An Analysis of the Department of Education Quality Survey and Its Efficacy

Andrew Bowen[1], Glen Dale Davis[1], Josh Forster[1], Shoshana Farber[1], & Charles Ugiagbe[1]

[1] City University of New York

Abstract

We present a study on the effectiveness of New York City School Quality Snapshot in predicting the 4-year college persistence rate for New York City high schools. These surveys are used in the deciding of educational policy across the United States. While input from educators and families can be invaluable, it often can not reflect the underlying factors that most influence academic performance. We also build a predictive model based on proxy socioeconomic factors (presence in temporary housing, economic need) that can also be used to predict the average college persistence of a high school in NYC. Our model based on socioeconomic indicators outperforms a model based solely on results of the school quality snapshot, which includes survey responses from students, parents, and educators within NYC public schools. This increased performance comes across several model diagnostic statistics, including root mean-squared error, Akaike and Bayesian Information Criteria, and adjusted R-squared. Additionally, a weighted least-squares model is created on the same set of proxy variables to compare modeling techniques. This model outperforms the ratings-based linear model, but introduces additional complexity for less explainability when compared with the direct proxy variable model.

Keywords: Educational Outcomes, School Quality, Education

An Analysis of the Department of Education Quality Survey and Its Efficacy

Introduction

The NYC School Survey seeks to collect data to provide an overview of New York City (NYC) Schools. First conducted in 2005, the survey gathers demographic and achievement data for NYC Public Schools and provides a standardized rating of various elements of school quality.

The survey has changed over the years. These changes have come from the recommendations of public policy analysts seeking to more accurately define the quality of schools *New York City Schools (2018)*. The 2020-21 academic year report provides a robust dataset of school-level observations of academic and socioeconomic data.

Research Question: Our analysis aims to determine whether NYC School Quality Survey ratings accurately reflect educational outcomes or if these outcomes could be better predicted by proxy variables related to the student body.

The primary measure of success we aim to predict is the 4-year college persistence rate for NYC high schools. This measure is defined as the percentage of students who graduate from a high school and eventually complete a 4-year college program. Identifying the key indicators of a school's ability to successfully prepare students for college could benefit the NYC Department of Education (DOE) and NYC Public Schools in several ways:

1. It would provide insights to the NYC DOE and NYC Public Schools which would enable them to tailor instructional approaches and develop targeted curricula that specifically address college preparedness.
2. It would allow for strategic allocation of resources to address identified areas that significantly impact college readiness, ensuring that resources are utilized efficiently to increase the percentage of college-ready students across NYC Public Schools.

It is well-established that attending 4-year institutions significantly enhances career potential earnings. Ensuring that high school students are adequately prepared for their college careers not only benefits their immediate educational success but also contributes to their long-term success in life.

Literature Review

One of the main predictors of academic performance is a student's socioeconomic background. According to the National Center for Education Statistics (NCES), students from low-income families are nearly four times more likely to drop out of high school than students from wealthy families *Education Statistics (2008).*

Several prior studies have made attempts to use more sophisticated modeling techniques, different data sources, and different predictor variables to predict educational outcomes similar to what we're trying to predict. In one such study, *Bernacki, Chavez, and Uesbeck (2020)* based their modeling on trying to predict educational achievement based on student digital behavior, rather than the social factors we intend to explore. The model in this study reached an accuracy of 75%, and was able to flag early interventions. This modeling technique attempts to predict a slightly different metric of student success than our modeling will, and the training data and predictor variables differ as well.

Similarly, *Musso, Cascallar, Bostani, and Crawford (2020)* attempted to train an artificial neural network (ANN) to identify relationships between variables and educational performance data. They modeled educational performance of Vietnamese students in grade five and included individual characteristics as well as information related to daily routines in their training data. This method uses a more sophisticated model, and resulted in an impressive prediction accuracy of $95 - 100$. However, as their training data comes from a different country with a different educational system and methods, it may not be prudent to compare the model's results to those of our model or of any other US-centric study.

In another study, *Yağcı (2022)* predicted final grade exams for Turkish students through machine learning models, using prior exam scores as their input variables. While this provides a valuable metric for academic performance, concerns arise regarding the direct correlation between good exam grades and later career success *Afarian and Kleiner (2003)*. However, a parent study found a correlation of up to 0.3 between academic grades and later job performance *Roth, BeVier, Switzer III, and Schippmann (1996)*, so it may be worthwhile to consider this metric as a measure to predict later success in life. Further analysis would have to be conducted in this respect.

Measuring which predictors impact educational outcomes and how much is a difficult task. There are generally many confounding variables related to the student body being observed, and causal relationships can be difficult if not impossible to establish.

Data Sourcing

The dataset used in this study is published in the NYC School Quality Report for the Academic Year 2020 - 2021. It consists of data from 487 NYC Public Schools, and there are 391 variable columns. The observations are all school-level, indexed by each school's *District Borough Number* (DBN).

In addition to the school quality ratings based on survey responses, average and raw academic performance data are included as well. There are also socioeconomic variables, such as a school's percentage of students in temporary housing services.

Methodology

Our primary interest is finding proxy variables within the data that can better serve as predictors of 4-year college persistence rates at a given NYC high school than the school survey ratings collected by the quality review. Toward this end, we will need to first construct a baseline model that predicts a school's college persistence rate.

We will attempt to use three variables as a proxy for the school's survey rating in predicting college persistence:

- `temp_housing_pct`: the percentage of students living in temporary housing
- `eni_hs_pct_912`: Economic Need Index: a measure of the percentage of students facing economic hardship at a school[1]
- `val_chronic_absent_hs_all`: the percentage of students who are chronically absent[2]

We begin by taking a look at a summary of the dataset's completeness.

Table 1

Completeness Summary

rows	487
columns	391
all_missing_columns	12
total_missing_values	47332
complete_rows	0

There are 12 columns that are completely devoid of data, so we identify and remove those.

[1] **noauthor_student_2021 (fix, not in references)** Economic hardship in this context is based on three criteria: whether the student is 1) eligible for public assistance from the NYC Human Resources Administration (HRA); 2) lived in temporary housing in the past four years; 3) is in high school, has a home language other than English, and entered the NYC DOE for the first time within the last four years.

[2] Chronic absenteeism is defined by the NYC DOE as "students who are absent 10 percent or more of the total days."

Table 2

All NA Columns
QR_1_1
QR_1_2
QR_2_2
QR_3_4
QR_4_2
QR_1_4
QR_1_3
QR_3_1
QR_4_1
QR_5_1
Dates_of_Review
principal

We create a 20% holdout set of data to be used later on in order to evaluate the efficacy of our model's predictive capability. The remaining 80% of the data is to be used for model training and exploratory data analysis (EDA).

For ease of single-node computation, we reduce our primary datasets to our primary variables of interest. Notably, these are the survey ratings, enrollment levels, and our preferred proxy variables for each school. We retain secondary datasets, which contain all variables for which there are no missing values, for later use when developing certain models.

We take a look at whether the reduced training dataset contains any missing values and what the spread is.

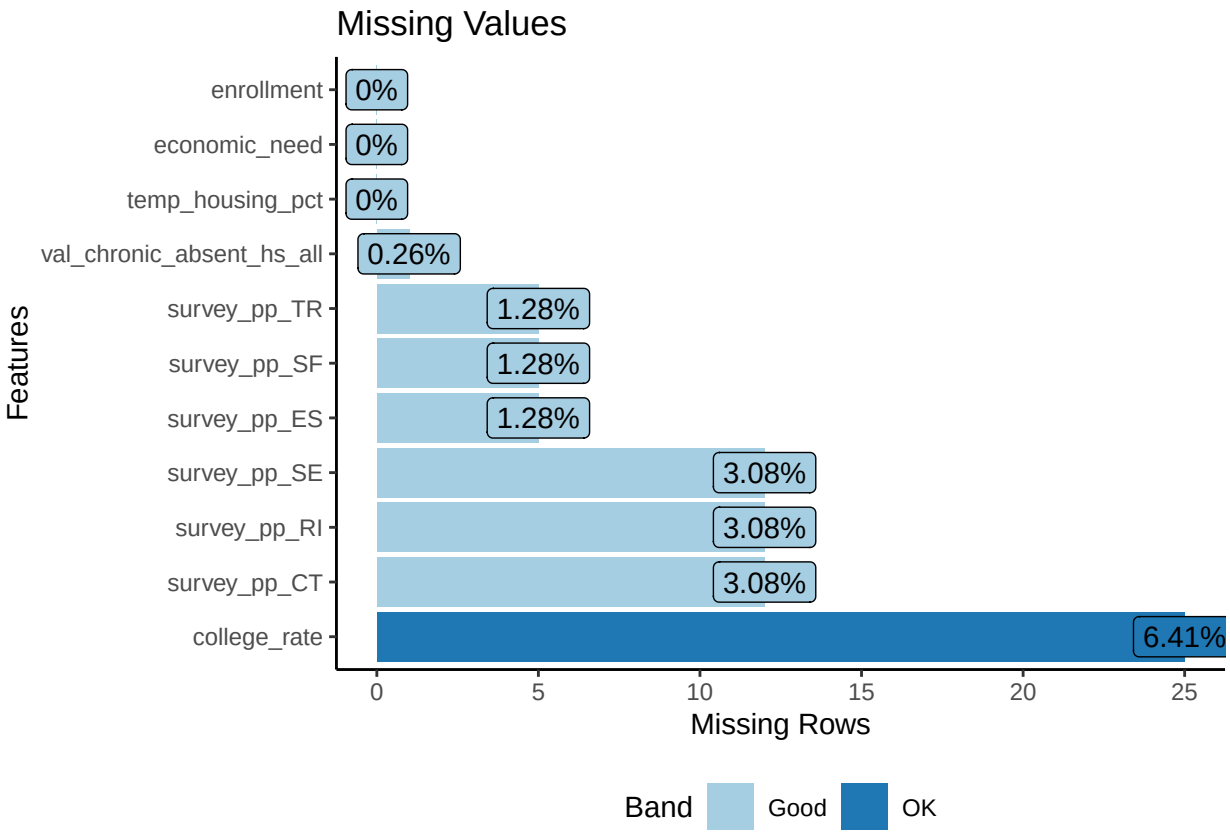


Figure 1

The variable with the most missing data is `college_rate`. Some schools are also missing some survey ratings, and a very small percentage of schools are missing chronic absenteeism values.

