

UC *Love Data Week*

Data in Context: Strategies for Evaluating and Utilizing Existing Datasets for Research

Wasila Dahdul



Pamela Reynolds



February 14, 2025 | UC Love Data Week

UC Love Data Week



Code of Conduct

By participating in this community, participants accept to abide by the UC Love Data Week (UCLDW) Code of Conduct (CoC) and accept the procedures by which any CoC incidents are resolved. Any form of behavior to exclude, intimidate, or cause discomfort is a violation of the CoC.

Expected Behavior

- Use welcoming & inclusive language
- Be respectful of different viewpoints & experiences
- Gracefully accept constructive criticism
- Focus on what is best for the community
- Show courtesy & respect towards other community members

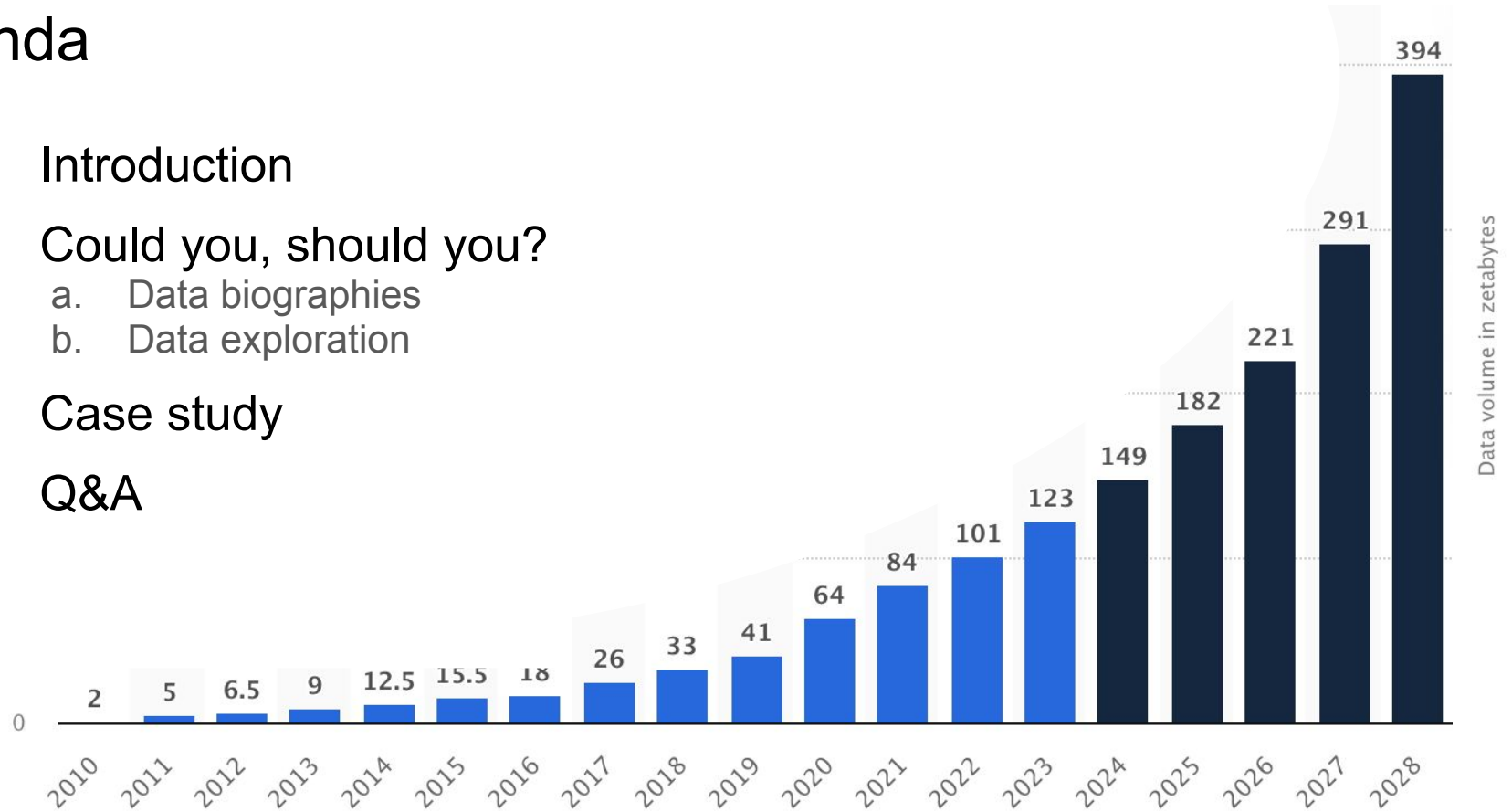
If you believe someone is violating the CoC we ask that you report it to the instructor/helper in your session, or to uclovedataweek@gmail.com.

