QUESTION 1

The dataset on American College and University Rankings contains information on 1302 American colleges and universities offering an undergraduate program. For each university, there are 17 measurements, including continuous measurements (such as tuition and graduation rate) and categorical measurements (such as location by state and whether it is a private or public school).

Note that many records are missing some measurements.

1. Remove all records with missing measurements from the dataset.

Answer: Removed all the missing measurements which resulted 471 records.

2. For all the continuous measurements, run K-Means clustering. Make sure to normalize the measurements. How many clusters seem reasonable for describing these data? What was your optimal K?

Answer: Pulling out the categorical measurements and standardizing the data and computing the distance. We find that 3 clusters are the reasonable for this data and the optimal K is 3.

3. Compare the summary statistics for each cluster and describe each cluster in this context (e.g., "Universities with high tuition, low acceptance rate...").

Answer: **Cluster 1 –** Universities with lowest applications received, accepted, Students enrolled from top 10 and 25 with mid-level of full time and part time students, lowest board and additional fees, lowest faculty with PHD and Graduation rate with second highest Student-faculty ratio.

Cluster 2 – Universities with highest application received, accepted, enrolled in Full time besides lowest in state tuition fees along with second in room, boarding and additional fees with highest books and personal costs and second highest with faculty holding PHD, sets highest student-faculty ratio with second low graduation rate.

Cluster 3– Universities with second highest applications received accepted and enrolled, with highest students from top 10 and 25 with lowest part time and highest in state tuition, room, boarding fees along with second highest additional fee and book costs and lowest personal cost, Holds highest number of PHD faculty with lowest student-faculty ration and highest graduation.

4. Use the categorical measurements that were not used in the analysis (State and Private/Public) to characterize the different clusters. Is there any relationship between the clusters and the categorical information?

Answer: Plotted the graph which represents all the public and private colleges along with its cluster. Each state has colleges a maximum of 2 out of the 3 clusters. Yes, there is a relationship between clusters and categorical information.

5. What other external information can explain the contents of some or all of these clusters?

Answer:

- Within cluster sum of squares with high ratio as possible
- Mean of distances between cluster centers with ration lower as possible
- Number of points in each cluster
- Cluster Centers
- The k value which the highest $withinss is the best choice, because we expect the within sum of squares ratio to be as lower as possible.

6. Consider Tufts University, which is missing some information. Compute the Euclidean distance of this record from each of the clusters that you found above (using only the measurements that you have). Which cluster is it closest to? Impute the missing values for Tufts by taking the average of the cluster on those measurements.

Answer : Its Closest to Cluster 2