PASSNYC Technical Report (excerpts) using clustering

Abstract

At the heart of this research was deciphering the relationships between the schools of New York City. Some of this research looks at economic need, test scores, demographics, and the surrounding community. To do this, linear discriminant analysis, k-means clustering, canonical correlation analysis, multiple linear regression, and principal component analysis were used. The results were sometimes expected (schools with a higher economic need tended to have lower test scores) and sometimes not (attendance rate has little bearing on average test scores). This data set was originally used to find under performing schools and send assistance to students there so that more students could apply for and be admitted to a specialized high school. Many of the methods used above were able to separate lesser performing schools and find correlations (not necessarily causeations) to those low performing schools. Canonical Correlation shows that race plays a significant part in the classroom. Particularly it was found that schools with African American students have lower standards in the classroom, have teachers that are not as committed to the success and improvement of their classroom and schools, and have less developed relationships with families, businesses, and community-based organizations. Using K-means, two clusters of schools in NYC were found: one with low Average Math and ELA Proficiency and one with high Average Math and ELA Proficiency. Those two clusters can be distilled further by demographics and environment. Using a multiple regression model it was found that 89% of the variation in Economic Need Index is explained by school income, strong family community ties, average Math performance rating, % White, % Hispanic students, student attendance rate, and % of students who are English Language Learners. Lastly, using a PCA, components were created to be used in further analysis that were successfully organized by race and school type without multicollinearity.

Method

Before analysis could be done, the data was preprocessed for stray symbols and missing data. All of the symbols and hidden dashes , a product of Excel, were removed. Missing values were also computed across all columns. Several columns shared missing values along the same row and around 30 of these were removed.

The main source of ‘NA’ values was the School Income Estimate column. Because there were around 400 missing values in this column, removing them was not an option as this would have cut down the number of observations by 30%. As a result, the average income within each of the 32 NYC school districts was used to fill in the missing values.

K-means clustering was used to examine the relationships between schools based on School Income Estimate, Average Math Proficiency, and Average ELA Proficiency. These three variables were chosen because they seemed like broad descriptors of what data set. Average Math & ELA Proficiency gives a general understanding of test scores at each school, while School Income Estimate provides a benchmark for the economic level of the community. Using average silhouette width, the ideal k value was k = 2 (figure 2.1), though clusters k = 2 through k = 5 were tested. The best cluster was for k = 2. The two clusters are sizes 386 and 781 (figure 2.2). The sum of squares between clusters minimized at 55.7%. The means for cluster 2 are School Income Estimate = 3.23, Average Math Proficiency = 6.90, and Average ELA Proficiency = 8.20. The means for cluster 1 are School Income Estimate = 1.96, Average Math Proficiency = 5.18, and Average ELA Proficiency = 6.48. Cluster 2- “High Score Cluster”- is schools with better Average Math & ELA Proficiency Scores. There is a slightly higher School Income Estimate as well, but the relationship is not as strong. Cluster 1 - “Low Score Cluster” - are the lower scoring schools with slightly lower income. Plotting the clusters in two dimensions (as in figure 2.2) shows that most of the data (81.2%) is explained in dimension 1, with only 16.9% explained in dimension 2. The two components together explain 98.1% of the variability.

These clusters divide the schools of New York City into two factions. One of which has lower Average Math and Average ELA Proficiencies. School Income is the least determining factor when shaping the clusters. This is supported in figures 2.3-2.5. Schools with high and low income estimates have both high and low Math/ELA Proficiency. This challenges an oft held assumption that poorer schools produce less intelligent students (or students that test worse). To harken back to the purpose of the PASSNYC data set orginginaly, it would seem that School Income Estimate is not a good factor when considering which schools need the most help. Since the schools are not highly correlated by income, a scatter plot was used to see if certain school districts could be found in only one cluster. As shown in figure 2.6, almost all of the school districts have schools in both clusters. About two thirds of the schools are in the lower scoring cluster.

