Step 1: Loading and cleaning the data

In [1]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import PolynomialFeatures
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error

%matplotlib inline
```

In [2]:
```
df = pd.read_csv("StudentsPerformance.csv")
df.head()
```

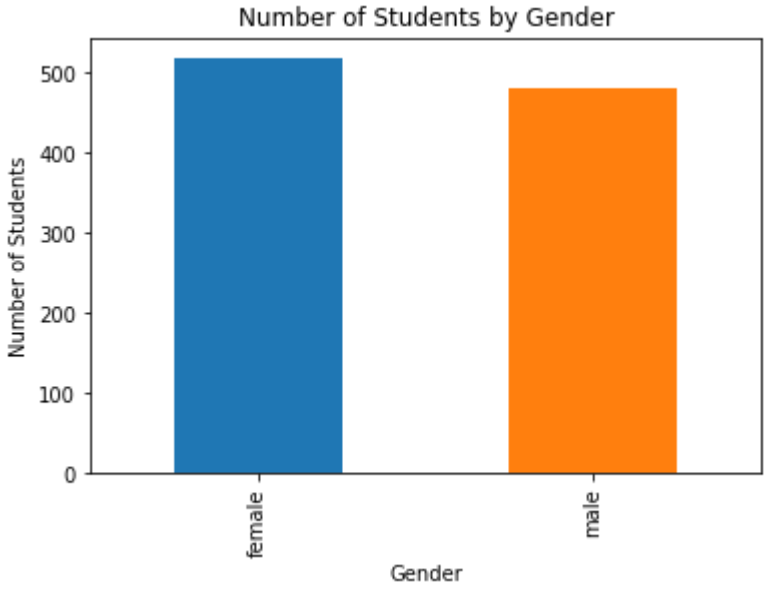
Out[2]:

	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

Step 2: Single Variable Distribution Plots

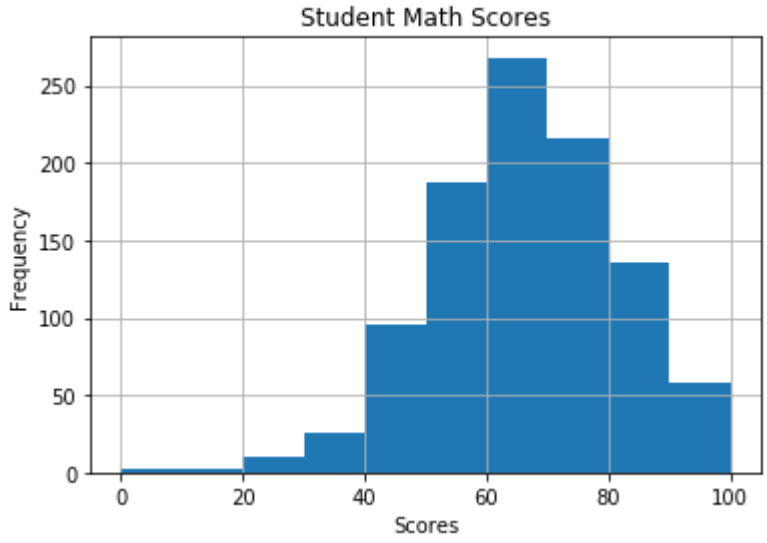
In [3]:
```
df_gender_counts = df["gender"].value_counts()
df_gender_counts.plot(kind = "bar")
plt.title("Number of Students by Gender")
plt.xlabel("Gender")
plt.ylabel('Number of Students')
```

Out[3]: Text(0,0.5,'Number of Students')



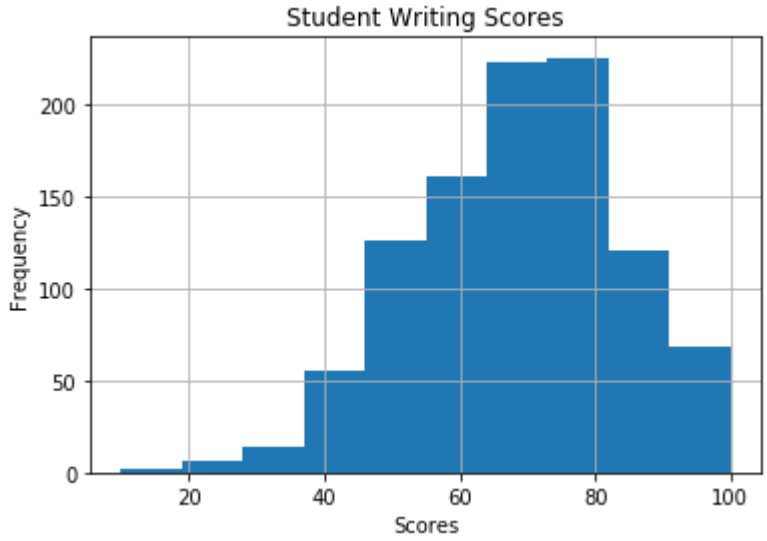
In [4]:
```
df["math score"].hist()
plt.title("Student Math Scores")
plt.xlabel("Scores")
plt.ylabel("Frequency")
```

Out[4]: Text(0,0.5,'Frequency')



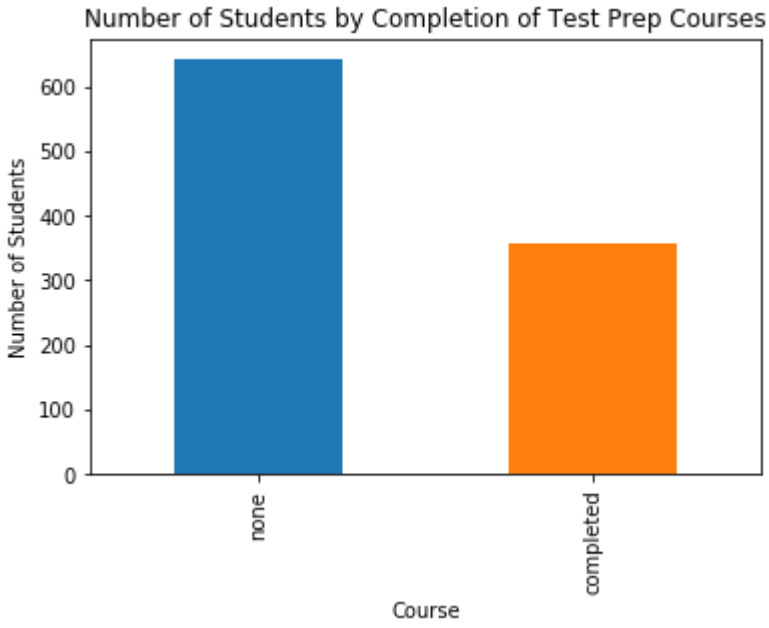
In [5]:
```
df["writing score"].hist()
plt.title("Student Writing Scores")
plt.xlabel("Scores")
