Capstone Project 1 - Milestone Report

Factors Influencing ACT Scores

1. Problem Statement

The ACT test is designed to assess high school students' general educational development and their ability to complete college-level work. ACT scores are also used to compare schools and districts on the basis of their average score and to evaluate state educational programs. Students can prepare for and take the test several times to attain their desired score. ACT test covers four subject areas: English, mathematics, reading, and science. Each subject area test receives a score ranging from 1 to 36.

In this study, I investigated the factors related to schools' performance on the ACT. In order to achieve this, I addressed the following questions.

1. What is the distribution of the ACT scores?
2. Which high schools received the highest and lowest overall ACT scores?
3. Which counties have the highest and lowest performing schools?
4. On which subjects are the schools performing better?
5. Is there an increase of the success level of schools with time?
6. What kind of factors influence success levels?
    - Ethnicity
    - Number of low income students
    - New teacher ratios
    - Population of the schools
6. Are the parents' and GreatSchools' ratings related to school success?

2. Dataset

The ACT data to be used is obtained from the California Department of Education (https://www.cde.ca.gov/ds/sp/ai/). Each dataset consists of a row for every accredited high school in California with its code number, school name, county name, enrollment based on the number of students in grade twelve, number of students, the average english, math, reading and science scores, the number of students who scored 21 points or more, the percent of students who scored 21 points. The data covers the period from 2014 through 2018.

The other two datasets showing the census information about the high schools in California are obtained from greatschools.org (https://www.greatschools.org/california) website. The census datasets include information about ID number, longitude, latitude, city, county, number of students enrolled, percentage of different ethnic backgrounds, percentage of teachers who are in their first or second years, percentage of low income students who have the right to receive reduced priced lunch of the high schools.

2.1. Data Wrangling

Data Wrangling Part 1:

For the ACT data, I had five different datasets belonging to different years ranging from 2014 to 2018. Before taking the cleaning steps, I added year columns to some of the datasets with no year column information.

1. **Cleaning steps:**

    a. To have consistent column names, I cleaned the extra white spaces in the column names and capitalized them to concatenate the datasets from all years.
    b. I constructed new columns by trimming the redundant ones.
    c. I formatted the year column and set it as index.
    d. I corrected the incompatible value types in columns.
    e. To eliminate empty spaces in the text of some columns, I used col.astype(str).str.strip() method.

2. **Missing Data:** I handled the missing data in two steps.

    a. In the data, the missing values were shown with ‘ * ’ sign. I changed them using na_values=['*'].
    b. Then, I dropped the rows with too many Nan values using .dropna(thresh=7) method.

3. **Outliers:** I applied the .describe() method on the data and inspected the outcomes. The max and min values of the columns were compatible with the other numbers.

Data Wrangling Part 2:

I had three csv files in this section: The ACT file, gsid file (school ID information used to retrieve other data from greatschools website) and the census data that I acquired using api request. The main process in this notebook was merging the datasets with as much common information as possible. The act data had school and county names, the gsid data had school and city names, and the census data had only the school names column that I could use in the merging process. Thus, I added county names to the gsid data with gsid['City'].map(city_cnty) method using the excel file which includes city names and their corresponding counties. Then, using the school ID numbers, I merged the census and gsid datasets and created census_id data with the command shown: census_id = pd.merge(census, gsid, left_on='Id', right_on='gsId', how ='inner', suffixes = ('','_gsid')).

Some school names were spelled differently in the ACT file and census_id data. To merge these files, I used the fuzzywuzzy library string matching. Another issue was that there were multiple schools with the same name in California. To fix this issue, I matched the schools names only if they were in the same county. I merged the ACT data with census_id data based on two columns (school names and county names) using the final_df = pd.merge(act, census_id, left_on=['Sname2', 'Cname'], right_on=['Sname', 'County'], how ='inner',suffixes = ('','3')) command.

