Analysis of Average School SAT Scores in New-York City

Ryad Taleb

1 Introduction

This paper will explore factors related to average SAT scores in New-York City. A total of 435 schools across all 5 Boroughs will be considered along with their average SAT scores for the 2014-2015 school year. The dataset includes variables such as enrollment and ethnic percentages, and our goal is to determine which of these factors are the best predictors of average school SAT scores. It will include descriptive statistics of the dataset used, a literary review of similar studies, the proposed methods, and a results section.

1.1 Descriptive Statistics

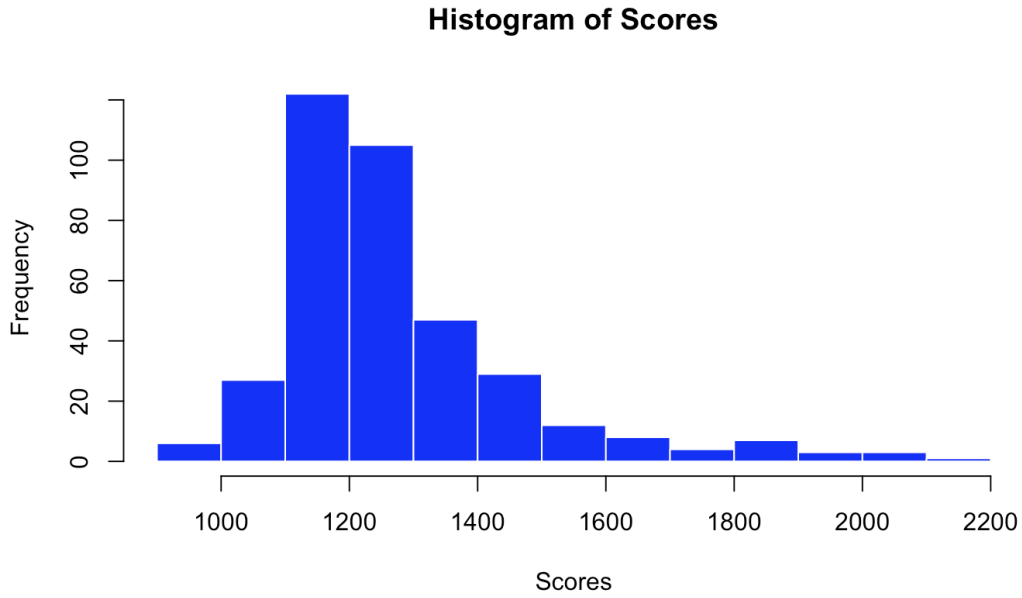


Figure 1: Histogram plot of average school SAT scores. There is a right skew

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

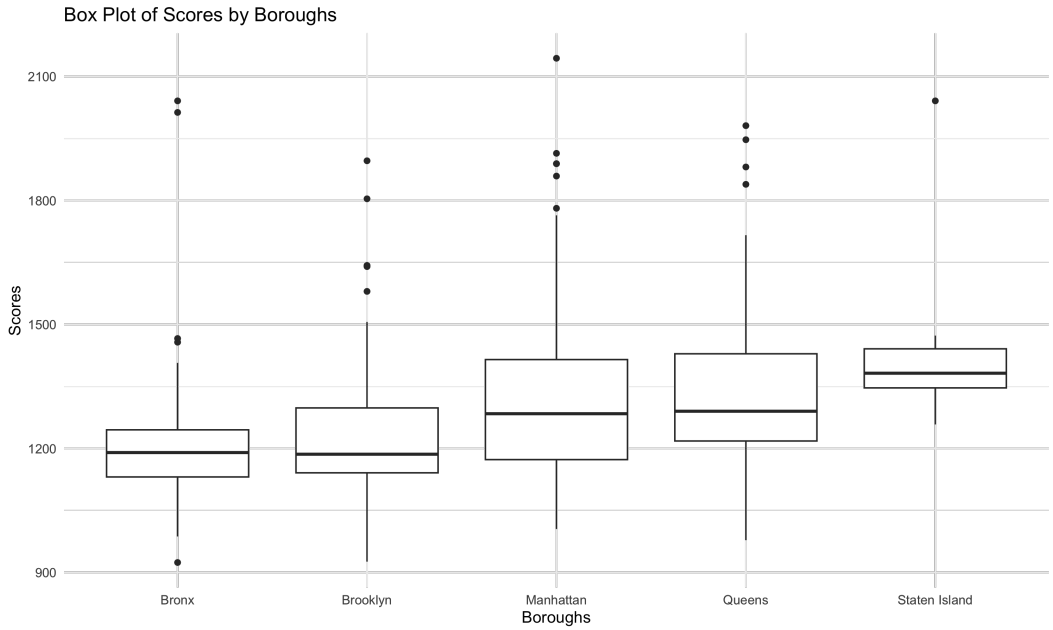
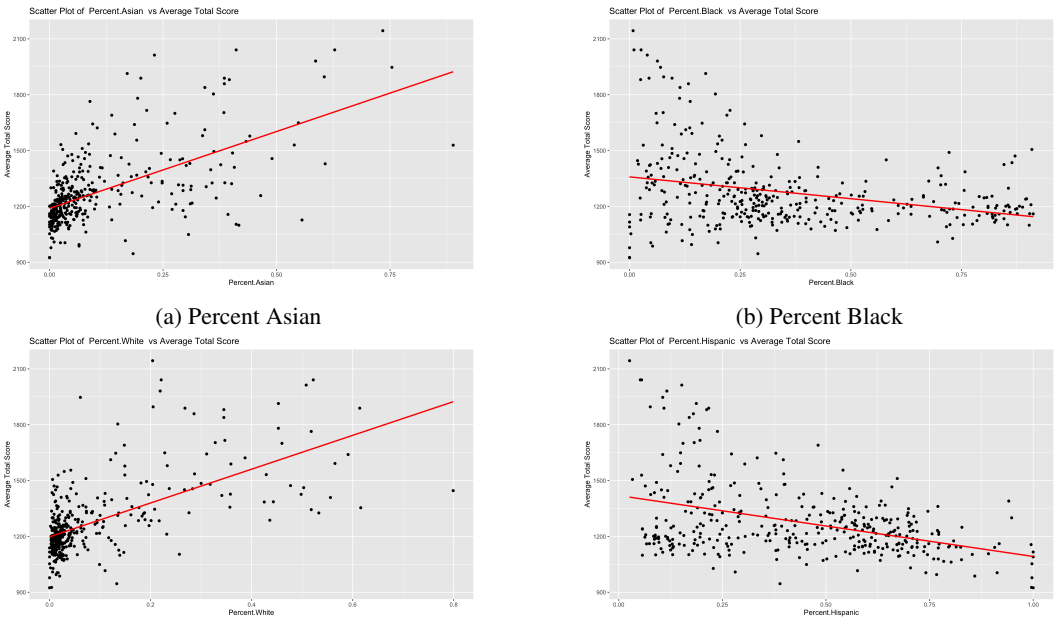


Figure 2: Average School SAT scores by Borough



(a) Percent Asian

(b) Percent Black

(c) Percent White

(d) Percent Hispanic

Figure 3: Average School SAT scores plotted Against ethnic percentages

1.2 Data Cleaning / Modification

Several changes were made to the dataset before using it for analysis. Schools' start and end times were given in hh:mm AM/PM format in 2 separate columns. School duration was then calculated and the 2 previous columns were dropped. Any row (school) with NA values was also removed. Total SAT score had to be calculated and made into its own column. (individual Reading/Writing/Math scores were originally given). Some column data-types had to be changed. The categorical variable of Borough location was dummy coded with Staten-Island being the excluded variable to avoid singularity.

The variables tested are then: School duration, School Enrollment, School Ethnic Percentages (Black, White, Asian, Hispanic), and Borough (Brooklyn, Queens, Manhattan, Staten Island, Bronx).

2 Literature Review

A study conducted by the NYC Data Science Academy analyzed New York Public School SAT scores over the same year (2014-2015). They used Multiple Linear Regression and Random Forest models.

