E-COMMERCE CUSTOMER SEGMENTATION ANALYSIS VIA UNSUPERVISED LEARNING AND MICROSOFT POWER BI

A Project Report
In the partial fulfilment of the award of the Industrial Training certificate of

DATA SCIENCE AND BUSINESS ANALYTICS

Under

Ardent Computech PVT LTD

Submitted by:

RHYTHAM SAHA

TECHNO INDIA UNIVERSITY

KOLKATA, WEST BENGAL

CERTIFICATE FROM THE MENTOR

This is to certify that RHYTHAM SAHA has successfully completed the project titled **"E-commerce customer segmentation via unsupervised learning and Microsoft Power BI"** under my supervision at Ardent Computech Pvt Ltd, as part of the Data Science and Business Analytics Internship Program. The project was carried out during the period June to July 2025, and involved real-world data analysis using Python, Power BI, and unsupervised machine learning techniques.

I hereby acknowledge the originality and effort demonstrated by the student throughout the project.

DATE: **Signature of the Mentor**

ACKNOWLEDGMENT

I would like to express my sincere gratitude to **Mr. Mahendra Dutta**, my project mentor at **Ardent Computech Pvt Ltd**, for his constant guidance, support, and encouragement throughout the duration of this project.

His insights into customer segmentation and data science methodologies were instrumental in shaping my understanding and approach toward real-world problem solving.

I also extend my thanks to the entire team at **Ardent Computech** for providing an excellent learning environment during the internship, which helped me enhance both my technical and analytical skills.

RHYTHAM SAHA

(Note: All entries of the proforma of approval should be filled up with appropriate and complete information of approval in any respect will be summarily rejected.)

1. Name of the Student: **RHYTHAM SAHA**

2. Title of the Project: **E-commerce customer segmentation analysis via unsupervised learning and Microsoft Power BI.**

3. Name and Address of the Guide: **MR. MAHENDRA DUTTA**

4. Educational Qualification of the Guide:

Ph. D	M. Tech	B.E/B. Tech	MCA	M. Sc
☐	☐	☑	☐	☐

5. Software used in the Project:

   a. **Jupyter Notebook**
   b. **Python**
   c. **Numpy, Pandas**
   d. **Matplotlib, Seaborn**
   e. **Scikit-Learn**
   f. **Microsoft Power BI**

Signature of the Guide

Date:

Name: Mr. Mahendra Dutta

Subject Matter Expert

Signature, Designation, Stamp of the Project Proposal Evaluator

APPROVED NOT APPROVED

SELF-CERTIFICATE

I, Rhytham Saha, hereby declare that the project report titled "**E-commerce Customer Segmentation via Unsupervised Learning and Microsoft Power BI**" has been completed by me as part of my Data Science and Business Analytics Internship at Ardent Computech Pvt Ltd. This report is an original piece of work carried out by me under the guidance of Mr. Mahindra Datta during the period of June to July 2025.

All data analysis, visualizations, and interpretations presented in this project are based on genuine efforts, using Python and Microsoft Power BI to derive actionable business insights through unsupervised machine learning techniques.

Name of the Student: Rhytham Saha

Signature of the Student:

RHYTHAM SAHA

CERTIFICATE BY GUIDE

This is to certify that this project entitled "**E-commerce Customer Segmentation via Unsupervised Learning and Microsoft Power BI**" submitted in partial fulfilment of the Internship certificate through Ardent Computech PVT LTD, done by the student is an authentic work carried out under my guidance & best of our knowledge and belief.

Signature of the student
Date:

Signature of the Guide
Date:

CONTENTS

1. ABSTRACT

In today's competitive digital marketplace, understanding customer behavior is crucial for driving targeted marketing and personalized experiences. This project, titled "E-commerce Customer Segment Analysis via Unsupervised Learning and Microsoft Power BI", focuses on leveraging data-driven techniques to identify meaningful customer segments from transactional data.

Using unsupervised machine learning techniques—primarily K-Means clustering—customers were grouped based on key behavioral and demographic features such as purchase frequency, recency, monetary value, and product categories. Python was used for data preprocessing, clustering, and exploratory analysis, while Power BI enabled dynamic and interactive visualization of customer segments, trends, and key performance indicators.

The analysis revealed distinct customer clusters, each with unique characteristics and purchasing patterns, helping stakeholders understand the value of segment-specific strategies. These insights can be used to enhance customer retention, improve marketing ROI, and support strategic decision-making in e-commerce operations. This project showcases the power of combining machine learning with modern business intelligence tools for effective customer analytics.

1. INTRODUCTION

In the era of data-driven decision making, the e-commerce industry stands at the forefront of leveraging customer data to gain strategic insights. As online retail platforms accumulate vast amounts of transactional and behavioral data, it becomes increasingly important for businesses to understand their customer base in a more nuanced and actionable manner. One of the most effective approaches to achieving this is through customer segmentation.

Customer segmentation enables organizations to categorize their customer base into distinct groups based on shared characteristics, preferences, or behaviors. This project focuses on applying unsupervised machine learning, specifically K-Means clustering, to segment e-commerce customers without relying on predefined labels. This technique uncovers natural groupings in the data, offering a deeper, unbiased understanding of consumer behavior.

To complement the analytical insights, Microsoft Power BI is utilized for data visualization and dashboard creation. Power BI facilitates the translation of complex analytical results into intuitive, interactive visuals that support informed business decisions.

The primary goal of this project is to identify key customer segments within an e-commerce dataset, analyze their characteristics, and present insights that can inform targeted marketing strategies, enhance customer satisfaction, and drive overall business performance.

2. RELATED WORKS

Customer segmentation has long been a cornerstone of marketing strategy, and recent advancements in data science have significantly enhanced its precision and effectiveness. Numerous studies and industry applications have demonstrated the value of machine learning—particularly unsupervised learning—in discovering meaningful customer patterns from large datasets.

Several research papers and commercial case studies have successfully applied K-Means clustering and other unsupervised learning techniques such as DBSCAN, hierarchical clustering, and PCA (Principal Component Analysis) for customer segmentation in retail and e-commerce sectors. These methods allow businesses to move beyond traditional RFM (Recency, Frequency, Monetary) analysis and adopt more scalable, flexible, and data-driven approaches.

For instance, publications in journals like Expert Systems with Applications and Decision Support Systems highlight the effectiveness of K-Means in identifying clusters that reveal customer loyalty, spending behavior, and product preferences. Additionally, many practitioners have emphasized the importance of data preprocessing—such as normalization and dimensionality reduction—to improve clustering accuracy and interpretability.

On the visualization front, tools like Microsoft Power BI, Tableau, and QlikView have been widely adopted to present segmentation results in an interactive and accessible format for business users. Power BI, in particular, offers seamless integration with Python scripts and SQL databases, making it an ideal tool for combining analytics with real-time business intelligence.

4. LITERATURE REVIEW

Customer segmentation has been a focal point in marketing and analytics research, especially with the increasing availability of large-scale e-commerce datasets. The literature reveals a growing emphasis on using unsupervised machine learning methods for uncovering latent customer groupings, offering a more dynamic alternative to traditional rule-based segmentation.

RFM (Recency, Frequency, Monetary) analysis, once the standard in segmentation, is often enhanced or replaced by clustering algorithms to derive deeper insights. Tsiptsis and Chorianopoulos (2009) demonstrated that clustering customers based on behavioral data could significantly improve campaign targeting in retail sectors. Similarly, Chen et al. (2012) used K-Means clustering to segment online shoppers and found that it led to more accurate personalization strategies.

K-Means clustering is frequently cited due to its simplicity, scalability, and effectiveness in partitioning large datasets. Studies such as Suryawanshi and Gaikwad (2017) highlight its application in e-commerce platforms for identifying high-value customers and tailoring promotions accordingly. However, researchers also note limitations, such as sensitivity to initial centroids and the need to predefine the number of clusters, which can be mitigated through methods like the Elbow Method or Silhouette Score.

In terms of visualization and reporting, business intelligence tools like Microsoft Power BI have gained traction for operationalizing machine learning insights. As per Microsoft Power BI White Papers (2020), the tool's integration with Python and R allows seamless modeling and visualization, enabling decision-makers to interact with data in real time.

Recent academic works also emphasize the importance of data preprocessing, such as feature scaling and outlier treatment, to enhance clustering accuracy. Integrating machine learning with interactive dashboards aligns with the trend toward automated, insight-driven business solutions.

This literature review underpins the current project by validating the effectiveness of K-Means clustering for customer segmentation and establishing Power BI as a robust platform for visualization and decision support.

5. PROBLEM STATEMENT

Task 1: Exploratory Data Analysis (EDA) and Business Insights 1. Perform EDA on the provided dataset.

2. Derive at least 5 business insights from the EDA.

○ Write these insights in short point-wise sentences (maximum 100 words per insight). Deliverables:

● A Jupyter Notebook/Python script containing your EDA code.

● A PDF report with business insights (maximum 500 words).

Task 2: Lookalike Model Build a Lookalike Model that takes a user's information as input and recommends 3 similar customers based on their profile and transaction history. The model should:

● Use both customer and product information.

● Assign a similarity score to each recommended customer. Deliverables:

● Give the top 3 lookalikes with there similarity scores for the first 20 customers (CustomerID: C0001 - C0020) in Customers.csv. Form an "Lookalike.csv" which has just one map: Map>

● A Jupyter Notebook/Python script explaining your model development

Evaluation Criteria:

● Model accuracy and logic.

● Quality of recommendations and similarity scores.

