Assignment5

March 21, 2025

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

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

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

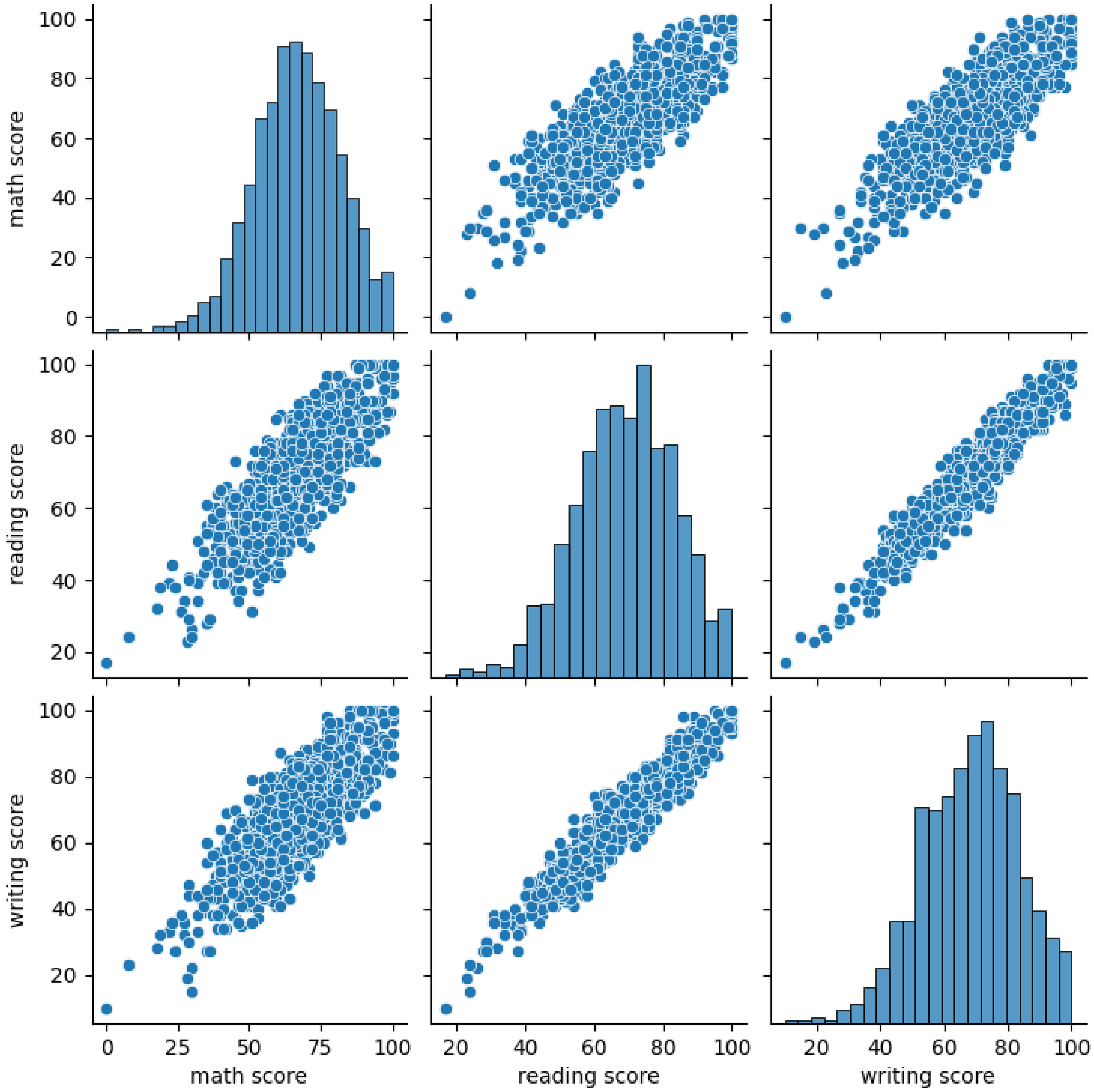
[ ]:
```
   gender race/ethnicity parental level of education         lunch  \
0  female        group B           bachelor's degree      standard
1  female        group C                some college      standard
2  female        group B             master's degree      standard
3    male        group 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
```

1. Perform multivariate analysis: • Identify patterns using techniques such as pair plots and matrix plots.

[ ]:
```
df_numeric = df[['math score', 'reading score', 'writing score']]
```

[ ]:
```
sns.pairplot(df_numeric)
plt.show()
```



2. Identify and summarize key insights from the dataset.

```
[ ]: df_numeric.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
```

