Exam Performance Project

PAYTON FALCONE AND ETHAN WU

Abstract

Purpose: To see the connections between student's grade and other parameters including but not limited to test preparation, parental level of education, etc.

Hypothesis: We believe students that prepare for the test should score higher overall than students who do not prepare.

Dataset

	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
369	female	group A	high school	standard	none	70	75	74
376	female	group A	high school	standard	none	85	91	89
388	male	group A	associate's degree	free/reduced	completed	82	82	78
399	female	group A	some college	free/reduced	completed	53	70	69
...	...	...	...	...	...	...	...	...
752	male	group E	high school	standard	none	70	62	61
777	male	group E	associate's degree	standard	none	78	72	68
783	male	group E	associate's degree	free/reduced	completed	63	60	63
273	female	group E	high school	standard	none	93	100	93
536	male	group E	some college	free/reduced	completed	83	63	63

Dataset Cleaning

```
# Checks if there is any NULL and NA values
print(examSet.isnull().sum(), "\n")
print(examSet.isna().sum())
```

```
gender                         0
race/ethnicity                 0
parental level of education    0
lunch                          0
test preparation course        0
math score                     0
reading score                  0
writing score                  0
dtype: int64

gender                         0
race/ethnicity                 0
parental level of education    0
lunch                          0
test preparation course        0
math score                     0
reading score                  0
writing score                  0
dtype: int64
```

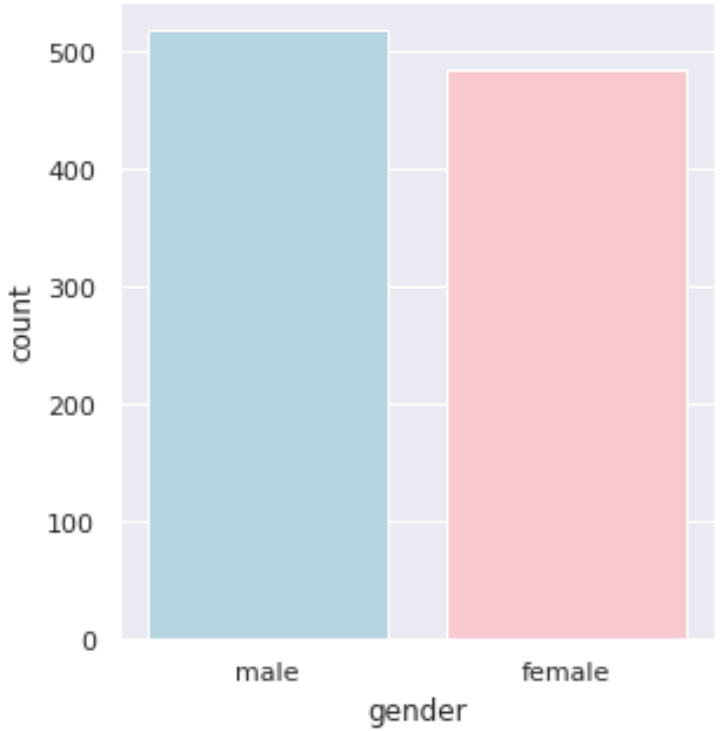
Dataset Additions

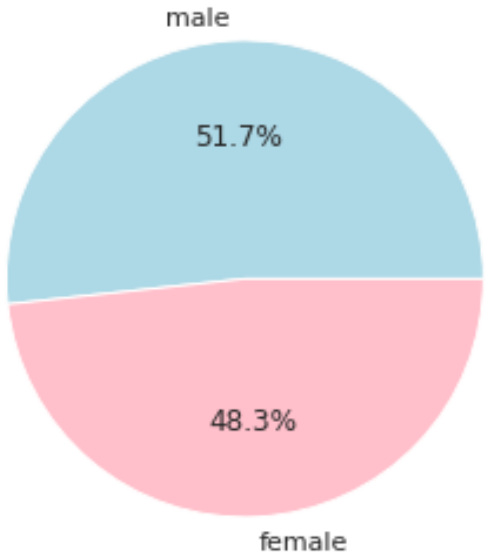
```
examSet["avgScore"] = examSet[["math score", "reading score", "writing score"]].mean(axis=1)
```

	gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score	avgScore
0	male	group A	high school	standard	completed	67	67	63	65.666667
369	female	group A	high school	standard	none	70	75	74	73.000000
376	female	group A	high school	standard	none	85	91	89	88.333333
388	male	group A	associate's degree	free/reduced	completed	82	82	78	80.666667
399	female	group A	some college	free/reduced	completed	53	70	69	64.000000
...	...	...	...	...	...	...	...	...	...
752	male	group E	high school	standard	none	70	62	61	64.333333
777	male	group E	associate's degree	standard	none	78	72	68	72.666667
783	male	group E	associate's degree	free/reduced	completed	63	60	63	62.000000
273	female	group E	high school	standard	none	93	100	93	95.333333
536	male	group E	some college	free/reduced	completed	83	63	63	69.666667

Core Questions

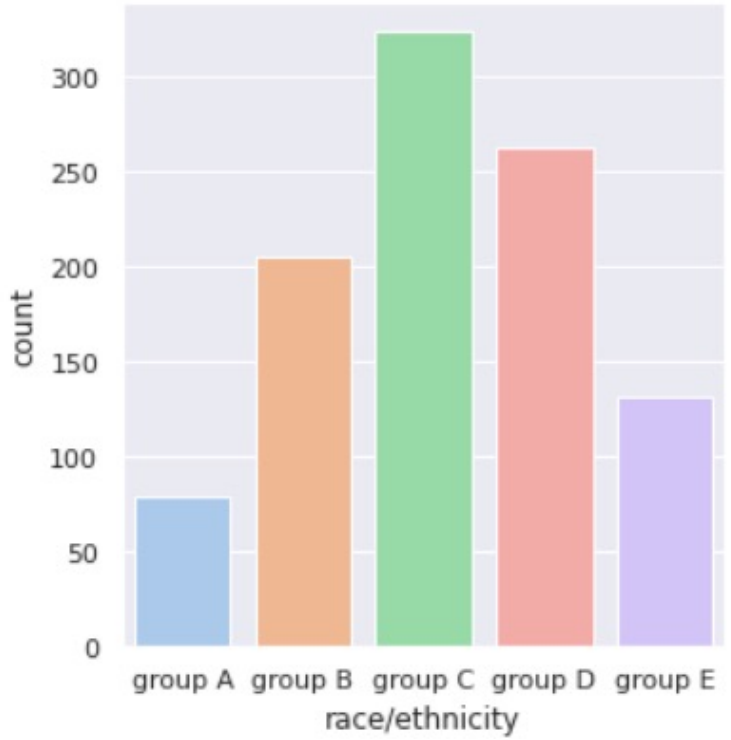
- What is the distribution between all scores?
- What is the correlation between grades and gender?
- What is the correlation between grades and parent's level of education?
- What is the correlation between grades and groups?
- What is the correlation between grades and test preparation?
- What is the correlation between grades and lunch?

