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 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	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.

[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'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.

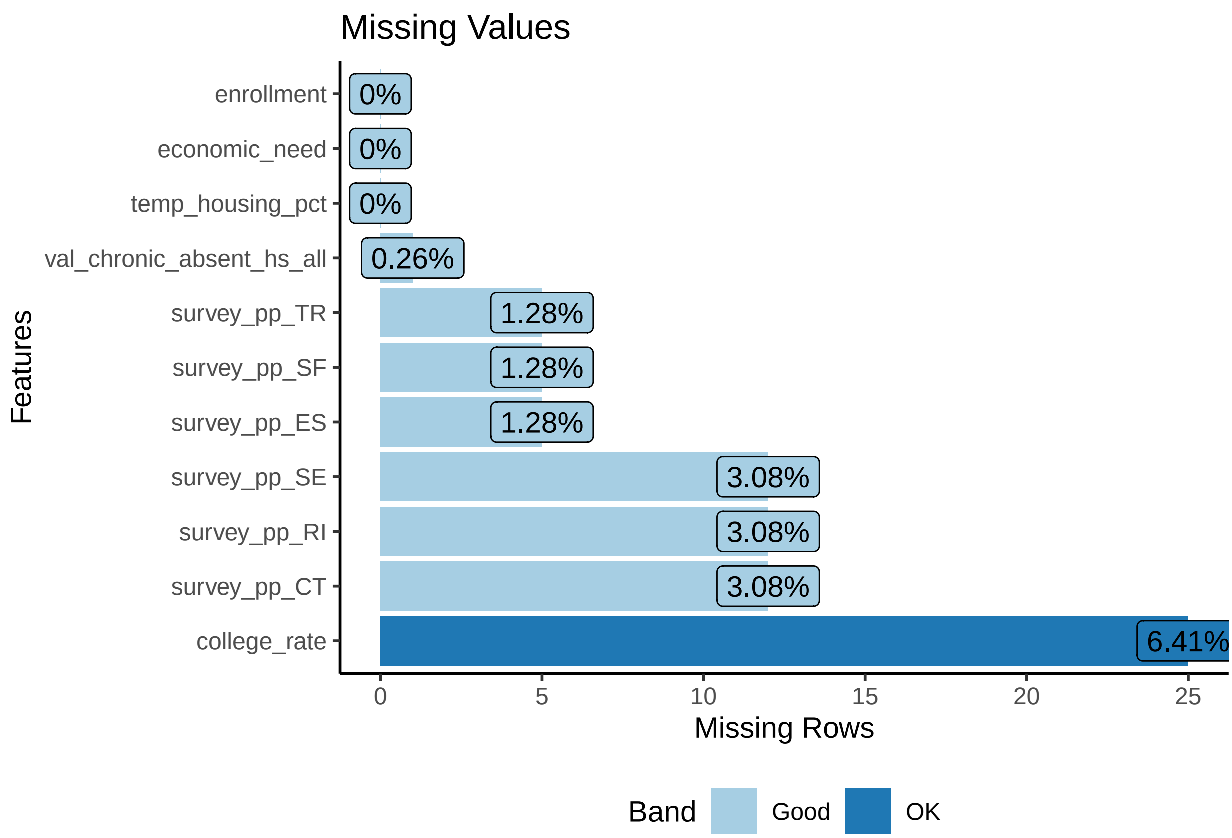


Figure 1

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.

