**Assessment**

**Applied Statistics and Data Visualisation**

**MSc Data Science**

Unveiling Economic Growth Patterns:

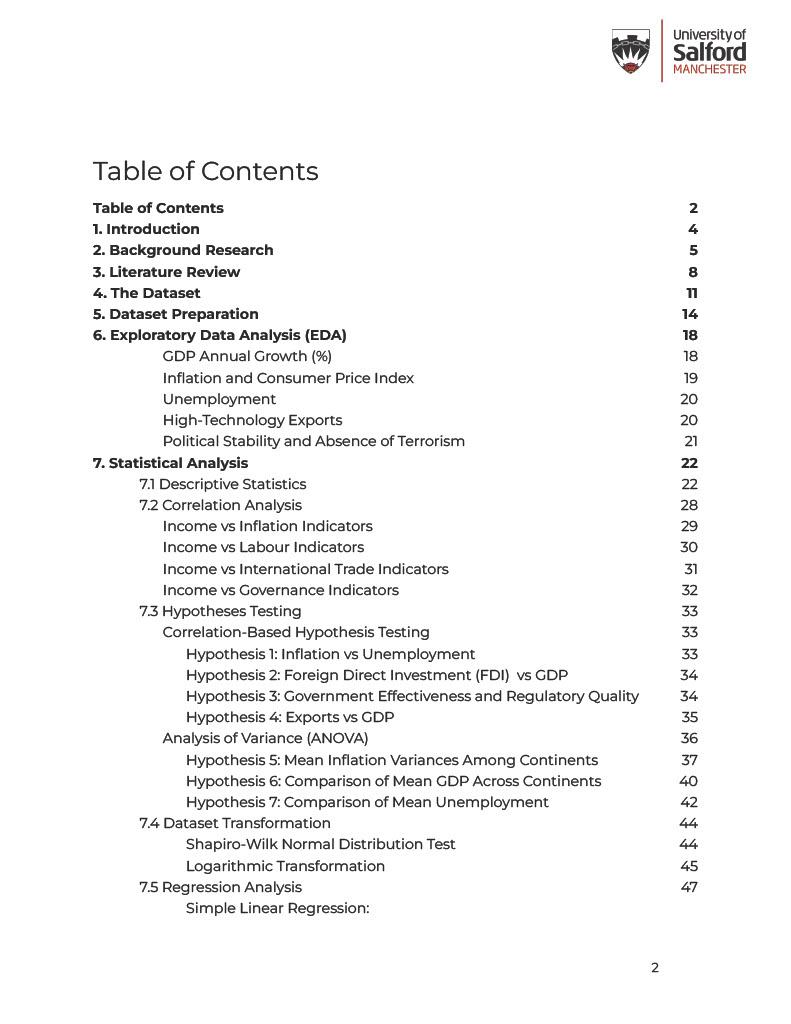
A Statistical and Visual Exploration of Global Trends

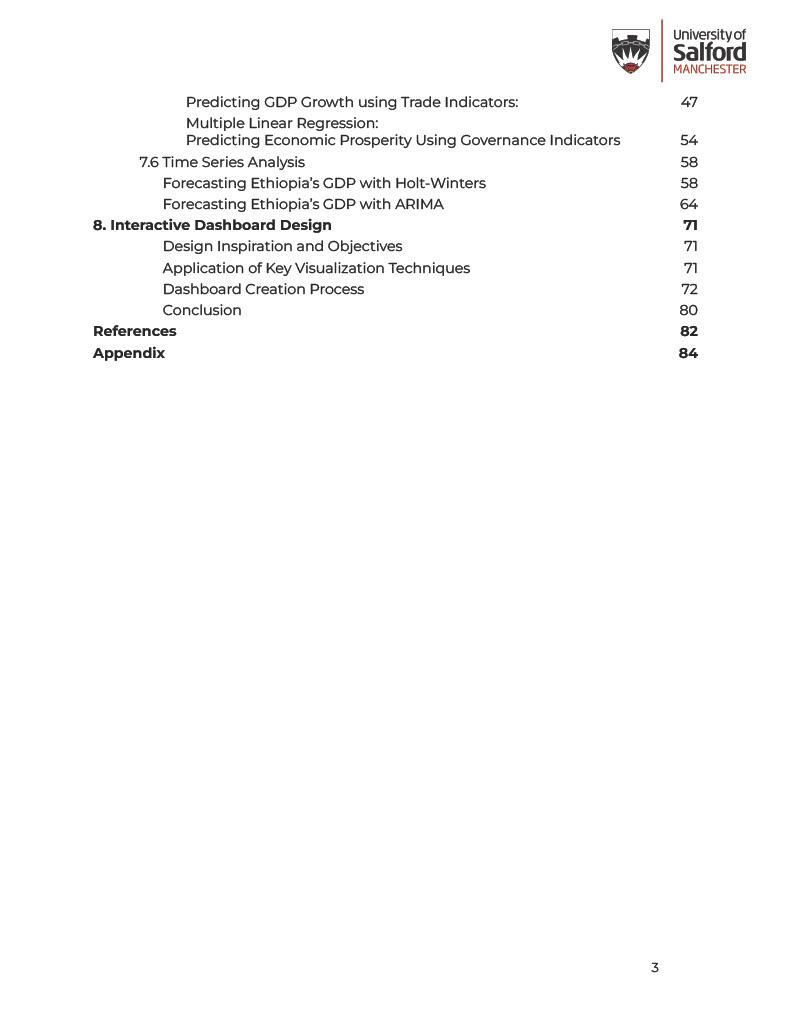
**Shola Ayeotan**

Student ID: @00607855

5th December, 2023

# 





# 1. **Introduction**

The interconnection between economic development and social progress is undeniable, and it is crucial for organisations looking towards global success to clearly understand these various economic indicators and their trends over the years.

As a Data Scientist at a non-governmental organisation, I have been working on a project that involves conducting a comprehensive statistical analysis and interactive dashboard development based on the World Bank's World Development Indicators (WDI) dataset.

This dataset spanning from 2008 to 2022 provides a wealth of information on key economic indicators across different countries.

Highlighted below are the objectives of the study:

**Objective 1:** Investigate the impact of Governance on Economic Prosperity

* What exact impact do indicators such as Political Stability, Corruption Control, Government Effectiveness and Rule of Law have on GDP per Capita?
* How do these factors collectively influence a nation's economic prosperity?
* What are even the interconnections between these indices?

These are just a snapshot of some research questions we will be seeking answers to in this study.

**Objective 2:** Explore the interconnectedness of Trade Balance and Economic Health. Our second objective for this study is to investigate the intricate relationship between Exports, Imports and GDP growth.

**Objective 3:** Our third objective focuses on unravelling the relationship between the labour force, unemployment rates, and GDP growth across continents. What role do labour market conditions play in shaping a country's economic performance, and how has this relationship evolved?

Having stated all our objectives, the subsequent sections of this report will present a comprehensive statistical analysis using R and an interactive dashboard design using PowerBI, each meticulously conducted to meet the defined research objectives.

# 2**. Background Research**

Various methodologies were applied to extract significant insights from the WDI dataset, with each method playing a crucial role in uncovering patterns, relationships, and trends within the data.

Briefly discussed below are some of these methods:

**Exploratory Data Analysis (EDA)**

This is a critical preliminary step in any data analysis process, conducted to develop a foundational understanding of any underlying patterns within a dataset. It serves as the initial diagnostic phase that provides an assessment of data semantics, quality, and inherent characteristics.

**Descriptive Statistics**

Similar to the EDA, this process involves a comprehensive examination and summary of key statistics relevant to the dataset and the case study under exploration. The Measure of Central Tendency is one of the most significant and commonly used forms of descriptive statistics. It includes measures such as the mean, median, and mode, which provide insights into a dataset's central values.

**Correlation Analysis:**

In our study, we used the Spearman method for Correlation analysis to determine the strength and direction of relationships between our key indicators. We chose this method over Pearson correlation as it is more suitable for non-normally distributed or ordinal data. The output of this method is known as the correlation coefficient, represented by 'r', and it ranges from -1 to 1. A value of -1 indicates a perfect negative correlation, 1 denotes a perfect positive correlation, and 0 suggests no correlation between the variables.

**Analysis of Variance**

ANOVA is a statistical methodology that enables us to compare means across multiple categories and types of categorical variables. It helps us to understand the level and extent of inherent variations by assessing whether there are statistically significant differences among the means of different categories present in a dataset. In our case study, ANOVA will be useful in investigating the observed differences across diverse countries and continents. This is particularly suitable when analysing continuous outcomes such as GDP, inflation, and other similar indicators.

**Regression Analysis:**

With a reputation as one of the most powerful statistical methods, Regression is usually used to model relationships between dependent and independent variables. Simple Linear Regression is used to assess the linear association between two variables - one dependent variable and one independent variable. While Multiple Linear Regression is used when there are multiple predictors to be analysed.

These two methods are denoted by the below equations :



**Time Series Analysis (ARIMA and Holt-Winters):**

Time Series Analysis is used to understand patterns and trends within variables over time. There are two main methods used in this analysis: AutoRegressive Integrated Moving Average (ARIMA) and Holt-Winters smoothing.

Whilst both methods are good forecast models, the ARIMA model is usually used to capture temporal trajectories using differencing across time, while the Holt-Winters technique helps to address seasonality and trends in data with the use of smoothing.

These models both provide a robust framework for predicting future values of variables based on historical trends and give insights into their future trajectories.

# 3**. Literature Review**

Political stability has always been a crucial factor in promoting economic development. Numerous studies have shown that there is a direct correlation between political stability and the economic growth of a country. For example, Baklouti & Boujelbene (2020) highlight that countries which achieve sustained political stability create an environment that is favourable for investment and business development, leading to long-term economic growth.

Similarly, foreign direct investment (FDI) is also attracted to countries that exhibit political stability. A study by Mohamed, M. ( 2015) emphasised the importance of political stability in building investor confidence in a country.

Moreover, political stability also affects other governance indices, which ultimately shape economic outcomes. Iman, M. (2021) suggests that corruption control, good governance, and adherence to the rule of law are integral components of political stability. Countries that have higher levels of corruption control and effective government institutions show better economic outcomes, according to this study.

Another study by Kaufmann et al. (2016) also found that these governance factors act together, influencing the overall economic environment.

**Exports, Imports and Economic Health**

Exports play a significant role in contributing to a nation's GDP growth and catalyse economic expansion, as per Mabrouki Mohamed's study in 2017. The research asserts that a strong export sector positively correlates with increased economic output (GDP).

In a comprehensive study conducted by Sayef, B. (2018), the complex relationship between imports and GDP growth is explored. The findings suggest that imports can contribute to economic growth by meeting domestic demand, but maintaining a balance is crucial to prevent adverse effects on trade balances.

A work by Garcia, M (2018) complements this perspective by examining the impact of trade imbalances on economic health. The study identifies potential risks associated with persistent trade deficits and underscores the significance of policy measures to address trade imbalances.

**Technology Exports**

Emrah Sofuoglu et al. (2015) highlighted in their study on the impact of technology exports on GDP growth and innovation that nations focusing on technology exports tend to experience higher rates of economic growth. They attribute this correlation to increased innovation and competitiveness in global markets.

Confirming this relationship, a report by the United Nations Department of Economic and Social Affairs (2019) noted that technology exports contribute not only to economic growth but also to the continuous development of a country's technological capabilities. The report emphasises the importance of fostering a conducive environment for technology exports to enhance economic health.

**Labour Force and Unemployment**

Numerous studies conducted in the past have highlighted the pivotal role played by the labour force in shaping a nation's economic trajectory. According to Karen A, et al. (2022), a skilled and adaptable labour force is a crucial driver of productivity growth. They argue that investments in education and training help enhance labour force dynamics, foster innovation, and contribute to economic development.

Furthermore, Brown A. (2016) emphasises the significance of labour force mobility in stimulating economic growth. Their longitudinal study reveals that regions with higher labour market fluidity experience more significant and sustained GDP growth.

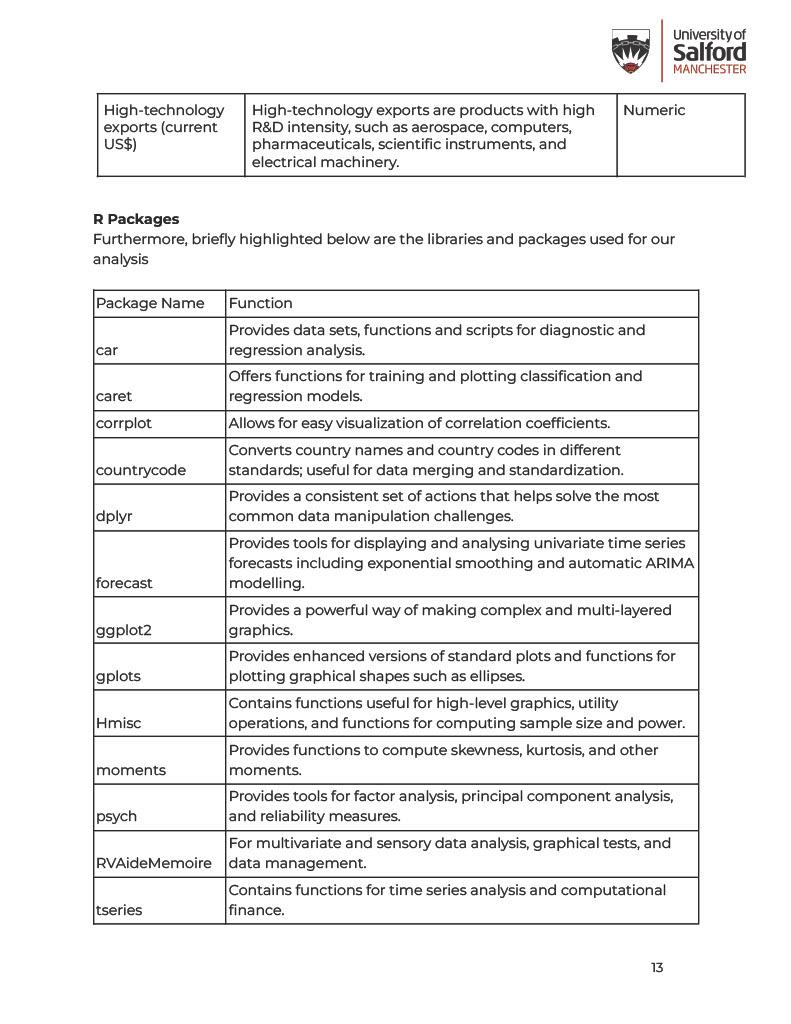
**Interactive Dashboard Design and Presentation**

Interactive dashboards have become an essential tool for visualising complex data using dynamic and user-friendly interfaces. When building a dashboard, it is crucial to take the time and effort to create a thoughtful layout that is easy to understand.

According to Meagan (2019), a well-designed dashboard should provide a clear overview of critical information at first glance. Placing important visualisations and widgets strategically contributes to user engagement and effective communication of insights. Helen, W. (2019) emphasises this importance and asserts that ensuring every pixel and block of the dashboard conveys essential information is crucial.

Clark, D. (2017) further highlights the need for simplicity, consistency, and user-centric design to facilitate seamless understanding. By following these guidelines, the dashboard will be more effective in presenting data and insights to any category of user and device.

# 4. **The Dataset**

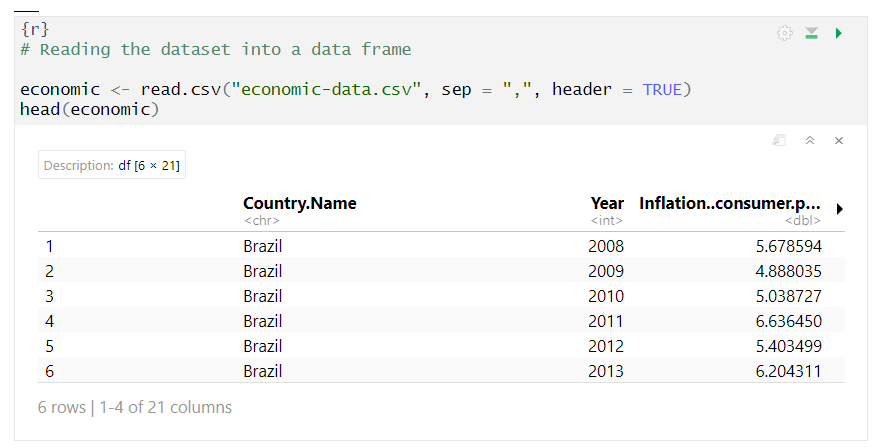


# 5. **Dataset Preparation**

To begin, the dataset underwent a meticulous preparation process to ensure its reliability for statistical analysis.

**Loading the dataset:**

The dataset, stored in a CSV file ("economic-data.csv"), was read into a data frame and named 'economic.' The structure of the dataset was subsequently explored using the head() and str() functions.



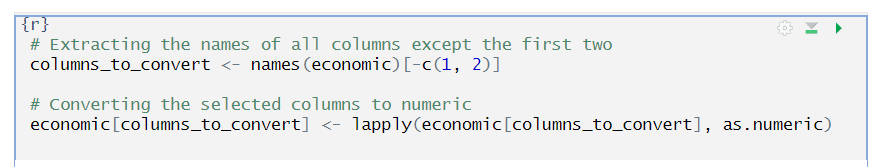
**Variable Renaming**

To ensure readability and easy referencing, the variables were renamed to concise and shorter versions.



**Data Type Conversion**

It was also observed that some of the indicators had inappropriate data types *(‘character’)* and there was a need to convert these data types to *Numeric*, paving the way for an error-free analysis.



**Detecting and Handling Missing Values:**

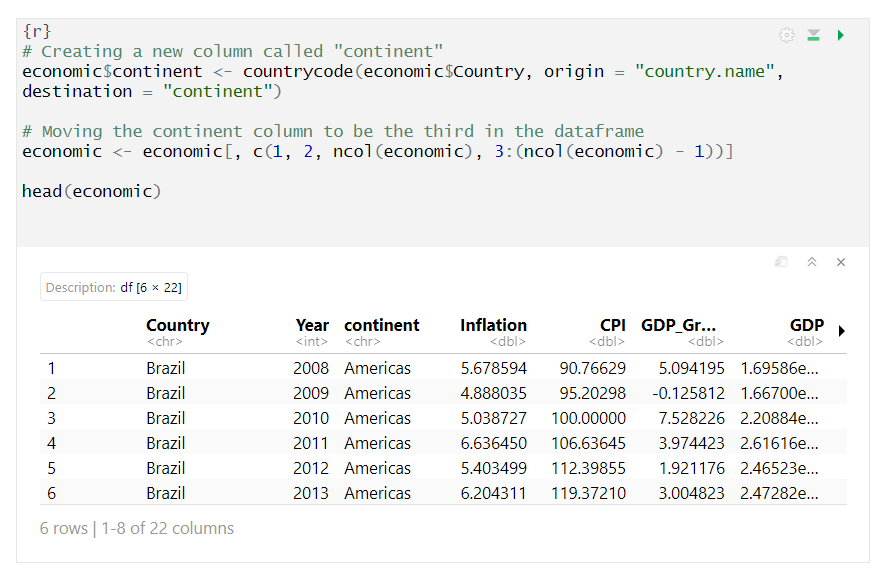
On closer inspection, some missing values were identified, and a function (replace\_missing\_values) was defined to impute them based on the mean values within corresponding countries.

