**Macroeconomic and Social Impacts on Global Trade**

*Liam Beresteanu, Liam Goodwin, Kiana Kazemi, Shibendu Mandal*

# **Abstract**

The global macroeconomic system is the result of complex interactions between endogenous financial, social, and technological variables. In our research, we sought to draw an outline for predicting one of those variables: trade. Specifically, we looked at how variables ranging from Gini index value, GNI per capita, inflation, proportion of women in parliament, and foreign direct investment could be used to predict trade as a proportion of GDP in a set of representative countries: China, the United States, Kenya, and Norway. To address the autocorrelated nature of our time series data, we constructed and optimized both linear and polynomial regression models using differenced data points, that is values representing the change in a variable from one period to the next. Through an analysis of mean squared error on a test set of data points, we found that our linear model had the best predictive capacity. Notably, among the limited variables included in the model, changes in GNI, the proportion of women in parliament, and changes in inflation were the strongest predictors of changes in trade. While these results are not generalizable and only involve a limited set of explanatory variables, they highlight some key relationships underlying trade and lay a structural groundwork for future empirical research on global macroeconomic variables.

# **Introduction**

In this project, we set out with the intention of familiarizing ourselves with macroeconomic data and determining meaningful relationships between macroeconomic variables. Specifically, we wanted to analyze variables which we consider the most relevant to the wellbeing of individuals. We ended up focusing on GNI per capita, inflation, total trade, gender equality, and wealth equality.

Of course, we recognize that macroeconomic systems and economic relationships vary across countries. To account for this, we structured our project as a representative case study. Specifically, we focused our analysis around four distinct countries from different regions of the world: China, Kenya, Norway, and the United States.

## **Statement of Purpose**

The purpose of this paper is to explore the relationships between key macroeconomic and social variables in representative countries - China, Kenya, Norway, and the United States.

## **Research Question**

Which social and economic variables are the most important predictors of the centrality of trade in an economy?

# **Methodology**

## **Data Set**

The data set originated from the World Bank Group’s Open Data [1], pulling in 8 variables, seen below.

library(tidyverse)  
  
CountryData <- read\_csv("CleanedCountryData.csv")  
  
CountryData= CountryData %>% rename(  
 'GNI' = 'GNI per capita growth (annual %)',  
 'Inflation'= 'Inflation, consumer prices (annual %)',  
 'PropW'= 'Proportion of seats held by women in national parliaments (%)',  
 'Gini'= 'Gini index',  
 'FDI'= 'Foreign direct investment, net inflows (% of GDP)',  
 'Energy'= 'Public private partnerships investment in energy (current US$)',  
 'Imports'= 'Imports of goods and services (% of GDP)',  
 'Exports'= 'Exports of goods and services (% of GDP)') %>%   
 mutate(TotalTrade = Imports + Exports) %>%   
  
dplyr::select(Year:Country, TotalTrade, GNI:Inflation, FDI:Gini)  
ViewData=CountryData %>% filter((Year== 1960)|(Year==2022))   
  
ViewData

## # A tibble: 8 × 8  
## Year Country TotalTrade GNI Inflation FDI PropW Gini  
## <dbl> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 1960 China 8.73 NA NA NA NA NA  
## 2 2022 China 38.1 NA 1.97 1.00 24.9 NA  
## 3 1960 Kenya 64.8 NA 1.24 NA NA NA  
## 4 2022 Kenya 33.7 3.04 7.66 0.347 23.4 NA  
## 5 1960 Norway NA NA 0.360 NA NA NA  
## 6 2022 Norway 83.0 NA 5.76 0.825 45.0 NA  
## 7 1960 United States NA NA 1.46 NA NA NA  
## 8 2022 United States NA NA 8.00 1.52 28.7 NA

## **Variables**

| Short Variable Name | Original Variable Name |
| --- | --- |
| GNI | GNI per capita growth (annual %) |
| Inflation | Inflation, consumer prices (annual %) |
| PropW | Proportion of seats held by women in national parliaments (%) |
| Gini | Gini index |
| FDI | Foreign direct investment, net inflows (% of GDP) |
| Energy | Public private partnerships investment in energy (current US$) |
| Imports | Imports of goods and services (% of GDP) |
| Exports | Exports of goods and services (% of GDP) |
| TotalTrade | Imports + Exports (% of GDP) |

### Explanation of Variables

#### **GNI**

GNI per capita growth (annual %) is an indicator of the percentage change in the gross national income (GNI) per capita over a year. It is calculated by dividing the difference between the GNI per capita of two years by the GNI per capita of the earlier year and multiplying the result by 100. GNI per capita is a measure of the average income a person has at their disposal. A higher GNI per capita is often associated with higher rates of literacy and access to hygienic and pure water for the population. Overall, GNI per capita growth provides insight into the changing standard of living for the average person in each country over the years.

#### **Gini**

The Gini Index is a measure of statistical dispersion that captures the relative amount of wealth disparity in a particular country. The values range from 0 to 100, with 0 indicating perfect equality and 100 indicating absolute inequality. The summary tables below show important descriptive statistics regarding the Gini index in China, Kenya, Norway, and the United States respectively.

#### **Proportion of Women in Parliament**

The Proportion of Women in Parliament is simply the percentage of seats held by women in single or lower parliament chambers.

#### **Inflation**

Inflation occurs when prices of goods and services increase over a large period, thus causing purchasing power to decrease. This means that a country’s currency would not be able to buy as much today as it would be able to in the past.

#### **Foreign direct investment -** Excluded from Data Exploration

Foreign direct investment (FDI) represents a specific category of investment into a country’s businesses/enterprises made by individuals, companies or governments from other countries. The specific metric we are utilizing expresses the net amount (Inflows-Outflows) of FDI as a percentage of a country’s Gross Domestic Product. This metric serves as an indicator in a country’s economic openness and a dependence on foreign investment to drive growth instead of trade, the variable of interest.

*Note*: Our group decided to not do data exploration on FDI as economic intuition indicates weak correlation with total trade (Exports + Imports). We still will implement FDI into our machine models to corroborate with the studies.

#### **Public private partnerships investment in energy -** Excluded from Data Exploration

Public Private Partnerships (PPP) specifically in the energy sector represents the collaborations between public and private entities to help finance, develop, and finally manage energy base projects. These projects can encompass developing green energy projects or establishing a robust energy infrastructure to mitigate blackouts. Additionally, there may be stipulations in place such as using domestic manufacturing for materials. The metric included in our dataset measures the total PPP investment in energy in terms of current United States Dollar value.

*Note*: We decided to not include this variable in our modeling process as it is a very limited dataset. With Norway and the United States having a total of 3 observations, this would shrink our test and training data set significantly and under-represent these countries. Not to mention the high probable chance of having models overfit the training data significantly.

## **Modeling and Analysis Plan**

The goal of our analysis is to find the strongest macroeconomic/social predictor(s) for total trade in the world economy. To do this, we will first begin by reviewing our selected predictors in an exploratory environment, using time series charts and graphing in comparison to total trade. This will help by giving us an initial insight into what variables may end up influencing total trade to the greatest extent. Next, we plan to execute a further analysis of the relationship between these predictors and total trade through more advanced machine learning models–particularly linear and polynomial regressions. They can help us establish a more grounded conclusion on our predictors and relationships. While the conclusion may not be completely distinct, it will provide stronger evidence of what may be most influential towards total trade in a nation.

Overall, this comprehensive analysis plan serves as the guide for extracting meaningful information into the relationships regarding trade and lays the groundwork for our interpretations and conclusions in the research paper.

