EDA-Outlier Detection

Outlier detection is a crucial step in Exploratory Data Analysis (EDA) that involves identifying and analyzing unusual or anomalous data points in your dataset. These outliers can be errors, noise, or even valuable insights waiting to be discovered.

Types of Outliers:

1. **Univariate Outliers**: Outliers that are significantly different from the mean value when looking at a single feature.
2. **Multivariate Outliers**: Outliers that don't fit the general trend or pattern in multiple features.

Common Techniques for Outlier Detection:

1. **Z-Score Method**: Calculate the Z-score for each data point, and if it's greater than 3 (or another threshold), flag it as an outlier.
2. **Density-Based Methods**: Use algorithms like DBSCAN to identify clusters and outliers based on density.
3. **Statistical Methods**: Use statistical tests like Kolmogorov-Smirnov or Shapiro-Wilk to detect outliers.

Example:

Suppose we have a dataset of car prices in the United States, with features like:

Price (thousands)	Engine Size (cubic inches)
25.5	1997
15.2	1973
40.8	3020
22.1	2500
99.9	4999

In this example, the price of $99.9 seems unusually high compared to the other values in the dataset.

Outlier Detection Technique:

Using the Z-score method with a threshold of 3:

- Calculate the mean price (μ) = 25.5
- Calculate the standard deviation (σ) = 8.4
- For each data point, calculate its Z-score: Z = (price - μ) / σ

For the data point with price $99.9:

Z = (99.9 - 25.5) / 8.4 ≈ 10.3

Since the Z-score is greater than 3, we can flag this data point as an outlier.

Next Steps:

- Investigate why this data point is so high.
- Verify if it's a genuine outlier or just an error.
- Decide whether to keep or remove the outlier from your analysis.

Remember, outlier detection is not always about removing outliers but also learning from them and making more accurate predictions.

Curated by Brajesh Kumar