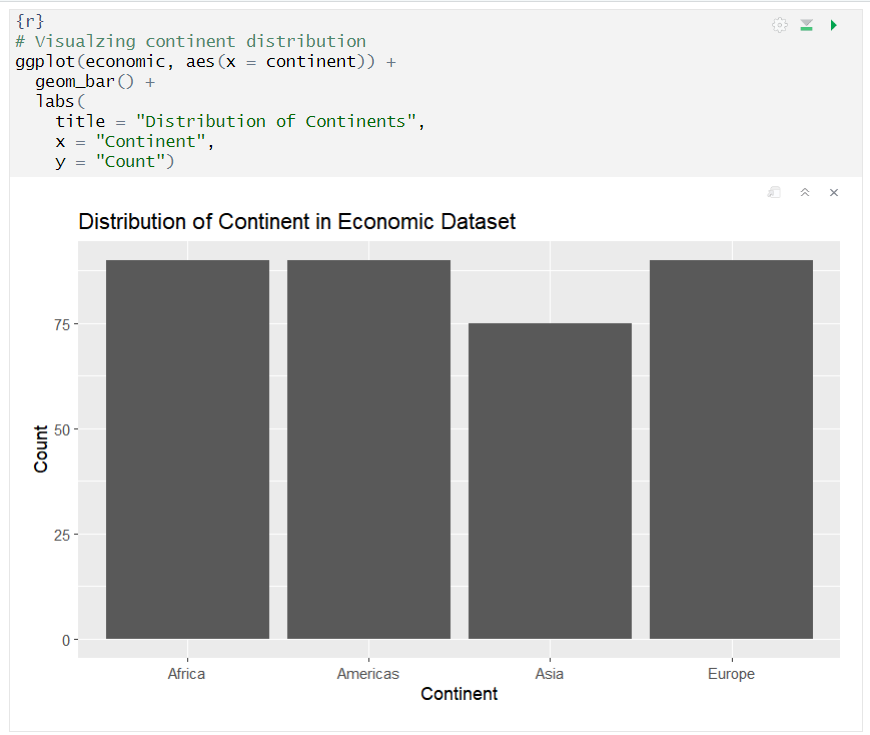


**Feature Engineering:**

A 'continent' column was added to the dataset to provide a geographical context for each country. This feature is crucial for comparative analysis across regions. A bar plot was created to visualise the distribution of continents in the dataset.

This data preparation phase ensured a coherent and consistent dataset, laying the foundation for robust statistical analysis. In the next section, we will explore the data and uncover key insights.





**Outlier Detection:**

To ensure the integrity of our analysis, we conducted outlier detection across key economic indicators and box plots were employed to visually identify these potential outliers.

**Observations:**

Upon careful examination, these significant outliers were identified across multiple variables, and they notably reflect the diverse economic performances of countries, ranging from exceptionally positive to challenging economic conditions.



**Decision on the Outliers:**

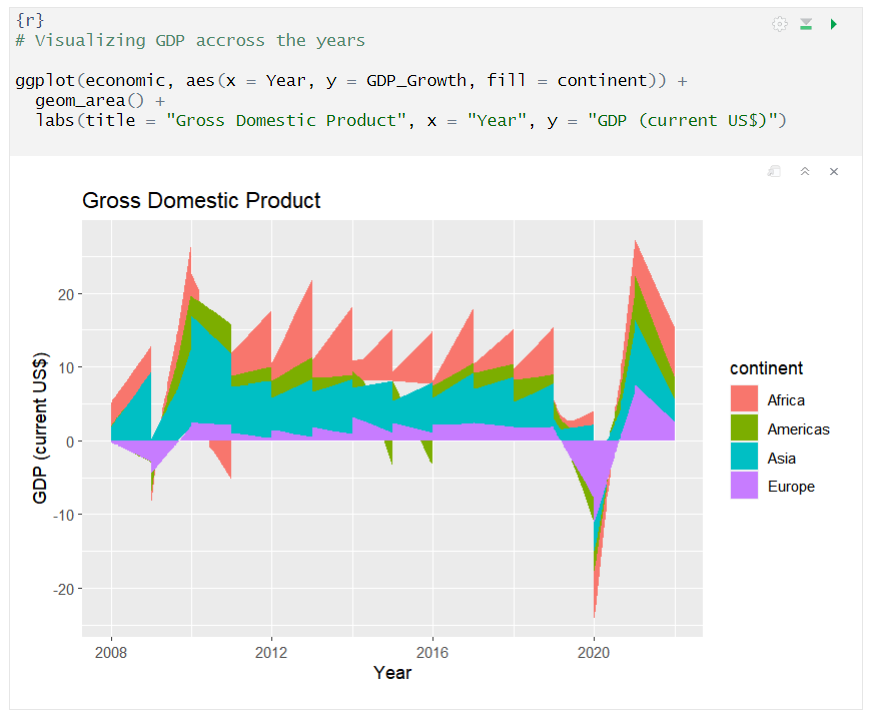
Given the real-world nature of economic indicators, a strategic decision was made not to remove the outliers. Instead, we chose to retain these data points. This decision aligns with the nuanced nature of economic data, where outliers can represent unique economic circumstances rather than mere data errors.

# 6. **Exploratory Data Analysis (EDA)**

In this section, we conducted a comprehensive EDA to uncover patterns and trends peculiar to each indicator in our dataset.

#### GDP Annual Growth (%)

The analysis began with a stacked area plot which visualised the *GDP\_Growth* across continents over the years. The R code used for this visualisation can be found below:

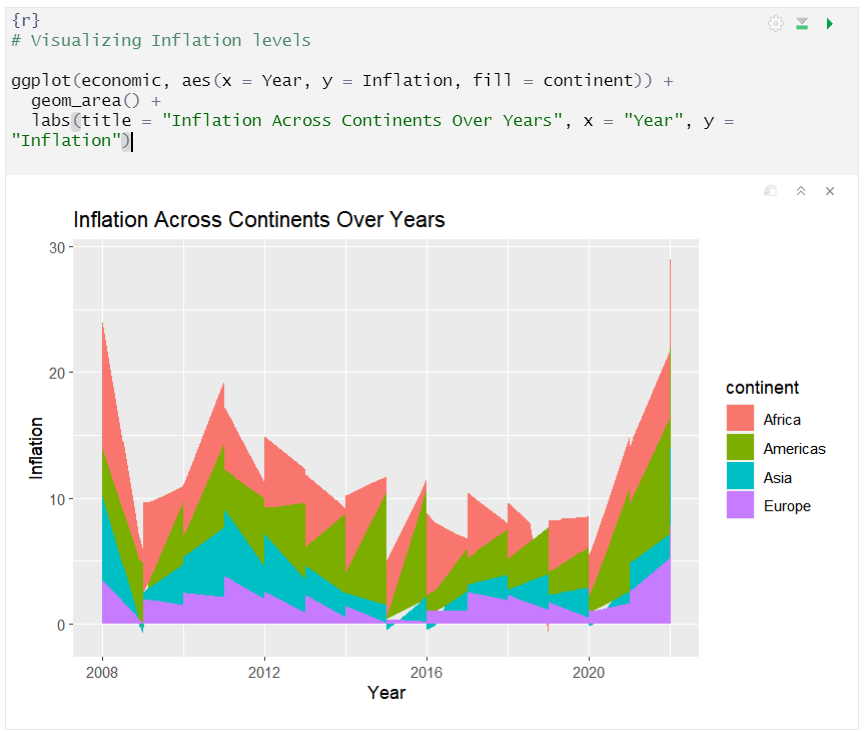


From the plot, we observed that over the last 15 years, there’s been substantial improvement in the GDP growth rate across African countries. This positive trend suggests economic progress and resilience within the continent.

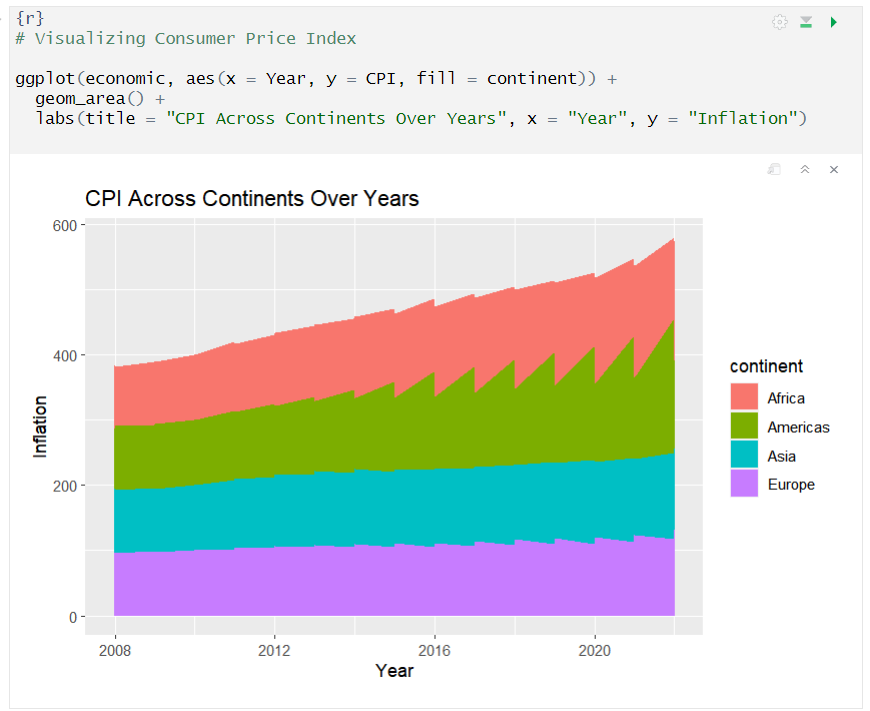
Furthermore, we observed a global slump in 2020, coinciding with the onset of the Covid-19 pandemic and the economic crisis that followed

#### Inflation and Consumer Price Index

Here we visualised how inflation rates vary across continents over the years.



As expected, there is noticeably high inflationary pressure within the African continent, a trend peculiar to developing economies and suggestive of harsh economic realities.



This revelation is supported by the International Monetary Fund (2022) report, indicating worrying decade-high levels of inflation, a trend exacerbated by political instability and worsening food insecurity.

#### Unemployment

Using a stacked area and scatter plot, we observed that Africa continues to grapple with the age-long problem of unemployment, followed by the Americas, whose figures seem to be heavily skewed towards the Latin American countries.



Countries like Spain and South Africa appear to have relatively high unemployment rates, whilst China exhibited a steady and low unemployment rate over the years, with only a slight increase in 2020 due to the pandemic.

#### High-Technology Exports

Furthermore, we looked at the regional trends shaping the global technological landscape, with a special focus on technological exports.



We observed that Asia has considerably cemented its position as the leading global technology force, with China leading the region in high-tech exports. Similarly, Europe remains a formidable contender, thanks to major players such as Germany, the United Kingdom, and Switzerland.

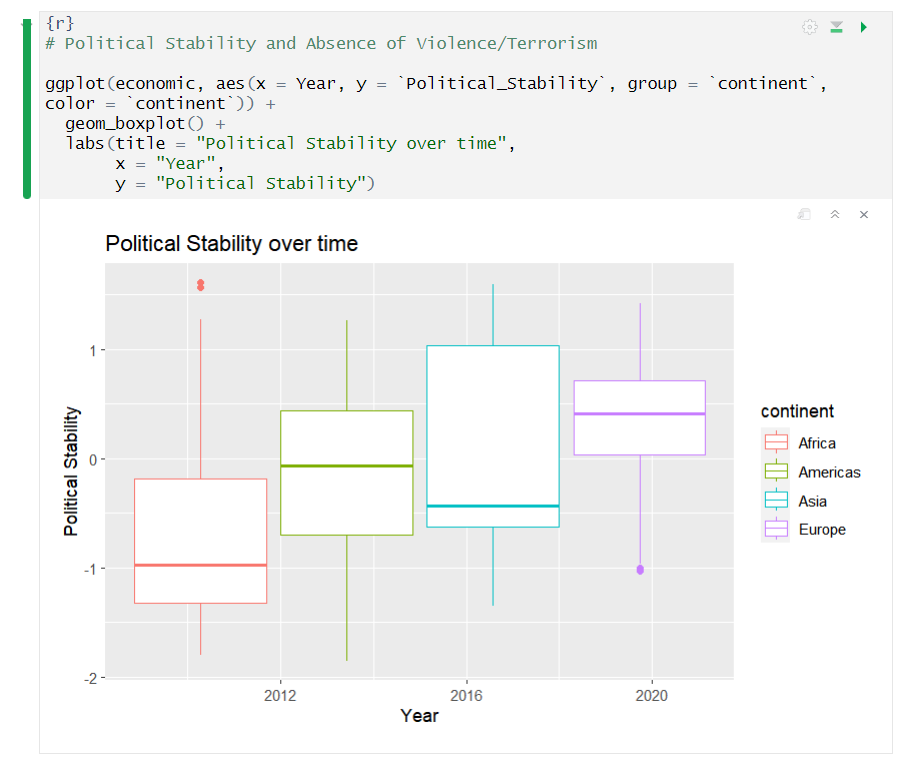
The Americas led by the United States also display a strong foothold as well, while African nations have lately begun to grow, indicating heightened technological investments in the region.

#### 

#### Political Stability and Absence of Terrorism

Using a box plot, we also explored the perceived likelihood of political instability and politically motivated violence across different continents, and Africa was revealed to be the most unstable and likely to resort to political violence, followed closely by Latin and North America.

Asia and Europe on the other hand appear to be the most relatively stable regions suggesting a shared perception of stability and non-violent instances.



# 7. **Statistical Analysis**

During our analysis, we went through several phases, starting with basic descriptive statistics and moving on to exploring time and trends using an Arima model. Throughout each phase, we carefully uncovered various patterns, relationships, and subtleties that were hidden within our data.

### 7.1 **Descriptive Statistics**

We started our analysis with some basic descriptive statistics, which is always a good place to start.

We used the 'describe' function from the 'psych' package, which gave us a neat summary of our data showing statistics such as mean, median, how spread out each variable is, and other interesting statistics like skewness and kurtosis.





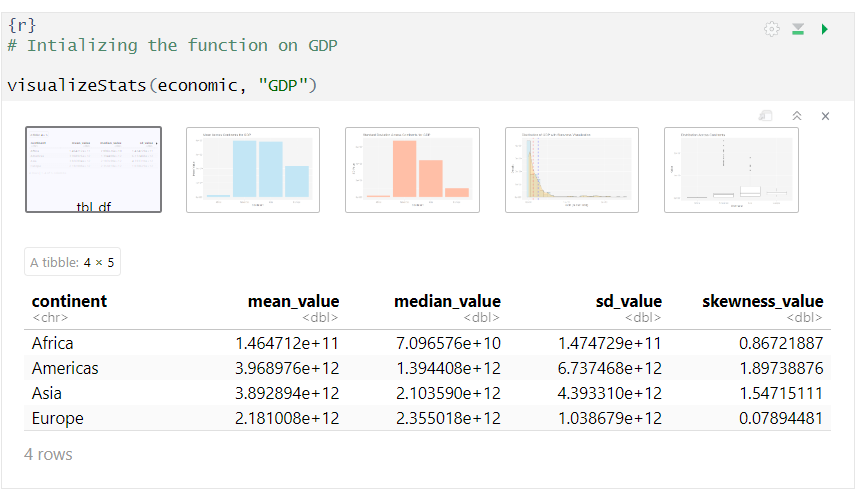
**Continent-wide Analysis**

To conduct a deeper analysis and investigate how the economic indicators vary across continents, we defined a custom function, ‘*visualizeStats’*. This function simply fetched the mean, median, standard deviation, and skewness of any indicator across each continent, and then presented them in a graph.



**Gross Domestic Product**

To begin, the function was initialised to dissect the "GDP" indicator across continents and it yielded the below insights.



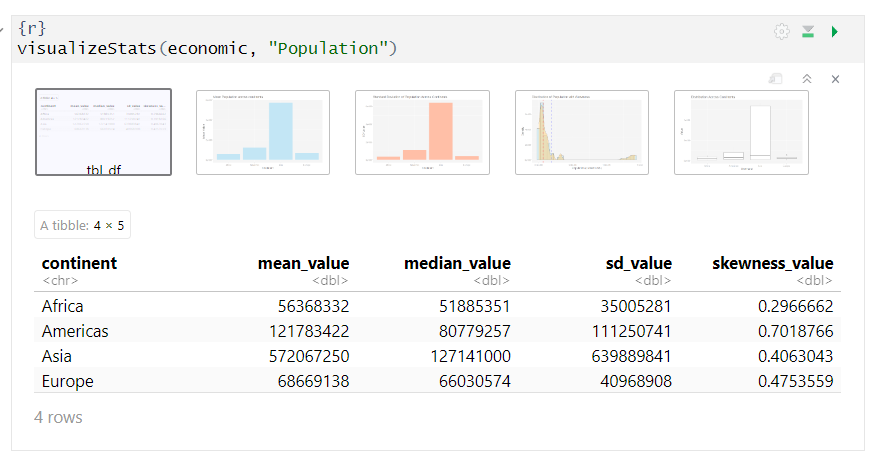


On close inspection, it can be inferred that countries in the Americas had the highest average GDP, followed by Asia, Europe, and Africa respectively. The Americas also showed the greatest variability, which indicates that its economy is more diverse than the other regions.

Whilst Asia exhibited a distribution suggesting a mix of economically advanced nations and those still developing; Europe depicts a contrasting picture with a skewness close to zero, indicating a relatively symmetric and balanced spread of GDP values.

