = 44.056, p-value = 1
alternative hypothesis: true difference in means between group Less_Developed and group
More_Developed is less than 0
95 percent confidence interval:
-Inf 0.716164
sample estimates:
mean in group Less_Developed mean in group More_Developed
2.036730 1.468013

```

The p-value is 1 which is more than 0.05, we will not reject the null hypothesis.

Apart from running a one-tail test that made the null hypothesis acceptable, other t-tests did not make the assumptions acceptable hence, we will run a non-parametric test alternative to t-tests.

```

##Non-parametric alternatives to T-tests
##Using Mann-Whitney
##Let's visually check for normality
M_developed <- WDI_New2$Inflation[WDI_New2$Country_Status=="More_Developed"]

ggplot(mapping = aes(sample = M_developed)) +
  stat_qq_point(size=2,color="blue") +
  stat_qq_line(color="orange") +
  xlab("Theoretical") + ylab("Sample")

options(scipen=999)
hist(M_developed)

L_developed <- WDI_New2$Inflation[WDI_New2$Country_Status=="Less_Developed"]

ggplot(mapping = aes(sample = L_developed)) +
  stat_qq_point(size=2,color="blue") +
  stat_qq_line(color="orange") +
  xlab("Theoretical") + ylab("Sample")

hist(L_developed)

```

Figure 2.38a: Q-plot of More developed countries

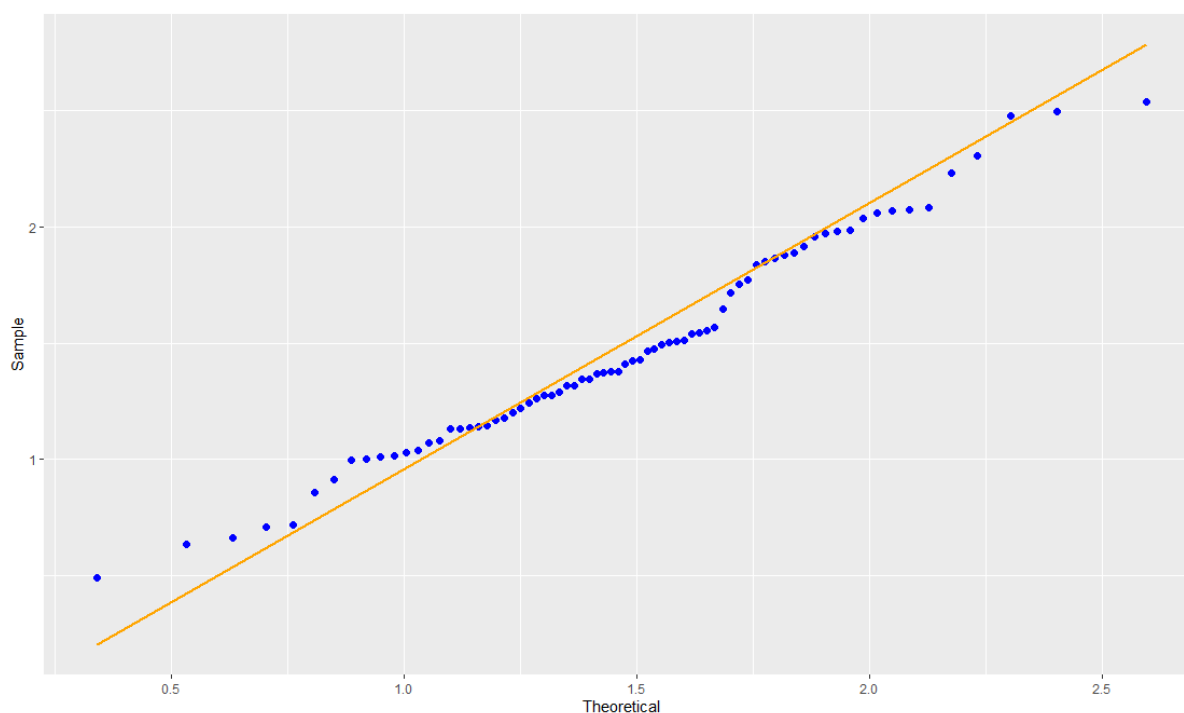


Figure 2.38b: Q-plot of Less developed countries

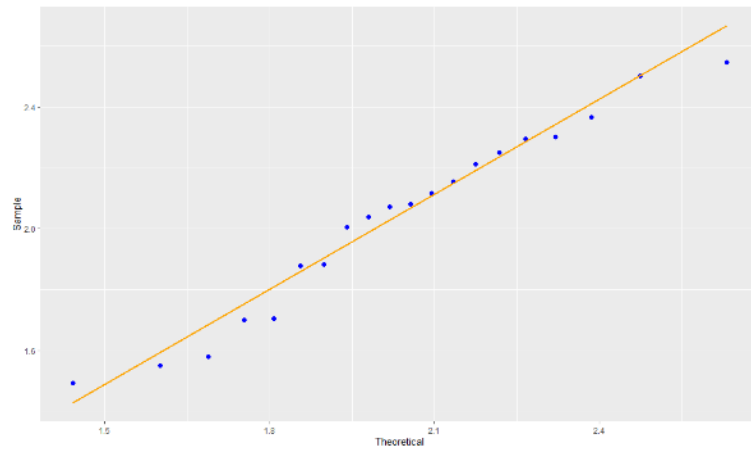


Figure 2.39a: Histogram of More developed countries

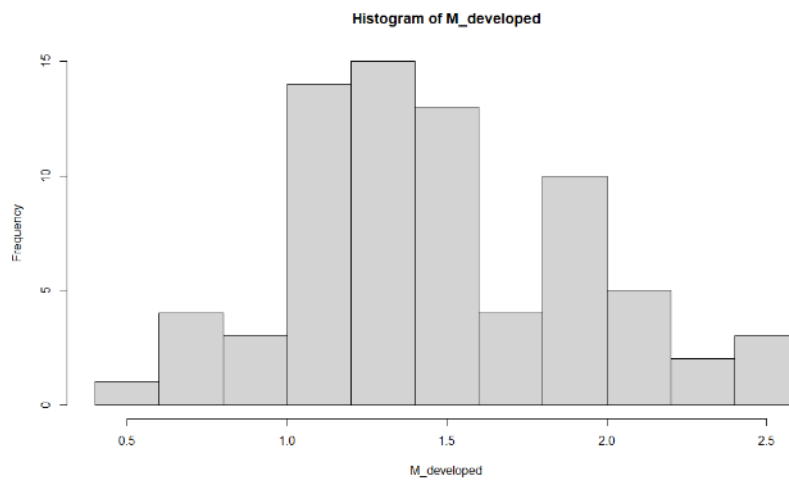
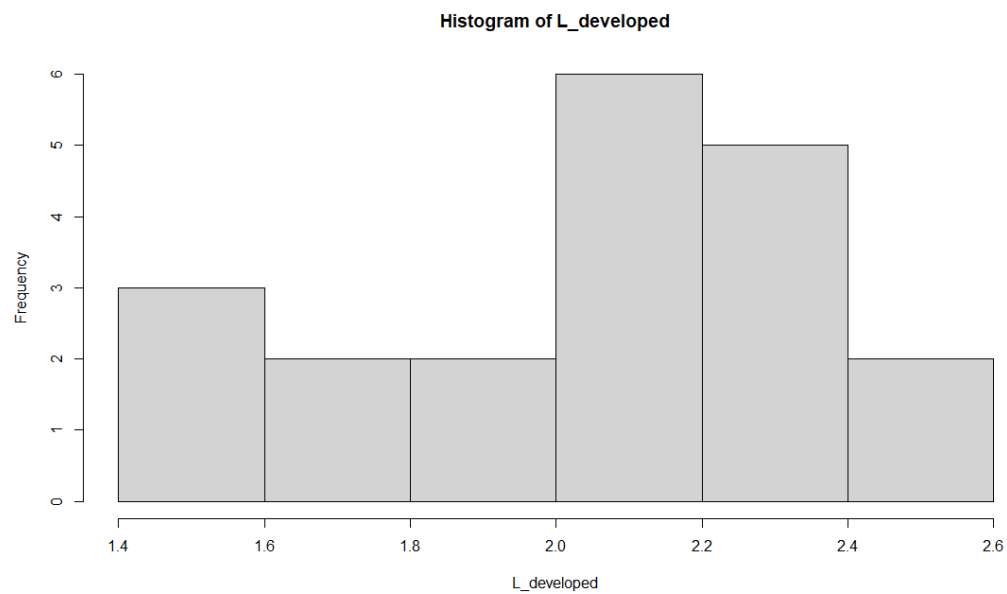


Figure 2.39b: Histogram of Less developed countries



The q-plots and histogram for more developed and less developed countries showed that the new column is not normally distributed.

