Final Project

Jinny Choi, LJ Flores, Megan Zhang, Nathan Kim

May 8, 2019

Introduction

Levels of educational equity (achievement and opportunity in education) vary greatly throughout public schools across the nation and are associated with poverty levels. To indicate poverty level, we can use the percentage of students eligible for free and reduced lunches as a proxy [1]. The two factors do differ - more students qualify for free and reduced lunches than those who live below the federal poverty line. This includes students whose families at up to 185 percent of the poverty line and students who meet additional eligibility requirements (foster children, Head Start Programs, etc). However, poverty level often cannot be measured at the school level while reduced lunch eligibility can be counted; and, since reduced lunch eligibility is derived from federal poverty level, they are closely related.

The dataset we are using contains information about school equity for the 2015-2016 school year and is taken from the National Center for Education Statistics (NCES) [2], a subdepartment of the Department of Education. The NCES releases a trove of data called the Common Core of Data (CCD) every year, and though the exact variables which are measured have changed since the beginning of collection through new standards of collection set by the Bush Administration through the No Child Left Behind Act, the CCD serves as a valuable, open, and comprehensive resource for many types of researchers.

The CCD doesn't hold information on every axis possible: we don't know about students' family income, how they get to school, about teacher salary, or even the graduation rates of schools. But it nonetheless hold valuable information on schools, and attempts to be incredibly thorough in recording ethnicity and race. With its limited power to gather information, the NCES instead uses proxy variables like participation in the Free and Reduced Lunch Program to measure larger concepts like poverty associated with a school.

Design & Questions

The analysis aimed to answer the following three questions:

How can variables be combined to concisely describe the data set?

It would be of great use for educators and policy makers to visualize where certain schools lie in comparison with other schools. Given a raw data set described in nine variables however, it was necessary to aggregate variables into components to generate a 2D graph plotting schools based on certain characteristics. This was accomplished using principal component analysis (PCA).

Does the school type have an impact on whether or not a school is a charter school?

It was of interest to determine if the types of schools had any impact on whether or not a school was a charter school, simply to better characterize the types of schools present. This was analyzed using multivariate analysis of variance (MANOVA).

How can one predict the type of school using variables other than the SCHOOL TYPE variable itself?

Finally, the group was interested in seeing if there were other variables that could be used to predict the type of school, which was accomplished using discriminant analysis.

[1] https://nces.ed.gov/programs/coe/indicator_clb.asp]
[2] https://nces.ed.gov/blogs/nces/post/free-or-reduced-price-lunch-a-proxy-for-poverty]

Descriptive Plots and Summary Statistics

Our group pulled this data to explore and understand schools on four axes: Absolute size, race and ethnicity of students, participation in the free and reduced lunch program, and the type of school.

NCESSCH	ASp	HIp	BLp	WHp	G01p	G07p	G09p	TOTFRLp	TOTAL	SCH_TYPE_TEXT	CHARTER_TEXT
10000500870	0.00	0.39	0.03	0.55	0.00	0.53	0.0	0.40	704	Regular School	No
10000500871	0.00	0.36	0.03	0.59	0.00	0.00	0.3	0.33	1290	Regular School	No
10000500879	0.01	0.42	0.04	0.52	0.00	0.00	0.0	0.44	766	Regular School	No
10000500889	0.00	0.51	0.02	0.44	0.00	0.00	0.0	0.51	863	Regular School	No
10000501616	0.00	0.52	0.03	0.42	0.00	0.00	0.0	0.49	503	Regular School	No
10000502150	0.00	0.49	0.03	0.44	0.48	0.00	0.0	0.47	1088	Regular School	No

These variables were pulled and merged from different datasets in the CCD using a unique school identifier column. After merging, we removed all rows containing NA or negative values (the CCD uses a few negative values to mark cells as NA for continuous variables) because the techniques and functions used in our analyses required complete cases for our data. Then, the proportion variables were calculated by dividing the count for a certain variable by the total number of students in that school. For example, if there were 50 Asian students in a school of 400, then the corresponding proportion value would be 0.125. This was implemented for the following variables: AS (Asian students), HI (Hispanic students), WH (White students), BL (Black students), TOTFRL (Total Students on Free Lunch), G01, G07, G09 (Number of 1st, 7th, and 9th Grade Students).

A simple summary of the continuous variable is shown below.

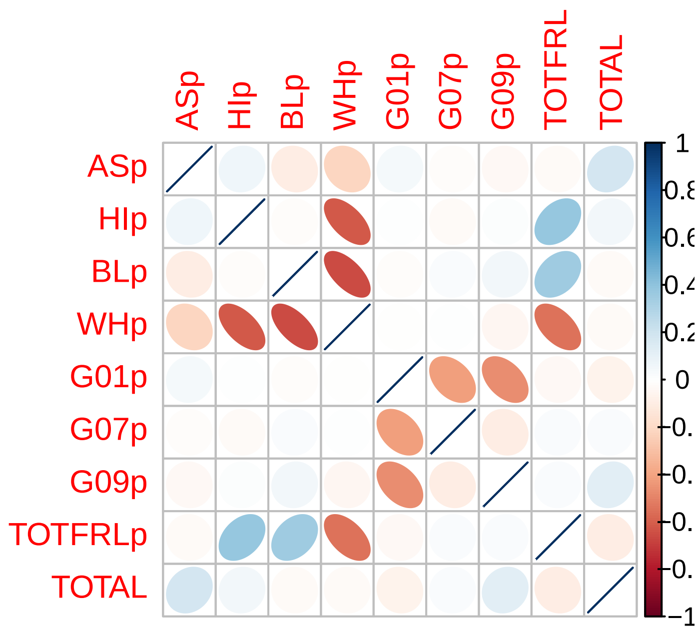
% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Wed, May 08, 2019 - 17:01:56

Table 1:

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
ASp	9,940	0.042	0.090	0.000	0.003	0.038	0.972
HIp	9,940	0.202	0.229	0.000	0.035	0.285	1.000
BLp	9,940	0.182	0.257	0.000	0.011	0.250	1.000
WHp	9,940	0.521	0.346	0.000	0.143	0.838	1.000
G01p	9,940	0.080	0.085	0	0	0.1	1
G07p	9,940	0.072	0.115	0.000	0.000	0.097	1.000
G09p	9,940	0.068	0.110	0	0	0.1	1
TOTFRLp	9,940	0.510	0.272	0.000	0.293	0.728	1.000
TOTAL	9,940	519.054	507.298	6	245	651.2	14,153

