Question 1: Assignment Summary

Briefly describe the "Clustering of Countries" assignment that you just completed within 200-300 words. Mention the problem statement and the solution methodology that you followed to arrive at the final list of countries. Explain your main choices briefly(what EDA you performed, which type of Clustering produced a better result and so on)

Note: You don't have to include any images, equations or graphs for this question. Just text should be enough.

→ This case study aims is to categorize the countries using some socio-economic and health factors that determine overall development of the country. So here we have performed Clustering technique to select the countries which are in direst need of aid by considering socio–economic factor in to consideration. This analysis make it easy and help an international humanitarian NGO to provide the top 5 backward countries to provide the basic amenities and relief during the time of disasters and natural calamities. HELP International is an international humanitarian NGO that runs a lot of operational projects from time to time along with advocacy drives to raise awareness as well as for funding purposes.

Below is the Analysis methodology:

Data collection and cleaning

- Import the data
- Identifying the data quality issues and clean the data

Outlier analysis and removal

- Removing the outlier where ever required as per understanding the problem statement.

Visualizing the data

- Visualizing few original data variables to look for any pattern or correlation.

Hopkins Statistics

- To check if data has tendency to form clusters

Scaling the data

- Standardizing all the continuous variables

K means clustering

- Identify the 'k' by using silhouette analysis and elbow curve graph
- Visualizing the clusters with various variables
- Analyzing the clusters
- Identifying the countries which require aid.

As per our K means clustering-Cluster - 1 are of concern due to:

•Low gdpp
•Low income
•High child mortality

Hierarchical Clustering

- Identify the 'n' via dendrogram.
- Forming n –clusters on PCA modified data
- Visualizing the clusters with various variables
- Analyzing the clusters
- Identifying the countries which require aid.
- As per our Hierarchical clusters-Cluster - O are of concern due to:
  •Low gdpp
  •Low income
  •High child mortality

Decision Making

- Identifying the countries which require aid by analyzing both K-means and Hierarchical Clustering results.

As by both k-means and Hierarchical clustering method-we have got same countries which require aid. The following are the countries which are in direst need of aid by considering socio–economic factor in to consideration:

The order of precedence given to the features is ggdp, child_mort and then the income and the top 5 countries are:

- ❑ Liberia
- ❑ Burundi
- ❑ Congo, Dem. Rep.
- ❑ Niger
- ❑ Sierra Leon

Question 2: Clustering

a) Compare and contrast K-means Clustering and Hierarchical Clustering.

→ In k-means clustering, we try to identify the best way to divide the data into k sets simultaneously. A good approach is to take k items from the data set as initial cluster representatives, assign all items to the cluster whose representative is closest, and then calculate the cluster mean as the new representative, until it converges (all clusters stay the same).

Hierarchical Clustering: The most important difference is the hierarchy. Actually, there are two different approaches that fall under this name:

top-down and bottom-up.

In top-down hierarchical clustering, we divide the data into 2 clusters (using k-means with k=2k=2, for example). Then, for each cluster, we can repeat this process, until all the clusters are too small or too similar for further clustering to make sense, or until we reach a preset number of clusters.
In bottom-up hierarchical clustering, we start with each data item having its own cluster. We then look for the two items that are most similar, and combine them in a larger cluster. We keep repeating until all the clusters we have left are too dissimilar to be gathered together, or until we reach a preset number of clusters.

K- means is a collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters. It is a division of objects into clusters such that each object is in exactly one cluster, not several.

In Hierarchical clustering, clusters have a tree like structure or a parent child relationship. Here, the two most similar clusters are combined together and continue to combine until all objects are in the same cluster.

