https://colab.research.google.com/drive/1Q0EU88srVNzRIcz9YC-gopW-PCYDFcht

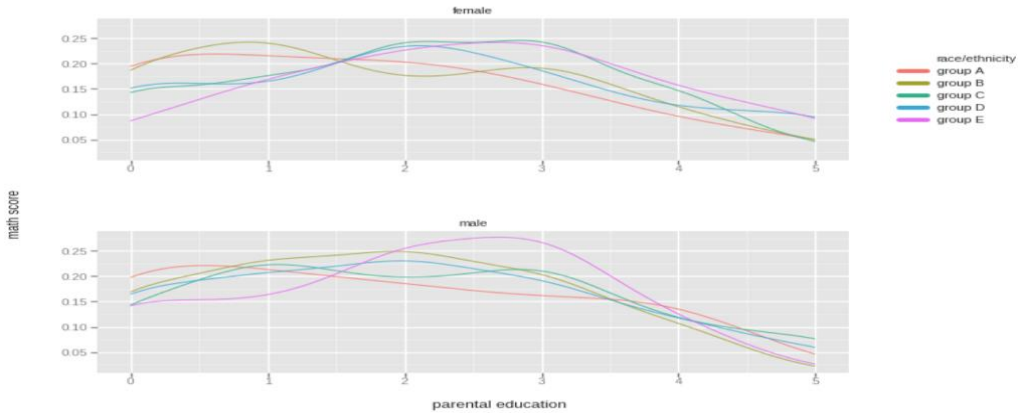
Overview

In this assignment, I explore students' performance data using different data visualization techniques and get insights from the data.

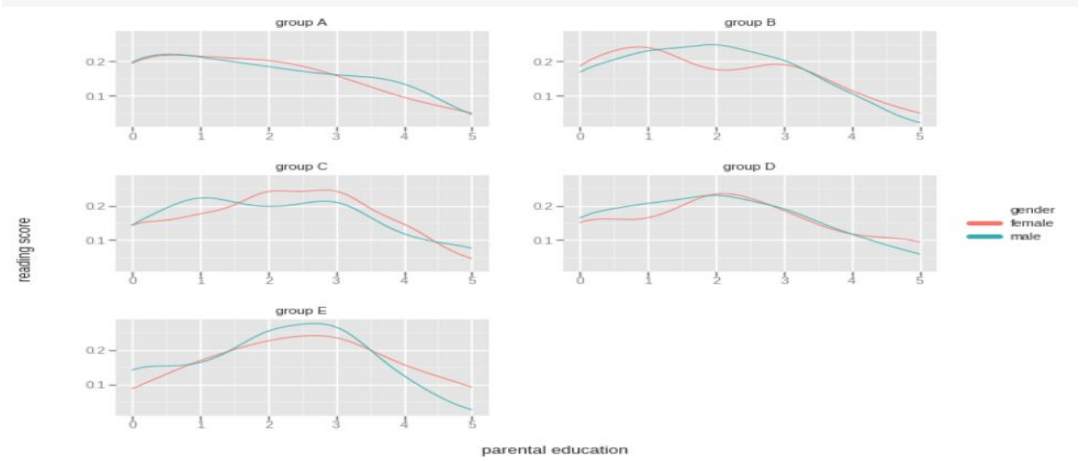
Technique

The students' exam performance data have eight variables in total, including five influential variables and three response variables. Influential variables are *gender*, *race*, *parental level of education*, *lunch* and *test preparation course*. math score, reading store, and writing score can be treated as response variables, but there might be relationships among these three variables.

