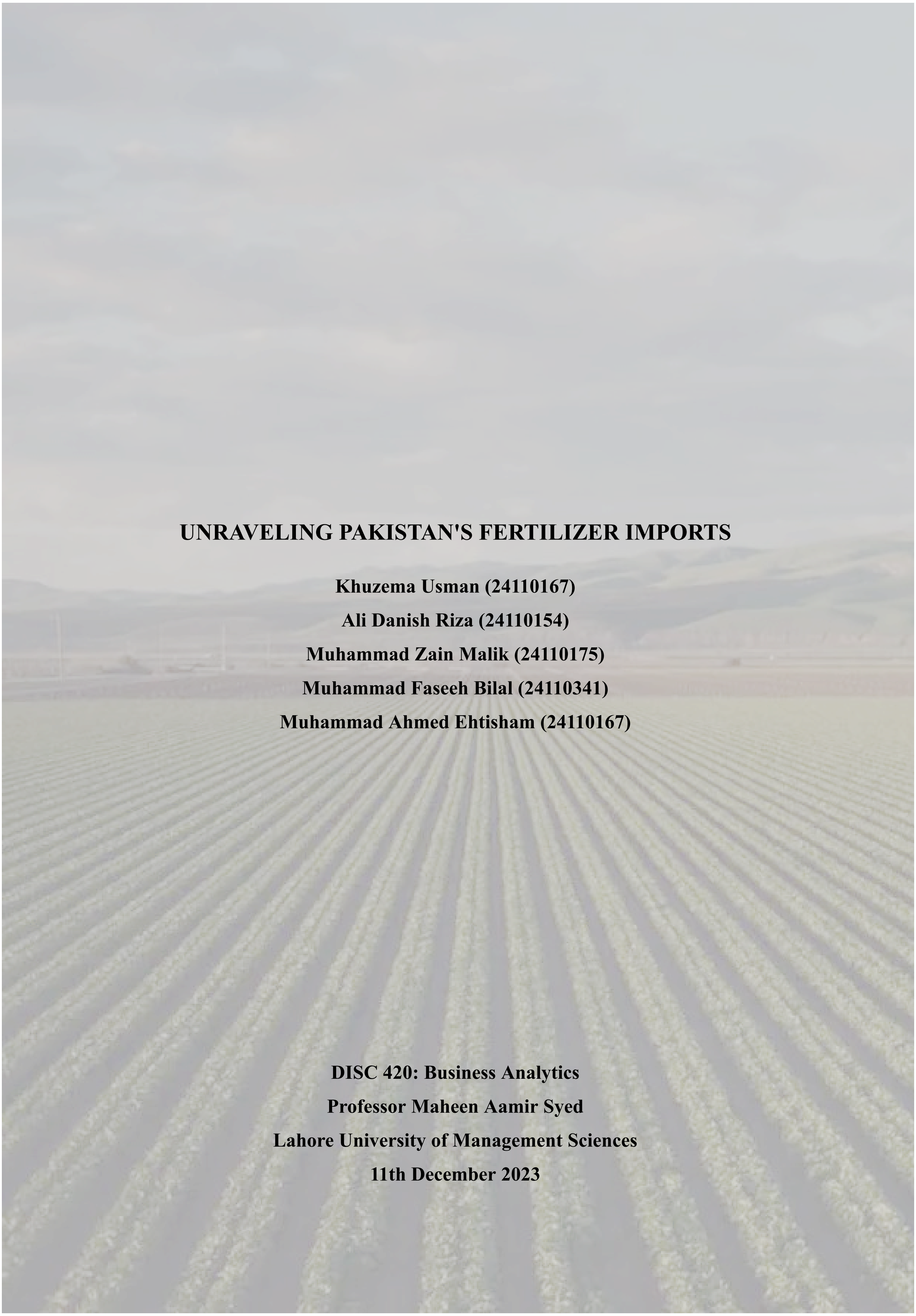


UNRAVELING PAKISTAN'S FERTILIZER IMPORTS

Khuzema Usman (24110167)

Ali Danish Riza (24110154)

Muhammad Zain Malik (24110175)

Muhammad Faseeh Bilal (24110341)

Muhammad Ahmed Ehtisham (24110167)

DISC 420: Business Analytics

Professor Maheen Aamir Syed

Lahore University of Management Sciences

11th December 2023

Table of Contents

1. Introduction

Pakistan relies heavily on import of fertilizers for its agricultural sustainability which helps the country meet the demands of an expanding population and keep steady crop production. The main idea behind this project is to examine the underlying trends of Pakistan’s fertilizer imports and make out how well they can forecast imports required. This research explores various literature works like PACRA’s and Tariq, Farid & others (2016), Overview of the Fertilizer Sector (2021), which reveals the relationships between seasonal amounts of imports and demand. The research recognizes the complex nexus of regional agricultural cycles, national and international trade complexities and financials which come into play in the import of fertilizers.

In addition, our project examines a dataset which has 23 variables in total, of which 8 are relevant to our scope of exploration. The primary dataset contains variables including HS Codes, item descriptions, country of origin, consigner names and importer, shipping value, quantity and shipment dates. Specific addresses, tax information, and port-specifics are examples of irrelevant variables that are removed from the analysis. Our null hypothesis refutes the presence of any patterns in the import trends while also casting doubt on the potential patterns too which could be used to predict import quantities. On the other hand, the alternative hypothesis suggests that the data contain significant structures that indicate possible correlations between important factors and the amounts of fertilizer imported. The use of sophisticated statistical models on the dataset, historical import data, macroeconomic indicators and the use of AI to address these claims. Our main motive is to close the knowledge gaps by examining unexplored areas in predictive modeling for these imports. The results are aimed at offering deep insights to stakeholders of the sector as whole.

2. Literature Review

Pakistan’s agricultural dynamics have seen remarkable development and adaptation, adopting cutting-edge techniques and technology to boost output. But despite these advancements, fertilizers still hold a very important position to boost yields. Pakistan’s dependence on fertilizer imports is a reflection of the complex interactions between internal and foreign policy, as well as the changing agricultural dynamics. These imports are influenced by crop seasons, which include the planting and harvesting cycles of staple crops throughout the year. Pakistan's imports are affected by price volatility in the global market, which comes up by variables such as raw material pricing, disruptions in the supply chain, and changes in the political spectrum. Moreover, Pakistan’s vulnerability to weather fluctuations, such as monsoons and droughts, can have a substantial impact on agricultural practices. Hence, it is critical for stakeholders in the agriculture sector to comprehend the complex interaction between these diverse elements.

Let's now explore some of the research that has been conducted on this issue. During a 20-year period, the first research we looked at provides important insights on the rising demand and usage of inorganic fertilizers like urea, DAP, SOP, and NOP. It hinted at potential areas where patterns might exist. The study touches upon seasonal variations in fertilizer consumption during the Rabi and Kharif seasons, hinting at varying fertilizer needs aligned with different crop planting periods.(Tariq, Farid & others, 2016) These variations potentially suggest cyclical patterns in fertilizer imports concerning specific agricultural seasons.Moreover, the study emphasized the positive correlation between fertilizer usage and crop productivity, indicating a possible relationship that could be leveraged for predictive purposes, albeit without detailed exploration into predictive modeling for import quantities, which is an area our research wishes to build upon. To comprehensively assess the extent of inherent patterns in fertilizer imports and their predictive utility, further analysis involving detailed historical import data, seasonal trends, and statistical modeling techniques would be necessary. Such an approach could ascertain the presence and effectiveness of these patterns.

The second work we examined is the Fertilizer Sector Overview, which provides us with distinct patterns of fertilizer imports in Pakistan, offering insights for predicting import quantities effectively. As stated above, the nation heavily relies on specific fertilizer types such as Urea, DAP, SOP, NPK, and NOP. Urea, in particular, maintains a dominant position in the market (PACRA, 2021). Seasonal variations during crop seasons (Rabi and Kharif) significantly impact fertilizer demands, notably urea, which is evenly utilized in both seasons. Keeping an eye on opening and closing inventory levels throughout the course of many crop cycles gives indications about future import requirements. Policies, subsidies, and modifications to tax laws all have a significant impact on fertilizer pricing, which in turn affects the choices and amounts of imports. Furthermore, domestic production capacities are greatly impacted by the price and availability of raw resources, especially natural gas, which may necessitate imports (PACRA, 2021). Accurately forecasting fertilizer import quantities over a range of timeframes can be facilitated by developing predictive models that take into account past data on crop yields, government regulations, market trends, and production capacities.

Subsequently, by delving further into Pakistan's fertilizer industry, including historical patterns, economic impacts, and worldwide effects, we expose an abundant terrain for investigating innate import tendencies. The fertilizer industry is a vital component of Pakistan's economy, contributing significantly to the country's Large-Scale Manufacturing (LSM) sector at a rate of about 4.4% and the GDP at a rate of 0.9% (Jahangir, 2023). Leading firms in the sector, such as Fatima Group, Engro Corporation, and Fauji Fertilizer Company, control the majority of the market and generate more than PKR 100 billion in tax income. According to Jahangir's (2023) research, historical growth data illustrates the evolution of the sector, from its initial slow development attributed to restricted investments and rising input costs to its enormous expansion driven by a range of variables and problems.

Meanwhile, the massive inflow of foreign investments—billion-dollar commitments from China, the US, the UK, Qatar, Saudi Arabia, and Japan, for example—that are targeted at bolstering Pakistan's food security portend a potential shift in the country's import profile for fertilizer (Jahangir, 2023). An emerging paradigm for import projections is presented by the combination of cutting-edge technology like artificial intelligence (AI) and precision agriculture, which have the potential to completely transform farming methods. By examining past import data in relation to economic, technological, and governmental factors, predictive models may be able to identify complex patterns directing import quantities over time and, as a result, predict future import trends for Pakistan's fertilizer industry.

In conclusion, there are observable trends in Pakistan's fertilizer imports that are impacted by regional agricultural cycles, national and international market movements, and economic variables. Previous studies, like PACRA's Fertilizer Sector Overview (2021) and Tariq, Farid & others (2016), show relationships between import volumes, crop cycles, and seasonal demand. Although these insights provide a starting point, more research employing sophisticated statistical models and thorough data analysis is required to accurately anticipate import quantities. While research suggests connections among fertilizer application, crop yield, and seasonal requirements, comprehensive forecasting is still lacking. Predictions can be made more accurately by combining historical data with macroeconomic indicators and new technologies like AI and precision farming. However, in order to properly utilize these patterns for predictive modeling, a thorough examination including a variety of parameters is required.

3. Research Question

"To what extent do inherent patterns emerge in the fertilizer imports of Pakistan, and can these patterns be effectively used to predict import quantities?"

Null Hypothesis (H0): There are no discernible inherent patterns in the fertilizer imports of Pakistan, and any observed patterns are due to random variability. Furthermore, the identified patterns are not sufficient to predict import quantities

Alternative Hypothesis (H1): Inherent patterns exist in the fertilizer imports of Pakistan, which establish meaningful structures or groupings in the data. Furthermore, these patterns demonstrate effectiveness in predicting import quantities, indicating a relationship between key variables and the quantity of fertilizer imports.

4. Data Selection

Pakistan, being an agricultural-based economy, heavily relies on fertilizer imports to meet the demands of its population and sustain consistent crop yields. The nation's agriculture sector plays a pivotal role in the economy, contributing significantly to employment and GDP. However, the dependency on imported fertilizers poses challenges in terms of economic stability, food security, and sustainability. Addressing this issue is crucial to shift towards self-sufficiency in fertilizer production, aligning with national objectives of reducing dependency on imports and enhancing local manufacturing capabilities.

Numerous studies in the field emphasize the critical role of fertilizer imports in Pakistan's agricultural sustainability. Existing literature, such as the research conducted by Tariq, Farid & others (2016), PACRA's Fertilizer Sector Overview (2021) and Asim Jahangir's work on the future of Agriculture (2023), etc highlight the correlation between fertilizer usage, seasonal demands, and import quantities. However, these studies mainly establish correlations and provide foundational insights, leaving a gap in developing predictive models to forecast import quantities accurately.

The dataset selected for this study contains relevant variables crucial for predictive modeling and analysis. It encompasses essential information such as HS codes, item descriptions, country of origin, importer and consigner names, shipment values, quantities, and shipment dates. This dataset provides a comprehensive foundation to explore and discern patterns in fertilizer imports, aiding in the development of predictive models. Analyzing this dataset aligns with the larger goal of reducing Pakistan's dependence on imported fertilizers by empowering local production. Leveraging historical import data and diverse variables within the dataset can help identify underlying patterns guiding import quantities. Utilizing this information in predictive modeling will aid in making informed decisions, optimizing local fertilizer production, and fostering sustainable agricultural practices. Additionally, given the presence of significant local fertilizer manufacturing companies like Engro Corporation, Fauji Fertilizer Company, and Fatima Group, the predictive model derived from this dataset can provide valuable insights for these industry players. Ultimately, this aligns with the broader national objective of bolstering self-sufficiency in fertilizer production and promoting a more sustainable agricultural economy in Pakistan.

5. Dataset Description

There are 23 total variables in the dataset out of which 8 are relevant to our research.

- **PCT:** PCT represents the HS Code assigned to each product. An HS code, or Harmonized System code, is a unique numerical code assigned to all internationally traded goods. It is used to classify traded products.
- **ITEM_DESC:** Item description is the description of the product provided to port authorities for shipment.
- **ORIGIN:** The country from which the shipment was sent.
- **IMPORTER_NAME:** Name of the firm that imported the shipment.
- **CONSIGNER.NAME:** Name of the firm that sold the shipment.
- **ASSVALD:** The value of the shipment at the time of receiving.
- **QTY.KG:** The total weight of the shipment product.
- **CASH_DATE:** The date at which the shipment was received at the port.

The irrelevant variables pertain to the LC date, agent name, importer and consignor addresses, taxation details and the specific port in Pakistan where the shipment was received.

6. Dataset Cleaning

1. Data Selection and Shortlisting

For shortlisting the data, a subset of columns was selected from the original .csv file, “df_fert” and the first column was assigned as character variable hence to avoid dropping trailing 0’s. The columns that are selected have data containing item description, hs-codes of the products, importer names, country of origin, quantity of product in kgs, and assessed value. Then a new column called “QTY.MT” is made in which the quantity in kgs columns is converted to metric tonnes because the value in kgs would become too large for easier understanding. Then columns are renamed for further clarification and a new column called “PRICE.PER.MT” is made which shows the price of the respective type of fertilizer for each metric ton during a specific time. Lastly, the weight column in kgs is removed as excess.

2. Convert Date Column to Date Format

The CASH.DATE column, shows the date of the transaction is converted to a date format from character which would be further used in the time series analysis along with regression.

3. Filtering and Standardizing Data Based on HS Codes

The HSCODE column which represents the specific import codes of the products were cleaned of any trailing 0’s. Making the codes more concise and easier to interpret along with the removal of excess products like Ammonium Nitrate which is an explosive not a fertilizer. Furthermore, specialty fertilizers were also removed as its usage isn’t widespread.
The ITEM.DESC column which describes the fertilizer products, is further cleaned based on the revised product codes to only allow the relevant products to stay.

4. Correcting Variable Classes

Several columns, including **HSCODE**, **ITEM.DESC**, **ORIGIN**, **IMPORTER.NAME**, and **CONSIGNERS.NAME**, are converted to the factor variable type for categorical data which would be further used in our analysis after classification and prediction techniques like clusters and regression are run on them.

5. Checking for NA Values

The revised data frame after the thorough data selection is checked for any NA values.

6. Removing Values After 2022

Rows with transaction dates beyond June 2nd ,2022, are filtered out from the dataframe, this is due to data unavailability and irregular data which was proving to be hindrance in running the ML techniques and giving inaccurate outputs.

7. Shortening Country Names

White Spaces are removed from the country-of-origin column, “ORIGIN”. Other countries like the United Arab Emirates, Islamic Republic of Iran, etc are shortened for easier interpretation.

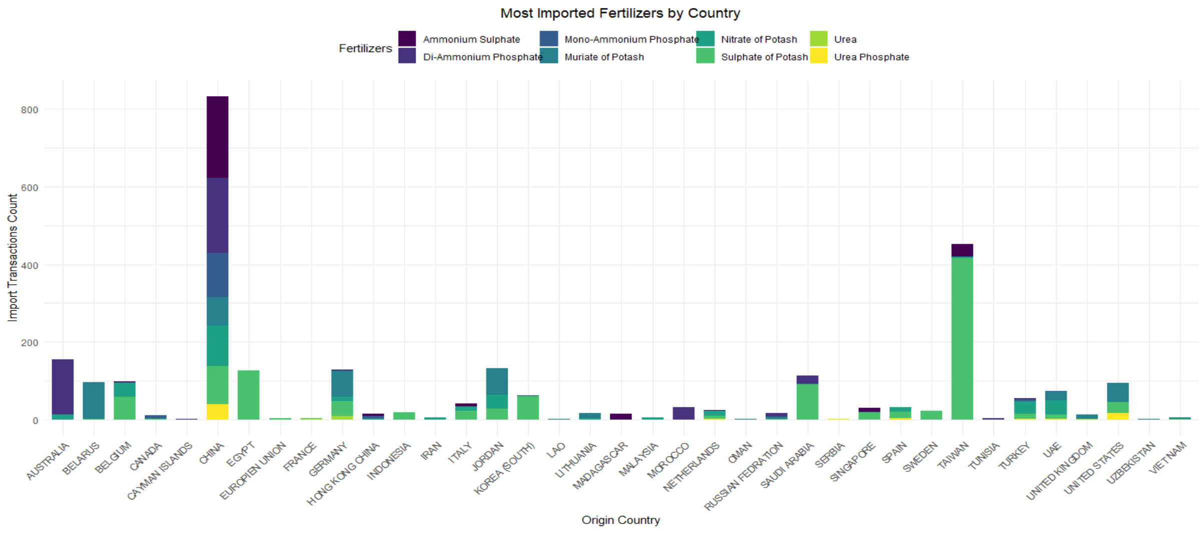
8. Removing Values Below 18 Tons

Rows with quantities less than or equal to 18 metric tons are filtered out, this is because a single container has 18 tons of fertilizer. Entries below are personal use samples which are not used by organizations enough to be non-negligible.

