National College of Ireland

BSc in Computing
Data Analytics
2020/2021
Tanya Beth Rosaldo
X17145937

“The Effects of Increasing CO2 and Fossil Fuel Consumption on car companies especially electric car manufacturers”

Technical Report

Contents

Executive Summary

Electric vehicles are thought to be the future vehicle, creating a never-ending debate about the benefits and drawbacks that this future vehicle could bring. The report outlines the most important aspects of the final year project. The project will complete a data analytics project that will carry out an analysis of the Electric Vehicle (EV) Industry and the role of carbon dioxide emissions affecting the environment. The project studies whether environmental concerns are causing or influencing the growth of the electric vehicle (EV) industry market, as well as the market for gasoline-powered vehicles. This project will focus on these elements and conduct research on forecasting the growth of EVs and how this affects the growth of gasoline-powered vehicles, CO2 emissions and fossil fuel consumption. Additionally, studying the relationship between stock market and CO2 emissions levels and attempting to assess the risk for retail investors involved in the electric vehicle (EV) market.

The approach for the study will be the CRoss Industry Standard Process for Data Mining. The Tesla dataset, CO2 emissions dataset, Fossil Fuel Consumption dataset, and other datasets such as crude oil and an automotive dataset (BMW) will be integrated and transformed during study. The details, along with the findings and information, will be visualised for the end-user. To that end, the project will use a variety of tools and technology such as Python, Excel, Google Colab, Pandas and other package libraries to transform and clean the data, as well as analyse and evaluate it. The project's goals, criteria, expectations, design, results, and findings will be outlined in this study.

1.0 Introduction

1.1. Background

Electric cars were introduced a century ago, but for a many reasons, they are gaining popularity today. Electric cars were produced after the first popular electric car was made in the 1890s, and it has continued to show a strong scale over the next few years. As gasoline-powered cars were introduced, however, the number of electric cars increased and decreased over time. According to a 2018 survey, carbon dioxide accounts for 81% of greenhouse gas emissions, followed by methane (10%), nitrous oxide (7%), and fluorinated gases (3%). Most oil is used for transportation, accounting for 75 percent, 10 percent for chemical feedstock, 7 percent for industry, and 8 percent for others. The rise in CO2 emissions and oil prices has developed an interest in seeking alternative fuel vehicles, as well as more research and development in the field of electric and hybrid vehicles. With organisation introducing awareness of the environment, a new interest in electric vehicles in the U.S. was developed with better features, such as speed and efficiency far similar to gasoline-powered vehicles. Since then, electric vehicles have begun to draw more interest, and more people continue to become aware of the environment.

This project will analyse whether CO2 emissions will cause or influence the increase in demand for electric vehicles and the value of the automotive industry, as well as the impact will it bring to crude oil and fossil fuel consumption. Data on CO2 emissions and fossil fuel consumption are used because this may influence the growth of demand for electric vehicles. Furthermore, data on electric and gasoline-powered vehicles was used to determine if there is a relationship or trend between CO2 emissions or crude oil. The project will graph the changes in the growth of electric vehicles, BMW, CO2, and fossil fuel consumption by displaying and contrasting the data discovered, as well as comparing it to other periods. To do this, the project will collect a significant amount of data, which will help to ensure that the project covers as much ground as possible and uses realistic time frames to demonstrate a fair and equitable difference in the patterns of CO2 emissions, fossil fuel consumption, and other datasets through the use of models such as ARIMA, LSTM, VAR, and GARCH that will help throughout the analysis. For research purposes, it is worthwhile to consider looking at companies that market electric vehicles, as Tesla is currently the only hundred percent electric vehicle company. This will help in measuring the adoption of other car company's in EV and if this will impact oil companies.

The aim of this project is to analyse to see whether CO2 emissions and fossil fuel consumption are causing or affecting the growth of the electric vehicle (EV) industry market, as well as the market for gasoline-powered vehicles. This report will analyse the requirements which will be put in place for analysing the data and how this will be achieved. The project will describe the methodology which will be implemented in the project.

1.2. Aims

The aim of this project is to conduct an analysis, and data relevant to the project will be chosen. CO2, Tesla, fossil fuel consumption, crude oil, and BMW datasets will be collected. The collected datasets would then be analysed to identify trends and patterns, which will be obtained when the datasets are pre-processed, such as cleaning and transformation to make them more appropriate for the project scope. The goal of this project is to examine how changes in CO2 emissions, fossil fuel consumption, and crude-oil prices affect EVs and gasoline-powered vehicles, and vice versa. With the CO2 and fossil fuel consumption, how does this impact the oil industry, gasoline-powered vehicles, and EVs, as well as the company's valuation, taking the adoption period into account. The collected datasets, as well as the associated changes and patterns, will be used to analyse these areas.

Aim 1: The first aim of this project is to choose a correct and relevant dataset for the main goal. Data on CO2, electric cars, fossil fuels, crude oil, and the BMW dataset will be collected in order to conduct data analysis.

Aim 2: Once appropriate datasets have been identified, the next goal is to correctly pre-process the data so that it is suitable for further analysis and Machine Learning implementation. Cleaning and filtering the datasets for specific and appropriate time frames is part of this method.

Aim 3: After pre-processing, the selected variable from datasets is chosen, and the data has been merged together (if merging is required) after the transformation phase. The next move is to conduct exploratory data analysis (EDA) to identify associations between datasets, such as a correlation between EVs, gasoline-powered vehicles, and crude oil, amount of fossil fuel consumption, and CO2 emissions, these are few examples of EDA that will be performed.

Aim 4: After completing exploratory data analysis (EDA), the project's next goal is to programmatically prepare the data for Machine Learning Models.

Aim 5: The final goal is to record all of the study's findings and visually represent the results once the objectives have been met. Evaluate the collected data, analyse, and draw a conclusion.

1.3. Technology

There are multiple technologies used for the development of this project that are acceptable and feasible and these include:

Python: Python is a programming language and has a basic syntax for the management of big data. While R is more dedicated to statistical analysis, because it has open tools but when it comes to combining statistics, in one perspective, Python is the best choice on this project because it is well suited for deploying machine learning and compatible with deep learning and machine learning libraries, all of which will be extremely useful for this project. Python was used to program machine learning models as well as to pre-process, transform data and statistical analysis.

Pandas: Along with Python, Pandas has been used, which is a Python Data Analysis Library and is used for anything from importing data from Excel spreadsheets to analysing time series data and more. In addition, with Panda data frames, cleaning and manipulation can also be achieved.

Other Python Packages: these packages are used for visualisation, high level mathematical functions, explore data, estimate statistical models, and perform statistical tests that are beneficial for the analysis and when creating models.

Google Colab: is a type of notebook and a Google Research product. It allows anyone to write and execute Python code via a browser. Instead of using a local machine, this is intended to use when coding because running Python scripts always takes a lot of computer power which can take time if personal computer is slow.

Excel: The data is stored in a comma separate value (.csv format) file. The project will use Microsoft Excel to manage the data. Microsoft Excel is a spreadsheet program that can be used to visualise data and run calculations.

1.4. Structure

This Data Analytics Report includes an executive summary that describes the specifics of the project and the goals, followed by introduction that discusses the context history or useful information of the project, proposed technologies that would be useful for the advancement to the project. Data and methodology are a significant part of the project, as it describes what datasets are used, how they are sourced and their types; the methodology part reflects research framework and is a tool used to classify, select, process and analyse the information collected about the subject followed by analysis after methodology is completed. This section analyses how data is pre-processed, cleaned and transformed and modelled. Testing explains how the analysis went and if the same results are being achieved after testing multiple times and to know if there are any performance issues that needs to deal with. Results explains the evaluation, visualisation, results from the analysis and if goals/aims are met. Lastly, is the conclusion and further progress and research that provides an understanding of the research. The issues faced and the project's shortcomings and what aspects of the research and findings can be accomplished in the immediate future. Finally, the appendices are the main body of the study containing additional details.

1.5. Definitions, Acronyms and Abbreviations

CRISP-DM: CRoss Industry Standard Process for Data Mining
EV: Electric Vehicle/s
CO2: Carbon Dioxide
ARIMA: Auto Regressive Integrated Moving Average
Time Series/Time Series Analysis: a statistical technique for dealing with time series results, also known as trend analysis. Time series data is data that is organised in a series of specific time periods or intervals.
LSTM: Long Short-Term Memory
GARCH: Generalised AutoRegressive Conditional Heteroskedasticity
VAR: Vector AutoRegression
Google Colab: allows you to write and run Python code in your browser and enable you to include executable code, rich text, images, HTML, LaTeX, and other elements in a single document
Python: Apart from web development, it is a general-purpose scripting language that can be used for other forms of programming and software development. This includes, among other aspects, back end creation, software development, data science, and writing machine scripts.
Tableau: In the Business Intelligence Industry, a data visualisation tool is used. It aids in the simplification of raw data into a more readable format.
CSV: Comma Separated Values
EDA: Exploratory Data Analysis
Machine Learning: a Data Analysis approach that allows computers to find insights and trends in data by automating analytical model building and employing algorithms.
BMW: Bayerische Motoren Werke GmbH is an abbreviation for the Bavarian Engine Works Company.
Tesla: a Palo Alto, California-based electric vehicle and renewable energy business.
Kaggle: An open source data collection website that provides free datasets.
API: Application Program Interface

2.0 Data

Multiple datasets are used in this analysis, including carbon dioxide emissions extracted from GitHub, fossil fuel consumption extracted from Our World in Data, crude oil extracted as a .csv file from Kaggle, and Tesla and BMW car companies extracted from the Yahoo Finance API. The project's use of stock prices is advantageous for retail investors who want to invest or are involved in the EV market because it assesses risk. Rising fossil fuel consumption and CO2 emissions concern people about climate change, global warming and other environmental issues. These datasets are important as it can show how CO2 emissions and fossil fuels have changed over the last few years, as well as the effects of these elements on the growth of car companies and whether or not this will or has affected the oil industry.

Table 1

Dataset	CO2 emissions	Crude oil	Fossil fuel consumption	Tesla	BMW
File Format	*CSV*	*CSV*	*CSV*	*Extracted from API*	*Extracted from API*
Number of attributes	*38*	*7*	*6*	*7*	*7*
No. of Records	*24016*	*6244*	*5190*	*2624*	*6170*
Size	*3.4 MB*	*395 KB*	*3 KB*	*290 KB*	*660 KB*

These open datasets are extracted from the site and downloaded as CSV files, as seen on Table 1 CO2 emissions is the biggest dataset since this will be the core of the project followed by fossil fuels and crude oil. Read_csv is used to load data using Google Drive, in this case Pydrive is used that simplifies common Google Drive API tasks, followed by authentication to allow Google to access Google Drive. Both TSLA and BMW are taken from Yahoo Finance using a Python pandas package called data reader that allows to create and read data frame from various data sources. The required libraries are imported such as numpy, matplotlib, seaborn, pandas that are need during this study. Line charts are used using matplotlib and seaborn library for heat maps as visualisations to analyse if there are any correlation and pattern/trend between these datasets (BMW, TSLA and crude oil; CO2 emissions and Fossil Fuel consumptions). In table 2 shows all dataset used displaying total numbers of rows and columns using .shape(). As observed column differs in table 1 and table 2, the reason is dates are not considered as a column when ran in Colab in this case, resetindex() is used and it is only mandatory to change columns into their correct data types since this will be useful for further analysis most especially when dealing with timeseries datasets.

Tesladf.shape	***Bmwdf.shape***	***Crudeoildf.shape***	***Fossildf.shape***	***Carbondf.shape***
2604,6	6107,6	6244,7	5190,6	24016,38

Table 2

In addition, .head() function is used that returns the dataset's first five observations. Similarly, .tail() command returns the last five dataset observations. Finding whether it includes null or missing values, the .info() is also used to understand the columns and their data types. If null values are found .dropna() is used and this is applied in some dataset used also, attributes and number of records are also removed and renamed on few datasets and this will also be discussed on methodology section.

3.0 Methodology

The project follows the Cross-InduStry Process for Data Mining (CRISP-DM) approach on this final project for its usefulness in arriving at a conclusion and versatility provided by being able to generalise the findings to other areas.

A. Business Understanding

This section summarises the research's purpose and objectives and aids in the creation of a technical flow for the analysis of analysing the impact of CO2 emissions and fossil fuel consumption on the growth of car companies and how can stock be used in evaluating the risk for retail investors. This phase will be able to take a model's core features and break it down into granular levels for easy comprehension along the way. In figure 1 shows the different sections of the methodology and understanding these steps is beneficial as this will guide in the completeness of the project.

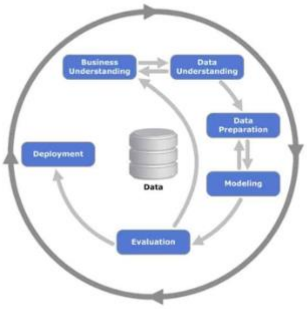


Figure 1 CRISP-DM Model Structure

B. Data Understanding

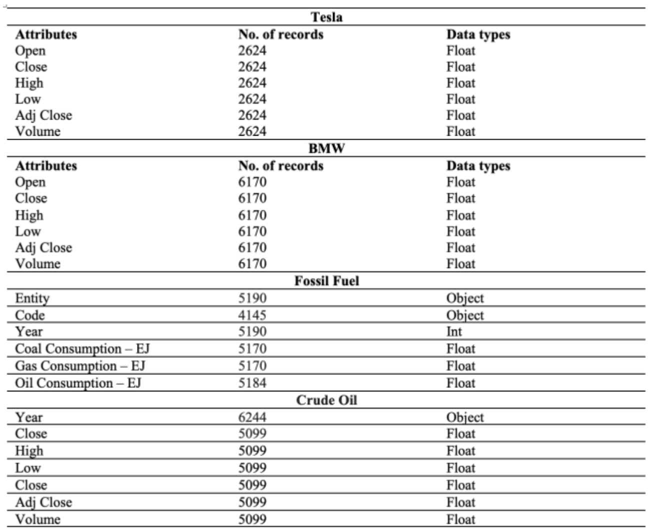
The next phase is to build a theory in order to gain a more in-depth understanding of what needs to be achieved and to perform data analysis. This entails comprehending the questions and metrics in the dataset. The data set collected is intended to obtain a better understanding of the data and to identify the problems. With the findings and conclusions this will help to explain whether the assumption is correct. Understanding units and general terms, like EJ, Tonnes, GDP, stock terms like open, close and so on, are important to learn and understand for this project. Looking at the raw data, it was discovered that the crude-oil dataset contains missing values, which must be removed because it can affect the accuracy of the result. There are also columns in fossil fuels and CO2 emissions that need to be renamed and removed; this will be addressed more during the data preparation phase. As previously stated, visualisations and modelling are needed to ensure that this results in more accurate results/outcomes. Visualisation of the datasets collected is required to gain a clearer understanding of the data through the use of heatmaps and line charts, as well as the use of Tableau to best represent the data.

C. Data Preparation

The data must be prepared for modelling based on the understanding gained in the previous phase. As there is a clear understanding of the project's aims, data collection is carried out to determine which datasets will be used and where these datasets will be extracted. Before data preparation, Tables 3 and 4 included a summary of all datasets. At this stage, feature engineering is used to assess the feasibility of the features, as well as to remove unnecessary and obsolete features in order to reduce data noise. The datasets are examined for issues such as missing values and to determine what data types each dataset contains. Since datasets are stock prices, it has the same attributes, however the number of records in fossil fuel, and CO2 datasets varies.

