PASSNYC Technical Report

Breanna Atwood, Will Chirciu, Amy Edwards, Jeffrey Moy, Julie Simms

# 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.

# Introduction

The NYC Schools dataset was obtained from Kaggle, which contains 1,272 schools. With over 158 variables, the primary focus was to figure out what which variables were categorical and continuous. After exploring the data, some of the continuous data had symbols (ex. $ for school spending and % for % of ethnicity), so data cleaning was performed in order to to utilize those variables. With a multitude of continuous and binary variables that could be used for the dependent variable, multiple techniques were selected to perform to explore hidden insights within the dataset. Those techniques include: Cluster Analysis, Canonical Correlation Analysis (CCA), Model Building and Multiple Regression, Principal Component Analysis (PCA), and Linear Discriminant Analysis (LDA). With these techniques, an explanation is compiled on how the technique is applied and the results that were found.

# Literature Review

Several researchers have explored the potential causes of the widening graduation gap in inner city high schools in New York. As overall graduation rates continue to climb, it has been extremely important to identify what factors are causing city schools to lag behind. The Office of the New York City Comptroller, recently released a study which quantified this gap, concluding that “while the highest performing quantile saw their graduation rates jump from 93% to 97%, those in the lowest quantile experienced an 11 drop, from 61% to 50%”(“Diploma Disparities”, 2016). These results reveal the need to quantify the challenges and needs of students in low performing schools, as presented by PASSNYC. Based on previous studies, PASSNYC has identified that data on English Language Learners, Students with Disabilities, Students on Free/Reduced Lunch, and Students with Temporary Housing, have been good indicators of the types of schools facing these challenges. Based on the analysis released by the Comptroller office, schools with the highest Economic Needs Index tend to have the lowest graduation rate, and vice versa. A lot of variables presented by the PASSNYC data contribute to factors that drive the Economic Needs Index. Studies show that “students in the lowest quantile of the Economic Need Index had a 51% graduation rate, while those in the highest quantile had a 96% graduation rate” (“Diploma Disparities”, 2016). Recent studies have also explored the disparities of graduation rates by race. In 2018 results, the reported graduation rates for Asian student is 88%, White students 83%, Black students 70%, and Hispanic students 68% (Colangelo, 2018). As challenges continue to face these low performing schools, the importance of actionable analysis surrounding these various demographics and school characteristics, such as those presented by PASSNYC, will continue to be the highest priority in determining what is most significantly driving these results.

# Methods

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.

Canonical Correlation analysis is a method that is used to explore the relationships between two sets of variables. It is often a curiosity how the demographics of an inner-city school relate to its performance and environment. Things like income and ethnic distribution might have a thing or two to say about test scores or whether the school reflects a supportive, collaborative community. With the School Explorer dataset, it was possible to do just this by analyzing canonical correlations between School Income, Economic Need index, %Asian/Black/Hispanic/White, and community school identity with performance metrics like Supportive Environment Rating, Collaborative Teachers Rating, Rigorous Instruction Rating, and average test scores.

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 runnings 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.

Linear Discriminant Analysis (LDA) is used to classify items, where there are two or more categories, by using a linear combination of related variables. The Economic Need Index, Rigorous Instruction Percent, Strong Family-Community Ties %, Average ELA Proficiency, and Average Math Proficiency, was used to predict the Student Achievement Rating.   
 According to the dataset information, the Student Achievement Rating is the weighted average score plus the closing the achievement gap score. Being able to predict the Achievement Rating would allow more insight on ways to increase the likelihood that students would be successful, especially in schools with a high economic need index. The Linear Discriminant Analysis will show whether or not different variables are strong enough to separate the Student Achievement categories. The findings could be applicable to several schools and would overall increase the success rate across the board.   
 The NYC School Explorer dataset provides an opportunity to explore how various demographics, and teacher/student attributes affect the overall Economic Need Index (the higher the index, the higher the need), and potentially use these predictors to drive change in schools and communities, and potentially increase graduation rates. Therefore, a Multiple Regression Model was applicable to this dataset as to explore what variables were significant in predicting Economic Needs Index. In order to build a model, several variables were considered during the data exploration phase including: community school indicator, school income, % of students who are English Language Learners, race, school attendance rate, % of students chronically absent, rigorous instruction rating, collaborative teacher rating, supportive environment rating, effective school leadership rating, strong family-community ties rating, trust rating, average English Language Arts performance rating, and average Math performance rating. All N/A values were removed from this dataset, except for N/A values in the school income variable. School income N/A values were replaced with applicable district averages.

In the NYC Schools dataset, there were over 158 variables. In particular, the interest was in the amount of students in various grades and ethnicity that scored a 4 on a math or ela test. With over 84 variables, it was worth checking for multicollinearity, so a correlation plot was done to calculate the total amount of variables that each variable was highly correlated with by testing with a p-value less than .01, which suggest that there is less than a 1% chance that the variables are insignificant with each other (beta coefficient =0). In doing so, the amount of variables that are correlated with other variables in the matrix was extreme. There were variables that were correlated with roughly 30 other variables. Thus, PCA was used to remove multicollinearity from the independent variables and see if the variables can be grouped into meaningful components.

PCA has two applications to it: Dimensionality Reduction and Latent Variable Discovery. Dimensionality Reduction is reducing the amount of features/variables within the dataset to remove complexity and keep only the variables that can explain the most amount of variance within the data. Latent Variable Discovery is finding the variable groupings by combining the correlated variables into components that are distinct from other components. Thus, multicollinearity is removed and all variables are kept within the component for more explanatory power.

Before PCA is performed on the data, factorability needs to be done. Three tests to check for factorability are Kaiser-Meyer-Olkin (KMO), Bartlett's Test of Sphericity, and Cronbach's Alpha. KMO tests for sampling adequacy to verify that the dataset is large enough. The test was performed and the value was .81, which is greater than .70. Thus, the data meets the sampling adequacy. Bartlett’s Test of Sphericity tests if there is enough shared variance within the dataset. The test results in a p-value less than .05, suggesting there is sufficient variance in the data to run a PCA. Lastly, Cronbach’s Alpha is used to test reliability, to see if the variables and the groupings make sense (in theory and conceptually) to be used for later analysis. The test results with a reliability value of .9, suggesting the groupings and variables make sense to run a PCA.

Now that the dataset has been tested for factorability and the assumptions of linear regression, PCA can be performed now.. The first thing needed to be done is determine the amount of components to be used in the model. There are three tests used to determine this: the amount of components in a scree plot that have an eigenvalue greater than 1, the amount of components at the knee of the scree plot that levels out, and by looking at the summary information and choose the number of components where at least 60-80% of the variance is explained. There is no definitive test to determine the amount of components. From the scree plot of all the components an eigenvalue greater than 1 (Figure 5.2), it appears that the first 10 components have an eigenvalue greater than 1. The amount of components at the scree plot (Figure 5.1), looks as if it levels off at 9 components. Lastly, let's check the % of variance of all the components (Figure 5.3), it looks that 9 components explains at 62.7% of the variance, which just reaches the cutoff of 60%-80%. From all 3 tests, it seems that 6 components is sufficient as a starting point to test towards a final model and verify if the groupings are interpretable.

Lastly, when performing the PCA, a varimax rotation will be used as it is assumed that the components are independent of each other. A correlation test against the components will be used to verify if this assumption holds true.

# Discussion and Results

### Canonical Correlation Analysis

The canonical variate pairs in a CCA analysis provide the ability to analyze the relationship between groups of independent variables and groups of dependent variables. The independent variables in this case were the various demographics of the school and the dependent variables were the various academic and environmental performance metrics. Figure 1.1 reveals that there were 5 significant canonical variate pairs obtained from the CCA. Because there were no significant structural loadings in the 5th pair, it was excluded from the analysis. Figure 1.2 depicts all of the Canonical Correlations corresponding to the 4 variate pairs examined.

The first canonical variate pair had the strongest relationship with a Canonical Correlation of 0.86. The helio-plot in Figure 1.3 is a visual representation of the structural loadings of the variables on the first canonical variate. Looking at this and the actual loadings in Figure 1.7 with a cutoff of 0.45, variables were selected to represent each variate. The variate for Demographics is best represented by all of the independent variables except Community School. The Quality variate is best represented by Average ELA/Math Proficiency Scores. The standardized coefficients were then analyzed to determine the relationship and impact the Demographic variate had on the Quality variate . The standardized coefficients for CV1 in Figures 1.9 and 1.10 show that the economic need index had the largest impact on test scores with a value of -0.87. Because the coefficients for test scores were positive, it was concluded that the economic need index has a strong negative relationship with average math/ELA proficiency scores. This would suggest that schools with a higher economic need index, or higher economic need, have fewer resources available to provide their students to better prepare them for these tests. Additionally, %Black/Hispanic/White also had negative relationships with test scores, with %black having the largest impact. This would suggest that schools with a larger population of black students tend to have lower test scores in the context of the NYC school system.

