NULL VALUE MATRIX REPORT

ON HCES 2023-24

Author: Rudra Prasad Bhuyan
Position: Data Scientist Intern

Organisation: SBC Labs
Guidance: Anushka Ashok, Alok Gangaramany

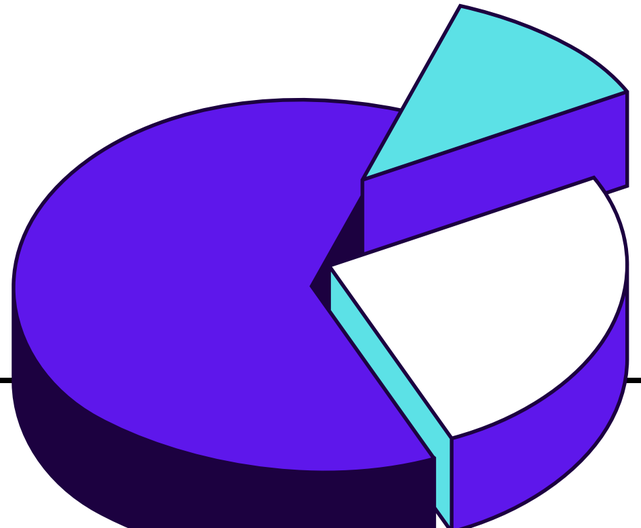
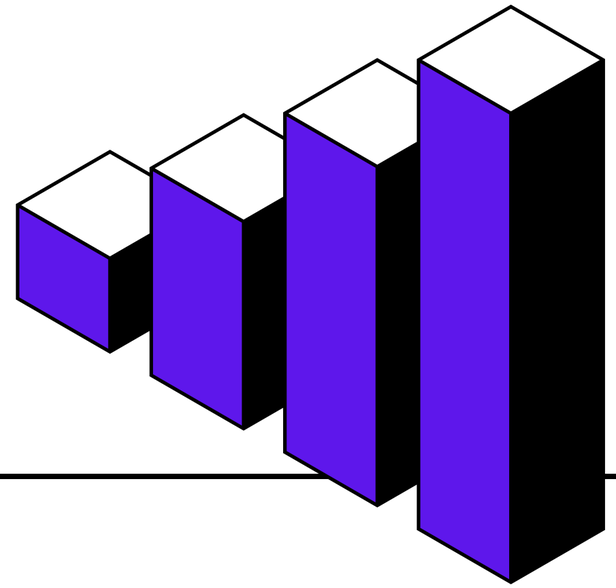


Table of Contents

Null Value Analysis

- Survey/Household
- Person
- Socio-Economic
- Ration/Online
- Food Consumption
- Consumption Value
- Benefits/Welfare
- Non-Food Items
- Non-Durabels
- Services/Utillities
- Assets/Possessions
- Clothing/Footwear
- Durable Goods
- Total Expenditure
- Summary/ Metadata

Filling Method Suggestions

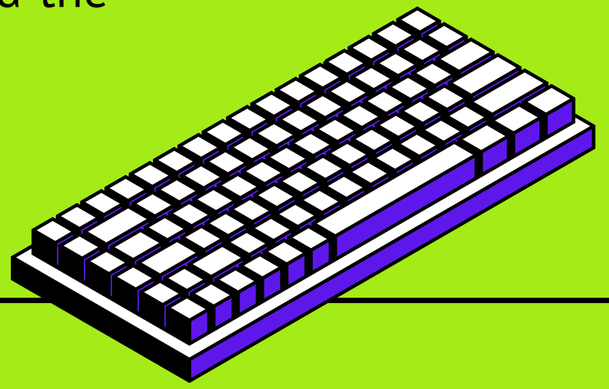


Introduction

- After encountering a large number of null values in the dataset, we decided that since this dataset is very important, we cannot compromise on data quality. Because most of the features (or columns) are checkbox-type, we can fill the missing values with null to maintain consistency.

- We can then perform feature engineering on the remaining features to achieve the desired output.

- In this report, I will present the missing values, the level names, and the suggested methods for filling in the missing data.



NULL Value Analysis

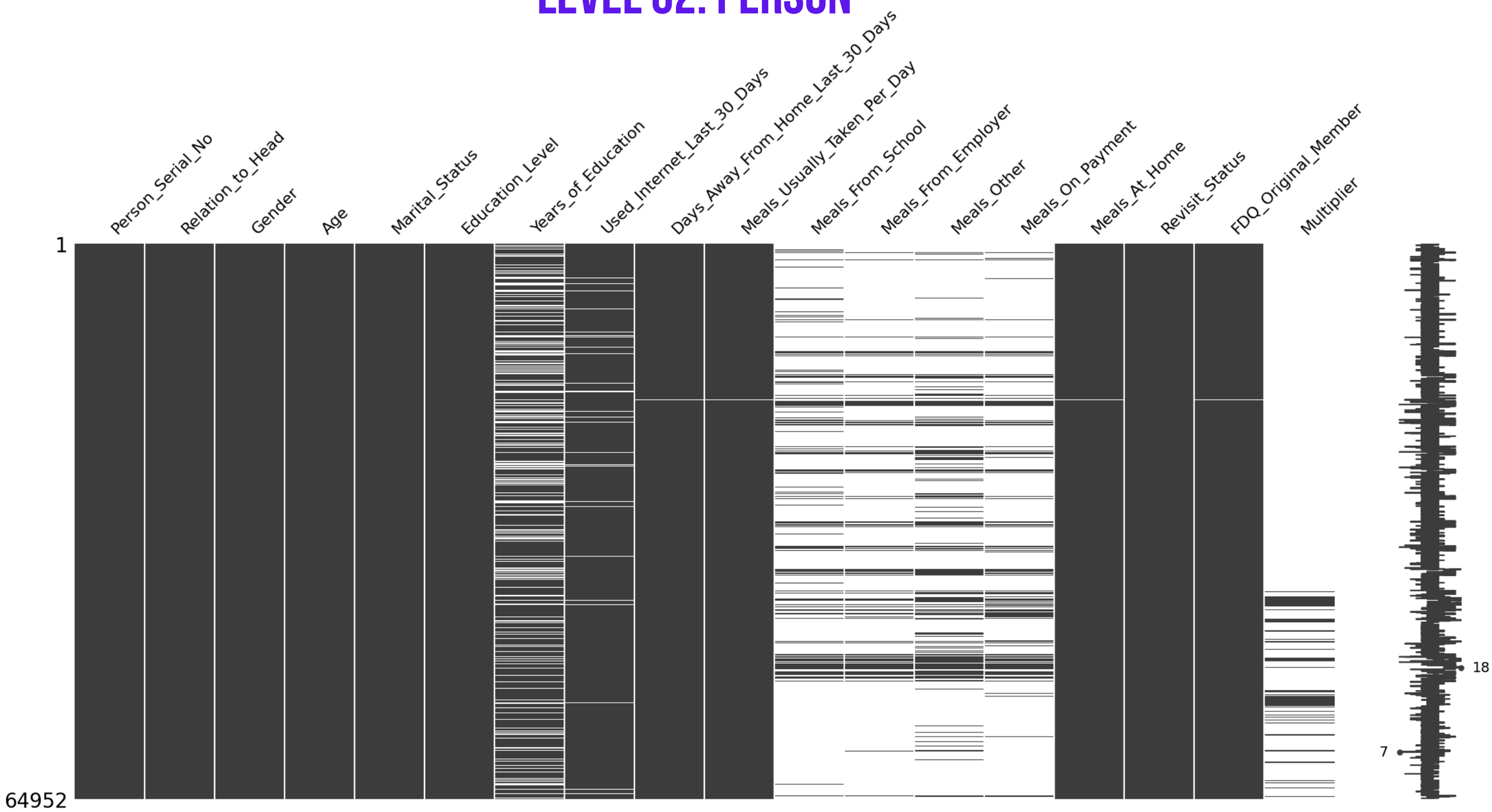
- I used Polars, and after filtering the data for the MP state (labeled as 23), I cast the selected features to Int16 to save storage space.

