Blow Moulding price prediction

Reesu Jagan

Agenda

Aim

Introduction

EDA

Pre-processing

Results

Aim

- The goal is to develop a predictive model that can estimate or forecast the price of "Domestic Market (Contract) Blow Moulding, Low."
- The term "Domestic Market" indicates that the focus is on the local market rather than international or global markets.

Introduction:

1.Predicting the price of blow moulding contracts can be valuable for various stakeholders, including manufacturers, suppliers, and customers.

2.Accurate price predictions can assist manufacturers in determining optimal pricing strategies, negotiating contracts, and managing production costs.

3.Suppliers can use price predictions to optimize their supply chain and inventory management.

4.Customers, on the other hand, can make informed decisions based on price forecasts when procuring blow moulding services.

About Dataset:

- The provided dataset consists of 276 rows and 50 columns, representing data on a monthly basis from the years 2000 to 2022.

- The provided data includes various columns with missing data. It appears to be a collection of different variables related to commodities, market prices, protests, and import/export information. Here's a breakdown of the available columns and their categories:

1. Commodity Prices:
   - WTISPLC: Spot Crude Oil Price: West Texas Intermediate (WTI)
   - MHHNGSP: Henry Hub Natural Gas
   - MCOILBRENTEU: Crude Oil Prices: Brent - Europe
   - GASREGM: US Regular All Formulations Gas Price
   - PRUBBUSDM: Global price of Rubber

2. Producer Price Indices:

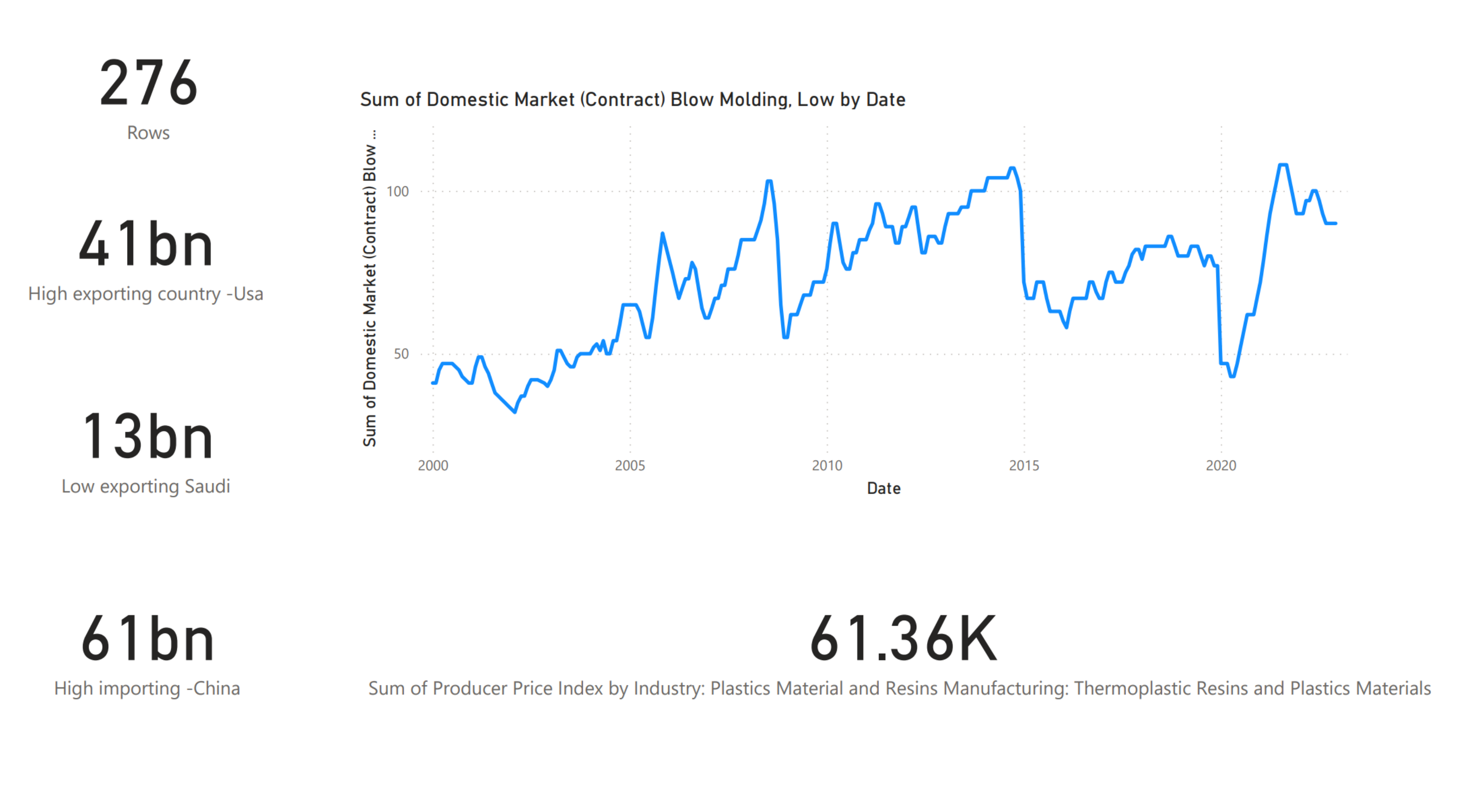
- WPUFD4111: Producer Price Index by Commodity for Final Demand: Finished Consumer Foods
- PCU325211325211: Producer Price Index by Industry: Plastics Material and Resins Manufacturing
- PCU32611332611301: Producer Price Index by Industry: Nonpackaging Plastics Film and Sheet Manufacturing
- WPU0915021625: Producer Price Index by Commodity: Pulp, Paper, and Allied Products: Other Polyethylene Bags, Pouches, and Liners
- PCU3261113261112: Producer Price Index by Industry: Plastics Bag and Pouch Manufacturing: Polyethylene Refuse Bags
- WPU072205011: Producer Price Index by Commodity: Rubber and Plastic Products: Unlaminated Polyethylene Film and Sheet
- PCU3252113252111: Producer Price Index by Industry: Plastics Material and Resins Manufacturing: Thermoplastic Resins and Plastics Materials
- PCU3252132521: Producer Price Index by Industry: Resin and Synthetic Rubber Manufacturing

3. Export and Import:

The data provides information on exports and imports for various countries, specifically related to a specific commodity (Ethylene polymers) with the HS code 390120. The countries include:

- Australia export/import
- Canada export/import
- Saudi export/import
- USA export/import
- India export/import
- Russia export/import
- SouthAfrica_export/import
- Turkey export
- Argentina export/import
- Brazil export
- Mexico export/import
- Italy export/import
- United Kingdom export/import
- China export/import
- Indonesia export/import
- Japan export/import
- South_Korea_import

Exploratory data analysis:



276

Rows

41bn

High exporting country -Usa

13bn

Low exporting Saudi

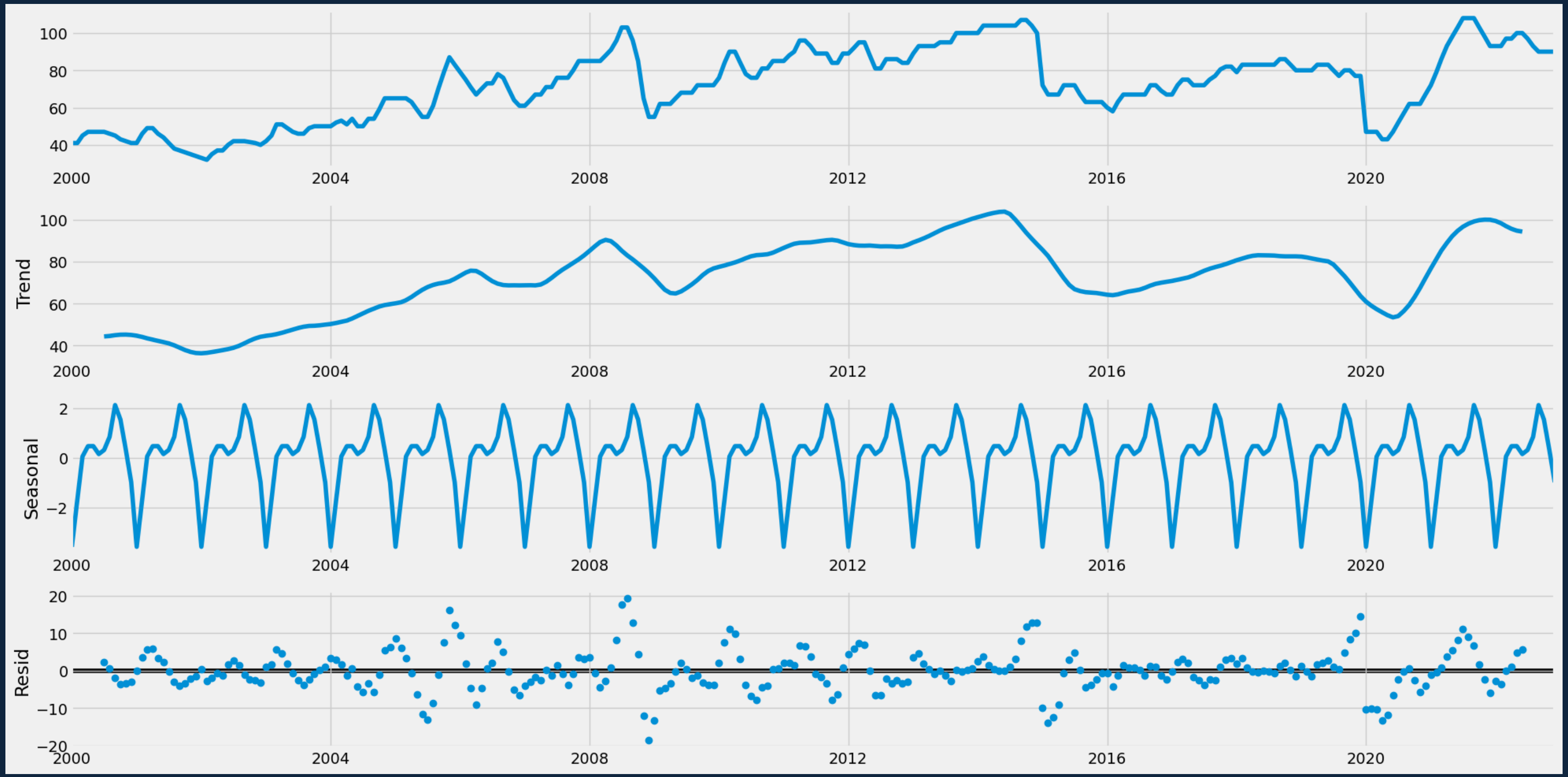
61bn

High importing -China

Sum of Domestic Market (Contract) Blow Molding, Low by Date

61.36K

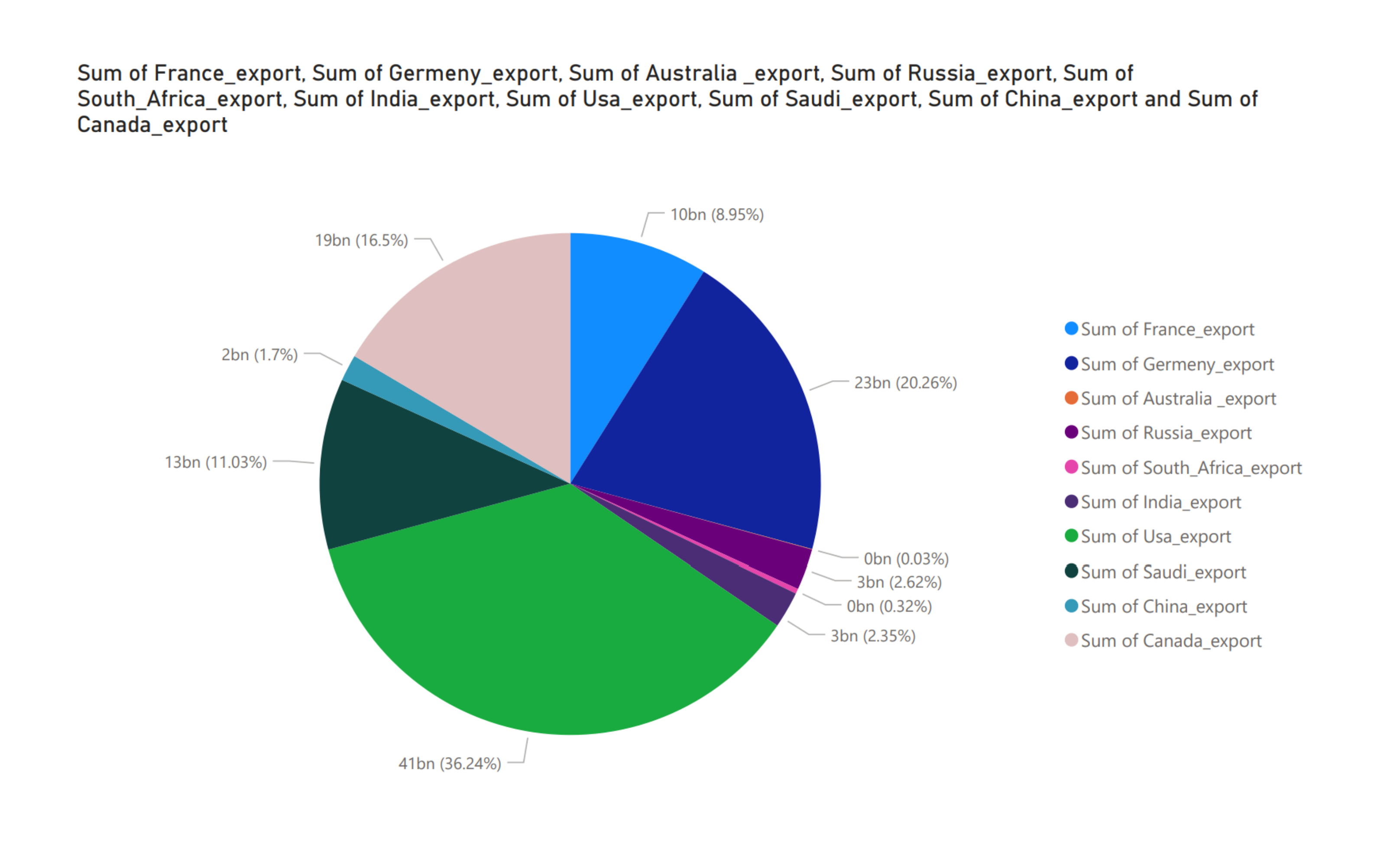
Sum of Producer Price Index by Industry: Plastics Material and Resins Manufacturing: Thermoplastic Resins and Plastics Materials



1. The time series exhibits seasonality patterns that repeat on a yearly basis. This suggests that there are recurring patterns or fluctuations within each year, indicating a seasonal component in the data.

2. A noticeable uptrend is observed in the time series until 2015, indicating a consistent increase in the values over time. However, starting from 2016, there is a subsequent uptrend, indicating a renewed upward movement in the values. This suggests a shift or change in the underlying trend of the time series.

Percentage of countries Exports



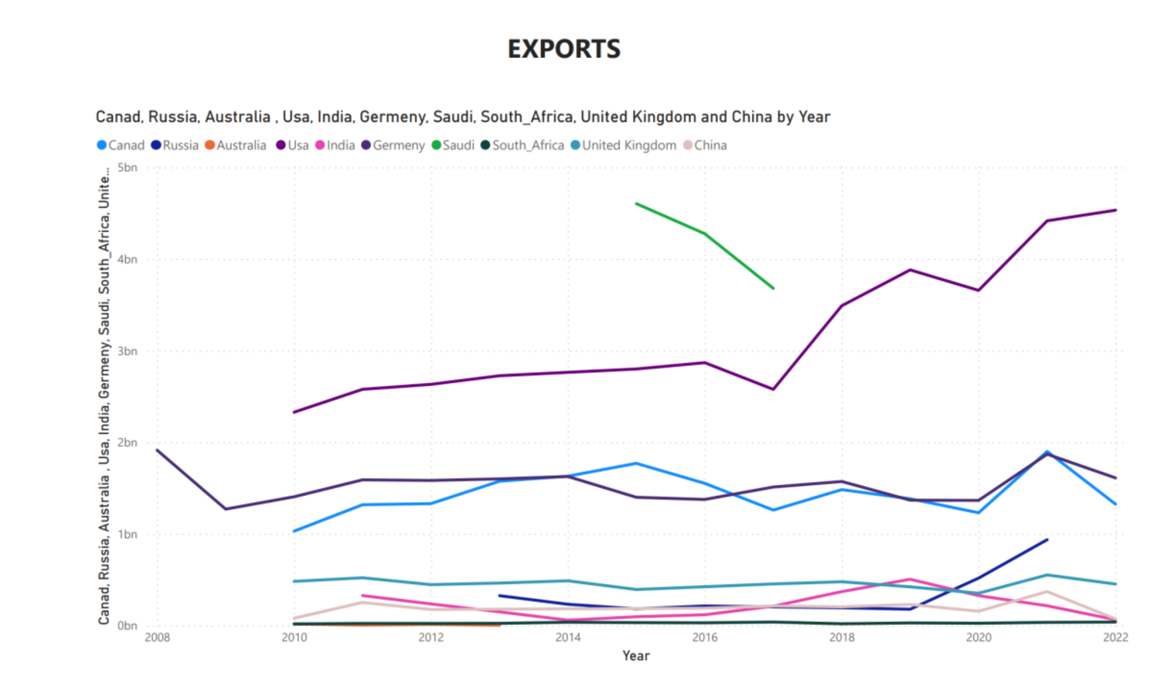
Sum of France_export, Sum of Germeny_export, Sum of Australia _export, Sum of Russia_export, Sum of South_Africa_export, Sum of India_export, Sum of Usa_export, Sum of Saudi_export, Sum of China_export and Sum of Canada_export

Insights:

1. 1. From 2000 to 2008, there were no exports reported. This suggests that during this period, there was a reliance on local goods rather than foreign goods for the production of blow moulding products.

