In [1]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
import warnings
warnings.filterwarnings("ignore")
```

In [2]:
```
exam = pd.read_csv("C:\\Users\\gacek\\OneDrive\\Documents\\Jupiter N\\data\\exams.csv")
```

In [3]:
```
exam.head(5)
```

Out[3]:

	gender	race/ethnicity	parent_education_level	lunch	test_prep_course	math	reading	writing
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

In [4]:
```
exam.tail(5)
```

Out[4]:

	gender	race/ethnicity	parent_education_level	lunch	test_prep_course	math	reading	writing
995	female	group E	master's degree	standard	completed	88	99	95
996	male	group C	high school	free/reduced	none	62	55	55
997	female	group C	high school	free/reduced	completed	59	71	65
998	female	group D	some college	standard	completed	68	78	77
999	female	group D	some college	free/reduced	none	77	86	86

In [5]:
```
exam.sample(3)
```

Out[5]:

	gender	race/ethnicity	parent_education_level	lunch	test_prep_course	math	reading	writing
144	male	group D	some college	standard	none	88	73	78
271	male	group C	some college	standard	none	58	49	42
41	female	group C	associate's degree	standard	none	58	73	68

In [6]:
```
exam.columns.to_list()
```

Out[6]:
```
['gender',
 'race/ethnicity',
 'parent_education_level',
 'lunch',
 'test_prep_course',
 'math',
 'reading',
 'writing']
```

In [7]:
```
exam.shape
```

Out[7]:
```
(1000, 8)
```

In [8]:
```
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 [9]:
```
exam.isna().sum()
```

Out[9]:
```
gender                    0
race/ethnicity            0
parent_education_level    0
lunch                     0
test_prep_course          0
math                      0
reading                   0
writing                   0
dtype: int64
```

Questions

- What are the average reading scores for students with/without the test preparation course?
- What are the average scores for the different parental education levels?
- Create plots to visualize findings for questions 1 and 2.
- [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).
- [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.
- Summarize your findings.

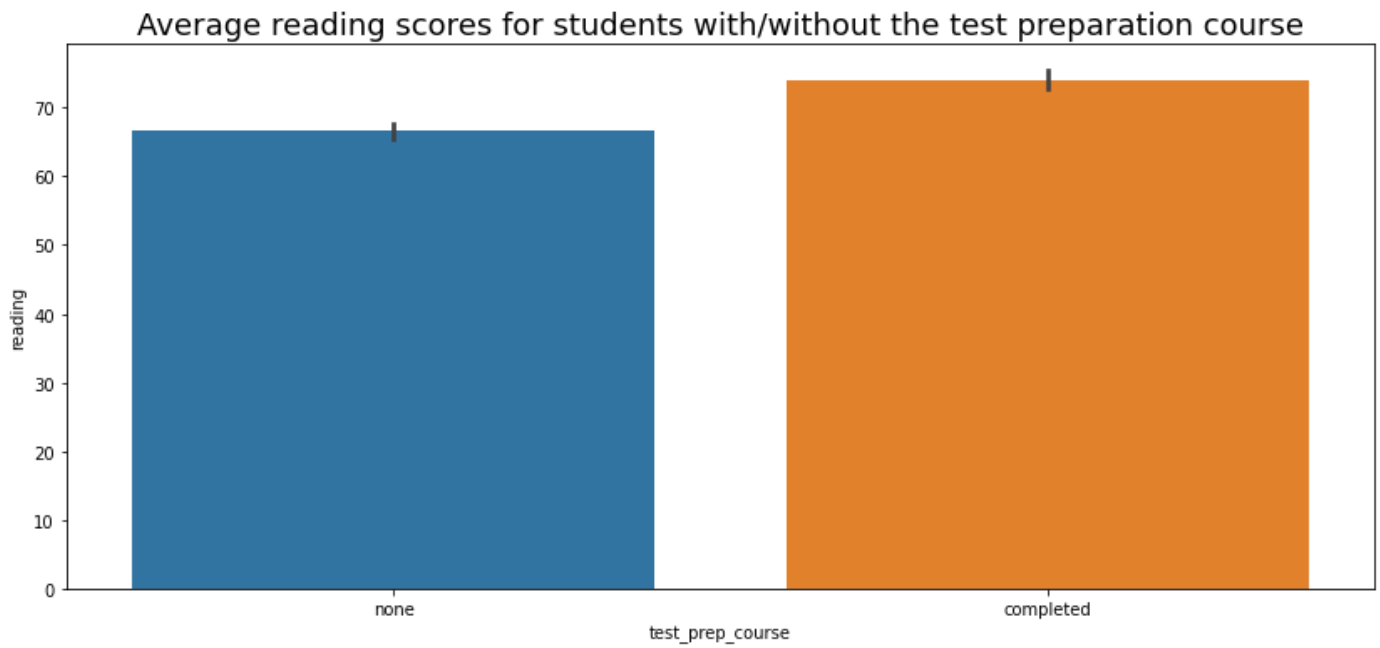
1. average reading scores for students with/without the test preparation course

