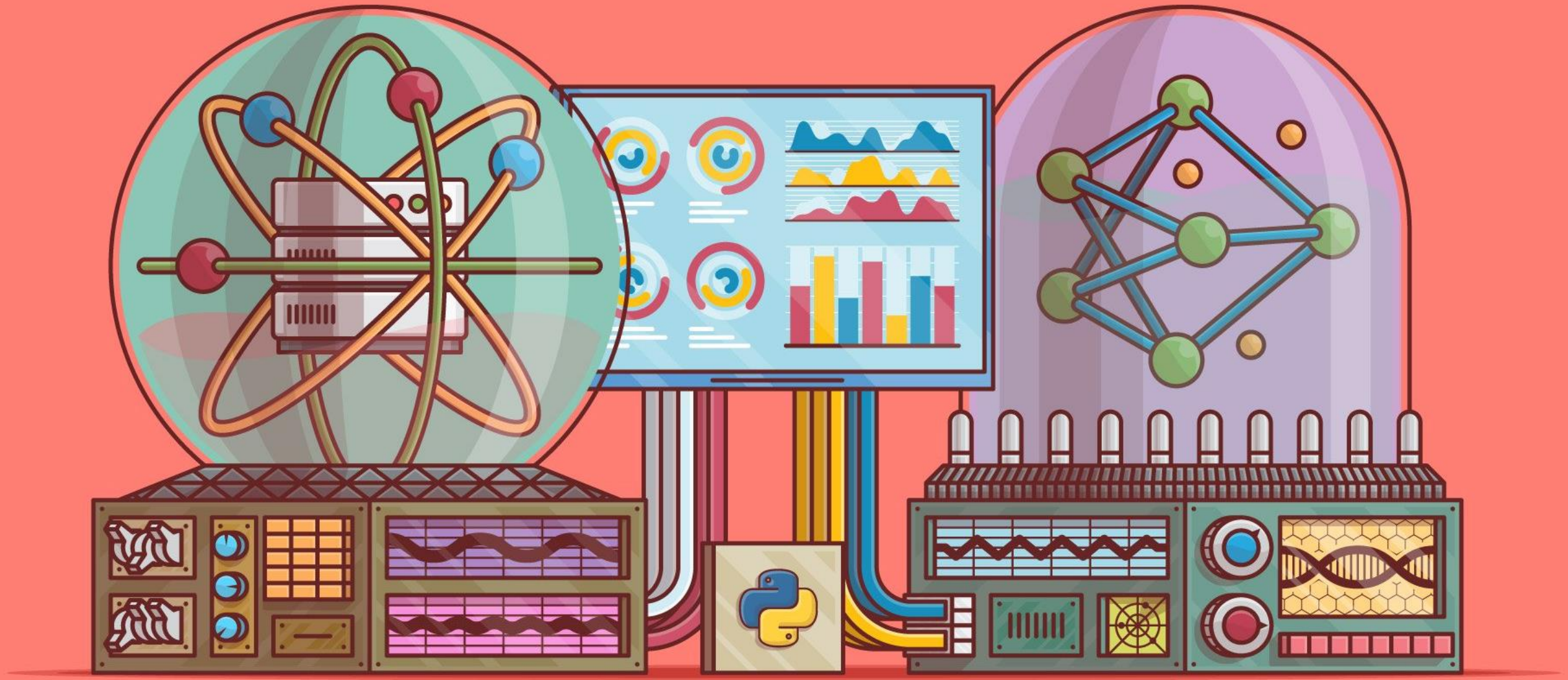
