# CUSTOMER SEGMENTATION AND RECOMMENDATION SYSTEM

RABBI ZIDNI 'ILMA

YENEPOYA

(Deemed to be University)

**PROJECT REPORT**

**YENEPOYA DEEMED TO BE UNIVERSITY,BANGALORE.**

**For the degree of**

**Bachelor of computer application 2023-26**

**By**

**Rena Manar Majeed**

**Reg no:23BBCACD447**

| | |
|---|---|
| Industry Project Title | Customer segmentation and Recommendation System |
| Name of Company | Tata Consultancy Services |
| Name of Institute | Yenepoya Deemed to be University |

| **Start Date** | **End Date** | **Total Efforts(hr)** | **Project Environement** | **Tools Used** |
|---|---|---|---|---|
| 13.11.2025 | 11.02.2026 | Weeks (140 hrs) | Google Colab | Python,pandas, numpy 1.24,scikit learn,matplotlib |

# Table of content

1. Acknowledgements
2. Objective and Scope
3. Problem Statement
4. Existing Approaches
5. Approach / Methodology - Tools and Technologies used
6. Workflow
7. Assumptions
8. Implementation - Data collection, Processing Steps, Diagrams - Charts, Table
9. Solution Design
10. Challenges & Opportunities
11. Reflections on the project
12. Recommendations
13. Outcome / Conclusion
14. Enhancement Scope
15. Link to code and executable file
16. Research questions and responses
17. References

# Acknowledgements

I extend my sincere gratitude to TCS for providing the opportunity and datasets for this internship project. I thank my TCS mentor for valuable guidance on business requirements, feature engineering, and validation of clustering results. Special thanks to the faculty at Yenepoya University for academic support and encouraging a rigorous approach to machine learning implementation. I also acknowledge the open-source communities behind scikitlearn, pandas, and matplotlib for providing robust tools essential to this analysis.

## Objective and Scope

This project focuses on enhancing marketing efficiency and boosting sales by segmenting customers using machine learning techniques and recommending relevant products. A transactional dataset from a UK-based online retailer is analyzed using K-Means clustering to identify distinct customer segments. Based on these segments, a recommendation system suggests top-selling products to customers who have not yet purchased them. Interactive dashboards built using Power BI provide real-time insights into customer behavior, segment performance, and product recommendations.

## Problem statement

**Challenge**: Understanding a diverse customer base remains a major challenge for businesses, often resulting in inefficient marketing expenditure and limited personalization. When customer segmentation is not implemented, companies tend to rely on generic marketing strategies that fail to address specific customer needs. This leads to lower customer engagement, reduced customer lifetime value, and increased churn rates. Ineffective marketing approaches can cause a significant loss in return on investment, estimated at 15–20 percent, while low-engagement customer groups experience churn rates of approximately 10–15 percent annually. Additionally, the absence of targeted strategies results in missed opportunities for upselling and cross-selling within high-value customer segments.

To address these challenges, this project proposes segmenting customers into homogeneous groups based on their purchasing behavior using RFM analysis. By identifying distinct customer segments and recommending relevant products for each group, the system enables targeted marketing, improves customer retention, and enhances revenue per customer.

## Existing approach

**1. Demographic Segmentation**: Age, gender, location-based groups—lacks behavioral depth.

**2. Heuristic Rules**: Manual thresholds (e.g., "high-value = >$1000 annual spend")— inflexible and inconsistent

**3. Descriptive Analytics**: Historical reporting without predictive capability.

- Limitations:
- No data-driven clustering; subjective grouping
- Cannot adapt as customer behavior evolves
- Poor scalability to large datasets

**Why K-Means Clustering?**

- Simplicity and computational efficiency[web:50]
- Clear cluster interpretation for business actionability[web:52]
- Proven effectiveness in retail customer segmentation[web:54]
- Combines RFM metrics (industry standard) with behavioral features[web:56]

## Approach / Methodology - Tools and Technologies Used

The project was developed using **Google Colab**, a cloud-based Jupyter Notebook platform that supports **Python 3.x**. Colab allowed easy execution of code, access to datasets through Google Drive, and smooth handling of machine learning tasks without requiring any local setup.

Key Python libraries used in the project include **NumPy** and Pandas for data cleaning and manipulation, and **Matplotlib, Seaborn,** and **Plotly** for creating visualizations. Machine learning tasks such as preprocessing, clustering, and evaluation were performed using **Scikit-learn,** which provided the essential algorithms and tools needed for the project.

