Capstone Project 1 - Exploratory Data Analysis

Factors Influencing ACT Scores

1. Questions

Performing Exploratory Data Analysis in order to address these 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. At which subject ACT tests are the schools performing better?
5. Is there an increase with the success level of the schools with time?
6. What kind of factors have an effect on the success level?
    - Ethnicity
    - Number of low income students
    - New teacher ratios
    - Population of the schools
7. Are the parent ratings and great schools ratings related to the school success?

2. Trends

The first step in the analysis was looking at the distribution of the average ACT scores of four subject tests. For that reason, I created a histogram and an ecdf graph of the average scores. Then I calculated the 2.5, 25, 50, 75, 97.5 percentiles of the average scores in order to determine the highest and lowest performing schools. For the highest performing schools I chose the schools which received 28.0 or higher on the ACT, which is the top 2.5% of the overall scores. For the lowest performing schools, I chose the schools which received 15.25 or lower on the test which is 2.5% of the average scores. Then, I showed these schools with their parent and GreatSchools ratings on the map using the folium package. Using these school names, I determined the counties that contain the highest and lowest performing schools.

In order to determine at which subject tests schools are performing best, I used histograms and showed all the subject tests on the same graph with their density functions and compared them in the low, middle and high score ranges. I also created a boxplot for more explicit visualization.

I showed mean scores for each of the subject tests (english, reading, math and science), as well as the overall mean score on the same plot for each of the years from 2014 to 2018 in order to illustrate the trend in the scores.

For each of the ethnic groups, I made histograms showing the number of schools with their percentage of the ethnic groups. I also looked at the relationship between the percentage of the ethnic groups and the average scores using sns.regplot method. I used the same

techniques for the percentage of free and reduced priced lunch students to represent the percentage of low income students, teachers who are in their first or second years, and the total number of students enrolled in these schools to investigate the relationship with the average scores.

I used scatter plot to show the relationship between the parent and Great Schools ratings and the average scores in order to envision how parent ratings and great schools ratings related to school success.

Finally, I made a correlation matrix for the variables to show the correlation coefficients to summarize the data.

3. Resulting visualizations and conclusions

I showed that the average scores are normally distributed, and this can also be seen by looking at the graphs.

About 4.5% of the schools received 28.00 (top 2.5 %) or above, 6.3% of the schools received 15.00 (2.5%) or less at least once between 2014 and 2018.

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, Los Angeles.

The schools in the low score range (≤18) are performing better in English. The schools in the middle score range (18-25) perform better in Math and Science, worse in English. The schools in the high score range (>=25) perform worse in Science. 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.

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 2013-2018.

As the average ACTscores are increasing, the Great Schools ratings also increased, but this was not true for parent ratings. We can conclude that there is a strong correlation between the Great School ratings and school success, but the parent ratings are not related to the success.

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

The Great Schools 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.