

# AI-Driven Demand Prediction for Smarter Retail

Presented by:

Nitin Manohar Mishra

Chris Joy

Sri Harshini

Swastik Roy Choudhury

# Introduction to the Project

- Project Overview:
- The AI-driven demand prediction project focuses on using advanced time series forecasting models to predict product demand for retail businesses, helping improve inventory management, optimize marketing efforts, and enhance customer satisfaction.
- Objectives:
- Using AI models to forecast demand for retail products based on historical sales data and external factors like Google Analytics.
- Motivation:
- Accurate forecasting allows businesses to manage stock levels more effectively, reduce overstock/stockouts, and improve customer satisfaction.

# Data Sources & Overview

- Data Description:
- The dataset used includes historical sales data (target variable: Quantity) and external factors such as Clicks and Impressions (from Google Analytics and social media).
- Data Structure:
- Target Variable: Quantity (sales data)
- Exogenous Variables: Clicks, Impressions (external factors influencing demand)
- Data Quality:
- Challenges included handling missing values and data cleaning.

# Step 1 : Data Pre Processing

BEFORE	AFTER
Day Index	Day Index
01-12-2021	2021-12-01 00:00:00
02-12-2021	2021-12-02 00:00:00
03-12-2021	2021-12-03 00:00:00
04-12-2021	2021-12-04 00:00:00
05-12-2021	2021-12-05 00:00:00
06-12-2021	2021-12-06 00:00:00
07-12-2021	2021-12-07 00:00:00
08-12-2021	2021-12-08 00:00:00
09-12-2021	2021-12-09 00:00:00
10-12-2021	2021-12-10 00:00:00
11-12-2021	2021-12-11 00:00:00
12-12-2021	2021-12-12 00:00:00
13-12-2021	2021-12-13 00:00:00
14-12-2021	2021-12-14 00:00:00
15-12-2021	2021-12-15 00:00:00
16-12-2021	2021-12-16 00:00:00

## Step 1: Convert the 'Day Index' column to a proper datetime format

Converting the 'Day Index' column to a proper datetime format ensures consistency, accuracy, and compatibility with date-based operations. It allows for easier analysis, time-based aggregations, and smooth data merging or filtering, while preventing errors or mismatches due to incorrect date formats.

## Step 2: Remove duplicate rows

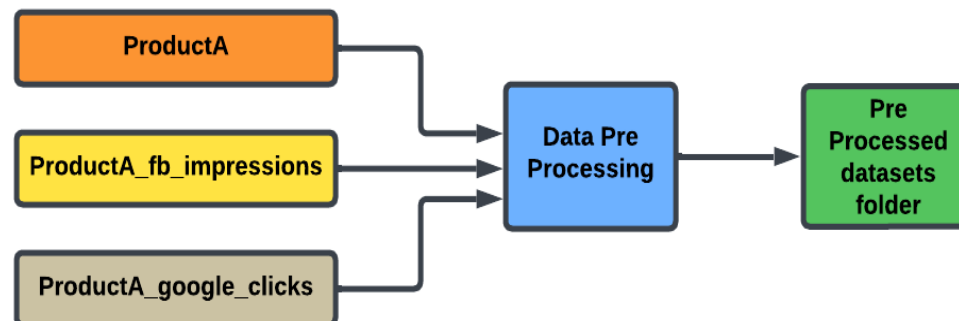
There were no duplicate rows in the dataset, so this step was not necessary.

## Step 3: Fill any missing values using forward fill

There were no missing values in the dataset, so this step was not needed.

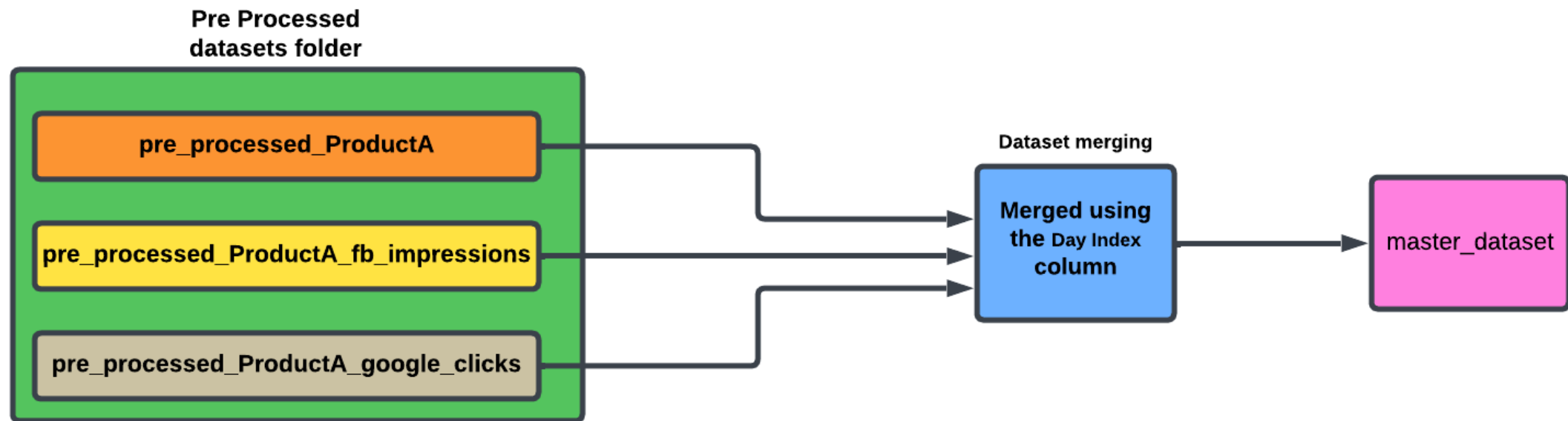
## Final Step :Output Data

Cleaned datasets are saved into the pre\_processed\_datasets folder, ready for merging.



# Step 2 : Dataset Merging

The process involves loading preprocessed datasets from the **pre\_processed\_datasets** folder, merging them based on the common column **Day Index**, and saving the final merged dataset to the **master\_dataset** folder.