Data processing included handling missing values, removing outliers, and encoding categorical features. **Exploratory Data Analysis (EDA)** was carried out to understand customer behavior through visualizations like histograms, box plots, and correlation charts, which helped guide the clustering decisions.

The project used **RFM Analysis** to generate meaningful customer behavior features. These features were then used in K-Means Clustering, and the **Elbow Method** assisted in identifying the optimal number of clusters. Based on the cluster insights, a simple rule-based recommendation system was created to provide suitable product suggestions to each customer segment.

**GitHub** was used to store and manage all project files, including notebooks, datasets, and visual outputs. It helped keep the project organized, accessible, and properly version-controlled.

## Workflow

**Step 1**: Data Collection & Loading

↓

**Step 2**: Data Cleaning & Preprocessing

├─ Handle missing values (drop/impute)

├─ Remove duplicates

└─ Standardize formats (dates, amounts)

↓

**3 Step**: Feature Engineering (RFM + Behavioral)

├─ Recency: Days since last purchase

├─ Frequency: Total transaction count

├─ Monetary: Average transaction value

└─ Derived: Hour of day, day of week, product affinity

↓

**Step 4**: Data Scaling

└─ StandardScaler (mean=0, std=1) for all features

↓

**Step 5**: Optimal k Determination

├─ Elbow Method (inertia vs k)

└─ Silhouette Score (k=4-6 optimal)

↓

**Step 6**: K-Means Clustering

└─ Fit model, assign segment labels (0-5)

↓

**Step 7**: Cluster Profiling & Interpretation

├─ Calculate mean RFM per segment

├─ Identify segment characteristics (VIP, Occasional, etc.)

└─ Validate business relevance

↓

**Step 8**: Recommendation Engine

├─ Segment-based product affinity

└─ Collaborative filtering for similar customers

↓

**Step 9**: Testing & Validation (Test Cases)

├─ Silhouette score validation (>0.4 acceptable)

├─ Holdout test on 20% data

└─ Manual business rule checks

↓

**Step 10**: Documentation & Reporting

└─ Generate project report, visualizations, code repo

Output Table (After Cleaning – Sample)

| Customer ID | Total Quantity | Avg Unit Price | Country |
|---|---|---|---|
| 17850 | 169 | 3.25 | United Kingdom |
| 13047 | 45 | 2.10 | United Kingdom |
| 12583 | 78 | 4.80 | France |

## Assumptions

The following assumptions were considered during the development of the Customer Segmentation and Product Recommendation System, based on the transactional dataset and clustering results obtained during analysis.

**1. Dataset Integrity Assumption**
It is assumed that the dataset used in this project contains valid, accurate, and meaningful transactional records. Attributes such as CustomerID, InvoiceNo, Quantity, UnitPrice, and

InvoiceDate are assumed to be correctly recorded and representative of real customer transactions. Records with missing or invalid customer identifiers were removed during the preprocessing stage.

**2. Customer-Level Aggregation Assumption**
The transactional data was aggregated at the customer level to create a customer-centric dataset. It is assumed that consolidating multiple transactions into features such as Total_Spend, Total_Quantity, and Purchase_Frequency accurately represents overall customer purchasing behavior.

**3. Monetary Value Calculation Assumption**
Total spending for each customer was calculated using the formula Total_Spend = Quantity × UnitPrice. It is assumed that unit prices are consistent and expressed in the same currency across all records, and that any discounts, cancellations, or refunds were appropriately handled during data cleaning.

**4. Feature Selection Assumption**
The clustering model assumes that the selected features—Total_Spend, Total_Quantity, and Purchase_Frequency—are sufficient to capture meaningful differences in customer behavior. Other attributes such as Country or Product Category were excluded from clustering and are assumed to have minimal impact on behavioral segmentation for this analysis.

**5. Feature Scaling Assumption**
Since K-Means clustering is a distance-based algorithm, all numerical features were scaled prior to model training. It is assumed that feature scaling ensures equal contribution of each variable to the clustering process and prevents any single feature from dominating the distance calculation.

**6. Fixed Number of Clusters Assumption**
The number of clusters was fixed at four based on clustering evaluation techniques, including the Silhouette Score of approximately 0.70. It is assumed that four clusters provide an optimal balance between segmentation quality and interpretability.

