Clustering and PCA Assignment

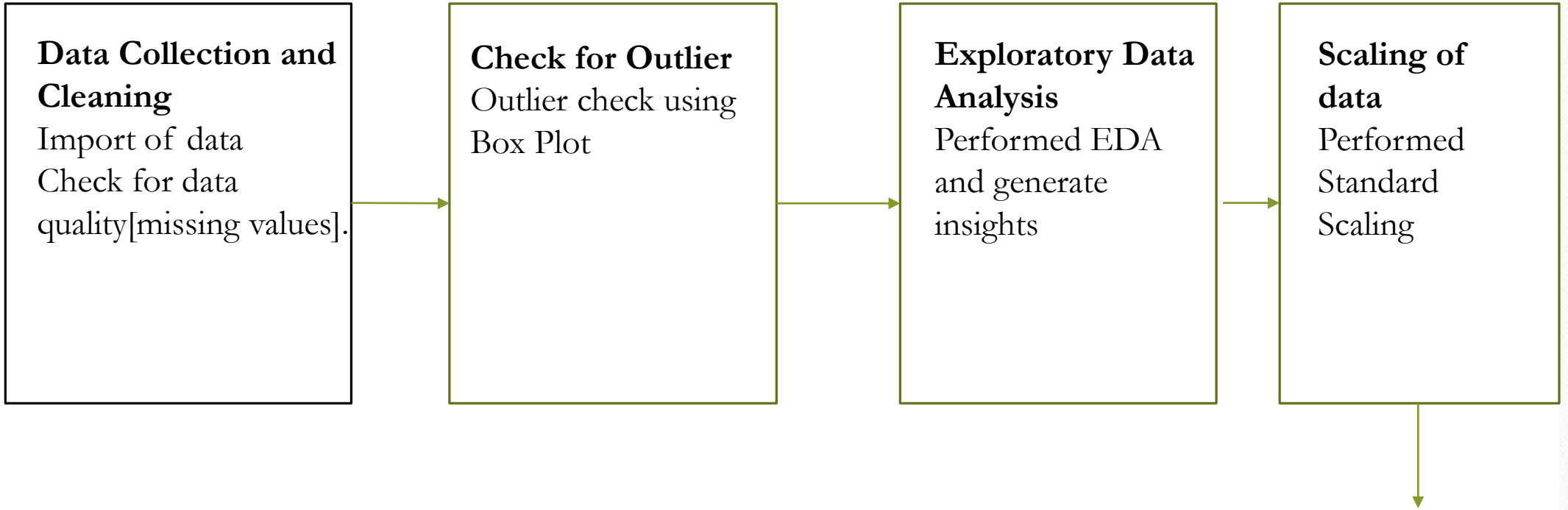
Clustering of Countries

Answer 1

- HELP International is an international humanitarian 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. It runs a lot of operational projects from time to time along with advocacy drives to raise awareness as well as for funding purposes. After the recent funding programmes, they have been able to raise around $ 10 million. The CEO of the NGO needs to decide how to use this money strategically and effectively. As a data analyst we need to analyze the countries that are in the direst need of aid.