# **Results**

## **Exploratory Data Analysis**

### Variables by Country over Time

#### **Gini Index Value**

#Plotting the Gini Index  
YearvGini= CountryData %>% filter(!is.na(Year) & !is.na(Gini))  
ggplot(data = CountryData, mapping = aes(x = Year,y = Gini)) + geom\_point()+geom\_smooth(se=FALSE, aes(color=Country))+  
facet\_wrap(~Country)+ labs(x= 'Year',y= 'Gini Index Value', title='Year vs Gini Index', subtitle= 'How has Wealth Disparity Changed Over Time?' )

A graph of different countries/regions

Description automatically generated

#### **GNI**

YearvGNI= CountryData %>% filter(!is.na(Year) & !is.na(GNI))  
ggplot(data = CountryData, mapping = aes(x = Year,y = GNI)) + geom\_point()+geom\_smooth(se=FALSE, aes(color=Country))+  
facet\_wrap(~Country)+ labs(x= 'Year',y= 'GNI per capita growth (annual %)', title='Year vs GNI', subtitle= 'How has Income Changed Over Time?' )

A graph of different countries/regions

Description automatically generated

#### 

#### **Proportion of Women in National Parliaments**

YearvWomen= CountryData %>% filter(!is.na(Year) & !is.na(PropW))  
ggplot(data = CountryData, mapping = aes(x = Year,y = PropW)) + geom\_point()+geom\_smooth(se=FALSE, aes(color=Country))+  
facet\_wrap(~Country)+ labs(x= 'Year',y= 'Proportion of Women in National Parliaments', title='Year vs Women in Parliament', subtitle= 'How has Gender Equality Changed Over Time?' )

A graph with different colored lines

Description automatically generated

#### 

#### **Inflation**

YearvInflation= CountryData %>% filter(!is.na(Year) & !is.na(Inflation))  
ggplot(data = CountryData, mapping = aes(x = Year,y = Inflation)) + geom\_point()+geom\_smooth(se=FALSE, aes(color=Country))+  
facet\_wrap(~Country)+ labs(x= 'Year',y= 'Inflation (annual %)', title='Year vs Inflation Rate', subtitle= 'How has Inflation Changed Over Time?')

A graph showing the difference between inflation and the united states

Description automatically generated

#### 

#### **Total Trade (Imports + Exports)**

YearvTotalTrade= CountryData %>% filter(!is.na(Year) & !is.na(TotalTrade))  
ggplot(data = CountryData, mapping = aes(x = Year,y = TotalTrade)) + geom\_point()+geom\_smooth(se=FALSE, aes(color=Country))+  
facet\_wrap(~Country)+ labs(x= 'Year',y= 'Total Trade of Goods and Services (% GDP)', title='Year vs Trade', subtitle= 'How has Trade Changed Over Time?')

A graph of different countries/regions

Description automatically generated

### **Relationships**

NorwayData <- CountryData %>%   
 filter(Country == "Norway")   
ChinaData <- CountryData %>%   
 filter(Country == "China")  
KenyaData <- CountryData %>%   
 filter(Country == "Kenya")  
USAData <- CountryData %>%   
 filter(Country == "United States")  
  
CountryData <- CountryData %>%   
 add\_column('TradeGrowth'=c(NA,diff(ChinaData$TotalTrade),NA,diff(KenyaData$TotalTrade),  
 NA,diff(NorwayData$TotalTrade),NA,diff(USAData$TotalTrade)))

#### Variables

***GNI***

To analyze GNI, we had to include a new variable, TradeGrowth. Since the GNI is depicted as % change annually, we want the response variable also to depict change over TotalTrade by finding change over the year.

##### *Summary Statistics*

GNISummary <- CountryData %>%   
 group\_by(Country) %>%   
 summarise(  
 Min\_GNI = min(GNI, na.rm = TRUE),  
 Max\_GNI = max(GNI, na.rm = TRUE),  
 Mean\_GNI = mean(GNI, na.rm = TRUE),  
 Median\_GNI = median(GNI, na.rm = TRUE),  
 StdDev\_GNI = sd(GNI, na.rm = TRUE)  
 )  
# Show the summary statistics  
print(GNISummary)

## # A tibble: 4 × 6  
## Country Min\_GNI Max\_GNI Mean\_GNI Median\_GNI StdDev\_GNI  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 China 1.47 14.1 8.03 7.96 2.42  
## 2 Kenya -12.5 17.9 1.55 1.15 4.67  
## 3 Norway -9.74 16.2 2.55 2.46 4.46  
## 4 United States -3.48 7.03 1.80 2.19 2.22

##### *Further Analysis*

When examining the time series graph alongside Gross National Income (GNI), it becomes evident that there is no discernible and consistent pattern in the GNI’s evolution across the countries we surveyed. This raises inquiries regarding the factors responsible for, or at least capable of forecasting, shifts in wealth inequality. The graph below delves into the role of trade as a potential predictor for GNI.

GNI\_Compare <- ggplot(CountryData, aes(x = GNI, y = TradeGrowth, color = Country)) +   
 geom\_smooth(se = TRUE, method = 'lm') + geom\_point(aes(shape=Country),color="black",size=0.75) +   
 labs(x= 'GNI per Capita Growth (annual %)',y= 'Total Trade Growth (annual %)',  
 title='GNI vs Total Trade Growth', subtitle= 'Can you predict Change in Total Trade from GNI?' )  
GNI\_Compare

A graph with black dots and lines

Description automatically generated

The graph displayed above highlights a distinct positive correlation between trade and GNI in China, Kenya, and the United States. It also indicates a somewhat weaker positive association in Norway, where the trend line remains relatively flat. These findings justify the need for additional investigation into the predictive role of trade in Gross National Income.

***Gini Index: Trade and Wealth Disparities***

##### *Summary Statistics*

GiniSummary <- CountryData %>%   
 group\_by(Country) %>%   
 summarise(  
 Min\_Gini = min(Gini, na.rm = TRUE),  
 Max\_Gini = max(Gini, na.rm = TRUE),  
 Mean\_Gini = mean(Gini, na.rm = TRUE),  
 Median\_Gini = median(Gini, na.rm = TRUE),  
 StdDev\_Gini = sd(Gini, na.rm = TRUE)  
 )  
  
# Show the summary statistics  
print(GiniSummary)

## # A tibble: 4 × 6  
## Country Min\_Gini Max\_Gini Mean\_Gini Median\_Gini StdDev\_Gini  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 China 32.2 43.7 39.1 38.9 3.11  
## 2 Kenya 36.2 57.5 44.0 43.1 6.95  
## 3 Norway 24.6 31.6 27.0 27.0 1.63  
## 4 United States 34.7 41.5 38.7 38.4 2.13

##### *Further Analysis*

As the above graphs reveal, there does not appear to be a consistent trend in how the Gini index changes over time among our survey of countries. This brings the question of what factors cause, or at least predict, changes in wealth inequality. The below graph explores the Gini index as a potential predictor for Total Trade.

ggplot(data = CountryData, mapping = aes(x = Gini,y = TotalTrade)) + geom\_point(aes(shape=Country)) + geom\_smooth(se=TRUE, method=lm, aes(color=Country))+ labs(x= 'Gini Index Value',y= 'Total Trade (%GDP)', title='Trade vs Gini Index', subtitle= 'The Relationship Between Trade and Wealth Disparity' )

A graph with different colored lines and dots

Description automatically generated

The above graph points at a clear positive relationship between trade and inequality in China, the United States, and Kenya as well as a very weak positive relationship between the variables in Norway. These results warrant further exploration of inequality as a predictor of trade.

***Women in Government***

As governments continue to grow, we find that the proportion of women in government can be an interesting factor towards government policies, development, and foreign relations. As such, it can be hypothesized that there is potentially a relationship between the proportion of women in government and trade. As seen below, we find that Norway has had the highest proportion of women in government, while Kenya has the least.

