Correlation Analysis

(부모의 학력과 자녀의 시험점수 상관 관계)

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
plt.style.use('dark_background')
sns.set(style="darkgrid", palette = "bright", font_scale=1.5)
```

```
df = pd.read_csv("../Desktop/data_science practice/correlation analysis/StudentsPerformance.csv")
df.head()
```

```
.dataframe tbody tr th {
    vertical-align: top;
}

.dataframe thead th {
    text-align: right;
}
```

	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

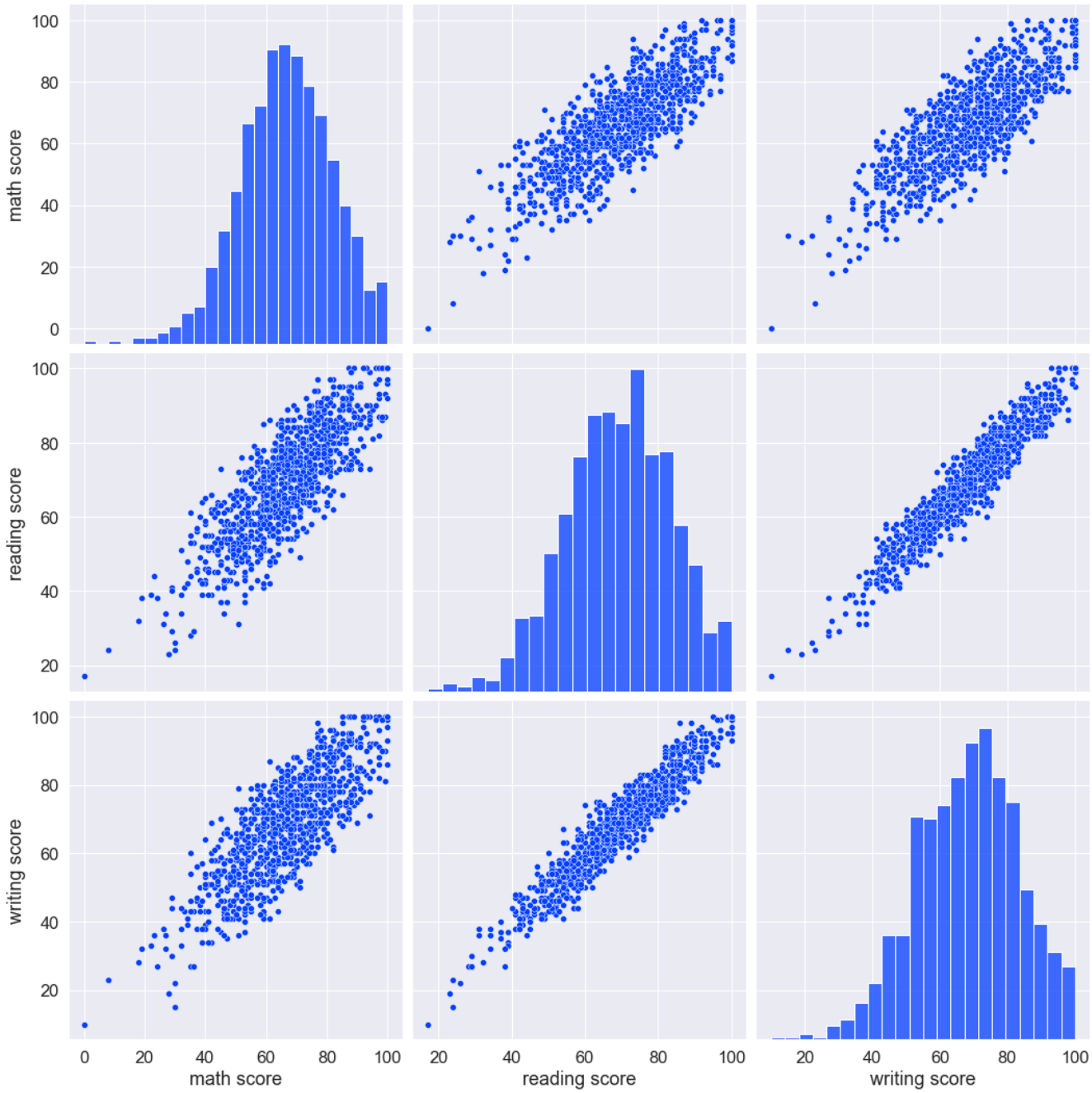
```
df.describe()
```

