DS8007 - Advanced Data Visualization

Project: Students Performance in Exam

Description of the problem:

Students performance in exam generally represent their skills on different subjects, however, there could be various influence factors that have an impact on their exam performance. I have worked on a project to visualize student's performance in exam. I have further tried to identify interesting correlations within the given dataset and understand the influence of various factors behind their exam performances.

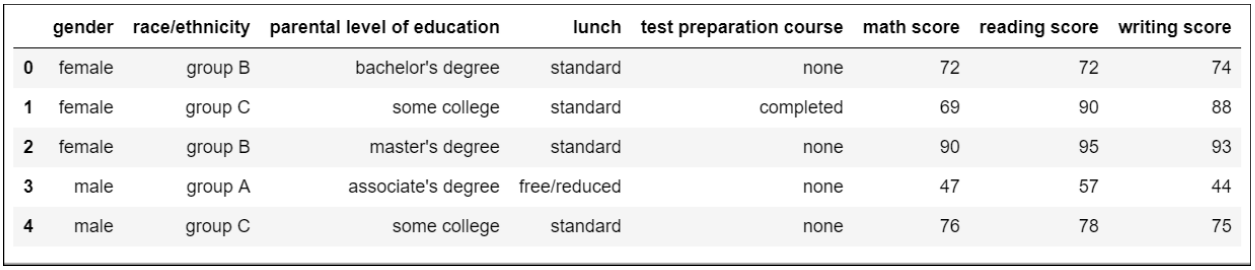
The dataset I have worked with consists of students exam score in different subjects along with their gender, ethnicity, parent's education and the type of lunch they used to take. Based on the given factors, I have visualized the score by different category, such as by subjects, genders etc. and then identified some correlation among the scores. In addition, I have identified some influencing factor that has impact on the students score.

Data source:

I have used the open source dataset from Kaggle site to get the students' scores in exam. Data source URL: http://roycekimmons.com/tools/generated_data/exams

This data set consists of the marks secured by the 1000 students in 3 different subjects, Math, Reading and Writing. In addition, it consists of some influencing factors such as gender, race, parents education, lunch type, and if the student took test preparation courses.

There is no missing value in the given dataset. Below is a glimpse of few data from the dataset:



	gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score
0	female	group B	bachelor's degree	standard	none	72	72	74
1	female	group C	some college	standard	completed	69	90	88
2	female	group B	master's degree	standard	none	90	95	93
3	male	group A	associate's degree	free/reduced	none	47	57	44
4	male	group C	some college	standard	none	76	78	75

Fig 1: Glimpse of dataset

Research Questions:

RQ1: How to visualize the dataset observations in different categories and students' scores in exam?
RQ2: Is there any correlation among the scores students achieved in exam for different subjects?
RQ3: What are the visible and influencing factors for the students' score in exam?

Technologies used:

I have implemented the project using Python Anaconda Notebook (python version 3.5). Specifically, I have used python pandas, numpy, seaborn and matplotlib libraries.

Data Analysis and Visualization:

In this section, I will elaborate my findings according to the Research Questions (RQs) stated above:

RQ1: Visualize Dataset

To answer RQ1 (How to visualize the dataset observations in different categories and students' scores in exam?), I will illustrate the dataset observations in different categories from high level.