Task 3: Customer Segmentation / Clustering Perform customer segmentation using clustering techniques. Use both profile information (from Customers.csv) and transaction information (from Transactions.csv).

● You have the flexibility to choose any clustering algorithm and any number of clusters in between(2 and 10)

● Calculate clustering metrics, including the DB Index(Evaluation will be done on this).

● Visualise your clusters using relevant plots. Deliverables:

● A report on your clustering results, including: ○ The number of clusters formed.

○ DB Index value.

○ Other relevant clustering metrics.

● A Jupyter Notebook/Python script containing your clustering code.

Evaluation Criteria:

● Clustering logic and metrics.

● Visual representation of clusters.

Task 4: Show Business Analysis Report on Microsoft Power BI Dashboard.

6. **OBJECTIVE**

The primary objective of this project is to perform a comprehensive customer segmentation analysis for an e-commerce platform using unsupervised machine learning techniques and visualize the results through Microsoft Power BI. This project aims to uncover distinct customer groups based on their demographic attributes and transactional behaviors, enabling personalized marketing strategies and improved business decision-making. The following sub-objectives outline the methodological approach:

1. Unify Heterogeneous Data Integrate customer profile information (Customers.csv) and transactional data (Transactions.csv) into a unified analytical dataset. This step ensures that all relevant behavioral and demographic attributes are linked at the individual customer level for accurate modeling.

2. Feature Engineering and Preprocessing Construct meaningful features such as Recency, Frequency, Monetary (RFM) scores, average transaction value, and product diversity. Standard preprocessing steps like handling missing values, normalization, and encoding categorical variables are applied to prepare the data for clustering algorithms.

3. Dimensionality Reduction Reduce the complexity of high-dimensional data using techniques like Principal Component Analysis (PCA). This helps in improving clustering performance, mitigating multicollinearity, and enhancing the interpretability of the clusters in visualizations.

4. Multi-Model Clustering Implement and compare multiple unsupervised clustering algorithms including K-Means, DBSCAN, and Hierarchical Clustering. The goal is to experiment with different models and determine which method most effectively segments customers based on intrinsic patterns in the data.

5. Robust Evaluation Evaluate the clustering models using quantitative metrics such as Davies-Bouldin Index (DBI), Silhouette Score, and Inertia to assess intra-cluster cohesion and inter-cluster separation. The evaluation will guide the selection of the optimal clustering technique and number of clusters.

6. Cluster Profiling Analyze and interpret the characteristics of each customer segment by profiling them based on behavioral and demographic attributes. Each cluster is assigned a business-relevant label (e.g., “High-Value Loyalists”, “Low-Spend Inactives”), enabling strategic targeting and actionable insights.

This objective-driven framework ensures a methodical, data-centric approach to customer segmentation that is scalable, interpretable, and valuable for business applications.

7. DATA COLLECTION

7.1 SOURCE OF DATASET:

Customers.csv: https://drive.google.com/file/d/1bu_--mo79VdUG9oin4ybfFGRUSXAe-WE/view?usp=sharing

Products.csv : https://drive.google.com/file/d/1IKuDizVapw-hyktwfpoAoaGtHtTNHfd0/view?usp=sharing

Transactions.csv : https://drive.google.com/file/d/1saEqdbBB-vuk2hxoAf4TzDEsykdKlzbF/view?usp=sharing

7.2 DATASET DESCRIPTION:

- **Customers.csv:** This dataset contains information about each individual customer registered on the e-commerce platform.

COLUMN	DESCRIPTION
CustomerID	Unique identifier for each customer
CustomerName	Full name of the customer
Region	Geographic region where the customer resides
SignupDate	Date when the customer registered on the platform

Total Records: 200 customers
Use: For demographic segmentation and mapping transactions to customers.

- **Product.csv:** This dataset includes product-level information used to enrich transactional data.

COLUMN	DESCRIPTION
ProductID	Unique identifier for each product
ProductName	Name or title of the product
Category	Product category
Price	Unit price of the product

Total Records: 100 products
Use: For analyzing customer preferences and product-level trends in transactions.

- **Transaction.csv:** This dataset captures all e-commerce transactions made by customers.

COLUMN	DESCRIPTION
TransactionID	Unique identifier for each transaction
CustomerID	ID of the customer who made the purchase
ProductID	ID of the product purchased
TransactionDate	Timestamp of when the transaction occured
Quantity	Number of items bought in the transaction
TotalValue	Total value of the transaction
Price	Unit price of the product

Total Records: 1,000 transactions
Use: For behavior analysis, RFM scoring, clustering, and building the lookalike model.

8. DATA CLEANING & PREPROCESSING

- **8.1. Data Shape and Structure Inspection:**

```
print(df1.info(), end='\n\n')
print(df2.info(), end='\n\n')
print(df3.info(), end='\n\n')
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 4 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   CustomerID    200 non-null    object
 1   CustomerName  200 non-null    object
 2   Region        200 non-null    object
 3   SignupDate    200 non-null    object
dtypes: object(4)
memory usage: 6.4+ KB
None

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100 entries, 0 to 99
Data columns (total 4 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   ProductID    100 non-null    object
 1   ProductName  100 non-null    object
 2   Category     100 non-null    object
 3   Price        100 non-null    float64
dtypes: float64(1), object(3)
memory usage: 3.3+ KB
None

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 7 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   TransactionID    1000 non-null   object
 1   CustomerID       1000 non-null   object
 2   ProductID        1000 non-null   object
 3   TransactionDate  1000 non-null   object
 4   Quantity         1000 non-null   int64
 5   TotalValue       1000 non-null   float64
 6   Price            1000 non-null   float64
dtypes: float64(2), int64(1), object(4)
memory usage: 54.8+ KB
None
```

```
print(df1.shape, end='\n\n')
print(df2.shape, end='\n\n')
print(df3.shape, end='\n\n')
```

```
(200, 4)

(100, 4)

(1000, 7)
```

The .shape function was used to inspect the dimensions of each dataset.

.info() was used to check data types and non-null values.

📌 Purpose: To verify successful data loading and ensure that each column had the expected type (e.g., string, numeric).

- ## 8.2. Missing Value Detection:

```
print(df1.isnull().sum(), end='\n\n')
print(df2.isnull().sum(), end='\n\n')
print(df3.isnull().sum(), end='\n\n')
```

```
CustomerID      0
CustomerName    0
Region          0
SignupDate      0
dtype: int64

ProductID      0
ProductName    0
Category       0
Price          0
dtype: int64

TransactionID      0
CustomerID         0
ProductID          0
TransactionDate    0
Quantity           0
TotalValue         0
Price              0
dtype: int64
```

Each dataset was checked for null (missing) values.

Columns with missing data were identified for future handling or imputation.

📌 Purpose: To assess data completeness and plan necessary data filling or cleaning steps.

- ## 8.3. Uniqueness and Duplicate Check:

```
print(df1.nunique(), end='\n\n')
print(df2.nunique(), end='\n\n')
print(df3.nunique(), end='\n\n')
```

```
CustomerID      200
CustomerName    200
Region            4
SignupDate      179
dtype: int64

ProductID      100
ProductName     66
Category         4
Price          100
dtype: int64

TransactionID      1000
CustomerID          199
ProductID           100
TransactionDate    1000
Quantity              4
TotalValue          369
Price               100
dtype: int64
```

Counted unique values in each column to identify:

Categorical columns

Potential duplicate entries

Primary identifiers (e.g., CustomerID, ProductID)

📌 Purpose: To ensure data integrity and prepare for grouping or joining.

- **Categorical Frequency Analysis:**

```
print('\nCustomers by Region: ')
df1['Region'].value_counts()
```

```
Customers by Region:
Region
South America     59
Europe            50
North America     46
Asia              45
Name: count, dtype: int64
```

```
print('\nProduct price summary : ')
df2['Price'].describe()
```

```
Product price summary :
count    100.000000
mean     267.551700
std      143.219383
min       16.080000
25%      147.767500
50%      292.875000
75%      397.090000
max      497.760000
Name: Price, dtype: float64
```

```
print('\nProduct by  category: ')
df2['Category'].value_counts()
```

```
Product by  category:
Category
Books          26
Electronics    26
Clothing       25
Home Decor     23
Name: count, dtype: int64
```

```
print('\nProduct by  productname: ')
df2['ProductName'].describe()
```

```
Product by  productname:
count                   100
unique                   66
top       ActiveWear Smartwatch
freq                      4
Name: ProductName, dtype: object
```

Value counts were used to examine common values in Region, Category, and ProductName.

📌 Purpose: To identify dominant categories and assess class imbalance.