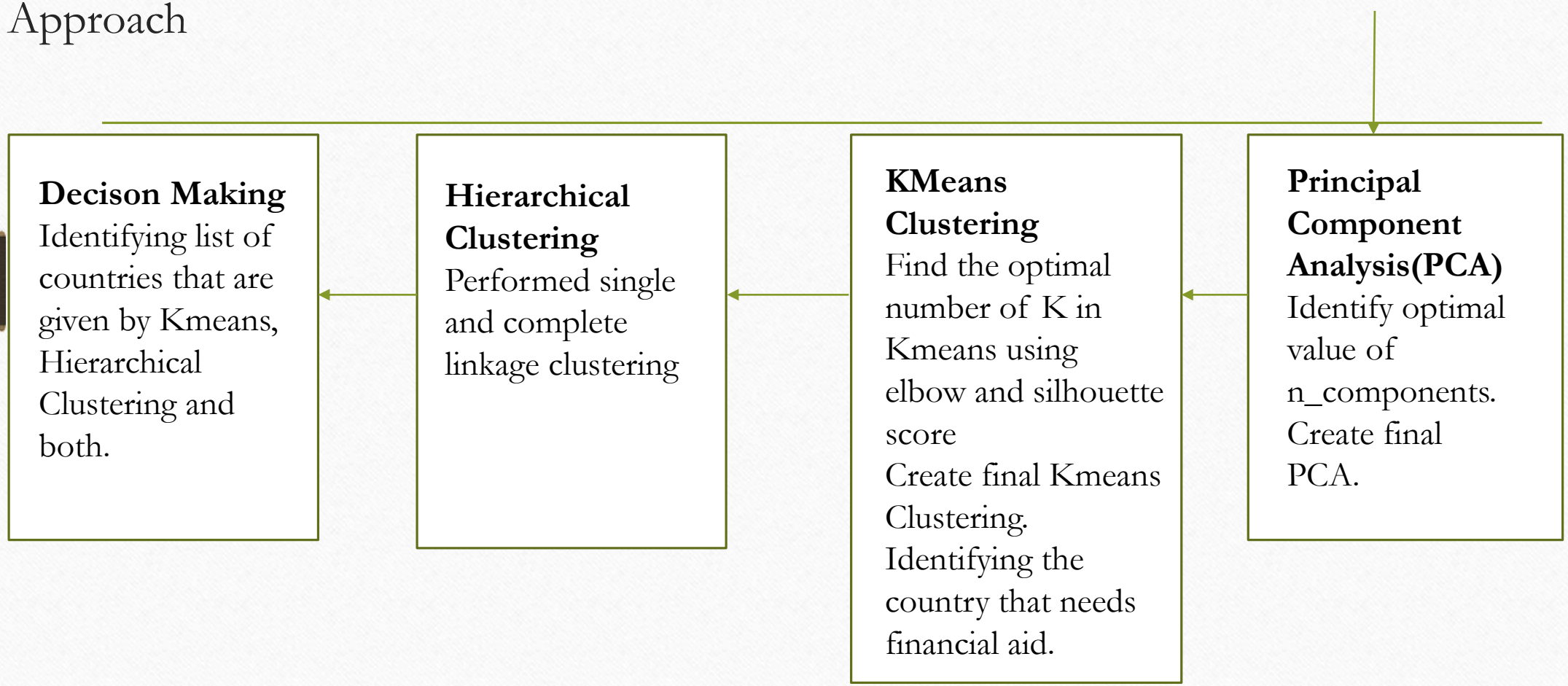
Answer 1

Approach



Answer 1

Approach



Answer 1

- As we can see from the heat map that
    - Income and gdpp is having strong positive correlation.
    - Child_mort and tot_fer is having strong positive correlation.
    - Child_more and life_expec is having strong negative correlation.
    - Life_expec and and tot_fer is having strong negative correlation.
- For outlier detection we have ploted the box plot. From the box plot we can see that there are many rows where we can see the outliers. In accordance to that, we have not removed the outlier because if we will remove then we will loose the data for that particular country. We have not imputed the data because the data was correct and cannot feed these trends in our cluster.
- We have taken 7 Principal Components because we are getting maximum variance(information gain) at 7. In the notebook, we have plotted a graph of variance at each number of Principal Component.
- For Clustering, we have taken value of k as 4 according to elbow method and silhouette analysis.

Answer 2

a: Difference between Hierarchical Clustering and K-means Clsutering are:

- Hierarchical Clustering is easy to implement.
- For larger dataset Hierarchical Clustering works faster as compare to K-means
- K-means clustering requires prior knowledge of K i.e number of cluster you want to divide.
- In K-means clustering, result are not reproducible since we select random data points for initialization whereas this problem not occurs in Hierarchical Clustering

b: Steps in Kmeans algorithm are as follows:

1. Choose any random K data point as centroids.

Answer 2

1. Create cluster for those point that are close to centroid(Find Ecludian distance between data point and centroid).
2. As we have new cluster, update the centroid.
3. Repeat step 2 and 4 until K centers have been chosen

Answer 2

c: For Clustering, we have taken value of k as 4 according to elbow method and silhouette analysis. In elbow method we have to plot intertia at each value and check at which value of k elbow is formed. In silhouette analysis we calculate the silhouette score and check at which value of K it is maximum.

d: Scaling is necessary step before clustering because clustering is done on the basis of distance between the feature. If the feature are on the differenct scale then cluster might not be correct. Standardization helps to make the relative weight of each variable equal by converting each variable to a unitless measure or relative distance.

e: The different kinds of linkage in Hierarchical Clustering are:

- Single Linkage
- Complete Linkage
- Average Linkage

Answer 3

a: Application of Principal Component Analysis(PCA) are:

- PCA is used for Dimensionality reduction(reduces number of features).
- PCA reduces multicollinearity.
- It is used for finding patterns in data of high dimension.

b: Basis Transformation: Transformation of the original dataset so that eigenvectors are the basis vectors and find new coordinated of the data points with respect to this new basis.

Variance Information: Variance is information gain from the dataset. It is the measure of variation. Higher the value, better it is. It can be checked by explained_variance_ratio.

Answer 3

c: Three shortcoming of Principal Component Analysis(PCA) are:

- According to PCA lower variance column are of no use but maybe in business perspective it might be useful.
- Components of PCA needs to be perpendicular.
- PCA is limited to linearity.
- Before applying PCA we need to scale our dataset.

Thank You