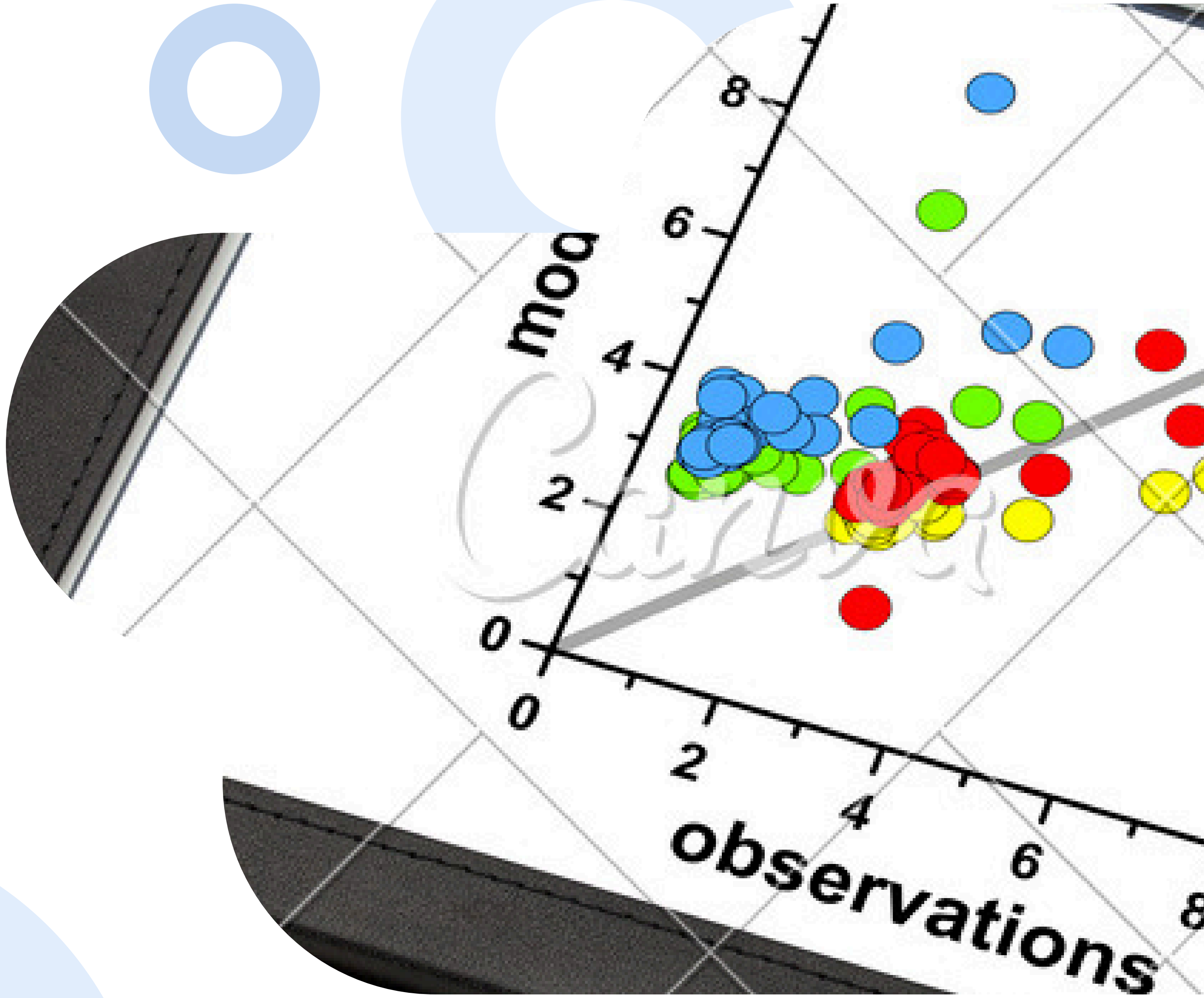
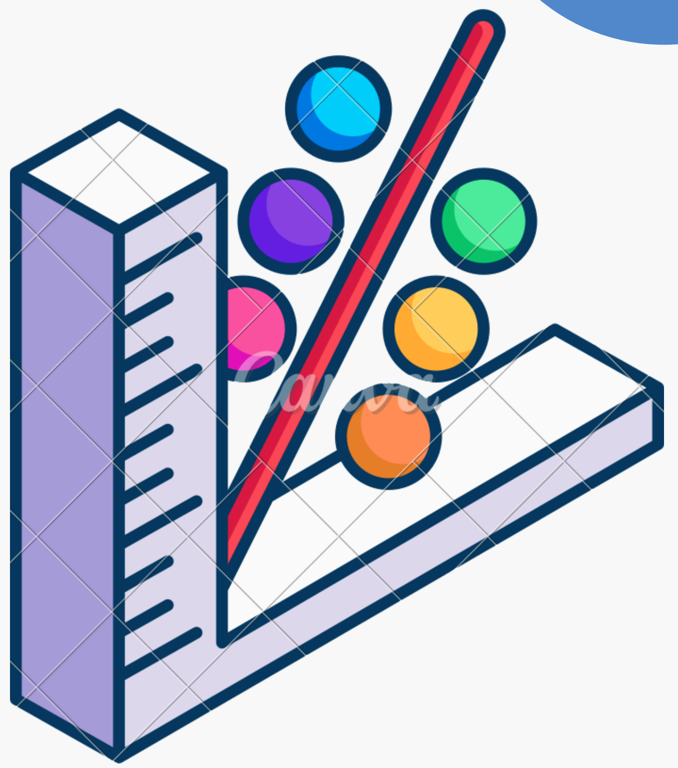
Covariance and Correlation Coefficient Calculator

“Interactive Python Tool for Statistical Analysis”

Presented by : Fatimah Alzarani
Supervised by : Dr.Amal Sulayman





Introduction

Covariance and correlation are fundamental concepts in statistics for understanding the relationship between two variables. While covariance measures how two variables vary together, the correlation coefficient normalizes this measure to provide a dimensionless value in the range of -1 to 1.

Objective

Visualize relationships

Visualize relationships between variables using scatter plots.

covariance and correlation

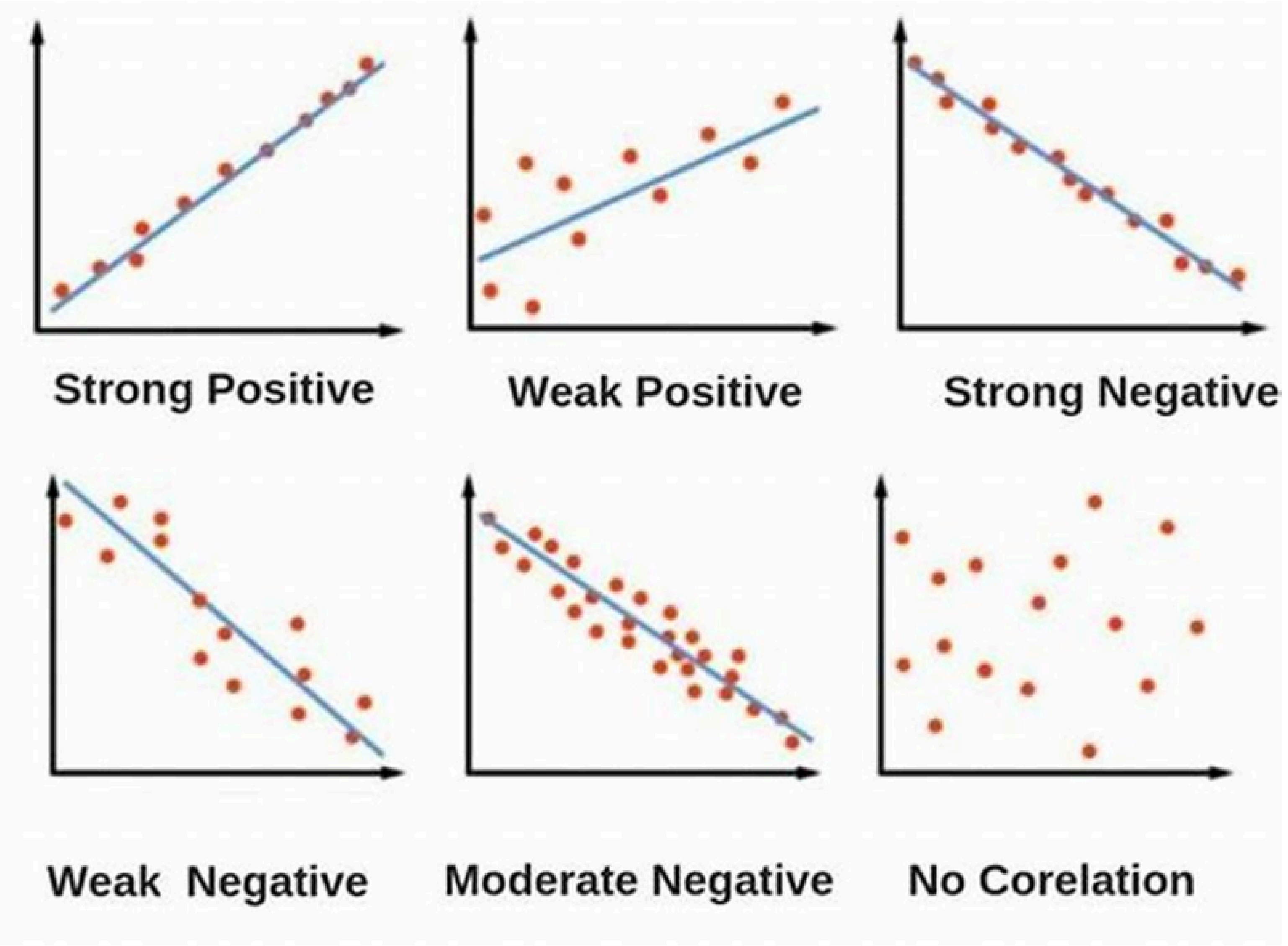
Learn how to calculate covariance and correlation coefficient.

interactive

Allow user interaction to explore different scenarios.

