بسم الله الرحمن الرحیم

مشکل چیست؟

در این مسئله عواملی که بر نتیجه ی 3 امتحان مختلف تاثیر دارد بررسی میشود و با توجه به نتیجه ی به دست آمده عواملی که تاثیر بیشتری روی نتیجه ی امتحانات دارد را بررسی کرده و بر این اساس برای بهبود نتایج امتحانات دانش اموزان از این فاکتور ها بهره می بریم.

سوال داده کاوی که برای حل مشکل مطرح شده چیست ؟(قرار است چه ستون یا اطلاعاتی را پیش بینی کنید؟قصد دارید از چه اقلام اطلاعاتی برای این پیش بینی استفاده کنید؟)

در این مسئله ستون مشخصی برای پیش بینی وجود ندارد و تنها عوامل موثر بر نتیجه ی امتحانات بررسی میشوند اما میتوان ستونی مبنی بر پاس شدن یا افتادن دانش اموز در درس را اضافه کرد و با استفاده از مدل مشخصی که از داده های اموزشی به دست می اید و تاثیر عوامل موجود در دیتاست بر نتیجه ی امتحان را بررسی میکند، این پیش بینی را انجام داد که با توجه به مقادیری که به فاکتورهای موثر بر نتیجه داده میشود، پاس شدن یا افتادن دانش اموز پیش بینی میشود.

داده های مسئله:

این دیتاست 3 ستون نمره شامل math score , reading score , writing score دارد و شامل ستون هایی به عنوان فاکتورهای موثر بر نتیجه ی امتحان است به عنوان مثال عوامل بررسی شده در این دیتاست شامل : امادگی شخص قبل از امتحان ، وضعیت ناهار(ناهار استاندارد خورده باشد یا ناهار نخورده باشد یا ناقص باشد) ، تحصیلات والدین دانش اموز، جنسیت ، قومیت

نحوه ی حل مسئله:

از انجاییکه هدف از این مسئله بررسی فاکتورهای موثر روی نتیجه ی امتحانات است ابزار اصلی مورد استفاده ی ما رسم نمودارها خواهد بود که با رسم نمودارهای مختلف ابعاد مختلف را بررسی و تاثیر عوامل مختلف را روی نتیجه هر امتحان به تفکیک و به طور کلی روی میانگین نمرات انها بررسی میکنیم .

در نهایت با بررسی مدل های مختلف پیش بینی، بهترین مدل را انتخاب و به منظور پیش بینی پاس شدن یا افتادن فرد در درس ان را روی داده ها اعمال میکنیم.

overview on dataset

```
df.head()
```

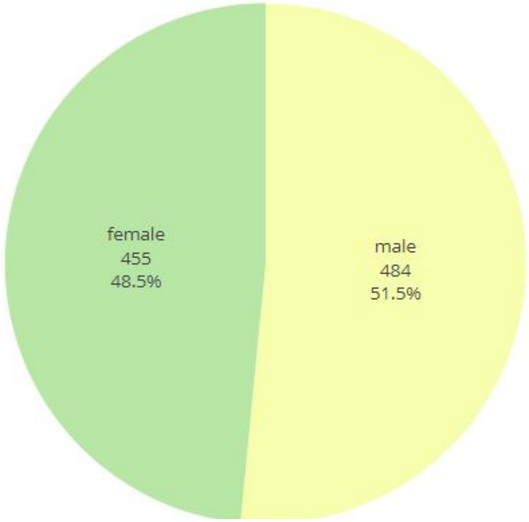
	gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score
0	female	group D	some high school	free/reduced	none	57	76	69
1	male	group D	high school	free/reduced	none	39	40	40
2	female	group C	some college	standard	none	66	67	66
3	female	group E	high school	standard	none	61	73	74
4	male	group A	some high school	standard	none	48	44	45

	math score	reading score	writing score
count	1000.000000	1000.000000	1000.000000
mean	66.476000	69.584000	68.480000
std	15.249064	14.447688	15.228575
min	12.000000	15.000000	20.000000
25%	56.000000	60.000000	58.000000
50%	67.000000	70.000000	68.500000
75%	78.000000	80.000000	79.000000
max	100.000000	100.000000	100.000000

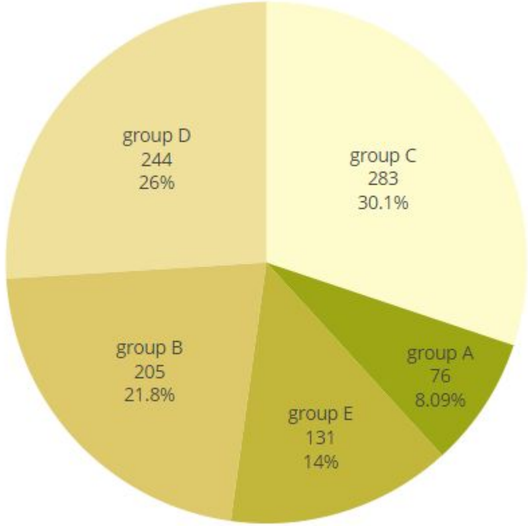
Visualization

count of each feature

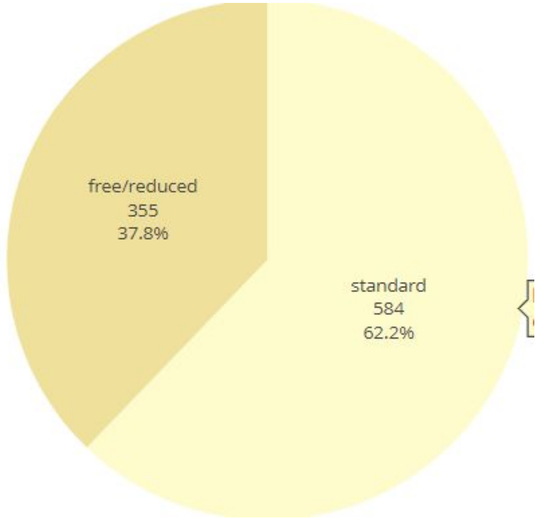
gender

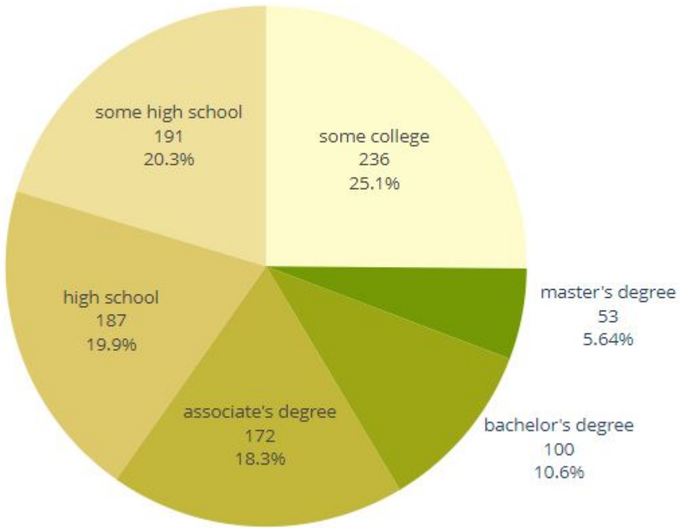


race/ethnicity

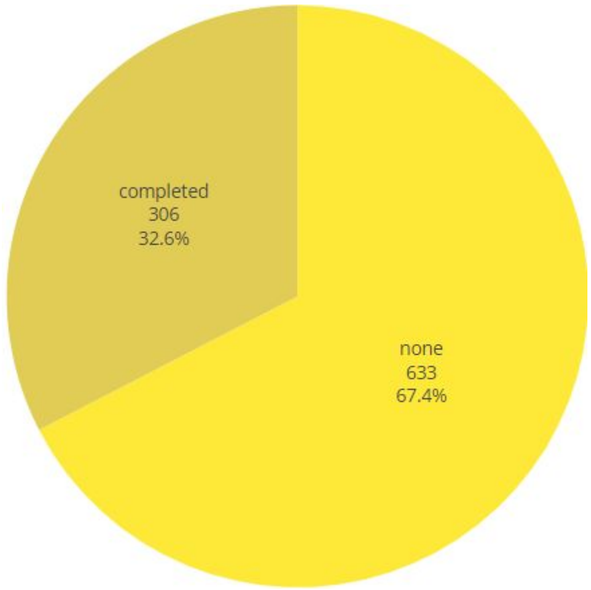


lunch





parental level of education

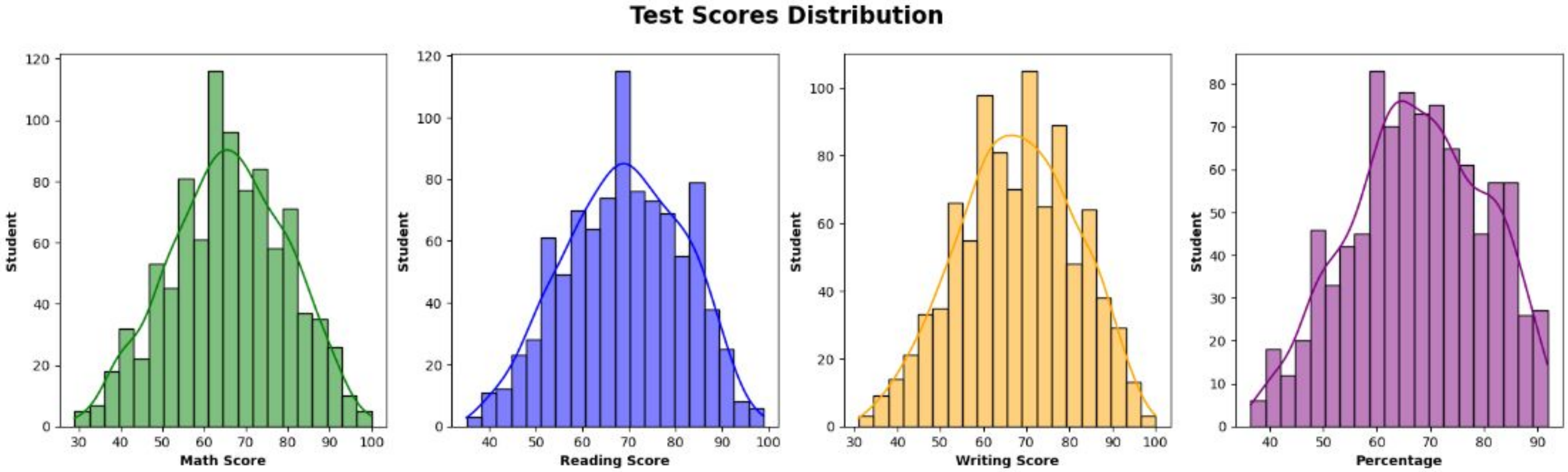


test preparation course

The Distribution of Student's Test Scores For Each Subject

- Distribution of scores
- Distribution of grades
- all Distributions together

The Distribution of scores



The Distribution of grades

marks >= 90 : grade = 'A'

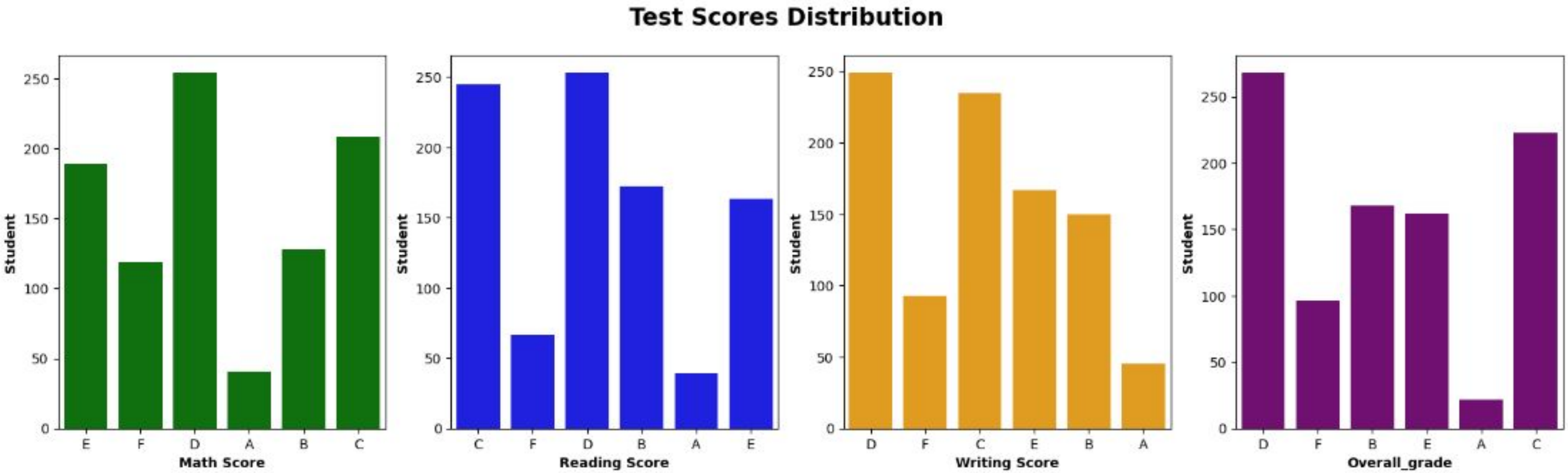
marks >= 80 : grade = 'B'

marks >= 70 : grade = 'C'

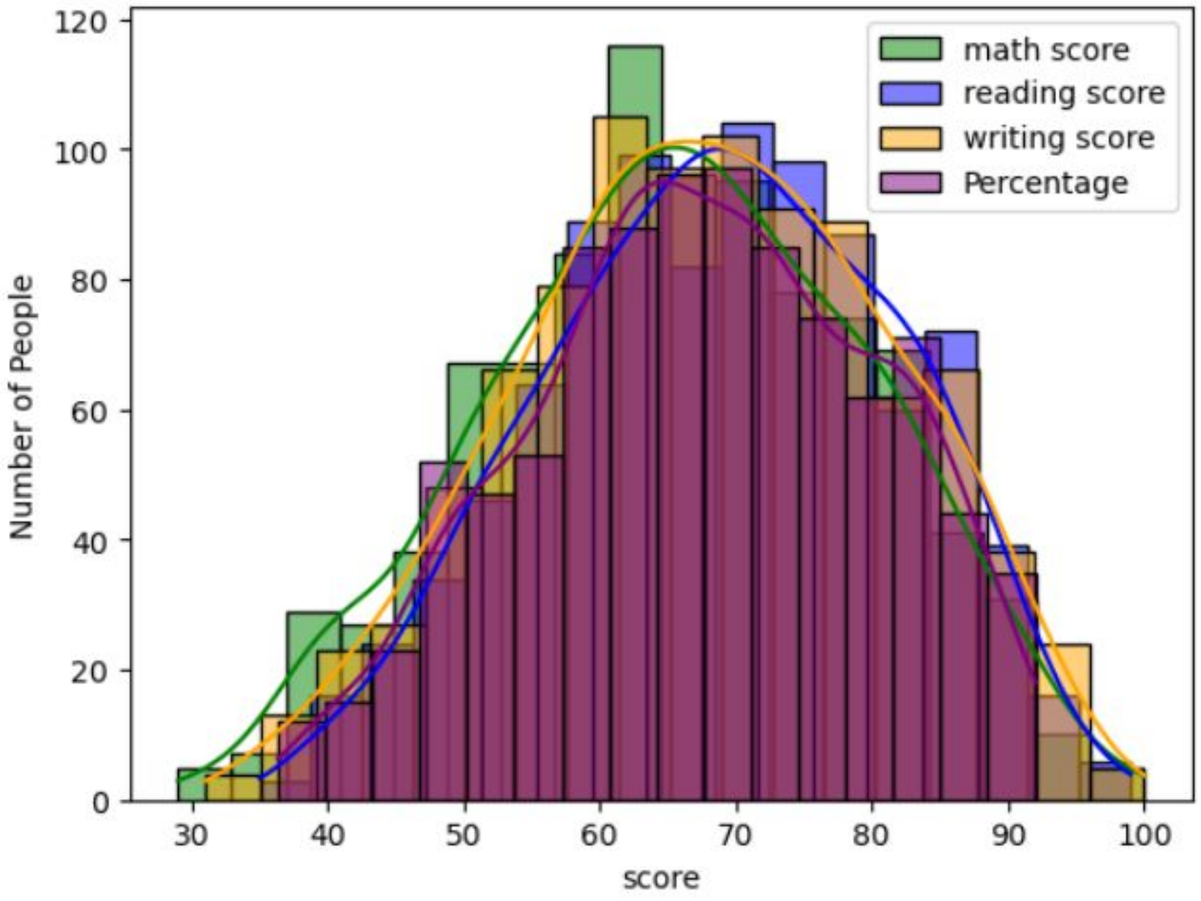
marks >= 60 : grade = 'D'

