DA-IICT

IT 462 EXPLORATORY DATA ANALYSIS

Prof. GOPINATH PANDA

Date: 19 September, 2024

Assignment 1

Group-24

Submitted by:
202201522 Arnold Mochahari
202411061 Rushali Shah
202421008 Vandhana Mahajan

Missing Data Model :

- A missing data model is a conceptual framework or approach that defines how and why data is missing in a dataset. Understanding the nature and mechanism of missing data is critical because it can affect the validity of any analysis or model built using the data.

- There are three main types of missing data mechanisms:

1) MCAR (Missing Completely at Random):

➔ Data is missing completely independently of both observed and unobserved data. The likelihood of missingness is the same for all observations. Example: A survey respondent accidentally skips a question.

2) MAR (Missing at Random):

➔ The missingness depends on observed data but not on the missing values themselves. Example: Older participants in a survey are more likely to skip questions about income, but within each age group, the missingness is random.

3) MNAR (Missing Not at Random):

➔ The missingness depends on unobserved data or the values that are missing. Example: Patients with more severe conditions are less likely to report certain symptoms, leading to missing values that are correlated with the severity of the condition.

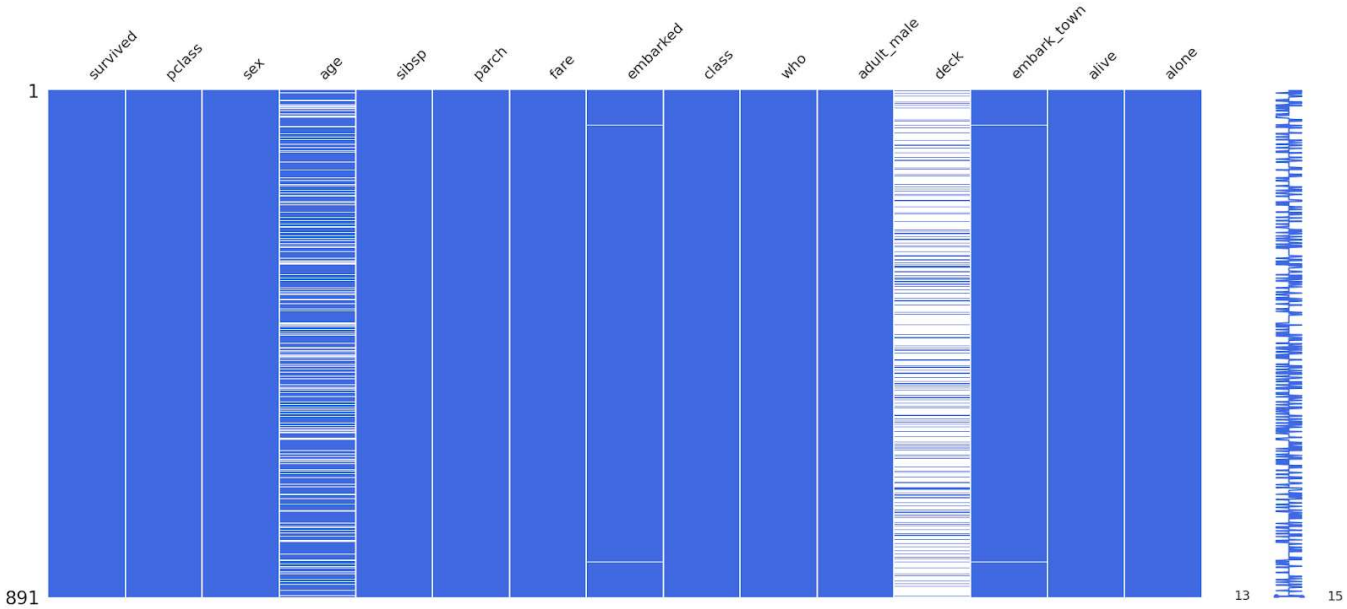
- Correctly identifying the missing data mechanism helps in choosing the appropriate imputation methods and determining the impact of missing data on analysis results.

missingno Package :

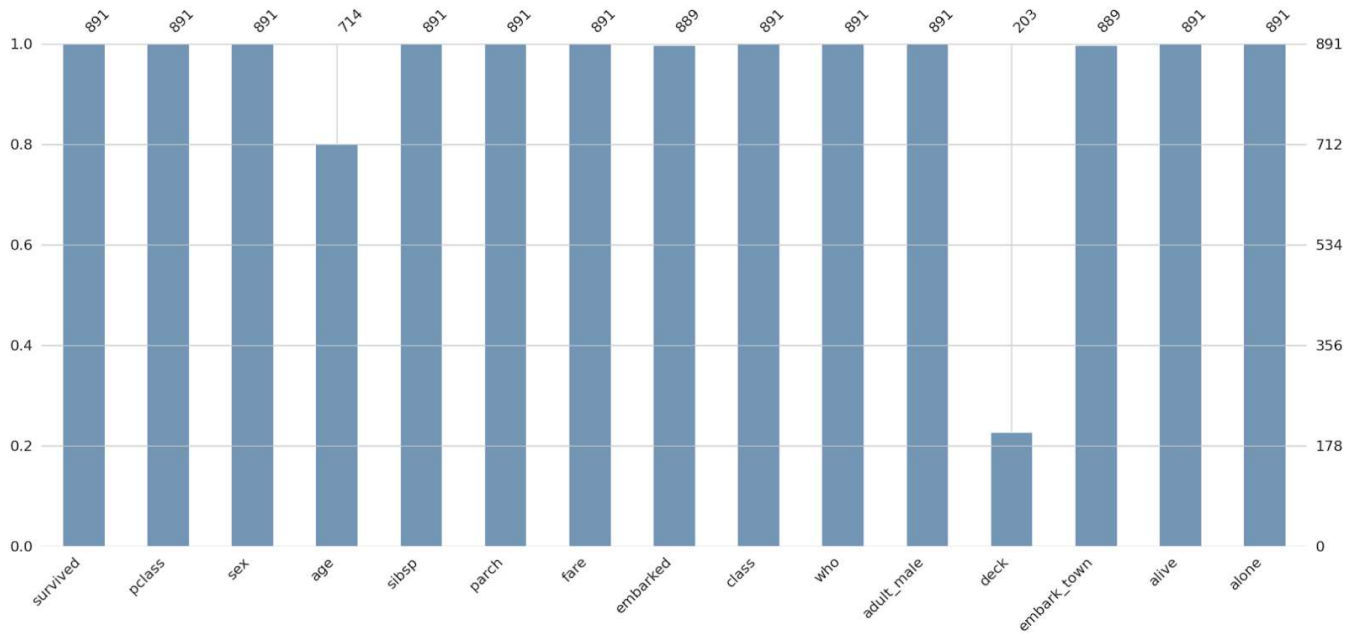
➔ The missingno package in Python provides a simple and flexible toolset for visualizing missing data within a dataset. By generating clear and concise visualizations, missingno helps us understand the structure of missing values in our data. It also helps identify patterns in missing data, which can guide your decision-making process regarding imputation or removal of missing values.

