Students Performance in Exams

Marks secured by the students in various subjects

09/21

—

Carolina Marques

—

Prologue

In this notebook, we will dive deep into understanding how such factors affect the performance of the students.
Dataset is about student performance in a different skills such as maths,reading and writing. It contains 1000 rows and 8 columns. Dataset has the columns named gender of a student, race/ethnicity,parental level of education,lunch time, completion status of test preparation course, maths,reading and writing score.

Initial plan for data exploration

The chosen dataset consists of various characteristics of the student, and it was taken from a free source in internet. We will be working with a data frame which has 1000 observations (rows) and 8 attributes (columns). All the data types share a different characteristic – numeric and categories. We can find dependencies even from looking at the variables.

FEATURES:

- **Gender** - Female/Male
- **Race/Ethnicity** - Group division from A to E
- **Parental Level of Education** - Details of parental education varying from high school to master's degree
- **Lunch** - Type of lunch selected
- **Test Preparation Course** - Course details
- **Math Score** - Marks secured by a student in Mathematics
- **Reading Score** - Marks secured by a student in Reading
- **Writing Score** - Marks secured by a student in Writing

ANSWER TO THE FOLLOWING QUESTIONS ARE GIVEN:

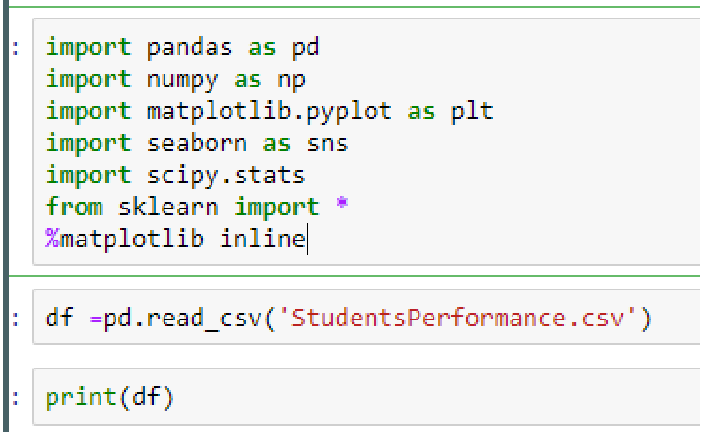
1. Does Gender have any relation with overall score of students in academics.
2. Is parental level of education effect the Students overall performance
3. Does Race have any realation with students performance .
4. We Will do all the Things That will Effect the Students Performance .

LIBRARIES:

Library **pandas** will be required to work with data in tabular representation.
Library **numpy** will be required to round the data in the correlation matrix.
Library **matplotlib, seaborn** required for data visualization.

Data cleaning

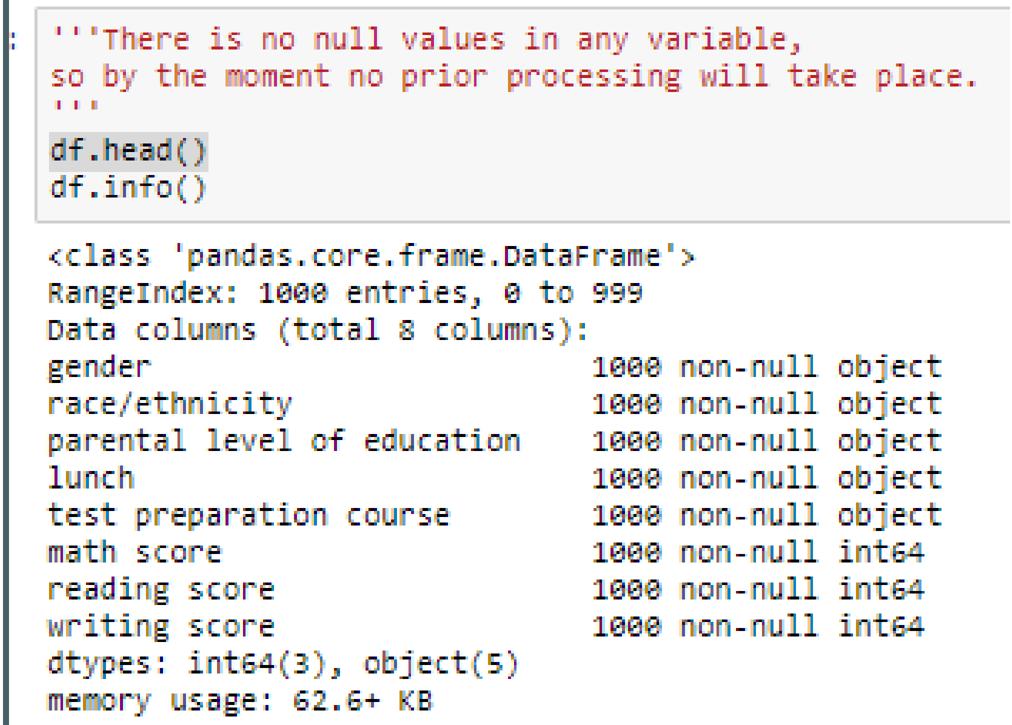
Upon loading the dataset, i proceed with a check of the types of the data



```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import scipy.stats
from sklearn import *
%matplotlib inline
```

```
df =pd.read_csv('StudentsPerformance.csv')
```

