Analysis of Post-Secondary Bound Graduation Rates in Pennsylvania Public Schools

by

Elijah Lovelace

BA Biochemistry, University of Colorado, 2014

Submitted to the Graduate Faculty of the

Department of Biostatistics

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2020

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This thesis was presented

by

Elijah Lovelace

It was defended on

April 20, 2020

and approved by

Committee Member: Jeanine Buchanich, PhD, MPH, MEd, Research Associate Professor, Biostatistics, Graduate School of Public Health, University of Pittsburgh

Committee Member: Douglas Landsittel, PhD, Professor, Biomedical Informatics, Graduate School of Medicine, University of Pittsburgh

Committee Member: Jenna Carlson, PhD, Assistant Professor, Biostatistics, Graduate School of Public Health, University of Pittsburgh

Thesis Advisor: Ada Youk, PhD, Associate Professor, Biostatistics, Graduate School of Public Health, University of Pittsburgh

Ada Youk, PhD

Analysis of Post-Secondary Bound Graduation Rates in Pennsylvania Public Schools

Elijah Lovelace, MS

University of Pittsburgh, 2020

Abstract

High school graduation rates have been increasing statewide in Pennsylvania in recent years. However, the rate of these graduates attending any form of post-secondary education remains inconsistent across the state and even within districts. Access to post-secondary education is important to public health because significant reductions in negative health outcomes have been observed in those with post-secondary education levels. This thesis analyzes the relationship between a school's post-secondary bound graduation rate and the distribution of the race and socioeconomic status of the student population. In order to quantify and test these relationships for statistical significance, we developed a mixed effects model relating demographic covariates and other school characteristics to multiple post-secondary bound graduation rates. In addition, we also utilized machine learning clustering techniques to categorize schools on student demographic data distributions and model the differences in post-secondary bound graduation rates between these groups. We observed that school-wide Title I status (as an indicator for socioeconomic status) had a negative effect on post-secondary bound rates. In addition, there a was positive relationship observed between the proportion of students from Historically Underserved Groups (HUGs) in a school's student population and post-secondary bound graduation rates. Through our cluster analysis, we found that the race/ethnicity distribution of students in individual schools fell into four categories. Further analysis using the results from the cluster analysis showed the previous

relationship between student HUG proportion and post-secondary graduation rates applied only toward schools with a majority Black student population and not schools with a majority Hispanic student population or schools with more diverse student populations. In conclusion, there is evidence the proportion of school's student population that is historically underserved may affect the post-secondary bound graduation rates of that school, however, this trend may not be similar in schools serving different demographics of students. The results of this analysis justify further quantitative and qualitative research into understanding what school-level qualities influence a student's access to post-secondary education.

Table of Contents

List of Tables

List of Figures

1.0 Introduction

Education has long been an important determinate in the overall health of an individual. Relationships between education level and several health outcomes have been thoroughly observed in multiple countries and time periods (Desjardins, 2008). Particularly, post-secondary education influences multiple factors that contribute to one's ability to maintain a healthy lifestyle including financial security, safer work environments and access to healthcare (Desjardins, 2008). Many factors contribute to one's access to post-secondary education; however, the public-school system bears most of the responsibility in preparing individuals for further education and contributes greatly to the accessibility of a post-secondary education. The responsibility of the public-school system is even more significant for students from historically underserved groups (HUG).

According to the 2019 National Vital Statistics Report, the age-adjusted mortality rate of those with a high school diploma or less is more than twice the rate of those with some college or more (National Vital Statistics Report, 2019). One could rightfully argue that many factors contribute simultaneously to increased education and decreased mortality, such as family income and childhood environment. That being said, there is still a significant relationship between education and health outcomes (Zajacova, 2018). Negative correlations in predicted probability of negative health outcomes (such as poor/fair health, multimorbidity and functional limitations) and education levels have been observed in men and women and among all race/ethnic groups (Zajacova, 2018). For example, the probability of reporting fair or poor health in White men and women without a high school diploma is 57% compared to only 9% in those with a college education (Zajacova, 2018).

One explanation for the above-described results is the positive relationship between post-secondary educations and accessibility of jobs with higher salaries. The U.S Bureau of Education reports that the median income of those with a bachelor's degree is 64% higher than that for those with a high school diploma. This difference increases further with higher levels of education. Higher income in adulthood increases accessibility to many items that contribute to a healthier lifestyle, such as healthy food, secure housing, protection against environmental shocks, and better healthcare (Bloom, 2005). Because of these and other factors, higher salary jobs tend to increase the employees' resilience to health setbacks (Bloom, 2005).

Mirowsky and Ross argue that although education's economic benefits strongly affect health, education itself is the main factor to a healthy life (Mirowsky, 2015). They propose that a college education helps people overcome the 'default American lifestyle;' which they describe as the consumption of engineered non-nutritious food, non-stimulating or unsatisfying employment, and reactive, rather than proactive, health related actions. This healthy lifestyle may be more influenced by the insight and knowledge provided by education to override this unhealthy standard of life (Mirowsky, 2015). In addition, higher levels of education lead to creative and fulfilling careers that not only might provide for better mental health but also a sense of control and optimism that inspires healthier choices (Mirowsky, 2015). This effect can be observed in the positive relationship between education levels and healthy behaviors (Cutler, 2006).

Because of the statistical and public health significance of the relationship between education and health, access to proper education could be seen as a health equity issue. This premise motivates the importance of the public education system in providing quality education equally to all of the residents of the associated jurisdiction. These public health concerns further motivate the need to monitor the performance of public schools and set goals for measurable

quality education to all students. The following section focuses on historically underserved students, as they represent a particularly vulnerable population in terms of both health outcomes and access to quality education.

1.1 Historically Underserved Students

Historically underserved (HUG) students are students of color, students from low income families, and students who speak English as a second language. According to the National Center for Education Statistics (NCES), 50% of students in the public education system are students of minority backgrounds. This figure is projected to grow to 55% percent by 2026 (NCES: Racial/Ethnic Enrollment in Public Schools). Students of minority backgrounds tend to attend schools with a majority population of minority students, in fact, over 50% of students of Black, Hispanic, and Pacific Islander descent attend schools with 75% or more minority student enrollment (NCES: Racial/Ethnic Enrollment in Public Schools). Although improving, there is still a significant discrepancy in the graduation rates between White students and students of color. In 2017, the nationwide high school graduation rate for White students was 89% compared to 78% for Black students and 80% for Hispanic students. However, 90% of all high school seniors plan to go college (Green, 2006).

The gap in high school graduation rates between White students and students of color is generally closing over time (McFarland, 2018). Although, this is certainly a step in the right direction, it is not necessarily indicative of the difference in the quality of the education that is being provided to White students and students of color. Graduation requirements may differ between school districts and school districts who are focusing on increasing the graduation rates

in their schools may relax their requirements in order to give the appearance that they are progressing (Murane, 2006).

1.2 Statement of the Problem

The purpose of this study is to test whether there are differences in the percent of post-secondary bound graduates in public schools that serve higher proportions of students from HUGs than those that do not in Pennsylvania from 2008 to 2016. This study aims to model the distribution of percent college bound and total-postsecondary bound graduates in relationship to proportion of HUG students, and student socioeconomic status at an individual school level using longitudinal regression analysis. The results of this regression analysis will be used to quantify the relationship, if any, between proportion HUG students served and percent of college-bound graduates.

Furthermore, this study is to analyze the distribution of student race/ethnicities in Pennsylvania public schools using cluster analysis. The results of this cluster analysis will be used to investigate post-secondary bound graduation rates in schools based on their student demographics in more detail than a single HUG student variable.

2.0 Methods

2.1 Data Management

2.1.1 Data sources

Post-secondary bound graduation rates and charter school status for 1628 public schools in Pennsylvania that have a graduating 12th grade class were pulled from the Pennsylvania Department of Education (PDE) public database found on their website (www.education.pa.gov) and was accessed on February 16, 2020 . School demographic data (race and sex distribution) were pulled from National Center for Education Statistics (NCES) Common Core of Data (CCD) Database found on their website (www.nces.ed.gov/ccd/) and was accessed on February 17, 2020. Data from school years 2007/2008 to 2015/2016 were used. Both databases are available to the public and were created for internal and external analysis use. All source datasets were separated by school year.

Data from NCES required substantial cleanup. Data from certain years from the NCES database were stored in txt format while others were stored as csv. Variable names from the NCES datasets were somewhat inconsistent across all tables and required some standardization. All NCES data were converted to wide data format (if not already) and combined to form a collective wide-format table. The data was standardized using NCES-defined variables that appeared in most of the datasets. Variables were created for student count in each grade, gender and race combination both by school year and by school. For example, a high school would have 40 covariates (4 grades * 2 sexes * 5 race/ethnicity categories). In the same example, the number of

male Asian 10th graders would have its own variable (e.g. 10ASM). Grade-level and school-level totals for sex and race were calculated and added to dataset. Furthermore, grade-level and school-level gender and race/ethnicity proportions were calculated and added to dataset.

In addition to race and sex data, other variables from the NCES datasets were kept in final table. An indicator variable for schoolwide Title I status (1= Title I, 0 = non-Title I) by year was kept to account for general family income levels of students in school while modeling. An indicator variable for charter school status was also kept (1= charter school, 0= non-charter school).

All data from the Pennsylvania Department of Education was available as Microsoft Excel Spreadsheets (.xlsx) and required some manual formatting in order to be readable by R. All manual formatting was purely for arrangement and no data were edited. Post-secondary bound graduation figures and rates were sourced from the PDE data and joined to the NCES data on school ID (assigned by state) and school year.

2.1.2 Post-secondary bound graduation rates

Two different rates are used in this analysis to measure a school's effectiveness in preparing students for a post-secondary education: college bound and total post-secondary bound graduation rates. College bound graduates (according to the Pennsylvania Department of Education) include any student attending a 2- or 4-year degree-granting college or university. A postsecondary graduate is any student defined as either meeting the definition of a college bound graduate or attending a specialized degree granting institution. A specialized degree granting institution is considered a non-degree granting institution (such as a trade school) or a specialized associate degree-granting institution (such as a medical assistant technician training program). The

denominator used for these rates is the total number of graduating 12th grade seniors for that academic year.

2.2 Statistical Analysis

2.2.1 Covariates

Race and ethnicity data were obtained from the NCES Common Core Database and categorized as White, Black, Asian, Hispanic/Latino, American Indian/Alaska Native, Pacific Islander, or two or more races. For analysis, race/ethnicity was further categorized by historically underserved (HUG) and non-historically underserved (non-HUG). Race/ethnicities categorized as HUG were Black, Hispanic/Latino, American Indian/Alaska Native and Pacific Islander. Races/ethnicities categorized as non-HUG were White and Asian. Percent HUG was then calculated by dividing the total number of HUG students by the total number of students enrolled in the school and multiplying by 100. Title I status was used as an indicator of overall general income level of student families. In order to qualify for Title I status, a school must have at least 40% of its student population be eligible for free and reduced lunch (U.S Dept. of Education, 2018). Charter School Status was provided by the NCES CCD Database.

2.2.2 Regression Analysis

2.2.2.1 Continuous outcomes

The two post-secondary bound graduation rates were treated as continuous percent ranging between 0 and 100%. Observations with missing outcomes were assumed to be missing completely at random (observations are considered at a school and year level) and were removed from the dataset.

2.2.2.2 Mixed effect regression models

Mixed effect regression models continuous outcomes on covariates of interest while accounting for changes in time. This allows for identification and quantification of fixed significant relationships between the outcome and covariates. The mixed effect regression model accounts for different individual intercepts and time trends, if they exist. The model achieves this by establishing an intercept parameter and slope parameter (if needed) for the change in outcome over time at both the individual and population level. A simple version of this model shows the outcome for individual *i* at time *j* based on the time effect for outcome y_{ij} (Hedeker, 2006):

$$y_{ij} = b_{0i} + b_1 t_{ij} + \epsilon_{ij}$$

$$b_{0i} = \beta_0 + v_{0i}$$

$$b_{1i} = \beta_1 + v_{1i}$$

Where b_{0i} represents the initial level for subject *i*, b_{1i} represents the slope for individual *i*, β_0 is the overall population intercept and β_1 is the overall population slope, v_{0i} is the intercept deviation for subject *i* and v_{1i} is the slope deviation for subject *i*. ϵ_{ij} represents the independent error term.

This model assumes that error term ϵ_{ij} is conditionally independent on v_{0i} and v_{1i} and normally distributed with mean 0 and variance σ^2. When there are two random individual-specific effects, the population distribution of the intercept and slope deviation are assumed to be bivariate normal with mean zero and variance-covariance matrix $\Sigma_v = \begin{bmatrix} \sigma^2_{v_0} & \sigma_{v_0 v_1} \\ \sigma_{v_0 v_1} & \sigma^2_{v_1} \end{bmatrix}$. The model also assumes that the change in outcome over time is linear (Hedeker, 2006).

When additional covariates are introduced to the model, the equation above can be re-written in matrix form as follows:

$$y_i = X_i\beta + Z_i v_i + \epsilon_i$$

Where X_i is the $n_i \times p$ covariate matrix, β is a $p \times 1$ vector of fixed regression parameters, Z_i is the $n_i \times r$ design matrix for the random effects and v_i is a $r \times 1$ vector of random individual effects and ϵ_i is the $n_i \times 1$ error vector. The assumptions for the random effects and errors are (Hedeker, 2006):

$$\epsilon_i \sim \mathrm{N}\left(0, \sigma^2 I_{n_i}\right),$$

$$v_i \sim N(0, \Sigma_v).$$

When the mixed effect model is fit for a set of observations, values in the β vector of fixed regression parameters can be interpreted as the fixed effect of the covariate of interest x_i on the average outcome *y* if the covariate is discrete, or per one-unit change in x_j if continuous.

Mixed effect regression models were used to model post-secondary bound graduation rates on covariates of interest while accounting for changes over time. This method allows us to utilize 8 years of data in hopes of modeling the underlying trend more accurately and allows the identification of possible time effects.

We fit a random effects models with a random intercept as well as models with both a random intercept and random slope at the individual school level. A random intercept model was used to assess the fixed effects of the covariates while accounting between school variability of the outcomes while the random effects model was used to assess the fixed effects of the covariates while accounting for between school variability in the outcomes and change in the outcomes over time (if any). Figure 1 shows the hierarchy structure and estimated count of each assuming no missing data.

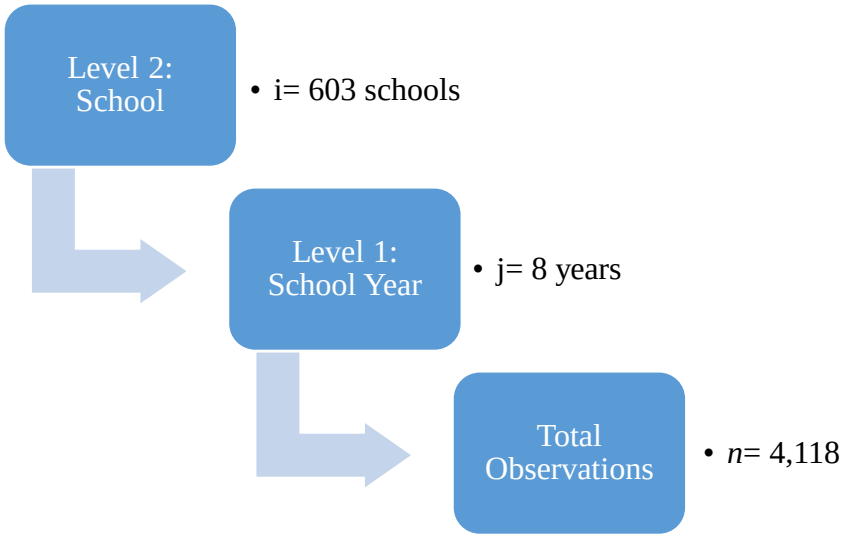

Figure 1 Hierarchical structure of mixed effect model

Title I status and charter school status were treated as time constant fixed effects while percent HUG was treated as a time-varying covariate.

$$y_{ij} = \beta_0 + \beta_1 year_j + \beta_2 HUG_{ij} + \beta_3 TitleI_i + \beta_4 CHRT_i + \zeta_{0i} + [\zeta_{1i} year_{ij}] + \epsilon_{ij}$$

Where:

y_{ij} = post-secondary graduation rate of interest for school *i* during year *j*

β_1 = overall mean effect of school year *j* on post-secondary graduation rate

HUG_{ij} = percent HUG student population for school *i* during year *j*

$\beta_2 =$ overall mean effect of percent HUG on outcome

$TitleI_i =$ Title I status for school *i* (1=Title 1, 0=non-Title 1)

$\beta_3 =$ overall mean effect of Title I status on post-secondary graduation rate

$CHRT_i =$ Charter school status for school *i* (1=charter school, 0=non-charter school)

$\beta_4 =$ overall mean effect of charter school status on post-secondary graduation rate

$\zeta_{0i} =$ school *i* intercept deviation

$\zeta_{1i} =$ school *i* slope deviation (not included in only random intercept models)

$\epsilon_{ij} =$ error term for school *i* during year *j*

Assuming:

$\zeta_{0j} \sim N(0, \psi_{00})$

$\zeta_{1j} \sim N(0, \psi_{11})$

$Cov(\zeta_{0j}, \zeta_{1j}) = \psi_{01})$

$\epsilon_{ij} \sim N(0, \theta)$

In addition, interaction between all covariates, including school year, were tested for significance in either model in order to assess changes in the covariates over time or whether there are interactions between covariates. Predicted residuals, random intercepts and random slopes were plotted to check normality according to model assumptions.

