CERTIFICATE OF APPROVAL

Using Machine Learning to Analyze Procurement Data

by

Sahal Abdul Ghafur

18001865

A project dissertation submitted to the
Information Technology Programme
Universiti Teknologi PETRONAS
in partial fulfilment of the requirement for the
Bachelor of Information Technology (Hons)
(BIT)

Approved by,

Dr. MANZOOR AHMED HASHMANI
Associate Professor
Computer and Information Sciences
Universiti Teknologi PETRONAS

(Dr. Manzoor Hashmani)

UNIVERSITI TEKNOLOGI PETRONAS

TRONOH, PERAK

January 2022

CERTIFICATE OF ORIGINALITY

This is to confirm that I am responsible for the work presented in this project, that the original work is mine except as noted in the references and acknowledgements, and that the original work included herein was not undertaken or completed by unnamed sources or individuals.

Sahal

____

(Sahal Abdul Ghafur)

ABSTRACT

International trade rules have recently received a lot of attention for restricting cross-border trade of vital goods. Forecasting future trade patterns is a high priority for policymakers all around the world because trade has such a significant impact on jobs and income.

A country's import data is an important element in projecting the country's GDP and can also be used to forecast future imports of a specific product. The reason for this is that the amount of imports a country purchases also indicates how much consumption occurs within the country, i.e. whether imports increase or decrease, It will also have an impact on that country's exports, and if exports are impacted in any way, there will be a clear link between those exports and the imports of a certain product.

As a result, the goal of this research is to develop a model for forecasting future trends using Qatar's import data. I also intend to demonstrate how Import and Export data benefit a country, as well as the data cleaning procedures and different models I used to predict the model. For each quarter of the year, the ultimate goal is to plot the actual data versus the forecasted data.

ACKNOWLEDGMENT

I would like to take this chance to thank Universiti Teknologi Petronas (UTP) for giving me with numerous opportunities throughout my studies, as well as numerous experiences that have aided in the development of my knowledge and skills, while leading me towards the achievement of my goal.

I would also like to take this opportunity to thank my Final Year Project supervisor, Dr Manzoor Ahmed Hashmani for guiding me during the project and helping me overcome the obstacles I faced along the way.

I would also like to thank my family for their infinite support and moral encouragement throughout my studies, especially during my Final Year Project, without your love and support I wouldn't be able to reach this far.

Finally, I would like to thank the Almighty for giving me the strength and guidance which lead me towards success.

Table of Contents

LIST OF FIGURES

Chapter 1: Introduction

1.1 Project Background

A country's import data is an important element in projecting the country's GDP and can also be used to forecast future imports of a specific product. The reason for this is that the imports that a country buys also tell us how much consumption occurs within the country, so if imports increase or decrease, it will have an impact on the nation's exports, and if exports are affected in any way, there will be a direct relationship between those exports and the imports of a specific product.

Many governments have been concerned in recent years about rising trade deficits (exports minus imports) and the implications for employment and wages. The United Kingdom is an example of this, as its trade deficit has risen over the past two years. In fact, the UK’s trade deficit is at a record high since it joined the European Union (EU) in 1973. The UK’s trade deficit is currently approaching £50 billion and was even larger before the global financial crisis in 2008.

Although the UK has some of the highest levels of trade deficits among developed countries, it is not alone. The United States is another high-deficit country; however, it is experiencing a surplus for the first time in about two decades. In fact, for some countries, such as Japan and Germany, it appears that their trade surpluses are only temporary and that they will experience deficits in the future.

International economics has a long history of helping us better understand the causes that cause trade and the implications of free commerce between countries. Nonetheless, recent shocks to the free-trade regime raise doubts about the accuracy of past forecasts and their application in significant trade disputes. Because of the close relationship between imports and exports, many countries are faced with the task of forecasting both at the same time, which can be difficult depending on what type of economic model is used. The most common method for predicting imports is to use an economic model that contains a sequence of variables that are projected to influence future imports. These variables can include but aren't limited to: GDP, population growth and consumption levels. However, these variables alone may not be enough for making accurate predictions about a country's imports soon, which is why several other factors need to be taken into consideration when creating a method for projecting imports.

High-quality import forecasts would help the country in considering future imports, and its impact on the country's exports. The economic effects of imports can be very significant on a nation’s overall economy. Changes to long-term trend in a country's imports is often a major indicator of whether that country is experiencing an economic downturn, which would ultimately lead to increased unemployment. This study identifies ML approaches applicable for international trade and demonstrates their validity in creating high-quality estimates to handle these difficulties. Transparency, which is crucial in the context of trade policymaking, has also benefited from recent technical developments in machine learning and data accessibility.

Borrowing from traditional forecasting methods, machine learning approaches are used to create high-quality import forecasts for countries involved in international trade disputes. Machine learning models have been shown to have efficiency in terms of adaptiveness and accuracy.

1.2 Problem Statement

A recent McKinsey & Company report stated that “The focus on analytics has become a major challenge for companies across all industries, as they try to keep pace with the continuing shift in consumer behavior, the growing complexity of their business models and ever-higher expectations from investors and regulators.”

The use of machine learning and data mining techniques across many disciplines has exploded in recent years with the field of educational data mining growing significantly in the past 15 years. But Machine Learning is only lately being applied to econometrics, according to a recent study by the National Bureau of Economic Research (NBER). ML has been used to solve problems in a variety of fields, including healthcare, education, and sports. So far, there are just a few applications for studying international trade trends.

1.3 Objective

The project demands to fulfill the objectives mentioned below: -

1. To Analyze the Import data, and analyze various algorithms required to give the best accurate result.

2. To Develop a Machine Learning model to accurately forecast National Import trends in Qatar.

3. To Verify and test the working of the model for accurate results.

1.4 Scope Of Study

Scope of the study refers to the elements that will be covered in a research project. It defines the boundaries of the research. It is often used in research projects based on quantitative data. The scope of study will depend on several factors related to the research topic and methodology. It is important to identify the factors that affect the scope to avoid any misconceptions or misinterpretation of results.

The scope of this project is limited to the data sourced from the Foreign Import Trade website of the State of Qatar. Foreign merchandise trade is the exchange of goods across international borders or territories. Foreign Merchandise Data is an important data since it represents a significant share of gross domestic product (GDP).

It is recommended by the United Nations, International Merchandise Trade Statistics Manual (IMTS) 2010-Concepts and Definitions that foreign merchandise trade statistics record all goods which add to or subtract from the stock of material resources of Qatar by entering or leaving the economic territory.

Chapter 2: Literature Review

This literature review examines the previous research for predicting import trends of International Trade , previous research on using Machine Learning for econometrics. Qatar is a country characterized by high and growing intra-industry trade. Qatar is a major supplier of natural gas and has also increased its exports of refined petroleum products to Europe. Qatar is not only a receiver but also an exporter depending on the commodity exported.

2.1 Using Data Mining on Linked Open Data for Analyzing E-Procurement Information

This report published by Eneldo Loza Menc´ıa and Simon Holthausen and Axel Schulz and Frederik Janssen explores how complex procurement information can be used to support strategic decision-making which is important with the increasing amount of information available on the world wide web. The study takes on the task of describing how data mining techniques can be used on data to estimate the number of bids in government contracts. They present a basic approach for converting linked data into a typical machine learning language, after which they use popular techniques like as discretization, text field processing, feature selection, state-of-the-art machine learning algorithms, and more.

Unfortunately, the accuracy of the algorithm was not satisfactory. The estimate of the accuracy was found to be around 32.69% (with a baseline of 30.34%)

2.1.1 Transformation of Data

The information in this study came from a publicly available dataset that included 1658 contracts for which tenders were issued. The report's purpose was to forecast the number of tenders. The goal values in the training set ranged from 0 to 73, however there were only 36 possible values in the data, with 1 being the most common. The RDF data was transformed into attribute-value data, also known as tabular data, which is the most often used data type in Machine Learning methods. Contract examples, which are normally based on publicly available information released by the European Commission, are usually the starting point of the change.

They were unable to determine the suitable type since all of the dataset's characteristics were not text strings. The RDF-file predicates led to an xsd:int literal, which was subsequently transformed into numeric attributes. The authors used a basic heuristic that they said worked well since predicates normally offer no information if the objects are in a closed value range. The following is the heuristic:

- It's a nominal candidate if the attribute-value of all occurrences is smaller than 20 characters.

- If the attribute is a nominal candidate and the total number of different values is less than 30%, it's a nominal attribute.

2.1.2 Preprocessing of Data

For all of their studies, the scientists employed the Weka package, which is a prominent machine learning framework that incorporates numerous cutting-edge algorithms. They also favoured the Weka package since the team was more interested in implementing existing algorithms than than inventing new ones. Missing values were replaced with a placeholder value that indicated the value was missing. Because the dataset included a large number of values in the form of RDF triples (text). Using a simple text processing method, the researchers attempted to leverage the new information. Feature selection may generally improve a model's prediction quality while also lowering its computing expenses. However, this feature selection approach came with a trade-off. It reduced the precision of the results by a huge margin. Moreover, by removing features, the model was "discriminating against" its training data and consequently learning irrelevant information. This kind of behaviour made it difficult to make use of useful information in some cases.

2.1.3 Best Approach

The team's cost sensitive ensemble had the best overall results, hence it was chosen to make predictions on the unlabeled data test set. The report shows the distribution of the projected number of tenders in contrast to the original distribution on the training data.

2.1.4 Conclusion

This research is the first step toward high-performance machine learning on semantically linked data. The team noticed a number of problems with data conversion into relational format and post-processing. The crew, however, had only taken a cursory look at the different choices and needed to do further research. The model they created projected an accuracy of 32.69 percent and a cost of 0.37. The team also believes that feature filtering may be improved.

2.1.5 Issues with the Report

- Predicates with multiple objects weren’t covered in the machine-learning model
- Some of the literals were required to have a same type
- The team could not solve the same as predicates and did not merge the resources
- The team failed to achieve a satisfactory accuracy level
- The team found it hard to create their own filters to resolve these ambiguities
- There was no set standard for filters, since they were not based on any legal statutes
- There was an incomplete solver model in the system that did not allow them to filter

2.2 Application of Machine Learning in Forecasting International Trade Trends

Feras A. Batarseh, Munisamy Gopinath, Ganesh Nalluru, and Jayson Beckman describe the use of machine learning in forecasting international trade patterns in this research study. Different machine learning models for international commerce scenarios were described in this research, along with concerns regarding their applicability and prediction quality. The methodologies used in this study allowed the most relevant economic variables that drive commodity trading to be extracted.

2.2.1 Transformation of Data

The data for this study came from the USDA's Foreign Agricultural Services' Global Agricultural Trade System (FAS - GATS) (USDA 2019). The GATS system is published by the United States Department of Agriculture. The World Bank's World Integrated Trade Solution (WITS 2019) and the United States International Trade Commission's Gravity Portal are both used to gather economic data (2019). The GATS approach is used for seven key commodities: wheat, milk, rice, corn, beef, soy, and sugar. After that, the economic and commodity data is merged in a SQL database. A R code is used to combine the year of economic data as well as country-to-country trade transactions. The data is joined using an Inner Join. More than 30 economic factors are examined for relationships.

The strongest economic relationships are discovered, for example, between population and whether a nation is an island, currency and GDP, and WTO membership and Free Trade Agreements, among other things. These findings suggest the use of machine learning approaches to see if predictions can be made that are better than traditional econometrics.