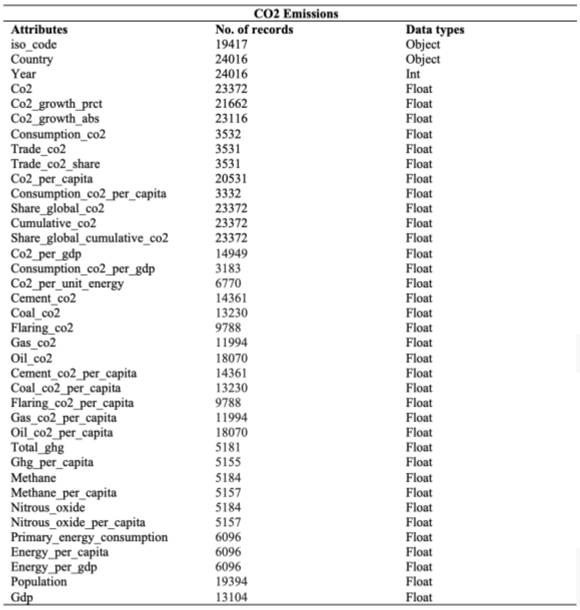
Prior to modelling, pre-processing and data transformation are also performed. This includes changing the data type by changing the basic form of features of interest. As seen in Figure 2, there are null values found by using a function called .isna() in the crude oil dataset, so these are removed. Dropping/removing null values may result in insufficient dataset and may impact results; however, after analysing the crude oil dataset, the presumption is that the stock market is closed on holidays and weekends. The description of the dataset after it has been transformed is shown in Table 5, this shows the changed data types and as mentioned removing null values. Furthermore, data types are modified, which is why the Date attribute can be seen in Tesla and BMW, totalling 7 attributes as seen in Table 1. This is also useful for visualisation purposes, as date/year is relevant when conducting line graphs. BMW dataset dates from 1996 to May 2020 are also excluded. The reason for the removal of crude oil and BMW dates is that the relationship between Tesla, BMW, and crude oil is being analysed, and in this case, Tesla has only began trading in 2010 hence why these dates has been removed. Figures 3 and 4 demonstrate how the fossil fuel and CO2 datasets were renamed and unnecessary columns were removed because only valid attributes are needed for a country chosen which is the United States.

Table 3 Before pre-processing

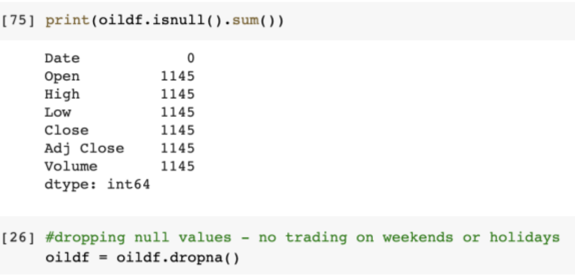


	Tesla	
Attributes	**No. of records**	**Data types**
Open	2624	Float
Close	2624	Float
High	2624	Float
Low	2624	Float
Adj Close	2624	Float
Volume	2624	Float
	BMW	
Attributes	**No. of records**	**Data types**
Open	6170	Float
Close	6170	Float
High	6170	Float
Low	6170	Float
Adj Close	6170	Float
Volume	6170	Float
	Fossil Fuel	
Entity	5190	Object
Code	4145	Object
Year	5190	Int
Coal Consumption – EJ	5170	Float
Gas Consumption – EJ	5170	Float
Oil Consumption – EJ	5184	Float
	Crude Oil	
Year	6244	Object
Close	5099	Float
High	5099	Float
Low	5099	Float
Close	5099	Float
Adj Close	5099	Float
Volume	5099	Float

Table 4 Before Pre-processing



CO2 Emissions		
Attributes	**No. of records**	**Data types**
iso_code	19417	Object
Country	24016	Object
Year	24016	Int
Co2	23372	Float
Co2_growth_prct	21662	Float
Co2_growth_abs	23116	Float
Consumption_co2	3532	Float
Trade_co2	3531	Float
Trade_co2_share	3531	Float
Co2_per_capita	20531	Float
Consumption_co2_per_capita	3332	Float
Share_global_co2	23372	Float
Cumulative_co2	23372	Float
Share_global_cumulative_co2	23372	Float
Co2_per_gdp	14949	Float
Consumption_co2_per_gdp	3183	Float
Co2_per_unit_energy	6770	Float
Cement_co2	14361	Float
Coal_co2	13230	Float
Flaring_co2	9788	Float
Gas_co2	11994	Float
Oil_co2	18070	Float
Cement_co2_per_capita	14361	Float
Coal_co2_per_capita	13230	Float
Flaring_co2_per_capita	9788	Float
Gas_co2_per_capita	11994	Float
Oil_co2_per_capita	18070	Float
Total_ghg	5181	Float
Ghg_per_capita	5155	Float
Methane	5184	Float
Methane_per_capita	5157	Float
Nitrous_oxide	5184	Float
Nitrous_oxide_per_capita	5157	Float
Primary_energy_consumption	6096	Float
Energy_per_capita	6096	Float
Energy_per_gdp	6096	Float
Population	19394	Float
Gdp	13104	Float

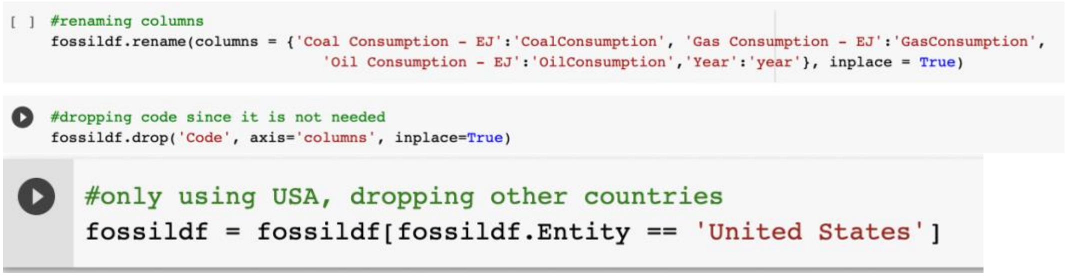


```
[75] print(oildf.isnull().sum())

     Date           0
     Open        1145
     High        1145
     Low         1145
     Close       1145
     Adj Close   1145
     Volume      1145
     dtype: int64

[26] #dropping null values - no trading on weekends or holidays
     oildf = oildf.dropna()
```

Figure 2 Checking for null values

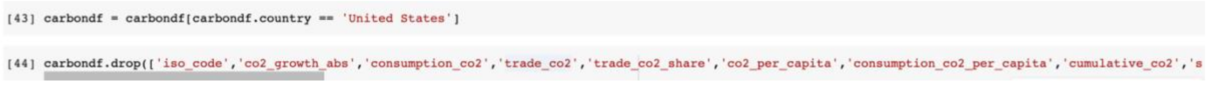


```
[ ]  #renaming columns
     fossildf.rename(columns = {'Coal Consumption - EJ':'CoalConsumption', 'Gas Consumption - EJ':'GasConsumption',
                               'Oil Consumption - EJ':'OilConsumption','Year':'year'}, inplace = True)

     #dropping code since it is not needed
     fossildf.drop('Code', axis='columns', inplace=True)

     #only using USA, dropping other countries
     fossildf = fossildf[fossildf.Entity == 'United States']
```

Figure 3 Renaming attributes and dropping columns - Fossil Fuel Dataset

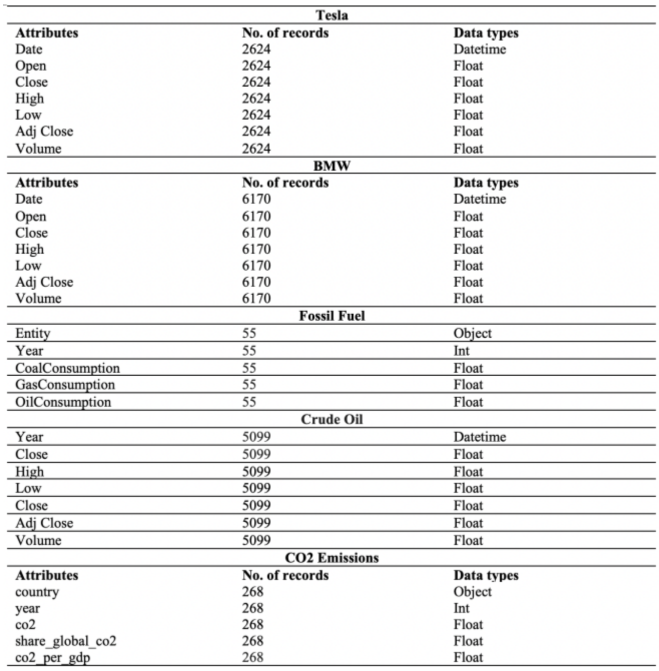


```
[43] carbondf = carbondf[carbondf.country == 'United States']

[44] carbondf.drop(['iso_code','co2_growth_abs','consumption_co2','trade_co2','trade_co2_share','co2_per_capita','consumption_co2_per_capita','cumulative_co2','s
```

Figure 4 Dropping unnecessary columns - CO2 dataset

Table 5 After Pre-processing



	Tesla	
Attributes	**No. of records**	**Data types**
Date	2624	Datetime
Open	2624	Float
Close	2624	Float
High	2624	Float
Low	2624	Float
Adj Close	2624	Float
Volume	2624	Float
	BMW	
Attributes	**No. of records**	**Data types**
Date	6170	Datetime
Open	6170	Float
Close	6170	Float
High	6170	Float
Low	6170	Float
Adj Close	6170	Float
Volume	6170	Float
	Fossil Fuel	
Entity	55	Object
Year	55	Int
CoalConsumption	55	Float
GasConsumption	55	Float
OilConsumption	55	Float
	Crude Oil	
Year	5099	Datetime
Close	5099	Float
High	5099	Float
Low	5099	Float
Close	5099	Float
Adj Close	5099	Float
Volume	5099	Float
	CO2 Emissions	
Attributes	**No. of records**	**Data types**
country	268	Object
year	268	Int
co2	268	Float
share_global_co2	268	Float
co2_per_gdp	268	Float

The CO2 dataset was also converted using the def covert function since it was originally in Tonnes and the fossil fuel dataset is originally in EJ. This is done to ensure that all units are the same and that no graphs are affected. Once conversion was complete, the CO2 converted dataset and the fossil fuel dataset are plotted to see if there are any patterns; however, the CO2 converted dataset has lower values than the fossil fuel dataset. The results of combining fossil fuel and the unconverted CO2 were comparable to the results of combining converted CO2 and fossil fuel since it has quite similar results. There are also issues found with plotting CO2 with BMW or TSLA and the same thing for fossil fuel with TSLA or BMW or even when trying to combine the dataset, the reason of doing this is to visualise if there are any interesting trends/patterns and if there are any correlations. The reason for this to be unsuccessful is there are not similar attributes to start with. As a result, CO2 and fossil fuels are the only datasets merged, the reason behind merging these datasets is to show the relationship between fossil consumption and carbon dioxide emissions and to see if there's a pattern. Since there are errors encountered with visualisation the graphs are then analysed separately instead of having these datasets combined and plotted. Merge was also used on TSLA, BMW, and crude oil using a heatmaps and line graphs to see whether there was any association between these datasets, and the same was done on fossil fuel datasets (oil, coal, and gas consumption) and CO2 datasets (co2, co2 per gdp, and share global co2).

D. Modelling

The modelling technique is used after the data has been prepared and pre-processed. This involves choosing a modelling approach that is appropriate for the project's goals. This involves training the data algorithm – to quantify a range of parameters, such as training error and predictive accuracy, to help the model learn and make more general predictions. This is needed to provide a more detailed analysis of the dataset as well as to compare and contrast which models are better suited for this analysis. The goal of this project is to determine whether CO2 emissions and fossil fuel consumption have an effect on electric vehicles, gasoline-powered cars, and the oil industry, as well as evaluate the risk of retail investors who are interested on EV market, so the Exploratory Data Analysis (EDA) method is used, and models are used to forecast growth, future values and stock volatility.

ARIMA (Auto-Regressive Integrated Moving Average), LSTM (Long Short-Term Memory), and VAR (Vector AutoRegression) models are used for time series analysis/forecasting, while GARCH (Generated Autoregressive Condition Heteroskedasticity) is a volatility model for stocks. The ARIMA model is applied to time series data in order to analyse or forecast future data points on a time scale. The LSTM model can learn a function that maps a sequence of previous observations as input to an output observation, while VAR is a forecasting algorithm that can be used when two or more time series influence each other. These three models are used for forecasting future values and growth. Finally, the GARCH model is a method for estimating volatility. LSTM gains the ability to predict and uses the K2 test to either reject or accept the hypothesis in the same way that ARIMA uses ADF (Augmented Dickey-Fuller Test), similarly with, VAR using Granger's Causality Test, ADF, and Durbin Watson, and the GARCH model uses PACF to find the best order that fits the model. All of these models were used to predict/forecast the growth of automobile manufacturers, as well as the growth of CO2 emissions and fossil fuel consumption. Furthermore, the models are intended to compare the performance and results of each model, as well as how they vary from one another.

E. Evaluation

After training the model, the final stage of CRISP-DM analyses the results from the project's implementation stages. Its aim is to assess efficiency and accuracy in order to ensure the objectives are met. This is accomplished by visually exploring the data and using visualisation techniques such as python libraries and Tableau.

3.1 Research

For the project, analysis is required, such as knowledge of the data obtained and stock knowledge, as this is required to understand how a business can be calculated using the Closing stock price. Exploring Tesla and BMW companies is also advantageous because it can provide insight into how well the companies have performed, and how CO2 and fossil fuels has evolved in the past few years, as well as understanding different machine learning models that were to be applied within the project and looking into how it can be incorporated to produce the best results possible. Investigating related works was also useful in deciding which models should be used. When the CO2 and fossil fuel datasets were extracted, one thing that was discovered was the different units since CO2 is in Tonnes and Fossil Fuel is in EJ; hence, research on how to convert Tonnes to EJ using Python was conducted.

4.0 Analysis

This section describes the analyses performed during the project's progress when applying the stages of the CRISP-DM Methodology. The initial stages of implementation included data selection, business and data understanding, and data pre-processing. The acquisition of important and useful datasets, as well as thorough analysis, took a considerable amount of time. To prepare for analysis and modeling, null values, column renaming, and attribute removal are performed. For data analysis/preprocessing, ARIMA, LSTM, VAR, and GARCH are used, and the same techniques are used for extracted data sets. Imported libraries are used to model and read data. Datareader is imported, which generates a data frame from internet resources, in this case, Yahoo Finance for Tesla and BMW, and read csv, which reads CSV files containing data extracted from Kaggle and World of Data. As datasets must be cleaned and transformed, pre-processing such as dropna is implemented to eliminate null values. Columns are also renamed using the rename function to avoid long column names and for feasibility purposes, and columns are omitted using the drop function because they are unimportant to the project's purpose.

4.1 Library Used

- Pandas
- Matplotlib
- Seaborn
- Numpy
- Datareader
- Keras
- Kearas-models
- Keras-layers
- matplotlib-inline
- pandas-datareader
- python-dateutil
- scipy

- sklearn
- sklearn-pandas
- statsmodels
- tensorflow
- tqdm
- google-colab
- oauth2client
- itertools
- math
- pmdarima

4.2 Functions Used

- to_csv()
- pct_change()
- dropna()
- arch_model()
- fit()
- forecast()
- filter()
- minMaxScaler
- def()
- reshape()
- sequential()
- astype()
- drop()
- rename()
- invert_transformation()
- fillna()

4.3 Initial Analysis

The stock datasets are the first to be analysed (TSLA, BMW, and crude oil). The dataset does not require much pre-processing because it is mostly clean data; however, in the case of crude oil, there are missing values; hence, these missing values are removed. Additionally, because date/year is needed for plotting, the index type is reset, the Date attribute is then set to DateTime. The data collected in TSLA from Yahoo Finance dates back from 2010 to 2020, while BMW is 1996 to 2020 and crude-oil is 2000 to 2020. Since these datasets are in different time frames when it began its stock exchange, it must have the same timeframe to be analysed, so other years and months was dropped to accommodate TSLA starting and ending dates.

