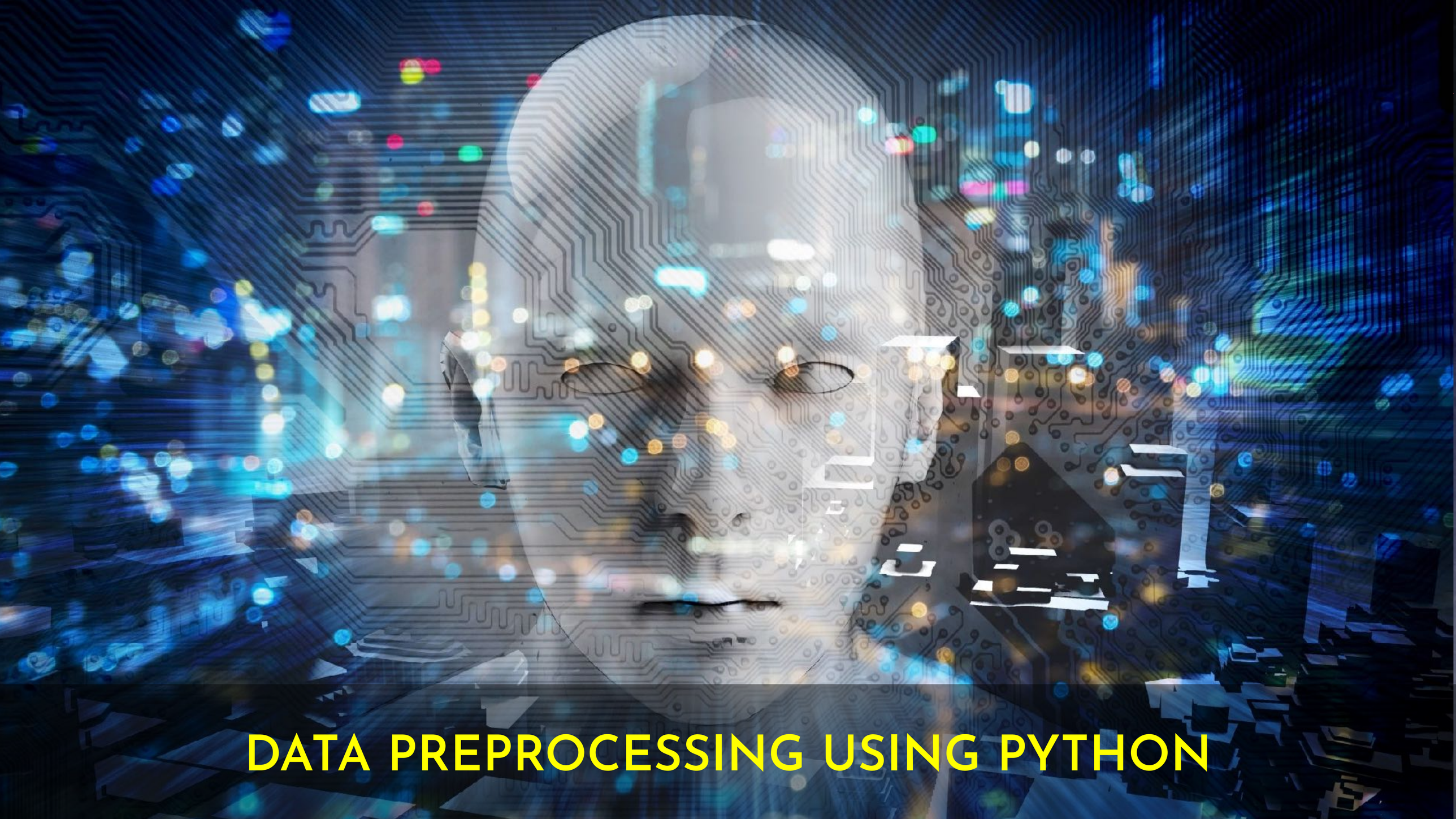
