**Data PreProcessing:**

Data PreProcessing is a very important step in machine learning. The goal of data preprocessing is to prepare and clean the raw dataset so that efficiency and accuracy of the machine learning algorithms can be maximized.

Similarly before processing our dataset through the different machine learning algorithms we had to prepare our dataset for optimal use.

The Processes we had to go through to prepare our dataset were:

1. Null Value Replacement: Real World datasets can have missing data which needs to be replaced by the mean value of the missing attribute. We checked our dataset for missing values and as the dataset did not have any missing values we did not have to do anything.
2. Skewness Reduction:

Skewness is the measure of asymmetry of the probability distribution of an attribute. Excessive skewness can lead to bias in the final model. For our dataset we first check each attribute for their skewness and for any attribute which has an absolute skewness value of greater than 1 we append that attribute column to a list named skewedCols.

Below is the list of 5 attributes each which have the highest and lowest skewness before skewness reduction:

Table

| **Attribute Name** | **Skew Value** |
| --- | --- |
| tqwt\_TKEO\_mean\_dec\_32 | 26.48258509147365 |
| tqwt\_TKEO\_std\_dec\_32 | 26.0620838755717 |
| tqwt\_TKEO\_mean\_dec\_33 | 24.9442663610479 |
| tqwt\_TKEO\_std\_dec\_33 | 24.283816068288253 |
| det\_TKEO\_mean\_3\_coef | 20.874872200923843 |
| det\_LT\_entropy\_shannon\_7\_coef | -21.4150979727166 |
| tqwt\_medianValue\_dec\_29 | -21.623643949849406 |
| tqwt\_skewnessValue\_dec\_24 | -22.684339987002808 |
| tqwt\_entropy\_shannon\_dec\_33 | -25.06135227711703 |
| tqwt\_entropy\_shannon\_dec\_32 | -25.672811274750888 |

Fig \*\*#\*\*. Attribute values with the highest and lowest skewness values before skewness removal