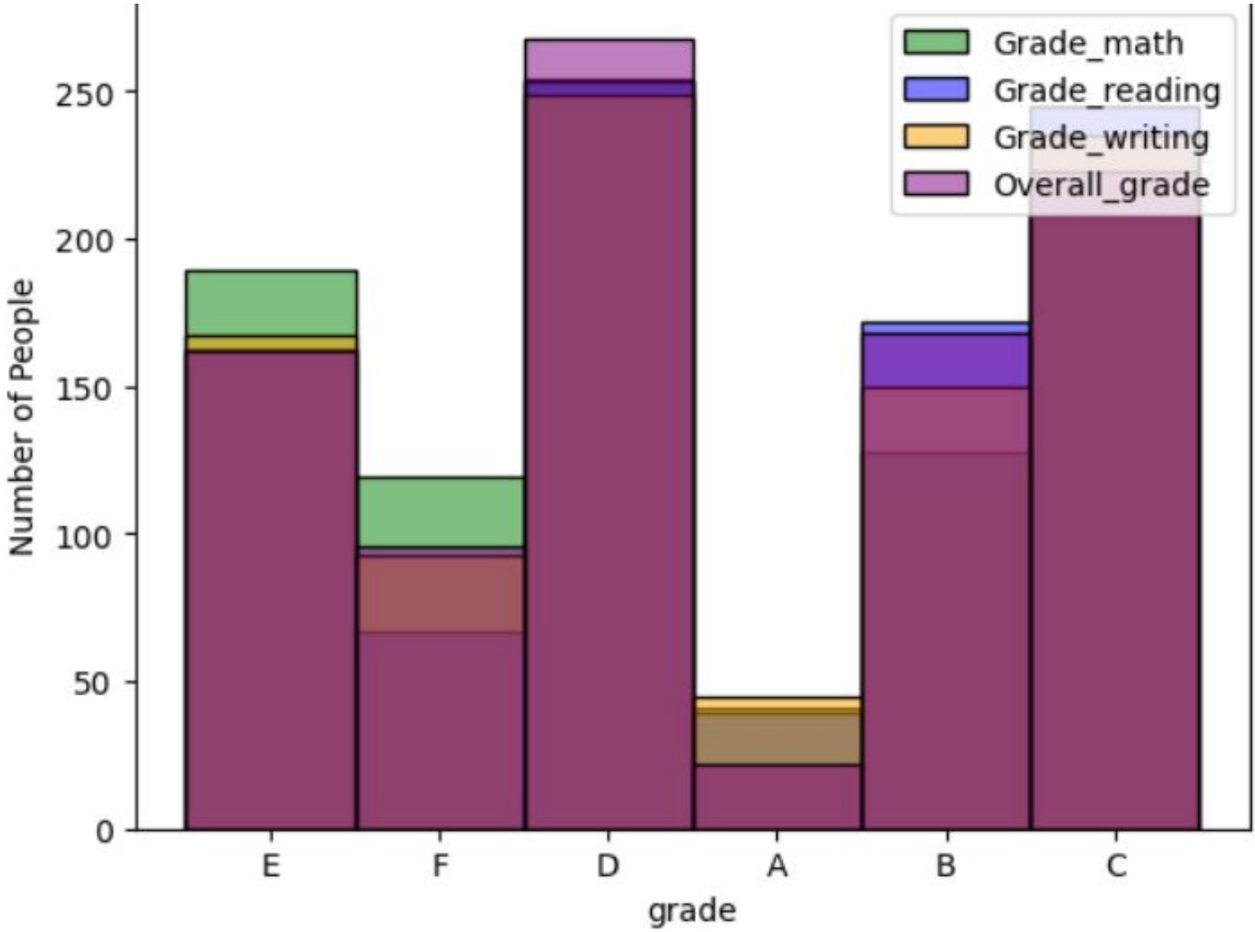
marks >= 50 : grade = 'E'

marks < 50 : grade = 'F'



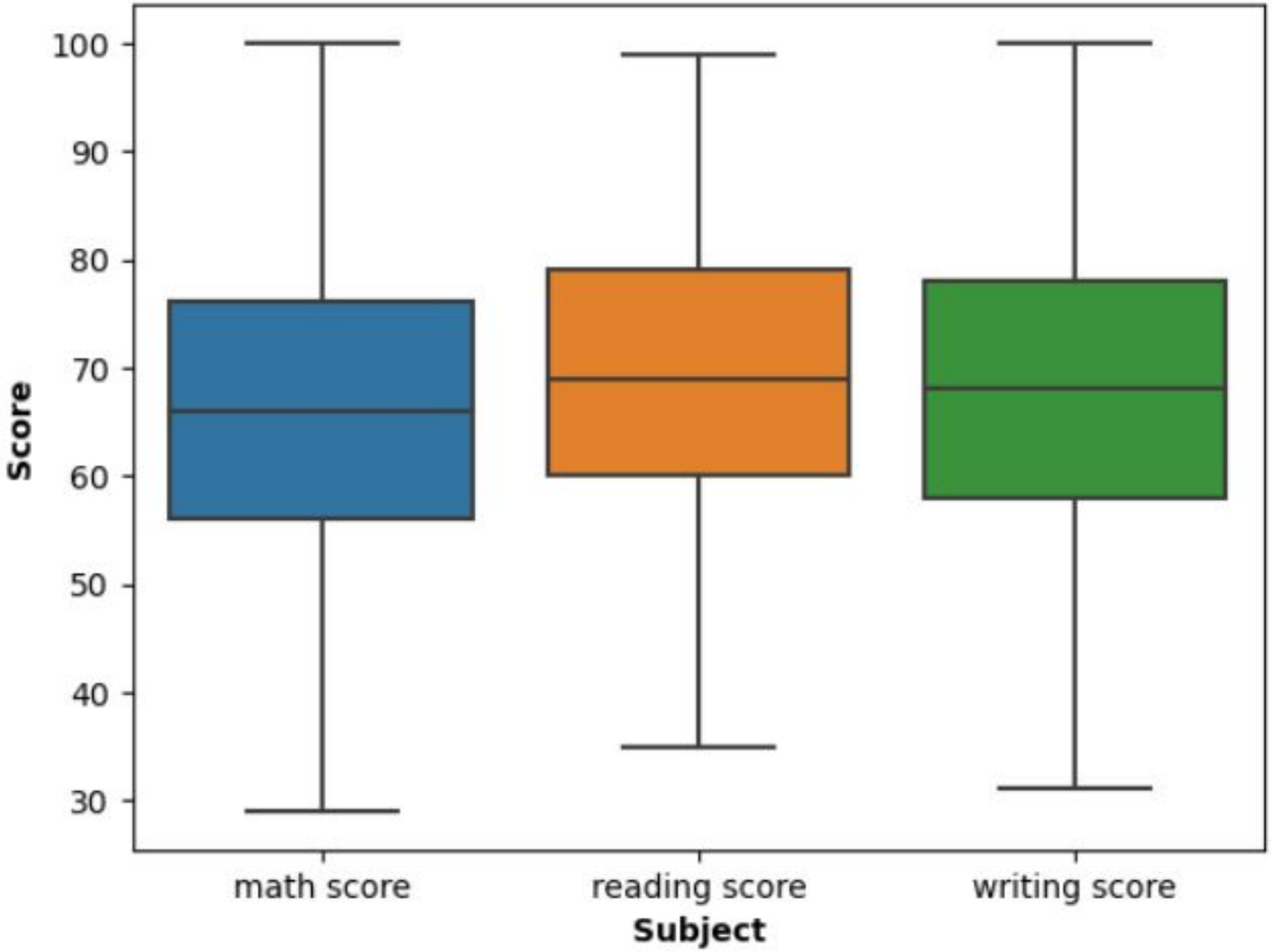
all the Distributions together



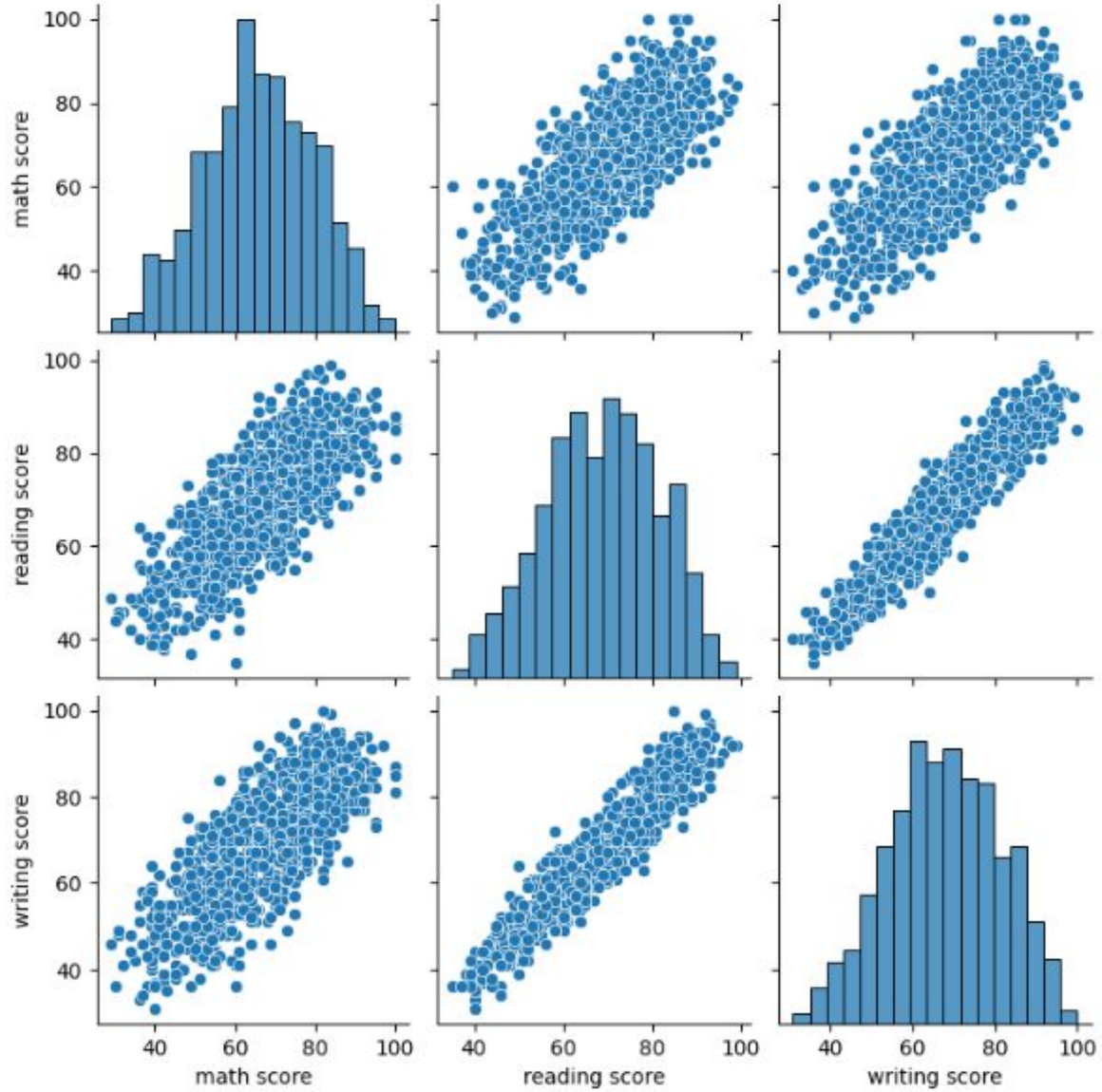


comparison of student test score between subjects

students performed worse in mathematics



correlation



corrcoef(reading score, writing score)= 0.94

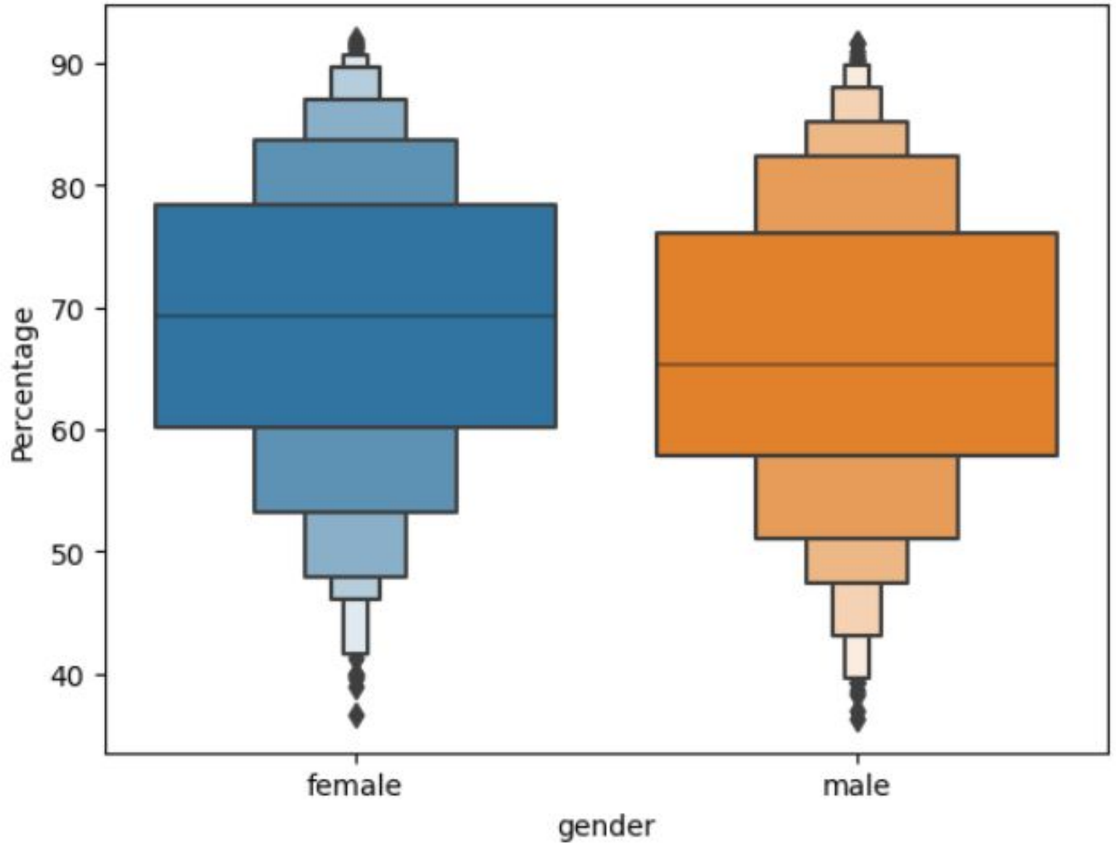
reading score and writing score are linearly related

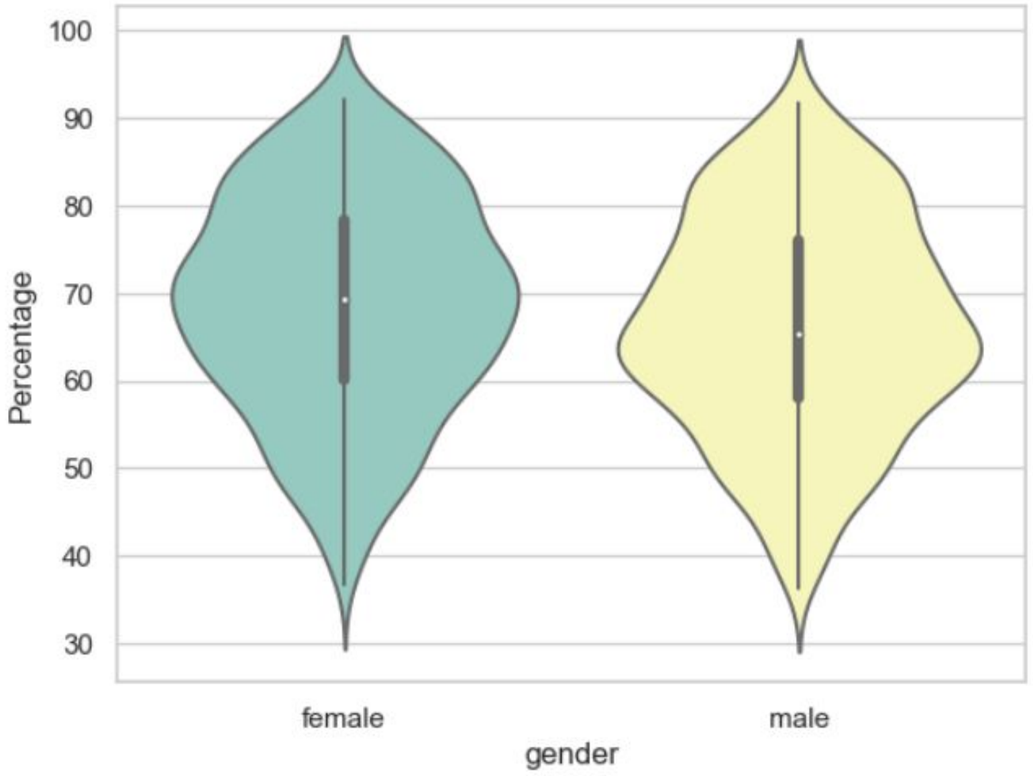
influence of different factors on students performance

- gender
- Race / Ethnicity
- lunch
- test preparation course
- parental level of education

influence of different factors on students performance gender

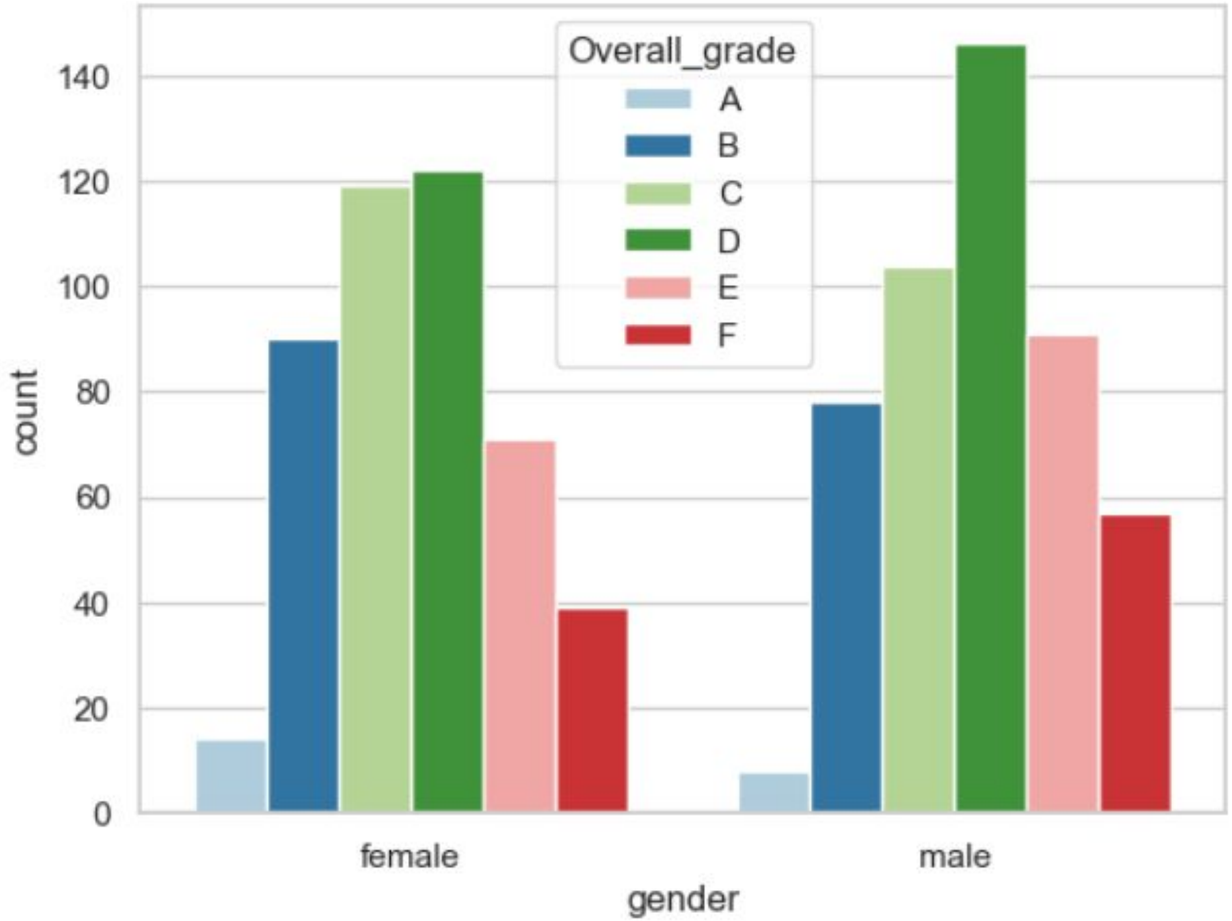
students performance on mean of 3 scores