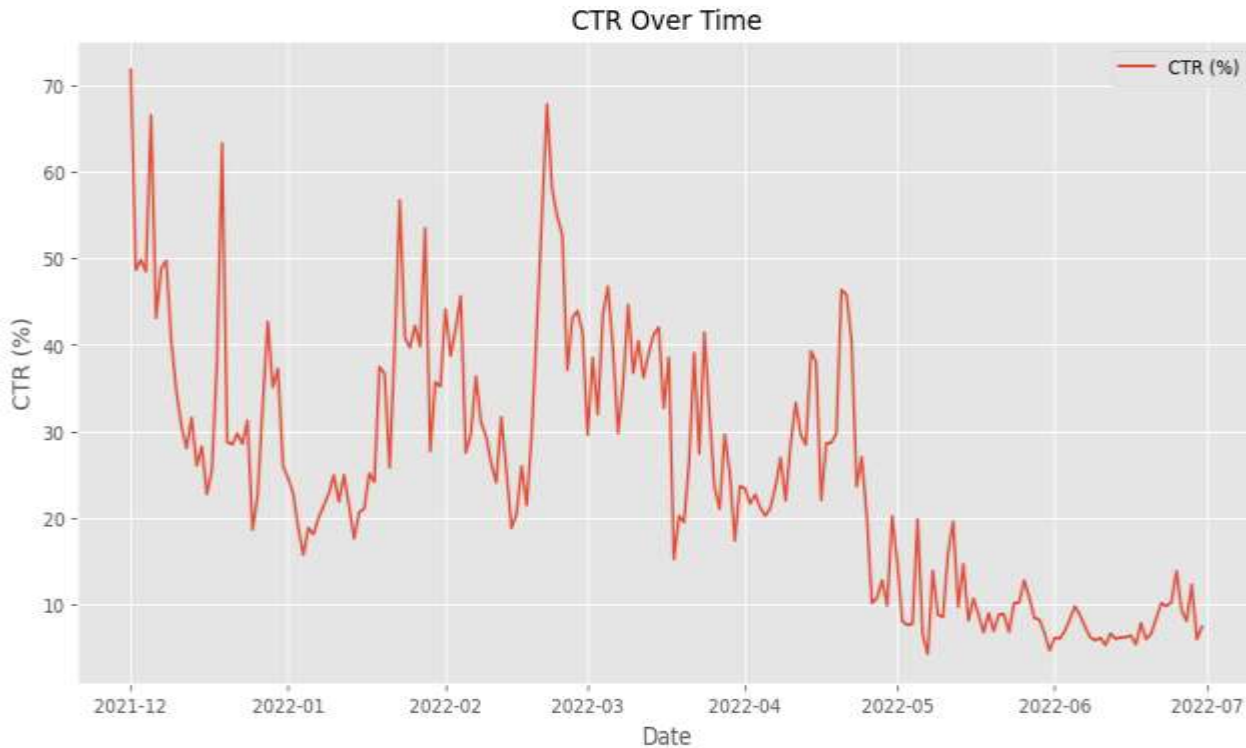
# Step 3: Data Analysis and Visualization

**Why We Need This Code:** The code is essential for analyzing the Master Dataset, as it transforms raw data into actionable insights. It helps uncover trends, correlations, and key metrics crucial for making informed decisions in the project.

## **Purpose of this code:**

- **Feature Engineering:** It calculates vital metrics such as Click-Through Rate (CTR), Conversion Rate, and Sales per Click, which enhance data comprehension.
- **Data Cleaning:** The code addresses missing values and infinities, ensuring that calculations and visualizations are robust and reliable.
- **Visualizations:** A series of plots and charts are generated to highlight relationships, growth patterns, and insights in the data, making it easier to identify trends over time.
- **Key Takeaways and Conclusion:** This code allows us to analyze and visualize key performance metrics, track trends over time, and identify factors influencing sales. It is a crucial step in understanding data, making it more actionable, and supporting better decision-making for demand prediction and business planning.

# Visual 1: Viewing CTR over Time



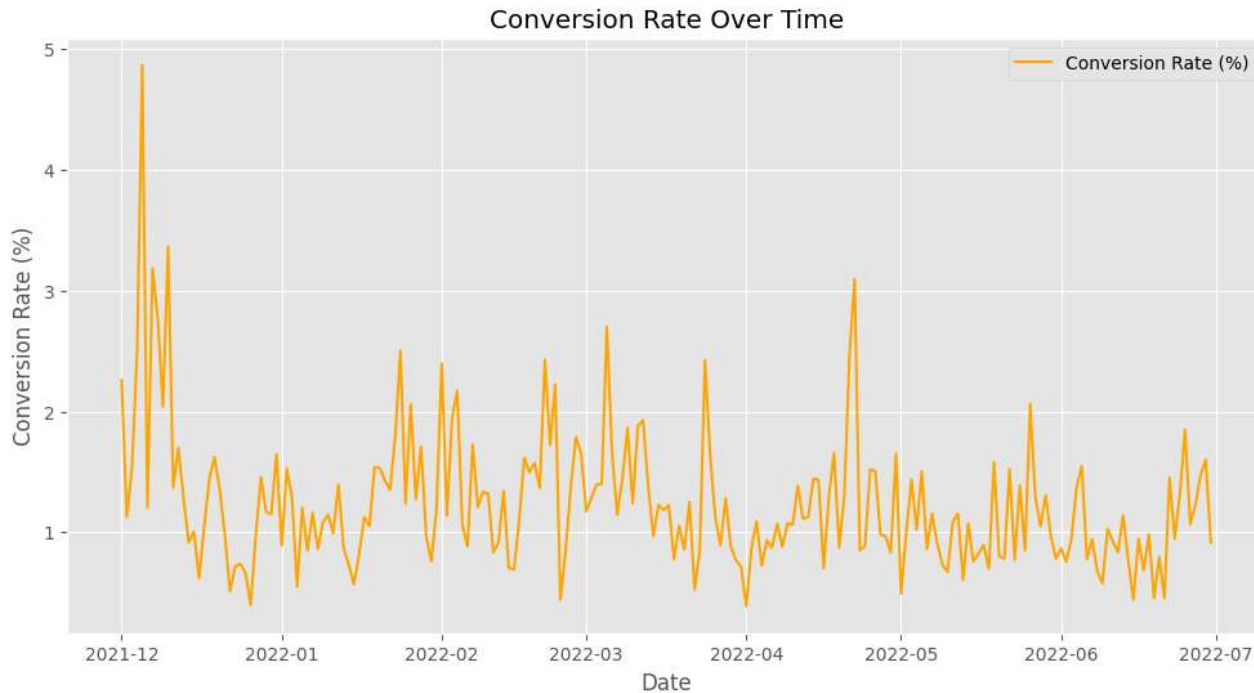
**Objective:** To see how CTR changes over time and also to check how clicks perform in different times.