##### *Summary Statistics*

PropWSummary <- CountryData %>%   
 group\_by(Country) %>%   
 summarise(  
 Min\_PropW = min(PropW, na.rm = TRUE),  
 Max\_PropW = max(PropW, na.rm = TRUE),  
 Mean\_PropW = mean(PropW, na.rm = TRUE),  
 Median\_PropW = median(PropW, na.rm = TRUE),  
 StdDev\_PropW = sd(PropW, na.rm = TRUE)  
 )  
  
# Show the summary statistics  
print(PropWSummary)

## # A tibble: 4 × 6  
## Country Min\_PropW Max\_PropW Mean\_PropW Median\_PropW StdDev\_PropW  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 China 20.2 24.9 22.4 21.8 1.70  
## 2 Kenya 2.97 23.4 12.1 9.82 7.47  
## 3 Norway 35.8 45.0 38.9 39.6 2.59  
## 4 United States 11.7 28.7 18.1 16.8 4.74

##### *Further Analysis*

We find that China, Kenya, and the United States are most similar to each other in terms of their median proportion of women in government. Norway appears to be a distinct outlier, having the highest minimum, maximum, mean, and median. To get further insight on how this can relate to the openness of a country, below is a chart with a simple comparison of total trade and the proportion of seats held by women in national parliaments.

ggplot(data = CountryData, mapping = aes(x = PropW,y = TotalTrade)) + geom\_point(aes(shape=Country)) + geom\_smooth(se=TRUE, method=lm, aes(color=Country))+ labs(x= 'Proportion of seats held by women in national parliaments (%)',y= 'Total Trade (%GDP)', title='Trade vs Proprtion of Women in Government', subtitle= 'The Relationship Between Trade and Women in Government' )

A graph of different colored lines and dots

Description automatically generated

Reflecting on this graph, it is not intrinsically clear that there is any direct relationship between the proportion of women in a government, and the total trade of the nation. However, we find for countries such as China and Kenya, a lower proportion is related to greater total trade. In contrast, the US and Norway display positive relationships – as the proportion of women increases, so does total trade. This can lead to further discussion on how the makeup of national parliaments can have implications for a nation’s trade.

***Inflation***

Below we provided some summary statistics for the four counties that we selected. The Min\_Inflation shows the year with the least inflation, if this value is negative then there was deflation (purchasing power increases) in that specific country. Max\_Inflation shows the year with the highest inflation. The Mean\_Inflation and Median\_Inflation are different measures for the average and middle years of inflation per country. Lastly the Std\_Dev\_Inflation is a measure of dispersion between the different values of inflation over all the years.

##### *Summary Statistics*

InflationSummary <- CountryData %>%   
 group\_by(Country) %>%   
 summarise(  
 Min\_Inflation = min(Inflation, na.rm = TRUE),  
 Max\_Inflation = max(Inflation, na.rm = TRUE),  
 Mean\_Inflation = mean(Inflation, na.rm = TRUE),  
 Median\_Inflation = median(Inflation, na.rm = TRUE),  
 StdDev\_Inflation = sd(Inflation, na.rm = TRUE)  
 )  
  
# Show the summary statistics  
print(InflationSummary)

## # A tibble: 4 × 6  
## Country Min\_Inflation Max\_Inflation Mean\_Inflation Median\_Inflation  
## <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 China -1.40 24.3 4.77 2.62  
## 2 Kenya -0.172 46.0 9.82 8.64  
## 3 Norway 0.360 13.6 4.42 3.25  
## 4 United States -0.356 13.5 3.76 3.02  
## # ℹ 1 more variable: StdDev\_Inflation <dbl>

##### *Further Analysis*

ggplot(data = CountryData, mapping = aes(x = TotalTrade,y = Inflation)) + geom\_point(aes(shape=Country)) + geom\_smooth(se=TRUE, method=lm, aes(color=Country))+ labs(x= 'Inflation, consumer prices (annual %)',y= 'Total Trade (%GDP)', title='Trade vs Inflation', subtitle= 'The Relationship Between Trade and Inflation in Consumer Prices' )

A graph of different colored lines and dots

Description automatically generated

In this graph, we can see the relationship between total trade and inflation for each country. Each country has a semi-clear relationship, yet due to relative differences between countries, there may be indication that inflation may not be the clearest predictor for trade.

### Exploration Conclusions

The exploration of our selected macroeconomic variables in comparison to total trade yielded mixed results. The weakest relationships were seen between inflation and GNI in comparison to total trade. We found a more convincing relationship between Gini index and trade. However, the strongest relationship was found in the proportion of women in parliament (PropW) and total trade. Thus, through implementing machine learning models, we planned to explore the relationship between trade and the predictors further in detail. Doing so would allow us to grasp a better understanding of the strength of relationships between these key socioeconomic variables and total trade.

## **Modeling**

Time series data is a unique data type due to its inherent dependence on time to gather observations. This dependence is often referred to as autocorrelation and can hinder results severely when applying the data to a traditional machine learning model such as linear regression. Some of the reason, include -

1. *Overfitting due to multicollinearity*: Correlated variables tend to provide redundant information making analyses harder to highlight and interpret variables affecting the response.
2. *Model instability*: The connecting nature of variables can make them very sensitive to changes in the model resulting in huge fluctuations.
3. *Violation of Independence Assumption*: The big assumption most models make, including linear regression, is that the data points are independent. Autocorrelation inherently breaks this assumption, leading to biased estimates.

To make sure our data lacked autocorrelation we decided to assess it by using the checkresiduals() function from the forecast library. It not only provides a visual inspection of the residuals, which could reveal patterns suggesting autocorrelation, but also provides a formal statistical test called the Ljung-Box test. In the hypothesis test, the null hypothesis is when the data is not correlated while the alternative hypothesis claims the data to be correlated. When we ran it for Gini, we noted the outliers in the residuals and a very small p-value, which allowed us to reject the null hypothesis of the Ljung-Box test. Recognizing the presence of correlation in our data, we opted to apply differentiation - an established technique in mitigating autocorrelation.

library(forecast)

## Warning: package 'forecast' was built under R version 4.3.2

## Registered S3 method overwritten by 'quantmod':  
## method from  
## as.zoo.data.frame zoo

modelData <- CountryData %>% na.omit()  
checkresiduals(modelData$Gini)

A graph of a line graph

Description automatically generated with medium confidence

##   
## Ljung-Box test  
##   
## data: Residuals  
## Q\* = 221.56, df = 10, p-value < 2.2e-16  
##   
## Model df: 0. Total lags used: 10

After dropping observations with missing data, we stratified the data by country. In each country, we added a dummy variable n to order the observations, then differenced the observations for each variable. Following that, we used rbind() to bring the data back together and again removed the new observations with missing data, leaving the differenced dataset in CountryDataDiff. We then removed the irrelevant variables. By differencing the dataset, we removed time from the data in order to address the issue of autocorrelation. Now, our observations represent changes in variables that we can use to predict a change in total trade, now represented as TTDiff.

library(broom)  
library(modelr)

##   
## Attaching package: 'modelr'

## The following object is masked from 'package:broom':  
##   
## bootstrap

CountryData=CountryData %>% drop\_na()  
glimpse(CountryData)

## Rows: 62  
## Columns: 9  
## $ Year <dbl> 1999, 2002, 2005, 2008, 2010, 2011, 2012, 2013, 2014, 2015…  
## $ Country <chr> "China", "China", "China", "China", "China", "China", "Chi…  
## $ TotalTrade <dbl> 33.52416, 42.74740, 62.20789, 57.61272, 50.71708, 50.74090…  
## $ GNI <dbl> 6.9779239, 8.7821830, 10.2735018, 9.5102843, 9.8106254, 8.…  
## $ Inflation <dbl> -1.4014727, -0.7319709, 1.7764141, 5.9252514, 3.1753248, 5…  
## $ FDI <dbl> 3.7489586, 3.6090814, 4.5542633, 3.7336107, 4.0035446, 3.7…  
## $ PropW <dbl> 21.819402, 21.782842, 20.268456, 21.325745, 21.325745, 21.…  
## $ Gini <dbl> 38.7, 42.0, 40.9, 43.0, 43.7, 42.4, 42.2, 39.7, 39.2, 38.6…  
## $ TradeGrowth <dbl> 1.09985714, 4.22004436, 2.70236865, -4.58064813, 5.5322072…

