"Sales Analytics: Exploring Online Retail Patterns for Business Growth"

Project submitted to the

SRM University – AP, Andhra Pradesh

for the fulfillment of the requirements to the end semester project of

Master of Technology

In

Computer Science and Engineering

School of Engineering and Sciences

Submitted by

Candidate Name

Shrishti Shiva (AP22122040009)

Deepali Kumari (AP22122040007)

Under the Guidance of **(Dr.**

Rajiv Senapati)

SRM University–AP

Neerukonda, Mangalagiri, Guntur

Andhra Pradesh – 522 240

[June, 2023]

Certificate

Date: 26-Jun-23

This is to certify that the work present in this Project entitled **"Sales Analytics: Exploring Online Retail Patterns for Business Growth"** has been carried out by **Deepali Kumari & Shrishti Shiva** under our supervision. The work is genuine, original, and suitable for submission to the SRM University – AP for the end semester project in Master of Technology in **School of Engineering and Sciences**.

Supervisor

Prof. / Dr. Rajiv Senapati

Assistant professor,

Department of computer science and engineering

Acknowledgements

I would like to express my sincere gratitude and appreciation to all those who have supported and contributed to the completion of this project.

First and foremost, I would like to thank our project supervisor Dr. Rajiv Senapati for their guidance, valuable insights, and continuous support throughout the project. Their expertise and feedback have been instrumental in shaping our analysis and enhancing the quality of our work.

I would also like to extend my gratitude to Kaggle for providing the Online Retail dataset used in this project. The availability of this dataset has been crucial in conducting our analysis and generating meaningful insights.

Additionally, I would like to thank Dr.Rajiv Senapati for providing the necessary resources and facilities that have enabled us to carry out this project effectively. Furthermore, I am grateful to my project partner Deepali Kumari/Shrishti Shiva for their collaboration, dedication, and teamwork throughout the project. Their contributions have been invaluable, and I appreciate their commitment to achieving our project objectives.

Finally, I would like to acknowledge the support and encouragement from our friends and family members who have provided motivation and understanding during the project's duration.

Overall, I am truly thankful to everyone who has played a part in this project, directly or indirectly. Your support, guidance, and contributions have been vital in making this project a success.

Table of Contents

Abstract

This project conducted a comprehensive analysis of sales data from an online retail dataset to gain insights into customer behavior and market trends. The data was preprocessed and stored in a data warehouse for efficient retrieval. Customer and product analyses were performed to understand purchasing patterns and identify top customers and best-selling products. Through various analyses, including customer analysis, product analysis, sales analysis, market basket analysis, time series forecasting, customer segmentation, and churn prediction, this project provides valuable information to understand sales patterns, customer behaviour, and potential business opportunities.

Sales trends and seasonal patterns were analyzed, and market basket analysis revealed product associations and cross-selling opportunities. Customer segmentation was achieved using clustering techniques. Data visualization techniques were employed to communicate the findings effectively.

The project's results provided valuable insights for improving sales performance, customer satisfaction, and marketing strategies. Further research opportunities include sentiment analysis and predictive modeling for sales forecasting and customer churn prediction. Overall, the project showcases the power of data analysis in the e-commerce industry.

This project equips businesses with valuable insights and actionable information to make informed decisions, optimize strategies, and drive revenue growth. By understanding customer behaviour, sales patterns, and potential churn, organizations can enhance customer satisfaction, streamline operations, and maximize profitability.

Abbreviations

MBA	Market Basket Analysis
DW	Data Warehouse
PM	Pattern Mining
ML	Machine Learning
RFM	Recency, Frequency, Monetary

List of Tables

List of Figures

Introduction

The rapid growth of e-commerce has transformed the way businesses operate and interact with customers. In this era of digital commerce, understanding customer behaviour and predicting revenue are crucial for the success of online retail businesses. This mini research paper explores various data analysis techniques and predictive models applied to a real-world online retail dataset to gain insights into customer purchasing patterns and forecast revenue. The study begins by preprocessing the dataset, which includes handling missing values, removing duplicates, and calculating the total revenue. With a clean and organized dataset, the analysis delves into identifying top customers based on their total revenue contribution and examining customer purchasing patterns over time.

