Project 3

Jinting Zhang

**Question 1:** Show the statistics of the dataset to describe the mean value, the standard deviation, etc. for all the development indicators of the EU region.

Answer:

dataset.describe()

**Question 2:** Separate the dataset's numerical indicators (X\_dataset) from the country names (y\_dataset).

Answer:

y\_dataset = dataset.pop("Country\_Name")

X\_dataset = dataset

**Question 3:** Find the normalized input features.

Answer:

X\_dataset\_mean = X\_dataset.mean()

X\_dataset\_std = X\_dataset.std()

X\_dataset\_norm = (X\_dataset - X\_dataset\_mean)/X\_dataset\_std

**Question 4:** Set the number of components in the PCA model to two and then fit the model with the normalized dataset.

Answer:

pca = PCA(n\_components=2)

pca.fit(X\_dataset\_norm)

**Question 5:** Using the found eigenvectors, name the top four development indicators from the original dataset that have the largest impacts on each of the two principal components?

Answer:

Two components give us

[[ 2.09769928e-01 4.97345498e-01 1.34095169e-01 4.88851672e-01

2.97829533e-01 2.71525482e-01 3.26397993e-01 -2.22714435e-01

-6.13966806e-02 3.59654170e-01]

[-3.72799823e-01 -4.67068052e-04 5.23711318e-01 1.40692553e-01

-4.21708884e-01 -6.61906157e-02 9.80983292e-02 1.18597791e-01

5.18470541e-01 3.03697438e-01]]. To determine the top four development indicators that have the largest impacts on each of the two principal components, we need to consider the absolute values of the elements in the given eigenvectors.

The four development indicators with the biggest impact on the 1st principal components:

Index(2)-GDP\_per\_capita\_KUSD-Value(4.97e-01)

Index(4)-GNI\_per\_capita\_KUSD-Value(4.88e-01)

Index(7)-Inflation\_consumer\_prices\_annual\_pct-Value(3.26e-01)

Index(10)-Life\_expectancy\_at\_birth\_total\_years-Value(3.59e-01)

The four development indicators with the biggest impact on the 2nd principal components:

Index(1)-GDP\_growth\_annual\_pct-Value(-3.72e-01)

Index(3)-GDP\_current\_TUSD-Value(5.23e-01)

Index(5)-Exports\_of\_goods\_and\_services\_pct\_of\_GDP-Value(-4.21e-01)

Index(9)-Total\_tax\_and\_contribution\_rate\_pct\_of\_profit-Value(5.18e-01)

**Question 6:** What is the total ratio of variance that has been retained using the two principal components?

Answer:

total\_explained\_variance\_ratio = sum(pca.explained\_variance\_ratio\_)

0.5657820657915604

**Question 7:** Transform the normalized dataset and reduce the dimensions using the PCA model.

Answer:

X\_pca = pca.transform(X\_dataset\_norm)

original shape: (27, 10)

transformed shape: (27, 2)

**Question 8:** Define the k-means clustering algorithm with the suitable number of clusters. Make sure that the random state is set to zero or random\_state=0. Then fit the model using the X\_dataset\_.

Answer:

kmeans = KMeans(n\_clusters=10, random\_state=0).fit(X\_dataset\_)

kmeans.labels\_

**Question 9:** Find the cluster index for each data point in the dataset.

Answer:

clusters = kmeans.predict(X\_dataset\_)

(10, 64)

**Question 10:** Find the accuracy score of the k-means clustering in grouping similar digits.

Answer:

kmeans\_accuracy =accuracy\_score(y\_dataset\_, labels)

0.7885364496382861

**Question 11:** According to the confusion matrix, which number was the most difficult for the model to cluster correctly?

Answer: To identify the most difficult number for the model to cluster correctly, we look for the lowest value on the diagonal, as this represents the class with the fewest correct predictions (true positives). The number '1' has 55 correct predictions, making it the most difficult for the model to cluster correctly. In off-Diagonal Elements, the number '8' has 99 instances misclassified as '1'. In summary, the number '1' is the most challenging for the model to classify correctly because it has the lowest true positive value compared to other classes in the confusion matrix.