NorwayData <- CountryData %>%   
 filter(Country == "Norway")  
NorwayData$n= seq(1,18,1)  
NorwayData=NorwayData %>% mutate(GiniDiff = NorwayData$Gini[n+1]-NorwayData$Gini[n])%>% mutate(InfDiff = NorwayData$Inflation[n+1]-NorwayData$Inflation[n])%>% mutate(GNIDiff = NorwayData$GNI[n+1]-NorwayData$GNI[n])%>% mutate(PropWDiff = NorwayData$PropW[n+1]-NorwayData$PropW[n])%>% mutate(TTDiff = NorwayData$TotalTrade[n+1]-NorwayData$TotalTrade[n]) %>% mutate(FDIDiff = NorwayData$FDI[n+1]-NorwayData$FDI[n])   
NorwayData

## # A tibble: 18 × 16  
## Year Country TotalTrade GNI Inflation FDI PropW Gini TradeGrowth  
## <dbl> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 2000 Norway 74.3 13.5 3.09 4.86 36.4 27.4 4.81   
## 2 2003 Norway 66.3 0.748 2.49 1.82 36.4 27.6 -1.20   
## 3 2004 Norway 68.7 6.51 0.454 1.85 38.2 31.6 2.35   
## 4 2005 Norway 70.5 9.21 1.53 3.24 37.9 30.6 1.80   
## 5 2006 Norway 71.7 5.45 2.33 3.42 37.9 26.4 1.19   
## 6 2007 Norway 71.9 0.620 0.713 6.16 36.1 27.1 0.282   
## 7 2008 Norway 74.8 4.59 3.75 4.43 36.1 27 2.84   
## 8 2009 Norway 67.5 -9.74 2.20 2.23 39.6 26.2 -7.32   
## 9 2010 Norway 68.2 2.38 2.42 5.02 39.6 25.7 0.775   
## 10 2011 Norway 69.6 3.25 1.28 2.05 39.6 25.3 1.36   
## 11 2012 Norway 68.5 2.07 0.697 5.44 39.6 25.7 -1.10   
## 12 2013 Norway 68.0 -0.155 2.12 -0.260 39.6 26.4 -0.526   
## 13 2014 Norway 69.1 0.579 2.04 0.646 39.6 26.8 1.16   
## 14 2015 Norway 70.4 -2.99 2.17 1.87 39.6 27.5 1.29   
## 15 2016 Norway 69.1 -3.05 3.55 -5.03 39.6 28.5 -1.32   
## 16 2017 Norway 69.1 2.74 1.88 1.47 41.4 27 -0.00806  
## 17 2018 Norway 70.0 3.91 2.76 -1.29 41.4 27.6 0.894   
## 18 2019 Norway 70.9 -3.67 2.17 4.00 40.8 27.7 0.893   
## # ℹ 7 more variables: n <dbl>, GiniDiff <dbl>, InfDiff <dbl>, GNIDiff <dbl>,  
## # PropWDiff <dbl>, TTDiff <dbl>, FDIDiff <dbl>

ChinaData <- CountryData %>%   
 filter(Country == "China")  
ChinaData$n= seq(1,15,1)  
ChinaData=ChinaData %>% mutate(GiniDiff = ChinaData$Gini[n+1]-ChinaData$Gini[n])%>% mutate(InfDiff = ChinaData$Inflation[n+1]-ChinaData$Inflation[n])%>% mutate(GNIDiff = ChinaData$GNI[n+1]-ChinaData$GNI[n])%>% mutate(PropWDiff = ChinaData$PropW[n+1]-ChinaData$PropW[n])%>% mutate(TTDiff = ChinaData$TotalTrade[n+1]-ChinaData$TotalTrade[n]) %>% mutate(FDIDiff = ChinaData$FDI[n+1]-ChinaData$FDI[n])  
ChinaData

## # A tibble: 15 × 16  
## Year Country TotalTrade GNI Inflation FDI PropW Gini TradeGrowth n  
## <dbl> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 1999 China 33.5 6.98 -1.40 3.75 21.8 38.7 1.10 1  
## 2 2002 China 42.7 8.78 -0.732 3.61 21.8 42 4.22 2  
## 3 2005 China 62.2 10.3 1.78 4.55 20.3 40.9 2.70 3  
## 4 2008 China 57.6 9.51 5.93 3.73 21.3 43 -4.58 4  
## 5 2010 China 50.7 9.81 3.18 4.00 21.3 43.7 5.53 5  
## 6 2011 China 50.7 8.40 5.55 3.71 21.3 42.4 0.0238 6  
## 7 2012 China 48.3 7.89 2.62 2.83 21.3 42.2 -2.47 7  
## 8 2013 China 46.7 6.43 2.62 3.04 23.4 39.7 -1.52 8  
## 9 2014 China 44.9 7.75 1.92 2.56 23.4 39.2 -1.84 9  
## 10 2015 China 39.5 5.78 1.44 2.19 23.6 38.6 -5.44 10  
## 11 2016 China 36.9 6.22 2.00 1.56 23.6 38.5 -2.57 11  
## 12 2017 China 37.6 6.69 1.59 1.35 24.2 39.1 0.738 12  
## 13 2018 China 37.6 5.93 2.07 1.69 24.9 38.5 -0.0666 13  
## 14 2019 China 35.9 5.75 2.90 1.31 24.9 38.2 -1.68 14  
## 15 2020 China 34.8 1.47 2.42 1.72 24.9 37.1 -1.14 15  
## # ℹ 6 more variables: GiniDiff <dbl>, InfDiff <dbl>, GNIDiff <dbl>,  
## # PropWDiff <dbl>, TTDiff <dbl>, FDIDiff <dbl>

KenyaData <- CountryData %>%   
 filter(Country == "Kenya")  
KenyaData$n= seq(1,5,1)  
KenyaData=KenyaData %>% mutate(GiniDiff = KenyaData$Gini[n+1]-KenyaData$Gini[n])%>% mutate(InfDiff = KenyaData$Inflation[n+1]-KenyaData$Inflation[n])%>% mutate(GNIDiff = KenyaData$GNI[n+1]-KenyaData$GNI[n])%>% mutate(PropWDiff = KenyaData$PropW[n+1]-KenyaData$PropW[n])%>% mutate(TTDiff = KenyaData$TotalTrade[n+1]-KenyaData$TotalTrade[n]) %>% mutate(FDIDiff = KenyaData$FDI[n+1]-KenyaData$FDI[n])  
KenyaData

## # A tibble: 5 × 16  
## Year Country TotalTrade GNI Inflation FDI PropW Gini TradeGrowth n  
## <dbl> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 1997 Kenya 54.1 -1.69 11.4 0.473 2.97 45 -3.25 1  
## 2 2005 Kenya 64.5 3.58 10.3 0.113 7.14 46.5 5.00 2  
## 3 2015 Kenya 40.3 4.07 6.58 0.884 19.7 40.8 -5.84 3  
## 4 2020 Kenya 27.2 -2.39 5.40 0.424 21.8 36.2 -4.52 4  
## 5 2021 Kenya 30.7 5.57 6.11 0.422 21.6 38.7 3.45 5  
## # ℹ 6 more variables: GiniDiff <dbl>, InfDiff <dbl>, GNIDiff <dbl>,  
## # PropWDiff <dbl>, TTDiff <dbl>, FDIDiff <dbl>