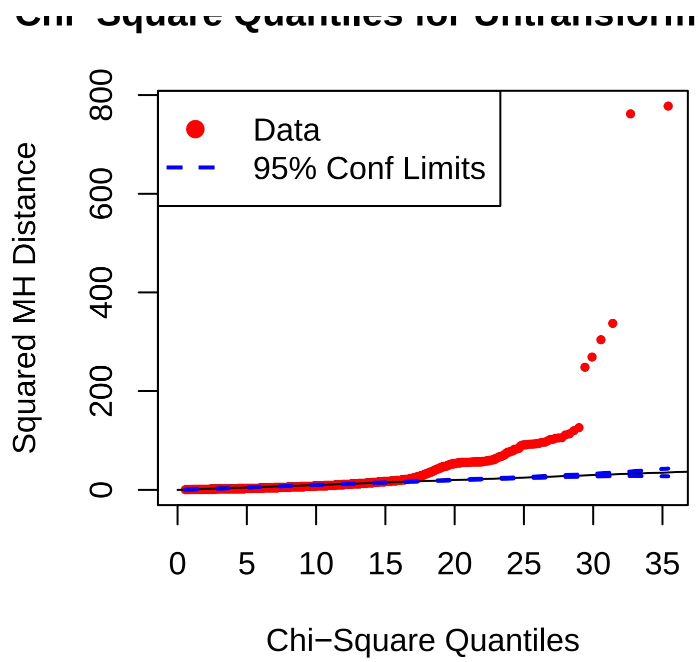
Correlations within the data

Many of the variables have predictably high correlations, as shown below. For example, Hispanic student population is negatively associated with White student population, but is positively correlated with TOTFRLp which acts as a proxy for poverty level. These trends will be further discussed in PCA.



Multivariate Normality

Our untransformed data is graphed below. There are indeed a few outliers with extremely high Mahalanobis distances from the mean, but even without these outliers our data does not come close to approaching multivariate normality. This did not change with any form of $log()$, $sqrt()$, x^2, and Box-Cox procedures we carried through, and so we chose to leave our data in this shape. Still, transformations will be applied for MANOVA and discriminant analysis, which require multivariate normality at least within groups or for a fewer set of variables.



The remaining two variables are categorical. SCH_TYPE_TEXT describes a "type" of school, and is a factor variable containing levels "Regular School" (traditional public schools), "Alternative Education School" (generally thought of as for students who have been identified by their districts or states as "not succeeding" in a regular school), "Vocational Education School" (schools training students more directly on career paths rather than just academics), and "Special Education School" (for students with special needs). CHARTER_TEXT is an indicator variable telling whether a school is a charter school or not. It has two levels: "Yes" and "No".

A simple breakdown of these categorical variables is shown below. They will be further discussed in MANOVA and Discriminant Analysis.

Variable	count
Alternative Education School	501
Regular School	8705
Special Education School	707
Vocational Education School	27
Not a Charter	8861
Charter	1079
Total	9940

SCH_TYPE_TEXT	CHARTER_TEXT	count
Alternative Education School	No	467
Alternative Education School	Yes	34
Regular School	No	7679
Regular School	Yes	1026
Special Education School	No	688
Special Education School	Yes	19
Vocational Education School	No	27

Results

Principal Components Analysis

We have several numerical variables that are inherently related; we know that by the mounting evidence of school segregation in America, schools with many Black students are unlikely to have many White or Asian students, and vice versa. We also know intuitively that schools with many grade 9 students are unlikely to have many grade 1 students, because schools in America are mostly divided into rigid separations of elementary-middle-high school.

Additionally, we are interested in how Free Lunch program participation ties to these other variables. PCA could give us solid evidence that this proxy for poverty is also tied to ethnicity.

The correlation diagram shown above also shows strong negative correlation (e.g. HIp and WHp, BLp and WHp, G09p and G01p) of 0.5 and above for a few pairs of variables, which leads us to believe that PCA will be useful in combining them.

```
## Importance of components:
##                           Comp.1    Comp.2    Comp.3    Comp.4    Comp.5
## Standard deviation     1.5282560 1.2651368 1.1296565 1.0501921 0.9881506
## Proportion of Variance 0.2595074 0.1778412 0.1417915 0.1225448 0.1084935
## Cumulative Proportion  0.2595074 0.4373486 0.5791401 0.7016850 0.8101785
##                            Comp.6     Comp.7     Comp.8      Comp.9
## Standard deviation     0.88850676 0.72882106 0.56928147 0.252364875
## Proportion of Variance 0.08771603 0.05902001 0.03600904 0.007076448
## Cumulative Proportion  0.89789449 0.95691451 0.99292355 1.000000000
##
## Loadings:
##         Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9
## ASp      0.080  0.071  0.643  0.233  0.213  0.651  0.128  0.013  0.189
```

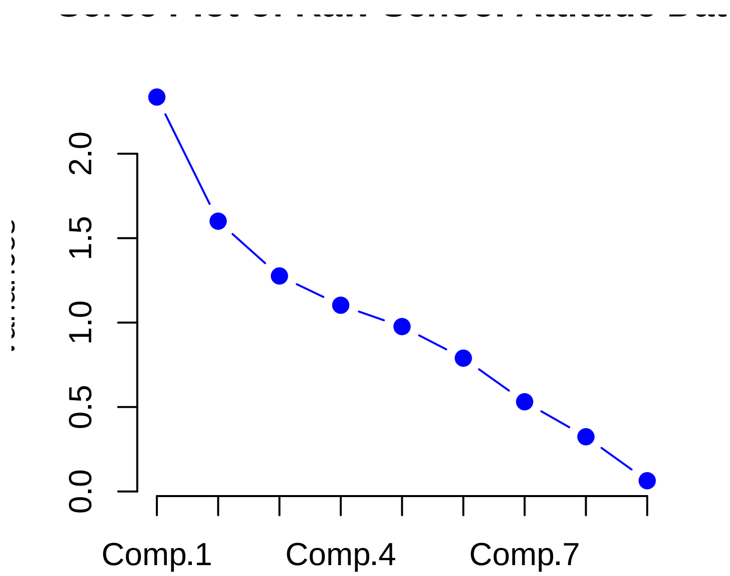


Figure 1: Screeplot of PCA

```
## HIp      0.424  0.069  0.256  0.073 -0.618 -0.196 -0.356  0.000  0.445
## BLp      0.427 -0.027 -0.311 -0.107  0.638 -0.017 -0.209 -0.049  0.506
## WHp     -0.620 -0.056 -0.075 -0.026 -0.099 -0.066  0.291  0.022  0.712
## G01p    -0.038  0.715  0.037 -0.059  0.109 -0.139  0.003  0.672  0.005
## G07p     0.011 -0.418 -0.146  0.740  0.047 -0.090 -0.059  0.492 -0.004
## G09p     0.067 -0.520  0.108 -0.612 -0.081  0.181 -0.023  0.548  0.004
## TOTFRLp  0.492  0.019 -0.166  0.021 -0.166 -0.030  0.836  0.035  0.045
## TOTAL    0.006 -0.174  0.597 -0.041  0.329 -0.687  0.164 -0.047 -0.031
```

The first three components have an eigenvalue above 1, however it requires five parameters to cover 80% of the variance in data and using only three components only explains about 57.9% of the variance in our data.