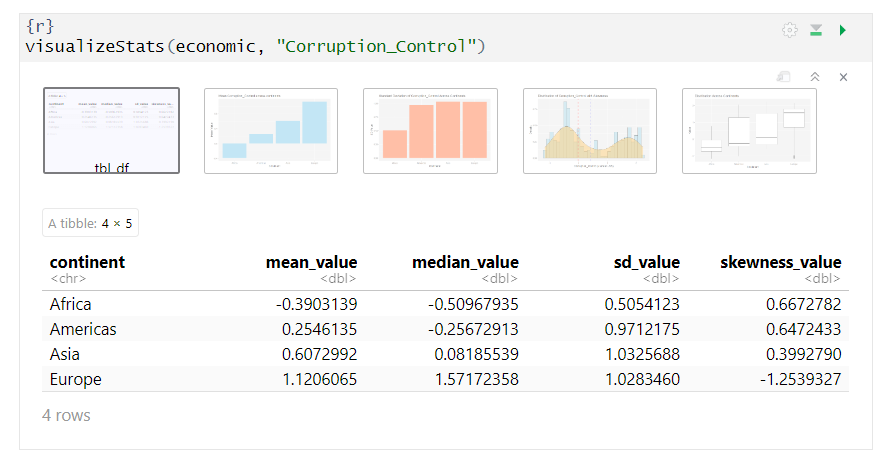
**Population**

Next up was population. No surprises here as Asia topped the charts, with the Americas close behind. Africa and Europe, however, showed smaller average populations but had an interesting skew towards larger sizes. It appears most countries were on the smaller side, and a few big ones tipped the scale.



**Corruption Control**

As can be visually observed, Europe stood out with the highest mean, indicating a high level of corruption control, while Asia looks relatively tougher on corruption control measures too.



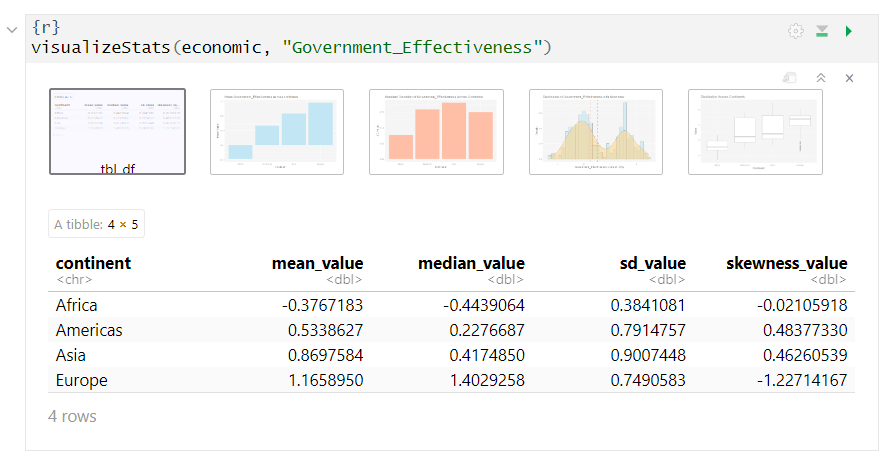
Africa on the other hand showcased negative mean and median values, suggesting a prevalence of corruption, which is supported by the study conducted by Simplice, A. (2013), where he revealed corruption to be the most daunting challenge for most African countries.



It can also be observed that anti-corruption policies have a bimodal distribution of success in fighting corruption. Countries with the highest and lowest corruption levels tend to experience the least positive effects, while those in the middle see the most benefits.

**Government Effectiveness**

Similar to the statistics revealed for Corruption Control, there is a notable lethargy in governance across Africa when compared against its counterparts from other continents.



In contrast, Europe stood out with a high level of effective governance, with its left-skewed distribution (1.28) indicating that a significant number of European countries had effective governments. Countries in Asia and the Americas also enjoy effective governance, with their wider standard deviation implying varying degrees of effectiveness within each region. 

### 7.2 **Correlation Analysis**

Delving deeper, I endeavoured to uncover the relationships between our indicators, and what influences possibly exist among them.

Using the *‘corrplot’* package in R, a correlation matrix was initialised to examine the relationship strength between the variables.

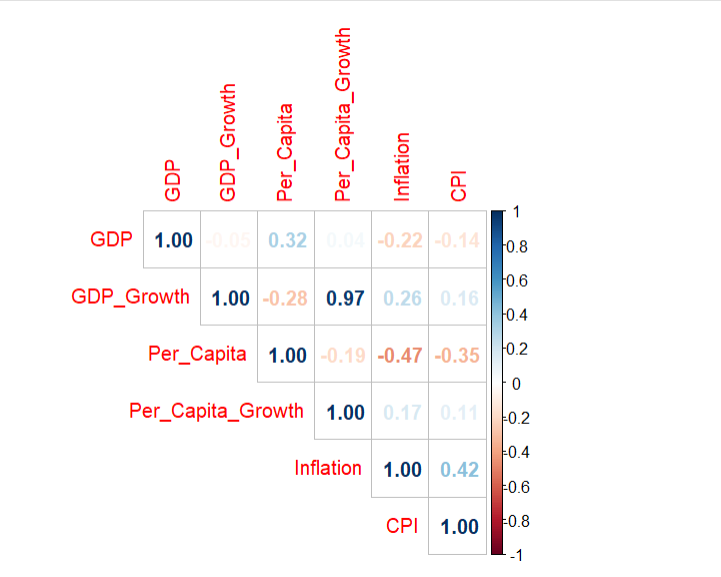


We can observe that there appear to be significant relationships between our indicators, with most of them showing positive and negative relationships, ranging from strong to weak.

To further break down these relationships, the indicators were grouped into categories based on their similarity.

#### Income vs Inflation Indicators

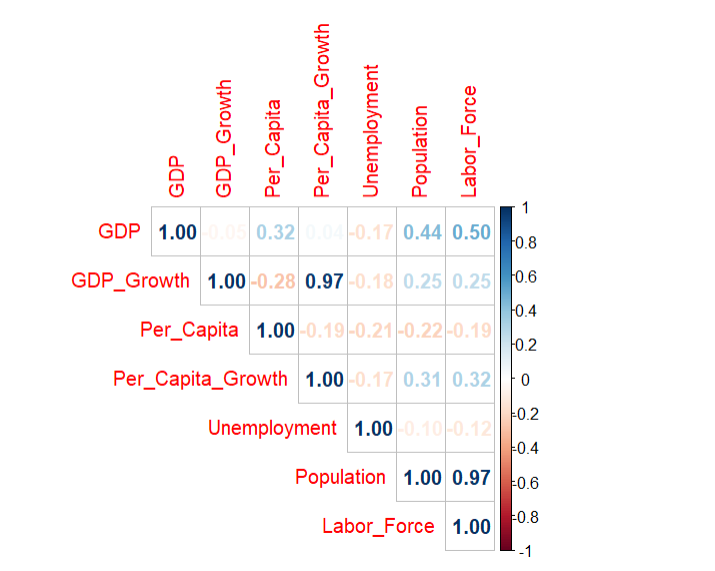
To explore the relationship between inflation and our income indicators, a subset was created. The key indicators observed were GDP, GDP Growth, Per Capita Income, Inflation Rate, and Consumer Price Index (CPI).



The figure above shows a negative correlation of -0.62 between inflation rates and per capita income, indicating the negative impact of inflation on individual economic well-being. On the positive side, there is a strong positive correlation of 0.61 between GDP and per capita income, and an even stronger relationship of 0.96 between the growth rates of GDP and per capita income, emphasising the interconnectedness of national and individual economic prosperity.

#### Income vs Labour Indicators

To add an extra layer to our correlation analysis, we explored how labour-related variables influence economic growth.

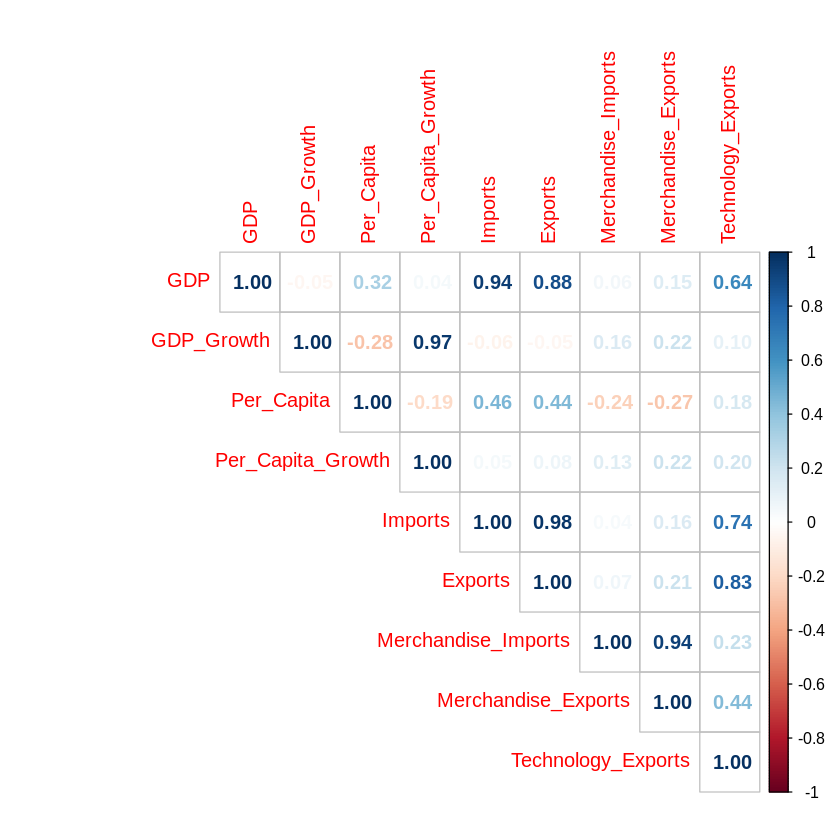


The matrix showed a strong positive correlation between GDP and variables like Population (0.68) and Labor Force (0.73). This showed that countries with larger populations and labour forces were more likely to experience high economic growth.

Interestingly, GDP Growth has a negative correlation (-0.17) with Unemployment, indicating that unemployment tends to decrease as prosperity increases. Lastly, a strong positive correlation of 0.99 between Population and Labor Force confirms the natural connection between these variables.

#### Income vs International Trade Indicators

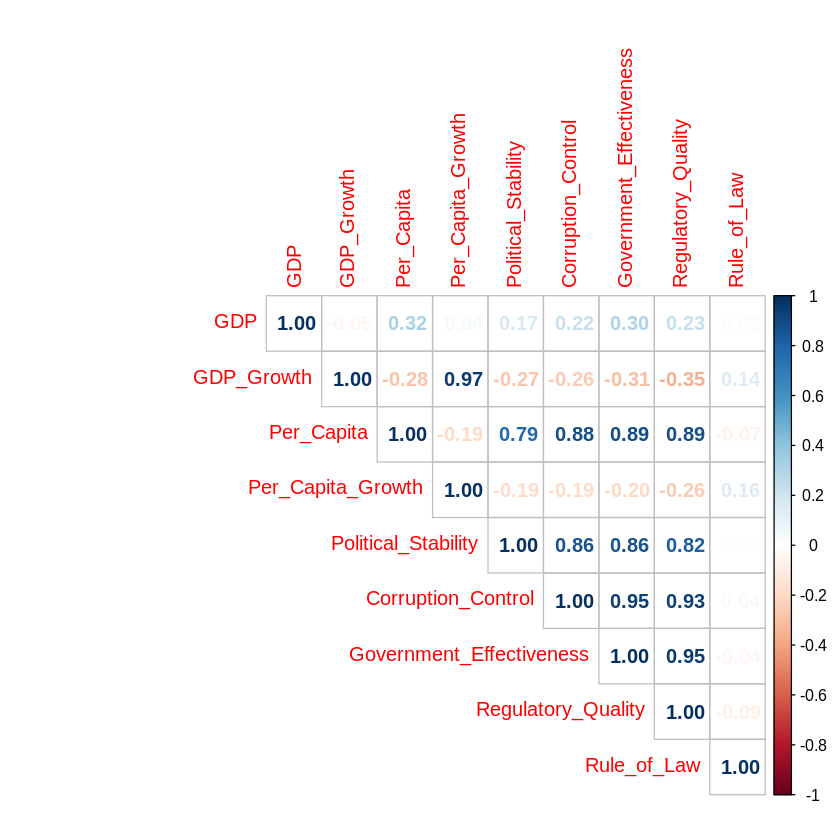
Subsequently, our focus shifted to trade, emphasising Imports, Exports, and Technology Exports.



A high positive correlation was observed between GDP and Imports (0.935) and Exports (0.920). This revelation strongly follows basic economic logic, indicating that the countries with larger trade balances tend to experience higher GDP. Notably, technology exports also exhibited strong relationships with GDP, imports, and exports, implying a connection between technological prowess and economic strength/trade.

#### Income vs Governance Indicators

To conclude our correlation analysis, a focused examination of income and governance-related variables was conducted and a positive correlation was observed between GDP and governance variables such as Per Capita, Political Stability, Corruption Control, Government Effectiveness, and Regulatory Quality. This indicated that as these governance-related factors improve, GDP tends to increase as well.



An even stronger relationship was observed between Per Capita income and the governance variables, suggesting that an improvement in a country’s governance is likely to result in improved quality of life and economic well-being of its residents.

Interestingly, the analysis revealed that Political Stability, Corruption Control, and Government Effectiveness exhibited strong positive correlations with each other, indicating their interconnectedness in influencing economic outcomes.

### 

### 7.3 **Hypotheses Testing**

This phase rigorously put all of our conjectures through statistical scrutiny and validation. It involved two approaches; hypothesis testing based on correlations and the ANOVA method. The correlation analysis tested the relationships between key economic indicators, while ANOVA explored and compared the indicators on key metrics.

#### Correlation-Based Hypothesis Testing

##### **Hypothesis 1:** Inflation vs Unemployment

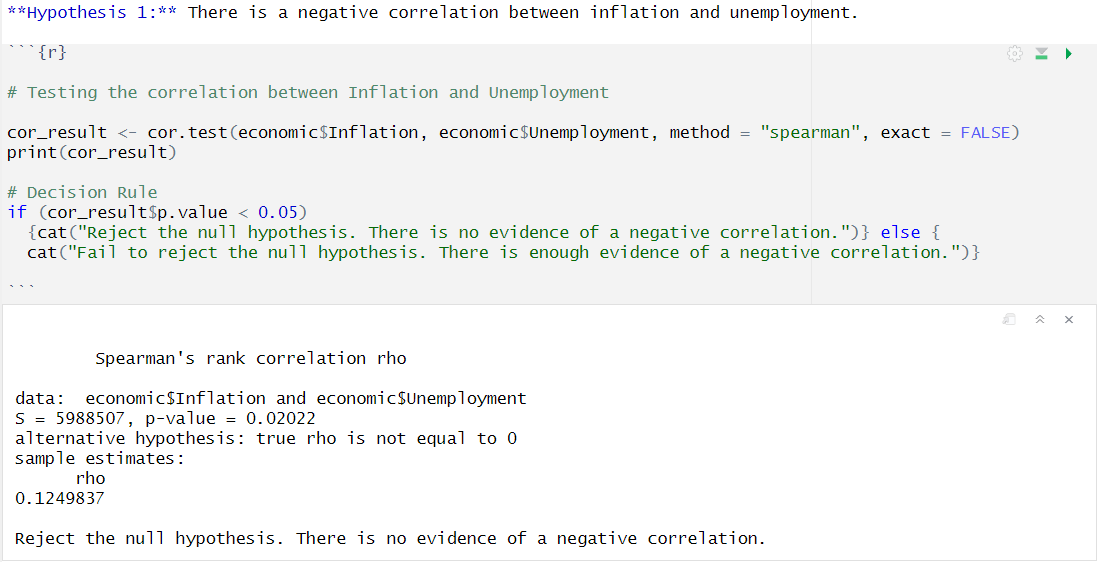
My first hypothesis was based on the probability of a relationship between Inflation and Unemployment. The inspiration was to test the Phillips curve Economic theory that suggests an inverse relationship between inflation and unemployment.

Based on this, we set our null and alternative hypotheses as follows:

**HO:** The *correlation coefficient (rho)* between both variables is negative

**H1:** The *correlation coefficient (rho)* between both variables is not negative

Subsequently, a Spearman correlation test was calculated and a decision rule was set using a significance level of 0.05.



Based on the result above, we reject the null hypothesis and conclude that there is no evidence of a negative correlation between inflation and unemployment. The p\_value is less than our preset significance level.

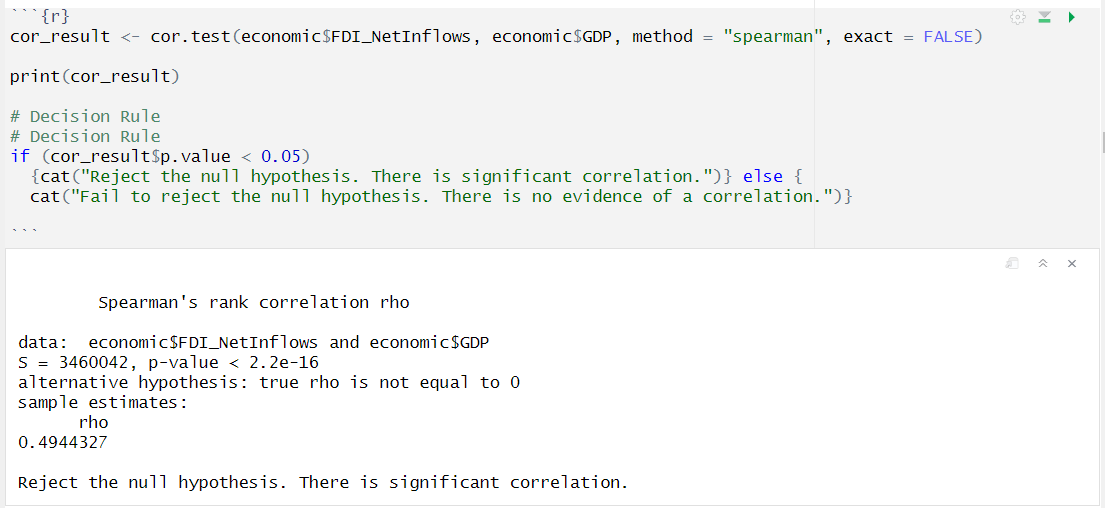
##### 

##### **Hypothesis 2:** Foreign Direct Investment (FDI) vs GDP

FDI has often been touted to be the driver behind capital and technology ingestion and is meant to be positively correlated with economic growth. However, to prove this claim, we set it to a hypothesis test:

**HO:** There is no relationship between FDI and GDP

**H1:** There is a significant positive relationship between FDI and GDP



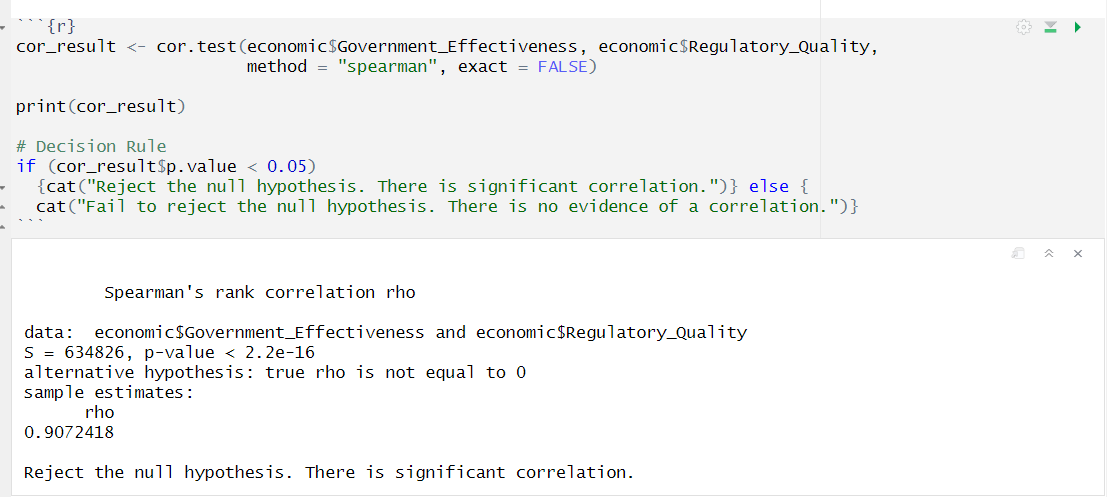
After conducting the test, we found that Spearman's rank coefficient was 0.4944 and the p-value was statistically lesser than our significance level. This led us to reject the null hypothesis. Our result supports the idea that foreign direct investment (FDI) drives economic growth.

