**ANL488 PROJECT Final Report**

**TIME SERIES FORECASTING OF CO2 EMISSIONS**

A picture containing logo

Description automatically generated

**Submitted by**

**Bryan Lim Tze Yuan**

**H1981079**

**SCHOOL OF BUSINESS**

**Singapore University of Social Sciences**

**Presented to Singapore University of Social Sciences**

**in partial fulfillment of the requirements for the**

**Degree of Bachelor of Science in Business Analytics**

**2022**

**ABSTRACT**

The problem of the earths rising temperatures can largely be attributed to the increased output of Greenhouse Gases (GHG) emitted by countries. With the population inevitably growing, the problem of rising temperatures is further exacerbated, to keep up with the population’s standard of living. This potential rise in earths temperature could further worsen the consequences that we are already facing due to global warming.

This paper thus aims to give an idea of future levels of GHG emissions by forecasting co2 emissions from the top 5 co2 emitting countries, giving insight as to how future co2 levels might look and their consequences. The dataset used for this forecast contains yearly co2 emission levels for all countries from the year 1960 to 2020.

The methods used in this study are multivariate LSTM, and univariate ARIMA, comparing a more recent Neural Network based model against a more traditional predictive modelling method. Comparing these 2 methods, we surmised that the more traditional method of forecasting, ARIMA, had better model accuracies overall.

The results of both the models showed that the United States and Japan were shown to continue the trend of increasing co2 emissions up to the year 2040, going above historical highs, while Russia was shown to have decreasing co2 emissions.

**TABLE OF CONTENTS**

[**Chapter 1.0** **Introduction** 4](#_Toc118651152)

[**Chapter 2.0** **Literature Review** 5](#_Toc118651153)

[**Chapter 3.0** **Data Understanding and Preparation** 9](#_Toc118651154)

[**Chapter 3.1** **Data Exploration** 10](#_Toc118651155)

[**Chapter 3.2** **Data Preparation** 12](#_Toc118651156)

[**Chapter 4.0** **Modelling and Evaluation/Discussion** 17](#_Toc118651157)

[**Chapter 4.1** **LSTM** 17](#_Toc118651158)

[**Chapter 4.2** **ARIMA** 28](#_Toc118651159)

[**Chapter 4.3** **Model Evaluation** 33](#_Toc118651160)

[**Chapter 5.0** **Conclusion** 34](#_Toc118651161)

[**Chapter 5.1** **Limitations** 34](#_Toc118651162)

[**References** 36](#_Toc118651163)

# **Chapter 1.0 Introduction**

Global warming, caused by concentrated GHG emissions, has caused an unprecedented rise in global temperatures in the recent decades. With an increase of about 0.85 °C of land and ocean temperatures from 1880 to 2012 (Tang, 2019), the earth is now the hottest it has ever been. Some reports have also stated that the worlds temperatures between 1983 and 2012 have been the warmest in 800 years (IPCC, 2018).

While greenhouse gasses are key in keeping the planet inhabitable through maintaining its temperature, the problem comes when there is an excess of said gasses. Excessive GHG emissions prevent heat from the Earth from escaping into space, due to its their heat trapping capabilities, which, over time, causes the aforementioned rise in temperature.

This rise in temperature affects not only us but also threatens the biodiversity of the planet. Effects like rising sea levels as well as lengthened wildfire seasons are only some of the few consequences resulting from an increase in global temperature.

Unfortunately, GHGs such as co2 are a by-product of many industries that are essential in maintaining our standard of living. For example, the agriculture industry, is one of the biggest contributors of GHG emissions, and is responsible for more than a quarter of global emissions (Homaira & Hassan, 2021). As the population inevitably grows, we are faced with the challenge of sustaining our necessities such as food and transport, whilst keeping rising temperatures in check.

While GHG’s such as methane have higher heat trapping capabilities when compared to others such as co2 (UCAR, n.d.) this study will focus solely on co2 emissions, due to it taking up the vast majority of overall GHG emissions, taking up almost 80% of total GHG emissions globally (C2ES, n.d.).

We have seen a meteoric rise in atmospheric co2 levels since the 1960s, largely due to industrialization as well as population growth. Emissions have risen from about 11 billion tons of co2 per year in the 1960s, to 35 billion tons of co2 per year in the 2010s (Lindsey, 2022), a trend which has shown little to no signs of stopping.

This report thus aims to discuss the future of co2 emissions, should we continue this upward trend, through the application of forecasting. More specifically, we will be conducting forecasting using 2 methods, Long Short-Term Memory (LSTM) and Autoregressive Integrated Moving Average (ARIMA). One of the goals of this study is to determine if newer models, such as LSTM, are better suited at forecasting than more traditional methods such as ARIMA. Through comparing the accuracy of both methods, we will be able to ascertain the best model to use to attain as accurate a forecast as possible. For the purposes of this study, we will be forecasting potential co2 emissions from the top 5 highest co2 producing countries over the next 20 years.

# **Chapter 2.0 Literature Review**

This chapter aims to discuss and review literatures pertaining to time series forecasting, as well as summarize their findings and insights on different forecasting methodologies.

Radojević et al. (2013) conducted a study on GHG emissions in Serbia using Artificial Neural Networks (ANNs). Using the countries performance indicators such as GDP and energy consumption as input parameters, Radojević et al. (2013) used ANN to forecast GHG emissions.

The resulting model showed satisfactory results, having an R-Squared value of 0.9125, in essence having a relative error rate of less than 10%. Radojević et al. (2013) states that ANN models are suitable for predicting GHG emissions, given the non-linear nature of the problem as well as not having strict mathematical relationships between the variables (Radojević et al., 2013)

However, Radojević et al. (2013) also states that other ANN architectures, such as Recurrent Neural Networks (RNN) should be tested to improve the quality of the models, due to their ability to factor in previous values in a time series. Thus, RNNs such as LSTM would suit our use case far better, given the upward trend of co2 emissions. Factoring in previous values in a time series may prove imperative in getting a more accurate forecast for future co2 emission levels.

Siami-Namini et al. (2018) carried out a comparison between traditional forecasting methods and deep-learning algorithms for time series forecasting, specifically Auto-Regressive Integrated Moving Average (ARIMA) and Long Short-Term Memory (LSTM) respectively.

Siami-Namini et al. (2018) used the 2 techniques and conducted data mining on stocks from various indexes on Yahoo finance, predicting the “Adjusted Close” price based on previous closing prices in the time series, at a ratio of 70% training data to 30% testing data for both techniques. The following results from the 2 techniques were then assessed using metrics such as Root-Mean-Square Error (RMSE), comparing the accuracy of the 2 techniques.

Siami-Namini et al. (2018) found that LSTM performed vastly superior to ARIMA, with an estimated 84% - 87% reduction in error rates, likely due to the iterative nature of deep-learning algorithms like LSTM.

Similarly, Ludwig (2019) did a similar comparison of three different methods of predictive modelling: Adaptive Neuro-Fuzzy Inference System (ANFIS), Recurrent Neural Network (RNN) and LSTM. These 3 techniques were chosen thanks to their nonlinearity, as compared to traditional forecasting techniques such as ARIMA.

Using time series data of GHG concentrations in 2,921 different grid cells in California, Ludwig (2019) used the 3 aforementioned techniques to forecast GHG emissions. Using evaluation measures such as Root Mean Squared Error (RMSE) and Mean Absolute Error

(MAE), Ludwig (2019) found both ANFIS and LSTM to perform favourably, with RNN performing the worst amongst the three.

