# 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. 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 these outcomes could actually be better predicted by proxy variables related to the student body.

The primary measure of success we are interested in predicting is the 4-year college persistence rate for an NYC high school. This measure is defined as the percentage of students who graduate from a high school and eventually go on to graduate from a 4-year college. Being able to identify the main indicators of a school's ability to successfully prepare students for college would benefit the NYC Department of Education (DOE) and NYC Public Schools in a couple of ways:

**improve bullet points below**

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)*.

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. *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'll be looking at. 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 are different 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. 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 prediction accuracy of $95 - 100$, higher than other modeling techniques. Since the training data comes from a different country and therefore a different educational system, comparing modeling results from our analysis (or results from other US-centric studies) to theirs may not be prudent.

*Yağcı (2022)* predicted final grade exams for Turkish students via machine learning models. Their input variables were prior exam grades, which can be good "vacuum" comparisons between one set of academic performance data and another. However, there is a concern that good exam grades in one or more subjects do not correspond to higher rates 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).* **The previous two sentences go a little off-topic since we're linking school demographics to college persistence, which is definitely a precursor to a career, but obviously different.**

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'll 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:

**improve these bullets**

- 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 (fix, not in references)**. 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.
- Chronic Absenteeism (`val_chronic_absent_hs_all`)

We take a look at a summary of the dataset's completeness.

Table 1

*Completeness Summary*

| | |
|---|---|
| rows | 487 |
| columns | 393 |
| all_missing_columns | 12 |
| total_missing_values | 47359 |
| complete_rows | 0 |

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

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'll select the variables of interest from our dataset. Notably, these are the survey ratings, enrollment levels, and our preferred proxy variables for each school.

