DS 740
Final Project

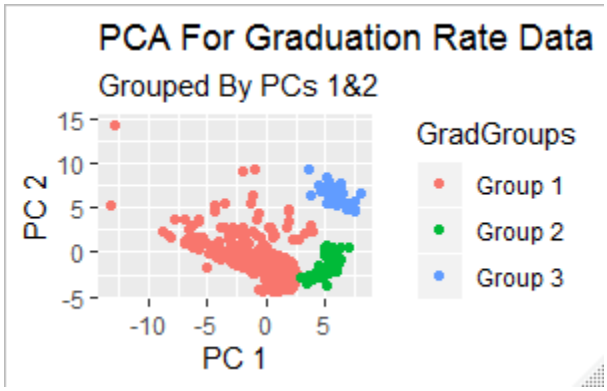
To help the Massachusetts Department of Education identify ways to improve the outcomes in the state's high schools, I analyzed a dataset of Massachusetts Public School data from 2017. My goal was to use principal components analysis (PCA) and association rule mining (using the apriori algorithm) to identify variables that may be associated with better outcomes for students in Massachusetts High Schools. For my analysis, the outcome variables that measure the quality of a student's outcome were:

- ***Graduation Rate***: % of students who successfully graduated from high school.
- ***College Attendance %:*** % of students who graduated in the prior year who are attending college.
- ***Average Total SAT Score***: The total of the average Math, Reading and Writing SAT scores taken by students at the school for the year.

I chose these three outcome variables because there is no single definition of a successful outcome for a student. Analyzing and reacting to a single metric can incentivize schools/teachers/administrators to take direct action to improve that metric without regard for if a student's experience is actually improved. By looking at three different outcomes variables and the input variables that are associated, we can search for intervention opportunities that could have a positive effect across many kinds of schools and students.

First, I used principal components analysis to cluster groups of schools together based on the variables in the data. For each outcome, I excluded the outcome from the data and clustered schools together based on the other variables in the data set. Once the clusters were formed, I split the schools into 3 groups and measured the average outcome variables for each of the three groups.

For graduation rate, we can see that two clusters (Groups 2 and 3) were clearly split out from the main group in our analysis, suggesting that they are more similar to each other that the other groups based on the data in the file.



After splitting the data into groups I found the mean graduation rate for each of the groups, as shown here:

Group 1 = 91.2%
Group 2 = 40.5%
Group 3 = 64.6%

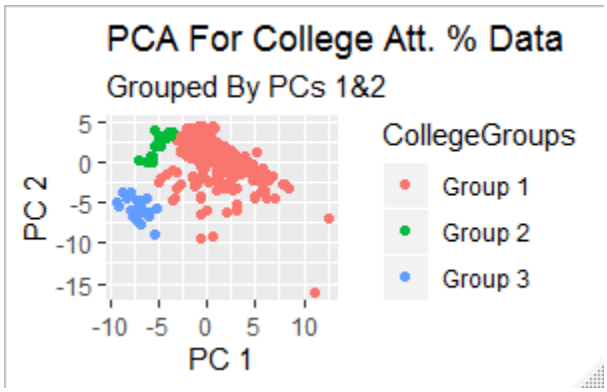
We can see that groups 2 and 3 showed drastically lower graduation rates than group 1. Knowing that, I looked at the variables that were most important in splitting groups two and three from group 1. I found that groups 2 and 3 were mostly defined by:

- Higher % of students whose first language was not English (group 3 especially).
- Higher % of students who were English language learners.
- Higher % of students who were classified as "High Needs".

- Higher % of students who were economically disadvantaged.
- Higher % of students whose first language was not English (group 3 especially).
- Higher % of students who have disabilities (group 2 especially).

I performed the same analysis on the College Attendance % and Average Total SAT Score:

College Attendance %:



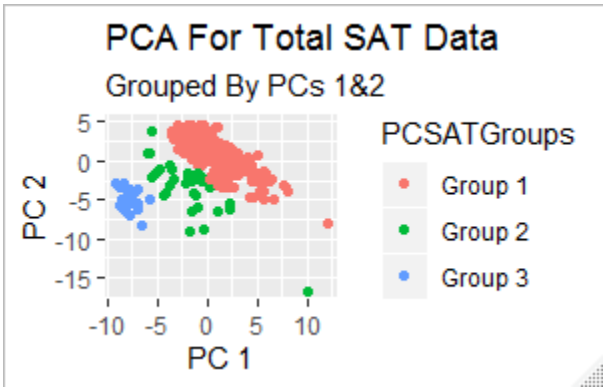
Average College Attendance % by Group:

Group 1 = 77.5%
Group 2 = 37.6%
Group 3 = 64.4%

Groups 2 and 3 were defined by:

- Fewer AP test taken. There were 0 AP tests taken by students at the schools in group 2.
- Higher % of students whose first language was not English.
- Higher % of students who are economically disadvantaged.
- Higher % of students who were classified as “High Needs”.

Total SAT Score:



Average Total SAT Score by Group:

Group 1 = 1,538
Group 2 = 1,305
Group 3 = 1,135

Groups 2 and 3 were defined by:

- Higher % of students whose first language was not English.
- Higher % of students who were classified as “High Needs”.
- Higher % of students who are economically disadvantaged.

Interestingly, the underperforming groups in all three analyses were also associated with higher average expenditures per pupil. So, as we dig further into interventions, we should note that increasing spending without targeting that spending will likely not change these outcomes significantly.

I also performed an association rules mining analysis to determine if there are associations between the input variables and the output variables that we are trying to target. In the output from the association rules analysis, I found many rules that showed similar associations to the associations we found in the PCA. A few examples:

- Several rules showing that very low percentages of students who are economically disadvantaged (less than 15%) are associated with very high percentages of students attending college (over 85%).

