↳ Handle missing values

↳ Outliers (Distribution of Data)

↳ Symmetric & Skewed

↳ outliers

↳ Median

(More robust to the outliers)

Relationship b/w mean, median & mode

↳ Categorical data

↳ mode

↳ Descriptive Statistics

Handle Categorical data

↳ numeric form (One hot encoding & Label encoding)

colors = [Green, Blue, Red, Green, Green, Red]

Encoding Techniques

One hot encoding

Dummy variables

Colors_Green	Colors_Blue	Colors_Red
1	0	0
0	1	0
0	0	1
0	0	0
1	0	0
1	0	0
0	0	1

Red → 1

Dummy Variable Trap

No multicollinearity

drop_first = True

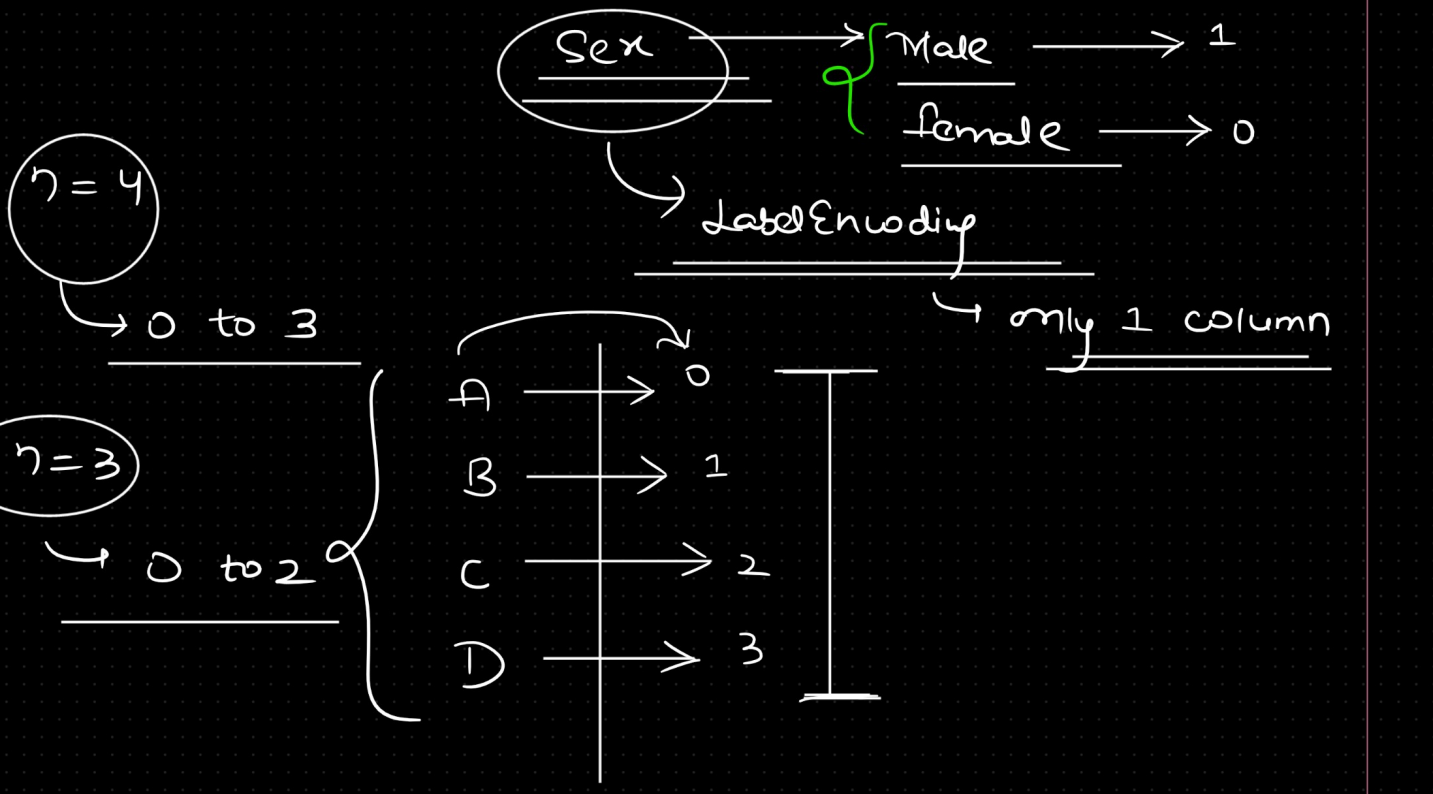
→ avoid the issue of dummy variable trap