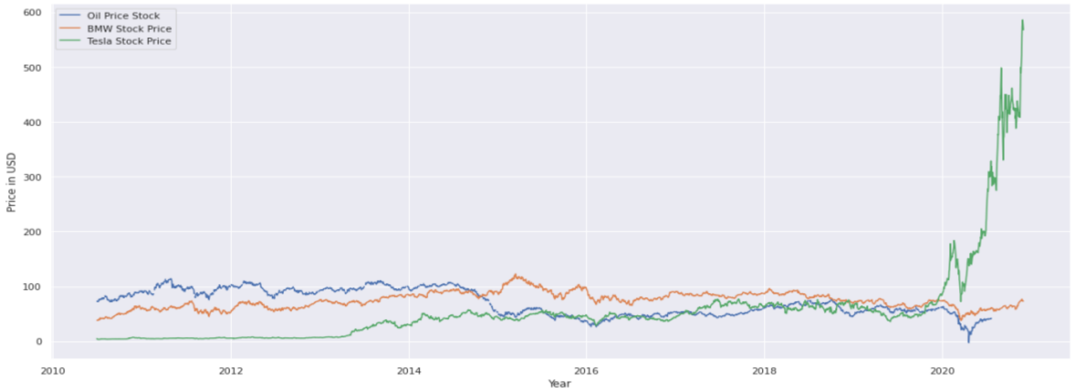


Figure 5 Correlation between Closing stock and three datasets (TSLA, BMW and Crude Oil)

Figure 5 illustrates the relation between the three datasets on the Closing stock, since this is where the stock's performance over time is determined. BMW and crude oil have a similar pattern from 2010 to the end of 2013, and both have better stock performance. In 2015, the price of crude oil dropped while the price of BMW rose, leaving Tesla and oil performance almost identical. Tesla had a negative trend on crude oil in 2017 and almost all stock exchanges are the same in 2019, followed by a major drop in the stock market in the first month of 2020 where three datasets stock close value had dropped, the cause may be the Covid-19 outbreak.

The data from Our World in Data (Fossil Fuel) and GitHub are the second datasets analysed (CO2). A lot of pre-processing time was spent on these data because they contain redundant columns, so they were removed. Additionally, the data includes other countries, and because the focus is on the United States, these countries were removed, as were the years before 1820 since CO2 per GDP only has a record during that period.

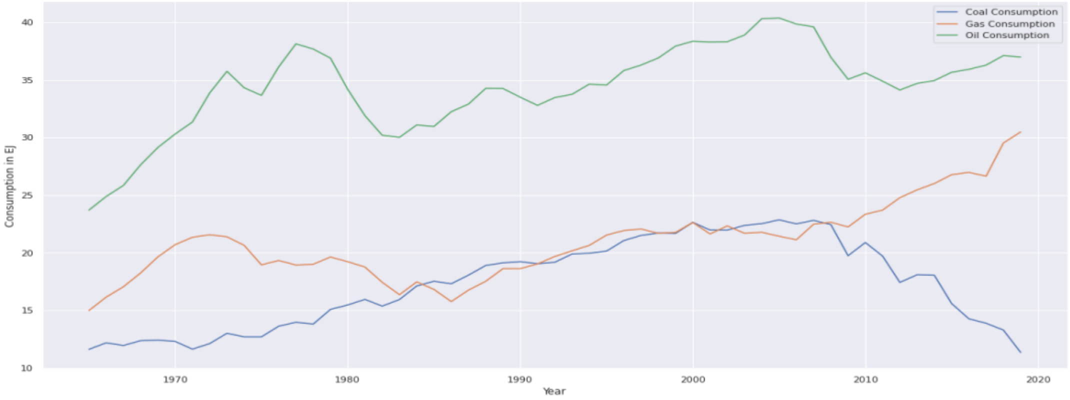


Figure 6 U.S.A Fossil Fuel Consumption

Fossil fuels are critical to global resources. From 1965 to 2019, coal ranged(blue) from 10 to 23 EJ. From 1965 to 2019, gas ranged(orange) from 15 to 30 EJ, while oil ranged(green) from 23 to 36 EJ. Coal, Oil, and Gas consumption has evolved over time, with coal and gas remaining low and oil remaining the most consumed fossil fuel. Coal and oil had the same pattern from the end of 1999 to the beginning of 2000, and then had an opposite trend around 2005. From 2011 to 2019, the demand for coal remained the lowest. Gas, on the other hand, increased in 2010 at a time when coal use is declining. Oil consumption is seen as strong, with a rise of 40 EJ in 2005 and continued to be on demand of about 30 to 35 EJ from 2010 to 2019.

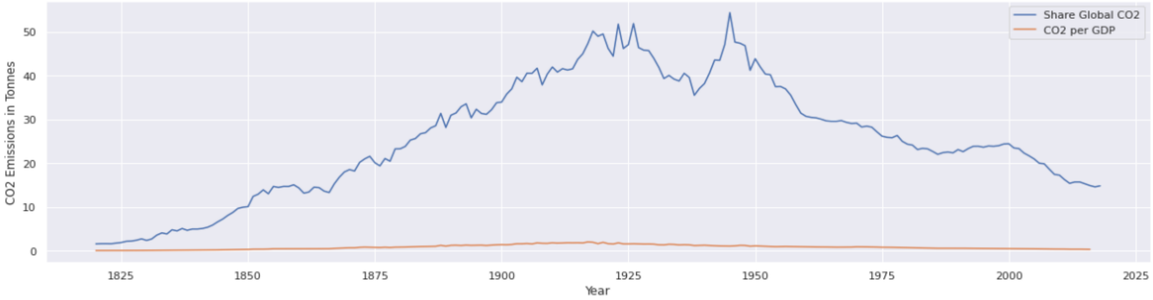


Figure 7 Share Global and CO2 per GDP

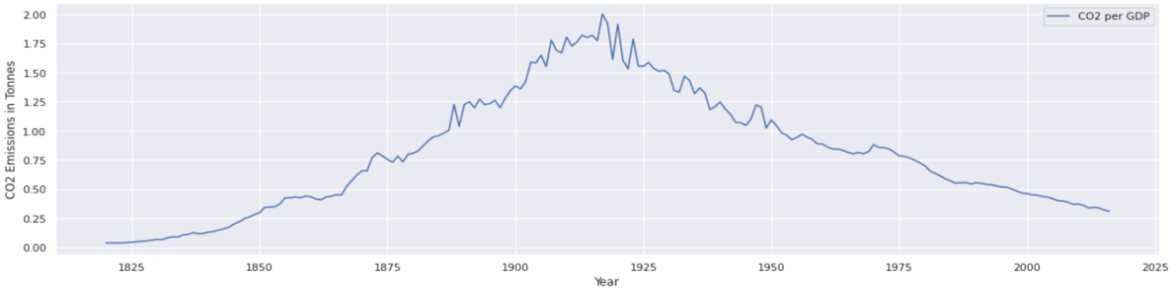


Figure 8 CO2 per GDP

Figure 7 shows a rise in global emissions from the 1830s to 2018. Global emissions peaked in 1948/49 and have steadily declined since then, with a steady decrease from 2000 to 2018. On the other hand, the highest trend in CO2 per GDP is 1917, and it has continued to fall since then, particularly in 1975, and has fallen even further in the 2000s and beyond, as shown in Figure 8. CO2 per GDP is graphed separately to gain a better understanding of the trend since it has smaller values than share global, which makes it difficult to interpret CO2 per GDP as seen in Figure 7. Figure 9 shows the rise of CO2 emissions in the United States from the graph; emissions were very low prior to the Industrial Revolution, which mostly began in the 1900s, and have risen over time, as seen in 1960/65, and have increased further to 2000 Tonnes in the year 2000, before dropping again. CO2 was graphed separately for the same purpose as CO2 per GDP and share global, as graphing the three attributes together would be difficult to analyse, as seen in Figure 10.

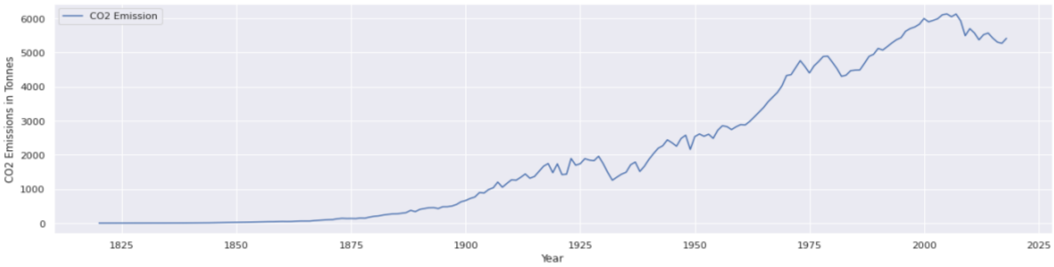


Figure 9 CO2 emissions

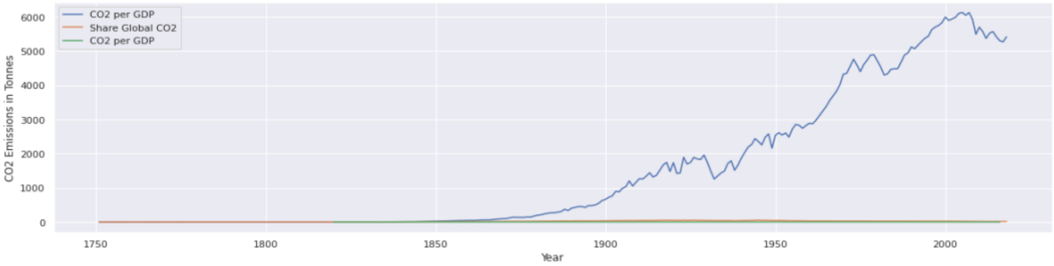
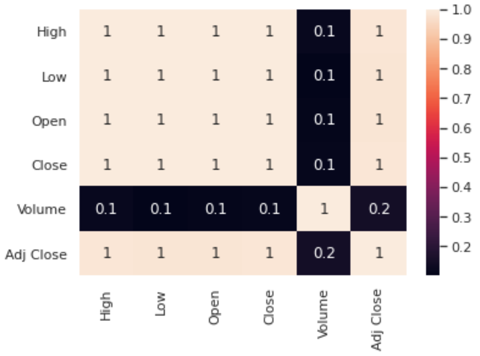


Figure 10 CO2, Share Global CO2 and CO2 per GDP

4.4 Merging Datasets

Merging the datasets was also done to find a correlation between the datasets; while graphical representation is good for visualising, it does not display how highly correlated they are; therefore, a heatmap is used. Figure 11 shows the relation between the stock datasets, and the results indicate that there is a close relationship for all of the attributes except Volume, but closing is the main focus since it shows how the company has performed. Figure 12 shows the correlation between the fossil fuel dataset; there is a clear correlation between the attributes year, gas, coal, and oil, with coal and gas showing the weakest correlation. Finally, figure 13 shows a strong correlation between co2 and year, gdp and global share, and a weak/negative correlation between co2 and co2 gdp.



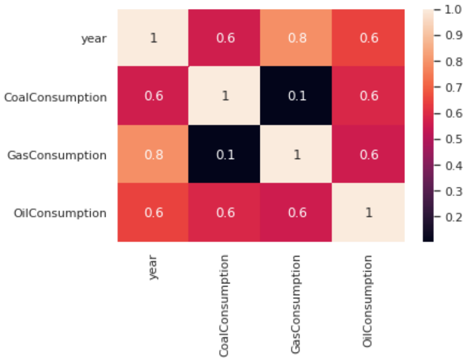
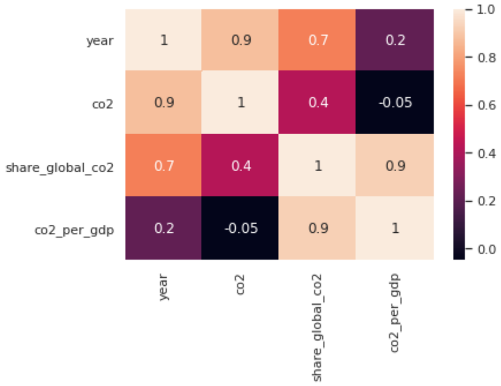


Figure 11 TSLA, BMW and Crude Oil

Figure 12 Oil Consumption, Gas Consumption and Coal Consumption



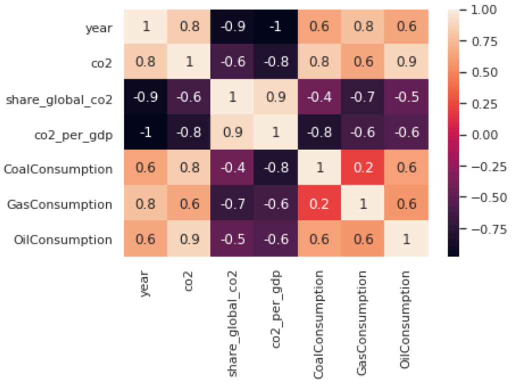


Figure 13 Share Global, CO2 per GDP and CO2

Figure 14 Converted CO2 and Fossil Fuel

CO2 must be converted into EJ since it was initially in Tonnes in order to find a link between CO2 and fossil fuel dataset. Figure 14 shows strong correlations between year and coal, oil and gas consumption, and co2 but a weak correlation between coal and gas consumption and a negative correlation with share global and co2 gdp with all the attributes but a strong correlation with one another (co2 gdp and share global). A graph was created by combining the datasets (CO2 and fossil fuel) for visualisation purposes; however, these were graphed separately because the converted CO2 dataset has lower values when converted into EJ and is difficult to interpret when graphed as a whole merge dataset. Figure 15 illustrates the unconverted CO2 and fossil fuel since a similar result was obtained (converted CO2 and fossil fuel), and the graph portrays the trend of the amount of energy from fossil fuels consumed per year in the US, as well as CO2 emissions, which have also been increasing and decreased in year 2010 onwards and the share global and CO2 per GDP has decreased further decreased over time compared to other periods in the past..

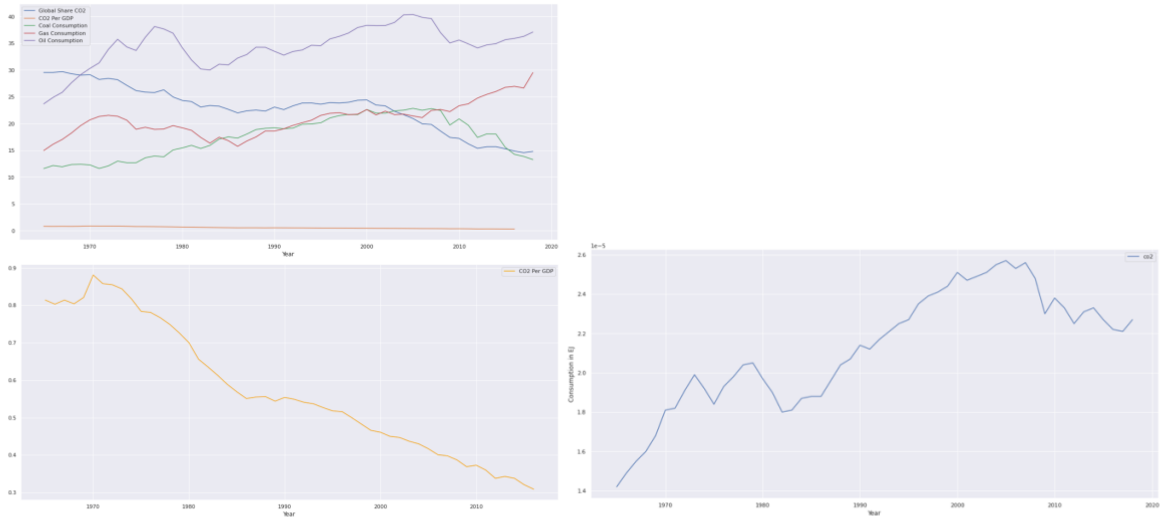


Figure 15 Merge dataset (CO2 and Fossil Fuel)

4.5 Machine Learning

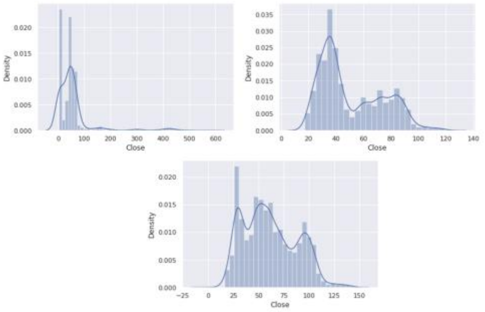
After all of the data has been collected, pre-processed, and explored, it can be incorporated into various machine learning models for the project. The stages leading up to the models being implemented are a critical part of the project which necessitated a significant amount of work in order for the models to be performed. In the case of CRISP-DM, the data preparation process was critical in order to transform the data so that different machine learning models could be applied. ARIMA, LSTM, VAR, and GARCH models were used in the project.

