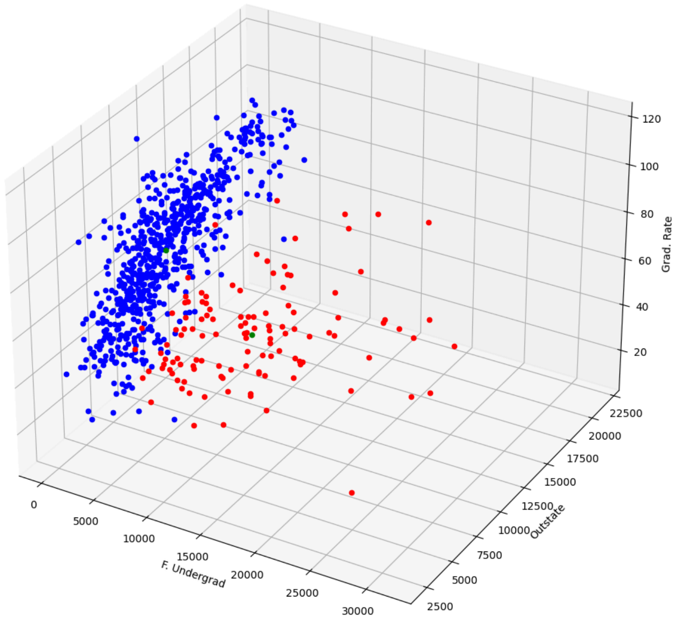
GoogleColab:
https://colab.research.google.com/drive/1cc7MgtACINtO7FqxDA5ZS13vQRfw6vFw?authuser=2#scrollTo=UE8KpyfHbCQE

A file called input.csv must be in the same directory as the ipython notebook for the code to run.

Note that when viewing the plots, it can be seen that the data assigned to the x and y axis are swapped from the initial initialization of the data dataframe. After trial and error, the plot of the original data set made it hard to decipher the clusters. By swapping the x and y axis, the plot is simply being viewed from a different angle and therefore has no effects on the clustering.

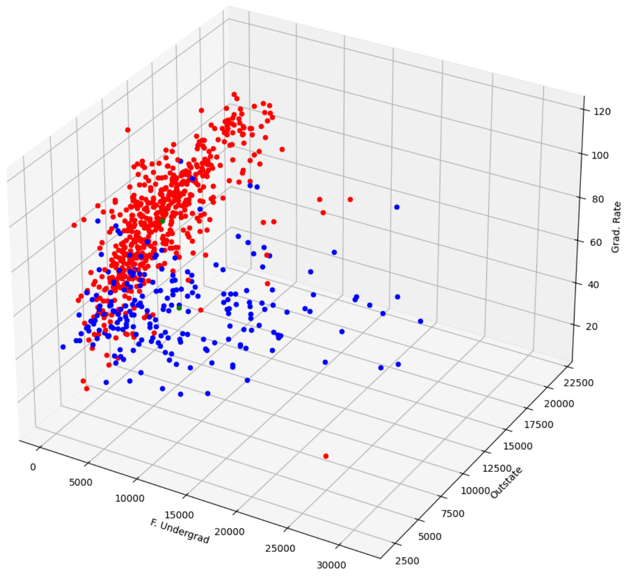
The coloring of clustering is random where all data points with each color are in the same color but one color does not necessarily always map to private or public in the real data.

Original Data Calculated Clustering:



Calculated Centroids: (7657.311, 13301.840, 58.109) and (10944.043, 1963.388, 66.793)

Original Data Actual Clustering:

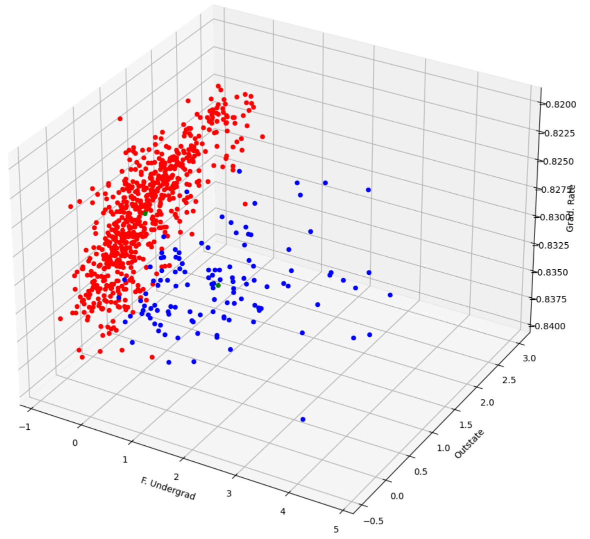


Real Centroids: (6813.410, 8571.005, 56.042) and (11801.694, 1872.168, 68.998)