1. **Cleaning steps:**

    a. I corrected the incompatible value types in columns (strings, integers, etc.).
    b. Then I removed the school names that caused duplicates in the final data.
    c. There were a few school names which were mismatched. Manually, I created a dictionary to correct the mismatched school names. Then, I dropped the redundant columns.

2. **Missing Data:** In the final data there wasn't much missing data because of the major cleaning steps that I've performed in the first data wrangling part. In the final data, I replaced the missing values with Nan and left them as is.

3. **Outliers:** I applied the .describe() method on the final data to be able to inspect the outliers. In the 'Enrollment' column, the maximum value seemed very high. So, setting up a threshold value for the number of students, I obtained the school names with high enrollment number using final_df[final_df['Enrollment']>4500] code. Those schools were the largest high schools in California according to the information on the website, and I kept them as is. The other columns didn't have any outliers.

3. Findings from Exploratory and Statistical Data Analysis

Through exploratory and statistical data analysis, significant relationships were found between several variables.

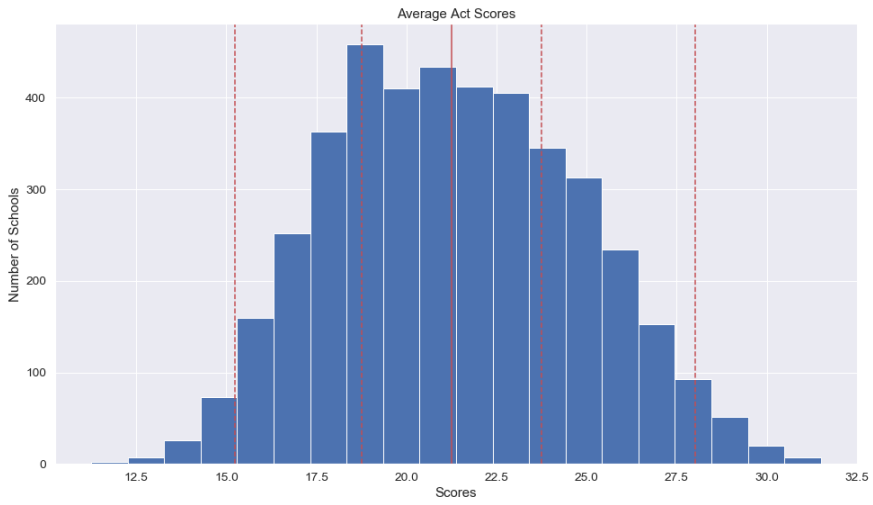

Figure 1. *Distribution of average ACT scores*

Figure 1 illustrates the distribution of the average ACT scores. About 4.5% of the schools received 28.00 (top 2.5 %) or above, and 6.3% of the schools received 15.00 (2.5%) or less at least once between 2014 and 2018.

