Dealing with outliers is the second part of the outlier handling process, after detection. The methods chosen depend on the outlier's cause and the goals of the analysis.

## Dealing with Outliers: Treatment Methods 🩹

The primary methods for addressing outliers are **Winsorization**, **Trimming**, and **Data Transformation**.

### 1. Winsorization (Capping)

**Theory:** Winsorization replaces extreme outlier values with the nearest values that are considered non-outliers. The technique caps the data, limiting the effect of the extreme values on statistical measures while retaining the majority of the data points.

* **How it Works:** You define a percentile range (e.g., 1st and 99th percentiles).
  + Any value below the lower percentile threshold is set equal to that threshold value.
  + Any value above the upper percentile threshold is set equal to that threshold value.

| Technique | Description | Rationale |
| --- | --- | --- |
| **Winsorization** | **Replaces** extreme values at the tails of the distribution with a less extreme value (a "cap"). | Reduces the skew and variance caused by outliers **without deleting** any data points. Best when the data points are considered valid but disproportionately large/small. |

Example:

In a Salary dataset:

* 1st percentile (Lower Cap): $30,000
* 99th percentile (Upper Cap): $250,000
* A salary of **$15,000** is replaced by **$30,000**.
* A salary of **$5,000,000** is replaced by **$250,000**.

### 2. Trimming (Truncation)

**Theory:** Trimming, or truncation, involves **removing** the extreme values at the tails of the distribution. This is equivalent to listwise deletion but specifically for outliers.

| Technique | Description | Rationale |
| --- | --- | --- |
| **Trimming** | **Deletes** the observations that fall outside a specified percentile range. | Most effective when outliers are clear measurement errors or recording mistakes. It's the simplest method but results in **loss of sample size** and potential **selection bias**. |

Example:

If you trim 5% of your data, you remove the bottom 2.5% and the top 2.5% of values.

* This is often done based on the **IQR method's fences**—if a value is outside the 1.5×IQR range, the entire row is deleted.

### 3. Transformations to Reduce Impact

**Theory:** Mathematical transformations are applied to the entire feature column to change the shape of its distribution, pulling large outliers closer to the mean and making the data more symmetrical (closer to normal). This is essential for models that assume normality (e.g., Linear Regression).

| Transformation | Formula | Effect on Outliers |
| --- | --- | --- |
| **Log Transformation** | new\_value=log(x) | **Most common.** Heavily compresses the upper end of a positively skewed (right-skewed) distribution, effectively dampening the impact of large positive outliers. |
| **Square Root** | new\_value=x​ | A gentler transformation than the log; also used for reducing positive skew and variance. |
| **Reciprocal** | new\_value=1/x | Reverses the direction of the relationship and heavily compresses large values (e.g., useful for very heavy-tailed distributions). |

Example:

If an outlier income is $5,000,000 and the mean is $60,000:

* The raw difference is huge.
* log(5,000,000)≈6.7
* log(60,000)≈4.8
* The log transformation dramatically reduces the magnitude of the difference, making the outlier less influential.

### 4. Domain-Specific Methods

These methods require **subject matter expertise** to determine the appropriate course of action, as some "outliers" are the most important data points.

| Domain Context | Description | Example |
| --- | --- | --- |
| **True Anomalies** | The outlier represents a genuine, high-impact, but rare event (e.g., a major fire, a viral sale, fraudulent activity). | In **Fraud Detection**, the rare transactions of $50,000 are the *target* of the analysis, not data to be cleaned. They are kept and flagged. |
| **Time Series** | Outliers that occur at specific, explainable times (e.g., holidays, system crashes). | A huge spike in retail sales on **Black Friday** is an outlier that must be retained but often modeled using a **dummy variable** (a feature indicating if it's a holiday). |
| **Error Verification** | Outliers are traced back to the source data collection or entry system. | A price of $999,999 for a coffee mug is clearly a typo. The domain expert may advise **deleting** the record or manually replacing it with a verified price. |