Education in Ghana in the 20th Century Datasheet*

The Education Level in Diiferent Age Group and Gender

Shengyi Dai Suofeiya Guo Zifeng Zhu

08 April 2022

Abstract

This report is going to find the relationship between age, gender and educational attainment in Ghana in 1998. After some analyses we get that men were generally more educated than women, and at the same time, older people were much less educated than younger people. Education level is very important in today's society, it can change a person's mind and life, and it can also change the economic level of a country. From this report we can see that in Ghana in 1998, most people did not review a higher education, but today, education has become widespread all over the world, and prejudice between men and women is gradually decreasing.

Contents

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 collect data on fertility, family planning, and maternal and child health.
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
   - The Ghana Statistical Service created the dataset.
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 funded the creation of dataset is United Nations Fund for Population Activities (UNFPA), the United States Agency for International Development (USAID), and other donor agencies to implement a number of population-related activities.
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.*

- The instances represent Ghana. The types are: Western, Central, Greater Accra, Volta, Eastern, Ashanti, Brong Ahafo, Northern, Upper West, and Upper East, corresponding to different regions of the country.

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

- There are no data show the number of instances.

*Code and data are available at: LINK.

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 does contain all possible instances.

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 5 variables: gender, age, residence, region and education level.

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.*

- There are no relationships between individual instances.

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.*

- There are no recommended data splits.

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

- There are no errors, sources of noise, or redundancies in the dataset.

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.

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.*

- There is no confidential data, and the dataset is available to public.

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.*

- In Table 2.4, the word "Father dead" and "Mother dead" might be offensive because the word "dead" might remind people who lost their parents in a negative way.

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 survey was designed to obtain completed interviews of 4,500 women age 15-49. In addition, all males age 15-59 in every third selected household were interviewed, to obtain a target of 1,500 men.

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.*

- It is not possible to identify individuals in any way.

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.*

- Sensitive columns may include but not limited to: "Father dead", "Mother dead" and "Both dead".

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.*

- These data were acquired from survey responses,, the data were validated since if there are any mistakes workers will come back and get them again.

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?*

- Manual human curation had been used.

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 sample was randomly selected from households.

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)?*

- There are a lot of workers that had involved in data collection, they got paid and trained by the government.

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.*

- Household listing began in August and lasted for about two months.

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.*

- Not sure.

7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*

- I collect the data from the DHS project 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.*

- The individuals voluntarily participated in the interview with data collectors. Also, the notice of data collection is not provided.

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 individuals consented to the collection and use of their data, however, the exact language to which the individuals consented is not available.

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).*

- A mechanism to revoke consent was not provided.

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.*

- An analysis of the potential impact of the dataset and its use on data subjects was not conducted.

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.*

- The data was obtained in PDF format originally. A summary table on page 12 from the survey PDF was converted to a useable data frame in R using the library pdftools.

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 obtained from the PDF is saved in inputs/data/raw_data.csv

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

- R software is available at https://www.R-project.org/

4. *Any other comments?*

- The library used to convert data from PDF is available at https://docs.ropensci.org/pdftools/

Uses

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

- The dataset has not been used for other tasks yet.

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.*

- http://dhsprogram.com/pubs/pdf/FR106/FR106.pdf

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

- The dataset can be used for analyzing educational level among all male and female age groups in Ghana in 1998.

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 cleaning process is very specific to consider how this table was formatted, in which case it might not work on other tables.

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

- The dataset would not be appropriate for purposes other than studying the educational level in 1998 Ghana.

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.*

- No, the dataset is available to public and being used for personal uses only.

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 will be distributed using Github.

3. *When will the dataset be distributed?*

- The dataset wwill be distributed in April 2022.

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.*

- The dataset will be released under MIT license.

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.*

- There are no restrictions.

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 such controls are applicable.

7. *Any other comments?*

- None.

Maintenance

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

- Ghana Statistical Service hosting the dataset.

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

- Can be contacted via github.

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

- There is no currently available 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)?*

- There is no plan to update the dataset.

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 applicable limits as people in Ghana voluntarily participate.

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 older versions will not be supported. Dataset consumers could check the updates through github commit history.

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.*

- Currently there is no mechanism for accepting other contributions.

8. *Any other comments?*

- None.