IMMUNE
CODING INSTITUTE

Supervised Machine Learning

Module 3.4

IMMUNE
CODING INSTITUTE

Module 3 Summary

SESSION	TITLE	TEACHER
1	ML Foundations	Juan
2	Regression Introduction and Practice	Juan
3	Classification Introduction and Practice	Carlos
4	**Feature Engineering and Selection for ML**	**Carlos**
5	Advanced Supervised Models 1	Carlos
6	Advanced Supervised Models 2	Carlos
7	Hands-on Practice	Carlos

IMMUNE
CODING INSTITUTE

Feature Engineering for Machine Learning

Feature Engineering

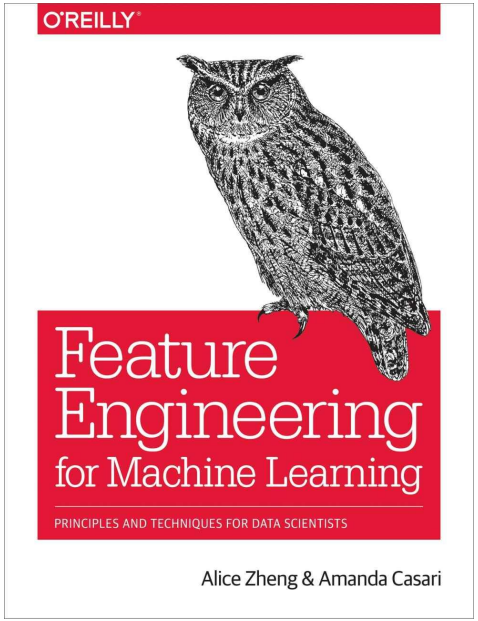
- Missing Imputation
- Categorical Encoding
- Variable transformation
- Filtering, Binning, Scaling
- Log transform and Power transform
- Outlier Engineering
- Datetime Engineering
- Discretisation

Feature Engineering

- **Missing Imputation**
- **Categorical Encoding**
- Variable transformation
- Filtering, Binning, Scaling
- **Outlier Engineering**
- Datetime Engineering
- Discretisation

Feature Engineering

- Missing Imputation
- Categorical Encoding
- Variable transformation
- Filtering, Binning, Scaling
- Log transform and Power transform
- Outlier Engineering
- Datetime Engineering
- Discretisation



Missing Data Imputation

- Is the act of replacing missing data with statistical estimates of the missing values

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

- The goal of any imputation technique is to complete a dataset without missing data so this features can be used to train a ML model

Missing Data Imputation

Numeric Variables

- Mean / Median Imputation
- Arbitrary value imputation
- End of tail imputation
- Correlation / Linear Models

Categorical Variables

- Frequency imputation
- Creating a missing category

Mean or Median imputation

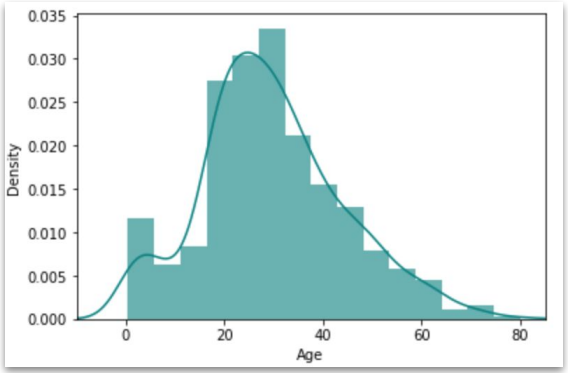
- Consist of replacing all occurrences of missing values (NA - Nulls) within a variable by the mean or median

- Titanic example:

```
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age            177
SibSp            0
Parch            0
Ticket           0
Fare             0
Cabin          687
Embarked         2
dtype: int64
```

Mean: 29.69

Median: 28



If the variable is normally distributed the mean and median are approximately the same