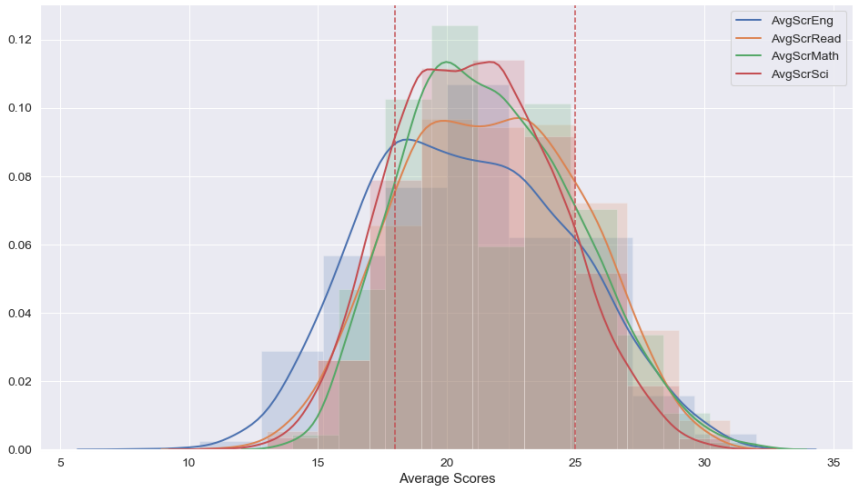

Figure 2. *Distribution of subject test scores*

The schools in the low score range (≤18) perform better in English. The schools in the middle score range (18-25) perform better on the Math and Science sections, worse in English. The schools in the high score range (>=25) perform worse on the Science section. About 60% of the schools received about 22 or less in English, and about 40% of the schools received 22.0 or more in Science. This number is less for the other subject tests.

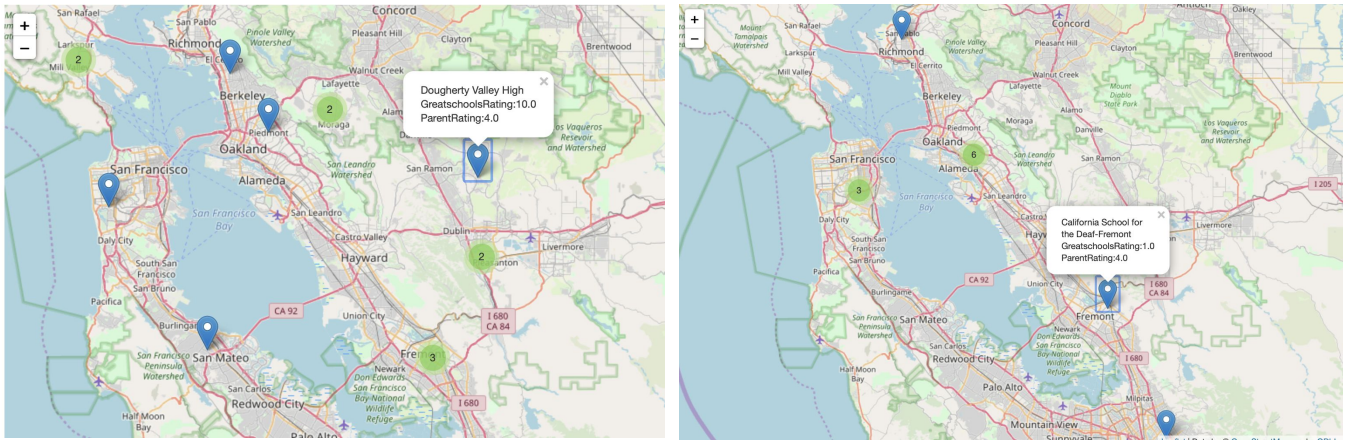

Figure 3. *Highest and lowest performing schools on map*

The highest performing schools are in the following counties: Santa Cruz, Alameda, Contra Costa, Marin, Sacramento, San Francisco, San Diego, Orange, San Mateo, Santa Clara, Los Angeles, and Nevada. The lowest performing schools are in the following counties: San Joaquin, Santa Clara, Kings, Alameda, Contra Costa, Tulare, Riverside, Fresno, Sacramento, San Francisco, San Diego, San Bernardino, and Los Angeles.