4.5.1 Implementing ARIMA

ARIMA was the first model to be introduced. This model is used to forecast the value of TSLA, BMW, and crude oil, as well as the increase in CO2 emissions and fossil fuel consumption. A rolling statistic is used in an ARIMA model to distinguish mean and standard deviation differences, and an additive model is used to provide interpretable patterns. Moving average, weighted average, and exponential smoothing was also computed. Statsmodel is imported, which performs rolling statistics using TSA (time series analysis). Decomposition is used to determine if seasonality exists in the data, followed by the Augmented Dickey-Fuller Test (ADF) to test the null hypothesis and the stationary of the data; stationary data is best suited for data analysis; if data is non-stationary, it is further differenced until it is stationary; non-stationary data leads to inaccurate and poor forecasting. The Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) are introduced by importing the plot_pacf and plot_acf libraries, whose functionality is to check if similar fluctuations in the lagged observation confirm the correlation, allowing the order of the ARIMA to be determined. Pmd autoARIMA, on the other hand, is a library that is installed and used to determine the order instead of PACF/ACF, as it selects the right model for ARIMA with the best AIC score. It is recommended that the dataset be divided into two parts: training and testing. The explanation for this is that when the dataset is divided into train and test sets, the training dataset would not contain enough data for the model to learn an accurate mapping of inputs to outputs. There will also be insufficient data in the test set to assess the model's success effectively. The model is then trained using the order of ARIMA to forecast predictions based on the test, and the model is retrained using the entire dataset to forecast future values.

4.5.2 Implementing LSTM

Similarly, LSTM is a time series analysis with implementations similar to ARIMA. In this case, the statistical normality test is performed in LSTM; it is the D'Agostino K2 test, which is similar to the Dickey-Fuller test in ARIMA and either rejects or supports the null hypothesis. For this test, SciPy is used instead of the stats model used in ARIMA. Kurtosis, Skewness, Histogram, and P-P Plot are also determined to decide if the data distribution deviates from the normal distribution and are represented as graphs for visualisation purposes. Filtering was used to evaluate only required columns, Numpy is used to convert a new data frame created with a value array, scaling, and MinMax was used to transform and train the model. Since the arrays are two-dimensional and three-dimensional is needed for LSTM, reshaping was also performed. The optimiser was used to improve the loss function as well as to assess how well the model has been conditioned. These measures are carried out again for the test model. The data is validated by re-creating a new array and using the same implementation, and plotting shows the eligible data, estimate, and validation in a graphic visualisation for better comprehension. A new data frame is also created to forecast a stock price in one day by transforming the test data into a numpy array, reshaping the data, obtaining the expected scaled price, and undoing scaling. The data reader function is used to equate the forecasted price for one day to the actual price.



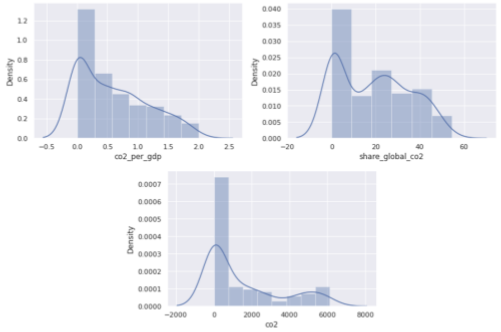


Figure 16 TSLA(left), BMW(right), Crude oil(bottom) *Figure 17 CO2 GDP(left), share global(right), CO2(bottom)*

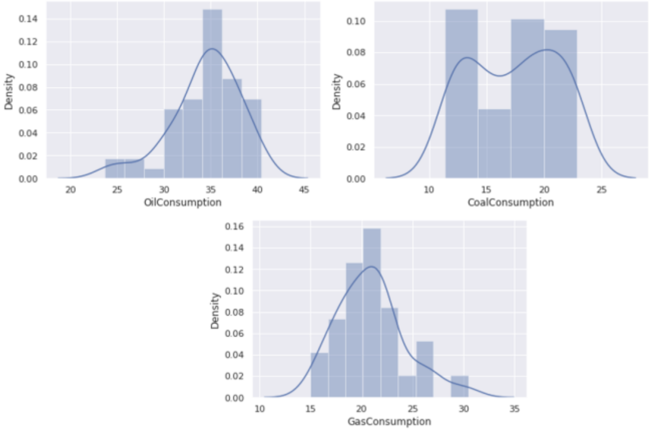
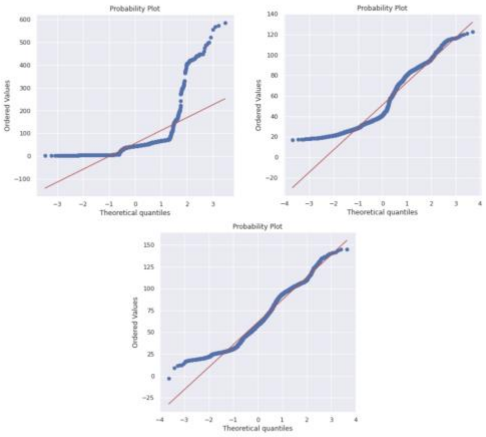


Figure 18 Oil Consumption(left), Coal Consumption(right), Gas Consumption(bottom)

Figures 16, 17, and 18 display the dataset’s kurtosis and skewness, with TSLA and Oil Consumption having heavier tails and BMW, Crude oil, CO2 attributes, Coal, and Gas Consumption having lighter tails. However, TSLA, CO2, and CO2 per GDP are extremely skewed, while BMW, Crude oil, Share global, oil and gas consumption are moderately skewed, and coal consumption is fairly skewed.



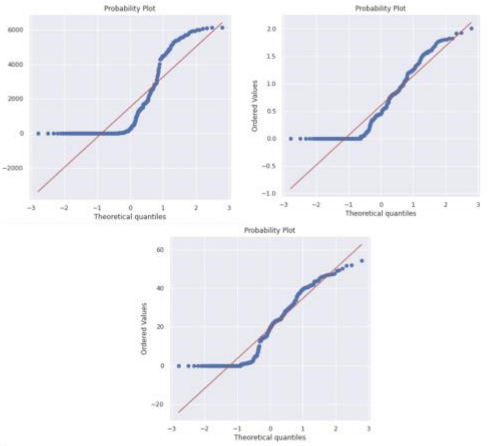


Figure 19 TSLA (left), BMW(right), crude oil (bottom) *Figure 20 CO2(left), CO2 per GDP (right) and share global(bottom)*

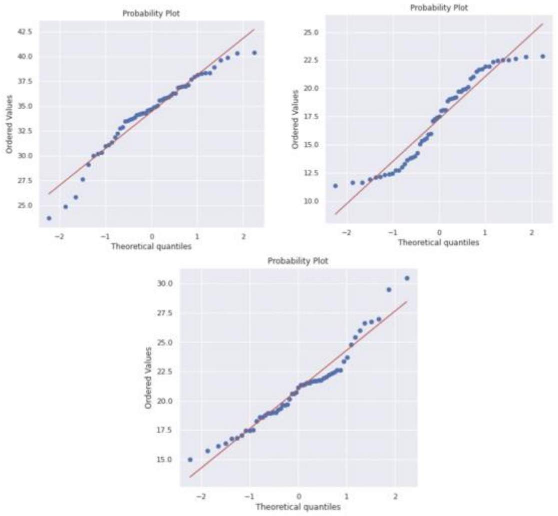
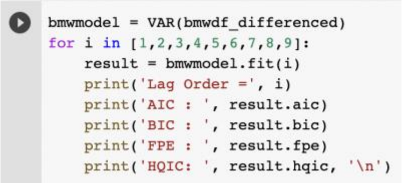


Figure 21 Oil(left), Coal(right), gas(bottom)

Figures 19, 20, and 21 illustrate the data distribution, and as can be seen, the BMW, crude oil, and fossil fuel datasets are not normally distributed, but also fall more on the line when compared to the TSLA and CO2 datasets.

4.5.3 Implementing VAR

It is a forecasting algorithm that can be used when two or more time-series influence each other. The Granger Causality Test is used to test the null hypothesis that the regression coefficients of past values are zero. The dataset is then divided into training and test results, after which the model is fitted on df_train and used to forecast the next four observations. Statsmodels is used to train and forecast the VAR model. Similarly, ADF is used to check for stationarity in ARIMA, and differencing is used to ensure that $p < 0.05$. To choose the best order, we iteratively match the model's increasing order and choose the order that produces the model with the lowest AIC value, as shown in Figure 22. The model is then trained using the model's chosen order. Durbin Watson is used to see if there are any remaining patterns in the residuals, ensuring that the model can adequately describe the variances and patterns in the time series. Invert transformation is used to return the forecast to its original scale in order to obtain the true forecast; plotting is then used to map the forecast and the actual values.



```
bmwmodel = VAR(bmwdf_differenced)
for i in [1,2,3,4,5,6,7,8,9]:
    result = bmwmodel.fit(i)
    print('Lag Order =', i)
    print('AIC : ', result.aic)
    print('BIC : ', result.bic)
    print('FPE : ', result.fpe)
    print('HQIC: ', result.hqic, '\n')
```

Figure 22 Choosing the order of model

4.5.4 Implementing GARCH

This model is used to predict a stock’s volatility, and it is primarily used for TSLA and BMW. The arch model is installed using the !pip function, allowing the GARCH process to run. Pct_change is used to calculate the percentage change in stock form the start to the end date.

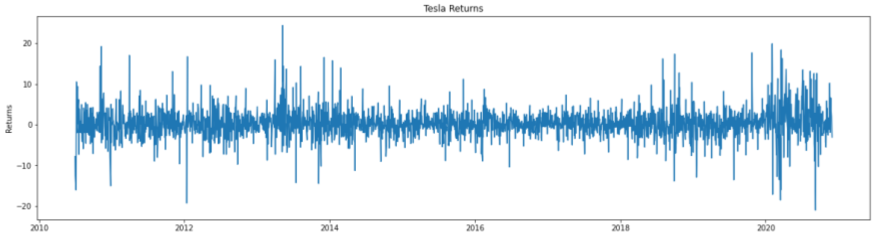


Figure 23 TSLA Volatility

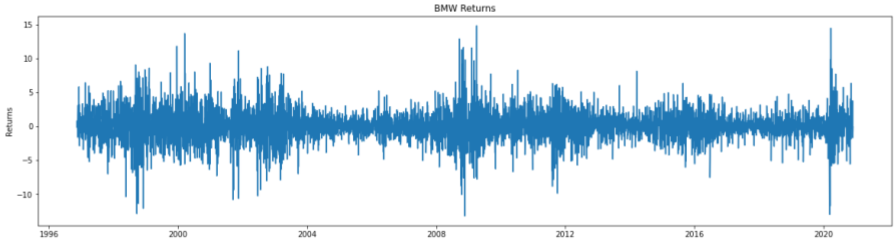


Figure 24 BMW Volatility

As seen in figures 23 and 24, there is clearly high volatility as compared to other times, so this is a good place to start when performing this GARCH process. The PACF function is used to find the order of the model, which aids in choosing the correct order by checking the coefficients and the p value. Then, a rolling forecast origin is used to forecast the origin of 365 days and constructing a model for each of the forecasting times, then predicting the next day and building a new GARCH method and predicting the next, and the process repeats, providing the prediction of the stock's volatility. The model fit.forecast function is used to forecast the stock's volatility for the next seven days.

These methods are used for the project since the aim is to forecast the growth of carbon emissions and fossil consumption, as well as the stock prices of Tesla and BMW. Moving average and exponential smoothing techniques are rather naive, and some features are already present in the ARIMA algorithm, and this method of modelling uses a linear regression technique to make future forecasting a reason, which is why keeping data stationary is key in eliminating seasonality and trends that can affect the model's performance. Since LSTM with keras is powerful in prediction problems and capable of storing past information, this is important because the previous price of the stock can be critical in predicting future prices. VAR is a multivariate time series, which means it has more than one dependent variable, and each variable has some dependency on other variables, which is then used to forecast future values. In other words, this model can understand and use the relationship between many variables, which is useful for describing the dynamic behaviour and provides a better forecasting result. Lastly, as previously said, GARCH is a model that predicts the volatility of the stock rather than the value of the stock. This is useful for those interested in investing in the EV market and other stock markets because it allows retailers to decide if it is worth buying or selling a stock based on whether it is going to be very volatile in a certain period or expected to be not volatile.

4.5.5 Implementing Tableau

In this project, Tableau is used to create an interactive visualisation. The CO2 and Fossil Fuel are used to view the differences in growth between the United States and other countries using a tree map and a filter feature that allows the user to click on the desired country. Figure 25 illustrates an interactive tree map, and it can be seen that, as predicted, Asia has the highest CO2 emissions followed by the United States, while the United States has the highest oil consumption.

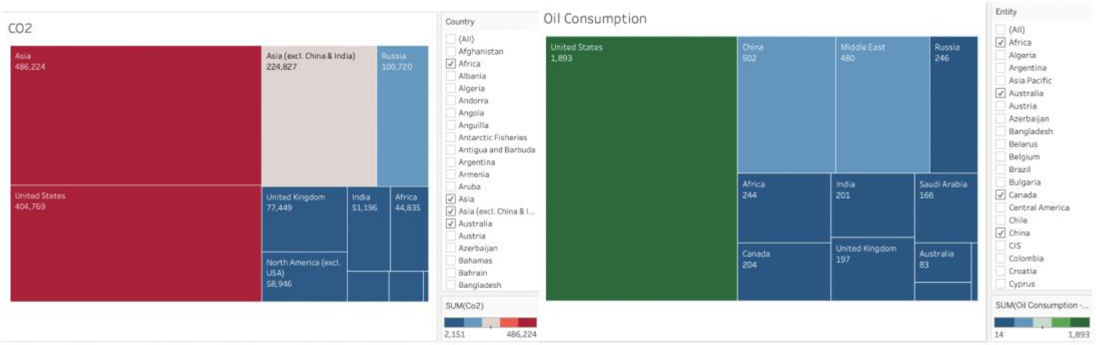


Figure 212 CO2 & Fossil Fuel

Figures 26 and 27 illustrate an interactive forecast for the three stock datasets, as well as CO2 and Fossil Fuel datasets, allowing the end-user to see forecast stock in a year as well as stock in a day.