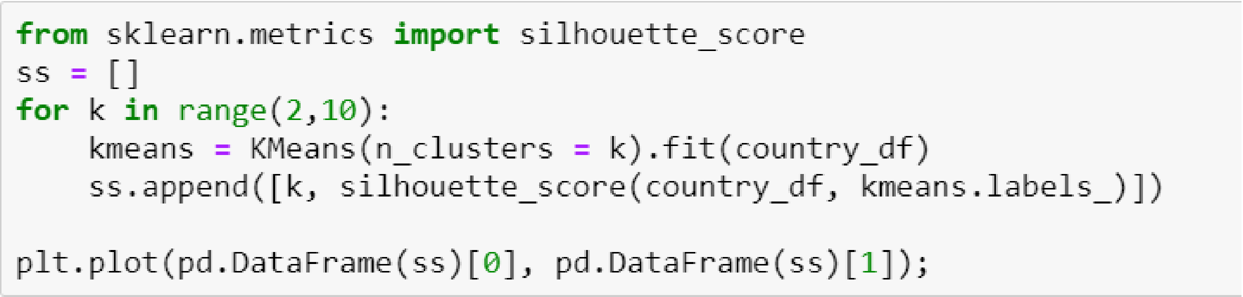
b) Briefly explain the steps of the K-means clustering algorithm.

The first step of this algorithm is creating, among our unlabeled observations, *c* new observations, randomly located, called ‘centroids’. The number of centroids will be representative of the number of output classes (which, remember, we do not know). Now, an iterative process will start, made of two steps:

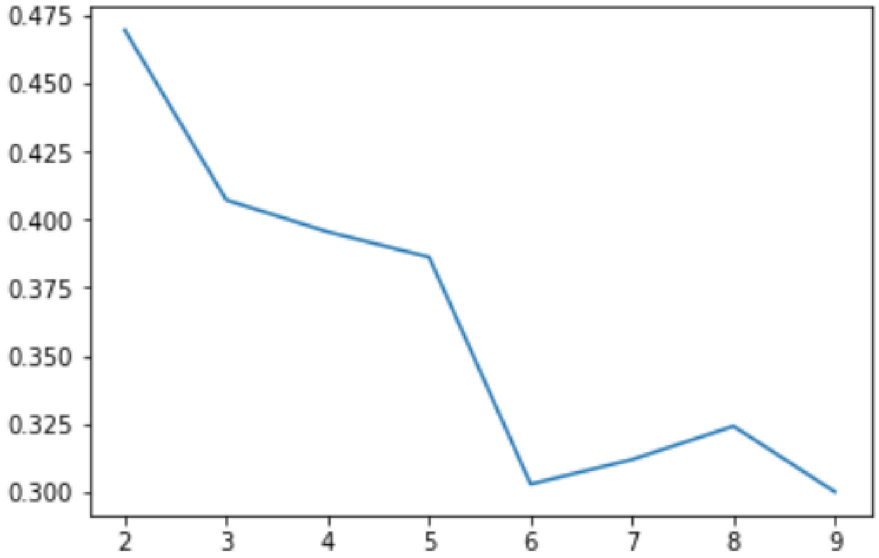
- First, for each centroid, the algorithm finds the nearest points (in terms of distance that is usually computed as Euclidean distance) to that centroid, and assigns them to its category;
- Second, for each category (represented by one centroid), the algorithm computes the average of all the points which has been attributed to that class. The output of this computation will be the new centroid for that class.

