SUSHMA ACHARY

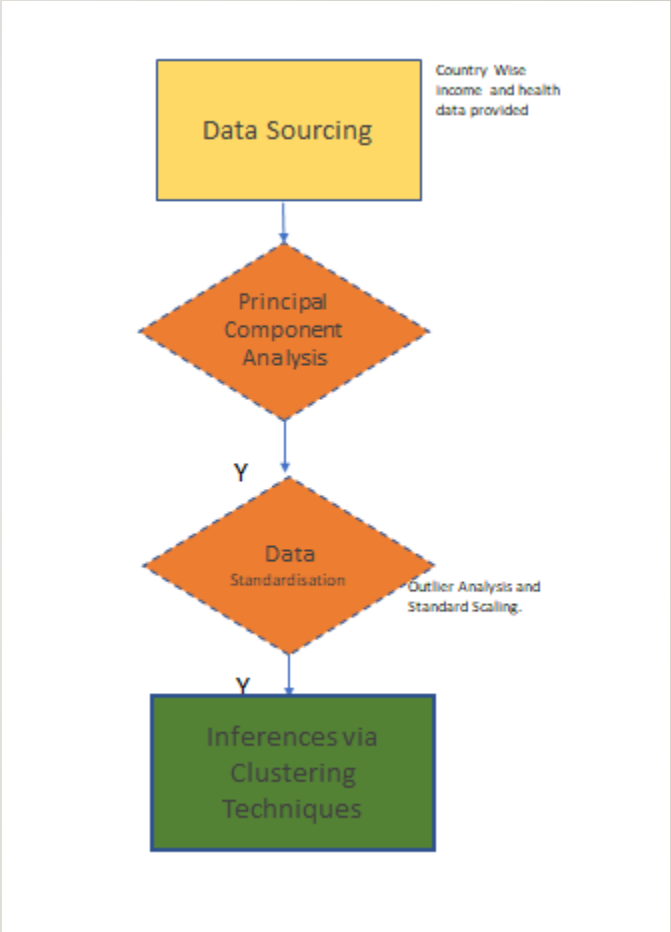
PCA AND CLUSTERING CASE STUDY

SUBMISSION

ABSTRACT

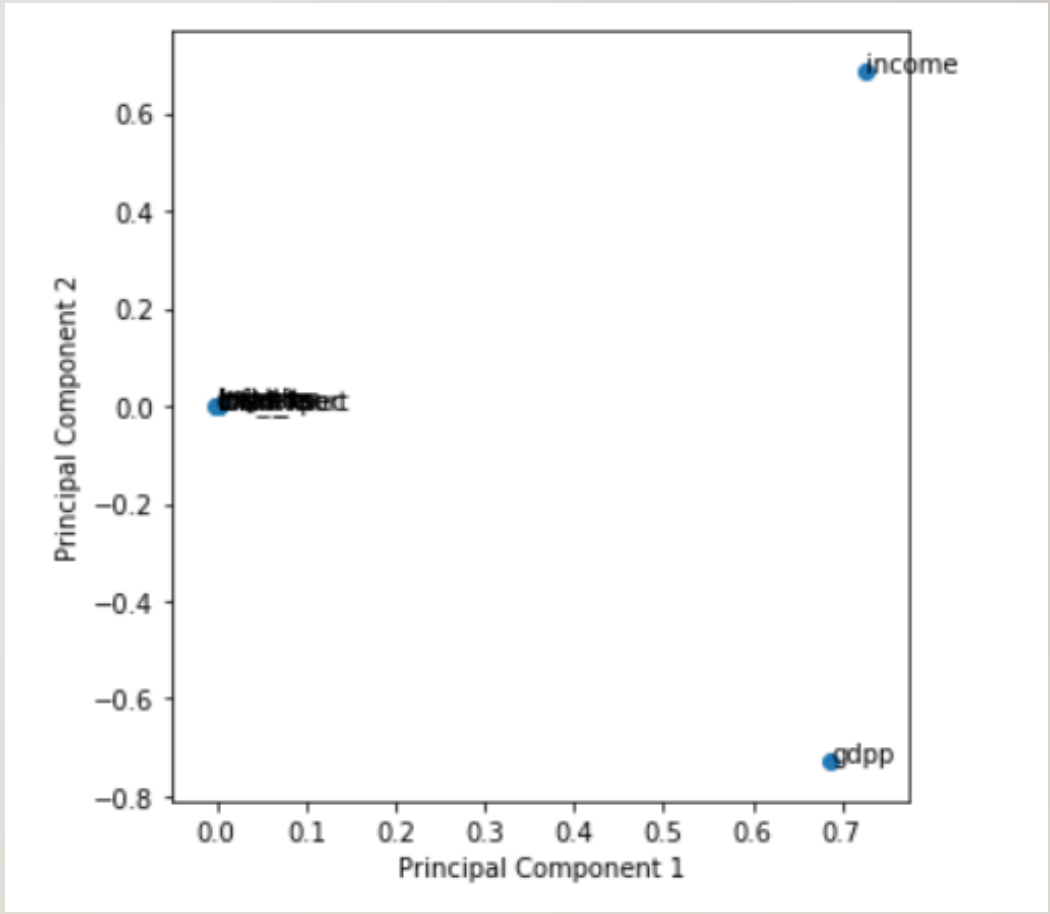
- To analyse the data provided in the case study
- To identify the social, economic and health backgrounds of each country
- Based on them, on the basic understanding of the data, perform Principal Component Analysis
- Using the components from PCA, perform Clustering