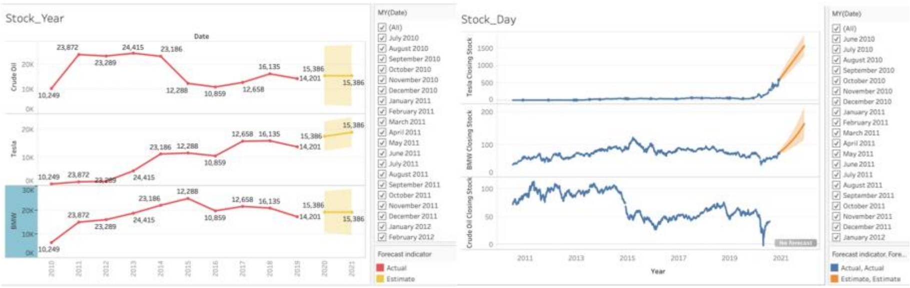


Figure 13 TSLA, BMW and Crude oil

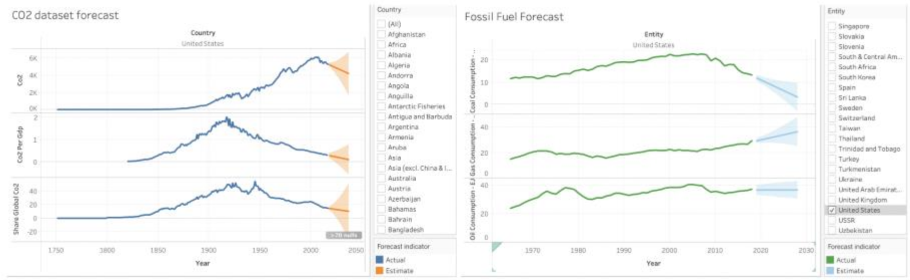


Figure 14 CO2 (left) and Fossil Fuel (right) Forecast

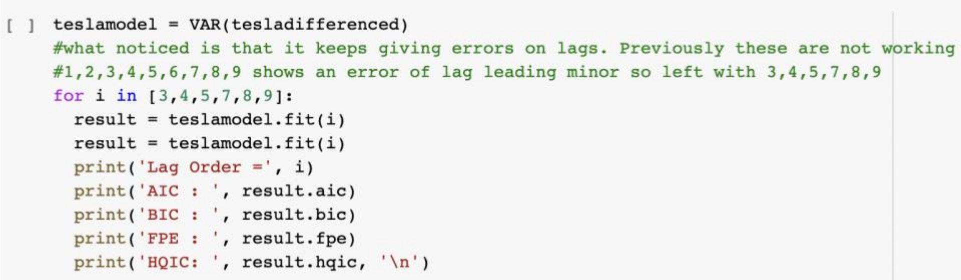
5.0 Testing

When implementing a project, testing is an important step to ensure that all of the project's components are working properly. Several tests will be carried out in order to put the project to the test. Each chart describes the function used during the test, the function's intent, the expected outcome of the test, and the actual outcome.

Table 6 Testing functions

Function	Purpose	Expected Outcome	Actual Outcome	Solution
Datareader	Load the data from Yahoo Finance	The required data will load into Google Colab	As expected	N/A
Pydrive	Authenticate and create the pydrive client	The user will be authenticated	As expected	N/A
Read_csv	Read the csv file from Google Drive – this works with Pydrive	The csv file will read it from Google Drive and load into Google Colab	As expected	N/A
model.predict	The model predicts the future values	The predicted values will remain the same after Google Colab gets reconnected	As expected	N/A
Model_fitted	Shows the summary of the model	Gives the same summary results even when Google Colab gets reconnected	As expected	N/A
Prct_change	Shows the percentage change of the model	Gives the same percentage change when Google Colab gets reconnected	As expected	N/A

As seen in Table 6, various tests were performed to ensure that it was functioning and producing the same results as when it was previously tested; these are only a few of the functions used. The results of the project as a whole have remained the same since the last time it was checked, such as the description, results of predicted values and future values, reshaped values, plotted graphs, and so on. However, there is a problem with the VAR model in that selecting the order of the model keeps changing in the sense that on the previous day the order of the model worked and if tested again, it will show which lag number it breaks and the order of lags must be modified. As seen in Figure 28, when the model was first implemented, numbers 1 through 9 worked and displayed their AIC scores; but, when tested/loaded the next day, only numbers 3 through 9 worked; this is fixed by excluding the number where it breaks. There is no real explanation for why this is happening, but this is the fastest solution. The 6th leading varies depending on where the code break occurs; it may be the 5th leading and so on, for example.



```
[ ]  teslamodel = VAR(tesladifferenced)
     #what noticed is that it keeps giving errors on lags. Previously these are not working
     #1,2,3,4,5,6,7,8,9 shows an error of lag leading minor so left with 3,4,5,7,8,9
     for i in [3,4,5,7,8,9]:
       result = teslamodel.fit(i)
       result = teslamodel.fit(i)
       print('Lag Order =', i)
       print('AIC : ', result.aic)
       print('BIC : ', result.bic)
       print('FPE : ', result.fpe)
       print('HQIC: ', result.hqic, '\n')
```

"LinAlgError: 6-th leading minor of the array is not positive definite"

Figure 15 VAR Model

6.0 Evaluation

The project has been evaluated in accordance with the CRISP-DM phases. Once the project reaches the evaluation phase, the model implemented will be evaluated based on the outcomes obtained and how well it worked. The models implemented for the project were aimed at forecasting the growth of CO2 emissions and Fossil Fuels, as well as the values of TSLA, BMW, and Crude Oil. All datasets sourced from Yahoo Finance, Kaggle, Our World in Data, and GitHub have been used in the project, including TSLA from 2010 to 2020, BMW from 1996 to 2020, Crude oil from 2000 to 2020, Fossil Fuel from 1965 to 2019, and CO2 from 1949 to 2018. The data has been extensively pre-processed and cleaned in preparation for modelling. The data was transformed for model implementations to produce outcomes and gain insight into the data through modelling stages. The CRISP-DM approach used in the project played a critical role in identifying each stage of project creation and was strictly adhered to. The models will be evaluated and compared based on the given results, model performance, and RMSE.

7.0 Results

The ARIMA model was the first Machine Learning model to be implemented into the project. Following the transformation of the data for the model and the implementation of the model to forecast future market prices, CO2 emissions, and fossil fuel consumption growth. As shown in figure 29, the predicted value (blue) and test (orange) are plotted to compare, and the model is not performing well based on the results of the predicted values. The model was then retrained as an entire dataset to forecast the values for the next thirty days. Figure 30 portrays the potential values for the next 30 days, and as can be seen, the closing stock appears to be performing well based on the model established. The same procedure was followed for BMW, Crude Oil, Fossil Fuel, and CO2, and the findings are shown in the graphs, indicating that the model is not performing well because it is far off from the actual values. However, what is very important to see is how it forecasts the 30 day future values; even if it is a bad model, it can be seen how the values shift to either increasing or decreasing values as seen in figures 30,32,34,36, and 38, and from these findings, the model is still able to forecast the values and growth as well as the future growth and values of the datasets.

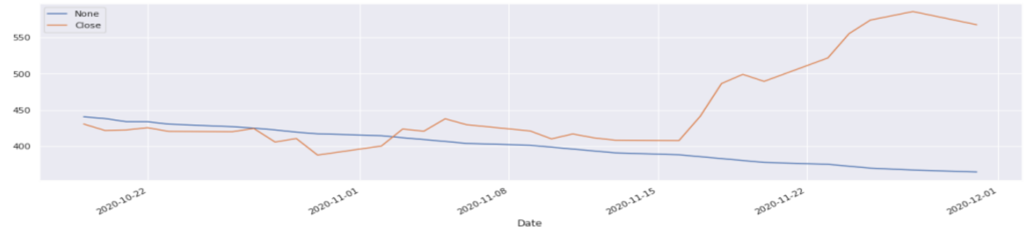


Figure 16 TSLA Test Model

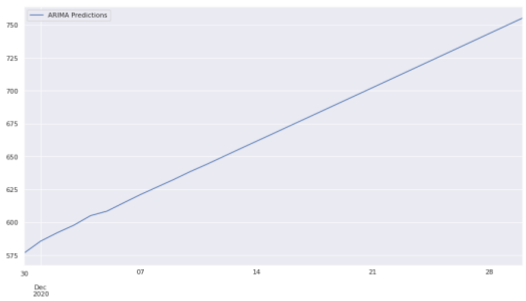


Figure 17 Future 30 days Predicted

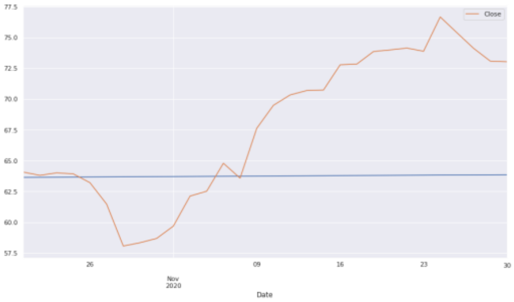


Figure 18 BMW Test Model

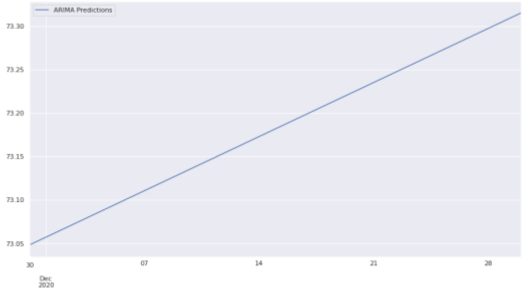


Figure 19 BMW 30 Day Future Prediction

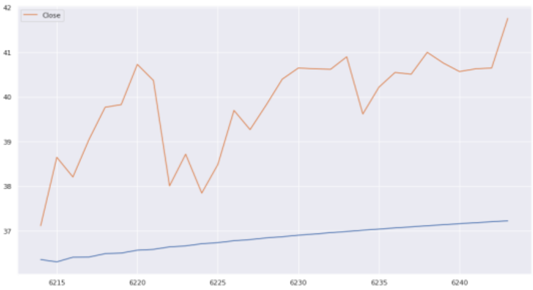


Figure 20 Crude Oil Test Model

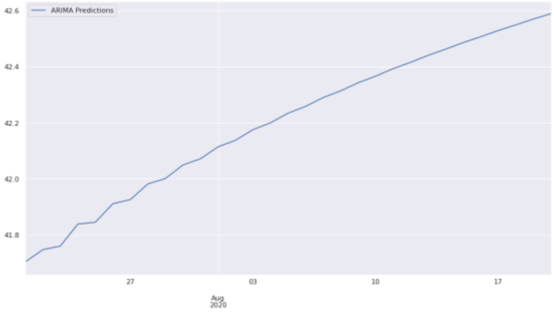


Figure 21 Crude Oil 30 Day Future Prediction

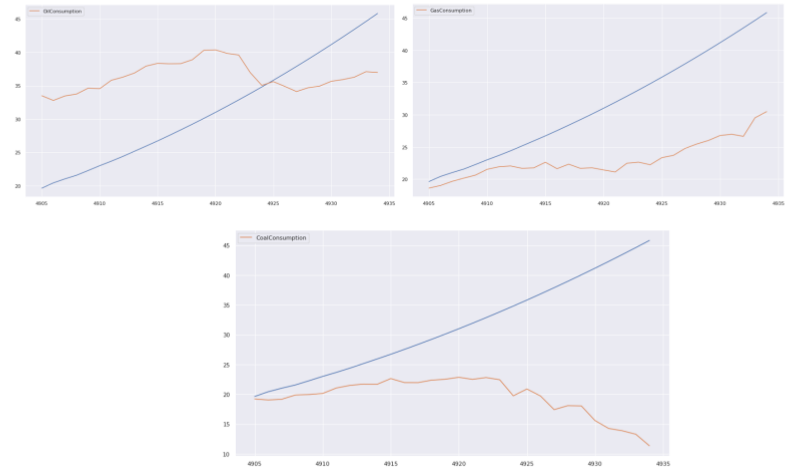


Figure 22 Oil(left), Gas(right), Coal(bottom)

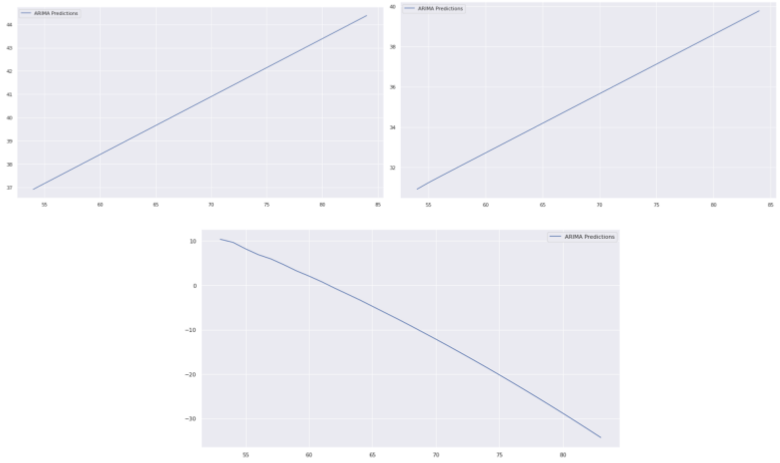


Figure 23 Oil(left), Gas (right), Coal (bottom) Future 30 day prediction

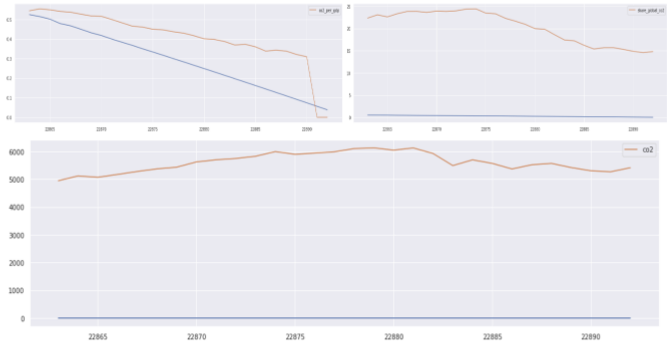


Figure 24 CO2per GDP (left), share global(right), CO2 (bottom)

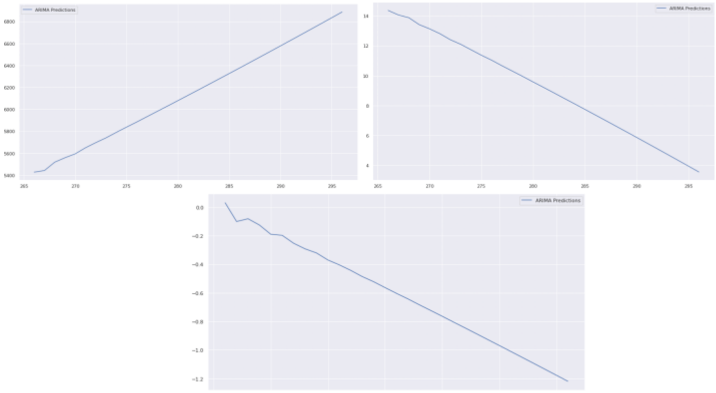


Figure 25 CO2(left), shareglobal(right), CO2 per GDP (bottom) 30 Day Future Prediction

The LSTM model was the second model to be implemented. As shown in figure 39, the TSLA model is performing well, with the predicted values being slightly lower than the actual values, but the model is still outperforming ARIMA. Figures 40 and 41 (BMW) and (Crude Oil) show that it outperformed TSLA in terms of predictions as it follows a similar pattern to the actual values. CO2 GDP, share global, and CO2 predicted values, on the other hand, are slightly off but still follow the same pattern as the actual values, as it has a very similar in patterns where the growth declines or increases. The difference between the LSTM and ARIMA models was in the estimation of future values; the ARIMA model can predict 30 days, while the LSTM model can only predict one day, but dates can be modified because it uses the datareader, which reads it from Yahoo Finance data. Following a successful LSTM implementation, the Fossil Fuel dataset was not successfully implemented because it failed to perform the modelling tasks even by using the same steps and making minor adjustments. The conversion of the training data into arrays, reshaping, and converting to 3D data has failed, which is why the dataset is not included in modelling implementation.

