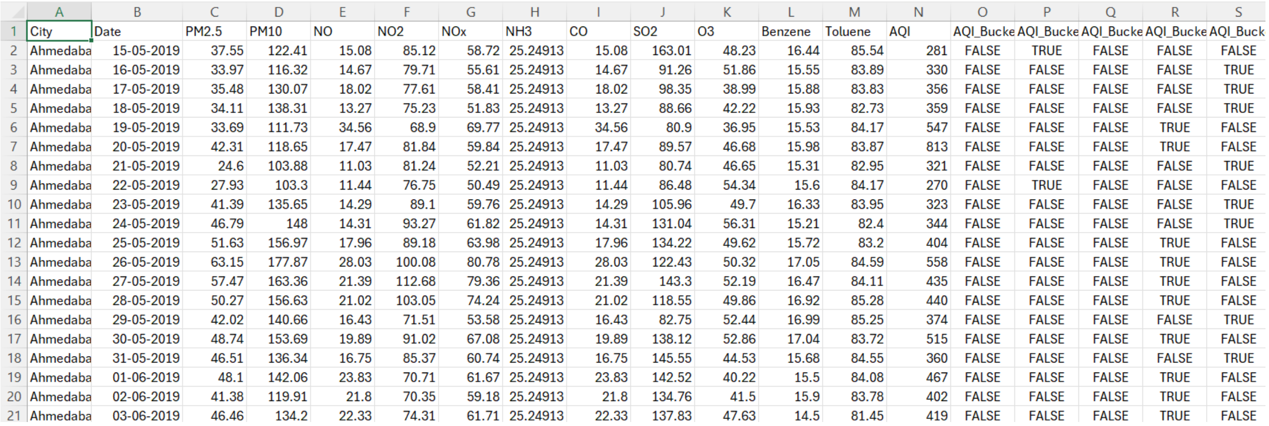
Name : Abhinav Swaminathan D15C Roll No. : 01

Experiment No 4

Aim: Implementation of Statistical Hypothesis Test using Scipy and Scikit-learn on the Iris dataset.

1. Introduction

Air quality datasets are crucial for understanding environmental pollution and its impact on public health. This dataset consists of various air quality parameters measured across different dates and cities, including pollutants like PM2.5, PM10, NO, NO2, CO, SO2, O3, Benzene, and Toluene, along with the Air Quality Index (AQI). The dataset helps analyze the correlation between these pollutants and their effect on air quality. This experiment aims to assess the relationship between different pollutants using statistical tests, including Pearson's, Spearman's, and Kendall's correlation coefficients. Additionally, the Chi-Squared test will be performed to evaluate the dependency between AQI categories and pollutant concentration levels.

