The Education Level is Unbalanced in Philippines in 1993*

Zihao Liu

2 April 2022

Abstract

The education level varies a lot in different ages and regions in Philippines in 1993. This report uses the education level data of males to do analysis, and found out that male adults who are youger have a longer time of education received, and are more likely to have a college degree. Also, the regions with a higher rate of having college degree have a lower rate of haivng no education experience.This report can be used to assist the social studies related to education, and can be used to do comparative study with the education level in other countries.

1 Introduction

In Metro,Manila in 1993, a significant number of people had college degree, but there are still a portion of the people who have never received education before ((NSO) and (MI) 1994). The education level varies a lot in different regions in the country, and the level is quite different between urban and rural areas ((NSO) and (MI) 1994). The education level is unbalanced in Philippines in 1993. And this report is going to explore where and how it is unbalanced.

The report uses the information from demographic survey data of Philippines in 1993 ((NSO) and (MI) 1994). An analysis was performed on the data, and I found out that the male adults who are younger are more likely to have a college degree or higher, the male adults who are older are more likely to have no education experience or only a education level of elementary school, and the younger the male adults are the longer education time they received. Besides the unbalancedness on age, the regions with higher rate of college degree have a lower rate of no education, and the rate of having college degree or higher in urban area is more than two times of the rate in rural area.

The setup of data will be in the Data section. The analysis and explanation will be in the Results section. The Discussion section will include a overview of the paper, summary of major points, limitation of the report and next steps. The report was created using R programming language (R Core Team 2020). R packages `tidyverse`(Wickham et al. 2019), `knitr`(Xie 2014), `dplyr`(Wickham et al. 2022), `ggplot2`(Wickham 2016),`here`(Müller 2020), and `gridExtra`(Auguie 2017) were used to create this report.

2 Data

The data for this report comes from a full page table in *National Demographic Survey Philippines in 1993*(NDS) ((NSO) and (MI) 1994). The table exhibits the education level of male Filipinos in 1993. The data for NDS were collected by surveys and interviews, the sample selection is planned according to the 14 regions, "using the 1990 Population Census data on population size" ((NSO) and (MI) 1994). The data might not be a complete representation of the nation, because one of the 15 regions in the country just formed and no sample were taken in that region ((NSO) and (MI) 1994). Therefore, the data is only a representative of the 14 regions, but not the whole population in Philippines in 1993.

The image of the table was transferred to text using `tesseract`(Ooms 2022). And the text was processed to raw datasets using `tidyverse`(Wickham et al. 2019), `janitor`(Firke 2021), `purrr`(Henry and Wickham

[*] Code and data are available at: https://github.com/ZihaoLiu2/Education_Statistics_Philippines_1993

2020),and `stringi`(Gagolewski 2021). The tests for the variables were done using `pointblank`(Iannone and Vargas 2022). Since the table exhibits the education level information by age, residence and regions, I made three datasets for each type. The datasets have 9 variables, which are age/residence/region, the percentage of no education, elementary school, high school, college or higher education, do not know or missing, and Total percentage, Total number, Median number of years of education. For age, there are 13 observations. For residence, there are 2 observations. For regions, there are 15 observations (14 regions and 14 regions in total).

Table 1: Table 1:First row of the three datasets(combined), education level of males by age/residence/region

Age	No education	elementary	high school	college+	dont know	Total	number	median years
6-9	38	59	0	0	3	100	3660	1

residence	No education	elementary school	high school	college+	dont know	Total	number	median years
Urban	6	39	33	22	0	100	13942	8

region	No education	elementary	high school	college+	dont know	Total	number	median years
MetroManila	4	26	39	31	0	100	3732	10

Table 1 is formed by the first row of the three datasets. In table 1, Age represents the age groups of the males, residence represents where the males are from, ether rural or urban area, and region represents the which region it is from the 14 regions. No education represents the percentage of males who have no education experience, elementary school represents the percentage of males with elementary school as their highest education level. High school, college or higher are similar to the representation of elementary school. Total is the total percentage, which is always 100.Number represents the total number of males for each age group/residential type/region. Median years represents the median number of years of education received.

