Market Analysis and Forecasting Using Machine Learning

Predicting Commodity Quantities and Prices

30-May-2024

Project by Konstantinos Soufleros

Introduction

Background:

Understanding patterns in time series data is crucial for market analysis and forecasting.Machine learning techniques allow for accurate predictions based on historical data.

Objective:

Develop a robust time series machine learning model to forecast market trends.Predict future quantities and prices of commodities.

Dataset Information

Dataset Overview:
Contains monthly market data spanning multiple years.Includes various regions, commodities, and pricing information.

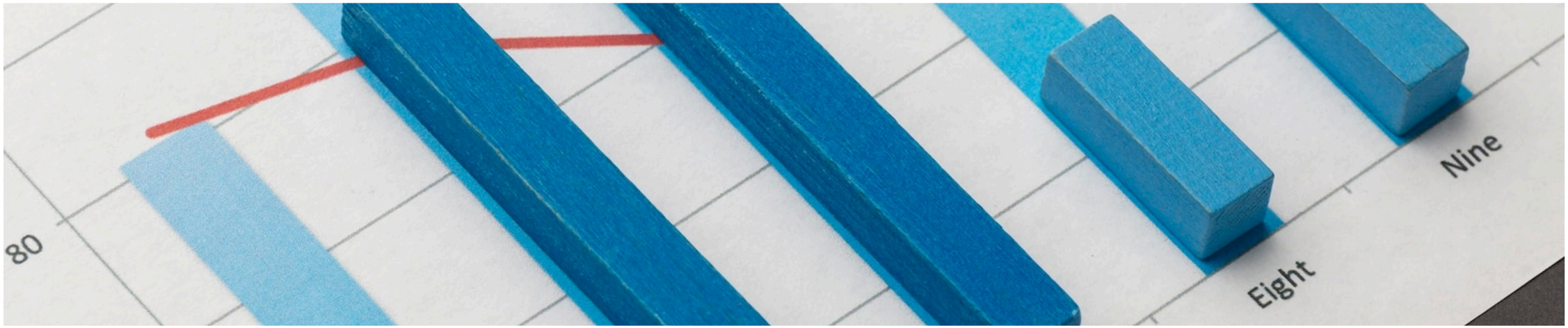
Key Columns:
market: Commodity or market.*month*: Month of the data.*year*: Year of the data.*quantity*: Quantity traded.*priceMin*: Minimum price.*priceMax*: Maximum price.*priceMod*: Mode price.*state*: State or region.*city*: City of the market.*date*: Specific date.

Research Question

"How can we develop a time series forecasting model that accurately predicts the future quantity and prices of commodities based on historical market data?"

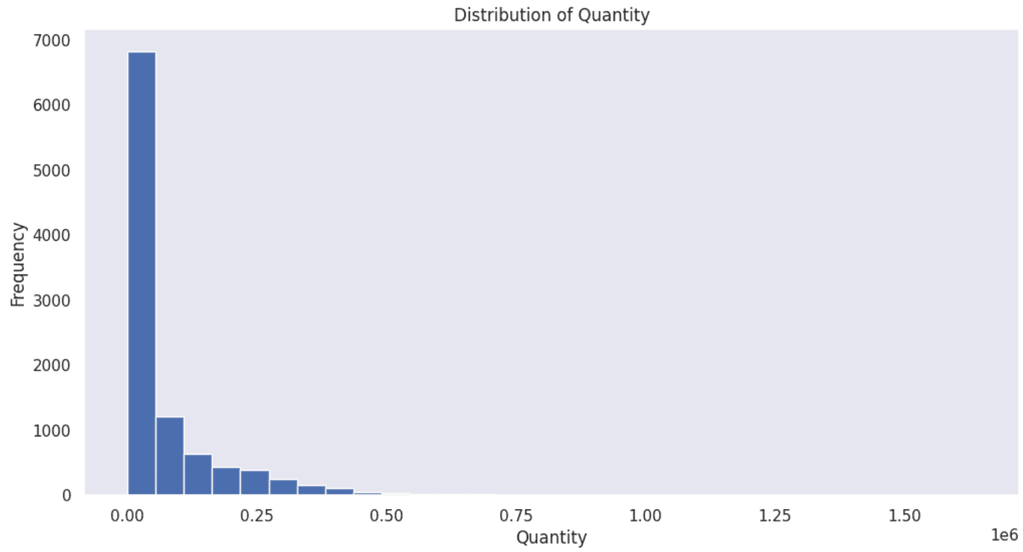
Approach of Analysis

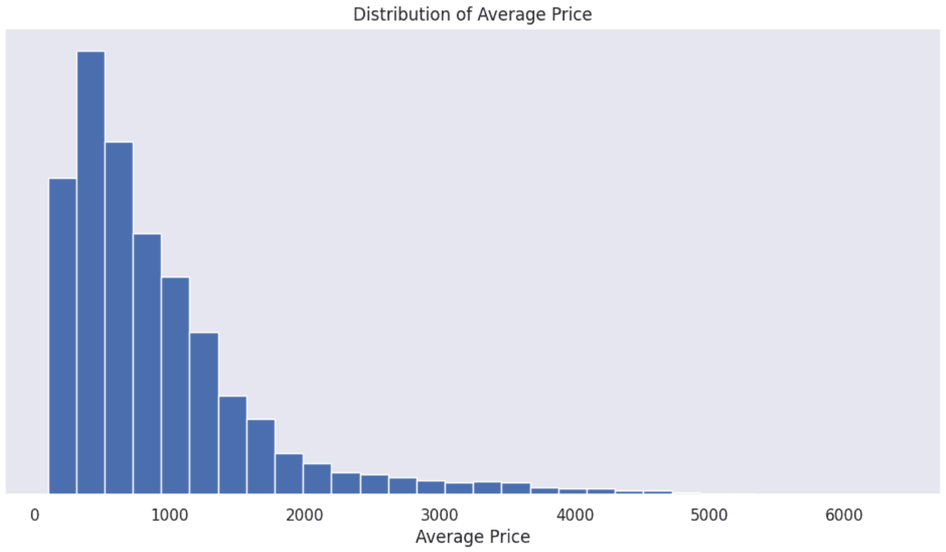
Data Preprocessing: Clean dataset. Handle missing values. Encode categorical variables. **Exploratory Data Analysis (EDA)**: Analyze temporal patterns. Identify seasonality, trends, and anomalies. **Feature Engineering**: Create lagged variables. Calculate rolling statistics. Extract seasonal indicators. **Model Selection and Training**: Evaluate ARIMA, SARIMA, Prophet, LSTM models. Train the selected model. **Model Evaluation**: Assess performance using MAE, MSE, RMSE. **Fine-tuning and Validation**: Fine-tune parameters. Validate on unseen data.