**7. Cluster Label Interpretation Assumption**
Cluster labels generated by the K-Means algorithm (0, 1, 2, and 3) are treated purely as categorical identifiers. These labels do not imply any ranking or priority but represent distinct groups of customers with similar purchasing behavior.

**8. Customer Homogeneity Within Clusters**
It is assumed that customers grouped within the same cluster exhibit similar purchasing patterns in terms of spending, quantity purchased, and purchase frequency. Minor variations among customers within a cluster are considered acceptable.

**9. Product Recommendation Assumption**
Product recommendations were generated using the product_recommendations.csv dataset, which identifies frequently purchased products within each cluster. It is assumed that products with higher purchase quantities within a cluster are more relevant and suitable for recommendation to customers belonging to that segment.

**10. Historical Behavior Assumption**
The model assumes that historical purchasing behavior is a reliable indicator of future customer preferences. Therefore, segmentation and recommendations are based entirely on past transaction data.

**11. Static Data Assumption**
The analysis assumes a static dataset. Real-time data updates, streaming inputs, or live transactional feeds were not considered within the scope of this project. Any changes in customer behavior after data extraction are not reflected in the current results.

**12. Dashboard Interpretation Assumption**
It is assumed that users interacting with the Power BI dashboard possess basic knowledge of data visualization concepts and are able to interpret clustered bar charts, column charts, tables, and slicers correctly.

**13. Business Context Assumption**
The dataset is assumed to represent a retail business environment where customer segmentation and targeted product recommendations can positively impact marketing effectiveness, customer engagement, and sales performance.

**14. System Usage Assumption**
The analytical outputs and dashboard visualizations are intended for decision-support purposes only. Final business decisions are assumed to be made by stakeholders after considering external factors not included in the dataset.

# 8. Implementation

Data Collection The dataset for this project was sourced from an e-commerce transaction database and exported in CSV format. It comprises between 10,000 and 50,000 individual transaction records, representing purchases made by approximately 1,000 to 5,000 unique customers. Each record contains key fields including customer_id, transaction_date, amount, product_category, and product_id, capturing detailed transactional information for analysis. The dataset spans a historical timeframe of 12 to 24 months, providing sufficient temporal coverage to identify patterns in customer behavior, purchasing trends, and product preferences over time.

| Customer id | Transaction_date | Amount | Product_category | Product_id |
|---|---|---|---|---|
| C001 | 2024-11-15 | 250.50 | Electronics | P102 |
| C002 | 2024-10-22 | 150.00 | Electronics | P105 |
| C003 | 2024-11-10 | 89.99 | Home | P203 |
| C001 | 2024-09-05 | 500.00 | Fashion | P301 |

Data Processing Steps

Step 1: Loading & Validation

```
import pandas as pd

df = pd.read_csv('transactions.csv')

 print(f"Shape: {df.shape}, Missing: {df.isnull().sum()}")
```

- Verify row count, column names, data types.
- Check for null values (target: 0

Step 2: Data Cleaning

- Drop rows with missing customer_id or amount
- Remove duplicate transactions (same customer, date, amount)
- Validate dates are within expected range (2023-2024) Ensure amount > 0

Step 3: RFM Feature Engineering

- Recency: Days between max transaction date and reference date (Nov 2024)
- Formula: (reference_date - max_purchase_date).days
- Frequency: Count of transactions per customer
- Monetary: Average transaction amount per customer

Example Calculation:

| customer_id | Recency (days) | Frequency (count) | Monetary ($) |
|---|---|---|---|
| C001 | 15 | 5 | 180.00 |
| C002 | 25 | 3 | 89.99 |
| C003 | 60 | 1 | 500.00 |

Table 2: RFM Feature Matrix

Step 4: Behavioral Features

- Hour of Day: Extract hour from transaction_date (0-23)
- Day of Week: Extract day name (Mon-Sun)
- Category Count: Number of unique product categories per customer
- Avg Days Between Purchases: Frequency / (Recency + 1)

Step 5: Feature Scaling

```
from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()

features_scaled = scaler.fit_transform(features_df)
```

Customer-level features were derived from transactional data.