The second canonical variate pair had a Canonical Correlation of 0.60. The helio-plot in Figure 1.4 is a visual representation of the structural loadings of the variables on the second canonical variate. The Demographics variate is best represented by %Asian and %Black. The Quality variate is best represented by Strong Family Community Ties. The standardized coefficients for CV2 show %Asian and %Black have positive coefficients with similar values within Demographics . Strong family community ties has a negative coefficient within Quality. Therefore, it was concluded that %Asian and %black are negatively related with Strong family community ties. This would suggest that schools with a larger number of Asian and Black students have less developed relationships with families, businesses, and community-based organizations.

The third canonical variate pair had a Canonical Correlation of 0.29. The helio-plot in Figure 1.5 is a visual representation of the structural loadings of the variables on the third canonical variate. The Demographics variate is best represented by %Black. The Quality variate is best represented by Supportive Environment and Strong Family Community Ties. The coefficients for CV3 in Figures 1.9 and 1.10 show %Black has a negative coefficient within the Demographics variate.. Strong family community ties and Supportive Environment have positive coefficients within the Quality variate. Therefore, it was concluded that %black is negatively related with strong family community ties and Supportive Environment. This would suggest that schools with a higher percentage of black students have less developed relationships with families, businesses, and community-based organizations and contain students who do not feel safe, supported, and challenged by their teacher and peers.

The fourth canonical variate pair had a Canonical Correlation of 0.17. The helio-plot in Figure 1.6 is a visual representation of the structural loadings of the variables on the fourth canonical variate. The variate for Demographics is best represented by Community School. The Quality variate is best represented Rigorous Instruction, Collaborative Teachers, and Strong Family Community Ties. The coefficients for CV4 in Figures 1.9 and 1.10 show Community School has a negative coefficient within Demographics. Rigorous Instruction, Collaborative Teachers, and Strong Family Community Ties are all positive within Quality. I conclude that Community School is negatively related with the variables in Quality for the fourth CV pair. This would suggest that community schools 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. This was not intuitive, as the whole purpose of a community school is to increase community involvement and provide additional resources to students and faculty.

### Clustering

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.

### Linear Discriminant Analysis

Within the Linear Discriminant Analysis, the data was separated into training and validation sets in a 70/30 split. The data was then plotted to look at the relationships between variables. The numeric variables were plotted (Figure 3.1) to show the relationships between the pairs of variables displaying correlation, histogram and scatter plots. The Economic Need Index was highly correlated with Average ELA Proficiency and Average Math Proficiency. In addition, Average ELA Proficiency and Average Math Proficiency were highly correlated with each other. The histogram shows that Strong Family-Community Ties % had a normal distribution. The other variables show distribution that were slightly skewed left or right. The scatterplots showed a linear relationship with the highly correlated variables. However, the variables correlated with Economic Need Index showed a negative linear relationship.   
 The Linear Discriminant Analysis was used to find a linear combination of the variables that will give the best separation between the achievement ratings. The Prior Probabilities of Groups shows 25% of the data belongs to Approaching Target, 54% to Meeting Target, 20% to Exceeding Target, and less than zero to Not Meeting Target. The group means for each rating and variable showed us the averages. The Coefficients of Linear Discriminants gave the coefficients for each of the 5 variables used in the discriminant function. And the Proportion of Trace under the first discriminant function, gives us 98.57% of the information. The linear discriminants are plotted (Figure 3.2) showing there is no clear separation between the categories.   
 The first discriminant was used to predict the rating with the training data. After plotting the results in a histogram (Figure 3.3), no clear separation between the rating is visible. The variables did not perform well in separating the categories. A bi-plot was created (Figure 3.4) which shows LD1 on the x-axis which gives 98.57% of the information. LD2 is shown on the y-axis which only provides 1.07% of the information. There is too much overlap between all the groups. Exceeding Target, shown in green, appears to have a little bit of separation if the value of the equation for linear discriminant 1 is approximately -5.  
 The Confusion Matrix (Figure 3.5) determines the accuracy of the model. It also shows how items are properly and improperly categorized. In column one, Approaching Target, 87 were categorized correctly and 120 were miscategorized as Meeting Target. The sum of the numbers on the diagonal divided by the total sum of the table gives you the accuracy percentage. The diagonal sum is 525 and the total sum is 816. The model is 64% accurate. The validation data was applied to the model which also resulted in a 64% accuracy rate.  
 In summary, the variables selected from the New York City School Explorer dataset to predict the School Achievement Rating don’t have enough information in them to separate the categories. Different combinations of other variables from the data set should be looked at or maybe more research needs to be done.

### Multiple Linear Regression

The null hypothesis of the multiple regression model was that no variable would have a significant effect on the Economic Needs Index. A correlation plot was created to explore the collinearity between variables (Figure 4.1). Based on the correlation plot, there did not appear to be any alarming problems with multicollinearity amongst the independent variables. A full model was ran, and the residual plots were analyzed to validate assumptions. Based on these plots there did not seem to be any obvious pattern, and the points looked to be scattered around the zero line (Figure 4.2). Influential points were detected (Figure 4.3), along with outliers (Figure 4.4) to identify any common observations that were significantly skewing the data set. Based on the results of cooks, and the potential outliers, observations 32, 44, and 1001 were removed from the dataset.  
 In order to improve the model, Stepwise was used as an automatic selection method to determine significant predictors. Based on the final stepwise selection step, using the training data set, the following predictors were determined to be significant in determining Economic Needs Index: school income, average English Language Arts performance rating, % White, % Hispanic, % of students chronically absent, % of students who are English Language Learners, student attendance rate, % Black or Hispanic, collaborative teacher rating, average Math performance rating, and strong family community ties rating (Figure 4.5). Therefore, the null hypothesis was rejected and it was determined there was at least one variable that was a significant predictor of Economic Needs Index.  
 Analyzing the VIF scores of these variables revealed there were a few variables within the model that may have been skewing the model results due to multicollinearity (Figure 4.6). Considering average English Language Arts performance rating had the highest VIF value, it was removed from the model first. Amongst the removed variables were % of students who are Black or Hispanic, and % of students chronically absent. % of student who are Black or Hispanic is being captured in the remaining variable % Hispanic. Similarly % of students chronically absent is highly correlated with the remaining variable Student Attendance rate. After removing variables with high VIFs, it was concluded that the remaining variables did not show indications of any multicollinearity issues within the model (Figure 4.7).  
 Based on p-value <.05, the significant predictors in predicting Economic Needs Index are: school income, strong family community ties, average Math performance rating, % White, % Hispanic students, student attendance rate, and % of students who are English Language Learners (Figure 4.8). Based on the F test statistic, GOF is concluded to be ok. The R-Squared and Adjusted R-Squared value was 0.89, indicating that 89% of the variation in Economic Need Index can be explained by its relationship with the remaining independent variables. The residuals were revisited at this point to ensure all assumptions were met, with no concerns (Figure 4.9).   
 The remaining 20% of the dataset was used for testing the final model. A correlation between actuals and predicted values of 0.89 was calculated, indicating that actuals and predicted values have similar directional movement. Min/max accuracy and mean absolute percentage error were also calculated to validate the model, 92% and 0.08 respectively. Considering a relatively high min/max accuracy, and low mean absolute percentage error, it was concluded that the final model was significant in predicting Economic Needs Index.

### Principal Component Analysis

A PCA model was computed with 6 components using a varimax rotation, which assumes that the components are independent from each other. This can be verified by running a correlation between the components (FIgure 5.4) to ensure there is no multicollinearity. The reason why a rotation was performed is because it makes the loadings easier to interpret and re-distributes the variance. In doing so, 6 components did not produce great results, as most of the groupings for grade, ethnicity, and test were overlapped with others. By increasing the amount of components to 9, the groupings of the variables to their respective component made much more sense. The interpretations of each component are listed in Figure 5.5 to view. Each component for the most part was broken up by race and grade(elementary school and middle school) except for Black or African American students. Also, looking at the loadings, a high value suggests a high influence towards the component’s overall score and a vast amount of loadings had a value in the 70%s.

Lastly, not all variables were placed into components due to low loadings/beta coefficients. Due to this, the PCA was performed again on the dataset excluding those variables. This time, all variables were mapped to the same component groupings as before, just the loadings/beta coefficients changed a bit. But no variable was left ungrouped this time. For a visualization, Figure 5.6 is a bar plot showing the most influential variable per component.  
 Overall, PCA did a great job creating components by race and school type and multicollinearity was removed within the independent variables by combining the correlated variables into their respective components. With this information, the components can be used for regression analysis, which includes all of the explanatory power of all the variables in the components without breaking the violation of multicollinearity.

