CHALLENGE

Create a report to answer the principal's questions. Include:

1. What are the average reading scores for students with/without the test preparation course?
2. What are the average scores for the different parental education levels?
3. Create plots to visualize findings for questions 1 and 2.
4. [Optional] Look at the effects within subgroups. Compare the average scores for students with/without the test preparation course for different parental education levels (e.g., faceted plots).
5. [Optional 2] The principal wants to know if kids who perform well on one subject also score well on the others. Look at the correlations between scores.
6. Summarize your findings.

In [1]:
```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
import warnings
warnings.simplefilter(action = 'ignore', category = FutureWarning)
```

In [2]:
```
exam = pd.read_csv("exams.csv")
```

In [3]:
```
exam
```

Out[3]:

	gender	race/ethnicity	parent_education_level	lunch	test_prep_course	math	reading
0	female	group B	bachelor's degree	standard	none	72	72
1	female	group C	some college	standard	completed	69	90
2	female	group B	master's degree	standard	none	90	95
3	male	group A	associate's degree	free/reduced	none	47	57
4	male	group C	some college	standard	none	76	78
...	...	...	...	...	...	...	...
995	female	group E	master's degree	standard	completed	88	99
996	male	group C	high school	free/reduced	none	62	55
997	female	group C	high school	free/reduced	completed	59	71
998	female	group D	some college	standard	completed	68	78
999	female	group D	some college	free/reduced	none	77	86

1000 rows × 8 columns

```
In [4]: exam.head(10)
```

```
Out[4]:
   gender  race/ethnicity  parent_education_level         lunch  test_prep_course  math  reading  wr
0  female         group B       bachelor's degree      standard              none    72       72
1  female         group C            some college      standard         completed    69       90
2  female         group B         master's degree      standard              none    90       95
3    male         group A      associate's degree  free/reduced              none    47       57
4    male         group C            some college      standard              none    76       78
5  female         group B      associate's degree      standard              none    71       83
6  female         group B            some college      standard         completed    88       95
7    male         group B            some college  free/reduced              none    40       43
8    male         group D             high school  free/reduced         completed    64       64
9  female         group B             high school  free/reduced              none    38       60
```

```
In [5]: exam.tail(10)
```

```
Out[5]:
     gender  race/ethnicity  parent_education_level         lunch  test_prep_course  math  reading
990    male         group E             high school  free/reduced         completed    86       81
991  female         group B        some high school      standard         completed    65       82
992  female         group D      associate's degree  free/reduced              none    55       76
993  female         group D       bachelor's degree  free/reduced              none    62       72
994    male         group A             high school      standard              none    63       63
995  female         group E         master's degree      standard         completed    88       99
996    male         group C             high school  free/reduced              none    62       55
997  female         group C             high school  free/reduced         completed    59       71
998  female         group D            some college      standard         completed    68       78
999  female         group D            some college  free/reduced              none    77       86
```

In [6]:
```
exam.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 8 columns):
 #   Column                Non-Null Count  Dtype 
---  ------                --------------  ----- 
 0   gender                1000 non-null   object
 1   race/ethnicity        1000 non-null   object
 2   parent_education_level  1000 non-null   object
 3   lunch                 1000 non-null   object
 4   test_prep_course      1000 non-null   object
 5   math                  1000 non-null   int64 
 6   reading               1000 non-null   int64 
 7   writing               1000 non-null   int64 
dtypes: int64(3), object(5)
memory usage: 62.6+ KB
```

In [7]:
```
exam.isnull().sum()
```

Out[7]:
```
gender                    0
race/ethnicity            0
parent_education_level    0
lunch                     0
test_prep_course          0
math                      0
reading                   0
writing                   0
dtype: int64
```

Question 1

- What are the average reading scores for students with/without the test preparation course?

