**✅ Anomaly**

An anomaly is **any data point or pattern that deviates from the norm** or expected behavior. In data analysis or machine learning, anomalies are often referred to as "outliers" or "weird data".

Examples:

* A sudden spike in sales on a random day.
* A negative age value in a dataset.
* A user who logs in 1000 times a day.
* Or yes — **duplicate records**, if they aren't supposed to be there.

**✅ Duplicate**

A duplicate is when **two or more identical records** appear in a dataset, either completely or partially. It’s one specific kind of anomaly — usually unintentional and needs cleaning.

**Summary:**

* **Anomalies** = unusual data (could be wrong, rare, or surprising)
* **Duplicates** = repeated data (can be a type of anomaly)

**Method to find anomalies in dataset:**

These assume your data follows a known distribution (like normal distribution).

* **Z-Score**:  
  Measures how far a value is from the mean in standard deviations.  
  z = (x - mean) / std\_dev Values with |z| > 3 are often considered anomalies.
* **IQR (Interquartile Range)**:  
  Based on percentiles. Outliers are typically: lower\_bound = Q1 - 1.5 \* IQR  
  upper\_bound = Q3 + 1.5 \* IQR