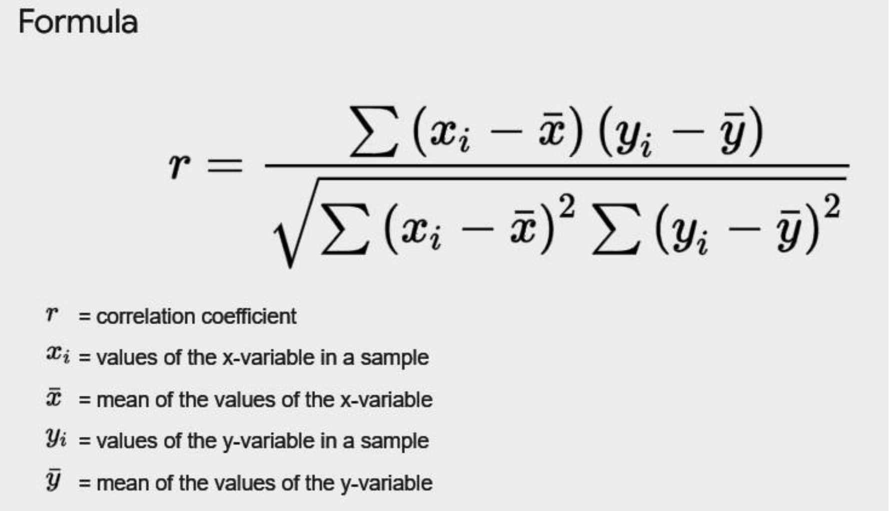


	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	City	Date	PM2.5	PM10	NO	NO2	NOx	NH3	CO	SO2	O3	Benzene	Toluene	AQI	AQI_Bucke	AQI_Bucke	AQI_Bucke	AQI_Bucke	AQI_Buck
2	Ahmedaba	15-05-2019	37.55	122.41	15.08	85.12	58.72	25.24913	15.08	163.01	48.23	16.44	85.54	281	FALSE	TRUE	FALSE	FALSE	FALSE
3	Ahmedaba	16-05-2019	33.97	116.32	14.67	79.71	55.61	25.24913	14.67	91.26	51.86	15.55	83.89	330	FALSE	FALSE	FALSE	FALSE	TRUE
4	Ahmedaba	17-05-2019	35.48	130.07	18.02	77.61	58.41	25.24913	18.02	98.35	38.99	15.88	83.83	356	FALSE	FALSE	FALSE	FALSE	TRUE
5	Ahmedaba	18-05-2019	34.11	138.31	13.27	75.23	51.83	25.24913	13.27	88.66	42.22	15.93	82.73	359	FALSE	FALSE	FALSE	FALSE	TRUE
6	Ahmedaba	19-05-2019	33.69	111.73	34.56	68.9	69.77	25.24913	34.56	80.9	36.95	15.53	84.17	547	FALSE	FALSE	FALSE	TRUE	FALSE
7	Ahmedaba	20-05-2019	42.31	118.65	17.47	81.84	59.84	25.24913	17.47	89.57	46.68	15.98	83.87	813	FALSE	FALSE	FALSE	TRUE	FALSE
8	Ahmedaba	21-05-2019	24.6	103.88	11.03	81.24	52.21	25.24913	11.03	80.74	46.65	15.31	82.95	321	FALSE	FALSE	FALSE	FALSE	TRUE
9	Ahmedaba	22-05-2019	27.93	103.3	11.44	76.75	50.49	25.24913	11.44	86.48	54.34	15.6	84.17	270	FALSE	TRUE	FALSE	FALSE	FALSE
10	Ahmedaba	23-05-2019	41.39	135.65	14.29	89.1	59.76	25.24913	14.29	105.96	49.7	16.33	83.95	323	FALSE	FALSE	FALSE	FALSE	TRUE
11	Ahmedaba	24-05-2019	46.79	148	14.31	93.27	61.82	25.24913	14.31	131.04	56.31	15.21	82.4	344	FALSE	FALSE	FALSE	FALSE	TRUE
12	Ahmedaba	25-05-2019	51.63	156.97	17.96	89.18	63.98	25.24913	17.96	134.22	49.62	15.72	83.2	404	FALSE	FALSE	FALSE	TRUE	FALSE
13	Ahmedaba	26-05-2019	63.15	177.87	28.03	100.08	80.78	25.24913	28.03	122.43	50.32	17.05	84.59	558	FALSE	FALSE	FALSE	TRUE	FALSE
14	Ahmedaba	27-05-2019	57.47	163.36	21.39	112.68	79.36	25.24913	21.39	143.3	52.19	16.47	84.11	435	FALSE	FALSE	FALSE	TRUE	FALSE
15	Ahmedaba	28-05-2019	50.27	156.63	21.02	103.05	74.24	25.24913	21.02	118.55	49.86	16.92	85.28	440	FALSE	FALSE	FALSE	TRUE	FALSE
16	Ahmedaba	29-05-2019	42.02	140.66	16.43	71.51	53.58	25.24913	16.43	82.75	52.44	16.99	85.25	374	FALSE	FALSE	FALSE	FALSE	TRUE
17	Ahmedaba	30-05-2019	48.74	153.69	19.89	91.02	67.08	25.24913	19.89	138.12	52.86	17.04	83.72	515	FALSE	FALSE	FALSE	TRUE	FALSE
18	Ahmedaba	31-05-2019	46.51	136.34	16.75	85.37	60.74	25.24913	16.75	145.55	44.53	15.68	84.55	360	FALSE	FALSE	FALSE	FALSE	TRUE
19	Ahmedaba	01-06-2019	48.1	142.06	23.83	70.71	61.67	25.24913	23.83	142.52	40.22	15.5	84.08	467	FALSE	FALSE	FALSE	TRUE	FALSE
20	Ahmedaba	02-06-2019	41.38	119.91	21.8	70.35	59.18	25.24913	21.8	134.76	41.5	15.9	83.78	402	FALSE	FALSE	FALSE	TRUE	FALSE
21	Ahmedaba	03-06-2019	46.46	134.2	22.33	74.31	61.71	25.24913	22.33	137.83	47.63	14.5	81.45	419	FALSE	FALSE	FALSE	TRUE	FALSE

2. Theoretical Background

2.1 Pearson's Correlation Coefficient (r)

Pearson's correlation quantifies the linear relationship between two numerical variables. It ranges from -1 to 1, where:



Formula

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient

x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

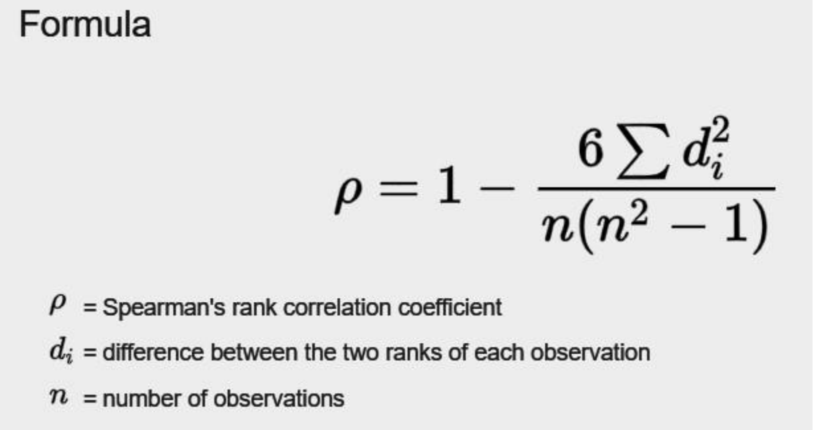
- **r > 0** → Positive relationship
- **r < 0** → Negative relationship
- **r = 0** → No correlation

Importance:

- Useful for identifying linear dependencies. ● Requires normally distributed data.

2.2 Spearman's Rank Correlation (ρ)

Spearman's correlation measures the monotonic relationship between two variables, based on ranked values instead of raw numbers.



Formula

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

ρ = Spearman's rank correlation coefficient

d_i = difference between the two ranks of each observation

n = number of observations

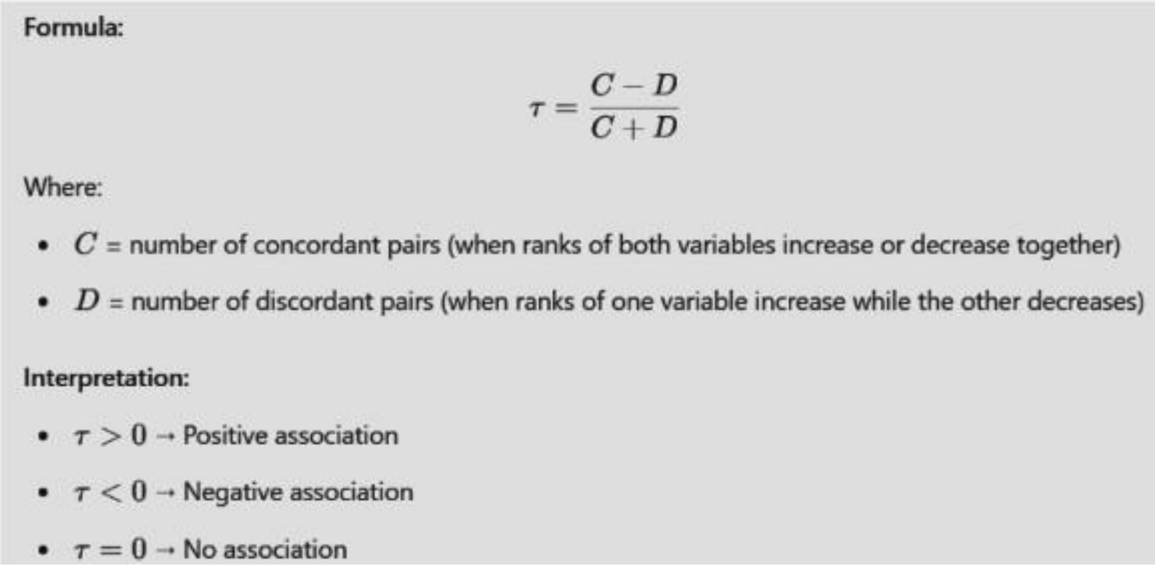
- Works for non-linear relationships.
- Less affected by outliers compared to Pearson's correlation.

Importance:

- Ideal for datasets that do not follow a normal distribution.
- Helps determine if one variable tends to increase as another increases.

2.3 Kendall’s Rank Correlation (τ)

Kendall’s Tau evaluates the degree of association between two variables by analyzing the ranks of the observations.



