Create Institution Segments based on various fields

Ziqi Liu

Intro/Background:

The primary purpose of the Institutional Scorecard was to help prospective students make informed admissions decisions. It provided information for students and families who compare college costs and outcomes as they weigh the trade-offs of different colleges, taking into account their own needs and educational goals. As a student who has also been involved in the school application process, I knew that the original data set contains too much information, making it difficult for students and parents to directly find the right school that meets their needs. Based on this, I was interested in separating these institutions primarily based on student earnings after graduation, so that when choosing schools, students and parents would know which schools can help them achieve their financial goals, while multiple schools in the same cluster would help give them more options.

The data I used was collected through the US Department of Education. Most attributes were provided through federal reporting from institutions, data on federal financial aid, and tax information. It mainly described seven aspects of each institution: **self-information, academic profile, admission, cost, financial aid, completion and post-graduation earning**. Based on that raw data I have, I manipulated the data through standardizing it, filtering out outliers, generating the feature selection and handling the missing value. Then I applied PCA and K Means algorithms to generate institution segmentation and visualize it.

Methods:

For the preprocessing steps, it mainly contains those four categories:

1. **Standardized data:**

After fixing structural errors, including removing spaces and checking for typos or case inconsistencies. I made sure that all values of the string are lowercase or uppercase, and that all values have a specific numerical unit of measure. In addition to this, I also normalized the data types by converting the types of some fields from "object" to "float" so that they could be used in the calculation of the relevant scores.

2. **Filter unwanted outlier:**

As a preprocessing step before reading the dataset, I filter out outliers in the original dataset by generating boxplots for visualization or calculating values above three standard deviations to exclude rows with very large or very small values.

3. **Calculate Correlation and Feature Selection:**

There were more than 2,000 columns included in the raw dataset, which described each feature in a very comprehensive way. Take earnings of students who graduated after 10 years as an example. It had 10th, 25th, 75th, 90th, median, average, standard deviation of earnings.
First of all, for the selection of key indicator, after comparing the distribution of students' earning 1,3,6,10 years after graduation, I thought the '10 years later' data is more representative, because earnings of graduated students didn't change too much in the early stage of their career and they might consider pursuing another degree or trying a variety of positions in a different industry, which made the earnings in the early years after graduation may not be indicative of longer-term earnings. After that,I calculated the correlation between the key metric (*MN_EARN_WNE_P10:mean earnings of students working and not enrolled 10 years after entry*) and other attributes, and generated an initial selection based this score, then if multiple features were included in the initial selection for each category, I only kept the feature with highest correlation score as the most representative factor of each feature. It also reduced collinearity in this way.

4. **Handle missing data**

After the final selection is complete, I filled in the missing value with the mean of each attribute column. But before that, I also checked the feature information to make sure the missing percentage of the selected columns is not too large so that the filled data wouldn’t cover the true underneath trend.

Algorithm:

In this project, in order to achieve the clustering purpose, I chose to use K-means.
K-means is an unsupervised algorithm, which means we don’t have labels or groups available as the reference from the dataset and we need to cluster those data points into different clusters based on similarity. The euclidean distance is used as the similarity metric in this case.
Firstly, we need to have a target cluster number K, which represents the number of centroids we need, and we randomly select the initial centroids. Then we calculate the distance between every

data point and those centroids based on similarity and assign the data points into the closest cluster centroid. After that, we re-generate the centroid of each cluster. We will repeat this processing until the centroids have stabilized, which means those centroids don't move anymore.

The space complexity of K-means clustering algorithm is O((n+K)d)
◦ *n = number of points,* ◦ *K = number of clusters,* ◦ *d = number of attributes.*
The time complexity of K-means is O(n*K*I*d)
◦ *n = number of points,* ◦ *K = number of clusters,* ◦ *I = number of iterations,* ◦ *d = number of attributes*

