141SL Final Project Report

*Yan Kang, Duyun Tan, Shuocheng Dong, Shicheng Chen,
Chen Chen, Chen Zhang, Menglu Wang, Yifei Du*

Table of Contents

Abstract

This analysis examined how successful students were after they finished their program in the UCLA Extension. We found that some of the data was of no use as some students skipped through while they took the survey, and it became a major difficulty of the analyzing process. After we got a cleaned dataset, we used K-means Cluster Analysis and separated the data into two groups. We found that the students in cluster 2 actually perform better, and we can conclude that students with similar characteristics as them in the cluster are likely to be more successful.

Introduction

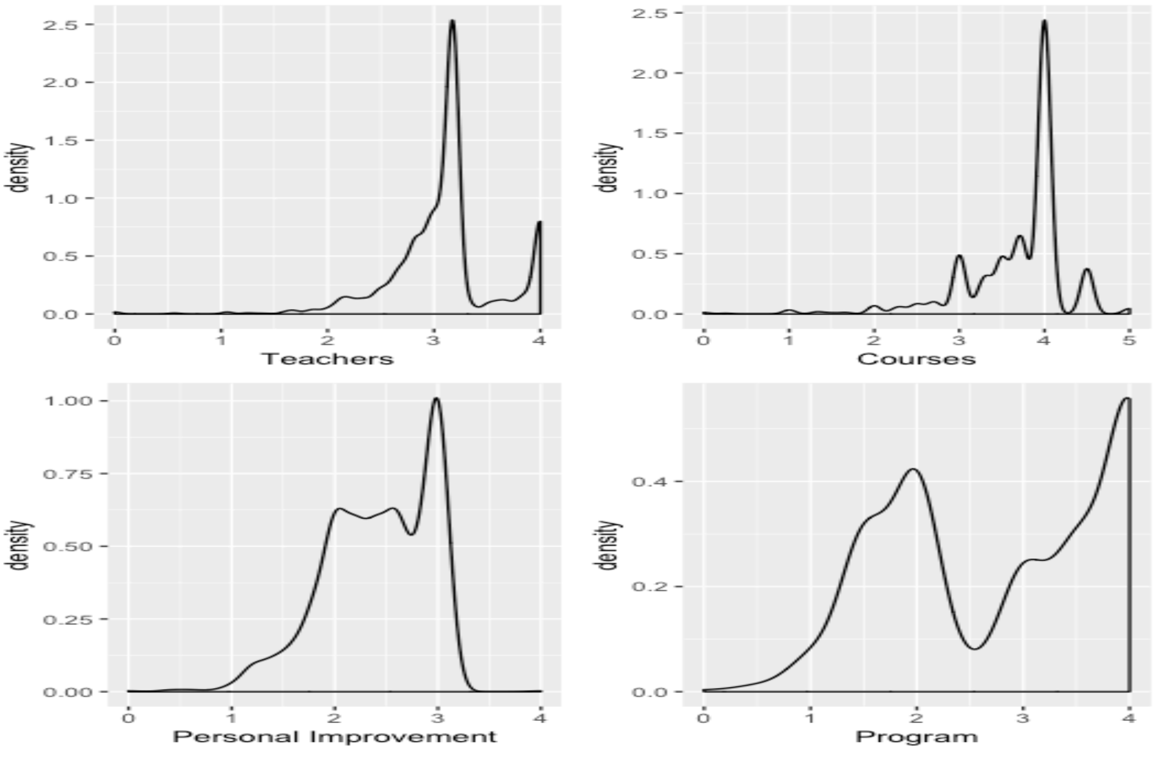
At the UCLA Extension, we want every student to be satisfied with their program experience and become successful academically as they exit the program. To better understand our students and provide better support in the program, we need to first define and validate "success" for our students. The data we received from UCLA extension contains 29 tables. These datasets include information on student background, placement test results, program evaluations, and market results in the year 2016 - 2018 (Include Summer Sessions). The goal of this project is to define and validate "success" for our students using the massive information provided in the datasets.

