eid	esal	location	purchased
1	NaN	Mumbai	0
2	2000	NaN	1
3	3000	Chennai	NaN
4	4000	Bangalore	0

Deal with Approximation

Numeric column!
Imputation
- mean
- median
- most freq

non-numeric column!
Two options
- most freq
* drop record

Label coln
- drop record.

ML Preprocessing Task

① Load the data
② Seperated the data as features and label. (numpy) (numpy)
2a missing data analysis and repair on non-numeric coln using Pandas.

③ Deal with missing Data (numeric column)
sklearn → Imputer (always accept numbers as input)

If a non-numeric column has a missing data and you are compelled to handle the same, recommended step is
a most frequent (mode)
use pandas ——— ②

④ Dealing with Categorical Data.
sklearn → LabelEncoder
OneHotEncoder → Dummy columns (Binary)

⑤ Handling Numeric Data.

All ML algo expect your (numeric) FEATURES to follow a common scale. To achieve this reqt. we use FEATURE SCALING.

① Rescale the feature in a limited range decided by you.

② Standardize the data such that mean is 0 and standard deviation is 1.

③ Normalize the data using

- ① L1 normalization
- ② L2 normalization

① Rescaling the feature.

Ideally all practitioners expect your data to be scaled at

(0, 1)

(-1, 1)

MinMax Scaler ⇒

$$x_i' = \frac{x_i - \min(x)}{\max(x) - \min(x)}$$

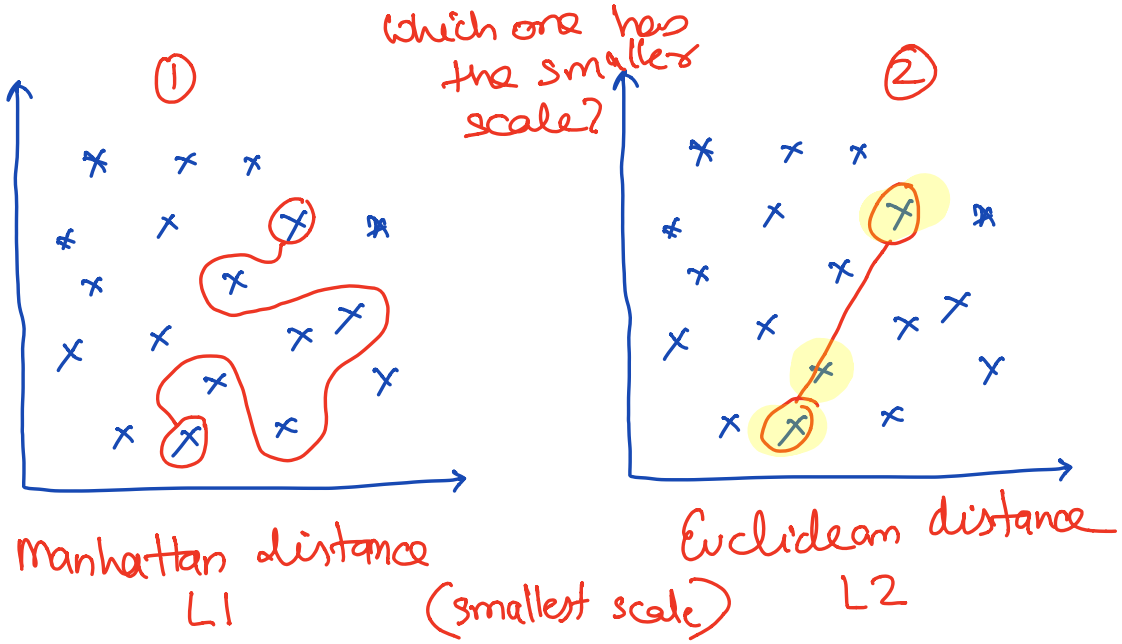
x → feature vector

x_i' → individual element calculated (result)

x_i → individual feature whose result to be calc.

Standardization:

$$x_i' = \frac{x_i - \bar{x}}{\sigma} \rightarrow \begin{cases} z \text{ score} \\ \text{statistics} \end{cases}$$



- your data weights is being maintained

- optimization is the key