##### **Hypothesis 3:** Government Effectiveness and Regulatory Quality

We believe that an efficient government is likely to implement and enforce better regulations, however, to validate this belief, I conducted a hypothesis test:

**HO:** No relationship exists between Government Effectiveness and Regulatory Quality

**H1:** A significant positive relationship exists between Government Effectiveness and Regulatory Quality



The p-value subsequently obtained was extremely low and indicated a strong correlation between the variables. The correlation coefficient (rho) was also significantly high at 0.9072418.

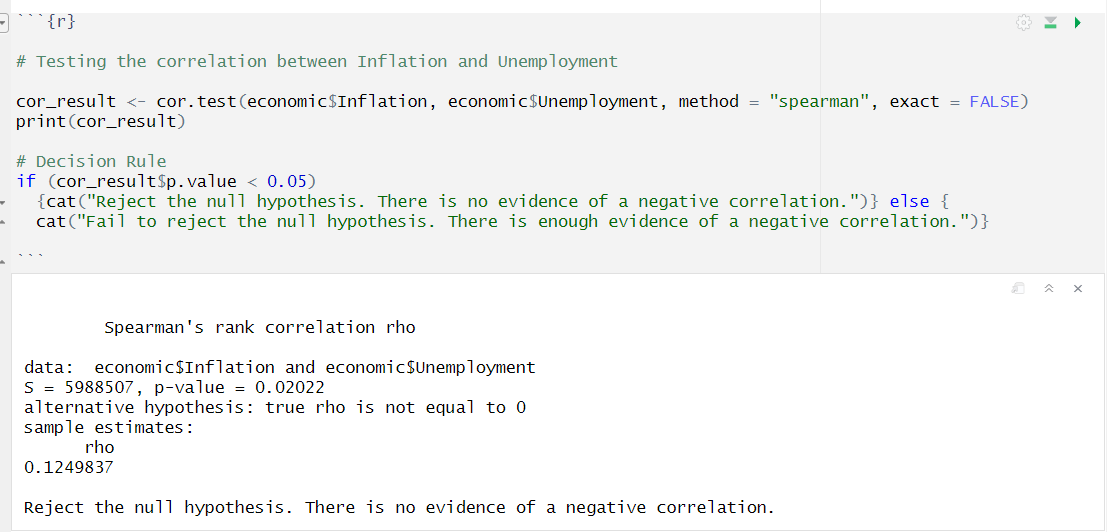
I, therefore, inferred that there was enough empirical evidence to reject the null hypothesis and support the idea that as governments improve, the quality of their regulation is also likely to improve.

##### **Hypothesis 4:** Exports vs GDP

Studies like Titus O. (2018) have found that export drives the economic growth of many countries across the world today. To test if this assertion holds, I employed hypothesis testing once again.

**HO:** No relationship exists between exports and economic growth

**H1:** A significant positive relationship exists between exports and economic growth



The test yielded a p-value less than our significance level of 0.05 and we simply rejected the null hypothesis, that there is no significant correlation between GDP and exports.

#### Analysis of Variance (ANOVA)

Furthermore, we conducted an in-depth analysis of variance with hypothesis testing to examine the differences in our key variables and potential disparities across regions.

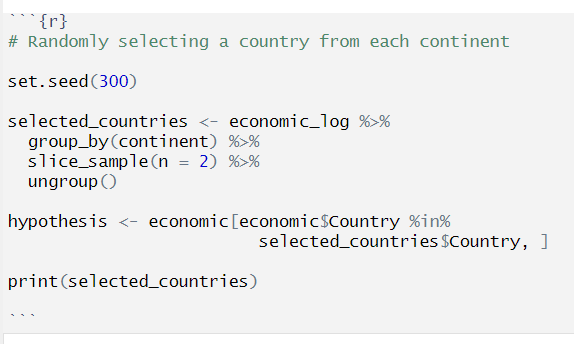
To avoid bias, we randomly selected a subset of countries from different continents:

- Americas: United States, Canada

- Africa: South Africa, Rwanda

- Asia: India, Japan

- Europe: Spain, United Kingdom



##### **Hypothesis 5:** Mean Inflation Variances Among Continents

This test aimed to examine the mean inflation across the four continents and see whether it exhibits significant changes.

We formulated a null and alternative hypotheses as follows:

**HO:** No significant difference in mean inflation across continents.

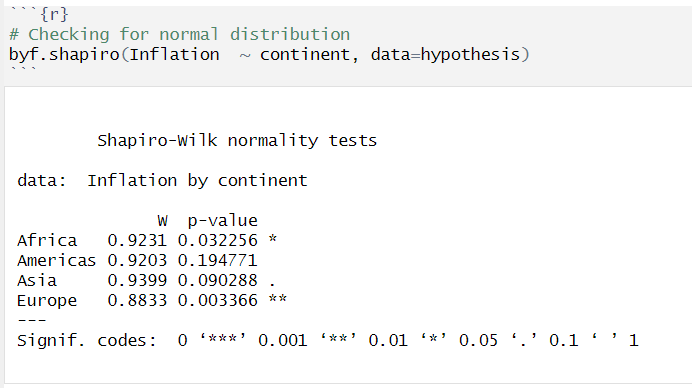
**H1:** Significant difference in mean inflation across continents.

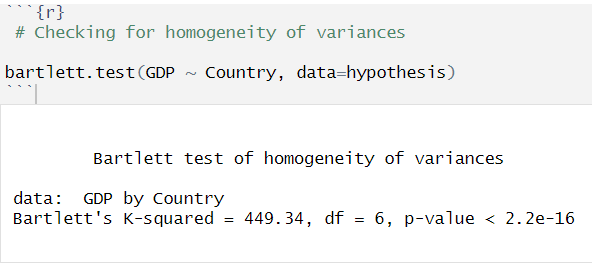
Before conducting the hypothesis test, we verified the following six ANOVA assumptions:

1. Inflation is a continuous variable.
2. The independent variable, Continent, has four categorical regions
3. Assumptions across each continent are independent.
4. There are no significant outliers in the variable. There were however two outlying data points in Africa, but they are non-significant and do not raise concerns.

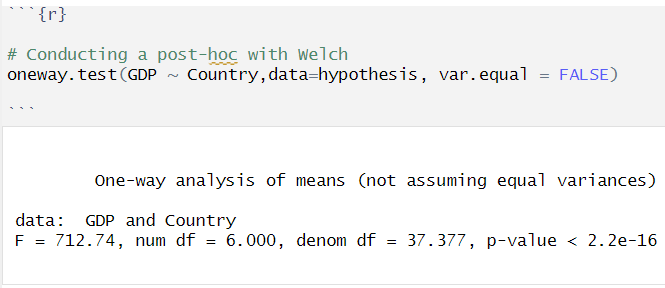


**Assumption 5:** To meet this assumption, the dependent variable needed to be approximately normally distributed. The Shapiro-Wilk normality test was conducted, and it revealed deviations from normality in Africa and Europe, with p-values (0.043943 and 0.003366) lower than the significance level.



**Assumption 6:** The final verification was to check the homogeneity of variances., and the Bartlett test was introduced. The results showed a p-value of 1.283e-12 which is less than the significance level, suggesting a difference in variances among the countries.

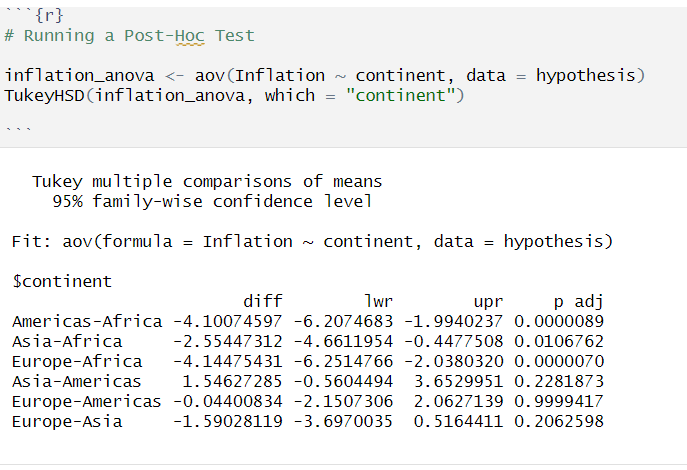
Because the last two assumptions were violated, we were then required to conduct a posthoc test using **Welch’s ANOVA** instead.



The post-hoc test returned an F-value of 35.064 and a low p-value of 5.45e-15. This meant that there was compelling evidence of differences in the mean inflation among the selected countries, and we proceeded to reject the null hypothesis (HO).

**Tukey HSD Test**

Afterwards, a Tukey HSD test was performed to understand the source of the variation between the continents and the results showed that the differences were emanating from the combinations of Americas and Africa, Asia and Africa as well as Europe and Africa.



##### **Hypothesis 6:** Comparison of Mean GDP Across Continents

Based on the sample drawn earlier, a comparison of GDP figures across each continent was conducted.

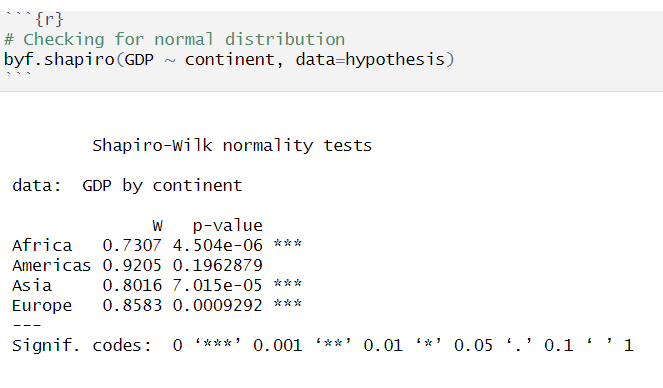
**HO:** No significant difference exists in the mean GDP across continents.

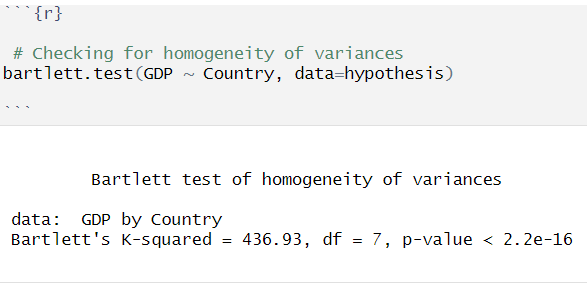
**H1:** A significant difference exists in the mean Inflation across continents.

We checked and verified the first three ANOVA assumptions. We also verified the spread of GDP values across the selected countries and found no outliers, confirming the fulfilment of the fourth assumption.



However, after conducting the Shapiro-Wilk normality test, we observed that the p-values for all continents were less than 0.05, indicating that the normality assumption was violated. Similarly, the Bartlett test showed a low p-value indicating that the variances of GDP across countries were not homogeneous.





Because the last three assumptions were violated, we had to carry out a post-hoc test using Welch's ANOVA instead.



Using Welch's test, it was found that there is strong evidence against the null hypothesis of equal mean GDP across countries. The test yielded a low p-value, which supports the idea that there are significant differences in mean GDP across continents.

##### **Hypothesis 7:** Comparison of Mean Unemployment

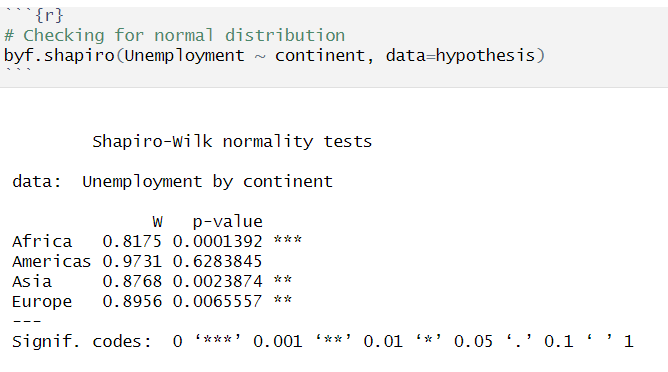
To compare the unemployment figures across each continent, the below hypotheses were formed:

**HO:** No significant difference exists in the mean unemployment rates across continents.

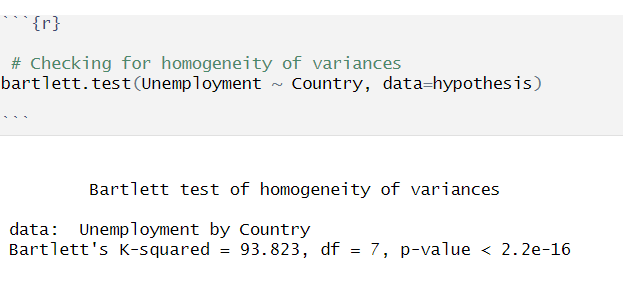
**H1:** A significant difference exists in the mean unemployment rates across continents.

As usual, we checked and fulfilled the first three ANOVA assumptions. I also found no significant outliers in the data. 

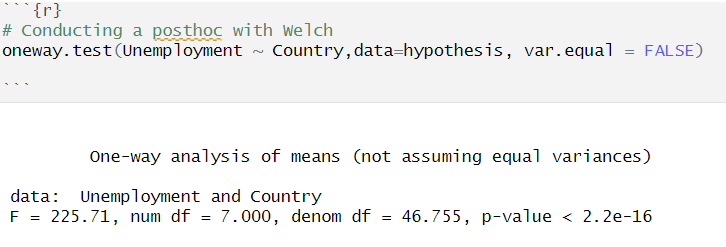
Furthermore, we conducted a normality test and discovered that the p-values were less than the significance level for Africa, Asia, and Europe, which meant that the test failed. However, the Americas showed a p-value of 0.6284, suggesting normality.



The Bartlett test of homogeneity of variances also showed that there were significant differences in variances across continents, with a p-value of less than 2.2e-16.

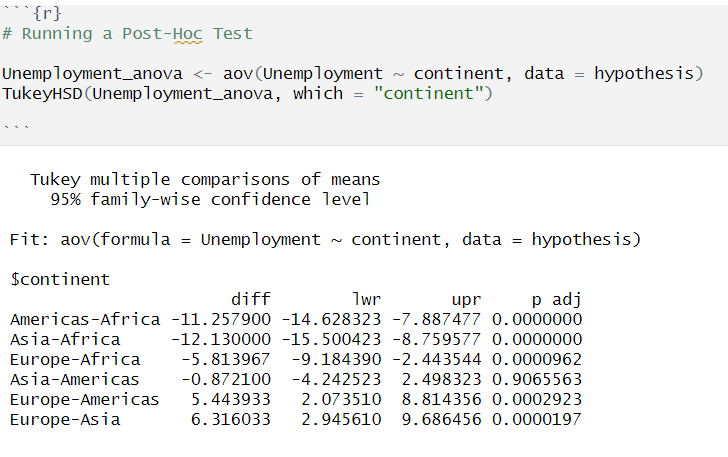


Subsequently, a posthoc test with Welsh was performed it showed that there was a significant difference in the mean unemployment rates.



Finally, the Tukey test was conducted and it revealed differences in the mean values across continents. The test showed significant differences in all the pairwise comparisons except for Asia and the Americas, leading us to reject the null hypothesis once again.

In conclusion, the results demonstrate that there is a significant difference in the mean unemployment rates across continents.



### 

### 

### 7.4 **Dataset Transformation**

In this phase, we needed to ensure that our dataset was appropriate for regression analysis by checking for normality and performing needed transformations.

##### **Shapiro-Wilk Normal Distribution Test**

To check for normal distribution, we designed a function called *'check\_normality'* and also implemented the Shapiro-Wilk test. Upon close inspection, we observed that the selected indicators did not appear to be normally distributed.





However, to statistically confirm this observation, we conducted the Shapiro-Wilk test on the entire dataset using a function called *'ultimate\_shapiro\_test'*. It delivered a decision on the normality of the entire dataset.



These results, complemented by the initial visualisations, led us to reject the Null Hypothesis and confirm that no indicator in the dataset was normally distributed. This was a crucial finding because we then needed to perform transformations to make the dataset conducive for regression analysis.

##### **Logarithmic Transformation**

To get the dataset closer to a normal distribution, we then applied Logarithmic transformation. This technique aimed to mitigate the non-normal distributions and prepare them for regression modelling.

Following the transformation, histograms were used to visualise and inspect the transformed indicators. Although perfect normal distributions weren’t achieved, we were successful in bringing the dataset closer to it.



### 

### 

### 

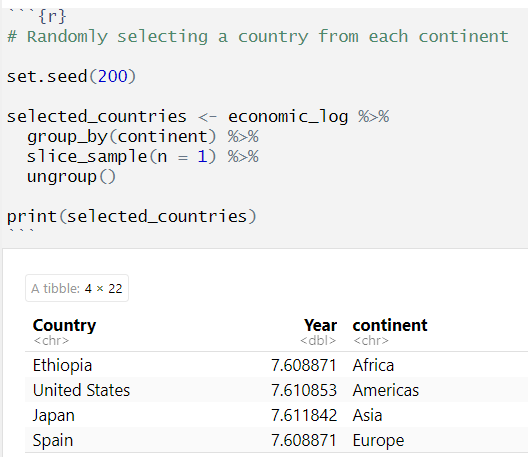
### 

### 

### 7.5 **Regression Analysis**

After the logarithmic transformation of the dataset, I used Simple and Multiple Linear Regression to create models for each continent based on a representative sample country. Then, I developed a more general model reflecting the global economic landscape.

