Missingno: a missing data visualization suite

Aleksey Bilogur[1]

1 Independent

DOI: 10.21105/joss.00547

Software

- Review
- Repository
- Archive

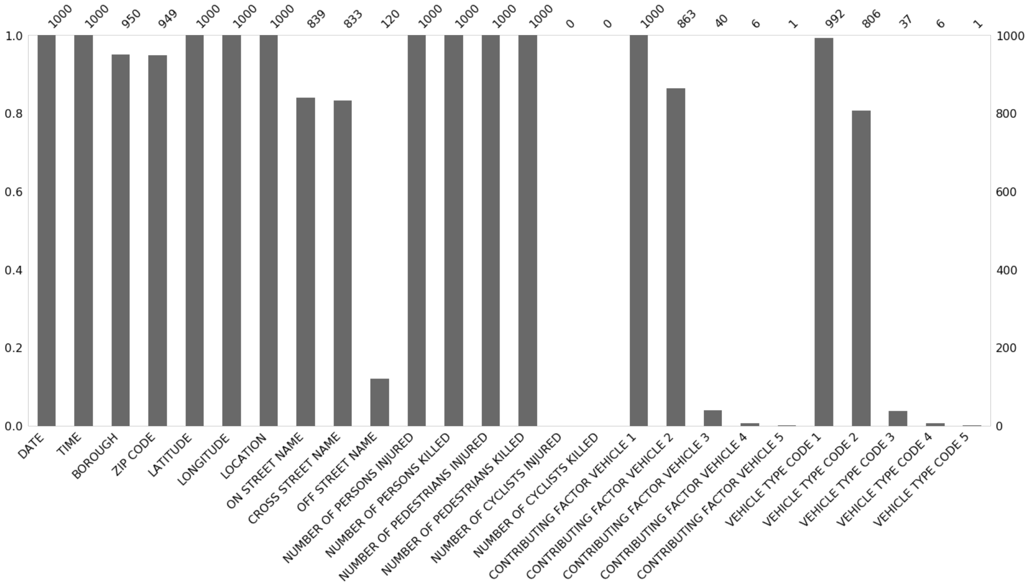
Submitted: 27 October 2017
Published: 27 February 2018

Summary

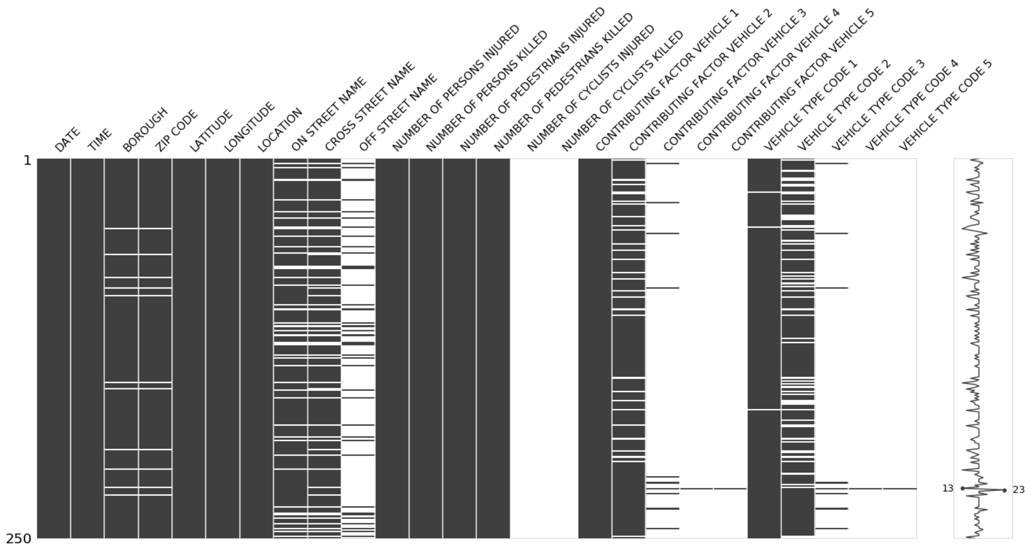
Algorithmic models and outputs are only as good as the data they are computed on. As the popular saying goes: garbage in, garbage out. In tabular datasets, it is usually relatively easy to, at a glance, understand patterns of missing data (or nullity) of individual rows, columns, and entries. However, it is far harder to see patterns in the missingness of data that extend between them. Understanding such patterns in data is beneficial, if not outright critical, to most applications.

missingno is a Python package for visualizing missing data. It works by converting tabular data matrices into boolean masks based on whether individual entries contain data (which evaluates to true) or left empty (which evaluates to false). This “nullity matrix” is then exposed to user assessment through a variety of special-purpose data visualizations.