# Conclusion

The analyses used to explore interrelationships between variables in the NYC School Explorer dataset gave insight into the meaning of the data. PASSNYC sought to target and support low performing schools so that more students could prepare for and apply to specialized high schools. Significant predictors such as family and community relationships, math scores, and racial demographics proved to be strong predictors of the school’s economic need.   
 Simultaneously, the data showed how racial demographics and a low economic need play a role in higher test scores. This sheds light on the huge disparity between higher and lower economic needs schools. Examining and manipulating the data to determine which variables were true predictors of the need, gives the high economic need schools the chance to compete for spots in the specialty high school.  
 In many instances, the data was subject to interpretation because the number of teachers, students and family/neighborhood not represented. Another hindrance, was the limited one-year timeframe from which the data was gathered. Overall, there was not enough information in the dataset to come to consistent conclusions about the school’s needs.

# Appendix

Appendix 1

Figure 1.1 : CCA Significance Test

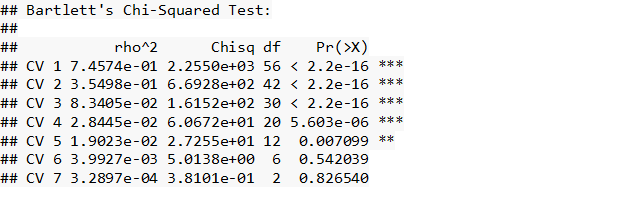


Figure 1.2 : Canonical Correlations

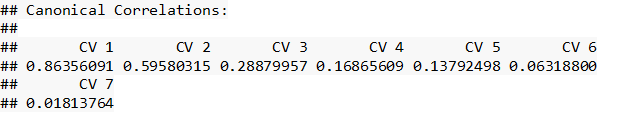


Figure 1.3 : Helio Plot for First Canonical Variate Pair

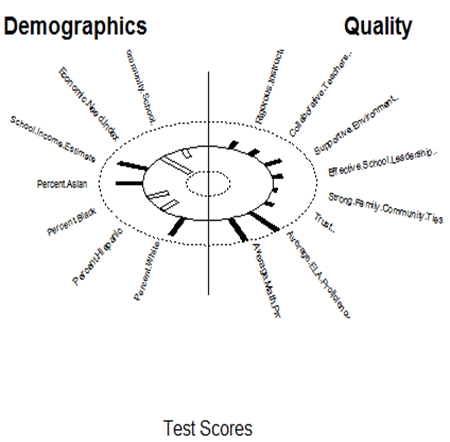


Figure 1.4 : Helio Plot for Second Canonical Variate Pair

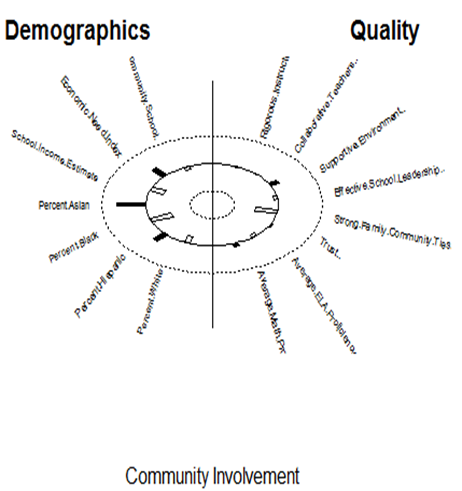


Figure 1.5 : Helio Plot for Third Canonical Variate Pair

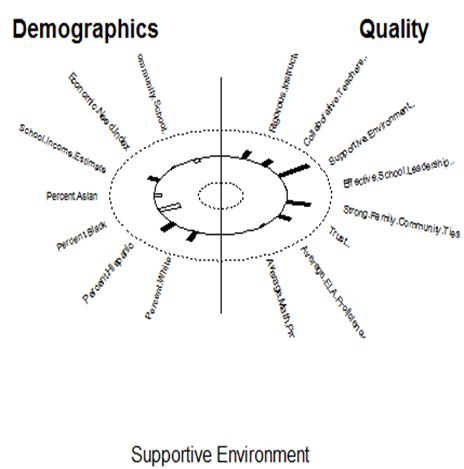


Figure 1.6 : Helio Plot for Fourth Canonical Variate Pair

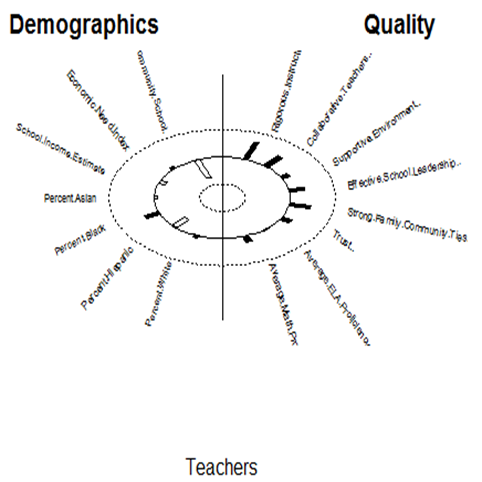


Figure 1.7 : Structural Loadings of Demographic Vars on Demographic Variates

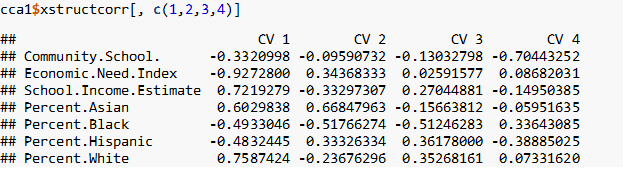


Figure 1.8 : Structural Loadings of Quality Vars on Quality Variates

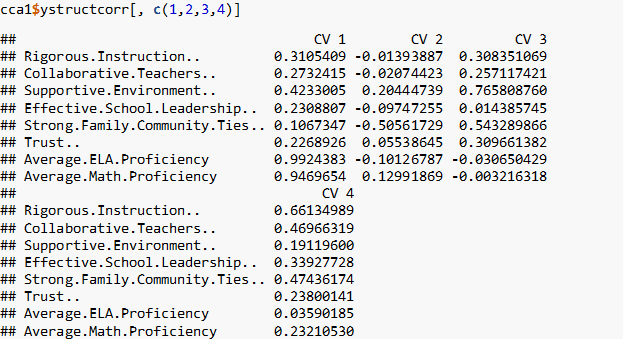


Figure 1.9 : Canonical Standardized Coefficients of Demographic Variates

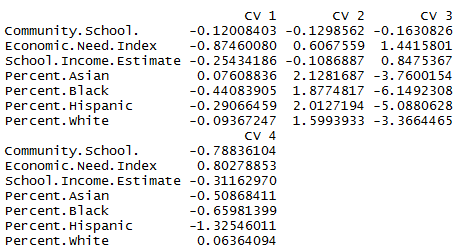


Figure 1.10 : Canonical Standardized Coefficients of Quality Variates

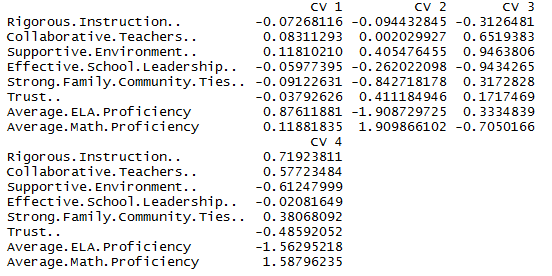


Figure 1.11 : Canonical Redundancies for Demographic Variables

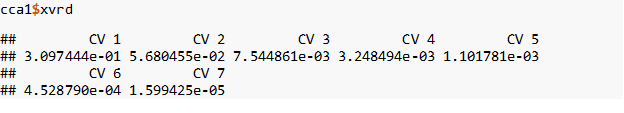
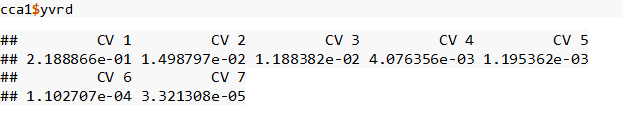


