Project Name – Demand Forecasting for E-Commerce

Week 1 Deliverables

Hypothesis List –

1. **Stationarity:** The time series data for sales, Google clicks, and Facebook impressions is stationary (i.e., mean and variance do not change over time)

ADF Test

```
def adfuller_test(series):
    result = adfuller(series, autolag='AIC')
    print('ADF Statistic:', result[0])
    print('p-value:', result[1])
    print('Critical Values:')
    for key, value in result[4].items():
        print(f'\t{key}: {value}')
    print("-"*50)

# Perform Dickey-Fuller test on the Sales data
adfuller_test(df['Quantity'])
adfuller_test(df['Clicks'])
adfuller_test(df['Impressions'])
```

✓ 0.0s

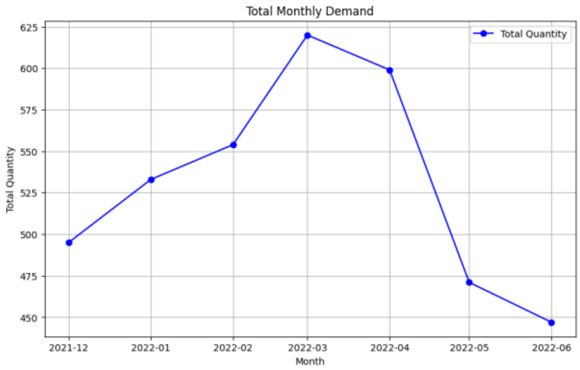
```
ADF Statistic: -4.445717448758575
p-value: 0.00024615679644112006
Critical Values:
	1%: -3.4621857592784546
	5%: -2.875537986778846
	10%: -2.574231080806213
--------------------------------------------------
ADF Statistic: -0.8705717270828215
p-value: 0.797509057499528
Critical Values:
	1%: -3.4620315036789666
	5%: -2.8754705024827127
	10%: -2.5741950726860647
--------------------------------------------------
ADF Statistic: -5.6962061101258685
p-value: 7.863580125889645e-07
Critical Values:
	1%: -3.46172743446274
	5%: -2.8753374677799957
	10%: -2.574124089081557
--------------------------------------------------
```

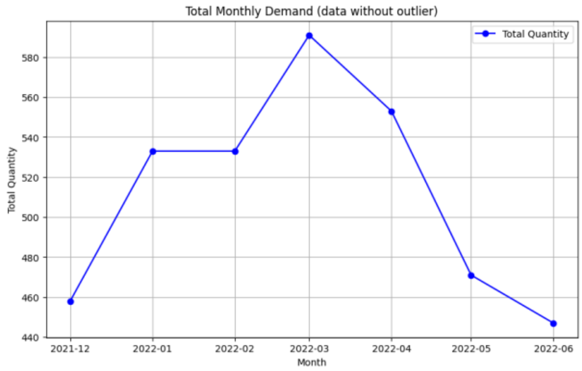
p-value is less than 0.05 (commonly used threshold), so reject the null hypothesis that the series has a unit root (i.e., the series is non-stationary). Hence, concluding that the series is stationary for Quantity.

2. **Trend:** There is a significant upward or downward trend in the sales data over time. This could indicate a growing or declining market.

   Plotting Monthly Sales Data:

   - The monthly Sales are added for each month
   - Then, Plot the Sales data over time.
   - **Formatting**:
     - Formats the date on the x-axis for better readability.
     - Sets titles, labels, and legend for the plot.



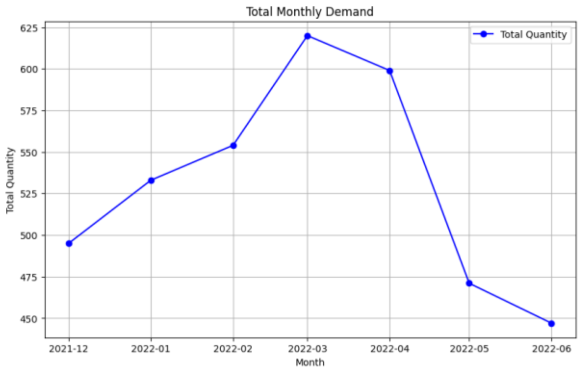


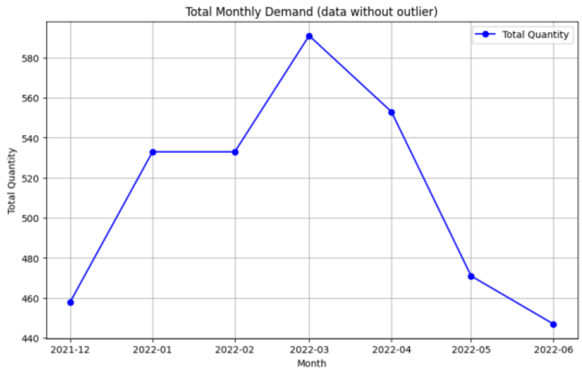
The total quantity demand of Product A significantly increases (projecting upward trend) until march and then declines rapidly (projecting downward trend).

