Handing Outliers

What are Outliers?
Outliers are values that lie an abnormal distance away from most other data points or the overall data pattern.

They can occur because of:

- Errors (e.g., typos or faulty sensors),
- Genuine rare events, or
- Unique contextual conditions.

Three Main Type of Outliers:

Global Outliers

- Definition: Values that are drastically different from all other data points.
- Example: A height of 7.9 meters in a dataset of human heights around 1.6–1.9 meters.
- Handling: Usually removed, as they often result from data entry errors or measurement mistakes.

Contextual Outliers

- Definition: Data points that are normal in one context but abnormal in another.
- Examples:
    - A sudden spike in movie sales years after release.
    - A 2.5-meter "mammal" that turns out to be a mouse — normal for mammals generally, but not for mice.
- Common In: Time-series or categorical data.
- Handling: Evaluate the context before deciding whether to keep or remove them.

Collective Outliers

- Definition: A group of data points that collectively behave abnormally but follow a similar pattern.
- Example: A parking lot full of cars after store hours — unusual unless there's an event happening.
- Handling: Often reveal special circumstances or events worth investigating.

Detecting Outliers

Visualization is one of the best tools — using line graphs, bar charts, or scatter plots to spot “spikes” or “dips” that stand out.

Dealing with Outliers

- Always analyze in context — determine whether the outlier reflects an error or an important real-world event.
- Consider ethical implications: removing an outlier might make your results “look cleaner” but could distort the truth.
- Example:
  - A retail company finds one month of unusually low sales.
  - Instead of deleting the data, the analyst investigates and learns that staff shortages and a new product launch caused the drop.
  - The insight helps the business plan better in the future.

Outliers are unusual data points that differ greatly from the rest of the dataset. There are three main types — global, contextual, and collective. Detecting and analyzing them through visualization and careful reasoning ensures that conclusions remain accurate and ethical. Instead of simply removing outliers, data professionals should investigate their causes and consider how they affect the story the data tells.