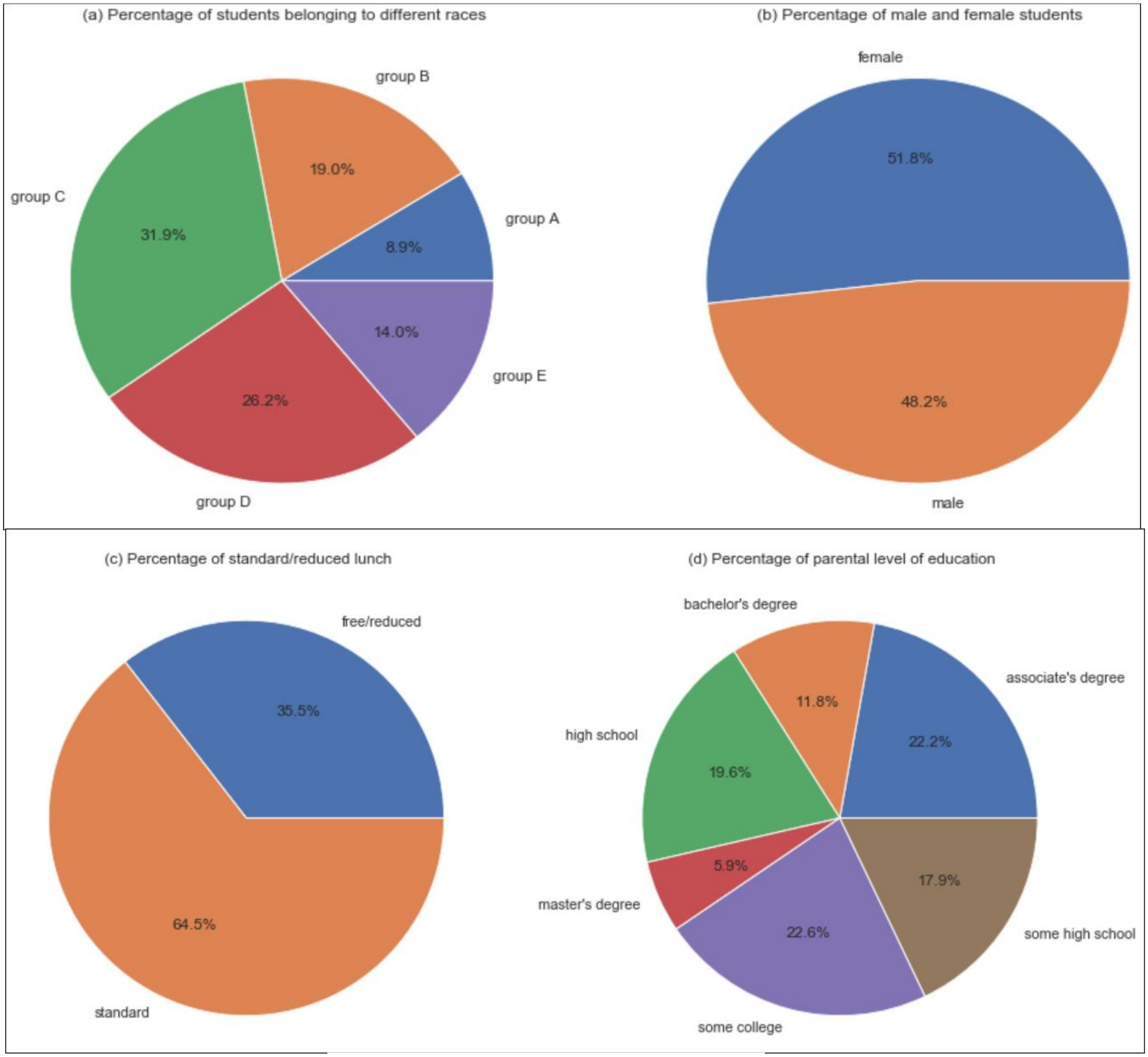


Fig 2: Data observations

It is always important to understand the data observations from different perspective in order to get better insights. The pie charts in Fig 2 shows the percentage of students falls under different category/factors. It is seen from the Fig 2(a) that the majority of the students belong to ethnic group C and group D, respectively 31.9% and 26.2%. We also see in Fig 2(b) the number of female students are slightly higher compared to male students. Among all the students, majority of the students take standard lunch than the reduced lunch as in Fig 2(c). Fig 2(d) shows the parents education level, and as seen only 6% of the parents have Masters' degree, whereas approximately 45% of parents have associate/college degree combined.