USAData <- CountryData %>%   
 filter(Country == "United States")  
USAData$n= seq(1,24,1)  
USAData=USAData %>% mutate(GiniDiff = USAData$Gini[n+1]-USAData$Gini[n])%>% mutate(InfDiff = USAData$Inflation[n+1]-USAData$Inflation[n])%>% mutate(GNIDiff = USAData$GNI[n+1]-USAData$GNI[n])%>% mutate(PropWDiff = USAData$PropW[n+1]-USAData$PropW[n])%>% mutate(TTDiff = USAData$TotalTrade[n+1]-USAData$TotalTrade[n]) %>% mutate(FDIDiff = USAData$FDI[n+1]-USAData$FDI[n])  
USAData

## # A tibble: 24 × 16  
## Year Country TotalTrade GNI Inflation FDI PropW Gini TradeGrowth n  
## <dbl> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 1997 United… 23.4 3.96 2.34 1.42 11.7 40.5 0.741 1  
## 2 1998 United… 22.8 4.41 1.55 2.33 11.7 40 -0.602 2  
## 3 1999 United… 23.3 3.29 2.19 3.24 13.3 40 0.488 3  
## 4 2001 United… 23.0 0.556 2.83 1.63 14.0 40.6 -2.14 4  
## 5 2002 United… 22.3 0.465 1.59 1.02 13.8 40.4 -0.681 5  
## 6 2003 United… 22.6 1.31 2.27 1.02 14.3 40.8 0.341 6  
## 7 2004 United… 24.4 2.84 2.68 1.75 14.9 40.3 1.82 7  
## 8 2005 United… 25.6 2.33 3.39 1.09 15.2 41 1.19 8  
## 9 2006 United… 27.0 2.53 3.23 2.16 16.3 41.4 1.34 9  
## 10 2007 United… 28.0 -0.249 2.85 2.39 16.8 40.8 1.04 10  
## # ℹ 14 more rows  
## # ℹ 6 more variables: GiniDiff <dbl>, InfDiff <dbl>, GNIDiff <dbl>,  
## # PropWDiff <dbl>, TTDiff <dbl>, FDIDiff <dbl>

CountryDataDiff= rbind(NorwayData,ChinaData,KenyaData,USAData)  
CountryDataDiff= CountryDataDiff %>% drop\_na() %>% dplyr::select(Country, GiniDiff, InfDiff, GNIDiff, PropWDiff, TTDiff, FDIDiff)  
CountryDataDiff

## # A tibble: 58 × 7  
## Country GiniDiff InfDiff GNIDiff PropWDiff TTDiff FDIDiff  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 Norway 0.200 -0.598 -12.8 0 -7.99 -3.04   
## 2 Norway 4 -2.03 5.76 1.82 2.35 0.0322  
## 3 Norway -1 1.08 2.71 -0.312 1.80 1.39   
## 4 Norway -4.20 0.797 -3.77 0 1.19 0.179   
## 5 Norway 0.700 -1.62 -4.83 -1.78 0.282 2.74   
## 6 Norway -0.100 3.04 3.97 0 2.84 -1.73   
## 7 Norway -0.800 -1.56 -14.3 3.55 -7.32 -2.20   
## 8 Norway -0.5 0.222 12.1 0 0.775 2.79   
## 9 Norway -0.400 -1.13 0.866 0 1.36 -2.97   
## 10 Norway 0.400 -0.588 -1.18 0 -1.10 3.39   
## # ℹ 48 more rows

Subsequently, upon re-evaluating the variable with the Ljung-Box test, the p-value became statistically insignificant, leading us to fail to reject the null. By addressing the autocorrelation through differencing, we can now proceed with a greater amount of confidence in building a robust and reliable machine learning model.

checkresiduals(CountryDataDiff$GiniDiff)

A graph of a graph

Description automatically generated with medium confidence

##   
## Ljung-Box test  
##   
## data: Residuals  
## Q\* = 4.3856, df = 10, p-value = 0.9283  
##   
## Model df: 0. Total lags used: 10

Next, we split the data into a training set and a test so that we have a way to gauge the predictive effectiveness of our model. Notably, to account for the differences in countries and the number of data-points we have available for each, we stratified the data, ensuring an equal proportion of random observations associated with each country are represented in our training and test sets.

set.seed(100)  
train1 <- CountryDataDiff %>% group\_by(Country) %>% sample\_frac(size = 0.8)  
test1 <- CountryDataDiff %>% setdiff(train1)

1. Linear Regression Model

The first model we fit was a simple linear regression model using our differenced training data. We then optimized the model using the stepAIC() function. To test the predictive accuracy of our model, we calculated the mean squared error for the test set and compared it to the MSE of the training set.

library(MASS)

##   
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':  
##   
## select

train1$Country <- as.factor(train1$Country)  
  
model1 <- lm(TTDiff ~ GiniDiff + InfDiff + PropWDiff + GNIDiff, data = train1)  
model\_selected <- stepAIC(model1, direction = "both")

## Start: AIC=116.71  
## TTDiff ~ GiniDiff + InfDiff + PropWDiff + GNIDiff  
##   
## Df Sum of Sq RSS AIC  
## <none> 467.93 116.70  
## - InfDiff 1 34.358 502.29 117.97  
## - GiniDiff 1 40.151 508.08 118.49  
## - GNIDiff 1 68.220 536.15 120.97  
## - PropWDiff 1 106.767 574.70 124.16

summary(model1)

##   
## Call:  
## lm(formula = TTDiff ~ GiniDiff + InfDiff + PropWDiff + GNIDiff,   
## data = train1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -7.8896 -1.5496 0.3868 1.7762 11.8163   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.1676 0.5519 0.304 0.76295   
## GiniDiff 0.7424 0.3958 1.876 0.06784 .   
## InfDiff 0.6477 0.3733 1.735 0.09024 .   
## PropWDiff -0.8477 0.2771 -3.059 0.00391 \*\*  
## GNIDiff 0.2923 0.1196 2.445 0.01887 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.378 on 41 degrees of freedom  
## Multiple R-squared: 0.5194, Adjusted R-squared: 0.4725   
## F-statistic: 11.08 on 4 and 41 DF, p-value: 3.485e-06

summary(model\_selected)

##   
## Call:  
## lm(formula = TTDiff ~ GiniDiff + InfDiff + PropWDiff + GNIDiff,   
## data = train1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -7.8896 -1.5496 0.3868 1.7762 11.8163   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.1676 0.5519 0.304 0.76295   
## GiniDiff 0.7424 0.3958 1.876 0.06784 .   
## InfDiff 0.6477 0.3733 1.735 0.09024 .   
## PropWDiff -0.8477 0.2771 -3.059 0.00391 \*\*  
## GNIDiff 0.2923 0.1196 2.445 0.01887 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.378 on 41 degrees of freedom  
## Multiple R-squared: 0.5194, Adjusted R-squared: 0.4725   
## F-statistic: 11.08 on 4 and 41 DF, p-value: 3.485e-06

test1

## # A tibble: 12 × 7  
## Country GiniDiff InfDiff GNIDiff PropWDiff TTDiff FDIDiff  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 Norway 0.700 -1.62 -4.83 -1.78 0.282 2.74   
## 2 Norway -0.100 3.04 3.97 0 2.84 -1.73   
## 3 Norway 0.600 0.890 1.17 0 0.894 -2.76   
## 4 China 3.30 0.670 1.80 -0.0366 9.22 -0.140   
## 5 China -1.10 2.51 1.49 -1.51 19.5 0.945   
## 6 China -1.30 2.38 -1.41 -0.00271 0.0238 -0.295   
## 7 Kenya -4.60 -1.18 -6.46 2.06 -13.1 -0.460   
## 8 United States 0.600 0.638 -2.73 0.690 -0.346 -1.61   
## 9 United States 0.400 0.684 0.841 0.460 0.341 0.00605  
## 10 United States -0.600 2.00 5.19 0 3.46 0.642   
## 11 United States -0.300 -1.50 -0.130 0.0445 -2.19 1.37   
## 12 United States 0.100 3.46 8.89 0.185 2.10 1.46

