Clustering Task

In the clustering task, a county dataset is provided which contains information on countries including their child mortality rate, inflation rate, exports, imports etc. The goal is to develop clustering models that will accurately categorize the data set for further analysis.The KMeans clustering algorithm was used to cluster the dataset.

The dataset was loaded onto a Jupyter notebook using the pandas library. There are 167 rows and 10 columns in the dataset and null values were absent.

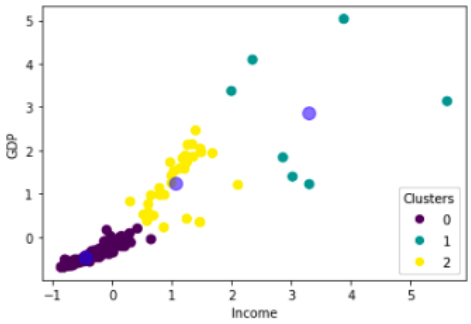
Visualization

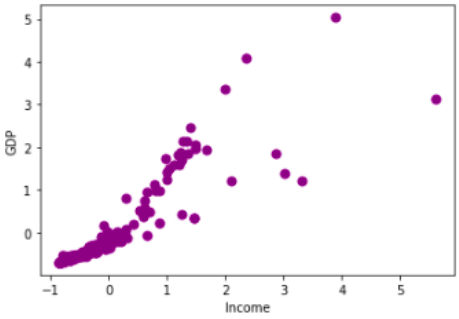
The correlation (.corr()) between the columns of the dataset was plotted and visualized using the seaborn library (heatmap). The GDP and income columns showed the highest correlation, therefore they were chosen as the first two features for clustering.
In a second iteration, three features were used for clustering to check for model improvements.

The scatterplot was also used to visualize the cluster model of 2 features and the 3D plot was used to visualize the cluster model of three features.

