Datasheet for IPUMS USA Dataset*

Xizi Sun

December 3, 2024

Extract of the questions from (Gebru et al. 2021).

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.*
    - This dataset was created to analyze the determinants of non-marriage in the United States. It addresses gaps in understanding how socio-demographic factors such as education, income, race, gender, and age influence marital status. Existing datasets lacked the level of detail and harmonization required for comprehensive modeling and analysis.
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 curated using the Integrated Public Use Microdata Series (IPUMS USA), a project of the Minnesota Population Center.
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 IPUMS USA project is supported by the National Institutes of Health (NIH) and the National Science Foundation (NSF).
4. *Any other comments?*
    - No.

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.*

[*] Code and data are available at: https://github.com/LilianS77/US_Marriage

- Each instance in the dataset represents an individual in the United States, including demographic, economic, and social variables such as marital status, education level, race, and income.

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).*

   - IPUMS USA's sampling strategy involves randomly generating a sample point for every hundred people in the census. The sampling strategy also ensures that dwellings have an equal probability of being included in the sample, regardless of their size.

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 record includes raw data (e.g., income in dollars, age in years) and processed features (e.g., binary marital status: "Married" or "Not Married").

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

   - The target variable is marital status, categorized as "Married" or "Not Married."

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.*

   - Missing data is minimal, with gaps primarily in income due to non-responses. The missing data are cleaned.

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.

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.*

- No specific splits are recommended for this dataset. However, for modeling purposes, researchers may opt to split the data into training, validation, and testing sets. A common practice would be an 80/10/10 split to ensure enough data for training while retaining sufficient records for validation and testing.

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

   - The dataset has been harmonized by IPUMS USA, minimizing errors and redundancies. However, self-reported variables like income and marital status may introduce noise due to inaccuracies or social desirability bias. Additionally, missing data in income and education fields could affect analyses, though imputation methods have been used to mitigate this.

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 external resources. It is available through the IPUMS USA platform, which provides archival access and ensures consistency over time. Usage is governed by the IPUMS user agreement, which prohibits redistribution and requires proper citation.

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.*

    - No, the dataset does not contain confidential information. It is anonymized and complies with legal and ethical standards for data protection. Identifiable information, such as names or addresses, has been removed by IPUMS to ensure privacy.

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 age, gender, race, income, and education level. These variables are explicitly labeled and allow for detailed demographic analyses.

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, the dataset is anonymized and does not allow direct or indirect identification of individuals. All personally identifiable information has been removed, and variables are aggregated or generalized to prevent re-identification.

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, income, and marital status. While these variables are anonymized and aggregated, they could still be sensitive due to their potential to reveal disparities or inequalities. No highly sensitive data, such as biometric information or government identifiers, is included.

16. *Any other comments?*

    - No.

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 primarily collected through survey responses as part of the American Community Survey (ACS), conducted by the U.S. Census Bureau. These responses were directly provided by individuals in the sampled households.

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 through a combination of mechanisms: Mail: Respondents received surveys to complete and return. Telephone: Follow-up interviews were conducted via phone for non-respondents. Personal Visits: Trained enumerators conducted in-person interviews for households that did not respond via mail or phone. Online Forms: Respondents could also complete the survey online. Validation was ensured through rigorous protocols, including automated consistency checks, interviewer training, and manual quality reviews.

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 is a 1% stratified random sample drawn from the larger ACS dataset. Stratification was based on demographic characteristics such as age, gender, race, and geographic location to ensure proportional representation of the U.S. population.

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)?*

   - Data collection was carried out by the U.S. Census Bureau, involving trained enumerators, survey administrators, and data processing staff. Enumerators and survey administrators are federal employees compensated according to U.S. government pay scales. No external contractors or crowdworkers were involved in this process.

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 for this study was collected as part of the 2023 American Community Survey. The ACS operates on a continuous basis, collecting data monthly throughout the year. The instances in this dataset reflect data collected from January to December 2023.

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.*

   - The ACS and its data collection procedures are subject to review by the U.S. Census Bureau and adhere to federal laws and guidelines, including the Confidential Information Protection and Statistical Efficiency Act (CIPSEA). These processes ensure compliance with ethical standards for data collection, privacy protection, and data use.

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 obtained from IPUMS USA, a third-party organization that harmonizes ACS data for research purposes. IPUMS USA receives data directly from the U.S. Census Bureau and applies its own harmonization procedures before making it available to researchers.