3. **Seasonality:** The sales data exhibits seasonality (e.g., weekly, monthly, or yearly patterns). This is common in retail data due to holidays, promotions, and other recurring events.

   Plotting Monthly Sales Data:

   - The monthly Sales are added for each month
   - Then, Plot the Sales data over time.
   - **Formatting**:
     - Formats the date on the x-axis for better readability.
     - Sets titles, labels, and legend for the plot.

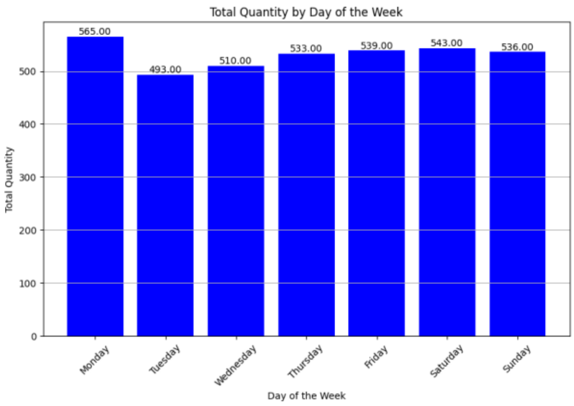


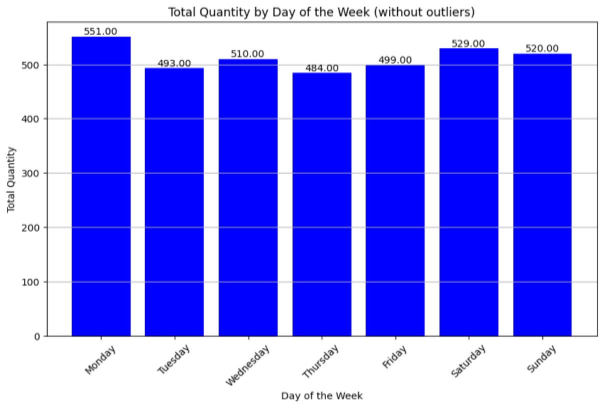


The quantity of Product A increases as March arrives and decreases after march.

Plotting days in a weekly basis: Sales Data:

- The Sales are added for each day in a week.
- Then, Plot the Sales data over days in a week.
- **Formatting**:
  - Formats the days on the x-axis for better readability.
  - Sets titles, labels, and legend for the plot.