Additionally, the research investigates the top-selling products, analysing their sales performance, and further explores product category sales to identify trends across different countries. The study also examines return rates or customer complaints, shedding light on customer satisfaction and product quality aspects of the business. To gain a comprehensive understanding of sales trends, the analysis focuses on analysing sales patterns over time and identifying seasonal trends in the online retail industry. Visualizations are utilized to illustrate the monthly sales trends and highlight the impact of various factors on revenue.

Furthermore, this research paper explores the application of market basket analysis to uncover associations and patterns among customer purchases, revealing insights into cross-selling and upselling opportunities. The findings from the market basket analysis contribute to optimizing product recommendations and enhancing the overall customer experience.

Moreover, this study demonstrates the utilization of time series forecasting techniques, specifically SARIMA modelling, to predict future sales. The accuracy of the forecast is evaluated using mean squared error, providing valuable insights for demand planning and inventory management. Customer segmentation is crucial in shaping marketing strategy since it helps marketers forecast sales as well as analyse consumer behaviour. K-means clustering is used in this study's customer segmentation to help firms find unique client groups that share common traits and target them with tailored marketing initiatives.

Project identifies valuable consumer groupings that may drive efficient marketing strategies and increase customer retention by giving RFM scores to customers based on their transaction history. In the context of an online retail firm, this research emphasises the significance of studying consumer behaviour and making revenue predictions. Businesses are enabled by prediction models and data analysis tools.

Dataset Description:

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/2010 8:26	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	12/1/2010 8:26	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/1/2010 8:26	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/1/2010 8:26	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12/1/2010 8:26	3.39	17850.0	United Kingdom

The dataset used in the project is an online retail dataset. It contains information about various transactions made by customers in an online retail business. The dataset includes the following columns:

- InvoiceNo: A unique identifier for each invoice or transaction.
- StockCode: The code or identifier for each product in the inventory.
- Description: The description or name of the product.
- Quantity: The quantity of each product purchased in a transaction.
- InvoiceDate: The date and time when the transaction was made.
- UnitPrice: The unit price of each product.
- CustomerID: The unique identifier for each customer.
- Country: The country where the customer is located

Methodology

Methodology starts from importing the necessary libraries.

- ***Data Collection:***

The research utilizes a real-world online retail dataset obtained from a specific source. The dataset includes information such as customer transactions, product details, invoice dates, quantities, and revenue.

- ***Data Preprocessing:***

The dataset undergoes preprocessing steps to ensure data quality and consistency. Missing values are identified and handled appropriately, either by imputation or by removing the corresponding records. Duplicate records are also removed to avoid skewing the analysis.

- ***Exploratory Data Analysis:***

The initial analysis involves gaining an overview of the dataset by examining its structure and content. Descriptive statistics and visualizations are used to understand the distribution of variables, identify outliers, and uncover any patterns or relationships.

- ***Customer Analysis:***

The research focuses on understanding customer behaviour and preferences. This involves identifying top customers based on their total revenue contribution and analysing their purchasing patterns over time. Customer segmentation techniques, such as K-means clustering or RFM analysis, may be applied to group customers with similar characteristics.

- ***Product Analysis:***

The study investigates product-related aspects, such as identifying top-selling products based on quantity sold and examining product performance across different categories. Market basket analysis techniques may be utilized to uncover associations and patterns among customer purchases.

- ***Sales Trends and Forecasting:***

Analysing sales trends over time is crucial for understanding seasonal patterns and identifying potential growth opportunities. Time series analysis techniques, such as SARIMA modelling, may be employed to forecast future sales. The accuracy of the forecast is evaluated using appropriate metrics, such as mean squared error.