3. Compute the correlation matrix for numerical attributes using: • Pearson correlation • Spearman correlation

```
[ ]: pearson_corr = df_numeric.corr(method='pearson')
     spearman_corr = df_numeric.corr(method='spearman')
```

```
[ ]: pearson_corr
```

```
[ ]:                math score  reading score  writing score
     math score       1.000000       0.817580       0.802642
     reading score    0.817580       1.000000       0.954598
     writing score    0.802642       0.954598       1.000000
```

1. Math & Reading Scores: High positive correlation (0.82–0.85) → Students who score well in Math tend to perform well in Reading.
2. The highest covariance is between Reading & Writing, confirming they are closely related.

```
[ ]: spearman_corr
```

```
[ ]:                math score  reading score  writing score
     math score       1.000000       0.804064       0.778339
     reading score    0.804064       1.000000       0.948953
     writing score    0.778339       0.948953       1.000000
```

1. Similar results but Spearman measures rank-based relationships, meaning it captures non-linear trends as well.
2. Reading & Writing have the strongest monotonic relationship, meaning higher reading scores always tend to imply higher writing scores.

4. Compute covariance for pairs of numerical attributes

```
[ ]: cov_matrix = df_numeric.cov()
```

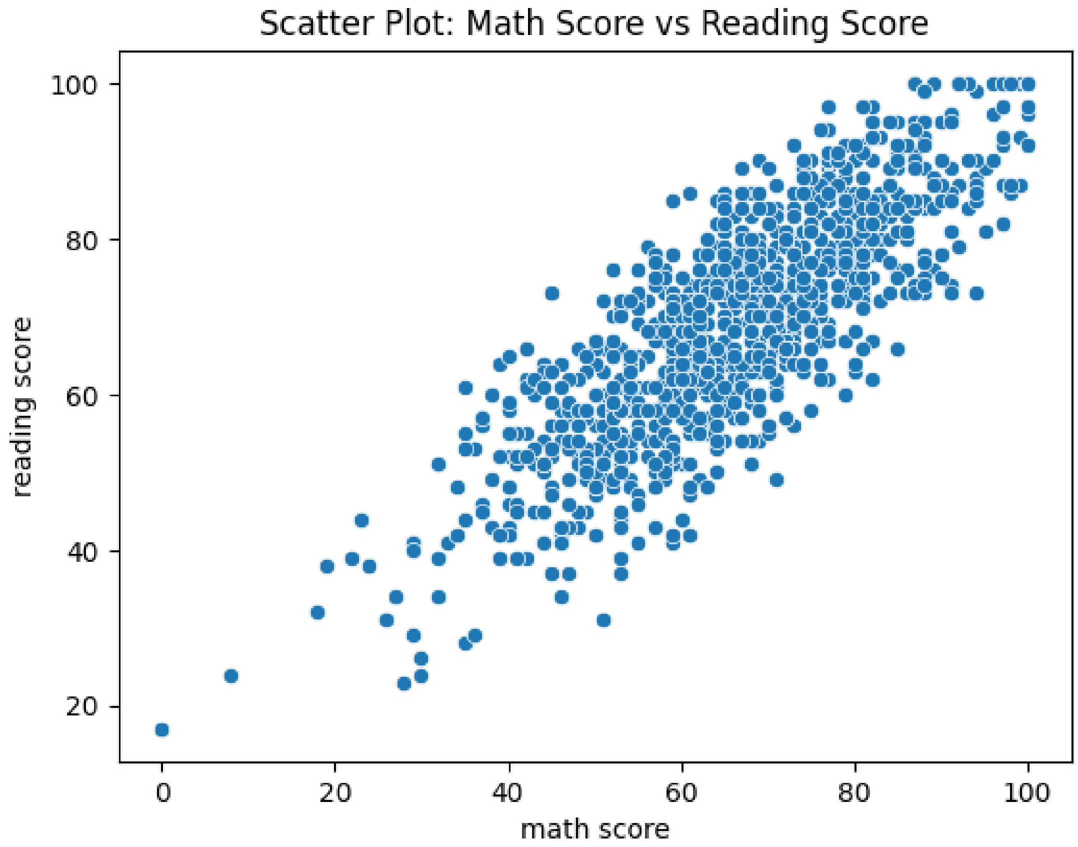
```
[ ]: cov_matrix
```

```
[ ]:                math score  reading score  writing score
     math score     229.918998     180.998958     184.939133
     reading score  180.998958     213.165605     211.786661
     writing score  184.939133     211.786661     230.907992
```

1. Positive covariance between all attributes → When one score increases, the other tends to increase as well.
2. The highest covariance is between Reading & Writing, confirming they are closely related.
3. Math has a slightly lower covariance with Writing, meaning writing performance may depend on more than just numerical skills.