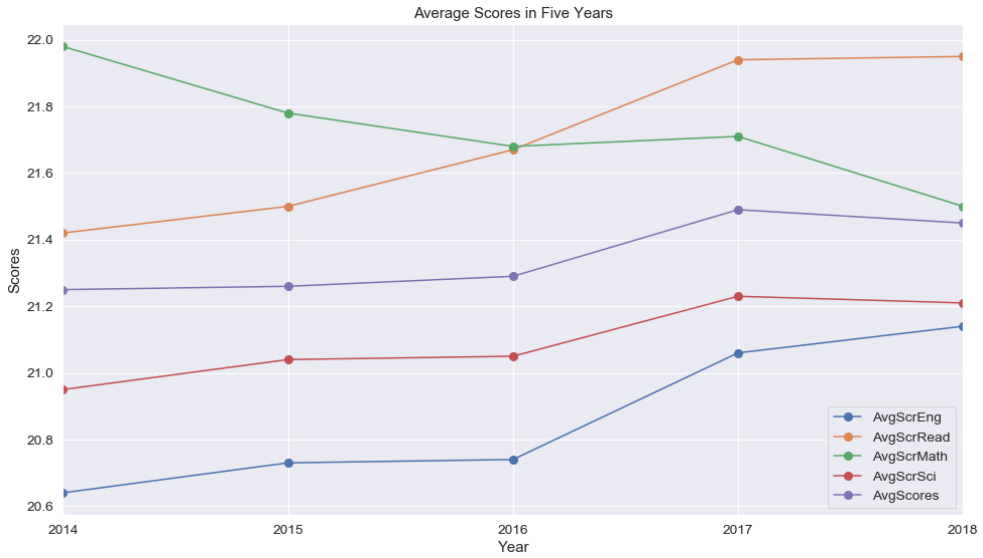

Figure 4. *Average scores in five years*

While average reading, english and science scores increased over five years, the average math scores decreased. On the other hand, overall scores also increased from 2014 to 2018.

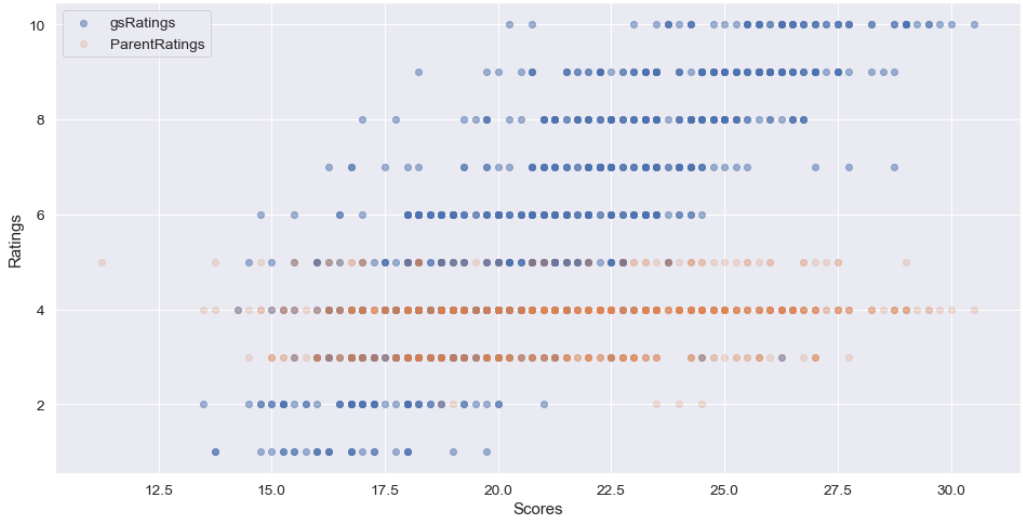

Figure 5. *Parent and Great Schools ratings*

As the average ACT scores increase, the GreatSchools ratings also increased, but this was not true for parent ratings. We can conclude that there is a strong correlation between the GreatSchool ratings and school success, but the parent ratings are not related to the success.

