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

Abstract coming soon!

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. Beginning in 2005, the survey looks to collect demographic and achievement data for New York City Public Schools, and provide a standardized rating of various elements of school quality.

The survey has changed over the years. This change has come from recommendations of public policy analysts in order to more accurately define the quality of schools *New York City Schools (2018).* The 2020-21 academic year report provides a robust dataset defined at the school level with academic and socioeconomic data provided.

Research Question: This study aims to determine whether the school ratings within the NYC School Quality Survey accurately reflect educational outcomes, or if other variables related to certain schools can be used as a better proxy.

In our case, we are interested in predicting the 4-year college persistence rate for an NYC high school. This measure is defined as the percent of students who graduate from a high school, and eventually go on to graduate from a 4-year colllege. Being able to identify the main indicators of a school's ability to successfully prepare students for college can benefit the NYC Department of Education, and New York City Schools along a couple of dimensions:

1. More directed instruction to enable useful skill transfer in preparatory courses
2. Better use of resourcing available to public schools to increase the percentage of college-ready students

For point 2 above, it's well correlated that students who attend 4-year institutions increase their career potential earnings significantly.

Literature Review

One of the main predictors of academic performance is the socioeconomic background of a student. 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)*.

Attempts to use more sophisticated modeling techniques and different sources datasets come from several prior studies. *Bernacki, Chavez, and Uesbeck (2020)* based their modeling off trying to predict based on student digital behavior, rather than social factors. The model in this study reached an accuracy of 75%, and was able to flag early interventions. While this modeling technique attempts to predict the same variable (educational achievement, albeit a different metric where we are predicting college attainment), the base dataset used to train the model and input variables are different.

Similarly, *Musso, Cascallar, Bostani, and Crawford (2020)* attempted to train an artificial neural network (ANN) to identify variable relationships to educational performance data. They modeled educational performance of Vietnamese students in grade 5. They 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 accuracy in prediction of $95 - 100$ higher than other modeling techniques. However, the training data came in that case from a different country (Vietnam, rather than the United States). Comparing modeling results from this (and other US-centric studies) may not be prudent.

Yağcı (2022) predicted final grade exams for Turkish students as well via machine learning models. Their input variables were prior exam grades. These can be a good "vacuum" comparison to compare one set of academci performance to another. However, there is a concern that good exam grades (even in one subject) do not correspond to a higher rate of career success later in life *Afarian and Kleiner (2003)*. Additionally, a parent study also found a correlation of up to 0.3 between academic grades and later job performance

Roth, BeVier, Switzer III, and Schippmann (1996).

Measuring the input variables that impact educational outcomes is a difficult task. With so many confounding variables, it can be difficult to determine direct causal relationships that have an outsized impact

Data Sourcing

The dataset used in this study is published from the NYC School Quality Report for the Academic Year 2020 - 2021. It consists of data from 487 New York City public schools, and 391 variables (in the form of columns). This dataset is defined at the school level, indexed by a school's *district borough number* (DBN).

In addition to the school quality ratings provided from survey responses in the data, there is average and raw academic performance data included. In addition to thesea academic indicators, there are socioeconomic variables included as well, such as the percentage of students at a given school in temporary housing services.

Methodology

Our primary interest is finding proxy variables within our dataset that can better serve as predictors of 4-year college persistence rates as a given NYC high school than the survey ratings collected by the school quality review. As a result, we'll need to first construct a "base" model that predicts a school's college persistence rate.

We can use two variables as a proxy for the school's survey rating in predicting college persistence:

- Percent of Students in Tempoarary Housing (`temp_housing_pct`)
- Economic Need Index (`eni_hs_pct_912`) - this is a measure of the percent of students facing economic hardship at a school ***(noauthor_student_2021?)***. This measures the economic hardship faced by students measured along a few criteria:

- The student is eligible for public assistance from the NYC Human Resources Administration (HRA)
- The student lived in temporary housing in the past four years
- The student 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.

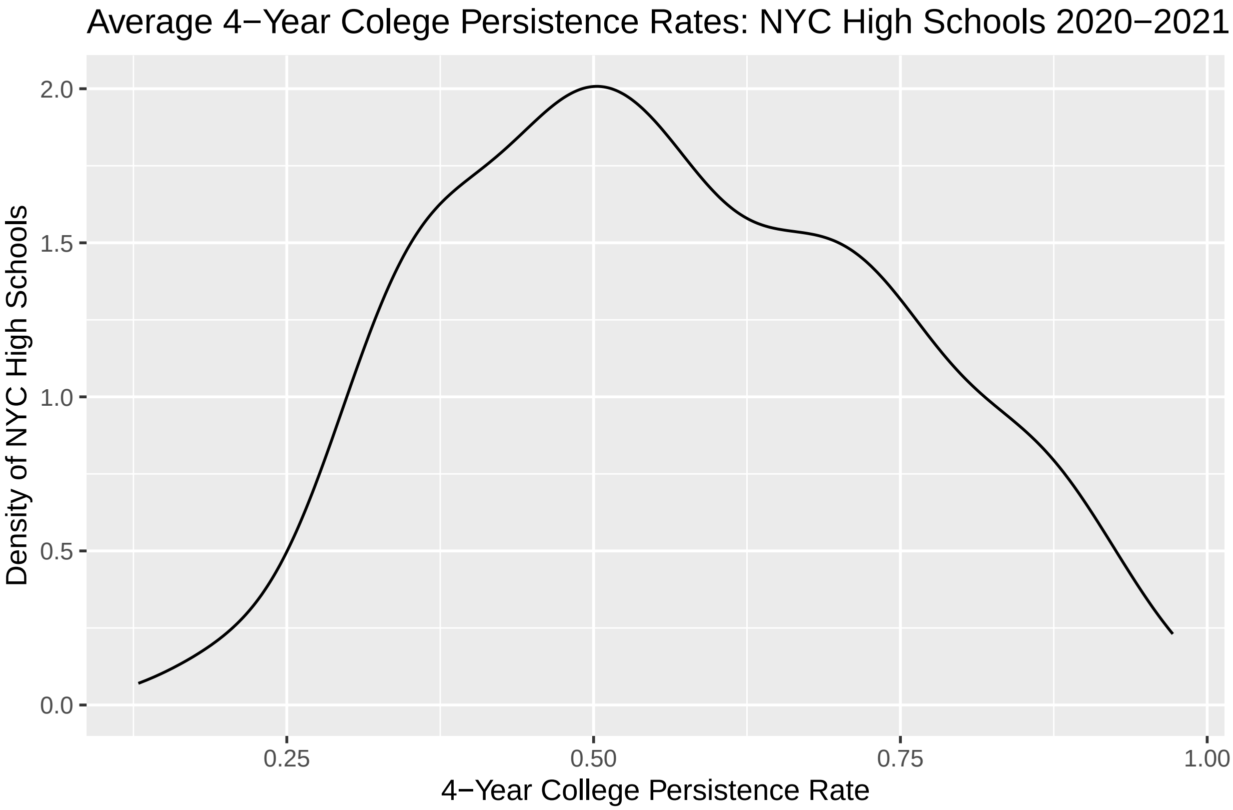
We create a 20% holdout set of data to be used later on in order to evaluate the efficacy of our model's predictive capability. The remaining 80% of the data is to be used for model training and exploratory data analysis (EDA).

