# pynhts: survey-weight-aware Python access to the US National Household Travel Survey

Seyed Saber Naseralavi[a,*], Akram Mazaheri[b]

[a]*Department of Civil Engineering, Shahid Bahonar University of Kerman, Kerman 7616914111, Iran*

[b]*Department of Civil and Environmental Engineering, Tarbiat Modares University, Tehran 1411944961, Iran*

**Abstract**

The US National Household Travel Survey (NHTS) is the main public source of US household travel data, and four waves (2001, 2009, 2017 and 2022) are freely available. Using them correctly is harder than it looks: weights and variance designs differ by wave, the same code can mean different things in different waves, skipped questions are stored as numbers, and trip weights give annual rather than daily totals. These details are rarely stated in papers, so published results are difficult to reproduce or to compare. pynhts is an open-source Python library that downloads and verifies the official files, converts them to typed Parquet, joins the four tables without silent errors, and computes estimates with the variance method of each wave (jackknife replicate weights for 2001, 2009 and 2017; Taylor linearisation for 2022). It adds validated crosswalks between waves, guards for top-coded and non-answer values, and a study module that turns a written research statement into an analysis file with a sample-flow table. Validation is part of the package: 142 of 175 figures printed in official reports reproduce completely, Taylor-series errors agree with R to 1e-9, and 22 published studies were rebuilt and 1,150 printed numbers graded. The rebuilding recovered unstated rules, such as the base of a percentage and the definition of a weekend, that move published values by several per cent. The package is young and has no external users yet; its contribution is a tested, documented way to state and check the assumptions of NHTS analyses.

*Keywords:* National Household Travel Survey, travel survey, replicate

*Corresponding author

*Email addresses:* saber_naseralavi@uk.ac.ir (Seyed Saber Naseralavi), akram.mazaheri.010@gmail.com (Akram Mazaheri)

weights, survey statistics, reproducibility, Python

## Required Metadata

### Current code version

| Nr. | Code metadata description | Please fill in this column |
|---|---|---|
| C1 | Current code version | v0.8.1 |
| C2 | Permanent GitHub link to code/repository used for this code version | `https://github.com/sabernaseralavi-60/pynhts` |
| C3 | Legal Code License | BSD 3-Clause License |
| C4 | Code versioning system used | git |
| C5 | Software code languages, tools, and services used | Python 3; NumPy, pandas, PyArrow, pooch, PyYAML; optional: matplotlib, PyTorch |
| C6 | Compilation requirements, operating environments & dependencies | Python 3.10 or newer on Linux, Windows or macOS (continuous integration on Linux and Windows); about 0.4 GB of Parquet per wave and 1.8 GB of raw CSV for 2017, which can be deleted after conversion |
| C7 | If available Link to developer documentation/manual | `https://github.com/sabernaseralavi-60/pynhts/tree/main/docs` |
| C8 | Support email for questions | saber_naseralavi@uk.ac.ir |

Table 1: Code metadata (mandatory)

## Main text

### 1. Motivation and significance

The National Household Travel Survey (NHTS), sponsored by the US Federal Highway Administration (FHWA), is the main public source of information on how US households travel: who makes trips, by which mode, for which purpose and over what distance [1]. Public-use files exist for 2001, 2009, 2017 and 2022, each with four linked tables (households, persons, trips, vehicles), sampling weights and, for three waves,

replicate weights. The files are used for many topics. Recent examples are access to health care [2], the mobility of older adults across waves [3, 4, 5], taxi and ridehailing use [6, 7], ridesharing and vehicle ownership [8], alternative fuel vehicles [9], vehicle miles travelled [10, 11], active travel [12], online shopping [13] and household trip generation by state [14].

