Train In Data

Missing Data

Missing Data: Definition

- Missing data, or missing values, occur when no data is stored for a certain observation in a variable.
- Missing data are a common occurrence in most datasets
- Missing data can have a significant effect on the conclusions that can be drawn from the data.

Missing Data: Causes

Lost

- A value is missing because it was forgotten, lost or not stored properly.

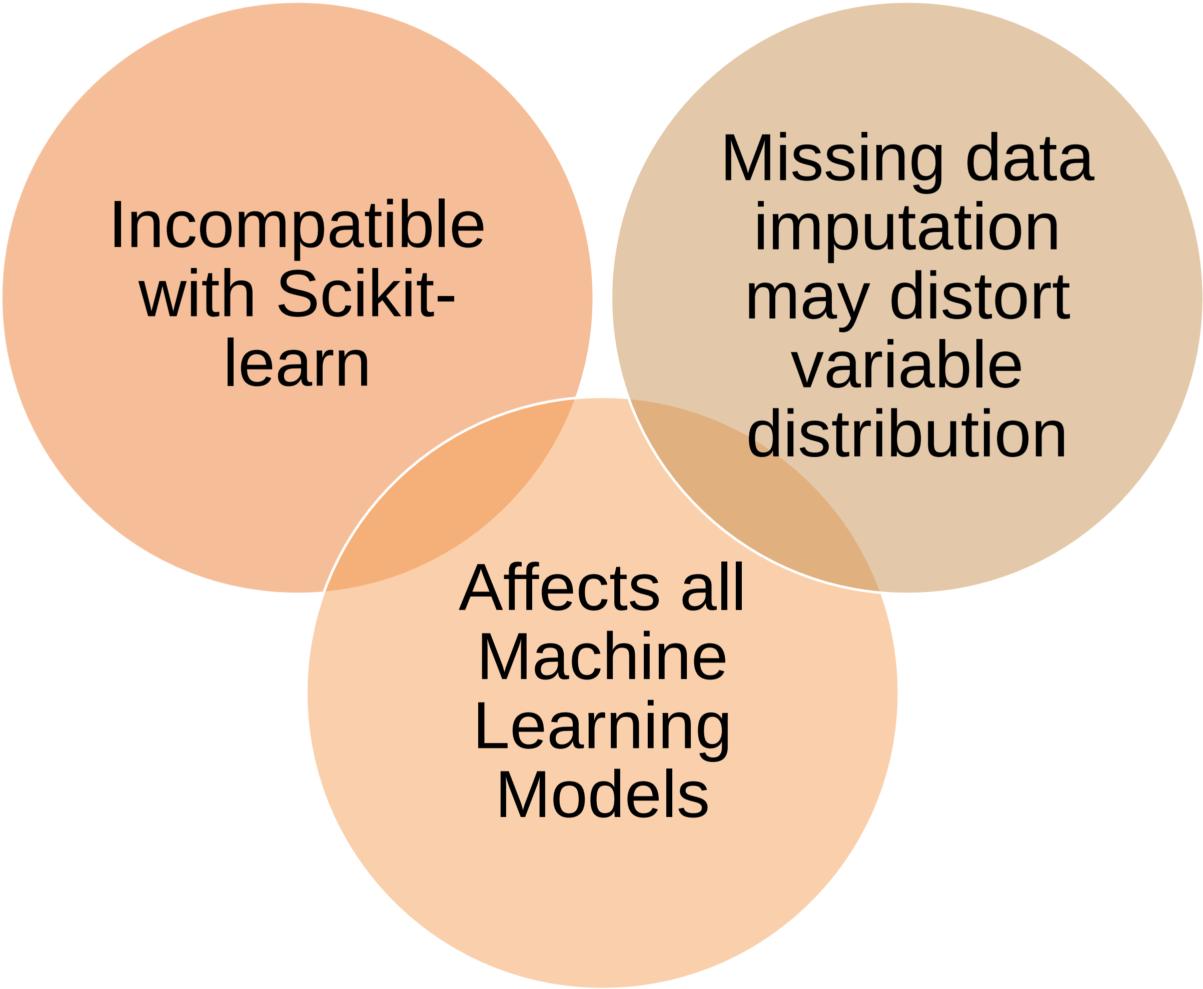
Don’t exist

- E.g., a variable is created from the division of 2 variables and the denominator takes 0.

Not found | Not Identified

- E.g., when matching data against postcode, or date of birth, to enrich with more variables, and the postcode or dob are wrong or don’t exist, the new variables will take NA.

