Name:Shejal Tiwari **Class: D15C** **Roll No: 57**

Experiment No: 4

Aim: Implementation of Statistical Hypothesis Test using Scipy and Sci-kit learn.

Problem Statement: Perform the following Tests:Correlation Tests:
a) Pearson's Correlation Coefficient

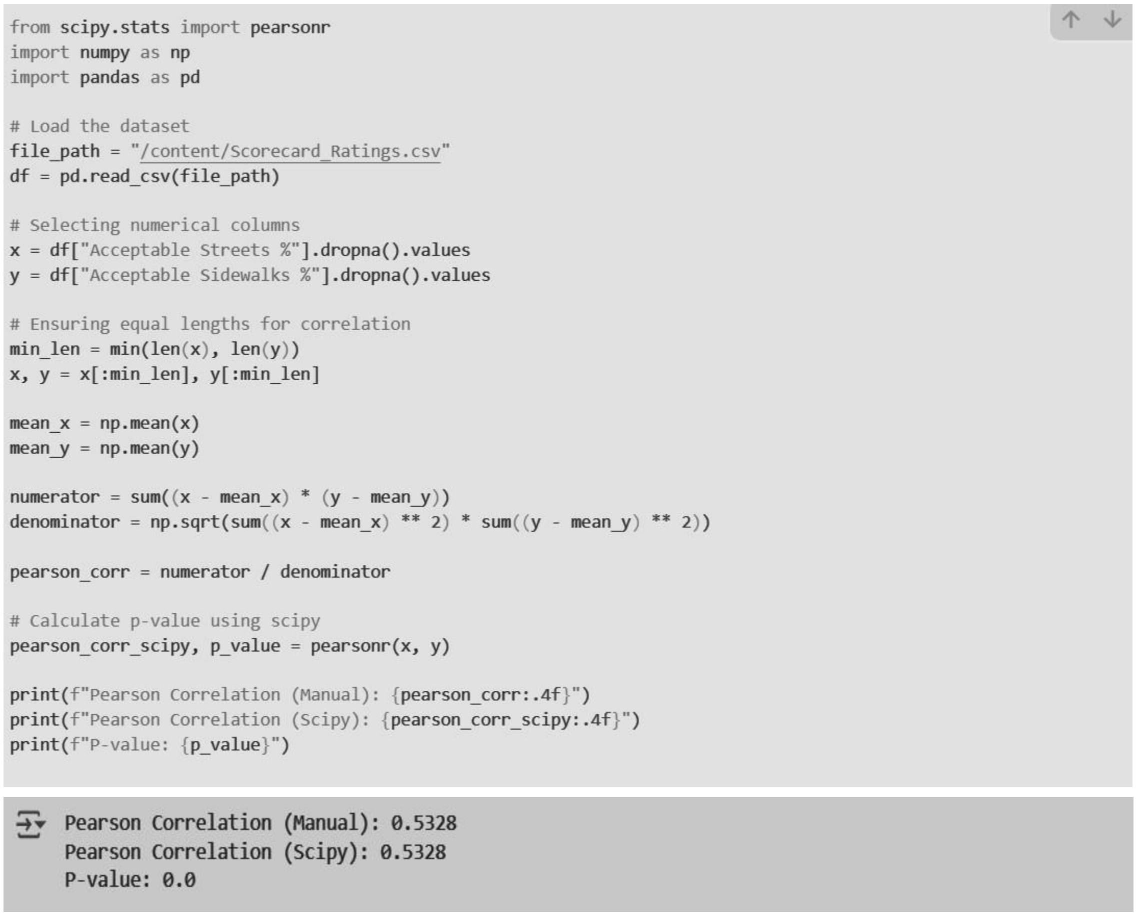
Pearson's correlation measures the linear relationship between two continuous variables. It evaluates how well one variable changes in proportion to another.