[Code of Conduct Full Document](#)

[Acknowledgements Full Document](#)

Agenda

1. Introduction
2. Could you, should you?
 - a. Data biographies
 - b. Data exploration
3. Case study
4. Q&A



Why reuse existing data?

- **It's already there!**
- **I can ask bigger questions!**
- **I can discover something novel!**

Where to find existing datasets

- Public data repositories
 - Discipline-specific repositories (e.g., ICPSR)
 - Generalist repositories (e.g., Zenodo, Dryad)
 - Search Re3sharing, FAIRsharing for discipline-specific repositories
- Government databases
 - Census Bureau, CDC, NOAA
 - Federal Departments (Transportation, Labor, etc.)
 - State Departments (Health, etc.)
- Research articles, data journals
- Libraries, archives
- Search engines (e.g., Google dataset search)

Librarians are a
great resource!

Check out
libguides, ex.:
<https://ucsd.libguides.com/usgov>

Criteria for data reuse evaluation

1. Legal
2. Responsible use
3. Quality and biases

How do you know?!

Can you reuse the data, legally?

Look for a license or data use agreement.

- Common open licenses:

- [Creative Commons](#)
(e.g., CC0, CC-BY)
- [Open Data Commons Attribution License](#)
(ODC-By)



Open Knowledge
FOR A FAIR, SUSTAINABLE AND OPEN FUTURE

- Data Use Agreements are sometimes required for access to restricted use or licensed datasets

Just because you can access a corpus does not mean you have the rights to use it in your NLP / LLM. If not easily discoverable, ask your librarian for help.

Should you reuse the data?

1. Legal
2. Responsible use -> Data Biography
3. Quality and biases -> Exploratory Data Analysis (EDA)

Conducting a Data Biography

Is there sufficient information available to understand the dataset and its context/origin?

1. Who...
 - Collected it? Funded it? Maintains it? Owns it?
 - Does it affect, and how?
2. Why does it exist?
 - Why was it collected and for what purpose?
 - What questions does/doesn't it answer?
3. How was it collected?
 - What are the methods behind the data collection design and process?
 - Which instruments were used? What settings/parameters?
 - When and where was it collected?
4. What does/doesn't it contain?
 - Parameters? Subpopulations? PII?
5. How is it being used?

Data Biography Synthesis

- Are the data well documented? Can you access the protocols?
- Is the information complete, understandable and consistent?
- Were appropriate ethical review processes conducted? (e.g., IRB)
- If it pertains to people, was there consent for collection and use?
- Is my research similar to the original research question?
- Are the collection and processing methods appropriate to answer my research question?
- What questions does this dataset NOT address?

Data Exploration

How complete, tidy, and consistent is the dataset?

Are your findings consistent with the metadata and your data biography?

1. Look at the data!
2. Examine the structure of the data
 - What are the observations, features, their data types and levels
 - Tidy data rules (for structured data)
 - Rows are observations
 - Columns are features
 - One value per cell
3. Calculate basic statistics
 - Length, ranges, mean, median, mode, etc.
4. Look for patterns of missingness
5. Explore known and test for random assumptions/correlations

Other considerations

Cite everything!

- data
- software
- techniques
- version

Alt. Resources for Gov't Data

- [ICPSR](#)
- [End of Term Web Archive](#)
- [Wayback Machine](#)
- [IPUMS](#)
- [Harvard Dataverse](#)

Also, see:

- Dataverse
- Dryad

Context for why data version matters

<https://libguides.ucmerced.edu/US-Federal-Data/overview>

<https://guides.library.fresnostate.edu/c.php?g=288876&p=10776458>

<https://libguides.lib.msu.edu/c.php?g=1451065>

Case study*

Casey Williams is a graduate student in Sociology. She is studying patterns of use at U.S. National Parks over time. Specifically, how different groups of people utilize the parks and how park usage reflects broader societal trends.

She would like to find and analyze data on public recreational use, private business activities, and on the primary residents of national parks. She begins data collection with the Visitor Use Data provided by the National Park Service.



