Assignment 4

In this question, you will apply unsupervised learning techniques to identify the segment of population that have greater than 50k earning per year. You are given two sets of data, Dataset 1. Population.csv contains information about the general population, while Dataset 2. more_than_50k.csv contains same information about the people who are making more than 50k per year. You are required to do feature selection, feature extraction and clustering. The goal of the problem is to cluster data in general population and more_than_50k population and analyze which clusters are over-represented in general population vs more_than_50k population and vice versa.

Steps you are required to do:
On Dataset Population.csv

3.1. Preprocessing

3.1.1. Replace missing data with NaN where the missing data is marked by '?'

3.1.2. Perform an assessment of how much missing data there is in each column of the dataset, based on that remove columns with more than 40% data missing.

3.2 Feature Analysis

3.2.1. Plot histogram of values for each feature (both categorical as well as numerical features)

3.2.2. Drop features in which most of data is in one column and there is almost no data in the remaining columns, for example feature 'GRINST'. Make sure you also convert numerical data to categorical data using bins for better analysis in later parts.

3.3. Imputation, Bucketization, One-Hot Encoding

3.3.1. Replace missing values in each column with mode for the column, make sure you store mode for each feature as you will need to replace missing features in more_than_50k dataset with same values.

3.3.2. Bucketize Numerical features

3.3.3. One hot encode features

3.4. Feature Transformation

3.4.1. Fit PCA and analyze cumulative variance vs number of components

3.4.2. Make your decision regarding the number of components.

3.4.3. Fit PCA again with the number of components you chose above.

3.4.4. Interpret Principal Components: Map weights for the first principal component to corresponding feature names and then plot the linked values, sorted by weight. Do this for first 3 principal components.

3.5. Clustering

3.5.1. Apply K-mean clustering with varying values of k in range [10,24] and draw avg within cluster distance vs number of clusters graph.

3.5.2. Based on elbow in the graph, chose best value for k

3.5.3. Apply K-means clustering with best value choosen above.

3.6. Handling more_than_50k data

3.6.1. Apply all the steps you did on general population data to more than 50k population data. While doing this make sure you dont perform operations on this data which which do not align with operations done with population data.

3.7. Compare more_than_50k data with Population Data

3.7.1. Compare the proportion of data in each cluster for the more_than_50k data to the proportion of data in each cluster for the general population.

3.7.2. Find out which clusters are over-represented in general population vs more_than_50k population and vice versa

3.7.3. What kinds of people are part of a cluster that is overrepresented in the more_than_50k data compared to the general population? For this you may need to inverse transform pca to map to orginal features and then analyze the value of centroid of the clusters. You may use features that have highest magnitude for first principal component to analyze the values for the centroid.

3.7.4. Similarly analyze cluster that is overrepresented in the more_than_50k data compared to the general population?

Note: It is advised to make extensive use of Pandas and Numpy