HYPOTHESIS TESTING ON STUDENTS PERFORMANCE IN EXAMS

Course project – Peer review

Wang Xu

2021.12.04.

PROJECT INTRODUCTION

- Case background:
- I am a data analyst and help to research a dataset of students performance in exam, and do the exploratory data analysis and hypothesis testing of significant features which concludes insights/ implications may help audience to make commercial decisions.
- Report to:
- Chief data officer of an education organization or CEO of a company in education industry.

BRIEF DESCRIPTION OF THE DATA SET AND A SUMMARY OF ITS ATTRIBUTES

- Students Performance in Exams from *Kaggle*
- https://www.kaggle.com/gaganmaahi224/clean-eda-queries-visualization-for-beginners/data
- Data has 1000 rows, 8 columns. (1000 students' scores)
- Numerical features: math score, reading score, writing score.
- Categorical features: Gender, race/ethnicity, parental level of education, lunch, test preparation course.

	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

INITIAL PLAN FOR DATA EXPLORATION

- Find data / Import data / Import libraries
- Understand data / Data visualization
- Exploratory data analysis / Data cleaning
- Hypothesis Testing
- Statistics analysis on hypothesis
- Insights / Conclusions

ACTIONS TAKEN FOR DATA CLEANING AND FEATURE ENGINEERING (UNDERSTAND DATA)

Understanding Data structure, and Check missing values.

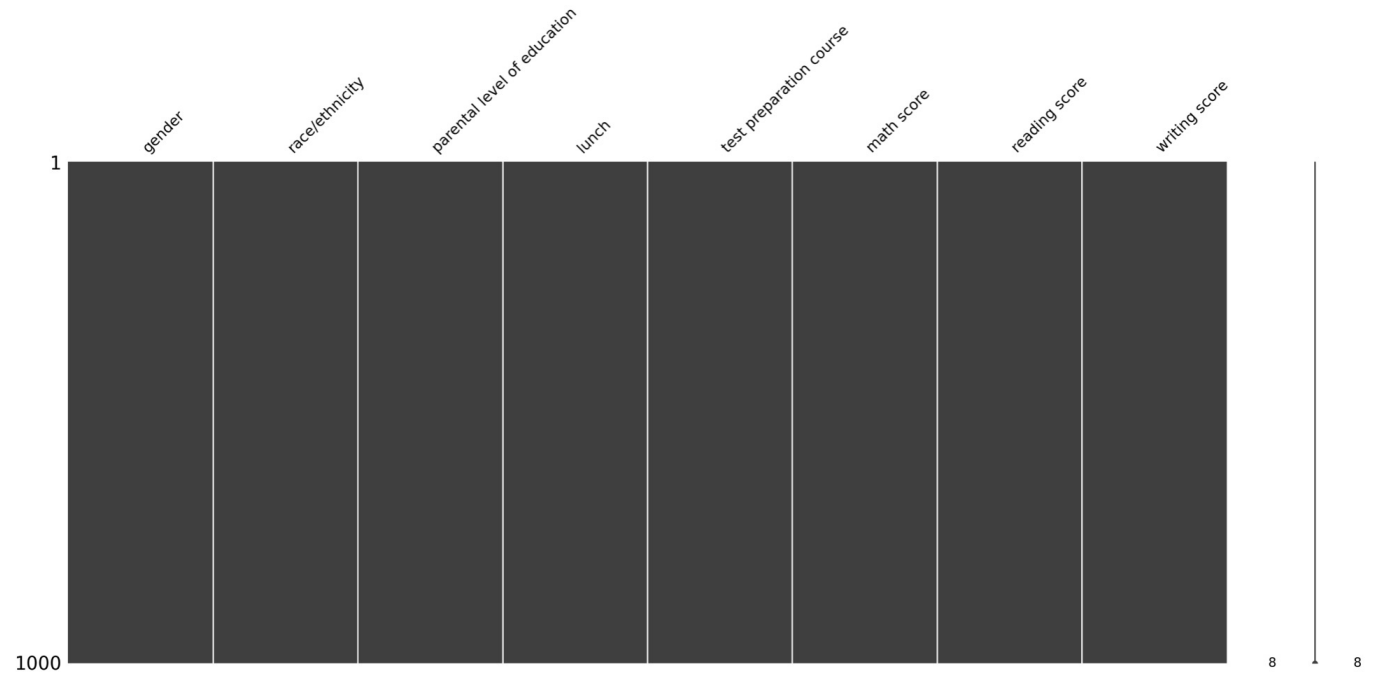
```
1 df.describe()
```

	math score	reading score	writing score
count	1000.00000	1000.000000	1000.000000
mean	66.08900	69.169000	68.054000
std	15.16308	14.600192	15.195657
min	0.00000	17.000000	10.000000
25%	57.00000	59.000000	57.750000
50%	66.00000	70.000000	69.000000
75%	77.00000	79.000000	79.000000
max	100.00000	100.000000	100.000000

2.2.4 check missing values

```
1 msno.matrix(df) #Check if has missing value
```

```
<AxesSubplot:>
```