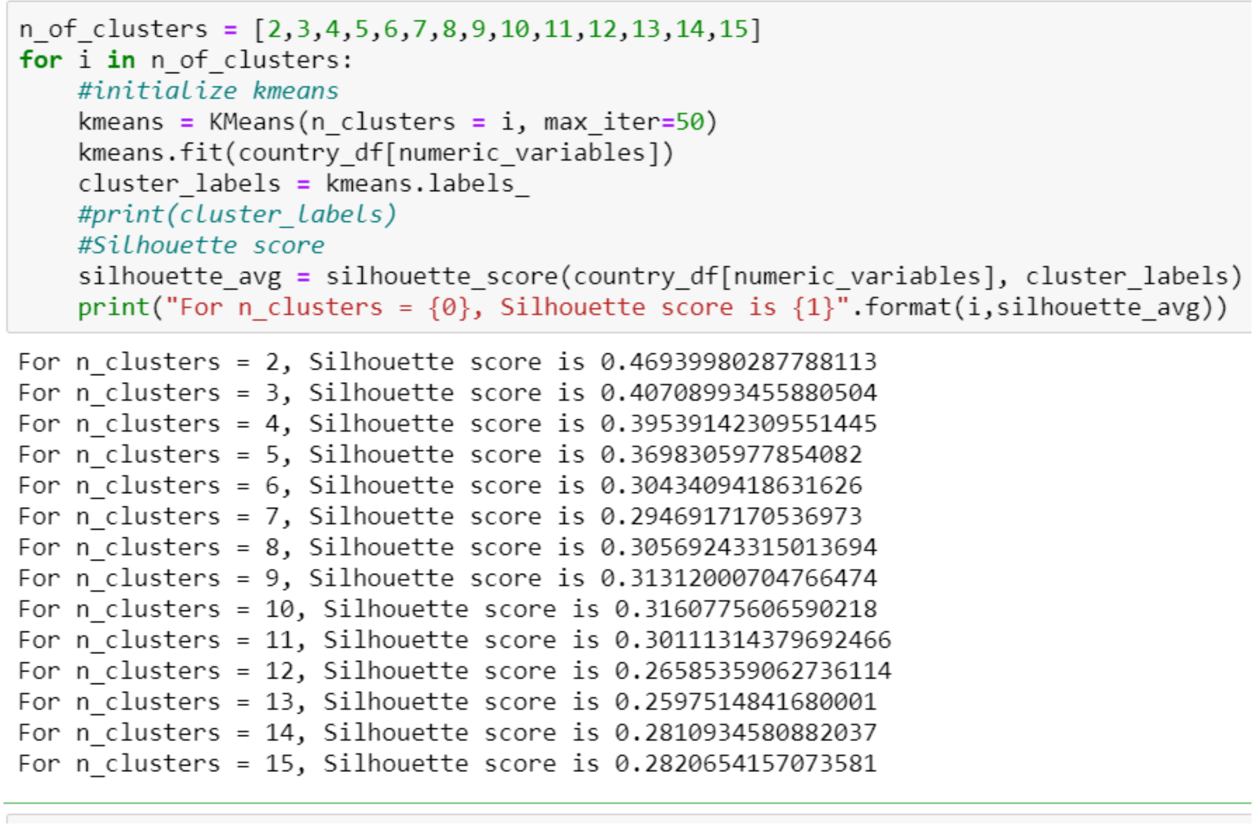
b) How is the value of ‘k’ chosen in K-means clustering? Explain both the statistical as well as the business aspect of it.

→ In K-means clustering the value of ‘k’ is chosen by using silhouette score analysis and elbow curve. We can see the code in below screen sort.



