Data Analysis of Parental Qualification On The Performance of Thier Children

1)Introdcution

[ENG]The data-set,"Student's Performance in Exams' has arisen my question on whether the level of parental's educational background should give an impact on their children's achievements in subjects. To answer the question, we will go through a series of steps including data manipulation, graphical representation, conclusions based on the results.

[KOR] "학생의 수학능력 수행"의 데이타 셋은 과연 부모의 학력이 자식들의 학업능력에 어떠한 영향을 미치는 지에 궁금중을 유발시켰습니다. 이러한 호기심을 충족시키려, 데이터 조작, 그래프, 결과에 대한 결론을 포함한 일련의 과정들을 수행해 볼 예정입니다.

[ENG]Visit the website given below to find raw data and full information about it.

[KOR]데이터 소스는 아래의 웹사이트를 방문하시고, 더 많은 정보를 얻도록 하세요.

Source: [https://www.kaggle.com/spscientist/students-performance-in-exams]

2) Data Manipulation

1. Data Preapreation

In [3]:
```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
plt.style.use('dark_background')
sns.set(style="darkgrid",palette="bright",font_scale=1.5)
df=pd.read_csv(r"C:\Users\DAVID SEO\Desktop\StudentsPerformance.csv")
df.head()
```

Out[3]:

	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

[ENG] The data is stored in the format of pandas for the convenience of data handling. With the use of DataFrame/Series.head() method, we can get a glimpse of the whole data: showing the first five entries of a input data.

[KOR] 데이터는 데이터 핸들링의 편의성을 의하여 pandas 형식으로 저장을 했습니다. head() 메소드를 통하여 우리는 데이터의 전체적인 측면을 볼 수가 있습니다. 이 메소드는 첫 5 개의 행을 반환합니다.

2. Checking The Missing Entries

[ENG] To check if there are missing entries emobided in the data-set, isnull().any() method is employed to detect them. However, the results return all False and we do not take extra measures to deal with them.

[KOR] 데이터의 내재된 결측값을 확인하기 위해서, 우리는 isnull().any() 메소드를 이용하여 찾아보았습니다. 이에 대한 결과로 모든 필드(field)에서 False값을 반환했기 때문에, 추가저인 조취는 필요해 보이지 않습니다.

In [4]:
```
df.isnull().any()
```

Out[4]:
```
gender                         False
race/ethnicity                 False
parental level of education    False
lunch                          False
test preparation course        False
math score                     False
reading score                  False
writing score                  False
dtype: bool
```

3.Filtering Out Irrevalent Infomration

[ENG]What we are interested in is, regardless of ethnicity or race, the influence of parent's education level on children's academic performace.Furthemore, the code code Furthermore, the code labels(ie 'A', 'B' and 'C') are not specifically designated names, leading to confusion and ambiguity to data analysis.Therefore we should take away the columns named "race/ethnicity

[KOR] 우리가 관심있는 것은 민족과 인종에 상관없이 부모의 학업수준이 아이들에게 미치는 영향입니다. 게다가, 코드 레벨값이 정확하게 주어지지 않아서, 오히려 데이터 분석에 혼란을 가중시킬 수 있습니다. 이러한 이후로 우리는 열 "race/enthnicity"를 배제하도록 합니다.

In [5]:
```
df=df.drop("race/ethnicity",axis=1)
df.head()
```

Out[5]:

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

4. Correlation [optional]

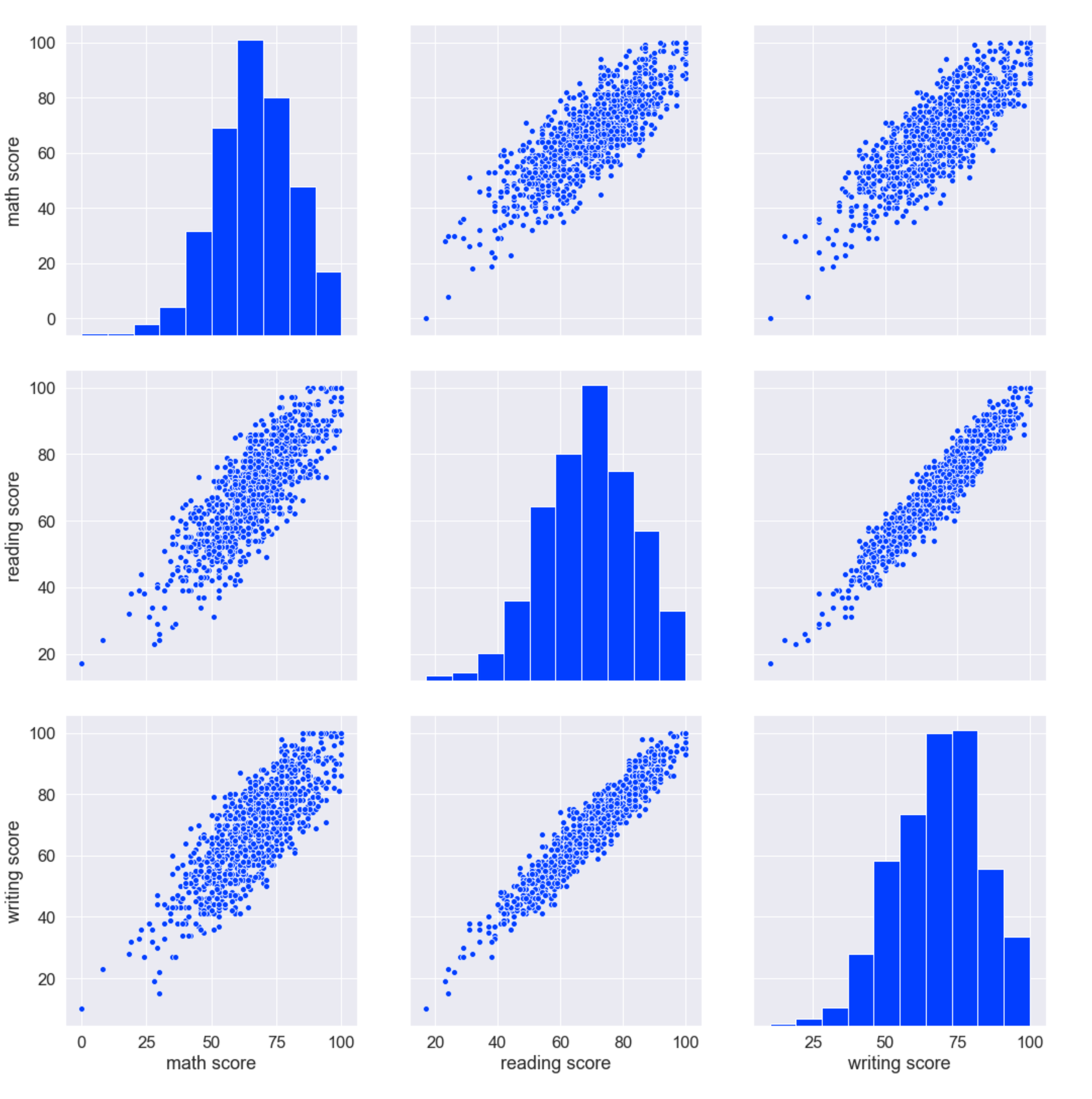
4.1 The Realtionship Among Subjects

[ENG] Let's find out any relationship existing among the three subjects. The best statistical measure for exploding it is a
correlationship, a measure of how two variables move together. Now, we have an ellipse as shown in the plot below. In
lossely term, the wider the curve is, the less apparent the relatinship seems to be. The oppsite is ture where the narrower the stronger it is.Based ont this, we could realize that two subjects (ie writing and reading) are closely related to one another.

[KOR] 세 가지 과목들 사이에서 관계가 존재하는지를 알아보도록 합시다. 이때, 가장 훌륭한 통계적 측정법은 바로 상관계수입니다. 두 변수의 관계등을 나타내는 측정법입니다.타원형 모양의 curve가 아래의 도표에 나타난 것을 볼 수가 있습니다. 만약 이 커브의 모양이 퍼지면 퍼질 수록 둘의 관계는 불명확하다는 것입니다. 반대로, 더 납작하면 납작할 수록 둘의 관계가 명확합니다. 이러한 특징을 인지할 때, writing과 reading 이 두과목이 상당히 밀접한 관계를 가지는 것을 인지할 수 있습니다.