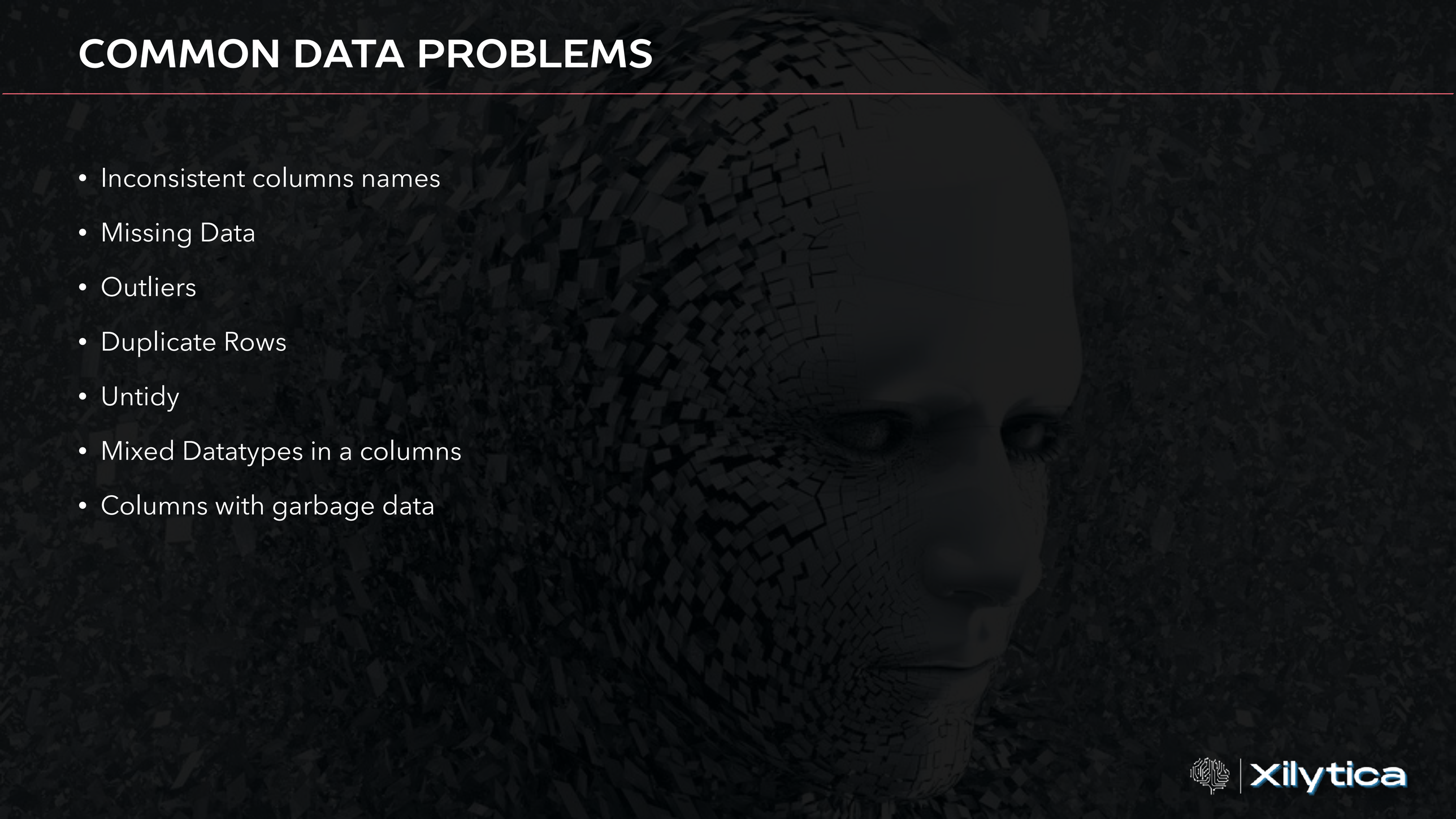