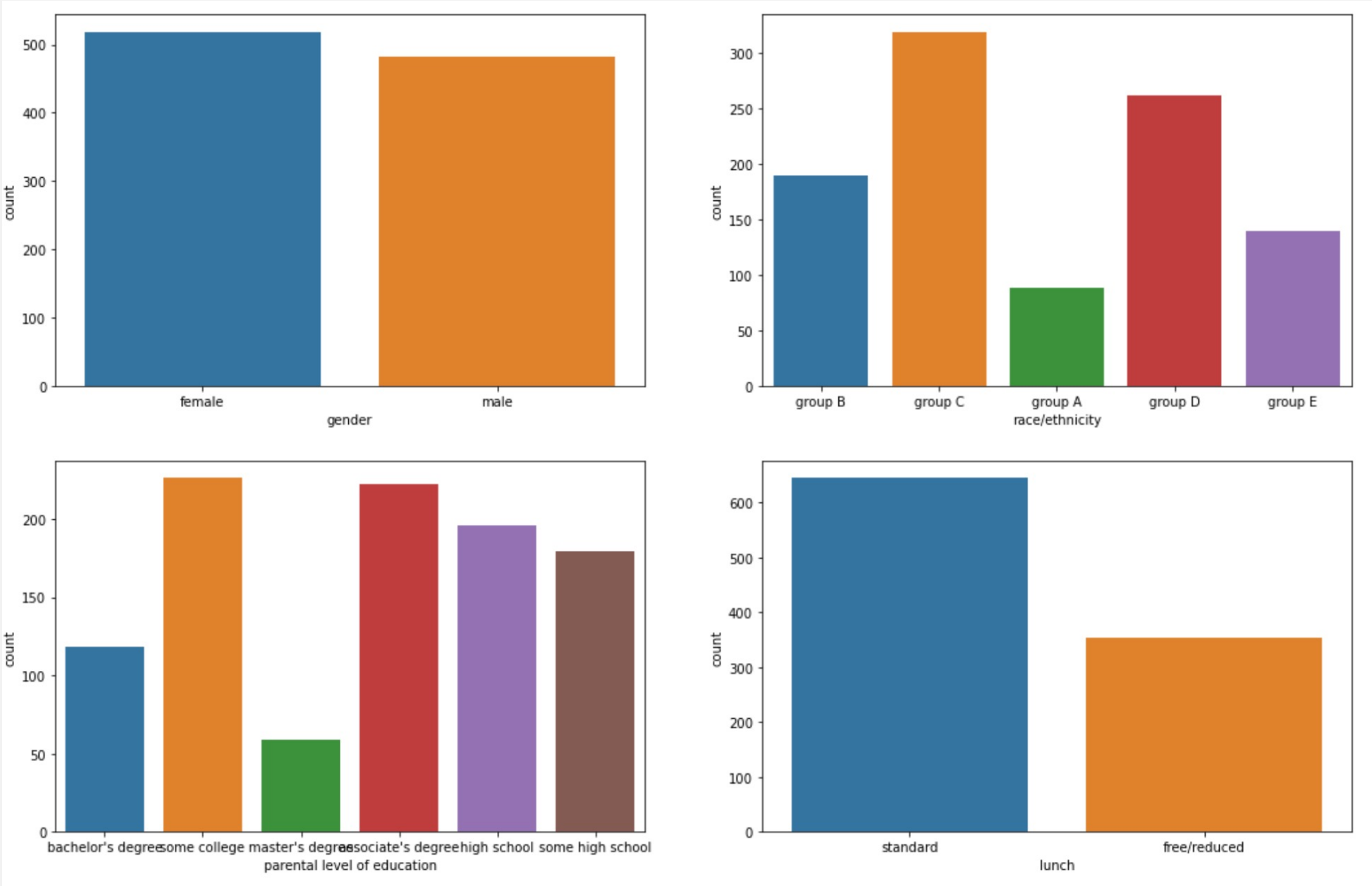
2.2.5 categorial data

```
for i in ['gender','race/ethnicity','parental level of education','lunch','test preparation course']:
        print(i+' distribution')
        print(df[i].value_counts())
        print('-----------------------------------------------')
```

```
gender distribution
female    518
male      482
Name: gender, dtype: int64
-----------------------------------------------
race/ethnicity distribution
group C    319
group D    262
group B    190
group E    140
group A     89
Name: race/ethnicity, dtype: int64
-----------------------------------------------
parental level of education distribution
some college          226
associate's degree    222
high school           196
some high school      179
bachelor's degree     118
master's degree        59
Name: parental level of education, dtype: int64
-----------------------------------------------
lunch distribution
standard        645
free/reduced    355
Name: lunch, dtype: int64
-----------------------------------------------
test preparation course distribution
none         642
completed    358
Name: test preparation course, dtype: int64
-----------------------------------------------
```

