Date: 19/07/2024

Infosys Springboard

Theme : AI First

Presentation on

Demand Forecasting for E-commerce

-By Vidhi Vaishnav

Problem statement

The realm of E-Commerce, demand forecasting plays a pivotal role in ensuring business success. This project aims to develop a demand forecasting model in an E-commerce business that predicts future product demand leveraging time series analysis and multivariate regression based on historical sales data, along with Google Analytics KPIs such as Google clicks and Facebook impressions, which are valuable indicators of customer interest.

- Objective: Develop a demand forecasting model for an E-commerce business.
- Purpose: Predict future product demand.
- Approach: Utilize time series analysis and multivariate regression.
- Data Sources: Historical quantity sales data, Google Analytics KPIs, Google clicks, and Facebook impressions.
- Significance: Leverage customer interest indicators to enhance business success.

Contents

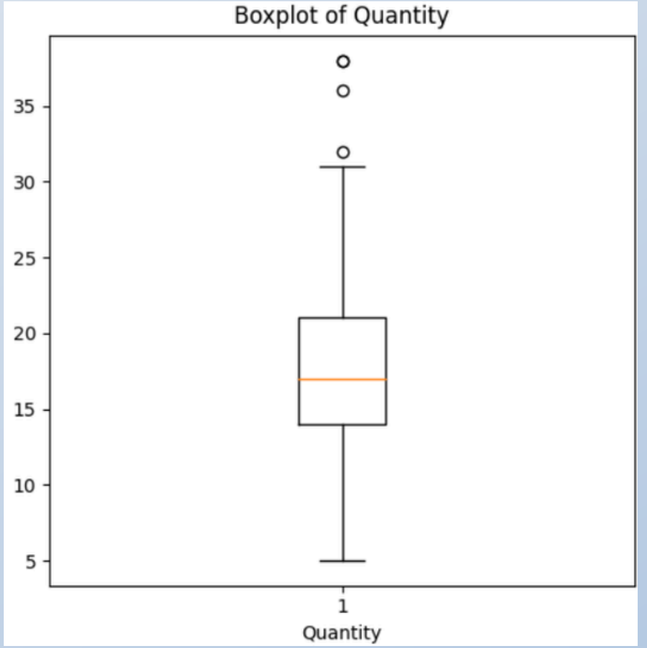
Exploratory Data Analysis

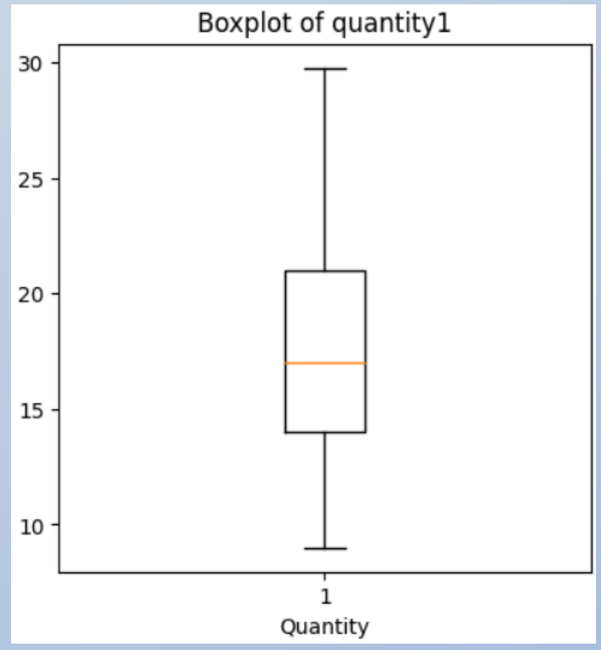
➢ Data Preprocessing:

1. Data Loading and Merging:
- Loaded three datasets: Quantity, Facebook Impressions, and Google Clicks.
- Merged datasets on the common column 'Day Index'.

2. Data Overview:
- Total Rows: 212 and Total Columns: 4
- Columns: Day Index, Quantity, Impressions, Clicks
- No Missing Values: All columns have zero missing values.

3. Outlier Detection and Imputation:
- Outliers: Identified in the 'Quantity' column (values above 30).
- Imputed outliers in 'Quantity' using outlier capping method and created a new column 'quantity1'.





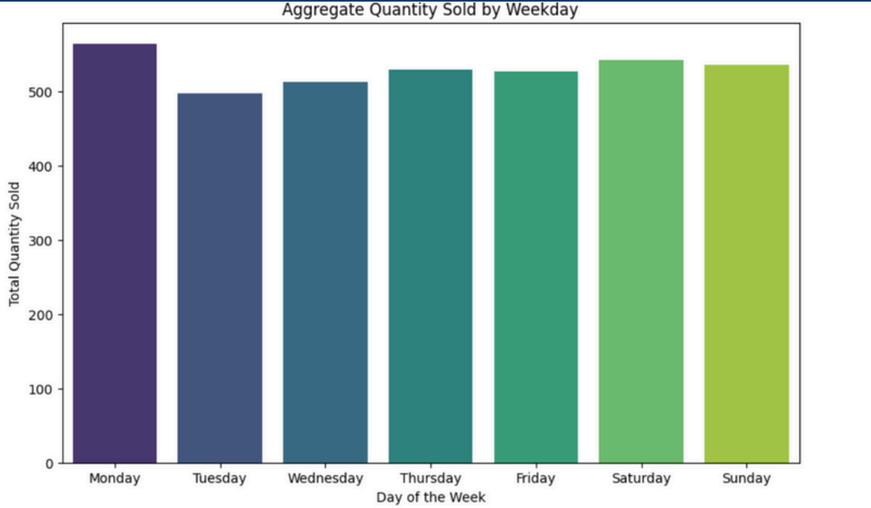
Exploratory Data Analysis

4. Correlation Analysis:
   - There is a strong positive correlation between Impressions and Clicks.
   - A moderate positive correlation exists between Quantity and Clicks.
   - A weak positive correlation is observed between Quantity and Impressions.