What Are Covariance and Correlation?

- **Covariance:** Measures how two variables vary together.
- **Correlation Coefficient:** Standardized measure (range: -1 to 1) of the relationship's strength and direction.
- **Positive values:** Variables move in the same direction.
- **Negative values:** Variables move in opposite directions.



Workflow

Generate data

Generate random data with a linear relationship and noise.

Calculate

Calculate covariance and correlation coefficient.

Visualize

Visualize the data using a scatter plot.

Code Breakdown

Generate Data

Creates two random variables with adjustable noise and slope.

```
[1]  import numpy as np

     def generate_data(seed=42, n=100, noise=0.5, slope=2):
         np.random.seed(seed)
         x = np.random.randn(n)
         y = slope * x + np.random.randn(n) * noise
         return x, y
```

Code Breakdown

Calculate Statistics

Computes covariance and correlation coefficient.

```
[2] def calculate_statistics(x, y):
        mean_x, mean_y = np.mean(x), np.mean(y)
        covariance = np.sum((x - mean_x) * (y - mean_y)) / (len(x) - 1)
        std_x, std_y = np.std(x, ddof=1), np.std(y, ddof=1)
        correlation = covariance / (std_x * std_y)
        return {
            "mean_x": mean_x,
            "mean_y": mean_y,
            "covariance": covariance,
            "std_x": std_x,
            "std_y": std_y,
            "correlation": correlation,
        }
```

Code Breakdown

Plot Data

Creates two random variables with adjustable noise and slope.

```
[3] import matplotlib.pyplot as plt

    def plot_data(x, y, stats):
        plt.figure(figsize=(10, 6))
        plt.scatter(x, y, alpha=0.7, label="📍 Data Points")
        plt.axhline(stats['mean_y'], color="red", linestyle="--", label="🔴 Mean Y")
        plt.axvline(stats['mean_x'], color="blue", linestyle="--", label="🔵 Mean X")
        plt.title(f"Scatter Plot 📊 (Correlation: {stats['correlation']:.4f})")
        plt.xlabel("X-axis ➡️")
        plt.ylabel("Y-axis ⬆️")
        plt.grid(True)
        plt.legend()
        plt.tight_layout()
        plt.show()
```

Code Breakdown

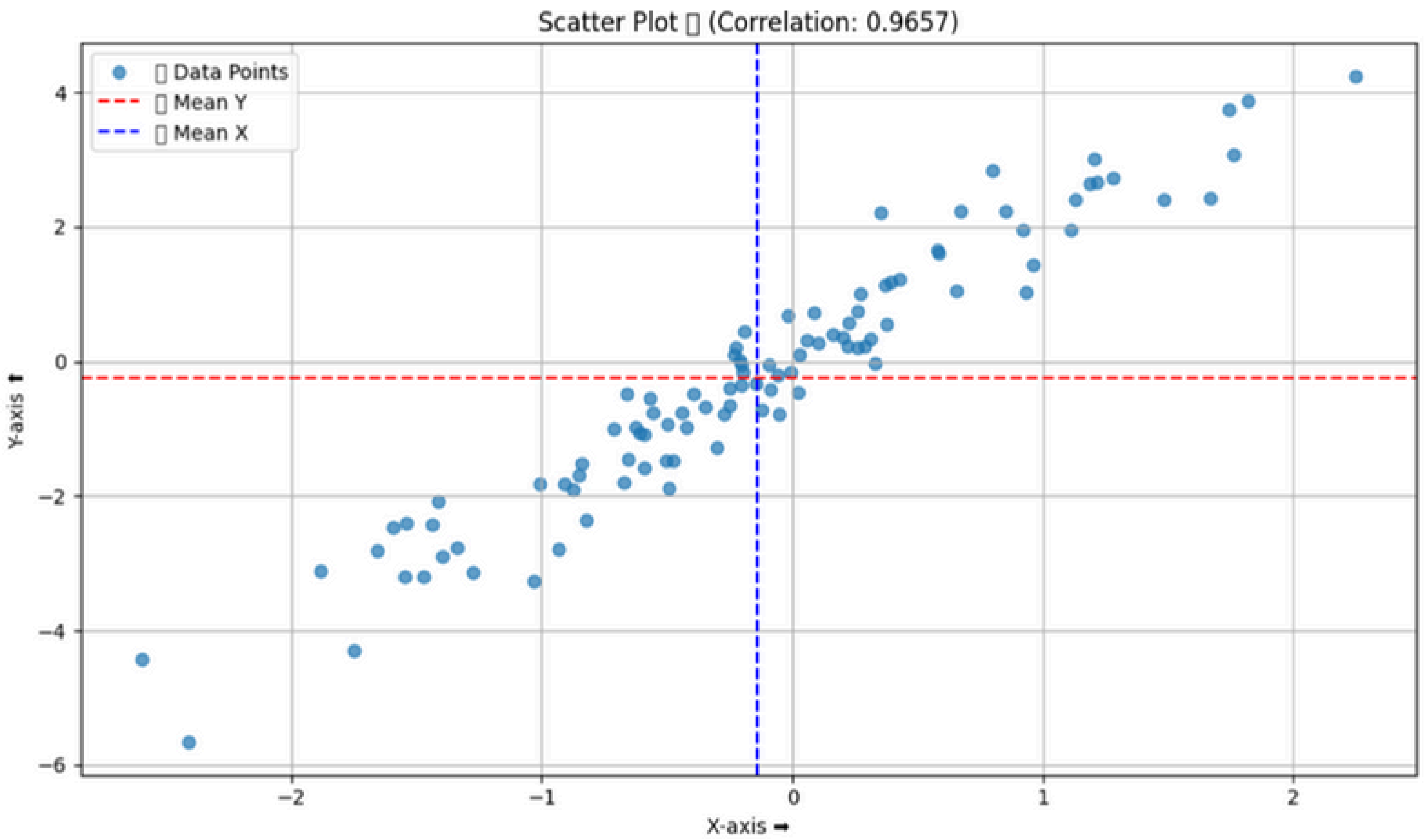
Main Program

interacts with the user to generate data, calculate statistics, and visualize results.

```
print("🌟 Welcome to the Covariance and Correlation Calculator 🌟")

n = int(input("🔢 Enter the number of data points (e.g., 100): "))
noise = float(input("🎛️ Enter the noise level (e.g., 0.5): "))
slope = float(input("📈 Enter the slope of the linear relationship (e.g., 2): "))
seed = int(input("🎲 Enter a random seed for reproducibility (e.g., 42): "))

x, y = generate_data(seed=seed, n=n, noise=noise, slope=slope)
stats = calculate_statistics(x, y)

print("\n📋 --Results:--")
print(f"   ➡️ Covariance: {stats['covariance']:.4f}")
print(f"   ➡️ Correlation Coefficient: {stats['correlation']:.4f}")

print("\n🔍 Visualizing your data...")
plot_data(x, y, stats)
```

Interactive Demo

```
🌟 Welcome to the Covariance and Correlation Calculator 🌟
🔢 Enter the number of data points (e.g., 100): 100
🔢 Enter the noise level (e.g., 0.5): 0.5
📈 Enter the slope of the linear relationship (e.g., 2): 2
🔮 Enter a random seed for reproducibility (e.g., 42): 45

📋 --Results:--
  ➡ Covariance: 1.7507
  ➡ Correlation Coefficient: 0.9657

🔍 Visualizing your data...
```



Enter the number of data points, noise level, slope, and seed.