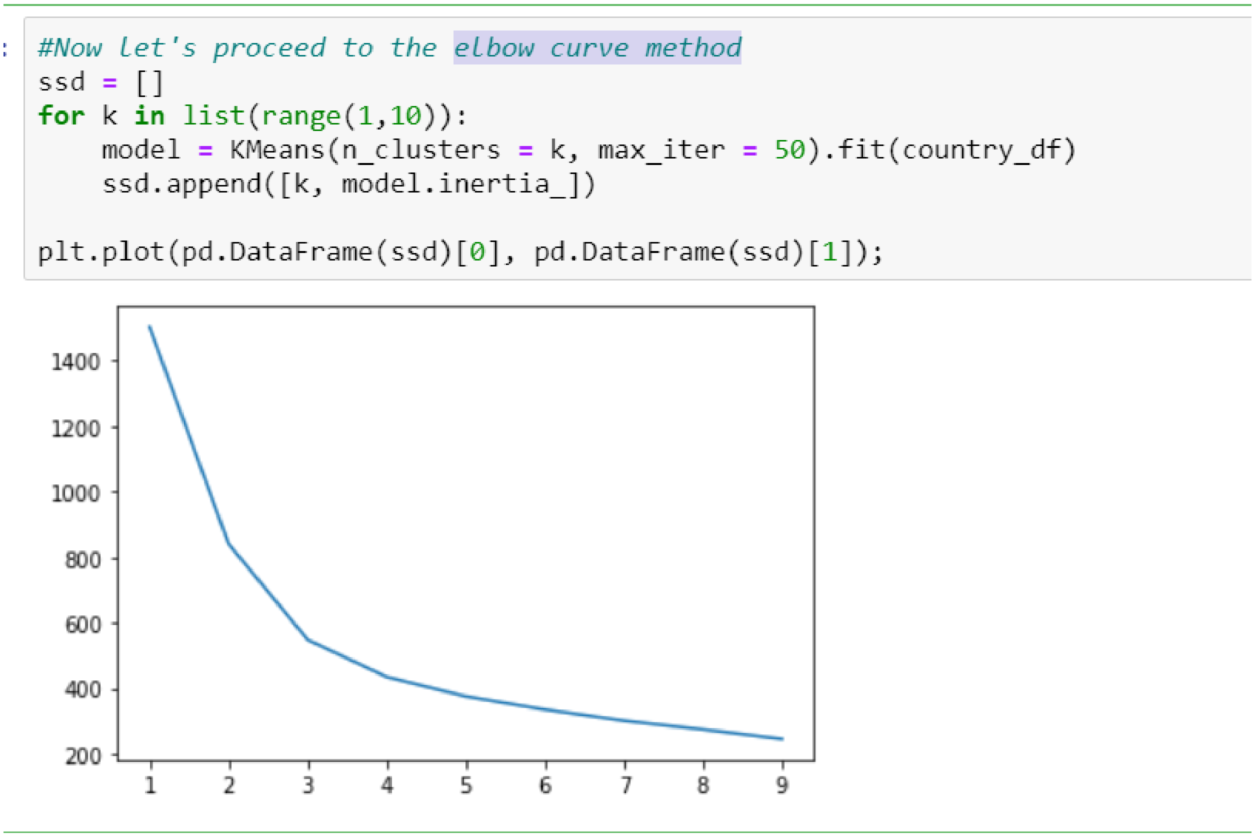
```
from sklearn.metrics import silhouette_score
ss = []
for k in range(2,10):
    kmeans = KMeans(n_clusters = k).fit(country_df)
    ss.append([k, silhouette_score(country_df, kmeans.labels_)])

plt.plot(pd.DataFrame(ss)[0], pd.DataFrame(ss)[1]);
```





```python
n_of_clusters = [2,3,4,5,6,7,8,9,10,11,12,13,14,15]
for i in n_of_clusters:
    #initialize kmeans
    kmeans = KMeans(n_clusters = i, max_iter=50)
    kmeans.fit(country_df[numeric_variables])
    cluster_labels = kmeans.labels_
    #print(cluster_labels)
    #Silhouette score
    silhouette_avg = silhouette_score(country_df[numeric_variables], cluster_labels)
    print("For n_clusters = {0}, Silhouette score is {1}".format(i,silhouette_avg))
```

```
For n_clusters = 2, Silhouette score is 0.46939980287788113
For n_clusters = 3, Silhouette score is 0.40708993455880504
For n_clusters = 4, Silhouette score is 0.39539142309551445
For n_clusters = 5, Silhouette score is 0.3698305977854082
For n_clusters = 6, Silhouette score is 0.3043409418631626
For n_clusters = 7, Silhouette score is 0.2946917170536973
For n_clusters = 8, Silhouette score is 0.30569243315013694
For n_clusters = 9, Silhouette score is 0.31312000704766474
For n_clusters = 10, Silhouette score is 0.3160775606590218
For n_clusters = 11, Silhouette score is 0.30111314379692466
For n_clusters = 12, Silhouette score is 0.26585359062736114
For n_clusters = 13, Silhouette score is 0.2597514841680001
For n_clusters = 14, Silhouette score is 0.2810934580882037
For n_clusters = 15, Silhouette score is 0.2820654157073581
```



```
#Now let's proceed to the elbow curve method
ssd = []
for k in list(range(1,10)):
    model = KMeans(n_clusters = k, max_iter = 50).fit(country_df)
    ssd.append([k, model.inertia_])

plt.plot(pd.DataFrame(ssd)[0], pd.DataFrame(ssd)[1]);
```