In [10]:
```
exam.groupby("test_prep_course")[["reading"]].mean()
```

Out[10]:

test_prep_course	reading
completed	73.893855
none	66.534268

In [11]:
```
#3a Visualization for Qstn 1.
plt.figure(figsize = (14,6))
sns.barplot(x = "test_prep_course", y = "reading", data = exam)
plt.title("Average reading scores for students with/without the test preparation course"
plt.show()
```



2. Average scores for the different parental education levels

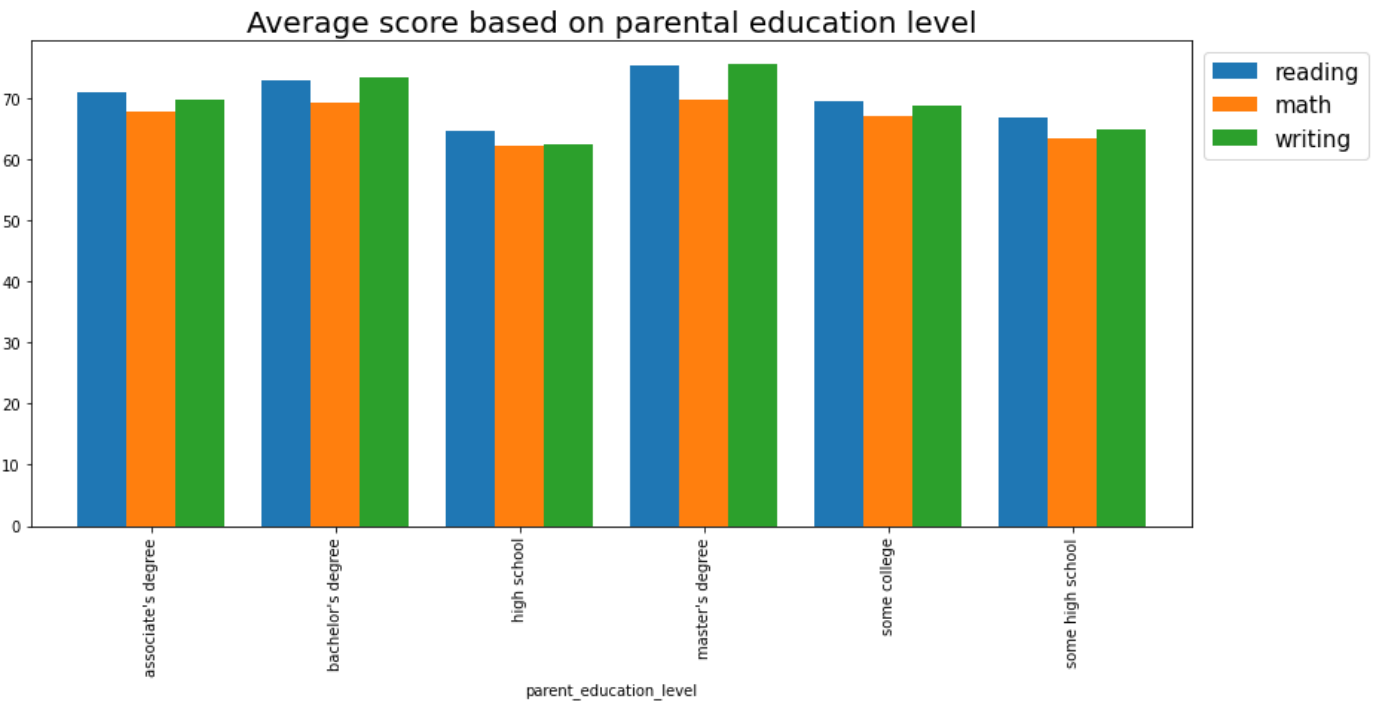
In [12]:
```
p = exam.groupby('parent_education_level')['reading','math', 'writing'].mean()
p
```