2.2.2.3 Mixed-effects polynomial regression models

It is often too simplistic to assume the change across time is linear, particularly for outcomes that are a proportion or a percent because floor or ceiling effects can occur (Hedeker, 2006). A curvilinear trend model would allow for this leveling off, as well as handle any

accelerated changes over time. The model achieves this with the addition of quadratic terms of time to the model. For example, the simple model considering only the effect of time on outcome y_{ij} can be written as:

$$y_{ij} = b_{0i} + b_{1i}t_{ij} + b_2 t_{ij}^2 + \epsilon_{ij}$$

Using multiples of the time covariate *t* can often result in collinearity problems which can be avoided by representing the polynomials in orthogonal form (Bock, 1975). This is accomplished by expressing the time covariate and associated polynomials in centered form.

Because our outcome is limited between 0% and 100% it is possible that a floor or ceiling effect could occur. For example, a school may reach close to 100% post-secondary bound graduates in the middle of the study window and remain there until the end of the study window. The result of a flooring or ceiling effect is a possible non-linear change across time of our outcome (Hedeker, 2006). Therefore, we also fit a curvilinear trend model regression model in addition to the linear mixed-effects model.

To avoid collinearity issues when adding polynomials of year, we centered the year and associated polynomial terms (Bock, 1975). Varying degree polynomial terms were added to the mixed effect model in a forward stepwise matter (starting with quadratic, then cubic, etc.). AIC, BIC and adjusted-R^2 was used to determine the optimal degree polynomial term to include.

2.2.2.4 Test for statistical significance and quantify effect on outcome

Coefficients of the covariate effects were tested for significance by two-tailed t-test with degrees of freedom equal to *n* -k-1 (where k is the number of variables) against the null hypothesis of $\beta_j = 0$. A p-value of 0.05 or less was considered statistically significant. The test statistic, *t*, was calculated by dividing the coefficient for covariate *j* (β_j) by the standard error (SE) of β_j. The effect

of the covariates was quantified from the relative mean-effect coefficients derived from the most accurate model which determined by evaluation of AIC, BIC and adjusted-R^2.

For continuous covariates (HUG), the interpretation of the mean-effect coefficient can be interpreted as a change in percent of post-secondary bound graduates per a one percent change in student HUG population. For binary covariates (charter status, title I status), the mean-effect coefficient can be interpreted as the difference in percent of post-secondary bound graduates between schools having with that covariate status and those that do not.

2.2.2.5 Categorical percent HUG covariate

In certain situations, continuous variables may be better represented as categorical variables to improve the interpretability of the effect of the variable on the outcome (DeCoster, 2011). We refit the final models from 2.2.2.3 using a categorical percent HUG variable that represents the quartile of a school's percent student HUG population in place of the continuous percent HUG variable.

2.2.2.6 Identifying outliers

Outliers were identified by plotting standardized residuals against fitted values from the final model. Observations with standardized residuals less than -2 or greater than 2 were considered outliers and identified. Cook's distance for each observation was then calculated and plotted to ensure that previously identified outliers were not too influential on the model, thus reducing the overall accuracy.

2.2.3 Cluster analysis

2.2.3.1 K-Means clustering

K-means clustering is a popular and effective method of unsupervised machine learning. It is used to partition a dataset into *K* distinct, non-overlapping clusters (James, 2013). Subgroups derived from K-means clustering can be explored to not only understand the data better, but if the data can be clustered into practical subgroups, these subgroup assignments can be used as a covariate when fitting statistical models.

The K-means algorithm works to minimize some within-cluster measure, $W(C_k)$, of the amount by which observations within a cluster differ from each other. K is the user-defined number of clusters.

$$\underset{C_1,\ldots,C_k}{\text{minimize}}\left\{\sum_{k=1}^{K} W(C_k)\right\}$$

Typically, squared Euclidean distance is used as this measure and is what will be used in this study.

$$W(C_K) = \frac{1}{|C_k|}\sum_{i,i'\in C_k}\sum_{j=1}^{p}\left(x_{ij} - x_{i'j}\right)^2$$

The K-means algorithm has essentially 2 steps when using Euclidian distance as the within cluster distance measure (James, 2013):

1. Each observation is randomly assigned a cluster number from 1 to K, where K is defined by the user.
2. The following is repeated until the individual cluster assignments do not change.
    a. A centroid is computed for each of the K clusters by calculating a vector of length *p* of the means of each feature observed in the *k*th cluster.

b. Assign each observation to the cluster whose centroid is the ‘closest’ as defined by Euclidean distance.

2.2.3.2 Within cluster sum of squares

A popular method to determine the optimal number of clusters (*K*) the data should be partitioned into is to optimize cluster size and within-cluster sum of squares (WSS). To do this optimization, the K-means algorithm must be used on the data for multiple values of *K*, typically 2- 20 times. For each value of *K*, the WSS should be calculated and plotted against *K*. The optimal number of clusters can then be determined by identifying the ‘elbow’ of the resulting curve or where the WSS stoops decreasing rapidly and begins leveling out (Kodinariya, 2013).

$$WSS = \sum_{k=1}^{K} \sum_{i \in S_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})$$

2.2.3.3 Clustering schools by race/ethnicity distributions

To expand on our model, we used K-means clustering to cluster the schools based on the student race/ethnicity distributions of each school’s student population. The mean proportion of students of each race/ethnicity were calculated for the all available time periods for each school. This averaged data was then used to cluster the schools using 2 to 20 centers (i.e. the K-means algorithm was used to separate the data into 2 clusters, then 3 clusters, then 4 clusters etc.); this range was chosen to ensure cluster numbers were assesses while being computationally efficient. The algorithm was used on the data 18 times with cluster amounts ranging from 2 to 20. The WSS was calculated at each number of clusters and then plotted against the number of clusters. The elbow method was then used to determine the optimal number of clusters. The distributions of

student race/ethnicity of schools in each cluster was then analyzed to understand how the representation of each cluster.

2.2.3.4 Modeling on clusters

The clusters were treated as a categorical variable representing racial/gender distributions. This categorical model was then used to replace the continuous student HUG proportion covariate in the linear mixed model. Then, a model was fit for each round of clustering. For example, the model for the data obtained from clustering on 3 clusters would look like:

$$y_{ij} = \beta_0 + \beta_1 year_{ij} + \beta_2 CLST2_i + \beta_3 CLST3_i + \beta_4 TitleI_i + \beta_5 CHRT_i + \zeta_{0i} + \zeta_{1j} year_{ij} + \epsilon_{ij}$$

Where:

$CLST2_i$ = indicator variable for school *i* if in cluster 2

$CLST3_i$ = indicator variable for school *i* if in cluster 3

The best fitting model based on number of clusters was determined by AIC, BIC and adjusted-R^2 evaluation.

For this study, mixed modeling was done in R (Version 3.6.2) using the package 'nlme' (Pinheiro et.al). Visualization were created in R using the 'ggplot2' package (Wickham et.al). K-means clustering was done using base R.

3.0 Results

3.1 Summary Statistics

After the data were cleaned, there were 4,118 total years of data for 601 individual schools. Table 1 displays the number of schools per number of years (1 year up to 8 years) of complete data available. There was an average of 6.85 years of data available per school. Out of the 601 schools, 199 were missing at least one year's worth of data because either the school only existed for a portion of the study timeframe (2008 - 20015) or the school had not reported data for that school year.

Table 1 Number of schools per number of years of data available

No. of years of data	No. of Schools (%)
8	402 (66.89%)
7	39 (6.50%)
6	38 (6.32%)
5	29 (4.83%)
4	31 (5.16%)
3	28 (4.66%)
2	14 (2.33%)
1	20 (3.33%)
Total	601

3.1.1 Race

The race category labels were used directly from the NCES Common Core Database. On average, schools had 74% White students, 16% Black students, 6% Hispanic students, 2% Asian students and less than 1% of the other racial categories. However, there was a large variation in the racial distribution across schools was observed, particularly in percent of White, Black and

Hispanic students. This implies variability of the racial distributions across schools and that there may be schools with significantly more students of one race/ethnicity than others. This was further explored in the cluster analysis.

3.1.1.1 Percent Historically Underserved

The distribution of race/ethnicity was then summarized by categorizing race/ethnicity as either HUG or not. Black, American Indian, Hispanic, Hawaiian/Pacific Islander were classified as HUG while White and Asian were not. Figure 2 shows the overall distribution of percent HUG across all schools by school year. There are no obvious changes in the distribution of HUG student percent over the course of the study timeframe.

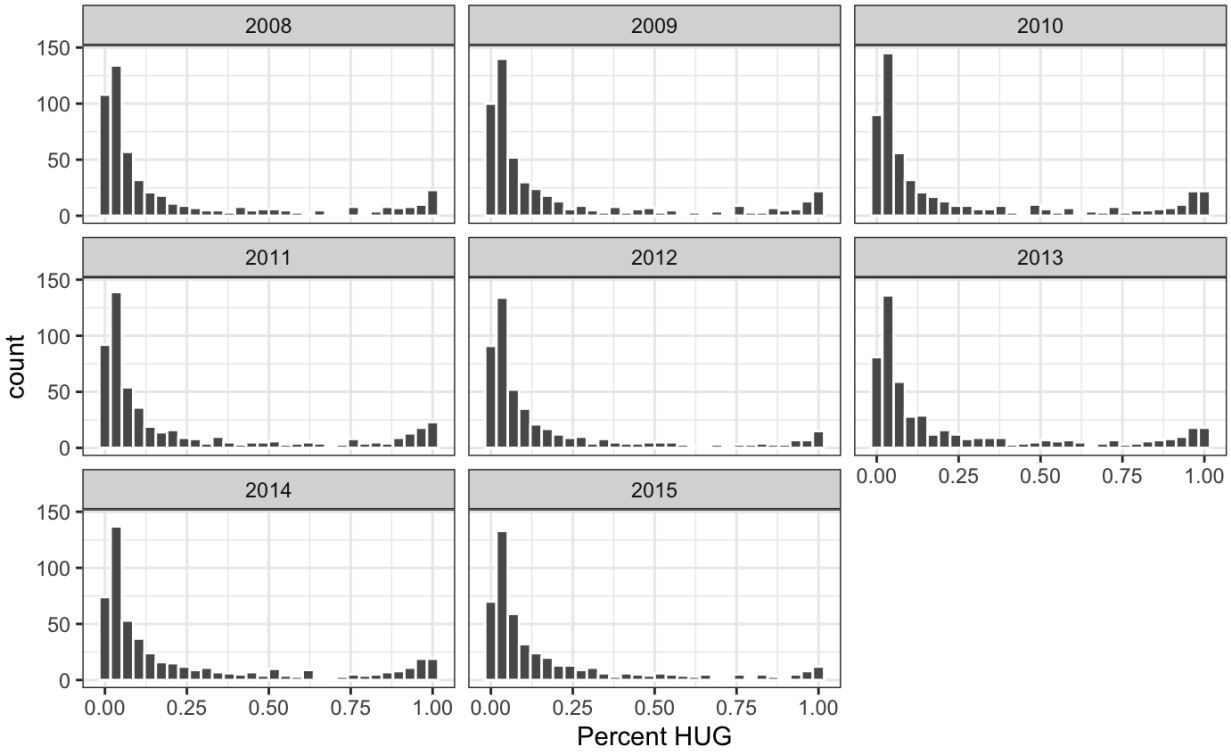

Figure 2 Distribution of school-level percent HUG students by school year

3.1.2 Title I and Charter School Status

No school had a change in school-wide title I status over the course of the study timeframe. Overall, there were 177 (29.45%) schools who held school-wide Title I status and 424 (70.55%) who did not. In addition, 74 (12.31%) of school were classified as charter schools while 527 (87.69%) of schools were not.

3.1.3 Post-Secondary Bound Graduation Rates

Figure 3 shows the overall distribution of post-secondary bound, college bound and specialized degree bound graduation rates for all schools over all time periods. Both total post-secondary bound and college bound graduation rates were bell shaped and symmetrically distributed and centered around means 73.25% and 70.10%, respectively. Specialized degree bound was very right skewed with a mean of 3.55%; for most schools, the majority of the post-secondary bound graduates went to 2- or 4-year colleges. There is some slight skewness in the total post-secondary bound and college bound rates due to outlying values below 30%. There was little variation in the distribution of each outcome over time.

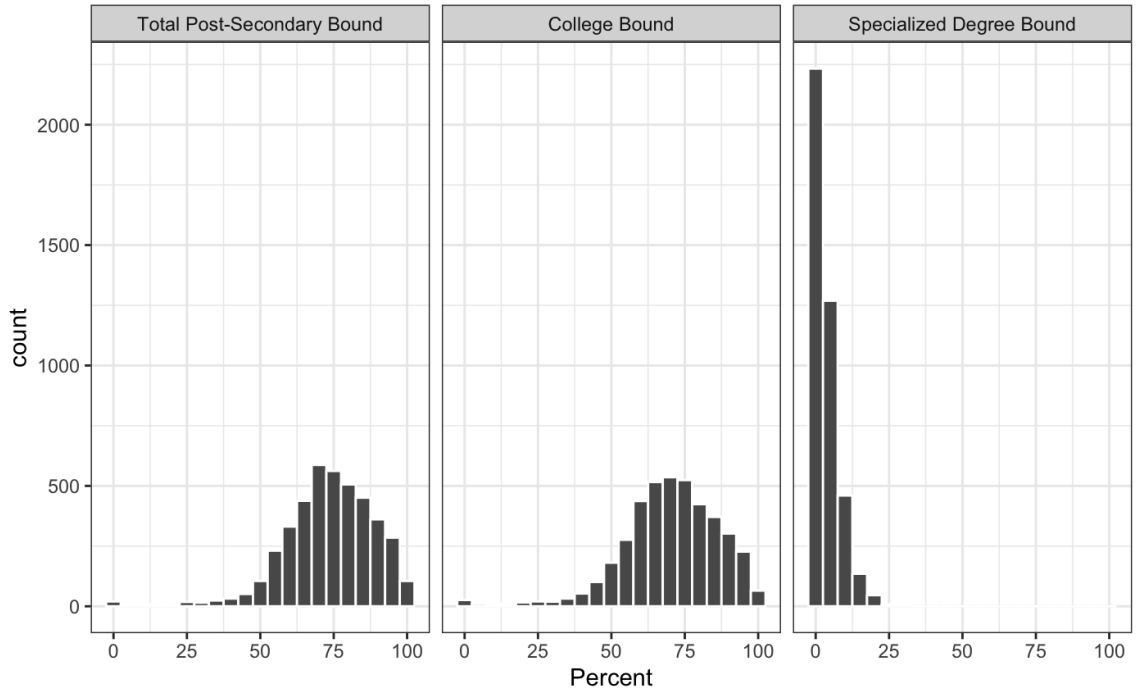

Figure 3 Overall distributions of total post-secondary bound, college bound and specialized degree bound distributions (all schools, over entire study timeframe)

3.2 Longitudinal Linear Regression Models

3.2.1 Modeling with Continuous HUG Covariate

Continuous percent HUG, Title I status and charter status were used as covariates to model total post-secondary bound and college bound graduation rates over time. Both random effect and random intercept models were tested. A random intercept model was used to assess the fixed effects of the covariates while accounting between school variability of the outcomes while the random effects model was used to assess the fixed effects of the covariates while accounting for

between school variability in the outcomes and change in the outcomes over time (if any). Stepwise variable selection was used to identify statistically significant covariates and to build the best fitting model. Percent HUG and Title I status were identified as statistically significant covariates. The random intercept model modeled the data as accurately as the random effects model according to AIC and BIC. In addition, the likelihood ratio testing the assumption that the random intercept model was nested in the random coefficient model was not statistically significant. Therefore, the random intercept model with HUG percent, Title I status and school year was chosen as the final model for both total post-secondary bound and college bound graduation rate outcomes. Table 2 describes the coefficients from the final models. The final models for both outcomes met all of the assumptions for a linear random intercept mixed model.

Table 2 Coefficients of random intercept model with continuous HUG, Title I status and School Year

Covariate		Coefficient [95% CI]	p-value
Total Post-Secondary Bound			
	Percent HUG	0.066 [0.03, 0.11]	0.001
	Title I Status	-6.70[-9.81, -3.59]	<0.0001
	School Year	-0.61[-0.74, -0.48]	< 0.0001
College Bound			
	Percent HUG	0.060 [0.018, 0.10]	0.006
	Title I Status	-7.36 [-10.73, -3.99]	<0.0001
	School Year	-0.54 [-0.67, -0.42]	<0.0001

The model presented in Table 2 shows total post-secondary bound and college bound graduation rates were estimated to increase 0.07 percentage points (pp) and 0.06pp, respectively, per 1pp increase in HUG, all other covariates constant. Schools with school-wide Title I status had significantly lower post-secondary bound (-6.70pp) and college bound (-7.36pp) graduation rates than schools who did not, all else constant. In addition, there was slight decrease in average total post-secondary bound and average college bound rates over time (0.61pp and 0.54pp annually respectively, adjusting for other covariates).

3.2.1.1 Mixed-effects polynomial regression models

To test whether there was a floor or ceiling effect, both quadratic and cubic variations of centered time (school year) were added to the random coefficient and random intercept models. School year was centered to avoid collinearity in the associated polynomials. The relative squared and cubed time variables were not statistically significant and did not increase the effectiveness of the model based on AIC and BIC and therefore final random effects models were kept the same (results not shown).

3.2.2 Modeling with Categorical HUG Covariate

As seen in Figure 2, The distribution of percent HUG is nearly U-Shaped with a majority of the concentrated towards 0 or 100%. We can see in Figure 4 that the schools are fairly concentrated in certain parts of the graph and it is difficult to see the linear relationship between percent student HUG and post-secondary bound graduation rates. Because of this, a linear relationship may not be the most effective way of quantifying the effect of percent HUG on either outcome. Therefore, percent HUG was categorized based on the quartiles of the overall distribution (over all school years) of percent HUG as follows: Q1 *(0 -2.40%),* Q 2 *(2.41% - 6.84%),* Q3 *(6.85% - 30.709%),* Q4 *(30.71% - 100%).* New random intercept models were derived with categorical percent HUG covariate replacing continuous percent HUG. The 0 – 2.40% quartile was used as the reference group in the models. Table 3 describes the coefficients of these models.