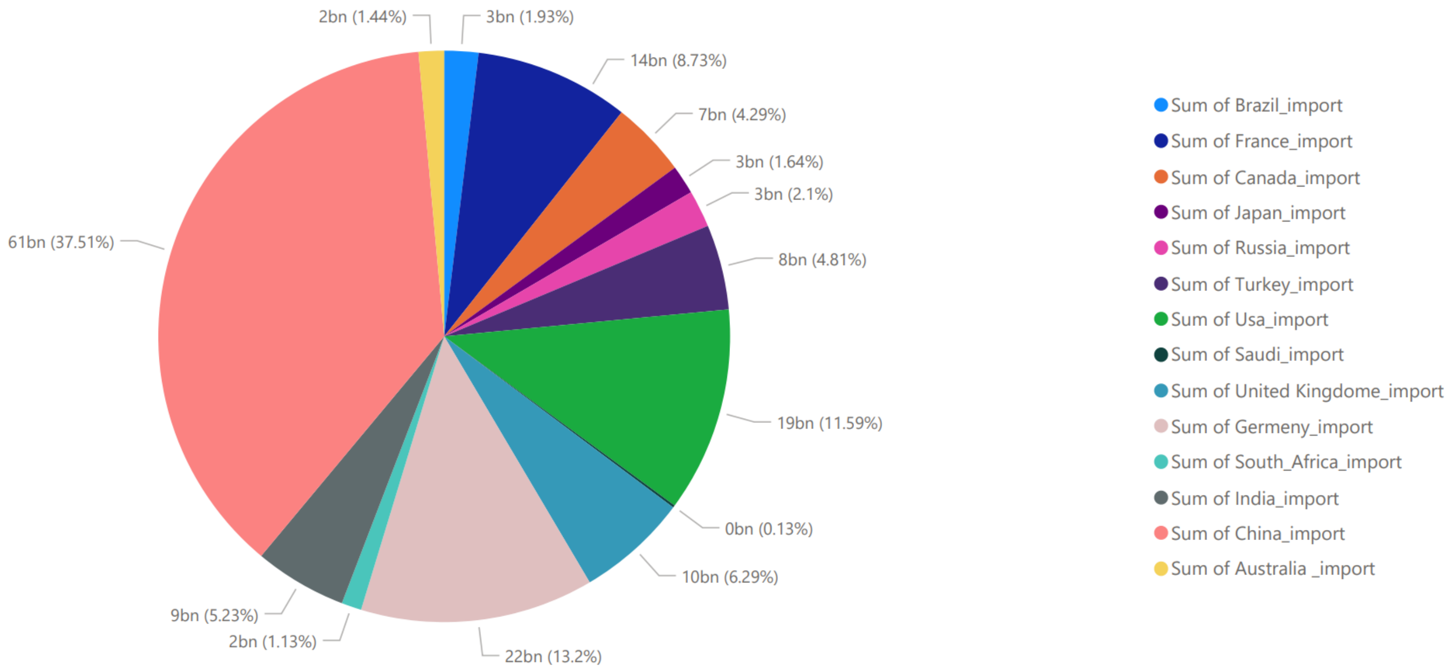
2. 2. In the case of the USA, there were consistent exports recorded from 2008 to 2022. This indicates a steady flow of exported goods from the USA throughout this time period.

3. 3. During the US recession, which occurred from January 2008 to May 2009, it was observed that only the USA had exports, while other countries did not report any exports. This highlights the impact of the recession on international trade, leading to a decline in exports from countries other than the USA.



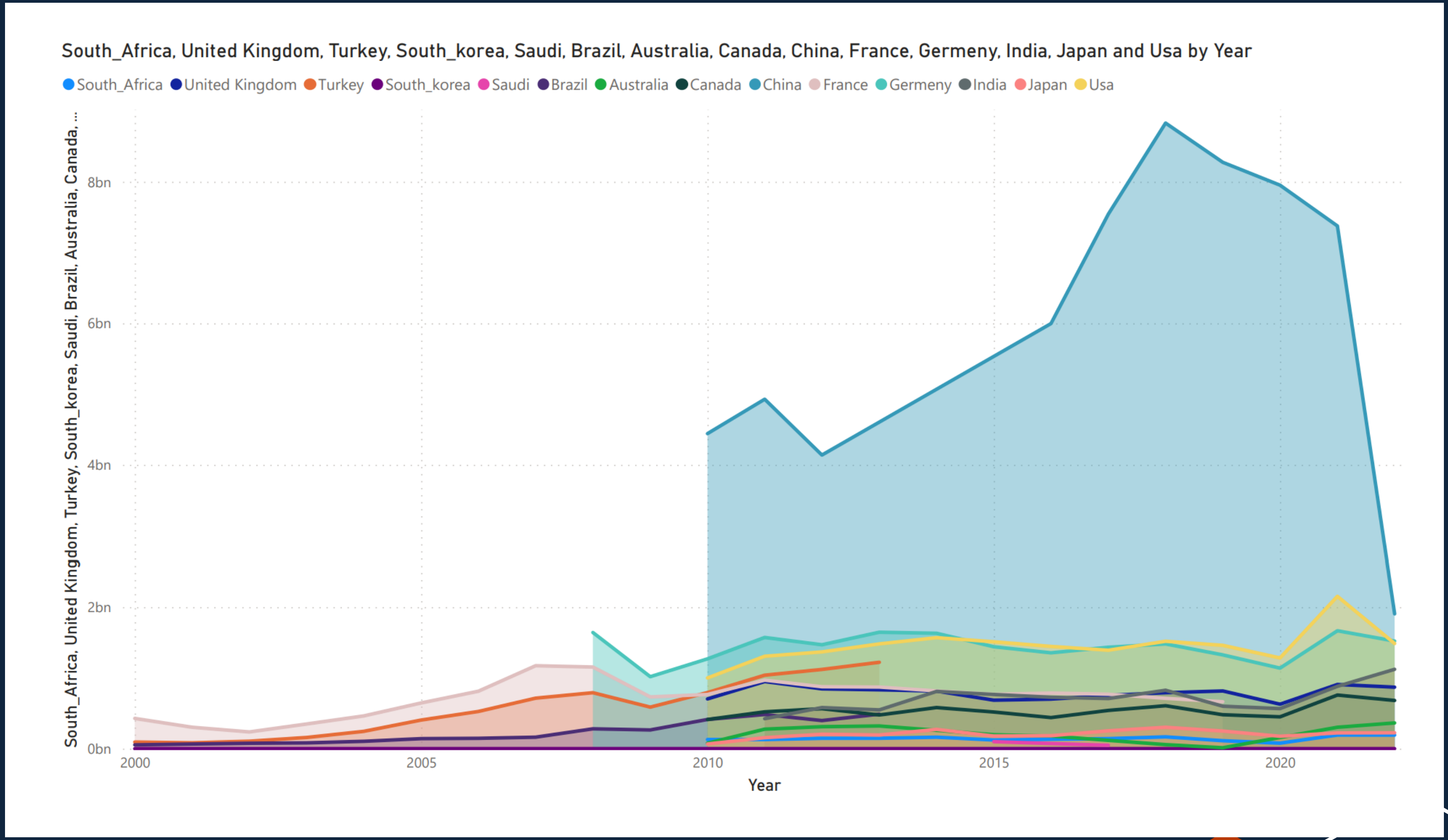
Percentage of countries Imports

Sum of Brazil_import, Sum of France_import, Sum of Canada_import, Sum of Japan_import, Sum of Russia_import, Sum of Turkey_import, Sum of Usa_import, Sum of Saudi_import, Sum of United Kingdome_import, Sum of Germeny_import, Sum of South_Africa_import, Sum of India_import, Sum of China_import and Sum of Australia _import



Insights:

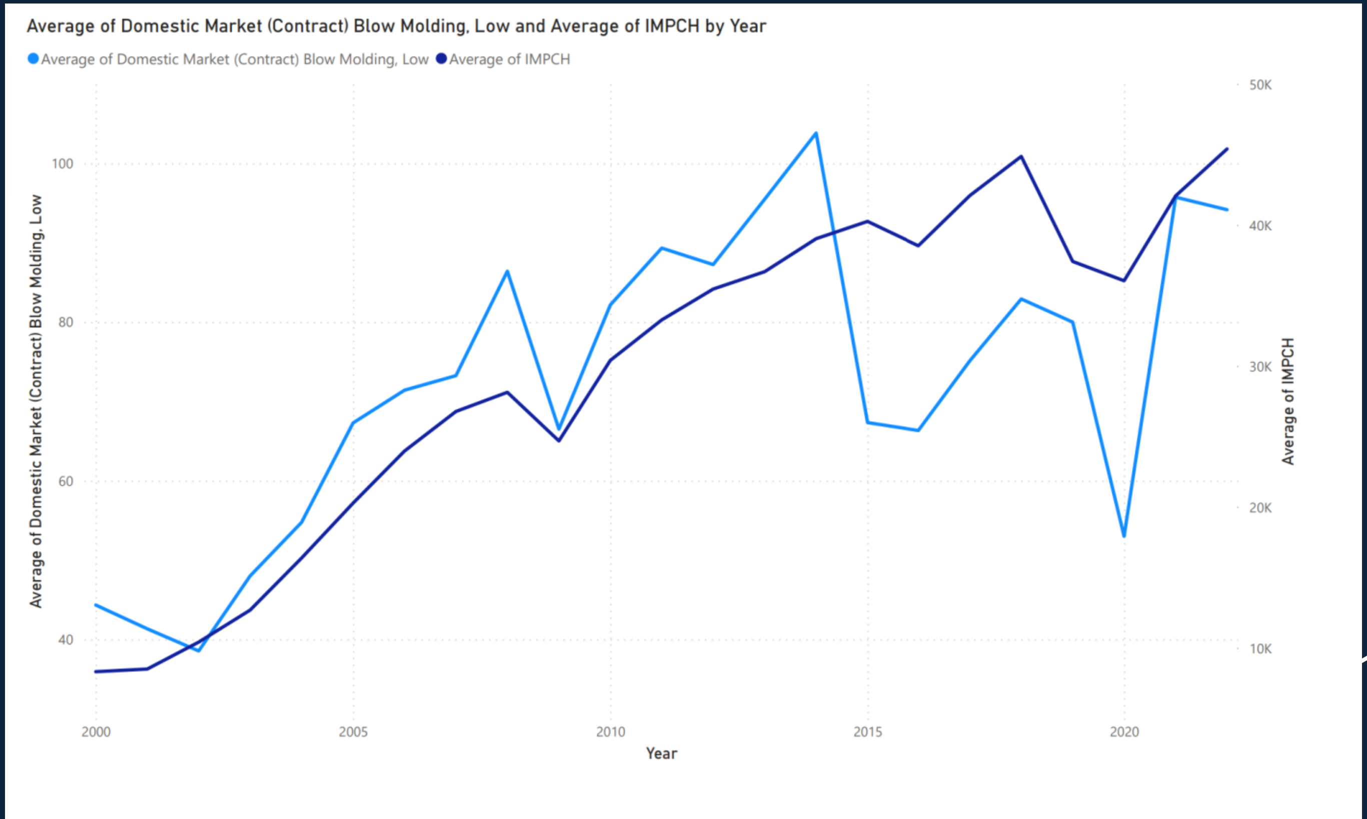
1. China is characterized by significant imports, whereas other countries have relatively low import levels.



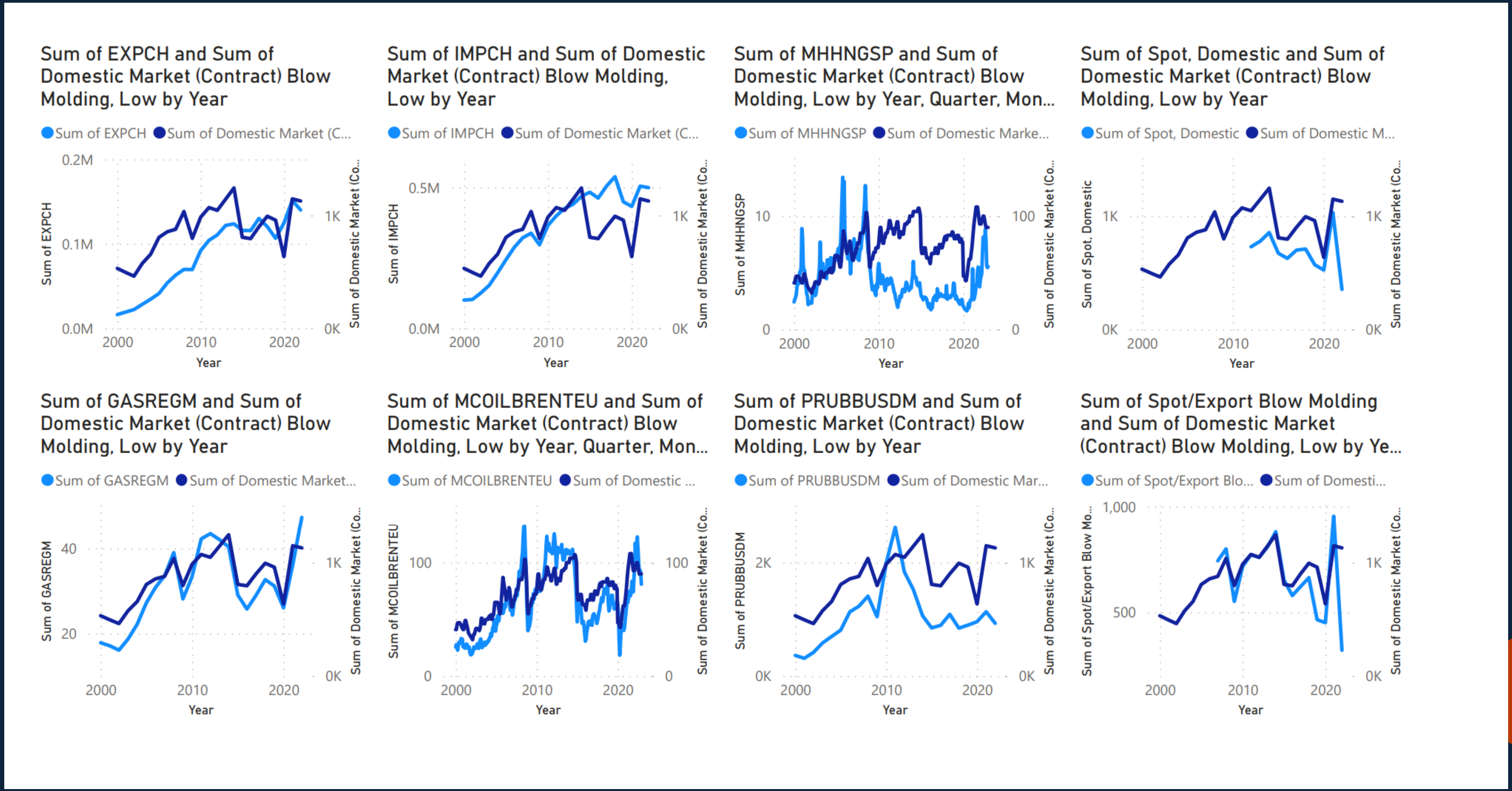
Blow moudling price and oil prices

- Upon examining the provided figure, a notable observation is that there is a similar pattern observed between moulding price and oil prices until 2014. Both prices appear to exhibit a similar trend and move in a somewhat synchronized manner during this period.

- However, a significant shift is observed starting from 2015. While there is an upward trend in oil prices during this period, there is a simultaneous downward trend in moulding prices. This indicates a divergence between the two prices, with oil prices continuing to rise while moulding prices experience a downward movement.



Sum of Oil/gas price and Moudling price on yearly



Sum of EXPCH and Sum of Domestic Market (Contract) Blow Molding, Low by Year

Sum of IMPCH and Sum of Domestic Market (Contract) Blow Molding, Low by Year

Sum of MHHNGSP and Sum of Domestic Market (Contract) Blow Molding, Low by Year, Quarter, Mon...

Sum of Spot, Domestic and Sum of Domestic Market (Contract) Blow Molding, Low by Year

Sum of GASREGM and Sum of Domestic Market (Contract) Blow Molding, Low by Year

● Sum of GASREGM ● Sum of Domestic Market...
40
20
Sum of GASREGM
1K
0K
Sum of Domestic Market (Co...
2000
2010
2020
Year

Sum of MCOILBRENTEU and Sum of Domestic Market (Contract) Blow Molding, Low by Year, Quarter, Mon...