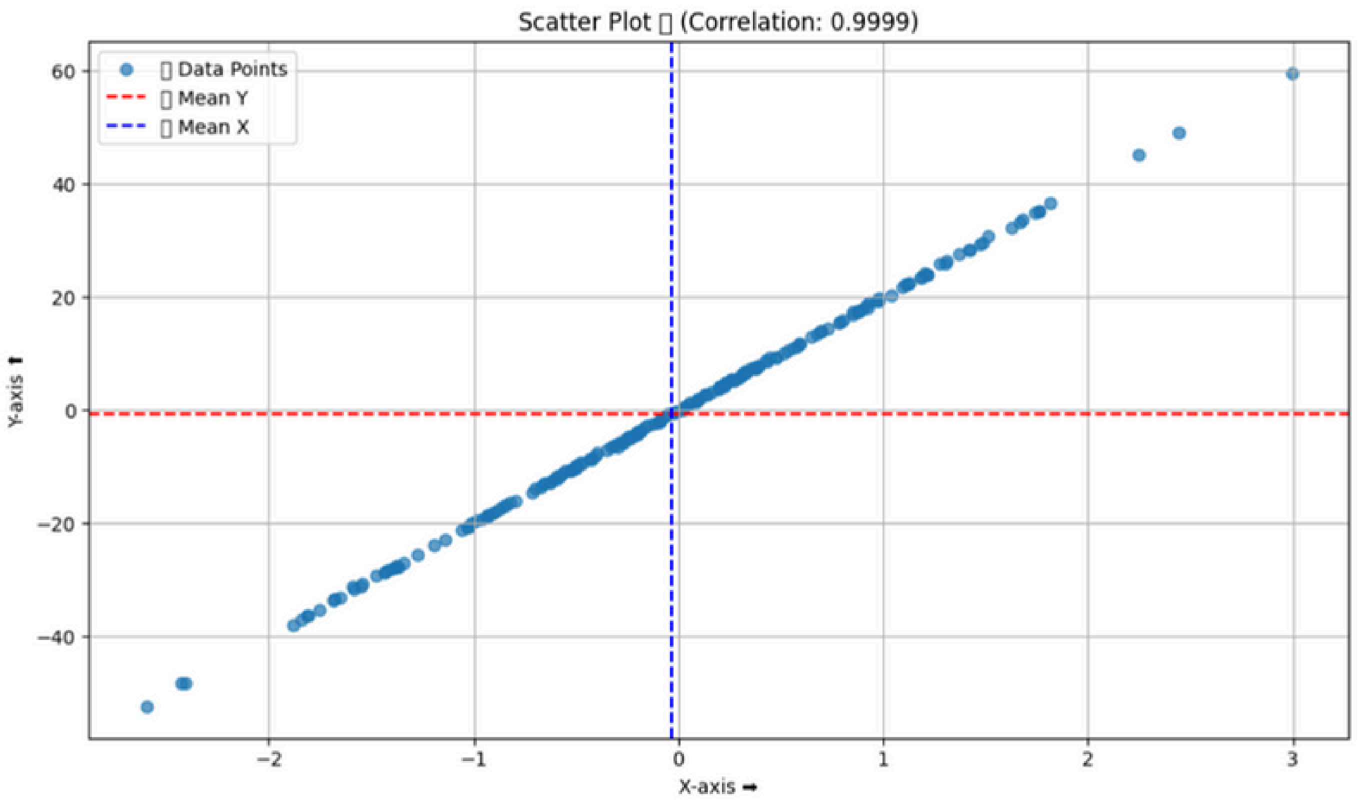
View results dynamically in both numerical and graphical form.

Explore how changes in input affect covariance and correlation.

Experiment With Relationships

Strong Positive Correlation

```
🌟 Welcome to the Covariance and Correlation Calculator 🌟
🔢 Enter the number of data points (e.g., 100): 200
🎛 Enter the noise level (e.g., 0.5): 0.2
📈 Enter the slope of the linear relationship (e.g., 2): 20
🎲 Enter a random seed for reproducibility (e.g., 42): 45

📋 --Results:--
  ➡ Covariance: 18.9964
  ➡ Correlation Coefficient: 0.9999

🔍 Visualizing your data...
```



