Preserving "Merit" and Redressing "Underrepresentation" in New York City's Specialized High Schools

Scott Karr

20 January 2021

Contents

Abstract **2**
Key Words: . . . 2

Introduction **3**

Literature Review **3**
Theoretical Constructs . . . 4
Methodology . . . 7
Findings . . . 7

Part 1: Data Exploration **8**
Data Summary . . . 8
Compositional Analysis . . . 8
Missing Data . . . 9
Descriptive Statistics . . . 10
Inferences from Theoretical Constructs . . . 11
Correlation Matrix . . . 18
Conclusion of Data Exploration . . . 19

Part 2: Data Preparation **20**
Step 1: Imputing Data . . . 20
Step 2: Converting Categorical Variables to Factors . . . 20
Step 3: Addressing Zero-Inflated skew . . . 20

Part 3: Model Building **21**
Model Preparation, . . . 21
Model 1: Logit Forward Selection + AIC . . . 21
Models 2,3,4: Poisson, Quasi-Poisson & Negative Binomial . . . 22
Models 5,6: Zero-Inflated Poisson & Negative Binomial . . . 24

Part 4: Model Selection **25**
Model Comparison . . . 25

Part 5: Analysis **27**
Discussion and Conclusions . . . 27

Part 6: Bibliography **29**

Part 7: Data Dictionary **30**

Abstract

The use of high-stakes testing as a fundamental determinant of students' future prospects in the public education system is a well established phenomenon. Jonathan Supovitz's article in The Journal of Educational Change offers four major theories that suggest why this is: *Motivational theory*, which argues that test-based accountability can catalyze improvement; *Alignment theory*, which argues that test-based accountability can enable structural consistency among major components of the educational system; *Information theory*, which tells us that analytics can be used as a feedback mechanism to drive performance improvements; and *Symbolism*, which emphasizes that systems of accountability signal important values to stakeholders [1].

New York City's nine specialized high schools (SPHS) have a long history serving the educational needs of students who excel academically and/or artistically. These schools are also a crucible for evaluating the effectiveness of high school placement exams in measuring merit and in providing access to qualified yet underrepresented populations of the city. All students eligible to attend city high schools have the option of applying to one of these specialized schools. Only applicants to the Fiorello H. LaGuardia High School of Music & Art and Performing Arts are evaluated on the basis of their portfolio, a studio audition and a review of their academic record. For the other eight specialized high schools, admissions is based soley on the Specialized High Schools Admissions Test (SHSAT) [2].

This paper looks at 570 middle schools and examines factors that contribute to students passing the New York City Specialized High School exam (SHSAT) which determines offers of acceptance to eight of these nine coveted schools. Also explored is the use of the New York State Tests as an altenative that is taken by a much larger proportion of the student population and is more closely aligned with the NYCDOE curriculum. The goal of this research is to predict both the likelihood a middle school will have SPHS acceptances, the number of SPHS acceptances and the factors that drive these offers. Predictive models were constructed based upon theoretical constructs (categorized by economic, behavioral, demographic, geographic and performance factors) to gauge their impact on SPHS acceptances. The objective is to model an SPHS admissions process that preserves academic merit and also represents the city's diverse popuation more equitably.

Key Words:

High-Stakes Testing, SHSAT, Ranked Choice Selection, School Choice, Predictive Modeling, Behavioral Economics

[1] Supovitz, J. (2010). Is high-stakes testing working? @Penn GSE A Review of Research, 7(2), 3-8. Retrieved from http://www.gse.upenn.edu/review/feature/supovitz

[2] "Specialized High Schools". NYC Department of Education. 2021. Retrieved January 04, 2021. Retrieved from https://www.schools.nyc.gov/enrollment/enroll-grade-by-grade/specialized-high-schools

Introduction

School choice in public education has historically been a matter of local control, but over time this has been challenged based on the notion that local schools supported by property taxes are inequitable because school quality becomes dependent on the "wealth of one's neighbors." [3] In recent decades the debate has shifted toward accountability and representation. While local control is still a factor, the New York City Department of Education (NYCDOE) for the past several years assigns students to high schools using an algorhithmic matching system based on several factors including preference, attendance record, geography, and performance on the New York State test scores [4]. The Specialized High School Admissions process in particular–which is soley based on the Specialized High School Admissions Test (SHSAT)–is indicative of this merit-based approach to assigning schools and has elevated the importance of testing. Given this context, the following research attempts to answer three significant questions:

- Is it possible to accurately predict SPHS feeder middle schools using factors derived from the Literature Review on high stakes testing?
- For middle schools predicted to be SPHS feeders, can the number of students accepted be accurately predicted?
- How can feeder middle schools be more representative of the city's diverse demographics and yet preserve their reputation of academic merit?

More broadly implied by these research questions is improvement of the NYCDOE's selection process for high school. For example, if a middle school is lacking in the number and diversity of students that get SPHS offers, this could signal an evaluation of how that school is aligned to the curriculum so as to improve upon the school's performance. In this way the high school selection can benefit from this information feedback loop to improve SPHS acceptances, diversity and improved academic acheivement.

Literature Review

The literature review identifies policy strategies implicit in the NYCDOE use of high-stakes testing to determine high school acceptances. Variations of these same strategies may be more broadly applicable to the overall high school selection process. Several such strategies were evalutated and modeled for predicting which feeder middle schools lead to SPHS acceptances. Of those reviewed, a few were chosen as being most tractable to the data set's predictor variables. This approach is the basis for testing theories as described in the literature and integrating them into the models that were chosen.

The literature presents challenges in translating to linkages with the data set but can be surmized by classifying the dataset's predictor variables and matching each to a theoretical contruct. The table below summarizes these relationships. To facilitate this research extensive use of the NYC Department of Education (NYCDOE) Information Hub. [5] and PASSNYC [6] was made to access data related to 591 middle school records that could potentially be feeder schools to the New York City Specialized High Schools. This data

[3]Fleeter, Howard B. "The Impact of Local Tax-Based Sharing on School Finance Equity in Ohio; Implementation Issues and Comparative Analysis." Journal of Education Finance 20, no. 3 (1995): 270-301. Accessed January 16, 2021. http://www.jstor.org/stable/40703928.

[4]Supovitz, J.A. & Klein, V. (2003). Mapping a course for improved student learning: How innovative schools systematically use student performance data to guide improvement. Philadelphia, PA: Consortium for Policy Research in Wiliam, D., & Leahy, S. (2006, April). A theoretical foundation for formative assessment. Paper presented at the National Council on Measurement in Education, San Francisco.

[5]"Test Results". NYC Department of Education. 2021. Retrieved January 04, 2021. Retrieved from https://infohub.nyced.org/reports/academics/test-results

[6]Yiping, Lai (2018) Target Schools & Action Recommended for PASSNYC. Retrieved January 04, 2021. "https://www.kaggle.com/laiyipeng/target-schools-action-recommended-for-passnyc

includes information on the number of testers, offers, demographics, economics, geography and performance factors and served as the basis for the research discussed herein.

Variable Classification

The predictor variables in this study break-down along lines of related qualatative measures. For example, all the behavioral predictors were compiled from family surveys issued by the NYCDOE. Behavioral measures also have significant zero-inflated distributions likely indicative of non-responsiveness to surveys.

Finally, the extent to which any particular predictor influences the overall model is indicated by the correlation matrix at the end of part 1 and the coefficients of the multiple regression models in part 3. Theoretical Constructs associated with specific predictors have been included to provide some explanatory value regarding how those predictors influence SPHS acceptances.

- **Geographic** - Most of these variables are political boundaries and their numerical values are not ranked catagorical measurements. 'LATITUDE', 'LONGITUDE', 'BOROUGH', 'DBN', 'ZIPCODE', 'COMMUNITY_SCHOOL'.
- **Economic** - These related variables reflect both the actual funding per school and the students' need for funding. Economic Need and Income are inversely related factors. 'ECONOMIC_NEED_INDX', 'INCOME'.
- **Behavioral** – These related variables are mainly responses based upon yearly survey sent to families by the NYC DOE. As indicated in part 2, the zero-inflated distributions likely reflect missing information on surveys. 'PCT_ATTENDANCE', 'PCT_TRUST', 'PCT_EFFECTIVE', 'PCT_SUPPORTIVE', 'PCT_ABSENCES', 'PCT_RIGOROUS', 'PCT_COLLABORATIVE', 'PCT_FAMILY_TIES', 'PTRATIO', 'CLASS_SIZE'.
- **Demographic** - These related variables are proxies for underrepresentation indicating racial, ethnic, gender and language breakdown per school. 'PCT_BLACK', 'PCT_HISPANIC', 'PCT_ELL', 'PCT_WHITE', 'PCT_ASIAN', 'PCT_FEMALE', 'SPHS_APPLICANTS', 'SPHS_TESTERS', 'PCT_4S_UNDRRP'.
- **Performance** - These related variables are measurable academic assesments. It should be noted that participation rates are much greater on State Tests than the SHSAT which is a high stakes single exam. 'PCT_4S', 'PCT_4S_UNDRRP', 'PCT_4S_ECNDSV', 'ELA_PROF', 'MATH_PROF', 'SPHS_FEEDER', 'SPHS_OFFERS'.

Theoretical Constructs

Adverse Selection

Concentrating SPHS admissions into a small number of feeder middle schools is a type of principal-agent problem where the agent (NYCDOE) has more information about school quality, student performance and academic options than the principals (the students). When such information asymmetries occur, they can lead to a system failure resulting in a few students that benefit from best schools while the rest of the are left with lower-quality options.

This outcome can be rectified by addressing the root of the assymetric information through publishing screens (student test scores, school evaluations) and signals (academic alternatives to specialized schools, test preparation policies).[7]

[7]George A. Akerlof, 1970. "The Market for"Lemons": Quality Uncertainty and the Market Mechanism," The Quarterly Journal of Economics, Oxford University Press, vol. 84(3), pages 488-500.

Rank Choice Matching

All prospective high school freshmen rank up to 12 schools they would like to attend for which they have met the requirements. A matching algorhithm then applies a minimization strategy to the students' selections that best match choices to actual outcomes. For unscreened high schools, only student preferences drive the algorithm, whereas in other schools, several factors determine admissions as described below.[8]

- Audition - Programs demonstrating proficiency typically in performing arts/visual arts.
- Educational Option - Programs designed for a normal distribution of students by State Test scores.
- Limited Unscreened - Programs prioritizing demonstrated interest.
- Screened - Programs ranking students based on grades, State Test scores and attendance.
- Test - Programs ranking students using the Specialized High Schools Admissions test (SHSAT)
- Unscreened - Programs that are ranked by student preference.
- Zoned - Programs ranking students by geographic zoning.

Research underlying this matching process shows that lottery systems based on school preference tend perform better on school tests and have reduced truancy rates. This research suggests students who are offered their top choice have increased motivatation to perform well [9].

Underrepresentation

The NYCDOE high-stakes testing policy for SPHS has become a focal point of criticism in that it has resulted in a lack of diversity at these elite schools which are predominately Asian and White and increasingly admit a smaller proportions of Black and Hispanic students.[10]

Alignment factors

Alignment Theory as applied to SPHS admissions look at linkages between standards, curricula, assessments, and instruction to achieve desired academic goals. To acheive these goals teaching and learning activities require performance assessments that feedback their effectiveness. [11] [12] [13].

