**Title / name of your software**
zipcodeR: Advancing the analysis of spatial data at the ZIP code level in R

**Authors / main developers**
Gavin C. Rozzi

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

**Abstract**
The United States Postal Service (USPS) assigns unique identifiers for postal service areas known as ZIP codes which are commonly used to identify cities and regions throughout the United States in datasets. Despite the widespread use of ZIP codes, there are challenges in using them for geospatial analysis in the social sciences. This paper presents zipcodeR, an R package that facilitates analysis of ZIP code-level data by providing an offline database of ZIP codes and functions for geocoding, normalizing and retrieving data about ZIP codes and relating them to other geographies in R without depending on any external services.

**Keywords:**
*ZIP code; R; ZCTA; ZIP code tabulation area; zipcodeR*

## 1. Introduction

zipcodeR is a package for the R statistical programming language [11] aimed to make research and geospatial analysis easier when analyzing data aggregated at the ZIP code level, which is one of the most common forms of geographic data encountered by researchers in the social sciences. While less ideal than other geographies, such as Census tracts, an individual's ZIP code can be used to predict social determinants of health and reveal inequalities in small areas [15,21]. zipcodeR makes working with these data easier by enabling users to rapidly acquire, geocode and relate ZIP code-level data to states, counties, Census tracts, and other geographies commonly encountered in social science research. zipcodeR contributes to the R data science ecosystem by integrating multiple open-source datasets and official government crosswalk files to provide data on over 41,000 ZIP codes that are suitable for integration into larger projects via datasets and wrapper functions. zipcodeR has been available on the Comprehensive R Archive Network (CRAN) since September, 2020 [14].

In addition to making these data sources available for integration, zipcodeR includes a suite of functions for programmatically retrieving data on ZIP codes by U.S. state, city, county, time zone and other search parameters to support the rapid analysis of data. Beginning in version 0.3.0 of the package, new geographic functions make it possible to calculate the geographic distance between ZIP codes in miles when only the ZIP codes are known, as well as searching all ZIP codes located in a specific radius around a given coordinate pair with a single function call in R. These features make it possible to easily

map ZIP code-level data without depending on external geocoding APIs, which makes it especially useful for working with restricted datasets that must not be sent outside of internal networks.

## 2. The zipcodeR package

### 2.1 Methodology

The zipcodeR package integrates multiple open data sources to provide maximum utility for researchers analyzing datasets aggregated at the ZIP code-level. Most of the data used in the package's data retrieval functions are supplied by the package's zip_code_db object, an R data frame that contains 41,877 observations of 24 variables. A data dictionary for all of the variables of data available in zip_code_db is provided by the package's documentation [12]. This data frame was built by adapting the approach used by Hu, (2020) and converting the dataset from an SQLite database into the native binary format used by R for integration into the package. A visualization of the spatial distribution of ZIP codes located within the continental U.S. by region using zip_code_db dataset is shown in Fig. 1.