Low Noise, High Slope

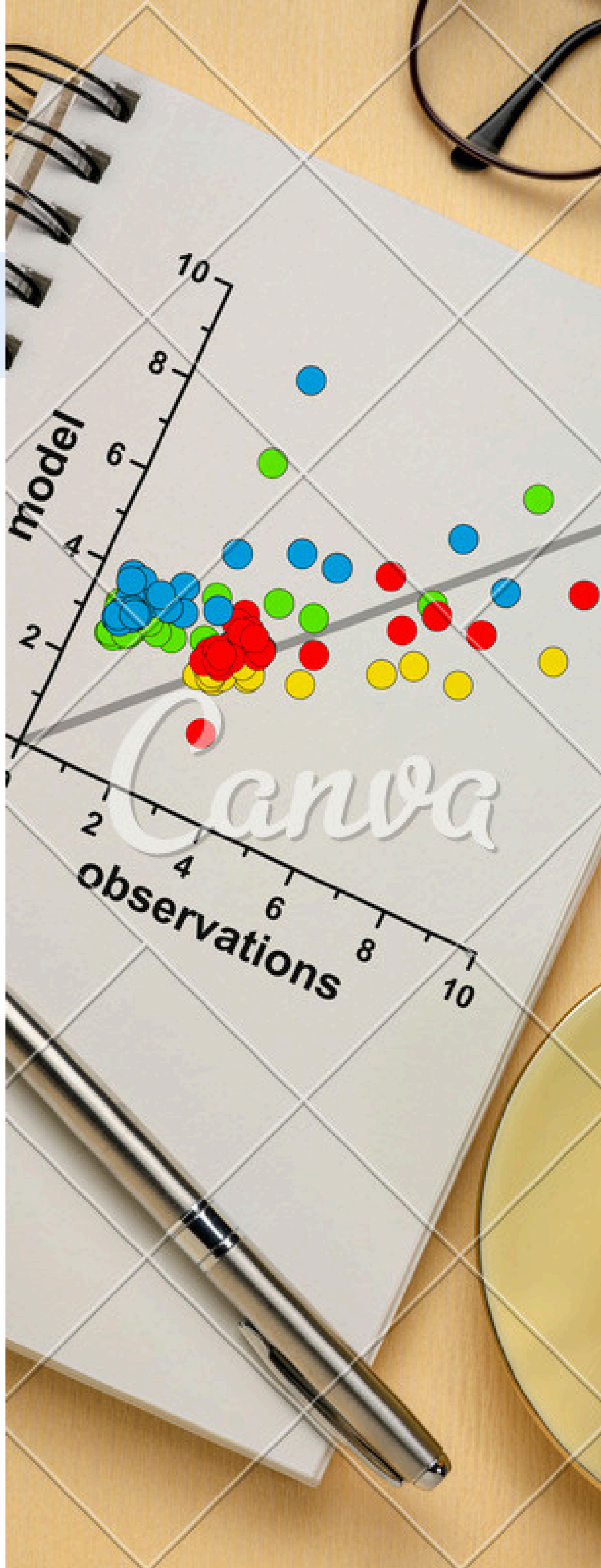
noise level : 0.2, slope : 20

Covariance

18.9964

Correlation

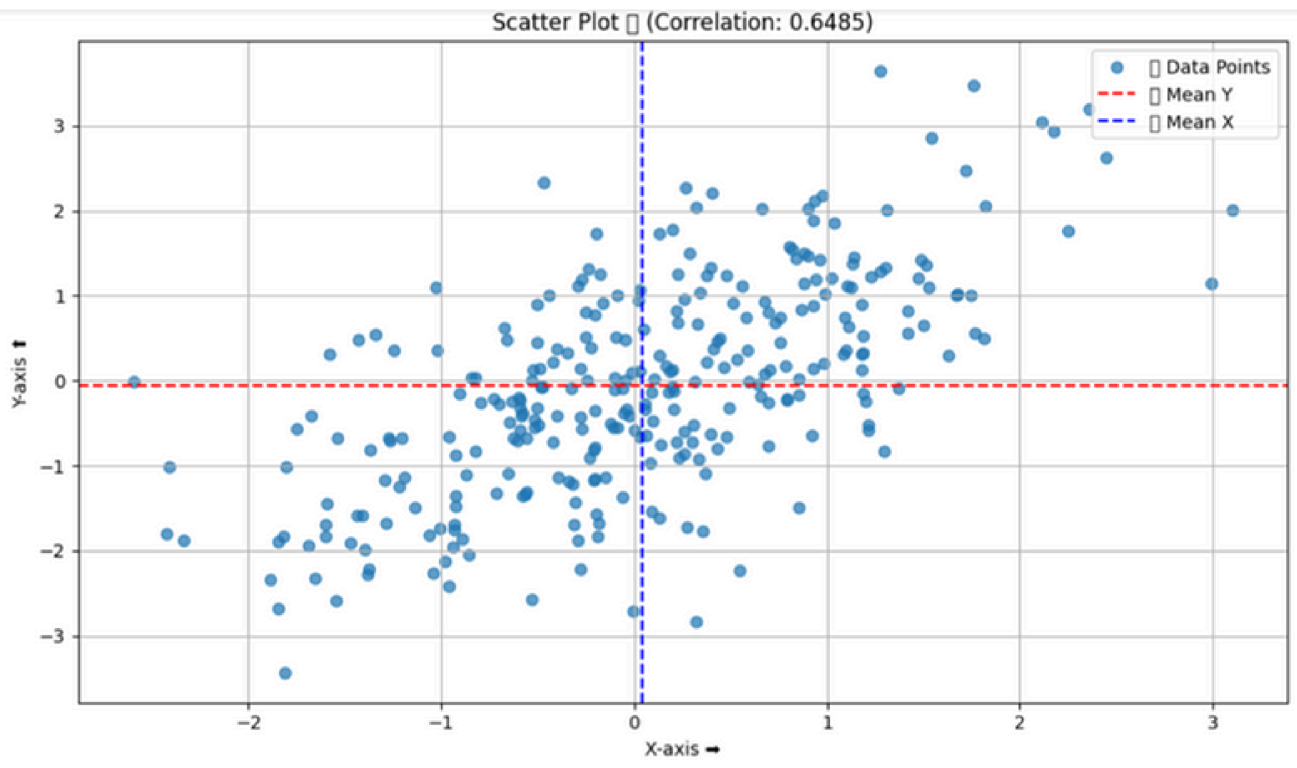
0.9999



Experiment With Relationships

Weak Positive Correlation

```
🌟 Welcome to the Covariance and Correlation Calculator 🌟
🔢 Enter the number of data points (e.g., 100): 300
🎛 Enter the noise level (e.g., 0.5): 1
📈 Enter the slope of the linear relationship (e.g., 2): 0.8
🎲 Enter a random seed for reproducibility (e.g., 42): 45

📋 --Results:--
   ➡ Covariance: 0.8058
   ➡ Correlation Coefficient: 0.6485

🔍 Visualizing your data...
```



High Noise, Low Slope

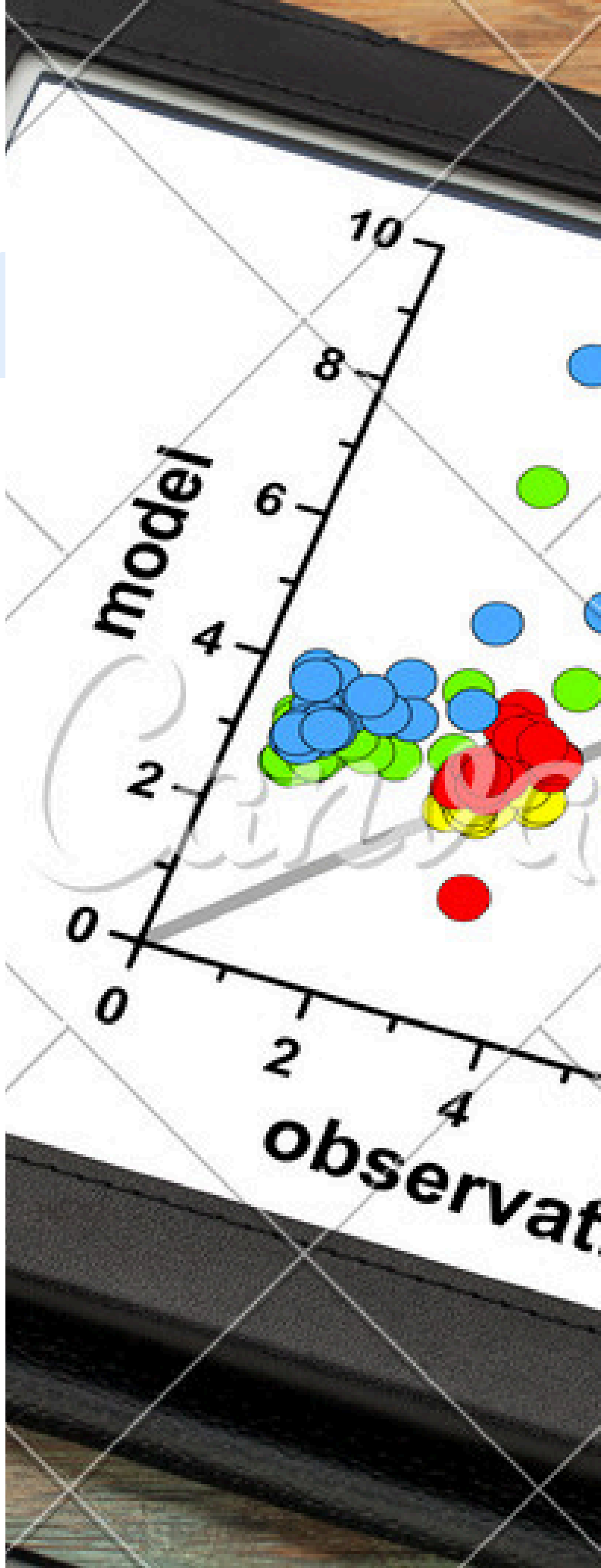
noise level : 1, slope : 0.8

Covariance

0.8058

Correlation

0.6485



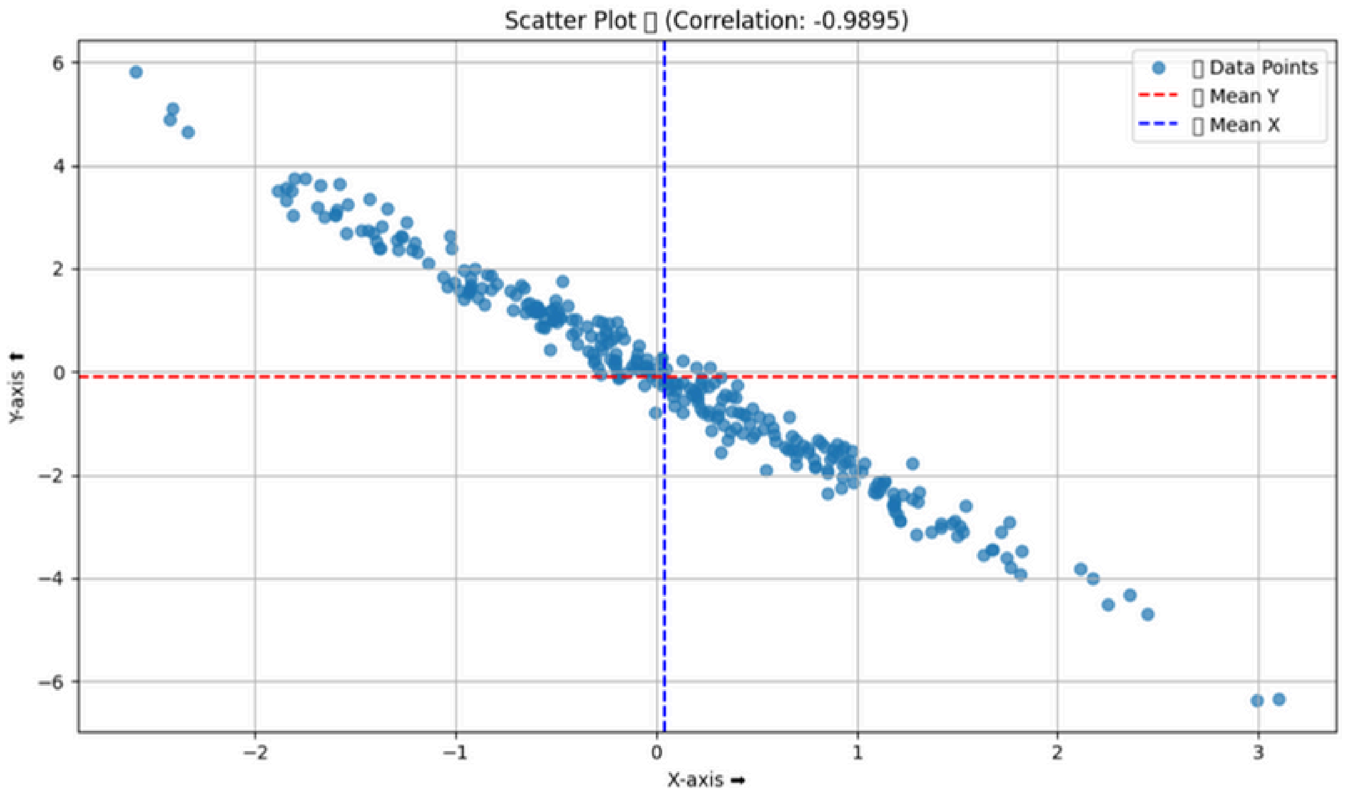
Experiment With Relationships

Strong Negative Correlation

☀️ Welcome to the Covariance and Correlation Calculator ☀️
🔢 Enter the number of data points (e.g., 100): 300
🎛 Enter the noise level (e.g., 0.5): 0.3
📈 Enter the slope of the linear relationship (e.g., 2): -2
🎲 Enter a random seed for reproducibility (e.g., 42): 45

📋 --Results:--
➡️ Covariance: -1.9449
➡️ Correlation Coefficient: -0.9895

🔍 Visualizing your data...



