```python
import pandas as pd

student = pd.read_csv("Student_Performance.csv")

student.head(10)
```

```
   student_id  age  gender school_type parent_education  
study_hours  \
0           1   14    male      public    post graduate          3.1
1           2   18  female      public         graduate          3.7
2           3   17  female     private    post graduate          7.9
3           4   16   other      public      high school          1.1
4           5   16  female      public      high school          1.3
5           6   19    male      public        no formal          3.8
6           7   14  female     private    post graduate          1.8
7           8   18  female     private    post graduate          5.6
8           9   15   other     private      high school          3.2
9          10   14  female      public          diploma          6.8

   attendance_percentage internet_access travel_time extra_activities  
\
0                   84.3             yes     <15 min              yes
1                   87.8             yes     >60 min               no
2                   65.5              no     <15 min               no
3                   58.1              no   15-30 min               no
4                   61.0             yes   30-60 min              yes
5                   69.6             yes     >60 min              yes
6                   81.6             yes   30-60 min               no
7                   59.4             yes     >60 min              yes
8                   89.6             yes   15-30 min              yes
9                   62.4             yes     >60 min               no
```

```
  study_method  math_score  science_score  english_score
overall_score  \
0        notes        42.7           55.4           57.0
53.1
1     textbook        57.6           68.8           64.8
61.3
2        notes        84.8           95.0           79.2
89.6
3        notes        44.4           27.5           54.7
41.6
4  group study         8.9           32.7           30.0
25.4
5     coaching        51.5           78.3           63.9
63.5
6     textbook        41.9           29.4           39.2
39.1
7  group study        56.7           60.1           53.4
69.6
8        mixed        54.1           59.5           38.3
55.2
9        mixed        71.9           70.4           81.3
69.6

  final_grade
0           e
1           d
2           b
3           e
4           f
5           d
6           f
7           d
8           d
9           d
```

```
# checking the data

student.shape
```

```
(25000, 16)
```

```
student.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 25000 entries, 0 to 24999
Data columns (total 16 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   student_id              25000 non-null  int64
 1   age                     25000 non-null  int64
 2   gender                 25000 non-null  object 