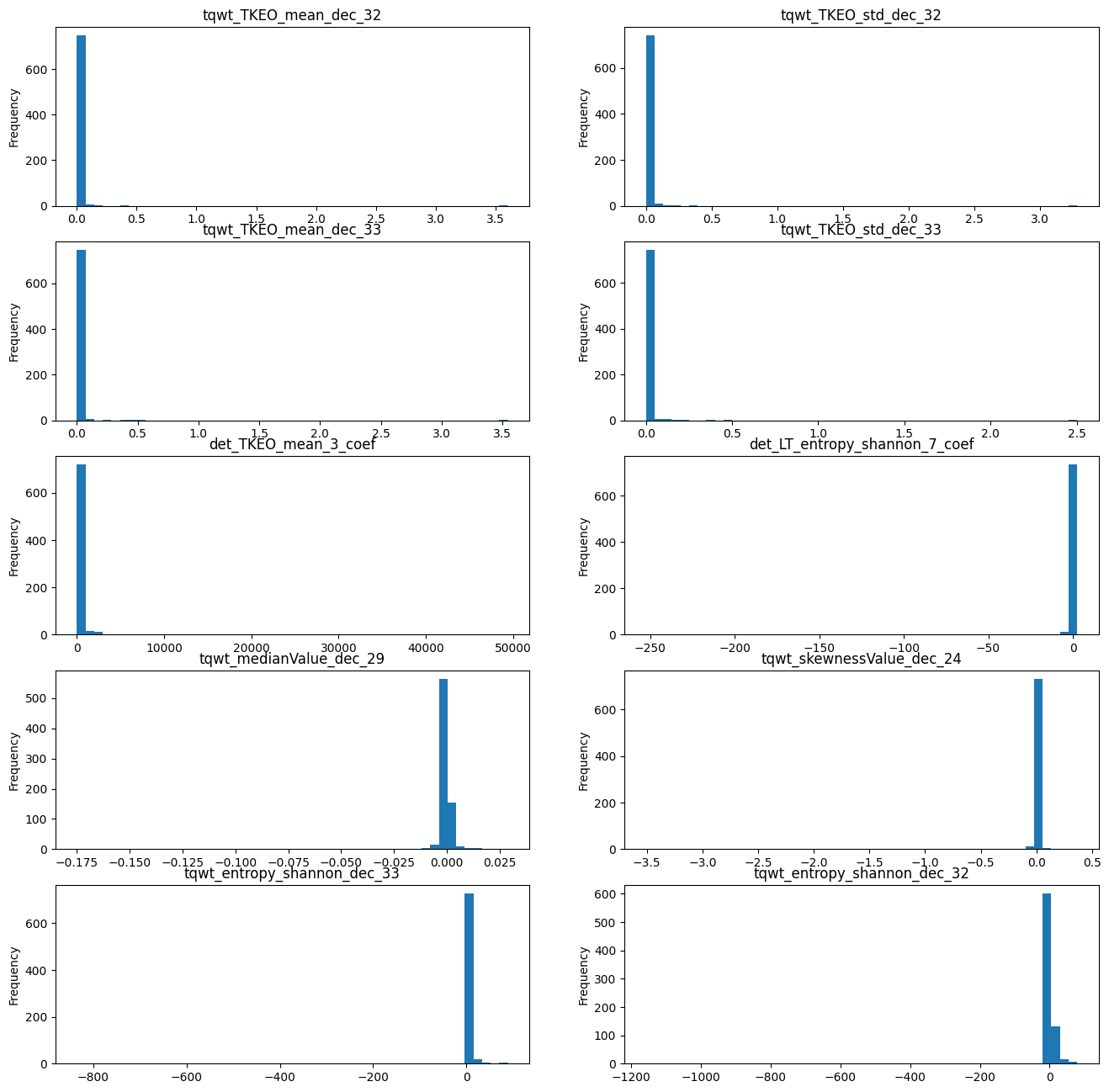


Fig. Attributes with the highest skewness before processing

Then we further classify the skewedCols based on whether that attribute contains any positive, zero or negative values into three separate lists named skewedCols\_PositiveVals, skewedCols\_ZeroVals, SkewedCols\_NegativeVals respectively.

Then for the attributes present in the list skewedCols\_PositiveVals we first use Box-Cox Transformation and to reduce the skewness of the attributes.

Box-Cox transformation works by applying a power function to the dependent variable which allows it to be transformed into a normal distribution and reduce its skewness.

We also use cube root transformation to reduce skewness of the attributes which are present in the lists skewedCols\_ZeroVals, SkewedCols\_NegativeVals. Cube Root transformation works by taking the cube root of each value of the attribute and making the attribute more closely resemble a normal distribution and reduce its skewness.

Below is the final skewness of those attributes:

Table

| **Attribute Name** | **Skew Value** |
| --- | --- |
| tqwt\_TKEO\_mean\_dec\_32 | 0.19846778387365632 |
| tqwt\_TKEO\_std\_dec\_32 | 0.01987384966738003 |
| tqwt\_TKEO\_mean\_dec\_33 | 0.3104628329189008 |
| tqwt\_TKEO\_std\_dec\_33 | 0.07507677689667193 |
| det\_TKEO\_mean\_3\_coef | 1.4434955464066024 |
| det\_LT\_entropy\_shannon\_7\_coef | -4.780452134423335 |
| tqwt\_medianValue\_dec\_29 | -0.31596805061181005 |
| tqwt\_skewnessValue\_dec\_24 | -1.690083153162617 |
| tqwt\_entropy\_shannon\_dec\_33 | -2.023628408215805 |
| tqwt\_entropy\_shannon\_dec\_32 | -2.120342043665047 |

Fig \*\*#\*\*. The attributes having the highest and lowest skewness after skewness reduction