In [6]:
```
sns.pairplot(df[['math score','reading score','writing score']],height=5)
```

Out[6]: <seaborn.axisgrid.PairGrid at 0x280b40421d0>



[ENG]Additionally,the field called "average score" is added to the existing data-set. I have implemented apply function, which applies the arbitary function to the data.

[KOR]추가적으로, "average score"의 명칭을 가진 열을 현재 데이터에 추가하였습니다. 이때,전 임의의 함수를 데이터에 적용하게 해주는 함수인 apply를 사용하였습니다.

In [7]:
```
def average_score(dt):
    return (dt["math score"]+dt["reading score"]+dt["writing score"])/3

df["average score"]=df.apply(average_score,axis=1)
df.head()
```

Out[7]:

	gender	parental level of education	lunch	test preparation course	math score	reading score	writing score	average score
0	female	bachelor's degree	standard	none	72	72	74	72.666667
1	female	some college	standard	completed	69	90	88	82.333333
2	female	master's degree	standard	none	90	95	93	92.666667
3	male	associate's degree	free/reduced	none	47	57	44	49.333333
4	male	some college	standard	none	76	78	75	76.333333

4.2 Parentel level of Education and School Meal Programs

[ENG]I have carefully examined a relationship between the educational level of parents and the support for school meals their children have.My initial expectation is that if the educational level of parents is higher. there should be a more chance of earning a great salary to support the family members.Taht is,the higher income could lead to their children being disqualified to have support or subsidy for the school meal. Let;s to find out if this is true

[KOR] 부모의 교육성취와 자식이 받는 무료급식 또는 감면혜택의 관계를 유심있게 살펴보았습니다. 저의 첫 견해는 만약 부모의 교육수준이 높다면 가족들을 부양하기 위한 돈을 많이 벌 수 있을 것이라고 생각했습니다. 즉,소득이 높다는 것은 그들의 자식들이 혜택을 받을 수 없다는 것을 의미하기도 합니다. 한번 이것이 사실인지 알아보도록 합시다.

In [10]:
```
#Extract the relevant fields from the table and
data=df.loc[:,("gender","parental level of education","lunch")]
final=data.groupby(["parental level of education","lunch"])["gender"].count().
unstack()

import copy
data["parental level of education"].unique()
index=['some high school','high school',"associate's degree",'some college',"b
achelor's degree","master's degree"]
# Rearrange the index of college degree from high school to master's degree
final=final.reindex(index)
#deepcopy the final
final2=copy.copy(final)
#add another coulmn "sum" along the colums
final2["sum"]=final.sum(axis=1)
final2
```

Out[10]:

lunch	free/reduced	standard	sum
parental level of education			
some high school	61	118	179
high school	70	126	196
associate's degree	77	145	222
some college	79	147	226
bachelor's degree	44	74	118
master's degree	24	35	59

In [12]:
```
final2["free/reduced (%)"]=round(final2["free/reduced"]/final2["sum"]*100,2)
final2["standard (%)"]=round(final2["standard"]/final2["sum"]*100,2)
final2
```

Out[12]:

lunch	free/reduced	standard	sum	free/reduced (%)	standard (%)
parental level of education					
some high school	61	118	179	34.08	65.92
high school	70	126	196	35.71	64.29
associate's degree	77	145	222	34.68	65.32
some college	79	147	226	34.96	65.04
bachelor's degree	44	74	118	37.29	62.71
master's degree	24	35	59	40.68	59.32

[ENG]Surprisingly,the proprtion of free/reduced has shown to be incresing as the parental level of education is higher. it seems to be necessary that further evidence is provided to demonstrate this movement.

[KOR]놀랍게도, 혜택비율은 부모의 교육수준이 높아지면 높아질 수록, 혜택의 비율은 높아지는 것을 알 수 있었습니다.우리는 이를 증명할 추가적인 자료가 필요해 보입니다.