 3   school_type            25000 non-null  object 
 4   parent_education       25000 non-null  object 
 5   study_hours            25000 non-null  float64
 6   attendance_percentage  25000 non-null  float64
 7   internet_access        25000 non-null  object 
 8   travel_time            25000 non-null  object 
 9   extra_activities       25000 non-null  object 
 10  study_method           25000 non-null  object 
 11  math_score             25000 non-null  float64
 12  science_score          25000 non-null  float64
 13  english_score          25000 non-null  float64
 14  overall_score          25000 non-null  float64
 15  final_grade            25000 non-null  object 
dtypes: float64(6), int64(2), object(8)
memory usage: 3.1+ MB
```

```
student.isnull().sum()
```

```
student_id               0
age                      0
gender                   0
school_type              0
parent_education         0
study_hours              0
attendance_percentage    0
internet_access          0
travel_time              0
extra_activities         0
study_method             0
math_score               0
science_score            0
english_score            0
overall_score            0
final_grade              0
dtype: int64
```

```
student.describe()
```

```
        student_id           age   study_hours  attendance_percentage 
\
count  25000.00000  25000.000000  25000.000000           25000.000000

mean    7493.04380     16.482760      4.253224              75.084084

std     4323.56215      1.703895      2.167541              14.373171

min        1.00000     14.000000      0.500000              50.000000

25%     3743.75000     15.000000      2.400000              62.800000
```

```
50%     7461.50000     16.000000      4.300000               75.100000

75%    11252.00000     18.000000      6.100000               87.500000

max    15000.00000     19.000000      8.000000              100.000000
```

```
         math_score  science_score  english_score  overall_score
count  25000.000000   25000.000000   25000.000000   25000.000000
mean      63.785944      63.745320      63.681948      64.006172
std       20.875262      20.970529      20.792693      18.932025
min        0.000000       0.000000       0.000000      14.500000
25%       48.300000      48.200000      48.300000      49.000000
50%       64.100000      64.100000      64.200000      64.200000
75%       80.000000      80.000000      80.000000      79.000000
max      100.000000     100.000000     100.000000     100.000000
```

```
(student['gender'] == 'other').mean()*100
```

```
np.float64(33.852)
```

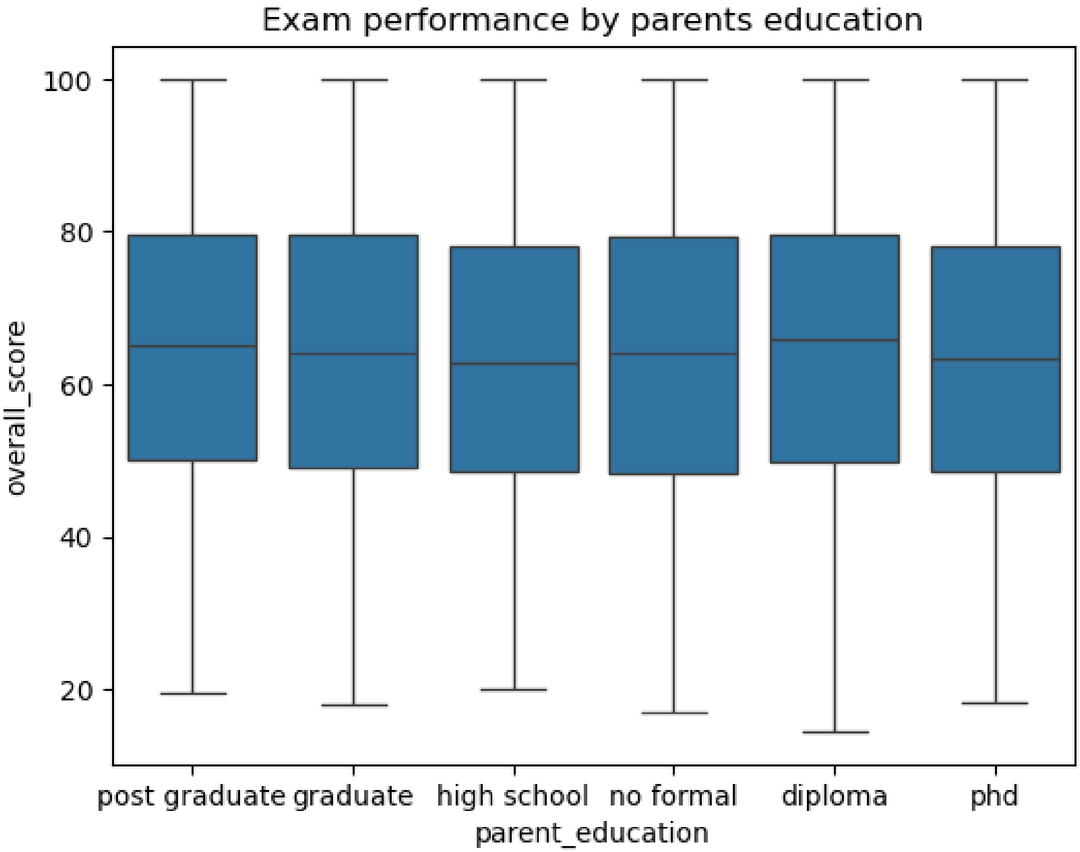
```
# since the data is pretty clean, I'm going to focus on answering my
questions
# doing this in python to develop my skills.

# Q1) Is there an association between parents’ education level and
student exam performance?

