Datasheet for IPUMS USA Dataset*

Jianing Li

November 26, 2024

This datasheet provides a detailed description of a dataset compiled to analyze factors influencing wages among the U.S. workforce, based on data from the 2023 American Community Survey (ACS). The dataset includes over 1% of the U.S. population, highlight individuals aged 18 to 65 who are employed. It focuses on key variables such as education, age, gender, race, region, work hours, and income, each selected for their relevance to understanding the determinants of wages. The data has been preprocessed to ensure clarity and consistency, making it suitable for advanced statistical analysis and modeling.

This datasheet extracts questions from Gebru et al. (2021) to provide a structured framework for documenting datasets, ensuring transparency, accountability, and fostering responsible data use.

Motivation

1. *For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*

    - The dataset was created to analyze the factors influencing wages among the U.S. workforce. The aim of the study was to use a multiple linear regression model to investigate key factors such as education, age, gender, race, region, and hours worked on income. It addresses the gap in existing research, which often lacks recent, nationally representative data suitable for diverse analyses. The 2023 IPUMS USA sample fills this gap.

2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*

    - The dataset was created by the IPUMS team under the University of Minnesota. Variables selected by the author of the paper at the time of download.

[*] Code and data are available at: https://github.com/JianingLi1225/Determinants-of-Wages.git

3. *Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.*

   - The creation of the dataset was funded by multiple organizations. Key contributors include the National Institutes of Health (NIH), specifically the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD) and the National Institute on Aging (NIA), which provided grants supporting various IPUMS projects. The National Science Foundation (NSF) has also been a significant funder of IPUMS initiatives. Additionally, the Bill and Melinda Gates Foundation has contributed funding to specific IPUMS projects. Alongside external funding, the University of Minnesota offers substantial institutional support to IPUMS, ensuring the sustainability and continued development of its datasets.

4. *Any other comments?*

   - None

Composition

1. *What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*

   - The instances represent individual members of the U.S. workforce. Each instance records attributes such as gender, age, race, education, region, hours worked, and income. The dataset does not contain multiple types of instances.

2. *How many instances are there in total (of each type, if appropriate)?*

   - The dataset is based on a 1% national sample from the 2023 American Community Survey (ACS)

3. *Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).*

   - The dataset is a 1% random sample from a larger set of U.S. population data. Sampling weights and adjustments ensure that the sample is representative of the national population.

4. *What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.*

- Each instance contains preprocessed features, such as income (INCWAGE), weekly work hours (UHRSWORK), gender, age, race, education level, and region. The data has been cleaned and recoded for analysis.

5. *Is there a label or target associated with each instance? If so, please provide a description.*

   - Yes, the target variable for each instance is income (INCWAGE), which is used to analyze the effects of various factors on earnings.

6. *Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.*

   - There is missing information. But the missing values are removed during the process of cleaning the data

7. *Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.*

   - The instances in the dataset are primarily independent individual records, with no explicit relationships between them.

8. *Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.*

   - The dataset is split into a training set (60%) and a testing set (40%). The training set is used for model training and parameter estimation, while the testing set evaluates the model's performance on unseen data.

9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*

   - Since the data is self-reported, there may be minor noise, such as inaccuracies in income or work hours. Additionally, smaller racial groups are categorized under "Other," which may lead to a loss of specific details.

10. *Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*

    - The dataset is self-contained and does not rely on any external resources.

11. *Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.*

    - The dataset is anonymized and does not include any information that could directly identify individuals, so it does not contain confidential data.

12. *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.*

    - No

13. *Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*

    - Yes, the dataset identifies sub-populations based on gender, race, and region. For example, racial groups include White, Black, Asian, and "Other."

14. *Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.*

    - No, it is not possible. The dataset has been strictly anonymized, making it impossible to directly or indirectly identify individuals.

15. *Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.*

    - Yes, the dataset includes sensitive information such as race, gender, education, and income. These variables are necessary for studying socioeconomic phenomena.

16. *Any other comments?*

    - None

Collection process

1. *How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*

- The data was collected through self-reported responses to the American Community Survey (ACS). To enhance accuracy, the ACS applies weighting methods to adjust for sampling and nonresponse biases and follows strict statistical methodologies in questionnaire design.

2. *What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?*

   - The data was collected using standardized survey questionnaires via online submissions, mail, and face-to-face interviews. The U.S. Census Bureau oversees the data collection process, ensuring that all procedures adhere to statistical standards.

3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?*

   - The dataset employs a 1% random sampling strategy from a larger U.S. population database. The sampling method ensures coverage of diverse geographic areas and population characteristics nationwide.

4. *Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?*

   - The data was collected by staff and field surveyors from the U.S. Census Bureau. They are salaried employees compensated according to government pay scales.