For ease of single-node computation, we'll select the variables of interest from our dataset. Notably, these are the survey ratings for each school, as well as our preferred froxy variables. Additionally, 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

```
##
##  iter imp variable
##   1   1  survey_pp_CT  survey_pp_RI  survey_pp_ES  survey_pp_SE  survey_pp_SF  survey
##   1   2  survey_pp_CT  survey_pp_RI  survey_pp_ES  survey_pp_SE  survey_pp_SF  survey
##   1   3  survey_pp_CT  survey_pp_RI  survey_pp_ES  survey_pp_SE  survey_pp_SF  survey
##   1   4  survey_pp_CT  survey_pp_RI  survey_pp_ES  survey_pp_SE  survey_pp_SF  survey
##   1   5  survey_pp_CT  survey_pp_RI  survey_pp_ES  survey_pp_SE  survey_pp_SF  survey
##   2   1  survey_pp_CT  survey_pp_RI  survey_pp_ES  survey_pp_SE  survey_pp_SF  survey
##   2   2  survey_pp_CT  survey_pp_RI  survey_pp_ES  survey_pp_SE  survey_pp_SF  survey
##   2   3  survey_pp_CT  survey_pp_RI  survey_pp_ES  survey_pp_SE  survey_pp_SF  survey
##   2   4  survey_pp_CT  survey_pp_RI  survey_pp_ES  survey_pp_SE  survey_pp_SF  survey
##   2   5  survey_pp_CT  survey_pp_RI  survey_pp_ES  survey_pp_SE  survey_pp_SF  survey
##   3   1  survey_pp_CT  survey_pp_RI  survey_pp_ES  survey_pp_SE  survey_pp_SF  survey
##   3   2  survey_pp_CT  survey_pp_RI  survey_pp_ES  survey_pp_SE  survey_pp_SF  survey
##   3   3  survey_pp_CT  survey_pp_RI  survey_pp_ES  survey_pp_SE  survey_pp_SF  survey
##   3   4  survey_pp_CT  survey_pp_RI  survey_pp_ES  survey_pp_SE  survey_pp_SF  survey
##   3   5  survey_pp_CT  survey_pp_RI  survey_pp_ES  survey_pp_SE  survey_pp_SF  survey
##   4   1  survey_pp_CT  survey_pp_RI  survey_pp_ES  survey_pp_SE  survey_pp_SF  survey
##   4   2  survey_pp_CT  survey_pp_RI  survey_pp_ES  survey_pp_SE  survey_pp_SF  survey
##   4   3  survey_pp_CT  survey_pp_RI  survey_pp_ES  survey_pp_SE  survey_pp_SF  survey
##   4   4  survey_pp_CT  survey_pp_RI  survey_pp_ES  survey_pp_SE  survey_pp_SF  survey
##   4   5  survey_pp_CT  survey_pp_RI  survey_pp_ES  survey_pp_SE  survey_pp_SF  survey
##   5   1  survey_pp_CT  survey_pp_RI  survey_pp_ES  survey_pp_SE  survey_pp_SF  survey
##   5   2  survey_pp_CT  survey_pp_RI  survey_pp_ES  survey_pp_SE  survey_pp_SF  survey
##   5   3  survey_pp_CT  survey_pp_RI  survey_pp_ES  survey_pp_SE  survey_pp_SF  survey
##   5   4  survey_pp_CT  survey_pp_RI  survey_pp_ES  survey_pp_SE  survey_pp_SF  survey
##   5   5  survey_pp_CT  survey_pp_RI  survey_pp_ES  survey_pp_SE  survey_pp_SF  survey

##
##  iter imp variable
##   1   1  survey_pp_CT  survey_pp_RI  survey_pp_SE  college_rate
##   1   2  survey_pp_CT  survey_pp_RI  survey_pp_SE  college_rate
##   1   3  survey_pp_CT  survey_pp_RI  survey_pp_SE  college_rate
##   1   4  survey_pp_CT  survey_pp_RI  survey_pp_SE  college_rate
##   1   5  survey_pp_CT  survey_pp_RI  survey_pp_SE  college_rate
##   2   1  survey_pp_CT  survey_pp_RI  survey_pp_SE  college_rate
##   2   2  survey_pp_CT  survey_pp_RI  survey_pp_SE  college_rate
##   2   3  survey_pp_CT  survey_pp_RI  survey_pp_SE  college_rate
##   2   4  survey_pp_CT  survey_pp_RI  survey_pp_SE  college_rate
##   2   5  survey_pp_CT  survey_pp_RI  survey_pp_SE  college_rate
##   3   1  survey_pp_CT  survey_pp_RI  survey_pp_SE  college_rate
##   3   2  survey_pp_CT  survey_pp_RI  survey_pp_SE  college_rate
##   3   3  survey_pp_CT  survey_pp_RI  survey_pp_SE  college_rate
##   3   4  survey_pp_CT  survey_pp_RI  survey_pp_SE  college_rate
##   3   5  survey_pp_CT  survey_pp_RI  survey_pp_SE  college_rate
##   4   1  survey_pp_CT  survey_pp_RI  survey_pp_SE  college_rate
##   4   2  survey_pp_CT  survey_pp_RI  survey_pp_SE  college_rate
##   4   3  survey_pp_CT  survey_pp_RI  survey_pp_SE  college_rate
##   4   4  survey_pp_CT  survey_pp_RI  survey_pp_SE  college_rate
##   4   5  survey_pp_CT  survey_pp_RI  survey_pp_SE  college_rate
##   5   1  survey_pp_CT  survey_pp_RI  survey_pp_SE  college_rate
##   5   2  survey_pp_CT  survey_pp_RI  survey_pp_SE  college_rate
##   5   3  survey_pp_CT  survey_pp_RI  survey_pp_SE  college_rate
##   5   4  survey_pp_CT  survey_pp_RI  survey_pp_SE  college_rate
##   5   5  survey_pp_CT  survey_pp_RI  survey_pp_SE  college_rate
```

