Clustering Assignment

SUBMITTED BY :
SHUBHANGI TRIVEDI

Problem Statement

HELP International is an NGO that is committed to fighting poverty and providing the people of backward countries with basic amenities and relief during the time of disasters and natural calamities. We need to categorise the countries using some socio-economic and health factors that determine the overall development of the country.

The company wants to provide relief fund of $ 10 Million to 5 countries that are in immediate need of aid.

Approach to Analysis

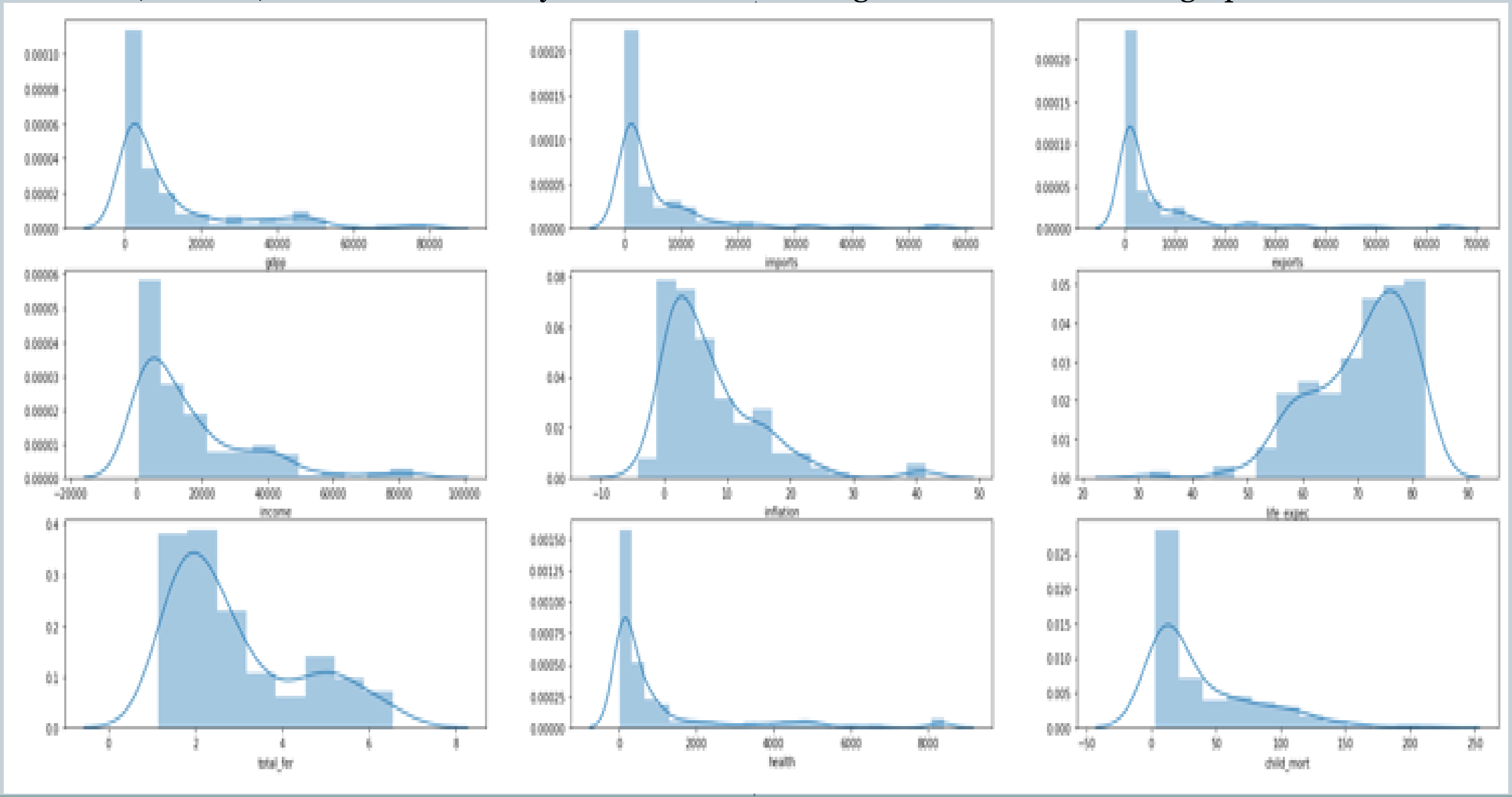
- Read and Visualize the data
  - Check for data type and shape of data.
  - Checked for nulls
  - Converted columns exports, imports and health to actual value from percentage of GDPP.
- Exploratory Data Analysis – For analysis used below approach
  - Univariate Analysis – Distplot
  - Multivariate Analysis – Pair Plot and Heat Map
- Outlier Treatment
  - For columns - 'gdpp', 'imports','exports','income', 'inflation','life_expec','total_fer', 'health', 'child_mort‘, changed upper outliers to 0.99 quantile value.
- Hopkins Score
- Clustering
  - Kmeans Algorithm
  - Hierarchical Algorithm
- Final Conclusion

Exploratory Data Analysis Results

Univariant Analysis- Dist Plot

POINTS CONCLUDED:

- GDPP, Income, Child_mort are in sync and their value range can be seen from the graph.

