GROKKERS
AI FOR EVERYONE

# Data Scaling

# Pre-reqs

- Python
- NumPy and PANDAS, SciPy, Visualizations
- Elementary stats and maths
- Some preprocessing steps – may need ML as well, for advanced topics

# Background

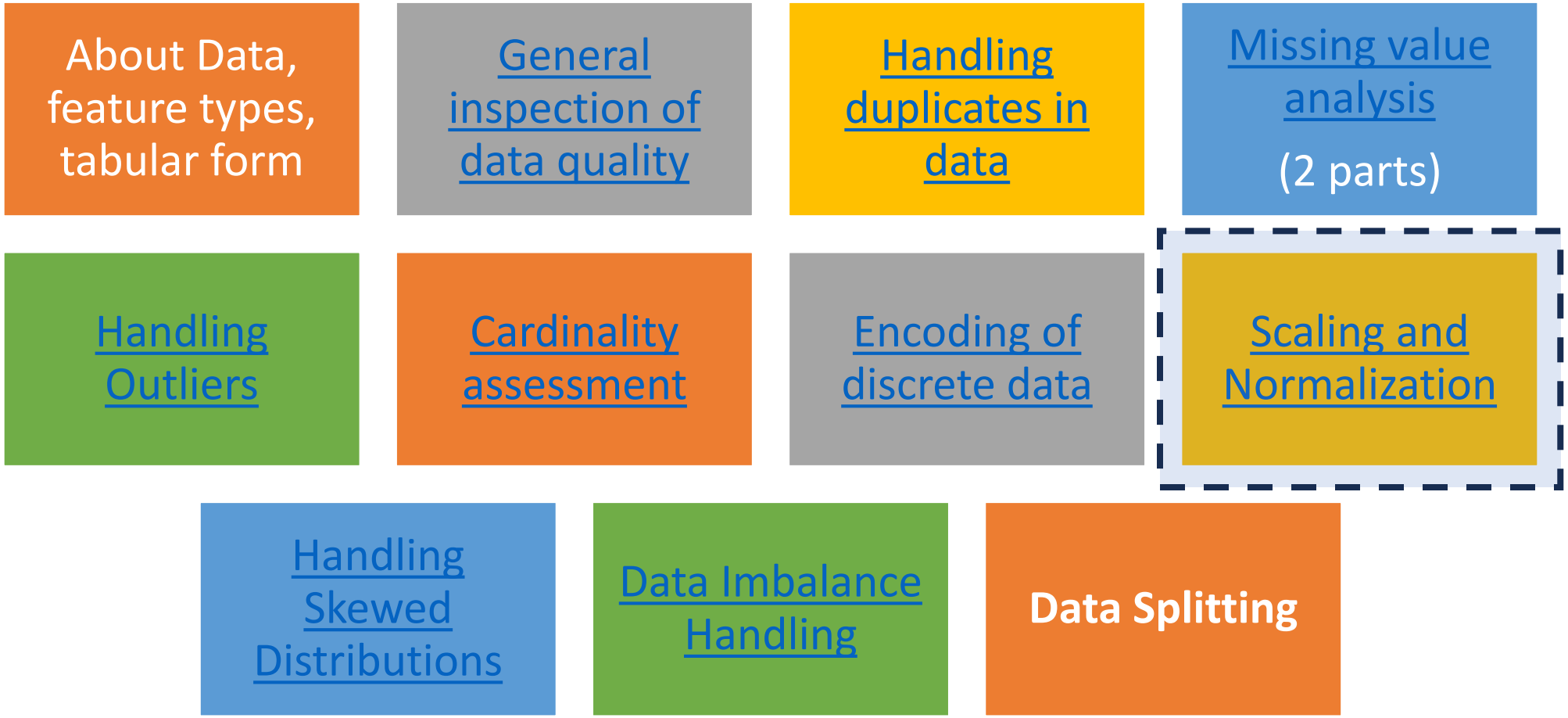
**Numeric Data:** Preprocessing involves handling missing values, scaling to a similar range, and possibly normalizing the distribution.

Text Data**:** Common preprocessing steps include text cleaning (removing stop words, punctuation, etc.), tokenization, and vectorization (converting text into numerical form, such as TF-IDF or word embeddings).

**Image Data:** Techniques like resizing, normalization of pixel values, and data augmentation (creating variations of existing images) are often used.

**Time Series Data**: Dealing with temporal aspects, handling missing values over time, and creating lag features are important steps in preprocessing time series data.

# Topics



# Scaling

## Definition

- Scaling involves transforming the values of a variable to a specific range, making it easier to compare and interpret.
- standardize the range of independent variables or features of the data.

## Methods:

- **Min-Max Scaling:** Scales the values between 0 and 1.
- **Standardization (Z-score Scaling):** Scales the values to have a mean of 0 and a standard deviation of 1.
- **Robust Scaling:** Scales the values based on the median and interquartile range.