- **Descriptive Statistical Analysis:**

```
df1.describe(include='all')
```

	CustomerID	CustomerName	Region	SignupDate
count	200	200	200	200
unique	200	200	4	179
top	C0001	Lawrence Carroll	South America	2024-11-11
freq	1	1	59	3

```
df2.describe(include='all')
```

	ProductID	ProductName	Category	Price
count	100	100	100	100.000000
unique	100	66	4	NaN
top	P001	ActiveWear Smartwatch	Books	NaN
freq	1	4	26	NaN
mean	NaN	NaN	NaN	267.551700
std	NaN	NaN	NaN	143.219383
min	NaN	NaN	NaN	16.080000
25%	NaN	NaN	NaN	147.767500
50%	NaN	NaN	NaN	292.875000
75%	NaN	NaN	NaN	397.090000
max	NaN	NaN	NaN	497.760000

```
df3.describe(include='all')
```

	TransactionID	CustomerID	ProductID	TransactionDate	Quantity	TotalValue	Price
count	1000	1000	1000	1000	1000.000000	1000.000000	1000.00000
unique	1000	199	100	1000	NaN	NaN	NaN
top	T00001	C0109	P059	2024-08-25 12:38:23	NaN	NaN	NaN
freq	1	11	19	1	NaN	NaN	NaN
mean	NaN	NaN	NaN	NaN	2.537000	689.995560	272.55407
std	NaN	NaN	NaN	NaN	1.117981	493.144478	140.73639
min	NaN	NaN	NaN	NaN	1.000000	16.080000	16.08000
25%	NaN	NaN	NaN	NaN	2.000000	295.295000	147.95000
50%	NaN	NaN	NaN	NaN	3.000000	588.880000	299.93000
75%	NaN	NaN	NaN	NaN	4.000000	1011.660000	404.40000
max	NaN	NaN	NaN	NaN	4.000000	1991.040000	497.76000

Summary statistics were generated for all datasets, including:

Mean, median, min, max for numeric values

Count, unique, top value for categorical columns

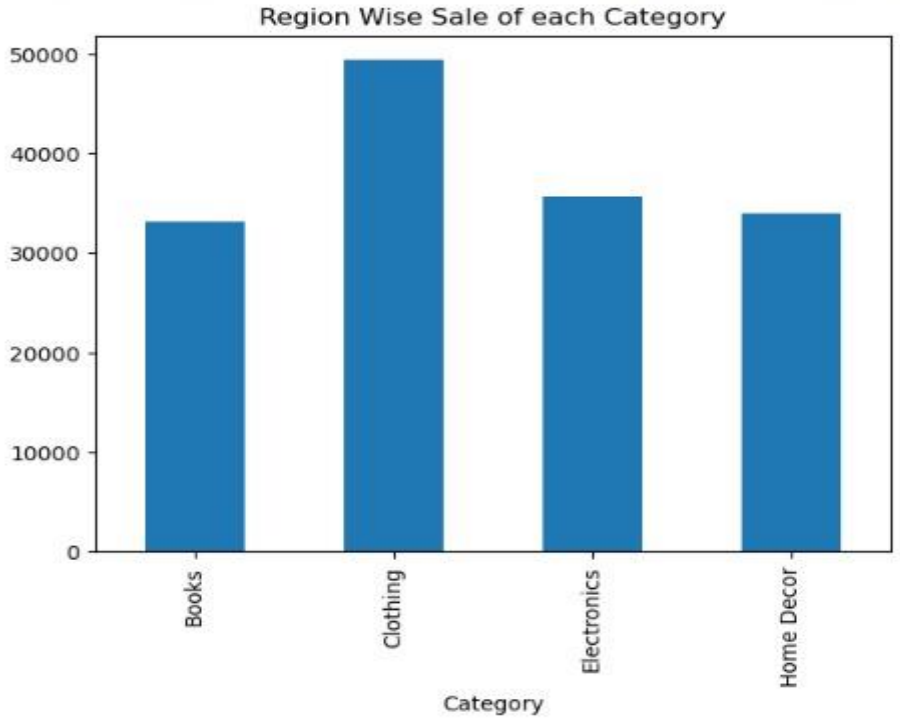
📌 Purpose: To detect outliers, skewed data distributions, and potential data quality issues.

9. EXPLORATORY DATA ANALYSIS (EDA)

- **Asia Region Sale of each Category:-**

```
#Region wise sale of each Category
#1
print("Asia")
category=merge_df[merge_df['Region'] == 'Asia'].groupby('Category')["TotalValue"].sum()
category.plot(kind='bar',title="Region Wise Sale of each Category")
```

```
Asia
<Axes: title={'center': 'Region Wise Sale of each Category'}, xlabel='Category'>
```



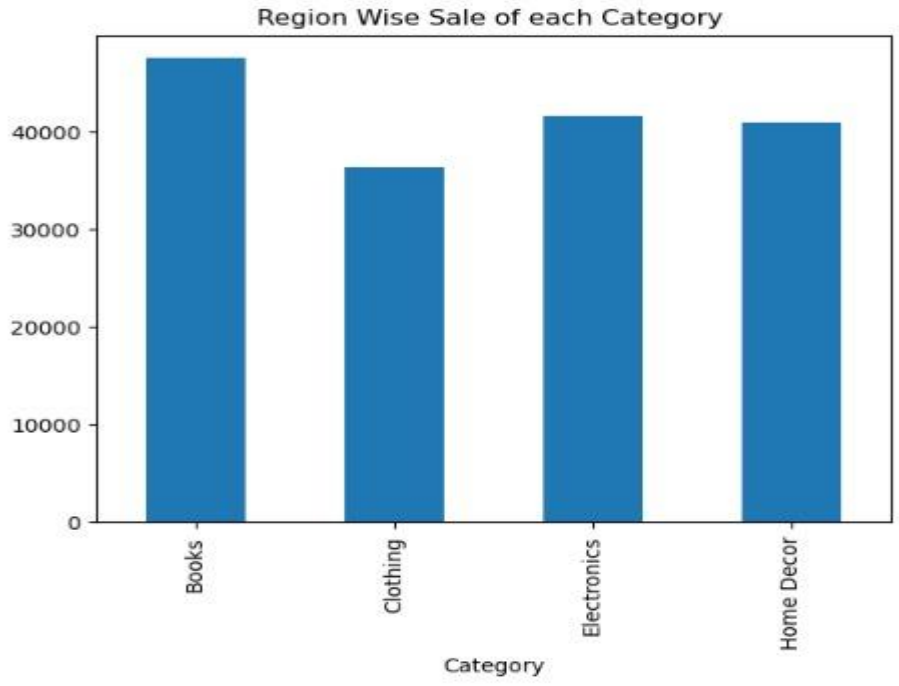
The goal of this analysis is to understand how different product categories are performing in terms of sales specifically within the Asia region. This insight will help in identifying customer preferences and potential focus areas for marketing and supply chain efforts.

- merge_df[merge_df['Region'] == 'Asia']: Filters the dataset to include only transactions from the Asia region.
- .groupby('Category')['TotalValue'].sum(): Groups the filtered data by product Category and calculates the total sales value for each category.
- category.plot(kind='bar'): Plots the resulting data as a bar chart with appropriate title.

- **<u>Europe Region Sale of each Category</u>:-**

```
#4
print("Europe")
category=merge_df[merge_df['Region']=='Europe'].groupby('Category')['TotalValue'].sum()
category.plot(kind="bar",title="Region Wise Sale of each Category ")
```

```
Europe
<Axes: title={'center': 'Region Wise Sale of each Category '}, xlabel='Category'>
```



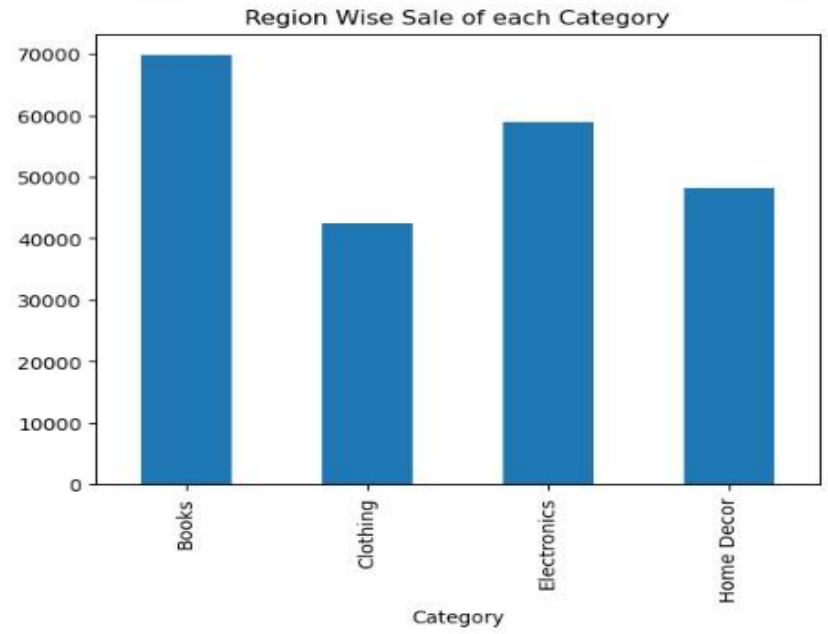
This analysis provides insights into the total sales value per product category specifically in the Europe region. The goal is to understand regional product performance and assist in inventory decisions, targeted marketing, and regional demand planning.

1. Books category leads in Europe, generating the highest total sales among all categories.
2. Electronics and Home Decor follow closely with similar sales volumes, indicating strong consumer interest in both tech and lifestyle segments.
3. Clothing records the lowest sales in the region, suggesting:
    - Lower demand,
    - Pricing issues,
    - Or potentially less variety in this segment.

