Datasheet for ‘The evolution of fertility rates in the United States’*

Yuchen Chen

April 3, 2024

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 creation of this dataset is to analyze the fertility rate in the United States, assist the government in formulating effective policies that may increase the country’s fertility rate. The main goal is to truly study how factors such as per capita GDP and female employment rate in the United States affect fertility rates. But this dataset only focuses on the United States and only covers the period from 2002 to 2021, so the results of this study have certain country and time limitations.
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
    - This dataset is sourced from the World Bank database and is compiled based on officially recognized international sources. It has been available since 1960 and is updated annually..
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.*
    - No information on website.

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/Victor1114/The-evolution-of-fertility-rates-in-the-United-States.git

- In studies like fertility rates in the United States, these examples represent the situation from 2002 to 2021, involving fertility rates, per capita GDP, and female employment rates..

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

    - 20 years.

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).*

    - For a country, there is only one annual fertility rate, per capita GDP, and female employment rate, thus including all possible examples.

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.*

    - fertility rate, year, per capita GDP, and female employment rate.

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

    - **Fertility Rate**: Fertility rate (per capita female fertility).
    - **Per Capita GDP**: The age of the voter at the time of the survey.
    - **Female Employment Rate**: per capita GDP (in current US dollars).

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.*

    - It does not exist because a country's fertility rate, per capita GDP, and female employment rate data are only available once a year.

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.*

    - Clear relationships refer to data from the same country at different times.

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 main objective of this paper is to analyze the influencing factors of fertility rate, specifically the relationship between fertility rate, per capita GDP, and female employment rate, rather than the predicted results. Therefore, this article does not do data splitting. But we did some testing to ensure that the dataset we cleaned is good and complete for use.

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

    - There may not be any noise present.

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.*

    - This article mainly utilizes the World Bank database, which rely on the websites World Bank database.

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.*

    - It is not be considered confidential, because everyone can get data from World Bank Dataverse.

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.*

    - This dataset does not contain offensive, insulting, threatening, or might otherwise cause anxiety, but social topic covered in the survey could potentially be sensitive or provoke strong emotional responses in individuals with differing views.

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.*

    - **Fertility Rate**: Fertility rate (per capita female fertility).
    - **Per Capita GDP**: The age of the voter at the time of the survey.
    - **Female Employment Rate**: per capita GDP (in current US dollars).

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.*

    - This is impossible because it is the overall data of a country.

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.*

    - **Per Capita GDP**: Economic data that includes social development.

16. *Any other comments?*

    - TBD

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 dataset was collected by the World Bank from various publicly available databases.

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 is compiled based on officially recognized international sources of data.

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

    - This is not sampling data.

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)?*

    - Compiled by the World Bank, the specific collection process is unknown.

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 is updated annually and updated year by year.

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 and its guide doesn't specifically mention the details of 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 data comes from third parties websites, the World Bank database.

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.*

   - The data does not involve individuals.

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.*

   - The data does not involve individuals.

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 data does not involve individuals.

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.*

    - No mentioned in website or materials, but usually academic studies incorporate ethical considerations and privacy protections.

12. *Any other comments?*

    - TBD

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.*

- here are no Preprocessing/Cleaning/Labeling of Data showing in website or materials , but I cleaned it when I use it to analyze. I only filtered the data from 2002 to 2021 in the new dataset.

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.*

   - https://data.worldbank.org.cn/indicator/SP.DYN.TFRT.IN?view=chart&locations=US
   - https://data.worldbank.org.cn/indicator/SL.EMP.TOTL.SP.FE.NE.ZS?locations=US
   - https://data.worldbank.org.cn/indicator/NY.GDP.PCAP.CD?view=chart&locations=US

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

   - https://www.r-project.org/

4. *Any other comments?*

   - TBD

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*

   - This data is widely used by researchers and policymakers.

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.*

   - Most people only use some or more data, and we will not give any examples for now.

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

   - Economic development and gender discrimination.

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?*

   - Will not affect future use.

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

- Not present.

6. *Any other comments?*

    - TBD

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 data is publicly available.

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

    - This data is obtained from the World Bank database and can be downloaded in CSV format.

3. *When will the dataset be distributed?*

    - 2024 update.

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.*

    - Any people can copy, modify, distribute and perform the work, even for commercial purposes, all without asking permission.

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 any limitation.

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 any limitation.

7. *Any other comments?*

    - TBD

Maintenance

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

    - World Bank.

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

    - On the website https://datatopics.worldbank.org/world-development-indicators/ , there is an option called "contact".

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

    - No findings about erratum in website.

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 is updated annually with new data for the current year, which can be directly obtained from the World Bank's official website.

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 data does not involve individuals.

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.*

    - The data will be maintained because a new year's data is added to the previous data every year.

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.*

    - They can try to contact website administrators.

8. *Any other comments?*

    - TBD