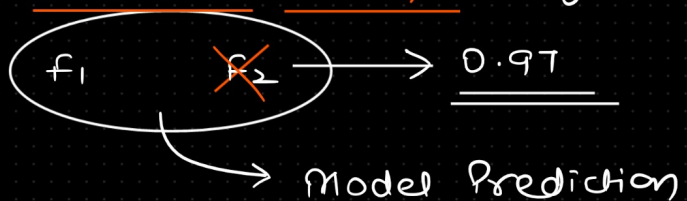
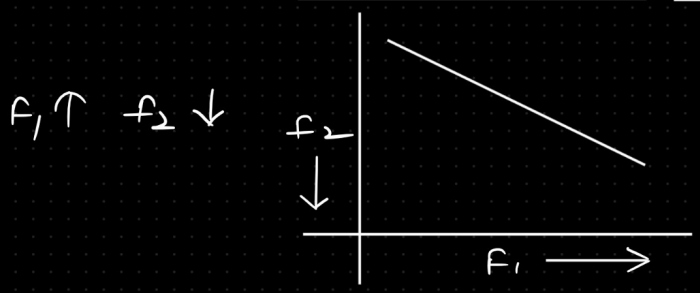
Correlation coefficient

↳ Range → -1 to +1

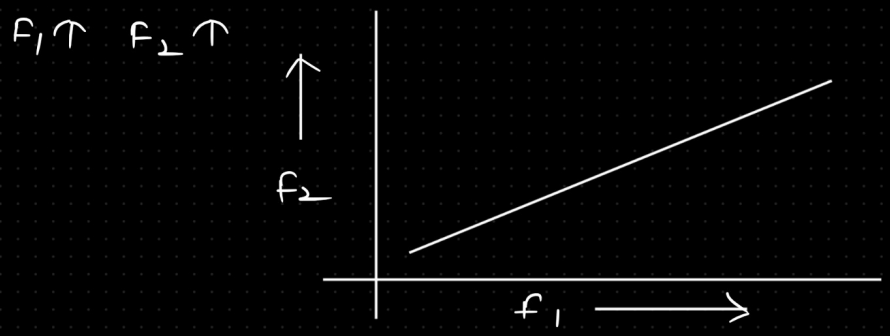
feature selection

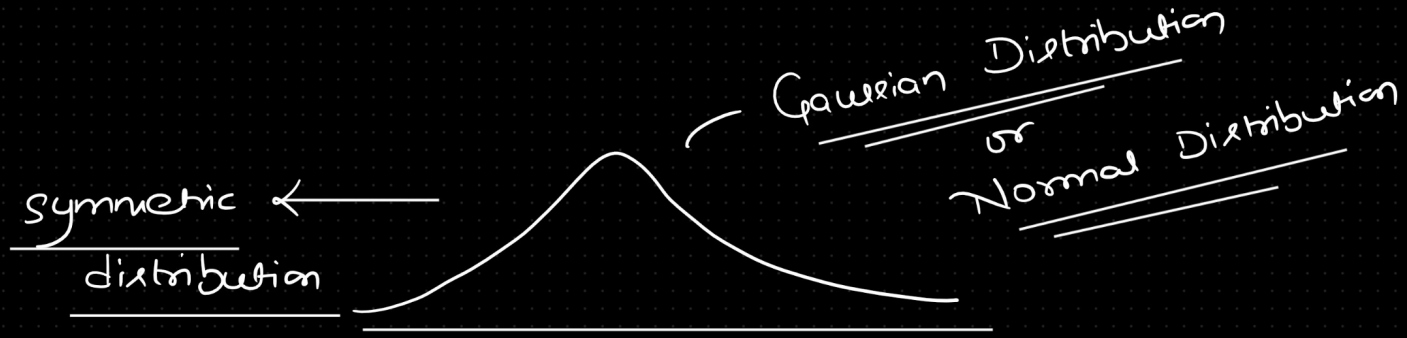



-ve correlation coefficient




+ve correlation coefficient

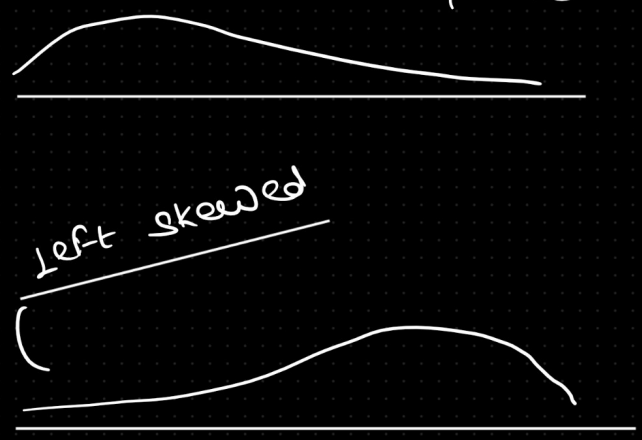



Standard Normal Distribution

$\mu = 0, \sigma = 1$

Non-Symmetric Distribution

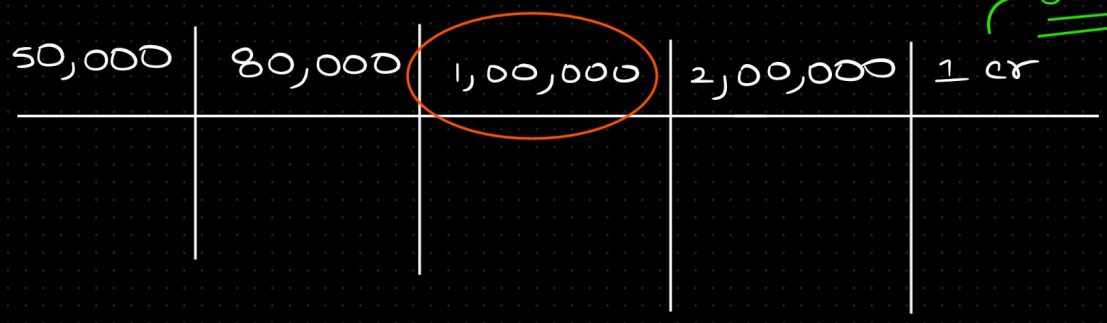



→ Outliers

Bengaluru — freshers

→ Data Scientist

$n = 5$



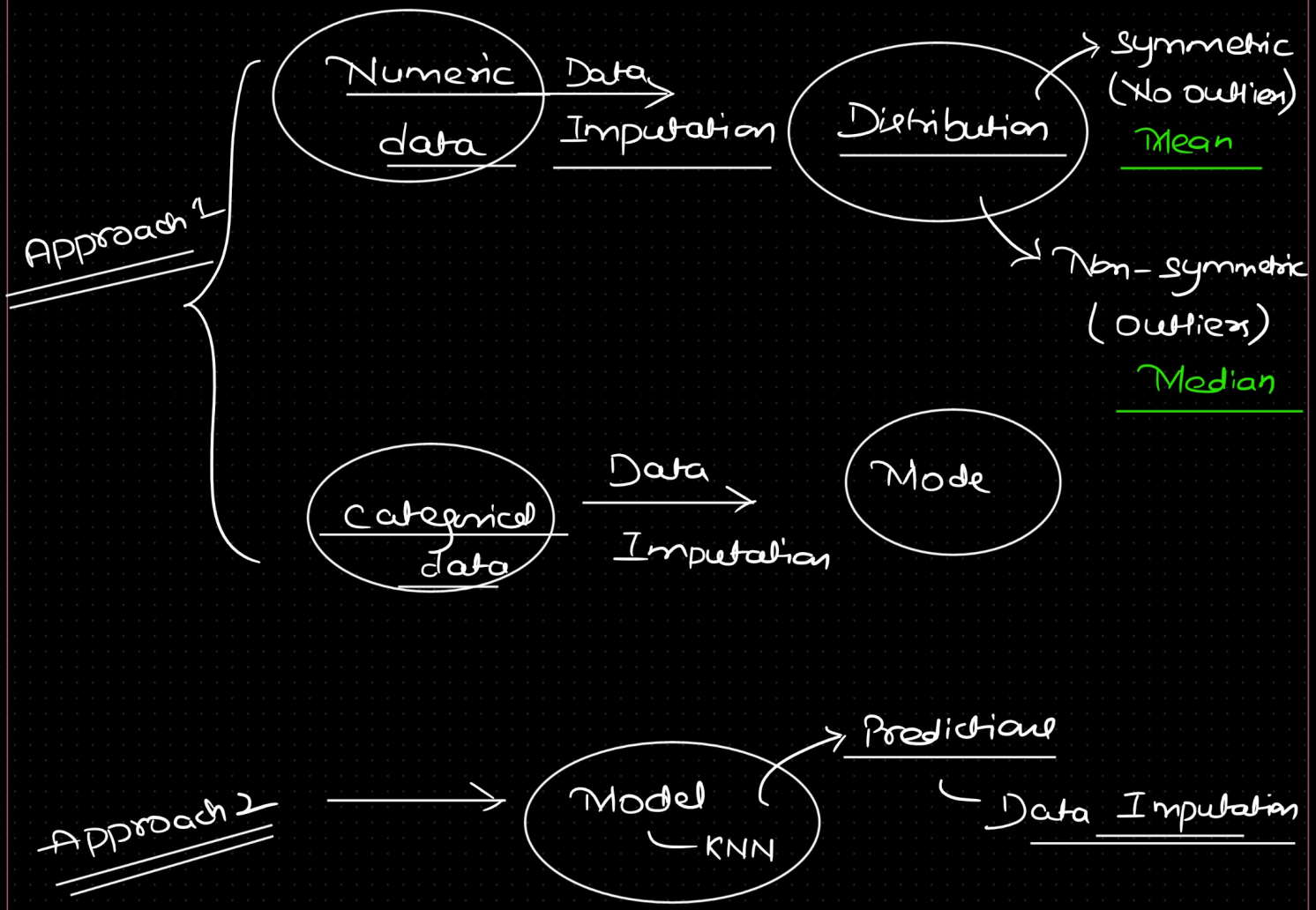

Mean → $\frac{50{,}000 + 80{,}000 + 1\text{Lakh} + 2\text{Lakh} + 1\text{cr}}{5}$

→ Shifts alot

Median → ① Sort (ascending)

② Middle → 1 Lakh

Median is more robust to outliers


Approach 1
Numeric data
Data Imputation
Distribution
Symmetric
(No outliers)
Mean
Non-symmetric
(Outliers)
Median
Categorical data
Data Imputation
Mode
Approach 2
Model
KNN
Prediction
Data Imputation