## **Key Insights:**

- We observed big variations in CTR, with abrupt spikes showing high efficiency on certain days.
- The trend line indicates a general downward trend over time, suggesting less user engagement or perhaps the campaign became less effective over time.
- The plot shows times of growth and decline, giving helpful ideas to improve strategies.

**Conclusion:** By looking at the CTR trends, we can find exact times that led to better engagement, allowing for more focused changes to marketing efforts.

# Visual 2: Viewing Conversion Rate over Time



**Objective:** Time-series analysis of the trend of Conversion Rate, understand customer behavior and the impression's effectiveness.

## **Key Findings:**

- Conversion Rate time-series has spikes at some intervals and thus reflects certain time periods that were more conversion-efficient.
- There is generally a pattern in the Conversion Rate; it goes up in some months and goes down in other months, probably because of seasonality or the effect of losing the effectiveness of campaigns.
- We know now because of the plot that there were instances when conversion efficiency was great, giving useful ideas for future marketing strategies.

**Conclusion:** From Conversion Rate trends, we can track successful periods that help to adjust marketing efforts according to customer engagement and craft even better campaign strategies in the future.



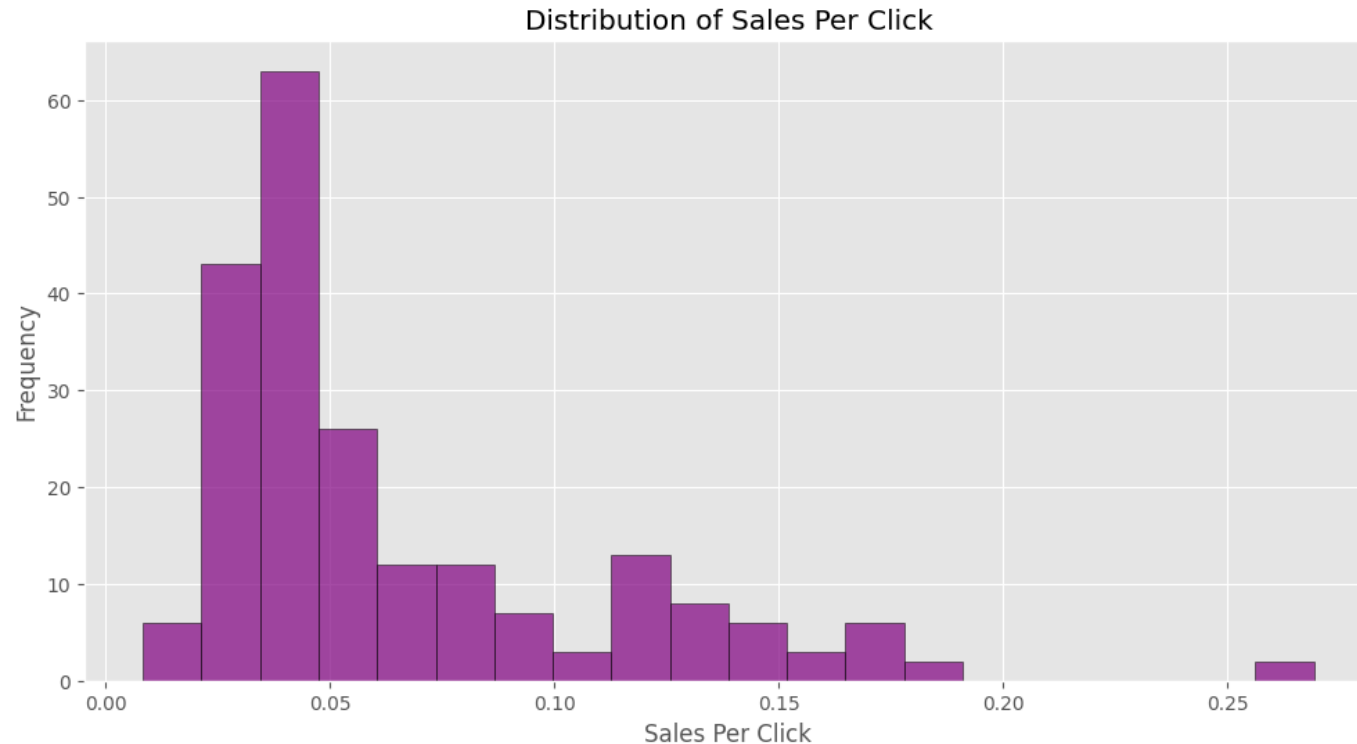
# Visual 3: Distribution of Sales Per Click

**Goal:** Use visualizations to show Sales Per Click that contributes to each click effectively for determining the best possible sales outcomes.