The scree plot below is too smooth to unambiguously cut and retain components, but we do see slight indications of bends after the first, third, and fifth components.

We can't conduct parallel analysis because we don't have multivariate normality, so our final method of determining which eigenvalues to keep was to observe the variables that they represented.

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
ASp	0.08	0.07	0.64	0.23	0.21
HIp	0.42	0.07	0.26	0.07	-0.62
BLp	0.43	-0.03	-0.31	-0.11	0.64
WHp	-0.62	-0.06	-0.07	-0.03	-0.10
G01p	-0.04	0.71	0.04	-0.06	0.11
G07p	0.01	-0.42	-0.15	0.74	0.05
G09p	0.07	-0.52	0.11	-0.61	-0.08
TOTFRLp	0.49	0.02	-0.17	0.02	-0.17
TOTAL	0.01	-0.17	0.60	-0.04	0.33

Component 1: Ethnicity & Lunch Variable The first component accounts mainly for ethnicity. Schools with a high component 1 value have high hispanic and black student proportions and high proportions of students on the free lunch program, while schools with low component 1 values have high white student proportions.

Component 2: Grade Level Variable The second component accounts for the number of students in the school by grade level. Schools with a high component 2 value have many grade 1 students, while those with low values have more grade 7 and 9 students.

Component 3: Asian & Lunch Variable The third component accounts for the proportion of Asian students in the school, with higher component 3 values indicating higher Asian proportions. Low component 3 values also indicate high use of free lunch.

Component 4: Grade Level Noise Variable The fourth component distinguishes schools with high grade 7 and low grade 9 student populations (high component values) from schools with low grade 7 and high grade 9 populations (low component values).

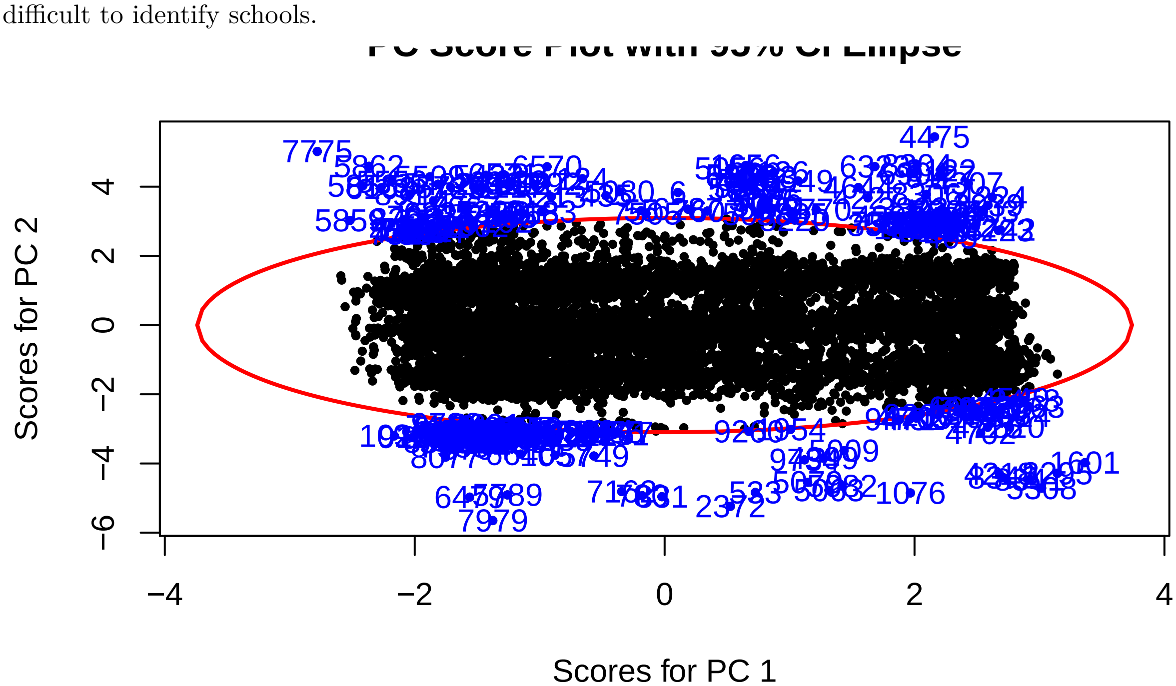
Component 5: Ethnicity Noise Variable The fifth component distinguishes schools with high hispanic and low black student populations (high component values) from schools with low hispanic and high black student populations (low component values).

Since the fourth and fifth component were interpreted as noise variables but the first three components were necessary to explain variances in all continuous, we decided to keep only the first three components.

Principal Component Analysis Plots

After obtaining this, we made simple graphs of the principal components against each other to ensure there was no colinearity here. There was not, as far as we could tell.

We then create a 95% confidence ellipse plotting the first vs. second components. We also created an integer ID column solely for this graph because the 12-digit NCESSCH ID took up too much space and made it difficult to identify schools.



Among this jumbled mess of a diagram, we do have some fairly definite conclusions. Firstly, there is few data at the edges of the ellipse (extreme values for PC 1), and there appears to be clumps of data at certain values along the PC 2 axis, which we called Grade Level before. We suspect that this may be because a school that has Grade 9 students is not likely to have Grade 1 students because it would be a high school instead of an elementary school.

Outliers are also fairly easily explainable, and our group took some time to explain them using the school profiles on the official NCES online database. Some schools had extremely small numbers of students (often in the single digits), which resulted in other extreme values such as either all students being of a certain ethnicity or no students being of that ethnicity, as well as all of the students being on the Free Lunch program.