Data Cleaning

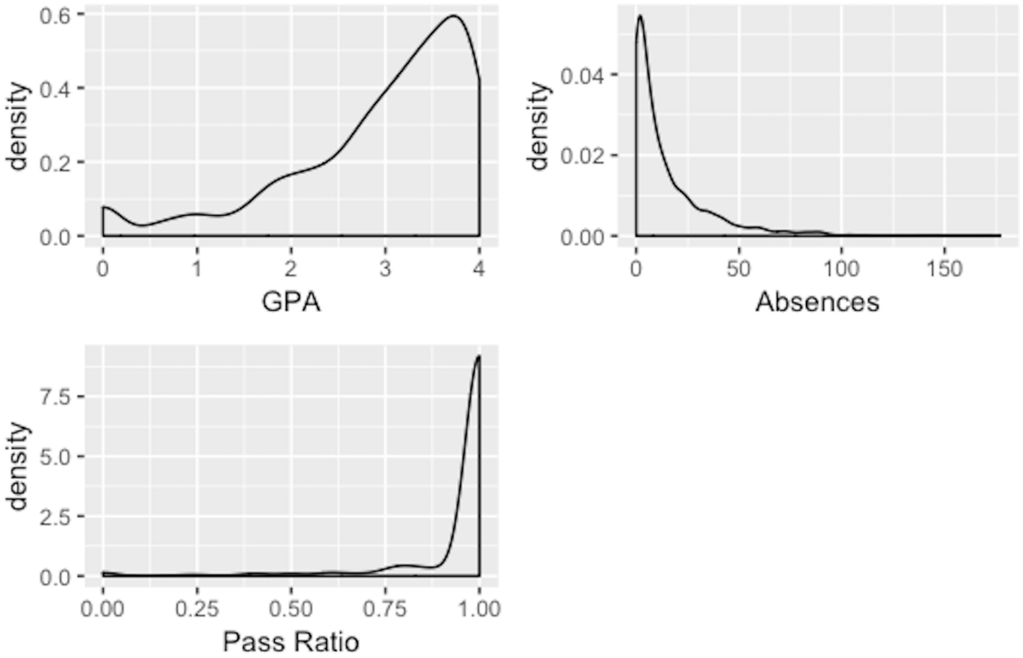
The datasets consist of information pertaining to students' opinions about their classes, professors, and the ALC program as well as their performances. We first dropped some variables that were considered unrelated to the topic. We re-coded some categorical variables to numeric variables as well as added some new variables. For the variables such as "my general evaluation of the teacher", students give response A (very good) - E (very bad), we converted these letters to a numeric variable where a higher score indicates a positive attitude. For the survey question, "will you recommend ALC to your friends", we set "yes" and "no" to binary variable 1 and 0. In this way, we were able to classify the variables on evaluations into four categories which we think would relate to the topic. These four categories are courses, teachers, personal improvement, and program, and we also calculated their averages. However, some datasets do not include questions in some of the four categories, so we did not use them. We renamed them as mean_score_teachers, mean_score_courses, mean_score_on_personal improvement, and mean_score_on_program and put them into use in the later discussion. Moreover, we selected, recoded and added some variable such as GPA, pass ratio and absence that demonstrate students performances during attending ALC programs. The way we created GPA and pass ratio variables was to use the grade variable to do some re-coding. For example, classes with letter grades, we transferred the letter grades into GPA scores range from 0.0 - 4.0 and calculated a cumulative GPA. For classes were taken with grade type of pass or fail, we converted it into a pass ratio where we divided the passed courses by the total number of courses that student took. There was a situation in which students repeated courses To resolve this, we called these repeat classes "duplicate observations". To cope with these observations, we averaged their scores in each category so that each student ID is unique.

