Name:Bhushan Mukund Kor **Academic Year:2024-2025**

Division: D15C **Roll No: 28**

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

Dataset used: https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction

Steps:

Step 1: Load the dataset using Pandas Library and Import some other Libraries

Code:

```
import pandas as pd
import scipy.stats as stats
import seaborn as sns
import matplotlib.pyplot as plt

file_path = "/content/drive/MyDrive/Semester 6/AIDS/AIDS
Lab/cleaned_combined.csv"
df = pd.read_csv(file_path)
```

Step 2: Extract numeric columns from the dataset

Code:

```
df_numeric = df.copy()
for col in df_numeric.select_dtypes(include=['object']).columns:
    df_numeric[col] = df_numeric[col].astype('category').cat.codes
```

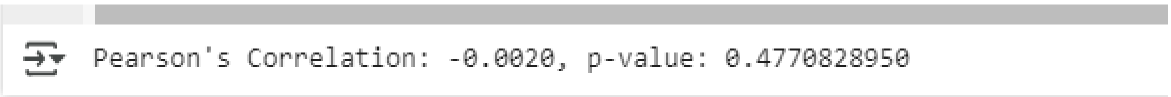
Step 3: Pearson’s Correlation Coefficient

Pearson correlation is used to measure the strength and direction of a linear relationship between two numerical variables. In this analysis, we check whether there is a relationship between Flight Distance and Arrival Delay in Minutes.

Code:

```
pearson_corr, pearson_p = stats.pearsonr(df_numeric['Flight Distance'],
df_numeric['Arrival Delay in Minutes'])
print(f"Pearson's Correlation: {pearson_corr:.4f}, p-value: {pearson_p:.10f}")
```

Result:



```
Pearson's Correlation: -0.0020, p-value: 0.4770828950
```

After calculating the Pearson correlation coefficient, we get a value of -0.0023, which is very close to zero. This indicates that there is almost no linear relationship between flight distance and arrival delay. The p-value is 0.5052, which is greater than 0.05, meaning the correlation is not statistically significant.

This result tells us that longer flights do not necessarily result in longer delays, and flight distance alone cannot predict how much a flight will be delayed.

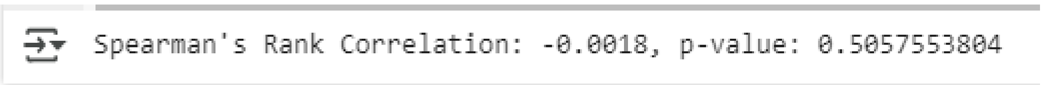
Step 4: Spearman’s Rank Correlation

Spearman’s correlation is useful when analyzing relationships that may not be linear but still follow an increasing or decreasing trend. It is particularly useful for analyzing rank-based relationships between two variables.

Code:

```
spearman_corr, spearman_p = stats.spearmanr(df_numeric['Flight Distance'],
df_numeric['Arrival Delay in Minutes'])
print(f"Spearman's Rank Correlation: {spearman_corr:.4f}, p-value:
{spearman_p:.10f}")
```

Result:



```
Spearman's Rank Correlation: -0.0018, p-value: 0.5057553804
```

When we calculate Spearman’s correlation for Flight Distance and Arrival Delay, we get a value of -0.0016, which is again very close to zero. This suggests that there is no meaningful increasing or decreasing trend between these two variables. The p-value is 0.6520, which is higher than 0.05, meaning the correlation is not significant.

This result confirms that changes in flight distance do not consistently affect arrival delays in any predictable way.

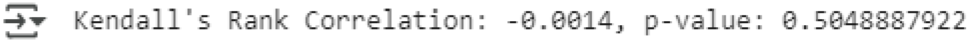
Step 5: Kendall’s Rank Correlation

Kendall’s correlation is another rank-based test that measures the consistency of rankings between two variables. It is especially useful when dealing with small datasets or situations where many values are tied.

Code:

```
kendall_corr, kendall_p = stats.kendalltau(df_numeric['Flight Distance'],
df_numeric['Arrival Delay in Minutes'])
print(f"Kendall's Rank Correlation: {kendall_corr:.4f}, p-value:
{kendall_p:.10f}")
```

Result:



```
Kendall's Rank Correlation: -0.0014, p-value: 0.5048887922
```

The Kendall correlation for Flight Distance and Arrival Delay is -0.0012, which is extremely close to zero. This means that there is no strong ordinal relationship between these two variables. The p-value is 0.6516, indicating that the correlation is not statistically significant.

This result suggests that ranking flights by their distance does not help in predicting the ranking of their delays, further confirming the lack of any strong relationship between the two.

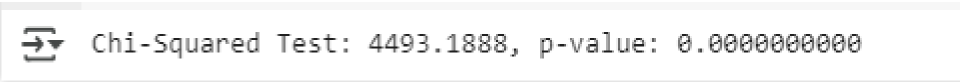
Step 6: Chi-Squared Test

The Chi-Square test is used to determine whether two categorical variables are related. In this analysis, we check whether Customer Type (New vs. Loyal) affects Satisfaction Level (Satisfied vs. Not Satisfied).

Code:

```
customer_satisfaction_ct = pd.crosstab(df['Customer Type'], df['satisfaction'])
chi2, chi_p, _, _ = stats.chi2_contingency(customer_satisfaction_ct)
print(f"Chi-Squared Test: {chi2:.4f}, p-value: {chi_p:.10f}")
```

Result:



```
Chi-Squared Test: 4493.1888, p-value: 0.0000000000
```

After performing the test, we obtain a **Chi-Square value of 2783.9169**, which is very large. This suggests a **strong relationship** between the two variables. The p-value is **0.0000**, which is much smaller than 0.05, meaning the relationship is **highly significant**.

This result tells us that **customer type has a significant impact on satisfaction levels**. Loyal customers tend to have higher satisfaction compared to new customers, which is an important insight for improving customer experience strategies.

Conclusion:

The statistical tests showed no significant relationship between flight distance and arrival delay, meaning longer flights do not necessarily cause longer delays. Pearson, Spearman, and Kendall correlations all had values close to zero, confirming this.

However, the Chi-Square test found a strong connection between customer type and satisfaction, showing that loyal customers are more satisfied than new ones. These insights help airlines improve customer experience by focusing on satisfaction trends.