Analysis and prediction of GDP for countries in European Union

Samuel Posejpal | 914829

December 2020

Department of Computer Science

Swansea University

"Project Dissertation submitted to Swansea University in Partial Fulfilment for the Degree of Master of Science."

Declarations/Statements

DECLARATION

This work has not previously been accepted in substance for any degree and is not being concurrently submitted in candidature for any degree.

STATEMENT 1

This dissertation is the result of my own independent work/investigation, except where otherwise stated. Other sources are acknowledged by giving explicit references. A bibliography is appended.

STATEMENT 2

I hereby give consent for my dissertation, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organizations.

Signature: Samuel Posejpal Date: 13 of December 2020

Abstract

As we live in the information age and the decision-makers and scientists aspire to gain valuable insights from data, it is necessary to investigate and provide solutions in order to replace ineffective techniques. Such practical methods of data handling and time series prediction will be put to the test. Data science is an interdisciplinary field that uses various scientific methods, processes, algorithms and systems to gain knowledge and insights from structured and unstructured data. This project specification focuses on the intersection of Computer Science and most precisely Data Science with Economy. These areas will be explored together with clear focus on time series forecasting using ARIMA method. Furthermore possible future pathways for research in these domains will be explored, especially transparency and how can one area implement the other.

Table of Contents

Contents

1 Introduction

1.1 Problem formulation

The purpose of this thesis is to identify a general model to forecast Gross Domestic Product (GDP) and its growth for the countries within European Union. If the model provides reliable results for this region, then the model should be able to forecast GDP growth for other developed countries of interest like the US. GDP is a monetary measure of the market value of all the final goods and services produced within a country and is usually measured in a yearly or quarterly period. It is commonly used to determine the economic performance of a country or regions. Time series provide the opportunity to forecast future values. Based on previous values, time series can be used to forecast trends in economics, weather or even capacity planning. The specific properties of time-series data require that specialized statistical methods are used. This work will begin by introducing and discussing the concepts of autocorrelation, stationarity, and seasonality, and proceed to apply one of the most commonly used method for time-series forecasting, known as ARIMA. The World bank database was used to obtain the dataset for modelling.

1.2 Motivation

GDP is one of the most important measures of economic well-being. Technological advances have contributed enormous benefits for human's life, as well as disappointment over measured productivity and output growth in recent years, have spurred widespread concerns about whether our statistical systems are capturing these improvements (Dynan, Sheiner, 2018) Given the importance of GDP data in modern society, from becoming an election narrative to influencing the global commodity market, it is imperative that proper research should be done for GDP modelling for countries during economic crisis. The purpose of this study is to analyse historic data and explore econometric models for time series analysis. Secondly, this study aims to undertake a comprehensive analysis of GDP forecasting and modelling for a developed country which could be used further as a template and reference for modelling GDP of European Union. Additionally, the transparency of previous theses will be explored, and these works recreated to address significant issue within Data Science. Such importance of research in GDP modelling for European Union countries serves as my main motivation to undertake this study. Aim is to come with a holistic approach towards modelling of GDP numbers and set a template which could be widely and easily used.

1.3 Literature review

The use of ARIMA models for GDP forecasting was started with the seminal Box and Jenkins (1976) paper. Zhang Haonan (2013) focused on a developed country Sweeden and results showed that 1st order ARMA had the most significant results. Dritsaki (2014) used Greece GDP data from 1980 to 2013 to fit an ARIMA model and the prediction showed an upward trend in the growth of GDP. Waboma et al (2015) predicted Kenya's GDP and found out that ARIMA fits their data best with forecast in sample being 5% close the actual numbers. According to these authors ARIMA is one of the most used and accurate methods. AR and ARIMA models are models from within the same class of models and in one sense they share the same degree of complexity and therefore a properly constructed ARIMA

model will always produce more optimal results when chosen correctly. ARIMA is more general than AR and ARMA and thus in majority cases it is a clear candidate.

1.4 Challenges

When trying to recreate former predictions made in academic papers, many academic papers did not clearly show the datasets used and many such sources are currently updated or do not exist. For example when following Varun Agrawal's thesis 'GDP Modelling And Forecasting Using Arima: An Empirical Study From India' (2018), certain datasets from Indian Central Bank that were declared to be the bases for the thesis could not be easily found as the link to some of them wasn't present in the thesis. On the other hand, 'Total Gross Domestic Product for India' data retrieved from FRED was clearly pointed out and the link to the source was shared. Furthermore, there were different datasets which used various levels of prices and according to Agrawal, Indian Central Bank changed the GDP calculation method in 2015 further complicating the continuity of time series analysis. (Agrawal,2018) Central European University, which enables access to this thesis also did not share the sources in form of for example csv files on its website. Jahnke and Asher (2012), explain that there are barriers to data curation, because only few researchers think about long-term preservation of the data. Metadata and documentation are not being strictly taken care in all the cases.

Another difficulty and barrier to transparency is use of different software throughout various theses. Agrawal (2018) used STATA software and for the detailed recreation and verification, a researcher would have to invest time into learning how to use this or other econometric software or specific programming language. While running projects in programming language Python online, most of the notebooks reported errors due to the fact that they were written using outdated library functions and older versions of programming language. Certain python libraries have also unclear documentation. Berger et.all (2019) confirmed association between certain programming languages and software defects in projects hosted on GitHub. Python has such association higher than average programming language and it was also found that defect types are strongly associated with languages. Amongst researchers there is therefore a need for more effective collaboration tools, or online spaces that support the volume of data generated and provide appropriate privacy and access controls (Jahnke and Asher, 2012).

For the visualisation, picking the correct plots was a challenge as some of the basic ones were not clearly visible. When using econometric software like SPSS, there wasn't freedom in interacting with the data and plots as SPSS aims to be user friendly by providing the end results in a compact form, but it hides the internal functionality and it is not clear to understand how the programs are being conducted. Furthermore, many datasets do not include all the data needed in correct format and in once place, therefore data preparation and manipulation took place. As the World Bank dataset was chosen, even this high-quality dataset does not include GDP data for certain Eastern European countries before 1990s and year 2020 is empty. First results of my work were wrong as the differencing and decomposition was not conducted correctly.

1.5 Results and Novelty Insights

The trend of a time series refers to the general direction in which the time series is moving and, in this case, GDP of EU has shown to have a linear positive trend in the long run. Data

visualisation and tests showed that time series is non-stationarity and therefore decomposition was successfully conducted by applying additive decomposition and differencing the time series. Visual check and p value lower than 0.05 confirmed stationarity. Correct parameters and ARIMA model was chosen by calculating Akaike Information Criterion (AIC) for each model and then choosing the model with the lowest (AIC) value. ARIMA model (5,2,0) demonstrated close relationship with real GDP and after predicted GDP per capita for the next 15 years until 2032. GDP per capita in US dollars will according to the chosen ARIMA model rise from 2017 value of 40 000 to 60 000 in 2030. Very close alignment was also confirmed by exactly confirming GDP prediction for years 2018 and 2019 where ARIMA model predicted correct values of 44 000 and 46 000 US dollars for these years.

Regression lines for prediction were also produced. During the work it was demonstrated that it is possible that methods used will be deprecated in the future. Furthermore the automatization of selecting data is tricky as for example European Union loses United Kingdom as its member and new members may join later. Code includes European Union as selected option for this thesis and for future reference, countries can be then added and selected while extracting dataset from The World Bank website . Furthermore, commenting of the code enables to use these methods later and quickly add additional years into the equation. When selecting data, it is useful to add countries whose economies are interconnected and corelation is clearly visible. From the analysis it could be seen that countries within European Union have corelated economies. European Union was chosen, because the economies corelate even more due to factors like the funding from richer countries towards poorer countries and relative similarities and close distance of these countries.

1.6 Thesis structure

Following the introduction, review of the background theory is introduced and discussed. This includes economic concepts like GDP and its historical development. Time series methods and principles of autocorrelation, stationarity, and seasonality will be explained and then the work proceeds to apply one of the most commonly used methods for time-series forecasting, known as ARIMA. The World bank database was used to obtain the dataset for the modelling and its data will be examined together with methodology. After Data pre-processing, data analysis will demonstrate properties of data. This will summarize large datasets and identify relationships, trends and anomalies. Historical data will be used to predict future GDP values by ARIMA model (Master's in data science, 2020). Visualising the data analysis and outputs will help to better showcase the results and confirm for example stationarity. This dissertation will combine computer science and economy, meaning that GDP will be predicted through programming the models and the forecasting power of the model will be gradually improved. Transparency and ethical issues within data management and accessibility of data are analysed and conclusion given in Transparency and Replication section.

2 Gross Domestic Product – GDP

2.1 Overview

GDP acts as an aggregate measure of total economic production for a country and showcases the market value of all goods and services produced by the economy during certain period. This market value consists of personal consumption, government spending, private inventories or the foreign trade balance, where the exports are added, while imports are subtracted. It is a central interest for most researchers in the field of business and especially economics. Amongst macro economy variables, GDP is the primary concern. Furthermore GDP related data are regarded as an important index for exploring the economic development of countries and it is vital factor when assessing the operating status of the whole macro economy.(Ning, Wei & Kuan-jiang, Bian & Zhi-fa, Yuan, 2010) If the GDP is rising, it means that incomes are growing, and consumers are purchasing more. All of this means a stronger economy. GDP is further limited by country's borders whereas Gross National Product (GNP) on the other hand takes into account the value of goods produced by a country's residents regardless of whether they live inside the country or abroad. GDP is measured and forecasted by most financial institutions like World Bank or certain governmental institutions like Office for National Statistics in the UK (Wei et.all, 2010). There is a difference between sizes of countries and thus comparing economy of USA and GBR does not indicate the actual living conditions. Therefore GDP per capita portrayed on the figure xx illustrates this indication better and it can be understood that technological advances and other factors in modern society help to increase productivity worldwide, even though surges are notable. Moreover, this line plot also shows interconnection between lines of Great Britain and European Union as countries within the EU are more correlated together.

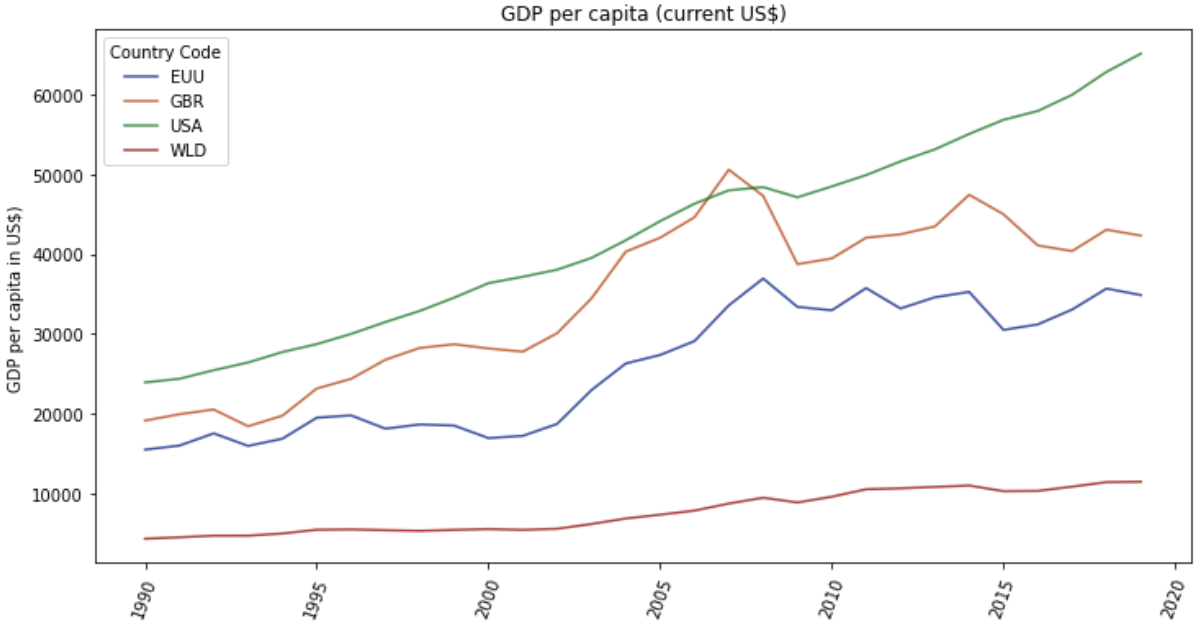

Figure 1. GDP per capita of selected countries

There are many types of GDP. For example, Real GDP is a calculation of GDP that is adjusted for inflation. Nominal GDP is calculated with inflation as all goods and services are calculated at current price levels. Actual GDP measures country's economy at the current

moment in time and finally Potential GDP is a calculation of a country's economy under perfect conditions, taking into account steady currency, low inflation, and full employment. On the plot below it's clear that the number of items or services a citizen can buy grows for economy of all actors presented and this is shown through the GDP per capita Purchasing power parity (GDP per capita PPP). This means that standard of living and productivity grows worldwide, not just in USA as it could be interpreted on the previous line chart. Comparisons using GDP PPP are therefore usually more useful than those using nominal GDP when assessing a nation's domestic economic status.

