Mechanisms of Missing Data

November 9, 2022 3:35 AM

- Exploratory Data Analysis
- Data Preprocessing
- feature Engineering.

1- Missing Completely At Random (MCAR)
2- Missing At Random (MAR)
3- Not Missing At Random (NMAR) or Missing Not At Random (MNAR)

Data belonging to each mechanism is handled differently. Let's understand these mechanisms first:

Biases

1- MCAR - means that missingness has the same correlation with all the categories of other features in the data. For example, when we take a random sample from a population, each member has an equal chance of being included. MCAR is the unobserved data of individuals in the population who were not included in the sample. This gives us the most confidence that we aren't systematically missing values from some of our respondents.

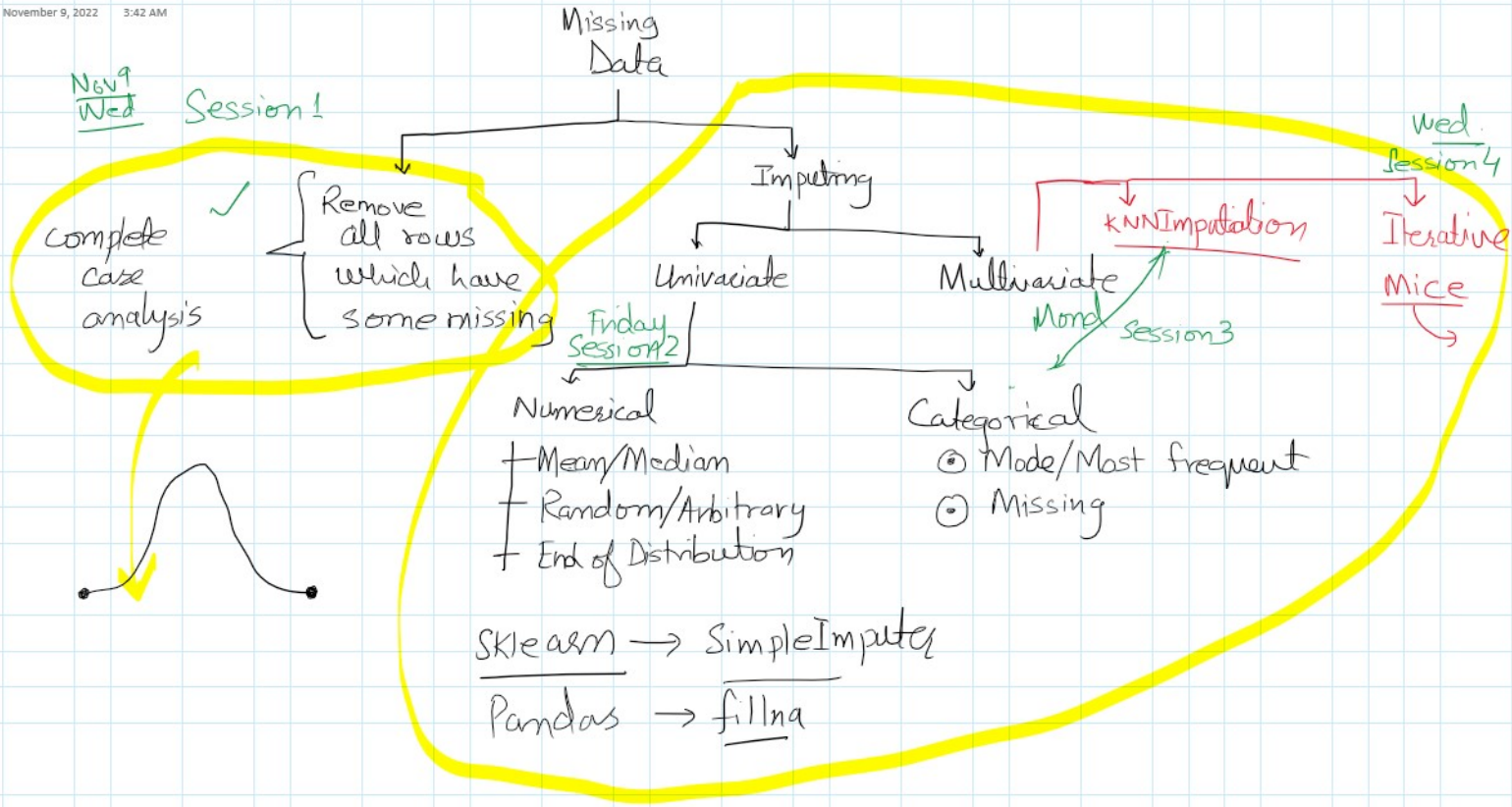
2- MAR - means there may be systematic disparities between missing and observed values, but these can be explained totally by other observed variables. For example, when we pick a sample from a population and the likelihood of inclusion is determined by some known property. Modern methods for missing data generally start from the MAR assumption.

3- NMAR - what it means is that missingness is due to unknown reasons and cannot be explained by any other variable in the data. For example, when we pick a sample from a population that is not representative of the overall population, like some minorities are not included in the population. This is the most challenging case for handling missingness, as one strategy to handle it would be to do what-if analyses for different scenarios or collect data that can explain the cause of missingness.

From <https://www.linkedin.com/feed/update/urn:li:activity:6994343161326178304/>

Strategies to Handle Missing Data

November 9, 2022 3:42 AM

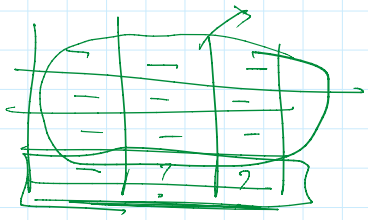


Complete Case Analysis → Removing missing data.

November 9, 2022 5:29 AM

Assumptions:

- MCAR
- Small amount of missing values. (< 5%)



Advantages:

- Easy to implement
- Preserves the distribution of the data.

Disadvantages:

- You can drop alot of key/valuable data.
- Excluded observations can be valuable.
- In production model will not know how to handle missing data.

When to use CCA

- MCAR
- < 5%