From both the methods above the optimal number of clusters that can be formed is 3.

The most common ways in which businesses segment their customer base are:

1. **Demographic information**, such as gender, age, familial and marital status, income, education, and occupation.

2. **Geographical information**, which differs depending on the scope of the company. For localized businesses, this info might pertain to specific towns or counties. For larger companies, it might mean a customer’s city, state, or even country of residence.

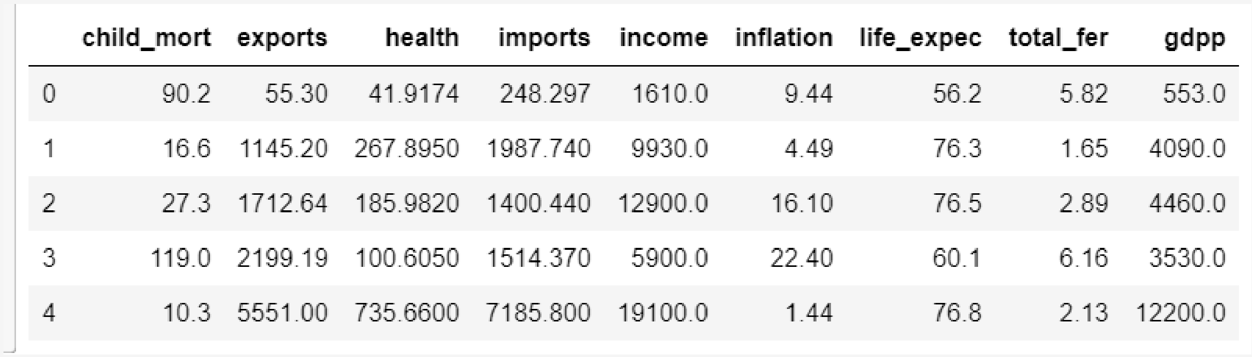
3. **Psychographics**, such as social class, lifestyle, and personality traits.

4. **Behavioral data**, such as spending and consumption habits, product/service usage, and desired benefits.

d) Explain the necessity for scaling/standardisation before performing Clustering.

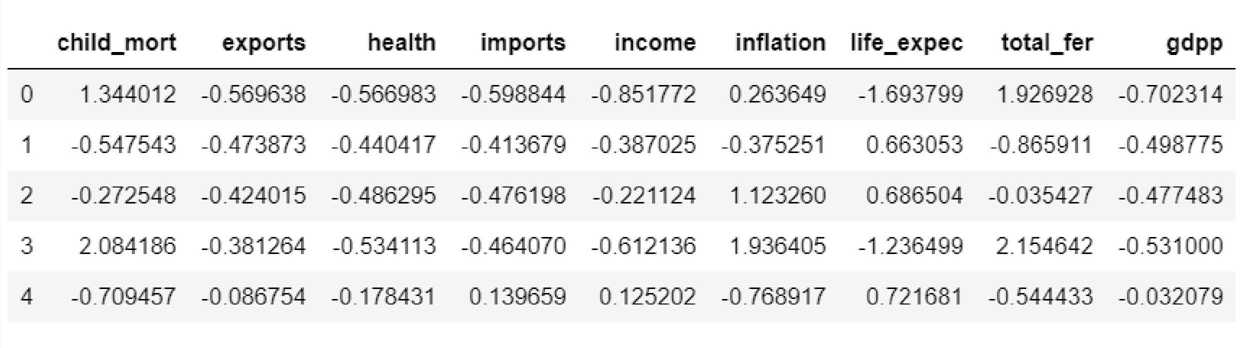
→ scaling/standardisation is the process of rescaling the values of the variables in data set so they share a common scale. Often performed as a pre-processing step, particularly for cluster analysis, standardization is important if we are working with data where each variable has a different unit , or where the scales of each of our variables are very different from one another (e.g., 0-1 vs 0-1000). The reason this importance is particularly high in cluster analysis is because groups are defined based on the distance between points in mathematical space. That's why performing scaling/standardisation before performing Clustering is better.