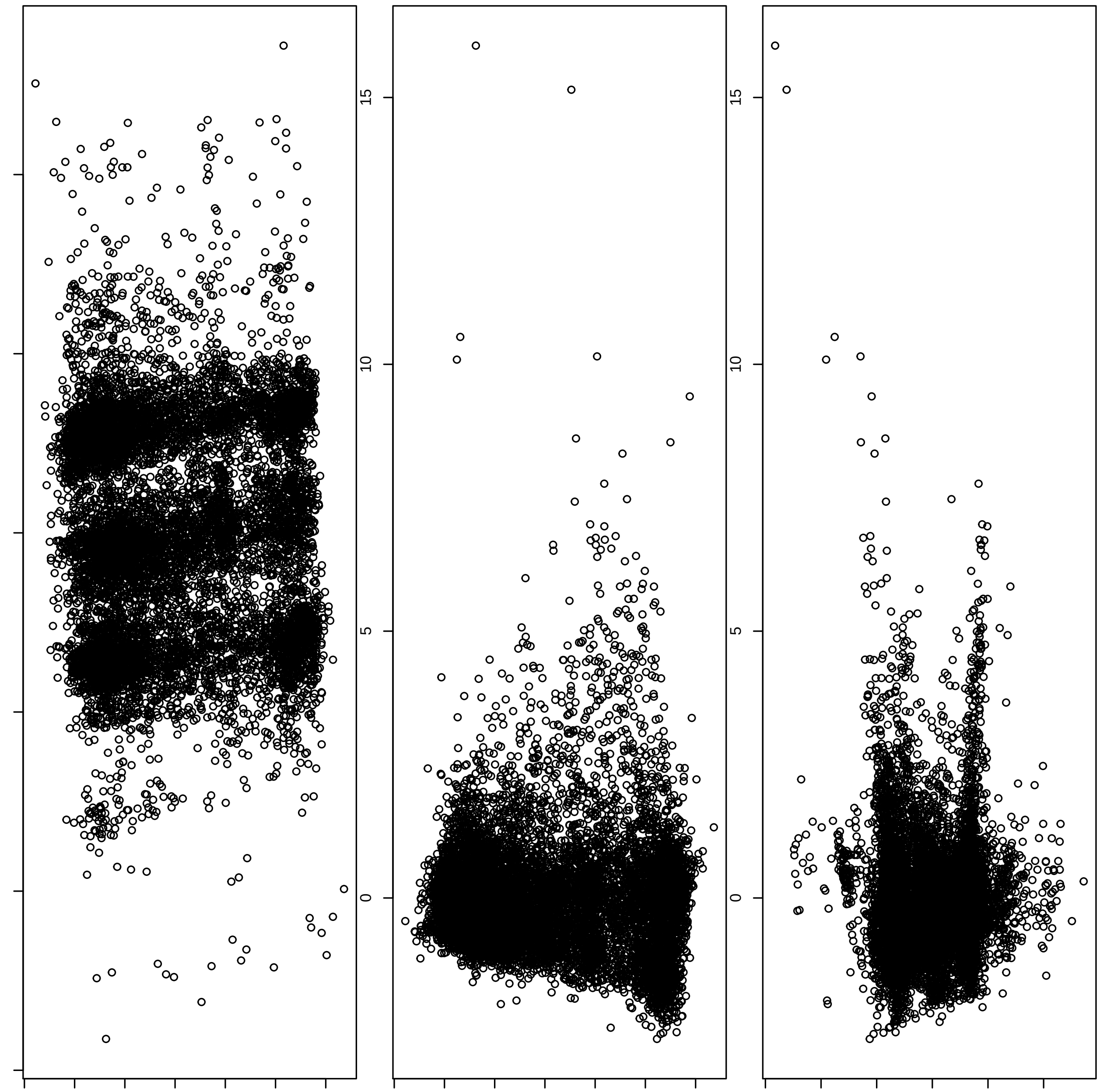


Figure 2: PCA Components

Other schools were even less traditional, such as the Electronic Classroom of Tomorrow (#7979, the lower left-hand-side of the scoreplot), which has over 13,000 students and operates mostly electronically.

