Project Report: Customer Segmentation Using RFM & K-Means Clustering

---

A. Problem Statement

Businesses need to categorize customers based on their purchasing behavior to optimize marketing strategies, improve customer retention, and increase revenue. Traditional segmentation methods are not data-driven, leading to inefficiencies. The goal is to implement customer segmentation using RFM (Recency, Frequency, Monetary) analysis and K-Means clustering.

B. Project Objective

1. Identify distinct customer segments based on RFM attributes.
2. Apply K-Means clustering to group customers with similar behaviors.
3. Derive actionable insights to optimize marketing strategies.
4. Improve customer engagement and retention through data-driven decisions.

C. Data Description

- **Dataset:** The dataset consists of e-commerce transactional data.
- **Key Attributes:**
    - `InvoiceNo`: Transaction ID
    - `StockCode`: Product ID
    - `Description`: Product description
    - `Quantity`: Number of units purchased
    - `InvoiceDate`: Date of purchase
    - `UnitPrice`: Price per unit
    - `CustomerID`: Unique ID for customers
    - `Country`: Country of purchase
    - `Revenue`: Total transaction value (Quantity × UnitPrice)
    - `Month`: Month of transaction

D. Data Pre-processing Steps and Inspiration

1. **Handling Missing Values:** Removing rows with null `CustomerID`.
2. **Feature Engineering:** Creating the `Revenue` column.
3. **Date Formatting:** Converting `InvoiceDate` to datetime format.
4. **RFM Calculation:**
    - **Recency:** Days since last purchase.
    - **Frequency:** Number of transactions.
    - **Monetary:** Total revenue generated.

5. **Data Normalization:** Scaling RFM values for better clustering results.

E. Choosing the Algorithm for the Project

- **K-Means Clustering** is selected due to its efficiency in partitioning customers into distinct groups based on their purchasing patterns.
- **Alternative Models Considered:**
    - Hierarchical Clustering (Less scalable for large datasets)
    - DBSCAN (Sensitive to parameter tuning, not suitable for RFM data)

F. Motivation and Reasons for Choosing K-Means

- Handles large datasets efficiently.
- Provides clear and interpretable segmentation.
- Works well with numerical RFM data.
- Allows easy visualization and analysis of customer segments.

G. Assumptions

1. Customers with the same RFM values have similar purchasing behaviors.
2. The dataset represents the entire customer base.
3. The transactional data is accurate and up-to-date.

H. Model Evaluation and Techniques

1. **Elbow Method:** Determines optimal `k` for clustering.
2. **Silhouette Score:** Measures cluster quality.
3. **Cluster Visualization:** Using scatter plots to analyze segment distribution.

I. Inferences from the Same

- **Cluster 0 (Loyal Customers):** Frequent buyers, high monetary value.
- **Cluster 1 (Churned Customers):** Long inactive periods, low purchases.
- **Cluster 2 (High-Value Customers):** Very high spending, must be retained.
- **Cluster 3 (Low-Value Customers):** Rarely purchase, minimal engagement needed.

J. Future Possibilities of the Project

1. **Personalized Marketing Strategies:** Tailoring promotions for each cluster.
2. **Predictive Modeling:** Using machine learning to predict churn and loyalty.
3. **Integration with Business Intelligence (BI) Tools:** Automating insights.
4. **Real-time Segmentation:** Updating clusters dynamically as new data arrives.

Conclusion

This project successfully segments customers using K-Means clustering on RFM attributes, providing valuable insights for targeted marketing strategies. The findings help businesses enhance customer retention and maximize revenue growth.