7. Data Exploration

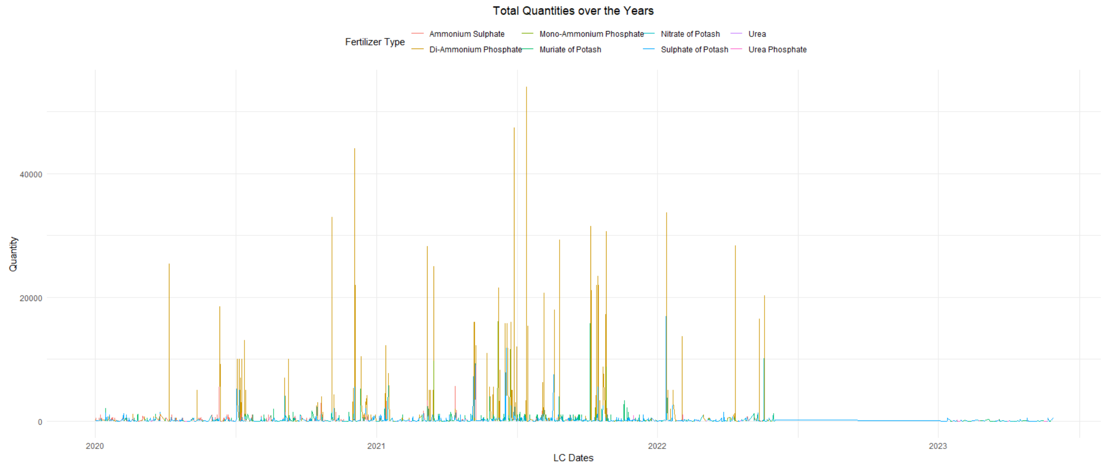
Exploratory Plots Analysis

1. Most Imported Fertilizers by Country



The bar graph shows the Item description of the different types of fertilizers being imported along with its count (Y-axis) and the country of Origin (X-axis). The bar graph shows the highest total count in the country of China with a count of approximately 800. China was also the country importing the most variety of fertilizers which include Ammonium Sulphate, Urea Phosphate, Nitrate of Potash, Muriate of Potash, Mono-Ammonium Phosphate, Di-Ammonium Phosphate, and Ammonium Sulphate. And the lowest being Cayman Island, Egypt, Iran, LAO, Serbia, and Uzbekistan with the count being almost negligible. The greatest item imported was Sulphate of Potash and most of this fertilizer was being imported from Taiwan with almost 420 counts.

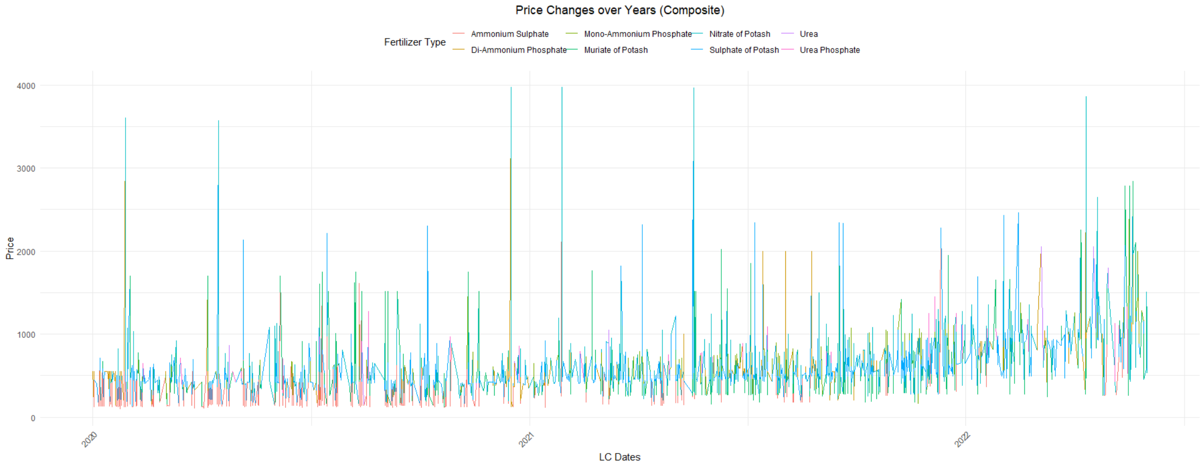
2. Total Fertilizer Import Quantities Over the Years



This plot shows a line chart for the respective item descriptions along with their quantities (Y-axis) over the years 2020-2022 (X-axis). The most imported items were a mix of

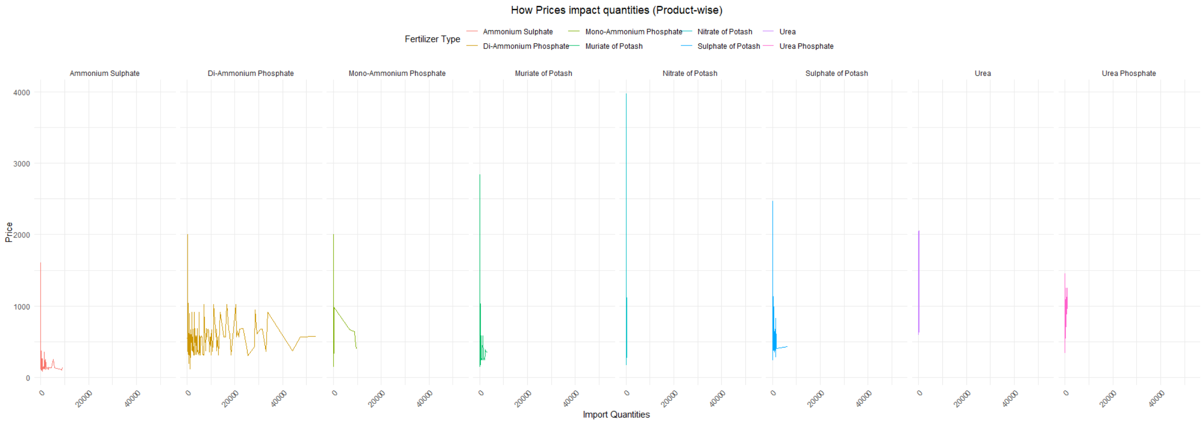
Ammonium Sulphate and Di-Ammonium Phosphate. There were almost negligible imports in the first few months of 2020 for nitrate and Muriate of potash and other fertilizers, with a large import of 25000 for Ammonium Sulphate and di-ammonium phosphorus. Overall, the import quantities are below 3000 and irregular surges in demands over the years are usually for the most imported items with some Mono-Ammonium Phosphate. The highest import quantity was a mix of Ammonium Sulphate and Di-Ammonium Phosphate reaching approximately 53000 in the second half of 2021.

3. Price Changes Over Years (Composite)



This line graph shows the variation in price of the different imported fertilizers (X-axis) over the year 2020-2022 (Y-axis). This graph provides a holistic view of how prices have evolved over the years, which indicates the market and economic conditions, along with reasons for constraints and the surges in demand. With an overall value between 200-600 for the first few months, the sudden surge in Mono-Ammonium Phosphate jumps the price to 3600pkr/-. For instance, the lowest prices were at the start of 2020 with spikes at random intervals. The highest price reached over the years was reached by Muriate of Potash and Nitrate of Potash, with the price reaching up to approximately 4000pkr/- at the start of 2021. The Average price of the fertilizers had relatively low fluctuations and were usually between 100 – 1000pkr/-.

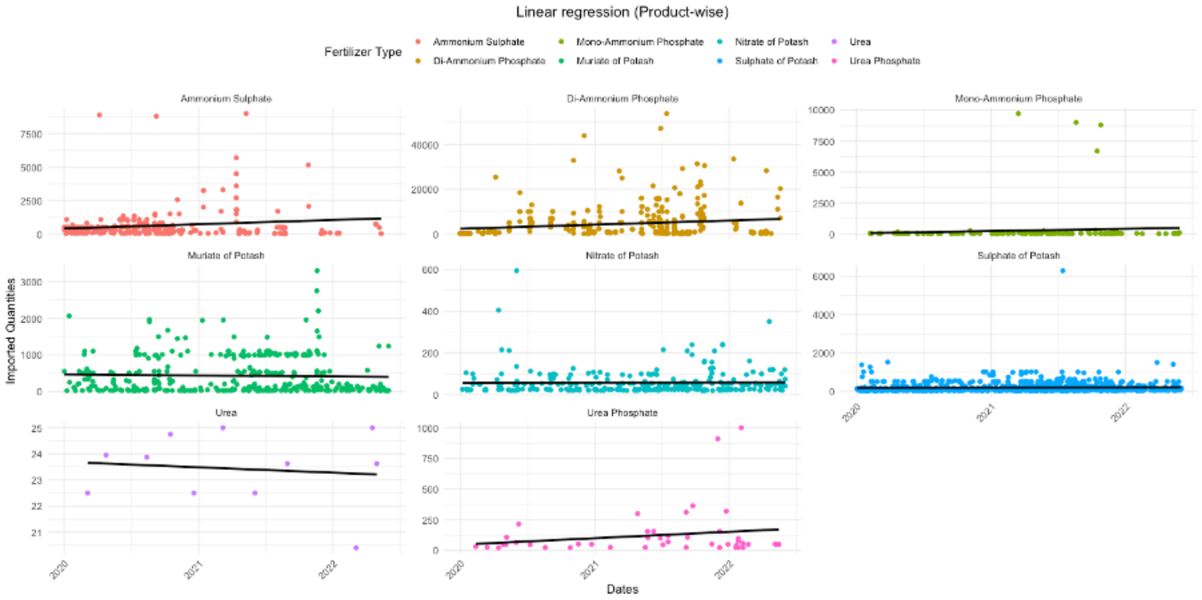
4. Impact of Prices on Import Quantities



The individual line graphs for each different type of fertilizer in the graph help us analyze the fluctuation in the demand (X-axis) compared with the LC dates (Y-axis). The highest total demand for the three years was for Di-Ammonium Phosphate, with the highest order of

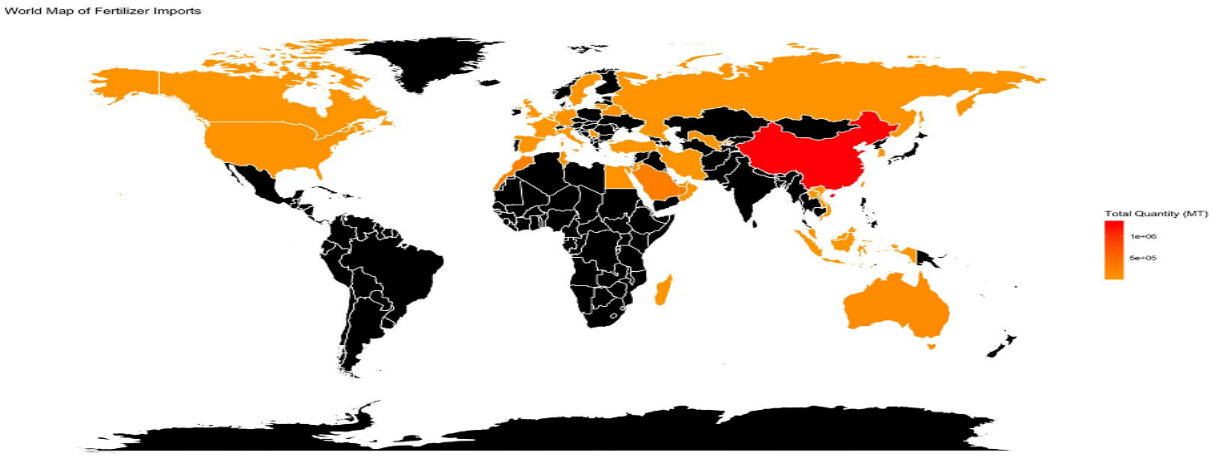
50000+ metric tonnes, while the lowest belonged to UREA. The low import of UREA is due to subsidized gas rates for its local production and the size of production over Pakistan. While fertilizers like Ammonium Sulphate had a quantity lower than 10,000, Muriate of Potash and Sulphate of Potash had an average quantity lower than 5000.

5. Product-wise Linear Regression



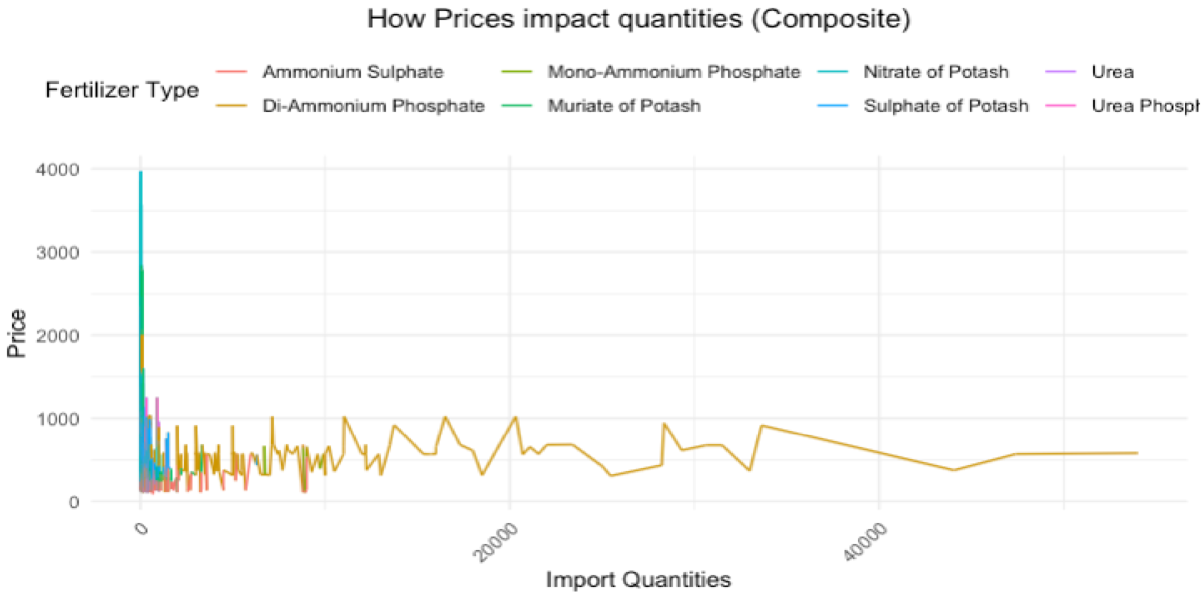
Linear Regression was run on the individual scatter plots for each fertilizer. With imported quantities on the x-axis and dates from the year 2020-2022 on the y-axis, trend lines were created for each plot for accurate prediction of the quantities that need to be imported.

6. World Map of Fertilizer Imports



The world map above shows the geographic location of the countries from whom the fertilizers were imported. The countries are color-coded according to different shades of orange and red which represent lower to higher values of **Total Metric Tonnes Quantity Supplied** by each country. The color black denotes the locations from which any fertilizer wasn't imported. The country with the highest total quantity is China which is denoted in bright red. While low quantity countries like the Cayman Islands, Egypt, and Iran are shown in light orange.

7. Impact of Price over Quantity (Composite)



The line graph above helps provide insights into the correlation between the price (X-axis) and the import quantities of the fertilizers (Y-axis). During the peak prices of every kind of fertilizer, the demand was the lowest, as can be seen at the start of the graph. However, there is one exception which is Di-Ammonium Phosphate. This fertilizer was imported in high quantities over all the years despite changes in price between 400-1000pkr/- with total import quantity reaching over 50,000 metric tonnes.

8. Choice of Technique

After completing literature review and dataset exploration a number of techniques for data mining were explored. The decision to select clustering and regression to analyze the fertilizer imports of Pakistan is grounded in the capabilities of these techniques to uncover the complex relationships and patterns within the data.

Since the dataset included different variables like origin, type of fertilizers, shipment values, dates etc., clustering handles this multidimensional data exceptionally well. It segments the data based on similarities across these multiple dimensions and provides insights revealing the underlying market dynamics. Clustering can also prove useful in uncovering patterns related to market shifts and policy changes. Since there is complexity influenced by local practices and global market conditions in the agricultural sector, robust analytical techniques are required. Clustering simplifies this complexity.

