HUMAN AND SOCIAL DATA SCIENCE MSc

Data-Driven Models for wealth inequality

SUPERVISOR NAME: Professor Erinco Scalas

WORD COUNT:

Candidate Number: 253696

Table Content

ABSTRACT

Wealth inequality is the difference between the rich and the poor (Brian, 2015). The distribution of wealth has become increasingly skewed as the rich have accumulated more wealth than the poor. Although Thomas Piketty has proposed several possible solutions to this problem, there are drawbacks to his approach because it is primarily subjective. We adopted an alternative statistical and data-centric approach by forecasting with an autoregressive integrated moving average (ARIMA) model. Our findings show that the Gini index, a commonly used measure of wealth inequality, has risen over time and that economic growth in developed countries has contributed to this increase. This trend will continue in the future if no policy interventions are put into place. We discovered that wealth clusters tend to grow larger over time, leading to an increased concentration of wealth in fewer hands.

INTRODUCTION

Wealth disparity has risen drastically worldwide over the last several decades (Sala-i-Martin, 2002). The wealthiest 1% of individuals hold more than half of the world's wealth, while the lowest 50% own less than 1%. (Zucman, 2019). The top income earners receive 15-20% of total income, while the lowest income earners receive less than2% illustrating how wealth concentration affects our society and economy (Potter, 2014). The gap between the affluent and poor in most nations in the Organization for Economic CO-OPERATION AND DEVELOPMENT (OECD) has reached its most significant level in 30 years. The wealthiest 10% of the OECD population makes 9.5 times the income of the poorest 10%; in the 1980s, this ratio was 7:1 and steadily climbed up (Cingano, 2014).

However, the increase in real income disparity is not (primarily) due to rising top income shares: incomes at the bottom generally expanded far slower during successful years and plummeted during downturns, placing relative (and, in some countries, absolute) income poverty on policymakers' radar. Wealth differs between the affluent and the poor (Brian, 2015). Often people use wealth inequality and income inequality interchangeably; however, they vary because income is a flow variable, whereas wealth is a stock variable (Easton & Harris, 1991). In other words, income is how much money is generated in a year, whereas wealth is what is an accumulation through time. Net worth or net worth per capita are two ways to quantify wealth. We can calculate Net worth by deducting liabilities from assets. Net worth per capita is computed by dividing total wealth by the number of people living in a particular country. In most cases, a high GDP and a high GDP per capita do not imply equal earnings. The Gini index is one of the most important indices for measuring income or wealth disparities. The OECD believes that the Gini coefficient is the best way to examine a tax's

distributional impact. For starters, it is not affected by population or economic size. Second, it evaluates inequality at low- and high-income levels, as opposed to other metrics such as the Palma ratio, which focuses solely on the transition between the low- and high-income groups (Basu & Stiglitz, 2016; Palma & Stiglitz, 2016).

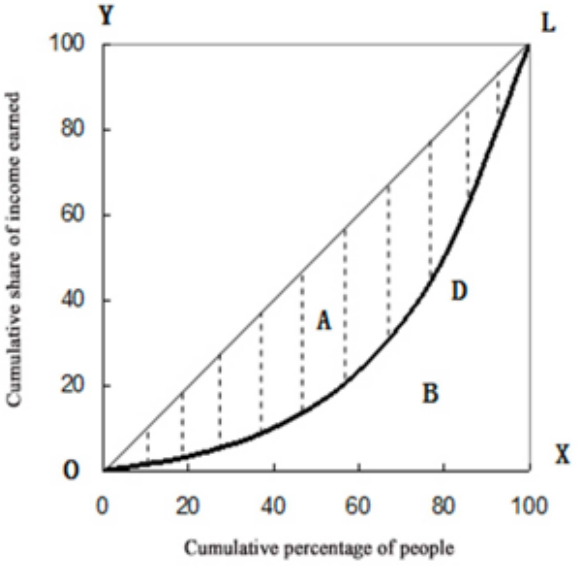
In his book Capital in the Twenty-First Century (2014), the French economist Thomas Piketty studied and popularized the phenomenon of wealth inequality. His book has been influential in many fields of study, from economics to sociology. It became a bestseller as it moved from classrooms to mainstream media outlets and sparked interest among scholars regarding its distribution of overtime. In the last two decades, models have been developed through various approaches, including stochastic or conventional, to derive the distribution of wealth. There are two different approaches to measuring inequality: the conventional method, which uses income or wealth distribution tables, and the synthetic approach, which uses the Gini coefficient. The conventional method is easy to understand and apply but suffers several problems. First, the number of observations differs between countries depending on the size of their economies. Second, the data are subject to reporting errors that may result in incorrect measurements of income or wealth by households or countries at different points in time. Third, even if all these problems are solved perfectly, it would still be necessary to find out what share each decile and centile has in total income or total wealth using this method. This answer is not provided directly by any distribution table but can only be obtained indirectly through formulaic calculation. The dissertation will focus on recent machine learning and statistical methods developed to model wealth distributions and distribution changes over time.

GINI INDEX

The Gini Index for wealth measures the distribution of wealth in an economy or country. A High Gini Index suggests more wealth inequality and vice versa. Income disparity, on the other hand, is frequently accompanied by wealth inequality. The Gini index is a statistical measure. It indicates how closely a distribution of income or wealth approximates a normal distribution, but it is not itself a normal distribution. The Gini coefficient is an unbiased estimator that measures inequality. It does not matter whether it is the United States or Denmark. Within each country, there will be no significant differences between people at different points based on income or wealth distribution.

LORENZ CURVE

Lorenz curve is a graphical representation of quantiles of cumulative distributions in economics. They are used to demonstrate disparity in some variables among certain groups. Lorenz curve, named after Max Oskar Lorenz (1876-1959), included them in his work in 1905. The Lorenz curve compares the cumulative percentages of total income or wealth to the total number of recipients or people assigned to each quantile. This curve shows how much of an economy's total wealth is held by the rich versus how much is held by the poor. The Gini index is calculated from this curve by measuring the area between the line of perfect equality (a 45-degree line) and the Lorenz curve. (Gastwirth, 1972).



The current wealth distribution model uses the Gini coefficient to measure inequality, which measures how equally or unequally a country's income distribution is spread on a scale of 0 to 1. The lower the Gini coefficient, the more equal a country's income distribution is. So, if we have a Gini coefficient of 0, everyone has the same amount of money! In contrast, if we have a Gini coefficient of 1, then one person has all the money, and everyone else has nothing. According to the United Nations, a Gini coefficient of less than 0.2 represents perfect equal income distribution; one between 0.2 and 0.3 represents equal distribution; one between 0.3 and 0.4 represents reasonable distribution; one between 0.4 and 0.5 represents unequal distribution; and one greater than 0.6 represents unequal distribution. We will use the Gini index and Gini coefficient interchangeably. We can get the Gini index by multiplying the GINI coefficient by 100.

Forecasting and Time Series Analysis

Forecasting and Time series analysis are a part of temporal data mining. Data mining is extracting and detecting patterns in vast amounts of data using methods from machine learning, statistics, and database systems(Lee & Siau, 2001). Large volumes of a dataset are collected within a uniform time interval, termed time series data(Esling & Agon, 2012b). The time can be represented as year, month, week, and day. The time series is analyzed to predict the changes within the given data and the changes that will happen in the future(Esling & Agon, 2012b; Gan et al., 2014). The characteristics of time series are analyzed to foresee future data by discovering patterns and relationships within a set of time-based events(Esling & Agon, 2012a; Jebb et al., 2015). It can be used to analyze future trends, predict changes in variables that are affected by time, and compare events that happened at different times but under similar conditions. For time series data, the pattern describes one or more recurring events that affect a particular variable within a given period. The period is the average length between events in each pattern. The confidence estimates how accurate the prediction is based on the pattern. The ending time is the last point at which a pattern occurs in the dataset. For time series analysis to be practical, it requires large amounts of data over many years to make accurate predictions about future occurrences. Data collection is usually done by performing surveys and other forms of research arch, then storing it in databases for future reference. There are several different methods for performing time series analysis. These include ARIMA, an acronym for Autoregressive Integrated Moving Average, and Long short-term memory (LSTM), a type of recurrent neural network (RNN) used in deep learning. It was developed in 1995 by Sepp Hochreiter and Jürgen Schmidhuber. Time series prediction is used in various fields, including finance and economics. The prediction can be made based on three

different time spaces: linear, nonlinear, and autoregressive, and specification testing may also be used as a tool for time series analysis(Teräsvirta et al., 2005).

ARIMA

ARIMA is an acronym that stands for Autoregressive Integrated Moving Average. The latter part of this acronym (MA) is sometimes substituted with seasonal (SA) and has led to the name Box-Jenkins approach. This approach has been prevalent among statisticians for the last 20 years (Fan et al., 2010). ARIMA modelling is primarily an experimental data-oriented technique with the flexibility of fitting a suitable model modified from the data's structure. The stochastic character of the time series may be approximated using the autocorrelation function and partial autocorrelation function, from which information such as trends, random fluctuations, periodic components, cyclic patterns, and serial correlation could be determined. Therefore, projections of the series' future values can be derived with some degree of accuracy. There are several methods for estimating the parameters of an ARIMA model. The Box-Jenkins (ARIMA) approach is widely established in the statistical literature. An iterative three-stage procedure of model identification, parameter estimation, and diagnostic check is required to assess the appropriateness of the suggested model (Ho & Xie, 1998). The ARIMA (Autoregressive Integrated Moving Average) model is more accurate than a standard moving average. It removes trend, seasonal, and irregular components from the data (Domingos et al., 2019; Saputro et al., 2022). The model has three parameters that are used to represent Autoregressive Model (AR), The Moving Average (MA), and Differencing order (Babu & Reddy, 2014; Elamin & Fukushige, 2018).

Deterministic and stochastic trends can both be identified in time series data. Deterministic trends are those with a constant level over long periods. Stochastic trends are those with random levels over short periods(Renard et al., 2013; Yue & Pilon, 2003). Methods applicable to both deterministic and stochastic trends are adopted to determine trends based on the scope of observation, which could be global or local.

RESEARCH OBJECTIVES

Using data science tools to mine data, understand and make predictions of wealth inequality using economic variables can be a unique opportunity with great potential. Through exploratory analysis and visualization, we can study the ability to collect data, understand patterns and make future predictions using data science tools. This ability helps to tackle the problem of wealth inequality better.

The project aims to achieve these objectives:

1. Assess the distribution of wealth and estimate countries' Gini coefficient using descriptive Analysis in Machine Learning Model

2. Testing whether wealth inequality is a local phenomenon or if there is evidence of a global concentration of wealth using the Statistical tools learning model and

3. Predicting the Future wealth Gini Coefficient from the previous wealth coefficient efficiency using a suitable Time series Model.

LITERATURE REVIEW/ RELATED WORK

Wealth Inequality has been systematically studied and popularised by Thomas Piketty in his recent book "Capital in the Twenty-First Century (Piketty, 2013)". The distribution of wealth also shows an evolution over time, as evidenced in Piketty's book, which showed that from 1810 to 2000, both income and wealth had increased for all but one percentile. (de Nardi et al., 2015) He mentioned the correlation between wealth and income in developed countries. He argued that the rise in inequality in recent decades could be explained by a rise in the return on capital and a slowdown in growth rates and not industrialization. The fact that the wealth distribution is right-skewed, unlike the normal distribution, and that it is right-tailed can be well approximated by a Pareto distribution (Pareto, 1964).

In economics, one benefit of Pareto distributions—or power law distributions, as they are sometimes called—is their use to model different income and wealth distributions(Arnold, 2015; Balakrishnan & Nevzorov, 2005; Forbes et al., 2010). Pareto distributions are also known as log-normal, heavy-tailed, or long-tailed distributions. A general lack of symmetry characterizes such distributions: there is no specific value that occurs with the highest frequency (and thus the "peak" of the curve). Instead, the distribution is said to have an "asymmetric peak," and its shape (in particular, how high it peaks) depends on the parameter(Blanchet et al., 2022; Geisser, 1984; STIGLITZ, 1981). The Burr distribution is also known as the multivariate Pareto distribution(Arnold, 2014). Arnold (2014) investigated more extensive families of multivariate Pareto distributions generated through multivariate geometric reduction. Bivariate or multivariate Pareto models may help assess the wages of

related individuals (such as employed husband-wife pairs) or individuals measured on two different dates(Arnold, 2015; Cook & Johnson, 1986; Hutchinson, 1979).