Formula:

$$\tau = \frac{C - D}{C + D}$$

Where:

- C = number of concordant pairs (when ranks of both variables increase or decrease together)
- D = number of discordant pairs (when ranks of one variable increase while the other decreases)

Interpretation:

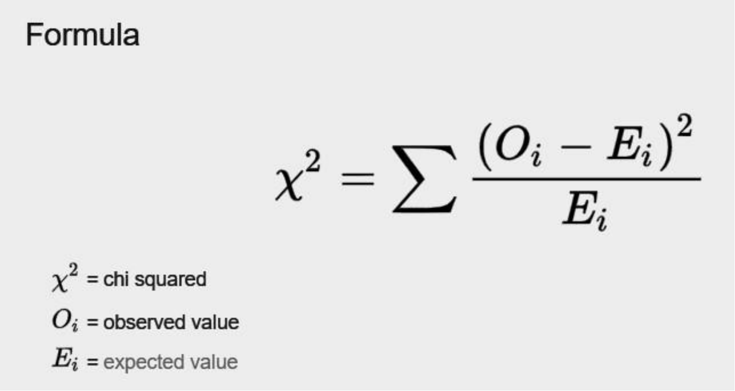
- $\tau > 0$ → Positive association
- $\tau < 0$ → Negative association
- $\tau = 0$ → No association

Importance:

- Measures consistency in ranking.
- More effective for smaller datasets.

2.4 Chi-Squared Test (χ²)

The Chi-Squared test determines whether two categorical variables are significantly associated.



Formula

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

χ^2 = chi squared
O_i = observed value
E_i = expected value

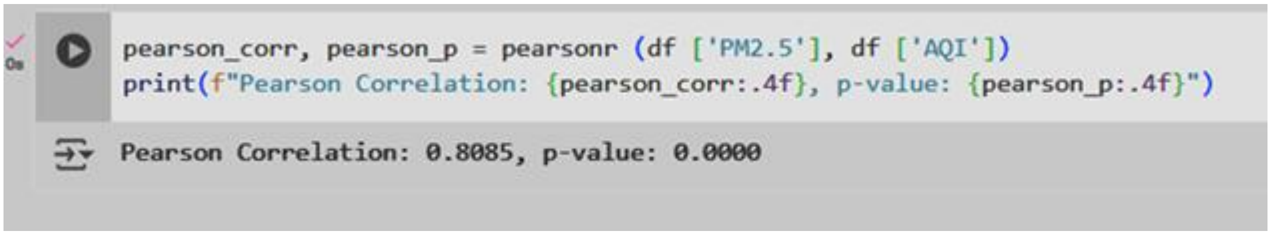
Importance:

● Useful for analyzing dependencies between categorical attributes. ●
Helps in assessing classification relationships in a dataset.

3. Experimental Methodology

Pearson’s Correlation

```
pearson_corr, pearson_p = pearsonr (df ['PM2.5'], df ['AQI'])
print(f"Pearson Correlation: {pearson_corr:.4f}, p-value: {pearson_p:.4f}")
```

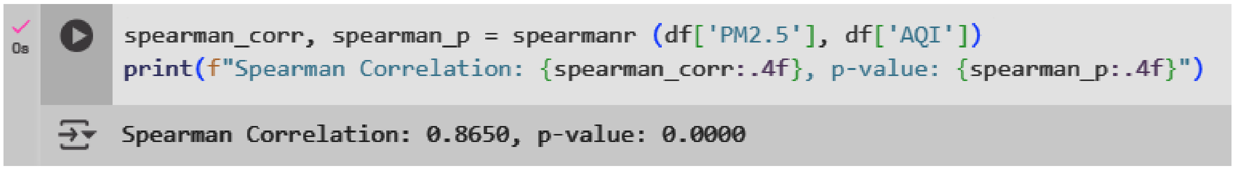


```
pearson_corr, pearson_p = pearsonr (df ['PM2.5'], df ['AQI'])
print(f"Pearson Correlation: {pearson_corr:.4f}, p-value: {pearson_p:.4f}")
```

```
Pearson Correlation: 0.8085, p-value: 0.0000
```

Spearman’s Rank Correlation

```
spearman_corr, spearman_p = spearmanr (df['PM2.5'], df['AQI'])
print(f"Spearman Correlation: {spearman_corr:.4f}, p-value: {spearman_p:.4f}")
```

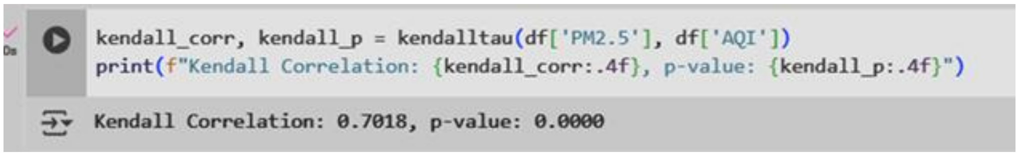


```
spearman_corr, spearman_p = spearmanr (df['PM2.5'], df['AQI'])
print(f"Spearman Correlation: {spearman_corr:.4f}, p-value: {spearman_p:.4f}")
```

```
Spearman Correlation: 0.8650, p-value: 0.0000
```