- **South America Region Sale of each Category:-**

```
#2
print("South America")
category=merge_df[merge_df['Region'] == 'South America'].groupby('Category')["TotalValue"].sum()
category.plot(kind='bar',title="Region Wise Sale of each Category")
```

```
South America
<Axes: title={'center': 'Region Wise Sale of each Category'}, xlabel='Category'>
```



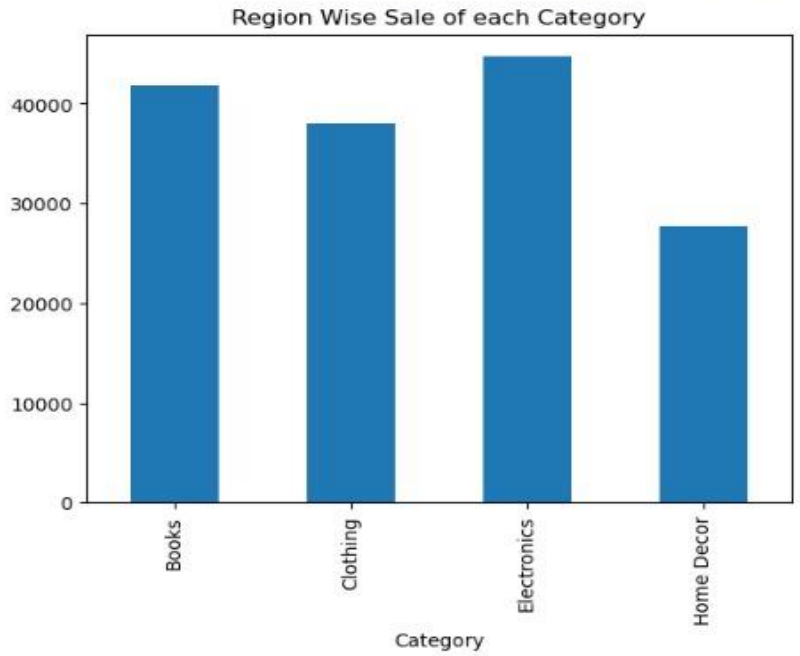
This analysis evaluates the sales performance across various product categories in the South America region. It helps to identify which categories are most popular in this market and prioritize efforts accordingly.

1. Books have the highest sales in South America, contributing significantly to total revenue.
    - This suggests that books are a dominant product in this region and could be a focus for future marketing and promotions.
2. Electronics comes in second, indicating strong demand in the tech and gadget market, but it is not as dominant as books.
3. Clothing and Home Decor have comparable sales, with Home Decor slightly lagging behind.
    - This suggests a diverse customer base in South America, with balanced demand across various product types.

- **North America Region Sale of Each Category:-**

```
#3
print("North America")
category=merge_df[merge_df['Region']=='North America'].groupby('Category')['TotalValue'].sum()
category.plot(kind='bar',title="Region Wise Sale of each Category")
```

```
North America
<Axes: title={'center': 'Region Wise Sale of each Category'}, xlabel='Category'>
```



1. Region Filtered:

The code filters the DataFrame merge_df to include only rows where Region == 'North America'.

2. Group & Aggregate:

It groups the filtered data by Category and calculates the sum of TotalValue for each category.

This means it computes the total sales for each product category in the North America region.

3. Bar Plot:

A bar chart is plotted to visualize these total sales for each category

☑ Chart Interpretation

From the bar chart:

- Electronics: Has the highest total sales in North America.
- Books: Are the second highest.
- Clothing: Comes next.
- Home Decor: Has the lowest sales among the categories.

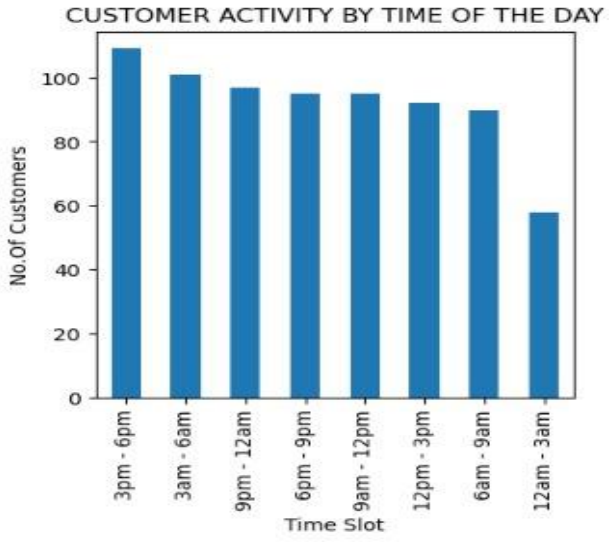
- **Customer Activity by Time of the Day:-**

```
active_time_period = df3.groupby('Time Slot')['CustomerID'].nunique()
active_time_period = active_time_period.sort_values(ascending=False)
```

```
plt.figure(figsize=(4,4))
active_time_period.plot(kind='bar')

plt.title('CUSTOMER ACTIVITY BY TIME OF THE DAY')
plt.ylabel('No.Of Customers')
```

```
Text(0, 0.5, 'No.Of Customers')
```



The purpose of this analysis is to identify which times of the day have the highest customer activity, measured by the number of unique customers active in each time slot. This helps businesses optimize marketing campaigns, ad placements, and support availability during peak activity periods.

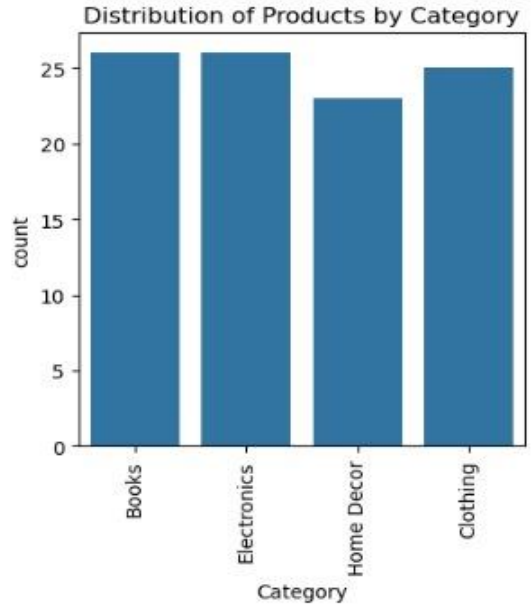
• groupby('Time Slot')['CustomerID'].nunique() counts the number of unique customers active in each defined time slot.

- The result is sorted in descending order to identify the most active periods.
- A bar chart is generated to visually represent customer counts by time slot.

1. Peak Hours:
   - 3 PM – 6 PM has the highest customer activity, making it the most strategic time for launching promotional campaigns or sending marketing emails.
2. Evening to Late Night Activity:
   - Time slots like 6 PM – 9 PM and 9 PM – 12 AM also show strong engagement, indicating high customer presence after work hours.
3. Late Night Dip:
   - 12 AM – 3 AM shows the lowest customer activity, suggesting this is a less effective time for targeting users.
4. Consistent Usage:
   - All time slots except 12 AM – 3 AM maintain a relatively stable range (80–110 customers), indicating a steady level of user interaction throughout the day.

- **Distribution of Products by Category:-**

```
plt.figure(figsize=(4,4))
sns.countplot(x='Category', data=df2)
plt.xticks(rotation= 90)
plt.title('Distribution of Products by Category')
plt.show()
```



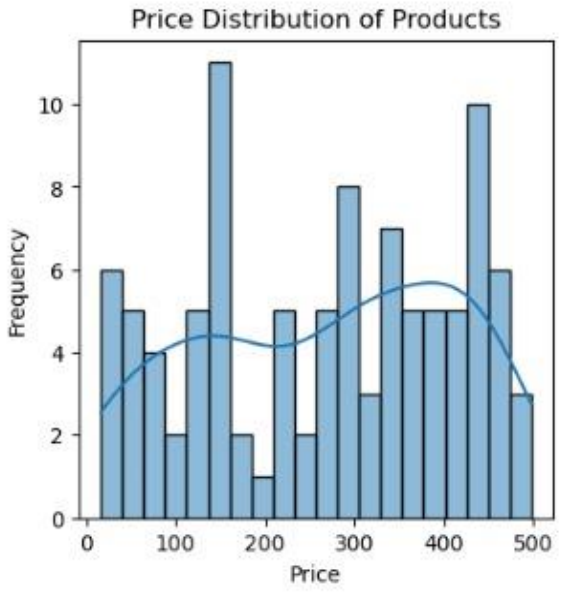
This analysis aims to explore the availability of products across different categories in the dataset. It helps determine whether the product distribution is balanced or skewed toward certain categories.
sns.countplot(x='Category', data=df2) : This line creates a bar chart showing the number of products in each product category (Books, Electronics, Home Decor, Clothing).

- The chart uses count as the y-axis and Category as the x-axis.
- plt.xticks(rotation=90) ensures category labels are vertical for readability.
- plt.title(...) adds a meaningful chart title.

1. Balanced Distribution:
   - All four categories — Books, Electronics, Home Decor, and Clothing — have comparable product counts, with values ranging from ~23 to 27.
   - This indicates a well-balanced product catalog, which is good for maintaining diversity in offerings.
2. Slightly Lower in Home Decor:
   - Home Decor has slightly fewer products than the other categories.
   - It may suggest an opportunity to expand product listings in this category to match others.
3. No Dominance:
   - No single category overwhelmingly dominates the product listings.
   - This suggests the company is targeting multiple customer segments and maintaining category parity.