The multivariant Burr distribution, first developed in 2006 by Takahashi, is a generalization of the Burr distribution. A. Burr introduced the Burr distribution in 1944 as a probability distribution for a sum of independent identically distributed random variables. The Burr distribution is based on the gamma function and is defined by the interval [0,1](Fisk, 1961; Tadikamalla & Johnson, 1982; Takahasi, 1965). The Burr distribution introduces an important concept: the origin of the random variable of interest must be between 0 and 1 (or -1 and 1 for the case of a right-skewed variable)(Rodriguez, 1977a; Singh & Maddala, 1976). The multivariant Burr distribution extends this concept to dependent random variables(Domma, 2009; Hose, 2005). It is a generalization of its predecessor because it allows more freedom concerning the shape of the cumulative density function (CDF) and adds new features such as skewness and kurtosis to the CDF(Keelin, 2016; Matis et al., 2009). This extension is significant because it opens possibilities for transformations not present in the original Burr distribution(Champernowne, 1952; Lomax, 1954; Tadikamalla, 1980). The use of transformations allows us to better fit certain distributions commonly found in nature, such as power-law distributions(Bennett, 1983; Burr, 1968). The multivariant Burr distribution is defined through a Gamma combination of independent Weibull random variables, which means that each random variable has its location parameter, but all have the same(Ahmad et al., 2009; Takahasi, 1965). Takahashi addresses that the Burr distribution while modelling heavy tails for return periods beyond the normal, does not model skewness(Ahmad et al., 2009; Rodriguez, 1977b; Team et al., 1993; Takahasi, 1965). He uses this as motivation to develop the multivariant Burr distribution using a gamma combination of separate Weibull

random variables(Taam et al., 1993; Takahasi, 1965). While Takahashi acknowledges that both the Burr distribution and the resulting multivariant Burr distribution are still not capable of considering skewness, he justifies his work by quoting Fama and French (1993). They state that "skewness has no implications for expected returns" (Fama & French, 1993).

Salas-Rojo et al. (202) present a study of inheritance and wealth distribution and the inequality of opportunity. Their study proved that developing a machine learning model that uses the inequality of opportunity approach to measuring the impact of inheritance and wealth helps overcome the previous limitations experienced with a traditional model.
For example, in the US, it has been shown that the top 1% of wealthiest families own about 40% of total wealth. In comparison, the bottom 80% only own about 7%, revealing that wealth has shown a lot of significant wealth mobility over time according to the stylized evidence (Nardi et al., 2015).

Mathematical Modelling of an ARIMA Model

To build an ARIMA model, we are concerned with the three values of the parameters. Let us express the values as (p, d, q).

P here represents the lags in the autoregressive model.

D here represents the integration /differencing order.

Q here represents the moving average lags.

These can be further represented below

Autoregressive Model: AR

$AR(1): Y_t = \mu + \phi * Y_{t-1}$.....(1)

Moving Average lags: MA

$MA(1): Y_t = \mu + \varepsilon_t + \phi_1 * \varepsilon_{t-1}$.....(2)

ARIMA

$ARIMA(1): Y_t = \mu + \phi * Y_{t-1} + \mu + \varepsilon_t + \phi_1 * \varepsilon_{t-1}$.....(3)

Therefore, from equation (1) above

Auto-Regressive (AR) Model

$$\hat{y}_t = \beta_1 y_{t-1} + \beta_2 y_{t-2} + \beta_3 y_{t-3} + \cdots + \beta_p y_{t-p}$$

The AR model assumes that the current value ($\hat{y}_t$) is **dependent on previous values** ($y_{t-1}, y_{t-2}, + \cdots + y_{t-p}$.).Partial Autocorrelation Function (PACF) is used to figure the order of an AR Model .n a Time Series Analysis, The Partial Autocorrelation between x_t and x_{t-k} is

defined as the conditional correlation between x_t Andx_{t-k} conditional on $x_{t-h+1},\ldots,x_{t-1}$ the observed time series data point was found between the two-time points t and $t-h$.

Also, From Equation (2)

Moving Average (MA) Model

$$\hat{y}_t = \varepsilon_t + \beta_1\varepsilon_{t-1} + \beta_2\varepsilon_{t-2} + \ldots + \beta_q\varepsilon_{t-q}$$

The MA model assumes that the current value ($\hat{y}_t$) is dependent on the error terms, including the current error ($\varepsilon_t, \varepsilon_{t-1}, \varepsilon_{t-2} + \ldots + \beta_q\varepsilon_{t-q}$)

Because error terms are random, there is no linear relationship between the current value and the error terms. The Autocorrelation Function (ACF) is used to figure out the order of the MA model. The autocorrelation function (ACF) is a statistical metric used to detect if a time series and a lagged version of itself are correlated. It determines whether a time series follows a random walk or a pattern. The autocorrelation function (ACF) depicts the correlation between observations at various time intervals. The autocorrelation function starts at a default value of lag 0, which is the correlation of the time series with itself and therefore results in a correlation of 1. The ACF plot can help to determine whether the time series is white noise or random. The Moving Average can modulate the relationship between observation and the observed time series.

Autoregressive Integrated Moving Average(ARIMA)

Finally, from equation (3), We can say that the time series y_t is an Autoregressive Integrated Moving Average *model of order p,d q*, ARMA(p,d, q), if:

$$y_t = \beta_1 y_{t-1} + \beta_2 y_{t-2} + \beta_3 y_{t-3} + \cdots + \beta_p y_{t-p} +$$

$$\varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}$$

Therefore, we can say that ARIMA (1,1,1), which means ARIMA model of order (1, 1, 1) where AR specification is 1, Integration /Differencing order is 1and Moving average specification is 1.

COMPARING ACF AND PACF

We also note that the ACF and PACF start with a lag of 0, which is the correlation of the time series with itself and therefore results in a correlation of 1. The difference between ACF and PACF is the inclusion or exclusion of indirect correlations in the calculation. Furthermore, you will see a blue area in the ACF and PACF plots, which depicts the 95% confidence interval and indicates the significance threshold. That means anything within the blue area is statistically close to zero. The ACF starts with lag 0 because it includes all correlations between any two observations (correlation between observations at lag 1). In contrast, PACF does not include these indirect correlations, so they are not included in its calculation. This can be seen by looking at both plots: The blue region in ACF starts at zero, whereas it starts at one in PACF.ACF and PACF assume that the underlying data is stationary. Stationarity Data

assumption is made, meaning that the mean and variance of a time series are constant for the whole series, no matter where you choose a period.

STATIONARITY TEST USING AUGMENTED DICKEY-FULLER

In 1979, economists James A. Dickey and Robert E. Fuller and statistician Nancy A. Reichardt developed the Augmented Dickey-Fuller (ADF) test to check for stationarity in time series data. It is applied to determine whether a time series has a unit root. If the p-value is equal to or less than 5% or 0.05, fail to reject the null hypothesis and infer that your data does not have a unit root; if the p-value is larger than 5% 0r 0.05, reject the null hypothesis and conclude that your data does have a unit root and is still stable.

The first step of the ARIMA model is to check whether the data are stationary or non-stationary. If they are non-stationary, they usually have trends and seasonality that need to be removed before fitting an ARIMA model. The simplest way to do this is through differencing, which involves subtracting each value from its corresponding value in the previous period.

Differencing

For example

the values of Y after 1-lag differencing

Y= [3, 6, 4, 7, 9, 7]

$Y^* = [-3, 2, -3, -2, 2]$

While the value of Y after 1-lag twice will give

$Y = [-3, 2, -3, -2, 2]$

$Y^* = [-5, 5, -1, -4]$

Therefore:

$Y_{t+1} = Y_t - Y_{t-1} = Y_t(1 - Y_t)$

APPLICATION OF ARIMA IN PRACTICE

This section shows the approach taken to collect data and understand the distinction between wealth inequality using exploratory analysis, diagnostic methods, and forecasting. We also used the Gini index to study concentration and forecast future wealth distribution from past and current data.

We must conduct a descriptive analysis to assess wealth distribution and estimate Gini coefficients. In this analysis, we will examine the current wealth inequality in countries to assess whether there is a global concentration of wealth or if there is evidence of a local concentration of wealth. We will use statistical tools to predict future Gini coefficients from previous ones; time series models are best suited for such predictions.

To achieve this objective, we have used two types of methods.

The first is an exploratory analysis which helps us answer questions like what kind of inequalities exist in a country? Is it increasing or decreasing? Do different countries have unequal distribution of wealth?

The second method is the diagnostic method which uses statistical tools such as the Gini index, or Gini coefficient of variance, and p-value to study concentration and forecast future wealth distribution from past and current data.

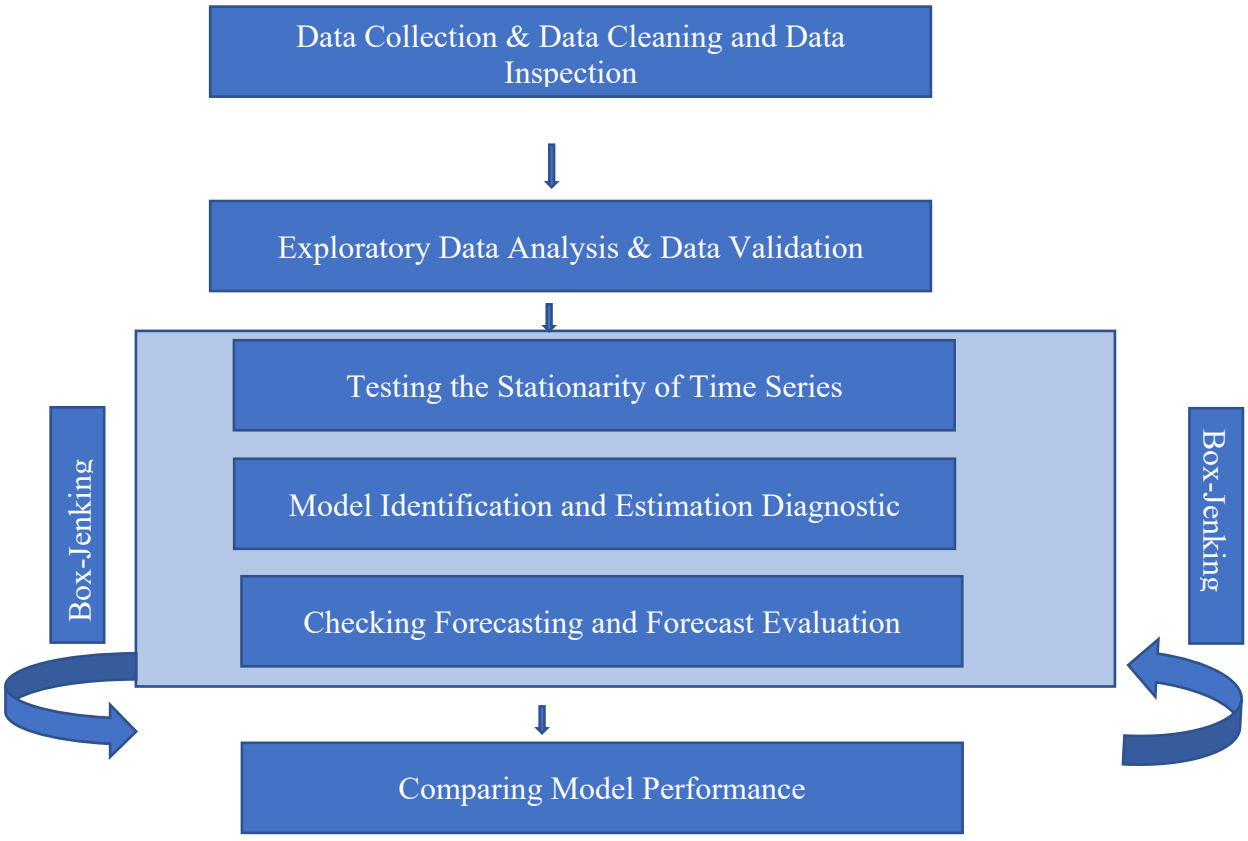
Autoregressive Integrated Moving Average ARIMA Time series forecasting

Algorithms (e.g., ARIMA) were used, and long short-term memory in an artificial neural network is popularly used in Arterial intelligence, Deep learning, etc. Though we have small observations, we could not achieve much with the LSTM model.

This section gives an overview of a primary ARIMA forecasting method. This approach is graphically depicted in Figure 1. It is essential to remember that this procedure does not follow a simple sequential approach but may entail repeating loops based on diagnostic and forecasting outcomes. The first stage is to collect and analyze the forecasted data visually and quantitatively. The second stage is to determine if the data remain stationary or if differencing is necessary. Once the data has been deemed stationary, the suitable ARMA model should be identified and estimated.

The Box-Jenkins methodology and penalty function criteria are both recognized as alternate approaches to model identification. Every discovered model must be subjected to diagnostic checks (often focused on evaluating the residuals) and sensitivity analysis. For instance, the calculated parameters must be very resilient to the selected time range. The process should be repeated if the diagnostic checks reveal issues with the specified model. Once a model or set has been selected, the models should be employed to predict the time series, typically using out-of-sample data to assess the model's forecasting ability.