In [8]:
```
# average reading scores for students with/without the test preparation course
exam.reading.mean()
```

Out[8]:
```
69.169
```

In [9]:
```
#1
exam.groupby("test_prep_course")[["reading"]].mean()
```

Out[9]:

	reading
test_prep_course	
completed	73.893855
none	66.534268

Question3a

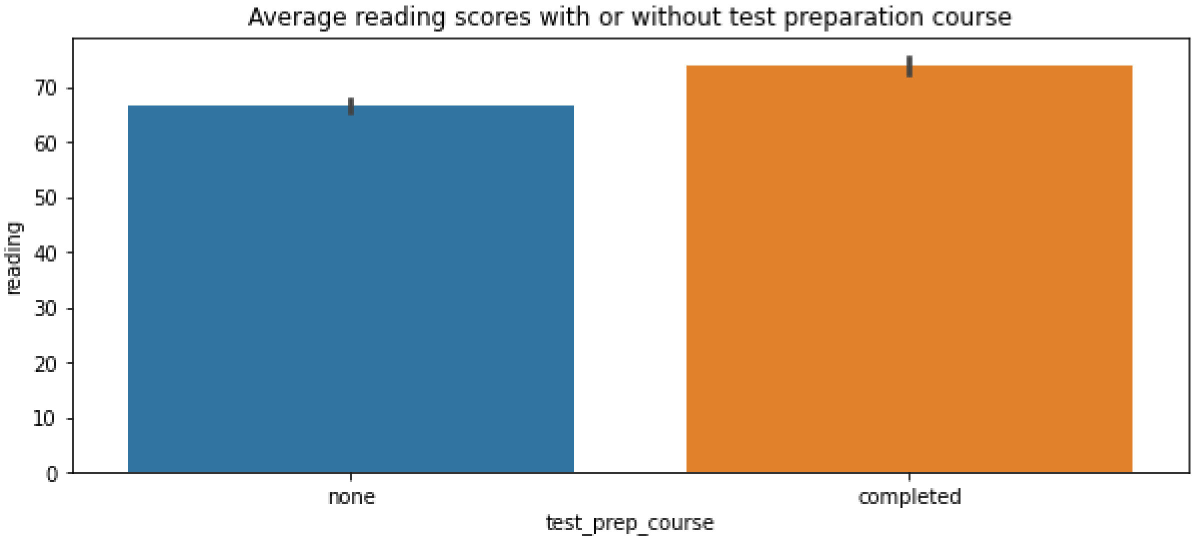
Create plots to visualize findings for questions 1

In [10]:

```
#3a
#graph
plt.figure(figsize = (10,4))
sns.barplot(x="test_prep_course", y="reading", data= exam)
plt.title('Average reading scores with or without test preparation course')
```

Out[10]:

```
Text(0.5, 1.0, 'Average reading scores with or without test preparation cours
e')
```



Question 2

- What are the average scores for the different parental education levels?

In [11]:

```
#2
# average scores for the different parental education levels
p=exam.groupby("parent_education_level").mean()
```