● Sum of MCOILBRENTEU ● Sum of Domestic ...
100
0
Sum of MCOILBRENTEU
100
0
Sum of Domestic Market (Co...
2000
2010
2020
Year

Sum of PRUBBUSDM and Sum of Domestic Market (Contract) Blow Molding, Low by Year

● Sum of PRUBBUSDM ● Sum of Domestic Mar...
2K
0K
Sum of PRUBBUSDM
1K
0K
Sum of Domestic Market (Co...
2000
2010
2020
Year

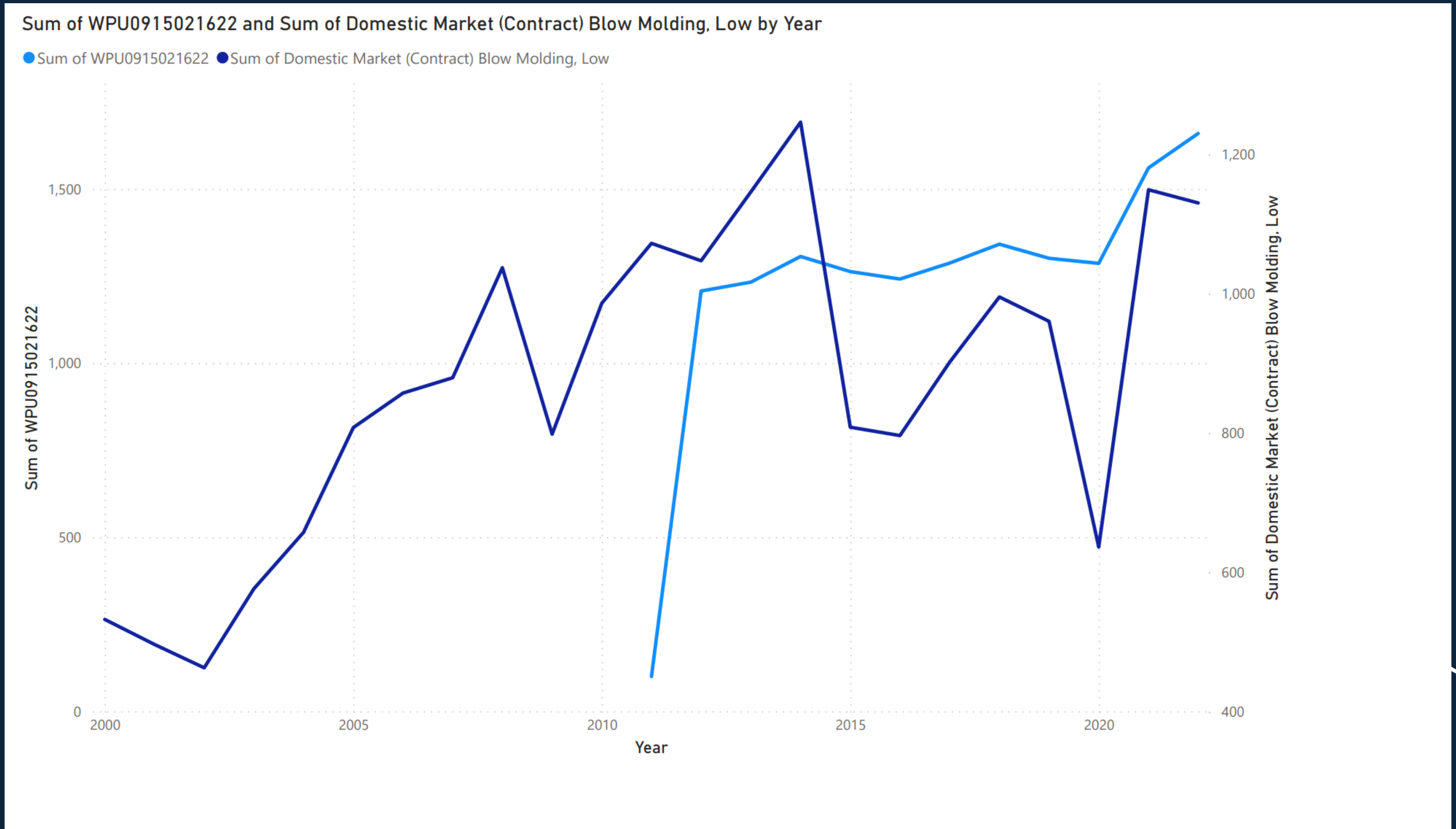
Sum of Spot/Export Blow Molding and Sum of Domestic Market (Contract) Blow Molding, Low by Ye...

● Sum of Spot/Export Blo... ● Sum of Domesti...
1,000
500
Sum of Spot/Export Blow Mo...
1K
0K
Sum of Domestic Market (Co...
2000
2010
2020
Year

Plastic/resins price and moudling price

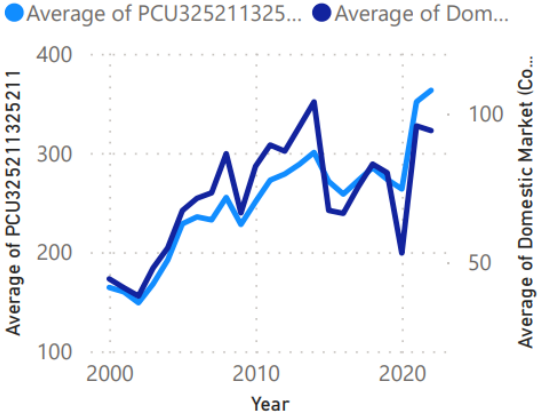
- During the period from 2000 to 2010, there was no production recorded for plastic/resins. This indicates a complete absence of data or activity related to plastic/resins production during that time frame.

- Furthermore, it was observed that there is no discernible pattern or correlation between the two prices being analyzed. This suggests that there is no meaningful relationship or association between the prices, indicating a lack of correlation between them.

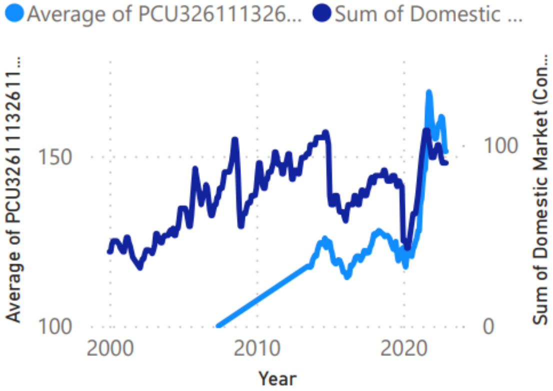


PLASTICS Price VS Domestic Market (Contract) Blow Molding, Low

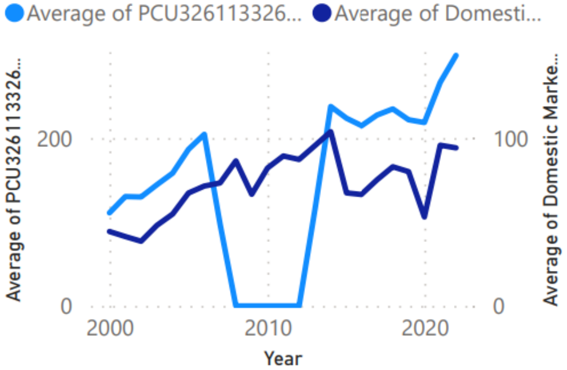
Average of PCU325211325211 and Average of Domestic Market (Contract) Blow Molding, Low by Year



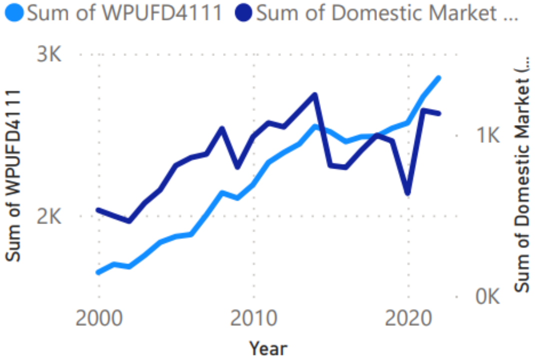
Average of PCU32611132611112 and Sum of Domestic Market (Contract) Blow Molding, Low by Year, Quarter, M...



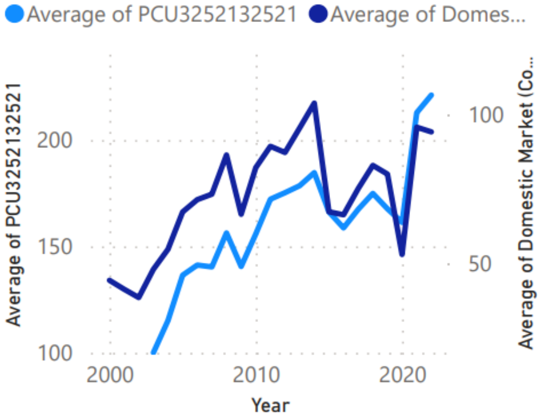
Average of PCU32611332611301 and Average of Domestic Market (Contract) Blow Molding, Low by Year



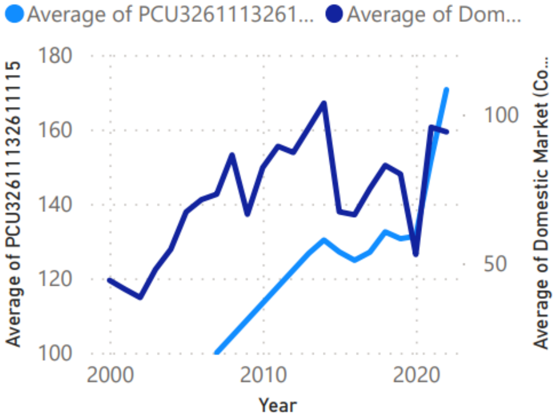
Sum of WPUFD4111 and Sum of Domestic Market (Contract) Blow Molding, Low by Year



