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)

Answer:

- Problem Statement: Our client-HELP International is an international humanitarian NGO and have raised around $ 10 million to help fight poverty and providing people of backward countries with basic amenities and relief during the time of disasters and natural calamities. The CEO of the NGO needs to decide how to use this money strategically and effectively.
- Solution: My analysis categorizes the countries using some socio-economic and health factors that determine the overall development of the country and suggest the countries which are in dire need of aid to the CEO to focus on the most.
- Methodology: I tried to solve the problem by clustering the countries based on various data-points available.
- Performing the data inspection and data cleaning to get rid of any missing or invalid values if any and filtering out the outliers- In this case the outliers weren't treated because removing them would mean interfering with the clusters that must form naturally in such a limited data.
- Performing the EDA to understand the distribution of various fields in our data set and to understand the relationship amongst the variables that will help understand the data more as well as point outliers if still it were unnoticed and will help make an informed decision while deriving insights from the model.
- Later Hypothesis testing using Hopkins statistics to check validity of our data to form clusters and then rescaled data for standardization and optimizing the models.
- After data preparation, clustered the data using KMeans: k=3 or 4, Hierarchical(single, closed(k=3 and 4)). How did I reach at k=3 and 4? Plotted ssd against k(elbow method) and Silhouette Score and coefficients.
- Based on the clusters formed by the above methods described, profiled the clusters- analyzing each one of them and the variables to conclude countries which are in dire need of financial-aid.

Question 2: Clustering

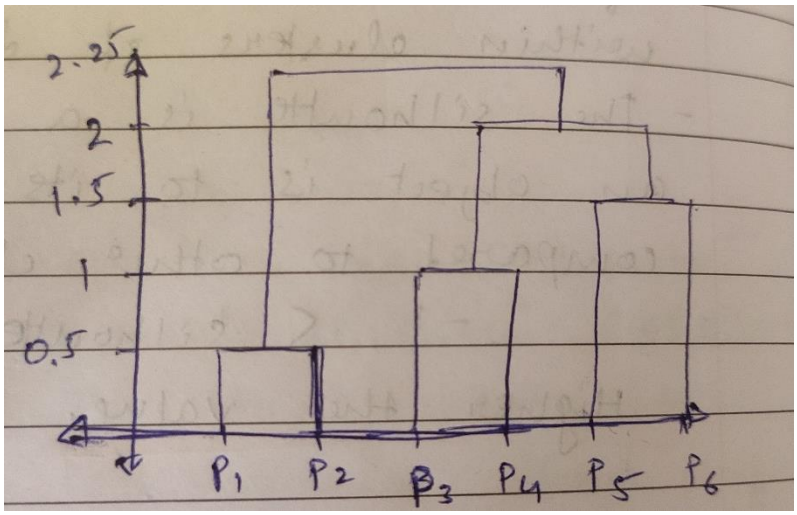
a) Compare and contrast K-means Clustering and Hierarchical Clustering.

Answer:

KMeans: In KMeans algorithm the relationship between the datapoints is not known. There are a lot of uncertainties in KMeans clustering like the number of clusters(k) in which data-points must be divided, the initial center points to be chosen before optimization starts and KMeans clustering

algorithm depends a lot on these parameters. Hence this method of clustering needs to happen in iterations to give the optimum result.

Hierarchical: In Hierarchical algorithm, clustering takes place because of relationship amongst the datapoints (Euclidean distance). Mostly occurs in 2 ways: Agglomerative and divisive clustering taking bottom to top and top to bottom approach respectively. In Agglomerative hierarchical clustering, each datapoints are considered as individual clusters and the clusters with shortest Euclidean distances tend to combine into one cluster. The process repeats itself unless all the data-points combines under a single cluster. This hierarchical linkage structure is called a dendrogram.



Differences:

- KMeans algorithm is mostly time efficient for bigger datasets compared to Hierarchical algorithm because Hierarchical algorithm involves datapoint to datapoint linkage from bottom to top.
- K needs to be assumed and given as an input in KMeans and not in Hierarchical.
- KMeans clusters are not well defined with outliers. Hierarchical clusters perform well with outliers.
- In KMeans clustering, assumption is made about centroids. Hence different centroids leads to different clusters unlike in Hierarchical clustering.