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

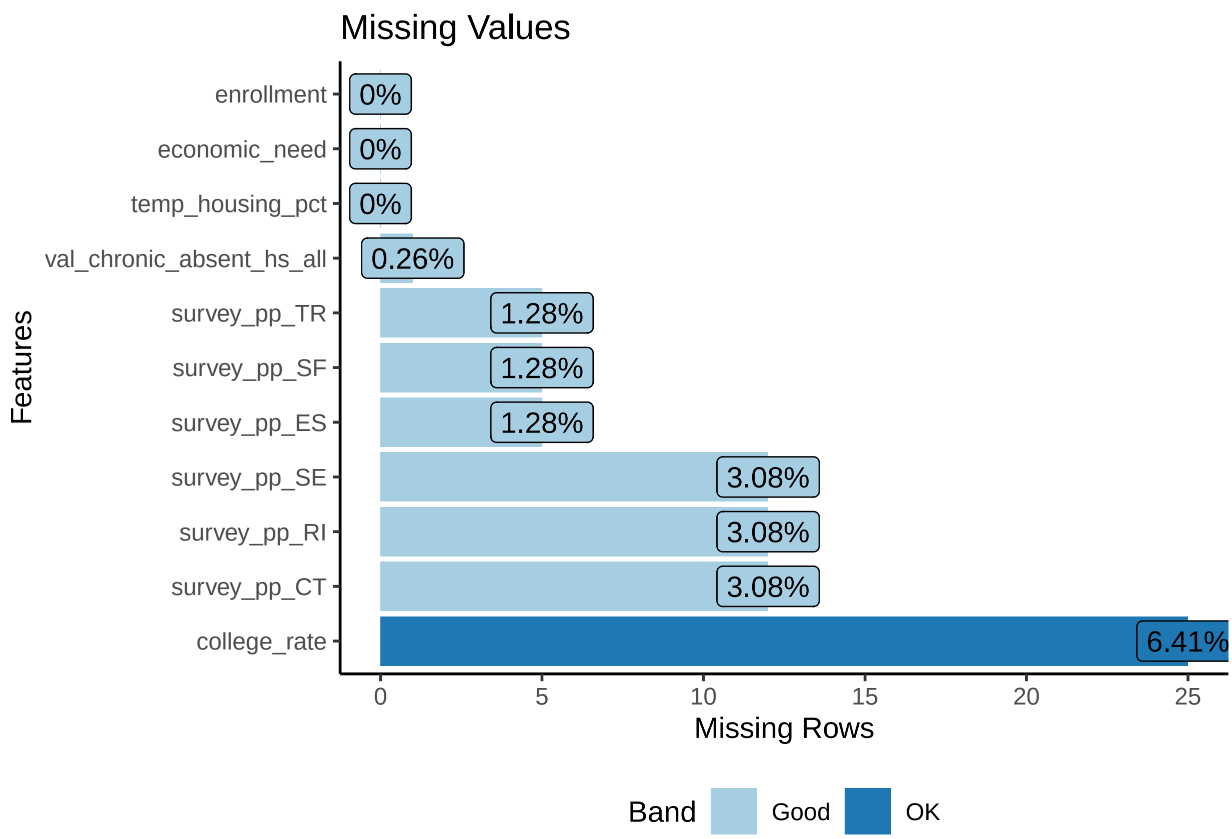


*Figure 1*

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.