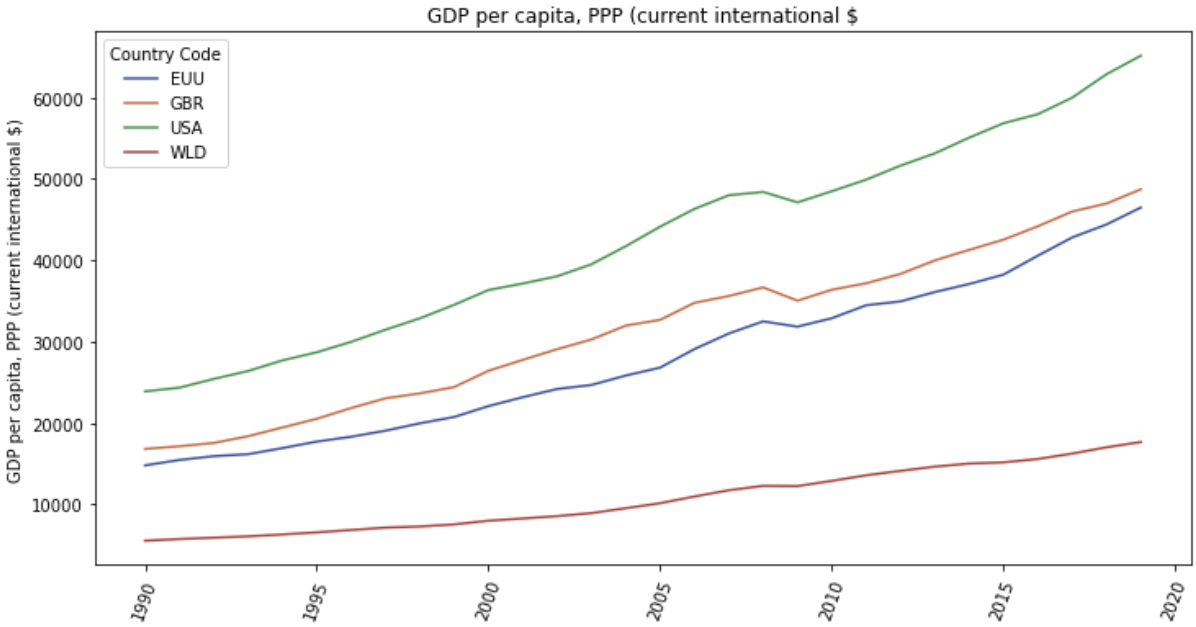

Figure 2. GDP per capita PPP of selected countries

2.2 GDP and economic growth

GDP is the market value of all finished goods and services, produced within a country in a year. As productivity raises over the years, there is a steady growth in GDP, and this almost never goes down throughout the human history. On the other hand, debt swings are appearing in cycles and they are influencing fluctuation in GDP growth. This is due to the human nature and due to how credit works. This results in short term GDP and spending boom and after that short-term GDP and spending drop. Credit can be considered a bad factor if it enforces overconsumption that cannot be paid back. This is called Short term debt cycle. Over long period of time the GDP goes up after each individual debt cycle but up goes also debt, which grows faster than GDP and is caused by human nature of spending more than a person can afford. This causes people not being able to pay their debts and stock market collapses. Then comes leveraging of the debt, deflation and depression. If deleveraging done correctly, it can achieve a slow debt reduction. This development can be noticed even in 2020s USA during the coronavirus and economic crisis. In February 2020 there was a biggest fall of stock prices since 2008 and global economic growth forecast fall. SARS outbreak in 2003 killed 774 people and cost the global economy an estimated US$50 billion. In case of COVID-19 though, ANU researchers estimate a global GDP loss of $2.4 trillion and death toll of 15 million in the best-case scenario (Business Insider, 2020). There was a sharp fall of all of manufacturing activity in china influencing European markets as well. This caused global

economic activities to slow down due to restrictions and supply chain disruptions. This resulted in market anomalies taking place due to the panic. (Chu, 2020). Government restrictions will affect all sectors and almost all businesses will be affected by the supply ability within micro-economic sector (Welmans, 2020).

2.3 Calculation of GDP

Country's GDP can be calculated based on expenditures, income or production. Expenditures consist of everything that is purchased within the country plus net exports to other countries. Net exports can be calculated as nation's total export goods and services minus the value of all the goods and services it imports. A country with positive net exports enjoys a trade surplus, while having negative net exports equals a trade deficit (Investopidia, 2020). This formula states that GDP = consumption (C) + investment (I) + government spending (G) + (exports (X) - imports (M)). It is worth noting that according to the formula, consumption covers the largest part of this calculation while excluding purchasing property. Investments consist of purchasing new equipment or materials, but not purchasing of stocks, which is considered to be filed under savings. Government spending then factors everything that government paid for excluding welfare or social security. The income of all the individuals and businesses within the country or domestic income is also used to calculate GDP. Furthermore, we can calculate GDP based on production, which is the market value of everything that is produced within the country.

2.4 GDP prediction

According to Marcellino, predicting the future evolution of GDP growth and inflation is the main concern in economics. Forecasts are usually produced either from economic theory-based models or from simple linear time series models. A time series model can provide a reasonable benchmark to evaluate the value added of economic theory relative to the explanatory power of the historical behaviour of the variable. Latest progress in time series analysis catalysed the fact that more sophisticated time series models could produce better benchmarks for economic models. Research then revealed that linear time series models are the most efficient if they are carefully specified. But a more complicated benchmark can alter the conclusions of economic analyses about the driving forces of GDP growth and inflation (Marcellino, 2007). Simple linear regression can roughly predict the growth in European countries, with specifically Germany and United Kingdom expecting sharper growth than countries of the former Soviet Comunnist Block. This is displayed by the regression lines useful for prediction. Overview of EU, its historic development and regional separation of EU will be discussed in the next section. Autoregressive model ARIMA was selected and will be further explained in Time Series section.

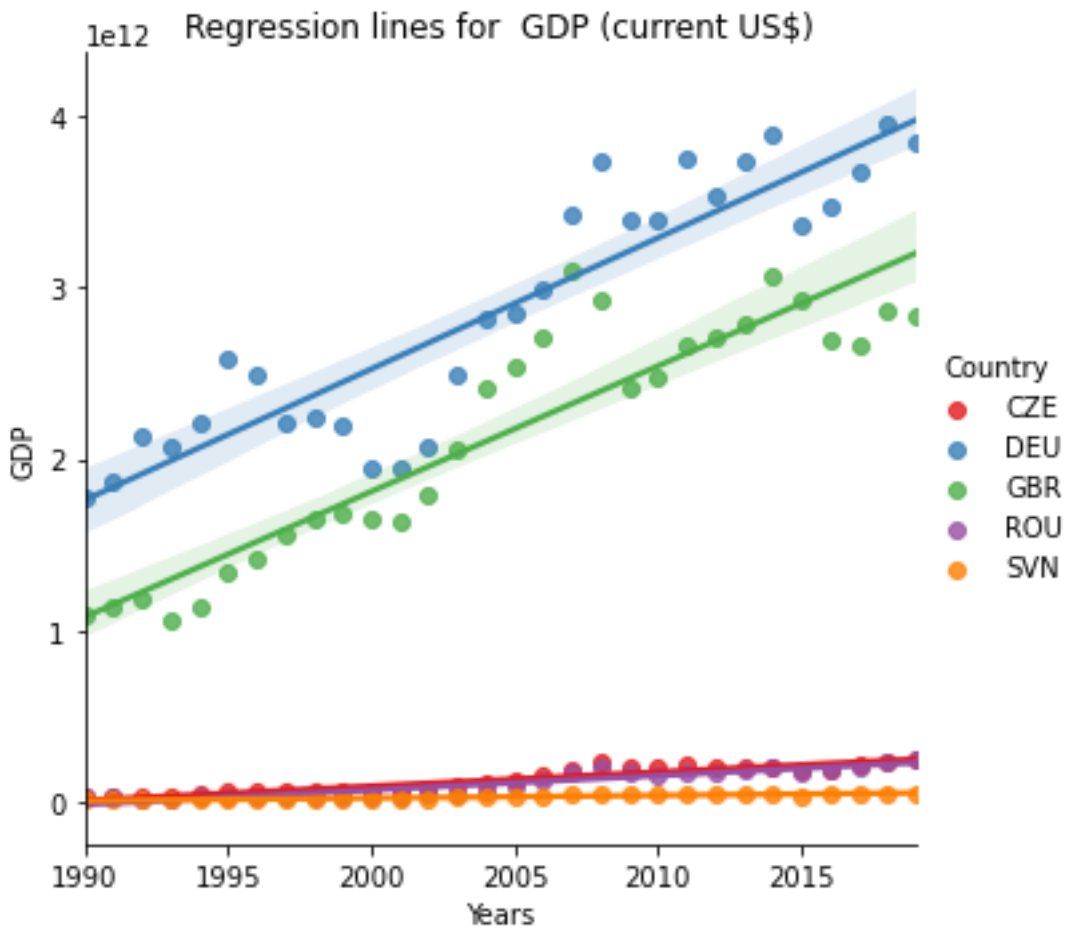

Figure 3. Regression lines for specific countries in EU

2.5 Historical development of European Union

Containing some 5.8% of the world population in 2020, the EU (excluding the United Kingdom) had generated a nominal gross domestic product of approximately US$15.5 trillion in 2019, which constitutes around 18% of global nominal GDP. Additionally, all EU countries have a very high Human Development Index according to the United Nations Development Programme. In 2012, the EU was awarded the Nobel Peace Prize and it maintains permanent diplomatic missions throughout the world and represents itself at the United Nations, the World Trade Organization, the G7 and the G20. Due to its global influence, the European Union has been described by some scholars as an emerging superpower. GDP will be analysed backwards and predicted, for countries currently in the European Union including UK and therefore there is an analysis before 1993 as well, which is year that EU was founded.

