**Principles of Data Science**

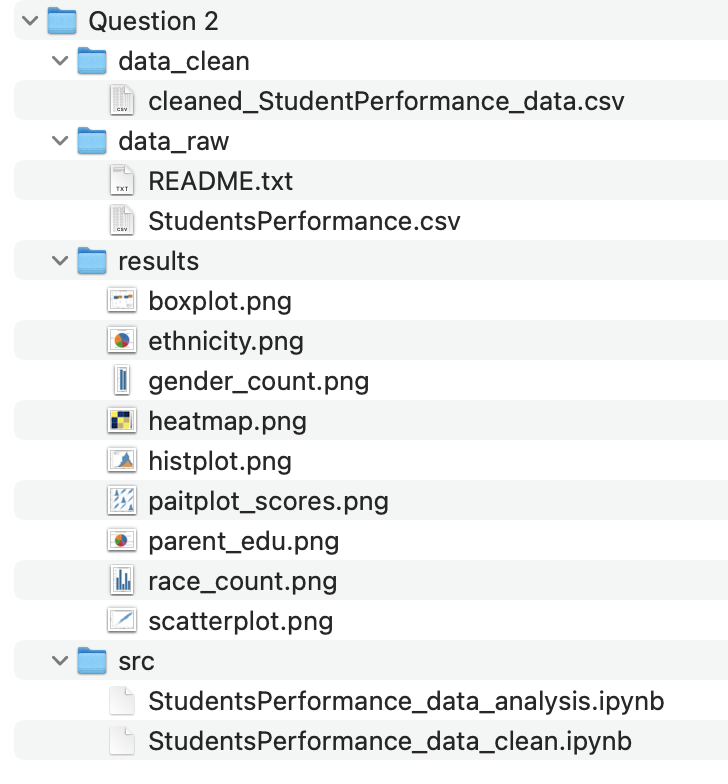
**Assignment-1**

Name: Sreevardhan Reddy Soma ID: 16352646

Question 2:

Step 1: Data Collection

* Given data set is stored in the form of CSV which makes analysis of the data easier as it will be in structured format.



* Here the data set is stored under data\_raw and README.txt file contains all the details about the dataset.

The dataset is Students performance which has all the details about their gender, race/ethnicity, parents education level, the type of lunch, test preparation course, math score, reading and writing scores. There are 1000 rows and 8 columns in the dataset.

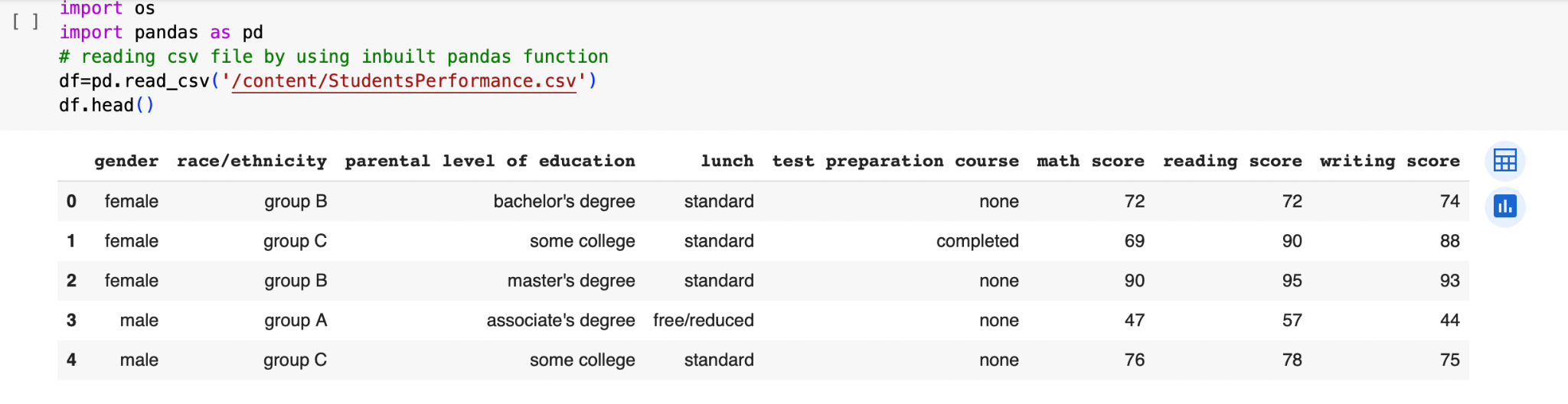
Description of dataset:

1. Gender: Gender of the student whether Male or female
2. Race/ethnicity: They are categorized into five groups namely Group A,B,C,D and E.
3. Parental level of education: Details of the students parent’s education like masters, some college, high school e.t.c.
4. Lunch: whether the lunch of the students is standard or free/reduced.
5. Test preparation course: whether student have completed the course or not
6. Math score: Score obtained by a student in math
7. Reading Score: Score obtained by a student in reading
8. Writing Score: Score obtained by a student in writing