- ***Churn Prediction:***

Predicting customer churn, or the likelihood of customers ceasing their relationship with the business, is vital for customer retention strategies. Predictive models, such as

logistic regression, can be employed to identify customers at risk of churn based on various customer attributes, such as total revenue, invoice frequency, and quantity.

- ***Evaluation and Interpretation:***

Throughout the analysis, appropriate evaluation techniques, such as metrics, statistical tests, and visualizations, are employed to assess the quality of the results and interpret the findings. The significance and implications of the findings are discussed in the context of the online retail business.

Steps Performed:

Import Libraries and Read Data:

1) This code cell imports necessary libraries such as pandas, datetime, matplotlib.pyplot, numpy, and seaborn.
2) It reads the 'OnlineRetail.csv' file using pd.read_csv() into a DataFrame called df.
3) The head() function is used to display the first few rows of the DataFrame.

Data Preprocessing:

1) This code cell performs data preprocessing tasks on the DataFrame df.
2) It checks for missing values using isnull().sum() and prints the total count of missing values for each column.
3) Rows with missing values are dropped using dropna().
4) Duplicate rows are removed using drop_duplicates().
5) The info() function is used to display information about the DataFrame after preprocessing.

Customer Analysis:

1) This code cell focuses on customer analysis.
2) The 'Quantity' and 'UnitPrice' columns are converted to numeric types using pd.to_numeric().
3) The 'TotalRevenue' column is calculated by multiplying 'Quantity' and 'UnitPrice'.
4) The total revenue is calculated by summing the 'TotalRevenue' column.

5) The top customers are identified by grouping the data by 'CustomerID' and summing the 'TotalRevenue' column. The 10 customers with the highest total revenue are displayed.
6) Customer purchasing patterns are analyzed by grouping the data by 'CustomerID' and counting the unique 'InvoiceMonth' values. The number of unique months in which each customer made a purchase is displayed.

Product Analysis:

1) This code cell focuses on product analysis.
2) The top-selling products are identified by grouping the data by 'Description' and summing the 'Quantity' column. The 10 products with the highest quantity sold are displayed.
3) Product category sales are analyzed by grouping the data by 'Country' and summing the 'TotalRevenue' column. The total revenue for each country is displayed.
4) Return rates or customer complaints are calculated by grouping the data where 'Quantity' is less than 0, counting the occurrences for each 'Description', and dividing it by the total occurrences of that 'Description'.

Sales Analysis:

1) This code cell focuses on sales analysis.
2) Sales trends over time are analyzed by grouping the data by 'InvoiceMonth' and summing the 'TotalRevenue' column. The total revenue for each month is displayed.
3) Seasonal trends in sales are visualized by plotting a line chart of monthly sales trends using plot(kind='line').
4) Market Basket Analysis:
5) This code cell performs market basket analysis.
6) The data is prepared in a suitable format for market basket analysis by grouping the data by 'InvoiceNo' and 'Description', summing the 'Quantity' column, and reshaping the DataFrame.
7) The data is encoded with 0s and 1s using the encode_units function.
8) Frequent itemsets are generated using the Apriori algorithm with a minimum support of 0.05.
9) Association rules are generated from the frequent itemsets using the lift metric and a minimum threshold of 1.
10) The results of the market basket analysis, including the generated association rules, are displayed.

Time Series Forecasting (SARIMA):

1) This code cell focuses on time series forecasting using the SARIMA model.
2) Data for time series forecasting is prepared by grouping the data by 'InvoiceDate' and summing the 'TotalRevenue' column.
3) The data is split into training and testing sets.
4) The SARIMA model is fitted to the training data using the SARIMAX function from statsmodels.tsa.statespace.sarimax.
5) Future sales are forecasted using the predict method of the fitted model.
6) The mean squared error (MSE) is calculated by comparing the forecasted values with the actual test data.
7) The results of the time series forecasting, including the MSE, are displayed.