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

Answer:

Considerations in KMeans clustering algorithm:

- The number of clusters(k) determined by Elbow method and Silhouette Analysis.
- The choice of initial cluster centres: random, kmeans++ etc.
- The dataset must pass the hypothesis test of Hopkins statistics for its validity of getting clustered.

Once the conditions are satisfied, we may proceed with KMeans clustering.

Intuition:

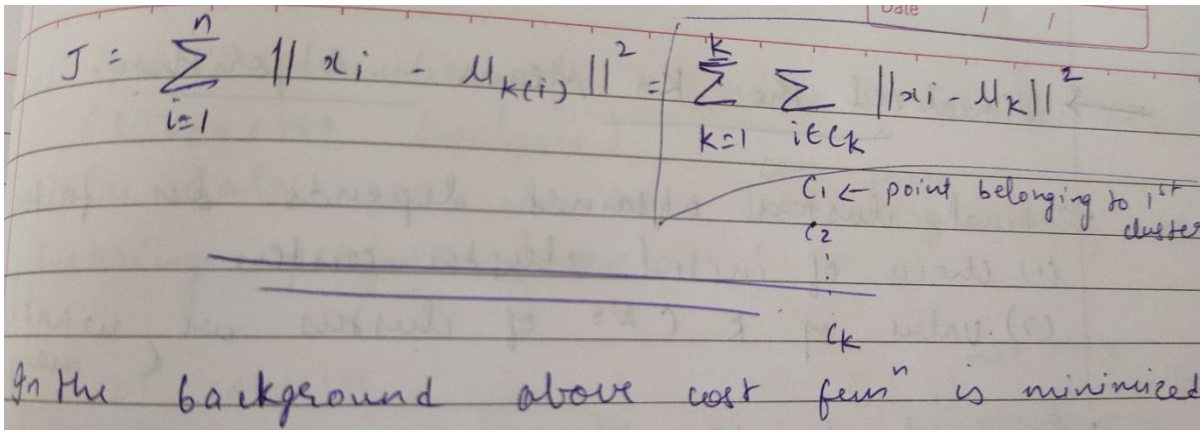
Step1: Assignment: In a space of data-points, choose one of the data-points as the center randomly. Based on this center another center will be chosen either randomly or as k-means++ as mentioned.

Now for all the datapoints **X**i , we will calculate Euclidean distance/Manhattan distance between the nearest center and **X**i and thus form the first cluster.

Step2: Optimization: We calculate a new center by finding the centroid of the cluster.

The assignment and optimization problem goes on and on until we end up unchanged centroids for 2-3 consecutive iterations.

The cost function to be optimized is: For every data point, the Euclidean distance between the data-point and the cluster center must be calculated and get squared. This is the cost function for the entire space that must be minimized.



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

Answer: Statistically K is derived using 2 methods:

1. Elbow Method
2. Silhouette Analysis

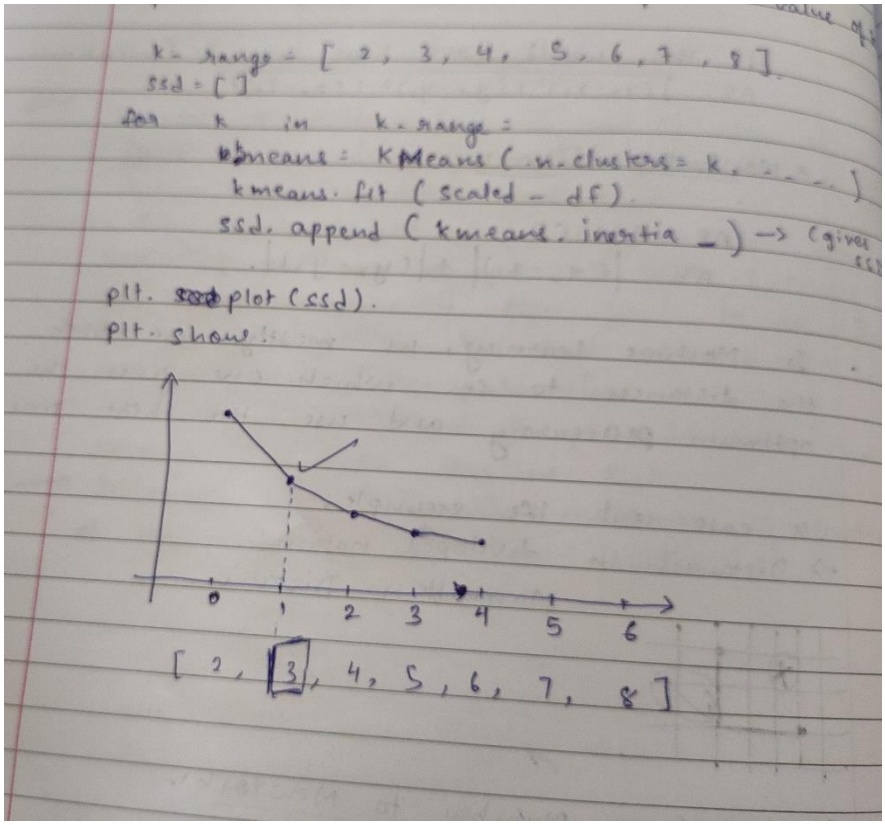
Elbow curve: Sum squared of distance(SSD) is the sum of the squared distances from the centroid and each member of the cluster.

For different values of k, we need to find the SSD and plot a graph against the number of clusters. The point where the abrupt decrease in SSD is less than the previous point, that is the optimum value of k.

For example:

K_range=[2,3,4,5,6,7,8]

Plot of SSD vs k_range:



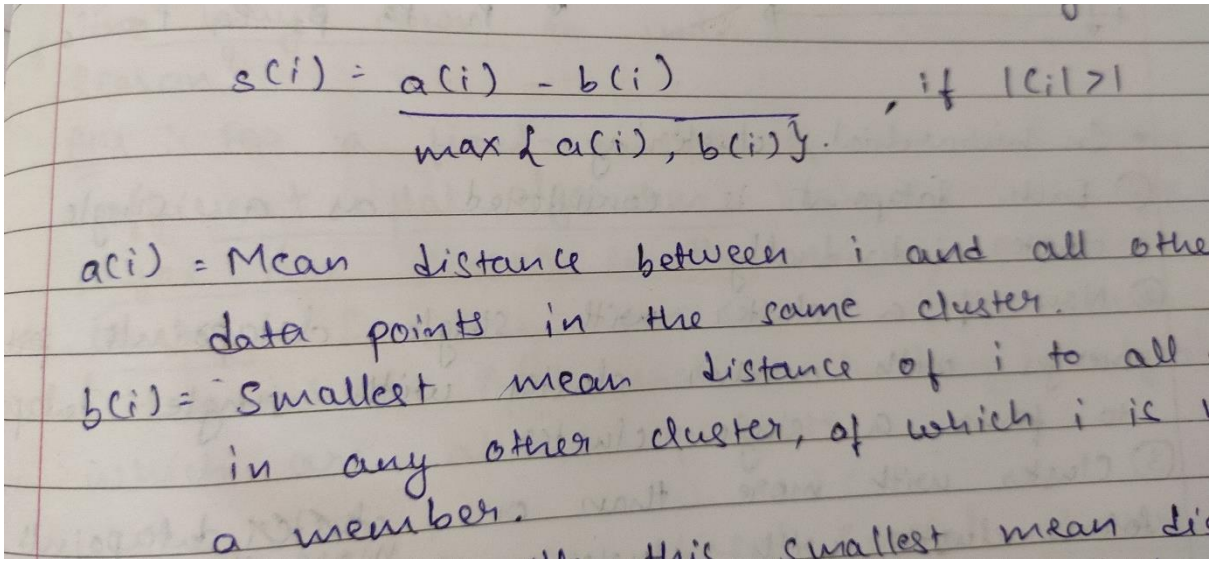
In this case, 3 can be considered as the optimum value of k.

