Feature Engineering

Feature Engineering

T	C	S	E
Transformation	Construction	Scaling	Extraction

Feature Engineering

- Scaling
  - Normalisation
  - Standardisation
  - Robust Scaler

This Scales the data

- Encoding
  - Ordinal Encoding
  - OneHot Encoding

This Encodes the data

- Mathematical Transformations
  - Log Transform
  - Reciprocal Transform
  - Square/Square Root Transform
  - Power Transform

This Normalises the data

What is Feature Scaling?

Feature Scaling is a technique to standardise the independent feature present in the data

Types of Feature Scaling

- Normalization (With MinMaxScaler)
- Standardization (StandardScaler)
- Robust Scaling (RobustScaler)

Normalization

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

The goal of normalisation is to change the values of numeric columns in the data set to a common scale without distorting the data or losing information

The data transformed in a 0 to 1 scale

Standardisation / Z Score Normalisation

$$x_{standardised} = \frac{x - x_{mean}}{standard\ deviation}$$

The new data formed will have its mean equal to zero and standard deviation equal to one

When to apply Scaling

- K-Means
- KNN
- PCA
- ANN
- Gradient Descent

Few Points to Remember

- Normalization is good to use when you know that the distribution of your data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbors and Neural Networks.

- Standardization, on the other hand, can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Also, unlike normalization, standardization does not have a bounding range. So, even if you have outliers in your data, they will not be affected by standardization.

- There is no any thumb rule to tell you when to normalize or standardize your data. You can always start by fitting your model to raw, normalized and standardized data and compare the performance for best results.

Robust Scaler

$$x_{scaled} = \frac{x_i - x_{median}}{IQR}$$

$$IQR = Q_3 - Q_1$$

It is called Robust because it is Robust to outliers

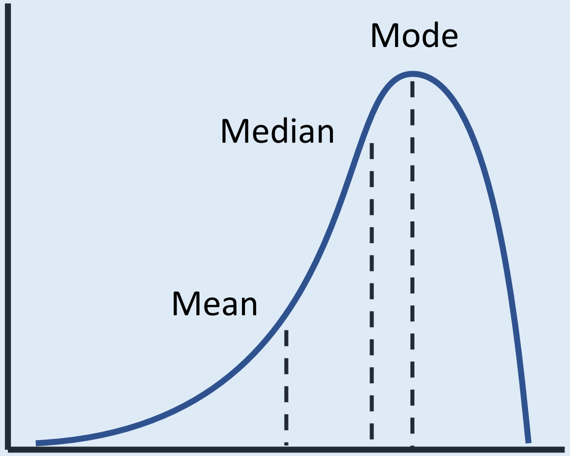
Feature Engineering

T	C	S	E
Transformation	Construction	Scaling	Extraction

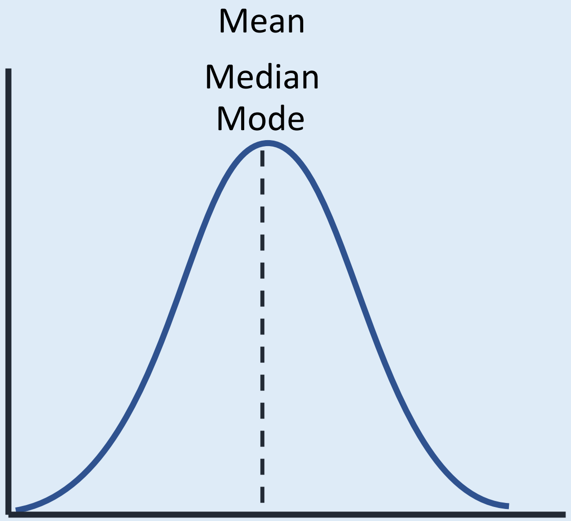
Mathematical Transformation

- These Transformations are used when the data is not normally distributed
- How to check if the data is normally distributed or not?
  - Use sns.distplot
  - Used pandas skew method
  - QQ Plots
  - Hypothesis Testing

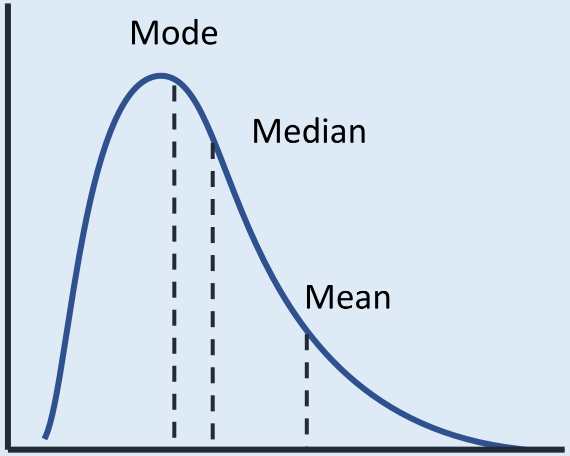
What is data skewness?