import matplotlib.pyplot as plt
import seaborn as sns

student['parent_education'].unique()
```

```
array(['post graduate', 'graduate', 'high school', 'no formal',
'diploma',
       'phd'], dtype=object)
```

```
student.groupby('parent_education')['overall_score'].agg(
    mean='mean',
    median='median',
    count='count'
)
```

```
                       mean  median  count
parent_education
diploma           64.651484    65.8   4314
graduate          63.956991    64.1   4127
high school       63.386492    62.8   4205
no formal         63.907085    64.1   4079
phd               63.538637    63.4   4079
post graduate     64.562917    65.1   4196
```

```
sns.boxplot (
    data=student,
    x='parent_education',
    y='overall_score'
)
plt.title('Exam performance by parents education')
plt.xlabel('parent_education')
plt.ylabel('overall_score')
plt.show()
```

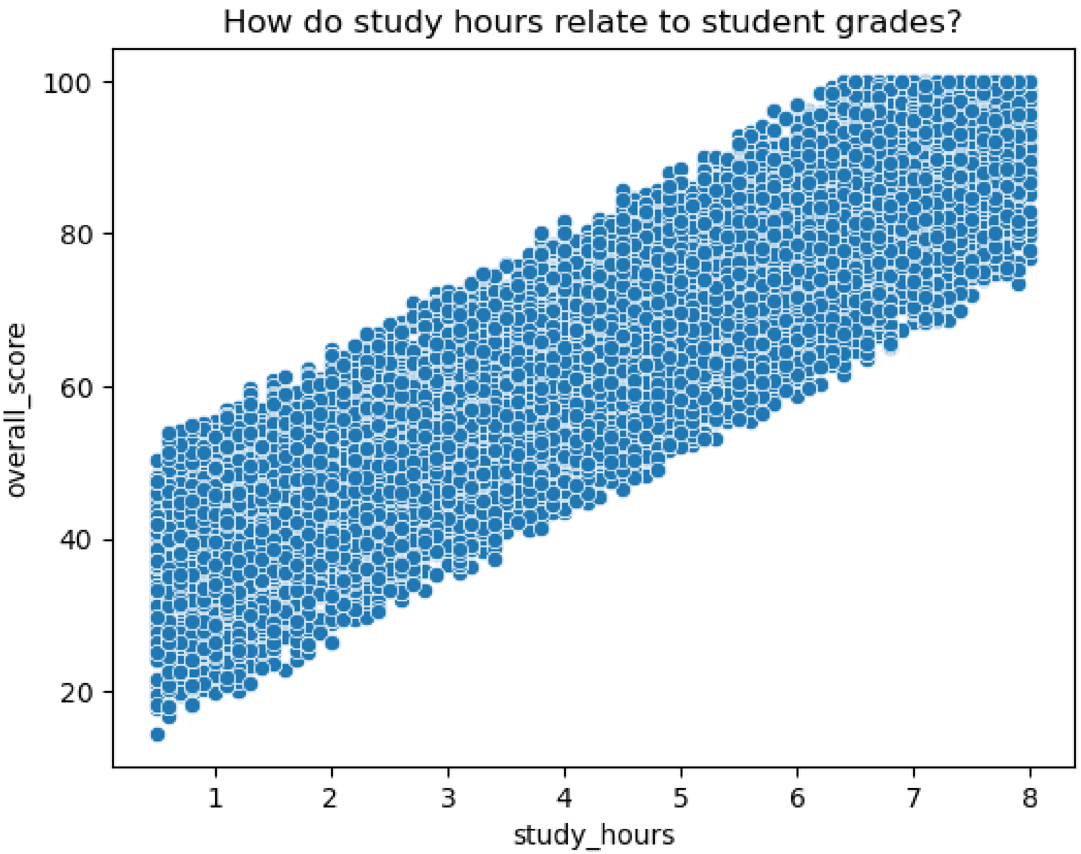