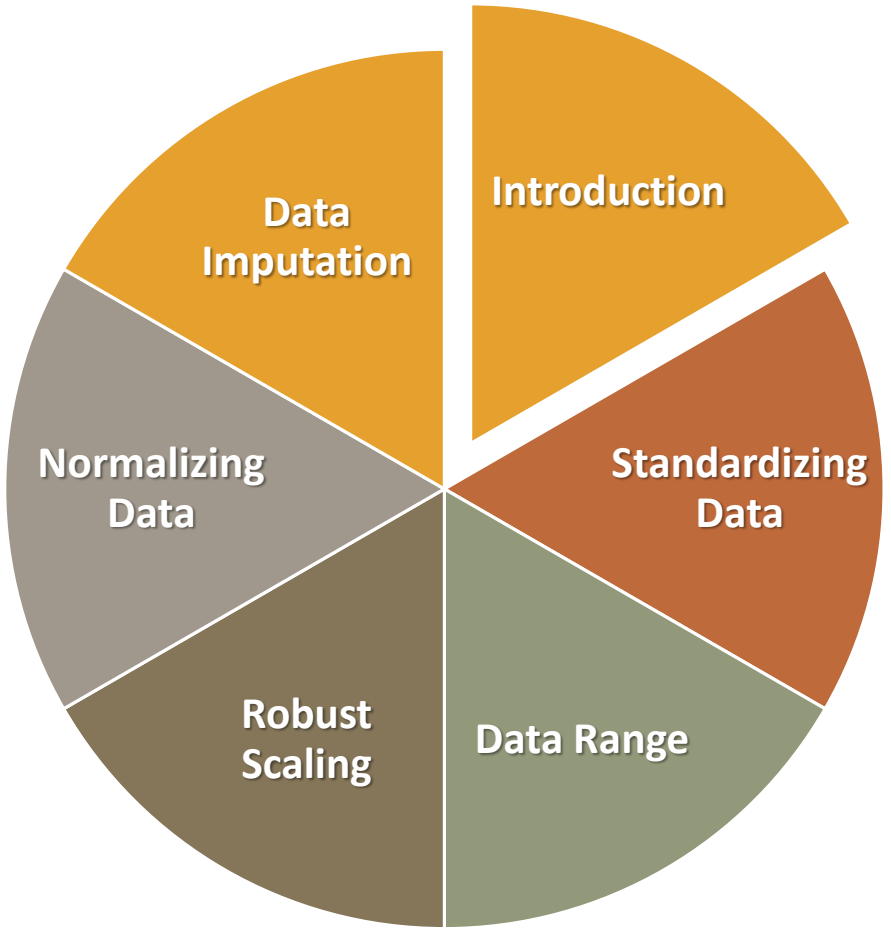


Data Preprocessing Using Scikit-Learn

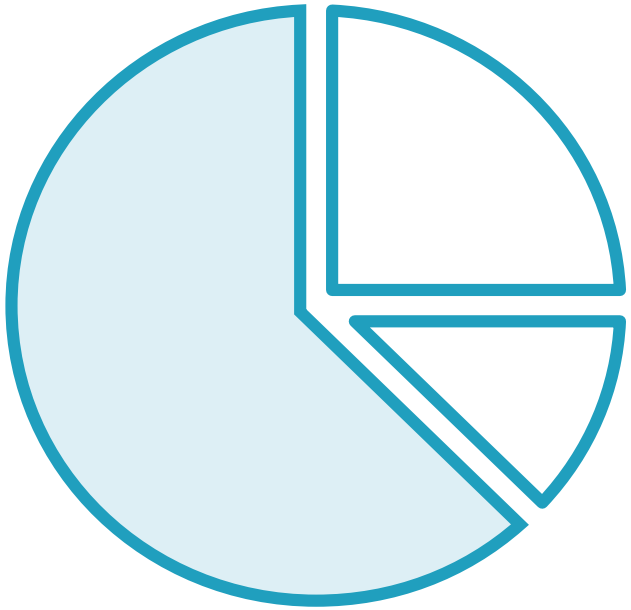
By Anil Kumar &
Team APSSDC

Data Preprocessing with Scikit-Learn

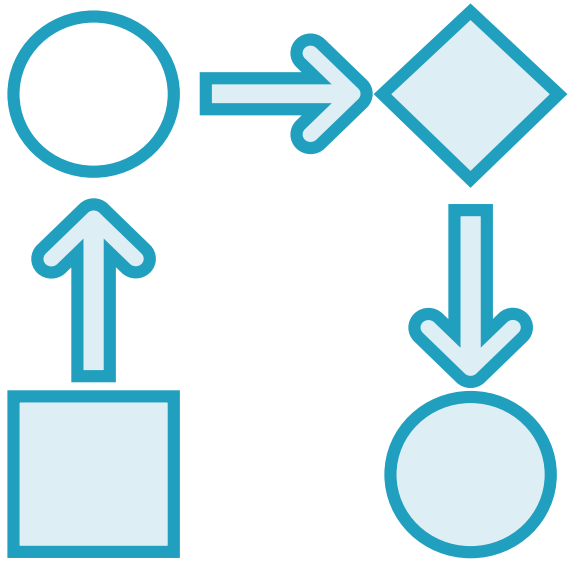


Two Hats of a Data Professional

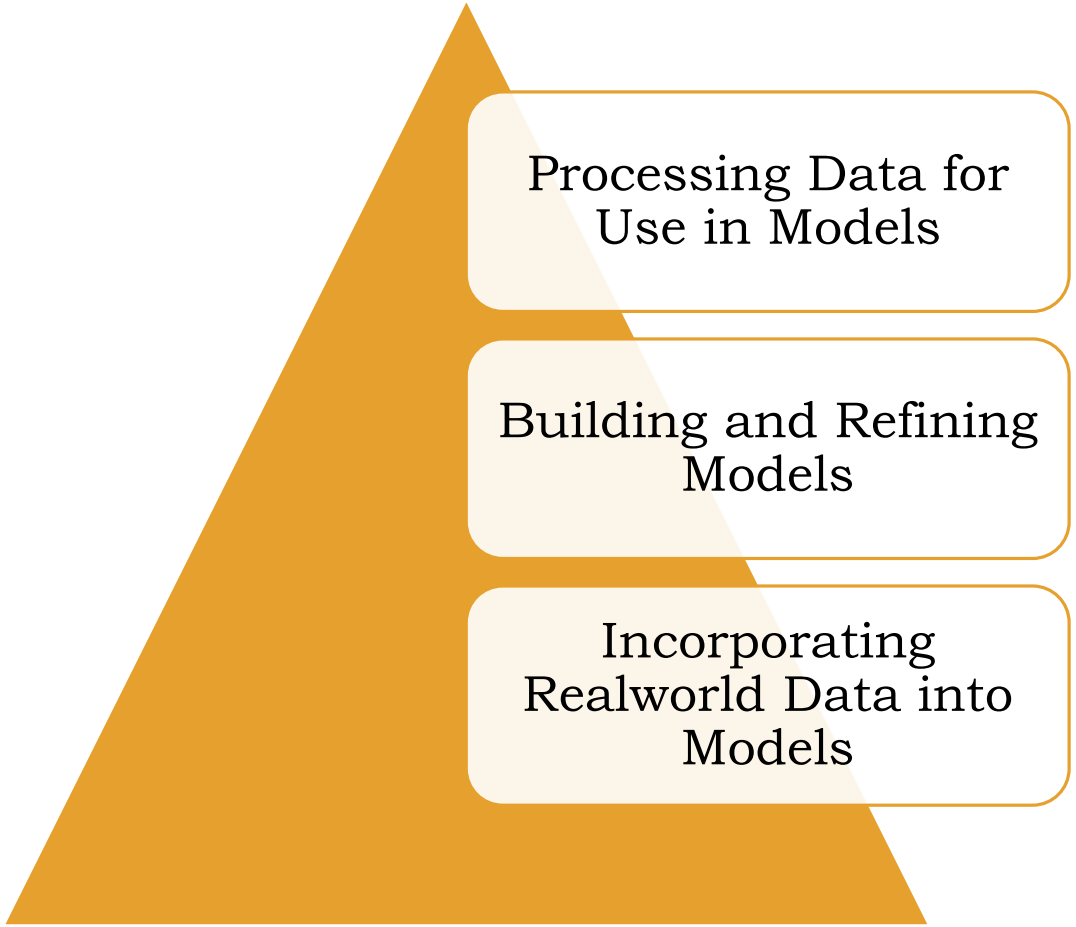
Find The Dots: Identify Important Elements In A Dataset



Connect The Dots: Explain Those Elements Via Relationships With Other Elements



Essential Steps in Connecting the Dots



Standardizing Data

$$\begin{bmatrix} X_{11} & & X_{1k} \\ X_{21} & \dots & X_{2k} \\ \dots & & \dots \\ X_{n1} & & X_{nk} \end{bmatrix}$$

$$\text{avg}(X_1) \quad \dots \quad avg(X_k)$$

$$\text{stdev}(X_1) \quad \dots stdev(X_k)$$

Standardizing Data

$$\begin{bmatrix} \frac{X_{11} - avg(X_1)}{stdev(X_1)} & \cdots & \frac{X_{1k} - avg(X_k)}{stdev(X_k)} \\ \cdots & \cdots & \cdots \\ \frac{X_{n1} - avg(X_n)}{stdev(X_n)} & & \frac{X_{nk} - avg(X_k)}{stdev(X_k)} \end{bmatrix}$$

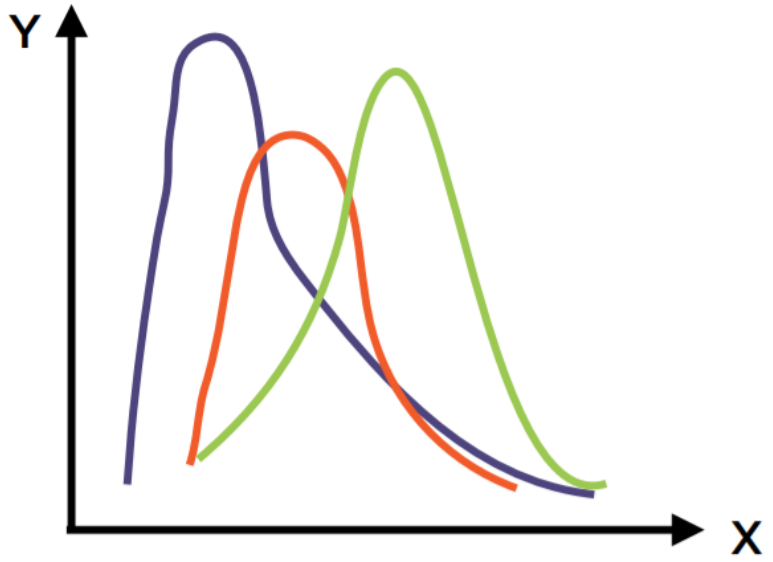
Each column of the standardized data has mean 0 and variance 1

Standardizing Data

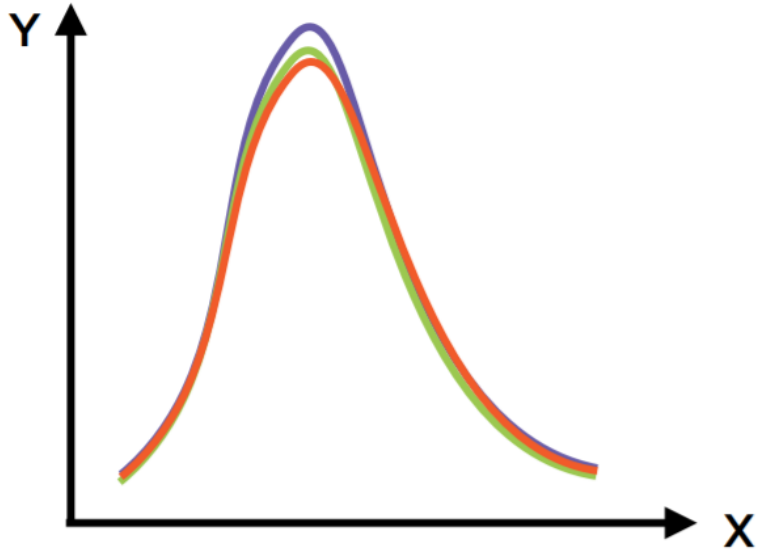
$$Z = \frac{X_i - mean(X)}{stdev(x)}$$

Standardization operates column-by-column and yields features with zero mean and unit variance

Standardizing Data



Before



After

Mean is a measure of central tendency and standard deviation is a measure of dispersion