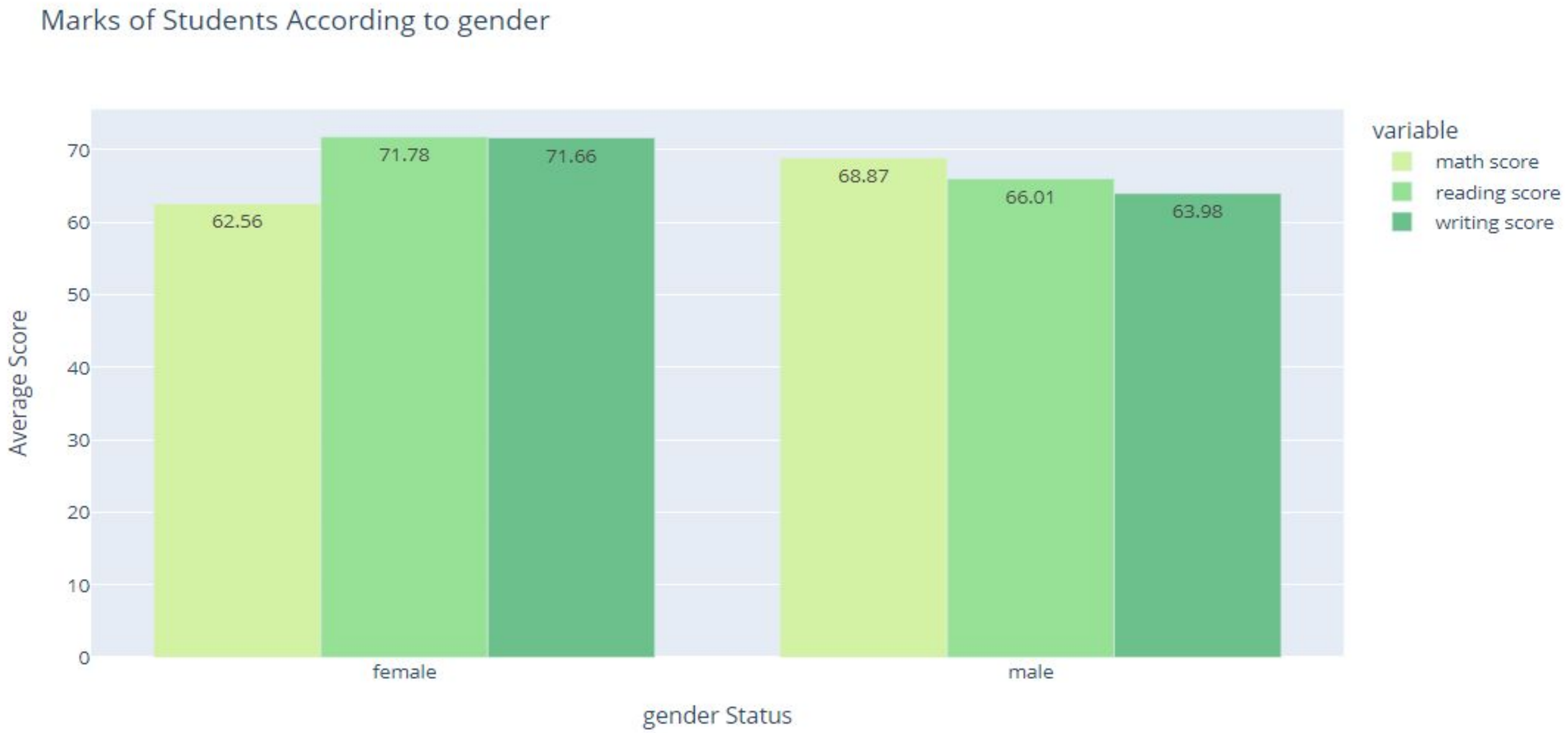
influence of different factors on students performance gender

students performance on mean of 3 scores



on average performance of female are better

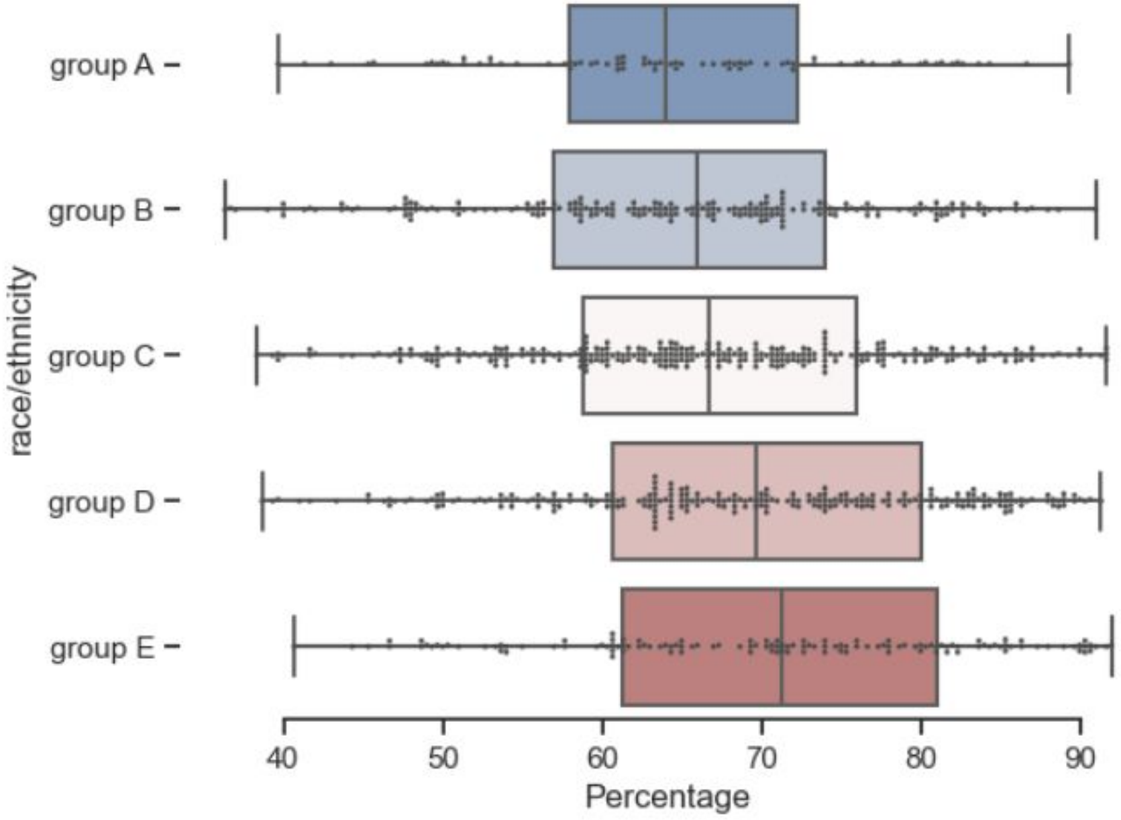
influence of different factors on students performance gender

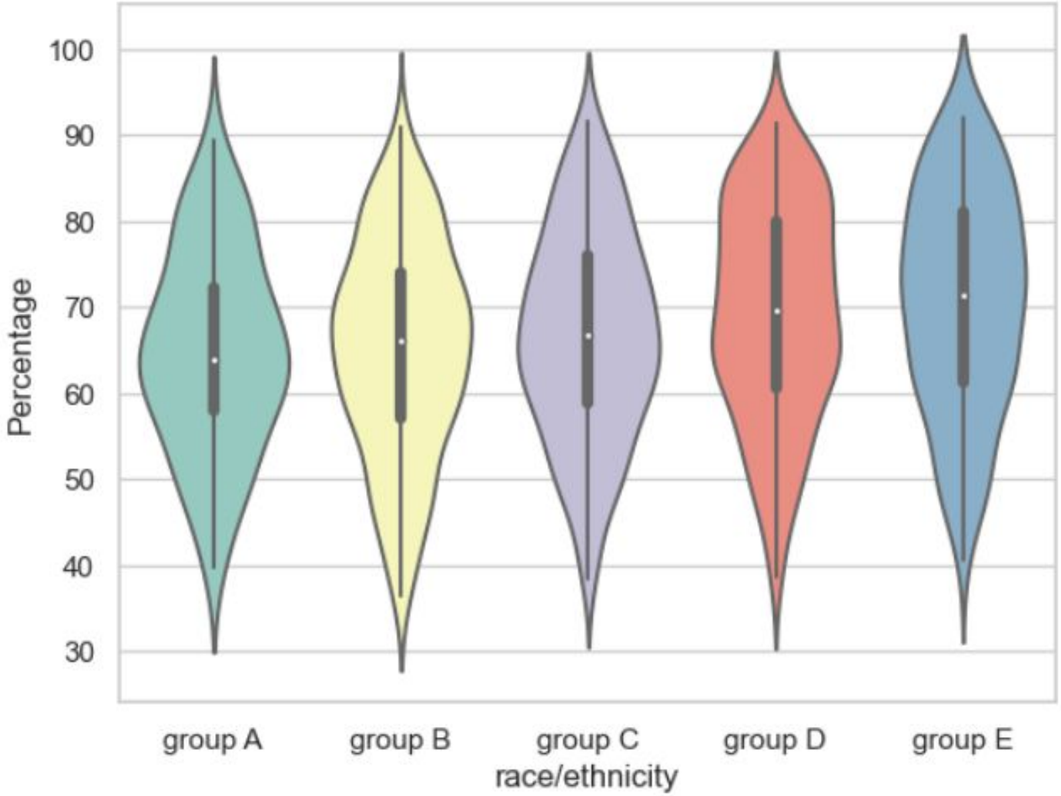


in math, male performed better
in writing, female performed better
in reading, female performed better

influence of different factors on students performance Race / Ethnicity

students performance on mean of 3 scores





on average performance of group E are better

influence of different factors on students performance Race / Ethnicity

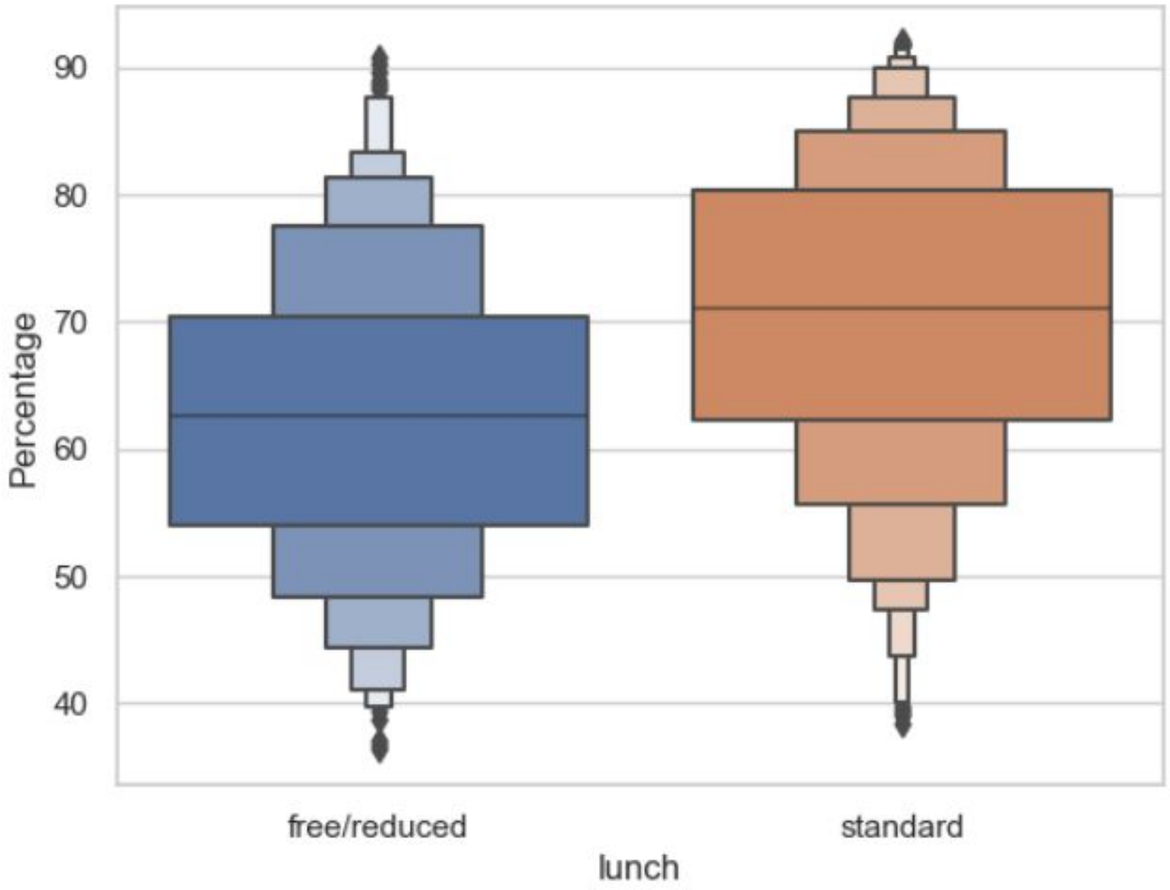
in math, group E performed better

in writing, group E performed better

in reading, group D performed better

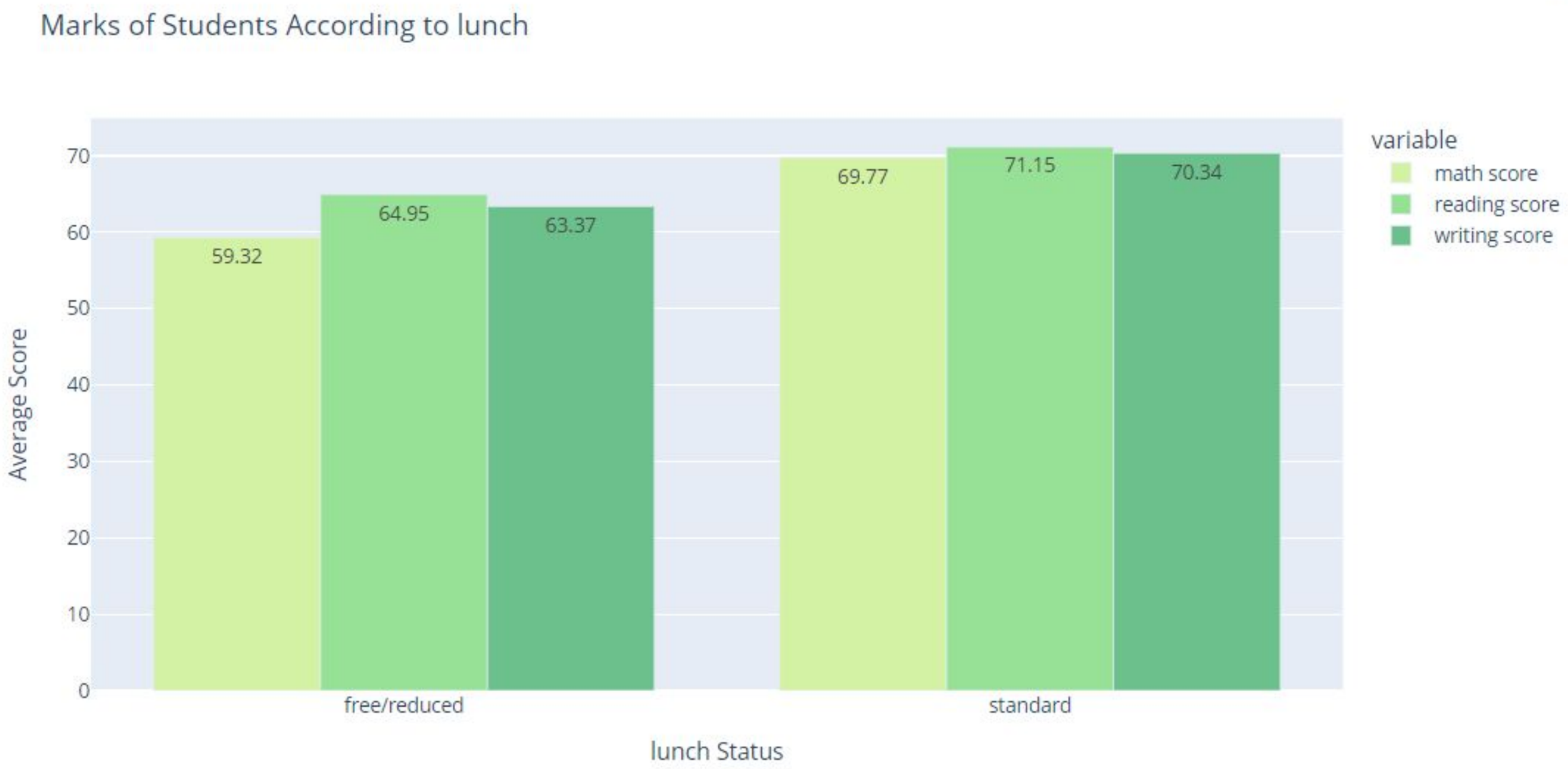
influence of different factors on students performance lunch