```
.dataframe tbody tr th {
    vertical-align: top;
}

.dataframe thead th {
    text-align: right;
}
```

	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

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

```
<seaborn.axisgrid.PairGrid at 0x19f25ab2370>
```



점 그래프의 분포가 얇을 수록 상관관계가 있다고 할 수 있다.

-> reading score 와 writing score 간의 상관관계가 있다고 할 수 있다.

학생의 평균 점수 구하기

```
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.apply() : 데이터 프레임에 함수를 일괄적으로 적용한다.
# axis=1 : 1번 축에 추가
```

```
df.head()
```

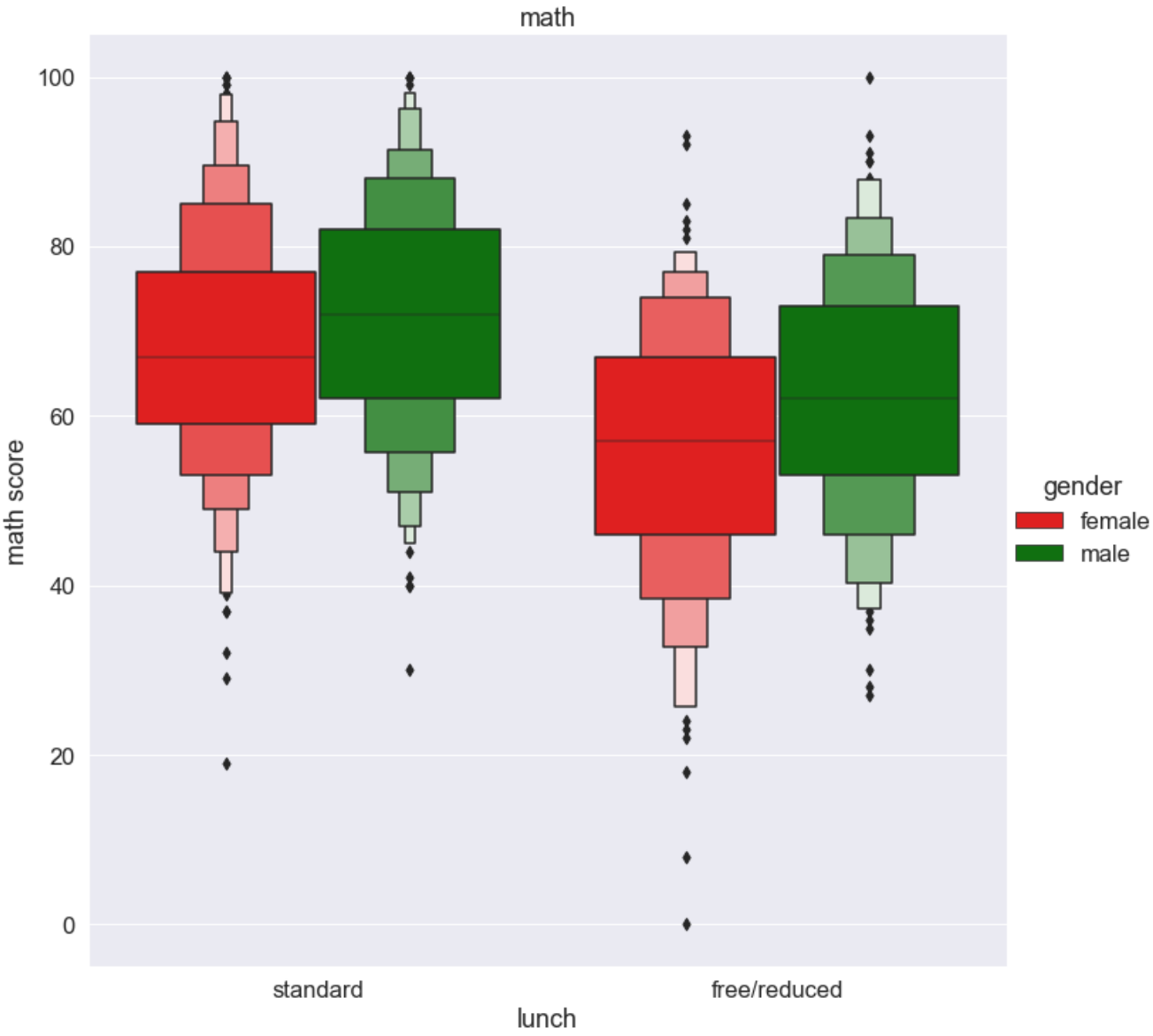
```
.dataframe tbody tr th {
    vertical-align: top;
}

