```
from google.colab import files
import pandas as pd

# Upload the CSV file
uploaded = files.upload()
```

Choose Files StudentsPerformance.csv

- **StudentsPerformance.csv**(text/csv) - 72036 bytes, last modified: 2/27/2023 - 100% done

```
Saving StudentsPerformance.csv to StudentsPerformance.csv
```

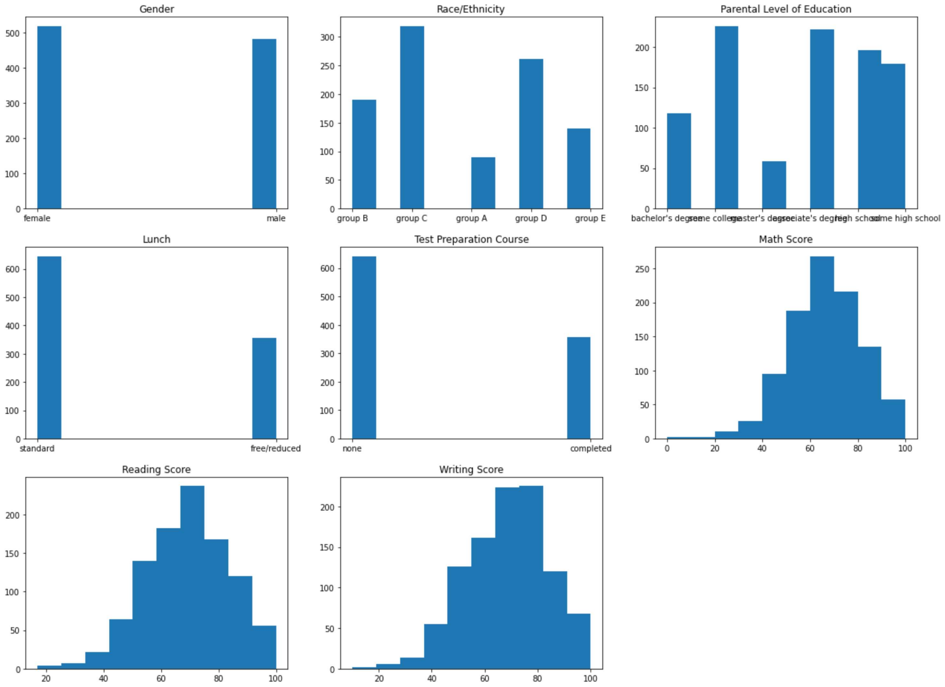
```
df = pd.read_csv('StudentsPerformance.csv')
# Print the original dataframe
#print(df.head())
num_rows = df.shape[0]

print("Number of rows in the DataFrame: ", num_rows)
# Remove rows with blank or NA values
Clean_StudentPerformance_data = df.dropna()

# Print the modified dataframe
#print(Clean_StudentPerformance_data.head())
num_rows = Clean_StudentPerformance_data.shape[0]
print("Number of rows in the DataFrame after clean: ", num_rows)
```

```
Number of rows in the DataFrame:  1000
Number of rows in the DataFrame after clean:  1000
```

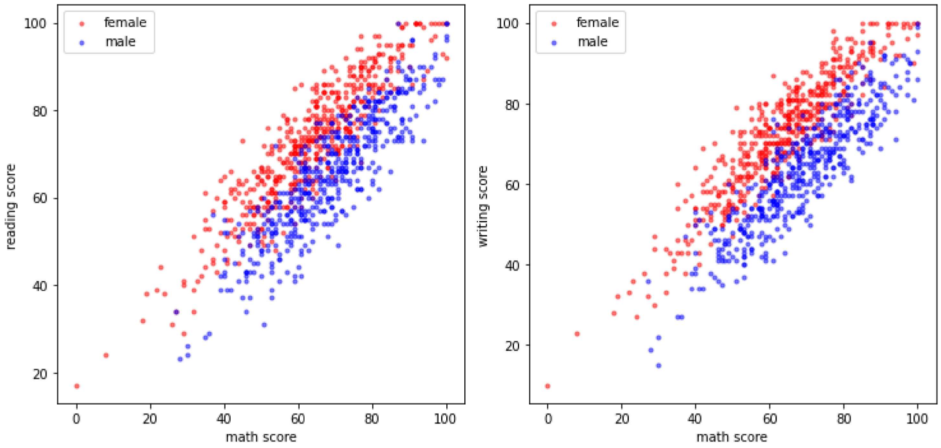
```
Clean_StudentPerformance_data.to_csv('Clean_StudentPerformance_data.csv', index=False)
```

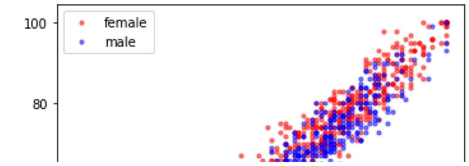
```
import matplotlib.pyplot as plt
```

```
data = Clean_StudentPerformance_data
# Plot histograms for gender, race/ethnicity, parental level of education, lunch, test preparation course, math score, reading score, writing
plt.figure(figsize=(20, 15))

plt.subplot(3, 3, 1)
plt.hist(data['gender'])
plt.title('Gender')

plt.subplot(3, 3, 2)
plt.hist(data['race/ethnicity'])
plt.title('Race/Ethnicity')

plt.subplot(3, 3, 3)
plt.hist(data['parental level of education'])
plt.title('Parental Level of Education')

plt.subplot(3, 3, 4)
plt.hist(data['lunch'])
plt.title('Lunch')

plt.subplot(3, 3, 5)
plt.hist(data['test preparation course'])
plt.title('Test Preparation Course')

plt.subplot(3, 3, 6)
plt.hist(data['math score'])
plt.title('Math Score')

plt.subplot(3, 3, 7)
plt.hist(data['reading score'])
plt.title('Reading Score')

plt.subplot(3, 3, 8)
plt.hist(data['writing score'])
plt.title('Writing Score')

plt.show()
```

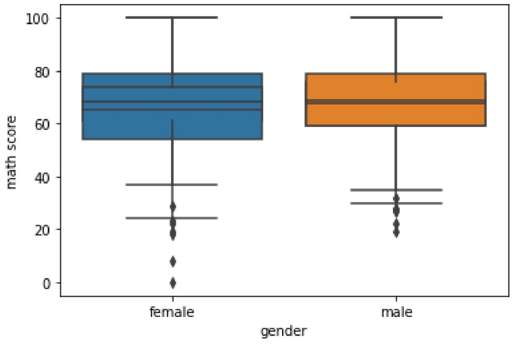


```
plt.figure(figsize=(15, 10))

vars = ['math score', 'reading score', 'writing score']
colors = ['blue', 'green', 'red']
gender_colors = {'male': 'blue', 'female': 'red'}

for i in range(len(vars)):
    for j in range(i+1, len(vars)):
        plt.subplot(2, 3, (i*2)+(j+1))
        for gender in data['gender'].unique():
            plt.scatter(data[data['gender'] == gender][vars[i]], data[data['gender'] == gender][vars[j]], s=10,
                        alpha=0.5, color=gender_colors[gender], label=gender)
        plt.xlabel(vars[i])
        plt.ylabel(vars[j])
        plt.legend()

plt.tight_layout()
plt.show()
```

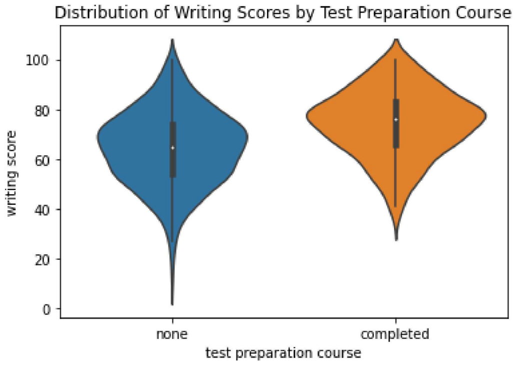




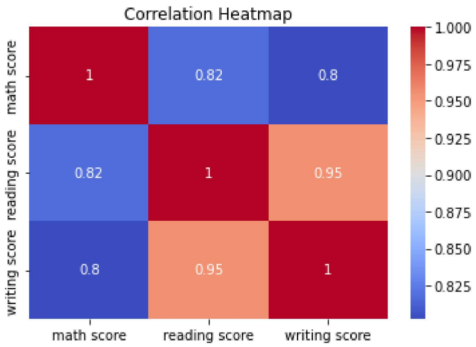
```
import seaborn as sns

sns.boxplot(x="parental level of education", y="math score", data=data)

# Box plot for gender vs math score
sns.boxplot(x="gender", y="math score", data=data)

# Show the plots
plt.show()
```



```
sns.violinplot(x="test preparation course", y="writing score", data=data)
plt.title('Distribution of Writing Scores by Test Preparation Course')
plt.show()
```



```
corr = data[['math score', 'reading score', 'writing score']].corr()
sns.heatmap(corr, annot=True, cmap='coolwarm')
plt.title('Correlation Heatmap')
plt.show()
```



✓ 0s completed at 9:05 PM