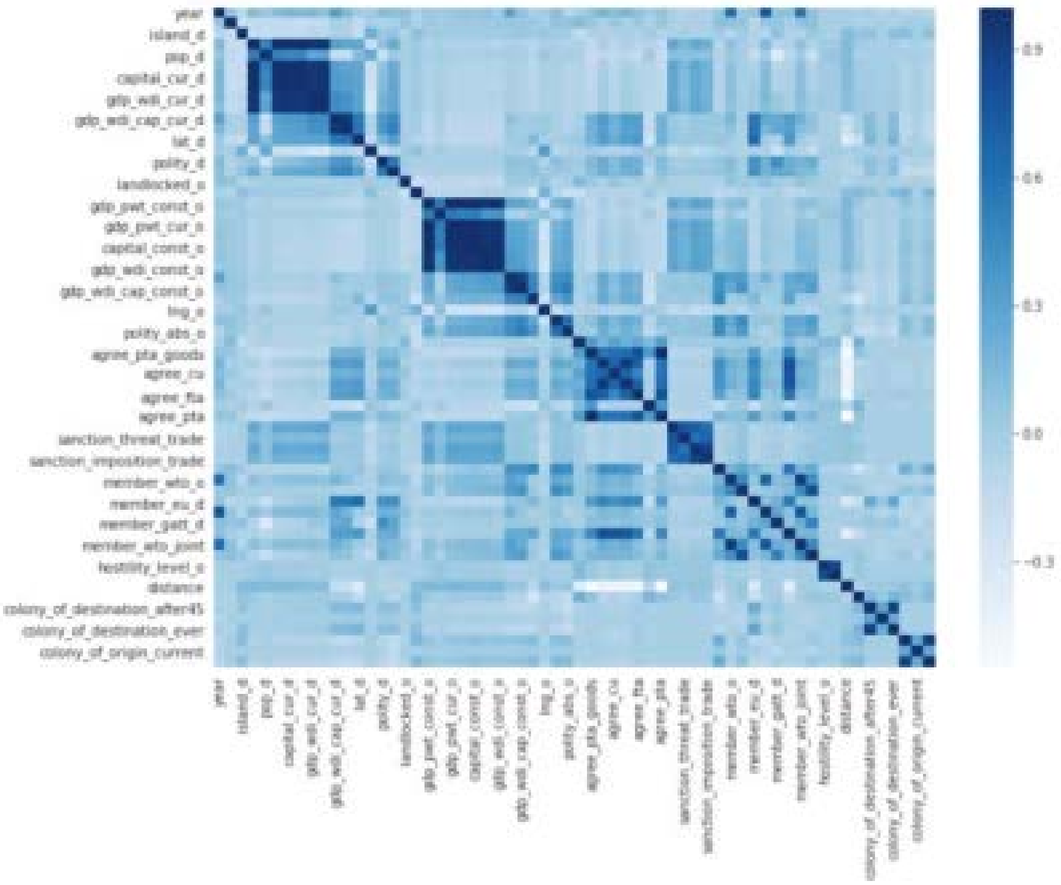


Figure 1: Correlations of 30+ Economic Variables

2.2.2 Machine Learning Methods used

The following supervised and unsupervised approaches were investigated in this study: Linear Regression, K-means Clustering, Pearson Correlations, Boosting, and Time Series such as Autoregressive Integrated Moving Average are some of the techniques used to analyse data (ARIMA). The seven primary commodities are subjected to simple linear regression modelling, with the goal of predicting imports and exports of a certain commodity.

The main commodity they chose for predicting their model is Beef.

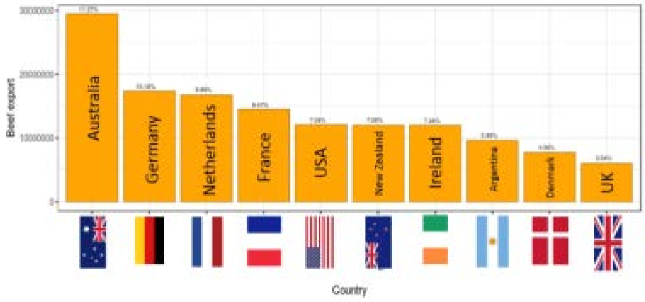


Figure 2: Top Beef Exporters

Beef trade data is available from 1989 to 2018; the years 2019-2021 are forecasted (red line in Figure 3). As seen in the graph, commerce between countries is varied and may fluctuate dramatically over time, even for a single item. As a result of the significant variation in the data, despite being supervised, a basic regression model yields straight-line pointers to the future of the beef trade (implying growth remains constant).

Other economic factors are gradually included to the modelling process after that, in addition to trade values. After all the economic factors have been incorporated, the goal is to determine which variables have the most impact on trade estimates, and which ones can be manipulated and tweaked to modify forecasts. Different commodities ranked economic variables differently, although distance (between the two nations conducting commerce), exporter population, and both countries' GDP had the most influence on whether two countries would trade one of the seven primary commodities or not. As a result, ARIMA is used in the cattle trade. ARIMA's benefit is that it delivers univariate forecasts that enhance output.

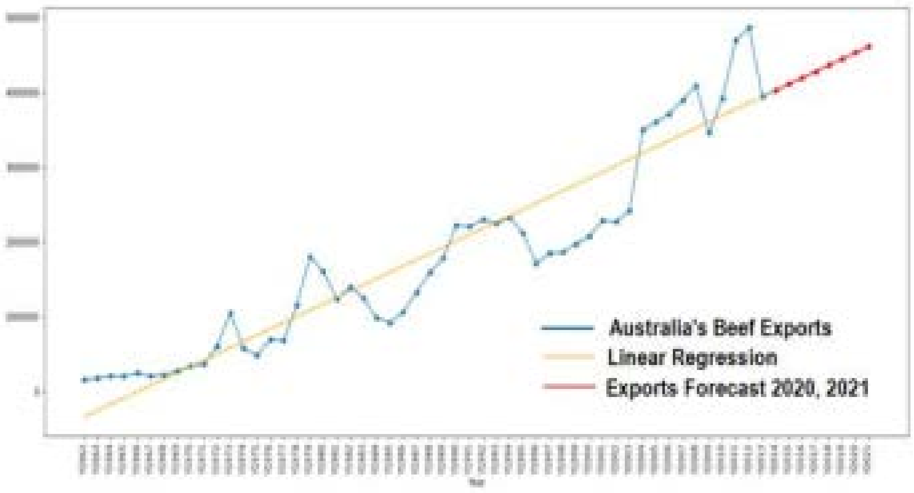


Figure 3: Australia’s Beef Exports 1988-2021

2.2.3 Results

Prediction is concerned with occurrences that will occur in the future. Machine learning algorithms boost the system's intelligence without requiring operator interaction. According to Ethem Alpaydin, "Machine Learning (ML) is utilized to optimize the performance criterion utilizing sample data or previous experience."

The quality of trade forecasts using the XGBoost Model was 69 percent, and the quality of trade predictions using LightGBM was 88 percent (in contrast, XGBoost scored the lowest of the three approaches). The number of leaves, the maximum depth of the tree, the learning rate, and the feature percentage are all parameters that may be tweaked in boosting models. With vast tree depths, small learning rates (0.01) are optimum. Using those parameters, LightGBM produced the greatest results. Sugar, for example, received a R 2 score of 0.73, whereas Beef received a R 2 value of 0.88, and Corn received a R 2 score of 0.66. These preliminary findings support the use of machine learning algorithms to forecast trade trends and show that they outperform traditional methods in terms of accuracy.

2.2.4 Conclusion

The results of this study show that machine learning is extremely useful for forecasting a variety of trade patterns with more accuracy than traditional methods. Their models are also a viable alternative to econometric methods, which are rarely cross validated. The authors learned that data collection and forecasting interact with each other: data collection will influence the selection of variables to be used by the machine learning algorithm, and forecasting will influence the choice of variables to be used.

Chapter 3: Methodology

3.1 Introduction

This chapter will go through the approach that will be employed and what will happen at each phase. The goals of the project might be fully articulated in the methodology section of the narrative plan. When it comes to attaining the provided goal of a project in a set amount of time, it is critical to prepare ahead of time and schedule properly so that the project may be completed on time. Excellent organization necessitates thorough planning and meticulous scheduling, which has an impact on the project's outcome.

For a project to be completed successfully, it is effective for all parties involved in the project to understand what is expected from them. This is typically communicated through a schedule and a detailed plan of action. This way, all parties involved in the project will be organized and prepared for the documentation of each phase's details. Communication is very important when it comes to project management. If there is a lack of communication surrounding a project, it could be detrimental to the project's success.

For a project of this magnitude to be successfully completed, the specific roles and responsibilities of each party involved need to be clearly defined. It is important that all of the parties involved in this project know who they are accountable for working with and get the correct resources they need to complete their tasks efficiently.

The results of each phase will be reviewed to see what worked, what didn't work, and how to improve for the next segment. This will help us steer the project in the appropriate direction to be successful. It is important to learn from mistakes that were made in earlier phases by analyzing the data that was collected during each milestone. Overall, this is an essential guide that can easily be adapted into any type of business or personal endeavor.

3.2 Research Methodology

In this project, Agile Software Development Life Cycle is chosen as the methodology. Agile software development methodology is one of the simplest methodologies to describe the details of the whole project. It is the fast, flexible, error-proof and simply a better way methodology to handle project.

Agile methodology is based upon several principles, including:

Iterative & Incremental Development - Developing the product in short cycles called iterations. Agile uses incremental approach to deliver solution. Each iteration must be completed within a time-boxed period called sprint. Typically, this time is between two weeks to four weeks depending on the project type. Within each sprint, the project team will have milestones marked to track how much work has been accomplished and how much work remains.

Customer Collaboration & Responsive - Agile methodology is based upon this customer collaboration concept. The customers are actively involved in the development process. Their feedback is required to ensure successful delivery of product at the end of project. There is a lot of communication between the team members and customers.

Individuals & Interactions Over Processes & Tools - It is better to focus on people than tools or processes, making sure they have the skills, knowledge, and behaviors they need to be successful in their work.

Simplicity & Evolving - The product must be simple and easy to understand, because it should be understood by everyone. The team should adopt a design thinking approach, which enables team members quickly to create new designs. It helps to evolve the product based on customer feedback rapidly.

3.3 Project Framework

3.3.1 Exploratory Analysis

As mentioned above the main framework that this project would be using is CRISP-DM which stands for and has six main phases, it can be noted that main of these

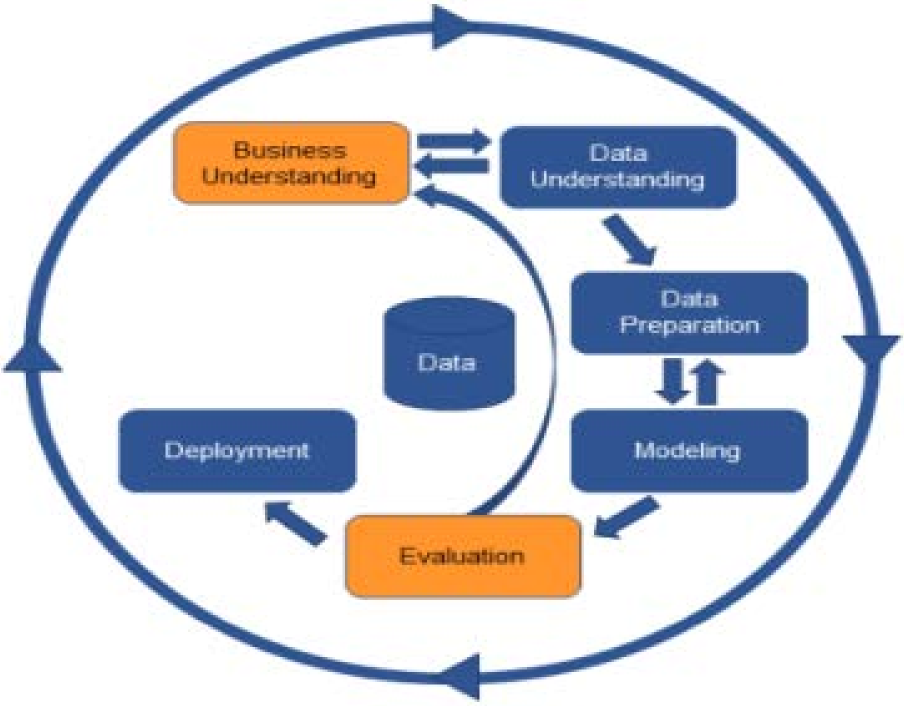


Figure 4: CRISP-DM Model

The stages involved in the data mining model include data understanding, preparation, modelling, evaluation, and deployment.

3.3.2 Data Understanding and Preparation

The data specified in the project resources must be acquired in the second stage of the CRISP-DM procedure. Data loading is included in the first collection if it is essential for data interpretation. In order to create an efficient machine learning model, it is critical to first analyse the data. I'll be describing data resources acquired from various locations, as well as the data format. Sourcing the data required for Machine Learning Analytics is one of the most important parts of a model.

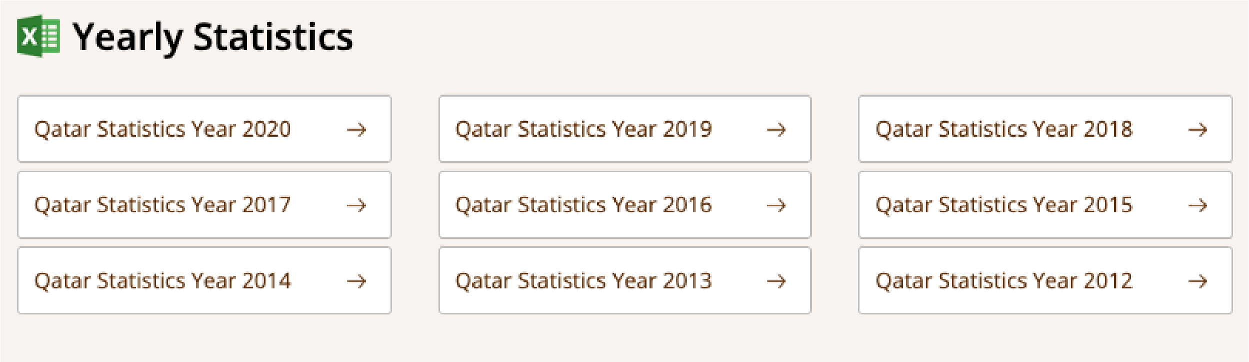
For my Machine Learning Model, I have procured my data from the Planning and Statistics Authority of Qatar (https://www.psa.gov.qa/en/statistics1/ft/Pages/default.aspx) The data I used for the model is the foreign merchandise trade data. Foreign merchandise trade is the exchange of goods across international borders or territories. Foreign Merchandise Data is an important data since it represents a significant share of gross domestic product (GDP). It is recommended by the United Nations, International Merchandise Trade Statistics Manual (IMTS) 2010-Concepts and Definitions that foreign merchandise trade statistics record all goods which add to or subtract from the stock of material resources of Qatar by entering or leaving the economic territory.

