K Means Clustering

Task:

For this project we will attempt to use KMeans Clustering to cluster Universities into to two groups, Private and Public.

It is \*\*very important to note, we actually have the labels for this data set, but we will NOT use them for the KMeans clustering algorithm, since that is an unsupervised learning algorithm. Use the label to vary model performance.

Data set:

When using the Kmeans algorithm under normal circumstances, it is because you don't have labels. In this case we will use the labels to try to get an idea of how well the algorithm performed (confusion matrix, accuracy and etc). We will use a data set with 777 observations on the following 19 variables.

1. Name of the college
2. **Private: labels (don’t use this feature in your model training!)**
3. Apps: Number of applications received
4. Accept: Number of applications accepted
5. Enroll: Number of new students enrolled
6. Top10perc: Pct. new students from top 10% of H.S. class
7. Top25perc: Pct. new students from top 25% of H.S. class
8. F.Undergrad: Number of fulltime undergraduates
9. P.Undergrad: Number of parttime undergraduates
10. Outstate: Out-of-state tuition
11. Room.Board: Room and board costs
12. Books: Estimated book costs
13. Personal: Estimated personal spending
14. PhD: Pct. of faculty with Ph.D.’s
15. Terminal: Pct. of faculty with terminal degree
16. S.F.Ratio: Student/faculty ratio
17. perc.alumni: Pct. alumni who donate
18. Expend: Instructional expenditure per student
19. Grad.Rate: Graduation rate