Before scaling the data



	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
0	90.2	55.30	41.9174	248.297	1610.0	9.44	56.2	5.82	553.0
1	16.6	1145.20	267.8950	1987.740	9930.0	4.49	76.3	1.65	4090.0
2	27.3	1712.64	185.9820	1400.440	12900.0	16.10	76.5	2.89	4460.0
3	119.0	2199.19	100.6050	1514.370	5900.0	22.40	60.1	6.16	3530.0
4	10.3	5551.00	735.6600	7185.800	19100.0	1.44	76.8	2.13	12200.0

After scaling the data



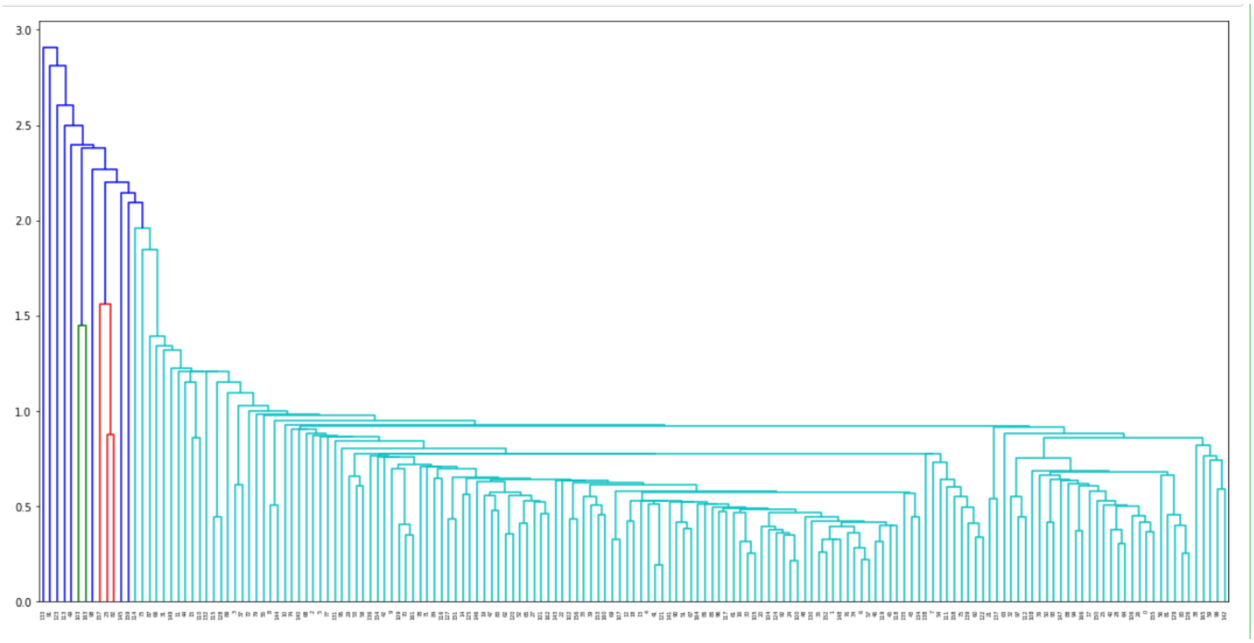
	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
0	1.344012	-0.569638	-0.566983	-0.598844	-0.851772	0.263649	-1.693799	1.926928	-0.702314
1	-0.547543	-0.473873	-0.440417	-0.413679	-0.387025	-0.375251	0.663053	-0.865911	-0.498775
2	-0.272548	-0.424015	-0.486295	-0.476198	-0.221124	1.123260	0.686504	-0.035427	-0.477483
3	2.084186	-0.381264	-0.534113	-0.464070	-0.612136	1.936405	-1.236499	2.154642	-0.531000
4	-0.709457	-0.086754	-0.178431	0.139659	0.125202	-0.768917	0.721681	-0.544433	-0.032079