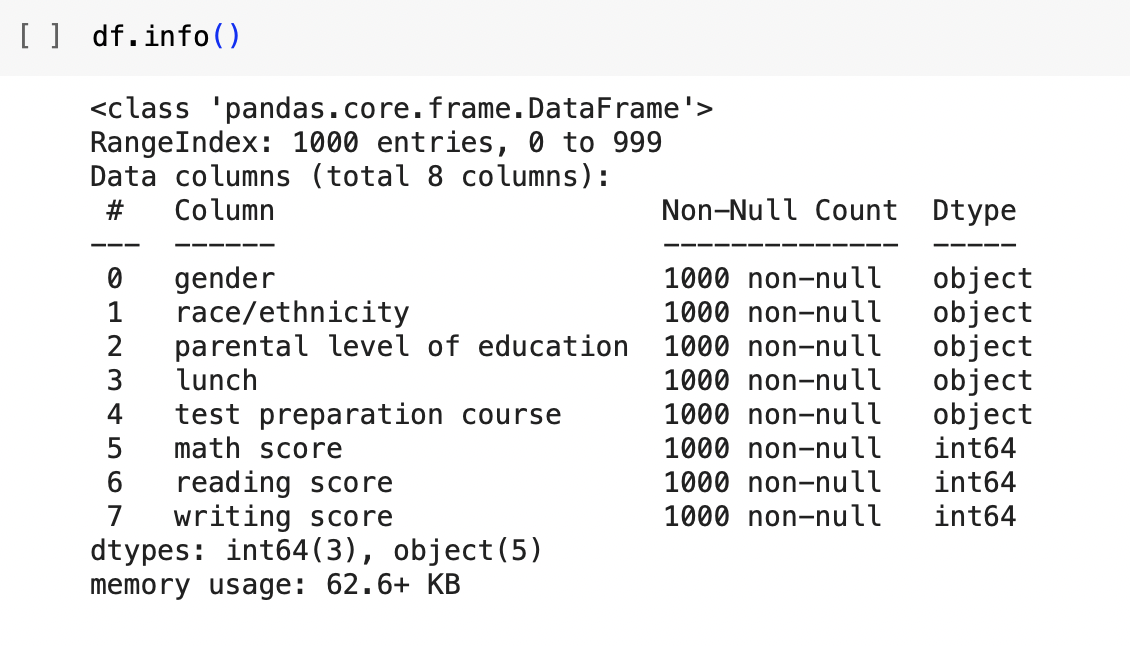
Data is in taken in CSV format

Step 2: Data Processing

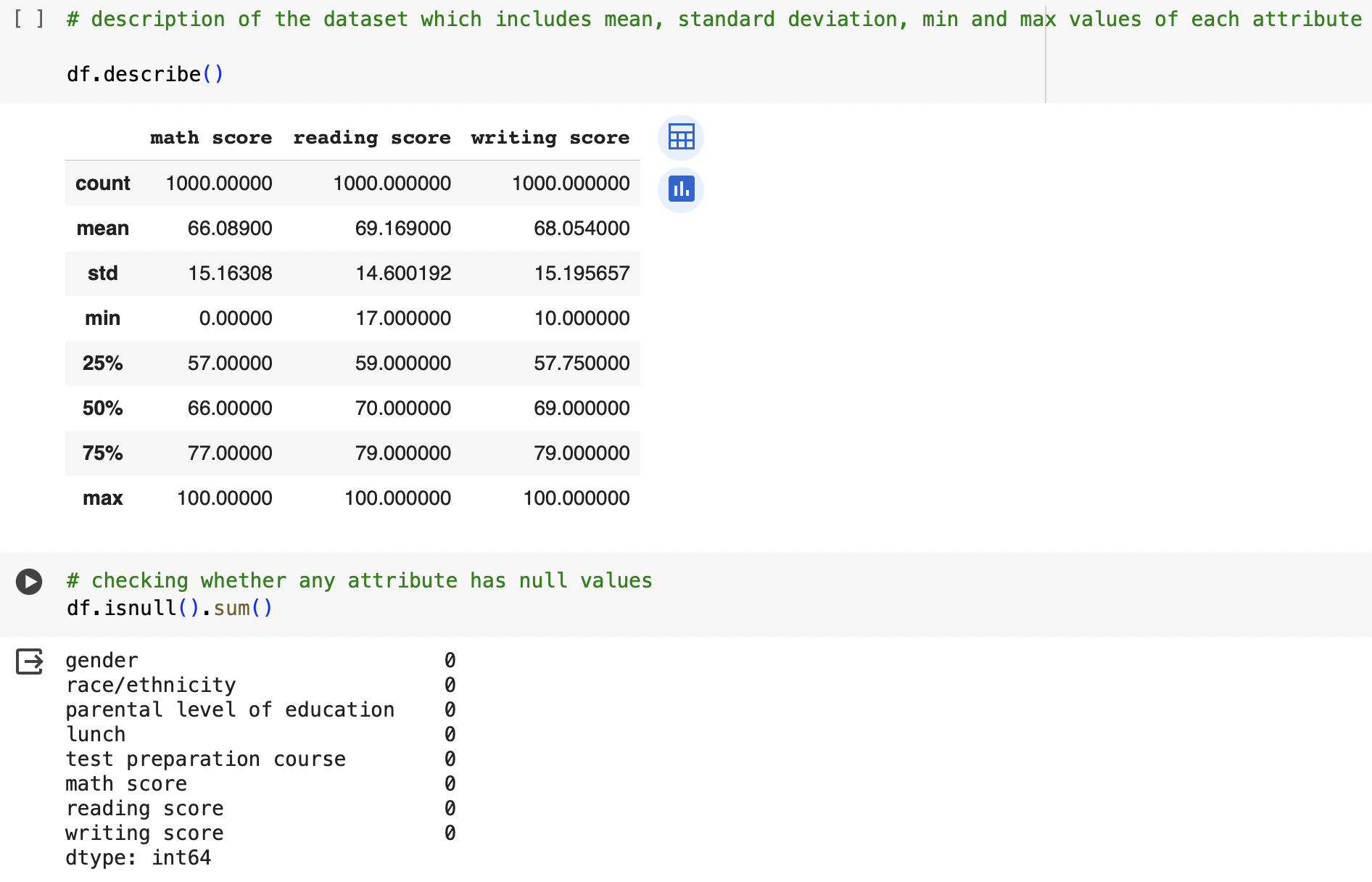
* Here we will perform some operations on the dataset to determine whether there were any null values or missing values present in our dataset. If so, we need to eliminate or replace them according to the problem statement.



