Report On Visualization Assignment -Lab2(a)

Dataset Topic: Life Expectancy Analysis

Source of Dataset: Kaggle

Link of Dataset: https://www.kaggle.com/code/yashgupta261100/life-expectancy-analysis/input?select=Life+Expectancy+Data.csv

Attributes:

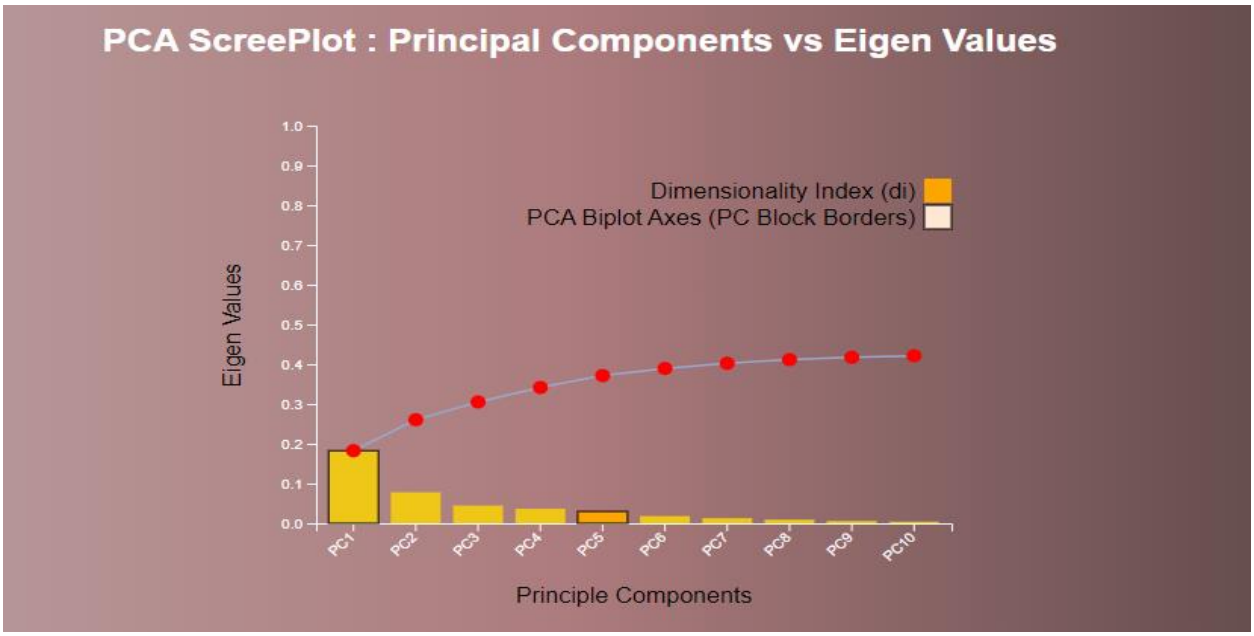
1. **Country:** Skipped here.
2. **Year:** Skipped here.
3. **Status:** Skipped here.
4. **Life Expectancy:** The average number of years a person is expected to live.
5. **Adult Mortality:** Probability of dying between 15 and 60 years per 1000 population.
6. **Infant Deaths:** Number of infant deaths per 1000 population.
7. **Alcohol:** Alcohol consumption measured in liters per capita.
8. **Hepatitis B:** Hepatitis B immunization coverage among 1-year-olds (%).
9. **BMI:** Average Body Mass Index of the entire population.
10. **Polio:** Polio immunization coverage among 1-year-olds (%).
11. **Total Expenditure:** Skipped here.
12. **Diphtheria:** Diphtheria immunization coverage among 1-year-olds (%).
13. **HIV/AIDS:** Deaths per 1000 live births due to HIV/AIDS (0-4 years).
14. **Human Development Groups:** Skipped here.
15. **Population:** Skipped here.
16. **Income Composition of Resources:** Skipped here.
17. **Schooling:** Number of years of Schooling.

This Lab works involves dimensionality reduction using PCA and making of PCA base Scree Plot, making of PCA based Biplot, clustering using K-means and making of K-means MSE Plot, and at last Scatterplot matrix by using all plots observation. Now, the followings are a brief overview of the tasks accomplished and the tools used.