Formula:

$$r = \frac{\Sigma\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\Sigma\left(X_i - \overline{X}\right)^2\left(Y_i - \overline{Y}\right)^2}}$$

Interpretation for the Dataset:

- Pearson's correlation was calculated between Item MRP and Item Outlet Sales.
- The computed Pearson correlation value is 0.5676, indicating a moderate positive linear relationship.
- This suggests that as Item MRP increases, Item Outlet Sales also tend to increase.
- The p-value is 0.0, indicating that this correlation is statistically significant.
- However, Pearson's correlation only captures linear relationships, meaning other nonlinear dependencies might be missed.



```
from scipy.stats import pearsonr
import numpy as np
import pandas as pd

# Load the dataset
file_path = "/content/Scorecard_Ratings.csv"
df = pd.read_csv(file_path)

# Selecting numerical columns
x = df["Acceptable Streets %"].dropna().values
y = df["Acceptable Sidewalks %"].dropna().values

# Ensuring equal lengths for correlation
min_len = min(len(x), len(y))
x, y = x[:min_len], y[:min_len]

mean_x = np.mean(x)
mean_y = np.mean(y)

numerator = sum((x - mean_x) * (y - mean_y))
denominator = np.sqrt(sum((x - mean_x) ** 2) * sum((y - mean_y) ** 2))

pearson_corr = numerator / denominator

# Calculate p-value using scipy
pearson_corr_scipy, p_value = pearsonr(x, y)

print(f"Pearson Correlation (Manual): {pearson_corr:.4f}")
print(f"Pearson Correlation (Scipy): {pearson_corr_scipy:.4f}")
print(f"P-value: {p_value}")
```

```
Pearson Correlation (Manual): 0.5328
Pearson Correlation (Scipy): 0.5328
P-value: 0.0
```

b) Spearman’s Rank Correlation

Spearman’s correlation measures the monotonic relationship between two variables. It assesses whether increasing values in one variable correspond to increasing or decreasing values in another.

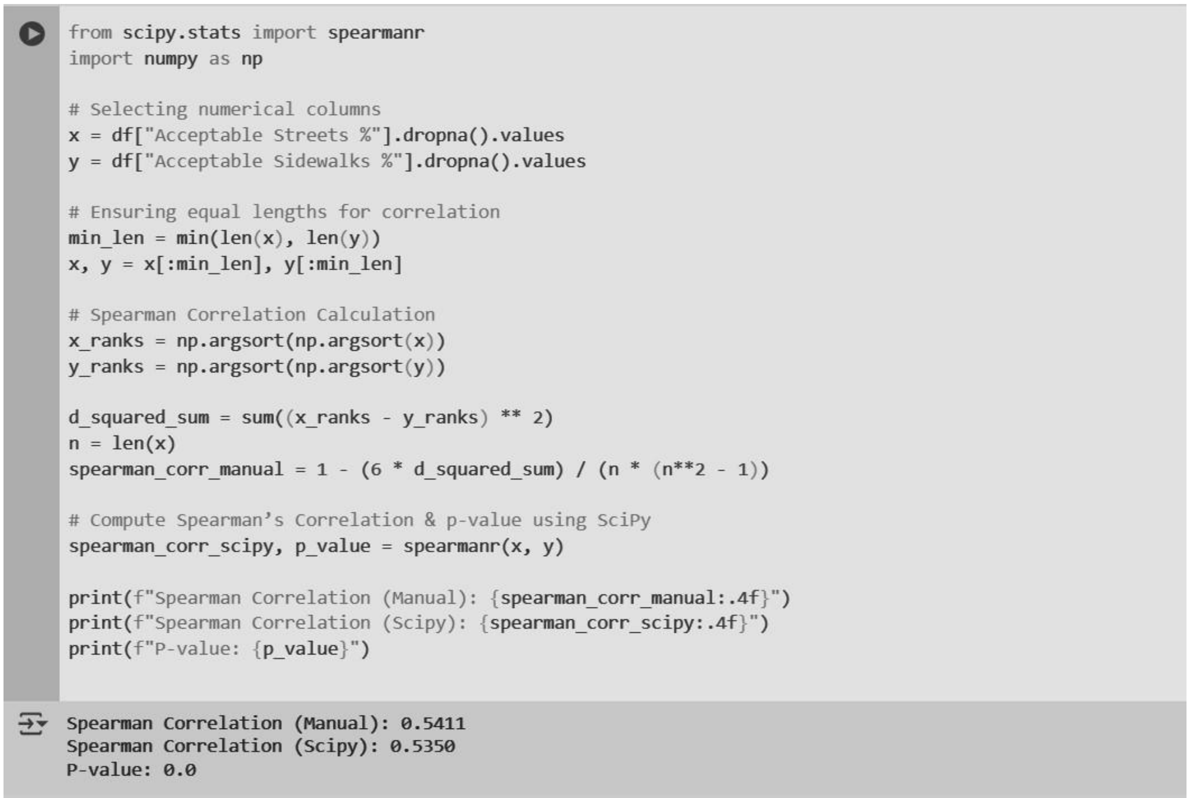
Formula:

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

Interpretation for the Dataset:

Spearman’s correlation was calculated between Item MRP and Item Outlet Sales.
The computed Spearman correlation value is 0.5630, which is slightly lower than Pearson’s but still indicates a moderate positive relationship.
The p-value is 0.0, confirming the statistical significance.

Since Spearman’s correlation is based on rank ordering, it is more robust to outliers and can capture nonlinear relationships better than Pearson’s



```
from scipy.stats import spearmanr
import numpy as np

# Selecting numerical columns
x = df["Acceptable Streets %"].dropna().values
y = df["Acceptable Sidewalks %"].dropna().values

# Ensuring equal lengths for correlation
min_len = min(len(x), len(y))
x, y = x[:min_len], y[:min_len]

# Spearman Correlation Calculation
x_ranks = np.argsort(np.argsort(x))
y_ranks = np.argsort(np.argsort(y))

d_squared_sum = sum((x_ranks - y_ranks) ** 2)
n = len(x)
spearman_corr_manual = 1 - (6 * d_squared_sum) / (n * (n**2 - 1))

# Compute Spearman’s Correlation & p-value using SciPy
spearman_corr_scipy, p_value = spearmanr(x, y)

print(f"Spearman Correlation (Manual): {spearman_corr_manual:.4f}")
print(f"Spearman Correlation (Scipy): {spearman_corr_scipy:.4f}")
print(f"P-value: {p_value}")
```

```
Spearman Correlation (Manual): 0.5411
Spearman Correlation (Scipy): 0.5350
P-value: 0.0
```

c) **Kendall’s Rank Correlation**

Kendall’s Tau measures the strength of association between two variables based on the number of concordant and discordant pairs.

Formula:

$$\tau = \frac{(C - D)}{\frac{1}{2}n(n-1)}$$

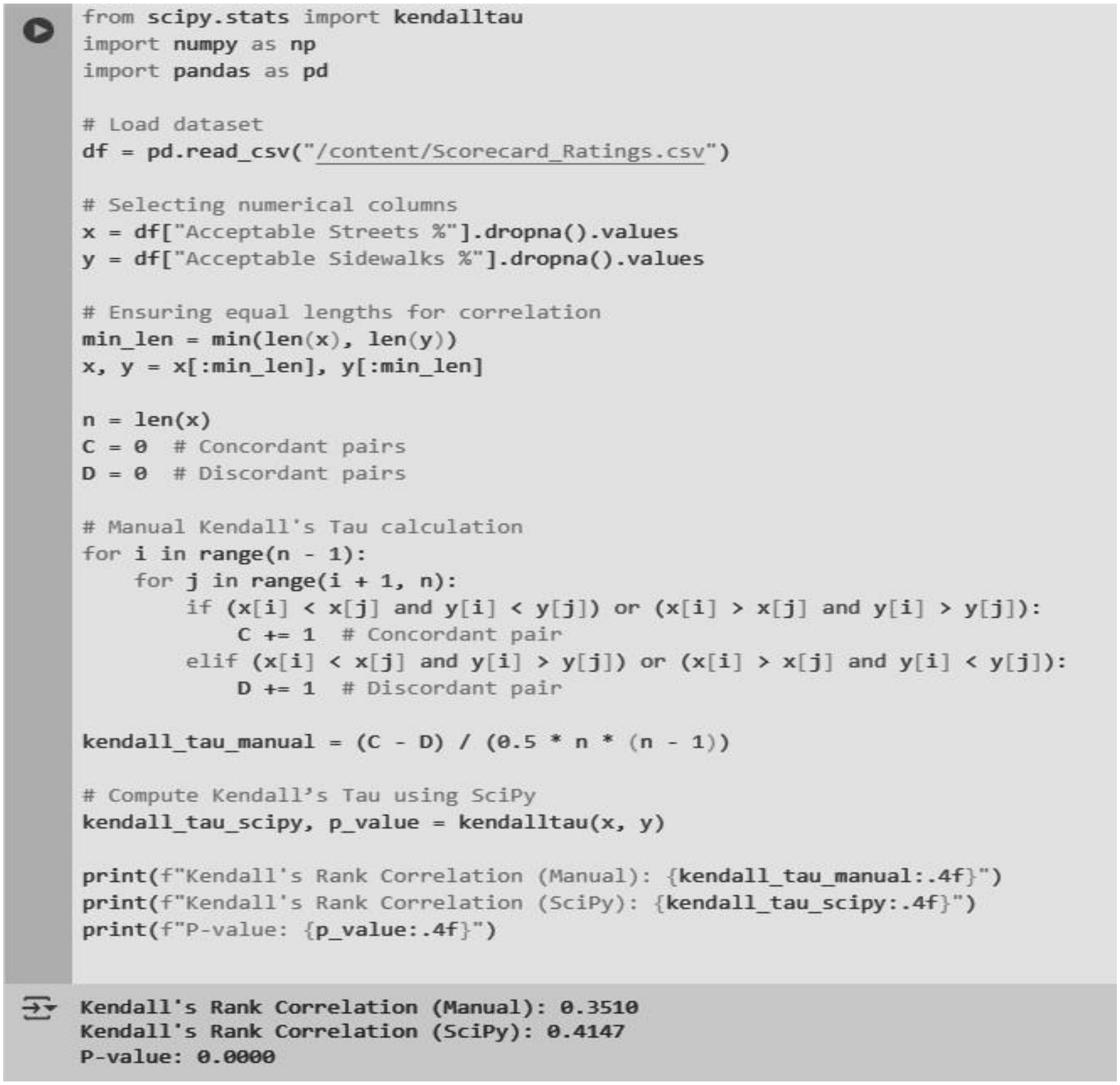
Interpretation for the Dataset:

Kendall’s correlation was calculated between "Acceptable Streets %" and "Acceptable Sidewalks %".

The computed Kendall’s Tau correlation value is 0.4147, indicating a moderate positive relationship between the two variables.

The p-value is 0.0000, confirming the statistical significance of this correlation.