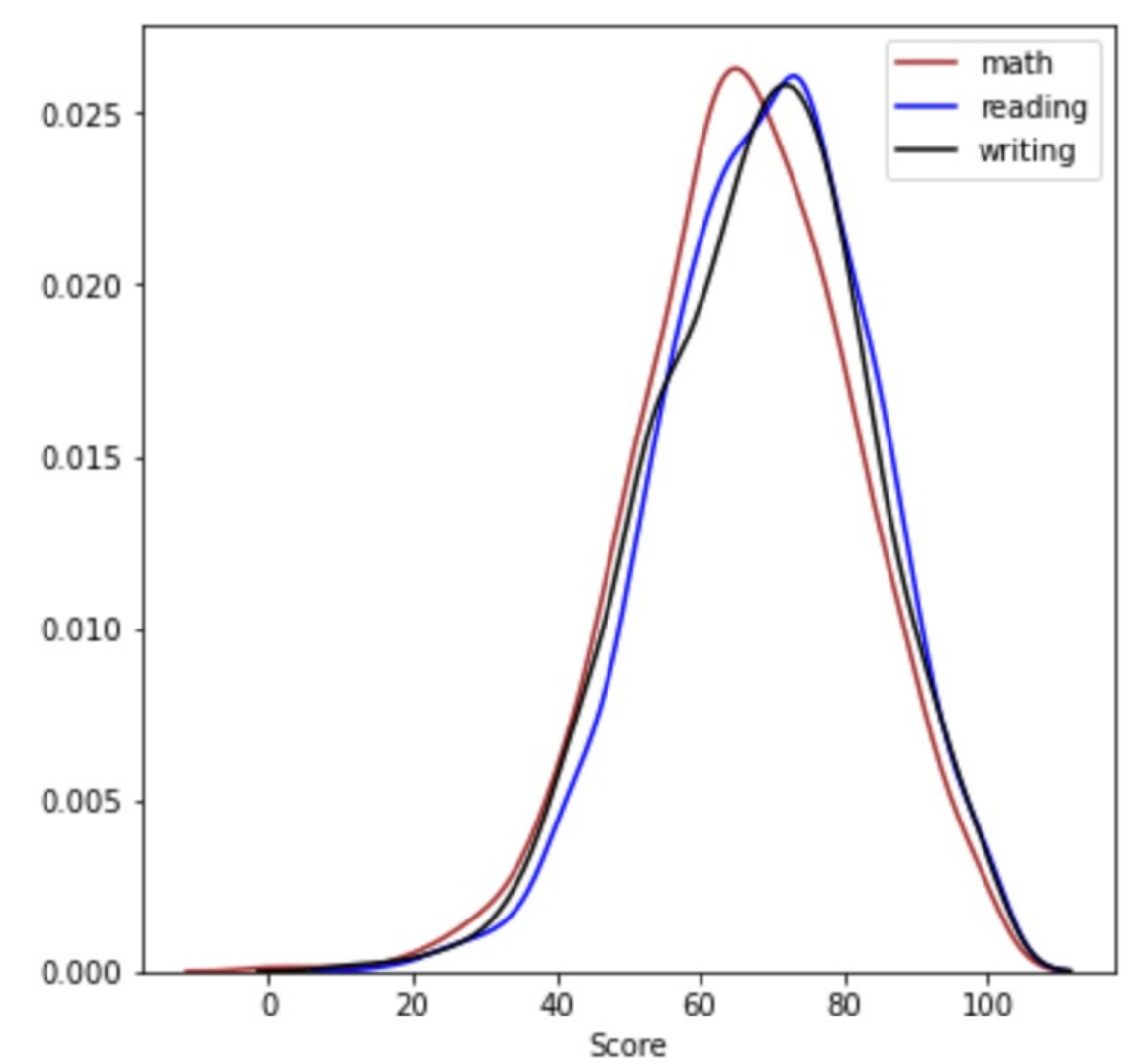
EXPLETORY DATA ANALYSIS

- Frequency on each feature

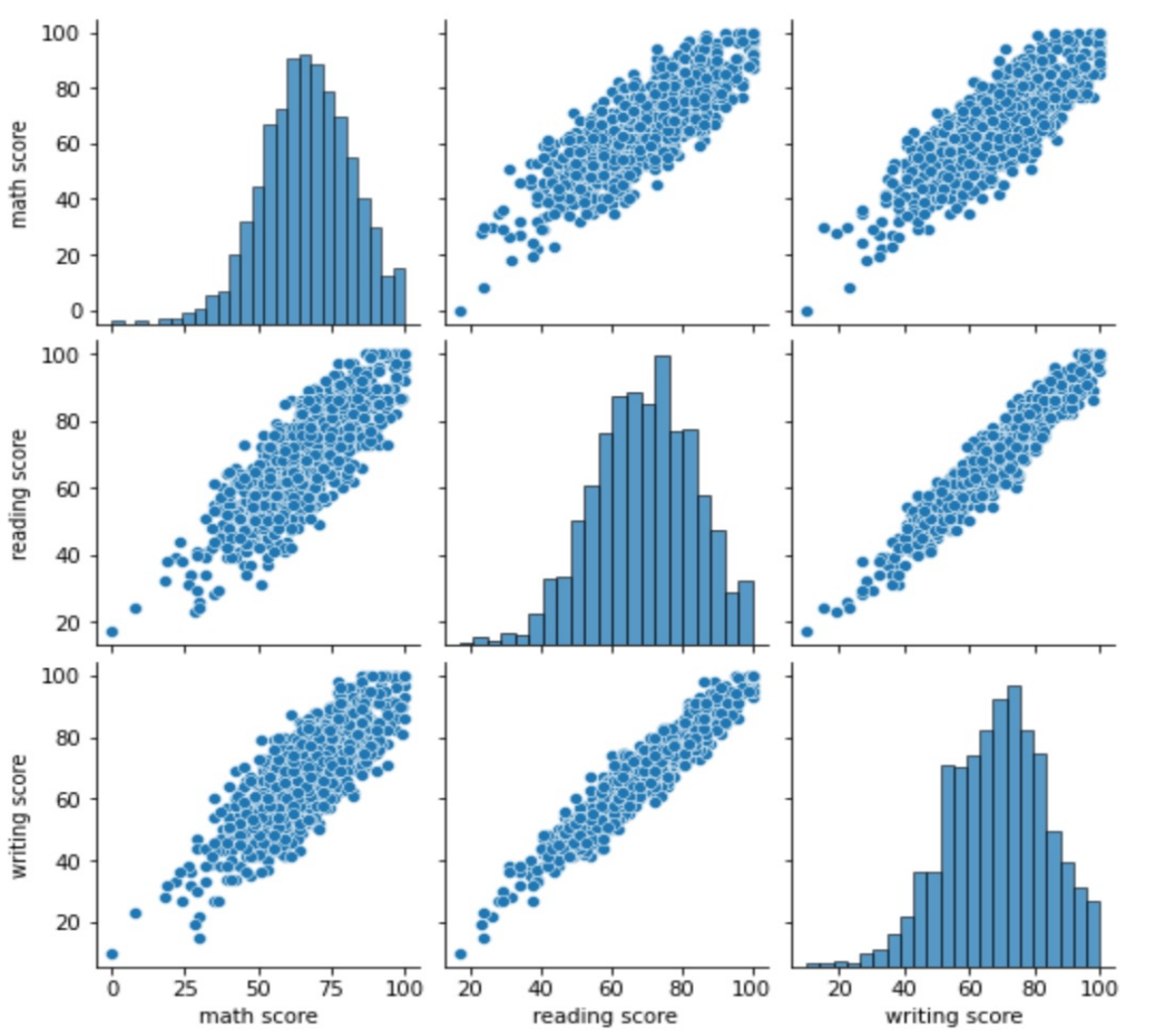


EXPLETORY DATA ANALYSIS

- Scores distribution



- Correlation between scores



EXPLETORY DATA ANALYSIS

- Correlation between categorical features with scores

gender	math score	reading score	writing score
female	63.633205	72.608108	72.467181
male	68.728216	65.473029	63.311203

race/ethnicity	math score	reading score	writing score
group A	61.629213	64.674157	62.674157
group B	63.452632	67.352632	65.600000
group C	64.463950	69.103448	67.827586
group D	67.362595	70.030534	70.145038
group E	73.821429	73.028571	71.407143

lunch	math score	reading score	writing score
free/reduced	58.921127	64.653521	63.022535
standard	70.034109	71.654264	70.823256

parental level of education	math score	reading score	writing score
associate's degree	67.882883	70.927928	69.896396
bachelor's degree	69.389831	73.000000	73.381356
high school	62.137755	64.704082	62.448980
master's degree	69.745763	75.372881	75.677966
some college	67.128319	69.460177	68.840708
some high school	63.497207	66.938547	64.888268

test preparation course	math score	reading score	writing score
completed	69.695531	73.893855	74.418994
none	64.077882	66.534268	64.504673

```
1 df.groupby(['parental level of education']).mean().style.background_gradient(cmap='PuRd')
```

