Final Project: Predicting SAT Scores in NYC

David Lattimer

8/2/2020

For the data I wanted to find out if there were ways to find out the factors that effect SAT scores. To do this, I found a data set of 2015 New York City SAT scores, which included information about each school, such as SAT scores for math, writing and reading, demographics of the school (percentage of kids that are white, Hispanic, black and Asian), percentage of kids tested, and a ton of location/school information that wasn't important in how it would effect these scores. I then decided to find another data set that had a bunch of New York City census results from 2015 as well. I decided to add some information that I thought might be significant in rises and falls in the scores on the SATs. Unfortunately, the census takes information into either boroughs which I felt was a little too big and census tracts that are usually up to 7,500 people and not sorted by zip code or anything else. So I had to use the borough information collected from all the census tracts in that borough. From there I found the average income of the borough using $Income_{Borough} = \frac{\Sigma Income_{Tract} * Population_{Tract}}{Population_{Borough}}$, I found the poverty percentage of the borough using $Poverty_{Borough} = \frac{\Sigma Poverty_{Tract} * Population_{Tract}}{Population_{Borough}}$ and lastly found the unemployment rate of the borough using $Unemployment_{Borough} = \frac{\Sigma Unemployment_{Tract} * Population_{Tract}}{Population_{Borough}}$. From there I combined the two data sets by splitting the SAT scores by borough, removing the rows with missing information and then adding in the information, which was the same for all schools in that borough. Then I added the rows back onto each other and had a data set with all the information I needed. The data set ended up looking like:

```
##    School ID Zip Code   Borough Percent White Percent Black Percent Hispanic
## 1     01M539    10002 Manhattan          28.6          13.3             18.0
## 2     02M294    10002 Manhattan          11.7          38.5             41.3
## 3     02M308    10002 Manhattan           3.1          28.2             56.9
## 4     02M545    10002 Manhattan           1.7           3.1              5.5
## 5     01M292    10002 Manhattan           3.9          24.4             56.6
## 6     01M696    10002 Manhattan          45.3          17.2             18.7
## 7     02M305    10002 Manhattan           2.7          41.9             49.2
## 8     01M509    10002 Manhattan           2.5          39.9             51.2
## 9     01M448    10002 Manhattan           3.3          25.0             41.1
## 10    02M543    10002 Manhattan           3.9          30.8             56.9
##    Percent Asian Average Math Score Average Reading Score Average Writing Score
## 1           38.5                657                   601                   601
## 2            5.9                395                   411                   387
## 3            8.6                418                   428                   415
## 4           88.9                613                   453                   463
## 5           13.2                410                   406                   381
## 6           17.1                634                   641                   639
## 7            5.8                389                   395                   381
## 8            5.8                438                   413                   394
## 9           29.9                437                   355                   352
## 10           5.9                381                   396                   372
##    Percent Tested Income Per Capita Poverty of Borough Unemployment of Borough
## 1            91.0          64995.14           17.87959                7.965419
```

```
## 2           78.9          64995.14           17.87959                7.965419
## 3           65.1          64995.14           17.87959                7.965419
## 4           95.9          64995.14           17.87959                7.965419
## 5           59.7          64995.14           17.87959                7.965419
## 6           70.8          64995.14           17.87959                7.965419
## 7           80.8          64995.14           17.87959                7.965419
## 8           35.6          64995.14           17.87959                7.965419
## 9           69.9          64995.14           17.87959                7.965419
## 10          73.7          64995.14           17.87959                7.965419
```

