

Customer Behavior Analysis and Outlier Detection in E-Commerce

Abstract

This project analyzes purchase patterns, segments customers, finds outliers, and mines association rules using customer transactional data from an e-commerce platform. The study offers practical insights into consumer behavior and product correlations by utilizing sophisticated data analysis and machine learning approaches, including association rule mining (Apriori and FP Growth algorithms), RFM analysis, K-Means clustering, Isolation Forest, and more. Key findings include customer segmentation into four distinct groups, identification of anomalous purchasing patterns, and the discovery of strong product associations for cross-selling opportunities. These findings provide useful tactics for improving e-commerce marketing, product bundling, and customer satisfaction.

Introduction

Understanding customer behavior and purchasing patterns is essential for e-commerce businesses to optimize marketing strategies and improve customer satisfaction. This project focuses on analyzing consumer purchasing habits using transactional data from an online retail store. The dataset contains detailed transactional information, including product descriptions, quantities, prices, and customer demographics.

Objectives

Goals

The main objective of this project are:

- **Customer Segmentation:** RFM analysis and K-Means clustering are used to segment the client base according to their purchase patterns.
- **Outlier Detection:** To use the Isolation Forest method to find odd consumer transactions or behaviors.
- **Association Rule Mining:** Using the Apriori and FP Growth algorithms, we aim to find important product relationships and frequently occurring itemsets for cross-selling opportunities.
- **Performance Comparison:** To evaluate and compare the efficiency of Apriori and FP Growth algorithms in association rule mining.

2. Data Collection

The dataset has been taken from Kaggle.

<https://www.kaggle.com/code/hellbuoy/online-retail-k-means-hierarchical-clustering/input>

Transactional data from an Online E-commerce store is included in the OnlineRetail.csv dataset. The dataset contains a total of 541,909 records and 8 features. The features include InvoiceNo, StockCode, Description, Quantity, UnitPrice, CustomerID, InvoiceDate, and Country.

3. Data Preprocessing

- **Handling Missing Values:** There were missing values in the dataset.

- CustomerID contained 135,080 missing values and the Description column had 1,454 missing values. These records were dropped from the analysis.
- **Filtered Negative or Zero Values:**
 - We removed transactions where the Quantity or UnitPrice was zero or negative, as these represented invalid or erroneous data.
- **Creating New Features:**
 - A new feature, TotalSales, was created by multiplying Quantity with UnitPrice. This helped to represent the monetary value of each transaction.
- **Date Transformation:**
 - The InvoiceDate was converted into a datetime format to facilitate time-based analysis.

Column	Count	Mean	Min	Max	25%	50%	75%
Quantity	397884	12.99	1	80995	2	6	12
UnitPrice	397884	4.25	0.01	39.99	1.90	3.50	5.99
TotalSales	397884	22.40	0.001	168469.6	4.68	11.80	19.80

Figure: Table showing statistics of some important features.

4. Exploratory Data Analysis (EDA)

Various analysis were performed on the dataset to understand the trends and draw meaningful insights. A summary of the EDA is mentioned below:

- **Missing Values:**
 - There are missing values in the dataset, in the CustomerID and Description columns.
 - CustomerID contained 135,080 missing values and the Description column had 1,454 missing values.

```

[5 rows x 8 columns]
InvoiceNo      0
StockCode      0
Description    1454
Quantity       0
InvoiceDate    0
UnitPrice      0
CustomerID    135080
Country        0
dtype: int64
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 541909 entries, 0 to 541908

```

Figure1: Image showing the missing values in the dataset.

- **Data Types:**

#	Column	Non-Null Count	Dtype
0	InvoiceNo	541909 non-null	object
1	StockCode	541909 non-null	object
2	Description	540455 non-null	object
3	Quantity	541909 non-null	int64
4	InvoiceDate	541909 non-null	object
5	UnitPrice	541909 non-null	float64
6	CustomerID	406829 non-null	float64
7	Country	541909 non-null	object

dtypes: float64(2), int64(1), object(5)
memory usage: 33.1+ MB

Figure2: Table showing the datatypes and features information before preprocessing the data.

- Quantity and UnitPrice are numerical characteristics.
- InvoiceDate was changed to datetime format.

- **Statistical Summary:**

	Quantity	...	TotalSales
count	397884.000000	...	397884.000000
mean	12.988238	...	22.397000
min	1.000000	...	0.001000
25%	2.000000	...	4.680000
50%	6.000000	...	11.800000
75%	12.000000	...	19.800000
max	80995.000000	...	168469.600000
std	179.331775	...	309.071041

Figure3: Table showing statistics of Quantity and TotalSales column.