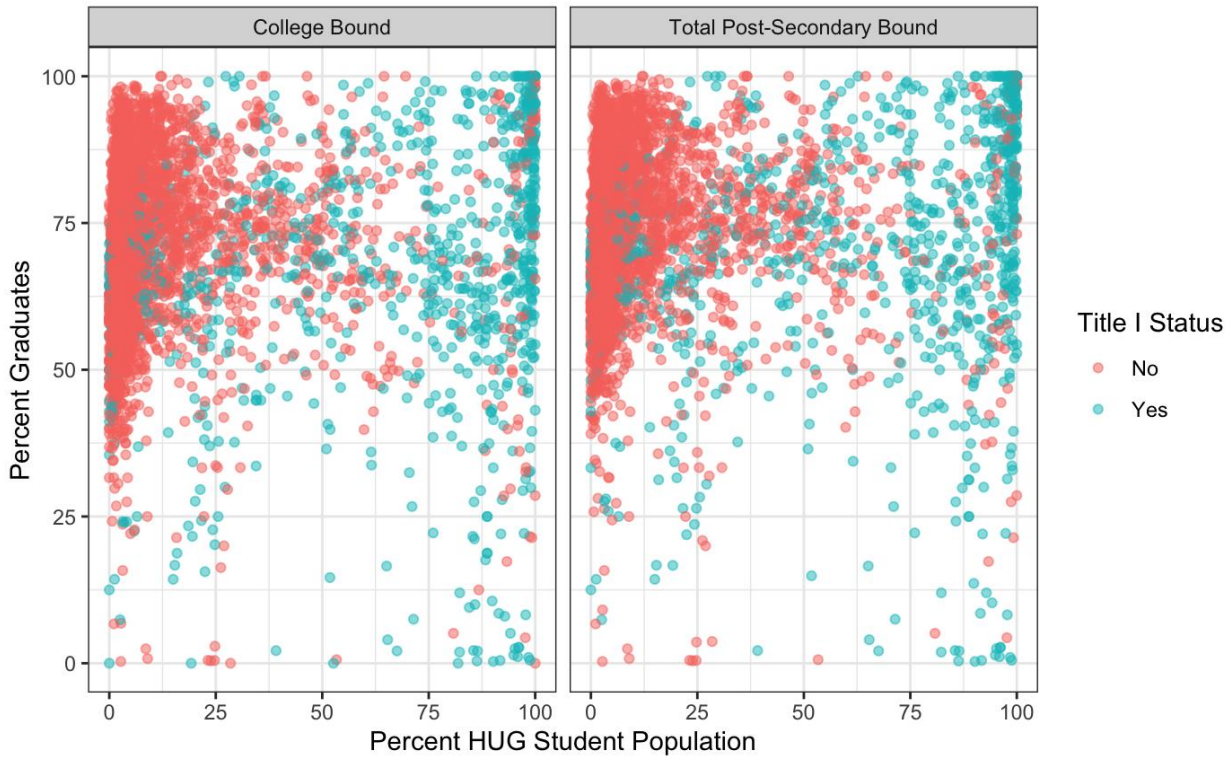

Figure 4 Plot of post-secondary outcomes against percent student HUG colored by Title I status

The models in Table 3 show that both school-wide Title I status and school year have nearly identical effects when being modeled with either continuous percent HUG (as presented in Table 2) or categorical percent HUG (as presented in Table 3). Schools in the third and fourth quartile of the percent HUG distribution had estimated total post-secondary and college-bound graduation rates 2.5pp and 5.5pp higher than those for schools with percent HUG in the first quartile, all else constant. There was not a statistically significant difference in school with percent HUG in the second quartile and those with percent HUG in the first quartile. A likelihood ratio test was used to test the overall statistical significance of the categorical HUG covariates by testing the models from Table 3 against random intercept models (for each outcome) with only the Title I status and school year covariates. The likelihood ratio test for the total post-secondary bound model had a p-

value of 0.0004 and the test for college bound model had a p-value 0.0005. As result, the categorical percent HUG covariate is statistically significant in each model for either outcome.

Table 3 Coefficients of random intercept model with categorical HUG, school-wide Title I status and school year

Covariate		Coefficient [95% CI]	p-value
Total Post-Secondary Bound			
	Percent HUG Q2	0.34 [-0.92, 1.60]	0.60
	Percent HUG Q3	2.33 [0.41, 4.24]	0.017
	Percent HUG Q4	5.51 [2.88, 8.13]	<0.0001
	Title I Status	-6.38[-9.18, -3.58]	<0.0001
	School Year	-0.63[-0.75, -0.50]	< 0.0001
College Bound Rate			
	Percent HUG Q2	0.25 [-1.02, 1.51]	0.70
	Percent HUG Q3	2.50 [0.54, 4.46]	0.012
	Percent HUG Q4	5.53 [2.79, 8.28]	<0.0001
	Title I Status	-7.31[-10.33, -4.299]	<0.0001
	School Year	-0.56 [-0.69, -0.43]	<0.0001

Quartile ranges: Q1- 0% to 2.4%, Q2- 2.41% to 6.84%, Q3- 6.85% to 30.709%, Q4- 30.71% to 100%

3.3 Cluster Analysis

3.3.1 Determining the Ideal Number of Clusters

Using school-level proportions of student race (White, Black, Hispanic, Asian, Hawaiian/Pacific Islander, American Indian and two or more races) as covariates, the K-means clustering algorithm was used on the data 18 times with number of clusters ranging from 2 to 20. Figure 4 shows the relationship of within sum of squares to the number of clusters. Using this plot,

the optimal number of clusters appears to be 4 using the elbow method (Kodinariya, 2013). As shown in Figure 4, the value of WSS drops substantially up 4 clusters and then begins to plateau.

Results for k=3, k=4, k=5 clusters were analyzed, particularly the distribution of student/race ethnicity of schools in each cluster. The results for k=3 clusters were determined to have too high of variability in regard to student race/ethnicity distributions and grouped together schools that we thought were too dissimilar in one cluster. The results for k=4 clustering were similar to the results of the k=3 clustering for 2 of the clusters, however, the extra center in the k=4 clustering allowed further stratification of the third cluster (the one with high-variability), reveling a distinct fourth cluster representing schools with a large majority of Hispanic students. The results of the k=5 clustering were very similar to the results of the k=4 clustering; however we felt the k=5 clustering unnecessarily stratified schools with large majorities of White students. Therefore, we decided that 4 clusters were the optimal amount to present and further investigate.

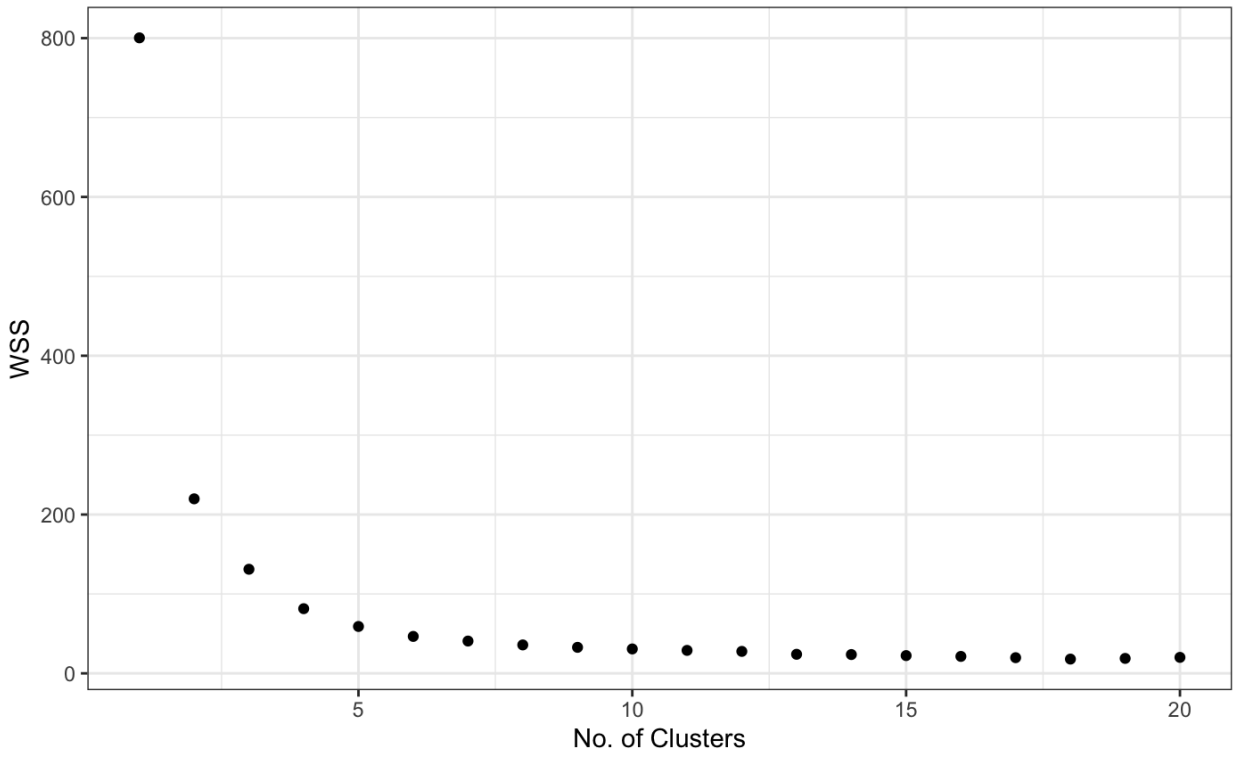

Figure 5 Number of clusters in k-means algorithm vs. WSS in order to determine optimal number of clusters

3.3.2 Cluster Characteristics

The distributions of student race/ethnicity in the resulting four clusters were identified in order to understand the school representation in each cluster. Table 4 shows the mean proportion of each race/ethnicity category and percent of schools with Title I status for each cluster. There appears to be differences in the racial distributions of each cluster, particularly in the proportion of White, Black and Hispanic students.

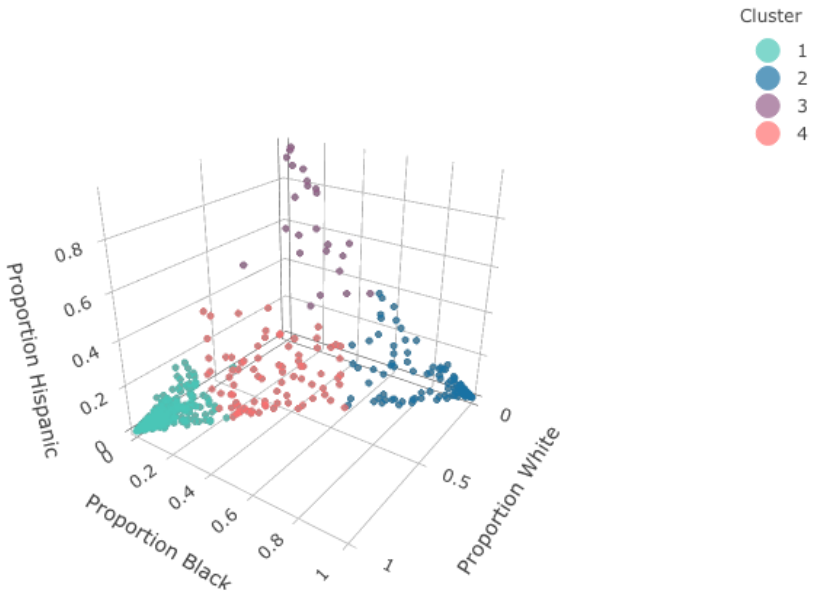

Figure 6 3D scatterplot of proportion of White, Black and Hispanic students by school (averaged over study period), colored by assigned cluster

The clusters represent four fairly distinct school-level race/ethnicity distributions as seen in table 4. This implies that the cluster labels would be useful as covariate in a model replacing percent HUG as each cluster represents a distinct race/ethnicity distribution. The first cluster contains schools with a large majority of White students and very small proportion of students from other/race ethnicities. The second cluster contains schools with a large majority of Black students with some variability in White and Hispanic proportions. The third cluster contained schools with a majority of Hispanic students with some variability in White and Black proportions. The fourth cluster contained schools with a more diverse student body, but still containing a majority of White students. There was substantial variability in all races in this fourth cluster. Cluster 1 was largest cluster as 65% of schools were assigned to cluster 1 while cluster 3 is the smallest cluster with only 4% of the schools were assigned to it. Figure 5 highlights this further

with a 3D scatterplot comparing the mean proportion over the time periods of these race/ethnicities in each school for each cluster.

As seen in Table 4, a large majority of the schools in cluster 1 and a small majority in cluster 3 did not have school-wide Title I status, whereas a majority of the schools in clusters 2 and 3 did. As a result of the larger proportions of Black and Hispanic students in clusters 2 and 3 (respectively), schools in these clusters had a higher average percent of HUG student population at 89% and 88% (respectively) compared to schools in cluster 1 (6%) and cluster 4 (40%).

Table 4 Average student race/ethnicity proportions, percent Title I status and average post-secondary bound graduation rates in each cluster in each cluster

	Cluster			
	1	2	3	4
Number of schools in cluster	390	105	24	82
Mean(sd) Percent of School-Level Student Population				
White	91.89(7.3)	7.09(9.11)	9.96(10.11)	51.57(13.65)
Black	3.32(3.79)	82.71(15.2)	20.16(13.01)	29(12.96)
Hispanic	2.45(3.45)	6.5(8.5)	67.55(17.74)	11.68(10.49)
Asian	1.46(2.38)	2.63(5.02)	1.41(1.61)	4.58(6.4)
Hawaiian/ Pacific Islander	0.04(0.14)	0.02(0.09)	0.04(0.11)	0.08(0.37)
American Indian/ Alaskan Native	0.15(0.28)	0.15(0.26)	0.07(0.11)	0.16(0.29)
HUG	5.97(5.84)	89.38(11.2)	87.81(10.62)	40.93(12.85)
Percent of Schools				
School-wide TITLE I Status	0.096	0.79	0.812	0.34
Mean(sd) School-level Percent Graduates				
College Bound	70.07(14.59)	71.17(23.41)	61.82(21.25)	71.52(16.59)
Total Post-Secondary Bound	73.12(13.73)	74.74(22.35)	67.12(21.61)	74.2(16.04)

Table 4 also describes the distribution of the total post-secondary bound and college bound graduation rates, respectively, for each cluster. Clusters 1, 2 and 4 had similar mean total post-secondary bound graduation rates around 74%, however, there was more variability in cluster

2. Cluster 3 (schools with a majority Hispanic student population) had the lowest average total post-secondary bound graduation rate at 67%. Total post-secondary bound rates in schools in cluster 1 (schools with a majority White student population) were normally distributed around mean 73.13%. In cluster 2 (majority Black student population), the total post-secondary bound rates were skewed left with mean 74.74%. Cluster 4 (schools with a more diverse student population) were normally distributed around mean 74.19%. Similar trends were observed in college bound graduation rates in each cluster however there was slightly more variability in this rate in cluster 2.

3.3.3 Longitudinal regression model using cluster as covariate

To address the clustering of the racial distributions in the student population across schools, the random intercept models were refit using the categorical cluster variable replacing any HUG related variable from previous models. Table 6 describes the coefficients resulting from these models.

In Table 6, school-wide Title I status and school year had similar effects as in previous models. Schools in cluster 2 had the highest total post-secondary bound and college bound graduations rates, 4.38pp higher on average than that for cluster 1 with all else constant. Schools in cluster 3 had the lowest rates in both outcomes, 1.1pp lower total post-secondary bound graduates on average and 3.55pp less college bound graduates on average than in cluster 1. There was not a statistically significant difference in either rate in cluster 3 or cluster 4 compared to cluster 1. The overall significance of the cluster variable was tested by likelihood ratio tests with relevant models of both outcomes using only Title I status and school year as covariates. The resulting p-values were 0.023 for the total post-secondary bound model and 0.005 for the college

bound model. This model reveals that the trend seen in previous models of increased post-secondary bound graduation rates in schools with higher proportion of HUG students may only apply to schools with a majority Black student population (schools in cluster 2) and not in schools with a majority Hispanic population (schools in cluster 3) or schools with a diverse student population (schools in cluster 4).

Table 5 Coefficients of random intercept model with cluster covariate (cluster 1 is reference), school-wide Title I status and school year

Covariate	Coefficient [95% CI]	p-value
Total – Post Secondary Bound		
Cluster 2	4.38 [1.20, 7.56]	0.01
Cluster 3	-1.10 [-6.30, 4.11]	0.20
Cluster 4	1.41 [-0.82, 3.63]	0.098
Title I Status	-5.36[-8.31, -2.40]	0.00019
School Year	-0.60 [-0.72, -0.47]	< 0.0001
College Bound Rate		
Cluster 2	4.38 [1.03, 7.73]	0.0071
Cluster 3	-3.55 [-9.02, 1.92]	0.678
Cluster 4	1.93[-0.35, 4.22]	0.21
Title I Status	-6.06 [-9.23, -2.89]	0.0004
School Year	-0.53[-0.66, -0.41]	< 0.0001

4.0 Discussion

This analysis modelled the relationship between the race/ethnicity distribution of a school student population and post-secondary bound graduation rates using a random intercept linear model. There were not substantial between-school differences in the changes of either outcome over time, therefore a random intercept model was adequate to fit the data. Three separate measurements to quantify a school's race/ethnicity distribution were modeled along with time and school-wide Title I status: continuous percent historically underserved (HUG), categorized percent HUG and student race/ethnicity cluster.