TT\_pred = predict(model\_selected, newdata=test1)  
TT\_pred

## 1 2 3 4 5 6 7   
## -0.2660839 3.2239817 1.5305706 3.6094772 2.6953034 0.3343444 -7.6456232   
## 8 9 10 11 12   
## -0.3563554 0.7638482 2.5315106 -1.1047864 4.9275228

TT\_pred\_test = predict(model\_selected, newdata=test1)  
sse\_test <- sum((TT\_pred\_test - test1$TTDiff)^2)  
mse\_test= sse\_test/nrow(test1)  
mse\_test

## [1] 29.45028

TT\_pred\_train = predict(model\_selected, newdata=train1)  
sse\_train = sum((TT\_pred\_train - train1$TTDiff)^2)  
mse\_train= sse\_train/nrow(train1)  
mse\_train

## [1] 10.17239

1. Polynomial Regression Model

The second model we fit was a polynomial regression model using our differenced training data. We then optimized the model using the stepAIC() function. In order to test the predictive accuracy of our model, we calculated the mean squared error for the the test set and compared it to the mse of the training set.

model\_poly <- lm(TTDiff ~ poly(GiniDiff, 2) + poly(InfDiff, 2) + poly(PropWDiff,2) + poly(GNIDiff, 2), data = train1)  
  
model\_selected\_poly <- stepAIC(model\_poly, direction = "both")

## Start: AIC=87.92  
## TTDiff ~ poly(GiniDiff, 2) + poly(InfDiff, 2) + poly(PropWDiff,   
## 2) + poly(GNIDiff, 2)  
##   
## Df Sum of Sq RSS AIC  
## - poly(GiniDiff, 2) 2 16.588 226.93 87.415  
## <none> 210.34 87.924  
## - poly(InfDiff, 2) 2 92.528 302.87 100.694  
## - poly(GNIDiff, 2) 2 99.528 309.87 101.745  
## - poly(PropWDiff, 2) 2 118.926 329.27 104.538  
##   
## Step: AIC=87.42  
## TTDiff ~ poly(InfDiff, 2) + poly(PropWDiff, 2) + poly(GNIDiff,   
## 2)  
##   
## Df Sum of Sq RSS AIC  
## <none> 226.93 87.415  
## + poly(GiniDiff, 2) 2 16.588 210.34 87.924  
## - poly(InfDiff, 2) 2 79.510 306.44 97.233  
## - poly(GNIDiff, 2) 2 132.747 359.67 104.602  
## - poly(PropWDiff, 2) 2 277.151 504.08 120.128

summary(model\_selected\_poly)

##   
## Call:  
## lm(formula = TTDiff ~ poly(InfDiff, 2) + poly(PropWDiff, 2) +   
## poly(GNIDiff, 2), data = train1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.9548 -1.2177 -0.2322 1.0804 8.3676   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.0110 0.3557 -2.843 0.00709 \*\*   
## poly(InfDiff, 2)1 5.2872 2.6421 2.001 0.05237 .   
## poly(InfDiff, 2)2 -8.6065 2.5908 -3.322 0.00195 \*\*   
## poly(PropWDiff, 2)1 -13.7341 2.6696 -5.145 7.93e-06 \*\*\*  
## poly(PropWDiff, 2)2 -13.2219 2.5682 -5.148 7.83e-06 \*\*\*  
## poly(GNIDiff, 2)1 11.1384 2.4164 4.609 4.26e-05 \*\*\*  
## poly(GNIDiff, 2)2 -3.2603 2.4886 -1.310 0.19783   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.412 on 39 degrees of freedom  
## Multiple R-squared: 0.7669, Adjusted R-squared: 0.7311   
## F-statistic: 21.39 on 6 and 39 DF, p-value: 6.181e-11

TT\_predpoly = predict(model\_selected\_poly, newdata=test1)

TT\_predpoly = predict(model\_selected\_poly, newdata=test1)  
ssepoly\_test <- sum((TT\_predpoly - test1$TTDiff)^2)  
msepoly\_test=ssepoly\_test/nrow(test1)  
msepoly\_test

## [1] 63.10175

TT\_predpoly\_train = predict(model\_selected\_poly, newdata=train1)  
ssepoly\_train = sum((TT\_predpoly\_train - train1$TTDiff)^2)  
msepoly\_train= ssepoly\_train/nrow(train1)  
msepoly\_train

## [1] 4.933201

***Polynomial Graphs***

plot\_graph <- function(data, var, title) {  
 model <- lm(TTDiff ~ poly(InfDiff, 2) + poly(PropWDiff, 2) + poly(GNIDiff, 2), data = data)  
 coefficients <- coef(model)  
  
 min\_value <- min(data[[var]], na.rm = TRUE)  
 max\_value <- max(data[[var]], na.rm = TRUE)  
 x <- seq(min\_value, max\_value, by = 0.1)  
 y <- coefficients[paste('poly(', var, ', 2)2', sep = '')] \* x^2 +   
 coefficients[paste('poly(', var, ', 2)1', sep = '')] \* x -   
 coefficients['(Intercept)']  
 data\_plot <- data.frame(x, y)  
 ggplot(data\_plot, aes(x, y)) +   
 geom\_line() +   
 ggtitle(title) +  
 xlab(var) +   
 ylab("y")  
}  
  
# Plot for InfDiff  
plot\_graph(CountryDataDiff, "InfDiff", "Impact of InfDiff on TTDiff")

A graph with a line

Description automatically generated

# Plot for PropWDiff  
plot\_graph(CountryDataDiff, "PropWDiff", "Impact of PropWDiff on TTDiff")

A graph with a curve

Description automatically generated

# Plot for GNIDiff  
plot\_graph(CountryDataDiff, "GNIDiff", "Impact of GNIDiff on TTDiff")

A graph with a line

Description automatically generated

## **Interpretation of Results**

To compare the models, we first looked at the training mean squared error (MSE) values. For the linear regression model, MSE was approximately 10.17, and for the polynomial model it was approximately 4.93. MSE tells us the average squared difference between the predicted value of TTDiff and the actual value; thus, it conveys how well the models fit the dataset that they were trained with. Because the polynomial model has a lower MSE, it better fits the training data. However, regarding the test set MSE values, the linear model was lower at 29.45028 and while the polynomial model was higher at 63.10175. This disparity conveys that the linear model was a better predictor for the test set despite not fitting the training set as well as the polynomial model, indicating that the latter was overfit. Looking inside the linear model, all of our predictors were significant at least at the ten percent level. GiniDiff, InfDiff, and GNIDiff had positive estimated coefficients of .7424, .6477, and .2923 respectively, while PropWDiff had an estimated coefficient of -.8477. This indicates that, among our selected countries, as Gini coefficients increase, Inflation growth accelerates and GNI per capita growth accelerates, trade as a percentage of GDP will also tend to increase. Conversely, as the proportion of women in national parliaments increases, trade as a percentage of GDP will tend to fall.

In our study, we employed a polynomial regression model to elucidate the relationships between several macroeconomic and social variables and their predictive power on the centrality of trade in an economy. Polynomial regression, an extension of linear regression, allows us to model the relationship between the independent variable(s) and the dependent variable as an nth degree polynomial. This approach is particularly useful in capturing the non-linear relationship often observed in economic data. The polynomial model can be represented by the equation y=ax2+bx+c. y is the dependent variable (in our case, the centrality of trade), x represents the independent variable, a and b are coefficients for the second and first degree of the polynomial respectively, and c is the intercept. In this model, the coefficients are estimated using a least squares approach, optimizing the fit of the polynomial curve to the observed data.