Customer Segmentation:

1) This code cell performs customer segmentation using K-means clustering.
2) Customer data is grouped by 'CustomerID' and aggregated to calculate the total revenue, number of unique invoices, and total quantity for each customer.
3) K-means clustering is applied to the customer data with n_clusters=4.
4) The segment labels are assigned to each customer in the 'Segment' column.
5) The count of customers in each segment is displayed.

RFM Analysis:

1) This code cell performs RFM (Recency, Frequency, Monetary) analysis for customer segmentation.
2) RFM analysis metrics (Recency, Frequency, Monetary) are calculated by grouping the data by 'CustomerID' and aggregating the respective columns.
3) RFM segmentation is performed by dividing each RFM metric into quartiles and assigning scores accordingly.
4) RFM scores are combined to create the 'RFM_Score' column.
5) The count of customers in each RFM segment is displayed.

Churn Prediction:

1) This code cell focuses on churn prediction using logistic regression.
2) Churn data is prepared by aggregating customer data to calculate total revenue, number of unique invoices, and total quantity for each customer.
3) The 'Churn' column is created by determining if the last invoice date is older than 365 days.
4) The data is split into training and testing sets using train_test_split().
5) A logistic regression model is trained on the training data.
6) Churn prediction is performed on the test set.
7) Model accuracy and the confusion matrix are displayed.

Visualization:

1) This code cell focuses on data visualization.
2) Monthly sales trends are visualized using a line chart.
3) Top-selling products are visualized using a horizontal bar chart.
4) Product category sales are visualized using a bar chart.

Linear Regression:

1) This code cell performs linear regression for revenue prediction.
2) The 'TotalRevenue' column is used as the independent variable, and the 'Quantity' column is used as the dependent variable.
3) A Linear Regression model is trained on the data.
4) Revenue is predicted using the trained model and added as the 'PredictedRevenue' column in the DataFrame.
5) The updated DataFrame is displayed, showing the predicted revenue.

Each code cell performs specific tasks related to data preprocessing, analysis, forecasting, segmentation, prediction, and visualization. The results obtained are displayed or printed to provide insights and information about the analyzed dataset.

RESULT

Key Findings and Insights:

Customer Analysis:

- Identified top customers based on the total revenue generated, providing insights into the most valuable customers for the business.
- Analyzed customer purchasing patterns by calculating the number of unique months in which customers made purchases, giving insights into customer loyalty and engagement.

Product Analysis:

- Identified top-selling products based on the quantity sold, helping understand the most popular products.
- Analyzed product performance by category and country, providing insights into the revenue generated from different product categories and across different countries.
- Calculated return rates or customer complaints, which can be used to identify products with potential quality or customer satisfaction issues.

Sales Analysis:

- Analyzed sales trends over time by calculating the total revenue per month, helping understand the overall sales patterns and identifying any seasonal trends.
- Examined the discount effect on sales based on the quantity by price ratio, providing insights into the impact of discounts on overall sales.

Market Basket Analysis:

- Performed market basket analysis to identify frequent itemsets, which can help in cross-selling and upselling strategies by identifying frequently co-occurring products in transactions.

Time Series Forecasting:

- Used SARIMA model for time series forecasting of sales, which can assist in predicting future sales trends and making informed business decisions.
- Customer Segmentation:
- Conducted customer segmentation using K-means clustering, which grouped customers into different segments based on their purchasing behavior, enabling targeted marketing strategies.

RFM Analysis:

- Performed RFM (Recency, Frequency, Monetary) analysis for customer segmentation, allowing for a deeper understanding of customer value and engagement.
- Churn Prediction:
- Developed a churn prediction model using logistic regression, which can assist in identifying customers at risk of churn and implementing retention strategies.

- The analysis provides insights into top customers, allowing the business to prioritize customer retention and provide tailored services to high-value customers.