*adapted from [Responsible Datasets in Context](#)

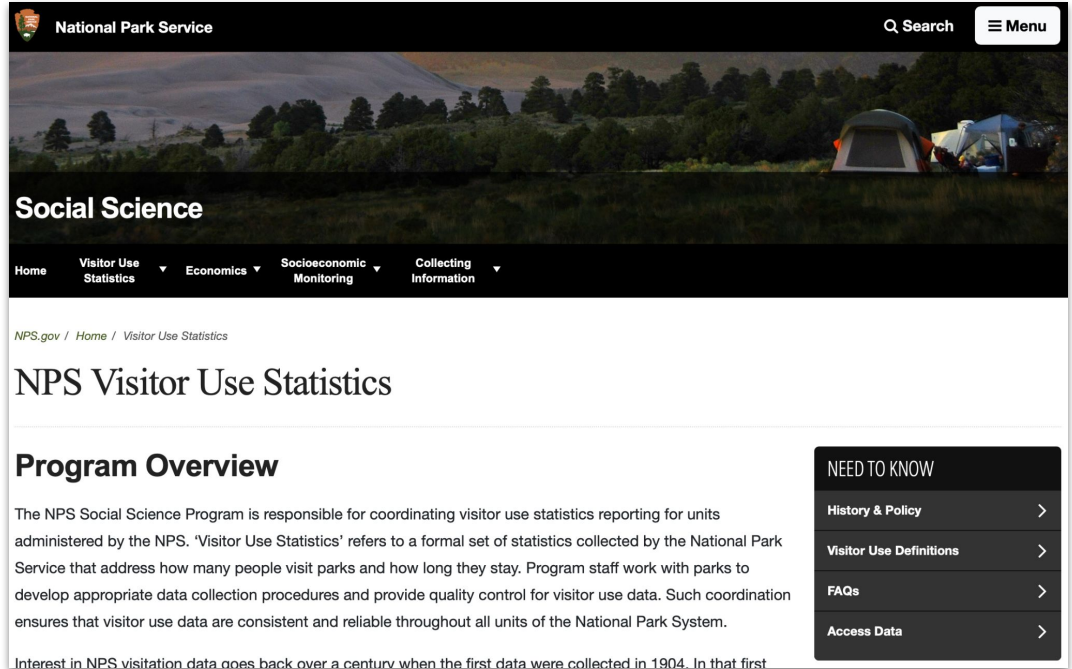
National Parks Data Exploration in R

[Download the R script](#) if you would like to follow along



NPS Social Science Program

- Collects visitor use data for US national parks, monuments, and other natural, historical, and recreational areas
- Park managers and the public can access datasets and documentation



The screenshot shows the National Park Service website's 'Social Science' section, specifically the 'Visitor Use Statistics' page. The header features the NPS logo, 'National Park Service' text, a search bar, and a menu icon. Below the header is a large banner image of a campsite in a forest. The 'Social Science' title is prominently displayed. A navigation bar includes links for Home, Visitor Use Statistics (selected), Economics, Socioeconomic Monitoring, and Collecting Information. The breadcrumb trail reads 'NPS.gov / Home / Visitor Use Statistics'. The main heading is 'NPS Visitor Use Statistics'. The 'Program Overview' section explains the NPS Social Science Program's role in coordinating visitor use statistics reporting. A 'NEED TO KNOW' sidebar on the right contains links for History & Policy, Visitor Use Definitions, FAQs, and Access Data. The bottom of the page begins with the text 'Interest in NPS visitation data goes back over a century when the first data were collected in 1904. In that first'.

National Park Service

Q Search Menu

Social Science

Home Visitor Use Statistics Economics Socioeconomic Monitoring Collecting Information

NPS.gov / Home / Visitor Use Statistics

NPS Visitor Use Statistics

Program Overview

The NPS Social Science Program is responsible for coordinating visitor use statistics reporting for units administered by the NPS. 'Visitor Use Statistics' refers to a formal set of statistics collected by the National Park Service that address how many people visit parks and how long they stay. Program staff work with parks to develop appropriate data collection procedures and provide quality control for visitor use data. Such coordination ensures that visitor use data are consistent and reliable throughout all units of the National Park System.