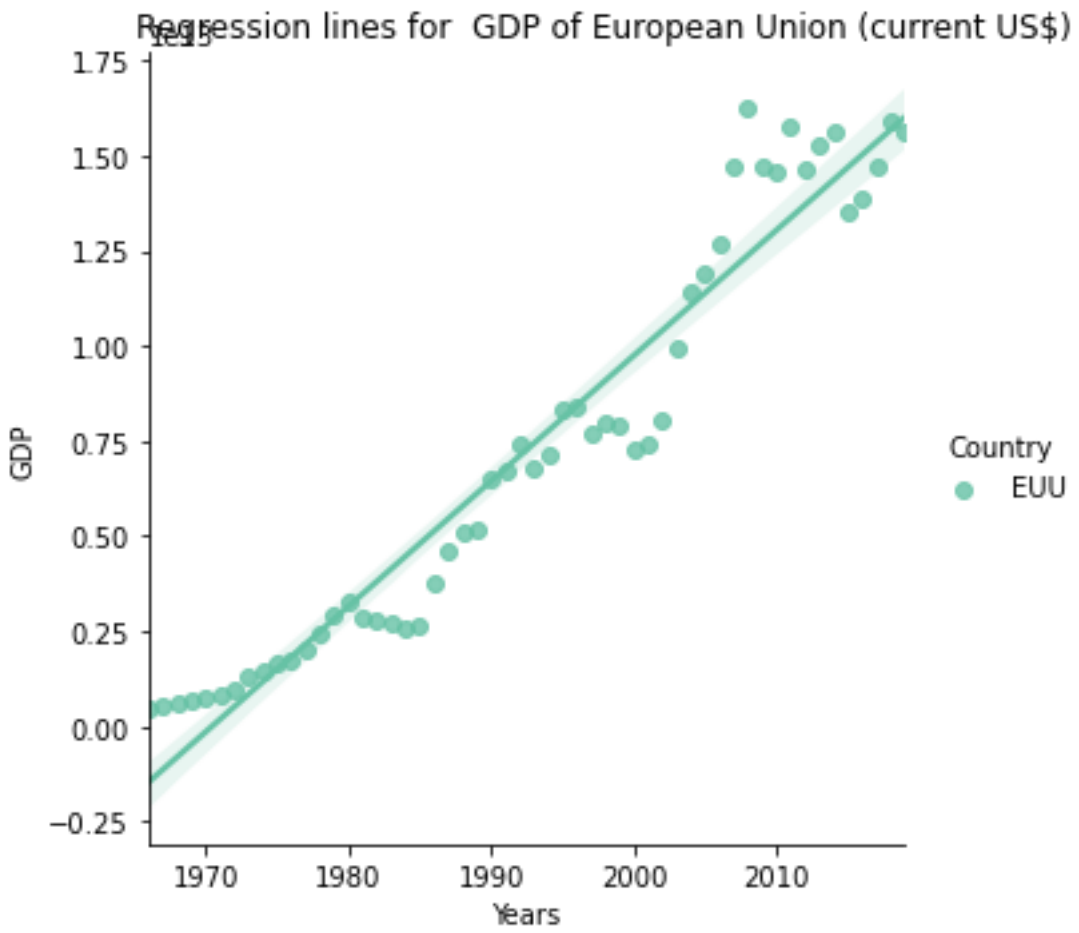

Figure 4. Regression lines for EU

2.6 GDP Historical growth in European Union

After the plateau in early 1960s, Europe's economy grew rapidly for many years from 1965, but now Europe is faced with problems as many large economies within Europe have failed to generate enough jobs or achieve high productivity. That is why according to Baily and Kirkegaard (2004) Europe's growth slowed. Furthermore, it is noted that Information technology makes high contribution to the growth and should be prioritised, while system with good work incentives and a high level of competitiveness will help to pursue economic transformation. The annual unemployment rate in the European Economic Community rose sharply in 1981 and continued rising to its 1984-85 plateau. The development of unemployment had clearly impact on GDP as it plateaued as well in this time period, which is seen on the figure below. This rise of unemployment in Europe has had grave consequences for individuals and countries as well. This slump presented difficult problems for economic analysis back in 1980s (Fitoussi,1980). European union experienced the sharpest GDP growth between years 2002 and 2008.

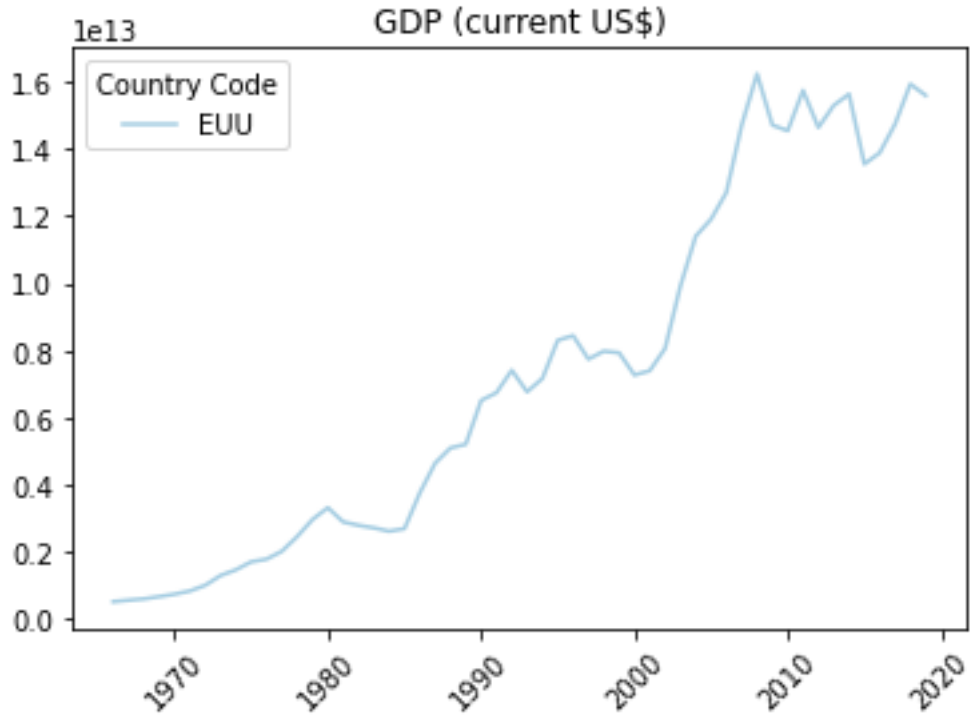

Figure 5. GDP of EU

It is worth noting that economic crises like the one in 2008 and pandemics like COVID-19 are harming the globalization efforts as there are barriers to trade. Similarly, immigration policies aimed at protecting citizens before workforce from foreign countries may cause worsening of economic state of that particular country due to lack of workers for given sectors (Ilter, 2017). On the figure below, a sharp dip in 2008 symbolises the economic crisis.

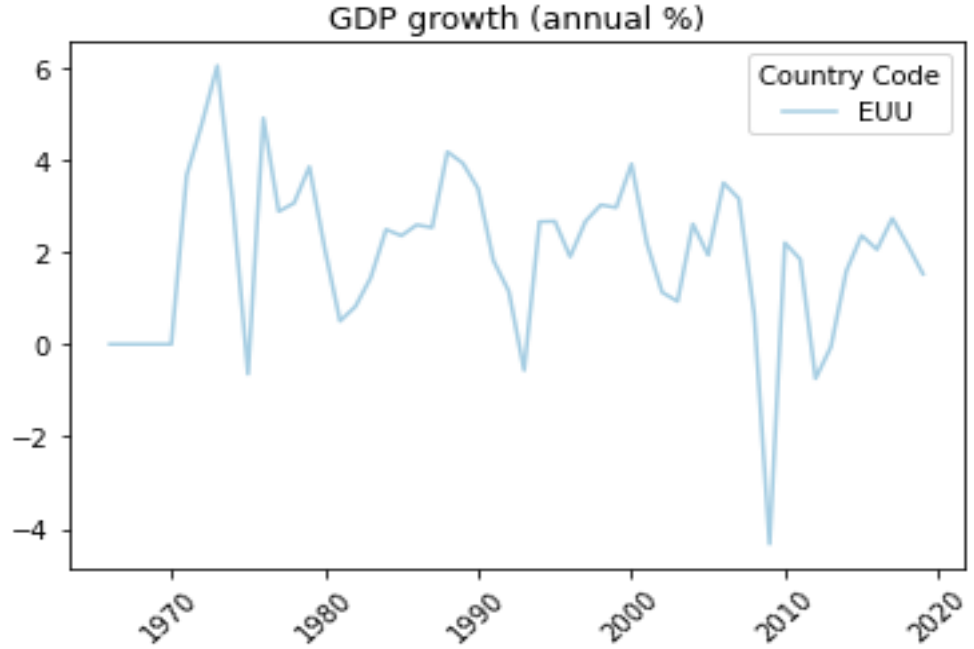

Figure 6. GDP growth of EU

2.7 Economic divide between western and eastern part of Europe

Economic dispersion is visible among countries all around the world and that is a fact even though globalization impacted and supposedly helped third world countries. The term globalization means that the size of the world is getting smaller with trade barriers, which are being lowered, and technology advances are more globally accessible. GDP states for a Gross Domestic Product and this figure shows the total of all goods and services produced in an economy during a period of one year. Furthermore GDP per capita is this sum divided by the population of the country and demonstrates a development in specific country. Developed countries in Europe have average GDP per capita between USD 30- 50,000. In terms of Europe there is not a huge divide, although western European countries have higher GDP and take a lead in factors such as better standard of living when compared to eastern Europe. (Ilter, 2017). There is a great divide between large countries in the west with high development and productivity and countries located eastern of Germany who are lagging behind in industry and are of smaller size. This pattern is portrayed by GDP plot below, where France and Great Britain stand in GDP size behind Germany, whereas eastern European countries have many times lower GDP.

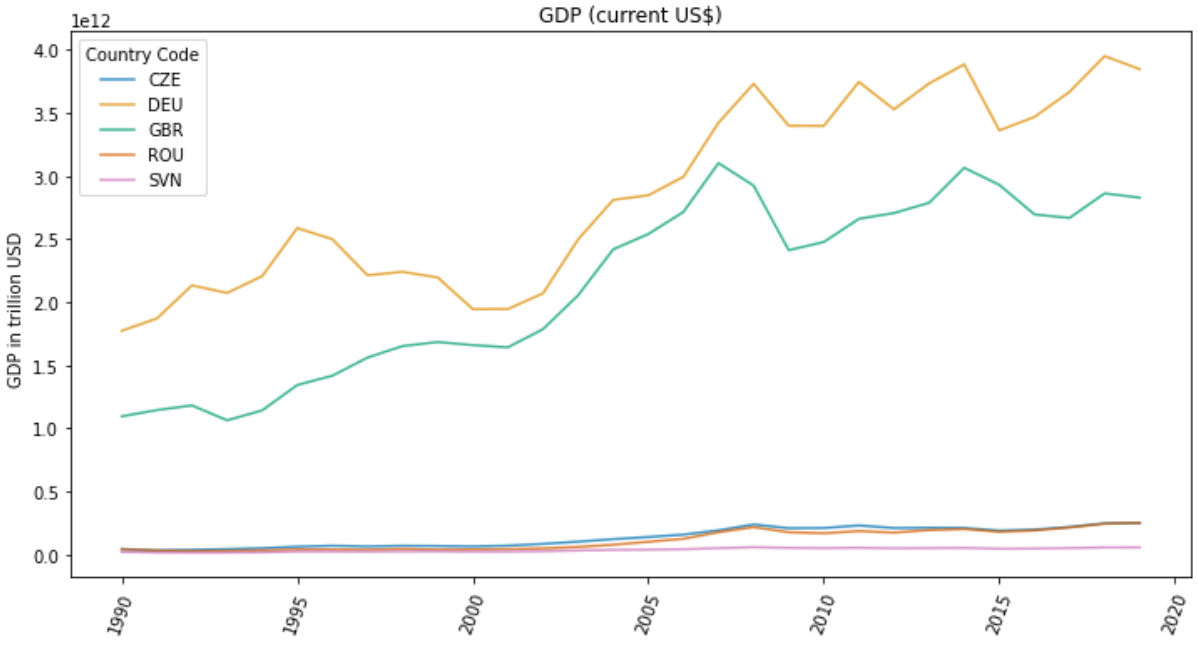

Figure 7. GDP of selected EU countries

Within European Union, Poland is the only one that did not experience a single year of recession during the significant period 2007–11. On the other hand Paul de Grauwe (2012) points out that several countries like France, Italy, Portugal, Romania, Spain or the United Kingdom experienced and suffered a double dip, with two years of recession as can be seen on figure below. Furthermore Greece, Ireland and Latvia experienced “multiple dips”, while the rest of the EU countries had at least one year of GDP contraction. These crises cause chain effects to erupt and one country will have influence on the others. It is obvious that countries within the same monetary union will catalyse changes on each other, but it is worth noting that during COVID-19, China strongly influenced economic situation all around the

world (Tridico, 2013). This similar pattern in GDP per capita growth can be seen on the figure below.