- **Price Distribution of Products:-**

```
plt.figure(figsize=(4,4))
sns.histplot(df2['Price'], kde=True, bins=20)
plt.title('Price Distribution of Products')
plt.xlabel('Price')
plt.ylabel('Frequency')
plt.show()
```



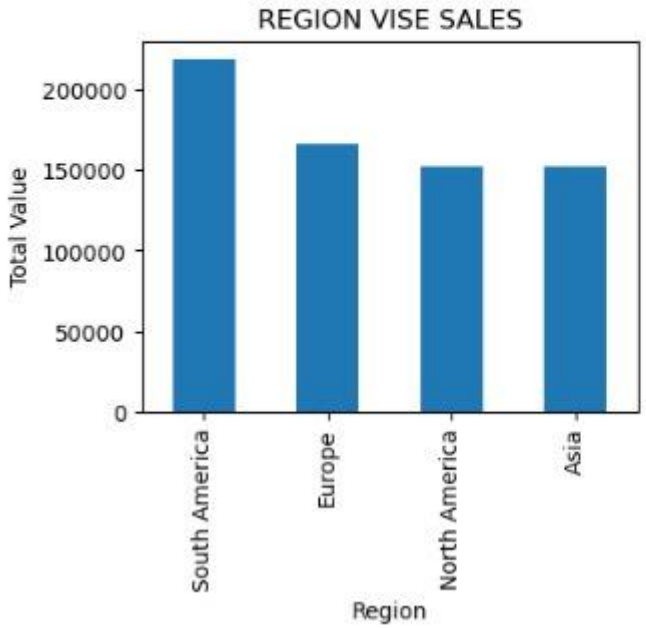
To analyze the spread and frequency of product prices in the dataset. This distribution helps identify pricing patterns and determine if the catalog is skewed toward low-end or high-end items.

1. Wide Price Range:
    - Product prices are distributed from around ₹0 to ₹500, indicating a diverse pricing strategy.
2. Moderate Skew:
    - The KDE curve shows multiple peaks, suggesting price clustering at specific levels (e.g., around ₹100, ₹250, and ₹400).
3. Popular Price Points:
    - The highest frequency of products falls roughly around ₹100 and ₹400, indicating these are common pricing brackets used in the product lineup.
4. Balanced Distribution:
    - The distribution isn't heavily skewed, suggesting the platform offers a balanced mix of low, mid, and high-priced products.

- **Region wise Sales:-**

```
regionvise_sales = merged_df1_df3.groupby('Region')['TotalValue'].sum().sort_values(ascending=False)
plt.figure(figsize=(4,3))
regionvise_sales.plot( kind='bar')
plt.title('REGION VISE SALES')
plt.ylabel('Total Value')
```

```
Text(0, 0.5, 'Total Value')
```



To analyze and compare the total sales value across different global regions, helping identify which geographic markets are performing best and where there may be opportunities for growth.

1. South America is the top-performing region, with the highest total sales.

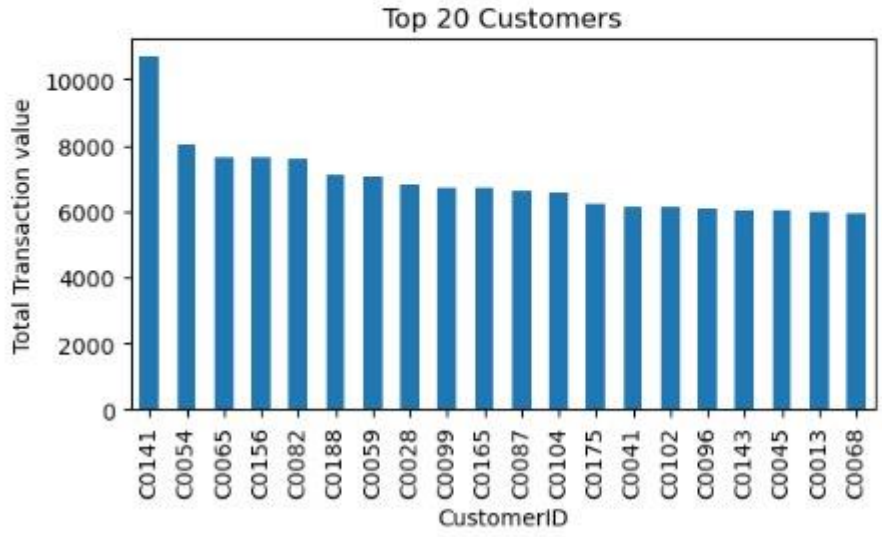
2. Europe, North America, and Asia follow, with relatively close sales figures.

3. Despite being a large market, Asia has the lowest total sales, suggesting:

- Under-penetration,
- Pricing mismatch,
- Or potential for market expansion.

- **<u>Top 20 Customers(Total Transaction Value)</u>:-**

```
customer_sales = merged_df1_df3.groupby('CustomerID')['TotalValue'].sum().sort_values(ascending=False)
plt.figure(figsize=(6,3))
customer_sales.head(20).plot(kind='bar')
plt.title('Top 20 Customers')
plt.ylabel('Total Transaction value')
plt.show()
```



To identify and analyze the most valuable customers based on their total purchase value, enabling the business to focus on high-value customer retention and loyalty strategies.

- Group by CustomerID: Aggregates the total transaction value (TotalValue) for each customer.
- Descending: Highlights customers with the highest lifetime value.
- Plot: Shows the top 20 customers using a vertical bar chart.

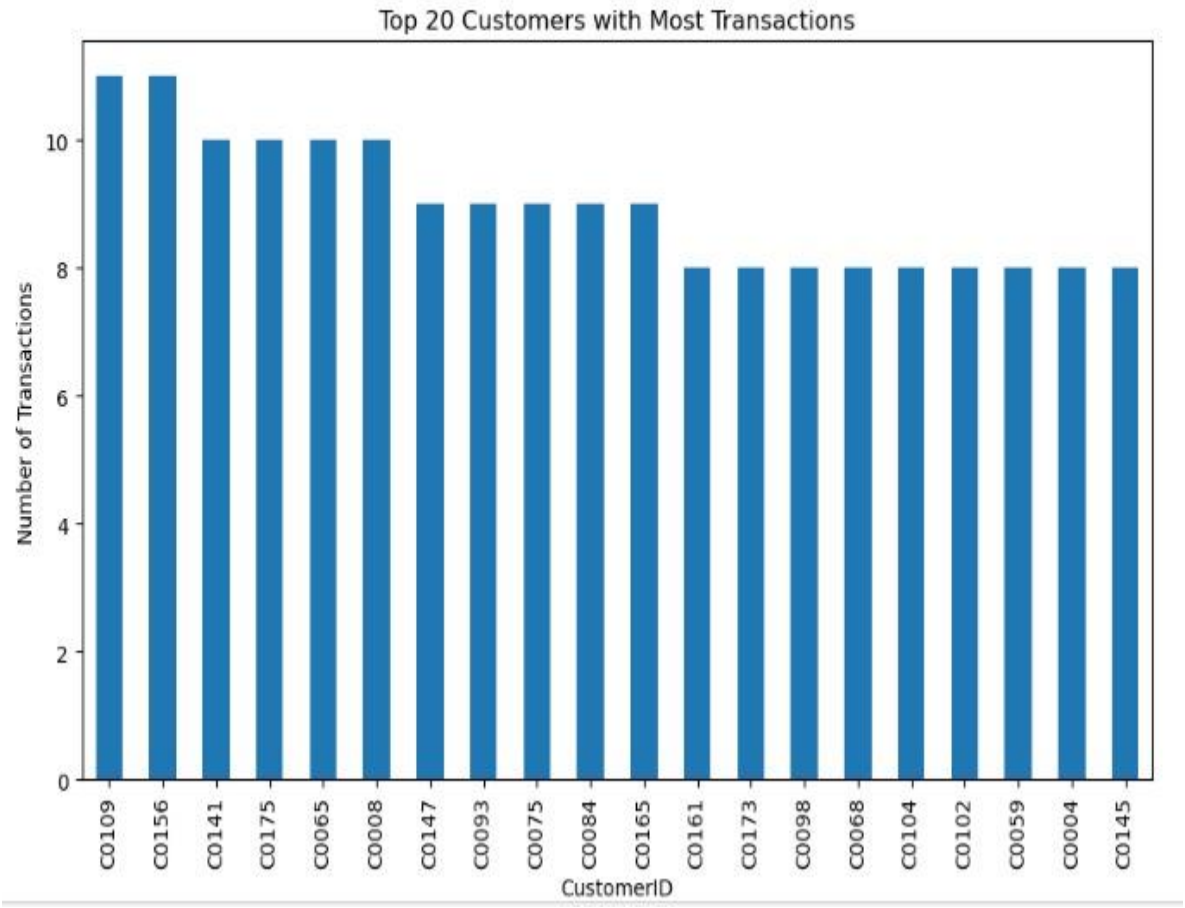
_____

☑ Insights from the Chart:

1. Customer C0141 is the highest spender, with a total transaction value exceeding ₹10,000.
2. The top 5 customers (C0141, C0095, C0156, C0082, C0108) contribute significantly more than the rest.
3. Although there's a gradual decline, all top 20 customers show substantial purchase values, making them prime targets for loyalty programs.

- **Top 20 Customers with Most Transactions:-**

```
plt.figure(figsize=(10, 6))
df3['CustomerID'].value_counts().head(20).plot(kind='bar')
plt.title('Top 20 Customers with Most Transactions')
plt.xlabel('CustomerID')
plt.ylabel('Number of Transactions')
plt.show()
```



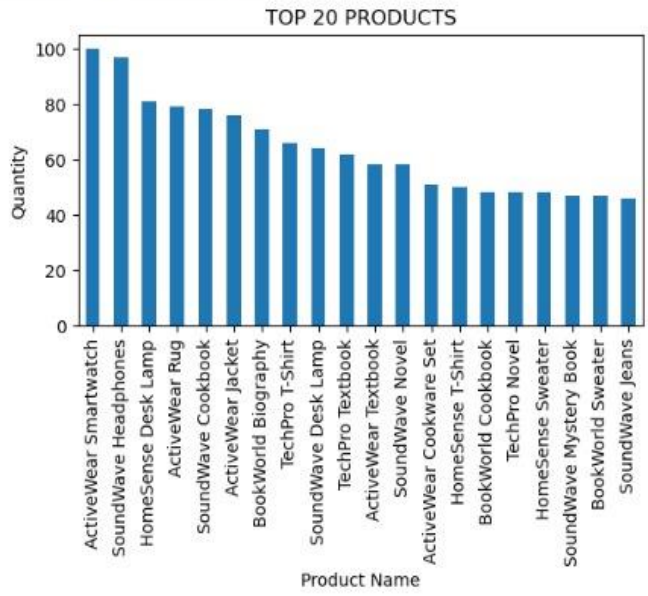
To identify the customers with the highest frequency of purchases, which provides insights into customer engagement, loyalty, and retention potential.

