Software Impacts 10 (2021) 100176

Contents lists available at ScienceDirect

Software Impacts

journal homepage: www.journals.elsevier.com/software-impacts

Original software publication

njtr1: An R package for researching road safety in New Jersey using open crash data

Gavin C. Rozzi

Rutgers Urban & Civic Informatics Lab, Edward J. Bloustein School of Planning & Public Policy, Rutgers The State University of New Jersey, 33 Livingston Ave., New Brunswick, NJ 08901, United States of America

ARTICLE INFO

Keywords:
njtr1
Road safety
Car crash data
Urban planning
Car accidents
NJTR-1
Motor vehicle crashes

ABSTRACT

New Jersey law requires police officers to document motor vehicle collisions on a standardized form known as NJTR-1. The data collected via this form contain detailed information about motor vehicle crashes as well as drivers, vehicles & pedestrians involved in crashes, a valuable, but often underutilized resource for studying road safety in New Jersey. This paper presents njtr1, an R package that enables road safety and urban planning research in New Jersey by facilitating the easy download, automated cleaning and analysis of the raw crash table data published by the New Jersey Department of Transportation using the R programming language.

Code metadata

Current code version	0.4.0.9000
Permanent link to code/repository used for this code version	https://github.com/SoftwareImpacts/SIMPAC-2021-141
Permanent link to reproducible capsule	https://codeocean.com/capsule/6782509/tree/v1
Legal code license	GPLv3
Code versioning system used	git
Software code languages, tools and services used	R
Compilation requirements, operating environments and dependencies	R > 3.5
If available, link to developer documentation/manual	https://gavinrozzi.github.io/njtr1/index.html
Support email for questions	gr@gavinrozzi.com

1. Introduction

Wherever vehicles are driven, collisions are inevitably likely to occur as a consequence of their operation by drivers. This is the unfortunate reality of car ownership. In New Jersey alone, a data analysis from the Insurance Institute for Highway Safety estimates that 559 deaths occurred from 525 fatal accidents in 2019 alone [1]. Like their counterparts in other jurisdictions, officials in New Jersey have adopted a standardized form for documenting the factual circumstances surrounding car crashes that took place within the state, which is known as form NJTR-1.

New Jersey police officers have historically filled out paper copies of the NJTR-1 form following accidents and they are often used by insurance companies processing claims arising out of the accidents. The standardized nature of the NJTR-1 crash report form and large volume of data collected through its implementations has made it a rich source of data for multiple stakeholders, including accident scene investigators, insurance companies handling claims arising out of car accidents, planners seeking to identify unsafe road conditions, as well as academic researchers seeking to study research questions that can be addressed from the data collected using this form.

The code (and data) in this article has been certified as Reproducible by Code Ocean: (https://codeocean.com/). More information on the Reproducibility Badge Initiative is available at https://www.elsevier.com/physical-sciences-and-engineering/computer-science/journals.

E-mail address: gr@gavinrozzi.com.

https://doi.org/10.1016/j.simpa.2021.100176
Received 25 October 2021; Received in revised form 10 November 2021; Accepted 11 November 2021

Table 1
Crash data tables available via the njtr1 package.

Table	Data description	Dates covered	Number of columns (2017 schema)
Accidents	Each row represents a single motor vehicle crash and environmental conditions at the time it occurred.	2001–present	51
Drivers	List of drivers involved in a crash with additional demographic data	2001–present	21
Vehicles	Vehicle-level data describing the make and model of all vehicles involved in a particular crash.	2001–present	40
Occupants	Describes demographic characteristics on the occupants of a vehicle involved in a particular crash.	2001–present	14
Pedestrians	Describes the characteristics of Pedestrians	2001–present	35

2. Data

As the state agency responsible for regulating transportation within the state, the New Jersey Department of Transportation (NJDOT) collects crash data and publishes multiples data tables derived from NJTR-1 data on their website [2].

Crash data published by NJDOT of five distinct tables containing records describing motor vehicle accidents, drivers, vehicles, vehicle occupants & pedestrians that were affected by motor vehicle crashes that occurred in New Jersey and were reported to law enforcement authorities. Table 1 presents an overview of the five tables of crash data which are made accessible via the njtr1 package. Each of the table names listed in Table 1 can be passed to the get_njtr1() function and the package will automatically generate the necessary queries to download the data, apply the appropriate schema based on the year selected and clean the data prior to storing it as an R dataframe.

