WIE3007 2023/2024 S1

Tutorial 7

Q1: Explain in your own words the importance of data quality in analytics. Why is it crucial to address missing data before proceeding with data analysis?

Data quality is important because it ensures that the information used to make key business decisions is reliable, accurate and complete. Before proceeding with data analytics, it is crucial to address missing data because this ensures the reliability of the insights, avoids biased conclusions, ensures no wasted resources and thus makes effective decision making in business.

Q2: Given a dataset from a survey, where younger participants tend to skip questions about retirement plans, while older participants often skip questions about recent technological trends:

a) Identify which type of missingness (MCAR, MAR, MNAR) applies to each scenario.

b) Provide a brief justification for your choices.

Younger participants skipping questions about retirement plans:

Missing At Random (MAR), *the missingness is related to an observed variable (age). Younger participants are more likely to skip questions about retirement plans, but this missingness is random within the groups defined by the observed variable (age). It's "at random" within the categories of age, making it MAR.*

Older participants skipping questions about recent technological trends:

Missing Not At Random (MNAR), *the missingness is not related to any observed variable; instead, it's associated with the values of the missing data itself. Older participants are more likely to skip questions about recent technological trends, and this missingness is not explained by any observed variables in the dataset. It's "not at random" because the probability of missingness depends on the unobserved variable of interest (interest or knowledge about recent technological trends), making it MNAR.*

Q3: You are given a dataset with 30% missing values for the variable "Income." You decide to remove all rows with missing "Income" values.

a) What potential problems might arise from this approach in your subsequent analysis?

- ***Biased sample.*** *The individuals with missing income values may have characteristics that differ systematically from those with reported incomes. This can lead to a sample that does not represent the overall population accurately.*
- ***Reduced sample size.*** *Deleting rows with missing values reduces the size of your dataset. A smaller sample size can lead to less precise estimates and wider confidence intervals, making it harder to draw meaningful and reliable conclusions from your analysis.*

b) How might the results of your analysis be biased?

- ***Selection bias.*** *The removal of rows with missing "Income" values creates selection bias if the likelihood of having missing income values is related to the variable of interest or the outcome you are studying.*
- ***Generalizability issues.*** *The results obtained from the analysis may not be generalizable to the broader population, as the removed data might represent a specific subgroup with distinct characteristics.*

Q4: You have visualized the missing data in a dataset using a missingness matrix (heatmap). The matrix shows clusters of missing values for certain variables.

a) What could be the potential reasons for these clusters of missing values?

- ***Survey design.*** *If the question is not set as a required question, the respondent might skip the question.*
- ***Participant Characteristics.*** *If participants of a certain demographic group are less likely to provide certain types of information, it could result in clusters of missing values.*

b) How would you determine if these clusters indicate MAR or MNAR?

- ***MAR (Missing at Random): I****f missingness can be explained by variables already in the dataset and not the missing values, it's more likely to be MAR.*
- ***MNAR (Missing Not at Random):*** *If the clusters of missing values are related to the values that are missing (e.g., high-income individuals not reporting income), it suggests MNAR.*

Q5: You've decided to use the mean imputation technique to handle missing values for a continuous variable in a dataset using SAS Enterprise Miner.

a) Describe the steps you would take in SAS EM to achieve this.

- ***Select data source and load into SAS enterprise miner.***
- ***Handle missing data.***
- ***Configure the handling missing values node.***
- ***Connect nodes.***
- ***Run the process.***

b) What are potential drawbacks of using mean imputation, and how might it affect the subsequent analysis?

- ***Introduction of Bias****. Mean imputation assumes that missing values are missing completely at random (MCAR). If this assumption is violated, the imputed values may not accurately represent the true underlying values, leading to biased results.*
- *L**oss of Information****. Mean imputation does not account for the uncertainty associated with imputed values. It essentially replaces missing values with a single point estimate (the mean), leading to a loss of information and potentially inaccurate standard errors.*