Mean or Median imputation

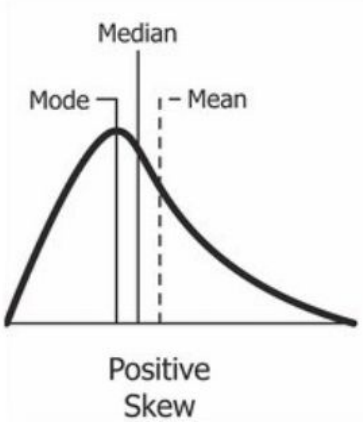
- Consist of replacing all occurrences of missing values (NA - Nulls) within a variable by the mean or median

- Titanic example:

```
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age            177
SibSp            0
Parch            0
Ticket           0
Fare             0
Cabin          687
Embarked         2
dtype: int64
```

Mean: 29.69

Median: 28



If the variable is skewed, the median is a better representation

Mean or Median imputation

Pros

- Easy to implement
- Fast way of obtaining complete datasets
- Can be integrated during model deployment

Cons

- Distortion of the original variable distribution
- Distortion of the original variance
- Distortion of the covariance with the rest of variables.

Mean or Median imputation

Final considerations

- The mean or median value should be calculated only in the train set and used to replace NA in both train and test sets.
- It is use when data is missing completely at random
- No more than 5% of the variable contains missing data*****

***** Depending on the problem

Arbitrary value imputation

- Consist of replacing all occurrences of missing values (Nulls - NA) within a variable by an arbitrary value.

- Typically used arbitrary values as: 0, 999, -999, -1 if the distribution is positive.

- This technique can be used for both numerical and categorical variables

Arbitrary value imputation

Pros

- Easy to implement
- Fast way of obtaining complete datasets
- Can be integrated during model deployment
- Captures the importance of being missing if there is one

Cons

- Distortion of the original variable distribution
- Distortion of the original variance
- Distortion of the covariance with the rest of variables.
- Need to be careful not to choose values too similar to the mean or median

Arbitrary value imputation

Final considerations

- Data are not missing at random.
- We want to flag the missing values with a different arbitrary value.

End of Tail imputation

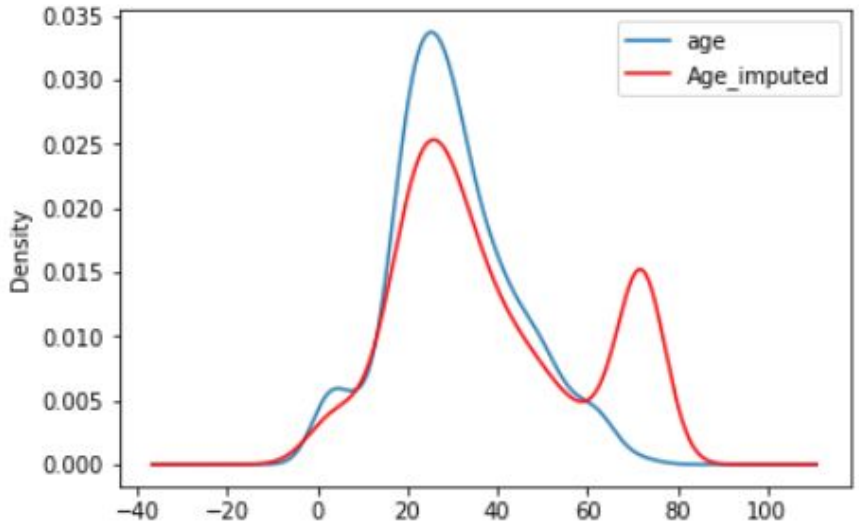
Is equivalent to arbitrary value imputation but automatically selecting arbitrary values at the end of the variable distribution.