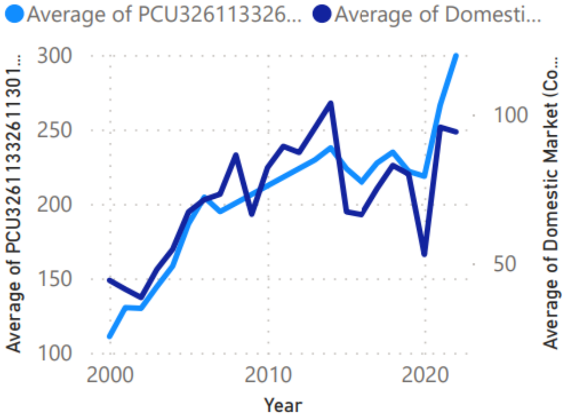
Average of PCU3252132521 and Average of Domestic Market (Contract) Blow Molding, Low by Year



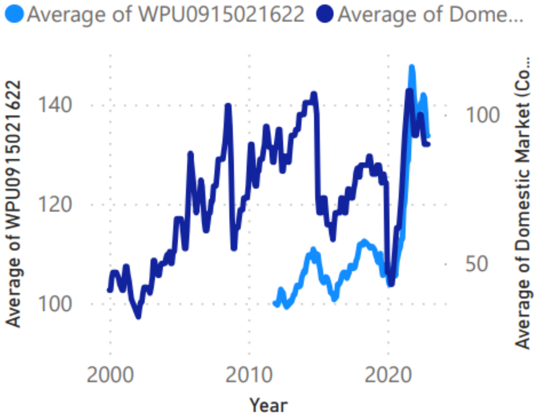
Average of PCU32611132611115 and Average of Domestic Market (Contract) Blow Molding, Low by Year



Average of PCU32611332611301.1 and Average of Domestic Market (Contract) Blow Molding, Low by Year



Average of WPU0915021622 and Average of Domestic Market (Contract) Blow Molding, Low by Year, Quarter, ...

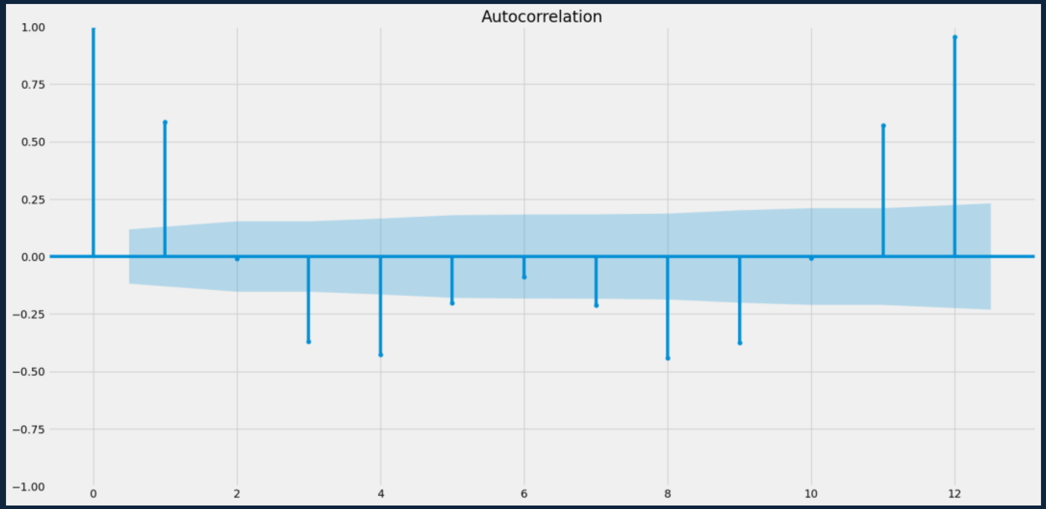


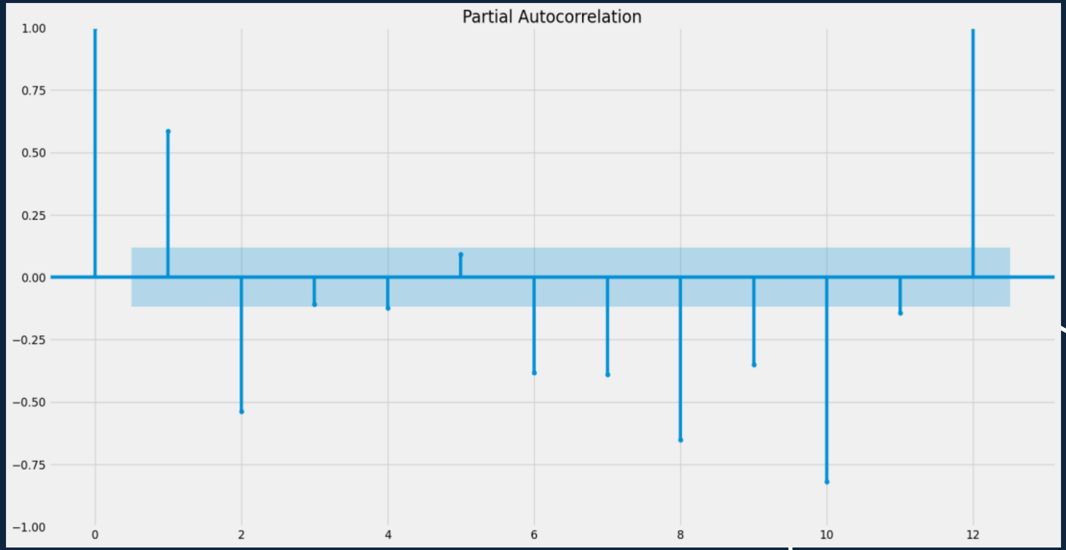
ACF-PACF Plots

- Based on the analysis of the partial autocorrelation function (PACF) values, it is observed that significant values persist up to lag 12. This indicates that an autoregressive order of p = 2 (ARIMA(p,0,0)) is an appropriate choice for the model.

- Similarly, by examining the autocorrelation function (ACF), it is determined that the significant values persist up to lag 2. Therefore, a moving average order of q = 1 (ARIMA(0,0,q)) is deemed suitable for the model.

- Considering these findings, the recommended ARIMA model configuration is ARIMA(2,0,1), with the inclusion of one differencing. This choice takes into account both the PACF and ACF analyses and aims to capture the autocorrelation and moving average components of the time series effectively.





Pre-processing

Missing value imputation:

1. In the dataset, there were several columns with null percentages exceeding 80%, making it challenging to handle such columns effectively. Consequently, the decision was made to remove these columns from the dataset, resulting in the deletion of 6 columns.

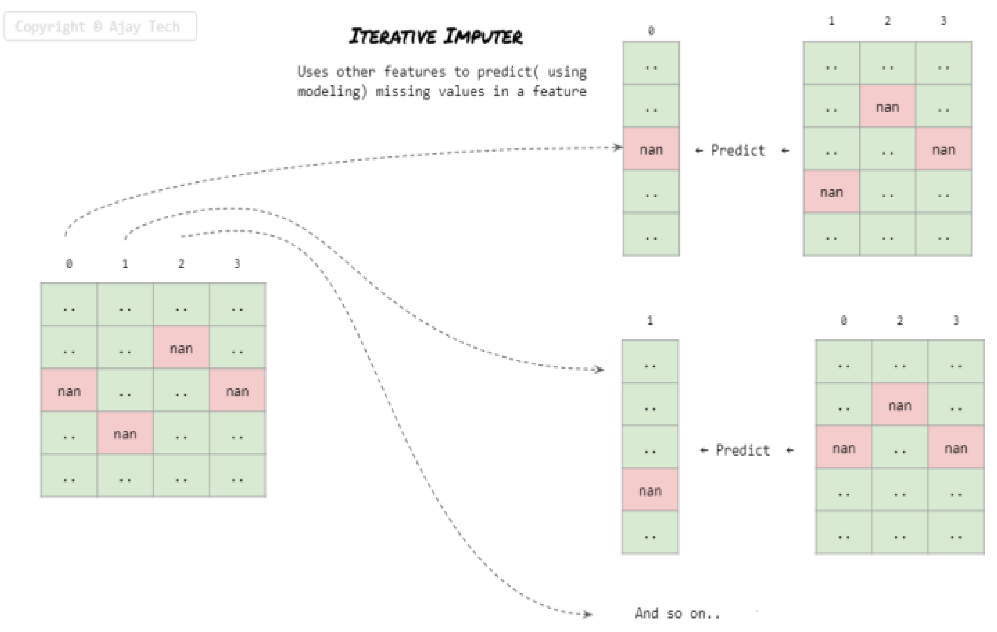
2. To address null values in the remaining columns, various techniques were attempted. For certain columns such as "oil/gas" and "plastic/resins," it was observed that their values consistently started with 100. Hence, null values in these columns were imputed with the value 100.

3. Additionally, different imputation methods were applied to handle null values, including backward fill, forward fill, KNN imputer, iterative imputer, and linear interpolation. After evaluating the performance of these techniques, it was determined that the iterative imputer yielded the most satisfactory results in terms of imputing missing values effectively.

