Dhirubhai Ambani
Institute of Information and Communication Technology

IT462: Exploratory Data Analysis

Group - 11

Assignment 1: Missingno Package

Dataset: Flight Delay and Cancellation

Group Details	
Name	Student ID
Akshat Kadia	202203029
Gunjan Sethi	202312112
Jaydeep Darji	202411029

Contents

1 Introduction

1.1 Purpose of Missing Data Visualization

Missing data can often be overlooked, but it's essential to identify it early in the analysis phase. Incomplete data can bias results, cause algorithms to fail, or reduce the accuracy of predictive models. The missingno package is designed to give a clear view of how much data is missing and its distribution within the dataset.

1.2 Why Use Missingno

The missingno package is used because it provides clear visualizations to identify and analyze missing data patterns in a dataset quickly. It helps understand the extent, distribution, and correlations of missing values, making it easier to decide on the best data-cleaning strategies.

2 Visualization of Missing Data

2.1 Bar plot:

- Each bar represents a column in your dataset.
- The height of the bar corresponds to the number of non-missing (i.e., present) data points in that column.
- The shorter the bar, the more missing values exist in that column.
- The Y-axis typically shows the count of non-missing values (i.e., how many values are available in the dataset for each column).
- Example, If the CANCELLATION_CODE column has a short bar, it indicates that the column is largely missing, perhaps because not many flights were canceled.

2.2 Matrix Plot:

2.2.1 Missing Data Pattern:

- Each row in the matrix plot represents a record (observation) in your dataset.
- White lines indicate missing values. Black or dark-colored lines represent non-missing values (or valid data).

2.2.2 Sparkline:

- To the right of the plot, there's a sparkline that summarizes the completeness of the dataset.
- The sparkline gives a condensed view, showing how many rows are fully populated (i.e., without missing values) and how many are partially complete.
- A more filled-in sparkline means that the dataset is relatively complete, while a broken sparkline suggests a higher percentage of missing values.
- You may observe that certain columns consistently have missing data across several rows. For example, if a column consistently has white lines in many rows, it might indicate that this feature is systematically missing and requires further analysis.
- Example, Flights that were canceled tend to have missing values in the columns related to delays (like ARR_DELAY or DEP_DELAY), because these delays are not recorded for canceled flights.

2.3 HeatMap:

- Heatmaps are a powerful way to visualize data relationships, trends, and missing data patterns. When analyzing flight delay and cancellation data, using a heatmap can help us to identify correlations between variables like delays, cancellations, or other operational metrics. This visualization helps you understand the structure and potential issues in your dataset quickly and clearly.

2.3.1 Correlation Calculation:

- The function calculates the correlation between the presence or absence of missing values in different columns. This correlation indicates how likely it is for a missing value in one column to coincide with a missing value in another.
- In a correlation heatmap, areas with dark colors indicate strong correlations, while lighter colors indicate weaker correlations.

2.3.2 Color-Coding:

- The core feature of a heatmap is the use of color to represent the magnitude of values in a dataset. Each cell in the heatmap corresponds to a pair of variables or features, and the color of the cell reflects the strength and direction of the relationship between them (e.g., correlation). In missing data visualization, colors typically represent the degree of missingness or correlation between missing values across features.
- Typically, a color scale is used to represent different ranges of values. For example:
  - Darker colors might represent higher values.
  - Lighter colors might represent lower values.
  - In some cases, a gradient from one color to another (e.g., blue to red) is used.

2.3.3 Axes and Cells:

- Heatmaps are often two-dimensional. The x-axis and y-axis represent two different variables or dimensions of your dataset.
- Each intersection of an x and y value forms a cell, which is color-coded based on the corresponding value from the data matrix.

2.3.4 Color Scales:

- Heatmaps use a continuous color scale or discrete colors.
- Continuous color scales (gradients from light to dark) are typically used for numerical data.
- Discrete colors (different categories represented by distinct colors) may be used for categorical data.

2.3.5 Finding Outliers:

- Outliers in your data often stands out as highly contrasting colors compared to surrounding cells.

2.4 Dendogram

- A dendrogram is a tree-like diagram that shows the hierarchical relationship between variables. It's often used in cluster analysis to visualize the arrangement of clusters formed by hierarchical clustering algorithms.

- The clusters represent groups of columns that have similar patterns of missing values.

2.4.1 Hierarchical Clustering:

- The function applies hierarchical clustering based on the presence or absence of missing values in the columns of your DataFrame.

- Columns with similar missing value patterns are grouped together, and the structure of the clustering is displayed as a dendrogram.

2.4.2 Branch Length:

- Shorter branches between two columns indicate a higher similarity in their missing data patterns (i.e., they often have missing values together).

- Longer branches indicate a lower similarity or more random missingness between columns.

2.4.3 Handling Missing Data:

- The dendrogram helps you visually detect relationships between variables.

- You can use the clustering information to guide how you handle missing data. If certain groups of columns have missing data together, you might want to impute them together or drop them if the data is systematically missing.

3 Interpreting the Visualizations

- From the visualization we decide that our data contains which type of missing values, MAR, MCAR, or MNAR.

- Example, If the matrix plot shows a pattern where missing values are concentrated around certain rows, this may indicate data collection issues or specific conditions that led to missing entries.

3.1 MCAR:

- Data is missing completely at random, with no relationship between the missing data and any other variables in the dataset.

- If missing values appear randomly distributed across rows and columns in a msno.matrix() or msno.bar() plot with no clear pattern, it could indicate MCAR.

- No correlation should be visible between missing data in the msno.heatmap().

3.2 MAR:

- Missing data is related to some observed variables but not the missing variable itself(i.e. Depends on other columns).
- If certain columns or rows have consistent patterns of missing data in the msno.matrix() plot, it may suggest MAR.
- In the msno.heatmap(), if the plot shows a correlation between missingness in different columns, it could imply that the missingness is dependent on other observed variables (MAR).

3.3 MNAR:

- Missing data depends on the value of the missing variable itself, meaning the reason for missingness is directly related to the unobserved values.
- Patterns that ignore correlation or relationships with other variables in the msno.heatmap() may suggest MNAR.

4 Conclusion

A missingno library is a valuable tool for visualizing missing data in datasets. It offers various plots, such as bar plots, matrix plots, heatmaps, and dendrograms, that provide intuitive insights into the distribution and patterns of missing values. These visualizations help detect whether missing data occurs randomly or follows a pattern, which can guide us in determining the type of missing data (MCAR, MAR, or MNAR). By quickly identifying missing data issues, missingno assists in making informed decisions about handling data gaps, ensuring cleaner and more reliable analysis.