Missing Data: Impacts

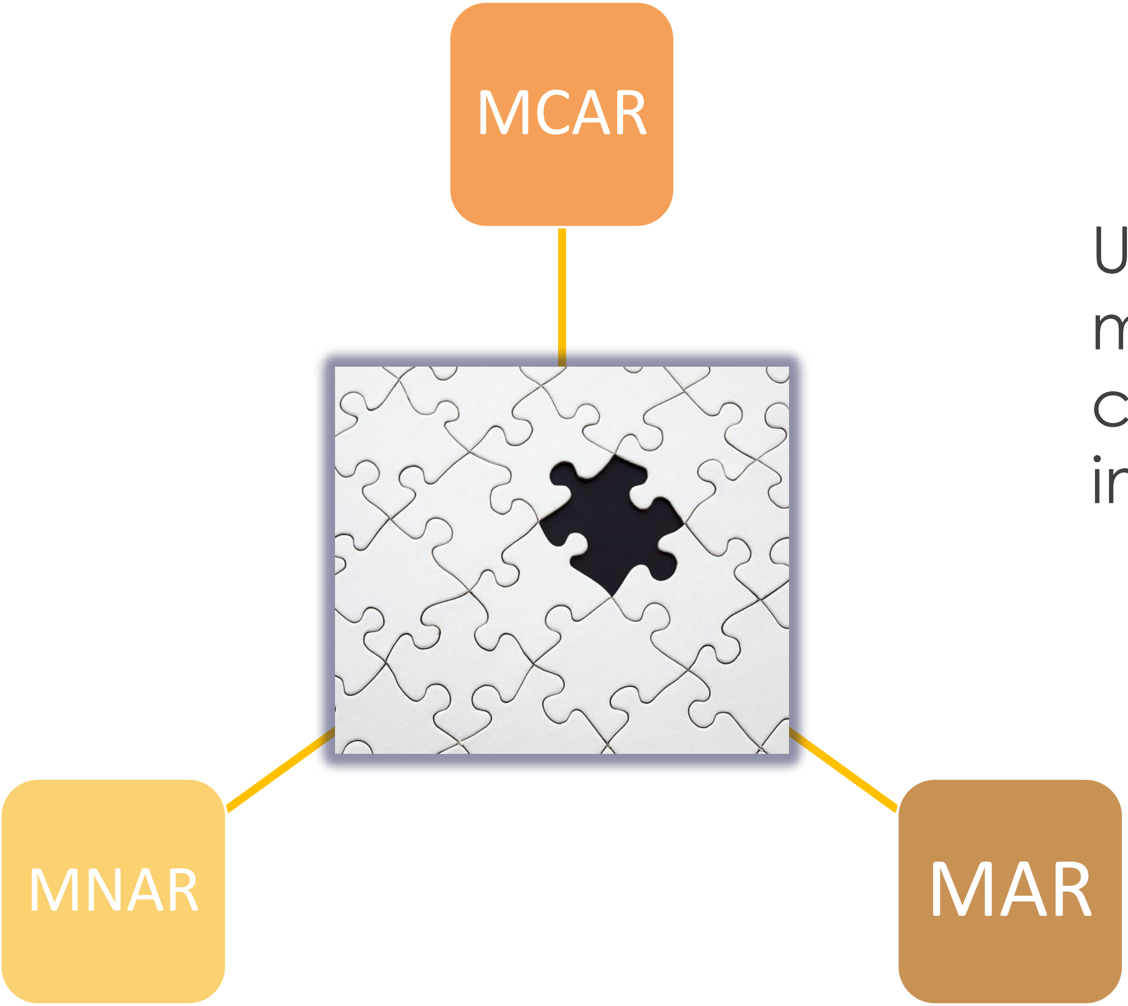


Incompatible with Scikit-learn

Missing data imputation may distort variable distribution

Affects all Machine Learning Models

Missing Data: Mechanisms



MCAR

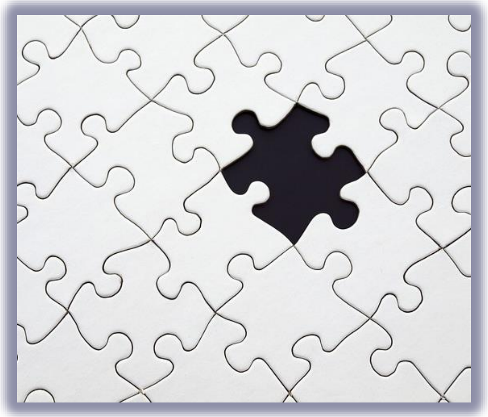
MNAR

MAR

Understanding the missing data mechanisms may help us choose the right missing data imputation technique

Image taken from here

Missing Data Completely at Random (MCAR)



- The probability of being missing is the same for all the observations
- There is absolutely no relationship between the data missing and any other values, observed or missing, within the dataset
- Disregarding those cases would not bias the inferences made

Image taken from here

Missing Data at Random (MAR)

- The probability of an observation being missing depends on available information

Gender	Weight
Male	60 kg
Male	NA
Male	NA
Male	77 kg
Male	80 kg
Male	62 kg
Female	NA
Female	NA
Female	60 kg
Female	55 kg
Female	NA
Female	58 kg

2 NA / 6 men = 33%

3 NA / 6 women = 50%

Missing Data not at Random (MNAR)

- there is a mechanism or a reason why missing values are introduced in the dataset.

Target = depression	No of clinic visits	No sports classes weekly
Yes	1	NA
Yes	NA	NA
Yes	NA	0
Yes	4	2
Yes	NA	1
Yes	3	NA
No	0	0
No	NA	5
No	1	2
No	1	1
No	2	1
No	NA	2

More NA overall for depressed patients

Less NA for non-depressed patients

In addition

- To understand the mechanisms by which missing data is introduced, we need to become familiar with the methods used for data collection.

- This is not always possible. However, it is a good idea to understand the methods of data collection as much as possible, to decide how best to engineer the features.

Accompanying Jupyter Notebook

- Read the accompanying Jupyter Notebook

- Examples of MCAR, MAR and MNAR
- Titanic dataset
- Loan Book from P-2-P company

THANK YOU

www.trainindata.com