Left Skewed / Negatively Skewed



Normally Distributed

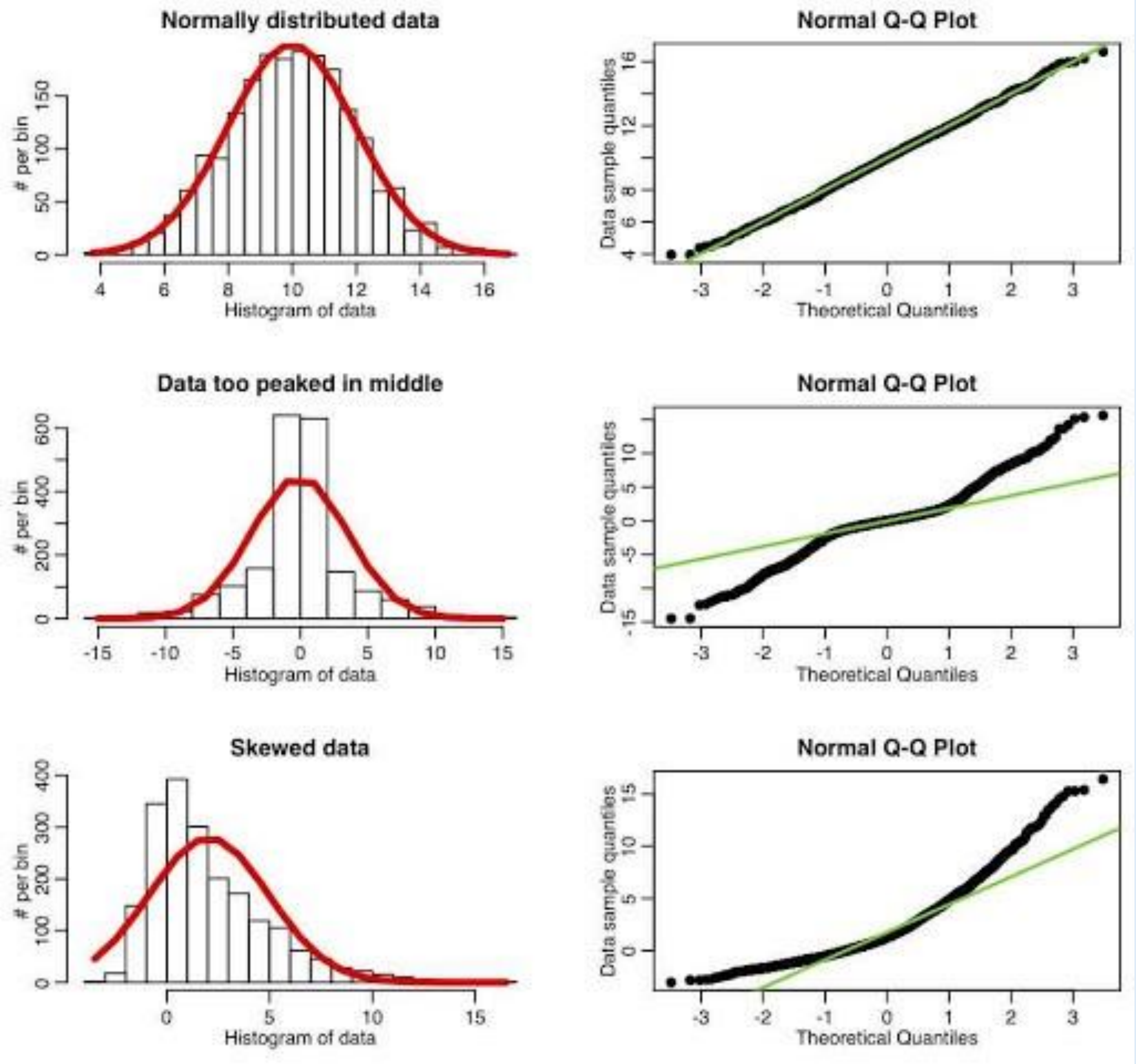


Right Skewed / Positively Skewed

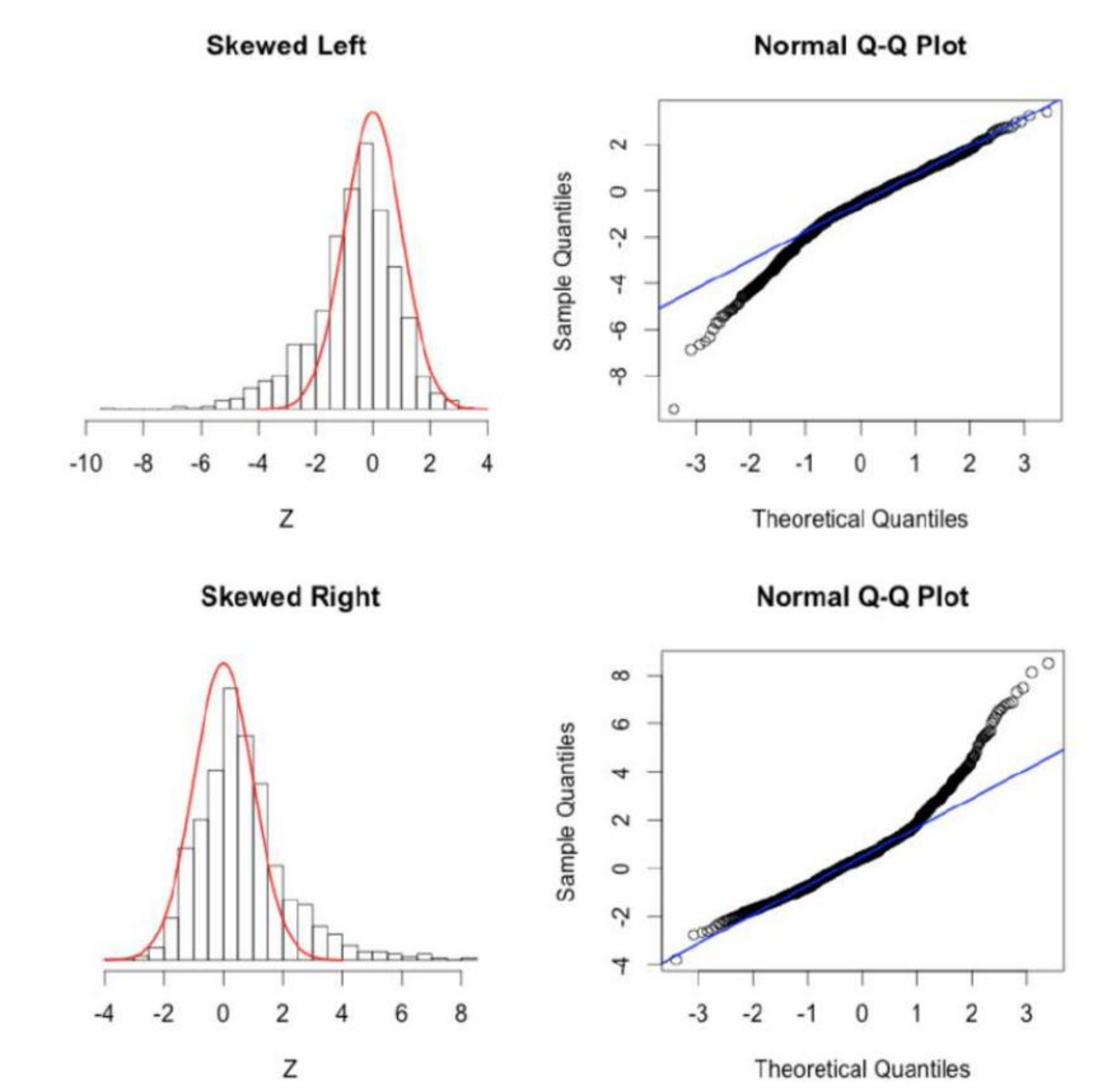
After Running pd.skew():

- If value is $-0.5\ to + 0.5$ the data is fairly normally distributed
- If the value is $-1\ to\ - 0.5$ the data is negatively skewed
- If the value is < -1 the data is highly negatively skewed
- If the value is $+0.5\ to\ 1.0$ the data is positively skewed distributed
- If the value $>\ +1$ the data is highly positively skewed

Data Distribution and QQ Plots



Data Distribution and QQ Plots



Hypothesis Testing

The statistical test conducted to check the normality of data are:

- Shapiro Test
- KS Test
- Normal Test

All these test assume the following null and alternate hypothesis as:

Null Hypothesis (H0): Data is normally distributed

Alternate Hypothesis (H1): Data is not normally distributed

We reject the null hypothesis if the P-Value is less than 0.05 and if it is greater than 0.05 technically we say that we do not have sufficient evidence to reject null hypothesis (that tantamount to rejecting the alternate and accepting the null hypothesis)

Which Mathematical Transformation to use?