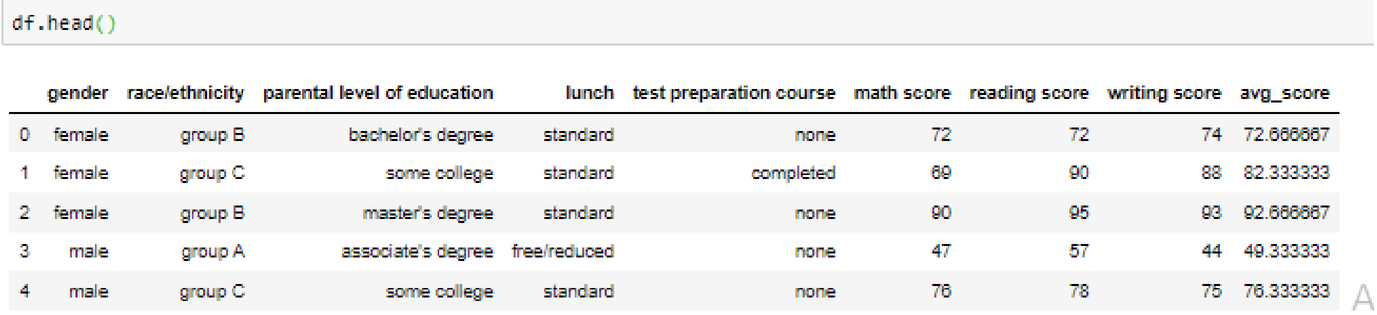
```
print(df)
```



```
'''There is no null values in any variable,
so by the moment no prior processing will take place.
'''
df.head()
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 8 columns):
gender                         1000 non-null object
race/ethnicity                 1000 non-null object
parental level of education    1000 non-null object
lunch                          1000 non-null object
test preparation course        1000 non-null object
math score                     1000 non-null int64
reading score                  1000 non-null int64
writing score                  1000 non-null int64
dtypes: int64(3), object(5)
memory usage: 62.6+ KB
```

Data contains 5 categorical columns, and 3 columns contains numeric values.

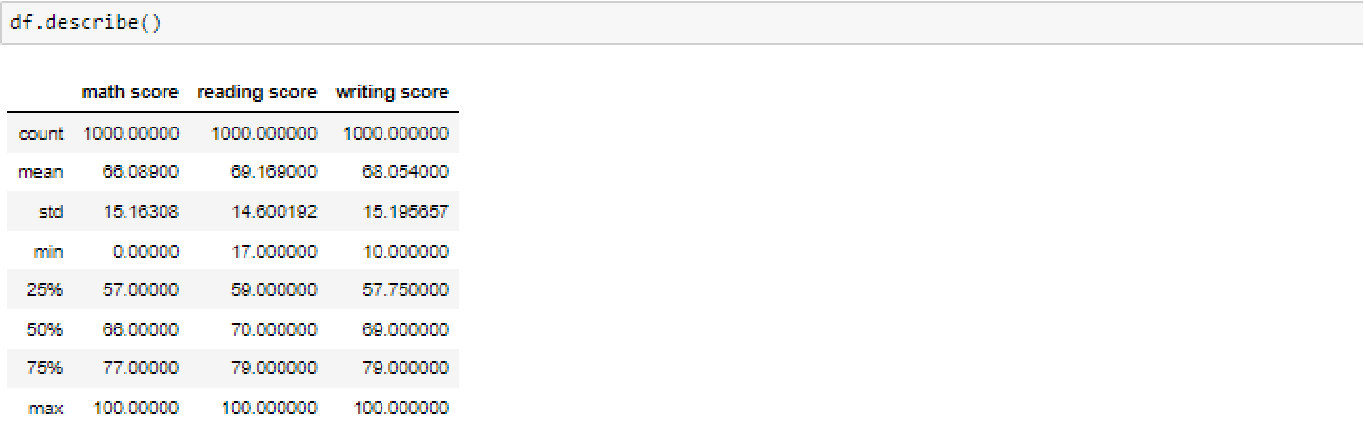


```
df.head()
```

	gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score	avg_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