Information factors

Applying Information Theory to the SPHS is that learning–measured as subjective entropy (subjective uncertainty)–diminishes over time as students master the NYCDOE curriculum and this reflects in rising test scores [14].

[8] https://ibo.nyc.ny.us/iboreports/preferences-and-outcomes-a-look-at-new-york-citys-public-high-school-choice-process.pdf

[9] Justine S. Hastings & Christopher A. Neilson & Seth D. Zimmerman, 2012. "The Effect of School Choice on Intrinsic Motivation and Academic Outcomes," NBER Working Papers 18324, National Bureau of Economic Research, Inc.

[10] Sean Patrick Corcoran & E. Christine Baker-Smith, 2018. "Pathways to an Elite Education: Application, Admission, and Matriculation to New York City's Specialized High Schools," Education Finance and Policy, MIT Press, vol. 13(2), pages 256-279, Spring.

[11] Biggs, John B.; Tang, Catherine Kim Chow (2011) Teaching for quality learning at university: what the student does. Maidenhead: McGraw-Hill. ISBN 9780335242757.

[12] Smith, Calvin (Novemeber 2008). "Design-focused evaluation". Assessment & Evaluation in Higher Education. 33 (6): 631–645. doi:10.1080/02602930701772762. S2CID 144731064.

[13] Resnick, L. B., Rothman, R., Slattery, J. B., & Vranek, J. L. (2003-2004). Benchmarking and Alignment of Standards and Testing. Educational Assessment, 9(1-2), 1–27. https://doi.org/10.1207/s15326977ea0901&2_1

[14] E. Pfaffelhuber (1972) Learning and Information Theory, International Journal of Neuroscience, 3:2, 83-88, DOI: 10.3109/00207457209147016

Motivational factors

Motivation is often cited as a driving factor in learning but interestingly, external rewards have been shown to be counterproductive to motivation particulary during early childhood development. Motivation is required to prepare for the SPHS entrance exam which pre-supposes middle schools support this initiative through the academic curriculum, communcation and test preparation. [15].

Symbolic factors

Signaling the value of merit in academic achievment is important to encourage participation in the competition for SPHS offers. Reflecting the diverse student population is also a primary value underlying the issue of SPHS underrepresentation and has been a major criticism of the current school choice process.

Mapping Theory to Models

The 'Corresponding Variables' listed above are meant to provide a mechanism for linking relevant theoretical strategies to predictor variables in corresponding models. While any predictive models we might choose to build would not be limited solely to the eight variables listed below, the intent is to choose the most representative indicators to model and provide explantion for SPHS outcomes.

Table 1: Theoretical Construct Mappings

Theoretical Construct	Source	Category	Corresponding Variable
Information Theory	Supovitz	Performance	SPHS_TESTERS, ELA_PROF, MATH_PROF, PCT_4S_UNDRRP
Alignment	Biggs, Smith	All 5 Catagories	SPHS_TESTERS, ELA_PROF, MATH_PROF, PCT_4S_UNDRRP PCT_WHITE, PCT_ELL
Motivational Theory	Deci	Behavioral	NONE
Symbolism	Supovitz	Demographic	PCT_WHITE, PCT_ELL, PCT_4S_UNDRRP
Underrepresentation	Corcoron, Baker-Smith	Demographic	PCT_BLACK, PCT_HISPANIC, PCT_WHITE, PCT_ASIAN, PCT_ELL
Ranked Choice Selection	Kapor, Neilson & Zimmerman	Behavioral, Performance	SPHS_TESTERS, ELA_PROF, MATH_PROF, PCT_4S_UNDRRP
Adverse Selection	Akerlof	Behavioral, Performance	SPHS_TESTERS, ELA_PROF, MATH_PROF, PCT_4S_UNDRRP

[15]Deci, E.L., Koestner, R. & Ryan, R.M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. Psychological Bulletin, 125, 627-668.

Methodology

Data Exploration

This section explores the characteristics of each individual variable contained within the NYCDOE data set, including data types, range of valid values, distributions and correlations with one another. In addition, variables with missing or invalid data values are investigated for imputation or removal.

Data Preparation

This section develops strategies for handling missing or invalid data values and the separation of the master data set into dedicated "regression modeling "Training" and "Evaluation" subsets for regression modelling. Creation of separate training and evaluation data sets was in consideration of testing the regression models with unbiased SPHS admissions records.

Regression Modeling

This section identifies, develops, and tests regression models that represent the SPHS admissions process. SPHS admissions are first predicted using a binary logistic distribution and then evaluated for effectiveness. The "best" model was selected on the basis of performance metrics including AIC score, AUC, accuracy, classification error rates, precision, specificity, sensitivity, and F1 scores. In addition, predictions of the likely number of admissions implied application of a count regression model and such models were fitted (e.g., Poisson, Negative Binomial, zero-inflated) to the school data set. These models were compared on the basis of AIC scores, log likelihood and whether or not they produced realistic predictions.

Model Selection

For those schools with signficant SPHS acceptances, the selected count regression model was then used to predict the likely number of acceptances. Two models were developed for each type of count model one with the *SPHS_OFFERS* response variables representing: high SHSAT scores; and high New York State test scores. The later is more widely used and therefore more reflective of the NYCDOE student population. Finally, the result of all regression models are summarized and their predictions evaluated against actual admissions data.

Findings

The binomial logistic regression for (Model1) predicted with 95% accuracy all feeder middle schools that used the SHSAT as the entrance test and for (Model2) predicted with 92% accuracy all feeder middle schools using the NY State Test as the entrance test.

Five count models were built to predict the number of SPHS offers per middle school. Negative binomial models were chosen over Poisson models because the Poisson's had dispersion that varied signficantly from 1 indicating a poor fit. Zero-inflated negative binomial (ZINB) models were ultimately chosen after testing for skew, log-likelihood, chi-squared and p-values that indicate best fit. The ZINB model for testing SHSAT offers still contains 18% zero-inflated data but its fit was still somewhat better then that of the ZINB model for testing NY State Test offers.

NY State Test scores being more closely aligned to the NYCDOE curriculum and taken by more students then the SHSAT resulted in more SPHS acceptances with a more dispersed distribution among schools. The salient point from this observation is that SPHS admissions based on the NY State Test addresses the selection bias criticism that qualified by Black and Hispanic students are underrepresented.

Part 1: Data Exploration

Data Summary

The data set in this study focuses on the likelihood of whether New York City's Middle Schools are feeders to the Specialized High Schools and to accurately predict the number of students sent. There are 591 rows of academic data, each representing geographic, economic, behavioral, demographic and performance attributes of a single NYCDOE middle school. For each acceptance record there are 29 attributes that could potentially be used as predictor variables and two response variable: *SPHS_FEEDER* which indicates if the school has five or more SPHS offers; and *SPHS_OFFERS* which indicates the actual number of offers.

Once the data set is complete, Generalized Linear Modelling (GLM) can be applied in a two-stage modeling approach. First model the response variable using a probability distribution, such as the binomial or Poisson distribution. Second, model the parameter of the distribution using a collection of predictors and a Logit form of multiple regression. This exercise builds an appropriate model that classifies SPHS offers using a subset of the 29 attributes to predict SPHS offers using a Logit model for multiple linear regression.

"The assumptions required for statistical tests in logistic regression are far less restrictive than those for Ordinary Least Squares regression. There is no formal requirement for multivariate normality, homoscedasticity, or linearity of the independent variables within each category of the dependent variable." However, logistical regression models should have little or no multicollinearity. That is that the independent variables should be independent from each other. [16]

Compositional Analysis

The NYCDOE administers the SHSAT to approximately 30% of eligible students each year which is roughly 25,000 eighth graders. However, offers are not distributed equally among students from different middle schools. In 2018 25% of SPHS offers went to the top 10 feeder middle schools and these offers skew heavily to Asian and White students [17]. The concentration of admissions is the crux of this paper, particularly the opportunities that exist for underrepresented high-achieving students. Currently, acceptance to the SPHS schools is solely determined by the SHSAT, a single high-stakes test administered at the end of 7th grade and for which the subject matter being tested is not directly taught in the NYCDOE curricula.

[16] http://www.sagepub.com/upm-data/5081_Spicer_Chapter_5.pdf, page 135

[17] Sean Patrick Corcoran & E. Christine Baker-Smith, 2018. "Pathways to an Elite Education: Application, Admission, and Matriculation to New York City's Specialized High Schools," Education Finance and Policy, MIT Press, vol. 13(2), pages 256-279, Spring.

Missing Data

There are 8 predictors in the data set that have missing values *PCT_4S*, *PCT_4S_UNDRRP*, *PCT_4S_ECNDSV*, *ELA_PREF*, *MATH_PREF*, *INCOME*, *PTRATIO* and *CLASS_SIZE*. There were also 21 of 591 original rows that had multiple missing values that were not considered to contain outliers of significance and these rows were deleted rather than imputed.

Table 2: Imputation & Removal

Variable	Stats / Values	Freqs (% of Valid)	Missing
INCOME [numeric]	Mean (sd) : 46242.2 (19615.7) min < med < max: 20931.7 < 41056.2 < 143926.5 IQR (CV) : 24215.1 (0.4)	210 distinct values	381 (64.47%)
PCT_4S [numeric]	Mean (sd) : 13.6 (16.2) min < med < max: 0 < 7.6 < 93.8 IQR (CV) : 14.7 (1.2)	392 distinct values	21 (3.55%)
PCT_4S_UNDRRP [numeric]	Mean (sd) : 2.2 (5.3) min < med < max: 0 < 0 < 57 IQR (CV) : 2.4 (2.3)	215 distinct values	21 (3.55%)
PCT_4S_ECNDSV [numeric]	Mean (sd) : 6.9 (8.3) min < med < max: 0 < 4.3 < 51.2 IQR (CV) : 7.6 (1.2)	339 distinct values	21 (3.55%)
ELA_PROF [numeric]	Mean (sd) : 2.5 (0.4) min < med < max: 1.8 < 2.5 < 3.9 IQR (CV) : 0.5 (0.1)	149 distinct values	6 (1.02%)
MATH_PROF [numeric]	Mean (sd) : 2.6 (0.5) min < med < max: 1.9 < 2.5 < 4.2 IQR (CV) : 0.7 (0.2)	178 distinct values	6 (1.02%)
CLASS_SIZE [numeric]	Mean (sd) : 24 (4.2) min < med < max: 11.8 < 23.9 < 45.6 IQR (CV) : 5.5 (0.2)	461 distinct values	113 (19.12%)
PTRATIO [numeric]	Mean (sd) : 13.4 (2.8) min < med < max: 4.1 < 13.4 < 24.2 IQR (CV) : 3.8 (0.2)	120 distinct values	112 (18.95%)

The remaining 570 rows in the data set required imputation of missing data for *INCOME*, *PTRATIO*, *CLASS_SIZE*. To do this, forward OLS models were built to predict the missing values using the remaining preditors in each row to impute them. This final data set containing 570 complete rows is used as the baseline for modeling in part 3. The rows that were removed and imputed data did not appear result in a loss of valid outliers that would change the models' predictive capability.