3. Challenges working with NJTR-1 data tables

New users seeking to research road safety using the NJTR-1 data tables often have a steep learning curve working with them because the files that are downloadable from the NJDOT website contain no column headers that denote what each column represents, and other issues associated with the format of the data can make it difficult to load the data into a statistical software environment such as R [3].

NJDOT does publish several PDF files that describe the content and proper name for each of the columns of data contained within each table, but these are not suitable for applying the column headers in bulk to a given data table [4]. For the accidents table alone, properly labeling the data would necessitate manually investigating and labeling 50 columns of data which is not an insignificant task and prone to human error. The names of the fields of data were extracted from these PDF files using Tabula [5], cleaned with the janitor R package [6] & bundled with the njtr1 package in order make these data tables more usable for R users. Furthermore, the interface of the NJDOT web interface does not make it feasible to download multiple years' worth of data or multiple tables at the same time — one must manually keep selecting tables, years and geographic locations to obtain data using this method.

4. The njtr1 package

The njtr1 package is available for download via the Comprehensive R Archive Network (CRAN) by running 'install.packages("njtr1")' at the R console [7]. The package's primary features are accessed by using the function 'get_njtr1()'. This function accepts the year and type of crash table as arguments, and the function will make the necessary requests to the NJDOT server, download & clean the data and store the result as an R dataframe suitable for further analysis and modeling. When the raw data is downloaded, the package automatically applies the correct schema and column headers based on the year of the data in order to provide clean and discernable column names. The package can also read offline versions of previously downloaded crash tables by using the 'read_njtr1()' function and pointing it to the path of the saved data.

njtr1 seeks to address the difficulties associated with working with these data by providing a suite of functions that can read, download and clean the motor vehicle crash data and get it in a format usable for road safety research in R. The package also bundles additional datasets used to facilitate analysis in one convenient wrapper. By eliminating the barriers inherent in working with New Jersey's car crash data, njtr1 enables researchers to focus on addressing their research questions rather than cleaning data. Among other applications, analyzing these data can inform data-driven transportation planning decisions and safety interventions to address unsafe road conditions within the state.

5. Motor vehicle accidents

To demonstrate how the package can be used to visualize spatial and temporal trends in car accidents within the state all accidents for the year 2019 were downloaded using the njtr1 package and clipped to a shapefile of New Jersey using R. This dataset consisted of 78,969 observations of 50 variables. The kernel density of the spatial distribution of the motor vehicle accidents was calculated with a 1000 foot search radius using the spatstat package for R [8] and the result is presented in Fig. 1 below.

Selected variables from this dataset were used to create a correlogram in Fig. 2, which shows how well variables within the 2019 car accidents dataset correlate with one another, a positive value indicates a strong correlation, and a negative value indicates a negative association. No statistically significant correlation was detected in the white squares.

A time series visualization of the total number of car accidents that occurred on each day of 2019, grouped by New Jersey County is presented below to demonstrate how the package can be used to study differences between counties regarding trends in motor vehicle crashes through space and over time (see Fig. 3).

In addition to the main dataset describing accidents, several other datasets can be downloaded that can be joined to this table. Each one of the additional tables available via the njtr1 package will subsequently be presented.

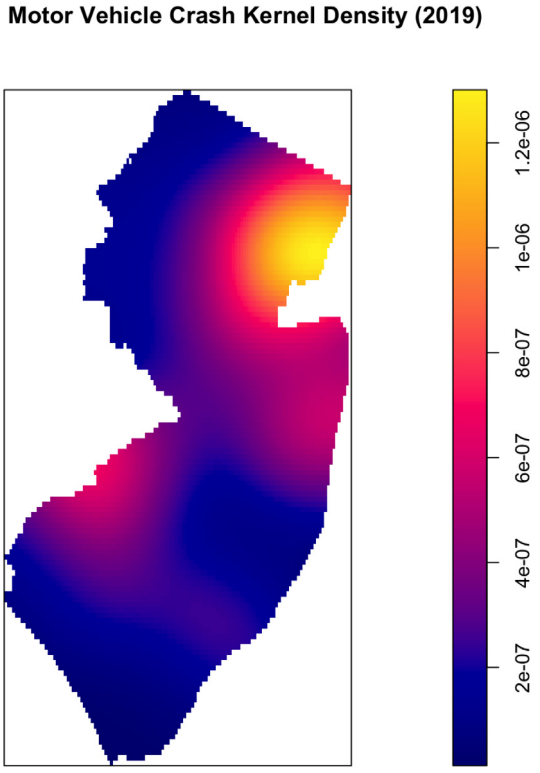

Fig. 1. Kernel density of 2019 geotagged accident records clipped to New Jersey state shapefile.

