Start-Tech Academy

Outlier Treatment

Outlier is a commonly used terminology by analysts and data scientists, Outlier is an observation that appears far away and diverges from an overall pattern in a sample.

Outlier Treatment

Reasons

- Data Entry Errors
- Measurement Error
- Sampling error etc

Impact

- It increases the error variance and reduces the power of statistical tests

Solution

- Detect outliers using EDD and visualization methods such as scatter plot, histogram or box plots
- Impute outliers

Outlier Treatment

Example

	Without Outlier	With Outlier
Data	6,6,6,4,4,5,5,5,5,7,7	6,6,6,4,4,5,5,5,5,7,7,300
Mean	5.45	30.0
Median	5	5.5
Mode	5	5
Standard deviation	1.04	85.03
Variance	1.08	7230.10

Outlier Treatment

Methods

1. Capping and Flooring

- Impute all the values above 3* P99 and below 0.3*P1
- Impute with values 3* P99 and 0.3*P1
- You can use any multiplier instead of 3, as per your business requirement

2. Exponential smoothing

- Extrapolate curve between P95 to P99 and cap all the values falling outside to the value generated by the curve
- Similarly, extrapolate curve between P5 and P1

3. Sigma Approach

- Identify outliers by capturing all the values falling outside $\boldsymbol{\mu \mp x\sigma}$
- You can use any multiplier as x, as per your business requirement