5.Graphcial Representation

5.1 Benefit for School meal and Subjects

[ENG] I have employed a catplot,a function that shows the relationshiop between a numerical and one or more categorical variable. Looking at three outcomes, we could obtain two findings helpful for our analysis.

First, assuming that we have divided the group into subgroups (ie standard and 'free/reduced', within each subgroup, male students have shown relatively better performance on a quantitative subject while female students are a little superior in linguistic subjects.

Secondly, students either female or male who receive benefits and whose academic achievements are not very pleasing score much lower marks than those in the standard. Regardless of gender and subjects, the rage of the score for students in the group of 'free/reduced ' is much broader than for those in 'standard'. Also, their minimum score for each subject is much lower.

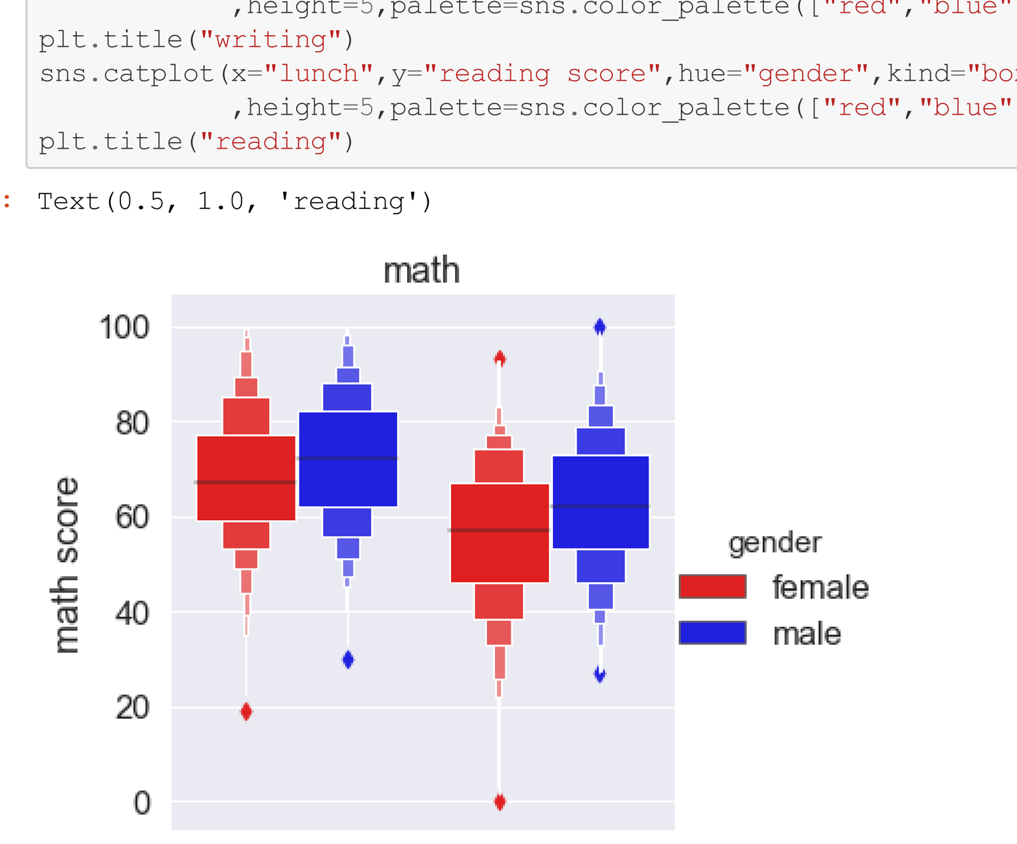
[KOR] 범주형자료와 숫자형 자료의 관계를 밝히는 분석함수인 catplot를 사용하여습니다. 그래프들을 잘 살펴보자면, 우리는 분석에 유용한 두 가지 사실들을 얻을 수 있습니다.

