Online Sales Analysis

Siti Asmaul Kusnah

INTRODUCTION

Background

The online sales dataset is an invaluable source of insights into customer behavior, sales trends, and opportunities for business growth. The dataset is extracted from Kaggle, consisting of two files:

- Orders.csv: Contains customer information, including order ID, order date, customer name, state, and city
- Details.csv: Provides details of all subsequent orders, including order ID and corresponding data

Objectives

1. To create insightful visualizatons for online sales performance
2. To analyze customer behavior using clustering (K-Means)
3. To conduct RFM (Recency, Frequecy, Monetary) analysis for customer segmentation

DATASET OVERVIEW

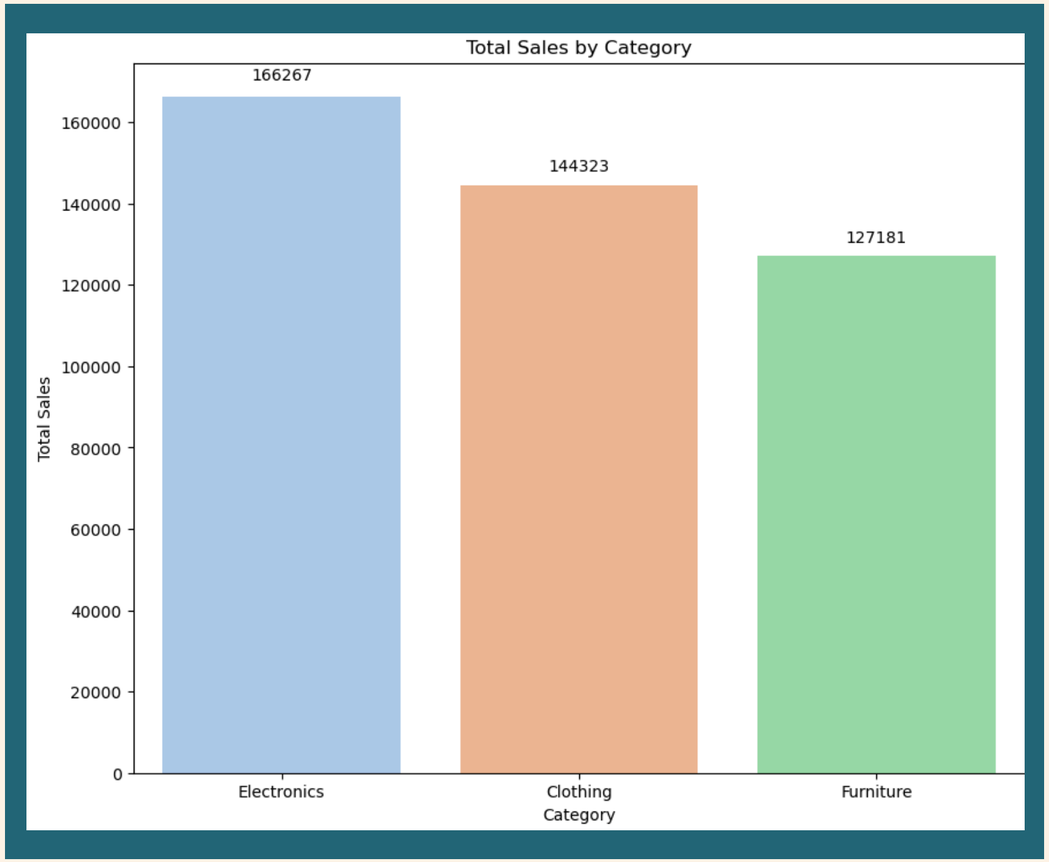
Orders.csv

- Number of Rows: 500
- Number of Columns: 5
- Variables:
  - Order ID: A unique identifier for each order (Primary Key)
  - Order Date: The data when the order was placed
  - Customer Name: The name of the customer who placed the order
  - State: The state where the customer resides
  - City: The city where the customer resides