e**) Explain the different linkages used in Hierarchical Clustering.**

→We have used below two linkage in the assignment:

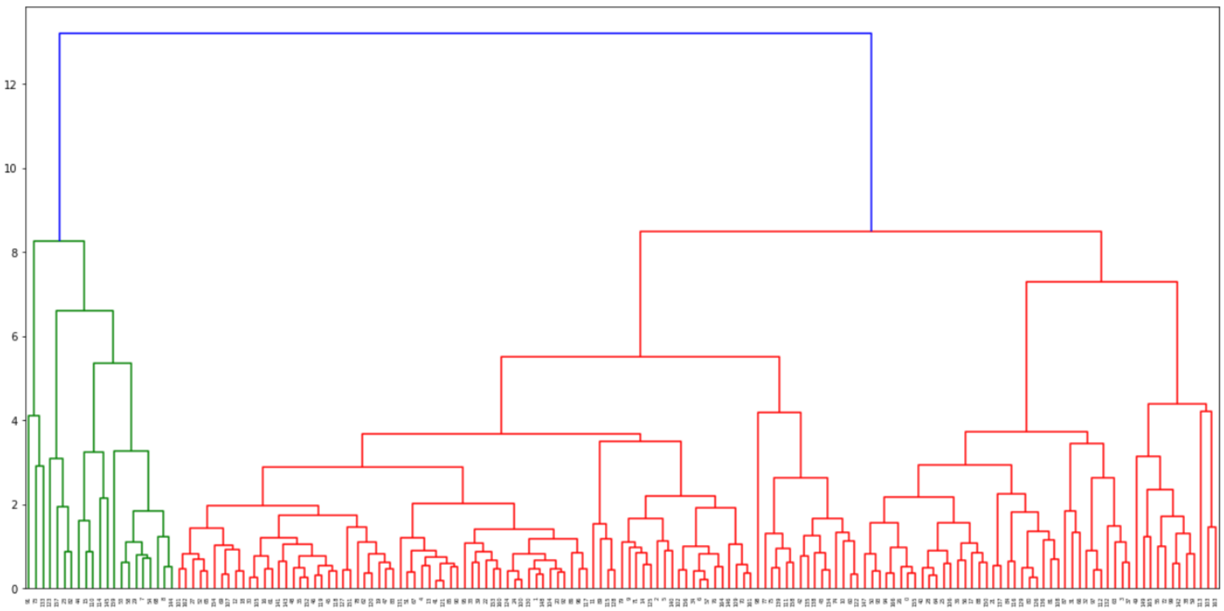
Single-Linkage

Single-linkage (nearest neighbor) is the shortest distance between a pair of observations in two clusters. It can sometimes produce clusters where observations in different clusters are closer together than to observations within their own clusters. After performing the analysis we can see that these clusters can appear spread-out.



Complete-Linkage

Complete-linkage (farthest neighbor) is where distance is measured between the farthest pair of observations in all the clusters. This method usually produces tighter clusters than single-linkage, but these tight clusters can end up very close together. It is one of the more popular distance metrics. As we can see below clusters can appear to be more readable.


12
10
8
6
4
2
0