Once I was able to add in the columns from the census data (which took a lot more effort and time than I would like to admit), I could find the information which was important and significant from these variables. Of course, now that we were looking to find trends in all three of the score variables, we have three different variables to try to predict. To do this we use a linear model to find a model that represents our data the best. For the math variable, two of the variables that we have in the data set are insignificant, which are "percent Asian" and "Poverty Rate of Borough" which both did not help our model enough to use. Math scores ended up looking like:

```
##
## Call:
## lm(formula = combined_df$"Average Math Score" ~ combined_df$"Percent White" +
##     combined_df$"Percent Black" + combined_df$"Percent Hispanic" +
##     combined_df$"Percent Tested" + combined_df$"Income Per Capita" +
##     combined_df$"Unemployment of Borough", data = combined_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -128.640  -19.687   -1.851   18.394  133.477
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       5.027e+02  2.211e+01  22.732  < 2e-16
## combined_df$"Percent White"      -7.595e-01  2.456e-01  -3.093  0.00213
## combined_df$"Percent Black"      -2.519e+00  1.594e-01 -15.806  < 2e-16
## combined_df$"Percent Hispanic"   -2.769e+00  1.727e-01 -16.030  < 2e-16
## combined_df$"Percent Tested"      8.845e-01  1.217e-01   7.268 2.20e-12
## combined_df$"Income Per Capita"   8.792e-04  1.701e-04   5.167 3.90e-07
## combined_df$"Unemployment of Borough"  5.832e+00  1.325e+00   4.402 1.41e-05
##
## (Intercept)                      ***
## combined_df$"Percent White"      **
## combined_df$"Percent Black"      ***
## combined_df$"Percent Hispanic"   ***
## combined_df$"Percent Tested"     ***
## combined_df$"Income Per Capita"  ***
## combined_df$"Unemployment of Borough" ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 37.25 on 368 degrees of freedom
## Multiple R-squared:  0.7362, Adjusted R-squared:  0.7319
## F-statistic: 171.2 on 6 and 368 DF,  p-value: < 2.2e-16
```

The other two scores, all of the variables used were significant and the model was made better by using

them, so their models looked like this for writing:

```
##
## Call:
## lm(formula = combined_df$"Average Writing Score" ~ combined_df$"Percent White" +
##     combined_df$"Percent Black" + combined_df$"Percent Hispanic" +
##     combined_df$"Percent Asian" + combined_df$"Percent Tested" +
##     combined_df$"Income Per Capita" + combined_df$"Unemployment of Borough" +
##     combined_df$"Poverty of Borough", data = combined_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -150.560  -19.328    1.388   19.261  132.513
##
## Coefficients:
##                                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        7.099e+02  1.325e+02   5.359 1.48e-07
## combined_df$"Percent White"       -3.044e+00  1.308e+00  -2.327 0.020485
## combined_df$"Percent Black"       -4.916e+00  1.287e+00  -3.820 0.000157
## combined_df$"Percent Hispanic"    -5.463e+00  1.261e+00  -4.332 1.91e-05
## combined_df$"Percent Asian"       -3.923e+00  1.287e+00  -3.048 0.002468
## combined_df$"Percent Tested"       8.847e-01  1.269e-01   6.973 1.46e-11
## combined_df$"Income Per Capita"    1.227e-03  2.314e-04   5.304 1.97e-07
## combined_df$"Unemployment of Borough"  1.861e+01  4.998e+00   3.723 0.000228
## combined_df$"Poverty of Borough"  -4.604e+00  1.731e+00  -2.659 0.008175
##
## (Intercept)                       ***
## combined_df$"Percent White"       *
## combined_df$"Percent Black"       ***
## combined_df$"Percent Hispanic"    ***
## combined_df$"Percent Asian"       **
## combined_df$"Percent Tested"      ***
## combined_df$"Income Per Capita"   ***
## combined_df$"Unemployment of Borough" ***
## combined_df$"Poverty of Borough"  **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 38.47 on 366 degrees of freedom
## Multiple R-squared:  0.6525, Adjusted R-squared:  0.6449
## F-statistic: 85.89 on 8 and 366 DF,  p-value: < 2.2e-16
```