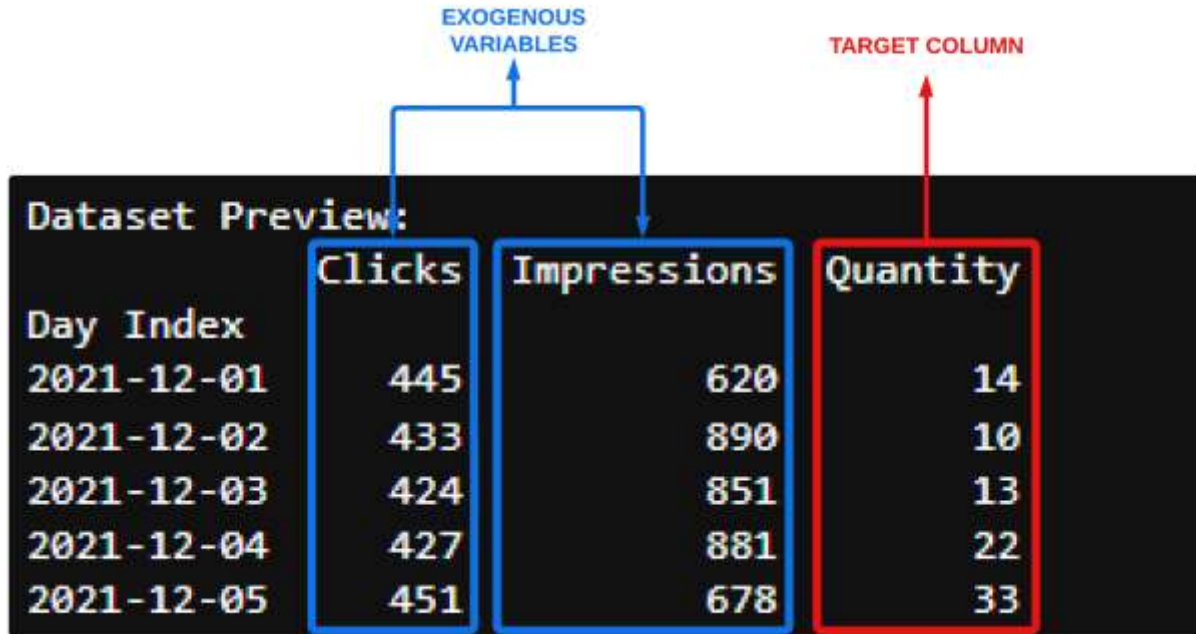
**Main Points:**

- The histogram has a giant peak at about 0.05 Sales Per Click—meaning most sales happen through rather low efficiency for each click; this would indicate improvements required in marketing efforts to help push clicks into actually going into sales.
- There is also a tail of lower frequency that emerges when values of Sales Per Click become too high. The reason is that all the clicks are not turning into substantive sales. This can be a result of misaligned targeting and messaging, which doesn't always turn clicks into meaningful sales.
- Spreads in the wide indicate that not all clicks are equal. Thus, there may be room for differentiation by segmentation of strategies to take more days or patterns on a higher conversion.

**Conclusion** :The data on Sales Per Click reveals that some clicks work really well, but most sales come from less effective clicks. Thus, finding and focusing on the more effective groups of clicks will allow marketing plans to be adjusted to boost overall sales and cut down on wasted efforts.



# Step 4 : Dataset Loading and Preparation



Dataset Preview:

Day Index	Clicks	Impressions	Quantity
2021-12-01	445	620	14
2021-12-02	433	890	10
2021-12-03	424	851	13
2021-12-04	427	881	22
2021-12-05	451	678	33

- Loaded the dataset using **pandas** (`read_excel`) and ensured correct file path.
- Key columns: Quantity, Clicks, and Impressions.
- Set the **Day Index** to organize the data in time series format for forecasting.
- **Exogenous Variables**: Identified and processed **Clicks** and **Impressions** to influence the target variable **Quantity**.
- **Data Alignment**: Ensured matching number of observations between target and exogenous variables for consistency in the forecasting model.

# Stationarity Check and Data Preparation

- **Stationarity Check and Data Preparation**
- **Objective:** Ensured the stationarity of the target column **Quantity** (required for time series modeling).
- **Steps Taken:**
  - **ADF Test:** Performed the Augmented Dickey-Fuller (ADF) test to check if the series was stationary.
    - **Dickey-Fuller Test Statistic:** -4.45
    - **p-value:** 0.00025 (indicating that the series is stationary as  $p \leq 0.05$ ).
  - **Differencing:** Since the series was stationary, no differencing was required. (If the series had been non-stationary, first-order differencing would have been applied.)
- **Outcome:**
  - The **target series (Quantity)** passed the stationarity test and was ready for further analysis.
  - Exogenous variables were adjusted to match the transformed target series.

# Implementing Time Series Models

**1. Objective :** Implement various time series forecasting models to predict the target variable **Quantity**.

**2. Models to be Implemented :**

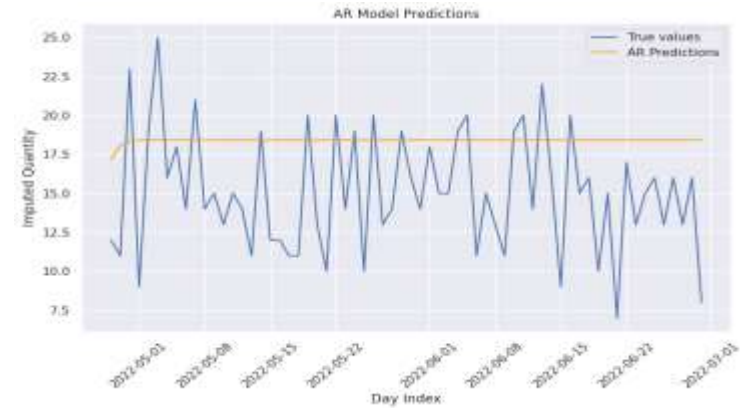
- Autoregression (AR)
- Moving Average (MA)
- ARIMA
- SARIMA
- ARIMAX
- SARIMAX

**3. Goal :**

- **Understand Model Performance:** Analyze how well each model captures patterns, trends, and seasonality in the data.
- **Evaluate External Variables:** Assess the impact of external factors (Clicks and Impressions) using models like ARIMAX and SARIMAX.
- **Select the Best Model:** Use performance metrics such as **Mean Squared Error (MSE)** and **R-squared ( $R^2$ )** to identify the most suitable model for **demand prediction**.

**4. Outcome :** The comparative analysis will help determine the best model for **forecasting product demand** in the FutureCart project.

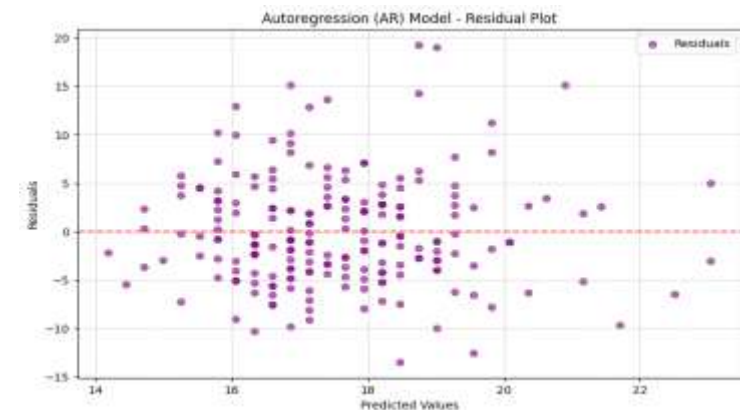
# Model 1: Autoregression (AR)



The AR model effectively captures daily fluctuations in the time series but struggles with larger variations, showing discrepancies between actual and predicted values for certain periods.

