

nycOpenData: The Data That Never Sleeps

Christian Martinez ¹

¹ The Open Data Lab, United States

DOI: [10.21105/joss.11339](https://doi.org/10.21105/joss.11339)

Software

- [Review](#) 
- [Repository](#) 
- [Archive](#) 

Editor: [Daniel S. Katz](#)  

Reviewers:

- [@danielskatz](#)

Submitted: 07 July 2026

Published: 25 September 2026

License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC BY 4.0](#)).

Summary

nycOpenData is an R package that provides streamlined access to the [New York City Open Data portal](#), enabling researchers, educators, students, and analysts to retrieve more than 2,000 public datasets, including continuously updated datasets, using a consistent tidy interface. By abstracting the underlying Socrata API and data infrastructure, the package reduces the technical barriers associated with discovering, accessing, and cleaning NYC Open Data.

Statement of Need

New York City maintains one of the most comprehensive open data portals in the world, providing thousands of public datasets spanning transportation, housing, public health, education, environmental monitoring, and city services. Despite the breadth of these resources, researchers, students, educators, journalists, and public-sector analysts often face technical barriers when accessing the data, including identifying dataset endpoints, constructing API requests, and cleaning raw JSON responses.

Historically, general-purpose tools for interacting with Socrata-hosted data portals, such as RSocrata ([Devlin et al., 2025](#)), have required users to manually identify dataset endpoints and interact directly with API infrastructure. While these tools provide flexible access across many Socrata implementations, they do not offer a streamlined workflow tailored to a specific municipal open data portal.

nycOpenData addresses these challenges by providing a unified interface for discovering and retrieving *all* publicly available datasets on the NYC Open Data portal through a consistent, tidy workflow. The package is also a core pedagogical component of the graduate course *Reproducible Research Using R*, where it serves as the primary tool for teaching reproducible data acquisition using live public datasets.

State of the Field

General-purpose R packages such as RSocrata ([Devlin et al., 2025](#)) have provided interfaces for downloading data from Socrata portals. These tools were designed to work across many different Socrata implementations but required users to manually locate dataset identifiers and construct queries for individual datasets.

In contrast, nycOpenData is specifically designed for the NYC Open Data ecosystem. Rather than functioning as a generic API wrapper, it provides integrated dataset discovery, human-readable dataset identifiers, standardized retrieval functions, automatic type conversion, and direct access to the complete NYC Open Data catalog through a single interface.

The architecture developed for nycOpenData has since provided the foundation for a broader ecosystem of portal-specific R packages. These packages retain a consistent approach to

dataset discovery and retrieval while adapting the interface to additional state, municipal, regional, and international open data portals.

Software design

Early versions of the package consisted of individual wrapper functions for specific datasets (e.g., `nyc_311()`). While straightforward, this architecture became increasingly difficult to maintain and could not reasonably scale to the more than 2,000 datasets available through NYC Open Data.

Following open peer review through the rOpenSci software review process (rOpenSci, 2026), the package architecture was substantially redesigned. Rather than maintaining individual functions for individual datasets, the package now centers around three general-purpose functions:

- `nyc_list_datasets()` — Retrieve a live catalog of available NYC Open Data datasets, including dataset titles, human-readable keys, Socrata UUIDs, endpoint URLs, and associated metadata.
- `nyc_pull_dataset()` — Download cataloged NYC Open Data datasets using either a human-readable key or dataset UUID, with support for filtering, ordering, date ranges, automatic type coercion, and optional column name cleaning.
- `nyc_any_dataset()` — Retrieve data directly from arbitrary NYC Open Data Socrata JSON endpoints without requiring inclusion in the package catalog.

This architecture allows the package to automatically adapt as the NYC Open Data portal evolves. Newly published datasets become immediately accessible through the package without requiring package updates or additional development effort.

This design shifts maintenance from tracking individual datasets to maintaining a stable interface for interacting with the NYC Open Data catalog as a whole.

The generalized architecture developed for `nycOpenData` has subsequently been reused to develop 12 additional R packages for interacting with state, municipal, regional, and international open data portals. These include `ausOpenData`, `chiOpenData`, `cincinnatiOpenData`, `coOpenData`, `ctOpenData`, `deOpenData`, `laOpenData`, `mtaOpenData`, `nolaOpenData`, `nyOpenData`, `SeattleOpenData`, and `sfOpenData`. The expansion of this architecture across 13 packages demonstrates that the underlying software design is extensible beyond a single city or open data portal.

Research impact statement

`nycOpenData` has been incorporated into educational, research, and public-facing projects that demonstrate its use for reproducible analysis of live public data. The package also serves as a core component of *The Open Data Lab*, an initiative focused on expanding access to public data through open-source software, education, and applied data projects.

Examples of work built around or supported by `nycOpenData` include:

- **NYC Open Data Student Gallery**, a published collection of graduate student research projects conducted in the M.S. in Psychological Research program at Brooklyn College, City University of New York (CUNY). This work was presented during NYC Open Data Week 2026.
- **Reproducible Research Using R**, an open educational textbook in which `nycOpenData` is used to introduce students to reproducible data acquisition and analysis using live public data.

- [Reproducible Research Using R: Student Work in Practice](#), a collection of student portfolios demonstrating reproducible analyses conducted using NYC Open Data.
- [NYC 311 Dashboard](#), a publicly deployed interactive dashboard built using nycOpenData that provides continuously updated analyses and visualizations of NYC 311 service request data.
- [OpenData Package Developer's Guide](#), an open guide documenting the framework used to develop additional OpenData R packages and extend the nycOpenData architecture to other public data portals.

Beyond these individual projects, the software architecture developed for nycOpenData has provided the foundation for a broader ecosystem of open data R packages developed through *The Open Data Lab* and its contributors. The *OpenData Package Developer's Guide* documents this framework and supports its extension to additional state, municipal, regional, and international open data portals.

More broadly, nycOpenData lowers barriers to working with public data by supporting reproducible research, classroom instruction, open-source software development, and public-facing data applications.

AI usage disclosure

Generative AI tools were used during the development of this software in two primary ways:

1. During software development, AI-assisted tools were occasionally used to help diagnose programming errors and resolve implementation issues.
2. AI-assisted tools were also used for grammar, spelling, and language refinement within the package documentation and manuscript.

All AI-generated suggestions were reviewed, modified where appropriate, and validated by the author prior to inclusion.

Acknowledgements

I am extremely grateful to the NYC Open Data team for maintaining the underlying data infrastructure that makes this work possible and for their continued commitment to making New York City's public data openly accessible.

I would also like to thank the students whose early contributions helped shape previous versions of the package. Although the software architecture has since evolved considerably, their work laid the foundation for the project.

Lastly, I am especially grateful to rOpenSci editor Ronny Hernandez Mora and reviewers Haolin Dong and Michael Pascale for the time, care, and thoughtful feedback they provided throughout the rOpenSci software peer review process. Their contributions substantially improved the design, functionality, documentation, and long-term maintainability of nycOpenData.

References

- Devlin, H., Tom Schenk, Jr., Leynes, G., Lucius, N., Malc, J., Silverberg, M., & Schmeideskamp, P. (2025). *RSocrata: Download or upload 'Socrata' data sets*. <https://doi.org/10.32614/CRAN.package.RSocrata>
- rOpenSci. (2026). *Software review: nycOpenData*. <https://github.com/ropensci/software-review/issues/750>