We impute both our training and evaluation datasets. Given we are dealing with continuous numeric (and not categorical variables), we use the *Predictive Mean Matching* imputation method native to the R `mice` package.

To check underlying modeling assumptions, we plot distributions and relationships of different variables. First, we plot the distribution of college persistence rates among NYC high schools to check for normality.

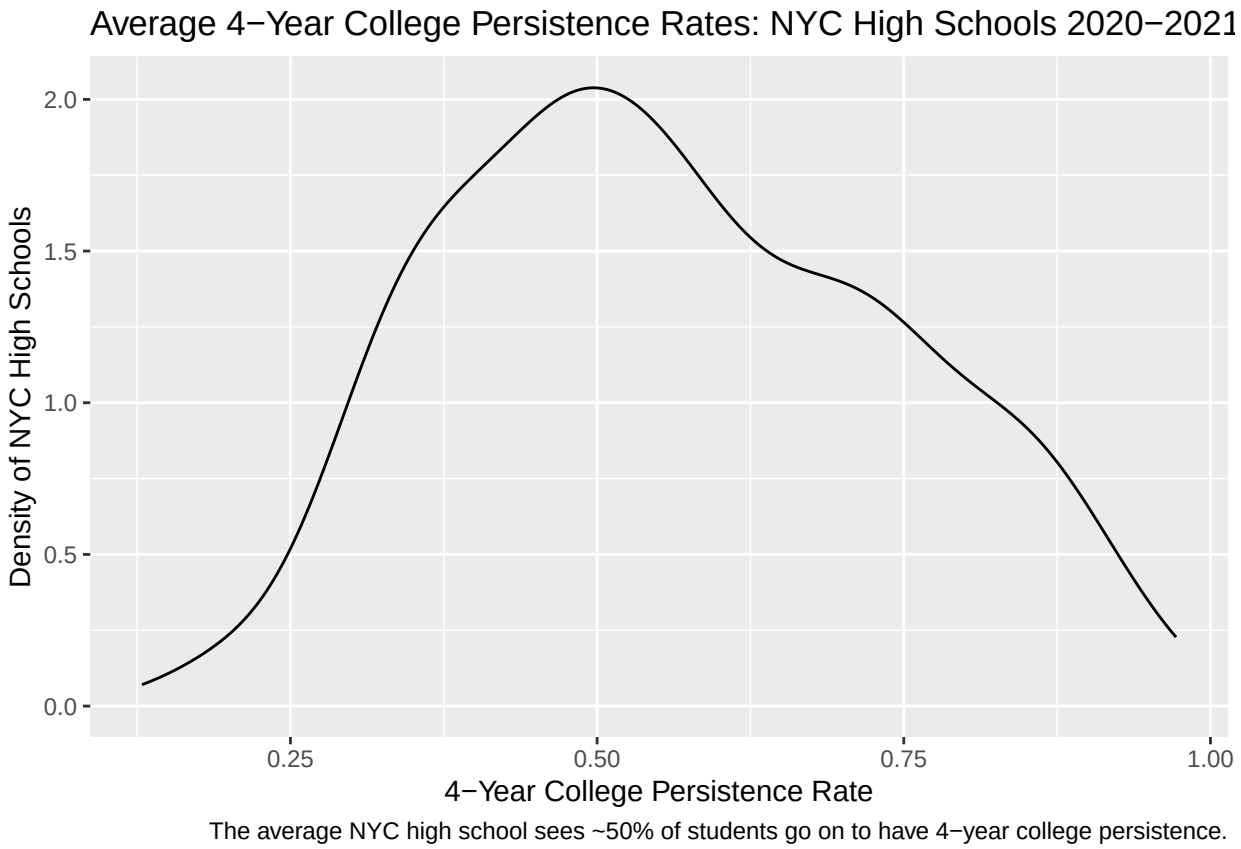


Figure 2

We see a relatively normal distribution of college persistence rates. In the case of NYC high schools, the peak is at around 50%. This is inline with national averages released by the *US Census Bureau (2023)*.

The below plot shows the raw correlation between each variable in our pared down dataset (*Collaborative Teaching*, *Trust*, etc) and the response variable of interest: *4-Year College Persistence Rate*.

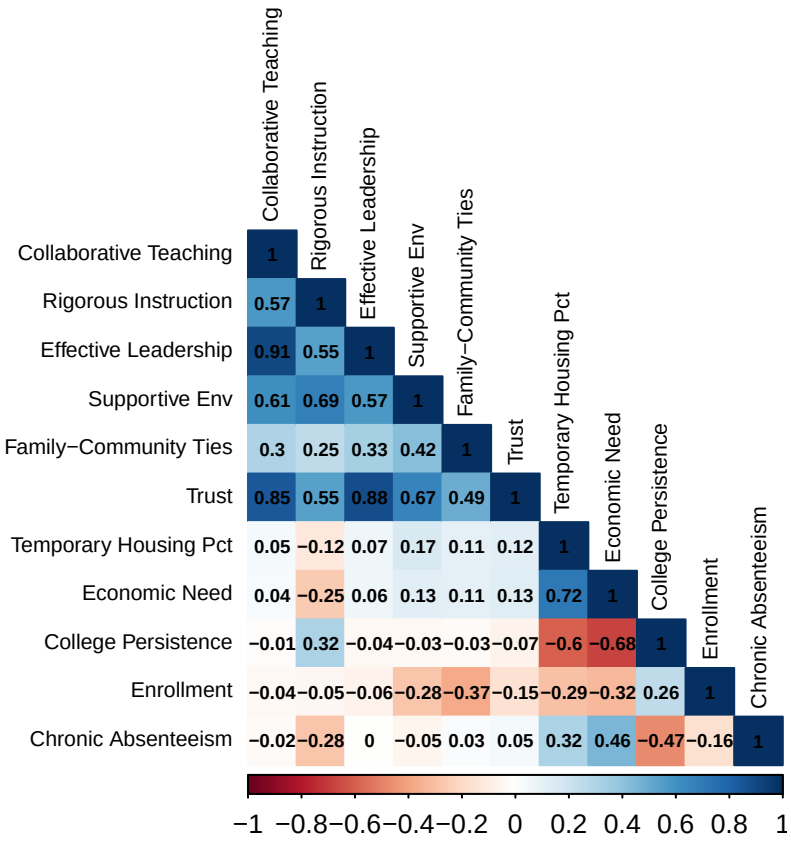


Figure 3

From our correlation plot above, we can see strong negative relationships between two of our proxy variables of interest (*Temporary Housing Rate* and *Economic Need Index*) and our target variable: *College Persistence Rate*. There is also a negative relationship between *Chronic Absenteeism* and *College Persistence Rate*, but to a lesser degree. This gives signal that constructing models based on these variables could give good insight into the factors that most influence college persistence.

Enrollment has only a slightly positive relationship with *College Persistence Rate*. We expected school size might be important when modeling, but that does not appear to be likely.

We also see that the survey ratings are all at least somewhat positively correlated with

one another, and the only survey rating that appears to have a relationship with *College Persistence Rate* is *Rigorous Instruction*. That relationship is only slightly positive. This signals that constructing a model based on one or more of the survey ratings might not give as much insight into college persistence as the proxy variables could.

Now we can plot the distributions of our proxy variables of interest.