DATA PREPROCESSING USING PYTHON

COMMON DATA PROBLEMS

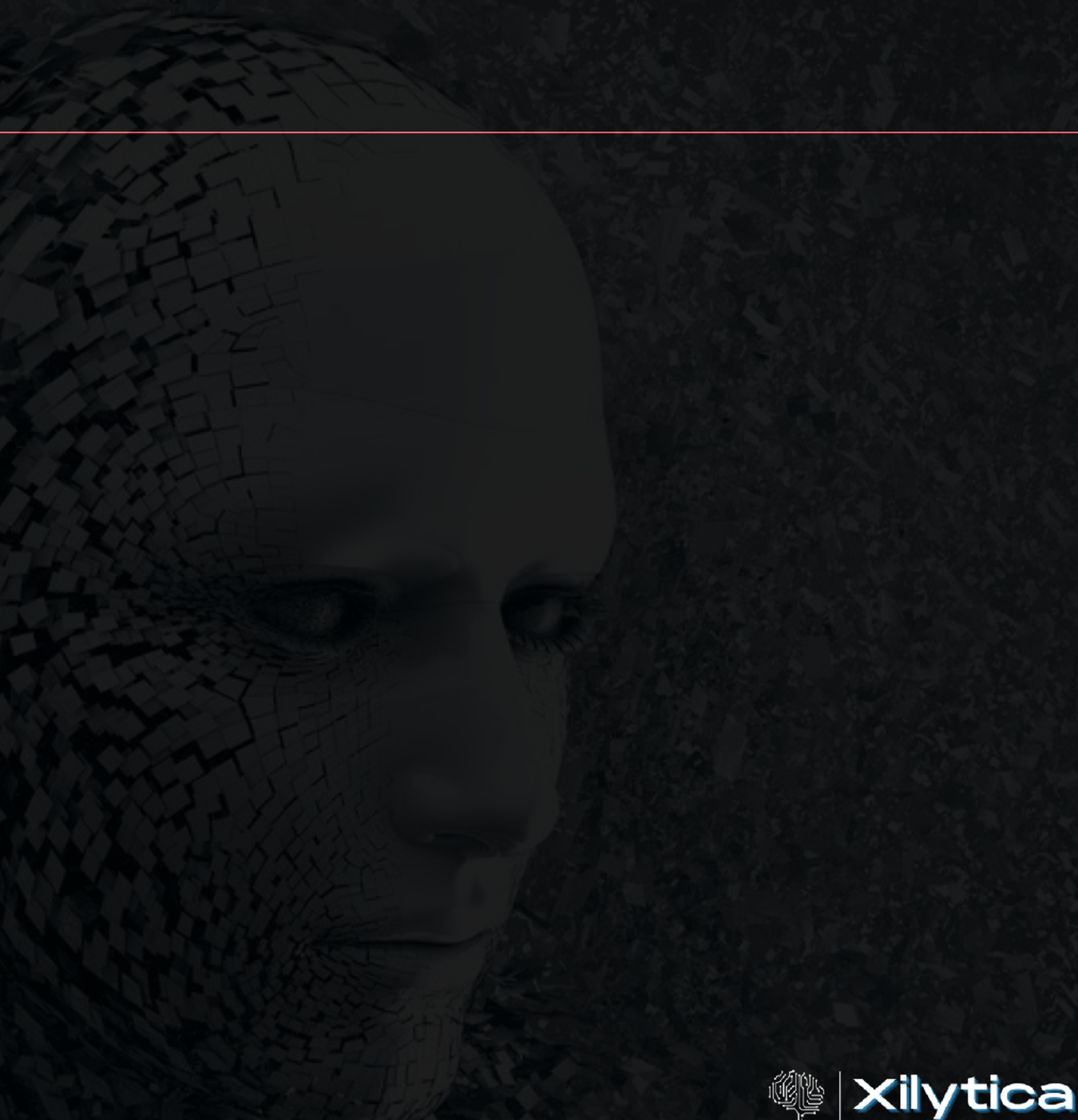
- Inconsistent columns names
- Missing Data
- Outliers
- Duplicate Rows
- Untidy
- Mixed Datatypes in a columns
- Columns with garbage data