Figure 1.12 : Canonical Redundancies for Quality Variables



## 

## Appendix 2

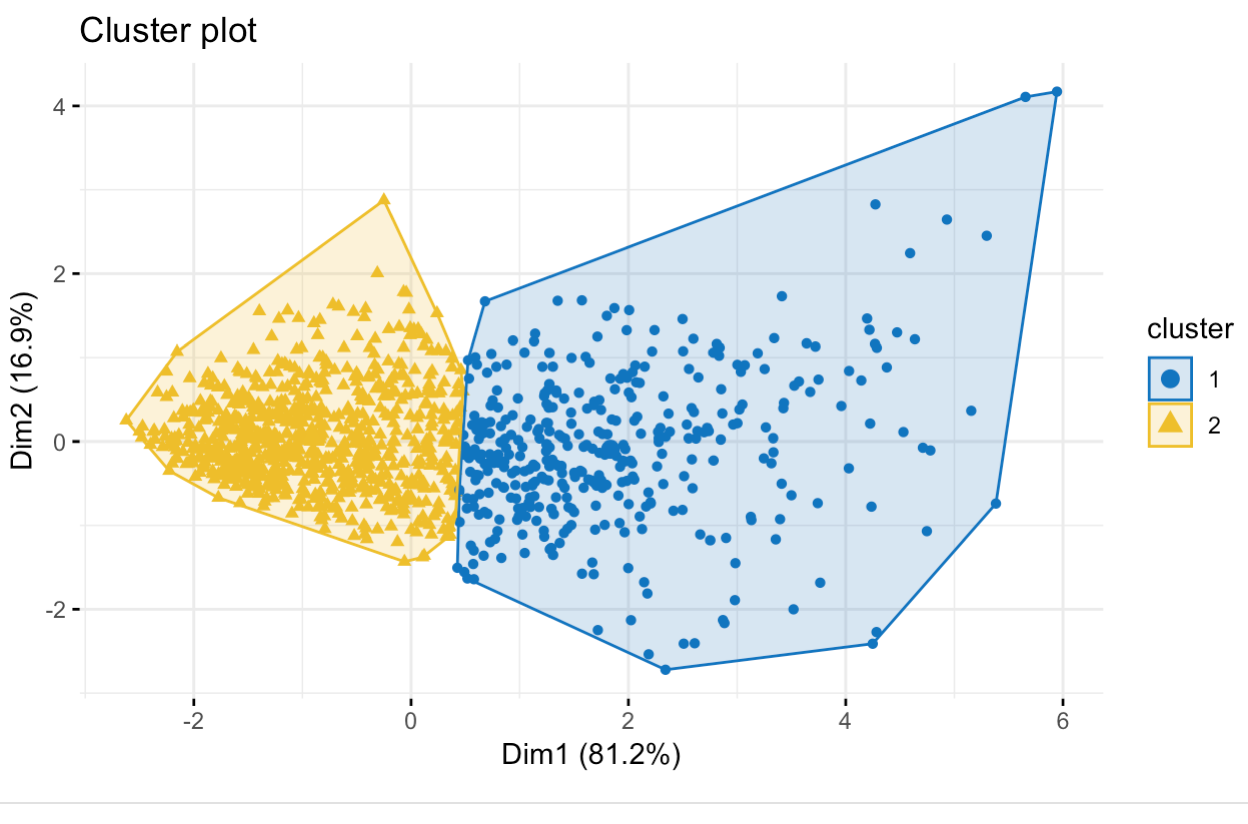
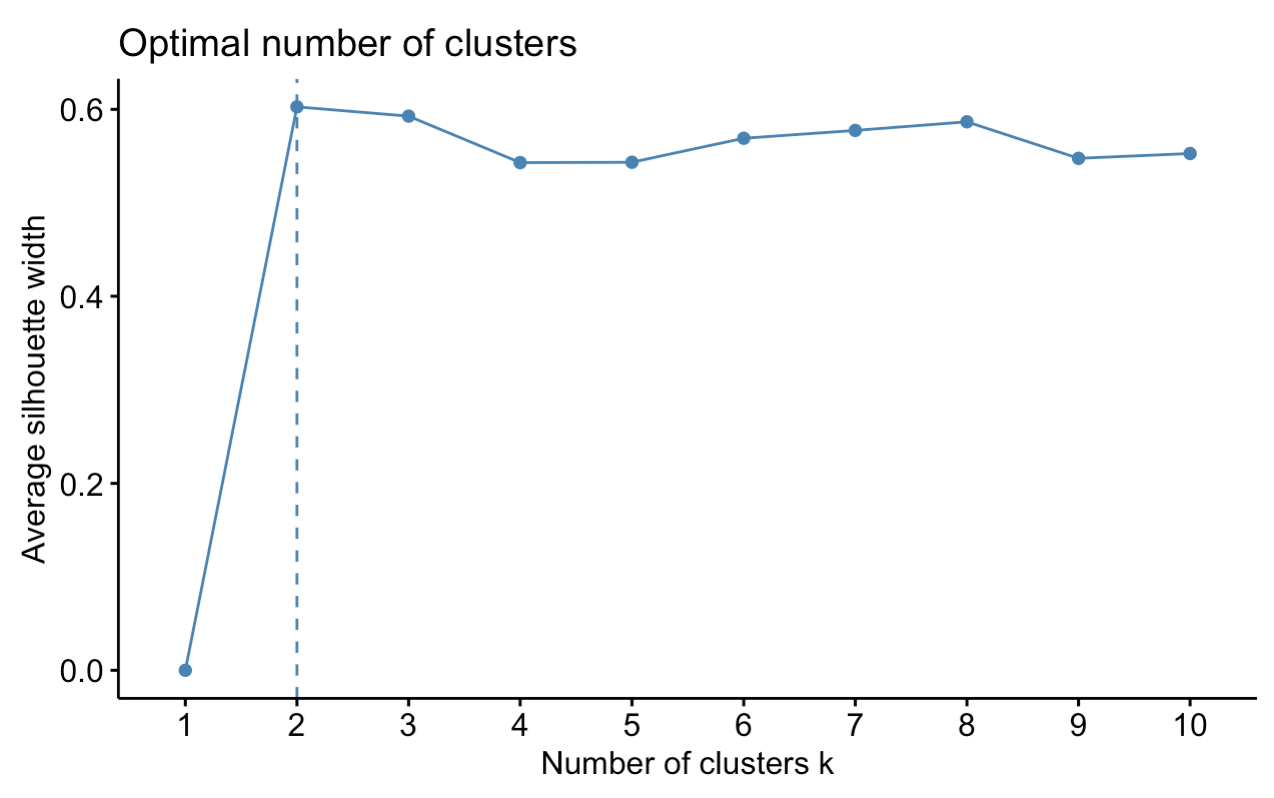