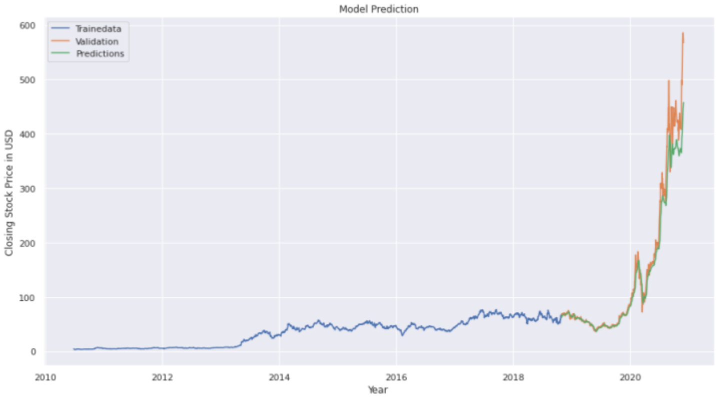


Figure 26 TSLA LSTM Prediction

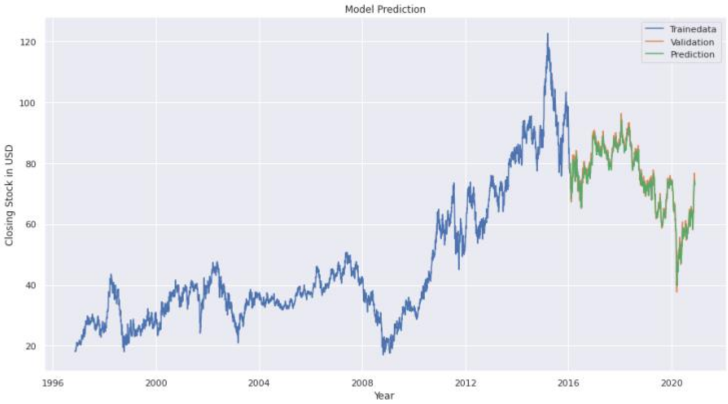


Figure 27 BMW LSTM prediction

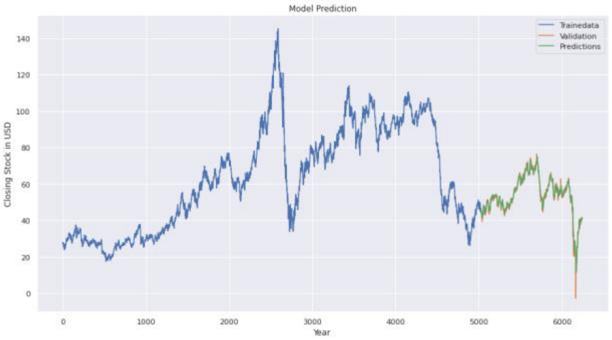


Figure 28 Crude Oil LSTM prediction

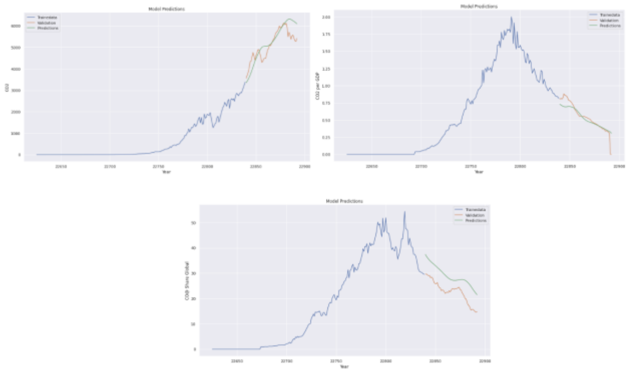


Figure 29 CO2(CO2(left), CO2 per GDP(right), Share Global(bottom) LSTM prediction

The third model implemented was VAR, which is a time series analysis with a function close to the ARIMA model. Figures 43 to 46 compare forecasted/predicted values to actual values, and the findings on the graph indicate that BMW (figure 43) outperformed the actual values, while TSLA (figure 44) actual values outperformed the forecasted values, indicating that both forecasted datasets are significantly off from the actual values.

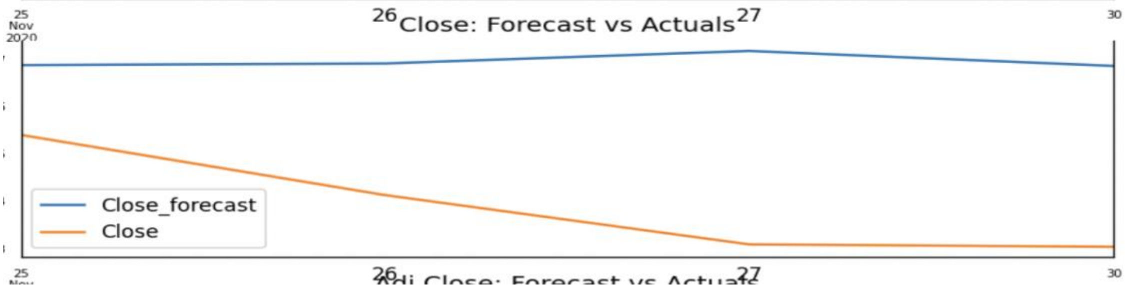


Figure 30 BMW Forecast

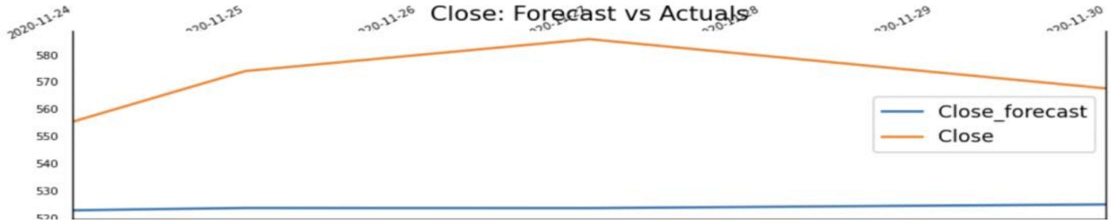


Figure 31 TSLA Forecast

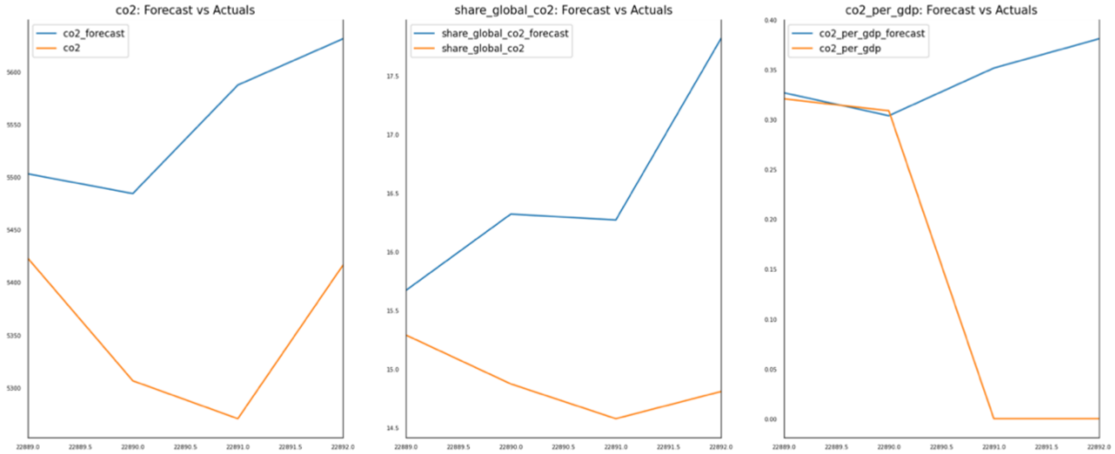


Figure 32 CO2 Forecasted

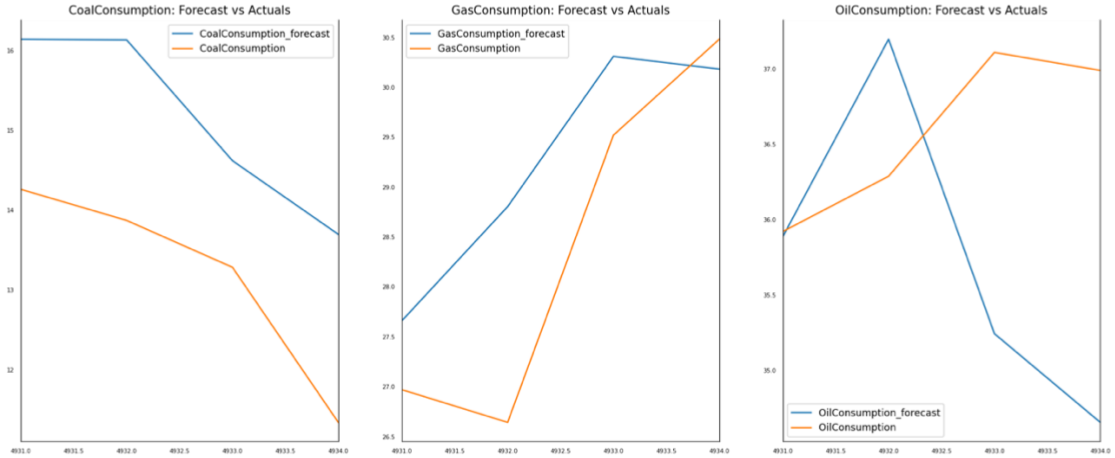


Figure 33 Fossil Fuel Forecasted

Figure 45 illustrates how the model performed in the CO2 dataset, with forecasted values outperforming actual values since the forecasted values have higher numbers than the actual values. On the other hand, fossil fuel has a slightly better performance on gas consumption, while coal consumption forecasted values are fairly high than actual values, and oil consumption has a slightly better performance as it fits the trend when actual values have also declined, as also observed the forecasted values are slightly following an opposite trend to the actual values. As discussed in the testing section, the VAR order of the model changes constantly, which mostly occurs in TSLA and could probably occur in CO2 and fossil datasets, which is why the crude-oil dataset is not implemented, as changing the number of lags from one to nine was also ineffective. Even, as noted, the model appears to be a poor model as compared to LSTM; however, in one's opinion, this is because LSTM was graphed from start to end date and VAR model was not graphed in the same format; thus, it appears to be a poor model; similarly, ARIMA model appears to be a poor model.

The final model implemented is the GARCH model, which measures the volatility of the stocks. The aim of this model was to supposedly apply the same method to non-stock datasets, but this cannot be achieved because it works best with stock datasets. GARCH is then only performed on TSLA and BMW because it determines market volatility, which is useful for retail investors. Figure 37 portrays BMW's 365-day rolling forecast, and the model is performing well because it follows the trend when the actual stock does not have high spikes (very high volatility) and changes when low volatility occurs, as observed when volatility occurred around March, which has the highest volatility, the predicted volatility (orange) also followed the same and same pattern has happened in the upcoming months. Figure 48 displays the seven-day volatility forecast, and as can be shown, there appears to be a trend, with high volatility and low volatility occurring every two days. As an example, the lowest volatility is on December 2, 2020, and a similar pattern is seen after two days.

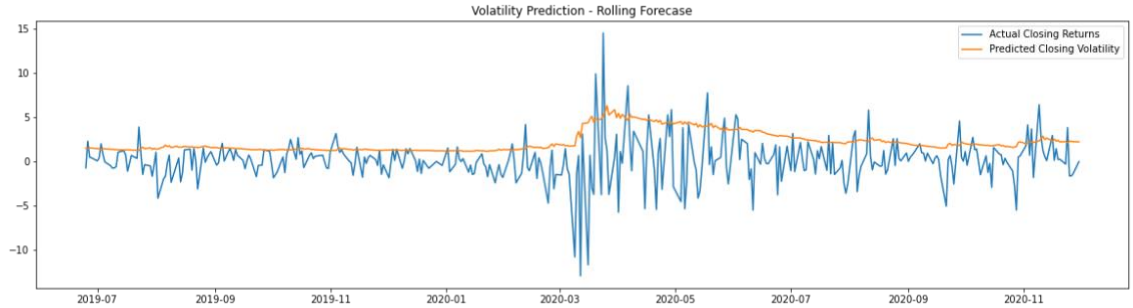


Figure 34 BMW Volatility Rolling Forecast

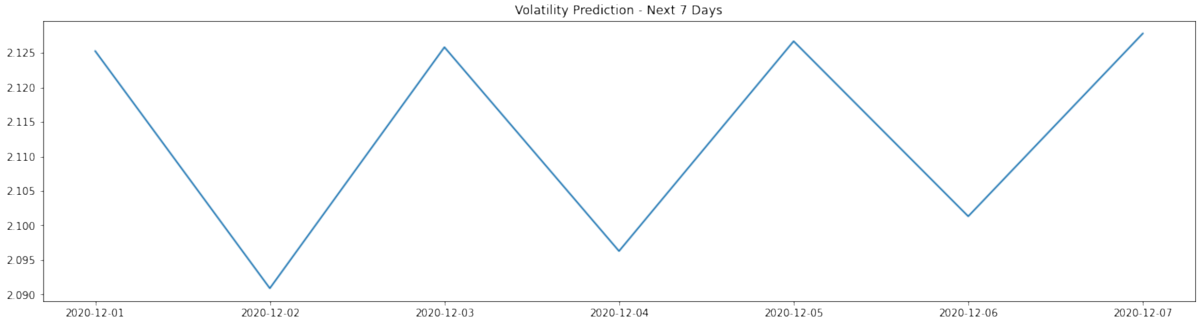


Figure 35 BMW 7 day volatility prediction

Figure 49 depicts TSLA's 365-day volatility forecast in comparison to figure 47; the volatility is more visible in figure 49 because TSLA tends to have more frequent higher spikes and the expected volatility follows the same pattern when this occurs. Figure 50 portrays the volatility of TSLA's stock for the next seven days, and it appears that 2 December 2020 has the highest volatility relative to BMW, with no trend observed. Furthermore, TSLA's volatility remained constant from the third to the seventh of December. Similarly, in Figure 51, crude oil volatility from start to end date shows the periods with the highest volatility, which was apparent as the expected volatility showed the highest swings, with the majority of the periods having low volatility. The model was also able to forecast seven-day volatility, and as shown in Figure 52, volatility is continuously increasing. One explanation may be that people buy and sell oil futures based on their hopes for the future. People would buy oil futures if they expect higher demand in the future, driving up the price. Similarly, if market sentiment shifts, investors would seek to sell oil futures ahead of prices. As a result, price fluctuations may have a greater momentum impact than the underlying economic cause. Oil demand is also affected by economic development. If incomes rise, there would be a greater demand for transportation, and thus for oil. As a result, demand for oil is extremely unpredictable.