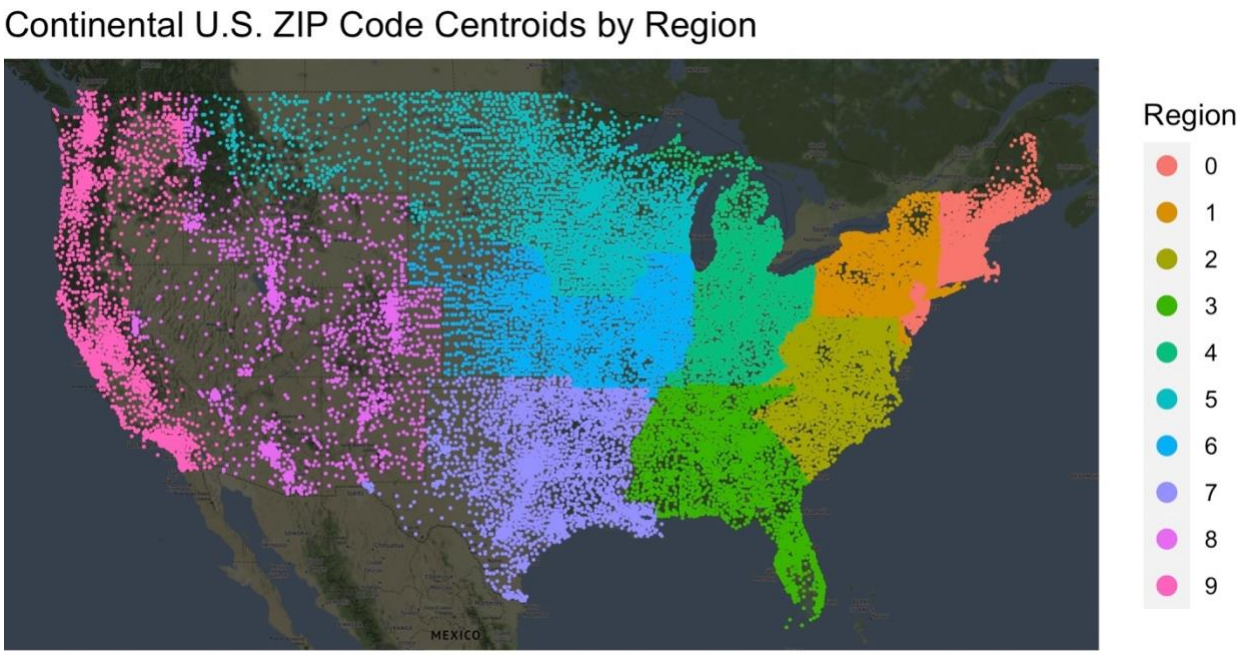


Fig. 1 A map showing centroids of each ZIP code contained within zipcodeR's zip_code_db dataset for the continental U.S., colorized by the region of the ZIP code as assigned by the USPS. The region is determined by the first character of the ZIP code as provided by USPS. This map was produced using the ggmap R package [7].

Similar work was undertaken to convert additional ZIP code crosswalk files produced by U.S. federal government agencies, including the U.S. Department of Housing & Urban Development (HUD) and the Census Bureau. These additional datasets were integrated to aid the task of relating ZIP codes to Census Tracts and related geographies [18,22].

ZIP codes are often used to compare differences between different regions of a state in social science research across one or more variables. Fig. 2 shows a visualization of the differences between ZIP code metadata contained within zip_code_db ordered by the population of each ZIP code. Table

plots are an effective method for visualizing the properties of large, multivariate datasets, such as zip_code_db. A table plot of the dataset was produced through the use of the tabplot R package [16]. These visualizations are produced by creating statistical bins from the observations contained within the dataset, with the bars representing the mean of numeric values and frequencies for categorical values. In Fig. 2, each bin contains 419 observations, about 10% of the full data frame. An analysis of this figure makes it possible to show that there are clear differences between both the distribution of the population of the United States and median home value across the 3 types and 9 regions of ZIP codes.

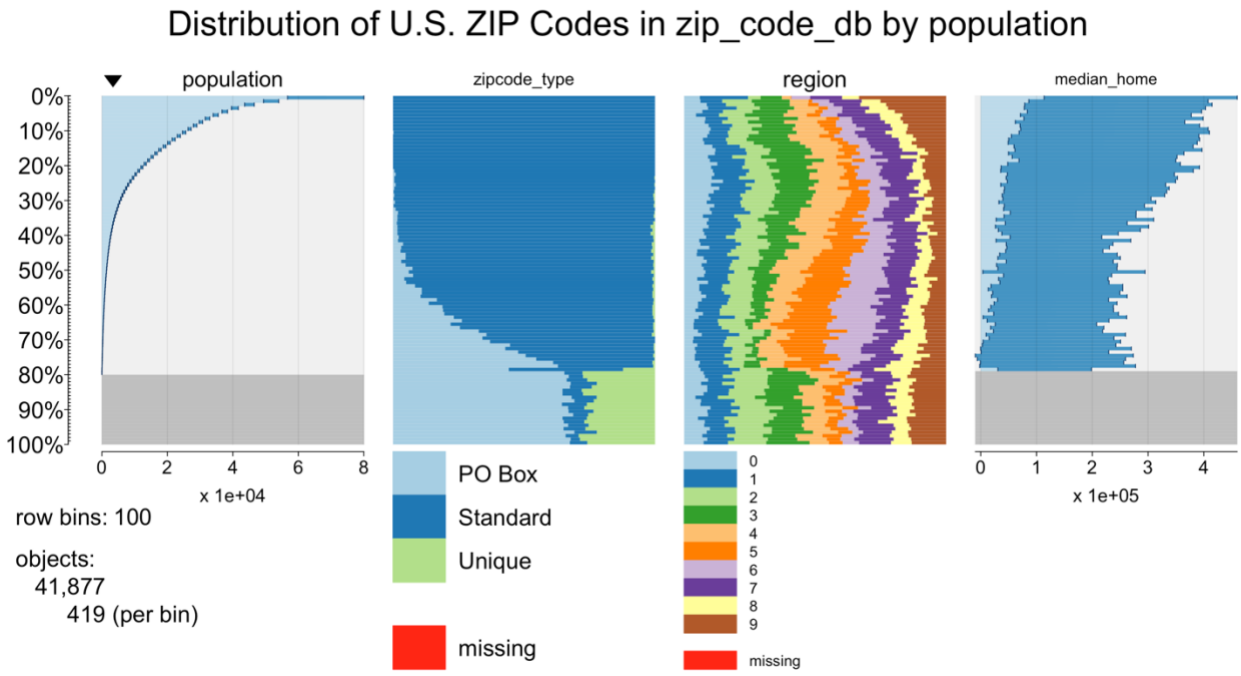


Fig 2. A table plot of the zip_code_db data frame provided by the zipcodeR package. This dataset provides data that is used to support most of the package's functions. In this visualization. statistical bins are ordered based on the population of each ZIP code contained within the dataset.