Robust Standardization

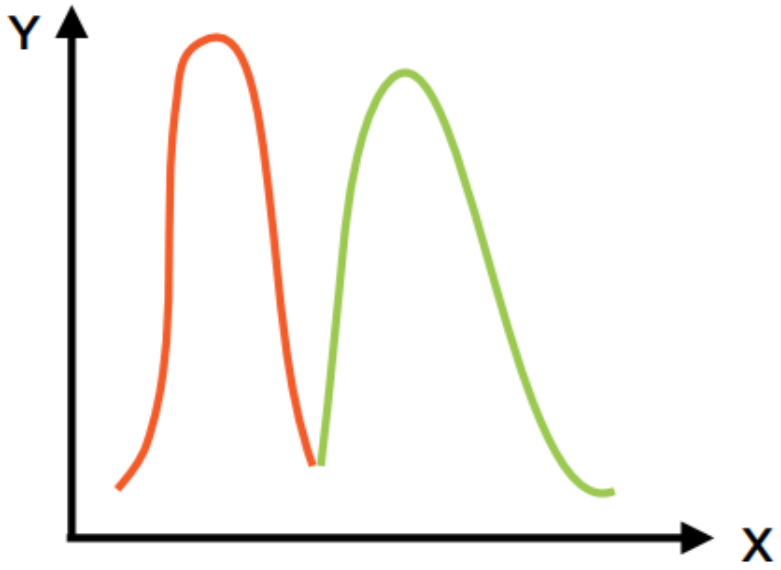
$$Z = \frac{X_i - median(X)}{stdev(x)}$$

Median is also a measure of central tendency and inter-quartile range is also measure of dispersion

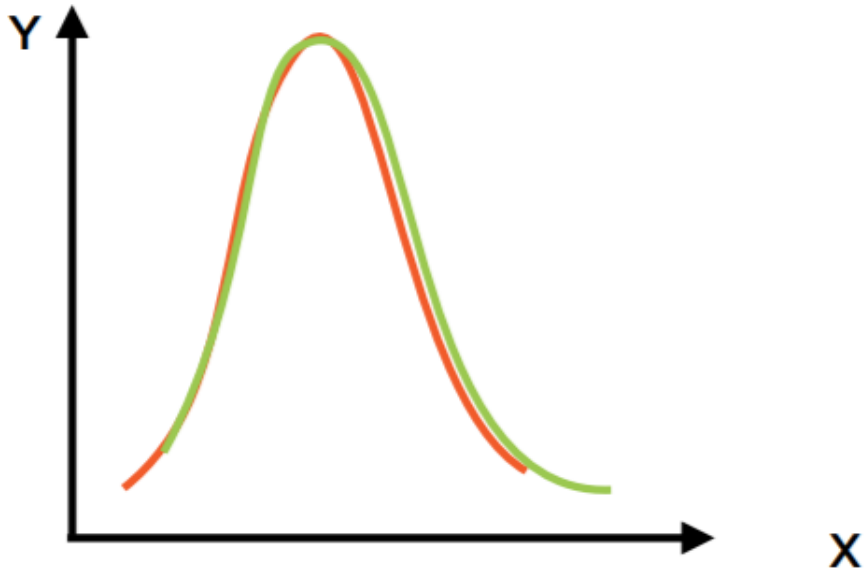
Output does not change much due to outliers

from sklearn.preprocessing **import** RobustScaler

Robust Standardization



Before



After

Data Range

we can also scale data by compressing it into a fixed range. One of the biggest use cases for this is compressing data into the range [0, 1].

$$x_P = \frac{x - d_{min}}{d_{max} - d_{min}}$$

from sklearn.preprocessing **import** MinMaxScaler

Normalization

Normalization Process of scaling input vectors individually to unit norm (unit magnitude), often in order to simplify cosine similarity calculations

from sklearn.preprocessing import Normalizer

$$X_{L2} = \left[\frac{x_1}{\ell}, \frac{x_2}{\ell}, \ldots, \frac{x_m}{\ell}\right], \text{where } \ell = \sqrt{\sum_{i=1}^{m} x_i^2}$$

Data Imputation

In real life, we often have to deal with data that contains missing values. Sometimes, if the dataset is missing too many values, we just don't use it.

There are many different methods for data imputation. In scikit-learn, the **SimpleImputer** transformer performs four different data imputation methods.

The four methods are:

1. Using the mean value
2. Using the median value
3. Using the most frequent value
4. Filling in missing values with a constant

```
from sklearn.impute import SimpleImputer
```