Exploratory Data Analysis (EDA)

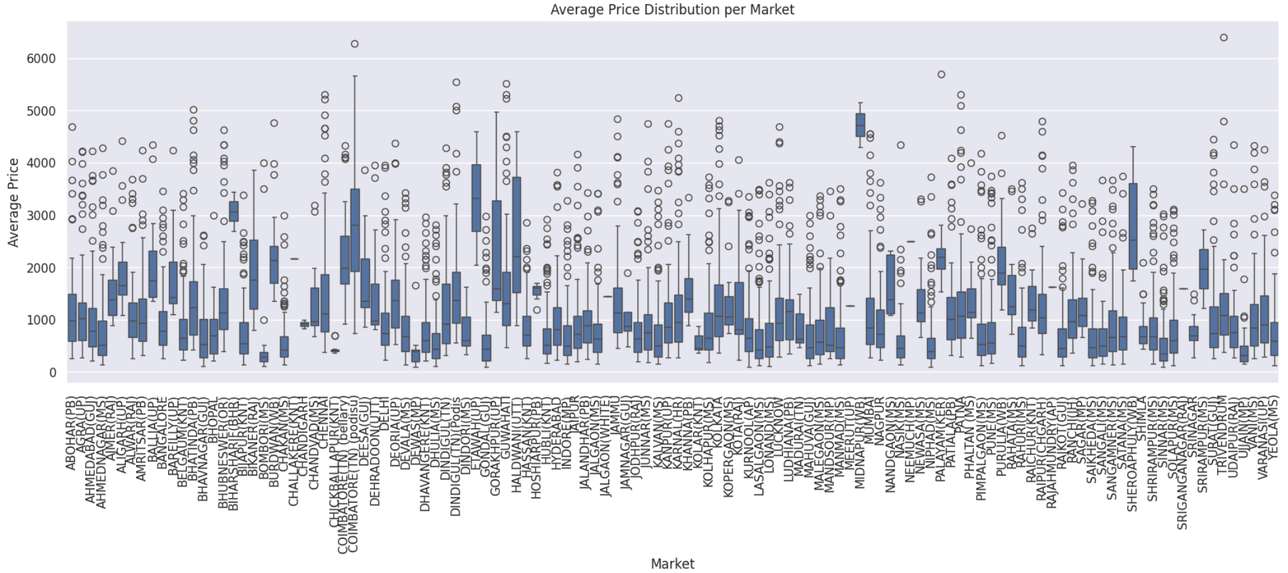
- The distribution of quantity is right skewed, indicating that there are a number of high quantity values that are less frequent but significantly larger than the rest.
- Average price is also right skewed, but with smaller tail indicating that the outliers are less here.
- This FacetGrid visualization helps to understand the pricing dynamics within each market

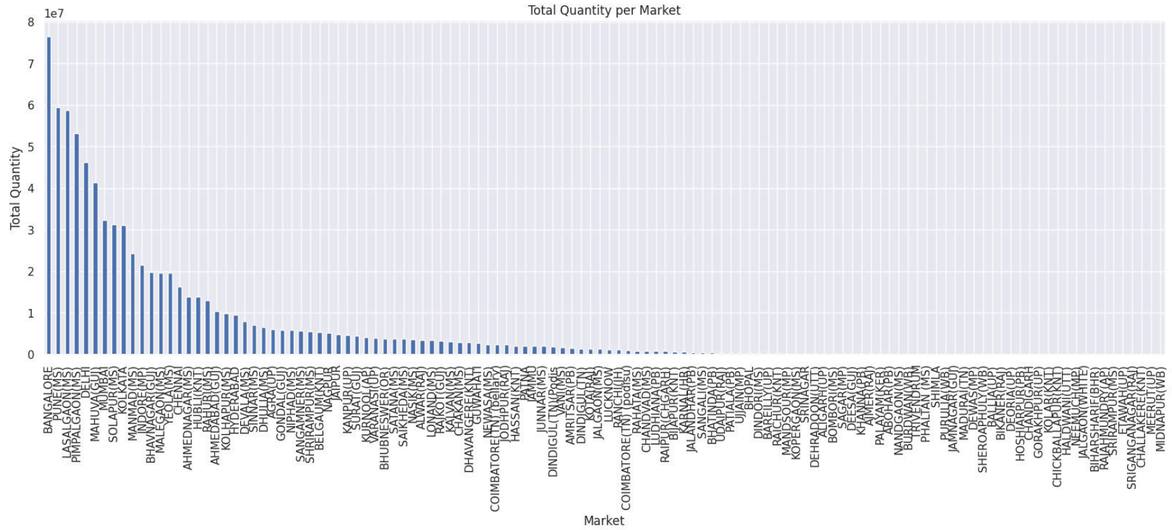


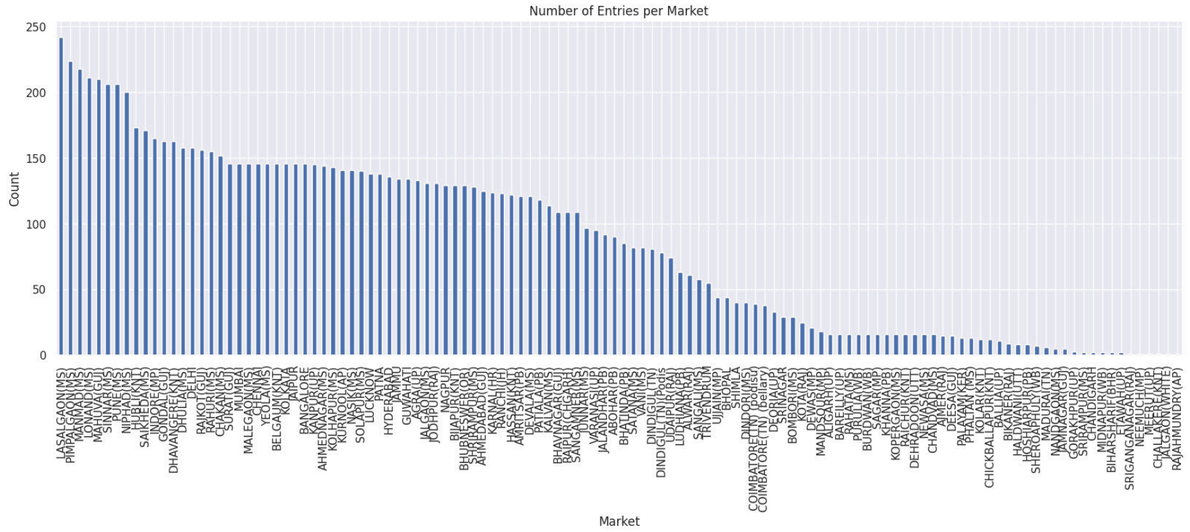


Exploratory Data Analysis (EDA)

- From the number of records per market above we observe that Lasalgaon, Pimpalgaon, Manmad are the top 3 markets were we have the most recorded observations.
- We observe the total quantities of commodities traded across different markets. Bangalore, Pune and Lasalgaon are the top 3 markets with the highest quantities traded.
- The boxplots here provide the variability and central tendency of prices within each market. We notice outliers in most of the markets.

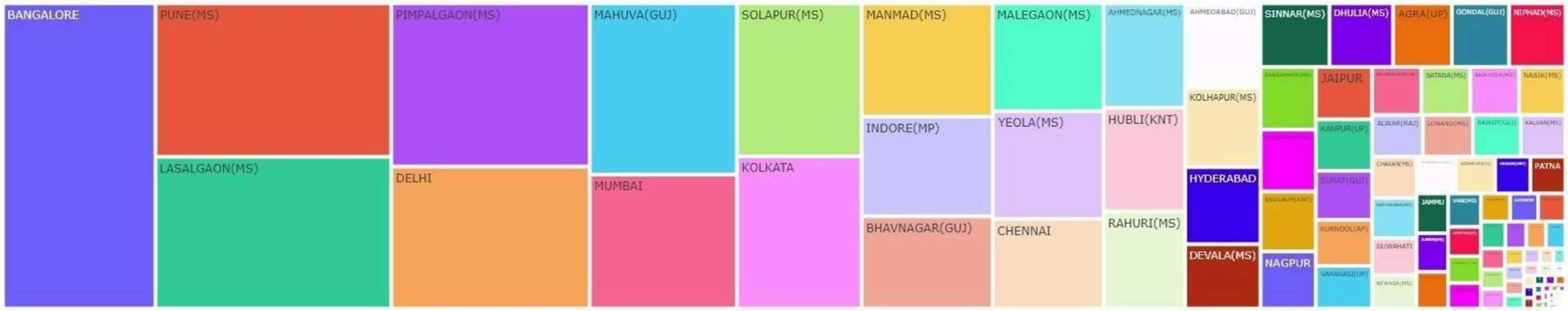




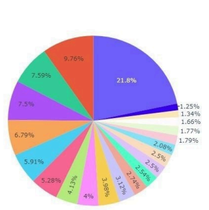


Exploratory Data Analysis (EDA)

Quantity Distribution Treemap



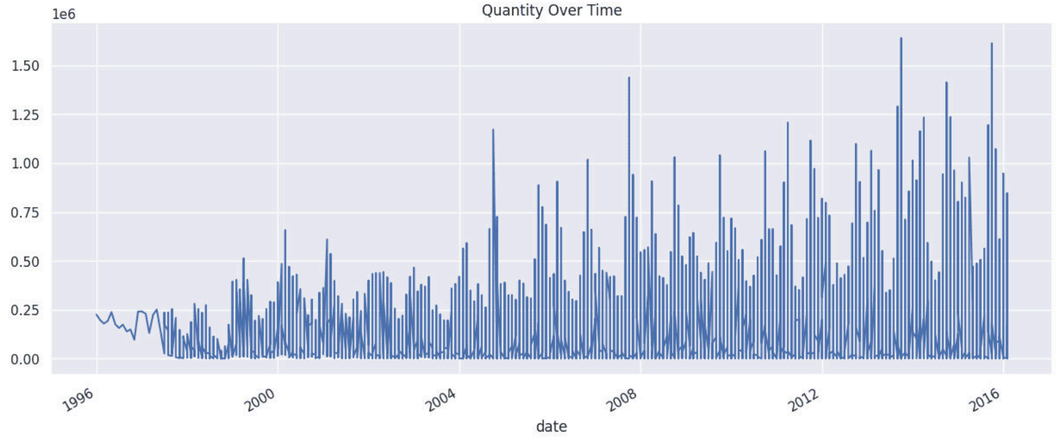
Average Price Distribution Treemap

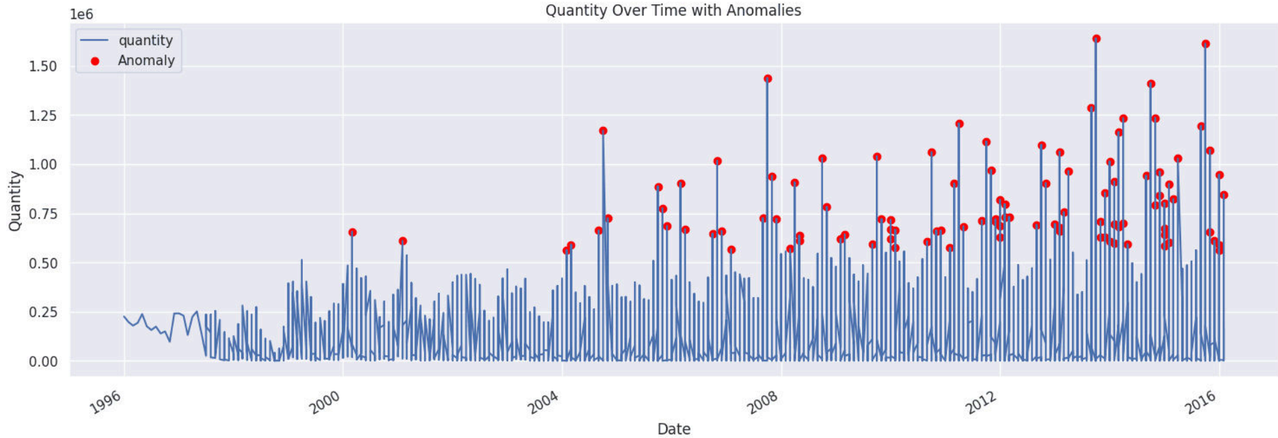


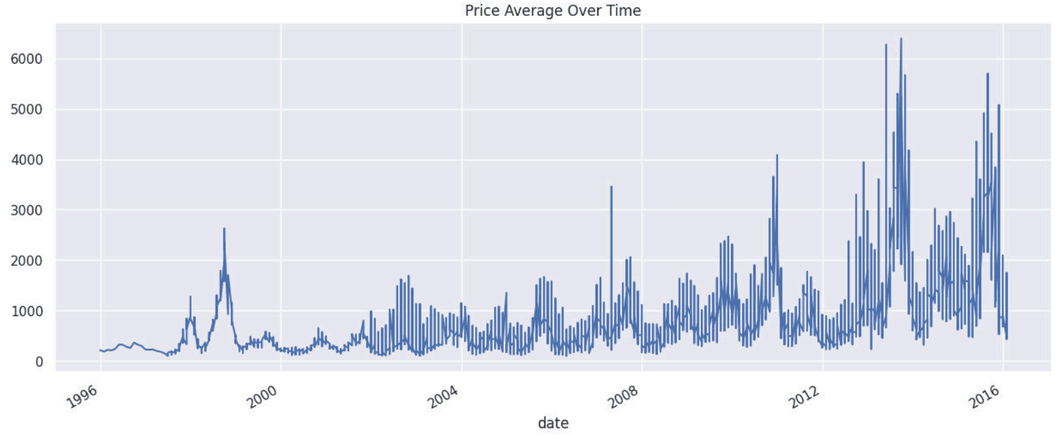
The pie chart visualizes the distribution of total quantities of commodities traded across major markets, with smaller markets grouped into an "Others" category. We can see this distribution more clearly in the next treemap, where each rectangle represents a market, and the size of the rectangle is proportional to the total quantity traded in that market. Also another treemap helps to quickly identify which markets have higher and lower average prices.

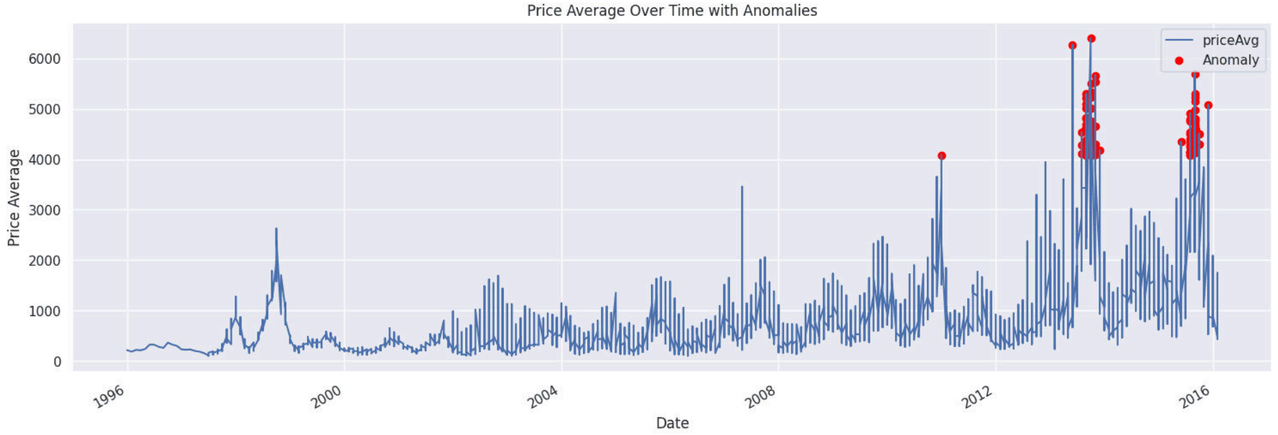
Exploratory Data Analysis (EDA)

- Plotted time series for Quantity and Average Price over Time
- Used Isolation Forest Algorithm to detect anomalies.
- Anomalies are detected as periodic spikes, indicating specific events or outliers that cause sudden increases in the quantity.
- In the average price, anomalies are consistently detected in certain periods, further investigation might be required to understand the underlying causes, such as data errors, market events etc.



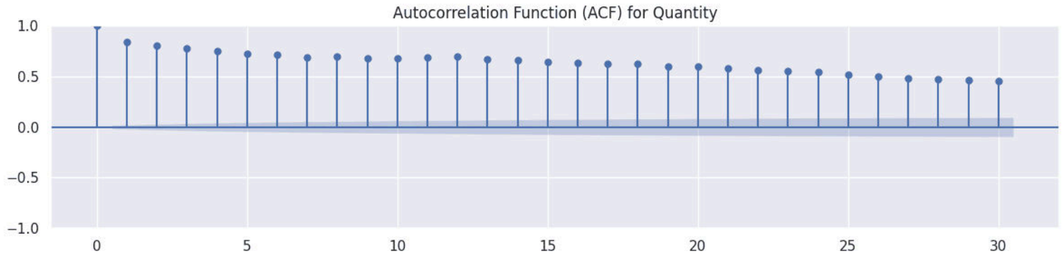


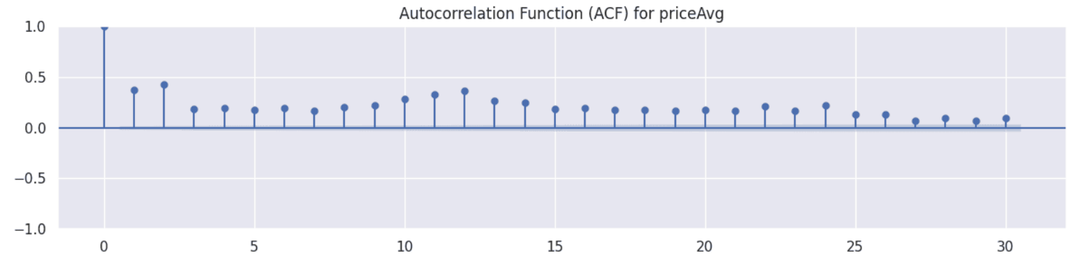


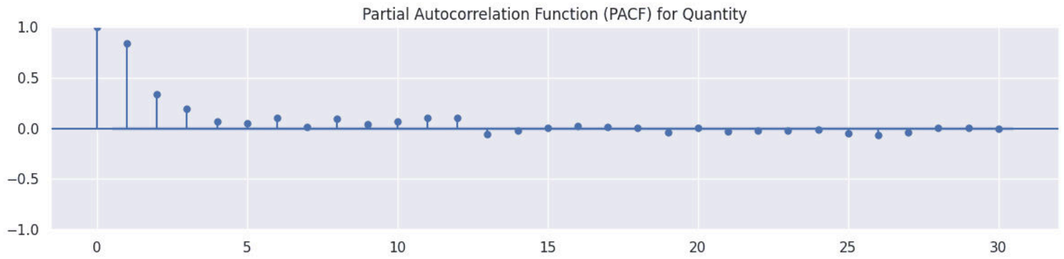


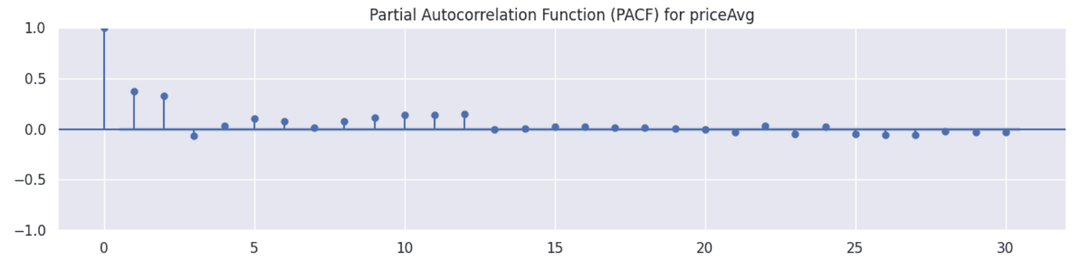
Exploratory Data Analysis (EDA)

- The Quantity ACF plot shows a slow decay, which suggests the presence of seasonality or a long-term trend in the data.The PACF plot shows significant spikes at the first few lags and then quickly drops off, it suggests that the data may follow an autoregressive process.
- The Average Price ACF plot shows a wave pattern with positive spikes, indicating a periodic or seasonal component. The PACF plot shows both positive and negative spikes at various lags. These alternating positive and negative spikes can indicate complex relationships between priceAvg values at different lags, possibly due to seasonality or other cyclical effects.

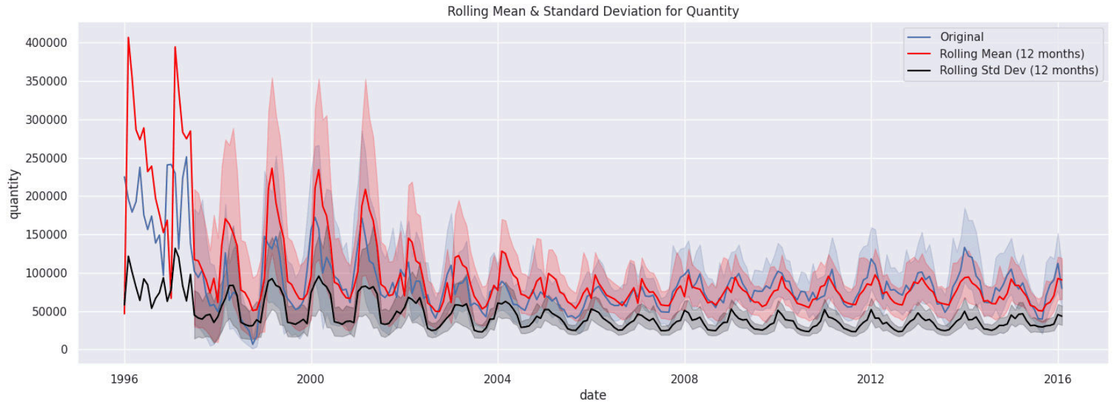




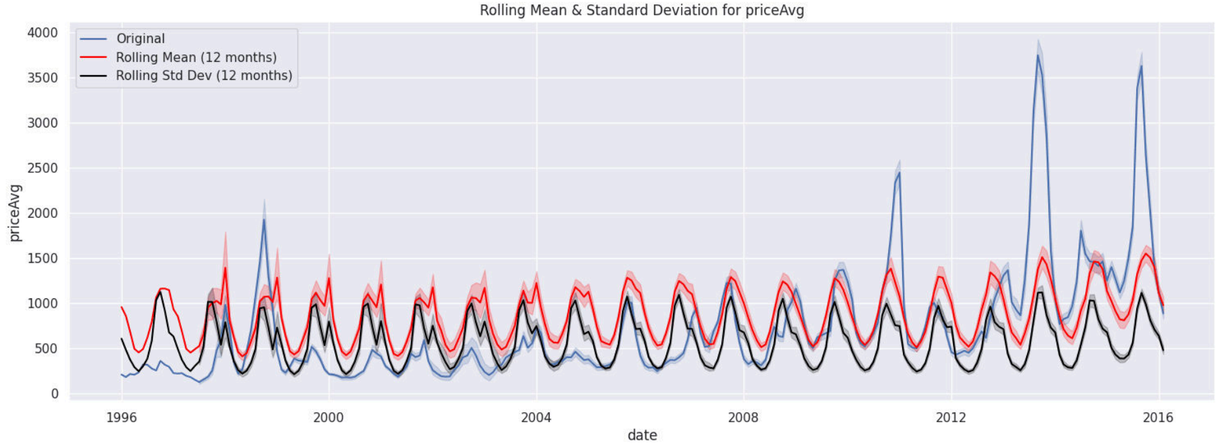




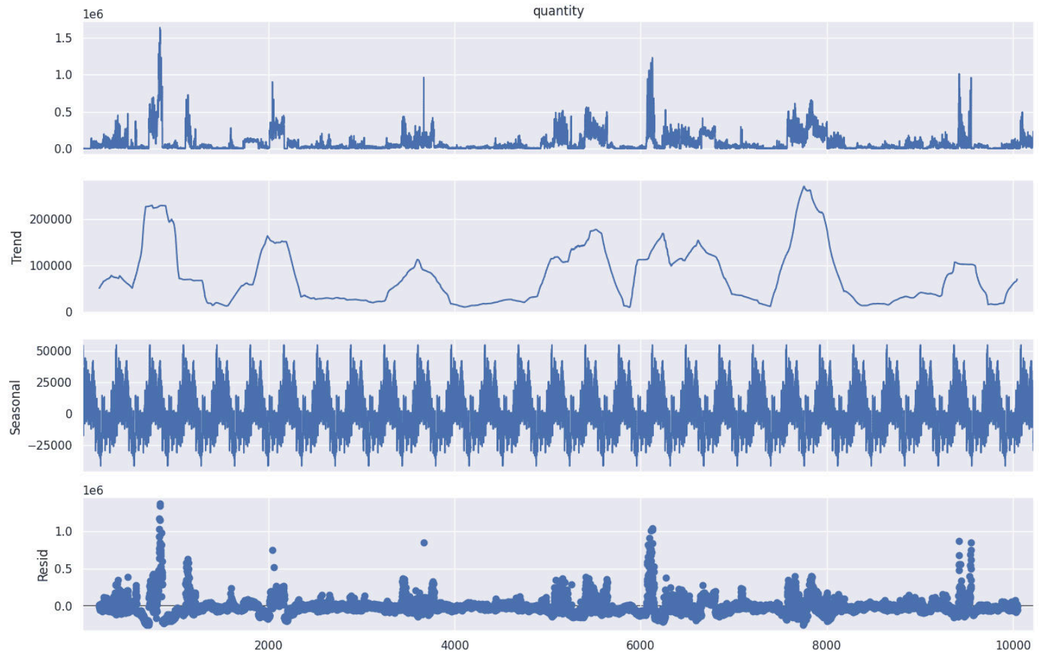
Feature Engineering



- The rolling mean (red line) in quantity shows an downward trend over time, which indicates that the average quantity is decreasing.The rolling standard deviation (black line) shows periodic spikes in the beggining, which indicates periods of high volatility in the quantity and later it smooths out.
- The slight upward trend in the rolling mean (red line) in AvgPrice indicates a consistent increase in average prices over time. This could be due to factors like inflation, increased demand, or changes in market conditions. The up and down spikes in the rolling standard deviation (black line) highlight periods of high and low volatility in prices. This could be due to seasonal effects, market disruptions, or other external factors affecting price stability.

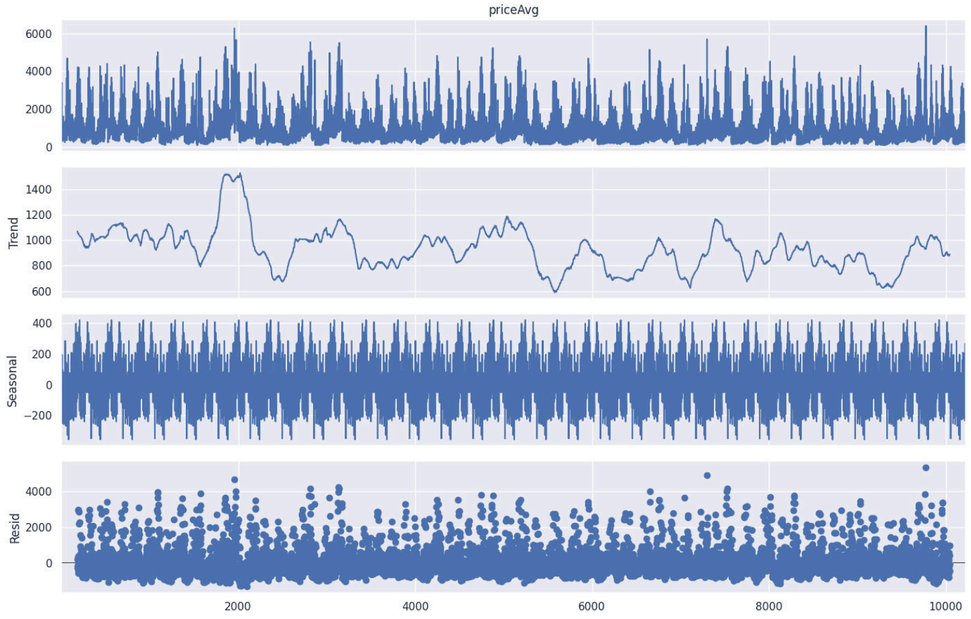


Feature Engineering



- Trend Analysis: The trend component shows an upward and downward trend, which indicates that the quantity of commodities is increasing and decreasing over time around a certain level.
- Seasonal Analysis: The seasonal component shows regular peaks and troughs, it indicates the presence of seasonality.
- Residual Analysis: The residual component helps to identify any anomalies or unusual patterns not captured by the trend or seasonal components.

- Trend Analysis: The trend component also here, shows an upward and downward trend, which indicates that the average price of commodities is increasing and decreasing over time around a certain level.
- Seasonal Analysis: The seasonal component shows regular peaks and troughs, it indicates the presence of seasonality.
- Residual Analysis: The residual component helps to identify any anomalies or unusual patterns not captured by the trend or seasonal components.

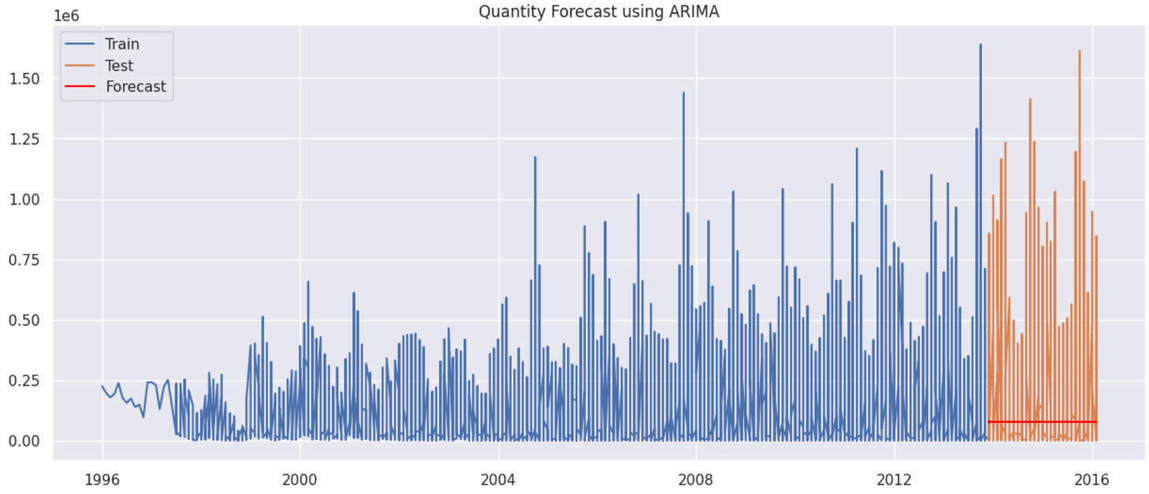


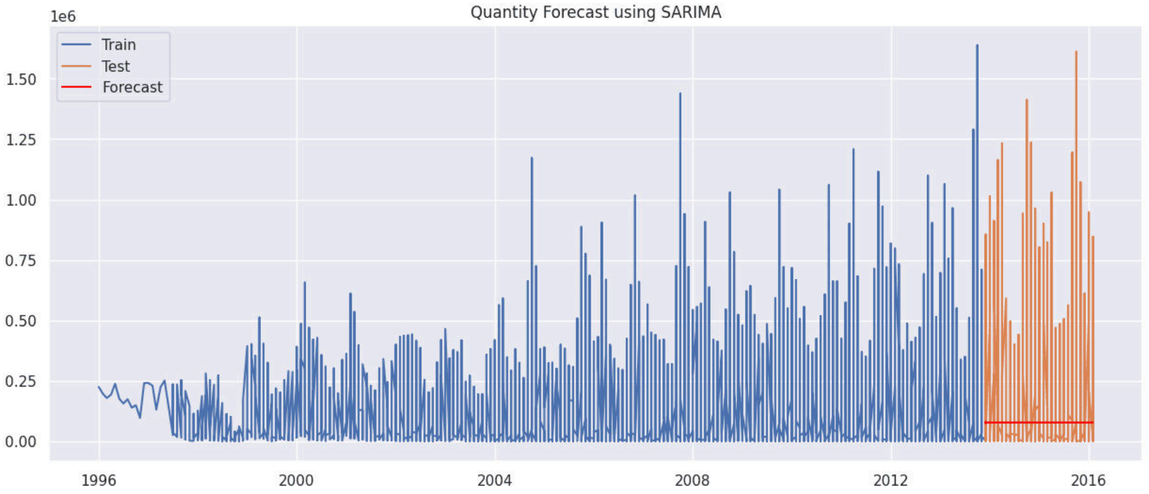
Model Selection and Training

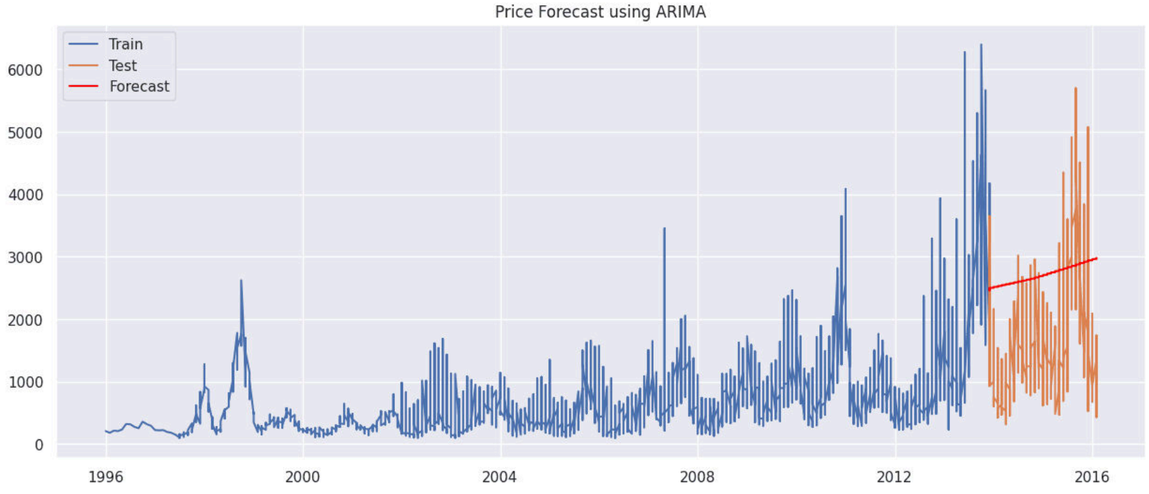
Evaluated Models:
ARIMA
SARIMA
Prophet
LSTM

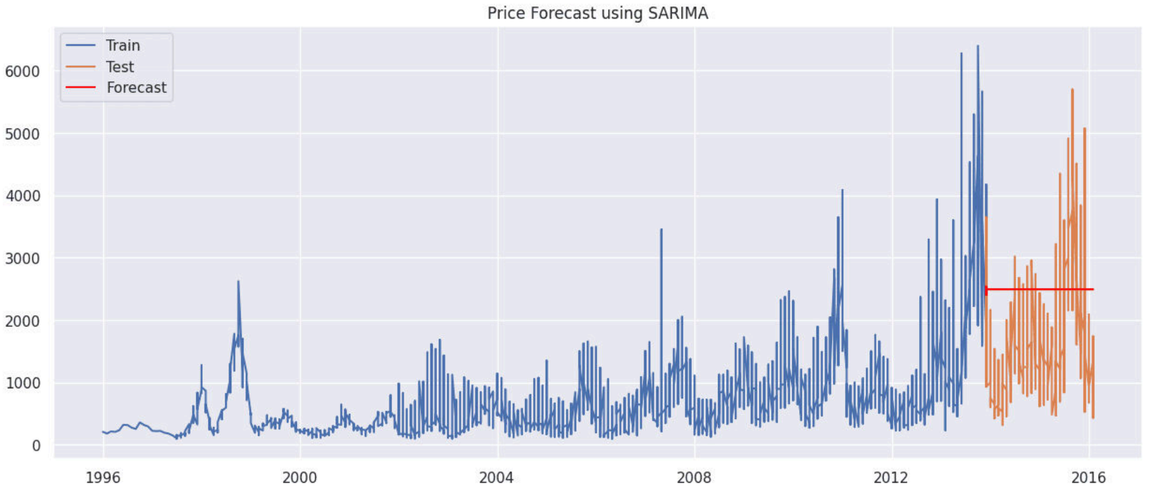
Training Process:
Split data into
training and testing sets.
Fit models and
tune parameters.

Model Selection and Training

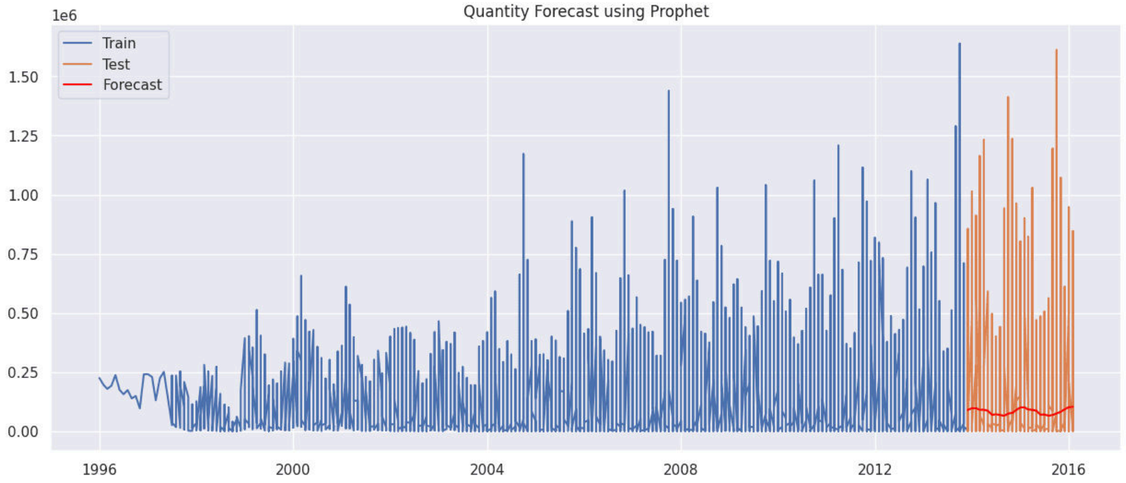


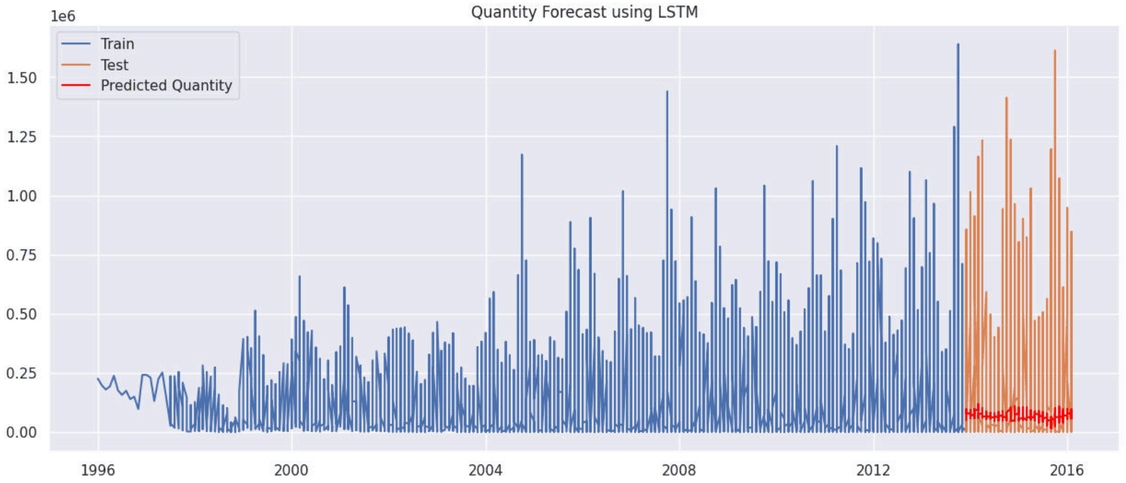


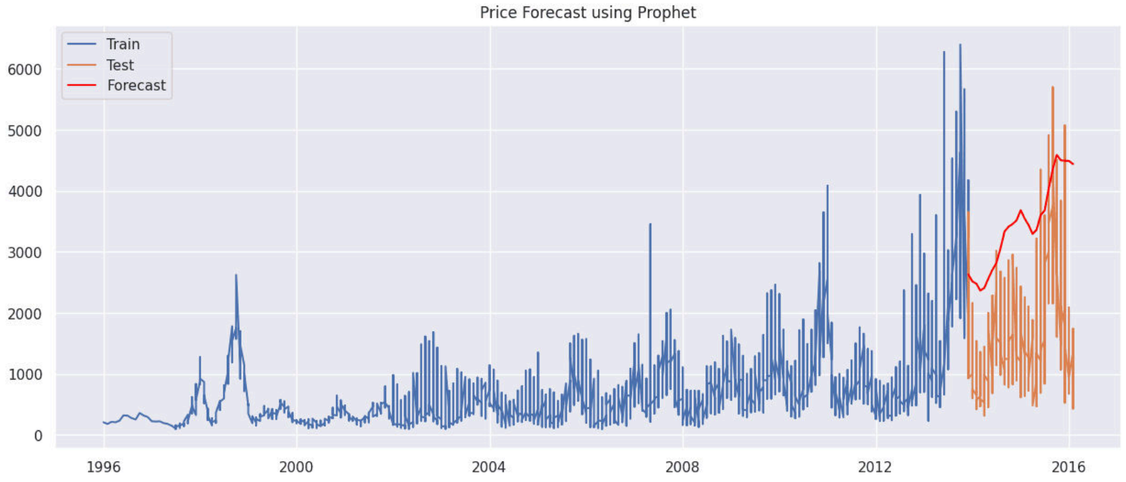


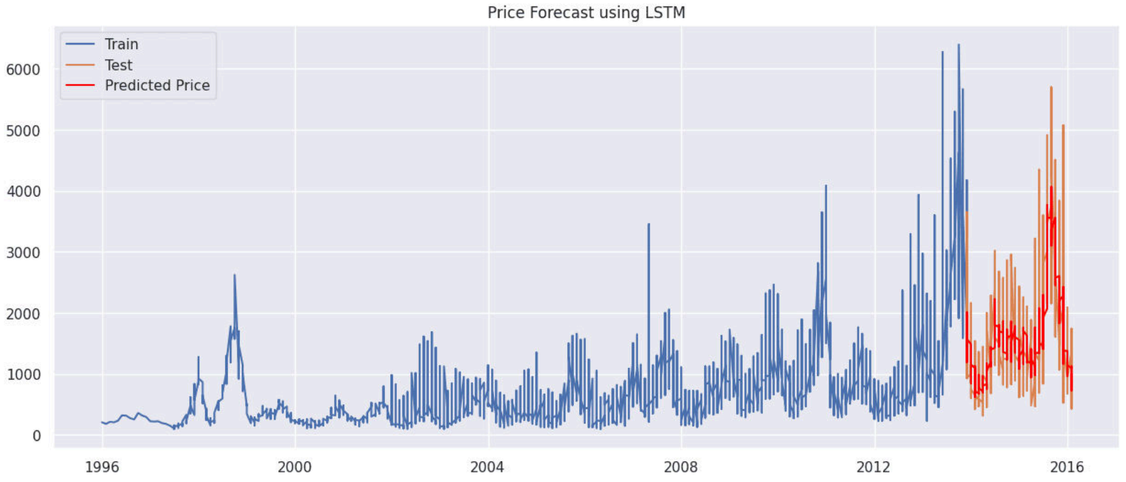


Model Selection and Training



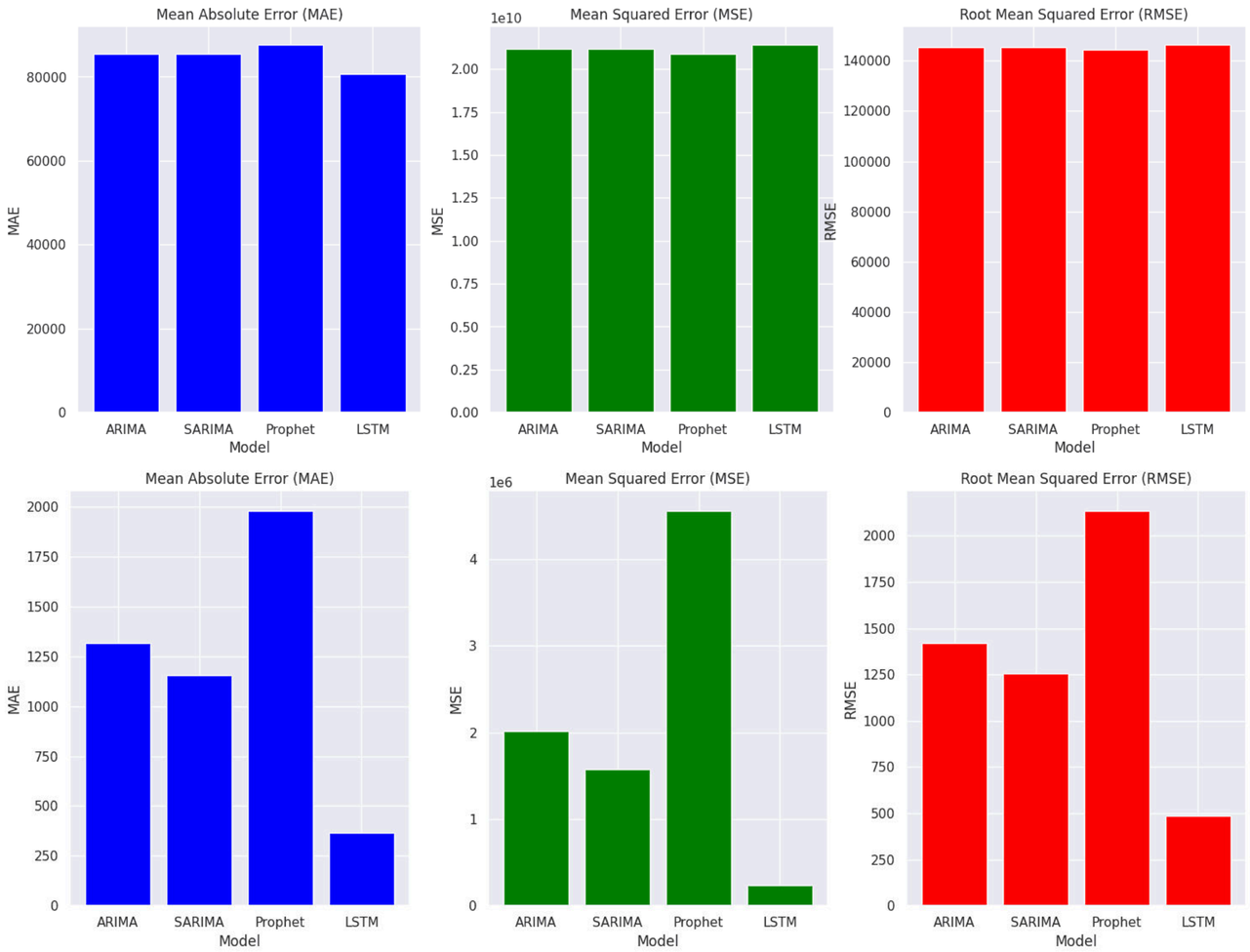






Performance Metrics:

- Mean Absolute Error (MAE)
- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)



Evaluating Model Performance

Quantity Forecasting: The LSTM model has the lowest MAE, indicating it may be more accurate on average for quantity forecasting. Prophet has the lowest MSE, suggesting it handles large deviations better. Prophet has the lowest RMSE, providing a balanced performance in both small and large errors.

Price Forecasting: The LSTM model outperforms all other models significantly in terms of MAE, MSE, and RMSE. This makes LSTM the most suitable model for price forecasting based on the given metrics.

Recommended Models: For quantity forecasting, the LSTM model is recommended for its lowest MAE. For price forecasting, the LSTM model is clearly the best performer across all metrics.

Fine-Tuning the Model

Fine-tuning:
Adjusted model parameters for better performance.
Validation:
Tested model on unseen data. Ensured robustness and accuracy.

Conclusions

Achievements:
Developed a robust forecasting model. Accurately predicted future quantities and prices.
Impact:
Informed decision-making for stakeholders. Optimized inventory management and pricing strategies.
Future Work:
Explore additional models. Incorporate more external data for improved accuracy.

Thanks!

Do you have any questions?

soufleros.kostas@gmail.com
in/konstantinos-soufleros
https://github.com/kostas696