pynhts: دسترسی پایتونیِ آگاه از وزن‌های نمونه‌گیری به پیمایش ملی سفر خانوار ایالات متحده

سید صابر ناصرعلوی*[a]، اکرم مظاهری[b]

[a]گروه مهندسی عمران، دانشگاه شهید باهنر کرمان، کرمان ۷۶۱۶۹۱۴۱۱۱، ایران

[b]گروه مهندسی عمران و محیط زیست، دانشگاه تربیت مدرس، تهران ۱۴۱۱۹۴۴۹۶۱، ایران

* نویسنده‌ی مسئول: saber_naseralavi@uk.ac.ir

چکیده

پیمایش ملی سفر خانوار ایالات متحده (NHTS) مهم‌ترین منبع عمومی داده درباره‌ی سفر خانوارهای آمریکایی است و چهار موج آن (۲۰۰۱، ۲۰۰۹، ۲۰۱۷ و ۲۰۲۲) آزادانه در دسترس است. استفاده‌ی درست از این داده‌ها دشوارتر از آن است که به نظر می‌رسد: وزن‌ها و طرح واریانس در هر موج فرق می‌کند، یک کد در موج‌های گوناگون معنای دیگری دارد، پرسش‌های پرش‌شده به‌صورت عدد ذخیره می‌شوند و وزن سفرها جمع سالانه می‌دهد نه روزانه. این جزئیات به‌ندرت در مقاله‌ها بیان می‌شود و به همین دلیل بازتولید یا مقایسه‌ی نتیجه‌های منتشرشده دشوار است. pynhts کتابخانه‌ای متن‌باز در پایتون است که پرونده‌های رسمی را دریافت و راستی‌آزمایی می‌کند، آن‌ها را به Parquet نوع‌دار تبدیل می‌کند، چهار جدول را بدون خطای خاموش به هم می‌پیوندد و برآوردها را با روش واریانس هر موج می‌سنجد (وزن‌های تکرارِ جک‌نایف برای ۲۰۰۱، ۲۰۰۹ و ۲۰۱۷ و خطی‌سازی تیلور برای ۲۰۲۲). افزون بر آن جدول‌های تناظر اعتبارسنجی‌شده میان موج‌ها، نگهبان‌هایی برای مقدارهای سقف‌گذاری‌شده و بی‌پاسخ، و ماژولی دارد که بیانیه‌ی پژوهشی نوشته‌شده را به پرونده‌ی تحلیل و جدول جریان نمونه تبدیل می‌کند. اعتبارسنجی بخشی از بسته است: ۱۴۲ عدد از ۱۷۵ عدد چاپ‌شده در گزارش‌های رسمی کاملاً بازتولید می‌شود، خطاهای استاندارد سری تیلور با R تا 1e-9 برابر است، و ۲۲ مطالعه‌ی منتشرشده بازسازی و ۱٬۱۵۰ عدد چاپ‌شده‌ی آن‌ها نمره‌گذاری شد. این بازسازی قاعده‌های نانوشته‌ای مانند مخرج یک درصد و تعریف آخر هفته را آشکار کرد که مقدارهای منتشرشده را چند درصد جابه‌جا می‌کنند. بسته هنوز جوان است و کاربر بیرونی ندارد؛ سهم آن راهی آزموده و مستند برای بیان و وارسی فرض‌های تحلیل‌های NHTS است.

واژه‌های کلیدی: پیمایش ملی سفر خانوار، پیمایش سفر، وزن‌های تکرار، آمار پیمایشی، بازتولیدپذیری، پایتون

فراداده‌های ضروری

شماره	توضیح فراداده	مقدار
C۱	نسخه‌ی کنونی کد	v0.8.1
C۲	پیوند دائمی گیت‌هاب برای این نسخه	https://github.com/sabernaseralavi-60/pynhts
C۳	پروانه‌ی کد	پروانه‌ی BSD 3-Clause
C۴	سامانه‌ی نسخه‌بندی	git
C۵	زبان‌ها، ابزارها و خدمات	پایتون ۳؛ NumPy، pandas، PyArrow، pooch، PyYAML؛ اختیاری: matplotlib، PyTorch
C۶	نیازها و محیط‌های اجرا	پایتون ۳٫۱۰ یا بالاتر روی لینوکس، ویندوز یا مک (یکپارچه‌سازی

		پیوسته روی لینوکس و ویندوز)؛ برای هر موج حدود ۰/۴ گیگابایت Parquet و برای ۲۰۱۷ حدود ۱/۸ گیگابایت CSV خام که پس از تبدیل می‌توان حذف کرد
C۷	پیوند مستندات توسعه‌دهنده	https://github.com/sabernaseralavi-60/pynhts/tree/main/docs
C۸	نشانی پشتیبانی	saber_naseralavi@uk.ac.ir

جدول ۱. فراداده‌های کد (الزامی)

متن اصلی

۱. انگیزه و اهمیت

پیمایش ملی سفر خانوار (NHTS) که اداره فدرال بزرگراه‌های ایالات متحده (FHWA) آن را حمایت می‌کند، مهم‌ترین منبع عمومی درباره‌ی چگونگی سفر خانوارهای آمریکایی است: چه کسی سفر می‌کند، با کدام وسیله، برای چه هدفی و تا چه مسافتی [1]. پرونده‌های عمومی برای ۲۰۰۱، ۲۰۰۹، ۲۰۱۷ و ۲۰۲۲ در دسترس است و هر یک چهار جدول پیوسته (خانوار، فرد، سفر، خودرو)، وزن‌های نمونه‌گیری و، برای سه موج، وزن‌های تکرار دارد. این پرونده‌ها در موضوع‌های بسیاری به کار می‌روند؛ نمونه‌های تازه عبارت‌اند از دسترسی به مراقبت سلامت [2]، تحرک سالمندان در موج‌های گوناگون [3-5]، استفاده از تاکسی و خودروهای درخواستی [6,7]، سواری اشتراکی و مالکیت خودرو [8]، خودروهای با سوخت جایگزین [9]، مایل‌های پیموده‌شده با خودرو [10,11]، جابه‌جایی فعال [12]، خرید اینترنتی [13] و تولید سفر خانوار به تفکیک ایالت [14].