1. Empirical Findings

***Impact of Inflation Differential (InfDiff) on Total Trade Differential (TTDiff)***

1. *Graphical Analysis:* The polynomial plot indicates a peak slightly above zero.
2. *Economic Interpretation:* This aligns with conventional economic theory, where positive inflation is beneficial for exporters but bad for importers because of favorable/unfavorable exchange rates. The opposite is true when the change in inflation is negative. Therefore, the “sweet spot” is zero, where neither importers or exporters are harmed.

***Impact of Gross National Income Differential (GNIDiff) on TTDiff***

1. *Graphical Analysis:* The curve exhibits a maximum slightly above zero.
2. *Economic Interpretation:* A modest year-over-year increase in GNI positively influences trade. However, an excessively rapid GNI growth can trigger inflation, speculative behaviors in financial markets, and resource constraints, ultimately impairing trade. This explains the tapering of the curve beyond a certain point.

***Impact of Proportion of Women Differential (PropWDiff) on TTDiff***

1. *Graphical Analysis*: The curve suggests a negative impact of increasing PropW on trade.
2. *Economic Interpretation:* We hypothesize this as a case of omitted variable bias. PropW might be acting as a proxy for a variable not included in our model, possibly reflecting a measure of equality. The data might indicate that countries with higher equality levels engage less in trade, though further investigation is required to confirm this hypothesis.

## **Post-Hoc Analysis**

1. Removing Outliers

Since our MSE even in our best model, the linear model, was high, we decided to further inquire as to why it is through looking at our residuals for specific points and see anomalous factors occurring. To summarize the residual graph above, we found that two data points had extremely high residuals when we plotted the residuals of our best model, Linear Regression. One of these points is from Kenya and has a residual difference close to 280. This difference is likely due to the fact that the data point is from a period between 2005 and 2015, which is a 10-year time gap. The other point is from China and has a residual difference of 150, which is due to a 4-year time gap. Our group believes that these data points should be considered outliers and excluded from the dataset due to the large time gaps in the data.

TT\_train = predict(model\_selected, newdata=train1)  
resid <- c(round((TT\_pred - test1$TTDiff)^2,1), round((TT\_train - train1$TTDiff)^2,1))  
plot(resid)

A graph with numbers and dots

Description automatically generated

***Effects on Models***

set.seed(100)   
CountryDataDiff1 <- CountryDataDiff %>% filter(TTDiff>-20,TTDiff<18)  
train2 <- CountryDataDiff1 %>% group\_by(Country) %>% sample\_frac(size = 0.8)  
test2 <- CountryDataDiff1 %>% setdiff(train2)

train2$Country <- as.factor(train2$Country)  
  
model2 <- lm(TTDiff ~ GiniDiff + InfDiff + PropWDiff + GNIDiff + FDIDiff, data = train2)  
model\_selected1 <- stepAIC(model2, direction = "both")

## Start: AIC=86.04  
## TTDiff ~ GiniDiff + InfDiff + PropWDiff + GNIDiff + FDIDiff  
##   
## Df Sum of Sq RSS AIC  
## - PropWDiff 1 3.284 240.01 84.645  
## <none> 236.72 86.039  
## - GiniDiff 1 17.262 253.99 87.136  
## - FDIDiff 1 20.322 257.05 87.663  
## - InfDiff 1 34.448 271.17 90.017  
## - GNIDiff 1 46.783 283.51 91.974  
##   
## Step: AIC=84.65  
## TTDiff ~ GiniDiff + InfDiff + GNIDiff + FDIDiff  
##   
## Df Sum of Sq RSS AIC  
## <none> 240.01 84.645  
## - GiniDiff 1 18.582 258.59 85.926  
## + PropWDiff 1 3.284 236.72 86.039  
## - FDIDiff 1 22.066 262.07 86.515  
## - InfDiff 1 36.631 276.64 88.895  
## - GNIDiff 1 52.893 292.90 91.408

summary(model2)

##   
## Call:  
## lm(formula = TTDiff ~ GiniDiff + InfDiff + PropWDiff + GNIDiff +   
## FDIDiff, data = train2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -7.3382 -0.6353 0.6562 1.1872 3.9782   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.31356 0.42617 -0.736 0.4664   
## GiniDiff 0.48970 0.29419 1.665 0.1042   
## InfDiff 0.67566 0.28732 2.352 0.0240 \*   
## PropWDiff -0.25648 0.35325 -0.726 0.4723   
## GNIDiff 0.27113 0.09894 2.740 0.0093 \*\*  
## FDIDiff 0.36181 0.20032 1.806 0.0788 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.496 on 38 degrees of freedom  
## Multiple R-squared: 0.4547, Adjusted R-squared: 0.3829   
## F-statistic: 6.337 on 5 and 38 DF, p-value: 0.0002276

summary(model\_selected1)

##   
## Call:  
## lm(formula = TTDiff ~ GiniDiff + InfDiff + GNIDiff + FDIDiff,   
## data = train2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -7.5684 -0.5847 0.5213 1.1915 4.2110   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.4452 0.3833 -1.161 0.25255   
## GiniDiff 0.5065 0.2915 1.738 0.09016 .   
## InfDiff 0.6940 0.2845 2.440 0.01935 \*   
## GNIDiff 0.2838 0.0968 2.932 0.00561 \*\*  
## FDIDiff 0.3754 0.1982 1.894 0.06572 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.481 on 39 degrees of freedom  
## Multiple R-squared: 0.4471, Adjusted R-squared: 0.3904   
## F-statistic: 7.885 on 4 and 39 DF, p-value: 9.309e-05

TT\_pred1 = predict(model\_selected1, newdata=test2)  
TT\_pred1

## 1 2 3 4 5 6 7   
## -1.5548966 2.0905906 1.7163394 2.1504458 -3.0592684 -1.7837286 0.9468321   
## 8 9 10 11 12   
## -1.0789995 0.4732127 2.3492568 -1.1618222 5.0796273

sse <- sum((TT\_pred1 - test2$TTDiff)^2)  
sse

## [1] 172.1618

TT\_pred\_test = predict(model\_selected1, newdata=test2)  
sse\_test <- sum((TT\_pred\_test - test2$TTDiff)^2)  
mse\_test= sse\_test/nrow(test2)  
mse\_test

## [1] 14.34682

TT\_pred\_train = predict(model\_selected1, newdata=train2)  
sse\_train = sum((TT\_pred\_train - train2$TTDiff)^2)  
mse\_train= sse\_train/nrow(train2)  
mse\_train

## [1] 5.454707

In comparing our linear model with and without the identified outliers, we observe notable improvements. The MSE values for both training (MSE = 5.46) and test (MSE = 14.34) are significantly lower with the exclusion of those data points, indicating a better fit to both datasets. It’s important to acknowledge that while the model still exhibits overfitting to the training data, the extent of overfitting has been reduced. Examining the updated linear model, we identify significant predictors below the alpha level of 0.1, namely GNIDiff, InfDiff, FDIDiff, and GiniDiff, with positive coefficients of 0.2838, 0.6940, 0.3754, and 0.5065, respectively. This contrasts with the original model, which included PropWDiff as a significant variable in predicting TTDiff but excluded FDIDiff due to stepAIC. The implications of this updated model suggest that, among the countries analyzed, whenever there is growth in the Gini coefficient, Inflation rate, Foreign Direct Investment, or Gross National Income per capita, Trade as a percentage of GDP will also increase.