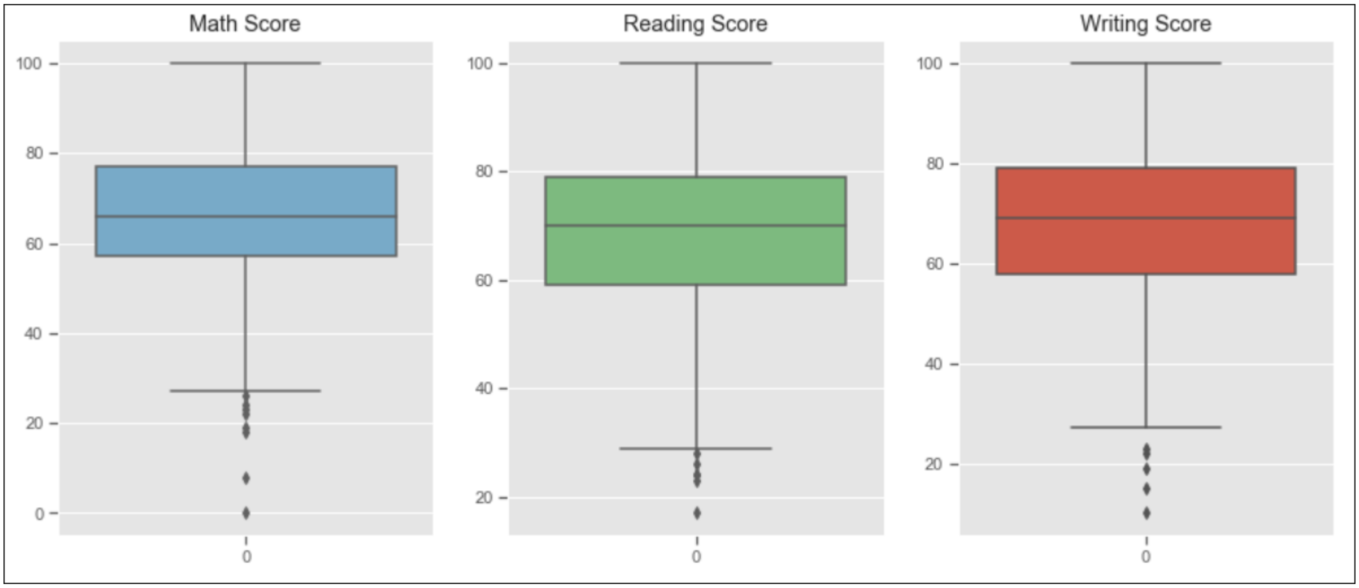


Fig 3: Score distribution in boxplots for Math, Reading & Writing

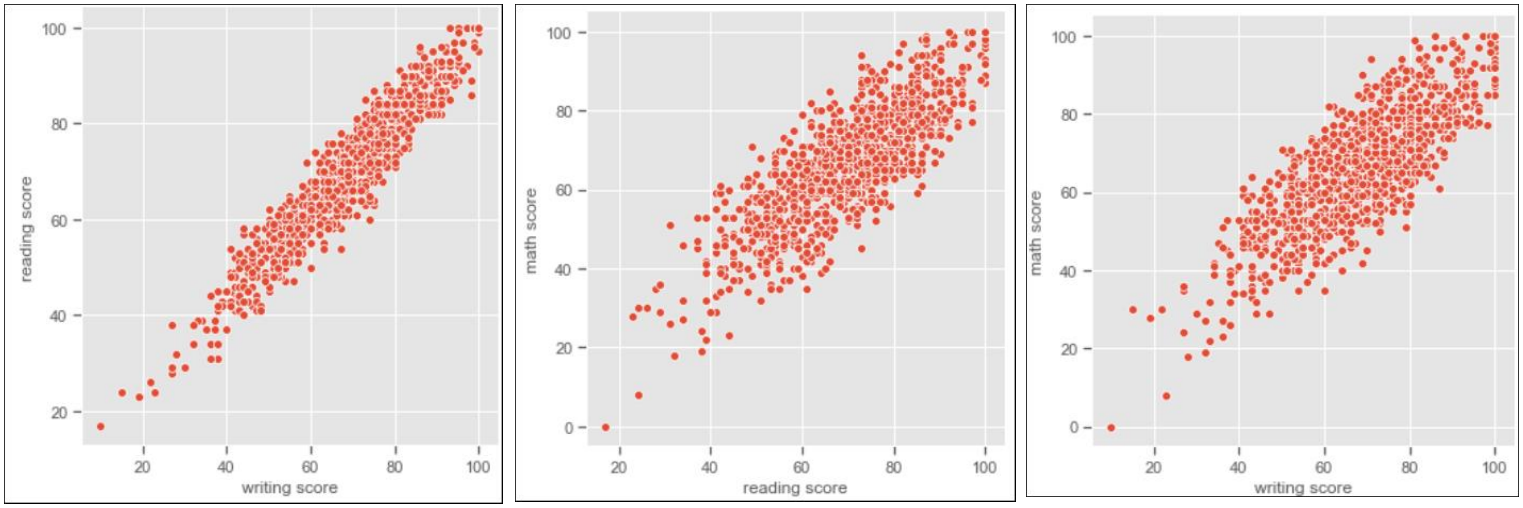
The above Fig 3 shows distribution of students' scores in Math, Reading and Writing in boxplots. It shows a similar distribution of scores in all three subjects with some minimal differences. For example, the Q1 for Reading score is lower than the others, but majority of the students scored between 60 and 80 for all three subject.