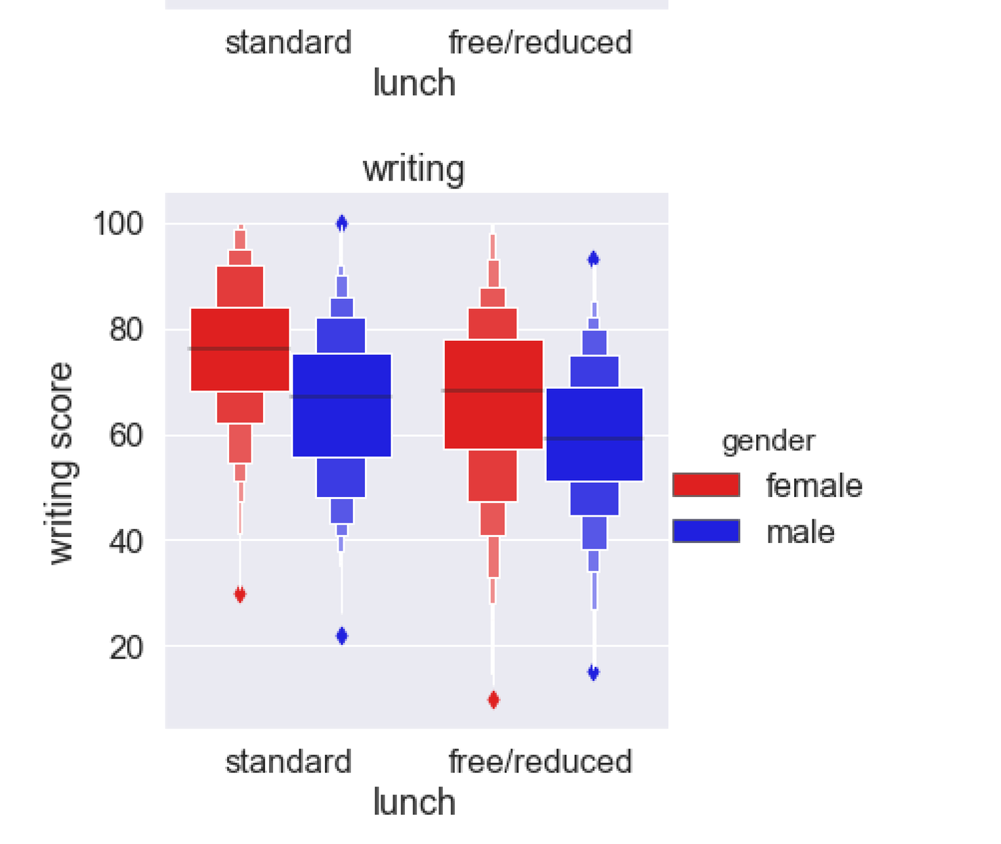
첫번째로,우리가 'standard'와 'free/reduced' 두 그룹으로 나누었을때, 각각의 그룹에서 계량과목에 대한 남학생의 성적 성취도가 여성보다 상대적으로 우수한것을 알 수 있었습니다.그에 반면에, 여성은 언어와 관련되 과목에서 우월성을 보여주었습니다.