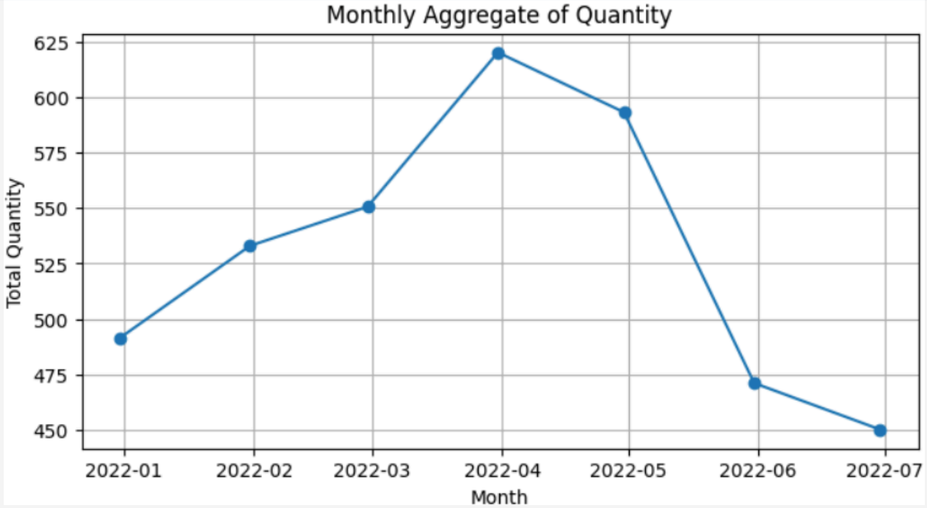
Feature Engineering:

- Extracted features from 'Day Index': Year, Month, Quarter, Weekday.
- Created new features: Is Weekend, Is Holiday, Sales_per_Impression, Sales_per_Click.
- Lag features: Quantity_Lag1, Quantity_Lag7.
- Created dummy variables for categorical features: Season, Day Type, Weekday, Month

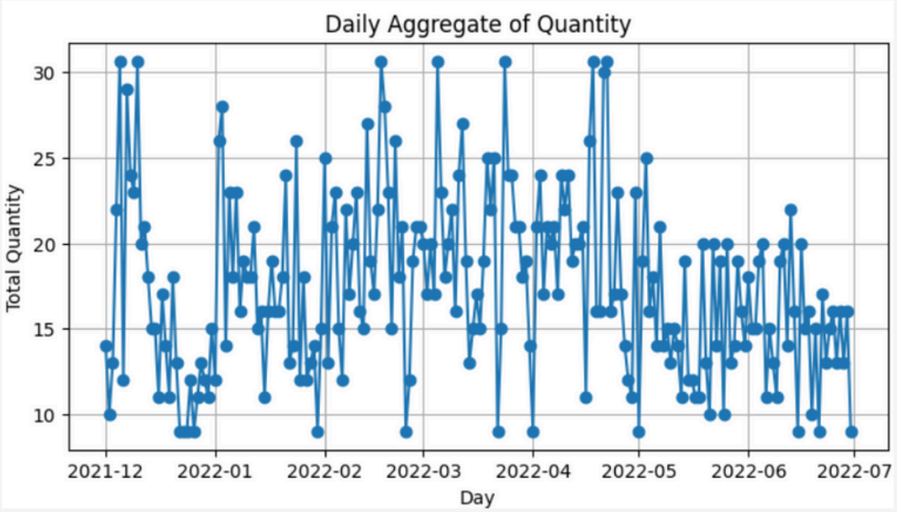
Visualisation & its Insights



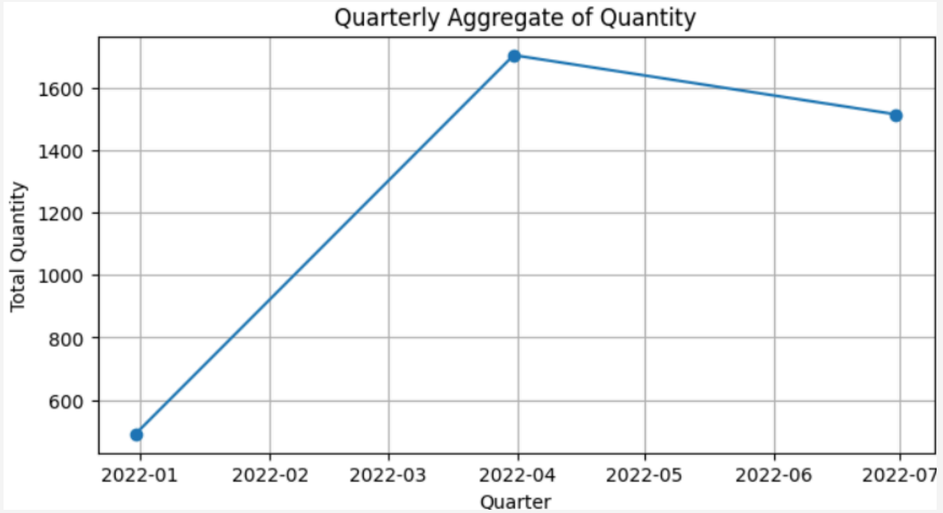
On weekdays, Monday has the highest sales volume, followed by Saturday and Sunday, with peak sales on Mondays and Sundays.