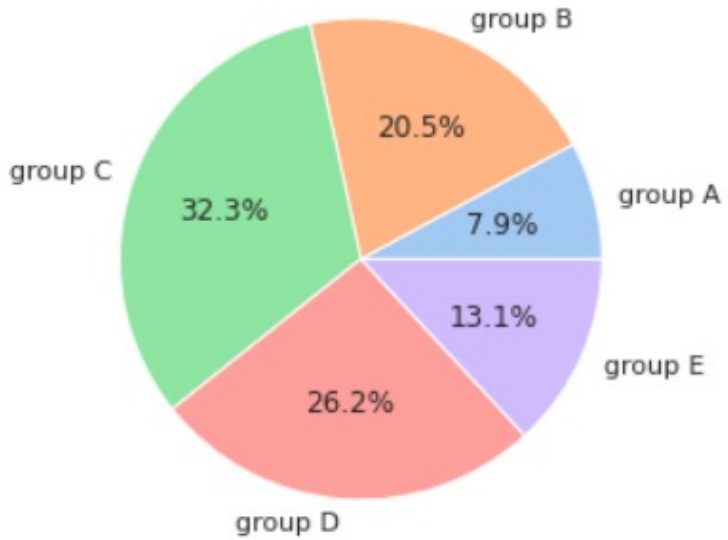




gender	
male	517
female	483

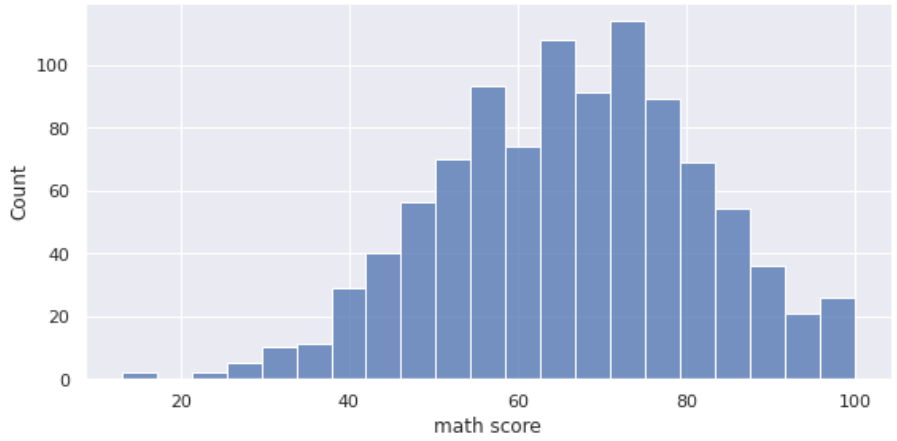
Diversity in Data

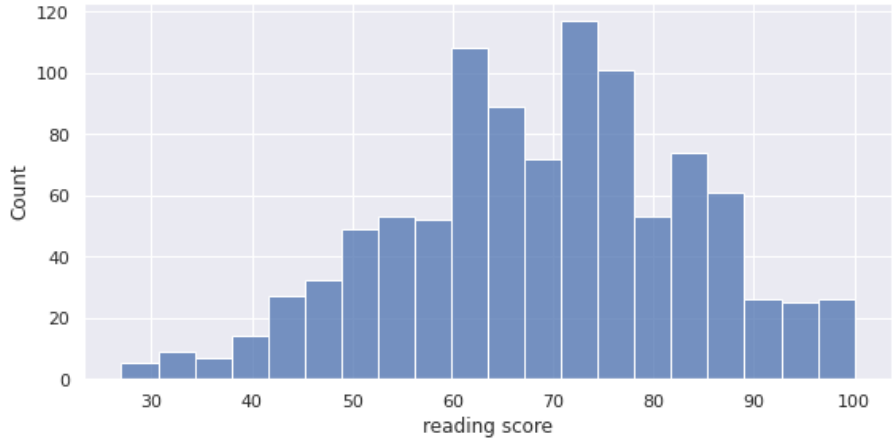


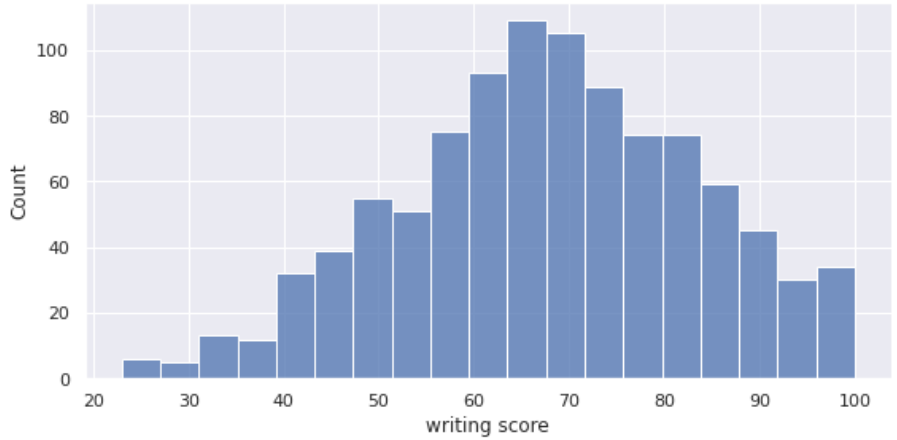


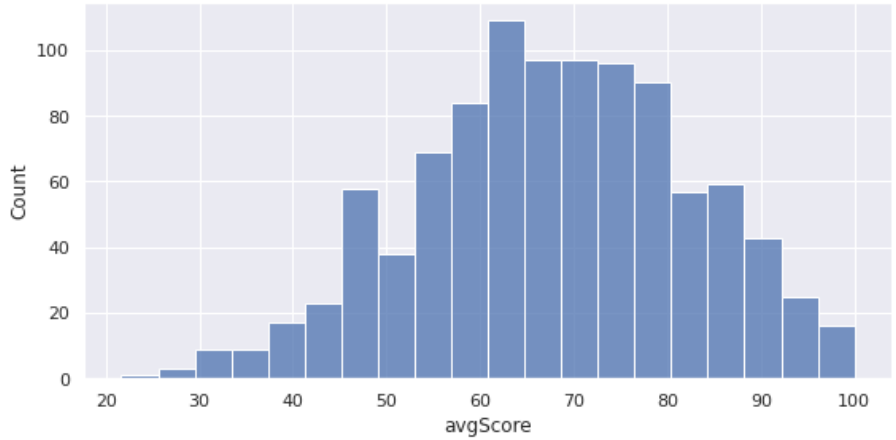
race/ethnicity	
group A	79
group B	205
group C	323
group D	262
group E	131

