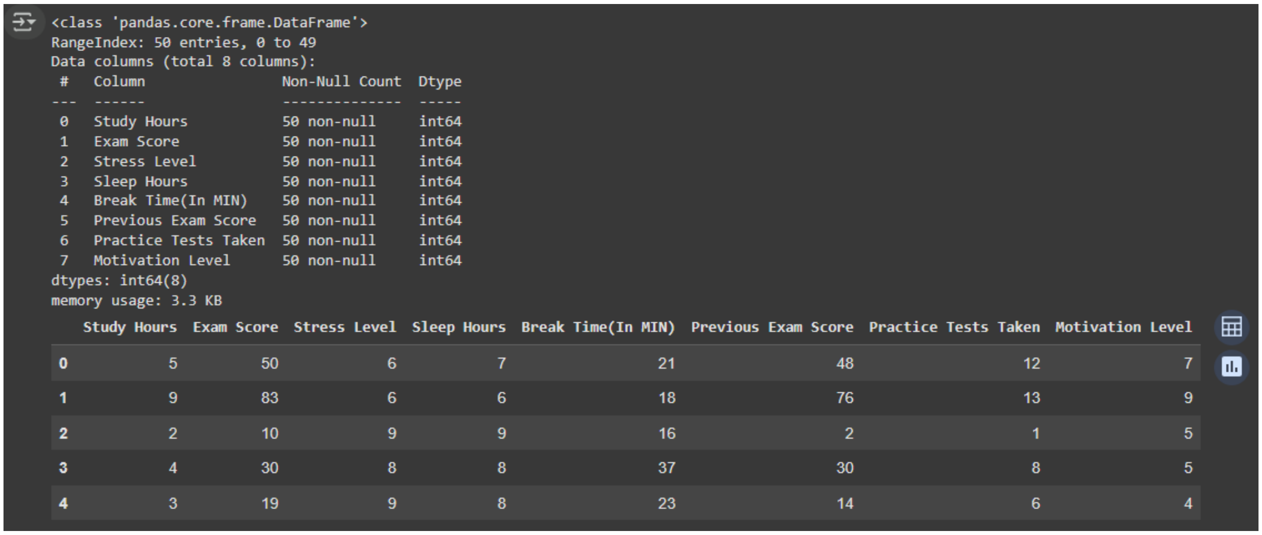
Experiment 4

Aim: Implementation of Statistical Hypothesis Test using Scipy and Sci-kit learn.
Perform the following correlation tests: a)
Pearson's Correlation Coefficient
b) Spearman's Rank Correlation
c) Kendall's Rank Correlation
d) Chi-Squared Test

Performance:

- Prerequisite: We import necessary libraries such as pandas for data manipulation, numpy for numerical operations, scipy.stats for statistical calculations, and seaborn and matplotlib.pyplot for visualization and load data into Pandas. To understand the dataset structure, we print its basic information using df.info() to check column types (numerical or categorical) and df.head() to preview the first few rows:

Command: import pandas as pd
import numpy as np
import pandas as pd
import scipy.stats as stats
import seaborn as sns
import matplotlib.pyplot as plt
df = pd.read_csv('set2.csv')
df.info()
df.head()



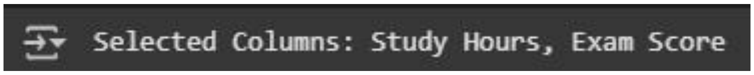
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 50 entries, 0 to 49
Data columns (total 8 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   Study Hours          50 non-null     int64
 1   Exam Score           50 non-null     int64
 2   Stress Level         50 non-null     int64
 3   Sleep Hours          50 non-null     int64
 4   Break Time(In MIN)   50 non-null     int64
 5   Previous Exam Score  50 non-null     int64
 6   Practice Tests Taken 50 non-null     int64
 7   Motivation Level     50 non-null     int64
dtypes: int64(8)
memory usage: 3.3 KB
```

	Study Hours	Exam Score	Stress Level	Sleep Hours	Break Time(In MIN)	Previous Exam Score	Practice Tests Taken	Motivation Level
0	5	50	6	7	21	48	12	7
1	9	83	6	6	18	76	13	9
2	2	10	9	9	16	2	1	5
3	4	30	8	8	37	30	8	5
4	3	19	9	8	23	14	6	4

To test correlations between features, we pick two numerical columns (col1 and col2) for Pearson, Spearman, or Kendall tests. We first check if these columns exist in the dataset to avoid errors. This makes sure we use the right variables for analysis.

Command:

```
col1 = 'Study Hours'
col2 = 'Exam Score'
if col1 not in df.columns or col2 not in df.columns:
   raise ValueError("One or both selected columns do not exist in the dataset!")
print(f"Selected Columns: {col1}, {col2}")
```

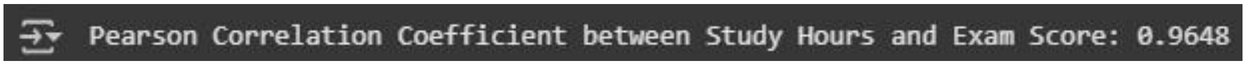


```
Selected Columns: Study Hours, Exam Score
```

a) Pearson's Correlation Coefficient:

Command:

```
pearson_corr, _ = stats.pearsonr(df[col1], df[col2])
print(f"Pearson Correlation Coefficient between {col1} and {col2}: {pearson_corr:.4f}")
```



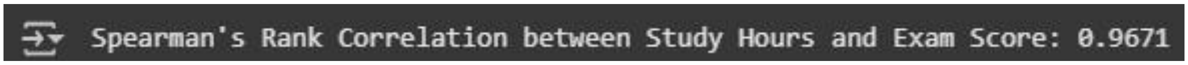
```
Pearson Correlation Coefficient between Study Hours and Exam Score: 0.9648
```

Pearson's correlation checks the linear relationship between two continuous variables. We calculate it using stats.pearsonr(df[col1], df[col2]), which gives a value between -1 (perfect negative) and +1 (perfect positive), with 0 meaning no correlation. A Pearson correlation of 0.9648 shows a strong positive link between Study Hours and Exam Score. This means more study hours usually lead to higher exam scores, following a nearly perfect straight-line trend.

b) Spearman's Rank Correlation:

Command:

```
spearman_corr, _ = stats.spearmanr(df[col1], df[col2])
print(f"Spearman's Rank Correlation between {col1} and {col2}: {spearman_corr:.4f}")
```



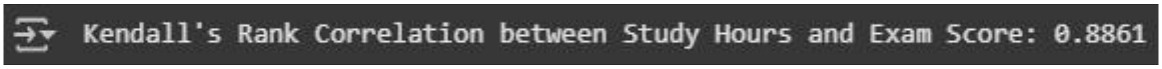
```
Spearman's Rank Correlation between Study Hours and Exam Score: 0.9671
```

Spearman's correlation measures the monotonic relationship between two variables, making it better for outliers and non-linear data than Pearson's. It ranks the values first and then calculates the correlation using `stats.spearmanr(df[col1], df[col2])`. Like Pearson's, it ranges from -1 to +1, with higher absolute values meaning stronger relationships. It's useful for non-normal data. A Spearman correlation of 0.9671 shows a strong monotonic link between Study Hours and Exam Score. Even if the relationship isn't perfectly linear, students who study more tend to rank higher in exam performance.

c) Kendall's Rank Correlation:

Command:

```
kendall_corr, _ = stats.kendalltau(df[col1], df[col2])
print(f"Kendall's Rank Correlation between {col1} and {col2}: {kendall_corr:.4f}")
```



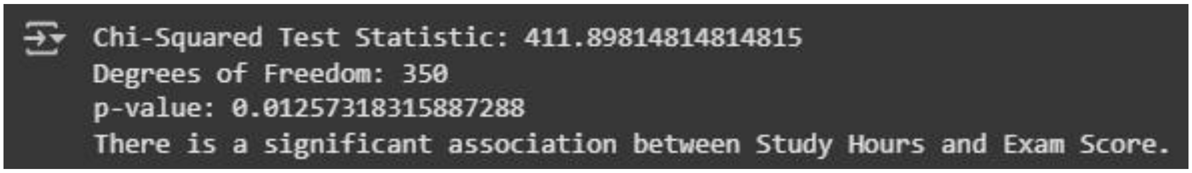
```
Kendall's Rank Correlation between Study Hours and Exam Score: 0.8861
```

Kendall’s correlation measures the strength of association between two variables using ranked data. It’s calculated with `stats.kendalltau(df[col1], df[col2])` and is ideal for small datasets or ordinal data. Like Pearson and Spearman, it ranges from -1 to +1, with values near ±1 showing stronger relationships. A Kendall correlation of 0.8861 indicates a strong agreement between Study Hours and Exam Score rankings. Students who study more consistently rank higher in exam performance, highlighting the predictability of scores based on study time.

d) Chi-Squared Test:

Command:

```
contingency_table = pd.crosstab(df[col1], df[col2])
chi2_stat, p_val, dof, expected = stats.chi2_contingency(contingency_table)
print(f"Chi-Squared Test Statistic: {chi2_stat}")
print(f"Degrees of Freedom: {dof}")
print(f"p-value: {p_val}") if p_val < 0.05:
    print(f"There is a significant association between {col1} and {col2}.")
else:
    print(f"There is NO significant association between {col1} and
{col2}.")
```



```
Chi-Squared Test Statistic: 411.89814814814815
Degrees of Freedom: 350
p-value: 0.01257318315887288
There is a significant association between Study Hours and Exam Score.
```

The Chi-Squared test checks if two categorical variables are related. It uses a contingency table created with pd.crosstab(df[col1], df[col2]) and calculates the test statistic, degrees of freedom, and p-value with stats.chi2_contingency(contingency_table). If the p-value is below 0.05, the variables are significantly associated; otherwise, they are independent. This test is helpful for analyzing grouped or categorical data. Chi-Squared test found a significant link ($p < 0.05$) between Study Hours and Exam Score, meaning study time likely affects exam performance.

Conclusion:

In this experiment, we learned to implement Statistical Hypothesis Tests using Scipy and Sci-kit learn. The Pearson correlation (0.9648) showed a strong positive linear relationship between Study Hours and Exam Score, meaning more study hours lead to higher scores. Spearman’s correlation (0.9671) indicated that students who study more tend to rank higher in exam performance, even if the relationship isn’t perfectly linear. Kendall’s correlation (0.8861) confirmed a strong agreement in rankings, reinforcing that more study time predicts better exam results. The Chi-Squared test ($\chi^2 = 411.90$, $p = 0.0126$) proved a significant association between Study Hours and Exam Score, highlighting the importance of study time in influencing performance. Overall, all tests confirmed a strong positive relationship, suggesting that increasing study hours is likely to improve exam scores.