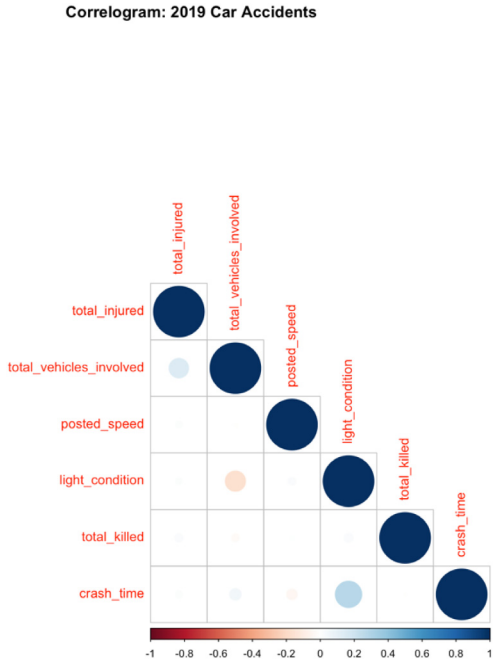

Fig. 2. Correlogram showing the relationship between a subset of variables from the 2019 NJTR-1 Accidents table obtained via the njtr1 R package.

6. Pedestrians involved in accidents

The second table of data accessible via the njtr1 package contains data on the pedestrians involved in car accidents. This table tracks demographic information about the pedestrians and whether they were charged with a violation in connection with the incident. Fig. 4 presents a histogram of the age of all pedestrians involved in 2019 motor vehicle accidents in New Jersey.

7. Drivers involved in accidents

Like the pedestrians table, this table tracks demographic information and provides the ZIP code of the residence of the drivers involved