EXPLETORY DATA ANALYSIS

- Data cleaning

parental level of education	math score	reading score	writing score
associate's degree	67.882883	70.927928	69.896396
bachelor's degree	69.389831	73.000000	73.381356
high school	62.137755	64.704082	62.448980
master's degree	69.745763	75.372881	75.677966
some college	67.128319	69.460177	68.840708
some high school	63.497207	66.938547	64.888268

parental level of education	math score	reading score	writing score
bachelor	67.895760	70.773852	70.201413
high school	62.786667	65.770667	63.613333
master	69.745763	75.372881	75.677966

```
df2=df.copy()
df2['parental level of education'] = df2['parental level of education'].replace({"some college":"bachelor",
                                        "associate's degree":"bachelor",
                                        "some high school":"high school",
                                       "bachelor's degree":"bachelor",
                                       "master's degree":"master"})
df2['parental level of education'].value_counts()
```

```
bachelor       566
high school    375
master          59
Name: parental level of education, dtype: int64
```

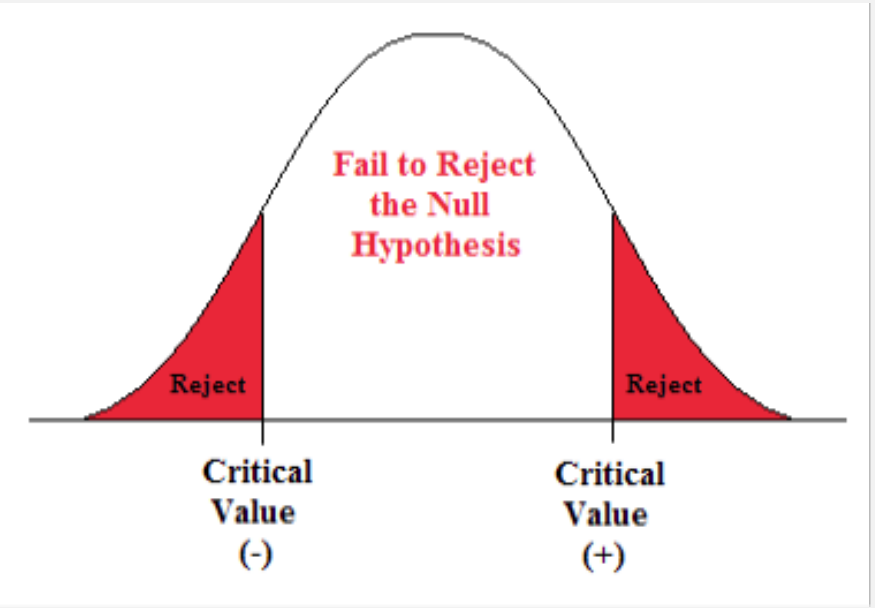
KEY FINDINGS AND INSIGHTS FROM EXPLETORY DATA ANALYSIS

- Strong correlations between math scores, reading score, writing scores.
- Mean math score of male students is higher than female.
- Mean reading score and writing score of female students are higher than male.
- Mean scores (math, reading, writing) of students whose parental with high school level of education are the lowest than others.
- Mean scores of students whom had standard lunch are higher than others whom had free/reduced lunch.
- Students who had completed test preparation courses have higher scores than others who had not.

FORMULATING AT LEAST 3 HYPOTHESIS ABOUT THIS DATA

3 hypothesis to define:

- Gender vs. Scores
- Parental level of education vs. Scores.
- Lunch vs. Scores



STATISTICS TEST ON HYPOTHESES 1 GENDER VS. SCORES

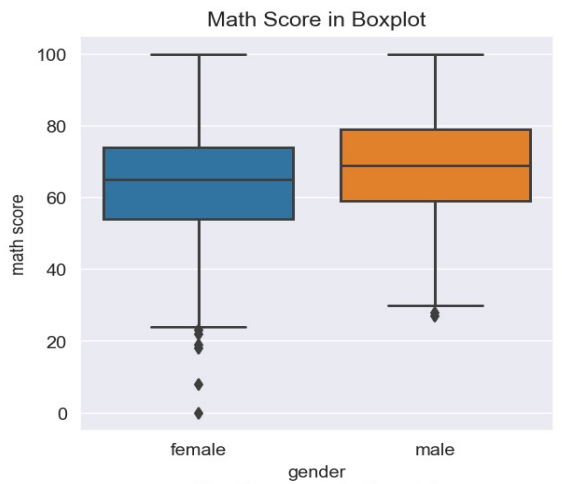
- On math scores:
- Male students are higher, female has more outliers. And distributions of male are more concentrated.

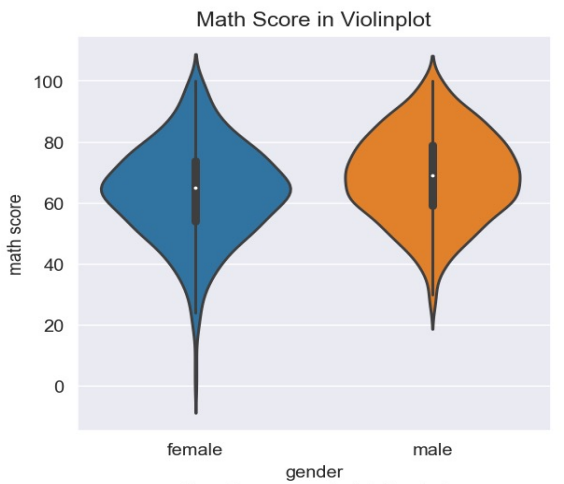
- On Reading scores:
- Female students are higher, but has more outliers with very low scores. Female students are more polarized than male students which have a wider distribution.

