Types of Missing Values

06 July 2023 20:49

The real-world data often has a lot of missing values. The cause of missing values can be **data corruption** or **failure to record data**. The handling of missing data is very important during the preprocessing of the dataset as **many machine learning algorithms do not support missing values**.

MCAR – Missing Completely At Random
MAR – Missing At Random
MNAR – Missing Not At Random

	Job Performance Ratings			
IQ	Complete	MCAR	MAR	MNAR
78	9	Missing	Missing	9
84	13	13	Missing	13
84	10	Missing	Missing	10
85	8	8	Missing	Missing
87	7	7	Missing	Missing
91	7	7	Missing	Missing
92	9	9	9	9
94	9	9	9	9
94	11	11	11	11
96	7	Missing	7	Missing
99	7	7	7	Missing
105	10	10	10	Missing
105	11	11	11	11
106	15	15	15	15
108	10	10	10	10
112	10	Missing	10	10
113	12	12	12	12
115	14	14	14	14
118	16	16	16	16
134	12	Missing	12	12

Note. MCAR = Missing Completely at Random. MAR = Missing at Random. MNAR = Missing Not at Random. For the MCAR data, there is no relationship between IQ and the job performance ratings. A case with a lower IQ is just as likely to be missing as a case with a higher IQ. With MAR data, all cases with missing job performance ratings are participants with a lower IQ. With the MNAR data, IQ can't be used to account for the pattern of missingness because IQ doesn't explain why the cases are missing. The examples are based on Enders (2010).

Suppose, in our dataset of 1000 people we have two fields:

Name	Age

In the age field, there are 50 values are missing. These missing values are randomly distributed. If we pick a value randomly from Age, the probability of that value being missing is 50/1000 = 0.05.
And probability of that picked value being Not-Missing is 950/1000 = 0.95

In this case, missing values are randomly distributed in such a way that we are not 100% sure that selected value is missing or not-missing. In this case the missingness is of **MCAR** type.

Let's say another dataset of 1000 person's(500 Male + 500 Female) age:

Name	Gender	Age

Imagine in the above datasets, out of 50 missing values, 40 missing values are in Female Category and 10 in Male Category.
Now probability of picking Missing value if the category is Female is (40/1000) more than missing value from Male category (10/1000). We can conclude that if we consider the observed values in Gender field then the probability of picking missing value is affected. Hence missingness of age depends on observed value in Gender. But in the above dataset we have nothing such field observed value of which affect missingness. In this case, the missingness is of **MAR** type.

MNAR: here we can say the number of missing values from both categories don't differ much and we can't say that missing values depends on any category.

But in the data we can notice that the missing values are distributed systematically, it means missing values are very close to each other or following some unknown way of spreading. It means probability of Missing values getting picked up is affected by some unobserved factors. Hence this is of **MNAR.**

1. **MCAR (Missing Completely At Random)**:

In this case, missing values are randomly distributed across the feature, and it is assumed that other features cannot predict the missing values.

Missing completely at random (MCAR) analysis assumes that missingness is nowhere dependent on observed and unobserved data.

Office Employees Data:

ID	Age	Salary/Yr (Lacs)
A1	22	12.0
B2	36	19.7
B5	28	26.5
A6	25	?

E3	45	12.0
F10	31	6.8
X1	23	5.6

2. **MAR (Missing At Random)**:

In MAR, missing values are randomly distributed, but these can be predicted by other features/data points.

This analysis assumes that missingness is only dependent on observed data but not on unobserved(Missing) data.

Office Data Scientist Salary Data:

ID	Experience(Yrs)	Salary/Yr (Lacs)
A1	2	15
B2	5	32
B5	3	21
A6	12	?
E3	4.5	30
F10	14	?
X1	0	8.5
R2	15	80
N13	7	50
K12	2	16

3. **MNAR (Missing Not At Random)**:

In this case, missing values are not random and missing systematically.

This analysis assumes that missingness is somehow dependent on unobserved(Missing) data.

Office Data Scientist Age Data:

ID	Gender	Age
A1	M	27
B2	F	23
B5	F	? - Absent/Don't want to tell
A6	M	42
E3	F	36
F10	M	33
X1	F	29
R2	M	22
N13	F	? - Absent/Don't want to tell
K12	F	26