Handling outliers is an important part of data cleaning as these extreme values can heavily distort statistical analyses and machine learning models. Detection typically involves a mix of visualization and statistical techniques.

## 1. Visualization Techniques for Outlier Detection 👀

Visual tools are the first step, as they allow for quick identification of abnormal data points and provide context that purely statistical methods might miss.

| Visualization | Theory/Principle | Outlier Appearance |
| --- | --- | --- |
| **Box Plot (Box-and-Whisker Plot)** | Displays the distribution of data based on the five-number summary: minimum, first quartile (Q1), median (Q2), third quartile (Q3), and maximum. It clearly marks values outside the whisker boundaries. | Points that lie **outside the whiskers** are considered potential outliers. |
| **Histogram** | Shows the frequency distribution of a single numerical variable. | Outliers appear as **tiny bars far away** from the main body of the distribution, often creating a long, sparse "tail." |
| **Scatter Plot** | Displays the relationship between two numerical variables. | Outliers appear as **points that are isolated** and lie far away from the general cluster or trend of the other data points. Essential for detecting **multivariate outliers** (points that are normal in one dimension but extreme in combination). |

## 2. Statistical Techniques for Outlier Detection 📊

Statistical methods provide objective, quantitative criteria for classifying a data point as an outlier.

### A. The Z-score Method

**Theory:** The Z-score measures how many **standard deviations** a data point is away from the mean (μ) of the data. This method assumes the data is **normally distributed**.

Z=σ(x−μ)​

* x: The individual data point.
* μ: The mean of the column.
* σ: The standard deviation of the column.

Detection Rule:

Any data point with an absolute Z-score ∣Z∣>3 is often considered an outlier. This means the value lies outside 3 standard deviations of the mean, encompassing roughly 99.7% of the data if it were perfectly normal.

Example:

If the mean annual income is $60,000 with a standard deviation of $10,000, an income of $95,000 would have a Z-score of 3.5. Since 3.5>3, it's flagged as an outlier.

### B. The Interquartile Range (IQR) Method

**Theory:** The IQR is the distance between the first quartile (Q1, the 25th percentile) and the third quartile (Q3, the 75th percentile). This method is **non-parametric**, making it robust to non-normal distributions and less sensitive to the outliers themselves.

IQR=Q3−Q1

Detection Rule (Tukey's Fences):

A value is considered an outlier if it falls outside the following range:

* **Lower Bound:** Q1−(1.5×IQR)
* **Upper Bound:** Q3+(1.5×IQR)

Example:

If the IQR for the House Price column is $100,000, and Q1=$200,000 and Q3=$300,000:

* Lower Bound: $200,000−(1.5×$100,000)=$50,000
* Upper Bound: $300,000+(1.5×$100,000)=$450,000

Any house price **below $50,000** or **above $450,000** would be flagged as an outlier.

The **IQR method** is generally preferred over the Z-score method in most data science scenarios because it's **less susceptible to distortion** by the extreme outliers themselves.