students performance on mean of 3 scores



on average performance of students who have standard lunch are better

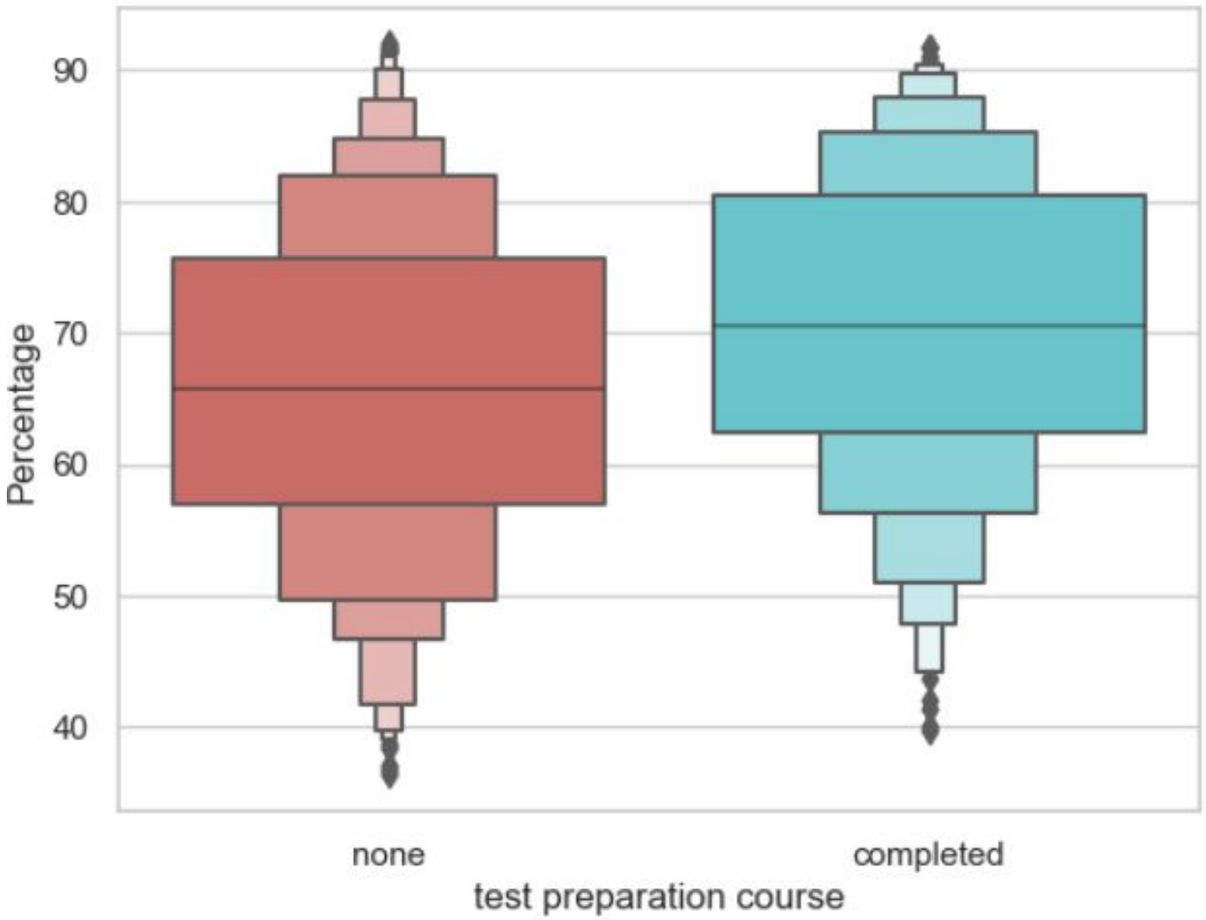
influence of different factors on students performance lunch



in all 3 exams students that have standard lunch performed better

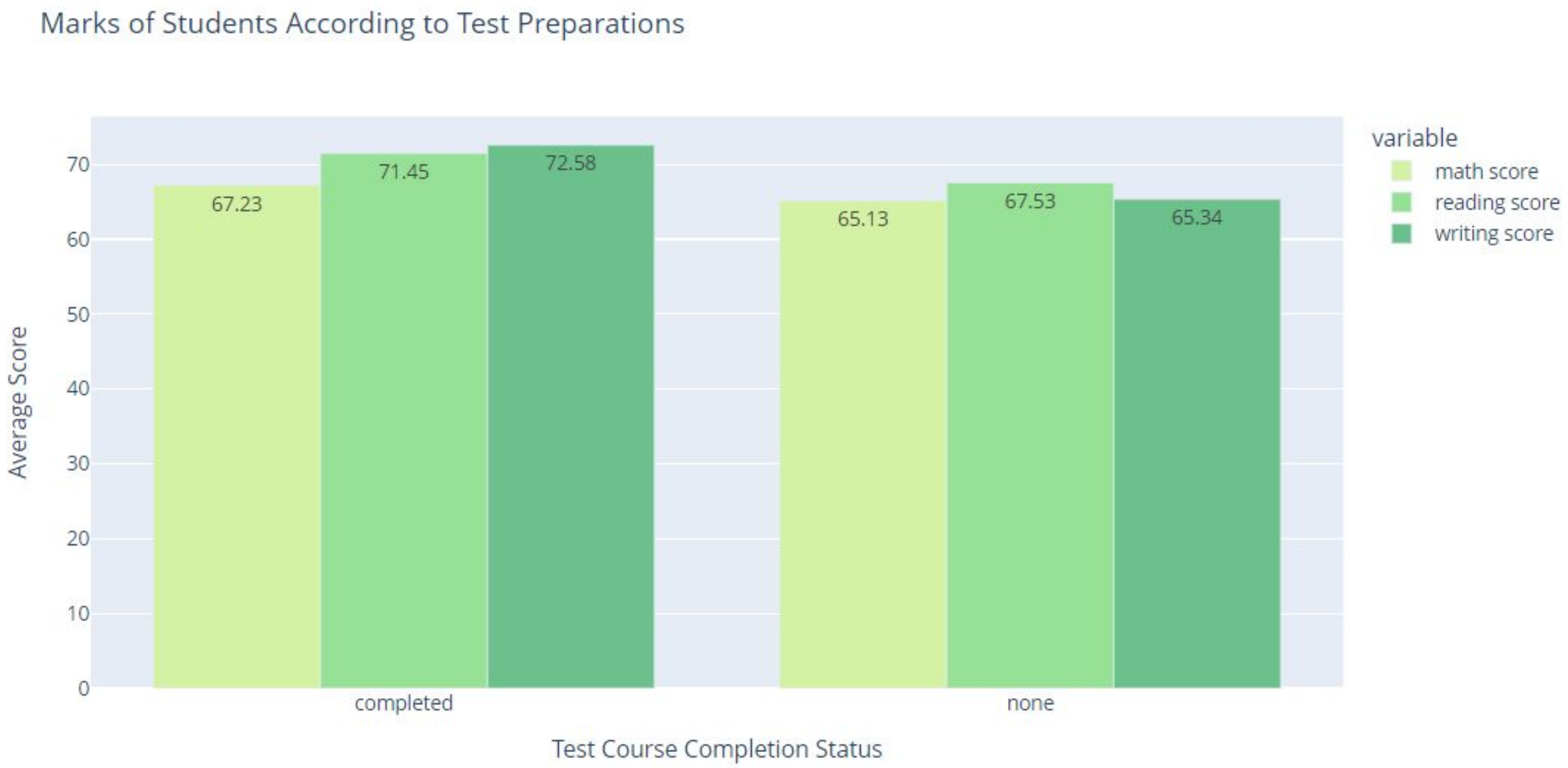
influence of different factors on students performance test preparation

students performance on mean of 3 scores



on average, performance of students who were completely prepared are better

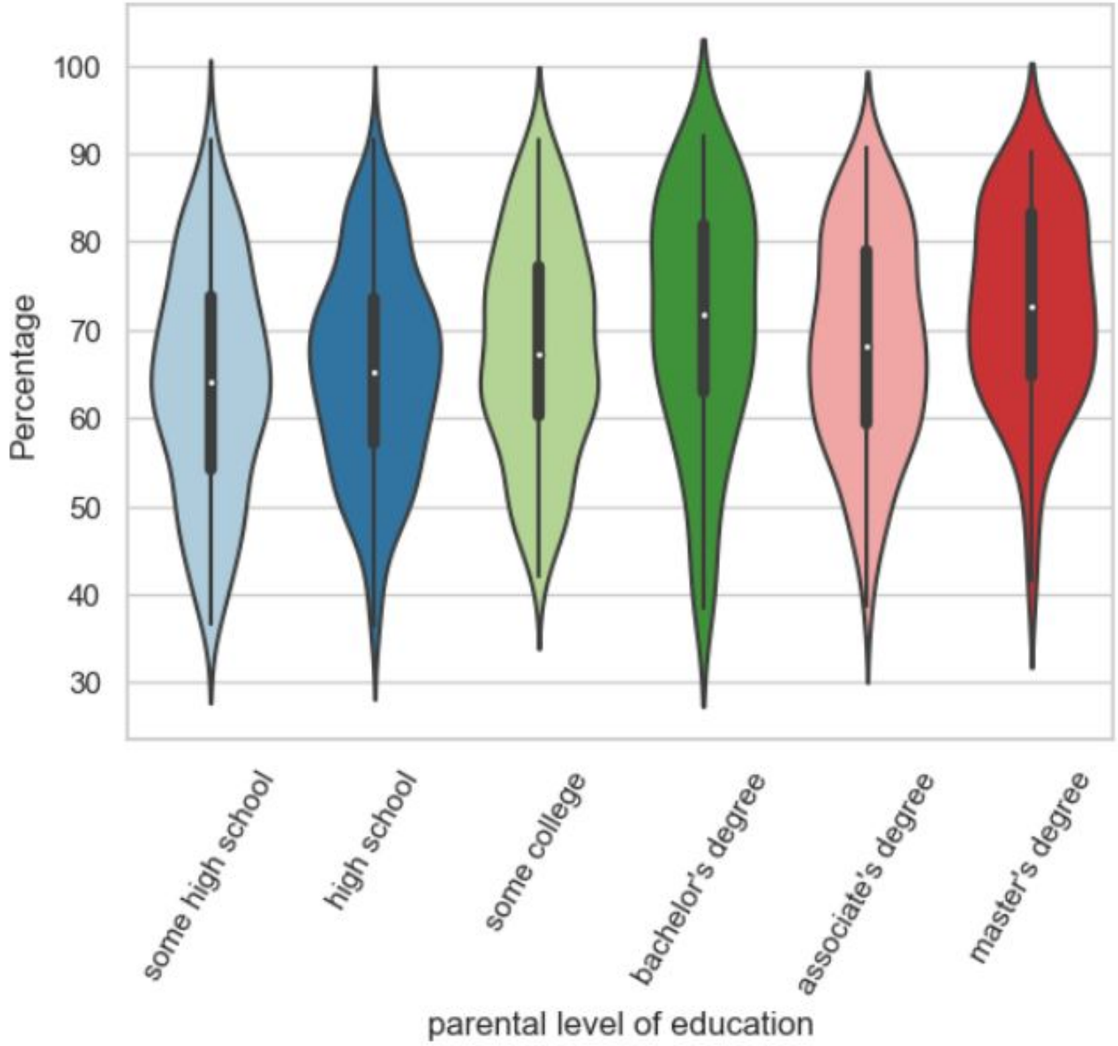
influence of different factors on students performance test preparation



in all 3 exams students that were completely prepared performed better

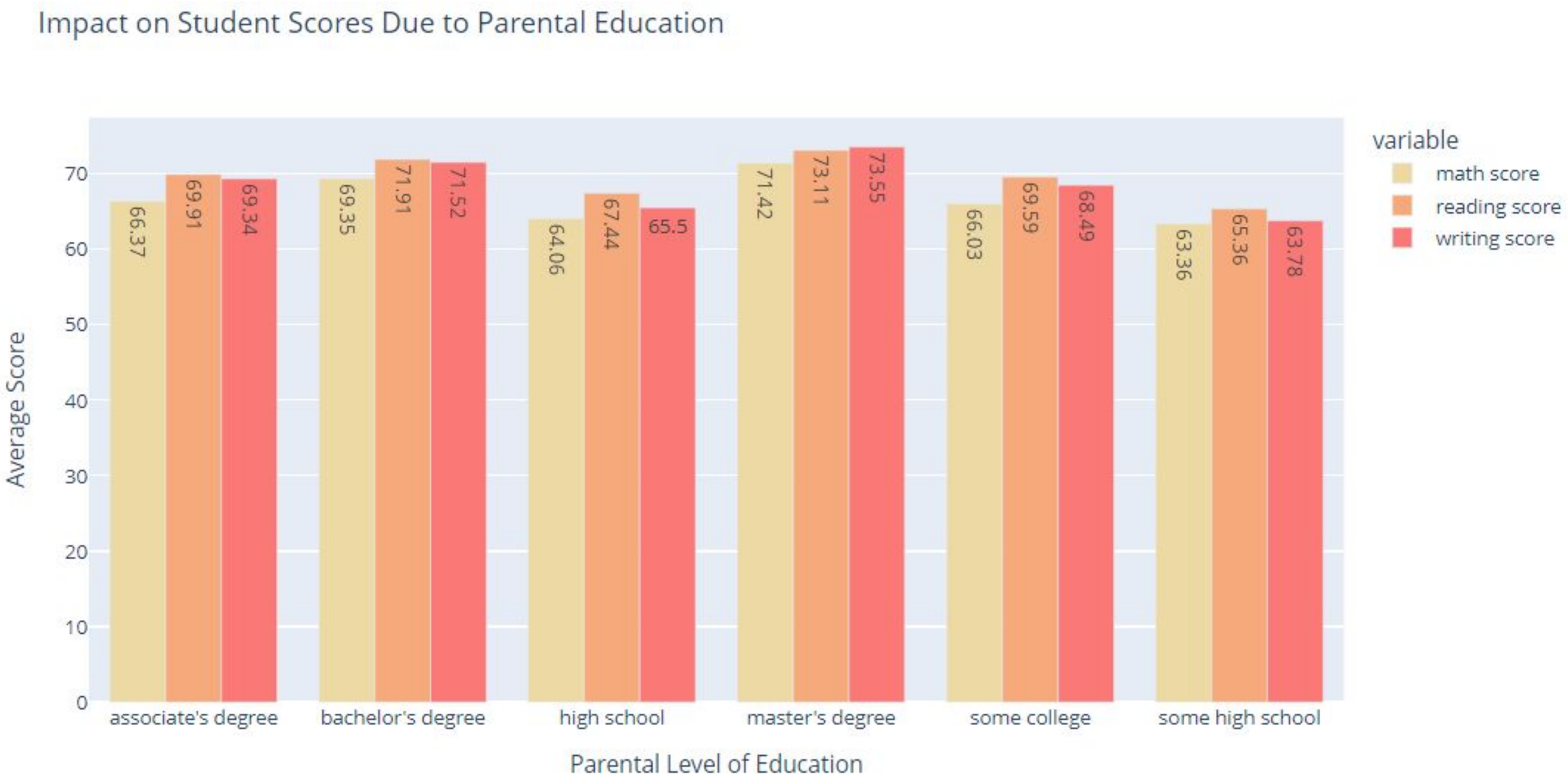
influence of different factors on students performance parental level of education

students performance on mean of 3 scores



on average, students whose parents have a master's degree or a bachelor's degree performed the best

influence of different factors on students performance parental level of education