1. Customer C0109 and C0156 made the most transactions (11+ each), reflecting high purchase frequency.
2. Several customers (e.g., C0141, C0175, C0005, etc.) consistently made 10 or more purchases, indicating strong engagement.
3. The remaining top customers made 8 to 9 transactions, still showing a high level of activity compared to the broader base.

- **Top 20 Products Quantity Wise:-**

```
top_products = merged_df2_df3.groupby('ProductName')['Quantity'].sum().sort_values(ascending=False)
plt.figure(figsize=(6,3))
top_products.head(20).plot(kind='bar')
plt.title('TOP 20 PRODUCTS')
plt.ylabel('Quantity')
plt.xlabel('Product Name')
```

```
Text(0.5, 0, 'Product Name')
```



To identify which products are most popular in terms of quantity sold across the dataset, helping uncover customer preferences and top-selling items.

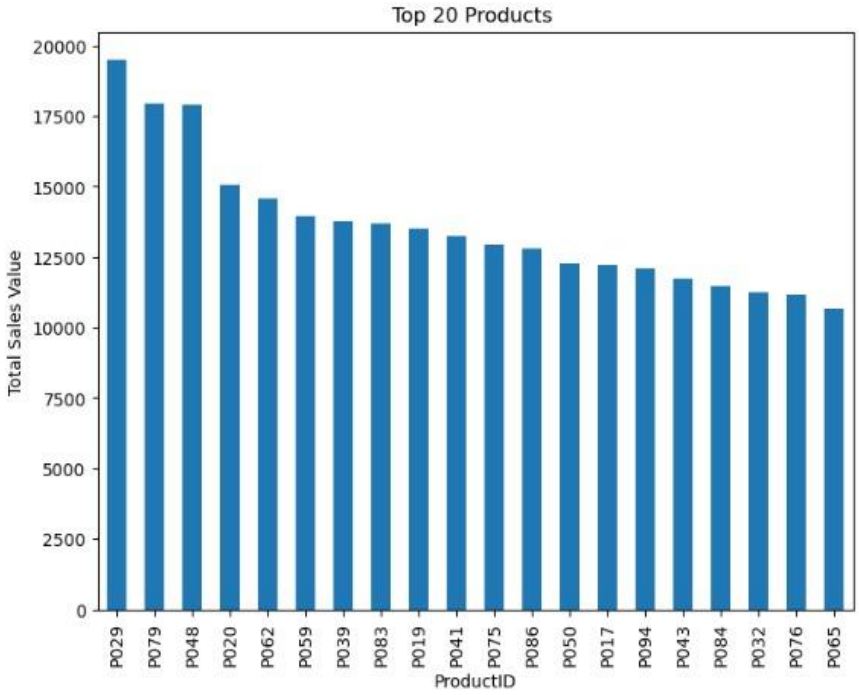
- Group by ProductName: Aggregates total units sold for each product.
- Sort Descending: Ranks products based on quantity sold.
- Bar Plot: Displays the top 20 products with the highest sales volume.

☑ Insights from the Chart:

1. Top Performers:
    - ActiveWear Smartwatch is the top-selling product with nearly 100 units sold.
    - SoundWave Headphones and HomeSense Bedding follow closely, reflecting high customer demand in both electronics and home decor.
2. Category Diversity:
    - The top 20 list includes products from diverse categories:
    - Electronics (e.g., Smartwatch, Headphones)
    - Home Decor (e.g., Rug, Bedding)
    - Books (e.g., Biographies, Novels)
    - Clothing (e.g., Jackets, Jeans)
3. Tech Products Dominate:
    - Multiple products from the SoundWave and ActiveWear series appear, suggesting strong brand or product line appeal.

- **<u>Top 20 Products Total Sales Value Wise</u>:-**

```
product_sales = df3.groupby('ProductID')['TotalValue'].sum().sort_values(ascending=False)
plt.figure(figsize=(8, 6))
product_sales.head(20).plot(kind='bar')
plt.title('Top 20 Products ')
plt.xlabel('ProductID')
plt.ylabel('Total Sales Value')
plt.show()
```



To determine the products that contribute the most to overall revenue, helping guide pricing, marketing, and inventory strategies.

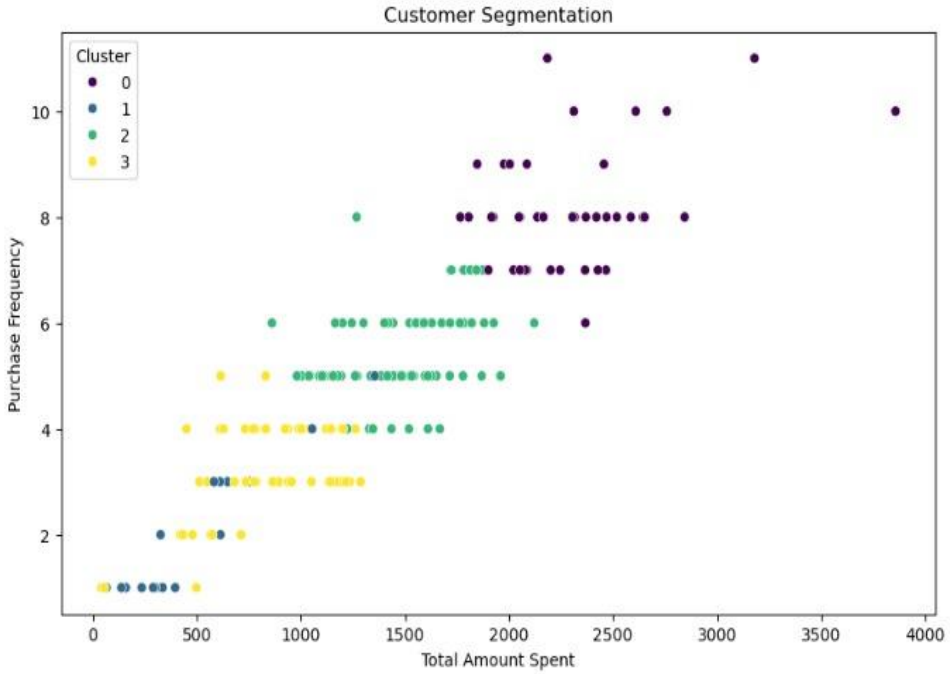
- Group by ProductID: Sums up total revenue (TotalValue) generated by each product.
- Sort Descending: Ranks products from highest to lowest revenue.
- Plot: Visualizes the top 20 revenue-generating products.
-

📈 Insights from the Chart:

1. Product P029 is the top revenue generator, earning close to ₹20,000 in sales.
2. Other high-value contributors include P079, P028, and P062, indicating strong performance both in sales volume and/or price.
3. The revenue difference is gradual but notable—the top 5 products are significantly ahead of the remaining 15.

- **Clustering:-**

```
# Visualizing customer segments
plt.figure(figsize=(10, 6))
sns.scatterplot(x=customer_data['TotalSpent'], y=customer_data['Frequency'], hue=customer_data['Cluster'], palette='viridis')
plt.xlabel('Total Amount Spent')
plt.ylabel('Purchase Frequency')
plt.title('Customer Segmentation')
plt.show()
```

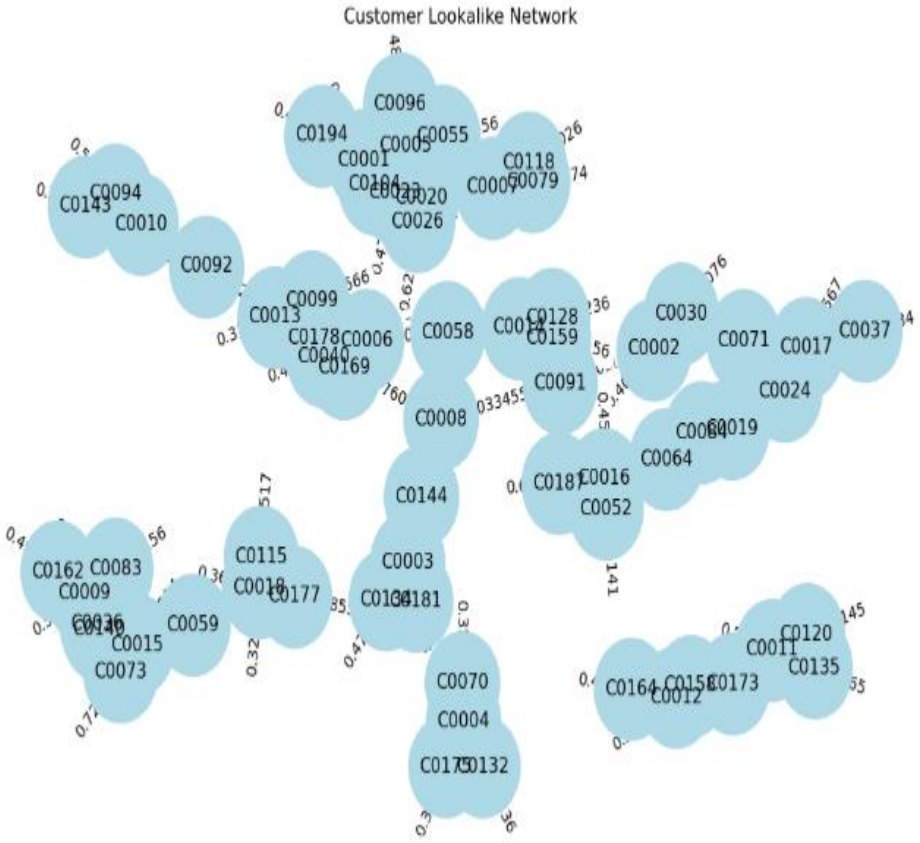


The aim of this visualization is to segment customers based on their total spending and purchase frequency, which are essential behavioral attributes for identifying different types of customers. This helps businesses tailor marketing strategies and resource allocation effectively.