The use of regression will allow predictive modeling in understanding how variables like fertilizer type, origin and value will influence the imports. It will also help in quantifying the strength and nature of the relationships between these variables and make insightful forecasts about future import quantities. Furthermore, time based variables like cash dates is another reason for using regression, allowing the analysis of trends and seasonal patterns in fertilizer imports, assessing the impact of changes on import quantities.

Among the un-supervised learning techniques, we were unable to implement K-means clustering because of the nature of the dataset, i.e. contained a lot of categorical and non-numerical values like HS codes and country names, which resulted in the algorithm rendered invalid for partitioning the data. Furthermore, Association rules are used for market basket analysis, and hence were not applicable on this dataset. Lastly, decision trees although are useful for classification and regression but they oversimplified the dataset at hand, and hence led to less accurate results.

9. Modeling

Hierarchical Clustering

1. Data Processing

To start with, we had to choose the appropriate columns to produce meaningful clustering results, and for that we chose 4 variables: "ORIGIN", "ITEM.DESC", "QTY.MT", "IMPORTER.NAME", "CONSIGNERS.NAME". Furthermore, we excluded the "HSCODE" and dates columns.

2. Number of Clusters

When deciding the number of clusters, we generated an elbow diagram to analyze where there was a significant drop in the Within Cluster Sum of Squares and that occurred at k=2. Intuitively as well, 2 clusters made more sense as it separated Di ammonium phosphate and rest of the products and in 3 clusters it only divided Di ammonium phosphate clusters. Cluster 1 has 403 members, and Cluster 2 has 2207 members.

3. Cluster Health and Membership

Then we moved onto the cluster size, information loss, overlap, and cluster height to determine which would be the best method to employ for the number of clusters, as shown in the dendrograms below.

4. Average Linkage

With this, the loss of information was a little lower than the single linkage method, however, the cluster size and the distances between them were great.

Hence, we went with **Complete Linkage**, the choice of linkage method can significantly affect the structure of the resulting dendrogram. In the dataset, the "complete" linkage method tends to find similar clusters, leading to more balanced dendrogram structures, especially when there are outliers or clusters of different sizes. Hence, we went for a lower loss of information and better cluster sizes which was provided to us by the complete method.

5. Cluster Profiling

The following countries and items are present in each cluster

Cluster 1:

Countries: Australia, Cayman Islands, China, Morocco, Russian Federation, Saudi Arabia, Singapore, Tunisia.

Items: Di-Ammonium Phosphate

Cluster 2:

Countries: Australia, Belarus, Belgium, Canada, China, Egypt, European Union, France, Germany, Hong Kong China, Indonesia, Iran, Italy, Jordan, Korea (South), Lao, Lithuania, Madagascar, Malaysia, Netherlands, Oman, Russian Federation, Saudi Arabia, Serbia, Singapore, Spain, Sweden, Taiwan, Turkey, UAE, United Kingdom, United States, Uzbekistan, Vietnam.

Items: Ammonium Sulphate, Mono-Ammonium Phosphate, Muriate of Potash, Nitrate of Potash, Sulphate of Potash, Urea, Urea Phosphate Now based upon these 2 criterions and our subjective interpretation, this is how we came up with the profiling:

Cluster 1: Di-Ammonium Phosphate Exporters

Counties in cluster 1 are the only ones Pakistan imports Di ammonium phosphate from. Countries exclusively in cluster 1 are the ones Pakistan only imports Di ammonium phosphate from Morocco which has the world's largest phosphorus deposits. The cluster mean, max and min value are all substantially different from that of cluster 2, which goes on to show how important this nutrient is as a fertilizer for Pakistani farmers.

Cluster 2: Other Fertilizer Exporters

Countries in cluster 2 are the ones from which Pakistan imports the rest of the fertilizers from. Countries in cluster 2 are those countries which are rich in nutrients like Potash and Nitrate and hence are major suppliers of the worldwide fertilizers for this particular nutrients.

Regression Analysis

1. Model Analysis

Model 1:

This regression model predicts quantity (QTY.MT) as a function of the types of fertilizers, assessed value (ASSDVAL), origin country (ORIGIN) and date (CASH.DATE).

Residuals (the difference between observed and predicted values) range between -13,979.2 to 18,110.2. This implies that there may be some large deviations, however, the median is close to zero suggesting a reasonable fit of the model. Residual standard error is observed to be relatively high which may indicate some inaccuracy in predictions. However, multiple R squared value is 0.91 which suggests that approximately 91% of the variability is explained by the model. A high F-statistic with low p-value indicates that the model is statistically significant.

Every coefficient represents the change expected in quantity with a unit change in the predictor variable while holding others constant. P-values will indicate if these coefficients are statistically significant (below 0.05)

Different types of fertilizers have varying impacts on the quantity of imports. Di-Ammonium Phosphate positively affects the quantity whereas Urea Phosphate negatively affects it. As expected, the assessed value also has a positive relationship with the value of imports and quantity imported, hence, every unit increase, also increases the import value. Origin Country also significantly affects the import quantity indicating that some countries are more crucial to fertilizer imports since there is more dependency. For example, China and Tunisia have a high positive impact, however, Belgium and Germany show low impacts. There is a negative influence of Cash Date (statistically significant) on quantity imported, suggesting that there has been a decrease over time.

Model 2:

This regression model predicts quantity (QTY.MT) as a function of the types of fertilizers (ITEM.DESC), assessed value (ASSDVAL), Price per metric ton (PRICE.PER.MT) and date (CASH.DATE).

Residuals (the difference between observed and predicted values) range between (-14,113.9 to 17,870.8. This implies that there may be some large deviations. The median is -47 suggesting a small bias in the model.

Residual standard error is observed to be relatively high which may indicate some inaccuracy in predictions. However, multiple R squared value is 0.907 which suggests that approximately 90.7% of the variability is explained by the model. A high F-statistic with low p-value (< 2.2e-16) indicates that the model is statistically significant.

Different types of fertilizers have varying impacts on the quantity of imports. The price variable signifies a negative influence on import quantities as expected, suggesting that as price increases, the quantity of imports will decrease. The rest of the variables are approximately similar to the first model.

Model 3 (Best Fit):

This refined regression model predicts quantity (QTY.MT) as a function of the types of fertilizers (ITEM.DESC), Price per metric ton (PRICE.PER.MT) and date (CASH.DATE). This analysis is conducted on a subset of data that includes only those fertilizers that proved to be significantly affecting the model (since they were imported more than others): "Di-Ammonium Phosphate," "Nitrate of Potash," "Sulphate of Potash," and "Ammonium Sulphate."

Residuals (the difference between observed and predicted values) range between -13,998.6 to 17,870.8. This implies that there may be some deviations from the predicted values, while the median of the residuals is -58.7, suggesting a slight bias.

Residual standard error is observed to be relatively high which may indicate some inaccuracy in predictions. However, the multiple R squared value is 0.907 which suggests that approximately 90.7% of the variability is explained by the model, indicating a strong fit.

A high F-statistic (3300) with a very low p-value indicates that the model is statistically significant.

Di-Ammonium Phosphate positively and significantly affects the quantity implying that is associated with higher quantity of imports making it a potentially lucrative product. "Nitrate of Potash" and "Sulphate of Potash" have negative coefficients, but they are not statistically significant, suggesting that they may not significantly impact profit in this context. The price variable with a coefficient of -5.037e-01 signifies a negative influence on import quantities as expected, suggesting that as price increases, quantity of imports will decrease. The negative impact of the price per metric ton on profit suggests the need for careful pricing strategies to maximize profit margins, especially for these specific fertilizer types.

As expected, the assessed value with a coefficient of 1.536e-03 also has a positive relationship with the value of imports and quantity imported, hence, every unit increase, also increases the import value. There is a negative influence of Cash Date with a coefficient of -4.332e-01 on quantity imported, suggesting that there has been a decrease over time which may also be attributed to the economic duress during the pandemic and the recovery period from the pandemic that has affected spending in the country.

2. Model Comparison

Model 1:

lm(QTY.MT ~ ITEM.DESC + ASSDVAL + ORIGIN + CASH.DATE, data = imports)

This model predicts the quantity of imports on the fertilizer, assessed value, origin and cash date and highlights the influence of these variables on import quantities. It has a high R-squared value, indicating a strong model fit which is useful for understanding how different variables impact the quantity of imports.

Model 2:

lm(QTY.MT ~ ITEM.DESC + PRICE.PER.MT + CASH.DATE + ASSDVAL, data = imports)

This model predicts the quantity of imports on the fertilizer, assessed value, price and cash date. It adds the dimension of price sensitivity to the model. It has a high R-squared value, indicating a strong model fit which is useful for understanding market dynamics and pricing strategies.

Model 3:

lm(QTY.MT ~ ITEM.DESC + PRICE.PER.MT + CASH.DATE + ASSDVAL, data = rel.prods)

This model predicts the quantity of imports on specific fertilizers, assessed value, price and cash date. It is focused on specific fertilizers and provides targeted insights. It has a high R-squared value, indicating a strong model fit which is useful for specific decision-making regarding the chosen fertilizers.

Model 4:

lm(ASSDVAL ~ ITEM.DESC + CASH.DATE + PRICE.PER.MT, data = imports)

This model predicts the assessed value on fertilizers, price and cash date. It has a low R-squared value, indicating a weak model fit which could be used for financial planning and budgeting.

Model 5:

lm(QTY.MT ~ CASH.DATE, data = imports)

This model predicts the quantity of imports solely on cash date. It shows the trend of import quantities over time. It has a low R-squared value, indicating a weak model fit which could be used only for trend analysis.

Model 6:

lm(QTY.MT ~ PRICE.PER.MT, data = imports)

This model predicts the quantity of imports solely on price. It examines the relationship between quantity of imports and price, also indicating price sensitivity. It has a very low R-squared value, indicating a weak model fit which could be used only to understand price impact on quantities.

3. Regression Insights

Model 1 and Model 3 are observed to be the best fitting models on a number of factors. They provide comprehensive insights into the factors affecting import quantities of fertilizers. The model is statistically significant with high F statistics and low p-values.

10. Recommendations

There is significant diversity in the import origins which shows a balanced approach by mitigating the risk of over-reliance on specific countries. This is a positive approach considering dynamic geopolitical conditions for Pakistan. The trends in import quantities and related shifts in prices show the high demand of fertilizers in Pakistan year round, making a primarily agricultural based country reliant on fertilizer imports. The results of cluster analysis show the significance of certain products such as Di-Ammonium Phosphate for Pakistan and help guide future procurement strategies for import of fertilizers.

This also shows certain concerning trends as it shows Pakistan's over reliance on the imports of Di-Ammonium Phosphate. For an agriculture based country, such a commodity product should be primarily produced within the country instead of being imported. However, Pakistan is reliant on its imports to such a scale that it is purchased in consistent quantities regardless of price. Another point of concern is that Nitrogen and Potassium based fertilizers are used in significantly higher quantities in the countries as these fertilizers are imported in relatively smaller quantities despite not being produced locally in sufficient quantities. Phosphorus fertilizers are integral for root development and fruit development in crops. This shows that the farmers currently do not use these fertilizers in sufficient quantities and policy makers should explore the causal factors of these results to improve Pakistan's agricultural yield.

In conclusion, these findings show current import patterns and provide suggestions for policy-making, trade negotiations, and forecasting fertilizer supply for the agriculture sector of Pakistan.

11. Limitations

Data Related Limitations

There are several limitations inherent to the data selected. The data was compiled by port authorities of Pakistan and at the time of the data collection, certain variables and data values were compromised due to human error which had to be removed during data cleaning thus resulting in a smaller dataset. The relevant variables in the data are mostly categorial with only two numeric variables i.e. quantity and assessed value. This means that many of the machine learning techniques such as Decision trees, K-means clustering and K-Nearest Neighbour cannot be used while taking into account many of the relevant variables. There was also a gap in the data post 2022 with certain product imports not being recorded resulting in the exclusion of data post 2022.

Models Related Limitations

1. Clustering

The nature of the data meant that clustering could only be used in limited capacity. Having mostly categorical variables limited the dataset to hierarchical clustering to take into account all the relevant variables. Having a significantly large dataset of transactions also resulted in an extensively wide dendrogram which is difficult to interpret. The results also have a limitation as the nature of Di-ammonium phosphate imports has a significant impact on the results on their own making the ideal clusters between Di-Ammonium Phosphate and other fertilizers despite the number of clusters made.

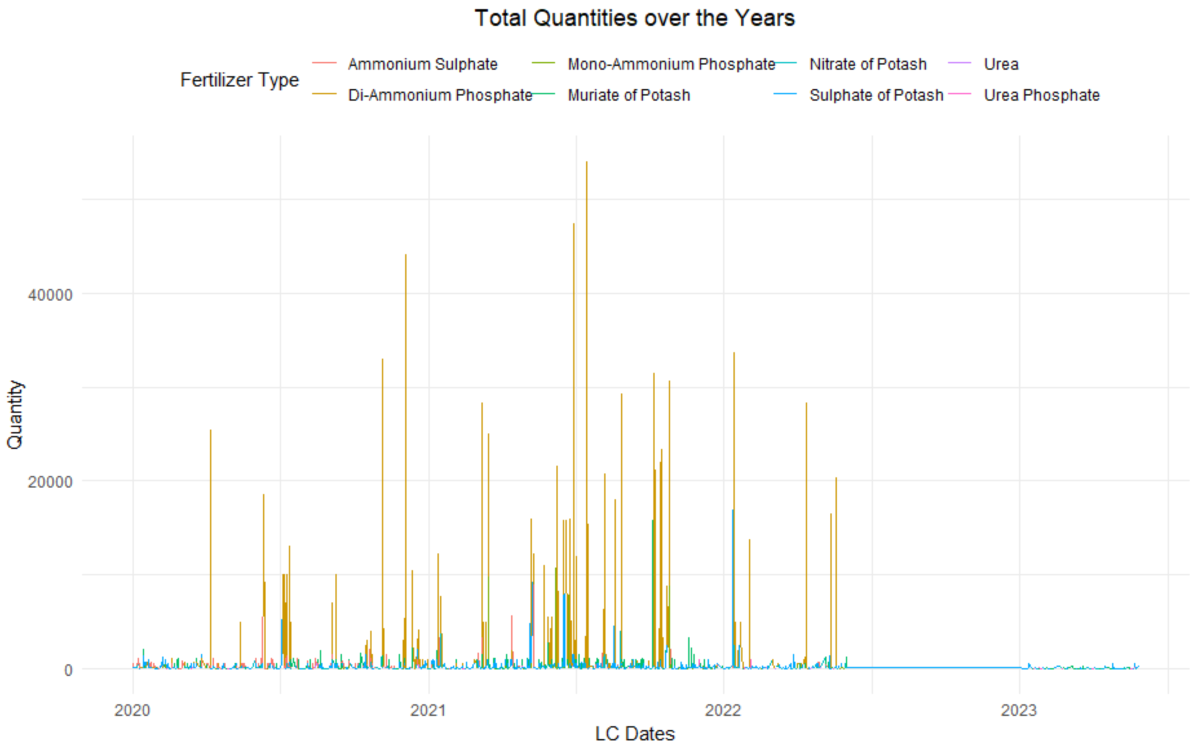
2. Regression

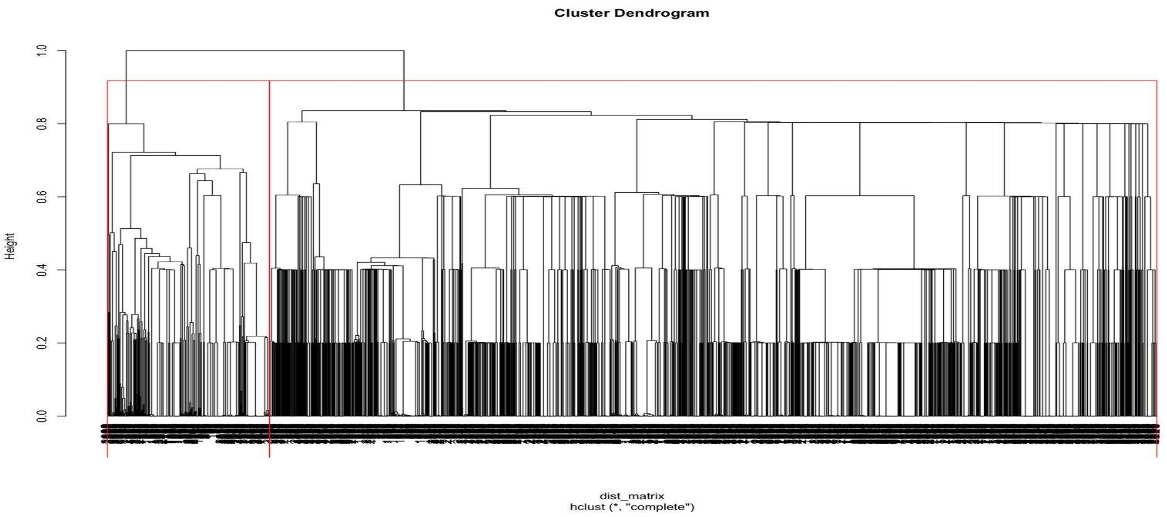
In the regression analysis certain variables are correlated such as origin and consigners and thus some models become sensitive to multicollinearity. Moreover, the use of linear and logistic regression gave us the same output which means that the relationship being formed is linear and is not taking into account the high number of outliers specifically accounting for the import quantities of Di-Ammonium Phosphate. Lastly, regression only takes into account certain variables from the data and does not explain causality behind the changes in quantity.

