Data Analysis and Visualization

CentraleDigitalLab@Nice

Deborah Dore - PhD - ddore@i3s.unice.fr
MARIANNE, I3S, CNRS, INRIA, Université Côte d'Azur

Outliers

Introduction

Definition: Outliers are data points that significantly deviate from the rest of the dataset

Significance:

- Skew results
- Affects the performance of predictive models.
- Distort statistical analysis

Examples:

- Age of 112 in a patient dataset represents an unusual high value
- An individual taking 10 years to complete a bachelor's degree, compared to the standard duration of 5 years in a university dataset, represents a significant deviation from the typical trend

Outliers

Common causes of outliers

- **Data entry errors**: Typos or measurement inaccuracies.
- **Natural variability**: Genuine deviations in data.
- **Sampling issues**: Sampling from different populations.
- **External factors**: Events or anomalies affecting data.

Outliers

Methods to Identify Outliers

Statistical Methods:

- Interquartile Range (IQR)
- Z-Score

Visualization Methods:

- Box Plots
- Scatterplots
- Histograms

Outliers

Statistical Methods to Identify Outliers: IQR

Interquartile Range (IQR) is a measure of statistical dispersion that measures the spread of the middle 50% of data.

Formula:

Given an even *2n* or odd *2n+1* number of values,

$Q_1 =$ median of the *n* smallest values

$Q_3 =$ median of the *n* largest values

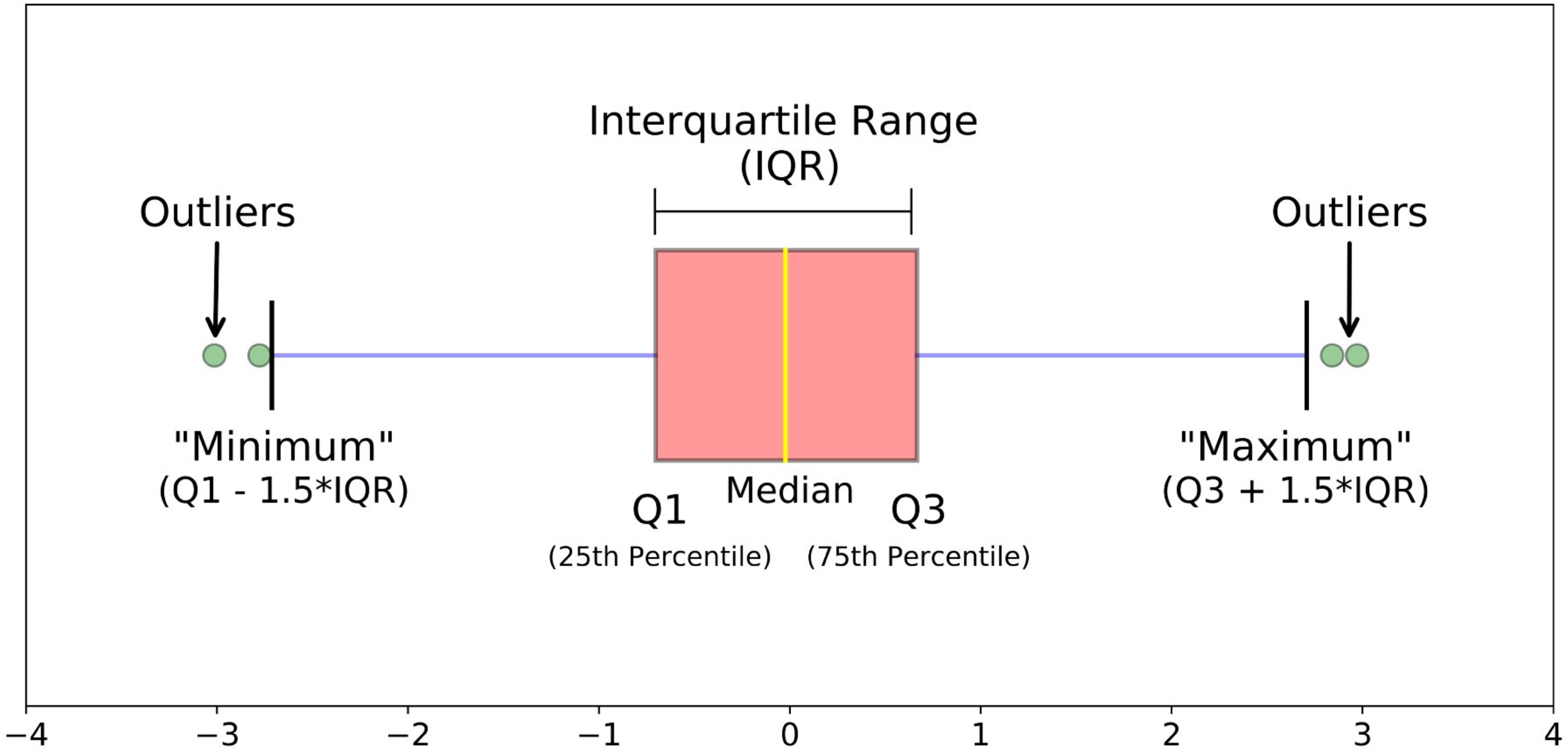
$IQR = Q_3 - Q_1$

Outliers lie below $Q_1 - (IQR * 1{,}5)$ and above $Q_3 + (IQR * 1{,}5)$

Outliers

Statistical Methods to Identify Outliers: IQR

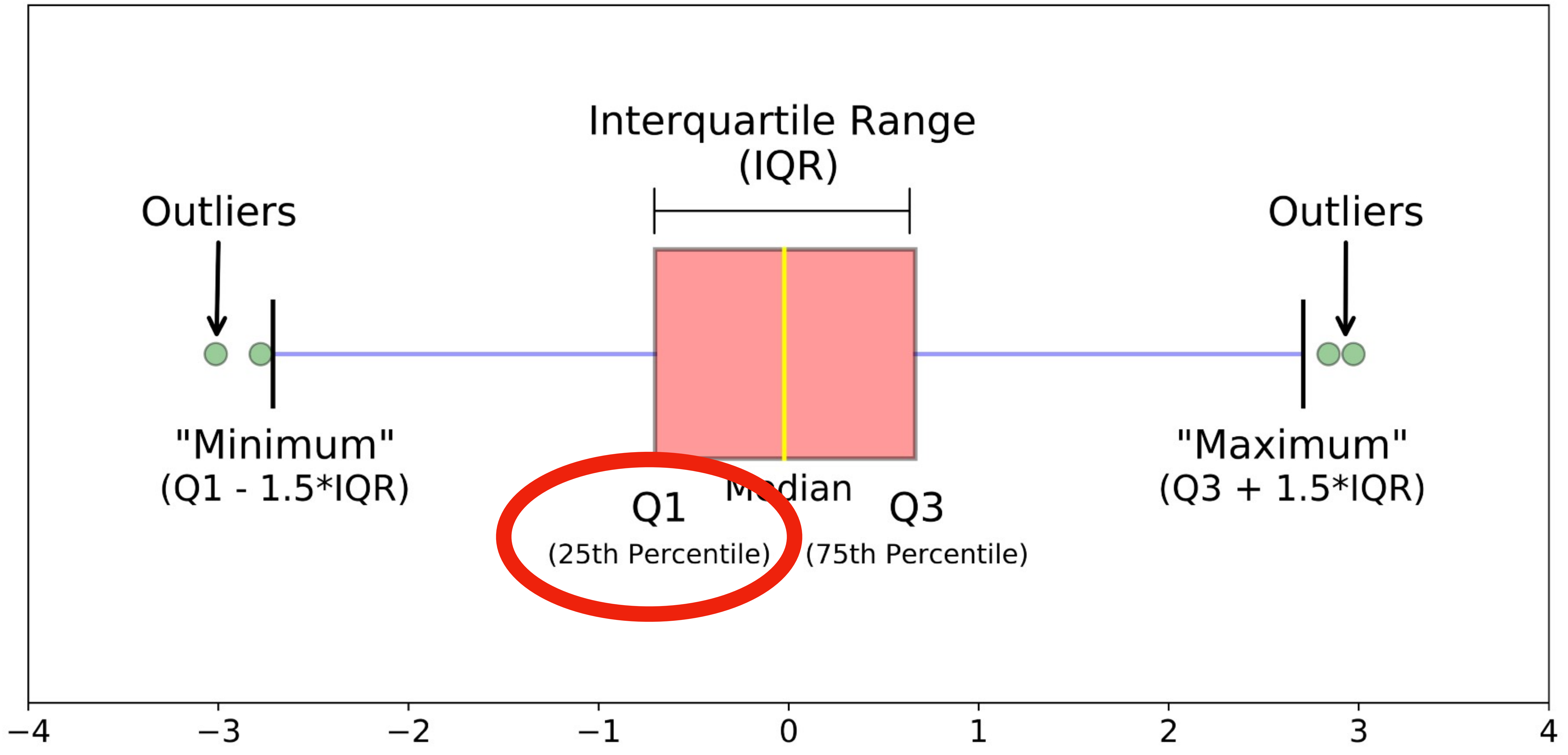
Interquartile Range (IQR) is a measure of statistical dispersion that measures the spread of the middle 50% of data



