

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:

1) Load the dataset using Pandas

```
import pandas as pd
import scipy.stats as stats
import seaborn as sns
import matplotlib.pyplot as plt
file_path = "/content/sample_data/cleaned_combined.csv"
df = pd.read_csv(file_path)
```

2) Extract numeric columns from the dataset

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

3) Perform Pearson Correlation.

The Pearson correlation coefficient (r) measures the linear relationship between two variables, ranging from -1 to 1. A value close to 1 indicates a strong positive correlation, -1 a strong negative correlation, and 0 no correlation. It is widely used in statistics for predictive analysis.

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")
```

Output:

```
→ 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
```

In the scipy library, we have a library function called as `scipy.stats` which we will be using to find the correlation coefficients and p-score of the columns mentioned. Here, we are finding the Pearson's Correlation Coefficient between the column Flight Distance and Arrival Delay in Minutes.

Flight Distance and Arrival Delay show almost no linear relationship, with a near-zero correlation (-0.0020) and an insignificant p-value (0.4771), indicating that changes in flight distance do not predict arrival delay. As $p_value > 0.05$, we are not able to reject the null hypothesis and hence, there is no linear relationship between Flight Distance and Arrival Delay

4) Perform Spearman's Rank Correlation.

Spearman's rank correlation coefficient (ρ) measures the monotonic relationship between two variables, assessing how well their ranks correspond. It ranges from -1 to 1, where 1 indicates a perfect increasing relationship, -1 a perfect decreasing relationship, and 0 no correlation. It is useful for nonlinear and ordinal data.

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")
```

Output:

```
➞ 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
```

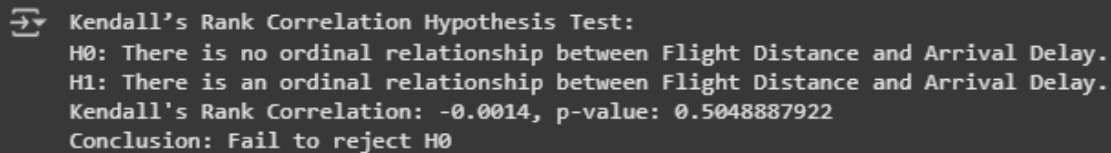
In the SciPy library, we use `scipy.stats` to compute the Spearman's Rank Correlation between Flight Distance and Arrival Delay in Minutes. The correlation coefficient (-0.0018) and p-value (0.5058) suggest no meaningful monotonic relationship, meaning changes in flight distance do not consistently influence arrival delay. As $p_value > 0.05$, we fail to reject the null hypothesis and hence, no monotonic relationship is observed between Flight Distance and Arrival Time.

5) Perform Kendall's Rank Correlation

Kendall's rank correlation coefficient (τ) measures the ordinal association between two variables. It evaluates the consistency of rank ordering between them. Ranging from -1 to 1, $\tau = 1$ indicates perfect agreement, -1 perfect disagreement, and 0 no correlation. It is robust for small datasets and tied ranks.

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")
```

Output:

```
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
```

Using the `scipy.stats` module, we determine the Kendall's Rank Correlation between Flight Distance and Arrival Delay in Minutes. The correlation (-0.0014) and p-value (0.5049) indicate no significant ordinal relationship, meaning ranking flight distances does not predict ranking arrival delays. Here, as $p_value > 0.05$, we fail to reject the null hypothesis and hence, observe no monotonic relationship between Flight Distance and Arrival Delay.

6) Perform Chi-Square Test

The Chi-Square test is a statistical test used to determine if there is a significant association between two categorical variables. It compares observed and expected frequencies in a contingency table. A higher Chi-Square value suggests a stronger relationship. It is widely used in independence testing and goodness-of-fit analysis.

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")
```

Output:

```
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 `scipy.stats` module is used to perform a Chi-Squared Test on Customer Type and Satisfaction. A high chi-square statistic (4493.1888) and near-zero p-value indicate a strong dependence between these variables, meaning customer type significantly influences satisfaction levels. As the $p_value < 0.05$, we have to reject the null hypothesis. Hence, there is a dependence between Customer Type and Satisfaction.

Conclusion:

The statistical hypothesis tests performed on the dataset provide insights into relationships between various airline passenger attributes. The Pearson's Correlation Coefficient between Flight Distance and Arrival Delay in Minutes was found to be -0.0020 with a p-value of 0.4771, indicating no significant linear relationship between these variables. Similarly, the Spearman's Rank Correlation was -0.0018 with a p-value of 0.5058, confirming that no monotonic relationship exists. Furthermore, the Kendall's Rank Correlation resulted in -0.0014 with a p-value of 0.5049, reinforcing the finding that changes in flight distance do not systematically influence arrival delays. In all three correlation tests, since the p-value was greater than 0.05, the null hypothesis was not rejected, meaning there is no significant association between flight distance and arrival delay. However, the Chi-Square Test between Customer Type and Satisfaction yielded a chi-square statistic of 4493.1888 and a p-value close to zero, indicating a strong dependence between these categorical variables. This suggests that passenger satisfaction is significantly influenced by customer type. Overall, while flight distance does not appear to predict arrival delays, customer type plays a crucial role in determining passenger satisfaction.