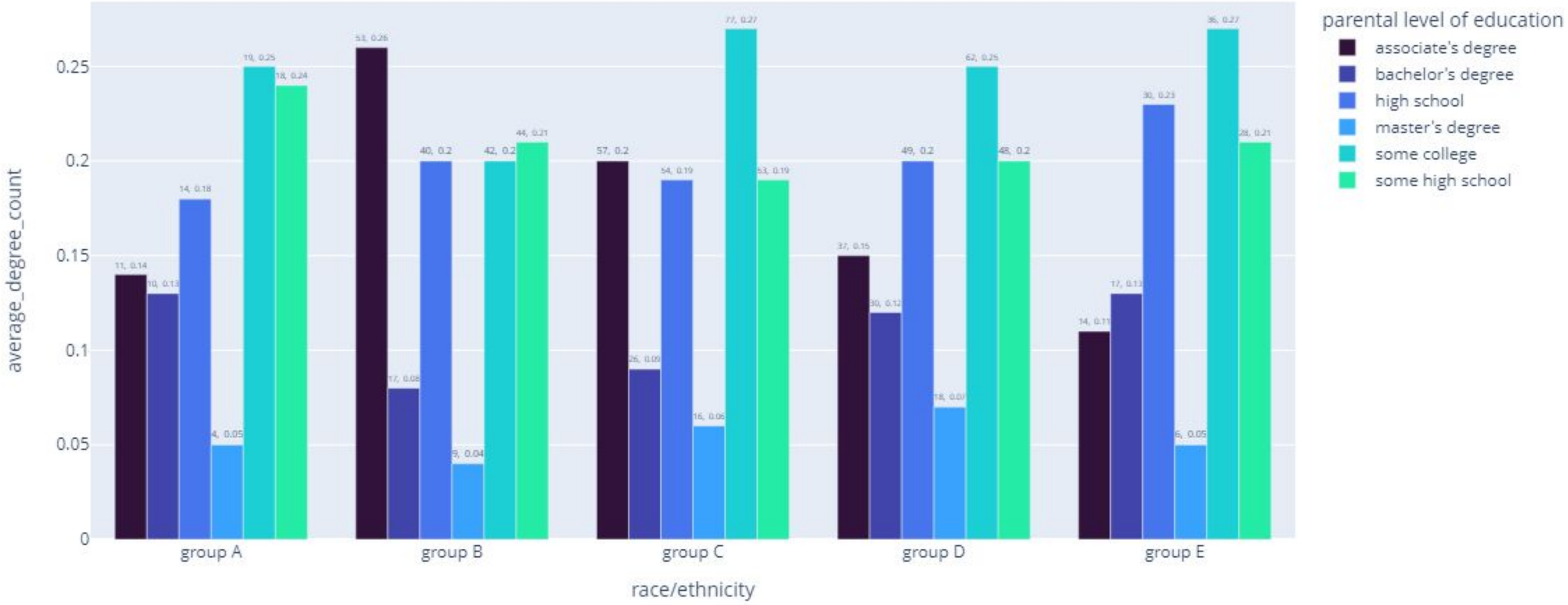


in all 3 exams students whose parents have a master's degree performed best

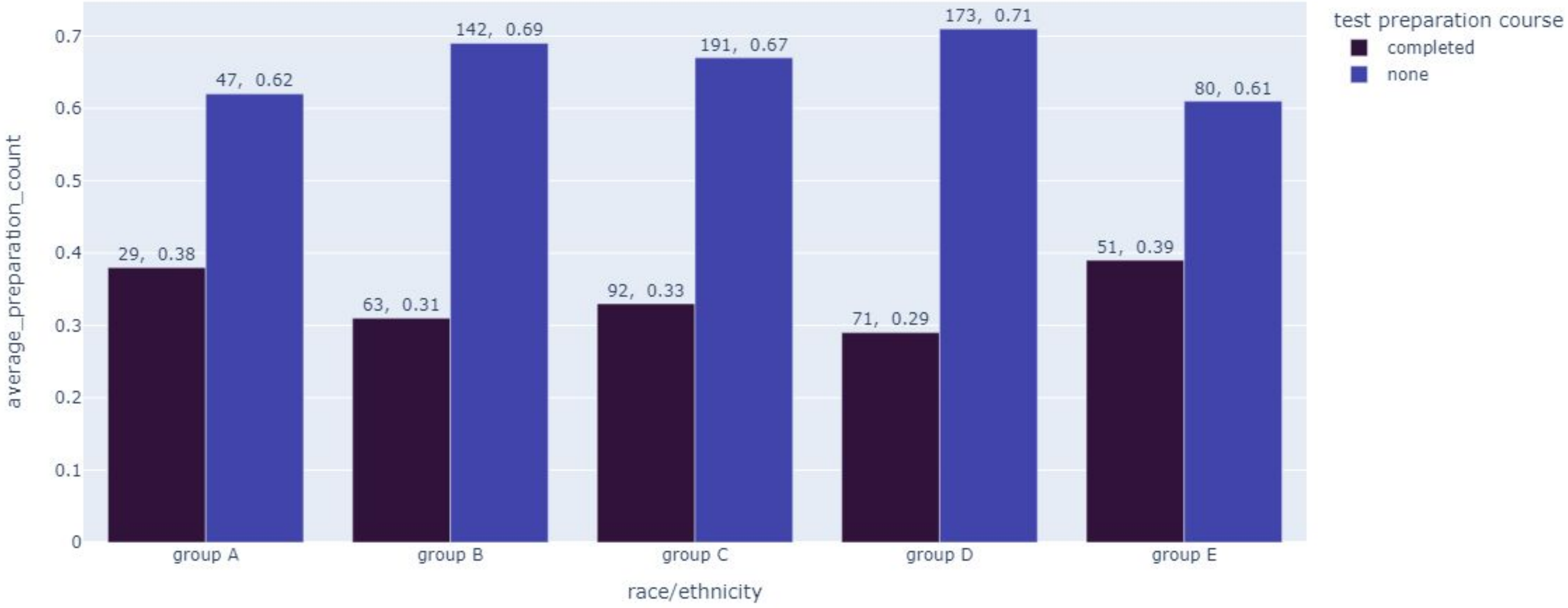
average count on "race/ethnicity"

- parental level of education
- test preparation course
- lunch
- gender

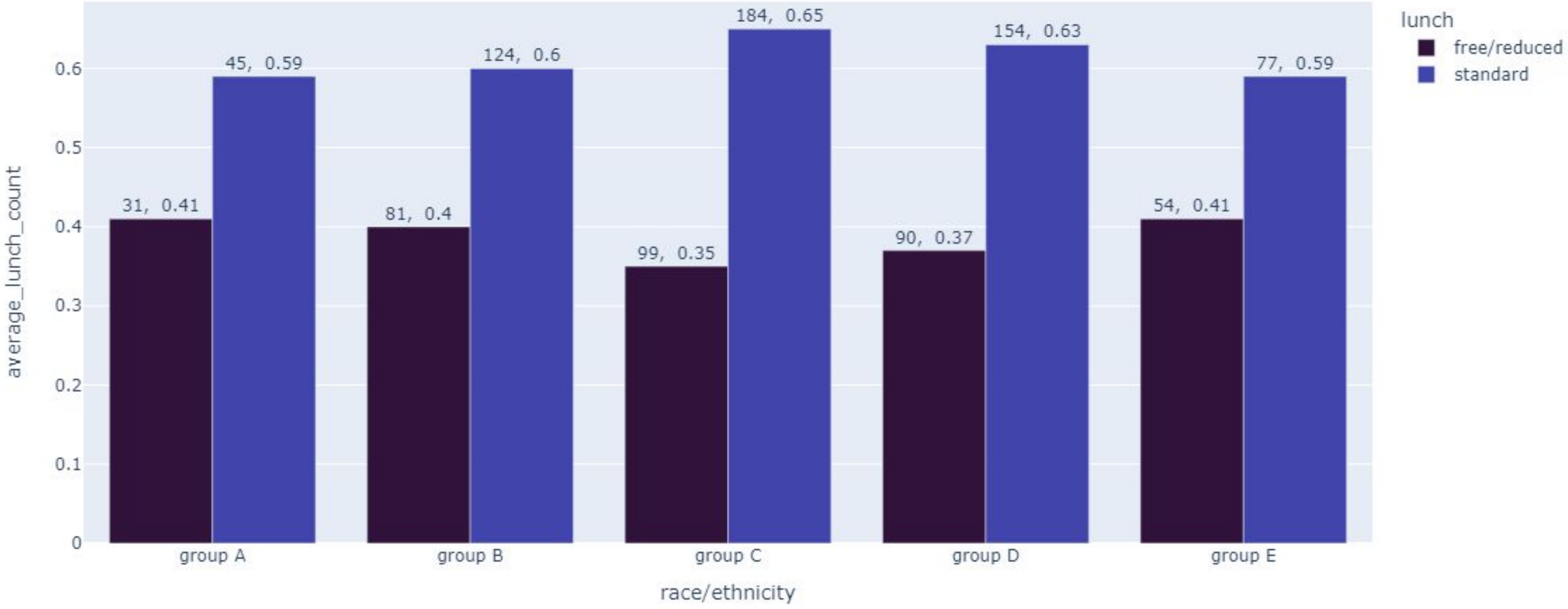
parental level of education



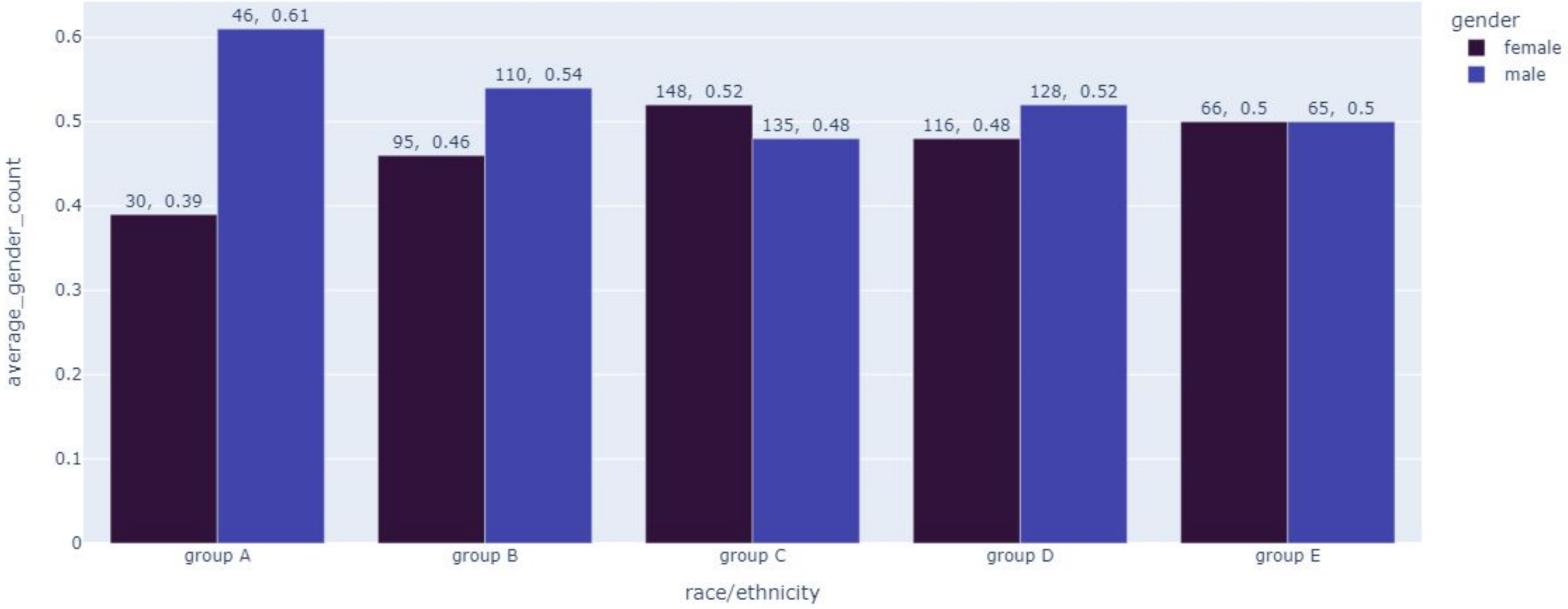
test preparation course



lunch


lunch
free/reduced
standard
average_lunch_count
0
0.1
0.2
0.3
0.4
0.5
0.6
31, 0.41
45, 0.59
81, 0.4
124, 0.6
99, 0.35
184, 0.65
90, 0.37
154, 0.63
54, 0.41
77, 0.59
group A
group B
group C
group D
group E
race/ethnicity

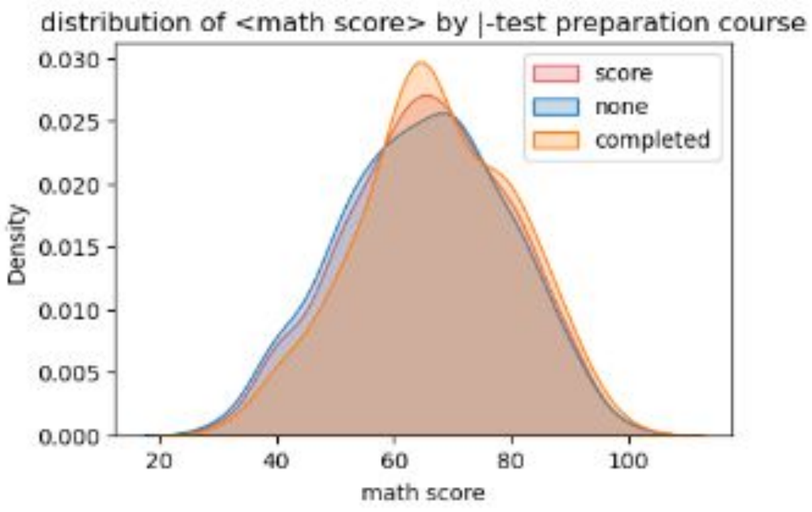
gender

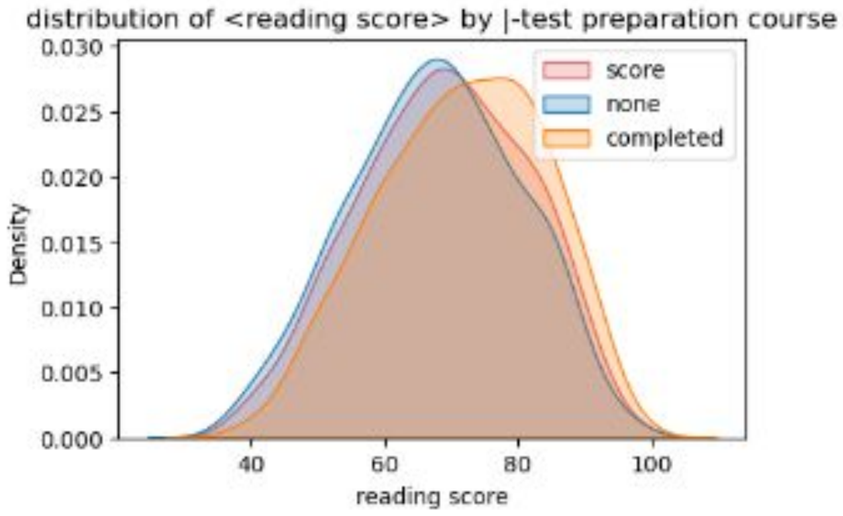

46, 0.61
110, 0.54
148, 0.52
128, 0.52
66, 0.5
65, 0.5
135, 0.48
116, 0.48
95, 0.46
30, 0.39
0.6
0.5
0.4
0.3
0.2
0.1
0
average_gender_count
group A
group B
group C
group D
group E
race/ethnicity
gender
female
male

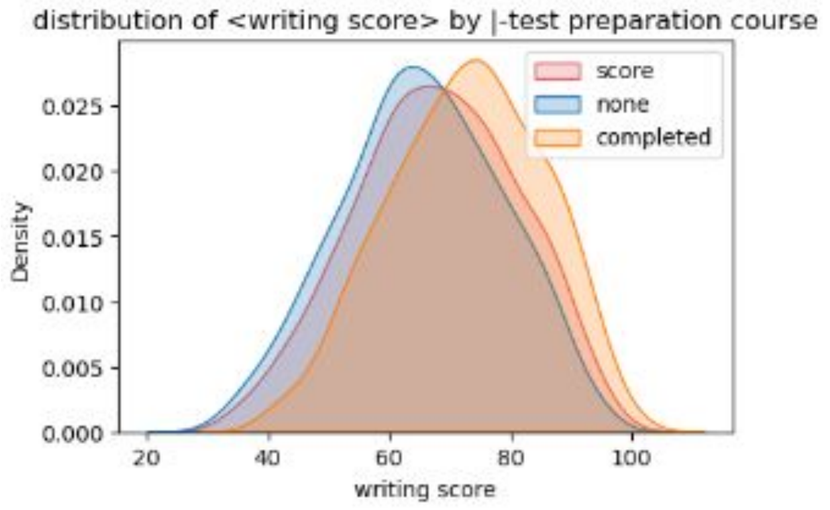
Distribution of scores by column for each test score

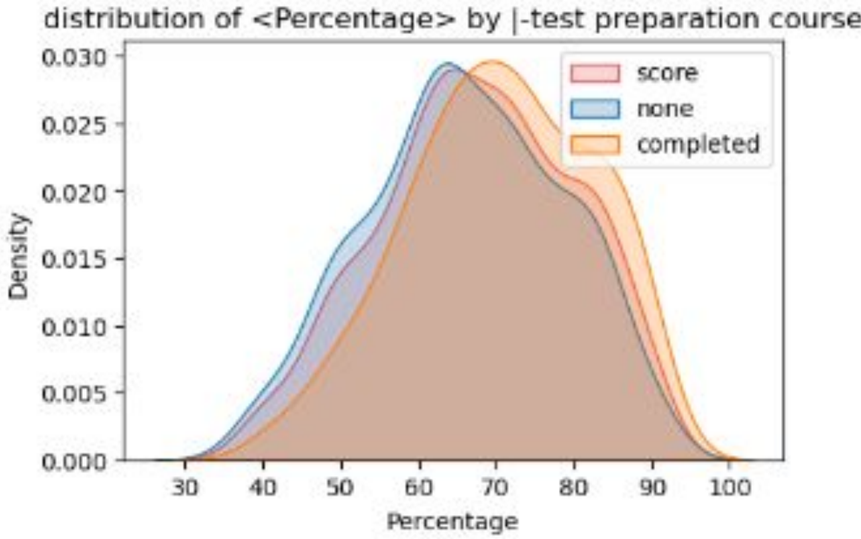
- test preparation course
- lunch
- gender
- parental level of education
- race/ethnicity

Distribution of scores by column for each test score test preparation course





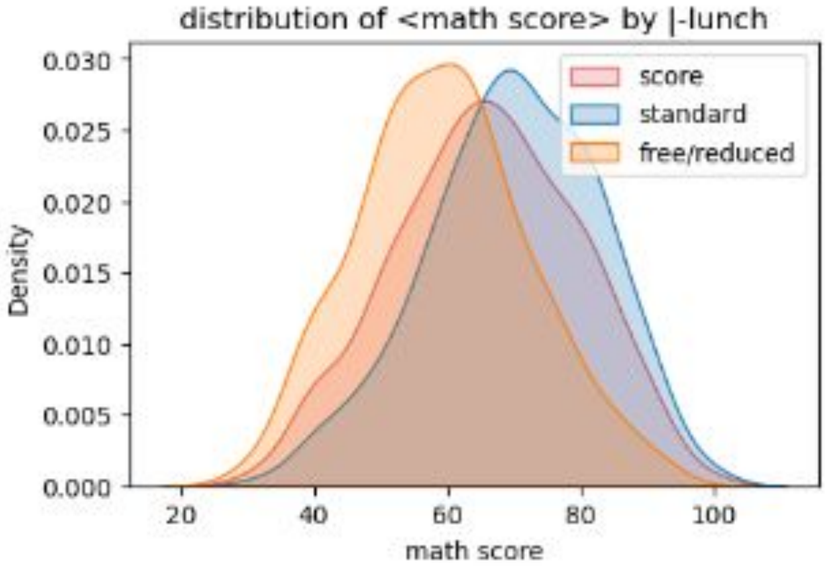


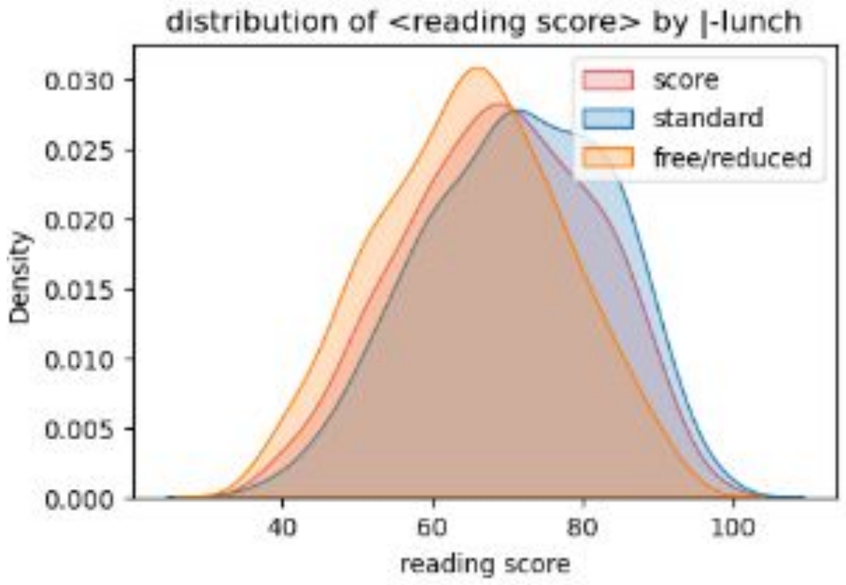


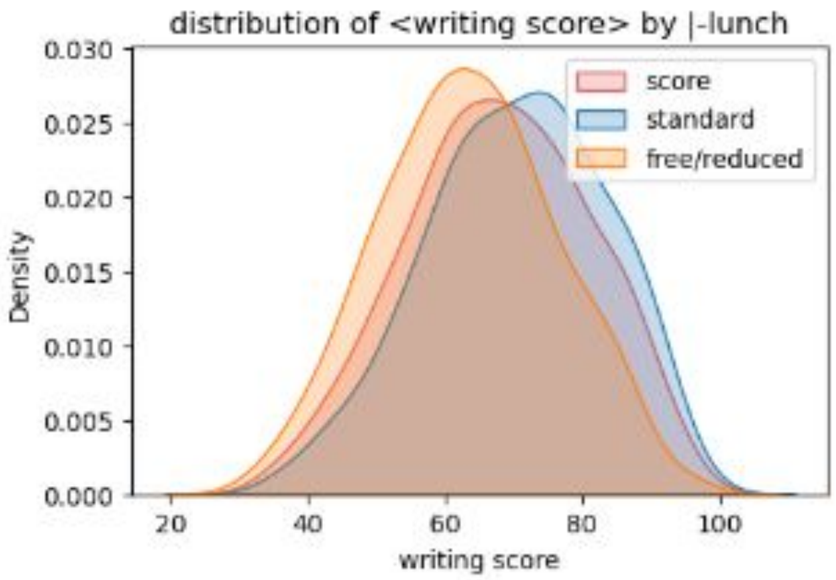
Exam preparation has a greater effect on the result of reading and writing scores and does not have a significant effect on the math score

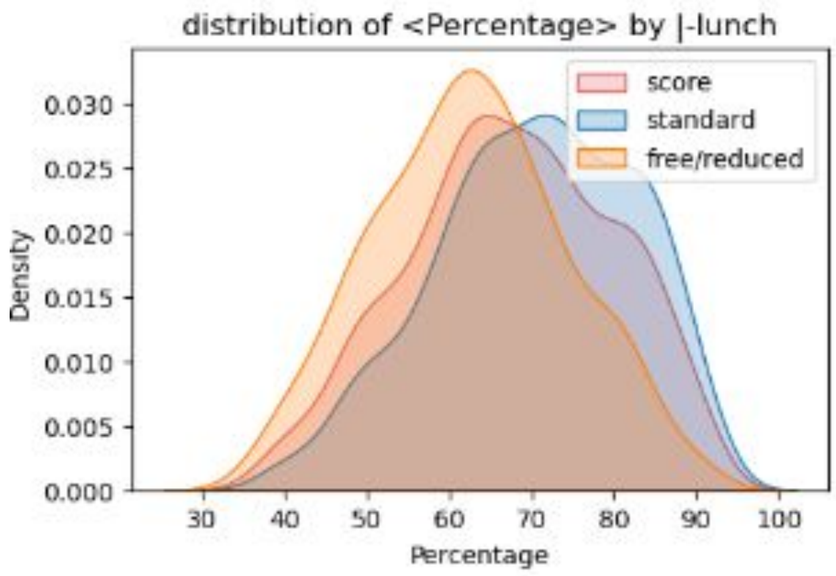
but on average has effect on performance

Distribution of scores by column for each test score lunch







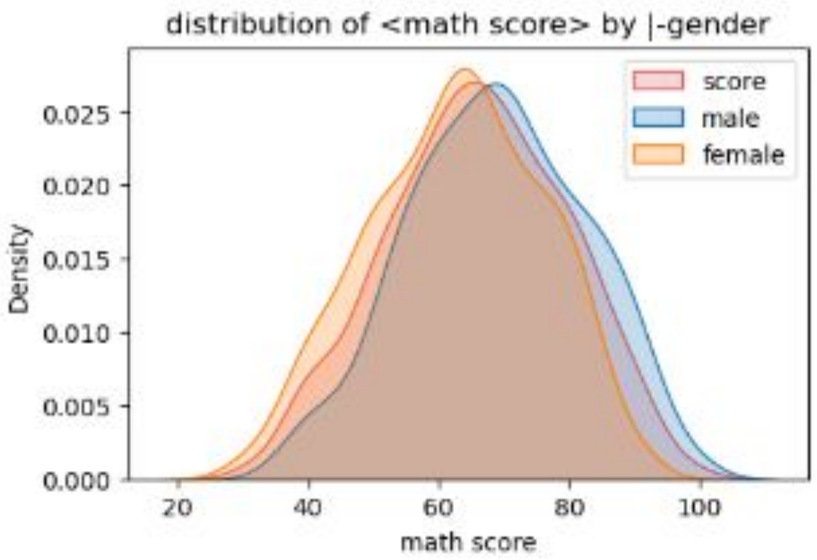


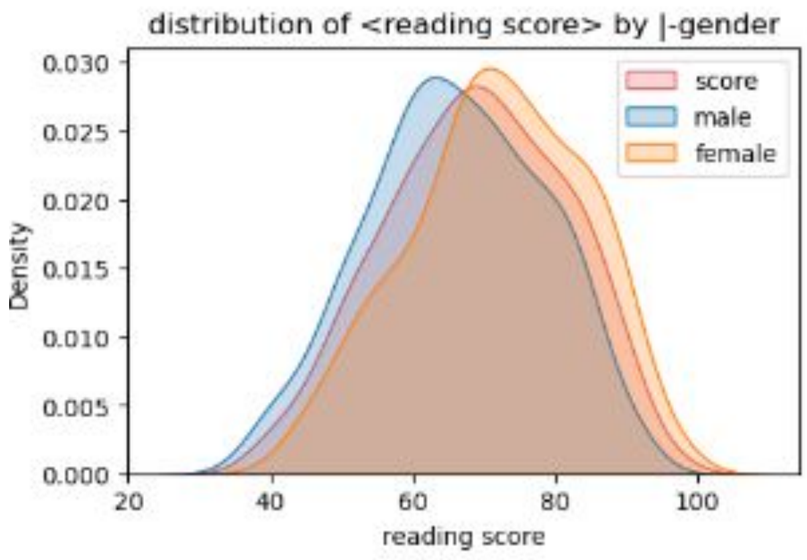
standard lunch has positive effect on the result of each 3 scores but have a significant effect on math score

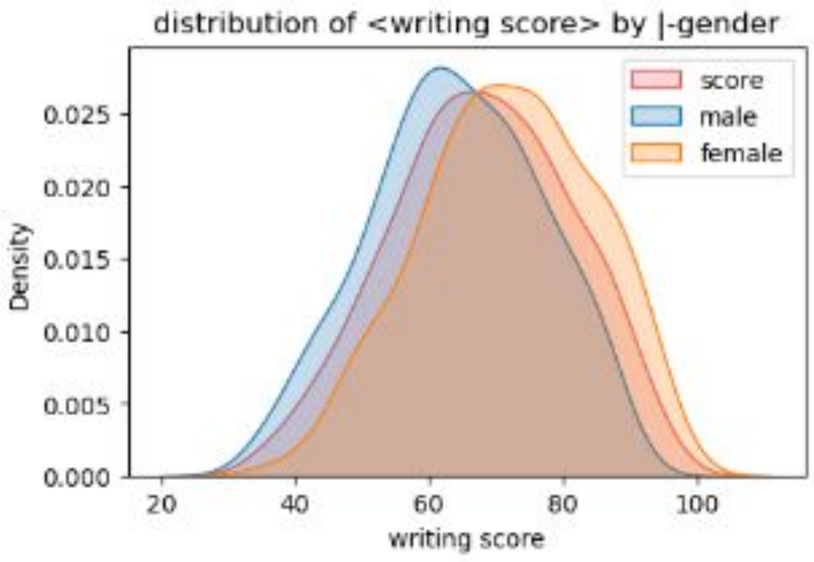
on average standard lunch has effect on performance

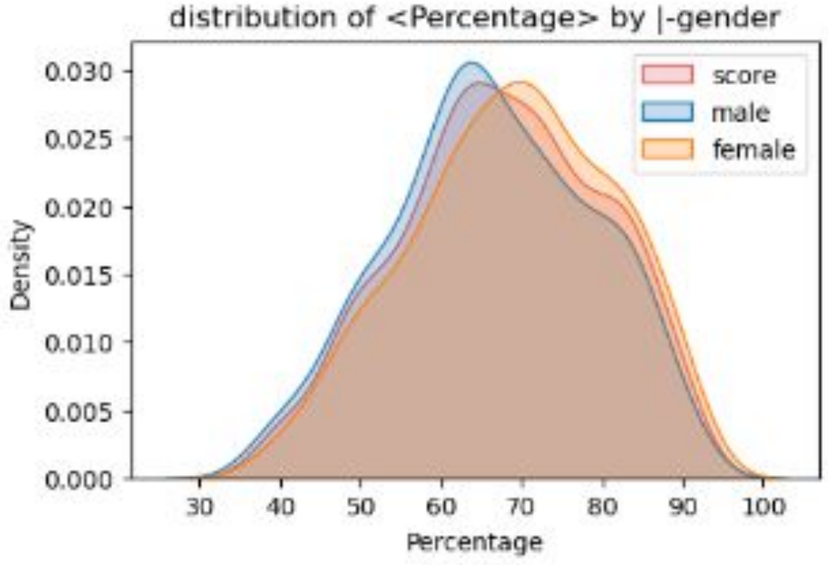
Distribution of scores by column for each test score

gender







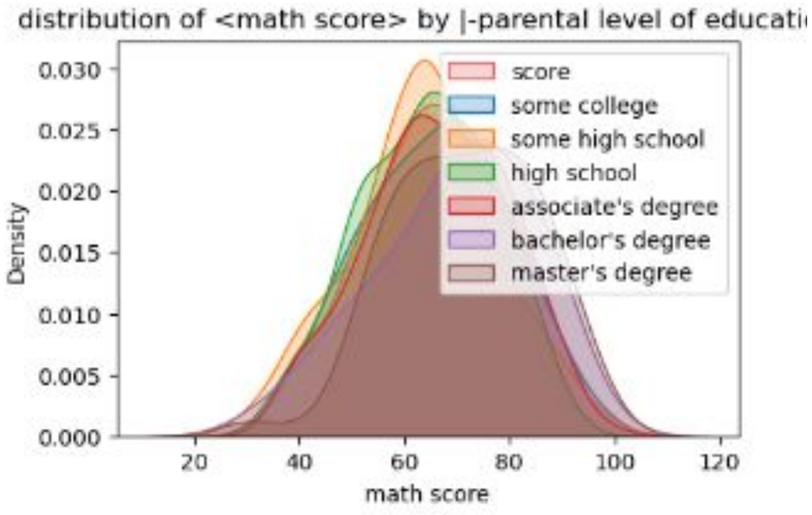


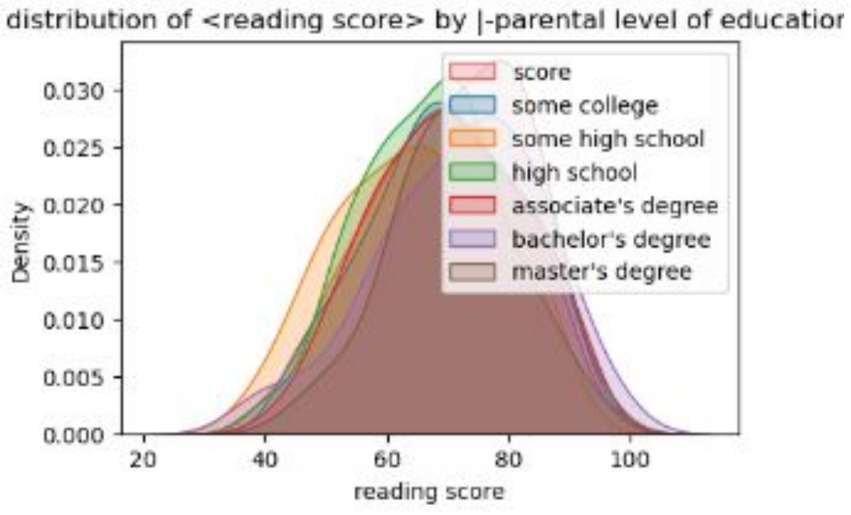
female performed better on reading and writing score but male performed better on math score

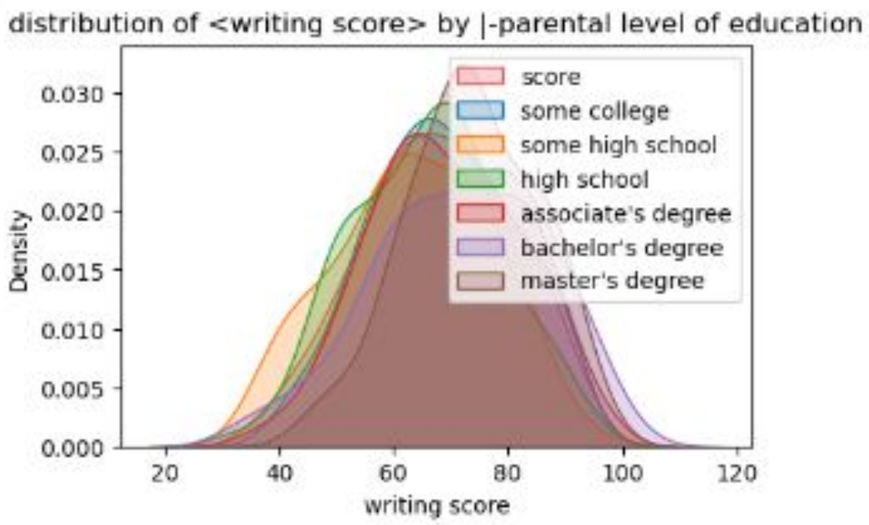
on average female performed better

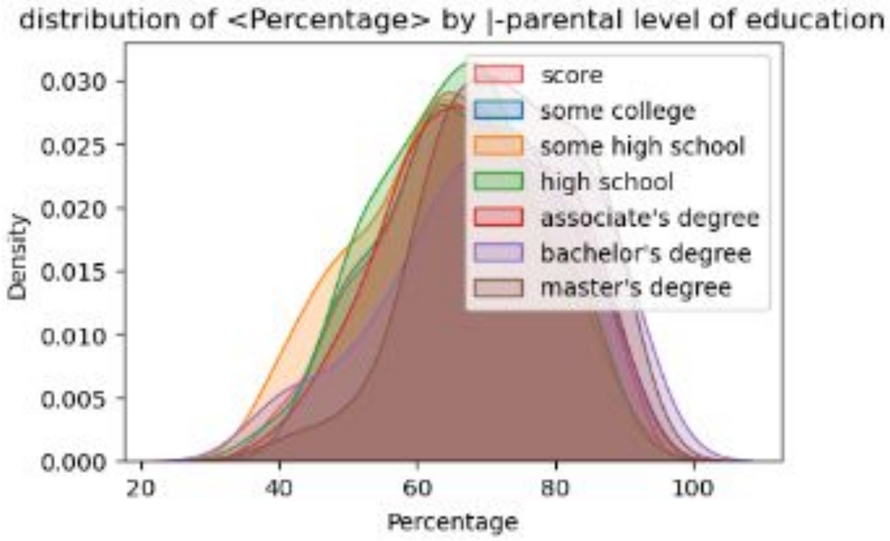
Distribution of scores by column for each test score

