**METHODS OF HANDLING OUTLIERS THEIR ADVANTAGES AND DISADVANTAGES**

**Outliers** are values within a dataset that vary greatly from the others. They may be either larger or significantly smaller compared to the rest. As the result, they can distort the analysis and modelling process.

Among the causes of outliers:

* Data entry errors
* Measurement errors; due to equipment or technique.
* Natural variability; Data can be collected accurately. Events or people variability
* Data processing errors; Sampling, transformation or imputation
* Data corruption

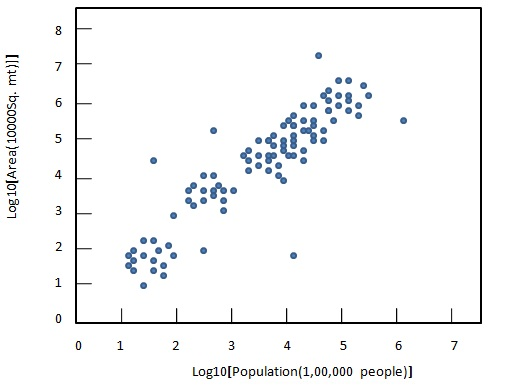
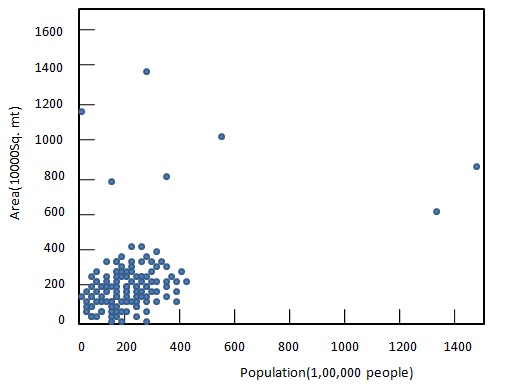
Types of outliers

* Univariate Outliers; variation of one variable e.g. height
* Multivariate outliers; variation of one of the two variables that correspond to each other e.g. weight and height 175cm to 50 kg
* Global outliers;
* Contextual outliers; variation of data in context of application
* Collective outliers; subset of data completely different from the entire data set
* Intentional outliers; Used to test outlier detecting methods

The following are among the methods of handling outliers:

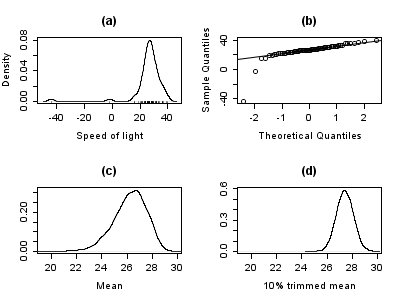
1. **Removing outliers**:
   * **Advantages**:
     + Can improve the performance of some algorithms by reducing noise or the sensitivity to outliers
     + Simplifies the dataset, leading to faster computation times.
   * **Disadvantages**:
     + Can lead to loss of valuable information if outliers contain important insights.
     + May not be appropriate for small datasets where each data point is valuable.
     + Can potentially introduce bias if outliers are removed without proper justification. Meaning there may be no clear reasons.
2. **Transformations** (e.g., log transformation, square root transformation) mathematical formulae or equations:
   * **Advantages**:
     + Can make the data distribution more symmetric and improve model performance.
     + Mitigates the impact of extreme values.

Log transformation example



* + **Disadvantages**:
    - Might not work well if the data contains zero or negative values.
    - Interpreting results becomes more complex due to the transformation.

1. **Binning**:
   * **Advantages**:
     + Reduces the impact of outliers by grouping values into bins.
     + Makes the data more robust to extreme values.
   * **Disadvantages**:
     + Loss of granularity in the data due to binning.
     + Optimal bin size selection can be subjective.
2. **Capping**:
   * **Advantages**:
     + Replaces extreme values with a predefined upper or lower bound.
     + Reduces the impact of outliers while retaining data integrity.
   * **Disadvantages**:
     + Can introduce bias if not applied carefully.
     + Choosing appropriate cut-off values can be challenging.
3. **Imputation**:
   * **Advantages**:
     + Fills in missing values with estimated values, which can reduce the impact of outliers.
     + Maintains the original data size and structure.
   * **Disadvantages**:
     + Imputed values might not accurately represent the true underlying data distribution.
     + Imputation methods might struggle with high-dimensional data.
4. **Robust statistical methods**:



* + **Advantages**:
    - Utilizes statistics less affected by extreme values (e.g., median instead of mean).
    - Provides more stable estimates even in the presence of outliers.
  + **Disadvantages**:
    - Some robust methods might be computationally intensive.
    - Results could be less intuitive for those not familiar with robust statistics.

The best technique to use depends on the characteristics and objectives of the data and the analysis. Factors such as the type and size of the data, the definition and impact of outliers and anomalies, and the trade-off between complexity and performance. Different techniques may work better for different types of data e.g. numerical, categorical or mixed data as well as different sizes of data.