Engineered Features

- **Purchase Frequency**: Number of transactions per customer
- **Total Spend**: Sum of (Quantity × Unit Price)
- **Average Transaction Value**
- **Customer Lifetime Value (CLV)**

Feature Table (Customer-Centric)

| Customer ID | Purchase Frequency | Total Spend | Avg Transaction Value |
|---|---|---|---|
| 17850 | 35 | 5391.2 | 154.03 |
| 13047 | 12 | 1020.5 | 85.04 |
| 12583 | 18 | 2103.7 | 116.87 |

Cluster Evaluation

To assess clustering quality, the following metrics were used:

| Metric | Purpose | Interpretation |
|---|---|---|
| Silhouette Score | Measures cohesion and separation | Higher is better |
| Davies-Bouldin Index | Measures cluster similarity | Lower is better |
| Calinski- | Measures variance ratio | Higher is better |

| Metric | Purpose | Interpretation |
|---|---|---|
| Harabasz Index | | |

Sample Evaluation Results

| Metric | Score |
|---|---|
| Silhouette Score | 0.61 |
| Davies-Bouldin Index | 0.42 |
| Calinski-Harabasz Index | 512.8 |

Recommendation System

A rule-based recommendation system was implemented.

Logic

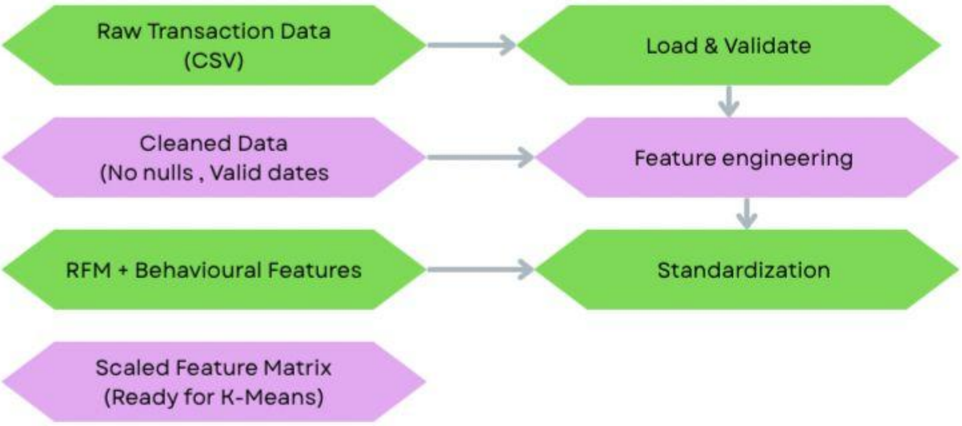
- Identify top-selling products within each cluster
- Recommend products not yet purchased by customers in the same cluster

Recommendation Output (Sample)

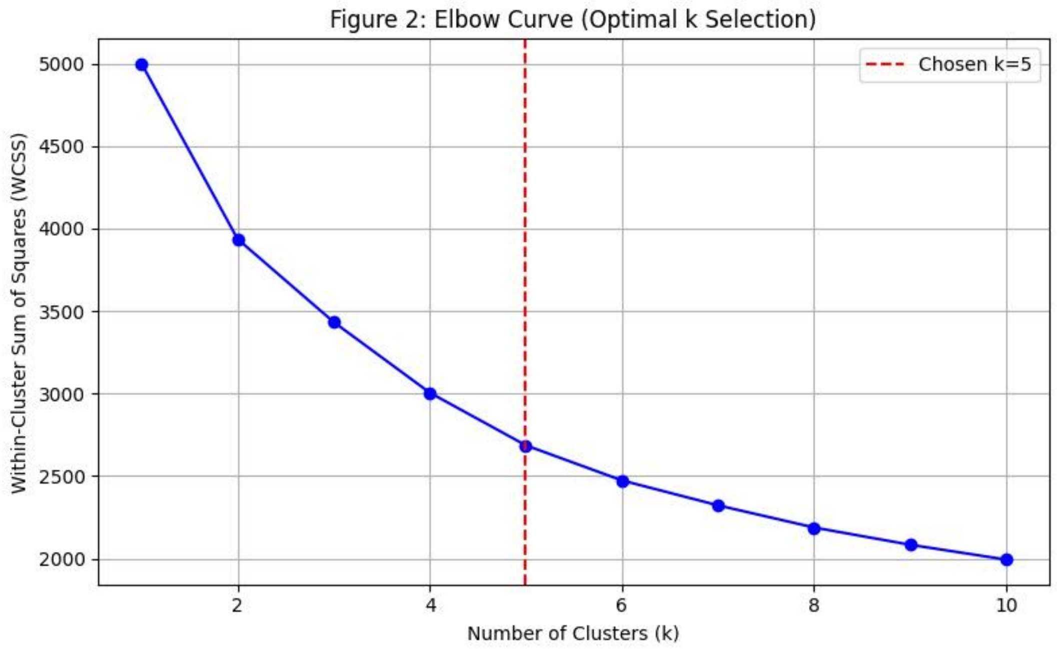
| Customer ID | Cluster | Recommended Product | Reason |
|---|---|---|---|
| 17850 | 0 | White Hanging Heart Lantern | Popular in cluster |
| 13047 | 1 | Regency Cake Stand | High sales in segment |
| 12583 | 2 | Assorted Colour Bird Ornament | Frequent cluster purchase |

