What is Correlation?

Correlation measures the strength and direction of a linear relationship between two continuous variables. The correlation coefficient ranges from -1 to 1.

1. A positive correlation (close to 1) indicates that as one variable increases, the other tends to increase.

2. A negative correlation (close to -1) indicates that as one variable increases, the other tends to decrease.

3. A correlation close to 0 suggests a weak or no linear relationship.

Example:

Consider a dataset of students where we have their hours of study and exam scores. If there's a positive correlation between hours of study and exam scores, it implies that, on average, students who study more tend to score higher on exams.

In [1]:
```
import pandas as pd

data = {'Hours_of_Study': [2, 5, 1, 6, 3],
        'Exam_Scores': [60, 85, 50, 90, 70]}

df = pd.DataFrame(data)

correlation = df['Hours_of_Study'].corr(df['Exam_Scores'])
print(f"Pearson correlation coefficient: {correlation}")
```

```
Pearson correlation coefficient: 0.9942734705398055
```

How to check if data is Linear and Non- Linear ?

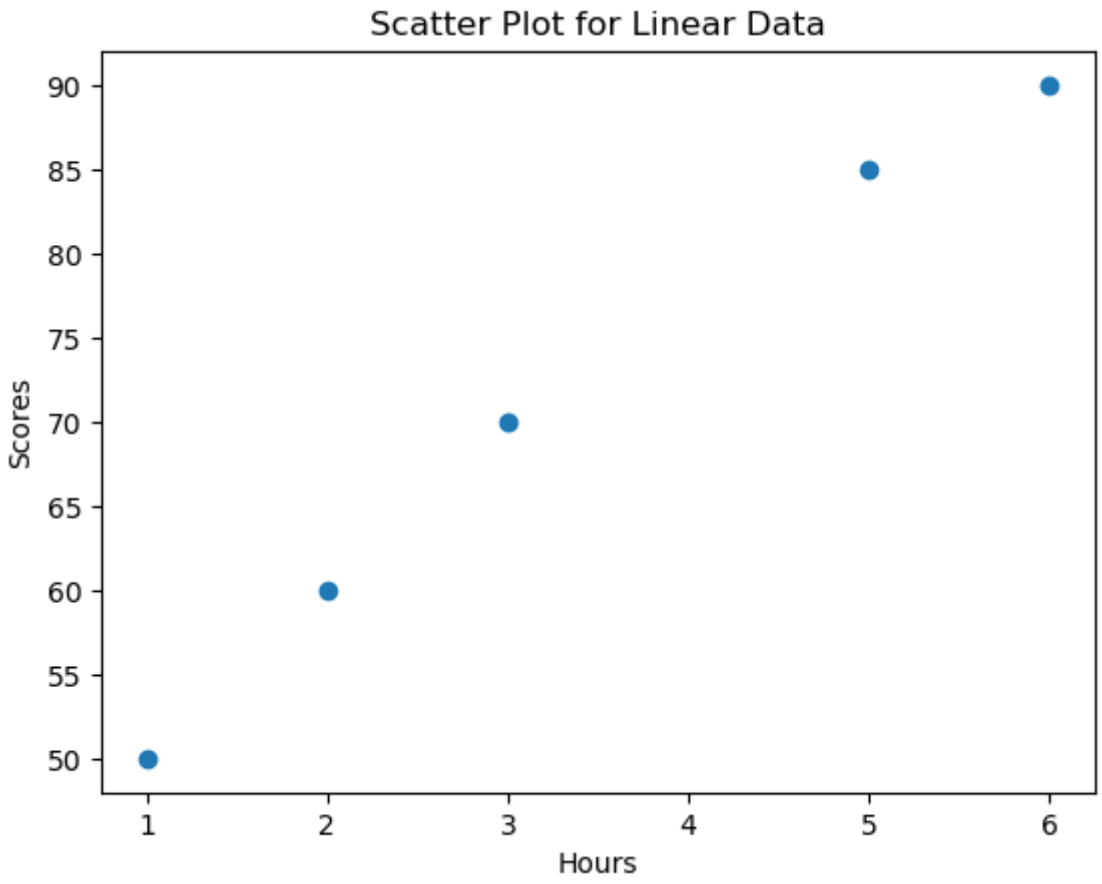
Here are some methods to help you determine linearity:

1. Scatter Plots:

Linear Relationship: In a scatter plot, points follow a clear pattern along a line.

Non-linear Relationship: Points deviate from a straight line, forming curves or other patterns.

In [2]:
```
import matplotlib.pyplot as plt
plt.scatter(df['Hours_of_Study'], df['Exam_Scores'])
plt.xlabel('Hours')
plt.ylabel('Scores')
plt.title('Scatter Plot for Linear Data')
plt.show()
```

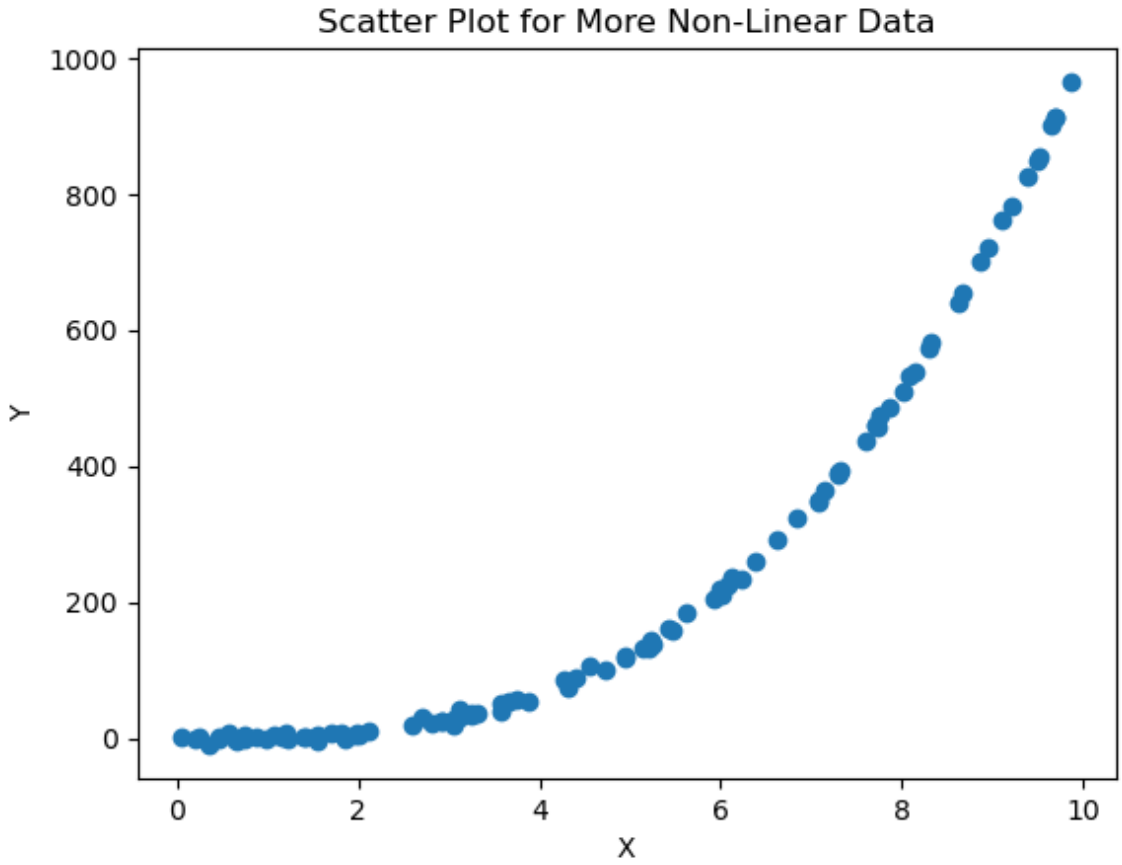


In [3]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

np.random.seed(42)
x = np.sort(10 * np.random.rand(100))
y = x**3 + np.random.normal(0, 5, 100)

df_nonlinear = pd.DataFrame({'X': x, 'Y': y})

plt.scatter(df_nonlinear['X'], df_nonlinear['Y'])
plt.title('Scatter Plot for More Non-Linear Data')
plt.xlabel('X')
plt.ylabel('Y')
plt.show()
```



2. Correlation Coefficient:

Linear Relationship: A high absolute value of the correlation coefficient (close to 1 or -1) indicates a strong linear relationship.

Non-linear Relationship: A correlation close to 0 suggests a weak or non-linear relationship.

In [4]:
```
correlation = df['Hours_of_Study'].corr(df['Exam_Scores'])
print(f"Pearson correlation coefficient on Linear Data: {correlation}")
```

```
Pearson correlation coefficient on Linear Data: 0.9942734705398055
```

In [5]:
```
correlation = df_nonlinear['X'].corr(df_nonlinear['Y'])
print(f"Pearson correlation coefficient on Non Linear Data: {correlation}")
```

```
Pearson correlation coefficient on Non Linear Data: 0.9166488959966173
```

If your data suggests a nonlinear relationship, you might consider using other correlation methods, such as Spearman's rank correlation (df['Variable1'].corr(df['Variable2'], method='spearman')).