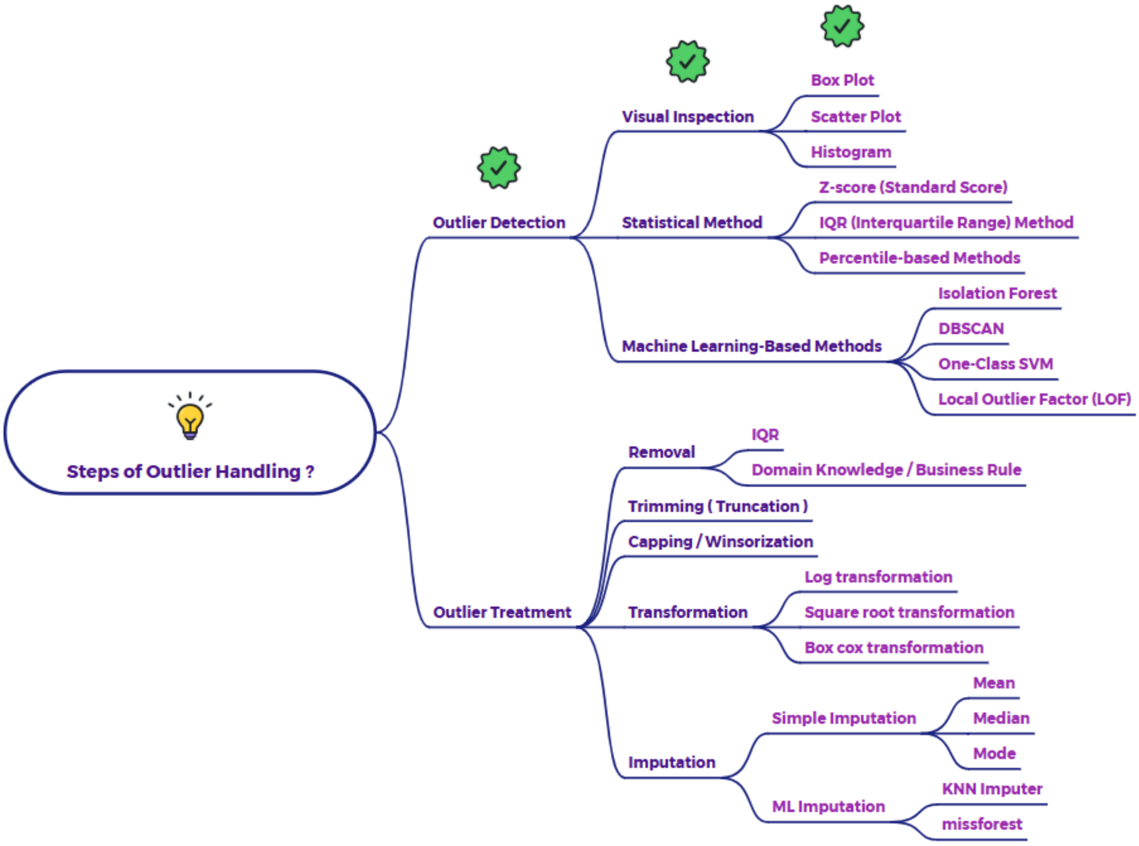
Explain Outlier detection through visual inspection (Box plot)



Outlier Detection: Visual Inspection - Box Plot

A box plot is a standardized way of displaying the distribution of data, making it easy to visually identify potential outliers. Outliers are points that fall outside the "whiskers" of the box plot.

Example:

Let's consider a dataset of the number of errors found per line of code during software testing for a small project:

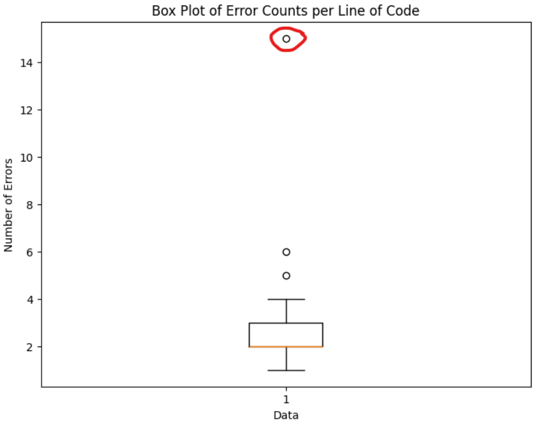
[2, 3, 1, 2, 4, 3, 2, 1, 5, 2, 3, 1, 2, 4, 3, 2, 1, 15, 2, 3, 1, 2, 4, 3, 2, 1, 6, 2, 3, 1]

In this dataset, the values 15 seem unusually high compared to the rest. Let's see how a box plot would highlight them.

1. **Calculate the Quartiles and IQR:**
   - First, sort the data: [1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 4, 4, 4, 5, 6, 15] (after removing duplicates for easier calculation, the principle remains the same for the full dataset)
   - Q1 (25th percentile) ≈ 1.75
   - Median (50th percentile) = 2.5
   - Q3 (75th percentile) ≈ 3.75
   - IQR = Q3 - Q1 = 3.75 - 1.75 = 2
2. **Determine the Whisker Limits:**
   - Lower whisker limit = Q1 - 1.5 * IQR = 1.75 - 1.5 * 2 = 1.75 - 3 = -1.25
   - Upper whisker limit = Q3 + 1.5 * IQR = 3.75 + 1.5 * 2 = 3.75 + 3 = 6.75
3. **Identify Outliers:**
   - Any data point below -1.25 or above 6.75 will be considered a potential outlier.
   - In our original dataset (and the simplified sorted version), the values **15** is greater than the upper whisker limit of 6.75.

In a box plot visualization of this data:

- The box would stretch approximately from 1.75 to 3.75.
- The median line would be around 2.5.
- The lower whisker would extend to the lowest value within the lower limit (likely 1).
- The upper whisker would extend to the highest value within the upper limit (likely 5).
- The values **15** would be plotted as individual point above the upper whisker, clearly indicating them as potential outliers.



Conclusion:

Box plots provide a simple yet effective visual way to identify data points that are significantly different from the central tendency and spread of the data, making outliers stand out beyond the whiskers.