Lab 5: Correlation Analysis – Quantifying Relationships

Prelab Questions

1. **Define correlation and explain its importance in data analysis.**

   Correlation is like seeing a connection between two things. For example, if you're tracking how much you study and how well you do on tests, you might find that the more you study, the higher your test scores tend to be. That's correlation. In data analysis, this is super helpful because it helps us spot patterns and understand how things are connected—without jumping to conclusions about *why* they're connected.

2. **What is the range of the correlation coefficient, and what does each extreme represent?**

The correlation coefficient is a number that shows how strong that connection is, and it's always between **-1** and **1**:

- **+1** means a perfect positive relationship: when one thing goes up, the other does too, exactly.
- **-1** means a perfect negative relationship: when one thing goes up, the other goes down, exactly.
- **0** means no relationship at all—one thing's changes don't predict the other's.
- Anything in between shows varying degrees of how strongly they're related.

3. **How does correlation differ from causation?**

   Here's the classic lesson: just because two things are related doesn't mean one caused the other. For instance, ice cream sales and shark attacks might be strongly correlated, but that doesn't mean eating ice cream causes shark attacks! It could be that both are tied to warmer weather. So, correlation is about patterns, while causation is about cause and effect.

4. **What are some common methods to calculate correlation?**

There are different ways to measure how two things are connected:

- **Pearson's correlation** is the go-to for checking how linear (straight-line) relationships look.
- **Spearman's rank** is used when the relationship isn't necessarily linear but still consistently goes up or down.
- **Kendall's tau** is another option, often used for smaller datasets or when you're dealing with ranked data.

5. **Why is it essential to check for outliers before calculating correlation?**

   Outliers are like that one person who shows up at a party and totally changes the vibe. If there's an extreme value in your data, it can skew the correlation and make it seem like there's a stronger (or weaker) relationship than there really is. That's why it's good to spot

those outliers and decide if they should be included or removed before calculating correlation.

In-Lab Details

Objective:
Quantify the relationship between two numerical variables using correlation analysis and visualize the results.

- Dataset: student_scores.csv with columns for study hours and test scores.
  SOURCE: https://www.kaggle.com/datasets/kamleshsam/student-scores

Expected Output:

1. **Correlation Coefficient**: A numerical value indicating the strength and direction of the relationship (e.g., 0.85 for a strong positive correlation).
2. **Scatter Plot**: A visual representation showing how test scores change with study hours.
3. **Heatmap**: A matrix displaying correlation coefficients for multiple variables.

PROGRAM CODE:

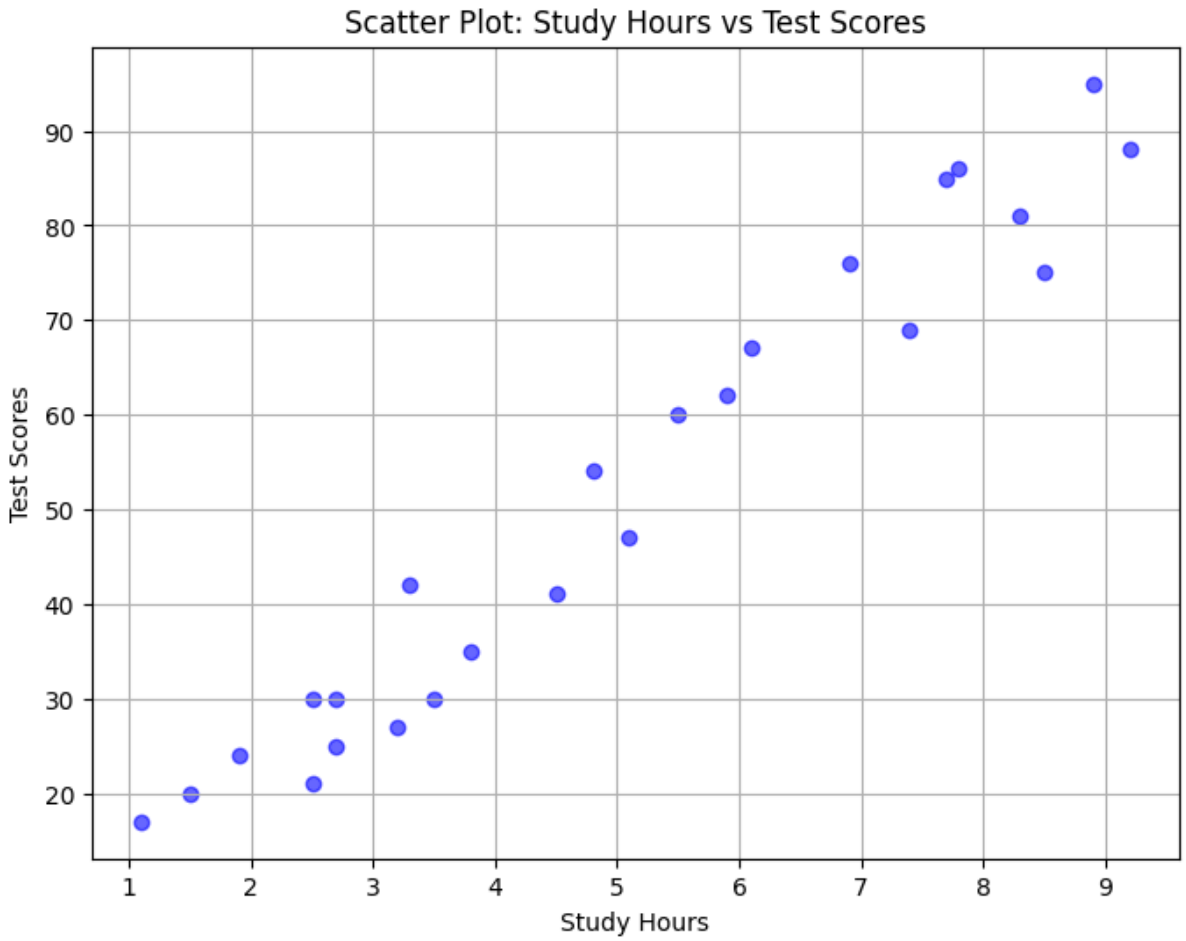
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from google.colab import files

# Step 2: Load the dataset
data = pd.read_csv('/content/student_scores.csv')
# Step 4: Calculate the correlation coefficient
correlation = data['Hours'].corr(data['Scores'])
print(f"\nCorrelation Coefficient: {correlation:.2f}")
```

OUTPUT:

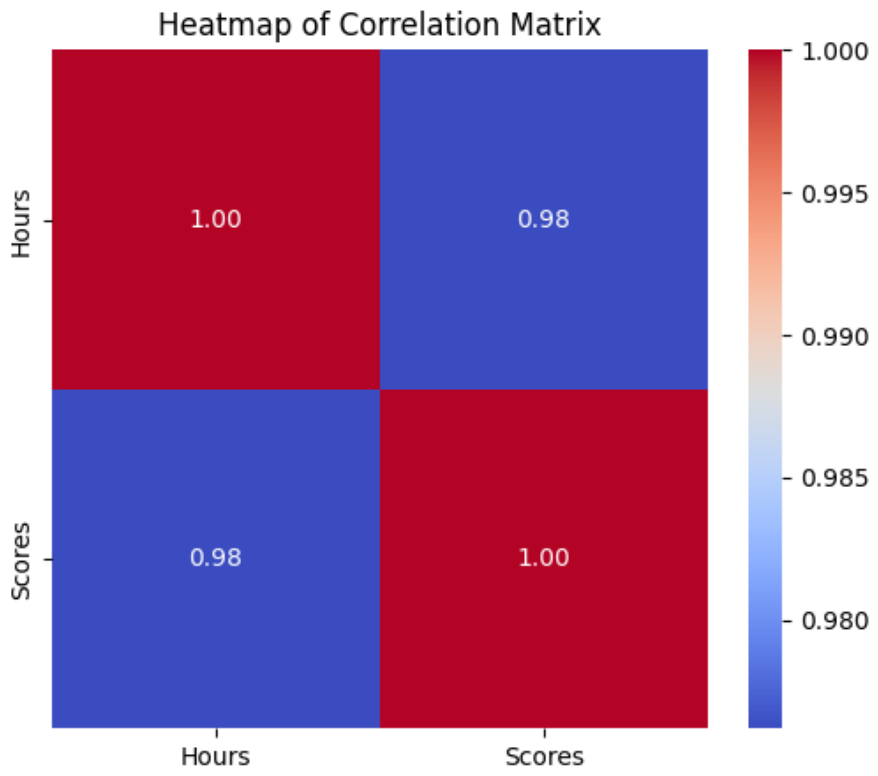
Correlation Coefficient: 0.98

```
# Step 5: Create a scatter plot
plt.figure(figsize=(8, 6))
plt.scatter(data['Hours'], data['Scores'], color='blue', alpha=0.6)
plt.title('Scatter Plot: Study Hours vs Test Scores')
plt.xlabel('Study Hours')
plt.ylabel('Test Scores')
plt.grid(True)
plt.show()
```

OUTPUT:




```
# Step 6: Create a heatmap for correlations
plt.figure(figsize=(6, 5))
correlation_matrix = data.corr()  # Compute correlation matrix
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".2f",
cbar=True)
plt.title('Heatmap of Correlation Matrix')
plt.show()
```



Postlab Questions

1. What does a correlation coefficient of 0.85 indicate about the relationship between study hours and test scores?

   A correlation coefficient of **0.85** indicates a **strong positive relationship** between the two variables (study hours and test scores). This means that, generally, as study hours increase, test scores also tend to increase. The value is close to 1, signifying a strong direct linear relationship, although it's not perfect (which would be 1).

2. Why is it important to visualize correlations in addition to calculating them?

- **Clarity**: Visualizations such as scatter plots or heatmaps help you quickly identify patterns, trends, or outliers that might not be immediately obvious from just a numerical correlation coefficient.
- **Context**: It gives a more intuitive understanding of how variables are related, showing if the relationship is linear or non-linear, and if there are exceptions or clusters in the data.
- **Communication**: Visuals are easier for non-experts to understand and interpret than raw correlation coefficients, making them effective for presenting findings.

3. How would you interpret a correlation coefficient of -0.3?

   A correlation coefficient of **-0.3** suggests a **weak negative relationship** between the two variables. This means that as one variable increases, the other tends to decrease, but the

relationship is not very strong. It’s an indication of a slight inverse relationship, but the correlation is weak, so it should be interpreted with caution.

4. Suggest scenarios where correlation analysis is not suitable.

   Correlation analysis is not suitable in the following scenarios:

- **Causal Relationships**: Correlation does not imply causation. If you're looking to establish a cause-and-effect relationship, other methods such as experiments or regression analysis are more appropriate.
- **Non-linear Relationships**: If the relationship between the variables is non-linear (e.g., exponential, quadratic), correlation might fail to capture this and provide misleading results.
- **Outliers**: Correlation is sensitive to outliers. In datasets with extreme values, the correlation coefficient may be skewed and not represent the true relationship.
- **Ordinal Data**: If your data consists of ordinal variables (ranked data), Pearson correlation may not be appropriate. Spearman's rank correlation is more suitable in such cases.

5. What are the limitations of using correlation as a measure of relationship?

   The limitations of correlation include:

- **No Causality**: Correlation only measures the strength and direction of the relationship between variables, not cause and effect. Two variables may be correlated due to a third variable influencing both (a spurious relationship).
- **Sensitive to Outliers**: Extreme values can significantly affect the correlation coefficient, potentially making the relationship seem stronger or weaker than it actually is.
- **Assumes Linear Relationships**: Pearson's correlation only detects linear relationships. Non-linear relationships may not be captured properly.
- **Range of Values**: Correlation measures the degree of linear relationship between -1 and 1, which may not fully represent the complexity of the relationship between variables.
- **Homogeneity**: Correlation assumes that the relationship between the variables is consistent throughout the data. It may not work well if the relationship changes at different values of the variables (heteroscedasticity).

DONE BY

Durga S – 22BCS026

Nandhini Priya K K -22BCS073

Nithiyaa T – 22BCS082

Priya Dharshini P – 22BCS094

Rithika S – 22BCS103

Sandhiya G – 22BCS109

Santhiya P – 22BCS111