- TotalSpent: Total monetary amount a customer has spent.
- Frequency: Number of times a customer has made purchases.
- Cluster: Cluster label assigned to each customer after applying a clustering algorithm (e.g., KMeans).
- The viridis palette is used to color-code the clusters.

This scatter plot visually distinguishes customer segments based on their spending patterns and frequency of purchases.

- **Lookalike:-**




```
import pandas as pd
import networkx as nx
import matplotlib.pyplot as plt
from sklearn.metrics.pairwise import cosine_similarity
from sklearn.preprocessing import StandardScaler

# Load datasets
customers = pd.read_csv("customers.csv")
products = pd.read_csv("products.csv")
transactions = pd.read_csv("transactions.csv")

# Extract first 20 customer IDs
target_customers = customers.iloc[:20]['CustomerID'].tolist()

# Merge transactions with products
txn_merged = transactions.merge(products, on="ProductID")

# Create a customer-product matrix
customer_product_matrix = txn_merged.pivot_table(index='CustomerID',
                                                 columns='ProductID',
                                                 values='Quantity',
                                                 aggfunc='sum',
                                                 fill_value=0)

# Standardize the matrix
scaler = StandardScaler()
customer_product_matrix_scaled = scaler.fit_transform(customer_product_matrix)

# Compute cosine similarity
similarity_matrix = cosine_similarity(customer_product_matrix_scaled)
similarity_df = pd.DataFrame(similarity_matrix,
                             index=customer_product_matrix.index,
                             columns=customer_product_matrix.index)

# Find top 3 similar customers for each target customer
lookalike_map = {}
for cust_id in target_customers:
    similar_customers = similarity_df.loc[cust_id].drop(cust_id).nlargest(3)
    lookalike_map[cust_id] = list(zip(similar_customers.index, similar_customers.values))

# Convert to DataFrame
lookalike_df = pd.DataFrame([(k, v[0], v[1]) for k, vals in lookalike_map.items() for v in vals],
                            columns=["CustomerID", "LookalikeID", "Score"])

# Save to CSV
lookalike_df.to_csv("Lookalike2.csv", index=False)
print("Lookalike.csv has been created successfully!")

# Visualize the data using a graph
G = nx.Graph()
for cust_id, lookalikes in lookalike_map.items():
    for lookalike_id, score in lookalikes:
        G.add_edge(cust_id, lookalike_id, weight=score)

plt.figure(figsize=(10, 6))
pos = nx.spring_layout(G, seed=42)
nx.draw(G, pos, with_labels=True, node_size=3000, node_color='lightblue', edge_color='gray')
labels = nx.get_edge_attributes(G, 'weight')
nx.draw_networkx_edge_labels(G, pos, edge_labels=labels)
plt.title("Customer Lookalike Network")
plt.show()
```

```
Lookalike.csv has been created successfully!
```

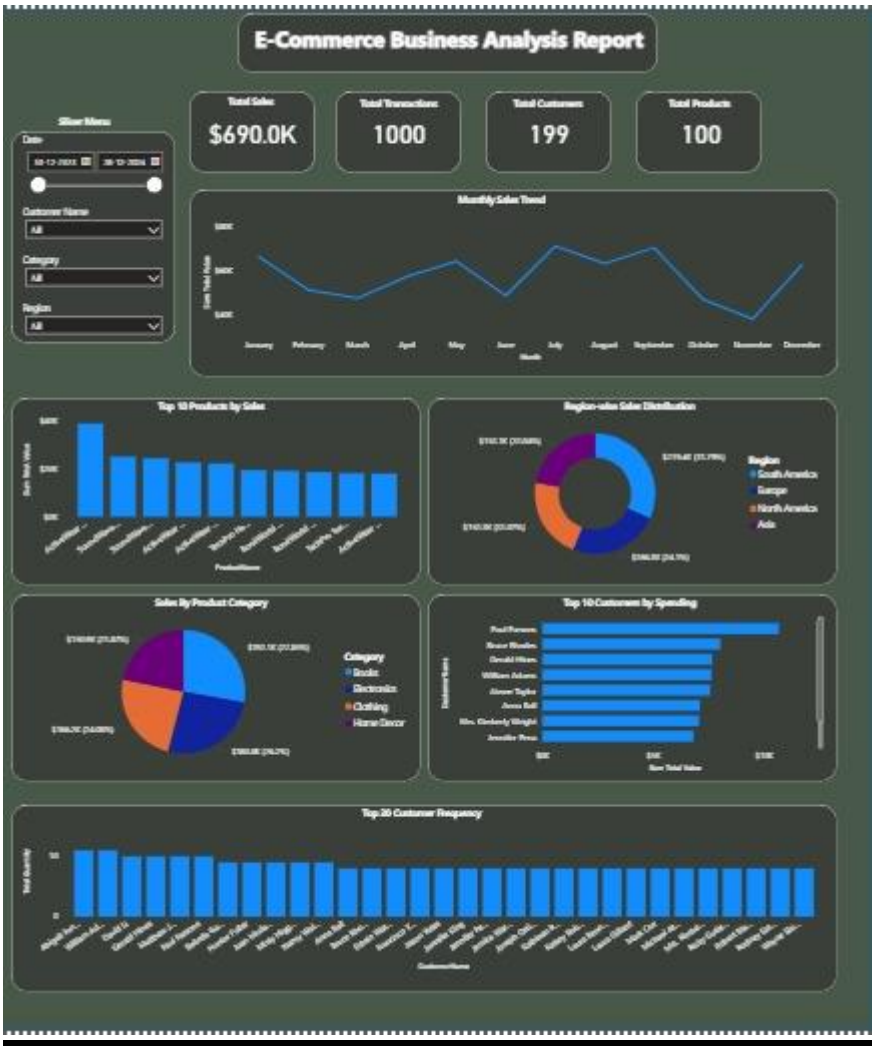
To identify customers with similar product preferences by building a lookalike network graph using cosine similarity. This model helps businesses make personalized recommendations and group similar customers for targeted campaigns.

1. Cluster Formation:

- Customers are naturally grouped into tight clusters, showing they purchased similar sets of products.
- These clusters can represent lookalike segments or communities.

2. Highly Connected Customers:
   - Central nodes with many edges (connections) are highly similar to many others.
   - These can be treated as representative customers for that segment.
3. Sparse Connections:
   - Some customers have fewer connections, indicating unique or less common preferences.

- **Microsoft Power BI Dashboard**:-



◈ **Top Summary Metrics (Top Row)**

- Total Sales: $690.0K
- Total Transactions: 1000
- Total Customers: 199
- Total Products: 100

➞ This gives a quick snapshot of the business performance.

🔷 Filters (Left Sidebar)

- Date Range
- Customer Name
- Age
- Category
- Region

➞ These slicers allow interactive filtering to drill down into specific subsets of the data.

📈 Monthly Sales Trend (Top Middle Graph)

- Line chart showing sales variation month-by-month.
- Observations:
- Sales were relatively low in the early months.
- Peak around July and September.
- Drop in November, then recovery in December.

🛒 Top 10 Products by Sales (Bar Chart - Bottom Left)

- Lists the highest-selling products.
- This helps identify which products contribute most to revenue.

🌍 Region-wise Sales Distribution (Doughnut Chart - Top Right)

- Shows the share of total sales by region:
- North America has the highest.
- Followed by Europe, South America, Asia, and Africa.

📦 Sales by Product Category (Pie Chart - Bottom Left)

- Product categories:
- Electronics (largest share)
- Followed by Books, Clothing, and Home Decor
- Matches the earlier bar chart EDA you shared.

👤 Top 10 Customers by Spending (Middle Right)

- Horizontal bar chart showing the highest-value customers.
- Useful for loyalty programs or targeted marketing.

🔁 Top 10 Customers by Frequency (Bottom Right)

- Shows customers who purchased most often.
- Some customers may buy frequently but spend less per transaction.

📊 Usefulness of the Dashboard

- For Marketing: Identify best customers, top products, and underperforming regions.
- For Inventory: Track product performance and seasonal trends.
- For Sales Strategy: Focus efforts on high-demand months and customer segments.

10. Five Key Business Insights per Notebook:-

After reviewing each dataset and corresponding notebook, the following insights were derived:

1. High Revenue Driven by a Few Key Products:

📊 **Based on: "Top 20 Products by Total Sales Value"**

- A small set of products (e.g., Product IDs like P029, P079, P028) account for a large share of total revenue.
- These products are not necessarily the most frequently sold but are likely higher in price or offer higher profit margins.
- Actionable Insight:
- Focus marketing and promotional efforts around these top-performing SKUs.
- Consider launching premium versions or accessories of these products.
- Secure inventory in advance to avoid stockouts.