Goods Included in foreign trade statistics are:

- Non-monetary gold.
- Banknotes and securities, and coins not in circulation.
- Goods traded in accordance with barter agreements.
- Goods which cross borders as a result of transactions between related parties.
- Gas and oil.
- Satellites and their launchers.
- Power lines, pipelines and undersea communications cables.
- Used Goods.
- Waste and scrap.
- Goods acquired by all categories of travelers.
- Migrants effects.
- Goods dispatched or received through postal or courier.
- Electricity and water.
- Goods traded on government account.
- Humanitarian aid, including emergency aid.
- Goods for military use.
- Media, whether or not recorded .
- Goods under financial lease.
- Goods received or sent abroad by international organizations.

- Goods delivered to or dispatched from offshore installations.
- Goods transferred from or to a buffer stock organization.
- Goods in electronic commerce.
- Gifts and donations.
- Fish catch, minerals from the seabed and salvage.

The website had given clear statistical data from 2012 until 2020

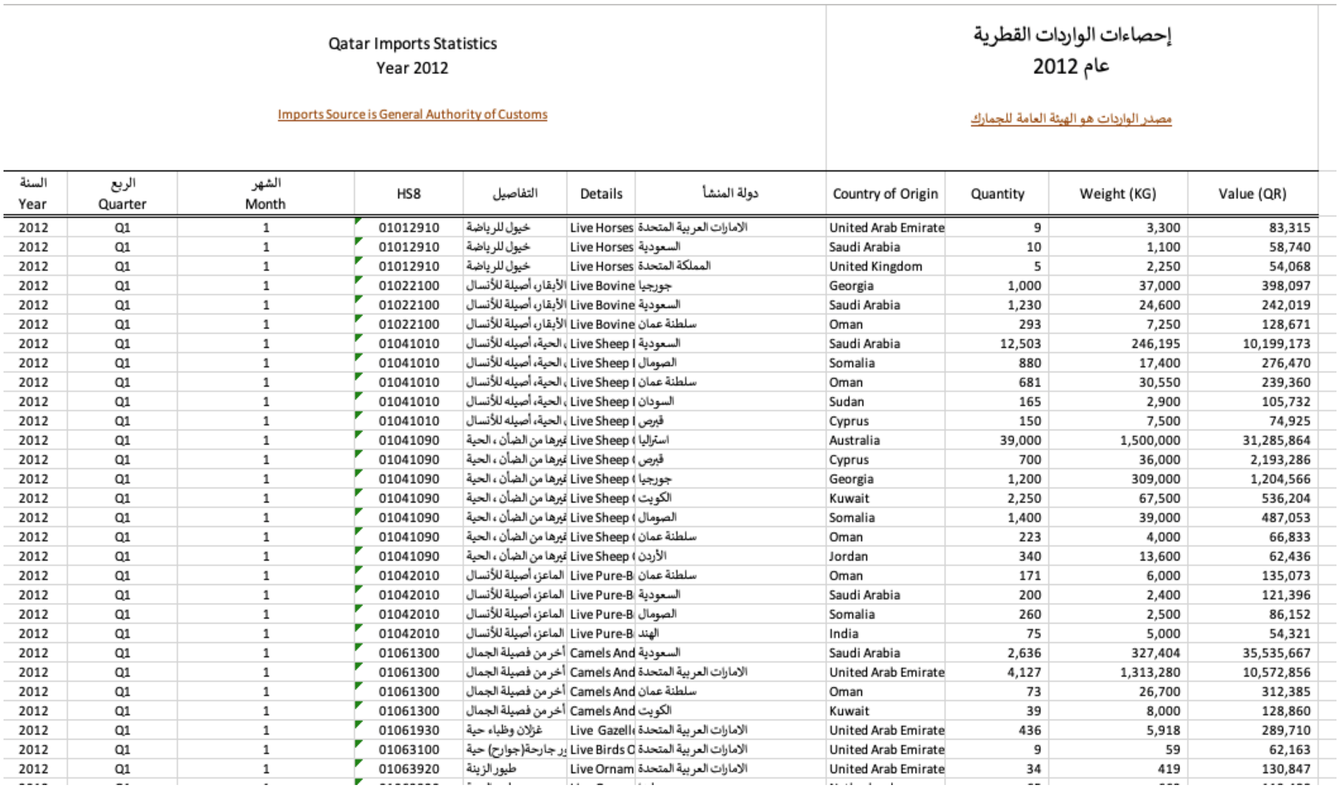


Yearly Statistics

Qatar Statistics Year 2020 →
Qatar Statistics Year 2019 →
Qatar Statistics Year 2018 →
Qatar Statistics Year 2017 →
Qatar Statistics Year 2016 →
Qatar Statistics Year 2015 →
Qatar Statistics Year 2014 →
Qatar Statistics Year 2013 →
Qatar Statistics Year 2012 →

Figure 5: Yearly Statistics from the PSA Website

Below is the screen capture of the first 20 values of the data



Qatar Imports Statistics
Year 2012

Imports Source is General Authority of Customs

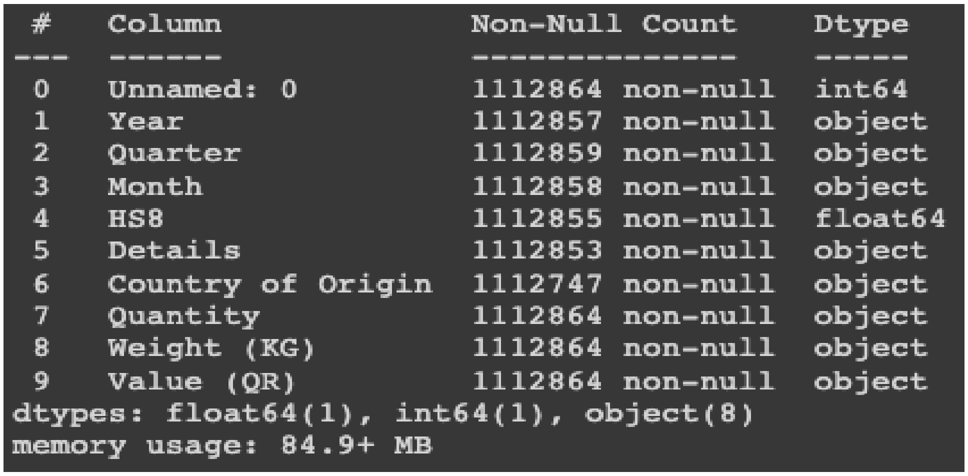
إحصاءات الواردات القطرية
عام 2012

مصدر الواردات هو الهيئة العامة للجمارك

السنة Year	الربع Quarter	الشهر Month	HS8	التفاصيل	Details	دولة المنشأ	Country of Origin	Quantity	Weight (KG)	Value (QR)
2012	Q1	1	01012910	خيول للرياضة	Live Horses	الامارات العربية المتحدة	United Arab Emirate	9	3,300	83,315
2012	Q1	1	01012910	خيول للرياضة	Live Horses	السعودية	Saudi Arabia	10	1,100	58,740
2012	Q1	1	01012910	خيول للرياضة	Live Horses	المملكة المتحدة	United Kingdom	5	2,250	54,068
2012	Q1	1	01022100	الأبقار، أصيلة للأنسال	Live Bovine	جورجيا	Georgia	1,000	37,000	398,097
2012	Q1	1	01022100	الأبقار، أصيلة للأنسال	Live Bovine	السعودية	Saudi Arabia	1,230	24,600	242,019
2012	Q1	1	01022100	الأبقار، أصيلة للأنسال	Live Bovine	سلطنة عمان	Oman	293	7,250	128,671
2012	Q1	1	01041010	الحية، أصيله للأنسال	Live Sheep I	السعودية	Saudi Arabia	12,503	246,195	10,199,173
2012	Q1	1	01041010	الحية، أصيله للأنسال	Live Sheep I	الصومال	Somalia	880	17,400	276,470
2012	Q1	1	01041010	الحية، أصيله للأنسال	Live Sheep I	سلطنة عمان	Oman	681	30,550	239,360
2012	Q1	1	01041010	الحية، أصيله للأنسال	Live Sheep I	السودان	Sudan	165	2,900	105,732
2012	Q1	1	01041010	الحية، أصيله للأنسال	Live Sheep I	قبرص	Cyprus	150	7,500	74,925
2012	Q1	1	01041090	يرها من الضأن ، الحية	Live Sheep (	استراليا	Australia	39,000	1,500,000	31,285,864
2012	Q1	1	01041090	يرها من الضأن ، الحية	Live Sheep (	قبرص	Cyprus	700	36,000	2,193,286
2012	Q1	1	01041090	يرها من الضأن ، الحية	Live Sheep (	جورجيا	Georgia	1,200	309,000	1,204,566
2012	Q1	1	01041090	يرها من الضأن ، الحية	Live Sheep (	الكويت	Kuwait	2,250	67,500	536,204
2012	Q1	1	01041090	يرها من الضأن ، الحية	Live Sheep (	الصومال	Somalia	1,400	39,000	487,053
2012	Q1	1	01041090	يرها من الضأن ، الحية	Live Sheep (	سلطنة عمان	Oman	223	4,000	66,833
2012	Q1	1	01041090	يرها من الضأن ، الحية	Live Sheep (	الأردن	Jordan	340	13,600	62,436
2012	Q1	1	01042010	الماعز، أصيلة للأنسال	Live Pure-B	سلطنة عمان	Oman	171	6,000	135,073
2012	Q1	1	01042010	الماعز، أصيلة للأنسال	Live Pure-B	السعودية	Saudi Arabia	200	2,400	121,396
2012	Q1	1	01042010	الماعز، أصيلة للأنسال	Live Pure-B	الصومال	Somalia	260	2,500	86,152
2012	Q1	1	01042010	الماعز، أصيلة للأنسال	Live Pure-B	الهند	India	75	5,000	54,321
2012	Q1	1	01061300	أخرمن فصيلة الجمال	Camels And	السعودية	Saudi Arabia	2,636	327,404	35,535,667
2012	Q1	1	01061300	أخرمن فصيلة الجمال	Camels And	الامارات العربية المتحدة	United Arab Emirate	4,127	1,313,280	10,572,856
2012	Q1	1	01061300	أخرمن فصيلة الجمال	Camels And	سلطنة عمان	Oman	73	26,700	312,385
2012	Q1	1	01061300	أخرمن فصيلة الجمال	Camels And	الكويت	Kuwait	39	8,000	128,860
2012	Q1	1	01061930	غزلان وظباء حية	Live Gazell	الامارات العربية المتحدة	United Arab Emirate	436	5,918	289,710
2012	Q1	1	01063100	ور جارحة(جوارح) حية	Live Birds O	الامارات العربية المتحدة	United Arab Emirate	9	59	62,163
2012	Q1	1	01063920	طيور الزينة	Live Ornam	الامارات العربية المتحدة	United Arab Emirate	34	419	130,847

Figure 6: Snapshot of Final Data

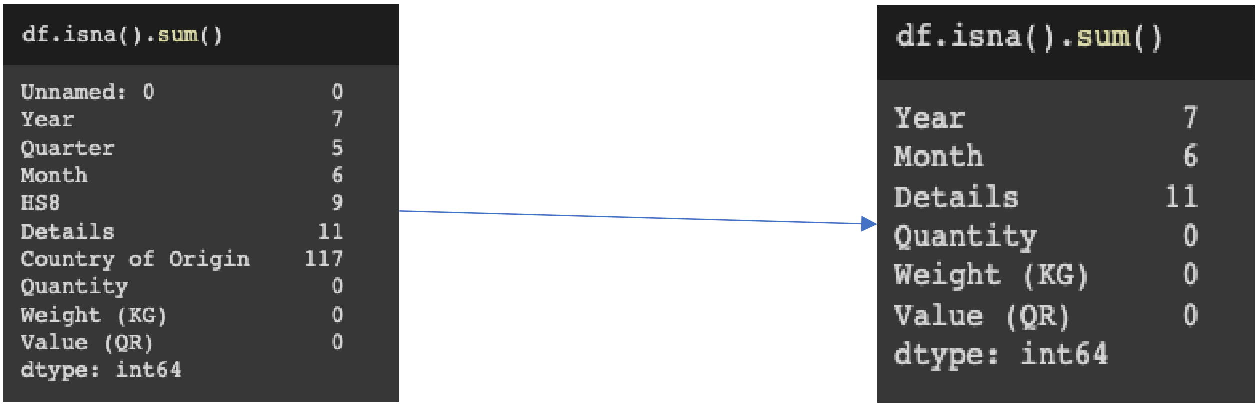
After converting the above .xlsx files to .csv, I've concatenated and combined the multiple csv files into a single csv file. The website psa.gov.qa contained data from the fiscal years 2012 until 2020 in .xlsx format , with the help of Jupyter Notebook and Pandas framework, I started on cleaning the data.



```
 #   Column             Non-Null Count    Dtype
---  ------             --------------    -----
 0   Unnamed: 0         1112864 non-null  int64
 1   Year               1112857 non-null  object
 2   Quarter            1112859 non-null  object
 3   Month              1112858 non-null  object
 4   HS8                1112855 non-null  float64
 5   Details            1112853 non-null  object
 6   Country of Origin  1112747 non-null  object
 7   Quantity           1112864 non-null  object
 8   Weight (KG)        1112864 non-null  object
 9   Value (QR)         1112864 non-null  object
dtypes: float64(1), int64(1), object(8)
memory usage: 84.9+ MB
```

Figure 7: Column Information of the dataset

From the data above we can see that there’s 9 columns of Data, Namely Unnamed:0 which is the S.No , Year , Quarter, Month , HS8, Details, Country of Origin, Quantity, Weight(KG), Value(QR). Primarily I had to find out the number of null values and delete them since some columns have different values.