- For normally distributed variables, we use the mean plus or minus 3 times the standard deviation

- For skewed distributions we can use the IQR proximity rule

End of Tail imputation

Titanic example



- ~20% of data is missing in Age

Original variable variance: 194
Variance after imputation: 427

Frequency Category Imputation

Mode imputation consist of replacing all occurrences of missing values (Null- NA) within a variable by the mode or the most frequent value

- For numerical and categorical variables
- Most common for categorical variables

Frequency Category Imputation

PassengerId	0
Survived	0
Pclass	0
Name	0
Sex	0
Age	177
SibSp	0
Parch	0
Ticket	0
Fare	0
Cabin	687
Embarked	2

S	644
C	168
Q	77

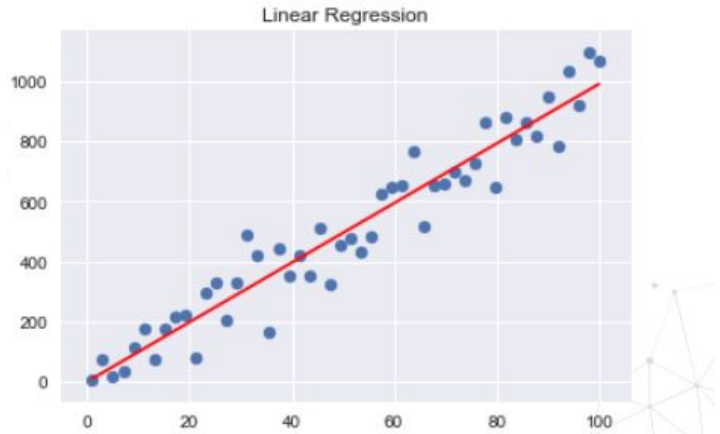
Missing Category Imputation

This method consists in treating missing data as an additional label or category of the variable.

- Missing observations are grouped in the newly created label “Missing”
- Is the most widely used method of missing data imputation for categorical variables.

Correlation Imputation

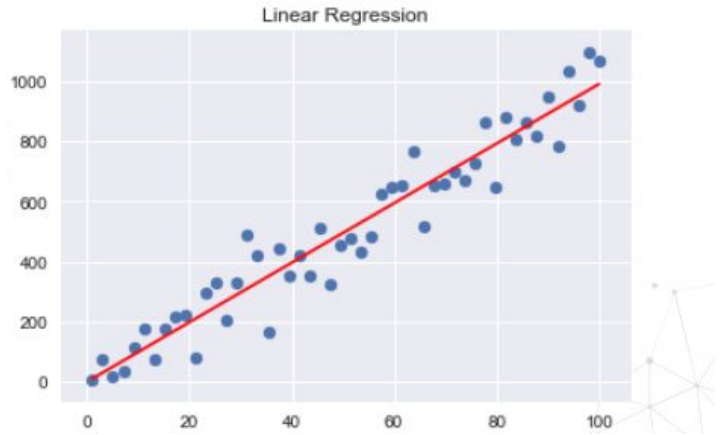
This method consists predict using correlation between variables the value of a missing variable



id	Age	Weight	Size
1	7	7.3	NA
2	4	NA	62cm
3	3	6.2	60cm
4	1	3	50cm
5	9	8.6	72cm

Correlation Imputation

This method consists predict using correlation between variables the value of a missing variable

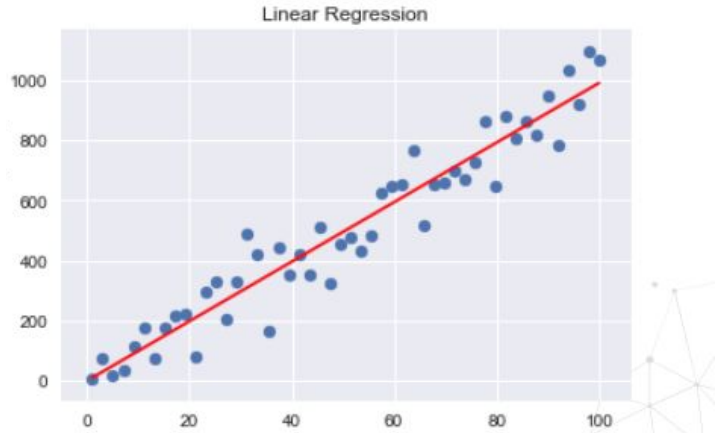


id	Age	Weight	Size
1	7	7.3	NA
2	4	NA	62cm
3	3	6.2	60cm
4	1	3.4	50cm
5	9	8.6	72cm

$$Size = \beta_0 + \beta_1 * Weight$$

Correlation Imputation

One of the variables must be eliminated to avoid collinearity in the final prediction.



$$Age = \beta_0 + \beta_1 * Size$$

id	Age	Weight	Size
1	7	7.3	68
2	4	NA	62cm
3	3	6.2	60cm
4	1	3.4	50cm
5	9	8.6	72cm

Categorical Encoding

Refers to replacing the category strings by a numerical representation.