* head() will give the first five observations of the data set.

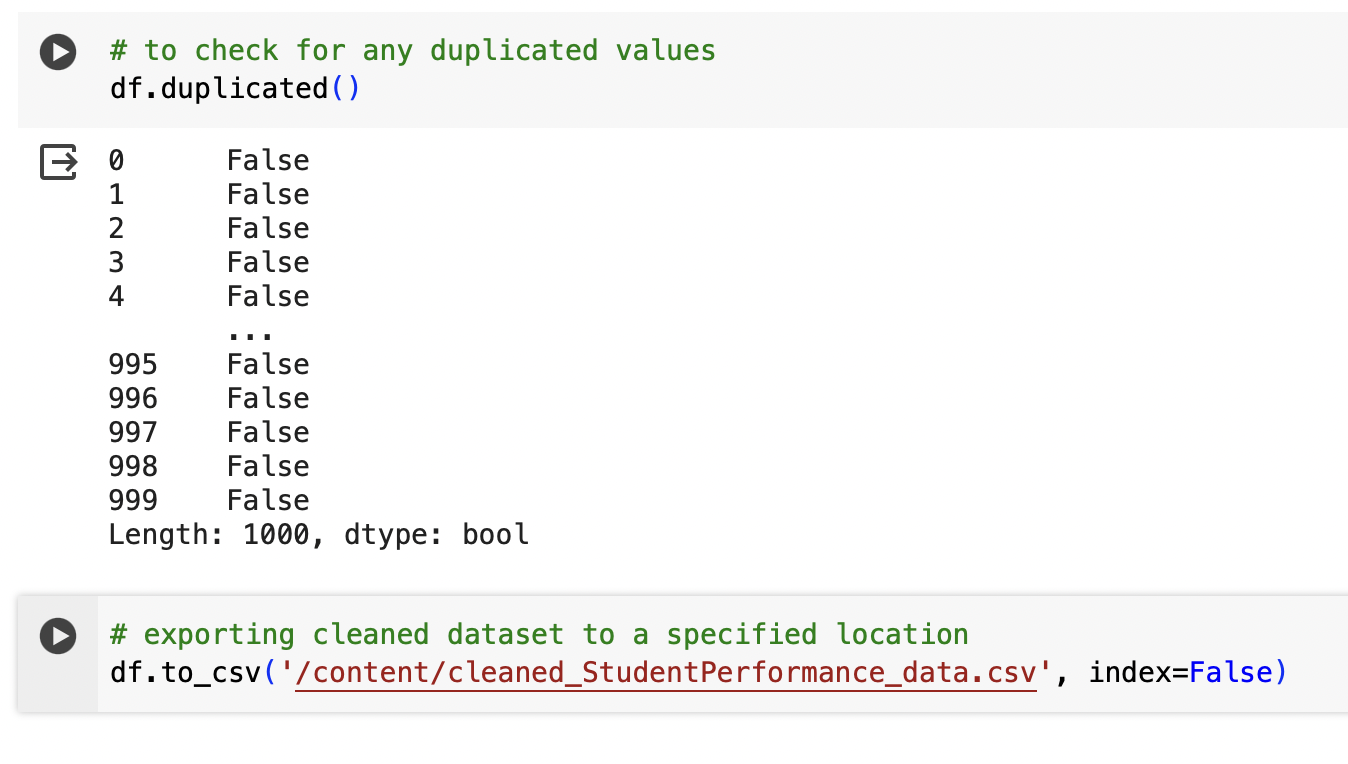


* By using info(), we will get to know about the data type of the attributes and also if there were any null values present in the dataset. Here we can see that there were no null values.