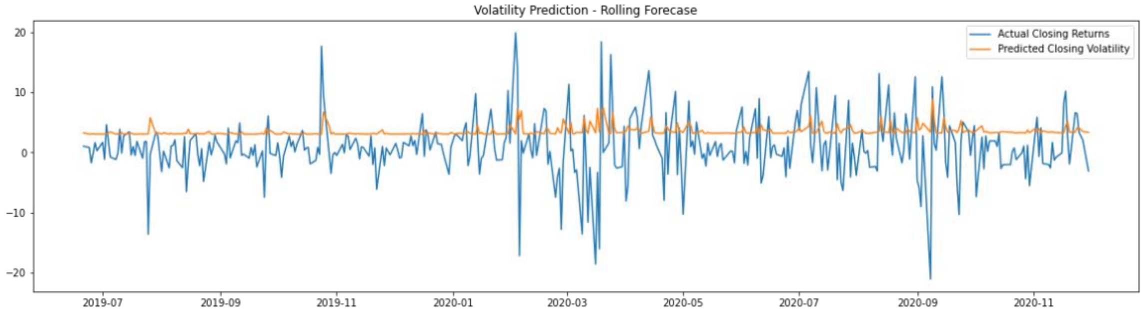


Figure 36 TSLA Volatility Rolling Forecast

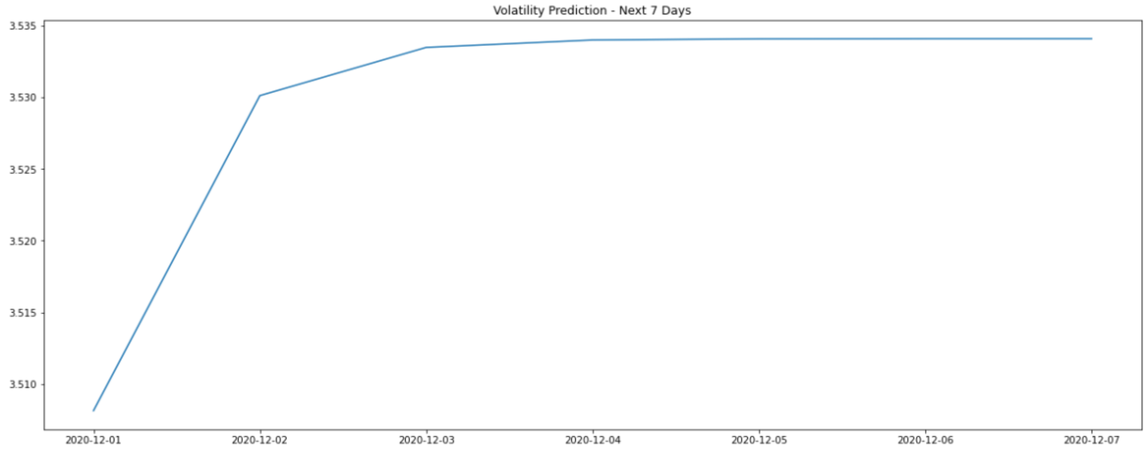


Figure 37 TSLA 7 day volatility prediction

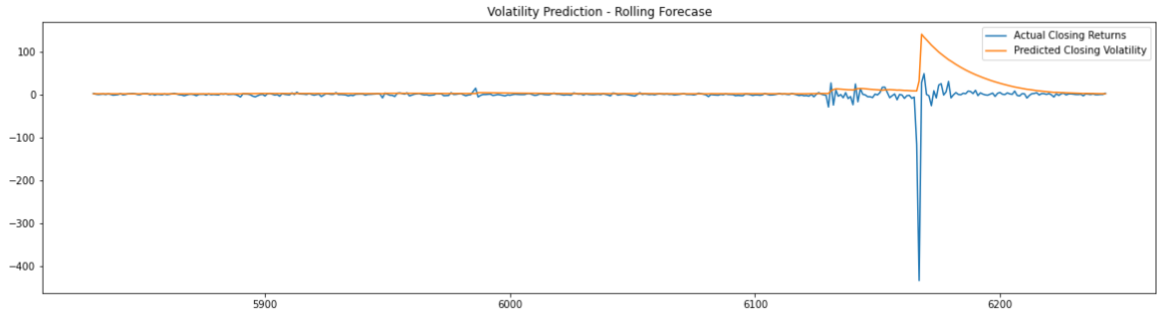


Figure 38 Crude Oil Volatility

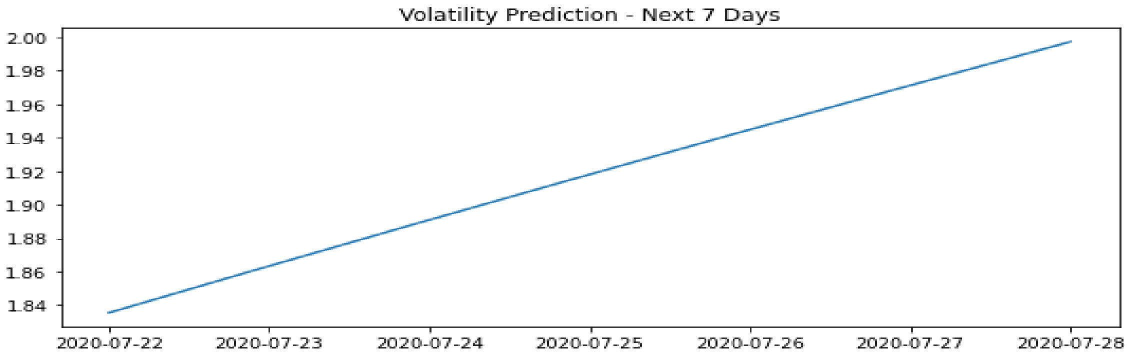


Figure 39 Crude oil Volatility for 7 days

Predicted Values Comparisons

		TSLA Predicted Values		
Year	**Actual**	**ARIMA**	**LSTM**	**VAR**
2020-11-24	521.849976	372.708327	406.004822	522.695493
2020-11-25	555.380005	370.080145	420.932800	523.538977
2020-11-27	574.000000	367.451767	436.076477	523.506541
2020-11-30	524.859791	364.820526	449.372498	524.859791
		BMW Predicted Values		
2020-11-25	75.400002	63.836616	74.866028	76.871882
2020-11-26	74.129997	63.844031	75.345467	76.907292
2020-11-27	73.089996	63.851445	74.931831	77.172011
2020-11-30	73.040001	63.858860	74.049164	76.854820
		Crude Oil Predicted Values		
2020-07-16	40.759998	37.141406	40.937550	
2020-07-17	40.570000	37.164959	41.061214	

2020-07-19	40.630001	37.186459	41.066471	
2020-07-20	40.650002	37.208556	41.053585	
2020-07-21	41.759998	37.229052	41.045158	
CO2 Emission				
2015	5422.966	5543.388019	5298.368326	5503.044709
2016	5306.662	5494.998070	5267.681246	5484.379233
2017	5270.749	5369.163487	5232.343602	5587.344617
2018	5416.278	5277.902844	5194.317641	5631.131621
Share Global				
2015	15.693	15.233424	24.039880	15.671687
2016	15.292	15.177928	23.430592	16.322975
2017	14.875	14.554619	22.856415	16.272250
2018	14.579	14.186330	22.317551	17.815818
CO2 GDP				
2015	0.321	0.319511	0.468442	0.327003
2016	0.309	0.315095	0.459687	0.304066
2017	NA	0.299501	0.450728	0.351860
2018	NA	0.094706	0.424467	0.381136
Oil Consumption				
2016	35.917927	36.238735		35.883652
2017	36.285099	35.993244		37.194733
2018	37.107278	36.692171		35.240089
2019	36.988039	37.581092		34.651478
Gas Consumption				
2016	26.96756	42.297827		27.656885
2017	26.63971	43.452675		28.800998
2018	29.51791	44.628841		30.307763
2019	30.47922	45.826382		30.180551
Coal Consumption				
2016	14.25867	42.297827		16.138601
2017	13.86933	43.452675		16.130741
2018	13.28205	44.628841		14.617291
2019	11.34064	45.826382		13.695051

When comparing expected past values to the graphs shown at the beginning, it is clear that the LSTM model is the better model. But, when looking at the results of values, VAR is the best option, followed by LSTM and ARIMA. LSTM showed a decent graph, so it appears to be the better model as compared to VAR; but, based on the graphs given, it only showed the first few months rather than the entire start date to end date. Despite the fact that VAR did not perform as well as the LSTM model in some of the datasets, VAR might still be the best model implemented.

8.0 Conclusions

Table 7 displays the RMSE results for all models, and the results show that VAR has the lowest RMSE value, followed by LSTM, and finally, ARIMA, implying that the VAR model outperformed the other two models (ARIMA and LSTM).

Table 7 RMSE for Models

TSLA RMSE	
ARIMA	87.464
LSTM	12.958
VAR	48.262
BMW RMSE	
ARIMA	6.957
LSTM	0.524
VAR	3.205
Crude Oil RMSE	
ARIMA	3.145
LSTM	0.205
VAR	N/A
CO2 RMSE	
ARIMA	5615.3038
LSTM	122.652
VAR	214.703
CO2 per GDP RMSE	
ARIMA	0.142
LSTM	0.019
VAR	0.259
Share Global CO2 RMSE	
ARIMA	20.353
LSTM	6.222
VAR	1.880
Oil Consumption RMSE	
ARIMA	9.269
LSTM	N/A
VAR	1.563
Gas Consumption RMSE	
ARIMA	9.889
LSTM	N/A
VAR	1.210
Coal Consumption RMSE	
ARIMA	15.512
LSTM	N/A
VAR	1.998

Modelling remains the most challenging task of the project; stock datasets work well with ARIMA, LSTM, and VAR, demonstrating the company's progress. With stock predictions, VAR is the best model because it has similar predictions to the actual values, followed by LSTM and ARIMA. Despite VAR's problems, this model will remain the best model for stock predictions/growth because it tests other variables that might be affecting each other. GARCH is often used for volatility and works better with stock datasets since running GARCH with non-stock datasets is not an optimal model or solution since it does not fit CO2 and fossil fuels dataset. As seen with TSLA and BMW, GARCH was helpful for determining market volatility; therefore, this will be useful for retail investors as it indicates when high or low volatility is possible. With regard to CO2 and fossil fuel use, I can still conclude that the VAR model is the best model for forecasting the growth of CO2 emissions and fossil fuel consumption. There are difficulties and problems in modelling since certain datasets do not function or experienced errors, with the exception of ARIMA, fossil fuel datasets do not seem to work in LSTM and Crude oil with VAR. Solutions were attempted to correct these errors, but they were unsuccessful and were left unfinished. More changes should be made in the future to ensure that the output models are better and that the majority of datasets work with the majority of the chosen models.

The project's strength demonstrates the company's valuation as well as the relationship between CO2 emissions and fossil fuels. Stock price prediction is also particularly useful for retail investors, as investors want to invest their money in something valuable and worthwhile, and with predictions and showing the growth of the business, investors can take advantage of, particularly the GARCH model. With all of the research and modelling, it is difficult to address whether environmental issues cause or impact the growth of car companies like electric car manufacturers because there is no association found from the datasets. While a relationship with TSLA, BMW, and crude oil was identified because some patterns have occurred, it is not a strong possibility. However, as people become more environmentally conscious, there is a chance that this will impact or influence the growth of car companies. As an example, graphs of fossil fuel consumption and CO2 emissions beginning in 2010 show a decline, and this is when TSLA began to release electric cars, implying that environmental problems may have an impact on the growth of car manufacturers. Oil companies may be impacted by the transition of cars to electric, but for car companies, the transition from traditional transportation to electric may take months or years. Furthermore, trucks and other forms of high-load transportation will continue to rely on fossil fuels.

In conclusion, the study shows that BMW and Tesla have an increase in their closing stocks in 2021, indicating the existence of environmental consciousness. The results indicate that electric cars will be the future car and will be in higher demand in the near future as countries such as the United Kingdom, California, and others begin to prohibit the use of diesel. This also means that companies that produce gasoline-powered cars are beginning to transition to electric vehicles, as BMW and Audi intend to introduce their electric vehicles in 2021. CO2 emissions are beginning to fall, as shown by the previous estimates, but oil consumption remains high. Oil has become the main energy source not only for transportation but also for homes, industries, and so on. Oil consumption is mostly driven by transportation, and reducing the amount of oil used by transportation will help to reduce CO2 emissions even more. The project has many drawbacks and limitations, such as the datasets provided not being large enough or fitting well into the model, and the chosen machine learning algorithm possibly not being the best model for the research or there are other approached not considered; therefore, the project needs a more in-depth analysis or more different approach.

9.0 Further Development or Research

More research is needed in this project to find the right dataset that will perform and fit well in the modelling process, as well as to look for different modelling techniques and/or how to better apply these models to the project. The project's aim is to build a GUI system for retail investors once the modelling methods have been mastered. An application that displays the results of the stock closure forecast based on the user's input. At present, the project can forecast closing market values for Tesla and BMW, which helps show the company's valuation, as well as predict the growth of CO2 and fossil fuels, but further research would be needed to determine if there is a correlation between these datasets and whether environmental issues really cause or affect the growth of car companies. The results would imply that an increase in CO2 emissions would result in an increase in green energy, which would replace oil dependence, resulting in the use of electric cars and other renewable energy sources. This would also increase demand for the production of electric vehicles and boost stock prices, resulting in the growth of car companies.

10.0 References

The History of the Electric Car", *Energy.gov*, 2020. [Online]. Available: https://www.energy.gov/articles/history-electric-car.

BAY.MOTOREN WERKE AG ST (BMW.DE) Stock Price, News, Quote & History - Yahoo Finance (no date). Available at: https://finance.yahoo.com/quote/BMW.DE?p=BMW.DE&.tsrc=fin-srch (Accessed: 7 May 2021).

Crude Oil Stock Price (no date). Available at: https://kaggle.com/awadhi123/crude-oil-stock-price (Accessed: 7 May 2021).

owid/co2-data (2021). Our World in Data. Available at: https://github.com/owid/co2-data (Accessed: 7 May 2021).

Ritchie, H. and Roser, M. (2017) 'Fossil Fuels', *Our World in Data*. Available at: https://ourworldindata.org/fossil-fuels (Accessed: 7 May 2021).

Tesla, Inc. (TSLA) Stock Price, News, Quote & History - Yahoo Finance (no date). Available at: https://finance.yahoo.com/quote/TSLA/ (Accessed: 7 May 2021).

11.0 Appendices

11.1. Project Plan

9.1.1 Objectives

I plan to create a project that will predict the development of the electric vehicle industry. I expect to find out how carbon dioxide emissions in the atmosphere and fossil fuels prices and the rise of the demand for electric cars can impact the valuation of the automotive industry. The project will concentrate mainly on CO2 emissions which also includes fossil fuel prices that impact or affect the development of the demand for electric cars. The project will also use data on electric vehicles and gasoline-powered cars to correspond on carbon emissions and fossil fuel data. The project will show and contrast the data found and compare it with other periods to graph the changes in the growth of automobile industry and CO2 emissions.

To achieve this, the project will be collecting considerable amount of data and this will help to ensure that the project will cover as much area as possible and use realistic time frames to illustrate a reasonable and equitable difference in the patterns in CO2 in the atmosphere and fossil fuels prices. During my research, I will also look at the percentage of other car companies that are electric cars or that are selling electric vehicles, as it is clear that Tesla is the only hundred percent electric car company that I am aware of. This will help me measure the adoption and will support my research question about what contributes most to an automotive company's valuation through the adoption of electric vehicles and a rise in carbon emissions, an increase in oil prices, and other potential factors. This is how the findings are to be done.