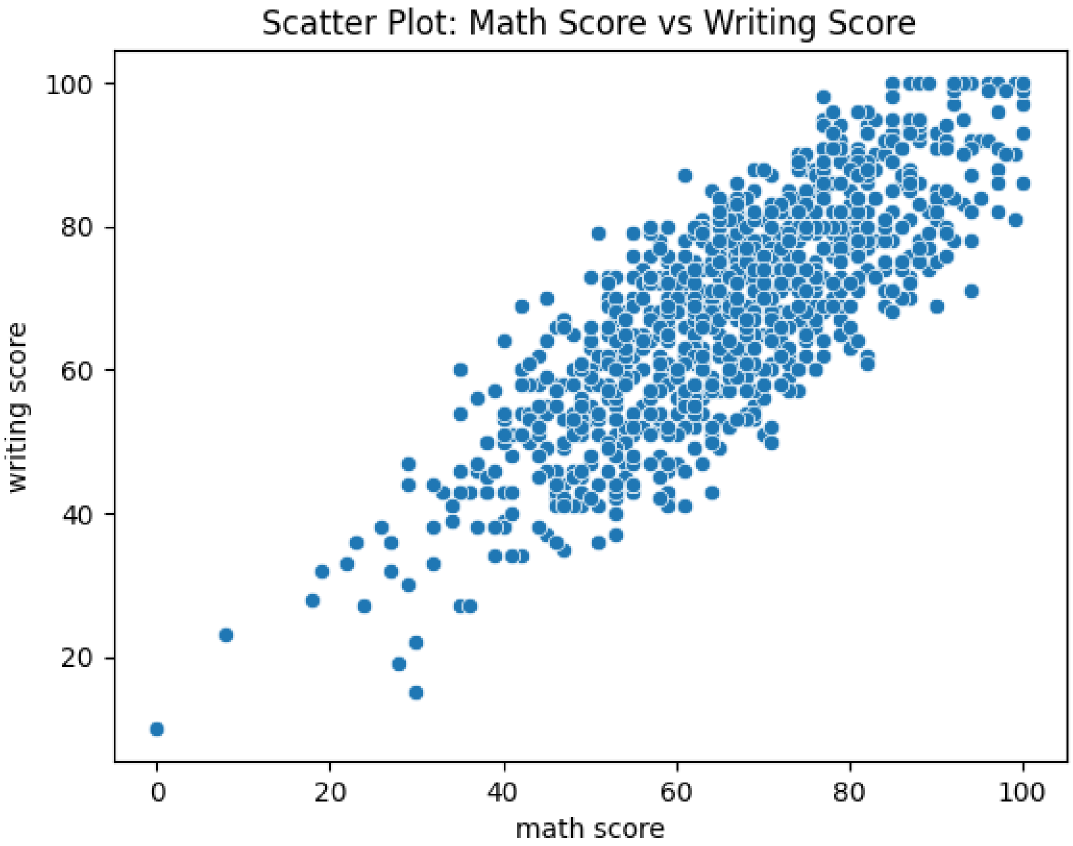
5. Visualize correlations using: Explore relationships between variables using • scatter plots • correlation plots • Heatmaps

[ ]:
```
sns.scatterplot(x='math score', y='reading score', data=df_numeric)
plt.title("Scatter Plot: Math Score vs Reading Score")
plt.show()
```