- **Distribution of Monetary Values:**

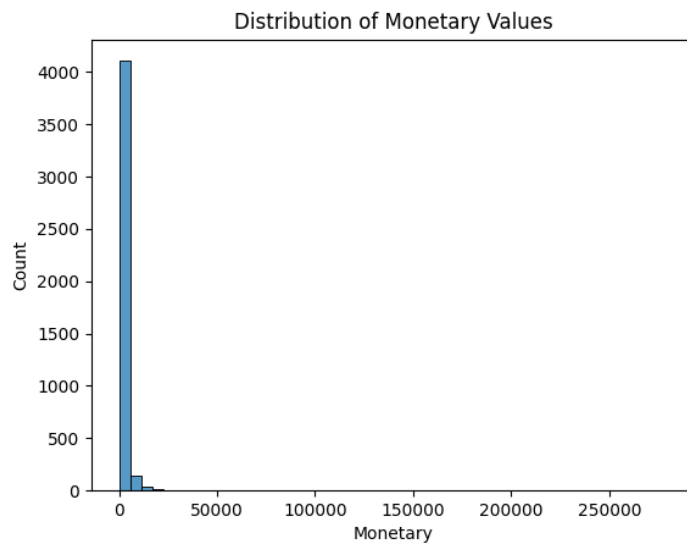


Figure4: The TotalSales histogram showed a highly skewed distribution, with a small number of transactions having extremely high values and the majority having low values.

- **Total Sales Distribution by Top 10 Countries**

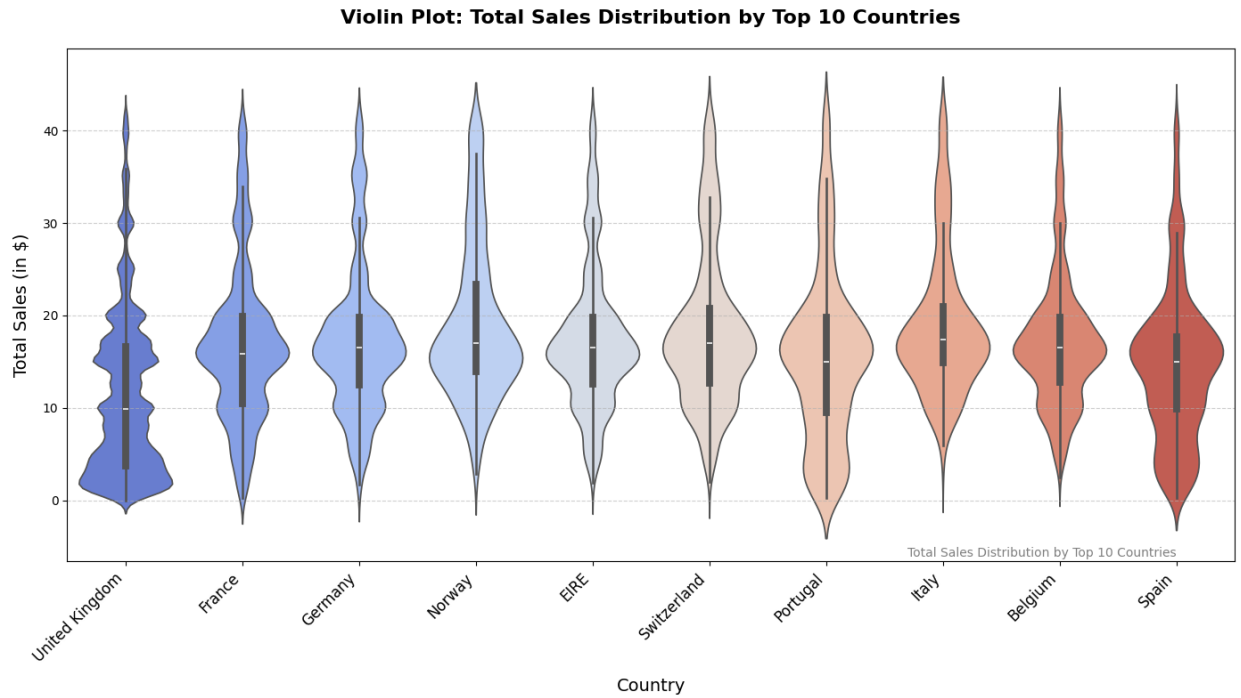


Figure5: Violin plot providing insights into the distribution of total sales across the top 10 countries.