```
##Running the hypothesis test
wilcox.test(Inflation~`Country_Status`,WDI_New2)

      wilcoxon rank sum test with continuity correction

data:  Inflation by Country_Status
W = 1257, p-value = 0.000001827
alternative hypothesis: true location shift is not equal to 0
```

We will still reject the null hypothesis that the inflation rate is higher in less developed countries.

5. DISCUSSIONS

The research showed that most of the indicators selected are positively and negatively correlated to each other. However, we were able to deduce our regression line for GDP with Inflation, Imports and Export being the predictor variables. Although they can only predict 91% variability in GDP but it is also on the high side and government can focus on these indicators for economic growth.

On the other end, our Time series analysis could forecast constant inflation rate for all the countries for the period of 10 years but it may not be absolute because it could only capture the information we have on our dataset.

Despite using a few method of hypothesis testing on our dataset even after using the cube-root transformation for normalization, our null hypothesis was still rejected with the exemption of the one-tailed test performed on the distribution. This shows that our assumptions are not accurate and our models is not the best fit to analyze the dataset.

6. CONCLUSIONS

Based on the objective of this analysis, we can conclude that the Multiple Linear Regression Model is the best fit that could explain and compare the interactions between the indicators. We also discovered that self-employment which could be in any aspect can single-handedly boost the GDP of any country which we can classify under the Simple Linear Regression equation.

In recent years, most countries have started encouraging start-ups which is insight to a better economy. With this, the level of unemployment will be reduced, production of goods and services will increase and this will at the long-run boost the revenue of the economy.

REFERENCES AND APPENDICES

- Bewick V, C. L., Ball J. (2003). Correlation and regression. *Statistics review* 7.
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC374386/#:~:text=The%20most%20commonly%20used%20techniques,the%20form%20of%20an%20equation>. (National Library of Medicine)
- Bobbitt, Z. (2020). How to Transform Data in R (Log, Square Root, Cube Root). *Statology*.
<https://www.statology.org/transform-data-in-r/>
- Coghlan, A. (2010). *A Little Book of R for Time Series!* <https://a-little-book-of-r-for-time-series.readthedocs.io/en/latest/>
- Gorton, D. (Updated November 30, 2022). Taxes Definition: Types, Who Pays and Why.
<https://www.investopedia.com/terms/t/taxes.asp>
- Group, S. F. (2003-2022). Time Series Analysis. <https://www.tableau.com/learn/articles/time-series-analysis>
- Group, T. W. B. D. (2022). *World Development Indicators*. <https://databank.worldbank.org/source/world-development-indicators>
- Learning, S. E. (2022). How to replace NA values in columns of an R data frame form the mean of that column. <https://www.tutorialspoint.com/how-to-replace-na-values-in-columns-of-an-r-data-frame-form-the-mean-of-that-column> (Tutorialspoint)
- Mauro, P., Romeu, R., Binder, A., & Zaman, A. (2015). . (2015). A mordern history of fiscal prudence and profligacy. *Monetary Economics*, 76, 55-70.
<https://www.imf.org/external/datamapper/datasets/FPP>
- Oner, C. (2022). Inflation: Prices On The Rise [Finance & Development]. *Back to Basics*.
<https://www.imf.org/en/Publications/fandd/issues/Series/Back-to-Basics/Inflation#:~:text=Inflation%20is%20the%20rate%20of,of%20living%20in%20a%20country>
- Roser, E. O.-O. a. M., & (2016). Government Spending [OWIDGovernmentSpending]. *Our World In Data*.
<https://ourworldindata.org/government-spending>
- Team, Y.-F. (2022). The History of BI Dashboards. <https://www.yellowfinbi.com/blog/history-of-bi-dashboards>
- Webster, R. (2015). *Selecting and removing rows in R dataframes*.
<https://www.youtube.com/watch?v=KXSPxijS8Fc>
- Wexler, S. (2022). *The Big Book of Dashboards* (Wiley, Ed.). <https://www.bigbookofdashboards.com/>
- Wikipedia. (2022a). *Gross Domestic Product*. https://en.wikipedia.org/wiki/Gross_domestic_product
- Wikipedia. (2022b). *Microsoft Power Bi*. Wikipedia, the free encyclopedia.
https://en.wikipedia.org/wiki/Microsoft_Power_Bi#:~:text=It%20was%20originally%20designed%20by,Power%20BI%20for%20Office%20365