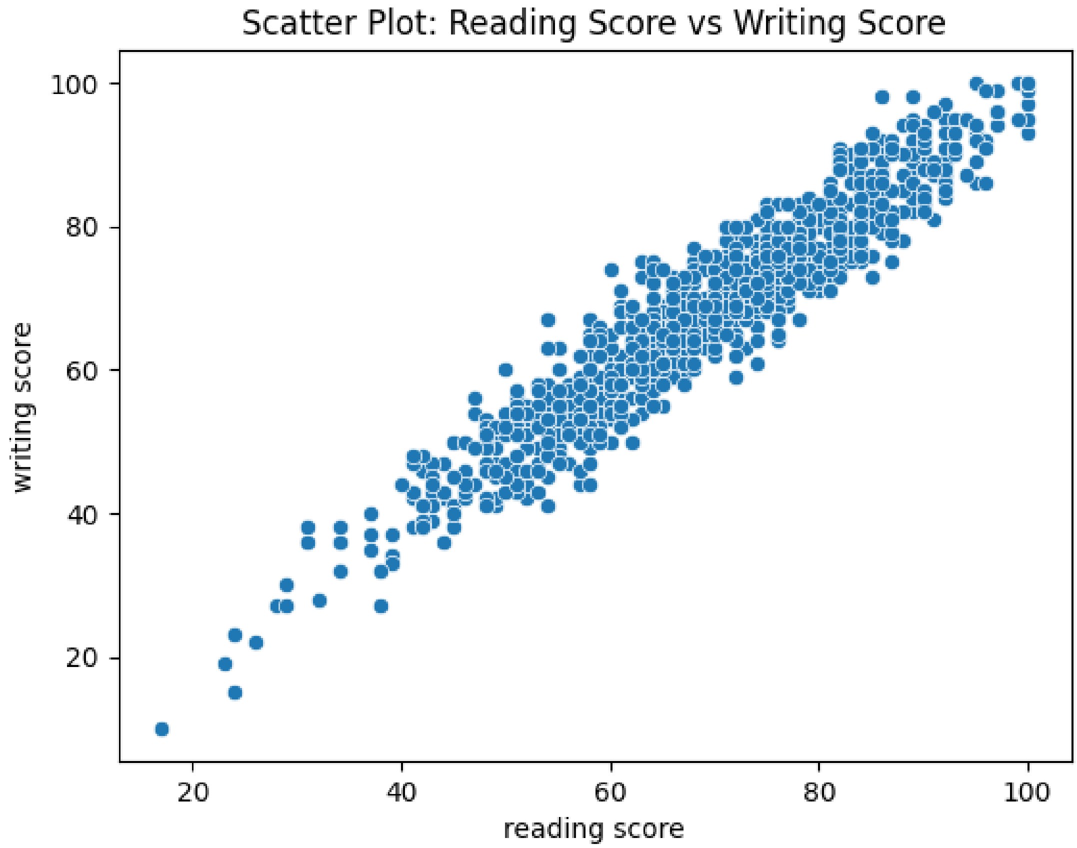
Some points deviate, suggesting that a few students excel in Math but not as much in Reading.

[ ]:
```
sns.scatterplot(x='math score', y='writing score', data=df_numeric)
plt.title("Scatter Plot: Math Score vs Writing Score")
plt.show()
```