Interest in NPS visitation data goes back over a century when the first data were collected in 1904. In that first

NEED TO KNOW

- History & Policy
- Visitor Use Definitions
- FAQs
- Access Data

<https://www.nps.gov/subjects/socialscience/visitor-use-statistics.htm>

Biography: NPS Visitor Use Statistics

1. Who...

- Individual National Parks collect visitor use data
- Aggregated by NPS Social Science Program

2. Why does it exist?

- NPS visit data is an important resource used in decision making for park maintenance, conservation, and local economies

3. How was it collected?

- Each of the 63 parks count visits differently using automatic traffic counters or manual counting
- Annual visit estimates are available from 1904-1978 and monthly visit estimates from 1979-present

Source: <https://www.nps.gov/subjects/socialscience/visitor-use-statistics.htm>

Biography: NPS Visitor Use Statistics

4. What does/doesn't it contain?

- Recreation and non-recreation visits (*not visitors*) counted
- Does not contain entry by NPS employees, their families, concession employees, members of cooperating associations, NPS contractors, and service personnel

5. How is it being used?

- Internal purposes— staffing, programming, trail and infrastructure maintenance and upgrades
- Planning for shared community and business resources
- Estimating economic impacts to advocate for more funding and support

Source: <https://www.nps.gov/subjects/socialscience/visitor-use-statistics.htm>

EDA and the NPS Visitor Use Data

Let's dive into the data and documentation to ask:

1. What's in the data? What “counts” as a visit?
2. How was the data collected?
3. What data is missing? How is uncertainty handled?

1. What is the data? What “counts” as a visit?

Select Year(s) Select Month(s)

Select Region(s) Select Park Type(s)

Select Park(s) Select Field Name(s)

Select Additional Field(s) Annual Summary Only ☒ True ☐ False

1 of 1 Find | Next

NPS Public Use Statistics Query Builder

Park	Region	State	Year	Recreation Visits
Pinnacles NP	Pacific West	CA	1979	153,717
Pinnacles NP	Pacific West	CA	1980	163,626
Pinnacles NP	Pacific West	CA	1981	155,205
Pinnacles NP	Pacific West	CA	1982	171,026
Pinnacles NP	Pacific West	CA	1983	170,631
Pinnacles NP	Pacific West	CA	1984	166,258
Pinnacles NP	Pacific West	CA	1985	191,578

Recreation Visits

- ☐ (Select All)
- ☒ Recreation Visits
- ☐ NonRecreation Visits
- ☐ Recreation Hours
- ☐ NonRecreation Hours
- ☐ Concessioner Lodging
- ☐ Concessioner Camping
- ☐ Tent Campers
- ☐ RV Campers
- ☐ Backcountry Campers
- ☐ NonRecreation Overnight Stays
- ☐ Miscellaneous Overnight Stays

Source: [Query Builder for Visitor Use Statistics \(1979 - Last Calendar Year\)](#)

Visitor Use Definitions

- **Recreation visits** - *visitors using the park 'as a park'*
 - The entry of a person onto lands or waters administered by the NPS except for non-reportable and non-recreation visits. Funeral parties at National Cemeteries, school groups, etc. are reportable as 'recreation visits' since their use aligns with the purpose for which the park was established. Visits originating on surface vehicles (trains, boats, other) and aircraft may be counted if they stop and disembark passengers on NPS administrated territory. The applicable rule is that one entrance per individual per day may be counted.
- **Non-recreation visits** - *visitors using park territory, roads, and facilities for their own convenience or as a part of their occupation.*
 - Persons going to and from inholdings across significant parts of park land;
 - Commuter and other traffic using NPS-administered roads or waterways through a park for their convenience;
 - Trades-people with business in the park;
 - Any civilian activity a part of or incidental to the pursuit of a gainful occupation (e.g., guides);
 - Government personnel (other than NPS employees) with business in the park;
 - Citizens using NPS buildings for civic or local government business, or attending public hearings;
 - Outside research activities (visits and overnights) if independent of NPS legislated interests (e.g. meteorological research).

Source: <https://www.nps.gov/subjects/socialscience/nps-visitor-use-statistics-definitions.htm>

2. How was the data collected?

- “Visitor Use Counting Procedures” are available in the [NPS Data Portal](#), for example:
 - [Joshua Tree National Park](#)
 - [Voyageurs National Park](#)
- Data collection issues are documented and tracked in the NPS data collection logs, for example:
 - [Crater Lake National Park](#)
 - [Carlsbad Caverns National Park](#)

3. What data is missing? How is uncertainty handled?

- Key summary statistics, such as max, min, or average values can reveal important patterns, problems, or inconsistencies in the data
- Let's look at a few examples for our NPS data in R

Takeaways from exploring the NPS visitor data

- Data is shaped by human decisions
 - e.g., what constitutes a “visit;” choosing among collection methods
- Data collection can be affected by technical and environmental factors
 - e.g., broken counters, weather related closures
- Recognizing the imperfections and approximations in data is critical for meaningful analysis and interpretation

Thank you!

Wasila Dahdul (UCI): wdahdul@uci.edu

Pamela Reynolds (UCD): plreynolds@ucdavis.edu



Love Data Week