Ludwig (2019) attributed this difference in performance to RNNs shortcomings, that being the network only remembering the past few steps in the data sequence and is thus unsuitable for longer time series data (Ludwig, 2019). As for LSTM and ANFIS, it was found that ANFIS outperformed the other 2 when using RMSE as an evaluation measure, and LSTM outperformed the other 2 when using MAE as an evaluation measure.

However, one notable observation is that LSTM took significantly longer to execute, as compared to ANFIS, with a run time of 150.39 seconds as compared to ANFIS’ 6.87 seconds (Ludwig, 2019).

While speed is generally an important factor when building models, this study prioritizes model accuracy over speed. The nature of this study aims to forecast co2 emissions for future years, and thus does not need any real-time updating of models, due to the annual nature of the data. Hence, LSTM is still an appropriate methodology to conduct forecasting with, despite Ludwig’s (2019) conclusion that LSTM is slower than other methods such as ANFIS.

Manowska (2020) used LSTM to conduct predictive modelling on electricity consumption in Poland. Using electrical consumption rates based on each different sector, Manowska (2020) was able to forecast electricity consumption with satisfactory results, using Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE), as performance metrics.

More specifically, Manowska (2020) found that the LSTM model had a MAPE of 1%, lower than the assumed 5% error rate. Manowska (2020) also finds that certain industries such as the industry sector may actually decline in overall energy consumption, linking it to other factors such as legal regulations that are outside of the scope of the model.

Manowska (2020) also states that the model operates under many assumptions and its accuracy is subject to uncertainties such as technological development and the aforementioned legal regulations that may affect the credibility of the model.

While not strictly pertaining to co2, insights gained from Manowska’s (2020) predictive modelling on electricity can still largely be used for our analysis, considering that co2 is a by-product of generating said electricity, accounting for about 38% of total energy-related co2 emissions in 2015 (Goh et al., 2018).

Homaira & Hassan (2021) used ARIMA, LSTM, as well as a simple linear regression model to predict agricultural emissions in Malaysia.

Homaira & Hassan (2021) used datasets consisting of values such as temperature and total emissions from 2000 to 2017 to build their models. Once again, evaluations measures such as RMSE, MAE and MAPE were used to calculate the accuracy of the forecast. For the ARIMA model, evaluation measures such as Akaike Information Criterion (AIC) were used to find the best order of the model.

All models were found to have satisfactory results, all of which forecasted a rising trend in emissions for Malaysia.

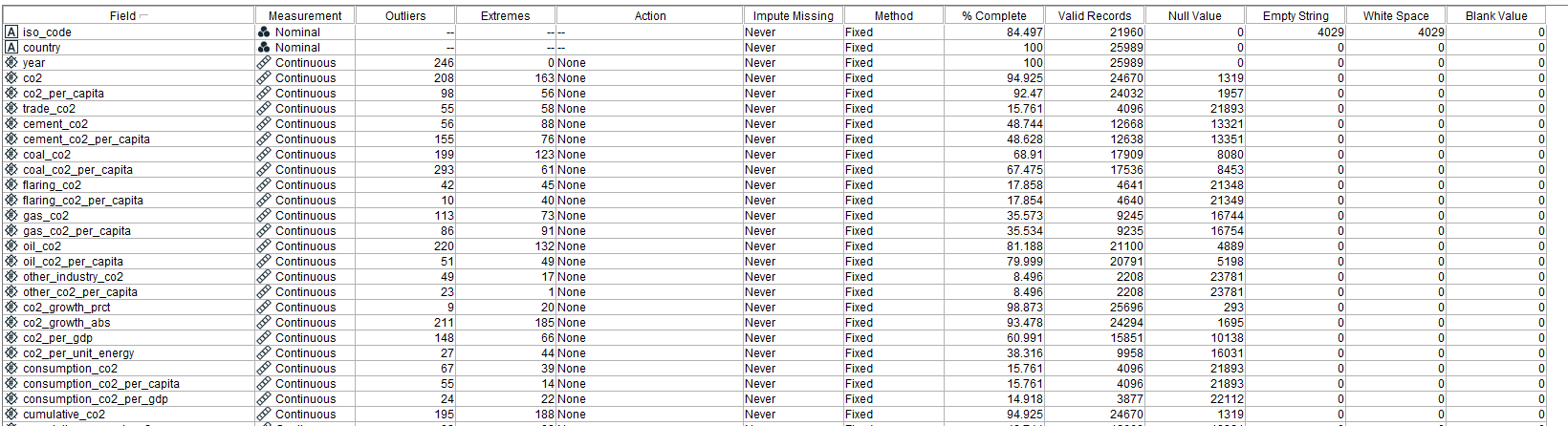
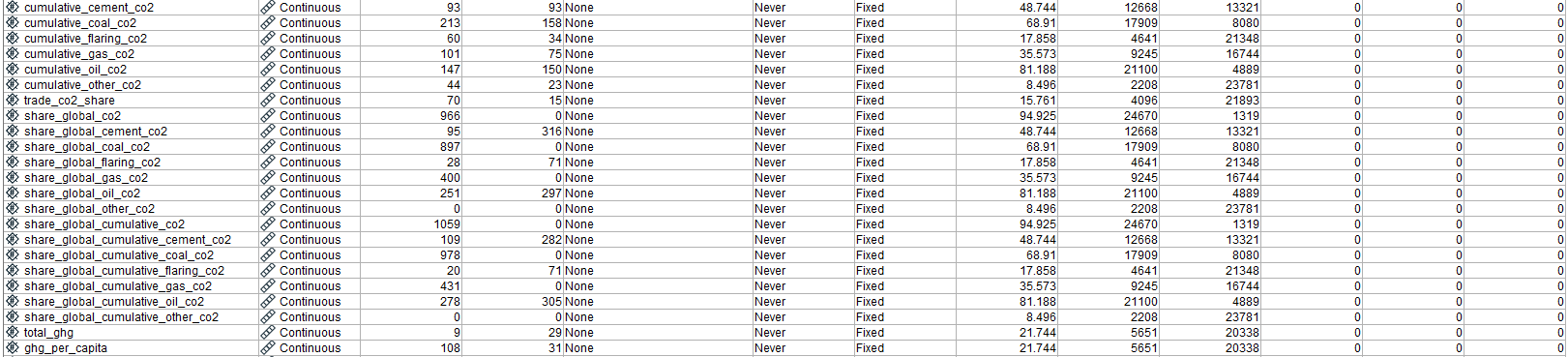
# **Chapter 3.0 Data Understanding and Preparation**

For the purposes of this study, we will be using one dataset, “owid-co2-data”. This dataset contains co2 and co2 equivalent emissions of 248 countries and regions. This dataset not only contains the total co2 emissions of every country, but also the amount of emissions from specific industries, such as oil and coal. This emission data is annual, containing each countries emission amounts for each year, spanning as far back as 1750, all the way up to the year 2020. A record of previous years data allows us to gain insight on the emission levels for each country and ascertain if there are any trends or seasonality with co2 emissions. Forecasting is also dependent on how much data we have to build the model off of.

The dataset is sourced from github and is a repository of emission data from various sources, such as Our World in Data and Global Carbon Project. Organisations such as Our World in Data collect and publish these metrics gathered from research articles and official channels such as government websites (Ritchie et al., 2022).

Through the use of this dataset, we can conduct forecasting to gauge the likely levels of co2 emissions for the next 20 years, by using historical levels of co2 emissions from the top 5 countries.

