Pandas Package

↓

Data Structure → ① Dataframe
└ 2D

② Series
└ 1D

Summary

↳ head(), tail(), info(), describe(),
dtypes, isnull().sum()

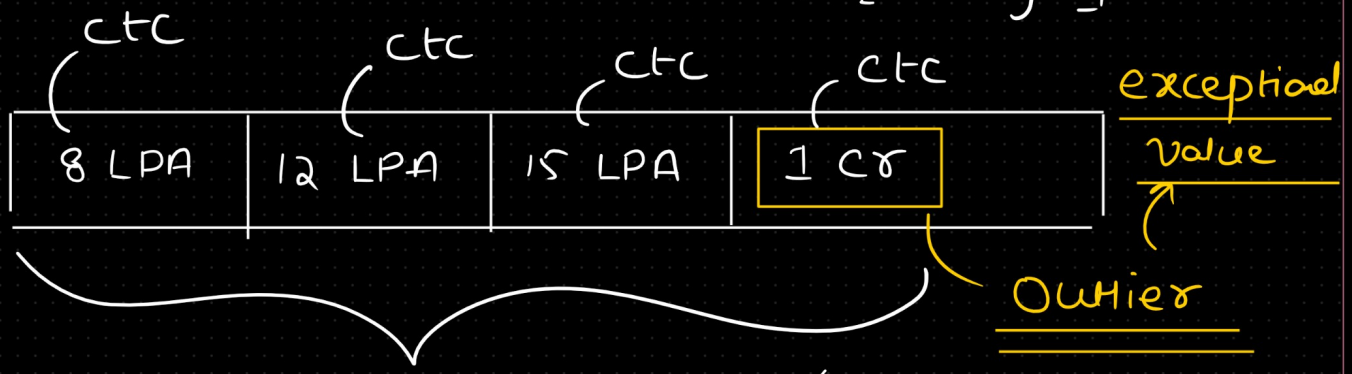
↳ 891 entries

↳ Loc vs iloc → Data Access

Outliers (Handle Missing Values)

Data Analytics in Bengaluru

(fresher) [0 - 2yrs]



mean = (8 LPA + 12 LPA + 15 LPA + 1cr)/4

$\Rightarrow$ 3375000

(33 LPA)

median $\Rightarrow$ 13.5 LPA

correct indicator

→ is more robust to the outliers

Numeric data

Outliers → Data Imputation

→ Median

Categorical data → Mode

How we detect the outliers ??

→ Data Distribution

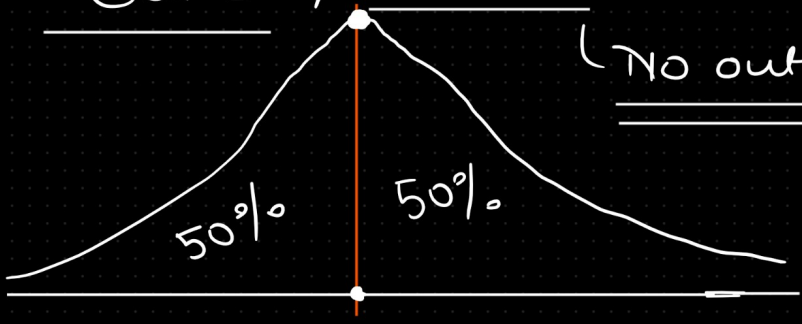
Symmetric Distribution

(Normal or Gaussian Distribution)

Non-Symmetric Distribution

Bell-shaped curve

└ No outlier



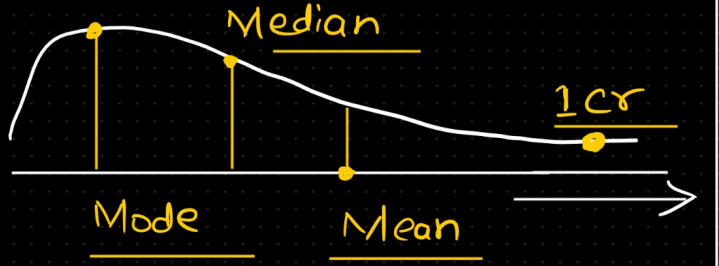
mean = median = mode

Relationship b/w mean, median & mode

Skewed Distribution

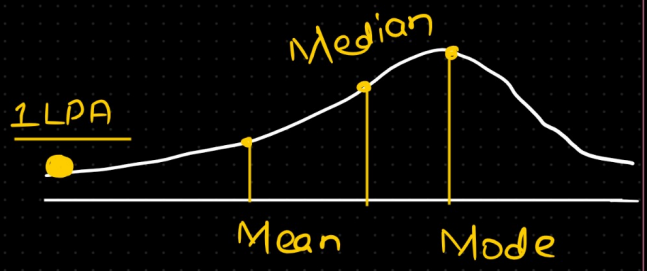
↳ Outliers

Right-skewed



Mean > Median > Mode

Left-skewed



Mean < Median < Mode

Age

23
33
44
57
65
76

$$\frac{44+57}{2} = \frac{101}{2} = 50.5$$

Categorical data → How we Can handle Categorical data?

1. One hot Encoding
2. Label Encoding → numeric data

Original Data

Team	Points
A	25
A	12
B	15
B	14
B	19
B	23
C	25
C	29

One-Hot Encoded Data

Team_A	Team_B	Team_C	Points
1	0	0	25
1	0	0	12
0	1	0	15
0	1	0	14
0	1	0	19
0	1	0	23
0	0	1	25
0	0	1	29

Team → Categorical data → [A, B, C]

increase in the num of features

One hot encoding

	Team_A	Team_B	Team_C
25	1	0	0
12	1	0	0
15	0	1 - B	0
14	0	1 - B	0
19	0	1 - B	0
23	0	1 - B	0
25	0	0	1
29	0	0	1

Dummy Variables → Team_A, Team_B, Team_C

Dummy variable Trap

→ Multicollinearity

Team_A	Team_B	Team_C
1	0	0
0	0	1