**Capstone Project Submission**

**Instructions:**

i) Please fill in all the required information.

ii) Avoid grammatical errors.

|  |
| --- |
| **Team Member’s Name, Email and Contribution:** |
| 1. Turab Ahmed ([thurab455@gmail.com](mailto:thurab455@gmail.com)) 2. Gottipati Gopichand ([gottipatigopichand0416@gmail.com](mailto:gottipatigopichand0416@gmail.com))  * Detected Null values, removed them from the dataset. * Performed EDA. * Applied Recency, Frequency and Monetary value to understand the customers contribution to the company/organization. * In RFM metrics, score of ‘1’ is best and ‘4’ is worst score. * Applied Log transformation to make the dataset as symmetric. * Performed data preprocessing and applied Scaler transformation. * Used Elbow method to find number of optimal clusters. * Applied K-means clustering algorithm to segment the customers. |
| **Please paste the GitHub Repo link.** |
| Turab Github Link:- [Thurab/Customer-Segmentation (github.com)](https://github.com/Thurab/Customer-Segmentation)  Gottipati Gopichand Link:- [gopichand0416/Customer-segmentation: Online retail customer segmentation (github.com)](https://github.com/gopichand0416/customer-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)** |
| **Problem Statement: -**  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.  **Approach :-**  Below are the approaches to analyze the customers and segment them,   * Detected Null values, removed them from the dataset. * Performed EDA. * Applied Recency, Frequency and Monetary value to understand the customers contribution to the company/organization. * In RFM metrics, score of ‘1’ is best and ‘4’ is worst score. * Applied Log transformation to make the dataset as symmetric. * Performed data preprocessing and applied Scaler transformation. * Used Elbow method to find number of optimal clusters. * Applied K-means clustering algorithm to segment the customers. |
| **Conclusions :-**   * We have a dataset of 541,908 rows and 8 columns. * The dataset contains transactions occurred between **01/12/2010 and 09/12/2011**. * The dataset contains ‘Null’ values of 132,220 rows, which we have dropped. After dropping the dataset contains **397,884** rows. * Then we performed Exploratory data analysis. Through EDA, we came to know that the dataset contains total of **37 countries**, where UK tops the list with 300k+ invoices and Saudi Arabia at the bottom with only 9 invoices. * The most selling products are ‘White Hanging heart T-light holder’, ‘Regency Cake Stand’ and ‘Jimbo bag red retrospect’ * When analyzed based on hourly bases, **noon session** is where most invoices generated, particularly between **12 PM and 3 PM** * When analyzed based on day, ’**Thursday**’ is the most invoices generated day. * When analyzed based on month, **’November’** month is the most invoice generated month. * Then, we applied Recency, Frequency and Monetary value metrics * We give number ‘1’ for best score and ‘4’ for worst score in RFM metrics * Based on RFM, the count of best customers are 440, Loyal customers are 1080, Aggressive spenders are 1085, Almost lost customers are 165 and Lost customers are 385. * For applying the model, the dataset must be symmetrical in shape for good results, when we look at distribution plot of our data, it was rightly skewed so we introduced **Log transformation** to result in symmetrical data. * For data transformation, we used Standard scaler transformation. * Then, we used **Elbow method** to determine number of optimal clusters. Which gave as cluster **value 4**. * By using **k-means** clustering, we were able to segment the customers in to 4 groups. |