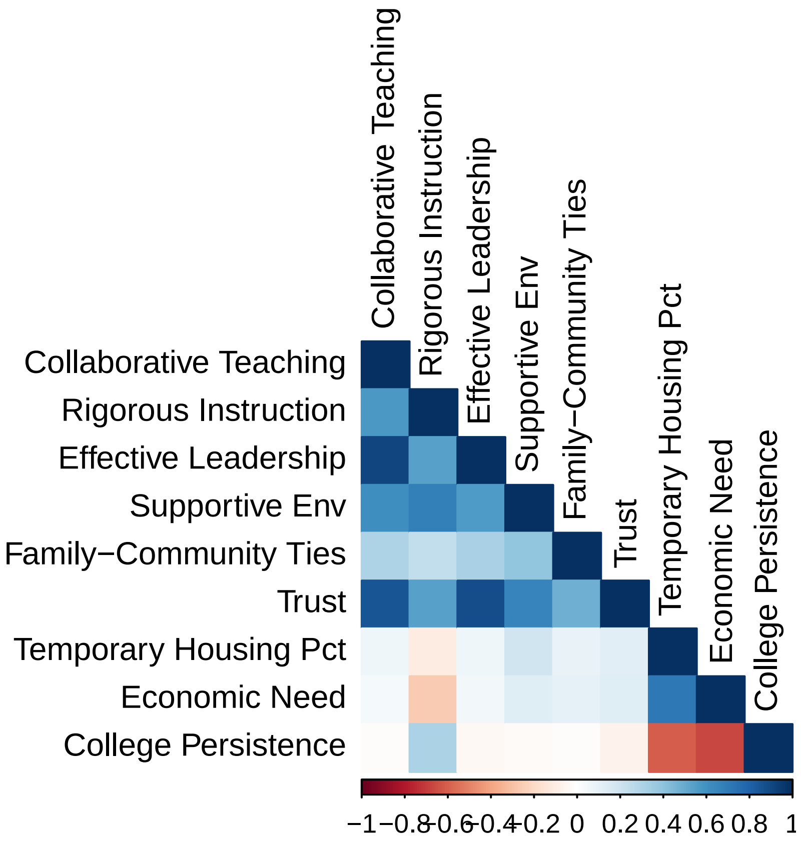
To check underlying modeling assumptions, we plot distributions and relationships of different variables. First, we plot the distribution of college persistence rates among NYC high schools to check for normality.


Average 4−Year Colege Persistence Rates: NYC High Schools 2020−2021

The average NYC high school sees ~50% of students go on to have 4−year college persistence.

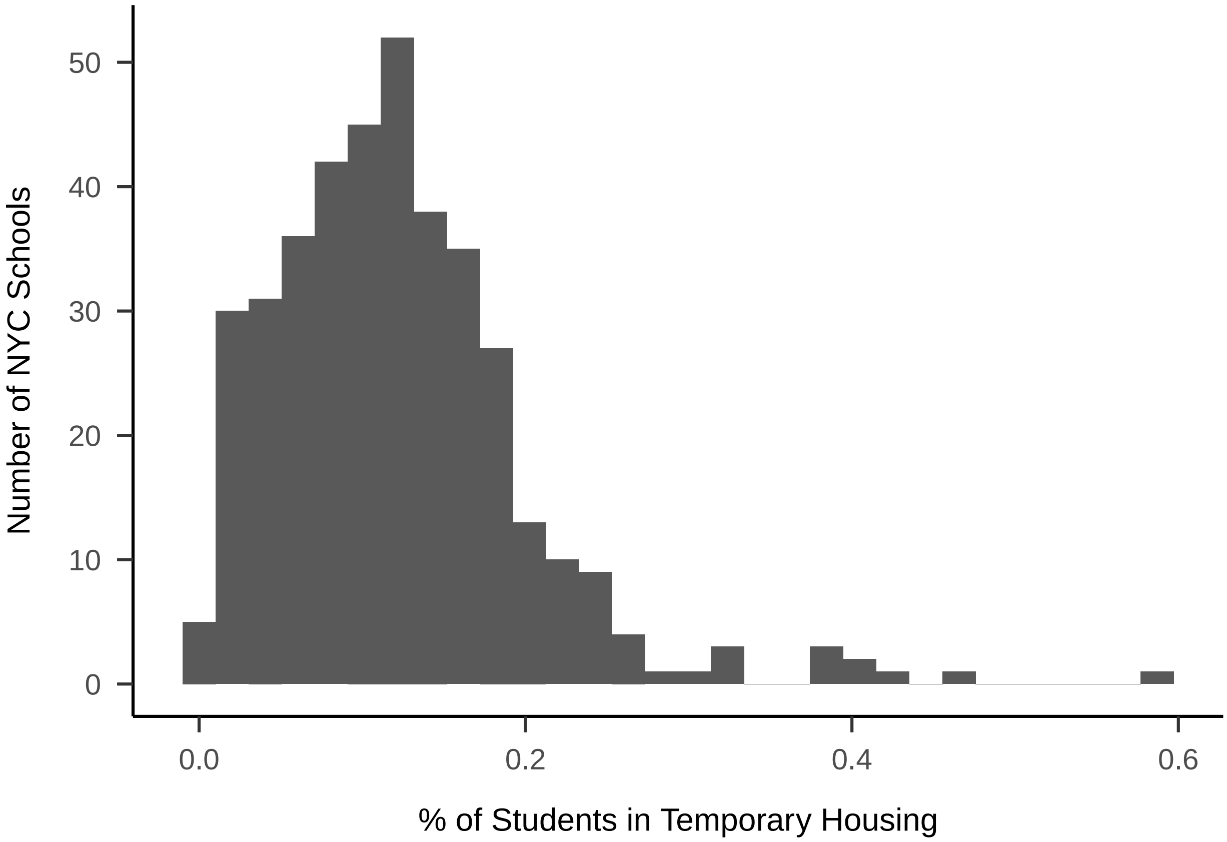
We see a relatively normal distribution of college persistence rates. In the case of NYC high schools, the peak is at around 50%. This is inline with national averages released by *US Census Bureau (2023)*