The goal of categorical encoding is:

- To produce variables that can be used to train machine learning models
- To build predictive features from categories.

One hot encoding.

Count/Frequency encoding

Ordinal encoding

Ordered label encoding

Mean encoding

Binary encoding

One hot encoding

Consist in encoding each categorical variable with a set of boolean variables wich take values 0 or 1, indicating if a category is present for each observation.

City
Madrid
Barcelona
Toledo
Valencia

Madrid	Barcelona	Toledo	Valencia
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1

One hot encoding (k-1 variables)

A categorical variable should be encoded by creating k-1 binary variables, where k is the number of distinct categories.

City
Madrid
Barcelona
Toledo
Valencia

Madrid	Barcelona	Toledo	Valencia
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	0

One hot encoding (k variables)

There are a few occasion when it is better to encode variables into k dummy variables:

- when building a tree based algorithm
- when doing feature selection by recursive algorithms
- when interested in determine the importance of each single category

One hot encoding

Pros

- Makes no assumption about the distribution or categories of the categorical variable
- Keeps all the information of the categorical variable
- Suitable for linear models

Cons

- Expands the feature space
- Does not add extra information while encoding
- Many dummy variables may be identical, introducing redundant information.

Label / integer encoding

Consist on replacing the categories by digits from 1 to n, where n is the number of distinct categories of the variable

City
Madrid
Barcelona
Toledo
Valencia

City
1
2
3
4

Label / integer encoding

Pros

- Easy to implement
- Does not expand the feature space
- Can work well enough with tree based algorithms

Cons

- Not suitable for linear models
- Does not handle new categories in test set automatically
- Does not add extra information while encoding

Count / frequency encoding

- Categories are replaced by the count or percentage of observations that show that category in the dataset.

- Captures the representation of each label in a dataset

- Very popular encoding method in Kaggle competitions.

- Assumption: the number observations shown by each category is predictive of the target.

Count / frequency encoding

City
Madrid
Barcelona
Toledo
Valencia
Madrid
Valencia
Valencia

City
2
1
1
3
2
3
3

Label / integer encoding

Pros

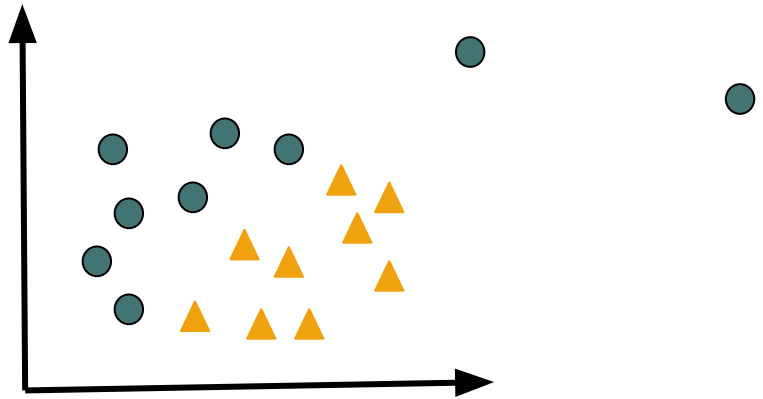
- Easy to implement
- Does not expand the feature space
- Can work well enough with tree based algorithms

Cons

- Not suitable for linear models
- Does not handle new categories in test set automatically
- If 2 different categories appear the same amount of times in the dataset, they will be replaced by the same number.

