Students' Performance in Exams

Introducing the problem

In this article, I will explore the exam results from various students and try to classify the different factors based on their exam scores. By doing this research I will be able to utilize classification algorithms to help better my understanding of the concept and exam performance based on different factors.

Introducing the data

The data utilized for this research is from Kaggle. Kaggle provides numerous available datasets for any kind of research. Specifically, the dataset used for this research is Students' Performance in Exams. This dataset offers a wide range of factors for each result of an exam—factors such as gender, race, and level of education, including a test preparation course. The results were checked through 3 math, reading, and writing exams.

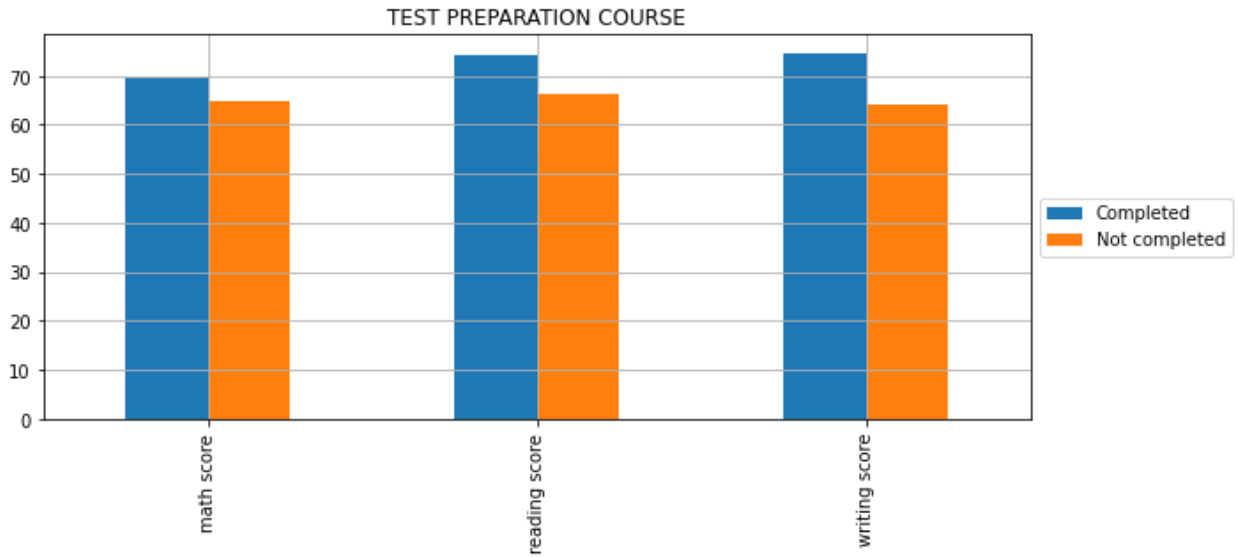
Pre-processing the data

Since this dataset had a relatively short amount of variables, there was no need to process the data. This dataset isn't too long or short nor has any missing values. The research was studied amongst 1000 students. Each data is crucial for looking at the effects on the whole dataset. The type of lunch they receive is just as important as the test preparation course;

