Airton Kamdem

**Objective**: Train a Machine Learning Model that can explain most of the randomness observed within a range of average sat scores outlines by school within the New York Public School district.

**WORK:** Through a combination of API calls and CSV downloads, I pulled a list of all of the public schools in New York City with corresponding average SAT test scores coming from the college board. The initial dataset included a number of factors that I believe could build a strong model with a low MSE and high R2 based on factors such as the total amount of **time the kids spend in school.** To calculate the time in school, I had to strip the AM/PM times from the csv, then subtract the provided end\_time from the provided start\_time and create a new column to house this. I then had to remove all NAN values along with their corresponding collums from the dataset, and make sure every value being fed into the model was numeric.

The initial model was relatively weak with a .6 R2 and .36 and a heavy bias towards the racial breakdown of the schools. In fact, removing race as a factor rendered the model useless with an R2 of 30 and MSE of 80. Diving into the racial element, I went back to the New York City public school databases and pulled additional surveys and assessments that covered everything from the Economic Need Index of each school, city, borough to the leadership effectiveness and parental involvement. Adding these factors required a series of inner mergers that aligned the data aligned the school\_id and school names, matching the sat scores, and individual school parameters accordingly. This helped us uncover that there were in fact a number of factors masking themselves as race or influencing what was initially coming across as race. Some of these included the Economic Need Index, whether or not the student was a native English speaker, if they had a disability. These all had a negative correlation, but among the strongest correlations were the student’s school attendance record, how supportive their environment is, and whether or not their parent responded to the school wide survey, as this indicates not just the presence of a parent or guardian, but it also indicates that this parent is an active participant in their child’s education.