Outliers

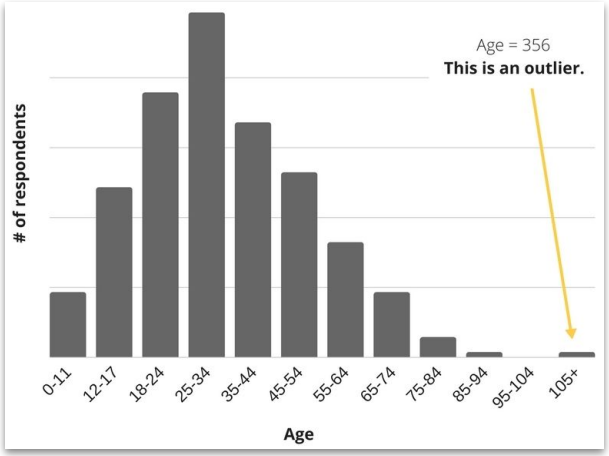
- An outlier is a data point which is significantly different from the remaining data.

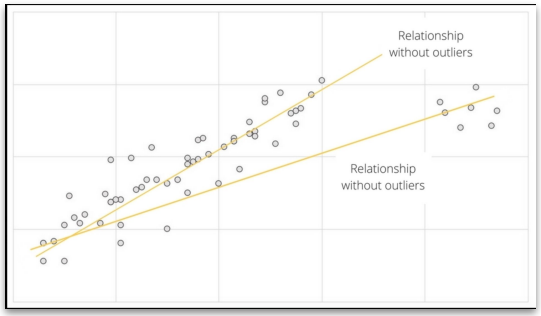


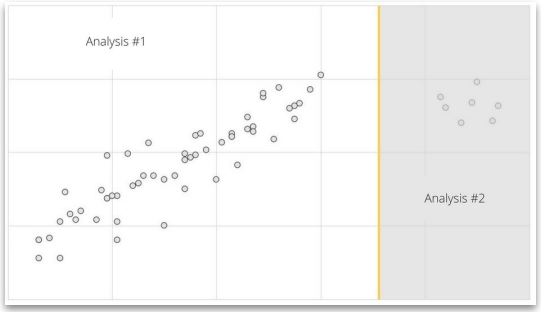
- “An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism.” [D. Hawkins. Identification of Outliers, 1980]

Should outliers be removed?

Depending on the context, outliers either deserve special attention or should be completely ignored







Handle outliers in cases where they may affect model performance

Automatic outlier detection

- **Out of scope: outlier detection**
- **References:**

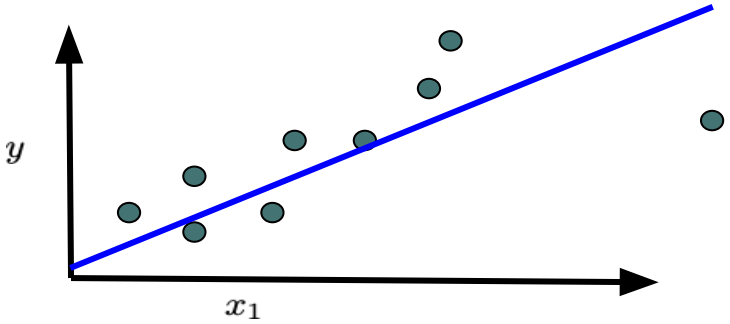
- 3 methods to deal with outliers [1]
- A brief overview of outlier detection techniques [2]
- How to identify outlier in your Data [3]
- Outliers detection techniques [4]

A massive fields with lots of techniques.