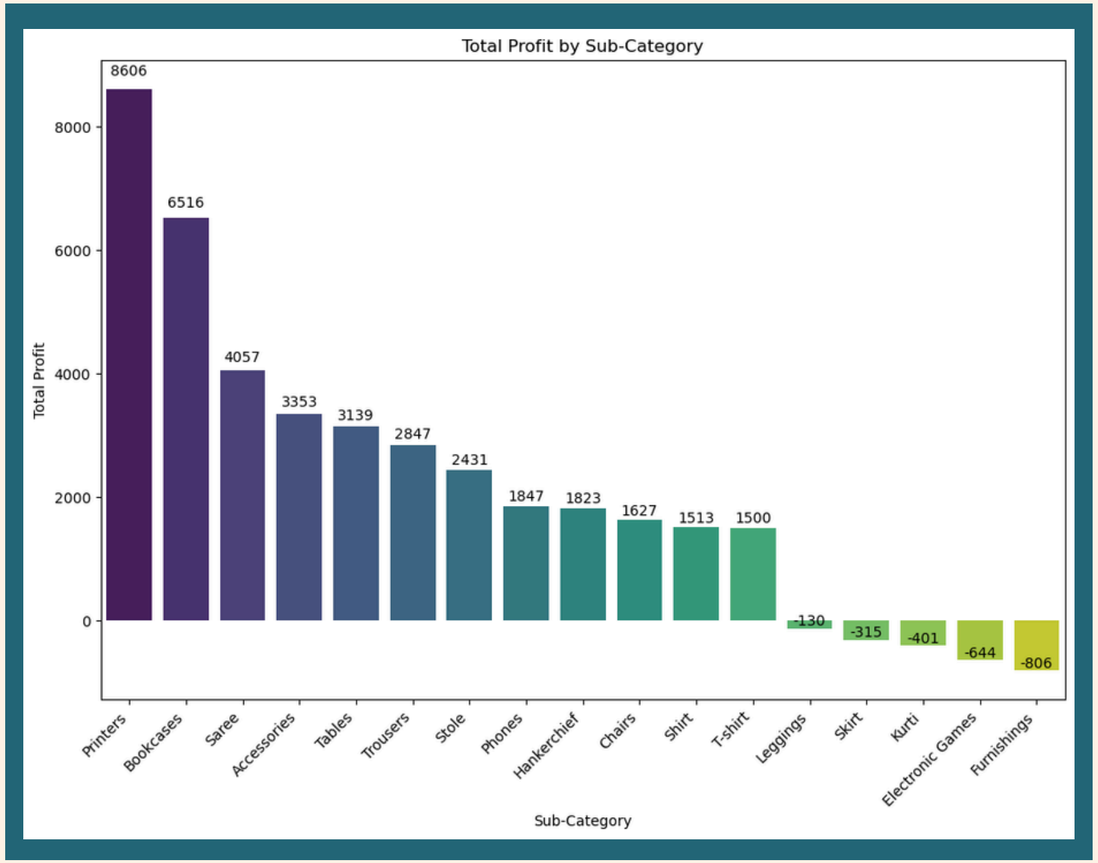
Details.csv

- Number of Rows: 1500
- Number of Columns: 7
- Variables:
  - Order ID: A unique identifier linking each order details to the corresponding order in Orders.csv
  - Amount The total monetary value of the order
  - Profit: The profit earned from the order after subtracting costs
  - Quantity: The number of items purchased in the order
  - Category: The broad category of the purchased product
  - Sub-Category: A more specific classification of the product within its category
  - Payment Mode: The payment method used by the customer