ARIMA Model Using Box-Jenking Methodology



The ARIMA fcasting process is shown in the Figurforabove is inspectingo visually inspect the data. Levels, logs, differences, and seasonal changes should all be explored. After that, the series should be plotted to evaluate if there are any structural breakdowns, outliers, or data issues. In this instance, intervention or dummy variables may be required. This phase may also reveal whether the time series has a significant seasonal pattern. O overfitting the model during identification is a common problem in ARIMA modelling because it improves the model's in-sample explanatory performance. However, model results for out-of-sample have poor predictive capabilities compared to a more parsimonious model. Whenever a model with many AR and MA delays predicts poorly, it may be advisable to return to the model identification stage and attempt a more parsimonious model.

The probability density is the relationship between observations and their probability. Some random variable outcomes will have a low probability density, and others will have a high probability density. The shape of the probability density is referred to as a probability distribution, and the calculation of probabilities for specific outcomes of a random variable is performed by a probability density function, or PDF for short. If a random variable is continuous, the probability can be calculated via probability density function, or PDF for short. The shape of the probability density function across the domain for a random variable is referred to as the probability distribution. Common probability distributions include uniform, normal, exponential, and so on.

Discussion of Data

The data set used for this study was collected from two sources. The Standardized World Income Inequality Database (SWIID) and The GapminderFoundation.The Standardized World Income Inequality Database (SWIID) is a longitudinal database containing aggregate information on income and inequality for more than 200 countries from 1800 onward. The data in SWIID are based on household surveys conducted by the World Bank, International Monetary Fund (IMF), United Nations (UN), OECD, and others. SWIID standardizes observations collected from the OECD Income Distribution Database, the Socio-Economic Database for Latin America and the Caribbean generated by CEDLAS and the World Bank, Eurostat, PovcalNet (which is o t.he sted by the World Bank), ,UN The standardized world income inequality database (swiid) contains variables such as country and year identifiers and four other variables: the Appendix contains the data descripti on and all the variable's name s.Economic Commission for Latin America, the Caribbean national statistical offices worldwide, and many other sources. Luxembourg Income Study data serves as the standard.

Dataset Description

Please refer to Table 1 in the Appendix Section for the SWIID sample dataset SWIID data has ten fields columns and 5625 rows of observation

Gamminder dataset

The Gap minder Foundation has collected and compiled a wide range of statistics about income distribution throughout history. The data were collected for three different purposes: 1) to depict trends over time, 2) to show how people's incomes have changed relative to each other, and 3) to show how these changes have affected different groups.

The sample dataset is presented in Appendix Table. The gapminder dataset contains eight columns or fields and 1704 rows of observation

Data Cleaning is done by removing unwanted columns and Null values: using pandas libraries, we were able to drop all unwanted columns and Null values. A total of 7 columns were dropped, and left only three relevant columns, which include country, year, and GINI index.

Feature engineering can be used to create new columns in a table. It is done by identifying the attributes which need to be captured and then creating a new column for each of them. Feature Engineering involves using domain knowledge to create features. Here we will discuss how feature engineering can be used to create new columns.

For example, we created a new column called Gini Index Level, denoted as *Gni_g_level,* in our data frame. To achieve this, the Gini coefficient, or Gini index, is classified into five different classes based on the recommendation from the United Nations and popular opinion. Based on this recommendation, a Gini index of less than 0.2 represents perfect equal income distribution; one between 0.2 and 0.3 represents equal distribution; one between 0.3 and 0.4 represents reasonable distribution; one between 0.4 and 0.5 represents unequal distribution; and on

0.2–0.3 represent relative equality,

0.3–0.4 represent adequate equality,

0.4–0.5 represent inequality

0.5 represents perfect inequality

Therefore, the warning level of the Gini index is 0.4. Please refer to the Appendix for the New Data sample

For a sample of the SWIID dataset. Please check **Appendix Table** for the dataset sample presentation

Performing Exploratory Data Analysis Using DESCRIPTIVE TIME SERIES

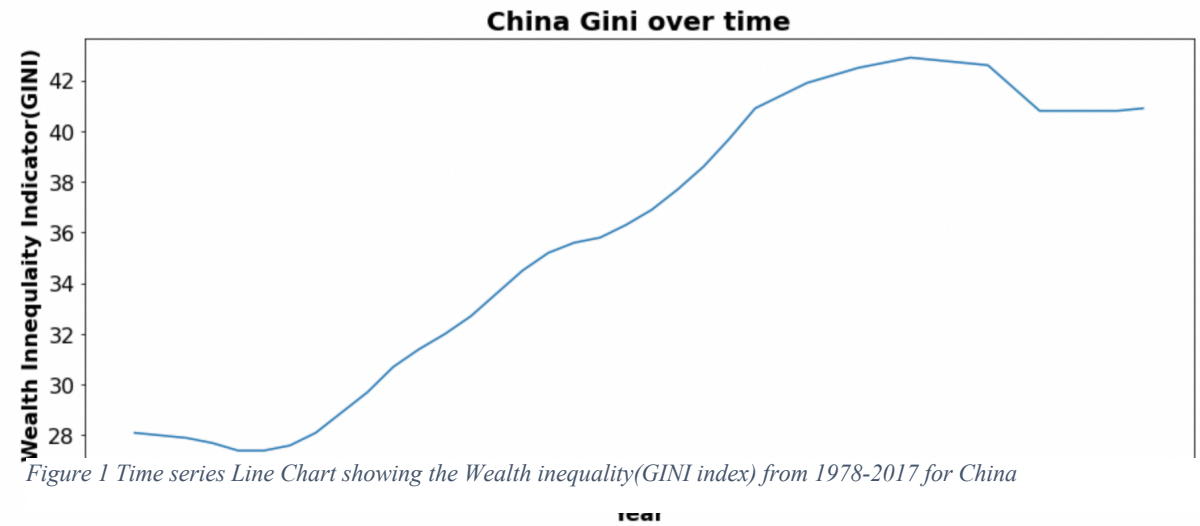


Figure 1 Time series Line Chart showing the Wealth inequality(GINI index) from 1978-2017 for China

To understand the change and t, trend, and pattern of wealth inequality (GINI index) throughout (1978 – 2017). From our chart above, we observe that there was a fall in wealth inequality (GINI index) from 28.1 to 27. 4 between (1978-1983) and then a rise from 27.4 to 42.9 (1984-200,8) and then another fall in wealth inequality (2009 -2017) based on the limit of our data.

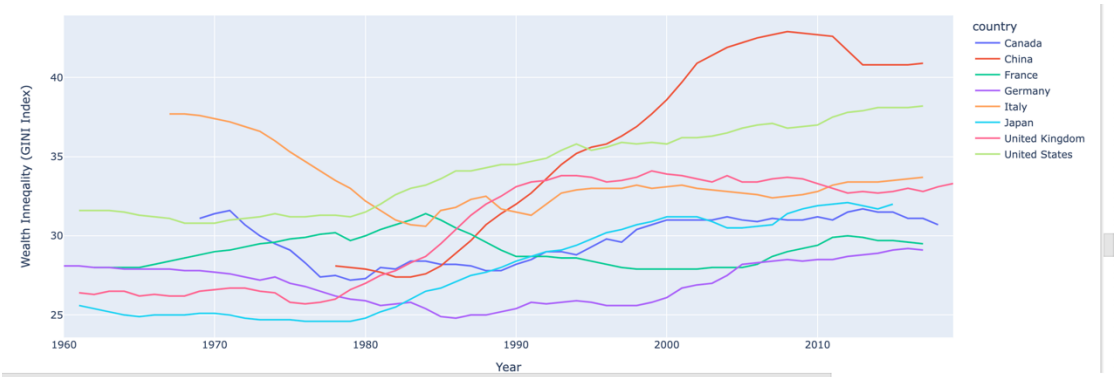


Figure 2 Time series Line Charts of the f G7 group and China showing the Wealth inequality(GINI index) from 1978-2017

To investigate China's performance in wealth inequality using the GINI index, we compare China and other Developed economies in Europe and North America. Our plot shows that China performed poorly in closing the wealth gap or wealth inequality, as reflected in the plot

Create a Line graph to study the trend of wealth gap over time for ten countries with the most Gini index (wealth inequality), 10

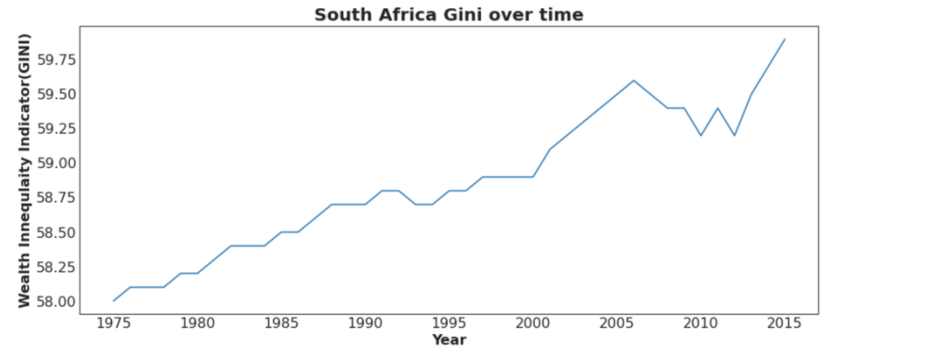


Figure 3 Time series Line Chart showing the Wealth inequality(GINI index) from 1975-2017 for South Africa

To understand the change, trend, and wealth inequality pattern (GINI index) throughout (1975 – 2015). South Africa exhibits a high level of wealth inequality categorized as perfect wealth inequality. Our chart above shows a slow rise in wealth inequality (GINI index) from 58:00 to 59. 75.

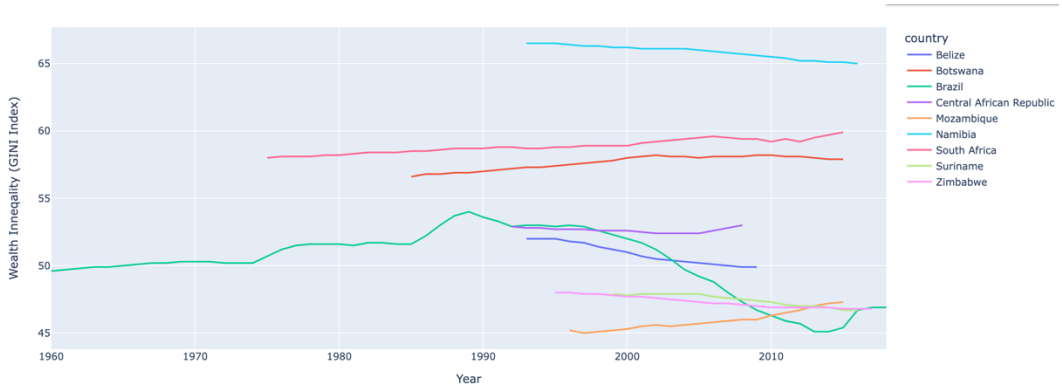


Figure 4 Figure 2 Time series Line Charts of the top countries with perfect Wealth inequality(GINI index) from 1960-2017

To Investigate the trend and wealth inequality gap among the nine countries with the most wealth inequality countries. 7 out of 9 countries are from Africa, while the other two are from South America

Almost all the countries rated as perfectly unequal or equal show little or no changes in closing the wealth gap, except Brazil from the South American continent. Out of 9 counters in this class from our data, seven are from Africa, while the other two are from South America. Apart from Namibia, South Africa

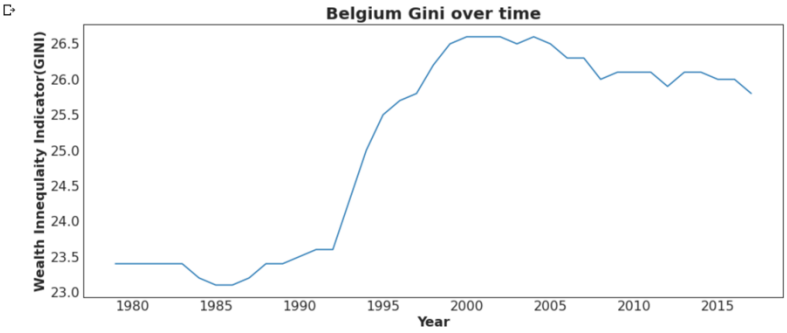


Figure 5 Time series Line Chart showing the Wealth inequality(GINI index) from 1978-2017 for Belgium