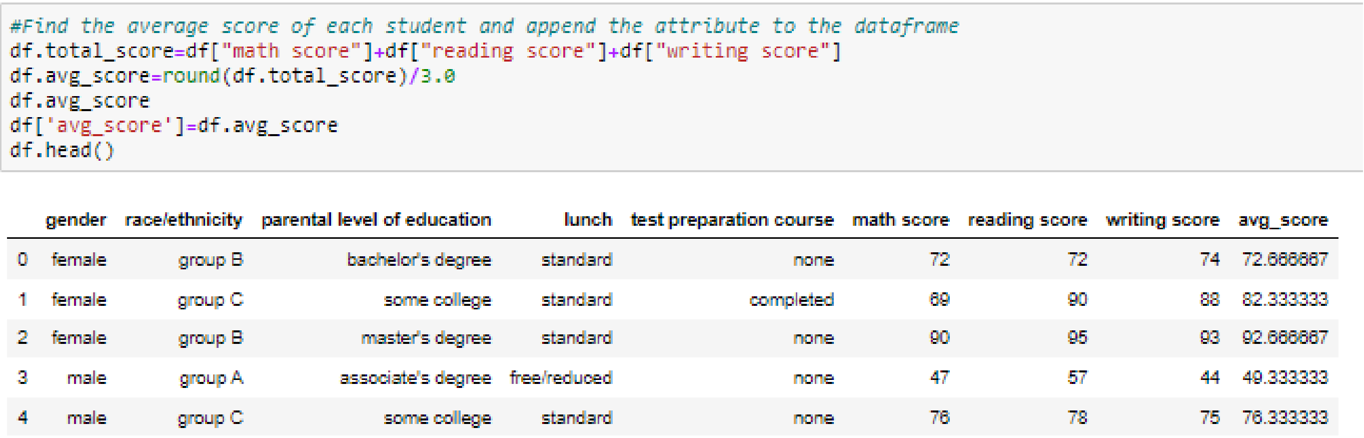
A

In the Column parent level of education. There are the values high school and some high school those are same, so we change some high school to high school. Adding total and average score of every student as separate column. Let us find the average score of each student to make a visualization easy. Below code will find the average score of each student and append it to the dataframe with a label 'avg_score'.



```
df.describe()
```

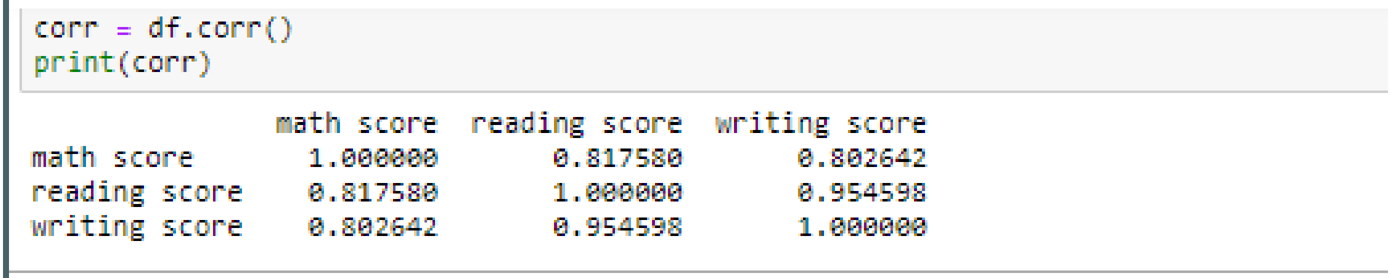
	math score	reading score	writing score
count	1000.00000	1000.000000	1000.000000
mean	66.08900	69.169000	68.054000
std	15.16308	14.600192	15.195657
min	0.00000	17.000000	10.000000
25%	57.00000	59.000000	57.750000
50%	66.00000	70.000000	69.000000
75%	77.00000	79.000000	79.000000
max	100.00000	100.000000	100.000000



```
#Find the average score of each student and append the attribute to the dataframe
df.total_score=df["math score"]+df["reading score"]+df["writing score"]
df.avg_score=round(df.total_score)/3.0
df.avg_score
df['avg_score']=df.avg_score
df.head()
```

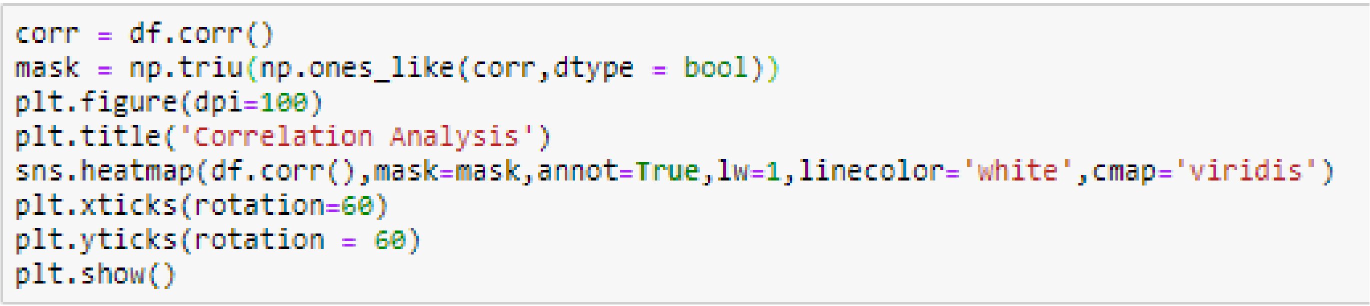
	gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score	avg_score
0	female	group B	bachelor's degree	standard	none	72	72	74	72.666667
1	female	group C	some college	standard	completed	69	90	88	82.333333
2	female	group B	master's degree	standard	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

