Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:

1. **Mohd Taufique**: taufiquemohd2@gmail.com
   1) Data Loading, understanding, analyzing data quality issues and Cleaning like Handling null and missing values, Handled cancelled orders etc.
   2) Feature Engineering: Introducing new features from invoice date, Created total_cost features from Quantity and Unit Price, Created RFM table etc.
   3) Performed EDA (Exploratory Data Analysis) like Top 5 Products, Top 5 Months, Top 5 Customer IDs, Distplot of all numerical features etc.
   4) Data Transformation: Log Transformation applied on Recency, Frequency and Monetary.
   5) Applied Machine Learning Clustering Algorithms:
      - K-Means with Silhouette score
      - K-Means with Elbow method
      - Agglomerative Hierarchical Clustering
   6) Project Summary, PPT, Colab grouping, Technical Documentation.

2. **Sonica Sinha**: sonicasinha2012@gmail.com
   1) Data Understanding, inspection and performed data wrangling.
   2) Feature Engineering:
      - Data Preprocessing
      - Introducing new features
   3) Performed EDA (Exploratory Data Analysis) like Top 10 countries, Customers over years etc.
   4) Data Transformation: Log Transformation, Splitting into quantiles etc.
   5) Applied Machine Learning Clustering algorithms:
      - K-Means Clustering
      - K-Means with Silhouette score and Elbow method
      - Hierarchical Clustering
   6) PPT, Technical Documentation, Project Summary.

Please paste the GitHub Repo link.

Mohd Taufique GitHub Link: - https://github.com/MOHD-TAUFIQUE/Customer-Segmentation-Unsupervised-ML-Capstone-Project-4

Sonica Sinha GitHub Link: - https://github.com/Soni-Test/Unsupervised-ML-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)

Online retail business is growing rapidly now-a-days, customers and wholesalers usually buy products of their interest because it's very easy to select and view more catalogues. It is very Important for companies to provide the products that their customers need and when they need and how much they have the capability to purchase? to deal with these problems, most of the retail companies use customer segmentation technique and I have provided with the dataset of such online retail company of UK which contains all the transactions occurring between 01/12/2010 and 09/12/2011. Our task is to identify major customer segments on this given dataset. The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers.

Our first step was to import the dataset through pandas 'read_excel' then did some data wrangling and analyzed data quality issue. We get into the situation where we have to deal with large NA values in CustomerID and Description column as we all know that customer IDs are uniquely assign to customers so we cannot impute it with other values. Therefore, we dropped the NA values.

Next, EDA (exploratory data analysis) in which visualization of different features has taken into account with bar-plot, distplot, heatmap and line chart. With the help of bar-plot we get the insights of top customers, top countries, top months of sales, top days of sale and top hours.

After that we performed feature engineering where we introduced new features like month, day, year, month_name, hour from invoice date column and then also created RFM (Recency, Frequency and Monetary) table to get more deep insights of particular customers. We assigned labels on each customer ID according to their purchasing power.

In next step, we calculated RFM Score and check features correlation, visualize features distribution, remove skewness and performed data normalization to make data features normally distributed as clustering algorithm require them to be normally distributed.

Finally, we Applied K-Means clustering, implemented K-Means with silhouette score, K-Means with elbow method, DBSCAN and then Agglomerative Hierarchical clustering. To get the better insight of clusters we plotted Dendogram for scaled RFM.

Conclusion:

- K-Means with Silhoutte_Score = Optimal Clusters: (2)
- K-Means with Elbow Method = Optimal Clusters: (2)
- Agglomerative Hierarchical Clustering = Optimal Clusters: (2)
- DBSCAN= Optimal Clusters: (3).