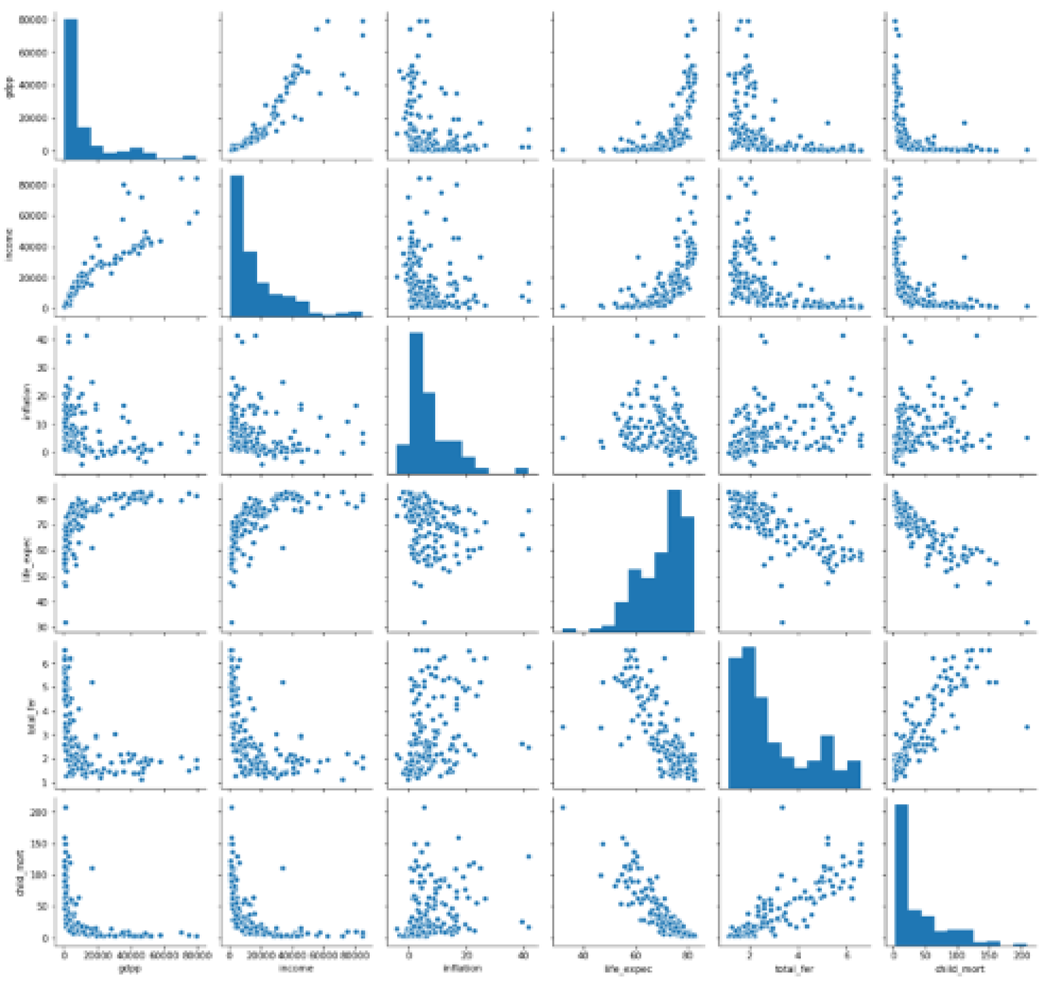


Multivariant Analysis-Pair Plot

POINTS CONCLUDED:

So from adjacent graph we can conclude that :

- Child Mortality & Fertility, GDPP & Income are directly proportional.
- GDPP & Child mortality, Income & Child mortality are inversely proportional.

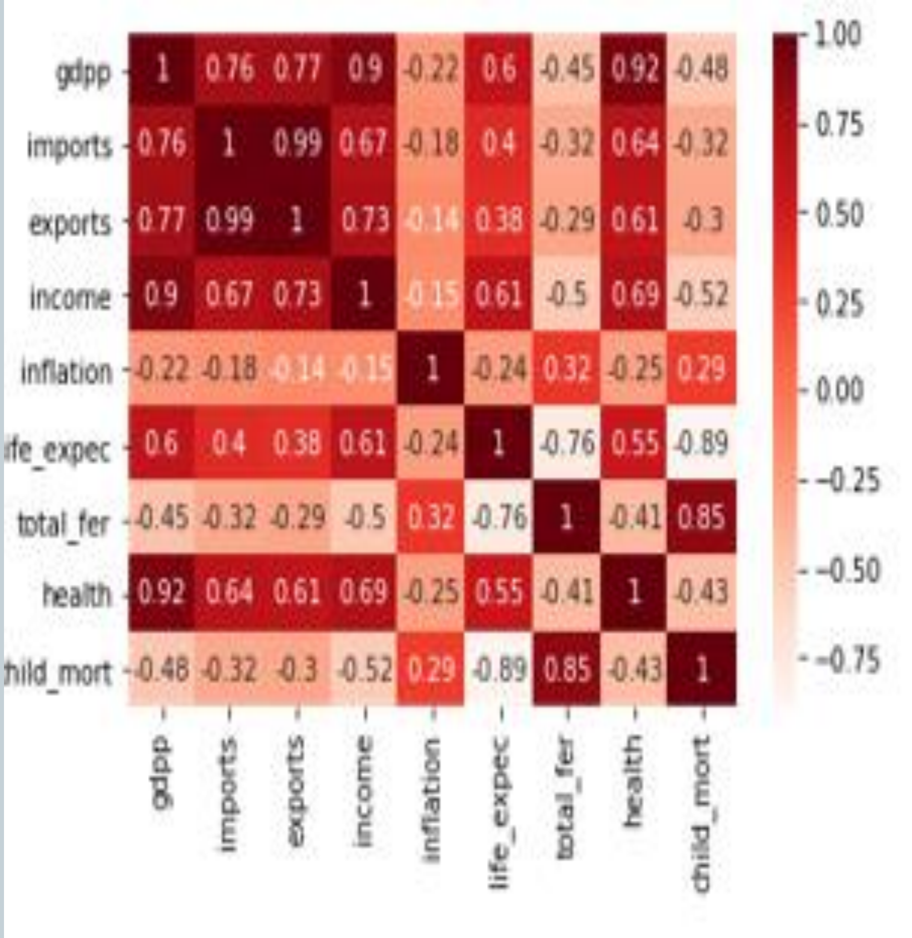


Multivariant Analysis-Heat Map

POINTS CONCLUDED:

We can conclude from adjacent heat map that -

- There is a strong correlation between child mortality, total fertility and life expectancy. However, life expectancy and child mortality are negatively correlated.
- Imports, exports and per capita income are also strongly correlated.
- Per capita health spending is strongly correlated with GDPP



