I. Problem Statement

The main goal of this project is to analyze the customer purchase data and categorize them into different segments based on their recency (how recent a customer made a purchase), frequency (how often a customer makes a purchase), and monetary value (how much money a customer spends). By analyzing these customer segments, we can identify the behavior of customers with similar patterns and target our marketing strategies accordingly.

II. Project Objective

The primary objective of this project is to identify customer segments based on their RFM scores. The project will involve pre-processing the data, training a machine learning model using the K-means algorithm, and evaluating the model's performance. The goal is to create a comprehensive customer segmentation report that can be used by businesses to make informed marketing decisions.

III. Data Description

The data used for this project consists of purchase information, including customer ID, invoice date, and total monetary value. The dataset is collected from a UK-based supermarket chain and covers a period of approximately 3 years.

IV. Data Pre-processing Steps and Inspiration

1. Remove duplicates: Duplicate rows in the dataset were removed to ensure consistency and accurate analysis.

2. Date parsing: The invoice date was parsed and transformed into a more readable format.

3. Recency calculation: For each customer, the most recent purchase date was calculated using the formula recency = current\_date - purchase\_date.

4. Frequency calculation: The number of purchases made by each customer was calculated and recorded.

5. Monetary value calculation: The total monetary value spent by each customer was calculated and recorded.

6. Feature scaling: To ensure consistent performance across all features, the RFM scores were scaled using the min-max scaler.

7. Missing value handling: No missing values were found in the dataset. However, if any were found, we would handle them by imputing the mean value for that feature.

V. Choosing the Algorithm for the Project

The K-means algorithm was chosen for this project because it is a popular and efficient clustering algorithm that can handle high-dimensional data. It works by partitioning the dataset into K clusters based on the mean distance between the data points within each cluster.

VI. Motivation and Reasons for Choosing the Algorithm

1. Handles high-dimensional data: K-means is well-suited for handling high-dimensional data, which is a common occurrence in customer segmentation problems.

2. Scalability: K-means is scalable and can handle large datasets.

3. Simplicity: The K-means algorithm is simple to understand and implement.

VII. Assumptions

1. Linearity: The K-means algorithm assumes that the relationship between the RFM scores and the cluster assignment is linear.

2. Homogeneity: Within each cluster, the customers have similar RFM scores and similar behavior.

VIII. Inferences from the Same

1. Cluster 0 ("0 Mean"): Customers in this cluster have relatively high recency, indicating that they haven't made a purchase recently. They have a moderate frequency of purchases and a moderate monetary value compared to other clusters.

2. Cluster 1 ("1 Mean"): Customers in this cluster have the lowest recency, suggesting that they are the most recent purchasers. They have a high frequency of purchases and the highest monetary value among all clusters, indicating high-value and frequent buyers.

3. Cluster 2 ("2 Mean"): Customers in this cluster have the lowest recency among all clusters, indicating that they are the most recent purchasers after Cluster 1. They have a moderate frequency of purchases and the highest monetary value among all clusters, indicating high-value customers.

4. Cluster 3 ("3 Mean"): These customers have a moderate recency, frequency, and monetary value compared to other clusters. They are neither the most recent purchasers nor the highest spenders.

5.Cluster 4 ("4 Mean"):

Customers in this cluster have the lowest recency and frequency of purchases.

However, they have a relatively high monetary value compared to other clusters, indicating that they make large purchases despite being less frequent.

Conclusion:

1.Successfully categorized customers into five segments based on recency, frequency, and monetary value using the K-means algorithm.

2.Insights gained from customer segmentation can inform targeted marketing strategies for businesses.

3.Choice of K-means algorithm justified by its scalability and simplicity.

4.Further research could enhance accuracy and reliability of predictions.

5.Overall, project offers valuable insights into customer behavior and opportunities for targeted marketing campaigns.