Total Sales by Category



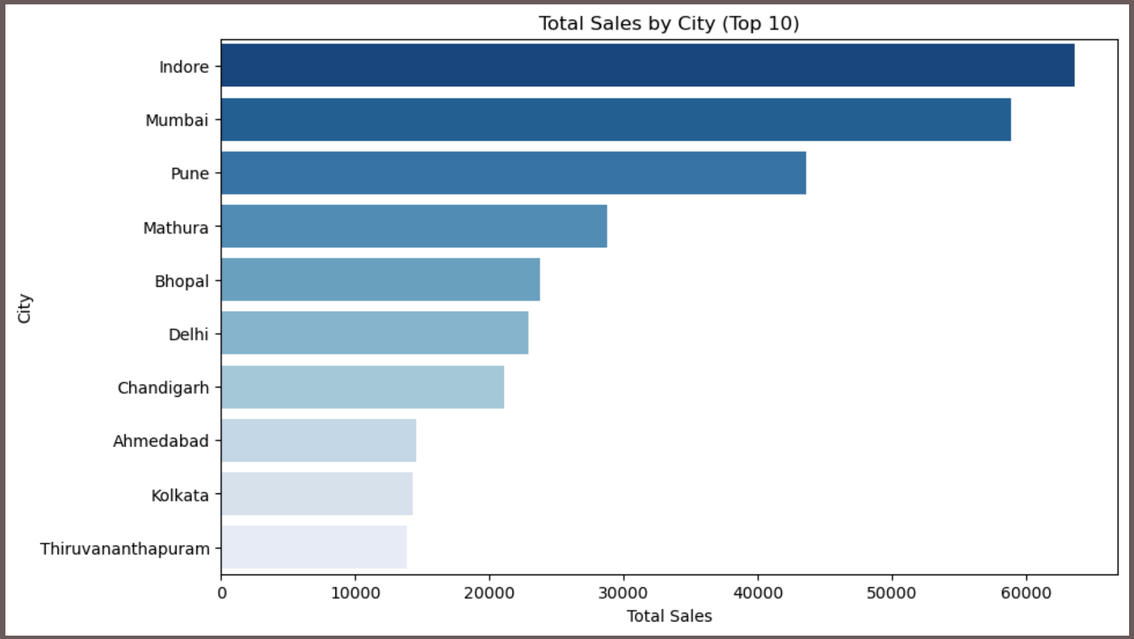
Total Profit by Sub-Category



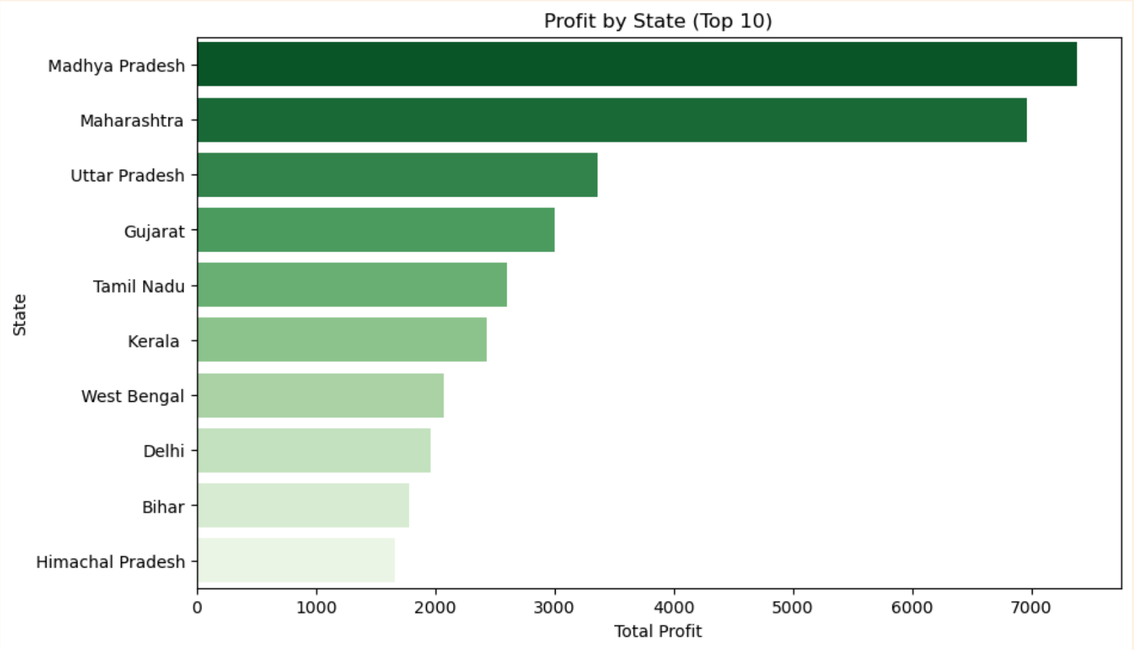
Both graphs show that:

- High sales do not always result in high profits. For example, the electronics category has the highest sales, but some of its sub-categories have low profits or even losses.
- Focusing on high-profit sub-categories, such as printers and sarees, can help improve overall profitability.
- Pricing strategies and operational efficiency should be improved in loss-making sub-categories such as leggings, skirts, kurti, electronic games, and furnishings.

Total Sales by City (Top 10)