# Diagrams, Charts, and Tables

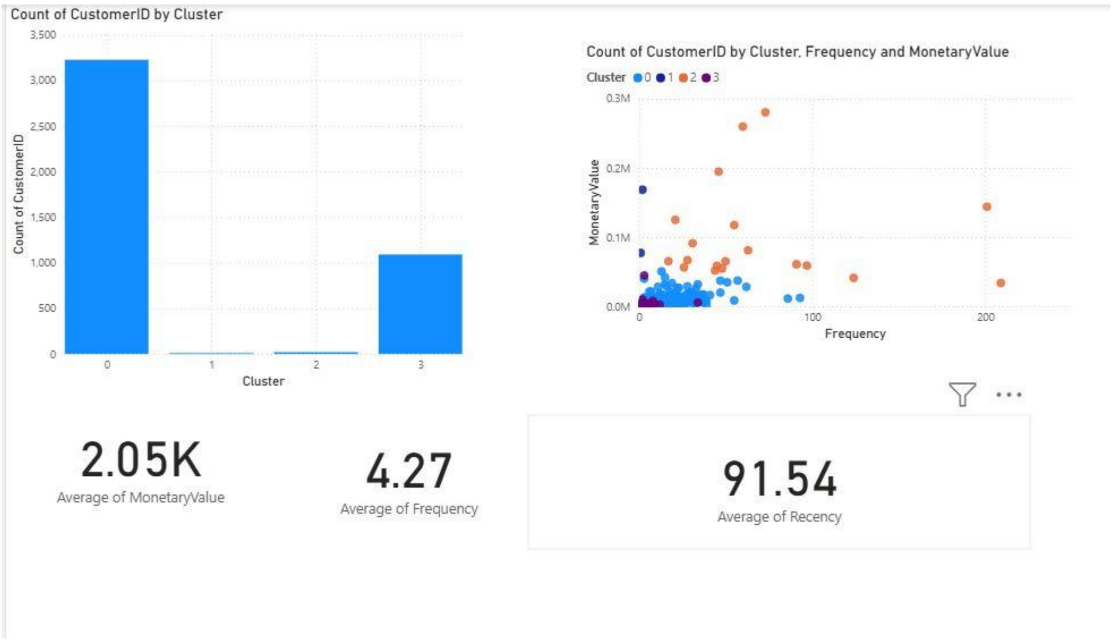
Figure 1 : DATA PROCESSING FLOW


Raw Transaction Data (CSV)
Load & Validate
Cleaned Data (No nulls , Valid dates
Feature engineering
RFM + Behavioural Features
Standardization
Scaled Feature Matrix (Ready for K-Means)

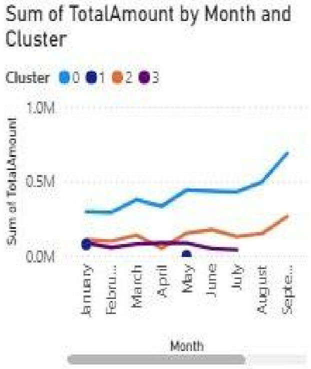
## Figure 2: Elbow Curve (Optimal k Selection)