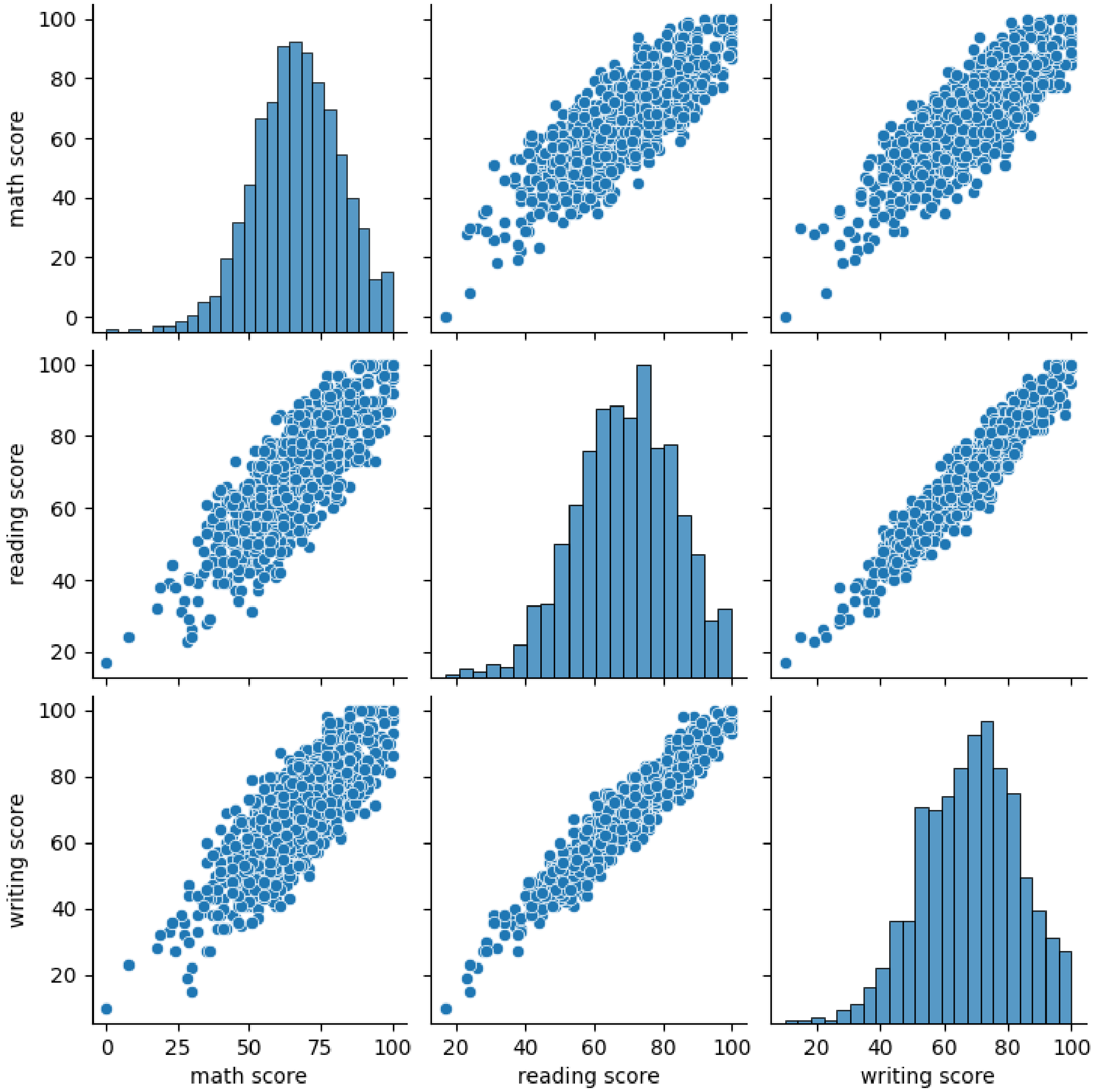
The trend is positive, but the correlation is weaker than Math & Reading.

```
[ ]: sns.scatterplot(x='reading score', y='writing score', data=df_numeric)
     plt.title("Scatter Plot: Reading Score vs Writing Score")
     plt.show()
```



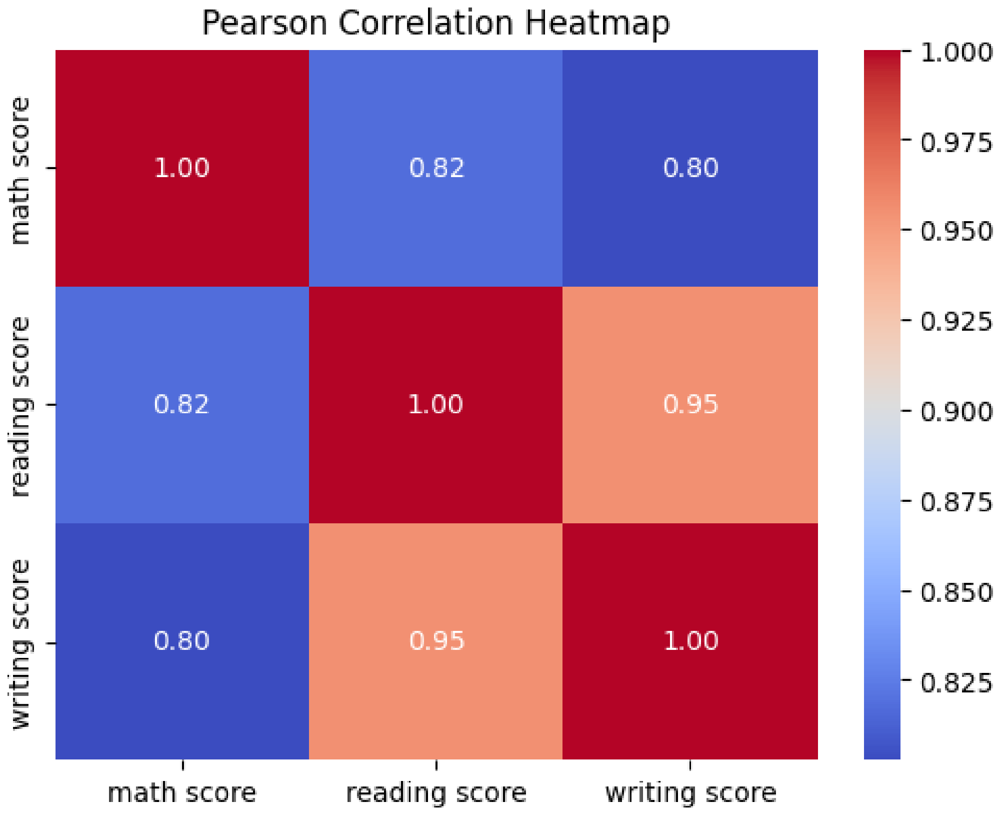
Almost a perfect linear relationship, meaning students who are good at Reading are almost always good at Writing.