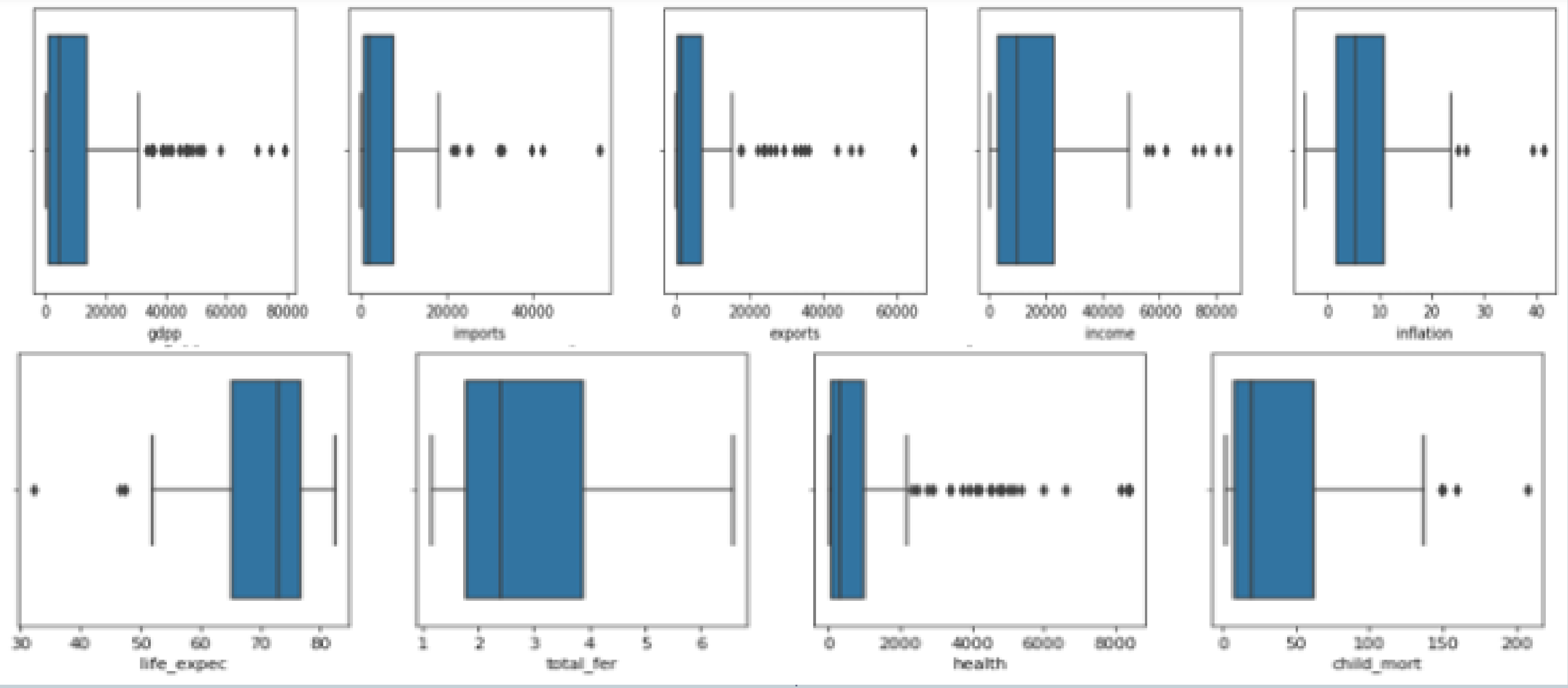
	gdpp	imports	exports	income	inflation	life_expec	total_fer	health	child_mort
gdpp	1	0.76	0.77	0.9	-0.22	0.6	-0.45	0.92	-0.48
imports	0.76	1	0.99	0.67	-0.18	0.4	-0.32	0.64	-0.32
exports	0.77	0.99	1	0.73	-0.14	0.38	-0.29	0.61	-0.3
income	0.9	0.67	0.73	1	-0.15	0.61	-0.5	0.69	-0.52
inflation	-0.22	-0.18	-0.14	-0.15	1	-0.24	0.32	-0.25	0.29
ife_expec	0.6	0.4	0.38	0.61	-0.24	1	-0.76	0.55	-0.89
total_fer	-0.45	-0.32	-0.29	-0.5	0.32	-0.76	1	-0.41	0.85
health	0.92	0.64	0.61	0.69	-0.25	0.55	-0.41	1	-0.43
hild_mort	-0.48	-0.32	-0.3	-0.52	0.29	-0.89	0.85	-0.43	1

Outlier Treatment

POINTS CONCLUDED:

In this case study, we need to find out countries that are actually in need of AID and that can be decided by high child mortality and low value of income, health, total fertility, gdpp, exports and imports , thus we cannot treat high value of child mortality and low value for all other columns