Diversity in Data

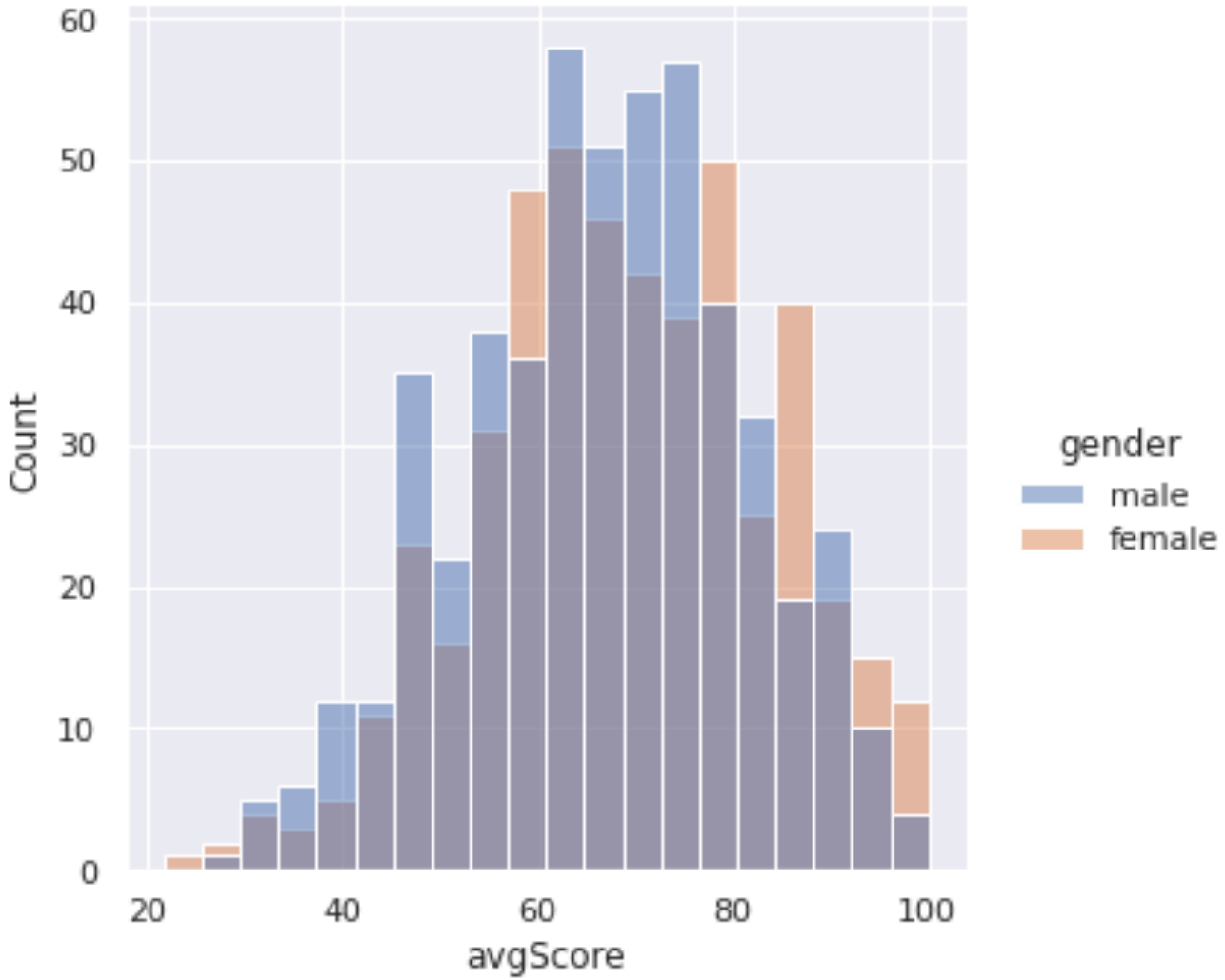




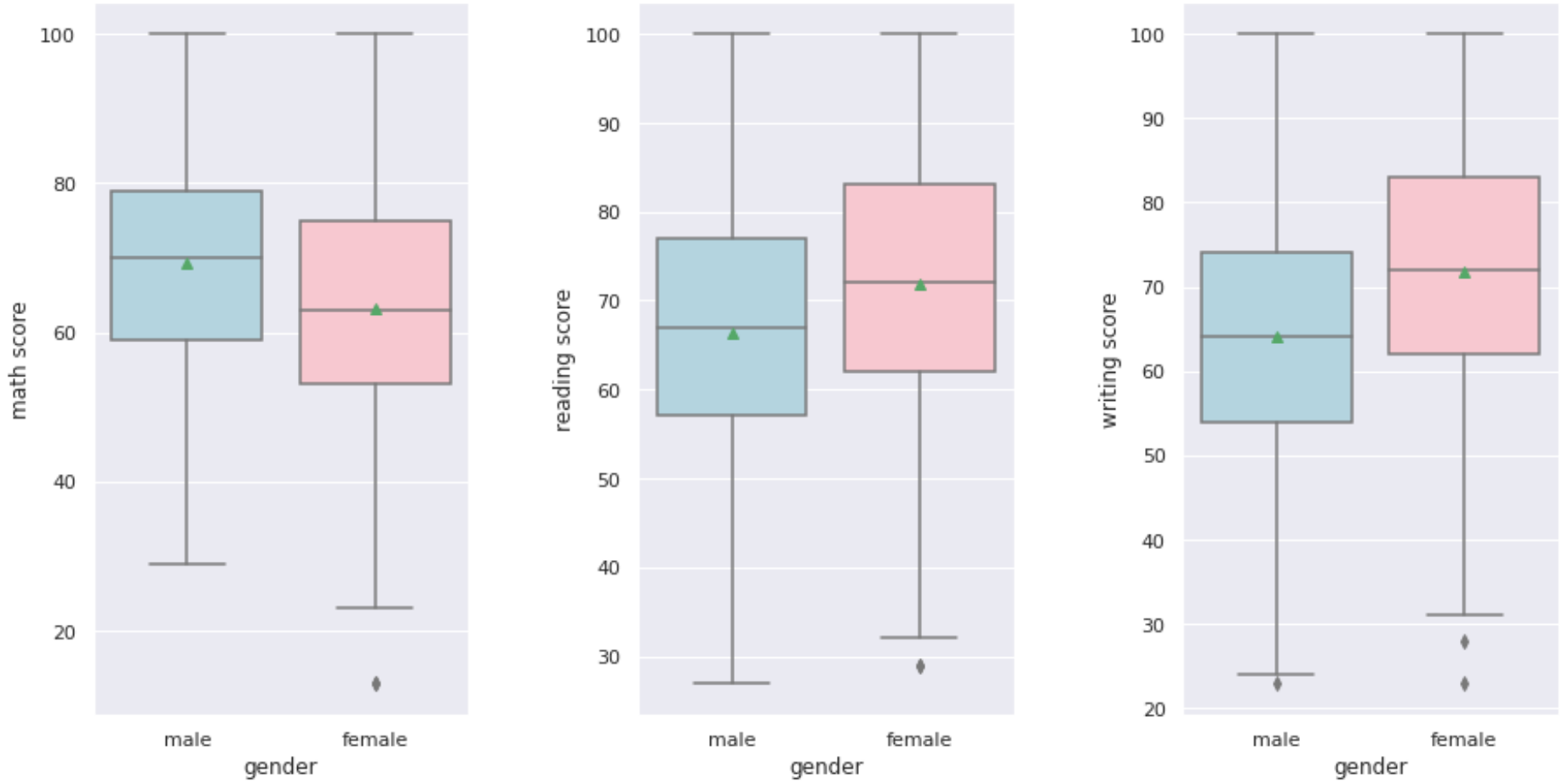




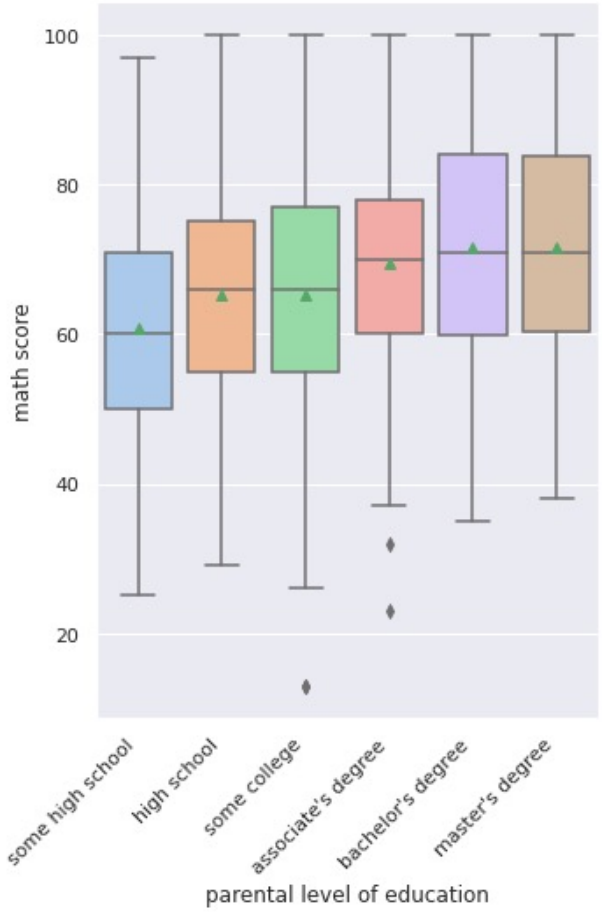
Distribution between Scores

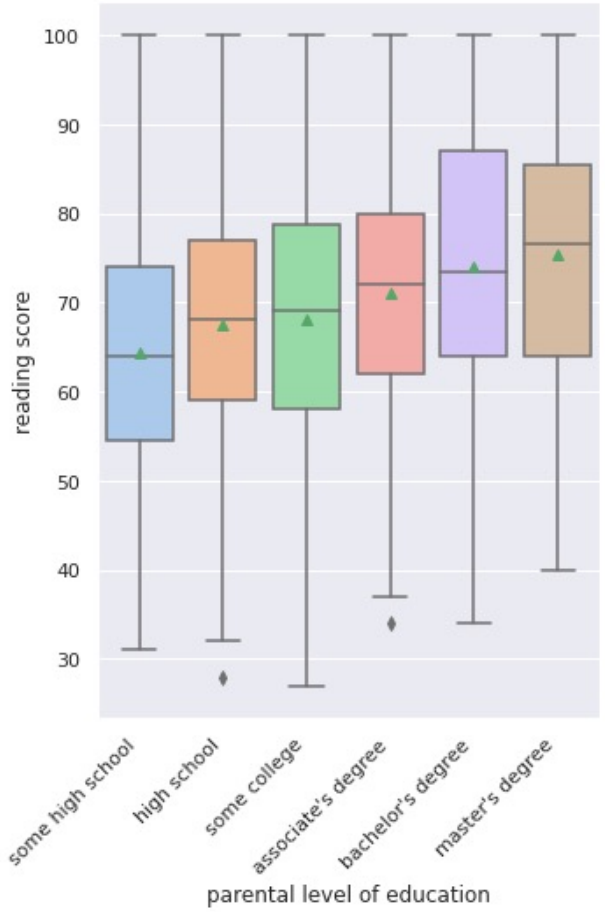


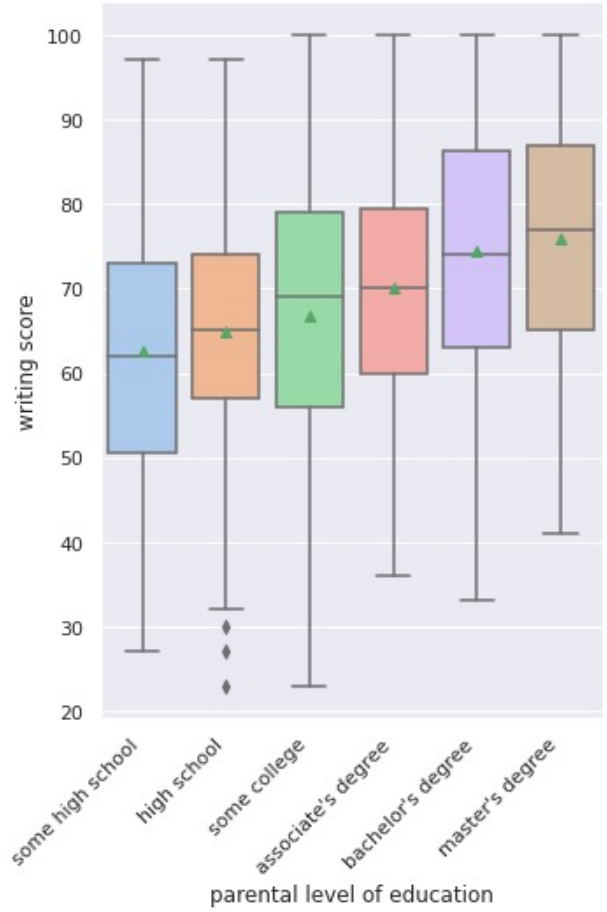
Average Score between Genders



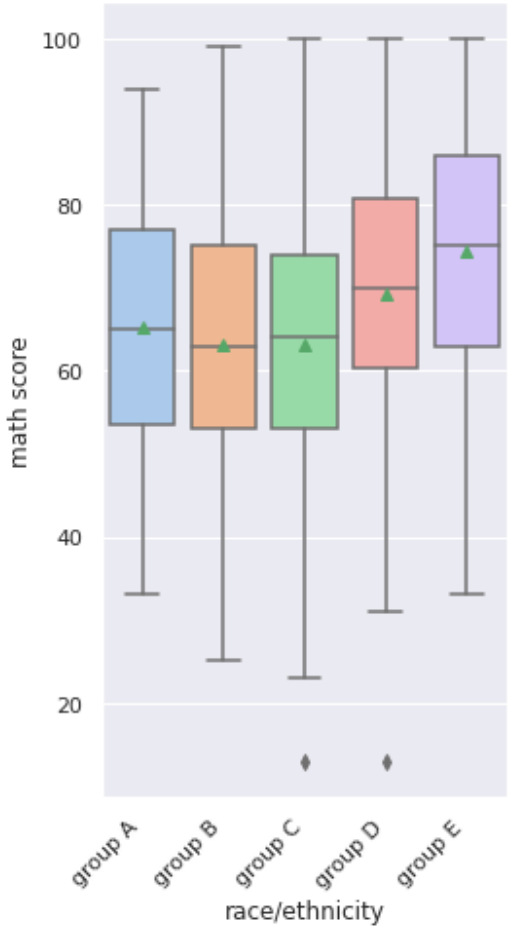
Gender and Scores

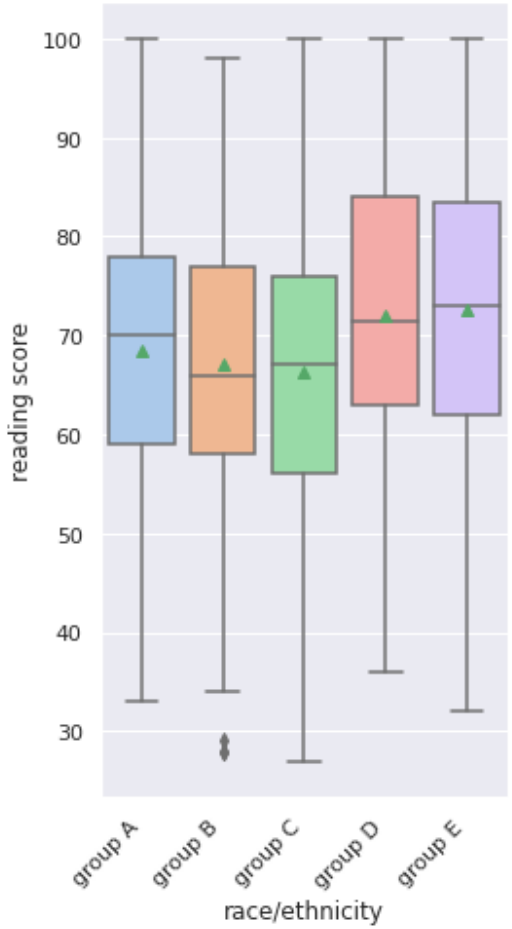


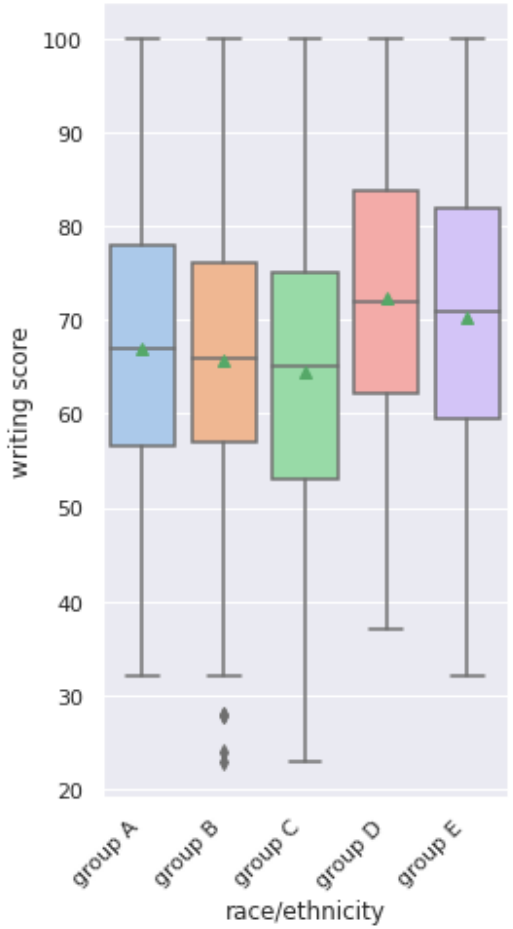


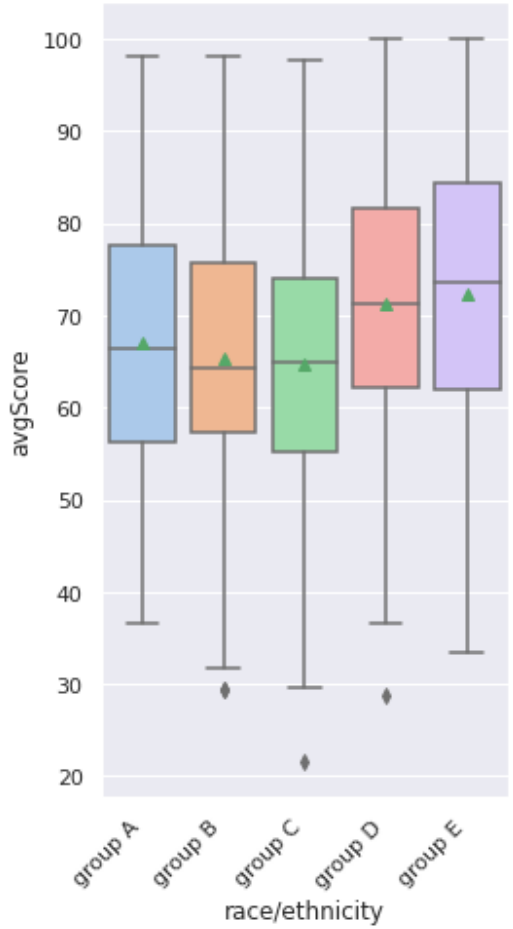


Parent Level of Education and Scores

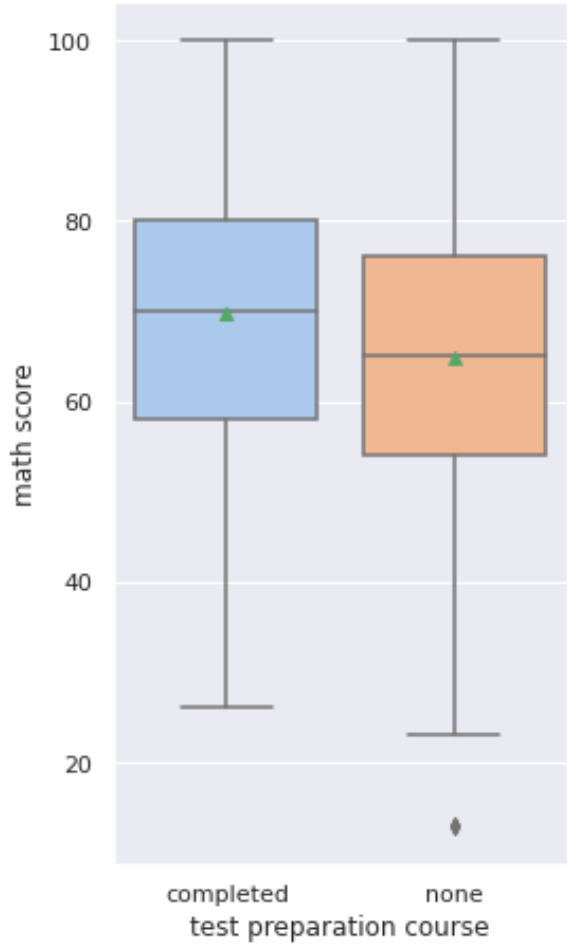


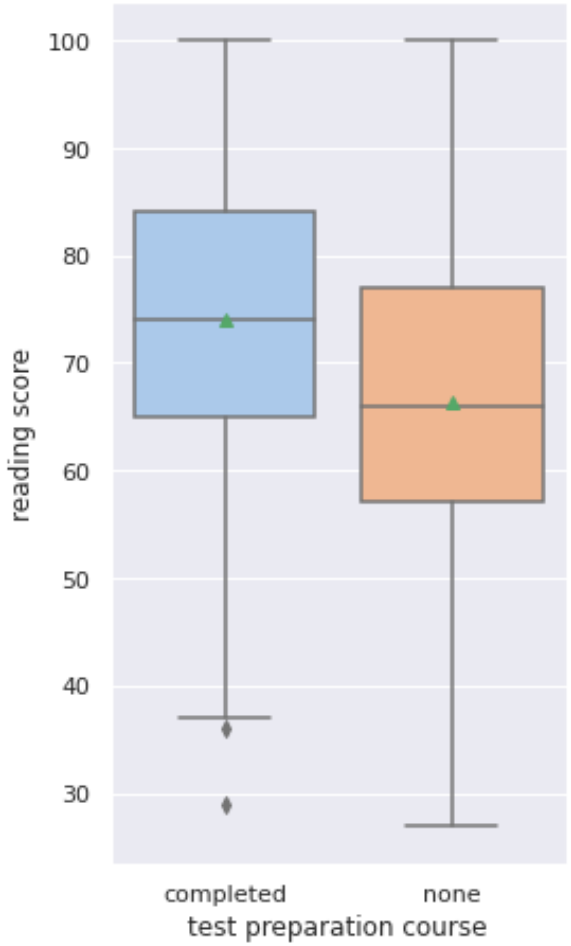


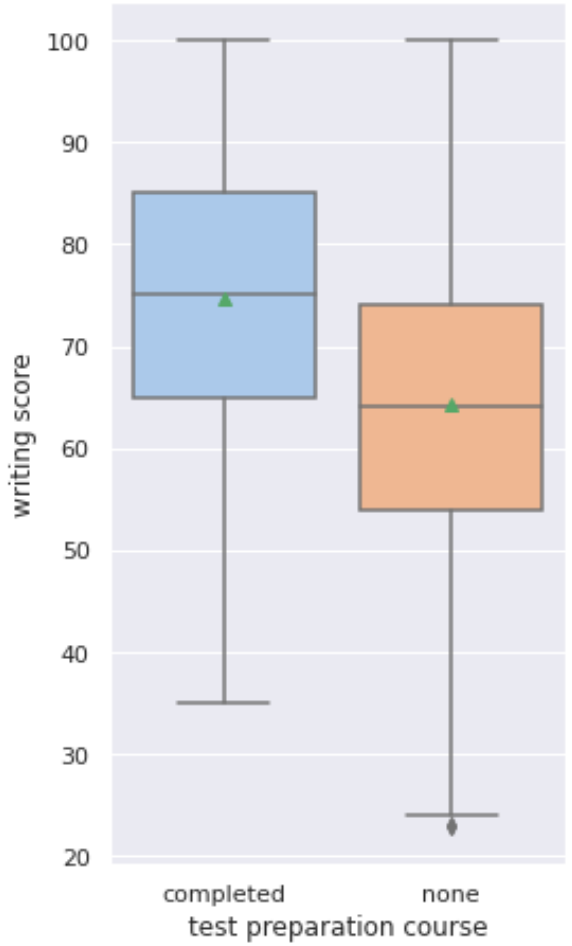




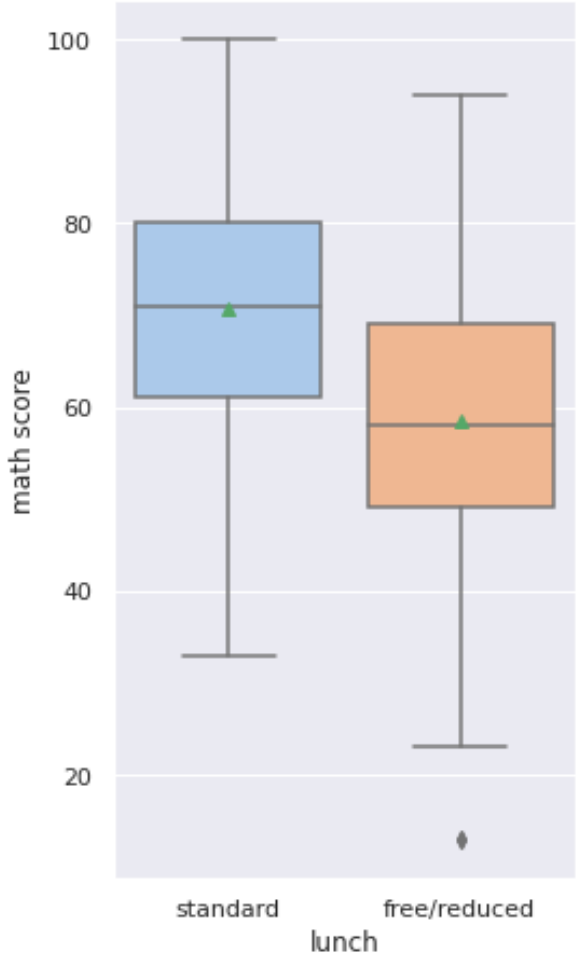
Groups and Scores

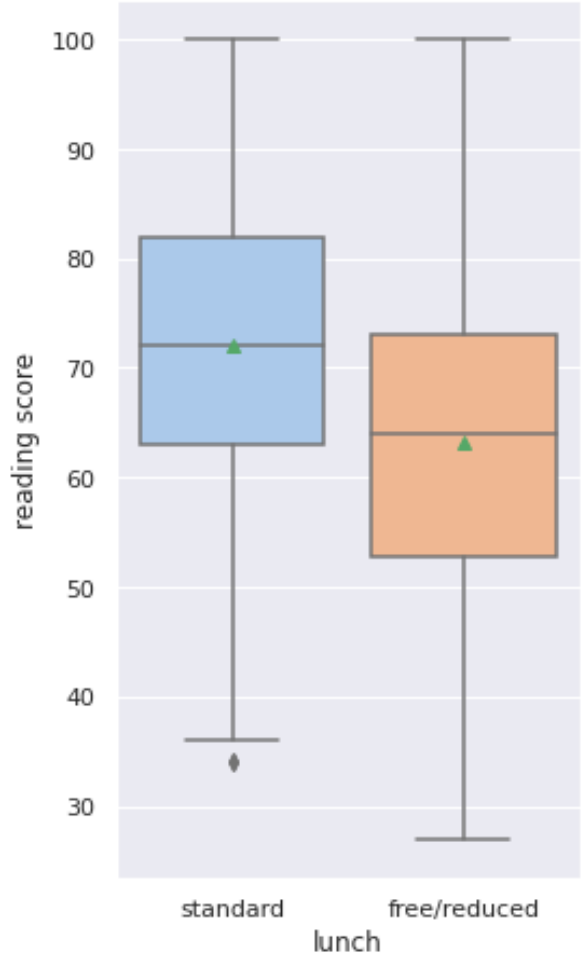


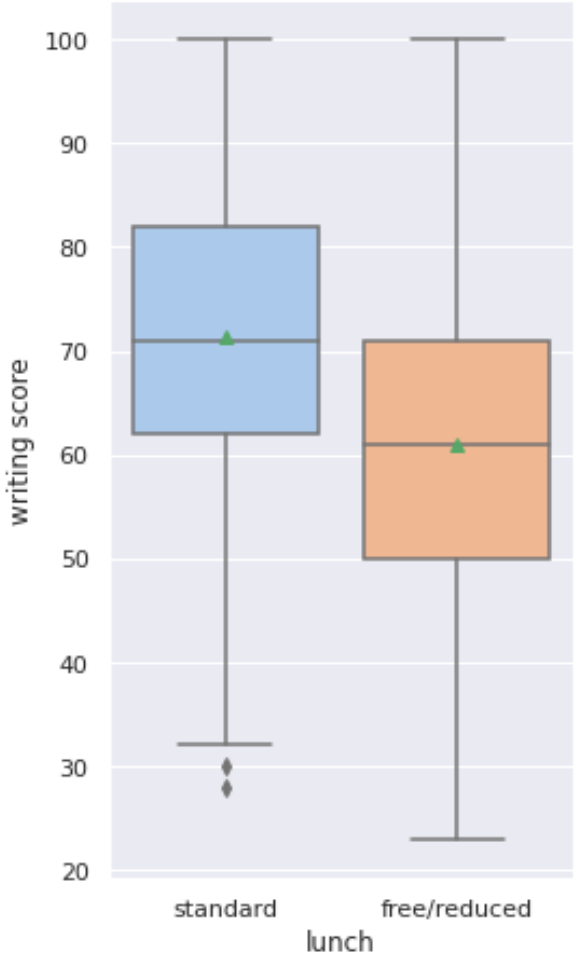




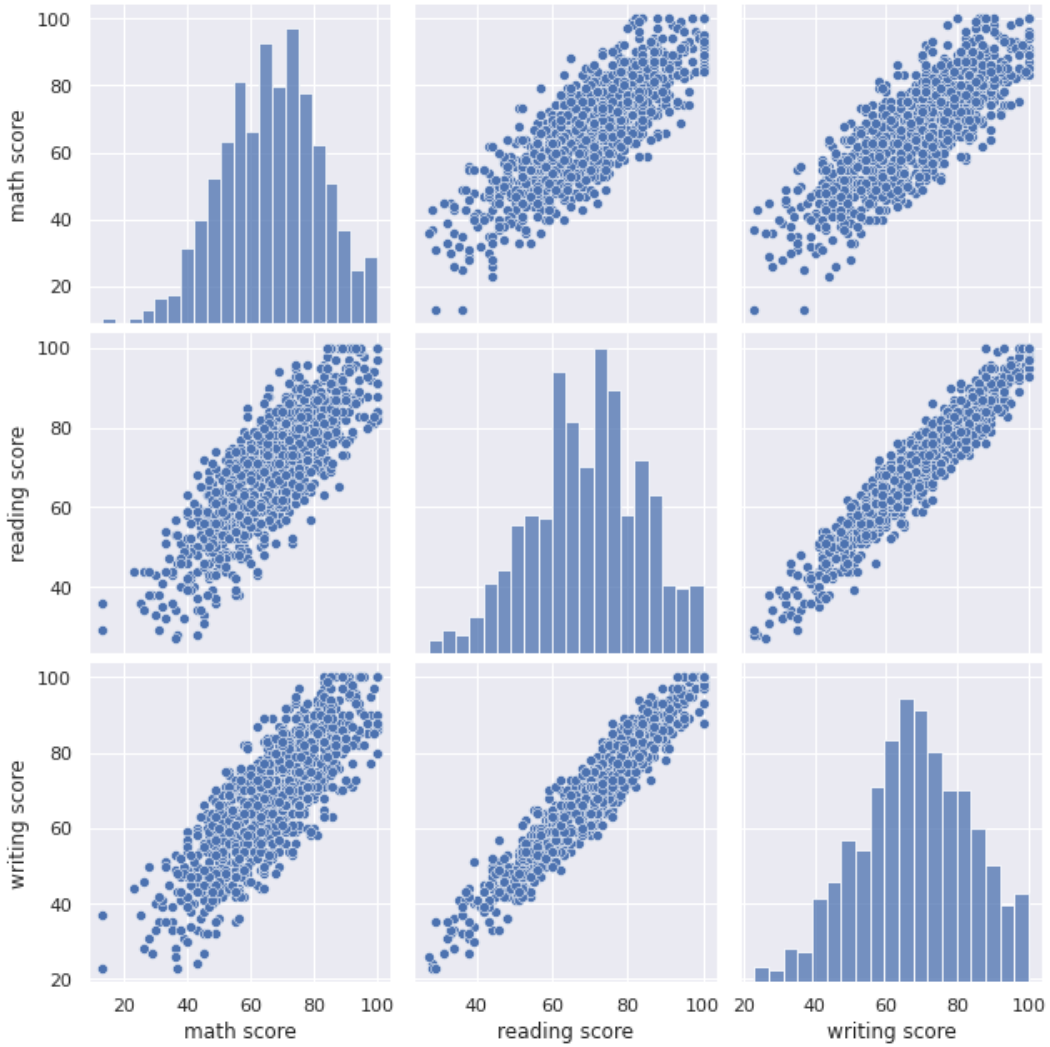
Test Preparation and Scores



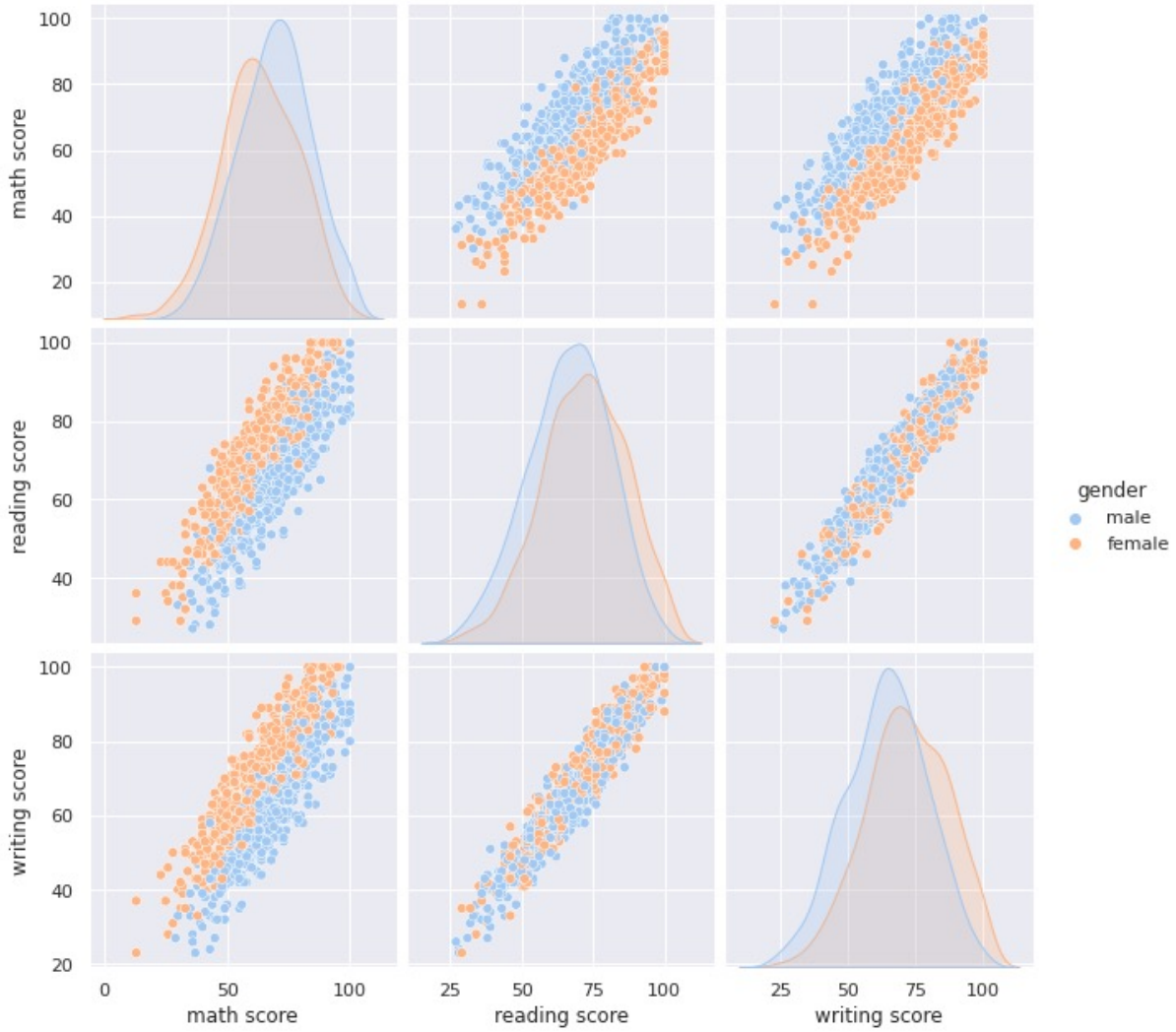




Lunch and Scores

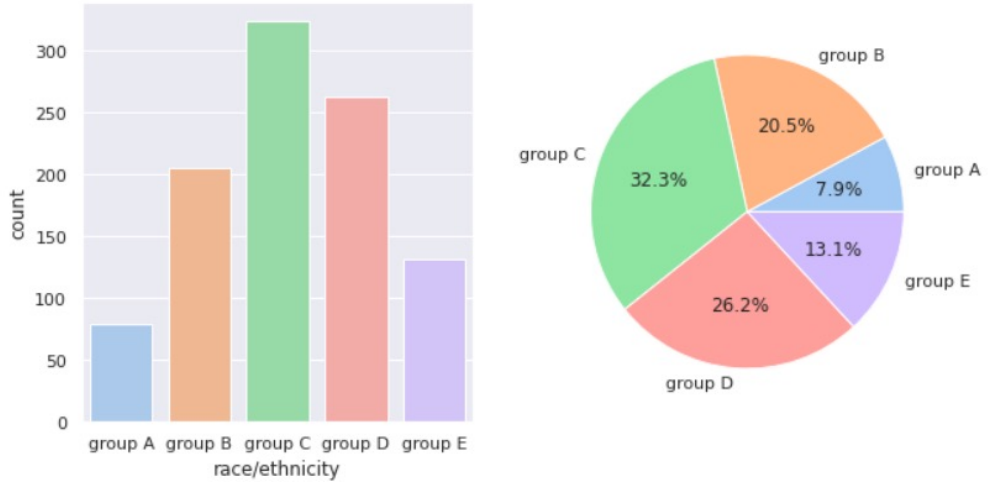


Correlation between Scores



Based on Gender

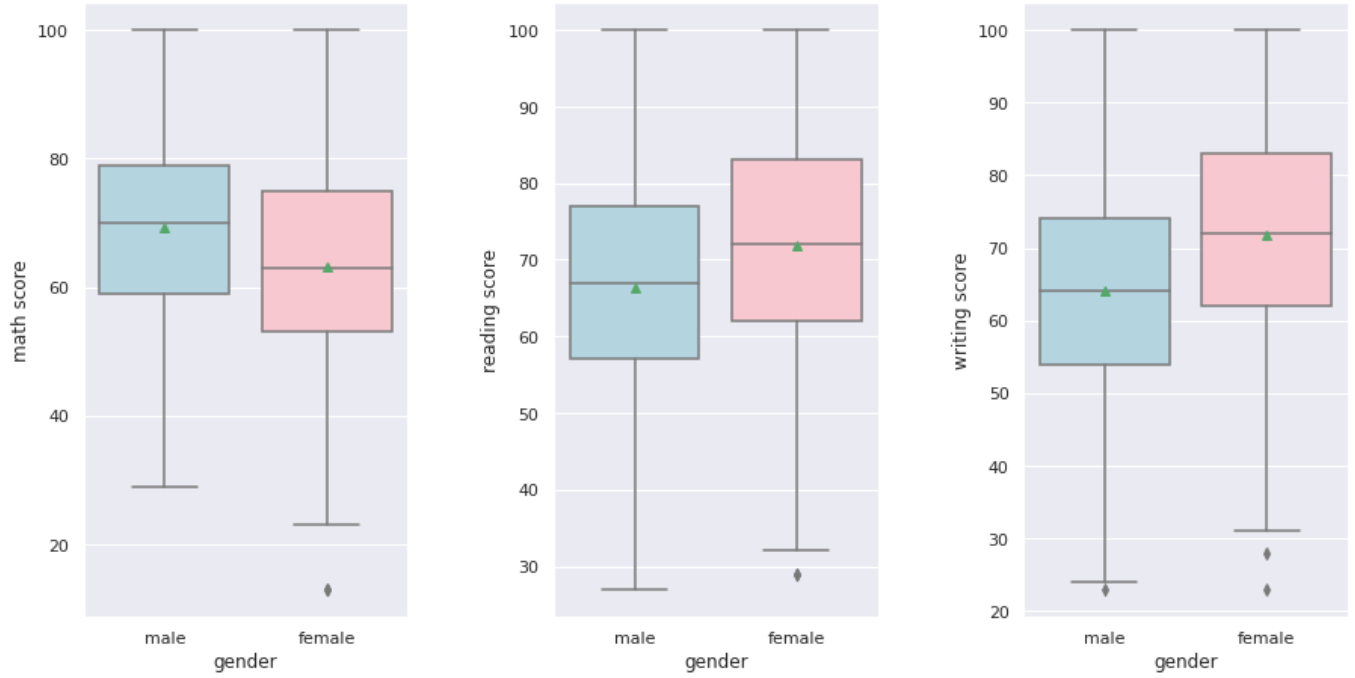
```
fig, ax = plt.subplots(1,2, figsize=(10,5))
sns.countplot(data=examSet,x="race/ethnicity", palette="pastel", ax= ax[0])
groupCount = examSet.value_counts("race/ethnicity").sort_index(0)
ax[1] = plt.pie(groupCount, labels=groupCount.index, colors=sns.color_palette("pastel"), autopct="%.1f%%")
plt.show()
```