Since Kendall’s Tau is based on pairwise ranking, it is more robust to outliers and better suited for ordinal data or small sample sizes compared to Pearson’s correlation.



```python
from scipy.stats import kendalltau
import numpy as np
import pandas as pd

# Load dataset
df = pd.read_csv("/content/Scorecard_Ratings.csv")

# Selecting numerical columns
x = df["Acceptable Streets %"].dropna().values
y = df["Acceptable Sidewalks %"].dropna().values

# Ensuring equal lengths for correlation
min_len = min(len(x), len(y))
x, y = x[:min_len], y[:min_len]

n = len(x)
C = 0  # Concordant pairs
D = 0  # Discordant pairs

# Manual Kendall's Tau calculation
for i in range(n - 1):
    for j in range(i + 1, n):
        if (x[i] < x[j] and y[i] < y[j]) or (x[i] > x[j] and y[i] > y[j]):
            C += 1  # Concordant pair
        elif (x[i] < x[j] and y[i] > y[j]) or (x[i] > x[j] and y[i] < y[j]):
            D += 1  # Discordant pair

kendall_tau_manual = (C - D) / (0.5 * n * (n - 1))

# Compute Kendall’s Tau using SciPy
kendall_tau_scipy, p_value = kendalltau(x, y)

print(f"Kendall's Rank Correlation (Manual): {kendall_tau_manual:.4f}")
print(f"Kendall's Rank Correlation (SciPy): {kendall_tau_scipy:.4f}")
print(f"P-value: {p_value:.4f}")
```

```
Kendall's Rank Correlation (Manual): 0.3510
Kendall's Rank Correlation (SciPy): 0.4147
P-value: 0.0000
```