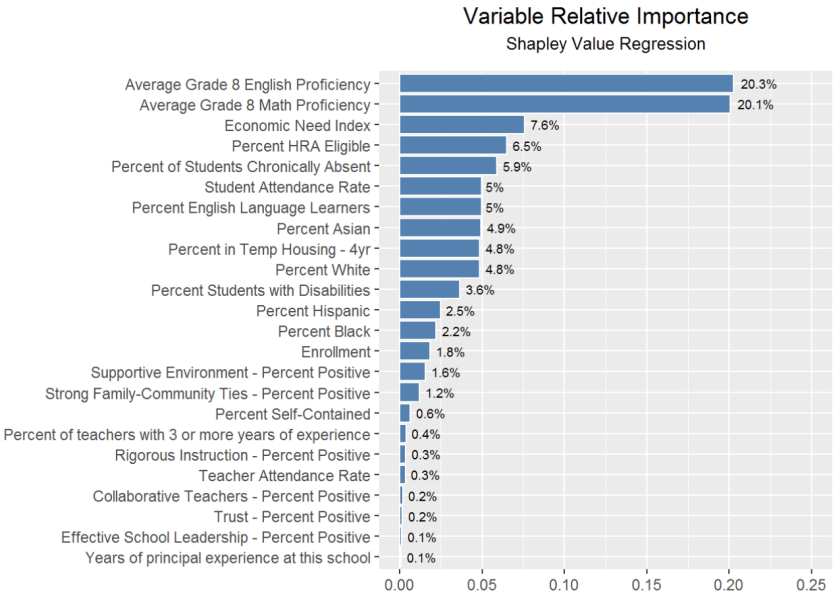


Figure 4: Graph from NY Data Science Academy showing Relative Importance of variables. The data set they used contained more variables, but we can see variables that we analysed such as the ethnic percentages and enrollment.

Another study by the Brookings Institution analyzed the racial disparities in SAT scores. They found average Math SAT scores were notably lower for Black and Hispanic students than White and Asian students. They mention socioeconomic factors and educational access as potential reasons.

3 Proposed Method

Due to the apparent linear dependence of SAT score on the variables, we will perform linear regression with Bayesian Estimation. To then determine the most important factors we will perform Bayesian Model Selection.

3.1 Bayesian Estimation

We will use the normal linear regression model:

$\epsilon_1, \ldots, \epsilon_n \sim$ i.i.d. $\mathcal{N}(0, \sigma^2)$
$Y_i = \beta^T x_i + \epsilon_i$

$$y \mid X, \beta, \sigma^2 \sim \mathcal{N}(X\beta, \sigma^2 I)$$

Our likelihood is then:

$p(y_1, \ldots, y_n \mid x_1, \ldots, x_n, \beta, \sigma^2) = \prod_{i=1}^{n} p(y_i \mid x_i, \beta, \sigma^2)$
$= \left(\frac{1}{(2\pi\sigma^2)^{n/2}}\right) \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta^T x_i)^2\right\}.$

Or in matrix notation:

$$p(y \mid X, \beta, \sigma^2) \propto \exp\left\{-\frac{1}{2\sigma^2}[y^T y - 2\beta^T X^T y + \beta^T X^T X \beta]\right\}.$$

Our prior for beta is also multivariate normal:

$\beta \sim \mathcal{N}(\beta_0, \Sigma_0)$

Then:
$p(\beta \mid y, X, \sigma^2) \propto p(y \mid X, \beta, \sigma^2) \times p(\beta)$
$\propto \exp\left\{\beta^T\left(\Sigma_0^{-1}\beta_0 + \frac{X^T y}{\sigma^2}\right) - \frac{1}{2}\beta^T\left(\Sigma_0^{-1} + \frac{X^T X}{\sigma^2}\right)\beta\right\}.$

We can recognize this as multivariate normal with:

$\text{Var}[\beta \mid y, X, \sigma^2] = \left(\Sigma_0^{-1} + \frac{X^T X}{\sigma^2}\right)^{-1}$
$E[\beta \mid y, X, \sigma^2] = \left(\Sigma_0^{-1} + \frac{X^T X}{\sigma^2}\right)^{-1}\left(\Sigma_0^{-1}\beta_0 + \frac{X^T y}{\sigma^2}\right)$

The semi-conjugate prior for σ^2 is an inverse gamma distribution:

let $\gamma = 1/\sigma^2$, and
$\gamma \sim \text{Gamma}\left(\frac{\nu_0}{2}, \frac{\nu_0\sigma_0^2}{2}\right)$
the posterior for σ^2 is :

$\sigma^2 \mid y, X, \beta \sim \text{Inverse-Gamma}\left(\frac{\nu_0+n}{2}, \frac{\nu_0\sigma_0^2 + \text{SSR}(\beta)}{2}\right)$

where $\text{SSR}(\beta) = (y - X\beta)^T(y - X\beta)$

We then perform Gibb's Sampling as follows:

Given current values $\{\beta^{(s)}, \sigma^{2(s)}\}$:

1. Updating β:

a) Compute $V = \text{Var}[\beta \mid y, X, \sigma^{2(s)}]$ and $m = E[\beta \mid y, X, \sigma^{2(s)}]$.

b) Sample $\beta^{(s+1)} \sim \mathcal{N}(m, V)$.

2. Updating σ^2:

a) Compute $\text{SSR}(\beta^{(s+1)})$ as $(y - X\beta^{(s+1)})^T(y - X\beta^{(s+1)})$.

b) Sample $\sigma^{2(s+1)} \sim \text{Inverse-Gamma}\left(\frac{\nu_0+n}{2}, \frac{\nu_0\sigma_0^2 + \text{SSR}(\beta^{(s+1)})}{2}\right)$.

3.2 Bayesian Model Selection

Let $\beta_j = z_j \times b_j, \quad$ where $z_j \in \{0, 1\}$. (The $z_j's$ act as on-off switches for the factors)

$y_i = z_1 b_1 x_{i,1} + \cdots + z_p b_p x_{i,p} + \epsilon_i$

We need to obtain a posterior distribution for our regression models:

$$p(\mathbf{z} \mid y, X) = \frac{p(\mathbf{z})p(y \mid X, \mathbf{z})}{\sum_{\tilde{\mathbf{z}}} p(\tilde{\mathbf{z}})p(y \mid X, \tilde{\mathbf{z}})}$$

However the denominator is too large to compute. Instead we will consider the ratio of 2 model probabilities:

$$\frac{p(\mathbf{z_a} \mid y, X)}{p(\mathbf{z_b} \mid y, X)} = \frac{p(\mathbf{z_a})}{p(\mathbf{z_b})} \times \frac{p(y \mid X, \mathbf{z_a})}{p(y \mid X, \mathbf{z_b})}$$

Posterior Odds = Prior Odds * Bayes Factor

We need to calculate Bayes Factor:

$$p(y \mid X, z) = \int\int p(y, \beta, \sigma^2 \mid X, z)\, d\beta\, d\sigma^2 = \int\int p(y \mid \beta, X) p(\beta \mid X, z, \sigma^2) p(\sigma^2)\, d\beta\, d\sigma^2$$

Which gives us:

$$p(y \mid X, z) = \pi^{-n/2} \frac{\Gamma\left(\frac{\nu_0+n}{2}\right)}{\Gamma\left(\frac{\nu_0}{2}\right)} (1+g)^{-p_z/2} \frac{(\nu_0\sigma_0^2)^{\nu_0/2}}{(\nu_0\sigma_0^2 + \text{SSR}_z)^{(\nu_0+n)/2}}$$

Where:

$$p_z = \sum_{i=1}^{p} z_j$$

and

$$\text{SSR}_z^g = y^T \left(I - \frac{g}{g+1} X_z (X_z^T X_z)^{-1} X_z^T \right) y.$$

Bayes Factor is then:

$$\frac{p(y \mid X, z_a)}{p(y \mid X, z_b)} = (1+n)^{\frac{p_{z_b} - p_{z_a}}{2}} \left(\frac{s_{z_a}^2}{s_{z_b}^2} \right)^{\frac{1}{2}} \times \left(\frac{s_{z_b}^2 + \text{SSR}_{z_b}^g}{s_{z_a}^2 + \text{SSR}_{z_a}^g} \right)^{\frac{n+1}{2}}$$

Let $\mathbf{z}_{-j}$ be the model $\mathbf{z}$ without factor j. We then calculate the conditional odds o_j that z_j is 1:

$$o_j = \frac{p(y \mid X, z_{-j}, z_j = 1)}{p(y \mid X, z_{-j}, z_j = 0)} \times \frac{\Pr(z_j = 1)}{\Pr(z_j = 0)}$$

And

$$\Pr(z_j = 1 \mid y, X, z_{-j}) = \frac{o_j}{1 + o_j}$$

We then construct a Gibb's Sampler:

Given $z^{(s)}$ generate $\{z^{(s+1)}, \sigma^{2(s+1)}, \beta^{(s+1)}\}$ as follows:

1. Set $z = z^{(s)}$.
2. For j in $\{1, \ldots, p\}$ in random order, replace z_j with a sample from $p(z_j \mid z_{-j}, y, X)$.
3. Set $z^{(s+1)} = z$.
4. Sample $\sigma^{2(s+1)} \sim p(\sigma^2 \mid z^{(s+1)}, y, X)$.
5. Sample $\beta^{(s+1)} \sim p(\beta \mid z^{(s+1)}, \sigma^{2(s+1)}, y, X)$.

3.3 Implementation

For Bayesian estimation our prior for beta is $\beta \sim MVN(0, \Sigma_0)$, where Σ_0 is a diagonal matrix with 100 for all diagonal elements. (meant to reflect a weak prior, our data is standardized). It was then run for 5000 MCMC steps.

Model Selection was run for 10,000 MCMC steps.

4 Data Analysis and Results

4.1 Bayesian Estimation Results

We compared the Bayesian estimation model to least squares and they were almost identical:

```
Call:
lm(formula = y ~ ., data = df_OLS)

Residuals:
    Min      1Q  Median      3Q     Max
-377.87  -60.79   -0.58   56.23  416.30

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           1275.348      6.028 211.579  < 2e-16 ***
Student.Enrollment      16.893      7.285   2.319 0.020952 *
Percent.White         -155.682     53.550  -2.907 0.003870 **
Percent.Black         -451.347     98.253  -4.594 6.01e-06 ***
Percent.Hispanic      -475.602     92.950  -5.117 5.05e-07 ***
Percent.Asian         -186.500     56.276  -3.314 0.001012 **
Duration                11.485      6.133   1.873 0.061925 .
BoroughBronx           -11.768      7.857  -1.498 0.135063
BoroughBrooklyn        -46.622      8.764  -5.320 1.82e-07 ***
BoroughQueens          -38.673      7.795  -4.961 1.08e-06 ***
`BoroughStaten Island`  -27.821      7.343  -3.789 0.000177 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 116.6 on 363 degrees of freedom
Multiple R-squared:  0.6517,    Adjusted R-squared:  0.6421
F-statistic: 67.93 on 10 and 363 DF,  p-value: < 2.2e-16
```

Figure 5: OLS model summary

Description: df [11 × 3]

	2.5% <dbl>	97.5% <dbl>	means <dbl>
Intercept	1263.3704654	1287.510408	1275.468
Enroll	2.4597970	31.009119	16.626
P.white	-258.1132511	-53.222830	-155.381
P.Black	-641.3748629	-263.104522	-450.917
P.Hisp	-653.8033025	-297.095517	-475.002
P.Asian	-294.5209321	-78.065471	-186.068
Duration	-0.3670514	23.970188	11.527
BoroughBronx	-27.3237417	3.421133	-11.685
BoroughBrooklyn	-63.4871561	-29.511997	-46.471
BoroughQueens	-54.2241542	-23.749506	-38.609

	2.5% <dbl>	97.5% <dbl>	means <dbl>
oroughStatenIsland	-42.6033251	-13.367183	-27.854

Figure 6: Bayesian Estimation model

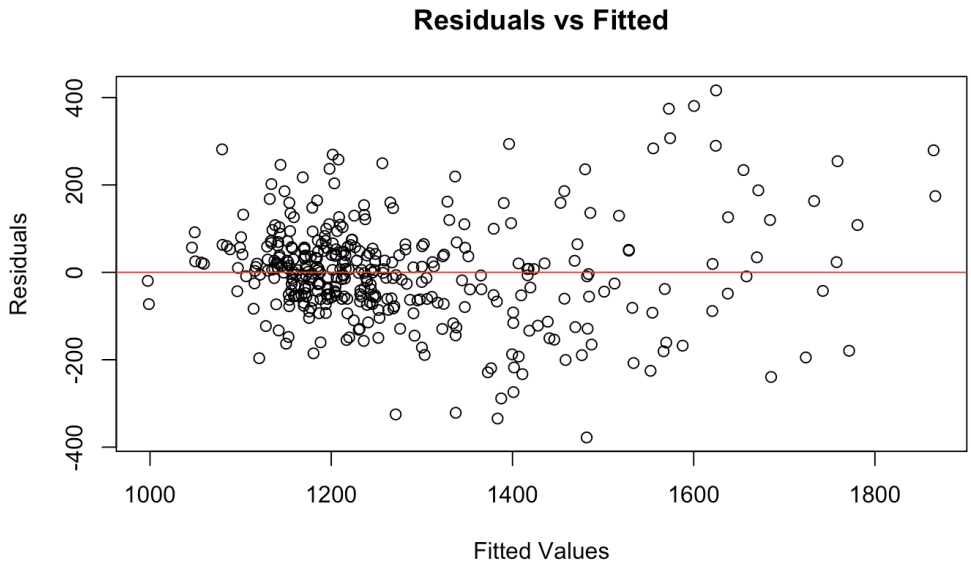


Figure 7: Residuals of the OLS model vs the fitted values

To account for the non-constant variance, a log transform was applied on y, however, there was still significant heteroscedasticity according to the Breusch-Pagan test:

OLS model: BP = 118.15, df = 10, p-value $<$ 2.2e-16

Log(y) model: BP = 69.445, df = 10, p-value = 5.673e-11

4.2 Model Selection Results

:

Description: df [11 × 2]

Variable <chr>	Inclusion Probability <chr>
Intercept	1
Enroll	0.1571
P.white	0.205
P.Black	0.999
P.Hisp	0.9999
P.Asian	0.2634
Duration	0.1647
BoroughBronx	0.1047
BoroughBrooklyn	0.9982
BoroughQueens	0.9949

Variable <chr>	Inclusion Probability <chr>
BoroughStatenIsland	0.863

Figure 8: variables and their inclusion probabilities

Percent White, Percent Asian, Duration, and Bronx identifier have low inclusion probabilities.

An 80-20 train-test split was then performed to test the performance of the full model (OLS) and a model only with variables above 0.85 inclusion probability (we removed the 5 variables). The results were:

RMSE for full model: 115.00
RMSE for selected model: 116.47

So with almost half the model complexity our RMSE only increased by 0.012

5 Conclusion

This study used Bayesian data analysis methods to explore the factors of average SAT scores among New York City schools. No significant difference was found between Ordinary Least Squares and Bayesian Estimation. Weak predictors were then removed with Bayesian Model Selection, and analysis between the full model and the reduced model showed a negligible increase in RMSE. (a simpler model can nearly match the performance of a more complex one). Future research might explore additional variables or alternative statistical methods to further refine these insights.

6 References

NYC Data Science. (2016, Oct 6). Data study on NYC public schools SAT scores. NYC Data Science Academy. https://nycdatascience.com/blog/student-works/data-study-on-nyc-public-schools-sat-scores/

Brookings Institution. (2017, Feb 1). Race gaps in SAT scores highlight inequality and hinder upward mobility. Brookings. https://www.brookings.edu/articles/race-gaps-in-sat-scores-highlight-inequality-and-hinder-upward-mobility/

Hoff, P. D. (2009). A first course in Bayesian statistical methods. Springer Science + Business Media. https://doi.org/10.1007/978-0-387-92407-6