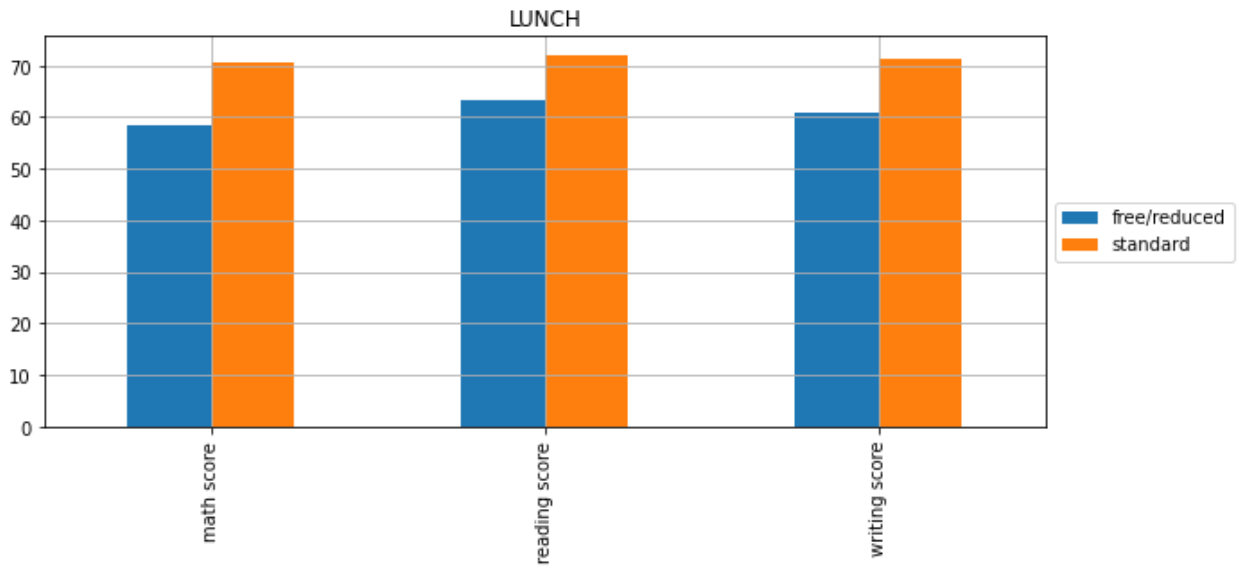
Data visualization

(**Data snippet available on the last page)**

Below are different types of graphs that help visualize the data in a helpful manner. This type of dataset allows for a very large amount of graphs for visualization. In this research I've selects graphs that will easily visualize the dataset. One factor I wanted to look at was the effect of the test preparation course. This graph shows the average test scores made by students who have taken the test preparation course. As the graph shows, students who have taken the course scored more on average.



The graph below shows the scores based on the types of lunch the students had. This is between 'free/reduced lunch' vs 'standard lunch'. We can see that it has a significant effect on the scores of the students.



Modeling

When it came to creating graphs that are usual, I had to refer back to my research question. Which was to see the different factors that affect how students score on exams. Also to classify the factors and utilize the algorithms we learned. In order to see the classification between the different factors, I used bar graphs. Attached below are snippets of the code utilized to create these graphs. In addition to bar graphs, a scatterplot is also useful in helping us visualize the distribution.

Evaluation

These two main models work well enough to show us the disparities created by the different factors. We see that there's a significant effect of both Lunches and whether they took the test preparations course or not. Looking at the scatterplots for the test preparations course, we can see that a higher percentage of students who took the course, scored higher. While the majority of the students who did not are marked with lower scores. This evaluations lead me to agree with my research assumptions, that

factors such as preparation course, lunch, parental level of education and etc are significant to test results.

Storytelling

Through this project, I've learned that there are many factors that affect test results. One surprising fact I learned was from the Free or standard lunch dataset. I never thought that the type of meal plan affects how well you perform in exams. This leads me to the analysis that there is more pressure or "seriousness" taken when education is costly.

Impact

I believe the impact of my project on readers, would be to reinforce the idea of effort into exams or anything education related. Based on the data, every factor is visible. I believe other factors such as sleep, time on the phone, and time spent studying are also relevant to the topic.

References

- EDA on students' performances

Code

In [5]:
```
data
```

Out[5]:

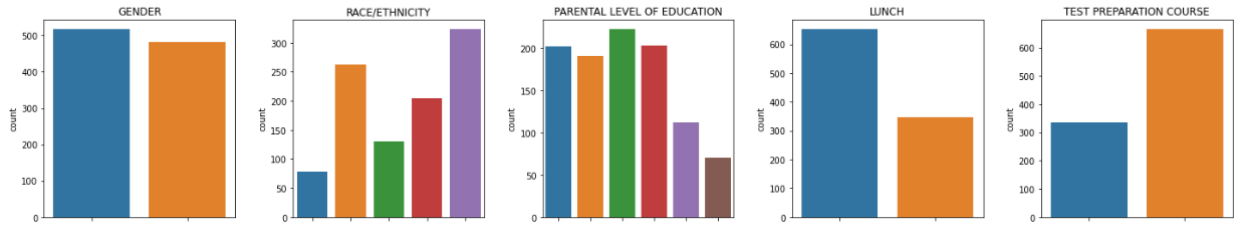
	gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score
0	male	group A	high school	standard	completed	67	67	63
1	female	group D	some high school	free/reduced	none	40	59	55
2	male	group E	some college	free/reduced	none	59	60	50
3	male	group B	high school	standard	none	77	78	68
4	male	group E	associate's degree	standard	completed	78	73	68
...	...	...	...	...	...	...	...	...
995	male	group C	high school	standard	none	73	70	65
996	male	group D	associate's degree	free/reduced	completed	85	91	92
997	female	group C	some high school	free/reduced	none	32	35	41
998	female	group C	some college	standard	none	73	74	82
999	male	group A	some college	standard	completed	65	60	62