Accuracy: 84.685%

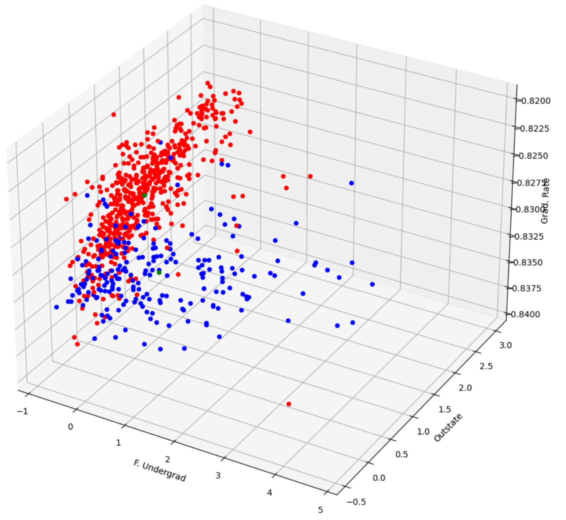
Observing the plot of the data with the real centroids, calculated based on the labels given in the data set, the reason for some discrepancies in the calculated clustering can be seen. In the K-Means clustering algorithm, the overlap of private and public school data points in the lower left hand corner will be grouped into the larger group of public schools. Due to this, the model is inaccurate on examples similar to those while quite reliable on the other samples, thus giving above a 80% accuracy rate.

Standardized Data Calculated Clustering:



Calculated centroids: (0.519, 1.522, -0.831) and (1.103, -0.492, -0.829)

Standardized Data Actual Clustering:

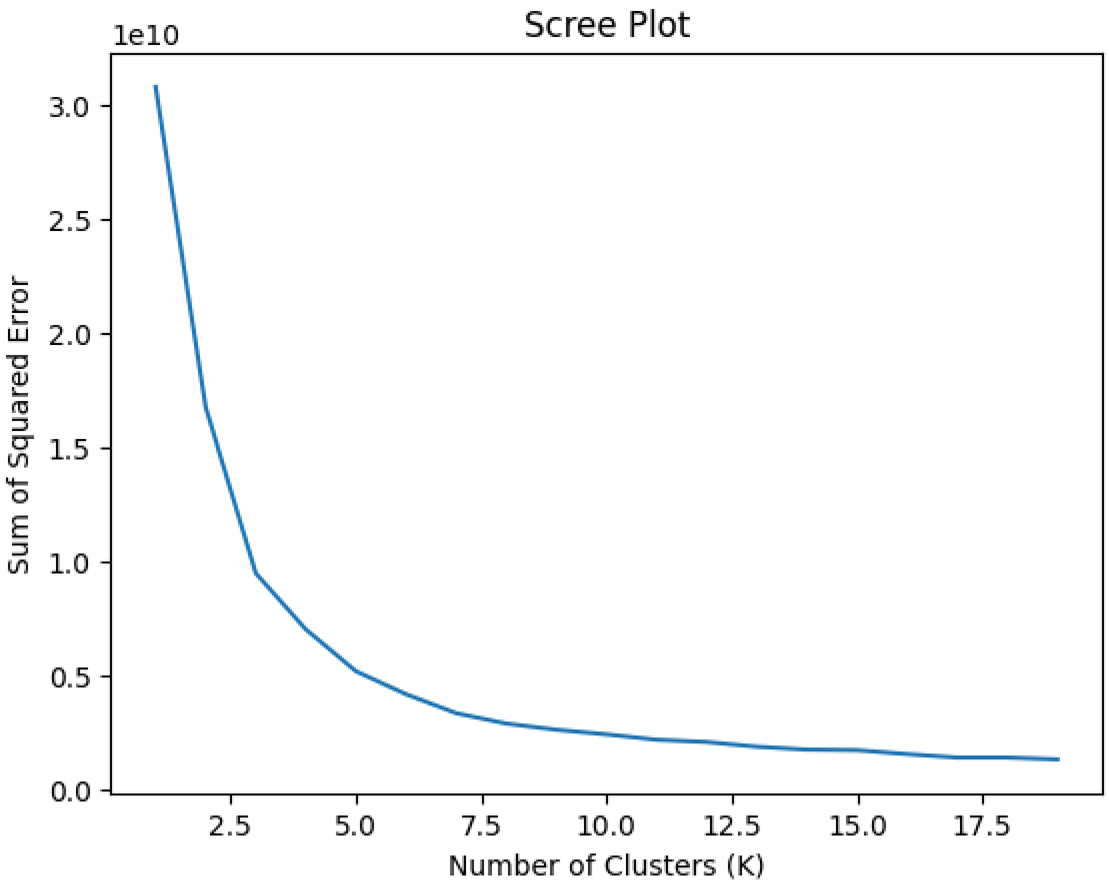


Real Centroids: (-0.902, 1.005, -0.549) and (0.339, -0.377, 0.206)

Accuracy: 84.685%

The standardized data resulted in almost identically the same clustering and accuracy as the original data.

Scree Plot:



The elbow does not occur at k=2. There appears to be somewhat of an elbow at k=6 where the accuracy of the clustering stops decreasing significantly, but does not flatten completely.

Contributions:

After discussing the logic and structure of the algorithm with Peter, we collectively implemented the functions to train and evaluate the model. Peter focussed on solving some issues with the clustering of the standardized data. I also created the scree plots to find the optimal k value.