The Journal of Open Source Software

zipcodeR: An all-in-one toolkit of functions & data for working with U.S. ZIP codes in R

Gavin C. Rozzi[1]

1 Rutgers Urban & Civic Informatics Lab, Edward J. Bloustein School of Planning & Public Policy, Rutgers, The State University of New Jersey, New Brunswick, NJ, USA

DOI: 10.21105/joss.0XXXX

Software

- Review
- Repository
- Archive

Editor: Editor Name

Submitted: 01 January XXXX
Published: 01 January XXXX

Summary

`zipcodeR` is a new R package intended to make working with datasets containing U.S. ZIP codes easier & simplify the process of relating ZIP code level data to other geographies in R. The package provides an offline database of ZIP codes & related socio-demographic data for over 41,000 ZIP codes in all 50 states. It enables the efficient retrieval of ZIP codes at the national, state & county level along with geocoding existing data containg ZIP codes for use in research. `zipcodeR` addresses limitations inherent in past packages to provide a new, general-purpose solution to assist with processing ZIP code-level data.

Statement of need

There has been a lack of a currently supported, lightweight & general-purpose library for working with U.S. postal codes in R that is available through the Comprehensive R Archive Network (CRAN). A popular, previously supported R library that had a degree of overlap with zipcodeR was the package `zipcode`, which has since been archived from the CRAN repository and no longer actively supported or updated (Breen, 2012). The removal of `zipcode` from CRAN has limited its reach & utility to the wider R community. Another key limitation of the package was its reliance on a data source that was last updated in 2004 according to its author, which was later integrated with an additional source in 2012 (Breen, 2012). Because ZIP code boundaries are based upon the mail delivery routes of the United States Postal Service - which can change over time - researchers cannot accurately rely upon older packages that have not been updated for use in research involving data at the ZIP code level.

`zipcodeR` goes further than previous packages and aims to address limitations in this area by including a more comprehensive set of data about each ZIP code, including among other things the radius in miles represented by the ZIP code, the county it is located in, whether or not that ZIP code exclusively represents post office boxes, along with selected population, income & property value statistics for each ZIP code along with numerous other indicators that could be useful to researchers when comparing observations across ZIP codes.

Other existing packages that overlap with `zipcodeR`, such as `choroplethrZip` are too large to be distributed via CRAN & were designed for specific use cases like mapping, which limits their reach & potential adoption by the wider R community (Lamstein, 2020). As this package is exclusively distributed via Github, it must be manually installed using a package such as `devtools` or `remotes`. Less experienced R users, which include students learning the language in an educational setting, may not be comfortable installing packages directly from Github or discover them without the visibility & ease of installation provided by being listed in CRAN

`zipcodeR` seeks to find a sensible middle ground between very large R packages designed with a specific use case in mind like `choroplethrZip` and leaner, but far more limited packages

like zipcode by including a comprehensive set of data bundled with the package, but not a large shapefile of polygons representing ZIP code boundaries that increases storage demands and forecloses the possibility of being published on CRAN due to their lack of support for large files bundled with R packages.

The aforementioned limitations inherent within the existing ecosystem of R packages necessitated the development of a new general-purpose solution like zipcodeR in order to ensure that a more up-to-date source of data about ZIP codes is available to facilitate research incorporating ZIP code-level data & eliminate repetitive tasks by focusing on common workflows used in the cleaning and analysis of data at the ZIP code level.

Data

The underlying database of ZIP codes builds upon related work in Python done by Sanhe Hu that was previously published under the MIT license. zipcodeR has adopted the same underlying SQLite database containing data on ZIP codes that was used to power Hu's Python package, uszipcode (Hu, 2020). However, it should be noted that this package differs in its approach from uszipcode by bundling a copy of the ZIP code database directly with the package rather than requiring it to be downloaded when first initializing the package.

Research & educational applications

In addition to being used by data science practitioners, zipcodeR has already been used in graduate-level courses in data science for relating ZIP codes to other data sources & administrative boundaries (Green et al., 2020), as well as graduate-level coursework in urban informatics focused on the analysis of administrative data produced by state & local government to evaluate land use patterns in cities (Wang, 2020).

Conclusion

This package contributes to the R package ecosystem by making it easier & more convenient to process data containing U.S. ZIP codes in R. This package will aid researchers in working with data at the ZIP code level using R without depending on online services or APIs.

Acknowledgements

The author would like to acknowledge Sanhe Hu, author of the Python package uszipcode, whose work developing the uszipcode Python package provided a foundation & inspiration for the development of this package.

References

Breen, J. (2012). *Zipcode: U.s. ZIP code database for geocoding*. https://CRAN.R-project.org/package=zipcode

Green, S., Gonsalves, M., Markowska-Desvallons, D., Khaimova, O., & Mazon, J. (2020). *DATA 607 Final Project - COVID rates vs. Election Results in NYC*. https://rpubs.com/OrliKhaim/DATA607_Final_Project

Hu, S. (2020). *uszipcode 0.2.4 documentation*. https://uszipcode.readthedocs.io/index.html

Lamstein, A. (2020). *choroplethrZip: Shapefile, metadata and visualization functions for US zip code tabulated areas (ZCTAs)*. https://github.com/arilamstein/choroplethrZip

Wang, J. (2020). *Revealing Knowledge – A closer view on the Building Permits – Seeing Boston Neighborhoods through Administrative Data*. https://sppua5262.wordpress.com/2020/10/07/revealing-knowledge-a-closer-view-on-the-building-permits/