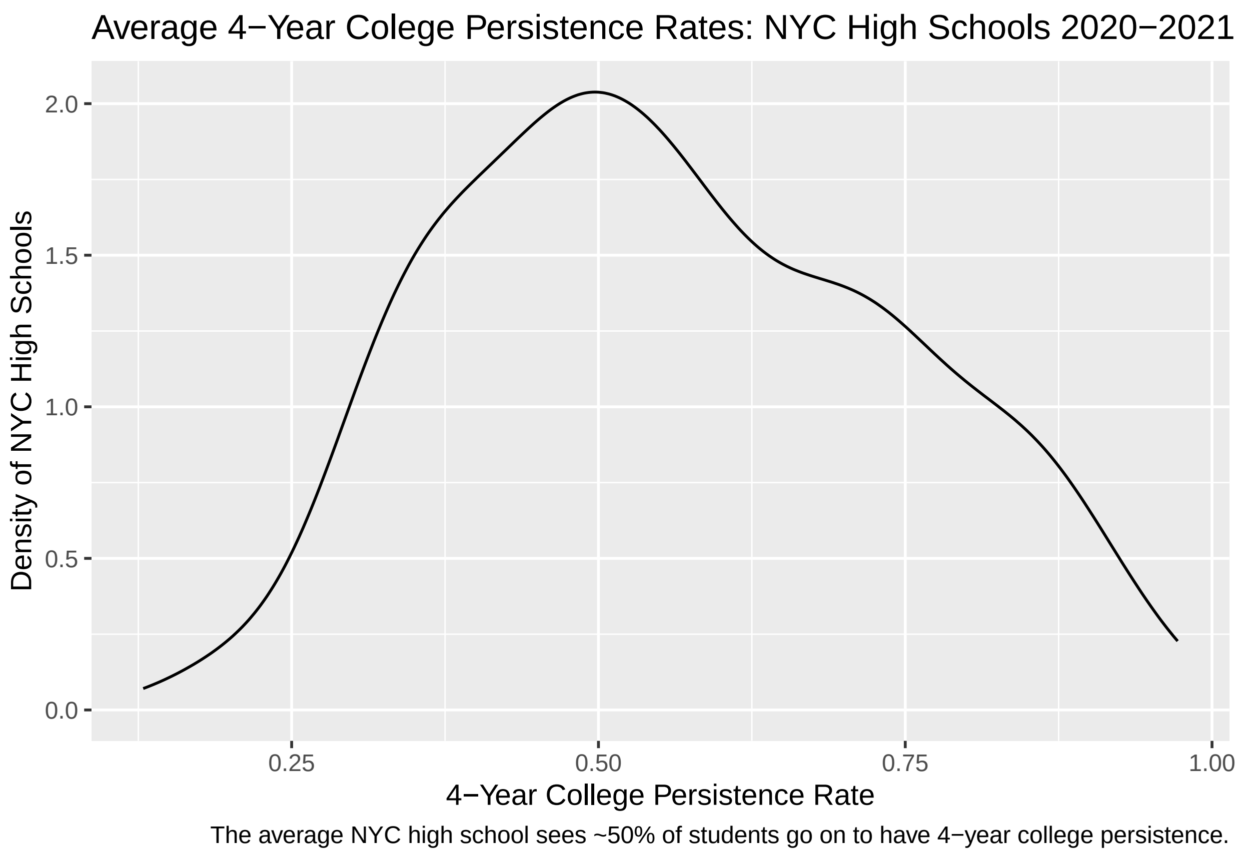


*Figure 2*

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 *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*.

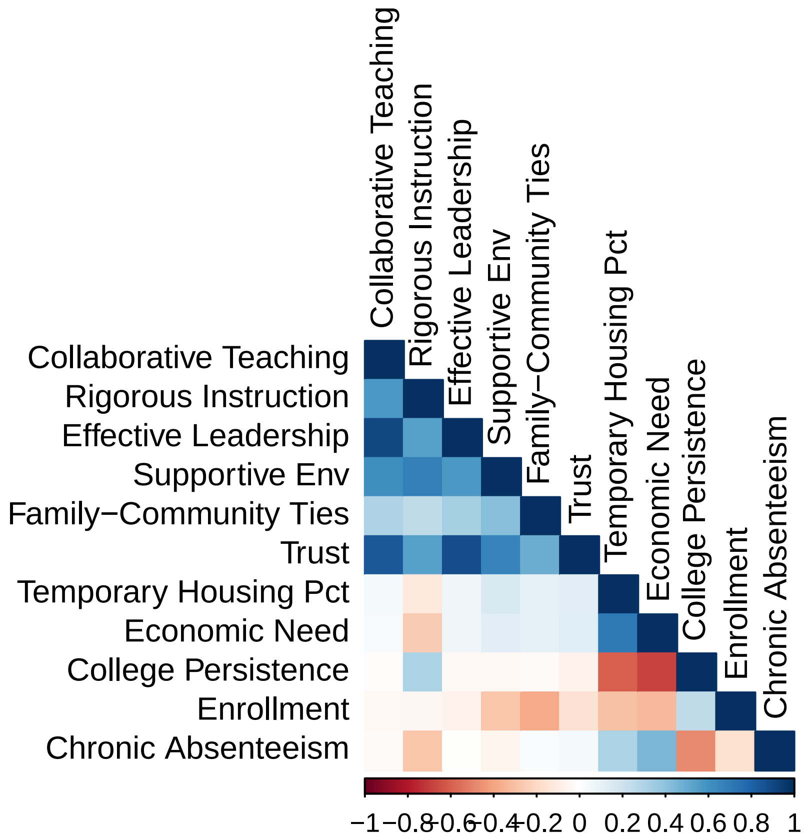


*Figure 3*

From our correlation plot above, we can see strong negative relationships between 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:

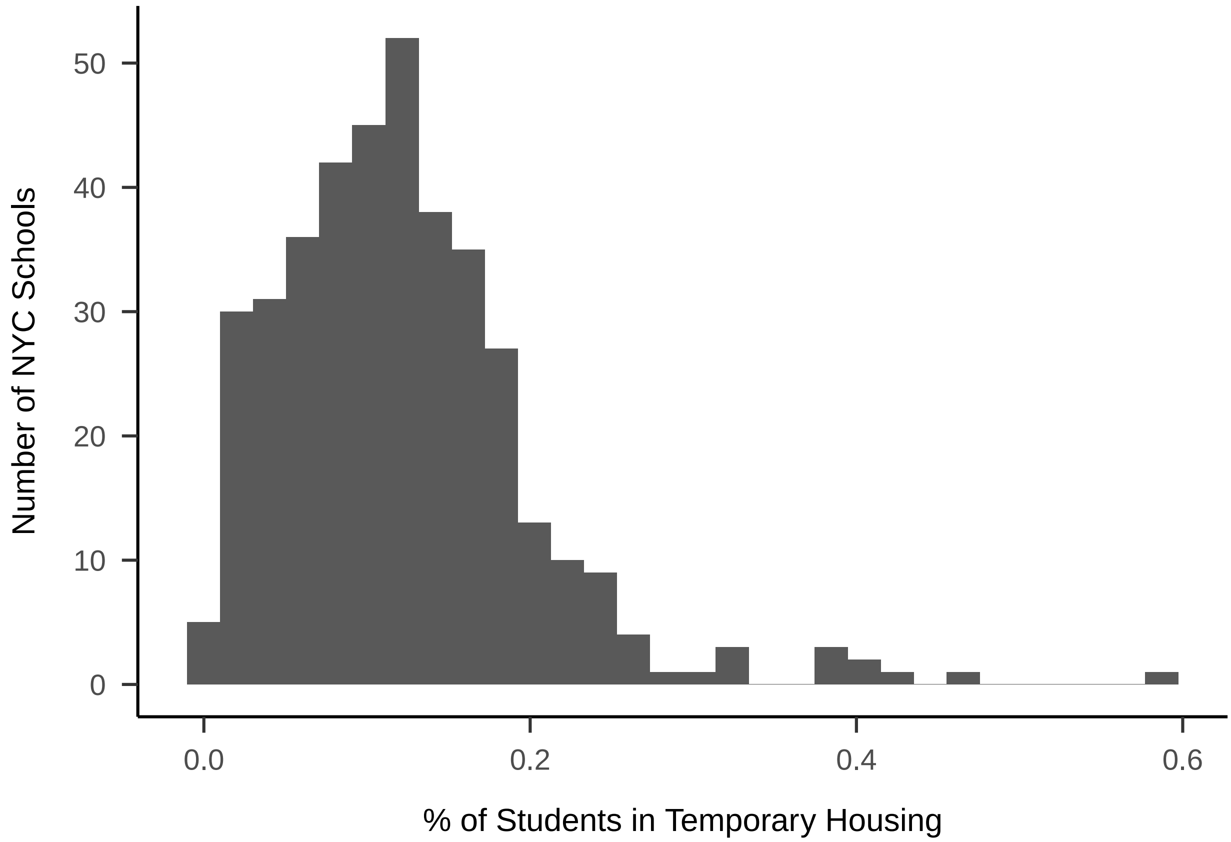


*Figure 4*

We see this distribution of the percentage of students in temporary housing per school to be skewed right. 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).