The percent of HUG students in a school's student population was also found to have statistically significant relationship with both total post-secondary bound and college bound graduation rates. Accounting for school-wide Title I status and school year, there was an observed increase of 0.07pp in total post-secondary bound and 0.06pp in college bound graduation rates per l% increase in percent student HUG. This implies that schools with higher percentages of a HUG students will have more of their graduates attend some form of post-secondary education including 2- or 4- year universities. However, we also observed that a majority of schools had either small percentages of HUG students (<5%) or large percentages of HUG students (>80%) with fewer schools in the middle of the range. This led us to believe that using a continuous percent HUG may not be the optimal method to quantify this relationship.

A categorical percent HUG variable based on the quartile of the student population was used in the model in place of the continuous percent HUG variable. In this model, the effect school-wide Title I status and school year did not change. However, the effect of percent HUG was more interpretable. It was observed that schools with percent student HUG in the fourth quartile (>31%)

had on-average 5.5pp higher total post-secondary bound and college bound graduations accounting for school-wide Title I status and school year. The 5.5pp increase in post-secondary bound graduation rates is fairly substantial and implies that graduates from schools with higher proportion of HUG students are attending some form of post-secondary education at higher rates than graduates from schools with small proportions of HUG students.

Interestingly, the effect of HUG student percent we observed in Pennsylvania schools contradicts college enrollments rates by race/ethnicity seen across the United States. In 2017 41% of White students 65% of Asian students were enrolled in college compared to 36% of Black students and 36% of Hispanic students (NCES College Enrollment Rates, 2019). We speculate that this may be due to that fact that there a significantly more schools in Pennsylvania with low proportions of HUG students than schools with high proportions of HUG students (50% of Pennsylvania schools have less than 7% HUG student enrollment). Because of the difference in sample size, there is a possibility of higher variability in post-secondary bound graduation rates in schools with low proportions of HUG student than those with high proportions of HUG students. To further investigate the trend observed in the analysis, qualitative research may be necessary to gain better insight on why schools in Pennsylvania with higher proportions of HUG students have higher post-secondary bound graduations rates, particularly in Title I schools. In addition, it would be beneficial to model the data from states with higher populations of HUG students to see if similar trends exist.

School-wide Title I status was observed to have a statistically significant effect on both total post-secondary and college bound graduation rates. Because at least 40% of a school's student population must be considered as coming from low-income families (NCES Fast Facts, 2019) to qualify for school-wide Title I status. Therefore, it is likely that the relationship we observed in

our model is due to the increasing financial barrier to post-secondary education access. The average cost of college attendance has increased over 170% since 1980, adjusting for inflation (NCES Fast Facts, 2019) and need-based financial aid has not increased to keep up with the increasing cost of college (College Board, 2019). This results in students, especially from low-income families, needing to borrow more money to afford college (Reimherr, 2013). Because of the increasing need to use loans to pay for college, there is large inferred risk in attending college that could stop more students from low income families from enrolling in college (Lim, 2019).

We used cluster analysis to further explore the racial distributions of schools. It was observed that schools fell into 4 distinct clusters based on the distribution of the race/ethnicity of their student population. Schools either had a large majority of White students, a larger majority of Black students, a large majority of Hispanic students or a more diverse student population with a smaller majority of White students.

As result, schools falling into the clusters with a large majority of Black students or a large majority of Hispanic students had substantially higher proportions of HUG students. This led us to using the categorical cluster assignment in place of categorical HUG in the modelling. The results of this model showed that schools with a large majority of Black students had, on average, larger percentages (by 4.4pp) of graduates attending some form of post-secondary education compared to schools with a majority of White students. This agreed with the earlier models, however, schools with a majority of Hispanic students and schools with a more diverse student population did not have statistically significant difference in either post-secondary bound graduation rate compared to schools with a majority of White students. This contradicts the relationship found between post-secondary bound graduation rates and student percent HUG found

in earlier models and implies that this relationship may only be observed for schools with a majority population of Black students.

Only 4% of the schools in our population had large majorities of Hispanic students compared the 17% of schools that had large majorities of Black students. This might indicate that percent HUG is too generalizing when describing the trends found in the previous models and may not be the most representative measurement when describing post-secondary bound outcomes to all HUG students. In the models that used percent HUG covariates, the percent student HUG covariate is more representative of schools with large majorities of Black students and less representative of schools with larger majorities of Hispanic students. This suggests that when assessing education equality, race/ethnicity distributions should be considered when modeling outcomes in addition or in place of a generalizing HUG covariate.

In addition, modeling multiple race/ethnicities under one category may over/underestimate unique barriers one culture may experience that others would not. For example, we speculate that one reason students in schools with a majority Hispanic population may experience different educational barriers than students in schools with majority Black student population could be that there is larger proportion of Hispanic students who are English Language Learners (ELLs). In 2015, 29.8% of ELL students in the United States were Hispanic, compared 2.4% who were Black (NCES English Language Learners in Public Schools, 2019). ELL students may face extra challenges to accessing college such as access to college preparatory classes (Perez, 2016).

We propose that additional qualitative research may be necessary to understand why there is a difference in post-secondary bound graduation rates in Pennsylvania schools with majority Black student population and schools with a majority White student population, particularly in schools with school-wide Title I status. A limitation in our study is that we did not consider overall

high school graduation rates in our models. It would be beneficial to model high school graduation rates as well as use high school graduation rates as a covariate when modeling post-secondary bound graduation rates allowing the model to account for the proportion of students that actually graduated. This would also help to identify any relationships between high school graduation rates and post-secondary bound graduation rates. Alternatively, the post-secondary bound graduation rates could be re-calculated with all seniors eligible to graduate as the denominator rather than graduating seniors only. Then model the re-calculated rates as the outcomes with similar covariates used in this study to observe any changes.

In conclusion, higher education rates have been observed to correlate with better health outcomes (Zajocova, 2018). It is the duty of the public education system to ensure that all students are prepared and provided resources to increase access to higher education levels. In order to ensure that every student is receiving equal opportunities for higher education, research must be conducted at multiple stages of a student's education in order to find areas that need support. One stage we analyzed was post-secondary bound rates at high school graduation. We found there is evidence that schools with different distributions of student race/ethnicities and socioeconomic status are having different rates of post-secondary bound graduates in Pennsylvania. We also found evidence that there may be a need to investigate schools with different student populations in a closer and qualitative manner to find areas that need support.

Appendix A R Code

Appendix A.1 Select R code for models and figures

Section 3.2.1 Random intercept models w/ continuous percent HUG:

```
#college bound grad rate outcome
ri.college <- lmer(college_bound.p ~ HUG_percent + SURVYEAR +STITLI_OVERALL
+ (1|SCHNO), df.arm1, REML = 0)

#total post-secondary bound grad rate
ri.total <- lmer(total_postsecondary_bound.p ~ HUG_percent + SURVYEAR
+STITLI_OVERALL  + (1|SCHNO), df.arm1, REML = 0)
```

Section 3.2.2 Random intercept models w/ categorical percent HUG:

```
#creating categorical variables
df.arm2$HUG_catq[df.arm2$HUG_prop < .0241] <- 1
df.arm2$HUG_catq[ df.arm2$HUG_prop >=.0241  & df.arm2$HUG_prop < .06853] <- 2
df.arm2$HUG_catq[df.arm2$HUG_prop >= 0.06853 & df.arm2$HUG_prop < .30710] <-
3
df.arm2$HUG_catq[df.arm2$HUG_prop >= .30710] <- 4

#modeling
ri.tot.c <- lmer(total_postsecondary_bound.p ~ factor(HUG_catq) + SURVYEAR
+STITLI_OVERALL  + (1|SCHNO), df.arm2, REML = 0)

ri.col.c <- lmer(college_bound.p ~ factor(HUG_catq) + SURVYEAR
+STITLI_OVERALL  + (1|SCHNO), df.arm2, REML = 0)

#likelihood ratio test testing the significance of categorical variable
#making null models
model.0.t <- lmer(total_postsecondary_bound.p ~ SURVYEAR + STITLI_OVERALL +
(1|SCHNO), df.arm2, REML =0)
model.0.c <- lmer(college_bound.p ~ SURVYEAR + STITLI_OVERALL + (1|SCHNO),
df.arm2, REML =0)
anova(model.0.t, ri.tot.c, test = 'LRT')
anova(model.0.c, ri.col.c, test = 'LRT')
```

Section 3.3.1 Cluster analysis:

```
#clustering and calculating WSS for k =2 through k=20
wss <- rep(NA,20)
clusters <- rep(NA,20)
for(k in 1:20){
  a <- kmeans(x = cluster_data[,-c(1:8)], centers =  k)
  wss[k] <- a$tot.withinss
```

```
  clusters[k] <- k
}


#3,4 or 5 clusters
k.4 <- kmeans(x = cluster_data[,-c(1:8)], centers =  4)
k.3 <- kmeans(x = cluster_data[,-c(1:8)], centers =  3)
k.5 <- kmeans(x = cluster_data[,-c(1:8)], centers =  5)

df.4cluster <- cbind.data.frame(cluster_data, cluster = k.4$cluster) #4421
```

Section 3.3.3 Random intercept model using cluster assignment covariate:

```
ri.model.cb <- lmer(college_bound.p ~ factor(clust_label) + SURVYEAR +
factor(STITLI_OVERALL) +(1|SCHNO), df.4cluster, REML = 0)

ri.model.psb <- lmer(total_postsecondary_bound.p ~ factor(cluster) + SURVYEAR
+STITLI_OVERALL  + (1|SCHNO), df.4cluster, REML = 0)

#likelihood ratio tests to test overall significance of clusters
model.0c.t <- lmer(total_postsecondary_bound.p ~SURVYEAR +STITLI_OVERALL  +
(1|SCHNO), df.4cluster, REML = 0)
model.0c.c <- lmer(college_bound.p ~SURVYEAR +STITLI_OVERALL  + (1|SCHNO),
df.4cluster, REML = 0)

anova(model.0c.t, ri.model.psb, test = 'LRT')
anova(model.0c.c, ri.model.cb, test = 'LRT')
```

Figure 2:

```
ggplot(data = df.arm1)+
        geom_histogram(aes(x=HUG_prop), col = 'white')+
        theme_bw()+
        facet_wrap('SURVYEAR')+
        xlab('Percent HUG')
```

Figure 3:

```
ggplot(data = df.arm1.ol)+
  geom_histogram(aes(x=Percent), col = 'white', binwidth = 5)+
  theme_bw()+
  facet_wrap('Outcome')
```

Figure 4:

```
ggplot(data = df.arm1.ol2[which(df.arm1.ol2$Outcome != 'Specialized Degree
Bound')])+
  geom_point(aes(x=(HUG_prop*100),y=Percent, color = STITLI_OVERALL), alpha =
0.5)+
  theme_bw()+
  xlab('Percent HUG Student Population')+
  ylab('Percent Graduates')+
  scale_color_discrete(name = 'Title I Status',
```

```
                      labels=c('1' = 'Yes', '2' = "No"))+
  facet_wrap('Outcome')
```

Figure 5:

```
ggplot()+
  geom_point(aes(x=clusters, y=wss))+
  theme_bw()+
  ylab('WSS')+
  xlab('No. of Clusters')
```

Figure 6:

```
fig <- plot_ly(data = df.4clusterm, x = ~WH_all.p, y = ~BL_all.p , z =
~HI_all.p, type = 'scatter3d', mode = "markers", color= ~
as.factor(clust_label2), size = 2, colors = c('#4AC6B7', '#1972A4',
'#965F8A', '#FF7070'))

fig <- fig %>% layout(scene = list(xaxis = list(title='Proportion White'),
                                   yaxis = list(title = 'Proportion Black'),
                                   zaxis = list(title = 'Proportion
Hispanic')),
                      #paper_bgcolor = 'rgb(243, 243, 243)',
                      #plot_bgcolor = 'rgb(243, 243, 243)',
                      annotations = list(
                       x = 1.1,
                       y = 1.05,
                       text = 'Cluster',
                       xref = 'paper',
                       yref = 'paper',
                       showarrow = FALSE
                       ))
```