To understand the change, trend, and pattern of wealth inequality (GINI index) of Belgium (1979 – 2017). Although Belgium is grouped under country with perfect wealth equality and distribution, a closer study from our chart above shows that there was a slow rise, almost constant, and sideways movement of trend in wealth inequality (GINI index) from 23.4 to 23.6 between (1979-1993) and then a sharp rise from 23.6 to 26.6 (1993-2002) and then another sideways movement from (2002 -2017) based on the year range of available data. Create a Line graph to study the trend of the wealth gap over time for ten countries with the least Gini index (wealth inequality), 10

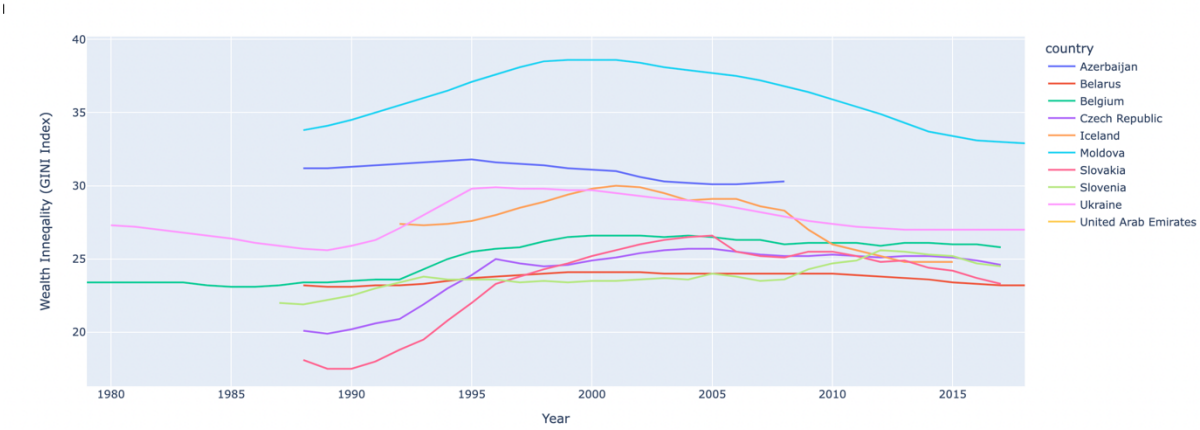


Figure 6 Time series Line Charts of the top countries with perfect Wealth equality(GINI index) from 1978-2017

To investigate the trend and performance of the wealth inequality time series chart among the top country classified as perfect equal or equal. All country in this category generally shows a concave chart to the origin which means that the chart of most country shows that most countries are either constant or rising slowly, which could be attributed to the general rise in world wealth inequality in the last 40 years according to our data.

To study the concentration of wealth in various continents and countries using two different time frames to study the deviation

Data Pre-processing

To Test whether wealth inequality is a local phenomenon or if there's evidence of a global concentration of wealth using, we created a tree map from the plotly -a visualization package in python.

Merging SWIID and Gapminder data to form a single dataset through an inner merge. The new dataset has 692 rows and ten columns. Please check the Appendix for the data sample table

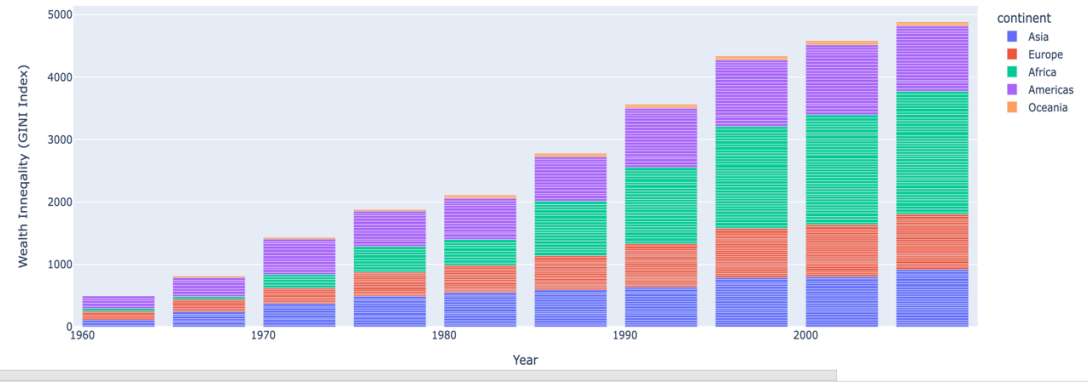


Figure 7 Bar Chart showing the proportion of wealth Inequality (GINI index) by continents in the last 70 years from 1960-2017

Figure 7 above shows that there has been a significant increase in global wealth inequality The y-axis shows the sum value of the GINI index across the country of the world, while the x-axis shows the year. The bar charts show that wealth inequality has increased significantly between 1960-2010

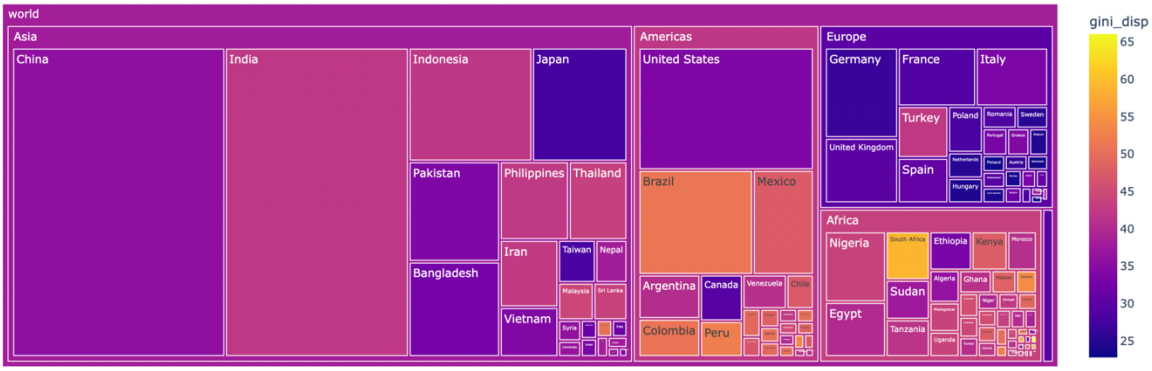


Figure 8 The treemap showing the concentration of Wealth inequality across the continents of the world

The general hierarchy of shades from darkest (lowest wealth inequality (GINI index) and highest equality through categorized as perfect equality to the mid-range of adequate equality to the light shades (highest wealth inequality (GINI index)).

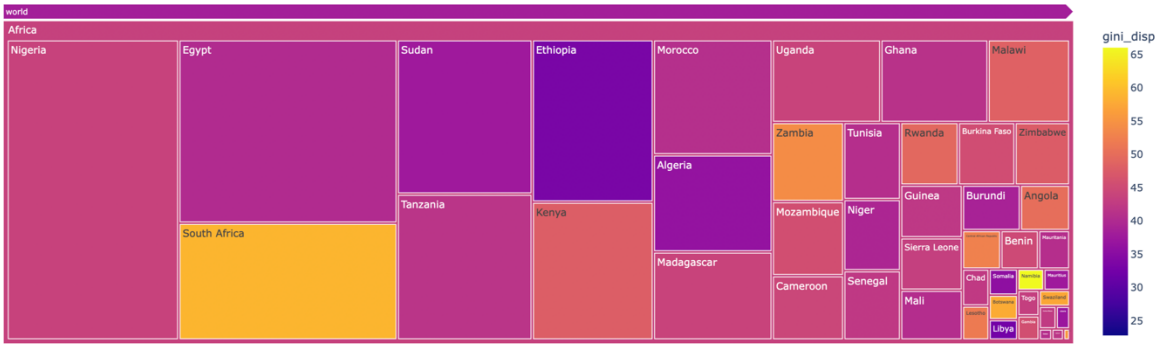


Figure 9 The tree map showing the concentration of Wealth inequality across the Africa

The overview of our tree maps shows that most countries in Africa show lighter shades indicating the highest level of wealth inequality, except for Ethiopia, Libya, and Algeria, which have dark shades revealing a relatively high inequality in Europe. Countries such as

Namibia, South Africa, Zambia, and Botswana show a very high level of wealth inequality from highest to lowest

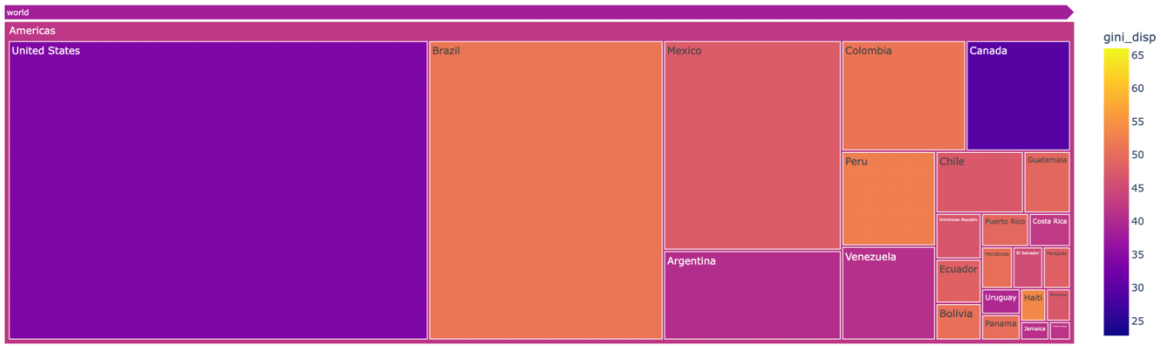


Figure 10 The tree map showing the concentration of Wealth inequality across the Americas

The overview of our tree maps shows that Canada and United States show darker shades indicating a low level of wealth inequality. In contrast, countries such as Brazil, Haiti, and Peru show a high level of wealth inequality.

The overview of our tree maps shows that most countries in Europe show darker shades indicating the lowest wealth inequality region except for Turkey, which is slightly light, revealing a relative high inequality in Europe

The overview of our tree maps shows that most countries in Europe show darker shades indicating the lowest wealth inequality region except for Turkey, which is slightly light, revealing a relative high inequality in Europe

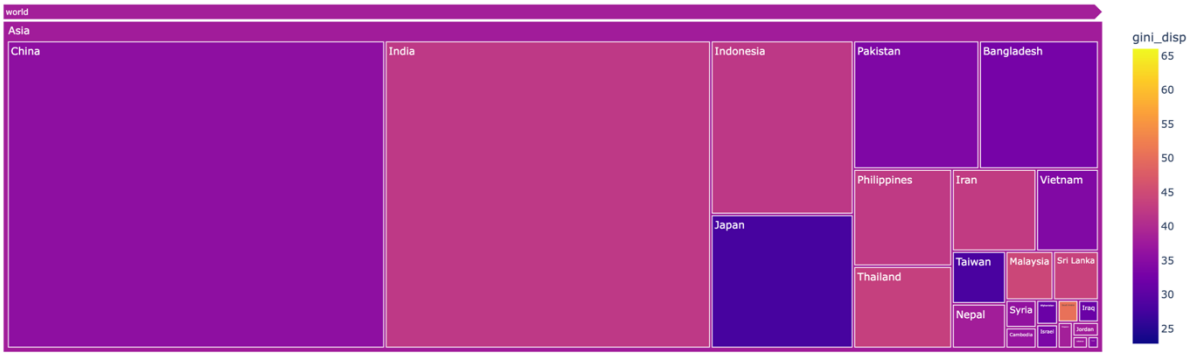


Figure 11 The tree map showing the concentration of Wealth inequality across the Asia

The overview of our tree maps shows that Japan and Taiwan show darker shades indicating the lowest wealth inequality region, While countries such as Vietnam, Bangladesh, Iraq, and Afghanistan reveal moderate wealth inequality. At the same time, Saudi Arabia, Iran, India, and Indonesiareis slightly light shades revealing a relative high wealth inequality in Asia.

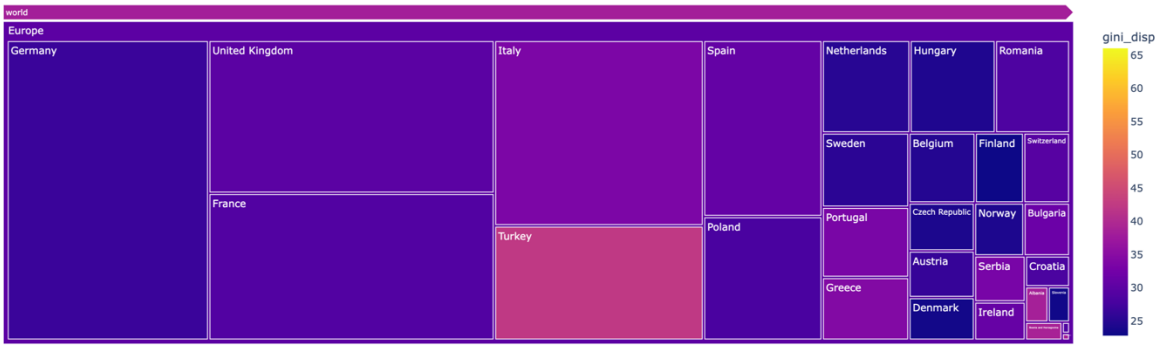


Figure 12 The tree map showing the concentration of Wealth inequality across Europe

The overview of our tree maps shows that most countries in Europe show darker shades indicating the lowest wealth inequality region except for Turkey, which is slightly light,t revealing a relatively high inequality in Europe.