```
# Q2) How do study hours relate to student grades?

student.groupby('study_hours')['overall_score'].agg(
    mean='mean',
    median='median',
    count='count'
)
```

```
                  mean  median  count
study_hours
0.5          34.649405    36.2    168
0.6          35.029595    34.7    321
0.7          36.224646    36.2    353
```

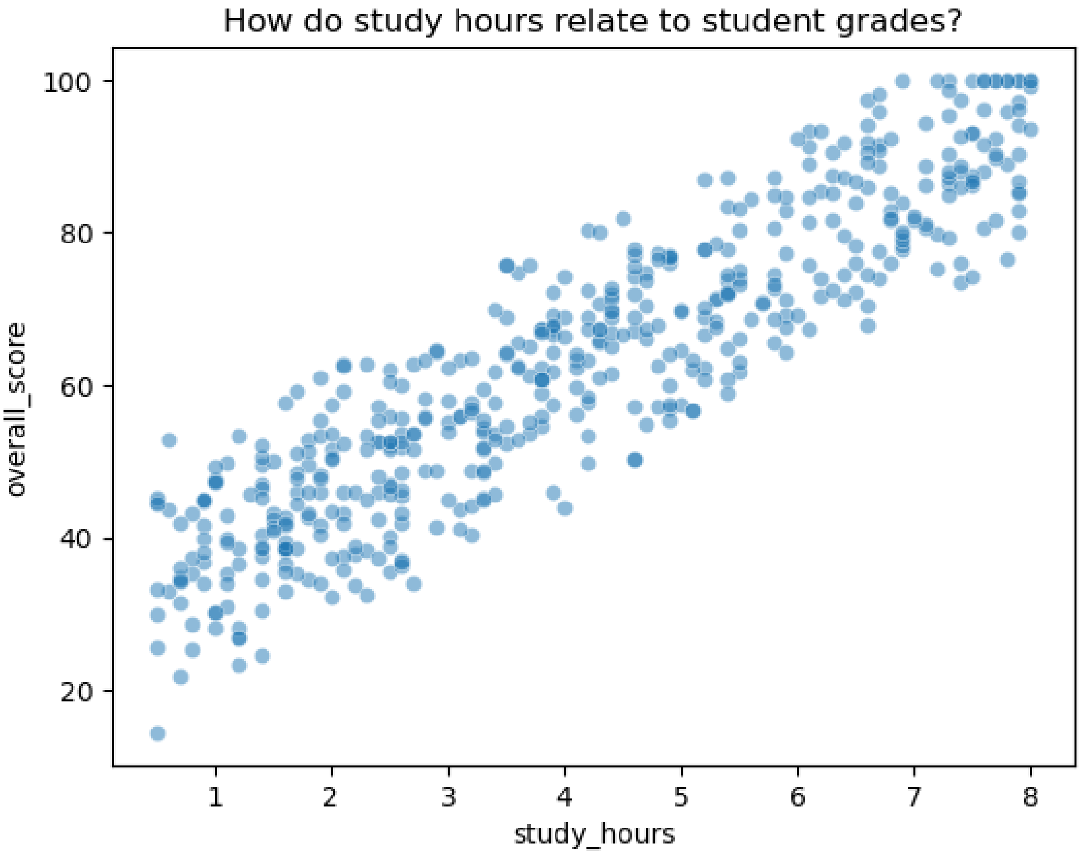
```
0.8          36.806885    38.4    305
0.9          37.490592    37.5    287
...                ...     ...    ...
7.6          91.012104    91.6    347
7.7          91.032877    91.9    365
7.8          91.748243    91.7    313
7.9          92.378736    93.3    348
8.0          93.521875    95.4    160

[76 rows x 3 columns]
```

```
sns.scatterplot(
    data = student,
    x='study_hours',
    y= 'overall_score'
)
plt.title('How do study hours relate to student grades?')