.dataframe thead th {
    text-align: right;
}
```

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

점심 식사 여부와 시험 점수와의 상관관계

```
#catplot : catagorical plot
#kind = 'boxen' : 대용량 데이터에 적합한 형태의 그래프
sns.catplot (x = 'lunch', y = 'math score', hue='gender', kind='boxen', data=df, height=10,
            palette = sns.color_palette(['red','green']))
plt.title('math')
```

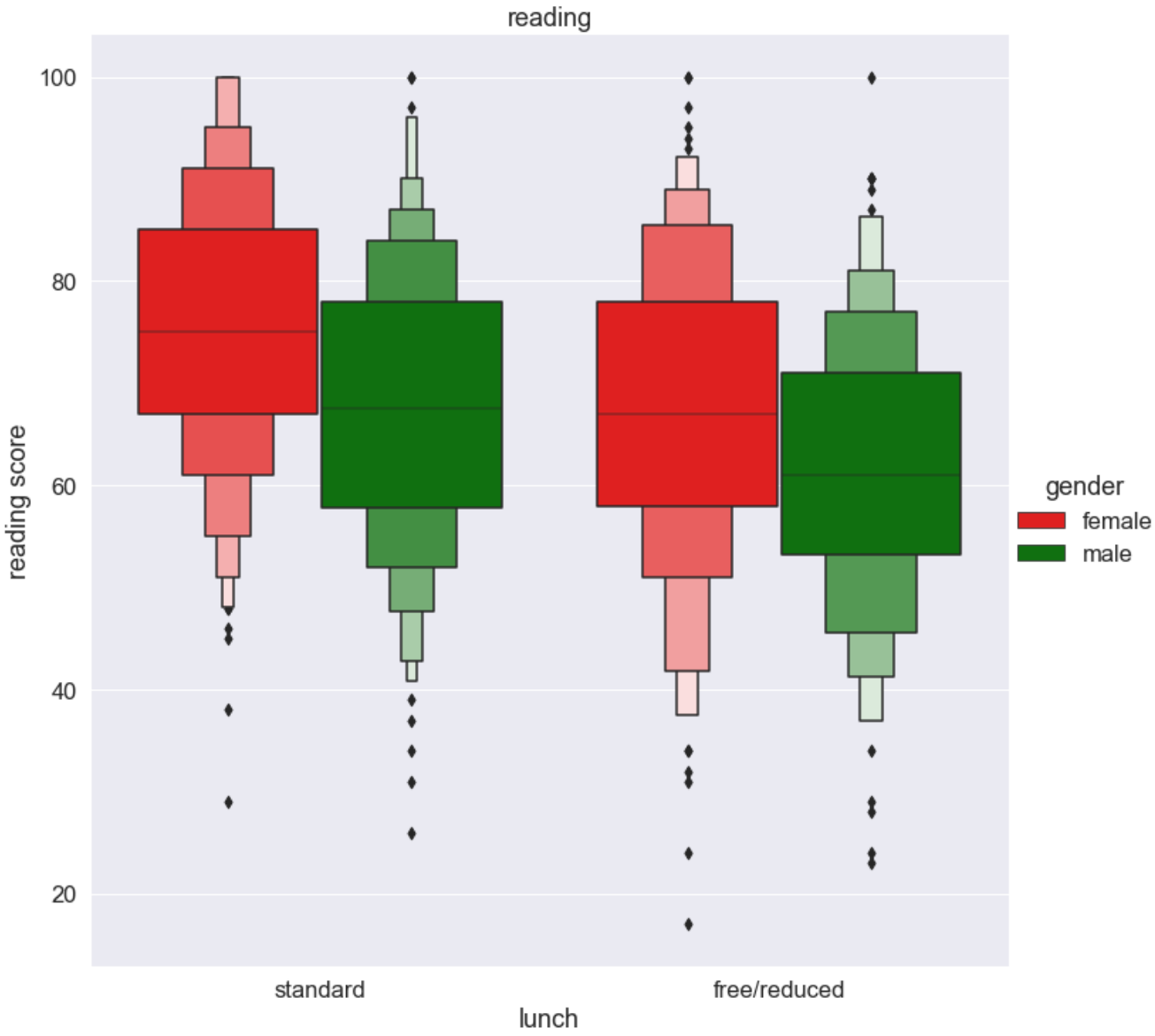
```
Text(0.5, 1.0, 'math')
```



- 남학생의 수학점수 평균이 여학생 수학점수 평균보다 높다.
- 점심을 먹은 학생들의 수학점수가 점심을 안먹은 학생보다 높다.