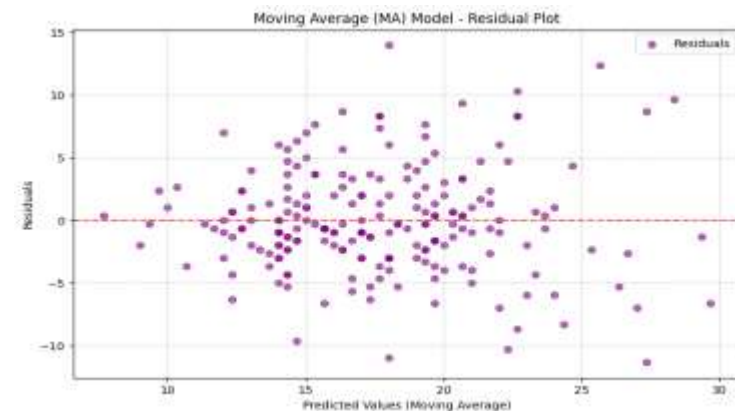
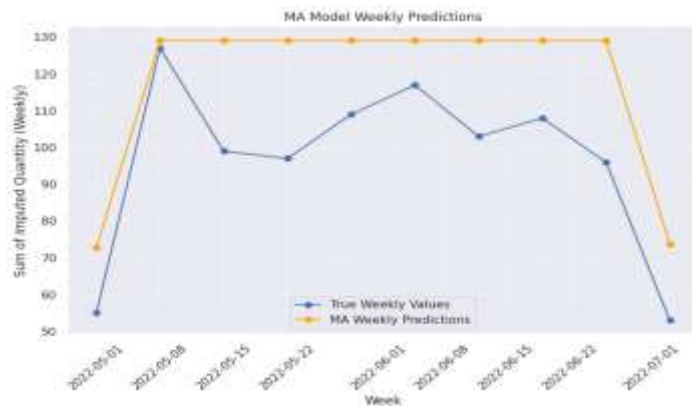
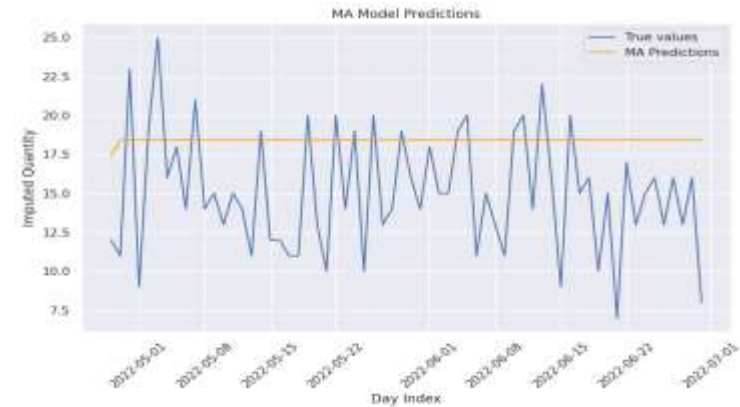


The AR model has difficulty capturing longer-term trends, as evidenced by the differences between actual and predicted weekly values. The model does not fully account for the overall seasonality in the data.



The residual plot suggests that the model captures some patterns but also has significant outliers, indicating potential issues in the model's ability to fully explain the data's variance. Randomness in residuals is not ideal.

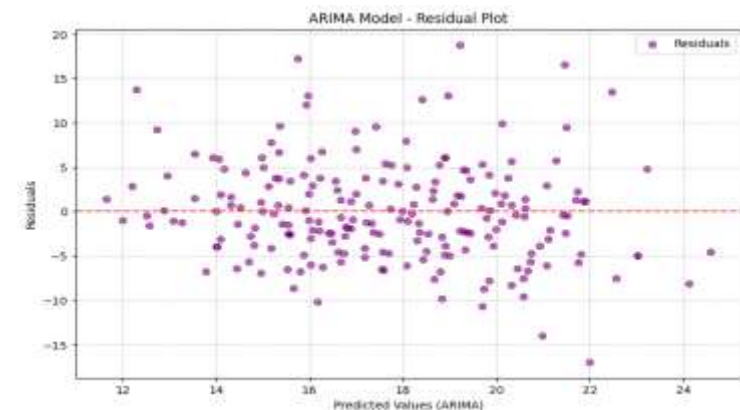
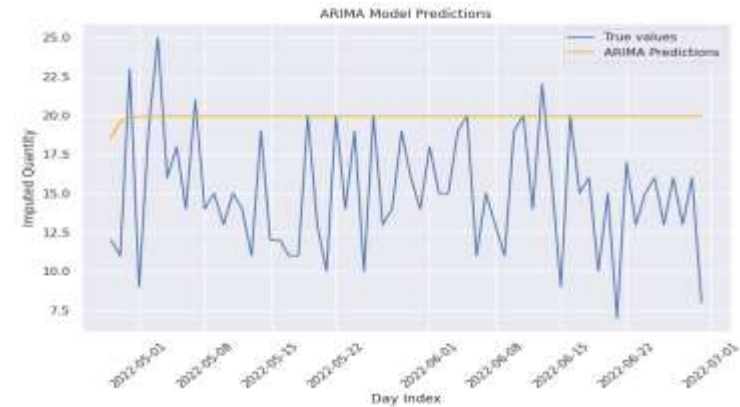
# Model 2: Moving Average (MA)



- The Moving Average (MA) model effectively smoothed the time series and identified short-term trends.
- The original series and predicted values demonstrated how well the MA model captured these trends with a rolling window.
- The model was able to approximate long-term trends when aggregated to weekly values.
- A clear comparison of the true and predicted weekly values showed how the MA model helped to smooth the variations in the data.
- A clear comparison of the true and predicted weekly values showed how the MA model helped to smooth the variations in the data.
- The residual plot showed some remaining patterns that the model did not capture, indicating areas where the model could be improved.