Xilytica

DATA PREPROCESSING

- Data Cleansing
- Missing Data
- Outliers
- Train Test Split
- Scaling Data
- Hot Encoding for Categorical Data



Xilytica

Missing Values Imputation

Imputation - Replacing missing data with statistical estimates of missing values. The goal of any imputation technique is to produce a complete dataset that can be used to train machine learning models

Numerical Variables	Categorical Variables	Both Numerical & Categorical Variables
Mean/ Median Imputation	Frequency Category Imputation	Complete Case Analysis
Arbitrary Imputation	Adding a 'missing' Category	Adding a 'missing' Indicator
End of Tail Imputation		Random Sample Imputation

Complete Case Analysis

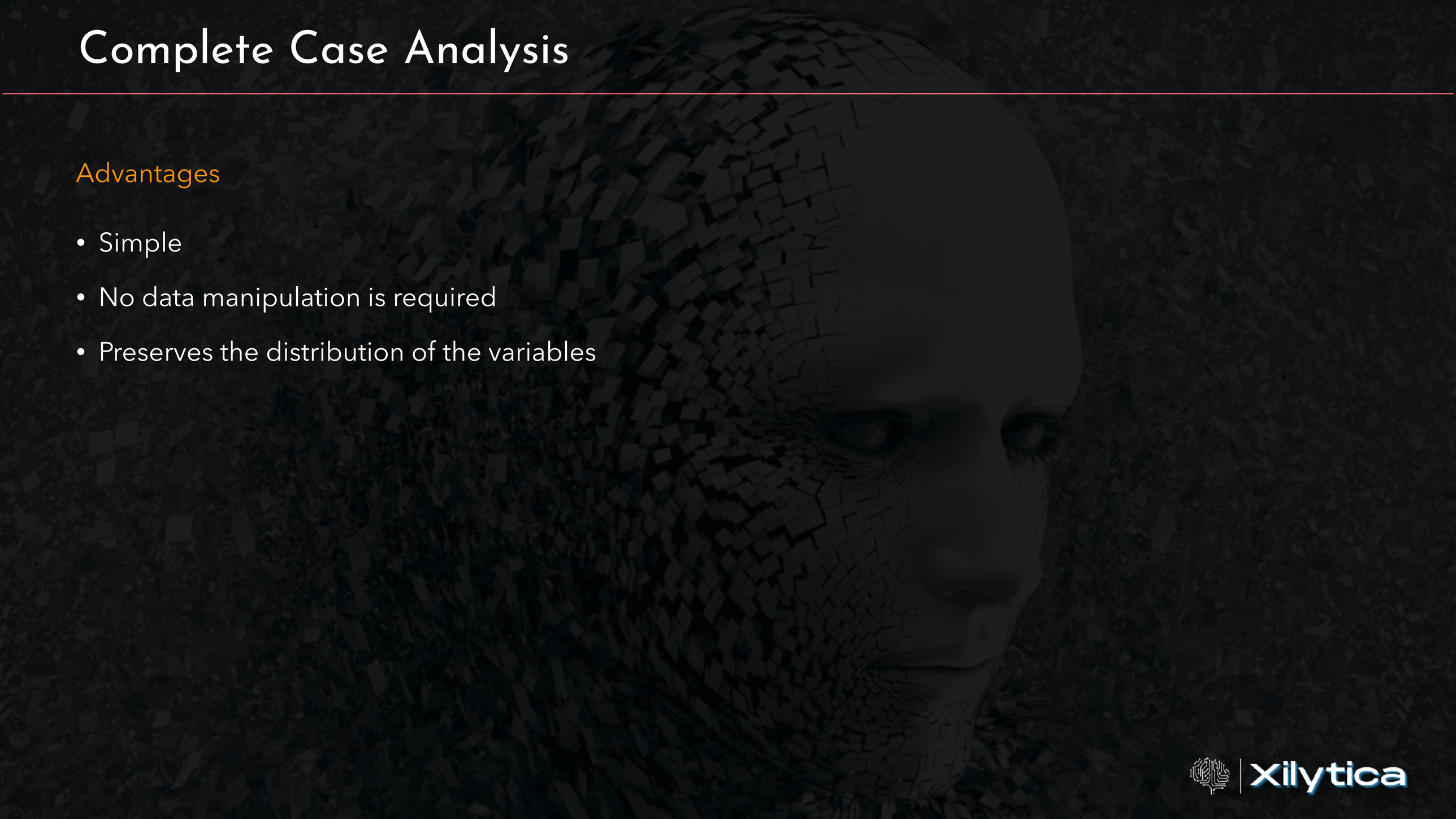
Complete Case - When we have data in all the columns and rows

