The Journal of Open Source Software

fars_cleaner: A Python package for downloading and pre-processing vehicle fatality data in the US

Mitchell Z. Abrams[1] and **Cameron R. Bass**[1]

1 Duke University, USA

DOI: 10.21105/joss.04678

Software

- Review
- Repository
- Archive

Editor: Mark A. Jensen

Reviewers:

- @ethanwhite
- @svburke

Submitted: 01 July 2022
Published: 26 October 2022

Summary

Historical vehicle safety analysis in the United States leans heavily on public datasets to determine factors leading to crash fatality. The Fatality Analysis Reporting System (FARS) is one such database collected by the National Highway Traffic Safety Administration (NHTSA) documenting all vehicle fatalities in the United States since 1975. The FARS dataset is used to inform safety decisions at the local, state and national levels, and provides key insights into the efficacy of changing vehicle and trafficway safety standards (National Highway Traffic Safety Administration, 2022). The FARS dataset is frequently used by vehicle safety researchers to track long-term trends in fatality outcomes. As the FARS data have evolved over time, the variable coding has changed, sometimes dramatically, in some cases making it difficult to compare and analyze data across decades.

The FARS dataset consists of up to 30 data files for each year, with hundreds of recorded variables for each crash. To decrease file storage requirements and simplify distribution of the data, the vast majority of data fields are coded numerically, and then converted by a user referencing the Analytical User's Manual provided by NHTSA (National Highway Traffic Safety Administration, 2022). As a continuously evolving database, the content of these data files and the specific numeric values associated with each field are constantly updated as needs change. Currently, researchers interested in exploring the data must manually download .zip files for each year of interest from the NHTSA website, and reference the annually-issued Analytical User's Manual to decode the downloaded files.

`fars_cleaner` is a Python package which aims to solve some of these issues. This package provides a simple API for downloading and pre-processing the FARS dataset in such a way that simplifies comparisons across time. Users can download the FARS data, and `fars_cleaner` delivers data to the user as Pandas DataFrames. Data are preprocessed within the subset of years requested, converting numerical values to categorical text fields. This reduces the burden on researchers seeking to utilize the FARS dataset.

Statement of need

The FARS dataset is in constant flux, placing a large burden on researchers seeking to conduct analyses of vehicle safety trends over many decades. This package simplifies the data intake and pre-processing process, leaving researchers with prepared pandas dataframes ready for any analysis with the FARS data. This package is similar in concept to the `stats-19` R package by Lovelace et al. (2019), but is developed for the US crash database.

`fars_cleaner` has been used in double-pair analyses of male and female relative fatality risk (Abrams & Bass, 2020, 2022a), as well as a matched study with multiple cause of death data in the US (Abrams & Bass, 2022b).

Usage

Downloading FARS data

The FARSFetcher class provides an interface to download and unzip selected years from the NHTSA FARS FTP server. The class uses pooch (Uieda et al., 2020) to download and unzip the selected files. By default, files are unzipped to your OS's cache directory.

```
from fars_cleaner import FARSFetcher

# Prepare for FARS file download, using the OS cache directory.
fetcher = FARSFetcher()
```

The user can optionally pass a download path for the data files, otherwise the fetcher defaults to the OS cache location. Passing project_dir will download files to project_dir/data/fars by default. This behavior can be overridden by setting cache_path as well. Setting cache_path alone provides a direct path to the directory you want to download files into.

```
from pathlib import Path
from fars_cleaner import FARSFetcher

SOME_PATH = Path(”/YOUR/PROJECT/PATH”)
# Prepare to download to /YOUR/PROJECT/PATH/data/fars
# This is the recommended usage.
fetcher = FARSFetcher(project_dir=SOME_PATH)

# Prepare to download to /YOUR/PROJECT/PATH/fars
cache_path = ”fars”
fetcher = FARSFetcher(project_dir=SOME_PATH, cache_path=cache_path)

cache_path = Path(”/SOME/TARGET/DIRECTORY”)
# Prepare to download directly to a specific directory.
fetcher = FARSFetcher(cache_path=cache_path)
```

Files can be downloaded in their entirety (data from 1975-2018), as a single year, or across a specified year range. Downloading all of the data can be quite time consuming. The download will simultaneously unzip the folders, and delete the zip files. Each zipped file will be unzipped and saved in a folder {YEAR}.unzip

```
# Fetch all data
fetcher.fetch_all()

# Fetch a single year
fetcher.fetch_single(1984)

# Fetch data in a year range (inclusive).
fetcher.fetch_subset(1999, 2007)
```

Processing FARS data

After defining the FARSFetcher instance, requested data can be passed to the load_pipeline method for preprocessing. load_pipeline returns fully preprocessed, concatenated pandas DataFrames for the year range requested, for the primary analysis files (Accident, Person, and Vehicle).

```
from fars_cleaner import FARSFetcher, load_pipeline

fetcher = FARSFetcher(project_dir=project_path, )
vehicles, accidents, people = load_pipeline(1975, 2020,
                                       fetcher=fetcher,
                                       first_run=True,
                                       target_folder=target_folder)
```

References

Abrams, M., & Bass, C. R. (2020). Female vs. Male Relative Fatality Risk in Fatal Crashes. *Proceedings of 2020 International Research Council on the Biomechanics of Injury*. http://www.ircobi.org/wordpress/downloads/irc20/pdf-files/13.pdf

Abrams, M., & Bass, C. R. (2022a). Female vs. Male Relative Fatality Risk in Fatal Crashes – United States and United Kingdom. *9th World Congress of Biomechanics*.

Abrams, M., & Bass, C. R. (2022b). Female vs. Male Relative Risk of Body System Injuries in Fatal and Non-Fatal Crashes. *Proceedings of 2022 International Research Council on the Biomechanics of Injury*. http://www.ircobi.org/wordpress/downloads/irc22/pdf-files/2212.pdf

Lovelace, R., Morgan, M., Hama, L., & Padgham, M. (2019). stats19: A package for working with open road crash data. *Journal of Open Source Software*, *4*(33), 1181. https://doi.org/10.21105/joss.01181

National Highway Traffic Safety Administration. (2022). *Fatality Analysis Reporting System (FARS) Analytical User's Manual, 1975-2020* (Report DOT HS 813 254).

Uieda, L., Soler, S. R., Rampin, R., Kemenade, H. van, Turk, M., Shapero, D., Banihirwe, A., & Leeman, J. (2020). Pooch: A friend to fetch your data files. *Journal of Open Source Software*, *5*(45), 1943. https://doi.org/10.21105/joss.01943