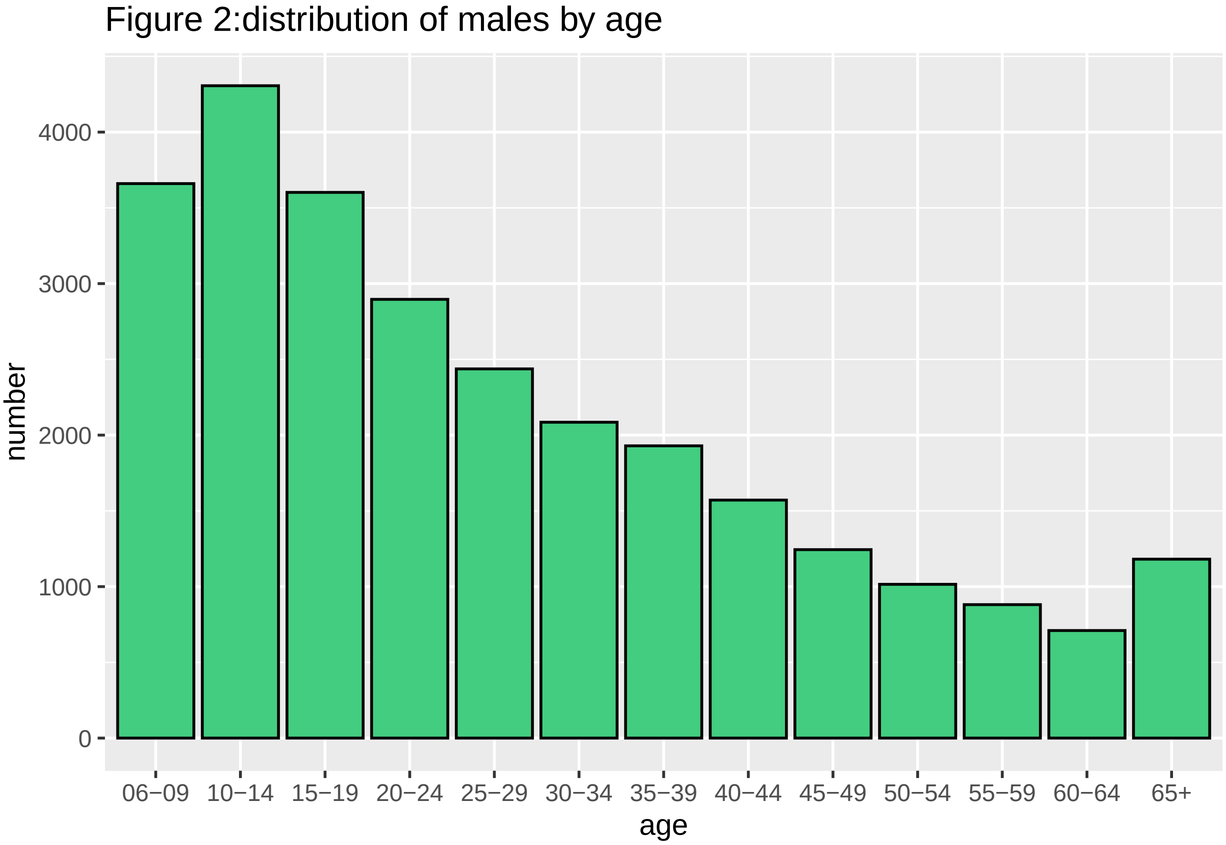


Figure 2 exhibits the distribution of number of males by age groups. The distribution peaks in the age group of 10-14 years old, and gradually decreasing as age gets bigger. The plot shows that a big portion of the males are youth and the relatively more aged group(50+) only occupies a small portion of the population in 1993.