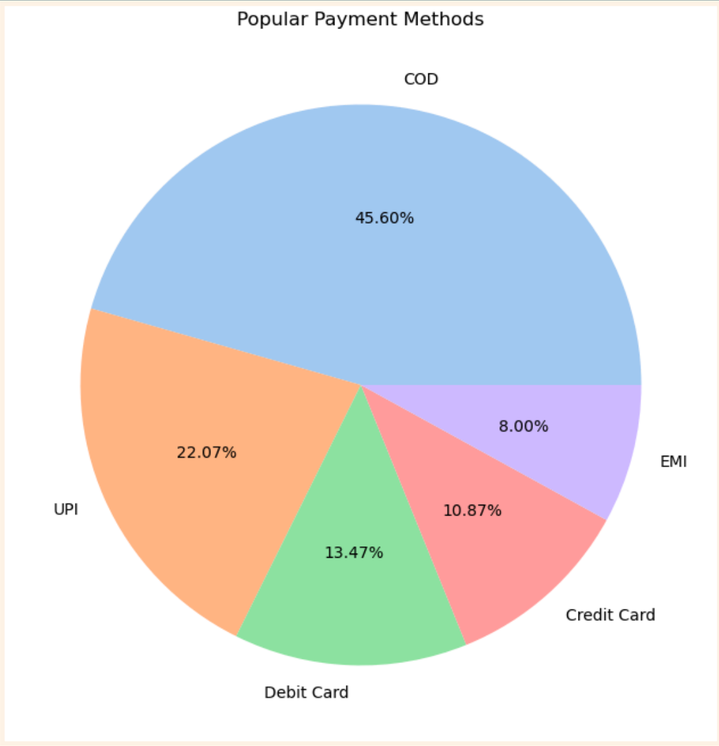


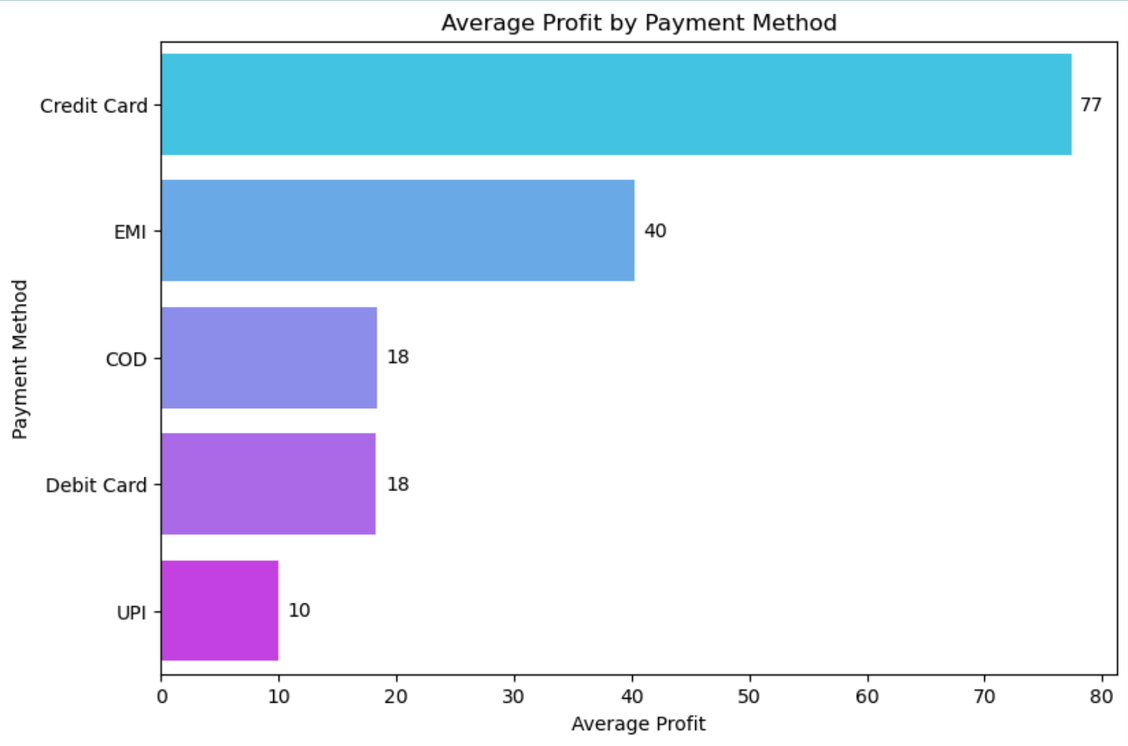
Profit by State (Top 10)



- Madhya Pradesh (with the city of Indore) and Maharashtra (with Mumbai and Pune) were the key markets that contributed the most profit and online sales. A focused digital marketing strategy in these regions could further improve performance.
- Profit does not always align with total revenue, suggesting the need to evaluate margins and operational efficiency in high revenue but low profit regions.
- Low contribution cities such as Thiruvananthapuram could be targeted for exploring new market penetration strategies to increase online sales.

PAYMENT METHOD





- COD is the most frequently used payment method but contributes less to average profit compared to credit card or EMI payments.
- Businesses might consider promoting credit card or EMI options to maximize profits.

K-MEANS ANALYSIS

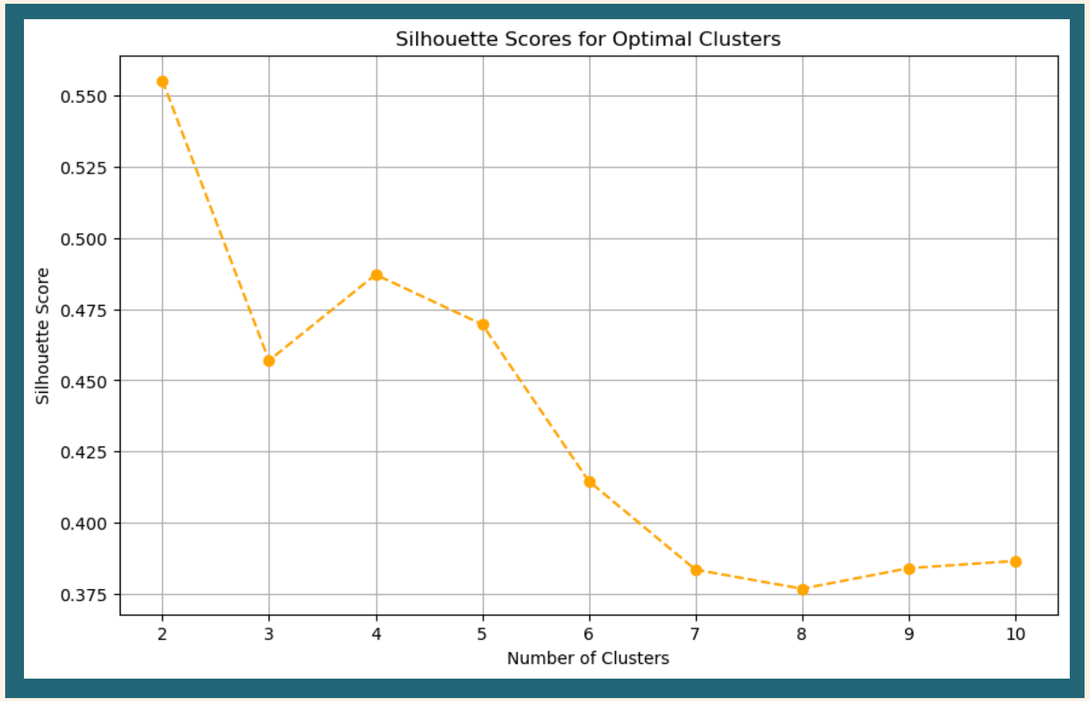
K-Means is used for customer segmentation by grouping customers based on purchasing patterns with the aim of:

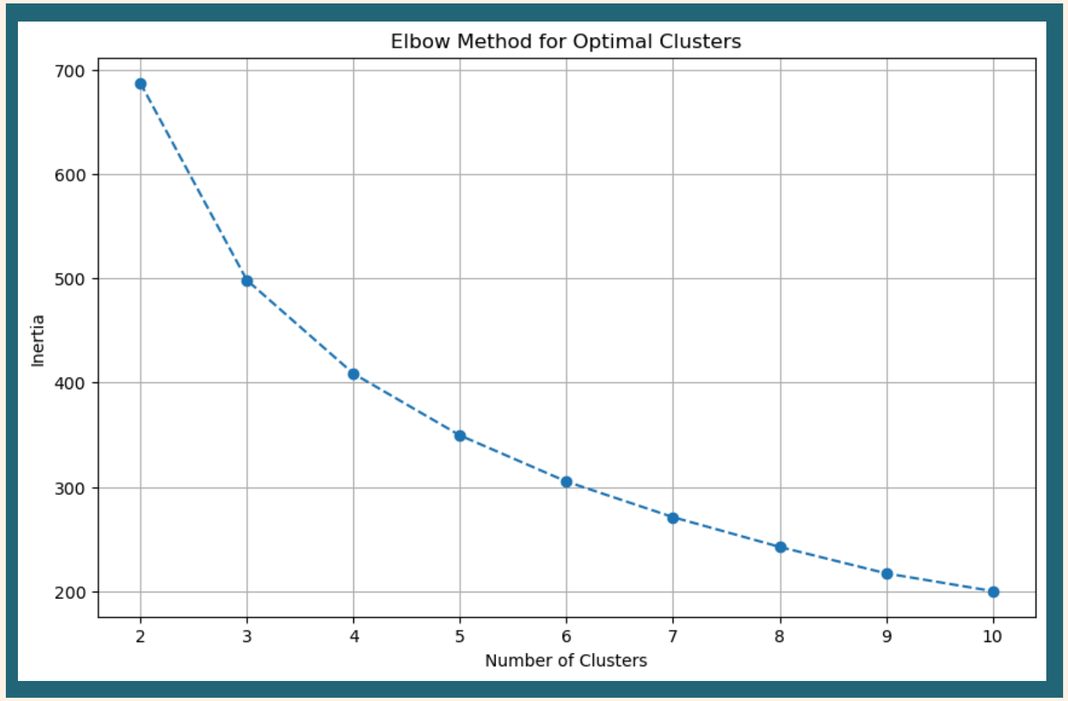
- Which customers are the most profitable.
- Customers who often buy in large quantities.
- Customers with high potential to improve profitability.

Metrics

- Amount: Total amount of money spent by the customer
- Profit: Total profit of the customer
- Order ID: The number of orders placed by the customer
- Quantity: Number of product units purchased by the customer