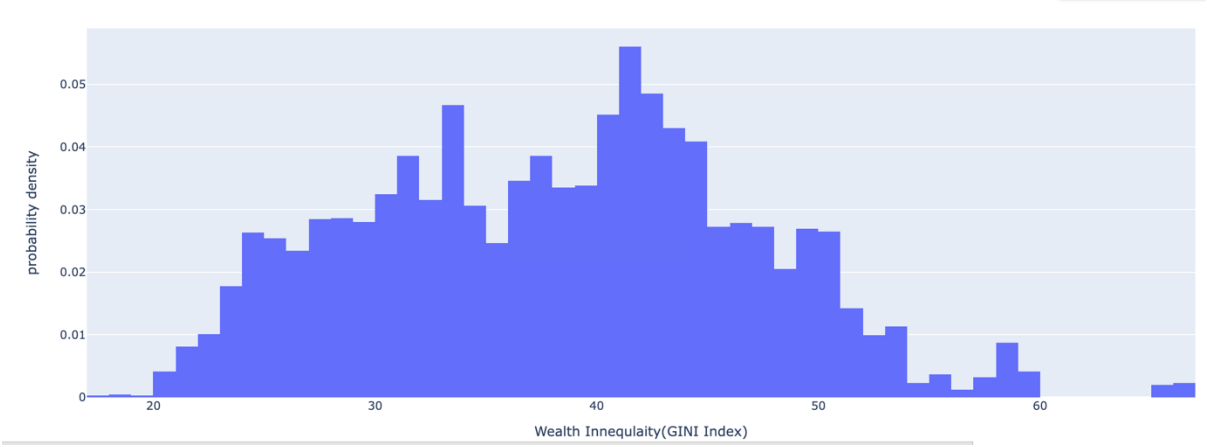
The higher GINI index (perfectly wealth inequality value from 55 and. Above indicated with light color e darker colour generally indicated country with low GINI index (perfectly wealth equality).

The result we get from the Data available shows that there is more Wealth inequality concentration in Africa and Latin American countries reflected, which was indicated by the high GINI index level.

In Europe, most countries with perfect equality, as indicated by the treemap, include Germany and GINI index ranges between.The result we get from the Data available shows that there is more Wealth inequality concentration in Africa and Latin American countries reflected, which was indicated by the high GINI index level

METHODOLOGY

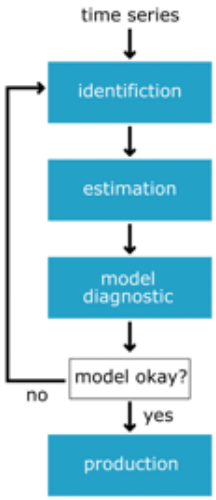
To assess the probability distribution of our dataset, we plot a Histogram



When we talk about a probability density function or PDF, we're referring to the relationship between , forvations and their probability. For example, for the GINI index for all the countries of the world which a continuous variable, such as a probability distribution on the interval (20, 60), the PDF's graph shows a bimodal distribution with a constant slope from 0 to 1.

The primary goal of this time series study is to utilize the ARIMA model to forecast the future value of the GINI index for two selected countries, China and the United States, and compare it to the actual values. It should be highlighted that we utilized multiple Python libraries for models, one of which is the stats model, which is commonly used for time series analysis.

So, to be more specific, ARIMA (1, 1, 1) denotes an autoregressive integrated moving average model with order or shift of integration of one and moving average of. 1. The primary goal of this time series study is to utilize the ARIMA model to forecast the future value of the Gini index in two selected countries—China and the United States—and compare it to actual values. It is crucial to mention that we utilized many Python libraries for models, one of which is statsmodels, commonly used for time series research. The Box-Jenkins methodology to determine the type of autoregressive moving average (ARMA) model to model a time series is based on plots of the sample autocorrelation, partial autocorrelation, and then the inverse autocorrelation function. In an ARMA(p,q) model, the sample autocorrelation is the correlation between values in consecutive lags, p and q. Using the lag operator, partial autocorrelations are measured by correlating values in different lags with a part of itself. The inverse autocorrelation function measures the correlations between values at different times. The Box-Jenkins methodology will evaluate plots of these functions and infer the proper ARIMA(p,q,d) model based on any patterns detected in these functions.



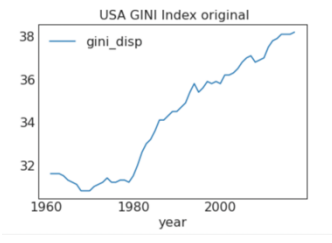
Box Jenkins Methodology Overview:

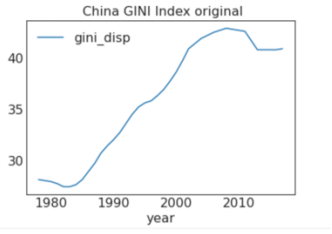
Source: Datacamp

The first step in the Box-Jenkins Methodology involves:

Testing for Stationarity of Time Series

Before Differencing





We applied the Augmented Dickey-Fuller (ADF) test to check for stationarity in the time series data of two data samples of the USA and China.

The output from the Application of ADF, the stat model library in python programming, is contained in the Appendix.

Summary Table

	China	USA	
p-value	0.39895198210341654	0.9925947221193557	
ADF	-1.7630548309555902	0.8634791417940741	
Num Of Lags :	1	0	

Number of observations used for ADF	38	56	
Critical value (5%)	-2.941262357486514	-2.9147306250000002	

- The null hypothesis is rejected if the Test Statistic is larger than the Critical Values.
- The null hypothesis was not rejected if the Test Statistic was smaller than the Critical Value. Furthermore, some interpretations of the dicky fuller test results can be seen, as the null hypothesis in the ADF test is that the data is not stationary.

A p-value of less than 0.05 is required to reject the null hypothesis and consider the data stationary. Thus, the results summary statistics are as follows:

China

the p-value is approximately 0.399

The ADF test statistics is -1.763

The critical value is -2.941

p-value = 0.339 > 0.05 then we fail to reject Null Hypothesis This means the time series is non-stationary. In other words, it has a time-dependent structure and does not have constant variance over time.

Since the ADF test statistics value = -1.763 is greater than the critical value (5%) = -2.941,

Showing that the data is not stationary.

USA

P-value is approximately 0.993

The ADF test statistics is 0.864

The critical value is -2.915

Decision

p-value = 0. 993 > 0.05, then we fail to reject Null Hypothesis. This means the time series is non-stationary. In other words, it has a time-dependent structure and does not have constant variance over time.

Since the ADF test statistics value =0.864 is greater than the critical value (5%) =-2.915,

Showing that the data is not stationary.

Now we must change the data to make it more stationary.

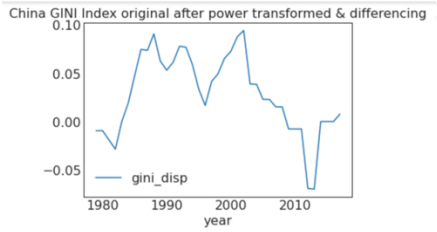
After Differencing to make data stationery

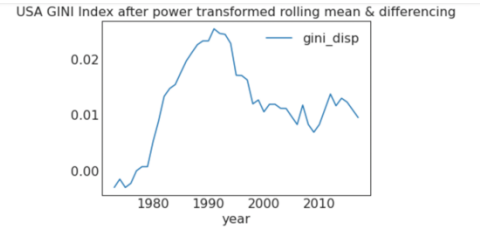
China

From the line chart above, both GINI index trend shows an upward trend. After performing a non-random first differencing, we observe some trends and seasonality in the GINI index of China every ten years or decade. The upward trend from (1980-1990), sideways movement from (1990-2000), then another downward trend from (2000-2010) and then an uptrend from (2010-2020).

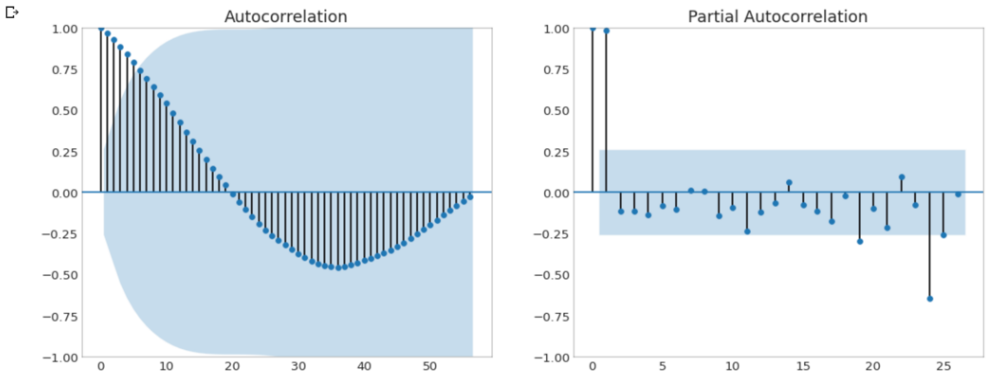
USA

From the line chart above, both GINI index trend shows an upward trend. After performing a non-random first differencing, we observe some trends and seasonality in the GINI index of the US. The upward trend from (1980-1990), downward movement from (1990-2000), then sideways movement from (2000-2010).





Box Jenkin Methodology for Model Identification and Estimation Diagnostic



In the above correlation, plot dotted lines represent the confidence band. The centre dotted line represents the mean, and the upper and lower dotted lines represent boundaries based on a 95% confidence interval.

We can see a positive correlation with delays up to 20, where the ACF plot breaks the upper confidence barrier. Although we have a decent correlation up to the 20th lag, we cannot utilize all of them since it would cause multi-collinearity. Therefore we use the PACF plot to retrieve only the most important lags.

An ARIMA model's AR and MA components may be determined using ACF and PACF graphs. The ACF and PACF plots may be used to calculate both the seasonal and non-seasonal AR and MA components.

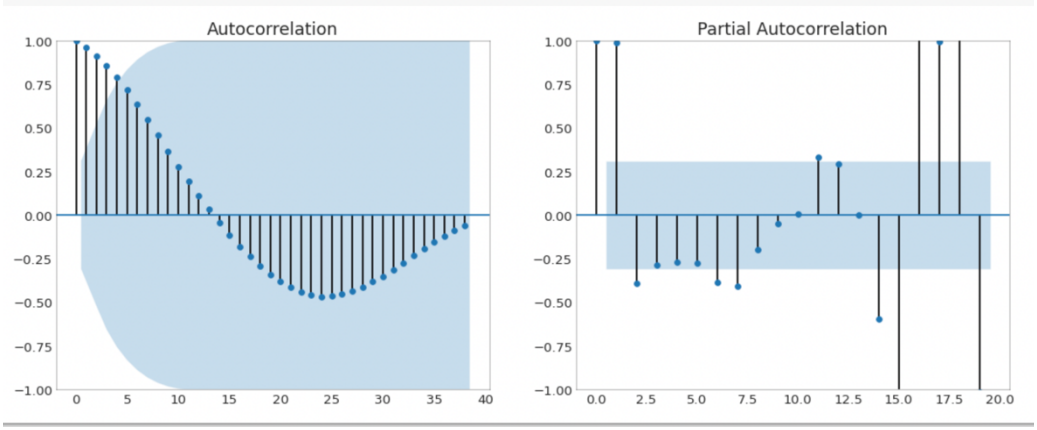
Since lag 0 is autocorrelation, we observe a significant Lag in ACF at 1, 2, 3,4,5,6

We can make the following observations:

- Several autocorrelations are significantly non-zero. Therefore, the time series is non-random.

- High degree of autocorrelation between adjacent (lag = 1,2,3,4,5,6,) and near-adjacent observations in PACF plot

- From the ACF and PACF plots, we can see a strong correlation with the adjacent observation (lag = 1) and also at a lag of 19,24,25, which is the value of T.



In the above correlation, plot dotted lines represent the confidence band. The centre dotted line represents the mean, and the upper and lower dotted lines represent boundaries based on a 95% confidence interval.

We can see a positive correlation with delays up to 15, where the ACF plot breaks the upper confidence barrier. Although we have a decent correlation up to the 20th lag, we cannot utilize all of them since it would cause multi-collinearity. Therefore we use the PACF plot to retrieve only the most important lags!