plt.ylabel("Frequency")
```

Out[5]: Text(0,0.5,'Frequency')



In [6]:
```
prep_counts = df["test preparation course"].value_counts()
prep_counts.plot(kind = "bar")
plt.title("Number of Students by Completion of Test Prep Courses")
plt.xlabel("Course")
plt.ylabel('Number of Students')
```

Out[6]: Text(0,0.5,'Number of Students')



Step 3: Multiple Variable Plots

In [7]:
```
df.head()
```

Out[7]:

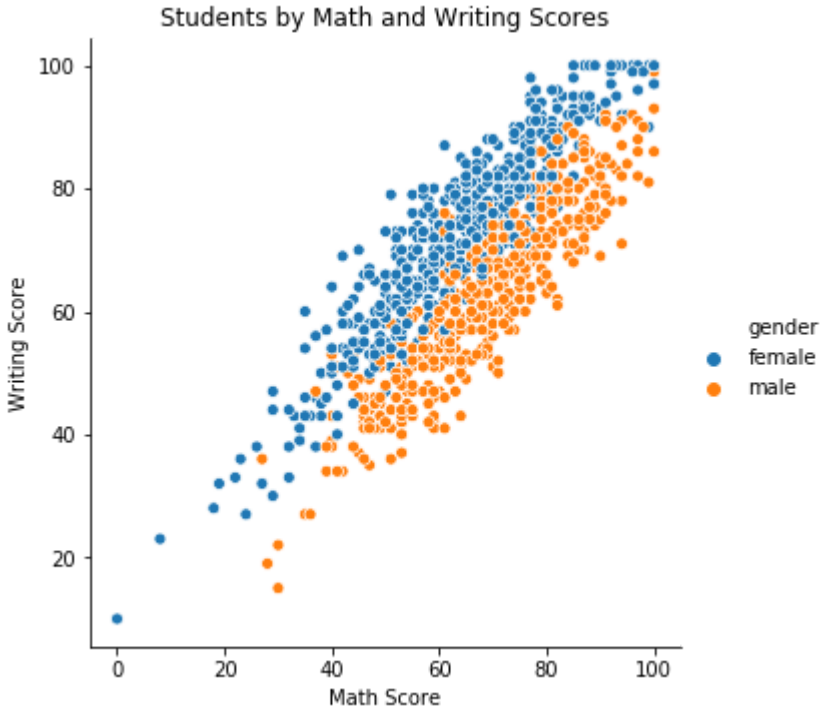
	gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score
0	female	group B	bachelor's degree	standard	none	72	72	74
1	female	group C	some college	standard	completed	69	90	88
2	female	group B	master's degree	standard	none	90	95	93
3	male	group A	associate's degree	free/reduced	none	47	57	44
4	male	group C	some college	standard	none	76	78	75

In [8]:
```
df.dtypes
```

Out[8]:
```
gender                         object
race/ethnicity                 object
parental level of education    object
lunch                          object
test preparation course        object