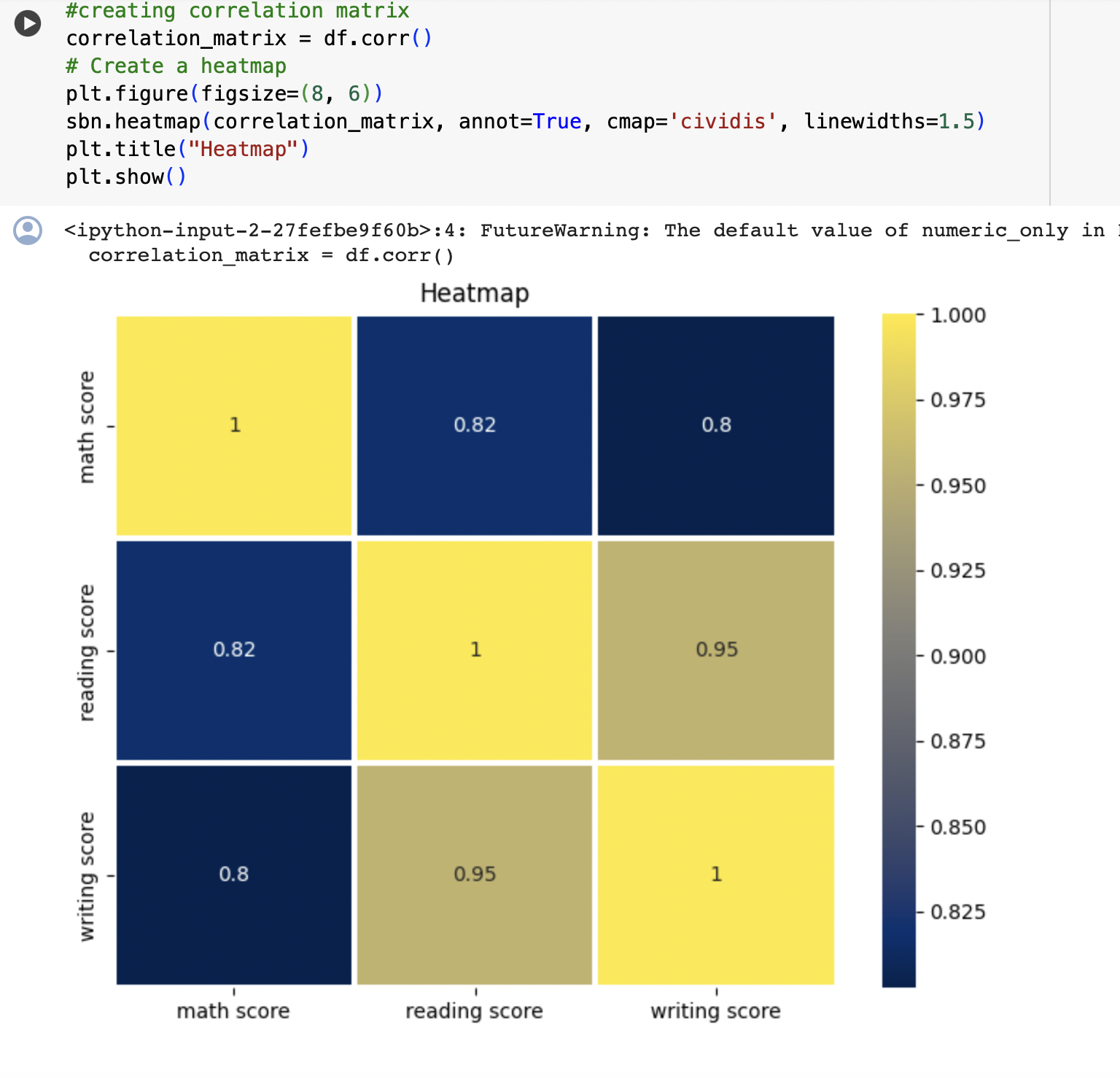
* describe() tells us the clear picture about the dataset that includes min and max values, standard deviation and also percentages at different stages which helps us in finding out the outliers.
* We can also check for null values using the isnull().sum().



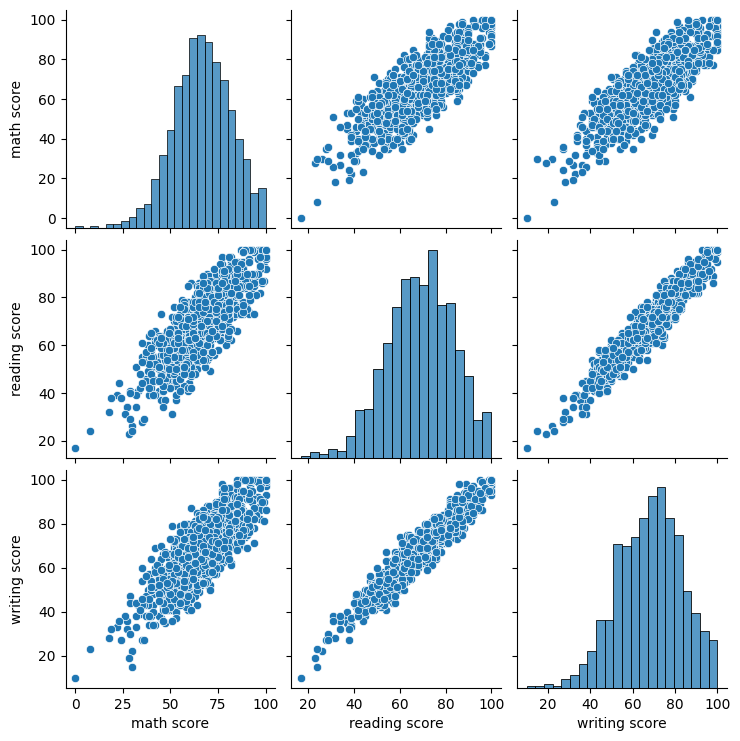


* isna() works the same as null function and gives us any null values present in the data set as true.
* After performing all these steps, we can download the cleaned dataset to the appropriate location.