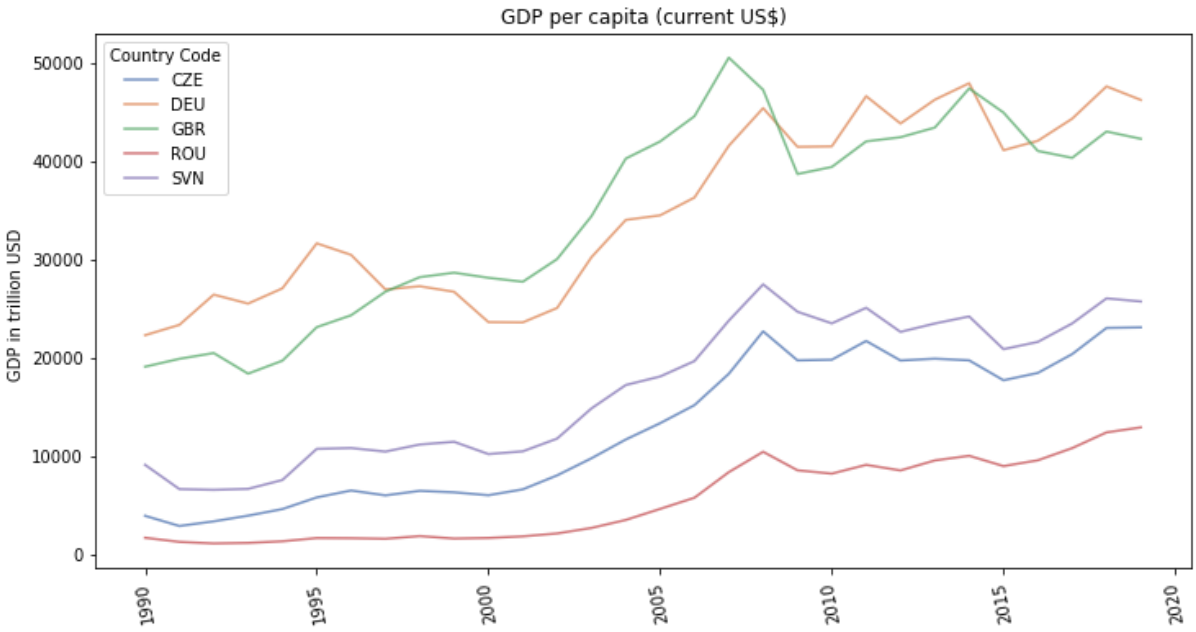

Figure 8. GDP per capita comparisson of selected countries

3 Time series

3.1 Overview of time series

A time plot displays values against time and can only display time on the x-axis. Timeplots are good for showing how data changes over time. The goal of time series analysis is to find patterns in the data and use the data for predictions. For example, if the data is affected by past data, one way to model this behaviour is through the ARIMA process. Time series patterns can be difficult to analyse because of noise. Analysis is based on the assumption that the data has equal intervals, for instance a yearly interval in case of GDP from World Bank database. In order to recognize trends, smoothing is used to create a line graph, also called a best fit line. Smoothing is also useful for spotting outliers. Seasonality refers to fluctuations in time series data that happens at regularly. While this traditionally means behaviour of seasons like Spring or Summer, this cyclical behaviour can happen yearly or every 10 years. Seasonality can cause issues with interpreting time series data and therefore must be included and analysed (Chatfield,1995).

3.2 Time series auto-regression methods

3.2.1 The AR model

An AR model is a representation of a type of a random process and has generally been applied to non-regular time events, such as stock market forecasting and macroeconomic forecasting including GDP. The AR model specifies that the output variable depends linearly on its previous values. In different words it's a regression of the variable against itself and this model contains a stochastic difference equation. Contrary to AR, in a multiple regression model, we forecast the focused variable by using a linear combination of predictors. To estimate AR model, we need to estimate the optimal number of lags. Not enough lags can cause omitting crucial information, and this will cause the residuals to become autocorrelated. Furthermore, everything that is not included as an independent regressor will be in the residual. (Hilde and Thorsrud, 2014).

3.2.2 The ARMA model

ARMA model is simply the combination of AR(p) and MA(q) models: AR(p) models try to explain the momentum and mean reversion effects often observed in trading markets. MA(q) models try to capture the shock effects observed in the white noise terms.

3.2.3 The ARIMA model

ARIMA is The Autoregressive Integrated Moving Average method, which is most frequently used in academic papers for GDP forecasting. This method models the next step in the sequence as a linear function of the differenced observations and residual errors at prior time steps. It is a combination of Autoregression (AR), and Moving Average (MA) models and it makes the sequence stationary. ARMA model is equivalent to an ARIMA model of the same MA and AR orders with no differencing. ARIMA contains extra I in its name and this stands for 'Integrated' and refers to how many times are needed to difference a series in order to achieve stationarity (Brownlee, 2020

3.2.4 Other methods for time series prediction

Machine learning algorithms often require a lot of data in order to recognize subtle patterns and trends, and this factor creates challenges when it comes to the gathering of domestic macroeconomic indicators. The range of these indicators has to be wide and includes Indicators GDP, inflation and unemployment rate. Problem is that other valuable indicators are not usually recorded or there are missing data in less developed countries. Other macroeconomic indicators useful for GDP growth prediction are for example employment rate, personal income and new job vacancies. Using multiple variables can be advantageous but makes the model narrower and more applicable to few limited scenarios.

Furthermore, for more complex multivariate time series, time-dependent variables are continuously added or removed to maximise the modelling accuracy. Basic methods like exponential smoothing, trend projection, mean model, naive forecast, random walk or drift method are often used as well. These methods are Long short-term memory neural networks or agent-based and system dynamic modelling. I can explore forecast function in Excel, Python or econometric software. There is an option of including G-Cubed and Globe model for Coronavirus impacts. Various verification techniques will be used as well.

3.3 Data

World Development Indicators (WDI) is the primary World Bank collection of development indicators, compiled from officially recognized international sources. It presents the most current and accurate global development data available, and includes national, regional and global estimates. Even though Global Development Finance (GDF) is no longer listed in the WDI database name, all external debt and financial flows data continue to be included in WDI. The GDF publication has been renamed International Debt Statistics (IDS), and has its own separate database, as well. This dataset is classified as Public under the Access to Information Classification Policy. Advantages of this dataset are good usability as it contains essential metadata, and it is in a clear and readable format. There is also an assurance that the dataset is well maintained. For the time series mostly GDP per capita in US dollars from year 1990 to 2017 will be utilised as important indicator of quality of life in EU. Starting year was selected to be 1990, because we have most data from this year onwards and ending year was 2017, which enables us to confirm prediction for next 2 years.

4 Research Design

Quantitative research, which is based on scientific method, was chosen for this dissertation. It's goal is to be as objective as possible, and is usually based on statistics or other measurable, empirical data. Conclusions are drawn from the analysis of things clearly measured (Mugenda, Mugenda,2003). From the beginning to the end of the assignment, predictive ability of ARIMA on GDP in European Union is explored. This central idea will be asserted, and the logical development of the argument will be pursued throughout every paragraph by providing another additional evidence or evaluation of the results. (University of Southampton, 2016) This thesis will combine economy and computer science fields as I will follow previous theses and work will be utilising both areas. Previous academic work, information and methods will be applied towards European countries. My approach will be quantitative as historical data from time series and correlation information will be taken into account. Qualitative approach would not make sense as it would be complicated to take opinions from various experts and include them in the model. Single time series means that a time series consists of single observations recorded sequentially over equal time periods. For univariate GDP forecasting in this thesis, there will be single time series of GDP alone as various social and macro-economic factors will not be taken into account, although they might be used for explanation in data analysis section. According to study conducted by Cenap Ilter (2017) these additional factors could be mostly population, transparency score and compulsory education. Amongst economic factors it is furthermore inflation, unemployment, consumer spending consumption and net exports. Methods and theories from background research will be applied to accomplish project aims while using collected data, analytics, modelling and visualisation techniques. My goal is to enable others to replicate my data and make my data and methods public by creating a prototype with methods and data for reproducibility on GitHub in repository. Verification and reflexion will be integral at any point of time during my dissertation. As for a main data science project management methodology, a classic linear methodology was chosen. This strategy consists of sequential phases and does not include any feedback loops that would massively change the project outline. The outcome of the project is not revealed until the final phase is reached. Further characteristic is a clearly defined goal, outcome and requirements. Duplication of previous work is mostly a routine and repetitive process as pre-established processes and methods are used. (Kokatjuhha, 2018).

5 Method and Implementation

5.1 Box-Jenkins methodology and ARIMA

The Box-Jenkins methodology will be used, and an Autoregressive Integrated Moving Average model will be utilised to predict GDP per capita of European countries. This methodology consists of Time series identification step where we have to find the appropriate values for p, d and q. The next step is Model estimation. After that the Model diagnostic takes place, meaning that we have to check that the model residuals follow a white noise distribution. Last step is Forecasting, which means that we can project statistical predictive inference. ARIMA identifying rules are useful for finding the order of differencing and the constant and therefore stationary test will be applied. Additionally, there is a possibility of identifying the numbers of AR and MA terms or the seasonal part of the model. (Robert Nau,2020) Furthermore an Akaike Information Criterion (AIC) criterion will be present, which is an estimator of out-of-sample prediction error and thereby relative quality of statistical models for a given set of data. Using this optimized ARIMA model, we'll generate a projection for the US GDP per capita in PPP from 2018 to 2032. Finally, after building optimized ARIMA model to project the world GDP per capita PPP.

5.2 Data Analysis

Economic development indicators were chosen for previous data analysis to demonstrate economic situation in the region and give us an idea of how the data look like. These indicators were GDP per capita (US$), GDP per capita growth (annual %), GDP growth (annual %) and GDP (current US$). Data cleaning, manipulation and data transformation was done with use of Pandas - easy-to-use data structures and data analysis tools for python. Additionally, visualizations were prepared with matplotlib. Matplotlib is the oldest Python plotting library, and it's still the most popular. It was created in 2003 as part of the SciPy Stack, which is an open source library. Regression plots and bar plots were created with seaborn library, which is an abstraction layer on top of Matplotlib and provides useful interface. Following visualisations for ARIMA prediction were produced with Plotly to enable more interactivity while working in Jupyter Notebook.

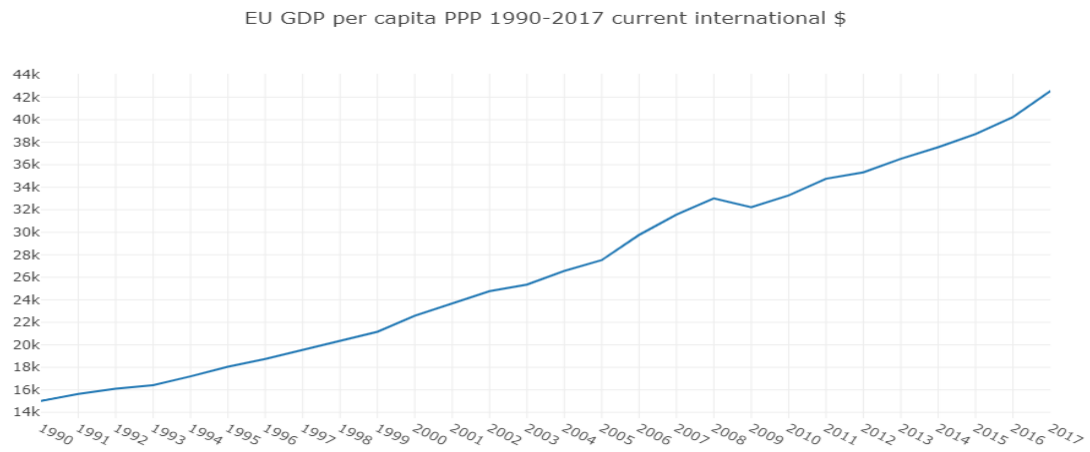

Figure 9. GDP of EU per capita PPP

6 Prediction using ARIMA

6.1 Overview of prediction

For autoregressive models like ARIMA, the rule of thumb is that there should be at least 50 but preferably more than 100 observations. Having this amount of observations is not possible due to the World Bank having GDP data from year 1971 for the EU and for other countries, mostly in the eastern europe, the data starts from year 1990 as discussed before. These eastern countries like The Czech Republic have much lower contribution of nominal GDP in comparison to bigger countries in the west, like UK, Germany or even France. With more countries planning to join the EU and Britain leaving, these models can partially account for that, if it is needed to predict purely EU's GDP. It should be noted that the predictions account for EU and Great Britain together into the future and the results should be interpreted as such. Even though there are only 27 observations from year 1990 to 2017, the following results and graphs show good predictive capability of the ARIMA model. Model predicts very accurately and in the future when more data are added, the predictive ability will be even higher.