Correlation with Original Data, We will use the Pandas function dataframe.corr()to find the correlation between numeric variables only.The return of this function give us a score ranging from -1 to 1that indicates if there is a strong linear relationship in a positive or negative direction.

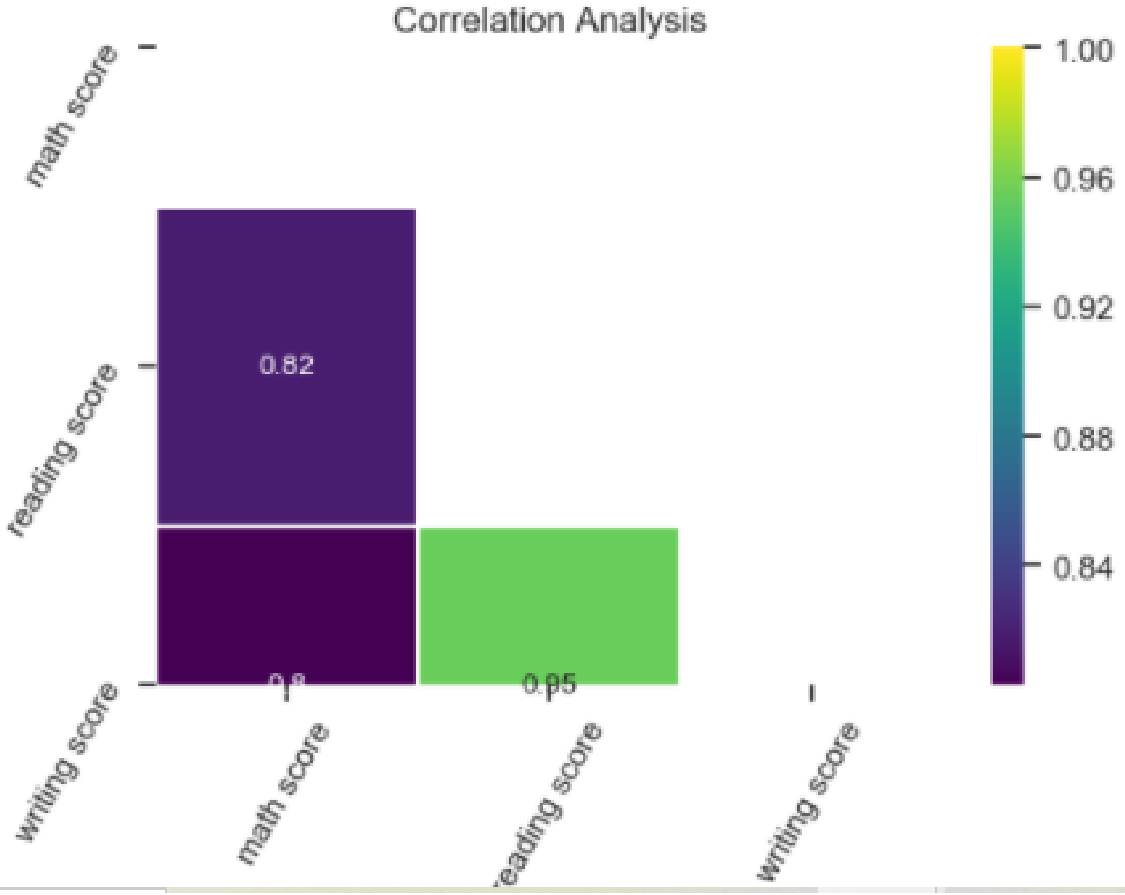


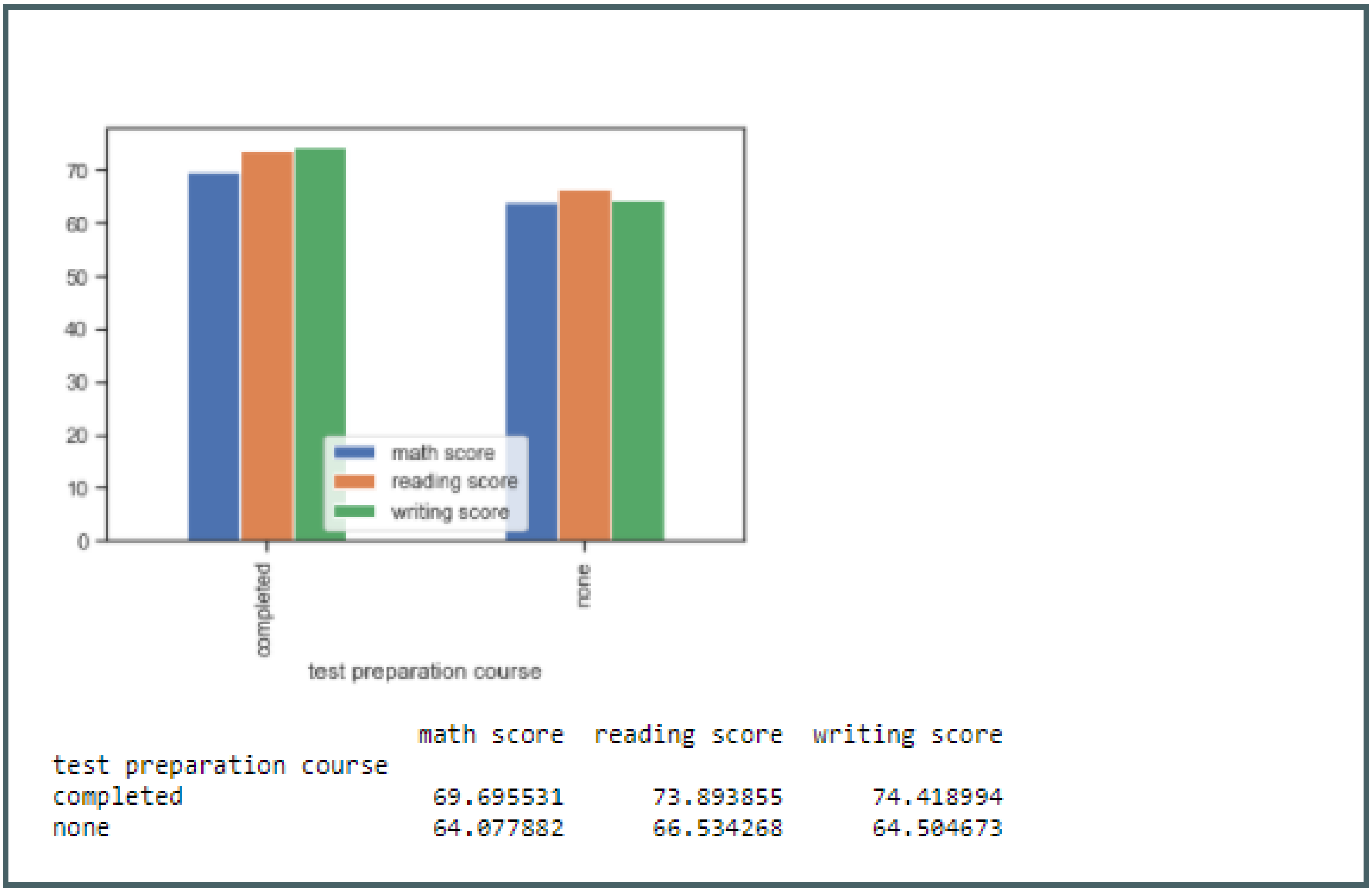
```
corr = df.corr()
print(corr)
```

```
               math score  reading score  writing score