1. PCA Analysis

In our supervised learning models, we did not specifically examine whether each country was distinct from each other. Instead, we attempted to consider our models by stratifying our test and training dataset. To thoroughly explore our dataset, our group decided to spur this minor question using Principal Component Analysis.

library(factoextra)

## Warning: package 'factoextra' was built under R version 4.3.2

## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa

library(cluster)

## Warning: package 'cluster' was built under R version 4.3.2

variablesdiff <- CountryDataDiff %>% dplyr::select(-Country)  
pca\_resultdiff <- prcomp(variablesdiff, scale = TRUE)  
t(pca\_resultdiff$rotation)

## GiniDiff InfDiff GNIDiff PropWDiff TTDiff FDIDiff  
## PC1 -0.45485029 -0.4093995 -0.3016777 0.4670928 -0.56241532 -0.002670247  
## PC2 0.13971203 0.2139563 -0.5407859 -0.1672793 -0.11390688 -0.775433246  
## PC3 -0.05203459 0.3757478 0.5911349 0.5633443 -0.07862243 -0.427933572  
## PC4 0.67427622 -0.6606853 0.2072953 0.1091845 -0.08386897 -0.216610148  
## PC5 0.46353299 0.3003789 -0.4558901 0.5952386 0.14376637 0.334807625  
## PC6 -0.31839113 -0.3445226 -0.1276460 0.2650564 0.79801628 -0.237808551

summary(pca\_resultdiff)

## Importance of components:  
## PC1 PC2 PC3 PC4 PC5 PC6  
## Standard deviation 1.5494 1.1123 0.9264 0.8910 0.62200 0.56831  
## Proportion of Variance 0.4001 0.2062 0.1430 0.1323 0.06448 0.05383  
## Cumulative Proportion 0.4001 0.6063 0.7494 0.8817 0.94617 1.00000

pca\_resultdiff$x %>% as\_tibble() %>% bind\_cols(CountryDataDiff) %>%   
 ggplot(aes(x = PC1, y = PC2, color = Country)) +   
 geom\_point()

A graph with different colored dots

Description automatically generated

Reviewing the tables outputted by the PCA, it is evident that the first three principal components account for 75% of the variance in the data. Additionally, from the proportion of variance row, we can highlight the elbow being at PC1 due to the sharp change in slope. As for looking at the linear combination of the principal components, i.e. rotations of the cartesian plane, we see a notable negative loading in GiniDiff, InfDiff, GNIDiff and TTDiff while strong positive loading for PropWDiff, which match the supervised linear regression performed when the outliers were included. Now turning our attention to the graphical output, the first thing to note is that all countries appear to cluster around the origin (0,0), with Norway and Kenya being more dispersed over the graph when compared to China and the US. These observations draw the conclusion that no country is unique when looking at the changes of variables i.e. growth. This makes sense as if countries opt for similar economic decisions, their growths would be in similar or same direction. The issue is that countries do not make the same decisions, so when looking at the non-differenced data we should see unique clusters.

variables <- CountryData %>% dplyr::select(-Country,-TradeGrowth,-Year)  
pca\_result <- prcomp(variables, scale = TRUE)  
t(pca\_result$rotation)

## TotalTrade GNI Inflation FDI PropW Gini  
## PC1 0.51916278 -0.02418397 -0.1888622 0.15533083 0.58077546 -0.57687816  
## PC2 0.01423883 -0.66286908 0.2862325 -0.66970997 0.07876667 -0.15413356  
## PC3 0.42776994 0.36511544 0.8239422 -0.01699941 -0.04455485 0.05048528  
## PC4 -0.03056871 0.62143034 -0.2658809 -0.71346835 0.18061426 0.02320900  
## PC5 -0.73847023 0.16117730 0.3581236 0.11284574 0.34607646 -0.40979023  
## PC6 0.03173988 0.12065088 -0.0676475 -0.07280563 -0.70860434 -0.68734197

summary(pca\_result)

## Importance of components:  
## PC1 PC2 PC3 PC4 PC5 PC6  
## Standard deviation 1.6567 1.2075 1.0363 0.73868 0.33947 0.25016  
## Proportion of Variance 0.4575 0.2430 0.1790 0.09094 0.01921 0.01043  
## Cumulative Proportion 0.4575 0.7005 0.8794 0.97036 0.98957 1.00000

pca\_result$x %>% as\_tibble() %>% bind\_cols(CountryData) %>%   
 ggplot(aes(x = PC1, y = PC2, color = Country)) +   
 geom\_point()

A graph with different colored dots

Description automatically generated

As mentioned before, there are 3 unique clusters of Norway, China and the US. Kenya can be debated in terms of it being its own cluster; however, if there were more data points, it would probably have its own cluster. This graphical output reveals that countries do not come to the same economic decisions, hence it being a unique cluster. In addition, there should not be an issue with correlation as it PCA converts any data into uncorrelated principal components due to the analysis applying orthonormality onto the data set.

# **Conclusion**

After assessing our data through exploratory modeling as well as more advanced machine learning techniques we have come to two solid conclusions. The first revolves around the relationship between the predictors of PropW, GNI, Gini index, and Inflation in correspondence to Total Trade. When not removing outliers, we find that the proportion of women in national parliaments is the strongest predictor of total trade. However, when removed, as seen in the Effects on Modeling section, we find GNI per capita to be the strongest predictor instead. The second conclusion we come to is regarding our PCA analysis. Through this modeling, we found that our non-differenced data had clear clustering, yet the differenced data did not. This indicates that if countries made the same decisions, they would have the same amount of total trade as one another. Our research lays a foundation for further exploration into the relationships between trade and macroeconomic and social variables. The findings, though our models, contribute to the ongoing discourse in global economic research and provide valuable insights for future empirical research in economics and data science.

## **Limitations**

During our research, we encountered several notable limitations that warrant discussion. A primary constraint was the incomplete nature of our dataset, exemplified by the inconsistent reporting of the Proportion of Women in Parliament (PropW) in earlier years, which posed challenges for conducting a complete historical analysis. This gap in the data may have potentially introduced biases and limits the generalizability of our findings. Furthermore, our team’s lack of prior experience with time series analysis presented a significant hurdle. As this was a crucial aspect of our study, the learning curve we faced may have affected the depth and rigor of our analysis, potentially impacting the robustness of our results. Another critical issue was the necessity to potentially overfit our models due to the limited availability of training data. This can lead to models that perform well on the training data but fail to generalize effectively to unseen data, thereby limiting the predictive power of our findings. Each of these limitations underscores the need for a cautious interpretation of our results and provides a valuable roadmap for future research endeavors.

## **Suggestions for Future Research**

When thinking about the future direction of the project, we propose several key directions aimed at refining our methodology. Firstly, the incorporation of ARIMA (Autoregressive Integrated Moving Average) would be beneficial. This approach would optimize the autocorrelation in time series data, thereby enhancing the accuracy of our forecasts. This is better than simply differencing the data like we did. However, it is crucial to acknowledge that interpreting ARIMA models requires a sophisticated understanding of their mechanics, especially as the data must be reverse-transformed for meaningful interpretation. Another future direction that could be pursued is classifier-based machine learning techniques. We would do this to deepen our understanding of the dataset by identifying and categorizing key economic indicators. The challenge here lies in our current limited economic expertise, particularly in determining appropriate categorizations for the data points. Lastly, addressing data quality and quantity issues is important. We aim to find datasets with minimal missing values (NAs) to ensure the reliability of our analysis. Additionally, expanding the size of our dataset, with more variables, is crucial to reduce the risk of overfitting, thereby enhancing the generalizability of our model. By pursuing these future directions, we anticipate a significant improvement in our ability to model and understand the intricate dynamics between trade and various economic and social factors.

1. **References**

[1] World Bank Open Data. (n.d.). https://data.worldbank.org/