Out[12]:

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

In [13]:
```
#3 b. Visualization for Qstn2

p.plot(kind = 'bar', width = 0.8, figsize = (14,6))
plt.legend(bbox_to_anchor =(1,1), fontsize = 15)
plt.title("Average score based on parental education level", fontsize = 20)
plt.show()
```



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).

In [14]:
```
# Make a copy of the dataframe.
exam1 = exam.copy()
```

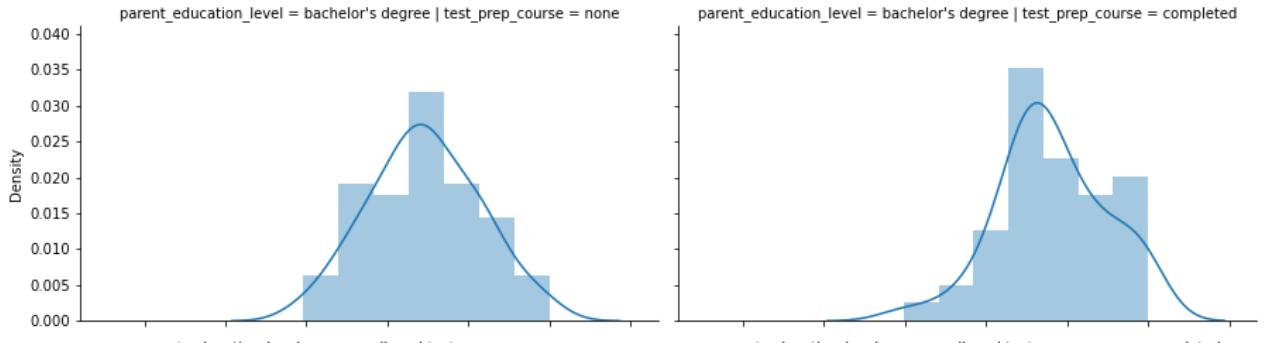
In [15]:
```
exam1['Average_score'] = round((exam1.math + exam1.reading + exam1.writing)/ 3,2)
exam1.sample(2)
```

