Capstone Project Submission

Instructions:

i) Please fill in all the required information.
ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:

Name - Vinay Kumar

Email – vkvinay6923@gmail.com

Contribution – 1. Data Description
2. Exploratory Data Anaysis
3. Feature Engineering
4.. Models
a. Recency and Monetary
b. Frequency and Monetary
5. Conclusion

Name – Nikhil Lamje

Email – Nikhillamje123@gmail.com

Contribution – 2. Data Description
2. Exploratory Data Anaysis
3. Data Modeling
4.. Models
a. Recency and Frequency
b. Recency , Frequency and Monetary
5. Conclusion

Please paste the GitHub Repo link.

Github Link:- https://github.com/vkvinay0627/Online-Retail-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)

In this project, our task is to identify customer segments within a transnational data set that contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail.

First, we started with data wrangling, checking all the null values that we did have in our data, and then we went for the outliers. We have also filtered out the invoice numbers of cancelled orders, which could be irrelevant for our model.

After that, we performed exploratory data analysis and did univariate, bivariate, and multivariate analyses and saw some interesting relationships between the variables. For example, the majority of customers purchased the gifts during the months of November, October, September, and December.

Using RFM (Recency, Frequency, and Monetary) modelling, we got Recency, Frequency, and Monetary values and found RFM scores and RFM groups by making bins at several quantiles, which are 0.25, 0.50, 0.75, and 1.

During the model development, we chose two machine learning algorithms, which are K-Means Clustering and DBSCAN (Density-based spatial clustering of applications with noise). To find the optimum number of clusters in the case of K-Means by Elbow Method and Silhouette Score on various models like RM, FM, RF, and RFM. and found the number of clusters coming out to be 2. But the number of clusters yielded was not clear and with a definite boundary.

So, after applying DBSCAN on all the models, the number of clusters was found to be 3, which has clear boundaries. Indeed, selecting the number of clusters varies from business to business case.