- Log Transform if the data is right skewed
- Square Root Transform if the data is left skewed
- Reciprocal Transform (Data is left Skewed)
- Square transform (Should be experimented)
- Cube Root Transform (Should be experimented)

Handling Skewness of data notes by ChatGPT

Left Skewed Data:

Handling left-skewed data can be a bit more tricky than handling right-skewed data, as there are not as many straightforward solutions. Here are a few strategies you can use to handle left-skewed data:

1. **Transformation:** A common method for handling skewness is to apply a transformation to the data. A left-skewed data is usually solved by applying a reciprocal or square root transformation which helps to reduce the skewness. However, it's worth noting that the log transformation should not be used since it will make the skew worse.

2. **Remove outliers:** Left-skewed data often contains outliers that skew the distribution. Removing these outliers can help to reduce the skewness and make the data more symmetric.

3. **Use a different statistical model:** Some statistical models, such as the Poisson or negative binomial regression, are specifically designed to handle left-skewed data. These models can be used instead of linear regression, which assumes that the data is normally distributed.

4. **Class imbalance:** If you are working with a categorical variable, you might want to consider oversampling the minority class, or the one with the fewer observations, or using a weighted loss function on the training process.

5. **Truncation**: If the data includes a small number of very large values, you can consider truncating the data to remove the extreme observations, but this is useful only when the outliers are "far" from the rest of the observations.

Handling Skewness of data notes by ChatGPT

It's worth noting that there is no one-size-fits-all solution for handling skewness, and the best approach will depend on the specific characteristics of your data and the research question you are trying to answer. In all cases, it's essential to understand and interpret the implications of the transformation in the context of the problem.

Handling Skewness of data notes by ChatGPT

Right Skewed Data:

1.Transformation: A common method for handling skewness is to apply a transformation to the data. A right-skewed distribution can be transformed to a more symmetric distribution by taking the natural logarithm (ln) of the data. It helps to make the larger values less extreme and to bring the data closer to a normal distribution. However, it's worth noting that when applying a log transformation on the data that contains 0 or negative values it is necessary to add a small constant value to each observation.

2.Remove outliers: Right-skewed data often contains outliers that skew the distribution. Removing these outliers can help to reduce the skewness and make the data more symmetric.

3.Use different statistical models: Some statistical models, such as the gamma or inverse Gaussian regression, are specifically designed to handle right-skewed data. These models can be used instead of linear regression, which assumes that the data is normally distributed.

4.Truncation: If the data includes a small number of very large values, you can consider truncating the data to remove the extreme observations, but this is useful only when the outliers are "far" from the rest of the observations.

5.Binning: another alternative is to convert continuous variable into categorical variable. Also, it can be useful to remove the outliers and to analyze the data in a more granular way.

It's important to note that, after applying any of the above-mentioned methods, it's critical to check if the data is normally distributed and check if the assumptions of the statistical methods being used are met. The best approach will depend on the specific characteristics of your data and the research question you are trying to answer.

Outliers

- Data points which are far away from observed values are called as outliers
- Unusual data points which differs significantly form other datapoints are called as outliers

How outliers are introduced in the dataset?

- Data entry error
- Intentional errors
- Instrumental Error
- Natural Error
- Sampling Errors (Mixing data from wrong sources)

How to detect outliers?