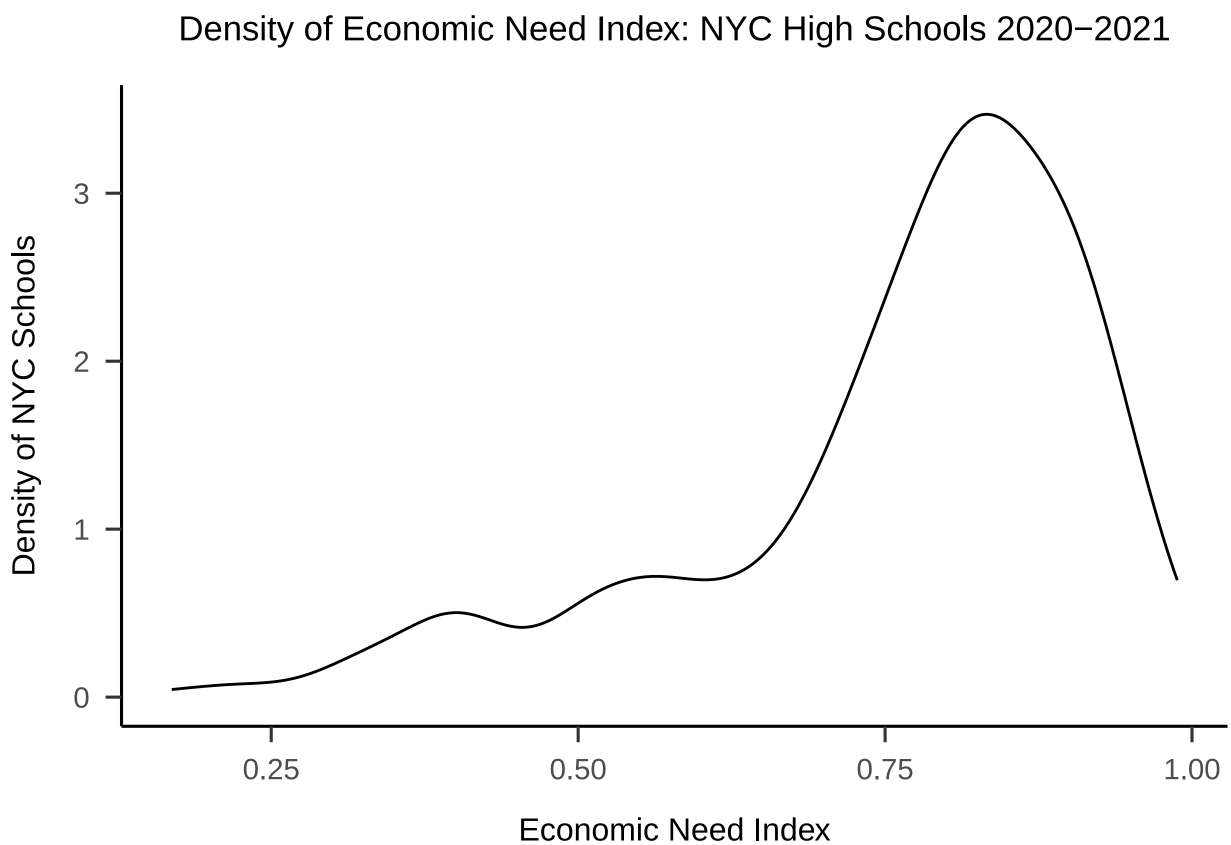


*Figure 5*

We see a left-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 predictors and our response variable. Here, we produce scatter plots of the response variable versus the percentage of students in temporary housing, the economic need index, the enrollment level, and chronic absenteeism.