The next step is delve more deeply into the interrelationship of the schools within each cluster. Data exploration was used to determine which variables to focus on (figure 2.7). The relationship between Rigorous Instruction and Strong Family Community Ties seemed worth exploring in particular because it had a small increasingly linear relationship. In order to see how all of the other variables related to one another, a logarithmic regression was produced to see the odds of being put into cluster 1 or cluster 2. Essentially, what contributes to being a school with lower scores rather than a school with higher scores? Variables omitted were Average ELA Proficiency, School Income Estimate and Average Math Proficiency, because those are the initial relationships from the first cluster. Grade specific categories were also ignored due to high multicollinearity. In subsequent running of the model, all variables except those with significance per $p < 0.05$ were removed. Because logarithmic regression has a binary dependent variable, if the result is 1, then the new school would be in cluster 1. If it is 0, then it would be in cluster 2. The resulting equation was as follows: cluster(1 or 0) = 13.416508 - 0.124842*Percent Asian + 0.136492*Percent ELL - 0.012566*Percent Hispanic - 0.087024*Percent White + 0.024627*Student Attendance Rate - 0.096939*Strong Family Community Ties - 0.054941*Rigorous Instruction. The largest odds ratio (aside from the intercept) was for Percent ELL: for a one unit increase in Percent ELL, the odds of being in cluster 1 (vs cluster 2) increase by a factor of 1.14 (figure 2.8).

The goal here is the ability to group like schools together, so that a similar educational policy may raise achievement scores. Through clustering we were able to ascertain two distinct test score groupings. Through logarithmic regression, we were able to determine the variables that have the most significant impact on achieving those groupings. Now, it will be taken one step further to break up those two large clusters into smaller fractions of schools that are most alike.

The variables used in the logarithmic equation are the most important factors in these two school groupings. The clusters were clustered again based on this criteria. This time PAM (or K medoids) was used because there are some outliers. For the first cluster of schools with lower average math and ela scores, the maximum silhouette width resulted in three clusters (figure 2.9). For the second cluster of schools with higher Average Math and ELA scores, the maximum silhouette width resulted in five clusters (figure 2.10). Using figures 2.11 and 2.12 it can be ascertained what makes these particular schools so alike and in the end will have 8 clusters of like schools, based on School Income Estimate, Average Math Proficiency, and Average ELA Proficiency, Percent Asian, Percent ELL, Percent Hispanic, Percent White, Student Attendance Rate, Strong Family Community Ties, and Rigorous Instruction.

Discussion and Results

After breaking the data into two clusters of schools using K-means clustering, it is known which ones had higher average Math/ELA scores and lower average Math/ELA scores. A logarithmic regression was used to find the significant variables. Using the variables from the logarithmic regression and k-medoids on the original two clusters gives 8 subsets of school with similar School Income Estimate, Average Math Proficiency, and Average ELA Proficiency, Percent Asian, Percent ELL, Percent Hispanic, Percent White, Student Attendance Rate, Strong Family Community Ties, and Rigorous Instruction.

Out of the schools that have high scores in Average Math Proficiency and Average ELA Proficiency, three clusters were able to emerge based on Strong Family Community Ties, Rigorous Instruction, Percent ELL, Percent Hispanic, Percent White, and Student Attendance Rate. All show that as the score for Strong Family

Community Ties increases, Rigorous Instruction also increases. The three clusters are different demographically. One cluster (shown in black on figure 2.11) has the lowest percentages of ELL, Hispanic, Asian, and White students. The second cluster (shown in red on figure 2.11) had a very even distribution of demographics, but has the most White students. The last cluster (shown in green on figure 2.11) has the highest percent of Hispanic and ELL students.

Out of the schools that have low scores in Average Math Proficiency and Average ELA Proficiency, five clusters were able to emerge based on Strong Family Community Ties, Rigorous Instruction, Percent ELL, Percent Hispanic, Percent White, and Student Attendance Rate. The cluster shown in black on figure 2.12 has the greatest percent of Hispanic students. The cluster in red in figure 2.12 has the highest percent of White students. The green cluster in figure 2.12 has the highest score on Strong Family Community Ties and the lowest percent of ELL students. The dark blue cluster in figure 2.12 has the highest percent of Asian students. Finally the light blue cluster in the same figure is a grouping of outliers that have a zero for Attendance Rate. Perhaps this is a case of no data, but the schools would disperse among the other clusters and we would have 4 total.

In finality, these 8 clusters show groupings of schools with similar average test scores, similar demographics, and similar community involvement. When enacting educational policy, it's important to consider all these factors. What may work for one school may not work for another, but since these schools are somewhat similar, they may have similar problems and solutions. This analysis could be taken much farther with a time series analysis. This data set limited by the year 2016, but should we implement educational policy or methods on any of the schools, it will be necessary to track the same metrics over time to determine if it was successful or not. Hard metrics such as demographics and test scores are less subject to interpretation, yet in order to truly understand where a school needs assistance the students and staff need to be surveyed. Another huge limitation is the incomplete data itself. There are variables for “American Indian or Alaskan Native” students who scored a 4 or higher on their proficiency exams, but no variable for “Percent American Indian” or “Percent Alaskan Native”. Future work can include a time series analysis, perhaps using neural networks. Performing a hierarchical cluster is the next step to explore this data with clustering to try building the model incrementally.

Appendix 2

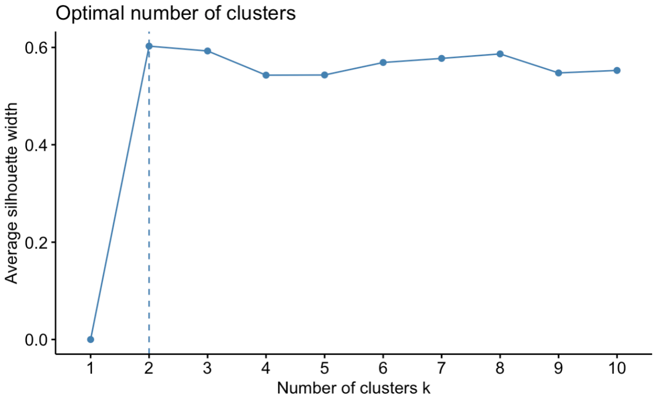


Figure 2.1: optimal number of clusters plot

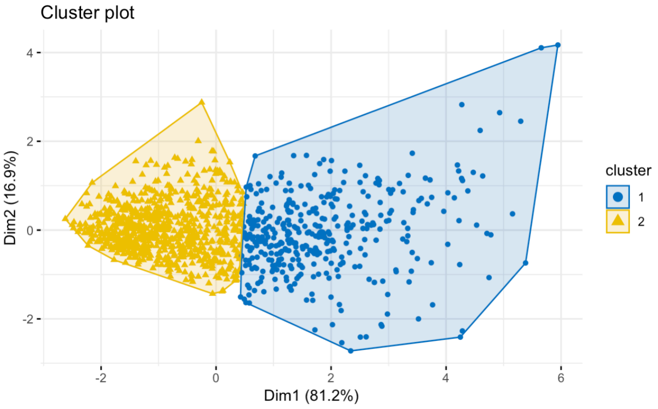


Figure 2.2: k = 2 cluster plot

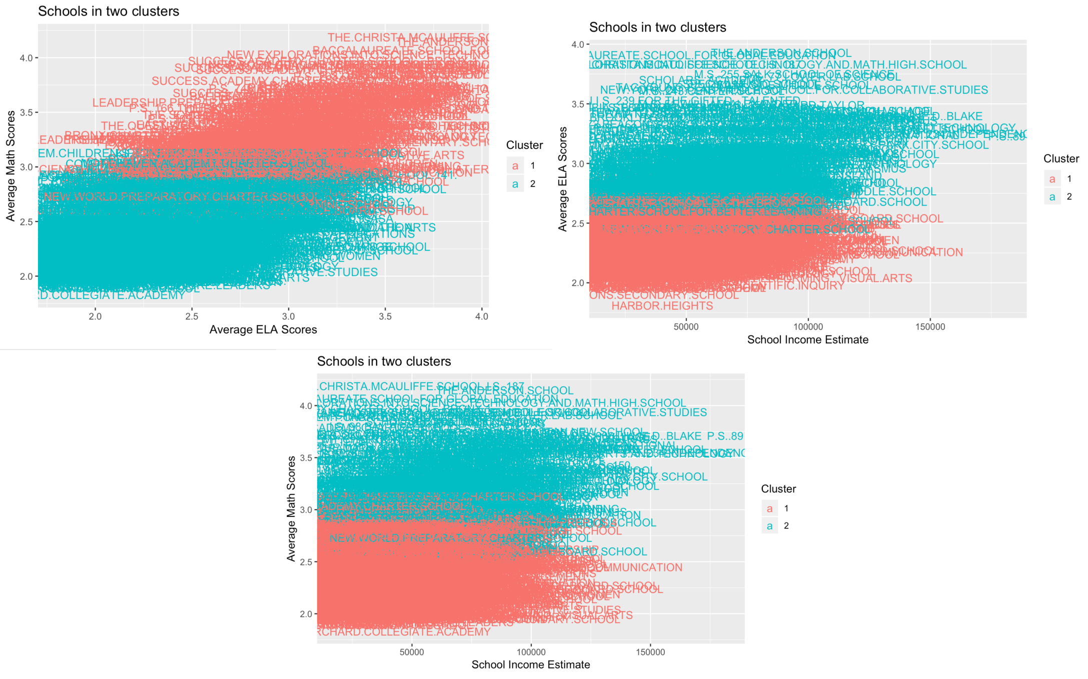


Figure 2.3 - 2.5: Plots of the variables used in the cluster analysis with respect to cluster number

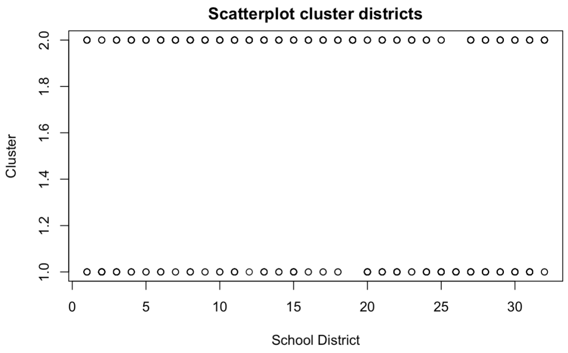


Figure 2.6: Cluster number VS school district

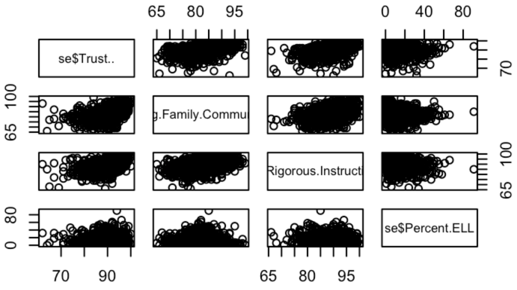
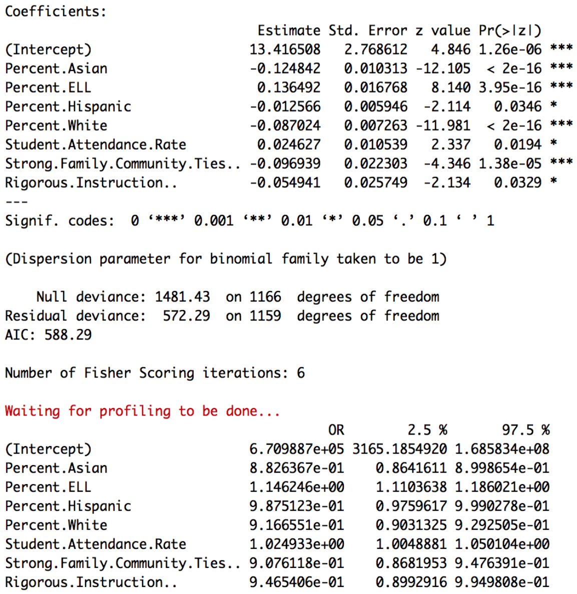