Figure 3:distribution of males by regions

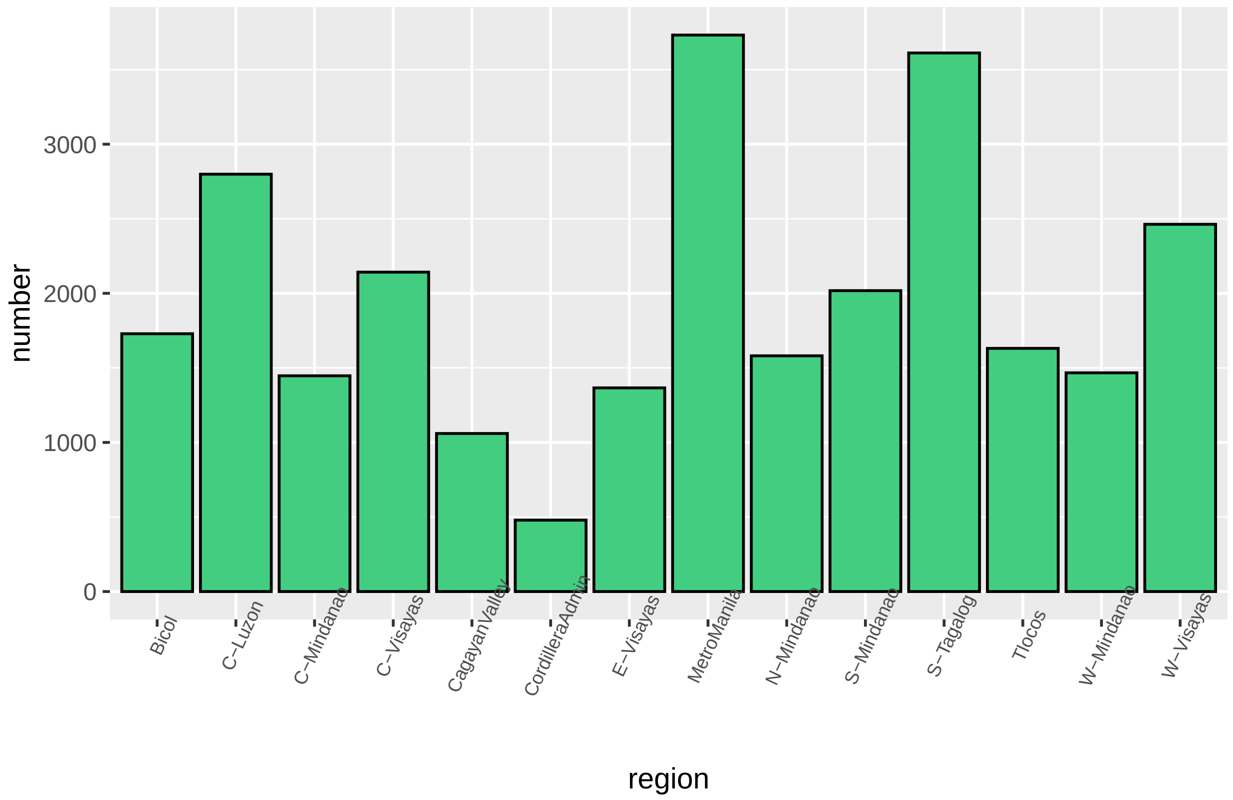


Figure 3 exhibits the distribution of number of males in the 14 regions in Philippines in 1993. From the plot, *Metro,Manila* has the highest number, whereas *Cordillera Admin.* has the lowest number. The distribution is not concentrated, there are only a few peaks, and the distribution is moderately spread out. For the distribution of residence, there are only two residential types, which are rural and urban. There are 13942 males from urban area and 13583 from rural area. The number is quite close, which might imply that approximately half of the males live in urbanized area in 1993.

