Question 1

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(why you took that many numbers of principal components, which type of Clustering produced a better result and so on)

Answer 1

Help International NGO needs assistance in determining few underdeveloped countries from a list of 167 overall countries so that they can allocate appropriate funds to these countries who are in dire need of aid.
In order to arrive at the final list of countries, we need to follow following steps:

- Perform EDA on the dataset
- Standardize/scale the data as there are quite a few columns that had very high values compared to few that are in single unit. This will ensure that entire dataset is on same scale.
- Perform PCA on the dataset to come up with principal components that can explain about 95% of the variance. We choose final number of PC as 5.
- Run Hopkins measure on PCA dataframe to find out if it's even feasible to get clusters out of the data. This score was ~ 86%. So we went ahead with both Kmeans and Heirarchical clusterings.
- For Kmeans we also did Scree plot and Silhouette analysis to come up with optimal number of cluster count which is a prerequisite for Kmeans algorithm unlike Hierarchical algorithm.
- We choose 5 clusters for kmeans algorithm.
- Upon plotting the dendrogram for hierarchical algorithm, we found that best choice of cluster formation would be 6 with 6th cluster having just one country that had all extreme values for all of its columns.
- Finally, we grouped the dataframe based on clusterID and took a mean of all the numerical variables to look at the pattern of how the clusters were formed for both the algorithms.
- Both algorithms formed clusters based on socio-economic conditions of the countires. That is, country GDP and income were negatively correlated to child mortality. Inflation and income or GDPP is also, inversely correlated. However, income and GDPP are

directly correlated. Health is positively correlated with GDP or Income. Countries also have high export and import, if they are more developed , i.e, they have high GDPP and Per capita income.

- Based on the clusters formed, id we plot the bar graph, it's very evident that Hierarchical clustering gives better results.
- Here are the final list of countries that were obtained from best of both the algorithms:

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	ClusterID
28	Cameroon	108.0	22.2	5.13	27.0	2660	1.91	57.3	5.11	1310	1
31	Central African Republic	149.0	11.8	3.98	26.5	888	2.01	47.5	5.21	446	1
32	Chad	150.0	36.8	4.53	43.5	1930	6.39	56.5	6.59	897	1
40	Cote d'Ivoire	111.0	50.6	5.30	43.3	2690	5.39	56.3	5.27	1220	1
63	Guinea	109.0	30.3	4.93	43.2	1190	16.10	58.0	5.34	648	1
106	Mozambique	101.0	31.5	5.21	46.2	918	7.64	54.5	5.56	419	1
112	Niger	123.0	22.2	5.16	49.1	814	2.55	58.8	7.49	348	1

Question 2

State at least three shortcomings of using Principal Component Analysis.

Answer 2

Three shortcomings of PCA are:

- If there's issue of class imbalance with the dataset which is very common in case of credit card fraud detection or span detection type datasets. PCA algorithm will suggest to ignore the variables that doesn't have more variance. This is something we don't want remove from our dataset.
- PCA always assumes that compounds are orthogonal or perpendicular. Hence, in situations where features are not orthogonal or aligned to the requirements of PCA, This algorithm might not be so much helpful.
- PCA is linear transformation of original features that works well in tandem with the the linear models like Linear or Logistic regressions. Its use becomes limited in case we are trying to solve non linear regression type problems. However PCA can still be used for computational efficiency.

Question 3

Compare and contrast K-means Clustering and Hierarchical Clustering.

Answer 3

- KMeans algorithm are computationally efficient as compared to Hierarchical clustering algorithm.
- Kmeans needs value of k=number of clusters before we can work with the algorithm. There's no such dependency with Hierarchical algorithm.
- Choice of initial number of cluster that we set in Kmeans has impact on the final number cluster composition.
- Hierarchical clustering is quite intuitive in determining the number of clusters based on various needs. Once algorithm runs, we can visualize it using dendrogram and cut the cluster from whichever point we want. This is not the case in KMeans. In KMeans we need to do some further work of determining Sum of squared distances/Inertia between clusters or Silhouette score to find the optimal value of K before Kmeans can be run.