PUI2015 Extra Credit Project Proposal
SAT Score Predictor
<Anita M Ahmed, ama908, N11949714>

Problem Description: SAT scores is one of the factors that affect admission decision for college entrant students. This factor also determines scholarships received by the student, therefore given the importance of the SAT score in determining student college eligibility, it is an interesting problem to determine what factor can affect the students SAT score.

Although there might be several approach to this problem. Given, the limited availability of data the research will be conducted to establish these following correlations at a school district level:

1) Does the High School attendance and class size have any correlation with SAT Scores?
2) Is there any correlation between students' grade 6 Mathematics performance and SAT Mathematics Score?
3) Is there any correlation between students' grade 6 English performance score and SAT English Score?

Data: Data was taken from NYC open data source Socrata.
1. SAT Results-The most recent school level results for New York City on the SAT. Results are available at the school level for the graduating seniors of 2012. Records contain 2012 College-bound seniors mean SAT scores taken during SY 2012. To perform the analysis this data will be aggregated at a school district level.   Mean SAT scored for Reading, Writing and Math Scores will be summed up to get a mean of overall SAT score at the school district level.

Link: https://data.cityofnewyork.us/Education/SAT-Results/f9bf-2cp4

2. School Attendance and Enrollment By District - 2010-11 - 2010- 11 Attendance & Enrollment (Unaudited) by District as of December 31, 2010. Charter Schools, Community-Based Organizations (Pre-K), Home Instruction, and Hospital Schools are not included. This dataset will provide average attendance at each districts level.

Link:https://data.cityofnewyork.us/Education/School-Attendance-And-Enrollment-By-District-2010-/rfpq-hs49

3. 2010-2011 Class Size - School-level detail Average class sizes for each school, by grade and program type (General Education, Self-Contained Special Education, Collaborative Team Teaching (CTT)) for grades K-9 (where grade 9 is not reported by subject area), and for grades 5-9 (where available) and 9-12, aggregated by program type (General Education, CTT, and Self-Contained Special Education) and core course (e.g. English 9, Integrated Algebra, US History, etc.).
Class size data is based on January 28, 2011 data. To perform analysis this data will be aggregated at a district level, only using 9-12 grade attendance data from the dataset.

Link - https://data.cityofnewyork.us/Education/2010-2011-Class-Size-School-level-detail/urz7-pzb3

4. NYS Math Test Results By Grade 2006-2011 - District - All Students New York City Results on the New York State Mathematics Tests, Grades 3 - 8 . The 2006 Grade 6 students' math test scores at each school district. Because it is assume students who were in grade 6 in 2006 took the SAT in 2012.

Link - https://data.cityofnewyork.us/Education/NYS-Math-Test-Results-By-Grade-2006-2011-District-/gyaz-82xj

5. English Language Arts (ELA) Test Results 2006-2012 - District - All Students Latest available data and trends in the state assessment results of English Language Arts for grades 3 through 8. Data are disaggregated by district. The 2006 Grade 6 students ELA test scores at each district. Because it is assume students who were in grade 6 in 2006 took the SAT in 2012.

Link - https://data.cityofnewyork.us/Education/English-Language-Arts-ELA-Test-Results-2006-2012-D/yhfh-vyns

Analysis: Python to merge datasets and perform correlation test on data.

References: The following papers are used as reference, to get an idea about statistical methods that was used to study some factors that might affect performance on SAT

1. An examination of the relationship between gender, race/ethnicity, socioeconomic status, and SAT performance by *St. Rose, Andresse*
   This paper examined the relationship between gender, race/ethnicity, socioeconomic status, and SAT math and verbal performance among college-bound seniors in 2004. Using multiple regressions this study looks at all three variables simultaneously to identify possible interrelationships between them.

Link: http://pqdtopen.proquest.com/doc/304645025.html?FMT=ABS

2. An Examination of the Relationship among Grade Point Average, Socioeconomic Status, Disability Status and ACT Performance by Tennessee Technological University.
   This study was to examine the relationship between students' Grade Point Average, socio-economic status, and disability status and ACT composite scores of 250 students enrolled in public high schools located in Marion County, Tennessee. Data was analyzed and interpreted through a multiple correlation.

Link:https://iweb.tntech.edu/jcbaker/Sample%20paper%201%20%20FOED%206920.pdf

Deliverable: Use NYC open data to and statistical methods to determine if SAT scores depends on average attendance, class size, and grade 6 performance in mathematics and ELA scores at a school district level.