```
sns.catplot (x = 'lunch', y = 'reading score', hue='gender', kind='boxen', data=df, height=10,
             palette = sns.color_palette(['red','green']))
plt.title('reading')
```

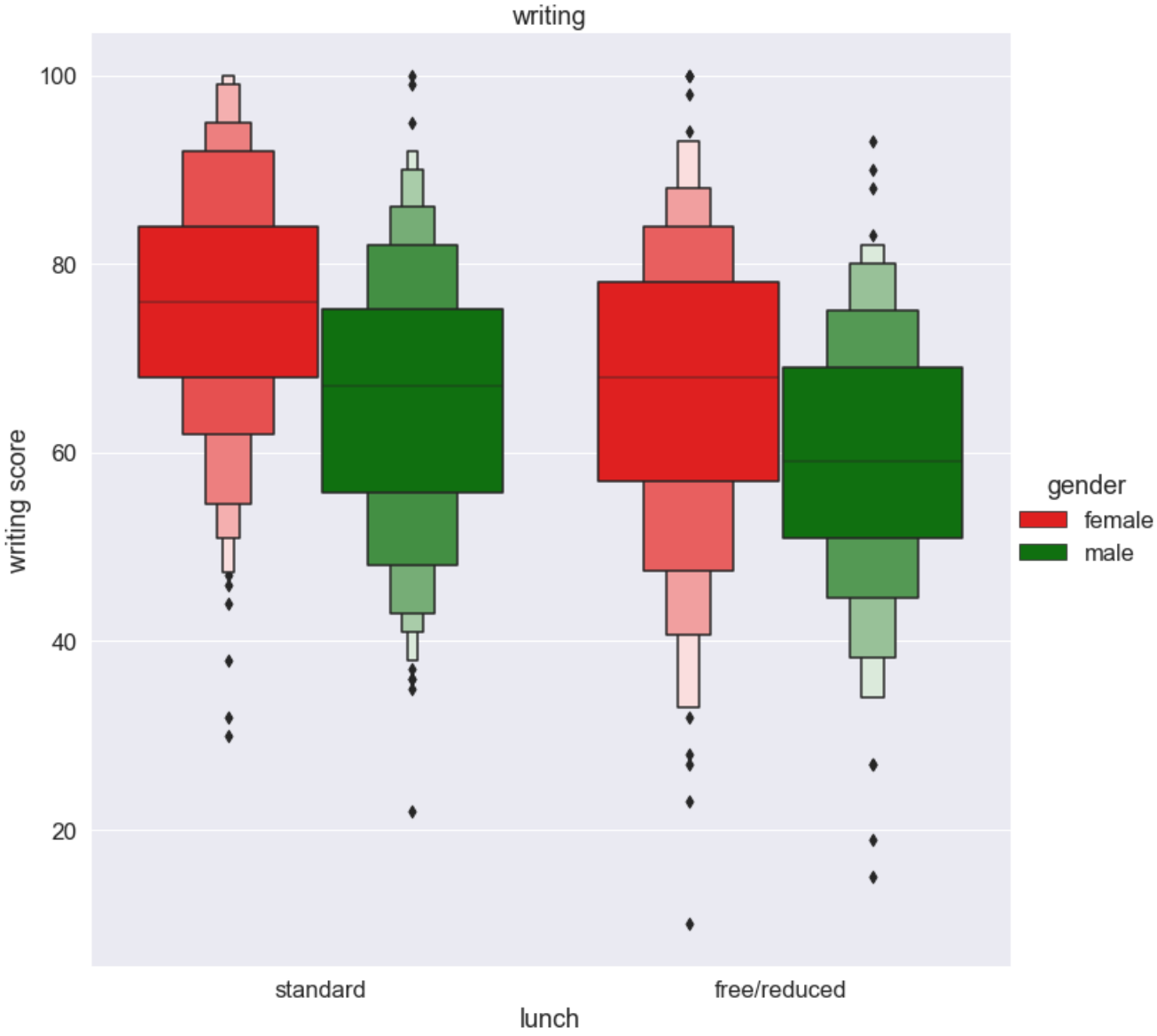
```
Text(0.5, 1.0, 'reading')
```



- 여학생의 수학점수 평균이 여학생 수학점수 평균보다 높다.
- 점심을 먹은 학생들의 수학점수가 점심을 안먹은 학생보다 높다.