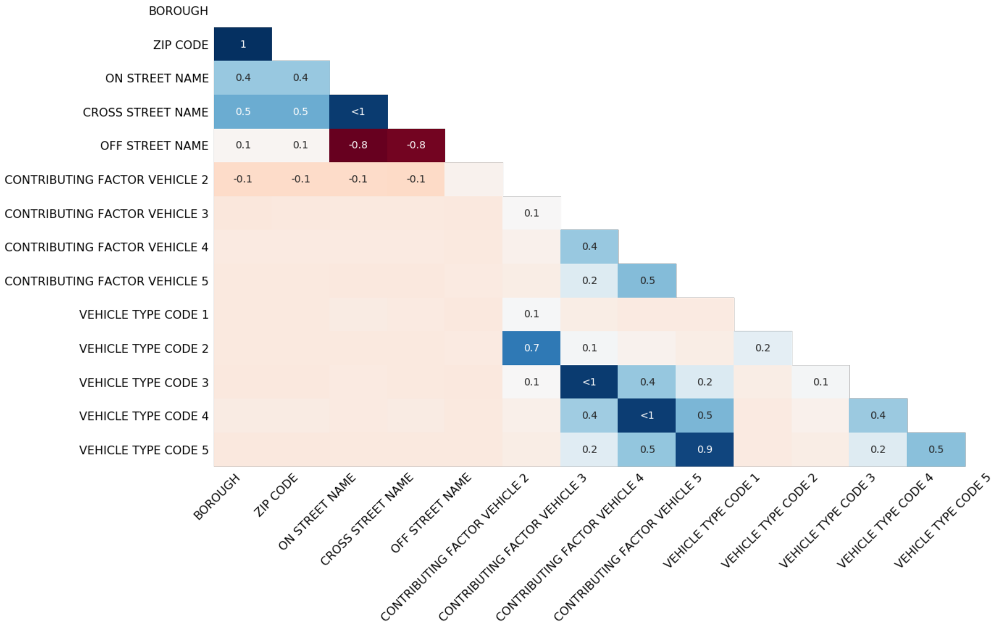
The simplest tool, the bar chart, is a snapshot of column-level information:



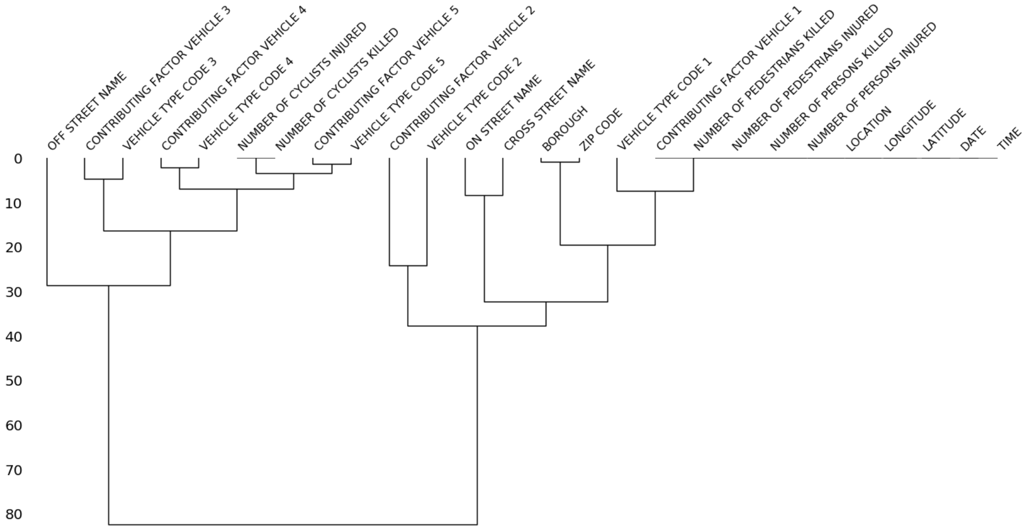
The matrix display provides a literal translations of a data table’s nullity matrix. It is useful for snapshotting general patterns:



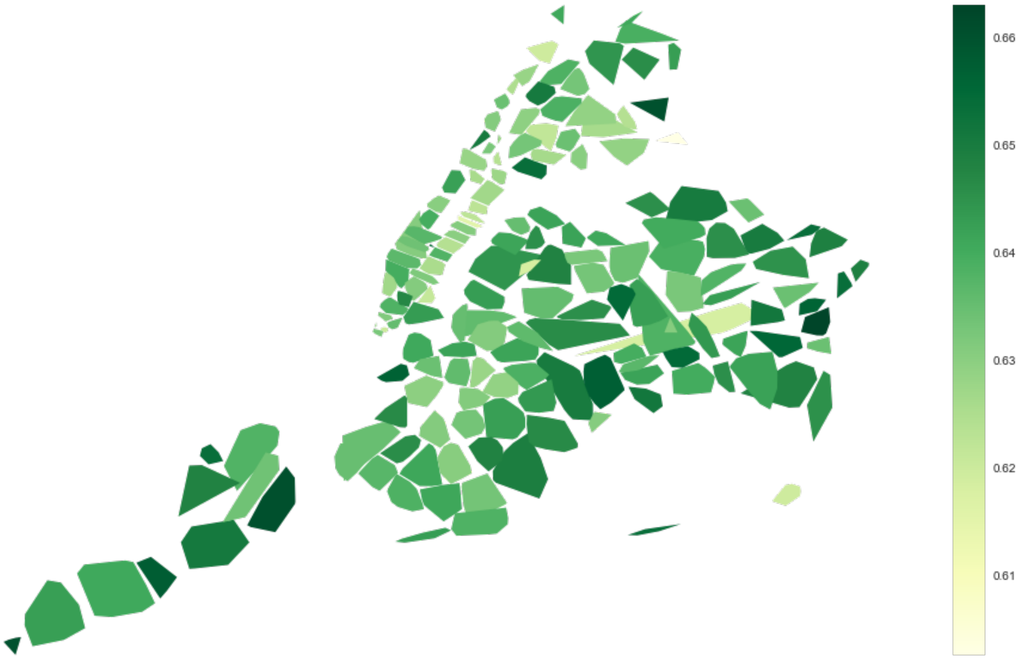
A heatmap provides a methodology for examining relationships within pairs of variables.



Higher-cardinality data nullity correlations can be understood using a hierarchically clustered dendrogram:



Finally, geospatial data dependencies are viewable using an approach based on the quadtree or convex hull algorithm:



The visualizations are consciously designed to be as effective as possible at uncovering missing data patterns both between and within columns of data, and hence, to help its users build more effective data models and pipelines. At the same time the package is designed to be easy to use. The underlying packages involved (NumPy (Stéfan van der Walt and Varoquaux 2011), pandas (McKinney 2010), SciPy (Jones et al. 2001–2001--), matplotlib (Hunter 2007), and seaborn (Waskom and others 2017)) are familiar parts of the core scientific Python ecosystem, and hence very learnable and extensible. missingno works “out of the box” with a variety of data types and formats, and provides an extremely compact API.

References

Hunter, John D. 2007. “Matplotlib: A 2D Graphics Environment” 9. https://doi.org/10.1109/MCSE.2007.55.

Jones, Eric, Travis Oliphant, Pearu Peterson, and others. 2001–2001--. “SciPy: Open Source Scientific Tools for Python.” http://www.scipy.org/.

McKinney, Wes. 2010. “Data Structures for Statistical Computing in Python.” *Proceedings of the 9th Python in Science Conference.*

Stéfan van der Walt, S. Chris Colbert, and Gaël Varoquaux. 2011. “The Numpy Array: A Structure for Efficient Numerical Computation” 13. https://doi.org/10.1109/MCSE.2011.37.

Waskom, Michael, and others. 2017. “Mwaskom/Seaborn: V0.8.1 (September 2017).” https://doi.org/10.5281/zenodo.883859.