RQ1: Finding Correlation in scores

To answer RQ2 (Is there any correlation among the scores students achieved in exam for different subjects?), I will show correlations among Reading, Writing and Math scores. Fig 4 shows correlations and Pearson co-efficient among these scores. We see in Fig 4(a) that the Reading and Writing scores are linearly correlated. It can be also verified from the Pearson's Correlation Co-efficient (r), which is a value between -1 to 1, where closer to 1 interprets more linear correlation. In this case the correlation co-efficient value for Reading and Writing is approximately 0.95, which is closer to 1, and hence, Reading and Writing scores exhibit higher linear correlation.

On the other hand, although scores for Reading & Math, and Writing & Math shows a linear correlation, however, these are less linear compared to the Reading & Writing scores, as indicated by the Pearson's Correlation Co-efficient shown in Fig 4 (b) (c).

Therefore, based on these observation, we can say Reading and Wring scores are most linearly correlated among others. In other words, students who get better score in Reading, also get better score in Writing.



r = 0.95
(a)

r = 0.82
(b)

r = 0.81
(c)

Fig 4: Correlation and Pearson co-efficient (r) among Reading, Writing and Math scores

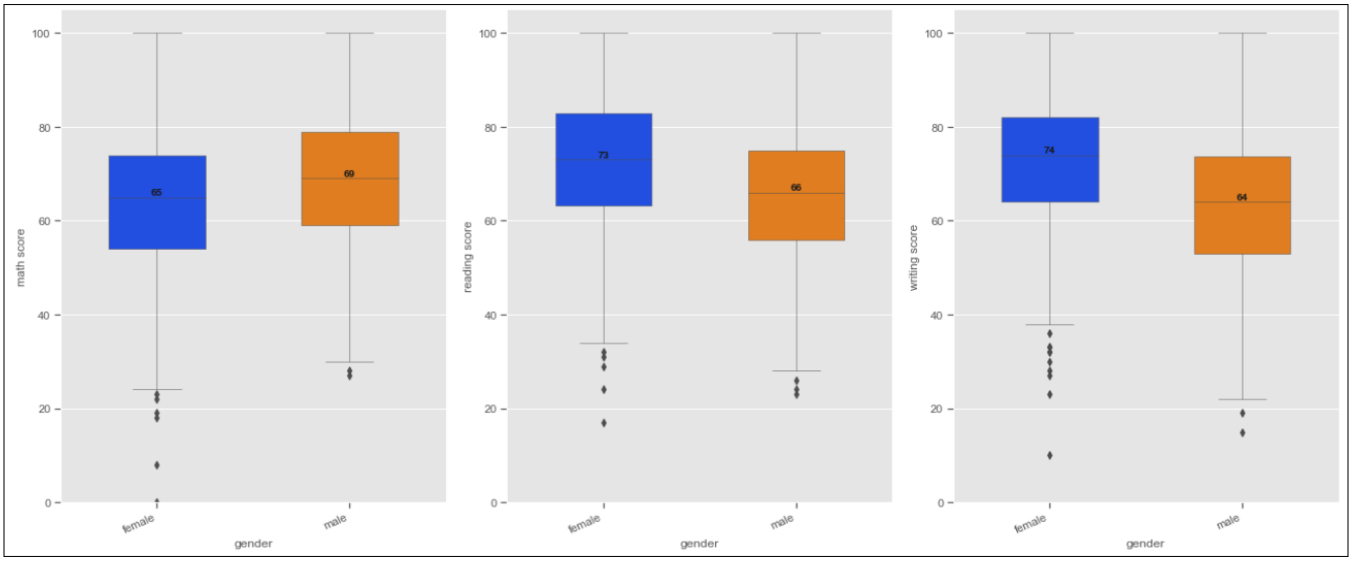


Fig 5: Score distribution based on gender

RQ3: Influencing Factors

To answer RQ3 (What are the visible and influencing factors for the students' score in exam), I will show?), I will present some insights of the score based on different factors.

To start with, let us take a look at the scores distribution for Reading, Writing and Math based on gender (male/female), which is depicted in Fig 5. As we see, on average, females get higher score than males in Reading and Writing, considering the median value and Q2/Q3 quartile in the boxplot. However, in Math, males tend to get better score than female. Therefore, we can say, students exam score differs based on gender.