Count Plot and Pie Chart Code Snippet

```
# Correlation between Gender and their Grades
fig, ax = plt.subplots(1,3, figsize=(15,7.5))
plt.subplots_adjust(wspace= 0.45)
sns.boxplot(data=examSet, x="gender", y="math score", palette=["lightblue","pink"], showmeans=True, ax=ax[0])
sns.boxplot(data=examSet, x="gender", y="reading score", palette=["lightblue","pink"], showmeans=True, ax=ax[1])
sns.boxplot(data=examSet, x="gender", y="writing score", palette=["lightblue","pink"], showmeans=True, ax=ax[2])

plt.show()
```



Box Plot Code Snippet

Conclusions

- Those with parents that had higher levels of education *usually* scored higher than those with parents of lower education
- Those who had a standard lunch *usually* scored higher than those who had a free / reduced lunch
- Those who prepared for the exam *usually* scored higher than those who did not prepare
- All test scores were proportional to each other (ex: as math scores increase, so does reading)

References/Libraries

Lecture Jupyter Notebooks

Seaborn Documentation: https://seaborn.pydata.org/tutorial.html

Matplotlib Documentation: https://matplotlib.org/stable/tutorials/index.html

Pandas Documentation: https://pandas.pydata.org/docs/

NumPy Documentation: https://numpy.org/doc/

Kaggle Dataset: https://www.kaggle.com/datasets/whenamancodes/students-performance-in-exams

GitHub Repository: https://github.com/wue1atwit/ExamDatatset