- I also filtered out the useful features to reduce computation time when generating the correlation plot, and calculated the percentage of null values using the Polars Python library.

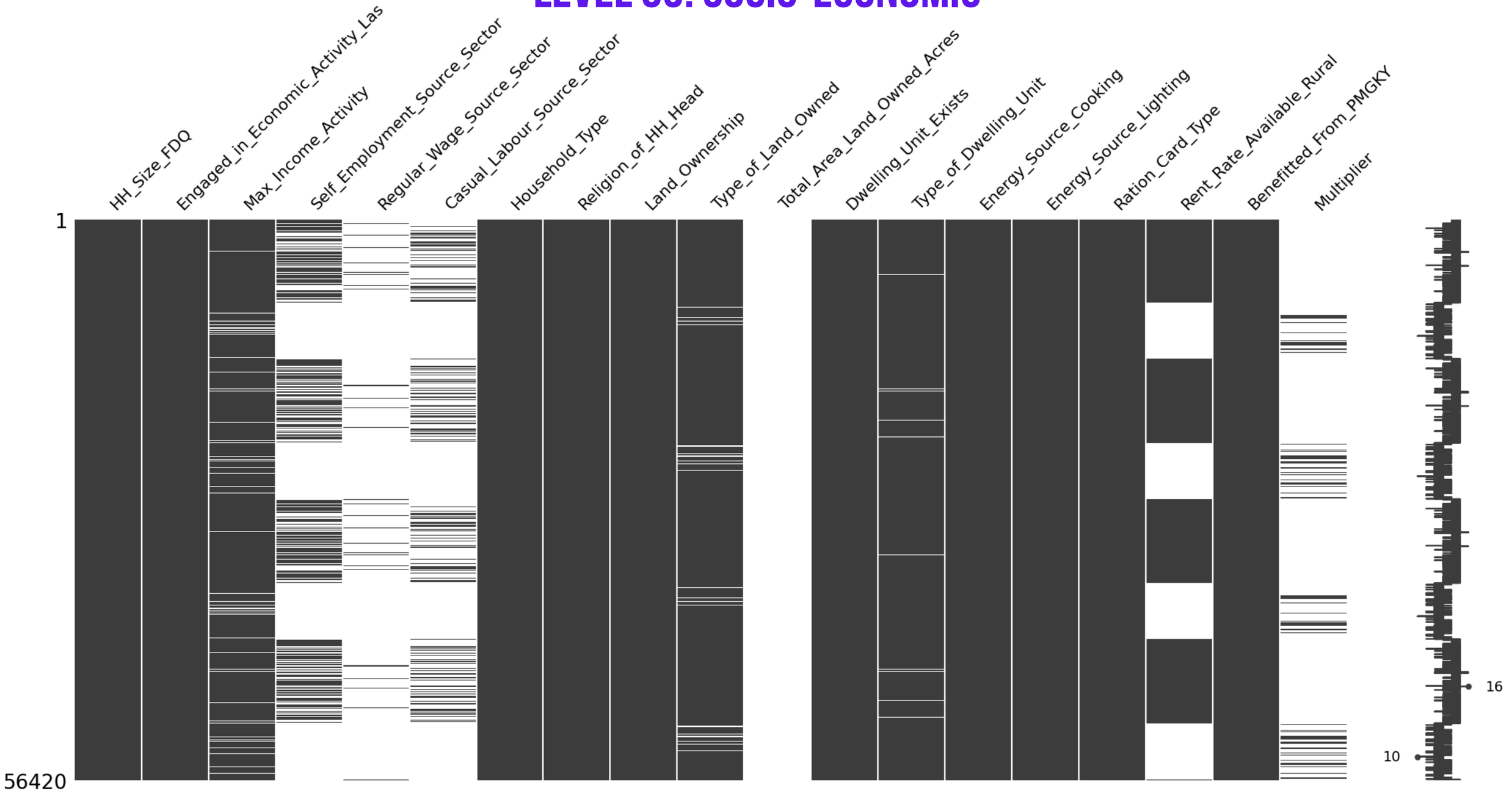
- After completing the state and feature selection, as well as the data type conversion, I converted the dataset into a Pandas DataFrame. Then, I used the Missingno Python library to visualize the missing values.