کار با این پرونده‌ها ساده‌تر از آن نیست که شمار این مطالعه‌ها نشان می‌دهد. چند نویسنده نشان داده‌اند که خود پیمایش بر نتیجه‌ها اثر می‌گذارد: پاسخ‌گویانی که موافق تماس دوباره می‌شوند با دیگران فرق دارند [15]، پاسخ‌های نیابتی زمان‌های حرکت و رسیدن گردشده‌تری دارند [16]، و شاخصی که از وزن‌های نمونه‌گیری ساخته شده بود در ۴۳ پیمایش سفر خانوار آمریکا، از جمله NHTS ۲۰۱۷، کم‌نمایی گروه‌های آسیب‌پذیر مسافر را نشان داد [17]. افزون بر ویژگی‌های خود پیمایش، کاربرد درست کارهایی لازم دارد که هر پروژه تکرار می‌کند و اشتباه در آن‌ها آسان است. پرسشِ پرش‌شده به‌صورت عدد 1- ذخیره می‌شود؛ پس میانگین مایل‌های رانده‌شده‌ی خودرو در ۲۰۰۹ (VMT_MILE) به‌جای ۹/۷۲ برابر ۵/۳۶ درمی‌آید و جمع آن ۷٪ کم می‌شود. یک کد در موج‌های گوناگون معنای دیگری دارد: وسیله‌ی «۰۱» در ۲۰۰۹ خودرو است و ۴۴٪ سفرها را می‌سازد، ولی در ۲۰۱۷ پیاده‌روی است و ۱۰/۵٪ سفرها. وزن سفرها جمع سالانه می‌دهد نه روزانه؛ پرونده‌ی ۲۰۰۱ کودکان ۰ تا ۴ ساله را دارد و بقیه ندارند؛ سن و سن خودرو سقف‌گذاری شده است؛ و ۲۰۲۲ طرح واریانسی دارد که با موج‌های پیشین فرق می‌کند. چنین اشتباه‌هایی برآورد را عوض می‌کنند و بیشتر وقت‌ها عدم‌قطعیت آن را کمتر از واقع نشان می‌دهند. مقاله‌ها به‌ندرت قاعده‌هایی را که به کار برده‌اند می‌نویسند و همین بازتولید عددهایشان را دشوار می‌کند.

اهمیت این موضوع از یک پیمایش فراتر می‌رود. فراخوان‌ها برای پژوهش بازتولیدپذیر در حمل‌ونقل کد باز و روش‌های مستند می‌خواهند [18,19] و کمبود جزئیات روش و به‌اشتراک‌نگذاشتن کد را از موانع می‌شمارند [19]؛ پژوهش تقاضای سفر مبتنی بر عامل نیز وقتی بر داده و ابزار اختصاصی تکیه دارد به‌ندرت بازتولیدپذیر است [20]. در جامعه‌ی گسترده‌تر

نرم‌افزار پژوهشی، اصول FAIR برای نرم‌افزار سازگار شده [21,22] و اصولی برای ارجاع به نرم‌افزار پیشنهاد شده است [23]. pynhts در عمل از آن‌ها پیروی می‌کند: هر انتشار با شناسه‌ی پایدار و پرونده‌ی ارجاع ماشین‌خوان بایگانی می‌شود.

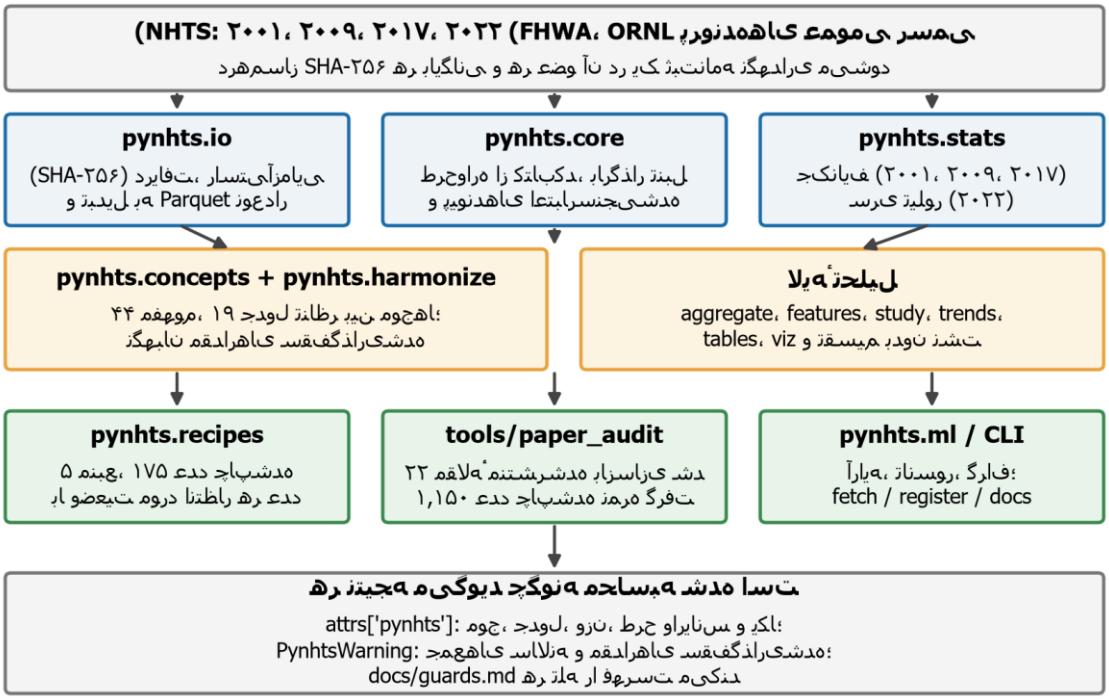
pynhts این مسئله‌ها را در یک بسته‌ی پایتونی پاسخ می‌دهد. (i) پرونده‌های رسمی را دریافت و راستی‌آزمایی می‌کند و آن‌ها را به‌صورت جدول‌های نوع‌دار عرضه می‌کند که نمی‌توان غلط به هم پیوستشان؛ (ii) هر برآورد را با روش واریانس موج خودش می‌سنجد؛ (iii) کدها را فقط وقتی میان موج‌ها نگاشت می‌کند که یکسانی معنا نشان داده شده باشد؛ (iv) از تحلیل‌گر می‌خواهد قاعده‌های نمونه، سیاست مقدار گمشده و وزن‌دهی را بیان کند؛ و (v) اعتبارسنجی خود را با عددهای منتشرشده همراه دارد. از نظر علمی، سهم کار برآوردگر تازه نیست، بلکه راهی است که فرض‌های یک تحلیل NHTS را آشکار، آزمودنی و دوباره‌به‌کارگرفتنی می‌کند. کاربر بسته را نصب می‌کند، یک موج را با یک فرمان دریافت می‌کند، آن را تنبل بارگذاری می‌کند و آمار وزنی، نرخ، دسته‌های هم‌تراز یا یک مطالعه‌ی کامل می‌خواهد (بخش ۳). نتیجه‌ها موج، وزن، طرح واریانس و یکا را با خود دارند و پانویس جدول را می‌توان از خود نتیجه نوشت.

پیشینه‌ی نرم‌افزاری. بسته‌ی R با نام summarizeNHTS [24] داده‌های NHTS را دریافت و خلاصه می‌کند و بسته‌ی survey در R [25] تحلیل عمومی با وزن‌های تکرار و سری تیلور عرضه می‌کند که کاربر پس از تنظیم طرح به‌دست خود می‌تواند بر NHTS به کار ببندد. بسته‌ی پایتونی samplics [26] برآورد نمونه‌ی پیچیده را به‌طور کلی پوشش می‌دهد ولی پرونده‌ها، کدها و وزن‌های NHTS را نمی‌شناسد. جست‌وجو در GitHub و PyPI در ۲۲ اکتبر ۲۰۲۶ (پرس‌وجوهای «NHTS python» و «national household travel survey» و نام‌های summarizenhts، nhts-py، nhts) فقط به اسکریپت‌های تحلیل برای پروژه‌های جداگانه رسید و بسته‌ی پایتونی عمومی دیگری برای دسترسی به NHTS نیافت. برآوردگرهای واریانس از نظریه‌ی استاندارد وزن‌های تکرار [27] و گزارش وزن‌دهی پیمایش ۲۰۱۷ [28] پیروی می‌کنند.

۲. توصیف نرم‌افزار

۲/۱ معماری نرم‌افزار

pynhts لایه‌لایه سازمان یافته است (شکل ۱). لایه‌های ورودی و آمار داده را مدیریت می‌کنند؛ لایه‌های هم‌ترازسازی و تحلیل به تحلیل‌گر کمک می‌کنند پرسش‌های درست بپرسد؛ و لایه‌ی اعتبارسنجی نتیجه را با عددهای منتشرشده می‌سنجد.



شکل ۱. معماری pynhts. آبی: دسترسی به داده و آمار؛ کهربایی: هم‌ترازسازی و تحلیل؛ سبز: اعتبارسنجی و پل‌ها. هر چارچوب نتیجه ثبت می‌کند چگونه محاسبه شده است.

pynhts.io یک ثبت‌نامه با درهم‌ساز SHA-256 هر بایگانی رسمی و هر عضو آن نگه می‌دارد، با pooch دریافت می‌کند و Parquet را به‌صورت قطعی می‌نویسد. کارگزار داده برای پرونده‌های ناموجود پاسخ HTTP 200 با یک صفحه‌ی HTML می‌دهد؛ از این رو نوع محتوا و اندازه بررسی می‌شود. pynhts.core طرح‌واره‌ای نوع‌دار از کتاب‌کد می‌سازد (هر ۳٬۸۰۷ مقدار کدشده‌ی ۲۰۱۷ و هر ۲٬۶۳۹ مقدار ۲۰۲۲ فراوانی‌های چاپ‌شده را بازتولید می‌کنند) و ستون‌ها را تنبل بارگذاری می‌کند و هر ستون را از جدولی که مالک آن است با پیوند چند-به-یکِ اعتبارسنجی‌شده می‌پیوندد. کد پرش در ستون‌های عددی به‌طور پیش‌فرض به مقدار گمشده تبدیل می‌شود و کدهای بی‌پاسخ در صورت درخواست پوشانده می‌شوند.

pynhts.stats جک‌نایف طبقه‌بندی‌شده‌ی طرح وزن‌دهی ۲۰۱۷ را اجرا می‌کند. با ۱۴ طبقه‌ی واریانس هر یک ۷ واحد، واریانس یک برآورد برابر ۷/۶ مجموع مجذور انحراف برآورد هر یک از ۹۸ تکرار از برآورد کامل است و حاشیه‌ی خطا (MOE) برابر چندک t با ۹۸ درجه‌ی آزادی ضرب در خطای استاندارد است [27,28]:

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

پرونده‌های ۲۰۰۹ و ۲۰۰۱ به‌ترتیب ۱۰۰ و ۹۹ تکرار با ضریب‌های خاص خود دارند. پرونده‌ی ۲۰۲۲ وزن تکرار ندارد؛ از این رو pynhts خطی‌سازی تیلور را با طبقه‌ها و خانوارها به کار می‌برد و خطاهای آن تا 1e-9 با بسته‌ی survey در R برابر است. همه‌ی سفرهای یک نفر یک وزن دارند؛ پس برآورد سفر روی واحدهای فرد-روز محاسبه می‌شود و به ماتریس ۹۸ ستونه برای ۲۶۴٬۲۳۴ واحد نیاز دارد، نه ۹۲۳٬۵۷۲ سفر. زیرجامعه‌ها به‌صورت دامنه مدیریت می‌شوند و هرگز با حذف ردیف‌ها نیستند.

pynhts.concepts و pynhts.harmonize ۴۴ مفهوم را در چهار موج توصیف می‌کنند، هر یک با درجه‌ی «قابل‌مقایسه»، «قابل‌مقایسه با احتیاط» یا «ناقابل‌مقایسه»، و ۱۹ جدول تناظر فراهم می‌کنند (هدف سفر، وسیله، نژاد و خاستگاه اسپانیایی‌تبار، بیماری و جز آن). جدول تناظر موج را از چارچوب می‌خواند و ناهمخوانی را نمی‌پذیرد؛ هر دسته «دقیق»، «تقریبی» یا «تودرتو» نشان داده می‌شود و شاهدش در داده است. لایه‌ی تحلیل شامل aggregate (سفر به فرد یا خانوار با سیاست صریح مقدار گمشده و صفرهای واقعی)، features (۹۳ متغیر مستند با تبارنامه)، study (در ادامه)، trends (برآورد هر موج؛ ادغام برای مفهوم‌های ناقابل‌مقایسه رد می‌شود)، transform (تقسیم بر پایه‌ی خانوار تا مجموعه‌های آموزش و آزمون خانوار مشترک نداشته باشند) و tables و viz برای گزارش‌دهی است. pynhts.recipes عددهای چاپ‌شده در گزارش‌های رسمی را با وضعیت مورد انتظار هر یک رمز می‌کند و یک گزارش تولیدشده هر شکاف را فهرست می‌کند. پلی کوچک جدول‌ها را به آرایه‌های NumPy، تانسورهای PyTorch و گراف‌های خانوار صادر می‌کند.

۲/۲ قابلیت‌های نرم‌افزار

کارکردهای اصلی عبارت‌اند از: دریافت راستی‌آزمایی‌شده و کار آفلاین (pynhts fetch, register, docs)؛ بارگذاری تنبل و نوع‌دار چهار جدول هر موج؛ جمع، میانگین، نسبت و نرخ وزنی با خطای استاندارد و حاشیه‌ی خطا (rates.daily_rate نرخ روزانه می‌دهد و می‌گوید کی یک جمع سالانه است)؛ جدول‌های تناظر میان موج‌ها و متغیر ترکیبی نژاد-و-خاستگاه؛ هشدار برای مقدارهای سقف‌گذاری‌شده، ستون‌های ویژه‌ی نمونه‌ی ملی و سقف‌ها؛ و ماژول مطالعه. یک شیء Study موج، واحد تحلیل، قاعده‌های مرتب نمونه (Keep, Drop, Complete)، متغیرها، وزن (یا «بدون وزن» صریح) و فرض‌های اعلام‌شده را می‌گیرد و پرونده‌ی تحلیل، جدول جریان نمونه با شمار حذف‌شده‌ی هر قاعده، جدول توصیفی، یک بند روش و مقایسه با عددهای منتشرشده را برمی‌گرداند که «دقیق»، «نزدیک»، «قابل‌قبول» یا «شکاف» نمره می‌گیرد. هرگز قاعده‌ای را برای رسیدن به یک عدد دست‌کاری نمی‌کند. جدول ۲ همه‌ی ماژول‌ها را با کارکردشان فهرست می‌کند. دو ابزار برای دستیارهایی است که کد تحلیل می‌نویسند: AGENTS.md ده قاعده‌ای را می‌گوید که از خطاهای خاموش بالا جلوگیری می‌کنند و agent_eval یک پرونده‌ی پاسخ را روی ۱۴ مسئله با همین تله‌ها نمره می‌دهد.

جدول ۲. ماژول‌های pynhts نسخه‌ی 0.8.1 و کار هر یک.

ماژول	کارکرد
pynhts.io	دریافت راستی‌آزمایی‌شده (SHA-256 هر بایگانی)، ثبت پرونده‌هایی که از پیش روی دیسک است، حافظه‌ی نهان Parquet قطعی.
pynhts.core	طرح‌واره‌ی نوع‌دار از کتاب‌کد، بارگذار تنبل با پیوندهای اعتبارسنجی‌شده، سیاست کد گمشده، پرچم سقف و ستون ویژه‌ی نمونه‌ی ملی.
pynhts.stats	جمع، میانگین، نسبت و کسر با خطای جک‌نایف (۲۰۰۱، ۲۰۰۹، ۲۰۱۷) یا سری تیلور (۲۰۲۲)؛ زیرجامعه‌ها به‌صورت دامنه.
pynhts.rates	سفر و مایل روزانه برای هر فرد یا خانوار با مسیر رسمی و یکای صریح.
pynhts.concepts	۴۴ مفهوم در چهار موج: جامعه، یکا، درجه‌ی قابلیت مقایسه، تله‌ها.
pynhts.harmonize	مقایسه‌ی دو موج؛ ۱۹ جدول تناظر (وسیله، هدف، درآمد، نژاد، خاستگاه اسپانیایی‌تبار، بیماری و جز آن) با وضعیت هر کد.
pynhts.aggregate, features	سفر به فرد و خانوار با سیاست صریح مقدار گمشده؛ ۹۳ ویژگی مستند با فرمول و تبارنامه.
pynhts.eda, quality	دلیل‌های مقدار گمشده، توصیف وزنی، گزارش یکپارچگی کلید و وزن.

`pynhts.transform, trends`	تبدیل‌های بدون نشت و تقسیم بر پایه‌ی خانوار؛ برآورد هر موج و ادغام نگهبانی‌شده.
`pynhts.tours`	تورهای خانه‌مبنای کاری، سایر و غیرخانه‌مبنا و زنجیره‌های سفر به سبک FHWA (فقط ۲۰۱۷).
`pynhts.study`	بیانیه‌ی پژوهشی به پرونده‌ی تحلیل، جدول جریان نمونه، جدول توصیفی، متن روش و نمره‌دهی.
`pynhts.tables, viz`	جدول به‌صورت HTML، Markdown، LaTeX یا Excel با پانویس روش؛ نمودارها از جمله دفتر فعالیت.
`pynhts.recipes`	جدول‌های منتشرشده با وضعیت مورد انتظار هر عدد و گزارش شکاف تولیدشده.
`pynhts.ml`	آرایه، تانسور PyTorch و گراف خانوار.
`Command line`	pynhts fetch, register, docs (مستندات رسمی با بررسی درهم‌ساز).

۳. مثال‌های توضیحی

مثال ۱ برآوردهای ملی پیمایش ۲۰۱۷ را با حاشیه‌ی خطای درست طرح تولید می‌کند.

```
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")
```

این مقدارها با عددهای چاپ‌شده در گزارش خلاصه‌ی رسمی [1] برابرند، از جمله حاشیه‌های خطا.

مثال ۲ نشان می‌دهد چرا مخرج باید بیان شود. Lidbe و همکاران [5] گزارش می‌کنند ۵٫۴۹٪ سفرهای ۲۰۰۱ را کسانی کرده‌اند که بیماری دشوارکننده‌ی سفر دارند. این پرسش فقط از افراد ۱۶ ساله و بزرگ‌تر شده است و افراد کوچک‌تر کد پرش دارند. اگر آن‌ها «بدون بیماری» شمرده شوند ۴٫۴۰٪ به دست می‌آید؛ با سفرهای کسانی که پاسخ داده‌اند مقدار چاپ‌شده برمی‌آید. جدول تناظر برای پرش مقدار گمشده برمی‌گرداند؛ پس مخرج دوم طبیعی است.

```
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()
```

مثال ۳ یک مطالعه‌ی کامل را به‌صورت داده بیان می‌کند. قاعده‌های نمونه‌ی Zhang و Zhang [8] (حذف ۱۵ ایالت کم‌کاربرد، فقط کسانی که پرسش سواری اشتراکی را پاسخ داده‌اند، متغیرهای کامل مدل، بدون وزن) یک Study می‌شود که جدول جریان نمونه و جدول توصیفی را می‌دهد. نتیجه ۲۱۸٬۸۳۱ نفر است در برابر ۲۱۹٬۰۲۶ چاپ‌شده (۰٫۰۹٪ کمتر)، پس از حذف ۹٬۱۴۲، ۲۷٬۰۹۰ و ۹٬۱۷۱ نفر با سه قاعده از ۲۶۴٬۲۳۴ نفر پرونده، و میانگین دفعات سواری اشتراکی در ۳۰ روز ۰٫۳۰۹ است، همان که چاپ شده.

```
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
```

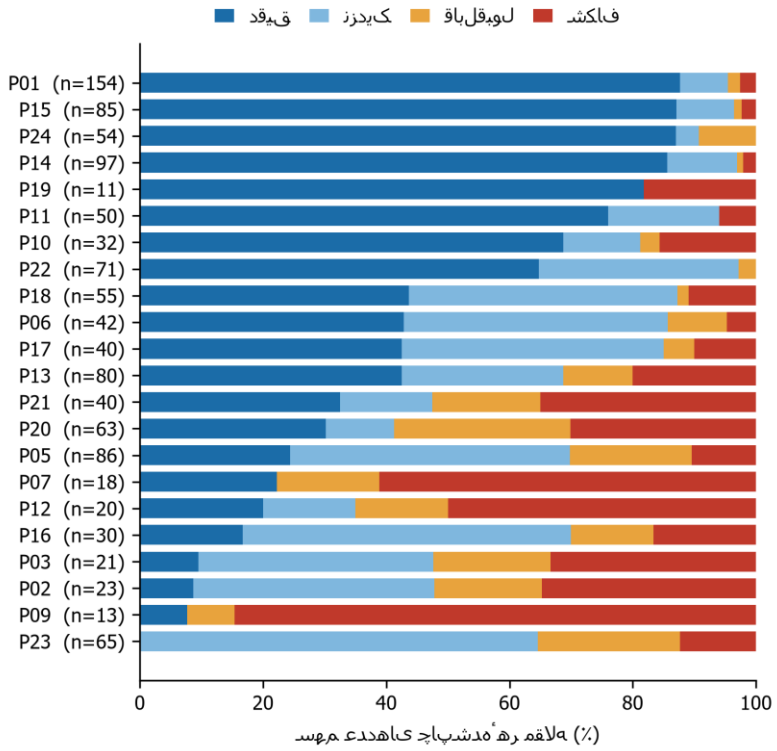
۴. اثر

اعتبارسنجی در برابر عددهای رسمی

پنج دستورالعمل (recipe) ۱۷۵ عدد چاپ‌شده در سندهای رسمی را رمز می‌کنند: جدول‌های ۱ و ۳ خلاصه‌ی روندهای سفر برای ۲۰۰۱، ۲۰۰۹ و ۲۰۱۷، جمع‌های کنترل راهنمای کاربر ۲۰۲۲ و یک مطالعه‌ی تور ترانزیتی. از این‌ها ۱۴۲ مورد کاملاً بازتولید می‌شود، ۸ مورد برآورد را بازتولید می‌کند ولی حاشیه‌ی خطای چاپ‌شده را نه (گزارش ضریب دیگری به کار برده) و ۲۵ مورد نمی‌خواند؛ هر شکاف در گزارش تولیدشده، هرجا علتش معلوم است، توضیح داده شده و هیچ قاعده‌ای برای پنهان‌کردن شکاف عوض نشده است. برای ۲۰۲۲ خطاهای سری تیلور با بسته‌ی survey در R برابر است. ۳۳۳ آزمون واحد بدون نیاز به داده و بیش از ۶۰۰ آزمون هنگام وجود پرونده‌های واقعی وجود دارد که روی لینوکس و ویندوز اجرا می‌شوند.

بازسازی مطالعه‌های منتشرشده

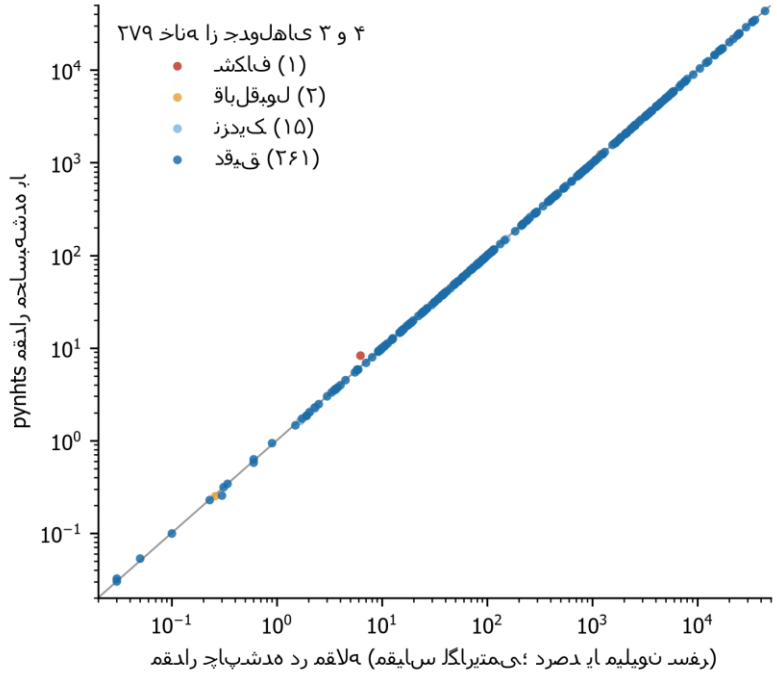
برای آزمودن اینکه بسته از تحلیل‌های واقعی پشتیبانی می‌کند، داده‌های ۲۲ مطالعه‌ی منتشرشده‌ی NHTS فقط با خود بسته بازسازی شد، یک اسکریپت برای هر مطالعه، و هر عددی که مطالعه درباره‌ی داده‌اش چاپ کرده بود با نتیجه سنجیده شد: دقیق (برابر در دقت چاپی)، نزدیک (تا ۲٪)، قابل‌قبول (تا ۵٪) یا شکاف. هیچ قاعده‌ای برای رسیدن به عدد دست‌کاری نشد. از ۱٬۱۵۰ مقدار نمره‌گذاری‌شده، ۶۱۸ دقیق، ۲۷۹ نزدیک، ۱۰۵ قابل‌قبول و ۱۴۸ شکاف است (شکل ۲). در نُه مطالعه داده‌های تحلیل را می‌شد به‌طور قانع‌کننده بازسازی کرد؛ در بقیه، عددهای چاپ‌شده به قاعده‌های نمونه یا متغیرهای بیرونی وابسته‌اند که مقاله نمی‌گوید یا پرونده‌های عمومی ندارند. هر شکاف را گزارشی درباره‌ی اطلاعات موجود در مقاله می‌دانیم، نه درباره‌ی کار نویسندگان آن.



شکل ۲. نتیجه‌ی ممیزی ۲۲ مطالعه‌ی منتشرشده‌ی NHTS (P01-P24؛ در مخزن فهرست شده، اسکریپت‌های ممیزی نخست). هر میله سهمی از عددهای چاپ‌شده‌ی مطالعه است که pynhts دقیق، نزدیک (۲٪)، قابل‌قبول (۵٪) یا ناهمخوان بازتولید می‌کند.

یکی از این ۲۲ مطالعه کار پیشین خود نویسندگان درباره‌ی تولید سفر خانوار به تفکیک ایالت است [14]؛ پس آن مورد مستقل نیست. آن مطالعه پاک‌سازی خود را فقط «حذف رکوردهای ناقص و پرت» می‌نویسد. قاعده‌هایی که جدول‌های چاپ‌شده‌ی آن القا می‌کنند (دست‌کم یک خودرو و یک سفر، درآمد، نژاد، چرخه‌ی زندگی و تصرف معتبر) ۱۰۳٬۳۱۷ خانوار باقی می‌گذارد در برابر ۱۰۶٬۲۸۷ چاپ‌شده (۲٫۸٪ کمتر)؛ از ۵۴ مقدار بررسی‌شده‌ی آن، ۴۷ در دقت چاپی دقیق، ۲ نزدیک و ۵ قابل‌قبول است. نمونه‌ی بازسازی‌شده همچنین نشان می‌دهد ۳۳ ایالت از ۵۱ واحد کمتر از ۵۰۰ خانوار دارند (میانه ۳۱۲)، که برای هر مدلی که ایالت به ایالت برازش شود اهمیت دارد.

این تمرین ۳۳ قرارداد بازیابی‌شده به بار آورد که با شاهدشان نوشته شده‌اند و چند تا از آن‌ها مقدارهای منتشرشده را جابه‌جا می‌کنند. بازبینی دوباره‌ی یک مطالعه [5] با جدول‌های تناظر و ماژول مطالعه این را نشان می‌دهد: ۲۶۲ خانه از ۲۸۰ خانه‌ی جدول‌های ۳ و ۴ آن دقیق بازتولید می‌شود (شکل ۳). درصدهای جدول ۳ آن مخرجی دارد که سفرهای افراد دارای پاسخ معتبر است (مثال ۲)؛ آخر هفته‌ی آن شنبه و یکشنبه‌ی متغیر روز سفر است، در حالی که پرچم آخر هفته‌ی NHTS بخشی از جمعه را هم دارد و مقدارها را ۵ تا ۸٪ بالاتر می‌دهد؛ مدت سفر متغیر دقیقه‌های محاسبه‌شده است؛ شش گویه درباره‌ی اینکه بیماری چگونه سفر را عوض می‌کند دقیق بازتولید می‌شود (۷۲ از ۷۲ خانه) از راه جدول تناظر اثر بیماری؛ و گروه‌های نژادی آن قاعده‌ای را دنبال می‌کند که خاستگاه اسپانیایی‌تبار فقط گروه‌های سفید، سیاه و آسیایی را بازنویسی می‌کند، که کار متغیر ترکیبی نژاد-و-خاستگاه نیست. در مطالعه‌ای دیگر [12] نمونه‌ی ۱۳۰٬۲۲۲ نفره به‌صورت یک Study دقیق بازتولید می‌شود و متغیر نژاد آن کد نژاد تکِ پاسخ‌گوی خانوار است؛ نگاشت آن از متغیر ترکیبی نمونه‌ای دیگر (۱۳۳٬۵۳۸ نفر) می‌دهد. این یافته‌ها به ویژگی‌های بسته (جدول‌های تناظر، سقف‌ها، هشدارها) یا مستندات تبدیل شدند.



شکل ۳. مقدارهای چاپ‌شده در جدول‌های ۳ و ۴ Lidbe و همکاران [5] در برابر مقدارهای محاسبه‌شده با pynhts (۲۷۹ خانه‌ی دارای مقدار مثبت؛ یک صفر چاپ‌شده نشان داده نشده است). قطر نشان‌دهنده‌ی برابری دقیق است.

پرسش‌های پژوهشی تازه و تغییر رویه

عدم‌قطعیت درست طرح برای هر موج، مقایسه‌های نگهبانی‌شده میان موج‌ها و جدول‌های جریان نمونه پرسیدن پرسش‌هایی را ارزان‌تر می‌کند که به روند و زیرجامعه‌ها بستگی دارد، و گزارش آن‌ها را چنان که دیگران بتوانند تکرار کنند. ماژول مطالعه می‌خواهد با واردکردن قاعده‌های نمونه و سیاست مقدار گمشده در کد، رویه‌ی روزانه‌ی تحلیل‌گران را عوض کند. هنوز تغییری در رویه نسنجیده‌ایم.

گستره‌ی استفاده

نخستین نسخه در اکتبر ۲۰۲۶ در PyPI منتشر شد و بسته در مجموع ۱۳ انتشار داشته است؛ نسخه‌ی 0.8.1 در Zenodo بایگانی شده است [29]. بسته کاربر بیرونی شناخته‌شده‌ای ندارد، هیچ نشریه‌ای از آن استفاده نکرده و کاربرد تجاری ندارد. PyPI دانلودها را می‌شمارد ولی آن‌ها آینه‌های خودکار را هم دربر می‌گیرند؛ پس آن‌ها را کاربر نمی‌شماریم. شاهد اثر بنابراین همان اعتبارسنجی و ممیزی است که هر کس با پرونده‌های عمومی می‌تواند تکرارش کند، و کاربرد مورد نظر پژوهشگرانی است که NHTS را تحلیل می‌کنند و کسانی که چنین تحلیل‌هایی را داوری می‌کنند.

۵. نتیجه‌گیری

pynhts به پژوهشگران دسترسی نوع‌دار و وزنی به چهار موج NHTS می‌دهد، با روش واریانس هر موج، جدول‌های تناظری که با داده سنجیده شده‌اند و راهی برای بیان نمونه و سیاست مقدار گمشده به‌صورت کد. اعتبارسنجی آن در برابر عددهای رسمی و مطالعه‌های منتشرشده نشان می‌دهد که بیشتر عددها را بازتولید می‌کند و اختلاف‌های باقی‌مانده معمولاً از قاعده‌هایی است که نشریه‌ها نمی‌نویسند.

محدودیت‌ها روشن است. بسته در مرحله‌ی آلفا است و رابط آن ممکن است تغییر کند. فقط پرونده‌های عمومی را پوشش می‌دهد؛ متغیرهایی که پرونده‌های محدودشده می‌خواهند (مثلاً شناسه‌ی بلوک سرشماری) در دسترس نیست. مدل برازش نمی‌دهد؛ بنابراین خطای استاندارد ضریب‌های رگرسیون که نیاز دارد وزن‌ها برای هر تکرار دوباره برازش شوند هنوز فراهم نیست. وزن‌های ملی نمونه‌ی ملی ۲۰۰۱ در دسترس نیست. بایگانی‌های وزن تکرار ۲۰۰۱ و ۲۰۰۹ بزرگ‌اند. گام‌های بعدی برآورد واریانس برای ضریب مدل‌ها، جدول‌های تناظر بیشتر و ممیزی تکرارشونده با پدیدآمدن مطالعه‌های تازه است.

سپاس‌گزاری

پیمایش ملی سفر خانوار را وزارت حمل‌ونقل ایالات متحده، اداره فدرال بزرگراه‌ها انجام می‌دهد و آزمایشگاه ملی اوک‌ریج آن را توزیع می‌کند. بسته داده‌های پیمایش را بازتوزیع نمی‌کند.

اعلان استفاده از هوش مصنوعی مولد و فناوری‌های یاری‌رسان در تهیه‌ی دست‌نوشته

در تهیه‌ی این اثر نویسندگان از Claude (Anthropic) به‌عنوان دستیار برای طراحی نرم‌افزار، نوشتن و اجرای کد و آزمون‌ها، سنجش کد با سندهای رسمی و داده‌ها، غربال ادبیات و پیش‌نویس و ویرایش متن استفاده کردند. پس از استفاده از این ابزار، نویسندگان مواد (کد، نتیجه‌ها، مرجع‌ها و متن) را بازبینی، راستی‌آزمایی و ویرایش کردند و مسئولیت کامل محتوای نشریه را می‌پذیرند. عددهای متن با کد از پرونده‌های نتیجه‌ی مخزن برداشته شده و مرجع‌ها با یک پایگاه کتاب‌شناختی سنجیده شده‌اند.

بیانیه‌ی مشارکت نویسندگان (CRediT)

سید صابر ناصرعلوی: مدیریت پروژه، راهنمایی (نظارت). اکرم مظاهری: مفهوم‌سازی، مدیریت داده، تحلیل صوری، پژوهش، روش‌شناسی، منابع، نرم‌افزار، اعتبارسنجی، تصویرسازی، نگارش (پیش‌نویس اصلی)، نگارش (بازبینی و ویرایش). نام نقش‌ها بر پایه‌ی طبقه‌بندی CRediT است.

تأمین مالی

این پژوهش از هیچ نهاد تأمین مالی در بخش دولتی، تجاری یا غیرانتفاعی کمک مالی ویژه‌ای دریافت نکرده است.

اعلان منافع رقابتی

نویسندگان اعلام می‌کنند هیچ منفعت مالی یا رابطه‌ی شخصی شناخته‌شده‌ای که بتواند بر کار گزارش‌شده در این مقاله اثر بگذارد ندارند.

دسترس‌پذیری داده

نرم‌افزار در https://github.com/sabernaseralavi-60/pynhts در دسترس است (بایگانی Zenodo، DOI 10.5281/zenodo.23575312). داده‌های عمومی NHTS را آزمایشگاه ملی اوک‌ریج (https://nhts.ornl.gov) توزیع می‌کند؛ pynhts آن‌ها را از همان‌جا دریافت می‌کند و بازتوزیع نمی‌کند.

مرجع‌ها (به انگلیسی، همان فهرست نسخه‌ی انگلیسی)

[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.

نسخه‌ی اجرایی کنونی نرم‌افزار

کاربرد ندارد: pynhts یک کتابخانه‌ی پایتون بدون برنامه‌ی اجرایی جداگانه است. نسخه‌ی ۰/۸/۱ را با pip install pynhts==0.8.1 نصب کنید.