Maragathavalli C S

Data Science Intern

Prodigy Info Tech

Task:2

Perform data cleaning and exploratory data analysis(EDA) on a dataset of your choice. Explore the relationships between variables and identify patterns and trends in the data.

Importing Necessary Libraries

In [4]:
```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

Loading the Dataset

In [7]:
```
df=pd.read_csv("StudentsPerformance.csv")
```

View Basic Structure

In [11]:
```
df.head()
```

Out[11]:

	gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score
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 [13]:
```
df.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   parental level of education  1000 non-null   object
 3   lunch                        1000 non-null   object
 4   test preparation course      1000 non-null   object
 5   math score                   1000 non-null   int64 
 6   reading score                1000 non-null   int64 
 7   writing score                1000 non-null   int64 
dtypes: int64(3), object(5)
memory usage: 62.6+ KB
```

In [15]:
```
df.size
```

Out[15]:
```
8000
```

Rename columns for ease

In [20]:
```
df.columns=[col.strip().replace(" ","_").replace("/","_").lower()for col in df.columns]
```

Check and handle missing values

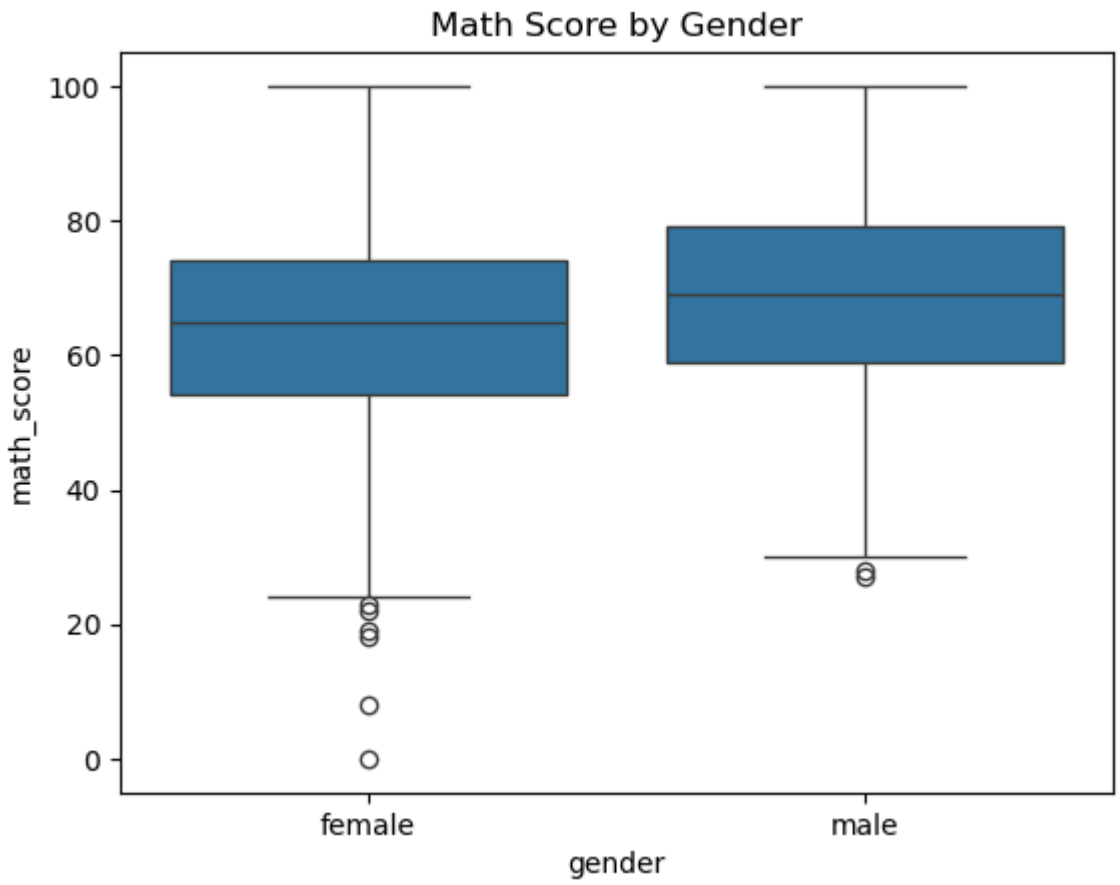
In [27]:
```
print("\nMissing Values:\n",df.isnull().sum())
```