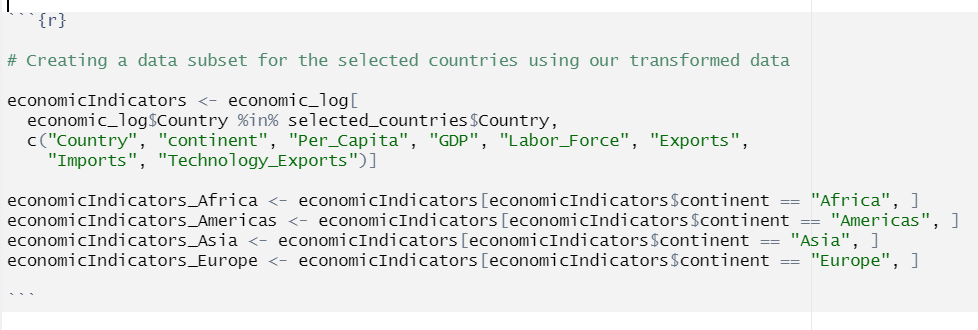
To begin, I randomly selected Ethiopia, the United States, Spain, and Japan as the focus of our regression models.



##### **Simple Linear Regression:** Predicting GDP Growth using Trade Indicators:

As per my research objective, I developed a model that shed light on how indices like exports, imports, and technology exports play a pivotal role in shaping a nation's economic growth.

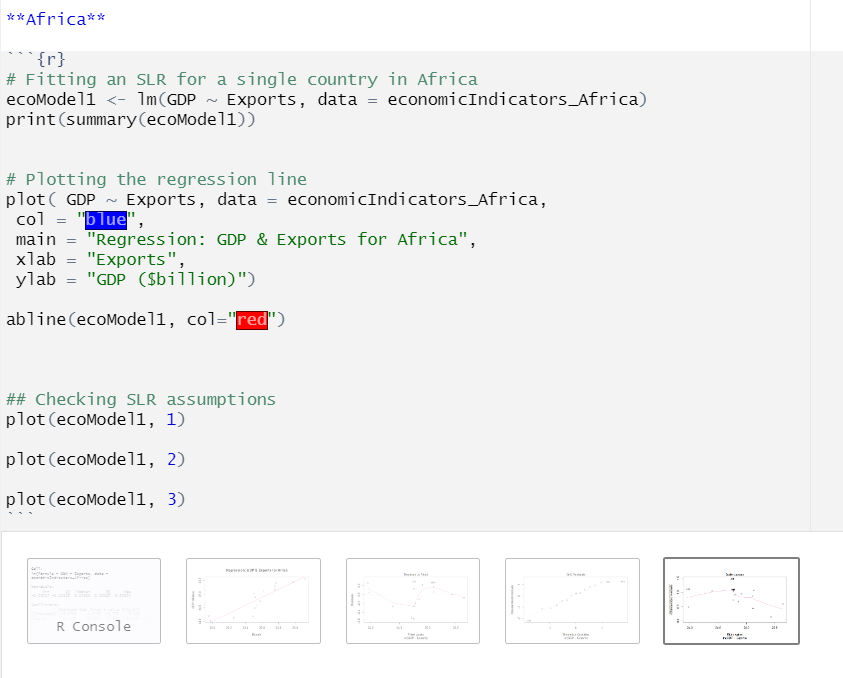
To start, I created subsets of *'economicIndicators'* for each continent by capturing only the relevant data points.



Once this was done, I proceeded to create continent-specific models using the representative country from each continent. This was done to capture the peculiarities that may be present in each continent and also mitigate the risks associated with fitting a generalised regression model.

**Continent-Specific Model**

Focusing on the relationship between GDP and Exports for the chosen African country (Ethiopia), the regression model was established as illustrated below:



The output of the model revealed a sufficient coefficient of determination (R2) of 0.77, a significant p-value less than 0.05, and an F-statistic of 48.05. The R2 indicated that a substantial portion of the GDP in Africa (about 77%) can be explained by the export figure.



However, so we could ensure the validity of the SLR model, we conducted checks on the key assumptions. Plots were also generated to assess the normality of residuals, homoscedasticity, and influential data points.

As observed below, the model still did not satisfy our assumptions.



Extending our analysis, we then fitted simple linear regression models for all the other continents (Americas, Asia, and Europe) and summarised below are some of the key observations from the results:

1. The models all exhibited returned good results in terms of the Adjusted R2 and the linearity of the resulting regression lines.
2. However, the models violated the rest of the SLR assumptions. The models showed clear deviations from normality, homoscedasticity, and independence assumptions.

Based on these violations, there was a need for further exploration with a more generalised model, instead of a continent-based model.

***NB:*** The plots visualising the assumption violations have been added in the appendix section.

**Generalised Model**

Expanding our focus to capture all the countries, we developed a general SLR model. It was hoped that this macro-level model would offer a better result, considering the collective impact of all the nations.

To this end, an SLR model was fitted on the entire dataset, and the resulting regression line is visualised below.

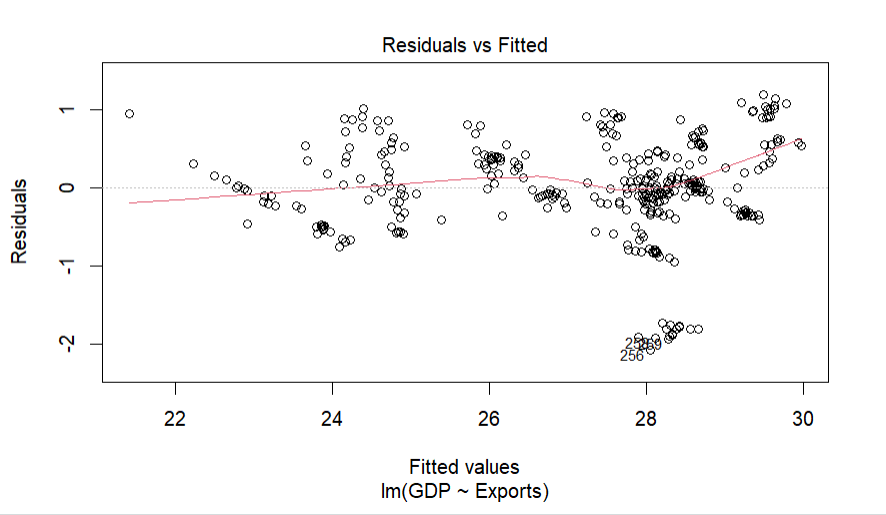


As can be observed in the above plot, the regression line appears linear, closely fitting to the observed data points and demonstrating the strength of the model.

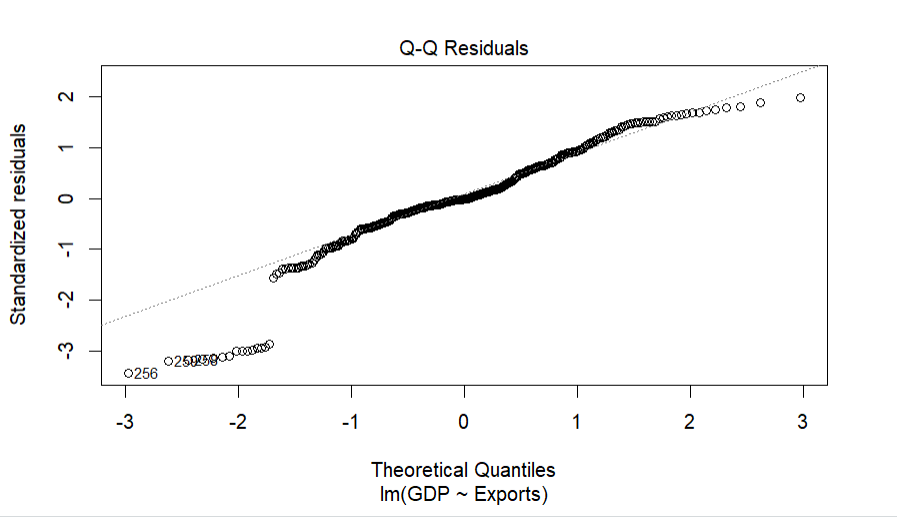
The model also showed a significant increase in R2 from 0.77 to 0.91 compared to the continent-specific model. The model also showed a significant p-value of less than 0.005, indicating a strong relationship between Exports and GDP.

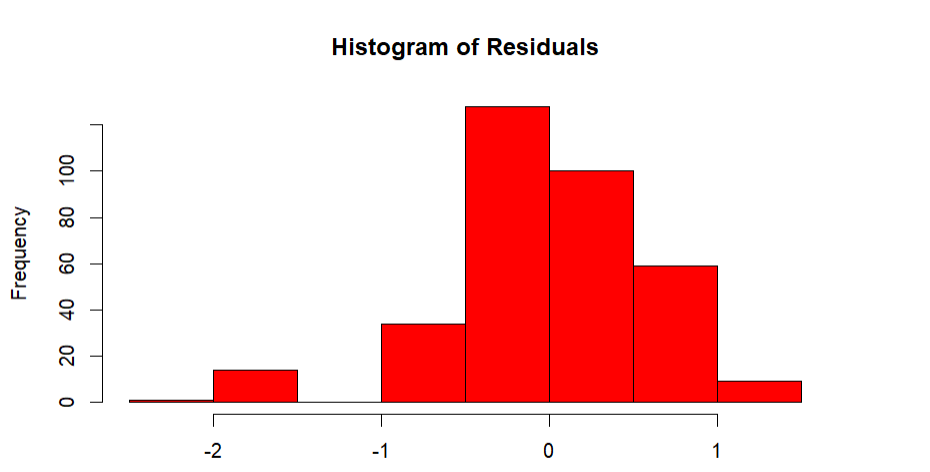
The model subsequently also fulfilled the mandatory requirements for homoscedasticity, normality, and residual independence.

Illustrated below are some of the assumption plots:

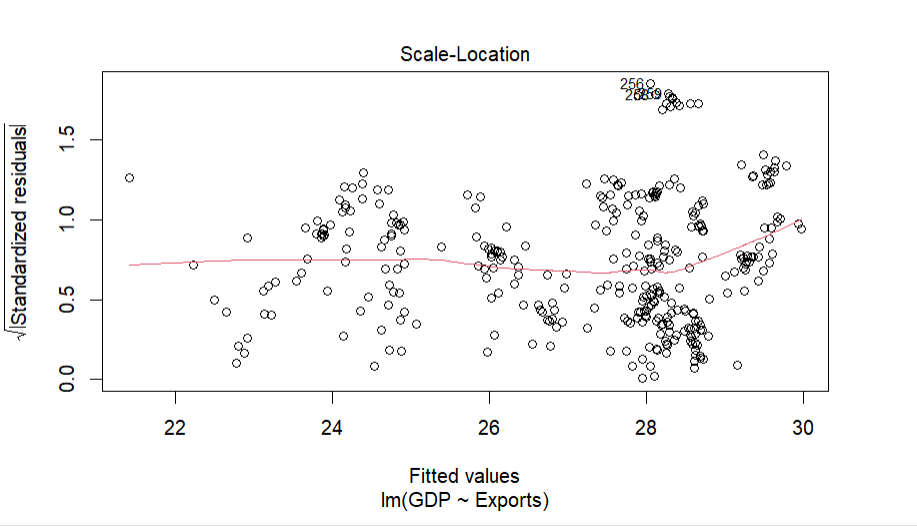


As observed, there was approximately zero correlation between the residual and fitted values, as they do not appear to have a pattern.





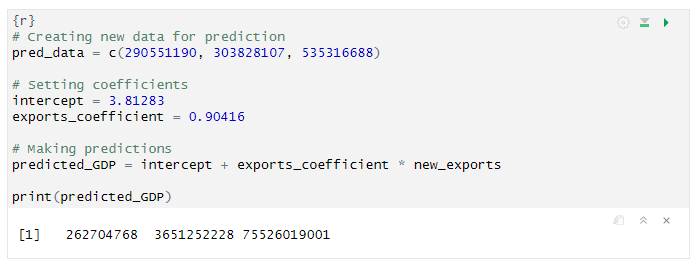
The residuals also appear to have a semblance of normality as the observations appear to be clustered around the line for the QQ plot and centred around the mean 0 on the histogram.



The Scale-Location plot also verified the Homoscedasticity assumption that the variance of the residuals was constant and not correlated. There is no clear pattern among the residuals.

**Making Predictions with the Model**

Subsequently, I made predictions based on the below equation:



Based on our model's predictions, we can observe that as export increases, the predicted GDP values also increase. This confirms the model’s effectiveness and the positive correlation between exports and GDP

##### **Multiple Linear Regression:** Predicting Economic Prosperity Using Governance Indicators

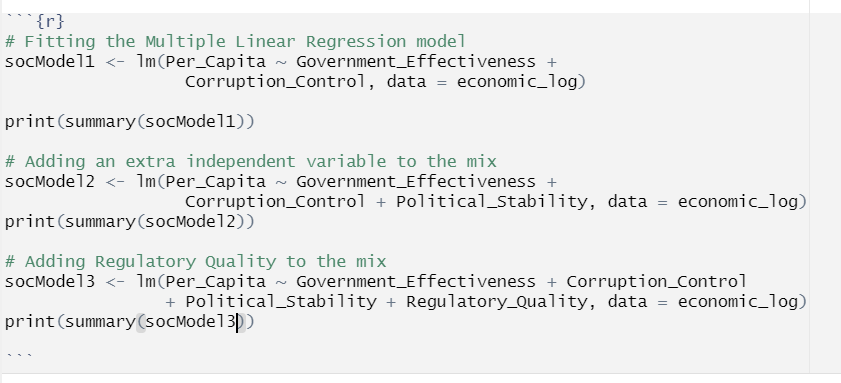
The inspiration behind this model is based on one of my major objectives, which was to examine the impact of a country’s governance on the economic prosperity of its residents.



As observed in the above pair plot and correlation matrix, there is a strong linear correlation between all the governance indicators and Per\_Capita income except for the Rule of Law.

To construct the MLR model, a Forward Stepwise approach was adopted, and we conducted up to five iterations of the MLR model, progressively adding and removing relevant governance indicators to the model.

Captured below is a snapshot of the models developed:



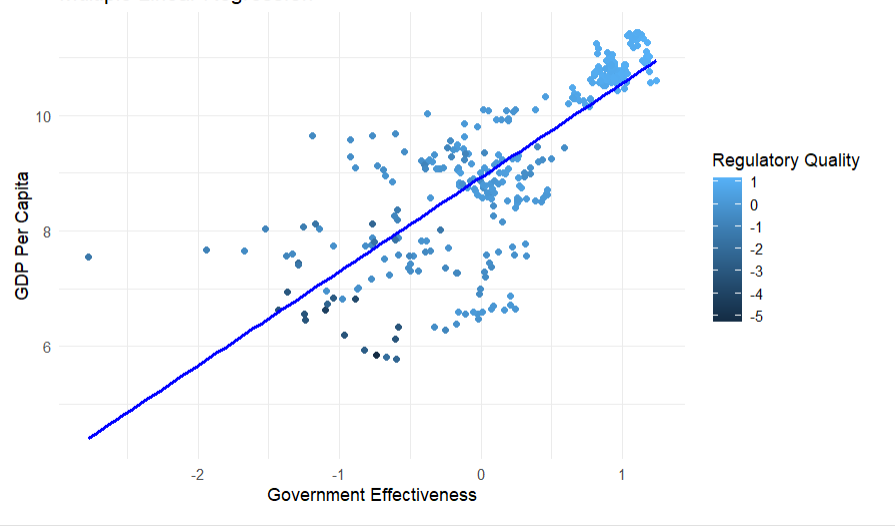
**First Model:** As a starting point, we used Government Effectiveness and Corruption Control as predictors. The results subsequently revealed a p-value less than the significance level and a coefficient of determination (R2) of 0.6388. This showed that this was a good model to build on.

**Second Model:** To strengthen the model, a new variable, Political Stability, was introduced and the results showed a slight but almost insignificant improvement to 0.6426.

**Third Model**: Subsequently, the addition of Regulatory Quality significantly improved the model showing an R2 of 0.71 and a p-value less than the significance level.

Based on the results, ‘*socModel3’* was chosen as our best model and formed the basis of our predictions.

Subsequently, we proceeded to fit our regression line and the data points appeared to be clustered around the regression line linearly.



**Assumptions Validation:**

Our assumption of linearity has already been satisfied, by the fitted regression line and also the initially constructed pairplot.

To check the other assumptions, the below plots were constructed

As can be observed, the model appears to pass the assumptions test as there is no significant correlation between the residual and fitted values. Additionally, the residuals also appear to be slightly normally distributed and the Scale-Location plot supported the assumption of Homoscedasticity showing that there is no correlation or pattern among the residuals.

**Making Predictions with the Model**

Based on our model, the regression equation is represented as below:

*Where;*

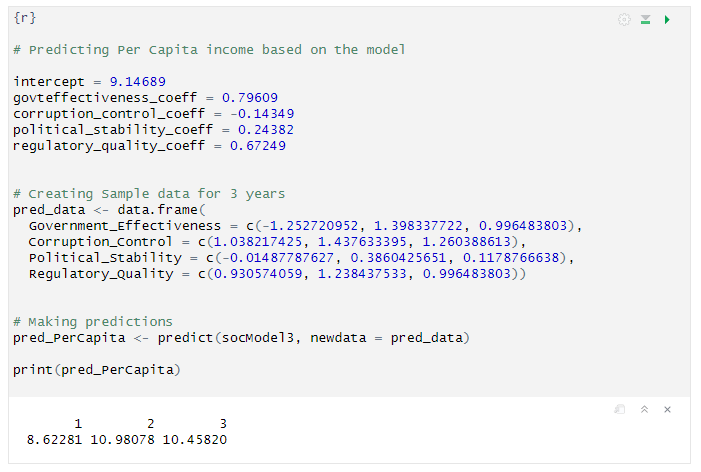
*X1 = Government Effectiveness*

*X2 = Corruption Control*

*X3 = Political Stability*

*X4 = Regulatory Quality*

Based on this equation, I proceeded to predict PerCapita for three different years on assumed values.