plt.xlabel('study_hours')
plt.ylabel('overall_score')
plt.show()
```



```
# since the visualisation shows a clear pattern
# however due to it looking messy i'll do a random sampling
```

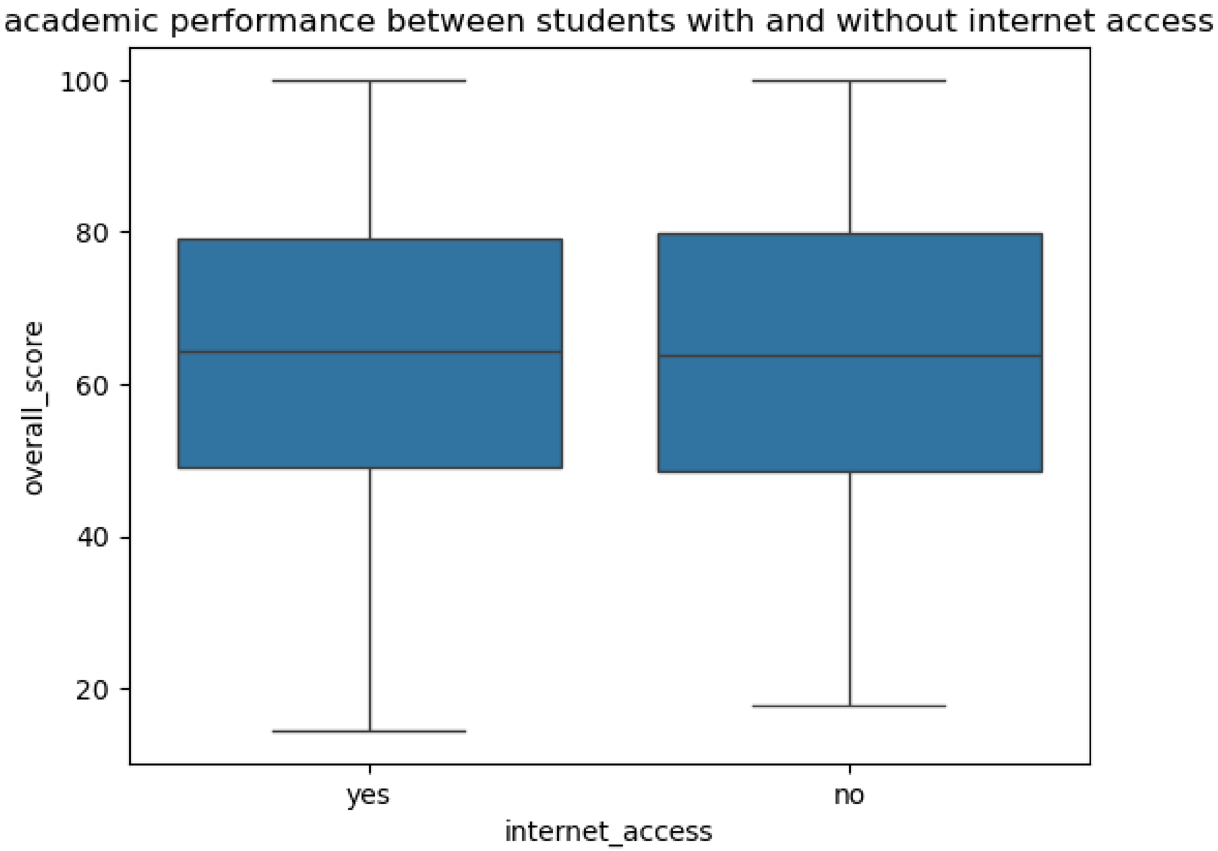
```
random_sample = student.sample(frac=0.02,random_state=1)

sns.scatterplot(
    data = random_sample,
    x='study_hours',
    y= 'overall_score',
    alpha = .5
)
plt.title('How do study hours relate to student grades?')
plt.xlabel('study_hours')
plt.ylabel('overall_score')
plt.show()
```



```
# Q3) Are there observable differences in academic performance between
students with and without internet access?