- Several rules showing that very low percentages of high needs students (less than 20%) were associated with very high (over 95%) graduation rates.

- Several rules showing that high percentages of economically disadvantaged students (over 50%), high percentages of high needs students (over 60%) were both associated with very low SAT scores.

We have seen from our analysis that a few variables are associated with poor outcomes in all of our analysis. Specifically, we see disproportionately poor outcomes for students who are:

- Economically Disadvantaged
- Students whose first language is not English
- Students who are classified as high needs

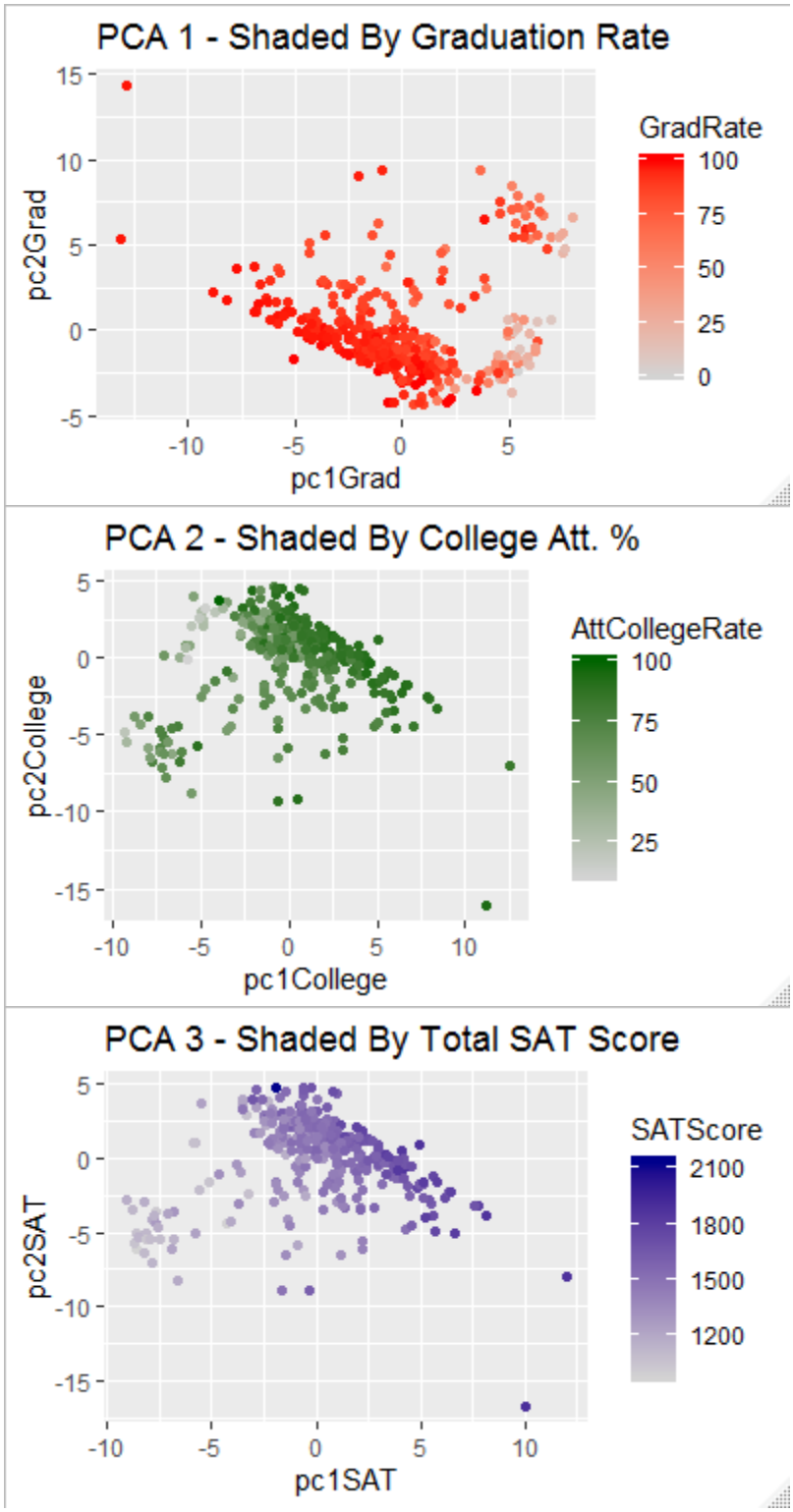
The Department of Education should explore programs/interventions that are targeted at improving outcomes for students in these groups because the interventions will improve student outcomes across all of our outcome measurements.

There are a few ways that the MA Department of Education could explore improving outcomes for students based on this analysis:

1. There could be an opportunity to improve the outcomes of students whose first language is not English, which is highly correlated with students with high needs (0.956), by building out the state's bilingual education offerings. Additionally, there is an opportunity (for all states/universities) to consider offering standardized tests in languages other than English. There are likely very capable students who are receiving low scores on the reading and writing portions of the SAT simply because English was their second language.
2. There could be an opportunity to improve outcomes for students who are economically disadvantaged by expanding the offerings of after school, mentoring and/or counseling available to students. Students who come from economically distressed families likely have parents at home who need to work long hours and/or multiple jobs to make ends meet. Can the schools expand offerings designed to help the students keep up with work and give them the support they need during the hours outside of school?

There is not a silver bullet that will allow the MA Department of Education to improve outcomes for all students; so by using the limited resources available to the schools to target the students with the most need they will be able to improve outcomes most effectively.

Appendix – Not considered part of the Executive Summary, I just thought these were interesting graphs from the PCA and wanted to include/share.



Thanks for a great semester, this was my favorite class in the program so far, lots of interesting stuff!