Capstone Project Submission

|  |
| --- |
| **Team Member’s Name, Email and Contribution:** |
| |  |  |  | | --- | --- | --- | | **NAMES** | **E-MAIL** | **CONTRIBUTION** | | Sunil Kumar | 14bbt1019@gmail.com | Entire Project work. | |
| **Please paste the GitHub Repo link.** |
| https://github.com/Sunilkumar17-design/Unsupervised\_Customer\_segmentation.git |
| **Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)** |

**Summary**

**I started with the loading, understanding, and exploring the dataset to see the major trends and insights from data regarding the purchases. I have made a few observations while performing the project on the given dataset. Which are below:**

**There were (541909, 8) rows and columns out of which I  saw that there were null values in the description and CustomerID columns, which were of float and object data types. I have removed the null values as imputing them with mode would not be meaningful.**

**After removal of the null values I had (406829, 8) observation and variables respectively.**

**I did see the overall data distribution and found few points as below:**

●    **In quantity we have values in negative and as well as in Unit Price.**

●    **Found positively skewed distribution of the dataset.**

●    **Once I started exploring further country wise, monthly basis, day basis and hourly basis and as per time zone my findings were below:**

●    **Countries with top customers are: United Kingdom, Germany, France, EIRE and Spain.**

●    **Most customers have purchased in the months of November, October, December and September.**

●    **Most of the customers have purchased the items on Thursday, Wednesday and Tuesday.**

●    **I have seen that afternoon timings are popular for purchasing items.**

●    **Especially 11-12-13-14-15 gave more numbers of customer purchasing.**

●    **Once I have seen the data and for its major minor trends, I then started with modeling techniques which are below:**

●    **Used RFM model to find out the valuable customers based on Recency, Frequency and Monetary values.**

● **While using this model I have seen a few points that there were customers which were having more Recency and more Monetary, more Recency and less monetary.**

● **Similarly, for these combinations I have checked for each customer and found the best set of customers after setting the threshold to 5 and 8 respectively given 1263 customers and 2587 customers with threshold of setting to 8.**

● **I then started with K-mean clustering to cluster the same set of customers and tried with 2 features and 3 features which were (RFM) Recency, Frequency and Monetary.**

● **I checked the cluster formation with the help of Silhouette score and elbow method and DBSCAN.**

● **I found DBSCAN performing good to find out the optimal clusters whereas K-mean clustering is not proven that well with elbow method and silhouette scores. After using all methods I have seen that most of the time optimal numbers of clusters were 2.**