student.groupby('internet_access')['overall_score'].agg(
    mean='mean',
    median='median',
    count='count'
)
```

```
                      mean  median  count
internet_access
no               63.774821    63.9   3773
yes              64.047294    64.3  21227
```

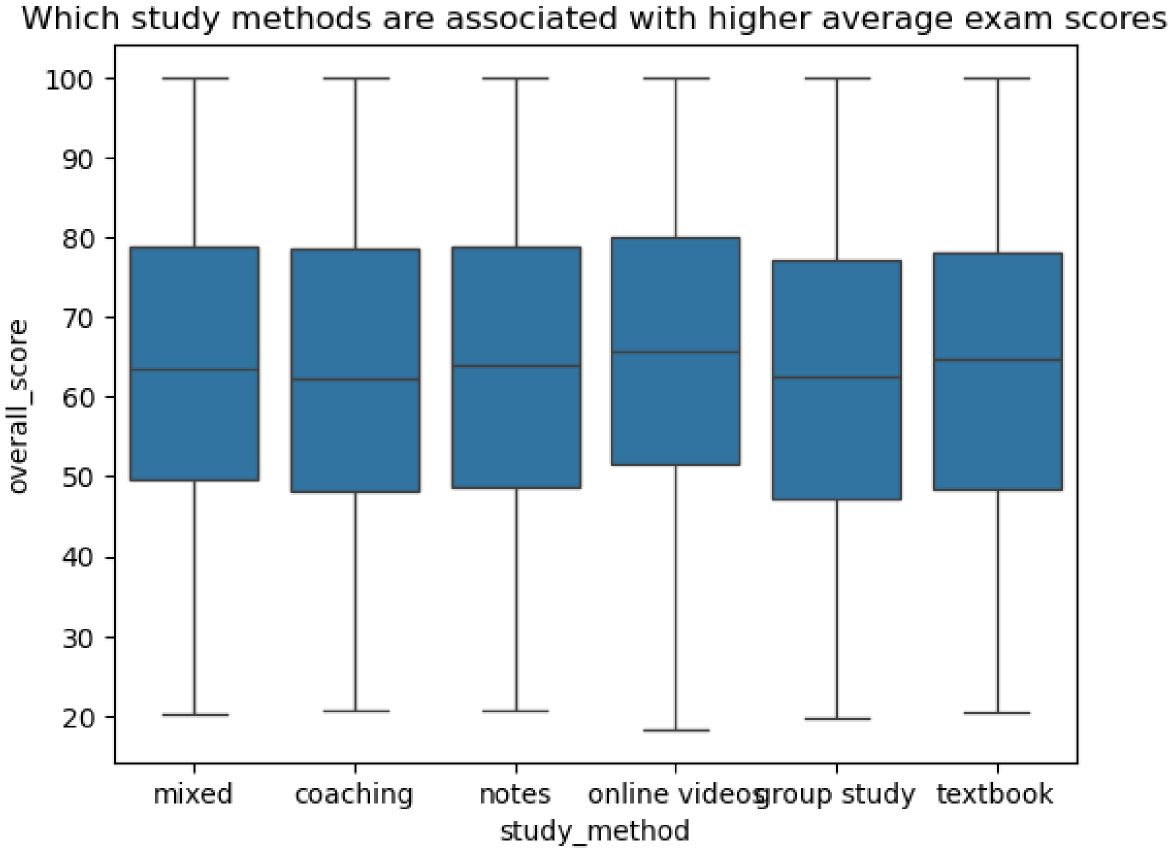
```
sns.boxplot(
    data=student,
    x='internet_access',
    y='overall_score'

)
plt.title('academic performance between students with and without
internet access')
plt.xlabel('internet_access')
plt.ylabel('overall_score')
plt.show()
```



```
# Q4) Which study methods are associated with higher average exam
scores?

student.groupby('study_method')['overall_score'].agg(
    mean='mean',
    median='median',
    count='count'
)
```

```
                    mean  median  count
study_method
coaching       64.368405    64.6   4026
group study    63.249487    62.8   4090
mixed          63.613960    63.4   4341
notes          63.895318    64.4   4165
online videos  64.686905    65.2   4139
textbook       64.238122    65.0   4239
```

```
sample_study = student.sample(frac=0.3, random_state=2)

sns.boxplot(
    data = sample_study,
    x='study_method',
    y='overall_score'
)
plt.title('Which study methods are associated with higher average exam scores')
plt.xlabel('study_method')
plt.ylabel('overall_score')
plt.show()
```


Which study methods are associated with higher average exam scores
overall_score
100
90
80
70
60
50
40
30
20
mixed
coaching
notes
online videos
group study
textbook
study_method