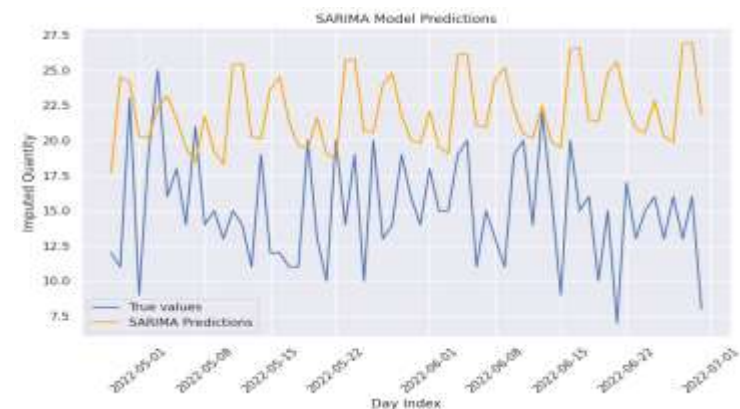


# Model 3: ARIMA (Autoregressive Integrated MA)



- The ARIMA model's hyperparameters were fine-tuned to find the optimal configuration for accurate forecasting.
- After testing multiple combinations of parameters ( $p$ ,  $d$ ,  $q$ ), the best model was identified with parameters ( **$p=1$ ,  $d=1$ ,  $q=2$** ), resulting in the lowest **Root Mean Square Error (RMSE) of 5.5470**
- The ARIMA model was used to predict daily values based on the optimized parameters.
- **Key Insight:** The ARIMA model accurately tracks daily trends and seasonal fluctuations, with some deviations between predicted and actual values.
- Predictions were aggregated on a weekly basis to capture longer-term trends. The ARIMA model displayed a reasonable fit, though some periodic divergences were observed.
- **Key Insight:** The ARIMA model successfully predicted general weekly trends, but minor discrepancies arose during certain periods.
- The residuals plot was used to check for randomness between actual and predicted values. A relatively random distribution around zero indicated that most patterns were captured by the model.
- **Key Insight:** While the model performed well, some small patterns remained unexplained, suggesting that more complex seasonality or external factors could improve the model.

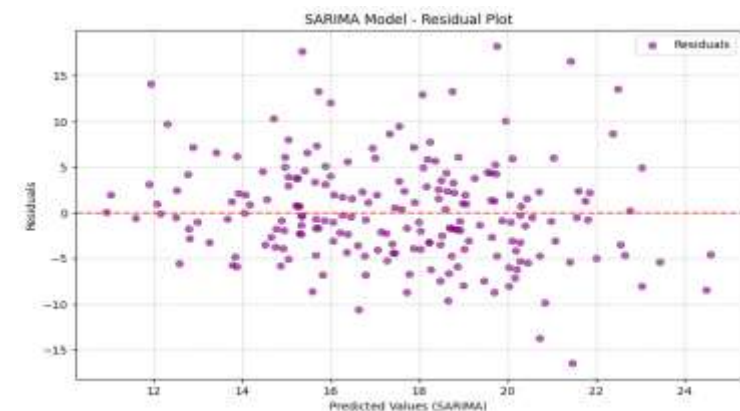
# Model 4: SARIMA (Seasonal ARIMA)



- Various combinations of seasonal and non-seasonal parameters ( $p, d, q$ ) and ( $P, D, Q$ ) were tested to find the optimal configuration. The best-performing model was identified with **parameters ( $p=2, d=1, q=2$ ), ( $P=2, D=0, Q=2, s=12$ )**, yielding an **RMSE of 5.5235**. This fine-tuning process helped improve the model's accuracy.
- The SARIMA model successfully captures daily fluctuations and seasonal patterns, providing accurate forecasts with minor deviations between predicted and actual values.



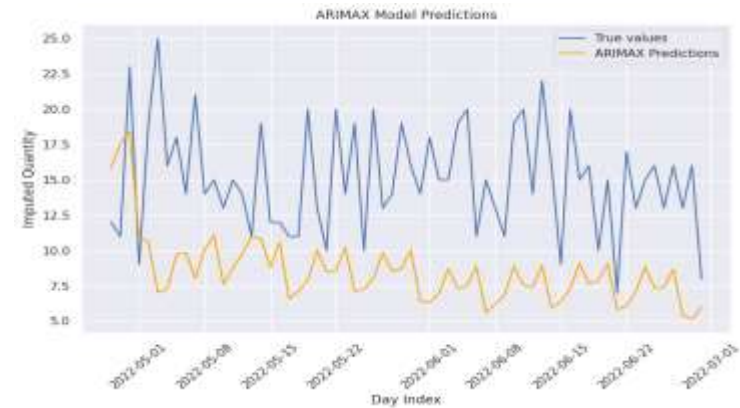
The SARIMA model excels in capturing longer-term trends and seasonality in the weekly data, showing good alignment with actual values, though small discrepancies appear during specific periods.



The residual plot indicates that the model effectively captures most patterns with minimal systematic bias. There are no significant outliers, suggesting that the model has adequately explained the variance in the data.



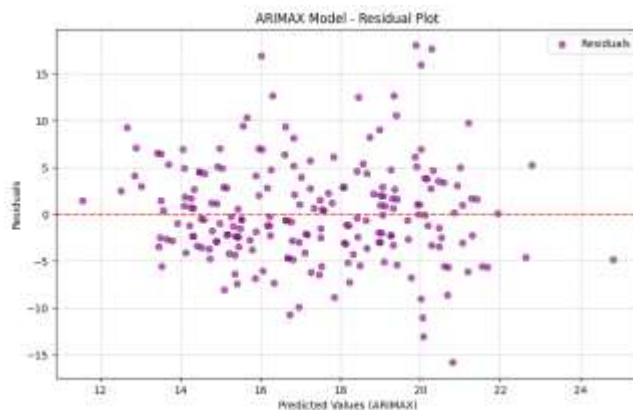
# Model 5: ARIMAX (ARIMA with Exogenous Variables)



- After testing various  $(p, d, q)$  combinations, the best parameters were  $(2, 1, 2)$  with an RMSE of 5.3600, resulting in the most accurate predictions. The SARIMA model successfully captures daily fluctuations and seasonal patterns, providing accurate forecasts with minor deviations between predicted and actual values.
- The ARIMAX model closely tracks daily trends, though minor deviations appear during sharp fluctuations. The addition of exogenous variables helps the model adjust to these variations.