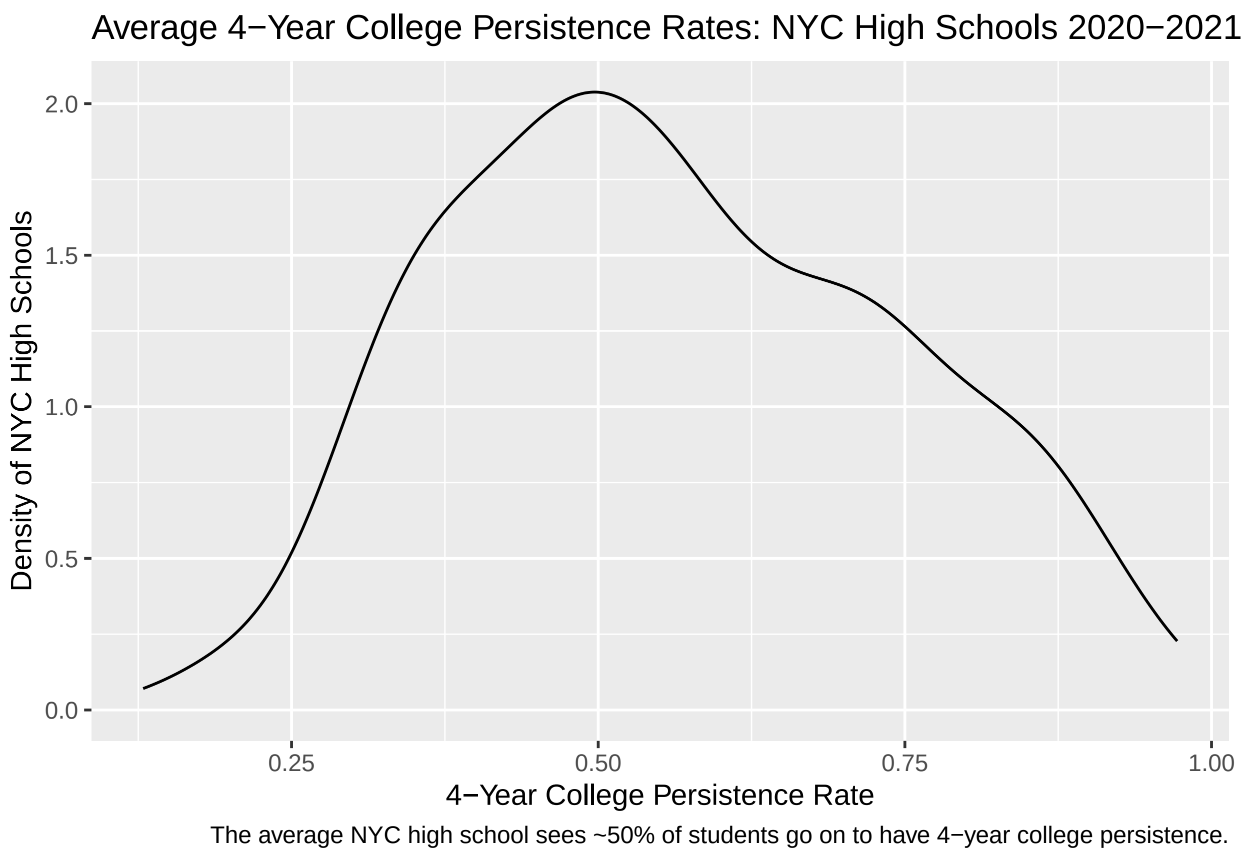


Figure 2

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 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*.