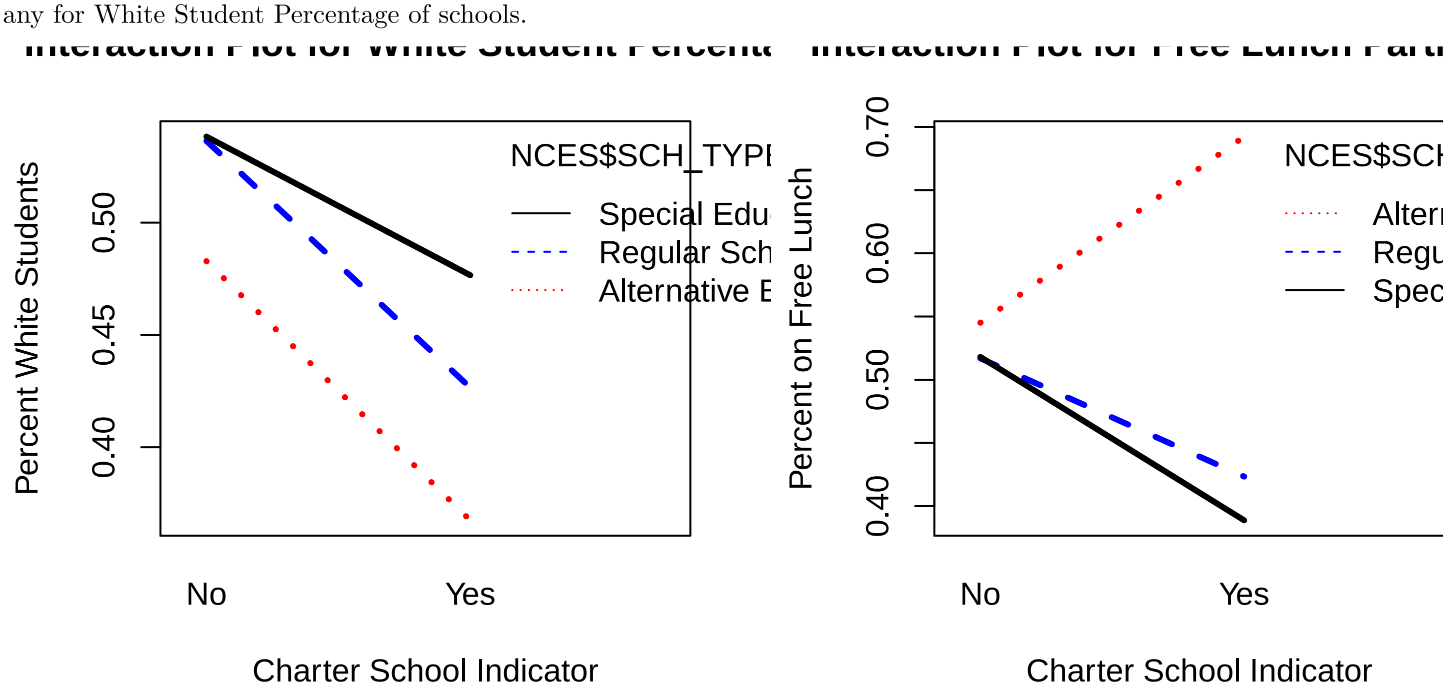
MANOVA

We also wanted to find relationships between types of schools and the poverty level of a school measured by the Free Lunch Program. We looked at types of schools along two axes: the school type variable marked by "SCH_TYPE_TEXT", and an indicator variable telling whether a school is a charter school or not. Because these are two different categorical variables predicting a single continuous response variable, MANOVA is aptly suited to discuss the problem at hand.

For this analysis, we removed Vocational Schools from the analysis because all 27 vocational schools in our dataset were not charter schools. To fit the assumptions of MANOVA requiring multivariate normality, we used only two response variables – WHp, representing percentage of white students, and TOTFRLp, representing Free Lunch Program participation – because more than two response variables always resulted in non-normal data.

Lastly, we transformed the WHp variable using a logit transformation. With the untransformed data, we found several "step"-shaped regions in the Chi-Square Quantile Plot of our residuals that we believed were due to having a large number of 0 and 1 values for WHp. Using the logit() function helped "smooth" out this data, and made sense given that we are dealing with percentage data.

We hypothesized that there would be an interaction with charter schools and school type, since charter schools are a heterogenous group of schools ranging from Montessori and "forest" schools to strict "No excuses" charter schools that have created controversy over the past twenty-five years. We believed that school types would "separate" these schools along teh White student percentage and Free Lunch participation axes, meaning that there would be some sort of interaction between the variables SCH_TYPE_TEXT and CHARTER_TEXT. But the interaction plots, shown below, tell a mixed story: there is definitely an interaction between the two variables for the Free Lunch Program participation, but a relatively small one if any for White Student Percentage of schools.



MANOVA analysis revealed mixed results as well. We used Pillai's trace because we are worried about the somewhat dubious distribution of our residuals, shown later

The multivariate analysis below reveals high F-statistics and low enough p-values for any reasonable threshold to reject the null hypothesis and conclude that there is a difference in the multivariate profiles in these types of schools. However, the Pillai statistic is remarkably small, indicating that the effects of each of the predictors in the model is not high.

Multivariate Contrasts

Predictors	Df	Pillai	approx F	num Df	den Df	Pr(>F)
NCES$SCH_TYPE_TEXT	2	0.0020493	5.08088	4	19814	0.0004329
NCES$CHARTER_TEXT	1	0.0428178	221.56317	2	9906	0.0000000
NCES$SCH_TYPE_TEXT:NCES$CHARTER_TEXT	2	0.0035704	8.85877	4	19814	0.0000004
Residuals	9907	NA	NA	NA	NA	NA

We worry that the significant p-values produced in the last column may have been produced because of the large sample size and not because of a real-world trend. To investigate this further, we performed univariate analyses on both White Student population and Free Lunch participation.

Predictors for White Student Pop.	Df	Sum Sq	Mean Sq	F value	Pr(>F)
NCES$SCH_TYPE_TEXT	2	1.2753341	0.6376670	5.3809893	0.0046167
NCES$CHARTER_TEXT	1	11.3642594	11.3642594	95.8979426	0.0000000
NCES$SCH_TYPE_TEXT:NCES$CHARTER_TEXT	2	0.0432243	0.0216122	0.1823754	0.8332913
Residuals	9907	1174.0159855	0.1185037	NA	NA

Predictors for Free Lunch Participation	Df	Sum Sq	Mean Sq	F value	Pr(>F)
NCES$SCH_TYPE_TEXT	2	1.170331	0.5851656	8.021595	0.0003304
NCES$CHARTER_TEXT	1	7.106327	7.1063272	97.415290	0.0000000
NCES$SCH_TYPE_TEXT:NCES$CHARTER_TEXT	2	1.809848	0.9049240	12.404921	0.0000042
Residuals	9907	722.703628	0.0729488	NA	NA

Here, results are consistent with our interaction plots. We see that the main effects produce significantly different values for both White student percentage and Free Lunch participation, but the interaction for White student percentage is found to be insignificant.

To investigate further into contrasts, we performed a pairwise comparison of all groups using Tukey's Honest Significant Differences, sorted so that variables with significant p-values are placed at the top of the following table.

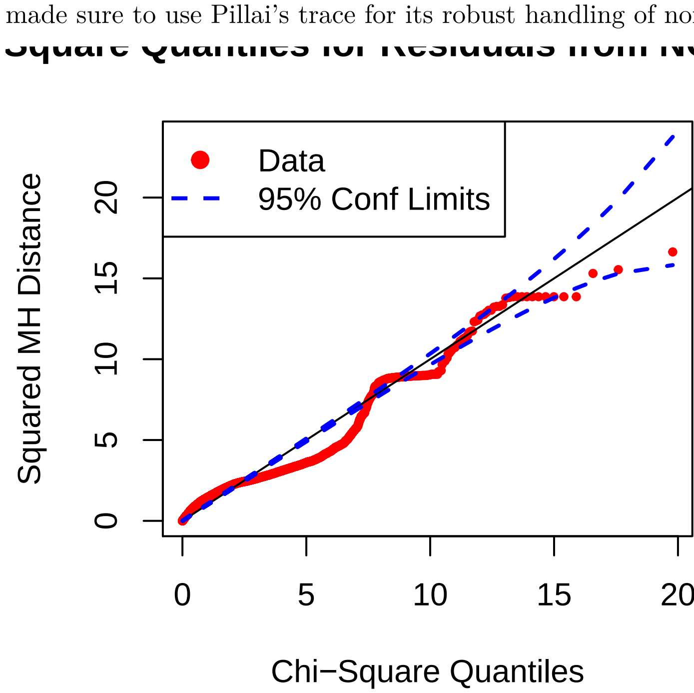
WHp Effect	p adj	diff	lwr	upr
Regular School:Yes-Regular School:No	0.000	-0.110	-0.142	-0.077
Regular School:Yes-Special Education School:No	0.000	-0.112	-0.160	-0.063
Regular School:No-Alternative Education School:No	0.014	0.054	0.007	0.100
Regular School:Yes-Alternative Education School:No	0.040	-0.056	-0.111	-0.002
Alternative Education School:Yes-Regular School:No	0.049	-0.169	-0.338	0.000
Alternative Education School:Yes-Special Education School:No	0.054	-0.171	-0.343	0.002
Special Education School:No-Alternative Education School:No	0.078	0.055	-0.003	0.114
Alternative Education School:Yes-Alternative Education School:No	0.410	-0.115	-0.290	0.059
Special Education School:Yes-Alternative Education School:Yes	0.879	0.109	-0.172	0.390
Regular School:Yes-Alternative Education School:Yes	0.923	0.059	-0.112	0.230
Special Education School:Yes-Special Education School:No	0.972	-0.062	-0.290	0.166
Special Education School:Yes-Regular School:No	0.975	-0.060	-0.285	0.166
Special Education School:Yes-Regular School:Yes	0.989	0.050	-0.177	0.277
Special Education School:Yes-Alternative Education School:No	1.000	-0.006	-0.236	0.223
Special Education School:No-Regular School:No	1.000	0.002	-0.037	0.041

Though a little hard to read, there are some notable trends from this analysis. For alternative and special education schools, the status of if it is a charter school does not seem to have a significant effect on the White student percentage. However, specific subgroups across both of these variables result in significant interactions. This is also congruent with the interaction plot and even with the MANOVA and ANOVA outputs from above, since there would need to be at least one group difference according to ANOVA and MANOVA.

TOTFRLp Effect	p adj	diff	lwr	upr
Regular School:Yes-Alternative Education School:No	0.000	-0.122	-0.165	-0.079
Regular School:Yes-Regular School:No	0.000	-0.094	-0.119	-0.068
Regular School:Yes-Special Education School:No	0.000	-0.095	-0.133	-0.057
Regular School:Yes-Alternative Education School:Yes	0.000	-0.269	-0.403	-0.135
Special Education School:Yes-Alternative Education School:Yes	0.001	-0.304	-0.524	-0.083
Alternative Education School:Yes-Regular School:No	0.002	0.175	0.043	0.308
Alternative Education School:Yes-Special Education School:No	0.003	0.174	0.039	0.310
Alternative Education School:Yes-Alternative Education School:No	0.026	0.147	0.010	0.284
Special Education School:Yes-Alternative Education School:No	0.132	-0.156	-0.337	0.024
Regular School:No-Alternative Education School:No	0.242	-0.028	-0.065	0.009
Special Education School:Yes-Regular School:No	0.306	-0.128	-0.305	0.049
Special Education School:Yes-Special Education School:No	0.310	-0.129	-0.308	0.050
Special Education School:No-Alternative Education School:No	0.549	-0.027	-0.073	0.019
Special Education School:Yes-Regular School:Yes	0.994	-0.035	-0.213	0.144
Special Education School:No-Regular School:No	1.000	0.001	-0.030	0.032

Again, we find congruent results with the interaction plots and the MANOVA/ANOVA outputs.

To finish MANOVA, we examine our assumption of normality for each group considered. This is tedious because of our high number of groups, so we graphed a Chi-Square Quantile plot of our residuals after our model creation. The “dubious” aspect that we mentioned earlier regarding normality is obvious, because although our data nearly follows the theoretical quanties on the whole, it does so in a very strange and “step”-like manner. Still, our group felt that this approximated multivariate normality well enough, and we made sure to use Pillai’s trace for its robust handling of non-normal data.



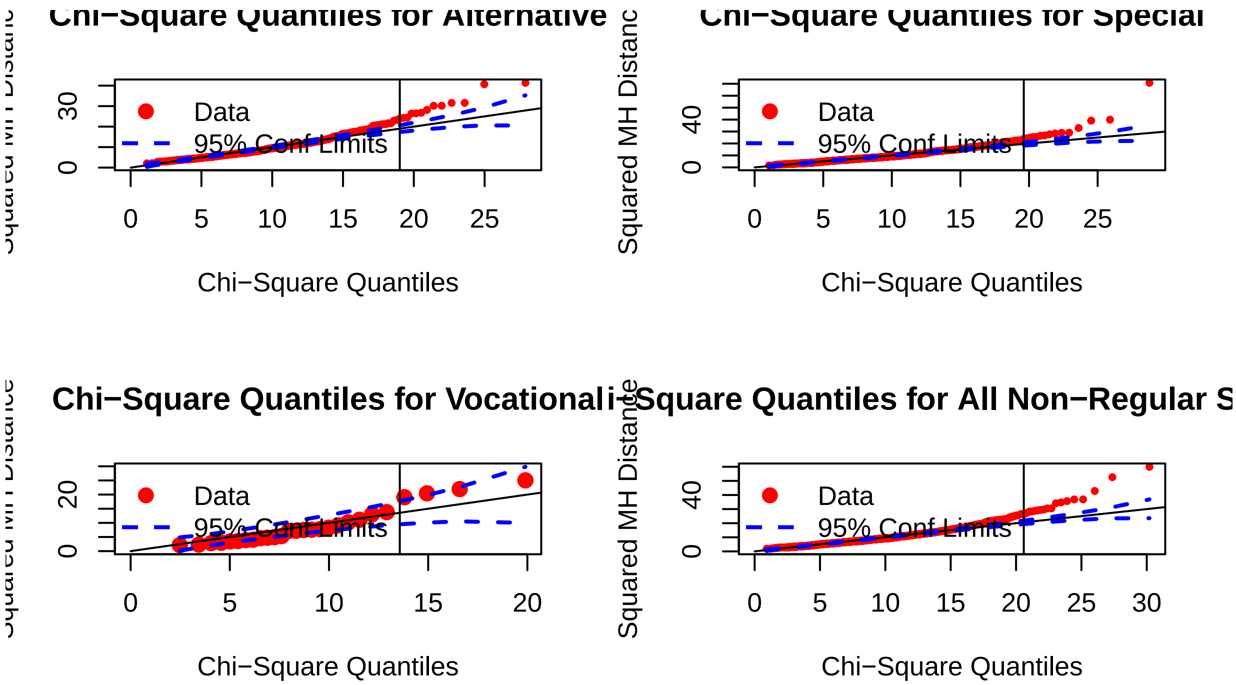
Discriminant Analysis

To finish off the report, our group used discriminant analysis to investigate the classification that the NCES uses for school types. What really makes a school "alternative," and can we predict this given the variables we have? Discriminant analysis might help.