math score       1.000000       0.817580       0.802642
reading score    0.817580       1.000000       0.954598
writing score    0.802642       0.954598       1.000000
```

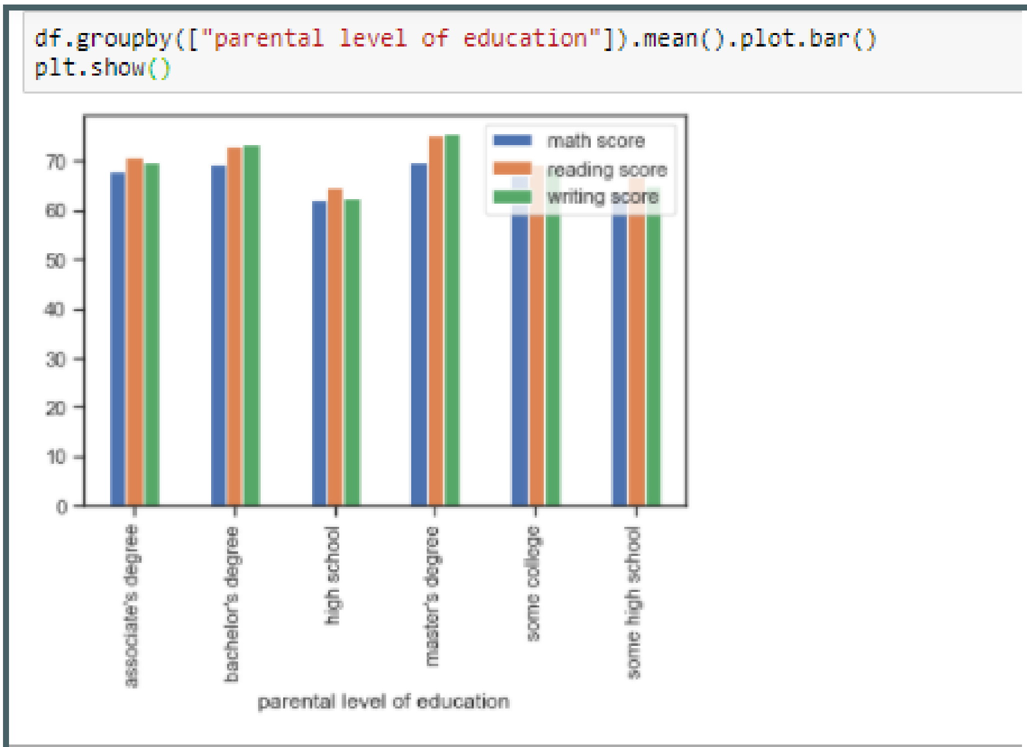


```
corr = df.corr()
mask = np.triu(np.ones_like(corr,dtype = bool))
plt.figure(dpi=100)
plt.title('Correlation Analysis')
sns.heatmap(df.corr(),mask=mask,annot=True,lw=1,linecolor='white',cmap='viridis')
plt.xticks(rotation=60)
plt.yticks(rotation = 60)
plt.show()
```