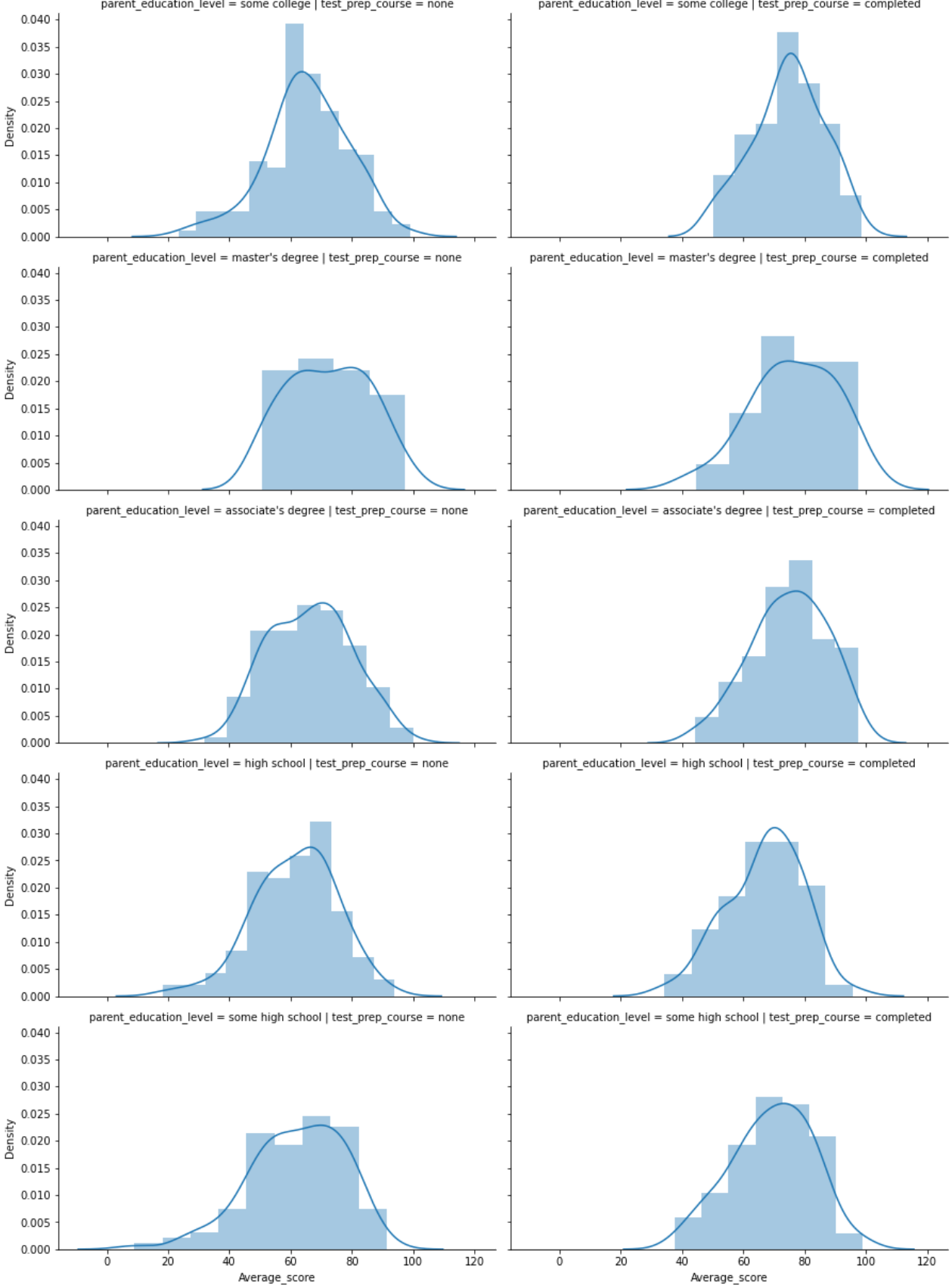
Out[15]:

	gender	race/ethnicity	parent_education_level	lunch	test_prep_course	math	reading	writing	Avera
164	female	group E	master's degree	standard	none	81	92	91	
225	female	group E	master's degree	free/reduced	none	45	56	54	

In [16]:
```
g = sns.FacetGrid(exam1, row = 'parent_education_level', col = 'test_prep_course', heigh
g = g.map(sns.distplot,'Average_score')
g.fig.subplots_adjust(top = 0.9)
g.fig.suptitle("Average scores for students with/without the test preparation course for
               fontsize = 18)
plt.show()
```

Average scores for students with/without the test preparation course for different parental education levels