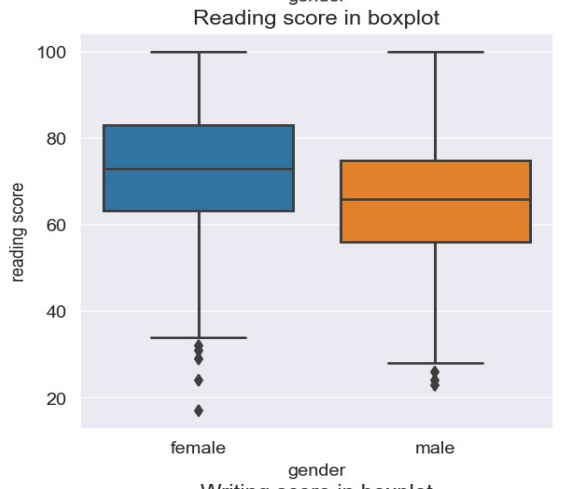
- On Writing scores:
- Very similar with reading scores situations.

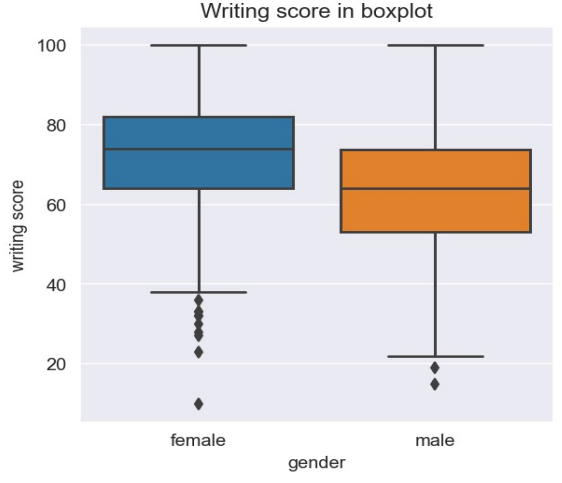
- *Does gender difference has impact on their scores?*

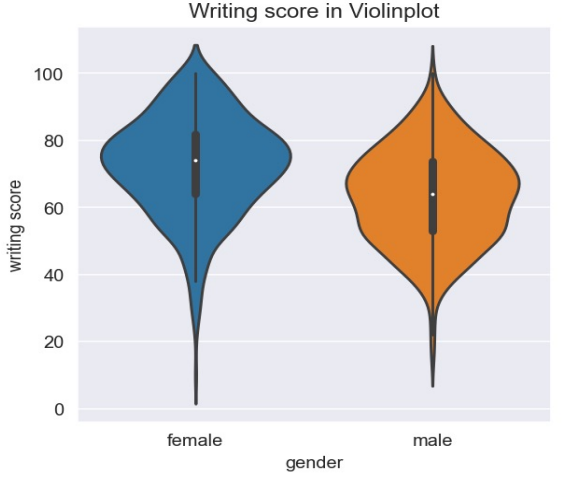
- T-test on two independent variables:
- females scores vs. males scores











Hypothesis testing 1 : Gender vs. Scores

H0: There is no impact on scores if different in gender.

H1: There is having impacts on scores if different in gender.

Conclusion: Gender has impacts on scores.

T-test on gender (male & female 2 independent variables) on math score:

```
1 scipy.stats.ttest_ind(df_f['math score'],df_m['math score'],equal_var=False)
```

```
Ttest_indResult(statistic=-5.398000564160736, pvalue=8.420838109090415e-08)
```

As P value < 0.05, therefore reject H0 and accept H1, which is there is difference between gender on math score. From the average score, male on math is considered better than female.

T-test on gender (male & female 2 independent variables) on reading score:

```
1 scipy.stats.ttest_ind(df_f['reading score'],df_m['reading score'],equal_var=False)
```

```
Ttest_indResult(statistic=7.9683565184844, pvalue=4.3762967534976715e-15)
```

As P value < 0.05, therefore reject H0 and accept H1, which is there is difference between gender on reading score. From the average score, female on reading is considered better than male.

T-test on gender (male & female 2 independent variables) on writing score:

```
1 scipy.stats.ttest_ind(df_f['writing score'],df_m['writing score'],equal_var=False)
```

```
Ttest_indResult(statistic=9.997718973491885, pvalue=1.7118093718497237e-22)
```

As P value < 0.05, therefore reject H0 and accept H1, which is there is difference between gender on writing score. From the average score, female on writing is considered better than male.

STATISTICS TEST ON HYPOTHESES 2 LEVEL OF EDUCATION VS. SCORES

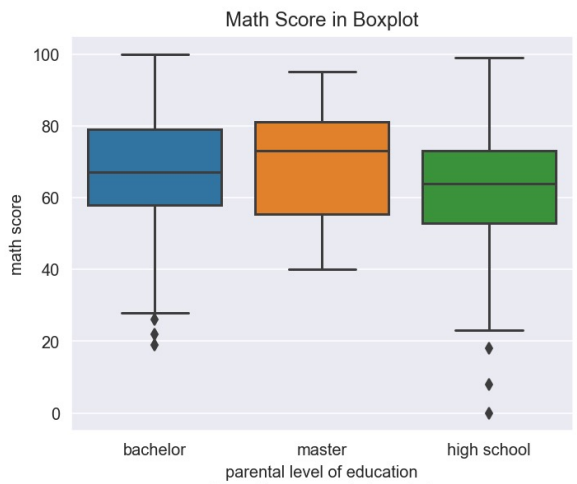
- Students whose parents have <u>master degree</u> have highest mean scores on math, reading, writing scores. Its distributions are more close than other two and more closely concentrated on higher scores with no outliers