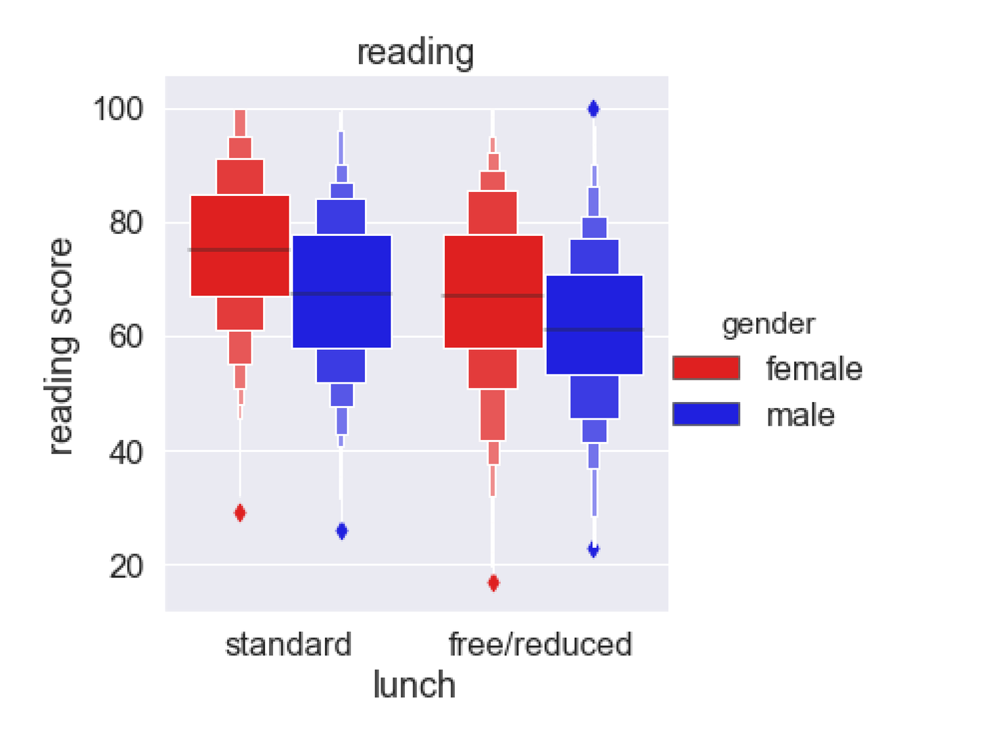
두번째로, 여학생이든지 남학생이든지 혜택을 받는 그룹에 속하며 학업성취도가 좋지 못한 학생들은 'standard'그룹에 속한 학생들 보다 훨씬 낮은 성적을 받았습니다. 성과 그들이 택한 과목에 상관없이, 이 학생들의 성적 범위(최댓값과 최솟값의 차이)가 상당히 크다는 것을 보여줍니다.

In [40]:
```
#Next, how the support would give an impact on each subject
sns.catplot(x="lunch",y="math score",hue="gender",kind="boxen",data=df
            ,height=5,palette=sns.color_palette(["red","blue"]))
plt.title("math")
sns.catplot(x="lunch",y="writing score",hue="gender",kind="boxen",data=df
            ,height=5,palette=sns.color_palette(["red","blue"]))
plt.title("writing")
sns.catplot(x="lunch",y="reading score",hue="gender",kind="boxen",data=df
            ,height=5,palette=sns.color_palette(["red","blue"]))
plt.title("reading")
```

Out[40]: Text(0.5, 1.0, 'reading')







5.2 The Impact of Parental Education On Children Academic Performance

[ENG] The examination of the boxplots gives us some decisive hints to conclude. The merits of the use of boxplots enable us to determine to some extent the shapes or distribution of a dataset. Every group has a symmetric distribution with a median nearly equal. Although a group of students whose parents hold advanced degrees tends to score a higher median, the difference is too small to conclude that the education background of parents is playing a determinant in deciding their children's academic performance.

Lasetly, The extreme vaues obtained by some groups would be examined to make sure that they are valid measurements. They could be either by chance or be critical to reaching our conclusion

[KOR] boxplot의 설명은 우리들에게 결정적인 힌트를 제공해줍니다. boxplot을 사용하는 큰 장점은 데이터의 분포를 어느정도 가늠하게 해준다는 것입니다. 각각의 그룹들은 어느 쪽으로 치우치는 현상을 보여주지 않기 때문에 대칭적이다라고 말할 수 있습니다. 또한, 물론 master group 이 가장 큰 수치를 나타냈지만,차이가 그리 크지않으므로 부모의 학력이 아이들의 학업성취도에 중요하게 기여한다고 결론을 내리기에는 무리가 있습니다.마지막으로 몇 그룹들에게 보여주는 극단적인 값들들을 유심히 살펴볼 필요가 있습니다. 이것은 순한 우연일 수도 있으며 결론을 내리는데 있어서 중요할 수 있습니다.

In [45]:
```
plt.figure(figsize=(16,6)) # setting seaborn plot size
sns.boxplot(x="parental level of education",y="average score",data=df,linewidt
h=1.5)
print("average")
```

```
average
```