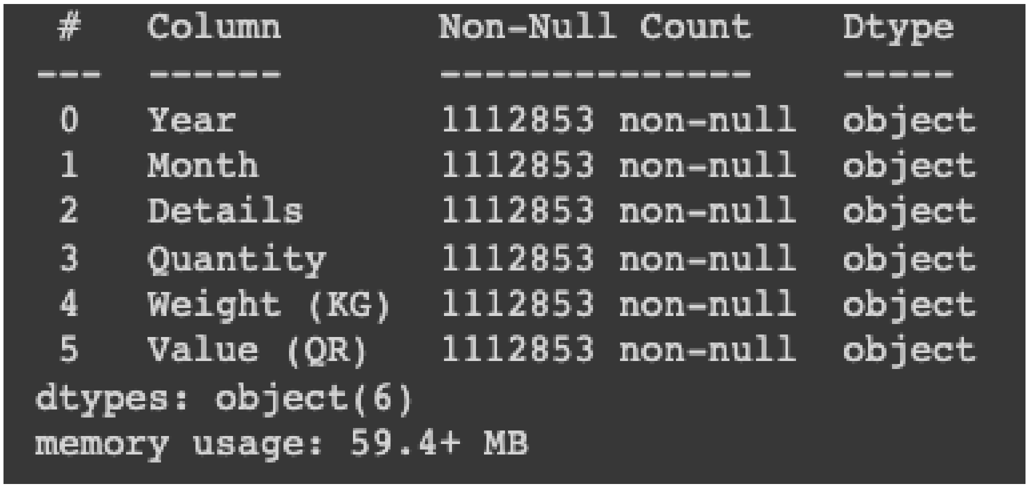


```
df.isna().sum()

Unnamed: 0           0
Year                 7
Quarter              5
Month                6
HS8                  9
Details             11
Country of Origin  117
Quantity             0
Weight (KG)          0
Value (QR)           0
dtype: int64
```

```
df.isna().sum()

Year          7
Month         6
Details      11
Quantity      0
Weight (KG)   0
Value (QR)    0
dtype: int64
```

Figure 8: Data Cleaning

I then dropped null values as they are very minimal compared to the total values (0.0009%) The data is now clear of null values and unwanted columns.



```
 #   Column       Non-Null Count    Dtype
---  ------       --------------    -----
 0   Year         1112853 non-null  object
 1   Month        1112853 non-null  object
 2   Details      1112853 non-null  object
 3   Quantity     1112853 non-null  object
 4   Weight (KG)  1112853 non-null  object
 5   Value (QR)   1112853 non-null  object
dtypes: object(6)
memory usage: 59.4+ MB
```

Figure 9: Final Dataset after Data Cleaning

An outlier is a piece of data that is an abnormal distance from other points. In other words, it's data that lies outside the other values in the set. If you had Pinocchio in a class of children, the length of his nose compared to the other children would be an outlier. STL stands for Seasonal and Trend decomposition using Loess. This is a statistical method of decomposing a Time Series data into 3 components containing seasonality, trend, and residual. With the help of STL Decomposition, we can remove residuals and other outliers to make our data prediction more accurate. Using STL Decomposition we can see that our model has a good trend and seasonal pattern, we then proceed to remove residuals.

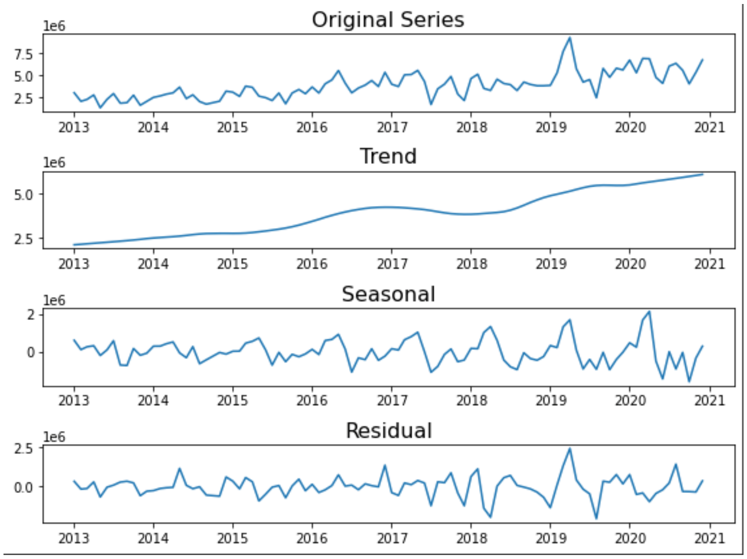


Figure 10: STL Decomposition

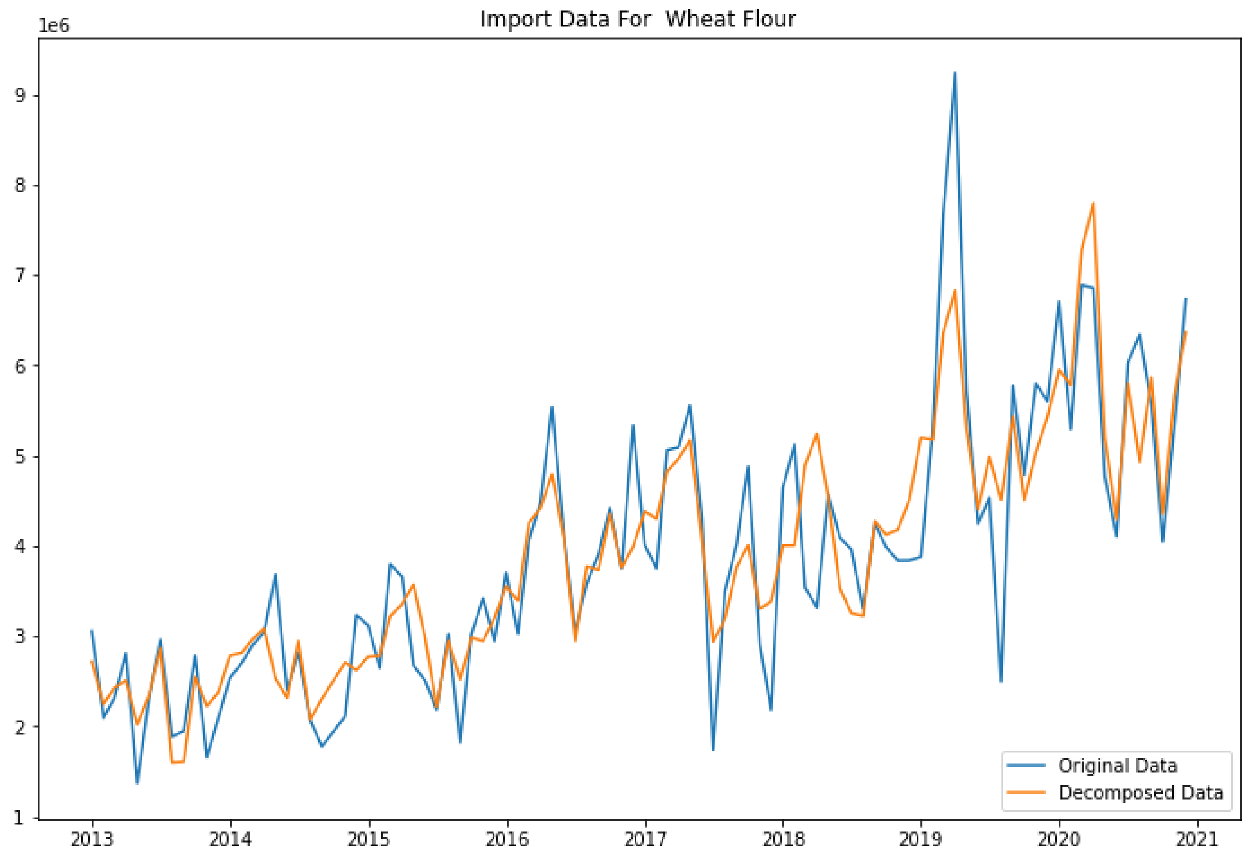
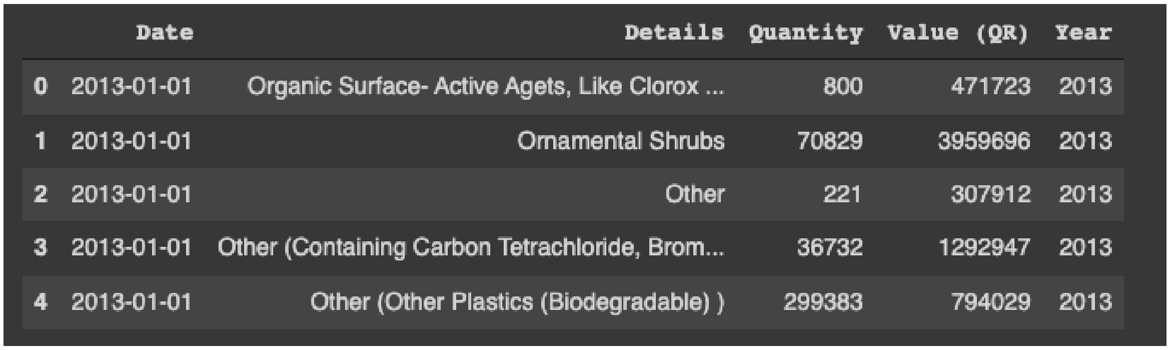


Figure 11: Decomposed Data

The above figure shows a clear representation of the data after removing outliers and residuals from the data. The data is somewhat noisy in the middle range (a very large part of the blue) but the outliers can be clearly seen as a clear sharp spike.

We then used RobustScaler to scale the data to remove more outliers which got across STL Decomposition. Robust Scaler algorithms scale features that are robust to outliers. The method it follows is almost like the MinMax Scaler, but it uses the interquartile range (rather than the min-max used in MinMax Scaler). The median and scales of the data are removed by this scaling algorithm according to the quantile range. Many machine learning estimators need the standardisation of a dataset. This is often accomplished by eliminating the mean and scaling to unit variance. Outliers, on the other hand, can frequently have a negative impact on the sample mean / variance. In such circumstances, the median and interquartile range are frequently more accurate.

In a time-series Data Model, the Date Variable is quite important, since our Data does not have a date time variable, we can combine the Month and Year column under a single Date column which would be a datetime datatype. We had to delete values of 2012 since they contained data for only the first 6 months. The Date Time variable will help us make the model a timeseries model which would help forecasting future trends easier and helpful.