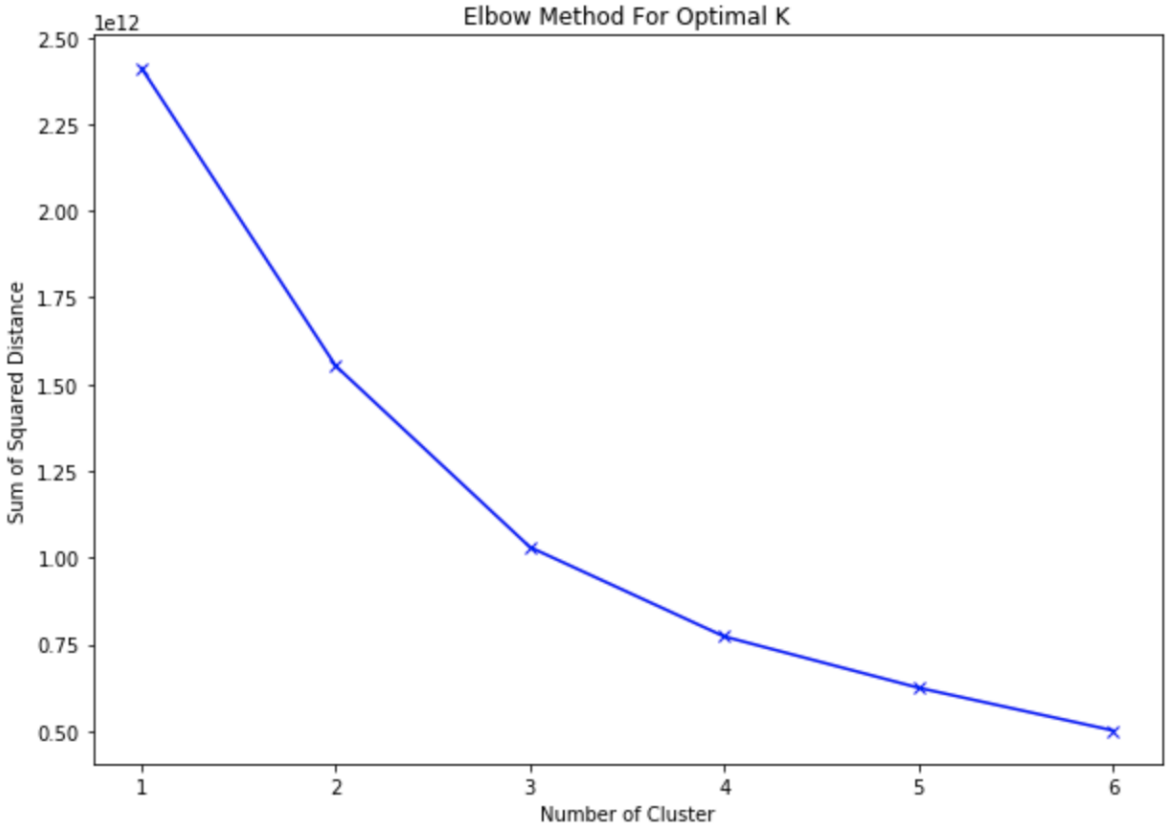
I also used PCA, which reduces the dimensions to two so it will be easier to generate a plot for cluster visualization.
Firstly, we need to standardize the range of variables so that each data point will have equal contribution to the analysis, then we generate the covariance matrix and compute the eigenvectors as well as the eigenvalues of the covariance matrix to get the principal components.

The complexity of the PCA algorithm is $O(p^2n+p^3)$, which includes covariance matrix computation as $O(p^2n)$ and eigen-value decomposition $O(p^3)$.
◦ *n = number of points,* ◦ *p = number of attributes.*

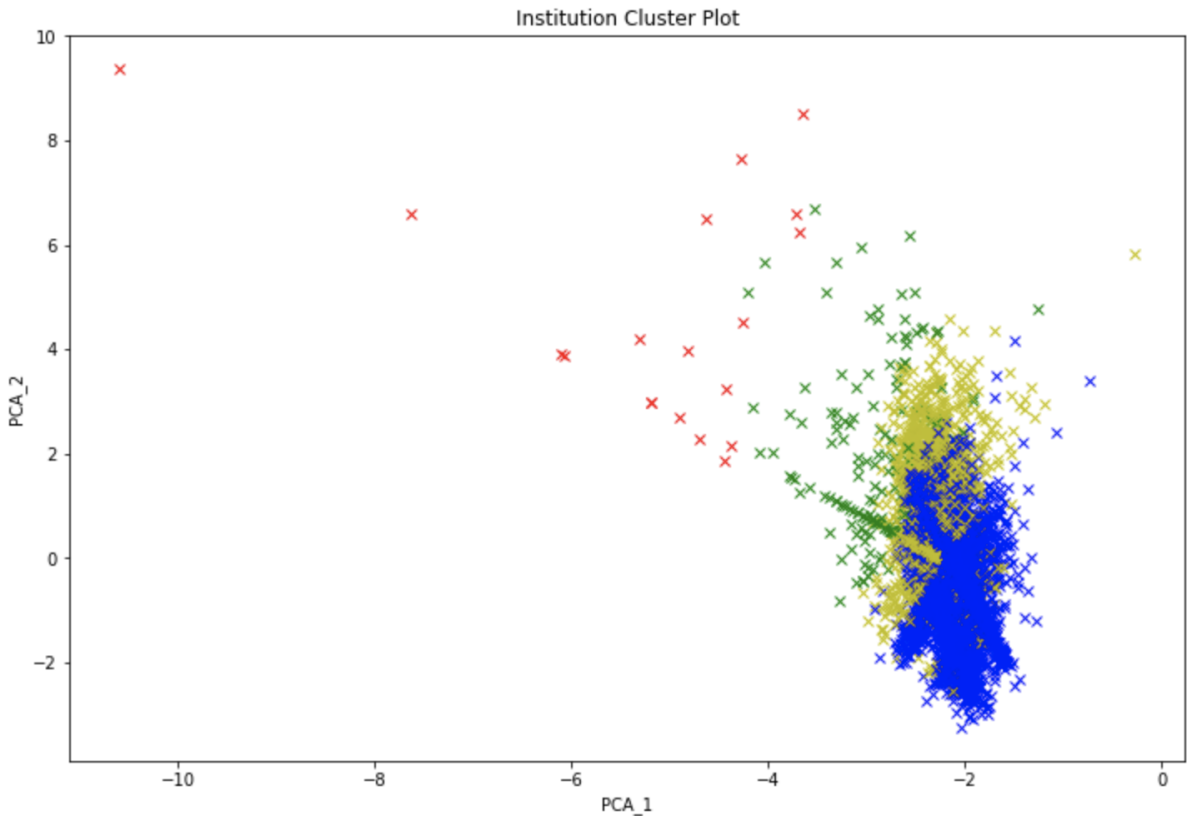
Results:

In the final model, I chose to use those predicted attributes as the primary representative of each cluster:
ACTMT25 - 25th percentile of the ACT math score
ADM_RATE - Admission rate
AVGFACSAL - Average faculty salary
INEXPFTE - Instructional expenditures per full time student
PPTUG_EF - Share of undergraduate, degree-/certificate-seeking students who are part-time
MALE_RPY_7YR_RT - Seven-year repayment rate for males

And set the number of clusters to 4 according to the elbow method plot. Obviously, 3 or 4 should be the ideal number of clusters. The final number is set to 4 after comparing the RSS for models with three or four clusters.



The final clustering results are shown in the figure below, it is obvious that the blue cluster contains the most institutions, while the red cluster contains fewer institutions. According to the detailed cluster information listed in the table below, the red cluster is cluster 4 and the blue cluster is cluster 1, indicating that the best performing institutions are less than the average institutions, and it also proves the scarcity and preciousness of high-quality educational resources



Feature/Cluster	Cluster 1	Cluster 2	Cluster 3	Cluster 4
ACT Math Score	19.33	20.00	21.35	21.97
Admission Rate	69.28%	66.16%	55.07%	52.82%
Average Faculty Income	$5656.58	$7270.50	$8638.98	$9489.43
Instru Expenditure	$4599.60	$12407.21	$37518.0	$118089.05
Part time Percentage	23.99%	16.37%	15.01$	15.19%
Male repayment rate in 7 years	59.76%	70.66%	68.07%	70.68%
SAT Score	1049.93	1078.78	1150.29	1170.94
Earning in 10 years	$34642.41	$45658.69	$67446.25	$88793.66

After checking the segmentation result, for the group which has the highest earning in 10 years, it shows that the institutions in this group with higher graduate incomes also have stricter admissions requirements, including higher standardized test scores (SAT and ACT) and lower acceptance rates. They also have stricter requirements on educational methods, such as less tolerance for part-time students, and a greater desire for students to focus on their studies. In addition to this, they also have better staff and student benefits such as higher faculty salary and higher instructional expenditure. While the opposite is true for the lowest earners of the decade, with lower entry requirements and less focus on academic progress, which also meets our expectations.

Conclusion and Discussion:

In summary, I chose to use Kmean to cluster institutions in this project. The clustering results provide us with an overarching view that helps students make admissions decisions to achieve their educational and financial goals. Knowing which group their target school belongs to, they can have more options to apply to some other schools with similar characteristics.

There is still something we can imporve in the future including the data set manipulation and feature selection.

- Since the dataset we have is yearly, it is difficult to merge information from multiple years due to the sparsity of the original data and the value may change from year to year.
- For feature selection, we can include some other methods to determine correlations and select predictive features, such as statistical tests or generative linear regression to identify the most influential attributes.

References:

Material Type	In-text Citation	Bibliography
A website	(Education Ecosystem 2018)	Education Ecosystem (Sep 12th, 2018) Understanding kmeans clustering in machine learning https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1
A website	(Zakaria Jaadi 2022)	Zakaria Jaadi (July 14th, 2022) A Step-by-Step Explanation of Principal Component Analysis (PCA) https://builtin.com/data-science/step-step-explanation-principal-component-analysis
A website	(Tola Alade 2018)	Tola Alade (May 27th 2018) How to determine the optimal number of clusters for k-means clustering https://blog.cambridgespark.com/how-to-determine-the-optimal-number-of-clusters-for-k-means-clustering-14f27070048f