Since lag 0 is autocorrelation, we observe a significant Lag in ACF at 1, 2, 3,4,

We can make the following observations:

- Several autocorrelations are significantly non-zero. Therefore, the time series is non-random.
- High degree of autocorrelation between adjacent (lag = 1,2,3,4) and near-adjacent observations in PACF plot
- From the ACF and PACF plots, we can see a strong correlation with the adjacent observation (lag = 1) and also at a lag of 16,17,17.5, which is the value of T.

SARIMA Model Summary Interpretation

SARIMAX stands for Seasonal Autoregressive Integrated Moving Average with Exogenous Regressors. The basics are very self-explanatory:

We are attempting to forecast the dependent variable.

Model - The type of model that we are using. ARIMA, MA, and AR Date - The date on which the model was run.

Time - The date the model was finished.

The data range is known as the sample.

No. Observations - The total number of observations recorded.

The constant beta represents the independent variables. From the equation above, the error term is sigma2 or epsilon. AR. L1, ar. L2, and ar. L3 are our lag variables. After ensuring that we did not make any fundamental errors with our model, The next step is to evaluate the level of significance of output. This involves testing if the relationship between each term in the model to the response is statistically significant; we compare each term's p-value (its probability of being significantly different from 0) to a significance threshold. If a term's p-value is less than this significance threshold, we can conclude that it is substantially different from 0. The threshold can be set at any value between 0 and 1, but many use 5% as their alpha or significance level. A significance level of 0.05 means there's a 5% chance of finding that the term is not significantly different from 0 when it is.

Then we can conclude that:

P-value ≤ α: When the term is statistically significant

If the p-value is less than or equal to the significance level, you can conclude that the coefficient is statistically significant.

P-value > α: When the term is not statistically significant

If the p-value exceeds the significance level, you cannot conclude that the coefficient is statistically significant. You may want to refit the model without the term

Using the Log-Likelihood test to evaluate model performance

The log-likelihood value is a more straightforward representation of the most significant likelihood estimate. It is constructed by taking logs of previous values. This number is meaningless; however, comparing different models can be beneficial. The higher the log-likelihood, in general, the better. However, this statistic should be used with other metrics to provide the optimal model performance.

AIC stands for Akaike's Information Criterion. It is a metric that can help you determine the strength of your model. It takes into account the most likely findings as well as the overall number of parameters. Because adding more parameters to your model always increases your maximum likelihood value, the AIC balances this by penalizing for the number of parameters, resulting in models with few variables but strong data fit. Examining the models with the lowest AIC is a clever method for determining the best one! This value decreases as the model perform better.

BIC (Bayesian Information Criterion) is similar to AIC in that it takes into account the BIC (Bayesian Information Criterion) is similar to AIC in that it considers the number of rows in your dataset. Again, the lower the BIC, the better your model's performance. In comparison to AIC, BIC penalizes models with complex parameters more. Both BIC and AIC are good feature selection numbers since they assist you in choosing the simplest version while simultaneously offering the most valid findings.

The Hannan-Quinn information criteria, abbreviated as HQIC, can also be used to select features. This is less prevalent than BIC or AIC. Let's have a look at the coefficients table now.

Each feature's relevance

the 'coef' column represents e.

• ar.L1 denotes the autoregressive term with a lag of one

• ma.L1 is moving average' terms with lags of 1. The ARMA equation includes all of these coefficients. A first-order model is used in this example. The greater the number of lags in our model, the longer the equation.

$$AR(1): Y_t = \mu + \phi * Y_{t-1}$$

$$MA(1): Y_t = \mu + \varepsilon_t + \phi_1 * \varepsilon_{t-1}$$

ARIMA

$$ARIMA(1): Y_t = \mu + \phi * Y_{t-1} + \mu + \varepsilon_t + \phi_1 * \varepsilon_{t-1}$$

The 'std err' columns estimate the predicted value's error. It indicates the magnitude of the residual error's influence on your calculated parameters (the first Column)

The **'z' is** equal to the values of 'coef' divided by 'std err'. It is thus the standardized coefficient.

The **P>|z|** column is the p-value of the coefficient.

It is imperative to check these p-values before you continue using the model. If any of these values are higher than your given threshold (usually 0.05), you might be using an unreliable

coefficient that might cause misleading results. In our example, all p-values are lower than 0.05, so this model looks good!

The last two columns represent the confidence intervals. In simple words, these values are the coefficient value minus (left Column) and (right Column) the given error margin.

Metrics	Standard Error	p-value (before Diff)	p-value(after) Diff	ARIMA Before Differencing	After Differencing
ar.L1	0.167	0.000	0.770	0.9110	0.3111
Ma.L1	0.409	0.638	0.932	0.1925	-0.0972
Log-Likelihood				3.709	67.170
AIC				12.581	-114.340
BIC				26.593	-100.328
HIQC				17.064	-109.858
Sample				31	31

China ARIMA Model and Summary Statistics

Metrics	Standard Error	p-value		ARIMA Before Differencing	After Differencing
ar.L1	0.274	0.000	0.860	0.9627	-0.0519
Ma.L1	0.236	0.001	0.000	-0.755	-0.7225
Log Likelihood				8.595	104.203
AIC				2.811	-188.407
BIC				20.653	-170.565
HIQC				9.427	-181.790
Sample				45	45
Sample				31	31

USA ARIMA Model and Summary Statistic

Model Result After Differencing

Our objective is to determine whether each of the coefficients in our model is statistically significant.

Hypothesis

Null Hypothesis

Each coefficient is not statistically significant.

Alternative Hypothesis

Each coefficient is statistically significant

Hypothesis testing entails determining if each item in the model is statistically significant because the Null hypothesis states that neither of the coefficients is statistically significant.

As a result, we want the p-value for each term to be less than 0.05 to reject the null hypothesis with statistically significant results.

From the Model summary result for the time series of the China GINI index in the table above, we find that some of the coefficients are not statistically significant, which include a.L1 and ma. L1 before differencing and a.L1 alone after differencing with 0.05 p-value threshold significant value.

While the Model summary result for the USA GINI index time series table above, we find that only ar. L1 is not statistically significant, while every other coefficient is statistically significant 0.05 p-value threshold significant value both before and after differences.

Using the Ljung-Box test on our statistical table

Instead of adopting the NULL hypothesis straight away, we check our model assumptions to ensure that our model fits the assumption of white noise, which is that all residuals are independent. The Ljung Box test, also known as the modified Box-Pierce test, is used to test that the errors are white noise

Above is a statistics table for the USA GINI index. At lag 1, the Ljung-Box (L1) (Q) test statistic is 0.18, as well as the p-value is 0.93 before differencing, the Prob(Q) is 0.01, and also the p-value is 0.93 after differencing. We cannot reject the null hypothesis that the errors are white noise because the probability is greater than 0.05.

Above is a statistics table for the China GINI index. The Ljung-Box (L1) (Q) test statistic at lag 1 is, before differencing, the Prob(Q) is 0.00, the p-value is 0.97, and after differencing, the Prob(Q) is 0.00, and the p-value is 1. We cannot reject the null hypothesis that the mistakes are white noise because the probability is greater than 0.05.

Using the Heteroscedasticity test on our statistical result

H =3.45, 0.02, 2.87, 0.05

1.99, 0.29, 1.71, 0.41

The homoscedasticity of the error residuals determines whether they are heteroscedastic or have shared variance. White's test is applied to the summary statistics. The summary statistics for the USA show that we reject the null hypothesis and have variance in our residuals in both scenarios, with a test statistic of 3.46 and a p-value of 0.02 before Differencing and a test statistic of 2.87 and a p-value of 0.05 after Differencing.

The heteroscedasticity of the error residuals determines whether they share the same variance or not. White's test is applied to the summary statistics. Before Differencing, our summary statistics for China show a test statistic of 1.99 and a p-value of 0.29, and after Differencing, a test statistic of 1.71 and a p-value of 0.41 after Differencing, indicating that we accept the null hypothesis and have no variation in our residuals in both scenario.

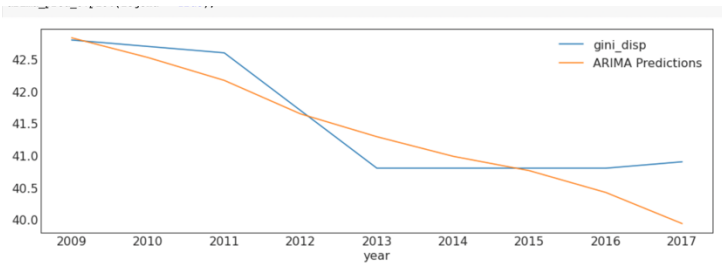
Jarque-Bera checks for error normalcy. It compares the null hypothesis that the data is regularly distributed to an alternate hypothesis of another distribution.

	CHINA		USA		
	Before Differencing	After Differencing	Before Differencing	After Differencing	
Jarque-Bera (JB):	4.23	5.88	3.76	3.38	
Prob(JB):	0.12	0.05	0.15	0.18	
Skew	-0.88	-0.87	-0.50	-0.45	
Kurtosis	3.56	**4.29**	4.03	4.02	

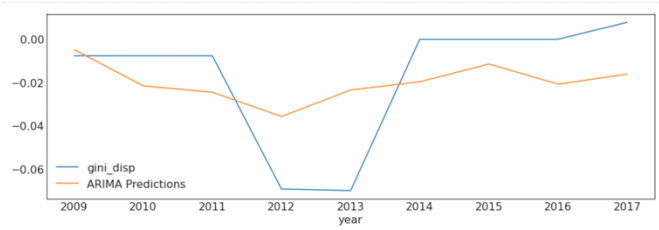
Before differencing, the Jarque-Bera (JB): test statistics is 4.23, with a corresponding p-value of 0.12. After differencing, the (JB): test statistics is 5.88, with a p-value of 0.05. In addition, the Jarque-Bera test reveals that the distribution for CHINA time series GINI index data shows a negative skew and a slight kurtosis. In the first scenario, we accept the NULL hypothesis, but after applying differencing, we reject it since the p-value is greater before and smaller after differencing. Before differencing, the Jarque-Bera (JB): test statistics is 3.38, and the associated p-value is 0.18, whereas the (JB): test statistics is 5.88 and the p-value is 0.05. In addition, the Jarque-Bera test reveals that the distribution has a negative skew and a small kurtosis for the CHINA time series GINI index data.

Fit our Analysis to model

China ARIMA Predicting of Wealth Inequality using past GINI

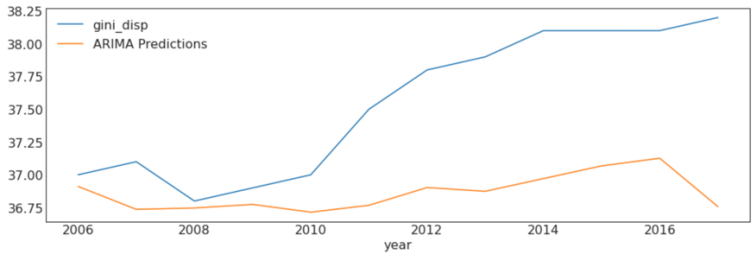


Line Plot showing Model Performance comparing Real data and ARIMA Prediction for China before Differencing the Dataset

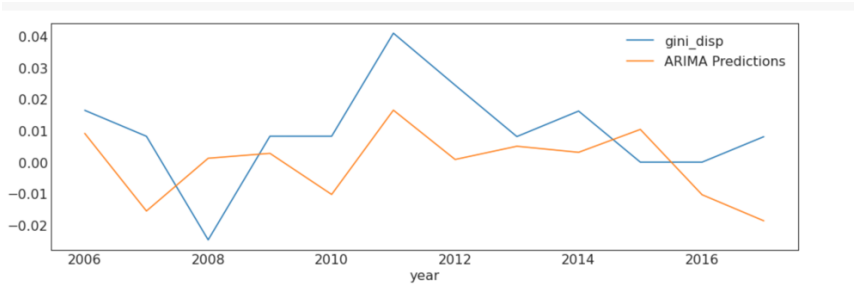


Line Plot showing Model Performance comparing Real data and ARIMA Prediction for China After Differencing Time series Dataset

USA ARIMA Predicting of Wealth Inequality using past GINI



Line Plot showing Model Performance comparing Rthe



Line Plot showing Model Performance comparing Real data and ARIMA Prediction for USA after Differencing the Dataset

The above plots show that differencing our time series dataset to remove non-stationarity helps the convergence of actual value estimates when fitting an autoregressive integrated moving average model to both China and USA data. The ARIMA model fitting better on USA data due to more data points or years of observations available, while in China we can see that it performs better with less data points or years of observations. We conclude that our model performs better with more data points or years of observations available.

Model Evaluation

To evaluate our model, we used the. The root mean square error (RMSE).