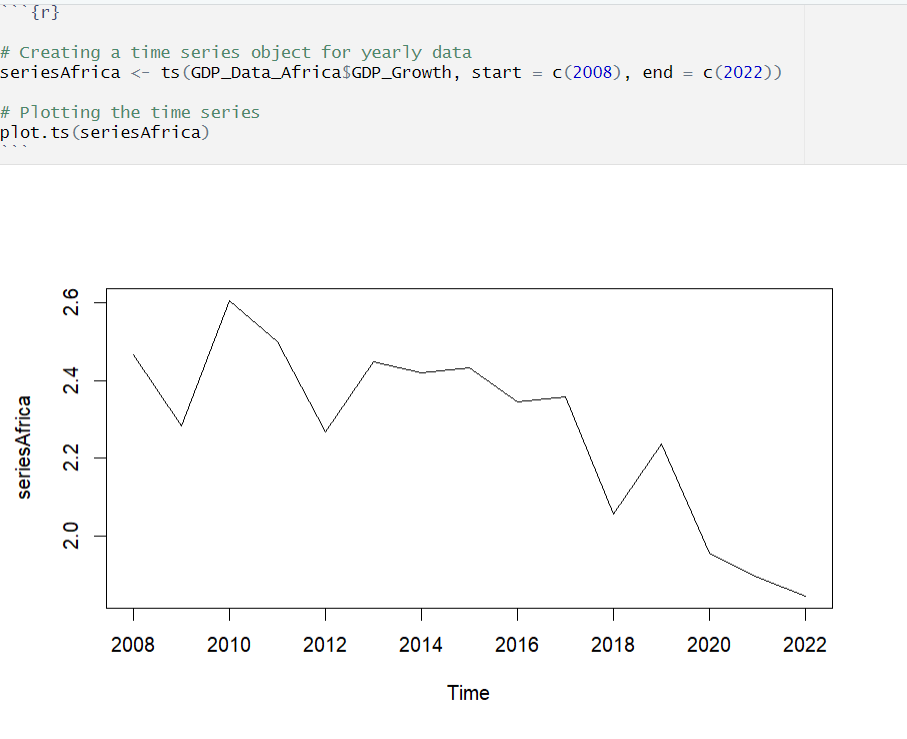
It can be observed that the model is robust and sensitive to changes because even small variations in the predictor variables led to significant differences in Per Capita income. The dependent variable tended to increase or decrease following the corresponding changes and correlation in the governance indicators.

### **7.6 Time Series Analysis**

In this phase, we conducted a time-series analysis of GDP across a selection of countries representing each continent. For granularity, our focus is narrowed to Ethiopia from Africa.

#### Forecasting Ethiopia’s GDP with Holt-Winters

We kickstarted the time-series analysis with Ethiopia and a data subset was created and converted to time-series, specifying the start and end dates.



As can be observed in the time-series plot for Ethiopia, there are discernable fluctuations, visible across the specified years, depicting periods of economic boom and downturns.

To smoothen out short-term fluctuations and correctly discern the longer-term trends, I applied SMAs with different window sizes of 3 and 8 respectively, and as we can observe in the resulting plots below, *SMA3* appears to paint a more realistic picture. We can also notably point out that the data has no discernable seasonal component.

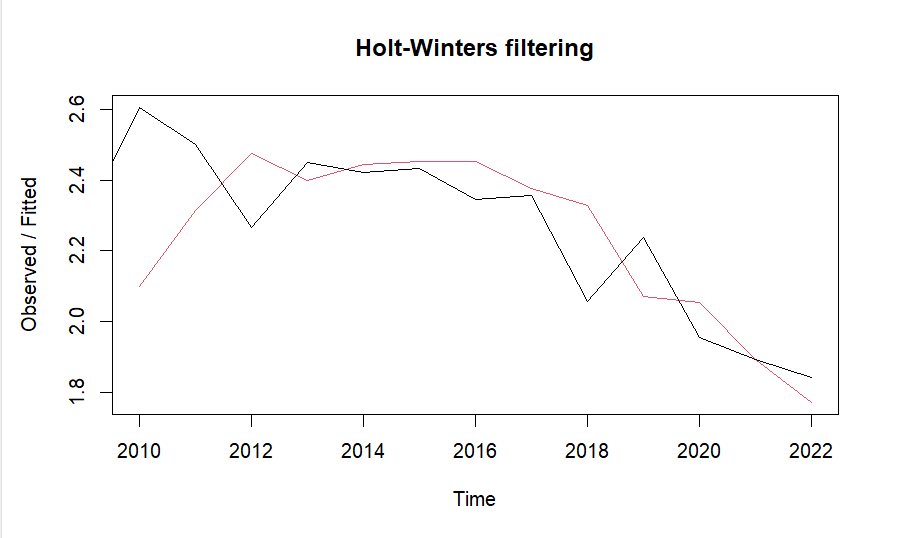


Using Holt-Winters, I proceeded to forecast the future values of Ethiopia’s GDP. The decision to use Holt-Winters is due to its suitability for datasets with no seasonal variations.



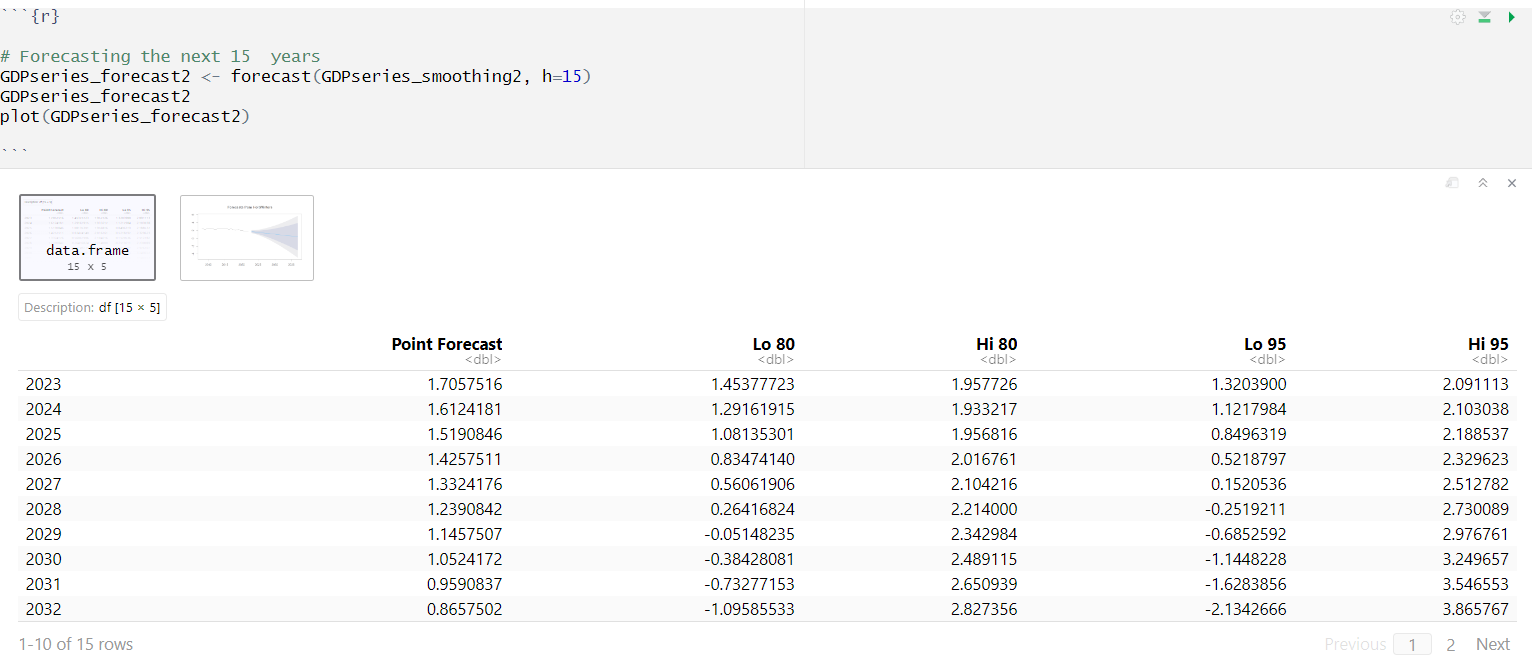
As can be observed in our result, the trend component (b) is negative and estimated as -0.09333349, indicating that the model expects the GDP for Africa to decrease slightly over time.

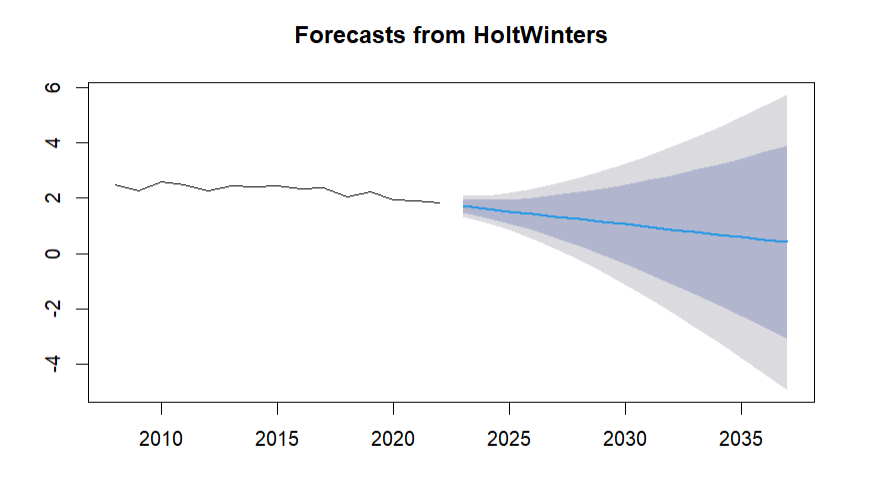
On plotting the smoothened series, we can also observe that the smoothened line in red relatively mirrors the actual trend in black.



To assess the efficiency of the forecast, I proceeded to compute the Sum of Squared Errors (SSE) for the smoothed series, and it returned a value of 0.4680147. Since this value is relatively low, we can infer that the model was a good fit.

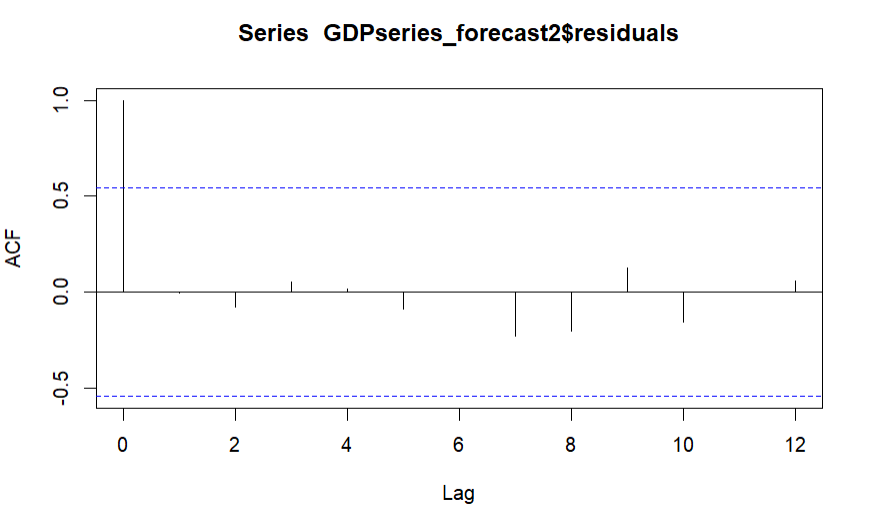
Subsequently, I forecasted 15 years into the future and below are the resulting figures at 80% and 95% confidence interval bounds.

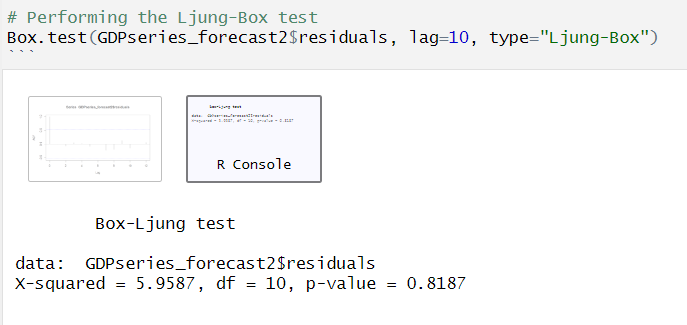




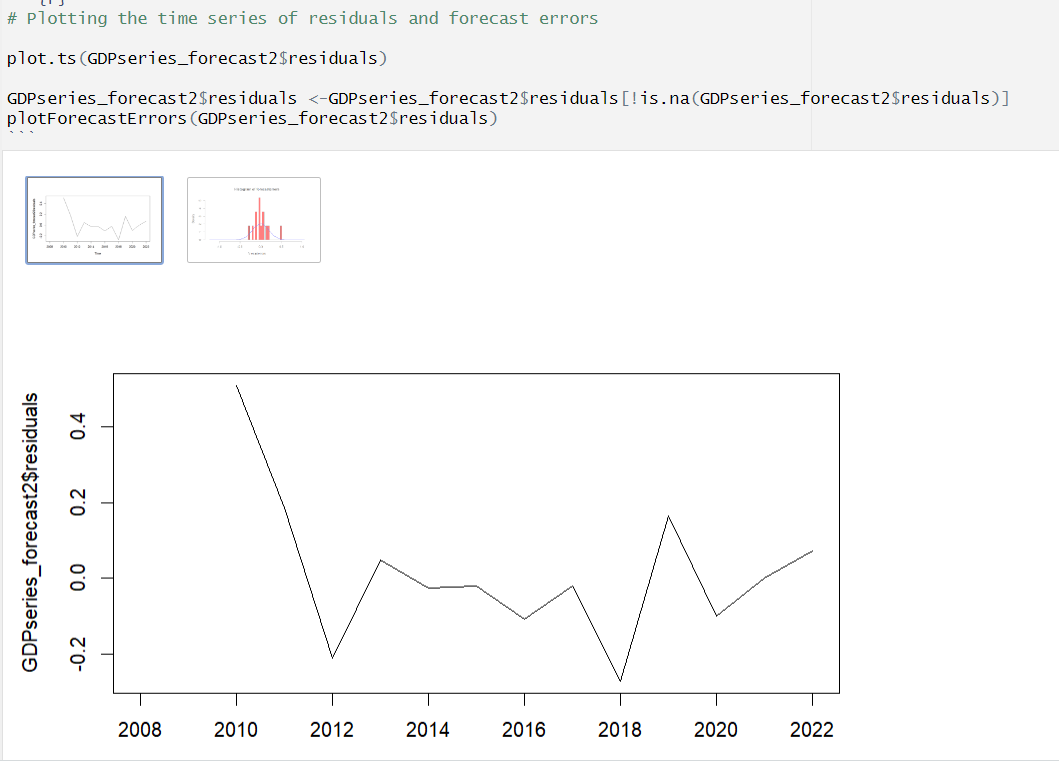
From the forecast, we can observe that the model predicts a continuous downward trend in GDP growth rates for Africa over the next 15 years. However, the widening confidence intervals suggest that although the model is fairly confident in the short term, its forecasts are uncertain for the distant future.

Using the ACF, I assessed the efficacy of the model, and the plot revealed that there was little to no autocorrelation in the residuals. Similarly, the Ljung-Box test confirmed the absence of autocorrelation in the residuals, showing an X-squared of 5.9587 and a p-value of 0.8187

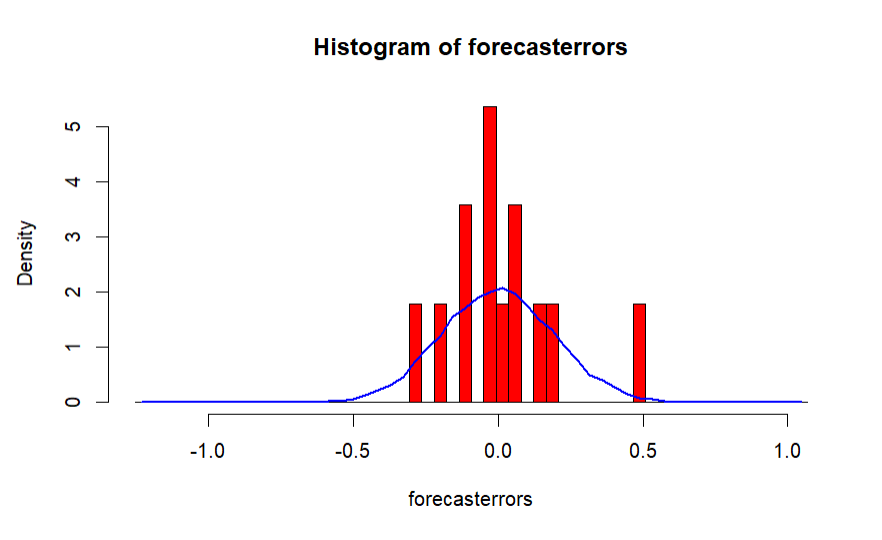




Finally, I visualised a time series plot of the residuals across the different periods and as hoped, the residuals appear not to show any systematic pattern or trend.



The histogram of forecast errors also looked roughly symmetric around zero, indicating the errors follow a slightly normal distribution.



#### Forecasting Ethiopia’s GDP with ARIMA

As a part of our analysis, We conducted two forecasts - one using the previous Smoothing method and another using ARIMA, a powerful forecast model known for its flexibility in capturing a wide range of time series behaviours.

To prepare the time-series data for ARIMA modelling, we tested it for stationarity and used differencing to achieve it. The results of the differencing are visualised below for comparison.



**Selecting the right model**

The next step was to select the appropriate autoregressive (AR) and moving average (MA) components for our ARIMA model, and this was done using the correlogram and partial correlogram to examine the differenced time series.

Visualised below are the individual plots.



We can see from the correlogram that the autocorrelation at lag1 (-0.642) exceeds the significance bounds.

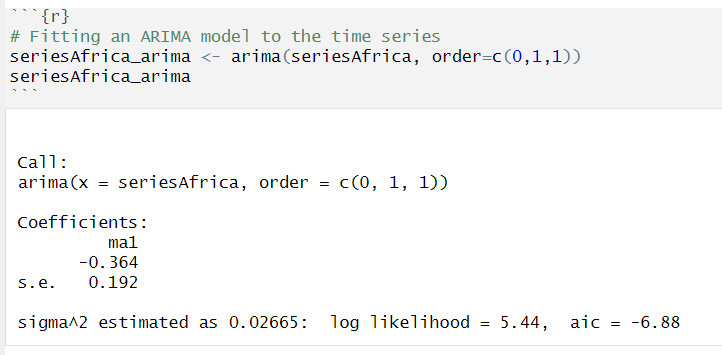


Similarly, the PACF was plotted and it showed that partial autocorrelation exceeds the significance bounds at *lags1* (-0.642), and slowly decreased in magnitude from there.