Figure 2.1: optimal number of clusters plot Figure 2.2: k = 2 cluster plot

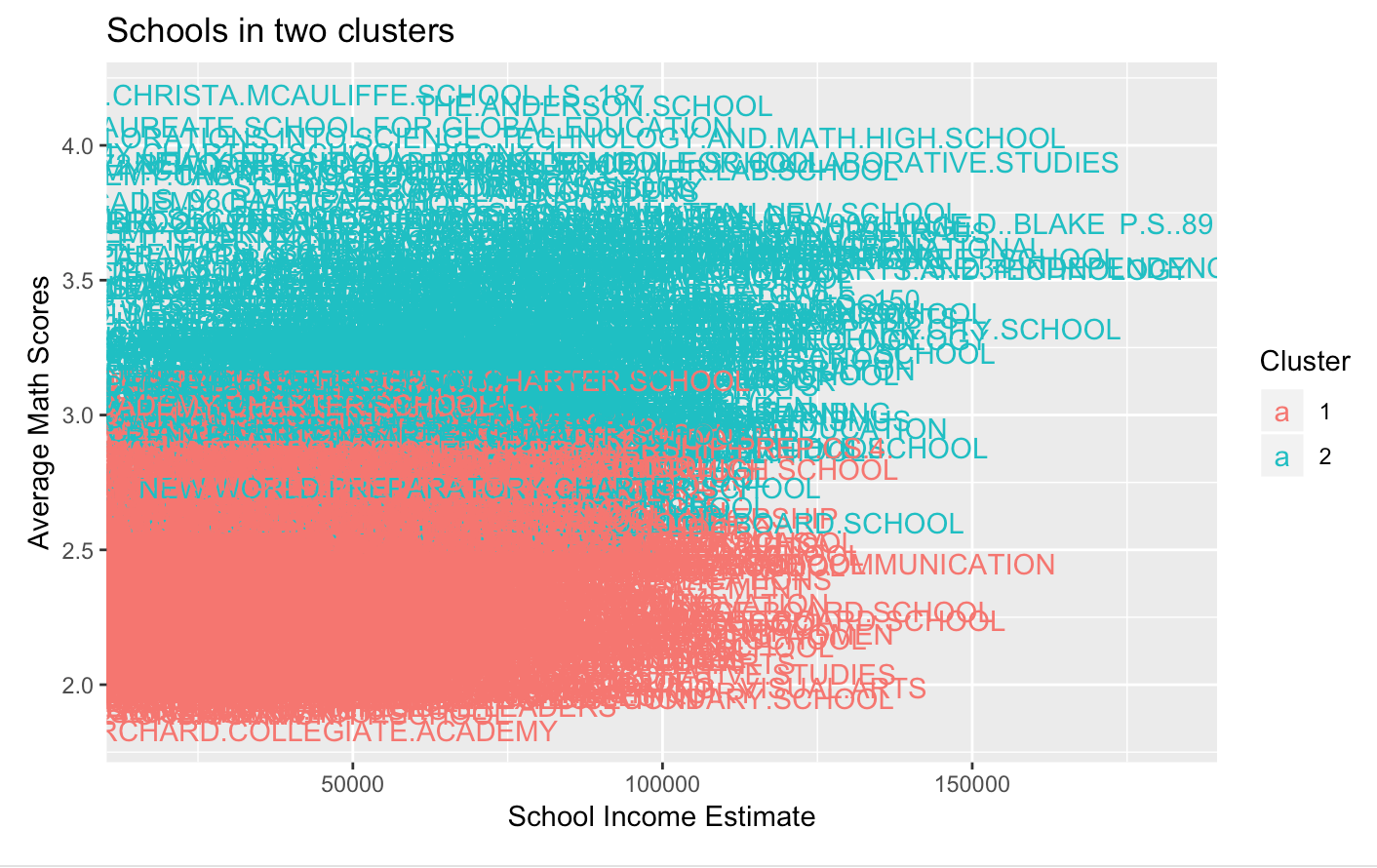
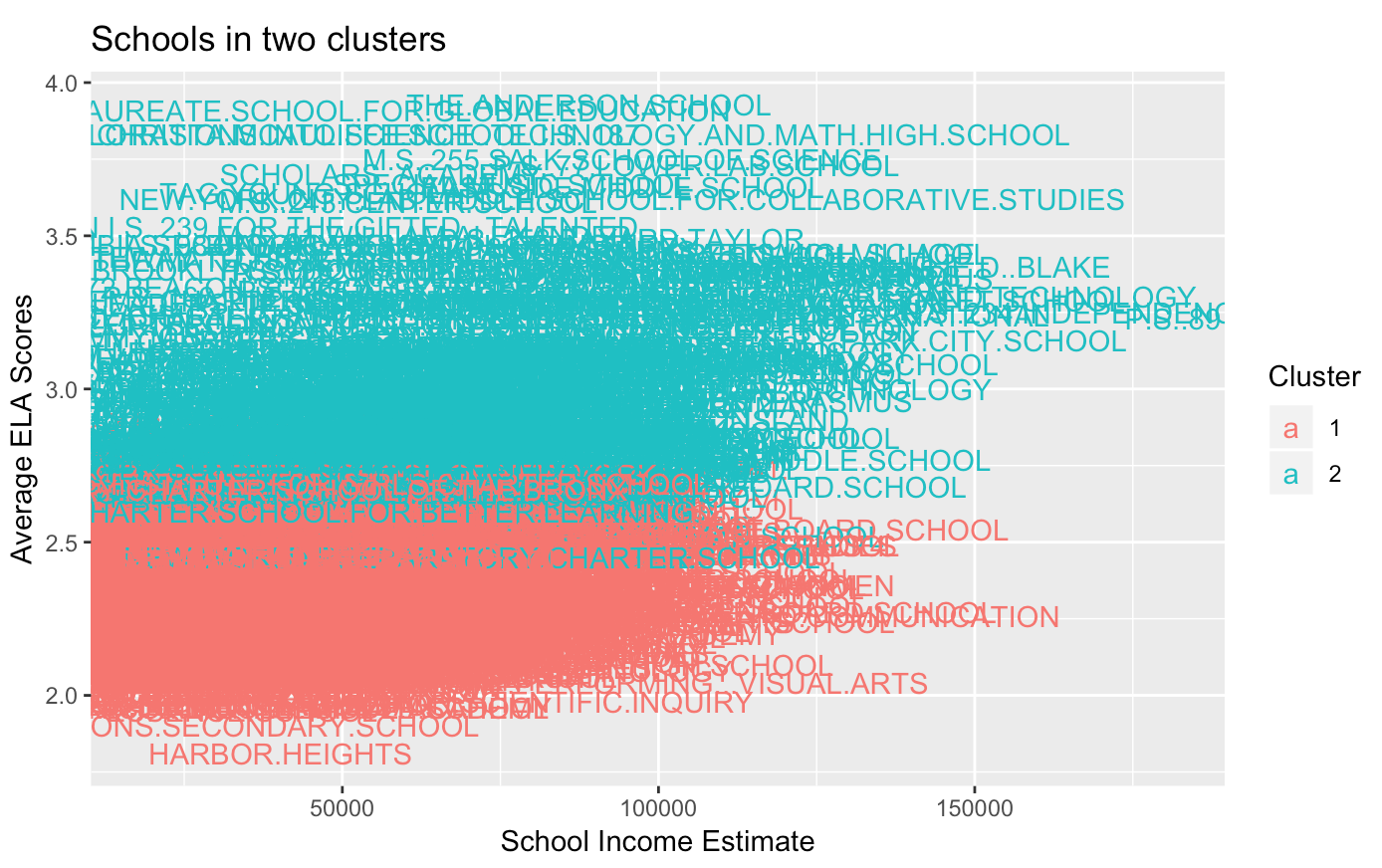
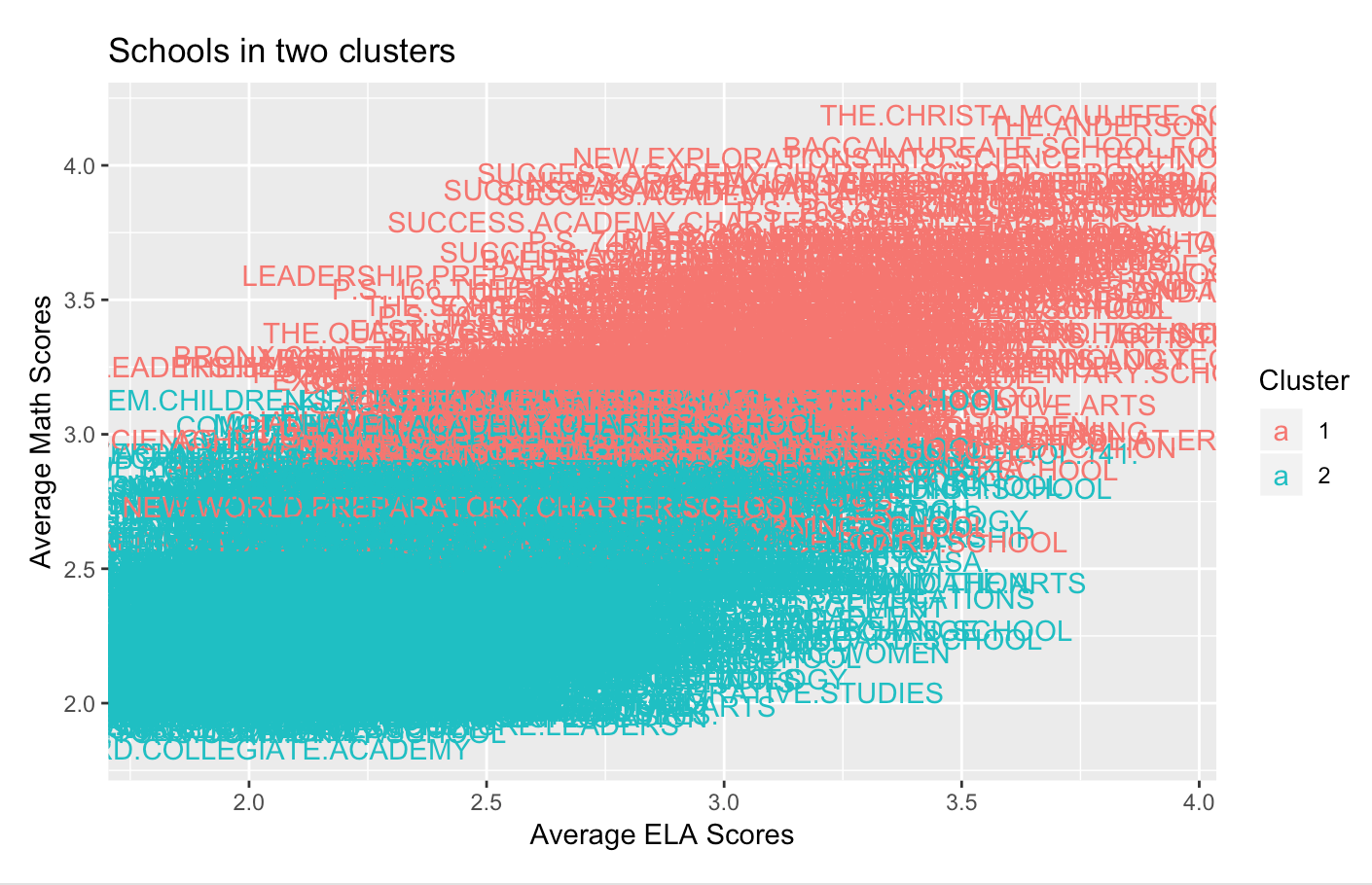