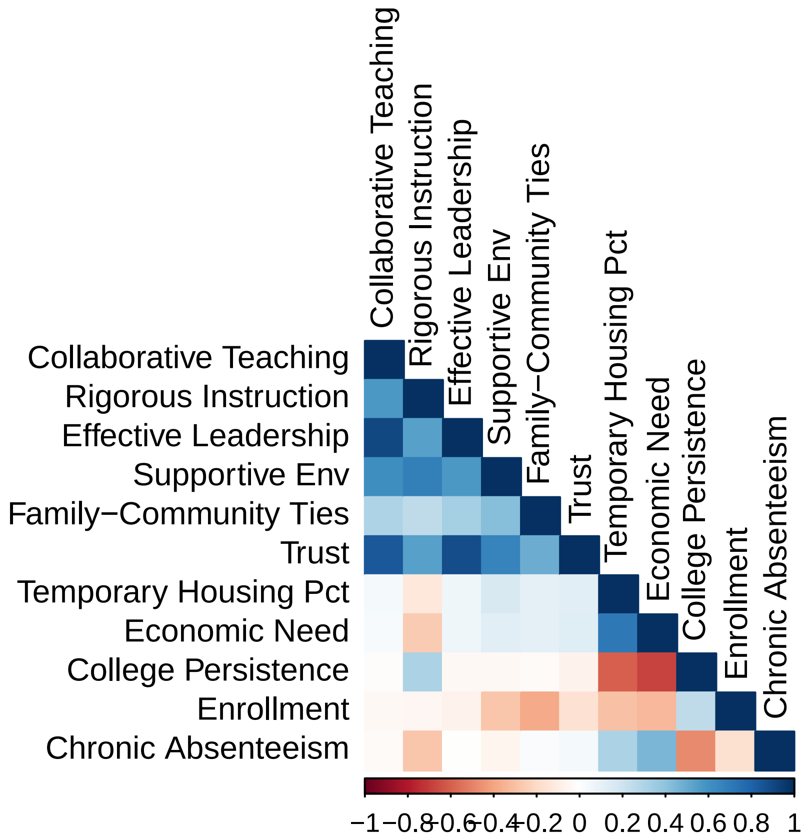


Figure 3

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.