In April 2022, highest sales; July 2022 had lowest. Seasonal trends show increased demand in spring and early summer.

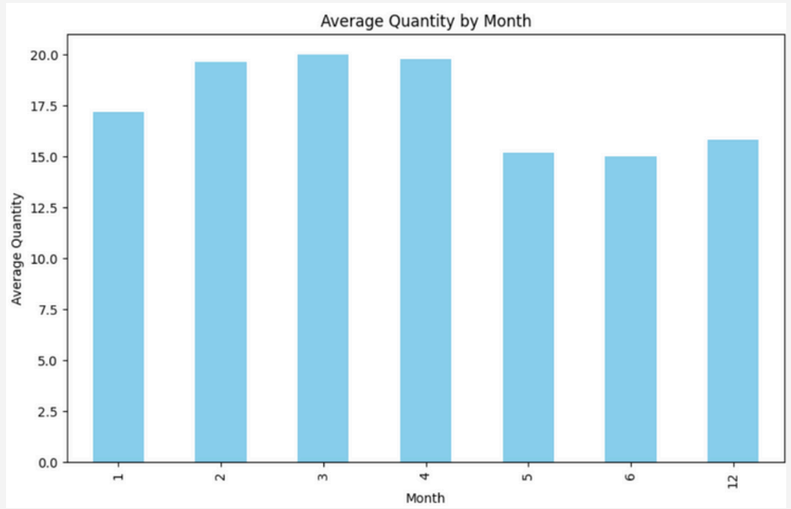


Daily sales peaked in December 2021, declined in January 2022, and followed a seasonal pattern with higher winter and lower summer demand from March to June 2022.

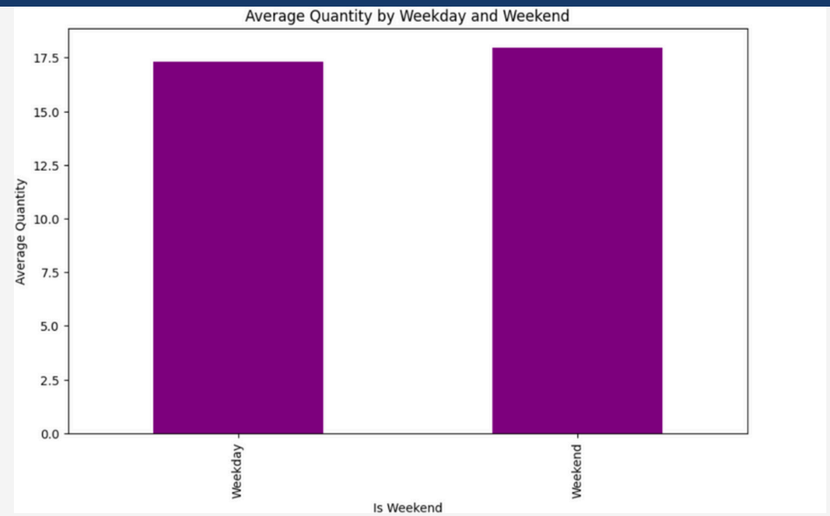


Quantity sold increased from Q1 to Q2 2022, peaked in Q3 2022, and slightly decreased in Q4 2022

