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's correlation is used to measure the strength and direction of a linear relationship between two numerical variables. Here, we test whether there is a relationship between Flight Distance and Arrival Delay in Minutes.

Hypothesis:

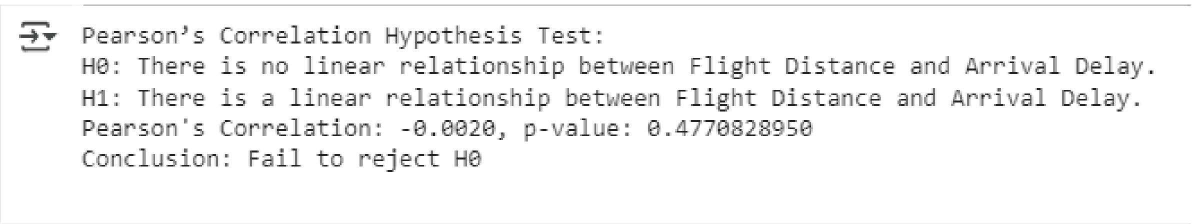
H0 (Null Hypothesis): There is no linear relationship between Flight Distance and Arrival Delay.
H1 (Alternative Hypothesis): There is a linear relationship between Flight Distance and Arrival Delay.

Code:

```
pearson_corr, pearson_p = stats.pearsonr(df_numeric['Flight Distance'],
df_numeric['Arrival Delay in Minutes'])
print("Pearson’s Correlation Hypothesis Test:")
print("H0: There is no linear relationship between Flight Distance and Arrival
Delay.")
```

```
print("H1: There is a linear relationship between Flight Distance and Arrival
Delay.")
print(f"Pearson's Correlation: {pearson_corr:.4f}, p-value: {pearson_p:.10f}")
print("Conclusion:", "Fail to reject H0" if pearson_p > 0.05 else "Reject H0",
"\n")
```

Result:



```
Pearson’s Correlation Hypothesis Test:
H0: There is no linear relationship between Flight Distance and Arrival Delay.
H1: There is a linear relationship between Flight Distance and Arrival Delay.
Pearson's Correlation: -0.0020, p-value: 0.4770828950
Conclusion: Fail to reject H0
```

The Pearson correlation coefficient is -0.0020, which is very close to zero, indicating almost no linear relationship between flight distance and arrival delay. The p-value is 0.47708, which is greater than 0.05, meaning the correlation is not statistically significant.

This result confirms that longer flights do not necessarily result in longer delays, and flight distance alone cannot be used to predict arrival delay.

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 measures how well the ranks of two variables correspond to each other.

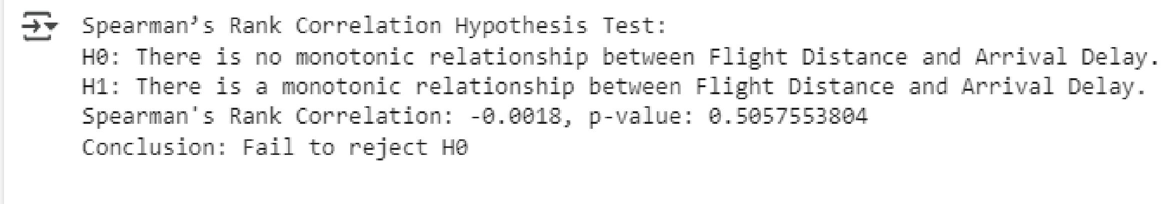
Hypothesis:

H0 (Null Hypothesis): There is no monotonic relationship between Flight Distance and Arrival Delay.
H1 (Alternative Hypothesis): There is a monotonic relationship between Flight Distance and Arrival Delay.

Code:

```
spearman_corr, spearman_p = stats.spearmanr(df_numeric['Flight Distance'],
df_numeric['Arrival Delay in Minutes'])
print("Spearman’s Rank Correlation Hypothesis Test:")
print("H0: There is no monotonic relationship between Flight Distance and
Arrival Delay.")
print("H1: There is a monotonic relationship between Flight Distance and
Arrival Delay.")
print(f"Spearman's Rank Correlation: {spearman_corr:.4f}, p-value:
{spearman_p:.10f}")
print("Conclusion:", "Fail to reject H0" if spearman_p > 0.05 else "Reject H0",
"\n")
```

Result:



```
Spearman’s Rank Correlation Hypothesis Test:
H0: There is no monotonic relationship between Flight Distance and Arrival Delay.
H1: There is a monotonic relationship between Flight Distance and Arrival Delay.
Spearman's Rank Correlation: -0.0018, p-value: 0.5057553804
Conclusion: Fail to reject H0
```

The Spearman correlation coefficient is -0.0018, which is very close to zero, meaning there is no meaningful increasing or decreasing trend between flight distance and arrival delay. The p-value is 0.505755, which is greater than 0.05, proving that the correlation is not statistically significant.

This result confirms that changes in flight distance do not consistently influence arrival delays.

Step 5: Kendall’s Rank Correlation

Kendall’s correlation measures the strength of the ordinal association between two variables. It is useful for analyzing ranked data, especially when there are many tied values.