Outliers

Statistical Methods to Identify Outliers: IQR

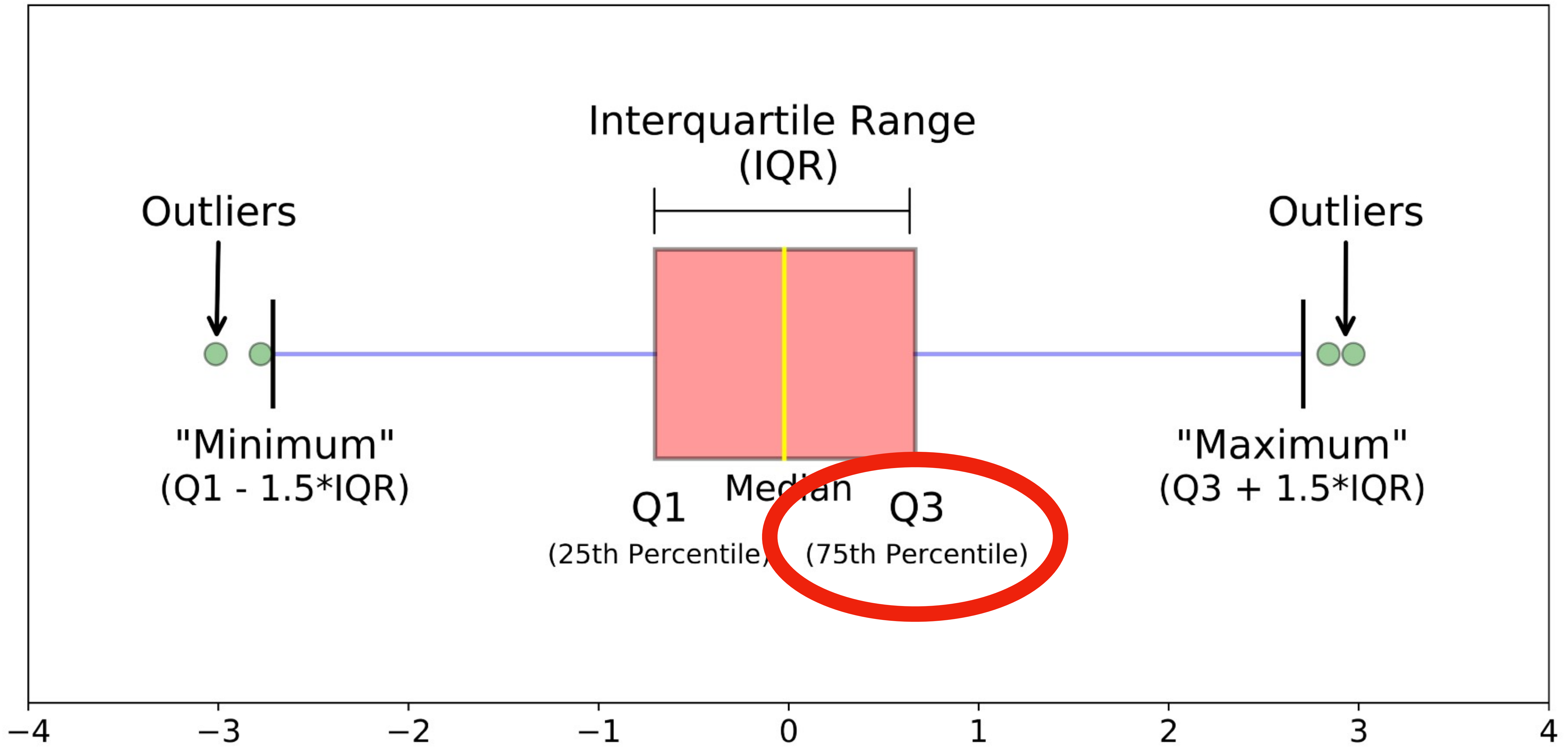
Interquartile Range (IQR) is a measure of statistical dispersion that measures the spread of the middle 50% of data



Outliers

Statistical Methods to Identify Outliers: IQR

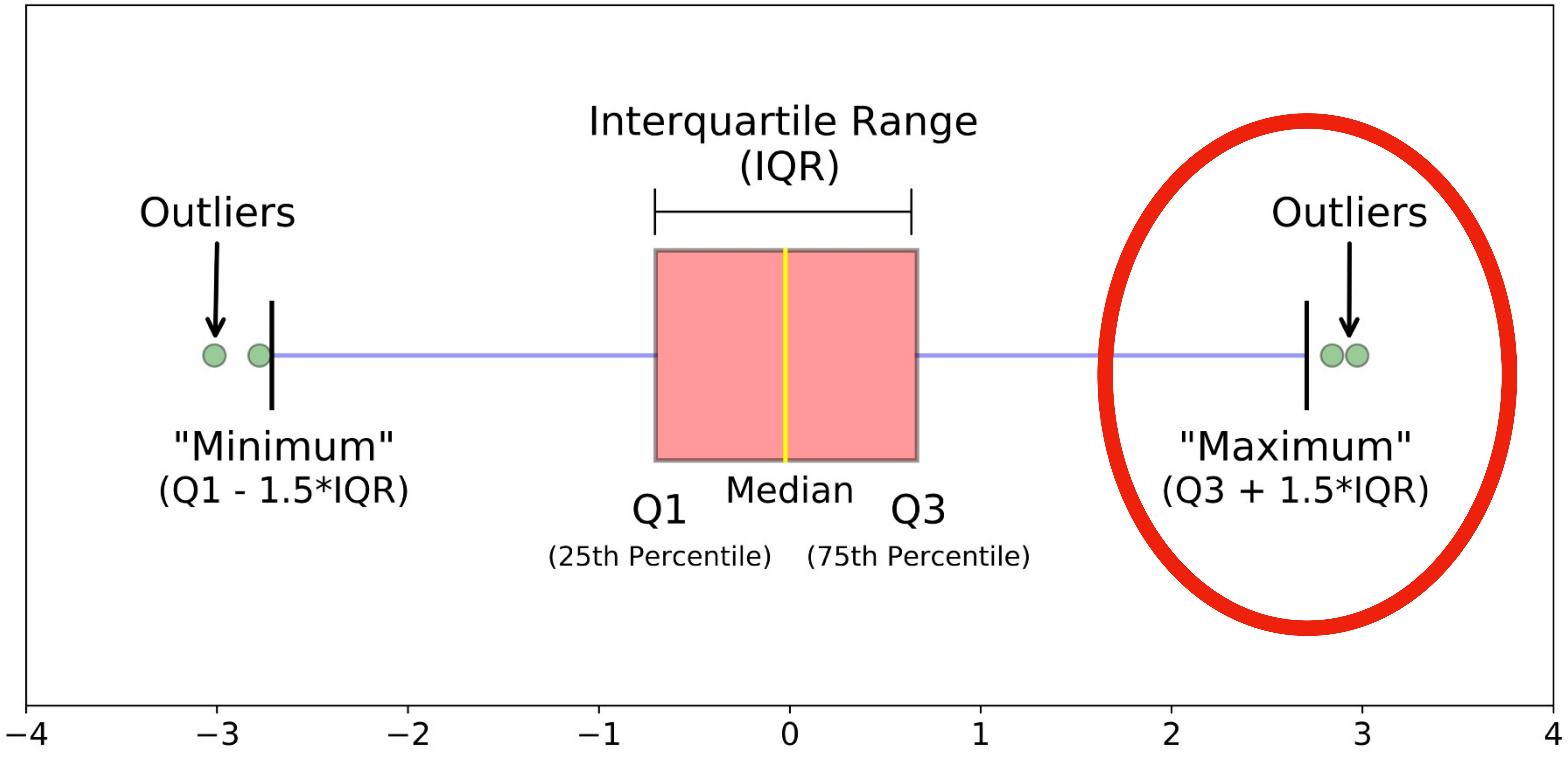
Interquartile Range (IQR) is a measure of statistical dispersion that measures the spread of the middle 50% of data