Exploratory Analysis

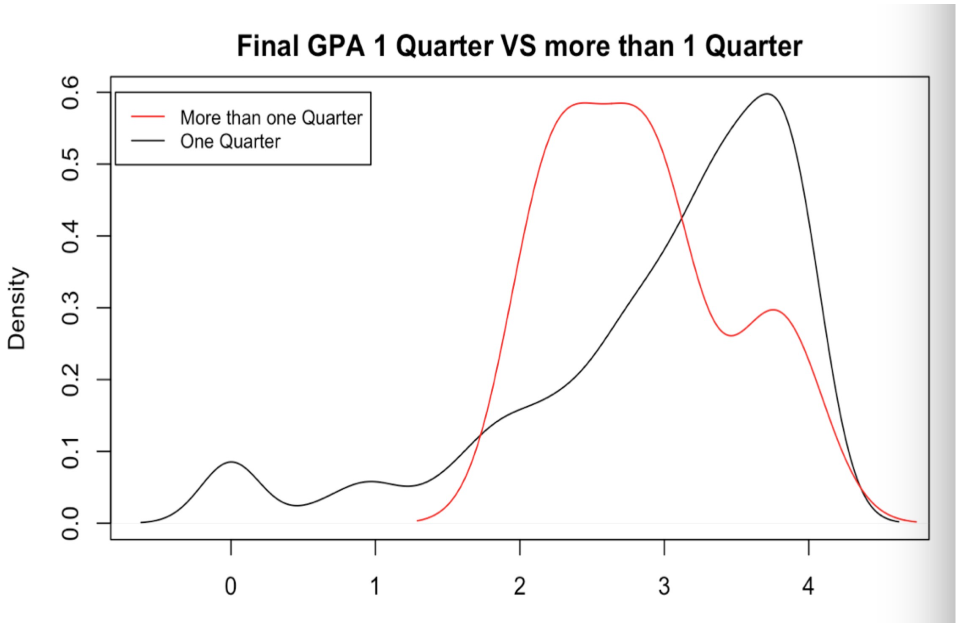
After the cleaning data process, we did some exploratory analysis. As we can visualize from the four plots below, most evaluations on teachers are around 3. Several very high scores and few very low scores; most evaluations on courses are around 4. However, The scores on evaluations on personal improvement and program are more spread out, which indicates they have relatively higher variances.



For the performance variable plots below, GPA is left skewed meaning that most students are able to maintain a GPA between 3 and 4. Absences are right skewed indicating that very few students have high frequencies of absences and most students' total number of absences was less than 25. For the last pass ratio plot, we can infer that most of the students have a pass ratio higher than 0.9.

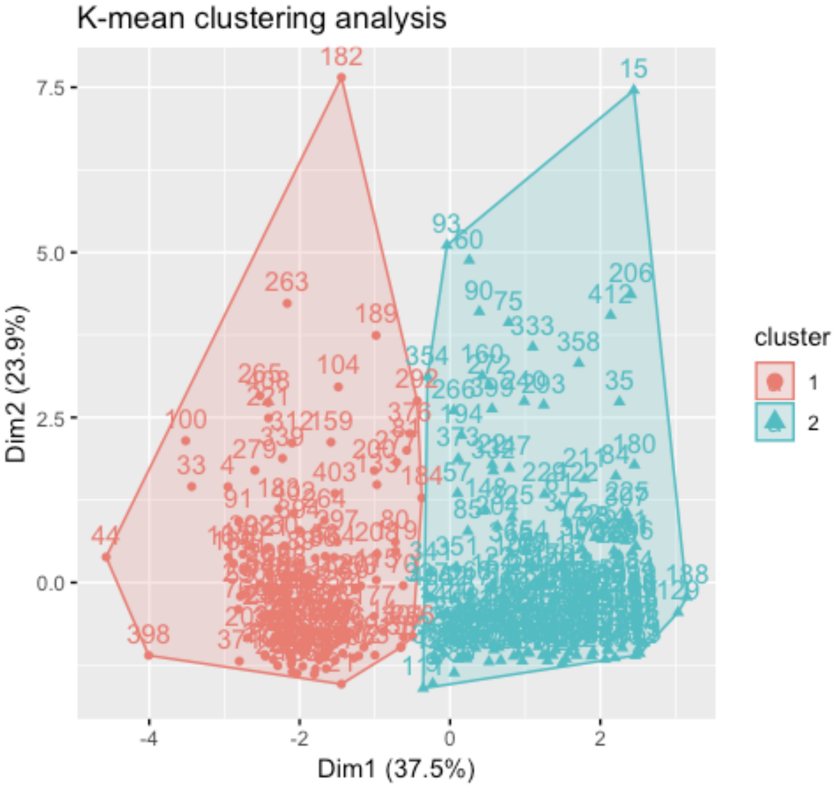


we were also interested in exploring if re-enrolled students have significant different GPA compared to students who enrolled one quarter. The graph below demonstrates final GPA of one-quarter students and final GPA of more than one-quarter students. The red curve shows re-enrolled students' non-first quarter GPA. We infer that students who took the courses for the second time would mostly end up receiving a GPA of 2-3 while student who took the courses once would mostly receive a GPA between 3 to 4.

