**Capstone Project Submission**

**Instructions:**

1. **Please fill in all the required information.**
2. **Avoid grammatical errors.**

|  |
| --- |
| **Team Member’s Name, Email and Contribution:** |
| **Name: Nitin B. Pawar**  **Email:** [**nitinpawar4315@gmail.com**](file:///C:\Users\Rakesh\Desktop\data%20analyst\credit%20card\nitinpawar4315@gmail.com)     * + Data Understanding   + Feature Analysis   + Feature Engineering   + Exploratory Data Analysis   + RFM model   + K-means   + Silhouette   + DBSCAN   + Elbow method   + Research Analytics |
| **Please paste the GitHub Repo link.** |
| **Github Link** :- <https://github.com/NITINPAWAR69/online_retail_costomer_segmentation> |
| **Please write a short summary of your Capstone project and its components.**  **Describe the problem statement, your approaches and your conclusions. (200-**  **400 words)** |
| Customer segmentation is the practice of dividing a company’s customers into groups that reflect similarity among customers in each group. The goal of segmenting customers is to decide how to relate to customers in each segment in order to maximize the value of each customer to the business. The contents of the dataset had features such as invoiceno., stockcode, description, quantity, unitprice, customerID, and country. The problem statement was to build an unsupervised machine learning algorithm to perform customer segmentation. We started with data wrangling in which we tried to handle null values, duplicates and performed feature modifications.  Next, we did some exploratory data analysis and tried to draw observations from the features we had in the dataset.  Next we perform a **RFM** model:  RFM stands for Recency, Frequency, and Monetary. RFM analysis is a commonly used technique to generate and assign a score to each customer based on how recent their last transaction was (Recency), how many transactions they have made in the last year (Frequency), and what the monetary value of their transaction was (Monetary).   * Recency = Latest Date - Last Invoice Data, * Frequency = count of invoice no. of transaction(s), * Monetary = Sum of Total   We implemented KMeans clustering algorithm on these features. We also performed silhouette and elbow method analysis to determine the optimal no. of clusters which was 2. We saw customers having high recency and low frequency and monetary values were part of one cluster and customers having low recency and high frequency, monetary values were part of another cluster. We saw higher values of frequency, monetary and low values of recency is deciding one class and low values of frequency, monetary and high values of recency is deciding other class. However, there can be more modifications on this analysis. One may choose to cluster into more numbers depending on company objectives and preferences.  **Conclusion:**   * K-Means with silhouette\_score of RFM Optimal\_Number\_of\_cluster are 2 * K-Means with Elbow methos of RFM Optimal\_Number\_of\_cluster are 2 * Although we didn't obtain two clearly separated clusters, we were able to build a model that can classify new customers into "low value" and "high value" groups. * Generally, if a customer only transacted with us a few times, they needed to be at least in the top 50th percentile in monetary spending to be considered a "high value customer". * The clusters assignments are muddled, which may be due to outliers that weren't removed. |