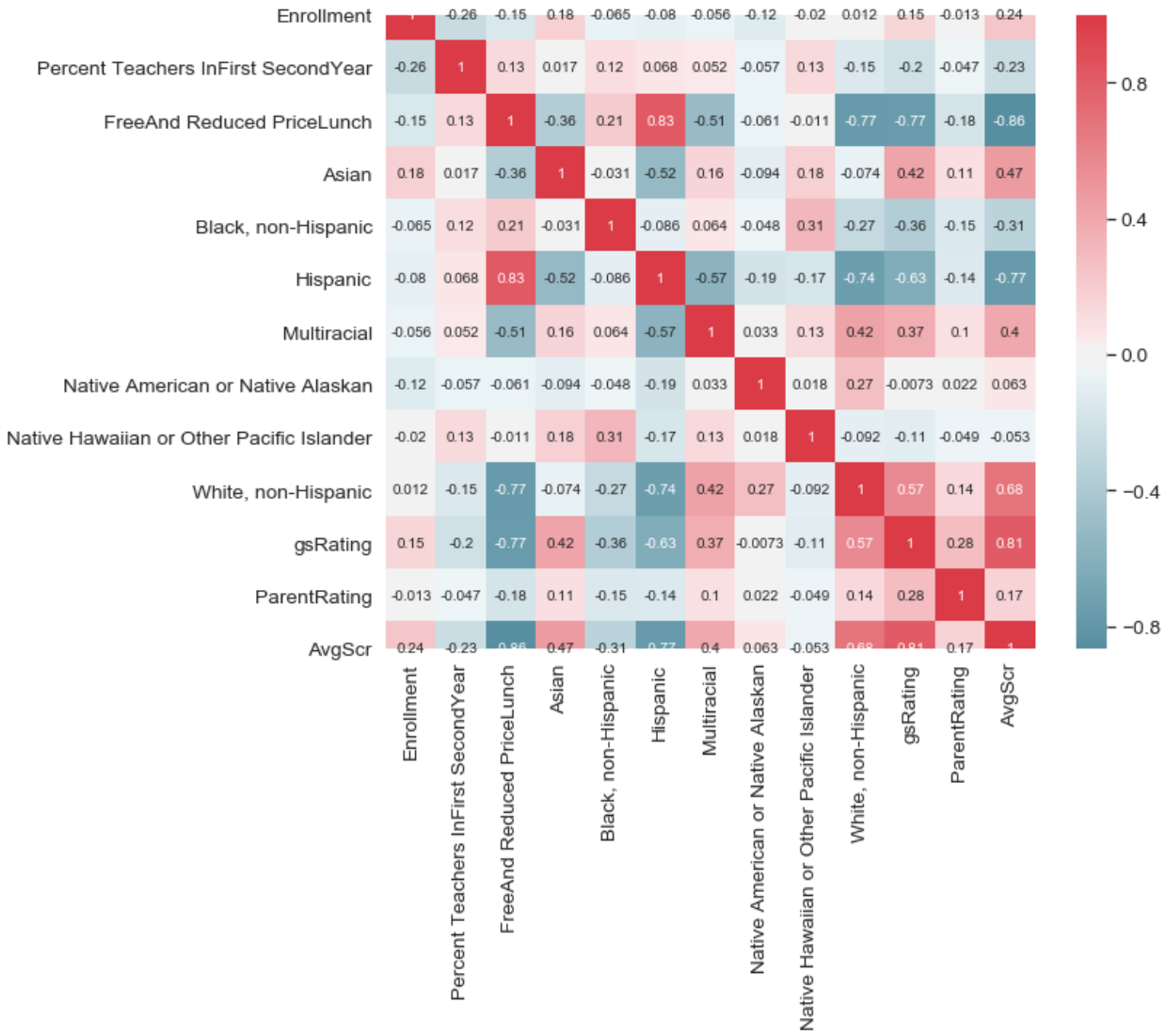

Figure 6. *Correlation Matrix*

The percentage of Hispanic students is positively correlated to the number of students who receive free and reduced price lunch.The percentage of White, non-Hispanic students is negatively correlated to the number of students who receive free and reduced price lunch.

The GreatSchools ratings are negatively correlated to the number of students who receive free and reduced price lunch, but positively correlated to the average Act scores.

Overall, average Act scores are negatively correlated to the number of students who receive free and reduced price lunch, and also are negatively correlated to the percentage of Hispanic students.