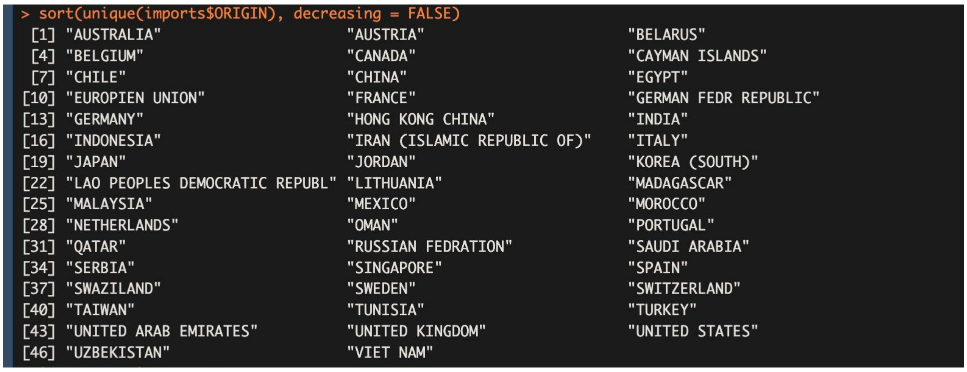
Findings Related Limitations

The results do not provide a holistic view to create a complete picture of the fertilizer imports market of Pakistan. For a better analysis we would view these findings with other qualitative variables such as government policies and international political environment. Moreover, this data was collected during the global pandemic which caused unprecedented change in the global trade and the findings need to be viewed in tandem with changes brought about by Covid 19. For a comparative analysis there would need to be a similar study conducted a few years after the mitigation of Covid 19 to get a holistic view of the fertilizer imports market.

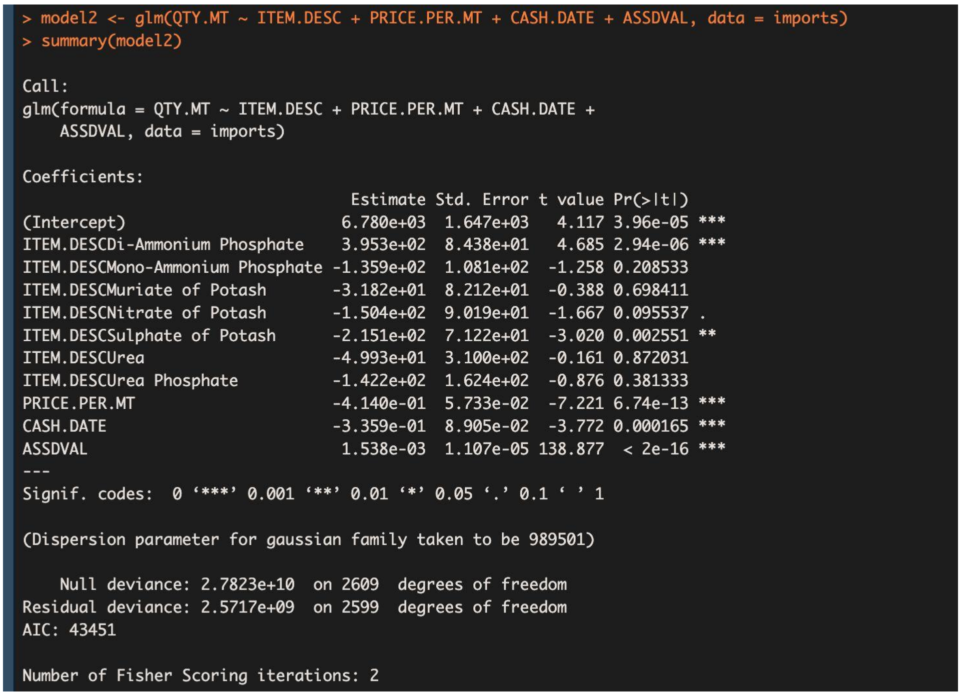
Appendix



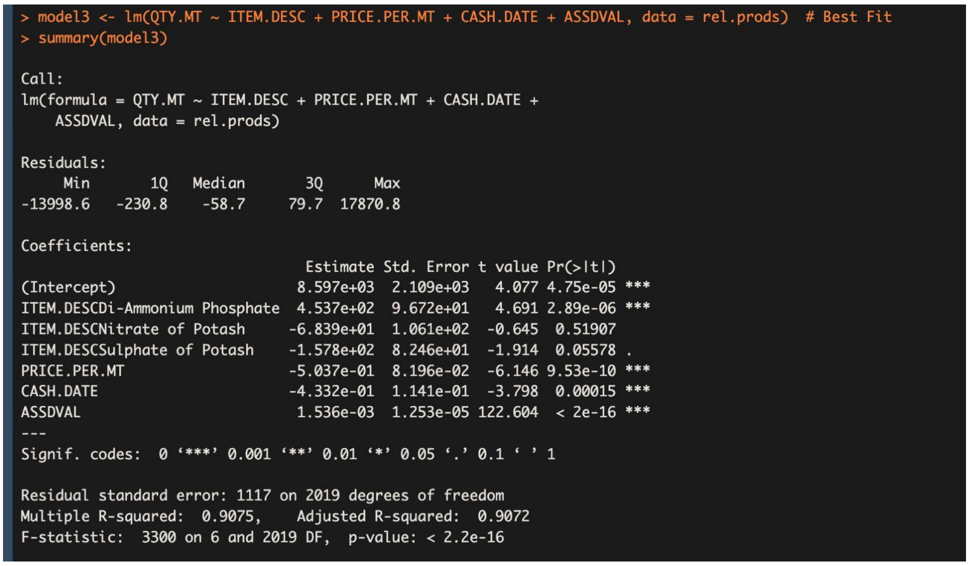




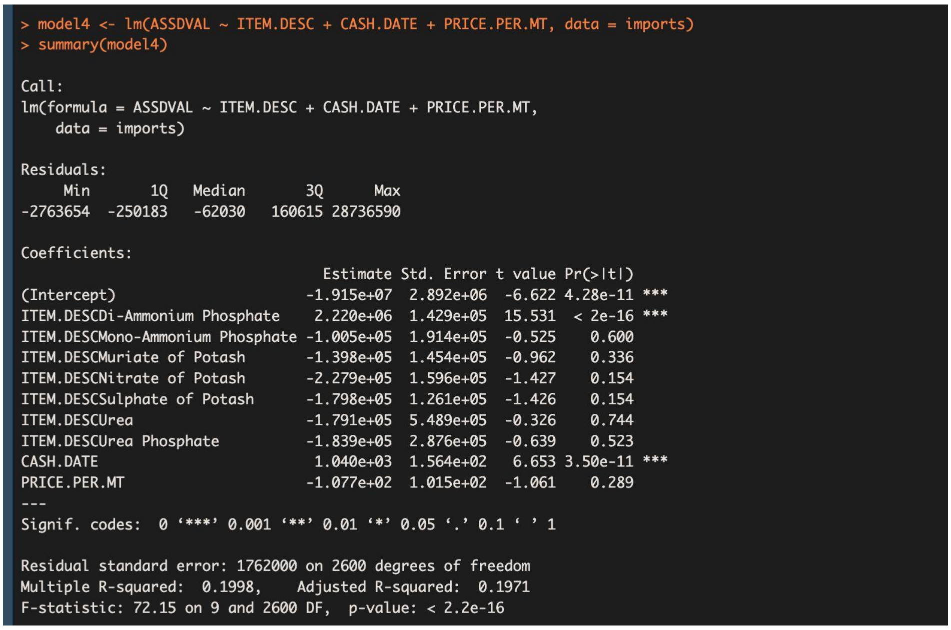
```
> sort(unique(imports$ORIGIN), decreasing = FALSE)
 [1] "AUSTRALIA"                       "AUSTRIA"                        "BELARUS"
 [4] "BELGIUM"                         "CANADA"                         "CAYMAN ISLANDS"
 [7] "CHILE"                           "CHINA"                          "EGYPT"
[10] "EUROPIEN UNION"                  "FRANCE"                         "GERMAN FEDR REPUBLIC"
[13] "GERMANY"                         "HONG KONG CHINA"                "INDIA"
[16] "INDONESIA"                       "IRAN (ISLAMIC REPUBLIC OF)"     "ITALY"
[19] "JAPAN"                           "JORDAN"                         "KOREA (SOUTH)"
[22] "LAO PEOPLES DEMOCRATIC REPUBL"   "LITHUANIA"                      "MADAGASCAR"
[25] "MALAYSIA"                        "MEXICO"                         "MOROCCO"
[28] "NETHERLANDS"                     "OMAN"                           "PORTUGAL"
[31] "QATAR"                           "RUSSIAN FEDRATION"              "SAUDI ARABIA"
[34] "SERBIA"                          "SINGAPORE"                      "SPAIN"
[37] "SWAZILAND"                       "SWEDEN"                         "SWITZERLAND"
[40] "TAIWAN"                          "TUNISIA"                        "TURKEY"
[43] "UNITED ARAB EMIRATES"            "UNITED KINGDOM"                 "UNITED STATES"
[46] "UZBEKISTAN"                      "VIET NAM"
```



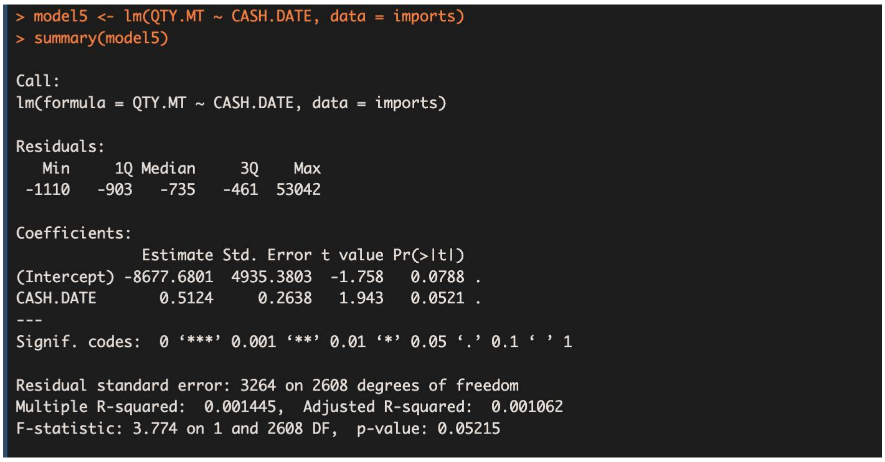
```
> model2 <- glm(QTY.MT ~ ITEM.DESC + PRICE.PER.MT + CASH.DATE + ASSDVAL, data = imports)
> summary(model2)

Call:
glm(formula = QTY.MT ~ ITEM.DESC + PRICE.PER.MT + CASH.DATE +
    ASSDVAL, data = imports)

Coefficients:
                                   Estimate Std. Error t value Pr(>|t|)
(Intercept)                       6.780e+03  1.647e+03   4.117 3.96e-05 ***
ITEM.DESCDi-Ammonium Phosphate    3.953e+02  8.438e+01   4.685 2.94e-06 ***
ITEM.DESCMono-Ammonium Phosphate -1.359e+02  1.081e+02  -1.258 0.208533
ITEM.DESCMuriate of Potash       -3.182e+01  8.212e+01  -0.388 0.698411
ITEM.DESCNitrate of Potash       -1.504e+02  9.019e+01  -1.667 0.095537 .
ITEM.DESCSulphate of Potash      -2.151e+02  7.122e+01  -3.020 0.002551 **
ITEM.DESCUrea                    -4.993e+01  3.100e+02  -0.161 0.872031
ITEM.DESCUrea Phosphate          -1.422e+02  1.624e+02  -0.876 0.381333
PRICE.PER.MT                     -4.140e-01  5.733e-02  -7.221 6.74e-13 ***
CASH.DATE                        -3.359e-01  8.905e-02  -3.772 0.000165 ***
ASSDVAL                           1.538e-03  1.107e-05 138.877  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 989501)

    Null deviance: 2.7823e+10  on 2609  degrees of freedom
Residual deviance: 2.5717e+09  on 2599  degrees of freedom
AIC: 43451

Number of Fisher Scoring iterations: 2
```



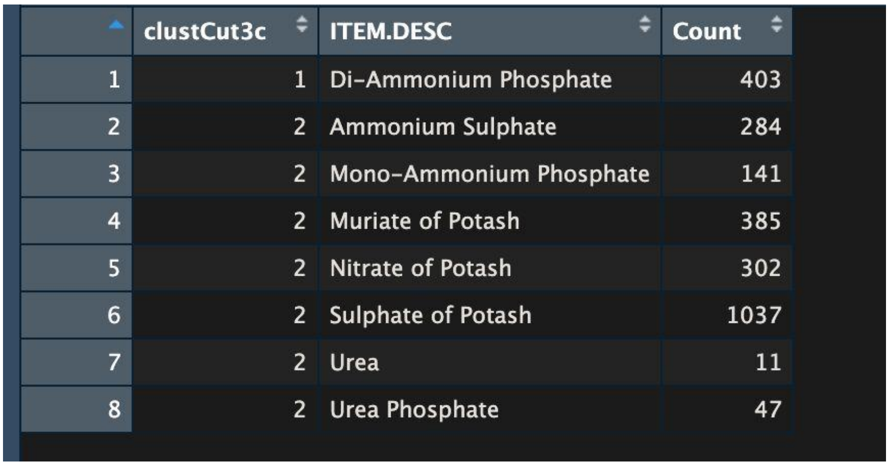
```
> model3 <- lm(QTY.MT ~ ITEM.DESC + PRICE.PER.MT + CASH.DATE + ASSDVAL, data = rel.prods)  # Best Fit
> summary(model3)

Call:
lm(formula = QTY.MT ~ ITEM.DESC + PRICE.PER.MT + CASH.DATE +
    ASSDVAL, data = rel.prods)

Residuals:
     Min       1Q   Median       3Q      Max
-13998.6   -230.8    -58.7     79.7  17870.8

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     8.597e+03  2.109e+03   4.077 4.75e-05 ***
ITEM.DESCDi-Ammonium Phosphate  4.537e+02  9.672e+01   4.691 2.89e-06 ***
ITEM.DESCNitrate of Potash     -6.839e+01  1.061e+02  -0.645  0.51907
ITEM.DESCSulphate of Potash    -1.578e+02  8.246e+01  -1.914  0.05578 .
PRICE.PER.MT                   -5.037e-01  8.196e-02  -6.146 9.53e-10 ***
CASH.DATE                      -4.332e-01  1.141e-01  -3.798  0.00015 ***
ASSDVAL                         1.536e-03  1.253e-05 122.604  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1117 on 2019 degrees of freedom
Multiple R-squared:  0.9075,	Adjusted R-squared:  0.9072
F-statistic:  3300 on 6 and 2019 DF,  p-value: < 2.2e-16
```