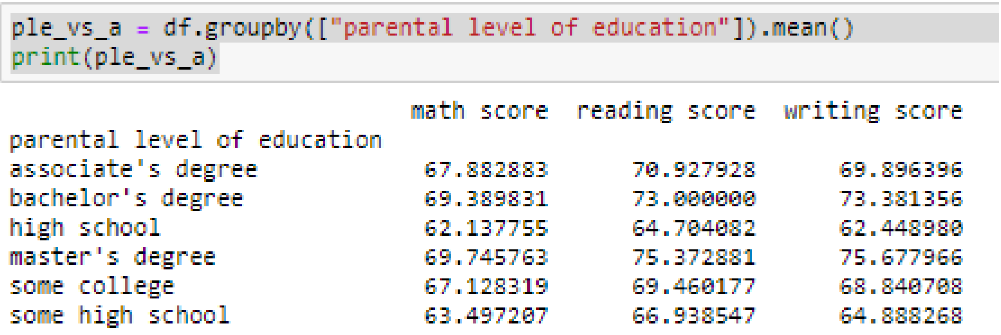




```
                         math score  reading score  writing score
test preparation course
completed                 69.695531      73.893855      74.418994
none                      64.077882      66.534268      64.504673
```



```
df.groupby(["parental level of education"]).mean().plot.bar()
plt.show()
```



```
ple_vs_a = df.groupby(["parental level of education"]).mean()
print(ple_vs_a)
```

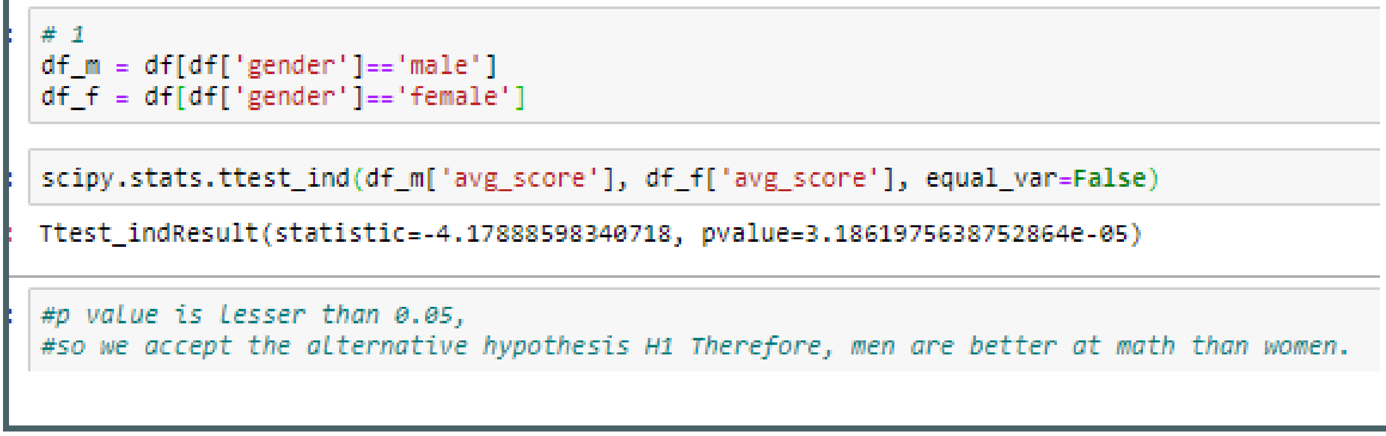
```
                            math score  reading score  writing score
parental level of education
associate's degree           67.882883      70.927928      69.896396
bachelor's degree            69.389831      73.000000      73.381356
high school                  62.137755      64.704082      62.448980
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
```

By observing the above graph we can make a hypothesis as below:

1 - Hypothesis test of average score by gender

Ho (null hypothesis) = There is no difference in math scores between genders.