Descriptive Statistics

Data Type Analysis reveals that the predictors *DBN*, *DISTRICT*, *ZIP*, *COMMUNITY_SCHOOL* and *SPHS_FEEDER* are all either binary or categorical variables and have been transformed to factors variables. Inclusion in the final dataset depends upon completeness of records as well as correlation with the response variables and non-collinearity with other predictors. The *Descriptive Statistics* below are following imputation of missing data and removal of redundant variables.

Some of the remaining predictors such as *ECONOMIC NEED*, *PCT_ELL*, *PCT_ASIAN*, *PCT_BLACK*, *PCT_WHITE*, *PCT_4S*, *PCT_UNDERREPRESENTED_4S*, *SPHS_TESTERS*, *SPHS_OFFERS* show evidence of significant skew as indicated by the large differences between their mean and median values.

Other predictors such as *PCT_ELL*, *PCT_ASIAN*, *PCT_WHITE*, *PCT_ATTENDANCE*, *PCT_RIGOROUS*, *PCT_COLLABORATIVE*, *PCT_SUPPORTIVE*, *PCT_EFFECTIVE*, *PCT_FAMILY_TIES*, *PCT_TRUST*, *PCT_ELA_4S*, *PCT_MATH_4S*, *SPHS_APPLICANTS*, *SPHS_TESTERS* and *SPSH_OFFERS* show evidence of potentially being problematic as evidenced by their large kurtosis and standard error values. The analysis that follows explores these observations in more detail

Table 3: SPHS Predictor Descriptive Statistics

	n	mean	sd	median	min	max	range	skew	kurtosis	se
LATITUDE	570	41	0	40.734	41	41	0	0	-1	0.004
LONGITUDE	570	-74	0	-73.920	-74	-74	1	0	2	0.003
ECONOMIC_NEED_INDX	570	67	19	72.150	10	94	83	-1	0	0.803
INCOME	570	45693	17971	41883.291	-5220	140085	145305	1	2	752.737
PCT_ELL	570	11	11	8.000	0	99	99	3	13	0.460
PCT_ASIAN	570	10	15	3.000	0	77	77	2	4	0.629
PCT_BLACK	570	35	29	26.500	0	97	97	1	-1	1.211
PCT_HISPANIC	570	42	26	36.000	2	100	98	0	-1	1.097
PCT_WHITE	570	11	18	2.000	0	88	88	2	4	0.751
PCT_ATTENDANCE	570	93	9	94.000	0	100	100	-9	88	0.385
PCT_ABSENCES	570	21	14	18.500	0	100	100	2	7	0.594
PCT_RIGOROUS	570	88	7	88.000	0	100	100	-4	53	0.277
PCT_COLLABORATIVE	570	88	8	89.000	0	100	100	-3	25	0.332
PCT_SUPPORTIVE	570	85	6	85.000	0	100	100	-5	71	0.250
PCT_EFFECTIVE	570	82	10	84.000	0	99	99	-2	9	0.403
PCT_FAMILY_TIES	570	81	7	80.500	0	99	99	-3	30	0.292
PCT_TRUST	570	90	6	90.000	0	100	100	-6	86	0.252
PCT_4S	570	14	16	7.580	0	94	94	2	4	0.681
PCT_4S_UNDRRP	570	2	5	0.000	0	57	57	5	37	0.221
PCT_4S_ECNDSV	570	7	8	4.293	0	51	51	2	6	0.349
SPHS_APPLICANTS	570	129	120	89.000	9	769	760	2	6	5.009
ELA_PROF	570	3	0	2.460	2	4	2	1	1	0.016
MATH_PROF	570	3	0	2.525	2	4	2	1	0	0.020
CLASS_SIZE	570	24	4	23.816	12	46	34	0	2	0.165
PTRATIO	570	13	3	13.582	6	24	18	0	1	0.109
PCT_FEMALE	570	50	10	49.049	0	100	100	1	19	0.400
SPHS_TESTERS	570	43	61	23.000	0	394	394	3	9	2.548
SPHS_OFFERS	570	7	22	0.000	0	205	205	5	30	0.912
SPHS_FEEDER	570	0	0	0.000	0	1	1	1	0	0.017

Note that the NY State Test distributions for *PCT_ELA_4S* and *PCT_MATH_4S* are leftward skewed but not so highly skewed as when *SPSH_OFFERS* represents students passing the SHSAT per school. The overconcentraion of SHSAT testers and offers raises the prospect that NY State Test scores might offer a more balanced approach determining *SPSH_OFFERS*.

Inferences from Theoretical Constructs

Ranked Choice Analysis

Student high school options are based on a centralized school choice model that assigns students to high schools based on a variety of published factors. For the SPHS schools the SHSAT test is the only factor. In binary logit models, only the variable *SPHS_TESTERS* shows up as being a significant influence on whether a middle school is likely to feed SPHS admissions in both models.

Model 2 uses the NY State Test to determine SPHS offers because it is more widely offered and more aligned to NYCDOE curricula. What is notable is that the signs in Model 2 are opposite those in Model 1. What this says is that if the NY State Test determines SPHS admissions, high concentrations of testers in particular schools negatively correlate with SPHS acceptances. This would be the case when high scorers are more uniformly distributed among schools.

Model 1	Variable	Model 2	Variable
- 30.5446	Intercept	+ 27.8199	Intercept
+ 2.9824	log(SPHS_TESTERS + 1)	- 3.5350	log(SPHS_TESTERS + 1)
+ 5.1919	ELA_PROF	- 1.5259	log(PCT_4S_UNDRRP + 1)
+ 0.8371	log(PCT_WHITE + 1)	- 7.4528	MATH_PROF
+ 0.9603	log(PCT_ELL + 1)		

Interpretation of Model 1 Coefficients . . .

- *One percent increase in SPHS TESTERS is associated with a (2.9824 / 100) unit increase being an SPHS FEEDER school.*
- *One percent increase in ELA PROFICIENCY is associated with a 5.1919 unit increase being an SPHS FEEDER school.*
- *One percent increase in PERCENT WHITE STUDENTS is associated with a (0.8371 / 100) unit increase being an SPHS FEEDER school.*
- *One percent increase in PERCENT ELL STUDENTS is associated with a (0.9603 / 100) unit increase being an SPHS FEEDER school.*

Interpretation of Model 2 Coefficients . . .

- *One percent increase in SPHS TESTERS is associated with a (-3.5350 / 100) unit decrease being an SPHS FEEDER school.*
- *One percent increase in UNDERREPRESENTED W SCORES on NY STATE TEST is associated with a (-1.5259 / 100) unit decrease being an SPHS FEEDER school.*
- *One percent increase in MATH PROFICIENCY is associated with a -7.4528 unit decrease being an SPHS FEEDER school.*

Adverse Selection Analysis

The existance of significant zero-inflation requires special modelling consideration because a properly fitted count model would show a proportionate number of SPHS feeders. Zero-inflation signifies an excessive number of non-SPHS feeder schools and the crowding-out many quality school–the so-called "middle school effect" that favors certain schools in the SPHS admissions process. [18]

[18] Sean Patrick Corcoran & E. Christine Baker-Smith, 2018. "Pathways to an Elite Education: Application, Admission, and Matriculation to New York City's Specialized High Schools," Education Finance and Policy, MIT Press, vol. 13(2), pgs. 256-279, Spring.

Alignment, Informational, Motivational & Symbolic Factors

High stakes testing addresses a few goals of the NYCDOE. It provides uniform measure of achievment based upon the stated academic standards. It provides information about student achievment to teachers, administrators, parents and students and perhaps most importantly, it signifies what the city expects a well educated student to know. The SHSAT is the traditional measure for SPHS acceptance is demographically unrepresentative of the student population. The NY State Exams test a larger proportion of the student population are more aligned with NYCDOE curriculum and are taken by a more representative student population.

Table 5: Contingency NY State ELA Test Score Distribution

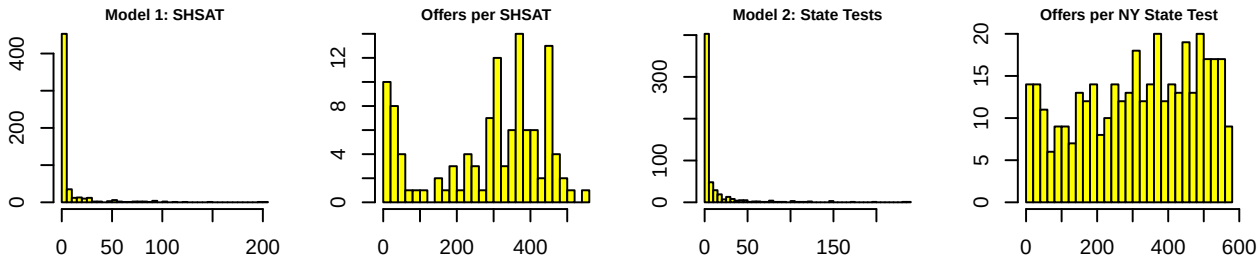
	0	1	Total		0	1	Total
2	8	0	8	2	1.000	0.000	1
3	431	67	498	3	0.865	0.135	1
4	14	50	64	4	0.219	0.781	1
Total	453	117	570	Total	0.795	0.205	1

Table 6: Contingency NY State Math Test Score Distribution

	0	1	Total		0	1	Total
2	136	1	137	2	0.993	0.007	1
3	294	66	360	3	0.817	0.183	1
4	23	50	73	4	0.315	0.685	1
Total	453	117	570	Total	0.795	0.205	1

This alternate prediction model will be developed in the model building section that includes a more disperssed set of potential SPHS feeder schools, thus addressing both underrepresentation and preserving merit as existing NYDOE metrics for SPHS selection. The alternate model still contains zero-inflated counts owing to the difficulty of acheiving 4s on the State Test but the exact policy threshold could be adjusted. As can be seen from the following histogram comparisons, the alternate model results in a signficantly higher yield of **SPHS_FEEDER** schools 189 (33.16%) vs. 117 (20.53%) in the base model. During model selection, predictors derived from 4s on the State Test scores will be removed due to tautological considerations of deriving the response variable from the same data.