1. United Kingdom shows the widest distribution, indicating high variability in sales.
2. France, Germany, and Norway have more balanced distributions, with sales clustering around the median.
3. Portugal and EIRE exhibit narrower ranges, suggesting more consistent sales patterns.
4. Spain and Belgium demonstrate slightly wider variability compared to others in their cluster.
5. The median sales (black bar) vary across countries, highlighting different central tendencies.

- **Distribution of Sales amount among the top 10 countries**

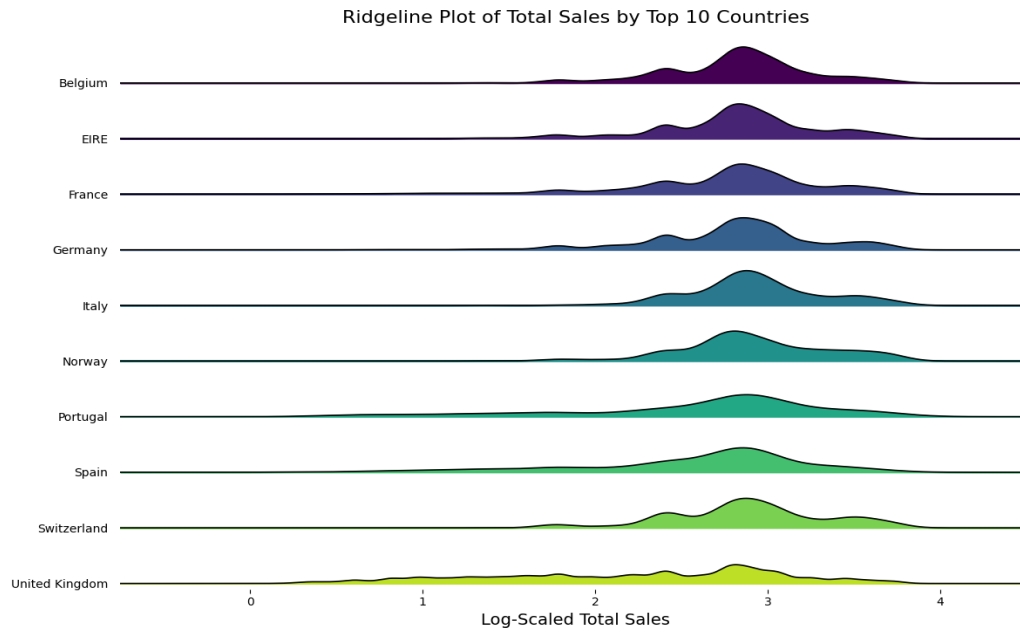


Figure6: The ridgeline plot shows the log-scaled total sales distributions for the top 10 countries. The United Kingdom has the widest range, indicating high sales variability, while countries like Belgium and EIRE show narrower, more consistent patterns.

● Distribution of Top 10 Customers by Total Sales

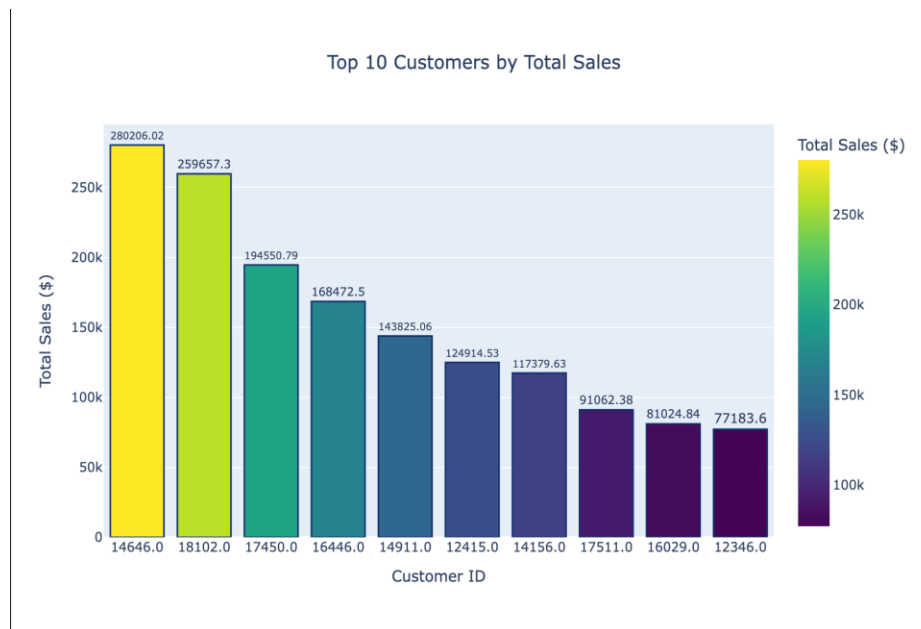


Figure6: Sales of the top 10 countires.

