0.0.1 Scaling

To normalize the data we first apply Z-Score standardization - we get the difference of each value from the mean of its column and divided by the standard deviation.

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

- **z** - the normalized value
- **x** - value of the row
- $\bar{x}$ - mean of the column
- σ - standard deviation

0.0.2 Outlier Detection and Clearing

Next we detect and clear some outliers in the dataset using Z-Score and a modified 5% 95% method. With the Z-Score method we're erasing rows containing score above or equal to 4. This is followed by a targetted clearing of outliers that have been discovered during Exploratory Data Analysis. Specifically we target columns Pixels_Areas, X_Perimeter, Y_Perimeter, Sum_of_Luminosity, Outside_X_Index, LogOfAreas and erasing rows above 98th percentile.

0.0.3 Feature Selection

At this point the data is split into Binary and Multiclass dataset. On the binary data we apply the Random Forest feature selection targetting the column Class to discover and rank the best non-linear predictors and discard the rest.