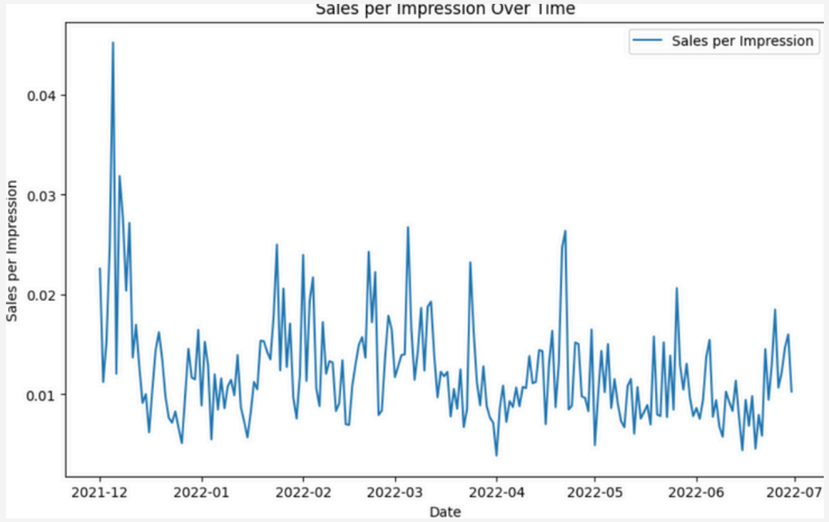
Visualisation & its Insights



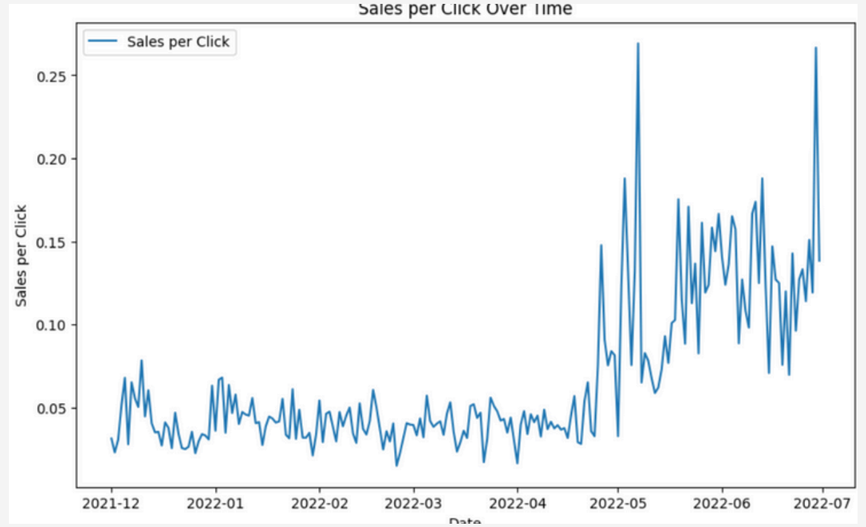
Highest sales were in March and April 2022, while December 2021 had the lowest sales.



On weekends, sales show a slightly higher average quantity sold compared to weekdays.



Sales per impression show a declining trend with fluctuations, reaching the lowest point in July 2022 at 0.01 and the highest in December 2021 at 0.05.



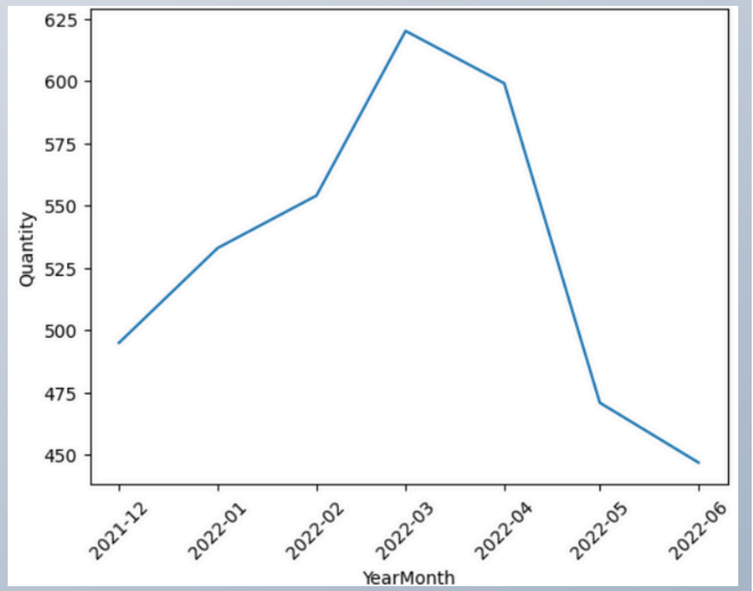
Sales per Click remained stable and low from December 2021 to March 2022, with a significant spike in May 2022 that remained high until July 2022.

Time series analysis

1. Stationarity:

ADF Test Results:

- Quantity: Stationary (ADF Statistic: -4.35, p-value: 0.00036)
- Impressions: Stationary (ADF Statistic: -5.70, p-value: 7.86e-07)
- Clicks: Non-stationary (ADF Statistic: -0.87, p-value: 0.7975)
- Clicks data became stationary after 1st order differencing
- (ADF Statistic: -14.22, p-value: 1.64e-26).