Figure 2.3 - 2.5: Plots of the variables used in the cluster analysis with respect to cluster number

## 

Figure 2.6: Cluster number VS school district Figure 2.7: Scatterplot Matrix (unclustered data)

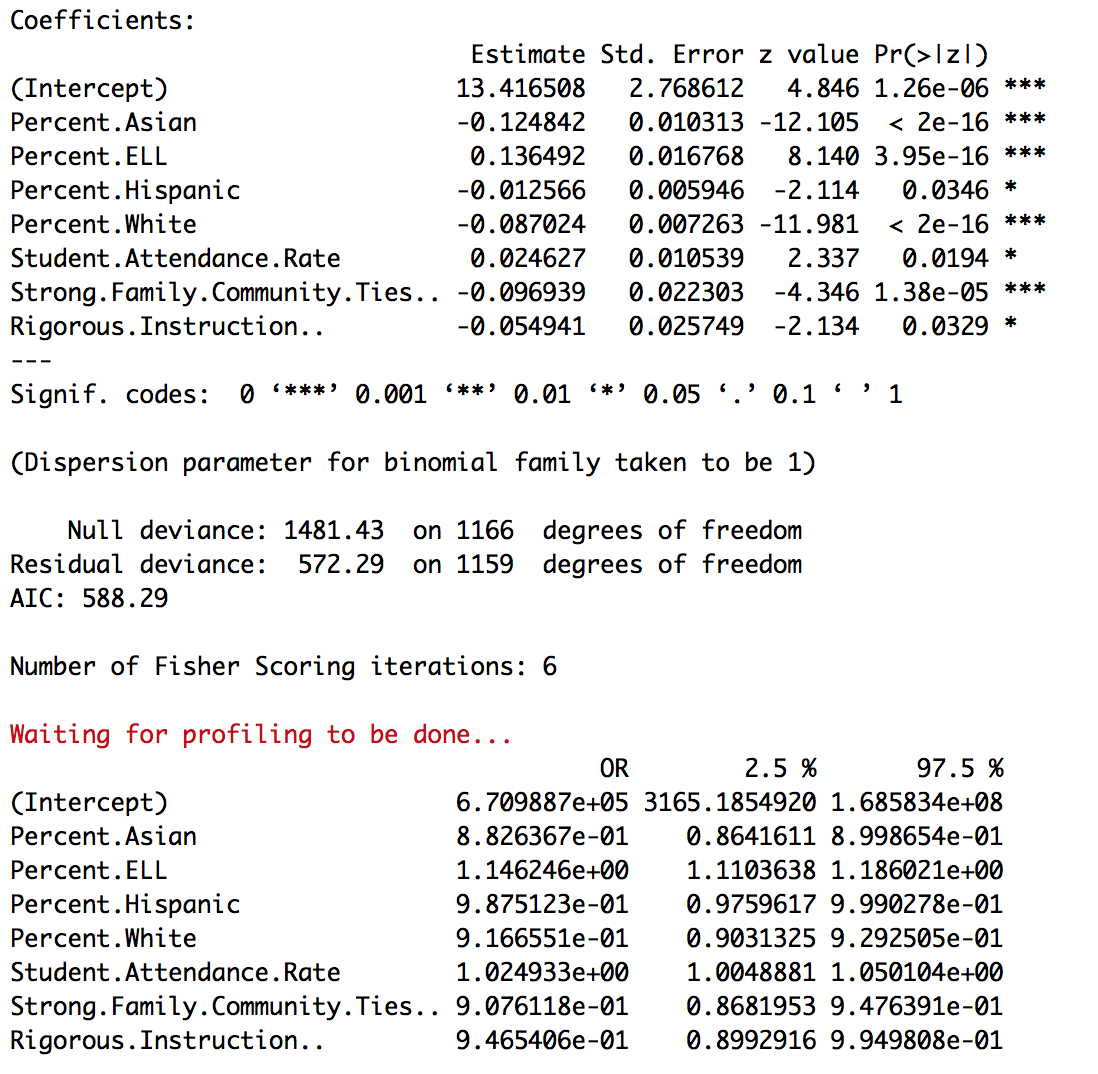


Figure 2.8: Summary of LogFit

## 

Figure 2.9: high math/ela cluster, partitioned around medoids Figure 2.10: low math/ela cluster, partitioned around medoids

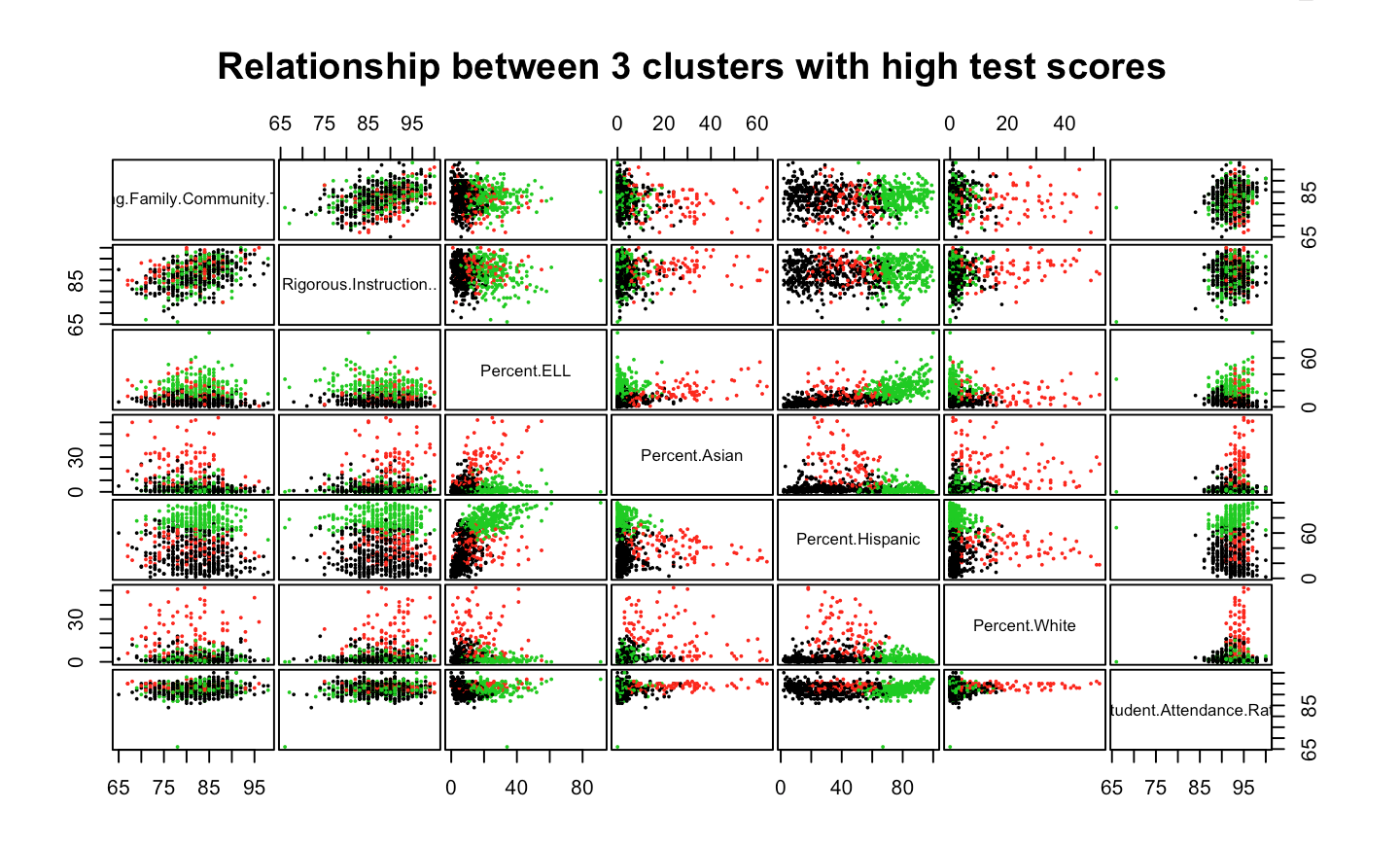
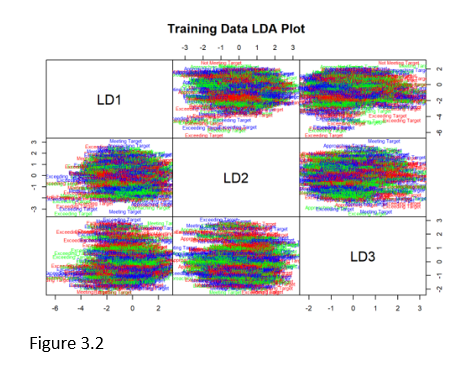
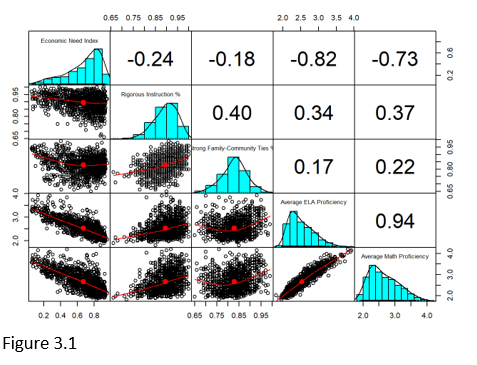