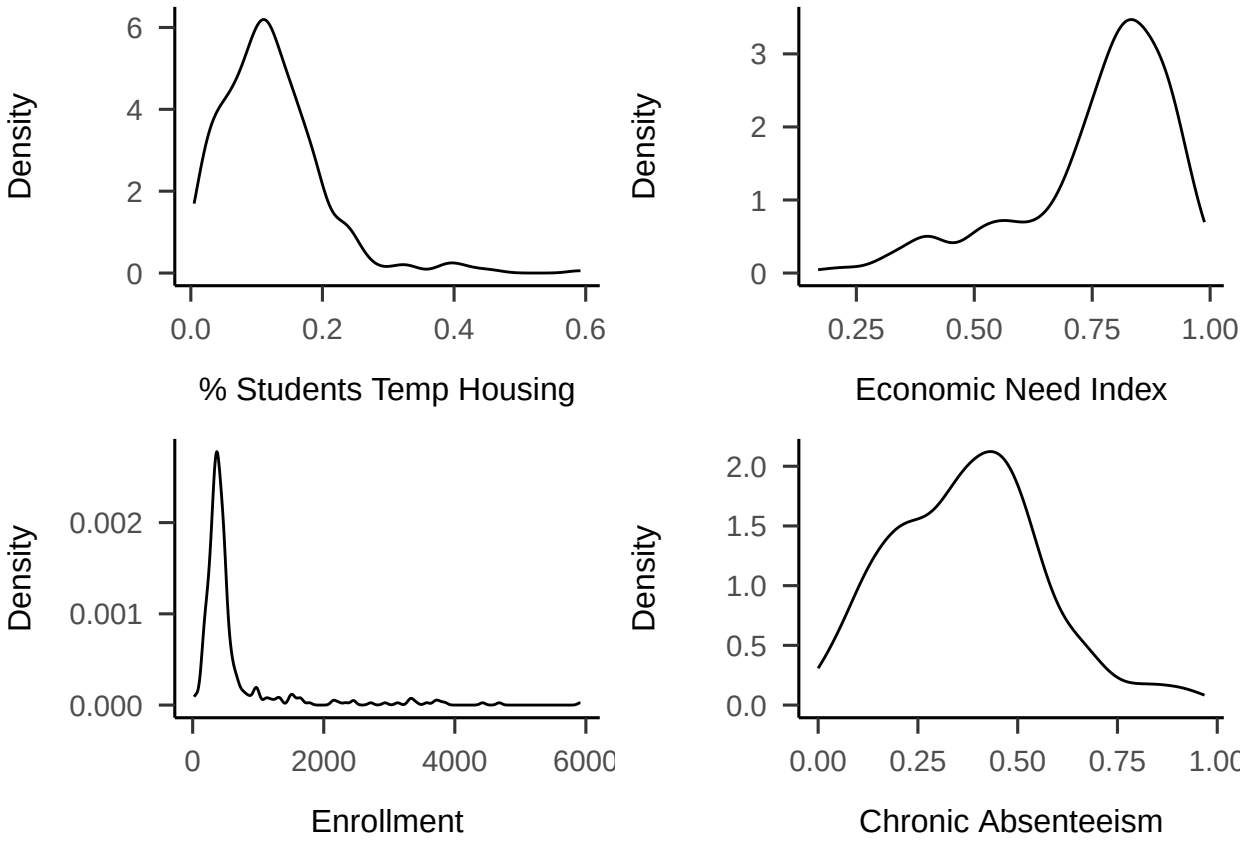


Figure 4

We see the distribution of *Temporary Housing Rate* is right-skewed. We also see the distribution of *Economic Need Index* is left-skewed. The closer the index is to 1, the more economic hardship students at that school face, so schools with high rates of students facing economic hardship are more prevalent than schools with low rates. These variables are both candidates for transformation due to their skew. Our model will not likely feature *Enrollment*, as observations are so concentrated at the low end, and we already noted it is

not as correlated with our target variable as the other proxies we're considering. *Chronic Absenteeism* is closer to a normal distribution than the other variables, but it is still slightly right-skewed, so there are more schools in this dataset with pretty low rates and fewer schools with pretty high rates.

We check an assumption of linearity between our proxy predictors and our response variable by producing scatter plots of the response variable versus each of the proxy predictors.

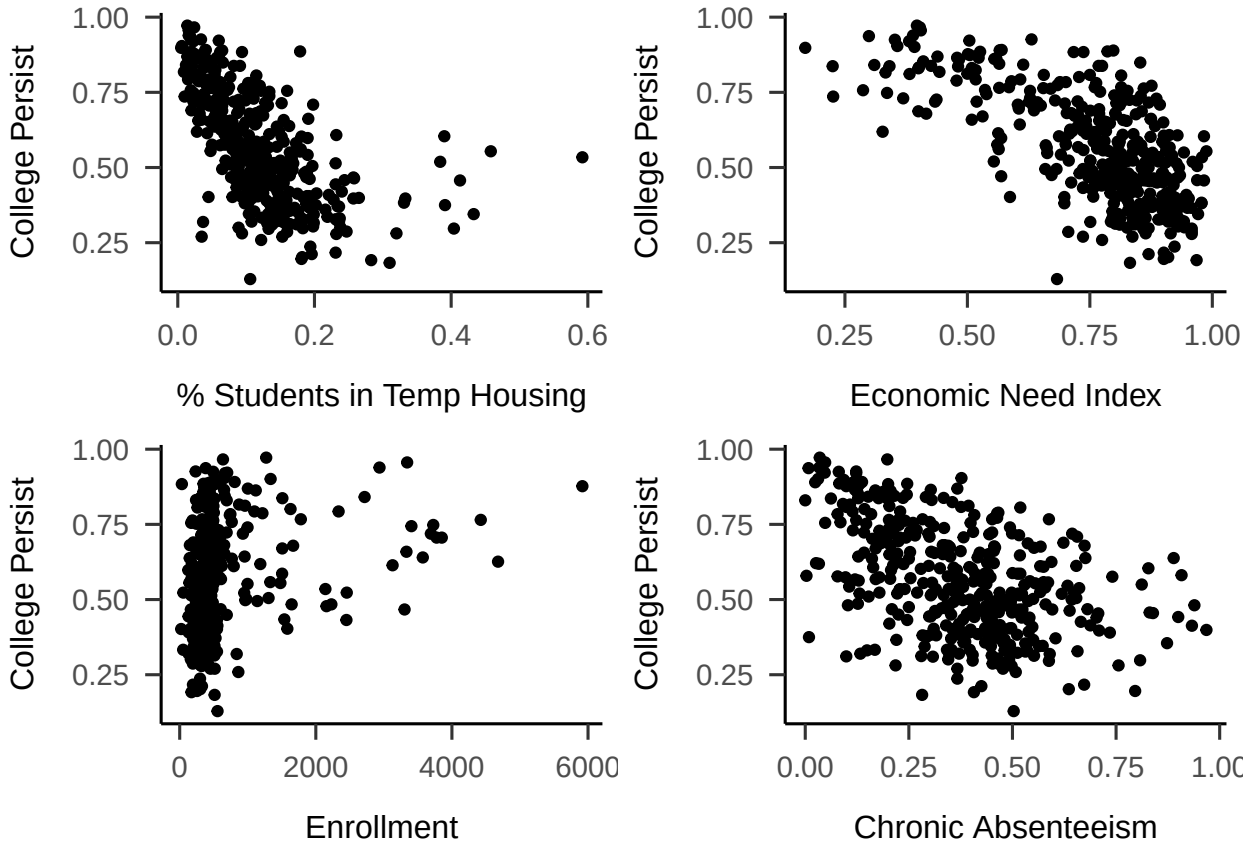


Figure 5

We see a generally negative linear relationship between the response variable and rates of students in temporary housing. As that rate increases, college persistence tends to decrease. However, that relationship does **not** appear to hold for schools with higher rates of

students in temporary housing. So the relationship cannot be completely captured by a linear trend.

We also see a non-linear relationship between the response variable and the economic need index.

Schools with lower enrollment levels have a wider range of college persistence rates than schools with higher enrollment levels.

Only one school where chronic absenteeism is greater than or equal to 50 percent achieves college persistence levels above 80 percent. However, college persistence varies widely at most chronic absenteeism levels.

Modeling

For evaluation purposes, we create a linear model based on the survey ratings present per school in our data. We fit this multiple least-squares model to predict the college persistence rate of a given high school. The model summary is printed below:

```
##
## Call:
## lm(formula = base_formula, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.5405 -0.1119  0.0053  0.1135  0.4303
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.5399     0.1976   2.732  0.00659 **
```

```
## survey_pp_CT   0.1150     0.2635   0.436  0.66281
## survey_pp_RI   2.1733     0.1976  11.001  < 2e-16 ***
## survey_pp_SE  -1.5105     0.2664  -5.671  2.8e-08 ***
## survey_pp_ES  -0.3090     0.2802  -1.103  0.27079
## survey_pp_SF   0.2349     0.2131   1.102  0.27109
## survey_pp_TR  -0.4708     0.4237  -1.111  0.26724
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1581 on 383 degrees of freedom
## Multiple R-squared:  0.2495, Adjusted R-squared:  0.2377
## F-statistic: 21.22 on 6 and 383 DF,  p-value: < 2.2e-16
```

We find our base model for the school survey ratings produces an adjusted R-squared of $R^2_{adj} = 0.24$. This is lower than the predictive model in *Roth et al. (1996)* produces. The two survey ratings that appear to be statistically significant to the model are *Rigorous Instruction*, which we expected based on our correlation analysis, and *Supportive Environment*, which we did not expect. We reduce the model via backward selection, and *Effective Leadership* becomes statistically significant as well. We reprint a summary below:

```
##
## Call:
## lm(formula = college_rate ~ survey_pp_RI + survey_pp_SE + survey_pp_ES,
##     data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.53159 -0.11178  0.00553  0.11225  0.46053
```