parental level of education



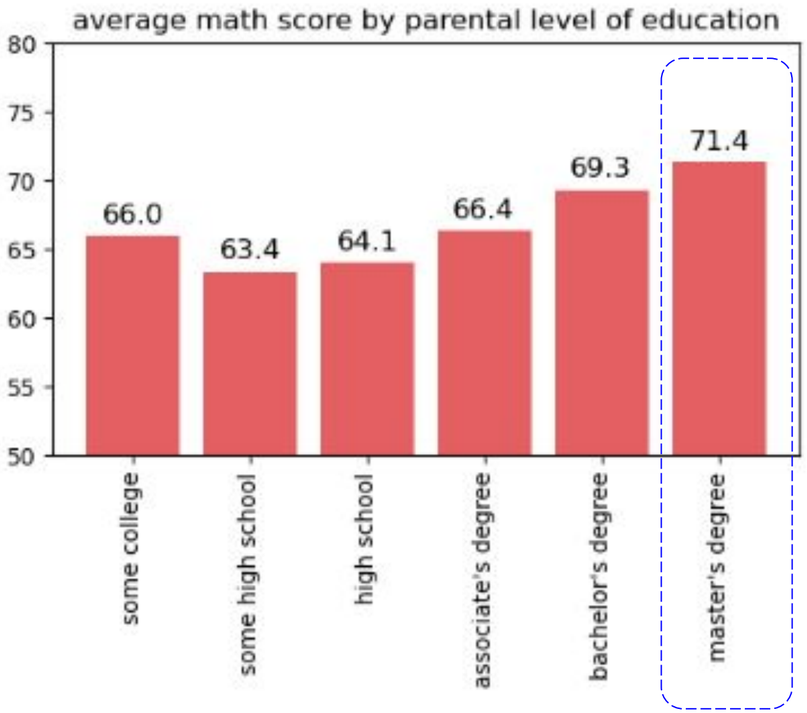






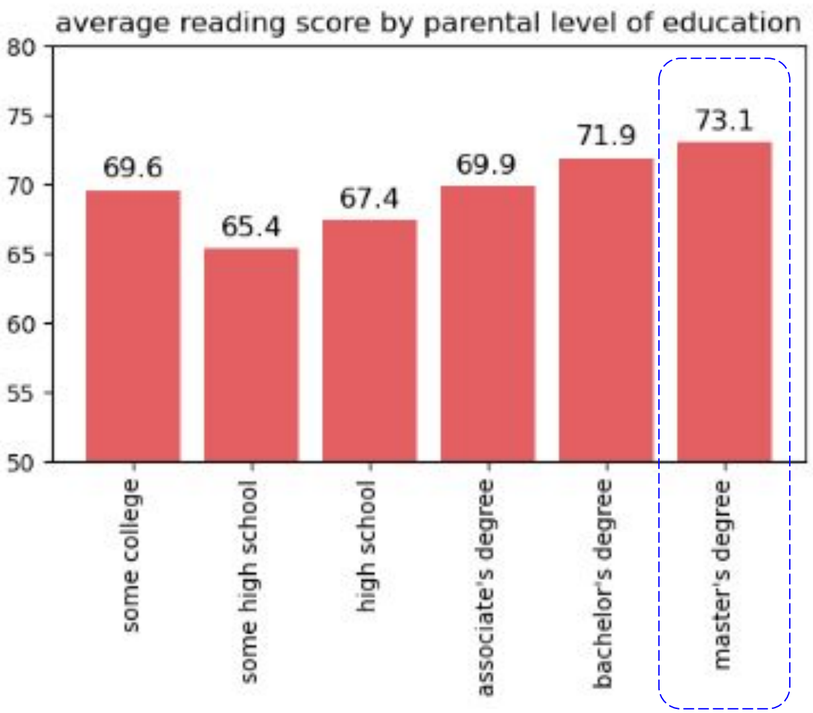
math score

```
< parental level of education >
some college :  66.03389830508475
some high school :  63.361256544502616
high school :  64.05882352941177
associate's degree :  66.36627906976744
bachelor's degree :  69.35
master's degree :  71.41509433962264
```



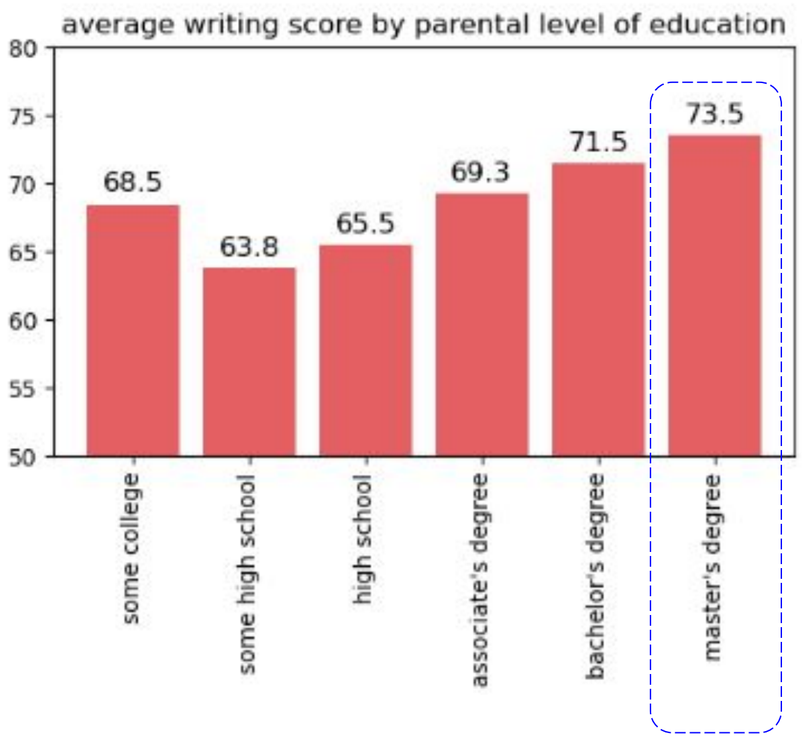
reading score

```
< parental level of education >
some college :  69.58898305084746
some high school :  65.35602094240838
high school :  67.44385026737967
associate's degree :  69.90697674418605
bachelor's degree :  71.91
master's degree :  73.11320754716981
```



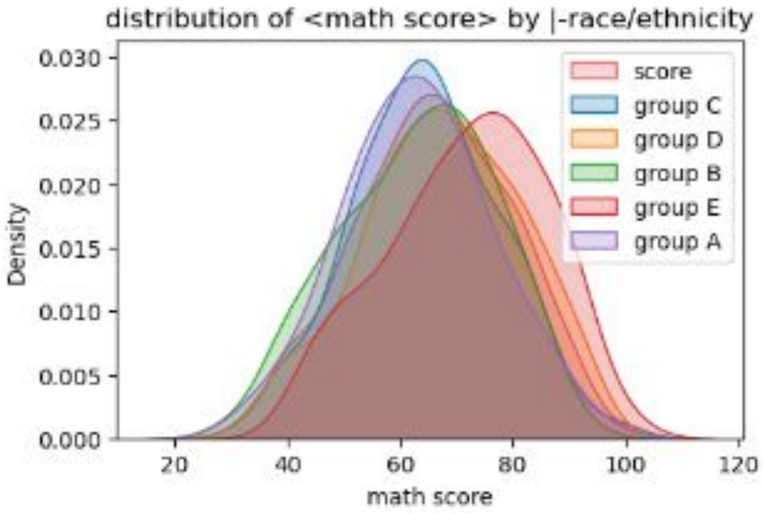
writing score

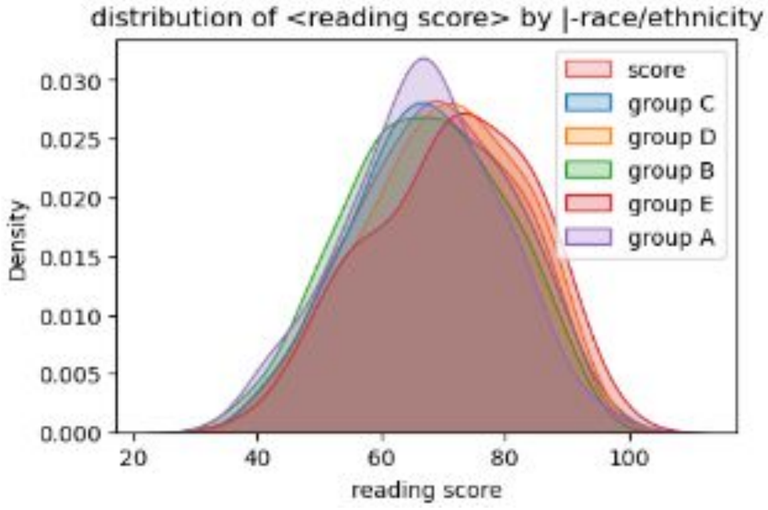
```
< parental level of education >
some college :  68.49152542372882
some high school :  63.78010471204188
high school :  65.50267379679144
associate's degree :  69.33720930232558
bachelor's degree :  71.52
master's degree :  73.54716981132076
```

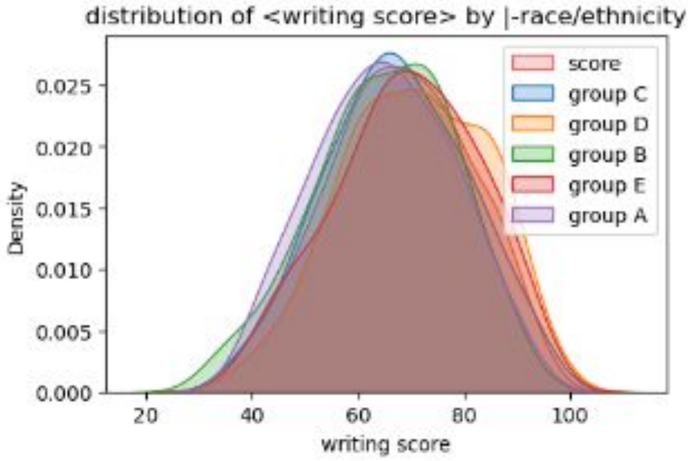


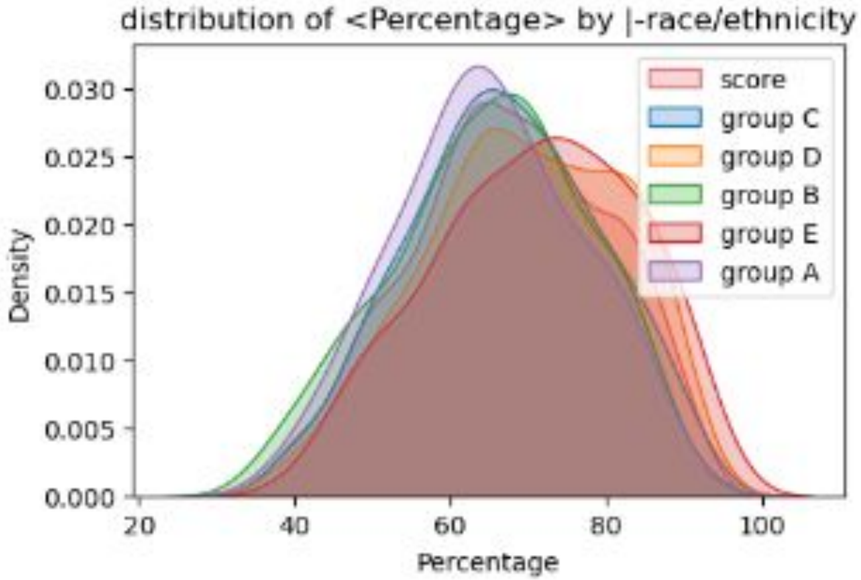
Distribution of scores by column for each test score

race/ethnicity



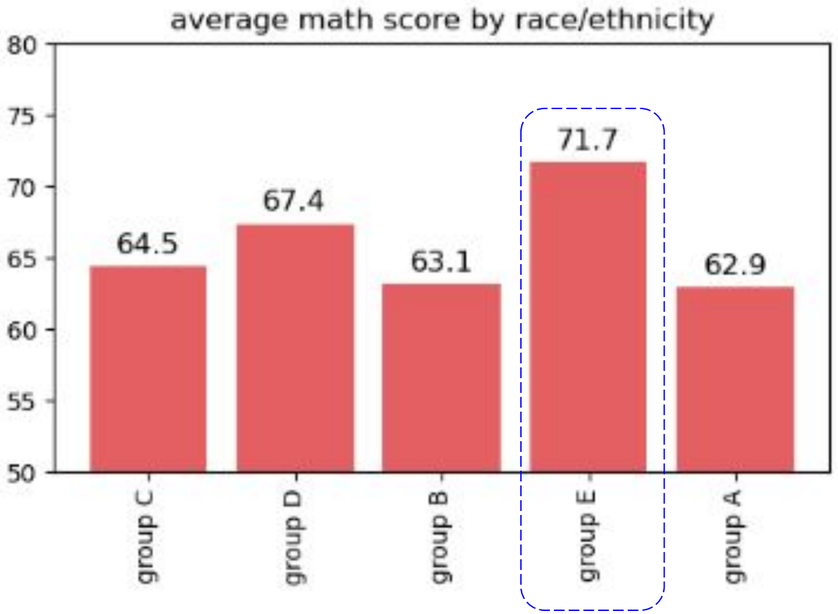






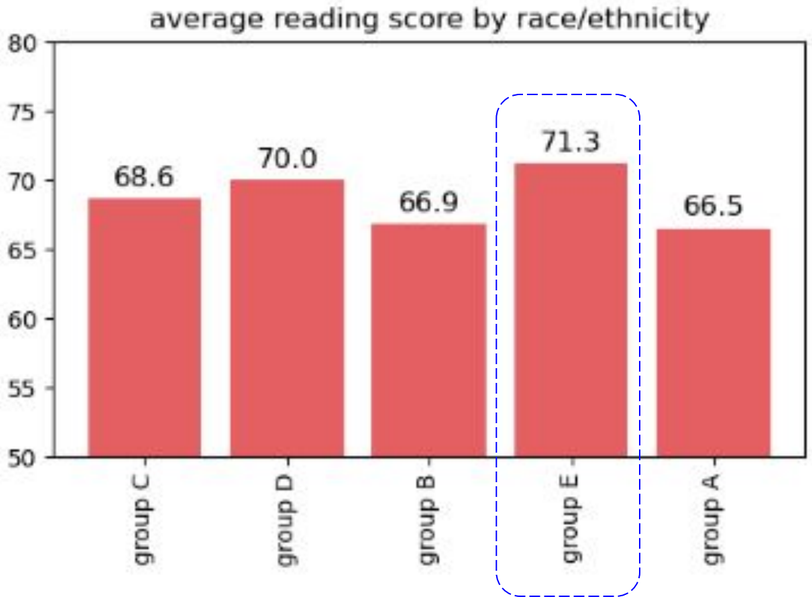
math score

```
< race/ethnicity >
group C :  64.45229681978799
group D :  67.39344262295081
group B :  63.136585365853655
group E :  71.70229007633588
group A :  62.89473684210526
```



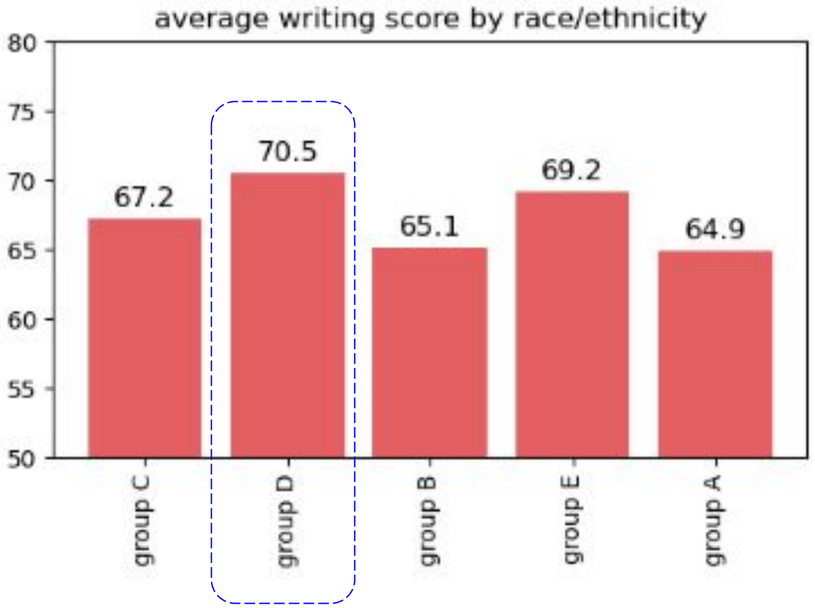
reading score

```
< race/ethnicity >
group C :  68.63604240282686
group D :  70.00819672131148
group B :  66.8780487804878
group E :  71.25954198473282
group A :  66.53947368421052
```



writing score

```
< race/ethnicity >
group C :  67.2155477031802
group D :  70.52459016393442
group B :  65.11219512195122
group E :  69.1526717557252
group A :  64.92105263157895
```



correlation

after Standardization of data and map a number to categorical features we can draw a heatmap correlation to understand relations between features