Appendix A.2 All R code used in analysis

loading libraries
```{r}
library(tidyverse)
library(data.table)
library(stringi)
library(tidyr)
library(reshape2)
library(ggplot2)
library(gganimate)
library(EnvStats)
library(nlme)
library(lme4)
library(lmerTest)
library(varhandle)
library(sjstats)
```

```
library(geepack)
library(cowplot)
library(plotly)
library(dotwhisker)
library(sjPlot)
library(sjlabelled)
library(sjmisc)
```
Reading in data
```{r}
#demographics
df.07_08 <- fread('07_08.txt', header = T) %>% subset(MSTATE07 == 'PA')
df.08_09 <- fread('08_09.txt', header = T) %>% subset(MSTATE08 == 'PA')
df.09_10 <- fread('09_10.txt', header = T) %>% subset(MSTATE09 == 'PA')
df.10_11 <- fread('10_11.txt', header = T) %>% subset(MSTATE == 'PA')
df.11_12 <- fread('11_12.txt', header = T) %>% subset(MSTATE == 'PA')
df.12_13 <- fread('12_13.txt', header = T) %>% subset(MSTATE == 'PA')
df.13_14 <- fread('13_14.txt', header = T) %>% subset(MSTATE == 'PA')
df.14_15 <- read.delim('14_15.txt', header = T) %>% subset(STATENAME ==
'PENNSYLVANIA')
df.15_16 <- fread('15_16.csv', header = T) %>% subset(STABR == 'PA')
df.16_17 <- fread('16_17.csv', header = T) %>% subset(STATENAME ==
'PENNSYLVANIA')
df.17_18 <- fread('17_18.csv', header = T) %>% subset(STATENAME ==
'PENNSYLVANIA')
#graduation data
grad_07_08 <- fread("grad_07_08.csv")
grad_08_09 <- fread("grad_08_09.csv")
grad_09_10 <- fread("grad_09_10.csv")
grad_10_11 <- fread("grad_10_11.csv")
grad_11_12<- fread("grad_11_12.csv")
grad_12_13<- fread("grad_12_13.csv")
grad_13_14 <- fread("grad_13_14.csv")
grad_14_15 <- fread("grad_14_15.csv")
grad_15_16 <- fread("grad_15_16.csv")
```
Aggregating Demographic Data
```{r}
#07,08,09 need to remove suffix
year <- c('07','08','09')
k=1
col_name_list=list(NA, NA, NA)
for (i in list(df.07_08,df.08_09,df.09_10)){
col_name <- colnames(i)
for (j in 1:length(colnames(i))){
if (stri_sub(col_name[j], -2, -1)==year[k]){
col_name[j] <- stri_sub(col_name[j],from = 1, to = -3)
}else{
col_name[j] <- col_name[j]
}
}
col_name_list[[k]]<- as.vector(rep(NA, times = length(col_name)))
col_name_list[[k]] <- col_name
k <- k+1
}
colnames(df.07_08) <- col_name_list[[1]]
colnames(df.08_09) <- col_name_list[[2]]
```

```
colnames(df.09_10) <- col_name_list[[3]]
#suffix removed
#adding survey year to 08/09 and 09/10
df.08_09$SURVYEAR <- 2008
df.09_10$SURVYEAR <- 2009
#fixing some survey years
df.14_15$SURVYEAR <- 2014
df.15_16$SURVYEAR <-2015
df.14_15$SURVYEAR <- as.numeric(df.14_15$SURVYEAR)
df.15_16$SURVYEAR <- as.numeric(df.15_16$SURVYEAR)
#fixing NCESSCH in 14_15
df.14_15$NCESSCH <- as.integer(df.14_15$NCESSCH)
#fixing colnames to match in 2014/2015
df.14_15$SCHNAM <- toupper(df.14_15$SCH_NAME)
df.14_15$SCHNO <- df.14_15$SCHID
df.14_15$LEANM <- df.14_15$LEA_NAME
df.14_15$SEASCH <- df.14_15$ST_SCHID
df.14_15$STID <- df.14_15$ST_LEAID
df.15_16$SCHNAM <- toupper(df.15_16$SCH_NAME)
df.15_16$SCHNO <- df.15_16$SCHID
df.15_16$LEANM <- df.15_16$LEA_NAME
df.15_16$SEASCH <- df.15_16$ST_SCHID
df.15_16$STID <- df.15_16$ST_LEAID
#matching all colnames
keep <- colnames(df.08_09)
for( i in keep[-which(keep %in% colnames(df.09_10))]){
df.09_10[,i] <- as.numeric(NA)
}
for( i in keep[-which(keep %in% colnames(df.10_11))]){
df.10_11[,i] <- as.numeric(NA)
}
for( i in keep[-which(keep %in% colnames(df.11_12))]){
df.11_12[,i] <- as.numeric(NA)
}
for( i in keep[-which(keep %in% colnames(df.12_13))]){
df.12_13[,i] <- as.numeric(NA)
}
for( i in keep[-which(keep %in% colnames(df.13_14))]){
df.13_14[,i] <-as.numeric(NA)
}
for( i in keep[-which(keep %in% colnames(df.14_15))]){
df.14_15[,i] <- as.numeric(NA)
}
for( i in keep[-which(keep %in% colnames(df.15_16))]){
df.15_16[,i] <- as.numeric(NA)
}
df.09_10.i <- as.data.frame(df.09_10)[,which(colnames(df.09_10) %in% keep)]
df.09_10.i <- as.data.frame(df.09_10)[,which(colnames(df.09_10) %in% keep)]
df.10_11.i <- as.data.frame(df.10_11)[,which(colnames(df.10_11) %in% keep)]
df.11_12.i <- as.data.frame(df.11_12)[,which(colnames(df.11_12) %in% keep)]
df.12_13.i <- as.data.frame(df.12_13)[,which(colnames(df.12_13) %in% keep)]
df.13_14.i <- as.data.frame(df.13_14)[,which(colnames(df.13_14) %in% keep)]
df.14_15.i <- as.data.frame(df.14_15)[,which(colnames(df.14_15) %in% keep)]
df.15_16.i <- as.data.frame(df.15_16)[,which(colnames(df.15_16) %in% keep)]
#turning LZIP to Numeric
df.08_09$LZIP <- as.numeric(df.08_09$LZIP)
df.08_09$LZIP4 <- as.numeric(df.08_09$LZIP4)
```

```
df.09_10.i$LZIP <- as.numeric(df.09_10.i$LZIP)
df.09_10.i$LZIP4 <- as.numeric(df.09_10.i$LZIP4)
df.10_11.i$LZIP <- as.numeric(df.10_11.i$LZIP)
df.10_11.i$LZIP4 <- as.numeric(df.10_11.i$LZIP4)
df.11_12.i$LZIP <- as.numeric(df.11_12.i$LZIP)
df.11_12.i$LZIP4 <- as.numeric(df.11_12.i$LZIP4)
df.12_13.i$LZIP <- as.numeric(df.12_13.i$LZIP)
df.12_13.i$LZIP4 <- as.numeric(df.12_13.i$LZIP4)
df.13_14.i$LZIP <- as.numeric(df.13_14.i$LZIP)
df.13_14.i$LZIP4 <- as.numeric(df.13_14.i$LZIP4)
df.14_15.i$LZIP <- as.numeric(df.14_15.i$LZIP)
df.14_15.i$LZIP4 <- as.numeric(df.14_15.i$LZIP4)
df.15_16.i$LZIP <- as.numeric(df.15_16.i$LZIP)
df.15_16.i$LZIP4 <- as.numeric(df.15_16.i$LZIP4)
df.overall <- as.data.frame(df.08_09) %>%
bind_rows(.,df.09_10.i) %>%
bind_rows(.,df.10_11.i) %>%
bind_rows(.,df.11_12.i) %>%
bind_rows(.,df.12_13.i) %>%
bind_rows(.,df.13_14.i) %>%
bind_rows(.,df.14_15.i) %>%
bind_rows(.,df.15_16.i)
#290 x 25627
```
Preparing DF overall
```{r}
df.hs <- df.overall
df.hs <- df.overall[df.overall$GSHI == '12' | df.overall$GSHI == 'N' |
df.overall$GSHI == 'UG' |df.overall$SURVYEAR == 2014|df.overall$SURVYEAR ==
2015,] ###########3 11031 x 290
#df.hs <- merge(x = df.overall, y = grad.overall, by =
c('SCHNO','SURVYEAR') ) #3957624 x 309
#df.hs <- unique(df.hs)
#changing negative values to missing
for(j in 31:290){
for(i in 1:length(df.hs$NCESSCH)){
if(df.hs[i,j] < 0 & is.na(df.hs[i,j]) == FALSE){
df.hs[i,j] <- 0
}
}
}
#totaling m+f for each race/grade
#G9
df.hs$AM09 <- df.hs$AM09F + df.hs$AM09M
df.hs$AS09 <- df.hs$AS09F + df.hs$AS09M
df.hs$HI09 <- df.hs$HI09F + df.hs$HI09M
df.hs$BL09 <- df.hs$BL09F + df.hs$BL09M
df.hs$WH09 <- df.hs$WH09F + df.hs$WH09M
df.hs$HP09 <- df.hs$HP09F + df.hs$HP09M
df.hs$TR09 <- df.hs$TR09F + df.hs$TR09M
df.hs$HUG09 <- df.hs$AM09 + df.hs$HI09 + df.hs$BL09 + df.hs$HP09
#G10
df.hs$AM10 <- df.hs$AM10F + df.hs$AM10M
df.hs$AS10 <- df.hs$AS10F + df.hs$AS10M
df.hs$HI10 <- df.hs$HI10F + df.hs$HI10M
df.hs$BL10 <- df.hs$BL10F + df.hs$BL10M
df.hs$WH10 <- df.hs$WH10F + df.hs$WH10M
```

```
df.hs$HP10 <- df.hs$HP10F + df.hs$HP10M
df.hs$TR10 <- df.hs$TR10F + df.hs$TR10M
df.hs$HUG10 <- df.hs$AM10 + df.hs$HI10 + df.hs$BL10 + df.hs$HP10
#G11
df.hs$AM11 <- df.hs$AM11F + df.hs$AM11M
df.hs$AS11 <- df.hs$AS11F + df.hs$AS11M
df.hs$HI11 <- df.hs$HI11F + df.hs$HI11M
df.hs$BL11 <- df.hs$BL11F + df.hs$BL11M
df.hs$WH11 <- df.hs$WH11F + df.hs$WH11M
df.hs$HP11 <- df.hs$HP11F + df.hs$HP11M
df.hs$TR11 <- df.hs$TR11F + df.hs$TR11M
df.hs$HUG11 <- df.hs$AM11 + df.hs$HI11 + df.hs$BL11 + df.hs$HP11
#G12
df.hs$AM12 <- df.hs$AM12F + df.hs$AM12M
df.hs$AS12 <- df.hs$AS12F + df.hs$AS12M
df.hs$HI12 <- df.hs$HI12F + df.hs$HI12M
df.hs$BL12 <- df.hs$BL12F + df.hs$BL12M
df.hs$WH12 <- df.hs$WH12F + df.hs$WH12M
df.hs$HP12 <- df.hs$HP12F + df.hs$HP12M
df.hs$TR12 <- df.hs$TR12F + df.hs$TR12M
df.hs$HUG12 <- df.hs$AM12 + df.hs$HI12 + df.hs$BL12 + df.hs$HP12
#adding all race categories
df.hs$AM_all <- df.hs$AM09 + df.hs$AM10 + df.hs$AM11 + df.hs$AM12
df.hs$AS_all <- df.hs$AS09 + df.hs$AS10 + df.hs$AS11 + df.hs$AS12
df.hs$HI_all <- df.hs$HI09 + df.hs$HI10 + df.hs$HI11 + df.hs$HI12
df.hs$BL_all <- df.hs$BL09 + df.hs$BL10 + df.hs$BL11 + df.hs$BL12
df.hs$WH_all <- df.hs$WH09 + df.hs$WH10 + df.hs$WH11 + df.hs$WH12
df.hs$HP_all <- df.hs$HP09 + df.hs$HP10 + df.hs$HP11 + df.hs$HP12
df.hs$TR_all <- df.hs$TR09 + df.hs$TR10 + df.hs$TR11 + df.hs$TR12
#addng all hug
df.hs$HUG_all <- df.hs$HUG09 + df.hs$HUG10 + df.hs$HUG11 + df.hs$HUG12
df.hs$HS_all <- df.hs$G09 + df.hs$G10 + df.hs$G11 + df.hs$G12
df.hs$HS_all_noNA <- df.hs$AM_all + df.hs$AS_all + df.hs$HI_all +
df.hs$BL_all + df.hs$WH_all + df.hs$HP_all + df.hs$TR_all
df.hs$HUG_prop <- df.hs$HUG_all/df.hs$HS_all_noNA
df.hs <- as.data.table(df.hs)
#adding title I status as of 2013 and 2009
#df.hs <- merge(x = df.hs , y = cbind.data.frame(SEASCH =
df.hs$SEASCH[which(df.hs$SURVYEAR == 2013)], TITLEI_2013 =
df.hs$TITLEI[which(df.hs$SURVYEAR == 2013)]), by = 'SEASCH', all.x = T)
#adding proportions of all races
df.hs$AM_all.p <- df.hs$AM_all/df.hs$HS_all_noNA
df.hs$AS_all.p <- df.hs$AS_all/df.hs$HS_all_noNA
df.hs$HI_all.p <- df.hs$HI_all/df.hs$HS_all_noNA
df.hs$BL_all.p <- df.hs$BL_all/df.hs$HS_all_noNA
df.hs$WH_all.p <- df.hs$WH_all/df.hs$HS_all_noNA
df.hs$HP_all.p <- df.hs$HP_all/df.hs$HS_all_noNA
df.hs$TR_all.p <- df.hs$TR_all/df.hs$HS_all_noNA
#adding all 12th graders
df.hs$all_G12 <- df.hs$AM12 + df.hs$AS12 + df.hs$HI12 + df.hs$BL12 +
df.hs$WH12 + df.hs$HP12 + df.hs$TR12
```
Preparing Graduation Data
```{r}
#adding Survey Year
grad_07_08$SURVYEAR <- 2007
grad_08_09$SURVYEAR <- 2008
```

```
grad_09_10$SURVYEAR <- 2009
grad_10_11$SURVYEAR <- 2010
grad_11_12$SURVYEAR <- 2011
grad_12_13$SURVYEAR <- 2012
grad_13_14$SURVYEAR <- 2013
grad_14_15$SURVYEAR <- 2014
grad_15_16$SURVYEAR <- 2015
#adding school number to 14/15 and 15/16
grad_14_15$`School Number` <- grad_14_15$`School Code`
grad_15_16$`School Number` <- grad_15_16$`School Code`
#fixing title of some columns in 14_15/15_16
names(grad_14_15)[names(grad_14_15) %in% c('Graduate Count',
'Total College Bound %',
'College Bound',
'2- Or 4-Year University %',
'Specialized Associate Degree
Granting Institution',
'Specialized Assocate Degree
Granting Institution %')] <- c('Total Graduates',
'Total College-Bound',
'Total College-Bound %',
'2- or 4-Year College or University %',
'Specialized Associate Degree-Granting Institution',
'Specialized Associate Degree-Granting Institution %')
names(grad_15_16)[names(grad_15_16) %in% c('Graduate Count',
'Total College Bound %',
'College Bound',
'2- Or 4-Year University %',
'Specialized Associate Degree
Granting Institution',
'Specialized Assocate Degree
Granting Institution %')] <- c('Total Graduates',
'Total College-Bound',
'Total College-Bound %',
'2- or 4-Year College or University %',
'Specialized Associate Degree-Granting Institution',
'Specialized Associate Degree-Granting Institution %')
#names(grad_15_16)[names(grad_15_16) %in% c('Total College Bound
%','College Bound')] <- c('Total College-Bound','Total College-Bound %')
#fixing columns to be the same
grad_07_08.i <- grad_07_08[,4:17]
grad_08_09.i <- grad_08_09[,5:18]
grad_09_10.i <- grad_09_10[,5:18]
grad_10_11.i <- grad_10_11[,5:18]
grad_11_12.i <- grad_11_12[,5:18]
grad_12_13.i <- grad_12_13[,5:18]
grad_13_14.i <- grad_13_14[,5:18]
grad_14_15.i <- grad_14_15[,5:18]
grad_15_16.i <- grad_15_16[,5:18]
#fixing some columns in some table so data types agree
grad_13_14.i$`Total College-Bound %` <- as.character(grad_13_14.i$`Total
College-Bound %`)
grad_13_14.i$`2- or 4-Year College or University %` <-
as.character(grad_13_14.i$`2- or 4-Year College or University %`)
grad_13_14.i$`Total Postsecondary Bound %` <-
as.character(grad_13_14.i$`Total Postsecondary Bound %`)
grad_13_14.i$`Non-Degree-Granting Postsecondary School %`<-
```

```
as.character(grad_13_14.i$`Non-Degree-Granting Postsecondary School %`)
grad_13_14.i$`Specialized Associate Degree-Granting Institution %`<-
as.character(grad_13_14.i$`Specialized Associate Degree-Granting
Institution %`)
#unioning all
grad.overall <- grad_08_09.i %>%
bind_rows(., grad_09_10.i) %>%
bind_rows(., grad_10_11.i) %>%
bind_rows(., grad_11_12.i) %>%
bind_rows(., grad_12_13.i) %>%
bind_rows(., grad_13_14.i) %>%
bind_rows(., grad_14_15.i) %>%
bind_rows(., grad_15_16.i)
#8674 x 20
grad.overall$SCHNO <- grad.overall$`School Number`
```
bringing outcomes and predictors together
```{r}
grad.overall$SCHNAM = toupper(grad.overall$School)
df.hs$SCHNAM <- toupper(df.hs$SCHNAM)
df.hs$SURVYEAR = as.factor(df.hs$SURVYEAR)
grad.overall$SURVYEAR = as.factor(grad.overall$SURVYEAR)
df.xynam <- merge(x = df.hs, y = grad.overall, by = c('SCHNAM','SURVYEAR')
) #5456
df.xynum <- merge(x = df.hs, y = grad.overall, by = c('SCHNO','SURVYEAR') )
#4364
df.xynam <- df.xynam[,-"SCHNO.y"]
names(df.xynam)[names(df.xynam) == "SCHNO.x"] <- "SCHNO"
df.xynum <- df.xynum[,-"SCHNAM.y"]
names(df.xynum)[names(df.xynum) == "SCHNAM.x"] <- "SCHNAM"
df.xy <- union(df.xynam, df.xynum) #5565
df.xy <-df.xy[,-333] # get rid of after fresh run
```
Preparing data
```{r}
#converting data to numeric that should be
#percents
df.xy$college_bound.p <- as.numeric(sub('%', '',df.xy$`Total College-Bound
%`))
df.xy$college_bound <- as.numeric(df.xy$`Total College-Bound`)
df.xy$nondegree_bound.p <- as.numeric(sub('%', '',df.xy$`Non-Degree-
Granting Postsecondary School %`))
df.xy$nondegree_bound <- as.numeric(df.xy$`Non-Degree-Granting
Postsecondary School`)
df.xy$total_postsecondary_bound.p <- as.numeric(sub('%', '',df.xy$`Total
Postsecondary Bound %`))
df.xy$total_postsecondary_bound <- as.numeric(df.xy$`Total Postsecondary
Bound`)
df.xy$specialized_degree_bound.p <- as.numeric(sub('%',
'',df.xy$`Specialized Associate Degree-Granting Institution %`))
df.xy$specialized_degree_bound <- as.numeric(df.xy$`Specialized Associate
Degree-Granting Institution`)
#creating charter variable
df.xy <- merge(df.xy, y = unique(cbind.data.frame(SCHNO =
df.xy$SCHNO[df.xy$SURVYEAR == 2008],CHARTER_2008 =
df.xy$CHARTR[df.xy$SURVYEAR==2008])), by = 'SCHNO', all.x = T)
df.xy <- merge(df.xy, y = unique(cbind.data.frame(SCHNO =
```

```
df.xy$SCHNO[df.xy$SURVYEAR == 2009],CHARTER_2009 =
df.xy$CHARTR[df.xy$SURVYEAR==2009])), by = 'SCHNO', all.x = T)
df.xy <- merge(df.xy, y = unique(cbind.data.frame(SCHNO =
df.xy$SCHNO[df.xy$SURVYEAR == 2011],CHARTER_2011 =
df.xy$CHARTR[df.xy$SURVYEAR==2011])), by = 'SCHNO', all.x = T)
df.xy <- merge(df.xy, y = unique(cbind.data.frame(SCHNO =
df.xy$SCHNO[df.xy$SURVYEAR == 2013],CHARTER_2013 =
df.xy$CHARTR[df.xy$SURVYEAR==2013])), by = 'SCHNO', all.x = T)
df.xy <- merge(df.xy, y = unique(cbind.data.frame(SCHNO =
df.xy$SCHNO[df.xy$SURVYEAR == 2014],CHARTER_2014 =
df.xy$CHARTR[df.xy$SURVYEAR==2014])), by = 'SCHNO', all.x = T)
df.xy <- merge(df.xy, y = unique(cbind.data.frame(SCHNO =
df.xy$SCHNO[df.xy$SURVYEAR == 2015],CHARTER_2015 =
df.xy$CHARTR[df.xy$SURVYEAR==2015])), by = 'SCHNO', all.x = T)
#making an overall charter variable
for(i in 1:length(df.xy$SCHNO)){
if(is.na(df.xy$CHARTER_2008[i])){
if(is.na(df.xy$CHARTER_2009[i])){
if(is.na(df.xy$CHARTER_2011[i])){
if(is.na(df.xy$CHARTER_2013[i])){
if(is.na(df.xy$CHARTER_2014[i])){
df.xy$CHARTER_OVERALL[i] <- df.xy$CHARTER_2015[i]
}
else{df.xy$CHARTER_OVERALL[i]<-df.xy$CHARTER_2014[i]}
}
else{df.xy$CHARTER_OVERALL[i] <- df.xy$CHARTER_2013[i]}
}
else{df.xy$CHARTER_OVERALL[i] <- df.xy$CHARTER_2011[i]}
}
else{df.xy$CHARTER_OVERALL[i] <- df.xy$CHARTER_2009[i]}
}
else{df.xy$CHARTER_OVERALL[i] <- df.xy$CHARTER_2008[i]}
}
for(i in 1:length(df.xy$SCHNO)){
if(is.na(df.xy$CHARTER_OVERALL[i])){
if(df.xy$SCHNAM[i] %like% "%CS%" | df.xy$SCHNAM[i] %like% "%CHARTER%"){
df.xy$CHARTER_OVERALL[i] <- 1
}
else(df.xy$CHARTER_OVERALL[i]<-2)
}
}
#creating TITLEI variable
df.xy <- merge(df.xy, y = unique(cbind.data.frame(SCHNO =
df.xy$SCHNO[df.xy$SURVYEAR == 2008],TITLEI_2008 =
df.xy$TITLEI[df.xy$SURVYEAR==2008])), by = 'SCHNO', all.x = T)
df.xy <- merge(df.xy, y = unique(cbind.data.frame(SCHNO =
df.xy$SCHNO[df.xy$SURVYEAR == 2009],TITLEI_2009 =
df.xy$TITLEI[df.xy$SURVYEAR==2009])), by = 'SCHNO', all.x = T)
df.xy <- merge(df.xy, y = unique(cbind.data.frame(SCHNO =
df.xy$SCHNO[df.xy$SURVYEAR == 2011],TITLEI_2011 =
df.xy$TITLEI[df.xy$SURVYEAR==2011])), by = 'SCHNO', all.x = T)
df.xy <- merge(df.xy, y = unique(cbind.data.frame(SCHNO =
df.xy$SCHNO[df.xy$SURVYEAR == 2013],TITLEI_2013 =
df.xy$TITLEI[df.xy$SURVYEAR==2013])), by = 'SCHNO', all.x = T)
df.xy <- merge(df.xy, y = unique(cbind.data.frame(SCHNO =
df.xy$SCHNO[df.xy$SURVYEAR == 2014],TITLEI_2014 =
df.xy$TITLEI[df.xy$SURVYEAR==2014])), by = 'SCHNO', all.x = T)
```

```
df.xy <- merge(df.xy, y = unique(cbind.data.frame(SCHNO =
df.xy$SCHNO[df.xy$SURVYEAR == 2015],TITLEI_2015 =
df.xy$TITLEI[df.xy$SURVYEAR==2015])), by = 'SCHNO', all.x = T)
#making an overall TITLE I variable
for(i in 1:length(df.xy$SCHNO)){
if(is.na(df.xy$TITLEI_2008[i])| df.xy$TITLEI_2008[i] == "N"){
if(is.na(df.xy$TITLEI_2009[i])| df.xy$TITLEI_2009[i] == "N"){
if(is.na(df.xy$TITLEI_2011[i])| df.xy$TITLEI_2011[i] == "N"){
if(is.na(df.xy$TITLEI_2013[i])| df.xy$TITLEI_2013[i] == "N"){
if(is.na(df.xy$TITLEI_2014[i])| df.xy$TITLEI_2014[i] == "N"){
df.xy$TITLEI_OVERALL[i] <- df.xy$TITLEI_2015[i]
}
else{df.xy$TITLEI_OVERALL[i]<-df.xy$TITLEI_2014[i]}
}
else{df.xy$TITLEI_OVERALL[i] <- df.xy$TITLEI_2013[i]}
}
else{df.xy$TITLEI_OVERALL[i] <- df.xy$TITLEI_2011[i]}
}
else{df.xy$TITLEI_OVERALL[i] <- df.xy$TITLEI_2009[i]}
}
else{df.xy$TITLEI_OVERALL[i] <- df.xy$TITLEI_2008[i]}
}
#creating School wide STITLI variable
df.xy <- merge(df.xy, y = unique(cbind.data.frame(SCHNO =
df.xy$SCHNO[df.xy$SURVYEAR == 2008],STITLI_2008 =