LEVEL 01: SURVERY/HOUSEHOLD

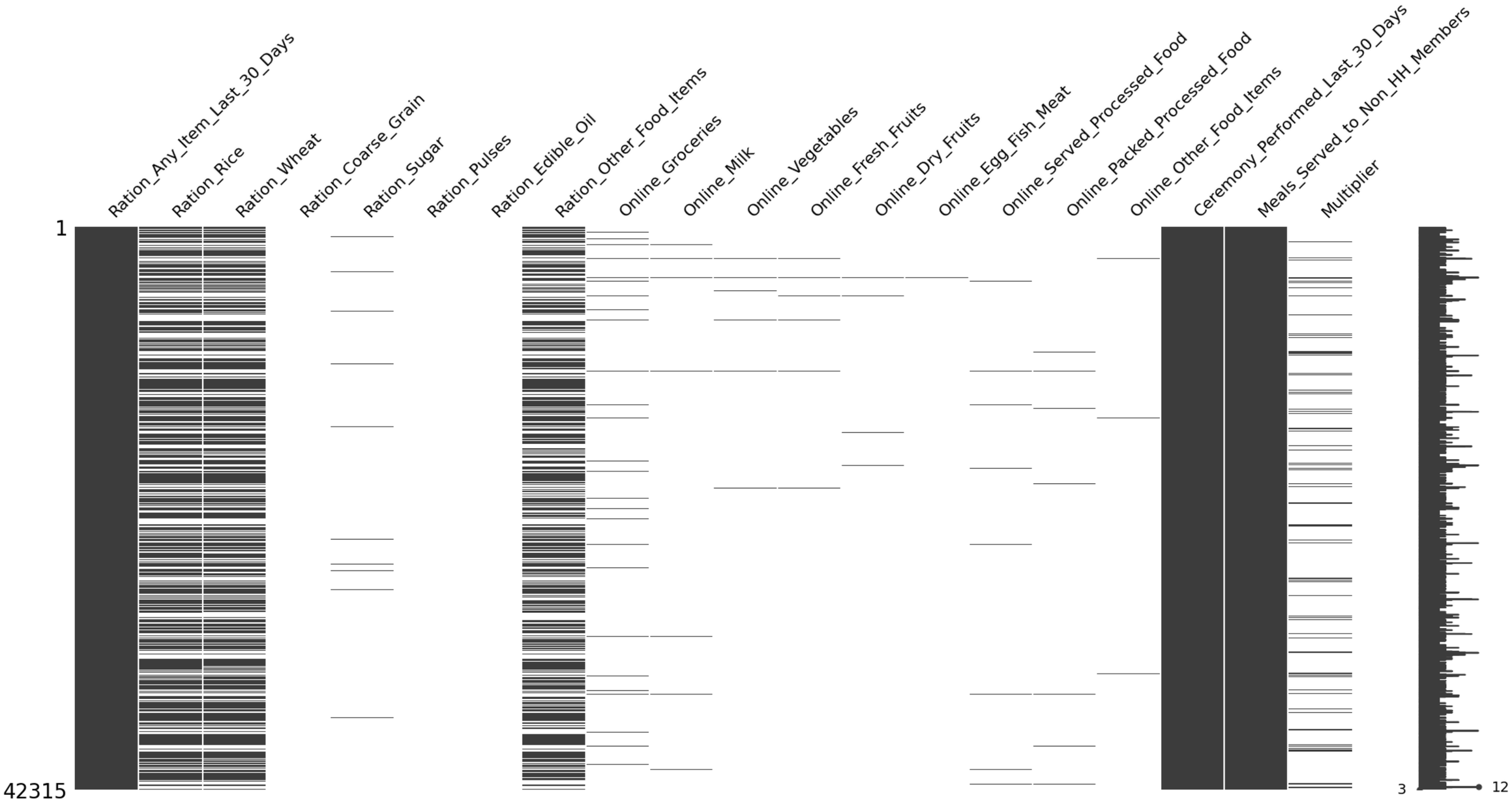
LEVEL 02: PERSON



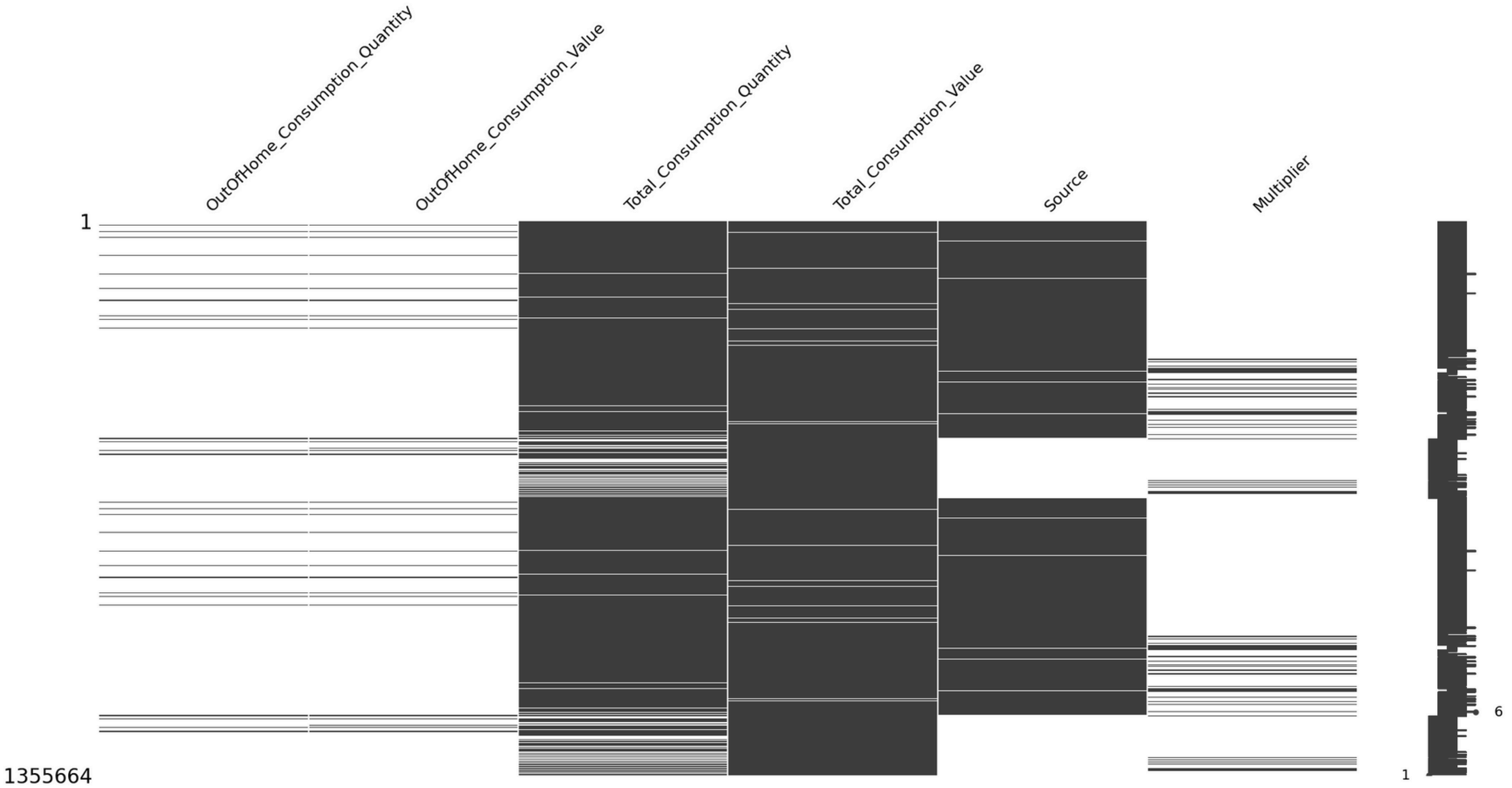
LEVEL 03: SOCIO-ECONOMIC



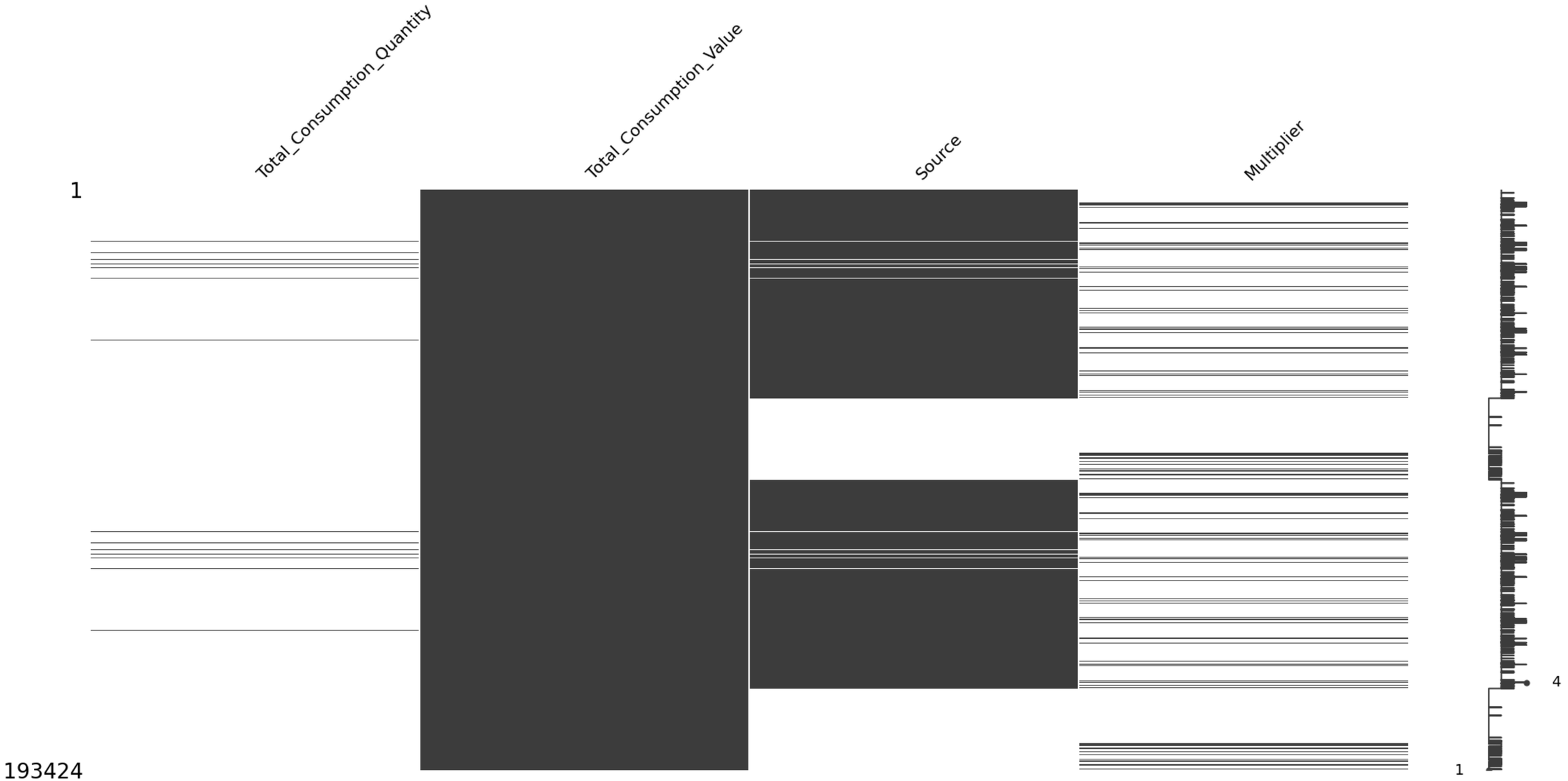
LEVEL 04: RATION/ ONLINE



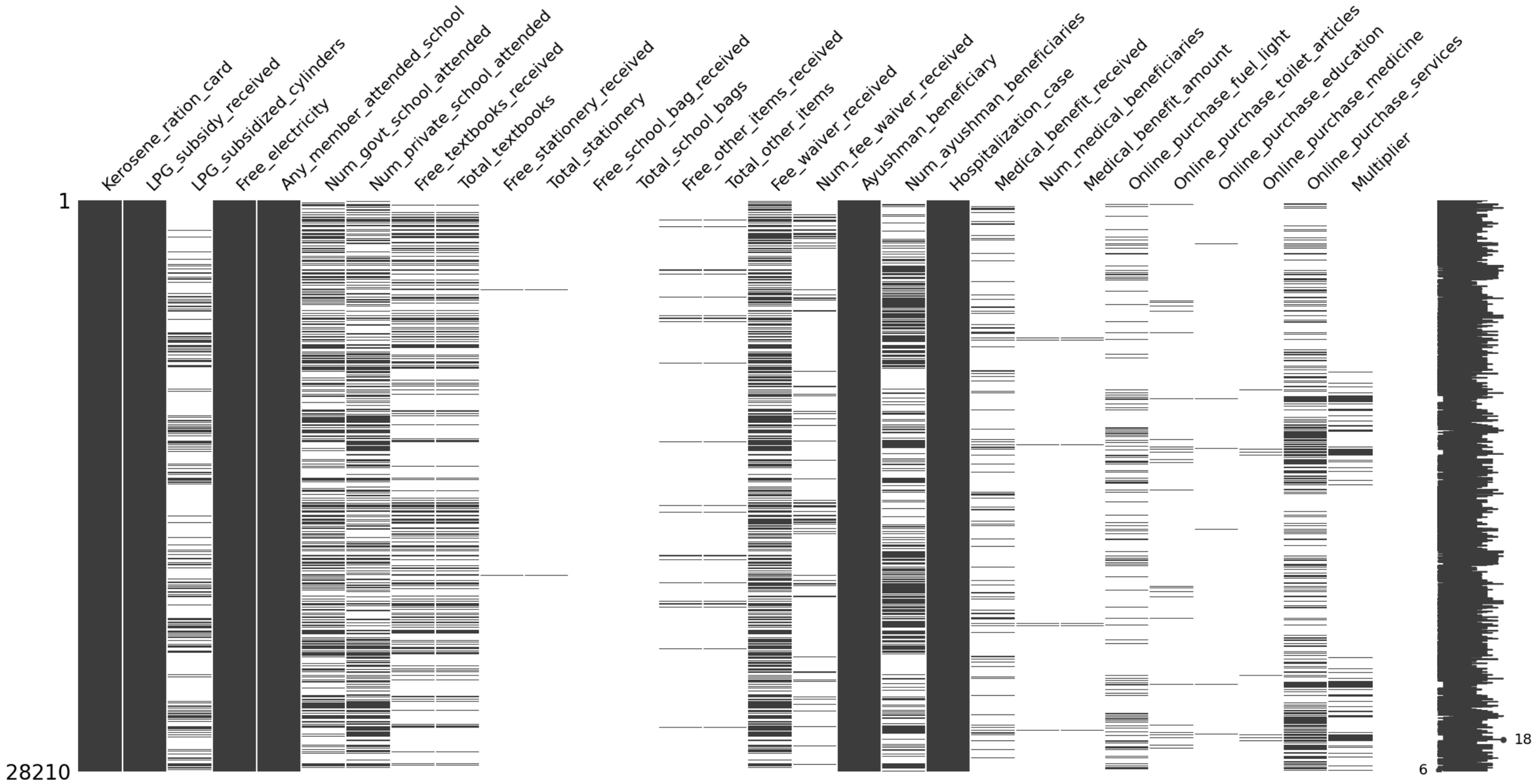
LEVEL 05: FOOD CONSUMPITION



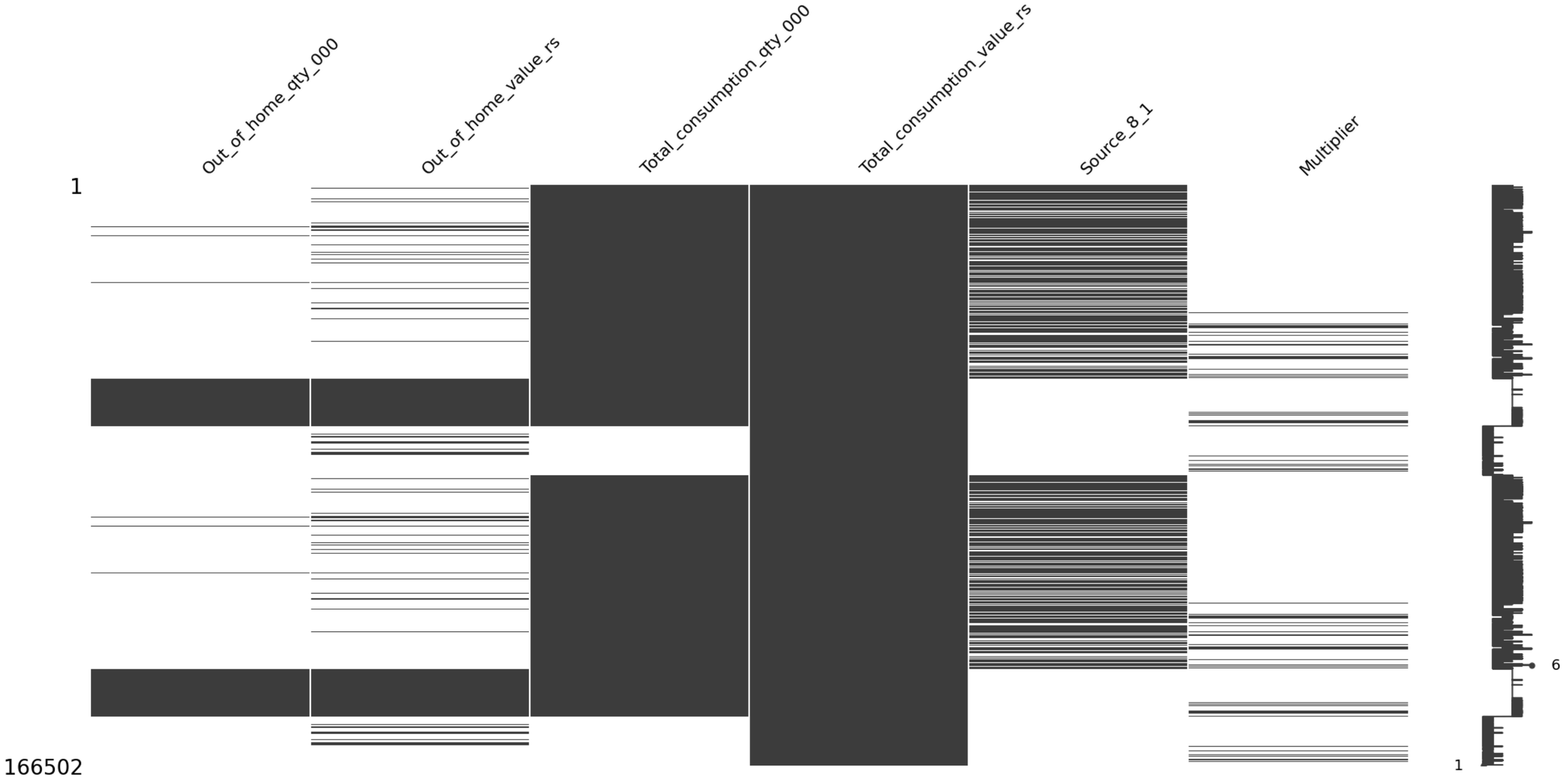
LEVEL 06: CONSUMPTION VALUE



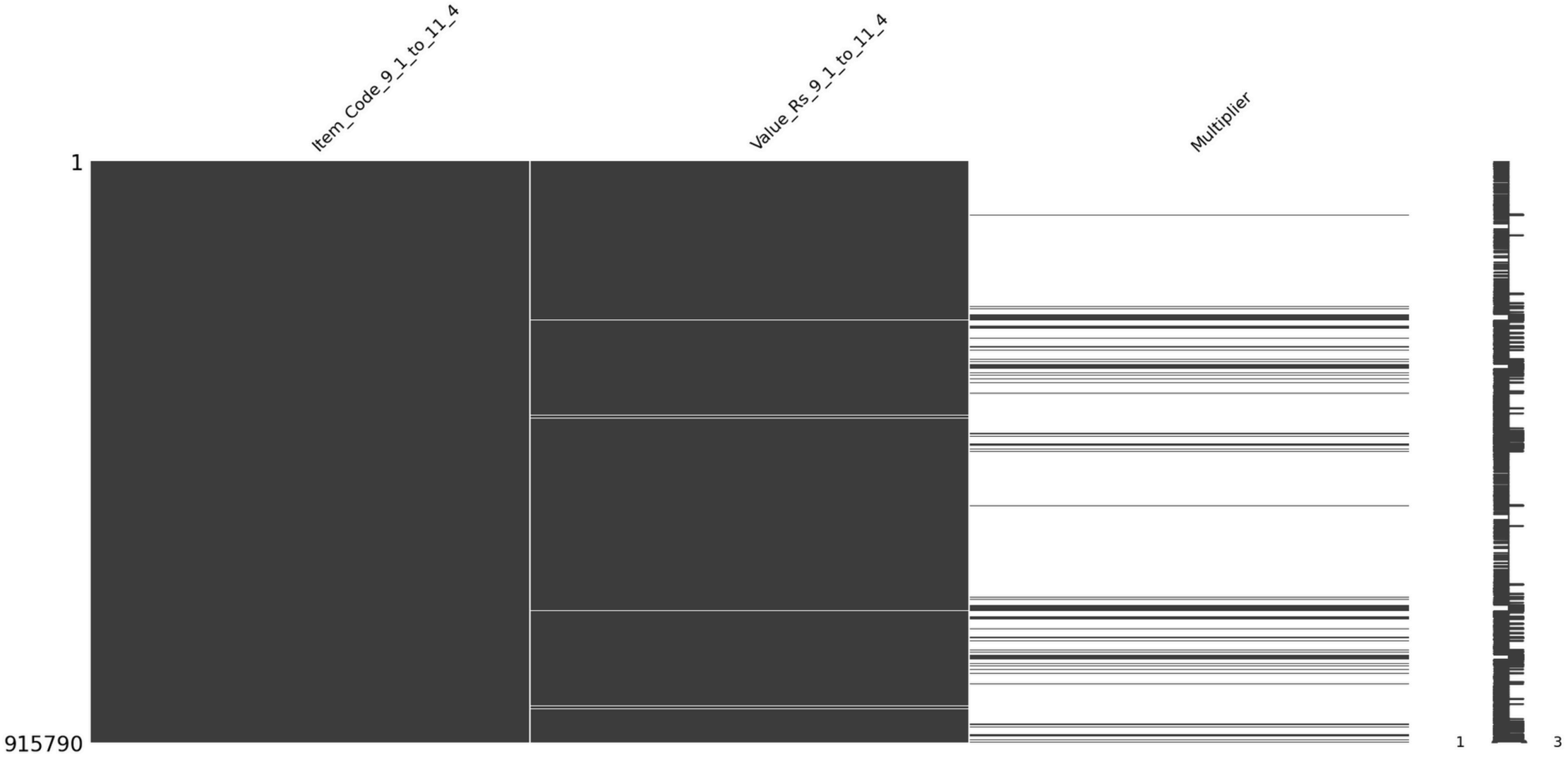
LEVEL 07: BENEFITS/ WALFARE



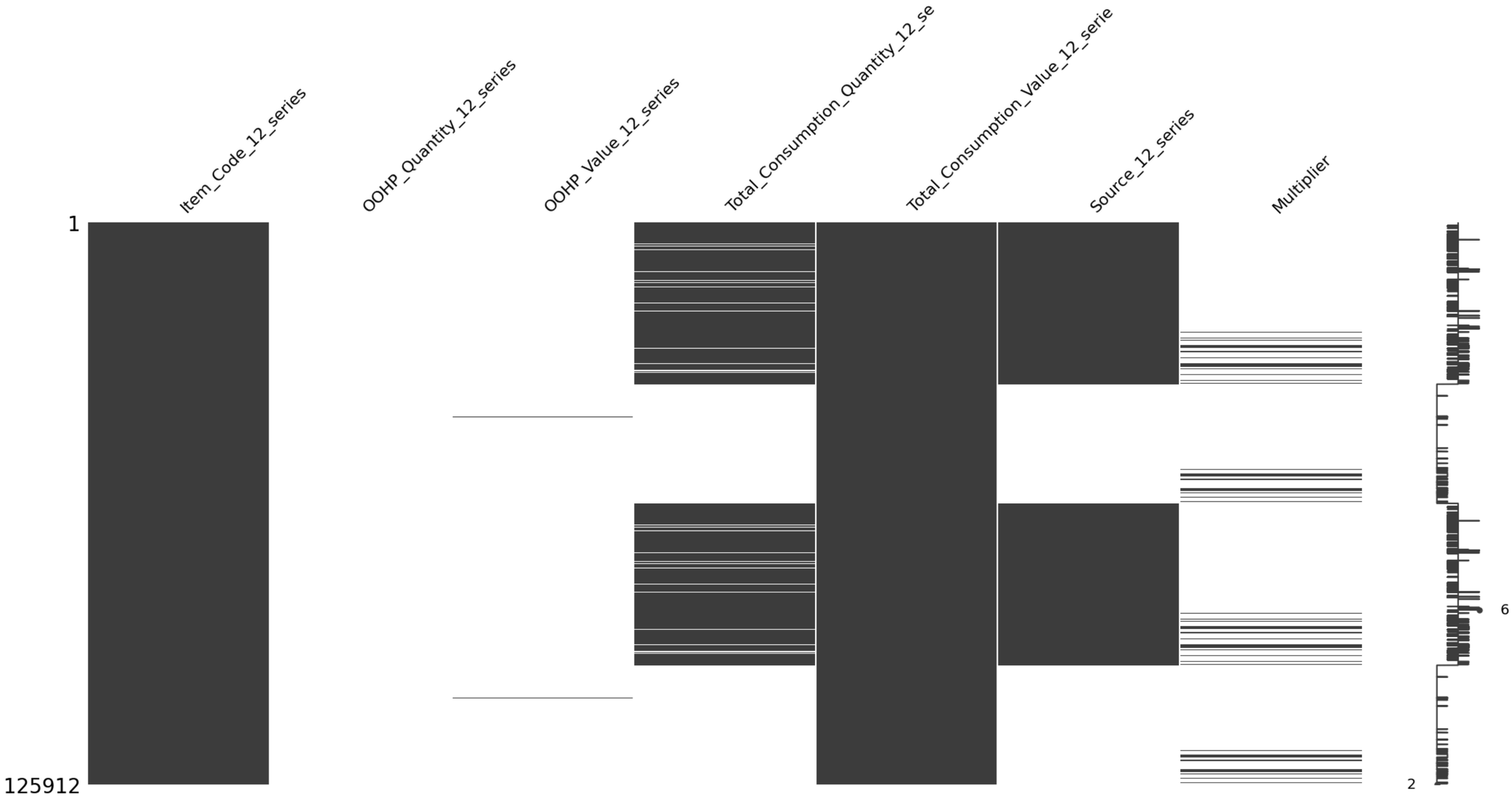
LEVEL 08: NON-FOOD ITEMS