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.5234     0.1369   3.824 0.000153 ***
## survey_pp_RI   2.1816     0.1951  11.182  < 2e-16 ***
## survey_pp_SE  -1.5291     0.2379  -6.426 3.86e-10 ***
## survey_pp_ES  -0.4134     0.1232  -3.355 0.000873 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1579 on 386 degrees of freedom
## Multiple R-squared:  0.2459, Adjusted R-squared:  0.2401
## F-statistic: 41.97 on 3 and 386 DF,  p-value: < 2.2e-16
```

The adjusted R-squared is the same due to rounding. We check for suspected multicollinearity within this model:

Table 3

Variance Inflation Factors

	VIF Value
survey_pp_RI	2.04
survey_pp_SE	2.12
survey_pp_ES	1.59

Surprisingly, none of the variance inflation factors are greater than five, so there are no multicollinearity issues to address for this model.

Let's look at some diagnostic plots for this model.

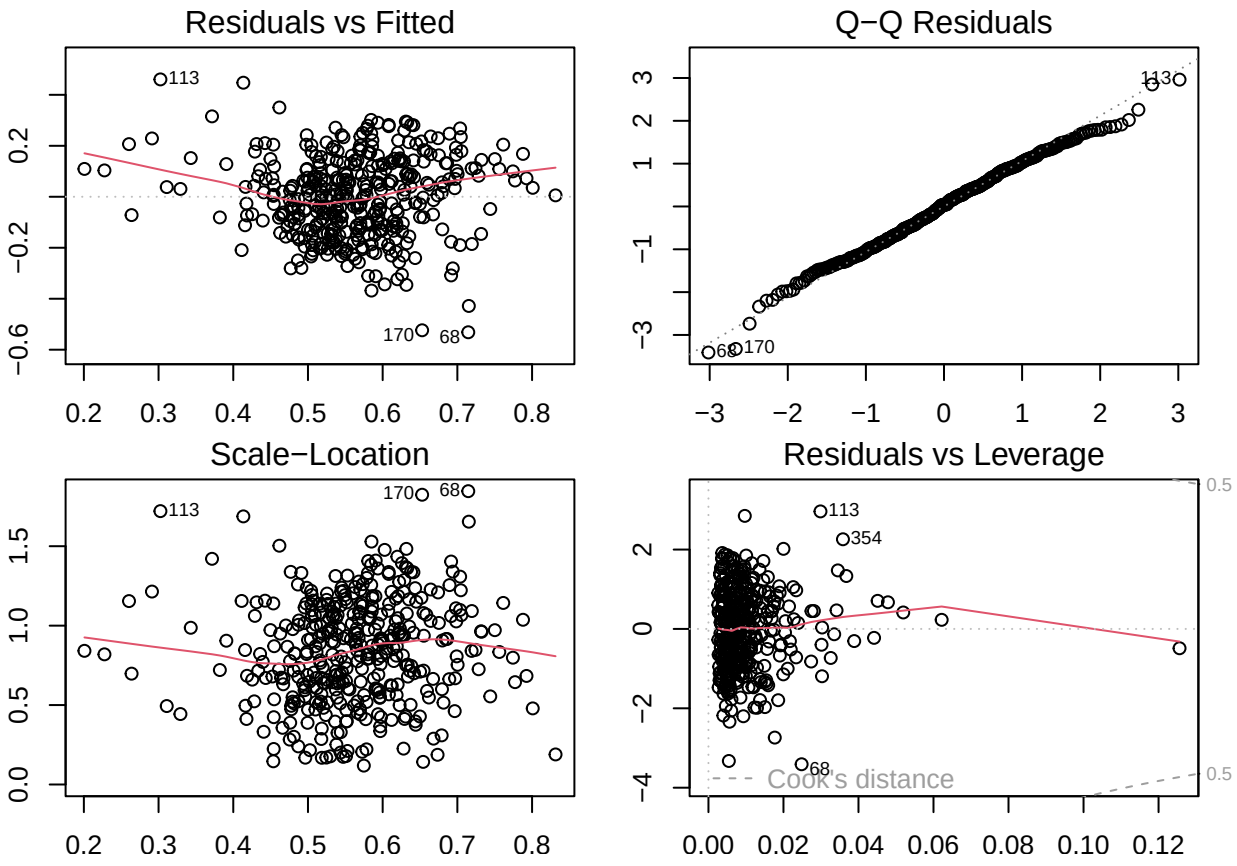


Figure 6

The residuals look relatively normal from the QQ-plot but there seems to be some clustering in the residuals vs fitted plot and we can see a slight curving.

We then create a basic multiple least squares linear model between the response and our three socioeconomic proxy variables: *Temporary Housing Rate*, *Economic Need Index*, and *Chronic Absenteeism*. We include *Enrollment* as well. The summary statistics of the socieoeconomic model are shown below.

```
##
## Call:
## lm(formula = proxy_formula, data = train)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.45147 -0.08833  0.00416  0.08316  0.31536
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  1.045e+00  3.776e-02  27.675  < 2e-16 ***
## temp_housing_pct            -5.449e-01  1.193e-01  -4.565 6.72e-06 ***
## economic_need               -4.528e-01  6.124e-02  -7.394 8.97e-13 ***
## val_chronic_absent_hs_all -2.008e-01  3.805e-02  -5.279 2.18e-07 ***
## enrollment                   6.561e-06  9.261e-06   0.708    0.479
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1261 on 385 degrees of freedom
## Multiple R-squared:  0.5198, Adjusted R-squared:  0.5149
## F-statistic: 104.2 on 4 and 385 DF,  p-value: < 2.2e-16
```

Enrollment is not statistically significant, so we remove it and reprint a summary.

```
##
## Call:
## lm(formula = college_rate ~ temp_housing_pct + economic_need +
##     val_chronic_absent_hs_all, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.45236 -0.08473  0.00471  0.08199  0.31786
```

```
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 1.05498    0.03494  30.192  < 2e-16 ***
## temp_housing_pct           -0.55304    0.11872  -4.659 4.39e-06 ***
## economic_need              -0.45928    0.06053  -7.588 2.46e-13 ***
## val_chronic_absent_hs_all -0.20141    0.03802  -5.298 1.97e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1261 on 386 degrees of freedom
## Multiple R-squared:  0.5192, Adjusted R-squared:  0.5155
## F-statistic:   139 on 3 and 386 DF,  p-value: < 2.2e-16
```

We find our proxy socioeconomic model produces an adjusted R-squared of $R^2_{adj} = 0.52$. We also check for multicollinearity within this model. Unlike with the base model based on the survey ratings, we do not expect any such issues with this model.

Table 4

Variance Inflation Factors

	VIF Value
temp_housing_pct	2.08
economic_need	2.35
val_chronic_absent_hs_all	1.26

None of the variance inflation factors are greater than five, so there are no multicollinearity issues to address for this model.

We produce diagnostic plots for the model below.

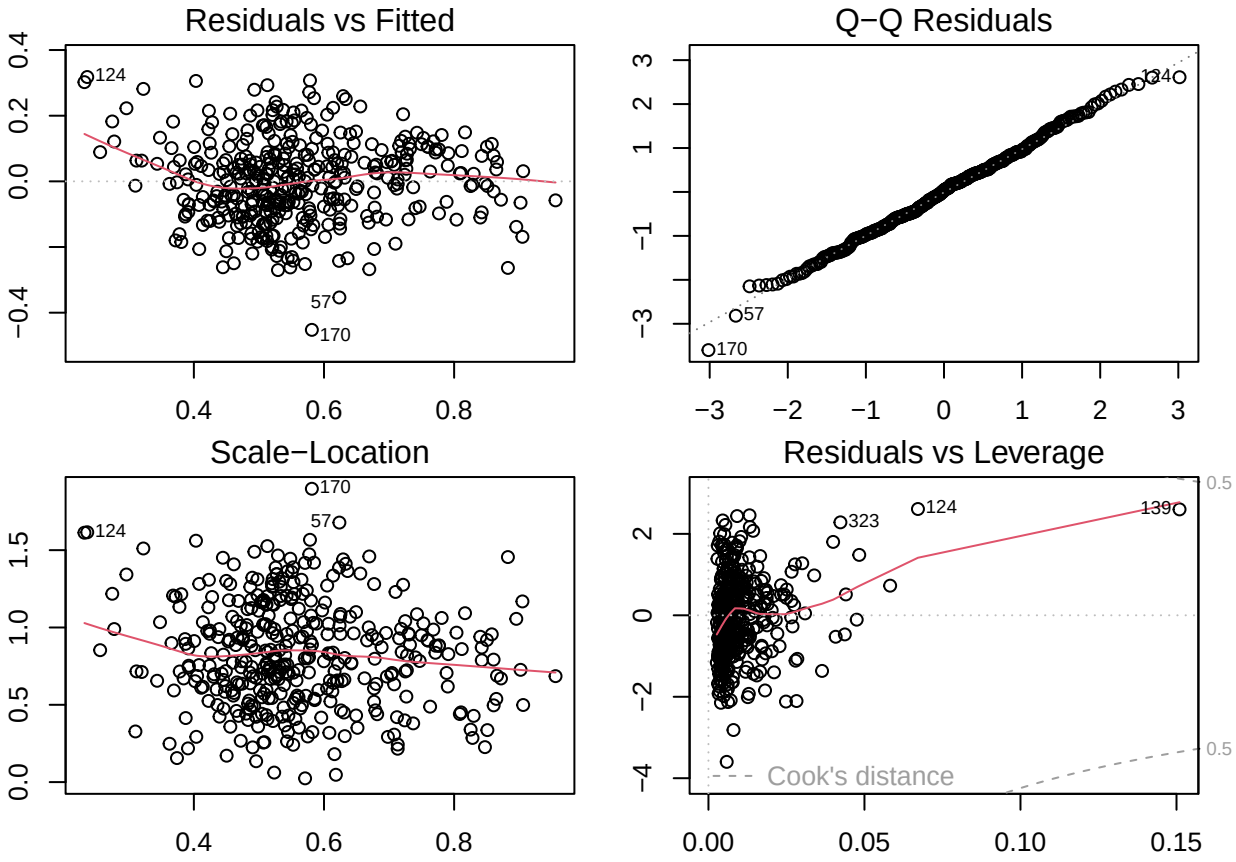


Figure 7

We see no strong trend in the residuals vs. fitted plot, indicating heteroscedasticity. The residuals also look to be normally distributed.

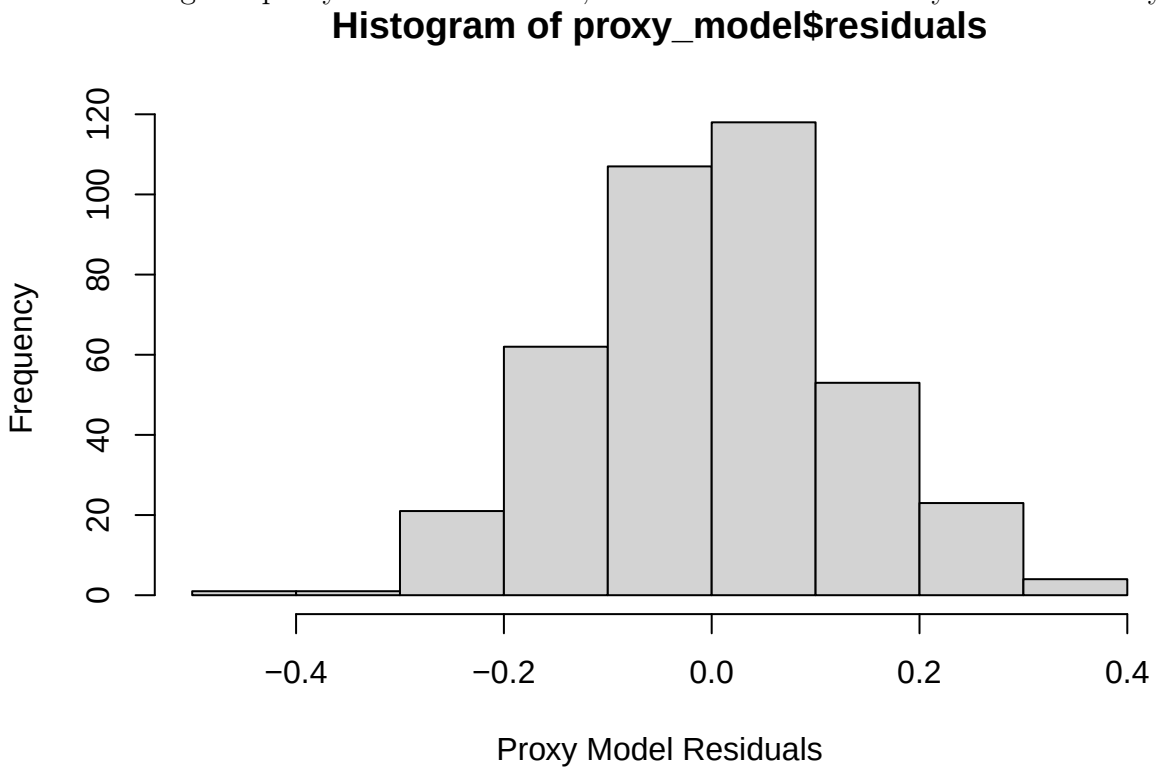
We can also test the assumption of normally-distributed residuals via a Shapiro-Wilk test for normality. Here we operate with the null H_0 and alternative hypotheses H_a:

- H_0: the error terms of the socioeconomic proxy model come from a normally-distributed population
- H_a: the error terms of the socioeconomic proxy model come from a population that is **not** normally distributed