df.xy$STITLI[df.xy$SURVYEAR==2008])), by = 'SCHNO', all.x = T)
df.xy <- merge(df.xy, y = unique(cbind.data.frame(SCHNO =
df.xy$SCHNO[df.xy$SURVYEAR == 2009],STITLI_2009 =
df.xy$STITLI[df.xy$SURVYEAR==2009])), by = 'SCHNO', all.x = T)
df.xy <- merge(df.xy, y = unique(cbind.data.frame(SCHNO =
df.xy$SCHNO[df.xy$SURVYEAR == 2011],STITLI_2011 =
df.xy$STITLI[df.xy$SURVYEAR==2011])), by = 'SCHNO', all.x = T)
df.xy <- merge(df.xy, y = unique(cbind.data.frame(SCHNO =
df.xy$SCHNO[df.xy$SURVYEAR == 2013],STITLI_2013 =
df.xy$STITLI[df.xy$SURVYEAR==2013])), by = 'SCHNO', all.x = T)
df.xy <- merge(df.xy, y = unique(cbind.data.frame(SCHNO =
df.xy$SCHNO[df.xy$SURVYEAR == 2014],STITLI_2014 =
df.xy$STITLI[df.xy$SURVYEAR==2014])), by = 'SCHNO', all.x = T)
df.xy <- merge(df.xy, y = unique(cbind.data.frame(SCHNO =
df.xy$SCHNO[df.xy$SURVYEAR == 2015],STITLI_2015 =
df.xy$STITLI[df.xy$SURVYEAR==2015])), by = 'SCHNO', all.x = T)
#making an overall School-wide TITLE I variable
for(i in 1:length(df.xy$SCHNO)){
if(is.na(df.xy$STITLI_2008[i])| df.xy$STITLI_2008[i] == "N" |
df.xy$STITLI_2008[i] == "M"){
if(is.na(df.xy$STITLI_2009[i])| df.xy$STITLI_2009[i] == "N"|
df.xy$STITLI_2009[i] == "M"){
if(is.na(df.xy$STITLI_2011[i])| df.xy$STITLI_2011[i] == "N"|
df.xy$STITLI_2011[i] == "M"){
if(is.na(df.xy$STITLI_2013[i])| df.xy$STITLI_2013[i] == "N"|
df.xy$STITLI_2013[i] == "M"){
if(is.na(df.xy$STITLI_2014[i])| df.xy$STITLI_2014[i] == "N"|
df.xy$STITLI_2014[i] == "M"){
df.xy$STITLI_OVERALL[i] <- df.xy$STITLI_2015[i]
}
else{df.xy$STITLI_OVERALL[i]<-df.xy$STITLI_2014[i]}
}
```

```
else{df.xy$STITLI_OVERALL[i] <- df.xy$STITLI_2013[i]}
}
else{df.xy$STITLI_OVERALL[i] <- df.xy$STITLI_2011[i]}
}
else{df.xy$STITLI_OVERALL[i] <- df.xy$STITLI_2009[i]}
}
else{df.xy$STITLI_OVERALL[i] <- df.xy$STITLI_2008[i]}
}
#creating managable sub-data set for arm 1
df.arm1 <- na.omit(df.xy[,c('SCHNAM',
'SCHNO',
'SURVYEAR',
'HUG_prop',
'TITLEI_OVERALL',
'STITLI_OVERALL',
'CHARTER_OVERALL',
'college_bound.p',
'nondegree_bound.p',
'specialized_degree_bound.p',
'total_postsecondary_bound.p')]) #4431
length(unique(df.arm1$SCHNO)) #610
df.arm1 <- unique(df.arm1[which(df.arm1$total_postsecondary_bound.p !=
0),])
df.arm1 <- unique(df.arm1[which(df.arm1$SCHNO != 16),])
df.arm1 <- unique(df.arm1[which(df.arm1$SCHNO != 3848),])
df.arm1 <- unique(df.arm1[which(df.arm1$SCHNO != 941),])
df.arm1 <- unique(df.arm1[which(is.na(df.arm1$HUG_prop) != T),]) #4118
length(unique(df.arm1$SCHNO)) #601
#unfactoring SURVYEAR
df.arm1$SURVYEAR <- unfactor(df.arm1$SURVYEAR)
#adding centered continuous features for polynomial models
df.arm1$SURVYEAR.c <- df.arm1$SURVYEAR - mean(df.arm1$SURVYEAR)
df.arm1$HUG_prop.c <- df.arm1$HUG_prop - mean(df.arm1$HUG_prop)
#adding polynomial continuous features
df.arm1$SURVYEAR.2 <- df.arm1$SURVYEAR.c^2
df.arm1$SURVYEAR.3 <- df.arm1$SURVYEAR.c^3
df.arm1$HUG_prop.2 <- df.arm1$HUG_prop.c^2
df.arm1$HUG_prop.3 <- df.arm1$HUG_prop.c^3
#making hug_prop a percent
df.arm1$HUG_percent <- df.arm1$HUG_prop*100
#creating functional data table with all variables of interest
df.dev <- na.omit(df.xy[,c('SCHNO',
'SCHNAM',
'SURVYEAR',
'Total Graduates',
'college_bound.p',
'nondegree_bound.p',
'specialized_degree_bound.p',
"total_postsecondary_bound.p",
'STITLI_OVERALL',
"HUG_prop",
'AM_all.p',
'AS_all.p',
'HI_all.p',
'BL_all.p',
'WH_all.p',
'HP_all.p',
```

```
'TR_all.p')]) #4431
df.dev <- unique(df.dev[which(df.dev$total_postsecondary_bound.p != 0),])
#4170
df.dev <- unique(df.dev[which(df.dev$SCHNO != 16),]) #4154
df.dev <- unique(df.dev[which(df.dev$SCHNO != 3848),]) #4118
df.dev <- unique(df.dev[which(df.dev$SCHNO != 941),])
#creating long forms of data
df.dev.l <- melt(df.dev,
# ID variables - all the variables to keep but not split apart on
id.vars=c("SCHNAM", "SURVYEAR"),
# The source columns
measure.vars=c("AM_all.p", "AS_all.p",
"HI_all.p","BL_all.p","WH_all.p","HP_all.p","TR_all.p", "HUG_prop",
"college_bound.p", "total_postsecondary_bound.p", 'STITLI_OVERALL'),
# Name of the destination column that will identify the original
# column that the measurement came from
variable.name="Race",
value.name="Percent"
)#29078 x 4
#long format just race
df.dev.r <- melt(df.dev,
# ID variables - all the variables to keep but not split apart on
id.vars=c("SCHNAM", "SURVYEAR"),
# The source columns
measure.vars=c("AM_all.p", "AS_all.p",
"HI_all.p","BL_all.p","WH_all.p","HP_all.p","TR_all.p"),
# Name of the destination column that will identify the original
# column that the measurement came from
variable.name="Race",
value.name="Percent"
)#29078 x 4
#long form of outcomes
df.arm1.ol <- melt(df.dev,
# ID variables - all the variables to keep but not split apart on
id.vars=c("SCHNAM", "SURVYEAR"),
# The source columns
measure.vars=c('total_postsecondary_bound.p', 'college_bound.p',
'specialized_degree_bound.p'),
# Name of the destination column that will identify the original
# column that the measurement came from
variable.name="Outcome",
value.name="Percent"
)#29078 x 4
#maybe merging df.xy with total number of grads?
df.dev <- merge(df.dev,df.hs[,c('SCHNO','SCHNAM','SURVYEAR','all_G12')], by
= c('SCHNO','SURVYEAR','SCHNAM'),all.x = T) #4118
df.dev <- merge(df.dev,df.xy[,c('SCHNO','SCHNAM','SURVYEAR','G12')], by =
c('SCHNO','SURVYEAR','SCHNAM'),all.x = T) #4118
```
Data exploration
```{r}
summary(df.hs[,323:332])
ggplot(data = df.hs)+
geom_bar(aes(x=factor(SURVYEAR), y = HUG_prop), stat = "summary", fun.y =
"mean")+
theme_bw()
ggplot(data = df.hs)+
```

```
geom_boxplot(aes(x=factor(SURVYEAR), y = HUG_prop))+
theme_bw()
ggplot(data = df.hs)+
geom_point(aes(x=HUG_prop, y = as.factor(TITLEI_2013)))+
theme_bw()
#merging outcome long with phug and title I
df.arm1.ol$SURVYEAR <- unfactor(df.arm1.ol$SURVYEAR)
df.arm1.ol2 <- merge(df.arm1.ol, df.arm1[,c('SCHNAM','SURVYEAR','HUG_prop',
'STITLI_OVERALL')], by = c('SCHNAM','SURVYEAR'))
df.arm1.ol2$Outcome <- factor(df.arm1.ol2$Outcome, levels =
c("college_bound.p", "total_postsecondary_bound.p",
"specialized_degree_bound.p"),
labels = c("College Bound", "Total Post-Secondary Bound",
"Specialized Degree Bound"))
#scatter plot of covariates vs outcome
ggplot(data = df.arm1.ol2[which(df.arm1.ol2$Outcome != 'Specialized Degree
Bound')])+
geom_point(aes(x=(HUG_prop*100),y=Percent, color = STITLI_OVERALL), alpha
= 0.5)+
theme_bw()+
xlab('Percent HUG Student Population')+
ylab('Percent Graduates')+
scale_color_discrete(name = 'Title I Status',
labels=c('1' = 'Yes', '2' = "No"))+
facet_wrap('Outcome')
ggplot(data = df.arm1)+
geom_point(aes(x=HUG_prop,y=total_postsecondary_bound.p, color =
STITLI_OVERALL))+
theme_bw()+
xlab('Proportion HUG Student Population')+
ylab('Percent Post-Secondary Bound Graduates')+
scale_color_discrete(name = 'Title Status',labels=c('1' = 'Yes',
'2' = "No"))+
scale_color_manual(values=c("#E69F00", "#56B4E9"))+
ggtitle("Total Post-Secondary Bound")
#scatter plot of college-bound rates
ggplot(data = df.arm1)+
geom_point(aes(x=HUG_prop,y=college_bound.p, color = STITLI_OVERALL))+
theme_bw()+
transition_time(as.integer(SURVYEAR)) +
labs(title = "Year: {frame_time}")+
ease_aes('linear')
#ease_aes('cubic-in-out')
p2 <- ggplot(data = ocdrug, aes(x = Tmnt, y = EE_Cmax, group = ID, colour =
Seq)) +
mytheme +
coord_trans(y="log10", limy=c(100,700)) +
labs(list(title = "Cmax", y = paste("EE","n","pg/mL"))) +
geom_line(size=1) +
geom_text(data=subset(ocdrug, ID %in% c(2,20)),
aes(Tmnt,EE_Cmax,label=ID)) +
theme(legend.position="none")
#interaction plot
interaction.plot(df.arm1$SURVYEAR,df.arm1$SCHNO,
df.arm1$total_postsecondary_bound.p, xlab="Year", ylab="Total Post-
Secondary College Bound", legend=F)
interaction.plot(df.arm1$SURVYEAR,df.arm1$SCHNO, df.arm1$college_bound.p,
```

```
xlab="Year", ylab="Percent College Bound", legend=F)
interaction.plot(df.arm1$SURVYEAR,df.arm1$SCHNO,
df.arm1$specialized_degree_bound.p, xlab="Year", ylab="Specialized Degree
Bound", legend=F)
```
Summary Stats
```{r}
#finding average obs per school
obvs_by_school <- (aggregate(x = df.arm1,
by = list(unique.values = df.arm1$SCHNO),
FUN = length))
table(obvs_by_school$SURVYEAR.c)
prop.table(table(obvs_by_school$SURVYEAR.c))*100
mean(obvs_by_school$SURVYEAR)
#mean/variance of race
mean(df.dev$AM_all.p)*100
mean(df.dev$AS_all.p)*100
mean(df.dev$HI_all.p)*100
mean(df.dev$BL_all.p)*100
mean(df.dev$WH_all.p)*100
mean(df.dev$HP_all.p)*100
mean(df.dev$TR_all.p)*100
var(df.dev$AM_all.p*100)
var(df.dev$AS_all.p*100)
var(df.dev$HI_all.p*100)
var(df.dev$BL_all.p*100)
var(df.dev$WH_all.p*100)
var(df.dev$HP_all.p*100)
var(df.dev$TR_all.p*100)
#table of TITLE I status by school
uniq_titleI <- unique(df.arm1[,c('SCHNO','STITLI_OVERALL')])
table(uniq_titleI$STITLI_OVERALL)
prop.table(table(uniq_titleI$STITLI_OVERALL))*100
#table of Charter School status by school
uniq_charter <- unique(df.arm1[,c('SCHNO','CHARTER_OVERALL')])
table(uniq_charter$CHARTER_OVERALL)
prop.table(table(uniq_charter$CHARTER_OVERALL))*100
#mean of outcomes
mean(df.arm1$total_postsecondary_bound.p)
mean(df.arm1$college_bound.p)
mean(df.arm1$specialized_degree_bound.p)
```
Exploring Distribution of Race in schools (mean/variance)
```{r}
#overall mean and variance (grouped by year)
ggplot(data = df.xyl)+
geom_bar(aes(x=Race, y=Percent, fill = SURVYEAR),
stat = "summary", fun.y = "mean", position = position_dodge())+
theme_bw()+
#theme(axis.text.x = element_text(angle = 90, hjust = 1, size = 4))+
ylab('Mean Percent')+
xlab('RACE')+
ggtitle('Mean Percents by Race')+
scale_x_discrete(labels=c("AM_all.p" = "American Indian",
"AS_all.p" = "Asian",
"WH_all.p" = "White",
"BL_all.p"="Black",
```

```
"HI_all.p" = "Hispanic",
"HP_all.p" = "Hawaiian/Pacific Islander",
"TR_all.p"="Two or More Races"))
ggplot(data = df.xyl)+
geom_bar(aes(x=Race, y=Percent, fill = SURVYEAR),
stat = "summary", fun.y = "var",position = position_dodge())+
theme_bw()+
theme(axis.text.x = element_text(angle = 90, hjust = 1))+
ylab('Variance of Percent')+
xlab('RACE')+
ggtitle('Percent Variance by Race')+
scale_x_discrete(labels=c("AM_all.p" = "American Indian",
"AS_all.p" = "Asian",
"WH_all.p" = "White",
"BL_all.p"="Black",
"HI_all.p" = "Hispanic",
"HP_all.p" = "Hawaiian/Pacific Islander",
"TR_all.p"="Two or More Races"))
#overall mean and variance over all years
ggplot(data = df.xyl)+
geom_bar(aes(x=Race, y=Percent),
stat = "summary", fun.y = "mean")+
theme_bw()+
theme(axis.text.x = element_text(angle = 45, hjust = 1))+
ylab('Mean Proportion')+
xlab('RACE')+
#ggtitle('Average Racial Distribution of Pennsylvania Public Schools from
2008 to 2016')+
scale_x_discrete(labels=c("AM_all.p" = "American Indian",
"AS_all.p" = "Asian",
"WH_all.p" = "White",
"BL_all.p"="Black",
"HI_all.p" = "Hispanic",
"HP_all.p" = "Hawaiian/Pacific Islander",
"TR_all.p"="Two or More Races"))
ggplot(data = df.xyl)+
geom_bar(aes(x=Race, y=Percent),
stat = "summary", fun.y = "var")+
theme_bw()+
theme(axis.text.x = element_text(angle = 45, hjust = 1))+
ylab('Variance of Proportion')+
xlab('RACE')+
#ggtitle('Variance of Racial Distribution of Pennsylvania Public Schools
from 2008 to 2016')+
scale_x_discrete(labels=c("AM_all.p" = "American Indian",
"AS_all.p" = "Asian",
"WH_all.p" = "White",
"BL_all.p"="Black",
"HI_all.p" = "Hispanic",
"HP_all.p" = "Hawaiian/Pacific Islander",
"TR_all.p"="Two or More Races"))
```
Checking Features/Linear Model Assumptions
```{r}
#looking at percent HUG
ggplot(data = df.arm1)+
geom_histogram(aes(x=HUG_prop), col = 'white')+
```

```
theme_bw()+
xlab('Percent HUG')
#looking at percent HUG by year
ggplot(data = df.arm1)+
geom_histogram(aes(x=HUG_prop), col = 'white')+
theme_bw()+
facet_wrap('SURVYEAR')+
xlab('Percent HUG')
#looking at total post secondary bound
pb <- ggplot(data = df.arm1)+
geom_histogram(aes(x=total_postsecondary_bound.p), col = 'white')+
theme_bw()+
xlab('Percent Total Post-Secondary Bound')
#over time
ggplot(data = df.arm1)+
geom_histogram(aes(x=total_postsecondary_bound.p), col = 'white')+
theme_bw()+
xlab('Percent Total Post-Secondary Bound')+
facet_wrap('SURVYEAR')
#boxplots
pb.bp <- ggplot(data = df.arm1, aes( y = total_postsecondary_bound.p))+
geom_boxplot()+
theme_bw()+
ylab('Percent Total Post-Secondary Bound')
cb <- ggplot(data = df.arm1)+
geom_histogram(aes(x= college_bound.p), col = 'white')+
theme_bw()+
xlab('Percent College Bound')
cb.bp <- ggplot(data = df.arm1, aes( y = college_bound.p))+
geom_boxplot()+
theme_bw()+
ylab('Percent College Bound')
plot_grid(pb, pb.bp, cb, cb.bp, nrow = 2, rel_widths = c(1,1,1,1))
#using the long data
levels(df.arm1.ol$Outcome) <- c('Total Post-Secondary Bound','College
Bound', 'Specialized Degree Bound')
ggplot(data = df.arm1.ol)+
geom_histogram(aes(x=Percent), col = 'white', binwidth = 5)+
theme_bw()+
facet_wrap('Outcome')
ggplot(data = df.arm1.ol, aes( y = Percent))+
geom_boxplot()+
theme_bw()+
facet_wrap('Outcome')
#looking at total college bound
ggplot(data = df.arm1)+
geom_histogram(aes(x=college_bound.p), col = 'white')+
theme_bw()
#also normal!
#looking at specialized degree bound
ggplot(data = df.arm1)+
geom_histogram(aes(x=specialized_degree_bound.p), col = 'white')+
theme_bw()
#not normal, trying transfrom
ggplot(data = df.arm1)+
geom_histogram(aes(x= sqrt(specialized_degree_bound.p+1)), col = 'white')
+
```

```
theme_bw()
#looking at charter status
ggplot(data = df.arm1)+
geom_histogram(aes(x=CHARTER_OVERALL), col = 'white')+
theme_bw()
#looking at School-wide Title I Status
ggplot(data = df.arm1)+
geom_bar(aes(x=STITLI_OVERALL, stat = 'count'), col = 'white')+
theme_bw()
#interaction plot
interaction.plot(df.arm1$SURVYEAR,df.arm1$SCHNO,
df.arm1$total_postsecondary_bound.p, xlab="Year", ylab="Percent College
Bound", legend=F)
```
Results Section I- Longitudinal and polynomial models
```{r}
xtabs(~ SCHNO + SURVYEAR, df.arm1)

rm.college <- lmer(college_bound.p ~ HUG_prop + SURVYEAR +STITLI_OVERALL +
(SURVYEAR|SCHNO), df.arm1, REML = 0)
rm.total <- lmer(total_postsecondary_bound.p ~ HUG_prop + SURVYEAR
+STITLI_OVERALL + (SURVYEAR|SCHNO), df.arm1, REML = 0)
#polynomial models (quadratic)
ri.college.2 <- lmer(college_bound.p ~ HUG_prop.c + HUG_prop.2 + SURVYEAR.c
+ SURVYEAR.2 +STITLI_OVERALL + (1|SCHNO), df.arm1, REML = 0)
ri.total.2 <- lmer(total_postsecondary_bound.p ~ HUG_prop.c + HUG_prop.2 +
SURVYEAR.c + SURVYEAR.2 +STITLI_OVERALL + (1|SCHNO), df.arm1, REML = 0)
summary(rm.model)
summary(ri.model)
summary(rm.model.2)
summary(ri.model.2)
performance::icc(rm.model)
performance::icc(ri.model)
performance::icc(rm.model.2)
performance::icc(ri.model.2)

#setting no title I as reference
df.arm1 <- within(df.arm1, STITLI_OVERALL <-
relevel(factor(STITLI_OVERALL), ref = 2))

#models for write-up
ri.college <- lmer(college_bound.p ~ HUG_percent + SURVYEAR +STITLI_OVERALL
+ (1|SCHNO), df.arm1, REML = 0)
ri.total <- lmer(total_postsecondary_bound.p ~ HUG_percent + SURVYEAR
+STITLI_OVERALL + (1|SCHNO), df.arm1, REML = 0)
summary(ri.college)
summary(ri.total)
confint(ri.college, method="profile", ## default
oldNames = FALSE)
confint(ri.total)
#checking residuals
#creating the residuals (epsilon.hat)
resid <- residuals(ri.model)
qqnorm(resid)
#creating the standardized residual (std epsilon.hat)
resid.std <- resid/sd(resid)
plot(df.arm1$SCHNO, resid.std, ylim=c(-10, 10), ylab="std epsilon hat")
```

```
abline(h=0)
#homoskedacity plots
plot(df.arm1$SURVYEAR, resid, ylim=c(-10, 10), ylab="epsilon.hat",
xlab="AGE")
abline(h=0)
plot(df.arm1$STITLI_OVERALL, resid, ylim=c(-10, 10), ylab="epsilon.hat",
xlab="AGE")
abline(h=0)
plot(df.arm1$HUG_prop, resid, ylim=c(-10, 10), ylab="epsilon.hat",
xlab="AGE")
abline(h=0)
#plotting results
dwplot(ri.total@frame)
plot_model(ri.total, show.values = TRUE, vline.color = 'grey', title =
'Total Post-Secondary Bound')+theme_bw()
dwplot(ri.college)
plot_model(ri.college, show.values = TRUE, vline.color = 'grey', title =
'College Bound')+theme_bw()
#LRT checking nested models
#rand intercepts vs random effects
anova(ri.total, rm.total, test = 'LRT')
anova(ri.college, rm.college, test = 'LRT')
#polynomial models vs ri models
anova(ri.total, ri.total.2, test = 'LRT')
anova(ri.college, ri.college.2, test = 'LRT')
#Looking at change in 10pp HUG
df.arm1$HUG_percent10 <- df.arm1$HUG_percent/10
ri.college10 <- lmer(college_bound.p ~ HUG_percent10 + SURVYEAR
+STITLI_OVERALL + (1|SCHNO), df.arm1, REML = 0)
ri.total10 <- lmer(total_postsecondary_bound.p ~ HUG_percent10 + SURVYEAR
+STITLI_OVERALL + (1|SCHNO), df.arm1, REML = 0)
summary(ri.college10)
summary(ri.total10)
plot_model(ri.total10, show.values = TRUE, vline.color = 'grey', title =
'Total Post-Secondary Bound')+theme_bw()
plot_model(ri.college10, show.values = TRUE, vline.color = 'grey', title =
'College Bound')+theme_bw()
```
Turning Prop HUG into a Categorical Variable
```{r}
#graphing box plot of hug plot
ggplot(data = df.arm1)+
geom_boxplot(aes(x= factor(SURVYEAR), y=HUG_prop))+
theme_bw()
table(df.arm1$SURVYEAR)
table(df.arm1$STITLI_OVERALL, useNA = 'ifany')
summary(df.arm2$HUG_prop)
#creating categorical variables
df.arm2 <- df.arm1
df.arm2$HUG_cat[df.arm2$HUG_prop < .05] <- 1