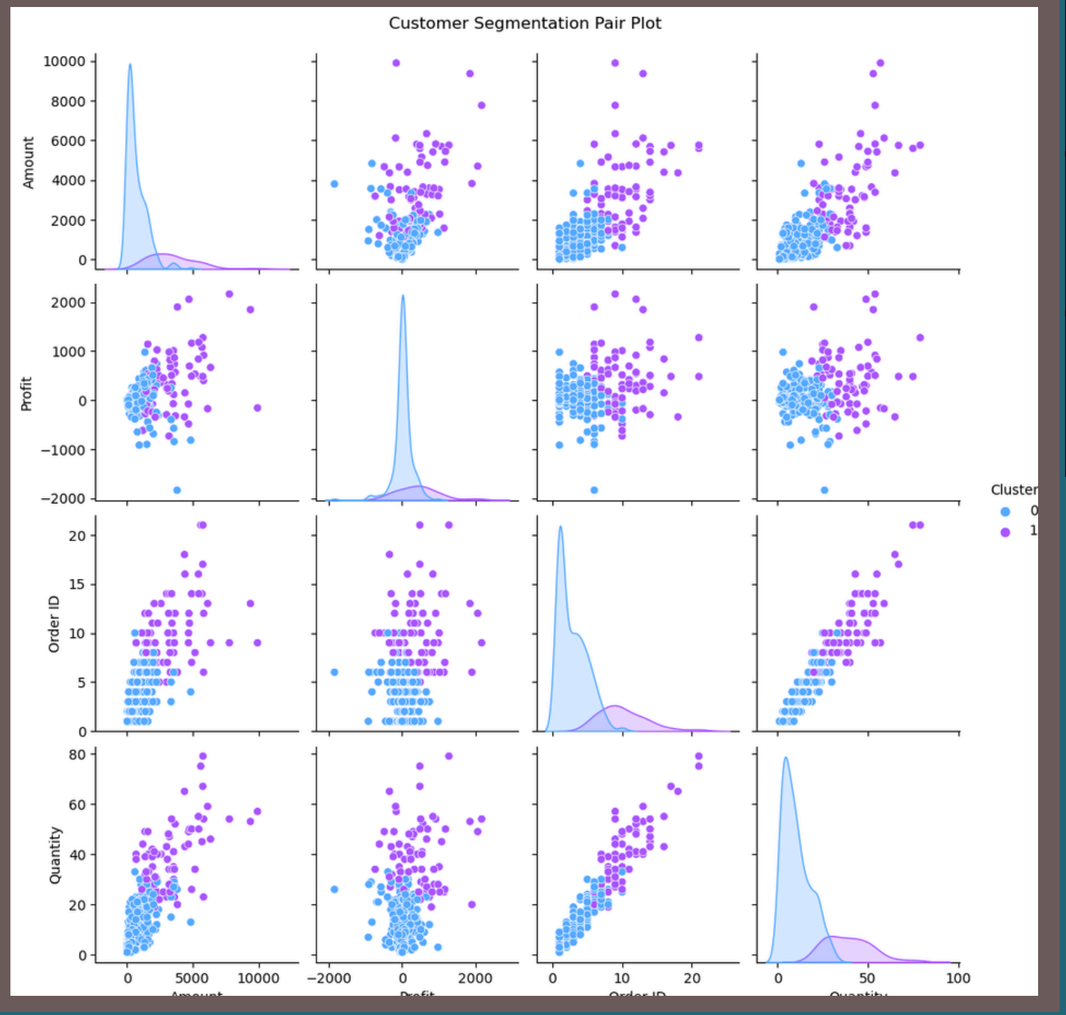
Determination of Number of Clusters





Determining the optimal number of clusters is based on Silhouette scores and the Elbow method. The optimal number of clusters based on the Silhouette score is 2 because it has a score of more than 0.5 while based on the Elbow method is 4. In this analysis, 2 clusters were selected.

CUSTOMERS SEGMENTATION PAIR PLOT



AVERAGE METRICS PER CLUSTER

Cluster	Amount	Profit	Order ID	Quantity
0	697.442748	17.622137	2.862595	10.335878
1	3446.500000	437.108108	10.135135	39.283784

Cluster 0 (Blue)

- Represent customers with lower spending (Amount), lower profit contributions, and smaller quantities purchased.
- These customers might be infrequent buyers or price-sensitive customers and targets for engagement strategies to increase order size and frequency

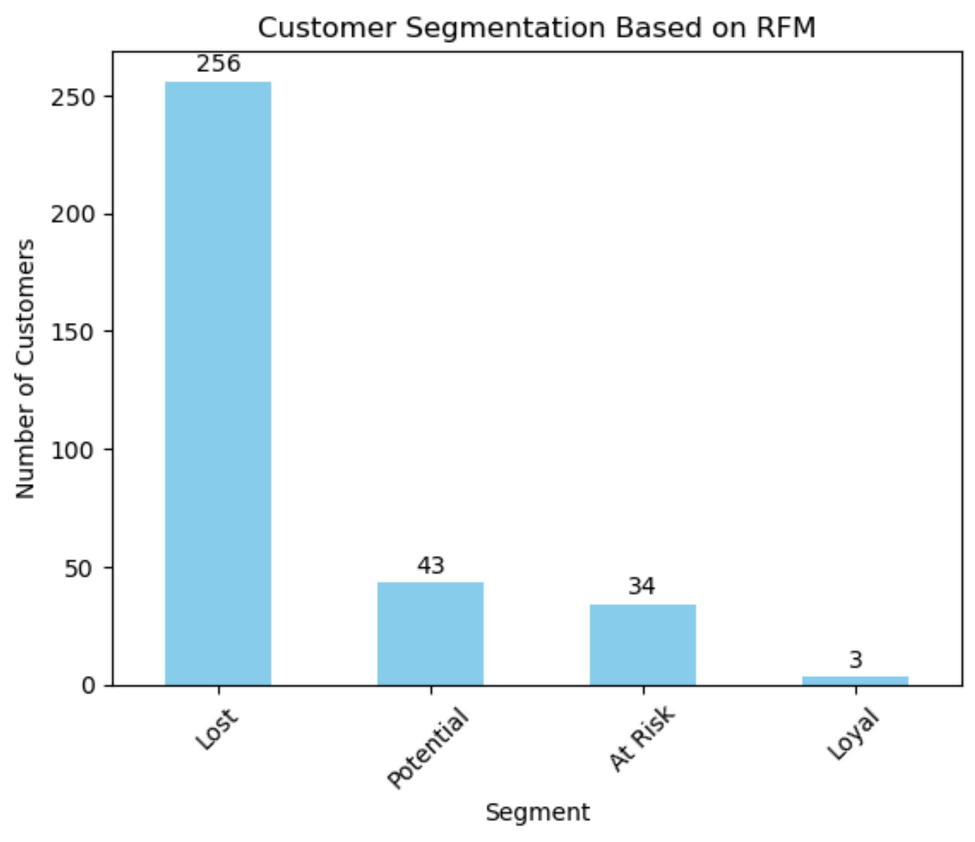
Cluster 1 (Purple)

- Represents customers with higher spending, higher profit contribution, and larger quantity purchased
- These customers are high-value customers that can be focused on for retention and loyalty strategies.