	Date	Details	Quantity	Value (QR)	Year
0	2013-01-01	Organic Surface- Active Agets, Like Clorox ...	800	471723	2013
1	2013-01-01	Ornamental Shrubs	70829	3959696	2013
2	2013-01-01	Other	221	307912	2013
3	2013-01-01	Other (Containing Carbon Tetrachloride, Brom...	36732	1292947	2013
4	2013-01-01	Other (Other Plastics (Biodegradable))	299383	794029	2013

Figure 12: Top 5 rows of the final data

After feature engineering the datetime variable, we can group the data by the commodity. First we can use our model to predict the Import Data trend for Ice Cream.This is the visualized Import Data for Gold Jewelry in Qatar from 2013 until 2021, we can predict use SARIMAX method of predicting the Import trend for Wheat Flour in Qatar

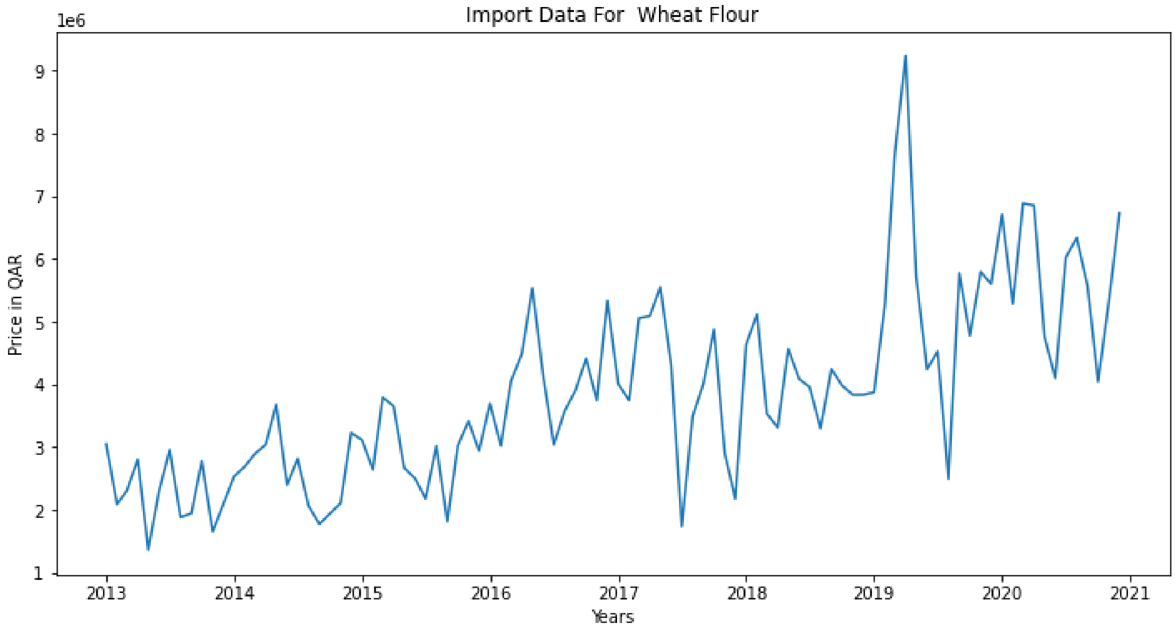


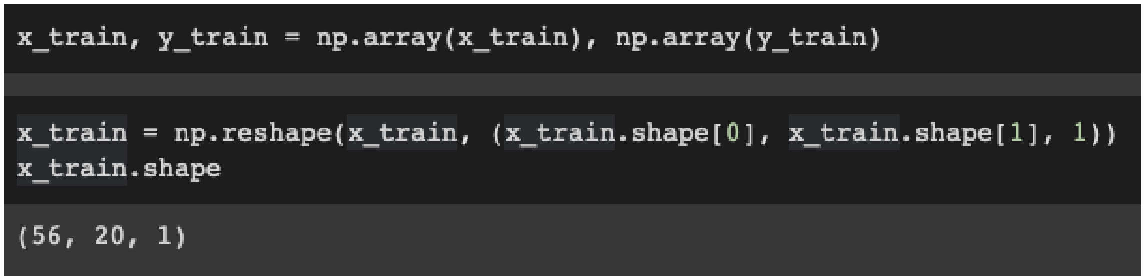
Figure 13: Import Data for Wheat Flour

3.3.3 Data Modeling

We’ve decided to use three different models to find which is the best suitable model for the data. The models we decided to use for the dataset are LSTM, ARIMA and SARIMAX

The first model that we are testing is the LSTM (Long short-term memory) Model. LSTM uses artificial recurrent neural networks (RNN) architecture used in deep learning to predict data. Since LSTM Models can store information over a period, it is extremely useful while dealing with Time Series Data. The second model that we are testing is the ARIMA (Auto regressive Integrated moving average) model. ARIMA models are useful to deal with numerical data and it is used to predict seasonal data like demand for winter coats and seasonal fruits.

The last model that we are testing is the SARIMAX (Sum of the squares auto regressive) Model. SARIMAX models are used to predict trend and seasonality using a linear combination of error terms.

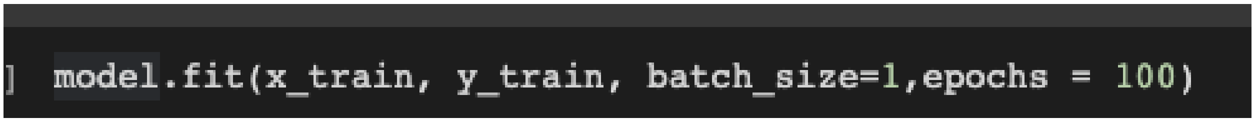


```
x_train, y_train = np.array(x_train), np.array(y_train)
```

```
x_train = np.reshape(x_train, (x_train.shape[0], x_train.shape[1], 1))
x_train.shape
```

```
(56, 20, 1)
```

Before, training our LSTM model we will be login into tensor board dashboard, to record the loss and the best seed value as the model trains over an epoch value of 100. An epoch is a phrase used in machine learning that denotes the number of runs the machine learning model has made across the full training dataset. Typically, datasets are organized into batches. These are also referred to as iterations.



```
]  model.fit(x_train, y_train, batch_size=1,epochs = 100)
```

LSTM stands for Long Short-Term Memory. LSTMs are an advanced version of recurrent neural networks. A recurrent neural network is a network which allows signals to flow in both directions. Recurrent networks are ideally suited to the problem of time series prediction because the order of data in a time series is critical. LSTMs add functionality to recurrent neural networks that allow it to remember information for very long periods of time (i.e., a week, month or even years). This has helped LSTM become extremely popular for many applications involving time series data like speech recognition and stock market prediction. For the LSTM Model we predict after Scaling the data using Robust Scaler and then running 100 epochs to find the model with the least loss.

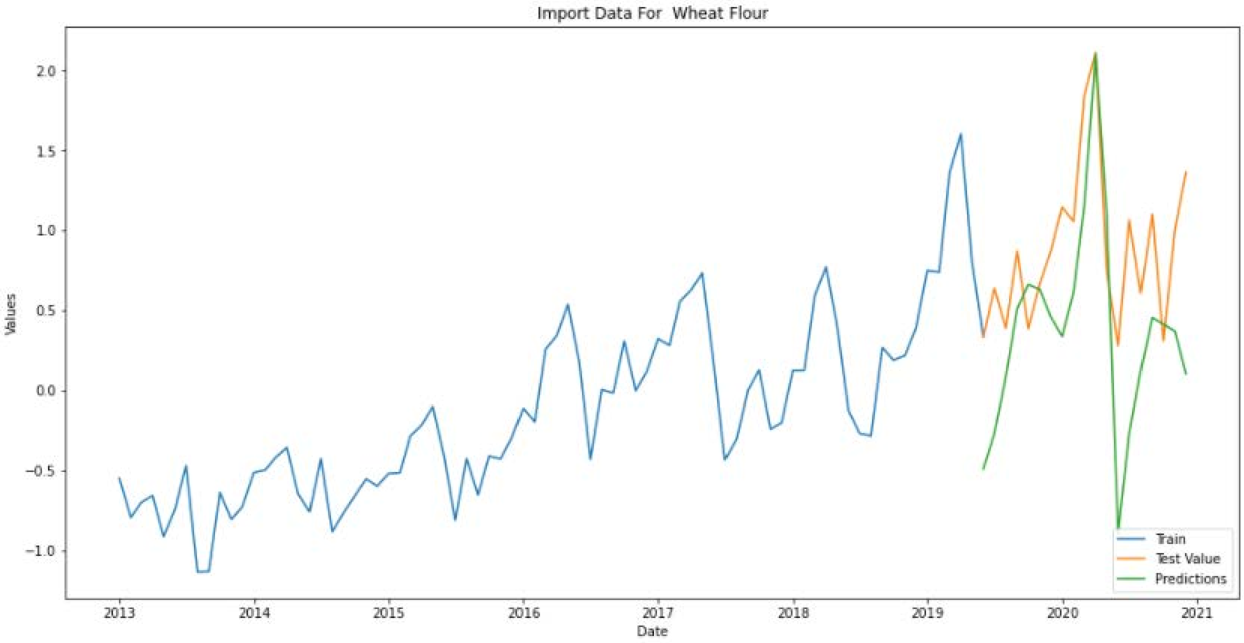


Figure 14: LSTM Model for Wheat Flour Prediction

The R^2 score for LSTM shows the Accuracy to be close to -0.1865. When the model chosen does not follow the trend of the data, the R^2 score is negative, resulting in a worse fit than the horizontal line. When there are limits on either the intercept or the slope of the linear regression line, this is frequently the case. This shows us that the LSTM Model isn't very accurate in predicting the model for Wheat Flour. The Mean Absolute Percent Error for LSTM is 2.145 and the Root Mean Squared Error is 0.697

The next model we're implementing is the ARIMA (Auto regressive Integrated moving average) Model. This is a type of time series that's used when you have a residual or error term. The ARIMA model essentially tries to address the problem of autocorrelation, which is where the eventual value over the period T depends on how previous data points behaved.

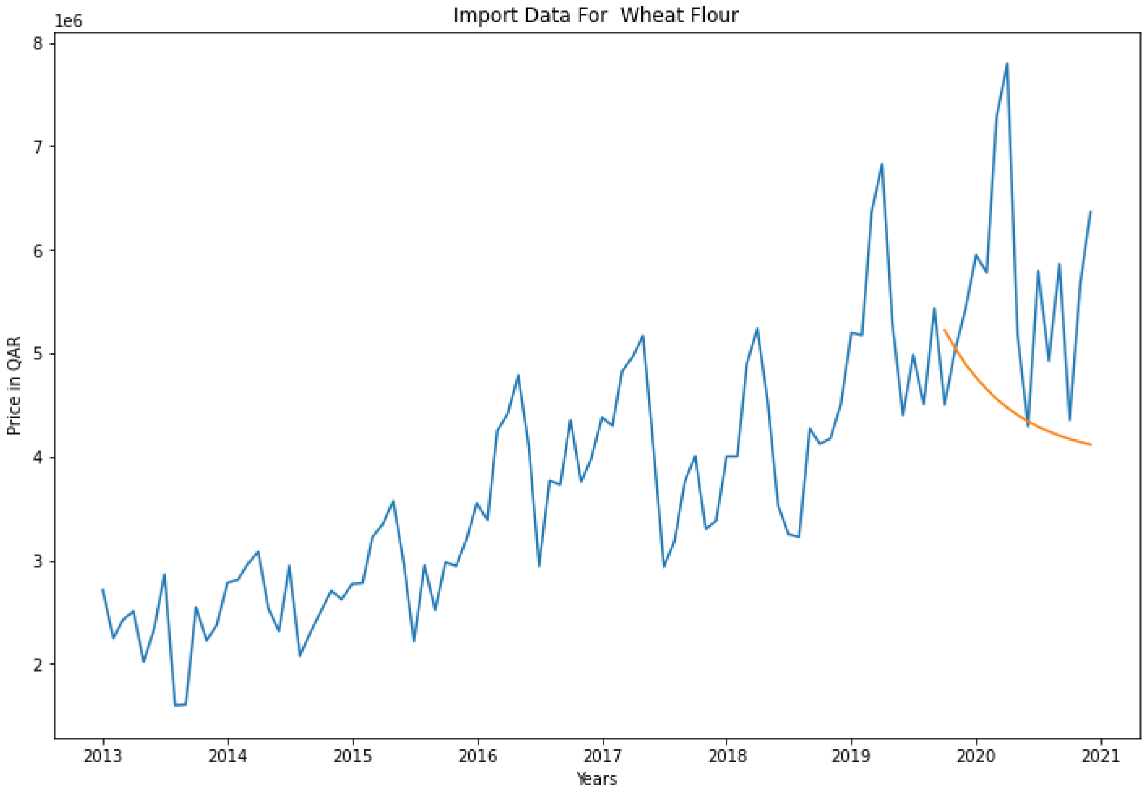
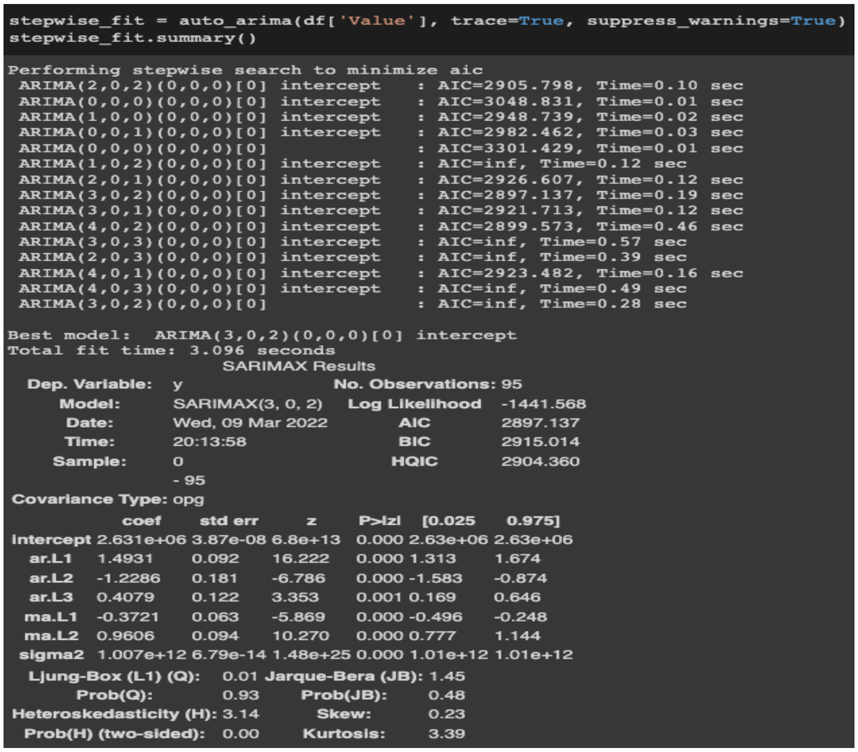