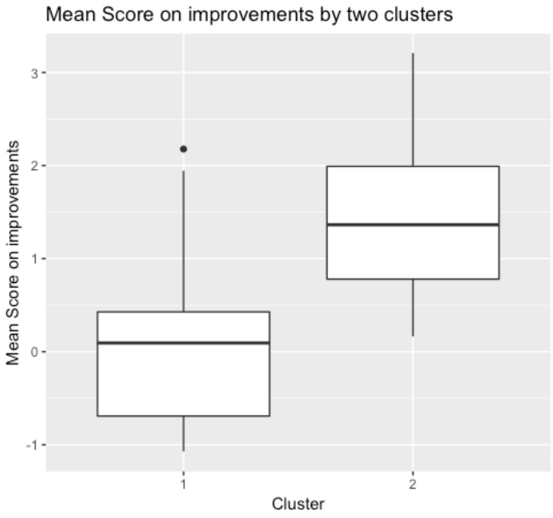


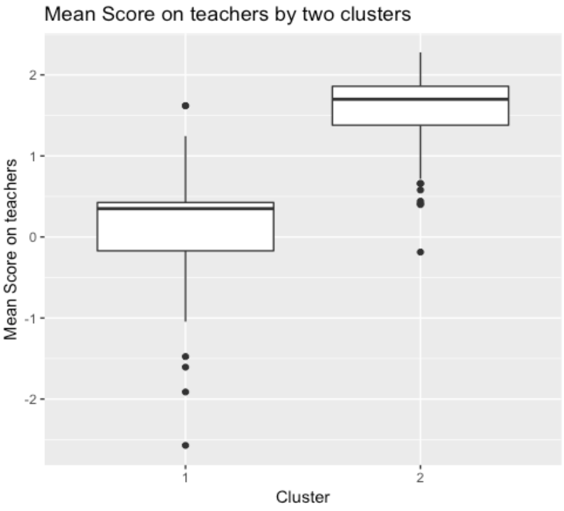
Methodology - K-means Clustering Analysis

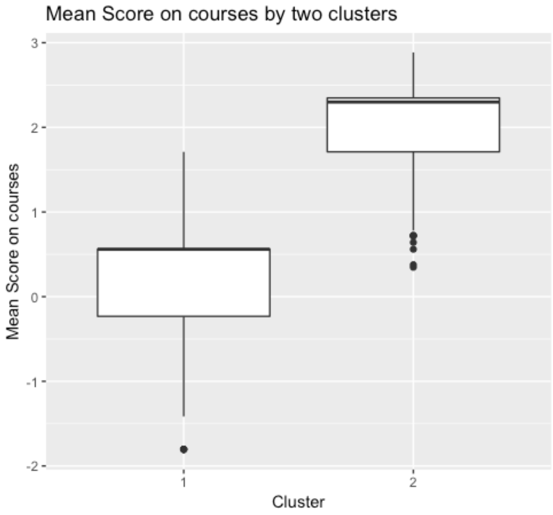
K-means clustering is a type of unsupervised learning for unlabeled data (i.e., data without defined categories or groups). The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable *K*. The algorithm works iteratively to assign each data point to one of *K* groups based on the features that are provided. Data points are clustered based on feature similarity (Trevino). For our dataset, we used the k-means clustering method to separate students into 2 groups based on our transformed variables. After applying this method, we found out that there are 132 students in cluster 1 and 202 students in cluster 2. The advantages of the K-means clustering method is to be able to determine what types of groups exist or to identify unknown groups in complex data sets. Once the algorithm has been run and the groups are defined, any new data can be easily assigned to the correct group.
The plot below shows the two clusters in our dataset:

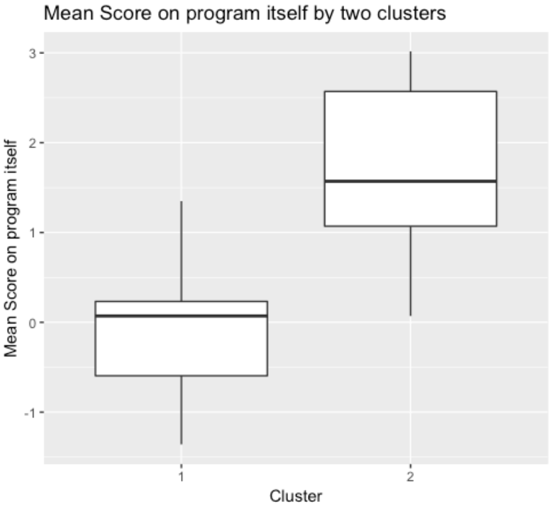


The four plots below demonstrate boxplots of mean score improvement, mean score teachers, mean score on courses, and mean score on the program based on two clusters.

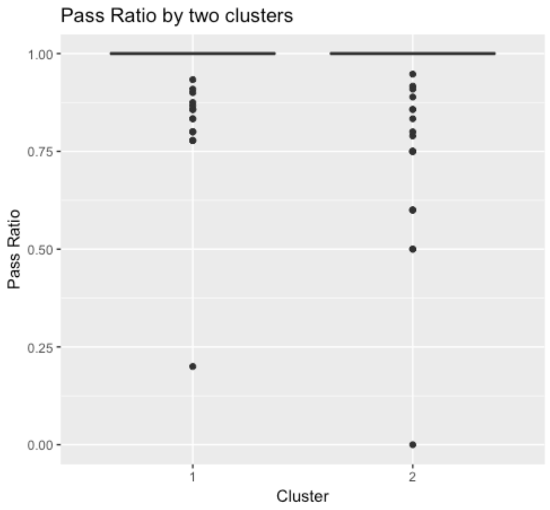


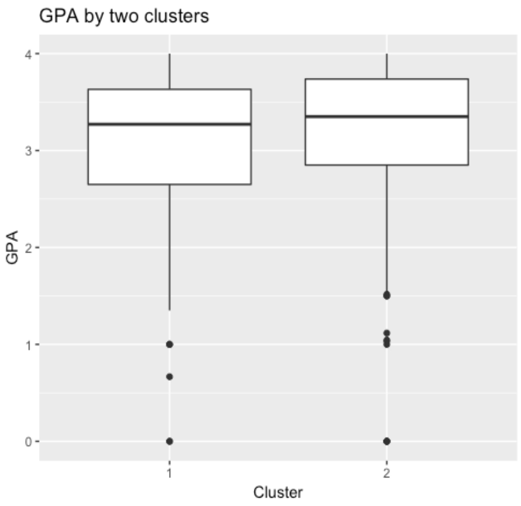




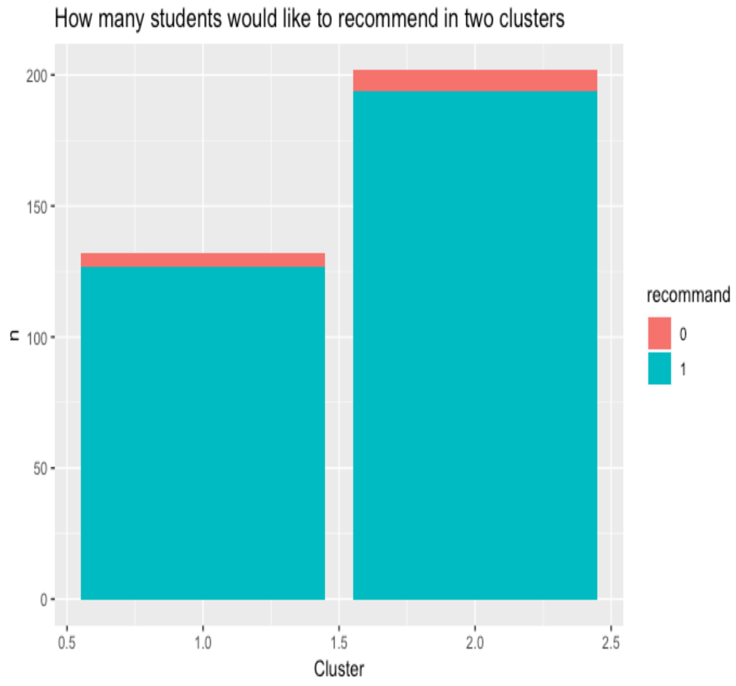


These two boxplots below show pass ratio and GPA by two clusters, respectively.



