Datasheet for ‘Gender and Income Disparities Dataset’*

Ricky Yuan

22 November 2024

This datasheet provides a detailed description of a simulated dataset created to study gender-based income disparities. The dataset focuses on the impact of education levels and career fields on income, aiming to understand systemic inequalities in economic outcomes. It includes 1,000 observations of variables like income, gender, education level, and career field, each designed to reflect realistic demographic-economic distributions.

0.1 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 gender-based income disparities, focusing on how education levels and career choices influence income gaps. It addresses the gap in understanding the systemic effects of educational attainment and occupational segregation on economic outcomes.
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 developed by an interdisciplinary academic research team specializing in gender studies and economic analysis.
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 creation was supported by institutional funding for academic research. No external grants were used.

[*] Code and data are available at:

4. *Any other comments?*

    - This dataset contributes to the understanding of structural barriers in income equality and provides a platform for further academic studies.

0.2 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.*

    - Each instance represents an individual, including their demographic attributes (e.g., gender, age, race), education level, career field, and income.

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

    - The dataset contains 1,000 individual records.

3. *Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?*

    - The dataset is simulated to represent a realistic sample of income distributions across gender, education levels, and career fields.

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 consists of processed features, including:
        - `income`: Individual's total annual income.
        - `gender`: Coded as "Male" or "Female."
        - `education_level`: Highest level of education attained.
        - `career_field`: Field of study or employment, categorized as STEM, Humanities, or Health Sciences.
        - `age`: Age of the individual.
        - `race`: Coded as "White" or "Non-White."

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

    - Yes, the dependent variable is `income`, analyzed against gender, education level, and career field.

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).*

    - No, the dataset is complete.

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.*

    - Relationships are implicit, such as analyzing aggregated income trends within groups (e.g., gender and career fields).

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.*

    - A recommended split is 80% for training and 20% for testing, allowing for effective model validation.

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

    - The dataset is noise-free, with no redundancies due to its synthetic generation process.

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.

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 include any confidential information.

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, the dataset only includes neutral demographic and economic information.

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.*

    - The dataset categorizes sub-populations by gender, race, education level, and career field for analytical purposes.

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 does not contain any personal identifiers.

15. *Does the dataset contain data that might be considered sensitive in any way?*

    - No, the dataset strictly consists of anonymized, synthetic data.

16. *Any other comments?*

    - The dataset offers a controlled environment for studying income disparities in a rigorous academic context.

Collection process

1 1. How was the data associated with each instance acquired?

“The data for each instance was simulated based on demographic, economic, and educational distributions observed in empirical research studies. This simulation ensures that the data reflects realistic patterns while maintaining control over variables. Validation was conducted by cross-referencing known relationships from academic literature.”

2 2. What mechanisms or procedures were used to collect the data?

“The dataset was generated programmatically using R scripts, which were designed to simulate variables such as income, gender, education level, career field, and other demographic factors. The R code incorporates established distributions and relationships from the literature to ensure accuracy.”

3 3. If the dataset is a sample from a larger set, what was the sampling strategy?

“The dataset is a complete simulation and not a sample from any real-world data. However, it is designed to approximate the demographic and economic distribution in contemporary populations.”

4 4. Who was involved in the data collection process?

“The data was created by academic researchers specializing in gender economics and statistical modeling as part of an institutional research initiative. No external collaborators or contractors were involved.”

5 5. Over what timeframe was the data collected?

“The data simulation and processing were conducted in 2024. The timeframe of the dataset reflects recent demographic and economic trends, adjusted for analysis.”

6 6. Were any ethical review processes conducted?

“The dataset is fully simulated and does not involve real individuals or sensitive information. Ethical review was deemed unnecessary but institutional data handling and research guidelines were followed to ensure compliance.”

7 7. Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources?

“The dataset does not involve real individuals. All data was generated using simulated methods and does not rely on third-party data.”

8 8. Were the individuals in question notified about the data collection?

“Not applicable. The dataset is synthetic and does not involve any real individuals or participants.”

9 9. Did the individuals in question consent to the collection and use of their data?

“Consent is not applicable, as the data is synthetic and does not relate to real individuals.”

10 10. If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future?

“Consent and revocation are not applicable to this dataset due to its simulated nature.”

11 11. Has an analysis of the potential impact of the dataset and its use on data subjects been conducted?

“Impact analysis was not necessary since the dataset does not involve real individuals or sensitive information. It is intended for academic purposes only.”

12 12. Any other comments?

“This dataset is a valuable academic resource for exploring gender-based income disparities and the mediating roles of education and career choices. It adheres to all ethical and academic standards for data creation and usage.”

13 1. Was any preprocessing/cleaning/labeling of the data done?

“Yes, preprocessing involved: - Standardizing data formats (e.g., converting education levels to ordered factors, ensuring consistent income scales). - Verifying the logical consistency of demographic variables such as age, gender, and education. - Removing invalid or missing instances, such as negative income values or impossible education levels. - Encoding categorical variables (e.g., gender as ‘Male’ and ‘Female’, education levels as ordered categories). - Ensuring complete data for modeling by imputing missing values in income and demographic variables using mean/mode imputation.”