5. *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.*

   - The data was collected in 2023 and aligns with the time frame of the data instances.

6. *Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*

   - Data collection was overseen by the U.S. Census Bureau, adhering to all relevant ethical standards, including privacy protection and data anonymization.

7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*

   - The data was directly collected from individuals through their responses to the American Community Survey.

8. *Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a*

link or other access point to, or otherwise reproduce, the exact language of the notification itself.

- Yes, respondents are informed before the survey begins, with detailed explanations about the purpose of the survey and an overview of how the data will be used.

9. *Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*

    - Yes, participation in the survey is voluntary, and respondents provide consent by completing the questionnaire.

10. *If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*

    - Since the data is collected anonymously, respondents cannot withdraw the information they have submitted.

11. *Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*

    - None

12. *Any other comments?*

    - None

Preprocessing/cleaning/labeling

1. *Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.*

    - The data cleaning process involved several steps to prepare the raw data for analysis. First, the dataset was filtered to include individuals aged 18 to 65, representing the working-age population, while records where the employment status was not "Employed" were excluded to focus on individuals with income. Categorical variables were recoded to enhance interpretability. For example, the education variable was grouped into five categories: "Below High School," "High School," "Some College," "Bachelor," and "Above Bachelor." The race variable was classified into "White," "Black," "Asian," and "Other." Regional data, based on state codes, was mapped into "Northeast," "Midwest," "South," and "West," and the gender variable was simplified to “Male” and “Female.” For numerical variables, outliers in work hours were addressed by removing zero or extreme values to ensure validity. These steps ensured a clean and well-structured dataset, suitable for further analysis.

2. *Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the “raw” data.*
    - The raw data is downloaded from the IPUMS website and saved locally on the computer.
3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*
    - Yes, data cleaning was done using R and related packages, such as dplyr and tidyverse.
4. *Any other comments?*
    - None

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*
    - The dataset was used to analyze key factors influencing wages among the U.S. workforce, evaluating the effects of variables such as education, age, gender, race, region, and work hours using a multiple linear regression model.
2. *Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.*
    - A GitHub repository is provided, and the code and data can be accessed through https://github.com/JianingLi1225/Determinants-of-Wages.git
3. *What (other) tasks could the dataset be used for?*
    - The dataset can be used for studying income inequality, analyzing the impact of education, assessing racial and gender disparities, and evaluating regional economic policies.
4. *Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?*

- Since the data is self-reported, it may introduce minor biases. Additionally, smaller racial groups were categorized as “Other,” which could limit detailed analysis.

5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*

    - The dataset is not suitable for tasks requiring personal identification or detailed tracking of individual changes, as the data is anonymized and lacks longitudinal information.

6. *Any other comments?*

    - None

Distribution

1. *Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.*

    - Yes, IPUMS USA provides public access, and any qualified researcher can apply to use the data.

2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*

    - The dataset is distributed via the IPUMS website, where users must create an account to download it.

3. *When will the dataset be distributed?*

    - The dataset is already available and updated annually with each new ACS release.

4. *Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/ or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*

    - The dataset is distributed under IPUMS terms of use and is restricted to non-commercial research. Users must comply with the usage agreement.

5. *Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*

    - No third-party restrictions, but users must agree to IPUMS terms and conditions.

6. *Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*

- No explicit mention of export controls.

7. *Any other comments?*

    - None

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*

    - The dataset is supported, hosted, and maintained by IPUMS, which is part of the Institute for Social Research and Data Innovation at the University of Minnesota.

2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*

    - IPUMS can be contacted via email at ipums@umn.edu for feedback, questions, or concerns.

3. *Is there an erratum? If so, please provide a link or other access point.*

    - IPUMS maintains a revisions page where updates and corrections to datasets are documented. For example, changes made to Version 7.4 before archiving Version 7.5 are detailed on the current revisions page. https://international.ipums.org/international-action/revisions

4. *Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?*

    - IPUMS datasets are continually revised to incorporate updates, corrections, or improvements. These changes are documented on the respective project's revisions page, such as for the IPUMS USA project.

5. *If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.*

    - IPUMS adheres to ethical standards and legal requirements concerning data retention. Its digital preservation framework emphasizes retaining data and metadata to meet archival requirements of funding agencies while ensuring data quality.

6. *Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.*

    - Yes, IPUMS archives annual snapshots of its datasets and documents changes between versions, ensuring that older versions remain accessible for research and reference purposes.

7. *If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.*

   - While IPUMS does not have a formal mechanism for external contributions to its datasets, it collaborates with national statistical offices and other organizations to collect and harmonize data.

8. *Any other comments?*

   - None

References

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. “Datasheets for Datasets.” *Communications of the ACM* 64 (12): 86–92.