Iterative Imputer

- Iterative Imputer is a machine learning technique used for imputing missing values in a dataset. It works by filling in the missing values with predicted values that are generated by iterating over the features of the dataset.

- The iterative imputer algorithm uses a regression model to predict the missing values for each feature, and then repeats this process until the missing values converge to a stable solution. The algorithm is based on the idea that the missing values in a dataset are not completely random, but rather are correlated with other features in the dataset.



Uses other features to predict(using modeling) missing values in a feature

0	1	2	3
..	..	..	..
..	..	nan	..
nan	..	..	nan
..	nan	..	..
..	..	..	..

0
..
..
nan
..
..

1	2	3
..	..	..
..	nan	..
..	..	nan
nan	..	..
..	..	..

1
..
..
..
nan
..

0	2	3
..	..	..
..	nan	..
nan	..	nan
..	..	..
..	..	..

Feature creation

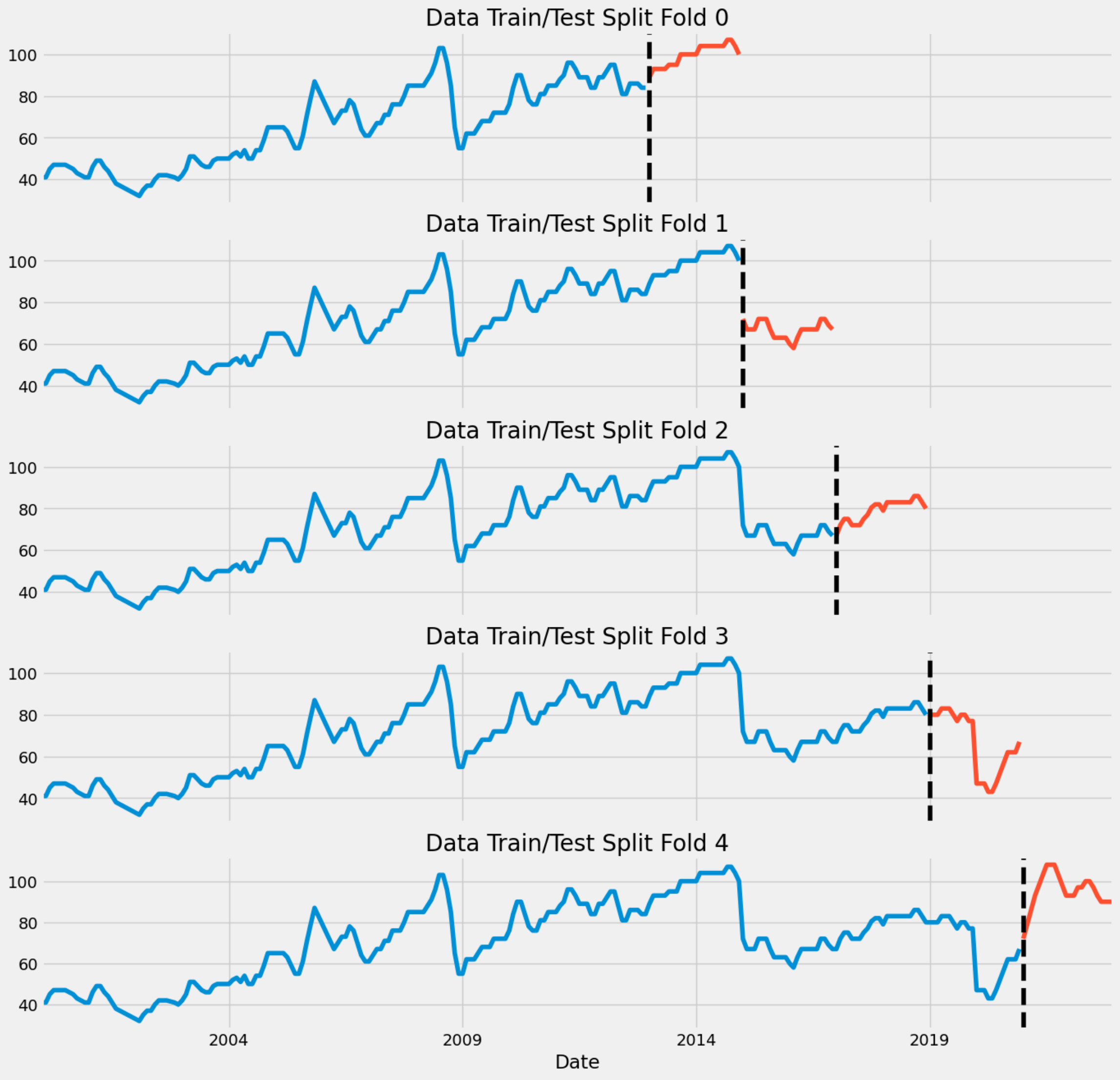
- From the given dataset, additional features were extracted to capture specific temporal information. This involved extracting the day, month, time, and week of the year from the data. These features provide more granular information about the timing and structure of the data, allowing the model to potentially capture seasonality and time-based patterns.

- Furthermore, lag features were created using the shift function. Lag features help incorporate information from previous time steps into the current observation. In this case, lag features with shift(1), shift(2), and shift(3) were generated, which means that the values from the previous three time steps were used to create new features. This enables the model to consider the historical behavior and trends of the data, potentially capturing dependencies and autocorrelation between time steps.

Time series split for CV

- The dataset was split into train and test sets, with 80% of the data allocated for training and 20% for testing.

- For time series cross-validation, 5 folds were used, with each fold having a test size of 56 data points and no gap between consecutive folds.



Model Building

Approach-1:

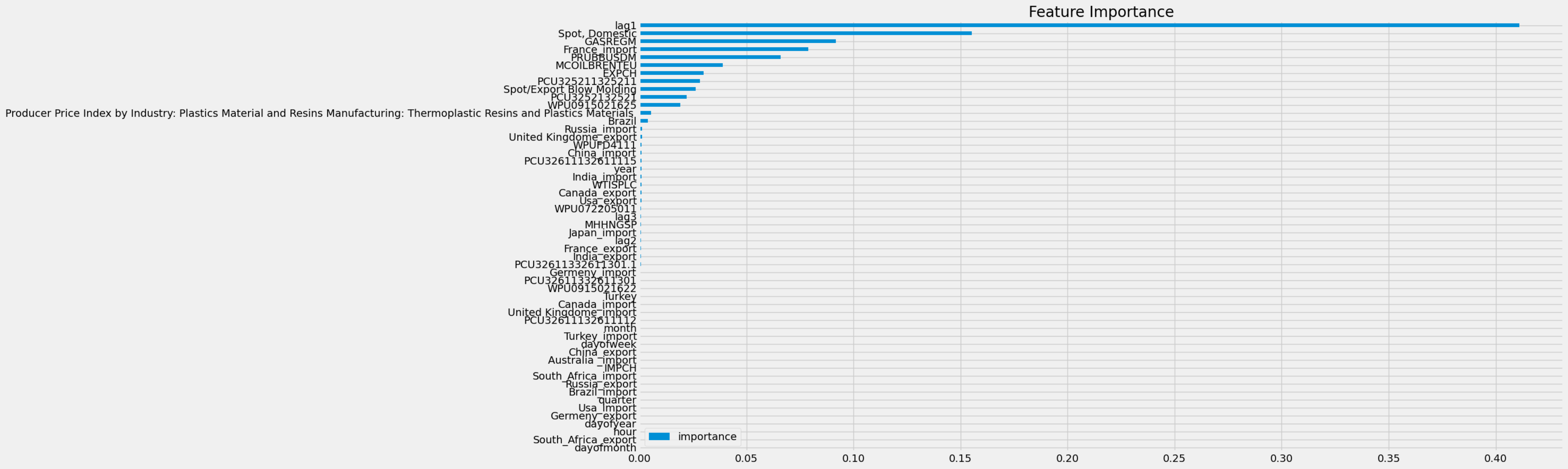
Univariate Analysis:

- For model building, the date and target variables were used. Various smoothing techniques such as rolling mean, exponential smoothing, and Facebook Prophet were employed to create models.

Bivariate Analysis:

- In addition to the date and target variables, exogenous columns and lag columns were included for model building. Several models were utilized, including Auto-ARIMA, SARIMA, LSTM, Lazy Predict AUTOML, and XGBoost.

