파이썬 문법 응용반

(To be renamed
Basics to Data Analysis (ML))

20240106_결측값 처리 (Missing valuesTreatment)

결측값이란? (Missing value)

결측값에 대해서 -> 결측값이 무엇인가?
- na값 ,np.nan 데이터를 봤을 때 빈 값들

MCAR : Missing Completely at Random (완전 무작위 결측)

순수하게 결측값이 -> 무작위로 된 결측값, 편향으로 인한 결측값이 아닌 정말 랜덤성 결측

MNAR : Missing at not Random (비무작위 결측

결측값이들이 변수의 하나의 값 , Na 의미가 있는 값
결혼 유무 , 결혼 안 함 na, 금융데이터 -> 민감질문 소득 이런 것들 답하기 싫으면 na

MAR : Missing at Random (무작위 결측)

다른 변수랑 관련이 있어서 결측치가 발생하는 경우
어떤 공장의 시스템 데이터가 있다.
오작동이 나는 경우에는 A 컬럼이 Na 데이터 수집이 되지 않는다.
B 컬럼이 오작동 유무를 판단하는 것
오작동인 경우 -> A 컬럼 Na 값으로 된다.

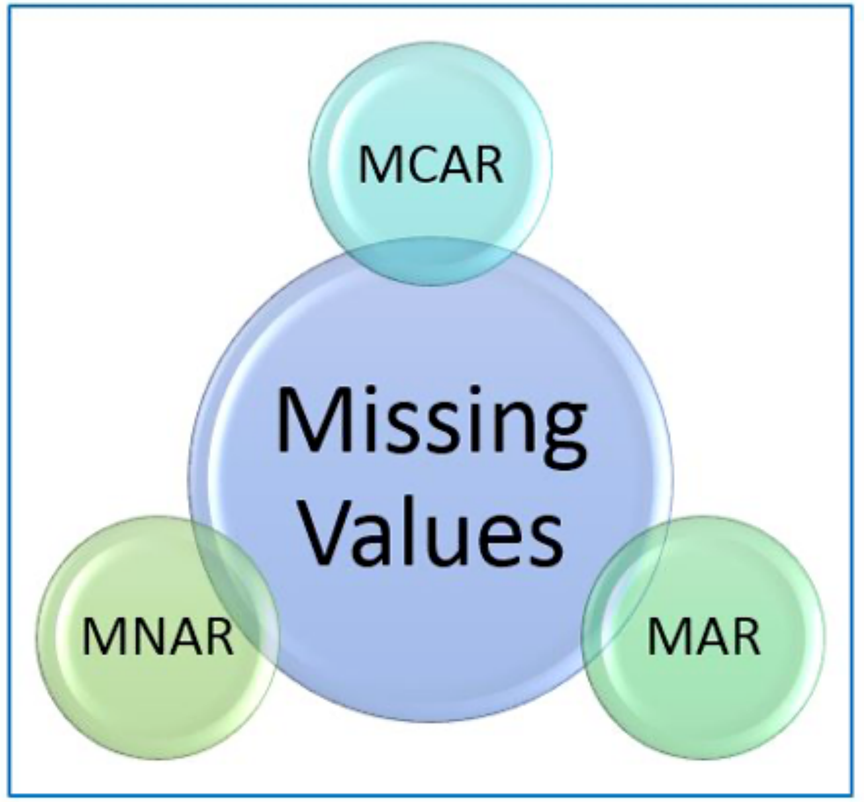


Figure 1 - Different Types of Missing Values in Datasets

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0.0	A/5 21171	7.2500		S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1			71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female		0	0.0	STON/O2. 3101282	7.9250		S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0.0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0.0	373450	8.0500		S
...	...	...	...	...	...	...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0.0	211536	13.0000		S
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0.0	112053	30.0000	B42	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female		1	2.0	W./C. 6607	23.4500		S
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0.0	111369	30.0000	C148	C
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0.0	370376	7.7500		Q