Figure 15: ARIMA Model for Wheat Flour Prediction

The R^2 score for the ARIMA model shows the accuracy to be close to -154.71. When the model chosen does not follow the trend of the data, the R^2 score is negative, resulting in a worse fit than the horizontal line. When there are limits on either the intercept or the slope of the linear regression line, this is frequently the case. This shows us that the ARIMA Model isn't very accurate in predicting the model for Wheat Flour. The Mean Absolute Percent Error for ARIMA is 58.3 and the Root Mean Squared Error is 0.806. A problem with ARIMA is that it does not support seasonal data. That is a time series with a repeating cycle. ARIMA has a limitation in that it does not accommodate seasonal data. That is a time series having a cyclical pattern. ARIMA anticipates data that is either not seasonal or has had the seasonal component removed, e.g., data that has been seasonally adjusted using methods such as seasonal differencing.

The last model we’re going to implement is the SARIMAX model. SARIMAX (Seasonal Auto-Regressive Integrated Moving Average with exogenous factors) is an updated version of the ARIMA model. SARIMAX, like SARIMA and Auto ARIMA, is a seasonal equivalent model. It's also capable of dealing with external influences. This aspect of the model sets it apart from others. It introduces four new hyper-parameters for defining the autoregression (AR), differencing (I), moving average (MA), and support for exogenous variables (X) for the seasonal component of the series, as well as an extra parameter for the seasonality period. The SARIMAX Model for my data produced some really good results. The SARIMA models have p,d,q and r values. Since we’re calculating data monthly, the r given is 12. To find the adequate p,q and q values we use the Auto Arima function that gives us the partial and autocorrelations values and find the adequate p,d and q values.



```
stepwise_fit = auto_arima(df['Value'], trace=True, suppress_warnings=True)
stepwise_fit.summary()
```

```
Performing stepwise search to minimize aic
 ARIMA(2,0,2)(0,0,0)[0] intercept   : AIC=2905.798, Time=0.10 sec
 ARIMA(0,0,0)(0,0,0)[0] intercept   : AIC=3048.831, Time=0.01 sec
 ARIMA(1,0,0)(0,0,0)[0] intercept   : AIC=2948.739, Time=0.02 sec
 ARIMA(0,0,1)(0,0,0)[0] intercept   : AIC=2982.462, Time=0.03 sec
 ARIMA(0,0,0)(0,0,0)[0]             : AIC=3301.429, Time=0.01 sec
 ARIMA(1,0,2)(0,0,0)[0] intercept   : AIC=inf, Time=0.12 sec
 ARIMA(2,0,1)(0,0,0)[0] intercept   : AIC=2926.607, Time=0.12 sec
 ARIMA(3,0,2)(0,0,0)[0] intercept   : AIC=2897.137, Time=0.19 sec
 ARIMA(3,0,1)(0,0,0)[0] intercept   : AIC=2921.713, Time=0.12 sec
 ARIMA(4,0,2)(0,0,0)[0] intercept   : AIC=2899.573, Time=0.46 sec
 ARIMA(3,0,3)(0,0,0)[0] intercept   : AIC=inf, Time=0.57 sec
 ARIMA(2,0,3)(0,0,0)[0] intercept   : AIC=inf, Time=0.39 sec
 ARIMA(4,0,1)(0,0,0)[0] intercept   : AIC=2923.482, Time=0.16 sec
 ARIMA(4,0,3)(0,0,0)[0] intercept   : AIC=inf, Time=0.49 sec
 ARIMA(3,0,2)(0,0,0)[0]             : AIC=inf, Time=0.28 sec

Best model:  ARIMA(3,0,2)(0,0,0)[0] intercept
Total fit time: 3.096 seconds
                          SARIMAX Results
  Dep. Variable:   y                    No. Observations: 95
      Model:       SARIMAX(3, 0, 2)   Log Likelihood   -1441.568
       Date:       Wed, 09 Mar 2022          AIC        2897.137
       Time:       20:13:58                  BIC        2915.014
     Sample:       0                        HQIC        2904.360
                   - 95
 Covariance Type: opg
               coef     std err      z      P>|z|  [0.025    0.975]
 intercept 2.631e+06 3.87e-08 6.8e+13  0.000 2.63e+06 2.63e+06
   ar.L1   1.4931     0.092    16.222   0.000 1.313    1.674
   ar.L2   -1.2286    0.181    -6.786   0.000 -1.583   -0.874
   ar.L3   0.4079     0.122    3.353    0.001 0.169    0.646
   ma.L1   -0.3721    0.063    -5.869   0.000 -0.496   -0.248
   ma.L2   0.9606     0.094    10.270   0.000 0.777    1.144
  sigma2   1.007e+12 6.79e-14 1.48e+25 0.000 1.01e+12 1.01e+12
  Ljung-Box (L1) (Q):    0.01 Jarque-Bera (JB): 1.45
        Prob(Q):         0.93     Prob(JB):     0.48
 Heteroskedasticity (H): 3.14       Skew:       0.23
  Prob(H) (two-sided):   0.00     Kurtosis:     3.39
```

Figure 16: Auto ARIMA Function

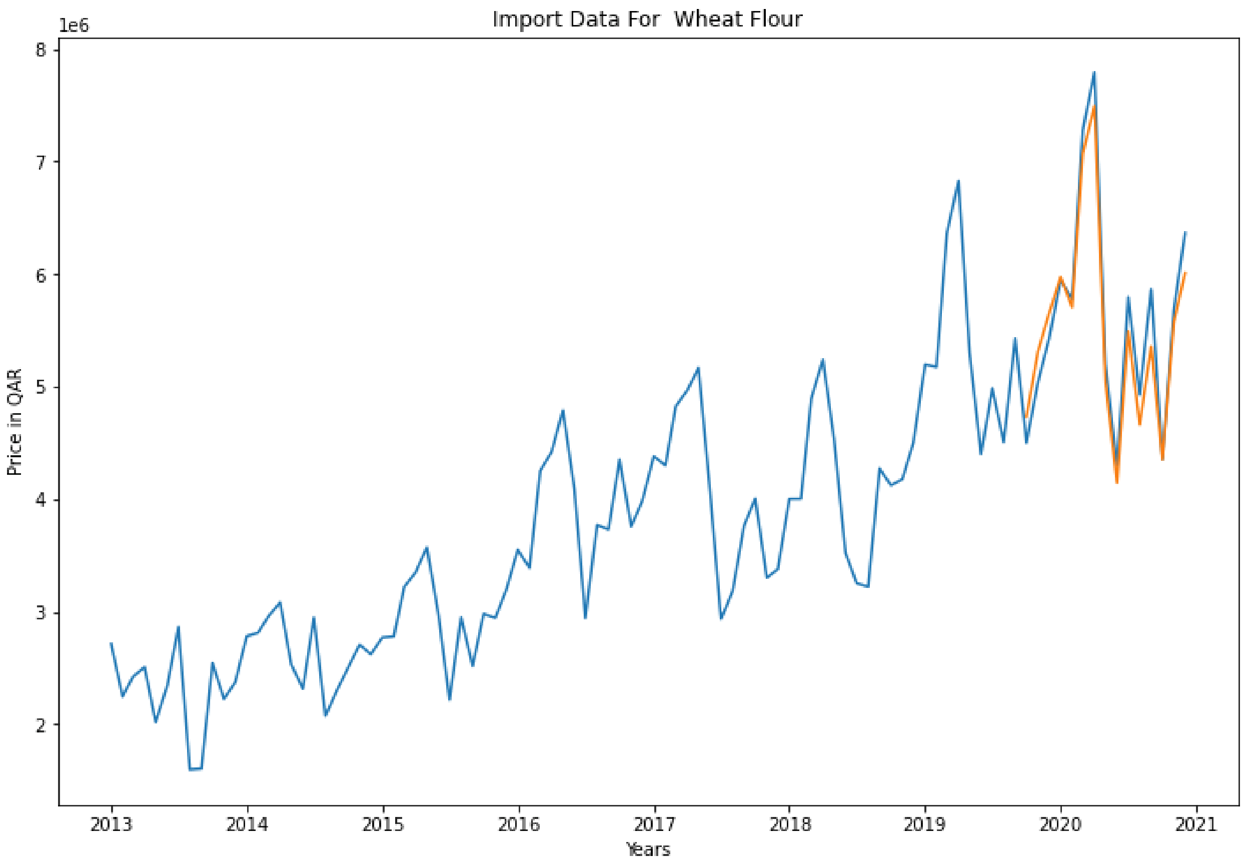


Figure 17: SARIMAX Model for Wheat Flour Prediction

The R^2 score for the SARIMAX model shows the accuracy to be close to 93.34% which is very accurate for our model. The SARIMAX model performed well to the data of the commodity 'Wheat Flour' . SARIMAX is used on data sets that have seasonal cycles. The difference between ARIMA and SARIMAX is the seasonality and exogenous factors. SARIMAX is much like ARIMA, but a little more complicated. Not only do you have to use a loop and grid search for the optimal values of p, d, and q, but you must also use a nested loop and grid search for the seasonal values for p, d, and q. There are also many more parameters in the SARIMAX function.

To verify the accuracy of the model we test the SARIMAX model with different commodities and calculate their R^2 score.

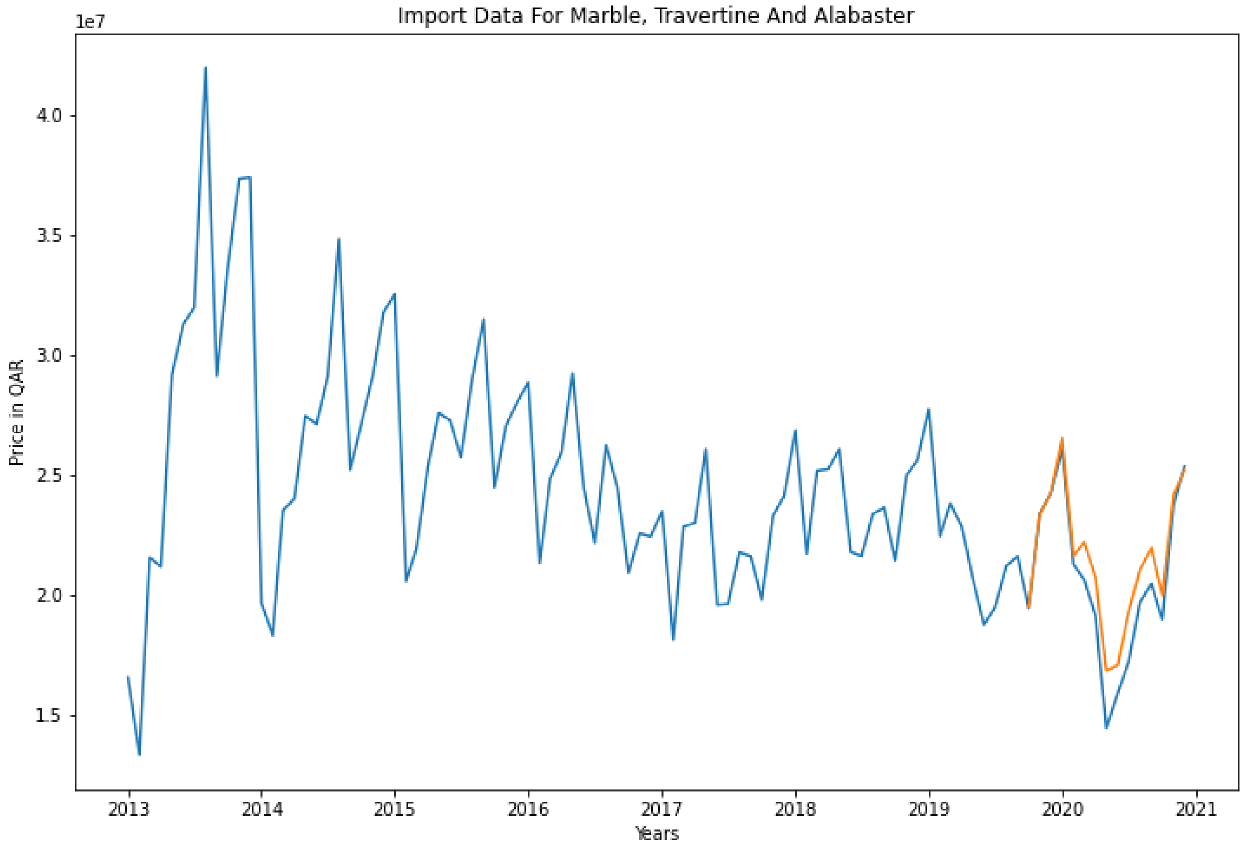


Figure 18: SARIMAX Model for Marble, Travertine and Alabaster

The R^2 score for the SARIMAX model shows the accuracy to be close to 86.42% which is very accurate for our model. The SARIMAX model performed well to the data of the commodity 'Marble, Travertine and Alabaster'.