6.2 Decomposition and Analysis

Time series models are different from other types of predictive models in that the target variable is both the object of the prediction for values in the future and an input feature of the model. The first steps in approaching a time series is to visualize and then decompose the data into trend and cyclical components. From there work on models like ARIMA will start although merely decomposing the series usually yields some important insights. All the following techniques make sense, because there is a clear degree of auto-correlation in the target variable. GDP is visibly correlated with itself from earlier periods as can be seen on autocorrelation plot below, even though certain outliers are present, which could demonstrate additional trend. This is because the GDP typically doesn't move down or up drastically between any given year.

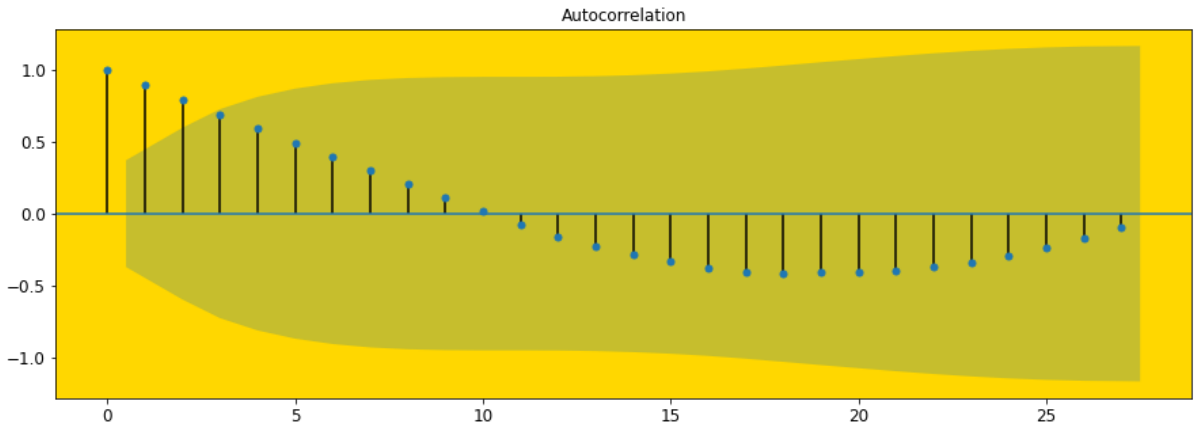

Figure 10. Autocorelation

We can say on 'Rolling Mean and Standard Deviation' line plot, that there is an overall positive trend. There is almost no changed variability in the trend. Yearly GDP doesn't show seasonality and this was also obvious, because the data is in yearly intervals. Patterns in the residuals also won't be considered. Decomposition must take place, because there is non-

stationarity in time series and it must be made stationary. There is a linear trend, which is constant over time and that is why additive model will be used for decomposition. Because the data is yearly, period will be 1. From the growth dynamic of the GDP time series, we can identify some cyclic patterns and certain sharp drops like in 2009.

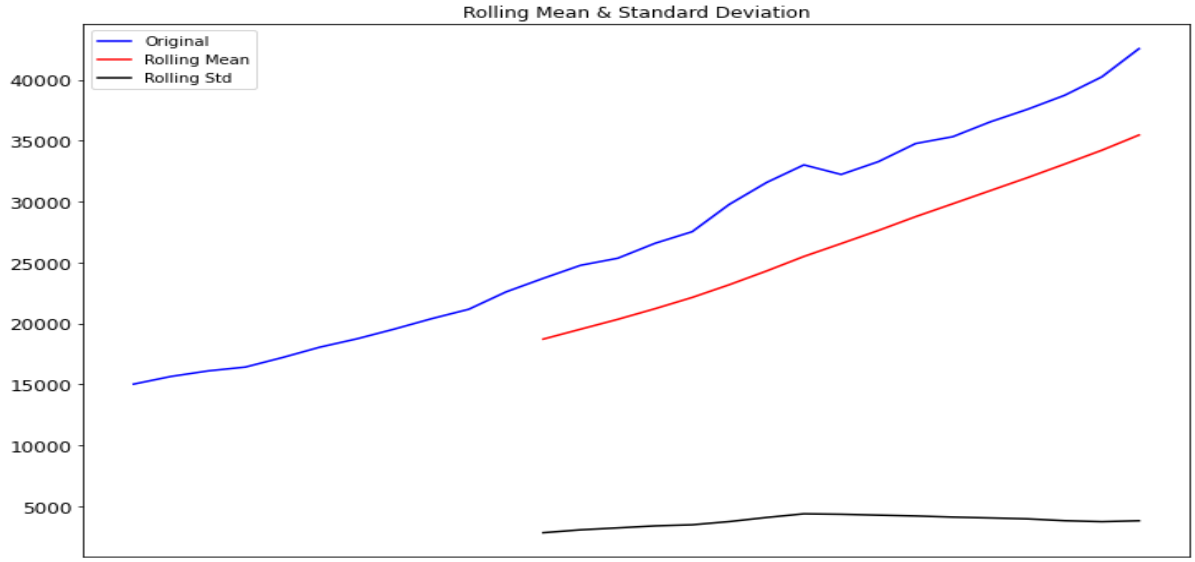

Figure 11. Rolling Mean and Standard Deviation

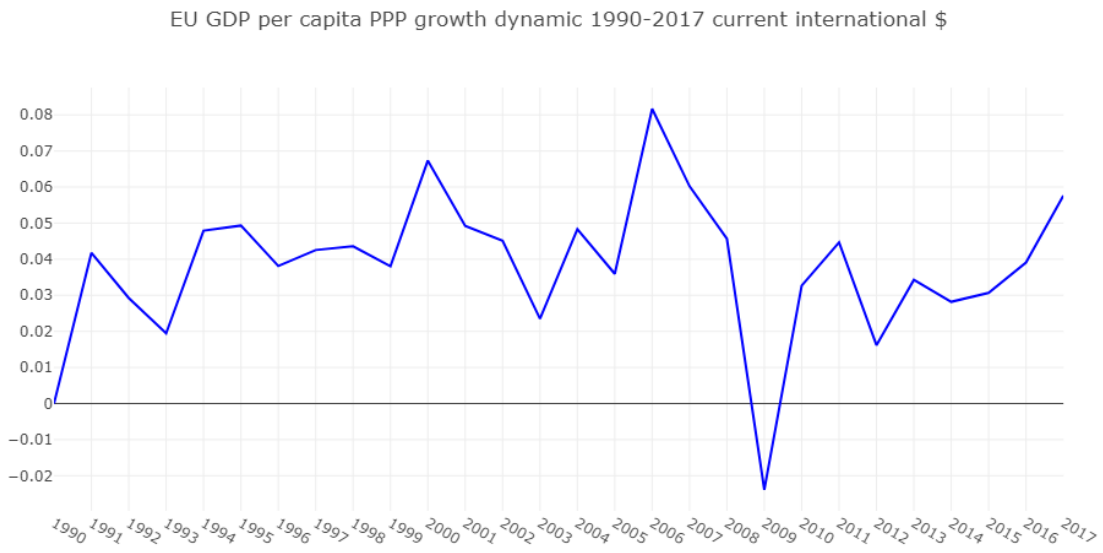

Figure 12. Growth dynamic of the GDP

6.3 Making the time series stationary

Already after first differencing, time series becomes stationary as p value (0.005834) is very close to 0.05. No more orders of differencing are needed as illustrated on the Rolling Mean and Standard Deviation figure below.

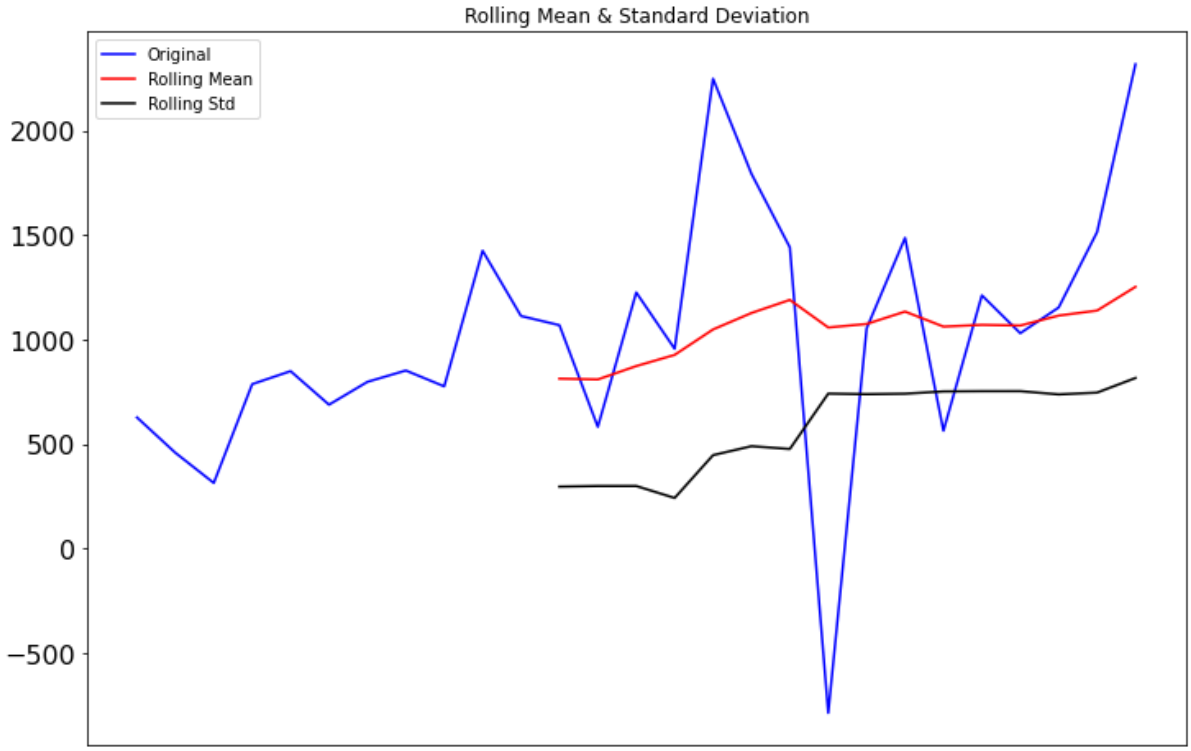

Figure 13. Stationary time series demonstration

Prediction using ARIMA

For ARIMA model, we need to find the best parameters by continuously generating various models and fitting data. Akaike Information Criterion (AIC) is developed for each of these models the aim is to choose model that generalises well with the lowest AIC, so that it can be widely used for developed countries like EU, North America or other developed economies all around the world. Histogram together with correlogram shows that the time series is normally distributed, because the time series residuals show low correlations with lagged residuals. Normal Q-Q plot also shows that this model performs well by presenting close fit of the line.

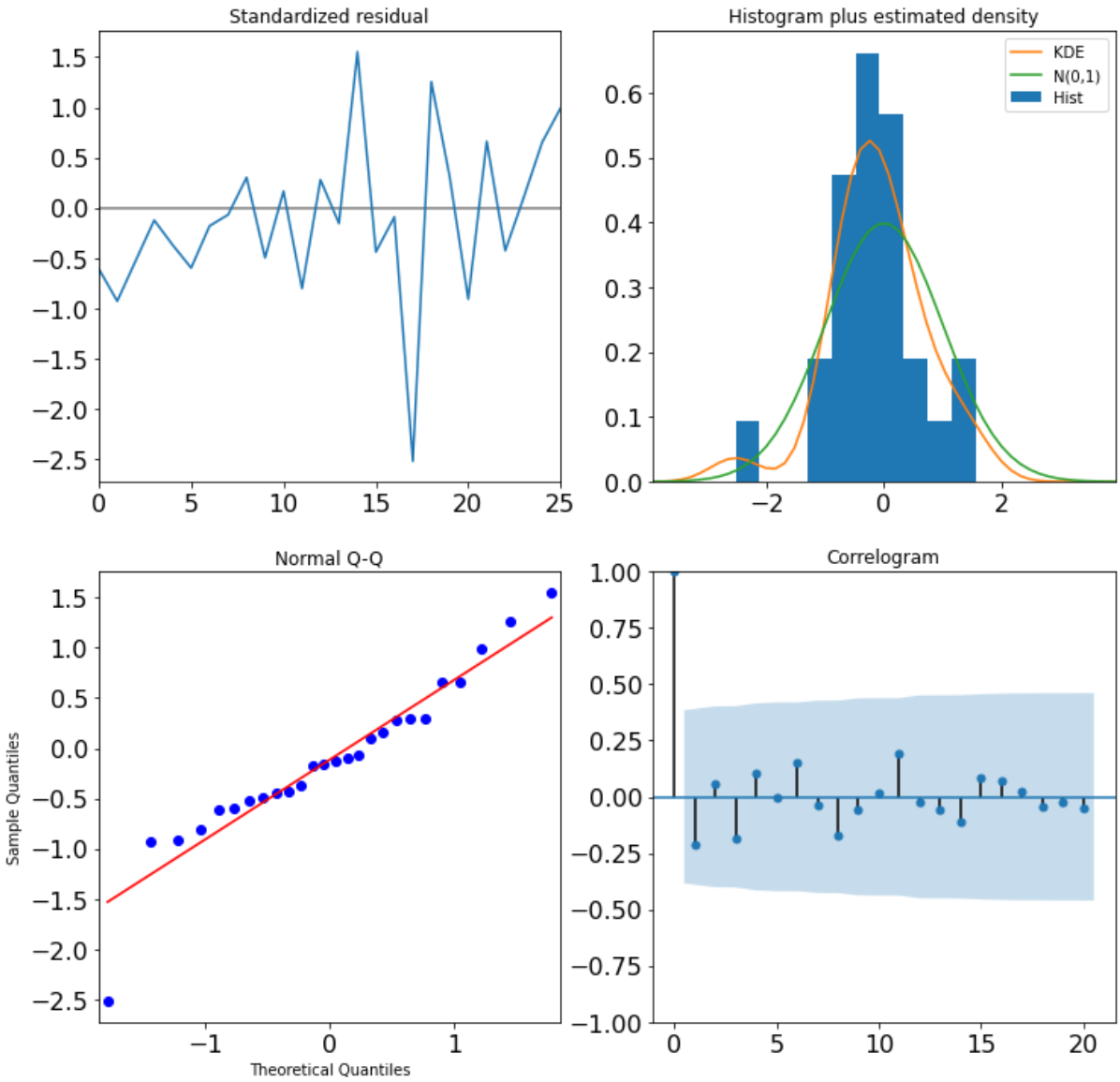

Figure 14. Demonstration of the accuracy of model

We see on the following figure that our in sample performance is very close to real figures until 2017 and we can be sure with prediction that the GDP per capita in European Union including Great Britain will grow from the current 41.000 US dollars in 2019 to 60.000 US dollars in 2030. The sudden changes in national investment strategy, stability of the region and other external and internal factors may cause drastic changes in terms of GDP per capita as it happened during pandemics and economic crisis before.

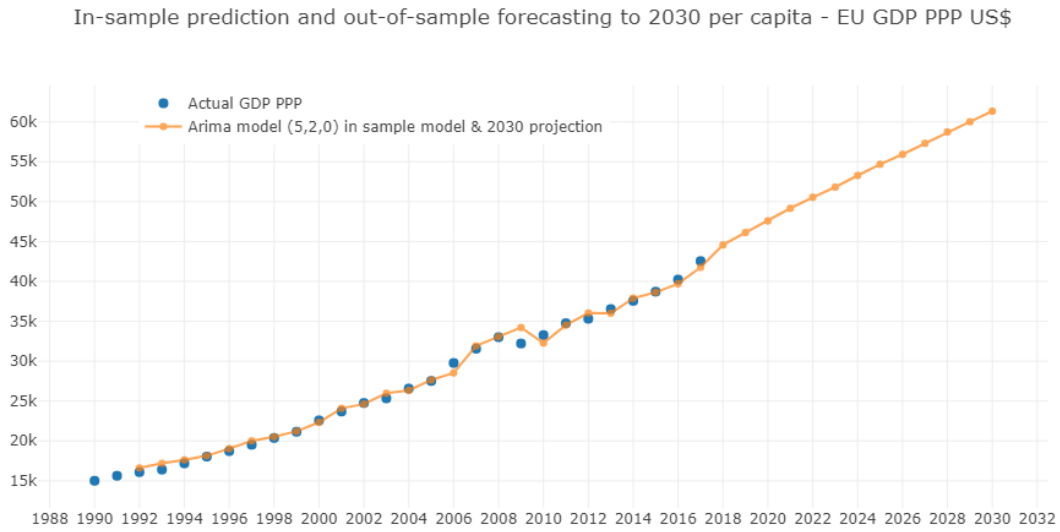

Figure 15. GDP of EU per capita PPP in US dollars

2018	44577.601153
2019	46130.829325
2020	47590.134605
2021	49170.544889
2022	50550.213495
2023	51824.715663
2024	53281.354634
2025	54682.058628
2026	55928.873936
2027	57310.944230
2028	58742.560089
2029	60015.120062
2030	61339.502003
2032	62776.102960

Figure 16. Prediction of GDP for 2018-2032 using ARIMA

7 Transparency and replication

As mentioned in the beginning, replication posed a challenge due to the age, state and accessibility of data and lack of unification of software and methods used. Jahnke and Asher (2012) think that there won't be a clear out-of-the-box solution that can be applied to the problem of data curation. The best way according to them would be to put emphasis on engagement with researchers. Discussion must be promoted in order to identify and construct the appropriate tools for a specific project, which will then result in better data curation. It is crucial to have access to appropriate networked storage and for instance multi-institutional research projects could be started by Universities and then further enhanced through excellent access to all the data and unified storage of all available material. Only by sharing research data and the results of research can new knowledge be transformed into socially beneficial goods and services.

In 2009 National Academy of Sciences (US), National Academy of Engineering (US) and Institute of Medicine (US) published a book called 'Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age'. In this publication it's argued that when research information is readily accessible, researchers and other innovators can use that information to create products and services in order to pursue improvements and extension of human capabilities. This concept is called 'Open data' and it's gaining high amount of traction in the recent years. It is defined as collection of data that is easily accessible, usable and shareable. The argument doesn't cover sensitive or personal data, which should be secured, but instead data, which could help society and humankind. As we live in a globalized world where there is tremendous potential of collaboration, locking up relevant data or putting this data behind expensive paywall can slow this progress down.

According to the research of Jahnke and Asher (2012), numerous researchers expressed concerns and uncertainty with the ethical use of data used for their research. Scientists and other people working with data can easily forget to backup data and collaboration solution is of great need in order for them to work with their team. Therefore, it is necessary that effort is put into establish the best practices in this research area, particularly for qualitative data sets. In order to produce the most efficient researchers, educational programs should focus on early intervention in career journeys, while teaching these researchers framework with clear boundaries for the greatest benefit. Berger et.all (2019) also reiterate the need for automated and reproducible studies as it is only through careful re-validation of studies that the broader community may gain trust in results of such research and also gain a better insight into the corelated problems.

Data degradation and data loss can be avoided by opening access and enabling researchers to make copies. If scientists are using data, then it remains relevant and new breakthroughs can result in profit for original data holder. Transparency is important for data integrity and easily shareable data result in further replicability and reproducibility. Peers can then easily suggest improvements and published work will have strong backing and it will help readers understand how the results were reached. In order to motivate research workers, it would help to publicly fund research and the risk of collusion decreases. Displayed on picture below is overview of key factors behind The European Commission's (EC) Horizon 2020 Programme, which is an EU initiative to spread the concept of open data. If the research is funded by programs like FP7, Horizon 2020, or FET, the data must be open. This will then moreover

shift the mentality of workers in academic and industry research areas. LabFolder (2020). To ensure that data is openly accessible it is important to have common policy and one central repository. In this paragraph it was shown that fear of misusing copyrighted content shows to be impediment to effectivity and progress of researchers.

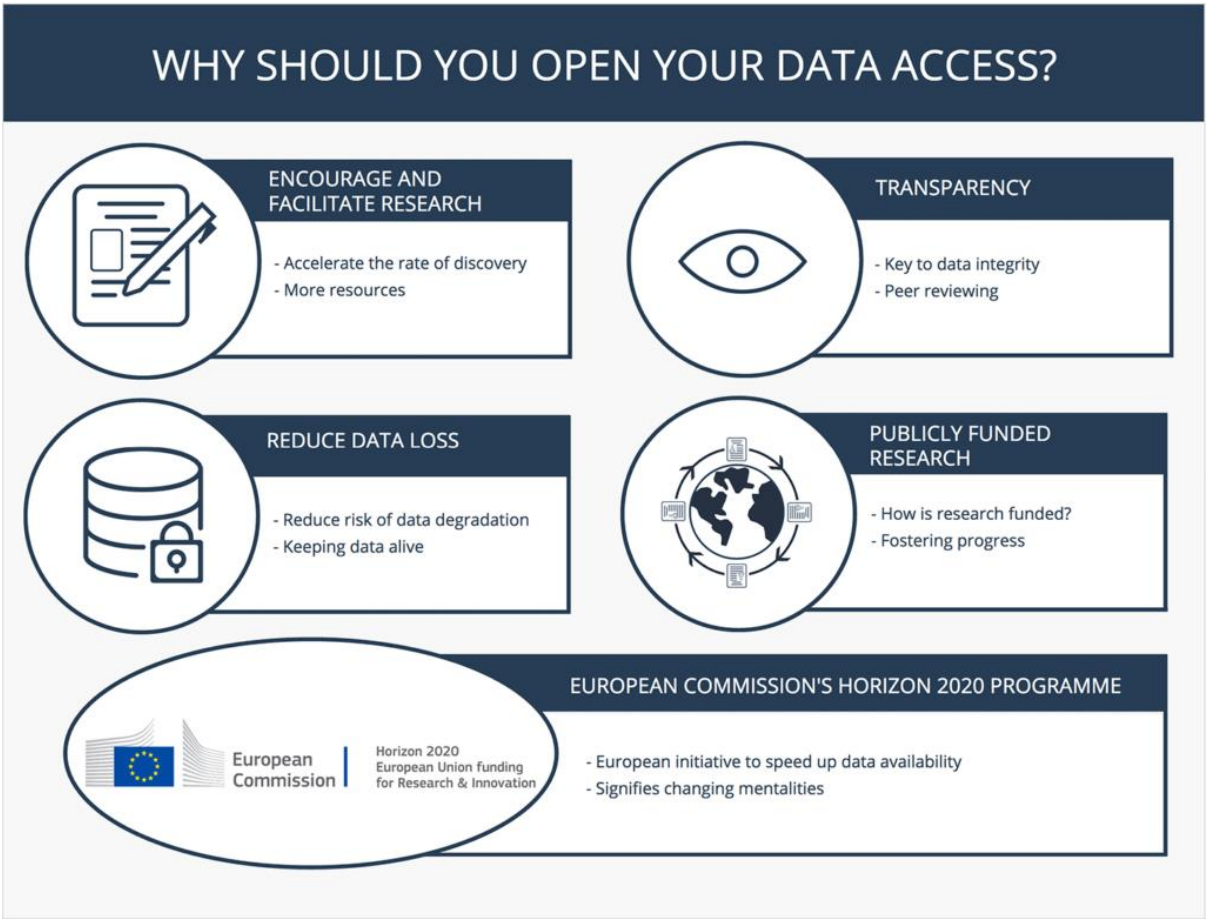

Figure 17. Goals of open data. Retrieved from https://www.labfolder.com/why-we-need-open-data-access/

8 Future work and Conclusion

Given the importance of GDP data in modern society, from becoming an election narrative to influencing the global commodity market, it is imperative that ample research should be done for GDP modelling for any country particularly for developed countries during economic crisis such as COVID-19. Gross domestic product is one of most important macroeconomic indicators in an economy. This paper attempted to analyze and explain historical growth of various types of GDP and explore ARIMA time series method in interactive programming in Python. Subsequently light was shed on data generation and cleaning process. Development of GDP worldwide and particularly in EU has shown a linearly increasing trend in the long run and therefore it is necessary to decompose time series using additive decomposition. For shorter time periods, the seasonal trend was also noticeable. Autoregressive models are very useful and correct for predicting GDP which was shown during literature and previous theses analysis. Using publicly available GDP data can be difficult, due to the large amount of data manipulation needed. More work can be done in terms of thesis replication, which proved to be problem due to limited transparency and the data were missing or proficiency with econometric software were not developed. During thesis replication, many methods and datasets were outdated, or appropriate sources and links were needed to overcome a transparency barrage and struggle with replication.

Bibliography

Online.stat.2020. Decomposition Models | STAT 510’. n.d. PennState: Statistics Online Courses. Accessed 13 December 2020. https://online.stat.psu.edu/stat510/lesson/5/5.1.

Otexts.2020. ‘Autoregressive Models | Forecasting: Principles and Practice’. n.d. Accessed 27 September 2020. https://Otexts.com/fpp2/.

European-Union. 2016. ‘The EU in Brief’. Text. European Union. 16 June 2016. https://europa.eu/european-union/about-eu/eu-in-brief_en.

Antoniou, Manos. 2018. ‘Forecasting GDP in the Eurozone’. Medium. 26 August 2018. https://towardsdatascience.com/forecasting-gdp-in-the-eurozone-54778f79912e.

Brownlee, Jason. 2016. ‘How to Check If Time Series Data Is Stationary with Python’. Machine Learning Mastery (blog). 29 December 2016. https://machinelearningmastery.com/time-series-data-stationary-python/.

Dynan, Karen, and Louise Sheiner. 2020. ‘GDP as a Measure of Economic Well-Being’

Wikipedia. 2020. European_Union. https://en.wikipedia.org/w/index.php?title=European_Union&oldid=984982761.

Fitoussi, J.-P., E. S. Phelps, and Jeffrey Sachs. 1986. ‘Causes of the 1980s Slump in Europe’. Brookings Papers on Economic Activity 1986 (2): 487. https://doi.org/10.2307/2534480.

Iwok, I. A., and A. S. Okpe. 2016. ‘A Comparative Study between Univariate and Multivariate Linear Stationary Time Series Models’. American Journal of Mathematics and Statistics 6 (5): 203–12.

Klenk, Jochen, Ulrich Keil, Andrea Jaensch, Marcus C. Christiansen, and Gabriele Nagel. 2016. ‘Changes in Life Expectancy 1950–2010: Contributions from Age- and Disease-Specific Mortality in Selected Countries’. Population Health Metrics 14 (May). https://doi.org/10.1186/s12963-016-0089-x.

Majaski, Christina. 2020. ‘What Is the Net Exports Formula?’ Investopedia. Accessed 15 November 2020. https://www.investopedia.com/terms/n/netexports.asp.

Marcellino, Massimiliano. 2008. ‘A Linear Benchmark for Forecasting GDP Growth and Inflation?’ Journal of Forecasting 27 (4): 305–40. https://doi.org/10.1002/for.1059.

Media. 2020. ‘Media_359359_smxx.Pdf’. n.d. Accessed 13 December 2020. https://www.gla.ac.uk/media/Media_359359_smxx.pdf.

Miller, Max. 2020. ‘The Basics: Time Series and Seasonal Decomposition’. Medium. 19 March 2020. https://towardsdatascience.com/the-basics-time-series-and-seasonal-decomposition-b39fef4aa976.

‘Mugenda OM Mugenda AG 2003 Research Methods Quantitative and Qualitative | Course Hero’. n.d. Accessed 15 November 2020. https://www.coursehero.com/file/p784s2n/Mugenda-OM-Mugenda-AG-2003-Research-Methods-Quantitative-and-Qualitative/.

Murray, Christian J, and Charles R Nelson. 2000. ‘The Uncertain Trend in U.S. GDP’. Journal of Monetary Economics 46 (1): 79–95. https://doi.org/10.1016/S0304-3932(00)00018-0.

Ning, Wei, Bian Kuan-jiang, and Yuan Zhi-fa. 2010. ‘Analysis and Forecast of Shaanxi GDP Based on the ARIMA Model’. Asian Agricultural Research 02 (01): 1–4.

Explorable. 2020. ‘Research Population - The Focus Group of a Scientific Query’. n.d. Accessed 23 October 2020. https://explorable.com/research-population.

People.duke. 2020. ‘Rules for Identifying ARIMA Models’. n.d. Accessed 15 November 2020. https://people.duke.edu/~rnau/arimrule.htm.

Statisticshowto. 2013. ‘Timeplot / Time Series: Definition, Examples & Analysis’. Statistics How To. 24 September 2013. https://www.statisticshowto.com/timeplot/.

Statisticshowto. 2015. ‘Autoregressive Model: Definition & The AR Process’. Statistics How To. 19 August 2015. https://www.statisticshowto.com/autoregressive-model/.

CLIR. 2020. ‘The Problem of Data: Data Management and Curation Practices Among University Researchers • CLIR’. 2012. CLIR (blog). 2012. https://www.clir.org/pubs/reports/pub154/problem-of-data/.

The World Bank. 2018. ‘The World Bank GDP Analysis Using Pandas and Seaborn Python Libraries’. 2018. Ermlab Software (blog). 10 July 2018. https://ermlab.com/en/blog/data-science/pandas-seaborn-world-bank-gdp-analysis/.

Wabomba, Musundi Sammy, M’mukiira Peter Mutwiri, and Mungai Fredrick. 2016. ‘Modeling and Forecasting Kenyan GDP Using Autoregressive Integrated Moving Average (ARIMA) Models’. Science Journal of Applied Mathematics and Statistics 4 (2): 64. https://doi.org/10.11648/j.sjams.20160402.18.

Labfolder. 2017. ‘Why Do We Need Open Data Access?’. https://www.labfolder.com/why-we-need-open-data-access/.

Zhang, Haonan, and Niklas Rudholm. n.d. ‘Modeling and Forecasting Regional GDP in Sweden Using Autoregressive Models’, 43. Accessed 19 October 2020. https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0101EN-SkillsNetwork/labs/Module%202/Data_Mining.md.html.

Appendix

.

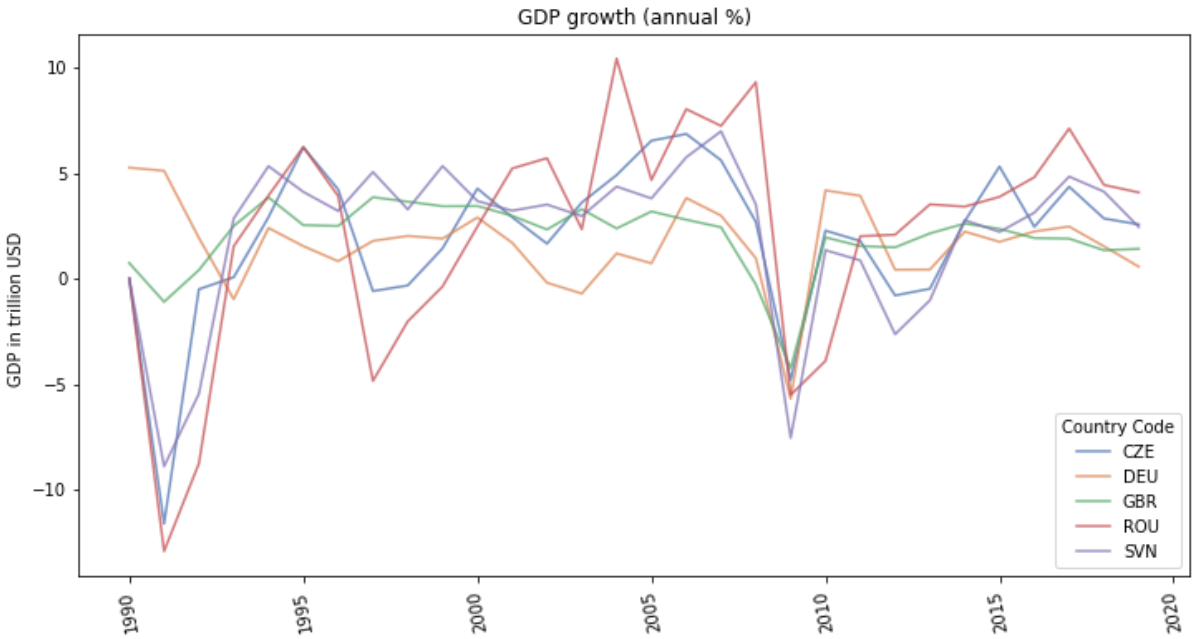

Figure 18. GDP growth of selected coutries

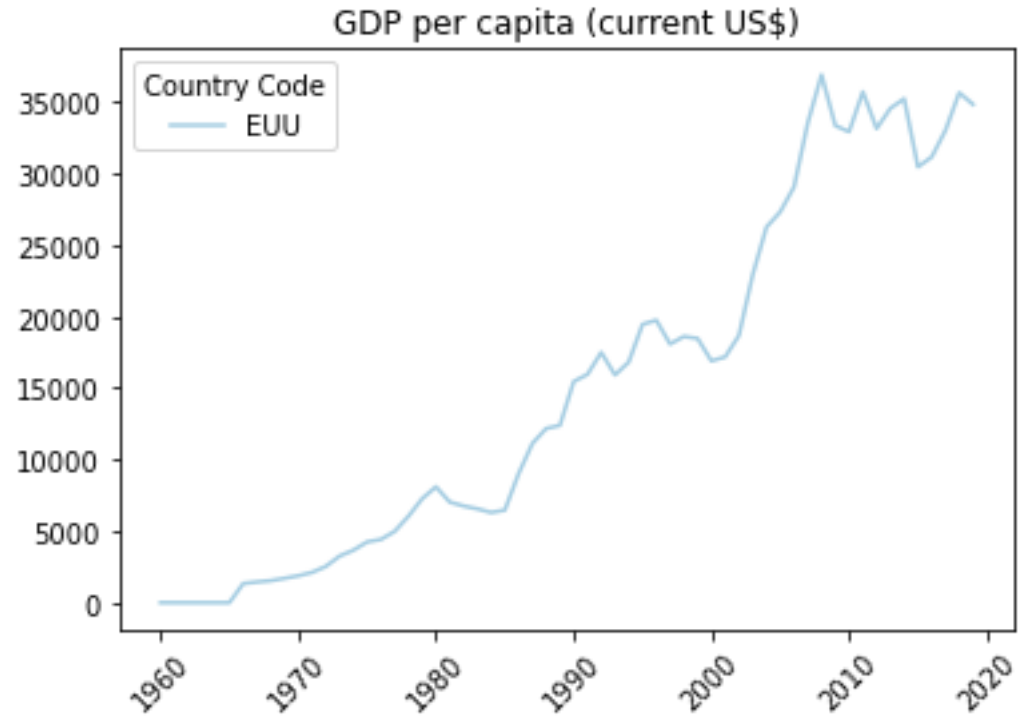

Figure 19. GDP per capita of EU

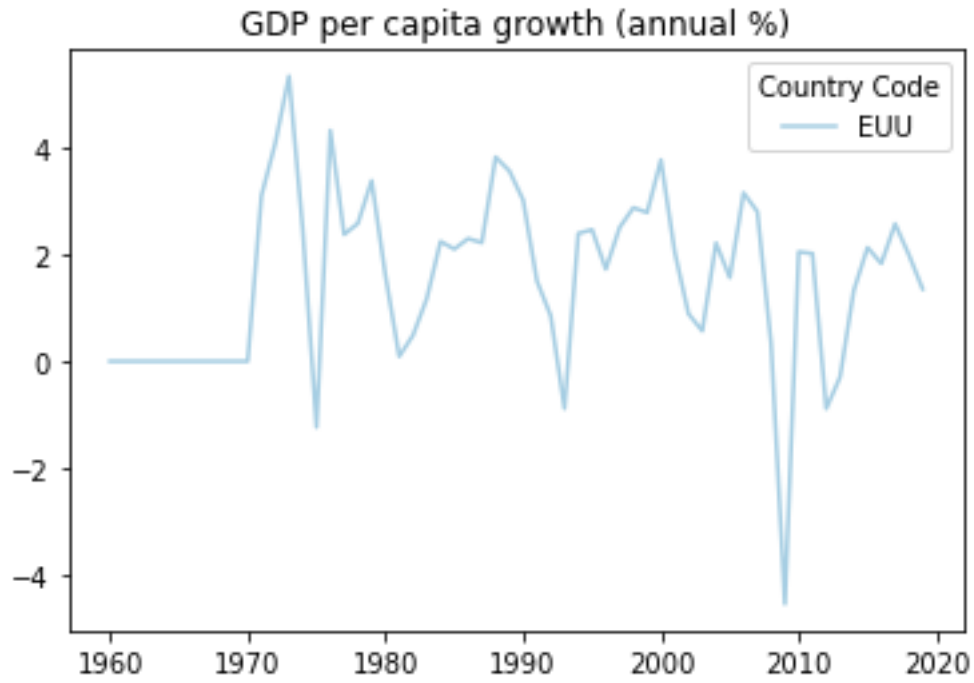

Figure 20. GDP per capita growth of EU

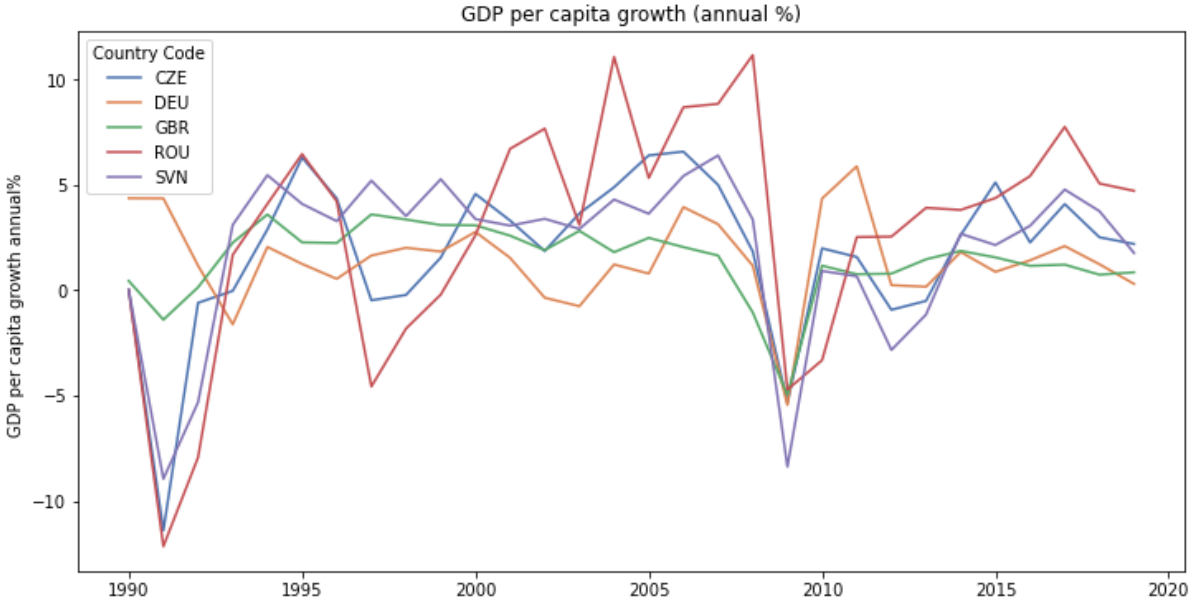

Figure 21.GDP per capita growth of selected countries

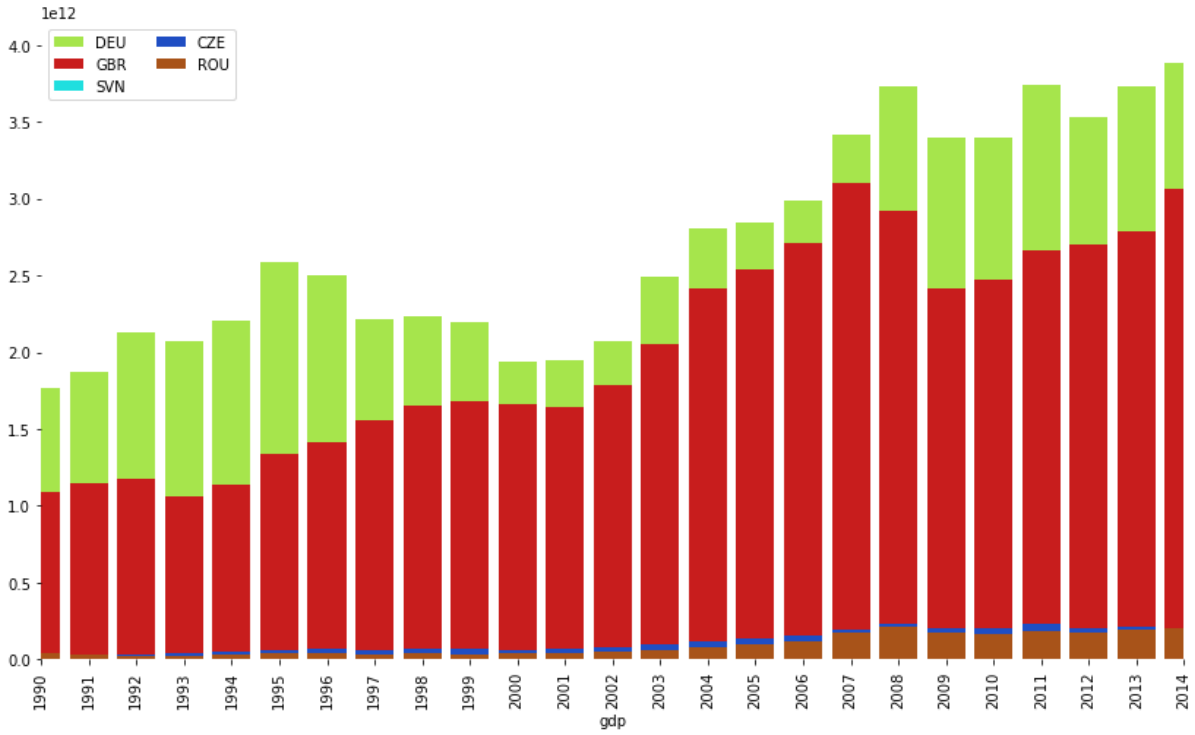

Figure 22. GDP divide between 2 large countries from Western Europe(DEU and GBR) and 3 countries from eastern part of Europe.

SARIMAX Results

Dep. Variable:	y	No. Observations:	28
Model:	SARIMAX(3, 2, 3)	Log Likelihood	-206.188
Date:	Sun, 13 Dec 2020	AIC	426.377
Time:	19:19:13	BIC	435.183
Sample:	0	HQIC	428.913
	- 28		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.0589	0.685	-0.086	0.931	-1.402	1.284
ar.L2	-0.5627	0.310	-1.818	0.069	-1.170	0.044
ar.L3	0.5200	0.482	1.079	0.280	-0.424	1.464
ma.L1	-0.2342	1.165	-0.201	0.841	-2.517	2.049
ma.L2	0.3061	0.663	0.462	0.644	-0.993	1.605
ma.L3	-0.9110	0.719	-1.267	0.205	-2.321	0.498
sigma2	6.142e+05	4.1e+05	1.500	0.134	-1.88e+05	1.42e+06

Ljung-Box (Q):	13.77	Jarque-Bera (JB):	4.14
Prob(Q):	0.97	Prob(JB):	0.13
Heteroskedasticity (H):	5.09	Skew:	-0.47
Prob(H) (two-sided):	0.02	Kurtosis:	4.72

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

Figure 23. SARIMAX results

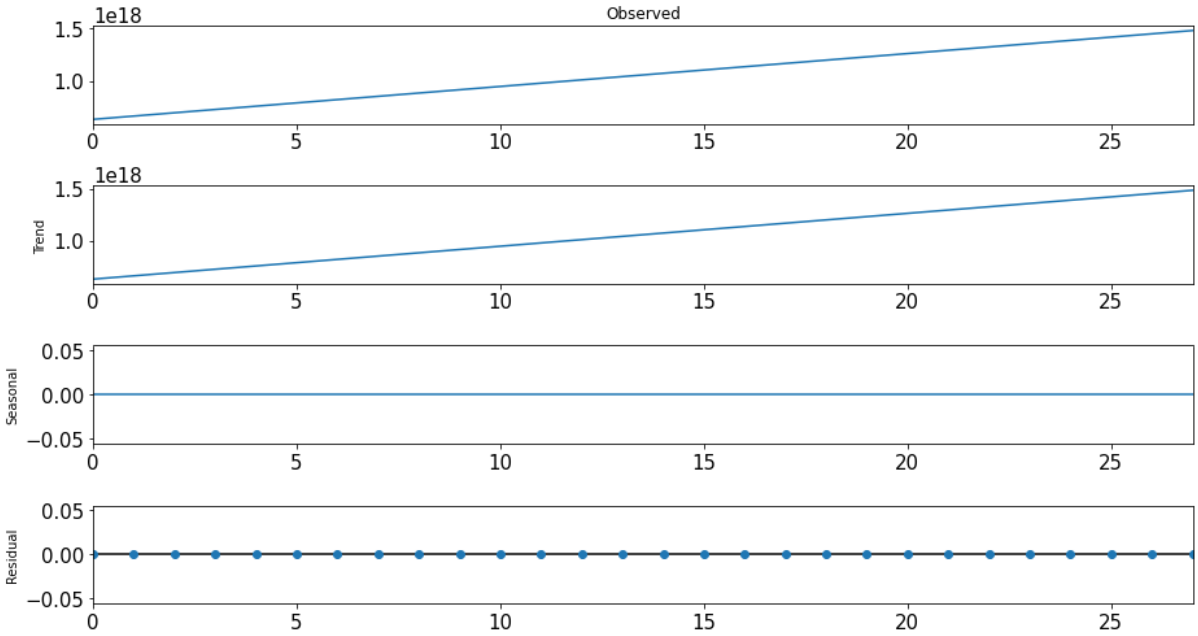

Figure 24. Decomposition

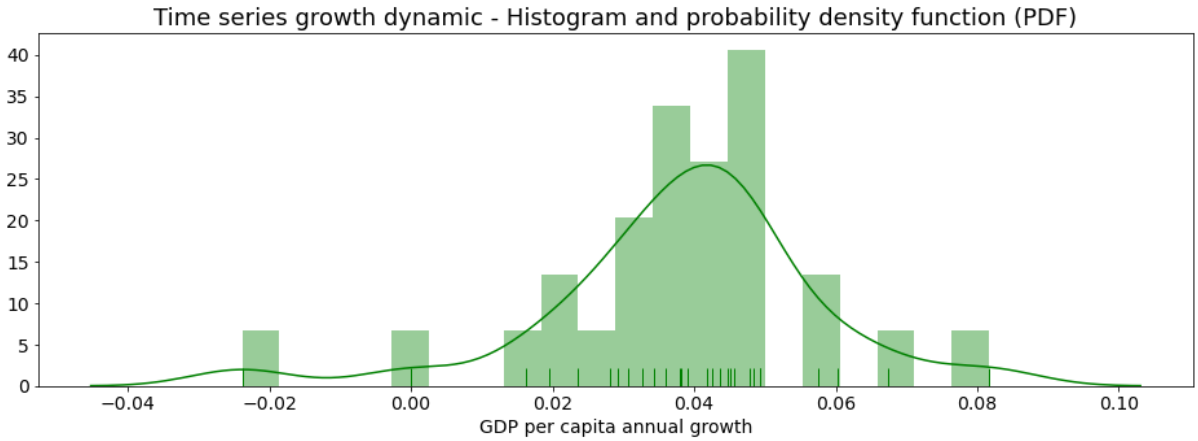

Figure 25. Time series growth dynamic