# EXAMPLE

- most machine learning algorithms take into account only the magnitude of the measurements, not the units of those measurements.

- one feature, expressed in a very high magnitude (number), may affect the prediction a lot more than an equally important feature.

- **Example**
- 2 lengths, L1 = 250 cm and L2 = 2.5 m.
- humans see these 2 are identical lengths (l1 = l2), but most ML algorithms interpret this differently.
- more weight to L1, just because it is expressed in a larger number , which, in turn is going to have a much larger impact on the prediction than l2.

# DO WE ALWAYS NEED TO SCALE?

- Certain types like Naive Bayes, Decision Trees, RF and XGB do not require feature scaling

- algorithms that exploit distances or similarities (e.g. in form of scalar product) between data samples, such as k-NN and SVM, often require feature scaling.

# Min-Max Scaling

- Min-Max Scaling, also known as Min-Max Normalization, is a technique used to transform the values of a numerical variable to a specific range, typically between 0 and 1.

$$X_{\text{scaled}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}}$$

  - $X$ is the original value of the variable.
  - $X_{\text{min}}$ is the minimum value of the variable in the dataset.
  - $X_{\text{max}}$ is the maximum value of the variable in the dataset.

# Example

$$X_{\text{scaled}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

Consider a dataset with the following values for a variable:

- [2,5,8,10,4]

$X_{min}$ = 2 (minimum value)

$X_{max}$ =10 (maximum value)

Using Min-Max Scaling:

- $X_{scaled}$ = ($X$−2)/ (10−2)
- For $X$ = 5: $X_{scaled}$ = 5−2/10−2 = 3/8
- The scaled values would be between 0 and 1.

**Resulting Scaled Values:**

[0.375,0.625,0.875,1.0,0.5]

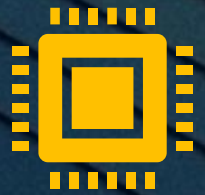
# Pros of Min-Max Scaling

**Preservation of Relationships**

Min-Max Scaling preserves the proportional relationships among the original values, ensuring that the order and relative distances between values are maintained.

**Simple Implementation**

Min-Max Scaling is straightforward and easy to implement



**Applicability to Algorithms**

useful for machine learning algorithms that are sensitive to the scale of features, such as support vector machines, k-nearest neighbors, and neural networks.

# Cons of Min-Max Scaling

**Sensitivity to Outliers:**

- Min-Max Scaling is sensitive to outliers in the data..

**Normalization to a Fixed Range:**

- fixed range of [0, 1] might not be suitable for certain datasets with specific characteristics.
- In some cases, the choice of a fixed range may not be optimal.

**Not Robust to Extreme Values:**

- Min-Max Scaling is not robust to extreme values, and the presence of outliers can affect the scaling, potentially leading to loss of information.

# Best Practices

17-08-2024

## Outlier Handling:

- If dataset contains outliers, consider applying outlier detection and handling techniques

## Normalization as a Choice:

- Evaluate whether normalizing to a fixed range is suitable for your specific problem

# Standard Scaler (Z-score Scaling)

The Standard Scaler, also known as Z-score Scaling or Z-score Normalization, is a technique used to transform the values of a numerical variable to have a mean of 0 and a standard deviation of 1

$$X_{\text{standardized}} = \frac{X - \text{mean}(X)}{\text{std}(X)}$$

- $X$ is the original value of the variable.
- mean($X$) is the mean (average) of the variable.
- std($X$) is the standard deviation of the variable.

# Examples

Consider a dataset with the following values for a variable:

[2,5,8,10,4]

$$X_{\text{standardized}} = \frac{X - \text{mean}(X)}{\text{std}(X)}$$

1. **Calculate Mean ($\text{mean}(X)$):**

$$\text{mean}(X) = \frac{2+5+8+10+4}{5} = \frac{29}{5} = 5.8$$

2. **Calculate Standard Deviation ($\text{std}(X)$):**

$$\text{std}(X) = \sqrt{\frac{(2-5.8)^2+(5-5.8)^2+(8-5.8)^2+(10-5.8)^2+(4-5.8)^2}{5}}$$

Calculating this gives $\text{std}(X) \approx 2.90593$.

3. **Apply Standard Scaler:**

For $X = 5$:

$$X_{\text{standardized}} = \frac{5-5.8}{2.90593} \approx -0.2758$$

Similarly, apply the formula for each value in the dataset.