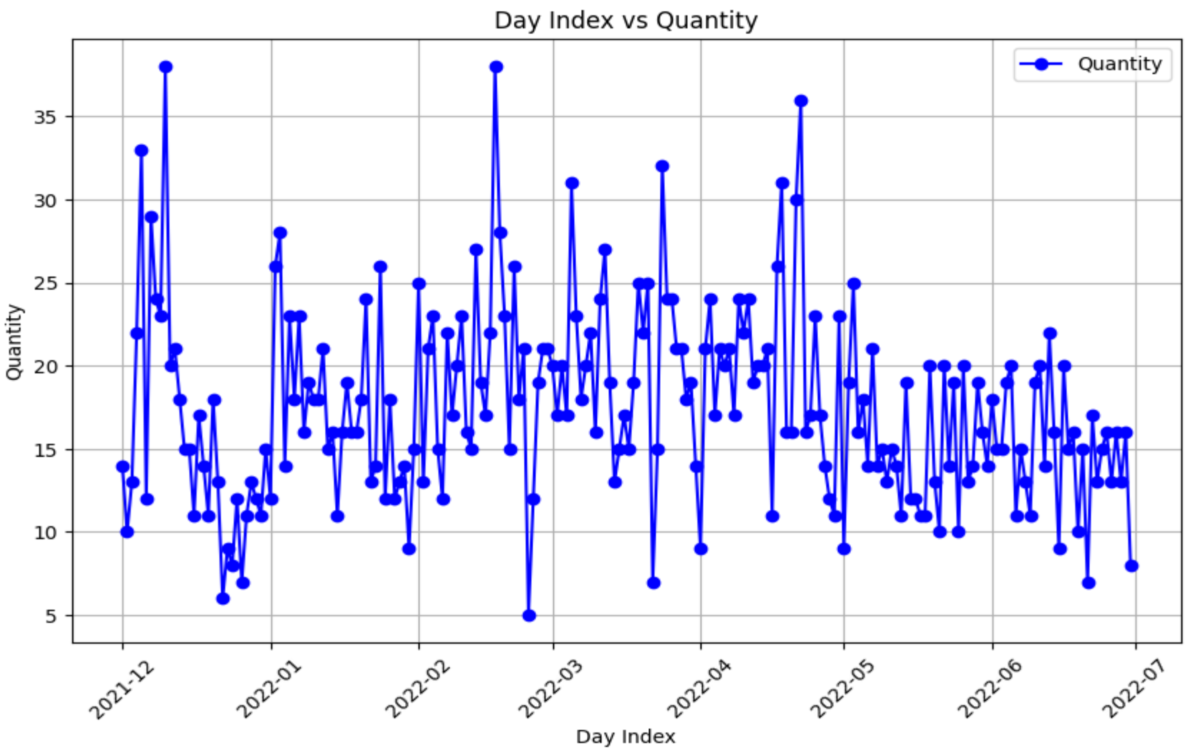




The Quantity demand is highest on Monday.

Plotting Daily Sales Data:

- Plot the Sales data over time for each day.
- **Formatting**:
  - Formats the date on the x-axis for better readability.
  - Sets titles, labels, and legend for the plot.

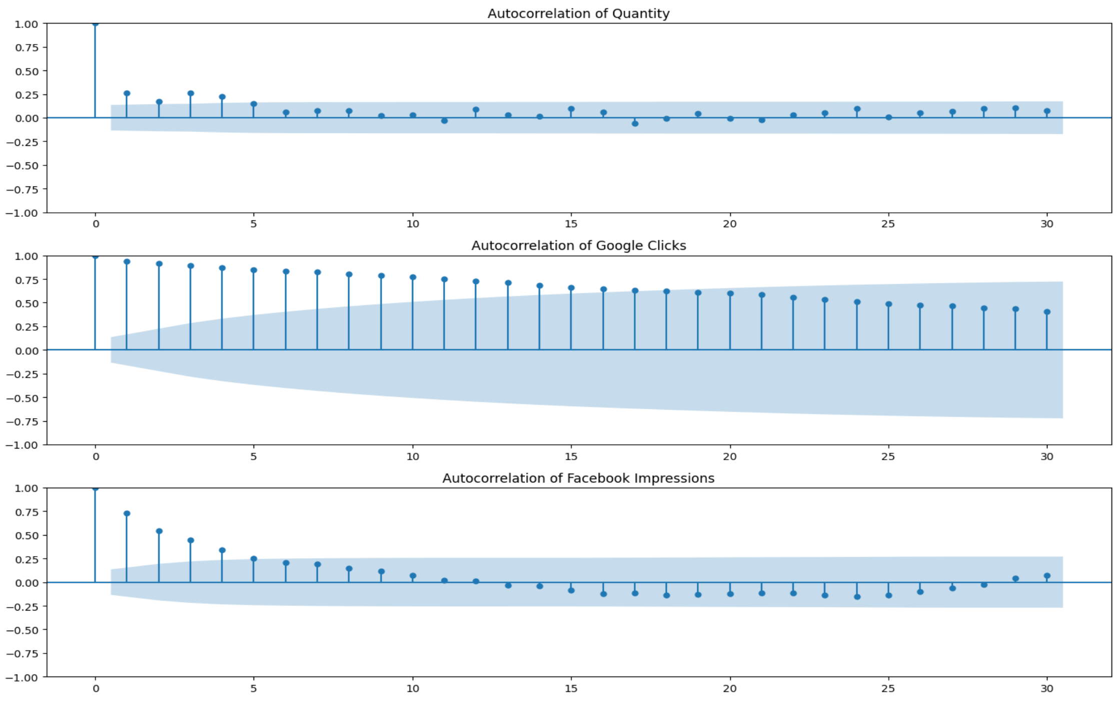


The days of the month do not have any relation to Product A Quantity.

4. **Autocorrelation:** There is a significant correlation between past and future values of sales, clicks, or impressions. This can be used to build autoregressive models (ARIMA).

Plotting Autocorrelation Functions (ACF):

- **Sales ACF**:
  - Create a subplot for Sales.
  - Plot the autocorrelation function (ACF) for Sales.
  - Set the title for the Sales ACF plot.
- **Google Clicks ACF**:
  - Create a subplot for Clicks.
  - Plot the ACF for Clicks.
  - Set the title for the Clicks ACF plot.
- **Facebook Impressions ACF**:
  - Create a subplot for Impressions.
  - Plot the ACF for Impressions.
  - Set the title for the Impressions ACF plot.
- **Display the Plots**:
  - Adjust subplot parameters for better visualization.
  - Show the plots.

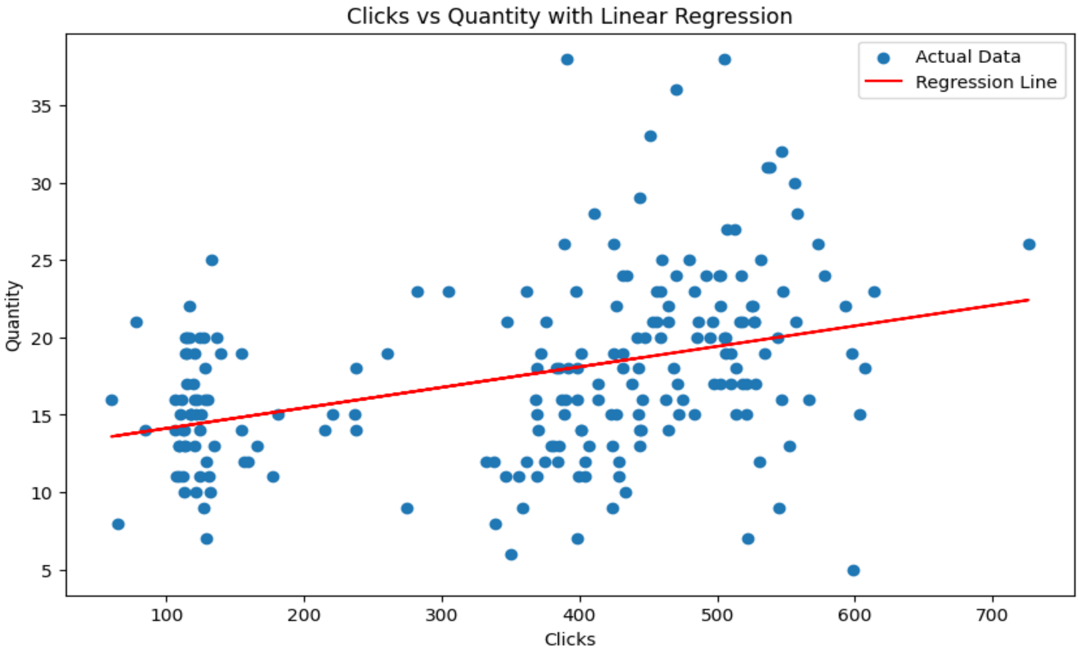


The correlation values lie between 0.25 to -0.25 showing no spikes, stating that there is no autocorrelation between past and future values of “Quantity” attribute of data.Meanwhile, there is a higher autocorrelation between past and future values in terms of “Clicks” and “Impressions” attributes of the data.

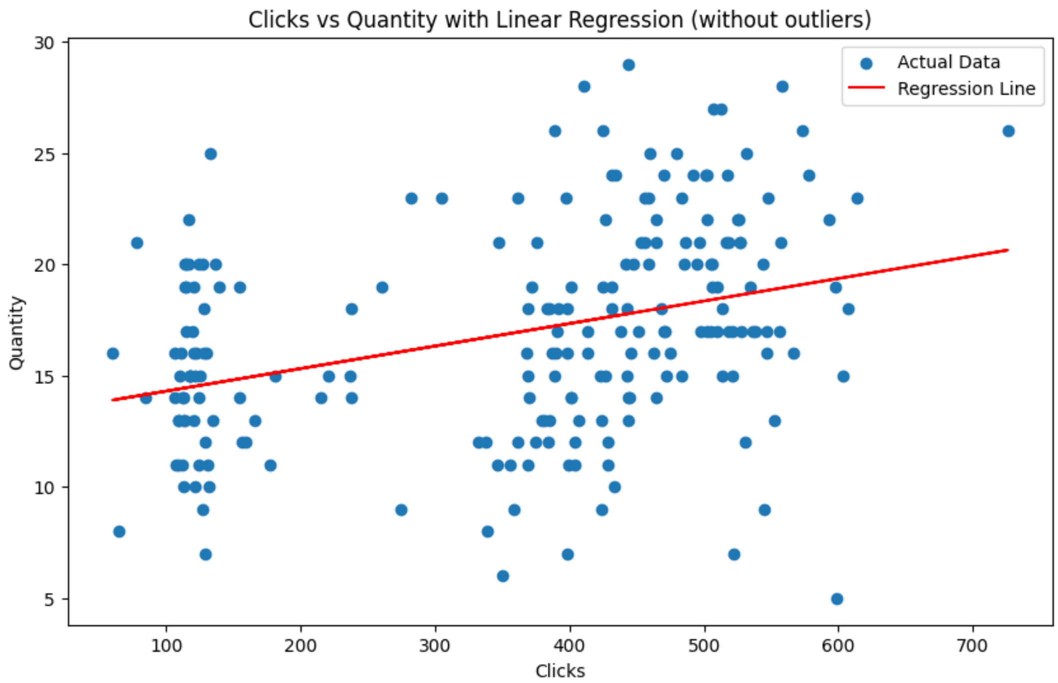
5. **Google Clicks Effect:** An increase in Google clicks is associated with an increase in sales. This tests the effectiveness of online advertising.

Linear Regression Analysis:

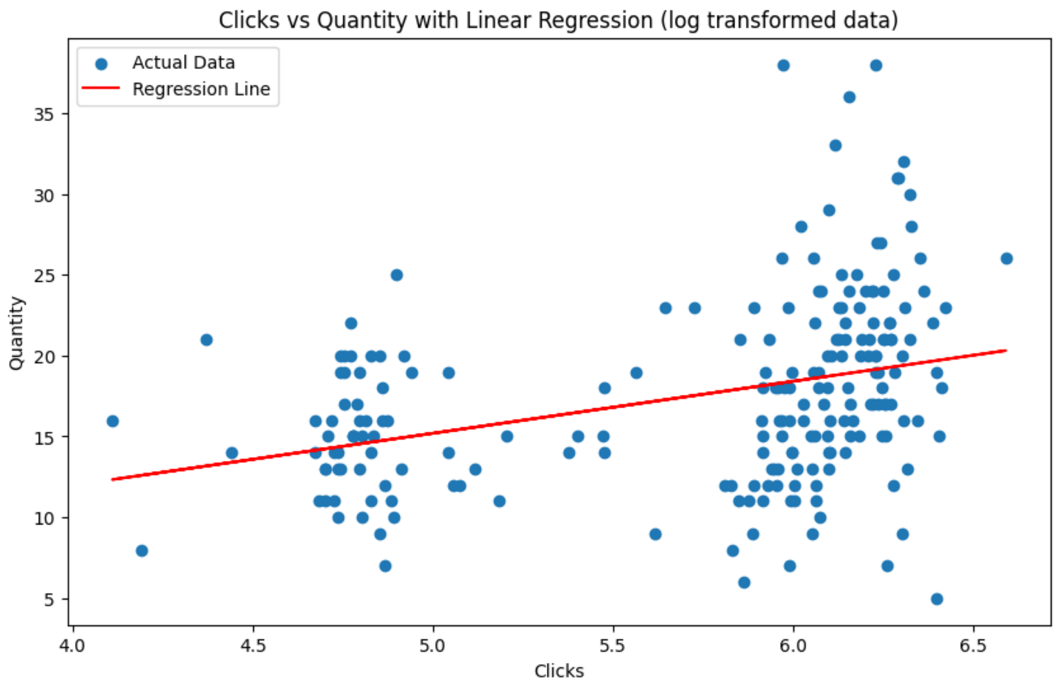
- **Import Linear Regression Model**:
  - Import the Linear Regression model from the scikit-learn library.
- **Prepare Data**:
  - Reshape the independent variable Clicks to a 2D array.
  - Assign the dependent variable Quantity to y.
- **Create and Fit Model**:
  - Create a Linear Regression model object.
  - Fit the model using the independent variable Clicks and the dependent variable Quantity.
- **Make Predictions**:
  - Predict Quantity values based on Clicks using the trained model.
- **Plot Results**:
  - Create a scatter plot of Clicks vs Quantity to visualize the actual data points.
  - Plot the regression line based on the predicted Quantity values.
  - Add titles, labels, and a legend to the plot.
  - Show the plot.
- **Calculate Slope**:
  - Retrieve the slope of the linear regression model.



The slope of the linear regression model is: 0.013234961483654805



The slope of the linear regression model is: 0.010120031987426179



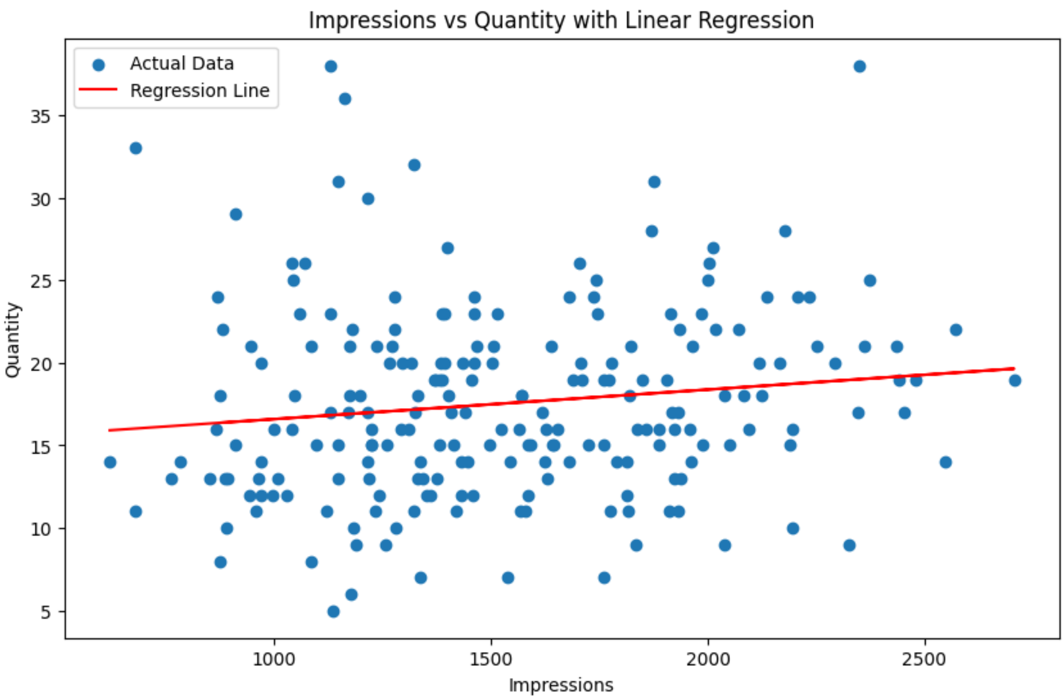
The slope of the linear regression model is: 3.2186587601137284

Hence, there is a relation with positive slope, i.e. Quantity increases as Clicks increases.

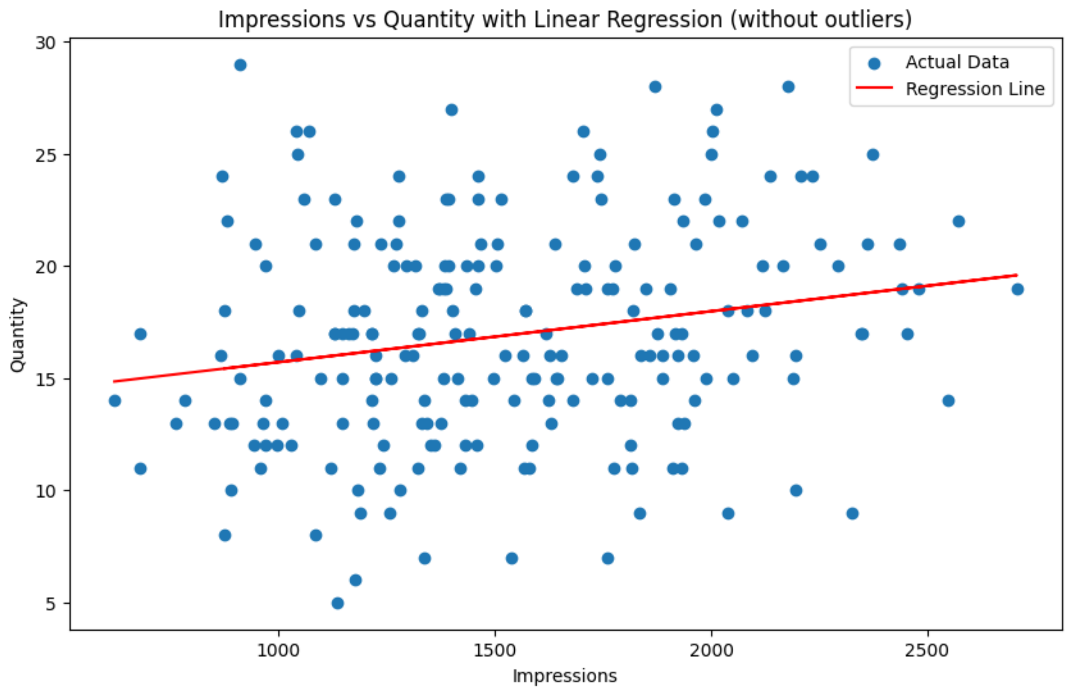
6. **Facebook Impressions Effect:** An increase in Facebook impressions is associated with an increase in sales.

Linear Regression Analysis:

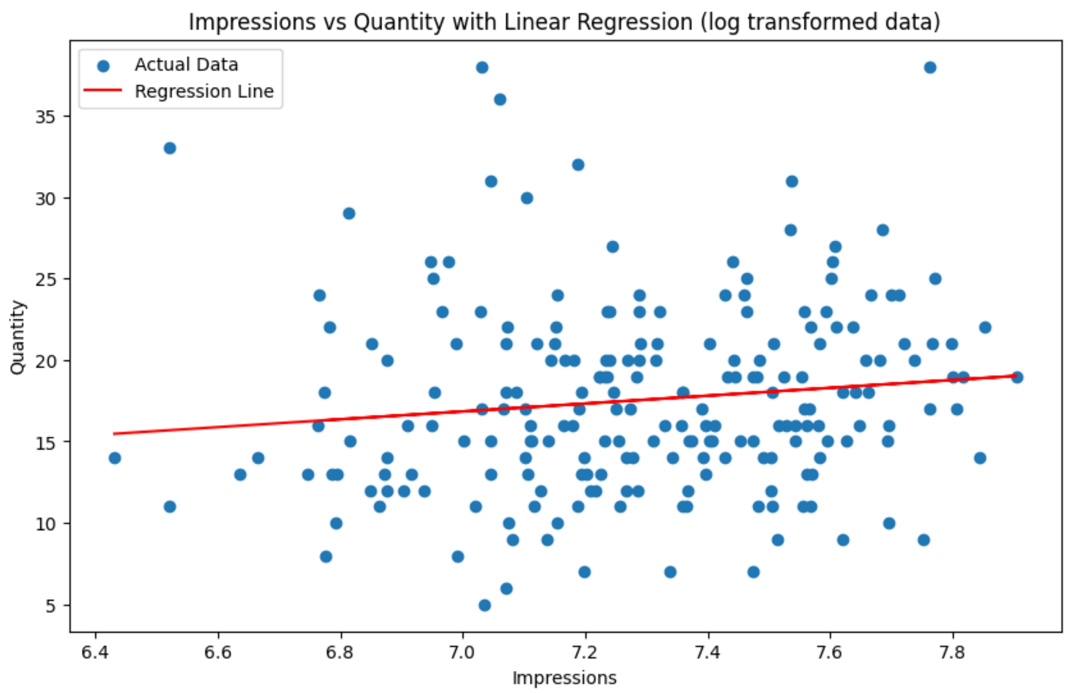
- **Import Linear Regression Model**:
  - Import the Linear Regression model from the scikit-learn library.
- **Prepare Data**:
  - Reshape the independent variable Impressions to a 2D array.
  - Assign the dependent variable Quantity to y.
- **Create and Fit Model**:
  - Create a Linear Regression model object.
  - Fit the model using the independent variable Impressions and the dependent variable Quantity.
- **Make Predictions**:
  - Predict Quantity values based on Impressions using the trained model.
- **Plot Results**:
  - Create a scatter plot of Impressions vs Quantity to visualize the actual data points.
  - Plot the regression line based on the predicted Quantity values.
  - Add titles, labels, and a legend to the plot.
  - Show the plot.
- **Calculate Slope**:
  - Retrieve the slope of the linear regression model.



The slope of the linear regression model is: 0.0017863293104571235



The slope of the linear regression model is: 0.002268934260063536



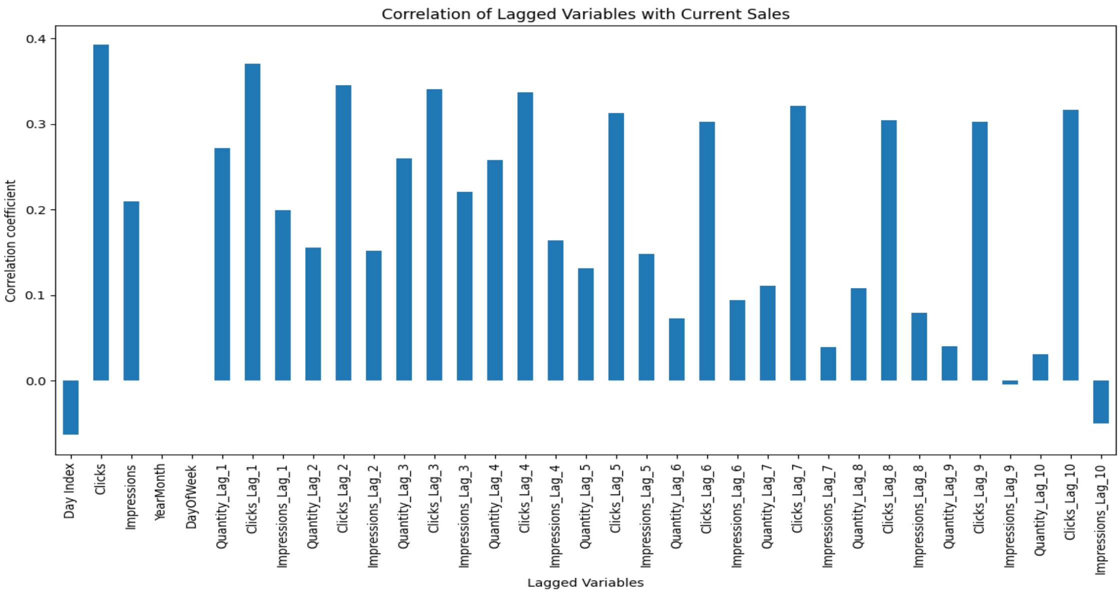
The slope of the linear regression model is: 2.409679402676977

Hence, there is a positive slope that means an increase in impressions can slightly increase the quantity demanded.

7. **Lagged Effects:** Past values of sales, clicks, or impressions have a significant impact on current sales. This could indicate a delayed response to marketing efforts or a carryover effect from previous periods.

   Analyzing Lagged Variables Correlation with Current Sales:

   - Create Lagged Variables:
     - Generate lagged versions of Sales, Clicks, and Impressions up to 10 lags.
     - Shift the original data by a specified number of periods to create lagged variables.
   - Correlation Plot:
     - Plot correlation coefficients of lagged variables with current Sales.
     - Visualize correlation strength and direction using a bar plot.
     - X-axis represents lagged variables, while Y-axis indicates correlation coefficient.
   - Interpretation:
     - Positive correlation implies higher lagged variable values correlate with higher current Sales, and vice versa for negative correlation.
     - Identify significant lagged variables impacting current Sales, aiding understanding of temporal relationship between past activities and current sales performance.
   - Correlation Values:
     - Display correlation values for each lagged variable with current Sales to enhance understanding.



- **Strong Positive Correlation**:
  - There is a moderate to strong positive correlation between the current Quantity and the lagged variables (Quantity_Lag_1 to Quantity_Lag_4), with correlation coefficients ranging from 0.155 to

0.271. This suggests that past values of Quantity have a moderate impact on the current Quantity.

- **Positive Correlation**:
  - The current Quantity also shows a positive correlation with the lagged variables for Clicks (Clicks_Lag_1 to Clicks_Lag_10) and Impressions (Impressions_Lag_1 to Impressions_Lag_7), indicating that past values of Clicks and Impressions are positively associated with the current Quantity, though to a lesser extent compared to lagged Quantity values.
- **Weak Correlation**:
  - For lagged variables beyond a certain lag (e.g., beyond Lag_7), the correlation coefficients start to decrease, indicating a weaker association with the current Quantity.
- **Negative Correlation**:
  - Some of the lagged variables (Impressions_Lag_9 and Impressions_Lag_10) exhibit negative correlation coefficients, indicating a weak negative association with the current Quantity. However, the magnitudes of these coefficients are relatively small, suggesting a weaker impact compared to the positive correlations observed with other lagged variables.

Overall, this analysis provides insights into the temporal relationship between past activities (represented by lagged variables) and the current Quantity. It suggests that past values of Quantity, Clicks, and Impressions have some influence on the current Quantity, with stronger associations observed for more recent lagged values.