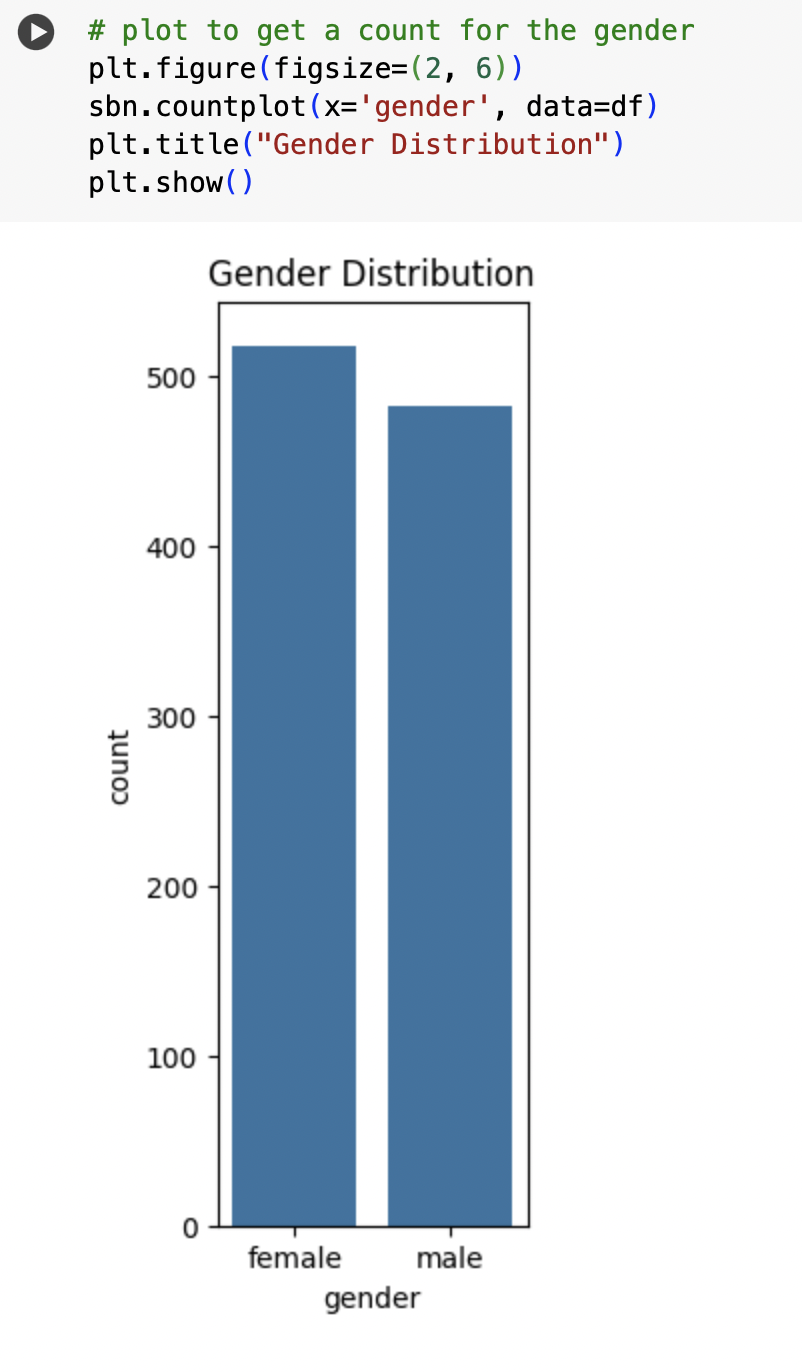
Step 3: Data Analysis



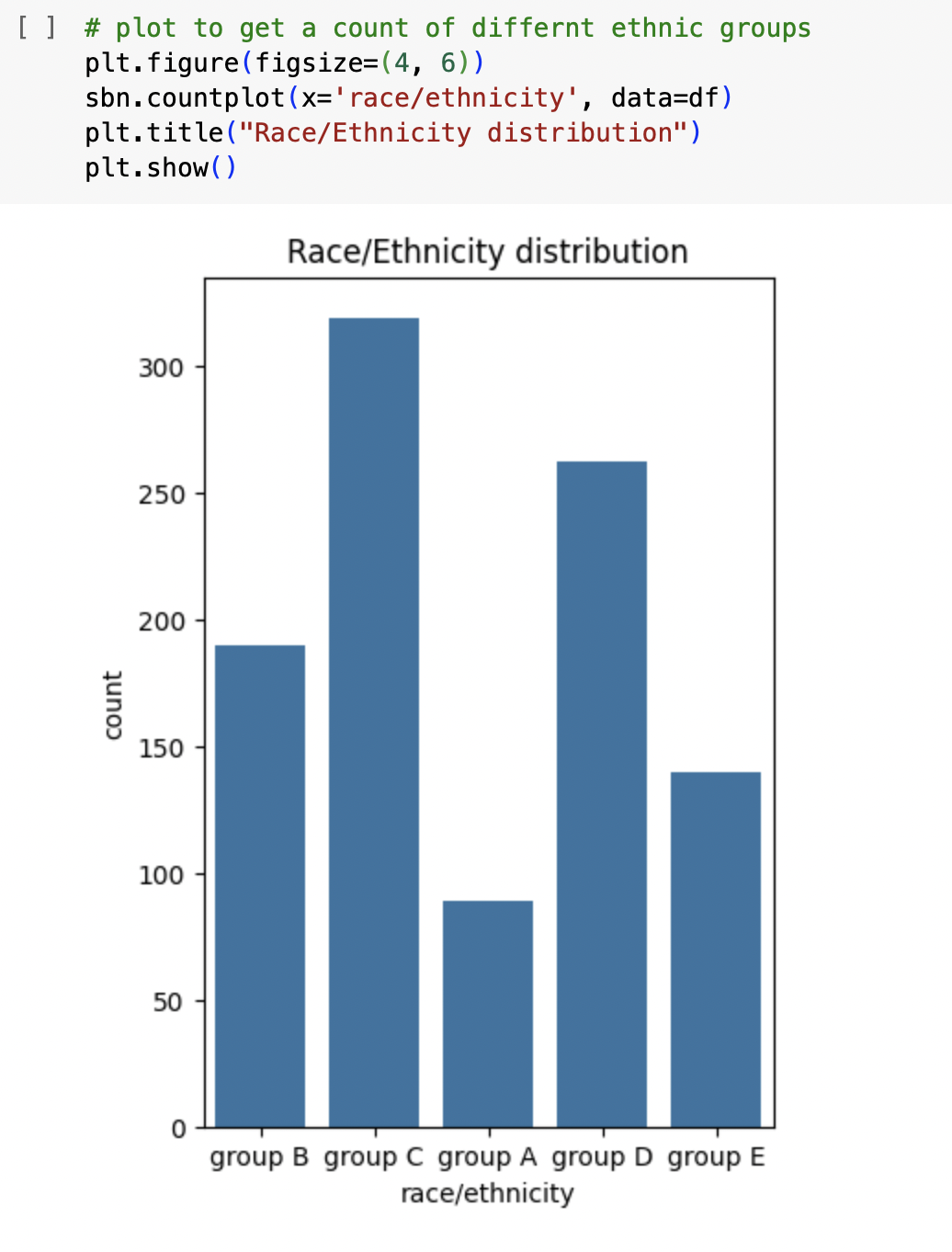
* The heat map provides us the relationship between the attributes. If the value in the heatmap is positive between two attributes, then we can say that those two attributes are positively correlated if not, negatively correlated. Here we can see that the value between reading score and writing score is 0.95 which is an obvious fact.