PROBLEM SOLVING METHODOLOGY



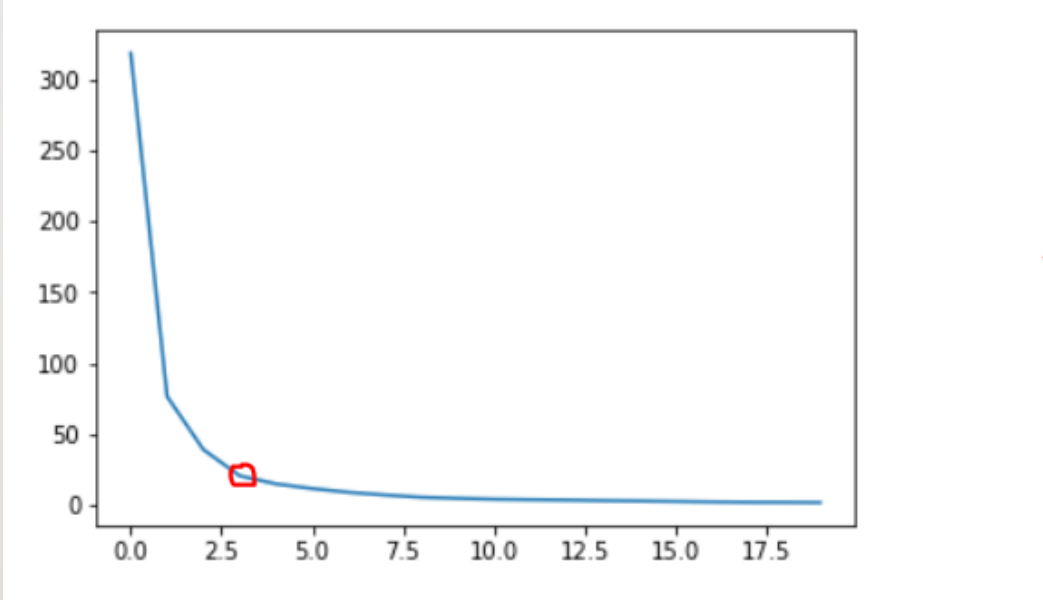
PRINCIPAL COMPONENT ANALYSIS

- After we do a basic PCA, we see that just one PCA alone shows about 95% accuracy, while PC1 and PC2 together show 100%.
- Thus we consider 2 components that we need to use as dimensions out of the 9 numeric columns available in the data set.
- These 2 dimensions are GDPP and Income.



