Finally, and this part is a little open ended: do a one-page write up answering this question: Do you think these predictions are good? Put this (preferably as a PDF) inside your ***Github repo*** *as well.*

Goal:

For my prediction model, I want to learn about at what school size has the best score performance.

Hypothesis:

The reason some school's SAT score are higher might be due to more students with competitive pressure, or maybe less students with more resources per each.

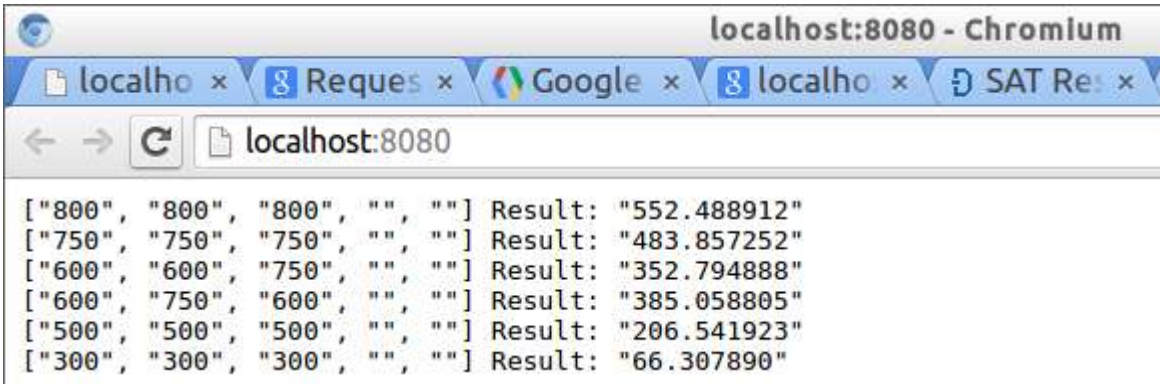
Design:

Regression numerical output, 3 SAT scores columns, and 2 label columns

Here was my training data:

	A	B	C	D	E	F
1	Num of SAT Test Tal	SAT Critical Reading	SAT Math Avg. Score	SAT Writing Avg. Sc	DBN	SCHOOL NAME
2	29	355	404	363	01M292	HENRY STREET SCHOOL FOR INTERNATIONAL STUDIES
3	91	383	423	366	01M448	UNIVERSITY NEIGHBORHOOD HIGH SCHOOL
4	70	377	402	370	01M450	EAST SIDE COMMUNITY SCHOOL
5	7	414	401	359	01M458	FORSYTH SATELLITE ACADEMY
6	44	390	433	384	01M509	MARTA VALLE HIGH SCHOOL
7	112	332	557	316	01M515	LOWER EAST SIDE PREPARATORY HIGH SCHOOL

Here are the test results:



Analysis:

From the results, it seems the best scores are schools with higher student test counts. Given the size of sample data is only 400, and the number of student tests count can range from 1 to infinite, the prediction might not completely accurate to predict the size of the school. We can transform the student count to lognormal for linear relationship. It does seem there is a correlation with the size. We can do an anova analysis to see the stats behind the data.

Further steps:

To improve the training model, more columns of data such as geography, income, and parent education might play a role in school score performance. Another study could also look at the name of the school might have an impact. For example, some schools might be only focus on science and math.