

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:

```
import pandas as pd
```

```
import numpy as np
```

```
import pandas as pd
```

```
import scipy.stats as stats
```

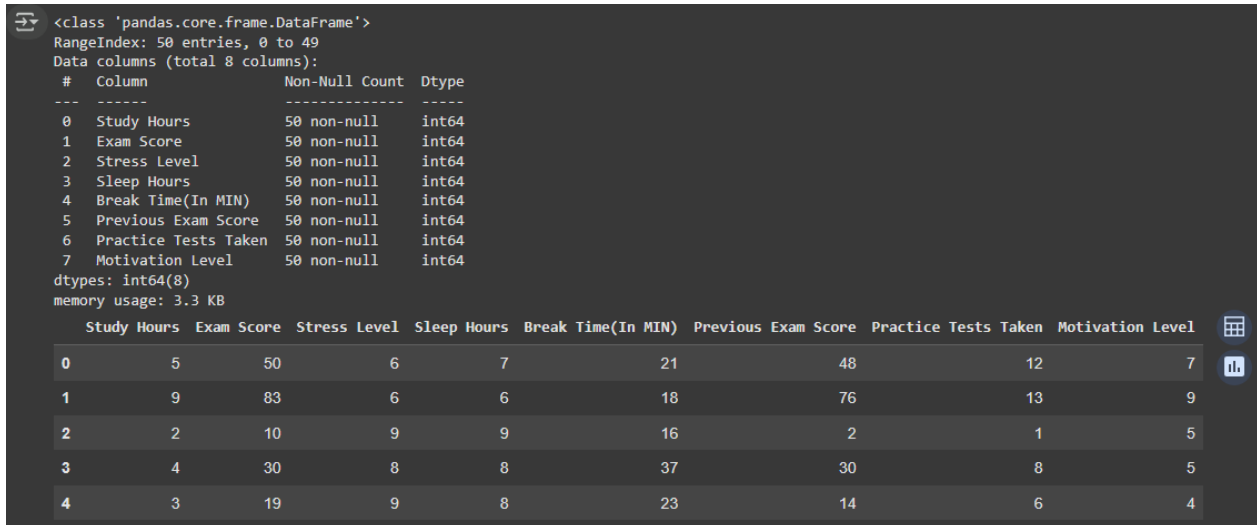
```
import seaborn as sns
```

```
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 for Pearson, Spearman, or Kendall tests. We first check if these columns exist in the dataset.

```
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. A Pearson correlation of 0.9648 shows a strong link between Study Hours and Exam Score. This means more study hours usually lead to higher exam scores.

Spearman's Rank Correlation:


```
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. A Spearman correlation of 0.9671 shows a strong monotonic link between Study Hours and Exam Score.

b) Kendall's Rank Correlation:

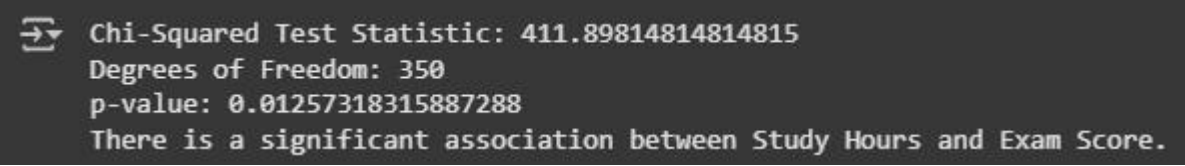
```
kendall_corr, _ = stats.kendalltau(df[col1], df[col2])
print(f"Kendall's Rank Correlation between {col1} and {col2}: {kendall_corr:.4f}")
```

 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. A Kendall correlation of 0.8861 shows students who study more consistently rank higher in exam performance.

c) Chi-Squared Test:

```
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}.")
```

A screenshot of a Jupyter Notebook cell showing the output of a Chi-Squared test. The output is displayed in a dark-themed console window with a magnifying glass icon on the left. The text shows the Chi-Squared Test Statistic as 411.89814814814815, Degrees of Freedom as 350, and a p-value of 0.01257318315887288. A final line states: "There is a significant association between Study Hours and Exam Score."

```
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.

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.