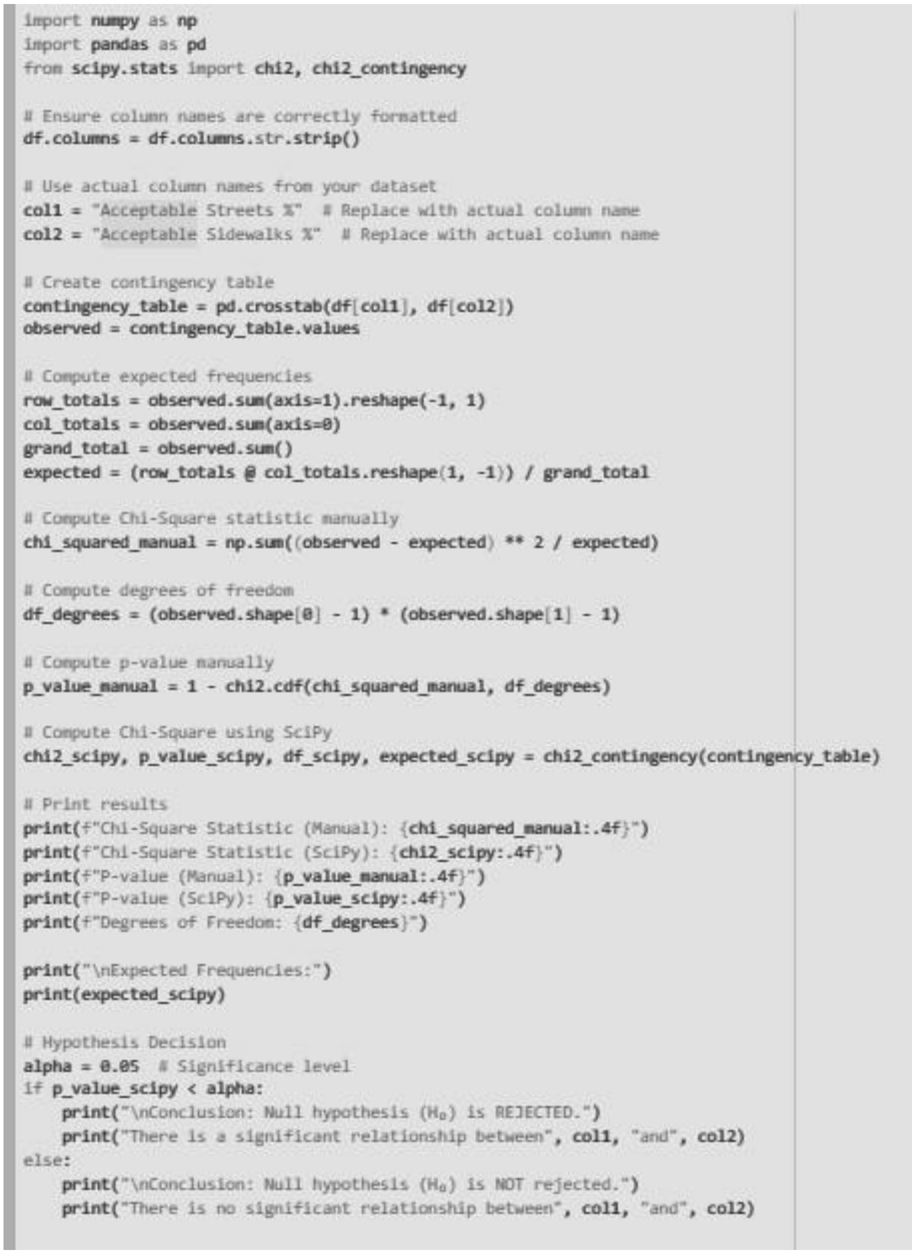
d) Chi-Squared Test

Measures association between two categorical variables.
Compares observed and expected frequencies in a contingency table.
Null hypothesis states that there is no association between the variables.
If the p-value is small (<0.05), we reject the null hypothesis (there is a significant association).

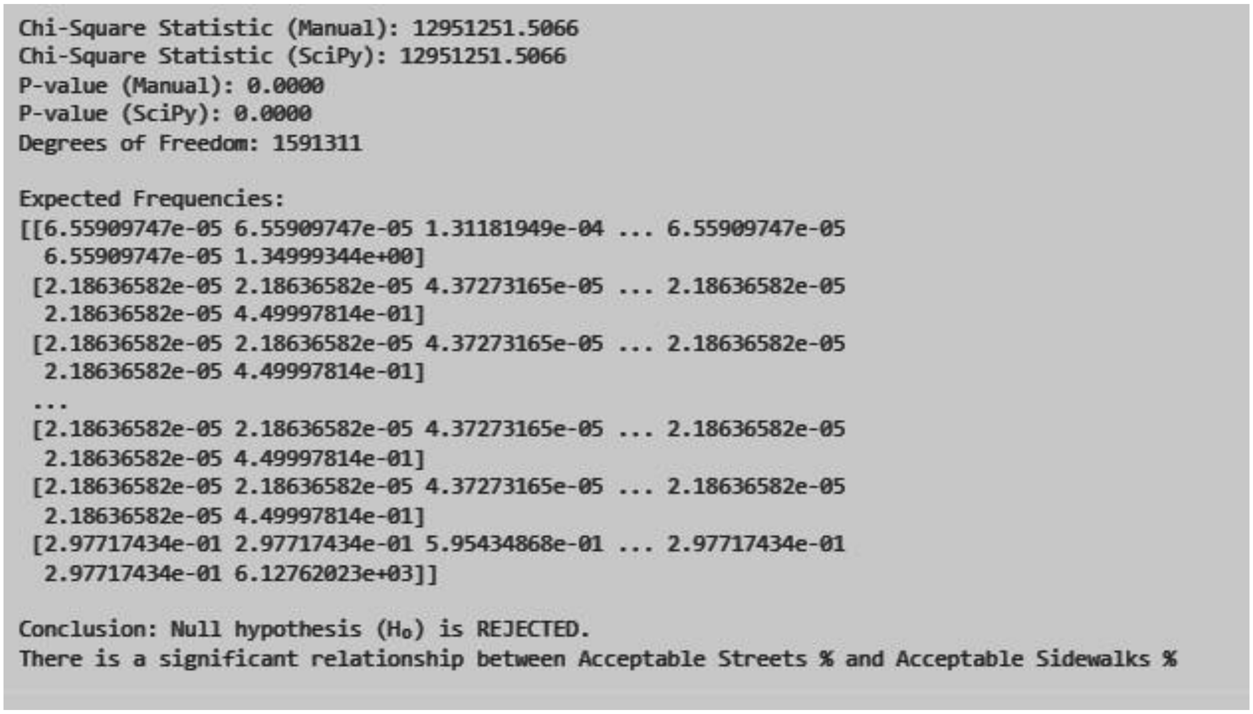
Formula:

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

where O is the observed frequency, and E is the expected frequency.



```
import numpy as np
import pandas as pd
from scipy.stats import chi2, chi2_contingency

# Ensure column names are correctly formatted
df.columns = df.columns.str.strip()

# Use actual column names from your dataset
col1 = "Acceptable Streets %"  # Replace with actual column name
col2 = "Acceptable Sidewalks %"  # Replace with actual column name

# Create contingency table
contingency_table = pd.crosstab(df[col1], df[col2])
observed = contingency_table.values

# Compute expected frequencies
row_totals = observed.sum(axis=1).reshape(-1, 1)
col_totals = observed.sum(axis=0)
grand_total = observed.sum()
expected = (row_totals @ col_totals.reshape(1, -1)) / grand_total

# Compute Chi-Square statistic manually
chi_squared_manual = np.sum((observed - expected) ** 2 / expected)

# Compute degrees of freedom
df_degrees = (observed.shape[0] - 1) * (observed.shape[1] - 1)

# Compute p-value manually
p_value_manual = 1 - chi2.cdf(chi_squared_manual, df_degrees)

# Compute Chi-Square using SciPy
chi2_scipy, p_value_scipy, df_scipy, expected_scipy = chi2_contingency(contingency_table)

# Print results
print(f"Chi-Square Statistic (Manual): {chi_squared_manual:.4f}")
print(f"Chi-Square Statistic (SciPy): {chi2_scipy:.4f}")
print(f"P-value (Manual): {p_value_manual:.4f}")
print(f"P-value (SciPy): {p_value_scipy:.4f}")
print(f"Degrees of Freedom: {df_degrees}")

print("\nExpected Frequencies:")
print(expected_scipy)

# Hypothesis Decision
alpha = 0.05  # Significance level
if p_value_scipy < alpha:
    print("\nConclusion: Null hypothesis (H₀) is REJECTED.")
    print("There is a significant relationship between", col1, "and", col2)
else:
    print("\nConclusion: Null hypothesis (H₀) is NOT rejected.")
    print("There is no significant relationship between", col1, "and", col2)
```