- Complete Case Analysis (CCA) also called list-wise deletion of cases - consists of discarding observations where values in any of the variables are missing.
- We analyse only those observations for which there is information in all of the columns in the dataset
- Suitable for both numerical & categorical variables
- We assume data is missing completely at random
- No more than 5% of the total dataset contains missing data

Complete Case Analysis

Advantages

- Simple
- No data manipulation is required
- Preserves the distribution of the variables



Complete Case Analysis

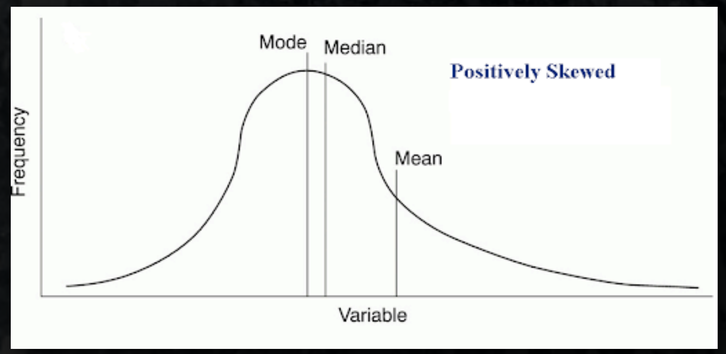
Limitations

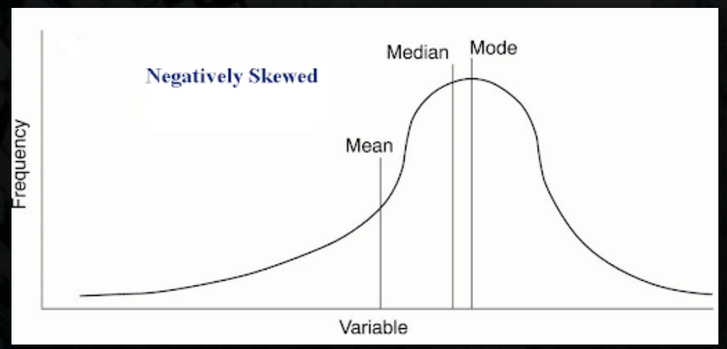
- It can exclude large fraction of the original dataset if missing data is in large number
- Excluded information could be informative for analysis if data is not missing at random
- CCA will create a biased dataset if the complete cases differ from the original data (https://stefvanbuuren.name/fimd/sec-MCAR.html)
- When using the model in production, model will not know how to handle missing data as we have not introduced any imputation technique

Mean / Median Imputation

Mean / Median Imputation - Replacing all the missing values, NA within a variable with the mean or median.

- Valid for for numerical columns
- If variable is normally distributed then the mean and median are approximately the same
- If the variable is skewed then the median is better representation as mean is influenced by the far end of the values





Xilytica

Mean / Median Imputation

Assumptions

- Data is missing at random
- The missing observations, most likely look like the majority of the observations in the variable
- No more than 5% of the variables contains missing data
- If data is missing completely as random, this would be captured by the mean / median, and if data is not missing by random then should be captured by additional missing indicator variable. Used in most of the data science competitions

Advantages

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

Mean / Median Imputation

Limitations

- Distortion of the original variable distribution
- Distortion of the original variance
- Distortion of covariance with the remaining variables of the dataset
- The higher the percentage of NA, the higher the distortions