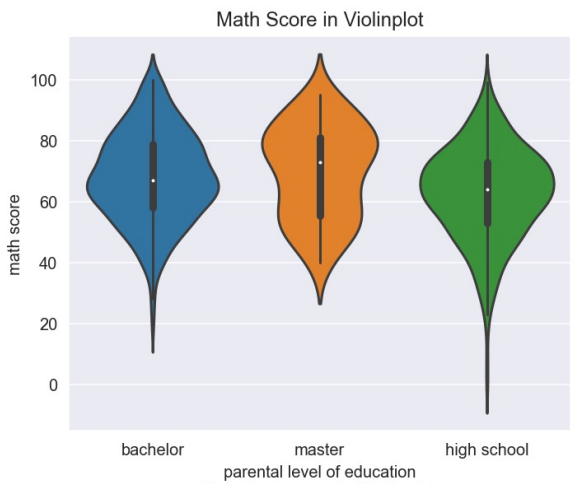
- Students whose parents have <u>bachelor degree</u> have the middle mean scores. Its distributions are relatively wider than "master" with few outliers.

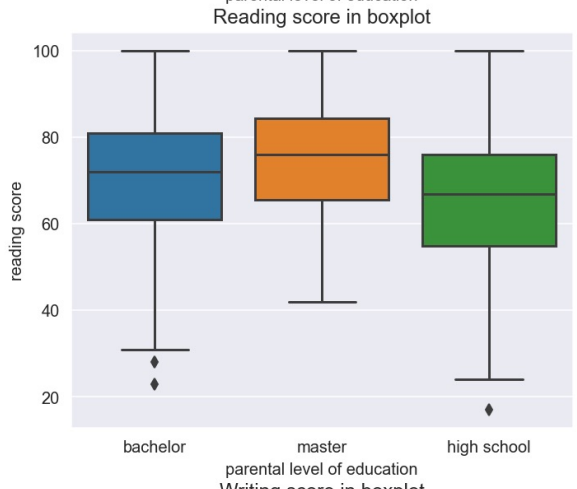
- Students whose parents have <u>high school level of education</u> have the lowest mean scores on each course. Its distribution is the widest and most polarized among all.

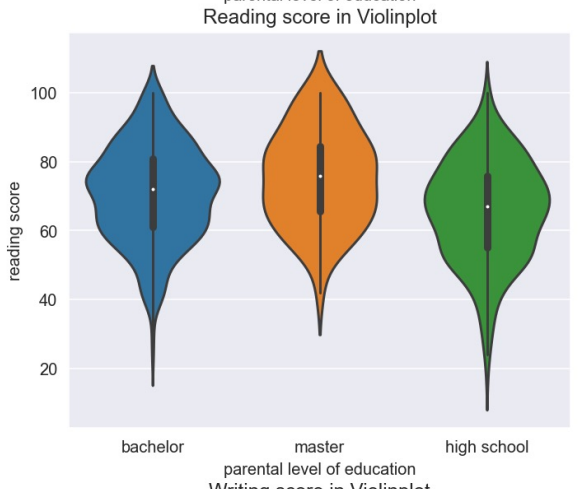
- It seems that parents' educational level has some impacts on students scores. So higher educational level of parents will deliver higher scores of their children? Is it TRUE?

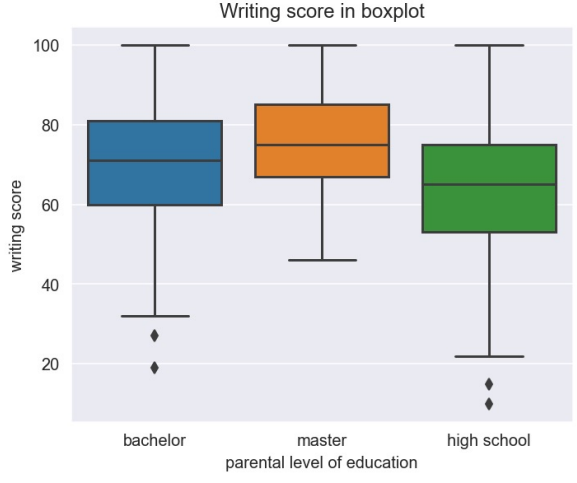
- T-test on each pair of variables on each course:
- Master vs. Bachelor
- Bachelor vs. high school
- Master vs. high school

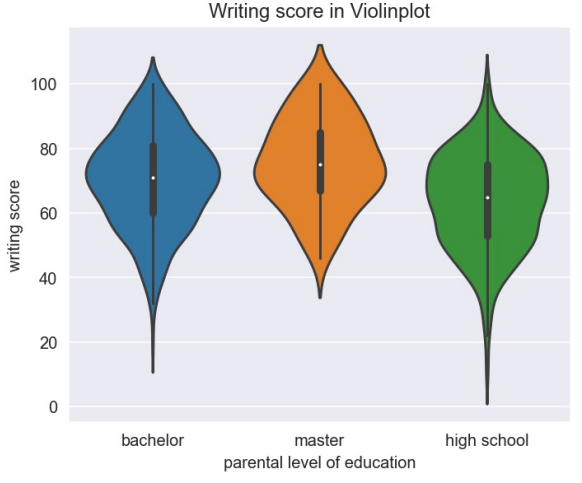












Hypothesis testing 2 : Parental level of education vs. Scores

H0:There is no impacts on scores if different in parental level of education.
H1:There is having impacts on scores if different parental level of education.

```
df2_ba = df2[df2['parental level of education']=='bachelor'] #parent has bachelor degree
df2_ma = df2[df2['parental level of education']=='master'] #parent has master degree
df2_hi = df2[df2['parental level of education']=='high school'] #parent has high school education
```

- On Math score:
- Master vs. Bachelor
- P-value>0.05, accept H0.
- No significant impacts!

- Bachelor vs. high school
- P-value<0.05, accept H1. Has impacts!

- Master vs. high school
- P-value<0.05, accept H1. Has impacts!

T-test on parent is bachelor or master level of education has impact on student math score.

```
scipy.stats.ttest_ind(df2_ba['math score'],df2_ma['math score'], equal_var=False)
```

```
Ttest_indResult(statistic=-0.8945328727367031, pvalue=0.3741037356058158)
```

As p-value > 0.05, so accept H0 (Either parental level of education is bachelor or master, no significant impact on student math score.

T-test on parent is high school or master level of education has impact on student math score.

```
scipy.stats.ttest_ind(df2_hi['math score'],df2_ma['math score'], equal_var=False)
```

```
Ttest_indResult(statistic=-3.277139142603498, pvalue=0.0015702752093248236)
```

As p-value<0.05, so accept H1. Parental level of education in master has positive impact on student math score than whose parent has high school education.

T-test on parent is high school or bachelor level of education has impact on student math score.

```
scipy.stats.ttest_ind(df2_hi['math score'],df2_ba['math score'], equal_var=False)
```

```
Ttest_indResult(statistic=-5.103857329611522, pvalue=4.179301907606294e-07)
```

As p-value<0.05, so accept H1. Parental level of education in bachelor has positive impact on student math score than whose parent has high school education.

Hypothesis testing 2 : Parental level of education vs. Scores

H0:There is no impacts on scores if different in parental level of education.
H1:There is having impacts on scores if different parental level of education.

- On Reading scores:
- all p-valus<0.05, reject H0.
- Has significant impacts on student reading scores!

- On Writing scores:
- Master vs. Bachelor
- p-value >0.05, accept H0.
- No significant impacts!

- Bachelor vs. high school
- Master vs. high school
- p-value<0.05, accept H1.
- Have significant impacts!

2.3.4.3 Parental level of education vs. Reading score

```
1 scipy.stats.ttest_ind(df2_ba['reading score'],df2_ma['reading score'], equal_var=False)
```

```
Ttest_indResult(statistic=-2.4354734401838174, pvalue=0.01737773810566553)
```

```
1 scipy.stats.ttest_ind(df2_hi['reading score'],df2_ma['reading score'], equal_var=False)
```

```
Ttest_indResult(statistic=-4.924988125014135, pvalue=4.412639124360135e-06)
```

```
1 scipy.stats.ttest_ind(df2_hi['reading score'],df2_ba['reading score'], equal_var=False)
```

```
Ttest_indResult(statistic=-5.174732319236381, pvalue=2.9123773693658355e-07)
```

2.3.4.4 Parental level of education vs Writing score

```
1 scipy.stats.ttest_ind(df2_ba['writing score'],df2_ma['writing score'], equal_var=False)
```

```
Ttest_indResult(statistic=-2.894432095872074, pvalue=0.005011521767456111)
```

```
1 scipy.stats.ttest_ind(df2_hi['writing score'],df2_ma['writing score'], equal_var=False)
```

```
Ttest_indResult(statistic=-6.1975271886286105, pvalue=2.2580401650022226e-08)
```

```
1 scipy.stats.ttest_ind(df2_ba['writing score'],df2_hi['writing score'], equal_var=False)
```

```
Ttest_indResult(statistic=6.659413040259989, pvalue=5.129182719510424e-11)
```

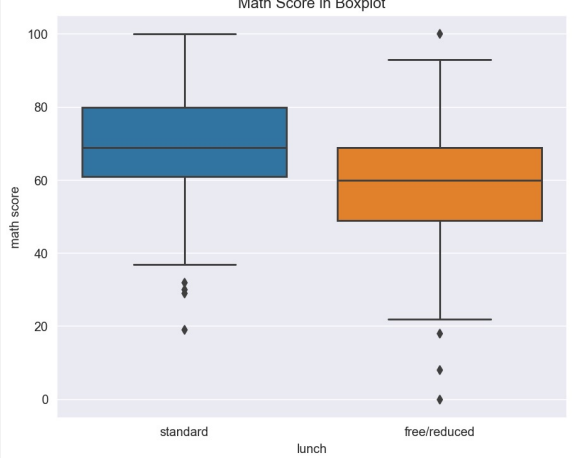
Parental level of education / Students scores	Math scores	Reading scores	Writing scores
Master vs. Bachelor	P>0.05, accept H0. Neither has significant impact.	P<0.05, reject H0.	P>0.05, accept H0.
Master vs. High school	P<0.05, reject H0. Has significant impacts on scores.	P<0.05, reject H0.	P<0.05, reject H0.
Bachelor vs. High school	P<0.05, reject H0.	P<0.05, reject H0.	P<0.05, reject H0.

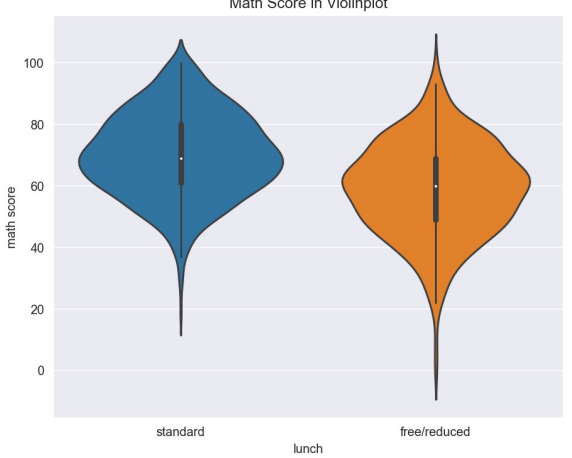
Conclusion:

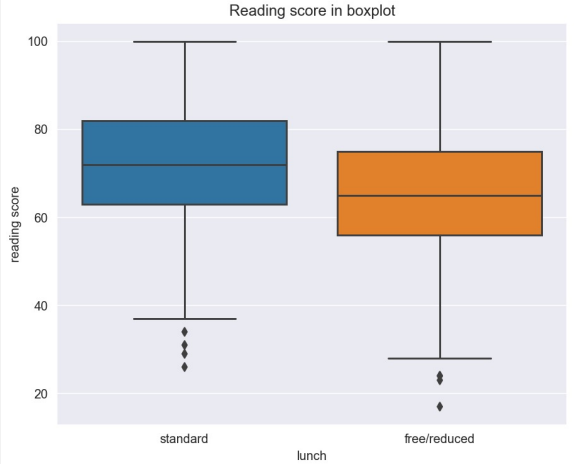
- Bachelor educational level of parents or above have significant positive impacts on children's scores.
- Inversely, high school level of parents have significant negative impacts on children's scores.

