Datasheet for 'A dataset'*

Dong Jun Yoon

Invalid Date

The income inequality in the United States from 2002 to 2019 will be observed using data from the United States Census and the World Inequality Database (WID). We will explore trends in income inequality, examining whether it is increasing or decreasing over time, and how the Great Recession of 2007-2009 affected income inequality. Investigating these trends is significant, as it could assist policymakers or the government in developing strategies that could decrease the inequalities occurring within the country. As a result, this report may be used to help promote social and economic equity in the United States.

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.*

    - The dataset "Income_inequality_social_class" and "Median_Income_Race" was created to enable analysis of income share and inequality within the United States. The dataset aims to analyze a detailed analysis of the income share and median income over different social classes and race.

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

    - The dataset is created from the World Inequality Database (WID) and United States Census Bureau.

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 dataset is funded by the public and non profit institutions.

4. *Any other comments?*

*Code and data are available at: https://github.com/DJY1231/Exploring-Income-Inequality-in-US.git.

- The paper focuses more on "Income_inequality_social_class" and does not really focus on the "Median_Income_Race" dataset.

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 in this dataset represent changing dynamics of income distribution across different social classes in the United States from 2002 to 2019. Specifically, the data shows an overall trend in the income share for the top 1%, top 10%, and bottom 50% of earners, providing insights into income distribution across different social groups.

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

    - The dataset has 54 instances.

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 contains a sample of instances than all possible instances that contributes the income share and income inequality. The sample is likely to be representative of a larger set, allowing to cover most of the social class for each income group.

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 is consist of year, social class and income share of corresponding gorup of people in the United States.

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

    - No

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.*

    - No

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 dataset does not build up any explicit model relationship between individual instances. Each instance in the dataset represents the total answer for each single respondent's survey answer.

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

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

    - Errors, sources of noise, or redundancies in the dataset may arise from sampling biases or errors that may occur within the report.

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.*

    - No, the dataset does not link to any external source. The dataset rely on suvery collection of individual participants.

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 data is protected by confidentiality rules that prevent individuals from being identified.

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.*

- No

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

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.*

    - The dataset contains income statistics over different social class of the United States. This data may be considered sensitive.

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 associated with each instances in the WID report, where the data was collected through survey responses directly by the subjects. The WID gathers information from outsources such as national annual reports of the income inequality worldwide and analyze it based on those reports.

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?*

    - Surveys and national accounts data are collected using questionnaires (both paper-based and electronic), interviews, and in some cases, digital entry forms. Countries may use computer-assisted personal interviewing (CAPI) systems or computer-assisted telephone interviewing (CATI) systems.

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

- Most income and wealth surveys use probabilistic sampling methods. This involves selecting a random, representative sample of the population, where every individual has a known probability of being included. Common techniques include stratified sampling, cluster sampling, and multi stage sampling.

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)?*

    - For national accounts and survey data, national statistical agencies typically carry out the collection of national accounts and survey data. Employees of these agencies are usually civil servants or government employees. In some cases, especially for specific surveys, external contractors or dedicated survey firms might be employed.

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 are usually collected annually or every few years, depending on the specific survey or national account reporting requirements. The data reflects economic activities within the specific year or period they were collected.

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 dataset does not provide any explicit details regarding any ethical review processes.

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 dataset was collected directly from individuals through survey responses.

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.*

    - For surveys, individuals are typically notified about the data collection directly at the time of the survey. The notification informs participants about the purpose of the survey, how the data will be used, and reassures them of confidentiality and voluntary participation.

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.*

- For survey data, which often involves direct interaction with respondents, informed consent is usually obtained at the beginning of the survey process. Respondents are informed about the purpose of the survey, how their data will be used, and their rights, including the right to withdraw from the survey at any time.

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).*

    - The dataset does not provide specfic information regarding to given mechanism. However, individuals participating in surveys usually have the option to withdraw from the survey at any time, and this option is typically explained at the time of obtaining consent.

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 research institution tend to assess the impact of the data collection on the participants. This includes considering privacy concerns, the risk of data breaches, and potential misuse of data.

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.*

    - No

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.*

    - No

3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*

    - No

4. *Any other comments?*

- No

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*
   - No
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.*
   - No
3. *What (other) tasks could the dataset be used for?*
   - This dataset could be used for analyzing the income inequality within the United States or comparing with different country.
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 may have sampling biases. The survey data may not perfectly represent the entire population, especially if certain groups are underrepresented due to non-response or accessibility issues. This could lead to skewed analyses and potentially unfair conclusions or policies based on this incomplete picture. So, researchers and policymakers using the data should apply statistical techniques to adjust the sampling biases.
5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*
   - No
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.*

- Yes, the dataset will be distributed to third parties outside of the entity. It will be open to individuals, researches, distribution, policy makers, and to other parties.

2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*
    - Data can be downloaded directly from the WID.world website, often in formats like CSV or Excel, which are useful for researchers and policymakers. WID.world might not necessarily assign a DOI to the entire database.
3. *When will the dataset be distributed?*
    - Every annual season when they will use the data to compare their model.
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.*
    - No
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
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
7. *Any other comments?*
    - No

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*
    - The World Inequality Database (WID) is supported, hosted, and maintained by a collaborative network of economists and researchers primarily affiliated with the Paris School of Economics, along with other academic institutions worldwide.
2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*

- Directly contact info@wid.world.

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

   - There is no erratum.

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)?*

   - The dataset will be updated annually to correct any errors.

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.*

   - There are no specific limits on the retention of individual data within the dataset.

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 generally kept accessible to ensure that researchers can replicate past analyses and compare trends over time. This is crucial for economic studies where longitudinal data consistency is necessary.

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.*

   - The guide doesn't not include a explicit mechanism for others to extend to the dataset.

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.