H1 (alternative hypothesis) = Differences in average scores between genders exist.



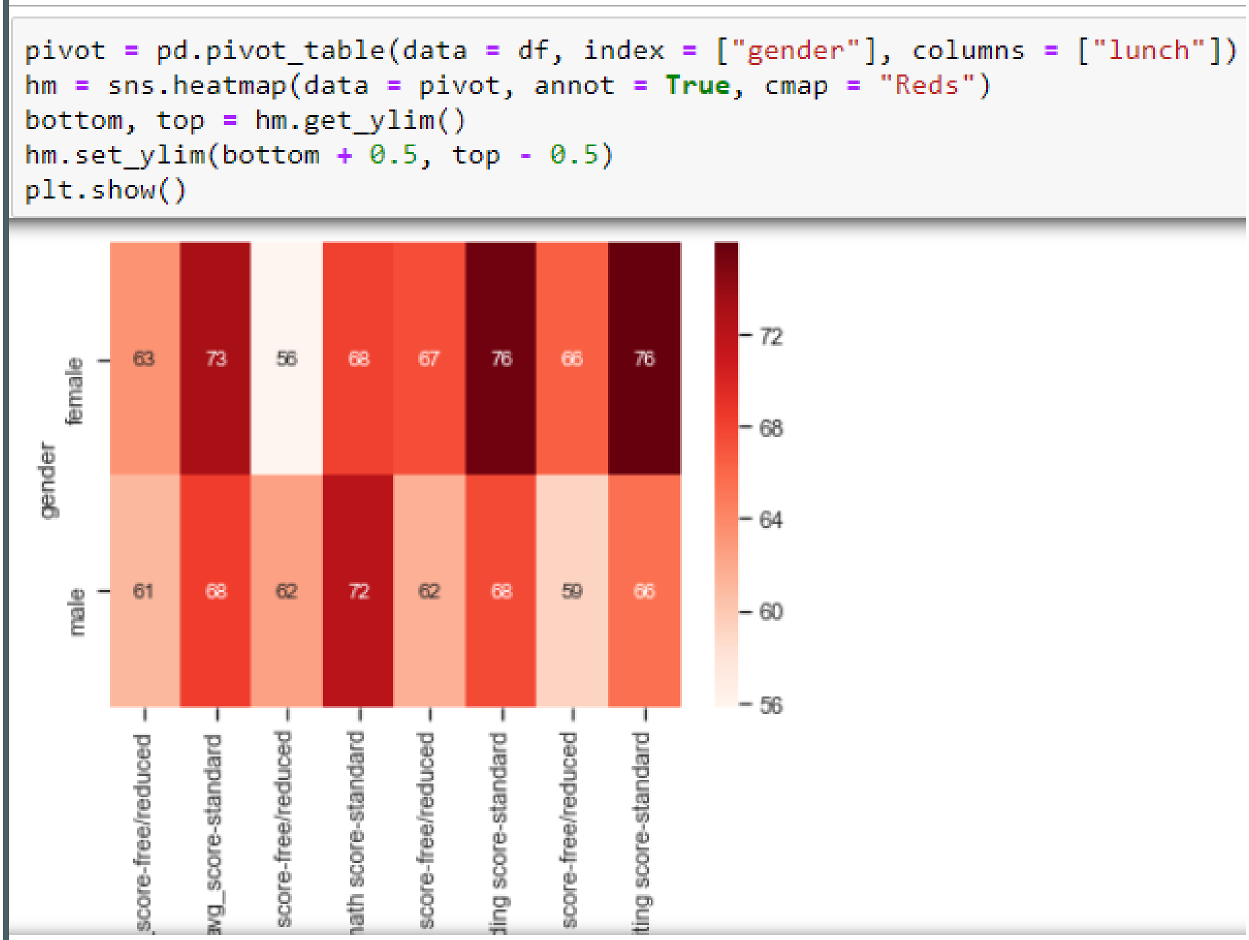
```
# 1
df_m = df[df['gender']=='male']
df_f = df[df['gender']=='female']
```

```
scipy.stats.ttest_ind(df_m['avg_score'], df_f['avg_score'], equal_var=False)
```

```
Ttest_indResult(statistic=-4.17888598340718, pvalue=3.1861975638752864e-05)
```