and this for reading:

```
##
## Call:
## lm(formula = combined_df$"Average Reading Score" ~ combined_df$"Percent White" +
##     combined_df$"Percent Black" + combined_df$"Percent Hispanic" +
##     combined_df$"Percent Asian" + combined_df$"Percent Tested" +
##     combined_df$"Income Per Capita" + combined_df$"Unemployment of Borough" +
##     combined_df$"Poverty of Borough", data = combined_df)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -144.09  -17.61    0.79   19.54  127.44
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          7.201e+02  1.305e+02   5.520 6.44e-08
## combined_df$"Percent White"         -3.084e+00  1.288e+00  -2.394 0.017170
## combined_df$"Percent Black"         -4.848e+00  1.267e+00  -3.826 0.000153
## combined_df$"Percent Hispanic"      -5.447e+00  1.242e+00  -4.386 1.51e-05
## combined_df$"Percent Asian"         -3.997e+00  1.267e+00  -3.154 0.001744
## combined_df$"Percent Tested"         8.036e-01  1.250e-01   6.431 3.96e-10
## combined_df$"Income Per Capita"      1.238e-03  2.279e-04   5.433 1.01e-07
## combined_df$"Unemployment of Borough" 1.812e+01  4.923e+00   3.681 0.000267
## combined_df$"Poverty of Borough"    -4.433e+00  1.705e+00  -2.600 0.009698
##
## (Intercept)                         ***
## combined_df$"Percent White"         *
## combined_df$"Percent Black"         ***
## combined_df$"Percent Hispanic"      ***
## combined_df$"Percent Asian"         **
## combined_df$"Percent Tested"        ***
## combined_df$"Income Per Capita"     ***
## combined_df$"Unemployment of Borough" ***
## combined_df$"Poverty of Borough"    **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 37.88 on 366 degrees of freedom
## Multiple R-squared:  0.6332, Adjusted R-squared:  0.6252
## F-statistic: 78.99 on 8 and 366 DF,  p-value: < 2.2e-16
```

From these three models, the important information is in the p value which the smaller the number, the more significant the variable is. This is how we removed "Poverty of Borough" and "Percent Asian" from the math data, but every other shows to be significant in all the models. The next important number from these summaries is the R^2 and adjusted R^2, which tell us how much of the variability of the estimates can be taken from the variables we used. We have an R^2 of 73.62% in the math model, 65.25% in the writing model and 63.32% in the reading model, which show a lot of the variability is covered by the data we used. It can't account for everything, but these are good values to have gotten. The last numbers are how these variables are related to the outcome. Finding their relationship is important and we find some interesting information from them. For all of the demographic variables, in all three of the models, we see something interesting. As all of these percentages rise, the scores on each of the SAT tests fall. We know this by seeing the negative relationships in all of the variables, which doesn't seem to make any sense for the data. But I think what this means is that diversity in schools leads to better test scores. Having a large number of any one race leads to lower test scores in general, and the more diverse a school is the better. Next, we get a positive relationship between percent tested and income per capita, which seems to make sense since the less a student has to worry about wealth and problems associated with that, the more they would be able to focus on school. Also a school encouraging a higher percentage of their kids to take the SATs will focus on the importance and help kids succeed which is shown from that relationship. Poverty of the borough has a negative relationship, so the higher rate of poverty in a borough, the lower the scores on the SATs. The last relationship is the only one that doesn't make a ton of sense, and that is the positive relationship between unemployment and test scores. Consistently, the higher the unemployment rate of a borough, the higher the test scores. This may be due to a lack of data or an inability to break up boroughs into zip codes for the census data, but it is the only one that doesn't make a ton of sense in my mind. One way we can check that the variables we measured are going to be useful is to check the confidence intervals of the variables.

Essentially what we are looking for in the 3 sets of confidence intervals below is small gaps between the sets of numbers and for them to not cross from negative to positive numbers.

```
##                                          2.5 %        97.5 %
## (Intercept)                          4.592268e+02 546.200693613
## combined_df$"Percent White"         -1.242430e+00  -0.276632089
## combined_df$"Percent Black"         -2.832853e+00  -2.205971277
## combined_df$"Percent Hispanic"      -3.108330e+00  -2.429031203
## combined_df$"Percent Tested"         6.451941e-01   1.123816926
## combined_df$"Income Per Capita"      5.446116e-04   0.001213785
## combined_df$"Unemployment of Borough" 3.227213e+00   8.437619962
```