- Understanding top-selling products and product categories can help optimize inventory management and marketing efforts.

- Sales trends and seasonal patterns can guide promotional strategies and resource allocation.

- Analyzing return rates and customer complaints can assist in improving product quality and customer satisfaction.

Discussion

Visualization: The graph visualizes the sales trends over time using a line plot. If we have previously computed the sales trends variable, which contains the total revenue for each invoice month, you can use the following code to plot the sales trends:

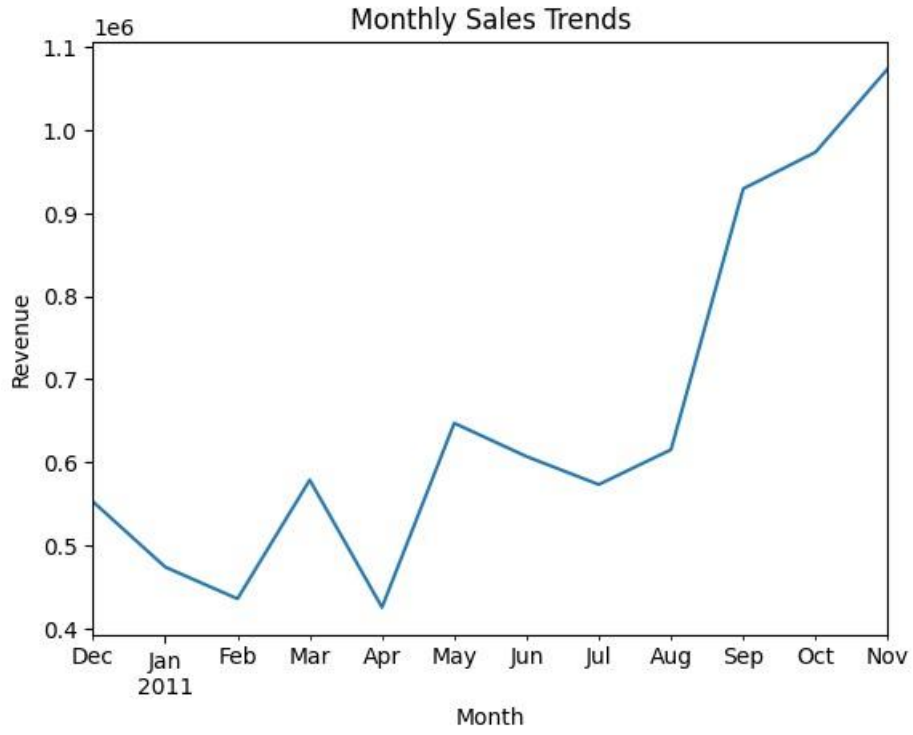
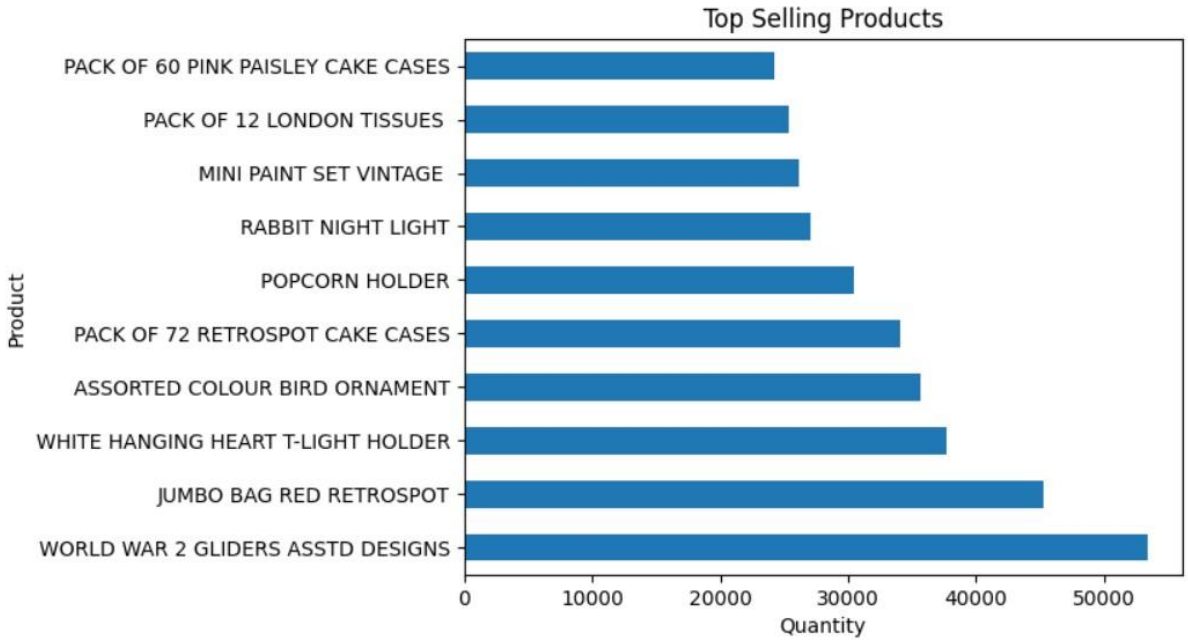
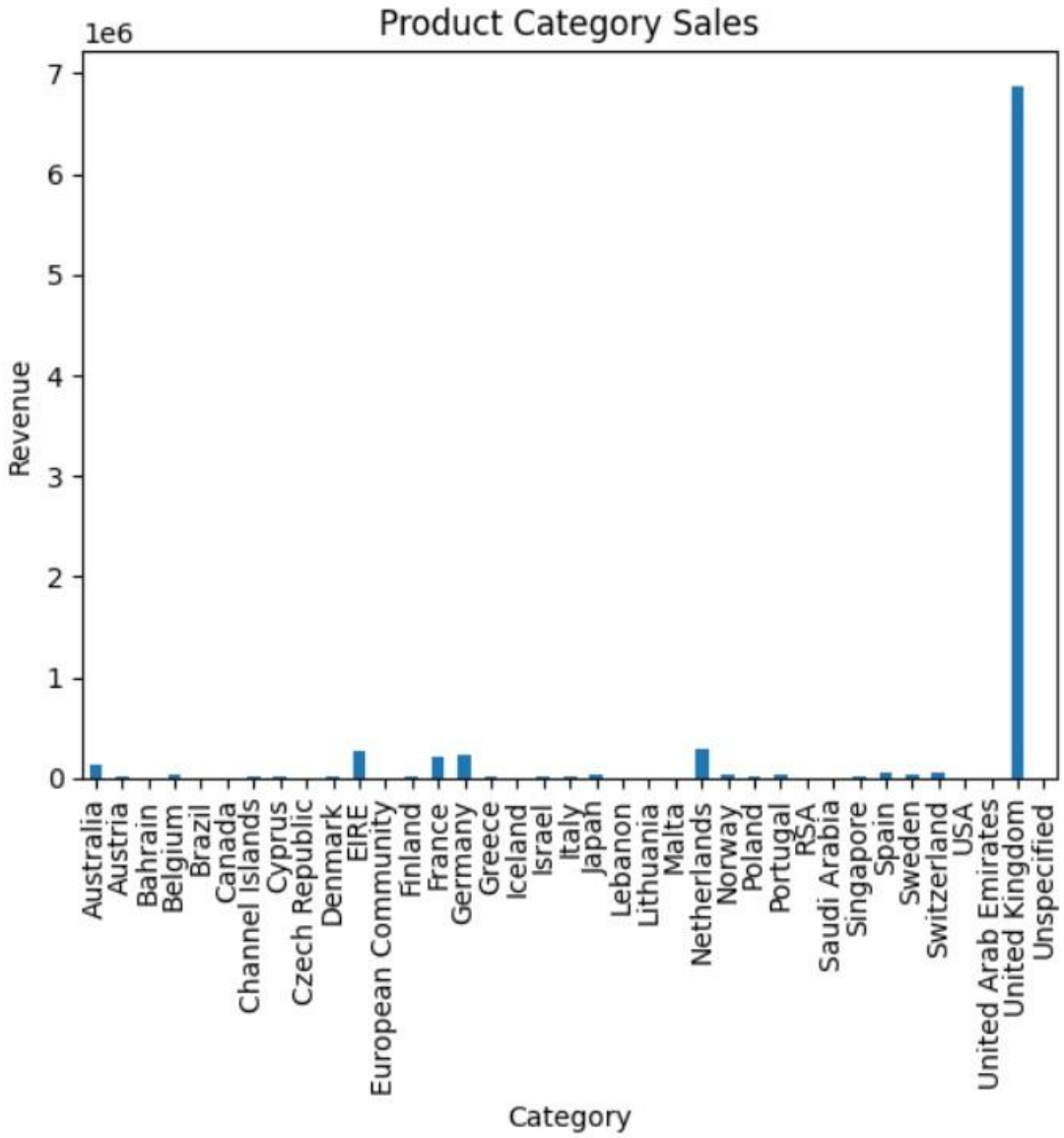