3 Results

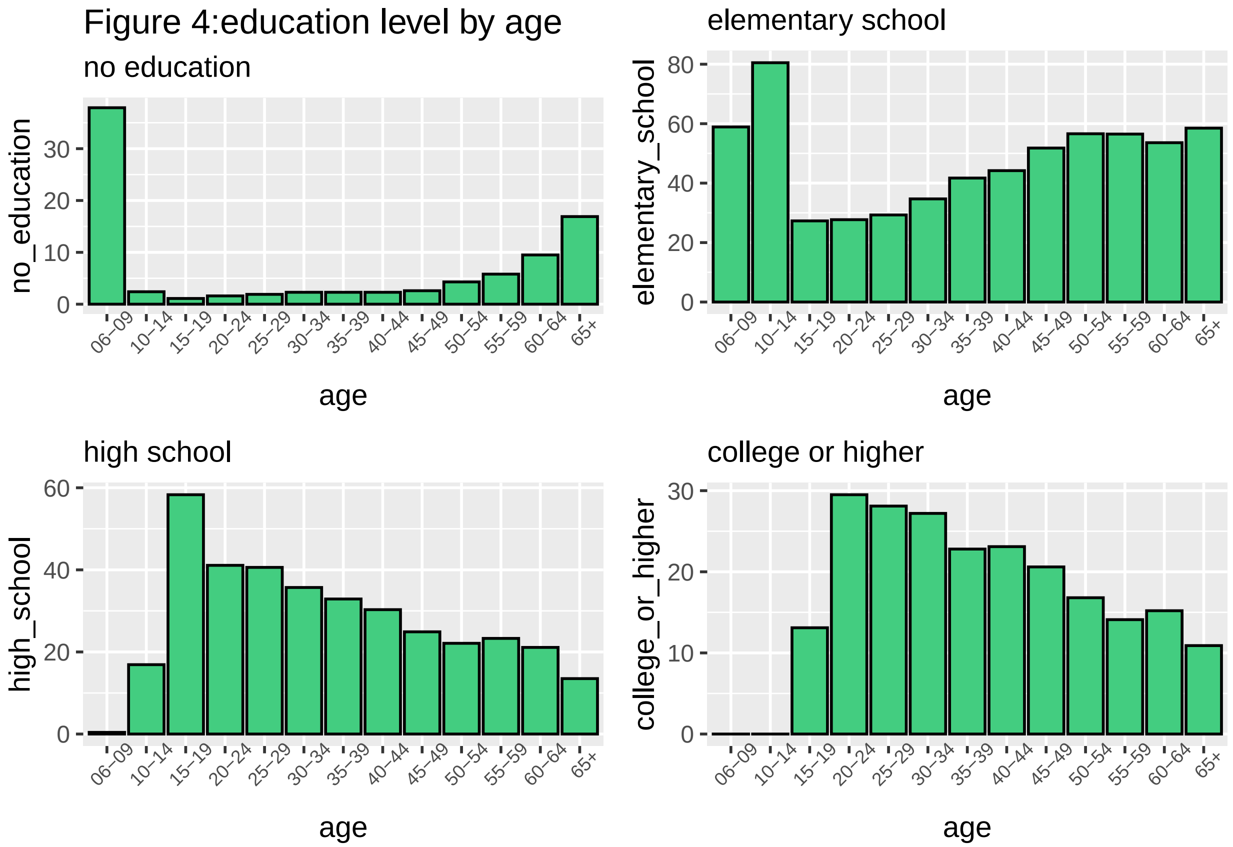


Figure 4:education level by age

no education

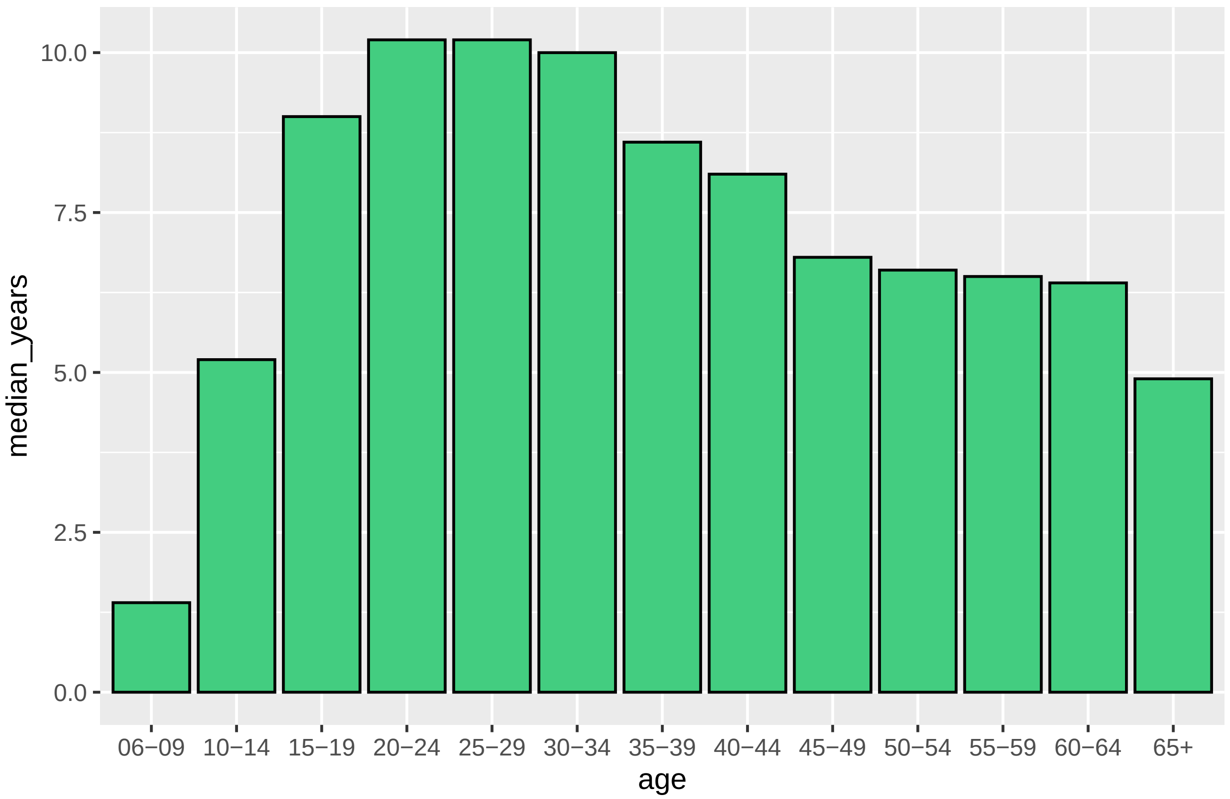
elementary school

high school

college or higher

Figure 4 exhibits the distributions of the education level of males in different age groups. From the top-left (no education) plot, age group of 6 to 9 years old has the highest percentage, which makes sense as some of them are not old enough to attend elementary school. But, besides 6-9 age group, the greater the age is, the higher the none-education rate is. From the top-right plot, age group 10-14 has the highest percentage of having elementary school as their highest degree, because this age group is right round the age for elementary school. But, besides 10-14 age group, the rate of having elementary school as highest degree goes up as the age goes up. From the bottom-left plot, besides age groups under the age of 19, the greater the age is, the less the percentage of having a high school degree is. From the plot for college or higher, for age groups more than 19 years old,the distribution is similar to the high school one. The rate of having a college degree or higher is decreasing as age goes up.

Figure 5:distribution of median years of education



In figure 5, besides age groups under 19, the less the age is, the longer the education they have received. In other words, younger male generation has a longer education time than the elder generation in Philippines in 1993.