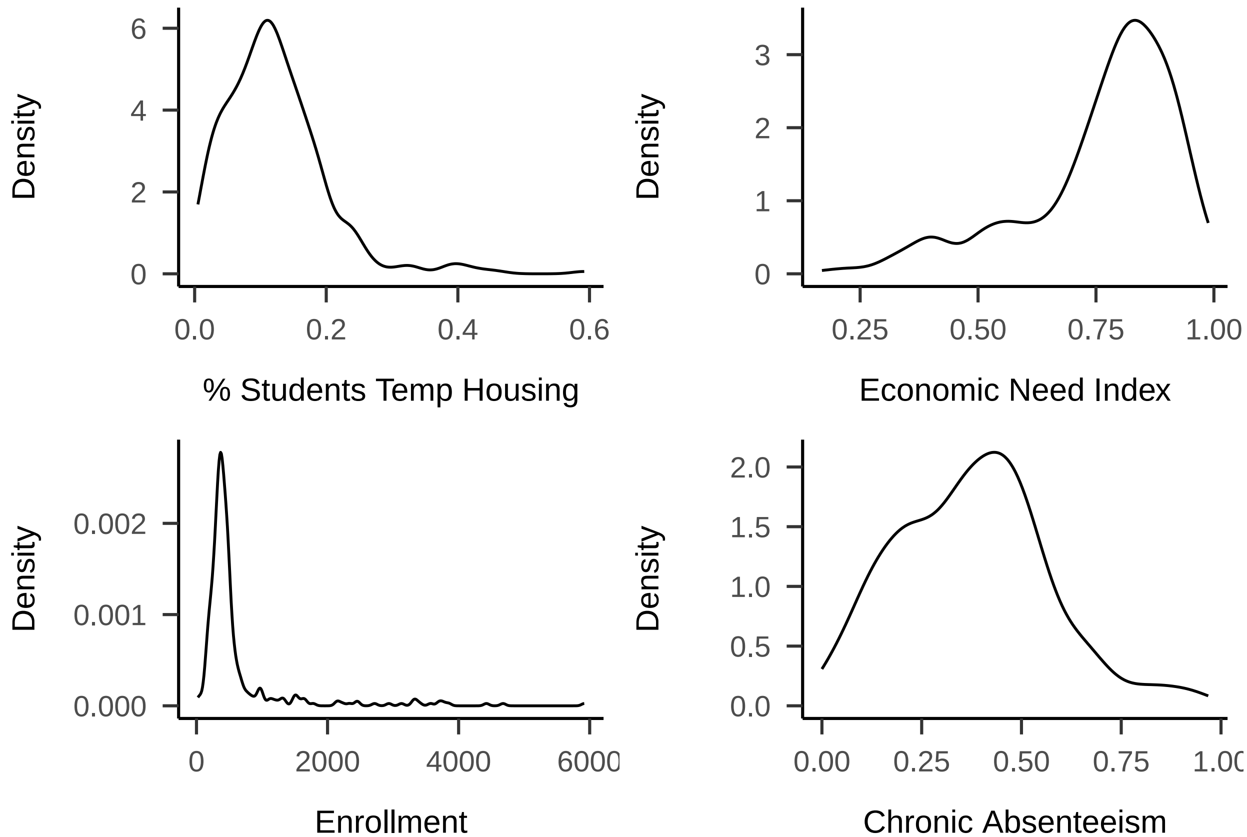


Figure 4

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.

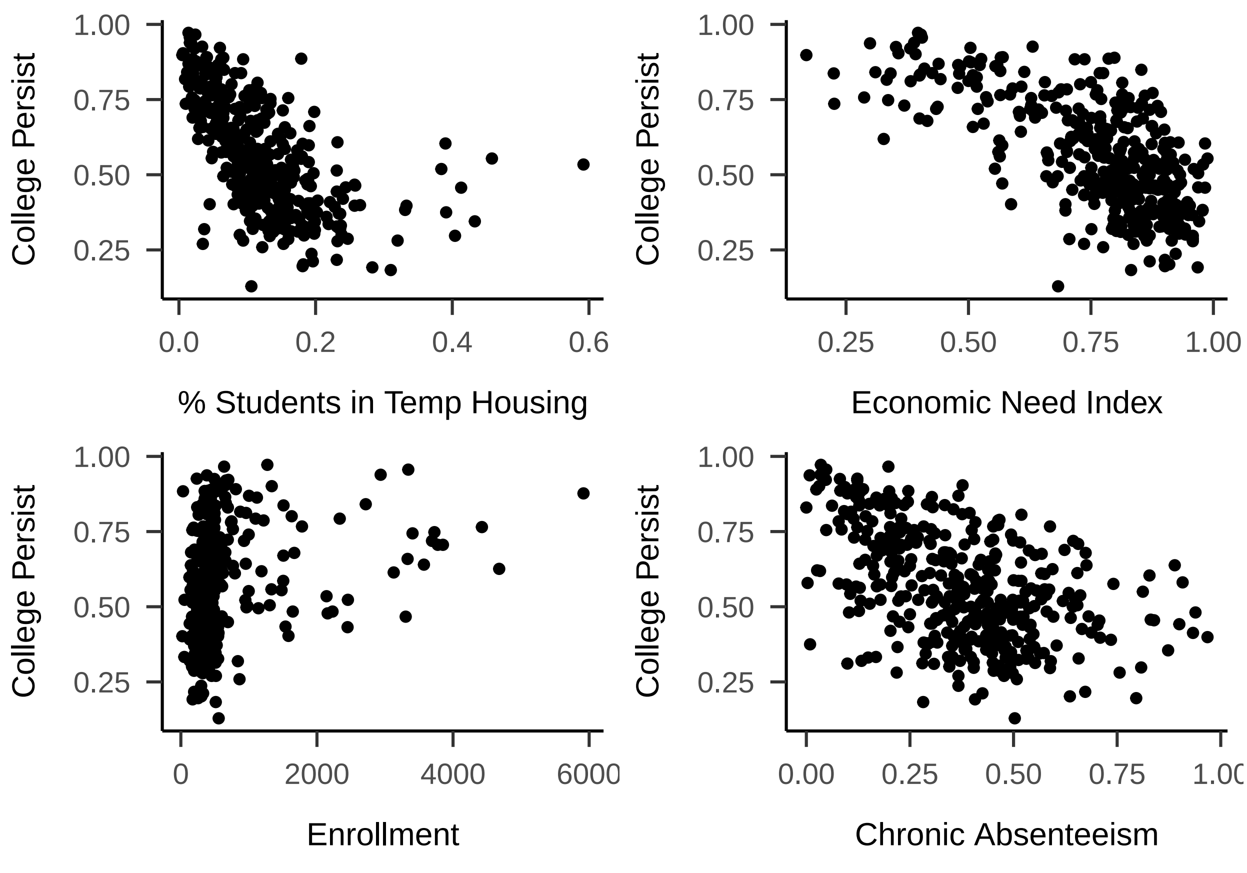


Figure 5

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.