Assumptions

Our data, as stated before, is not multivariately normal. But after transformations ("log+1" transformations on all continuous variables), we were able to observe roughly normal data for three of our four groups; only regular schools did not have close to normal data.

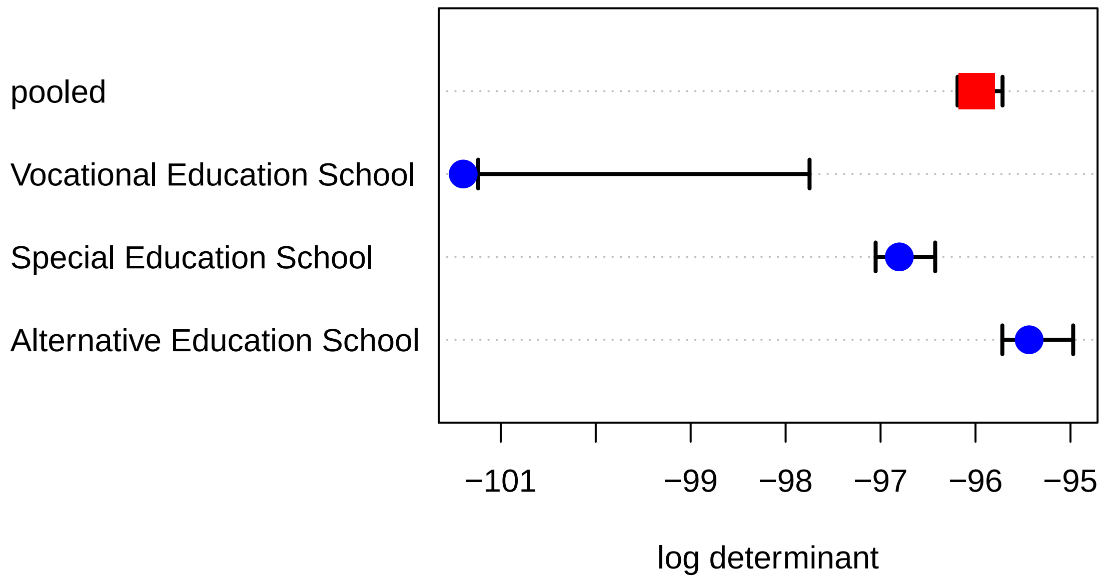
Our group proceeded with this, after reading online on discriminant analysis' robustness to slight violations to normality.




```
## # A tibble: 3 x 2
##   SCH_TYPE_TEXT                 count
##   <fct>                         <int>
## 1 Alternative Education School    501
## 2 Special Education School        707
## 3 Vocational Education School      27
```

Covariance Matrices

There are too many variables (columns of the covariance matrices) and groups (number of covariance mtrices) to make for an easy visual comparison, but running the Box's M test and visualizing the log determinants showed that the covariance matrices are not equal, although they are close to equal.



We thus decided to use quadratic discriminant analysis .

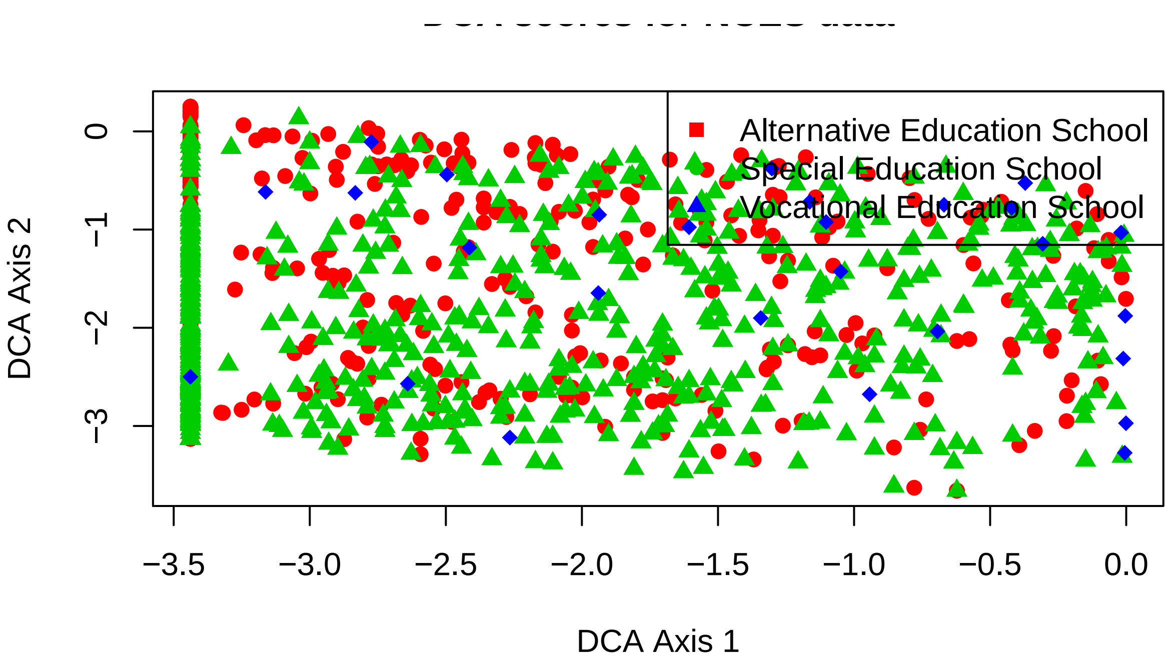
QDA Results

```
## Call:
## qda(NCES[, 2:10], grouping = NCES$SCH_TYPE_TEXT)
##
## Prior probabilities of groups:
## Alternative Education School     Special Education School
##                   0.40566802                   0.57246964
##  Vocational Education School
##                   0.02186235
##
## Group means:
##                                       ASp         HIp         BLp      WHp
## Alternative Education School 2.785253e-06 0.010604450 0.003548467 36.47083
## Special Education School     2.463625e-06 0.012278787 0.003180947 39.47919
## Vocational Education School  1.428671e-06 0.008618049 0.002290824 21.96301
##                                   G01p         G07p         G09p  TOTFRLp
## Alternative Education School 0.1614768 0.0008880901 0.0003935502 15.61302
## Special Education School     0.1564469 0.0008558899 0.0004836812 14.71927
## Vocational Education School  0.1727749 0.0012282336 0.0001890242 18.51450
##                                   TOTAL
## Alternative Education School 0.03149361
## Special Education School     0.03134702
## Vocational Education School  0.01733858
```

```
## [1] "Raw Result Correct Rate = 0.740890688259109"
```

```
## [1] "Cross-validated Results Correct Rate = 0.715789473684211"
```

Upon performing quadratic discriminant analysis, the resulting function was able to correctly predict a school's type 74.1% of the time based on raw results, and 71.6% of the time with cross validated results.