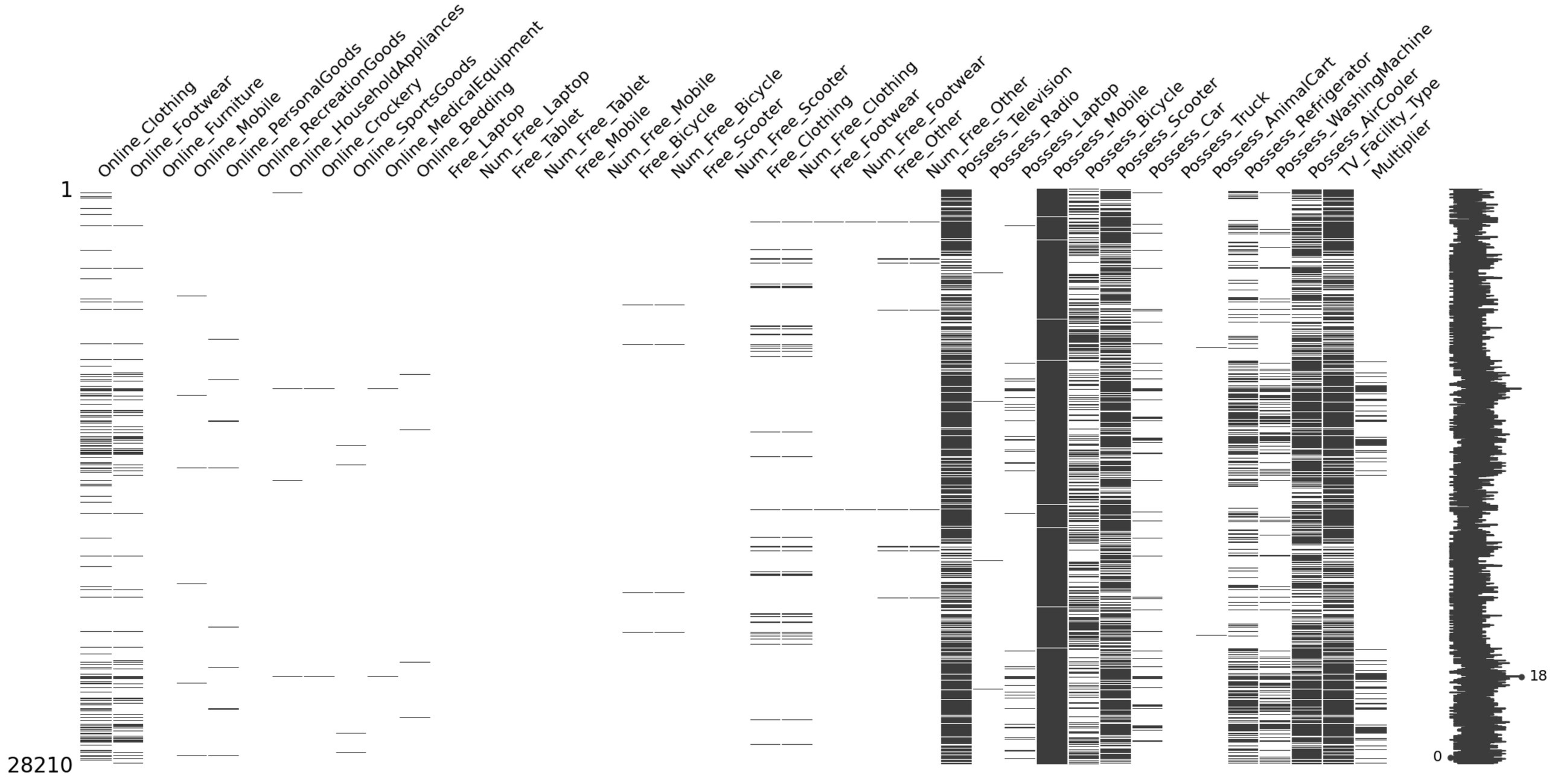
LEVEL 09: NON-DURABLES



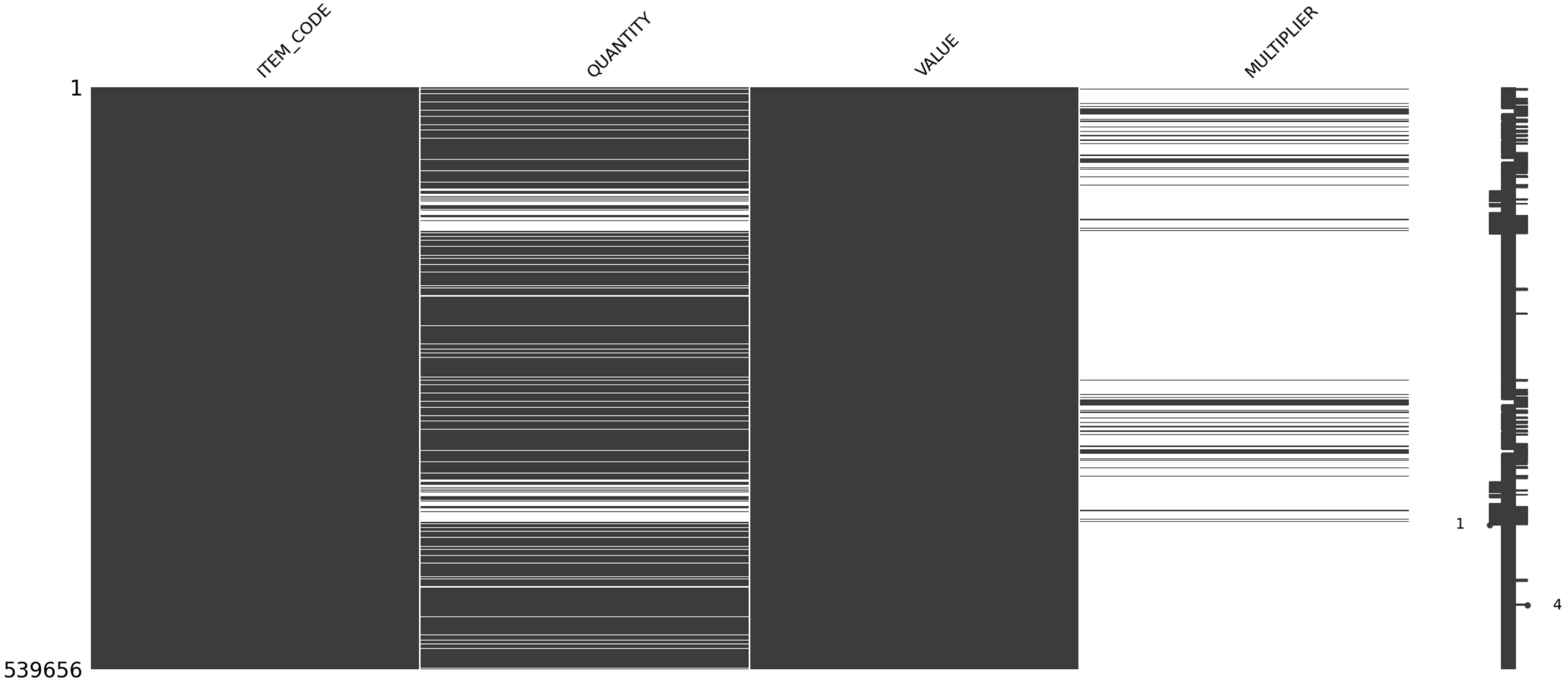
LEVEL 10: SERVICES/UTILLITIES



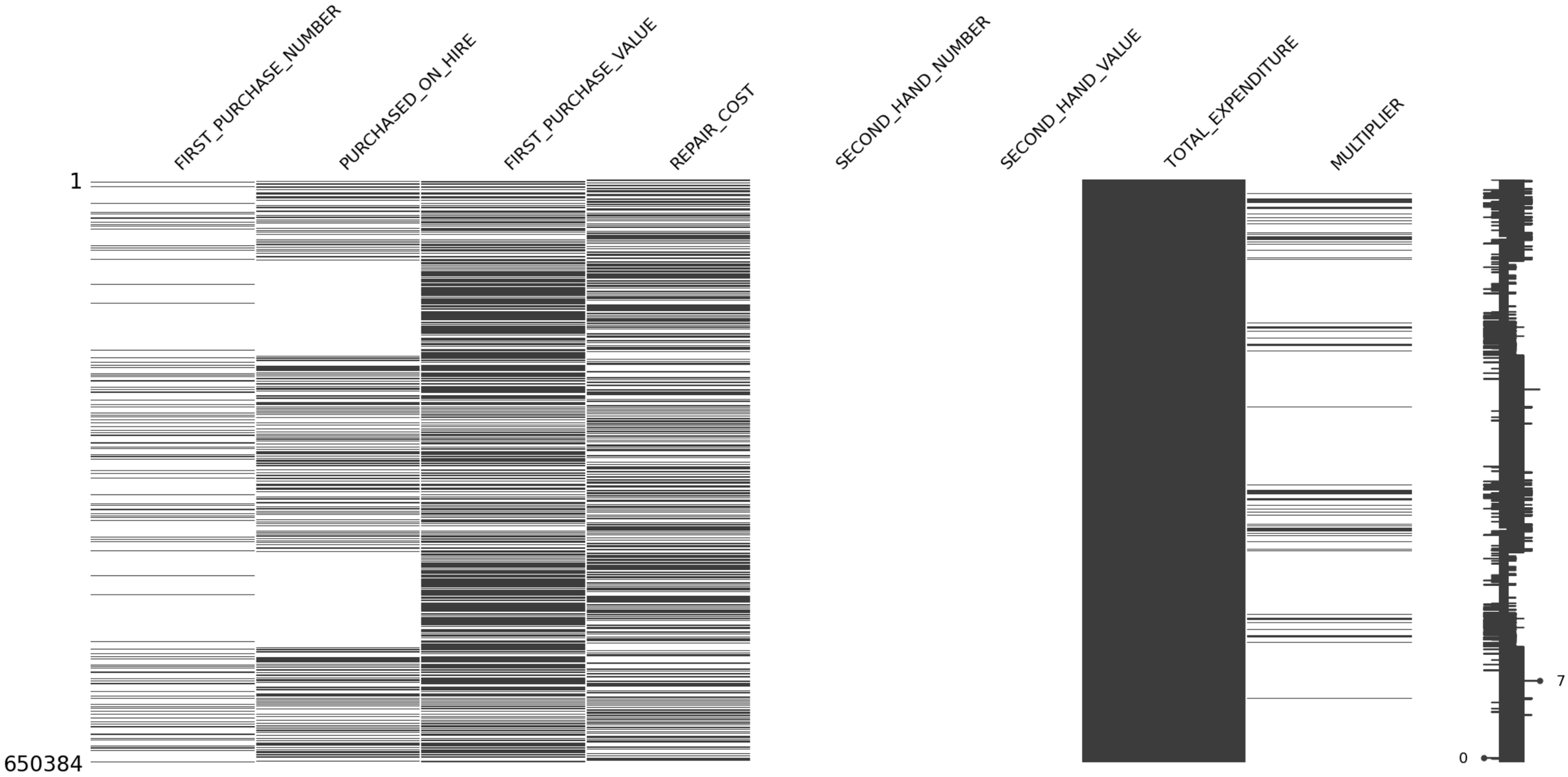
LEVEL 11: ASSETS/POSSESSIONS



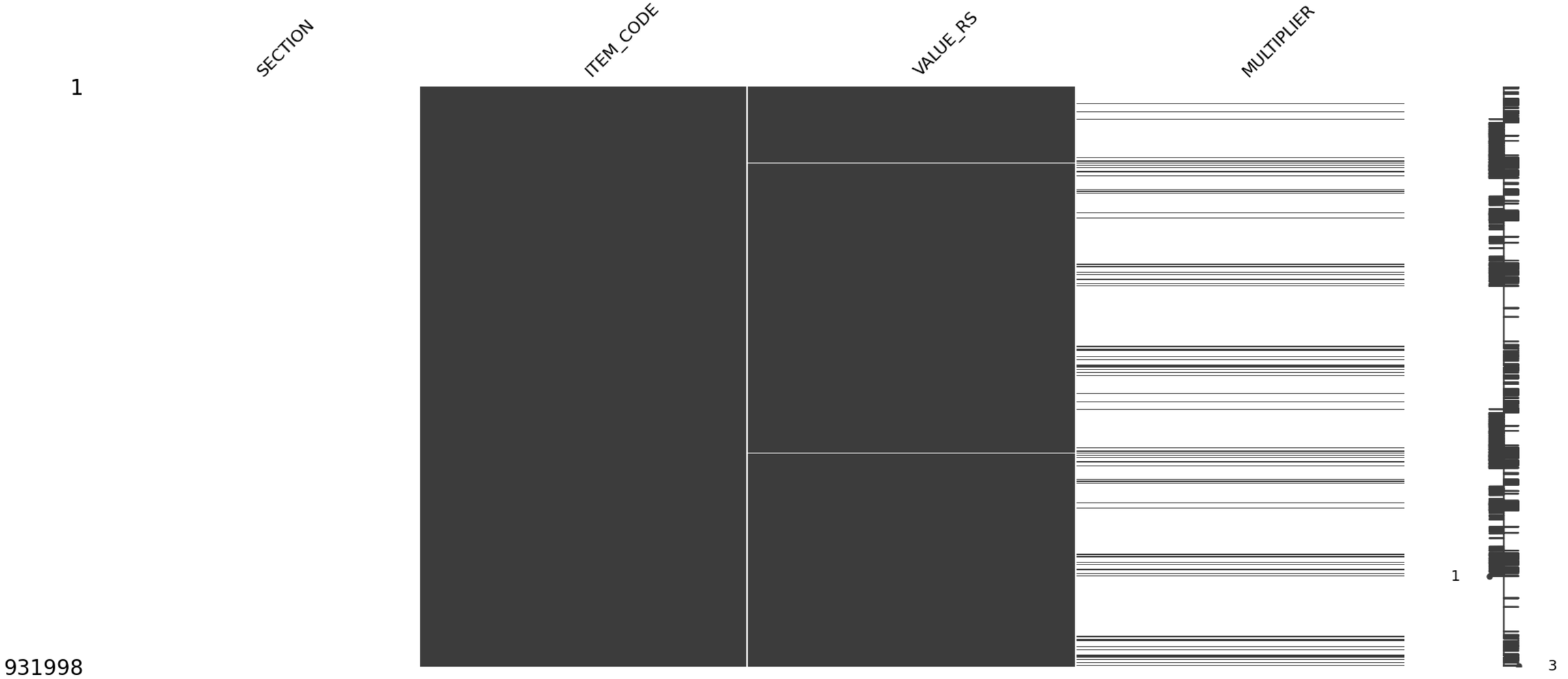
LEVEL 12: CLOTHING/FOOTWEAR



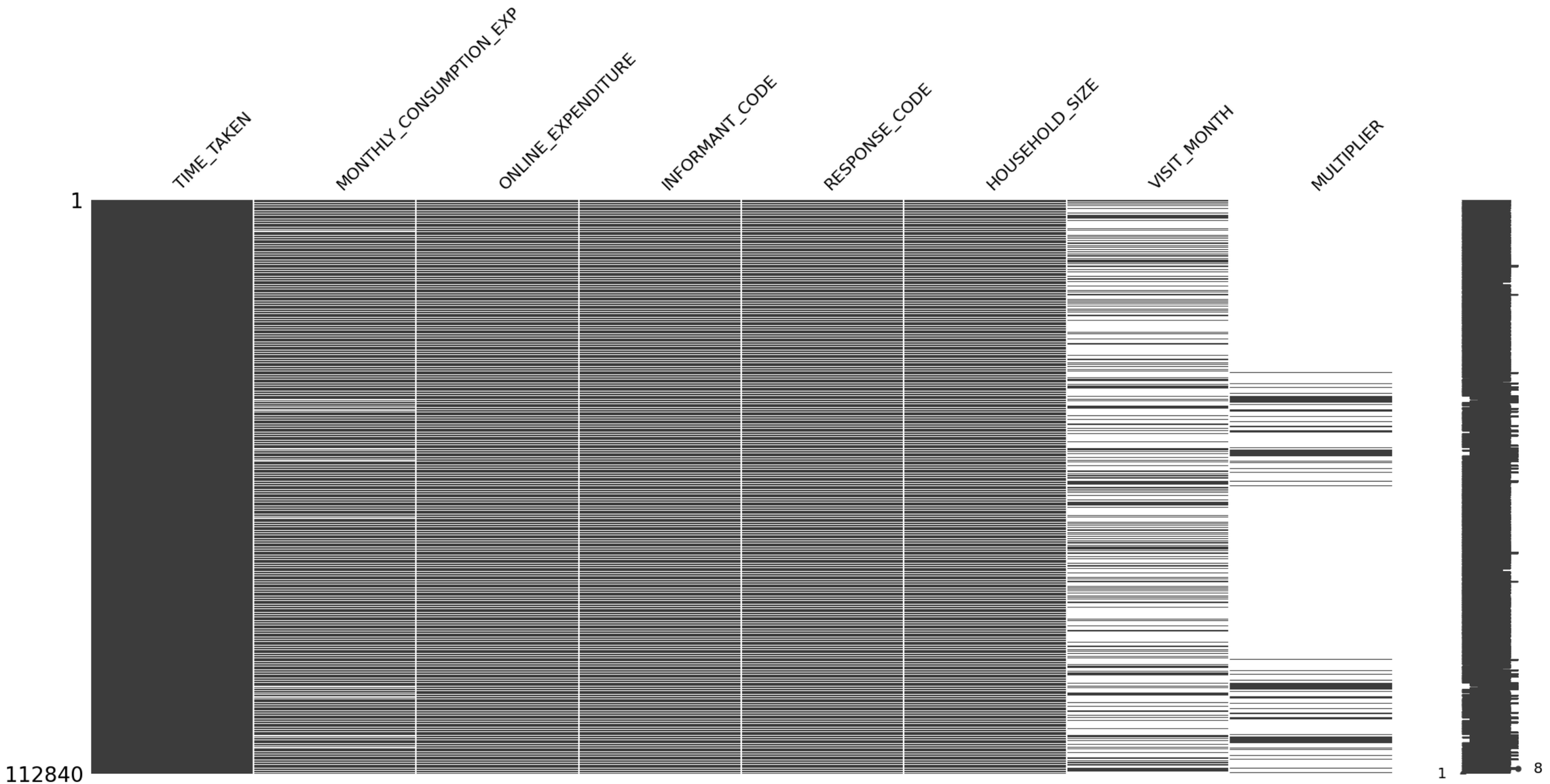
LEVEL 13: DURABLE GOODS



LEVEL 14: TOTAL EXPENDITURE



LEVEL 15: SUMMARY/ METADATA



Filling Method Suggestions

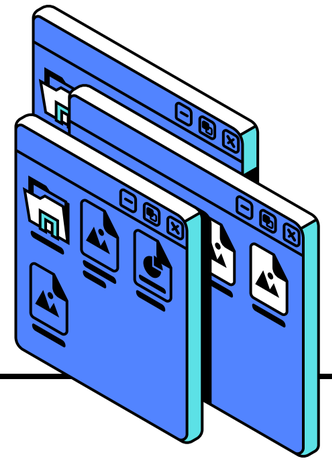
- If these features are checkboxes in the survey, fill missing values with 0 (No).