The root mean square error (RMSE) is a statistic that shows us, on average, how much our projected values differ from our actual values in a model. It is computed as follows:

$$\mathbf{RMSE} = \sqrt{[\ \Sigma(P_i - A_i)^2 / n\ ]}$$

where:

- Σ means "sum"
- P_i is the predicted value of the i^{th} observation
- A_i is the actual/ observed value of the i^{th} observation
- n stands for sample size

Source: Statology.org

Root Mean square error result is thus:

	ARIMA	ARIMA (After Differencing)	LSTM
China	0.418	0.024	5.430
USA	0.817	0.018	

DISCUSSION & CONCLUSION

Overall, my data strongly support that global and Local wealth inequality distribution can be describe, assessed, and predicted using statical and data science approach. It appears that our statistical tool was able to study the trends of Wealth inequality in various country using GINI index time series data and the concentration of wealth by combining GINI index and other important Macro Economics variables like GDP per capita, Population, Life expectance to show different clusters of wealth inequality.

We Forecasting Models also performed favourable when tested on two countries Time series data but was limited because the of lack of sufficient time series data. While also try to use multivariant regression, we experienced some difficulty due to the omission of some time series data. We can easily see how wealth inequality is changing over time in various countries by comparing different models applied on various variables like Gini Index or income inequality or Concentration of wealth or life expectancy etc. It appears that we have successfully used a statistical tool to assess changes in global wealth inequality over last 20 years where we could see that there has been a significant change in the distribution between rich and poor countries with more than 75% poor countries having less than 1% wealthiest country during 1990-2010 period.Our research objective was to examine the distribution of wealth and estimate countries' Gini coefficient using descriptive analysis in a machine-learning mode. We used statistical tools to test whether wealth inequality is a local phenomenon or if there is evidence

of a global concentration of wealth using the Statistical tools learning model, which can then be used to make predictions about future wealth Gini coefficients.

Although great progress has been made in the study of wealth distribution, there is still a need to conduct more research on this topic. For example, Pickety and others have made great strides in their studies, but it is important not to overlook subjectivity when trying to understand inequality trends. Creating a technique that better captures subjective aspects of social phenomena would be very beneficial for future research on this issue. Some traditional model selection techniques are subjective, and the reliability of the chosen model may be based on the forecaster's aptitude and expertise (although this criticism is also applicable to other modelling approaches as well). It is not based on any theoretical paradigms or structural connections. As a result, the economic significance of the chosen model is unknown. ARIMA models, unlike structural models, cannot be used to simulate policy (Meyler et al., 1998). ARIMA models are backwards-looking, cannot predict turning points, and do not usually anticipate long-run equilibrium. However, ARIMA models are highly robust and forecast short-run inflation more accurately than more complicated structural models. In this study, the ARIMA forecasting approach will be used as a standard against which other forecasting techniques may be evaluated and as an input into forecasting. These models also tend to rely heavily on past data so if the economy is about to undergo a major shift in fundamentals (e.g., if interest rates are about to be cut dramatically), an ARIMA model may not anticipate the change. ARIMA models cannot handle nonlinear relationships between variables or outliers such as extreme values. Finally, ARIMA does not consider seasonality or trend. It may overfit when applied to seasonally adjusted data. A large time series of data is

required. Box Jenkins and objective penalty function techiques are difficult to use with too little data. If a large time series is excluded from a statistical break, you may just have to analyze a subset of the entire data series or rely on dummy variables. As a result, avoiding structural breaks and working with a smaller data set may conflict with robustness and statistical validity. In The future, it would be helpful if we have more time series data available on various countries for further analysis like China, India and Brazil. In conclusion I believe that statistical methods are a very useful tool in studying wealth inequality especially when we have sufficient time series data available for analysis.

Ahmad, S., Abdollahian, M., Zeephongsekul, P., & Abbasi, B. (2009). Multivariate nonnormal process capability analysis. *The International Journal of Advanced Manufacturing Technology 2009 44:7*, *44*(7), 757–765. https://doi.org/10.1007/S00170-008-1883-9

Arnold, B. C. (2014). Pareto Distribution. *Wiley StatsRef: Statistics Reference Online*. https://doi.org/10.1002/9781118445112.STAT01100

Arnold, B. C. (2015). Pareto Distribution. *Wiley StatsRef: Statistics Reference Online*, 1–10. https://doi.org/10.1002/9781118445112.STAT01100.PUB2

Babu, C. N., & Reddy, B. E. (2014). A moving-average filter based hybrid ARIMA-ANN model for forecasting time series data. *Applied Soft Computing Journal*, *23*, 27–38. https://doi.org/10.1016/j.asoc.2014.05.028

Balakrishnan, N., & Nevzorov, V. B. (2005). Pareto Distribution. *A Primer on Statistical Distributions*, 133–138. https://doi.org/10.1002/0471722227.CH15

Bennett, S. (1983). Log-Logistic Regression Models for Survival Data. *Applied Statistics*, *32*(2), 165. https://doi.org/10.2307/2347295

Blanchet, T., Fournier, J., & Piketty, T. (2022). Generalized Pareto Curves: Theory and Applications. *Review of Income and Wealth*, *68*(1), 263–288. https://doi.org/10.1111/ROIW.12510

Burr, I. W. (1968). On a General System of Distributions: III. The Sample Range. *Journal of the American Statistical Association*, *63*(322), 636. https://doi.org/10.2307/2284034

Champernowne, D. G. (1952). The Graduation of Income Distributions. *Econometrica*, *20*(4), 591. https://doi.org/10.2307/1907644

Cingano, F. (2014). *Trends in income inequality and its impact on economic growth*.

Cook, R. D., & Johnson, M. E. (1986). Generalized burr-pareto-logistic distributions with applications to a uranium exploration data set. *Technometrics*, *28*(2), 123–131. https://doi.org/10.1080/00401706.1986.10488113

Domingos, D. S., de Oliveira, J. F. L., & de Mattos Neto, P. S. G. (2019). An intelligent hybridization of ARIMA with machine learning models for time series forecasting. *Knowledge-Based Systems*, *175*, 72–86. https://doi.org/10.1016/j.knosys.2019.03.011

Domma, F. (2009). Some properties of the bivariate Burr type III distribution. *Http://Dx.Doi.Org/10.1080/02331880902986547*, *44*(2), 203–215. https://doi.org/10.1080/02331880902986547

Elamin, N., & Fukushige, M. (2018). Modeling and forecasting hourly electricity demand by SARIMAX with interactions. *Energy*, *165*, 257–268. https://doi.org/10.1016/j.energy.2018.09.157

Esling, P., & Agon, C. (2012a). Time-Series data mining. *ACM Comput. Surv. 45, 1, Article*, *12*. https://doi.org/10.1145/2379776.2379788

Esling, P., & Agon, C. (2012b). Time-series data mining. *ACM Computing Surveys*, *45*(1). https://doi.org/10.1145/2379776.2379788

Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, *33*(1), 3–56. https://doi.org/10.1016/0304-405X(93)90023-5

Fan, R. Y. C., Ng, S. T., & Wong, J. M. W. (2010). Reliability of the Box–Jenkins model for forecasting construction demand covering times of economic austerity. *Http://Dx.Doi.Org/10.1080/01446190903369899*, *28*(3), 241–254. https://doi.org/10.1080/01446190903369899

Fisk, P. R. (1961). The Graduation of Income Distributions. *Econometrica*, *29*(2), 171. https://doi.org/10.2307/1909287

Forbes, C., Evans, M., Hastings, N., & Peacock, B. (2010). Pareto Distribution. *Statistical Distributions*, 149–151. https://doi.org/10.1002/9780470627242.CH34

Gan, M., Cheng, Y., Liu, K., & Zhang, G. L. (2014). Seasonal and trend time series forecasting based on a quasi-linear autoregressive model. *Applied Soft Computing Journal*, *24*, 13–18. https://doi.org/10.1016/J.ASOC.2014.06.047

Geisser, S. (1984). Predicting Pareto and exponential observables. *Canadian Journal of Statistics*, *12*(2), 143–152. https://doi.org/10.2307/3315178

Ho, S. L., & Xie, M. (1998). The use of ARIMA models for reliability forecasting and analysis. *Computers & Industrial Engineering*, *35*(1–2), 213–216. https://doi.org/10.1016/S0360-8352(98)00066-7

Hose, G. C. (2005). Assessing the need for groundwater quality guidelines for pesticides using the species sensitivity distribution approach. *Human and Ecological Risk Assessment*, *11*(5), 951–966. https://doi.org/10.1080/10807030500257788

Hutchinson, T. P. (1979). Four Applications of a Bivariate Pareto Distribution. *Biometrical Journal*, *21*(6), 553–563. https://doi.org/10.1002/BIMJ.4710210605

Introduction to Multiple Time Series Analysis - Helmut Lütkepohl - Google Books. (n.d.). Retrieved August 17, 2022, from https://books.google.co.uk/books?hl=en&lr=&id=qJXsCAAAQBAJ&oi=fnd&pg=PA1&dq=The+prediction+can+be+done+based+on+three+different+time+spaces:+linear,+nonlinear,+and+autoregressive.+Specification+testing+may+also+be+used+as+a+tool+for+time+series&ots=vRT8aqdtZT&sig=9V6qtW9WsWc3XtMLID1moynT4gw&redir_esc=y#v=onepage&q&f=false

Jebb, A. T., Tay, L., Wang, W., & Huang, Q. (2015). Time series analysis for psychological research: Examining and forecasting change. *Frontiers in Psychology*, *6*(JUN), 727. https://doi.org/10.3389/FPSYG.2015.00727/ABSTRACT

Keelin, T. W. (2016). The Metalog Distributions. *Https://Doi.Org/10.1287/Deca.2016.0338*, *13*(4), 243–277. https://doi.org/10.1287/DECA.2016.0338

Lee, S. J., & Siau, K. (2001). A review of data mining techniques. *Industrial Management and Data Systems*, *101*(1), 41–46. https://doi.org/10.1108/02635570110365989/FULL/PDF

Lomax, K. S. (1954). Business Failures: Another Example of the Analysis of Failure Data. *Journal of the American Statistical Association*, *49*(268), 847. https://doi.org/10.2307/2281544

Matis, J. H., Kiffe, T. R., van der Werf, W., Costamagna, A. C., Matis, T. I., & Grant, W. E. (2009). Population dynamics models based on cumulative density dependent feedback: A link to the logistic growth curve and a test for symmetry using aphid data. *Ecological Modelling*, *220*(15), 1745–1751. https://doi.org/10.1016/J.ECOLMODEL.2009.04.026

Meyler, A., Kenny, Quinn, G., & Terry. (1998). *Forecasting irish inflation using ARIMA models*.

Renard, P., Alcolea, A., & Ginsbourger, D. (2013). Stochastic versus Deterministic Approaches. *Environmental Modelling: Finding Simplicity in Complexity: Second Edition*, 133–149. https://doi.org/10.1002/9781118351475.CH8

Rodriguez, R. N. (1977a). A Guide to the Burr Type XII Distributions. *Biometrika*, *64*(1), 129. https://doi.org/10.2307/2335782

Rodriguez, R. N. (1977b). A guide to the Burr type XII distributions. *Biometrika*, *64*(1), 129–134. https://doi.org/10.1093/biomet/64.1.129

Saputro, D. R. S., Pratiwi, N. B. I., & Kusumawati, R. (2022). *Logistic smooth transition autoregressive model parameter estimation using Gauss Newton*. 020031. https://doi.org/10.1063/5.0100105

Singh, S. K., & Maddala, G. S. (1976). A Function for Size Distribution of Incomes. *Econometrica*, *44*(5), 963. https://doi.org/10.2307/1911538

STIGLITZ, J. E. (1981). Pareto Optimality and Competition. *The Journal of Finance*, *36*(2), 235–251. https://doi.org/10.1111/J.1540-6261.1981.TB00437.X

Taam, W., Subbaiah, P., & Liddy, J. W. (1993). A note on multivariate capability indices. *J Appl Stat*, *20*(3), 339–351. https://doi.org/10.1080/02664769300000035

Tadikamalla, P. R. (1980). A Look at the Burr and Related Distributions. *International Statistical Review / Revue Internationale de Statistique*, *48*(3), 337. https://doi.org/10.2307/1402945

Tadikamalla, P. R., & Johnson, N. L. (1982). Systems of Frequency Curves Generated by Transformations of Logistic Variables. *Biometrika*, *69*(2), 461. https://doi.org/10.2307/2335422

Takahasi, K. (1965). Note on the multivariate Burr's distribution. *Ann Inst Stat Math (Tokyo)*, *17*(1), 257–260. https://doi.org/10.1007/bf02868169

Teräsvirta, T., van Dijk, D., & Medeiros, M. C. (2005). Linear models, smooth transition autoregressions, and neural networks for forecasting macroeconomic time series: A re-examination. *International Journal of Forecasting*, *21*(4), 755–774. https://doi.org/10.1016/J.IJFORECAST.2005.04.010

Yue, S., & Pilon, P. (2003). Interaction between deterministic trend and autoregressive process. *Water Resources Research*, *39*(4). Ihttps://doi.org/10.1029/2001WR001210

CODE REFERENCE

https://medium.com/swlh/visualizing-taxes-impact-on-social-inequality-using-python-c5c2f2951e45

https://www.kaggle.com/code/iamleonie/time-series-interpreting-acf-and-pacf

https://www.bauer.uh.edu/rsusmel/phd/ec2-4.pdf

https://2019.qmplus.qmul.ac.uk/pluginfile.php/1974029/mod_resource/content/14/notes.pdf

https://iopscience.iop.org/article/10.1088/1742-6596/1879/3/032008/pdf

https://www.youtube.com/watch?v=IKY_uDiSe8U

https://machinelearningmastery.com/time-series-trends-in-python/

- **Machine Learning Mastery: How to Check if Time Series Data is Stationary with Python**

https://medium.com/analytics-vidhya/interpreting-arma-model-results-in-statsmodels-for-absolute-beginners-a4d22253ad1c

https://support.minitab.com/en-us/minitab/21/help-and-how-to/statistical-modeling/time-series/how-to/arima/interpret-the-results/key-results/?SID=117600&SID=117600#step-2-determine-how-well-the-model-fits-the-data

APENDIX

Table 1 SWIID Data Dictionary

Variable Name	Description
gini_disp:	Estimate of Gini index of inequality in equivalized disposable income,
gini_mkt:	Estimate of Gini index of inequality in equivalized market income,
abs_red:	Estimated absolute redistribution, the number of Gini-index points market income is reduced due to taxes and transfers
rel_red:	Estimated relative redistribution, where is the difference between the gini_mkt and gini_disp divided by, multiplied by 100.