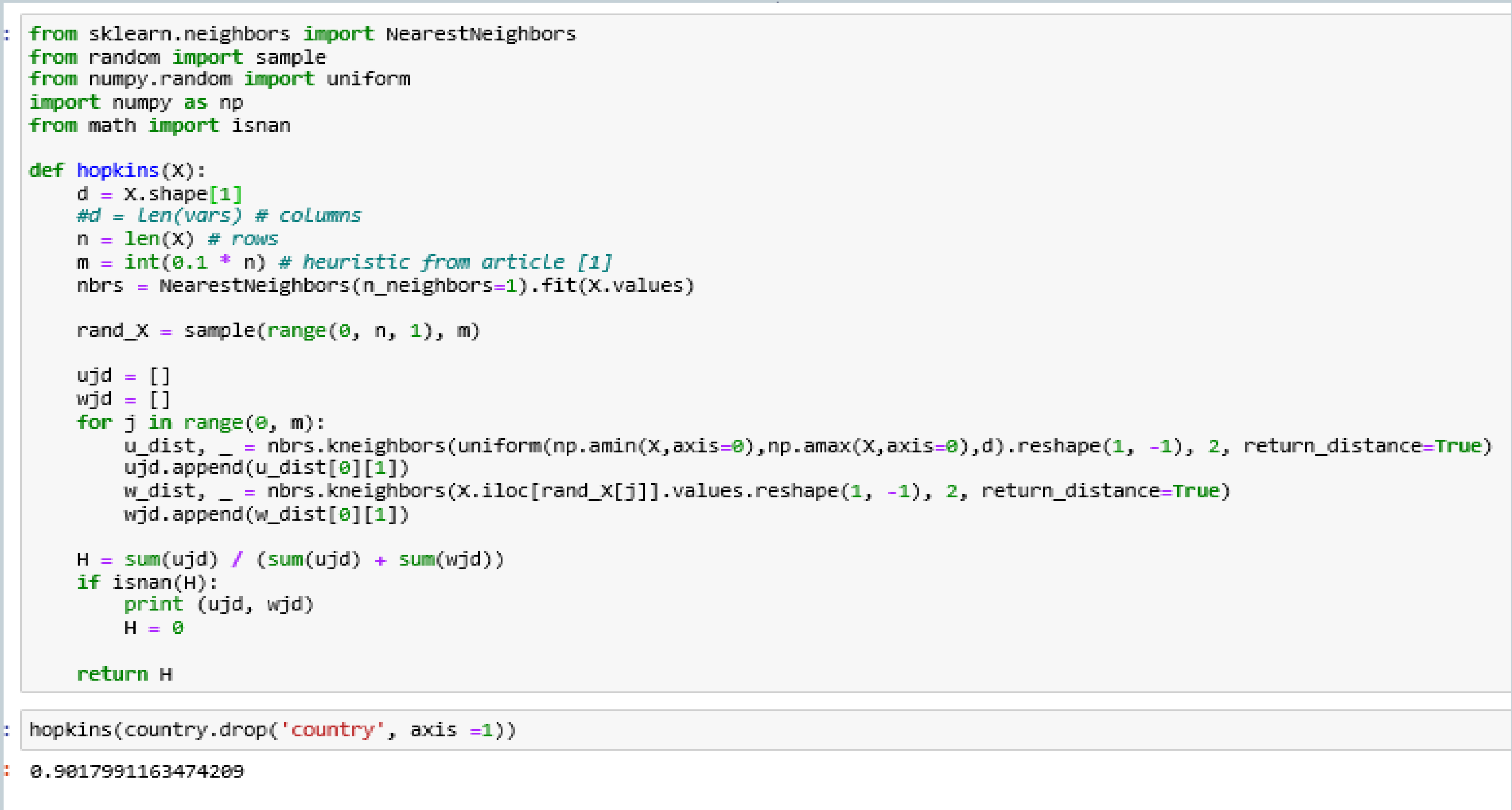
- For all columns except, child mortality, we will treat only higher outliers and will refrain from treating lower values.
- For child mortality, we will treat only lower values and will refrain from treating higher values.



HOPKINS SCORE

POINTS CONCLUDED:

Since Hopkins score is coming more than 90% that means that clustering is feasible on the available data



```
from sklearn.neighbors import NearestNeighbors
from random import sample
from numpy.random import uniform
import numpy as np
from math import isnan

def hopkins(X):
    d = X.shape[1]
    #d = len(vars) # columns
    n = len(X) # rows
    m = int(0.1 * n) # heuristic from article [1]
    nbrs = NearestNeighbors(n_neighbors=1).fit(X.values)

    rand_X = sample(range(0, n, 1), m)

    ujd = []
    wjd = []
    for j in range(0, m):
        u_dist, _ = nbrs.kneighbors(uniform(np.amin(X,axis=0),np.amax(X,axis=0),d).reshape(1, -1), 2, return_distance=True)
        ujd.append(u_dist[0][1])
        w_dist, _ = nbrs.kneighbors(X.iloc[rand_X[j]].values.reshape(1, -1), 2, return_distance=True)
        wjd.append(w_dist[0][1])

    H = sum(ujd) / (sum(ujd) + sum(wjd))
    if isnan(H):
        print (ujd, wjd)
        H = 0

    return H
```

```
hopkins(country.drop('country', axis =1))
```

```
0.9017991163474209
```

K-Mean Clustering Algorithm

ALGORITHMS STEPS :

