Handling Missing Values

1) Drop missing Values

2) Impute missing Value

Numerical	Categorical
1) Mean	1) Mode
2) Median (Outliers)	2) New Category
3) Mode	
4) Constant	

Time-series → BFill, F-Fill, Simple AVG, WA

5 10 (15) 20 1000 mean = 500
median = 15

* Python for Imputing :

1) Fillna( )

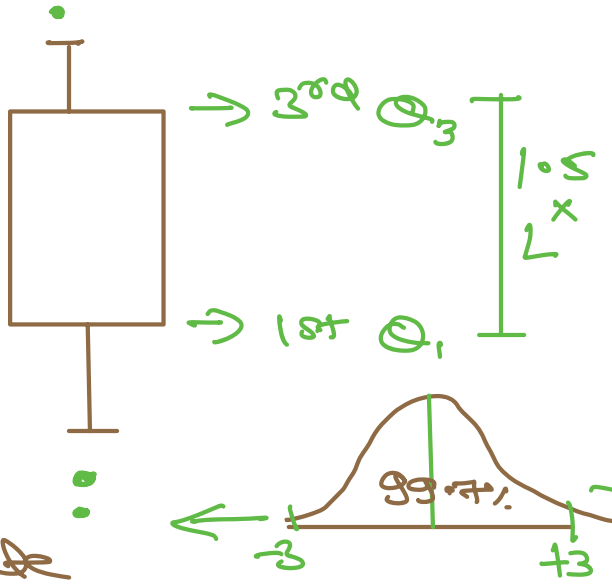
2) Simple Imputer

Outlier Treatment

Visualization

Uni-variate

1. Box Plot
2. Histogram



Bivariate / Multi-Variate

1. Scatterplot
2. DSSCAN
3. Kmeans

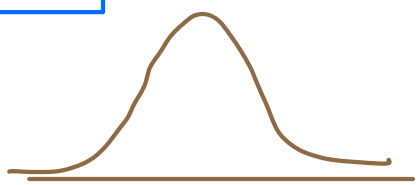
Methods to Find Outliers

1. IQR Based

$$IQR = Q_3 - Q_1$$

2. Z-score Based

$$\text{Z-score}_i = \frac{x_i - \bar{x}}{\sigma}$$



Categorical to Numerical

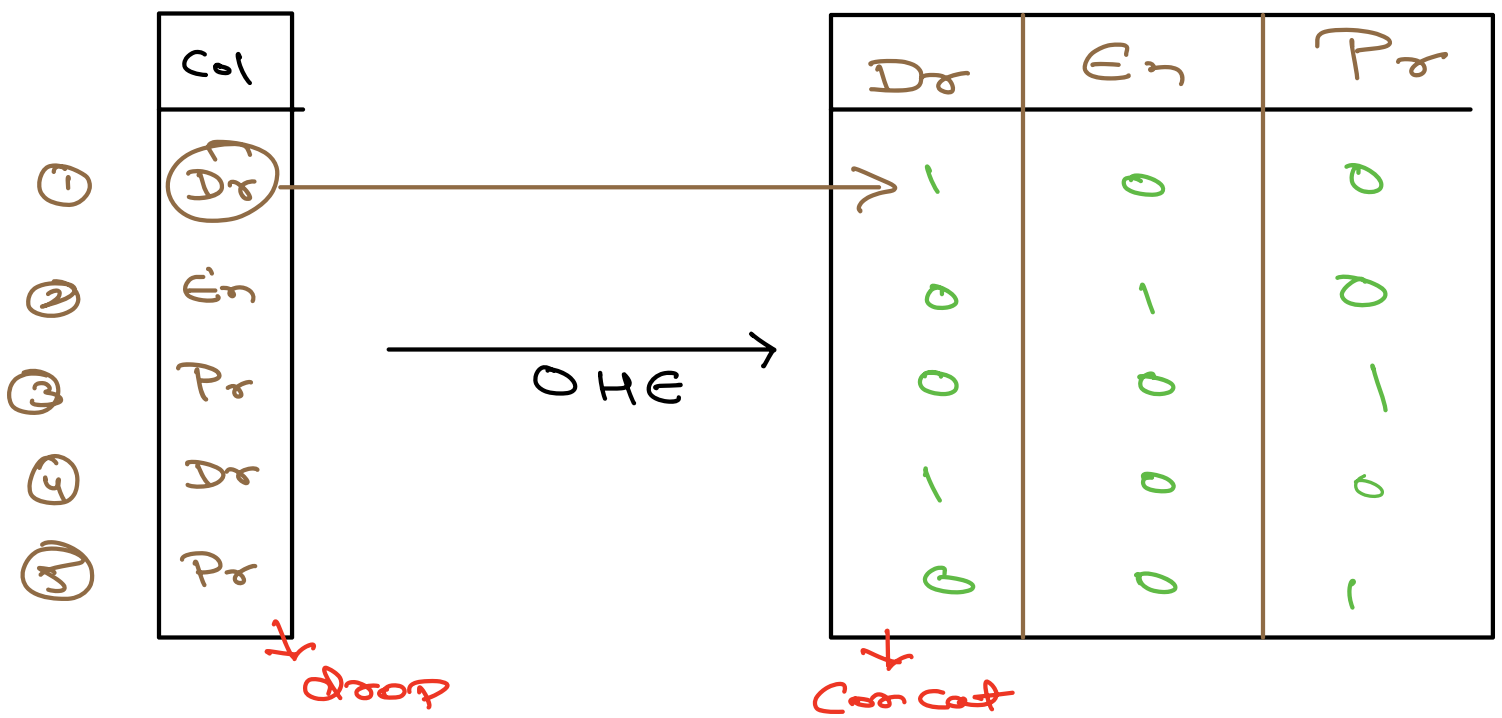
* ML Models process Data only in Numerical format.