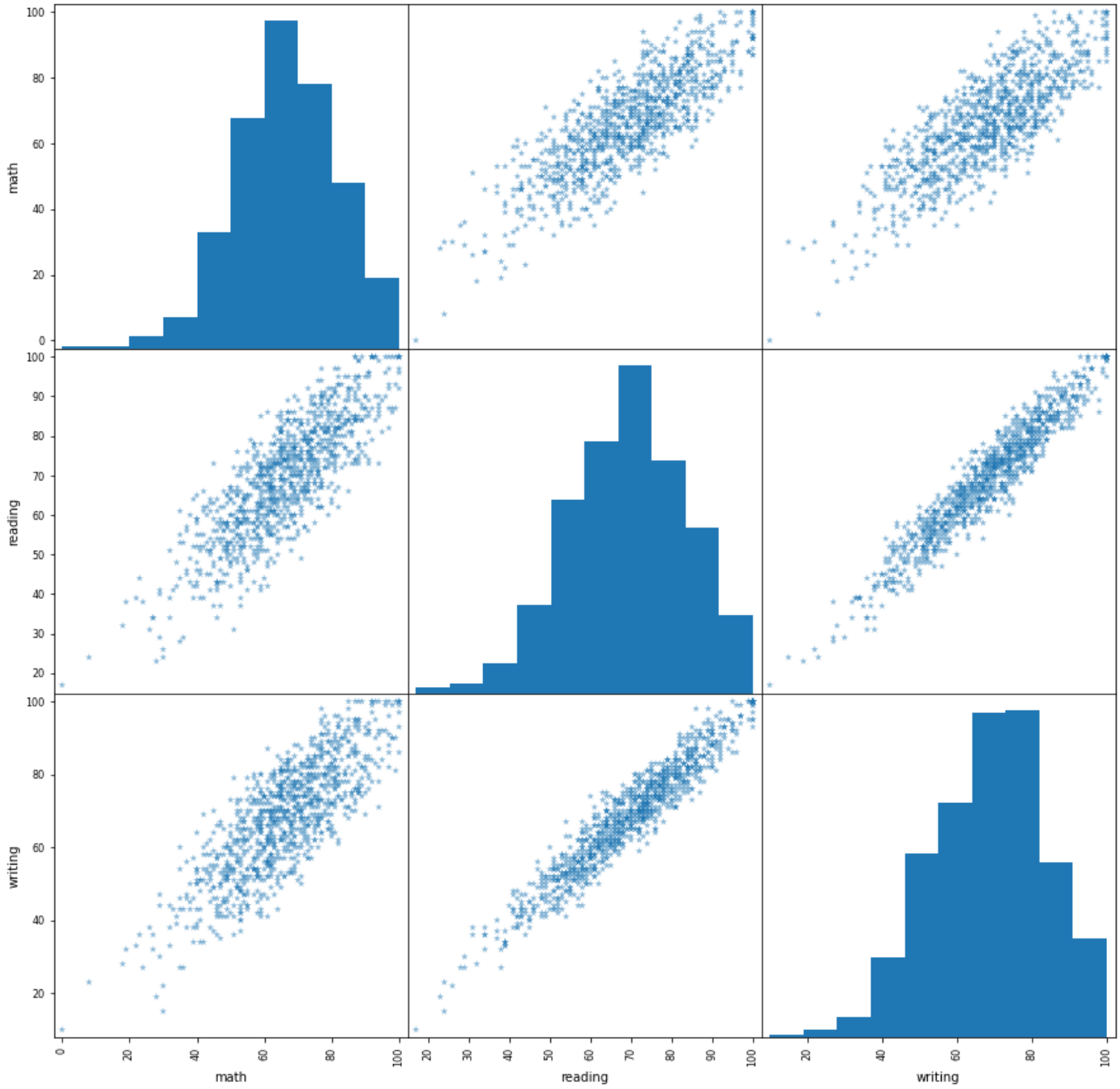


5. [Optional 2] The principal wants to know if kids who perform well on one subject also score well on the others.

Look at the correlations between scores.

In [17]:
```
pd.plotting.scatter_matrix(exam,figsize = (15,15),marker= '*');
plt.suptitle("Correlations between subject scores", fontsize = 18)
plt.show()
```

Correlations between subject scores



6. Summarize your findings.

- The students who completed the test preparation scores 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.
- 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 score.

- The student's scores are positively correlated for the different subjects; a student who performs well on one subject also performs well on the others and a student performing poorly in one subject performs poorly on the others.

In [ ]: