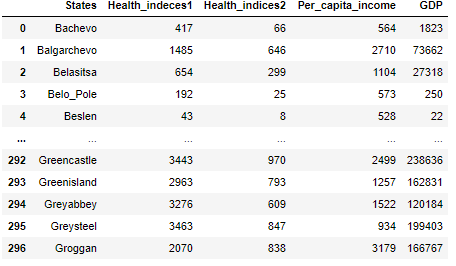
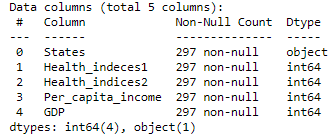
**Clustering – Hierarchical and K Means**

The dataset given is about the Health and economic conditions in different States of a country. The Group States based on how similar their situation is, so as to provide these groups to the government so that appropriate measures can be taken to escalate their Health and Economic conditions.

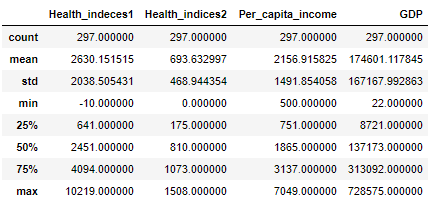
**1. Read the data and do exploratory data analysis.**



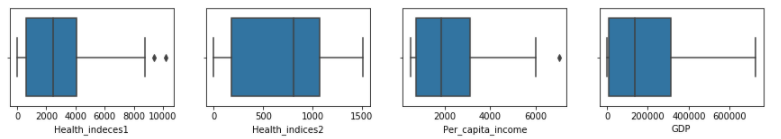
The dataset has 297 records, and 5 features. Health\_indices1, Health\_indices2, Per \_capita\_income and GDP are integers, while the data type of ‘States’ is object. States is categorical feature with 296 unique values, but we will be **dropping the ‘States’** feature as in case of clustering, we have to calculate the distances between data points, and numerical data would be preferred.



**Exploring the dataset:**



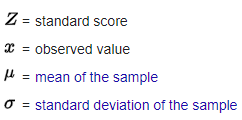
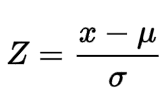
As we can see from above data, all 4 features have varying scales, and Health\_indices1, and GDP having extremely high standard deviation, almost equal to the mean. We will tackle this problem shortly.

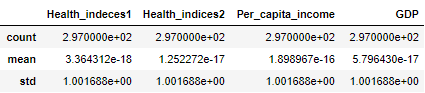


Health\_Indeces1 and Per\_capita\_income have 2 and 1 outlier each, which is not good for clustering, so we will be replacing the outliers with their respective median values.

2. **Scaling of data**

Yes scaling is required in case of clustering, as we calculate the distance between the data points, and cluster on the basis of the distance. Here we have mixed numerical data with varied values and varied scales, therefore values won’t be comparable. We will go for z- standardization, where we will be centering the features at mean = 0 and standard deviation as 1.

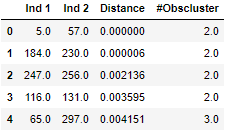




Standardized data, with mean = 0 at standard deviation = 1

**3. Apply hierarchical clustering to scaled data.**

In Agglomerative hierarchical clustering technique, each point is initially considered a single cluster and merge with other clusters in each iteration to eventually form 1 cluster. To calculate the proximity, we will consider ‘ward’ linkage method. Ward linkage calculates the root mean square distance to measure the similarity between two clusters. Metric used will be Euclidean.

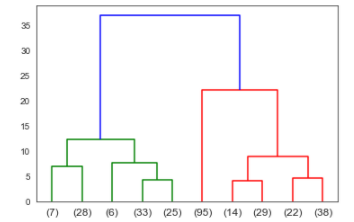


The first row is the first iteration, where observations from Ind 1 and Ind 2 are combined to form a cluster, Distance gives the Euclidean distance between the two points, and #Obscluster gives the number of observations at that time in this cluster.



This is the last row, where after n-1 iterations, 1 cluster is formed, which contains all n observations. Next we will create a dendrogram, which is a visual tree like representation of the clustered data.

**Dendrogram:**



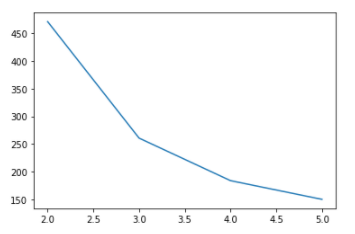
X axis shows the clusters and number of observations in those clusters, y axis shows the distance. Each data point is initially considered as a cluster. Based on similarity of clusters( linkage conditions defined) 2 points merge at a vertical distance to create a new cluster, which further merge to create a new cluster, and so on to create the above clusters( in green and red). These clusters further merge to create 1 cluster which contains all the data (blue line).

Based on the above dendrogram, and taking the distances into consideration, we will cut the dendrogram at distance = 15, thereby dividing the data into 3 clusters.

We get 3 clusters, which can be divided into categories, **high vulnerability, medium vulnerability, low vulnerability** with 96, 104 and 100 observations each.

**4. Apply K-Means clustering on scaled data**

In K- Means clustering, we need to identify the number of clusters, i.e. k in which our dataset would be clustered. The algorithm identifies K number of Centroids and then allocates all data points to the nearest cluster. To identify the ideal number of k, we calculate the inertia of varied values of k i.e. within cluster sum of squares and plot an elbow curve.



A low inertia value is preferred as the total intra cluster variance is minimized, which would make the cluster compact. Looking at the elbow curve, we should select a value of k, where inertia is low and also there is not much change in the inertia value with increasing k. We get the elbow like curve at 3, so we will take k =3, though we also have an elbow at 4, but the elbow at 3 is sharper.

The Silhouette Coefficient is calculated using the mean intra-cluster distance ( a ) and the mean nearest-cluster distance ( b ) for each sample. The Silhouette Coefficient for a sample is (b - a) /max(a, b)

**Silhouette score (for k=3) = 0.535**

**5. Recommend different priority based actions.**

**Hierarchical Clustering**

Below table gives the average value across all features for each cluster.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **#** | **Average** | | | | **#** | **Categories** |
| **Cluster** | **Health\_indeces1** | **Health\_indices2** | **Per\_capita\_income** | **GDP** | **# of obsrv** |
| **2** | **400** | **104** | **680** | **5400** | **96** | **High Vulnerability** |
| 3 | 2480 | 750 | 2350 | 136000 | 104 | Medium Vulnerability |
| 1 | 5000 | 1200 | 3375 | 377132 | 100 | Low Vulnerability |

According to Hierarchical Clustering, 96 observations categorized in cluster 2 having lowest Health\_Indeces1, Health\_indices2, per\_capita\_income and GDP have priority 1 vulnerability, followed by cluster 3. Government should focus on cluster 2 states and improve health conditions.

**K- Means Clustering**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **#** | **Average** | | | | **#** | **Categories** |
| **Cluster** | **Health\_indeces1** | **Health\_indices2** | **Per\_capita\_income** | **GDP** | **# of obsrv** |
| **1** | **500** | **116** | **693** | **9428** | **102** | **High Vulnerability** |
| 3 | 2597 | 783 | 2500 | 140000 | 102 | Medium Vulnerability |
| 2 | 4930 | 1212 | 3385 | 380000 | 96 | Low Vulnerability |

According to K means Clustering, 102 observations categorized in cluster 1 having lowest Health\_Indeces1 , Health\_indices2 , per\_capita\_income and GDP have priority 1 vulnerability, followed by cluster 3. Government should focus on cluster 1 states and improve health conditions.