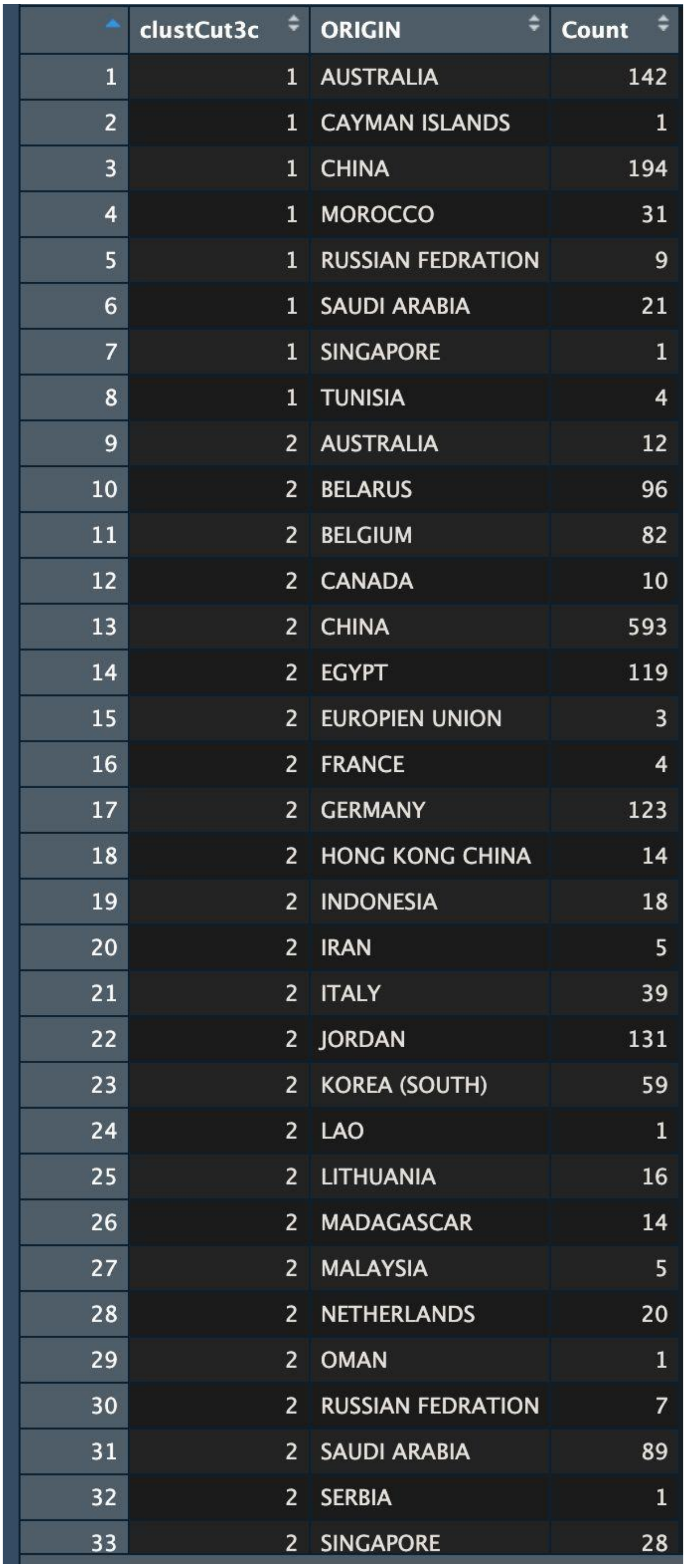
```
> model4 <- lm(ASSDVAL ~ ITEM.DESC + CASH.DATE + PRICE.PER.MT, data = imports)
> summary(model4)

Call:
lm(formula = ASSDVAL ~ ITEM.DESC + CASH.DATE + PRICE.PER.MT,
    data = imports)

Residuals:
     Min       1Q   Median       3Q      Max
-2763654  -250183   -62030   160615 28736590

Coefficients:
                                   Estimate Std. Error t value Pr(>|t|)
(Intercept)                      -1.915e+07  2.892e+06  -6.622 4.28e-11 ***
ITEM.DESCDi-Ammonium Phosphate    2.220e+06  1.429e+05  15.531  < 2e-16 ***
ITEM.DESCMono-Ammonium Phosphate -1.005e+05  1.914e+05  -0.525    0.600
ITEM.DESCMuriate of Potash       -1.398e+05  1.454e+05  -0.962    0.336
ITEM.DESCNitrate of Potash       -2.279e+05  1.596e+05  -1.427    0.154
ITEM.DESCSulphate of Potash      -1.798e+05  1.261e+05  -1.426    0.154
ITEM.DESCUrea                    -1.791e+05  5.489e+05  -0.326    0.744
ITEM.DESCUrea Phosphate          -1.839e+05  2.876e+05  -0.639    0.523
CASH.DATE                         1.040e+03  1.564e+02   6.653 3.50e-11 ***
PRICE.PER.MT                     -1.077e+02  1.015e+02  -1.061    0.289
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1762000 on 2600 degrees of freedom
Multiple R-squared:  0.1998,    Adjusted R-squared:  0.1971
F-statistic: 72.15 on 9 and 2600 DF,  p-value: < 2.2e-16
```



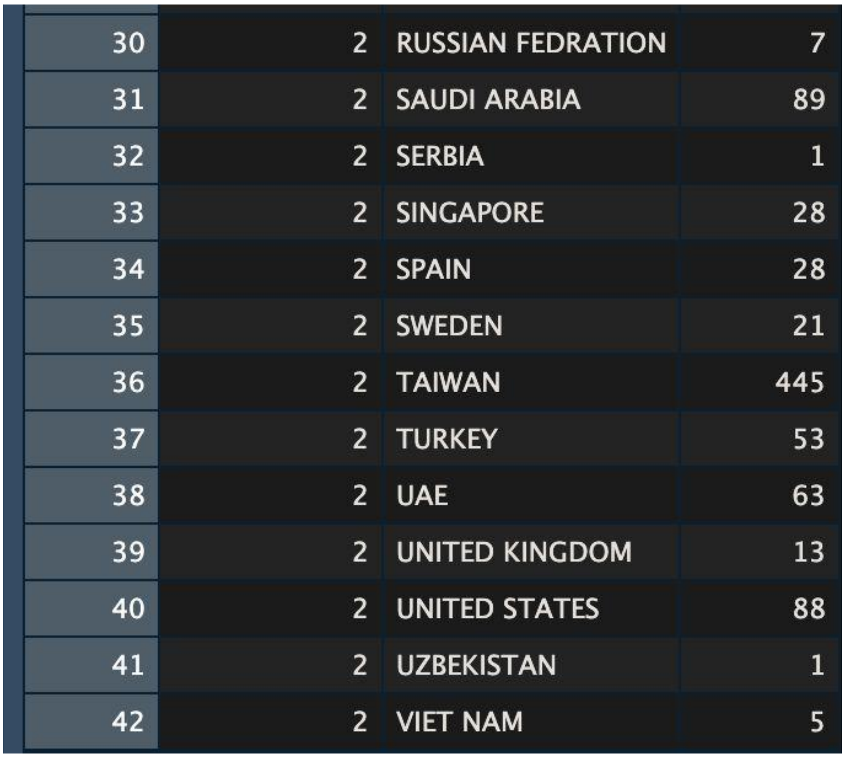
```
> model5 <- lm(QTY.MT ~ CASH.DATE, data = imports)
> summary(model5)

Call:
lm(formula = QTY.MT ~ CASH.DATE, data = imports)

Residuals:
   Min     1Q Median     3Q    Max
 -1110   -903   -735   -461  53042

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -8677.6801  4935.3803  -1.758   0.0788 .
CASH.DATE       0.5124     0.2638   1.943   0.0521 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3264 on 2608 degrees of freedom
Multiple R-squared:  0.001445,  Adjusted R-squared:  0.001062
F-statistic: 3.774 on 1 and 2608 DF,  p-value: 0.05215
```



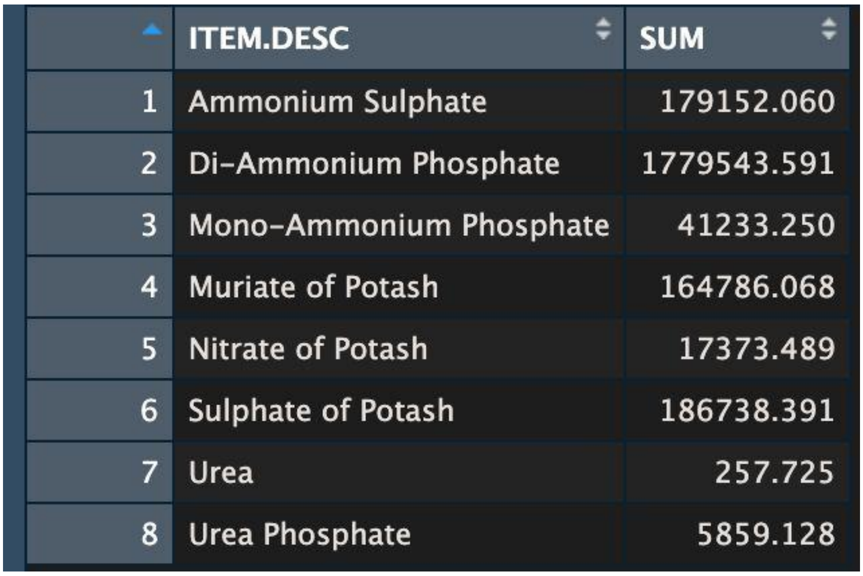
	clustCut3c	ITEM.DESC	Count
1	1	Di-Ammonium Phosphate	403
2	2	Ammonium Sulphate	284
3	2	Mono-Ammonium Phosphate	141
4	2	Muriate of Potash	385
5	2	Nitrate of Potash	302
6	2	Sulphate of Potash	1037
7	2	Urea	11
8	2	Urea Phosphate	47



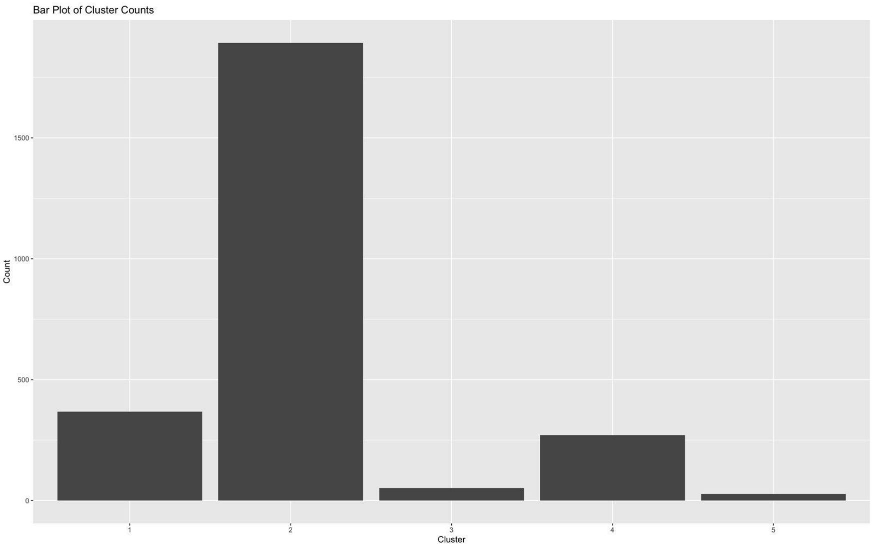
	clustCut3c	ORIGIN	Count
1	1	AUSTRALIA	142
2	1	CAYMAN ISLANDS	1
3	1	CHINA	194
4	1	MOROCCO	31
5	1	RUSSIAN FEDRATION	9
6	1	SAUDI ARABIA	21
7	1	SINGAPORE	1
8	1	TUNISIA	4
9	2	AUSTRALIA	12
10	2	BELARUS	96
11	2	BELGIUM	82
12	2	CANADA	10
13	2	CHINA	593
14	2	EGYPT	119
15	2	EUROPIEN UNION	3
16	2	FRANCE	4
17	2	GERMANY	123
18	2	HONG KONG CHINA	14
19	2	INDONESIA	18
20	2	IRAN	5
21	2	ITALY	39
22	2	JORDAN	131
23	2	KOREA (SOUTH)	59
24	2	LAO	1
25	2	LITHUANIA	16
26	2	MADAGASCAR	14
27	2	MALAYSIA	5
28	2	NETHERLANDS	20
29	2	OMAN	1
30	2	RUSSIAN FEDRATION	7
31	2	SAUDI ARABIA	89
32	2	SERBIA	1
33	2	SINGAPORE	28

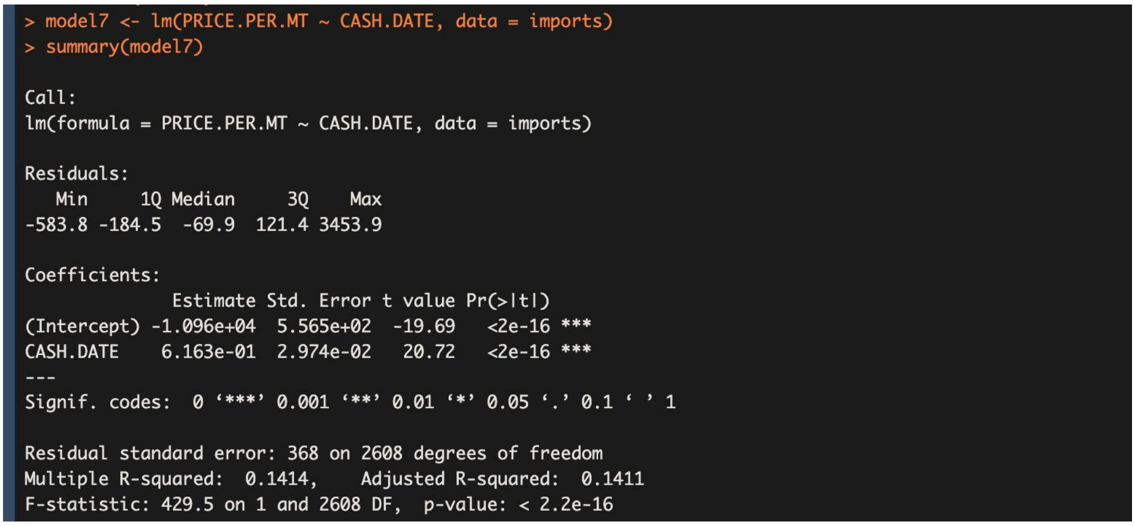


30	2	RUSSIAN FEDRATION	7
31	2	SAUDI ARABIA	89
32	2	SERBIA	1
33	2	SINGAPORE	28
34	2	SPAIN	28
35	2	SWEDEN	21
36	2	TAIWAN	445
37	2	TURKEY	53
38	2	UAE	63
39	2	UNITED KINGDOM	13
40	2	UNITED STATES	88
41	2	UZBEKISTAN	1
42	2	VIET NAM	5

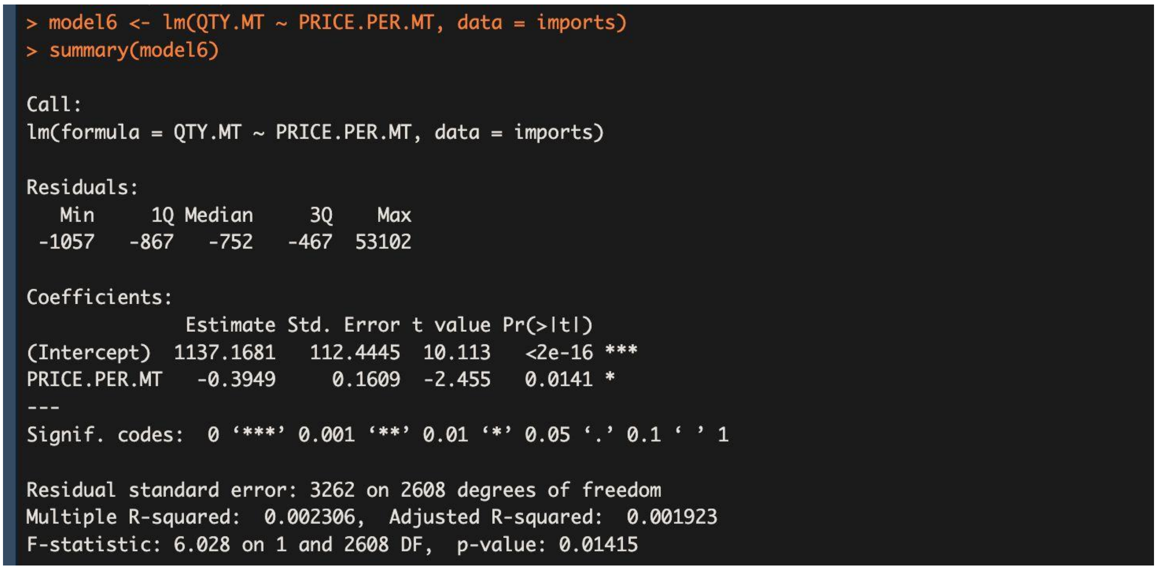


	ITEM.DESC	SUM
1	Ammonium Sulphate	179152.060
2	Di-Ammonium Phosphate	1779543.591
3	Mono-Ammonium Phosphate	41233.250
4	Muriate of Potash	164786.068
5	Nitrate of Potash	17373.489
6	Sulphate of Potash	186738.391
7	Urea	257.725
8	Urea Phosphate	5859.128