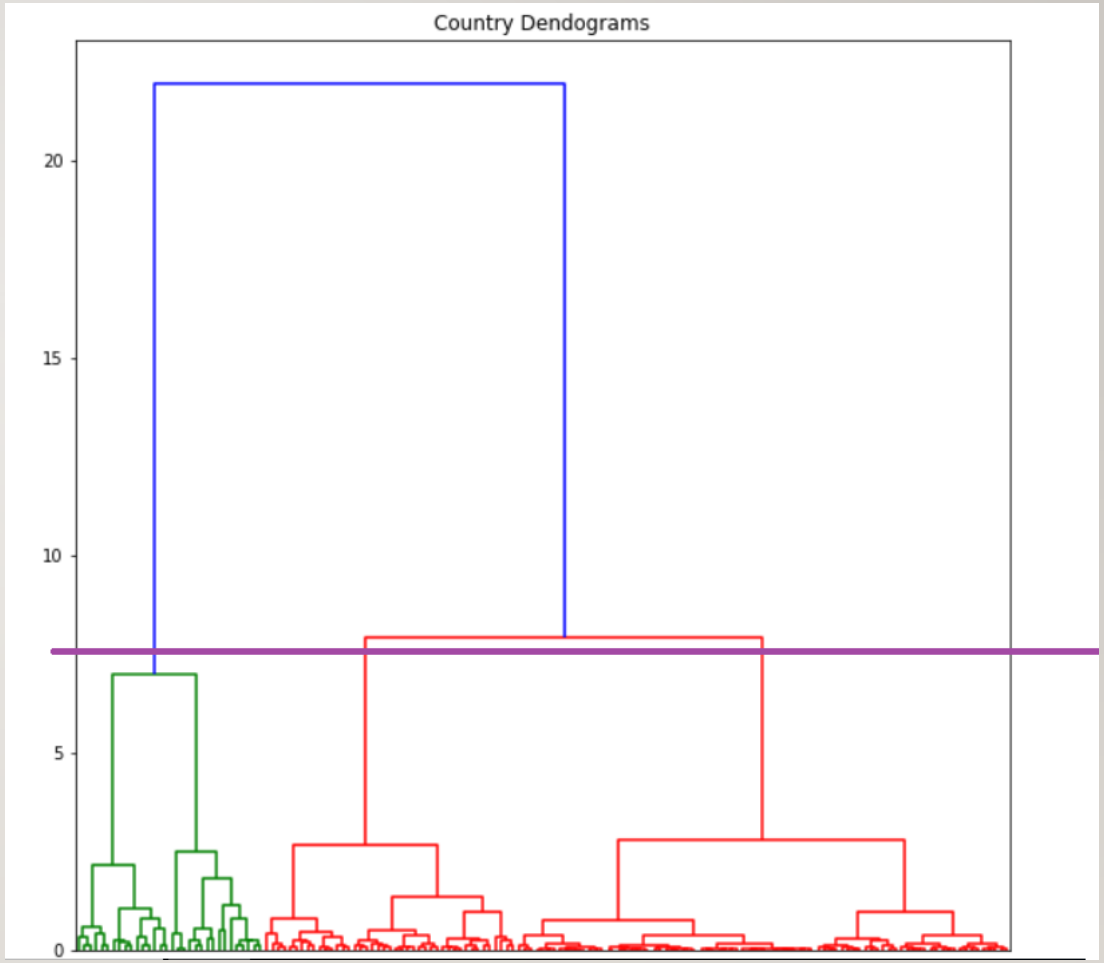
K MEANS CLUSTERING

- Our elbow curve says that at a point slightly after 2.5,as per the curve, is the K value that must be considered. Hence we select K = 3 for our clustering process. Using this, we fit the country data that was standardized with 4 iterations each
- The silhouette score for the same is ~0.62 for K = 3.



HIERARCHICAL CLUSTERING

- After K Means, we try to do a hierarchical clustering.
- Using dendrogram, when we cut the graph at a height of about 7.5, we get K as 3.
- When we plot an agglomerative clustering model, with no. of centroids as 3, we see that the data is distributed with at least 5% in each cluster.



DATA ANALYSIS AFTER CLUSTERING

- Cluster 5 shows a pattern where the health index is the highest meaning a huge chunk of their GDPP goes into health of the population of that country.
- Further, it is not only the income or the GDP ratio that needs to be looked on by the NGO but also the social factors.
- If we are to invest on building a healthy set of generations, spend fair amount on having healthy off-springs in turn increasing the overall life expectancy, we get a good set of population who are young, who can be educated enough to give back to the society in ways innumerable. This would be a long term planning investment option which would eventually boost the economy of a country just on the lines of what the Saudi Arabian prince did.

SCATTER PLOT OF CLUSTER 5

TOP PICK FOR COUNTRIES

- Incidentally, the countries with highest mortality rates also fall within Cluster 5 (which was chosen earlier)
- The country with the highest mortality rate is "Haiti" which is in the NAM, but mostly has African descent.
- There are also mostly countries from the African Continent like Sierra Leone, Chad, CAR, Niger, Mali, Congo etc where they still struggle to have their basic needs fulfilled and there is a lot of limelight in the recent past on how difficult it is to survive with extreme levels of economic and malnutrition issues.

INFERENCES – CLUSTERING