Outliers

Outliers: are the extreme values for the specific column which affects the generalization of the data and model.

- to find outliers:

1. Calculate the first quartile (Q1) and third quartile (Q3) of the data for each column.
2. Calculate the IQR for each column by subtracting Q1 from Q3.
3. Calculate the lower bound and upper bound for detecting outliers for each column.
   - Lower bound: Q1 - 1,5 * IQR
   - Upper bound: Q2 + 1,5 * IOR
4. Identify any out points that fall outside the lower and upper bounds for any of the columns. These can be considered as outliers.

Data Scaling

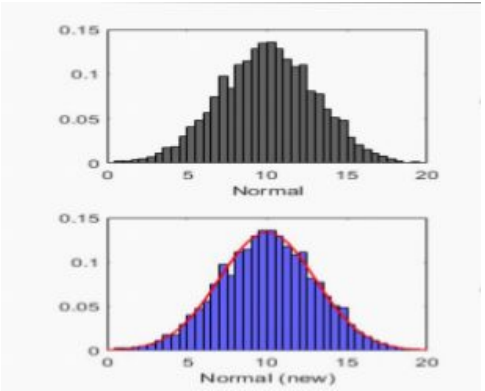
- Types:

1. **Standerscaler**:

It calculates the mean and standard deviation of the data set and normalize it by subtracting the mean and dividing by standard deviation.

$$z = \frac{x - \mu}{\sigma}$$

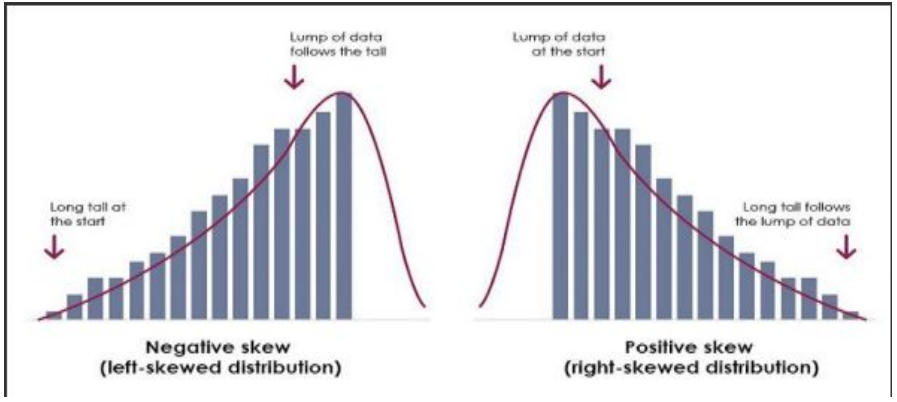
- it is often used for data, that have normal distribution. (Regression task)



2. **MinMax Scaler**:
IT is scaled the data set between 0 and 1, the maximum and minimum values in the scaled data set are 1 and 0.

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

- it used for data that is skewed distribution.



3. **Robust Scaler:**
It removes median and scale the data according to quantile range IQR.

IQR: is the range between the 1st quartile(25th quantile) and 3rd quartile(75th quantile).

· It is often used for data that has outliers, or is heavily skewed.

* both minmax and Robust can be use for classification task.

$$X_{\text{scale}} = \frac{x_i - x_{\text{med}}}{x_{75} - x_{25}}$$