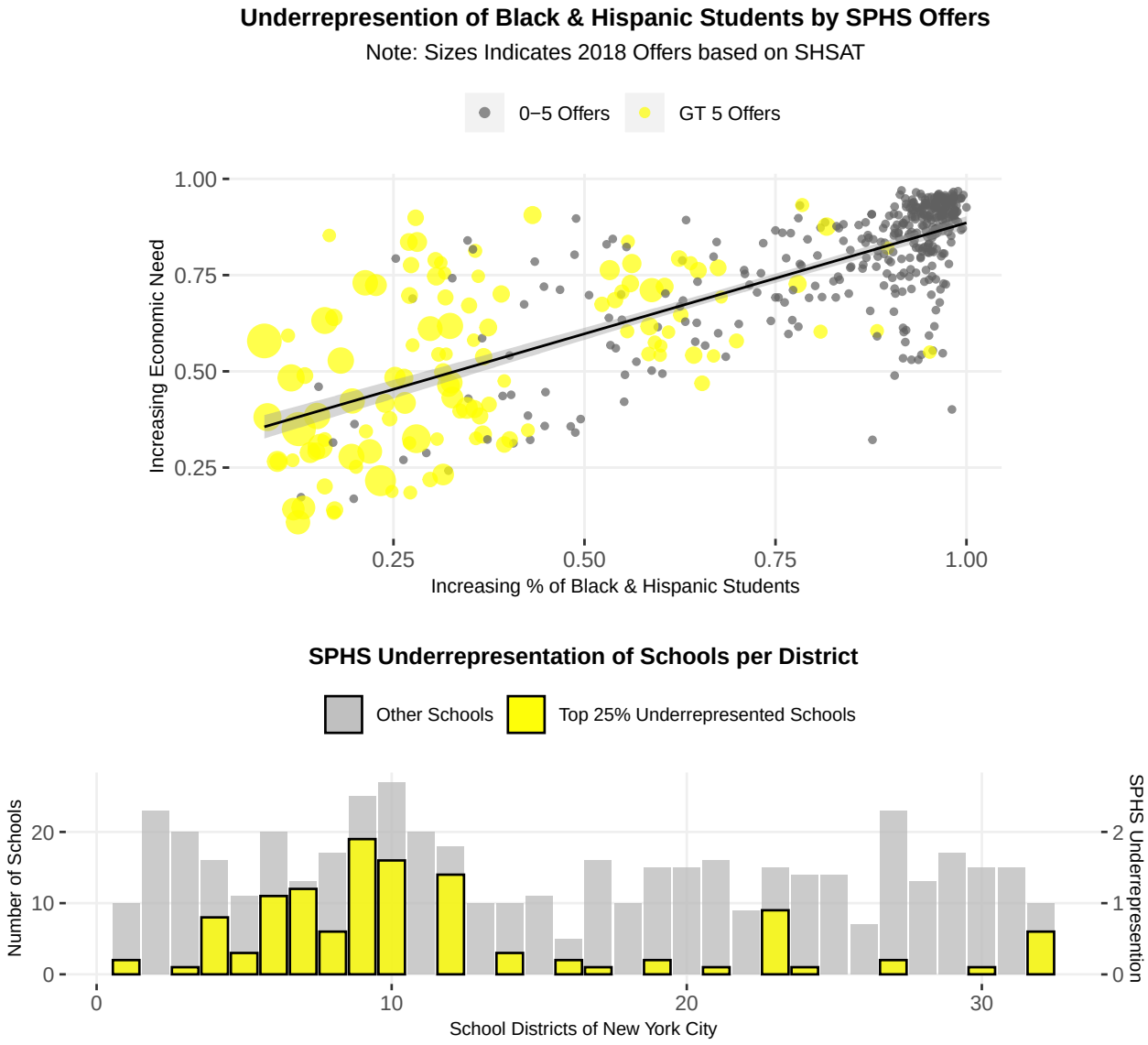
SPHS OFFERS



Underrepresentation

This is the key demographic predictor of SPHS offers and as such warrants special consideration. The top 25% schools with the highest Underrepresentation represent 20 out of 32 school dbns in New York City. Later analysis will show alternatives to the SHSAT and other mitigating factors that may well change the dynamic of this underrepresented population.

To examine whether the most successful feeder middle school are compositionally different from other schools, the following density plot was developed to illustrate *Underrepresentation* specifically of Black & Hispanic SPHS students using the concentration of ethnic makeup per school and the school's Economic Need Index [19] [20] [21].



[19] Sean Patrick Corcoran & E. Christine Baker-Smith, 2018. "Pathways to an Elite Education: Application, Admission, and Matriculation to New York City's Specialized High Schools," Education Finance and Policy, MIT Press, vol. 13(2), pages 256-279, Spring.

[20] https://www.kaggle.com/laiyipeng/target-schools-action-recommended-for-passnyc

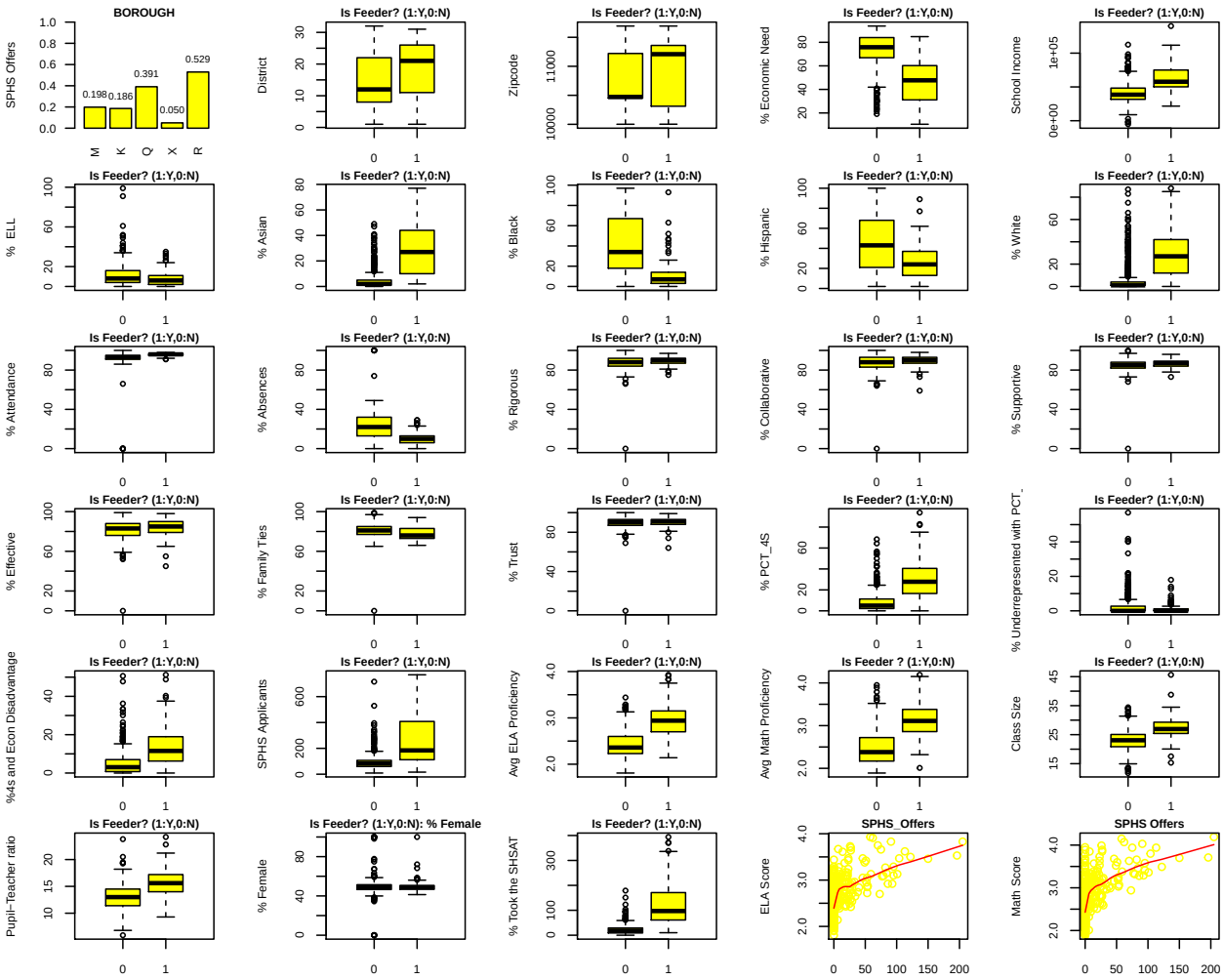
[21] (calculated as % temp housing + % HRA eligible * 0.5 + % free lunch eligible * 0.5)

Evaluating Catagorical Relationship

Prior to developing a multivariate model it is useful to analyze individual relationships between each predictor and the response variable and to evaluate the magnitude and direction of their relationship.

- **Boxplots** evaluating Predictor variables relative to the binary response variable s one way in which we can begin to gain insight into the predictive of each variable to the response variable.
- **Bargraphs** evaluating Predictor variables relative to the catagorical response variable are similar but more dispursed distribution showing the relative performance each catagory to the response variable.
- **Lineplots** evaluating predictor variables to the a continuous response variable can provide trendline insights of their relationship.

Catagorical Relationships



Response Variables *SPHS_FEEDER* is the binary catagorical response variable evaluated for each predictor as illustrated in the table above. Observations of of significance follow.

- **SPHS_FEEDER** = 0: school does not send 5 or more students to an SPHS school.
- **SPHS_FEEDER** = 1: school successfully sends 5 or more students to an SPHS school.
- **SPHS_OFFERS** = #: school's count of successful SPHS offers.

By Geography: *BOROUGH*, *DISTRICT*, *ZIPCODE*, *LONGITUDE* and *LATITUDE* are not `ranked` catagorical predictors, so while range differences are meaningful for schools that have SPHS offers, median differences are not.

- 5 *BOROUGH's* ordered by increasing % of schools with *SPHS_OFFERS*: the Bronx (5% of 139), Brooklyn (18.6% of 183), Manhattan (19.8% of 121), Queens (39.1% of 110) and Staten Island (52.9% of 17).
- 32 *DISTRICT's* show the 25% most underrepresented schools by SPHS OFFERS are concentrated in 20 of DOE's 32 *DISTRICT's*. The Inter Quartile Range (IQR) of offers is nearly the same for SPHS feeder and non-feeder schools so roughly an equal amount of schools per district are SPHS feeders although some districts send far fewer students per school.
- 146 *ZIPCODE's* have schools with *SPHS_TESTERS*. Note though distribution skews toward the low end of the range between 0 and 207 offers with a median of 7 and a mean of 27.32. *ZIPCODE's* sending more than 50 students may be considered outliers.

By Economics: *ECONOMIC_NEED_INDX*, *INCOME* are numeric predictors under consideration.

- *ECONOMIC_NEED_INDX* has a higher median (75) and IQR (70-80) for non feeder schools than median (45) and IQR (30-60). We can deduce unsurprisingly that schools with high economic need are far more likely to not to be feeder schools.
- *INCOME* has a lower median ($40,000) and IQR ($30-$50,000) for non feeder schools than median ($60,000) and IQR ($50-$80,000). We can deduce unsurprisingly that schools with high economic need are far more likely to not to be feeder schools.

By Behavioral: *PCT_ATTENDANCE*, *PCT_TRUST*, *PCT_EFFECTIVE*, *PCT_SUPPORTIVE*, *PCT_RIGOROUS*, *PCT_COLLABORATIVE*, *PCT_FAMILY_TIES*, *PTRATIO*, *CLASS_SIZE* are numeric predictors based on survey response none of which show much difference in their descriptive statistics when comparing SPHS feeder and non-feeder schools.

- *CLASS_SIZE* & *PTRATIO* (Pupil Teacher ratio) are averages per school and surprisingly slightly higher for SPHS feeder schools than for non-feeder schools. Perhaps this has something to do with high demand and crowding in the feeder schools
- *PCT_ABSENCES* has a higher median (20) and IQR (15-30) for non feeder schools than median (10) and IQR (5-15). We can deduce unsurprisingly that schools with high absenteeism are far more likely to not to be feeder schools.

By Demographic: Ethnic, Gender and Language breakdown per school *PCT_ELL*, *PCT_FEMALE* are numeric predictors based on survey response none of which show much difference in their descriptive statistics when comparing SPHS feeder and non-feeder schools. Only racial predictors show in this catagory show signficant differences between SPHS feeder and non-feeder schools.

