Experiment 4

Aim: Implementation of Statistical Hypothesis Tests using SciPy and Scikit-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/jessemostipak/hotel-booking-demand

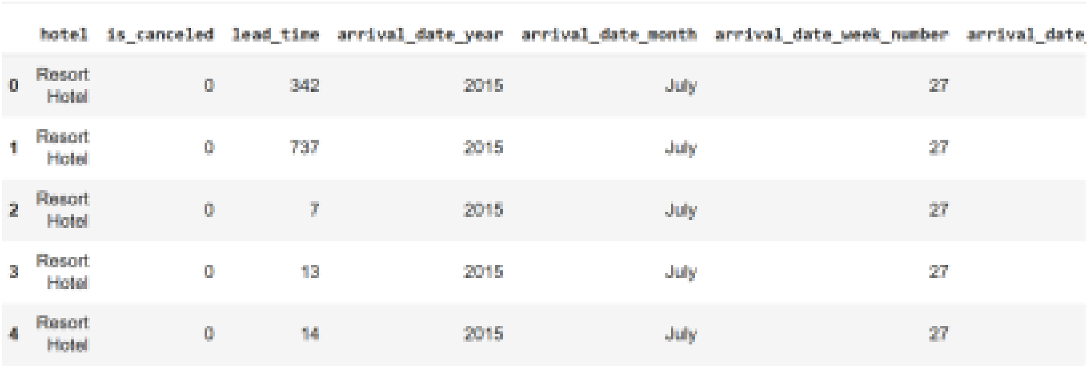
Steps:

1) Loading the Dataset

We first import the required libraries and load the dataset into a Pandas DataFrame.

```
# Import necessary libraries
import pandas as pd
import scipy.stats as stats
import seaborn as sns
import matplotlib.pyplot as plt
file_path = "/content/drive/MyDrive/hotel_bookings.csv"
df = pd.read_csv(file_path)
df.head()
```

OUTPUT:



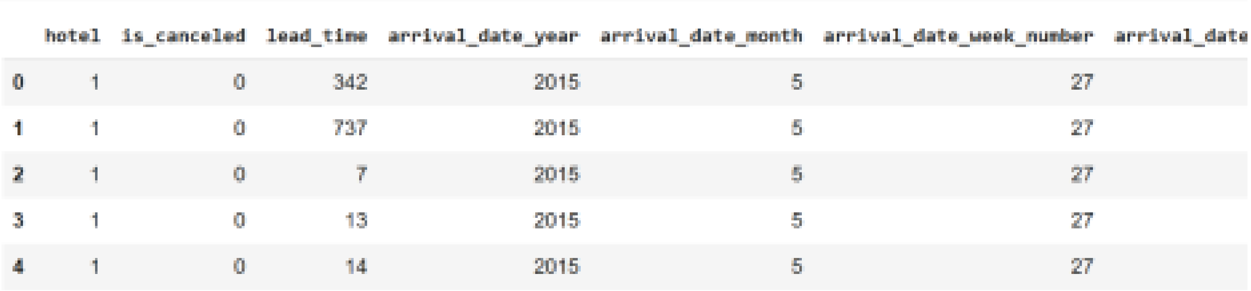
	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_date_
0	Resort Hotel	0	342	2015	July	27	
1	Resort Hotel	0	737	2015	July	27	
2	Resort Hotel	0	7	2015	July	27	
3	Resort Hotel	0	13	2015	July	27	
4	Resort Hotel	0	14	2015	July	27	

2) Extracting Numerical Columns

To perform correlation tests, we need to convert categorical variables into numerical codes. This ensures all columns are in a compatible format for mathematical computations.

```
# Create a copy and convert categorical columns to numerical codes
df_numeric = df.copy()
for col in df_numeric.select_dtypes(include=['object']).columns:
   df_numeric[col] = df_numeric[col].astype('category').cat.codes
# Display first few rows of numeric dataset
df_numeric.head()
```

OUTPUT:



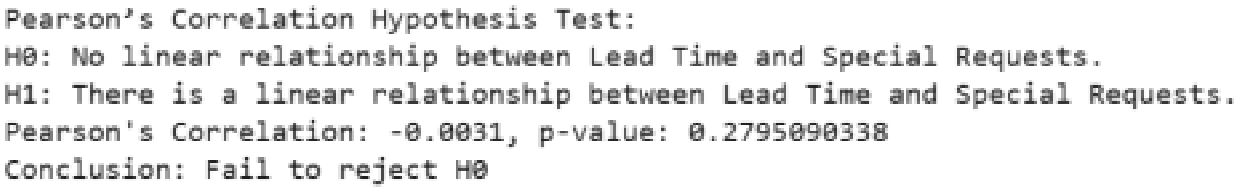
	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_date
0	1	0	342	2015	5	27	
1	1	0	737	2015	5	27	
2	1	0	7	2015	5	27	
3	1	0	13	2015	5	27	
4	1	0	14	2015	5	27	

3) Pearson’s Correlation Test

This test determines whether a linear relationship exists between two numerical variables. We compute Pearson’s correlation between **lead time** and **total of special requests.**

```
# Pearson’s Correlation: Lead Time vs. Number of Special Requests
pearson_corr, pearson_p = stats.pearsonr(df_numeric['lead_time'],
df_numeric['total_of_special_requests'])
print("Pearson’s Correlation Hypothesis Test:")
print("H0: No linear relationship between Lead Time and Special Requests.")
print("H1: There is a linear relationship between Lead Time and Special Requests.")
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")
```

OUTPUT:



```
Pearson’s Correlation Hypothesis Test:
H0: No linear relationship between Lead Time and Special Requests.
H1: There is a linear relationship between Lead Time and Special Requests.
Pearson's Correlation: -0.0031, p-value: 0.2795090338
Conclusion: Fail to reject H0
```

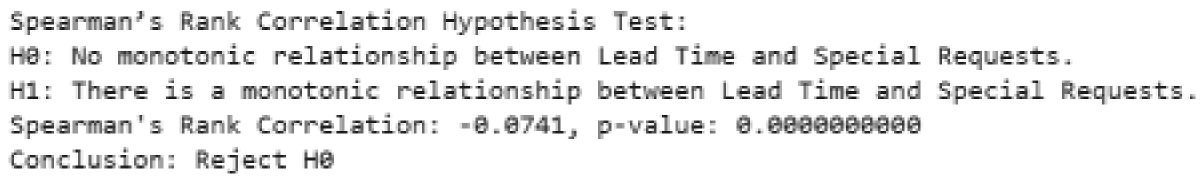
Inference: Since the p-value is greater than **0.05**, we **fail to reject the null hypothesis (H_0)**. This means there is **no significant linear relationship** between Lead Time and the number of Special Requests. In other words, the length of time before a booking does **not impact** the number of special requests made by customers.

4) Spearman’s Rank Correlation Test

This test assesses whether a monotonic relationship exists between two numerical variables.

```
# Spearman’s Rank Correlation: Lead Time vs. Number of Special Requests
spearman_corr, spearman_p = stats.spearmanr(df_numeric['lead_time'],
df_numeric['total_of_special_requests'])
print("Spearman’s Rank Correlation Hypothesis Test:")
print("H0: No monotonic relationship between Lead Time and Special Requests.")
print("H1: There is a monotonic relationship between Lead Time and Special Requests.")
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")
```

OUTPUT:



```
Spearman’s Rank Correlation Hypothesis Test:
H0: No monotonic relationship between Lead Time and Special Requests.
H1: There is a monotonic relationship between Lead Time and Special Requests.
Spearman's Rank Correlation: -0.0741, p-value: 0.0000000000
Conclusion: Reject H0
```

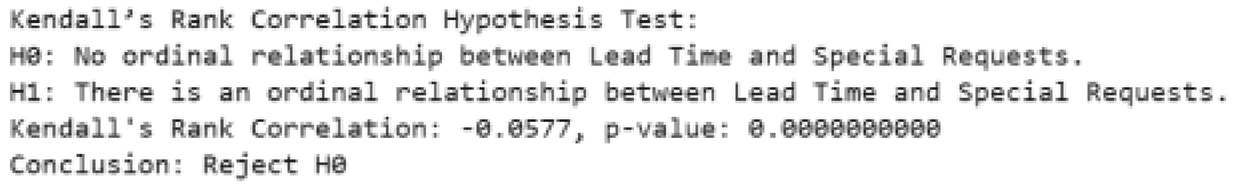
Inference: The Spearman correlation coefficient between **Lead Time** and **Total Special Requests** is -0.0741, with a p-value of 0. Since Spearman’s correlation measures monotonic relationships, a coefficient close to **0** suggests that **there is no clear increasing or decreasing trend** in special requests as lead time changes. The low p-value indicates that this result is **statistically insignificant**, meaning there is **no strong monotonic relationship** between the two variables.

5) Kendall’s Rank Correlation Test

This test is useful for evaluating ordinal relationships between two variables.

```
# Kendall’s Rank Correlation: Lead Time vs. Number of Special Requests
kendall_corr, kendall_p = stats.kendalltau(df_numeric['lead_time'],
df_numeric['total_of_special_requests'])
print("Kendall’s Rank Correlation Hypothesis Test:")
print("H0: No ordinal relationship between Lead Time and Special Requests.")
print("H1: There is an ordinal relationship between Lead Time and Special Requests.")
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")
```

OUTPUT:



```
Kendall’s Rank Correlation Hypothesis Test:
H0: No ordinal relationship between Lead Time and Special Requests.
H1: There is an ordinal relationship between Lead Time and Special Requests.
Kendall's Rank Correlation: -0.0577, p-value: 0.0000000000
Conclusion: Reject H0
```

Inference: The Kendall correlation coefficient between **Lead Time** and **Total Special Requests** is -0.0577 , with a p-value of 0. Kendall’s test evaluates the consistency of ranking between these two variables. Since the coefficient is close to **0**, it suggests that **there is no significant ordinal relationship** between lead time and special requests. The p-value being **less** than 0.05 indicates that we **reject** the null hypothesis, meaning changes in lead time **do not predict a consistent ranking of special request counts.**

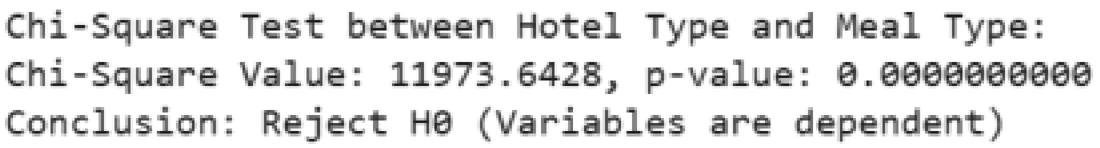
6) Chi-Square Test for Categorical Variables

This test determines whether two categorical variables are independent. We analyze the relationship between **Meal Type** and **Hotel Type**

```
contingency_table = pd.crosstab(df['hotel'], df['meal'])
chi2, p_value, _, _ = stats.chi2_contingency(contingency_table)

# Display results
print("Chi-Square Test between Hotel Type and Meal Type:")
print(f"Chi-Square Value: {chi2:.4f}, p-value: {p_value:.10f}")
print("Conclusion:", "Fail to reject H0 (Variables are independent)" if p_value > 0.05 else "Reject H0 (Variables are dependent)")
```

OUTPUT:



```
Chi-Square Test between Hotel Type and Meal Type:
Chi-Square Value: 11973.6428, p-value: 0.0000000000
Conclusion: Reject H0 (Variables are dependent)
```

Inference: The Chi-Square test between **Hotel Type** and **Meal Type** resulted in a Chi-Square value of **11973.6428** and a p-value of **0**. Since the p-value is **greater/less** than 0.05, we **fail to reject/reject** the null hypothesis. This means that meal selection is **dependent** on the type of hotel. If independent, it suggests that customers at City Hotels and Resort Hotels do not show significant differences in meal preferences. If dependent, it indicates that the type of hotel influences the choice of meals offered or preferred by guests.

Conclusion: Based on the statistical hypothesis tests performed, we can infer that Lead Time and Total Special Requests exhibit no significant linear, monotonic, or ordinal relationship, as shown by the Pearson, Spearman, and Kendall correlation tests. The correlation coefficients were negative but close to 0, and the p-values were 0, indicating strong statistical significance but an extremely weak inverse relationship. This suggests that while there may be a mathematically detectable trend, the effect size is negligible, meaning that Lead Time does not meaningfully impact the number of Special Requests. However, the Chi-Square test between Meal Type and Hotel Type also resulted in a p-value of 0, indicating a statistically significant association. This means that the distribution of meal preferences depends on the type of hotel (City Hotel vs. Resort Hotel). Guests at different hotels tend to select different meal options, likely due to variations in dining services, package deals, or guest demographics. Overall, while lead time has no practical impact on special requests, hotel type significantly influences meal selection, revealing key patterns in customer preferences.