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*.

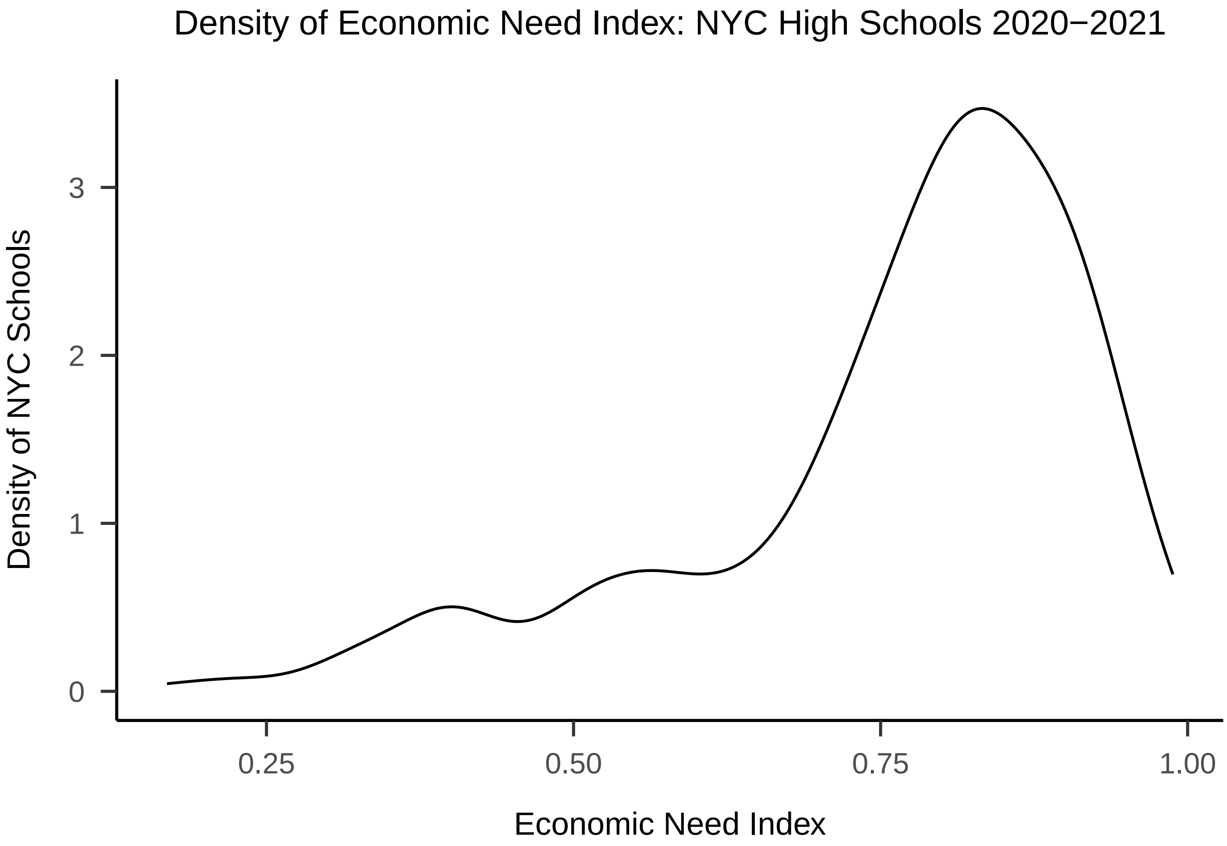


From our correlation plot above, we can see strong negative relationships between our proxy variables of interest (*Temporary Housing Rate* and *Economic Need Index*) and our target variable: *College Persistence Rate*. This gives signal that constructing models based on these variables could give good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

Now we can plot the distributions of our proxy variables of interest. First we can plot the temp housing rate:

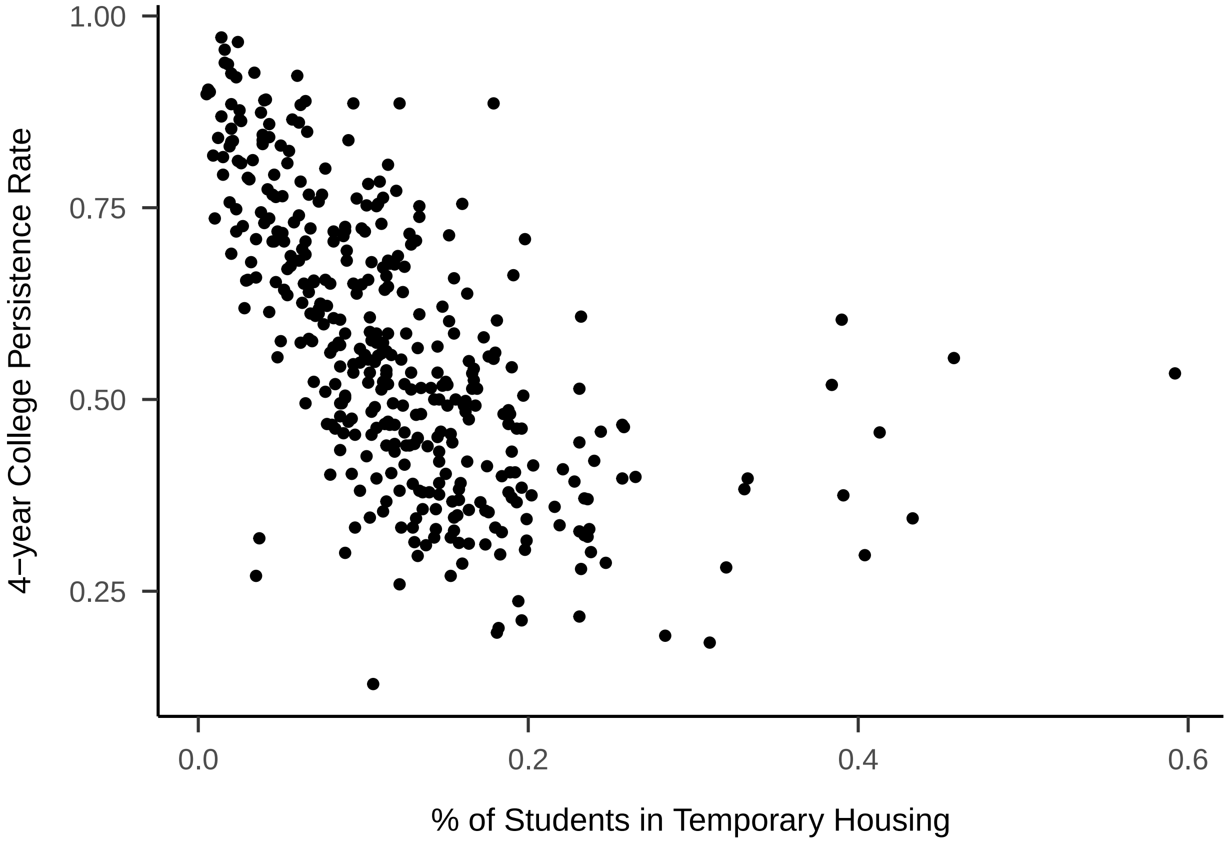


We see this distribution of the percentage of students in temporary housing per school to be skewed left. This will be an important piece of information as we model these relationships later. We also show the distribution of schools' economic need indices (also between 0 and 1). This index is closer to 1 the more economic hardship a student at a school faces (temporary housing use or food assistance, for instance).



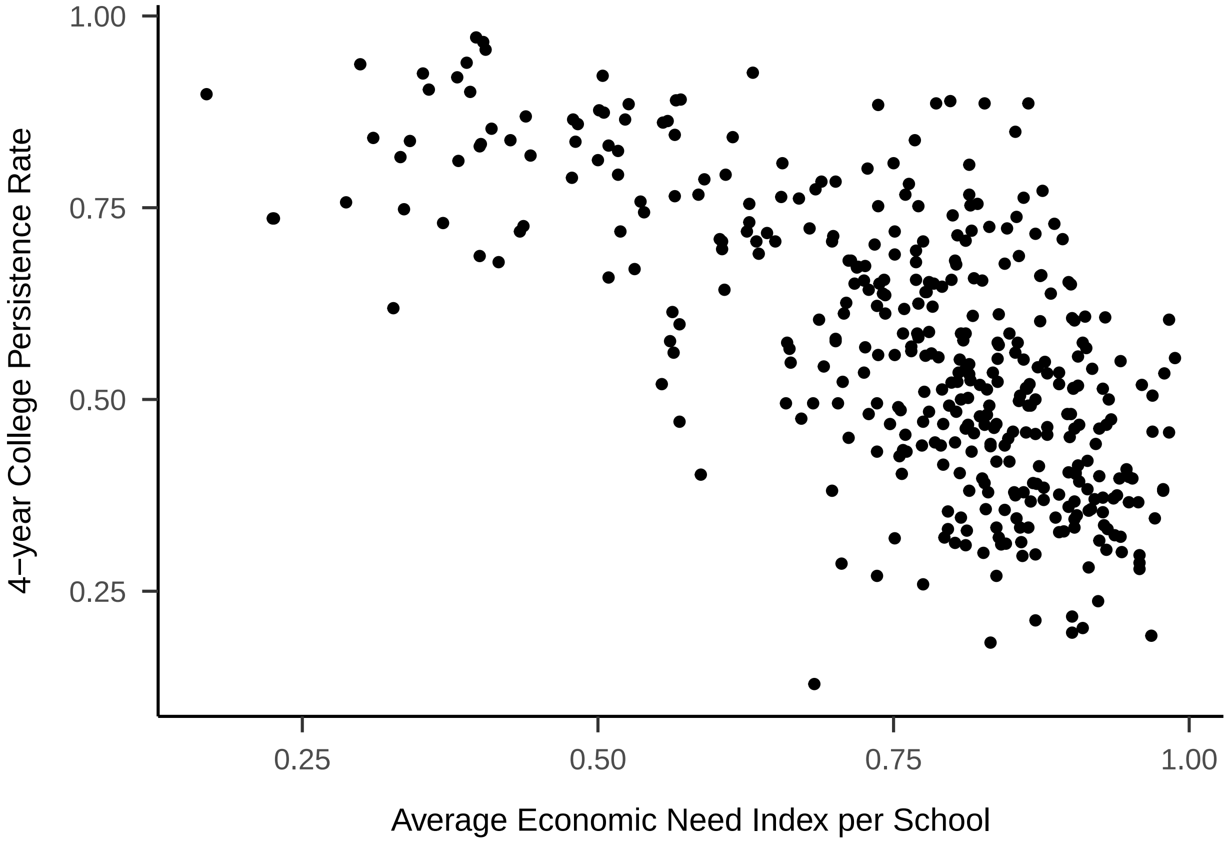
We also see a skewed distribution for our economic need index. This is a candidate for transformation before feeding into our proxy variable model.