Outliers

Statistical Methods to Identify Outliers: IQR

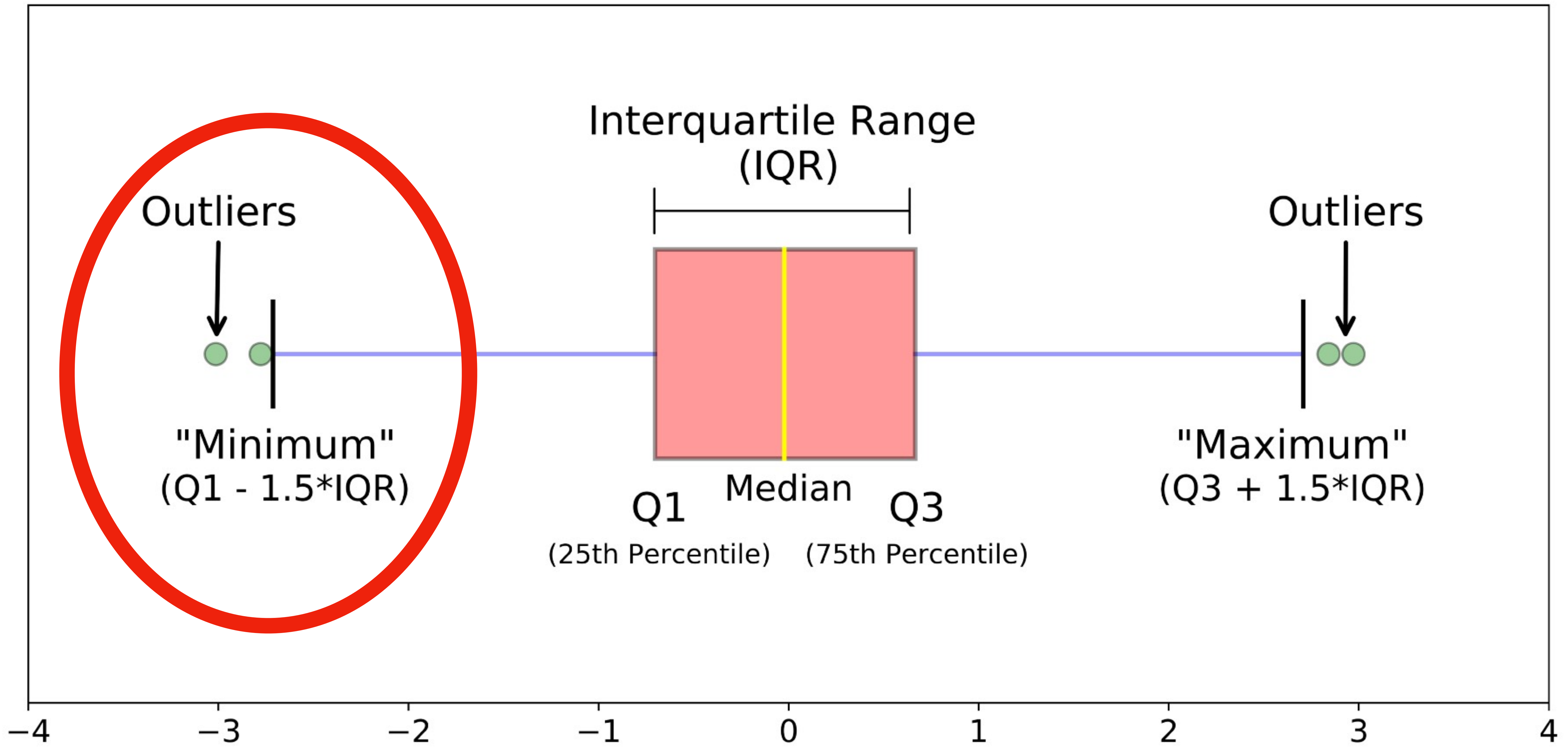
Interquartile Range (IQR) is a measure of statistical dispersion that measures the spread of the middle 50% of data



Outliers

Statistical Methods to Identify Outliers: IQR

Interquartile Range (IQR) is a measure of statistical dispersion that measures the spread of the middle 50% of data



Outliers

Statistical Methods to Identify Outliers: IQR

Interquartile Range (IQR) is a measure of statistical dispersion that measures the spread of the middle 50% of data

Example:

1. Dataset = $[0,110,5,100,200,-1]$ →*order it* → $[-1,0,5,100,110,200]$
2. Find the median value: in this case, the numbers are odd so there is no specific "value" but we can still divide our dataset into two parts: $p_1 = [-1,0,5]$ and $p_2 = [100,110,200]$
3. *Find the median value of both parts:* $Q_1 = median(p_1) = 0$ and $Q_3 = median(p_2) = 110$
4. $IQR = Q_3 - Q_1 = 110 - 0 = 110$
5. Outliers lie in the range $(-\infty, Q_1 - (IQR * 1{,}5)] = (-\infty, -165]$ and $[Q_3 + (IQR * 1{,}5), +\infty) = [275, +\infty)$

Outliers

Statistical Methods to Identify Outliers: IQR

Interquartile Range (IQR) is a measure of statistical dispersion that measures the spread of the middle 50% of data

Example:

1. Dataset = $[0,110,5,100,200,-1,50]$ →*order it* → $[-1,0,5,50,100,110,200]$
2. Find the median value: $median(dataset) = 50$. We can divide our dataset into two parts: $p_1 = [-1,0,5]$ and $p_2 = [100,110,200]$
3. *Find the median value of both parts:* $Q_1 = median(p_1) = 0$ and $Q_3 = median(p_2) = 110$
4. $IQR = Q_3 - Q_1 = 110 - 0 = 110$
5. Outliers lie in the range $(-\infty, Q_1 - (IQR * 1{,}5)] = (-\infty, -165]$ and $[Q_3 + (IQR * 1{,}5), +\infty) = [275, +\infty)$

Outliers

Statistical Methods to Identify Outliers: Z-Score

Z-Score measures how many standard deviations a data point is from the mean

Formula:

$$ZScore = \frac{x_i - mean}{standard\ deviation}$$

Define a threshold t, usually ± 3.0

Outliers

Statistical Methods to Identify Outliers: Z-Score

Z-Score measures how many standard deviations a data point is from the mean

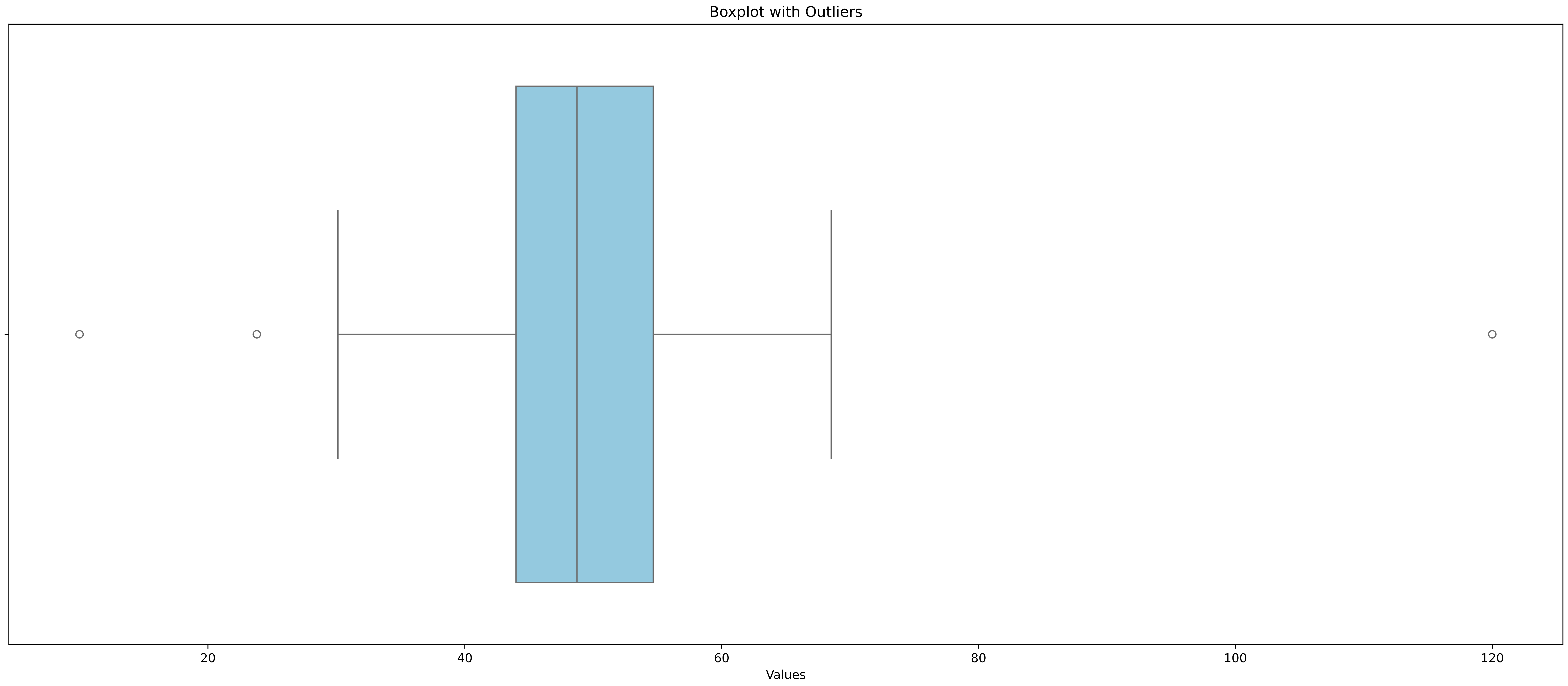
Example:

1. Dataset = $[0,10,5,15,4,1000]$, with threshold ± 2.0
2. Calculate the mean: $mean(dataset) = mean([0,10,5,15,4,1000]) = 172$
3. *Calculate std:* $s = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \overline{x})^2}{N}} = 370$
4. *Calculate the Score for each point:*
   $Dataset_{zscores} = [-0.46554667, \ -0.43853235, \ -0.45203951, \ -0.4250252, \ -0.45474094, \ 2.23588466]$
5. The ZScore of 1000 is 2.2, <u>above</u> our threshold. Therefore it's considered an outlier

Outliers

Visualization Methods to Identify Outliers: Box-Plot

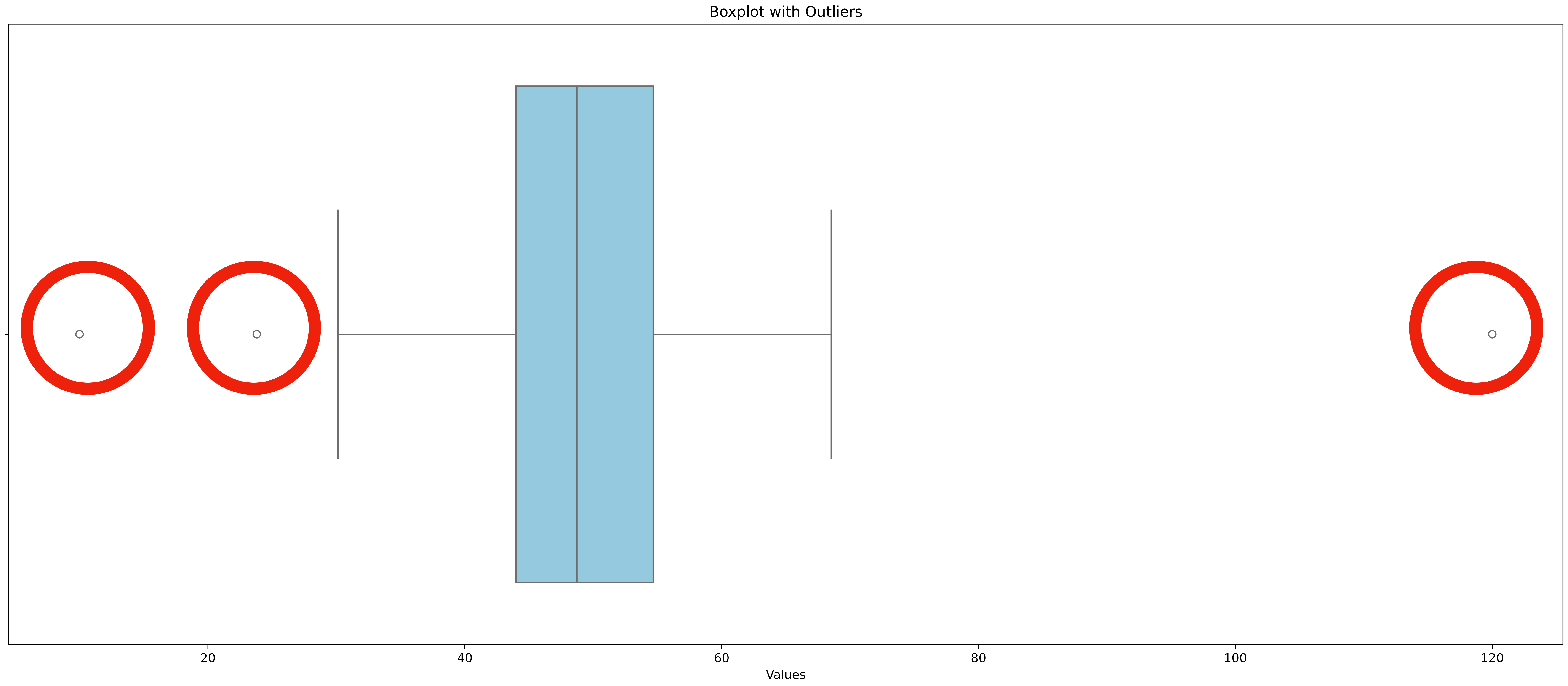
A **Box Plot** highlights outliers as individual points



Outliers

Visualization Methods to Identify Outliers: Box-Plot

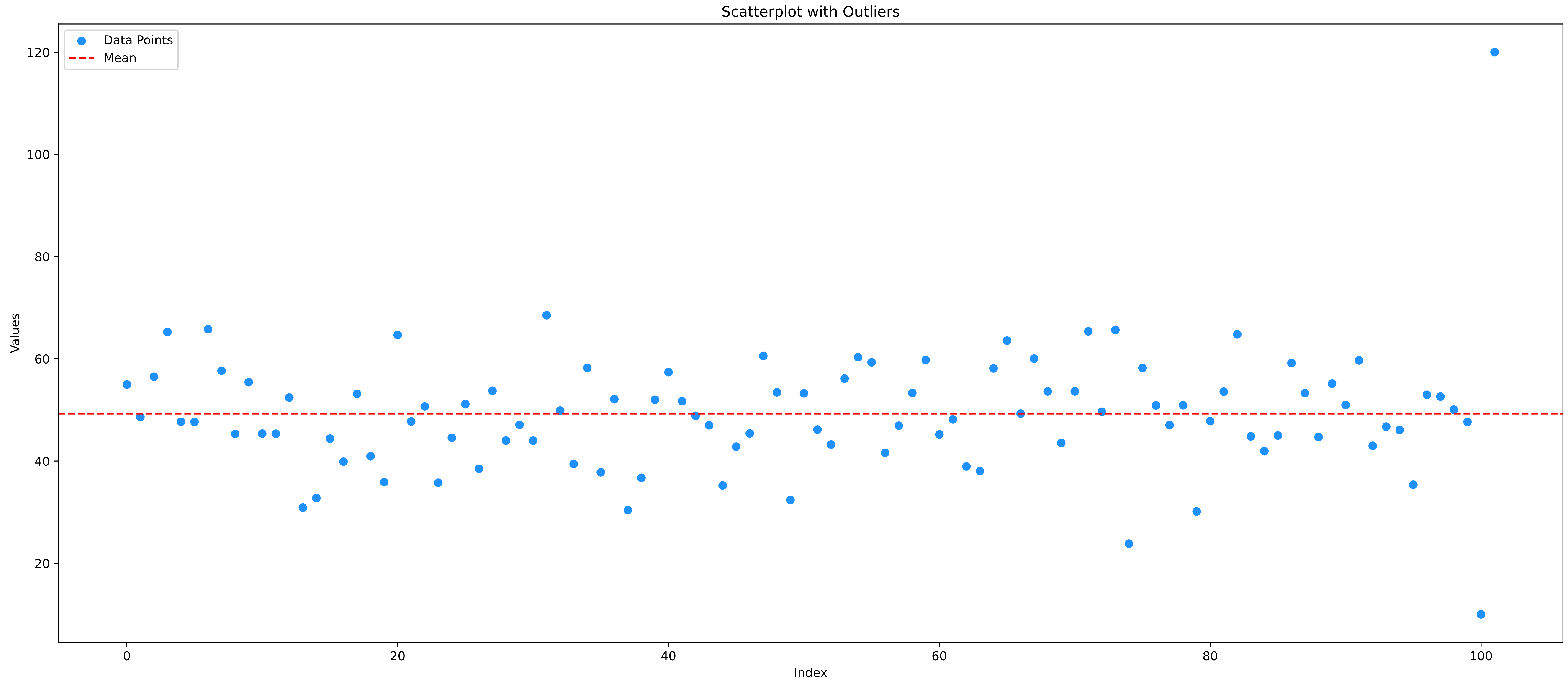
A **Box Plot** highlights outliers as individual points



Outliers

Visualization Methods to Identify Outliers: Scatter Plots

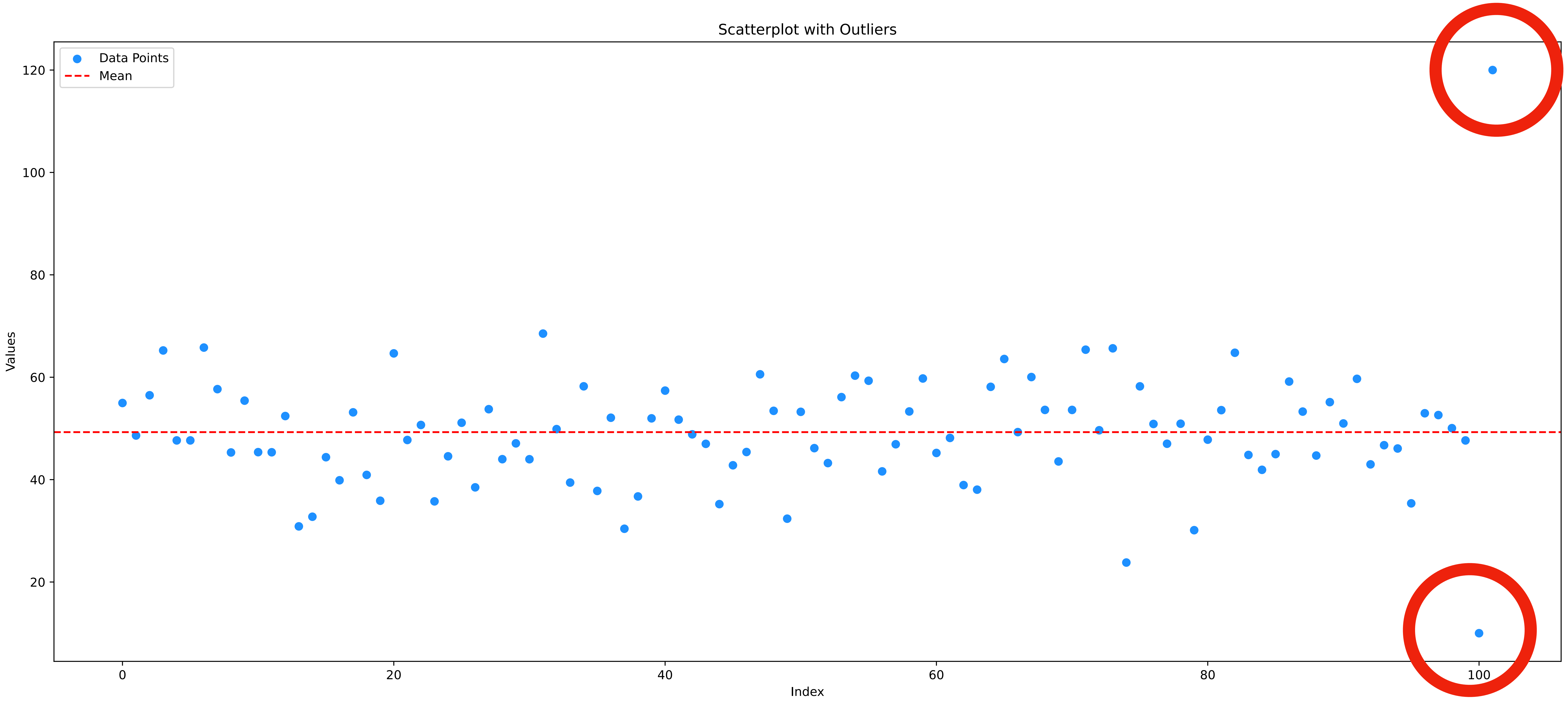
A **Scatterplot** identifies anomalies in bivariate relationships



Outliers

Visualization Methods to Identify Outliers: Scatter Plots

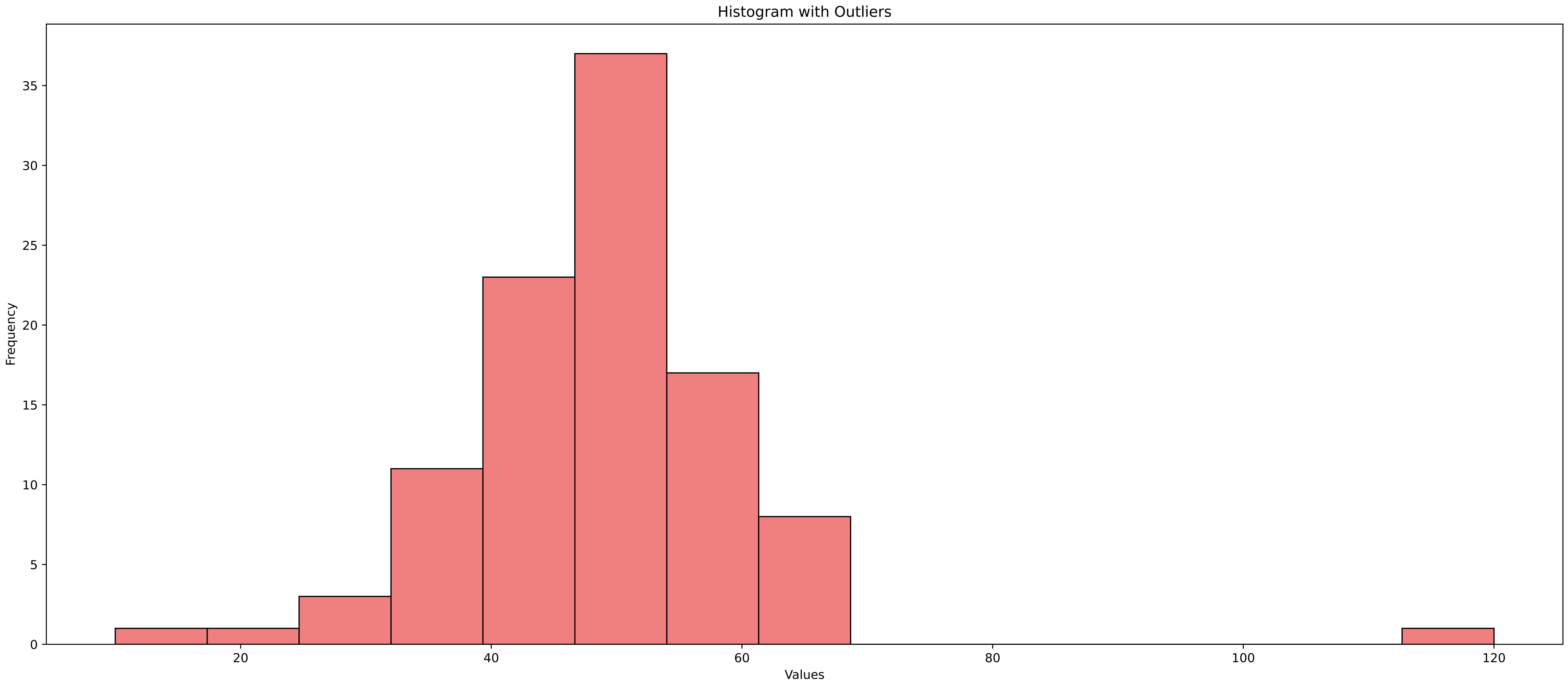
A **Scatterplot** identifies anomalies in bivariate relationships



Outliers

Visualization Methods to Identify Outliers: Histograms

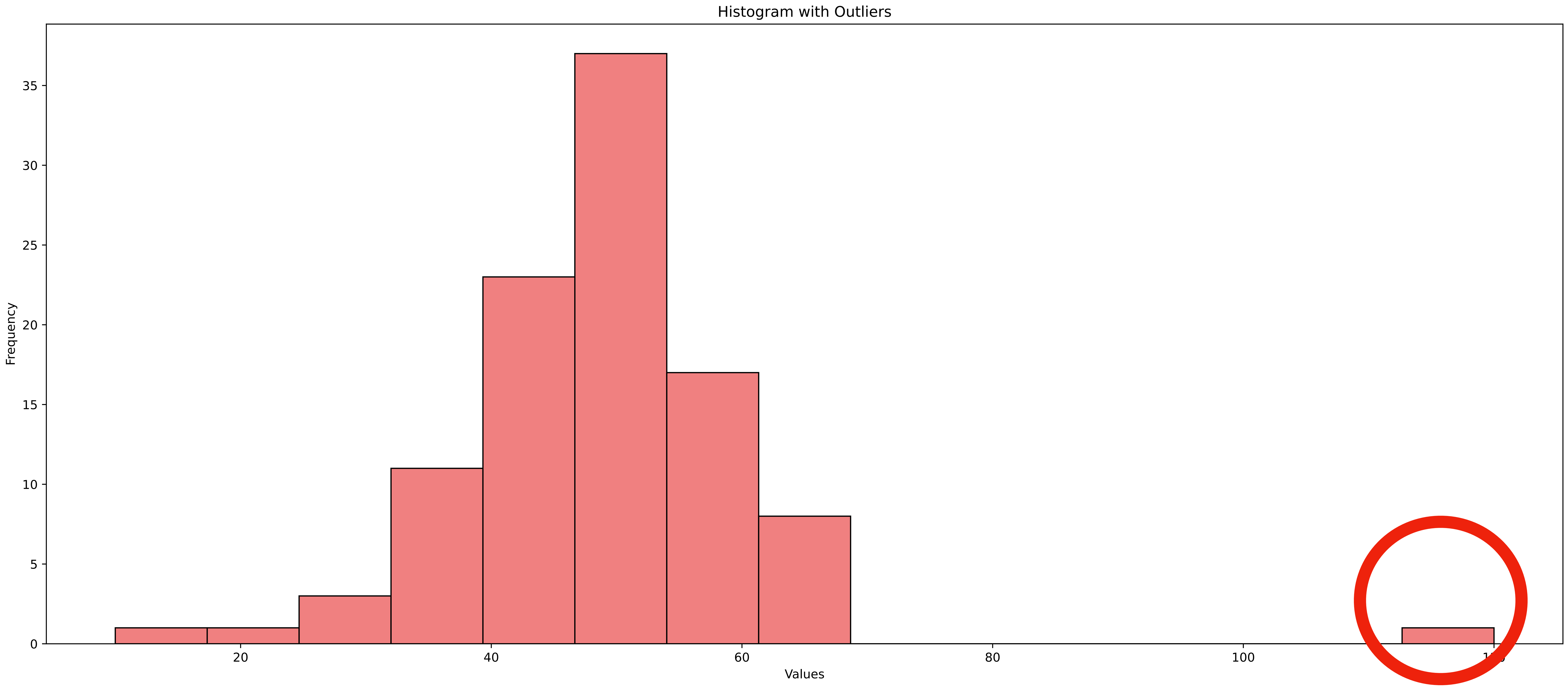
An **Histogram** shows extreme values in distributions



Outliers

Visualization Methods to Identify Outliers: Histograms

An **Histogram** shows extreme values in distributions



Outliers

Strategies to Handle Outliers

How to mitigate the effects of outliers?

- **Transformation:** Apply mathematical transformations to reduce the impact of outliers (log, square root ..)
- **Removal**: Eliminate data points that are significantly different from others. Ensure you're not discarding meaningful data!
- **Capping**: Replace extreme values with boundary values

Data Scaling

Introduction

Definition: Adjusting the range of data to bring features to a comparable scale

Significance:

- Brings features to a common scale
- Improves performance of machine learning algorithms
- Reduces bias caused by scale differences

Data Scaling

Standardization

Standardization rescales data to have a mean of 0 and standard deviation of 1

Formula: $z = \dfrac{(x - \mu)}{\sigma}$

where: μ is the *mean of the data points*

σ is the *standard deviation*

Example: Height in cm converted to standard units.

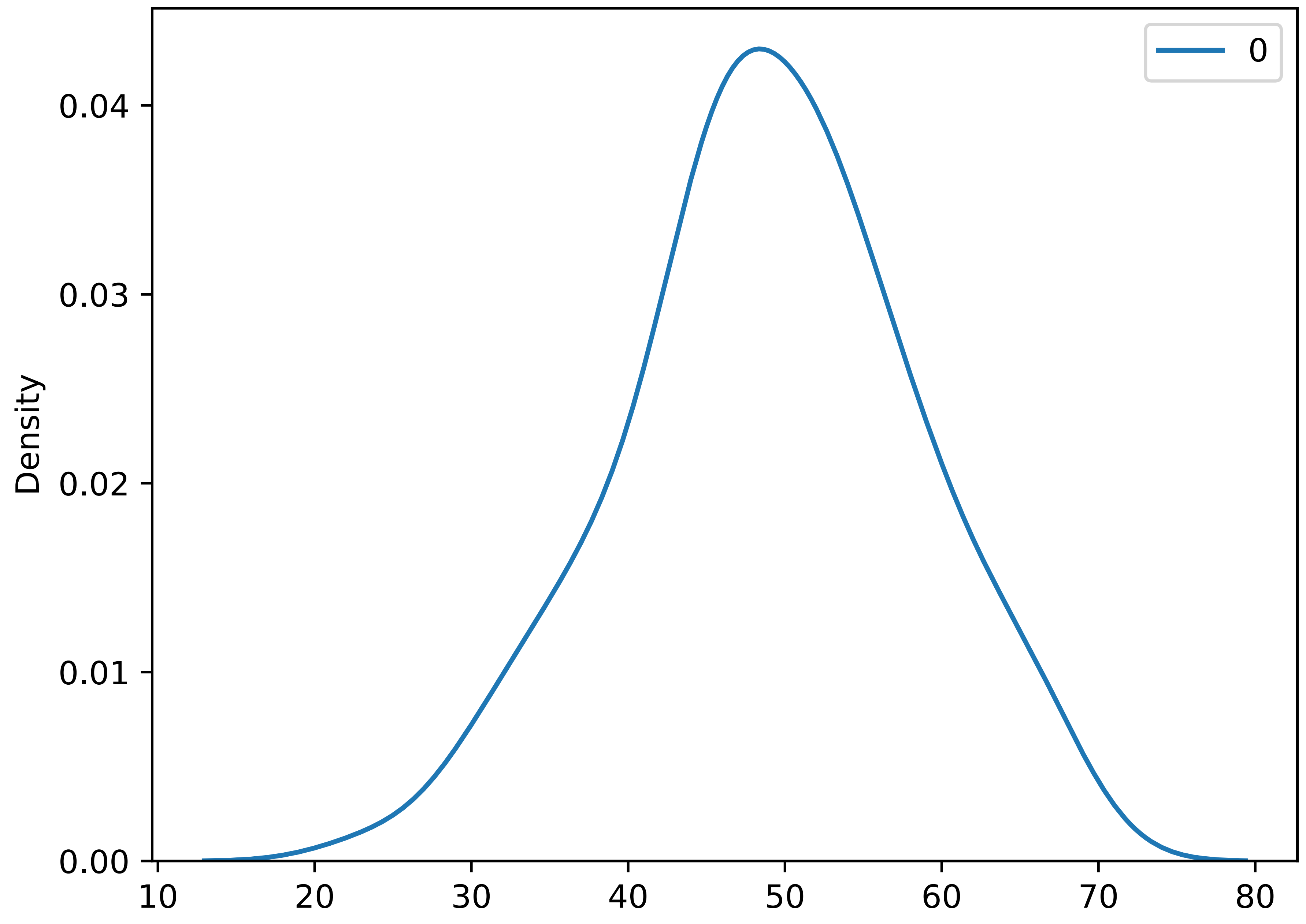
Data Scaling

Standardization

Standardisation is used when data needs to conform to a Gaussian distribution

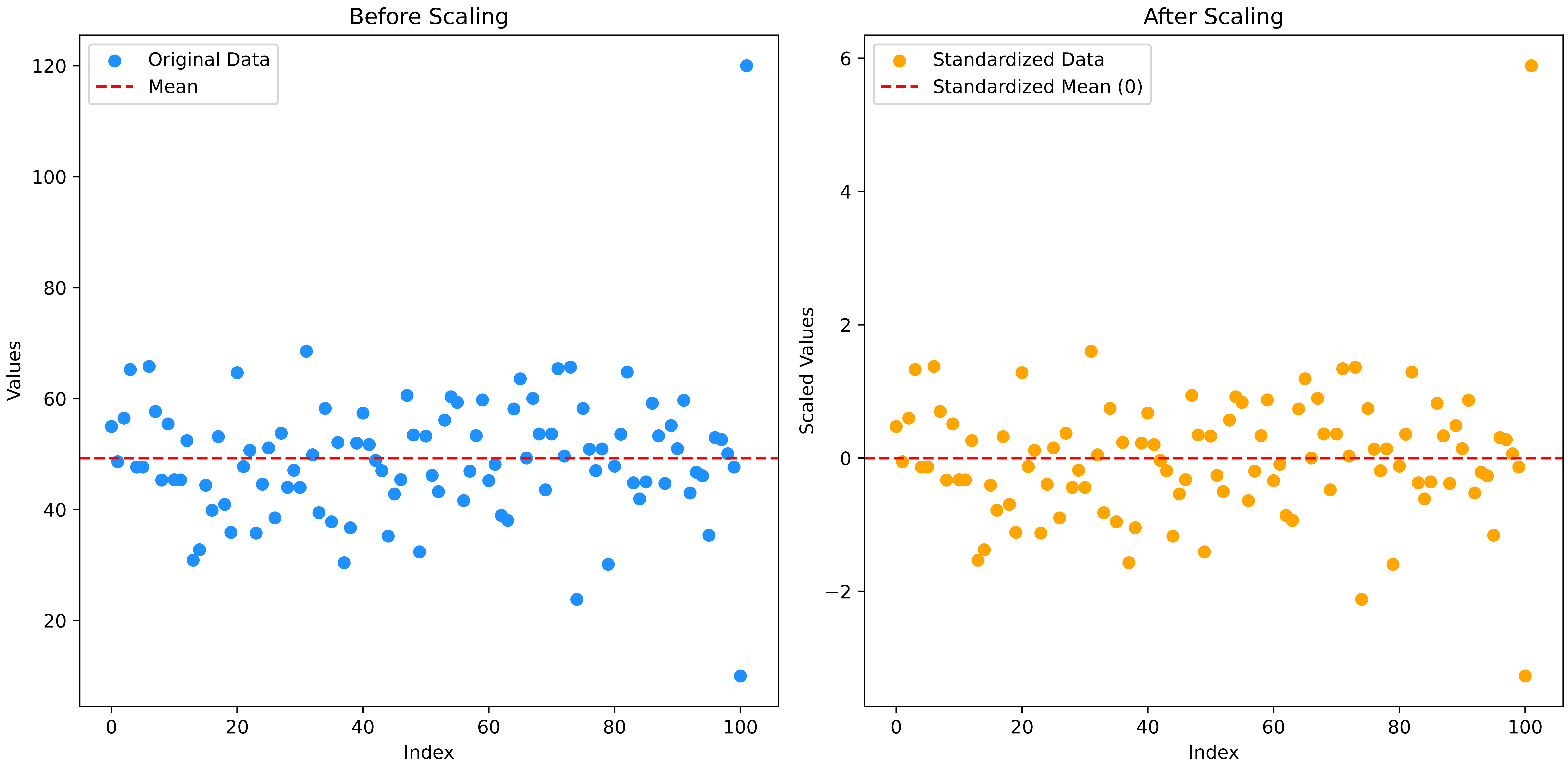
Key Points:

- Suitable for algorithms assuming Gaussian distribution
- Used when features have different variances



Data Scaling

Before & After Standardisation



Data Scaling

Standardization

Example:

Dataset: $[54.96714153,\ 48.61735699,\ 56.47688538,\ 65.23029856,\ 47.65846625]$

Mean: 54.59002974325087

Standard deviation: 6.334621540984489

Dataset standardized: $[0.05953186,\ -0.94286181,\ 0.29786399,\ 1.67970079,\ -1.09423482]$

Data Scaling

Normalization

Normalization rescales data to fit within a [0, 1] range.

Formula: $x' = \dfrac{x - x_{min}}{x_{max} - x_{min}}$

Where: x_{min} is the smallest value in the dataset

x_{max} is the largest value in the dataset

Example: Rescaling monthly sales figures.

Data Scaling

Normalization

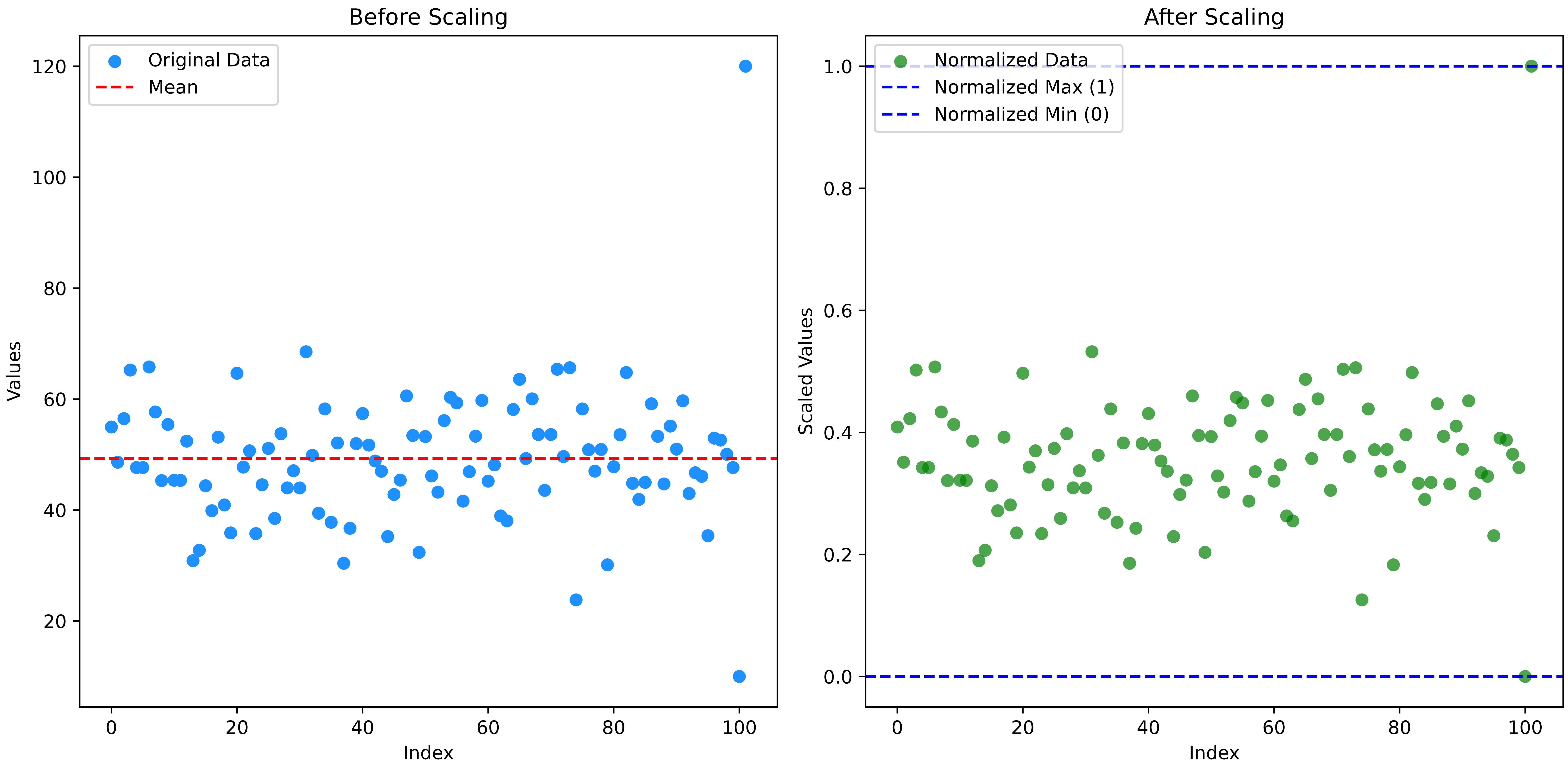
Normalisation is best suited for algorithms relying on distances

Key Points:

- Suitable for KNN, Neural Networks
- Ensures fair comparison of distances

Data Scaling

Before & After Normalisation



Data Scaling

Normalization

Example:

Dataset: [54.96714153, 48.61735699, 56.47688538, 65.23029856, 47.65846625]

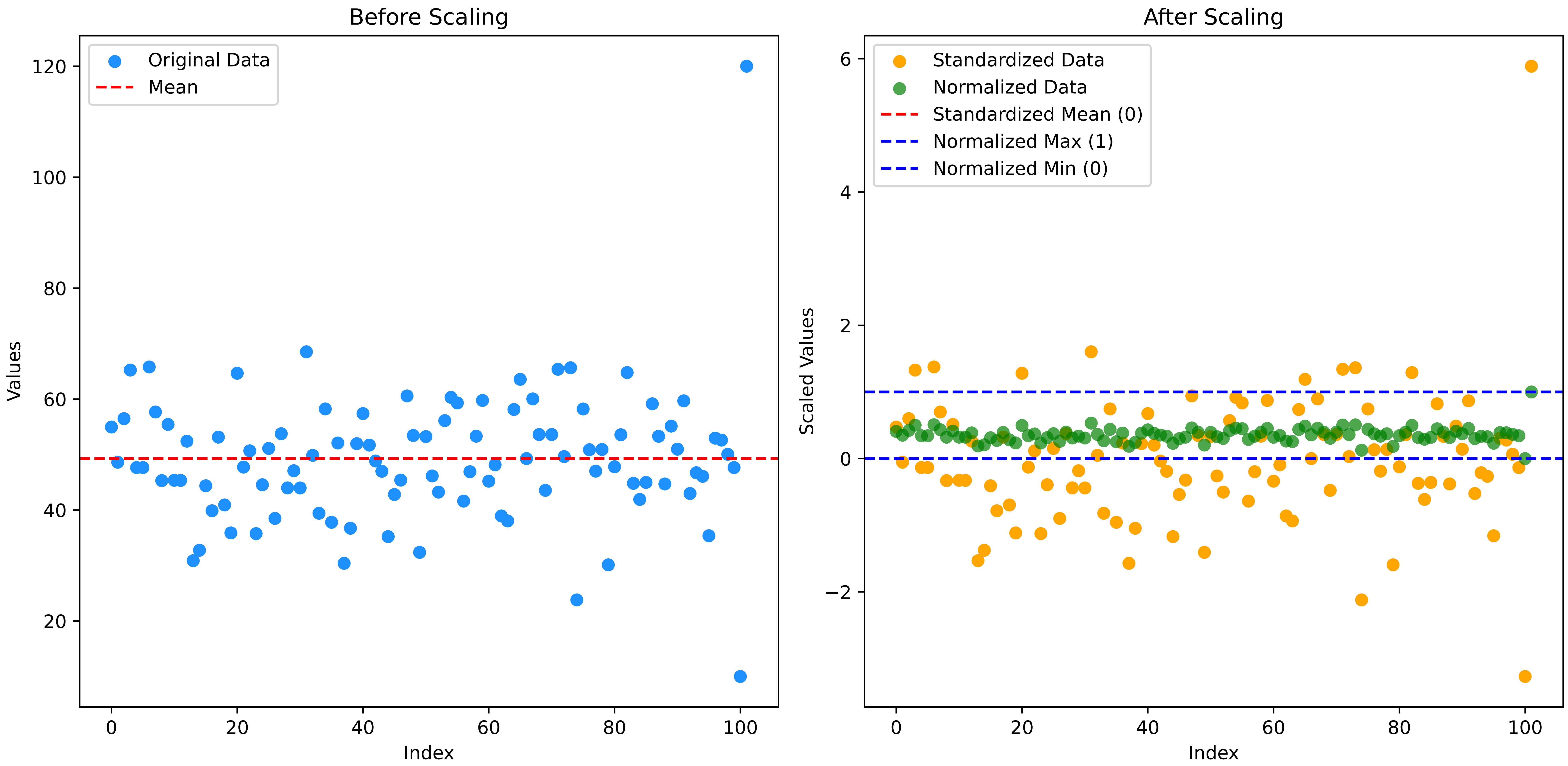
Max: 65.23029856408026

Min: 47.658466252766644

Dataset normalized: [0.41593131, 0.05456976, 0.50184972, 1, 0]

Data Scaling

Scaling vs Normalisation



Data Scaling

Good to Know

Algorithms Sensitive to Scaling:

- k-Nearest Neighbors (KNN)
- Support Vector Machines (SVM)
- Principal Component Analysis (PCA)
- Gradient Descent-based models

Data Scaling

Case Study: standardisation

Dataset: Heights and Weights

Problem: Large variance in features

Data Scaling

Case Study: standardisation

Dataset: Heights and Weights

Problem: <u>Large variance in features</u>

Solution: Apply z-score **standardization**

Data Scaling

Case Study: normalisation

Dataset: E-commerce user behavior

Problem: Features on different scales

Data Scaling

Case Study: normalisation

Dataset: E-commerce user behavior

Problem: Features on different scales

Solution: **Normalize** purchase frequency and session time

Data Scaling

Combining Scaling and Outlier Handling

Workflow:

- Detect and handle outliers.
- Apply appropriate scaling technique.
- Train machine learning models.

Data Scaling

Common Pitfalls

Key Points:

- Scaling before handling outliers
- Using the wrong scaling technique
- Forgetting to scale test data

Demo with
Notebook_Outliers_and_Data_Scaling.ipynb

Useful Links

- https://scikit-learn.org/stable/modules/outlier_detection.html
- https://scikit-learn.org/stable/modules/unsupervised_reduction.html