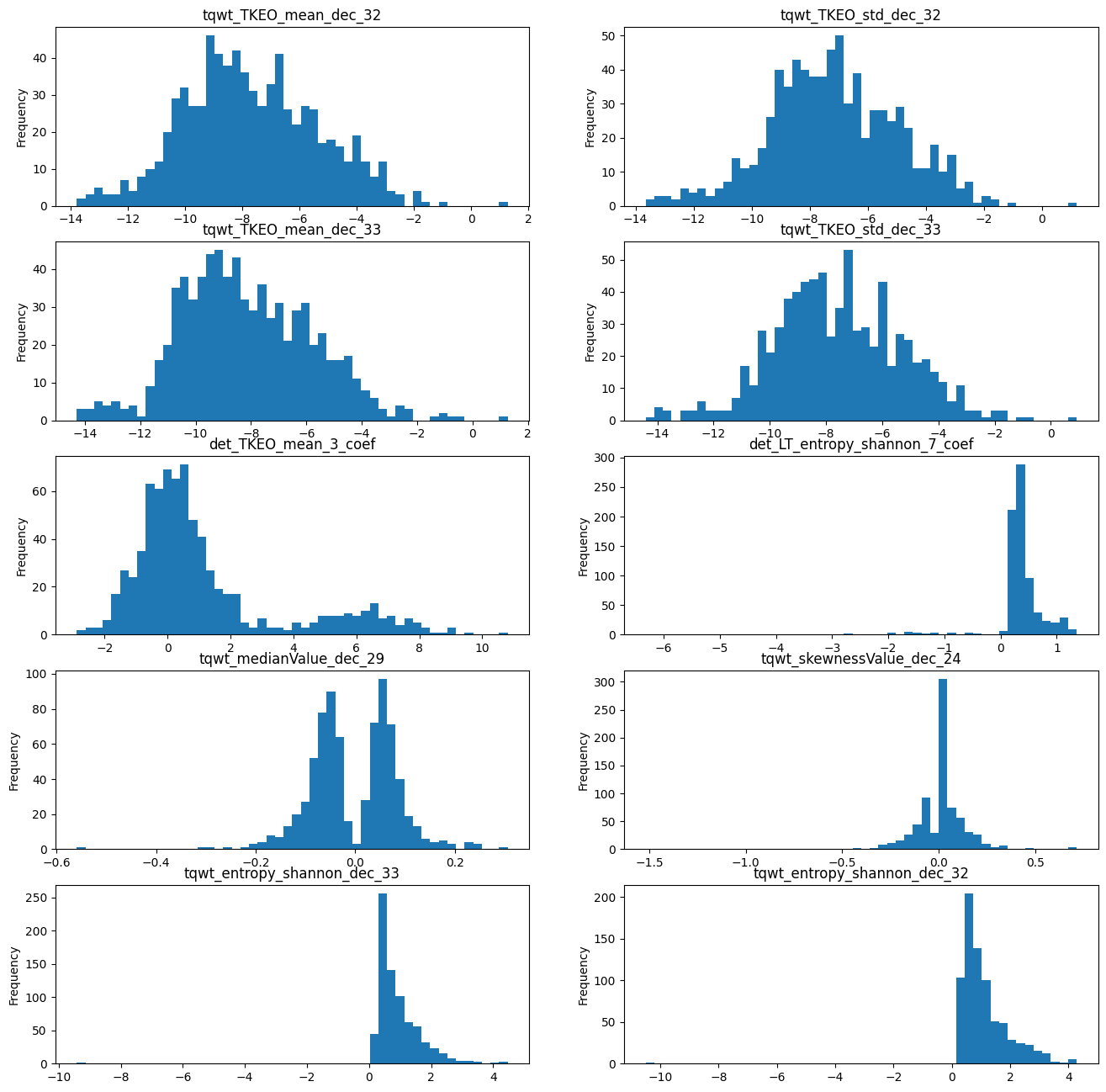


Fig. Attributes with the highest skewness after skewness Reduction

1. Kurtosis Reduction:

Kurtosis of an attribute is the measure of the peakedness/flatness of the probability distribution when measured relative to the normal distribution. The kurtosis of the normal distribution is considered zero. High kurtosis can lead to biases in the final model and hence we need to reduce kurtosis before processing our data. For our dataset we check for attributes which have kurtosis value greater than 3. For our dataset we did not have any such attribute and hence we did not need to do any kurtosis reduction.

1. Outlier Detection:

Outliers are data points in an attribute which are significantly different from the rest of the data. Outliers are bad for machine learning as they can bias the results of the final model. For our model We check each attribute and look for data points which are lower than 25 percentile or higher than 75 percentile of that attribute any data point which falls outside of this criteria is replaced by the mean of that attribute.