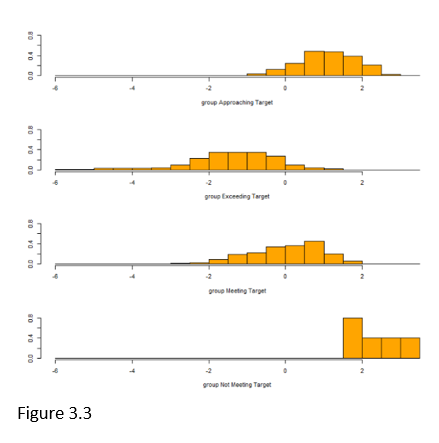
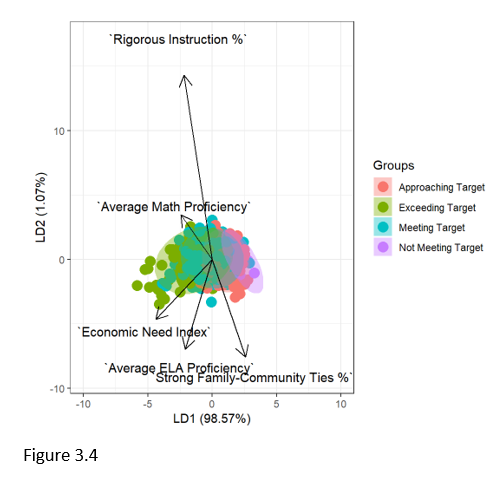
Figure 2.11: high math/ela cluster chosen variables plotted against one another, k = 3

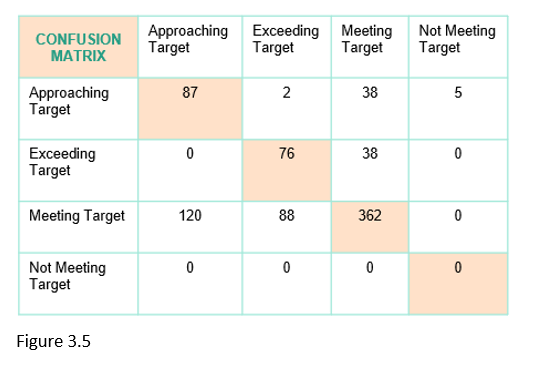
## 

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

## Appendix 3





## Appendix 4 - Breanna

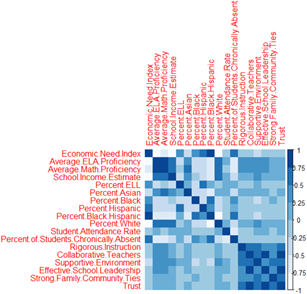
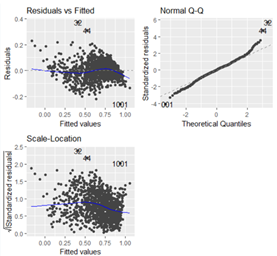
 