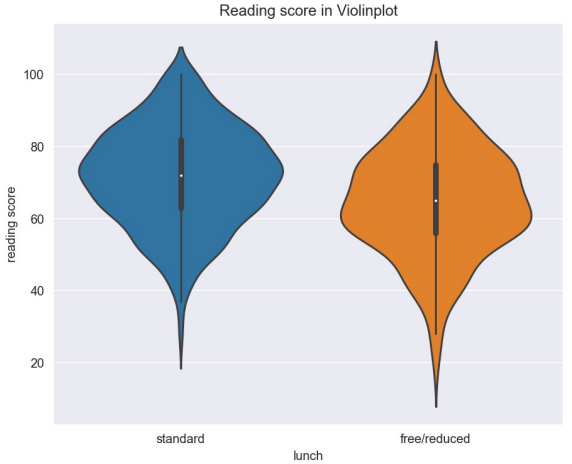
STATISTICS TEST ON HYPOTHESES 3 LUNCH VS. SCORES

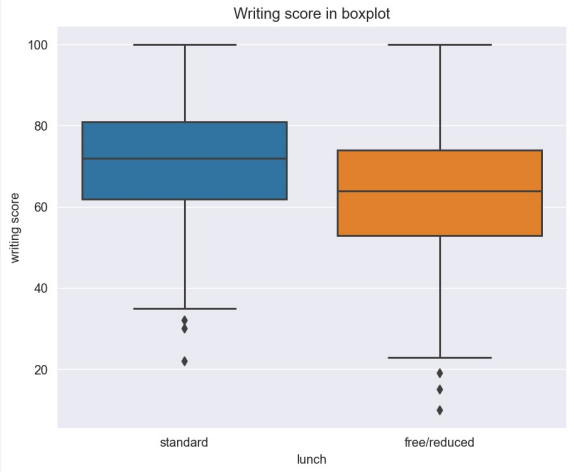
- From boxplot and violinplot, it seams students who have free/reduced lunch get lower mean scores in every course with some outliers and wider distribution.
- On the contrary, students who have standard lunch get higher scores and more concentrated distribution with few outliers.
- Having paid for lunch seems as a very unlikely factor to affect on students scores, however, it can be proved wrong.
- T-test on two independent variables:
- ‘Free/reduced lunch’ vs. ‘standard lunch’

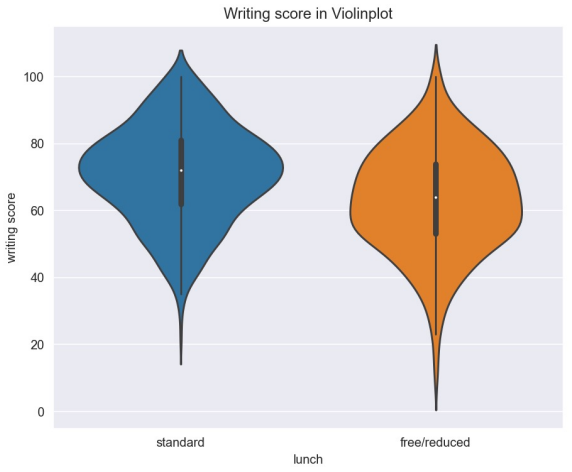












Hypothesis testing 3 : Lunch vs. Scores

H0:There is no impacts on scores if different in lunch type.
H1:There is having impacts on scores if different parental lunch type.

All p-value<0.05, reject H0 and accept H1.
There is significant impacts on scores if lunch is free/reduced or standard.
In particular, with standard lunch students have greater scores than whom has free/reduced lunch.

T-test on lunch vs. math score, reading score, writing score (H0: No impact on score if lunch is free or standard.)

```
1 scipy.stats.ttest_ind(df_fl['math score'],df_sl['math score'],equal_var=False)
```

```
Ttest_indResult(statistic=-11.484100293169273, pvalue=5.539584943965394e-28)
```

```
1 scipy.stats.ttest_ind(df_fl['reading score'],df_sl['reading score'],equal_var=False)
```

```
Ttest_indResult(statistic=-7.29261459119927, pvalue=8.421688691948049e-13)
```

```
1 scipy.stats.ttest_ind(df_fl['writing score'],df_sl['writing score'],equal_var=False)
```

```
Ttest_indResult(statistic=-7.840866279153781, pvalue=1.7161468025322293e-14)
```

SUGGESTIONS FOR NEXT STEPS IN ANALYZING THIS DATA

More data cleaning
More hypothesis tests
Commercial and valuable approaches
Predictable algorithms

- Parental level of education may need to divide into two categories to prove that parents educational level is significantly impact on children's scores.
- Define more hypothesis on Race/ethnicity and Test preparation course to prove how they have significant impacts on students.
- From each hypothesis, some implications and its corresponding approach on commercial moves will be discussed with audiences.
- Furthermore, appropriate algorithms will be needed to predict in new students' score, according to new students' corresponding features.

A PARAGRAPH THAT SUMMARIZES THE QUALITY OF THIS DATA SET AND A REQUEST FOR ADDITIONAL DATA IF NEEDED

- **Summary**
- This dataset has very well quality with no missing values and easy to understand each feature.
- **More to research (if needed)**
- According to three hypothesis tests above, if data gathering on parents' information, such as *annual revenue, occupations, locations* will help to locate more target parents to improve students' scores.
- About the lunch feature, it needs more evidence to prove that economic situation of parents will affect student score performances.
- Also, parents with high school education level feature also affects the children scores which give more space for org./company to show/educate targeted parents to pay more attention on importance of student scores improvement to earn higher level of education in future.

THANK YOU!

Course project - Peer review – EDA & hypothesis testing on dataset

Wang Xu

2021.12.04.