Low Noise, High Slope

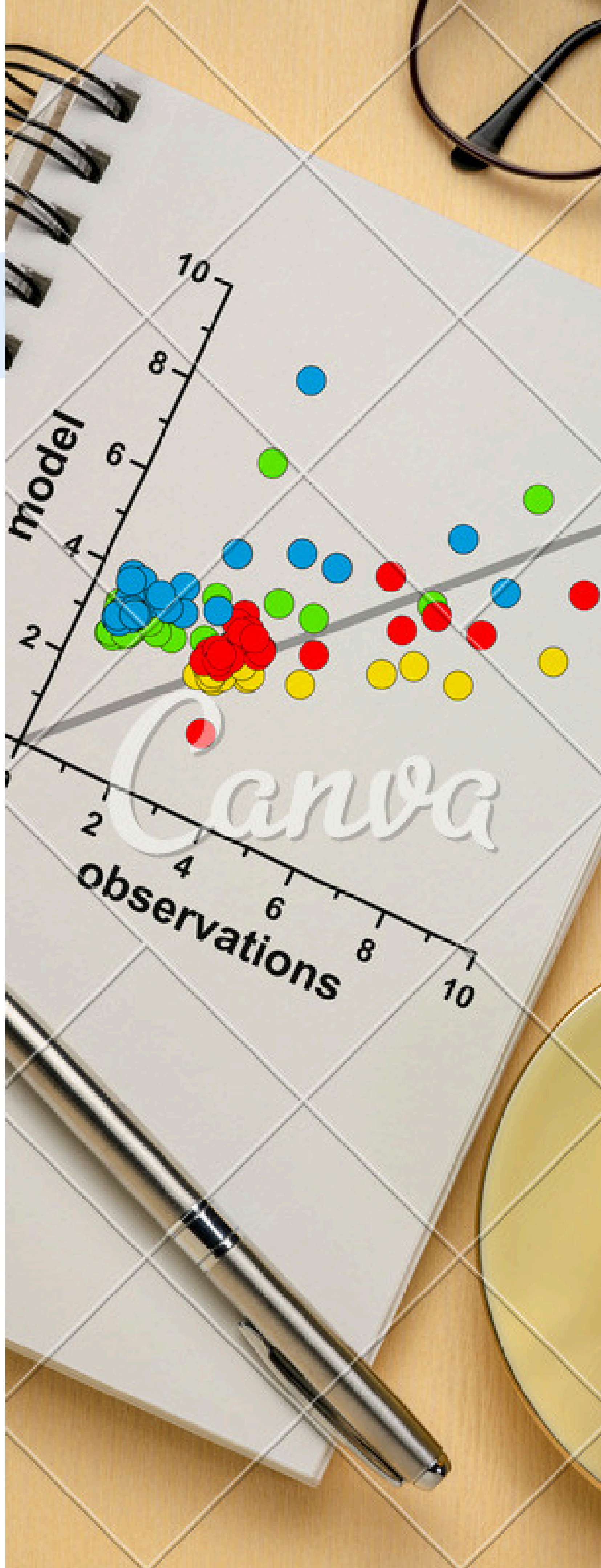
noise level : 0.3, slope : -2

Covariance

-1.9449

Correlation

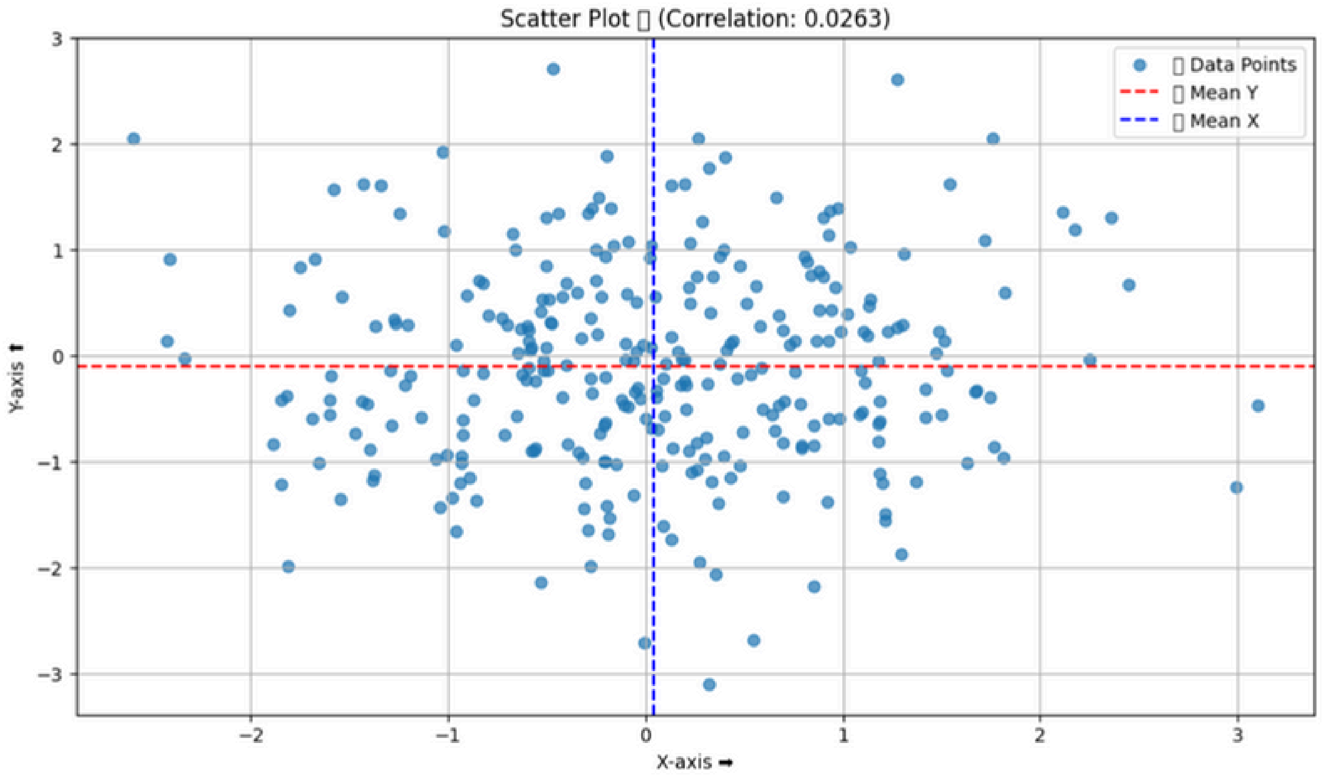
-0.9895



Experiment With Relationships

No Correlation

```
🌟 Welcome to the Covariance and Correlation Calculator 🌟
🔢 Enter the number of data points (e.g., 100): 300
🎛 Enter the noise level (e.g., 0.5): 1
📈 Enter the slope of the linear relationship (e.g., 2): 0
🎲 Enter a random seed for reproducibility (e.g., 42): 45

📋 --Results:--
   ➡ Covariance: 0.0249
   ➡ Correlation Coefficient: 0.0263

🔍 Visualizing your data...
```