Working with the files is less simple than the number of such studies suggests. Several authors have shown how the survey itself shapes results: respondents who agree to be contacted again differ from those who do not [15], proxy responses contain more rounded departure and arrival times [16], and a representation measure built from sampling weights found under-representation of vulnerable traveller groups in 43 US household travel surveys, the 2017 NHTS among them [17]. Beyond the properties of the survey, correct use requires work that every project repeats and that is easy to get wrong. A skipped question is stored as the number -1, so the mean of the 2009 vehicle miles driven (VMT_MILE) is 5.36 instead of 9.72 and its total is 7 % low. The same code means different things in different waves: transport mode "01" is a car in 2009, where it is 44 % of trips, and walking in 2017, where it is 10.5 %. Trip weights add up to annual, not daily, totals; the 2001 file contains children aged 0–4 and the others do not; ages and vehicle ages are top-coded; and 2022 uses a different variance design from the earlier waves. Such mistakes change estimates and, more often, understate their uncertainty. Papers rarely state the rules they applied, which makes their numbers hard to reproduce.

This matters beyond one survey. Calls for reproducible transportation research ask for open code and documented methods [18, 19], and name insufficient methodological detail and the lack of code sharing among the barriers [19]; agent-based travel-demand research is rarely replicable when it rests on proprietary data and tools [20]. In the wider research-software community, the FAIR principles have been adapted to software [21, 22] and principles for software citation have been proposed [23]. pynhts follows them in practice: every release is archived with a persistent identifier and a machine-readable citation file.

pynhts addresses these problems in one Python package. It (i) fetches and verifies the official files and exposes them as typed tables that cannot be merged wrongly, (ii) computes every estimate with the variance method of its wave, (iii) maps codes between waves only where the meaning is shown to be the same, (iv) requires the analyst to state sample rules, missing-value policy and weighting, and (v) ships its own validation against published numbers. Scientifically, the contribution

is not a new estimator but a way to make the assumptions of an NHTS analysis explicit, testable and reusable. A user installs the package, downloads a wave with one command, loads it lazily, and asks for weighted statistics, rates, harmonised categories or a complete study (Section 3). Results carry the wave, weight, variance design and unit, so a table footnote can be written from the result itself.

Related work. The R package summarizeNHTS [24] downloads and summarises NHTS data, and the R survey package [25] offers general replicate-weight and Taylor-series analysis that users can apply to NHTS data after setting up the design themselves. The Python package samplics [26] covers complex-sample estimation in general but does not know the NHTS files, codes or weights. A search of GitHub and PyPI on 22 October 2026 (queries "NHTS python" and "national household travel survey", and the package names nhts, nhts-py and summarizenhts) found analysis scripts for single projects and no other general Python package for NHTS access. The variance estimators follow the standard replicate-weight theory [27] and the weighting report of the 2017 survey [28].

## 2. Software description

### 2.1 Software architecture.

pynhts is organised in layers (Fig. 1). The input and statistics layers handle the data; the harmonisation and analysis layers help the analyst to ask correct questions; the validation layer checks the result against published numbers.

pynhts.io keeps a registry with the SHA-256 hash of every official archive and member, downloads with pooch, and writes Parquet deterministically. The data server answers HTTP 200 with an HTML page for files that do not exist, so content type and size are checked. pynhts.core derives a typed schema from the codebook (all 3,807 coded values of 2017 and all 2,639 of 2022 reproduce the printed frequencies) and loads columns lazily, joining each requested column from the table that owns it through a validated many-to-one join. The skip code is converted to a missing value in numeric columns by default, and non-answer codes are masked on request.

pynhts.stats implements the stratified jackknife of the 2017 weighting design. With 14 variance strata of 7 units, the variance of an estimate is $6/7$ times the sum over the 98 replicates of the squared deviation of the replicate estimate from the full estimate, and the margin of error (MOE) is the t quantile with 98 degrees of freedom times the standard error [27, 28]:

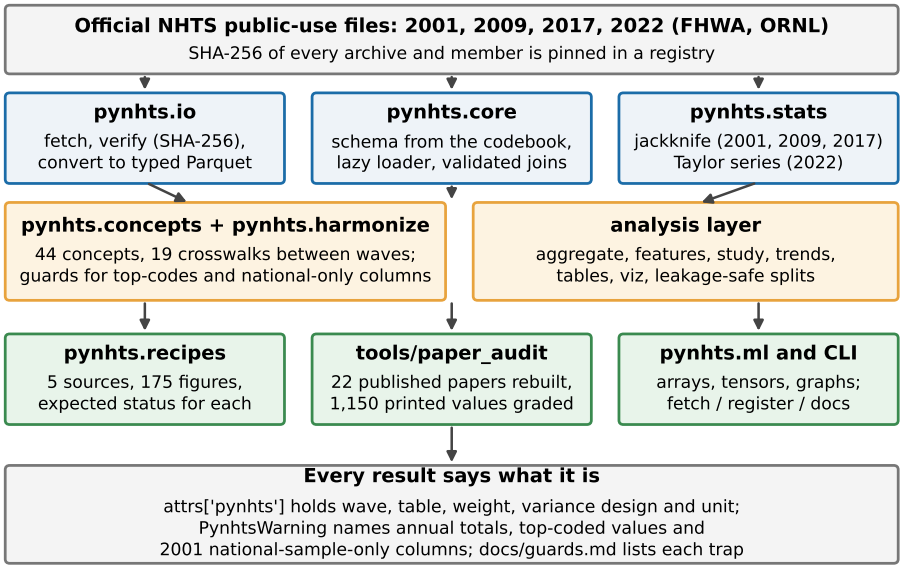



Figure 1: Architecture of pynhts. Blue: data access and statistics; amber: harmonisation and analysis; green: validation and bridges. Every result frame records how it was computed.

$$\hat{V}(\hat{\theta}) = \frac{6}{7}\sum_{r=1}^{98}(\hat{\theta}_r - \hat{\theta})^2, \qquad \mathrm{MOE} = t_{0.975,\,98}\,\sqrt{\hat{V}(\hat{\theta})}. \tag{1}$$

The 2009 and 2001 files have 100 and 99 replicates with their own factors. The 2022 file has no replicate weights, so pynhts uses Taylor linearisation with strata and households; its errors equal those of the R survey package to 1e-9. All trips of a person share one weight, so trip estimates are computed on person-day units and need a 98-column matrix for 264,234 units rather than 923,572 trips. Subpopulations are handled as domains, never by dropping rows.

pynhts.concepts and pynhts.harmonize describe 44 concepts across the four waves, each graded as comparable, comparable with care, or not comparable, and provide 19 crosswalks (trip purpose, mode, race and Hispanic origin, medical condition, and others). A crosswalk reads the wave from the frame and refuses a mismatch; each category is marked exact, approximate or nested, with the evidence in the data. The analysis layer contains aggregate (trips to persons or households with an explicit missing-value policy

and true zeros), features (93 documented variables with lineage), study (below), trends (per-wave estimates; pooling is refused for non-comparable concepts), transform (splits by household so that train and test sets do not share a household) and tables and viz for reporting. pynhts.recipes encodes the numbers printed in official reports together with the expected status of each, and a generated report lists every gap. A small bridge exports tables to NumPy arrays, PyTorch tensors and household graphs.

2.2 **Software functionalities.** The main functions are: verified download and offline use (pynhts fetch, register, docs); typed lazy loading of the four tables of each wave; survey totals, means, ratios and rates with standard errors and margins of error (rates.daily_rate gives daily rates and states when a total is annual); crosswalks between waves and a combined race-and-ethnicity variable; warnings for top-coded values, national-sample-only columns and caps; and the study module. A Study object takes the wave, the unit of analysis, ordered sample rules (Keep, Drop, Complete), variables, the weight (or an explicit none) and stated assumptions. It returns the analysis file, a sample-flow table with the number removed by each rule, a descriptive table, a methods paragraph and a comparison with published numbers graded exact, close, acceptable or gap. It never tunes a rule to reach a number. Table 2 lists every module with its function. Two aids are meant for assistants that write analysis code: AGENTS.md states the ten rules that prevent the silent errors above, and agent_eval grades an answers file on 14 tasks that contain those traps.

## 3. Illustrative examples

Example 1 produces national estimates of the 2017 survey with design-correct margins of error.

```
import pynhts
from pynhts import rates
d = pynhts.load(2017)
s = d.survey("person", ["DRIVER"])
# 223.3 million drivers, MOE 0.83 million
s.total(domain=s.rows["DRIVER"] == "01")
# 10.70 miles per trip, MOE 0.40
d.survey("trip", ["TRPMILES"]).mean("TRPMILES")
# 3.37 daily trips per person, MOE 0.04
rates.daily_rate(d, "trips", per="person")
```

These values equal the figures printed in the official summary report [1], including the margins of error.

Table 2: Modules of pynhts 0.8.1 and what each does.

| Module | Function |
|---|---|
| `pynhts.io` | Verified download (SHA-256 of every archive), registration of files already on disk, deterministic Parquet cache. |
| `pynhts.core` | Typed schema from the codebook, lazy loader with validated joins, missing-code policy, top-code and national-sample-only flags. |
| `pynhts.stats` | Totals, means, proportions and ratios with jackknife (2001, 2009, 2017) or Taylor-series (2022) errors; subpopulations as domains. |
| `pynhts.rates` | Daily trips and miles per person or household by the official route, with the unit stated. |
| `pynhts.concepts` | 44 concepts in four waves: universe, unit, comparability grade, pitfalls. |
| `pynhts.harmonize` | Comparison of two waves; 19 crosswalks (mode, purpose, income, race, Hispanic origin, medical condition, and others) with the status of each code. |
| `pynhts.aggregate, features` | Trips to persons and households with a stated missing-value policy; 93 documented features with formula and lineage. |
| `pynhts.eda, quality` | Reasons for missing values, weighted descriptives, key and weight integrity report. |
| `pynhts.transform, trends` | Leakage-safe transformations and split by household; per-wave estimates and guarded pooling. |
| `pynhts.tours` | Home-based work, other and non-home-based tours and FHWA-style trip chains (2017 only). |
| `pynhts.study` | Research statement to analysis file, sample-flow table, descriptive table, methods text and grading. |
| `pynhts.tables, viz` | Tables as LaTeX, Markdown, HTML or Excel with a methods footnote; plots, including activity diaries. |
| `pynhts.recipes` | Published tables with the expected status of each figure and a generated gap report. |
| `pynhts.ml` | Arrays, PyTorch tensors and household graphs. |
| `Command line` | pynhts fetch, register and docs (official documents, hash-checked). |

Example 2 shows why a stated base matters. Lidbe et al. [5] report that 5.49 % of 2001 trips were made by persons with a medical condition that makes travel difficult. The question was asked only of persons aged 16 or older; younger persons carry the skip code. Counting them as having no condition gives 4.40 %; using the trips of persons who answered gives the printed value. The crosswalk returns a missing value for the skip, so the second base is the natural one.

```
from pynhts import harmonize
d = pynhts.load(2001)
p = d.persons(["MEDCOND"], missing="keep", skip="keep")
```

```
p["med"] = harmonize.apply_column(
    p, "MEDCOND", "medical_condition")
# the skip and the non-answers are now missing, not "no"
k = ["HOUSEID", "PERSONID"]
t = d.trips(["TRPMILES"]).merge(p[k + ["med"]], on=k)
w = t["WTTRDFIN"]; yes = t["med"].eq("yes")
# 4.399: all trips as the base
100 * w[yes].sum() / w.sum()
# 5.489: trips of persons who answered
100 * w[yes].sum() / w[t["med"].notna()].sum()
```

Example 3 states a whole study as data. The sample rules of Zhang and Zhang [8] (15 low-use states excluded, only persons who answered the ridesharing question, complete model variables, no weights) become a Study, which returns the sample-flow table and the descriptive table. The result has 218,831 persons against 219,026 printed (0.09 % fewer), after removing 9,142, 27,090 and 9,171 persons by the three rules from the 264,234 in the file, and the mean number of ridesharing uses in 30 days is 0.309 as printed.

```
from pynhts import study as S
spec = S.Study.from_yaml(
    "docs/examples/ridesharing_35_states.yaml")
res = spec.run(pynhts.load(2017))
res.flow     # persons before and after each rule
res.table()  # mean, sd, min, max, share of zeros
S.compare(res, [{"variable": "ride", "stat": "mean",
                 "value": 0.309}])   # EXACT
```

## 4. Impact

### Validation against official figures

Five recipes encode 175 figures printed in official documents: Tables 1 and 3 of the Summary of Travel Trends for 2001, 2009 and 2017, the control totals of the 2022 user guide, and a transit-tour study. Of these, 142 reproduce completely, 8 reproduce the estimate but not the printed margin of error (the report uses another multiplier), and 25 do not; each gap is explained in the generated report where a cause is known, and no rule was changed to hide one. For 2022, the Taylor-series errors equal those of the R survey package. There are 333 unit tests that need no data and more than 600 tests when the real files are present, run on Linux and Windows.

### Rebuilding published studies

To test whether the package supports real analyses, the data of 22 published NHTS studies were rebuilt with the package only, one script per study, and every number the study prints about its data was graded against the result: exact (equal at the printed precision), close (within 2 %), acceptable (within 5 %) or gap. No rule was adjusted to reach a number. Of 1,150 graded values, 618 are exact, 279 close, 105 acceptable and 148 gaps (Fig. 2). In nine studies the analysis data could be rebuilt convincingly; in the others the printed numbers depend on sample rules or external variables that the paper does not state or that the public files do not contain. We treat each gap as a statement about the information available in the paper, not about its authors' work.

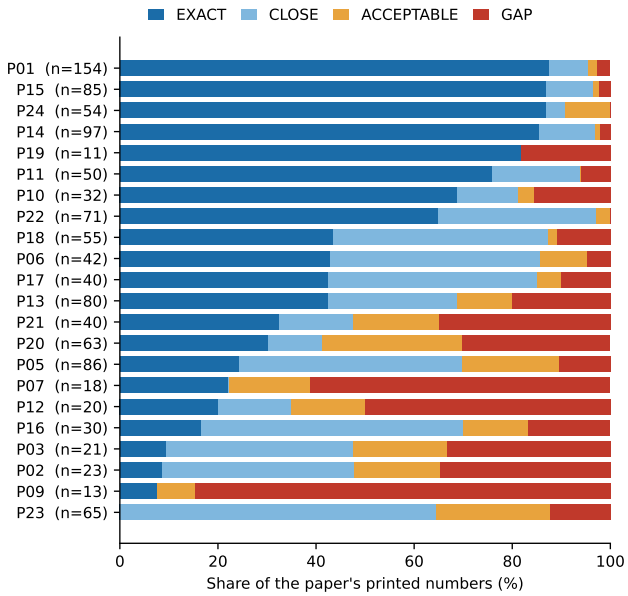



Figure 2: Outcome of the audit of 22 published NHTS studies (P01-P24; listed in the repository, first-audit scripts). Bars show the share of each study's printed numbers that pynhts reproduces exactly, closely (2 %), acceptably (5 %) or not.

One of the 22 studies is the authors' own earlier work on household trip generation by state [14], so that case is not independent. It states its cleaning only as removal of incomplete records and outliers. The rules that its printed tables imply (at least one vehicle and one trip, valid income, race, life cycle and tenure) leave 103,317 households against 106,287 printed (2.8 % fewer); of its 54 checked values, 47 are exact at the printed precision, 2 close and 5 acceptable. The rebuilt sample also shows that 33 of the 51 states contribute fewer than 500 households (median 312), which bears on any model fitted state by state.

The exercise produced 33 recovered conventions, written down with their evidence, and several of them change published numbers. A recheck of one study [5] with the crosswalks and the study module illustrates this: 262 of 280 cells of its Tables 3 and 4 are reproduced exactly (Fig. 3). The percentages in its Table 3 use the trips of persons with a valid answer as the base (Example 2); its weekend is Saturday and Sunday of the travel-day variable, whereas the NHTS weekend flag also contains part of Friday and gives values 5 to 8 % higher; trip duration is the calculated minutes variable; the six items on how the condition changes travel are reproduced exactly (72 of 72 cells) through the condition-effect crosswalk; and its race groups follow a rule in which Hispanic origin overrides only the White, Black and Asian groups, which is not what a combined race-and-ethnicity variable does. In another study [12] the sample of 130,222 persons is reproduced exactly as a Study, and its race variable is the household respondent's single race code; mapping it through the combined variable would give a different sample (133,538 persons). These findings became package features (crosswalks, caps, warnings) or documentation.

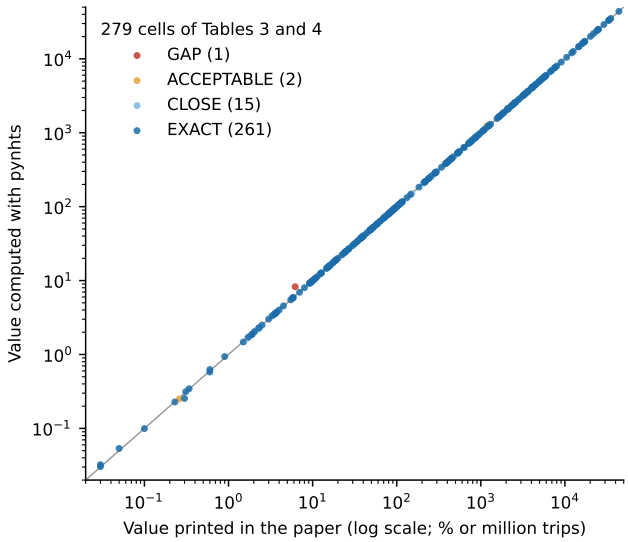



Figure 3: Values printed in Tables 3 and 4 of Lidbe et al. [5] against the values computed with pynhts (279 cells with a positive value; one printed zero is not shown). The diagonal is exact agreement.

### New research questions and changed practice

Design-correct uncertainty for every wave, guarded comparisons across

waves, and sample-flow tables make it cheaper to ask questions that depend on trends and on subpopulations, and to report them so that others can repeat them. The study module is meant to change the daily practice of analysts by making the sample rules and the missing-value policy part of the code. We have not measured any change in practice yet.

**Extent of use**

The first version was released on PyPI in October 2026 and the package has had 13 releases in total; version 0.8.1 is archived on Zenodo [29]. The package has no known external users, no publication has used it, and it has no commercial use. PyPI records downloads, but they include automated mirrors, so we do not report them as users. The evidence of impact is therefore the validation and the audit, which can be repeated by anyone with the public files, and the intended use is by researchers who analyse the NHTS and by those who review such analyses.

5. **Conclusions**

pynhts gives researchers typed, survey-weighted access to the four NHTS waves with the variance method of each wave, crosswalks that are checked against the data, and a way to state the sample and the missing-value policy as code. Its validation against official figures and published studies shows both that it reproduces most numbers and that the remaining differences usually come from rules that publications do not state.

Limitations are clear. The package is at an alpha stage and its interface may change. It handles the public-use files only; variables that need the restricted files (for example census-tract identifiers) are out of reach. It does not fit models, so standard errors of regression coefficients, which need the weights to be refitted for each replicate, are not yet provided. National weights of the 2001 national sample are not available. The replicate-weight archives of 2001 and 2009 are large. Next steps are variance estimates for model coefficients, more crosswalks, and a repeated audit as new studies appear.

## Acknowledgements

The NHTS is conducted by the US Department of Transportation, Federal Highway Administration, and distributed by Oak Ridge National Laboratory. The package does not redistribute the survey data.

## Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work the authors used Claude (Anthropic) as an assistant for software design, writing and running code and tests, checking the code against the official documents and data, literature screening, and drafting and editing the text. After using this tool, the authors reviewed, verified and edited the material (code, results, references and text) and take full responsibility for the content of the publication. Numbers in the text are taken by code from the result files of the repository, and the references were checked against a bibliographic database.

## CRediT authorship contribution statement

Seyed Saber Naseralavi: Project administration, Supervision. Akram Mazaheri: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Validation, Visualization, Writing - original draft, Writing - review and editing.

## Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

## 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.

## Data availability

The software is available at https://github.com/sabernaseralavi-60/pynhts (archived at Zenodo, DOI 10.5281/zenodo.23575312). The NHTS public-use data are distributed by Oak Ridge National Laboratory (https://nhts.ornl.gov); pynhts downloads them from there and does not redistribute them.

## References


[1] McGuckin N, Fucci A. Summary of travel trends: 2017 National Household Travel Survey. Report FHWA-PL-18-019. Washington (DC): US Department of Transportation, Federal Highway Administration; 2018. https://nhts.ornl.gov/assets/2017_nhts_summary_travel_trends.pdf [accessed 22 October 2026].

[2] Labban M, Chen CR, Frego N, Nguyen DD, Lipsitz SR, Reich AJ, et al. Disparities in travel-related barriers to accessing health care from the 2017 National Household Travel Survey. JAMA Netw Open 2023;6:e2325291. https://doi.org/10.1001/jamanetworkopen.2023.25291.

[3] Buehler R, Pucher J, Wittwer R, Gerike R. Travel behavior of older adults in the USA, 2001-2017. Travel Behav Soc 2024;36:100783. https://doi.org/10.1016/j.tbs.2024.100783.

[4] Akinlotan M, Khodakarami N, Primm K, Bolin J, Ferdinand AO. Travel for medical or dental care by race/ethnicity and rurality in the U.S.: findings from the 2001, 2009 and 2017 National Household Travel Surveys. Prev Med Rep 2023;35:102297. https://doi.org/10.1016/j.pmedr.2023.102297.

[5] Lidbe A, Adanu EK, Penmetsa P, Jones S. Changes in the travel patterns of older Americans with medical conditions: a comparison of 2001 and 2017 NHTS data. Transp Res Interdiscip Perspect 2021;11:100463. https://doi.org/10.1016/j.trip.2021.100463.

[6] Conway MW, Salon D, King DA. Trends in taxi use and the advent of ridehailing, 1995-2017: evidence from the US National Household Travel Survey. Urban Sci 2018;2:79. https://doi.org/10.3390/urbansci2030079.

[7] Rafiq R, McNally MG. An exploratory analysis of alternative travel behaviors of ride-hailing users. Transportation 2023;50:571-605. https://doi.org/10.1007/s11116-021-10254-9.

[8] Zhang Y, Zhang Y. Examining the relationship between household vehicle ownership and ridesharing behaviors in the United States. Sustainability 2018;10:2720. https://doi.org/10.3390/su10082720.

[9] Li X, Liu C, Jia J. Ownership and usage analysis of alternative fuel vehicles in the United States with the 2017 National Household Travel

Survey data. Sustainability 2019;11:2262. https://doi.org/10.3390/su11082262.

[10] Alberini A, Teja Burra L, Cirillo C, Shen C. Counting vehicle miles traveled: what can we learn from the NHTS? Transp Res Part D Transp Environ 2021;98:102984. https://doi.org/10.1016/j.trd.2021.102984.

[11] Wu X, MacKenzie D. Assessing the VMT effect of ridesourcing services in the US. Transp Res Part D Transp Environ 2021;94:102816. https://doi.org/10.1016/j.trd.2021.102816.

[12] Yang Y, Diez Roux AV, Bingham CR. Variability and seasonality of active transportation in USA: evidence from the 2001 NHTS. Int J Behav Nutr Phys Act 2011;8:96. https://doi.org/10.1186/1479-5868-8-96.

[13] Zhou Y, Wang X(C). Explore the relationship between online shopping and shopping trips: an analysis with the 2009 NHTS data. Transp Res Part A Policy Pract 2014;70:1-9. https://doi.org/10.1016/j.tra.2014.09.014.

[14] Naseralavi S, Soltanirad M, Ranjbar E, Lucero M, Gorzin F, Hakiminejad Y, et al. Machine learning modeling of household trip generation by state using NHTS data. Urban Sci 2025;9:353. https://doi.org/10.3390/urbansci9090353.

[15] Wang X, Shaw FA, Mokhtarian PL, Watkins KE. Response willingness in consecutive travel surveys: an investigation based on the National Household Travel Survey using a sample selection model. Transportation 2023;50:2339-73. https://doi.org/10.1007/s11116-022-10312-w.

[16] Yoshikawa S, Maruyama T. Proxy responses contain more rounded trip departure/arrival times: evidence from the 2017 US National Household Travel Survey. J JSCE 2023;11. https://doi.org/10.2208/journalofjsce.23-00137.

[17] Esekhaigbe EO, Bills TS, Broaddus A. Group representation of vulnerable travelers in household travel surveys: an assessment of 43 U.S. surveys. Transportation 2026. https://doi.org/10.1007/s11116-026-10795-x.

[18] Zheng Z. Reasons, challenges, and some tools for doing reproducible transportation research. Commun Transp Res 2021;1:100004. https://doi.org/10.1016/j.commtr.2021.100004.

[19] Wood J, van Schalkwyk I. Reproducibility in transportation research: importance, best practices, and dealing with protected and sensitive data. J Transp Technol 2025;15:179-202. https://doi.org/10.4236/jtts.2025.151010.

[20] Hörl S, Balać M. Synthetic population and travel demand for Paris and Île-de-France based on open and publicly available data. Transp Res Part C Emerg Technol 2021;130:103291. https://doi.org/10.1016/j.trc.2021.103291.

[21] Lamprecht AL, Garcia L, Kuzak M, Martinez C, Arcila R, Martin Del Pico E, et al. Towards FAIR principles for research software. Data Sci 2020;3:37-59. https://doi.org/10.3233/ds-190026.

[22] Barker M, Chue Hong NP, Katz DS, Lamprecht AL, Martinez-Ortiz C, Psomopoulos F, et al. Introducing the FAIR principles for research software. Sci Data 2022;9:622. https://doi.org/10.1038/s41597-022-01710-x.

[23] Smith AM, Katz DS, Niemeyer KE; FORCE11 Software Citation Working Group. Software citation principles. PeerJ Comput Sci 2016;2:e86. https://doi.org/10.7717/peerj-cs.86.

[24] Westat Transportation. summarizeNHTS: R package for summarizing NHTS data [software]. https://github.com/Westat-Transportation/summarizeNHTS [accessed 22 October 2026].

[25] Lumley T. Analysis of complex survey samples. J Stat Softw 2004;9(8). https://doi.org/10.18637/jss.v009.i08.

[26] Diallo M. samplics: a Python package for selecting, weighting and analyzing data from complex sampling designs. J Open Source Softw 2021;6:3376. https://doi.org/10.21105/joss.03376.

[27] Wolter KM. Introduction to variance estimation. 2nd ed. New York: Springer; 2007. https://doi.org/10.1007/978-0-387-35099-8.

[28] Roth SB, DeMatteis J, Dai Y. 2017 NHTS weighting report. Westat, prepared for the Federal Highway Administration; 2017. https://nhts.ornl.gov/assets/2017%20NHTS%20Weighting%20Report.pdf [accessed 22 October 2026].

[29] Naseralavi SS. pynhts: survey-weight-aware access to US National Household Travel Survey microdata. Version 0.8.1 [software]. Zenodo; 2026 Oct 22. https://doi.org/10.5281/zenodo.23575312.

### Current executable software version

Not applicable: pynhts is a Python library without a separate executable. Install version 0.8.1 with pip install pynhts==0.8.1.