RFM (RECENCY, FREQUENCY, MONETARY) ANALYSIS

RFM is an analytical method for customer segmentation based on three main metrics:

- Recency (R): Measures the time since the last transaction made by the customer. The more recent the transaction, the higher the value
- Frequency (F): Measures how often a customer makes a transaction within a certain period. The more often the transaction is made, the higher the value
- Monetary (M): Measures the total value of money spent by the customer. The larger the amount, the higher the value.

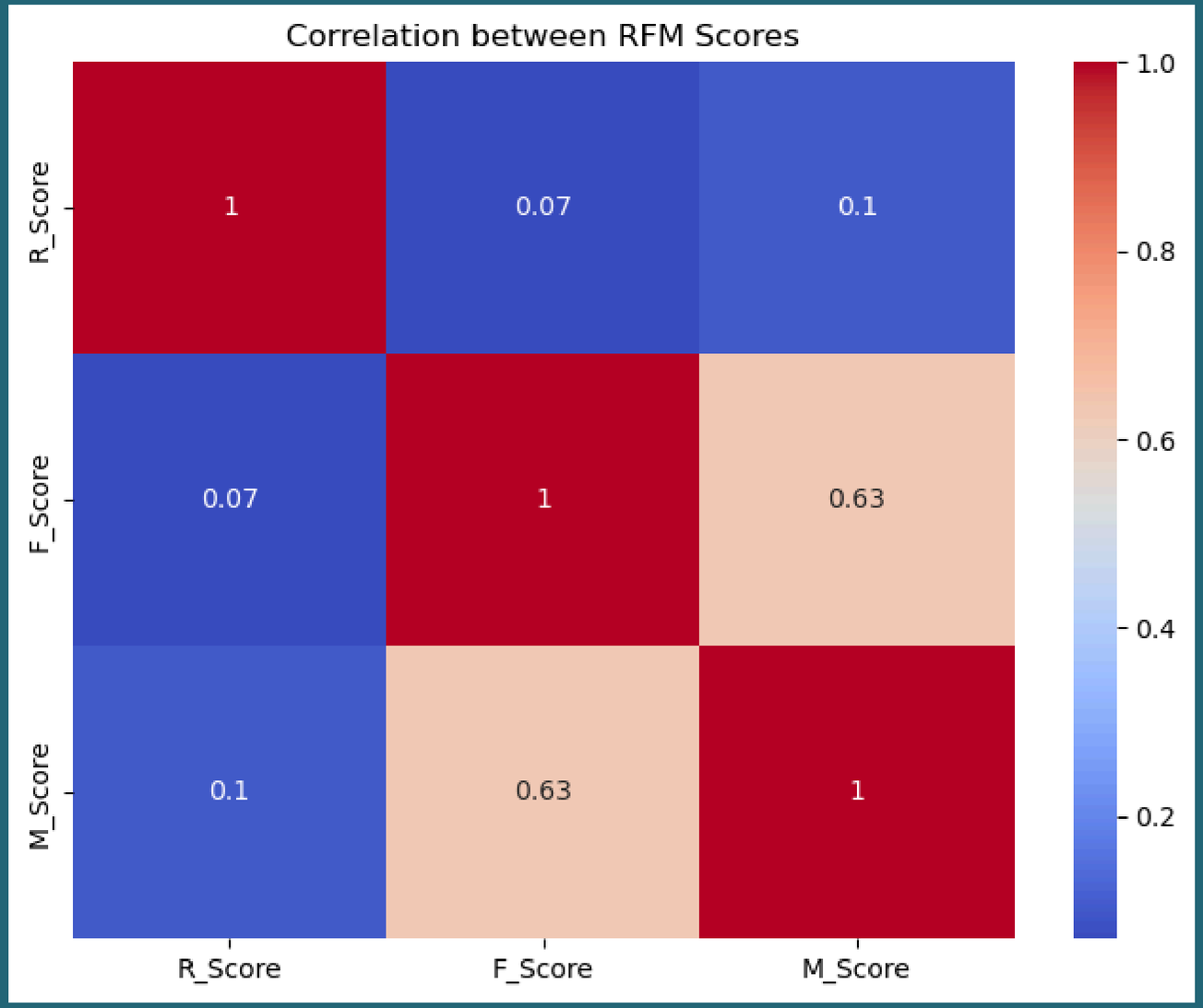


Average for RFM Metrics per Segment			
Segment	Recency	Frequency	Monetary
At Risk	264.823529	8.794118	2695.294118
Lost	165.000000	2.703125	737.238281
Loyal	47.333333	17.333333	5734.333333
Potential	104.953488	10.627907	3260.348837

- Most customers are in the Lost segment indicating many customers are no longer actively transacting
- The Potential segment shows customers with the potential to become more active
- The At Risk segment shows 34 customers who are at risk of no longer transacting
- The Loyal segment of 3 customers shows very few customers who are truly loyal, customers who are frequent and make significant contributions