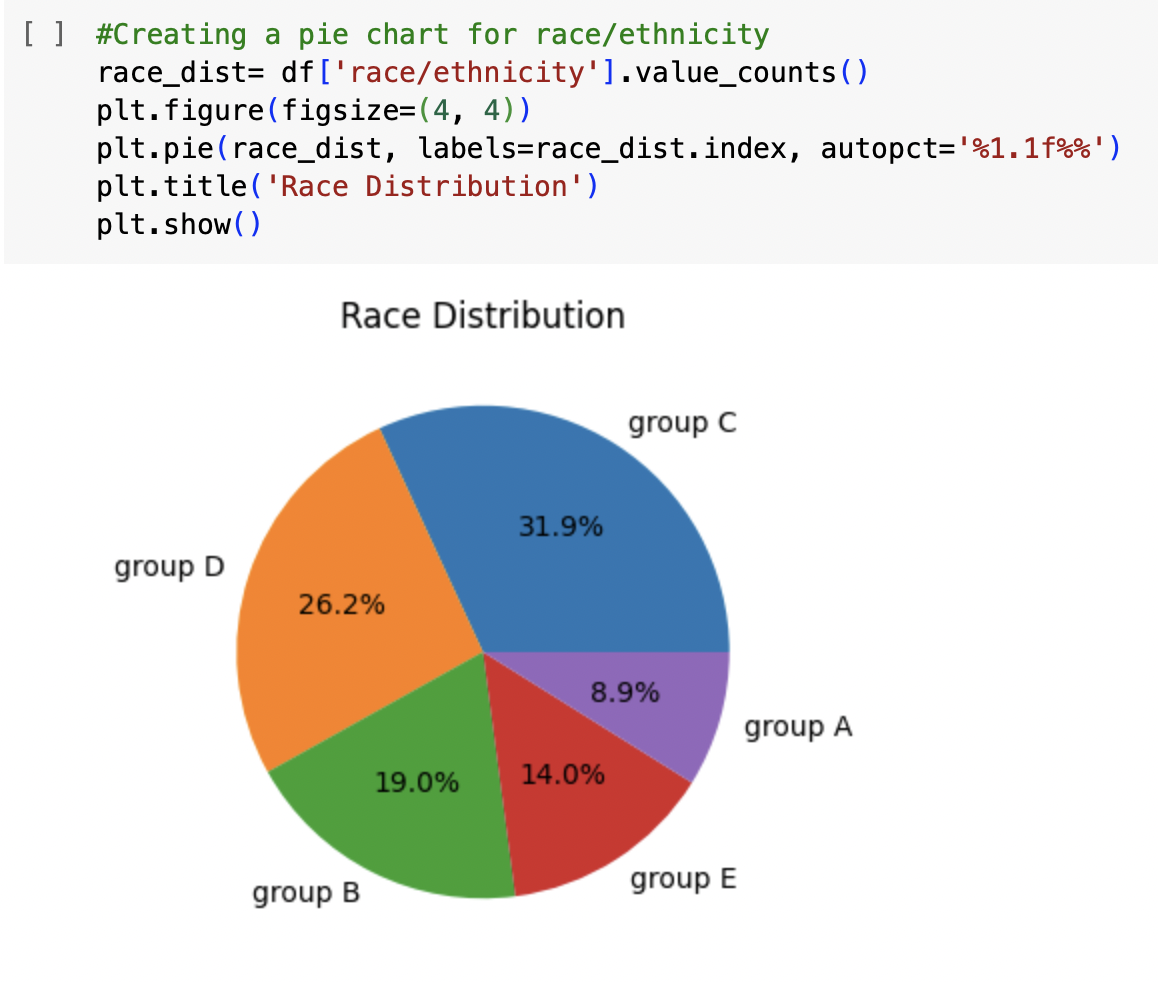
* In the pairplot, different scatter and histoplots are obtained which shows the relationship between math score, writing and reading score. Usually students with a high reading score also have a high writing score. Most of the dataset can be found between 75-85 in all the three scores.