- *PCT_BLACK* has a higher median (20) and IQR (20-70) for non feeder schools than median (10) and IQR (5-15). We can deduce that schools with higher *PCT_BLACK* population are far more likely to not to be feeder schools. This population is signficantly represented in the underserved population and is therefore a focus of alternate paths in applying and gaining admittance.
- *PCT_HISPANIC* has a higher median (20) and IQR (20-70) for non feeder schools than median (10) and IQR (10-30). We can deduce that schools with higher *PCT_HISPANIC* population are far more likely to not to be feeder schools. This population is signficantly represented in the underserved population and is therefore a focus of alternate paths in applying and gaining admittance.

- *PCT_WHITE* has a higher median (20) and IQR (5-10) for non feeder schools than median (10) and IQR (15-40). We can deduce that schools with higher % White population are far more likely to be feeder schools.
- *PCT_ASIAN* has a higher median (20) and IQR (5-10) for non feeder schools than median (10) and IQR (15-45). We can deduce that schools with higher % Asian population are far more likely to be feeder schools.
- *SPHS_APPLICANTS* & *SPHS_TESTERS* not surprisingly, schools with more SPHS applications and test takers have much more likely to be feeder schools.

By Performance: *PCT_4S*, *PCT_4S_UNDRRP*, *PCT_4S_ECNDSV*, *ELA_PROF*, *MATH_PROF*, *PTRATIO*, *CLASS_SIZE*, *SPHS_APPLICANTS*, *SPHS_TESTERS*

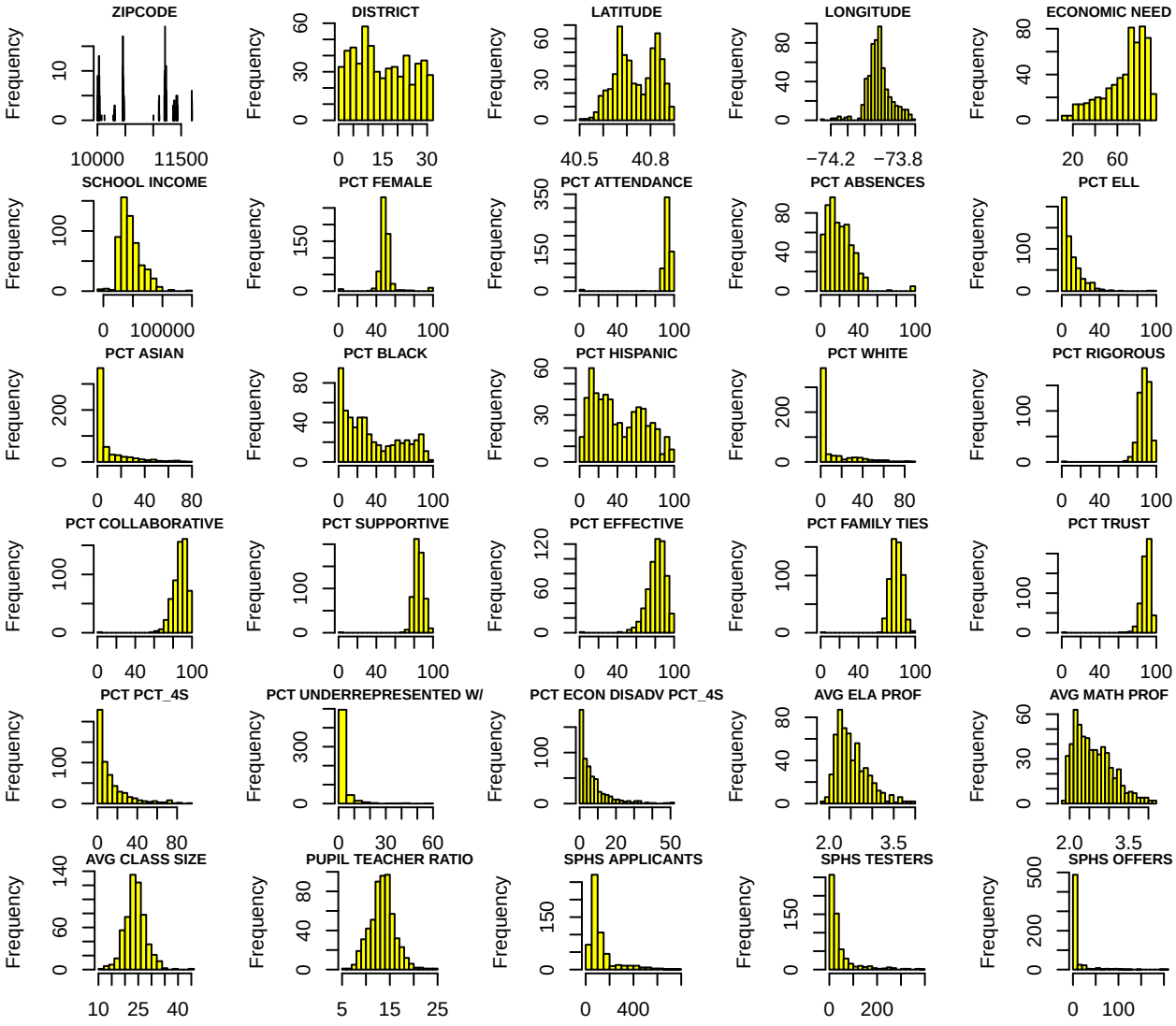
- *PCT_4S* or student's scoring 4's on the New York State Test, have a lower median (10) and IQR (5-10) for non feeder schools than median (25) and IQR (20-40). We can deduce that high State Test scores correlate to being an SPHS feeder schools. Since significantly more students take the state test then the SHSAT, this may well be the basis for getting to a representative pool of applicants. Both tests Math and English proficiency but the State Tests subject matter is more closely aligned with NYCDOE Regents standards and curricula.
- *PCT_4S_UNDRRP* or schools with underrepresented students that score 4's on the New York State Test showed little difference between those in feeder schools and those that are not. This further underscores the notion that there are schools with high acheiving students that are not SPHS feeder schools. The respresentation of this group may very well improve if the State Tests were to be used as the SPHS enrollment criteria.
- *PCT_4S_ECNDSV* are similar to *PCT_4S*
- *ELA_PROF* & *MATH_PROF* the average ELA and Math scores for SPHS feeder scores are ~3 and for non SPHS feeder schools ~2.4.

Evaluating Skew and Clustering

Distributions of variables show the existance of signficant skew for **Economic**, **Demographic** and **Performance** catagories of predictors as well as the **Response** variables. Some skewed predictors are be dominated by high concentration around specific values. For example all **Behavioral** predictors are zero-inflated because their data is survey driven and zeroes likely indicate non-responsiveness.

- **Histograms** evaluating predictor variables population density indicates anomolies such as skew, kurtosis and clustering of numeric variables.

Population Density Distributions



Skew by racial composition or English Language Learners illustrates true variance in the NYCDOE's composition and therefore should not be altered. Much of the skew in **Performance** predictors is explainable by observing that only a fraction of the NYC school population scores 4 on the NY State exams or passes the SHSAT. **Economic** predictors skew is similarly explainable by schools' variance in economic need. The concentration of a few feeder middle schools that disproportionatly account for SPHS offers correlates with economic and demographic factors. Finally, the catagorical variable *COMMUNITY_SCHOOL* contains very little data variation and appeared to add little value for modeling and was therefore removed.

Variable Anomolies

Understanding the nature of high frequencies concentrations and skew for specific predictor variables should not be assumed to be an anomoly that needs to be fixed. High frequency data concentrations cannot easily be transformed without information loss. For example zero-inflated ethnic data may indicate the school actually has no representation for a particular ethnic group or that the groups data went unreported. As such, transformation of such values during data preparation must only be attempted after careful consideration of the consequences.

- **Concentration** of six demographic variables *PCT_ELL*, *PCT_ASIAN*, *PCT_BLACK*, *PCT_HISPANIC*, *PCT_WHITE*, *PCT_ASIAN*, are zero-inflated to a degree reflecting varying proportions of representation by race, ethnicity, gender and English language proficiency. The distributions represent the diverse mix of schools and should not be transformed as this risks data loss and dilutes their predictive value. The last variable, *COMMUNITY_SCHOOL* contains binary categorical data which has little variance and was eliminated as it was of little use in modeling the likelihood of an SPHS offer.

In addition, performance variables *PCT_4S*, *PCT_UNDERREPRESENTED_PCT_4S*, *CLASS SIZE*, *PTRATIO*, *SPHS_TESTERS*, and *SPHS_OFFERS* all exhibit a concentration of values at zero that significantly skew their distributions. This zero inflation is indicative of a significant percentage of students that opt out of the SPHS test influenced by geographic proximity to these schools and preparation resources for the SHSAT provided by many underserved schools.

- **Skew** in Geographic predictors *BOROUGH*, *ZIPCODE* and *DISTRICT* are politically determined and have implications economically and in the SPHS representation. These boundaries should not be transformed as they were deliberately chosen and have direct correlation as they are to SPHS representation.
- **Imputation** of data was considered for eight variables with missing values *INCOME* 381 rows (64.47%), *PCT_4S* 21 rows (3.55%), *PCT_4S_UNDRRP* rows 21 (3.55%), *PCT_4S_ECNDSV* rows 21 (3.55%), *ELA_PROF* 6 rows (1.02%), *MATH_PROF* rows 6 (1.02%), *CLASS_SIZE* rows 113 (19.12%) and *PTRATIO* 112 rows (18.95%). Examination the 21 rows with missing data for **Performance** predictors suggests either making multiple imputations or to remove the rows. These rows do not appear to contain critical outliers that would siginficantly diminish modelling. Removing these 21 rows also has the added benefit of limiting the remaining imputations to *INCOME* 361 rows (63.33%), *CLASS_SIZE* rows 92 (16.14%) and *PTRATIO* 92 rows (16.14%).

Correlation Matrix

Directionality of each predictor against the response variable *SPHS_OFFERS* is provided in the table below. As can be seen, variables showing unusually strong corrleation values are from **Demographics** and **Performance** catagories. It is not suprising to find that *SPHS_TESTERS* has the highest colinearity with *SPHS_OFFERS* since students self-select to take the SHSAT and passing is the only requirement for an offer. Variables exhibiting lower or negative correlations with the response variable are **Behavioral**, **Geographic** and **Economic** categories but are still of value as modelling factors.

It is notable that the **Performance** predictors *ELA*_PROF, *MATH_PROF*, *PCT_4S* AND *PCT_4S_ECNDSV* are correlated strongly to *SPHS_OFFERS* which suggests NY State tests as a good candidate to replace the SHSAT. The NY State tests have two principal advantages over the SHSAT, they have near universal participation and they are an annual evaluation starting in grade 3. The implication is that they are more representative of the NYCDOE's student population but still measure academic merit. In fact the NY State tests may evaluate merit to a higher degree because they are part of schools' curricula.

Ranking relative correlations can be used during model building for selecting significant predictors based on correlation to the response variable while removing variables exhibiting signficant collinearity among the predictors.

