# Technical Design Document

## Customer Segmentation using RFM Analysis

## and K-Means Clustering

### 1. Introduction

The primary purpose of this project is to perform customer segmentation utilizing the RFM (Recency, Frequency, Monetary) approach. The project aims to analyze customer transaction data and categorize customers into distinct segments based on their purchase behaviors and interactions with the business.

**2. Toolset and Coding Language**

* *Python*: For data preprocessing, analysis, and segmentation using pandas and NumPy libraries.
* *Excel*: For storing cleaned, preprocessed, and segmented data.
* *RapidMiner*: For potential K-Means clustering and visualization.

**3. Data Models**

* *Original Dataset*: Contains customer transaction data with columns like Customer ID, Item Code, Invoice Number, Date of Purchase, Quantity, Price, etc. (Provided by Imarticus Institute & KPMG)
* *RFM Table\_Scaled*: Scaled values Recency, Frequency, and Monetary data per customer.
* *Segmented RFM Table*: RFM segments assigned to each customer.

**4. Data Volume**

* *Original Dataset*: 537979 Rows X 12 Columns
* *RFM Table\_scaled*: 395865 Rows X 04 Columns
* *Segmented RFM Table*: 395865 Rows X 07 Columns

**5. Technical Workflow**

* *Data Cleaning:* Using pandas, remove duplicates, handle missing values, and eliminate negative quantities.
* *RFM Analysis:* Calculate Recency, Frequency, and Monetary metrics per customer using pandas.
* *Data Preprocessing*: Log-transform and scale the RFM data using NumPy and sklearn.
* *RFM Segmentation*: Assign RFM segments and scores using pandas and NumPy based on quartiles.
* *Data Export*: Save the segmented data to Excel files.
* *K-Means Clustering in RapidMiner*:
* Import the cleaned and scaled data into RapidMiner.
* Utilize the K-Means operator for clustering based on RFM metrics.
* Configure input attributes, number of clusters, and distance measure.
* Analyze and visualize the clustering results within RapidMiner.
* *Interpretation and Labeling:* Analyze the clusters obtained from K-Means, interpret their characteristics, and assign meaningful labels to each segment.

**Contributed By (Group 03)**

|  |  |
| --- | --- |
| **Name** | **SID** |
| Ruchika Gupta | 200559617 |
| Bhavesh Natwarlal Waghela | 200532173 |
| Deema Sami Barham Al Dogom | 200548717 |
| Ashtha Ravikant Parmar | 200537002 |
| Navid Saleh Sadik | 200544271 |
| Olalekan Apantaku | 200560337 |