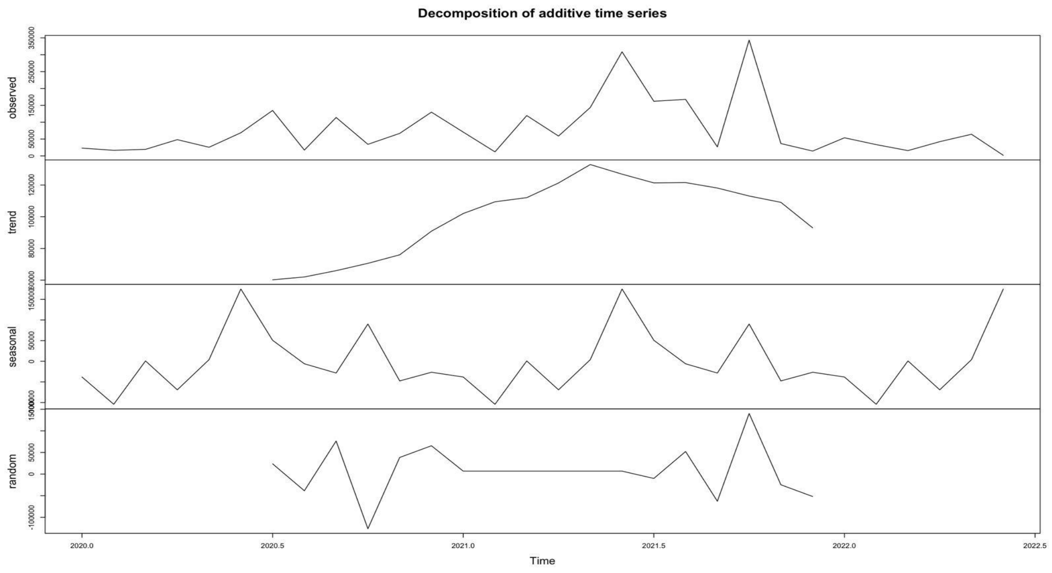


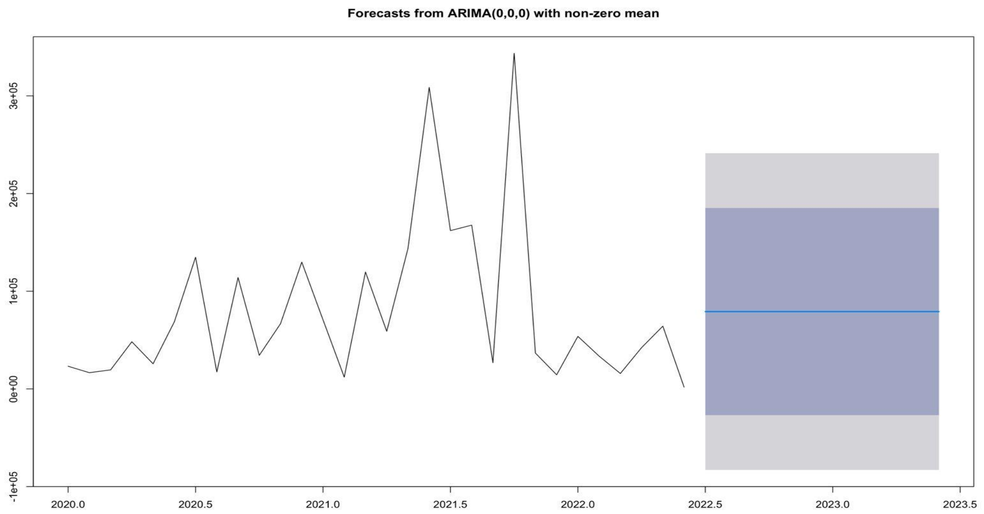


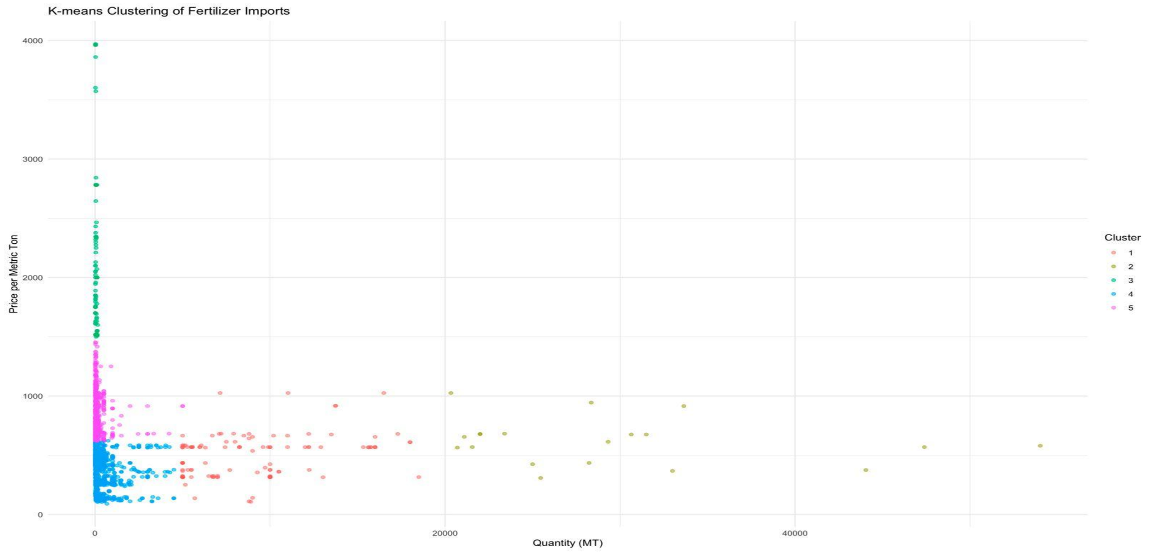
```
> model7 <- lm(PRICE.PER.MT ~ CASH.DATE, data = imports)
> summary(model7)

Call:
lm(formula = PRICE.PER.MT ~ CASH.DATE, data = imports)

Residuals:
   Min     1Q Median     3Q    Max 
-583.8 -184.5  -69.9  121.4 3453.9 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.096e+04  5.565e+02  -19.69   <2e-16 ***
CASH.DATE    6.163e-01  2.974e-02   20.72   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 368 on 2608 degrees of freedom
Multiple R-squared:  0.1414,	Adjusted R-squared:  0.1411 
F-statistic: 429.5 on 1 and 2608 DF,  p-value: < 2.2e-16
```

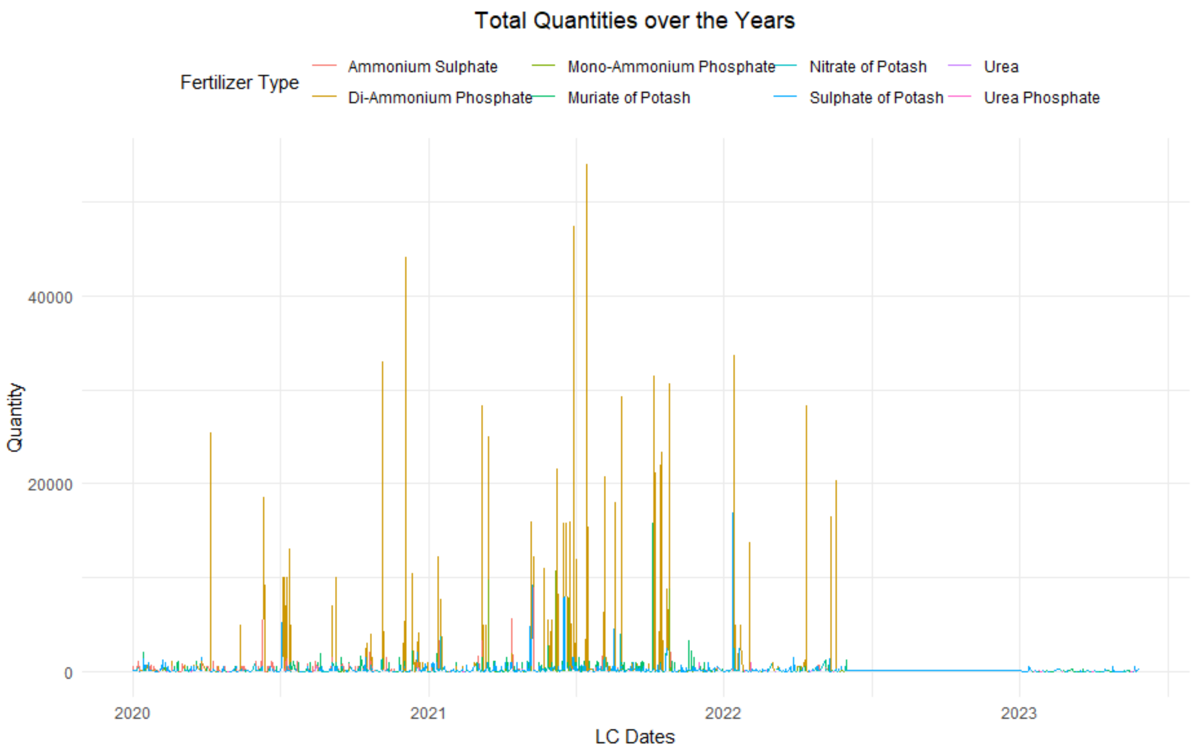


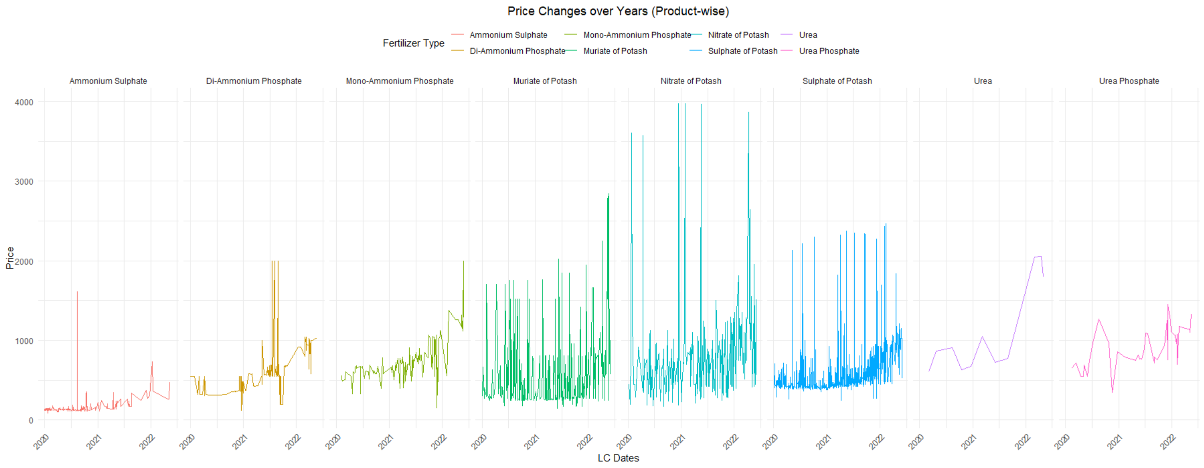
```
> model6 <- lm(QTY.MT ~ PRICE.PER.MT, data = imports)
> summary(model6)

Call:
lm(formula = QTY.MT ~ PRICE.PER.MT, data = imports)

Residuals:
   Min     1Q Median     3Q    Max 
 -1057   -867   -752   -467  53102 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1137.1681   112.4445  10.113   <2e-16 ***
PRICE.PER.MT   -0.3949     0.1609  -2.455   0.0141 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3262 on 2608 degrees of freedom
Multiple R-squared:  0.002306,	Adjusted R-squared:  0.001923 
F-statistic: 6.028 on 1 and 2608 DF,  p-value: 0.01415
```

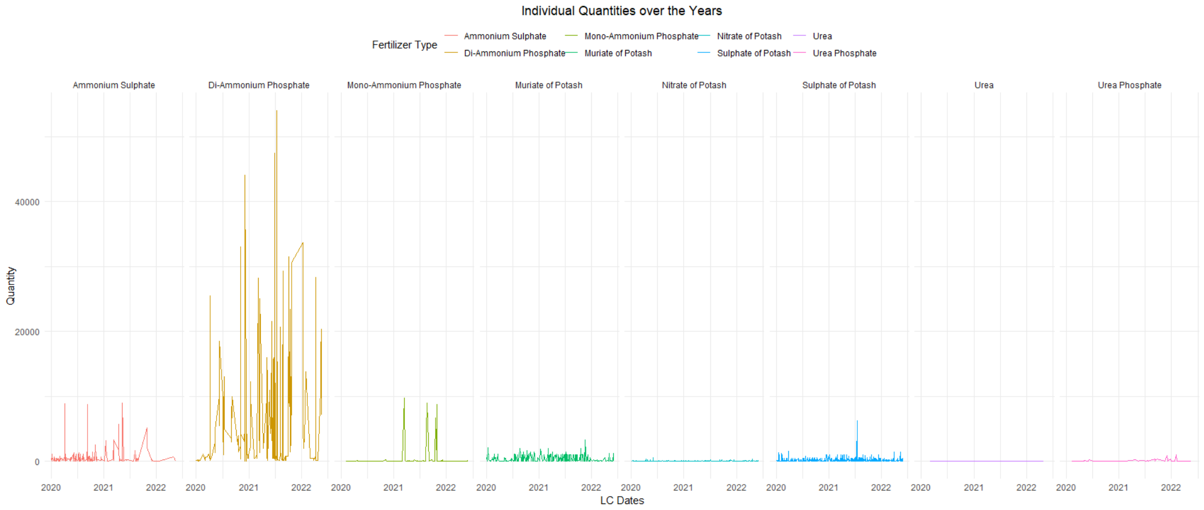

Decomposition of additive time series
observed
0
50000
150000
250000
350000
trend
80000
100000
120000
seasonal
-50000
0
50000
150000
random
-100000
0
50000
2020.0
2020.5
2021.0
2021.5
2022.0
2022.5
Time

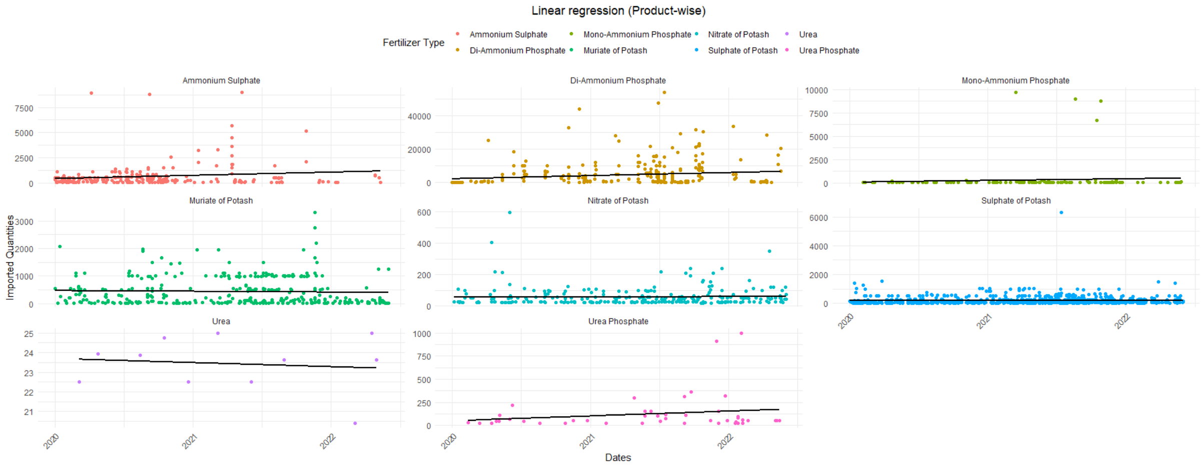

Forecasts from ARIMA(0,0,0) with non-zero mean
-1e+05
0e+00
1e+05
2e+05
3e+05
2020.0
2020.5
2021.0
2021.5
2022.0
2022.5
2023.0
2023.5

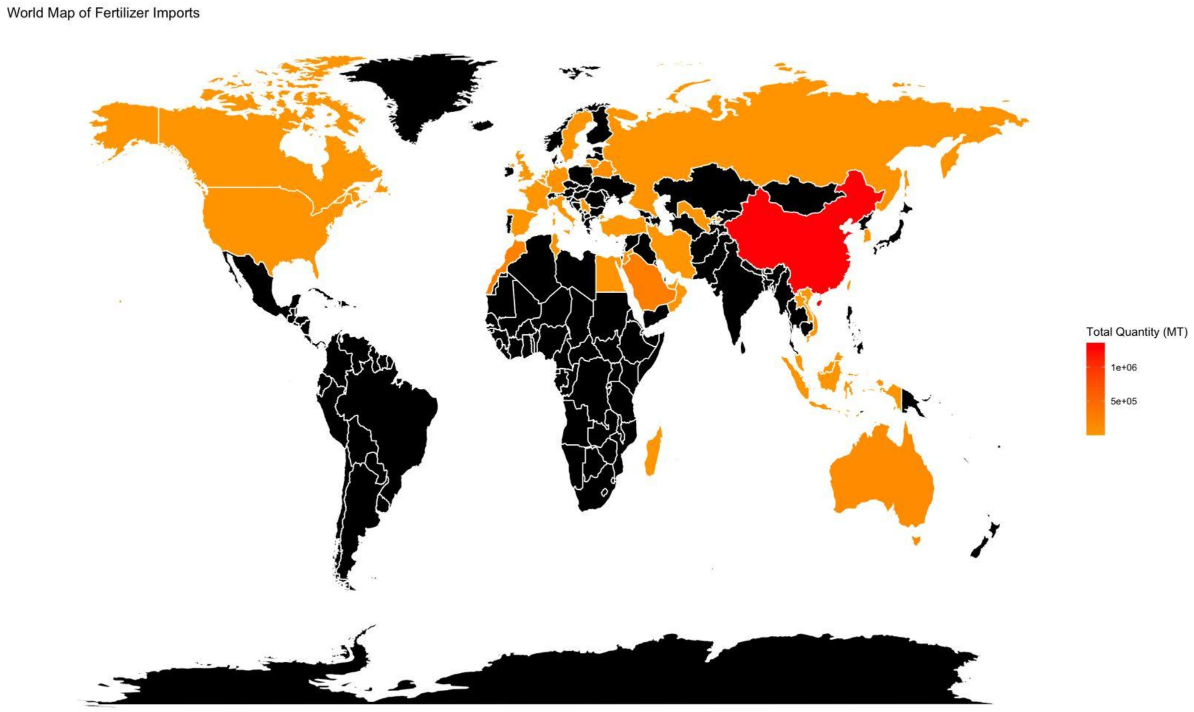

K-means Clustering of Fertilizer Imports
Price per Metric Ton
0
1000
2000
3000
4000
Quantity (MT)
0
20000
40000
Cluster
1
2
3
4
5


Total Quantities over the Years
Fertilizer Type
Ammonium Sulphate
Di-Ammonium Phosphate
Mono-Ammonium Phosphate
Muriate of Potash
Nitrate of Potash
Sulphate of Potash
Urea
Urea Phosphate
Quantity
40000
20000
0
2020
2021
2022
2023
LC Dates


Price Changes over Years (Product-wise)
Fertilizer Type
Ammonium Sulphate
Di-Ammonium Phosphate
Mono-Ammonium Phosphate
Muriate of Potash
Nitrate of Potash
Sulphate of Potash
Urea
Urea Phosphate
Price
4000
3000
2000
1000
0
2020
2021
2022
LC Dates


Individual Quantities over the Years
Fertilizer Type
Ammonium Sulphate
Di-Ammonium Phosphate
Mono-Ammonium Phosphate
Muriate of Potash
Nitrate of Potash
Sulphate of Potash
Urea
Urea Phosphate
Quantity
40000
20000
0
2020
2021
2022
LC Dates


Linear regression (Product-wise)
Fertilizer Type
Ammonium Sulphate
Di-Ammonium Phosphate
Mono-Ammonium Phosphate
Muriate of Potash
Nitrate of Potash
Sulphate of Potash
Urea
Urea Phosphate
Ammonium Sulphate
0
2500
5000
7500
Di-Ammonium Phosphate
0
20000
40000
Mono-Ammonium Phosphate
0
2500
5000
7500
10000
Muriate of Potash
0
1000
2000
3000
Nitrate of Potash
0
200
400
600
Sulphate of Potash
0
2000
4000
6000
Urea
21
22
23
24
25
Urea Phosphate
0
250
500
750
1000
Imported Quantities
2020
2021
2022
Dates


World Map of Fertilizer Imports
Total Quantity (MT)
1e+06
5e+05

Works Cited

The future of the fertilizer sector in Pakistan (2023) by Amir Jahangir: Of daily narratives, *Narratives Magazine*. Available at: https://narratives.com.pk/featured/the-future-of-the-fertilizer-sector-in-pakistan/ (Accessed: 11 December 2023).