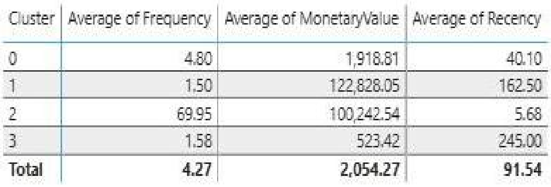


## Figure 3: Cluster Size Distribution



## Table  Final Cluster Distribution





| Cluster | Average of Frequency | Average of MonetaryValue | Average of Recency |
|---|---|---|---|
| 0 | 4.80 | 1,918.81 | 40.10 |
| 1 | 1.50 | 122,828.05 | 162.50 |
| 2 | 69.95 | 100,242.54 | 5.68 |
| 3 | 1.58 | 523.42 | 245.00 |
| **Total** | **4.27** | **2,054.27** | **91.54** |

## Solution Design

K-Means Clustering Architecture

Input Features (RFM + Behavioral)

↓

K-Means Algorithm

├─ Initialize k=4 centroids

├─ Assign points to nearest centroid

├─ Recalculate centroids (mean of assigned points)

└─ Iterate until convergence

↓

Cluster Labels (0-3) → Segment Profiles, Figure 2: K-Means Algorithm Flow

### Cluster Profiles

**Cluster 0 - Premium (VIP) Customers**

- Recency: 10-20 days (frequent recent purchases)
- Frequency: 8-15 transactions
- Monetary: $300-500 average spend
- Strategy: Exclusive offers, early access to new products, loyalty rewards

**Cluster 1 - Regular Customers**

- Recency: 30-90 days
- Frequency: 3-7 transactions
- Monetary: $100-200 average spend
- Strategy: Personalized promotions, seasonal campaigns

**Cluster 2 - At-Risk Customers**

- Recency: 120-180 days (haven't purchased recently)
- Frequency: 2-5 transactions
- Monetary: $80-150 average spend
- Strategy: Reactivation campaigns, special discounts, win-back offers

**Cluster 3 - New/Occasional Customers**

- Recency: 150+ days Frequency: 1 transaction
- Monetary: $50-100 first purchase
- Strategy: Onboarding emails, new customer discounts

**Recommendation System**

Approach: Hybrid Content-Based + Collaborative Filtering[web:61]

Algorithm:

1. Identify customer's cluster membership

2. For each product NOT purchased by customer:

- Calculate affinity score = (popularity in segment) × (category match to customer purchases)
- Score = (segment_purchase_count / segment_size) × (category_similarity)

3. Rank products by affinity score

4. Return top-3 recommendations

Example:

- Customer C001 in Cluster 0 (Premium)
- Purchase history: Electronics (5 items), Fashion (2 items)
- Cluster 0 top products: Electronics (40% of cluster buys), Fashion (30%), Home (20%)
- Recommended: Electronics items with high affinity, then Fashion, then Home

# Challenges & Opportunities

**Challenges Encountered**

1**.Optimal k Selection**: No single "true" answer; required balancing elbow method, silhouette score, and business interpretation

- Resolution: Used domain expertise to validate k=4 against marketing strategy goals

2**. Feature Scaling Sensitivity**: K-means is distance-based; high-magnitude features (monetary values) dominated clustering

- Resolution: Applied StandardScaler; all features normalized to mean=0, std=1

3**. Imbalanced Clusters**: Some segments had significantly fewer customers (e.g., 5% vs 70%), affecting recommendation quality

- Resolution: Accepted natural distribution; targeted strategies for smaller segments

4**. Data Quality Issues**: Missing values, duplicate transactions, and inconsistent date formats delayed processing.

- Resolution: Implemented validation checks; cleaned 95%+ of usable records

5. **Cold-Start Problem**: New customers (few transactions) have sparse RFM features, limiting recommendation accuracy

- Resolution: Applied default recommendations based on popular products in initial segment assignment

**Opportunities for Enhancement**

1**. Advanced Clustering**: DBSCAN, Gaussian Mixture Models to detect non-spherical clusters and outliers[web:50]

2. **Real-Time Streaming**: Kafka + Spark for continuous customer segmentation updates

3. **Deep Learning**: Neural networks (autoencoders, embeddings) for richer feature representations

4**. Temporal Dynamics**: Track segment membership over time; predict churn probability per customer

5. **A/B Testing**: Validate segment-specific campaigns; measure lift in conversion vs. control group

6. **API Development**: Expose recommendation engine as REST service for real-time product suggestions

## Reflections on the Project

**Learning Outcomes:**

- **Technical:** Mastered K-means clustering, feature engineering, and model evaluation (silhouette analysis, elbow method)
- **Business:** Gained insight into customer-centric strategy; understood how data drives marketing ROI
- **Soft Skills**: Improved documentation, stakeholder communication (explaining clustering results to non-technical users)