```
##
##  Shapiro-Wilk normality test
##
## data:  proxy_model$residuals
## W = 0.99654, p-value = 0.5668
```

Running a Shapiro test for normality at a 95% threshold, we receive a p-value of 0.5848, higher than our threshold, so we cannot reject our null hypothesis.

Plotting our proxy model's residuals, we can confirm normality as well visually:



We also fit a weighted-least squares (WLS) model to our proxy variables to account for unqual variances among measurements' residuals. The approach can be used to mitigate the effects of heteroscedastic data when modeling.

```
##
## Call:
```

```
## lm(formula = proxy_formula, data = train, weights = weights)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4992 -0.8375  0.0549  0.8169  3.1001
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 1.03959    0.03078  33.776  < 2e-16 ***
## temp_housing_pct           -0.68658    0.12684  -5.413 1.09e-07 ***
## economic_need              -0.41046    0.05692  -7.211 2.96e-12 ***
## val_chronic_absent_hs_all -0.21632    0.03831  -5.646 3.19e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.244 on 386 degrees of freedom
## Multiple R-squared:  0.5589, Adjusted R-squared:  0.5554
## F-statistic:   163 on 3 and 386 DF,  p-value: < 2.2e-16
```

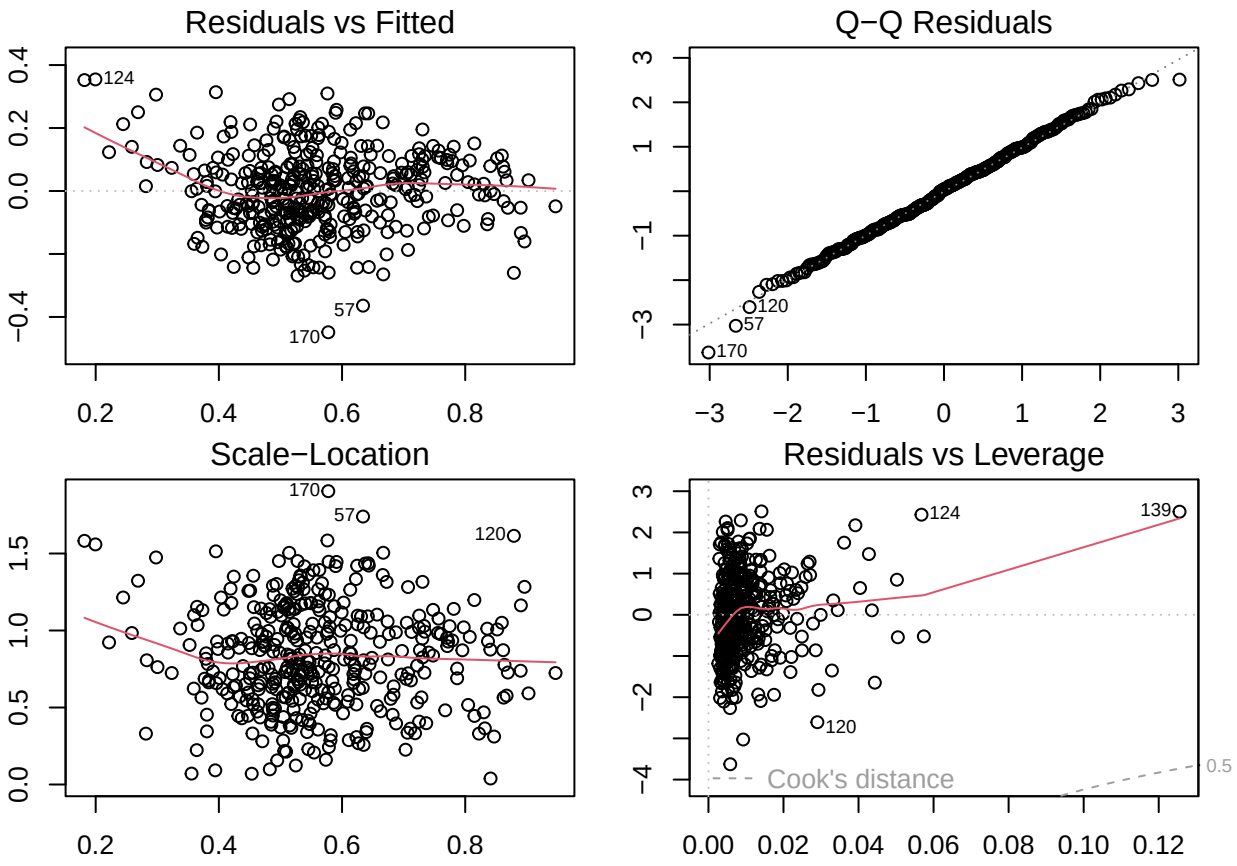


Figure 8

From our diagnostic plots of the WLS model above, we see no general pattern in the *Residuals vs Fitted* plot. In addition, our Q-Q plot shows general normality behavior, with some behavior off the trendline near the tails of the distribution. Similar to above, we can run a Shapiro test on the residuals of the WLS model to assess normality of residuals, with the same alternative and null hypotheses as above:

```
##
##  Shapiro-Wilk normality test
##
## data:  wls_model$residuals
## W = 0.99724, p-value = 0.7594
```

Again, our p-value of 0.7594 indicates that the underlying distribution of residuals is normally distributed.

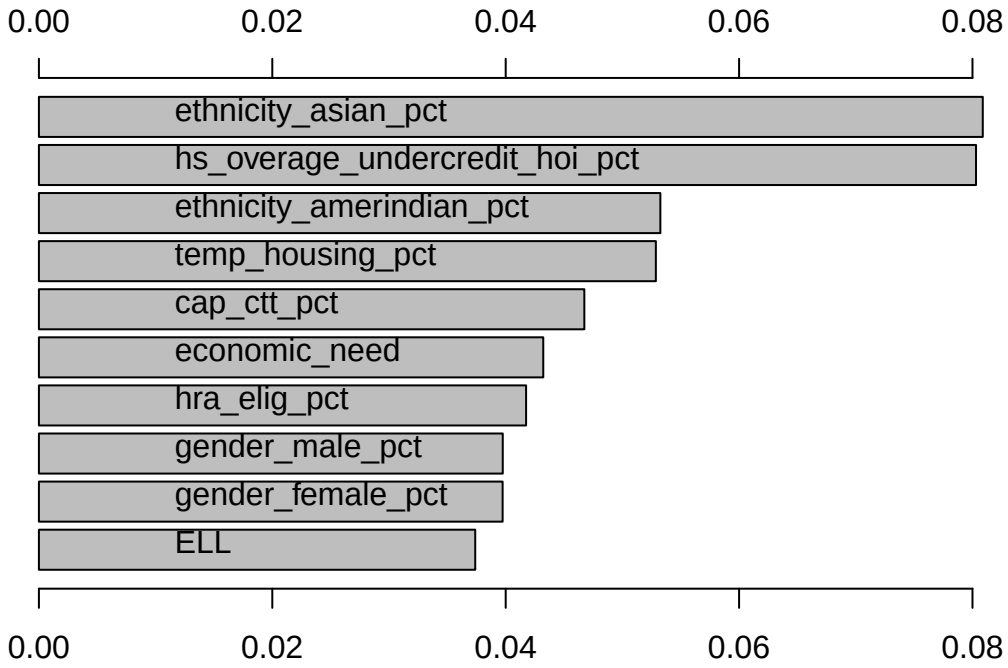
In the interest of determining whether there are other important variables in the original data that we have not considered, or whether a nonlinear approach could reveal relationships in the data that our linear models have not captured so far, we train a Support Vector Machine (SVM): Radial Basis (RB) model on centered and scaled data. SVMs effectively find hyperplanes that divide classes within data well, and they are particularly useful for datasets with large numbers of features. (Predictors with `NA` values in the original dataset have been excluded from consideration due to the undesirably large amount of imputation that would have been required, as SVMs can't handle missing data.) Radial Basis is simply one of several kernel functions we could have chosen, and it effectively defines the shape of the classification boundaries the model makes.

A summary of the ideal tuning parameters and R-Squared value for the SVM:RB model is below:

Table 5

Model	sigma	C	R-Squared
SVM:RB	0.025	0.5	0.6619

A summary of the estimated feature importance for the ten most important features in this SVM:RB model is below:



We can see that the SVM:RB model includes two of our proxy variables in its top ten most important features list: `temp_housing_pct` and `economic_need`. However, the single most important feature to the SVM:RB model is `ethnicity_asian_pct`, and the model considers other race and gender percentages important in predicting `college_rate` as well. It is not surprising that the student body race and gender percentages play a role in college persistence rates, and that could be a fruitful alternative avenue of analysis.

We build a multiple linear regression model using the ten most important features identified in the SVM:RB model, reduced via backward selection.