Table 9: Correlation against Response Variable

	SPHS_FEEDER_M1	SPHS_FEEDER_M2
SPHS_TESTERS	0.691	0.453
ELA_PROF	0.604	0.266
PCT_4S	0.598	0.138
MATH_PROF	0.569	0.129
SPHS_APPLICANTS	0.564	0.068
PCT_WHITE	0.534	0.039
INCOME	0.459	-0.090
CLASS_SIZE	0.454	-0.091
PTRATIO	0.441	-0.097
PCT_4S_ECNDSV	0.440	-0.102
DISTRICT	0.181	-0.115
PCT_ATTENDANCE	0.166	-0.135
PCT_SUPPORTIVE	0.118	-0.137
ZIPCODE	0.108	-0.138
PCT_COLLABORATIVE	0.098	-0.171
PCT_RIGOROUS	0.098	-0.232
PCT_EFFECTIVE	0.088	-0.266
PCT_TRUST	0.080	-0.280
PCT_FEMALE	0.014	-0.300
PCT_4S_UNDRRP	-0.099	-0.306
PCT_ELL	-0.133	-0.351
LONGITUDE	-0.136	-0.393
PCT_FAMILY_TIES	-0.184	-0.470
LATITUDE	-0.203	-0.472
PCT_HISPANIC	-0.275	-0.481
PCT_ABSENCES	-0.375	-0.517
PCT_BLACK	-0.427	-0.566
ECONOMIC_NEED_INDX	-0.554	-0.603

Conclusion of Data Exploration

Data exploration identified 8 predictors as candidates for either imputation or removal from the data set. Missing data was addressed by removing 21 of 591 rows (3.56%) and imputing data for the 3 variables that remained with missing values. In addition, several **Performance** and **Demographic** predictors were discovered that highly correlate to *SPHS_OFFERS*. Notable among the **Demographic** predictors were those describing a school's racial composition which appears to be a key determinant in SPHS underrepresentation.

Underrepresentation in particular is shown by *PCT_BLACK* and *PCT_HISPANIC* negatively correlating with *SPHS_OFFERS* while *PCT_ASIAN* and *PCT_WHITE* positively correlate and strongly so. Since *SPHS_APPLICANTS* are self-selecting and underrepresented students in non-feeder schools do score highly on the NY State tests, the implication the NY State tests are a viable alternative to the SHSAT as the determinant of SPHS. This will be explored further in during the model building phase.

Finally, *COMMUNITY_SCHOOL* was eliminated as a predictor variable based on it lacking enough data variation to inform a model.

Part 2: Data Preparation

Data Preparation efforts focused on variable relationships and considered the possibility of transforming one or more of the predictor variables with skewed distributions. It was decided not to apply any such transforms prior to model building since normal distributions aren't necessarily required for logistical regression modeling. Transforms can be applied if the marginal model plots for a logistic regression model show evidence of deviance between the modeled data and the actual data, but aren't required prior to model building.

Step 1: Imputing Data

During data exploration, work showed evidence of 570 out of 591 school records with complete data sets that could be used for modelling. Within the remaining records three variables: *INCOME*, *PTRATIO* AND *CLASS_SIZE* had missing data that was imputed using GLM modeling. Use of the standard predict function to deduce the imputed values was chosen to anomolies of overconcentrated values in the resulting distributions.

Missing values were imputed for *INCOME*, *CLASS_SIZE*, and *PTRATIO* variables using a linear regression approach recommended by Faraway (p.201) and Fox (p.611). We are not using the mean or median as a replacement value for imputation since regression yields values that are much more consistent with the actual distribution of the data without introducing bias.

Step 2: Converting Categorical Variables to Factors

Data Type Analysis reveals that the predictors *DBN*, *DISTRICT*, *BOROUGH*, *ZIP*, *COMMUNITY_SCHOOL* and *SPHS_FEEDER* are all either binary or categorical variables and are transformed to factors. *DBN* is simply the unique identifier of each school and was converted to a factor. *BOROUGH* is a subset of *DBN* coded as one of the 5 NYC Boroughs K Brooklyn, M Manhattan, Q Queens, R Bronx and X Staten Island and converted to a factor. *ZIP* although nominally numeric is actually a non-ranked catagorical varable but converted to a numeric to show data in histogram form. *COMMUNITY_SCHOOL* is a catagorical binary variable but with very little variation in data. It was therefore removed from the dataset. *SPHS_FEEDER* is a binary catagorical variable based upon conversion of *SPHS_OFFERS* and indicates whether a school has greater than or less then 5 SPHS offers. The *SPHS_FEEDER* variable was transformed via simple threshholding where all greater-than-zero values were converted to '1', yielding the following interpretation:

- **SPHS_OFFERS** n = 0: the school has **no** 0-5 SPHS offers.
- **SPHS_OFFERS** n = 1: the schools **has** greater than 5 SPHS offers.

Step 3: Addressing Zero-Inflated skew

The **BEHAVIORAL** and **PERFORMANCE** predictor catagories which represent "*survey responses on academic related attributes of schools*" and "*high scoring test percentages of schools* respectively." The analysis of the variable's boxplots indicate non-responses to survey questions showing up as excessive zeroes. The two parts of the a zero-inflated model are a binary model, usually a logit model to which processes the zero outcome and a non-zero outcome that is associated with and a count model.

Part 3: Model Building

Modelling Binomial Logistic Distributions

Two distinct binary logistic regression models were constructed for purposes of predicting whether or not a school was likely to be an SPHS feeder. The models' performance metrics were subsequently compared against each other to allow for selection of the "best" binary logistic regression model for purposes of making predictions of SPHS feeders for the Evaluation data set. Both models used the data set's **SPHS_FEEDER** attribute as the dependent response variable, while various subsets of the potential predictor variables were used as independent variables. A detailed discussion of the models can be found in the following section.

Modelling Count Distributions

Response Variable is a Count: This indicates that the model selection process favors a distribution that uses discrete positive integers dispersed in a way that fits the TARGET distribution.

Poisson regression: Poisson regression is often used for modeling count data because it uses a no_n-negative distribution that only puts mass at integer values. In addition it favors data that is not over-dispersed since variance is bound by the mean.

Negative binomial regression: Negative binomial regression can be used for over-dispersed count data, that is when the conditional variance exceeds the conditional mean. It can be considered as a generalization of Poisson regression since it has the same mean structure as Poisson regression and it has an extra parameter to model the over-dispersion. If the conditional distribution of the outcome variable is over-dispersed, the confidence intervals for Negative binomial regression are likely to be narrower as compared to those from a Poisson regression.

Quasi-Poisson model: Another way of dealing with over-dispersion is to use the mean regression function and the variance function from the Poisson GLM but to leave the dispersion parameter unrestricted. Thus, dispersion the parameter is not assumed to be fixed at 1 but is estimated from the data. This strategy leads to the same coefficient estimates as the standard Poisson model but inference is adjusted for over-dispersion.

Zero-inflated regression model: Zero-inflated models attempt to account for excess zeros. In other words, two kinds of zeros are thought to exist in the data, "true zeros" and "excess zeros". Zero-inflated models estimate two equations simultaneously, one for the count model and one for the excess zeros. OLS regression - Count outcome variables are sometimes log-transformed and analyzed using OLS regression. Many issues arise with this approach, including loss of data due to undefined values generated by taking the log of zero (which is undefined) and biased estimates. [22]

Model Preparation,

Model 1: Logit Forward Selection + AIC

These two binary logistic regresssion models made use of forward selection to yield the lowest AIC value possible when only statistically significant predictor variables were considered. The forward selection process used the correlation matrix rankings from **Part 1**. The two models both included log-transformqtions of several predictor variables which were chosen based correlation with the **SPHS_FEEDER** response variable. For Model 1 the response variable represents passing the SHSAT whereas Model 2 the response variable represents scoring a 4 on the NY State test.

[22] https://cran.r-project.org/web/packages/pscl/vignettes/countreg.pdf

Table 10: Logit Model Coefficients' Performance

Model 1	Variable	Model 2	Variable
- 30.5446	Intercept	+ 27.8199	Intercept
+ 2.9824	log(SPHS_TESTERS + 1)	- 3.5350	log(SPHS_TESTERS + 1)
+ 5.1919	ELA_PROF	- 1.5259	log(PCT_4S_UNDRRP + 1)
+ 0.8371	log(PCT_WHITE + 1)	- 7.4528	MATH_PROF
+ 0.9603	log(PCT_ELL + 1)		

*Note that log(x + 1) transformations will assign any 0 values that existed or were imputed to 0 in the transformation log(0 + 1) => 0. This transformation changes the skew of log(PCT_4S_UNDRRP + 1) ~ SPHS_OFFERS and the correlation between PCT_4S_UNDRRP ~ SPHS_OFFERS.

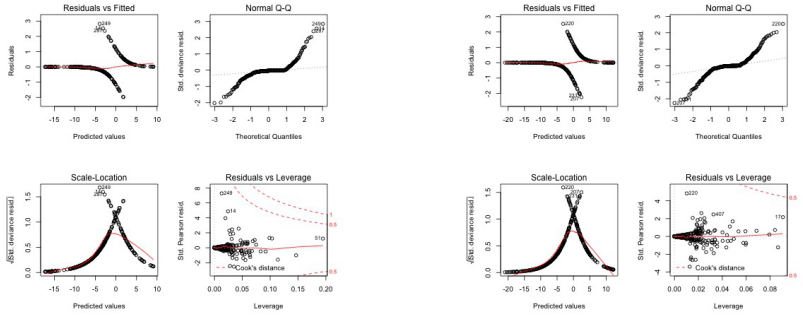
For Level-Log Regression dy=(B1/100)%dx If we increase x by one percent, we expect y to increase by (B1/100) units of y.

Table 11: Logit Model Performance

Model 1	Value	Model 2	Value
Number of Predictors	4	Number of Predictors	3
AIC	120.4	AIC	138.4
Accuracy	0.9448622	Accuracy	0.9398496
Classification Error Rate	0.0551378	Classification Error Rate	0.0601504
Precision	0.8658537	Precision	0.9090909
Sensitivity	0.8658537	Sensitivity	0.9090909
Specificity	0.9652997	Specificity	0.9550562
F1 Score	0.8531469	F1 Score	0.9090909
AUC	0.9156000	AUC	0.9321000

Both models' statistics show a good fit although the error distribution and Q-Q diagrams also show a clear bifurcation in outcomes *0/1* for SPHS_OFFERS. The models also don't show excessive leverage for particular datapoints.