Recommendations

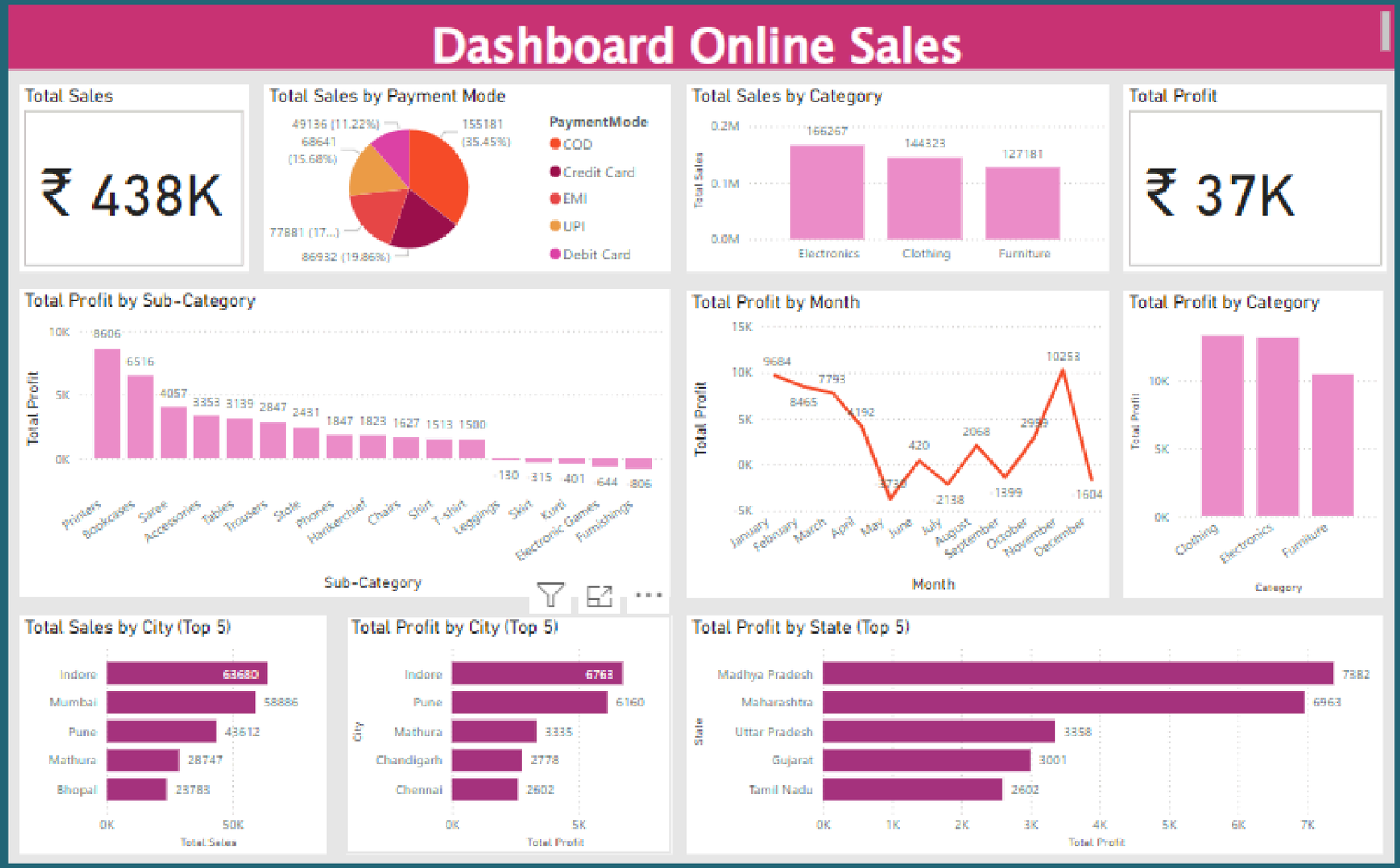
- Customers are dominated by the Lost segment, so reactivation strategies are needed to try to revive the relationship with them.
- The main focus should be on Loyal customers, as they contribute the highest frequency and value of transactions.
- Potential Customers have good recency and monetary value, so an engagement strategy is needed to turn them into loyal customers.
- At Risk customers need attention to prevent them from moving to Lost customers.



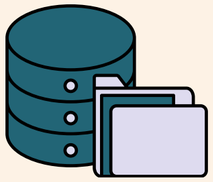
CORRELATION BETWEEN RFM METRICS

- The correlation between R_Score and F_Score is 0.07, meaning that the relationship between Recency and Frequency is almost non-existent. How often customers buy (Frequency) is not related to the last time they bought (Recency).
- The R-Score and M_Score correlation of 0.1 indicates that there is no relationship between the last time of purchase (Recency) and the amount of money spent (Monetary).
- F_Score and M_Score correlation of 0.63 indicates that customers who shop frequently (Frequency) tend to spend more money (Monetary)

VISUALIZATION USING POWER BI



MORE INFORMATION



Online Sales Dataset

Repository GitHub