## **Chapter 3.1 Data Exploration**

While this project aims to use Python to build an LSTM and ARIMA model for forecasting, preliminary data exploration will be done using SPSS Modeller.

*Figure 3.0 Data exploration results*

Using the Data Audit Node, we find that the dataset has only 3.33% of the fields being complete, with several missing, extreme, as well as outlier values. The purpose of this study is to forecast future co2 emissions based on historical data. However, simply using the past co2 levels to forecast may not be entirely accurate, as there are likely other factors that affect co2 emissions, such as the Gross Domestic Product (GDP) and population of a country. Hence, for a more holistic approach to forecasting, we will be using a total of 5 fields, those being: “country”, which describes which country or region the co2 emission comes from, “year” which is the date of the recorded co2 level, “co2” which measures each countries production-based CO2 emissions in million tonnes annually, “gdp” which is the country’s GDP for that year, and “population” which is the country’s population for that year.

*Figure 3.1 Data exploration results cont.*

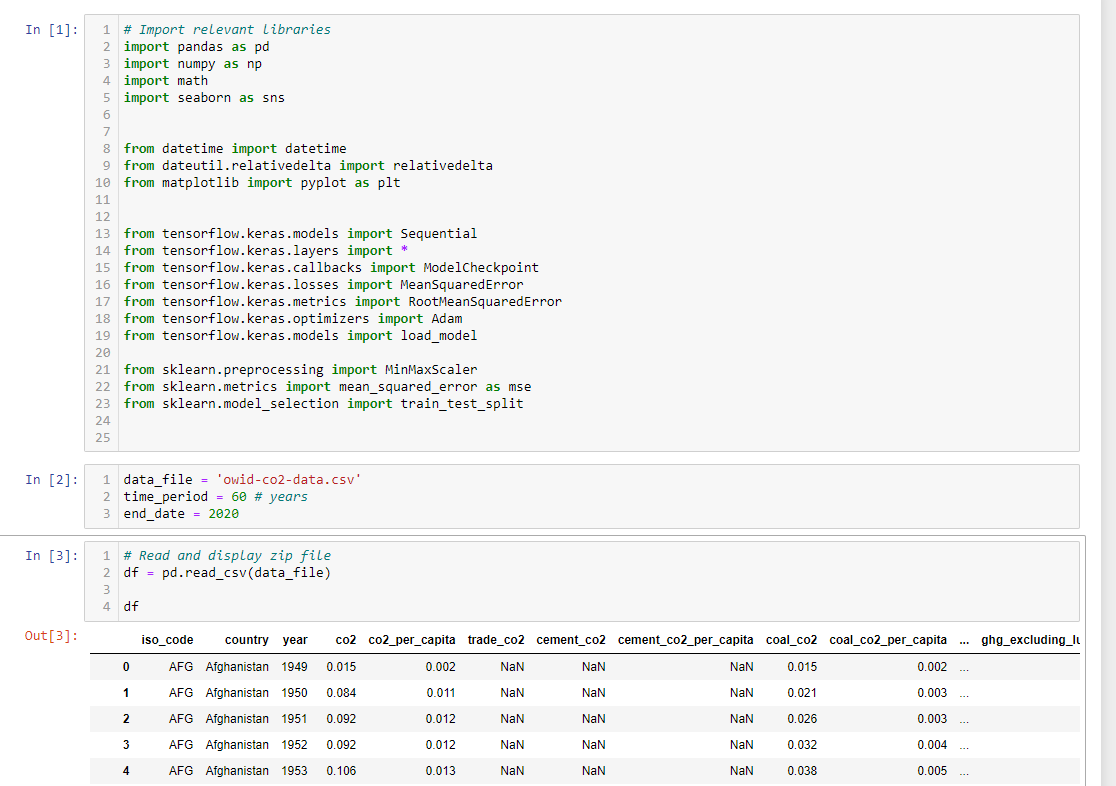
For the “co2” field, there are a number of outliers and extreme values (208 and 163 respectively). This is likely due to the fact that there is CO2 emission data for groups of countries, such as “Upper-middle-income countries”, as well as global CO2 emission levels. Emission data like these are a cumulation of multiple countries and will thus have much higher values than emissions from a single country.

“Population” and “gdp” also have their data issues, with them only being 88% and 51% complete respectively. This is likely due to the fact that some rows may only contain co2 data for that country for that year and may not always have their gdp and population numbers available.

## **Chapter 3.2 Data Preparation**

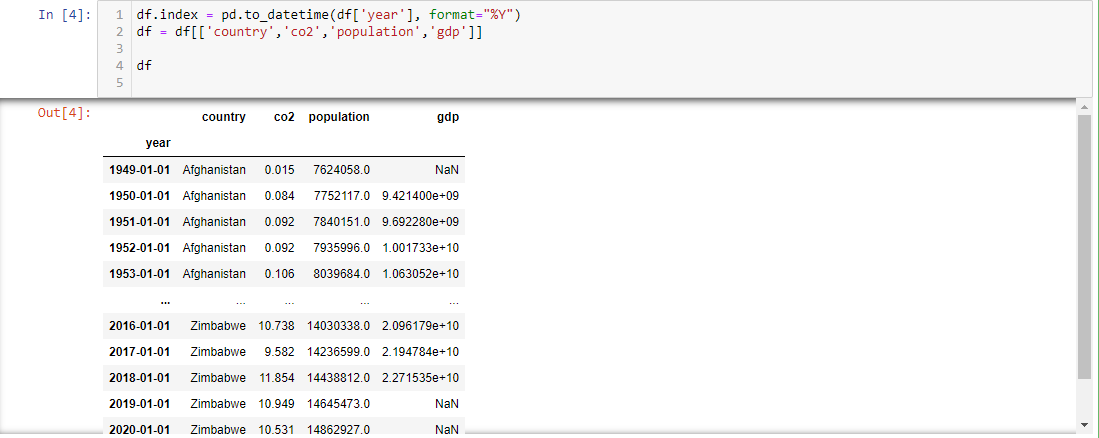
Because the goal of this study is to forecast emissions for the top 5 co2 producing countries, the data will need to be prepared before conducting forecasting. This data cleaning process applies to both the LSTM and ARIMA models, as they will both be using the same cleaned dataset.

After first importing the relevant libraries, the csv file is read and put it into a dataframe.



*Figure 3.2 Importing and reading of data file*

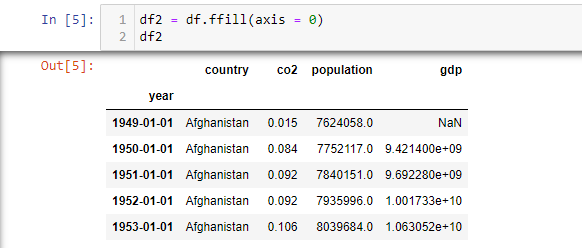
This study aims to conduct multivariate LSTM and univariate ARIMA for co2 emissions. Hence, unneeded columns are filtered out as such:



*Figure 3.3 Filtering out unnecessary fields*

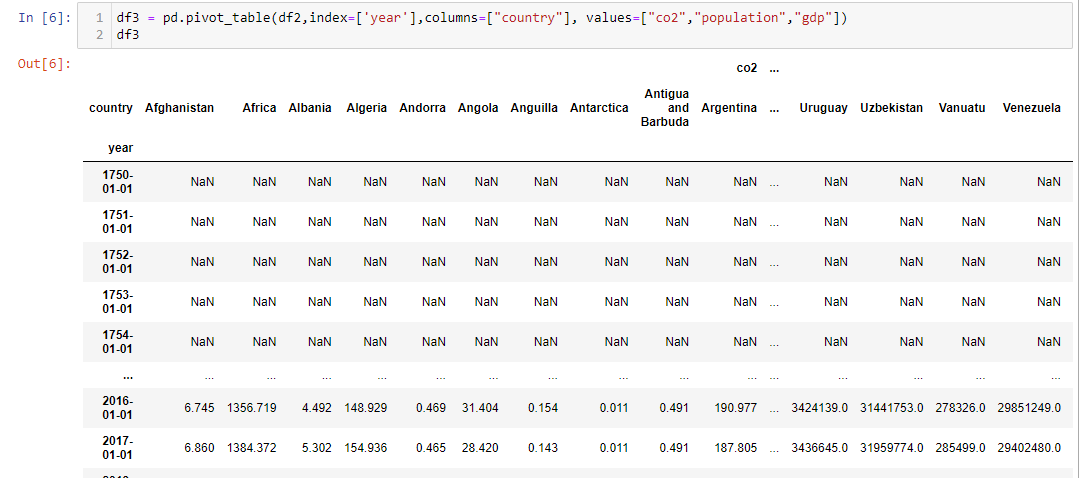
This leaves the country, the year, the co2 emission level that year, the population and GDP.

Rows that contain a null or blank value for co2, population and gdp are filled with the previous year’s value. This will give a more accurate forecast, as null values may affect the overall accuracy of the model. Co2, population and GDP are also factors that are unlikely to have large changes within a short span of time, hence using the previous year’s entry is suitable.



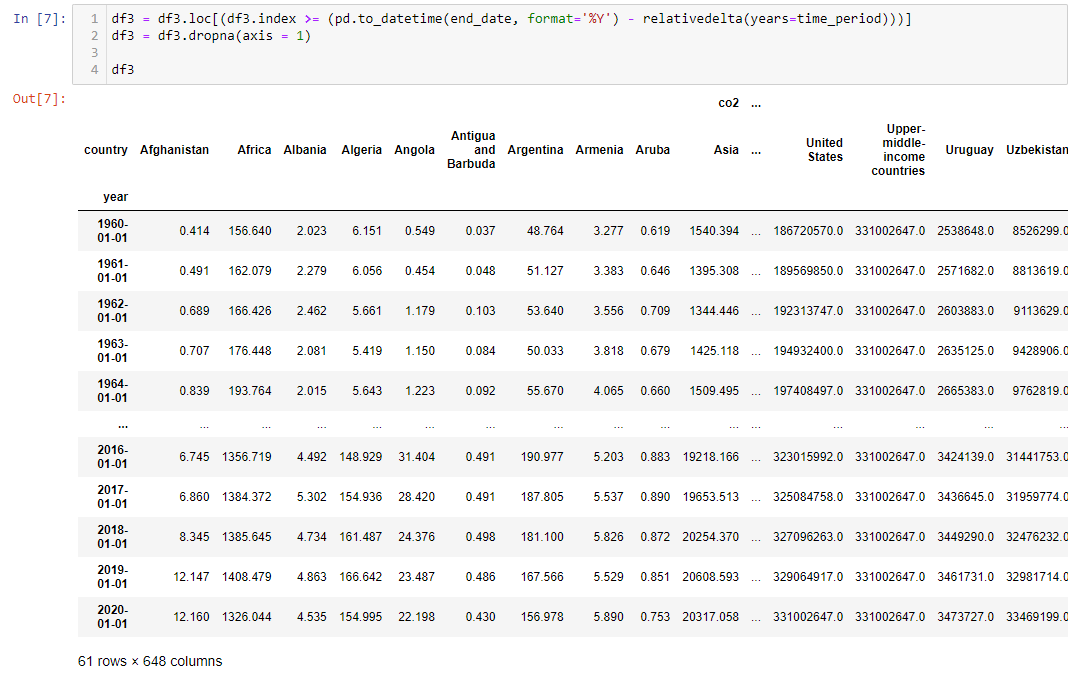
*Figure 3.4 Forward fill data*

Afterwards, the dataframe is pivoted, giving a better look at the CO2 emissions from each country year over year:



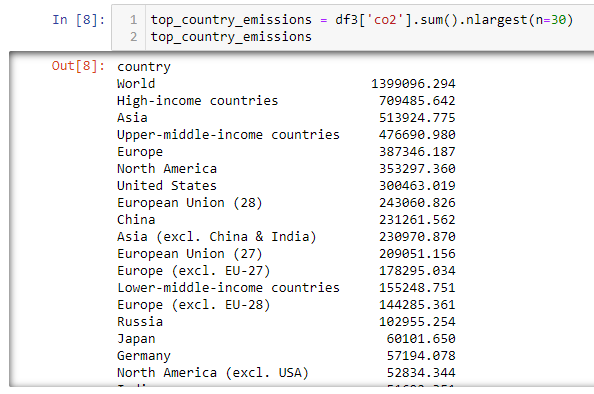
*Figure 3.5 Pivoting table*

As mentioned earlier, we have seen an unprecedented rise in co2 emissions starting from around the year 1960. Hence, we will be using data from the year 1960, up to the latest year of record, that being the year 2020.



*Figure 3.6 Filtering out data before 1960*

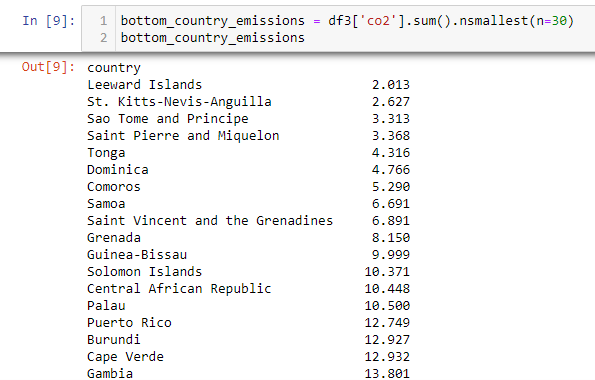
Finding the top 5 GHG emitters, we sum up the total CO2 emissions for the past 60 years and list the highest results.



*Figure 3.7 Top emitting countries and regions*

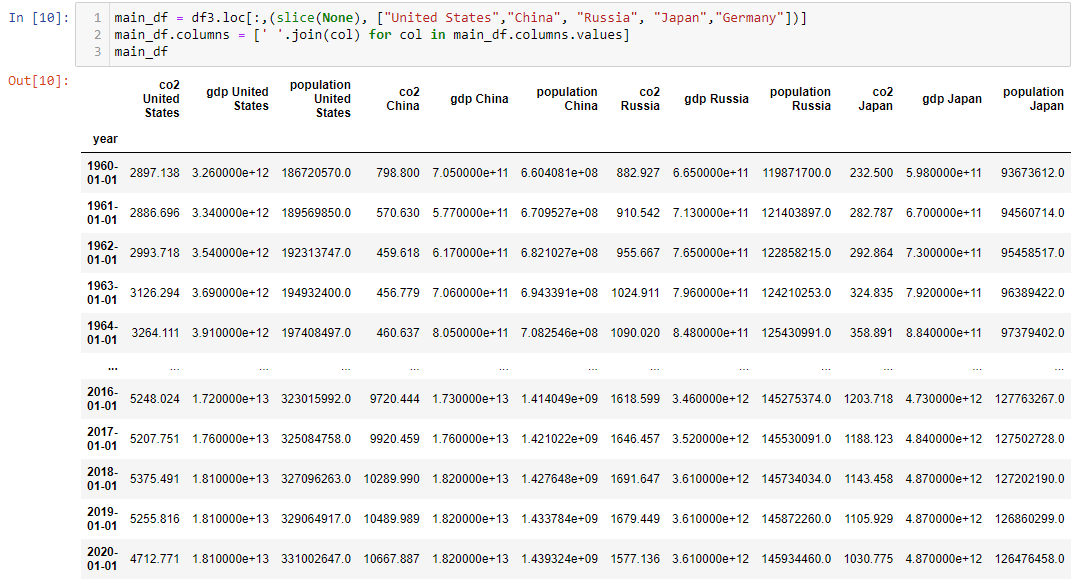
Unfortunately, due to the dataset containing global data as well as aggregated emissions, only rows that are part of the 195 countries (Worldometer, n.d.). will be selected. Doing this, the top 5 countries are: United States, China, Russia, Japan and Germany.