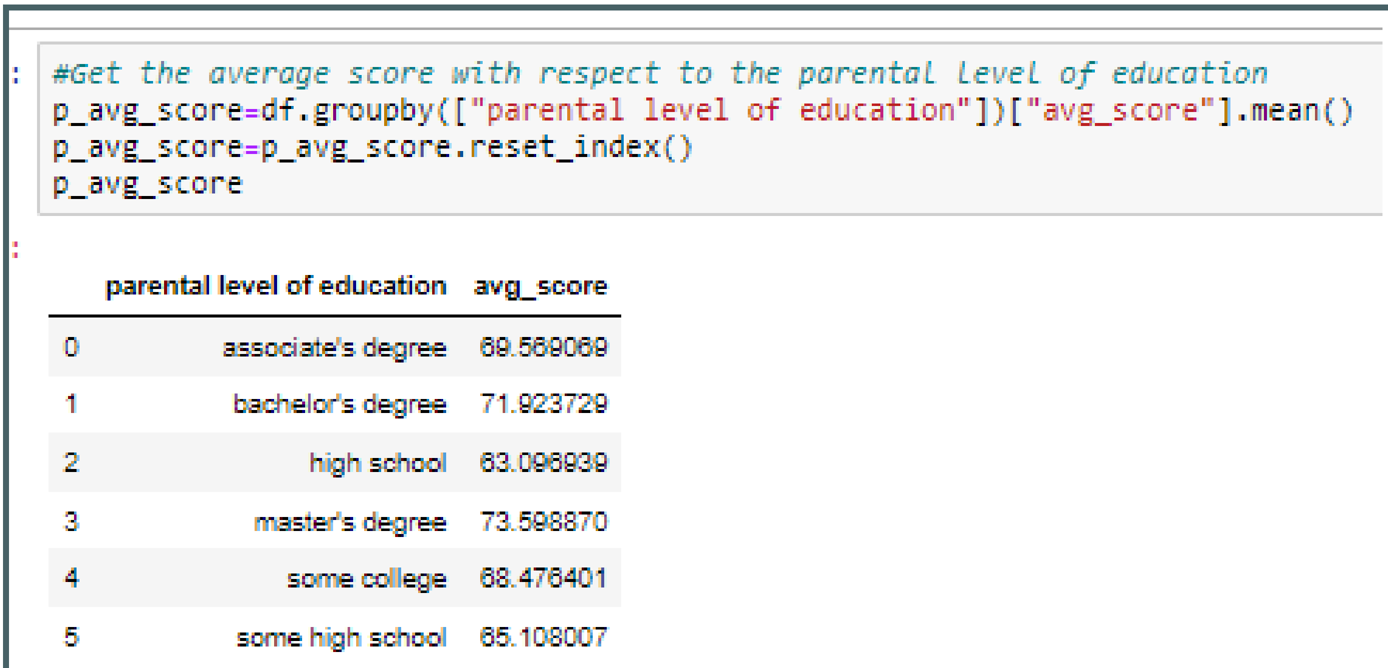
```
#p value is lesser than 0.05,
#so we accept the alternative hypothesis H1 Therefore, men are better at math than women.
```

2 - Hypothesis test of student Type of lunch selected by gender.
Ho = There is no difference in Type of lunch selected between genders.
H1 = Differences in Type of lunch selected between genders exist.



```
pivot = pd.pivot_table(data = df, index = ["gender"], columns = ["lunch"])
hm = sns.heatmap(data = pivot, annot = True, cmap = "Reds")
bottom, top = hm.get_ylim()
hm.set_ylim(bottom + 0.5, top - 0.5)
plt.show()
```

3 - Hypothesis test of average score by parental education level
Ho = There is no difference in average score between parental education level.
H1 = Differences in average scores between parental education exist.



```
#Get the average score with respect to the parental level of education
p_avg_score=df.groupby(["parental level of education"])["avg_score"].mean()
p_avg_score=p_avg_score.reset_index()
p_avg_score
```

	parental level of education	avg_score
0	associate's degree	69.569069
1	bachelor's degree	71.923729
2	high school	63.096939
3	master's degree	73.598870
4	some college	68.476401
5	some high school	65.108007

-By observing the correlation table, it would be interesting to explore deeply how 'gender', 'race/ethnicity', 'lunch' and 'test preparation course' influence our three scores: math, reading and writing. We will leave out 'parental level of education' as its correlation levels seem to be negligible.

-If we go back to the correlation table, we can observe that math score has a slight positive correlation with gender, whereas reading and writing have a slight negative one.

-Students that have completed the test preparation course tend to score better on all three areas: math, reading and writing (seems reasonable, but never forget correlation does not imply causation!). Although records are very unbalanced: there are many more students who have not taken the course than those who have.

Conclusions

As shown in graphs, the highest correlation between variables / features are:

- Writing score and Gender (also Math and Reading but slightly smaller)
- Race/Ethnicity group E and Math Score
- Parental Level of Education High School with Writing and Reading Scores
- Lunch plan and Math Score (also and Reading and Writing but slightly smaller)
- Test preparation course and Writing and Reading Score (also Math but slightly smaller)