How would student enrollment, borough, and median household income potentially correlate with SAT performance?

By Yuxuan Yang

Introduction

In this paper, we will use the dataset Average SAT Scores for NYC Public Schools from NYC Open Data to investigate the mean SAT scores in each NYC public school, and how the means of these scores of each school can be correlated with the school's location and student enrollment. In general, I would address these main questions in the paper:

How are the locations(borough) of each school correlated with the school's overall performance in the SAT Exam?

Is the overall performance on SAT related to a school's number of student enrollment?

where the Y variable would be the mean of the total SAT score, and the X variables would be the location (borough) and student enrollment of each school. Since I will also scrape extra data of median household income of different districts at NYC and combine it with my chosen X variables, I will also address a subsequent economic question: **Does median household income correlate with the mean SAT score of a school?**

To have a better understanding of this project, readers need to know the basic background of this paper: SAT is a pencil-and-paper test administered by the College Board with the purpose of measuring a high school student's readiness for college, and provide colleges with one common data point that can be used to compare all applicants.

Literature Review:

Sackett, Kuncel, and several other researchers have investigated the relationship between students' socioeconomic status and their academic performances such as SAT scores. In their article, they reached a key finding that students' general academic achievements are highly correlated with their socioeconomic status (Sackett et al., 2012).

Other researchers such as Dixon-Román also mentioned in their article that richer students tend to achieve high SAT scores more easily because they could pay for tuitions of better-funded schools with better academic support and they could also afford to take the SAT Exam several times (Dixon-Román et al., 2013).

While these authors all focus on the general impact of socioeconomic backgrounds on students' academic performances such as SAT scores and college GPA across America as a whole, in this paper, my analysis differs from the above articles in the way that I will focus on the correlation between economic status and students' mean SAT score in New York City specifically. I will also add insights to the existing literature through joint investigations of student enrollment and median household income and how they could be correlated with students' SAT performances.

Since residents in different regions have varying economic conditions, this can be illustrated by different boroughs, which is our X variable; the variation in the scale of the school indicates potential educational funds and resources of a school, which could also be demonstrated by our X variable, student enrollment.

Overall, I will provide different measures of analysis including summary statistics and visualizations such as boxplot, scatterplot, heat maps, simple linear regression, and multivariate linear regressions in order to add economic insights to the research question "How would student enrollment, borough, and median household income potentially correlate with SAT performance". I will focus on the discussion of how the X variables would correlate with the mean SAT score.

Section Data

The dataset I will use is Average SAT Scores for NYC Public Schools which comes from NYC Open Data; it can also be found on the Kaggle website: https://www.kaggle.com/nycopendata/high-schools. I will also use extra data on median household income information at the zip code level of NYC which comes from the Zip atlas website: http://zipatlas.com/us/ny/new-york/zip-code-comparison/median-household-income.htm.

The Average SAT Scores for NYC Public Schools dataset has observation levels of each public school in NYC. It includes variables such as borough information for each school, zip code for each school, student enrollment for each school, and percent of students taking the SAT in each school.

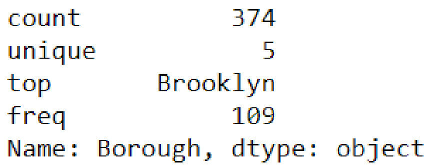
The extra data, on the other hand, includes observations at the zip code level. It includes variables such as median household income for each zip code level in NYC and the location of each zip code (longitude and latitude).

Section Summary Statistics

In Project One, I read the Average SAT Scores for NYC Public Schools data from the local computer and conducted data cleaning on it. I also selected variables of interest, including student enrollment, zip code, borough, percent of students tested, and all the SAT Section Score information, to form a data frame called *df1_SAT* to be used later. What's more, I renamed some variables and added some new variables to *df1_SAT,* including Total SAT Score and Number of students taking the SAT in each school.

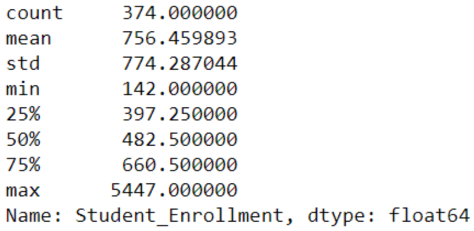
I also included summary statistics for the X and Y variables I'll investigate later. Below are summary statistics for the borough, student enrollment, and mean Total SAT Score of each school.

Borough:



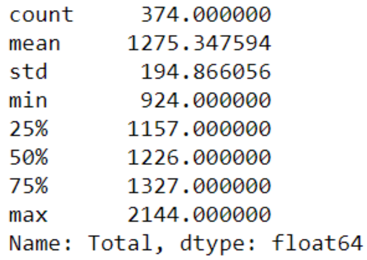
```
count          374
unique           5
top       Brooklyn
freq           109
Name: Borough, dtype: object
```

Student enrollment:



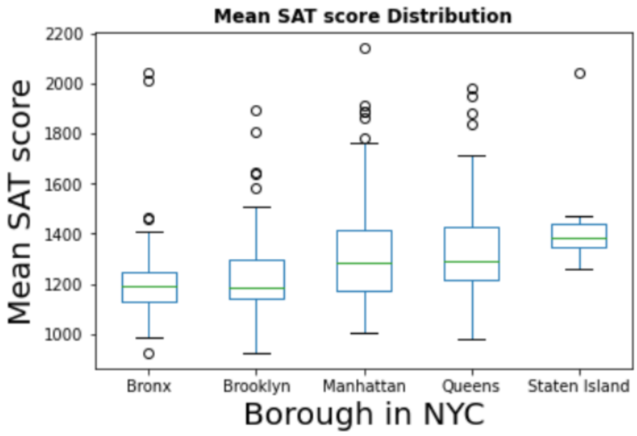
```
count     374.000000
mean      756.459893
std       774.287044
min       142.000000
25%       397.250000
50%       482.500000
75%       660.500000
max      5447.000000
Name: Student_Enrollment, dtype: float64
```

Mean of Total SAT Score:



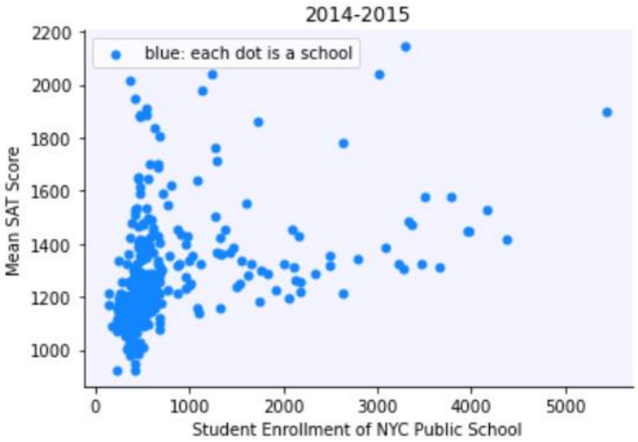
```
count     374.000000
mean     1275.347594
std       194.866056
min       924.000000
25%      1157.000000
50%      1226.000000
75%      1327.000000
max      2144.000000
Name: Total, dtype: float64
```

While in Project Two, I created a boxplot called *Mean SAT score Distribution* to demonstrate the distribution of mean SAT Total Score in each borough:

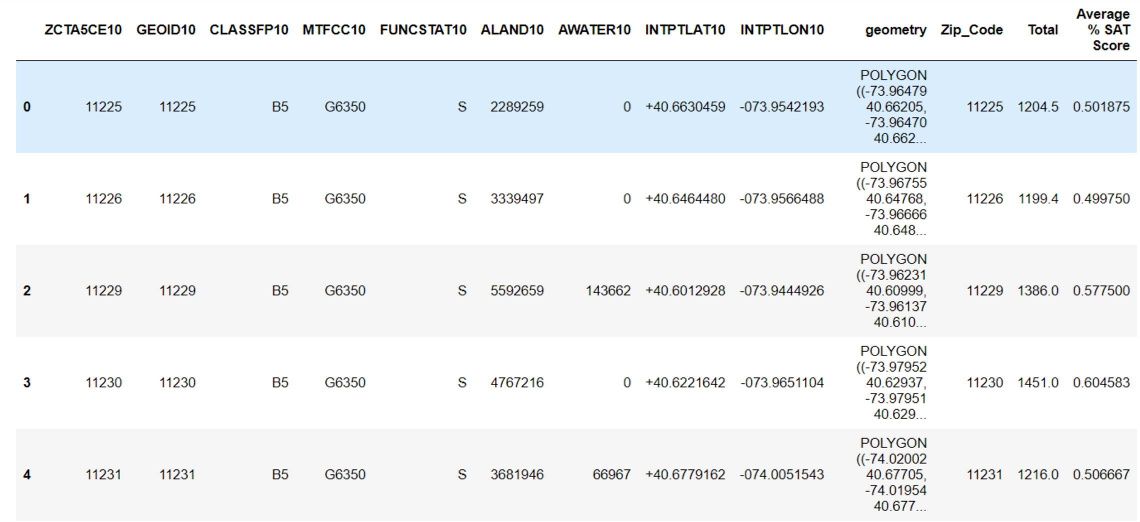


From the above boxplot, we could see the distribution of the mean SAT score of each borough.

I also created a scatterplot that shows student enrollment as the X variable and mean SAT score as the Y variable:

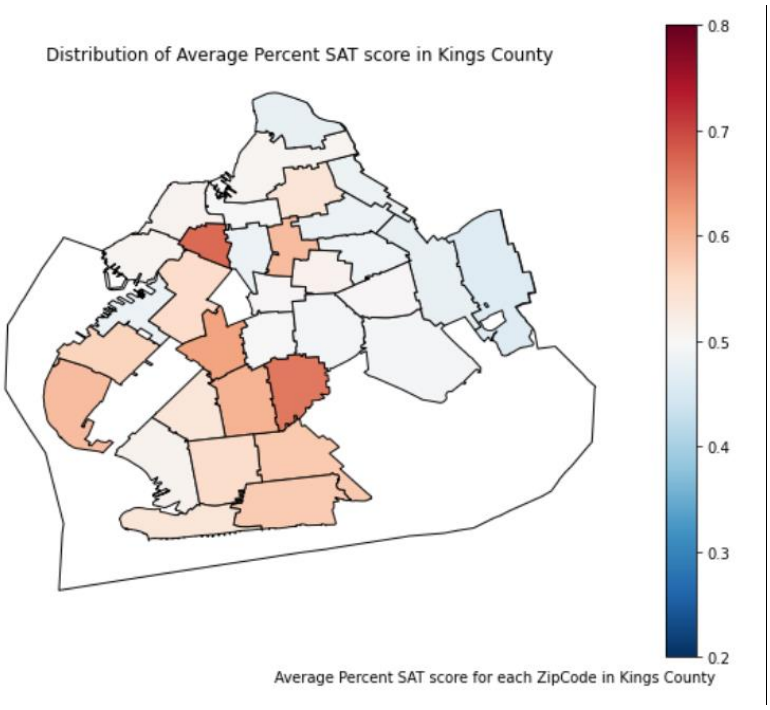


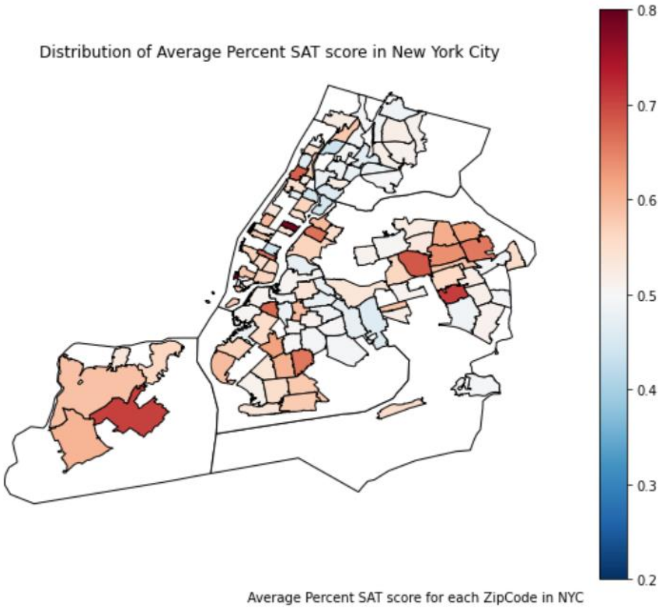
In addition, I imported shapefiles of US county and US zip code and converted US zip code shapefile into a data frame called *us_zip_code_df* to be used later. Then I created a new data frame out of df1_SAT called *mean_each_Zip* to include only the zip code and the mean of Total SAT Score of each school at the zip code level to be merged with *us_zip_code_df* later. Below is a shortcut of the merged data frame, *merge_zipcode_score*:



	ZCTA5CE10	GEOID10	CLASSFP10	MTFCC10	FUNCSTAT10	ALAND10	AWATER10	INTPTLAT10	INTPTLON10	geometry	Zip_Code	Total	Average % SAT Score
0	11225	11225	B5	G6350	S	2289259	0	+40.6630459	-073.9542193	POLYGON ((-73.96479 40.66205, -73.96470 40.662...	11225	1204.5	0.501875
1	11226	11226	B5	G6350	S	3339497	0	+40.6464480	-073.9566488	POLYGON ((-73.96755 40.64768, -73.96666 40.648...	11226	1199.4	0.499750
2	11229	11229	B5	G6350	S	5592659	143662	+40.6012928	-073.9444926	POLYGON ((-73.96231 40.60999, -73.96137 40.610...	11229	1386.0	0.577500
3	11230	11230	B5	G6350	S	4767216	0	+40.6221642	-073.9651104	POLYGON ((-73.97952 40.62937, -73.97951 40.629...	11230	1451.0	0.604583
4	11231	11231	B5	G6350	S	3681946	66967	+40.6779162	-074.0051543	POLYGON ((-74.02002 40.67705, -74.01954 40.677...	11231	1216.0	0.506667

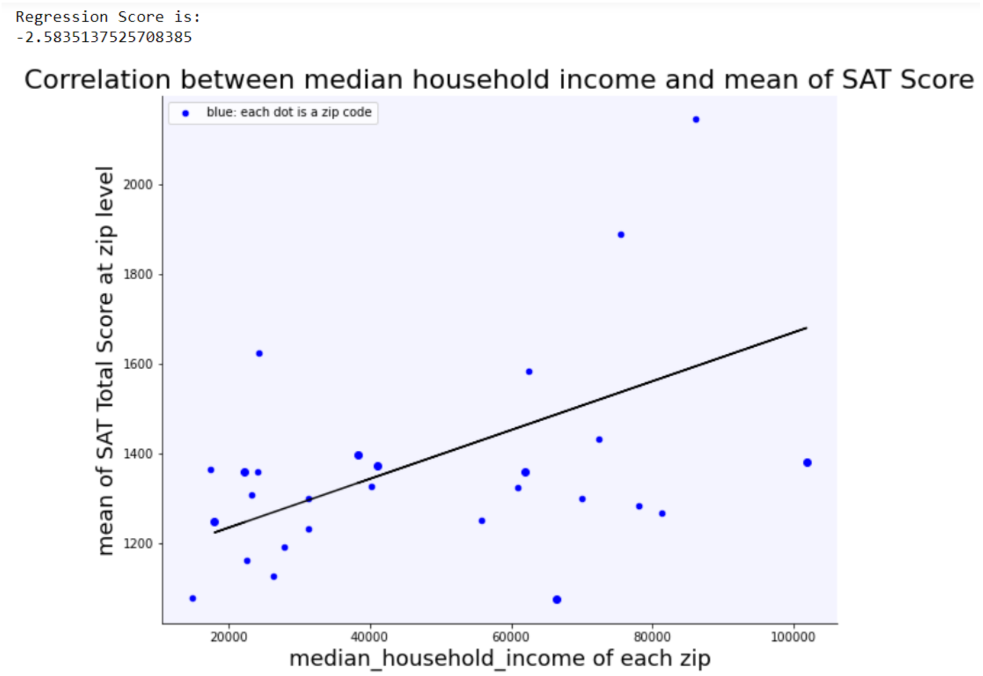
Next, we’ll use this merged data frame to plot two heat maps illustrating the distribution of the mean of SAT Scores at the zip code level, for both Brooklyn Borough and the whole of New York City. Below are heatmaps for both Brooklyn Borough and New York City.


Distribution of Average Percent SAT score in Kings County
0.8
0.7
0.6
0.5
0.4
0.3
0.2
Average Percent SAT score for each ZipCode in Kings County


Distribution of Average Percent SAT score in New York City
0.8
0.7
0.6
0.5
0.4
0.3
0.2
Average Percent SAT score for each ZipCode in NYC

From the above heat maps, we could reach a conclusion that different boroughs in NYC have different mean SAT scores, and Staten Island has the best overall performance in mean SAT scores.

In Project Three, I scraped extra data on median household income from the Zip atlas website to conduct a simple linear regression plot of the correlation between median household income and the mean of total SAT score at each zip code level. I first used web scraping techniques to extract columns of zip code information and median household income at the zip code level. Then I merged these pieces of information with the original *df1_SAT* dataset, and plotted median household income on the X-axis and mean of total SAT score on the Y-axis. Below is the output for this linear regression plot.

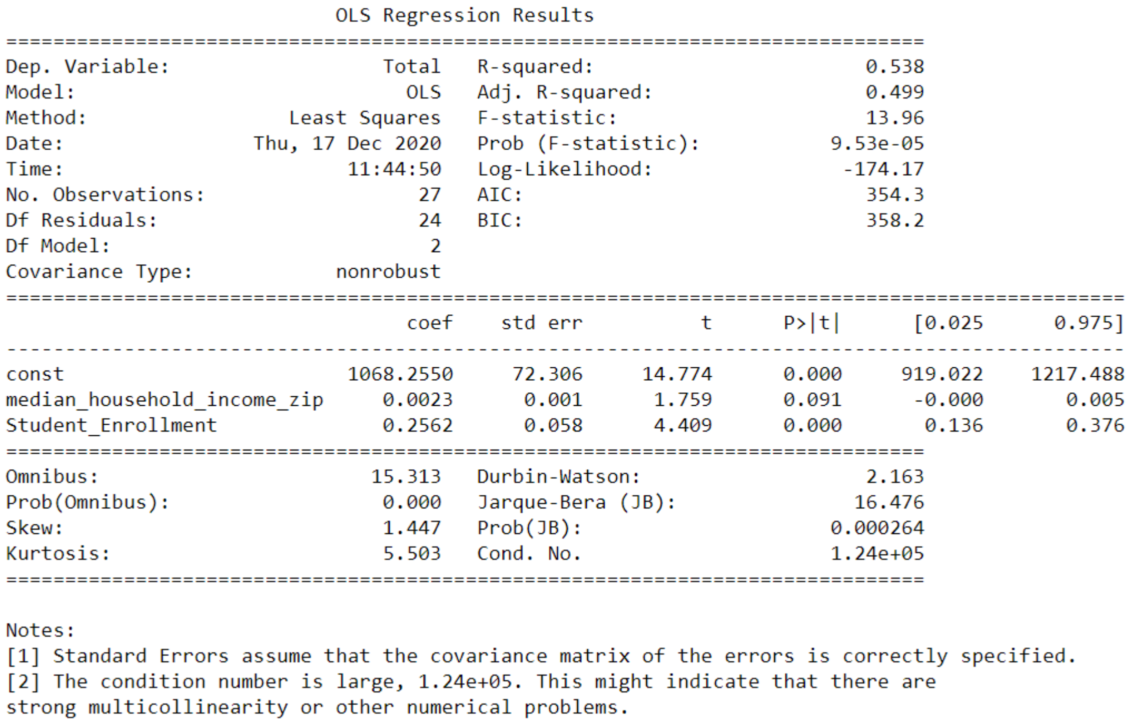


```
Regression Score is:
-2.5835137525708385
```

From the plot and the regression score, we could observe that there exists a relatively moderate linear relationship between the median household income and SAT score of each zip code region. This indicates that the potential X-variable, household income, might actually be correlated with or have influences on the Y variable, mean of SAT Total Score.

Section Results

In the Final Project, I ran four separate multivariate regressions on different variables in order to examine the joint effect of two X variables on my Y variable, the mean of SAT Total Score. For the first regression model, I chose student enrollment and median household income as my explanatory variables. Below is the summary table for the first regression.



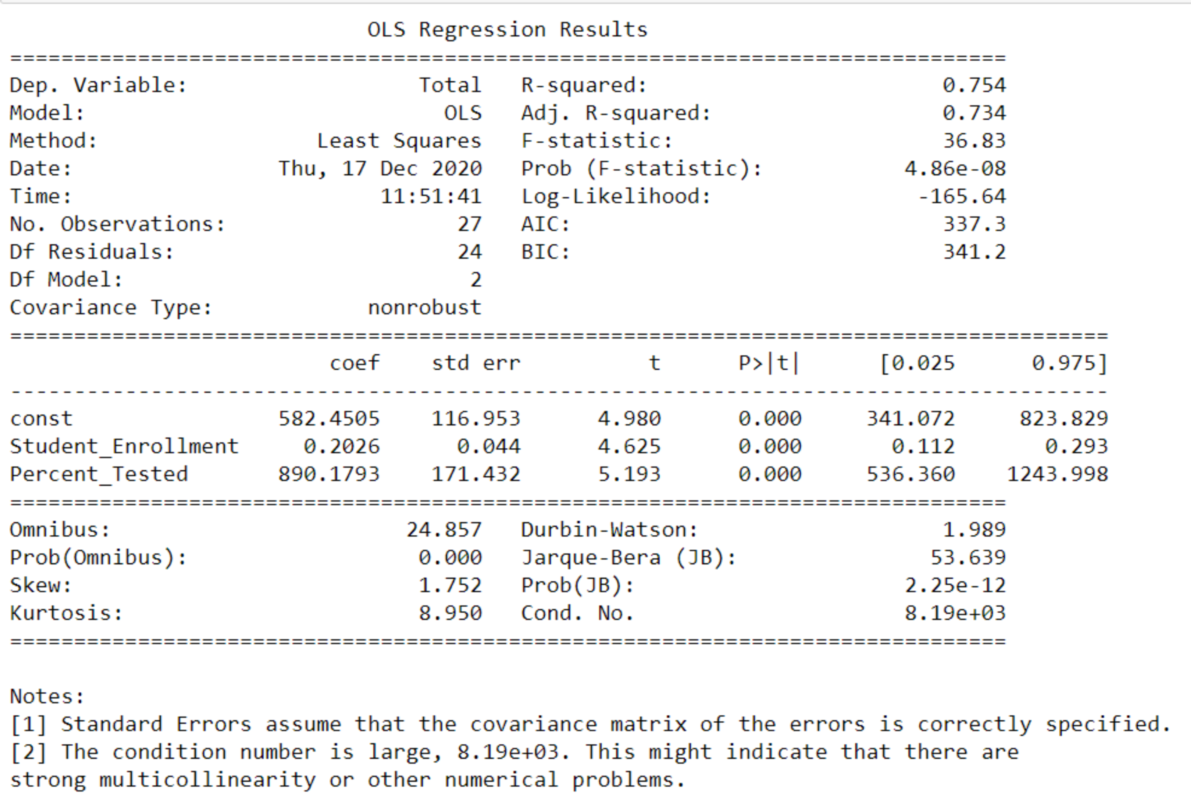
```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:                  Total   R-squared:                       0.538
Model:                            OLS   Adj. R-squared:                  0.499
Method:                 Least Squares   F-statistic:                     13.96
Date:                Thu, 17 Dec 2020   Prob (F-statistic):           9.53e-05
Time:                        11:44:50   Log-Likelihood:                -174.17
No. Observations:                  27   AIC:                             354.3
Df Residuals:                      24   BIC:                             358.2
Df Model:                           2                                         
Covariance Type:            nonrobust                                         
===============================================================================================
                                  coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------------
const                        1068.2550     72.306     14.774      0.000     919.022    1217.488
median_household_income_zip     0.0023      0.001      1.759      0.091      -0.000       0.005
Student_Enrollment              0.2562      0.058      4.409      0.000       0.136       0.376
==============================================================================
Omnibus:                       15.313   Durbin-Watson:                   2.163
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               16.476
Skew:                           1.447   Prob(JB):                     0.000264
Kurtosis:                       5.503   Cond. No.                     1.24e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.24e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

From this table, we could observe the results that for median household income, a P-value of 0.091 is less than 0.1 but greater than 0.05, indicating that the effect of median household income on the mean of Total SAT score is moderately strong (there is a moderate correlation between median household income and mean SAT Score). While for student enrollment, a P-value of 0.000 is significant, indicating a strong correlation between student enrollment and the mean SAT Total Score. An adjusted R-squared of roughly 0.5 indicates a moderate correlation between the joint performance of student enrollment and median household income and mean of SAT Total Score.

An economic explanation for this moderate correlation between student enrollment, median household income, and mean SAT Score would be that more student enrollment and more household income together would indicate that a school seems more attractive to high-income families who value education; both high income and value on education are necessary factors that would provide better academic resources for students and thus lead to better SAT performances.

The second regression I ran is a multivariate regression that combines both student enrollment and percent of students testing SAT with the mean of SAT Total score. Below is the summary table for this regression.



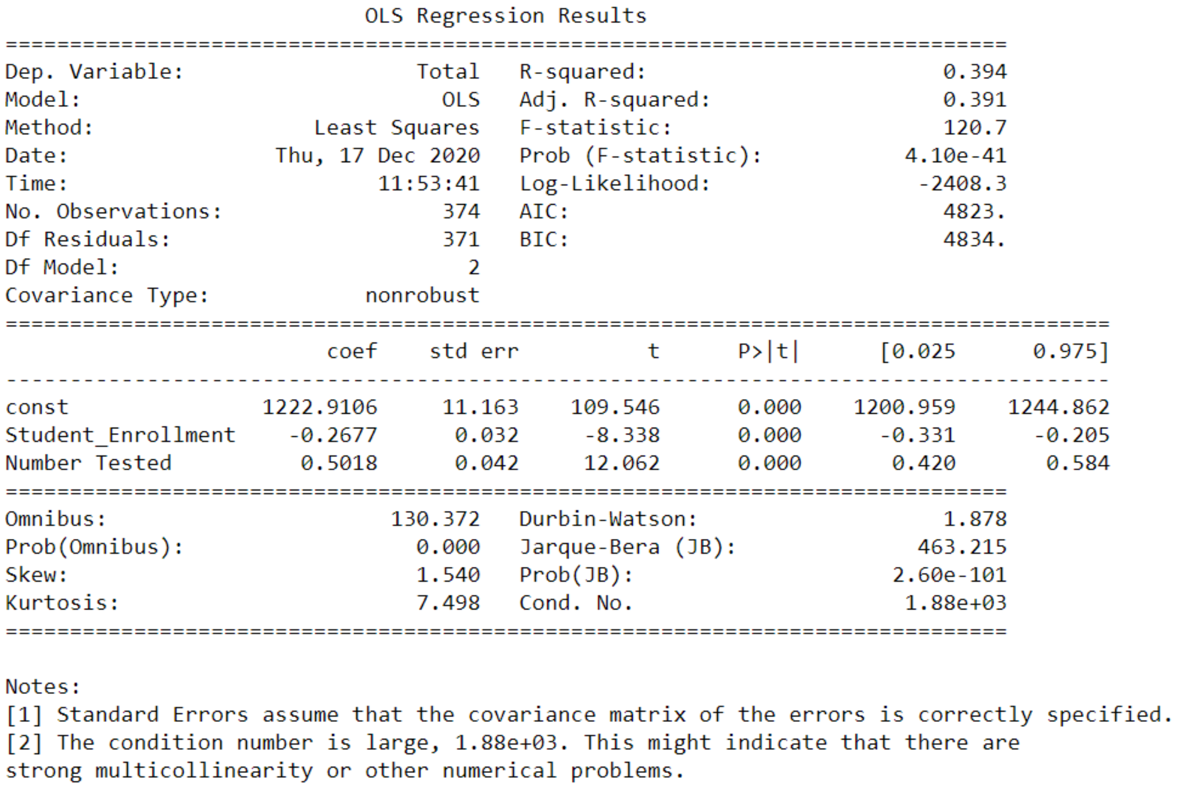
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Total   R-squared:                       0.754
Model:                            OLS   Adj. R-squared:                  0.734
Method:                 Least Squares   F-statistic:                     36.83
Date:                Thu, 17 Dec 2020   Prob (F-statistic):           4.86e-08
Time:                        11:51:41   Log-Likelihood:                -165.64
No. Observations:                  27   AIC:                             337.3
Df Residuals:                      24   BIC:                             341.2
Df Model:                           2
Covariance Type:            nonrobust
======================================================================================
                         coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------
const                582.4505    116.953      4.980      0.000     341.072     823.829
Student_Enrollment     0.2026      0.044      4.625      0.000       0.112       0.293
Percent_Tested       890.1793    171.432      5.193      0.000     536.360    1243.998
==============================================================================
Omnibus:                       24.857   Durbin-Watson:                   1.989
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               53.639
Skew:                           1.752   Prob(JB):                     2.25e-12
Kurtosis:                       8.950   Cond. No.                     8.19e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 8.19e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

From this table, we could observe these results: For student enrollment, a P-value of 0.000 is significant, indicating a strong correlation between it and mean SAT Score. As student enrollment increases by 1, the mean SAT score tends to increase by 0.2. While for percent of students taking the SAT, a P-value of 0 is also significant, indicating a relatively strong effect of percent of students taking SAT on mean SAT Score. An adjusted R-squared value of 0.73 also confirms the strong correlation between these three variables.

The economic intuition behind these results is that as the fraction of students taking the SAT becomes higher and student enrollment becomes larger, it indicates that students at this school value academic performances more and are in need of SAT scores in order to apply for colleges. These students might come from high-income families that value education a lot and could pay for expensive college tuitions, which further implies the relatively high socioeconomic status of their families. High income and economic status would bring students more educational resources that build towards the SAT Exam. Also, schools with more students taking the SAT would be better motivated to fund their teaching facilities and hard wares, which further contributes to success in SAT.

In the third multivariate regression, I changed my X variables into student enrollment and the number of students taking SAT to investigate their joint effect on the mean of SAT Total Score. Below is the summary table for this regression:



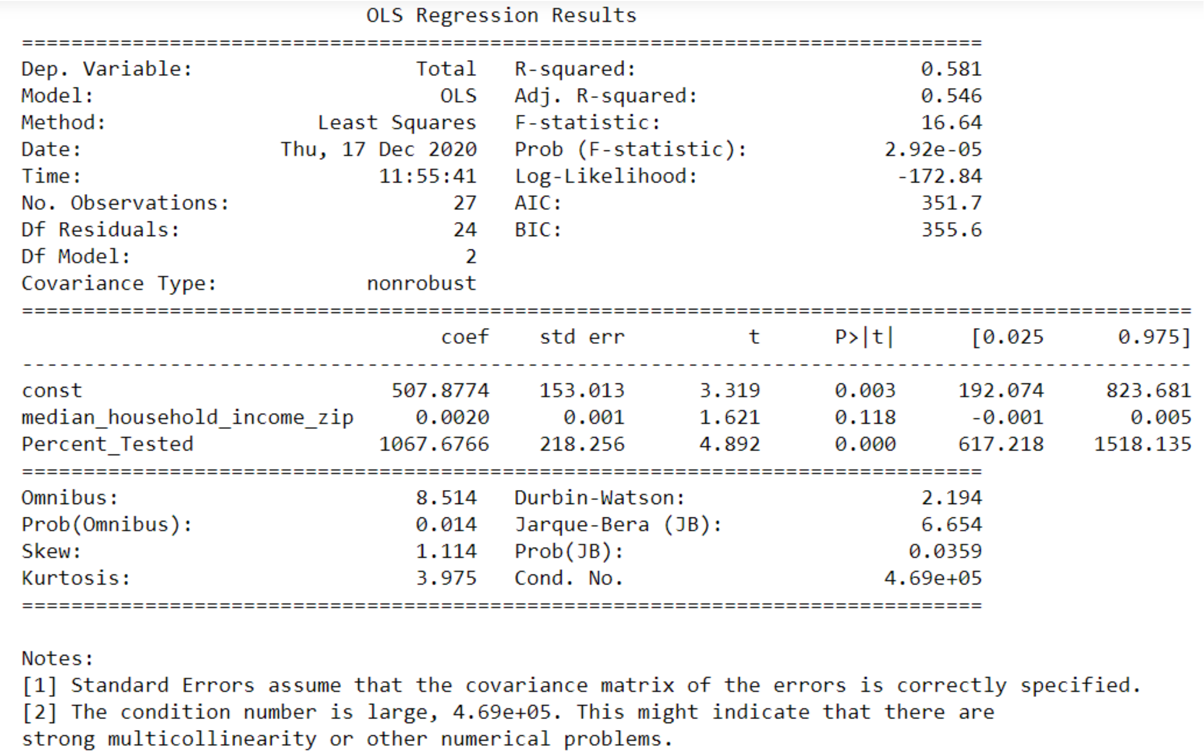
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Total   R-squared:                       0.394
Model:                            OLS   Adj. R-squared:                  0.391
Method:                 Least Squares   F-statistic:                     120.7
Date:                Thu, 17 Dec 2020   Prob (F-statistic):           4.10e-41
Time:                        11:53:41   Log-Likelihood:                -2408.3
No. Observations:                 374   AIC:                             4823.
Df Residuals:                     371   BIC:                             4834.
Df Model:                           2
Covariance Type:            nonrobust
======================================================================================
                         coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------
const               1222.9106     11.163    109.546      0.000    1200.959    1244.862
Student_Enrollment    -0.2677      0.032     -8.338      0.000      -0.331      -0.205
Number Tested          0.5018      0.042     12.062      0.000       0.420       0.584
==============================================================================
Omnibus:                      130.372   Durbin-Watson:                   1.878
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              463.215
Skew:                           1.540   Prob(JB):                     2.60e-101
Kurtosis:                       7.498   Cond. No.                     1.88e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.88e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

From this table we observed the following results: For student enrollment, a P-value of 0 indicates a strong effect of student enrollment on mean SAT Score. A coefficient of -0.27 indicates a slightly negative correlation between student enrollment and the mean SAT Score. For the number of students taking the SAT, a P-value of 0 also indicates it has a relatively strong correlation with the mean of total SAT score. Yet an adjusted R-squared value of 0.391 indicates that the joint effect of these variables on the mean of total SAT score is a bit weak.

An economic explanation for these results would be that the number of students taking the SAT in a school has correlations with the economic conditions of students' overall family incomes. A larger number of students taking the SAT means better overall economic conditions of students’ families and a larger ability to finance students’ SAT studies, which would potentially facilitate SAT performances.

In the last multivariate regression model, I changed my X variables again into median household income and percent of students taking SAT to examine their joint effect on mean SAT Score. Below are the summary statistics for this regression.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Total   R-squared:                       0.581
Model:                            OLS   Adj. R-squared:                  0.546
Method:                 Least Squares   F-statistic:                     16.64
Date:                Thu, 17 Dec 2020   Prob (F-statistic):           2.92e-05
Time:                        11:55:41   Log-Likelihood:                -172.84
No. Observations:                  27   AIC:                             351.7
Df Residuals:                      24   BIC:                             355.6
Df Model:                           2
Covariance Type:            nonrobust
===============================================================================================
                                  coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------------
const                         507.8774    153.013      3.319      0.003     192.074     823.681
median_household_income_zip     0.0020      0.001      1.621      0.118      -0.001       0.005
Percent_Tested               1067.6766    218.256      4.892      0.000     617.218    1518.135
==============================================================================
Omnibus:                        8.514   Durbin-Watson:                   2.194
Prob(Omnibus):                  0.014   Jarque-Bera (JB):                6.654
Skew:                           1.114   Prob(JB):                       0.0359
Kurtosis:                       3.975   Cond. No.                     4.69e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 4.69e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