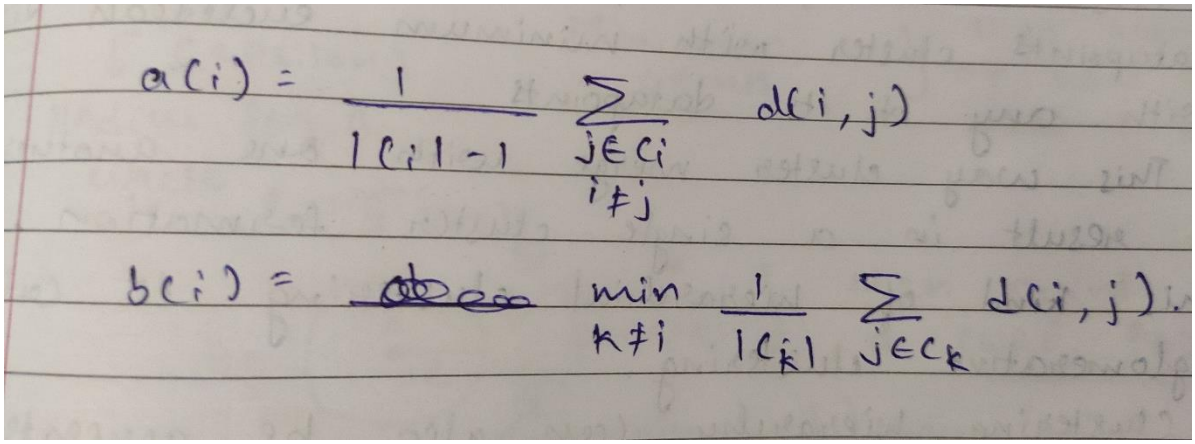
Silhouette Analysis: Silhouette refers to a method of interpretation and validation of consistency within clusters of data.

The Silhouette is a measure of how similar an object is to its own cluster(cohesion) and how dissimilar it is compared to other cluster(separation).

-1 < Silhouette < 1

Higher the value, better the clustering.





We will get different Silhouette values for different k values. Graphs can be plotted to see if for a cluster a(i)<<b(i) isn't there.

As far as **Business** aspects are concerned, with large datasets- dividing the entire datasets into small number of clusters such as 2, 3 most of the times might not make much sense because special attention won't be given to a lot of data-points in the group. Also number of clusters more than 15 are also not advisable in the business scenario because the data-points cannot be treated differently in so many different ways. Hence based on the statistics and business requirements a decision is arrived upon such that where statistics may guide us where to lead our decision to but it surely won't dictate the decision. It is the business and it's requirements that will take the final call based on the necessity.

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

Answer: Clustering technique uses calculation of distances using either the Euclidean or Manhattan methods. Now as the datapoints are spread across the spaces, Standardization is much required because if there are many variables with different values and different units, the distances calculated while clustering would be incorrect. Hence scaling is a technique to Standardize all the variables.

Also Standardizing all the variables, helps in optimizing the algorithm since minimizing the cost function happens much faster after such Standardizing techniques.

There are 2 scaling techniques:

1. Normalization/ MinMax Scaling
2. Standardization.

Difference Between Normalised scaling and standardized scaling:

- Formulas are different.
- Normalised scaling: Min value is 0 and maximum value is 1.
- Standardized scaling: Range is spread. Mean is 0 and standard deviation is 1.
- Normalised: Outliers are treated. In case of extreme outliers cannot be used.
- Standardized: Outliers are spread. In case of extreme outliers can be used.

e) Explain the different linkages used in Hierarchical Clustering.

Answer: In Agglomerative hierarchical clustering, each datapoints are considered as individual clusters and the clusters with shortest Euclidean distances tend to combine into one cluster. The process repeats itself unless all the data-points combines under a single cluster. This hierarchical linkage structure is called a dendrogram.

Single Linkage: Now in Single Linkage, the Euclidean distances calculated as mentioned above- we combine in each step the two clusters whose two closest members have the smallest distance.

Complete Linkage: In complete linkage, the distance between 2 clusters is defined as the maximum distance between any 2 points in the clusters.