Missing data

Reasons of missing data :-

→ human error
→ data corrupt

# of rooms	Price of house
1	28L
2	30L
3	50L
—	60L
5	70

* Missing data occurs in a dataset when some of the information is not saved / stored.

① Missing Completely at Random (MCAR)

→ The data missing is independent of observed and Unobserved data / missing data.

* Survey : if a person is suffering from a disease or not

Patient	disease	
A	Yes	Here missing value is not depending on either of observed value or other missing values.
B	NO	
C	Yes	
D	—	
E	Yes	
F	—	
G	Yes	

→ Every patient is independent of each other.

② Missing at Random (MAR)

→ The missing data depends on observed data but not on the missing data itself.

Survey To know the income of people.

family → income of all family member.

→ One of the family member don't want to disclose the income due to age/experience/gender or any other factor but not due to their income level.

③ Missing data not at Random (MNAR)

→ Missing values depends on the value of missing data itself.

Eg: income and job satisfaction of employees in a company.

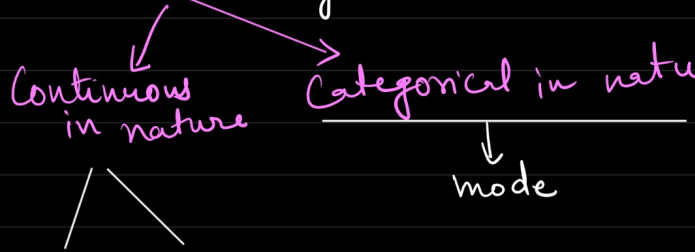
→ The employees who are not satisfied in the company will refuse to report their income, then data is Missing data not at Random.

* Missing value treatment (should be asked from business team)

Why? → For analysis & ML prediction it is important to deal with missing value.

① if the data is huge and missing value is less than 1% of data, you can drop the missing value.

② impute (Replace) the missing values



mean ↙ ↘ median

* if the outlier treatment is not done then impute the missing value with median.

③ if any column has missing value greater than 40%, drop the column.

④ impute the missing value with a random number or with a value which is not possible.

Marks
100
50
60
—
90
70
30

-1 / Ab / Na