```
sns.catplot (x = 'lunch', y = 'writing score', hue='gender', kind='boxen', data=df, height=10,
            palette = sns.color_palette(['red','green']))
plt.title('writing')
```

```
Text(0.5, 1.0, 'writing')
```

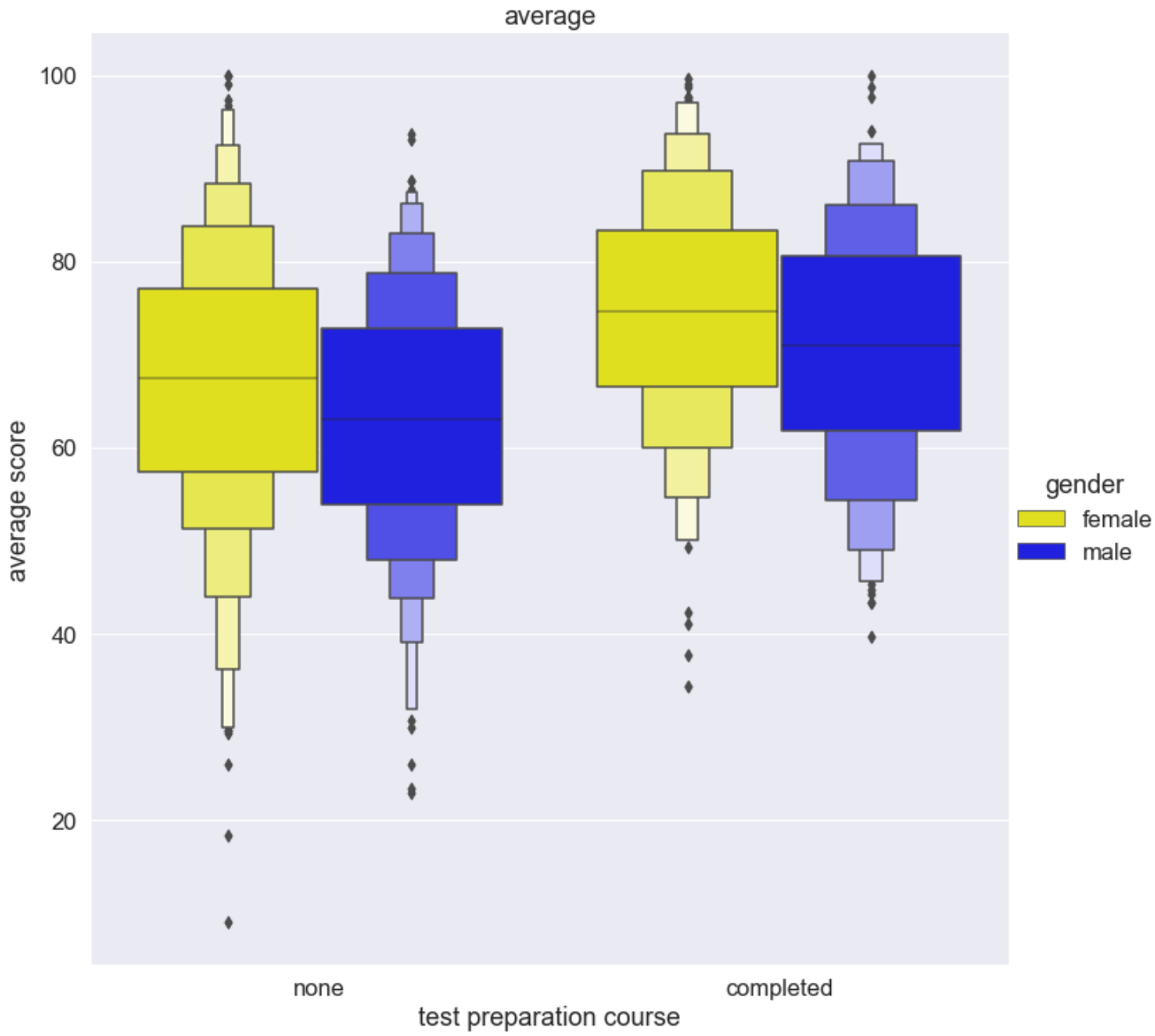


- 여학생의 수학점수 평균이 여학생 수학점수 평균보다 높다.
- 점심을 먹은 학생들의 수학점수가 점심을 안먹은 학생보다 높다.

Test Preparation Course와 평균 점수의 상관 관계

```
sns.catplot(x='test preparation course', y='average score', hue='gender', kind='boxen', data=df, height=10,
palette=sns.color_palette(['yellow', 'blue']))
plt.title('average')
```

