Exploratory Data Analysis (EDA) Report

Project Title: Customer Segmentation Using Online Retail Data

Author: Ojeranti Oluwateniola Toluwalase

Date: 9/10/2024

# 1. Introduction

Objective:

This project aims to explore the Online Retail dataset to uncover patterns in customer behavior, identify top-selling products, and analyze sales trends over time. The insights will help guide marketing strategies and customer segmentation efforts.

Dataset Description:

- Source: Kaggle

- Rows: 541909

- Columns: 8

- Timeframe: 01/12/2010 and 09/12/2011

- Key Columns: InvoiceNo, StockCode, Description, Quantity, InvoiceDate, UnitPrice, CustomerID, Country

# 2. Data Cleaning & Preparation

Steps Taken:

- Removed null and duplicate records

- Converted date fields to datetime objects

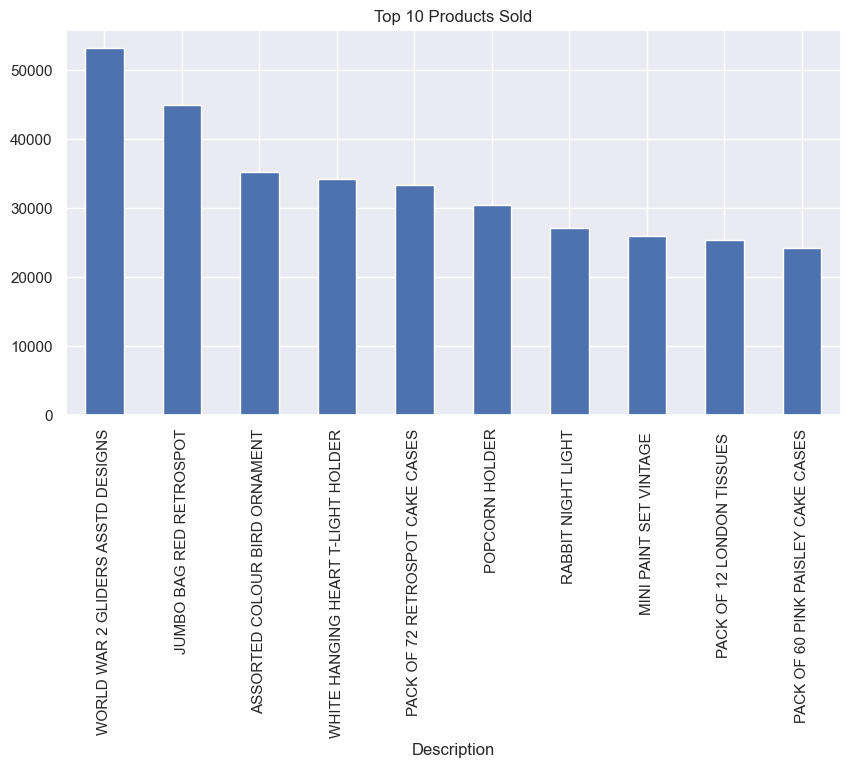
- Created new features: TotalPrice = Quantity \* UnitPrice, Month

- Filtered out canceled orders and negative quantities

# 3. Descriptive Analysis

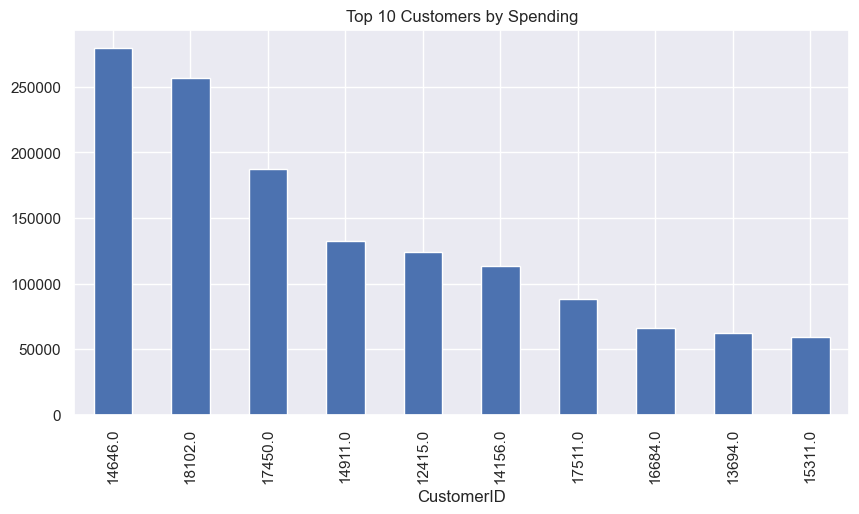
🔹 Top 10 Products Sold

* WORLD WAR 2 GLIDERS ASSTD DESIGNS sold 53119 units.( highest-selling item)
* JUMBO BAG RED RETROSPOT sold 44963 units.
* ASSORTED COLOUR BIRD ORNAMENT sold 35215 units.
* WHITE HANGING HEART T-LIGHT HOLDER sold 34128 units.
* PACK OF 72 RETROSPOT CAKE CASES sold 33386 units.
* POPCORN HOLDER sold 30492 units.
* RABBIT NIGHT LIGHT 27045 units.
* MINI PAINT SET VINTAGE sold 25880 units.
* PACK OF 12 LONDON TISSUES sold 5305 units.
* PACK OF 60 PINK PAISLEY CAKE CASES sold 24129 units.