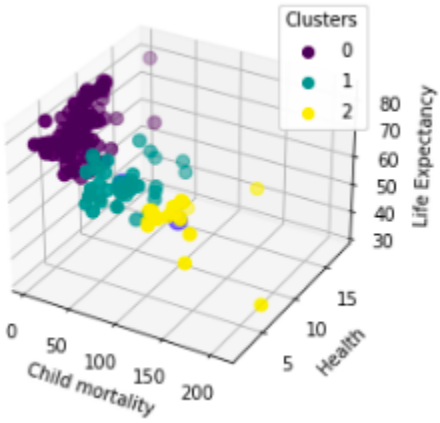
Heat map showing correlation

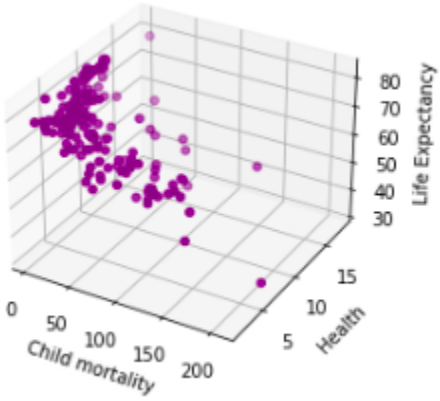
Scatterplot using 2 features





3D plot of three features

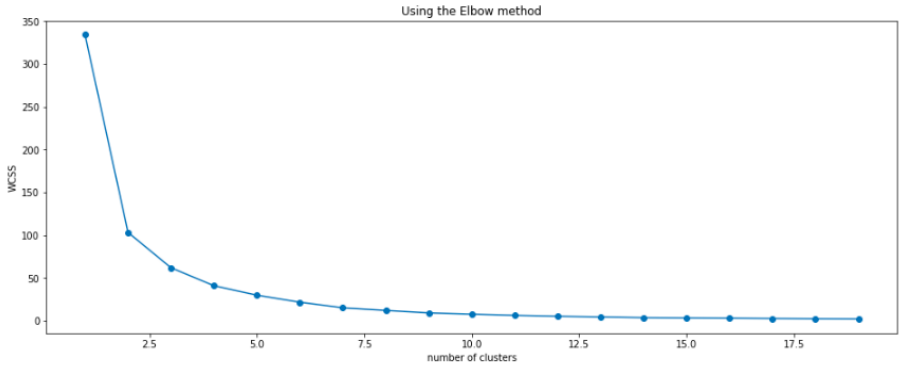




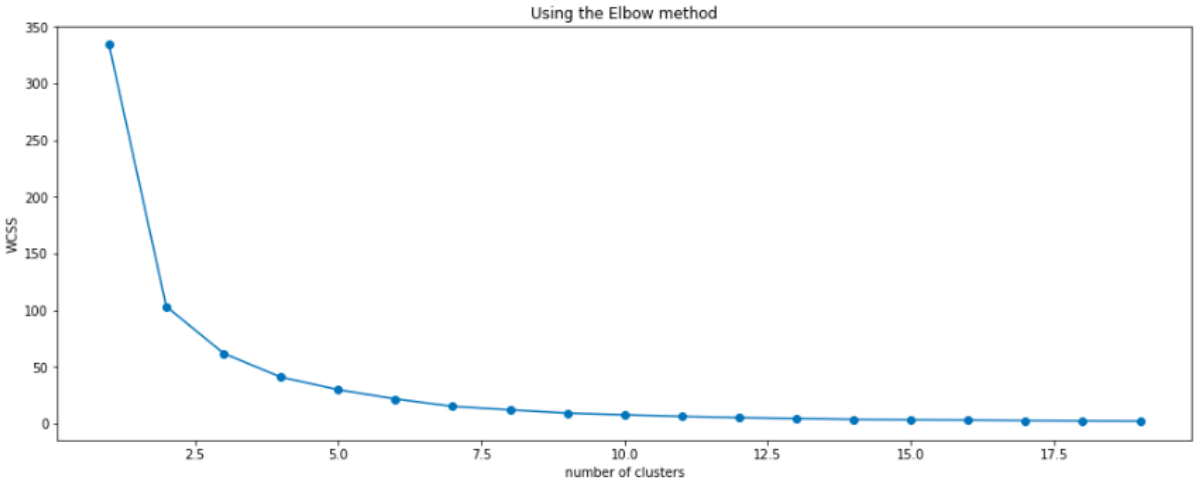
Did clustering change?

The elbow method was used to determine the number of clusters. Even though more features were added, the number of clusters suggested by the elbow method was only slightly impacted and remained largely the same when all features were used for the clustering.

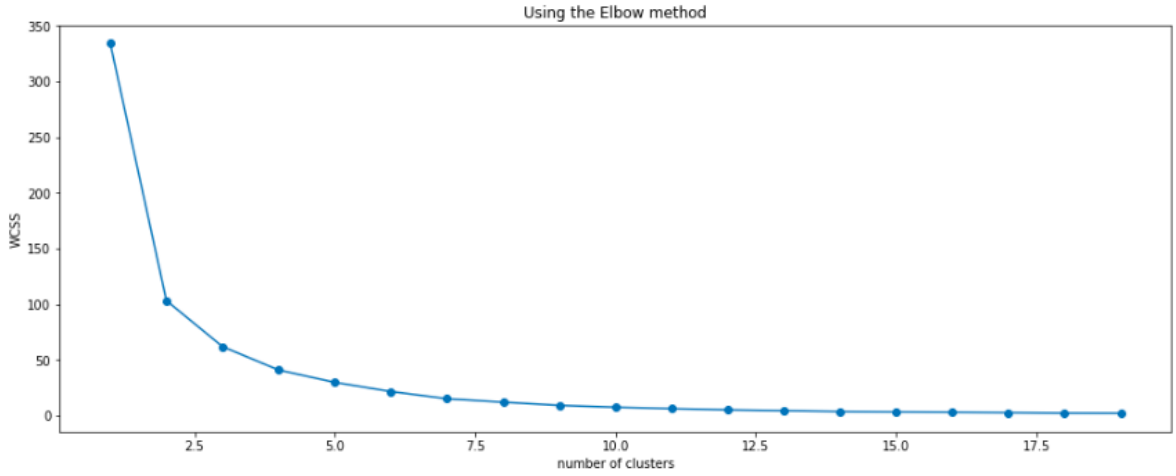
Elbow method at two features



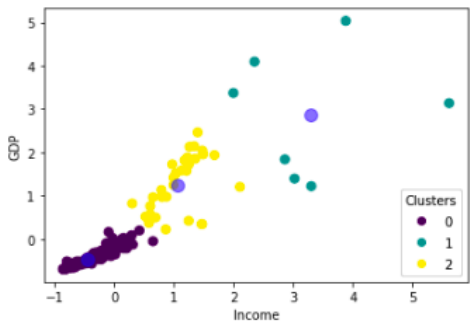
Elbow method at three features



Elbow method at all features



Conclusion



The elbow method predicted the number of clusters to be 3 after fitting all the features into the KMeans algorithm. The data points were clustered into 3 and a column ‘cluster’ was appended to the dataframe to visualize the clusters that each feature fell into. This helped us to better understand the reason for the clusters.

```
In [131]: #showing the reason behind the clustering algorithm
          kmeans_mean_cluster = pd.DataFrame(round(dc.groupby('cluster').mean(),1))
          kmeans_mean_cluster
```

Out[131]:

cluster	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
0	47.4	35.9	6.3	45.9	8569.2	9.0	67.9	3.3	4438.4
1	6.2	96.4	5.7	68.6	80342.9	6.4	80.3	1.8	65442.9
2	8.8	50.0	9.0	46.0	37621.9	3.4	79.2	1.9	35587.5

From here we can conclude that the data points in cluster 0 have higher child mortality, higher inflation, lower income and lower GDP than the other clusters, while the data points in cluster 1 have lower child mortality, lower inflation, higher income and higher GDP than the other clusters. We can further conclude that the countries in cluster 0 are low-income countries while the countries in cluster 1 are high income countries.

Recommendation

In the context of the data provided, economic factors such as a country's export and import, for instance, when rightly coordinated can serve to raise a company’s GDP.