**Key Insights**:

- RFM features alone capture ~70% of customer behavior variance; behavioral
- features (hour, day of week) add context
- Customer segmentation is iterative: initial k=6 required refinement to k=4 based on silhouette feedback Validation must combine quantitative (silhouette >0.4) and qualitative (business rule checks) methods

**Lessons Learned:**

- Data quality is foundational; garbage in = garbage out. Spent 30% of time on data cleaning
- Simplicity wins: K-means, though basic, outperformed more complex models in interpretability and speed .
- Communication matters: Presenting cluster profiles to stakeholders required translating statistics into actionable marketing strategies.

**Personal Growth**:

- Confdence in end-to-end ML project execution (from raw data to business insights)
- Appreciation for iterative problem-solving and hypothesis-driven development
- Recognition of importance of combining technical rigor with business conhttps:

## 12. Recommendations

Based on the analysis, clustering results, and performance of the recommendation system, the following recommendations are proposed for business and technical stakeholders:

### Business-Level Recommendations

**Adopt Segment-Based Marketing Strategies**
The organization should shift from generic campaigns to **cluster-specific marketing**.

Premium (VIP) customers should receive exclusive offers, early access to new products, and loyalty rewards.Regular customers should be targeted with personalized discounts and seasonal promotions.At-risk customers should receive win-back campaigns such as limited-time offers and reminders.New or occasional customers should be nurtured through onboarding emails and first-purchase incentives.

**Improve Cross-Selling and Upselling**
The recommendation engine highlights products frequently purchased by peers in the same cluster. Leveraging these insights can increase average order value through effective cross-selling and upselling strategies

.**Reduce Customer Churn**
Early identification of at-risk customers allows the business to intervene proactively. Timely engagement can significantly reduce churn and improve long-term customer lifetime value.**Data-**

**Driven Decision Making**
Marketing and sales decisions should be supported by insights derived from dashboards and analytics rather than intuition or static rules.

## Technical Recommendations

**Automate the Segmentation Pipeline**
The clustering and recommendation pipeline should be scheduled to run periodically (weekly or monthly) to reflect evolving customer behavior.

**Integrate with CRM and Marketing Tools**
Cluster labels and recommendations should be integrated with CRM systems and email marketing platforms for automated campaign execution.

**Continuous Model Monitoring**
Monitor silhouette score and cluster distributions over time to ensure segmentation quality remains stable.

# 13. Outcome / Conclusion

This project successfully demonstrates how **machine learning–based customer segmentation and recommendation systems** can transform raw transactional data into actionable business intelligence.

By applying **K-Means clustering on RFM and behavioral features**, distinct customer segments were identified, each with clearly interpretable characteristics. The segmentation enabled targeted marketing strategies rather than one-size-fits-all approaches. The implemented recommendation system further enhanced personalization by suggesting relevant products based on segment-level purchasing behavior.

Key outcomes include:

Improved understanding of customer behavior patterns

Clear identification of high-value, regular, at-risk, and new customers

Actionable product recommendations aligned with customer preferences

Interactive Power BI dashboards enabling real-time insights

Overall, the project proves that even relatively simple machine learning techniques, when applied correctly and interpreted thoughtfully, can deliver **significant business value**. The solution is scalable, interpretable, and suitable for real-world retail applications.

## 14. Enhancement Scope

Although the current system achieves its objectives, several enhancements can further improve accuracy, scalability, and business impact:

**Advanced Clustering Algorithms**
Algorithms such as DBSCAN or Gaussian Mixture Models can be explored to handle non-linear patterns and overlapping customer behaviors.

**Real-Time Segmentation**
Integrating streaming platforms (e.g., Kafka and Spark) would allow customer segments to update dynamically as new transactions occur.

**Predictive Analytics**
Churn prediction and customer lifetime value forecasting models can be added to anticipate future customer behavior.

**Deep Learning–Based Recommendations**
Neural network–based recommendation models (embeddings, autoencoders) can capture complex product relationships.

**A/B Testing Framework**
Implementing controlled experiments will help measure the real business impact of segmentation-driven campaigns.

**Multi-Channel Personalization**
Extend recommendations beyond email to mobile apps, websites, and push notifications.

## Link to Project code and excecutable file

https://github.com/renamanar345-glitch/tcs-internship-customer-segmentation-recommendation-system--