2. Top Customers Drive Disproportionate Revenue:

📊 **Based on: "Top 20 Customers by Transaction Value" and "Top 20 by Transaction Count"**

- A small percentage of customers (e.g., C0141, C0056, C0086) are responsible for a significant portion of revenue.
- Some customers are repeat purchasers with lower order values (high frequency), while others purchase infrequently but with higher ticket sizes.
- Actionable Insight:
  - Segment and personalize retention strategies:
  - Loyal high spenders: Provide exclusive deals, loyalty points.
  - Frequent low spenders: Offer bundle discounts or upselling.

3. Customer Segmentation Reveals Four Distinct Buying Behaviors

📊 **Based on: Customer Segmentation Clustering (TotalSpent vs Frequency)**

- Customers fall into 4 clusters:
- Cluster 0: High spend, high frequency (most valuable)
- Cluster 1: Moderate spenders
- Cluster 2: Low spend, but relatively frequent
- Cluster 3: Low spend, low frequency

- Actionable Insight:
- Target Cluster 0 with loyalty programs and early access to sales.
- Re-engage Cluster 3 using remarketing emails or special limited-time offers.
- Use different messaging per segment in ads or email marketing.

4. Sales Performance Varies Significantly by Region and Category:

📊 Based on: Region-wise Sale of Each Category

- Books perform best across all regions, especially in South America and Europe.
- Clothing leads in Asia, while Home Decor performs moderately well across all.
- Actionable Insight:
  - Prioritize Books and Electronics inventory in Europe and South America.
  - Create region-specific campaigns to promote best-performing categories.
  - Consider localizing promotions based on cultural or regional preferences.

5. Customer Activity Peaks Between 3 PM to 6 PM:

📊 Based on: "Customer Activity by Time of the Day"

- The highest number of unique customer interactions occur between 3 PM and 6 PM, followed by 3 AM to 6 AM and 6 PM to 9 PM.
- Activity drops significantly between 12 AM and 3 AM.
- Actionable Insight:
- Schedule marketing emails, push notifications, and product launches during peak hours.
- Optimize server load and ad bidding strategies during high-activity time slots.
- Consider running flash sales or live events in the most active windows.

11. **RESULTS & EVALUATION**

This project successfully delivered a comprehensive analysis of customer behavior, product performance, and regional sales trends using real-world e-commerce data. Through the integration of Exploratory Data Analysis (EDA), Unsupervised Learning (Clustering), and Recommendation Systems, we were able to extract actionable insights to support data-driven business decisions.

Key deliverables include:

- Region-wise sales performance across product categories.
- Customer segmentation based on spending and frequency behavior.
- Identification of top-selling and top-revenue-generating products.
- Network-based Lookalike Model for customer recommendations.
- Visual dashboards for business intelligence (Power BI or matplotlib-based).

11.1 Evaluation of Methods and Results:

COMPONENT	EVALUATION
Data Preprocessing & Feature Engineering	Successfully handled missing values, merged datasets, engineered features like TotalValue, Frequency, and Time Slot, ensuring clean and analysis-ready data.
EDA and Visualization	Charts clearly revealed category-wise, time-wise, and region-wise performance. Helped in deriving 5+ strategic business insights.
Clustering for Customer Segmentation	The KMeans model revealed 4 distinct customer segments based on purchase behavior, allowing targeted marketing strategies.
Lookalike Recommendation Network	Implemented using cosine similarity and network visualization. Identified customers with similar product preferences for cross-selling.
Sales Analysis	Identified top products and customers by value and volume. Also uncovered regional trends in consumer demand.
Total Efficiency	Python (Pandas, Seaborn, Matplotlib, NetworkX) was effectively used for analysis and modeling. Power BI can be used for live dashboarding.

12. **CONCLUSION**

The goal of this project was to analyze customer transaction data to uncover meaningful patterns and support strategic business decisions through customer segmentation, sales performance insights, and data visualization. Using a combination of Exploratory Data Analysis (EDA), unsupervised machine learning (clustering), and network-based recommendation modeling, the project provided a comprehensive view of customer behavior and product performance.

1. **Customer Behavior Varies Significantly:**

• Through KMeans clustering, we identified four distinct customer groups based on total spend and purchase frequency.

• High-value customers were distinguishable from low-engagement users, which allows businesses to prioritize retention efforts, personalized offers, and targeted engagement.

2. **Product-Level Analysis Revealed Revenue Concentration:**

- A handful of products contribute disproportionately to total revenue, emphasizing the importance of inventory control, pricing strategy, and marketing focus on top-performing SKUs.

- Additionally, product sales quantity did not always correlate with revenue, highlighting the impact of unit pricing.

3. **Regional and Category Trends Uncovered:**

• Different regions favored different product categories. For example, Books performed best in Europe and South America, while Clothing dominated sales in Asia.

• These insights enable tailored regional campaigns and localized inventory management.

4. **Customer Engagement Peaks at Specific Times:**

• Analysis of customer activity by time slots showed clear behavioral patterns, with the most engagement occurring between 3 PM – 6 PM.

• This supports optimized marketing scheduling, such as when to send promotional emails or run flash sales.

5. **Lookalike Modeling Supports Cross-Selling:**

• A customer-product similarity network was built using cosine similarity, enabling identification of similar customers.

• This method supports recommendation systems, helping suggest relevant products to customers who exhibit similar purchasing patterns.

13. **FUTURE SCOPE**

1**. Integration of Time-Series Forecasting**

- Objective: Predict future sales trends for products and regions.
- Tools: ARIMA, Prophet, LSTM models.
- Impact: Supports inventory planning, promotional scheduling, and demand forecasting.

2. **Customer Lifetime Value (CLV) Prediction**

- Objective: Estimate the total expected revenue from a customer over their lifetime.
- Tools: RFM (Recency, Frequency, Monetary) analysis, survival analysis, or regression models.
- Impact: Helps businesses focus on high-value customers and plan long-term engagement strategies.

3. **Real-Time Personalization & Recommendation Systems**

- Objective: Deliver product suggestions to users dynamically based on real-time interactions and lookalike modeling.
- Tools: Collaborative filtering, content-based filtering, deep learning-based recommender systems.
- Impact: Enhances user experience, improves conversion rate, and boosts average order value.

4. **Churn Prediction Modeling**

- Objective: Identify customers likely to stop purchasing and implement retention strategies.
- Tools: Classification models like Logistic Regression, Random Forest, XGBoost.
- Impact: Reduces customer attrition and improves lifetime customer engagement.

5**. A/B Testing for Marketing Strategy**

- Objective: Experiment with different marketing tactics (email campaigns, product bundles, discounts).
- Tools: Statistical hypothesis testing.
- Impact: Data-backed decisions on campaign effectiveness and customer response optimization.

6**. Incorporating External Data Sources**

- Objective: Enrich analysis with external factors like seasonality, economic trends, or competitor pricing.
- Tools: APIs, web scraping, market data platforms.
- Impact: Improves the accuracy of forecasts and broadens context for strategic decisions.

14. **LIMITATIONS**

1. Lack of Real-Time Data

• The analysis is based on historical, static datasets and does not account for real-time customer behavior or evolving trends.

• This limits the ability to track live performance, identify emerging product interests, or respond to customer actions dynamically.

🛠 **Mitigation:** Future integration of real-time data pipelines can improve responsiveness and automation.

2. Missing Demographic & Behavioral Attributes

• The dataset does not include customer demographics (e.g., age, gender, location) or behavioral data (e.g., clickstream, session time).

• As a result, segmentation is limited to transactional features like purchase frequency and amount, missing deeper personalization opportunities.

🛠 **Mitigation:** Enrich datasets with CRM, website analytics, or external demographic data.

3. Product Names vs. Product IDs

• In some analyses (e.g., top-selling products), product IDs were used instead of descriptive names.

• This reduces interpretability for stakeholders who want clear insights into specific product lines or categories.

🛠 **Mitigation:** Merge product metadata (name, brand, price) for clearer visualizations and storytelling.

4. No Profit Margin or Cost Consideration

• Sales analysis is based on revenue (TotalValue), without incorporating cost of goods sold (COGS) or profit margins.

• This may lead to overvaluing high-revenue but low-margin products or customers.

🛠 **Mitigation:** Include cost and margin data to shift analysis from sales-driven to profit-driven.

15. **BIBLIOGRAPHY**

1) Hughes, A. M. (1994). Strategic Database Marketing: The Masterplan for Starting and Managing a Profitable, Customer-Focused Database Marketing Program. McGraw-Hill.
2) Jolliffe, I. T. (2002). Principal Component Analysis (2nd ed.). Springer.
3) van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. Journal of Machine Learning Research, 9(Nov), 2579–2605.
4) MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability (Vol. 1, pp. 281–297). University of California Press.
5) Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD '96) (pp. 226–231).
6) Comaniciu, D., & Meer, P. (2002). Mean shift: A robust approach toward feature space analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(5), 603–619.
7) Zhang, T., Ramakrishnan, R., & Livny, M. (1996). BIRCH: An efficient data clustering method for very large databases. In Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data (pp. 103–114).
8) McLachlan, G., & Peel, D. (2000). Finite Mixture Models. John Wiley & Sons.
9) Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53–65.
10) Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-1(2), 224–227.
11) Calinski, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. Communications in Statistics, 3(1), 1–27.
12) Wedel, M., & Kamakura, W. A. (2000). Market Segmentation: Conceptual and Methodological Foundations (2nd ed.). Springer.