```
##
## Call:
## lm(formula = college_rate ~ ethnicity_asian_pct + hs_overage_undercredit_hoi_pct +
##     cap_ctt_pct + hra_elig_pct + gender_male_pct + ELL, data = train_svm)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -0.49534 -0.05954  0.00093  0.06959  0.26903
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      1.00518    0.03444  29.186  < 2e-16 ***
## ethnicity_asian_pct              0.23434    0.04685   5.002 8.93e-07 ***
## hs_overage_undercredit_hoi_pct -1.19117    0.12917  -9.221  < 2e-16 ***
## cap_ctt_pct                    -0.43102    0.12800  -3.367 0.000841 ***
## hra_elig_pct                   -0.30990    0.04482  -6.914 2.17e-11 ***
## gender_male_pct                -0.22074    0.04067  -5.427 1.06e-07 ***
## ELL                            -0.08866    0.04444  -1.995 0.046818 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1033 on 358 degrees of freedom
## Multiple R-squared:  0.6838, Adjusted R-squared:  0.6785
## F-statistic:   129 on 6 and 358 DF,  p-value: < 2.2e-16
```

The only predictors that remain after backward selection are `ethnicity_asian_pct`, `hs_overage_undercredit_hoi_pct`, `cap_ctt_pct`, `hra_elig_pct`, `gender_male_pct`, and `ELL`. A higher percentage of Asian students has a positive impact on college persistence rates, while having a higher percentage of overage students who are undercredited, having a higher percentage of students recommended for integrated co-teaching, having a higher rate of HRA eligible students, having a higher percentage of male students, and having a higher percentage of students whose language spoken at home is not English all have a negative impact on college persistence rates. (Note that while `economic_need` was not significant to this model after backward selection, some variables that contribute to this composite variable were: `hra_elig_pct` and `ELL`.)

We find our important features model produces an adjusted R-squared of $R^2_{adj} = 0.68$. We also check for multicollinearity within this model.

Table 6

Variance Inflation Factors

	VIF Value
ethnicity_asian_pct	1.48
hs_overage_undercredit_hoi_pct	2.04
cap_ctt_pct	2.02
hra_elig_pct	1.80
gender_male_pct	1.08
ELL	2.27

None of the variance inflation factors are greater than five, so there are no multicollinearity issues to address for this model.

We produce diagnostic plots for the model below.

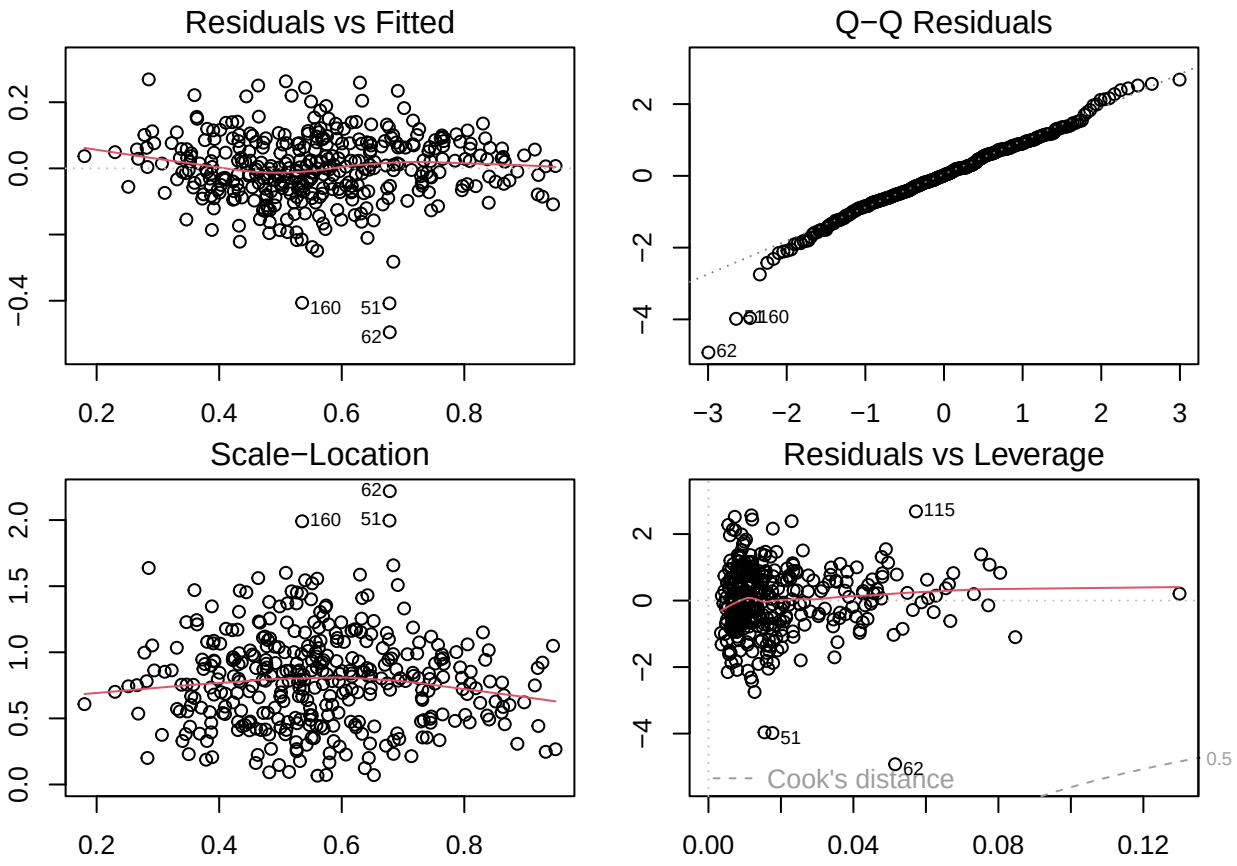


Figure 9

There are some issues revealed by the diagnostic plots. The residuals vs. fitted values aren't evenly distributed around zero, and the Q-Q residuals deviate from the normal line at the low end. There are no high leverage points, but point 62 comes closest and is one of the problematic points in all the diagnostic plots. We run a Shapiro test on the residuals of the Important Features model to assess normality of residuals, with the same alternative and null hypotheses as above.

```
##
##  Shapiro-Wilk normality test
##
## data:  imp_feat_model$residuals
## W = 0.97229, p-value = 1.914e-06
```

The low p-value here suggests the residuals do not come from a normal distribution. This is not surprising after reviewing the diagnostic plots, and it makes sense that building a multiple linear regression model with variables selected by a nonlinear model could produce non-normal residuals.

Lastly, we train a lasso model as an alternative means of feature selection. The ridge-regression penalty λ is fixed at 0, and we will tune over a number of lasso penalties. These lasso penalties penalize the model according to the sum of the absolute values of weights in it, and as such, we end up with a model where many less important variables are given zero weight. (Predictors with `NA` values in the original dataset have again been excluded from consideration due to the undesirably large amount of imputation that would have been required.)

A summary of the ideal tuning parameters and R-Squared value for the lasso model is below:

Table 7

Model	lambda	lasso penalty	R-Squared
Lasso	0	0.05	0.6894

A summary of the estimated feature importance for the ten most important features in this lasso model is below:

Table 8

Predictor	Importance
hs_overage_undercredit_hoi_pct	100.00
temp_housing_pct	87.81
economic_need	76.78
ELL	69.99
hra_elig_pct	68.80
val_attendance_remote_hs_all	50.39
ethnicity_asian_pct	45.16
ethnicity_white_pct	40.30
IEP	39.84
cap_ctt_pct	33.81

In the lasso model, `temp_housing_pct` and `economic_need` are relatively more important than any race or gender percentage variables, unlike in the SVM:RB model. The most important variable is `hs_overage_undercredit_hoi_pct`.

Here is a summary of the predicted coefficients in the lasso model, including those given zero weight:

Experimentation and Results

Model Evaluation. To evaluate our models' quality and performance, we will separate the models into two groups and measure Predictive R-Squared and RMSE using the holdout test data. The first group of models will be the Survey Ratings, Proxy Variables, and Weighted Least-Squares models, which all use a reduced dataset that only includes features we selected based on our research question. This dataset has `NA` values imputed.

The second group of models will be the SVM:RB, Important Features, and Lasso models. This group uses the full dataset (excluding predictors that had `NA` values, as well as observations that had `NA` values in the response). It is important to compare them separately because of the differences in their underlying data. We will only measure AIC and BIC for the first group, which are all linear models for which these metrics make sense.

Table 9

Group	Model	Predictive R-Squared	RMSE
Group 1	Survey Ratings	0.0115	0.1645
Group 1	Proxy Variables	0.2575	0.1426
Group 1	Weighted Least-Squares	0.2483	0.1435

Group 1 Models: In Group 1, the Proxy Variables model has the highest Predictive R-Squared and the lowest RMSE.

We can also use the Akaike and Bayesian Information Criterion for evaluating the complexity of the Group 1 models.