Figure 4.1: Correlation Figure 4.2: Full Model - Residual Plots

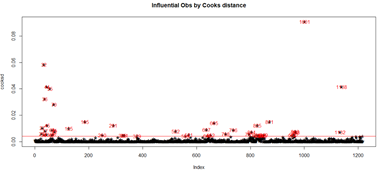


Figure 4.3: Influential Points

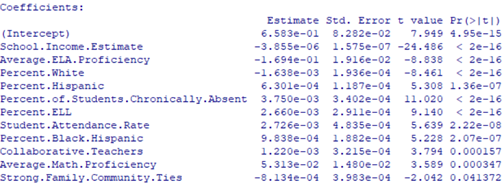
 

Figure 4.4: Outliers Figure 4.5: Stepwise Results

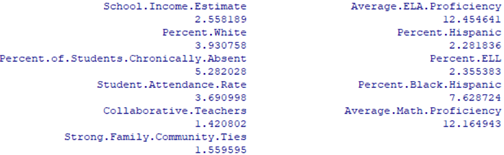


Figure 4.6: Stepwise - VIFs

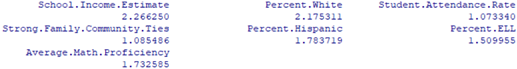


Figure 4.7: Final Model - VIFs

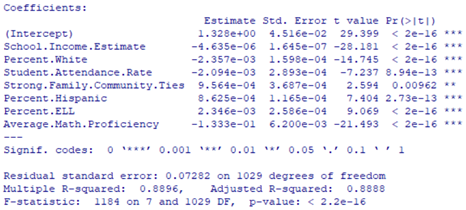


Figure 4.8: Final Model

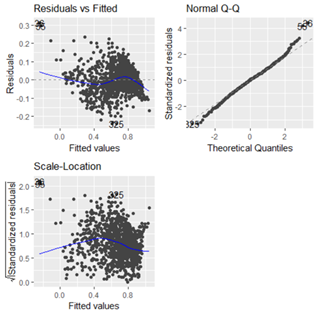


Figure 4.9: Final Model - Residual Plots

## Appendix 5

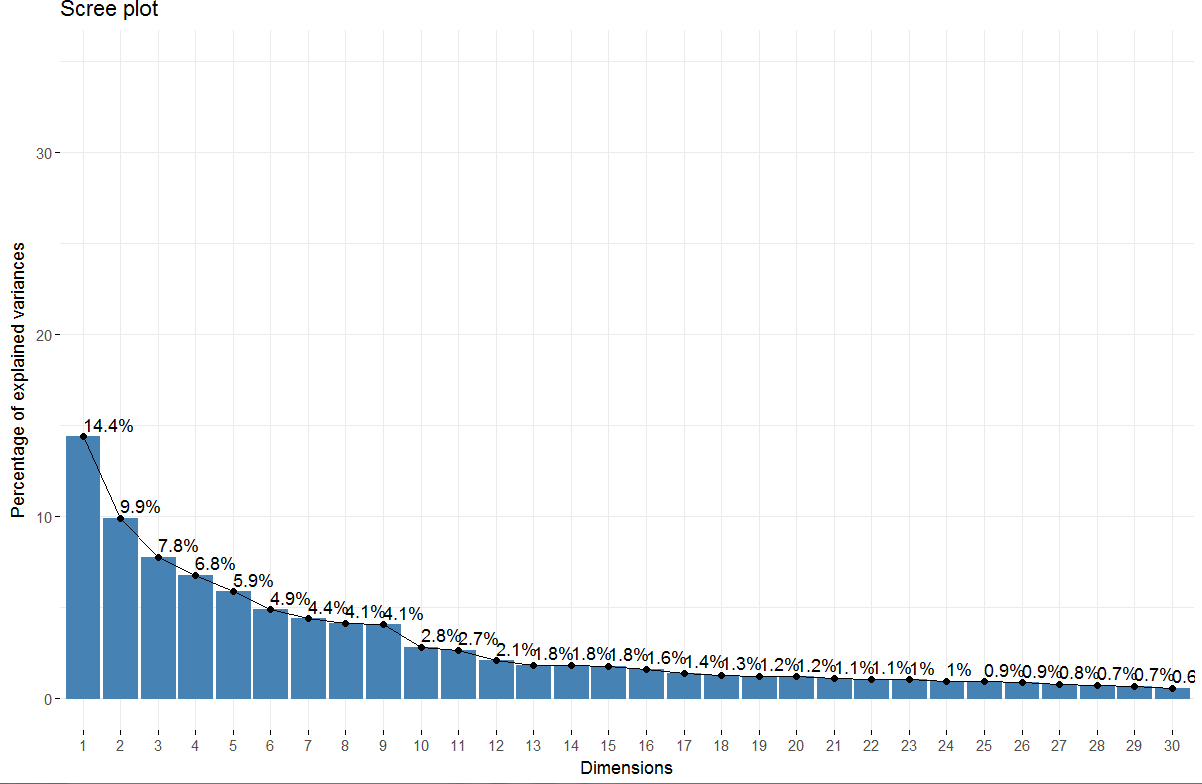


Figure 5.1: Screeplot that shows the percentage of explained variance

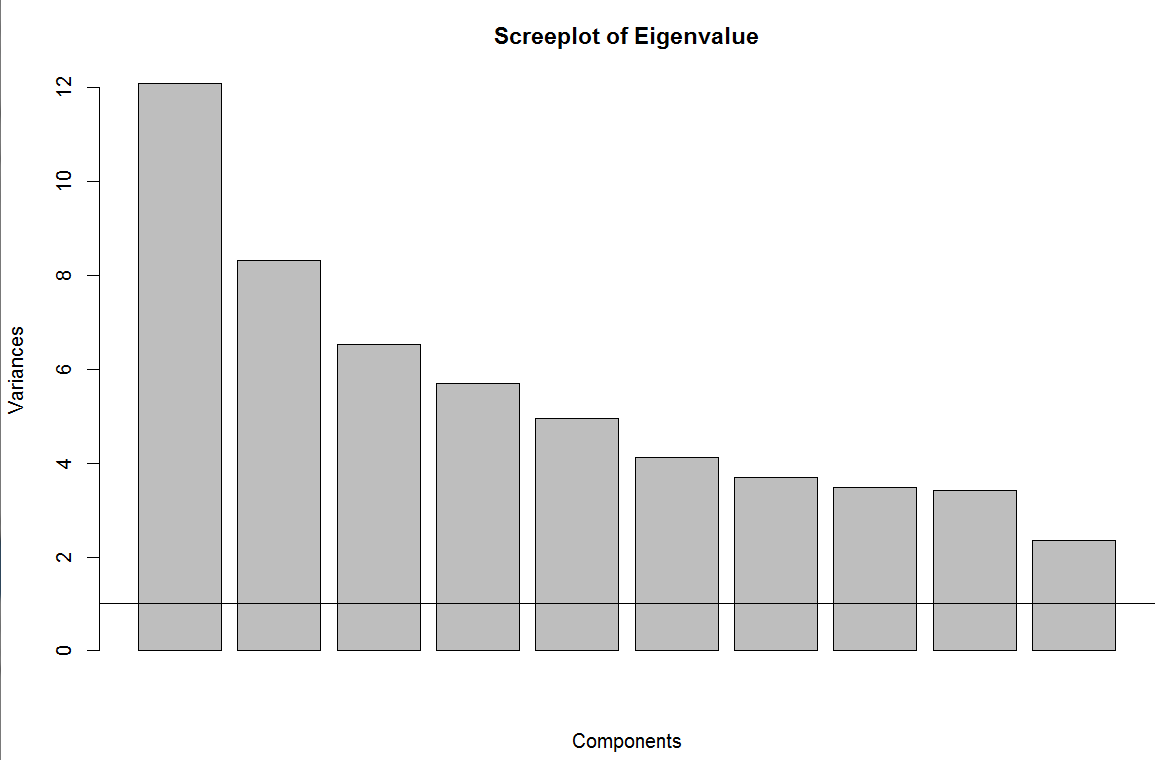


Figure 5.2: Screeplot that shows Eigenvalues greater than 10

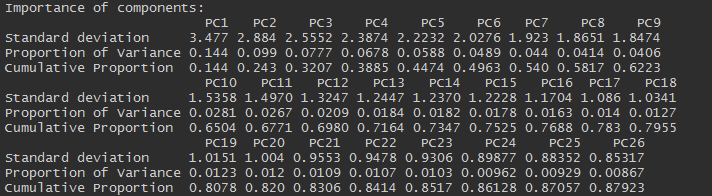


Figure 5.3: Summary of PCA showing the cumulative proportion of variance for each additional component

## 

Figure 5.4: Correlation plot that shows varimax rotation was the proper rotation technique and components are independent.

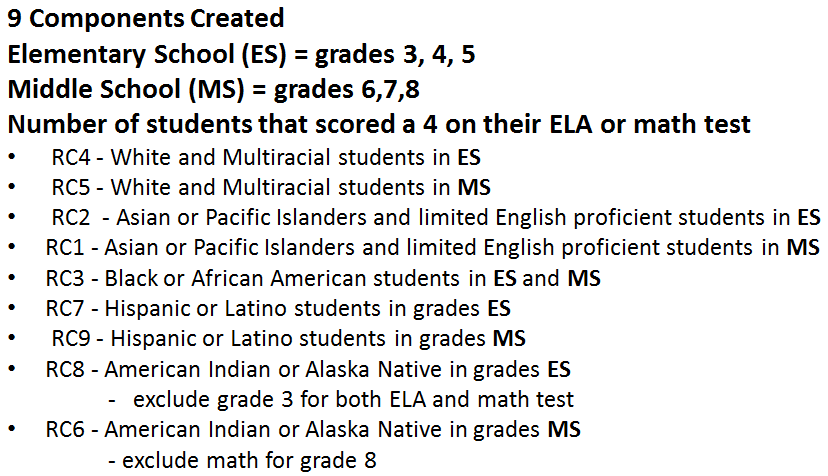


Figure 5.5: Summary of 9 interpretable components from PCA analysis

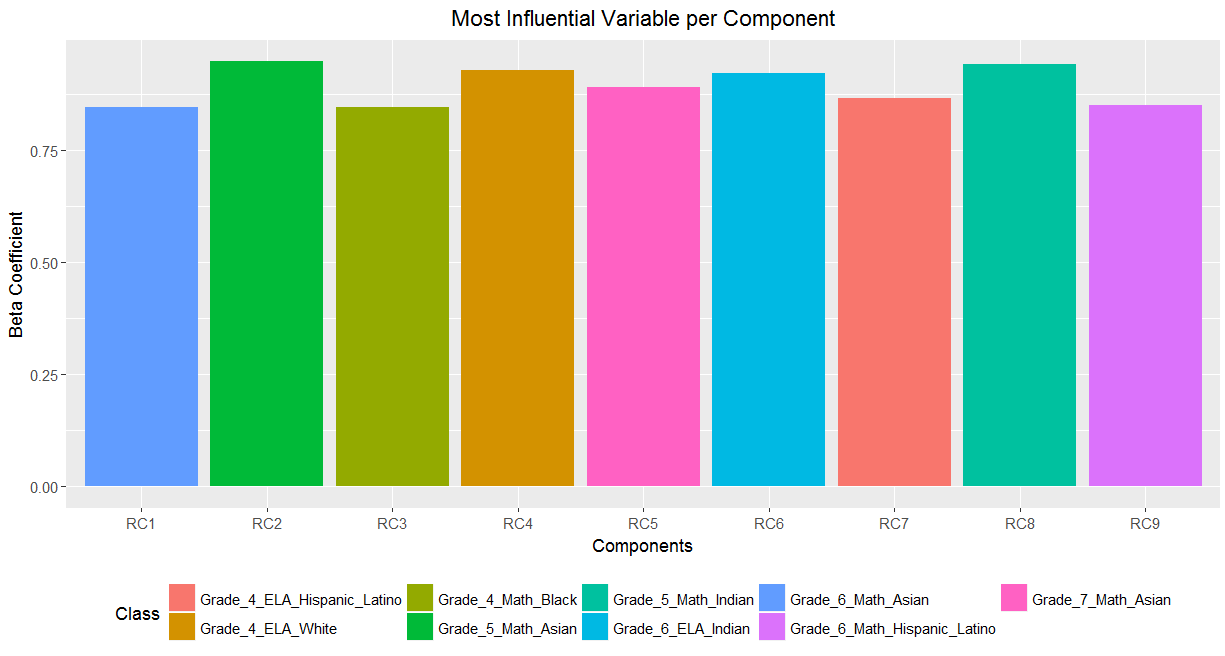


Figure 5.6: Barplot of most influential variable per component

## References

Diploma Disparities: High School Graduation Rates in New York City. (2016, September). Retrieved from <https://comptroller.nyc.gov/wp-content/uploads/documents/Graduation_Rate_Brief.pdf>.

Colangelo, L. L. (2018, February 08). City graduation rates show 1.2% rise for HS students. Retrieved from https://www.amny.com/news/nyc-high-school-graduation-rates-1.16606103.

Helen L. St. Clair-Thompson, and Susan E. Gathercole. Executive functions and achievements in school: Shifting, updating, inhibition, and working memory.

Retrieved from <https://journals.sagepub.com/doi/10.1080/17470210500162854>