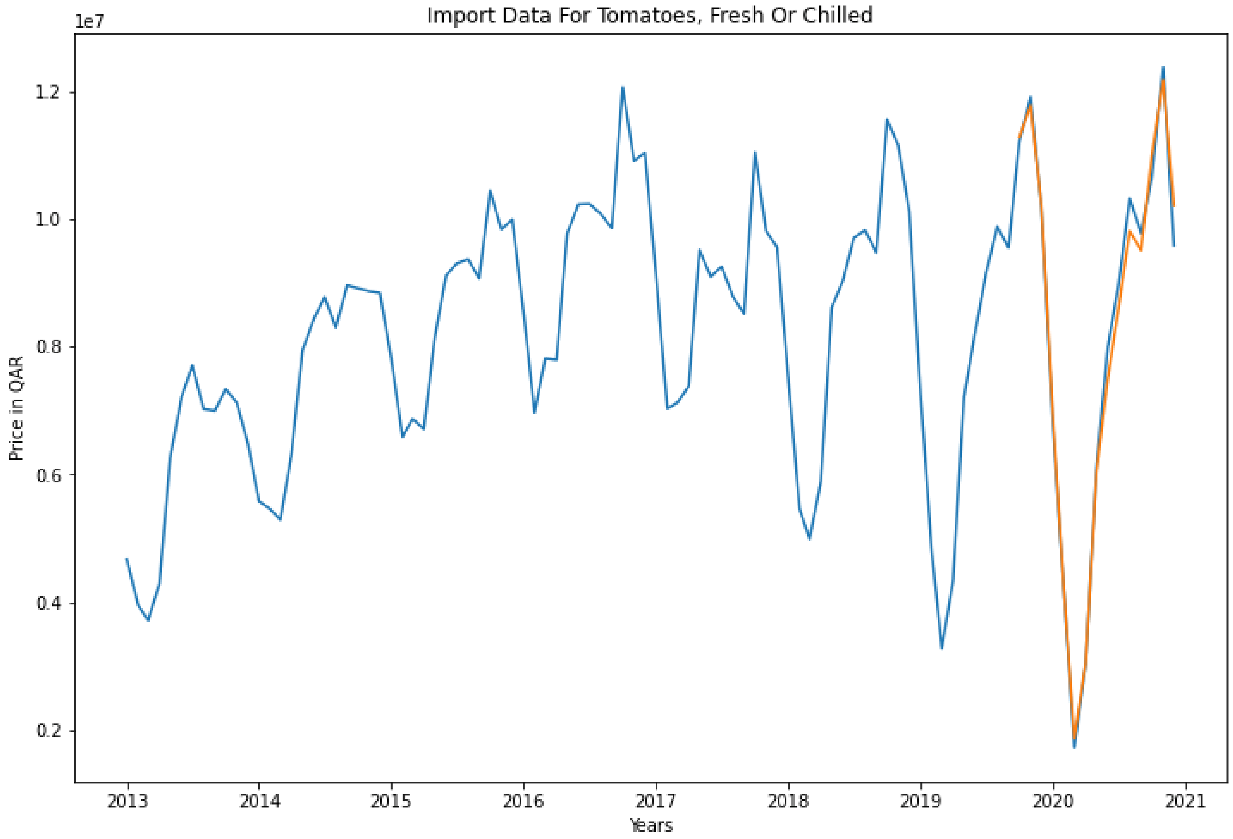


Figure 19: SARIMAX Model for Tomatoes, Fresh Or Chilled

The R^2 score for the SARIMAX model shows the accuracy to be close to 99.10% which is extremely accurate for our model. The SARIMAX model performed well to the data of the commodity ‘Tomatoes, Fresh Or Chilled’.

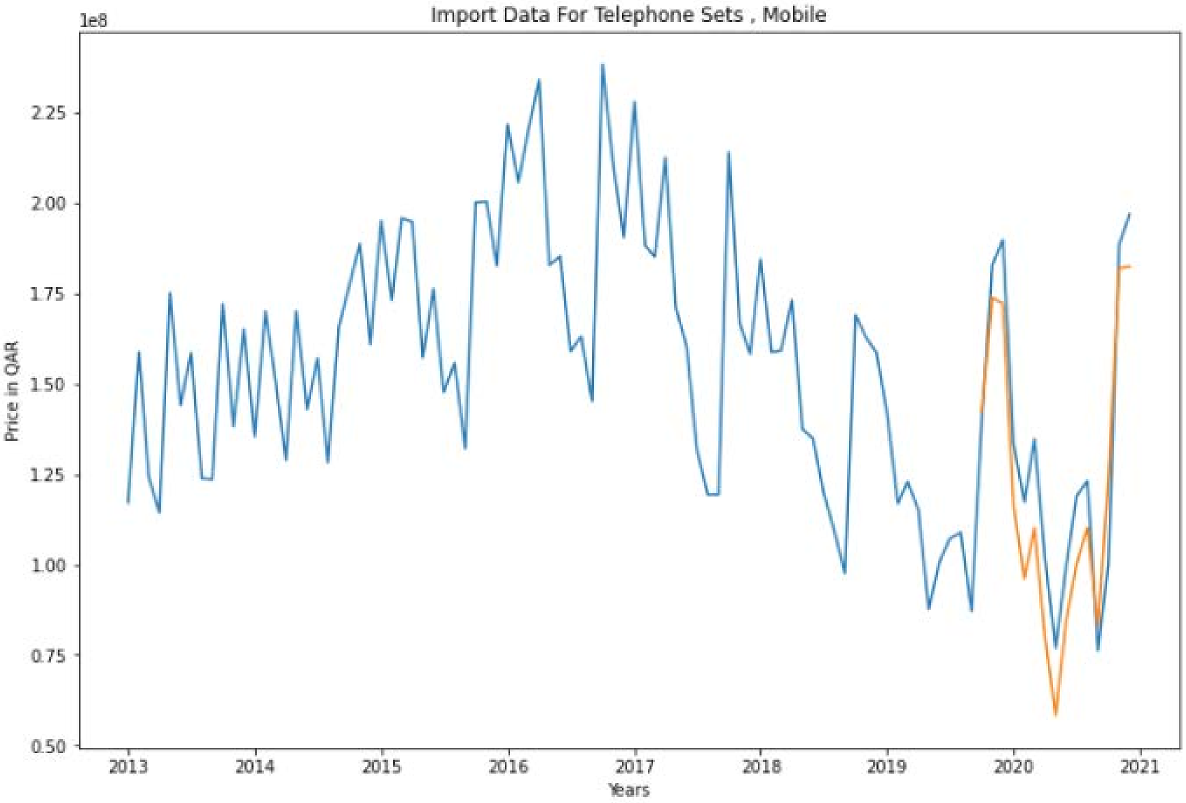


Figure 20: SARIMAX Model for Telephone Sets

The R2 score for the SARIMAX model shows the accuracy to be close to 81.92% which is extremely accurate for our model. The SARIMAX model performed well to the data of the commodity ‘Telephone Sets’. The R2 score is calculated, showing a very high percentage of accuracy in all variables and coefficients.

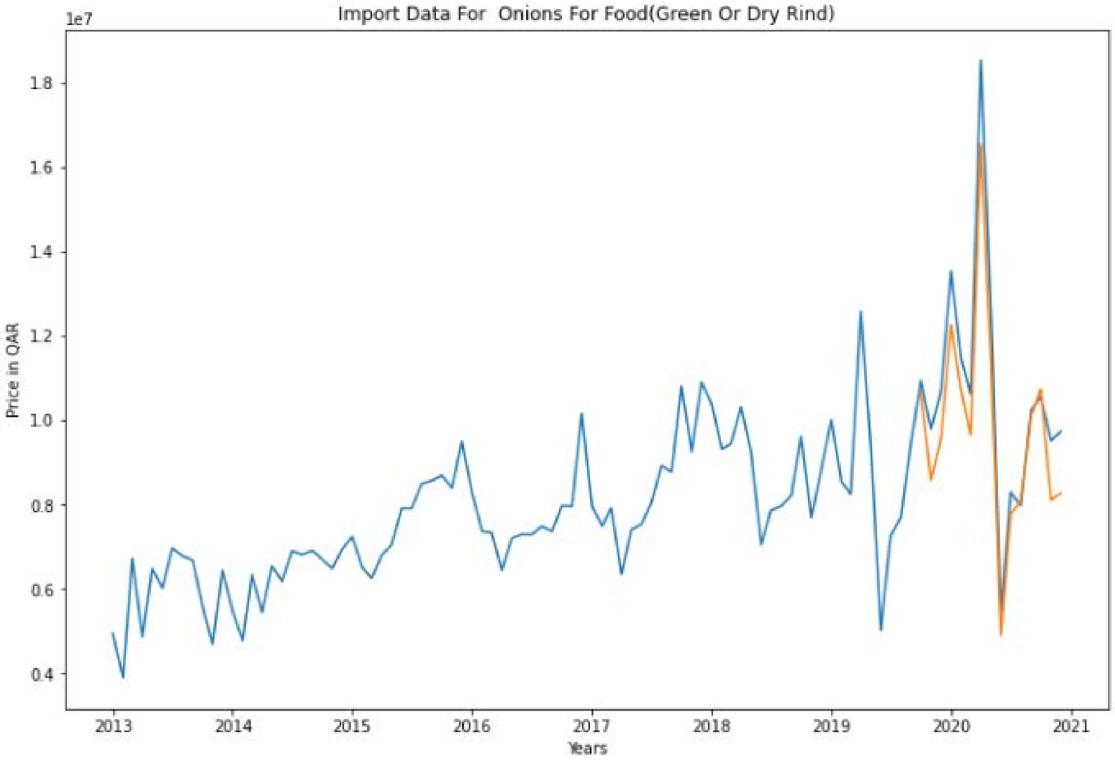


Figure 21: SARIMAX Model for Onions

The R^2 score for the SARIMAX model shows the accuracy to be close to 86.42% which is extremely accurate for our model. The SARIMAX model performed well to the data of the commodity Onions.

The four different models give us accuracies of 93.34, 86.42, 99.1,81.92 and 86.42. It also implies that perishable goods like fruits and vegetable have predictable seasonal data where commodities like mobile phones have lesser accuracy in prediction. The SARIMAX model was also found to have a low standard error. This is because the variables came from a large, varied set of data and the nonlinearity of the model tends to smooth out many of the errors within them. The use of non-linearity allows for a much more accurate model, which is helped by the high mean absolute error (MAE). The MAE in this case works in favor of the SARIMAX model because it can show that complex relationships are present between all variables.

After the model is evaluated and a satisfactory result is obtained the last phase of the framework would be implemented which is the deployment phase.

The table below shows the commodity name and their respective R^2 Score of each model.

S.NO	Commodity Name	R^2 Score
LSTM Model		
1	Wheat Flour	-0.0745
2	Marble, Travertine and Alabaster	-0.8952
3	Tomatoes	69%
4	Cell Phones	-31.22
5	Onions	-2.178
ARIMA Model		
1	Wheat Flour	-1.547
2	Marble, Travertine and Alabaster	-1.035
3	Tomatoes	0.0004
4	Cell Phones	-0.4374
5	Onions	-0.7419
SARIMAX Model		
1	Wheat Flour	93.34%
2	Marble, Travertine and Alabaster	86.42%
3	Tomatoes	99.108%
4	Cell Phones	81.922%
5	Onions	86.43%

3.3.4 Deployment Phase

The application would be deployed for usage by the target market after an iterative process of data preparation and modelling that leads to evaluation. Monitoring and maintenance become vital if the data mining results become part of the day-to-day operations of the company and its surrounds. Preparing a maintenance strategy ahead of time will help you avoid using erroneous data mining results for long periods of time. Model deployment as a service is the best way to deploy the completed project; nevertheless, if time and resources are available, deployment is surely achievable. Predicting the import and export trend of different commodities can really help the suppliers and business owners in analyzing the trend and make the good move for their business. And with the help of data mining, it helps the software engineers in finding out possible importing and exporting product to new markets.