Plot 1: Logit Model Dynamics



Models 2,3,4: Poisson, Quasi-Poisson & Negative Binomial

The Poisson models show under-dispersion (.330) for Model 1 and over-dispersion (1.248) for Model 2 which eliminates them as viable options in the selection process. This is based on signficant dispersion of variance from 1. As an alternive Quasi-Poisson models were fit which do not have a restrictive dispersion requirement

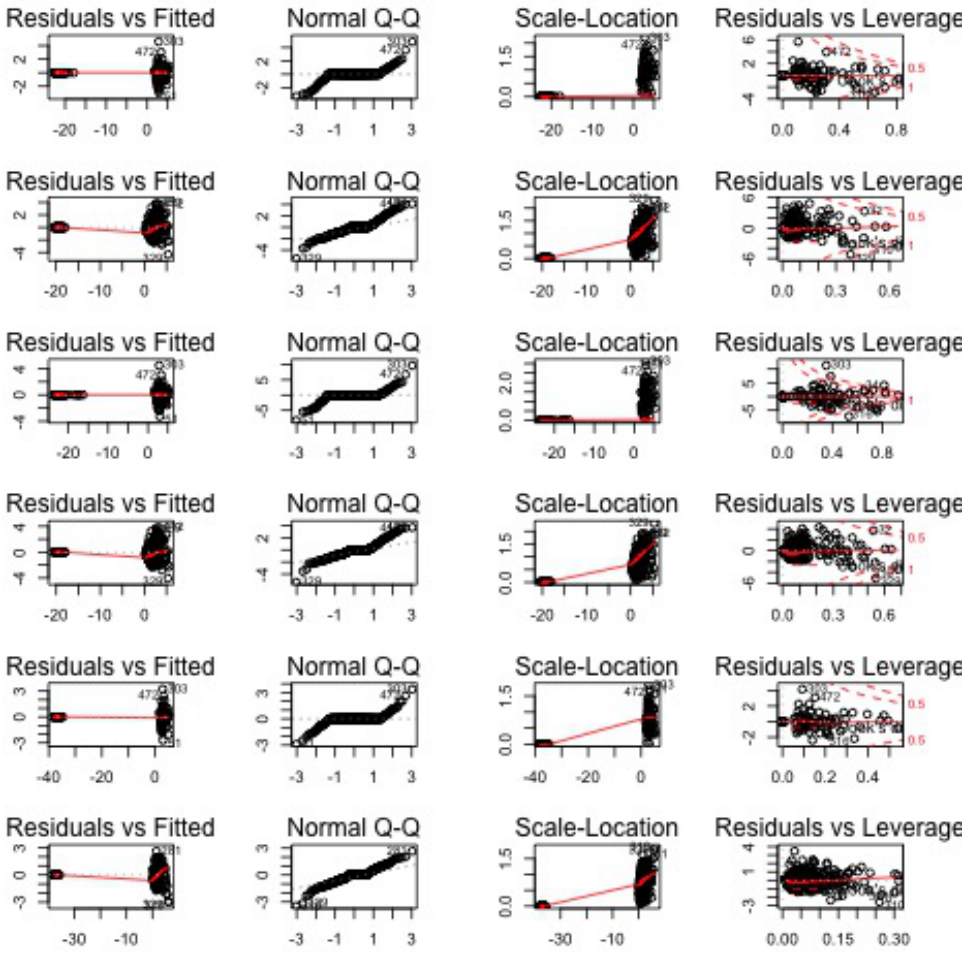
and they yield the best RMSE fit. Lastly Negative Binomial models were fit which also have no dispersion restrictions and have slightly higher RMSE values.

Table 12: Count Model Performance

Model	AIC	BIC	R2	RMSE	Sigma	Score_log	Dispersion
pm.1	565.40	649.17	1.00	3.03	0.57	-0.66	.330
pm.2	1463.99	1559.72	1.00	5.63	1.12	-1.77	1.271
qp.1			1.00	2.99	0.55	-0.65	-
qp.2			1.00	5.59	1.13	-1.77	-
nb.1	562.86	646.63	1.00	3.52	0.48	-0.83	73.30
nb.2	1343.91	1423.69	1.00	21.86	0.74	-1.81	11.85

The model plots show a clear bifurcation in outcomes *0/1* for SPHS_OFFERS. The models don't show excessive leverage for particular datapoints.

Plot 2: Count Model Dynamics



Models 5,6: Zero-Inflated Poisson & Negative Binomial

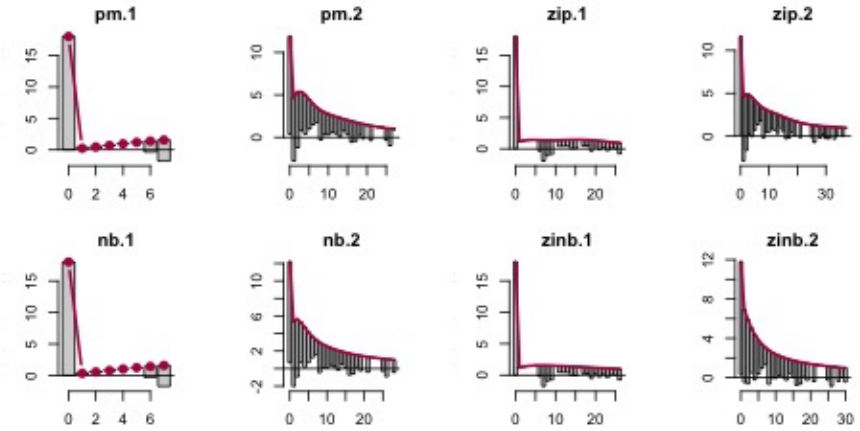
The final set of models test for zero inflation in the response variable *SPHS_OFFERS* which–per Section 2–represent different sources of response data for Model 1 and 2. The Model 1 SPHS offer threshold is the number of students passing the SHSAT whereas Model 2 this threshold is based on passing the NY State test with a score of 4. Zero-Inflation in these models, address observed excesses in the number of schools with zero offers, owing to factors other than test scores, i.e. low test taking percentages at those schools.

Table 13: Zero Inflated Count Model Performance

Model	AIC	R2	R2 (adj.)	RMSE	Score_log	Score_spherical
zip.1	1041.90	1.00	1.00	2.54e+06	-1.25	0.04
zip.2	3521.39	0.99	0.99	16.83	-4.38	0.03
zinb.1	814.54	1.00	1.00	76.77	-1.03	0.04
zinb.2	1801.27	1.00	1.00	43.96	-2.27	0.03

For the zero-inflated Poisson models, the predicted-to-observed ratio for Model 1 is 82% and for Model 2 is 62% indicating both models are likely to be underfitting zeros. For the zero-inflated Negative Binomial models, the predicted-to-observed ratio for Model 1 is 82% and for Model 2 is 101% indicating Model 1 is likely to be underfitting zeroes whereas Model 2 is likely predicting zeroes within expected range.

Plot 3: Count Model Dynamics



Part 4: Model Selection

Model Comparison

Table 14: Model1 Comparison

Model1	Type	AIC	RMSE	Sigma	Score_log	Score_spherical	BIC	BF	p
pm.1	glm	565.40	3.03	0.57	-0.66	0.05	649.17	1.00	
qp.1	glm		2.99	0.55	-0.65	0.05		1.00	0.550
nb.1	negbin	562.86	3.52	0.48	-0.83	0.05	646.63	> 1000	
zip.1	zinfl	1041.90	2.54e+06		-1.25	0.04			
zinb.1	zinfl	814.54	76.77		-1.03	0.03			

Table 15: Model2 Comparison

Model2	Type	AIC	RMSE	Sigma	Score_log	Score_spherical	BIC	BF	p
pm.2	glm	1463.99	5.63	1.12	-1.77	0.04	1559.72	1.00	
qp.2	glm		5.59	1.13	-1.77	0.04		1.00	0.957
nb.2	negbin	1343.91	21.86	0.74	-1.81	0.03	1423.69	> 1000	
zip.2	zinfl	3521.39	16.83		-4.38	0.03			
zinb.2	zinfl	1801.27	43.96		-2.27	0.03			

First, the structure of this output displays each model listed, followed by a table that shows each model's degrees of freedom and loglikelihood. To the right of these values, we are given the degrees of freedom of the test statistic, the value of the test statistic, and then the p-value.

Based on the results, we will reject the null hypothesis at the .05 significance level. The p-values for Models null hypothesis would be rejected for #'s 4 & 5 and we use the least log likelihood values for both models which would be the zero-inflated negative binomial models zinb.1 & zinb.2.

Table 16: Model Log Liklihood Comparison

Model1	#Df	LogLik	Df	Chisq	Pr(>Chisq)	Model2	#Df	LogLik	Df	Chisq	Pr(>Chisq)
1	21	-261.70				1	24	-707.99			
2	31	0	10			2	30	0	6		
3	21	-260.43	-10			3	20	-651.96	-10		
4	21	-499.95	0	479.04	< 2.2e-16 ***	4	12	-1748.69	-8	2193.5	< 2.2e-16 ***
5	14	-393.27	-7	213.36	< 2.2e-16 ***	5	08	-892.63	-4	1712.1	< 2.2e-16 ***

Binomial Logistic Regression Model Predictions

Two binary logistic regression models were constructed for purposes of predicting whether SPSH offers and used the data set's **SPHS_OFFERS** attribute as the dependent response variable, with various subsets of the potential predictor variables were used as independent variables. These models proved to be statistically significant with similar performance statistics although Model 1 exhibited somewhat higher number of SPHS offers predicted as shown in the summary tables shown below.

Table 17, 18: Count Model Confusion Matrix

Training Set

Model1	Observed	-	Model2	Observed	
Predicted	0	1	Predicted	0	1
0	306	11	0	255	12
1	11	71	1	12	120

Evaluation Set

Model1	Observed	-	Model2	Observed	
Predicted	0	1	Predicted	0	1
0	128	4	0	109	6
1	2	37	1	9	47

Zero-Inflated Negative Binomial Model Predictions

Analysis of the 'SPHS_OFFERS' variable revealed that it was zero-inflated and its mean was not nearly equal to its variance, thereby allowing us to quickly rule out the use of either Poisson or standardcount regression models for purposes of predicting likely SPHS offers. Instead, a zero-negative binomial count regression model was pursued.

An initial set of modeling iterations led to the removal of a few statistically insignificant predictors but yielded a model whose *SPHS_OFFERS* predictions were wildly inaccurate, with some predictions exceeding one trillion possible line items. Further investigation revealed that the 'COST' variable was the source of the problem: that variable's very large variance and outliers were causing the negative binomial model to generate wildly inaccurate predictions.

Table 19: Count Model Performance

Model 1	Variable	Model 2	Variable
Count Component			
+ 9.660478	Intercept	- 4.600193	Intercept
- 0.035459	PCT_ELL	+ 0.024771	PCT_ELL
+ 0.007172	PCT_ASIAN	+ 0.016472	PCT_ASIAN
- 0.020130	PCT_BLACK	- 0.044485	PCT_FAMILY_TIES
- 0.078413	PCT_RIGOROUS	+ 3.564050	ELA_PROF
+ 0.054207	PCT_COLLABORATIVE	+ 0.430324	Log(theta)
- 0.062838	PCT_FAMILY_TIES		
+ 0.025919	PCT_4S		
+ 1.811568	Log(theta)		
Zero Inflated Component			
+ 12.68920	Intercept	+ 42.23800	Intercept
- 0.08351	PCT_ASIAN	- 19.19000	ELA_PROF
- 0.03232	PCT_WHITE		
- 2.17053	ELA_PROF		
- 0.14575	CLASS_SIZE		

Generate SPHS Count Predictions

1) Load training and evaluation data set
2) Perform any necessary transforms on data
3) Build selected negative binomial regression model
4) Perform any necessary transforms on copy of eval data
5) use **predict** function to get required SPHS offers
6) Save predicted SPHS_OFFERS data to a file

write.csv files . . .