Although no clear groups were observed between alternative, special, and vocational schools based on the discriminant axes as shown in the plot above, the group believes that the model is still capable of decently predicting school type based on the correctness rate calculated.

Further analysis below, using stepwise DA to find a model with few variables, indicates that this can best be predicted with the G09p variable and the TOTFRLp variable, comprising about 63% of the total variance.

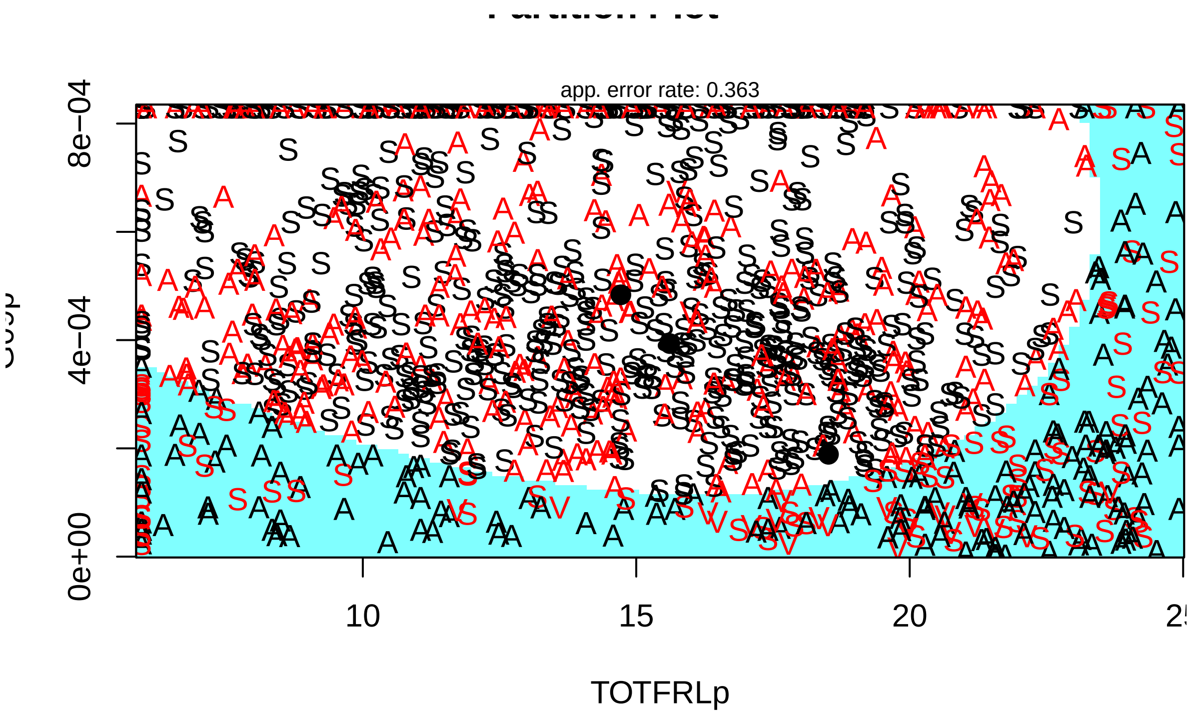
```
## correctness rate: 0.62024;  in: "G09p";  variables (1): G09p
## correctness rate: 0.63077;  in: "TOTFRLp";  variables (2): G09p, TOTFRLp
##
##  hr.elapsed min.elapsed sec.elapsed
##       0.000       0.000      52.188
```

```
## crossval.rate      apparent
##     0.6307692     0.3627530
```

```
## SCH_TYPE_TEXT ~ G09p + TOTFRLp
## <environment: 0x7fa1dbc20518>
```

This is particularly interesting to our group, as seemingly unrelated variables would be able to predict school type. We think Grade 9 makes sense as a strong predictor, because we often hear of specialized and differentiated high schools instead of specialized middle or elementary schools.

A partition of space according to this model is shown below:



This is incredibly hard to read because of the large sample size, but there are definitely distinct regions corresponding to special and alternative high schools – the blue region indicates alternative, while the white region indicates special. The 27 vocational high schools included in the analysis were not allocated space here, presumably because the group size was too small.

Comparison of Different Methods

The three tests the group implemented on the data set all had different aims in analyzing the data, and tackled different aspects of the data set as well. PCA was mainly a descriptive tool that allowed the group to see which factors and variables contributed the most variance between schools. MANOVA directly answered if the school type had an impact on if the school was a charter school or not, whereas discriminant analysis built a model for categorizing schools.

Conclusion

Schools across the United States were analyzed based on NCES data, thereby tackling the following questions:

How can variables be combined to concisely describe the data set? Does the school type have an impact on whether or not a school is a charter school? How can one predict the type of school using variables other than the SCHOOL TYPE variable itself?

It was observed that the data could be described using three main variables: (1) Ethnicity and Lunch, (2) Grade Level, and (3) Asian and Lunch, with two other noise variables distinguishing grade level and ethnicity. PCA verified that ethnicity and use of free lunch are intimately tied, and that these two variables account for most of the variance in data. These results reflect interesting correlations between variables:

1. Schools with high proportions of Hispanic and Black students have high free lunch use and low proportions of white students,
2. Students schools with more first grade students have fewer seventh and ninth grade students, and
3. Schools with higher Asian proportions have lower proportions of students on the free lunch program

The MANOVA analysis showed that school type and charter status can both serve as predictors of White student percentage and Free Lunch Program participation, but not all of the group means are separate – for example, alternative schools do not have a significantly different white student percentage under charter schools or non-charter schools.

Finally, the discriminant analysis was able to predict school type from the proportion of ninth grade students (G09p) and proportion of students on free lunch (TOTFRLp). This is an interesting result considering that school type seems unrelated to G09p and TOTFRLp, and a good way of characterizing schools whose types we may not know given its other data.

Points for Further Analysis

The analysis generally characterized the schools in the NCES data set, drawing conclusions about the relation of ethnicity and use of free lunch, school type and charter schools, and school type and G09p and TOTFRLp. However, the results provide little insight that could concretely inform policy decisions or motivate changes in schools. It would be useful for future analyses to look into current policy data, and possibly tie these to school demographics analyzed in this project, in order to to understand what action to be done and which schools to target for implementation of certain reforms.