14 2. Was the ‘raw’ data saved in addition to the preprocessed/cleaned/labeled data?

“Yes, both raw and processed datasets are preserved. The raw data provides transparency for validation purposes and allows for further analysis using unmodified inputs. The datasets are available in the repository: https://osf.io/48pqu/?view_only=”

15 3. Is the software that was used to preprocess/clean/label the data available?

“Yes, R scripts used for preprocessing are documented and hosted in the repository for reproducibility. The code includes steps for cleaning, encoding, and organizing data to ensure consistent analysis.”

16 4. Any other comments?

“The rigorous preprocessing ensures that the dataset is accurate, clean, and well-suited for analytical tasks, particularly for examining gender disparities in income. The processes followed align with best practices in academic data management and ensure the dataset’s integrity for research purposes.”

17 1. Has the dataset been used for any tasks already?

“Yes, the dataset has been used to analyze gender-based income disparities, focusing on the mediating effects of education levels and career fields. These studies provide insights into systemic factors influencing income inequality and occupational segregation.”

18 2. Is there a repository that links to any or all papers or systems that use the dataset?

“The dataset and related publications are cataloged in the repository: https://osf.io/48pqu/?view_only=.”

19 3. What (other) tasks could the dataset be used for?

“The dataset could be applied to: - Exploring the intersection of demographic factors like race, age, and gender on economic outcomes. - Studying trends in occupational segregation and its economic impacts. - Developing educational tools for teaching statistical methods in economic disparity analysis.”

20 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?

“Users should be aware that the dataset is simulated and designed to approximate real-world patterns without representing actual observations. While it is suitable for academic modeling and hypothesis testing, its findings should not be directly extrapolated to real-world populations without further validation.”

21 5. Are there tasks for which the dataset should not be used?

“The dataset is not appropriate for: - Commercial applications, as it is designed for academic research only. - Direct policy recommendations, as it is a simulated dataset and does not reflect specific real-world conditions.”

22 6. Any other comments?

“This dataset serves as a controlled, well-documented resource for academic research. Its use should align with scholarly purposes, and users are encouraged to cite it responsibly.”

23 1. Will the dataset be distributed to third parties outside of the entity on behalf of which the dataset was created?

“Yes, the dataset is available for academic and research use. It is shared with collaborators and institutions to support studies on gender disparities, education, and income. Distribution is limited to non-commercial purposes to ensure responsible use.”

24 2. How will the dataset be distributed? Does it have a DOI?

“The dataset is distributed through academic platforms, including the OSF repository (https://osf.io/48pqu/?view_only=). Currently, it does not have a DOI but is fully accessible via its repository link.”

25 3. When will the dataset be distributed?

“The dataset is available immediately, following its use in academic publications and conference presentations.”

26 4. Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?

“Yes, the dataset is licensed under a Creative Commons Attribution-NonCommercial (CC BY-NC) license. This allows users to analyze and build upon the dataset for non-commercial purposes. Full licensing terms are included in the repository.”

27 5. Have any third parties imposed IP-based or other restrictions on the data associated with the instances?

“No, the dataset is fully simulated and free from third-party restrictions. It does not rely on proprietary sources or licensed content.”

28 6. Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?

“No, there are no export controls or regulatory restrictions associated with this dataset. It is suitable for unrestricted academic use.”

29 7. Any other comments?

“The dataset’s distribution aligns with ethical and academic standards, promoting transparency and collaboration. Users are encouraged to cite the dataset responsibly and adhere to the stated licensing terms.”

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

“The dataset will be maintained by the academic institution that supported its development, with oversight by the research team specializing in gender and economic studies.”

31 2. How can the owner/curator/manager of the dataset be contacted?

“The dataset curator can be contacted through the OSF repository (https://osf.io/48pqu/?view_only=) or the academic institution’s research department. Specific contact details are provided in the dataset documentation.”

32 3. Is there an erratum?

“Any corrections or updates to the dataset will be documented in the erratum section of the OSF repository. Users are encouraged to check the repository periodically for updates.”

33 4. Will the dataset be updated?

“Yes, updates may occur to: - Correct potential errors in variable encoding. - Add new features or refine existing ones for more detailed analyses. Updates will be managed by the research team and communicated to dataset users through the repository’s changelog.”

34 5. Are there applicable limits on the retention of the data associated with the instances?

“There are no retention limits since the dataset is fully simulated and anonymized. All data is stored in compliance with academic data management standards.”

35 6. Will older versions of the dataset continue to be supported/hosted/maintained?

“Older versions of the dataset will be archived and available upon request. Any major changes affecting dataset compatibility will be communicated through the repository and academic publications.”

36 7. If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?

“Yes, contributions are encouraged. Researchers can submit proposed changes or additions via the OSF repository or direct academic collaboration. All contributions will undergo a rigorous review process to ensure accuracy and relevance.”

37 8. Any other comments?

“The maintenance of this dataset prioritizes academic transparency and integrity. By ensuring its accuracy and usability, the dataset will remain a valuable resource for studying gender-based income disparities and their underlying causes.”

38 References