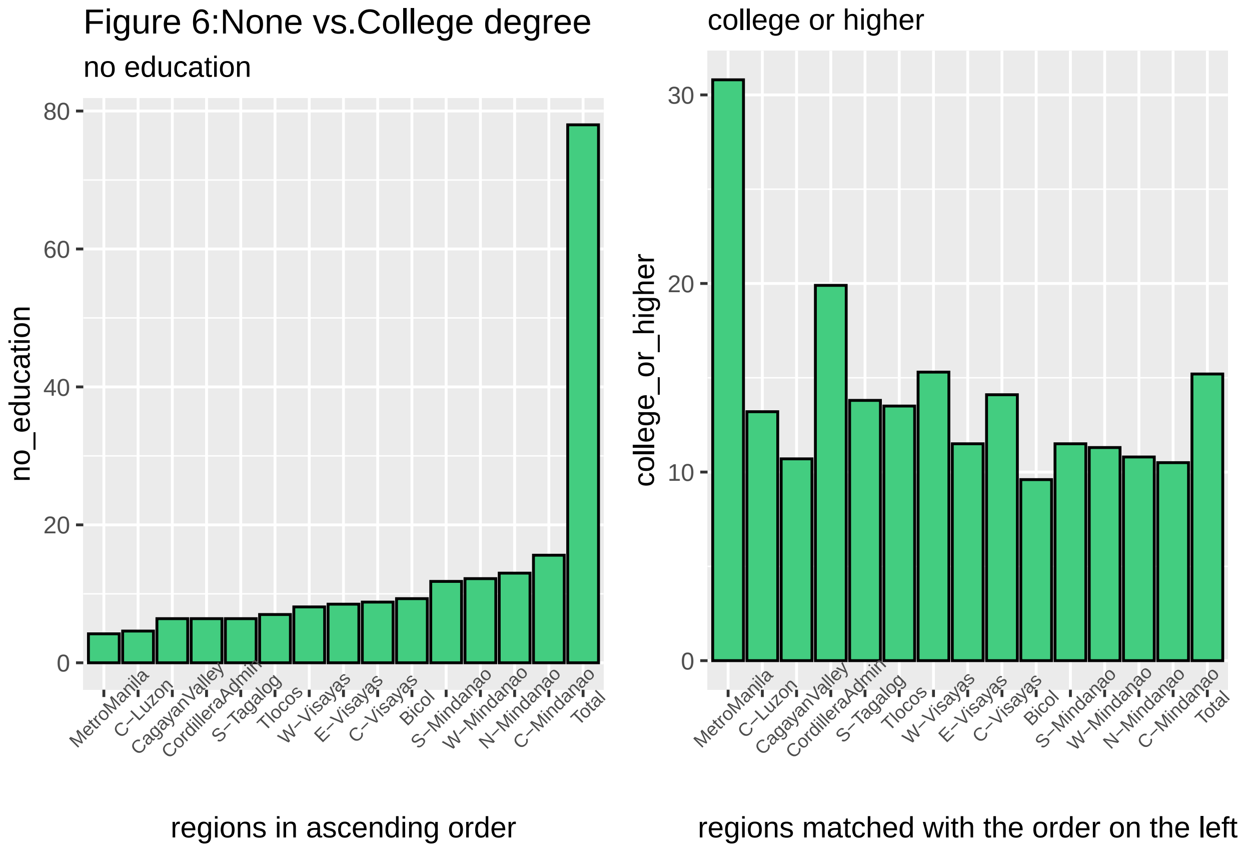


Figure 6 compares the None-education rate with the rate of having a college degree or higher in the 14 regions. The regions on x-axis of both plot are matched. The plot on the left has a increasing trend, and the plot on the right has a moderately decreasing trend. In general, the region that has a higher rate of having college degree or higher would have a lower rate of none-education.

For different types of residence, 56.9% percent of the rural male residents have a highest degree of elementary school. Only 8.1% of the rural residents have a college degree or higher, whereas the percentage for urban residents is 22%, which is more than two times of the percentage in rural area.