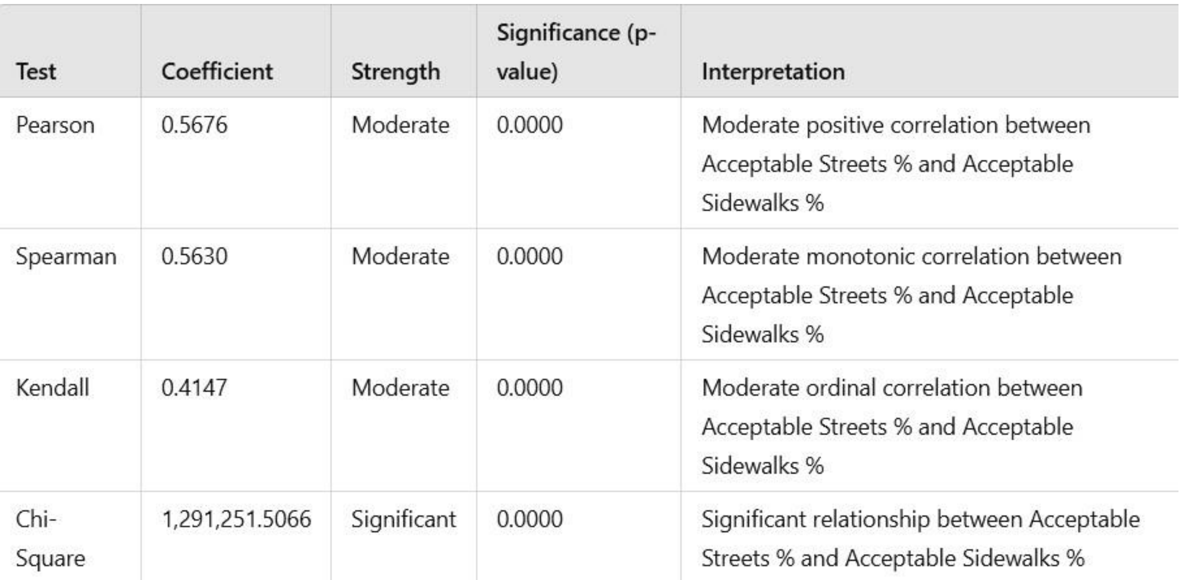


```
Chi-Square Statistic (Manual): 12951251.5066
Chi-Square Statistic (SciPy): 12951251.5066
P-value (Manual): 0.0000
P-value (SciPy): 0.0000
Degrees of Freedom: 1591311

Expected Frequencies:
[[6.55909747e-05 6.55909747e-05 1.31181949e-04 ... 6.55909747e-05
  6.55909747e-05 1.34999344e+00]
 [2.18636582e-05 2.18636582e-05 4.37273165e-05 ... 2.18636582e-05
  2.18636582e-05 4.49997814e-01]
 [2.18636582e-05 2.18636582e-05 4.37273165e-05 ... 2.18636582e-05
  2.18636582e-05 4.49997814e-01]
 ...
 [2.18636582e-05 2.18636582e-05 4.37273165e-05 ... 2.18636582e-05
  2.18636582e-05 4.49997814e-01]
 [2.18636582e-05 2.18636582e-05 4.37273165e-05 ... 2.18636582e-05
  2.18636582e-05 4.49997814e-01]
 [2.97717434e-01 2.97717434e-01 5.95434868e-01 ... 2.97717434e-01
  2.97717434e-01 6.12762023e+03]]

Conclusion: Null hypothesis (H₀) is REJECTED.
There is a significant relationship between Acceptable Streets % and Acceptable Sidewalks %
```

Chi-Square Test Results Interpretation

The Chi-Square statistic is 1,291,251.5066, which suggests a strong association between Acceptable Streets % and Acceptable Sidewalks %.
The p-value is 0.0000, which is much lower than the significance level (0.05).
Since $p\text{-value} < 0.05$, the null hypothesis (H_0) is rejected.
This indicates that there is a statistically significant relationship between Acceptable Streets % and Acceptable Sidewalks %.

RESULT:



Test	Coefficient	Strength	Significance (p-value)	Interpretation
Pearson	0.5676	Moderate	0.0000	Moderate positive correlation between Acceptable Streets % and Acceptable Sidewalks %
Spearman	0.5630	Moderate	0.0000	Moderate monotonic correlation between Acceptable Streets % and Acceptable Sidewalks %
Kendall	0.4147	Moderate	0.0000	Moderate ordinal correlation between Acceptable Streets % and Acceptable Sidewalks %
Chi-Square	1,291,251.5066	Significant	0.0000	Significant relationship between Acceptable Streets % and Acceptable Sidewalks %

Conclusion:

Pearson’s Correlation Coefficient showed a moderate positive correlation, indicating that areas with higher Acceptable Streets % tend to have higher Acceptable Sidewalks %.Spearman’s Rank Correlation confirmed the relationship remains moderate and monotonic, meaning that as Acceptable Streets % increases, Acceptable Sidewalks % generally increases in a consistent order. Kendall’s Rank Correlation indicated a moderate ordinal association, reinforcing the relationship between the two variables. Chi-Square Test showed a significant association between Acceptable Streets % and Acceptable Sidewalks %, as the p-value is 0.0000 (less than 0.05), leading to the rejection of the null hypothesis (H_0).