The lowest emitters are as follows:



*Figure 3.8 Lowest emitting countries and regions*

Given this, we extract the co2, population and GDP values of the top 5 countries and put them into a single dataframe. The final cleaned dataframe used to conduct time series forecasting on looks as such:



*Figure 3.9 Final dataframe*

# **Chapter 4.0 Modelling and Evaluation/Discussion**

This chapter goes through the building process and the results for the LSTM and ARIMA models, followed by evaluation and comparison of both models.

Model evaluation will be done in a similar manner to the previously discussed literature. Using Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE) as performance metrics, we should be able to ascertain the overall accuracy of both the LSTM and ARIMA model.

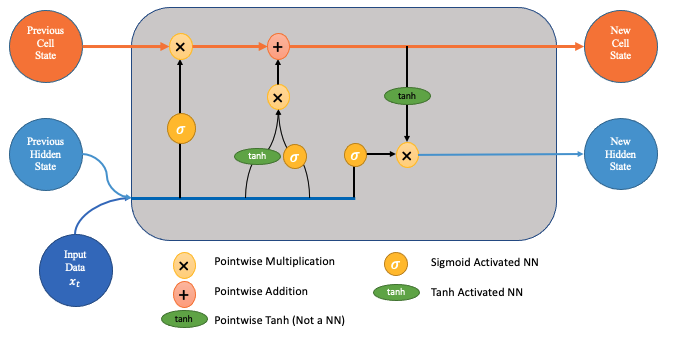
## **Chapter 4.1 LSTM**

As evidenced by the literature reviews, LSTM was shown to be an appropriate predictive modelling technique to use, given the nonlinear nature of co2 emissions, as well as the LSTMs ability to handle sequential data.

LSTM is a type of Recurrent Neural Network (RNN), which aims to predict the next value based on a sequence of past datapoints, by learning from them. Different from traditional feed-forward neural networks, RNNs hidden layers store information of the earlier sequence, allowing them to leverage past datapoints to predict future values (Siami-Namini et al., 2018). However, typical RNNs suffer from what is called a vanishing gradient problem, whereby the network has trouble remembering datapoints the earlier on in the sequence, basing its predictions on only more recent datapoints (Kostadinov, 2017).

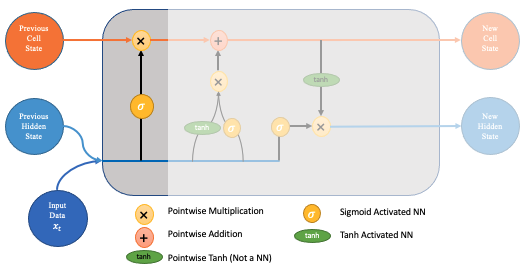
LSTM thus aims to solve this memory problem through additional features, such as each cell having gates that either drop, filter or add data for the subsequent cell (Siami-Namini et al., 2018). This forms something of a transport line between cells, transferring and gathering data. The three gates that alter each cells state are the Forget Gate, Memory/Input Gate, as well as the Output Gate.

The diagram below from Dolphin (2020) shows the process behind each cell in an LSTM network.



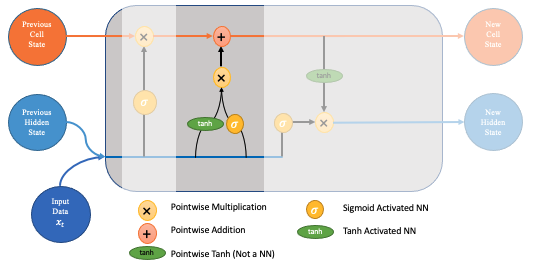
*Figure 4.1 LSTM Diagram*

The Forget Gate defines how much information to forget and remember, and outputs a number between 1 and 0, with 1 being “remember everything” and 0 being “forget everything”, this value output is dependent on how relevant the Forget Gate deems the new input data to be, given the previous cells state. This value is then multiplied with the previous cells state, with a number close to 0 having a smaller effect (Dolphin, 2020).



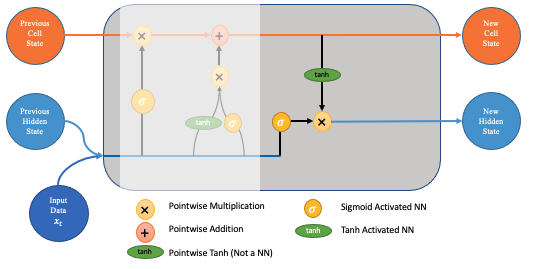
*Figure 4.2 Forget Gate*

The Memory/Input Gate then chooses what information is to be added or updated in the cell, by combining the new input data with the previous cells hidden state to get an “update vector”. This update vector is then added to the current cell state, giving us a new cell state (Dolphin, 2020). Similar to the Forget Gate, the Memory and Input Gate filter out how much of the update vector is worth keeping, using tanh and sigmoid activation functions.



*Figure 4.3 Memory/Input Gate*

Finally, the Output Gate determines the new hidden state of the cell by once again filtering the newly updated cell state from the previous step.

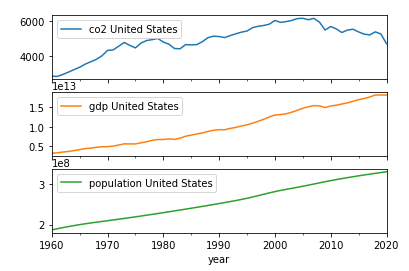
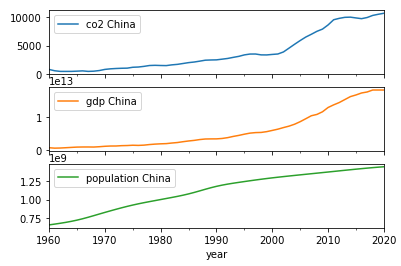


*Figure 4.4 Output Gate*

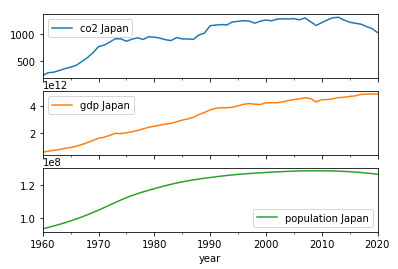
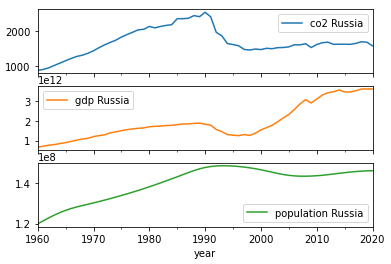
The LSTM model will be using 3 inputs: namely the co2 levels of the country, the population as well as the GDP of said country. Predictive modeling and forecasting are conducted on all 3 inputs, but for the purposes of this study only the forecasted co2 levels will be looked at.

Multivariate LSTM is used in this case to provide a more holistic approach to forecasting of co2 values, as population and GDP of a country are likely factors in determining a country’s co2 emission levels.

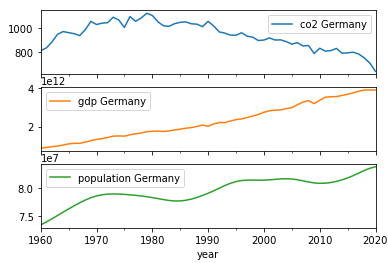
Below are the co2, GDP and population numbers from the year 1960 to 2020 for the United States, China, Russia, Japan and Germany respectively:

****

*Figure 4.5 United States Figure 4.6 China*

****

*Figure 4.7 Russia Figure 4.8 Japan*



*Figure 4.9 Germany*

The LSTM model is then trained and tested against the historical data for all 5 countries.

Normalization is done before fitting the data to the model, changing all of the values to be between 0 and 1. With the smallest value being normalized to 0, and the highest value being normalized to 1. This is applied to all 3 of the inputs. This is done to improve the accuracy of the model, preventing the difference in scale between the numbers from affecting the model, whilst maintaining the ratio and distribution between the different fields. This is especially so in the case of GDP, which can range in the millions, whereas co2 emissions mostly range within the thousands.

After the normalization, the data is then sequenced appropriately to fit the LSTM model. The model will use 5 inputs (timesteps) with 3 features (co2, gdp & population) to predict the next values of the 6th timestep. Essentially, our input shape will be: [[a1,b1,c1],[a2,b2,c2],[a3,b3,c3],[a4,b4,c4],[a5,b5,c5]]

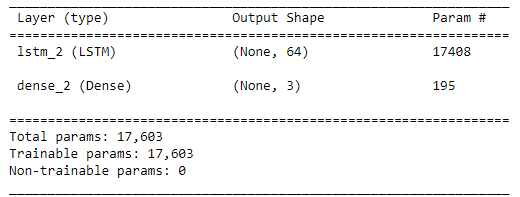
With our ouput shape being:

[a6,b6,c6]

whereby we take the co2, GDP and population of the first 5 years from 1960 to 1964 to predict the values in 1965, then using the values from 1961 to 1965 to predict the values in 1966, so on and so forth.

The model is then fit against these input and output values, learns and finally attempts to predict the values for each years given the previous 5 values. This predicted value is then compared with the actual value for that year, whereby we ascertain the accuracy of the model using the aforementioned RMSE and MAPE.

The LSTM model is chosen to have 64 memory inputs, with an output shape of 3, giving us the co2, GDP and population.



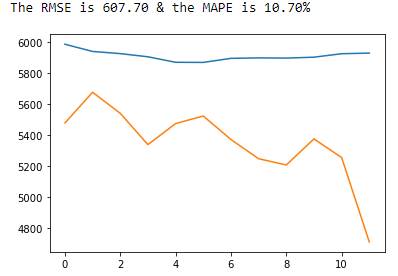
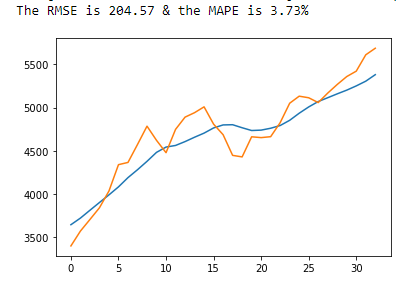
*Figure 4.10 Model Summary*

The model is compiled using “Adam” as the optimizer with a default learning rate of 0.001, through 20 epochs.

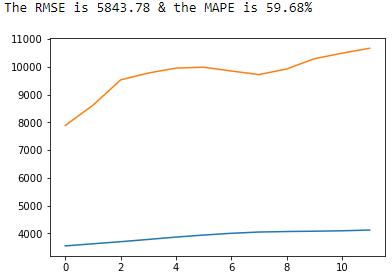
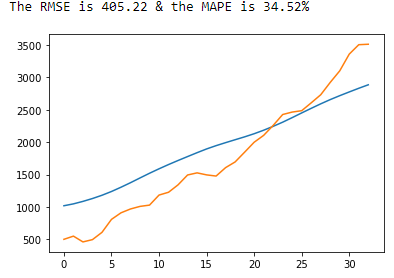
The Adam optimizer, one of the most popular algorithms, is chosen as it has a multitude of benefits, such as being computationally efficient, as well as being appropriate for non-stationary objectives (<https://machinelearningmastery.com/adam-optimization-algorithm-for-deep-learning/>), which the target is.

To train the model, the data is split 60% for training and 20% for validation, with the remaining 20% being used to test the model as it is unseen data.

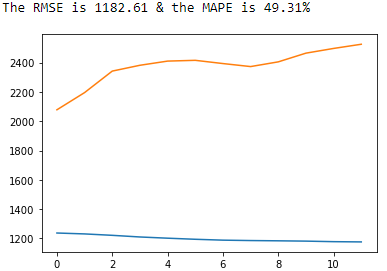
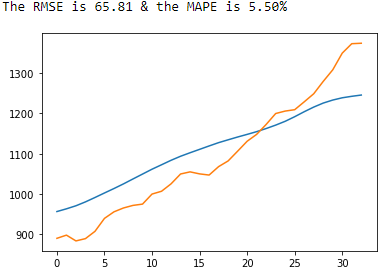
Fitting the models for all 5 countries, we get the models accuracy for both the training data as well as the testing data.



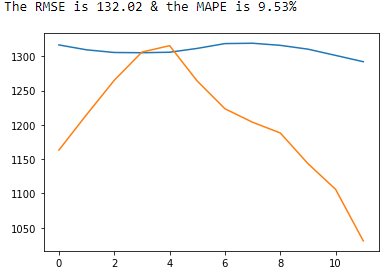
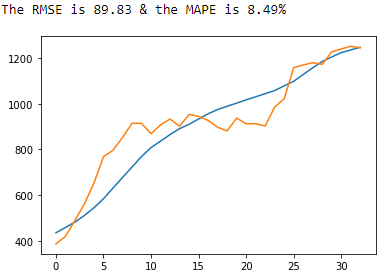
*Figure 4.11 United States training accuracy (left) and United States testing accuracy (right)*



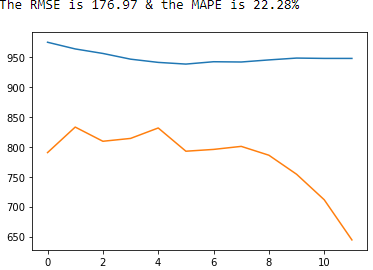
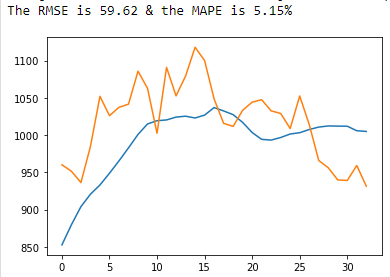
*Figure 4.12 China training accuracy (left) and China testing accuracy (right)*



*Figure 4.13 Russia training accuracy (left) and Russia testing accuracy (right)*



*Figure 4.14 Japan training accuracy (left) and Japan testing accuracy (right)*



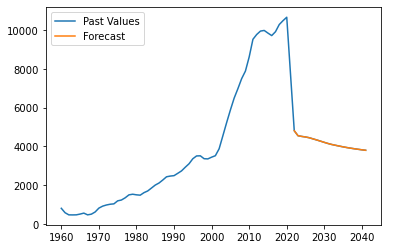
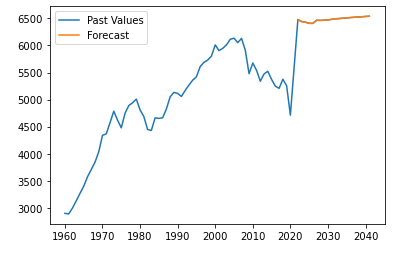
*Figure 4.15 Germany training accuracy (left) and Germany testing accuracy (right)*

Using the MAPE of the testing datasets, the LSTM models have varying levels of accuracy for all models. Both United States and Japan models accurately predicted the test data, with an MAPE of around 10% for both, meaning that each predicted value only deviated from the actual co2 value by about 10% on average.

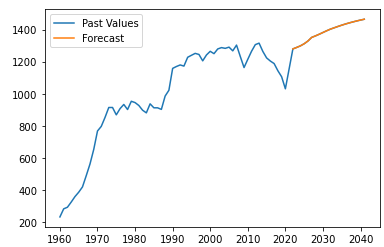
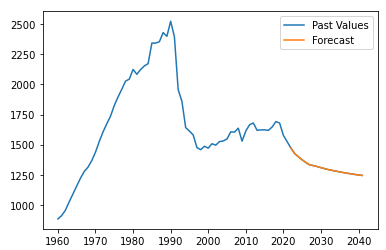
However, the same cannot be said for China, Russia and Germany models, the least accurate of which is the China model, with an MAPE of almost 60%.

This inaccuracy can likely be attributed to the sudden spikes and drops in co2 emissions from these three countries, with the United States and Japan’s co2 emissions being more stable.

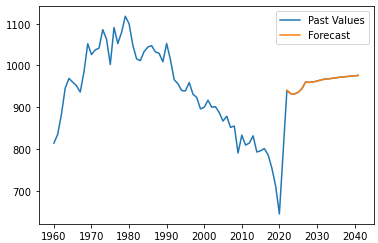
Using the model to forecast co2 values up till the year 2040, the model outputs the following results for the 5 countries:



*Figure 4.16 United States forecast (left) and China forecast (right)*



*Figure 4.17 Russia forecast (left) and Japan forecast (right)*



*Figure 4.18 Germany forecast*

Looking at the forecasts, both China and Russia are expected to decrease co2 emissions from the year 2020 to 2040, with a co2 levels of 3795 million tons and 1244 million tons of co2 emitted in the year 2040 respectively.

For the US and Japan, the model forecasts an increase in co2 emissions, with a forecasted 6540 million tons and 1464 million tons of co2 emitted in the year 2040 respectively.

For Germany, the model forecasts an increase in co2 emissions, but does not forecast it to hit levels higher than its peak at around the year 1980. Germany is forecasted to emit 976 million tons of co2 in the year 2040.

## **Chapter 4.2 ARIMA**

ARIMA gets its name from its 3 components: AutoRegressive (AR), Integrated (I) and Moving Average (MA).

Autoregression deals with finding a correlation between a specific value and a prior/lagged value, essentially seeing if a variable, which in this case is the co2 emission level, has any correlation to its past values.

Integrated deals with making data stationary, essentially ensuring that properties of the data, such as mean and variance, are constant over time. This is useful in removing trend from a time series, such as in the case with the co2 emission levels, giving us a more accurate forecast.

Moving Average in an ARIMA model finds the dependency between a specific value, and the error from a moving average model applied to previous values.

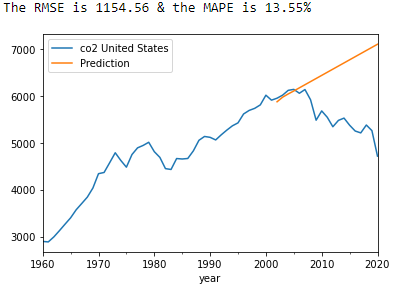
ARIMA models are useful in forecasting time series data and is especially useful when trying to predict time series data that is non-stationary. For this study, Seasonal ARIMA is not used, due to our co2 emissions data being annual, and not displaying any seasonal effects.

This study will use the “pmdarima” library to build the ARIMA model. The first step is to ascertain the best parameters for the ARIMA model, that being the p, d and q values. The p value indicates the lag order, d indicates the number of times the time series is differenced, and q is the moving average window size.

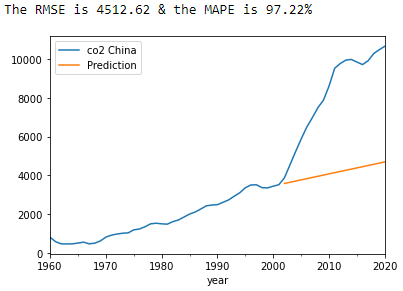
The pmdarima library uses Akaike Information Criterion (AIC) to determine the best p, d and q order. This process runs through different p, d and q combinations and finds the values that give us the lowest AIC value, this combination is then deemed to be the best order. The library also uses Augmented Dicky-Fuller (ADF) test to determine the best d order for differencing.

After determining the best order for all 5 models, the data is then split into training and testing datasets of 70% and 30% respectively. The model is fit to the training data, and we ascertain the model’s accuracy by once again using RMSE and MAPE, comparing the predicted co2 values with the actual co2 values in the test dataset.

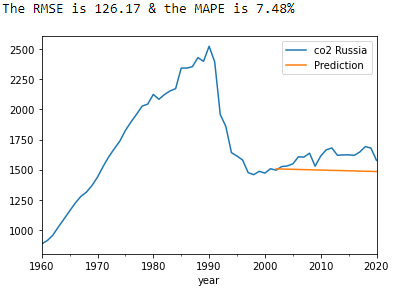
Doing this, we get the following predictions:



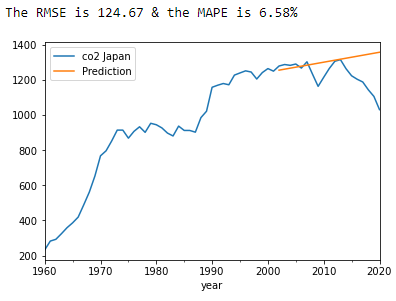
*Figure 4.19 United States predictions with ARIMA (2,2,2)*



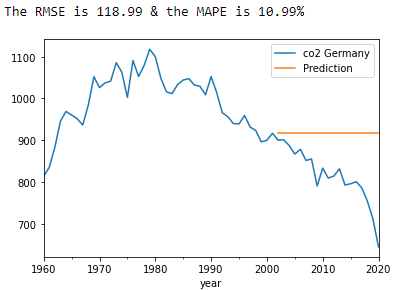
*Figure 4.20 China predictions with ARIMA (0,2,2)*



*Figure 4.20 Russia predictions with ARIMA (0,2,1)*



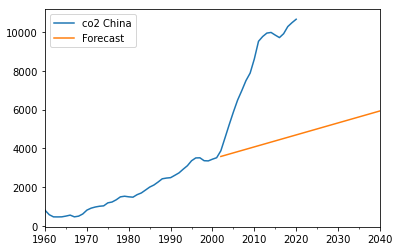
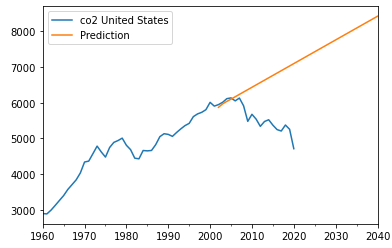
*Figure 4.21 Japan predictions with ARIMA (0,2,1)*



