**The 1.5 \* IQR Rule:**

The **1.5 \* IQR rule** is a commonly used method for identifying outliers in a dataset. It helps detect values that are significantly lower or higher than the majority of the data points, ensuring a more accurate understanding of the data distribution.

**The 1.5 \* IQR Rule:**

Outliers are typically defined as points that fall **below the lower bound** or **above the upper bound** of the dataset, calculated as follows:

* **Lower Bound** = Q1−1.5×IQRQ1 - 1.5 \times \text{IQR}Q1−1.5×IQR
* **Upper Bound** = Q3+1.5×IQRQ3 + 1.5 \times \text{IQR}Q3+1.5×IQR

Where:

* **Q1** is the 1st quartile (25th percentile).
* **Q3** is the 3rd quartile (75th percentile).
* **IQR** (Interquartile Range) = Q3−Q1Q3 - Q1Q3−Q1

**Purpose of the 1.5 \* IQR Rule:**

The reason for using this rule is to detect outliers in a systematic way. Here's why it's useful:

1. **Detects Extreme Values:**
   * Any value outside the range defined by Q1−1.5×IQRQ1 - 1.5 \times \text{IQR}Q1−1.5×IQR and Q3+1.5×IQRQ3 + 1.5 \times \text{IQR}Q3+1.5×IQR is considered **too far from the typical values**.
   * These values are seen as extreme and are likely to be **outliers**.
2. **Excludes Values that Could Skew Analysis:**
   * Outliers can distort statistical analyses, such as calculating the mean or variance, leading to misleading results. The 1.5 \* IQR rule ensures these extreme values are flagged for further investigation or exclusion.
3. **IQR Focuses on the Middle 50% of Data:**
   * The **IQR** measures the spread of the middle 50% of data, avoiding sensitivity to extreme values.
   * By extending the range of acceptable values by 1.5 times the IQR, the rule allows flexibility but still catches unusually large or small numbers.
4. **Robust to Non-Normal Data:**
   * Unlike the mean and standard deviation, which assume the data follows a normal distribution, the IQR is non-parametric. It doesn’t rely on the shape of the distribution, making the 1.5 \* IQR rule applicable in more general cases.

**Why 1.5 Specifically?**

* The **1.5 multiplier** is a balance between identifying outliers and allowing for natural variability in the data.
* The multiplier 1.5 is chosen because it allows for data to vary beyond the typical range without labeling too many values as outliers.
  + Multipliers smaller than 1.5 might mark too many points as outliers, while larger values could miss genuine outliers.

**Example:**

Let’s assume we have a dataset with these quartiles:

* **Q1 (25th percentile)** = 60
* **Q3 (75th percentile)** = 80
* **IQR** = Q3 - Q1 = 80 - 60 = 20

Now, we calculate the outlier boundaries:

* **Lower Bound** = Q1−1.5×IQR=60−1.5×20=60−30=30Q1 - 1.5 \times \text{IQR} = 60 - 1.5 \times 20 = 60 - 30 = 30Q1−1.5×IQR=60−1.5×20=60−30=30
* **Upper Bound** = Q3+1.5×IQR=80+1.5×20=80+30=110Q3 + 1.5 \times \text{IQR} = 80 + 1.5 \times 20 = 80 + 30 = 110Q3+1.5×IQR=80+1.5×20=80+30=110

In this example, any data point **below 30** or **above 110** would be considered an outlier based on the 1.5 \* IQR rule.

**How to Handle Outliers:**

* **Investigate the Cause**: Sometimes outliers can result from errors in data entry or measurement issues.
* **Decide to Keep or Remove**: Depending on the context, you can decide whether to exclude outliers from analysis or keep them if they represent valid extreme cases.

**Why 1.5 \* IQR is Important:**

* It’s a **simple** and **effective** method for identifying outliers.
* It avoids **overreacting** to slight variations in data but catches significant deviations that could distort analysis.
* It’s **non-parametric**, meaning it can be used regardless of the data distribution, unlike methods that rely on the mean and standard deviation (which assumes normality).