Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:
Neha Pasi nehapasi05@gmail.com
Please paste the GitHub Repo link.
Github Link:- https://github.com/haynapasi050505/Unsupervised-ML
Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)
We have dataset of shape(541909 , 8) There is some null value present in our dataset like in CustomerID and Description. We have to drop some InvoiceNo which are starts with 'c' because 'c', it indicates a cancellation. Then I calculated Top product based on maximum selling, Bottom 5 Product based on the selling, Top 5 Stock name based on selling. most of the customers are from United Kingdom ,Germany ,France ,EIRE and Spain Least number of customers from Lithuania, Brazil, Czech Republic, Bahrain and Saudi Arabia. Then I converted InvoiceDate columns into date time format. I created a new features from Invoicedate. most numbers of customers have purchase in the month of November ,October and December September and less numbers of customers have purchase in the month of April ,January and February Afternoon Time most of the customers have purchase the item Most of the customers have purchase the items in Afternoon ,moderate numbers of customers have purchase the items in Morning and least numbers of customers have purchase the items in Evening Then I created the RFM model (calculating the Recency, Frequency, Monetary value) And Splited them into four segments using quintiles I calculated Add R, F and M segment value columns

I calculated and Add RFMGroup and RFM score
value column showing combined concatenated score of RFM then I perform log transformation to bring into normal.

Then I done K-Means clustering by Applying silhouette Score Method on Frequency and Monetary & Recency and Frequency.
I done k – Means clustering by applying Elbow method on Recency, Frequency and Monetary.

Then I Used the dendogram to find the optimal number of clusters

Then I done hierarchical clustering.
By applying different clustering algorithm to our dataset .we get the optimal number of cluster is equal to 2.