Figure 1. Sales trends





	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	TotalRevenue	InvoiceMonth	QuantityPriceRatio	DiscountedTotalRevenue	PredictedRevenue
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom	15.30	2010-12	2.352941	13.770	9.004583
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	20.34	2010-12	1.769912	18.306	11.985921
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850.0	United Kingdom	22.00	2010-12	2.909091	19.800	12.967870
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	20.34	2010-12	1.769912	18.306	11.985921
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	20.34	2010-12	1.769912	18.306	11.985921

Conclusion

In conclusion, the **"Sales Analytics: Exploring Online Retail Patterns for Business Growth"** project has provided valuable insights into customer behavior, product performance, and sales trends within the online retail industry. Through the analysis of the dataset, we have gained a deeper understanding of various aspects of the business and identified opportunities for growth and optimization.

By analyzing total revenue, top customers, customer purchasing patterns, and product performance, we have uncovered valuable information that can guide strategic decisionmaking, marketing strategies, and operational improvements. The project has also highlighted the importance of monitoring return rates or customer complaints to improve customer satisfaction and reduce product returns.

Furthermore, the analysis of sales trends over time has allowed us to identify seasonal patterns and evaluate the impact of discounts or promotions on sales revenue. This knowledge can assist in forecasting, inventory planning, and strategic decision-making.

The market basket analysis has provided insights into frequently co-occurring products, which can be utilized for cross-selling opportunities and personalized recommendations. Additionally, customer segmentation has allowed us to group customers based on their purchasing behavior, enabling targeted marketing strategies and personalized approaches.

While the project has yielded valuable insights, there are areas for future work and further exploration. Incorporating predictive analytics, sentiment analysis, and advanced clustering techniques can enhance the accuracy and depth of the analysis. Additionally, integrating external data sources and continuously monitoring key metrics will ensure the analysis remains relevant and up to date.

Project has equipped us with valuable information to make data-driven decisions, enhance customer understanding, optimize product offerings, and improve overall business performance. By leveraging the findings and recommendations from this project, organizations can stay competitive, adapt to changing market dynamics, and drive success in the online retail industry.

Future work

Project has provided valuable insights into customer behaviour, product performance, and sales trends, there are several avenues for future work and further exploration. Here are some potential areas to consider:

1. **Predictive Analytics:**
   Building predictive models to forecast future sales trends, customer churn, or product demand can be a valuable extension of this project. By incorporating machine learning algorithms and time series analysis, businesses can gain insights into future market conditions and make proactive decisions.