Remember

- The mean / median value should be calculated only in the train set & used to replace NA in both train & test set. This is done to avoid over fitting

Arbitrary Value Imputation

- Arbitrary value imputation consists of replacing all occurrences of missing values (NA) within a variable by an arbitrary value decided by the user
- Typically used arbitrary values are 0, 999, -999 (or other combinations of 9s) or -1(if the distribution is positive)
- For categorical variables we use typically - Missing
- Works very well with trees but not so well with linear models

Arbitrary Value Imputation

Assumptions

Data is not missing at random. That means for values that are not missing at random, the arbitrary value should not be like the majority of values. We want to flag the missing values with a different (arbitrary)value, instead of replacing those occurrences with the mean or median, which represent the most common value

Arbitrary Value Imputation

Advantages

- Easy to implement
- Fast way of obtaining complete datasets
- Can be integrated in production (During model deployment)
- Captures the importance of being ‘missing’ if there is one

Arbitrary Value Imputation

Limitations

- Distortion of the original variable distribution
- Distortion of the original variance
- Distortion of the covariance with the remaining variables of the dataset
- If the arbitrary value is at the end of the distribution it may mask or create outliers
- Need to be careful not to choose the arbitrary value to similar to the mean or median or any other common value of the variable distribution
- The higher the percentage of NA, the higher the distortions

Complete Case Analysis

Limitations

- It can exclude large fraction of the original dataset if missing data is in large number
- Excluded information could be informative for analysis if data is not missing at random
- CCA will create a biased dataset if the complete cases differ from the original data (https://stefvanbuuren.name/fimd/sec-MCAR.html)
- When using the model in production, model will not know how to handle missing data as we have not introduced any imputation technique

Outliers

- An outlier is a datapoint which is significantly different from the remaining data. Population of India & China compared to population of other countries of same size.

- May affect performance of certain algorithms such as regression & ada boost

Outliers

Handling Outliers in a dataset

Trimming

Removing outliers from the dataset

Missing Data

Treat outliers as missing data & perform missing data imputation

Discretisation

Put outliers in the lower or upper bins

Censoring

Capping or limiting the minimum or the maximum value a variable can take. Outliers will be replaced by minimum or the maximum values that variable can take

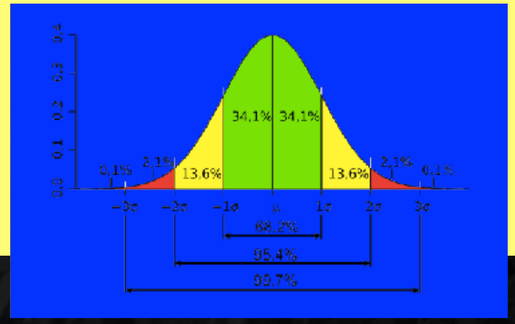
Outliers

Detecting Outliers

For Extreme values multiply IQR by 3 instead of by 1.5

Gaussian Distribution for Normal Distribution

99% of values lies with in 3 standard deviations around the mean. Any value that lies outside this are considered outliers



Inter-quantal Range Proximity Rule for Skewed Distribution

Values above - 75th Quantile + 1.5* IQR

Values below - 25th Quantile - 1.5*IQR

Quantiles

Values below 5th quantile & values above 95th quantile are considered outliers

Xilytica

Feature Scaling

- Feature Scaling - Method used to normalise the range of values of independent variables
- Set the feature value range within a similar scale
- Magnitude of the features we need to consider for training our machine learning models
- The regression coefficients are directly influenced by the scale of the variables
- Variables with bigger value range dominate over the ones the with smaller value range
- Gradient descent converges faster when features are on similar scales

Feature Scaling

Models Sensitive to Magnitude

- Linear & Logistic Regression
- Neural Network
- Support Vector Machine
- KNN
- K - Means Clustering
- Linear Discrimination Analysis
- Principal Component Analysis

Models Insensitive to Magnitude - Tree Based

- Classification & Regression Trees
- Random Forest
- Gradient Boost Trees