Missing Data

Intro

What is missing data?

Some of the Causes:

- incomplete information provided by participants.
- non-response from those who decline to share information.
- poorly designed surveys.
- removal of data for confidentiality reasons.

How the missing data is represented?

- the common placeholder for missing values is **NaN.**
- NumPy uses the IEEE Standard for Binary **Floating**-Point for Arithmetic (IEEE 754). This means that Not a Number is not equivalent to infinity.
- **NaN** and **NAN** are aliases of **nan**.
- Pandas has a method called `fillna()` that can help us with handling `NaN` values.
- However, sometimes the missing values are represented as `?`, `:`, `*`, etc.

Why do we need to handle the missing data?
To avoid:

- bias the conclusions.
- leading the business to make wrong decisions.

Types of missing data:

Missing Completely at Random (MCAR)

- We are assuming that whether or not the person has missing data is completely unrelated to the other information in the data.
- Examples: mistakes in data entry, temporary sensor failures, etc.
- The amount of missingness is low.
- All the variables and observations have the same probability of being missing.

ID	Gender	Age	Income
1	Male	Under 30	Low
2	Female	Under 30	Low
3	Female	30 or more	High
4	Female	30 or more	**.....?**
5	Female	30 or more	High

Missing at Random (MAR)

- Missing data will have some systematic relationship with the other observed features.
- the value that is missing based on the other data.
- Examples: data missing from observational equipment during scheduled maintenance breaks, a poll with different missing data based on gender.

Missing Not at Random (MNAR) - nonignorable missing data

- missing values may exist in large amounts, and the reason for the missingness is associated with factors beyond our control or knowledge.
- we cannot confidently make any conclusions about the likely value of missing data.
- For example, it is possible that people with very low incomes and very high incomes tend to refuse to answer.
- we cannot use any of the standard methods for dealing with missing data
- any standard calculations give the wrong answer.
- poll with different missing data based on the name.
- depends on how much data is true.

Handling Missing Data

Data Dropping

Pros

- Straightforward and simple to use.
- Beneficial when missing values have no importance.

Cons

- lead to information loss, which can introduce bias to the final dataset.
- This is not appropriate when the data is not missing completely at random.
- Data set with a large proportion of missing values can be significantly decreased, which can impact the result of all statistical analysis on that data set.

Imputation

- Mean, Median, and Mode Imputation
- Constant Value
- Forward Filling
- Backward Filling
- Random Sample Imputation

Mean/Median Imputation

- The normal distribution is the ideal scenario, but it is not always the case.

Pros

- Simplicity and ease of implementation
- no additional data is required as the imputation is performed using the existing information.
- Good for normally distributed data, and skewed data.

Cons

- only work for numerical data.
- Mean imputation is sensitive to outliers and may not be a good representation of the central tendency of the data.
- median also may not better represent the central tendency.
- makes the assumption that the data is MCAR, which is not always true.

Random Sample Imputation

Steps:

- Creating two subsets from the original data.
  - first contains all the observations without missing data.
  - second contains those with missing data.
- select random observation from each subset.
- missing data is replaced with the first subset.
- the process continues until there is no more missing information.

Pros

- easy and straightforward technique.
- for both numerical and categorical data types.
- less distortion in data variance

- it preserves the original distribution of the data, which is not the case for mean, median, and more.

Cons

- Randomness does not necessarily work for every situation, and this can add noise to the data.
- assumes that the data is missing completely at random (MCAR).

Best Practices

Adopting the right approach can save from introducing bias in the data and making wrong decisions.

Type of missing data	Imputation method
Missing Completely At Random	Mean, Median, Mode, or any other imputation method
Missing At Random	Multiple imputation, Regression imputation
Missing Not At Random	Pattern Substitution, Maximum Likelihood estimation

Resources:

- https://chen5595.cems.umn.edu/SameStats-DifferentGraphs.pdf
- https://matplotlib.org/2.1.1/api/_as_gen/matplotlib.pyplot.plot.html
- Wes McKinney, 2018, Python for Data Analysis, 2ed edition, O'Reilly Media, US.
- https://www.idrc.ca/sites/default/files/sp/Documents%20EN/10-data-visualization-tips-en.pdf
- https://www.datacamp.com/tutorial/techniques-to-handle-missing-data-values
- https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.fillna.html
- https://www.displayr.com/different-types-of-missing-data/#:~:text=There%20are%20four%20qualitatively%20distinct,as%20missing%20not%20at%20random).