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 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.

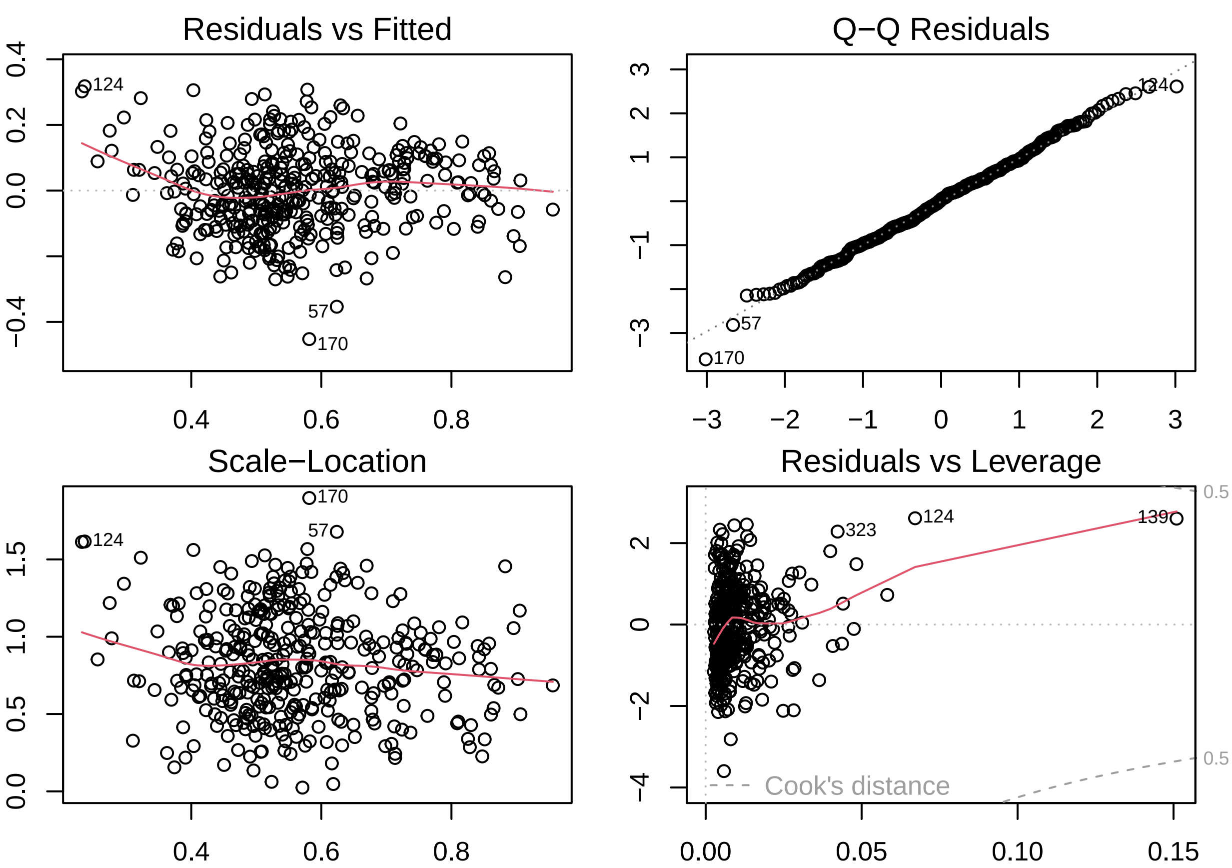


Figure 6

We see no strong trend in the residuals vs. fitted plot, indicating heteroscedasticity.

