Exploratory Data Analysis - IT 462

Assignment - 1

Missingno Package

Group-ID: 17

Team Members:

1. Manish Prajapati - 202411012
2. Yash Mehta - 202201309
3. Bhavish Hiranandani - 202312070

Outline:

Missingno is a Python Library and compatible with Pandas. Missingno Library offers a good way to visualize the distribution of NaN values. In the case of a real-world dataset, it is very common that some values in the dataset are missing. We represent these missing values as NaN (Not a Number) values. Missingno library is used for visualizing the same missing values in the dataset.

Installing Library:

>> pip install missingno

Functions:

Importing library

>> import missingno as ms

1. barplot

- Provides a bar chart of missing data by column.
- Shows how many data points are non-null per column and gives a quick overview of data completeness.
- E.g., ms.bar(dataFrame)

2. Matrix plot

- Displays a matrix visualization of missing data.
- Highlights the distribution of missing values across rows and columns, helping to detect patterns or clustering of missing data.
- The plot appears white wherever there are missing values.
- The **sparline** on the right gives an idea of the general *shape of the completeness* of the data and points out the row with the minimum nullities and the total number of columns in a given dataset, at the bottom.
- E.g., ms.matrix(dataFrame)

3. Heatmap

- Displays a heatmap of nullity correlations (missing data correlations) between columns. E.g., ms.heatmap(dataFrame)
- Helps to identify relationships between missing data in different columns.
- Values close to +1 indicate that the presence of null values in one column is correlated with the presence of null values in another column.
- Values close to -1 indicate the presence of null values in one column is anti-correlated with the presence of null values in another column.
- Values close to 0 indicates that there is little or no relationship between the presence of the null values in one column compared to another.

4. **Dendrogram**

- The dendrogram plot provides a tree-like graph through hierarchical clustering and groups together columns that have strong correlations in nullity.
- Useful for identifying columns with similar missing data patterns and for potential imputation strategies.
- E.g., ms.dendrogram(dataFrame)

Advantages

- Quickly assess the extent and pattern of missing data in a dataset.
- Ease in understanding the missing data relationships to guide imputation decisions.
- Ease in identifying the patterns of missing data that might indicate systematic collection problems.
- Helpful in analyzing the missing data trends over time in time-series datasets.

Disadvantages

- For very large datasets, missingno visualizations can become slow or unresponsive, especially for functions like ms.matrix and ms.heatmap
- Inability to visualize Non-missing Data Relationships, because the missingno only focuses on missing data, meaning it doesn’t provide insights into relationships between non-missing values, which can be important for data analysis.
- Limited customization of plots provided.

<><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><>
<><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><>