math score                      int64
reading score                   int64
writing score                   int64
dtype: object
```

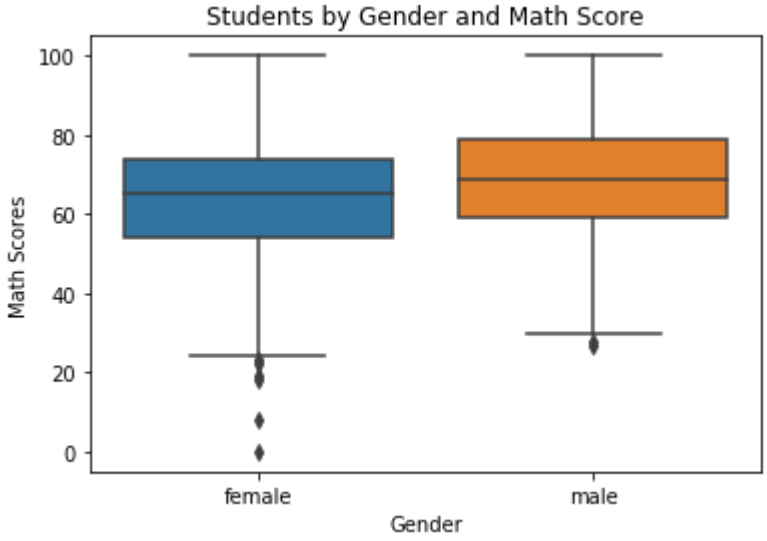
In [9]:
```
sns.relplot(x = "math score", y = "writing score", hue = "gender", data = df)
plt.title("Students by Math and Writing Scores")
plt.xlabel("Math Score")
plt.ylabel('Writing Score')
```

Out[9]: Text(30.8646,0.5,'Writing Score')



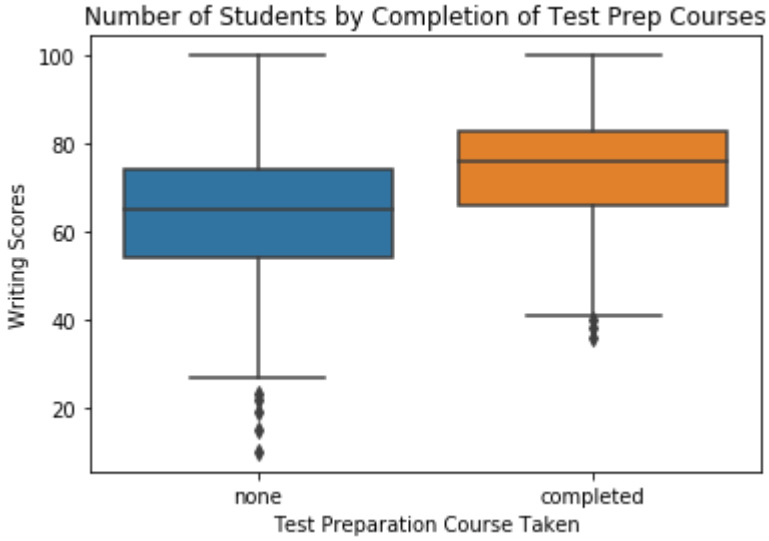
In [10]:
```
sns.boxplot(x = "gender", y = "math score", data = df)
plt.title("Students by Gender and Math Score")
plt.xlabel("Gender")
plt.ylabel('Math Scores')
```

Out[10]: Text(0,0.5,'Math Scores')



In [11]:
```
sns.boxplot(x = "test preparation course", y = "writing score", data = df)
plt.title("Number of Students by Completion of Test Prep Courses")
plt.xlabel("Test Preparation Course Taken")
plt.ylabel('Writing Scores')
```

Out[11]: Text(0,0.5,'Writing Scores')



In [12]:
```
sns.pairplot(data = df)
```

Out[12]: <seaborn.axisgrid.PairGrid at 0x7f4c9e126f60>