The zip_code_db dataset serves as a basis for many of the functions that were developed to obtain data on user-provided ZIP codes. Most of the functions provided by the package return a subset of these data based upon user-supplied search criteria. For example, the function `reverse_zipcode()` will return all 24 columns of data about a ZIP code contained within zip_code_db when provided with a ZIP code by the user. Many of the package's functions build upon the dplyr and raster packages for data manipulation & spatial calculations [5,20]. A full reference of the functions provided by zipcodeR may be viewed via the package's documentation [13]. Other functions provided by the package which do not rely on zip_code_db include the normalize_zip() function which relies on custom logic to clean ZIP codes that are messy or non-standardized, a common task encountered by researchers working with ZIP code-level data.

### 2.2 Comparison with existing R packages

Previous R packages have sought to address some of the challenges being addressed by zipcodeR but have suffered from drawbacks inherent in existing approaches. There has been a lack of a

currently supported and general-purpose library for working with U.S. ZIP codes in R that is available through the CRAN repository and covers most general use cases for exploratory data analysis and research applications. 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 updated by its maintainer [1]. The archival of the zipcode package from CRAN and lack of further support has limited its reach and utility to the wider R community, creating a gap in functionality. Another key limitation of the zipcode package was its sole 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. 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 newly collected data at the ZIP code level.

Other currently supported packages that overlap with zipcodeR, such as choroplethrZip, while useful for their intended applications, are too large to be distributed via CRAN and were designed for specific use cases like mapping [8]. As the choroplethrZip package is exclusively distributed via GitHub, it must be manually installed using a package such as remotes or devtools, limiting the discoverability of these types of packages for users new to R.

The zipcodeR package seeks to achieve 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 dataset and wrapper functions for subsetting data, 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 very large files included with packages.

**3. Impact Overview**

The zipcodeR package was originally developed to eliminate repetitive tasks for some of the workflows for using ZIP code level data that are common in social science research workflows using the R statistical programming language.

Since the package's initial release in late 2020, zipcodeR has enabled the rapid prototyping of research and data science projects and has been implemented in a diverse range of R projects, both published and those still under development. For example, an interactive dashboard bult in R Shiny was published by a data analytics firm showcasing an income tax dataset that shows economic data by ZIP code and state through an analysis of administrative data from the Internal Revenue Service of the United States aggregated from individual tax returns [9].

Furthermore, numerous academic and public health research projects that have successfully implemented zipcodeR in their analyses, as shown by a search of publicly available GitHub repositories. These have included efforts by academic researchers to create composite metrics of data collected during the COVID-19 pandemic, a study of physician mental health in partnership with the American Medical Association, as well as well as an analysis of how equitable COVID-19 vaccine distribution was in the state of Texas [3,17,19].

zipcodeR has also seen adoption in educational contexts, specifically in data science and urban informatics courses and projects. Several student projects & practicums were published on GitHub that

implemented the package, demonstrating its use in projects exploring relationships between COVID-19 case counts and election data at the ZIP code-level in New York City [4]. Another project that implemented zipcodeR in their analysis included a program evaluation at a university cancer center [2]. Another educational application of the package was seen via Tidy Tuesday [10], a weekly coding challenge intended to provide a safe environment for learning R and data analysis skills supported by the data science community. For the week of May 11th, 2021, Tidy Tuesday implemented the zipcodeR package as a basis for an analysis of broadband accessibility in the United States using ZIP code-level data. This demonstrated the package's potential in supporting the analysis of commonly available administrative data & eliminating repetitive tasks, especially for less-experienced R users.

## 4. Conclusion and Future Work

This paper introduced the zipcodeR package for streamlining the analysis of ZIP code-level data in social science research. The package has now had three major iterations published on CRAN since its initial release and is gradually becoming more comprehensive as additional user feedback is considered for future iterations. Contributions of code and feedback from researchers who have implemented zipcodeR have also been very helpful for the development of the package and further contributions from the community are welcome. While the package is relatively feature-complete at this stage, future iterations may see the integration of additional ZIP code crosswalk datasets available from the U.S. federal government, as well as the improvements to the existing functions for increased efficiency and broader applicability to research workflows.

## Declaration of Competing Interest

The author declares that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## References


[1] J. Breen, zipcode: U.S. ZIP Code database for geocoding, (2012).