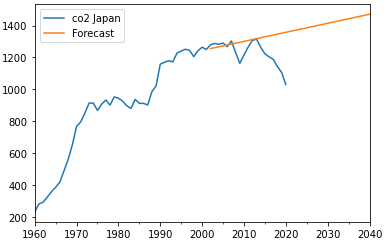
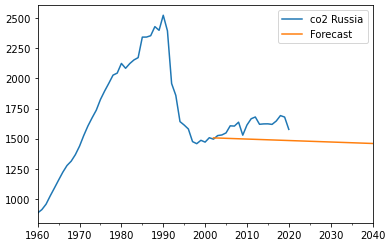
*Figure 4.22 Germany predictions with ARIMA (0,1,0)*

Using the MAPE values, the most of the ARIMA models have acceptable levels of accuracy, with most of them having an MAPE below 15%. However, the ARIMA model fails to accurately predict the co2 emission levels for China. This is likely attributed to the large spike in co2 emissions around the early 2000s, causing the models MAPE to be around 97%.

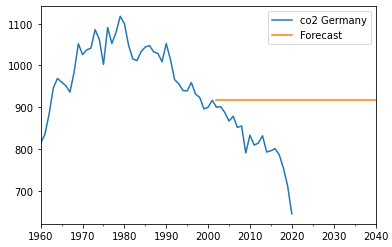
The forecast is then generated for all 5 of the countries:



*Figure 4.23 United States forecast (left) and China forecast (right)*



*Figure 4.24 Russia forecast (left) and Japan forecast (right)*



*Figure 4.25 Germany forecast*

Looking at the forecasts, the United States, China and Japan are forecasted to have a rise in co2 emissions from the years 2021 to 2040, with Russia being the only one to decrease their co2 emissions, emitting 1459 million tons of co2 in the year 2040.

For Germany, the model forecasts that its co2 emissions will stay the same, at a level of 916 million tons of co2 emitted every year until the year 2040. However, the reason for this stagnation is because it is an ARIMA model of (0,1,0) meaning that the time series is only differenced.

## **Chapter 4.3 Model Evaluation**

Looking at both model accuracies, overall, the ARIMA model performed better than the LSTM model when strictly looking at the MAPE for the testing data. However, given the fact that the ARIMA model is univariate, and is also only predicting a linear rise/fall in co2 emission levels, LSTM may prove to be a more accurate model overall.

Both the LSTM and ARIMA model both struggle with data with large spikes and dips between values, as evidenced by both models’ inability to properly predict China’s co2 emissions.

# **Chapter 5.0 Conclusion**

The problem of rising GHG emissions is a hard one to tackle, especially given the potential factors that affect emission levels in a country, such as a countries level of development.

While both models had varying forecasts for the 5 different countries, the most concerning countries are the United States and Japan, with both country’s co2 emission levels forecasted to rise up till the year 2040, hitting historically high levels of emissions. Russia was also the only country in which both models had predicted a decrease in co2 levels.

Predictive modelling and forecasting have enabled us to see the potential levels of co2 emissions by the year 2040. Insights like these allow us to ascertain which country may need to further their efforts in reducing co2 emissions, and potentially tackle underlying issues that cause excess co2 emissions.

## **Chapter 5.1 Limitations**

There are two main limitations to this study.

The first limitation is that the ARIMA model is univariate, only using the historical co2 emission levels to forecast future levels. This could potentially affect the accuracy of the model, as it does not take external factors such as a country’s GDP and population into account.

The second limitation comes from the fact that co2 emission levels are highly dependent on a number of factors, and forecasting using historical data co2 emission levels may not yield the best results. External factors that affect co2 emissions could present a problem when trying to attain as accurate of a forecast as possible, one example being a country’s policies and laws on GHG emissions. Elements like these play a large role in the amount of GHGs produced by a country, with no discernable way to factor them into our model, unlike GDP and population which can be quantified.

# **References**

C2ES (n.d.) *Global emissions*, *Center for Climate and Energy Solutions*. Available at: https://www.c2es.org/content/international-emissions/# (Accessed: November 6, 2022).

Dolphin, R. (2020) *LSTM networks | A detailed explanation | towards data science*, *Towards Data Science*. Available at: https://towardsdatascience.com/lstm-networks-a-detailed-explanation-8fae6aefc7f9 (Accessed: November 6, 2022).

Goh, T., Ang, B. W., & Xu, X. Y. (2018*). Quantifying drivers of CO2 emissions from electricity generation–Current practices and future extensions.* Applied Energy, 231, 1191-1204.

Homaira, M., & Hassan, R. (2021). Prediction of Agricultural Emissions in Malaysia Using Machine Learning Algorithms. *International Journal on Perceptive and Cognitive Computing*, *7*(1), 33-40.

IPCC. (2018). *Climate Change 2013: The Physical Science Basis*. Retrieved August 14, 2022, from https://www.ipcc.ch/report/ar5/wg1/.

Kostadinov, S. (2017). *How recurrent neural networks work*. Medium. Retrieved August 14, 2022, from https://towardsdatascience.com/learn-how-recurrent-neural-networks-work-84e975feaaf7

Lindsey, R. (2022) *Climate change: Atmospheric carbon dioxide*, *NOAA Climate.gov*. Available at: https://www.climate.gov/news-features/understanding-climate/climate-change-atmospheric-carbon-dioxide# (Accessed: November 6, 2022).

Ludwig, S. A. (2019). Comparison of time series approaches applied to greenhouse gas analysis: ANFIS, RNN, and LSTM. In *2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)* (pp. 1-6). IEEE.

Manowska, A. (2020). Using the LSTM network to forecast the demand for electricity in Poland. *Applied Sciences*, *10*(23), 8455.

Radojević, D., Pocajt, V., Popović, I., Perić-Grujić, A., & Ristić, M. (2013). Forecasting of greenhouse gas emissions in Serbia using artificial neural networks. *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects*, *35*(8), 733-740.

Ritchie, H., Roser, M., Mathieu, E., Macdonald, B. & Rosado, P. (2022). Data on CO2 and Greenhouse Gas Emissions by Our World in Data. Github. Retrieved August 14, 2022, from <https://github.com/owid/co2-data>

Siami-Namini, S., Tavakoli, N., & Namin, A. S. (2018). A comparison of ARIMA and LSTM in forecasting time series. In *2018 17th IEEE international conference on machine learning and applications (ICMLA)* (pp. 1394-1401). IEEE.

Tang, K. H. D. (2019). Climate change in Malaysia: Trends, contributors, impacts, mitigation and adaptations. *Science of the Total Environment*, *650*, 1858-1871.

UCAR (n.d.) *Some Greenhouse Gases Are Stronger than Others*, *Some Greenhouse Gases Are Stronger than Others | Center for Science Education*. Available at: https://scied.ucar.edu/learning-zone/how-climate-works/some-greenhouse-gases-are-stronger-others (Accessed: November 6, 2022).

Worldometer. (n.d.). *Alphabetical list of countries*. Worldometer. Retrieved August 14, 2022, from https://www.worldometers.info/geography/alphabetical-list-of-countries/