```
##
## Model selection based on AICc:
##
##                          K    AICc Delta_AICc AICcWt Cum.Wt     LL
## Weighted Least-Squares 5 -508.84       0.00   0.96   0.96 259.50
## Proxy Variables        5 -502.43       6.41   0.04   1.00 256.29
## Survey Ratings         5 -326.91     181.93   0.00   1.00 168.53

##
## Model selection based on BIC:
##
```

```
##                          K     BIC Delta_BIC BICWt Cum.Wt     LL
## Weighted Least-Squares 5 -489.16      0.00  0.96   0.96 259.50
## Proxy Variables        5 -482.76      6.41  0.04   1.00 256.29
## Survey Ratings         5 -307.24    181.93  0.00   1.00 168.53
```

From our tables above, we can see smaller values of corrected AIC (which accounts for smaller sample sizes) from the proxy and WLS models than our base survey rating model. This implies better predictive performance for our WLS and proxy-variable models.

Table 10

Group	Model	Predictive R-Squared	RMSE
Group 2	SVM:RB	0.559	0.1097
Group 2	Important Features (MLR)	0.4779	0.1194
Group 2	Lasso	0.5579	0.1098

Group 2 Models: In Group 2, the SVM:RB model has the highest Predictive R-Squared and the lowest RMSE, but it's a very close match between this model and the Lasso model.

Conclusion

Overall, our model to predict a high school's college persistence rate based on socioeconomic proxy variables outperformed the NYC Schools Open Survey Quality ratings of schools. This is not to say that school ratings based on teacher, student, and parent responses are not valuable inputs. However, they should not be the sole basis upon which educational policy decisions are made, considering the collective socioeconomic factors that most influence a school's performance.

Some limitations of our approach would come from conflation between socioeconomic

factors, as well as lacking a more robust imputation method. While our proxy variable model predicts college persistence better than one based off survey ratings, there could be error via omission of unseen variables that are collinear with these inputs. This stems from the availability of data in our source data. We used a *predictive mean matching* imputation method, native to the `mice` R package (*Buuren (2018)*). While this allows for realistic imputed values (no imputed value will fall outside the range of observed data), the underlying population distribution of those values could be non-normal.

Future work could include joining in other academic performance metrics (average SAT/ACT scores, etc.) to see if our proxy variables also have predictive power. The dataset provided is indexed on a high school's *district borough number* (DBN), which is present in several NYC Open Data datasets on education in New York City. As mentioned above, joining in other data sources to augment this data could be a good way to address the omission of any variables that better correlate with college persistence rates.

Overall, identifying the factors that most strongly correlate with academic performance and college persistence can improve in educational policy design. In addition, the public availability of educational data through open data platforms only serves to augment the relationships that help this decision-making.

References

Afarian, R., & Kleiner, B. (2003). The relationship between grades and career success. *Management Research News*, *26*, 42–51. https://doi.org/10.1108/01409170310783781

Bernacki, M. L., Chavez, M. M., & Uesbeck, P. M. (2020). Predicting achievement and providing support before STEM majors begin to fail. *Computers & Education*, *158*, 103999. https://doi.org/https://doi.org/10.1016/j.compedu.2020.103999

Buuren, S. van. (2018). *Fleixble imputation of missing data*. Retrieved from https://stefvanbuuren.name/fimd/sec-pmm.html

Education Statistics, N. C. for. (2008). *Percentage of high school dropouts among persons 16 through 24 years old*. Retrieved from https://nces.ed.gov/programs/digest/d08/tables/dt08_110.asp

Musso, M. F., Cascallar, E. C., Bostani, N., & Crawford, M. (2020). Identifying reliable predictors of educational outcomes through machine-learning predictive modeling. *Frontiers in Education*, *5*. https://doi.org/10.3389/feduc.2020.00104

New York City Schools, T. R. A. for. (2018). *Redesigning the Annual NYC School Survey: Lessons from a Research-Practice Partnership*. https://steinhardt.nyu.edu/sites/default/files/2021-01/Lessons_from_a_Research-Practice_Partnership.pdf.

Roth, P. L., BeVier, C. A., Switzer III, F. S., & Schippmann, J. S. (1996). Meta-analyzing the relationship between grades and job performance. *Journal of Applied Psychology*, *81*(5), 548–556. https://doi.org/10.1037/0021-9010.81.5.548

US Census Bureau. (2023). *Census Cureaur Releases New Educational Attainment Data*. Retrieved from https://www.census.gov/newsroom/press-releases/2023/educational-attainment-data.html

Yağcı, M. (2022). Educational data mining: Prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments*, *9*(1), 11. https://doi.org/10.1186/s40561-022-00192-z

Appendices

Below is the code used to generate this report. It's also available on GitHub here.

```
knitr::opts_chunk$set(echo = FALSE, warning = FALSE, message = FALSE)
library(tidyverse)
library(gridExtra)
library(glue)
library(mice)
library(corrplot)
library(caret)
library(modelr)
library("papaja")
library(DataExplorer)
library(cowplot)
library(car)
library(AICcmodavg)
library(rminer)
library(elasticnet)
r_refs("r-references.bib")
# Read in our dataset from GitHub
# https://www.opendatanetwork.com/dataset/data.cityofnewyork.us/bm9u-cvch
df <- read.csv("https://data.cityofnewyork.us/api/views/26je-vkp6/rows.csv?date=20231108
label_cols <- c("dbn", "school_name", "school_type")
# Convert needed columns to numeric typing
df <- cbind(df[, label_cols], as.data.frame(lapply(df[,!names(df) %in% label_cols], as.


df$college_rate <- df$val_persist3_4yr_all
```

```r
df <- df |>
    select(-val_persist3_4yr_all)
df$economic_need <- df$eni_hs_pct_912
df <- df |>
    select(-eni_hs_pct_912)
remove <- c("discrete_columns", "continuous_columns",
            "total_observations", "memory_usage")
completeness <- introduce(df) |>
    select(-all_of(remove))
apa_table(t(completeness), caption = "Completeness Summary", placement = "H")

find_all_na_cols <- function(dframe){
    col_sums_na <- colSums(is.na(dframe))
    all_na_cols <- names(col_sums_na[col_sums_na == nrow(dframe)])
    all_na_cols
}
all_na_cols <- find_all_na_cols(df)
df <- df |>
    select(-all_of(all_na_cols))
all_na_cols <- as.data.frame(all_na_cols)
colnames(all_na_cols) <- c("All NA Columns")
apa_table(all_na_cols, placement = "H")

set.seed(42)

# Adding a 20% holdout of our input data for model evaluation later
train <- subset(df[sample(1:nrow(df)), ]) %>% sample_frac(0.8)
```

```r
train_svm <- train |> select(-all_of(c("dbn", "school_name", "school_type")))
na_count <-sapply(train_svm, function(y) sum(length(which(is.na(y)))))
na_count <- data.frame(na_count)
na_count <- na_count |>
    filter(na_count == 0)
incl <- rownames(na_count)
train_svm <- train_svm |>
    select(all_of(c("college_rate", incl))) |>
    drop_na(college_rate)
test  <- dplyr::anti_join(df, train, by = 'dbn')
test_svm <- test |> select(-all_of(c("dbn", "school_name", "school_type")))
test_svm <- test_svm |>
    select(all_of(c("college_rate", incl))) |>
    drop_na(college_rate)

cols <- c("survey_pp_CT", "survey_pp_RI",
          "survey_pp_ES", "survey_pp_SE",
          "survey_pp_SF", "survey_pp_TR",
          "temp_housing_pct", "economic_need",
          "college_rate", "enrollment",
          "val_chronic_absent_hs_all")
train_data <- train[, cols]
p1 <- plot_missing(train_data, missing_only = FALSE,
                   ggtheme = theme_classic(), title = "Missing Values")

# Plot missing value percentages by cols of interest
p1 <- p1 +
```

```
    scale_fill_brewer(palette = "Paired")
p1
imp <- mice(train_data, method="pmm", seed=42, printFlag = FALSE)
train <- complete(imp)
test_data <- test[, cols]
imp <- mice(test_data, method="pmm", seed=42, printFlag = FALSE)
test <- complete(imp)
# Plot target variable distribution
ggplot(train, aes(x=college_rate)) +
    geom_density() +
    labs(x="4-Year College Persistence Rate",
         y="Density of NYC High Schools",
         title="Average 4-Year College Persistence Rates: NYC High Schools 2020-2021",
         caption="The average NYC high school sees ~50% of students go on to have 4-year


theme_set(theme_apa())
# Renaming training dataframe for correlation plot
train_renamed <- train %>%
  rename("Collaborative Teaching"=survey_pp_CT,
         "Rigorous Instruction"=survey_pp_RI,
         "Supportive Env"=survey_pp_SE,
         "Effective Leadership"=survey_pp_ES,
         "Family-Community Ties"=survey_pp_SF,
         "Trust"=survey_pp_TR,
         "Temporary Housing Pct"=temp_housing_pct,
         "Economic Need"=economic_need,
         "College Persistence"=college_rate,
```

```
          "Enrollment"=enrollment,
          "Chronic Absenteeism"=val_chronic_absent_hs_all)

# Create correlation plot between vars of interest
corMatrix <- cor(train_renamed)
corrplot(corMatrix, method="color", type="lower", tl.col="black", addCoef.col = "black",

# Plot temp housing rates
pa <- ggplot(train, aes(x=temp_housing_pct)) +
    geom_density() +
    labs(x="% Students Temp Housing", y="Density")
# Plot economic need index
pb <- ggplot(train, aes(x=economic_need)) +
    geom_density() +
    labs(x="Economic Need Index", y="Density")
# Plot enrollment
pc <- ggplot(train, aes(x=enrollment)) +
    geom_density() +
    labs(x="Enrollment", y="Density")
# Plot chronic absenteeism
pd <- ggplot(train, aes(x=val_chronic_absent_hs_all)) +
    geom_density() +
    labs(x="Chronic Absenteeism", y="Density")
p <- plot_grid(pa, pb, pc, pd, nrow = 2, ncol = 2, align = "hv", axis = "t")
p