[ ]:
```
sns.pairplot(df_numeric)
plt.show()
```



[ ]:

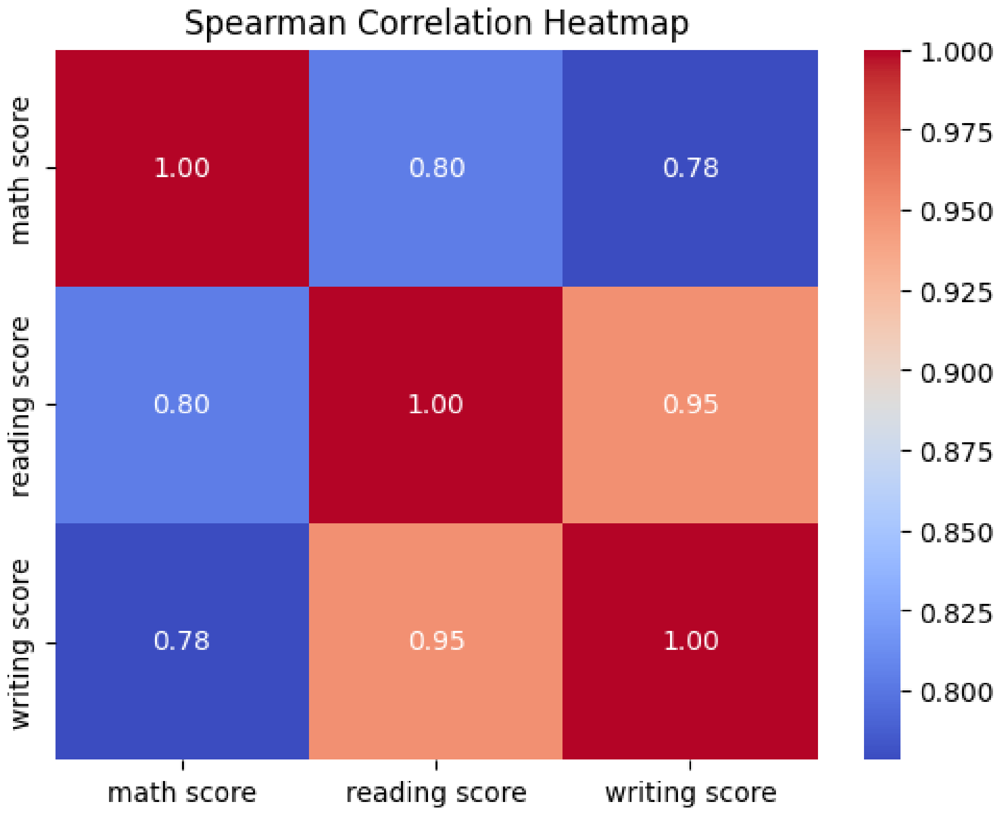
```
sns.heatmap(pearson_corr, annot=True, cmap="coolwarm", fmt=".2f")
plt.title("Pearson Correlation Heatmap")
plt.show()
```



All correlations are above 0.80, indicating strong relationships between all three subjects.

The heatmap confirms that Reading and Writing have the highest correlation (~0.9), meaning they are almost dependent on each other.

[ ]:
```
sns.heatmap(spearman_corr, annot=True, cmap="coolwarm", fmt=".2f")
plt.title("Spearman Correlation Heatmap")
plt.show()
```



Reading and writing scores (0.95) are highly correlated, indicating a strong relationship. Math has a slightly lower correlation with reading (0.82) and writing (0.80) , improving reading skills could significantly impact writing performance, while math is somewhat independent but linked to overall academic performance.