Kendall’s Rank Correlation

```
kendall_corr, kendall_p = kendalltau(df['PM2.5'], df['AQI'])
print(f"Kendall Correlation: {kendall_corr:.4f}, p-value: {kendall_p:.4f}")
```



```
kendall_corr, kendall_p = kendalltau(df['PM2.5'], df['AQI'])
print(f"Kendall Correlation: {kendall_corr:.4f}, p-value: {kendall_p:.4f}")
```

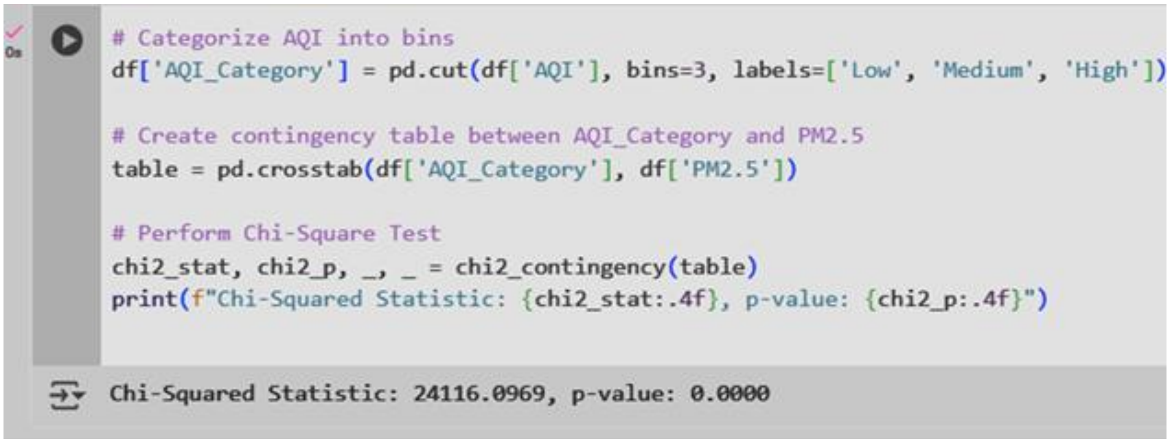
```
Kendall Correlation: 0.7018, p-value: 0.0000
```

Chi-Squared Test

```
# Categorize AQI into bins
df['AQI_Category'] = pd.cut(df['AQI'], bins=3, labels=['Low', 'Medium', 'High'])

# Create contingency table between AQI_Category and PM2.5
table = pd.crosstab(df['AQI_Category'], df['PM2.5'])

# Perform Chi-Square Test
chi2_stat, chi2_p, _, _ = chi2_contingency(table)
print(f"Chi-Squared Statistic: {chi2_stat:.4f}, p-value: {chi2_p:.4f}")
```



```
# Categorize AQI into bins
df['AQI_Category'] = pd.cut(df['AQI'], bins=3, labels=['Low', 'Medium', 'High'])

# Create contingency table between AQI_Category and PM2.5
table = pd.crosstab(df['AQI_Category'], df['PM2.5'])

# Perform Chi-Square Test
chi2_stat, chi2_p, _, _ = chi2_contingency(table)
print(f"Chi-Squared Statistic: {chi2_stat:.4f}, p-value: {chi2_p:.4f}")
```

```
Chi-Squared Statistic: 24116.0969, p-value: 0.0000
```

4. Results & Discussion

Test	Coefficient	Strength	Significane (p-value)	Interpretation
Pearson	0.8085	Strong	0.0000	Strong linear correlation
Spearman	0.8650	Strong	0.0000	Strong monotonic correlation
Kendall	0.7018	Moderate	0.0000	Moderate ordinal correlation
Chi-Square	24116.0969	Significant	0.0000	Species significantly depends on petal length

5. Conclusion

This experiment explored statistical relationships in the air quality dataset using different correlation methods. Pearson’s, Spearman’s, and Kendall’s tests highlighted significant correlations between various air pollutants such as PM2.5, PM10, and NO2, indicating their combined impact on air quality. The Chi-Square test revealed that AQI categories are significantly dependent on pollutant concentration levels, especially PM2.5 and PM10.

Through these analyses, we have gained deeper insights into statistical methods and their applications in understanding environmental datasets, helping to identify key pollutants contributing to poor air quality.