From this table, we observed the following results: For median household income, a P-value of 0.12 is a bit large, indicating a moderate effect of household income on mean SAT Score. A coefficient of 0.002 indicates a slightly positive correlation between median household income and mean SAT Score. While for percent of students taking the SAT, a P-value of 0 is significant and indicates it has a relatively strong correlation with the mean of total SAT score.

The economic intuition behind these results is that a larger fraction of students taking the SAT indicates better economic conditions of students' families to put money and resources into the preparation for the SAT. What's more, when a larger fraction of students take the SAT, schools might tend to hire more good teachers to teach SAT materials, which also facilitates students' overall SAT performances.

At the end of these four regressions, we understand an overall picture that there exists a moderate correlation between student enrollment, median household income and mean SAT score of schools. As student enrollment and median household income increases, the mean SAT score also tends to increase as well. We also understand that these two variables also seem to have moderate to strong effects on the SAT performance of a school when combined with other factors.

Conclusion

In this paper, I defined two X variables: student enrollment and borough, together with an imported variable from the extra data I scraped, median household income at the zip code level, to answer our economic question "How would student enrollment, borough, and median household income potentially correlate with SAT performance". I conducted a detailed analysis of these variables using plots, heat maps, and both simple linear and multivariate regression models.

What's more, we also reached some main results that make our paper distinct from other papers in the same literature. For example, we specifically conducted regression analysis on our X and Y variables in NYC, instead of vaguely concluding a large overall pattern of SAT performances and socioeconomic status. We also created detailed heat maps to clearly demonstrate the distribution of mean SAT score across boroughs and zip codes in NYC. Overall, we have reached valid findings to answer our economic question which states the correlation between student enrollment, borough, median household income, and mean of SAT Total score.

Although our research is valid and distinct, there are still some questions unanswered in our paper. For example, we didn't investigate the correlation between different districts (such as boroughs) and the median household income of each district. We didn't talk about economic features like demographic structure or government intervention on school education either, which could also probably be correlated with the SAT performance of a school. There are also apparent limitations of my work. For instance, there are inconsistencies in the direction of the coefficient in my first, second, and third regressions for the variable student enrollment. In future research, we could try to add more data on government intervention or demographic structure; we could also scrape more data on other economic features and involve more complicated regression models to form a more comprehensive paper.

References

Sackett, P. R., Kuncel, N. R., Beatty, A. S., Rigdon, J. L., Shen, W., &amp; Kiger, T. B. (2012). The Role of Socioeconomic Status in SAT-Grade Relationships and in College Admissions Decisions. *Psychological Science*, *23*(*9*), 1000-1007. doi:10.1177/0956797612438732

Dixon-Román, E., Everson, H.T., & Mcardle, J. (2013). Race, Poverty and SAT Scores: Modeling the Influences of Family Income on Black and White High School Students' SAT Performance. *Teachers College Record, 115*.