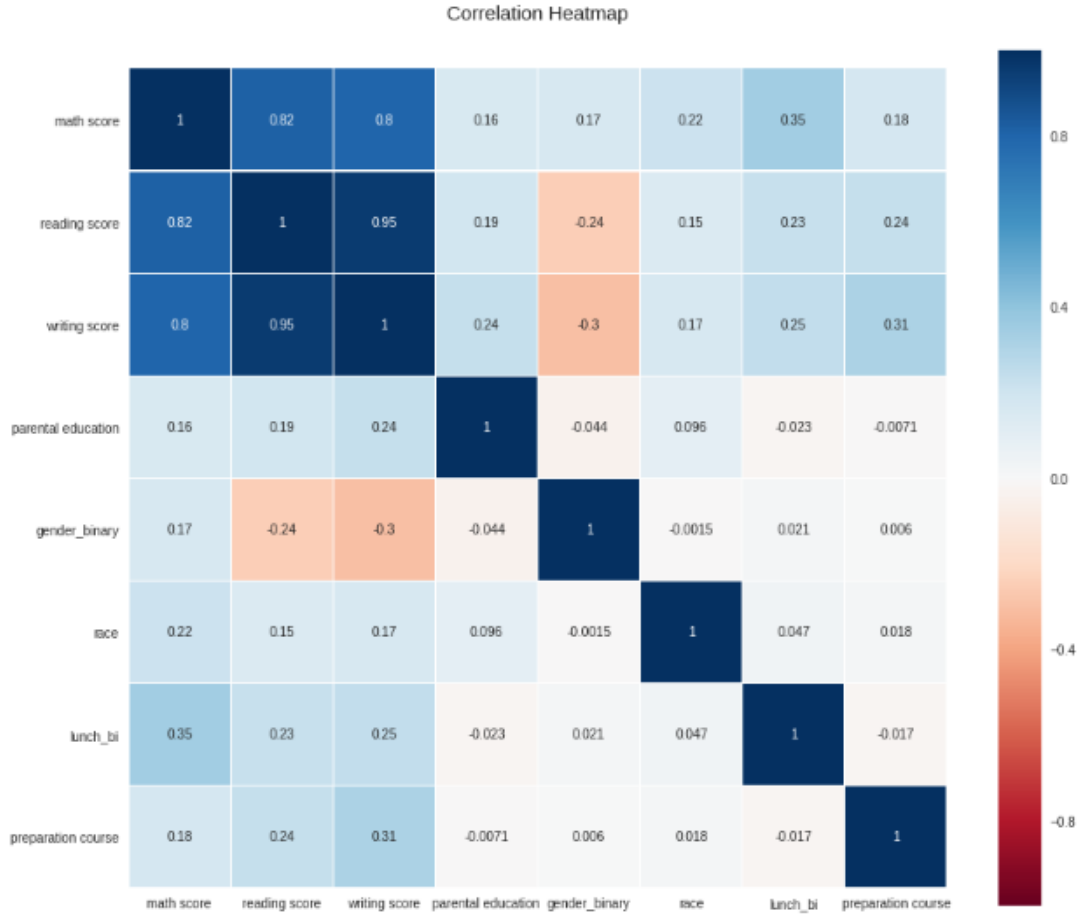
The first task is using *ggplot* to create faceted plots. After installing and importing necessary packages, I start processing raw data. First, I check whether there are missing values and find this data set is pretty neat and do not have missing values. Then, I try to factorize *parental level of education*: I assign a value of 0~5 to parents' education levels from low to high. The assigned variable is named *parental education*. Then I use *facet_wrap* function in *ggplot* to draw this diagram. The dimension I choose is *gender*, so the plot is going to represent the relationship between *math score* and *parental education level,* and each race has one line. One plot only shows the data of one gender.



Next, I compare math scores of students who have different genders but are in the same



race. I use the same function and get the following plots:
The second task is creating a heatmap using seaborn package. So, the first step is converting all character variables to numerical variables. Then, I use *heatmap* function creating this plot:



The third task is creating my own test and training sets. For this task, I use the *train_test_split* function in *sklearn.model_selection* package and *Standard Scaler* function in *sklearn.preprocessing* package.

Conclusion

Generally speaking, the educational level of parents will affect the children's math scores, and the education level of the parents whose children have the highest math scores is mostly concentrated in associate's degree and some college. Interestingly, the students whose parents have the highest education level got the lowest math scores. For female students, Group B and Group C have better grades in math. For male students, Group B and Group E have better grades in math. Within the same ethnic group, there is no significant difference in math achievement between boys and girls.

From the heatmap, we can observe that the two other grades have the highest correlation with math scores. This may be because students with higher scores usually have better performances in every subject. Another variable that correlated strongly with math scores is the students' lunch level.