**Data Science and Big Data Analysis**

**Course**

**Sample Solution of Assignment**

**Name: Sushant Khanal**

**LU Id: L20551749**

Chosen Dataset:

Dataset: <https://www.kaggle.com/datasets/vjchoudhary7/customer-segmentation-tutorial-in-python/download>

Data Analysis:

Number of Instances: 200

Number of Attributes: 5

There are two types of data

*Ratio:* Age, Income, Spending Score

*Nominal:* Gender

Details description: <https://www.kaggle.com/datasets/vjchoudhary7/customer-segmentation-tutorial-in-python?resource=download>

Implementation plan:

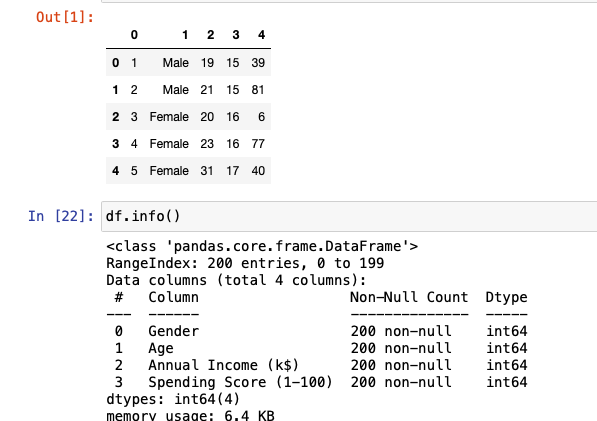
1. Data Exploration
2. Handle missing values
3. Normalize data
4. Plot the correlation matrix
5. Plot the data to search for the elbow to find the appropriate number of clusters
6. Perform k means clustering with “sklearn” library for different clusters
7. Get the summary for the best value
8. Use different distances to get k-Means and purity score with “pyclustering” library
9. Implement clustering again, this time with DBSCAN

Implementation steps:

1. Load the dataset
2. Replace ‘?’ with NA to handle missing values
3. Remove objects with missing nominal attributes
4. Export sample dataset as csv file
5. Convert all possible numeric data attribute to numeric type for further analysis
6. Normalize data
7. Save the correlation matrix as cor.png
8. Perform K means clustering and DBSCAN clustering techniques

Summary of Dataset

This is a hypothetical data taken from Kaggle. The attributes are Customer ID, age, gender, annual income and spending score. All of these attributes have non null values. There are 200 rows.



Summary Statistics:

Table

Description automatically generated

Correlation matrix:

Chart

Description automatically generated with medium confidence

Data Cleaning:

The column “Customer Id” attribute was removed. And in the column “Gender”, 1 was used for male and 0 for female. There were no missing values. If there were any, we would deleted the object or replaced the missing values with the mean values.

Text

Description automatically generated

Table

Description automatically generated

**Normalization:**

Standard Scalar is used for data processing, which normalizes our data.

Graphical user interface, text, application

Description automatically generated

K-means Clustering

First we find the elbow:

In the figure below, we can see that appropriate no. of clusters is somewhere between 5 and 6.

Chart, line chart

Description automatically generated

We plot the clusters(with k = 5and k = 6) and save them as 5ClustersK\_Means.png and 6ClustersK\_Means.png.

­Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Since the plot with 5 clusters is showing a clear segregation, we use k = 5 to get the following table

Table

Description automatically generated

Cluster 0 has high income and high spending score objects. Cluster 1 has objects with medium income and medium spending scores. Cluster 2 show high income but low spending scores. Cluster 3 contain high spending scores and low income. Cluster 4 shows data with low income and low spending scores.

To analyze the affect of different distance functions, “pyclustering” library is used.

Text

Description automatically generated

DBSCAN Clustering Evaluation:

DBSCAN is density based clustering, it has two parameters, minPts and eps. Here we have used eps value of 12 after some trial and error with values from 2 to 20. Minimum point values of 10, 11 and 12 were tried.

We plotted different minimum sample values against number of clusters to predict the appropriate number of minimum samples.

Chart, line chart

Description automatically generated

Therefore, we use DBSCAN to make plots with minimum sample values of 10, 11 and 12.

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

In the first plot, three clusters are identified. Cluster 2 matched with cluster 0 of the K mean Clustering figure. It represents high income objects with high spending scores. Cluster 1 is matching in both cases representing mid incomes with mid spending scores. Cluster 0 here in the above plot matches with cluster 3 in the “5ClustersK\_Means.png” plot.

DBSCAN Clustering summary table(min-samples10):

Table

Description automatically generated