First, we should check an assumption of linearity between our predictor and response variables. In this case this a scatter plot of the percentage of students in temporary housing



We see a general linear relationship for schools with lower rates of students in temp housing. However, this linear relationship does **not** visually hold for schools with higher rates of temp housing use.

Plotting the relationship below between a school's economic need index



Again, we see a non-linear relationship between our predictor (*Economic Need Index*) and Outcome Variable (*College Persistence Rate*)

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

```
##
## Call:
## lm(formula = base_formula, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.54578 -0.11206  0.00304  0.11205  0.41671
##
## Coefficients:
```

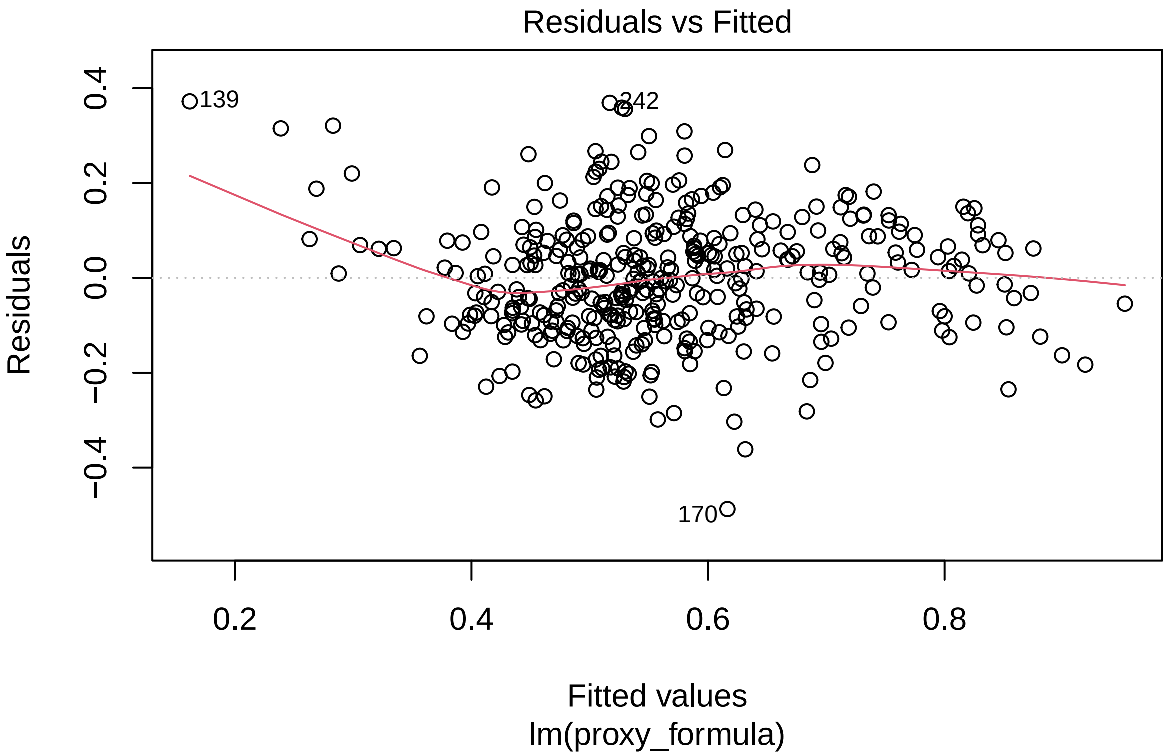
```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.49700    0.20023   2.482   0.0135 *
## survey_pp_CT  0.04036    0.27214   0.148   0.8822
## survey_pp_RI  2.10716    0.19902  10.587  < 2e-16 ***
## survey_pp_SE -1.41047    0.26337  -5.356 1.47e-07 ***
## survey_pp_ES -0.22948    0.28878  -0.795   0.4273
## survey_pp_SF  0.26272    0.21171   1.241   0.2154
## survey_pp_TR -0.48790    0.42585  -1.146   0.2526
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1592 on 383 degrees of freedom
## Multiple R-squared:  0.2348, Adjusted R-squared:  0.2228
## F-statistic: 19.58 on 6 and 383 DF,  p-value: < 2.2e-16
```

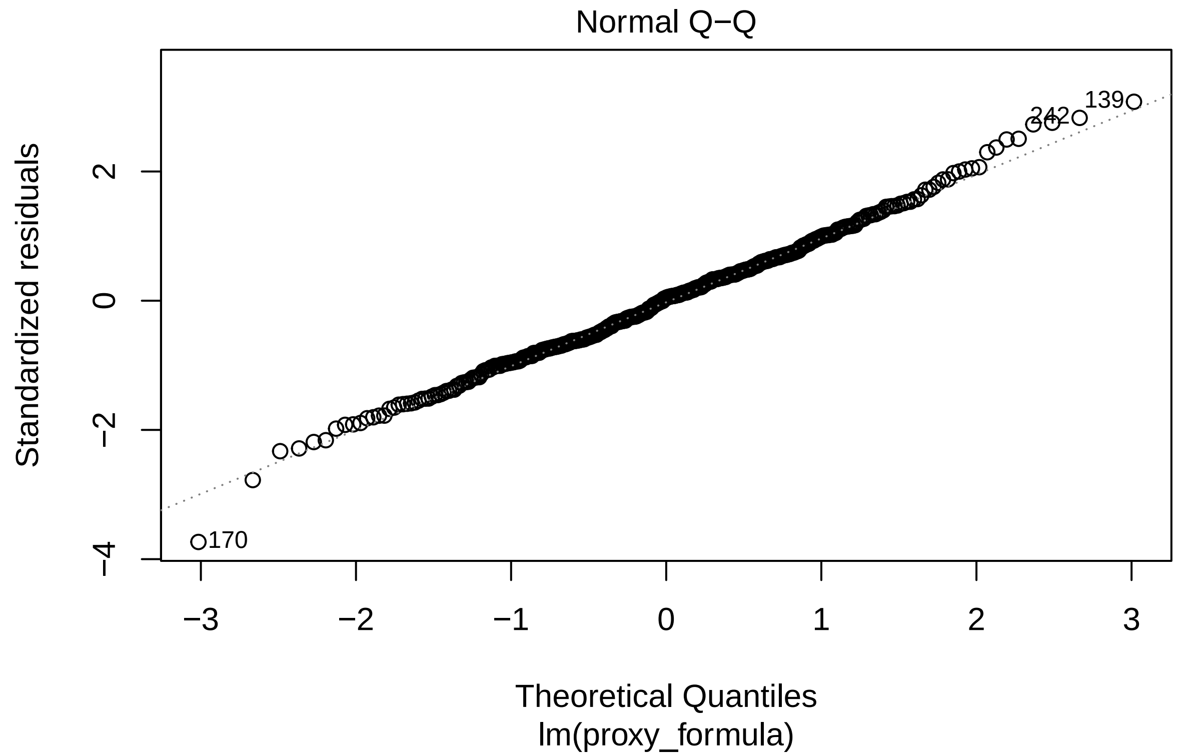
We find our base model for the school survey ratings produces an adjusted R-squared of $R^2_{adj} = 0.22$. This is lower than the predictive model produces in *Roth et al. (1996).*