1. Highest Sales Contribution:

- The customer with ID 14646 . 0 has the highest total sales at \$280,206.02, significantly outperforming others.

2. Distribution of Sales:

- The top two customers (IDs 14646 . 0 and 18102 . 0) dominate the chart, contributing over \$250,000 each.
- The remaining customers exhibit a more gradual decline in sales totals, with the 10th-ranked customer (ID 12346 . 0) generating \$77,183.6.

3. Color Gradient:

- The color intensity represents the sales magnitude, transitioning from bright yellow (highest sales) to deep purple (lower sales), helping to visually distinguish the contribution levels.

4. Sales Gap:

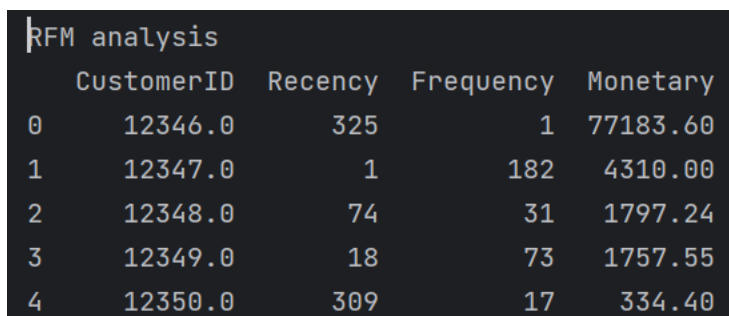
- There is a noticeable gap between the top customers and those at the lower end of the top 10 list, indicating a disparity in sales contributions.

5. Feature Engineering: RFM Analysis

RFM (Recency, Frequency, and Monetary) analysis was conducted on the cleaned dataset:

- **Recency:** The number of days since the customer's last purchase.
- **Frequency:** The number of purchases made by the customer.
- **Monetary:** The total monetary value spent by the customer.

The RFM analysis grouped the data by CustomerID, generating key insights on customer behavior and spending patterns.



	CustomerID	Recency	Frequency	Monetary
0	12346.0	325	1	77183.60
1	12347.0	1	182	4310.00
2	12348.0	74	31	1797.24
3	12349.0	18	73	1757.55
4	12350.0	309	17	334.40

Figure7: Image showing results of RFM Analysis. (Top customers)

6. Outlier Detection

```
number of outliers
Outlier
1      4294
-1      44
Name: count, dtype: int64
```

Figure8: Result showing the number of outliers detected.

Using the Isolation Forest algorithm, we identified outliers based on the RFM features. The algorithm detected:

- 4,294 normal customers (Outlier = 1)
- 44 outliers (Outlier = -1)

These outliers represent customers whose purchasing behavior is significantly different from the majority.

7. Customer Segmentation (Clustering)

To segment customers based on RFM values, we used the K-Means clustering algorithm, scaled using StandardScaler:

- We chose 4 clusters, which segmented customers into distinct groups based on their recency, frequency, and monetary values.
- **Silhouette Score for clustering** was 0.60, indicating a relatively good clustering result.
- The resulting customer segments were visualized using a scatter plot, showcasing the relationship between Recency and Monetary values for each segment.

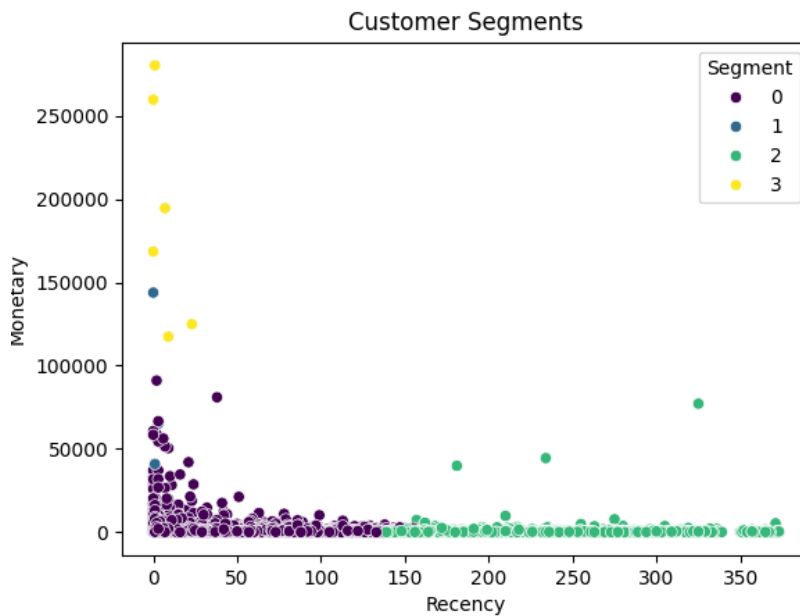


Figure 9: Customer Segmentation after clustering.

