DATA SCIENCE LAB EXPERIMENT 4

Name: Atharva Patil D15C Roll no: 39

Aim: Implementation of Statistical Hypothesis Test using Scipy and Sci-kit learn.
Perform the following Tests:Correlation Tests:
a) Pearson's Correlation Coefficient
Pearson's Correlation Coefficient (r) measures the linear relationship between two variables. It quantifies how strongly and in which direction (positive or negative) two variables are related.

Interpretation of r:

- +1 → Perfect positive correlation (as X increases, Y also increases)
- -1 → Perfect negative correlation (as X increases, Y decreases)
- 0 → No correlation (X and Y are unrelated)

b) Spearman's Rank Correlation
Spearman's rank correlation measures the monotonic relationship between two variables. It is a non-parametric alternative to Pearson's correlation and is useful when data is not normally distributed or when the relationship is not linear.

Interpretation of ρ\rhoρ:

- +1 → Perfect positive monotonic relationship
- -1 → Perfect negative monotonic relationship
- 0 → No correlation

c) Kendall's Rank Correlation
Kendall's τ (tau) coefficient measures the ordinal association between two variables. It evaluates the strength and direction of the monotonic relationship, similar to Spearman's correlation, but it considers the number of concordant and discordant pairs rather than rank differences.

DATA SCIENCE LAB EXPERIMENT 4

Name: Atharva Patil D15C Roll no: 39

Interpretation of τ\tauτ:

- +1 → Perfect positive correlation (both variables rank identically)
- -1 → Perfect negative correlation (inverse ranking)
- 0 → No correlation

d) Chi-Squared Test
The Chi-Square test is a statistical test used to determine if there is a significant association between two categorical variables. It compares the observed and expected frequencies in different categories.
Interpretation:

- If p-value < 0.05 → Reject H0H_0H0, meaning there is a significant relationship.
- If p-value > 0.05 → Fail to reject H0H_0H0, meaning no significant relationship.

Steps:
1) Load the dataset using Pandas

DATA SCIENCE LAB EXPERIMENT 4

Name: Atharva Patil D15C Roll no: 39

```
import pandas as pd
import numpy as np

df = pd.read_excel("/content/sample_data/Bengaluru_House_Data.xlsx")
print(df.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 13320 entries, 0 to 13319
Data columns (total 9 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   area_type     13320 non-null  object
 1   availability  13320 non-null  object
 2   location      13319 non-null  object
 3   size          13304 non-null  object
 4   society       7818 non-null   object
 5   total_sqft    13320 non-null  object
 6   bath          13247 non-null  float64
 7   balcony       12711 non-null  float64
 8   price         13320 non-null  float64
dtypes: float64(3), object(6)
memory usage: 936.7+ KB
None
```

2) Extract numeric columns from the dataset

```
# Function to handle 'total_sqft' column (convert ranges to average values)
def convert_sqft(value):
    try:
        # If the value contains a range (e.g., "2100 - 2850"), take the average
        if '-' in value:
            low, high = map(float, value.split('-'))
            return (low + high) / 2
        return float(value)  # Convert to float if it's a single value
    except:
        return np.nan  # Return NaN for invalid values

# Apply the conversion function
df['total_sqft'] = df['total_sqft'].astype(str).apply(convert_sqft)

# Drop rows with NaN values (if conversion fails)
df.dropna(subset=['total_sqft'], inplace=True)
```

DATA SCIENCE LAB EXPERIMENT 4

Name: Atharva Patil D15C Roll no: 39

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

```
corr, p_value = pearsonr(df['total_sqft'], df['price'])
print(f"Pearson Correlation: {corr}, P-value: {p_value}")
```

```
Pearson Correlation: 0.5755591870157646, P-value: 0.0
```

The Pearson correlation coefficient between total square footage and house price is 0.576, with a p-value of 0.0. This indicates a moderate positive correlation, meaning that as the total square footage increases, the house price also tends to increase. The p-value of 0.0 suggests that this correlation is statistically significant, implying that the relationship is unlikely to have occurred by random chance. However, since the correlation is not close to 1, other factors besides square footage may also play a significant role in determining house prices.

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.

```
corr, p_value = spearmanr(df['total_sqft'], df['price'])
print(f"Spearman Correlation: {corr}, P-value: {p_value}")
```

```
Spearman Correlation: 0.736118018937836, P-value: 0.0
```

DATA SCIENCE LAB EXPERIMENT 4

Name: Atharva Patil D15C Roll no: 39

The Spearman correlation coefficient between total square footage and house price is 0.736, with a p-value of 0.0. This indicates a strong positive monotonic relationship, meaning that as total square footage increases, house prices tend to increase consistently. Compared to the Pearson correlation (which was 0.576), the higher Spearman correlation suggests that the relationship between these variables may not be strictly linear but still follows a clear increasing trend.

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, τ = 1 indicates perfect agreement, -1 perfect disagreement, and 0 no correlation. It is robust for small datasets and tied ranks.

```
corr, p_value = kendalltau(df['total_sqft'], df['price'])
print(f"Kendall Correlation: {corr}, P-value: {p_value}")
```

```
Kendall Correlation: 0.5658054300704137, P-value: 0.0
```

The Kendall correlation coefficient between total square footage and house price is 0.566, with a p-value of 0.0. This indicates a moderate to strong positive ordinal association, meaning that as square footage increases, house prices tend to increase as well. Since Kendall's tau is designed for ranked data, this result confirms that the relationship is consistently increasing, even if the exact numerical differences between values vary
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.

DATA SCIENCE LAB EXPERIMENT 4

Name: Atharva Patil D15C Roll no: 39

```
# Convert categorical columns to string if needed
df['area_type'] = df['area_type'].astype(str)
df['availability'] = df['availability'].astype(str)

# Create a contingency table
contingency_table = pd.crosstab(df['area_type'], df['availability'])

# Perform the Chi-Squared test
chi2, p, dof, expected = chi2_contingency(contingency_table)

print(f"Chi-Squared Statistic: {chi2}, P-value: {p}")
```

```
Chi-Squared Statistic: 799.5963325974401, P-value: 2.486010365948735e-61
```

The Chi-Squared test for independence between area type and availability resulted in a Chi-Squared statistic of 799.60 and a p-value of 2.49e-61. Since the p-value is extremely small (close to zero), we reject the null hypothesis, which assumes no association between these two categorical variables. This suggests a strong dependence between area type and availability, meaning that certain area types are significantly more or less available compared to others.

DATA SCIENCE LAB EXPERIMENT 4

Name: Atharva Patil D15C Roll no: 39

Conclusion:The statistical analysis of the Bengaluru house prices dataset reveals key insights into the relationship between various housing attributes and prices. The Pearson correlation (0.575) indicates a moderate positive linear relationship between total square footage and price, while the Spearman (0.736) and Kendall (0.566) correlations suggest a stronger ordinal relationship, implying that larger homes generally have higher prices, even if the increase is not perfectly linear. Furthermore, the Chi-Squared test confirms a significant dependency between area type and availability (p-value ≈ 0), meaning that certain area types tend to have more or fewer available properties. Overall, the results suggest that square footage is a crucial determinant of house prices, and availability varies significantly by area type. These findings provide valuable insights for potential buyers, real estate developers, and policymakers in understanding price trends and market dynamics in Bengaluru.