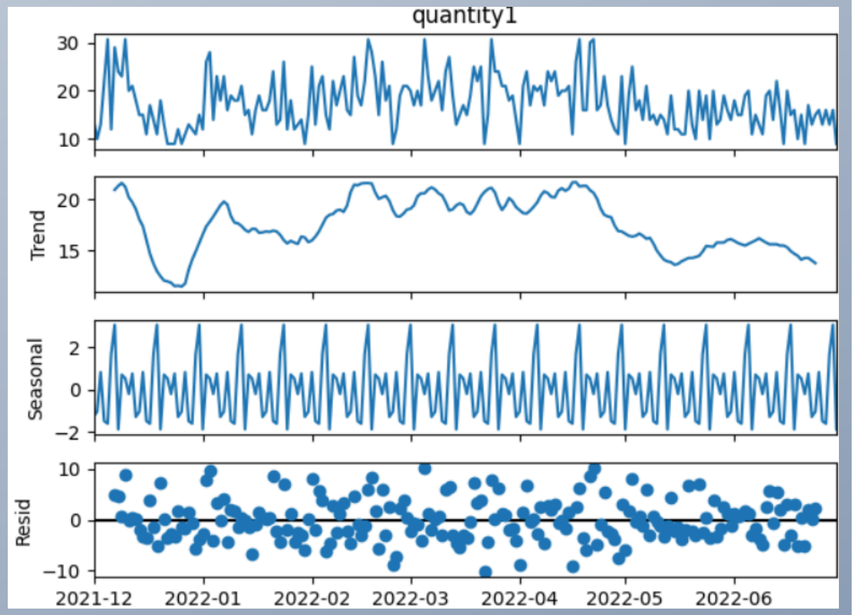


2. Trend analysis:

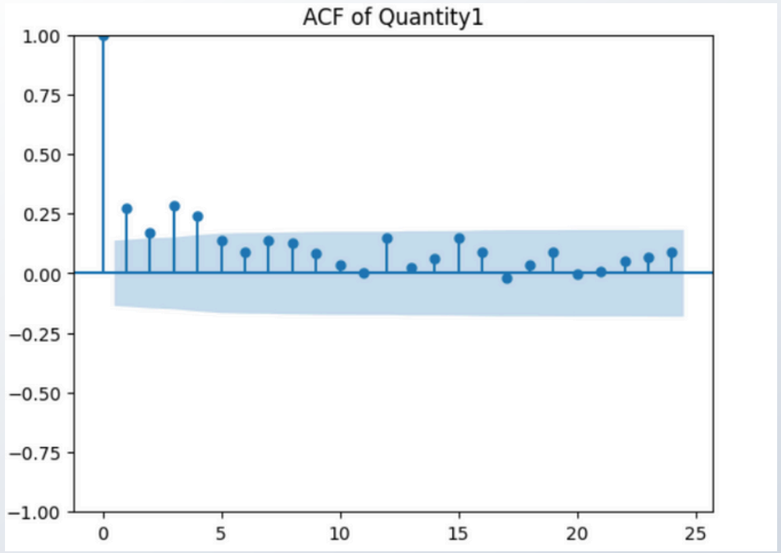
- Upward Trend: From December 2021 to March 2022, peaking in March 2022.
- Sharp Decline: From March 2022 to May 2022.
- Slow Decline: From May 2022 to June 2022.

3. Trend, Seasonal, and Residual Components:

- Trend Component: Shows a general decrease in quantity over time.
- Seasonal Component: Displays a slight, repeating pattern.
- Residuals: Appear random, indicating the model captures most of the variability in the time series.

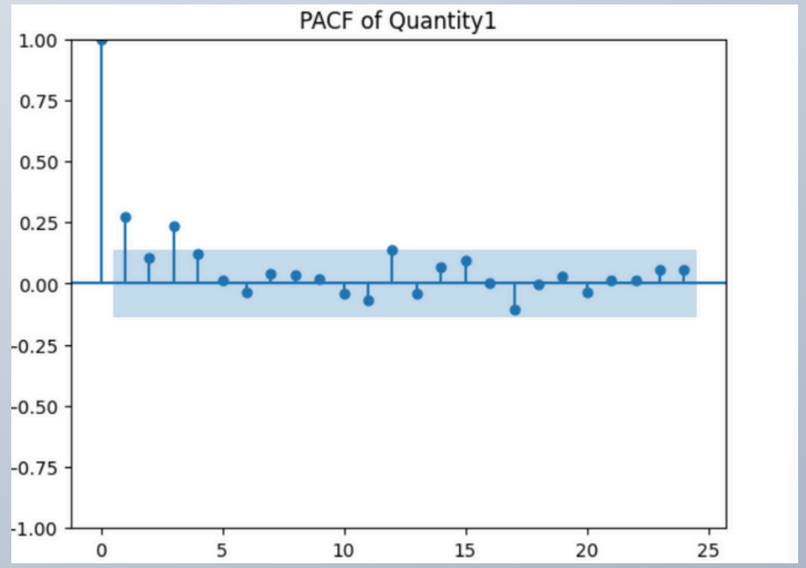


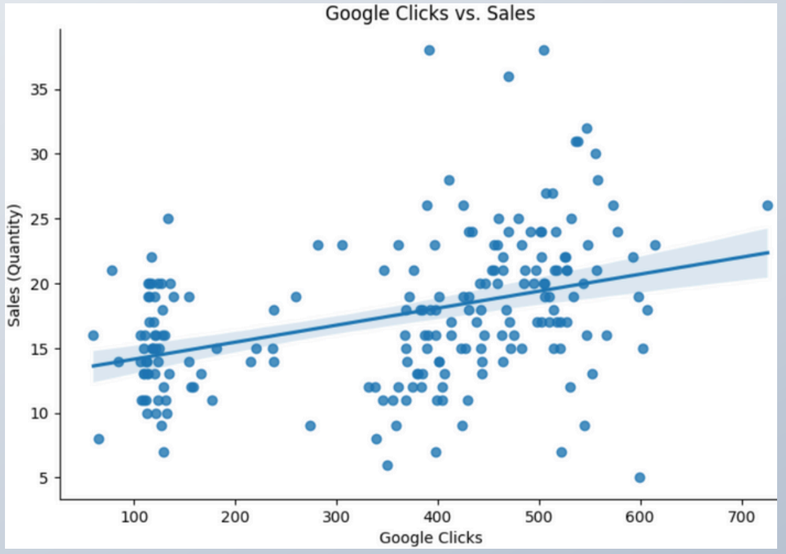
Time series analysis



4. Autocorrelation:

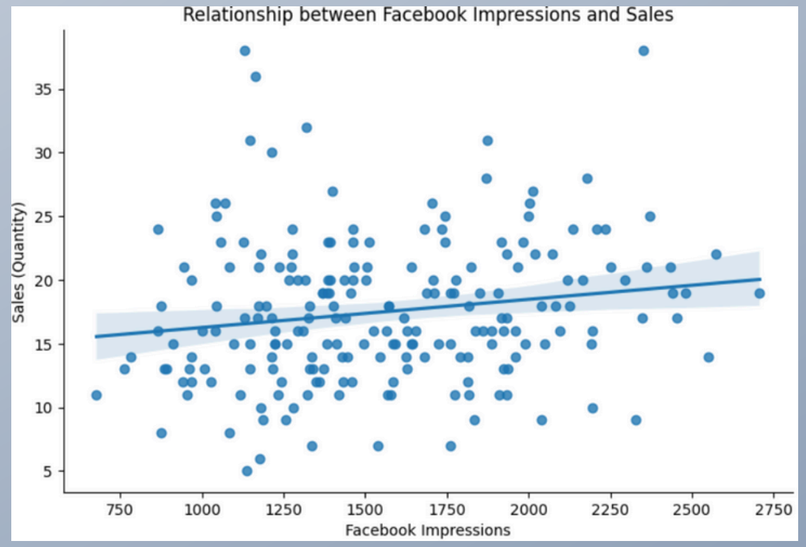
- Autocorrelation plot shows strong positive autocorrelation at lag 1, indicating stationarity.
- Partial autocorrelation plot reveals significant autocorrelation at lag 1, with subsequent autocorrelations declining to zero.