PACRA, P.C.R.A. (2021) *Fertilzer Sector (An Overview)* , *https://www.pacra.com*. Available at: https://www.pacra.com/sector_research/Fertilizer%20Draft%20V.1_Updated_1611841077.pdf (Accessed: 10 December 2023).

PBIT, P.B. of I. and T. (2018) *Fertilizers Sector of Pakistan*, *pbit.gov.pk*. Available at: http://pbit.gop.pk/system/files/Fertilizer%20Sector%20Report%20.pdf (Accessed: 09 December 2023).

Tariq, M.B., Fareed, R. & others (2016) *Fertilizer Production and Import in Pakistan. Available at: https://www.academia.edu/30778347/Fertilizer_Production_and_Import_in_Pakistan (Accessed: 11th December 2023)*

R-Code

```
## Group Project
## Group 5
## Members: Ali Danish Riza (24110154)
##

## Setting Directory -----

getwd()
setwd("/Users/ahmedehtisham/Downloads")

## Packages -----

library(dplyr)
library(ggplot2)
library(dslabs)
library(ggthemes)
library(ggrepel)
library(tidyr)
library(stringr)
library(NbClust)
library(factoextra)
library(cluster)
library(viridis)

## Importing CSV files -----

df_fert <- read.csv("Pak_Fert_Import.csv", header = TRUE, stringsAsFactors = FALSE, colClasses = c(PCT = "character"))   # Used colClasses = c(PCT = "character") to avoid trailing 0s from being removed
View(df_fert)
str(df_fert)


## A. Data Cleaning -----

# 1. Data Selection and shortlisting

columns <- c("PCT", "ITEM_DESC", "ORIGIN","IMPORTER_NAME", "CONSIGNERS.NAME", "ASSVALD", "QTY.KG", "CASH_DATE")

imports <- df_fert %>%
  select(columns) %>%                # Selected shortlisted columns
  mutate(QTY.MT = QTY.KG/1000) %>%    # Added quantity by metric tons column
  rename(HSCODE = PCT, ASSDVAL = ASSVALD, IMPORTER.NAME = IMPORTER_NAME, ITEM.DESC = ITEM_DESC, CASH.DATE = CASH_DATE) %>%      # Renamed columns for easier interpretation
  mutate(PRICE.PER.MT = ASSDVAL/QTY.MT) %>%    # Added price per metric ton column
  select(-QTY.KG)                     # Removed quantity by kgs column

View(imports)
str(imports)


# 2. Convert Date column to Date format

imports$CASH.DATE <- as.Date(imports$CASH.DATE, format = "%d-%b-%y")


# 3. Filtering data based on HS Codes

imports <- imports %>%
  mutate(HSCODE = str_trim(HSCODE))
HSC.List <- sort(unique(imports$HSCODE), decreasing = FALSE)
HSC.List <- as.data.frame(HSC.List)
View(HSC.List)

imports <- imports %>%
  mutate(HSCODE = case_when(
    HSCODE == "3102.1" ~ "3102.1000", # Corrected values where trailing 0s had been removed
    HSCODE == "3102.21" ~ "3102.2100",
    HSCODE == "3102.3" ~ "3102.3000",
    HSCODE == "3102.4" ~ "3102.4000",
    HSCODE == "3102.6" ~ "3102.6000",
    HSCODE == "3102.9" ~ "3102.9000",
    HSCODE == "3104.2" ~ "3104.2000",
    HSCODE == "3104.3" ~ "3104.3000",
    HSCODE == "3104.9" ~ "3104.9000",
    HSCODE == "3105.1" ~ "3105.1000",
    HSCODE == "3105.2" ~ "3105.2000",
    HSCODE == "3105.3" ~ "3105.3000",
    HSCODE == "3105.4" ~ "3105.4000",
    HSCODE == "3105.59" ~ "3105.5900",
    HSCODE == "3105.9" ~ "3105.9000",
    TRUE ~ HSCODE
  ))

Hscodes <- c("3102.1000",             # Removed irrelevant products such as 3102.3000 which is Ammonium Nitrate an explosive not a fertilizer
             "3102.2100",
             "3102.4000",
             "3102.6000",
             "3102.9000",
             "3104.2000",
             "3104.3000",
             "3104.9000",
             "3105.1000",
             "3105.2000",
             "3105.3000",
             "3105.4000",
             "3105.5100",
             "3105.5900",
             "3105.9000")

imports <- imports %>%
  filter(HSCODE %in% Hscodes)


# 4. Standardizing Item descriptions based on HS Codes

imports <- imports %>%
  mutate(ITEM.DESC = case_when(
    HSCODE == "3102.1000" ~ "Urea",
    HSCODE == "3102.2100" ~ "Ammonium Sulphate",
    HSCODE == "3102.4000" ~ "Speciality Fertilizer",
    HSCODE == "3102.6000" ~ "Speciality Fertilizer",
    HSCODE == "3102.9000" ~ "Speciality Fertilizer",
    HSCODE == "3104.2000" ~ "Muriate of Potash",
    HSCODE == "3104.3000" ~ "Sulphate of Potash",
    HSCODE == "3104.9000" ~ "Muriate of Potash",
    HSCODE == "3105.1000" ~ "Speciality Fertilizer",
    HSCODE == "3105.2000" ~ "Nitrate of Potash",
    HSCODE == "3105.3000" ~ "Di-Ammonium Phosphate",
    HSCODE == "3105.4000" ~ "Mono-Ammonium Phosphate",
    HSCODE == "3105.5100" ~ "Speciality Fertilizer",
    HSCODE == "3105.5900" ~ "Urea Phosphate",
    HSCODE == "3105.9000" ~ "Speciality Fertilizer",
    TRUE ~ ITEM.DESC
  ))

View(imports)

Hscodes <- c("3102.1000",             # Removed speciality fertilizers as are not used entirely by farmers
             "3102.2100",
             "3104.2000",
             "3104.3000",
             "3104.9000",
             "3105.2000",
             "3105.3000",
             "3105.4000",
             "3105.5900")

imports <- imports %>%
  filter(HSCODE %in% Hscodes)


# 5. Converting variable classes

str(imports)

imports$HSCODE <- as.factor(imports$HSCODE)
imports$ITEM.DESC <- as.factor(imports$ITEM.DESC)
imports$ORIGIN <- as.factor(imports$ORIGIN)
imports$IMPORTER.NAME <- as.factor(imports$IMPORTER.NAME)
imports$CONSIGNERS.NAME <- as.factor(imports$CONSIGNERS.NAME)
imports$HSCODE <- as.factor(imports$HSCODE)

str(imports)

# 6. Checking for NA Values

table(is.na(imports))


## 7. Shortening country names

imports <- imports %>%
  mutate(ORIGIN = str_trim(ORIGIN))
sort(unique(imports$ORIGIN), decreasing = FALSE)

imports <- imports %>%
  mutate(ORIGIN = case_when(
    ORIGIN == "LAO PEOPLES DEMOCRATIC REPUBL" ~ "LAO",
    ORIGIN == "UNITED ARAB EMIRATES" ~ "UAE",
    ORIGIN == "IRAN (ISLAMIC REPUBLIC OF)" ~ "IRAN",
    ORIGIN == "GERMAN FEDR REPUBLIC" ~ "GERMANY",
    TRUE ~ ORIGIN))

imports$ORIGIN <- as.factor(imports$ORIGIN)

# 8. Removing values below 18 tons

```

```
imports <- imports %>%                # A single container has 18 tons of fertilizer. Entries below that are personal use samples
  filter(QTY.MT > 18)
str(imports)
View(imports)


## B. Exploratory Plots -----


# 1. Which fertilizers are imported mostly from certain countries?

ggplot(imports, aes(x = ORIGIN, fill = ITEM.DESC)) +
  geom_bar(position = "stack", stat = "count", width = 0.7) +
  scale_fill_viridis_d() +
  labs(title = "Most Imported Fertilizers by Country",
       x = "Origin Country",
       y = "Import Transactions Count",
       fill = "Fertilizers") +
  theme_minimal() +
  theme(legend.position = "top",
        axis.text.x = element_text(angle = 45, hjust = 1),
        plot.title = element_text(hjust = 0.5))


# 2. Are there any visible trends in total fertilizer import quantities over the years

#         i) Composite

ggplot(imports, aes(x = CASH.DATE, y = QTY.MT, color = ITEM.DESC, group = 1)) +
  geom_line() +
  labs(title = "Total Quantities over the Years",
       x = "LC Dates",
       y = "Quantity",
       color = "Fertilizer Type") +
  theme_minimal() +
  theme(legend.position = "top",
        plot.title = element_text(hjust = 0.5))

imports <- imports %>%                # Removing values after 2022 are there is a gap in the recorded data
  filter(CASH.DATE < "2023-01-01")

#         ii) Product-wise

```

```
ggplot(imports, aes(x = CASH.DATE, y = QTY.MT, color = ITEM.DESC, group = 1)) +
  geom_line() +
  labs(title = "Individual Quantities over the Years",
       x = "LC Dates",
       y = "Quantity",
       color = "Fertilizer Type") +
  theme_minimal() +
  theme(legend.position = "top",
        plot.title = element_text(hjust = 0.5)) +
  facet_grid(~ITEM.DESC)


# 3. Do Prices change over time?

#         i) Composite

ggplot(imports, aes(x = CASH.DATE, y = PRICE.PER.MT, color = ITEM.DESC, group = 1)) +
  geom_line() +
  labs(title = "Price Changes over Years (Composite)",
       x = "LC Dates",
       y = "Price",
       color = "Fertilizer Type") +
  theme_minimal() +
  theme(legend.position = "top",
        plot.title = element_text(hjust = 0.5),
        axis.text.x = element_text(angle = 45, hjust = 1))

#         ii) Product-wise

ggplot(imports, aes(x = CASH.DATE, y = PRICE.PER.MT, color = ITEM.DESC, group = 1)) +
  geom_line() +
  labs(title = "Price Changes over Years (Product-wise)",
       x = "LC Dates",
       y = "Price",
       color = "Fertilizer Type") +
  theme_minimal() +
  theme(legend.position = "top",
        plot.title = element_text(hjust = 0.5),
        axis.text.x = element_text(angle = 45, hjust = 1)) +
  facet_grid(~ITEM.DESC)


# 4. How Prices Impact Import Quantities ?

#         i) Composite
```

```

ggplot(imports, aes(x = QTY.MT , y = PRICE.PER.MT, color = ITEM.DESC, group = 1)) +
  geom_line() +
  labs(title = "How Prices impact quantities (Composite)",
       x = "Import Quantities",
       y = "Price",
       color = "Fertilizer Type") +
  theme_minimal() +
  theme(legend.position = "top",
        plot.title = element_text(hjust = 0.5),
        axis.text.x = element_text(angle = 45, hjust = 1))

#         ii) Product-wise

ggplot(imports, aes(x = QTY.MT , y = PRICE.PER.MT, color = ITEM.DESC, group = 1)) +
  geom_line() +
  labs(title = "How Prices impact quantities (Product-wise)",
       x = "Import Quantities",
       y = "Price",
       color = "Fertilizer Type") +
  theme_minimal() +
  theme(legend.position = "top",
        plot.title = element_text(hjust = 0.5),
        axis.text.x = element_text(angle = 45, hjust = 1)) +
  facet_grid(~ITEM.DESC)


# 5. Does product-wise linear regression show any recognizable patterns in the data?

ggplot(imports, aes(x = CASH.DATE, y = QTY.MT, color = ITEM.DESC)) +
  geom_point() +  # Use points to plot the actual data
  geom_smooth(aes(group = 1), method = "lm", se = FALSE, color = "black") +
  labs(title = "Linear regression (Product-wise)",
       x = "Dates",
       y = "Imported Quantities",
       color = "Fertilizer Type") +
  theme_minimal() +
  theme(legend.position = "top",
        plot.title = element_text(hjust = 0.5),
        axis.text.x = element_text(angle = 45, hjust = 1)) +
  facet_wrap(~ITEM.DESC, scales = "free_y")
```

```



## C. Clustering -----

# 1. Variable selection

str(imports)

dist_matrix_data <- imports %>%       # Excluded date and
  select(ORIGIN, ITEM.DESC, QTY.MT, IMPORTER.NAME, CONSIGNERS.NAME, ASSDVAL, QTY.MT)


# 2. Calculating Distance Matrix

dist_matrix <- daisy(dist_matrix_data, metric = "gower")


# 3. Hierarchical Clustering

#         i) Testing Linkages

hc.c <- hclust(dist_matrix, method = "complete") # Creates a better interpretable deprogram
plot(hc.c, hang = -1, main = "Complete linkage Dendrogram", cex.axis = 0.7)

hc.a <- hclust(dist_matrix, method = "average")
plot(hc.c, hang = -1, main = "Average linkage Dendrogram", cex.axis = 0.7)

summary(hc.c)

#        ii) Testing Number of Clusters

#             a) Creating an Elbow Diagram

fviz_nbclust(dist_matrix_data, FUN = hcut, method = "wss")

#             b) Testing with 2 clusters

Clusters2 <- cutree(hc.c,k=2)
Clusters2
table(Clusters2)

rect.hclust(hc.c, k=2, border = "red")