Algorithms susceptible to outliers

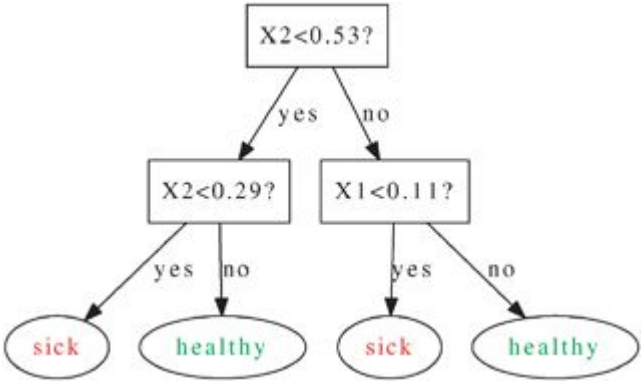
Bad

- Linear models (Linear regression, Logistic regressions)
- Adaboost



Good

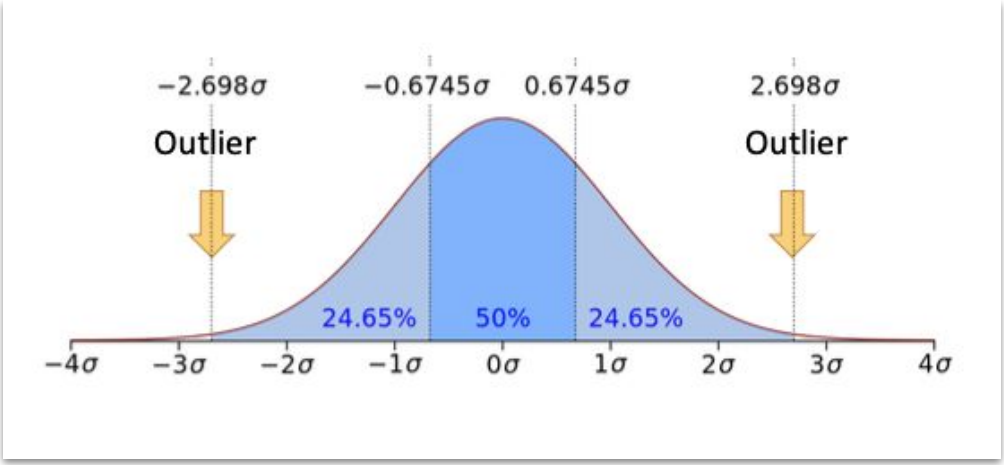
- Tree based models: Random forest, XGboost, etc



Normal distributions

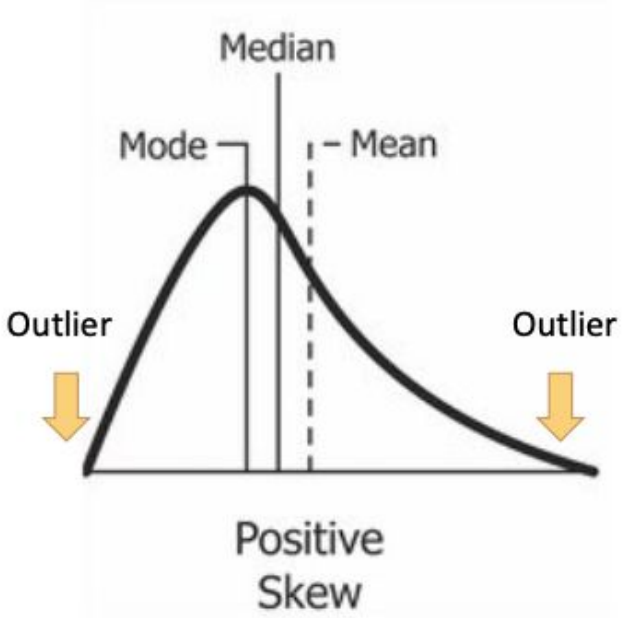
- 99% of the observation of a normally distributed variable lie within the mean and $\pm 3\sigma$