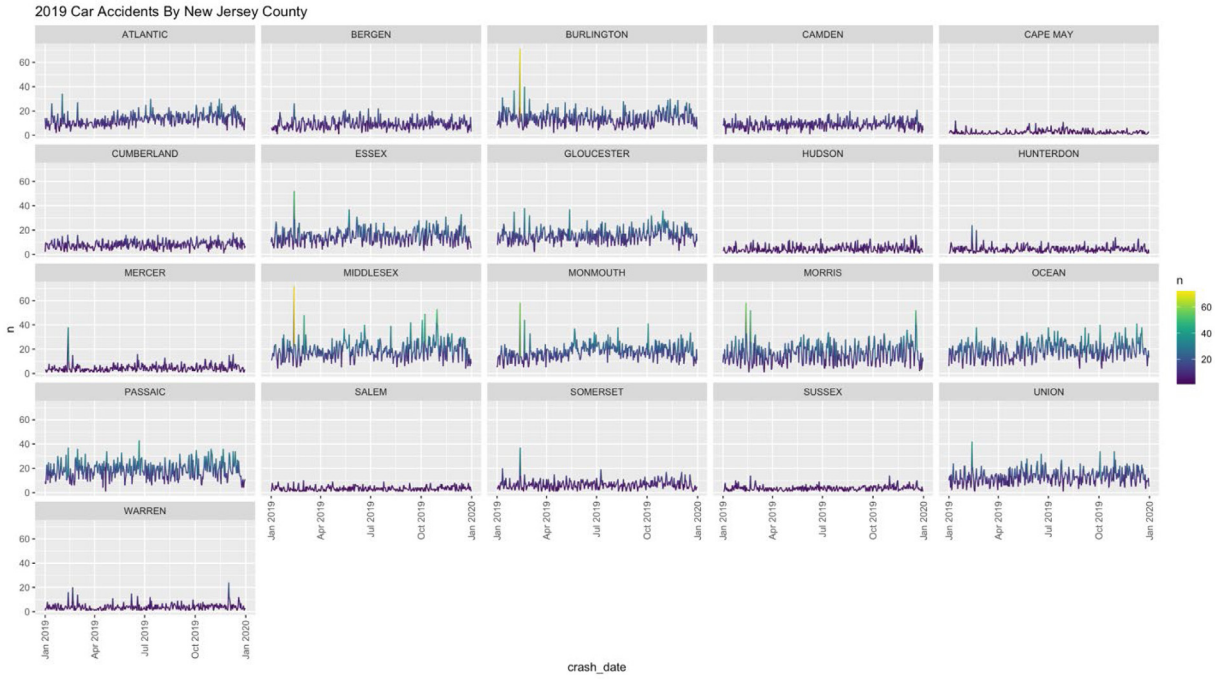

Fig. 3. Line plot of total number of motor vehicle accidents reported in NJTR-1 data per day during 2019, by New Jersey county. These plots are based on the Accidents table available via the package.

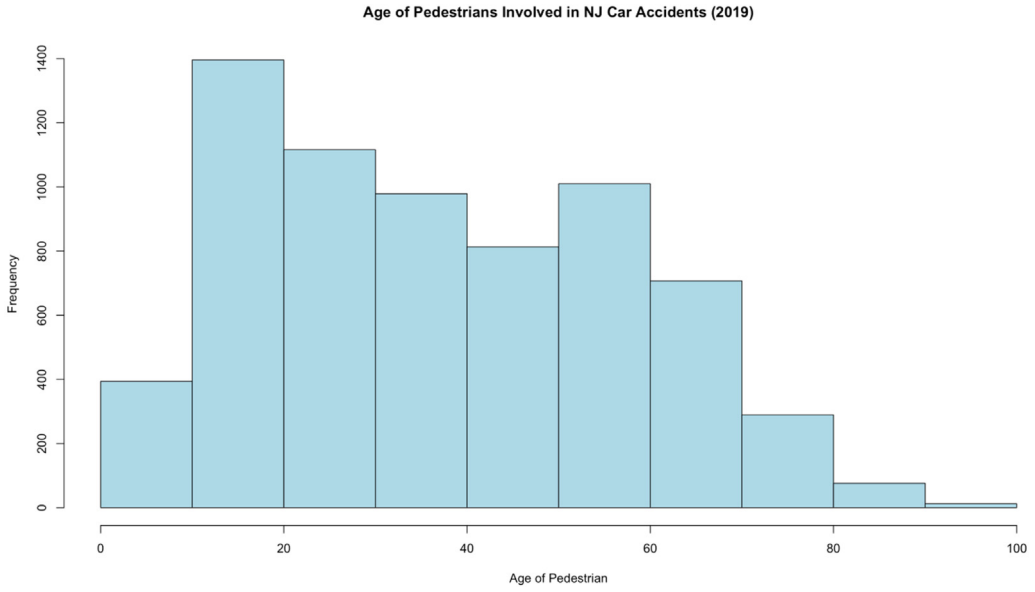

Fig. 4. Histogram of pedestrian age for pedestrians involved in New Jersey car accidents (2019).

in the accidents. If any charges were filed against the driver arising out of the incident, that information is also available in this table.

8. Vehicles involved in accidents

The last table of data that can be acquired using the njtr1 package tracks the vehicles that were involved in car accidents. This table lists the make and model of the vehicle, as well as further details about where it was damaged in the incident.

9. Impact overview

To the best of the author's knowledge, njtr1 is the first complete package of its kind that specifically focuses on New Jersey motor vehicle accident data and enables the exploration of research questions specifically targeted to communities located in the state. The package fills an important void by focusing on the state-specific dataset maintained by New Jersey officials. By providing the ability to rapidly acquire, clean and analyze motor vehicle crash records in New Jersey, the njtr1 package enables researchers using R to add another valuable layer of data to their research that can potentially generate additional insights regarding road safety within New Jersey and the causes of fatal car accidents in communities. This package contributes to road safety research by enabling reproducible analyses to be carried out, as the package provides a standardized method for acquiring and cleaning these data. Furthermore, the package makes an underutilized public dataset more usable and discoverable for the research community.