df.arm2$HUG_cat[ df.arm2$HUG_prop >=.05 & df.arm2$HUG_prop < .25] <- 2
df.arm2$HUG_cat[df.arm2$HUG_prop >= .25 & df.arm2$HUG_prop < .50] <- 3
df.arm2$HUG_cat[df.arm2$HUG_prop >= .50 & df.arm2$HUG_prop < .75] <- 4
df.arm2$HUG_cat[df.arm2$HUG_prop >= .75 & df.arm2$HUG_prop <= 1] <- 5
#same but with quartiles
df.arm2$HUG_catq[df.arm2$HUG_prop < .0241] <- 1
df.arm2$HUG_catq[ df.arm2$HUG_prop >=.0241 & df.arm2$HUG_prop < .06853] <-
```

```
2
df.arm2$HUG_catq[df.arm2$HUG_prop >= 0.06853 & df.arm2$HUG_prop < .30710]
<- 3
df.arm2$HUG_catq[df.arm2$HUG_prop >= .30710] <- 4
table(df.arm2$HUG_cat)
table(df.arm2$HUG_catq)
#modeling
ri.tot.c <- lmer(total_postsecondary_bound.p ~ factor(HUG_catq) + SURVYEAR
+STITLI_OVERALL + (1|SCHNO), df.arm2, REML = 0)
ri.col.c <- lmer(college_bound.p ~ factor(HUG_catq) + SURVYEAR
+STITLI_OVERALL + (1|SCHNO), df.arm2, REML = 0)
ri.modelq <- lmer(total_postsecondary_bound.p ~ factor(HUG_catq) + SURVYEAR
+STITLI_OVERALL + (1|SCHNO), df.arm2, REML = 0)
summary(ri.tot.c)
summary(ri.col.c)
confint(ri.tot.c, method="profile", ## default
oldNames = FALSE)
confint(ri.col.c)
confint(ri.col.c, method="profile", ## default
oldNames = FALSE)
plot_model(ri.tot.c, show.values = TRUE, vline.color = 'grey', title =
'Total Post-Secondary Bound')+theme_bw()
plot_model(ri.col.c, show.values = TRUE, vline.color = 'grey', title =
'College Bound')+theme_bw()
#likelihood ratio test testing the significance of categorical variable
#making null model
model.0.t <- lmer(total_postsecondary_bound.p ~ SURVYEAR + STITLI_OVERALL +
(1|SCHNO), df.arm2, REML =0)
model.0.c <- lmer(college_bound.p ~ SURVYEAR + STITLI_OVERALL + (1|SCHNO),
df.arm2, REML =0)
anova(model.0.t, ri.tot.c, test = 'LRT')
anova(model.0.c, ri.col.c, test = 'LRT')
```
Clustering the data based on race distribution
```{r}
set.seed(101)
cluster_data <- na.omit(df.xy[,c('SCHNO',
'SURVYEAR',
'college_bound.p',
'nondegree_bound.p',
'specialized_degree_bound.p',
"total_postsecondary_bound.p",
'STITLI_OVERALL',
"HUG_prop",
'AM_all.p',
'AS_all.p',
'HI_all.p',
'BL_all.p',
'WH_all.p',
'HP_all.p',
'TR_all.p')]) #4431
cluster_data<-
unique(cluster_data[which(cluster_data$total_postsecondary_bound.p != 0),])
#4170
cluster_data <- unique(cluster_data[which(cluster_data$SCHNO != 16),])
#4154
cluster_data <- unique(cluster_data[which(cluster_data$SCHNO != 3848),])
```

```
#4136
cluster_data <- unique(cluster_data[which(cluster_data$SCHNO != 941),])
#4118
wss <- rep(NA,20)
clusters <- rep(NA,20)
for(k in 1:20){
a <- kmeans(x = cluster_data[,-c(1:8)], centers = k)
wss[k] <- a$tot.withinss
clusters[k] <- k
}
ggplot()+
geom_point(aes(x=clusters, y=wss))+
theme_bw()+
ylab('WSS')+
xlab('No. of Clusters')
#3,4 or 5 clusters
k.4 <- kmeans(x = cluster_data[,-c(1:8)], centers = 4)
k.3 <- kmeans(x = cluster_data[,-c(1:8)], centers = 3)
k.5 <- kmeans(x = cluster_data[,-c(1:8)], centers = 5)
df.4cluster <- cbind.data.frame(cluster_data, cluster = k.4$cluster) #4421
df.allcluster <- cbind.data.frame(df.4cluster, cluster3 = k.3$cluster)
df.allcluster <- cbind.data.frame(df.allcluster, cluster5 = k.5$cluster)
df.4cluster$SURVYEAR <- unfactor(df.4cluster$SURVYEAR)
#releveling so that cluster 4 (all white) is reference (only if all white
cluster is not number1)
df.4cluster <- within(df.4cluster, cluster <- relevel(factor(cluster), ref
= 4))
#looking into number of schools that fit into each
just.schools <- unique(cbind.data.frame(SCHNO = df.4cluster$SCHNO, cluster
= df.4cluster$clust_lab))
table(just.schools$cluster)
prop.table(table(just.schools$cluster))*100
# 1=403 (62%), 2=116 (17.85%), 3=104 (16%), 4=27 (4.15%)
#adding labels to cluster
df.4cluster$clust_label[df.4cluster$cluster == 4] <- 'white'
df.4cluster$clust_label[df.4cluster$cluster == 1] <- 'black'
df.4cluster$clust_label[df.4cluster$cluster == 2] <- 'mix'
df.4cluster$clust_label[df.4cluster$cluster == 3] <- 'hisp'
df.4cluster$clust_label[df.4cluster$cluster == 4] <- 1
df.4cluster$clust_label[df.4cluster$cluster == 1] <- 4
df.4cluster$clust_label[df.4cluster$cluster == 2] <- 3
df.4cluster$clust_label[df.4cluster$cluster == 3] <- 2
df.4cluster <- within(df.4cluster, cluster <- relevel(factor(clust_label),
ref = 1))
#making long version fr graphing
df.4clusterl <- melt(df.4cluster,
# ID variables - all the variables to keep but not split apart on
id.vars=c("SCHNO", "SURVYEAR", "cluster","clust_label"),
# The source columns
measure.vars=c("AM_all.p", "AS_all.p",
"HI_all.p","BL_all.p","WH_all.p","HP_all.p","TR_all.p"),
# Name of the destination column that will identify the original
# column that the measurement came from
variable.name="Race",
value.name="Percent"
)
df.allclusterl <- melt(df.allcluster,
```

```
# ID variables - all the variables to keep but not split apart on
id.vars=c("SCHNO", "SURVYEAR", "cluster","clust_label","cluster3",
"cluster5"),
# The source columns
measure.vars=c("AM_all.p", "AS_all.p",
"HI_all.p","BL_all.p","WH_all.p","HP_all.p","TR_all.p"),
# Name of the destination column that will identify the original
# column that the measurement came from
variable.name="Race",
value.name="Percent"
)
```
Graphing Cluster Data
```{r}
#looking at race dsitributions by cluster
ggplot(data = df.4clusterl)+
geom_bar(aes(x=Race, y=Percent),
stat = "summary", fun.y = "mean")+
theme_bw()+
theme(axis.text.x = element_text(angle = 45, hjust = 1, size = 7))+
ylab('Mean Percent')+
xlab('RACE')+
#ggtitle('Mean Percents by Race')+
scale_x_discrete(labels=c("AM_all.p" = "American Indian",
"AS_all.p" = "Asian",
"WH_all.p" = "White",
"BL_all.p"="Black",
"HI_all.p" = "Hispanic",
"HP_all.p" = "Hawaiian/Pacific Islander",
"TR_all.p"="Two or More Races"))+
facet_wrap('clust_label')
#3 clusters
ggplot(data = df.allclusterl)+
geom_bar(aes(x=Race, y=Percent),
stat = "summary", fun.y = "mean")+
theme_bw()+
theme(axis.text.x = element_text(angle = 45, hjust = 1, size = 7))+
ylab('Mean Percent')+
xlab('RACE')+
#ggtitle('Mean Percents by Race')+
scale_x_discrete(labels=c("AM_all.p" = "American Indian",
"AS_all.p" = "Asian",
"WH_all.p" = "White",
"BL_all.p"="Black",
"HI_all.p" = "Hispanic",
"HP_all.p" = "Hawaiian/Pacific Islander",
"TR_all.p"="Two or More Races"))+
facet_wrap('cluster3')
#5 clusters
ggplot(data = df.allclusterl)+
geom_bar(aes(x=Race, y=Percent),
stat = "summary", fun.y = "mean")+
theme_bw()+
theme(axis.text.x = element_text(angle = 45, hjust = 1, size = 7))+
ylab('Mean Percent')+
xlab('RACE')+
#ggtitle('Mean Percents by Race')+
```

```
scale_x_discrete(labels=c("AM_all.p" = "American Indian",
"AS_all.p" = "Asian",
"WH_all.p" = "White",
"BL_all.p"="Black",
"HI_all.p" = "Hispanic",
"HP_all.p" = "Hawaiian/Pacific Islander",
"TR_all.p"="Two or More Races"))+
facet_wrap('cluster5')
ggplot(data = df.4clusterl)+
geom_bar(aes(x=Race, y=Percent),
stat = "summary", fun.y = "var")+
theme_bw()+
theme(axis.text.x = element_text(angle = 45, hjust = 1, size = 7))+
ylab('Variance of Percent')+
xlab('RACE')+
scale_x_discrete(labels=c("AM_all.p" = "American Indian",
"AS_all.p" = "Asian",
"WH_all.p" = "White",
"BL_all.p"="Black",
"HI_all.p" = "Hispanic",
"HP_all.p" = "Hawaiian/Pacific Islander",
"TR_all.p"="Two or More Races"))+
facet_wrap('clust_label')
#lookng at percent hug by cluster
ggplot(df.4cluster, aes(clust_label, HUG_prop)) +
stat_summary(fun.y=mean, geom="bar")+
theme_bw()+
xlab('Cluster')+
ylab('Mean Percent HUG')
ggplot(df.4cluster, aes(clust_label, HUG_prop)) +
stat_summary(fun.y=var, geom="bar")+
theme_bw()+
xlab('Cluster')+
ylab('Variance Percent HUG')
#looking at Title I status by cluster
ggplot(data = df.4cluster)+
geom_bar(aes(x=STITLI_OVERALL, stat = 'count'), col = 'white')+
theme_bw()+
facet_wrap('clust_label')
#looking post-secondary bound rates by cluster
ggplot(data = df.4cluster)+
geom_histogram(aes(x=total_postsecondary_bound.p), col = 'white',
binwidth = 5)+
theme_bw()+
xlab('Total Post-Secondary Bound Rate')+
facet_wrap('clust_label')
ggplot(data = df.4cluster)+
geom_histogram(aes(x=college_bound.p), col = 'white', binwidth = 5)+
theme_bw()+
xlab('College Bound Rate')+
facet_wrap('clust_label')
ggplot(data = df.4cluster)+
geom_bar(aes(x=clust_label, y=total_postsecondary_bound.p),
stat = "summary", fun.y = "mean")+
theme_bw()+
#theme(axis.text.x = element_text(angle = 90, hjust = 1, size = 4))+
ylab('Mean Percent')+
```

```
xlab('Cluster')#+
#ggtitle('Post-Secondary bound percent by cluster')
ggplot(data = df.4cluster)+
geom_bar(aes(x=clust_label, y=total_postsecondary_bound.p),
stat = "summary", fun.y = "var")+
theme_bw()+
#theme(axis.text.x = element_text(angle = 90, hjust = 1, size = 4))+
ylab('Variance of Percent')+
xlab('Cluster')#+
#ggtitle('Post-Secondary bound percent by cluster')
```
3d plot for clusters
```{r}
#aggregating for mean race by school
df.4clusterm<- (aggregate(x = df.4cluster,
by = list(unique.values = df.4cluster$SCHNO),
FUN = mean))
#using mode
#making mode function
Mode <- function(x) {
ux <- unique(x)
ux[which.max(tabulate(match(x, ux)))]
}
df.4clusterMode <- (aggregate(x = df.4cluster,
by = list(unique.values = df.4cluster$SCHNO),
FUN = Mode))
df.4clusterm$clust_label2 <- df.4clusterMode$clust_label
fig <- plot_ly(data = df.4clusterm, x = ~WH_all.p, y = ~BL_all.p , z =
~HI_all.p, type = 'scatter3d', mode = "markers", color= ~
as.factor(clust_label2), size = 2, colors = c('#4AC6B7', '#1972A4',
'#965F8A', '#FF7070'))
#fig <- fig %>% add_markers()
fig <- fig %>% layout(scene = list(xaxis = list(title='Proportion White'),
yaxis = list(title = 'Proportion
Black'),
zaxis = list(title = 'Proportion
Hispanic')),
#paper_bgcolor = 'rgb(243, 243, 243)',
#plot_bgcolor = 'rgb(243, 243, 243)',
annotations = list(
x = 1.1,
y = 1.05,
text = 'Cluster',
xref = 'paper',
yref = 'paper',
showarrow = FALSE
))
fig
#looking at how many schools in each cluster
table(df.4clusterm$clust_label2)
prop.table(table(df.4clusterm$clust_label2))*100
```
Modeling Cluster Data
```{r}
#modeling the 4 cluster data
#df.4cluster$SURVYEAR <- unfactor(df.4cluster$SURVYEAR)
#setting no title 1 status a reference
```

```
df.4cluster <- within(df.4cluster, STITLI_OVERALL <-
relevel(factor(STITLI_OVERALL), ref = 2))
ri.model.cb <- lmer(college_bound.p ~ factor(clust_label) + SURVYEAR +
factor(STITLI_OVERALL) +(1|SCHNO), df.4cluster, REML = 0)
ri.model.psb <- lmer(total_postsecondary_bound.p ~ factor(cluster) +
SURVYEAR +STITLI_OVERALL + (1|SCHNO), df.4cluster, REML = 0)
ri.model.sdb <- lmer(specialized_degree_bound.p ~ factor(cluster) +
SURVYEAR +STITLI_OVERALL + (1|SCHNO), df.4cluster, REML = 0)
summary(ri.model.cb)
confint(ri.model.cb)
summary(ri.model.psb)
confint(ri.model.psb)
summary(ri.model.sdb)
#plotting coefficients
plot_model(ri.model.psb, show.values = TRUE, vline.color = 'grey', title =
'Total Post-Secondary Bound')+theme_bw()
plot_model(ri.model.cb, show.values = TRUE, vline.color = 'grey', title =
'College Bound')+theme_bw()
#likelihood ratio tests to test overall significance of clusters
model.0c.t <- lmer(total_postsecondary_bound.p ~SURVYEAR +STITLI_OVERALL +
(1|SCHNO), df.4cluster, REML = 0)
model.0c.c <- lmer(college_bound.p ~SURVYEAR +STITLI_OVERALL + (1|SCHNO),
df.4cluster, REML = 0)
anova(model.0c.t, ri.model.psb, test = 'LRT')
anova(model.0c.c, ri.model.cb, test = 'LRT')
```
```{r}
ggplot(data = df.xyl[df.xyl$SCHNAM == 'SCHENLEY HS'])+
geom_bar(aes(x=Race, y=Percent),
stat = "summary", fun.y = "mean")+
theme_bw()+
#theme(axis.text.x = element_text(angle = 90, hjust = 1, size = 4))+
ylab('Percent')+
xlab('RACE')+
ggtitle('Race Distribution-SCHENLEY HIGH')
```
Race distribution in depth by cluster
```{r}
ggplot(data = df.4clusterl[df.4clusterl$clust_label == 2])+
geom_histogram(aes(x=Percent))+
theme_bw()+
facet_wrap('Race')
ggplot(data = df.4clusterl[df.4clusterl$clust_label == 'white'])+
geom_histogram(aes(x=Percent))+
theme_bw()+
facet_wrap('Race')
ggplot(data = df.4clusterl[df.4clusterl$clust_label == 'hisp'])+
geom_histogram(aes(x=Percent))+
theme_bw()+
facet_wrap('Race')
ggplot(data = df.4clusterl[df.4clusterl$clust_label == 'mix'])+
geom_histogram(aes(x=Percent))+
theme_bw()+
facet_wrap('Race')
```
Figure 2 race/ethnicity mean/ var prop
```{r}
```

```
#creating summary data set of long data
df.dev.r$Percent = df.dev.r$Percent*100
df.dev.r.summary <- df.dev.r %>% # the names of the new data frame and the
data frame to be summarised
group_by(Race,SURVYEAR) %>% # the grouping variable
summarise(mean_Percent = mean(Percent), # calculates the mean of each
group
sd_Percent = sd(Percent), # calculates the standard deviation
of each group
n_Percent = n(), # calculates the sample size per group
SE_Percent = sd(Percent)/sqrt(n())) # calculates the standard
error of each group
#line plot with error bars (too much clutter)
ggplot(df.dev.l.summary, aes(x = unfactor(SURVYEAR), y=mean_Percent, color
= factor(Race)))+
geom_line(stat = "identity")+
geom_errorbar(aes(ymin = mean_Percent - sd_Percent, ymax = mean_Percent +
sd_Percent), width = 0.2, alpha = 0.3) +
theme_bw()+
scale_color_discrete(name = 'Race/Ethnicity',labels=c("AM_all.p" =
"American Indian",
"AS_all.p" = "Asian",
"WH_all.p" = "White",
"BL_all.p"="Black",
"HI_all.p" = "Hispanic",
"HP_all.p" = "Hawaiian/Pacific Islander",
"TR_all.p"="Two or More Races"))+
xlab('School Year')+
scale_x_continuous(n.breaks = 7)
#no error bars (mean)
ggplot(df.dev.l.summary, aes(x = unfactor(SURVYEAR), y=mean_Percent, color
= factor(Race)))+
geom_line(stat = "identity")+
theme_bw()+
scale_color_discrete(name = 'Race/Ethnicity',labels=c("AM_all.p" =
"American Indian",
"AS_all.p" = "Asian",
"WH_all.p" = "White",
"BL_all.p"="Black",
"HI_all.p" = "Hispanic",
"HP_all.p" = "Hawaiian/Pacific Islander",
"TR_all.p"="Two or More Races"))+
xlab('School Year')+
scale_x_continuous(n.breaks = 7)+
ylab('Overall Mean Percent of Students')
#no error bars (sd)
ggplot(df.dev.l.summary, aes(x = unfactor(SURVYEAR), y=sd_Percent, color =
factor(Race)))+
geom_line(stat = "identity")+
theme_bw()+
scale_color_discrete(name = 'Race/Ethnicity',labels=c("AM_all.p" =
"American Indian",
"AS_all.p" = "Asian",
"WH_all.p" = "White",
"BL_all.p"="Black",
"HI_all.p" = "Hispanic",
"HP_all.p" = "Hawaiian/Pacific Islander",
```

```
"TR_all.p"="Two or More Races"))+
xlab('School Year')+
scale_x_continuous(n.breaks = 7)+
ylab('Standard Deviation of Percent')
#old version of plot
ggplot(df.dev.l, aes(x = unfactor(SURVYEAR), y=Percent, color =
factor(Race)))+
geom_line(stat = "summary", fun ="mean")+
geom_segment(aes(x = unfactor(SURVYEAR),
y = Low, yend = High), hjust = 4) +
theme_bw()+
scale_color_discrete(name = 'Race/Ethnicity',labels=c("AM_all.p" =
"American Indian",
"AS_all.p" = "Asian",
"WH_all.p" = "White",
"BL_all.p"="Black",
"HI_all.p" = "Hispanic",
"HP_all.p" = "Hawaiian/Pacific Islander",
"TR_all.p"="Two or More Races"))+
xlab('School Year')+
scale_x_continuous(n.breaks = 7)
```
Summary Stats by Year Table
```{r}
#creating summary data set of long data
df.dev.l$Percent = df.dev.l$Percent*100
df.dev.l.summary <- df.dev.l %>% # the names of the new data frame and the
data frame to be summarised
group_by(Race,SURVYEAR) %>% # the grouping variable
summarise(mean_Percent = mean(Percent), # calculates the mean of each
group
sd_Percent = sd(Percent), # calculates the standard deviation
of each group
n_Percent = n(), # calculates the sample size per group
SE_Percent = sd(Percent)/sqrt(n())) # calculates the standard
error of each group
```
Summary Stats of Clusters
```{r}
#making long version of cluster data
df.4clusterl.all <- melt(df.4cluster,
# ID variables - all the variables to keep but not split apart on
id.vars=c("SCHNO", "SURVYEAR", "cluster","clust_label"),
# The source columns
measure.vars=c("AM_all.p", "AS_all.p",
"HI_all.p","BL_all.p","WH_all.p","HP_all.p","TR_all.p", 'HUG_prop',
'STITLI_OVERALL', 'college_bound.p', 'total_postsecondary_bound.p' ),
# Name of the destination column that will identify the original
# column that the measurement came from
variable.name="Var",
value.name="Percent"
)
#summarizing everything
df.4clusterl.all$Percent = df.dev.l$Percent*100
df.4clusterl.summary <- df.4clusterl.all %>% # the names of the new data
frame and the data frame to be summarised
group_by(Var,clust_label) %>% # the grouping variable
```

```
summarise(mean_Percent = mean(Percent), # calculates the mean of each
group
sd_Percent = sd(Percent), # calculates the standard deviation
of each group
n_Percent = n(), # calculates the sample size per group
SE_Percent = sd(Percent)/sqrt(n())) # calculates the standard
error of each group
```
Spaghetti Plots
```{r}
p <- ggplot(data = df.4cluster, aes(x = SURVYEAR, y =
total_postsecondary_bound.p, group = SCHNO))
p+geom_line(aes(col = factor(clust_label)),alpha = 0.3)+
facet_wrap('STITLI_OVERALL')+theme_bw()
```
Univariate Plots
```{r}
test <- lmer(total_postsecondary_bound.p ~ HUG_percent + (1|SCHNO),
df.arm1, REML = 0)
summary(test)
test <- lmer(total_postsecondary_bound.p ~ STITLI_OVERALL+ (1|SCHNO),
df.arm1, REML = 0)
summary(test)
test <- lmer(total_postsecondary_bound.p ~ SURVYEAR + (1|SCHNO), df.arm1,
REML = 0)
summary(test)
```