```
Missing Values:
 gender                         0
race_ethnicity                 0
parental_level_of_education    0
lunch                          0
test_preparation_course        0
math_score                     0
reading_score                  0
writing_score                  0
dtype: int64
```

Explore with Visualizations

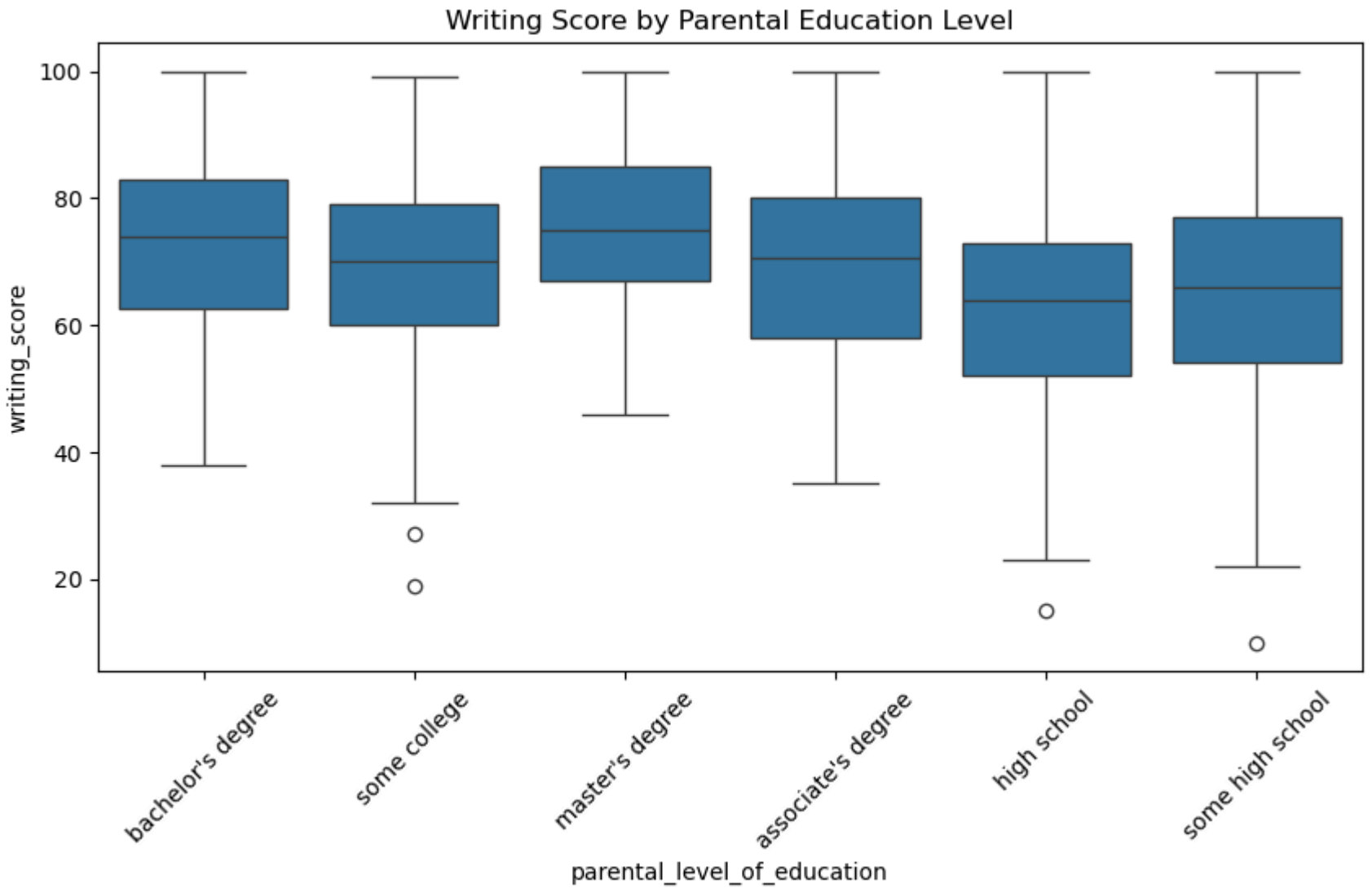
(i) Gender Vs Math Score

In [31]:
```
sns.boxplot(data=df, x='gender',y='math_score')
plt.title("Math Score by Gender")
plt.show()
```



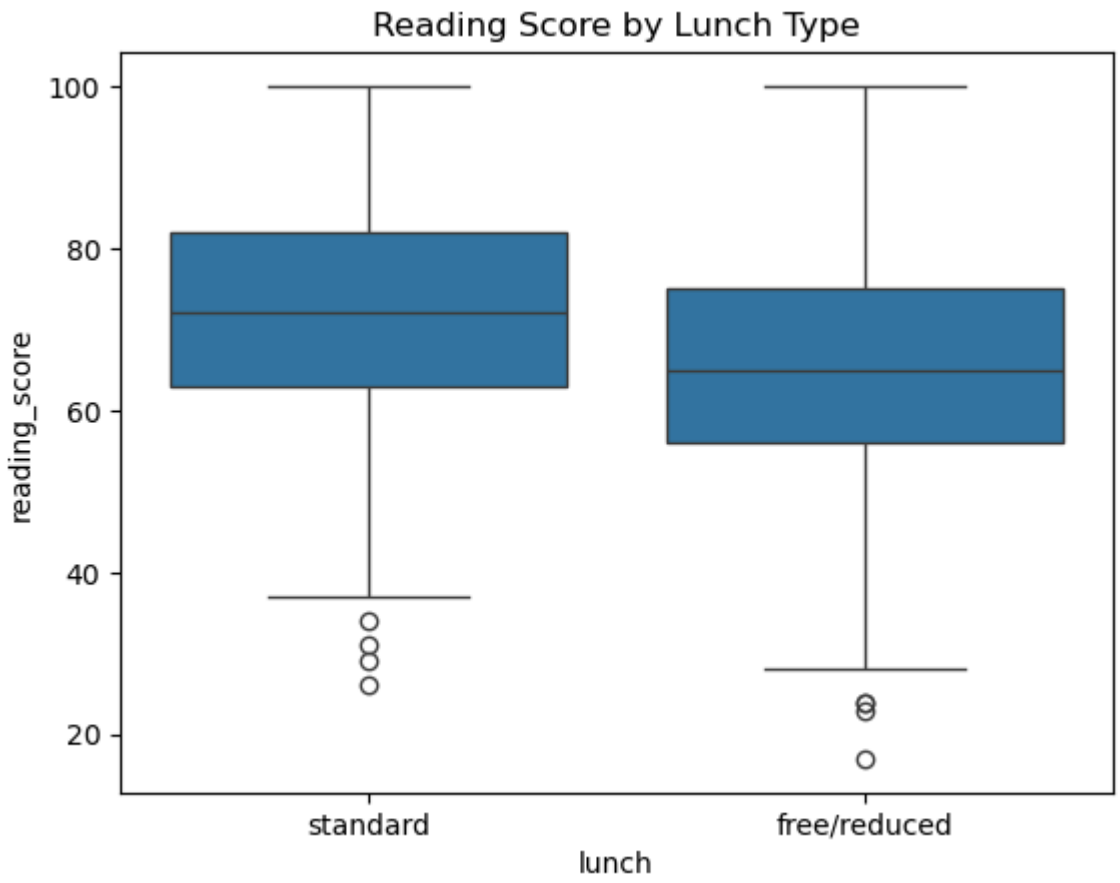
(ii) Parental Education Vs Writing Score

In [34]:
```
plt.figure(figsize=(10,5))
sns.boxplot(data=df, x='parental_level_of_education',y='writing_score')
plt.title("Writing Score by Parental Education Level")
plt.xticks(rotation=45)
plt.show()
```



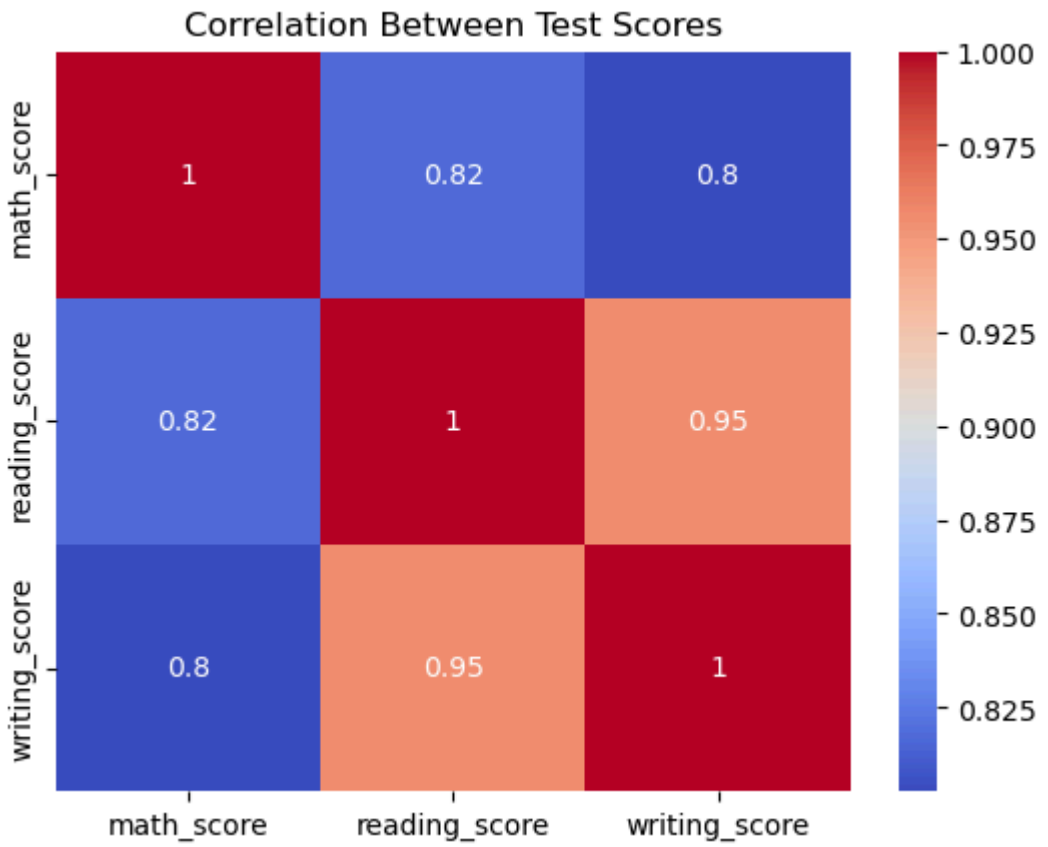
(iii) Lunch Type Vs Reading Score

In [37]:
```
sns.boxplot(data=df,x='lunch',y='reading_score')
plt.title("Reading Score by Lunch Type")
plt.show()
```



(iv) Correlation Heatmap

In [43]:
```
sns.heatmap(df[['math_score','reading_score','writing_score']].corr(),annot=True,cmap='coolwarm')
plt.title("Correlation Between Test Scores")
plt.show()
```



Add Average Score

In [46]:
```
df['average_score']=df[['math_score','reading_score','writing_score']].mean(axis=1)
```

Group analysis

In [51]:
```
grouped=df.groupby('gender')['average_score'].mean()
print("\nAverage Score by Gender:\n",grouped)
```

```
Average Score by Gender:
 gender
female    69.569498
male      65.837483
Name: average_score, dtype: float64
```

In [55]:
```
grouped2=df.groupby('test_preparation_course')['average_score'].mean()
print("\nAverage Score by Test Preparation Course:\n",grouped2)
```

```
Average Score by Test Preparation Course:
 test_preparation_course
completed    72.669460
none         65.038941
Name: average_score, dtype: float64
```

Save Results

In [58]:
```
df.to_csv("cleaned_students_data.csv",index=False)
```

Interpretation based on Students Performance Dataset

- Female students generally scored higher in reading and writing, while math scores were more balanced between genders.
- Students who completed a test preparation course scored significantly higher across all subjects.
- Those who had standard lunch outperformed those with free/reduced lunch, suggesting a possible link between nutrition and academic performance.
- Students whose parents had higher education levels tended to score better, especially in reading and writing.
- There is a strong positive correlation between scores in math, reading, and writing, indicating that students who perform well in one subject tend to perform well in others.