In [12]:

```
p
```

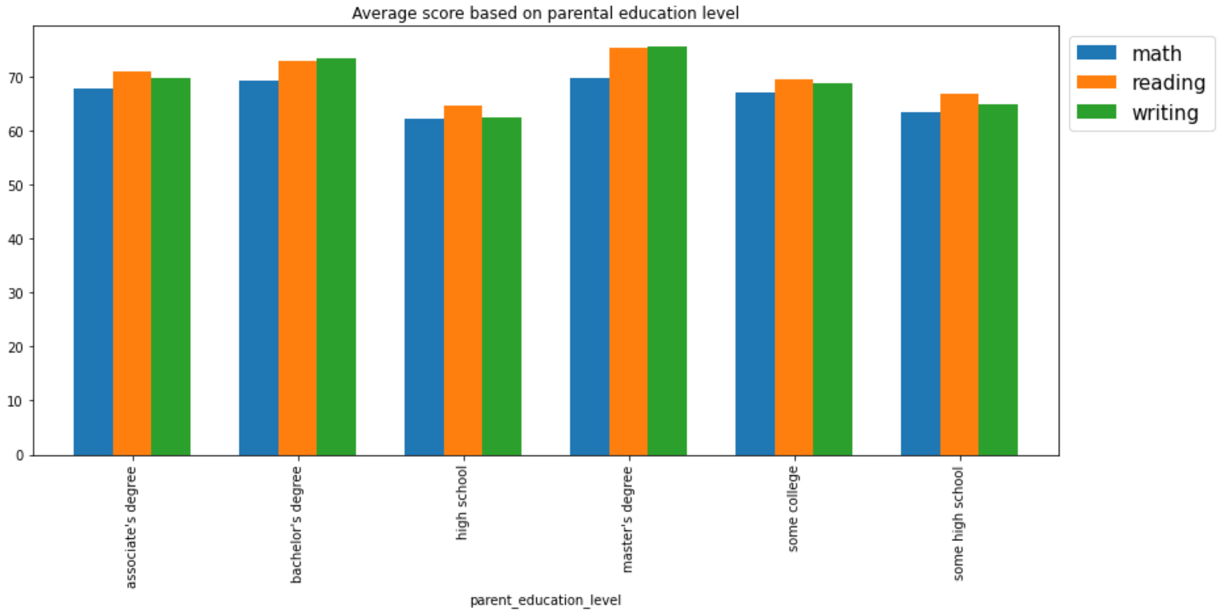
Out[12]:

	math	reading	writing
parent_education_level			
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

Question3b

Create plots to visualize findings for questions 2

In [13]:
```
p.plot(kind = 'bar', width = 0.7, figsize = (14,6) )
plt.legend(bbox_to_anchor = (1,1), fontsize = 15)
plt.title("Average score based on parental education level")
plt.show()
```



Question4

- Look at the effects within subgroups.
- Compare the average scores for students with/without the test preparation course for different parental education levels (e.g., faceted plots).

In [14]:
```
av_score = pd.pivot_table(exam, values="math",
                          index="parent_education_level",
                          columns="test_prep_course",
                          aggfunc=np.mean)
```

In [15]:

```
av_score
```

Out[15]:

test_prep_course	completed	none
parent_education_level		
associate's degree	71.829268	65.571429
bachelor's degree	73.282609	66.902778
high school	65.000000	60.992857
master's degree	70.600000	69.307692
some college	71.454545	64.892617
some high school	66.701299	61.078431

In [16]:

```
av_score2 = pd.pivot_table(exam, values=["math", "reading", "writing"],
                           index="parent_education_level",
                           columns="test_prep_course",
                          aggfunc=np.mean)
av_score2
```

Out[16]:

	math		reading		writing	
test_prep_course	completed	none	completed	none	completed	none
parent_education_level						
associate's degree	71.829268	65.571429	76.170732	67.857143	76.817073	65.842857
bachelor's degree	73.282609	66.902778	76.739130	70.611111	78.695652	69.986111
high school	65.000000	60.992857	67.839286	63.450000	68.053571	60.207143
master's degree	70.600000	69.307692	78.250000	73.897436	80.100000	73.410256
some college	71.454545	64.892617	75.987013	66.087248	76.519481	64.872483
some high school	66.701299	61.078431	70.948052	63.911765	70.363636	60.754902

Question4

Look at the effects within subgroups. Compare the average scores for students with/without the test preparation course for different parental education levels (e.g., faceted plots).