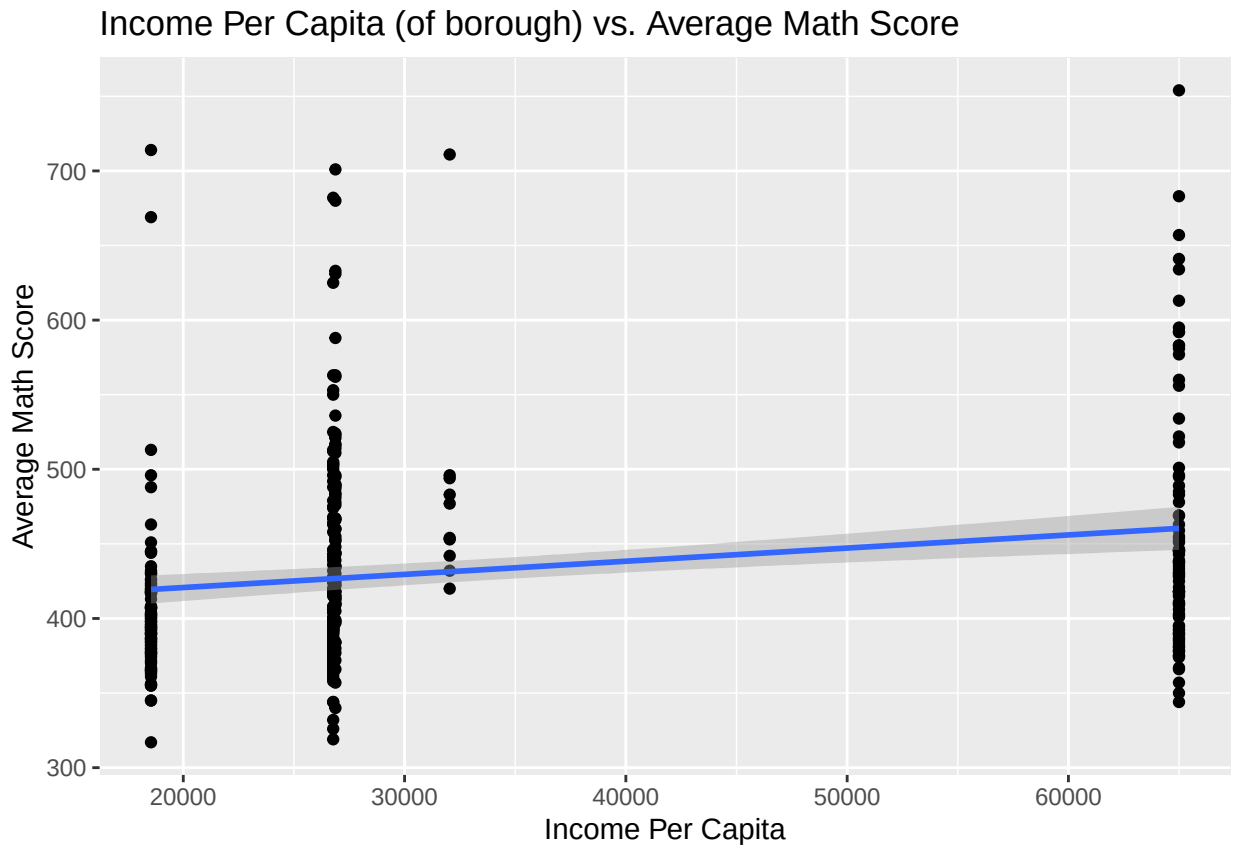
```
##                                          2.5 %        97.5 %
## (Intercept)                          4.494246e+02 970.406842598
## combined_df$"Percent White"         -5.616261e+00  -0.472171364
## combined_df$"Percent Black"         -7.446069e+00  -2.385341701
## combined_df$"Percent Hispanic"      -7.942933e+00  -2.983437677
## combined_df$"Percent Asian"         -6.453420e+00  -1.392240565
## combined_df$"Percent Tested"         6.351872e-01   1.134165818
## combined_df$"Income Per Capita"      7.721549e-04   0.001682091
## combined_df$"Unemployment of Borough" 8.778553e+00  28.436153573
## combined_df$"Poverty of Borough"    -8.008761e+00  -1.199465277
```

```
##                                          2.5 %        97.5 %
## (Intercept)                          4.635478e+02 976.640118254
## combined_df$"Percent White"         -5.616879e+00  -0.550692087
## combined_df$"Percent Black"         -7.340305e+00  -2.356217851
## combined_df$"Percent Hispanic"      -7.888843e+00  -3.004454873
## combined_df$"Percent Asian"         -6.489450e+00  -1.504918103
## combined_df$"Percent Tested"         5.578918e-01   1.049313850
## combined_df$"Income Per Capita"      7.899086e-04   0.001686065
## combined_df$"Unemployment of Borough" 8.441532e+00  27.801434167
## combined_df$"Poverty of Borough"    -7.786562e+00  -1.080387397
```

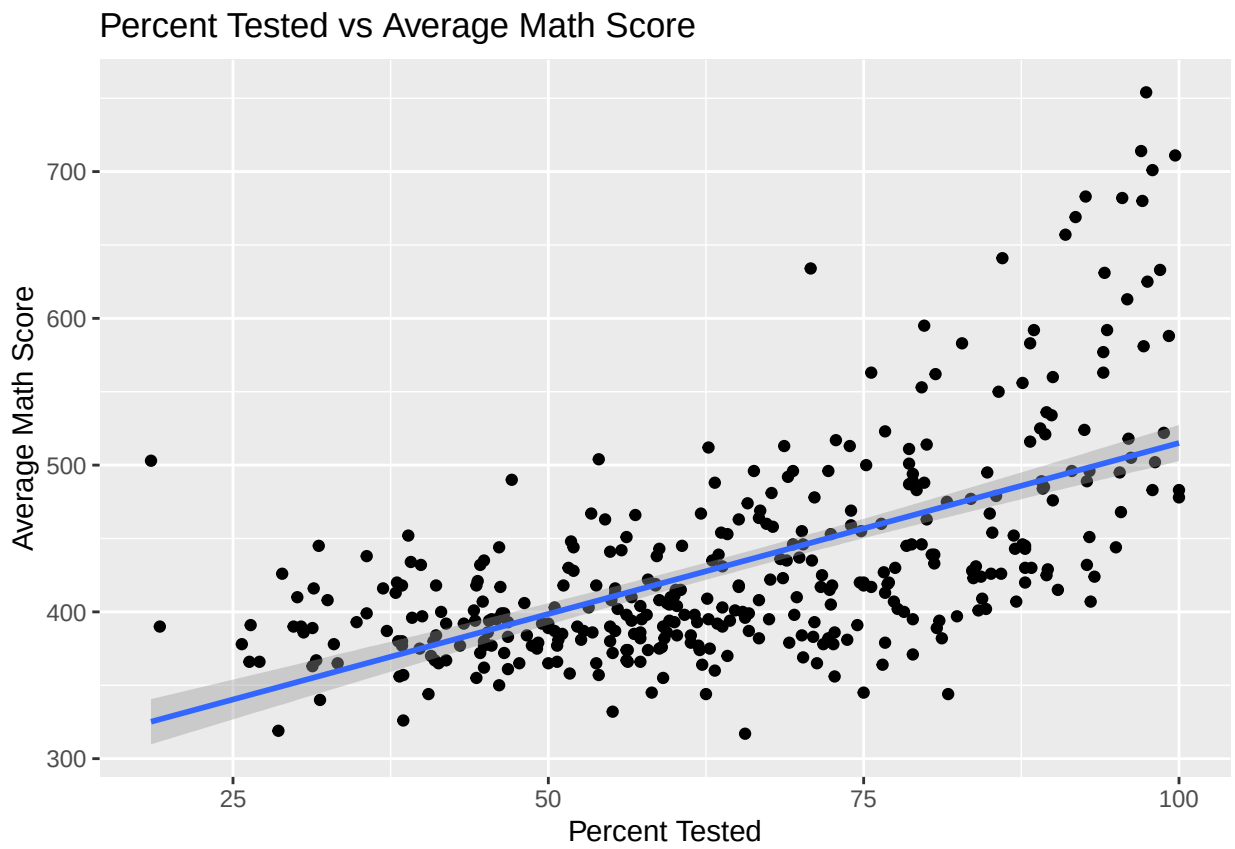
When looking at these intervals we got what we wanted. None of the negative relationships cross into positive numbers, and the same vice versa. We also see pretty small gaps between the numbers showing that our model will work for most sets of data.

Essentially for this data, we were able to find a lot of the variability in the SAT scores and find some interesting insights into what can help increase test scores. Obviously money and help from the school in encouraging testing will lead to higher scores as shown here:

```
## `geom_smooth()` using formula 'y ~ x'
```

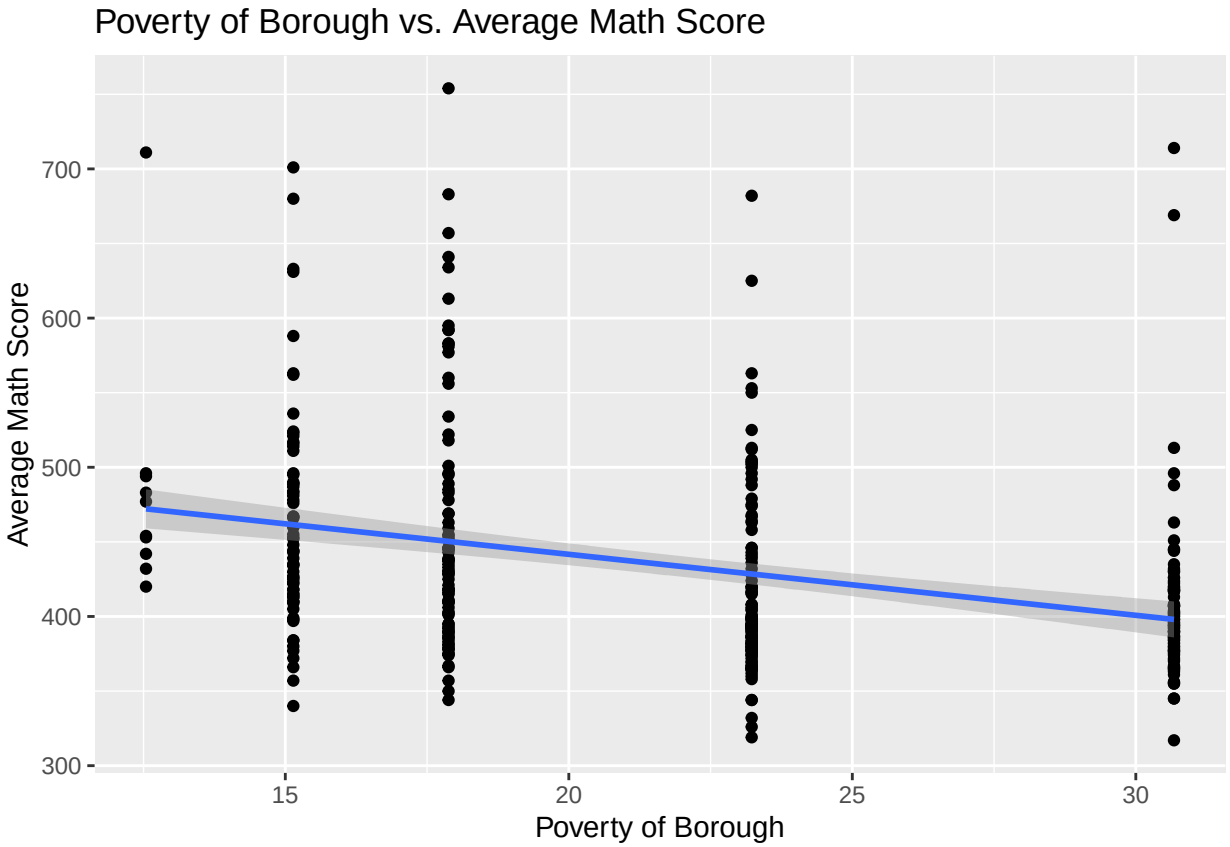



```
## `geom_smooth()` using formula 'y ~ x'
```

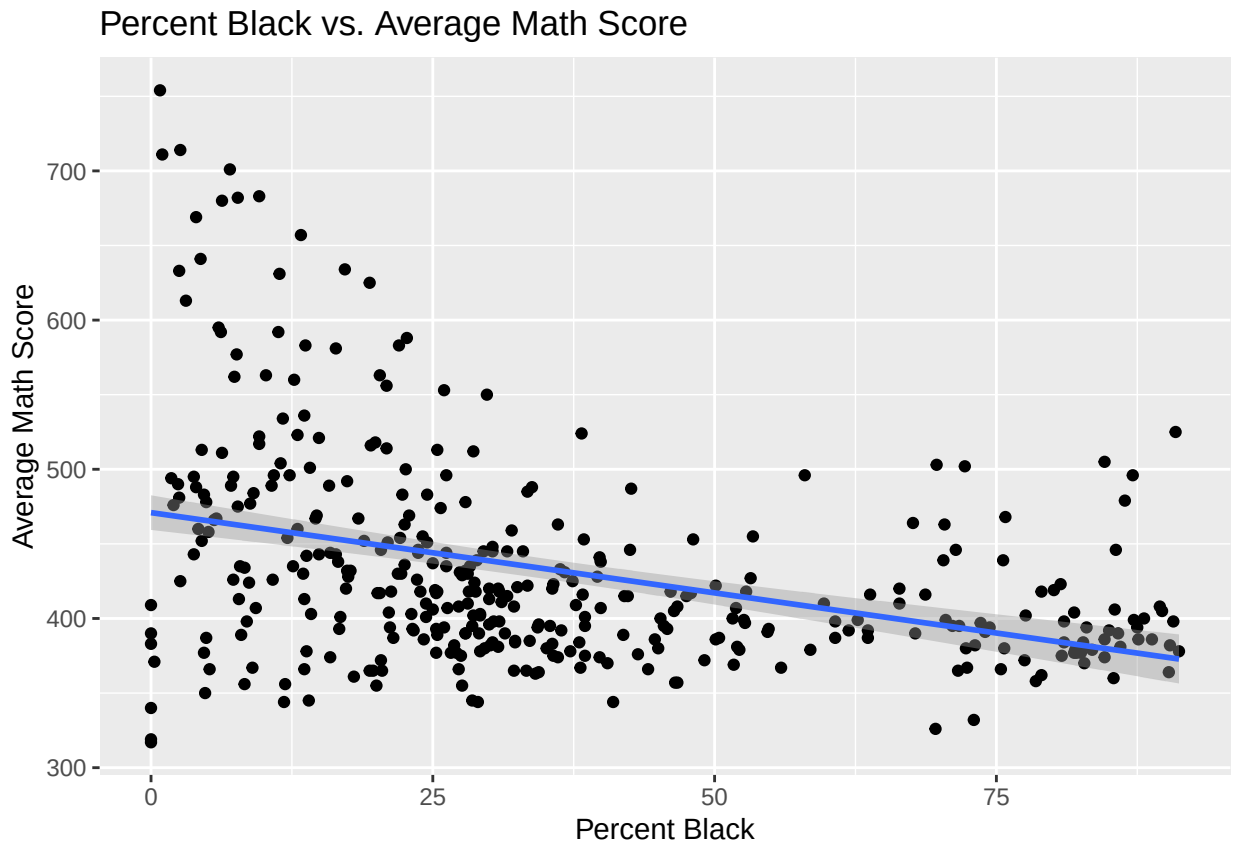


Sorry the income graph looks a little weird since the income variables are based off the average income of the borough so all the points are under 5 different incomes. You can still see the positive relationship, just like the graph of percent tested and math scores. Now the variables that had a negative relationship were any of the demographic percentages and poverty rate as shown here:

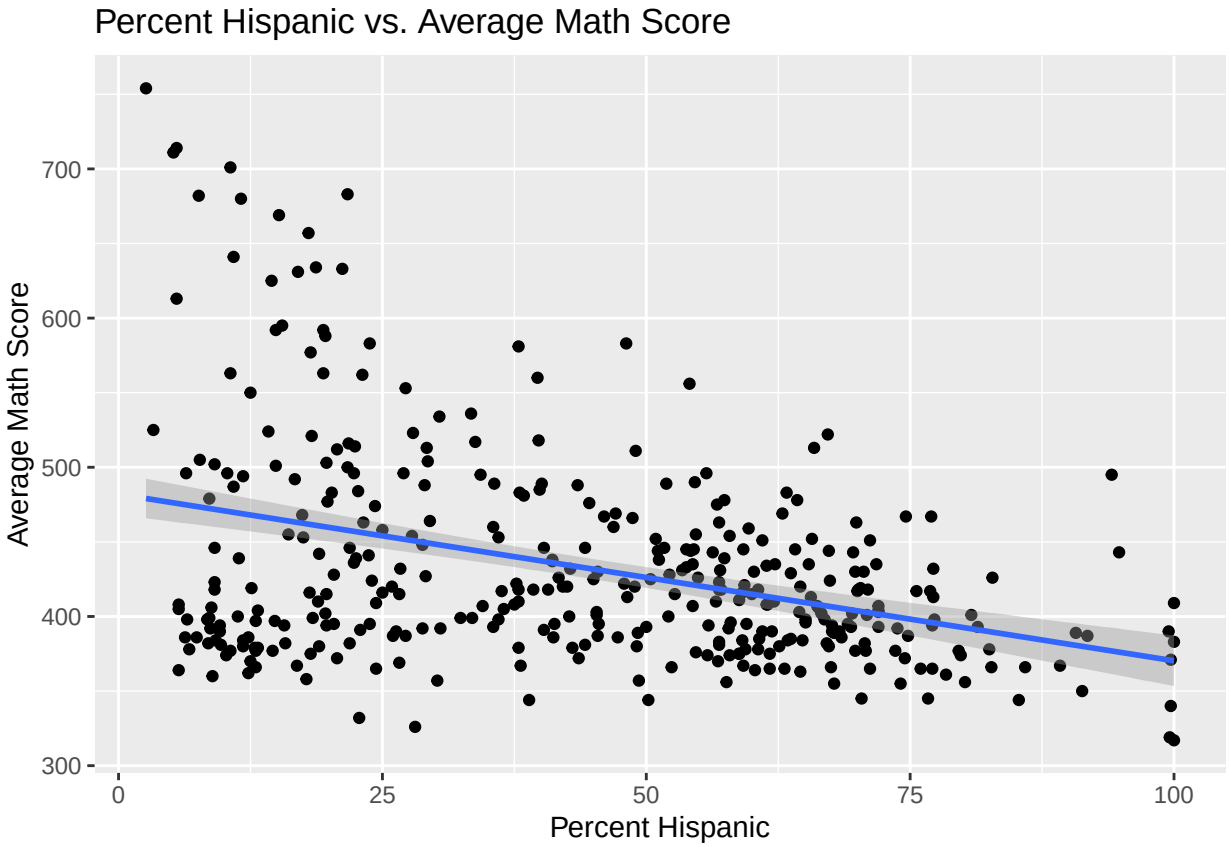
```
## `geom_smooth()` using formula 'y ~ x'
```




```
## `geom_smooth()` using formula 'y ~ x'
```




```
## `geom_smooth()` using formula 'y ~ x'
```



After looking into this subject you can see some of the variables and how they effect the SAT scores of a school. If I were to do this again in a perfect world I would have information based off zip code so we could narrow in more closely on the the different areas and find their rates of poverty and income on a smaller scale. But overall everything should be here.

(Disclaimer: if you want to see everything that was done for this project, open this in RMD and see all the steps taken to combine data frames and check information. Not all of it was important and useful, but there is a lot more there than just this write up. Thanks!)