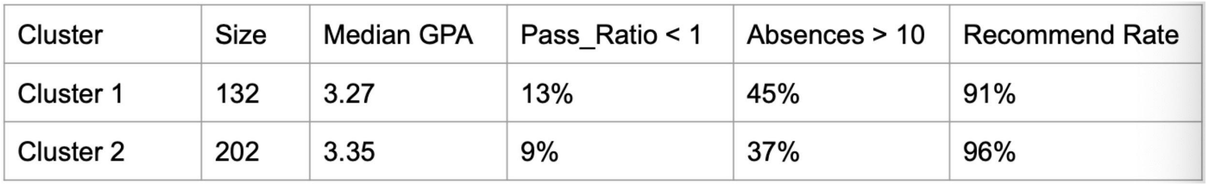


The bar plot below shows the number of students who would like to recommend the program against the number of students who would not recommend based on two clusters.



From these plots above, we can infer that there are no significant differences in GPA, times of absences and willingness to recommend between two clusters. Attitudes toward courses, attitudes toward teachers, attitudes toward the program, and attitudes toward self-improvements show significant differences between two clusters. However, we noticed that there are some influential outliers in the plots of pass ratio by two clusters and GPA by two clusters. Therefore, we did more explorations on these variables. Besides, we would also like to know the percentage of each group in terms of absence more than ten times as well as the actual recommend rate in each group.

The chart below demonstrates the median GPA, the percentage of low pass ratio (Since most of the students passed their courses, we considered the pass ratio less than 1 is low pass ratio), the percentage of absences more than ten times, and recommend rate between two clusters:



Cluster	Size	Median GPA	Pass_Ratio < 1	Absences > 10	Recommend Rate
Cluster 1	132	3.27	13%	45%	91%
Cluster 2	202	3.35	9%	37%	96%

132 students are assigned to cluster 1, and 202 students are assigned to cluster 2. Median GPA for cluster 1 is 3.27 while it is 3.35 for cluster 2. The percentage of students who have a pass ratio lower than 1 is 13% in cluster 1 and 9% in cluster 2. The percentage of students who were absent more than 10 times is 45% in cluster 1 and 37% in cluster 2. The “recommending rate” for cluster 1 is 91% and 96% for cluster 2. There is no significant difference in median GPA between two clusters. Cluster 2 has a relatively smaller percentage of low pass ratio than Cluster 1. Students in Cluster 2 are relatively more likely to attend classes than students in Cluster 1. The majority of students in both clusters are willing to recommend the program.

Conclusion

We came to the conclusion that students in cluster 2 are considered “more successful” in terms of their performances in the ALC program. If we were to get any new data in the future, based on their response to the specific questions on the survey, we should be able to assign them to the correct group.

Limitations

The first limitation of our project is that the number of observations is not large enough. Clustering would likely perform better with more observations included. Secondly, “Success” of a student can be defined on a wide of range of aspects. Moreover, the aspects of success reflected in the datasets are very limited, and these aspects are also useful for clustering. Thus, the variables we included in the clustering capture a very limited part of “success”. According to the comparison graph of GPA of the two clusters, we can see that there are still a few outliers (students with low GPA) in the “more successful cluster”. We suspect there are students who do not take the evaluation survey seriously and their answers do not accurately represent their real attitudes towards teachers, courses, personal improvement and the program. As a result, both response bias (Students who did not take surveys seriously) and non-response bias (Students who refused to answer some of the questions) are included in our findings. These bias would potentially harm our model efficiency.

Recommendation

We would recommend staffs to use consistent surveys for each quarter so that all observations will be included in the model. In addition, we would recommend staffs to not forcing students to take surveys or give them incentives for the purpose of contributing to the most accurate result. Students who click through all questions will be included in our model which could affect the precision of the final result. One last recommendation is not asking too many similar questions. While we were aggregating columns, we noticed many questions are asking the same aspects. Students are less likely to answer the same questions which will reduce their incentives to seriously answer questions. The issue will enlarge the response bias which could reduce the model performance.

Source

Trevino, Andrea. Introduction to K-means clustering. *Oracle*. 23 Mar. 2019. https://www.datascience.com/blog/k-means-clustering