Exploratory Data Analysis - IT 462

Assignment - 1

Missingno Package

Group-ID: 12

Team Members:

1. Preet Shah - 202411053
2. Neerav Sharma - 202312028

Outline:

Missingno is a Python library designed to work seamlessly with Pandas. It provides an efficient way to visualize the presence of missing data in datasets. In real-world scenarios, it's quite common to encounter datasets where certain entries are absent, represented as NaN (Not a Number) values. The Missingno library helps in graphically depicting these NaN values, making it easier to understand their distribution across the dataset.

Installing Library: >> pip install missingno

Functions:

Importing library

>> import missingno as ms

1. **Barplot**
    - The Barplot function generates a bar chart displaying the amount of missing data in each column of your dataset.
    - This visualization provides an overview of data completeness by illustrating the count of non-null entries per column.
    - For example, to use the Barplot function, you would run: `ms.bar(dataFrame)`
2. **Matrixplot**
    - Matrixplot offers a visual representation of missing data within a matrix format.
    - It shows the distribution of missing values across both rows and columns, which helps in detecting any patterns or clusters of missing data.
    - Missing values are represented in white, making them easy to spot. A sparline on the side summarizes the overall data completeness and highlights rows with minimal missing values.
    - You can create a Matrixplot with: `ms.matrix(dataFrame)`
3. **Heatmap**
    - The Heatmap function reveals the correlation between missing values across different columns.
    - It highlights the relationships between missing data in various columns: a value close to +1 suggests a strong correlation (i.e., missing data in one column is likely to occur in another), while a value near -1 indicates an inverse relationship. A value near 0 means there is minimal to no correlation.
    - To generate a Heatmap, use: `ms.heatmap(dataFrame)`
4. **Dendrogram**
    - The Dendrogram creates a tree-like diagram through hierarchical clustering, grouping columns that exhibit similar missing data patterns.
    - This visualization is particularly useful for identifying columns with similar patterns of missing data and can assist in deciding on potential imputation strategies.
    - To produce a Dendrogram, execute: `ms.dendrogram(dataFrame)`

Advantages:

- Facilitates a rapid assessment of both the extent and distribution of missing data within a dataset.
- Simplifies the understanding of how missing data is related, which can guide decisions on data imputation.
- Helps uncover patterns of missing data that might suggest systemic issues in data collection.
- Useful for tracking trends in missing data over time in time-series datasets.

Disadvantages:

- For exceptionally large datasets, visualizations from Missingno can become slow or unresponsive, especially for the `ms.matrix` and `ms.heatmap` functions.
- The library does not provide insights into the relationships between non-missing data, which could be crucial for a comprehensive data analysis.
- Offers limited customization options for the visualizations, which might restrict the ability to tailor plots to specific needs.