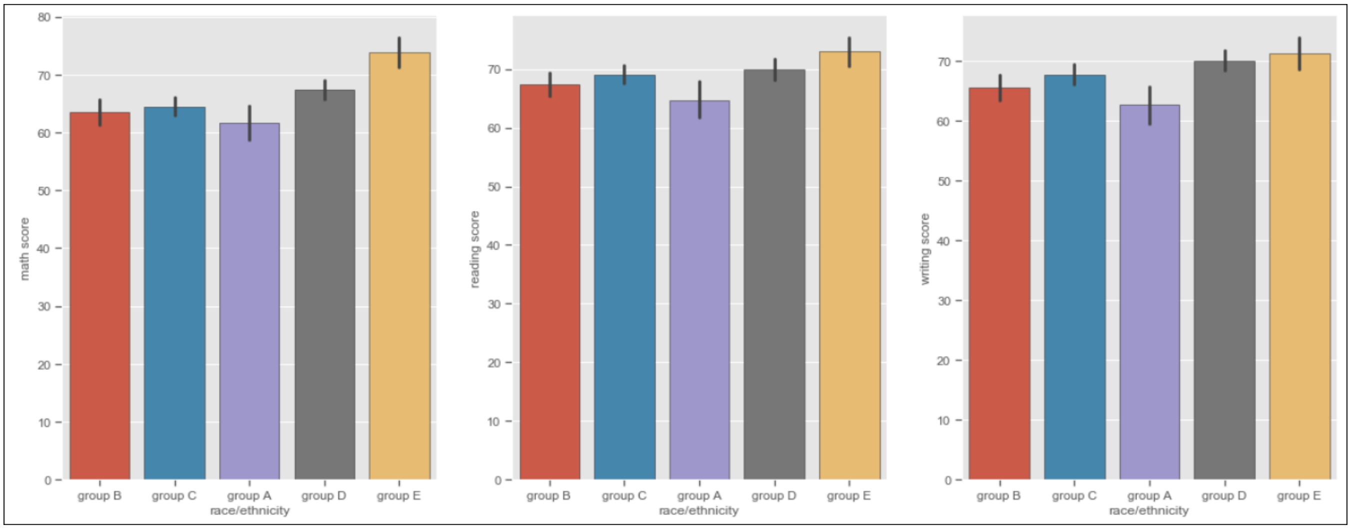


Fig 6: Score in Math, Reading and Writing based on race/ethnicity

Now, we would like to see exam scores in Reading, Writing and Math, based on the race/ethnicity, which is shown in Fig 6. We can students from group E scores better in all three subjects than the other groups. Students in group D also gets better score than the remaining group, except it is only behind than group E. So, in general, score shows some kind of relationship with the students race/ethnicity, where certain ethnicity group are performing better than other groups. This behavior is also true in terms of average score a student achieve in Math, Reading and Writing score, which is shown in the violin plot of Fig 7

