

Pandas Package



Data Structure → ① Dataframe

↳ 2D

② Series

↳ 1D

Summary

{ → head(), tail(), info(), describe(),
dtype, isnull().sum()

→ 891 entries

↳ loc vs iloc } → Data Access

Outliers (Handle Missing Values)

↓
Data Analytics in Bengaluru

(fresher) [0 - 2 yrs]

ctc	ctc	ctc	ctc	exceptional value
8 LPA	12 LPA	15 LPA	1 cr	

Outlier

$$\underline{\text{mean}} = (8 \text{ LPA} + 12 \text{ LPA} + 15 \text{ LPA} + 1 \text{ cr}) / 4$$

$$\Rightarrow \underline{\underline{3375000}} \\ \underline{\underline{(33 \text{ LPA})}}$$

correct indicator

$$\underline{\underline{\text{median}}} \Rightarrow \underline{\underline{13.5 \text{ LPA}}}$$

→ is more robust

Numeric data

to the outliers



Categorical data → Mode

How we detect the outliers??

→ Data Distribution

Symmetric

Distribution

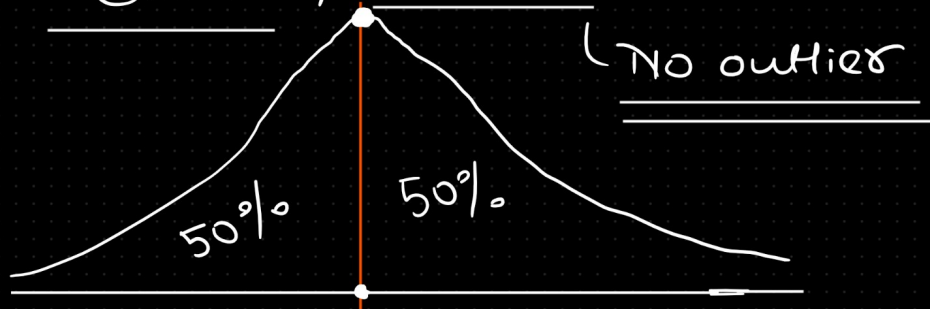
(Normal or

Gaussian Distribution)

Non-Symmetric

Distribution

Bell-shaped curve



mean = median = mode

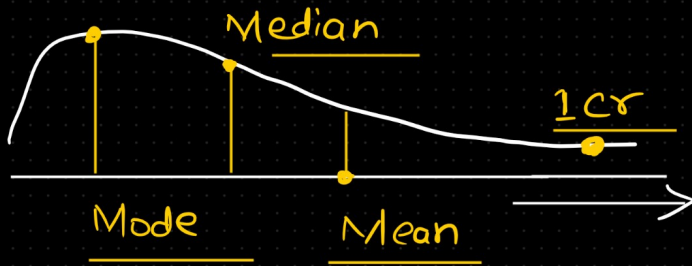
Relationship b/w

mean, median

& mode

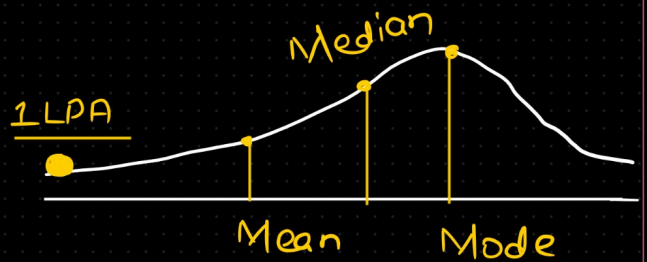
Skewed Distribution

Right-skewed



$$\text{Mean} > \text{Median} > \text{Mode}$$

→ Outliers
Left-skewed



$$\text{Mean} < \text{Median} < \text{Mode}$$

Age

23

33

44

57

65

76

$$44 + 57$$

2

$$= \frac{101}{2}$$

50.5

Categorical data

How we

Can handle
categorical

① One hot Encoding

② Label Encoding

numeric data

data?

Original Data		One-Hot Encoded Data			
Team	Points	Team_A	Team_B	Team_C	Points
A	25	1	0	0	25
A	12	1	0	0	12
B	15	0	1	0	15
B	14	0	1	0	14
B	19	0	1	0	19
B	23	0	1	0	23
C	25	0	0	1	25
C	29	0	0	1	29

Team → Categorical

✓ ✓ ✓ data
A, B, C

increase in the num of features

One hot encoding

	<u>Team_A</u>	<u>Team-B</u>	<u>Team-C</u>
25	<u>1</u>	0	0
12	<u>1</u>	0	0
15	<u>0</u>	1 - B	0
14	<u>0</u>	1 - B	0
19	<u>0</u>	1 - B	0
23	<u>0</u>	1 - B	0
25	<u>0</u>	0	1
29	<u>0</u>	0	1

Dummy → Team-A, Team-B,
Variables Team-C

Dummy variable Trap

→ Multicollinearity

A diagram illustrating a dummy variable trap. It consists of a 2x3 grid of cells. The columns are labeled 'Team-A', 'Team-B', and 'Team-C' at the top, each enclosed in a yellow oval. A yellow bracket above the first two columns indicates a relationship. The first row contains the values 1, 0, and 0, each underlined twice. The second row contains the values 0, 0, and 1, each underlined twice. The grid is defined by white lines on a black background.

Team-A	Team-B	Team-C
<u>1</u>	<u>0</u>	<u>0</u>
<u>0</u>	<u>0</u>	<u>1</u>