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: The Pearson correlation coefficient between Lead Time and Total Special Requests is -0.0031, indicating an almost nonexistent linear relationship. This means that as lead time increases or decreases, special requests remain nearly unchanged.The p-value is 0.27950, which is greater than 0.05, meaning the result is not statistically significant. Since we fail to reject the null hypothesis (H_0), there is no evidence of a linear relationship between these two variables. This suggests that lead time does not meaningfully impact the number of special requests in a predictive way.

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, indicating a very weak negative monotonic relationship. This means that as lead time increases, special requests tend to decrease slightly, but the effect is negligible.The p-value is extremely small (less than 0.01), which suggests statistical significance—meaning we reject the null hypothesis (H_0) that there is no relationship. However, despite the statistical significance, the correlation is so weak that it has little to no practical significance in predicting special requests based on lead time.

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's Tau correlation coefficient between Lead Time and Total Special Requests is -0.0577, indicating a very weak negative ordinal relationship. This means that as lead time increases, the ranking of special requests slightly decreases, but the effect is extremely small.The p-value is effectively 0 (extremely small, below any reasonable significance threshold), meaning the result is statistically significant and we reject the null hypothesis (H_0). However, despite this statistical significance, the correlation is so weak that it has no meaningful impact in practice. In other words, while a relationship technically exists, lead time is not a useful predictor of special requests.

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 statistic for the relationship between Hotel Type and Meal Type is 11,973.6428, with a p-value effectively 0 (extremely small). Since the p-value is far below 0.05, we reject the null hypothesis (H_0), meaning there is a statistically significant association between hotel type and meal type.This suggests that meal preferences are not independent of hotel type—guests at different hotel types (City Hotel vs. Resort Hotel) tend to choose meals differently. This dependency could be due to differences in hotel dining options, guest demographics, or package inclusions affecting meal choices.

Conclusion: Based on the statistical hypothesis tests performed, we analyzed the relationships between Lead Time and Total Special Requests using Pearson, Spearman, and Kendall correlation tests and examined the association between Hotel Type and Meal Type using the Chi-Square test.

The Pearson correlation coefficient (-0.0031) and p-value (0.27950) indicate no significant linear relationship between Lead Time and Total Special Requests, meaning that lead time does not predict the number of special requests a customer makes. Similarly, the Spearman (-0.0741) and Kendall (-0.0577) correlation coefficients suggest a very weak negative monotonic and ordinal relationship, with extremely small p-values confirming statistical significance. However, despite

this significance, the correlation values are so close to zero that the effect is negligible in practice. This implies that while there may be a detectable relationship, lead time is not a meaningful factor in determining special requests.

On the other hand, the Chi-Square test ($\chi^2 = 11{,}973.6428$, p-value ≈ 0) between Hotel Type and Meal Type confirms a strong dependency between the two variables, leading us to reject the null hypothesis. This means that meal preferences are significantly influenced by hotel type, possibly due to differences in dining services, guest demographics, or package offerings at City Hotels versus Resort Hotels.