- Z-Score method
- IQR Method
- Boxplot (Visualization Tool)
- Scatter Plot

Z - Score

$$x_i' = \frac{x_i - x_{mean}}{Standard\ Deviation}$$

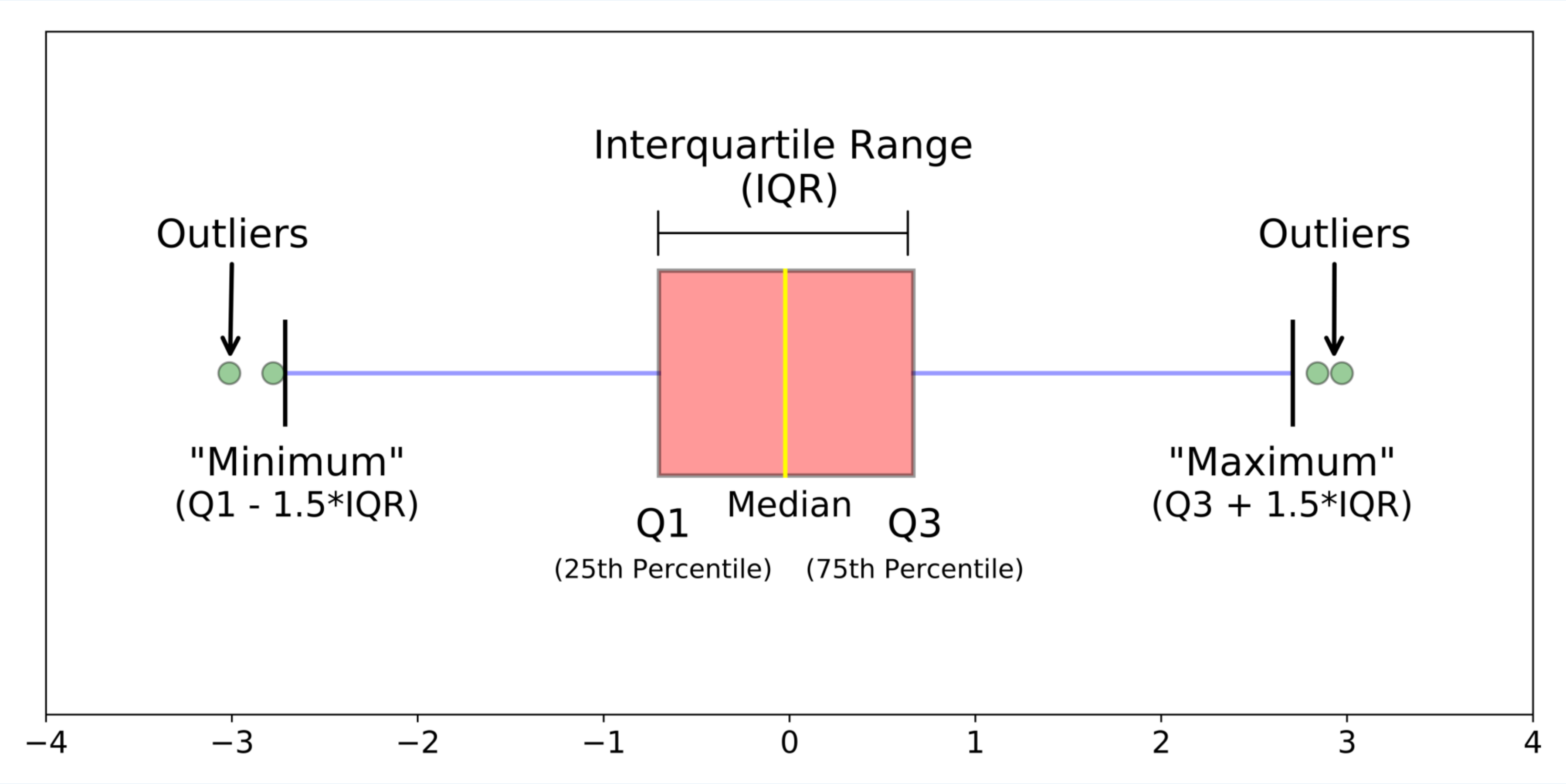
- Find the z score transform of the array / pandas column
- If, for a value in an array/column the Z – Score is less than -3 or greater than +3 that value shall be considered as an outlier
- In other words if the value is greater than (mean + 3 x standard deviation) or less than (mean – 3 x standard deviation) the value should be considered as an outlier

IQR Method

- First find the 1st quantile and 3rd quantile of the data
- Later, (3rd Quantile – 1st Quantile) = IQR
- If a value is greater than (3rd Quantile + 1.5*IQR) or less than

(1st Quantile – 1.5*IQR) the value is an outlier

Box Plot


Interquartile Range
(IQR)
Outliers
Outliers
"Minimum"
(Q1 - 1.5*IQR)
Median
Q1
(25th Percentile)
Q3
(75th Percentile)
"Maximum"
(Q3 + 1.5*IQR)
−4
−3
−2
−1
0
1
2
3
4

Scatter Plot

- This visualization tools doesn’t serve much purpose, better to use box plot

How to handle outliers?

- Delete Observation
- Imputation Technique
  - Mean
  - Median
  - Minimum Value
  - Maximum Value
  - Upper Tail
  - Lower Tail
  - Static Value
- Transformation Technique
  - Log
  - Square Root
  - Reciprocal
  - Cube Root

Which Algorithms are impacted by outliers?

- Linear Regression
- Logistic Regression
- K-Nearest Neighbour
- Support Vector Machine
- K-Means Clustering

Which algorithms are not impacted by outliers?

- Decision Tree
- Random Forest
- AdaBoost
- Gradient Boosting
- XGBoost
- Naïve Bayes