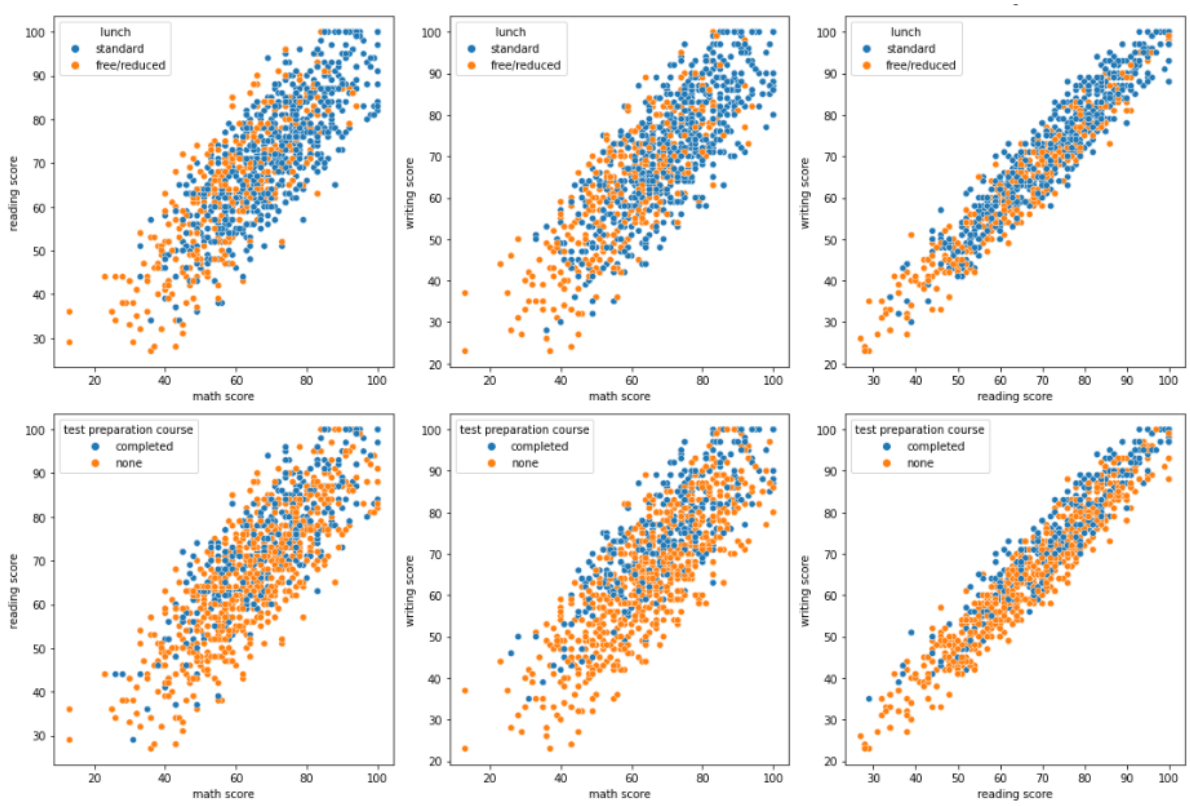
1000 rows × 8 columns

```
fix, ax = plt.subplots(figsize=(20,5), nrows=1, ncols=5)
cat_cols = ['gender', 'race/ethnicity', 'parental level of education', 'lunch', 'test preparation course']
num_cols = ['math score', 'reading score', 'writing score']

for i,col in enumerate(cat_cols):
    sns.countplot(data=data, x=col, ax=ax[i])
    ax[i].tick_params(axis='x', rotation=45)
    ax[i].set_title(col.upper())

plt.tight_layout()
plt.show()
```





In [16]:
```
num_pairs = [[num_cols[0], num_cols[1]],
             [num_cols[0], num_cols[2]],
             [num_cols[1], num_cols[2]]]
fix, ax = plt.subplots(figsize=(15,25), nrows=5, ncols=3)

for i,p in enumerate(num_pairs):
    for j, hue in enumerate(cat_cols):
        sns.scatterplot(data=data, x =p[0], y=p[1], ax=ax[j,i], hue=hue)
plt.tight_layout()
plt.show()
```