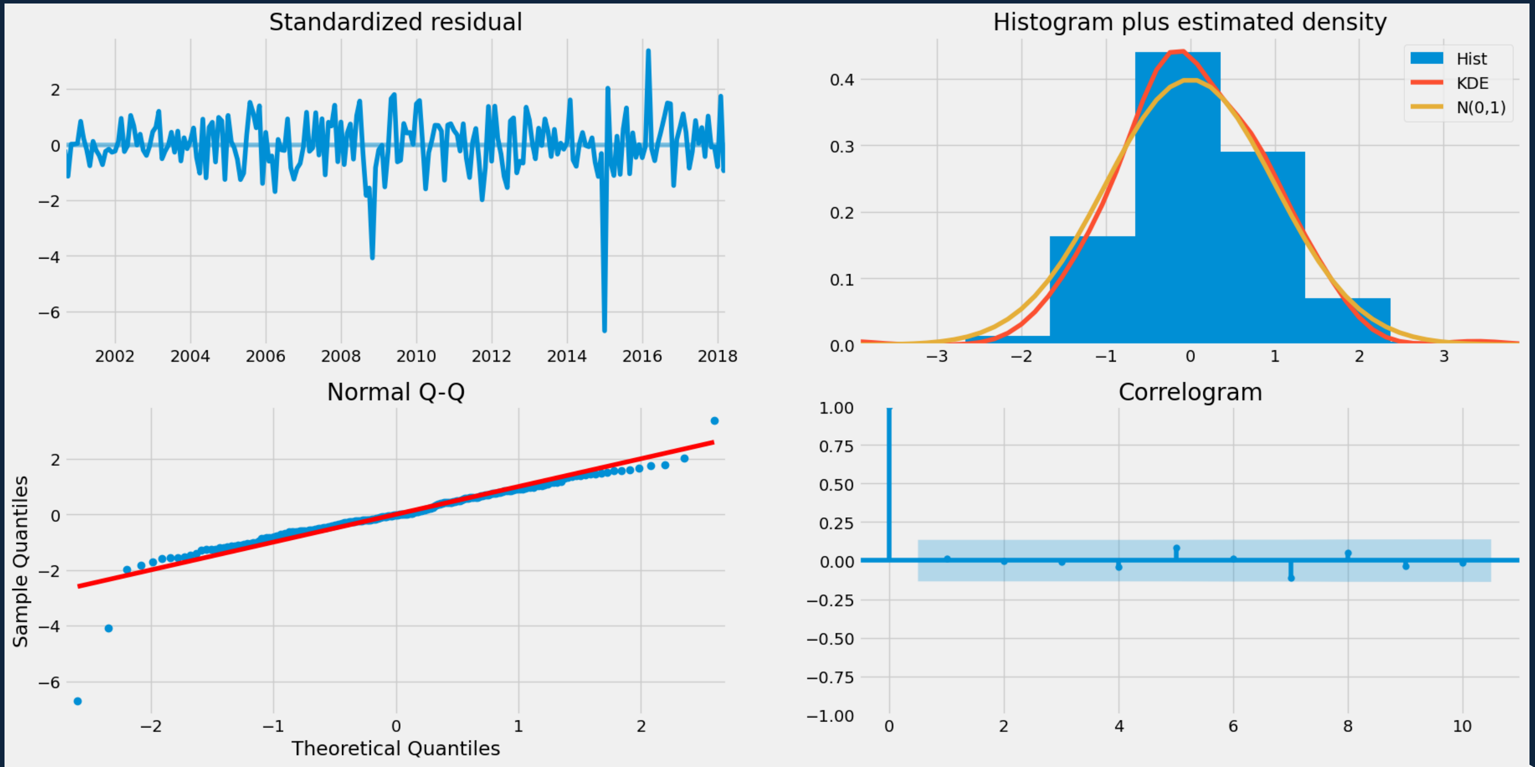
Feature Importance



Insights: Upon examining a dataset with 50 columns, it was observed that only a few columns exhibited a significant influence on the target variable. Specifically, the column labeled "Lag1" displayed a strong correlation with the target variable.

Diagnostic Plots

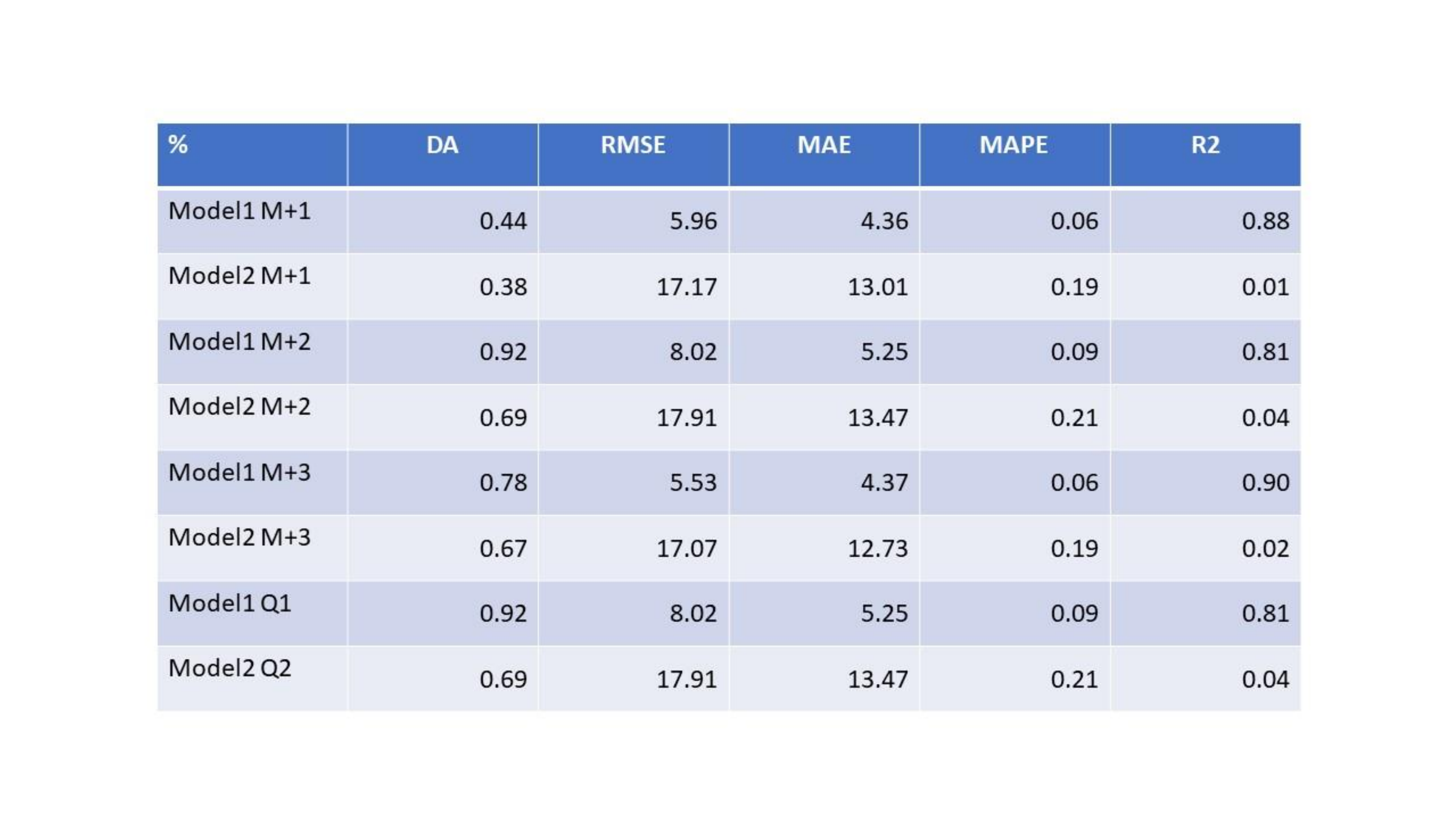
Diagnostic plots were generated using exponential smoothing. Upon analysis, it was discovered that outliers were present in the years 2009 and 2010. Furthermore, the errors were found to follow a normal distribution.



Results

Here

- Model1- xgboost
- Model2- fb prophet
- M-month
- Q- quarter



%	DA	RMSE	MAE	MAPE	R2
Model1 M+1	0.44	5.96	4.36	0.06	0.88
Model2 M+1	0.38	17.17	13.01	0.19	0.01
Model1 M+2	0.92	8.02	5.25	0.09	0.81
Model2 M+2	0.69	17.91	13.47	0.21	0.04
Model1 M+3	0.78	5.53	4.37	0.06	0.90
Model2 M+3	0.67	17.07	12.73	0.19	0.02
Model1 Q1	0.92	8.02	5.25	0.09	0.81
Model2 Q2	0.69	17.91	13.47	0.21	0.04

Areas of focus

Handling missing values:

The statement highlights the presence of missing values in the dataset and mentions the use of a missing iterative imputer to address this issue. Imputing missing values is crucial for ensuring complete and reliable data for modeling

Time series cross-validation:

Time series cross-validation is employed to evaluate the performance of the model. This type of validation is specifically designed for time-dependent data to assess how well the model generalizes to unseen future data.

Summary

- The given dataset contains missing values, which were imputed using a technique called missing iterative imputer.
- Additionally, feature creation was performed to enhance the dataset.
- To evaluate the performance of the model, time series cross-validation was used in conjunction with the XGBoost algorithm.
- The resulting root mean squared error (RMSE) for the XGBoost model was found to be 4, which indicates good predictive accuracy compared to other models.
- Furthermore, out of the total 50 features, only 8 were identified as having the most significant influence on the target variable.

Thank you

Reesu Jagan

reesujagan42@gmail.com

codejay12.github.io