Figure 2.7: Scatterplot Matrix (unclustered data)



```
Coefficients:
                                Estimate Std. Error z value Pr(>|z|)    
(Intercept)                    13.416508   2.768612   4.846 1.26e-06 ***
Percent.Asian                  -0.124842   0.010313 -12.105  < 2e-16 ***
Percent.ELL                     0.136492   0.016768   8.140 3.95e-16 ***
Percent.Hispanic               -0.012566   0.005946  -2.114   0.0346 *  
Percent.White                  -0.087024   0.007263 -11.981  < 2e-16 ***
Student.Attendance.Rate         0.024627   0.010539   2.337   0.0194 *  
Strong.Family.Community.Ties.. -0.096939   0.022303  -4.346 1.38e-05 ***
Rigorous.Instruction..         -0.054941   0.025749  -2.134   0.0329 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1481.43  on 1166  degrees of freedom
Residual deviance:  572.29  on 1159  degrees of freedom
AIC: 588.29

Number of Fisher Scoring iterations: 6

Waiting for profiling to be done...
                                         OR        2.5 %       97.5 %
(Intercept)                    6.709887e+05 3165.1854920 1.685834e+08
Percent.Asian                  8.826367e-01    0.8641611 8.998654e-01
Percent.ELL                    1.146246e+00    1.1103638 1.186021e+00
Percent.Hispanic               9.875123e-01    0.9759617 9.990278e-01
Percent.White                  9.166551e-01    0.9031325 9.292505e-01
Student.Attendance.Rate        1.024933e+00    1.0048881 1.050104e+00
Strong.Family.Community.Ties.. 9.076118e-01    0.8681953 9.476391e-01
Rigorous.Instruction..         9.465406e-01    0.8992916 9.949808e-01
```