9.1.2 Background

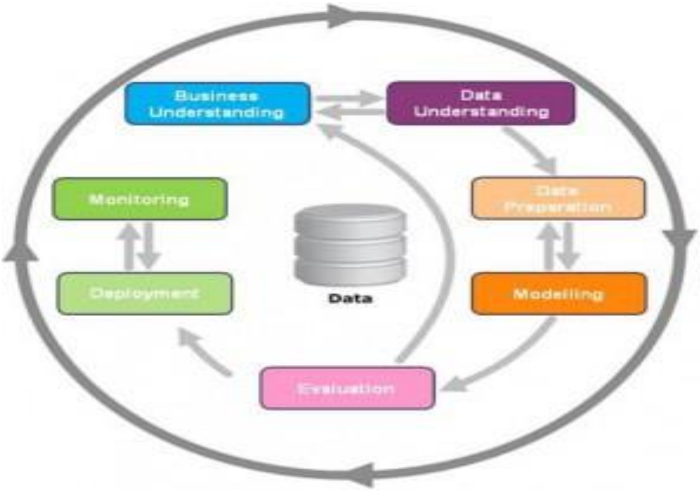
Electric cars were launched a hundred years ago, but for so many reason, they are seeing an increase in popularity today. Electric cars were manufactured after the first popular electric car made around 1890s and over the next few years around 1900s, and the continued to show a strong scale over the next 10 years. As stem and petrol/gasoline-powered cars have been introduced, electric cars have been growing and dropping over the years. However, it was not easy as it took manual effort to drive and was not easy to change gears. Electric cars seem to not have these problems since they are quiet and quick to drive and didn't emit pollutants like the others in this period. This led to the success of electric cars in such way that, in 1898, Porsche produced a hybrid electric car.

However, as it is widely available and more affordable and with other advances, Henry Ford Model T blow off electric cars in 1908, this also led to the decline of electric vehicles, especially with the discovery of crude oil, particularly not all cities across the countries had electricity at that time with this result that electric cars then disappeared by 1935. The internal combustion engines continued to be improved by cheap and plentiful coal. The rise in oil prices, however, has developed interest in seeking alternative fuel vehicles as well as more research and development in the field of electric and hybrid vehicles. With the introduction of the Clean Air Amendments and Energy Policy Act in 1990 and 1992, a new interest in electric vehicles in the U.S was developed with better features such as speed and efficiency far similar to gasoline-powered vehicles. Since then, electric vehicles have begun to draw more interest as people continue to become aware of the environment and the climate.

The aim of this project is to carry out data analytics on the role of rising of fossil fuels prices and carbon dioxide and adoption in the growth of the demand for electric vehicles. I will examine whether the increased demand in the atmosphere for fossil fuels and carbon dioxide affects the adoption or development of electric cars and the related changes and trends.

9.1.3 Technical Approach

Cross Industry Standard Process for Data Mining also known as CRISP-DM will follow the methodology for the Data Analytics Project. In order to evaluate the data, this kind of approach has many steps to follow, including business understanding, data understanding, data preparation, modelling, evaluation and deployment. The method is to discover valuable information from a certain amount of information. The project will apply this methodology in the following steps:



Data Understanding: This first step is to get a clear understanding of what we want to achieve and to examine the current situation by considering the issue, the effect and the consequences of the proposal. At this point, I'm also developing a hypothesis.

Data Preparation: I am taking the data from various tools to this. I want to make sure that the assumption I want to target for this project is clear within the data, and this is the start of testing my hypothesis and looking at the relationship that could exist.

At this point, I'm also looking for issues like missing values, duplicates, etc. Transformation, data extraction, data collection, data comprehension, integration is also handled at this stage where processing is important. It is at this point that the relevant data sources are blended together in such a way for data cleaning. I'm going to do an Exploratory Data Analysis in Data Preparation stage, where I'm going to look at the connection and relationship between variables, use statistical methods, and find trends that will help me understand the data and determine what I want to show. Feature Engineering tests the feasibility and value of features and eliminates unnecessary and irrelevant features to eliminate data noise.

Modeling: It is where we apply modeling techniques and also the core of the data life cycle. At this point, I'm going to choose a machine learning algorithm that is suitable for what I'm trying to achieve. This is where we have the training data algorithm and optimise the training process. With a training algorithm, it calculates a number of metrics, such as training error and predictive accuracy, which allows the model to learn well and generalise well to make predictions about unknown data.

Evaluation: It's the penultimate stage. After my model has been trained, I need to assess efficiency and accuracy to make sure I have achieved goals. Various multiple models can be created to evaluate effectiveness and to evaluate errors. This is to ensure that there is a clear understanding and assessment of the model metrics and, lastly, to prepare accordingly for how to execute the model delivery.

Deployment/Monitoring: This is the final step, after the model has been trained, tuned and tested, to be deployed in production and to make predictions. At this point after it has been deployed, it must be constantly checked, because the consistency of the model can be less reliable in time, to ensure that the output is still in place.

9.1.4 Special Resources Required

I will require no special resources.

9.1.5 Proposed Technologies

Here I will list and briefly explain the technologies that are acceptable and feasible that I will use for the final software project.

Python: Python is a programming language that supports and has a basic syntax for the management of big data. While R is more dedicated to statistical analysis, because it has open tools and even when it comes to combining statistics, Python is the best choice for me to use on this project because it has image analysis or control of physical experiments that would be very useful for this project.

Pandas: I'm going to use Pandas, along with Python, which is the Python Data Analysis Library and is used for anything from importing data from Excel spreadsheets to analysing time series data. In addition, with Panda data frames, cleaning and manipulation can also be achieved.

Google Colab: Colab or Colaboratory is a type of notebook and a Google Research product. What it does is allow anyone to write and execute Python code via a browser. Instead of using my own machine, I intend to use this program because running Python scripts always takes a lot of computer power which can take time and my machine output won't drop when running Python scripts.

Excel: The data is stored in a separate value (.csv format) file. The project will use Microsoft Excel to manage the data. Microsoft Excel is a spreadsheet program that can be used in spreadsheets to visualise data and run calculations.

Jupyter Notebook: In terms of configuration and organisation of the user interface, it is a web-based IDE and scalable to accommodate a broad variety of workflows such as data science and machine learning. In this project Jupyter will also be useful because is it efficient, scalable and shareable, as well as providing the ability to visualise data and build own documents from codes to reports.

9.1.6 Evaluation

This project will use many different methods of testing the data in order to validate the information. It helps you to answer questions related to data preparation, and model development by testing. This is carried out to check whether the visualisation graphs represent information correctly, whether it is possible to enhance it, and whether the models work without latency. This includes unit testing, integration testing, system testing, and acceptance testing.

Unit Testing can be checked if the components used work appropriately for individual methods and functions. Failure to carry out this test can lead to unexpected outcomes or incorrect data can lead to incorrect business choices. It becomes difficult to understand the mistake without data testing and where it happened, which makes it much harder to fix the issue.

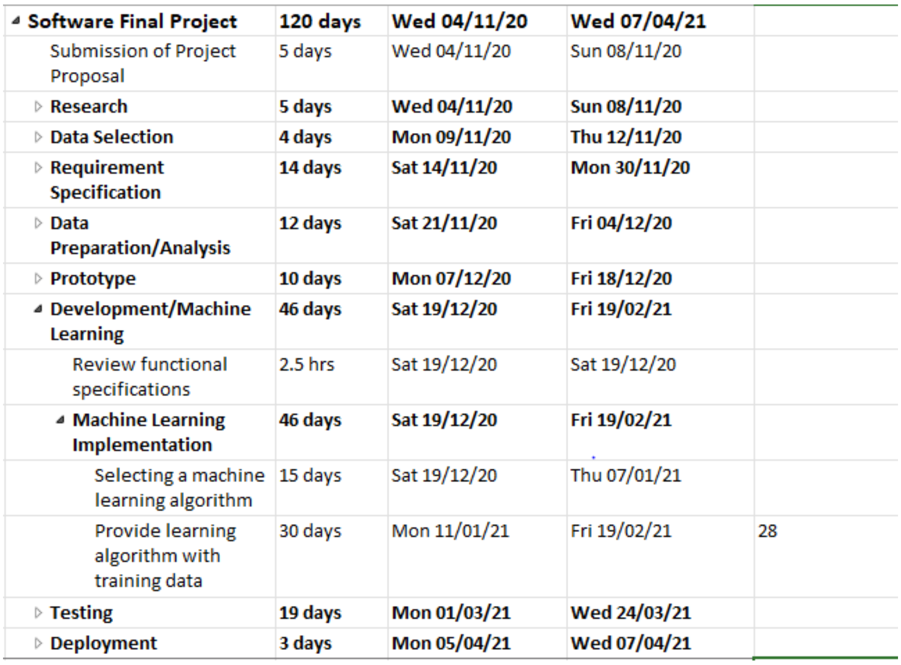
Integration Testing is testing all combined units and tested to check whether they are operating as they expect to when they were incorporated. This is to ensure that when various attributes are examined or tested, there are no faults. As there may be a chance during unit testing that anything was missing.

System Testing is where the entire program is checked to ensure that it is completely compatible with the specifications and to verify if other bugs are present or missing from the requirements.

Acceptance Testing is the final stage of testing. The project will ensure that the framework is thoroughly tested and developed in this context. The project will be reviewed to ensure that is is compatible and usable to and end user who is testing the program externally.

These are the tests that will be used to assess and evaluate the project. To ensure that it is compliant and meets the deliverable criteria set out. As well as ensuring that consistency, transformation of data, implementations, computations, scope and more are defined.

9.1.7 Project Plan



Software Final Project	**120 days**	**Wed 04/11/20**	**Wed 07/04/21**	
Submission of Project Proposal	5 days	Wed 04/11/20	Sun 08/11/20	
Research	**5 days**	**Wed 04/11/20**	**Sun 08/11/20**	
Data Selection	**4 days**	**Mon 09/11/20**	**Thu 12/11/20**	
Requirement Specification	**14 days**	**Sat 14/11/20**	**Mon 30/11/20**	
Data Preparation/Analysis	**12 days**	**Sat 21/11/20**	**Fri 04/12/20**	
Prototype	**10 days**	**Mon 07/12/20**	**Fri 18/12/20**	
Development/Machine Learning	**46 days**	**Sat 19/12/20**	**Fri 19/02/21**	
Review functional specifications	2.5 hrs	Sat 19/12/20	Sat 19/12/20	
Machine Learning Implementation	**46 days**	**Sat 19/12/20**	**Fri 19/02/21**	
Selecting a machine learning algorithm	15 days	Sat 19/12/20	Thu 07/01/21	
Provide learning algorithm with training data	30 days	Mon 11/01/21	Fri 19/02/21	28
Testing	**19 days**	**Mon 01/03/21**	**Wed 24/03/21**	
Deployment	**3 days**	**Mon 05/04/21**	**Wed 07/04/21**	

I created a Gantt chart for the project plan to visualise the key objectives and timeline for the project. By the second week of November, I set out to get the chosen data identified for the project. It might be difficult to find specific data that would fit well and be applicable to the project. With the support of my supervisor, I hope to find the right dataset during this week, because I am thinking of using Yahoo Finance as one of my sources and Kaggle and other possible sources. Going forward, I feel that while I was not where I initially set out to be at the beginning of the project, I am now more optimistic going into the second semester that I will be back on track with the plan.

11.2. Reflective Journals

9.2.1 October Journals

The first month focused more on research and the development of an idea. Since I had my six-month work placement in the area of child education and psychology, I have decided to link the idea to my work placement. Furthermore, I used to baby sit kids, and I personally have a younger sister, and I was curious about how children acquire such habits that could be dependent on what was going on within the home, school, and more.

I have decided to focus on how parenting affects the actions of children, and vice versa, and how the behaviour of parents or children relates to each other. At this point, there are no changes to the idea, as I am still waiting for feedback from my supervisor if there is anything that needs to be changed. As of now, finding an appropriate dataset about child behaviour is the biggest concern I have come across.

What I've done so far is research, and I conducted a brief parent survey last summer just to get an idea and where I could concentrate. Moving forward, I want to focus on finding the right dataset that will be very useful for the analysis, as this is the main element of the project and probably concentrate on few child behaviours, because parenting and child behaviour is a broad theme.

9.2.2 November Journal

This second month, after discussing with my supervisor, I decided to change my project idea because the first idea comes under a high-risk type of ethics. This was relatively challenging, since I've had my idea sorted out since summer of that needs to be targeted in the whole project and I find it very innovative. At the end of the day, I managed to think of a new idea.

My supervisor was incredibly helpful in offering, feedback, guidance and relevant information that I might probably consider for the software project. I haven't done a lot of work on what I've wanted to target since I've been occupied with module projects. What I've done so far is research work on my idea of possible technologies that I might potentially use, reading some research work, and finding the right dataset.

I have chosen to use CRISP-DM as my methodology, and knowing this method will help complete my project. Besides, the DAD project I submitted for my CA1 is similar to this final year project, so this is extremely helpful as I get an insight into how I can start the project. The next step I want to work on is to start coding, complete the requirements and have a prototype completed.

9.2.3 December Journal

I have been busy with completing the mid-implementation for the final year project and to finish much as I can. For the mid-implementation I have performed time series analysis on most of my datasets and graphs however, I have encountered issues when dealing with my non-stocks dataset ie fossil, carbon emissions and crude oil (crude oil is a stock data however inn some reasons it isn't working with the same analysis performed for real time stocks data taken from Yahoo Finance – Tesla and BMW). I have performed ARIMA and Linear Regressions, both are suitable for predictions, I have only successfully predicted the growth of my real time stock datasets but with the rest of my datasets they don't fit. This will be my next steps to have a more detailed research, either the datasets I have is not fitting well with the analysis I am trying and find another dataset or find a better solution.

The next steps is to do more research, more analysis to provide a better answer on my research question. This will be challenging since either the analysis or answer I can provide at the end of the day will provide an answer to what I am looking for or the opposite.

9.2.4 January Journal

This month was more smoother compared to semester one, after a busy previous semester with projects and TABA's. This month was a relief since there are only two modules a five credit module with Advanced Business Data Analysis and a ten credit module with Data and Web Mining. Data and Web Mining I find it really interesting since it can help me for my final project as it is learning about techniques and implementation.

After doing the mid-implementation submission it allows me to get an insight on my project. Implementing the following models – ARIMA and LSTM I have gathered some insights since ARIMA works best for stocks and for my other datasets like carbon emissions and fossil consumption using these types of models are not fitting well. For this month I have been looking for quite similar research papers to my project idea to understand how they have come up with solutions, methodologies used and the results and if these types of methodologies can be applied on my project idea. I have looked at some models and learning how it can be achieved and how to apply these on my final project additionally, includes understanding ARIMA and LSTM better. This month was more focused on research works.

9.2.5 February Journal

This month was spent on refining things however, I have spent the most time trying to resolve an issue with converting units and looking for good tutorials and reading materials for my ARIMA modelling, and trying to learn Random Forest, this has taken most of my time this month. Next month onwards the plan is to focus on modelling and finishing off the project and perform some testing since two projects on my other modules are both due in March.

What I am attempting to finish off is to try and solve the issue with my conversion of units if there is a more effective way instead of converting units manually once this is resolved the following thing to focus on will be documentation, machine learning/modeling, and visualisations. Using Random Forest and Tableau is still a thought to be considered. I have also been working with my showcase profile and updating based on the feedback I was given with.

9.2.6 March Journal

This month was devoted to working with unit conversions, and I struggled to figure this out and find a solution; however, I have now gotten this to work. I made some minor changes to my Data Analysis, such as adding corr plots, etc. In addition, I've been experimenting with Tableau since I'm considering using it for different types of visualisation when I've been using line graphs in Python and it might be nice to have a different type of visualisation. On the downside, I haven't made much of a difference in my modeling since I made the mistake of focusing more on Tableau; I've been busy with group projects and working with my CA, which I recently submitted but I've been working on my modeling using

ARIMA after submission and this week and next will be solely dedicated to modeling, with a week left before submissions to work on documentation and Tableau.

9.2.7 April Journal

This month was spent focusing on the Software Project as well as other CA/TABAs. I have finished my ARIMA model and LSTM, but there is a problem with one last dataset in LSTM (either it is not working with that dataset or something I am missing). I have also been working on the VAR model, but I have stopped working on it because it does not fit with most of my datasets and only works well with my Car stock datasets. My goal for the next few days and weeks, following the submission of my other module, is to possibly look at the GARCH model, work on Tableau, and begin working on the report and showcase. This will be my priority before the deadline.