STEP ONE: INITIALIZE CLUSTER CENTRES

STEP TWO: ASSIGN OBSERVATIONS TO THE CLOSEST CLUSTER CENTRE

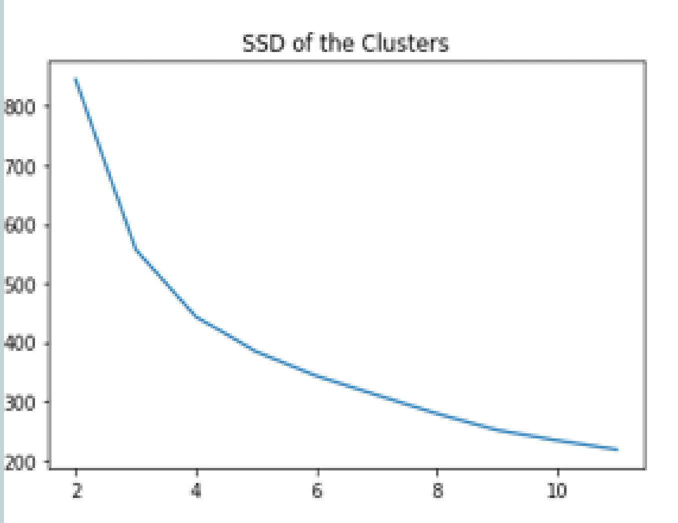
STEP THREE: REVISE CLUSTER CENTRES AS MEAN OF ASSIGNED OBSERVATIONS

STEP FOUR: REPEAT STEP 2 AND STEP 3 UNTIL CONVERGENCE

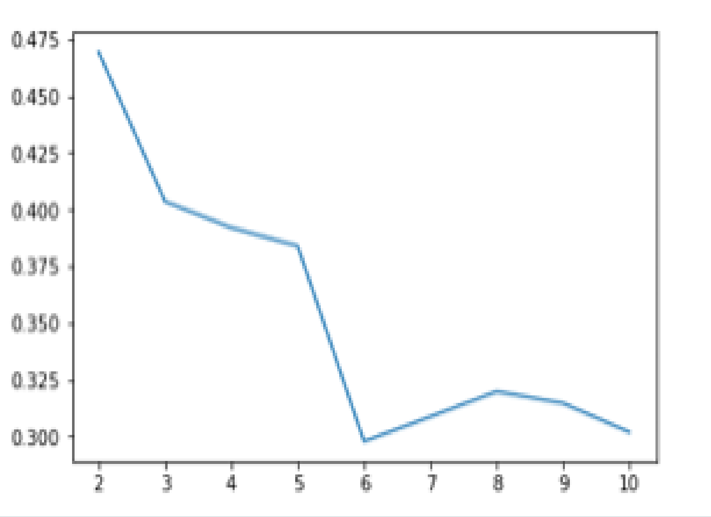
Finding number of clusters(K-Value)

POINTS CONCLUDED:

- Looking at Silhouette Score, we found out that max value is achieved at k = 2, but we will not choose 2 as this is logically like dividing clusters into 2 equal parts, so we will go with next highest value that is 3.
- Looking at elbow curve, we can conclude that elbow is forming at 3 and thus we will consider 3 as the value of clusters



Elbow Curve

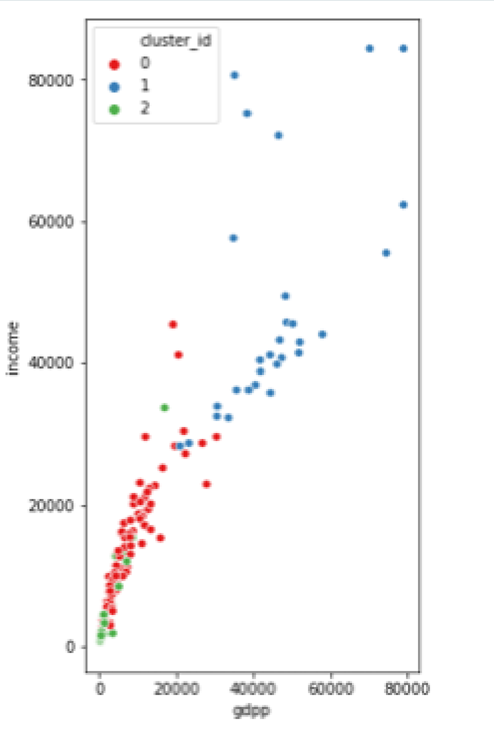


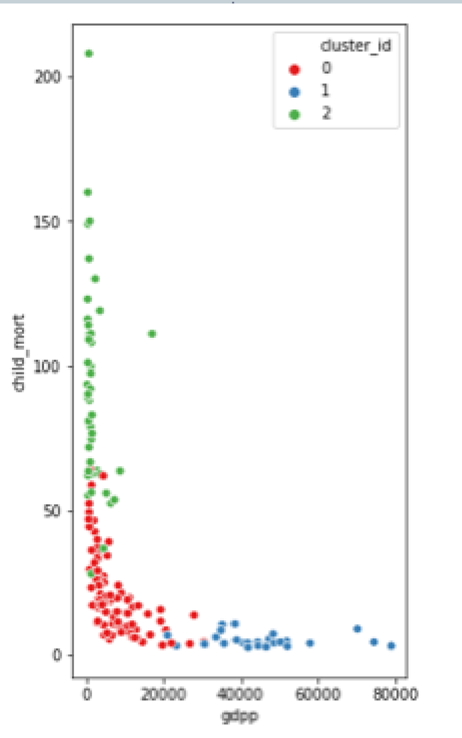
Silhouette Score

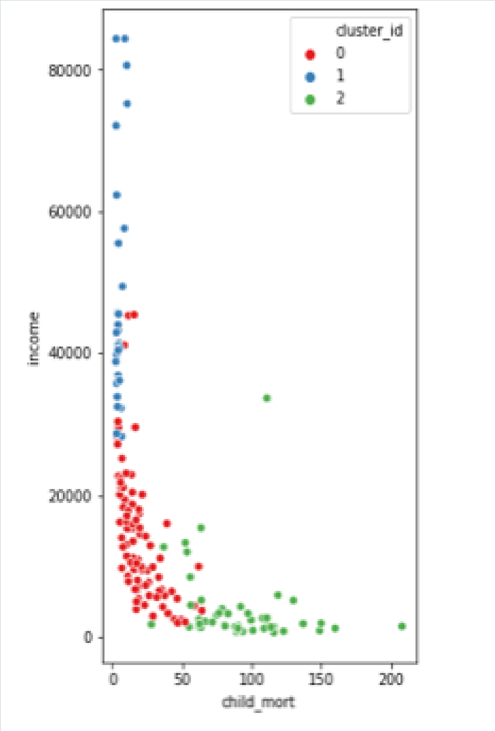
Plotting the Clusters

POINTS CONCLUDED:

- From Graph 1, we conclude that GDPP & Income are directly proportional to each other, that means if for a country income increases GDPP also increases & vice-versa.
- From Graph 2, we conclude that GDPP & Child mortality are inversely proportional
- From Graph 3, we conclude that Income & Child mortality are inversely proportional. that means if for a country income decreases child mortality increases & vice-versa







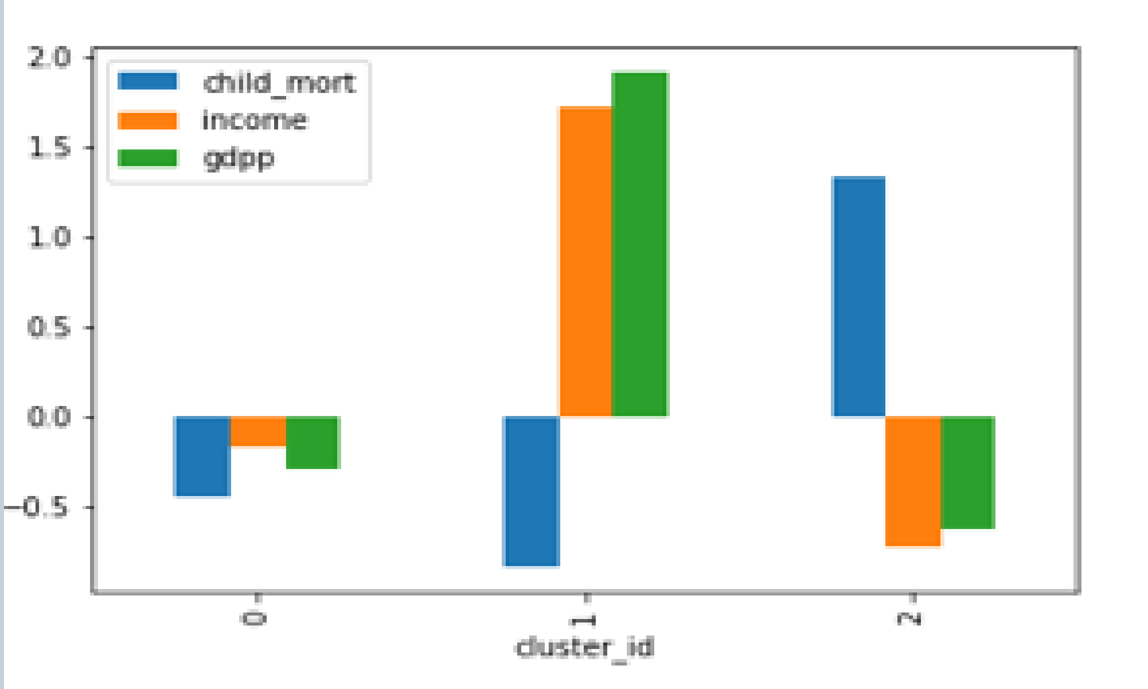
Cluster Profiling

POINTS CONCLUDED:

So from below graph we can conclude that :

- - Cluster 0 have countries which have child_mort, income and gdpp negative
- - Cluster 1 have countries with negative child_mort, and positive income and gdpp
- - Cluster 2 have countries with negative income and gdpp and positive child_mort

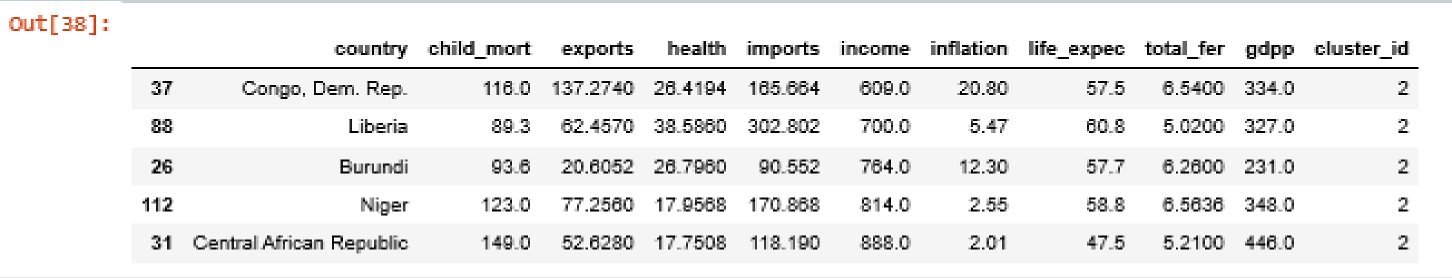
So, top 5 countries that need urgent AID are countries falling in cluster label 2



Final Conclusion-K Mean

So, we can conclude that countries list below are the countries that needs urgent AID :

1. Congo, Dem. Rep.
2. Liberia
3. Burundi
4. Niger
5. Central African Republic



Out[38]:

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	cluster_id
37	Congo, Dem. Rep.	116.0	137.2740	26.4194	165.664	609.0	20.80	57.5	6.5400	334.0	2
88	Liberia	89.3	62.4570	38.5860	302.802	700.0	5.47	60.8	5.0200	327.0	2
26	Burundi	93.6	20.6052	26.7960	90.552	764.0	12.30	57.7	6.2600	231.0	2
112	Niger	123.0	77.2560	17.9568	170.868	814.0	2.55	58.8	6.5636	348.0	2
31	Central African Republic	149.0	52.6280	17.7508	118.190	888.0	2.01	47.5	5.2100	446.0	2

Hierarchical Clustering Algorithm

ALGORITHMS STEPS :

STEP ONE: COMPUTE THE PROXIMITY MATRIX

STEP TWO: LET EACH DATA POINT BE A CLUSTER

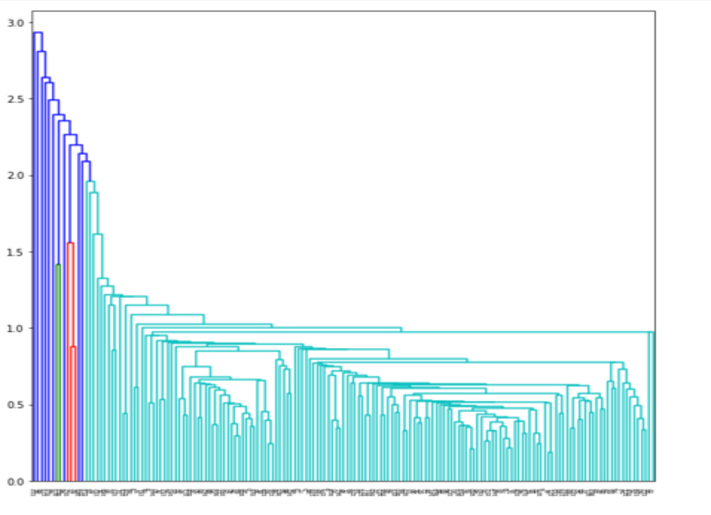
STEP THREE: REPEAT: MERGE THE TWO CLOSEST CLUSTERS AND UPDATE THE PROXIMITY MATRIX