Figure 2.8: Summary of LogFit

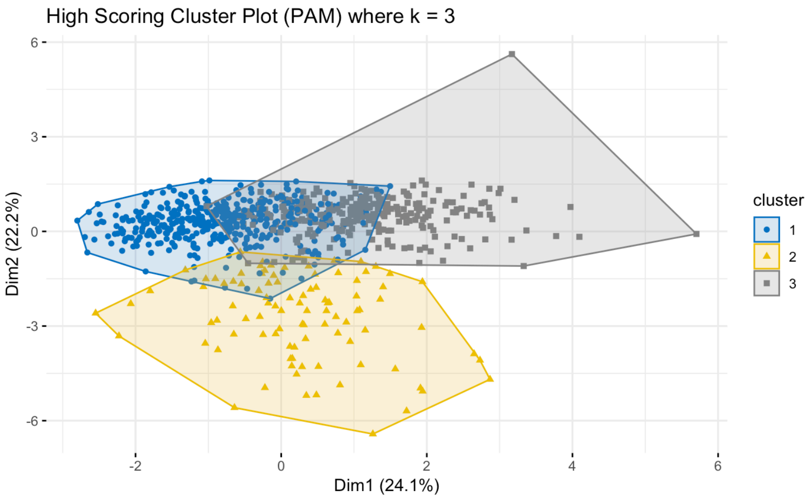


Figure 2.9: high math/ela cluster, partitioned around medoids

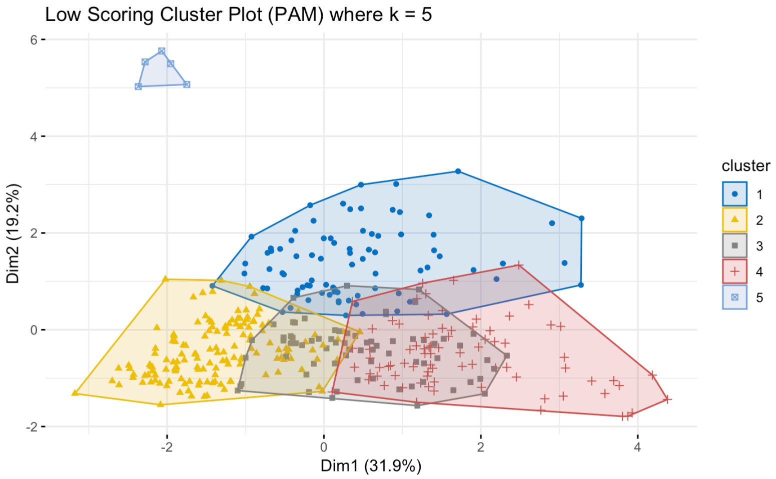


Figure 2.10: low math/ela cluster, partitioned around medoids

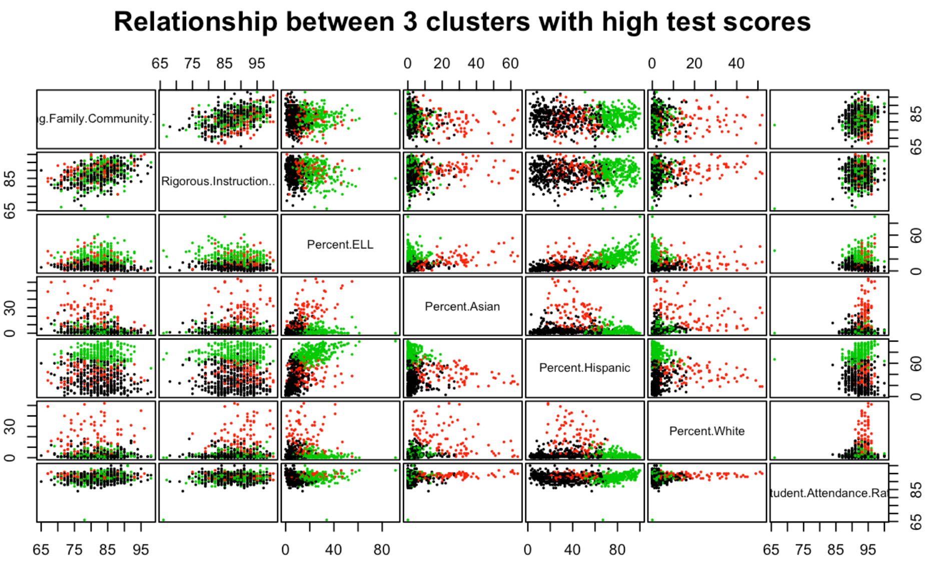


Figure 2.11: high math/ela cluster chosen variables plotted against one another, k = 3

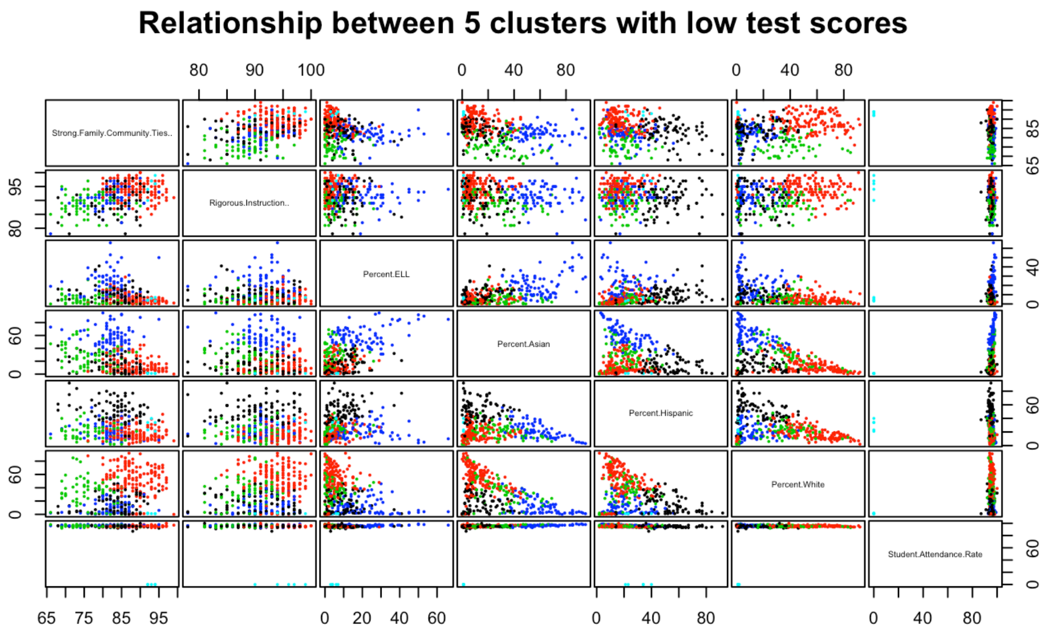


Figure 2.12: low math/ela cluster chosen variables plotted against one another, k = 5