- If these features are categorical, create bins or categories as needed.
- If these features are numerical, handle missing values using methods such as KNN imputation, mean/median imputation, dropping rows, or iterative imputation.

Source: https://github.com/Rudra-G-23/Play-with-Data/blob/main/Data_Accessing_%26_Cleaning/Handling%20Missing%20Values%20Chart.md

Filling Method Suggestions

Method	When to Use	Pros	Cons
KNN Imputer	<30% missing, correlated features	Preserves relationships	Slow for many features
Iterative Imputer (MICE)	<50% missing	Captures multivariate patterns	Computationally heavier
Model-based Imputation (Random Forest / XGBoost)	Important columns with 40–70% missing	High accuracy	Requires training per column
Median/Mode per Group	When natural grouping exists	Keeps context	Simpler but fast
Drop Columns	>70% missing	Keeps dataset clean	Loss of info

Source: https://github.com/Rudra-G-23/Play-with-Data/blob/main/Data_Accessing_%26_Cleaning/Handling%20Missing%20Values%20Chart.md



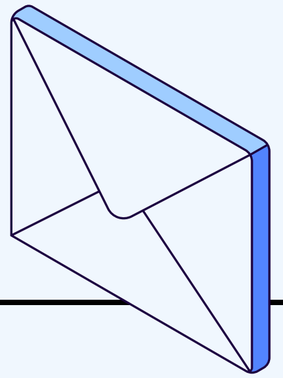
Conclusion

- After analyzing the dataset, it was observed that several features contained missing (null) values that could potentially affect the accuracy and reliability of the model. To address this issue, appropriate imputation techniques were suggested based on the nature of each feature.

- For checkbox-type survey features, missing values can be reasonably replaced with 0 (No), assuming non-selection indicates a negative response. For categorical variables, grouping or binning similar categories helps maintain data consistency while reducing sparsity.

- For numerical features, more sophisticated imputation methods such as K-Nearest Neighbors (KNN) imputation, mean or median imputation, iterative imputation, or row dropping were proposed depending on the proportion of missing data and the importance of the feature.

- These methods aim to minimize data loss, preserve distribution characteristics, and improve the overall quality and performance of subsequent analyses or predictive modeling.



Thank You

Rudra Prasad Bhuyan