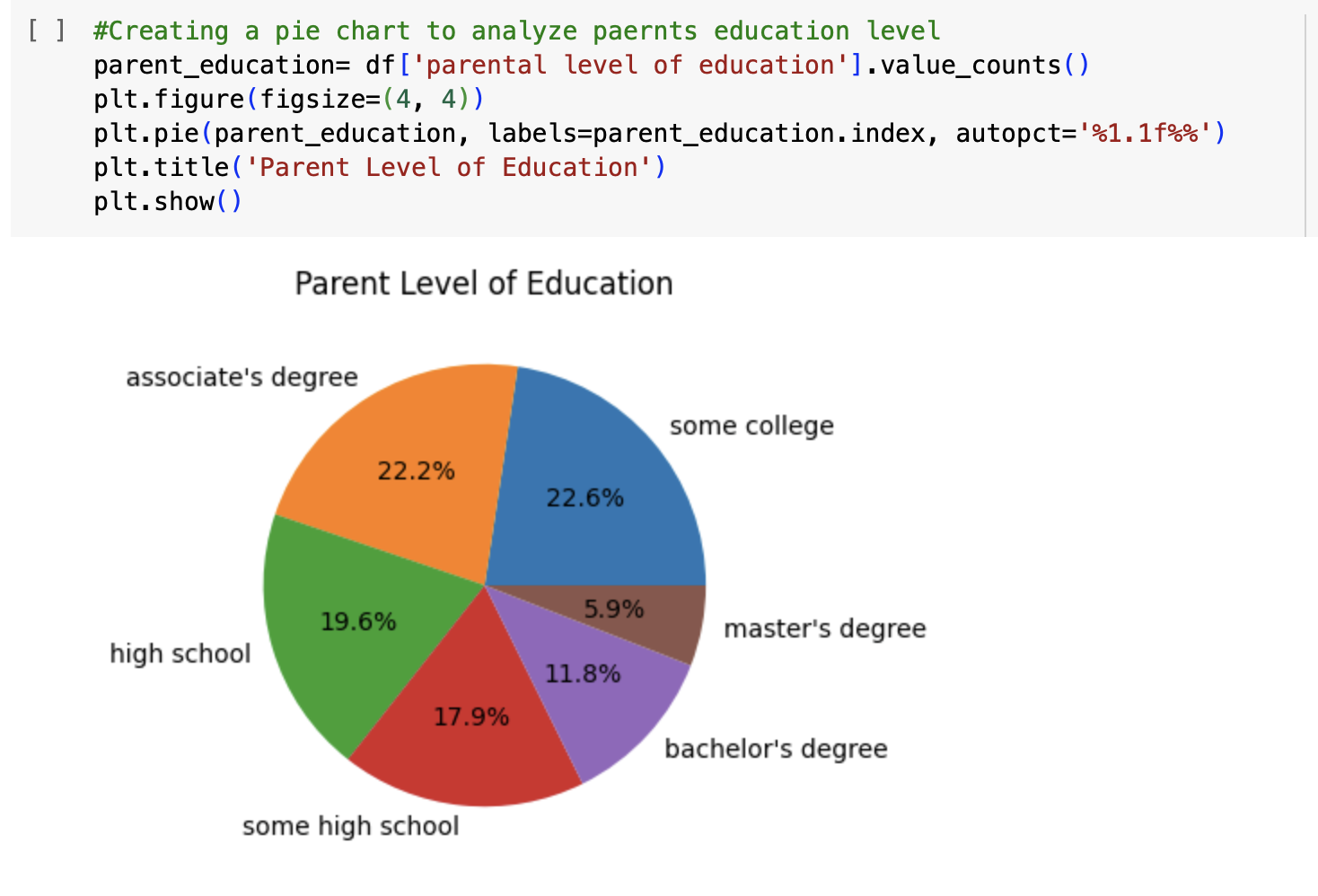
* From the dataset, we can obtain the count of male and female by the countplot. Female count is higher than male count. Females are around 525 and males are around 475.



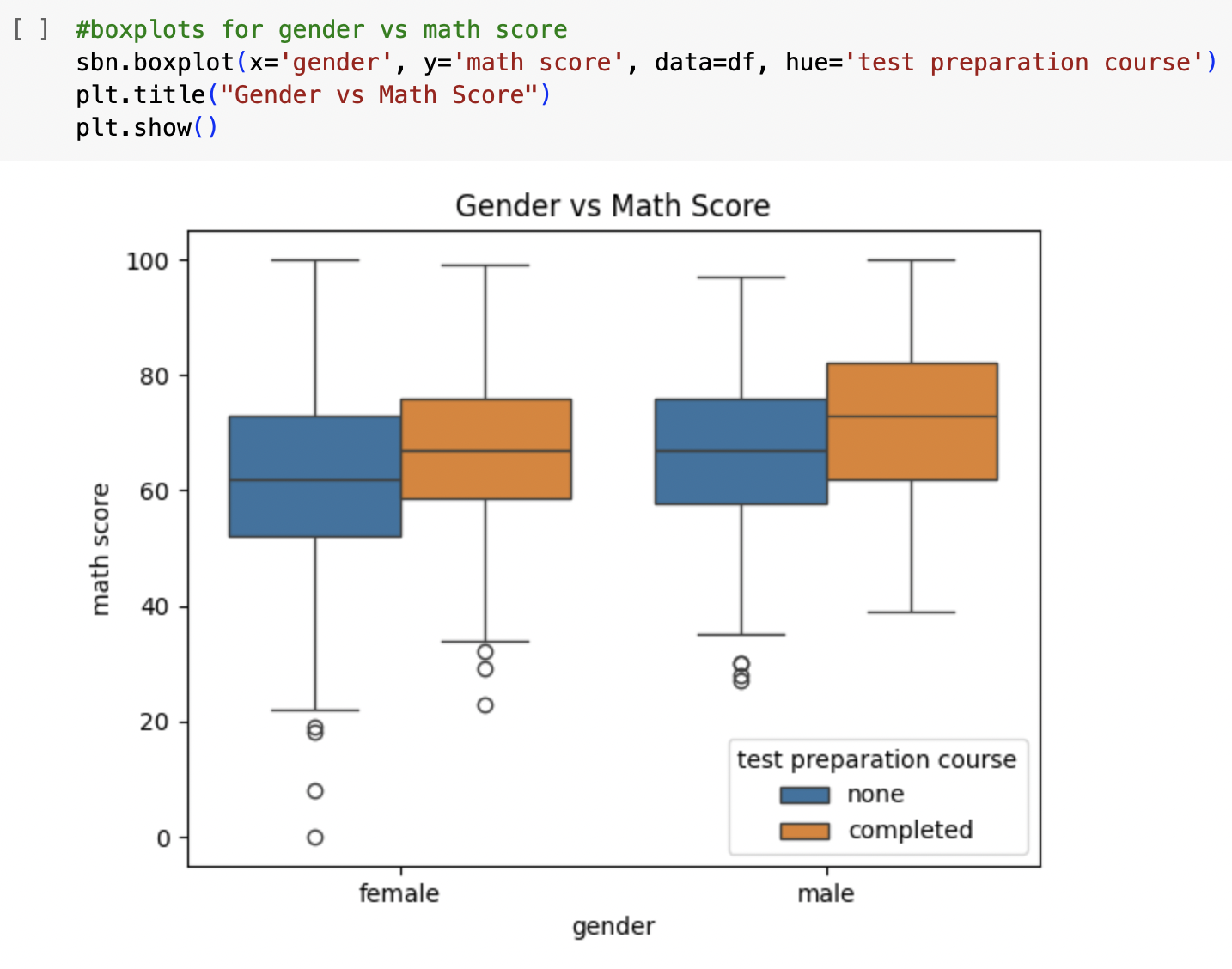
* From this plot, we can infer that among all the ethnic groups, group C is higher in number followed by group D, group B, group E and group A with the lowest. This gives us an idea of the students belonging to different ethnic groups.



