# **Baacumen Data Science and Analytics Bootcamp**

# **Statistics for Data Analytics Capstone Project**

# **Group 2 Members**

1. Eba Fekadu
2. Abenezer Mitiku
3. Mikiyas Abebe
4. Kumilachew Taye
5. Amanuel Mintesnot
6. Mekurianeh Tilahun
7. Yonas Negese
8. Leul Mintesinot

Submission Date - October 11, 2025

**Table Of Contents**

[1. Data Preparation and Quality](#_quz7v9qh4e17) 3

[Handling Missing Values and Columns](#_s6cjhbw9j7aa) 3

[Identifying and Addressing Outliers](#_84cy3j96j5fd) 4

[2. Descriptive Statistics and Exploratory Analysis](#_g4wxfjqx8kgl) 5

[3. Probability and Hypothesis Testing](#_9ujvqcr4l2ax) 11

[4. Confidence Intervals](#_unyl40e8mych) 11

[5. Correlation and Regression Analysis](#_ydbr92uq7fcs) 11

[6. Time Series and Seasonality](#_smi66ar9kgxa) 12

[7. Business Implications and Recommendations](#_yuxlnbtgiwsl) 13

[8. Conclusion](#_o6o6xg41cf47) 14

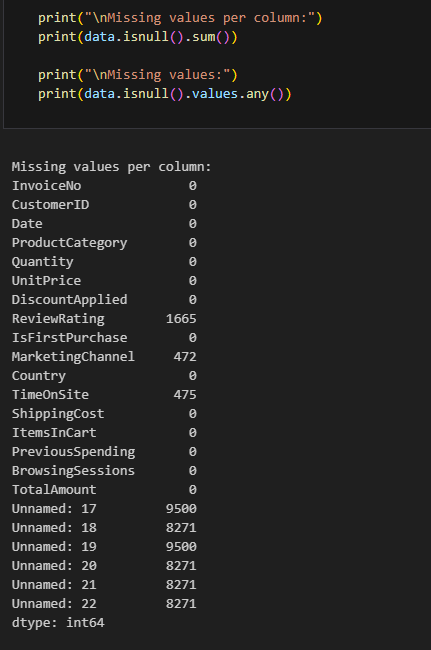
# **Interpretive Commentary & Business Implications**

## **E-Commerce Sales Analytics Capstone Group 2**

## **Data Preparation and Quality**

### **Handling Missing Values and Columns**

The initial phase of the analysis focused on ensuring the integrity and usability of the dataset. Unnecessary columns were removed, and missing values were addressed systematically. Using EDA, the dataset’s columns which have missing values were determined.

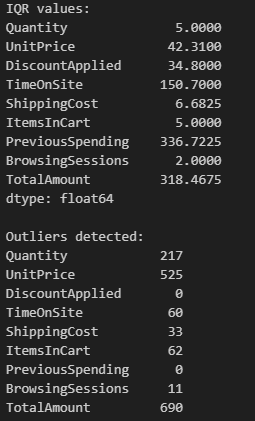


After observing the missing values of columns, it can be seen that there are unnamed columns in the dataset. Since there is no way of knowing what these columns represent, they can not be used in the analysis and therefore must be removed from the dataset.

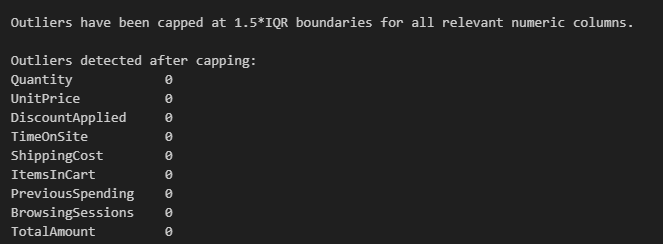
For categorical variables such as ReviewRating and MarketingChannel, missing entries were imputed using the mode, ensuring that the most common values filled the gaps. For the continuous variable TimeOnSite, the mean was used for imputation, preserving the overall distribution.