5. Linear regression:

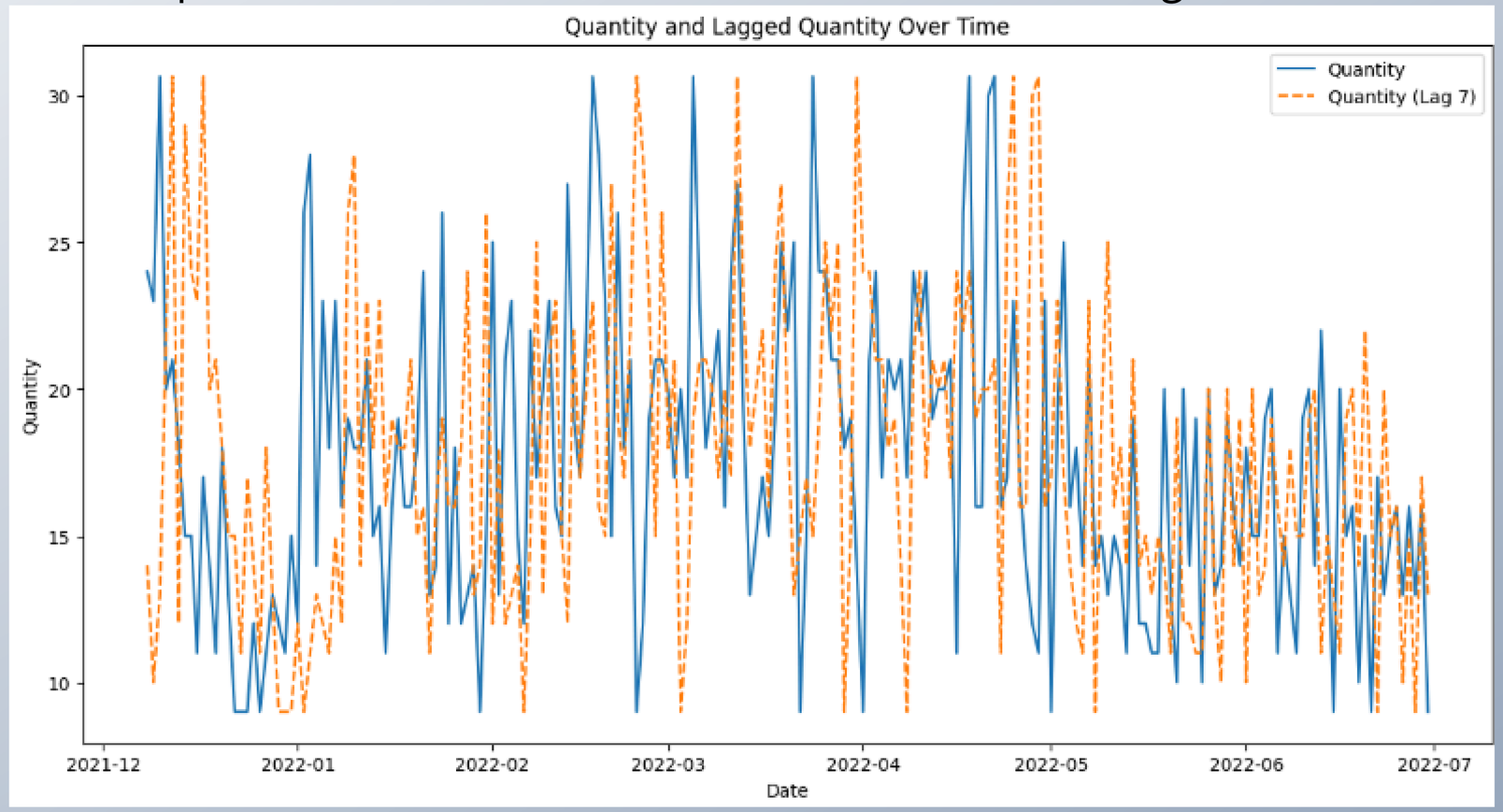
Positive correlations exist between sales and Google clicks (Pearson Correlation: 0.376) and between sales and Facebook impressions (Pearson Correlation: 0.135), with Google clicks showing a stronger impact on sales quantity than Facebook impressions.



Time series analysis

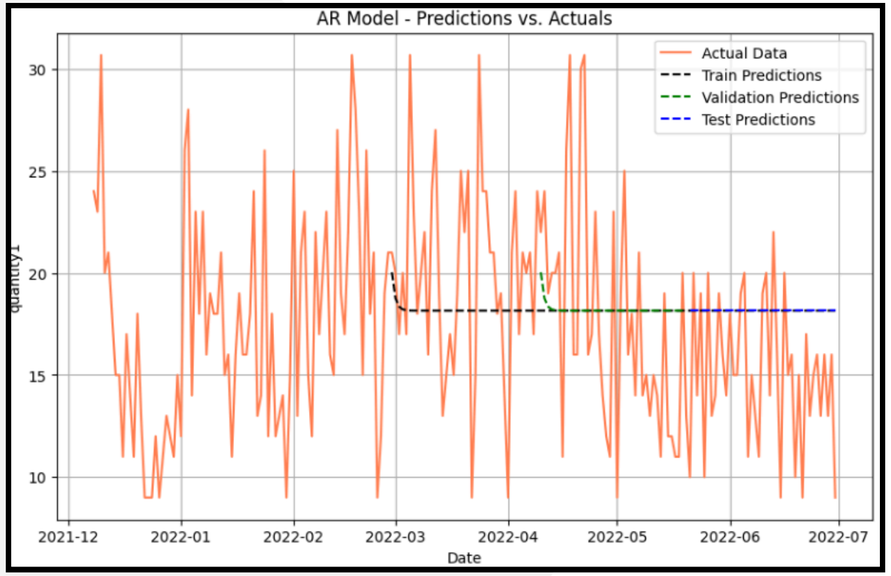
5. Quantity & lagged quantity over time:

- The plot shows the original quantity and lagged quantity (lag 7) over time.
- Pattern Similarity: The lagged quantity (lag 7) follows the same pattern as the original quantity.
- Autocorrelation Insight: This suggests that the series is autocorrelated, indicating that past values can be used to predict future values, which is useful for forecasting.



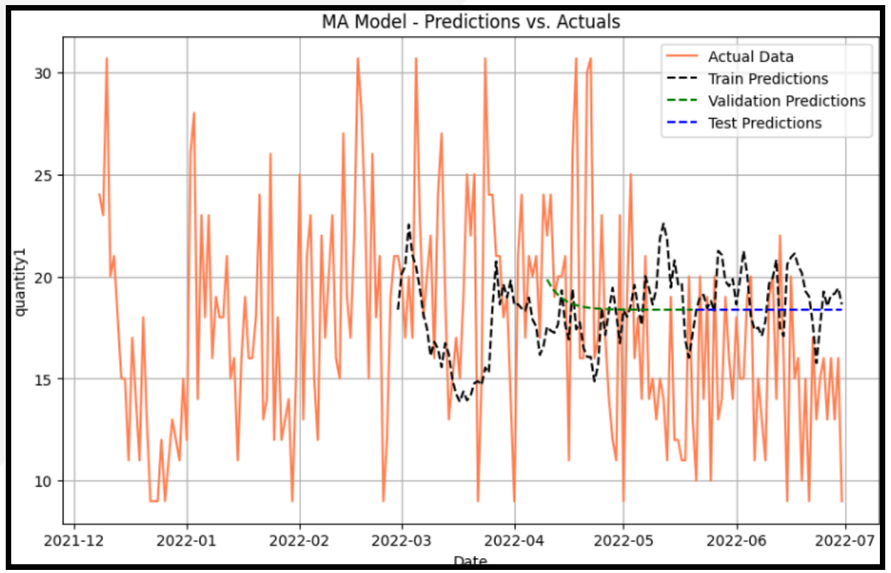
Implementation of Models

AR model



Metric	Validation Set	Test Set
RMSE	5.598	4.628
MAPE	29.445%	31.526%
MAE	4.625	3.858
R2	0.004	-0.772
Adjusted R2	-0.021	-0.817

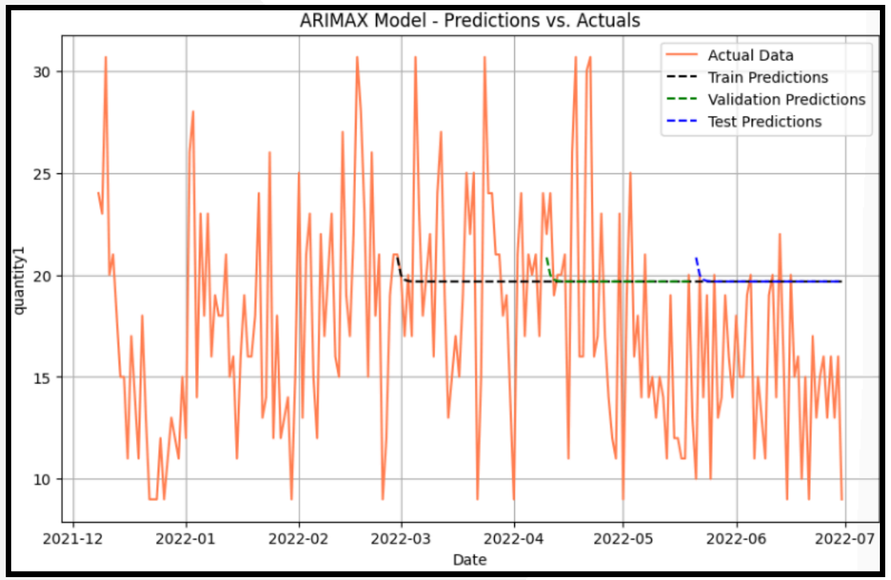
MA model



Metric	Validation Set	Test Set
RMSE	5.592	4.774
MAPE	29.871%	32.513%
MAE	4.617	3.969
R2	0.006	-0.885
Adjusted R2	-0.019	-0.934