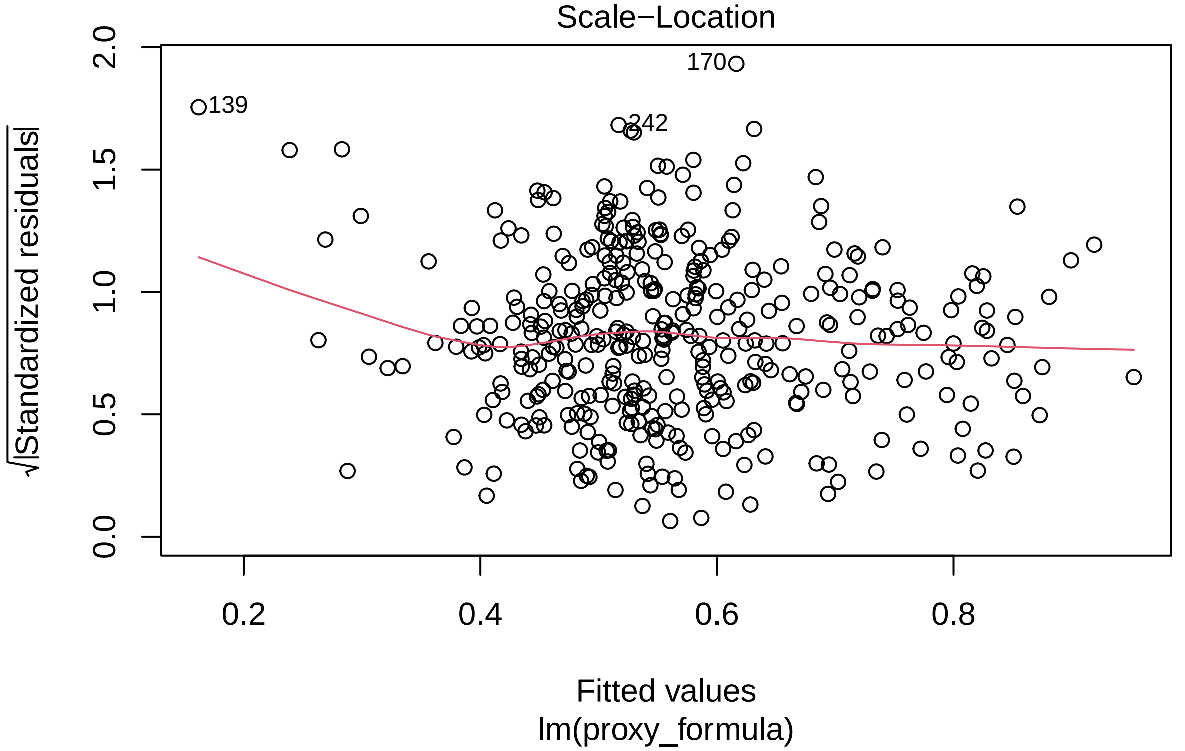
We then create a basic multiple least squares linear model between our two socioeconomic proxy variables: *Temporary Housing PErcentage of a School* and *Average Economic Need Index.*