# Plot temp housing percentage vs college persistence rate
```

```
pa <- ggplot(train, aes(x=temp_housing_pct, y=college_rate)) +
  geom_point() +
  labs(x="% Students in Temp Housing",
       y="College Persist")
# Plot ENI vs college persistence rate
pb <- ggplot(train, aes(x=economic_need, y=college_rate)) +
  geom_point() +
  labs(x="Economic Need Index",
       y="College Persist")
pc <- ggplot(train, aes(x=enrollment, y=college_rate)) +
  geom_point() +
  labs(x="Enrollment",
       y="College Persist")
pd <- ggplot(train, aes(x=val_chronic_absent_hs_all, y=college_rate)) +
  geom_point() +
  labs(x="Chronic Absenteeism",
       y="College Persist")
p <- plot_grid(pa, pb, pc, pd, nrow = 2, ncol = 2, align = "hv", axis = "t")
p


base_formula <- college_rate ~ survey_pp_CT + survey_pp_RI + survey_pp_SE + survey_pp_ES
rating_model <- lm(base_formula,
                   train)
summary(rating_model)


rating_model <- update(rating_model, ~ . - survey_pp_CT - survey_pp_SF - survey_pp_TR)
summary(rating_model)
```

```r
vif_df <- as.data.frame(vif(rating_model))
colnames(vif_df) <- c("VIF Value")
apa_table(vif_df, caption = "Variance Inflation Factors", placement = "H")
par(mfrow=c(2,2))
par(mai=c(.3,.3,.3,.3))
plot(rating_model)

# Create OLS linear model based on our proxy variables: no transforms
proxy_formula <- college_rate ~ temp_housing_pct + economic_need + val_chronic_absent_hs
proxy_model <- lm(proxy_formula, train)
summary(proxy_model)

proxy_model <- update(proxy_model, ~ . - enrollment)
summary(proxy_model)

vif_df <- as.data.frame(vif(proxy_model))
colnames(vif_df) <- c("VIF Value")
apa_table(vif_df, caption = "Variance Inflation Factors", placement = "H")

par(mfrow=c(2,2))
par(mai=c(.3,.3,.3,.3))
plot(proxy_model)

# Test proxy model for normality of residuals
shapiro.test(proxy_model$residuals)
hist(proxy_model$residuals, xlab="Proxy Model Residuals")
```

```r
# Calculating weights for WLS
weights <- 1 / lm(abs(proxy_model$residuals) ~ proxy_model$fitted.values)$fitted.values^

#perform weighted least squares regression
proxy_formula <- proxy_model$call$formula
wls_model <- lm(proxy_formula, data = train, weights=weights)

summary(wls_model)
par(mfrow=c(2,2))
par(mai=c(.3,.3,.3,.3))
plot(wls_model)
shapiro.test(wls_model$residuals)
svmRBTuned <- train(train_svm |> select(all_of(incl)), train_svm$college_rate,
                    method = "svmRadial",
                    preProc = c("center", "scale"),
                    tuneLength = 14,
                    trControl = trainControl(method = "cv"))

svm_summ <- c("SVM:RB",
              round(svmRBTuned$bestTune$sigma, 4),
              svmRBTuned$bestTune$C,
              round(svmRBTuned$results |>
                  filter(C == svmRBTuned$bestTune$C) |>
                  select(Rsquared) |> as.numeric(), 4))
svm_summ <- as.data.frame(t(svm_summ))
cols <- c("Model", "sigma", "C", "R-Squared")
colnames(svm_summ) <- cols
```

```r
apa_table(svm_summ, placement = "H")


y <- train_svm$college_rate
names(y) <- "y"
dat = cbind(train_svm |> select(all_of(incl)), y)
svmRBFit <- fit(y~., data = dat, model = "svm",
                kpar = list(sigma = 0.0244), C = 1)
svmRB.imp <- Importance(svmRBFit, data = dat)
L = list(runs = 1,sen = t(svmRB.imp$imp),
         sresponses = svmRB.imp$sresponses)
sen_vec <- as.numeric(L[["sen"]])
copy <- L
delete <- c()
sort <- sort(sen_vec, decreasing = TRUE)
for (i in 1:length(sen_vec)){
    if (sen_vec[i] > sort[11]){
        next
    }else{
        delete <- append(delete, i)
    }
}
copy[["sen"]] <- t(as.matrix(copy[["sen"]][, -delete]))
copy[["sresponses"]] <- copy[["sresponses"]][-delete]
names <- c()
for (i in 1:length(copy[["sresponses"]])){
    n <- copy[["sresponses"]][[i]][["n"]]
    names <- append(names, n)
```

```
}
mgraph(copy, graph = "IMP", leg = names, col = "gray",
       PDF = "")

imp_feat_form <- college_rate ~ ethnicity_asian_pct + hs_overage_undercredit_hoi_pct + e
imp_feat_model <- lm(imp_feat_form, train_svm)
imp_feat_model <- update(imp_feat_model, . ~ . - ethnicity_amerindian_pct - gender_femal
summary(imp_feat_model)

vif_df <- as.data.frame(vif(imp_feat_model))
colnames(vif_df) <- c("VIF Value")
apa_table(vif_df, caption = "Variance Inflation Factors", placement = "H")

par(mfrow=c(2,2))
par(mai=c(.3,.3,.3,.3))
plot(imp_feat_model)

shapiro.test(imp_feat_model$residuals)
train_lasso <- train_svm
test_lasso <- test_svm
lassoGrid <- expand.grid(.lambda = c(0),
                         .fraction = seq(.05, 1, length = 20))
ctrl <- trainControl(method = "cv", number = 10)
lassoTune <- train(train_lasso |> select(all_of(incl)),
                   train_lasso$college_rate,
                   method = "enet",
                   tuneGrid = lassoGrid,
```

```
                    trControl = ctrl,
                    preProc = c("center", "scale"))

lasso_summ <- c("Lasso",
                lassoTune$bestTune$lambda,
                lassoTune$bestTune$fraction,
                round(lassoTune$results |>
                    filter(fraction == lassoTune$bestTune$fraction) |>
                    select(Rsquared) |> as.numeric(), 4))
lasso_summ <- as.data.frame(t(lasso_summ))
cols <- c("Model", "lambda", "lasso penalty", "R-Squared")
colnames(lasso_summ) <- cols
apa_table(lasso_summ, placement = "H")

lasso_imp <- varImp(lassoTune, scale = TRUE)
cols <- c("Predictor", "Importance")
lasso_imp <- lasso_imp$importance |>
    rownames_to_column()
colnames(lasso_imp) <- cols
lasso_imp <- lasso_imp |>
    arrange(desc(Importance)) |>
    top_n(10)
apa_table(lasso_imp, placement = "H")

lasso_coefs <- as.data.frame(predict.enet(lassoTune$finalModel, s=lassoTune$bestTune[1,
lasso_coefs <- lasso_coefs |>
    rownames_to_column()
```

```r
cols <- c("Predcitor", "Coef")
colnames(lasso_coefs) <- cols
lasso_coefs <- lasso_coefs |>
    arrange(desc(abs(Coef)))


test_pred1 <- predict(rating_model, test |> select(-college_rate))
test_rsq1 <- as.numeric(R2(test_pred1, test$college_rate, form = "traditional"))
test_rmse1 <- as.numeric(RMSE(test_pred1, test$college_rate))
row1 <- cbind("Survey Ratings",
              as.character(round(test_rsq1, 4)),
              as.character(round(test_rmse1, 4)))
test_pred2 <- predict(proxy_model, test |> select(-college_rate))
test_rsq2 <- as.numeric(R2(test_pred2, test$college_rate, form = "traditional"))
test_rmse2 <- as.numeric(RMSE(test_pred2, test$college_rate))
row2 <- cbind("Proxy Variables",
              as.character(round(test_rsq2, 4)),
              as.character(round(test_rmse2, 4)))
test_pred3 <- predict(wls_model, test |> select(-college_rate))
test_rsq3 <- as.numeric(R2(test_pred3, test$college_rate, form = "traditional"))
test_rmse3 <- as.numeric(RMSE(test_pred3, test$college_rate))
row3 <- cbind("Weighted Least-Squares",
              as.character(round(test_rsq3, 4)),
              as.character(round(test_rmse3, 4)))
tbl <- as.data.frame(rbind(row1, row2, row3))
cols <- c("Model", "Predictive R-Squared", "RMSE")
colnames(tbl) <- cols
tbl <- tbl |>
```

```
    mutate(Group = "Group 1") |>
    select(Group, everything())
apa_table(tbl, placement = "H")

model_names <- c("Survey Ratings", "Proxy Variables", "Weighted Least-Squares")
model_list <- list(rating_model, proxy_model, wls_model)

# Print AIC results
aictab(model_list, modnames=model_names)

# Print BIC for each model
bictab(model_list, modnames=model_names)
test_pred1 <- predict(svmRBTuned, test_svm |> select(-college_rate))
test_rsq1 <- as.numeric(R2(test_pred1, test_svm$college_rate,
                           form = "traditional"))
test_rmse1 <- as.numeric(RMSE(test_pred1, test_svm$college_rate))
row1 <- cbind("SVM:RB",
              as.character(round(test_rsq1, 4)),
              as.character(round(test_rmse1, 4)))
test_pred2 <- predict(imp_feat_model, test_svm |> select(-college_rate))
test_rsq2 <- as.numeric(R2(test_pred2, test_svm$college_rate,
                           form = "traditional"))
test_rmse2 <- as.numeric(RMSE(test_pred2, test_svm$college_rate))
row2 <- cbind("Important Features (MLR)",
              as.character(round(test_rsq2, 4)),
              as.character(round(test_rmse2, 4)))
test_pred3 <- predict(lassoTune, test_lasso |> select(-college_rate))
```

```
test_rsq3 <- as.numeric(R2(test_pred3, test_lasso$college_rate,
                           form = "traditional"))
test_rmse3 <- as.numeric(RMSE(test_pred3, test_lasso$college_rate))
row3 <- cbind("Lasso",
              as.character(round(test_rsq3, 4)),
              as.character(round(test_rmse3, 4)))
tbl <- as.data.frame(rbind(row1, row2, row3))
cols <- c("Model", "Predictive R-Squared", "RMSE")
colnames(tbl) <- cols
tbl <- tbl |>
    mutate(Group = "Group 2") |>
    select(Group, everything())
apa_table(tbl, placement = "H")
```