* From this pie chart, we can also get the same inference as same as the one we have seen above.



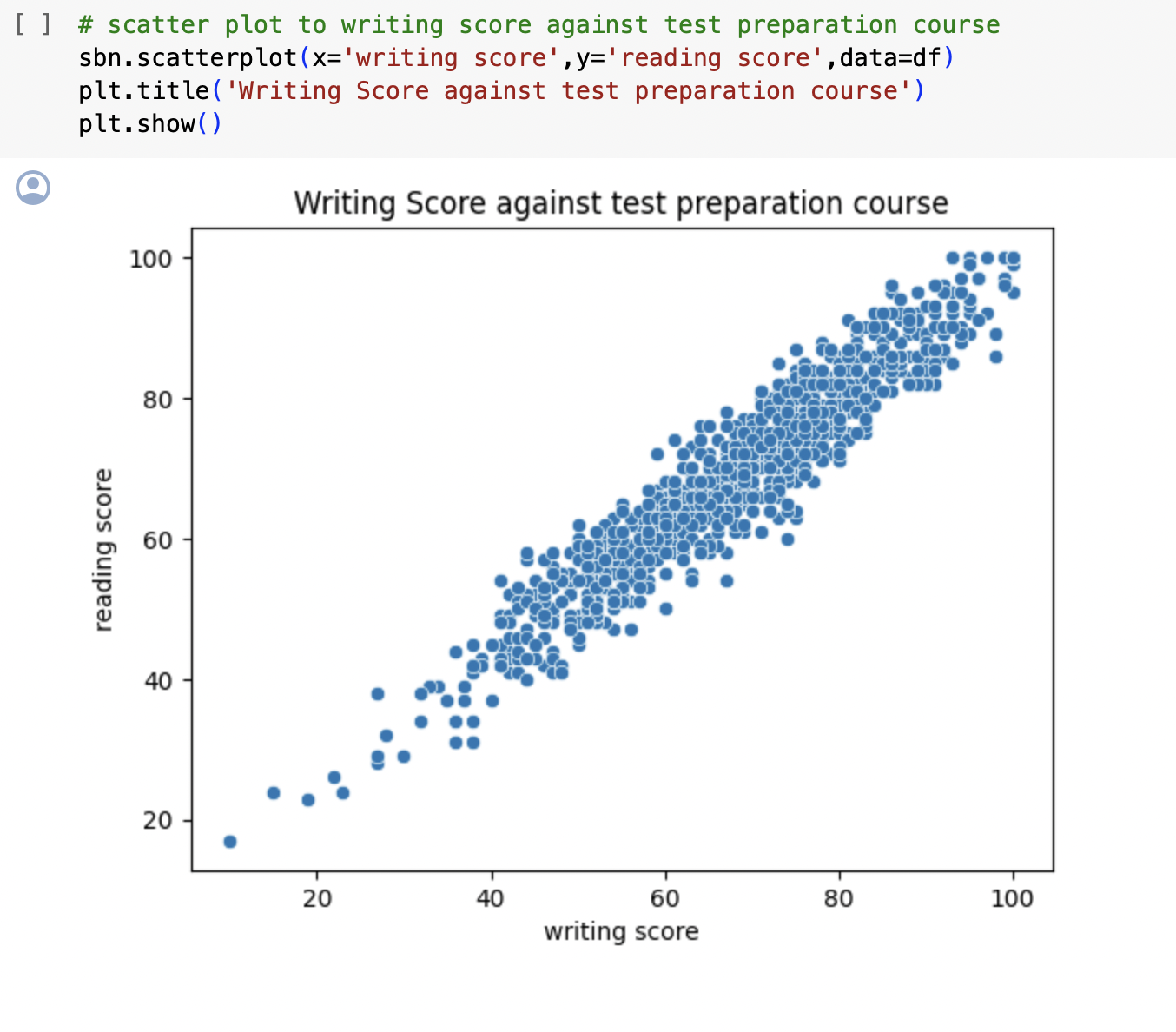
* From this pie chart, most of the parents of the students have an education level of some college or associate’s degree. THis factor might have created an impact on the students' performance in difference tests.



* While coming to the math course, students who have completed the test preparation course have scored more in females with an average score of 60-75 whereas while coming to males, this value lies around 60-80.



* In this plot, we can see that most of the students in the writing course have scored more even by not completing the course. This is an interesting observation.



* We can infer that these two scores are positively correlated as we have seen in the heat map. As writing score increases, reading score also increases and vice-versa.