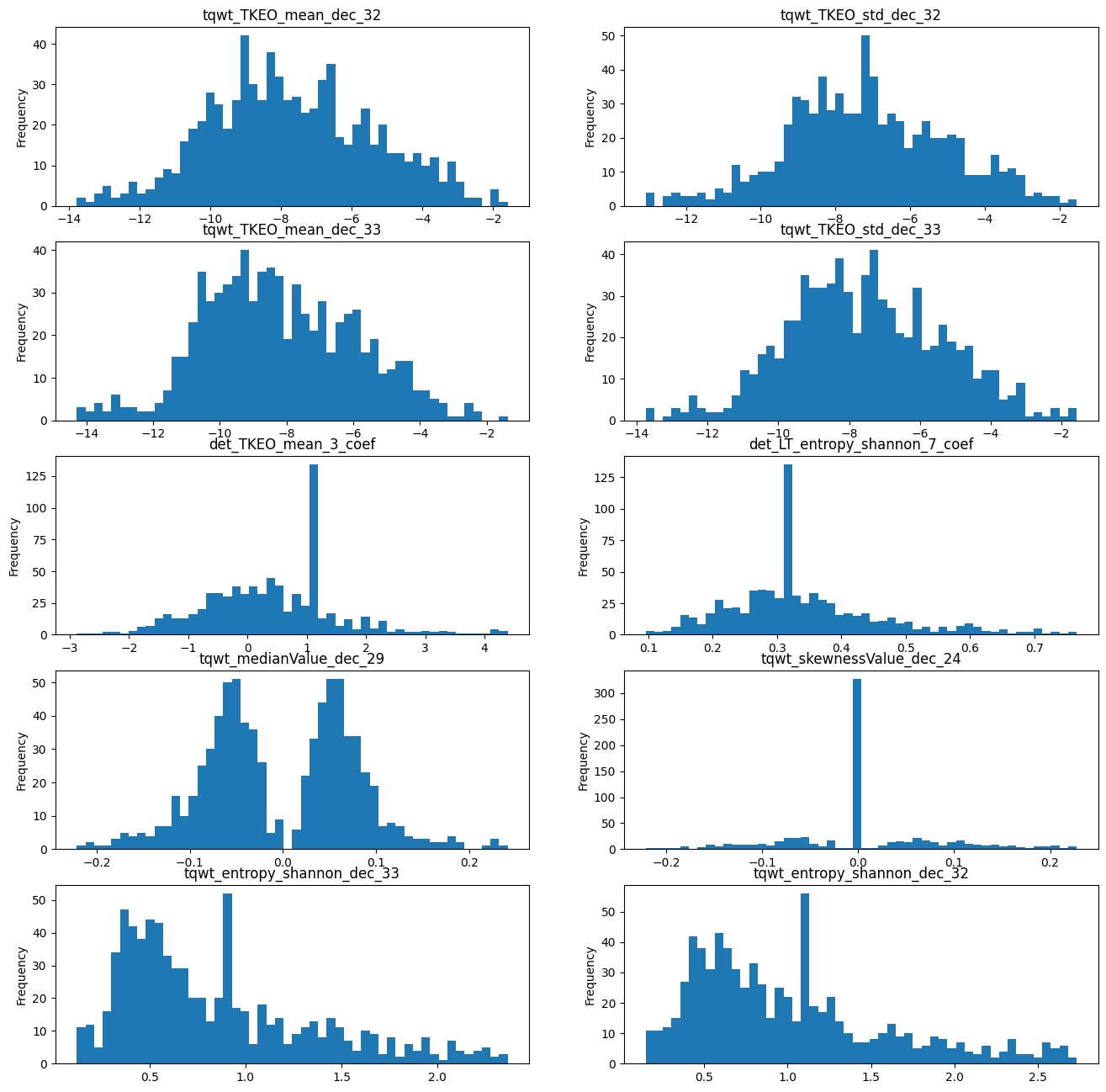


Fig. Attributes with the highest skewness after outlier Removal

1. PCA:

Principal Component Analysis or PCA in short is a widely used statistical technique used to reduce the dimensionality of a dataset. Dimension Reduction is a very important problem in machine learning as for smaller datasets the number of attributes can be higher than the number of instances which can lead to overfitting. PCA can be used to solve this problem by transforming the dataset into a different coordinate system where the dataset can be represented in a lower dimensional space.

PCA works by representing the original dataset in a new coordinate system where each dimension is a linear combination of the original attributes. This new coordinate system is selected in such a way that the first component captures the maximum variance in the dataset, the second component captures the maximum variance that is orthogonal to the first component and so on.

To perform PCA we first standardize the data points by subtracting the mean and divide by the standard deviation. We then compute the covariance matrix of the dataset which measures the relative variance of the attributes of the dataset.

Next we compute eigenvectors and eigenvalues of the covariance matrix. The eigenvectors represent the directions of the original feature space along which the variance of the data points are the highest, the eigenvalues represent how much the data points vary.

Finally we select the highest N eigenvalues and the corresponding eigenvectors and use them as the basis of the new coordinate system. We then project the data along the new coordinate system to obtain the lower dimensional representation of the dataset.

The algorithm for computing PCA can be represented as:

1. Standardize the data by subtracting mean and dividing by standard deviation.
2. Compute the covariance matrix of the standardized dataset.
3. Compute the eigenvalues and eigenvectors of the covariance matrix.
4. Sort the eigenvalues and eigenvectors according the eigenvalues in descending order
5. Select N largest eigenvalues and corresponding eigenvectors and use them as the basis for new coordinate system to represent the dataset in a lower dimensional feature space.

For our dataset we loop through different number of principal components and finally got 148 principal components which was giving us the optimal results.