3.4 Tools used

3.4.1 Google Colab

Figure 22: Google Colab Logo

Google Colaboratory is a free Jupyter notebook environment that runs on Google's cloud servers and allows users to take advantage of backend hardware such as GPUs and TPUs. Colab notebooks let you blend executable code and rich text, as well as graphics, HTML, LaTeX, and more, in a single document. Colaboratory supports Python 3 and has a REST API for accessing data. The best benefit of Colab notebooks is that it can be shared and run in the cloud, on mobile devices, and on local machines.

3.4.2 Python Language

Figure 23: Python Logo

Python is a scripting language that is high-level, interpreted, interactive, and object-oriented. Python is intended to be extremely readable. It commonly employs English terms rather than punctuation, and it has fewer syntactical structures than other languages. Python was mainly designed to be easy to understand and easy to modify (remodel). Python interpreters are available for many operating systems.

3.4.3 Statsmodels Framework

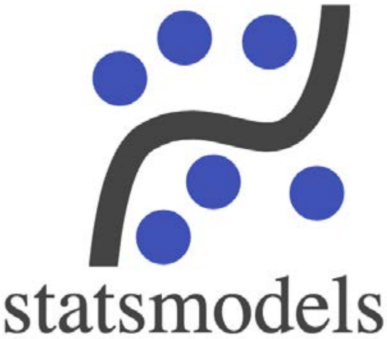


Figure 24: statsmodels Logo

Statsmodels is a Python package that lets you explore data, estimate statistical models, and run statistical tests. Statsmodels is built on NumPy, SciPy, and matplotlib, but it includes more complex statistical testing and modelling capabilities that you won't find in NumPy or SciPy. For each type of data and estimator, a comprehensive set of descriptive statistics, statistical tests, charting tools, and outcome statistics is given. We will be using statsmodel to implement the SARIMAX and the ARIMA model

3.4.4 Scikit-learn Framework

Figure 25: scikit-learn Logo

Scikit-learn is a free machine learning library for Python. It features various algorithms like support vector machine, random forests, and k-neighbors, and it also supports Python numerical and scientific libraries like NumPy and SciPy. Scikit-learn is available in many Linux distributions and it can also be installed on Windows, Mac OS X and R. We will be using sklearn to implement feature scaling into our model and to implement LSTM into the model

Chapter 4: Results and Discussions

In this project, we have used Machine Learning to help predict procurement data which would help in analyzing and forecasting future data trends so the procurement team could effectively plan accordingly. The data used here was a representative sample of how the actual data would look like. We did not want to overfit the model and cause it to be non-generalizable as well as we wanted the model to be able to make accurate predictions on unseen cases. My goal was not only to optimize performance but also get an idea on the effectiveness of using Artificial Intelligence and Machine Learning in predicting procurement data.

4.1 Project Limitations

Lack of reliable data – Most of the data for procurement are classified since companies do not let procurement data out due to its proprietary nature. This prevented me from getting a large sample of clean data which could have helped optimize the performance of my model and potentially improve results. Even though I was able to source data, some of the data points were corrupted. This problem could have been solved by getting more reliable data which would have probably resulted in better performance for the model.

No access to historical bidding data – Bid data is not available since it is proprietary information. This prevented me from conducting a test on actual data and lead to use of simulated data to train the model.

Lack of Accuracy parameters - Since my problem was to use regression on a time-series data, I couldn't use common accuracy parameters like accuracy_score. For this project I've used R2 score but R2 score does not clearly calculate the accuracies for models that do not fit. Therefore, I am using the average of Mean Absolute percentage error (MAPE), Mean Absolute Error (MAE) and MinMax error instead.

Chapter 5: Conclusion and Recommendations

5.1 Conclusion

The objective of this project was to predict trends in procurement data. The data for this project was the import data for Qatar from 2013 until 2020. It can be stated that the objective of the project was achieved. We had to eliminate some outliers because most of the outliers are external factors that machine learning can't predict. For example, in all the charts above you can see there has been a sharp fall in imports of commodities in mid-2017. Saudi Arabia, the United Arab Emirates, Yemen, Egypt, the Maldives, and Bahrain all declared independently on June 5 and 6, 2017 that they were breaking diplomatic ties with Qatar owing to diplomatic concerns. Saudi Arabia has shut down its border with Qatar. Qatar's only land border was with Saudi Arabia, which was a huge setback for commodities imports into the nation.

The experimental work in this study shows that machine learning is very relevant for forecasting a variety of trade patterns with more accuracy than traditional methodologies. Machine learning models will come to the rescue when traditional methods fail to categorize data adequately. The differentiation between various categories of goods which are classified under various sub-categories is also useful in understanding the processes of trade. The use of this model also helps in monitoring and predicting macroeconomic events in the country with their potential consequences on import and export volumes. This can be used for the purpose of forecasting future trends in import and export volumes to make the economy more efficient.

In conclusion, it is not possible to predict sudden rises and falls in the imports data by countries around the world. These are factoring that machine learning can't predict. However, we have managed to make a predict the sales trend of commodities without external environment factors affecting it

5.2 Recommendations

Our model can be made more accurate if the data received was more accurate and reliable, the data had corrupted and missing values, moreover if the data had more exogenous variables, the accuracy of the model can be increased drastically. If a new variable is added to the model and if it is having strong correlation with any of the other variables, then it can be a very good variable to be included as an exogenous variable. Production of new variables for the model will help us predict the probability of it occurring. This can also help us study other economic indicators and predict future trends.

Ranking of variables by their standard deviation in different industries is useful for the model to be more accurate. It will also be very useful for the modeler to determine if any of the variables used in modeling is having a strong correlation with another variable that should not be included in the training data. This will enable us to predict which variables are having strong correlation with other variables and which ones are not.

The model can be made more accurate by using a neural network which is capable of accurately modeling complex non-linear relationships. We had only one variable for import price prediction. If there were many variables included in our model, the accuracy of prediction relies on the model being able to capture these complex non-linear relationships. To make the model more accurate, we must develop a better understanding of these complex non-linear relationships through experimentation and data analysis. The best way to understand these complex non-linear relationships is to use a neural network.

References

BBC. (2017, July 19). *Qatar crisis: What you need to know*. BBC News. Retrieved March 14, 2022, from https://www.bbc.com/news/world-middle-east-40173757

Biedron, W. by R. (2021, December 13). *Machine learning in Procurement*. PLANERGY Software. Retrieved March 14, 2022, from https://planergy.com/blog/machine-learning-in-procurement/

Eneldo Loza Mencía Knowledge Engineering Group, Mencía, E. L., Group, K. E., Group, S. H. K. E., Holthausen, S., Lab, A. S. T., Schulz, A., Lab, T., Group, F. J. K. E., Janssen, F., Bari, U. of, Economics, U. of, Economics, P. U. of, & Metrics, O. M. V. A. (2013, September 1). *Using data mining on linked open data for analyzing e-procurement information: Proceedings of the 2013 International Conference on data mining on Linked Data - volume 1082*. Guide Proceedings. Retrieved March 14, 2022, from https://dl.acm.org/doi/10.5555/3053776.3053785

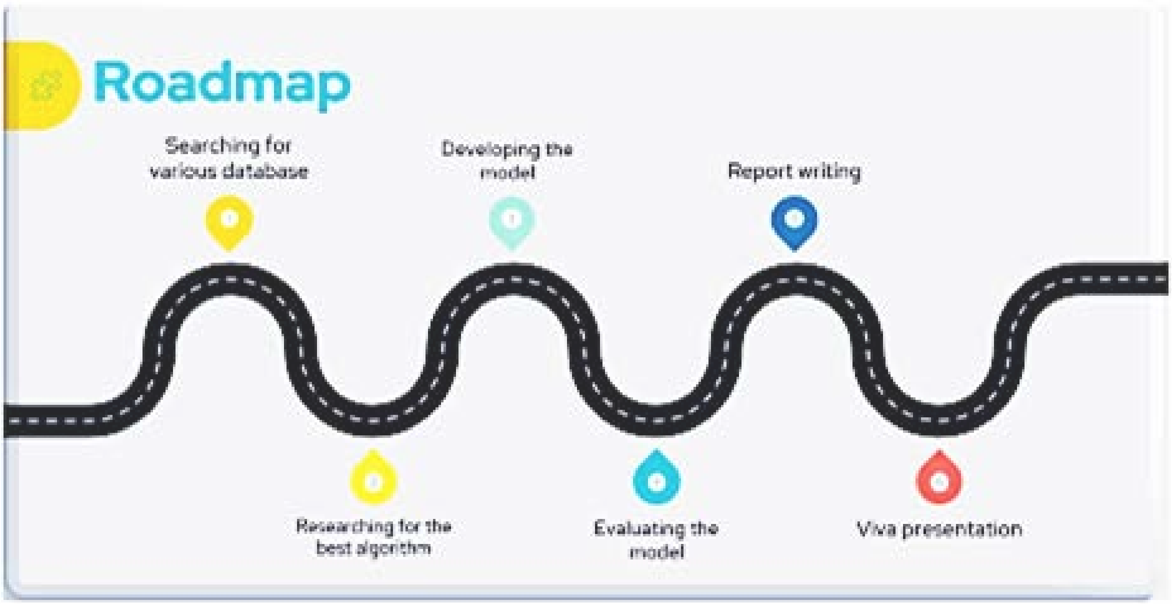
Key contacts Stephen Resar National Supply Chain & Network Operations Leader sresar@deloitte.ca . (2020, May 20). *The AI opportunity in sourcing and Procurement*. Deloitte Canada. Retrieved March 14, 2022, from https://www2.deloitte.com/ca/en/pages/deloitte-analytics/articles/ai-opportunity-sourcing-procurement.html

Suler, P., Rowland, Z., & Krulicky, T. (2021). Evaluation of the accuracy of machine learning predictions of the Czech Republic's exports to the China. *Journal of Risk and Financial Management*, *14*(2), 76. https://doi.org/10.3390/jrfm14020076

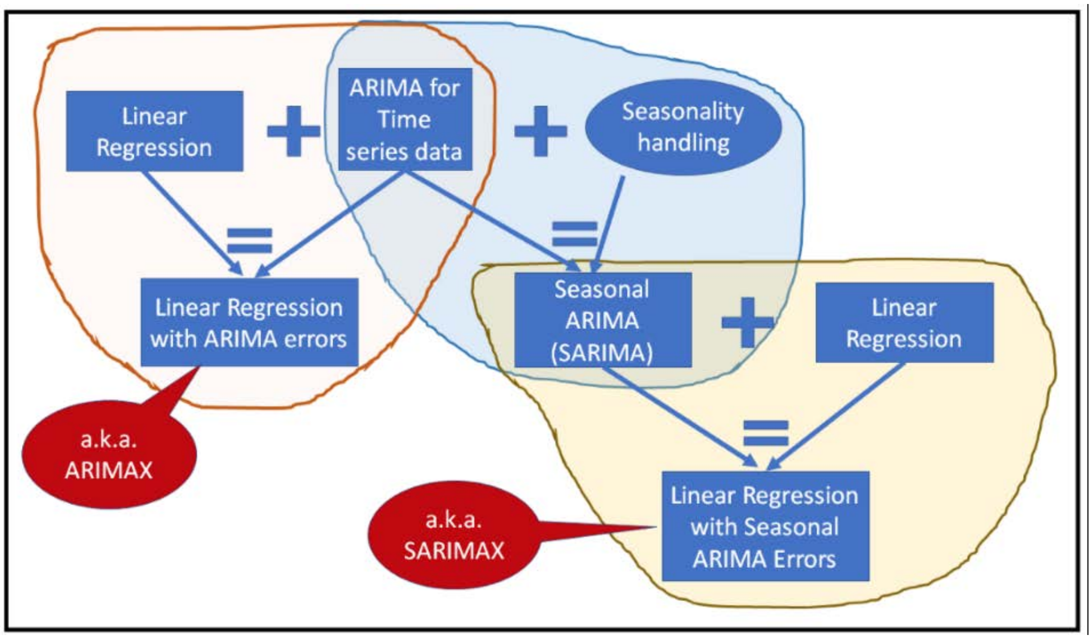
Feras Batarseh & Munisamy Gopinath & Ganesh Nalluru & Jayson Beckman. (2019, February 2). *Application of machine learning in forecasting international trade tre*. Papers. Retrieved March 14, 2022, from https://ideas.repec.org/p/arx/papers/1910.03112.html

Qatar Monthly Statistics. Planning and Statistics Authority Home Page. (n.d.). Retrieved March 14, 2022, from https://www.psa.gov.qa/en/Pages/default.aspx

Appendix



Appendix 1: Roadmap



Appendix 2: Visual Representation of SARIMAX and ARIMA