STEP FOUR: UNTIL ONLY A SINGLE CLUSTER REMAINS

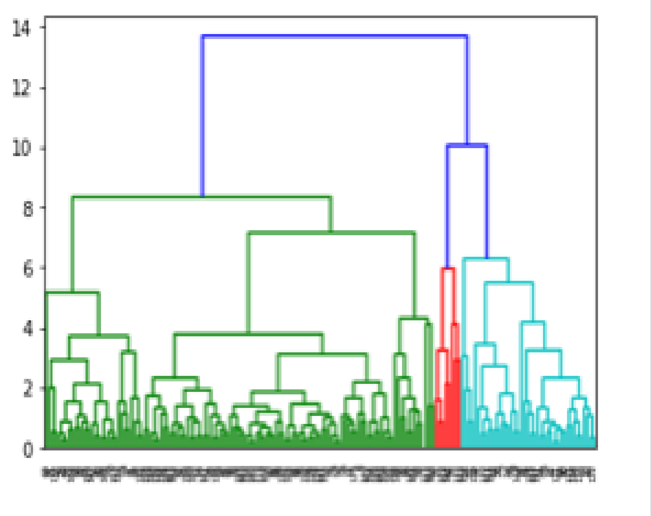
Dendogram-Type of Linkages

POINTS CONCLUDED:

- From below dendogram(single linkage/mimimum linkage) we are not able to see the levels we will go for a complete linkage.
- From below dendogram(Complete/Maximum Linkage), we can pick number of clusters as 3 for further analysis



Single/Minimum Linkage

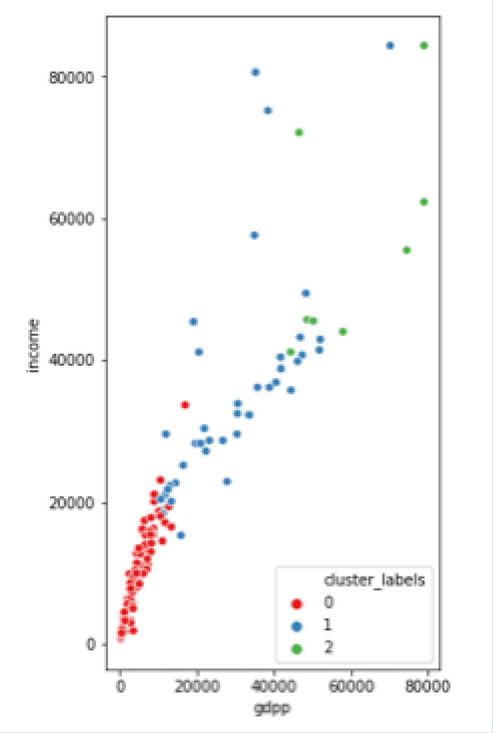


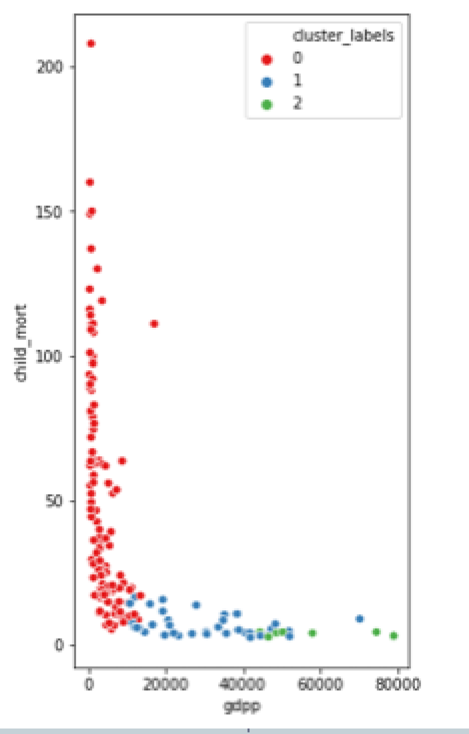
Complete/Maximum Linkage

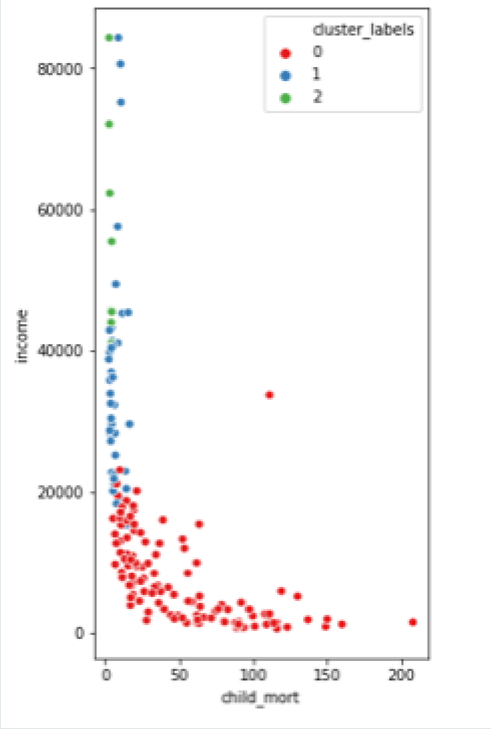
Plotting the Clusters-Hierarchical

POINTS CONCLUDED:

- From Graph 1, we conclude that GDPP & Income are directly proportional to each other, that means if for a country income increases GDPP also increases & vice-versa.
- From Graph 2, we conclude that GDPP & Child mortality are inversely proportional
- From Graph 3, we conclude that Income & Child mortality are inversely proportional. that means if for a country income decreases child mortality increases & vice-versa