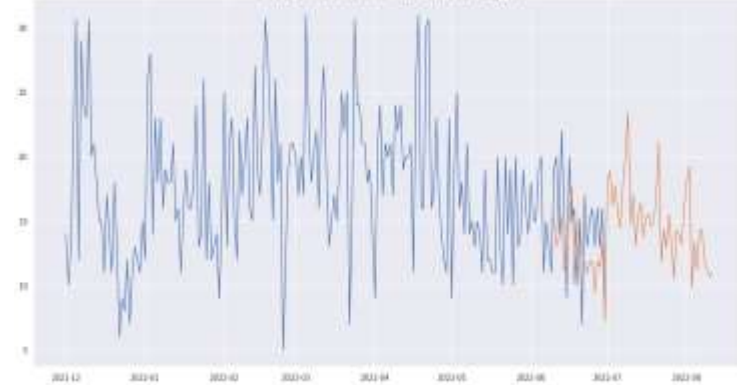
Weekly aggregation analysis shows that ARIMAX successfully captures larger trends, even in the presence of external factors influencing the target variable.



The residuals plot shows that the model has minimal systematic bias, indicating that it fits the data well with no significant error patterns.

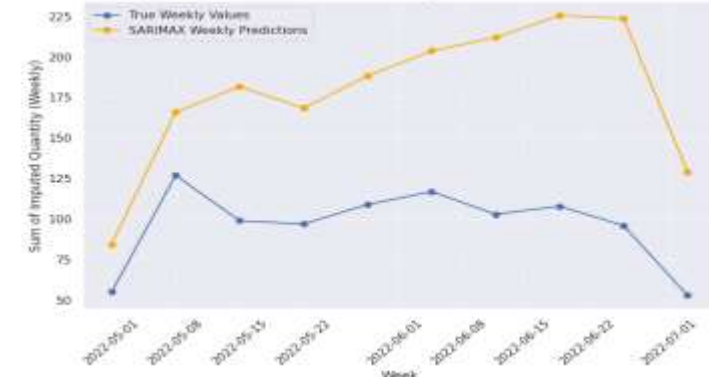
# Model 6: SARIMAX (Seasonal ARIMAX)

Future Forecast for next 6 weeks(Daily)



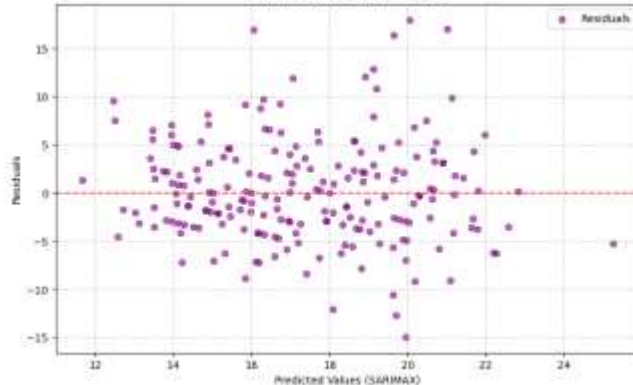
- After extensive testing, the best SARIMAX parameters were identified as (2, 1, 1, 2, 0, 1) with an RMSE of 5.3326. The ARIMAX model closely tracks daily trends, though minor deviations appear during sharp fluctuations. The addition of exogenous variables helps the model adjust to these variations.
- After tuning, the SARIMAX model was refitted, and predictions for daily values were generated and compared to actual data. The SARIMAX model closely tracks daily fluctuations, with a minor lag during certain peak periods.

SARIMAX Model Weekly Predictions



The weekly analysis involved aggregating both actual and predicted values to assess the model's performance over longer periods. The SARIMAX model successfully captured broader trends and seasonal variations, although it showed some limitations in predicting major spikes accurately.

SARIMAX Model - Residual Plot



Residuals were calculated by subtracting the predicted values from the actual values and plotted to assess the model's accuracy. The analysis revealed minimal systematic errors and no significant biases, highlighting the robustness of the model.

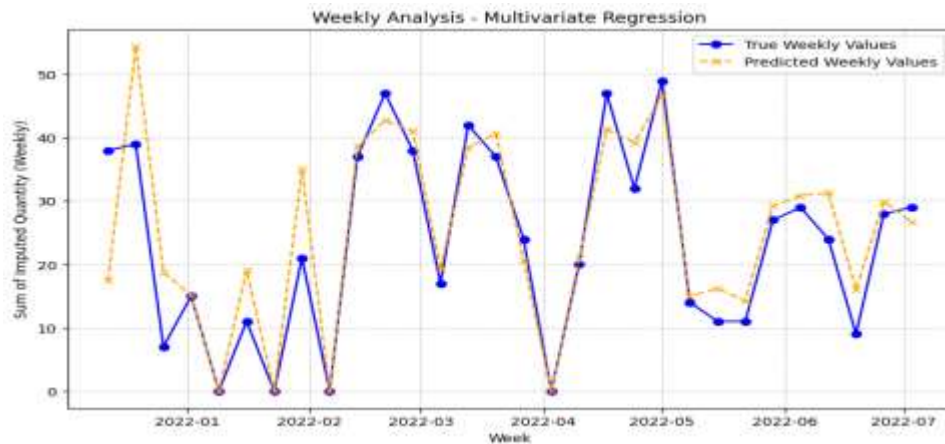
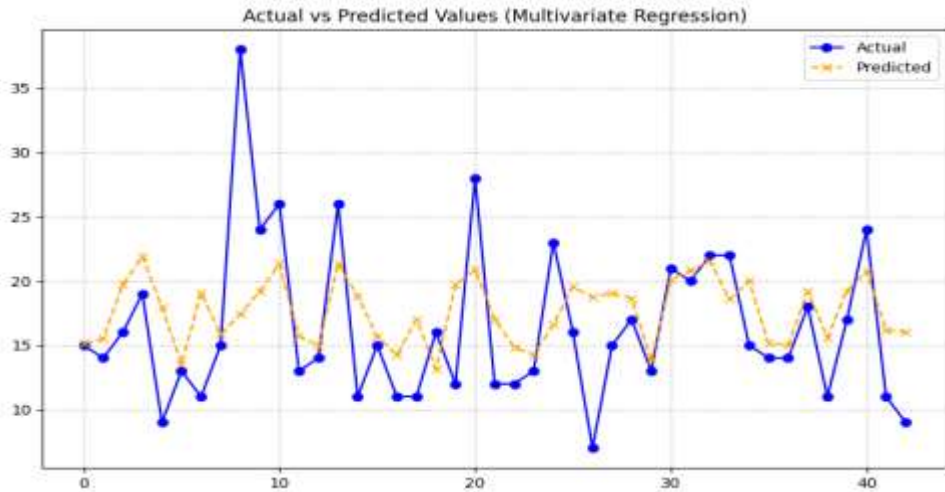
# Insights and Conclusion