In [17]:

```
# making a copy of the dataframe
exam2 = exam.copy()
exam2['Average_Score'] = round((exam2.math + exam2.reading + exam2.writing)/3,2)
exam2
```

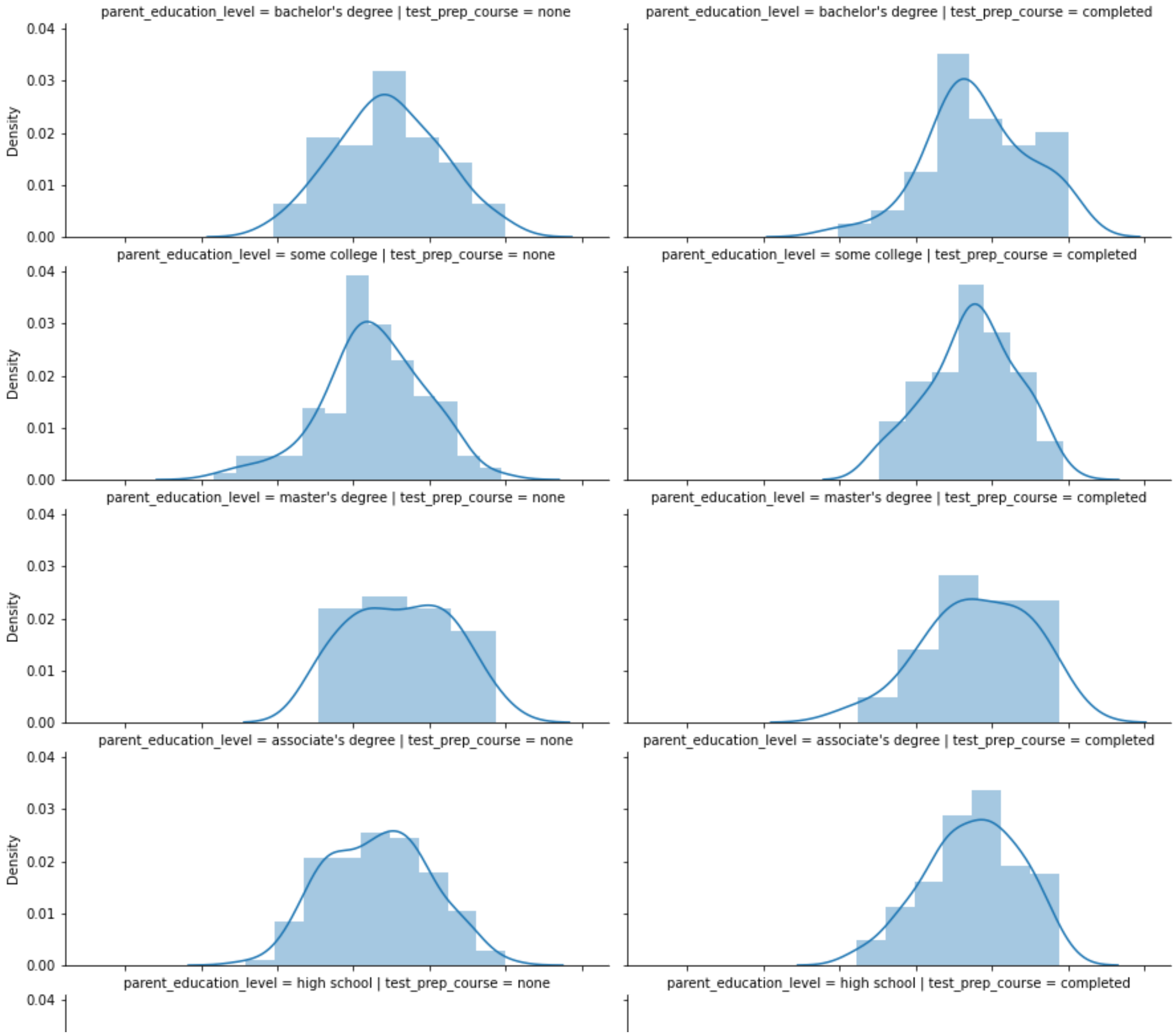
Out[17]:

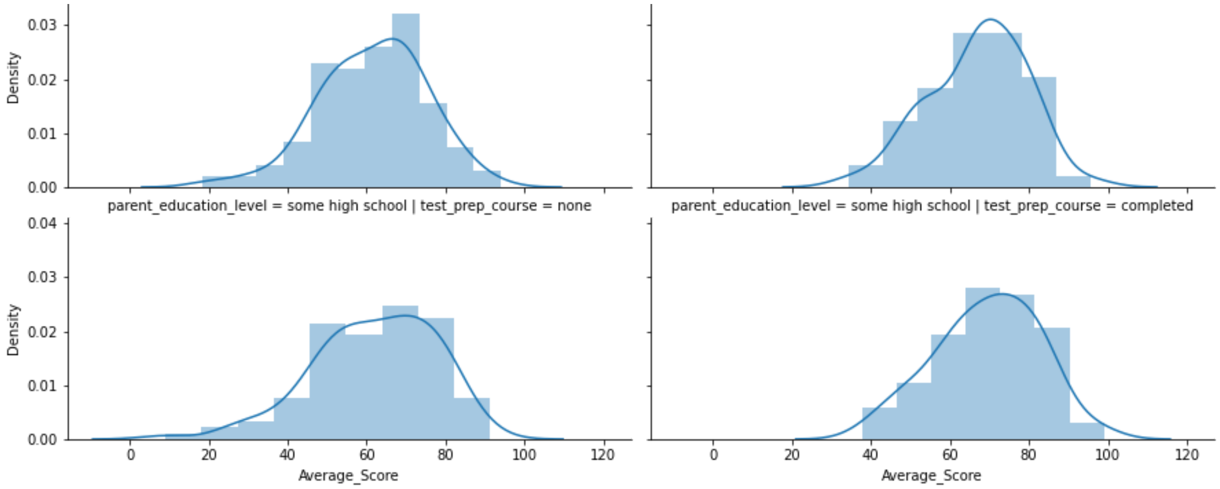
	gender	race/ethnicity	parent_education_level	lunch	test_prep_course	math	reading
0	female	group B	bachelor's degree	standard	none	72	72
1	female	group C	some college	standard	completed	69	90
2	female	group B	master's degree	standard	none	90	95
3	male	group A	associate's degree	free/reduced	none	47	57
4	male	group C	some college	standard	none	76	78
...	...	...	...	...	...	...	...
995	female	group E	master's degree	standard	completed	88	99
996	male	group C	high school	free/reduced	none	62	55
997	female	group C	high school	free/reduced	completed	59	71
998	female	group D	some college	standard	completed	68	78
999	female	group D	some college	free/reduced	none	77	86

1000 rows × 9 columns

```
In [22]: g2 = sns.FacetGrid(exam2, col = 'test_prep_course', row = 'parent_education_level
         g2 = g2.map(sns.distplot,'Average_Score')
         g2.fig.subplots_adjust(top = 0.7)
         g2.fig.suptitle("Average Score for students with/without the test preparation cou
                         fontsize = 18)
         plt.show()
```

Average Score for students with/without the test preparation course for parental education





Question5

- The principal wants to know if kids who perform well on one subject
    - also score well on the others. Look at the correlations between scores.

In [19]:
```
pd.plotting.scatter_matrix(exam, figsize = (15,15), marker = '*');
plt.suptitle("correlations between subject scores", fontsize =18)
plt.show()
```

Question6

Summarize your findings

- The students who completed the test preparation course had a higher reading average score compared to those who did not do the test preparation course.
- The exam scores for math, reading and writing are correlated with the parent education levels. Master's degree students acheive best average scores for the three exams while high school

acheive the lowest average scores for the exams. The difference between the average scores for different parental education levels is not very wide. Average scores for different parent education levels increase with the educational level exposure; the more students advance in their education, the better their scores.

- Average scores for different parent education levels for those that completed the test preparation course and those that did not take the test preparation course show a symmetrical distribution. This shows that the test preparation course has no effect on the student's scores. After estimating the mean and the median, they all occured at the same point for the different levels, thus a symmmetrical distribution.
- The student's scores are highly positively correlated for the different subjects; a student who performs well on one subject also performs well on the others and likewise a student performing poorly in one subject performs poorly on the others.

In [ ]: