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 Danish:** **Email:** mdanish63364@gmail.com
   1) Data Cleaning:
      - Handling null and missing values.
      - Handling cancelled orders.
   2) Feature Engineering:
      - Introducing new variables with DateTime
      - Introduced Total amount with Quantity and UnitPrice
   3) EDA(Exploratory Data Analysis):
      - Barplots(Top 5 Products, Top 5 Months, Top 5 Customer IDs)
      - Boxplot(handling RFM Outliers)
      - Distplot(All numerical features)
      - Correlation map between all features.
   4) Data Transformation:
      - Introducing RFM table
      - Log Transformation applied on Recency, Frequency and Moetary
   5) Machine Learning Clustering Algorithms
      - K-Means
      - K-Means with Silhouette score
      - K-Means with Elbow method
      - Hierarchical Clustering
   6) Group Colab
2) **Abdul Rahman Talha:** **Email:** rahman88talha@gmail.com
   1) Feature Engineering:
      - Data Preprocessing
      - Introducing new features
   2) EDA(Exploratory Data Analysis):
      - Barplots(Top 10 countries, Customers over years)
      - Distplot
      - Correlation map
   3) Data Transformation:
      - Log Transformation
      - Splitting into quantiles
   4) Machine Learning Clustering algorithms
      - K-Means Clustering
      - K-Means with Silhouette score and Calinski Harabasz
      - Hierarchical Clustering
   5) PPT

Please paste the GitHub Repo link.

Github Link:- Mdanish2020/Customer_Segmentation_Project (github.com)

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 its 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 first step was to import the dataset through pandas 'read_csv' then did some data wrangling and feature engineering. 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 barplot, distplot, heatmap and boxplot. With the help of barplot we get the insights of top customers, top countries, top months of sales, top days of sale and top hours.

After that we created RFM(Recency, Frequency and Monitory) table to get more deep insights of particular customers. We assigned labels on each customer ID according to their purchasing power.

We possessed with some outliers in RFM table, in order to remove that IQR(inter quantile range) has been applied. After that data has been transformed into standard scale.

Applied K-Means clustering, implemented K-Means with silhouette score, K-Means with elbow method and then Hierarchical clustering. To get the better insight of clusters we plotted Dendogram for scaled RFM. Then I cut-off the Dendogram at the threshold of 50 Euclidian distance.

Conclusion:

- **K-Means** = Optimal Clusters(**3**)
- **K-Means with Silhoutte** = Optimal_Clusters: (**2**)
- **K-Means with Elbow Method** = Optimal_Clusters: (**4**)
- **Hierarchical Clustering** = Optimal_Clusters: (**2**)
- **Hierarchical Clustering with cut-off** = Optimal_Cluster: (**3**)