- file = "/Users/scottkarr/IS628Fall2020/Project/M1_SPHS_COUNT_PREDS.csv"
- file = "/Users/scottkarr/IS628Fall2020/Project/M2_SPHS_COUNT_PREDS.csv"

Table 20: Models Descriptive Statistics

vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se	model
1	171	6.058	20.018	0	0.927	0.000	0	150	150	4.548	23.046	1.531	Model 1 Offers
1	171	5.497	15.657	0	1.401	0.000	0	112	112	4.194	20.047	1.197	Model 1 Predicted Offers
1	171	11.906	31.530	2	4.088	2.965	0	231	231	4.438	22.137	2.411	Model 2 Offers
1	171	13.275	36.655	2	4.891	2.965	0	308	308	5.062	29.700	2.803	Model 2 Predicted Offers

Part 5: Analysis

Discussion and Conclusions

The purpose of this research was to predict both the likelihood of an SPHS admissions and the anticipated number of admissions sent from feeder middle schools. Data related to 570 middle school records from the academic year 2017-18 was used as the basis for constructing and evaluating predictive models. To faciliate predictions, two binary logistic regression model were developed with an accuracy of 94.49% and 93.98% respectively. Model 1 (SHSAT test) predicted 41 (24%) SPHS feeder schools of 171 schools chosen at random from the data set. Model 2 (NY State test) predicted 53 (31%) SPHS feeder schools of 171 schools chosen at random from the data set. Perhaps the most important finding is that for Model 2, SPHS testers negatively correlates to SPHS offers. This tells us that SPHS admissions currently originate with a few feeder middle schools that have high concentrations of offers, thus underrepresenting a more dispersed set of students that would otherwise be admitted on the basis of their NY State test scores if that was the standard.

Prediction of SPHS offers led to the development of a zero inflation negative binomial regression model. A comparison between actual versus predicted offers shows a good fit for both models (Model 1: 20.0, 15.7 Model 2: 11.9, 13.3) means (Model 1: 6.1, 5.5 Model 2: 11.9, 13.3) and errors (Model 1: 1.5, 1.2 Model 2: 2.4, 2.8) for predictions.

- So it is possible–using binary Models 1 (offers based on SHSAT score) & 2 (offers based on NY State Test score)–to develop accurate predictions of SPHS feeder middle schools using factors derived from the Literature Review on high stakes testing is possible.
- For middle schools predicted to be SPHS feeders, the number of SPHS acceptances was able to be accurately predicted using a zero-inflated binomial model which discounts for overinflation of non-feeder schools
- Replacing the SHSAT with the NY State test as the determinate for SPHS acceptance is shown by this study to improve acceptance rates at underrepresented schools. The distinction between the SHSAT and

NY State Test that matters is accessiblity and preparation for the test. It should be noted that both tests measure academic merit but the SHSAT only caters to a select portion of the student population.

Using theoretical constructs as proxies in the binary logistic regression modelling allows for a variety of inferences about these theories as strategies from the research literature. The inferences from theoretical constructs desribed in the **Literature Review** section are discussed below:

1. **Ranked Choice Analysis:** The NYCDOE's widespread use of this algorhthim for matching students has the benefit of enhancing a students motivation to acheive when they are accepted to one of their top choice schools. Since SPHS admissions is determined by one test, schools where students prepare for and take the test weigh heavily on SPHS acceptances. SPHS testers do rank their choices of schools if they are able to gain entry and this competition does motivate the subset of students that are aware of the SHSAT test and who are able to prepare for it outside the standard NYCDOE curricula.
2. **Adverse Selection Analysis:** Significant zero-inflation exists in SPHS offers skewing heavily away from underrepresented demographic groups. Logit Model 1 confirms underrepresentation demographic groups and the "middle school effect" phenomenon that assigns a disproportionate number of SPHS admissions to just a few middle schools. Logit Model 2 however, suggests a more dispersed population of high performing and underrepresented students that do score highly on the NY State test, likely owing to more students' access and preparation. Broadening SPHS offers lessens the adverse selection problem where academic success is focused on only a few outperforming schools.
3. **Underrepresentation:** Addressing underrepresentation in SPHS admissions signals that SPHS admissions is an attainable goal for high acheiving students even if they wouldn't have traditionally considered this option. Logit Model 2 suggests that by using a more accessible test and by preparing students through the NYCDOE curricula, the issue of underrepresentation in underserved schools can be lessened.
4. **Information_Theory:** To the extent that more students are tested, more dispersed schools offer tests and schools prepare their students for the content of these tests then SPHS offers will increasingly reflect the student demographics of the city while still preserving academic merit. This isn't despositive on factors involved in passing the exam but the models in this study shows a relationship.
5. **Alignment:** This study shows determining SPHS admissions using the test more closely aligned with the school curriculum and the ones students have more exposure to taking results in higher admissions rates among a more dispersed set of schools.
6. **Motivational_Theory:** The effects of motivation are inconclusive in this study but to the extent that students make the effort to take the SPHS admissions test is a self-selecting factor in admissions. This observation favors the NY State test because of higher test awareness and alignment with the NYCDOE curricula.
7. **Symbolism:** This is the most subjective of theoretical constructs to measure, however the problem of underrepresentation is clearly present in the SPHS schools and is clearly addressible by reaching more underrepresented and academically capable students.

The conclusions of this study is limited to the middle schools tested in this data set during the 2018 academic year. Additional research would be required to determine whether the same conclusions might apply to other schools and other tests as the basis for admittance. Furthermore, the data provided was limited to high stakes testing at the school level and not the level of individual students. What this study does demonstrate is a clear relationship between underrepresentation students in the SPHS admissions process and both test access and preparation.

Part 6: Bibliography

[22] Achim, Zeileis & Kleiber, Christian & Jackman, Simon. (2008). Regression Models for Count Data in R. Journal of Statistical Software. 27. 1-25. 10.18637/jss.v027.i08. Retrieved from https://cran.r-project.org/web/packages/pscl/vignettes/countreg.pdf

[7] Akerlof, George A., 1970. "The Market for"Lemons": Quality Uncertainty and the Market Mechanism," The Quarterly Journal of Economics, Oxford University Press, vol. 84(3), pages 488-500.

[11] Biggs, John B.; Tang, Catherine Kim Chow (2011). Teaching for quality learning at university: what the student does. Maidenhead: McGraw-Hill. ISBN 9780335242757.

[10,17,18,19] Corcoran, Sean Patrick & E. Baker-Smith,Christine 2018. "Pathways to an Elite Education: Application, Admission, and Matriculation to New York City's Specialized High Schools," Education Finance and Policy, MIT Press, vol. 13(2), pages 256-279, Spring.

[15] Deci, E.L., Koestner, R. & Ryan, R.M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. Psychological Bulletin, 125, 627-668. (10)

[3] Fleeter, Howard B. "The Impact of Local Tax-Based Sharing on School Finance Equity in Ohio; Implementation Issues and Comparative Analysis." Journal of Education Finance 20, no. 3 (1995): 270-301. Accessed January 16, 2021. http://www.jstor.org/stable/40703928.

[9] Hastings, Justine S. & Neilson, Christopher A. & Zimmerman, Seth D., 2012. "The Effect of School Choice on Intrinsic Motivation and Academic Outcomes," NBER Working Papers 18324, National Bureau of Economic Research, Inc.

[9] E. Pfaffelhuber (1972) Learning and Information Theory, International Journal of Neuroscience, 3:2, 83-88, DOI: 10.3109/0020745720914701

[8,20] Przemyslaw Nowaczyk and Joydeep Roy, "Preferences and Outcomes: A Look at New York City's Public High School Choice Process," New York City Independent Budget Office (October 2016).

[2] "Specialized High Schools". NYC Department of Education. 2021. Retrieved January 04, 2021. Retrieved from https://www.schools.nyc.gov/enrollment/enroll-grade-by-grade/specialized-high-schools

[5] "Test Results". NYC Department of Education. 2021. Retrieved January 04, 2021. Retrieved from https://infohub.nyced.org/reports/academics/test-results

[13] Resnick, L. B., Rothman, R., Slattery, J. B., & Vranek, J. L. (2003-2004). Benchmarking and Alignment of Standards and Testing. Educational Assessment, 9(1-2), 1–27. https://doi.org/10.1207/s15326977ea0901&2_1

[12] Smith, Calvin (Novemeber 2008). "Design-focused evaluation". Assessment & Evaluation in Higher Education. 33 (6): 631–645. doi:10.1080/02602930701772762. S2CID 144731064.

[16] Spicer, John. Making Sense of Multivariate Data Analysis: An Intuitive Approach. India: SAGE Publications, 2005, p 135. Retrieved from http://www.sagepub.com/upm-data/5081_Spicer_Chapter_5.pdf

[1,14] Supovitz, J. (2010). Is high-stakes testing working? @Penn GSE A Review of Research, 7(2), 3-8. Retrieved from http://www.gse.upenn.edu/review/feature/supovitz

[4] Supovitz, J.A. & Klein, V. (2003). Mapping a course for improved student learning: How innovative schools systematically use student performance data to guide improvement. Philadelphia, PA: Consortium for Policy Research in Wiliam, D., & Leahy, S. (2006, April). A theoretical foundation for formative assessment. Paper presented at the National Council on Measurement in Education, San Francisco.

[6] Yiping, Lai (2018). Target Schools & Action Recommended for PASSNYC. Retrieved January 04, 2021. "https://www.kaggle.com/laiyipeng/target-schools-action-recommended-for-passnyc

Part 7: Data Dictionary

Variable	Description
DBN*	dbn Borough Number (NYC Department of Education school identifier)
SCHOOL_NAME*	School Name
BOROUGH*	NYC Borough M-Manhattan, K-Brooklyn, Q-Queens, X-Bronx, R-Staten Island
DISTRICT	School District 1-32
ADDRESS*	Middle School Address
ZIPCODE	Zipcode
LATITUDE	Latitude
LONGITUDE	Longitude
ECONOMIC_NEED_INDX	Economic Need Index
COMMUNITY_SCHOOL*	Community School
INCOME	School Income
PCT_ELL	Percent of schools English Language Learners
PCT_ASIAN	Percent of schools Asian Students
PCT_BLACK	Percent of schools Black Students
PCT_HISPANIC	Percent of schools Hispanic Students
PCT_WHITE	Percent of schools White Students
PCT_ATTENDANCE	Percent of schools Attendance
PCT_ABSENCES	Percent of schools Absences
PCT_RIGOROUS	Percent Rigorous
PCT_COLLABORATIVE	Percent Collaborative
PCT_SUPPORTIVE	Percent Supportive
PCT_EFFECTIVE	Percent Effective
PCT_FAMILY_TIES	Percent Family Ties
PCT_TRUST	Percent Trust
PCT_4S	Percent of school's students scoring 4s on the State tests
PCT_4S_UNDDRP	Percent of school's underrepresented students scoring 4s on State tests
PCT_4S_ECNDSV	Percent of school's economically disadvantaged scoring 4s on State tests
SPHS_APPLICANTS	Number of school's students taking SHSAT (SPHS applicants)
ELA_PROF	School's average ELA school's proficiency
MATH_PROF	School's average MATH school's proficiency
CLASS_SIZE	School's class size
PTRATIO	School's pupil-teacher ratio
PCT_FEMALE	Percent Female
SPHS_TESTERS	Number of school's SPHS testers
SPHS_OFFERS**	Number of school's SPHS offers
SPHS_FEEDER**	Is middle school an SPHS feeder school (sends at least 5 students)

* variables removed due to low predictive value

** response variables