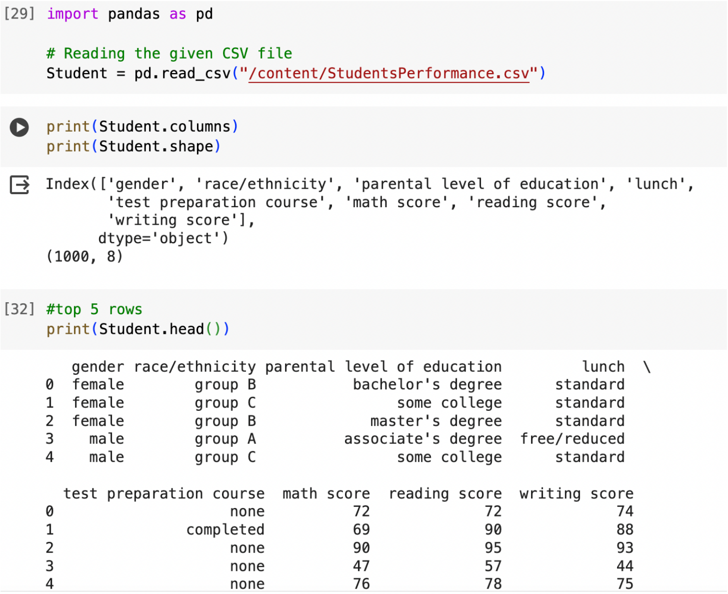
Question 2 -16345434

Data Cleaning:



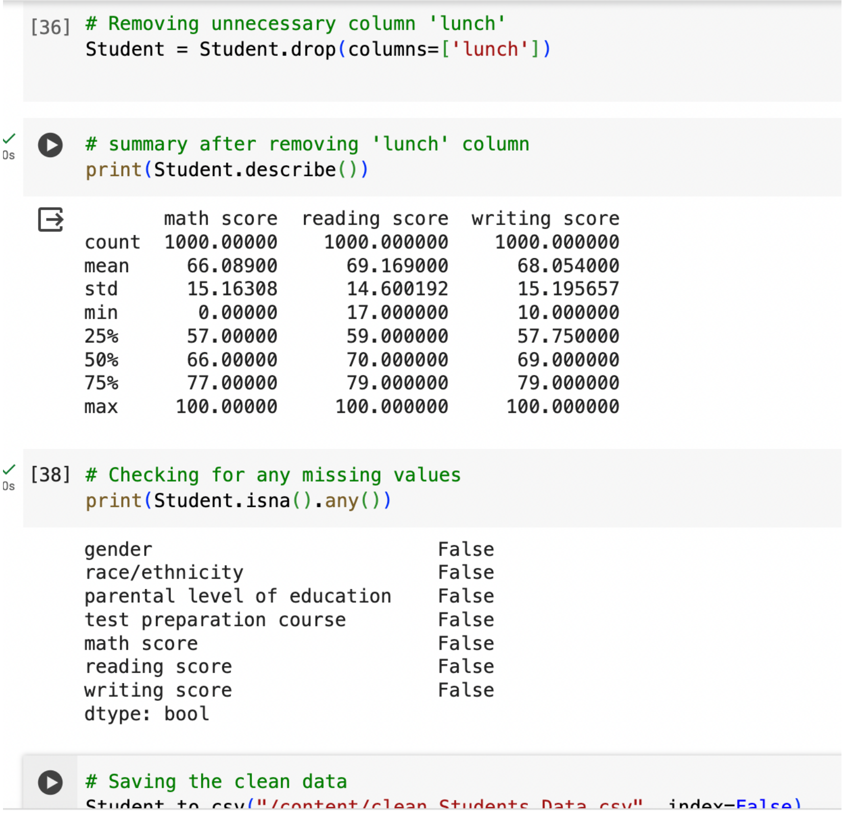
```
[29] import pandas as pd

     # Reading the given CSV file
     Student = pd.read_csv("/content/StudentsPerformance.csv")
```

```
print(Student.columns)
print(Student.shape)
```

```
Index(['gender', 'race/ethnicity', 'parental level of education', 'lunch',
       'test preparation course', 'math score', 'reading score',
       'writing score'],
      dtype='object')
(1000, 8)
```

```
[32] #top 5 rows
     print(Student.head())
```

```
   gender race/ethnicity parental level of education         lunch  \
0  female        group B           bachelor's degree      standard
1  female        group C                some college      standard
2  female        group B             master's degree      standard
3    male        group A          associate's degree  free/reduced
4    male        group C                some college      standard

  test preparation course  math score  reading score  writing score
0                    none          72             72             74
1               completed          69             90             88
2                    none          90             95             93
3                    none          47             57             44
4                    none          76             78             75
```



```
[36] # Removing unnecessary column 'lunch'
     Student = Student.drop(columns=['lunch'])
```

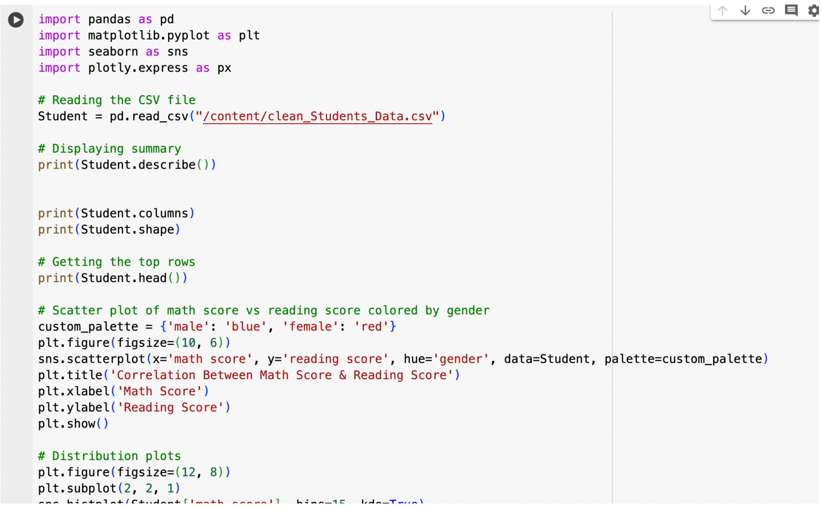
```
0s   # summary after removing 'lunch' column
     print(Student.describe())
```

```
        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
```

```
0s [38] # Checking for any missing values
        print(Student.isna().any())
```

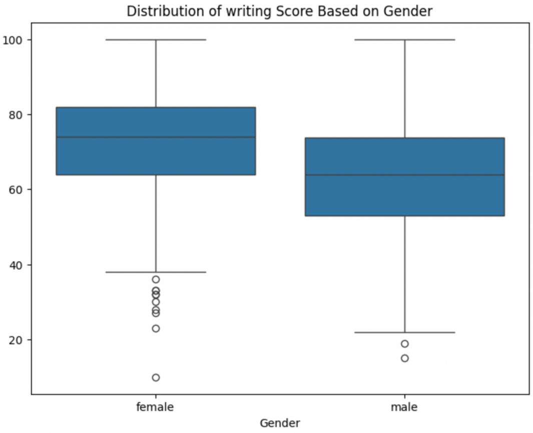
```
gender                         False
race/ethnicity                 False
parental level of education    False
test preparation course        False
math score                     False
reading score                  False
writing score                  False
dtype: bool
```

```
# Saving the clean data
Student.to_csv("/content/clean_Students_Data.csv", index=False)
```

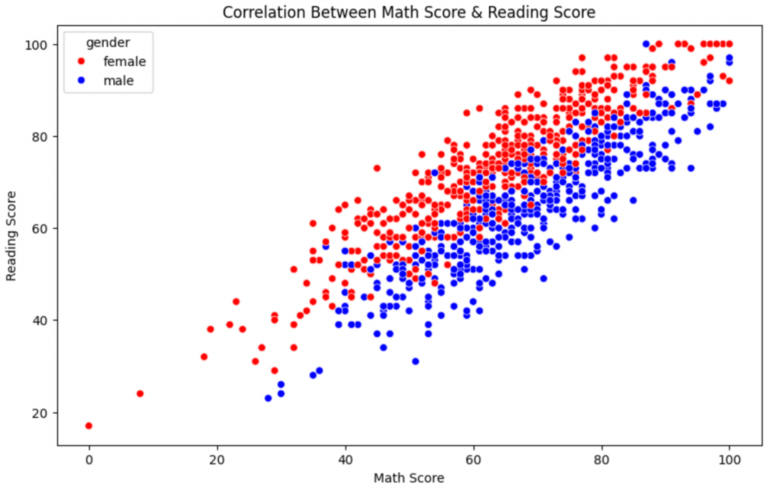


```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.express as px

# Reading the CSV file
Student = pd.read_csv("/content/clean_Students_Data.csv")

# Displaying summary
print(Student.describe())


print(Student.columns)
print(Student.shape)

# Getting the top rows
print(Student.head())

# Scatter plot of math score vs reading score colored by gender
custom_palette = {'male': 'blue', 'female': 'red'}
plt.figure(figsize=(10, 6))
sns.scatterplot(x='math score', y='reading score', hue='gender', data=Student, palette=custom_palette)
plt.title('Correlation Between Math Score & Reading Score')
plt.xlabel('Math Score')
plt.ylabel('Reading Score')
plt.show()

# Distribution plots
plt.figure(figsize=(12, 8))
plt.subplot(2, 2, 1)
sns.histplot(Student['math score'], bins=15, kde=True)
```

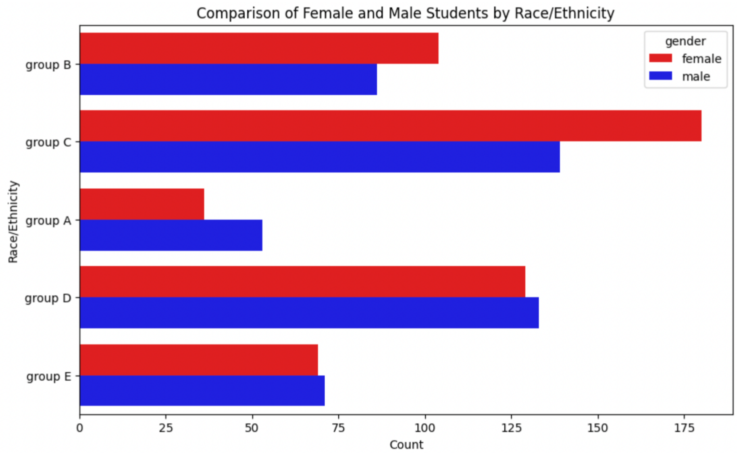
Distribution of writing score based on gender:



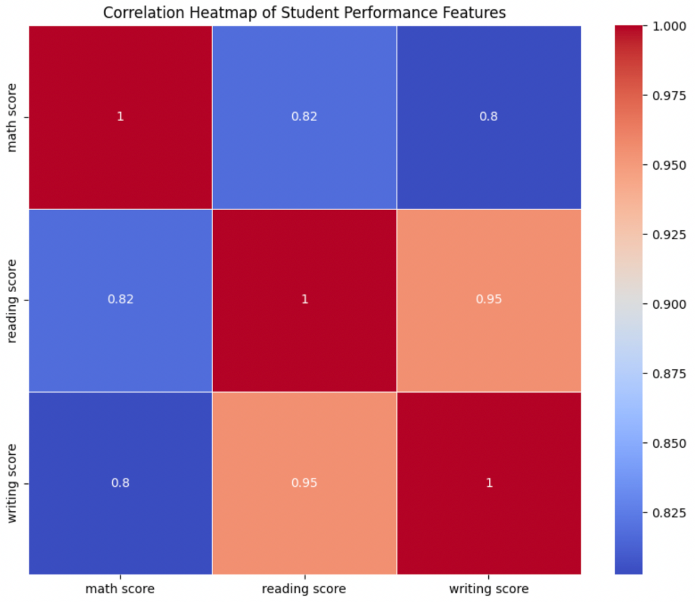
With the box plot, it becomes easier to compare the distribution of writing scores between male and female students. It facilitates the identification of any gender-based differences or similarities in writing performance, including the presence of outliers.



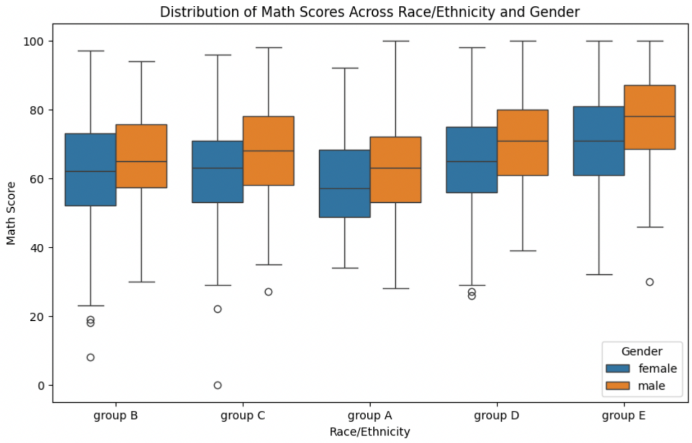
With this scatter plot, it becomes easier to visually identify any patterns or trends in the relationship between math and reading scores across different genders. It facilitates comparative analysis between male and female students in terms of their performance in math and reading.



This visualization makes it easier to compare the gender composition within each race/ethnicity group. It facilitates the identification of any gender imbalances or patterns across different racial or ethnic backgrounds.



By visually inspecting the heatmap and focusing on cells with higher absolute correlation coefficients, analysts can quickly identify which features are strongly related to each other and which are not. This aids in understanding the interdependencies and relationships between various features, guiding further analysis and modeling decisions.



This visualization makes it easier to compare the distribution of math scores across different race/ethnicity groups while also considering the gender of students within each group. It provides a clear visual representation of any variations or differences in math performance based on race/ethnicity and gender.