### **Identifying and Addressing Outliers**

Outliers in key numeric columns, including Quantity and UnitPrice, were detected using the interquartile range (IQR) method.



The outliers found were capped at 1.5 times the IQR. This approach minimized the influence of extreme values without discarding potentially valuable data. As a result, the cleaned dataset provided a robust foundation for subsequent statistical analysis and modeling.



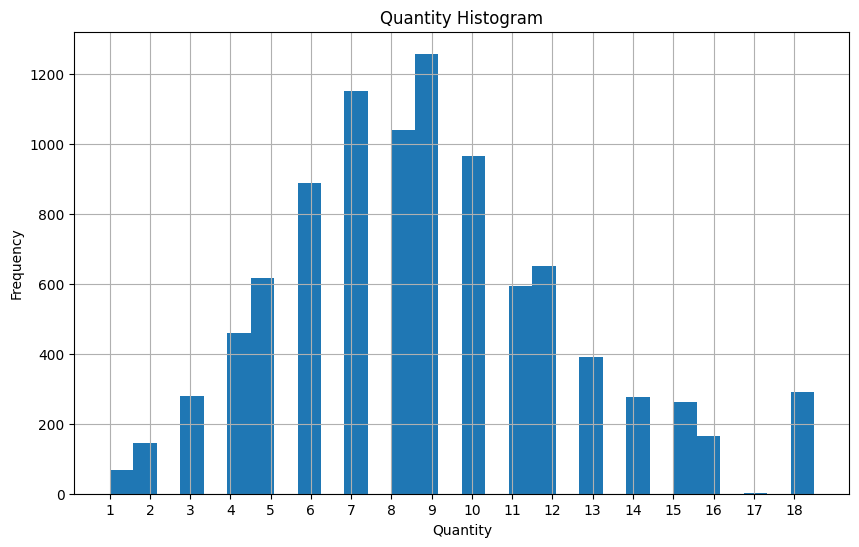
## **Descriptive Statistics and Exploratory Analysis**

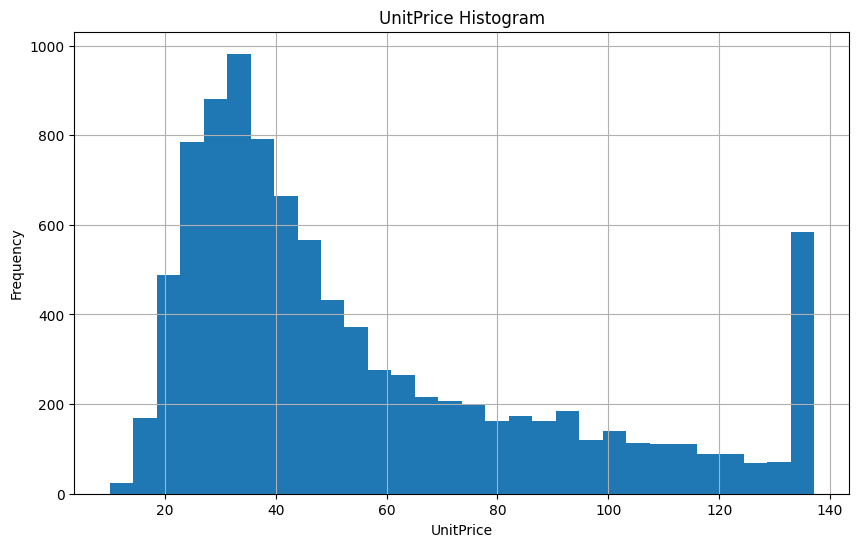
Descriptive statistics revealed important characteristics of customer purchasing behavior. The mean, median, and mode for variables such as Quantity, UnitPrice, and TotalAmount indicated that most transactions involved a modest number of items and moderate prices, but the presence of a right-skewed distribution in TotalAmount and UnitPrice suggested occasional high-value purchases. The standard deviation and IQR values confirmed substantial variability, particularly in spending and item quantities. After outlier capping, the range of these variables became more representative of typical customer behavior.

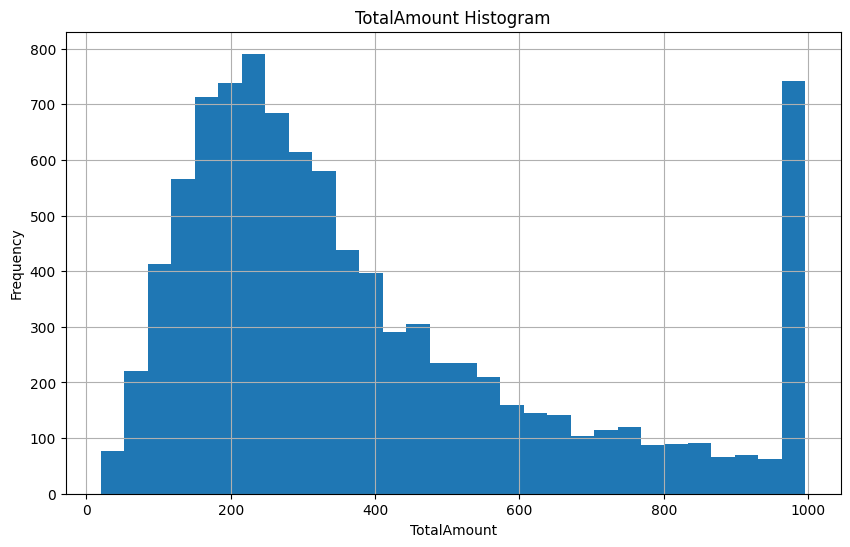
| Variables | count | mean | std | min | 25% | 50% | 75% | max | mode | IQR | variance | range |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Quantity | 9500 | 8.842578947 | 3.578394696 | 1 | 6 | 9 | 11 | 18.5 | 9 | 5 | 12.8049086 | 17.5 |
| UnitPrice | 9500 | 56.33489105 | 33.78522945 | 10.06 | 31.35 | 43.675 | 73.66 | 137.125 | 137.125 | 42.31 | 1141.441729 | 127.065 |
| DiscountApplied | 9500 | 18.1814 | 18.53147467 | 0 | 0 | 13.6 | 34.8 | 55 | 0 | 34.8 | 343.4155536 | 55 |
| TimeOnSite | 9500 | 299.2955413 | 114.7594242 | 30 | 224.4 | 299.5678781 | 375.1 | 601.15 | 299.5678781 | 150.7 | 13169.72544 | 571.15 |
| ShippingCost | 9500 | 10.03552803 | 4.857920442 | 0 | 6.63 | 9.96 | 13.3125 | 23.33625 | 0 | 6.6825 | 23.59939102 | 23.33625 |
| ItemsInCart | 9500 | 9.661894737 | 3.498762898 | 1 | 7 | 9 | 12 | 19.5 | 9 | 5 | 12.24134181 | 18.5 |
| PreviousSpending | 9500 | 209.0902958 | 161.3187197 | 0 | 0 | 241.93 | 336.7225 | 632.02 | 0 | 336.7225 | 26023.72931 | 632.02 |
| BrowsingSessions | 9500 | 3.063578947 | 1.437016768 | 1 | 2 | 3 | 4 | 7 | 3 | 2 | 2.065017193 | 6 |
| TotalAmount | 9500 | 392.9106361 | 262.6932294 | 20.26 | 199.9175 | 309.135 | 518.385 | 996.08625 | 996.08625 | 318.4675 | 69007.73276 | 975.82625 |

Revenue analysis by product category showed that Electronics and Clothing were the leading contributors to total sales, highlighting these as strategic categories for marketing and inventory focus. When examining revenue by country, the USA and UK emerged as the top markets, suggesting that region-specific campaigns and logistics optimizations could yield significant returns. Analysis of average order value by marketing channel revealed that Email and Ads channels generated the highest order values, indicating that these channels are particularly effective for driving high-value transactions.

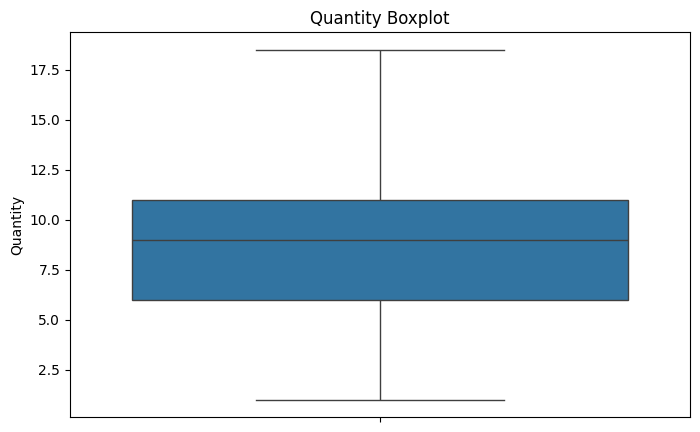
Visualizations such as histograms and boxplots illustrated the distributions and confirmed the effectiveness of outlier treatment. Some of the histograms include -

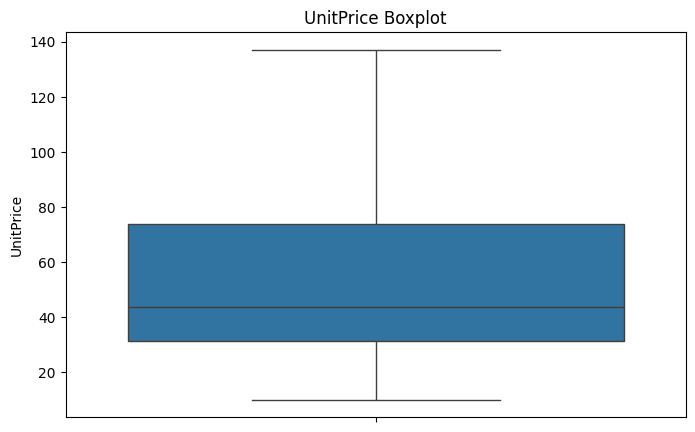


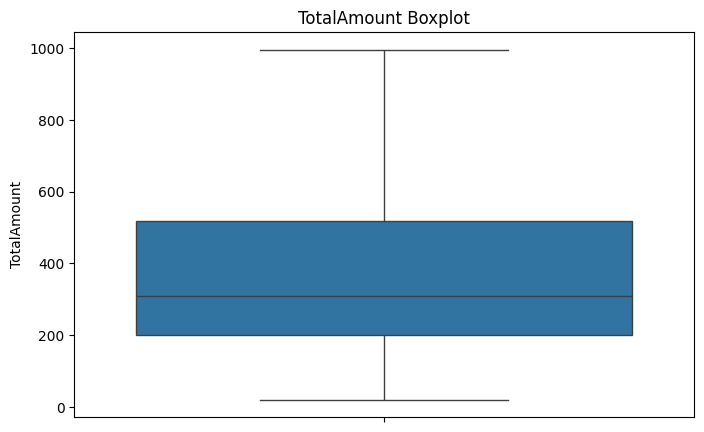




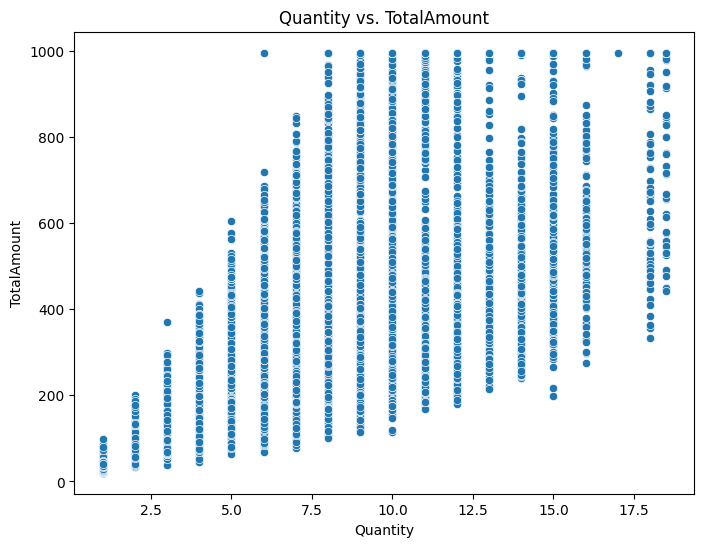
Some of the box plots(which are less relevant since outliers were already taken care of) include the following -

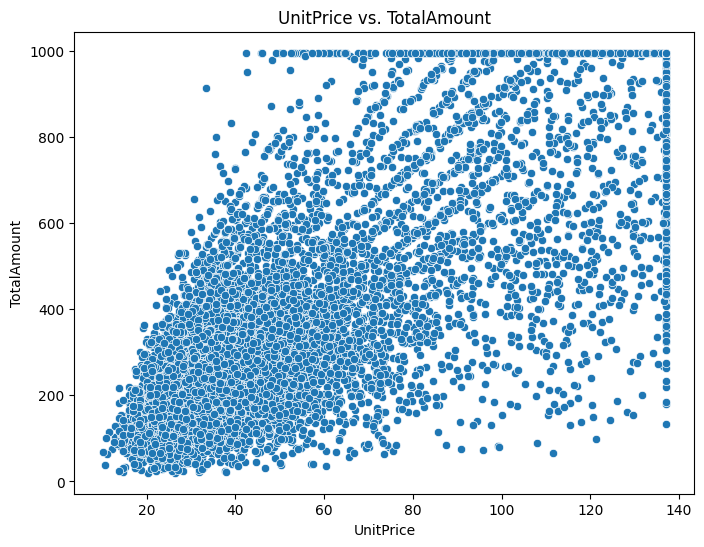


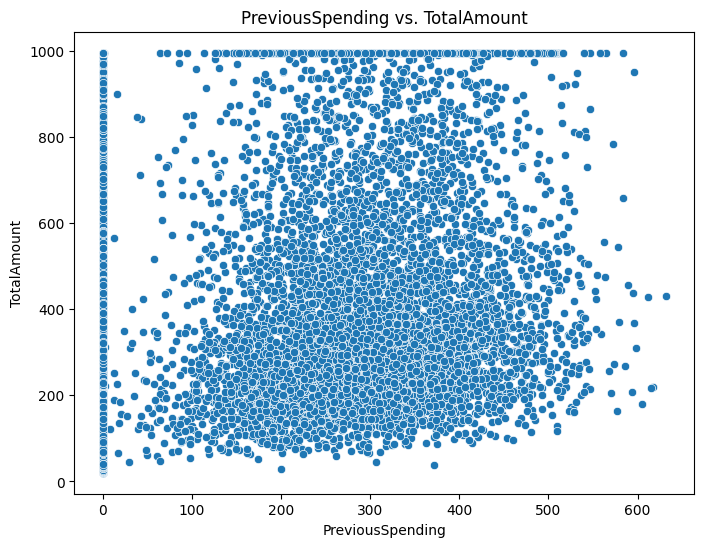




Scatterplots of predictors versus TotalAmount highlighted positive relationships between TotalAmount and variables like Quantity, UnitPrice, and PreviousSpending, reinforcing their importance as revenue drivers.







## **Probability and Hypothesis Testing**

On this dataset, probability tests were used to identify how likely a value is compared to others, which were given here as specified in the question. The probability of receiving a 5-star review was found to be substantial, with approximately 22% of all reviews at the highest rating. This suggests a generally positive customer experience, though there remains room for improvement. The probability of an order exceeding $1,000 was low, around 2%, confirming that while high-value transactions are rare, they do occur and can significantly impact revenue.

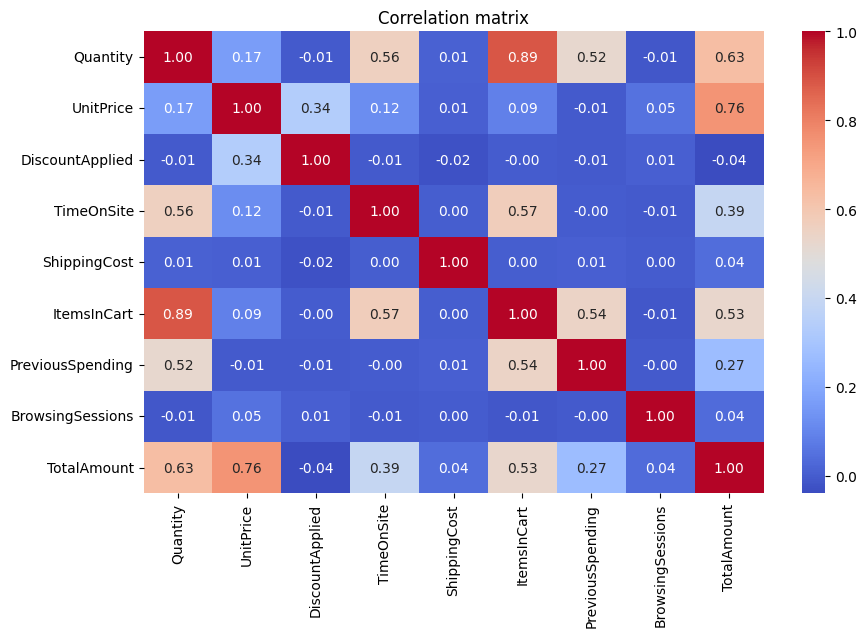
Hypothesis testing provided further actionable insights. A two-sample t-test comparing mean spending between first-time and returning customers revealed that returning customers spend significantly more per transaction. This underscores the value of customer retention strategies, such as loyalty programs and personalized offers. ANOVA results indicated statistically significant differences in average spending across countries, suggesting that localized pricing, promotions, or product assortments could further optimize revenue. The chi-square test for association between marketing channel and review rating was also significant, implying that certain channels are more likely to attract satisfied customers. This finding supports the strategic allocation of marketing resources to channels that not only drive sales but also foster positive customer experiences.

## **Confidence Intervals**

The 95% confidence interval for average daily revenue provided a reliable range for expected daily sales, supporting more accurate forecasting and resource planning. Similarly, the confidence interval for average review rating confirmed that customer satisfaction is stable, with the mean rating consistently above four stars. These intervals offer management a clear understanding of the variability and reliability of key business metrics, enabling more informed decision-making.

## **Correlation and Regression Analysis**

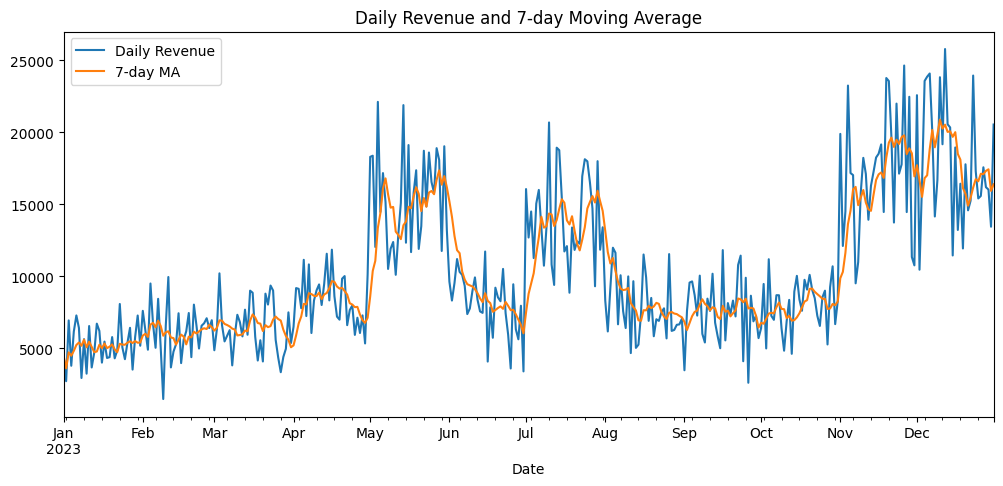
Correlation analysis revealed that TotalAmount is most strongly associated with Quantity, UnitPrice, and PreviousSpending. This finding was reinforced by multiple linear regression, where these variables emerged as the strongest numeric predictors of revenue. The regression model also identified significant effects from categorical variables such as ProductCategory, MarketingChannel, and Country, indicating that both customer behavior and business context play critical roles in determining transaction value.