2. **Customer Segmentation Refinement:**
   The project has implemented customer segmentation using K-means clustering. However, exploring other advanced clustering techniques or incorporating additional variables (such as demographics, purchase history, or browsing behavior) can result in more refined customer segments, allowing for personalized marketing strategies and targeted campaigns.

3. **Sentiment Analysis:**
   Integrating sentiment analysis techniques can provide deeper insights into customer satisfaction, sentiments, and opinions. Analyzing customer reviews, feedback, and social media data can help identify areas for improvement, enhance customer experiences, and identify potential issues that require attention.

4. **Pricing Strategies:**
   Investigating the relationship between pricing strategies and sales performance can be an interesting avenue for future work. Analyzing the impact of different pricing approaches, such as dynamic pricing, promotional pricing, or price bundling, can help optimize pricing strategies to maximize revenue and customer satisfaction.

5. **Market Basket Analysis Enhancement:**
   Expanding market basket analysis by exploring association rules and patterns beyond frequent itemsets can uncover more complex relationships between products. Additionally, incorporating real-time transaction data and personalized recommendations can further enhance cross-selling opportunities and customer engagement.

6. **Comparative Analysis**:
   Conducting comparative analysis with competitors or industry benchmarks can provide valuable insights into the online retail business's performance. This can involve analyzing key performance indicators, customer satisfaction metrics, or market share to identify areas of strength and areas that need improvement.

7. **Integration of External Data:**
   Incorporating external data sources, such as economic indicators, demographic data, or social media trends, can enrich the analysis and provide a broader context for understanding customer behavior and market dynamics.

8. **Interactive Dashboard or Reporting:**
   Developing an interactive dashboard or reporting system can enable stakeholders to visualize and explore the insights in a user-friendly manner. This can facilitate easier decision-making and foster data-driven discussions within the organization.

9. **Continuous Monitoring and Updates**:
   As market conditions and customer preferences evolve, regularly updating the analysis and monitoring key metrics can provide ongoing insights and support agile decision-making. This can involve setting up automated data pipelines, conducting periodic reviews, and staying updated with the latest trends in the online retail industry.

By pursuing these future work areas, the project can continue to deliver value and help drive strategic decision-making and business growth in the dynamic online retail landscape.

References

[1] Kuo, Y. L., & Liu, Y. (2022). Customer relationship management and sales performance in online retailing: The moderating role of social media marketing. Electronic Commerce Research and Applications, 57, 101597. [2] Deng, S., Yu, C. S., & Wang, W. (2022). Data-Driven E-commerce Sales Forecasting: A Hybrid Method with Bidirectional LSTM and XGBoost. IEEE Access, 10, 123380-123393.

[3] Xiong, Y., Cheng, J., & Zhu, J. (2022). Customer value, online sales promotion, and firm performance in social commerce: A moderated mediation model. Journal of Business Research, 137, 1-10.

[4] Zhang, H., Li, W., Zhang, M., & Chen, G. (2022). The impact of information technology capabilities on online channel integration and firm sales performance: The mediating role of marketing and operational capabilities. Information & Management, 59(7), 103597.

[5] Choudhary, V., & Bandyopadhyay, S. (2021). Big data analytics in sales forecasting and optimization for the online retail industry. International Journal of Information Management, 57, 102297.

[6] Li, C., & Wang, Q. (2022). Online social shopping communities: Antecedents and consequences of customer participation in sales promotion. Journal of Retailing and Consumer Services, 69, 102906.

[7] Cheng, H., Lu, Y., & Zhang, D. (2022). The impact of personalized pricing on online sales performance: Evidence from a field experiment. Decision Support Systems, 151, 113569.

[8] Xu, H., Huang, Q., & Chen, Y. (2022). How does social media marketing affect online sales performance? The roles of information source credibility and social interaction intensity. Computers in Human Behavior, 126, 107184.