~~Sex_Male~~

Sex_female

1 — female

0 — male



mapping

fit_transform()

fit() → {male, female}
male → 1
female → 0

transform()

complete feature 'sex' will be transformed as per the above mapping

Types of Categorical data

more than 2 classes

- Nominal — one hot encoding
  - Color = {Red, Green, Blue}
  - Embarked = {S, C, Q}
- Ordinal — Label Encoding
  - Same sort of order (Priority)
  - Education = {B.Tech (0), M.Tech (1), PhD (2)}
  - Ranking = {1, 2, 3, 4 -- 25}
    - I
    - II
    - III

Correlation coefficient (Heatmap)

→ -1 to 1

feature selection

→ Curse of Dimensionality

f_1 f_2 f_3 f_4 f_5 ——— f_{300}

300 Dimensional

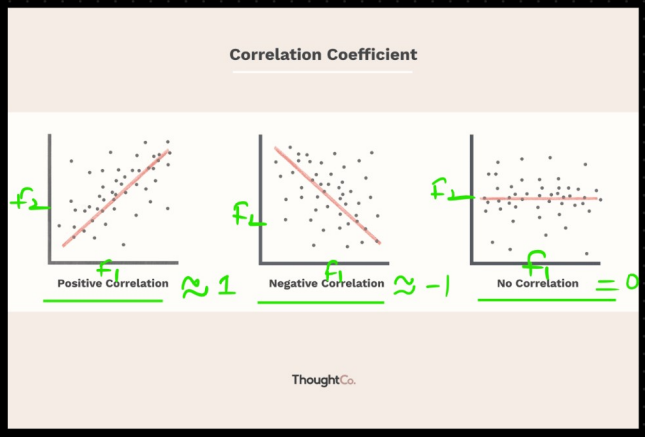
(Complex

Subset of the above features → feature selection

Capture the important info from all the above features & recreate new features all together → PCA, tSNE

Negative Correlation	Positive Correlation
↑ f_1 f_2 ↓	f_1 ↑ f_2 ↑ → 0.98
near to -1	near to +1

Redundant information



> 85% similar/redundant info

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

numeric values

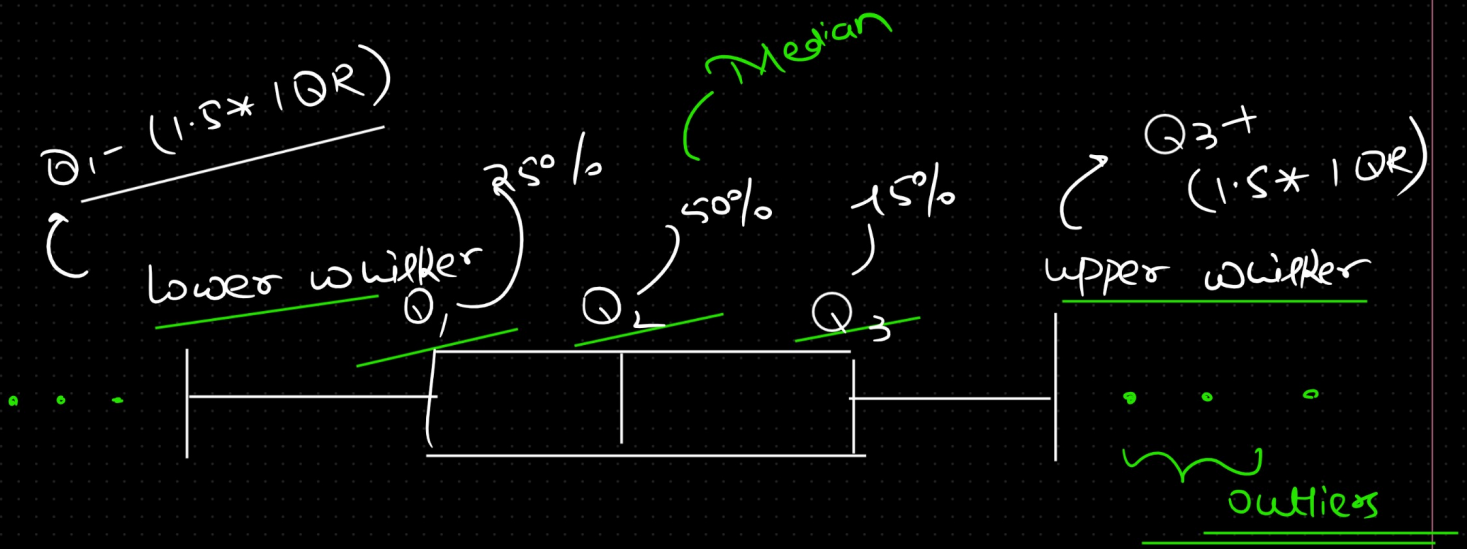
- $x \rightarrow f_1$
- $y \rightarrow f_2$

Outliers Detection

↳ Box Plot

$$\text{Range} = \text{Max} - \text{Min}$$

$$\text{IQR} = Q_3 - Q_1$$



m = 14

0 1 2 3 4 5 6 7 8 9 10 11

(5, 40, 42, 46, 48, 49, 50, 52, 53, 55, 56, 58)

~~(5, 102)~~ 12 13

Median → Sort the data (ascending order)

50% — $Q_2 = \frac{50+52}{2} = 51$

25% — $Q_1 = 46$

75% — $Q_3 = 56$

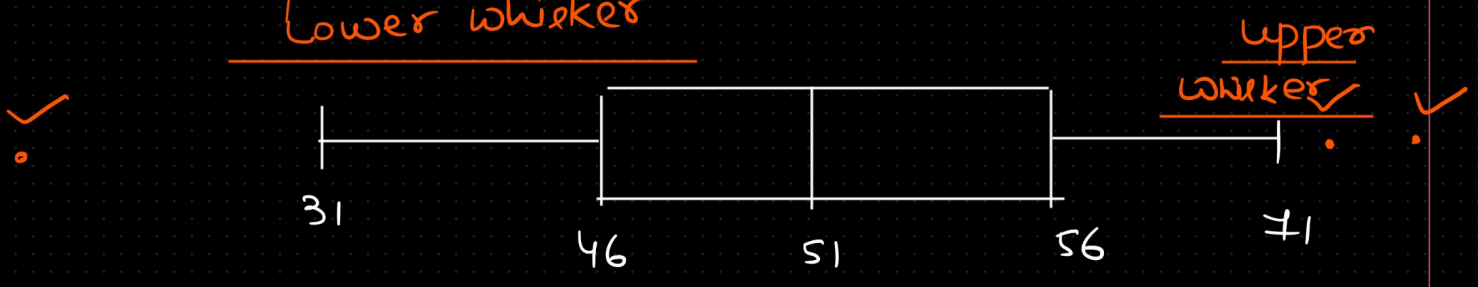
$IQR = Q_3 - Q_1 = 56 - 46 = 10$

Lower whisker $= 46 - (1.5 * 10)$ Research

$\Rightarrow 46 - 15 \Rightarrow 31$

Upper whisker $= 56 + (1.5 * 10)$

$\Rightarrow 56 + 15 \Rightarrow 71$



Grouping & Aggregation

↳ SQL