- Therefore, values outside this boundaries are considered outliers



Skewed distributions

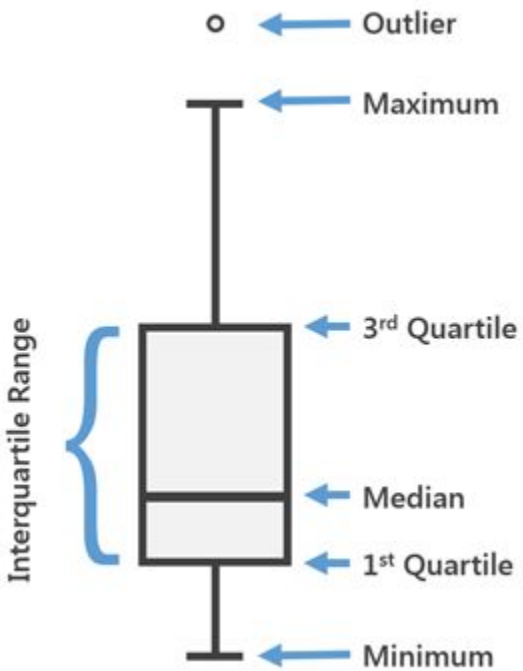
- For skewed distributions the approach is to calculate the quantiles, and then the inter-quantile range (IQR):
- IQR = 75^{th} Quantile - 25^{th} Quantile
- Upper limit = 75^{th} Quantile + IQR x 1.5
- Lower limit = 25^{th} Quantile - IQR x 1.5

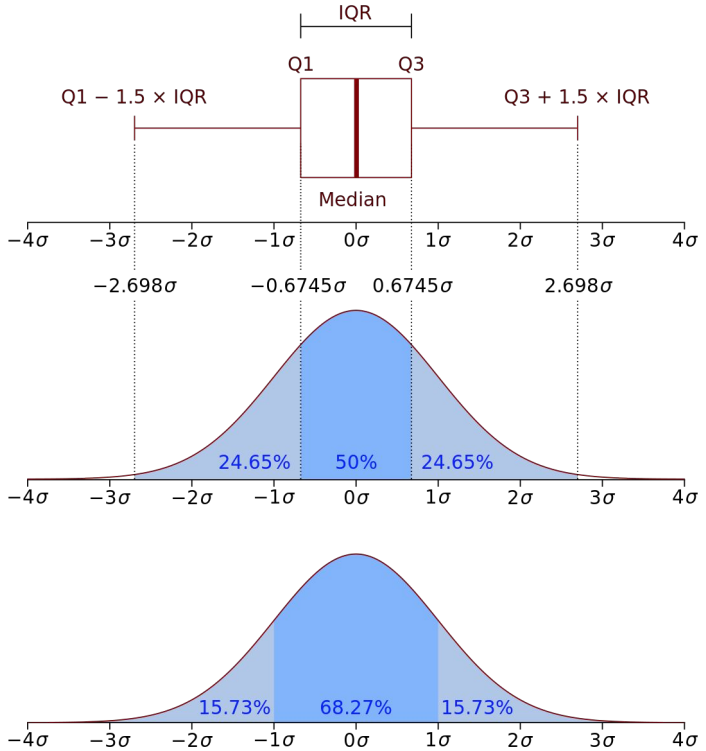


Notes on quantiles

- Quartiles = dividing the distribution in 4 parts
- Quantiles = dividing the distribution into 100 parts
- 1^{st} Quartile = 25^{th} Quantile
- 3^{rd} Quartile = 75^{th} Quantile
- 2^{nd} Quartile = 50^{th} Quantile = Median
- IQR = 75^{th} Quantile - 25^{th} Quantile = 3^{rd} Quartile - 1^{st} Quartile

Visualising outliers - Boxplots






IMMUNE
CODING INSTITUTE

References

- https://towardsdatascience.com/8-clutch-ways-to-impute-missing-data-690481c6cb2b
- https://towardsdatascience.com/feature-engineering-for-machine-learning-3a5e293a5114
- https://towardsdatascience.com/understanding-feature-engineering-part-1-continuous-numeric-data-da4e47099a7b
- https://towardsdatascience.com/understanding-feature-engineering-part-2-categorical-data-f54324193e63
- https://sebastianraschka.com/Articles/2014_about_feature_scaling.html
- https://towardsdatascience.com/ways-to-detect-and-remove-the-outliers-404d16608dba
- https://dl.acm.org/doi/book/10.5555/3239815