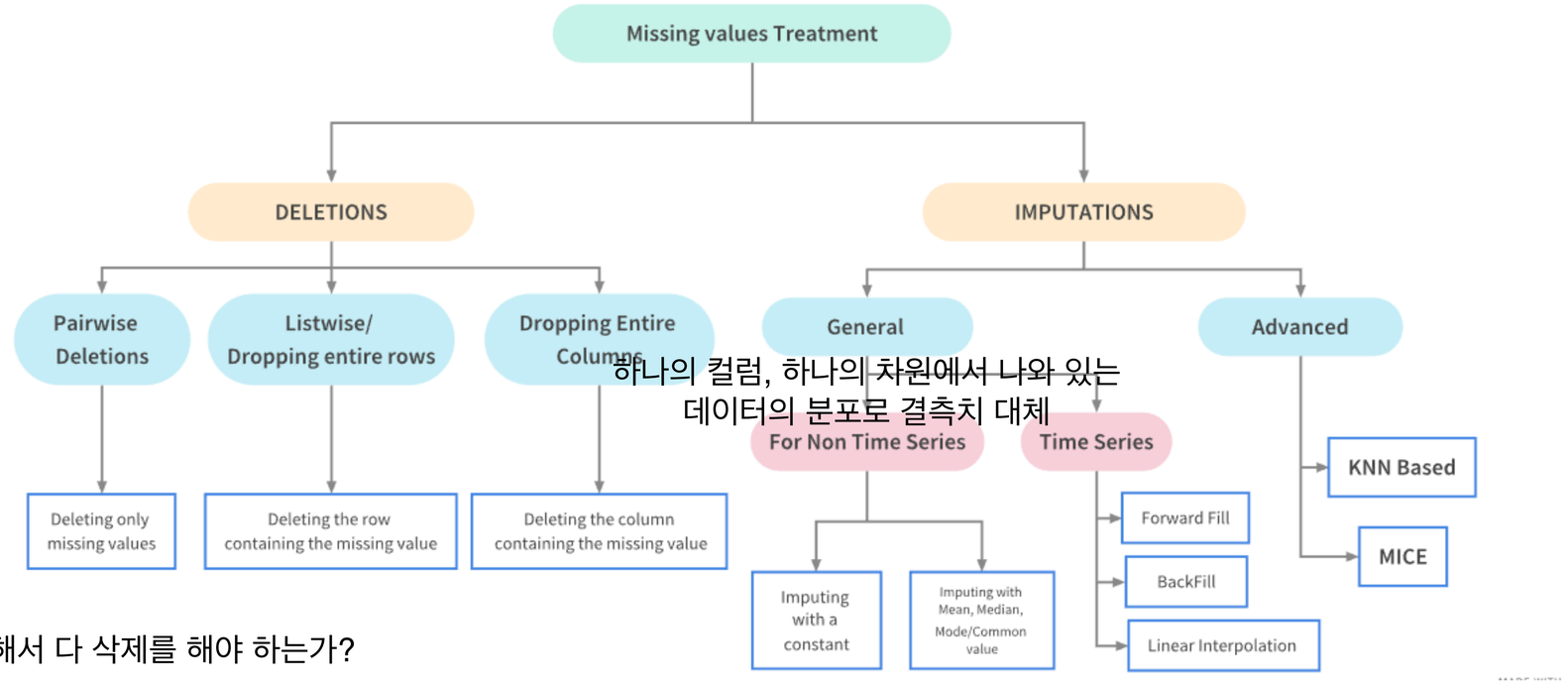
891 rows × 12 columns

연속형, 범주형에 따른 처리 방법도 다를 것

어떤 식으로 처리해야 할까?

실무에서 Missing Value는?

Missing Values Treatment



하나의 컬럼, 하나의 차원에서 나와 있는
데이터의 분포로 결측치 대체

Missing Value라고 해서 다 삭제를 해야 하는가?

전체 데이터로 봤을 때 비중이 최소 1%미만 정도면 삭제를 해도 괜찮을 수 있다.
삭제를 했을 때 중요한 건 이 컬럼이 영향을 줄 수 있는 컬럼인가?

Missing Value라고 해서 다 대체를 해야 하는가?

대체 하는 방법도 정말 다양하고
대체 방법에 따라 성능이나 어떤 영향을 줄 수 있는 게 크다.

Handling Missing Data