🔹 Top 10 Customers by Revenue

|  |
| --- |
| CustomerID Revenue |
| 14646.0 279489.02 |
| 18102.0 256438.49 |
| 17450.0 187322.17 |
| 14911.0 132458.73 |
| 12415.0 123725.45 |
| 14156.0 113214.59 |
| 17511.0 88125.38 |
| 16684.0 65892.08 |
| 13694.0 62690.54 |
| 15311.0 59284.19 |
|  |

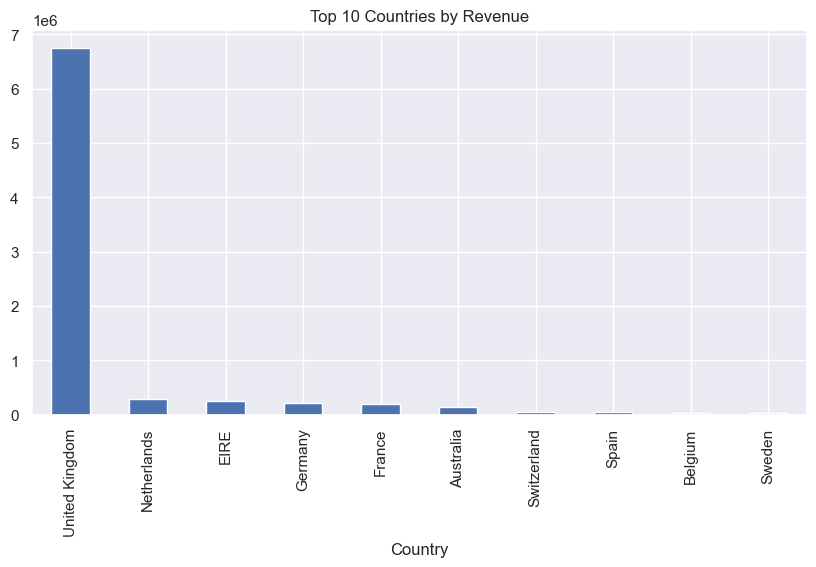


🔹 Monthly Sales Trend



The sales increased from the beginning of they year 2011 till the November with recorded the highest number of sales at   1126815.070 units but took a sudden drop by December which recorded the lowest number of sales at  341539.430 units.

🔹 Country-Wise Revenue



United Kingdom has provided the highest revenue with 6747156.154 units and Saudi Arabia withe the lowest revenue of  131.170 units.

# 4. Pattern & Anomaly Detection

- Outlier transactions identified using box plots

- Unusually high-value orders flagged for further review

- Some customers showed extremely high purchase frequency

# 5. Hypothesis Testing

\*\*Example Hypotheses:\*\*

- H1: Higher quantity orders lead to higher spending (confirmed via correlation)

- H2: Sales peak during the holiday season (confirmed by date-based grouping)

Total Price and Quantity have a high positive correlation of 0.91613. This means that there is a strong positive relationship between the total price and the quantity sold. A correlation of 0.91613 indicates that about 91.6% of the variation in total sales (Total Price) can be explained by the variation in the quantity sold, suggesting they move in the same direction most of the time.

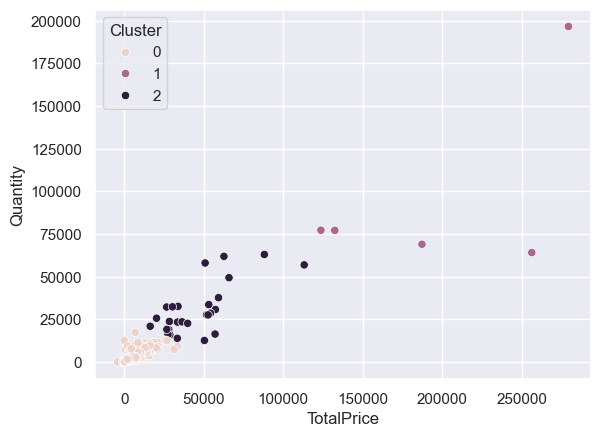
# 6. Customer Segmentation (Optional Advanced Section)

Technique Used: K-Means Clustering

- Clustered customers by Total Price and Total Quantity

- Identified 3 segments: Low, Medium, High-value customers

- Scatter plot of clusters included



# 7. Key Insights

- Majority of revenue comes from a small group of customers

- Seasonal trends significantly impact sales

- Certain products are consistently popular and could be promoted more

- Customer segmentation reveals opportunities for personalized marketing

# 8. Recommendations

- Target high-value customers with loyalty programs

- Optimize inventory for top-selling items

- Increase marketing in high-revenue months

- Explore fraud detection for high-value outliers

# 9. Appendix

- Libraries Used: pandas, seaborn, matplotlib, scikit-learn

- Notebook Link: <https://github.com/Teniola2022/Exploratory-Data-Analysis>

- Data Source: Kaggle