Cat_str ⟶ Cat_num

Methods to Convert

1. One-Hot Encoding
2. Label Encoding
3. Target Encoding

One Hot Encoding



Issue: Many Sparse columns are added leading to "Curse of Dimensionality"

Label Encoder

Col
Dr
En
Pr
Dr
Pr

→ Label Encoder

Col
0
1
2
0
2

map ह Dr : 0
En : 1
Pr : 2 }

Issue : Nominal Categorical Data → Ordinal Categorical Data

Target Encoding

mean Encoding

Col
Dr
En
Pr
Dr
Pr

→ Target Encoder

New-Col
Avg (Dr-T)
Avg (En-T)
Avg (Pr-T)
Avg (Dr-T)
Avg (Pr-T)

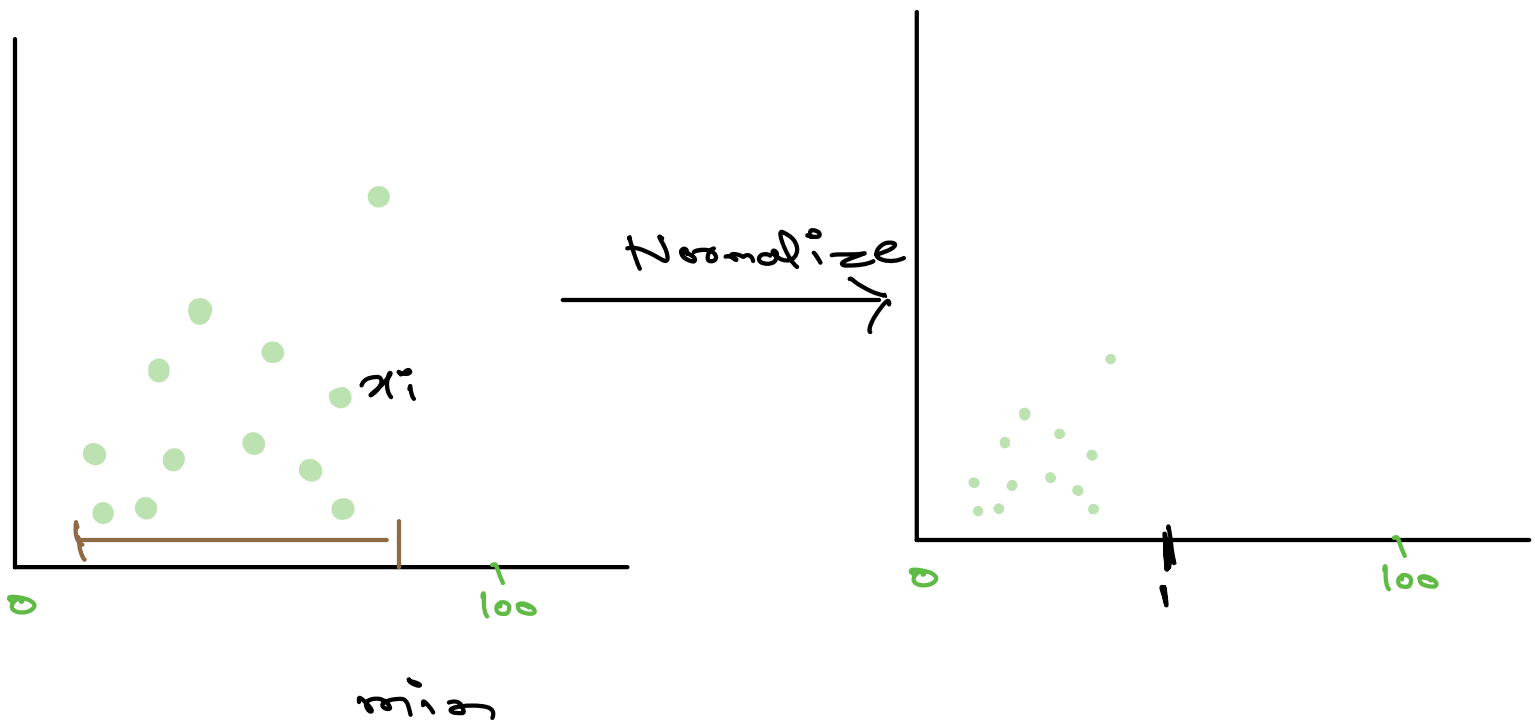
Scaling the Data

- Feature 1 → (0, 100)
- Feature 2 → (0, 10000)

↓

Bring all feature to similar Scale (Scaling)

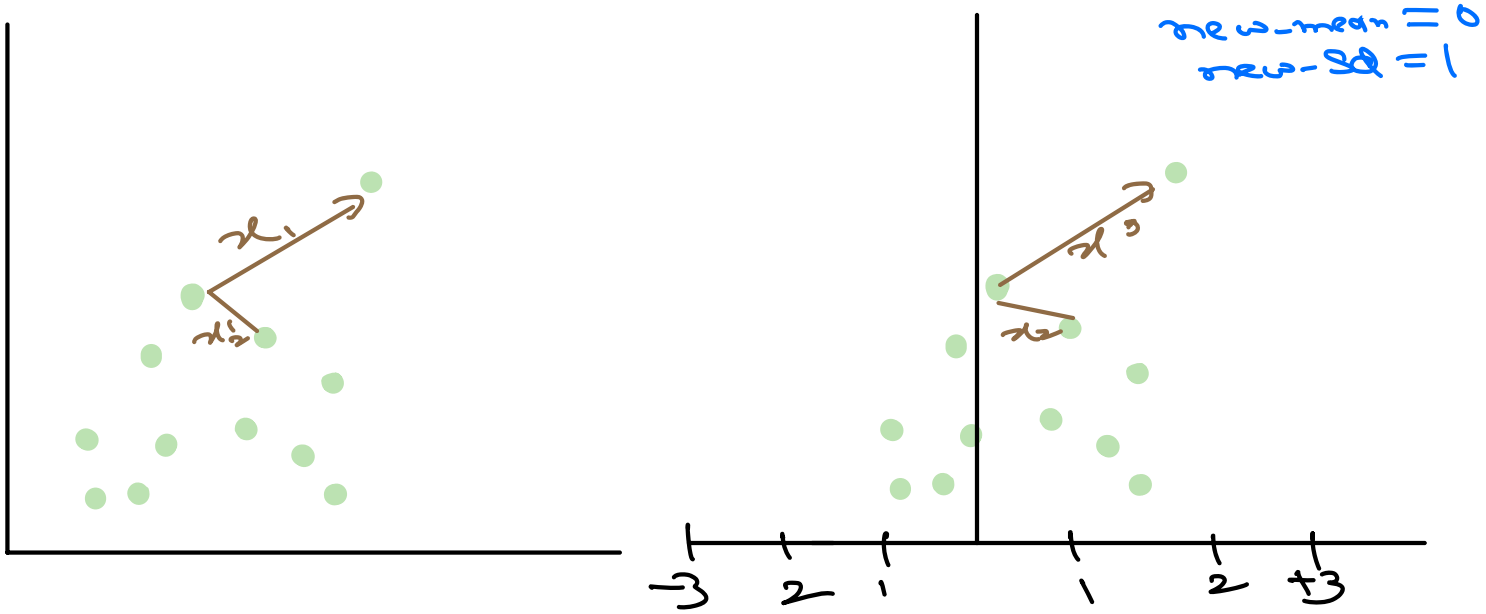
- Normalization → (0, 1)



$$x_{i\_new} = \frac{x_i - min}{max - min}$$

* Normalize is preferred Non-Gaussian Variable

② Standardization



$$x_{i\_new} = \frac{x_i - \bar{x}}{\text{Std}(x)}$$