Standardization

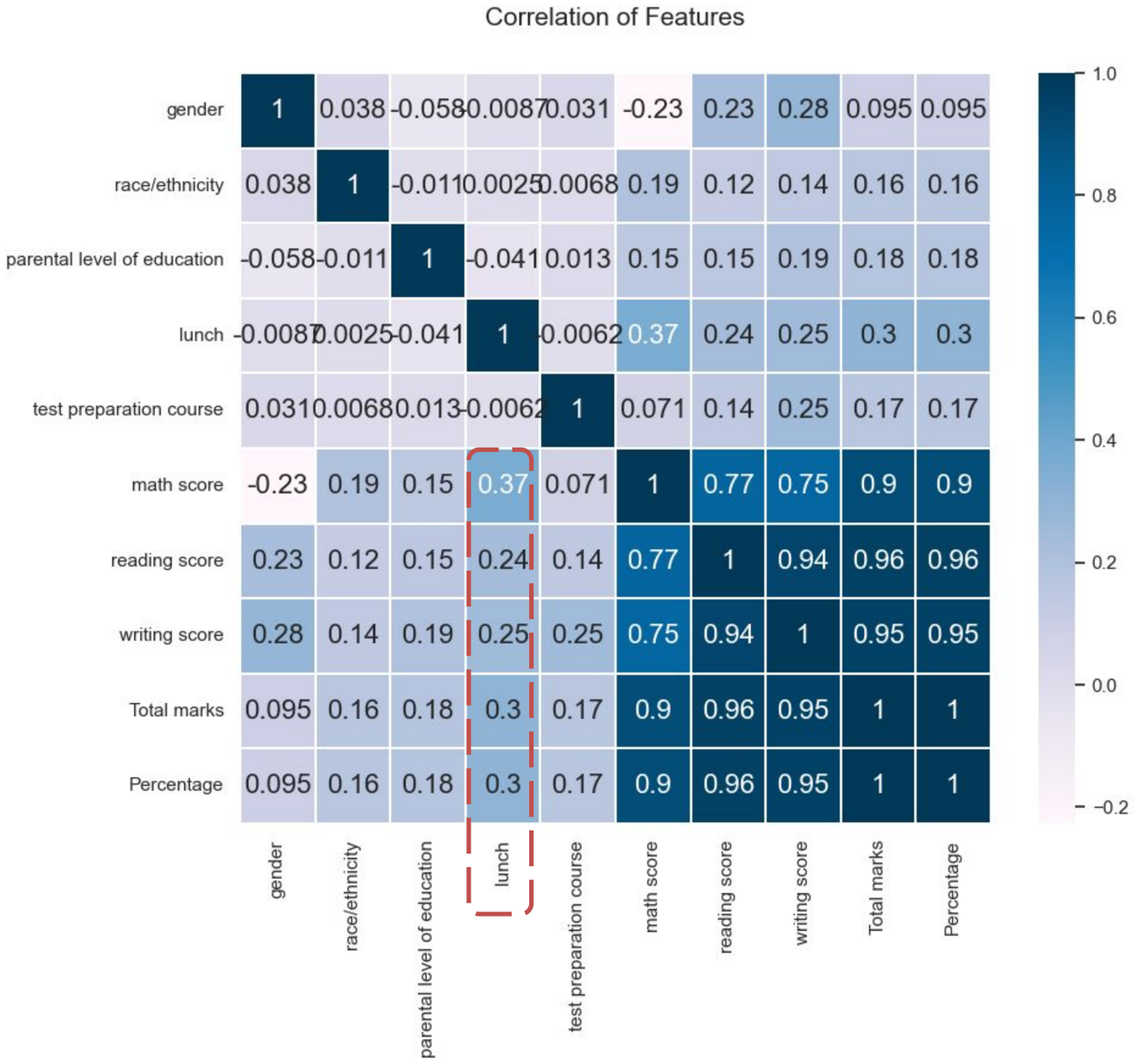
	gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score	Total marks	Percentage
0	1	0.8	0.2	0	0	0.57	0.76	0.69	202	0.673333
1	0	0.8	0.0	0	0	0.39	0.40	0.40	119	0.396667
2	1	0.6	0.4	1	0	0.66	0.67	0.66	199	0.663333
3	1	1.0	0.0	1	0	0.61	0.73	0.74	208	0.693333
4	0	0.2	0.2	1	0	0.48	0.44	0.45	137	0.456667

correlation heatmap

bold results:

according to heatmap, lunch has most effect in scores and specially on math score

test preparation has least effect on math score compared to other scores and most effect on writing score



machine learning (and results)

machine learning

- add column pass_or_nor
- split : data & target / test & train
- baseline
- Compare models using k-fold cross validation
- GaussianNB , SVC, KNN
- results

add column pass_or_nor

1 if Percentage >= 0.6 else 0

Percentage	pass_or_not
0.673333	1
0.396667	0
0.663333	1
0.693333	1
0.456667	0

split : data & target / test & train

target

```
0    1
1    0
2    1
3    1
4    0
```

features

	gender	race/ethnicity	parental level of education	lunch	test preparation course
0	1	0.8	0.2	0	0
1	0	0.8	0.0	0	0
2	1	0.6	0.4	1	0
3	1	1.0	0.0	1	0
4	0	0.2	0.2	1	0

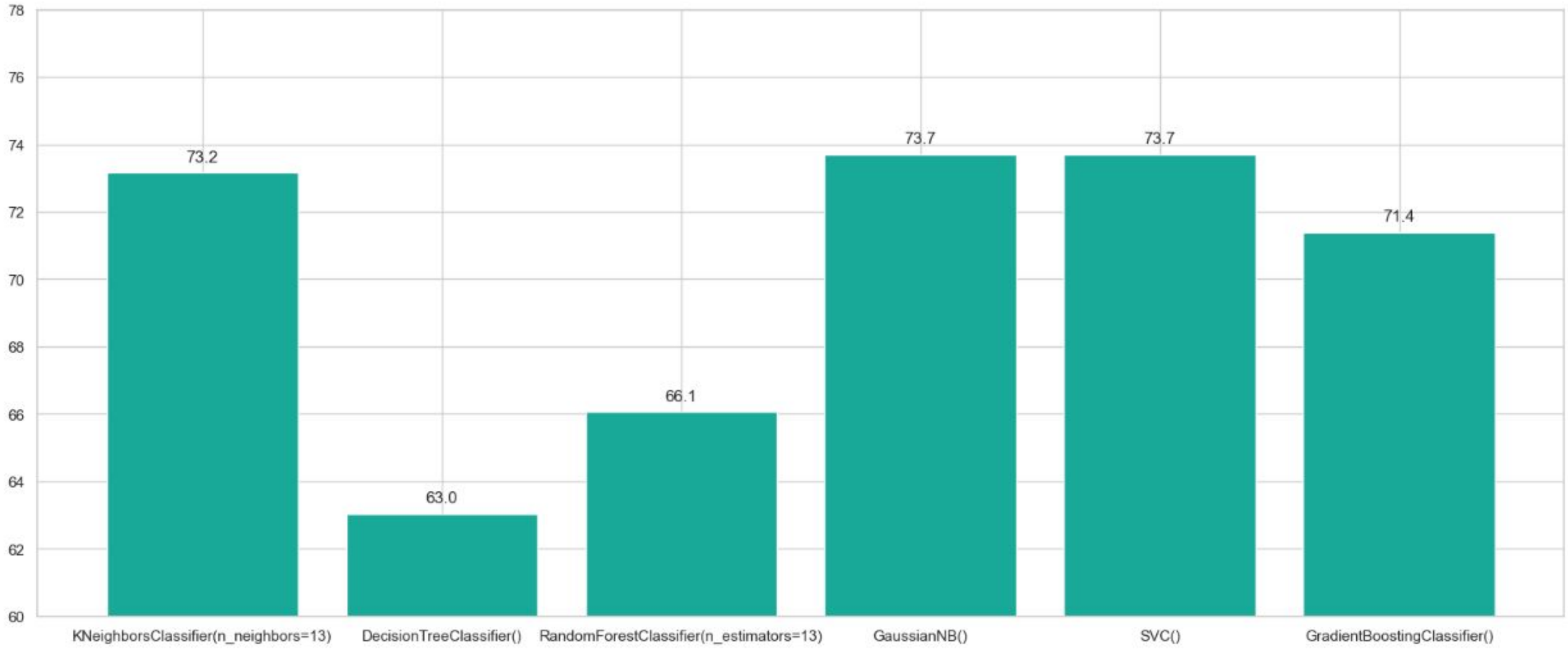
Decision Tree Classifier

Train Accuracy: 0.8063943161634103

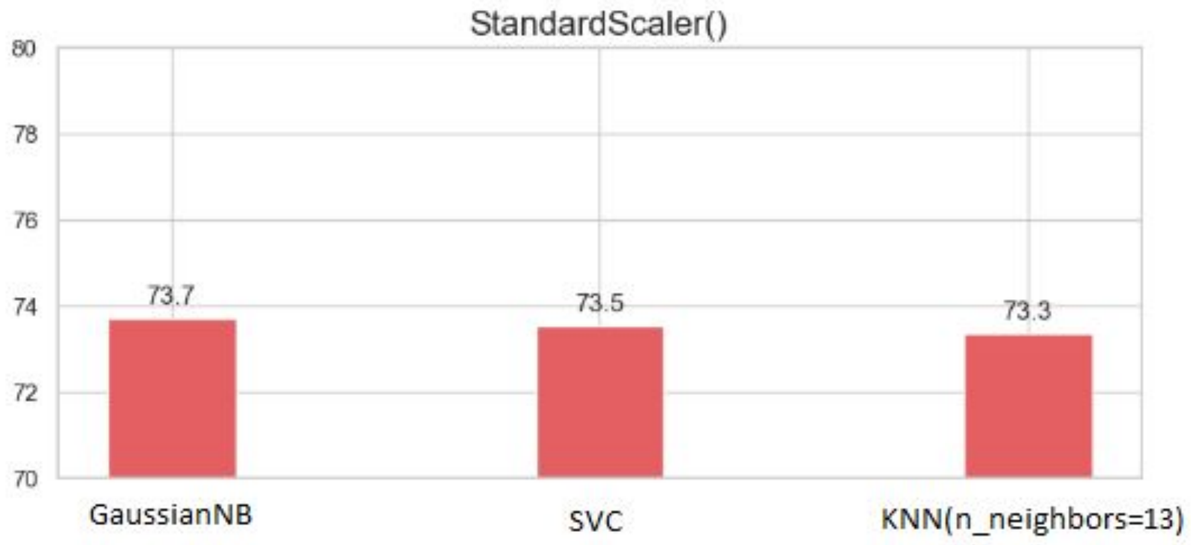
Test Accuracy: 0.648936170212766

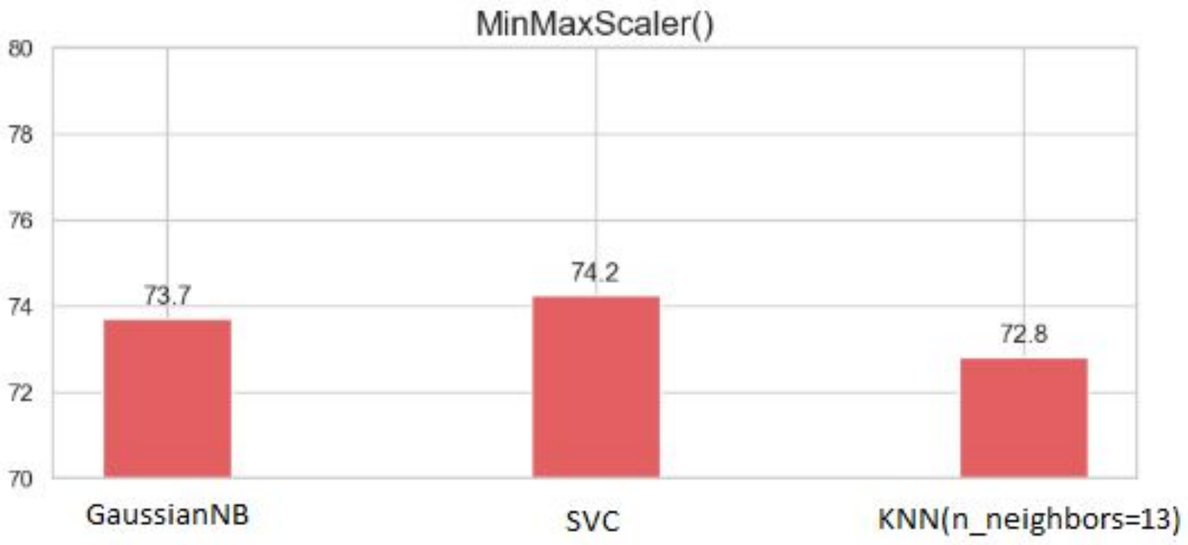
Compare models using k-fold cross validation

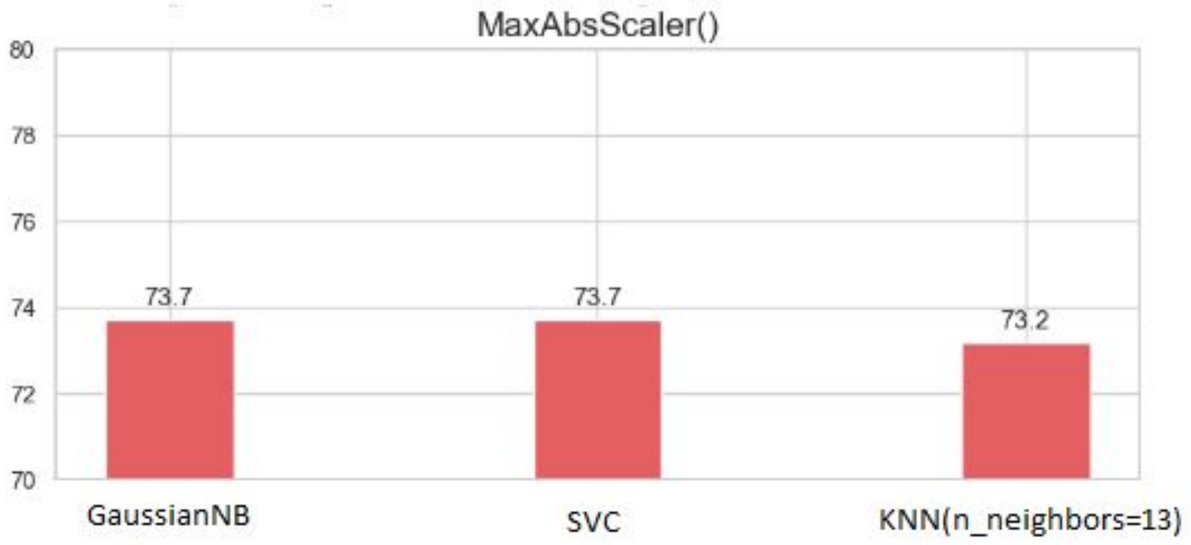
GaussianNB , SVC, KNN have highest score in machine learning models with default tuning

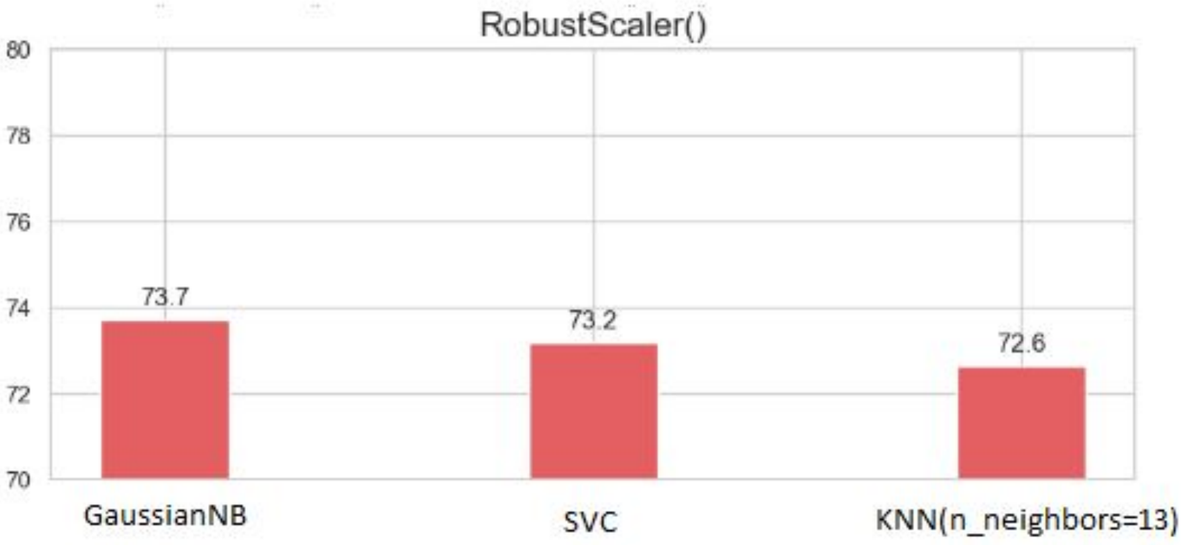


GaussianNB , SVC, KNN









GaussianNB , SVC, KNN

As a result of scaling, it can be seen that the score of GNB is not significantly affected by the scaling model, but nevertheless obtains a high score. Although the SVC and KNN machine learning models were affected, it can be seen that both SVCs scored higher.

only the GNB and SVC models was an efficient method Through the Data scaling, Got the same score in GNB Model.

models

Gaussian Naive Bayes

Random Forest Classifier

according to characteristic of this model and features of this dataset RFC can be prepare model by setting optimal parameter

RFC vs GNB

RFC vs GNB

The previous results show that the Naive Bayes model scored the highest score without tuning, and also the accuracy obtained by finding the optimized tuning value of the Naive Bayes model showed higher accuracy than the tuned Random Forest Classifier. Therefore, since the results may vary depending on the tuning method and parameter setting method of each model, so it is the most efficient way to find a machine learning model that is best suited to the data by improving understanding of the characteristics and performance of each model.

conclusion

conclusion

1. reading score and writing score are linearly related with coef=0.94
2. on average female performed better than male but male has better performance in math score and female has better performance in reading/writing score.
3. group E performed better in math and reading score and group D performed better in writing score . generally group E has best performance .
4. lunch has the **most** effect on scores comparing with other features, **especially on math score**.
5. test preparation has the least effect on math score and most effect on writing score .but generally people who prepared completely for test performed better.
6. students whose parents have a master's degree performed better

What would be the best way to improve student scores on each test?

overall lunch has the best effect on scores, so according to results having a standard lunch has positive effect on scores improvement.

references

http://roycekimmons.com/tools/generated_data/exams

https://www.kaggle.com/code/nihar14/analysis-on-factors-affecting-students-scores

https://github.com/AzT3Risk/Students-performance-in-Exams

https://www.kaggle.com/code/kagleo123/student-perform-in-exam-eda-ml-prediction

https://www.kaggle.com/code/victorferino/student-s-performance-in-exams-eda-ml#Multiple-Linear-Regression-Model