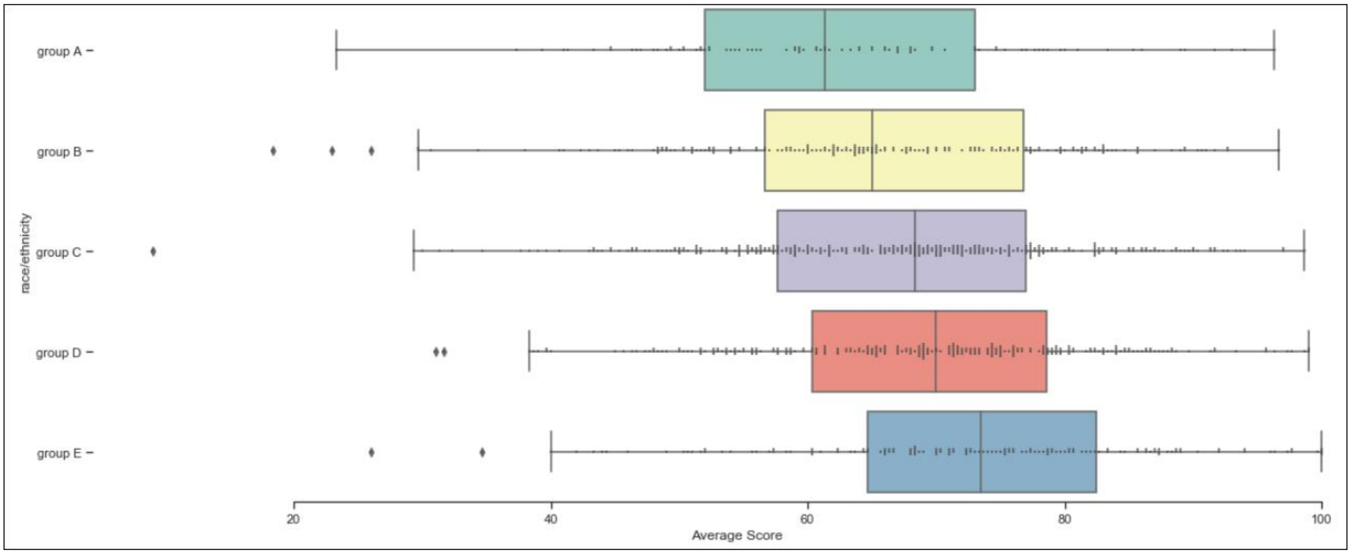


Fig 7: Average scores of students based on race/ethnicity

In Fig 8, I have shown the average scores in Math, Reading and Writing, based on the whether a student completed test preparation or not. We can clearly observer that, the students who completed the test preparation, get higher score in all the subjects, and this difference is higher in Writing. Therefore, completing a test is a crucial factor for students to get higher score in the exam.

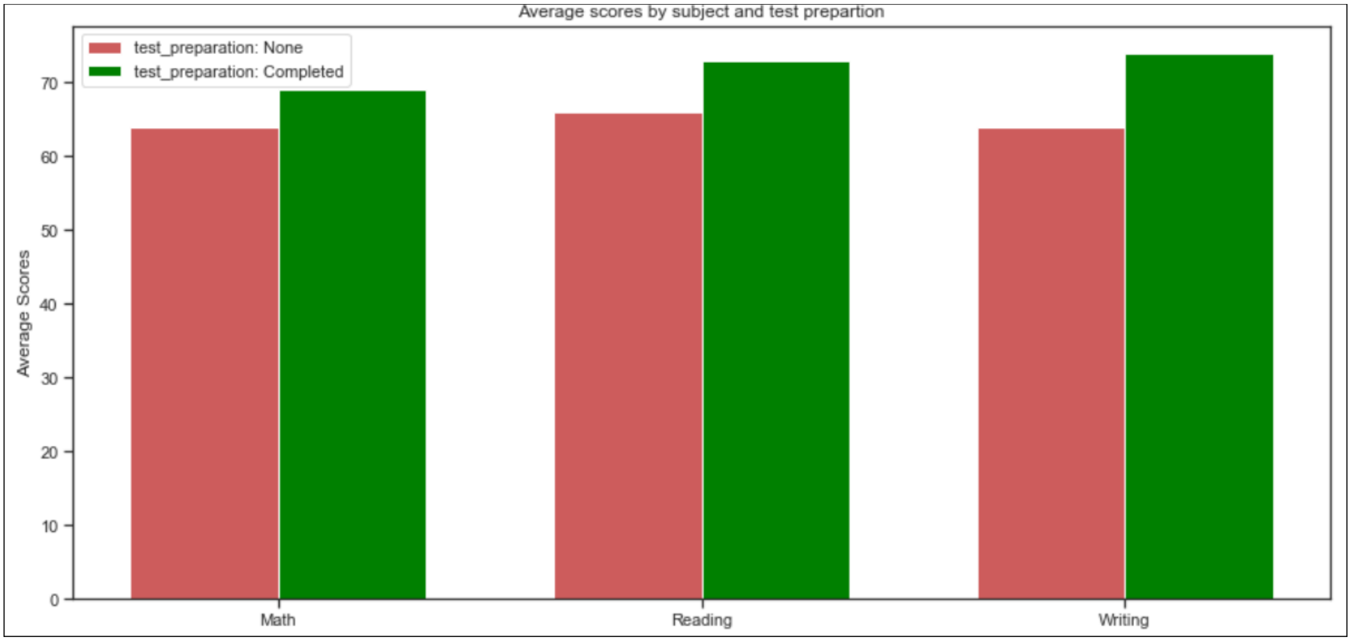


Fig 8: Average scores in Math, Reading and Writing based on test preparation

Another very important factor impacting student's score in the exam is their parents level of education, as depicted in Fig 9. Students whose parents have 'Master Degree' have performed better in all subjects and this behavior can be also observed according to the level of degree, since students have their parents with only a high school degree, performed worst among the others. We can consider from here that the parents who went to university are most likely better able to help their children in their studies. Therefore, we can conclude that student's parents who have a higher degree have a chance of getting better marks.