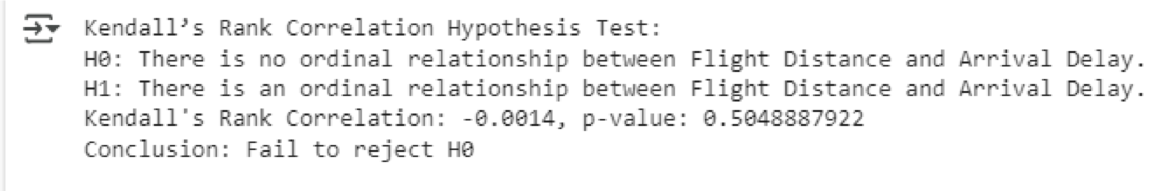
Hypothesis:

H0 (Null Hypothesis): There is no ordinal relationship between Flight Distance and Arrival Delay.
H1 (Alternative Hypothesis): There is an ordinal relationship between Flight Distance and Arrival Delay.

Code:

```
kendall_corr, kendall_p = stats.kendalltau(df_numeric['Flight Distance'],
df_numeric['Arrival Delay in Minutes'])
print("Kendall’s Rank Correlation Hypothesis Test:")
print("H0: There is no ordinal relationship between Flight Distance and Arrival
Delay.")
print("H1: There is an ordinal relationship between Flight Distance and Arrival
Delay.")
print(f"Kendall's Rank Correlation: {kendall_corr:.4f}, p-value:
{kendall_p:.10f}")
print("Conclusion:", "Fail to reject H0" if kendall_p > 0.05 else "Reject H0",
"\n")
```

Result:



```
Kendall’s Rank Correlation Hypothesis Test:
H0: There is no ordinal relationship between Flight Distance and Arrival Delay.
H1: There is an ordinal relationship between Flight Distance and Arrival Delay.
Kendall's Rank Correlation: -0.0014, p-value: 0.5048887922
Conclusion: Fail to reject H0
```

The Kendall correlation coefficient is -0.0014, which is very close to zero, meaning there is no strong ordinal relationship between flight distance and arrival delay. The p-value is 0.504888, 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.

Step 6: Chi-Squared Test

The Chi-Square test is used to check whether two categorical variables are dependent on each other. Here, we test if Customer Type (New vs. Loyal) affects Satisfaction Level (Satisfied vs. Not Satisfied).

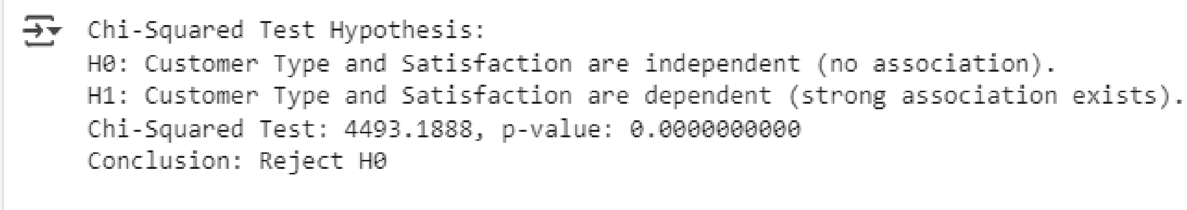
Hypothesis:

H0 (Null Hypothesis): Customer Type and Satisfaction are independent.
H1 (Alternative Hypothesis): Customer Type and Satisfaction are dependent.

Code:

```
customer_satisfaction_ct = pd.crosstab(df['Customer Type'], df['satisfaction'])
chi2, chi_p, _, _ = stats.chi2_contingency(customer_satisfaction_ct)
print("Chi-Squared Test Hypothesis:")
print("H0: Customer Type and Satisfaction are independent (no association).")
print("H1: Customer Type and Satisfaction are dependent (strong association
exists).")
print(f"Chi-Squared Test: {chi2:.4f}, p-value: {chi_p:.10f}")
print("Conclusion:", "Fail to reject H0" if chi_p > 0.05 else "Reject H0", "\n")
```

Result:



```
Chi-Squared Test Hypothesis:
H0: Customer Type and Satisfaction are independent (no association).
H1: Customer Type and Satisfaction are dependent (strong association exists).
Chi-Squared Test: 4493.1888, p-value: 0.0000000000
Conclusion: Reject H0
```

The Chi-Square value is 4493.1888, which is very large, indicating a strong relationship between customer type and satisfaction. The p-value is 0.0000, meaning the result is highly significant.

This result proves that customer type has a significant impact on satisfaction, with loyal customers showing higher satisfaction levels than new customers.

Conclusion:

The Pearson, Spearman, and Kendall correlation tests all showed that there is no significant relationship between flight distance and arrival delay. This confirms that longer flights do not necessarily lead to longer delays.

However, the Chi-Square test found a strong relationship between customer type and satisfaction, showing that loyal customers tend to be more satisfied than new ones. These insights can help airlines improve their customer service strategies based on passenger satisfaction trends.