Consequently, an ARMA(0,1) model, with a moving average model of order q=1, was decided as the best model.

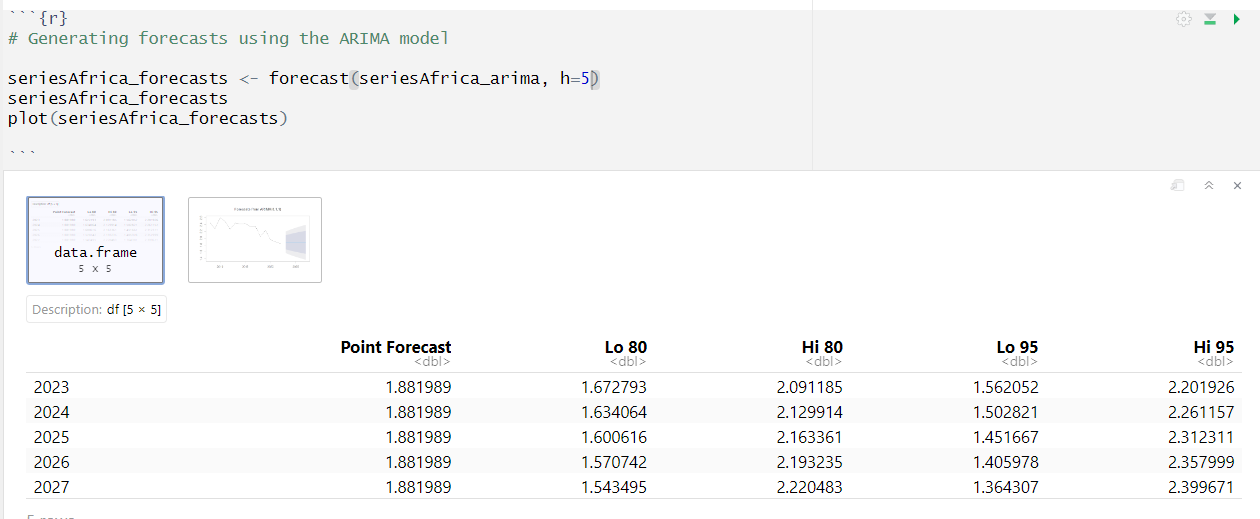
**Fitting the ARIMA Model**

The chosen Arima model was then fitted to our time-series data and the results showed a negative coefficient *(ma1)* for the first moving average term and a standard error of 0.192.



**Forecasting with the ARIMA Model**

For prediction, the ARIMA model was used to look into the future, 5 years from the present and forecast the likely GDP values for Ethiopia.



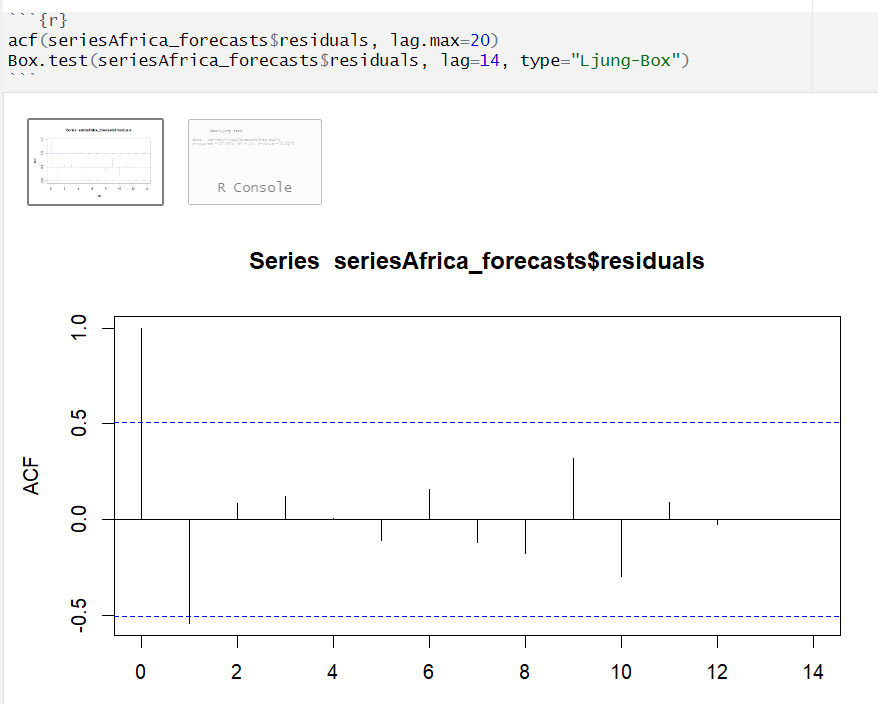


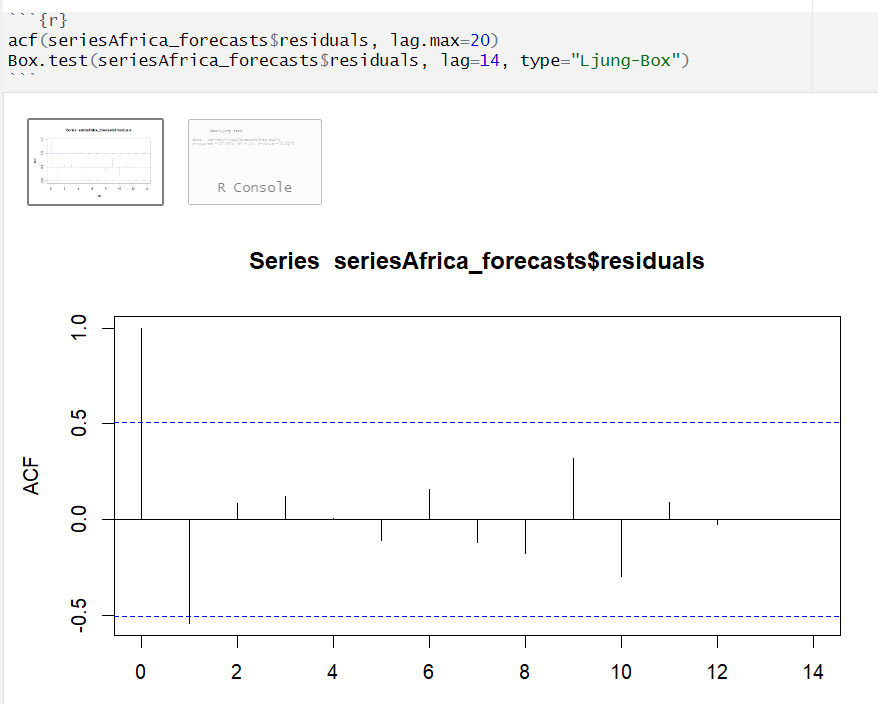
The results revealed the expected growth rates at 80% and 95% confidence intervals. From the plot, we can observe the blue flat line, implying that the model predicts the GDP growth rate to remain at a steady rate into the next 5 years.

As we moved further into the future, the confidence intervals became wider, indicating a certain level of uncertainty in the forecast. This uncertainty could be due to the potential for variation and unforeseen events that neither we nor the model is currently aware of.

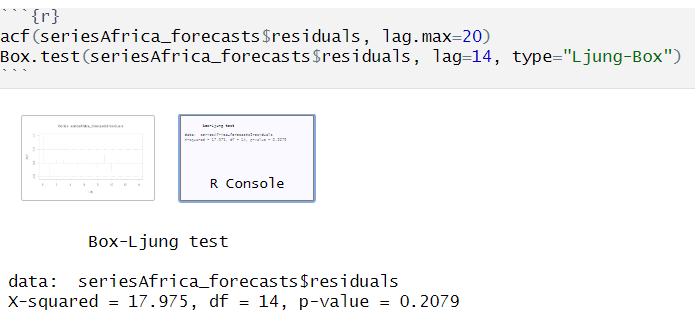
**Evaluating the ARIMA Model**

To assess the residuals of the ARIMA model, I plotted the autocorrelation function (ACF) plot and a Ljung-Box test.

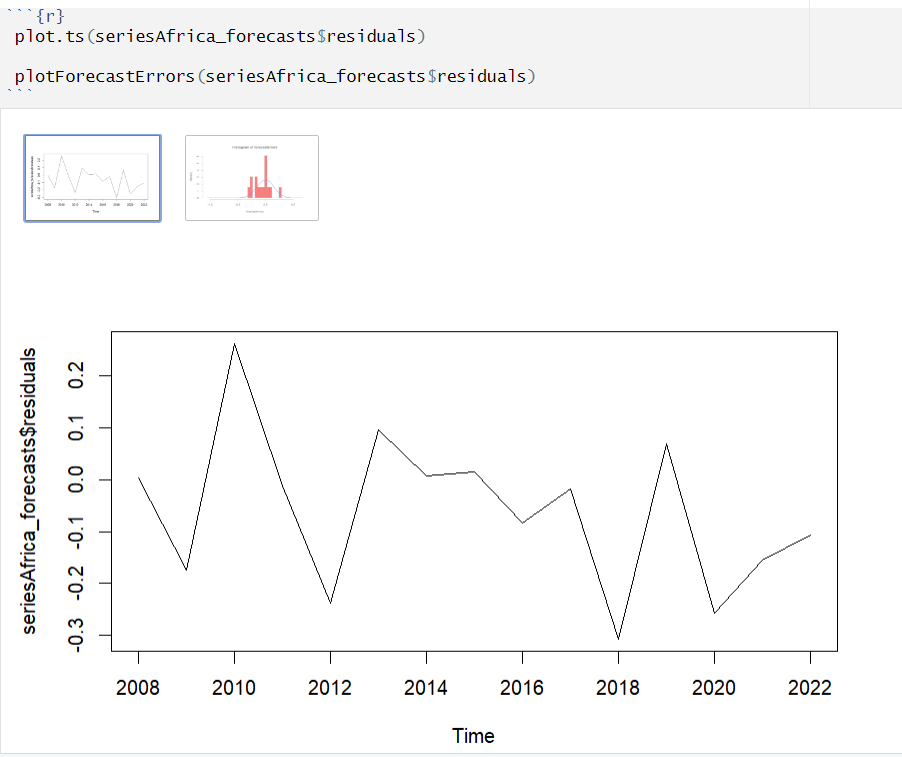




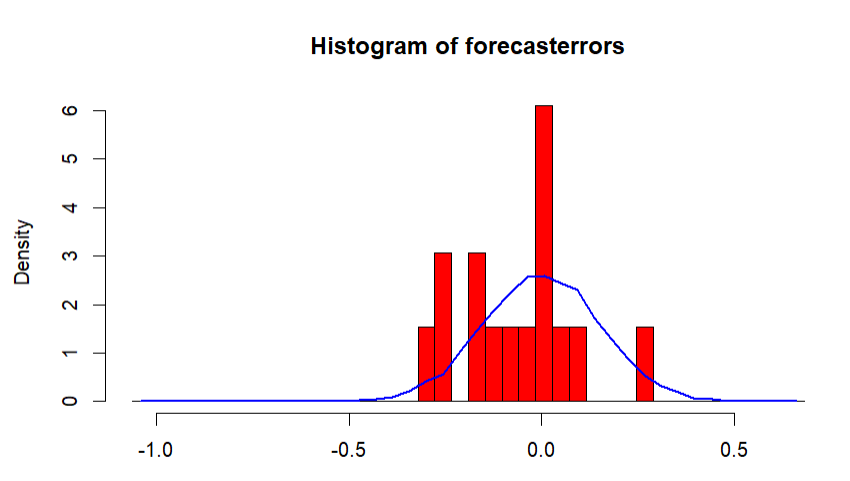
As can be observed in the plot above, except for a slight break at lag1, there were no significant spikes outside the confidence bounds. This implies that there is little evidence of a substantial autocorrelation in the residuals.



Similarly, the Ljung-Box test results showed a p-value of 0.2079 at 14 degrees of freedom, which is consistent with the results of the ACF plot. We hereby fail to reject the null hypothesis and conclude that there is no significant evidence of autocorrelation in the residuals.



Finally, we plotted a time series of the residuals and observed that the forecast errors remained roughly constant over time, with no clear consistent pattern. The histogram also showed that the forecast errors were approximately normally distributed with a mean close to zero, indicating that they may follow a normal distribution with constant variance.



# 8. **Interactive Dashboard Design**

#### 

#### Design Inspiration and Objectives

Living in the current era, data has become an integral part of our lives and data visualization has evolved from being just a means of presenting data to becoming a narrative that shapes our understanding and influences decision-making.

As part of my coursework, I designed 'The Global Economic Scorecard' dashboard, which not only presents a compelling story about the world's economic landscape from 2008 to 2022, focusing on countries from four continents.

The dashboard encapsulates a wide range of variables, providing a multifaceted view of the economic health of each continent. In this report, I discuss the visualization techniques, theoretical frameworks, and advanced features of PowerBI that were considered in designing the dashboard.

#### Application of Key Visualization Techniques

As I designed my dashboard, my primary goal was to make it easy for users to understand complex data patterns and relationships without feeling overwhelmed.

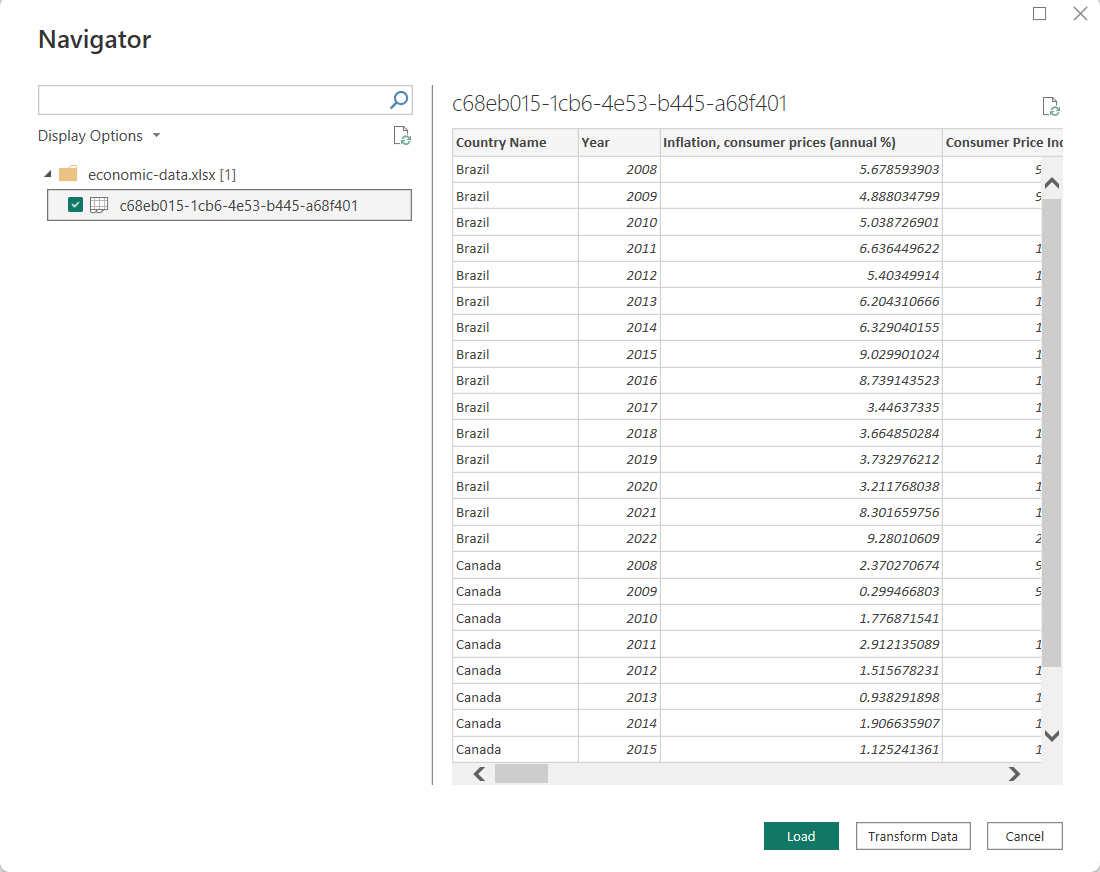
To achieve this, I used familiar visual structures and consistent colour coding. I also applied various techniques, and carefully selected simple visualizations such as line charts, bar graphs, bubble plots, and card visuals, to convey different variables. For example, I used a multi-line chart to display time series data of governance and political indicators, allowing for an easy comparison of trends over time across multiple regions. This highlighted periods of convergence and divergence in governance quality.

#### Dashboard Creation Process

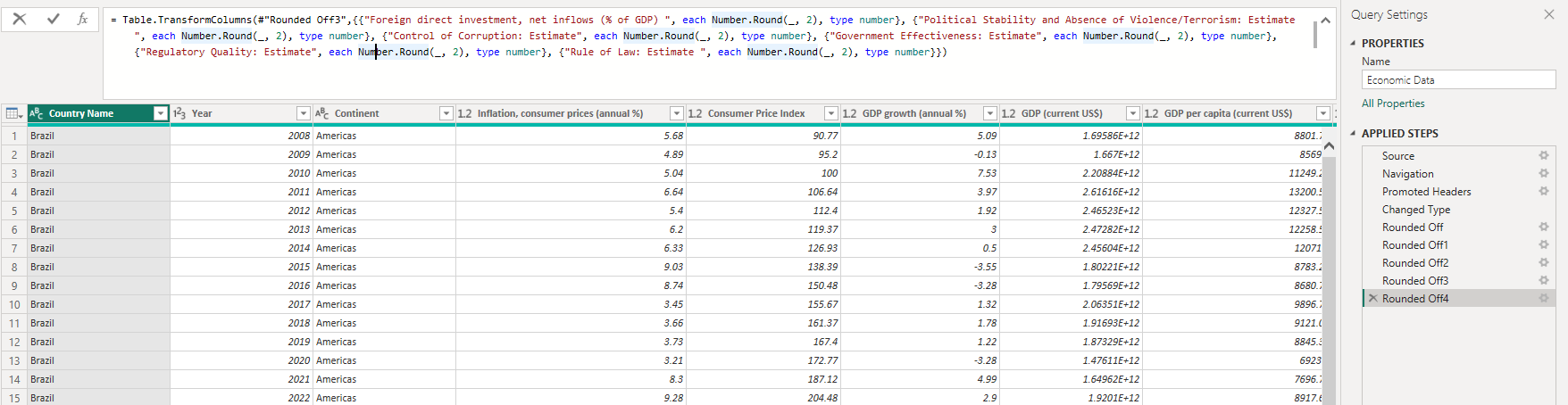
Documented in this section are the processes followed for constructing the dashboard, from data ingestion to relationship management and visual development.