```
Text(0.5, 1.0, 'average')
```



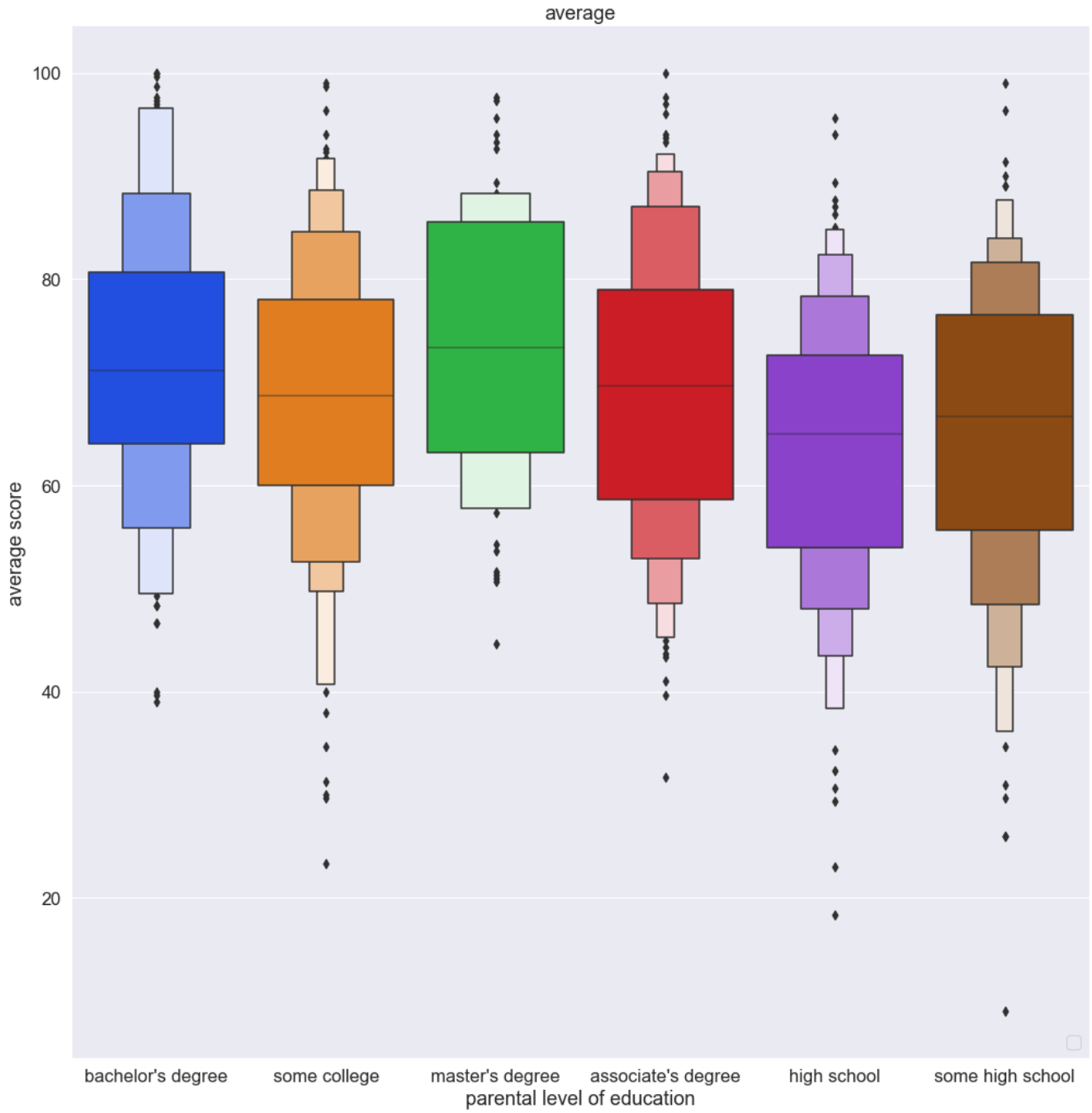
- 여학생의 평균점수가 남학생 평균점수보다 높다.
- test preparation course를 들은 학생의 평균점수가 더 높다.

부모의 학력과 평균 점수의 상관 관계

```
sns.catplot(x='parental level of education', y='average score', kind='boxen', data=df, height=14)
plt.title('average')
plt.legend(loc='lower right')
```

```
No handles with labels found to put in legend.
```

```
<matplotlib.legend.Legend at 0x19f28d6cee0>
```



- 석사학위를 가진 부모의 자녀의 평균이 가장 높다.
- 고등학교 학위를 가진 부모의 자녀의 점수가 가장 밑의 성적을 가지고 있다.
- 상관관계가 전체적으로 크지는 않다.