High Noise, No Relationship

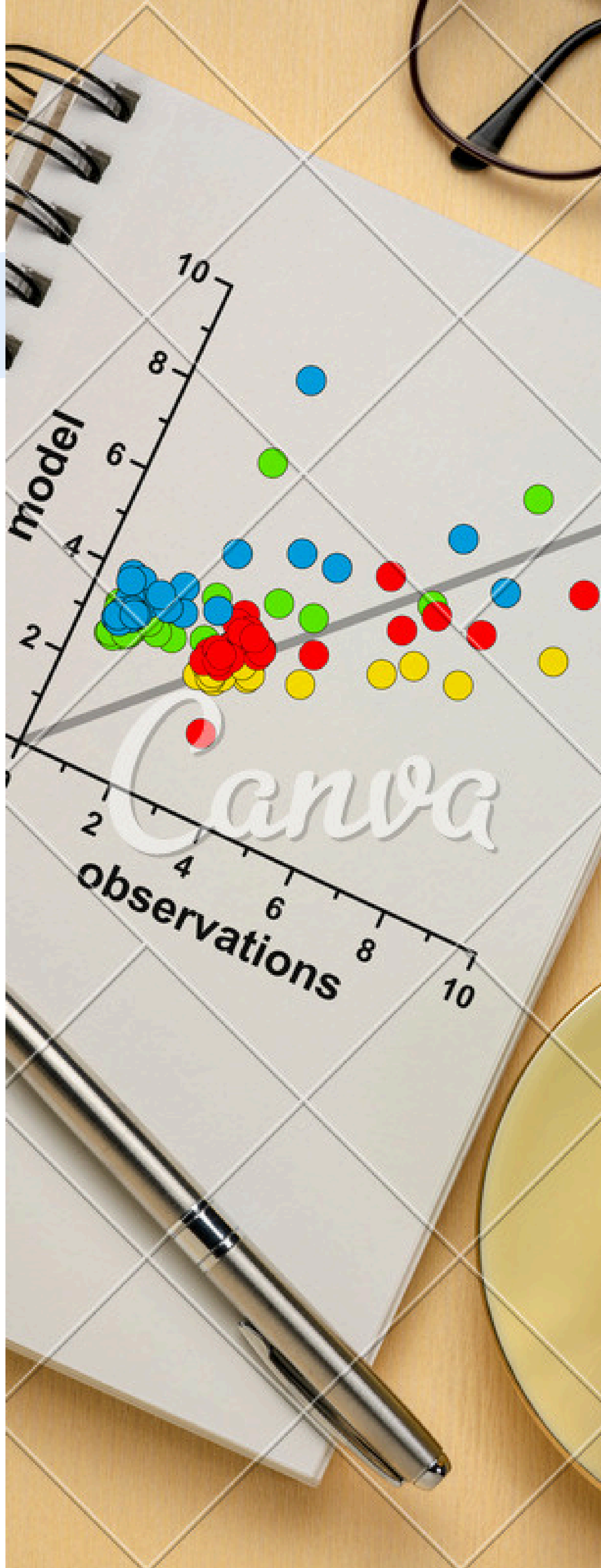
noise level : 1, slope : 0

Covariance

0.0249

Correlation

0.0263

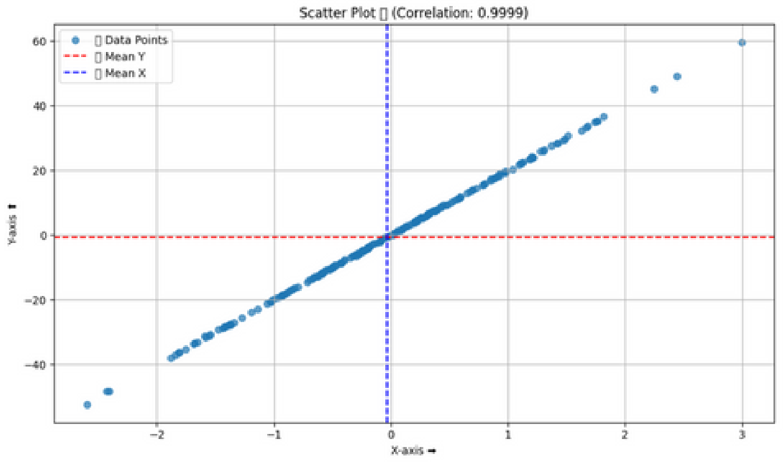


Questions

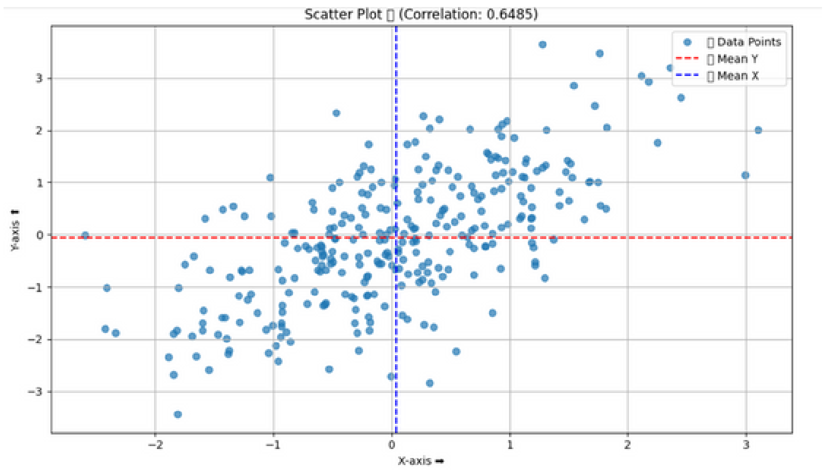
1. How does the calculated correlation coefficient relate to the visual pattern you observe in the scatter plot? Explain any discrepancies or confirmations you notice.

The correlation coefficient quantifies the strength and direction of the linear relationship between two variables, and this value is reflected in the scatter plot:

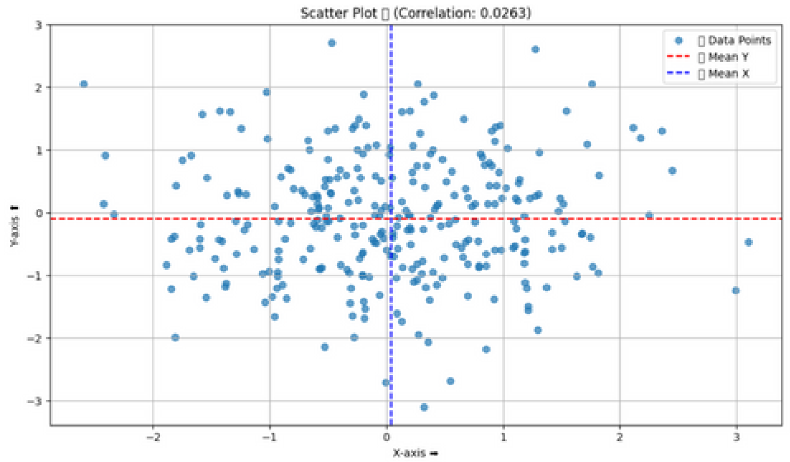
positive strong correlation coefficient



positive weak correlation coefficient



no correlation coefficient



Discrepancies: Small discrepancies may arise due to randomness or noise in the data. For instance, even in cases like Figure 3, where the correlation coefficient is moderate, noise can obscure the apparent linearity.

Questions

2. Compare the magnitude of the covariance with the correlation coefficient. Why might the correlation coefficient be more useful for interpreting the relationship between variables?

The scatter plots in Figures demonstrate how covariance and correlation coefficient differ:

- **Covariance:** Measures the joint variability of two variables.

- **Correlation Coefficient:** Standardizes covariance to produce a dimensionless measure.

The correlation coefficient is more useful because it normalizes covariance, allowing for direct comparison regardless of the units of the variables

Questions

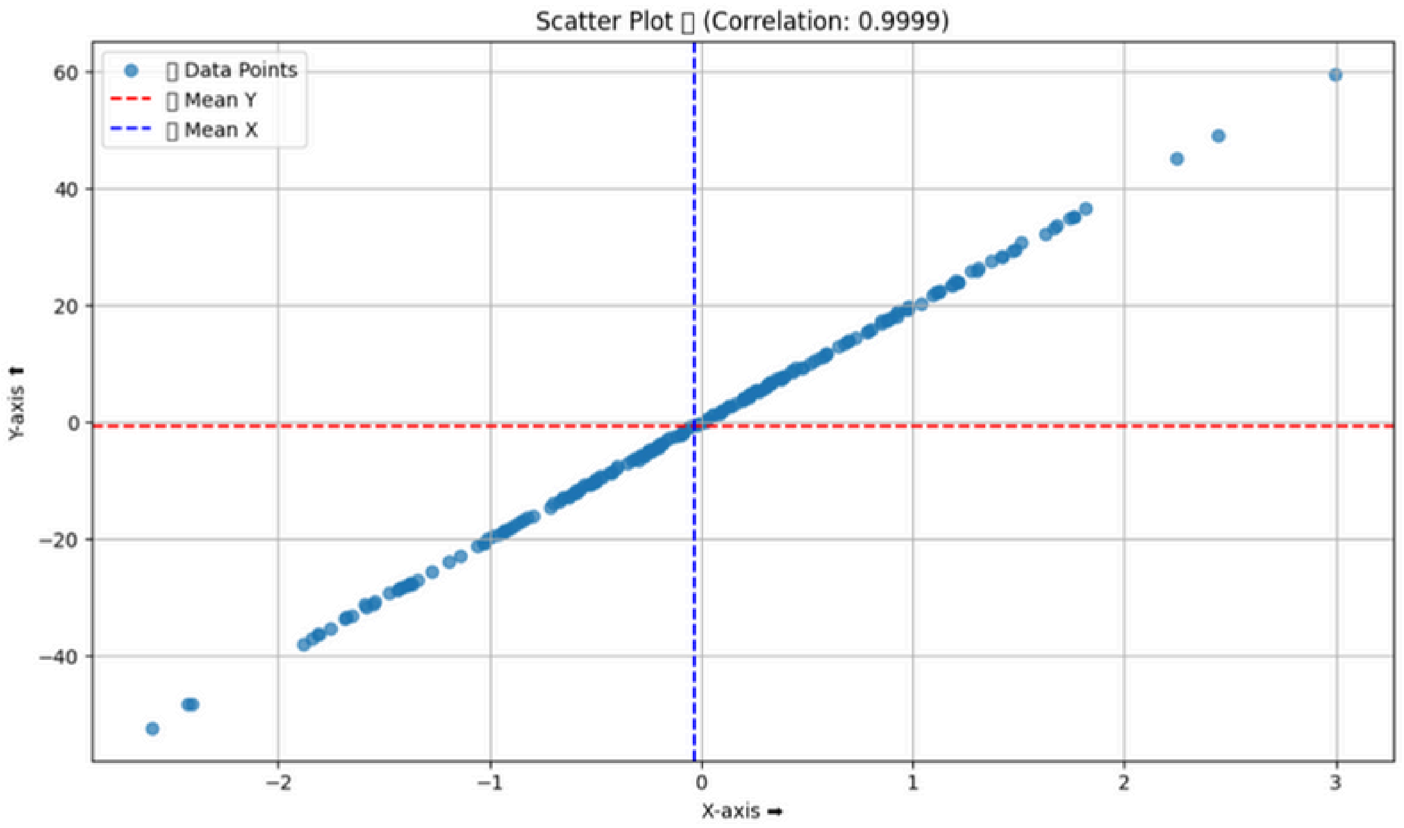
3. If you were to modify the script to generate data with a stronger or weaker relationship between X and Y, how would you expect the covariance, correlation coefficient, and scatter plot to change?

• **Stronger Relationship:**

– Covariance increases in magnitude, similar to what is observed in Figure .

– Correlation coefficient approaches 1 (positive) or -1 (negative).

– Scatter plot points align more closely along a straight line, as seen in Figure .



Questions

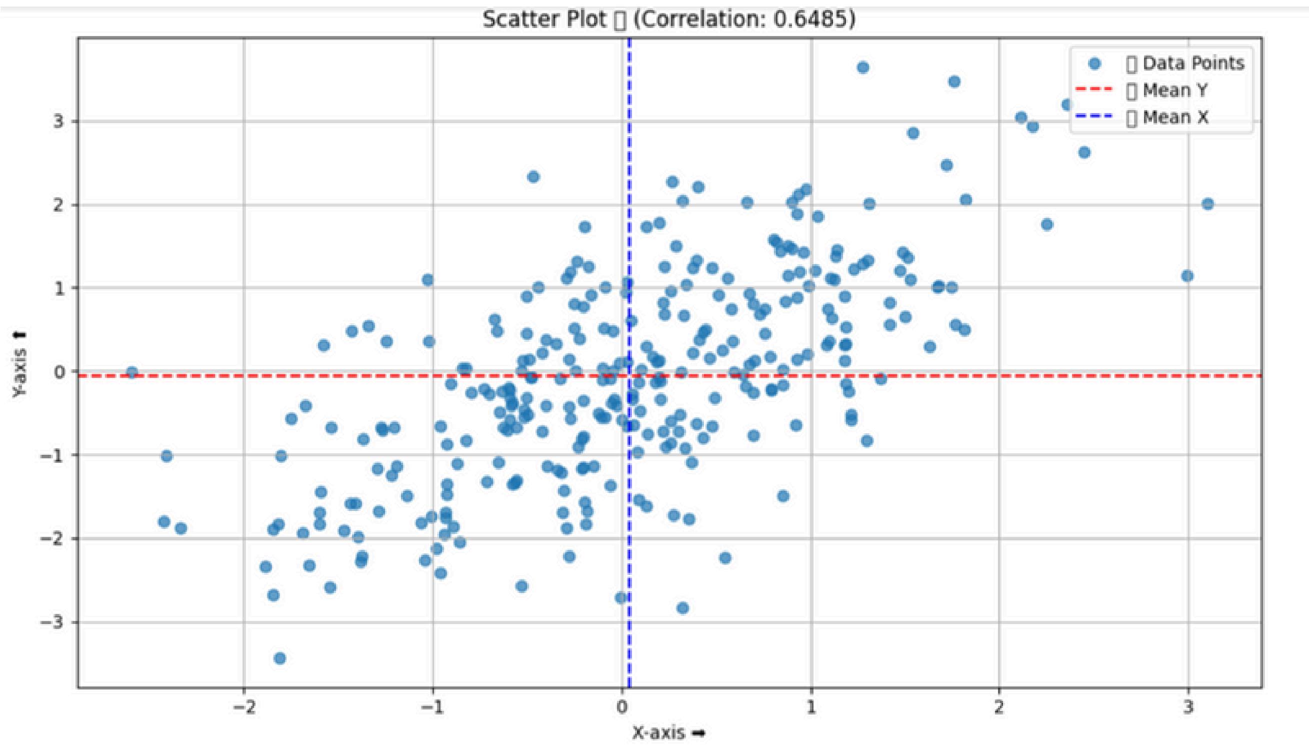
3. If you were to modify the script to generate data with a stronger or weaker relationship between X and Y, how would you expect the covariance, correlation coefficient, and scatter plot to change?

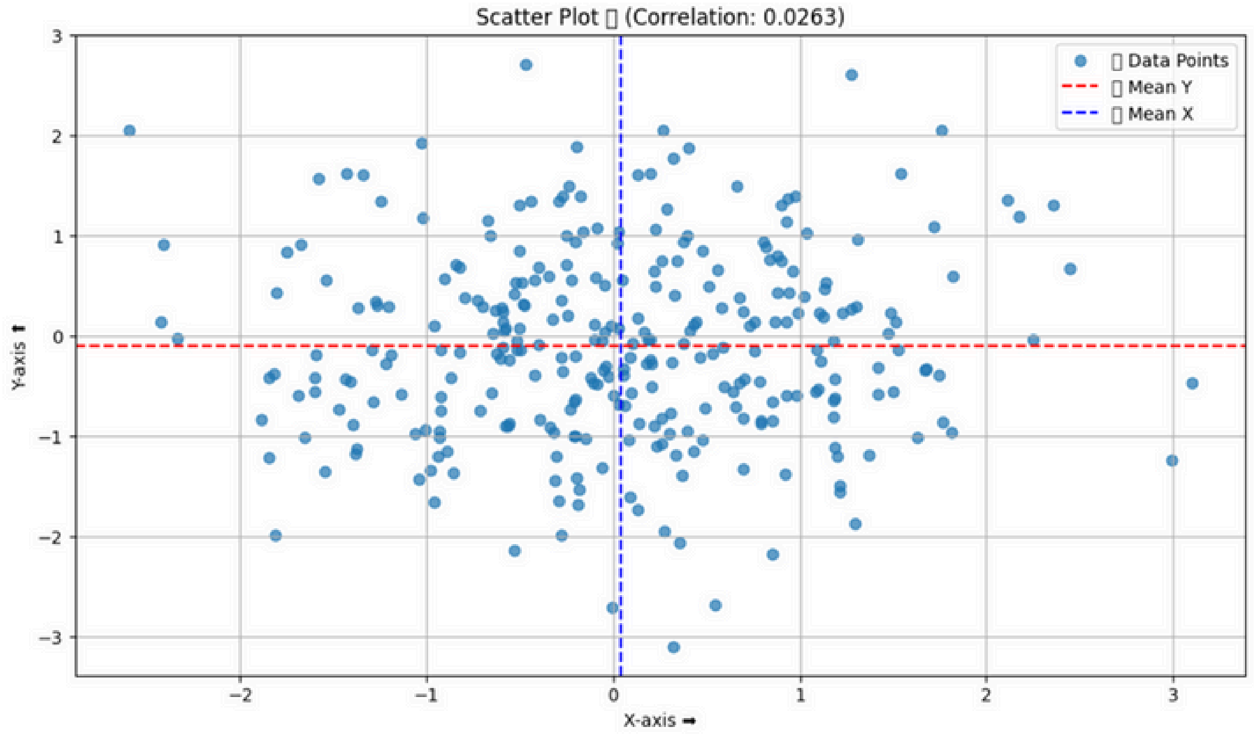
• **Weaker Relationship:**

– Covariance decreases in magnitude, approaching 0, as in Figures .

– Correlation coefficient approaches 0, indicating less linear association.

– The scatter plot points appear more dispersed, as shown in Figures .





Questions

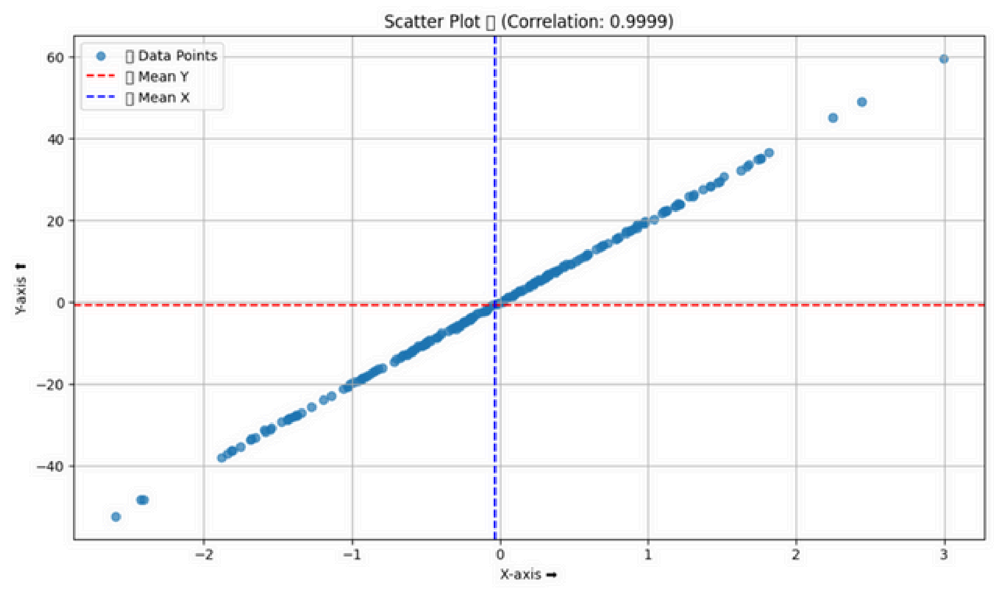
4. Based on the scatter plot and the calculated correlation coefficient, would you describe the relationship between X and Y as positive, negative, or no correlation? Justify your answer.

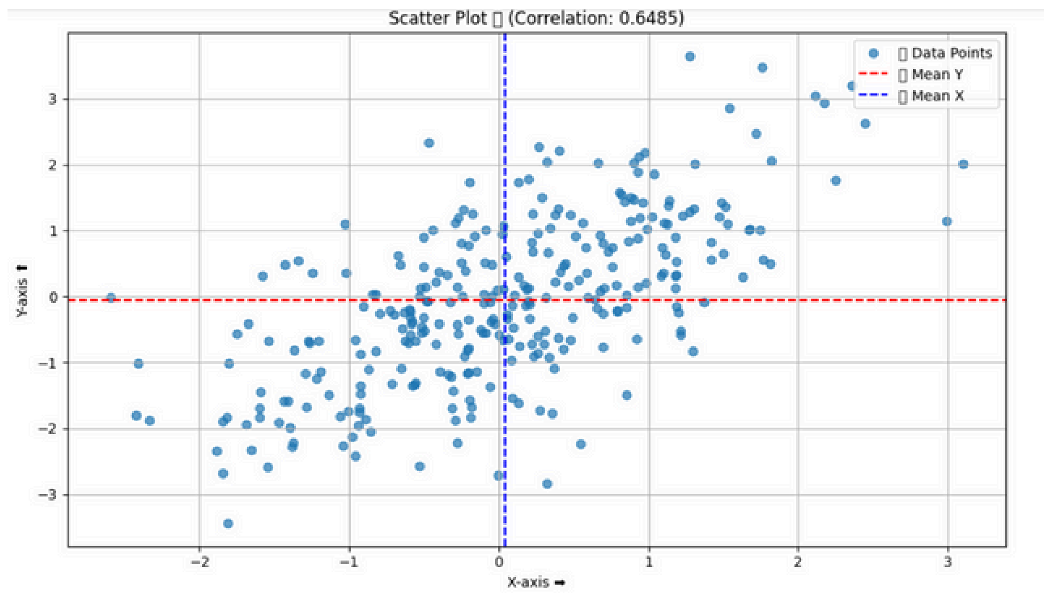
The relationship can be classified as follows:

• **Positive Correlation:** As in figure in left side, where the correlation coefficient is near 1, the scatter plot shows a clear upward trend.

• **Negative Correlation:** This would look opposite to figure in left side, with a downward trend.

• **No Correlation:** As in right side figure, the scatter plot appears random, with no discernible linear trend.





Questions

5. How might outliers in the generated data affect the calculated covariance and correlation coefficient? Can you identify any potential outliers in the scatter plot?

Outliers can significantly distort the calculated metrics:

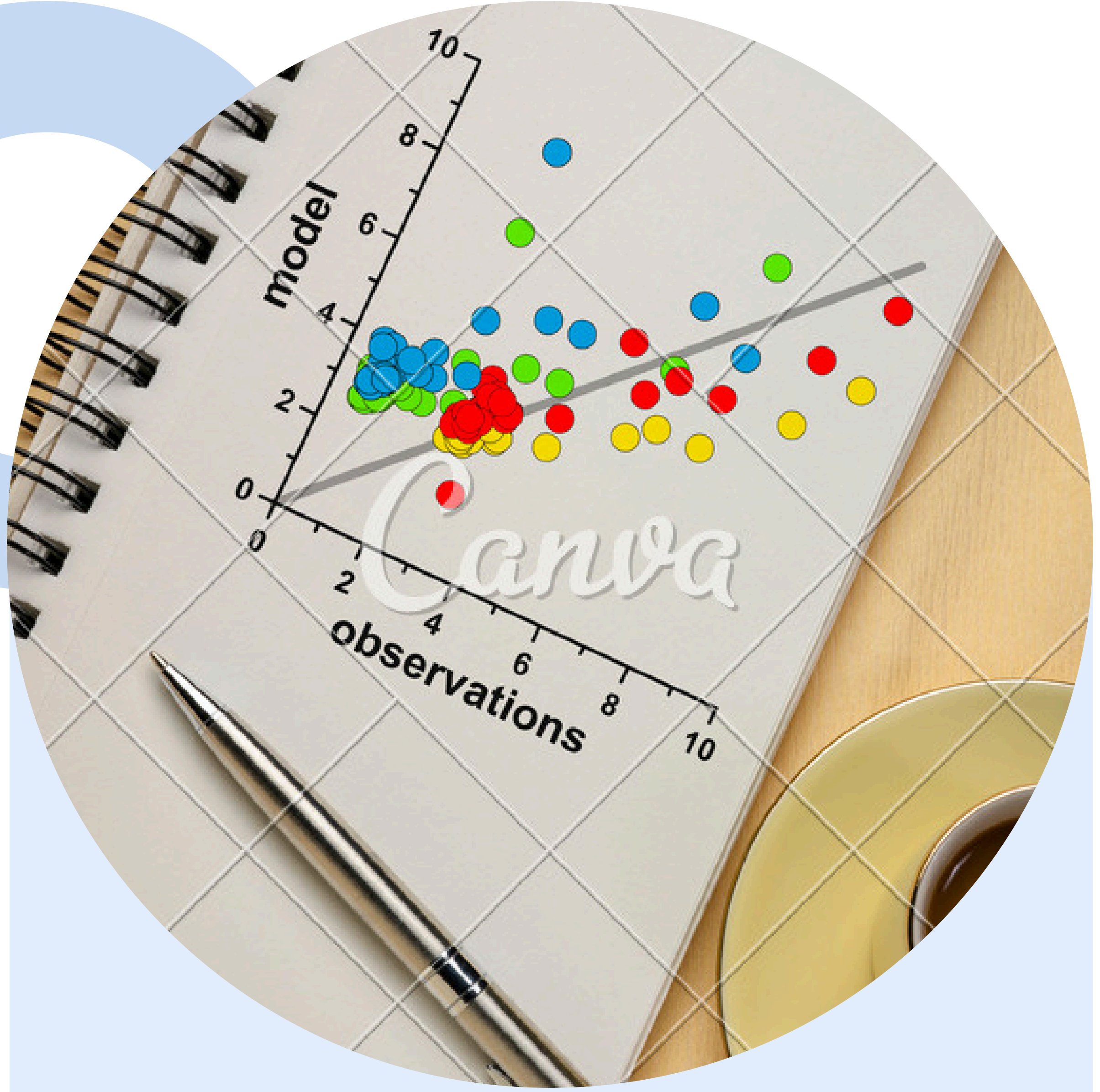
- **Covariance:** Outliers can inflate or skew the magnitude of covariance

- **Correlation Coefficient:**
  - Reduce a strong correlation
  - Artificially inflate correlation if the outlier aligns with the trend.

Identifying Outliers: Outliers are points that deviate significantly from the main trend. In the scatter plots, any point far from the cluster of data

Conclusion

- Covariance shows direction but not strength of a relationship.
- Correlation is a more interpretable measure of strength and direction.
- Visualizations help make statistical concepts more intuitive.



THANK YOU!