Datasheet for the child care of working mothers dataset*

Pascal Lee Slew Yunkyung Park

03 April 2022

Abstract

The datasheet serves the purpose of documenting the creation of a dataset. It has two main objectives, one for data creators and the other for data consumers. For data creators, it allows the reflection on the data gathering and cleaning process. While for data consumers, it provides an understanding of how the dataset was created. The key idea behind it would be transparency and accountability.

The questions were extracted 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 was created to enable a descriptive analysis of the child care situation of working mothers in Ghana. The dataset would be used to look at the trends concerning child caretaking across different factors such as region, work status or occupation area. We were unable to find a publicly available dataset.
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 Yunkyung Park and Pascal Lee Slew from the University of Toronto.
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 funding was obtained/required for the creation of the dataset.
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 dataset consists of categories: residence (urban or rural), regions of Ghana, mothers' education level, work status, occupation and employment status. For each category, we have different features: child status (whether families have children under six years old), childcare preferences and the number of employed women.
2. *How many instances are there in total (of each type, if appropriate)?*

*Code and data are available at: https://github.com/Pascal-304/dhs_analysis.

- There are 26 rows and 17 columns in the dataset. The residence category has 2 rows, regions has 10 rows, mothers' education level has 4 rows, work status has 3 rows, occupation has 2 rows and employment status has 4 rows.

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 all the data possible.

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 category consists of already processed data. The observations are mostly percentages which represent the proportions of categories.

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.*

   - Yes, some participants did not respond to all survey questions, leading to some missing data in the data collection process.

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.*

   - No

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.*

   - There may be possible sampling error.

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 was extracted from a publicly available pdf at https://dhsprogram.com/publications/publication-FR106-DHS-Final-Reports.cfm

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 results of the survey used to create the dataset were made confidential so that we are unable to identify the participants.

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.*

- None

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 population would be the working mothers. Sub-populations would be identified by the categories mentionned above such as employment status, occupation, mothers' education, region, residence and work status.
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.*
    - No
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.*
    - No, the data are percentages of categories based on results from the 1998 Ghana Demographic and Health Survey responses.
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 by extracting data from a pdf. In our case, we extracted data from a table located at page 45 in the 1998 Ghana Demographic and Health Survey pdf available at the link: https://dhsprogram.com/publications/publication-FR106-DHS-Final-Reports.cfm
    - The data collection was done using R (R Core Team 2020) and R packages tidyverse (Wickham et al. 2019), pdftools (Ooms 2022) and stringi (Gagolewski 2021).
    - We used the R package pointblank (Iannone and Vargas 2022) to set up tests and check on our dataset.
3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?*
    - No
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 raw survey data was collected by teams consisting of a team supervisor, a field editor, three interviewers and a driver. They were trained for 3 weeks. No mention of compensation was made.
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 raw GDHS data was collected from mid-November 1998 to mid-February 1999.

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 GDHS procedures and survey questionnaires were reviewed by the ICF Institutional Review Board (IRB) of Ghana.
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 GDHS raw data was collected from individuals directly while the dataset created was created from a publicly available pdf on the DHS website.
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, the individuals in questions were informed since data collection of the GDHS survey was in person. However, for the "raw" data from the pdf, it did not involve individuals directly since their responses were already processed together.
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 the GDHS survey, the participants consented to the collection and use of their data. Their data would be made anonymous and public but they were not explicitly informed that it would be used in this way.
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).*
    - N/A
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.*
    - N/A
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.*
    - Yes. The "raw" data obtained from the pdf was cleaned using the tidyverse r package (Wickham et al. 2019) in R (R Core Team 2020).
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.*
    - Yes, both the clean and "raw" data are saved and available at the link: https://github.com/Pascal-304/dhs_analysis/tree/main/outputs/data

3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*
    - We used the statistical software R (R Core Team 2020).
4. *Any other comments?*
    - None

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*
    - The dataset has been used to conduct a descriptive analysis of the child care situation of working mothers in Ghana. We used the dataset to create graphs.
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.*
    - The repository containing all the files pertaining to the dataset is available at https://github.com/Pascal-304/dhs_analysis
3. *What (other) tasks could the dataset be used for?*
    - None since the dataset pertained to the childcare situation of working mothers in Ghana only at one time point(1998), so it cannot be used to generalize in other countries or other time points.
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?*
    - N/A
5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*
    - N/A
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, the dataset is publicly available on the internet.
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 can be obtained by running the r script "01-gather_data.R" available at the link: https://github.com/Pascal-304/dhs_analysis/blob/main/scripts/01-gather_data.R
3. *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

4. *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.*
    - Unknown to the authors of the dataset
5. *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.*
    - Unknown to the authors of the dataset

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*
    - Yunkyung Park and Pascal Lee Slew
2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*
    - If anyone want to reach out to us about the dataset, you can contact us at clara.park@mail.utoronto.ca or pascal.leeslew@mail.utoronto.ca
3. *Is there an erratum? If so, please provide a link or other access point.*
    - No
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)?*
    - If there is any error made in the creation of the dataset, the dataset will be updated by either Yunkyung Park or Pascal Lee Slew. Any updates will be communicated as a note in the Readme file at the link below: https://github.com/Pascal-304/dhs_analysis/blob/main/README.md
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.*
    - N/A
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 dataset has just been created. If there are any updates, older versions will still be kept for consistency.
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.*
    - Others may do so and should contact the original authors about incorporating fixes/extensions.
8. *Any other comments?*

- None

References

Gagolewski, Marek. 2021. *Stringi: Fast and Portable Character String Processing in r.* https://stringi.gagolewski.com/.

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.

Iannone, Richard, and Mauricio Vargas. 2022. *Pointblank: Data Validation and Organization of Metadata for Local and Remote Tables.* https://CRAN.R-project.org/package=pointblank.

Ooms, Jeroen. 2022. *Pdftools: Text Extraction, Rendering and Converting of PDF Documents.*

R Core Team. 2020. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.