Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:

1. **Name → Meenakshi**
   Email → meenakshicuul@gmail.com
   Role :
   - **Data Munging**
     1. **Date, Month and Year are extracted from "InvoiceDate" column**
     2. **Checking for duplicate and null values**
     3. **Log transformation**
     4. **Scaling transformation**
   - **Data Visualization**
     1. **Distribution plot of numerical columns**
     2. **Heat map**
     3. **Bar plot**
     4. **Pie Chart**
     5. **Elbow plot**
     6. **Scatter plot**
     7. **Point plot**
     8. **Line plot**
   - **Correlation Analysis between independent variables**
   - **RFM(Recency Frequency Monetary) model**
   - **Building Model(Clustering)**
     1. **K-Means with Silhouette Analysis**
     2. **K-Means with Elbow Method**
     3. **Agglomerative Hierarchical Clustering**
   - **PPT**
   - **Group Colab Notebook**

2. **Name → Tushar R. Wagh**
   Email → waghtushar7276@gmail.com
   Role :
   - **Data Munging**
     1. **Date, Month and Year are extracted from "InvoiceDate" column**
     2. **Checking for duplicate and null values**
     3. **Log transformation**
     4. **Scaling transformation**
   - **Data Visualization**
     1. **Distribution plot of numerical columns**
     2. **Heat map**
     3. **Scatter plot**
     4 **Line plot**
     5 **Count plot**
     6 **Calanski Harabasz Score Elbow Plot for K-Means Clustering**
     7 **Bar plot**

- **Correlation Analysis between independent variables**
- **RFM(Recency Frequency Monetary) model**
- **Building Model(Clustering)**
    1. **K-Means with Silhouette Analysis**
    2. **K-Means with Elbow method**
    3. **Agglomerative Hierarchical Clustering**
- **Technical Documentation**

3. **Name → Aditya Singh Thakur**
   Email → imchillingadi@gmail.com
   Role :
    - **Data Munging**
        1. **Checking for unique values**
        2. **Checking for duplicate and null values**
        3. **Log transformation**
    - **Data Visualization**
        1 **Box Plot**
        2 **Elbow Plot**
        3 **Pair plot**
        4 **Scatter plot**
        5 **Line plot**
    - **Correlation Analysis between independent variables**
    - **RFM(Recency Frequency Monetary) model**
    - **Regression Analysis**
        1 **K-Means Clustering with Silhouette analysis**
        2 **K-Means Clustering with Elbow method**
        3 **Agglomerative Hierarchical Clustering**
    - **PPT**

Please paste the GitHub Repo

Github Link :
https://github.com/meena25091992/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)

Business all over the world are growing today.With the help of technology,they have access to a wider market and hence, a large customer base.Customer Segmentation refers to categorizing customers into different groups with similar characteristics .It can help businesses focus on each customer group in different way ,in order to maximize benefits for customers as well as for the business.This project is mainly deals in segmenting the customers of a online retail stores in UK.

The main objective of our project is to segment customers into different groups on the basis of their similarities in same group and difference in other group. The members of one group are

different from other cluster on the basis of their properties and nature.

In this project,our task is to identify major customer segments on a transactional dataset which contain all the transactions occuring between 01/12/2010 and 09/12/2011 for a UK based and registered non-store online retail.The company mainly sells unique all-occasion Gifts.Many customers of the company are wholesalers.

I have applied various Clusters Models in our Online Retail Customer Segmentation such as follows:-

1. K-Means Clustering with Silhouette Analysis.
2. K-Means Clustering with Elbow Method.
3. Agglomerative Hierarchical Clustering.

RFM Model:-

We have used RFM (Recency Frequency Monetary) Model to segments customers into different groups based on their preferences .It is powerful tool to segment the customers.
Recency signifies the days since order, **frequency** signifies the number of times the customer Is been billed and **monetary** signifies the sales each customer has provided.

Some insights:-

1. Optimal no. of cluster is 2 with all models.
2. We have seen that there are both null and duplicate values. So, we drop them.
3. We have used dendograms in hierarchical clustering to find the optimal no. of clusters.
4. Dendogram is a tree-like diagram that records the sequences of merge or splits.More the distance of vertical lines in the dedogram,more the distance between those clusters.
5. We have used Agglomerative clustering with different threshold value and see how clusters differ and find optimal no. of clusters.
6. With the help of distribution plot, we see that our data is positively skewed. So, we apply some kind of transformation i.e. Log Transformation to convert it into a normal distribution.

Some Conclusions:-

1. K-Means Elbow and K-Means silhouette analysis have 2 as optimal no. of clusters
2. We have implemented Cross Validation on different algorithms as CV performs better on small datasets. But, the result is nearly same.
3. The optimal no. of Agglomerative Hierarchical Clustering is 2
4. Ward linkage is best among all linkages method in hierarchical clustering.