```

```
dataframe <- data.frame(dist_matrix_data, Clusters2)

View(dataframe)

Clustered_data_Origin <- dataframe %>%
  group_by(Clusters2, ORIGIN) %>%
  summarise(Count = n())

View(Clustered_data_Origin)

Clustered_data_Products <- dataframe %>%
  group_by(Clusters2, ITEM.DESC) %>%
  summarise(Count = n())

View(Clustered_data_Products)

#             c) Testing with 3 clusters

Cluster3 <- cutree(hc.c,k=3)
Cluster3
table(Cluster3)

dataframe2 <- data.frame(dist_matrix_data, Cluster3)

View(dataframe2)

Clustered_data_Origin <- dataframe %>%
  group_by(Clusters2, ORIGIN) %>%
  summarise(Count = n())

View(Clustered_data_Origin)

Clustered_data_Products <- dataframe %>%
  group_by(Clusters2, ITEM.DESC) %>%
  summarise(Count = n())

View(Clustered_data_Products)

#           d) Plotting silhouettes

plot(silhouette(cutree(hc.c, k = 2), dist_matrix))
plot(silhouette(cutree(hc.c, k = 3), dist_matrix))


# 4. Finalized Deprogram

hc.f <- hclust(dist_matrix, method = "complete")
plot(hc.f, hang = -1, main = "Complete linkage Dendrogram")

Clusters.f <- cutree(hc.f,k=2)
Clusters.f
table(Clusters.f)

rect.hclust(hc.f, k=2, border = "red")

dataframe3 <- data.frame(dist_matrix_data, Clusters.f)

# 5. Analysis

Clustered_data_Origin <- dataframe3 %>%
  group_by(Clusters.f, ORIGIN) %>%
  summarise(Count = n())

View(Clustered_data_Origin)

Clustered_data_Products <- dataframe3 %>%
  group_by(Clusters.f, ITEM.DESC) %>%
  summarise(Count = n())

View(Clustered_data_Products)

dataframe3 %>%
  group_by(Clusters.f) %>%
  summarize(Mean.QTY.MT = mean(QTY.MT), Mean.ASSDVAL = mean(ASSDVAL))

dataframe3 %>%
  group_by(Clusters.f) %>%
  summarize(Max.QTY.MT = max(QTY.MT), Max.ASSDVAL = max(ASSDVAL))

dataframe3 %>%
  group_by(Clusters.f) %>%
  summarize(Min.QTY.MT = min(QTY.MT), Min.ASSDVAL = min(ASSDVAL))


## D. Regression -----

str(imports)

# 1. Model 1

model <- lm(QTY.MT ~ ITEM.DESC + ASSDVAL + ORIGIN + CASH.DATE, data = imports)
summary(model)

# 2. Model 2

model2 <- lm(QTY.MT ~ ITEM.DESC + PRICE.PER.MT + CASH.DATE + ASSDVAL, data = imports)
```

```
summary(model2)

# 3. Model 3 (Best Fit)

total.qt <- imports %>%
  group_by(ITEM.DESC) %>%
  summarise(SUM = sum(QTY.MT))

View(total.qt)

relevant <- c("Di-Ammonium Phosphate", "Nitrate of Potash", "Sulphate of Potash", "Ammonium Sulphate")

rel.prods <- imports %>%
  filter(ITEM.DESC %in% relevant)
unique(rel.prods$ITEM.DESC)

model3 <- lm(QTY.MT ~ ITEM.DESC + PRICE.PER.MT + CASH.DATE + ASSDVAL, data = rel.prods)
summary(model3)

# 4. Model 4

model4 <- lm(ASSDVAL ~ ITEM.DESC + CASH.DATE + PRICE.PER.MT, data = imports)
summary(model4)

ggplot(imports, aes(x = CASH.DATE, y = QTY.MT, color = ITEM.DESC, group = 1)) +
  geom_line() +
  labs(title = "Quantities against Dates",
       x = "Dates",
       y = "Quantity") +
  theme_minimal() +
  abline(model5)

# 5. Model 5

# QTY v DATE + ABLINE

model5 <- lm(QTY.MT ~ CASH.DATE, data = imports)
summary(model5)

ggplot(imports, aes(x = CASH.DATE, y = QTY.MT, color = ITEM.DESC, group = 1)) +
  geom_line() +
  labs(title = "Quantities against Dates",
       x = "Dates",
       y = "Quantity") +
  theme_minimal() +
  abline(model5)
```

```

# 6. Model 6
# QTY v PRICE + ABLINE

model6 <- lm(QTY.MT ~ PRICE.PER.MT, data = imports)
summary(model6)

ggplot(imports, aes(x = PRICE.PER.MT, y = QTY.MT, color = ITEM.DESC, group = 1)) +
  geom_point() +
  labs(title = "Quantities against Price",
       x = "Price",
       y = "Quantity") +
  theme_minimal() +
  abline(model6)

# 7. Model 7
# PRICE v DATE + ABLINE

model7 <- lm(PRICE.PER.MT ~ CASH.DATE, data = imports)
summary(model7)

ggplot(imports, aes(x = CASH.DATE, y = PRICE.PER.MT, color = ITEM.DESC, group = 1)) +
  geom_line() +
  labs(title = "Price against Dates",
       x = "Dates",
       y = "Price") +
  theme_minimal() +
  abline(model7)

# EDA - Summary statistics for each cluster
cluster_summary <- clustered_data %>%
  group_by(Cluster) %>%
  summarise(mean_QTY.MT = mean(QTY.MT), mean_ASSDVAL = mean(ASSDVAL))

# Visualization - Box plot of QTY.MT by cluster
ggplot(clustered_data, aes(x = as.factor(Cluster), y = QTY.MT)) +
  geom_boxplot() +
  labs(title = "Box Plot of QTY.MT by Cluster") +
  xlab("Cluster") +
  ylab("QTY.MT")

# Visualization - Bar plot of cluster counts
ggplot(clustered_data, aes(x = as.factor(Cluster))) +
  geom_bar() +
  labs(title = "Bar Plot of Cluster Counts") +
  xlab("Cluster") +
  ylab("Count")
```

```
# Decision Tree ####
# Select a target variable ('QTY.MT') and predictors
target_variable <- "QTY.MT"
predictor_variables <- c("PRICE.PER.MT", "ASSDVAL", "CASH.DATE")

# Install the rpart package if not already installed
if (!requireNamespace("rpart", quietly = TRUE)) {
  install.packages("rpart")
}

# Load the rpart package
install.packages("rpart")
install.packages("rpart.plot")
library(rpart)
library(rpart.plot)

# Build decision tree for Cluster 1
tree_cluster1 <- rpart(formula(paste(target_variable, "~", paste(predictor_variables, collapse = "+"))),
                       data = filter(clustered_data, Cluster == 1))

# Visualize decision tree for Cluster 1
rpart.plot(tree_cluster1)


# Time Series analysis ####

# Assuming you have a date column (CASH.DATE) and a quantity column (QTY.MT) in your 'imports' dataframe

library(ggplot2)
install.packages("forecast")
install.packages("tseries")
library(forecast)
library(tseries)

# Convert the date to a Date object if not already
imports$CASH.DATE <- as.Date(imports$CASH.DATE, format = "%d-%b-%y")

# Aggregate data by month for time series analysis
monthly_data <- aggregate(QTY.MT ~ format(CASH.DATE, "%Y-%m"), data=imports, sum)
colnames(monthly_data) <- c("Month", "TotalQuantity")

# Convert to time series object
ts_data <- ts(monthly_data$TotalQuantity, start=c(as.numeric(substr(min(monthly_data$Month), 1,
4)), as.numeric(substr(min(monthly_data$Month), 6, 7))), frequency=12)

# Decompose the time series to observe trends and seasonality
decomposed_ts <- decompose(ts_data)
plot(decomposed_ts)


## Mapping ####
# Assuming you have a country of origin column (ORIGIN) in your 'imports' dataframe

library(ggplot2)
library(maps)

# Aggregate data by country
country_data <- aggregate(QTY.MT ~ ORIGIN, data=imports, sum)
names(country_data) <- c("region", "TotalQuantity")  # Make sure the column names match for merging

# Merge with world map data
world_map <- map_data("world")
merged_data <- merge(world_map, country_data, by="region", all.x=TRUE)

# Plot
ggplot(merged_data, aes(x = long, y = lat, group = group, fill = TotalQuantity)) +
  geom_polygon(color = "black") +
  scale_fill_continuous(low = "blue", high = "red", na.value = "grey50", name="Total Quantity (MT)") +
  labs(title = "World Map of Fertilizer Imports", x = "", y = "") +
  theme_minimal() +
  theme(legend.position = "bottom")



# Load necessary libraries
library(ggplot2)
library(maps)
library(dplyr)

# Load your data (assuming it's already loaded in 'imports')
# imports <- read.csv("your_data.csv")

# Aggregate data by country
country_data <- imports %>%
  group_by(ORIGIN) %>%
  summarise(TotalQuantity = sum(QTY.MT)) %>%
  ungroup()

# Load world map data for comparison
world_map <- map_data("world")
map_countries <- unique(world_map$region)

# Checking for country name mismatches
data_countries <- unique(country_data$ORIGIN)
mismatched_countries <- setdiff(data_countries, map_countries)

# Print mismatched country names for manual correction
print(mismatched_countries)

# Manual adjustments for country names
imports$ORIGIN[imports$ORIGIN == "KOREA (SOUTH)"] <- "South Korea"
imports$ORIGIN[imports$ORIGIN == "UNITED STATES"] <- "USA"
imports$ORIGIN[imports$ORIGIN == "RUSSIAN FEDRATION"] <- "Russia"
imports$ORIGIN[imports$ORIGIN == "UNITED KINGDOM"] <- "UK"
imports$ORIGIN[imports$ORIGIN == "CHINA"] <- "China"
imports$ORIGIN[imports$ORIGIN == "AUSTRALIA"] <- "Australia"
imports$ORIGIN[imports$ORIGIN == "BELARUS"] <- "Belarus"
imports$ORIGIN[imports$ORIGIN == "BELGIUM"] <- "Belgium"
imports$ORIGIN[imports$ORIGIN == "CANADA"] <- "Canada"
# ... continue for other countries

# After adjusting the names, aggregate the data by country
country_data <- imports %>%
  group_by(ORIGIN) %>%
  summarise(TotalQuantity = sum(QTY.MT)) %>%
  ungroup()

# Merge with world map data
world_map <- map_data("world")
merged_data <- merge(world_map, country_data, by.x = "region", by.y = "ORIGIN", all.x = TRUE)

# Plot the map
ggplot(merged_data, aes(long, lat, group = group, fill = TotalQuantity)) +
  geom_polygon() +
  scale_fill_continuous(low = "white", high = "red") +
  labs(title = "World Map Showing Fertilizer Imports by Country", fill = "Total Quantity (MT)") +
  theme_void()


# Install and load necessary packages
if (!requireNamespace("dplyr", quietly = TRUE)) install.packages("dplyr")
if (!requireNamespace("ggplot2", quietly = TRUE)) install.packages("ggplot2")
if (!requireNamespace("maps", quietly = TRUE)) install.packages("maps")

library(dplyr)
library(ggplot2)
library(maps)

# Mapping table (example: you need to complete this table based on your data)
country_mapping <- data.frame(
  OriginalName = c("CHINA", "UNITED ARAB EMIRATES", "KOREA (SOUTH)", "UNITED KINGDOM", "GERMANY", "UAE", "USA", "SOUTH KOREA"),
  MapName = c("China", "UAE", "South Korea", "UK", "Germany", "UAE", "United States", "South Korea")
)

# Join your imports data with this mapping table
mapped_imports <- imports %>%
  left_join(country_mapping, by = c("ORIGIN" = "OriginalName"))

# Replace NA in MapName with ORIGIN
mapped_imports$MapName <- ifelse(is.na(mapped_imports$MapName), mapped_imports$ORIGIN, mapped_imports$MapName)

# Create a total quantity per country
country_data <- mapped_imports %>%
  group_by(MapName) %>%
  summarise(TotalQuantity = sum(QTY.MT)) %>%
  na.omit()

# World map plot
world_map <- map_data("world")
ggplot() +
  geom_polygon(data = world_map, aes(x = long, y = lat, group = group), fill = "gray", color = "white") +
  geom_point(data = country_data, aes(x = MapName, y = TotalQuantity), size = 3, color = "blue") +
  theme_minimal()

# Summarize total quantity per country
country_totals <- imports %>%
  group_by(ORIGIN) %>%
  summarise(TotalQuantity = sum(QTY.MT)) %>%
  na.omit()

# World map data
world_map <- map_data("world")

# Prepare data for plotting
plot_data <- merge(world_map, country_totals, by.x = "region", by.y = "ORIGIN", all.x = TRUE)

# Plotting
ggplot(data = plot_data, aes(x = long, y = lat, group = group, fill = TotalQuantity)) +
  geom_polygon(color = "white") +
  scale_fill_continuous(low = "gray", high = "red", na.value = "gray", guide = "colorbar") +
  theme_minimal() +
  labs(fill = "Total Quantity (MT)", title = "World Map Highlighting Fertilizer Imports")

# Load necessary packages
library(dplyr)
library(maps)

# Assuming 'imports' is your dataset with a column 'ORIGIN'

# Get unique country names from your dataset
unique_countries_imports <- unique(imports$ORIGIN)

# Get unique country names from the world map dataset
world_map <- map_data("world")
unique_countries_world_map <- unique(world_map$region)

# Compare the country names
mismatched_countries <- setdiff(unique_countries_imports, unique_countries_world_map)

# Output mismatched countries
print(mismatched_countries)

# Create a mapping of mismatched country names
country_mapping <- c("SINGAPORE" = "Singapore", "TAIWAN" = "Taiwan", "SPAIN" = "Spain", "TURKEY" = "Turkey",
                     "HONG KONG CHINA" = "China", "NETHERLANDS" = "Netherlands", "EGYPT" = "Egypt",
                     "UAE" = "United Arab Emirates", "GERMANY" = "Germany", "JORDAN" = "Jordan",
                     "SWEDEN" = "Sweden", "EUROPIEN UNION" = NA, "SAUDI ARABIA" = "Saudi Arabia",
                     "ITALY" = "Italy", "VIET NAM" = "Vietnam", "TUNISIA" = "Tunisia",
                     "MOROCCO" = "Morocco", "INDONESIA" = "Indonesia", "UZBEKISTAN" = "Uzbekistan",
                     "LITHUANIA" = "Lithuania", "FRANCE" = "France", "LAO" = "Laos", "IRAN" = "Iran",
                     "MADAGASCAR" = "Madagascar", "MALAYSIA" = "Malaysia", "OMAN" = "Oman",
                     "SERBIA" = "Serbia", "CAYMAN ISLANDS" = "Cayman Islands")

# Update country names in the dataset
imports$ORIGIN <- ifelse(imports$ORIGIN %in% names(country_mapping), country_mapping[imports$ORIGIN], imports$ORIGIN)

# Now let's try plotting again with the corrected country names
library(ggplot2)
library(maps)

world_map <- map_data("world")

# Group the data by ORIGIN and calculate total import quantity for each country
country_imports <- imports %>%
  group_by(ORIGIN) %>%
  summarise(TotalQuantity = sum(QTY.MT)) %>%
  filter(!is.na(ORIGIN))

# Join the world map data with your country import data
world_map <- world_map %>%
  left_join(country_imports, by = c("region" = "ORIGIN"))

# Plot the map
ggplot() +
  geom_polygon(data = world_map, aes(x = long, y = lat, group = group, fill = TotalQuantity), color = "white") +
  scale_fill_continuous(low = "grey", high = "red", na.value = NA) +
  labs(title = "World Map of Fertilizer Imports", fill = "Total Quantity (MT)") +
  theme_void()

```

```
# Load and prepare the world map data
world_map <- map_data("world")

# Create a mapping of mismatched country names
country_mapping <- c("SINGAPORE" = "Singapore", "TAIWAN" = "Taiwan", "SPAIN" = "Spain", "TURKEY" = "Turkey",
                     "HONG KONG CHINA" = "China", "NETHERLANDS" = "Netherlands", "EGYPT" = "Egypt",
                     "UAE" = "United Arab Emirates", "GERMANY" = "Germany", "JORDAN" = "Jordan",
                     "SWEDEN" = "Sweden", "EUROPIEN UNION" = NA, "SAUDI ARABIA" = "Saudi Arabia",
                     "ITALY" = "Italy", "VIET NAM" = "Vietnam", "TUNISIA" = "Tunisia",
                     "MOROCCO" = "Morocco", "INDONESIA" = "Indonesia", "UZBEKISTAN" = "Uzbekistan",
                     "LITHUANIA" = "Lithuania", "FRANCE" = "France", "LAO" = "Laos", "IRAN" = "Iran",
                     "MADAGASCAR" = "Madagascar", "MALAYSIA" = "Malaysia", "OMAN" = "Oman",
                     "SERBIA" = "Serbia", "CAYMAN ISLANDS" = "Cayman Islands")

# Update country names in the dataset
imports$ORIGIN <- ifelse(imports$ORIGIN %in% names(country_mapping), country_mapping[imports$ORIGIN], imports$ORIGIN)

# Group the data by ORIGIN and calculate total import quantity for each country
country_imports <- imports %>%
  group_by(ORIGIN) %>%
  summarise(TotalQuantity = sum(QTY.MT), .groups = 'drop') %>%
  filter(!is.na(ORIGIN))

# Merge the world map data with your country import data
world_map_imports <- world_map %>%
  left_join(country_imports, by = c("region" = "ORIGIN"))

# Plot the map
ggplot() +
  geom_polygon(data = world_map_imports, aes(x = long, y = lat, group = group, fill = TotalQuantity), color = "white") +
  scale_fill_continuous(low = "orange", high = "red", na.value = "black") +
  labs(title = "World Map of Fertilizer Imports", fill = "Total Quantity (MT)") +
  theme_void()


# K-mean Clustering ####

# Ensure required packages are installed and loaded
if (!requireNamespace("cluster", quietly = TRUE)) install.packages("cluster")
library(cluster)

# Prepare data for clustering
clustering_data <- imports %>%
  select(QTY.MT, PRICE.PER.MT) %>%
  na.omit()

# Scale the data
clustering_data_scaled <- scale(clustering_data)

# K-means clustering
set.seed(123)  # Setting seed for reproducibility
kmeans_result <- kmeans(clustering_data_scaled, centers = 5, nstart = 25)

# Add cluster information to the original data
imports$Cluster <- kmeans_result$cluster

# Visualizing the clusters
ggplot(imports, aes(x = QTY.MT, y = PRICE.PER.MT, color = as.factor(Cluster))) +
  geom_point(alpha = 0.5) +
  labs(title = "K-means Clustering of Fertilizer Imports",
       x = "Quantity (MT)",
       y = "Price per Metric Ton",
       color = "Cluster") +
  theme_minimal()


# Comparative Analysis ####
# Comparative analysis by country
country_analysis <- imports %>%
  group_by(ORIGIN) %>%
  summarise(TotalImports = sum(QTY.MT),
            AvgPricePerMT = mean(PRICE.PER.MT, na.rm = TRUE)) %>%
  arrange(desc(TotalImports))

# Visualization
```

```
ggplot(country_analysis, aes(x = reorder(ORIGIN, TotalImports), y = TotalImports, fill = AvgPricePerMT)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  scale_fill_viridis_c() +
  labs(title = "Comparative Analysis of Fertilizer Imports by Country",
       x = "Country",
       y = "Total Imports (MT)",
       fill = "Average Price Per MT") +
  theme_minimal()
```