Bibliography

Bloom, D. E. (2005). Education and Public Health: Mutual Challenges Worldwide. *Comparative Education Review*, *49*(4), 437–451. https://doi.org/10.1086/454370

Bock, M. E. (1975). Minimax Estimators of the Mean of a Multivariate Normal Distribution. *The Annals of Statistics*, *3*(1), 209–218. https://doi.org/10.1214/aos/1176343009

College Enrollment Rates (The Condition of Education, p. 4). (2019). National Center of Education Statistics. Retrieved April 10, 2020, from https://nces.ed.gov/programs/coe/pdf/coe_cpb.pdf

Cutler, D., & Lleras-Muney, A. (2006). *Education and Health: Evaluating Theories and Evidence* (No. w12352; p. w12352). National Bureau of Economic Research. https://doi.org/10.3386/w12352

DeCoster, J., Gallucci, M., & Iselin, A.-M. R. (2011). Best Practices for Using Median Splits, Artificial Categorization, and their Continuous Alternatives. *Journal of Experimental Psychopathology*, *2*(2), 197–209. https://doi.org/10.5127/jep.008310

Desjardins, R. (2008). Researching the Links between Education and Well-being: European Journal of Education. *European Journal of Education*, *43*(1), 23–35. https://doi.org/10.1111/j.1465-3435.2007.00333.x

Genesee, F., Lindholm-Leary, K., Saunders, W., & Christian, D. (2005). English Language Learners in U.S. Schools: An Overview of Research Findings. *Journal of Education for Students Placed at Risk (JESPAR)*, *10*(4), 363–385. https://doi.org/10.1207/s15327671espr1004_2

Green, D. (2006). Historically underserved students: What we know, what we still need to know. *New Directions for Community Colleges*, *2006*(135), 21–28. https://doi.org/10.1002/cc.244

Hedeker, D., & Gibbons, R. D. (2006). *Longitudinal data analysis*. Wiley-Interscience.

Hummer, R. A., & Hernandez, E. M. (2013). The Effect of Educational Attainment on Adult Mortality in the United States. *POPULATION BULLETIN*, 20.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning* (Vol. 103). Springer New York. https://doi.org/10.1007/978-1-4614-7138-7

Kodinariya, Trupti & Makwana, P.R.. (2013). Review on Determining of Cluster in K-means Clustering. *International Journal of Advance Research in Computer Science and Management Studies*. 1. 90-95.

Lim, H., Lee, J. M., & Kim, K. T. (2019b). What Factors Are Important in Aversion to Education Debt? *Family and Consumer Sciences Research Journal*, *48*(1), 5–21. https://doi.org/10.1111/fcsr.12324

McFarland, J. (2018). *Trends in High School Dropout and Completion Rates in the United States: 2018*. 101.

Measuring the value of education: Career Outlook: U.S. Bureau of Labor Statistics. (2018). Retrieved February 1, 2020, from https://www.bls.gov/careeroutlook/2018/data-on-display/education-pays.htm

Mirowsky, J., & Ross, C. E. (2015). Education, Health, and the Default American Lifestyle. *Journal of Health and Social Behavior*, *56*(3), 297–306. https://doi.org/10.1177/0022146515594814

Murnane, R. J. (2013). U.S. High School Graduation Rates: Patterns and Explanations. *Journal of Economic Literature*, *51*(2), 370–422. https://doi.org/10.1257/jel.51.2.370

Perez, C. P., & Morrison, S. S. (2016). *Understanding the Challenges of English Language Learners and Increasing College-Going Culture: Suggestions for School Counselors*. 12.

Racial/Ethnic Enrollment in Public Schools (The Condition of Education, p. 4). (2017). National Center of Education Statistics.

Reimherr, P., Harmon, T., Strawn, J., & Choitz, V. (2013). *How to Simplify Tax Aid and Use Performance Metrics to Improve College Choices and Completion*. 80.

Richard Kitchen, & Sarabeth Berk. (2016). Educational Technology: An Equity Challenge to the Common Core. *Journal for Research in Mathematics Education*, *47*(1), 3. https://doi.org/10.5951/jresematheduc.47.1.0003

Richard Kitchen, & Sarabeth Berk. (2017). Keeping the Focus on Underserved Students, Privilege, and Power: A Reaction to Clements and Sarama. *Journal for Research in Mathematics Education*, *48*(5), 483. https://doi.org/10.5951/jresematheduc.48.5.0483

Trends in Student Aid 2019 (Trend in Higher Education, p. 36). (2019). College Board.

Zajacova, A., & Lawrence, E. M. (2018). The Relationship Between Education and Health: Reducing Disparities Through a Contextual Approach. *Annual Review of Public Health*, *39*(1), 273–289. https://doi.org/10.1146/annurev-publhealth-031816-044628