8. Association Rule Mining

8.1.1 Apriori Algorithm

Several frequent itemsets were found by the Apriori algorithm; each itemset is a group of products that occur together in transactions at a specific frequency.

Support	Itemsets
0.018888	frozenset({'12 PENCIL SMALL TUBE WOODLAND'})
0.015111	frozenset({'12 PENCILS SMALL TUBE SKULL'})
0.014571	frozenset({'12 PENCILS SMALL TUBE RED RETROSPOT'})
0.012952	frozenset({'10 COLOUR SPACEBOY PEN'})
0.010793	frozenset({'12 MESSAGE CARDS WITH ENVELOPES'})

Figure10: Top 5 frequent itemsets based on their support values.

8.1.2 Association Rules

The Apriori algorithm generated association rules, which provide insights into the likelihood that certain products are purchased together.

Antecedents	Consequents	Zhang's Metric
frozenset({'60 CAKE CASES DOLLY GIRL DESIGN'})	frozenset({'PACK OF 60 PINK PAISLEY CAKE CASES'})	0.953142
frozenset({'HERB MARKER PARSLEY'})	frozenset({'HERB MARKER ROSEMARY'})	0.969161
frozenset({'60 CAKE CASES DOLLY GIRL DESIGN'})	frozenset({'12 PENCIL SMALL TUBE WOODLAND'})	0.975923
frozenset({'PACK OF 60 SPACEBOY CAKE CASES'})	frozenset({'12 PENCILS SMALL TUBE RED RETROSPOT'})	0.978619
frozenset({'60 CAKE CASES VINTAGE CHRISTMAS'})	frozenset({'12 MESSAGE CARDS WITH ENVELOPES'})	0.958781

Figure11: Figure showing Association Rules.

8.2. FP Growth Algorithm

The FP Growth algorithm was applied to the dataset to find frequent itemsets more efficiently. The key results from the FP Growth algorithm include frequent itemsets and association rules:

Support	Itemsets
0.069501	frozenset({'LUNCH BAG RED RETROSPOT'})
0.046838	frozenset({'JAM MAKING SET PRINTED'})
0.026926	frozenset({'PACK OF 72 SKULL CAKE CASES'})
0.025092	frozenset({'SET OF 12 FAIRY CAKE BAKING CASES'})
0.024660	frozenset({'PINK BLUE FELT CRAFT TRINKET BOX'})

Figure12: The top 5 frequent itemsets identified by FP Growth

8.2.2 Association Rules

Antecedents	Consequents	Zhang's Metric
frozenset({'WHITE HANGING HEART T-LIGHT HOLDER'})	frozenset({'LUNCH BAG RED RETROSPOT'})	0.440737
frozenset({'LUNCH BAG RED RETROSPOT'})	frozenset({'JAM MAKING SET PRINTED'})	0.423280
frozenset({'JUMBO BAG RED RETROSPOT'})	frozenset({'LUNCH BAG RED RETROSPOT'})	0.808118
frozenset({'LUNCH BAG RED RETROSPOT'})	frozenset({'SET OF 12 FAIRY CAKE BAKING CASES'})	0.793497
frozenset({'JAM MAKING SET PRINTED'})	frozenset({'PINK BLUE FELT CRAFT TRINKET BOX'})	0.909060

Figure13: Association rules generated from FP Growth Algorithm

The results from both algorithms were saved as CSV files.

9. Performance Evaluation

- **Silhouette Score for clustering:** 0.60, indicating a moderate clustering quality.
- **Visualization:** We used both static (Seaborn) and interactive (Plotly) visualizations to showcase customer segments.
- **Association Rule Mining:**
 - The FP Growth algorithm demonstrated better computational efficiency and scalability compared to Apriori, making it suitable for larger datasets.

10. Conclusion

In order to identify trends in product purchasing behaviour, this research used association rule mining and K-Means clustering to effectively divide clients into meaningful groups. The association rules' outcomes can yield useful information for product suggestions and cross-selling. The methods FP Growth and Apriori both successfully found worthwhile correlations, however FP Growth has superior scalability for big datasets. These models' findings can be used to improve in-store positioning tactics, product bundling, and marketing efforts, which will ultimately increase sales and consumer happiness.