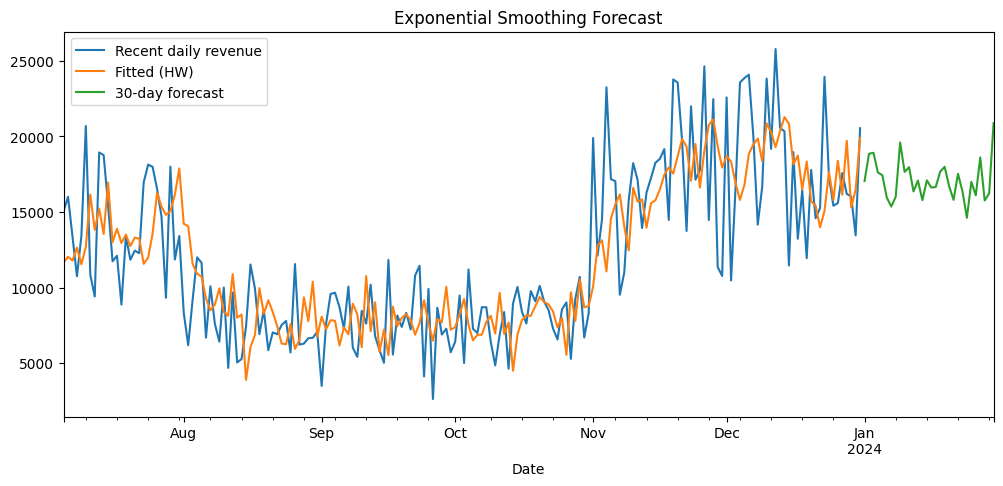


Standardized coefficients from the regression highlighted that increasing the quantity of items per order, raising unit prices (where justified by value), and targeting customers with a history of higher spending are the most effective levers for boosting revenue. The importance of categorical predictors suggests that optimizing product mix, refining marketing strategies by channel, and tailoring approaches to specific countries can further enhance performance. The regression summary and correlation heatmap provide a visual and statistical foundation for these recommendations.

## **Time Series and Seasonality**

Time series analysis of daily revenue revealed clear seasonal peaks in May, July, November, and December, aligning with common retail cycles such as summer sales and holiday shopping periods. Moving averages and exponential smoothing forecasts indicated that these seasonal trends are likely to persist, with anticipated spikes in revenue during these months.





The time series plots and forecast diagrams offer actionable insights for inventory planning, staffing, and promotional scheduling. By aligning operational and marketing efforts with these predictable peaks, the business can maximize sales and customer satisfaction during high-demand periods.

## **Business Implications and Recommendations**

The analysis demonstrates that customer retention is a key driver of revenue, as returning customers consistently spend more per transaction. Investments in loyalty programs, personalized marketing, and post-purchase engagement are likely to yield substantial returns. Channel optimization is also critical; focusing on Email and Ads channels, which generate higher order values and positive reviews, can improve both sales and customer satisfaction. Geographic analysis suggests that tailoring strategies to top-performing countries, particularly the USA and UK, will further enhance results. Seasonal planning is essential, with inventory and marketing resources concentrated around identified peak months. Finally, prioritizing high-revenue product categories such as Electronics and Clothing for promotions and stock allocation will ensure that the business capitalizes on its strongest offerings.

## **Conclusion**

This comprehensive analysis provides a data-driven roadmap for maximizing revenue, improving customer satisfaction, and optimizing business operations. By leveraging the insights from descriptive statistics, hypothesis testing, regression modeling, and time series forecasting, the company can make informed decisions that align with customer behavior and market dynamics. The inclusion of visualizations at key points in the report will further support these findings and facilitate communication with stakeholders.