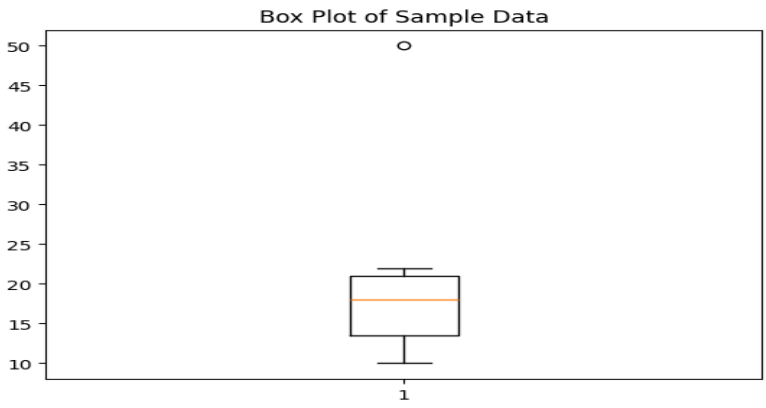
# Identifying the outliers in the dataset

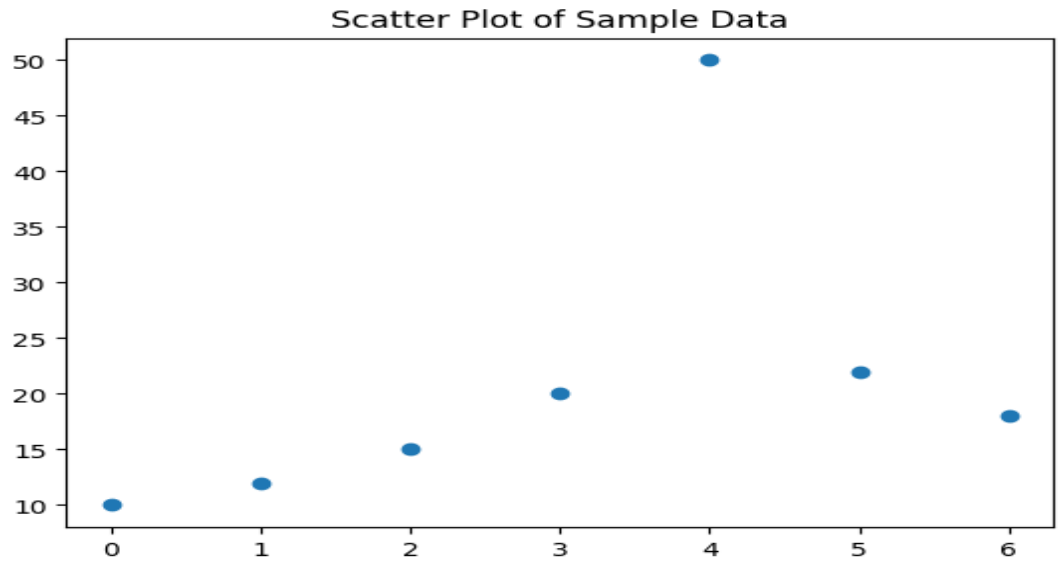
Outliers are data points that deviate significantly from the overall pattern of the rest of the data. They can be extremely high or extremely low compared to the majority. There's no single rule to define outliers, but common methods include:

1. **Visualization:**
**Box Plot**: This plot shows the median, quartiles (Q1 and Q3), and whiskers extending to potential outliers. Points outside the whiskers are candidates for outliers.
**Scatter Plot:** This plot allows you to visually inspect data points for deviations from the main cluster.
2. **Statistical Methods:**
**Z-scores:** These measure the number of standard deviations a data point is away from the mean. Values with absolute z-scores exceeding a threshold (e.g., 3) are considered outliers.
**Interquartile Range (IQR):** This is the difference between Q3 and Q1. Points below Q1 - 1.5IQR or above Q3 + 1.5IQR are potential outliers.

**Python Code Examples:**
**Visualization with libraries like matplotlib or seaborn:**

```
import matplotlib.pyplot as plt
# Sample data with an outlier
data = [10, 12, 15, 20, 50, 22, 18]
# Box plot
plt.boxplot(data)
plt.title("Box Plot of Sample Data")
plt.show()
# Scatter plot
plt.scatter(range(len(data)), data)
plt.title("Scatter Plot of Sample Data")
plt.show()
```






**Z-scores:**

```
import numpy as np
data = [10, 12, 15, 20, 50, 22, 18]
mean = np.mean(data)
std = np.std(data)
z_scores = (data - mean) / std
# Identify outliers based on a threshold (e.g., 3)
outliers = [data[i] for i, score in enumerate(z_scores) if abs(score) > 3]
print("Outliers based on Z-scores:", outliers)
```

**Interquartile Range (IQR):**

```
import numpy as np
data = [10, 12, 15, 20, 50, 22, 18]
q1 = np.percentile(data, 25)
q3 = np.percentile(data, 75)
iqr = q3 - q1
# Identify outliers based on IQR threshold (1.5 * IQR)
lower_bound = q1 - 1.5 * iqr
upper_bound = q3 + 1.5 * iqr
outliers = [x for x in data if x < lower_bound or x > upper_bound]
print("Outliers based on IQR:", outliers)
```