Task 1: PCA-based Dimension Reduction and Visualization

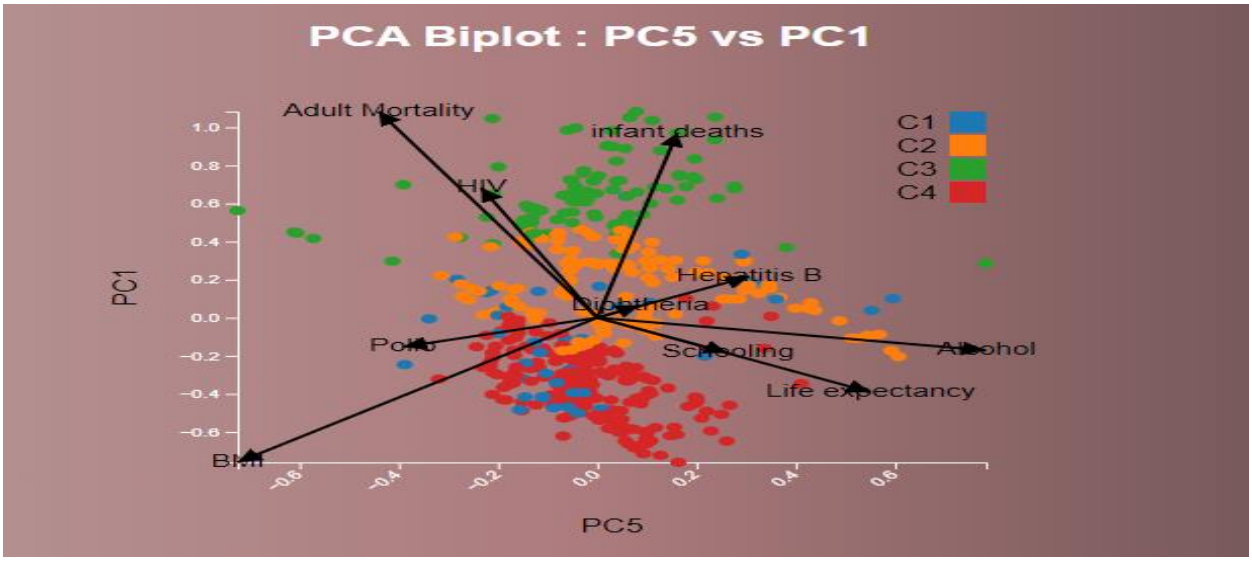
1.1 PCA Scree Plot: Here, we used PCA to compute the Eigenvectors of the data and visualize the Eigenvalues as a scree plot. We added an interaction on click element into the scree plot that allows the user to mark and that can automatically control the intrinsic dimensionality index (di).

The first and most interesting observation is that no one component contributes excessively to the life expectancy. We can see that the first principal component only explains around some percentage of the data. This means it is not a predominant factor. If we increase the dimensionality index (di) by clicking on our own choosing as how many numbers of components we want to see, we can see that a total of the selective principal components is required to explain the data.



1.2 PCA Biplot

We can observe from the biplot that the highest four attributes from the sum of all squared loadings have high loading on x-axis principal component, while other six attributes have high loading on y-axis principal component.

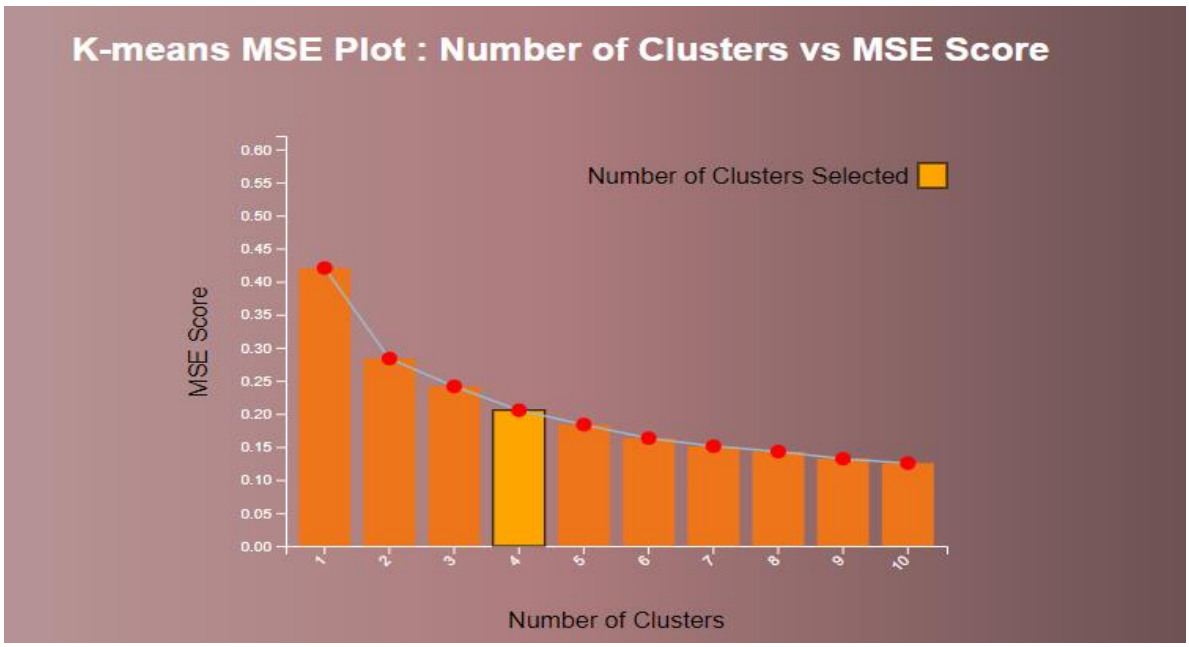


Task 2: Scatterplot Matrix Visualization: This part visualizes the data with a scatterplot matrix. Here, we used the PCA components ≤di to obtain the 4 attributes with the highest squared sum of PCA loadings and list them in a table on the webpage. We can see that the four attributes with the highest squared sum of PCA loadings are always in the scatterplot matrix. Here, the attributes can be chosen using the initial intrinsic dimensionality and the points are colored according to the initial k

Task 3: K-means Clustering and Visualization

3.1 Elbow Method for K-means: Here, we used of the elbow method to determine the optimal number of clusters (k) for K-means clustering.

3.2 Clustering Visualization: Here, we can see the points are colored by selecting the number of clusters (by clicking on the cluster bar) in the PCA based biplot based on the chosen value of k. So, we can see the impact of changing the value of k on the clustering visualization.



4. Additional observations of Implementation

User can choose a different intrinsic dimensionality in the scree plot by coloring the respective bar. Also, user can choose a different k in the k-means MSE plot, again by coloring the respective bar. Also, user can select their desired two principal components for Biplot and it can be nicely observed here. User can choose them in the scree plot by coloring the respective two bars or choosing a different visual marking.