Key Functionalities of the missingno Package :

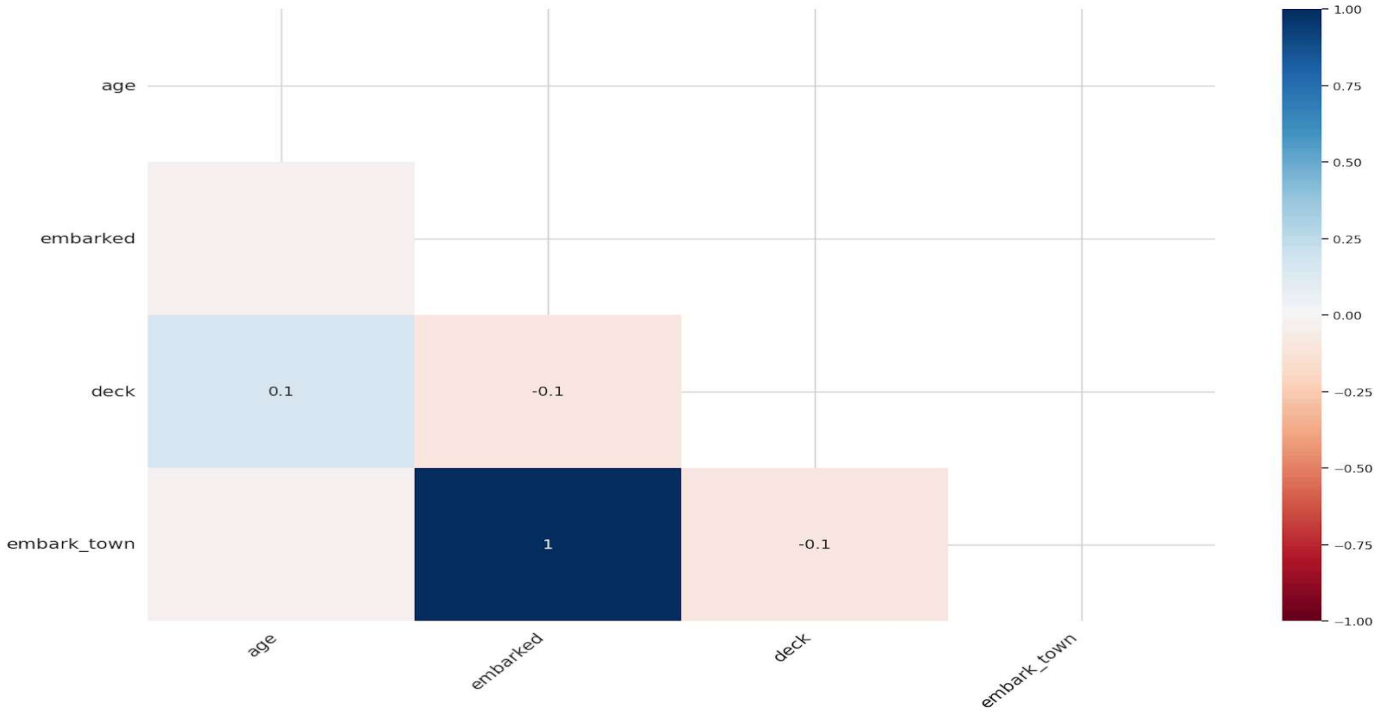
1) **missingno.matrix(df):** Provides a data matrix visualization of missing values. Shows the distribution and positions of missing values in the dataset. Allows you to quickly identify which columns have missing data and how much is missing.



2) **missingno.bar(df):** Creates a bar chart that shows the percentage of missing data in each column. Helpful for quickly assessing how much data is missing from each variable.



3) **missingno.heatmap(df):** Visualizes the correlation of missingness between columns. Shows relationships between missing data in different columns to detect patterns. The heatmap highlights variables that might have a strong correlation in terms of missing values.



4) **missingno.dendrogram(df):** Generates a dendrogram showing hierarchical clustering of columns based on their missing value structure. Helps identify groups of columns with similar missing data pattern