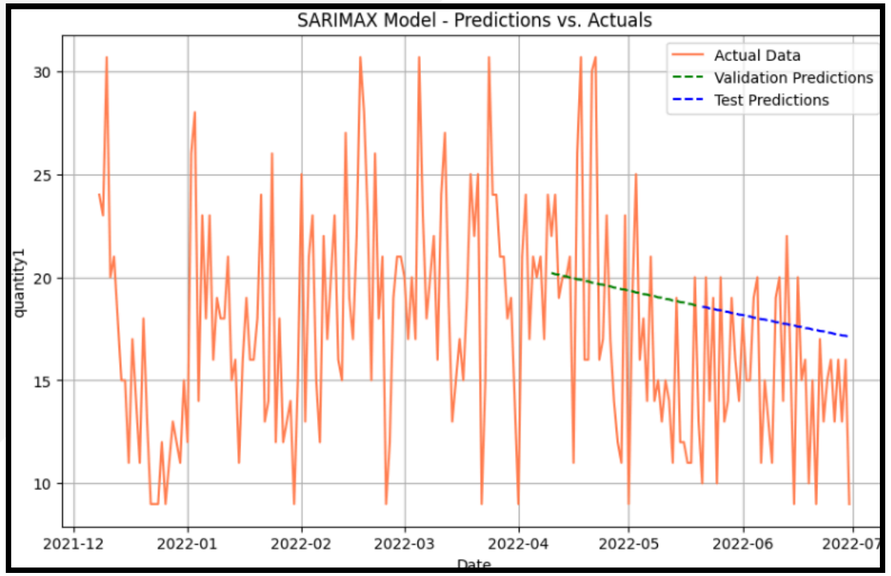
Implementation of Models

ARIMAX model



Metric	Validation Set	Test Set
RMSE	5.969	5.800
MAPE	34.036%	39.418%
MAE	5.016	4.793
R2	-0.132	-1.783
Adjusted R2	-0.161	-1.854

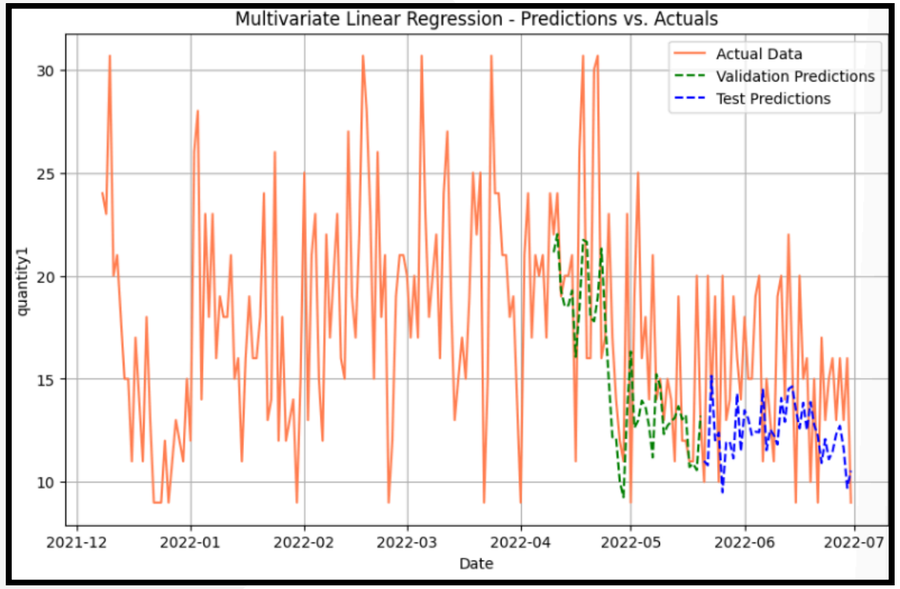
SARIMAX model



Metric	Validation Set	Test Set
RMSE	5.5926	4.7743
MAPE	29.8710%	32.5133%
MAE	4.6177	3.9698
R2	0.0061	-0.8857
Adjusted R2	-0.0193	-0.9340

Implementation of Models

Multivariate linear regression model



Metric	Validation Set	Test Set
RMSE	5.430	4.632
MAPE	21.22%	23.970%
MAE	3.993	3.804
R2	0.062	-0.775
Adjusted R2	-0.209	-1.290

Comparison of Models

Model	Set	RMSE	MAPE	MAE	R2	Adjusted R2
AR Model	Validation	5.598	29.445%	4.625	0.004	-0.021
	Test	4.628	31.526%	3.858	-0.772	-0.817
MA Model	Validation	5.592	29.871%	4.617	0.006	-0.019
	Test	4.774	32.513%	3.969	-0.885	-0.934
ARIMA Model	Validation	5.592	29.871%	4.617	0.006	-0.019
	Test	4.906	33.096%	4.014	-0.991	-1.042
ARIMAX Model	Validation	5.969	34.036%	5.016	-0.132	-0.161
	Test	5.800	39.418%	4.793	-1.783	-1.854
SARIMAX Model	Validation	5.681	32.047%	4.773	-0.025	-0.052
	Test	4.382	29.362%	3.582	-0.589	-0.629
Multivariate Linear Model	Validation	5.430	21.220%	3.993	0.062	-0.209
	Test	4.632	23.970%	3.804 acres	-0.775	-1.290

Conclusion: Multivariate Linear Regression Model

Model Selection: Multivariate Linear Regression was chosen for its ability to incorporate multiple predictors, including historical sales data, Google clicks, and Facebook impressions.

Metric	Validation Set	Test Set
RMSE	5.430	4.632
MAPE	21.22%	23.970%
MAE	3.993	3.804
R2	0.062	-0.775
Adjusted R2	-0.209	-1.290

- Analysis:
  - RMSE and MAE values indicate acceptable prediction error.
  - MAPE values show percentage error, with test set error slightly higher than validation set.
  - R^2 and Adjusted R^2 metrics indicate variance explained by the model, with better performance on validation data.

- Implications:
  - The model effectively uses multiple data sources to improve demand forecasting accuracy.
  - Despite some limitations in R^2, the error metrics (RMSE and MAE) show the model's usefulness in predicting future demand, aiding inventory management and strategic planning.

Thank You!