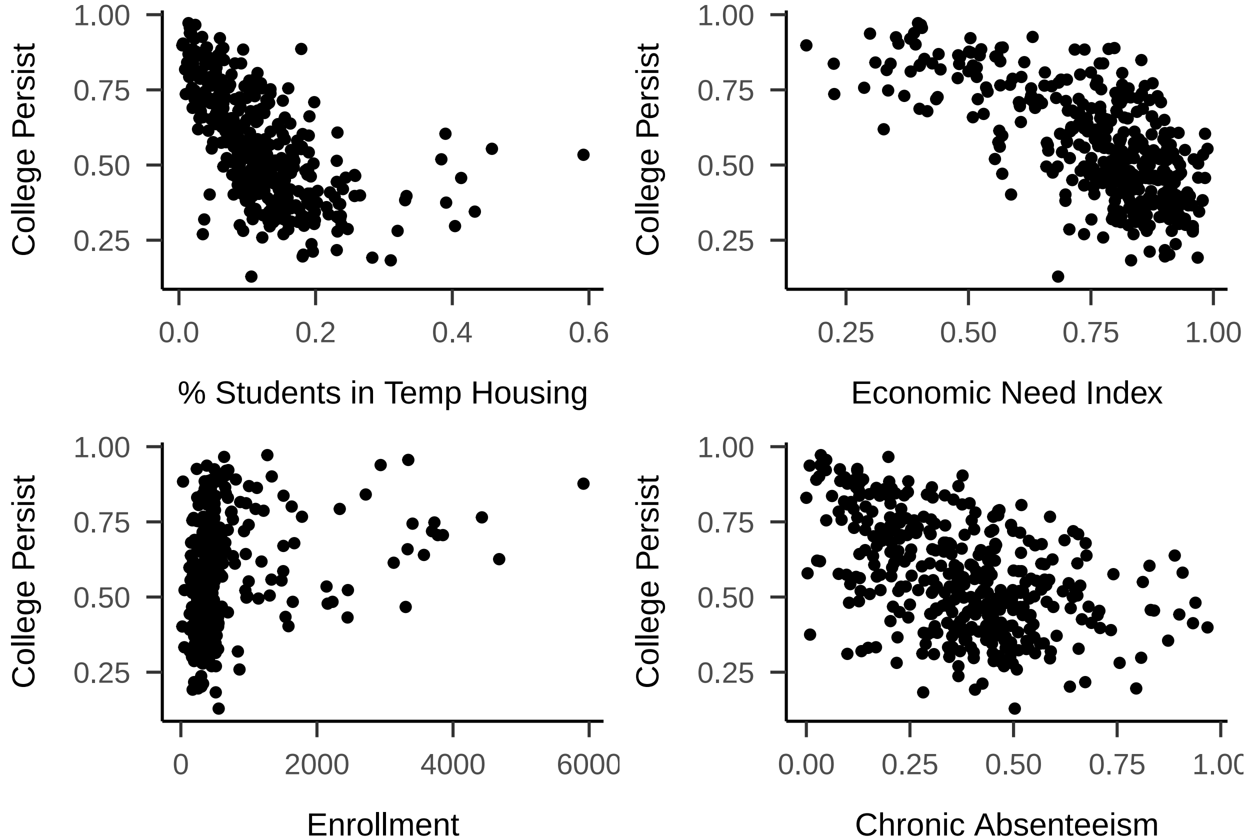


*Figure 6*

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 1 school where chronic absenteeism is greater than or equal to 0.5 achieves college persistence levels above 80 percent. However, college persistence varies widely at most chronic absenteeism levels. **Investigate why this variable can take such high rates and whether there's anything that can collectively be said about the 12 schools with values greater than 0.8 for this variable. Make sure we understand what it's measuring correctly.**

**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
## -0.5405 -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 $R^2_{adj} = 0.22$. This is lower than the predictive model in *Roth et al. (1996)* produces.

We then create a basic multiple least squares linear model between the response and 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
## -0.48637 -0.08608  0.00436  0.08041  0.36890
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.06204    0.03612  29.405  < 2e-16 ***
## temp_housing_pct -0.54224    0.12278  -4.416  1.3e-05 ***
## economic_need    -0.56982    0.05877  -9.696  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1304 on 387 degrees of freedom
## Multiple R-squared:  0.4843, Adjusted R-squared:  0.4816
```