After evaluating all the models, the **SARIMAX model** emerged as the most suitable choice for this project:

- **Seasonality Capture:** It effectively captures seasonality, essential for predicting recurring demand patterns.
- **Incorporation of Exogenous Variables:** It integrates key external factors like **Clicks** and **Impressions**, which influence the target variable (Quantity).
- **Superior Accuracy:** Compared to simpler models (AR, MA), SARIMAX consistently showed higher accuracy and a more comprehensive data understanding.
- **Conclusion:** SARIMAX meets the project's forecasting needs, making it the recommended approach for reliable time series analysis and demand prediction.

# Final Step: Splitting Data and Running Multivariate Regression with Visualization

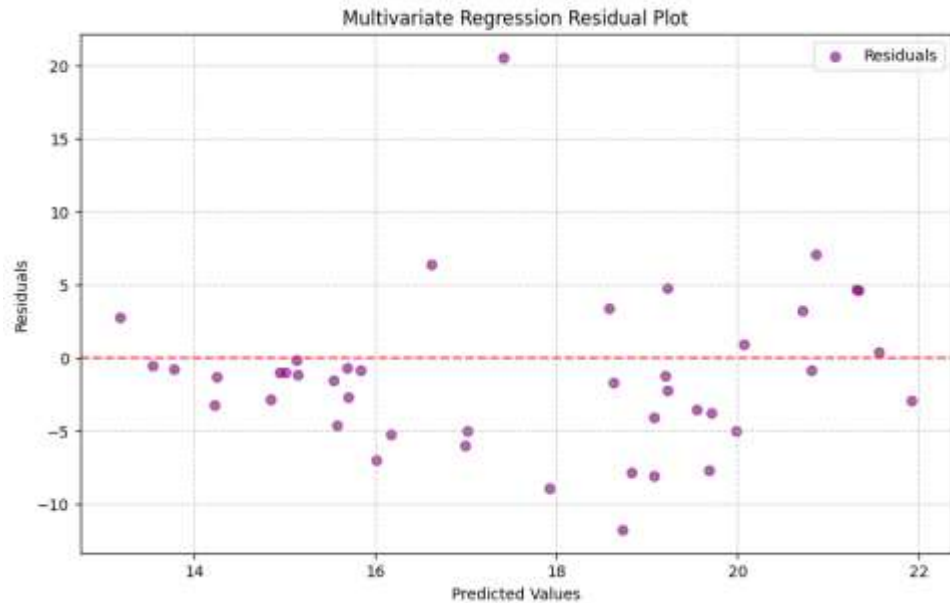
**Hyperparameter Tuning:** The multivariate regression model's hyperparameters were fine-tuned using GridSearchCV to optimize parameters such as `fit_intercept` and `copy_X`. After testing multiple configurations, the best model was identified with parameters `{'copy_X': True, 'fit_intercept': True}`. The evaluation metrics showed an RMSE of 30.13 and an R-squared value of 0.17.



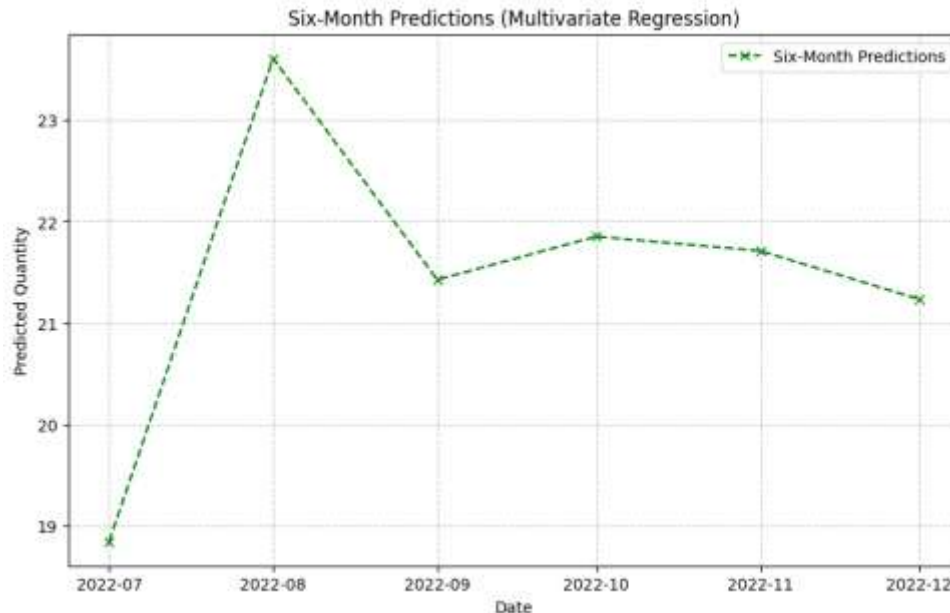
The multivariate regression model was used to predict daily values based on the tuned parameters. Key Insight: The model captures daily fluctuations but struggles with significant spikes, leading to slight discrepancies between predicted and actual values.

Residuals were calculated by subtracting the predicted values from the actual values and plotted to assess the model's accuracy. The analysis revealed minimal systematic errors and no significant biases, highlighting the robustness of the model.

## Final Step (Continued)



**Residual Analysis:** The residuals plot was used to check for randomness between actual and predicted values. **Key Insight:** Residuals show minor biases, but no significant outliers, suggesting that the model largely captures patterns.



The model was used to generate predictions for the next six months based on simulated future values of exogenous variables (Clicks and Impressions). **Key Insight:** The model forecasts a significant increase in the predicted quantity during the initial months, followed by a stabilization. While the trend generally follows expected patterns, further refinement is needed for more precise forecasting.