[2] R. Dario Herrera, GitHub - UACC-renedherrera/UAZCC_COE_Program_Evaluation: Implementation of different evaluation tools to measure program effectiveness., (2021).

[3] P. Ganguly, S. Mukherjee, A.S. Kumar, GitHub - abinesh-23/Physician-mental-health-analysis-AMA-: This repository contains the code for the research project partnered with American Medical Association (AMA) to understand physician mental health working as a frontline worker during COVID pandem, (2021).

[4] S. Green, M. Gonsalves, D. Markowska-Desvallons, O. Khaimova, J. Mazon, DATA 607 Final Project - COVID rates vs. Election Results in NYC, (n.d.).

[5] R.J. Hijmans, raster: Geographic Data Analysis and Modeling, (2021).

[6] S. Hu, uszipcode 0.2.4 documentation, (2020).

[7] D. Kahle, H. Wickham, ggmap: Spatial Visualization with ggplot2, R J. 5 (2013) 144–161.

[8] A. Lamstein, choroplethrZip: Shapefile, Metadata and Visualization Functions for US Zip Code Tabulated Areas (ZCTAs), (2020).

[9] D. Lucey, Introducing the Redwall IRS SOI Tax Dashboard - Redwall Analytics, Redwall Anal. (2021).

[10] T. Mock, Tidy Tuesday: A weekly data project aimed at the R ecosystem, (2021).

[11] R Core Team, R: A Language and Environment for Statistical Computing, (2021).

[12] G.C. Rozzi, ZIP Code Database — zip_code_db • zipcodeR, ZipcodeR Doc. (2021).

[13] G.C. Rozzi, Function reference • zipcodeR, (2021).

[14] G.C. Rozzi, Data & Functions for Working with US ZIP Codes [R package zipcodeR version 0.3.0], (2021).

[15] E. Sokol, How Geographic Data Can Help Address Social Determinants of Health, Heal. IT Anal. (2019).

[16] M. Tennekes, E.D. Jonge, P. Daas, Visualizing and Inspecting Large Datasets with Tableplots, J. Data Sci. 11 (2013) 43–58.

[17] UNC School of Government ncIMPACT Initiative, GitHub - ncIMPACT/covid-keys-impact: Examining composite variables for COVID-19 Keys to Economic Recovery project, (2021).

[18] United States Census Bureau, Relationship Files, (n.d.).

[19] L.B.J.S. of P.A. University of Texas at Austin, Texas COVID-19 Vaccine Tracker: Explore Equity, (2021).

[20] H. Wickham, R. François, L. Henry, K. Müller, dplyr: A Grammar of Data Manipulation, (2021).

[21] T.C. Yang, S. Kim, Y. Zhao, S. won E. Choi, Examining spatial inequality in COVID-19 positivity rates across New York City ZIP codes, Heal. Place. 69 (2021) 102574.

[22] HUD USPS ZIP Code Crosswalk Files | HUD USER, (n.d.).

**B- Required Metadata**

**B1 Current executable software version**

*Table 1 – Software metadata*

| Nr | (executable) Software metadata description | ***Please fill in this column*** |
|---|---|---|
| S1 | Current software version | *0.3.0* |
| S2 | Permanent link to executables of this version | *https://github.com/gavinrozzi/zipcodeR/releases/tag/0.3* |
| S3 | Permanent link to Reproducible Capsule | *https://codeocean.com/capsule/4509180/tree/v1* |
| S4 | Legal Software License | *GNU GPL* ≥ 3 |
| S5 | Computing platform / Operating System | *Linux, macOS, Windows, Unix-like* |
| S6 | Installation requirements & dependencies | *R 3.5 or greater, dplyr, tidycensus, udunits2, raster, rlang, magrittr* |
| S7 | If available Link to user manual - if formally published include a reference to the publication in the reference list | *https://gavinrozzi.github.io/zipcodeR/index.html* |
| S8 | Support email for questions | gr@gavinrozzi.com |

**B2 Current code version**

*Table 2 – Code metadata*

| Nr | Code metadata description | ***Please fill in this column*** |
|---|---|---|
| C1 | Current Code version | *Rolling release*<br>*commit 3f982b808c1ac2c4ed777c70c7cc0b3ec3e5bedc* |
| C2 | Permanent link to code / repository used of this code version | *https://github.com/gavinrozzi/zipcodeR* |
| C3 | Permanent link to Reproducible Capsule | *https://codeocean.com/capsule/4509180/tree/v1* |
| C4 | Legal Code License | *GPLv3* |
| C5 | Code Versioning system used | *git* |
| C6 | Software Code Language used | *R* |
| C7 | Compilation requirements, Operating environments & dependencies | *R ≥ 3.5* |
| C8 | If available Link to developer documentation / manual | |
| C9 | Support email for questions | gr@gavinrozzi.com |