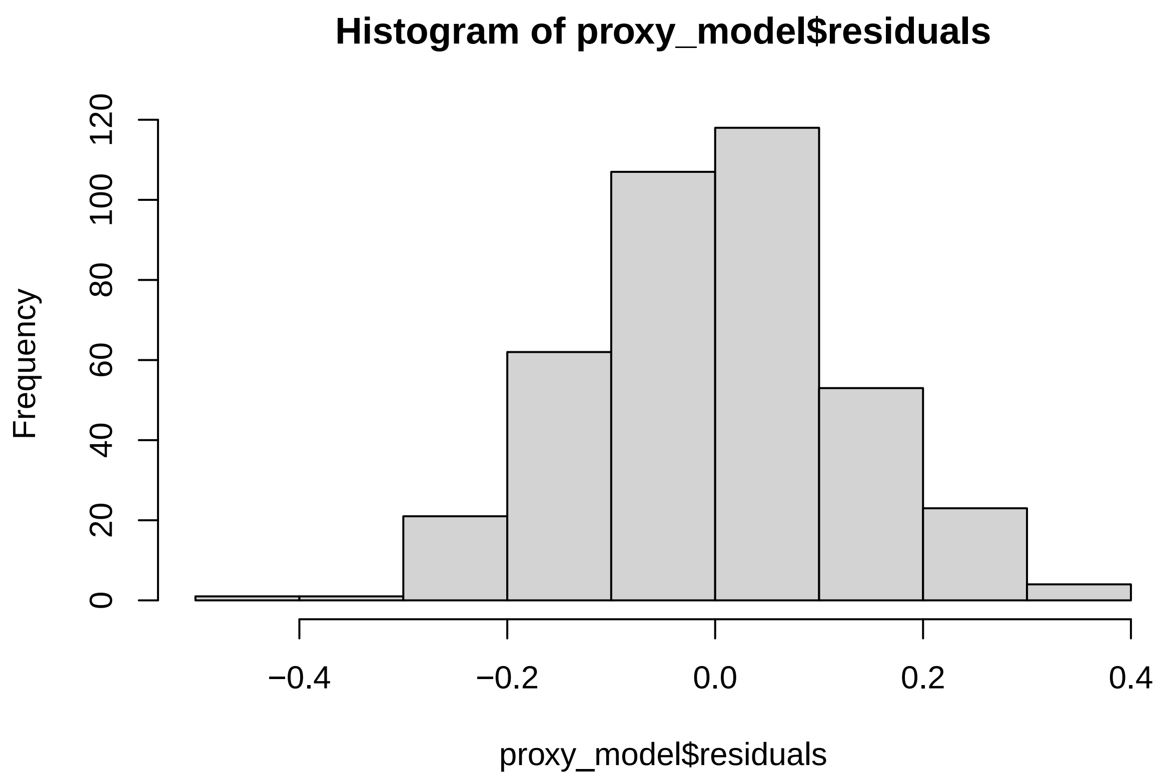
We can also test the assumption of normally-distributed residuals via a Shapiro-Wilk test for normality. Here we operate with the null H_0 and alternative hypotheses H_a:

- 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:




```
##
## 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
```

Experimentation and Results

Model Evaluation.

```
## [1] 0.1645167
```

```
## [1] 0.1425847
```

```
## [1] 0.1434627
```

We can also use the Akaike and Bayesian Information Criterion for evaluatng the complexity of our models.

```
## AIC for base model (rating results): -327.068181870516
```

```
## AIC for proxy variable model: -502.587529886915
```

```
## AIC for WLS model: -508.994310071707
```

```
## BIC for base model (rating results): -307.237448174897
```

```
## BIC for proxy variable model: -482.756796191296
```

```
## BIC for WLS model: -489.163576376089
```

Conclusion

TODO

- 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)
library(car)
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")

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



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