```
## F-statistic: 181.7 on 2 and 387 DF,  p-value: < 2.2e-16
```

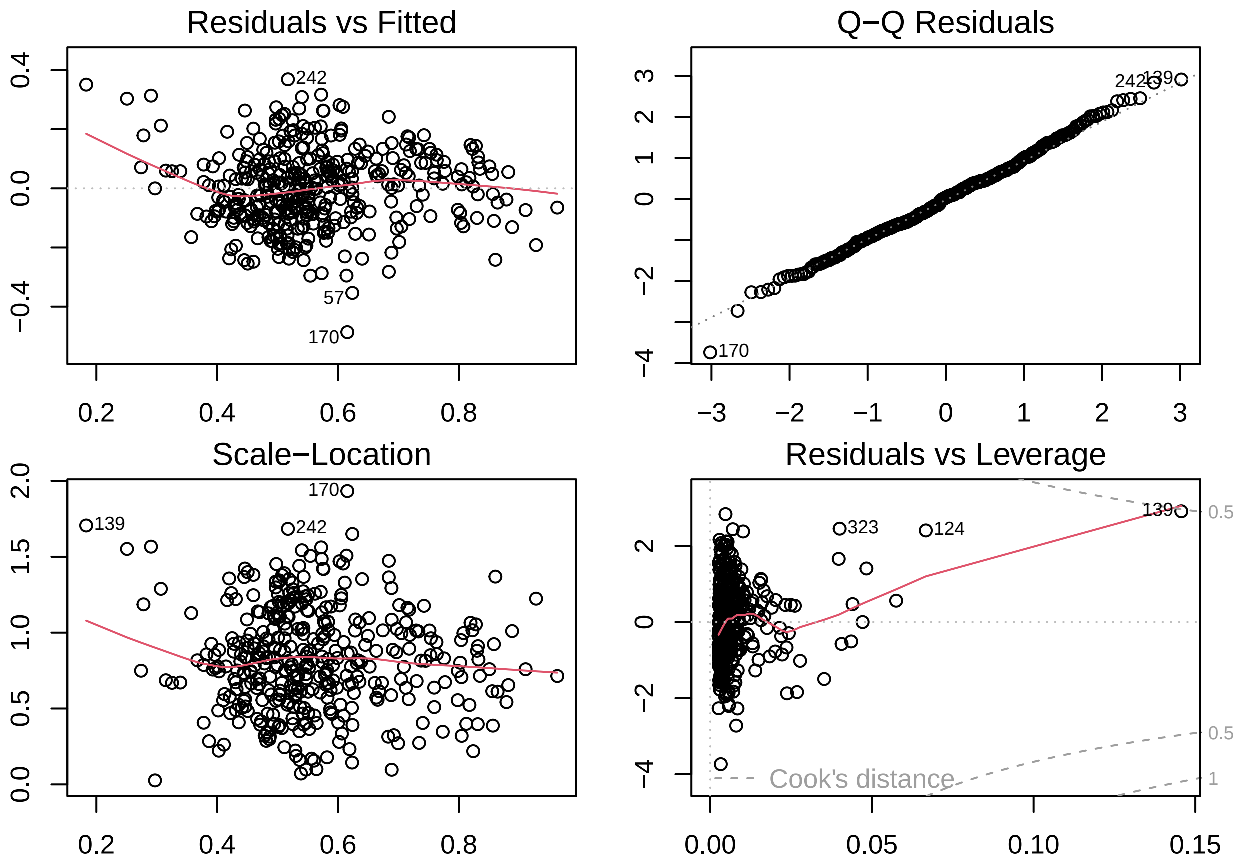


*Figure 7*

Given the

```
##
## Call:
## lm(formula = proxy_formula, data = train, weights = weights)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -4.7873 -0.8136  0.0381  0.7939  3.5364
##
```

```
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.05037    0.03324  31.603  < 2e-16 ***
## temp_housing_pct -0.63748    0.12934  -4.929 1.23e-06 ***
## economic_need    -0.53930    0.05640  -9.562  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.254 on 387 degrees of freedom
## Multiple R-squared:  0.5101, Adjusted R-squared:  0.5076
## F-statistic: 201.5 on 2 and 387 DF,  p-value: < 2.2e-16
```

## Experimentation and Results

### Model Evaluation.

```
## [1] 0.1662838
```

```
## [1] 0.1394362
```

```
## [1] 0.1395916
```

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): -322.903528891558
```

```
## AIC for proxy variable model: -477.210070495472
```

```
## AIC for WLS model: -480.874854356931
```

```
## BIC for base model (rating results): -291.174354978569
```

```
## BIC for proxy variable model: -461.345483538977
```

```
## BIC for WLS model: -465.010267400436
```

# 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")
library(DataExplorer)
library(cowplot)
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=2023110
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
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)
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", "enrollment",
          "val_chronic_absent_hs_all")
train_data <- train[, cols]
```

```
p1 <- plot_missing(train_data, missing_only = FALSE,
                   ggtheme = theme_classic(), title = "Missing Values")

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 Colege 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")
# 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
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)
# 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)
par(mfrow=c(2,2))
par(mai=c(.3,.3,.3,.3))
plot(proxy_model)
# Calculating weights for WLS
weights <- 1 / lm(abs(proxy_model$residuals) ~ proxy_model$fitted.values)$fitted.values
```

```r
#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)}"))
```