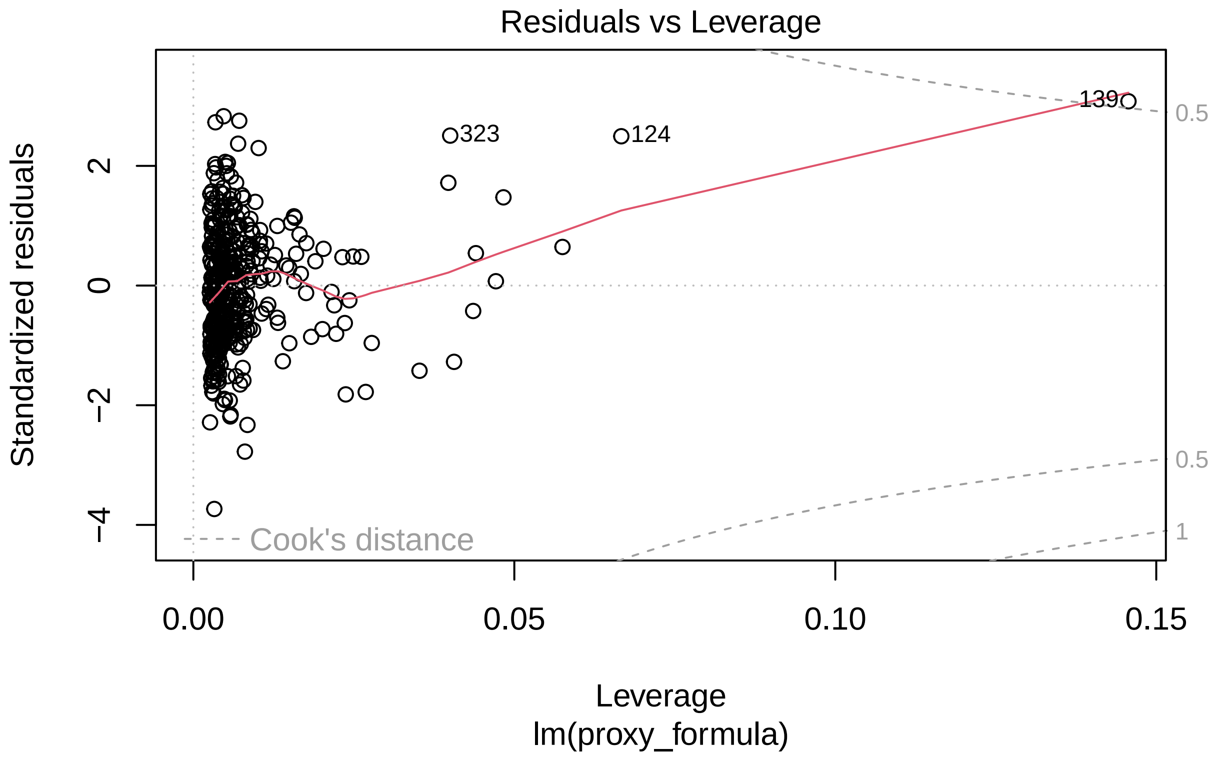
```
##
## Call:
## lm(formula = proxy_formula, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.48742 -0.09012  0.00566  0.08315  0.37216
```

```
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        1.04561    0.03621  28.873  < 2e-16 ***
## temp_housing_pct -0.61032    0.12311  -4.958 1.07e-06 ***
## economic_need    -0.53367    0.05893  -9.056  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1307 on 387 degrees of freedom
## Multiple R-squared:  0.4786, Adjusted R-squared:  0.4759
## F-statistic: 177.6 on 2 and 387 DF,  p-value: < 2.2e-16
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(proxy_formula)
139
242
170
−4
−2
0
2
−3
−2
−1
0
1
2
3


Scale−Location
$\sqrt{|\text{Standardized residuals}|}$
Fitted values
lm(proxy_formula)
170
139
242
0.0
0.5
1.0
1.5
2.0
0.2
0.4
0.6
0.8



Given the

```
##
## Call:
## lm(formula = proxy_formula, data = train, weights = weights)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8222 -0.8301  0.0452  0.7834  3.5360
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.02985    0.03260  31.595  < 2e-16 ***
## temp_housing_pct -0.74319    0.13149  -5.652 3.08e-08 ***
## economic_need    -0.49180    0.05594  -8.791  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.256 on 387 degrees of freedom
## Multiple R-squared:  0.5104, Adjusted R-squared:  0.5078
## F-statistic: 201.7 on 2 and 387 DF,  p-value: < 2.2e-16
```

Experimentation and Results

Model Evaluation.

```
## [1] 0.1653904
```

```
## [1] 0.1366313
```

```
## [1] 0.1372144
```

We can also use the Akaike and Bayesian Information Criterion for evaluatng the complexity of our models. We're using fewer variables in our proxy and WLS models, so we'd expect better results (minimized values) for each of those criteria

```
## AIC for base model (rating results): -317.495489747942
```

```
## AIC for proxy variable model: -475.126800916888
```

```
## AIC for WLS model: -480.729858621265
```

```
## BIC for base model (rating results): -285.766315834952
```

```
## BIC for proxy variable model: -459.262213960393
```

```
## BIC for WLS model: -464.86527166477
```

Conclusion

TODO

- Merge/Join in ACT/SAT information by DBN
- Model Selection

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

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")
r_refs("r-references.bib")
# Read in our dataset from GitHub
# https://www.opendatanetwork.com/dataset/data.cityofnewyork.us/bm9v-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
df$economic_need <- df$eni_hs_pct_912
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)
test  <- dplyr::anti_join(df, train, by = 'dbn')

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")
train_data <- train[, cols]
imp <- mice(train_data, method="pmm", seed=42)
train <- complete(imp)
test_data <- test[, cols]
imp <- mice(test_data, method="pmm", seed=42)
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="Av

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)
```

```
# Create correlation plot between vars of interest
corMatrix <- cor(train_renamed)
corrplot(corMatrix, method="color", type="lower", tl.col="black")
# Plot temp housing rates
ggplot(train, aes(x=temp_housing_pct)) +
  geom_histogram() +
  labs(x="% of Students in Temporary Housing", y="Number of NYC Schools")
# Plot economic need index
ggplot(train, aes(x=economic_need)) +
  geom_density() +
  labs(x="Economic Need Index", y="Density of NYC Schools",
       title="Density of Economic Need Index: NYC High Schools 2020-2021")
# Plot temp housing percentage vs college persistence rate
ggplot(train, aes(x=temp_housing_pct, y=college_rate)) +
  geom_point() +
  labs(x="% of Students in Temporary Housing",
       y="4-year College Persistence Rate")
# Plot ENI vs college persistence rate
ggplot(train, aes(x=economic_need, y=college_rate)) +
  geom_point() +
  labs(x="Average Economic Need Index per School",
       y="4-year College Persistence Rate")
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)
```

```r
# Create OLS linear model based on our proxy variables: no transforms
proxy_formula <- college_rate ~ temp_housing_pct + economic_need
proxy_model <- lm(proxy_formula, train)
summary(proxy_model)
plot(proxy_model)
# Calculating weights for WLS
weights <- 1 / lm(abs(proxy_model$residuals) ~ proxy_model$fitted.values)$fitted.values^

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

summary(wls_model)
# Compute RMSE for each model on our testing data
# TODO: Put in table with AIC and BIC results
rmse(rating_model, test)
modelr::rmse(proxy_model, test)
modelr::rmse(wls_model, test)
# Print AIC for each model type
print(glue("AIC for base model (rating results): {AIC(rating_model)}"))
print(glue("AIC for proxy variable model: {AIC(proxy_model)}"))
print(glue("AIC for WLS model: {AIC(wls_model)}"))

# BIC results
print(glue("BIC for base model (rating results): {BIC(rating_model)}"))
print(glue("BIC for proxy variable model: {BIC(proxy_model)}"))
print(glue("BIC for WLS model: {BIC(wls_model)}"))
```