**Final Project Proposal**

# Name of the dataset: California SAT score reports

Data Source Location: <https://data.world/data>

Description of how data source was found: I have searched in data world with the key term SAT scores, then I found the above data set which have met the requirements of the rubric.

Cost incurred to download the dataset: Zero, the dataset is free to download.

Legality of the data set: The data set is from a free source and can be accessible by everyone and the data is legal to use.

Documented Exploratory Data Analysis: I have searched a lot on google but did not find any previously documented EDA on the datasets.

Data types in the dataset: Numeric: Int, Categorical; Datetime

Project proposal relevant to 601 objectives: The project that I have selected is relevant to the objective of data601

Description (Stories in the Data): The project is regarding multiple datasets which contains California SAT scores from 1998 to 2016. I am aiming to combine the datasets from 2010 to 2016 and their number of rows are 2431, 2473, 2595, 2293, 2332 and 2335 rows. The common columns the datasets contained are County number, District number, School number, School name, grade 12th marks, mean score, average score in math and verbal.

Data size: The combined dataset occupied is 3-4MB. The data combined won’t have more than 2600 rows and the common columns are 7 in number.

Hypothesis:

1. I am assuming that the mean score has increased by atleast 10 percent for every years.
2. I am also assuming that the number of takers are increasing by 15 percent.

Questions to be answered:

1. Which county presented the largest change in the increase of takers
2. Which county presented the mean scores and are they same?