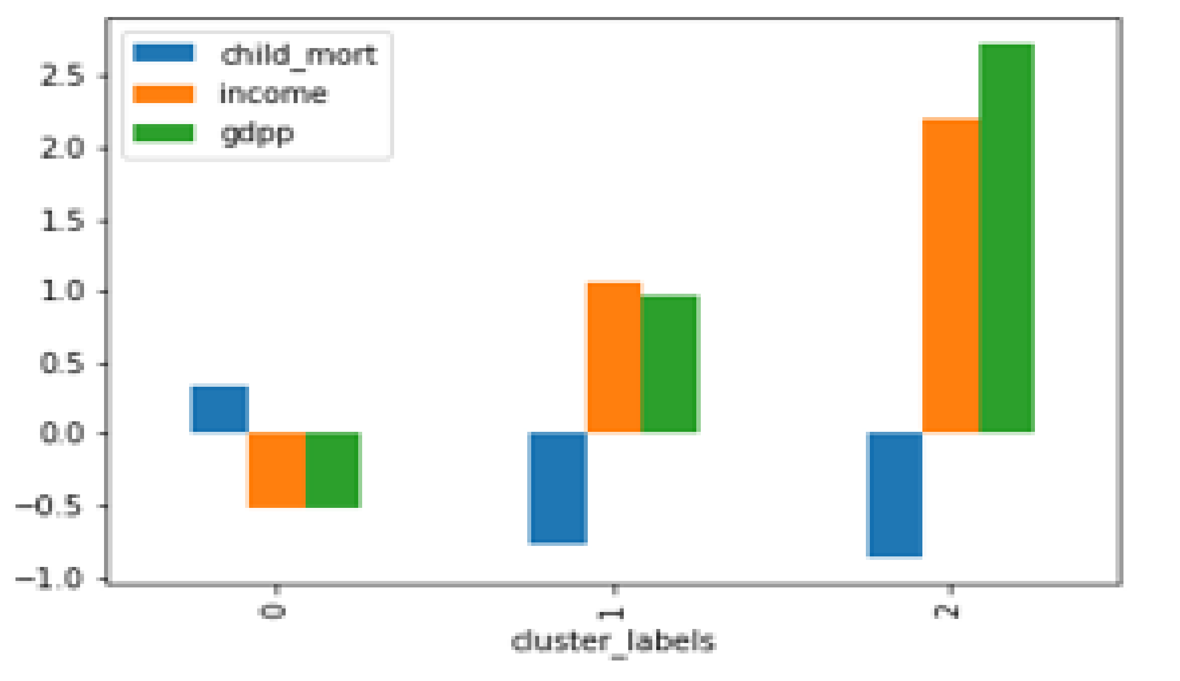
Cluster Profiling-Hierarchical

POINTS CONCLUDED:

So from above graph we can conclude that :

- Cluster 0 have countries which have positive child_mort, income and gdpp negative
- Cluster 1 have countries with negative child_mort, and positive income and gdpp
- Cluster 2 have countries with negative child_mort, and large positive income and gdpp

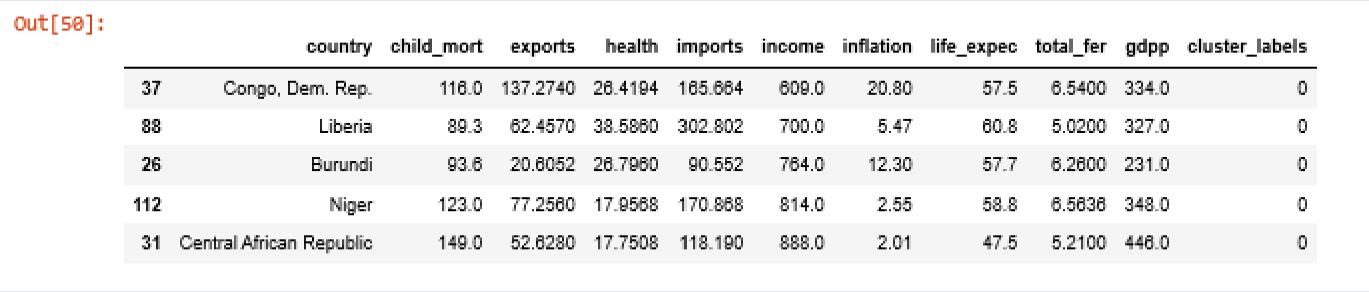
So, top 5 countries that need urgent AID are countries falling in cluster label 0



Final Conclusion-Hierarchical

So, we can conclude that countries list below are the countries that needs urgent AID :

1. Congo, Dem. Rep.
2. Liberia
3. Burundi
4. Niger
5. Central African Republic



Out[50]:

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	cluster_labels
37	Congo, Dem. Rep.	116.0	137.2740	26.4194	165.664	609.0	20.80	57.5	6.5400	334.0	0
88	Liberia	89.3	62.4570	38.5860	302.802	700.0	5.47	60.8	5.0200	327.0	0
26	Burundi	93.6	20.6052	26.7960	90.552	764.0	12.30	57.7	6.2600	231.0	0
112	Niger	123.0	77.2560	17.9568	170.868	814.0	2.55	58.8	6.5636	348.0	0
31	Central African Republic	149.0	52.6280	17.7508	118.190	888.0	2.01	47.5	5.2100	446.0	0

Final Conclusion

From above analysis, results from both the algorithms were same but from above analysis following points can be concluded:

- Hierarchical clustering can’t handle big data well but K Means clustering can. This is because the time complexity of K Means is linear i.e. O(n) while that of hierarchical clustering is quadratic i.e. O(n2).
- In K Means clustering, since we start with random choice of clusters, the results produced by running the algorithm multiple times might differ. While results are reproducible in Hierarchical clustering.
- K Means is found to work well when the shape of the clusters is hyper spherical (like circle in 2D, sphere in 3D).
- K Means clustering requires prior knowledge of K i.e. no. of clusters you want to divide your data into. But, you can stop at whatever number of clusters you find appropriate in hierarchical clustering by interpreting the dendrogram.