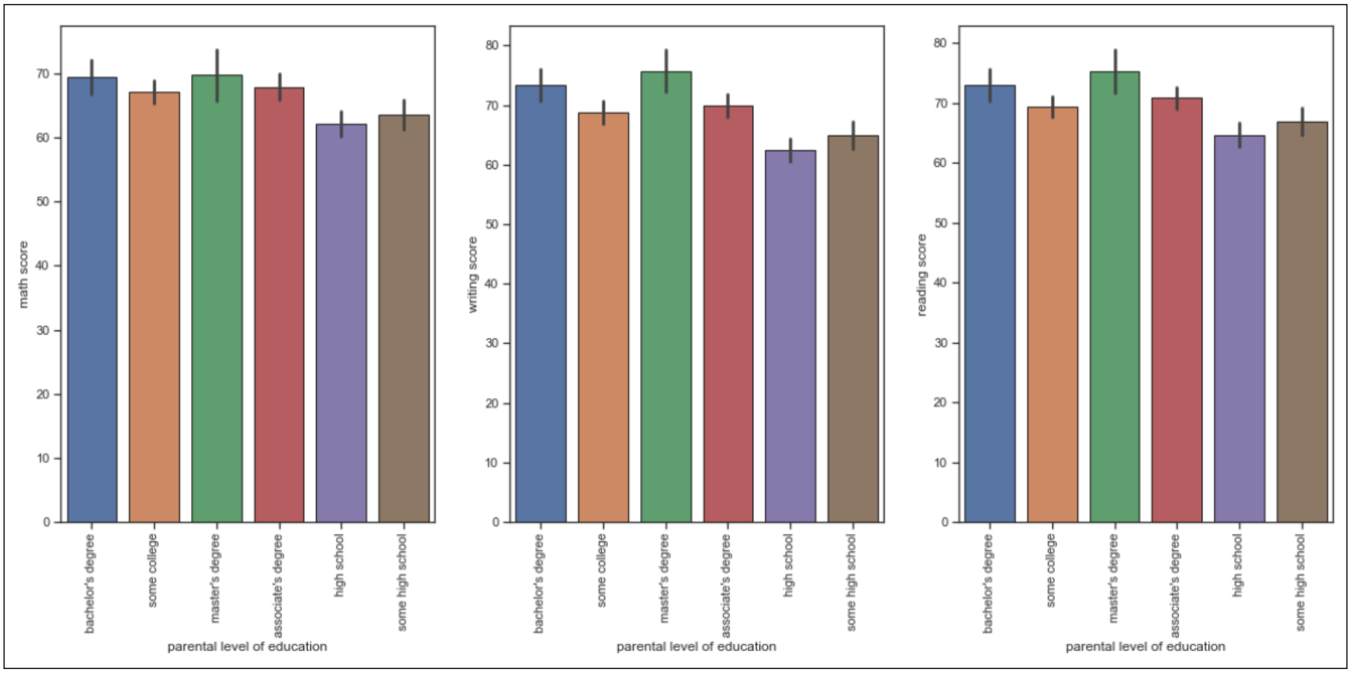


Fig 9: Scores in Math, Reading and Writing based on parents level of education

I also wanted to look into the “lunch” factor for students in getting a better score and it is shown in Fig 10. Similar to the test preparation, it also shows an impact on the score who takes the standard lunch compared to those who take the reduced lunch. The students taking the standard lunch tend to be getting higher scores in exam.

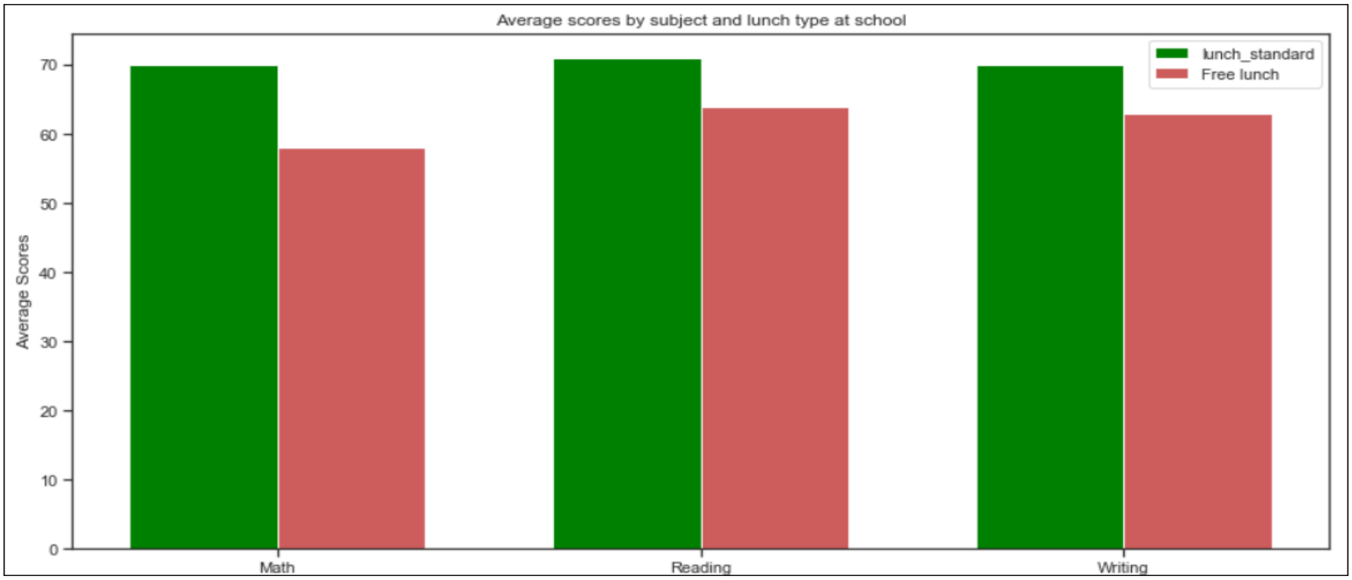


Fig 10: Scores in Math, Reading and Writing based on parents level of education

Now, since we have observed parent's level of education and the type of lunch student used to get have an impact on the students score in exam, I further wanted to see if there is concrete correlation between the parent's education level and the type of lunch their children get at school. Fig 11 shows the findings, where each stacked bar shows the number of students used to take free or standard lunch, for each category of parental education level. However, we do not see a relation between parents' education level and the lunch their children used to take at school. So we can conclude here, although the students score are related to their parents education level and the type of lunch, however, lunch and parents education are not related in terms of the score.

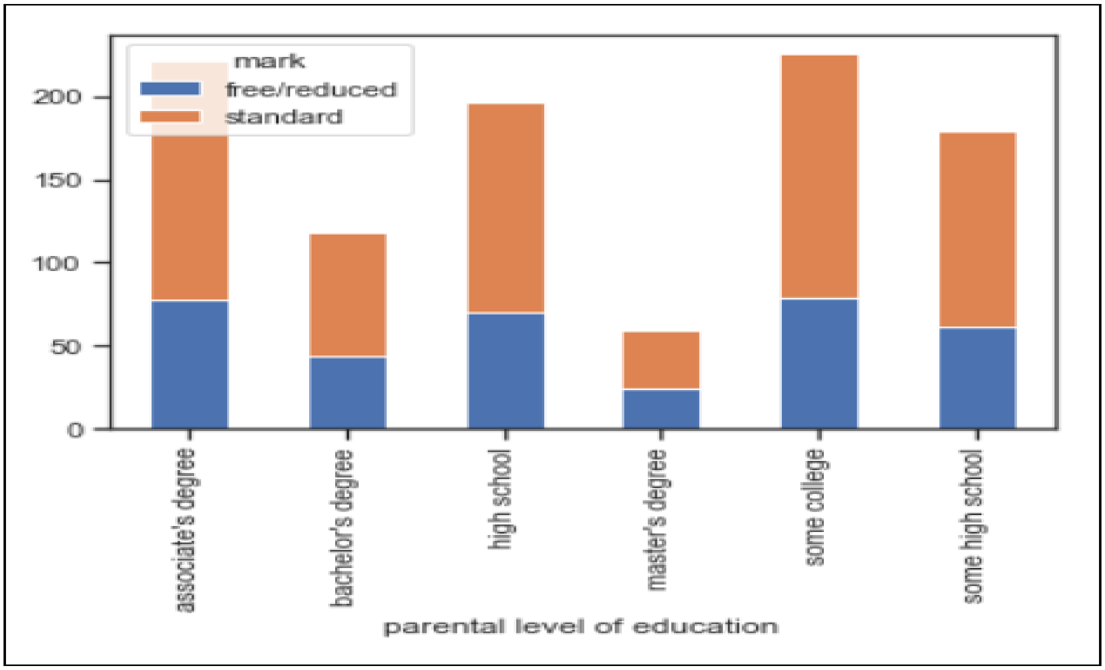


Fig 11: Student's type of lunch based on parent's education level

Conclusion & Lessons Learned:

In this project, I visualized students score in exam in different categories and identified correlation and influencing factors affect the score on the given dataset. We have seen Reading and Writing scores are linearly correlated compared to Reading vs. Math and Writing vs. Math. We have also seen students score is dependent on the gender and race/ethnicity. Parents' level of education and the type of lunch students get at school are major influencing factors in terms of achieving higher score in exam.