10. Related work

Packages of this class have proven highly valuable for R users seeking to study trends in road safety using crash records maintained by public authorities. The most similar package to njtr1 is the package stats19, which covers car accidents that took place within the United Kingdom [9]. Both njtr1 and stats19 provide comparable functionality and enable tackling similar research questions, with the key difference being that stats19 supports data acquisition for an entire country, whereas njtr1 is specifically targeted to support only New Jersey's data. It should also be noted that other developers have previously published ad-hoc scripts in other languages such as Python to download NJTR-1 data [10], but those prior approaches fall short of providing data cleaning & feature completeness of the njtr1 R package.

11. Limitations and future work

One key limitation of the package is that the data it can provide to users can only be as current as what is publicly released by the NJDOT. This results in a lag between when the latest available version of crash data was collected and when it became available for download. For example, at the time of writing in 2021, the latest year of NJTR-1 data available only covers up to the year 2019, and data for the past year, 2020 is not yet available, nor is any 2021 data. Considering this limitation, the package will not be suitable for short-term or real-time study of motor vehicle accidents but should be fairly comprehensive when used to pull data from the available years (2001–2019 as of November 2021).

It should also be noted that the usefulness of these data is largely going to depend on the quality of how it was gathered by police personnel who generated the data. There still may be typographical errors in the data, and not every accident record is geocoded, necessitating further work to refine the data cleaning process and extract further insights from the data, some of which may be attempted by adding new functionality to the package in the future.

A further limitation exists in the inconsistent manner in which police officers have entered citations to the legal statutes that constitute motor vehicle violations. The inconsistent usage has made it difficult to quantify the exact number of a particular motor vehicle violation as a result. This problem can be addressed via the development of a probabilistic parser or other tool to extract components of statute citations that appear in the data to label the components of a legal statute citation used to identify chargers against a driver.

12. Conclusion

This paper presented the njtr1 R package, which enables road safety research in New Jersey by presenting an interface to easily download, clean & analyze car crash data maintained by the New Jersey Department of Transportation for the R language & environment for statistical computing. This package is the first of its kind designed exclusively for studying road safety in New Jersey and can support future research into traffic safety in the state using R.

Declaration of competing interest

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

Appendix A. Supplementary code

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.simpa.2021.100176.Supplemental code used to generate the figures presented in this paper is included with the article.

References

[1] Insurance institute for highway safety, Fatality facts 2019 state by state, 2009, https://www.iihs.org/topics/fatality-statistics/detail/state-by-state. (Accessed 9 November 2021).

[2] New jersey department of transportation, 2001 to current crash tables, crash records, reference/links, 2021, https://www.state.nj.us/transportation/refdata/accident/rawdata01-current.shtm. (Accessed 19 October 2021).

[3] R Core Team, R: A language and environment for statistical computing, 2021, https://www.r-project.org/.

[4] New Jersey Department Of Transportation, Master file layout, crash records, reference/links, 2021, https://www.state.nj.us/transportation/refdata/accident/masterfile.shtm. (Accessed 9 November 2021).

[5] M. Aristarán, M. Tigas, J.B. Merrill, J. Das, D. Frackman, T. Swicegood, Tabula: Extract tables from PDFs, 2018, https://tabula.technology/. (Accessed 9 November 2021).

[6] S. Firke, Simple tools for examining and cleaning dirty data [R package janitor version 2.1.0], 2021, https://cran.r-project.org/package=janitor. (Accessed 9 November 2021).

[7] G.C. Rozzi, Download, analyze & clean new jersey car crash data [R package njtr1 version 0.3.1], 2021, https://cran.r-project.org/package=njtr1. (Accessed 19 October 2021).

[8] A. Baddeley, R. Turner, spatstat: An r package for analyzing spatial point patterns, J. Stat. Softw. 12 (2005) 1–42, http://dx.doi.org/10.18637/JSS.V012.I06.

[9] R. Lovelace, M. Morgan, L. Hama, M. Padgham, stats19: A package for working with open road crash data, J. Open Source Softw. 4 (2019) 1181, http://dx.doi.org/10.21105/joss.01181.

[10] J. Reiser, NJToolbox/DownloadCrashData.py at master · johnjreiser/NJToolbox · GitHub, 2015, https://github.com/johnjreiser/NJToolbox/blob/master/dot_crash/DownloadCrashData.py. (Accessed 9 November 2021).