Standardized Dataset ≈ [−0.9635, −0.2758, 0.4119, 0.9538, −0.1264]

# Pros of Standard Scaler

| | |
|---|---|
| **Mean and Standard Deviation Preservation** | Standard Scaler preserves the mean and standard deviation of the original data, making it suitable for data with a Gaussian distribution. |
| **Outlier Insensitivity** | Standardization is less sensitive to outliers compared to Min-Max Scaling, making it more robust in the presence of extreme values. |
| **Applicability to Algorithms** | Standard Scaler is widely applicable to machine learning algorithms that assume normally distributed data, such as linear regression and k-means clustering. |

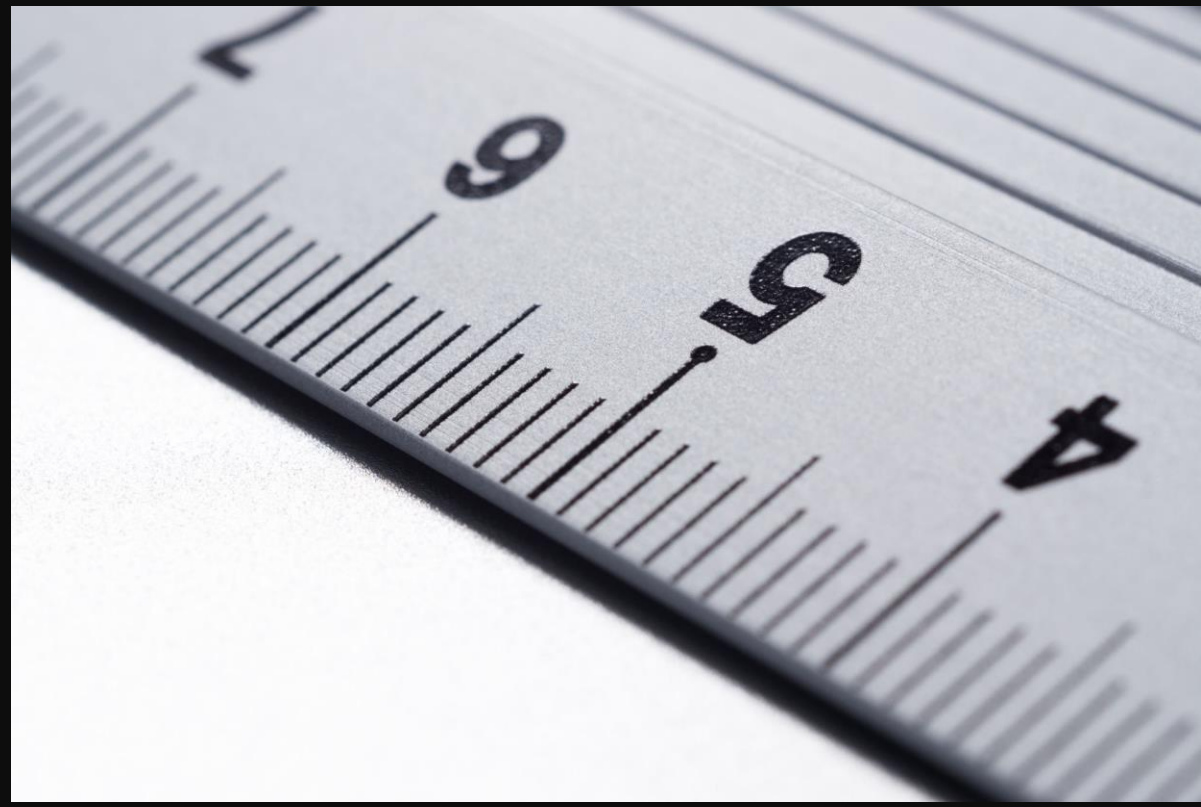
# Cons of Standard Scaler

- Standardization assumes that the data is approximately normally distributed.

- In cases where this assumption is violated, other scaling methods might be more appropriate.

# Robust scaler

- scale the values of a numerical variable in a way that is robust to the presence of outliers.
- useful when the data contains extreme values that can significantly impact the scaling process.
- scales the data by subtracting the median and dividing by the interquartile range (IQR).
- makes it less sensitive to outliers compared to other scaling methods.

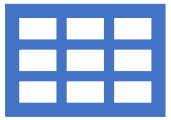


# Formula

$$X_{\text{robust-scaled}} = \frac{X - \text{median}(X)}{\text{IQR}(X)}$$

where:

- $X$ is the original value of the variable.
- median($X$) is the median of the variable.
- QR($X$) is the interquartile range of the variable.



# Example

- Consider a dataset with the following values for a variable:
  - [2,5,8,10,4,100]

- **Calculate Median (median(*X*))**
  - median(*X*) = 6
- **Calculate Interquartile Range (IQR(*X*))**
  - IQR(*X*) = *Q*3 − *Q*1
  - IQR(*X*) = 9 − 4 = 5
- For *X*=5:
  - $X_{\text{robust-scaled}}$ = (5−6) /5 = −0.2
- Similarly, apply the formula for each value in the dataset.

**Resulting Robust Scaled Values:**
= [−0.8,−0.2,0.4,0.8,0,19.6]

# Demo using python/sklearn

# (scaling data)

# Thanks !!