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 were notified as part of the ACS process. The U.S. Census Bureau provides clear information about the purpose of the survey, the voluntary nature of participation, and how the data will be used. This information is communicated via mailed instructions, online forms, and during personal interviews.

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.*

   - Participation in the ACS is required by law under Title 13 of the United States Code. While respondents are informed about the mandatory nature of the survey, they are also assured that their data will be used strictly for statistical purposes and protected under federal confidentiality laws.

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 participation in the ACS is legally mandated, there is no mechanism for respondents to revoke their consent. However, individuals are not required to answer every question, and any unanswered questions are handled through imputation or estimation.

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.*

    - The U.S. Census Bureau regularly conducts Privacy Impact Assessments (PIAs) to evaluate the risks associated with data collection and use. These assessments focus on maintaining respondent confidentiality and preventing misuse of data. IPUMS USA further ensures that datasets are anonymized and stripped of personally identifiable information before distribution.

12. *Any other comments?*

    - No.

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.*

- Yes, preprocessing was performed on the raw dataset obtained from IPUMS USA to prepare it for analysis. The missing data are removed. The gender variable was cleaned to male and female. The education variable was grouped into five categories: "Below High School," "High School," "Some College," "Bachelor," and "Above Bachelor."

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.
3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*
   - The preprocessing was conducted using R, with packages such as dplyr for data manipulation and arrow for efficient file handling. The code for all preprocessing steps is included in the project's GitHub repository, providing full reproducibility.
4. *Any other comments?*
   - No.

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*
   - Yes, the dataset was used to analyze the socio-demographic determinants of non-marriage in the United States, focusing on factors such as education, income, race, gender, and age.
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.*
   - Yes. Code and data are available at: https://github.com/LilianS77/US_Marriage.
3. *What (other) tasks could the dataset be used for?*
   - The dataset could be used for various tasks, including: Studying income inequality and its impact on social behaviors. Analyzing trends in educational attainment across different demographic groups. Exploring intersections of race, gender, and economic outcomes
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?*

- The dataset contains sensitive demographic variables such as race and income, which, if misused, could lead to stereotyping or biased interpretations. To mitigate these risks, users should follow ethical guidelines and contextualize findings appropriately.

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

    - The dataset should not be used for: Making deterministic conclusions about individuals. Stereotyping or reinforcing biases against specific demographic groups. Commercial purposes that violate the IPUMS terms of use.

6. *Any other comments?*

    - No

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.*

    - No, the dataset itself is not directly redistributed. Interested researchers can access the original data via IPUMS USA following their terms of use.

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 through the IPUMS USA platform in formats such as CSV and Parquet. It does not have a specific DOI but is accessible via the IPUMS website.

3. *When will the dataset be distributed?*

    - The IPUMS USA data is updated annually, with the current dataset representing data collected in 2023

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.*

    - Yes, the dataset is distributed under the IPUMS terms of use, which prohibit redistribution and require proper citation. There are no fees associated with academic use.

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.*

- Yes, the IPUMS terms of use restrict commercial use and redistribution of the dataset.

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 export controls or regulatory restrictions apply. The dataset is anonymized and adheres to all U.S. legal standards for public use data.

7. *Any other comments?*

   - No.

Maintenance

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

   - The dataset is hosted and maintained by the Minnesota Population Center through the IPUMS USA platform.

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

   - The dataset curators can be contacted via the IPUMS support email at ipums@umn.edu or through the support page on their website.

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

   - IPUMS USA regularly updates its documentation and data to address potential issues. Any errata are published on the IPUMS USA website under the "Revisions and Errata" section.

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)?*

   - Yes, the dataset is updated annually to include the latest ACS data. Updates are managed by IPUMS USA, and changes are communicated through the platform's release notes and mailing list.

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.*

   - The dataset contains anonymized data, and there are no specific retention limits since no personally identifiable information is included. The IPUMS platform ensures long-term availability for research purposes.

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.*

    - Older versions of the dataset are archived and accessible via the IPUMS platform. Researchers can access historical data spanning from 1850 to the most recent release to support longitudinal studies.

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 researchers cannot directly contribute to the IPUMS dataset, they can suggest improvements or report errors via the IPUMS feedback form. Extensions or augmentations must be managed independently and cannot be integrated into the IPUMS dataset due to its standardized format.

8. *Any other comments?*

    - No.

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.