**Loading the Dataset**

The process began with the ingestion of the economic indicators dataset structured across various countries and years.



Following that, the data was preprocessed and transformed by addressing missing values, normalizing data type formats, and rounding numerical figures using the 'Transform Columns' function for consistency and readability.



**Creation of Tables, Relationships and Hierarchy**

After transforming the dataset, we created new tables to better organize the data based on their natural groupings. The following tables were constructed: Economic Indices, Government Indices, Trade Indices, Year, and Country.

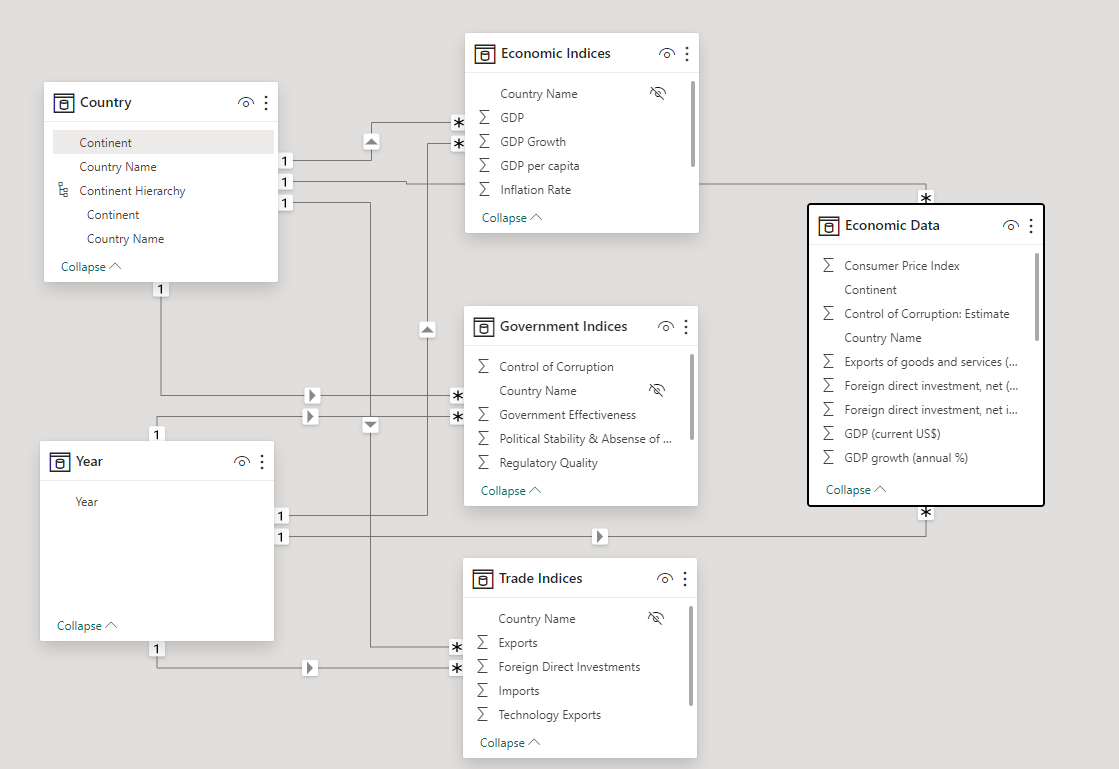
We then used DAX commands to construct the tables and renamed each column concisely for better readability.



Subsequently, a hierarchy was established on the country table to ensure that users could seamlessly navigate from continent-wide data down to individual country statistics.



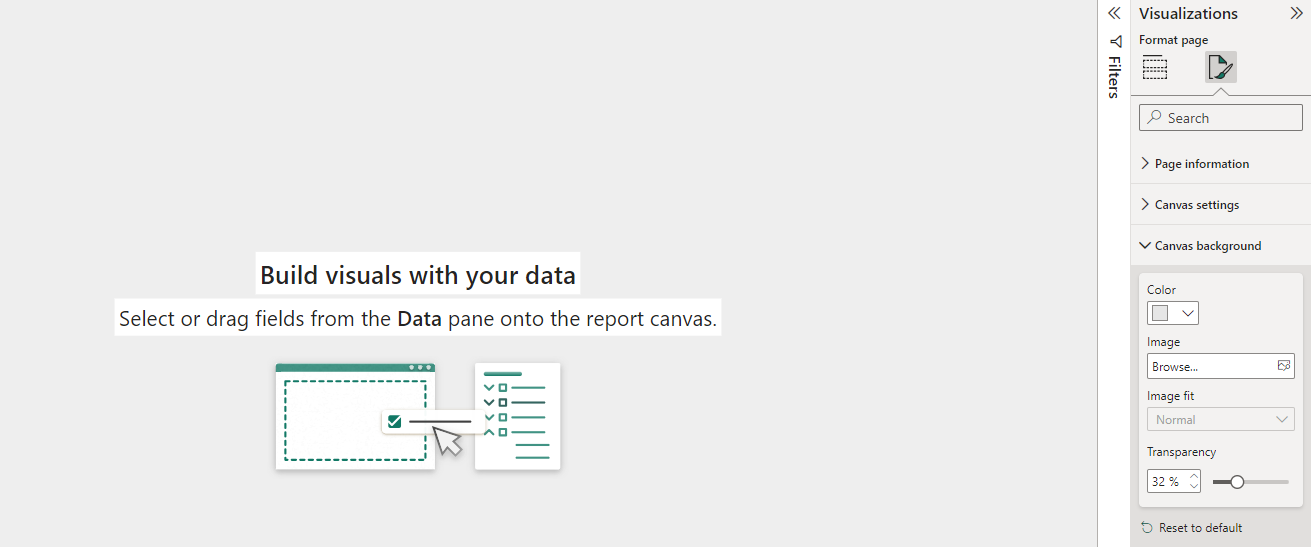
With the tables in place, the next crucial step was to create relationships that allow for the cross-referencing of data points. Each country and year was defined as a one-to-many connection, allowing for the aggregation of data at different levels of granularity.



**Canvas Preparation**

The canvas size and background were set with an appropriate size and colour ratio to ensure adequate space while maintaining a clean feel. A thematic template consistent with my objectives was then selected, along with a colour scheme that is accessible to all users, including those with colour vision deficiencies.

To structure the narrative effectively, the canvas was divided into distinct sections, each dedicated to a particular aspect of the economic data.



**Introduction of Title and Slicers**



A dashboard header and well-crafted title were then introduced with an accessible background colour. The title was chosen to indicate the dashboard’s purpose and scope.

Subsequently, a Year and Continent slicer was introduced and placed strategically in the corners, ensuring that they were easily accessible without obstructing the view of the main visualisations

**Construction of Metric Cards**



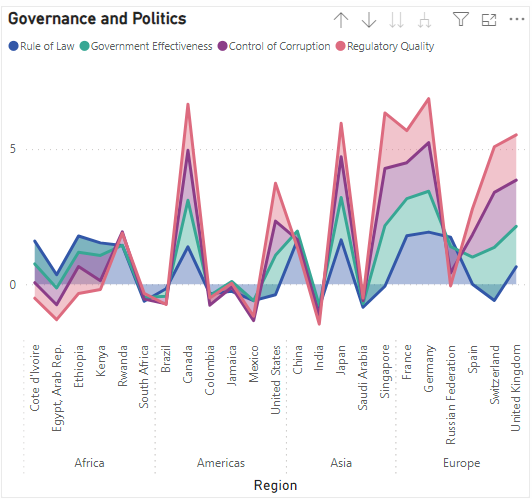
Afterwards, I created metric cards for key variables like GDP, Labor Force, Inflation, etc. Each card was associated with the corresponding data measure and formatted to display data in a user-friendly way. The cards were then collectively customized to maintain a consistent aesthetic and strategically placed at the top of the dashboard for maximum visibility.

**Visual 1:** Governance and Politics Indicators

Using an area chart, I visualized the quality of governance across different regions. This chart type was chosen to show both the individual scores of each governance indicator and the cumulative consequence when they are layered together.

To facilitate a comparative continental analysis, the countries were categorized by continents (Africa, Americas, Asia, and Europe). This categorization was crucial for the subsequent steps of visual encoding.

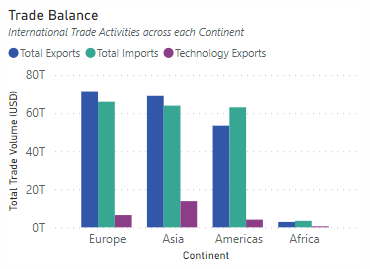
I subsequently assigned unique colours to each governance indicator and added a legend for easy recognition. Interactive tooltips were also used to provide more granular information without cluttering the primary visual.



**Visual 2:** Trade Balance

For this, I used a simple clustered bar chart to display international trade activities, side-by-side, across continents. Each bar represents a trade activity for each continent, visually coded for easy interpretation. Furthermore, an appropriate title, subtitles and labels were added to enhance clarity without overwhelming the visual.

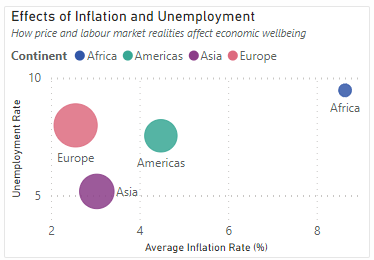
An interactive functionality with tooltips was also introduced along with a new measure that shows up whenever users hover over a bar.



**Visual 3:** Inflation vs Unemployment

A bubble chart was created to present a multidimensional view of the interplay between inflation rates, unemployment rates, and economic output across different continents.

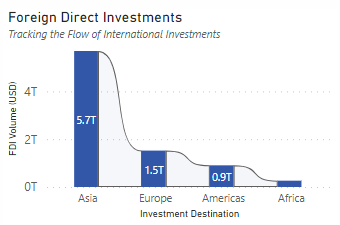
The x-axis represented the average inflation rate, the y-axis represented the average unemployment rate, and the size of the bubbles represented the Per Capita GDP across each of the continents.

Each bubble was adjusted to accurately reflect the Per Capita GDP, providing an immediate visual indication of the economic well-being of residents

**Visual 4:** Foreign Direct Investment

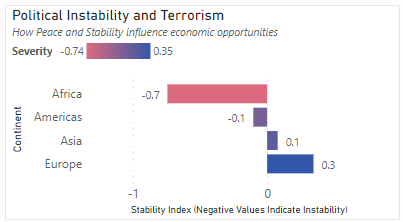
The FDI Ribbon Chart is a vivid representation of investment appeal and flow across continents. It provides a comparative analysis that goes beyond numbers to tell a story of economic allure and investor confidence.

The height of each ribbon element was deliberately encoded to leverage pre-attentive processing, allowing the viewer to gauge investment volume rapidly without conscious effort.



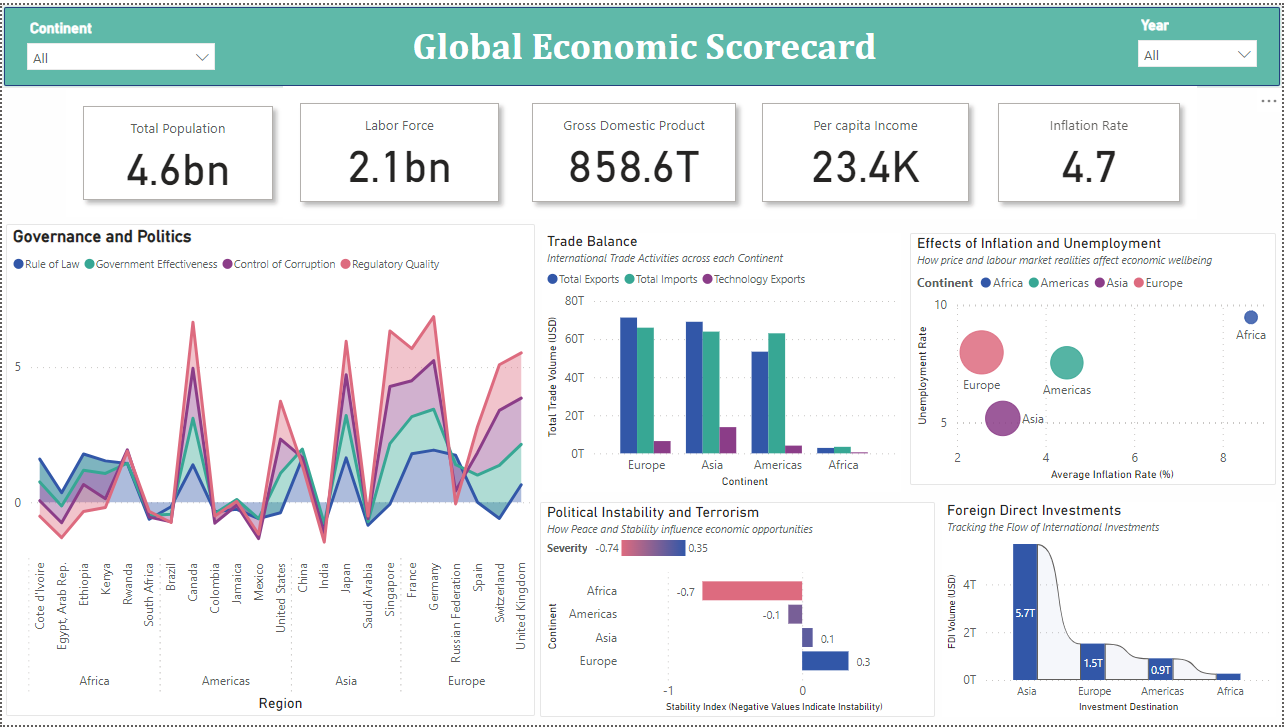
**Visual 5:** Political Instability and Terrorism

A clustered bar chart was chosen to visualize the impact of instability and terrorism on per capita income across different economic climates. The choice was because of its straightforward representation of comparative data.

The stability index, comprising positive and negative values was coded with gradient colours and used to differentiate severity levels. GDP per capita for each continent was also introduced into tooltips, providing instant context about how the index affects income.

**Final Dashboard**

The dashboard presents a comprehensive *Global Economic Scorecard* that encapsulates a wealth of data through an array of carefully constructed visualizations. Each chart serves its distinct purpose, from offering a glance at key metrics to delving deeper into the nuances of governance, trade balance, and economic health.



#### Conclusion

Throughout the design process, I followed core principles to ensure that the dashboard was user-friendly and easy to understand.

I prioritized clarity and focus, by carefully selecting simple charts to present information as clearly and efficiently as possible, avoiding cognitive overload. I used visualisations that facilitated side-by-side comparisons of different economic indicators across countries and continents.

I also considered pre-attentive processing and the effects of cognitive load in my design. I used a simple and clean dashboard layout with a logical flow, grouping related metrics and using a visual hierarchy to guide attention. I also introduced interactive elements, such as filters and drill-down functionality, to allow users to interact with the dashboard at their own pace. Similarly, I carefully maintained consistent scales, ensuring that the comparisons held statistical significance and visual integrity, without losing detail or context.

Studies like Lace Padilla (2019) showed me that when information is displayed in a way that aligns with human cognitive capabilities, the speed and accuracy of data comprehension are significantly enhanced.

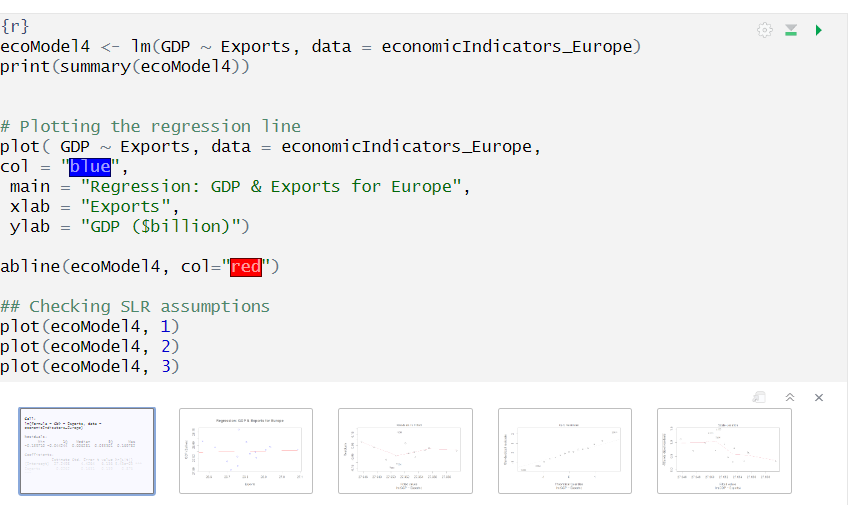
By structuring my work around these principles, I was able to ensure that each visualization constructed was accessible and comprehensible.

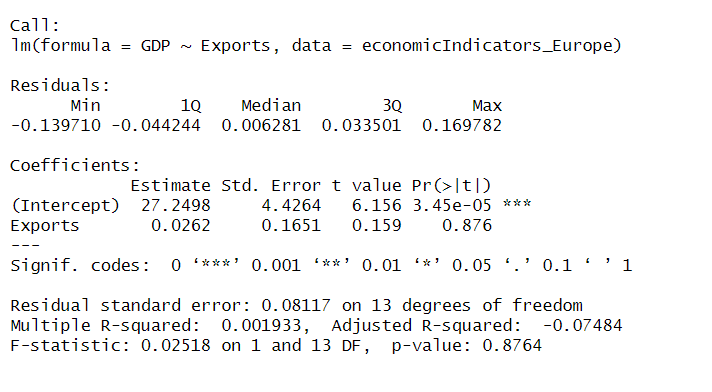
# 

# 

# **Appendix**

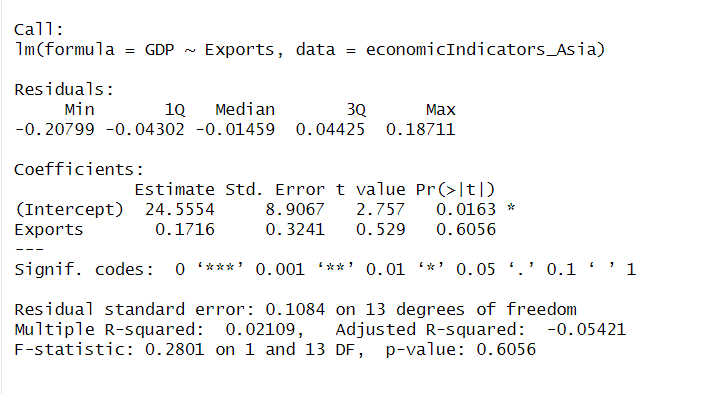
**Appendix 1: Continent-Specific SLR Model for Europe**







**Appendix 2: Continent-Specific SLR Model for Asia**





**Appendix 3: Continent-Specific SLR Model for America**