Table 2 Data Sample

	country	year	gini_disp	Gini_g_level
3054	Maldives	2004	39.3	adequate equality
5330	United States	1967	31.1	relative equality
69	Andorra	2013	30.6	relative equality
1684	Georgia	2007	39.6	adequate equality
1757	Ghana	1990	40.0	adequate equality
437	Belgium	1982	23.4	perfect equality
2314	Ireland	2012	30.2	relative equality
374	Barbados	1989	43.9	adequate equality
3440	Netherlands	1999	24.5	perfect equality
3	Afghanistan	2010	31.7	relative equality

The Gapminder;

Table 3 Gapminder Data Dictionary

country	
continent	
year	
lifeExp	life expectancy at birth **The Average number of years a child is expected to live if the current rate of mortality pattern remains constant**

pop	Total population
GDP per cap	per-capita GDP

Table 4 Gapminder Data Sample

	country	continent	year	lifeExp	pop	gdpPercap	iso_alpha	iso_num
267	Chad	Africa	1967	43.601	3495967	1196.810565	TCD	148
1372	Slovak Republic	Europe	1972	70.350	4593433	9674.167626	SVK	703
1641	Venezuela	Americas	1997	72.146	22374398	10165.495180	VEN	862
103	Bangladesh	Asia	1987	52.819	103764241	751.979403	BGD	50
21	Albania	Europe	1997	72.950	3428038	3193.054604	ALB	8

Table 5 Merge Dataset(SWIID + Gapminder) Data Sample

```
[ ] allData.sample(4)
```

	country	year	gini_disp	Gini_g_level	continent	lifeExp	pop	gdpPercap	iso_alpha	iso_num
226	Guatemala	2007	47.2	innequality	Americas	70.259	12572928	5186.050003	GTM	320
166	Egypt	2002	41.6	adequate equality	Africa	69.806	73312559	4754.604414	EGY	818
214	Greece	1977	36.9	adequate equality	Europe	73.680	9308479	14195.524280	GRC	300
592	Syria	2007	36.3	adequate equality	Asia	74.143	19314747	4184.548089	SYR	760

Table 6 USA GINI Index TimTime-series ADF Python Output:

```
from statsmodels.tsa.stattools import adfuller
dftest = adfuller(usa.gini_disp, autolag = 'AIC')
print("1. ADF : ",dftest[0])
print("2. P-Value : ", dftest[1])
print("3. Num Of Lags : ", dftest[2])
print("4. Num Of Observations Used For ADF Regression and Critical Values Calculation :", dftest[3])
print("5. Critical Values :")
for key, val in dftest[4].items():
    print("\t",key, ": ", val)
```

```
1. ADF :  0.8634791417940741
2. P-Value :  0.9925947221193557
3. Num Of Lags :  0
4. Num Of Observations Used For ADF Regression and Critical Values Calculation : 56
5. Critical Values :
         1% :  -3.552928203580539
         5% :  -2.9147306250000002
         10% :  -2.595137155612245
```

China

Table 7 China GINI index Time Series ADF Python Output

```
from statsmodels.tsa.stattools import adfuller
dftest = adfuller(china.gini_disp, autolag = 'AIC')
print("1. ADF : ",dftest[0])
print("2. P-Value : ", dftest[1])
print("3. Num Of Lags : ", dftest[2])
print("4. Num Of Observations Used For ADF Regression and Critical Values Calculation :", dftest[3])
print("5. Critical Values :")
for key, val in dftest[4].items():
    print("\t",key, ": ", val)
```

```
1. ADF :  -1.7630548309555902
2. P-Value :  0.39895198210341654
3. Num Of Lags :  1
4. Num Of Observations Used For ADF Regression and Critical Values Calculation : 38
5. Critical Values :
         1% :  -3.6155091011809297
         5% :  -2.9412623574865l4
         10% :  -2.6091995013850418
```

Table 8 USA ARIMA model result Without Differencing

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.9627	0.274	3.514	0.000	0.426	1.500
ma.L1	-0.7551	0.236	-3.200	0.001	-1.218	-0.293
ar.S.L12	-0.4263	767.000	-0.001	1.000	-1503.718	1502.866
ar.S.L24	-0.0195	682.157	-2.86e-05	1.000	-1337.023	1336.984
ar.S.L36	-0.4559	396.383	-0.001	0.999	-777.351	776.440
ar.S.L48	-0.9565	313.140	-0.003	0.998	-614.699	612.786
ma.S.L12	-0.3042	2060.689	-0.000	1.000	-4039.180	4038.571
ma.S.L24	-0.3558	7679.059	-4.63e-05	1.000	-1.51e+04	1.51e+04
ma.S.L36	-0.2768	5000.373	-5.53e-05	1.000	-9800.828	9800.274
sigma2	0.0017	2.569	0.001	0.999	-5.034	5.037

Ljung-Box (L1) (Q):	0.18	**Jarque-Bera (JB):**	3.76
Prob(Q):	0.67	**Prob(JB):**	0.15
Heteroskedasticity (H):	3.45	**Skew:**	-0.50
Prob(H) (two-sided):	0.02	**Kurtosis:**	4.03

Table 9 USA ARIMA Model result with Differencing

ar.L1	-0.0519	0.294	-0.177	0.860	-0.627	0.524
ma.L1	-0.7225	0.205	-3.528	0.000	-1.124	-0.321
ar.S.L12	0.1324	0.128	1.033	0.301	-0.119	0.384
ar.S.L24	0.0390	0.162	0.240	0.810	-0.279	0.357
ar.S.L36	-0.3980	0.573	-0.695	0.487	-1.520	0.724
ar.S.L48	-0.0910	0.139	-0.657	0.511	-0.363	0.181
ma.S.L12	0.0841	0.149	0.566	0.571	-0.207	0.375
ma.S.L24	-0.1161	0.178	-0.653	0.514	-0.465	0.233
ma.S.L36	-0.4287	0.258	-1.664	0.096	-0.934	0.076
sigma2	0.0002	0.000	0.842	0.400	-0.000	0.001

Ljung-Box (L1) (Q):	0.01	**Jarque-Bera (JB):**	3.38
Prob(Q):	0.93	**Prob(JB):**	0.18
Heteroskedasticity (H):	2.87	**Skew:**	-0.45
Prob(H) (two-sided):	0.05	**Kurtosis:**	4.02

Table 10 China ARIMA Model Without Differencing

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.9110	0.167	5.466	0.000	0.584	1.238
ma.L1	0.1925	0.409	0.471	0.638	-0.608	0.993
ar.S.L12	0.1069	5812.259	1.84e-05	1.000	-1.14e+04	1.14e+04
ar.S.L24	-0.3982	5434.992	-7.33e-05	1.000	-1.07e+04	1.07e+04
ar.S.L36	0.5826	3357.167	0.000	1.000	-6579.343	6580.509
ar.S.L48	0.2579	6983.873	3.69e-05	1.000	-1.37e+04	1.37e+04
ma.S.L12	0.1252	9157.719	1.37e-05	1.000	-1.79e+04	1.79e+04
ma.S.L24	-0.2278	5702.259	-3.99e-05	1.000	-1.12e+04	1.12e+04
ma.S.L36	0.3953	2658.296	0.000	1.000	-5209.768	5210.559
sigma2	0.0086	23.340	0.000	1.000	-45.738	45.755

Ljung-Box (L1) (Q):	0.00	**Jarque-Bera (JB):**	4.23
Prob(Q):	0.97	**Prob(JB):**	0.12
Heteroskedasticity (H):	1.99	**Skew:**	-0.88
Prob(H) (two-sided):	0.29	**Kurtosis:**	3.56

Table 11China ARIMA Model Result with Differencing

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.3111	1.064	0.292	0.770	-1.774	2.396
ma.L1	-0.0972	1.133	-0.086	0.932	-2.317	2.123
ar.S.L12	-0.1336	0.245	-0.546	0.585	-0.614	0.346
ar.S.L24	-0.2983	0.495	-0.603	0.547	-1.268	0.671
ar.S.L36	0.0236	0.204	0.116	0.908	-0.376	0.423
ar.S.L48	0.0081	0.300	0.027	0.979	-0.580	0.597
ma.S.L12	-0.1133	0.141	-0.801	0.423	-0.391	0.164
ma.S.L24	-0.3487	0.292	-1.196	0.232	-0.920	0.223
ma.S.L36	0.0757	0.170	0.445	0.657	-0.258	0.409
sigma2	0.0002	0.000	1.053	0.292	-0.000	0.001

Ljung-Box (L1) (Q):	0.00	**Jarque-Bera (JB):**	5.88
Prob(Q):	1.00	**Prob(JB):**	0.05
Heteroskedasticity (H):	1.71	**Skew:**	-0.87
Prob(H) (two-sided):	0.41	**Kurtosis:**	4.29

Table 12 USA Evaluation Metrics Result Before Differencing

SARIMAX Results

Dep. Variable:	gini_disp	**No. Observations:**	45
Model:	SARIMAX(1, 1, 1)x(4, 0, [1, 2, 3], 12)	**Log Likelihood**	8.595
Date:	Sat, 27 Aug 2022	**AIC**	2.811
Time:	09:24:28	**BIC**	20.653
Sample:	0	**HQIC**	9.427
	- 45		
Covariance Type:	opg		

Table 13 USA Evaluation Metrics After Differencing

Model:	SARIMAX(1, 1, 1)x(4, 0, [1, 2, 3], 12)	**Log Likelihood**	104.203
Date:	Sat, 27 Aug 2022	**AIC**	-188.407
Time:	09:25:42	**BIC**	-170.565
Sample:	0	**HQIC**	-181.790
	- 45		
Covariance Type:	opg		

Table 14 China Evaluation Metrics Before Differencing

SARIMAX Results

Dep. Variable:	gini_disp	**No. Observations:**	31
Model:	SARIMAX(1, 1, 1)x(4, 0, [1, 2, 3], 12)	**Log Likelihood**	3.709
Date:	Sat, 27 Aug 2022	**AIC**	12.581
Time:	10:38:11	**BIC**	26.593
Sample:	0	**HQIC**	17.064
	- 31		
Covariance Type:	opg		

Table 15 China Evaluation Metrics After Differencing

SARIMAX Results

Dep. Variable:	gini_disp	**No. Observations:**	31
Model:	SARIMAX(1, 1, 1)x(4, 0, [1, 2, 3], 12)	**Log Likelihood**	67.170
Date:	Sat, 27 Aug 2022	**AIC**	-114.340
Time:	09:53:38	**BIC**	-100.328
Sample:	0	**HQIC**	-109.858
	- 31		
Covariance Type:	opg		

A correlation among the numeric variables was done and plotted using a heatmap from the seaborn library, and we discovered a strong correlation among variables

Using the Gini index, GDP per capita, and life expectancy to predict wealth inequality using

```
ax = sns.heatmap(allDataNumeric1.corr(), annot=True)
```