Clustering Assignment

Your Name Vishnu Sharma

Overview

Explanation to various concepts:

- Data understanding and Data Cleaning
- EDA & Data Visualization
- K-Mean & Hierarchical clustering
- Conclusion

Data Understanding & Cleaning

Data Understanding:

- Read the data and understand the columns based on the check the null values
- Convert the export, health and imports to their actual values.
- After reading the data using pandas, we can see the data based on country and how the child death are there in each country
- As we can see in info there is no null values in the given data.

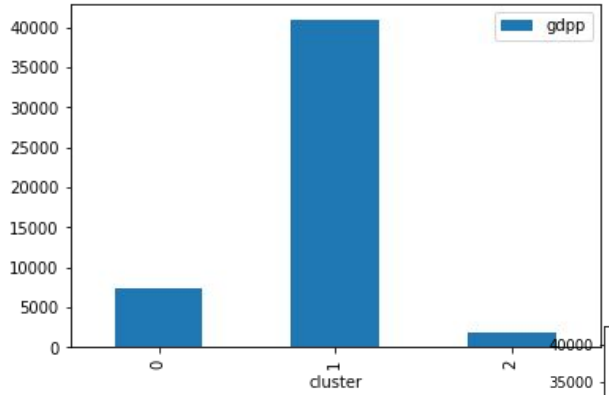
EDA & Data Visualization

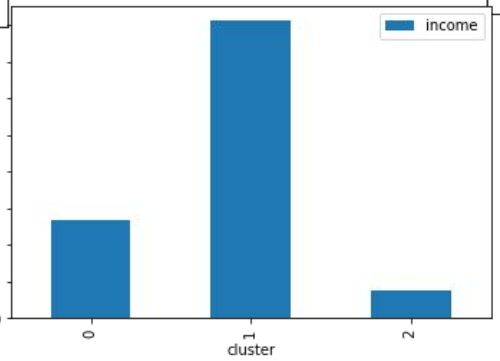
- The first step to the EDA is plotting the data into subplot and as we can see there are most of the columns having same line of graph.
- Since our analysis is towards gdpp, income, child_mort. So we will be closely focus on the three graphs.
- As can see in the scatter plot using all three fields and can understand the data and increasing/decreasing based on gdpp, income, child_mort.
- Also in boxplot we can see there are upper outlier in all three columns. So I have treated upper outlier using .95.
- I have not treated upper outlier of child_mort because most of the country could have high child death with could be removed.

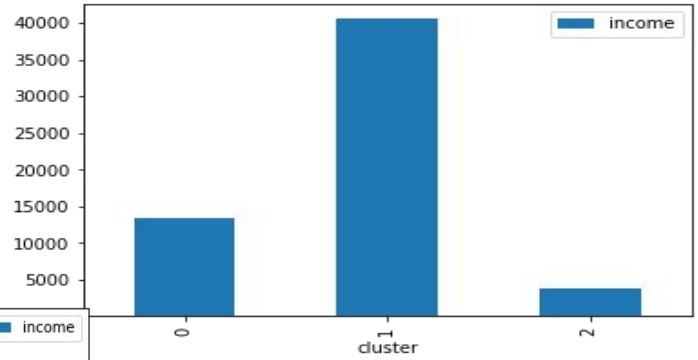
K-Mean Clustering

- Using K-mean clustering we are trying to find out which top 5 country are having high child mortality
- So first we are trying to find the number of k using Silhouette score and Elbow-curve ssd.
- Also we are checking the hopking score based on that we are trying to find out whether our data is good for clustering or not.
- After using all these steps we can see the best k value is 3.
- After performing the k-mean algo, we are visualizing the cluster with gdpp, income, child_mort
- Finally we can see that the low gdpp, low income and high child mortality top 5 country.

K-Mean Clustering Graph



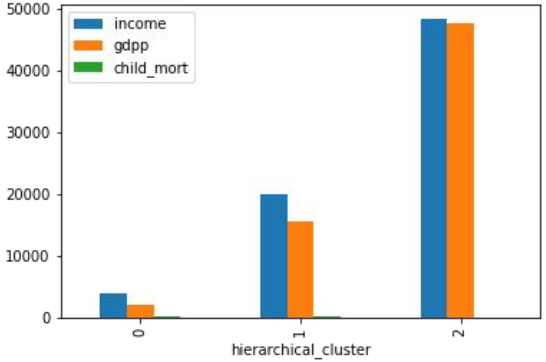


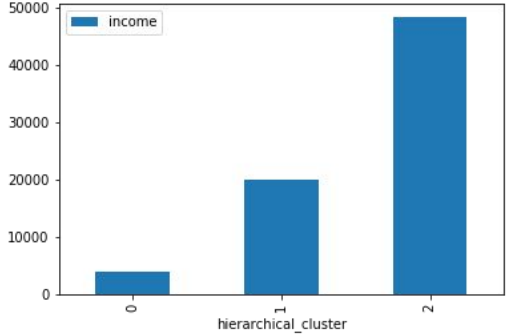


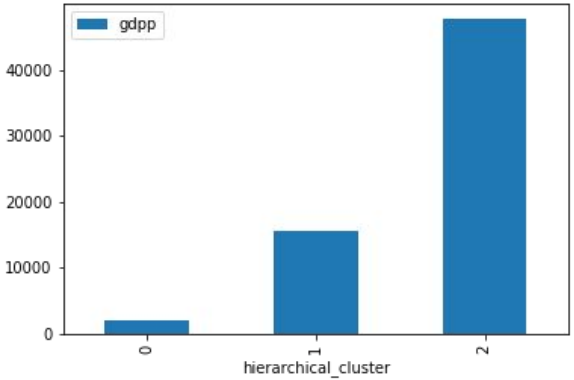
Hierarchical Clustering

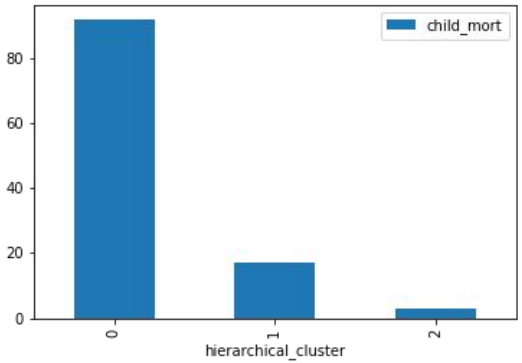
- Using hierarchical clustering we are trying to find out which top 5 country are having high child mortality
- So first we are trying to find the number of k using single/complete linkage.
- After that we can see the complete linkage having the best cut_tree as k = 3.
- We are visualizing the hierarchical cluster with gdpp, income, child_mort
- Finally we can see that the low gdpp, low income and high child mortality top 5 country.

Hierarchical Clustering Graph









K-Mean top 5

Filter K-Mean Cluster top 5 country

- Low income, Low GDP and High Child_mort
- Filter the data for that clsuter

[229]:
```
# Top 5 Country with low income, low gdpp and high child_mort

df_kmean[df_kmean['cluster'] == 2].sort_values(by =
                                               ['income', 'gdpp', 'child_mort'],
                                               ascending = [True, True, False]).head(5)
```

[229]:

	country	child_mort	income	inflation	life_expec	total_fer	gdpp	exports_actual	health_actual	imports_actual	cluster
37	Congo, Dem. Rep.	116.0	609.0	20.80	57.5	6.54	334	137.2740	26.4194	165.664	2
88	Liberia	89.3	700.0	5.47	60.8	5.02	327	62.4570	38.5860	302.802	2
26	Burundi	93.6	764.0	12.30	57.7	6.26	231	20.6052	26.7960	90.552	2
112	Niger	123.0	814.0	2.55	58.8	7.49	348	77.2560	17.9568	170.868	2
31	Central African Republic	149.0	888.0	2.01	47.5	5.21	446	52.6280	17.7508	118.190	2

Hierarchical Clustering top 5

Filter hierarchical cluster Data

- Low income, Low GDP and High Child_mort
- Filter the data for that clsuter

46]:

```
# Top 5 Country with low income, low gdpp and high child_mort

df_kmean[df_kmean['hierarchical_cluster'] == 0].sort_values(by =
                                                ['income', 'gdpp', 'child_mort'],
                                                ascending = [True, True, False]).head(5)
```

46]:

	country	child_mort	income	inflation	life_expec	total_fer	gdpp	exports_actual	health_actual	imports_actual	cluster	hierarchical_cluster
37	Congo, Dem. Rep.	116.0	609.0	20.80	57.5	6.54	334	137.2740	26.4194	165.664	2	0
88	Liberia	89.3	700.0	5.47	60.8	5.02	327	62.4570	38.5860	302.802	2	0
26	Burundi	93.6	764.0	12.30	57.7	6.26	231	20.6052	26.7960	90.552	2	0
112	Niger	123.0	814.0	2.55	58.8	7.49	348	77.2560	17.9568	170.868	2	0
31	Central African Republic	149.0	888.0	2.01	47.5	5.21	446	52.6280	17.7508	118.190	2	0

Conclusion

1. After performing the k-mean and hierarchical clustering.
2. K-mean clustering the cluster 2 has the low gdpp, low income and high child-mortality
3. Hierarchical clustering the cluter 0 has the low gdpp, low income and high child-mortality