4 Discussion

In the process of writing this report, I transferred data in the form of image to text, and then constructed usable datasets from the text. Then I corrected the misidentified information in the datasets and cleaned the datasets for further analysis. Moreover, I described the method of collecting data from the original NDS report((NSO) and (MI) 1994). And some of the weaknesses of the method were discussed in the data section. In addition, I ran a brief analysis of the three datasets, one for the education level in different age groups, one for the education level in different residential types, and the other one for the education level in different regions. After showing a sample of the three datasets, I went through the details of each variables, and the distribution of the number of males in different age groups, residential types, and regions. For the results section, a further step was taken. With the help of plots, I found out the relationship between age groups and their highest education degree, the relationship between none-education rate and the rate of having a college degree or higher, and some differences between urban and rural residents.

One of the major findings of this report is the relationship between education level and age in Philippines in 1993. For elementary education levels(no education,elementary school), the younger male adults are less likely to have a highest degree of “elementary education levels”.For higher education levels(high school, college and higher), the older male adults are less likely to have a highest degree of “higher education levels”. In addition, by comparing the median number of years of having education in different age groups, younger adults have a relatively longer time of receiving education. In simple words, the male adults who are younger have received longer education, and are more likely to have a college degree or higher.

Another major findings of this report is the relationship between the rate of no education and the rate of having college degree or higher in different regions. By putting the regions in the same order, the relationship can be observed from the plots, one plot exhibits the no education rate in the regions, and another one exhibits the rate of having college degree or higher in the regions. After the comparative method, I found that the regions with a higher rate of college degree have a lower rate of none-education. In the region of Metro,Manila, the difference between the two rates is extreme, a fairly low non-education rate with a extremely high college degree rate. Moreover, the difference between education level in urban and rural area is extreme as well. The rate of college degree for urban residents is more than two times of the rate for rural residents. And 56.9% of the rural residents have a elementary degree only.

4.1 Weaknesses and next steps

As mentioned, the datasets used in the report is transferred from image and the method used to do the transferring is called Optical Character Recognition(OCR). In the process of OCR, a lot of the characters in the image are misidentified, and I had to correct them based on the image, which potentially decreases the accuracy of the datasets. Also, the original data are from 14 regions of Philippines, and one region was not included, because it just formed in 1993. So the data might not be representative of the whole male population in Philippines. In the future, I will try to find better methods to transfer image into text, and I will try to use more recent data for similar studies. For the next steps, this report can be used to do comparative study with other countries. In addition, this report is based on the data from 1993, so there are potentials to do similar reports for other years.

Appendix

Extract of the questions from Gebru et al. (2021)

The datasheets is written using R programming language and a R Markdown file (R Core Team 2020).

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 the education level of male in Philippines in 1993. The dataset was generated from a table in a demographic report of Philippines in 1993 ((NSO) and (MI) 1994).
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?* -The original table is created by National Statistics Office in Philippines and Macro International Inc. in USA.

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.*
    - Macro International Inc. provided the funding for the project that includes the original table ((NSO) and (MI) 1994).

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 is regarding the education level of male Filipinos in 1993, by age groups, region, and residence.
2. *How many instances are there in total (of each type, if appropriate)?*
    - There are mainly two instances, the male Filipinos and Regions in Philippines.
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 contain a sample of the whole population of Philippines. It is a representative of the larger set, in terms of age groups, and geographic coverage. The number of samples for each region and age group is planned according to the demographic distribution.
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.*
    - The data is just regarding the education level of male Filippinos and the level in different age groups and regions.
5. *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 generated from a table in NDS of Philippines in 1993. The report is published and will exist and remain constant.
6. *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
7. *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

8. *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 is regarding the education level of male in Philippines in 1993.
9. *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.*
    - Not possible

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 was collected by doing interviews and survey with selected individuals ((NSO) and (MI) 1994).
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 surveys and interviews.
3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?*
    - Probabilistic and proportion to the size of the regions ((NSO) and (MI) 1994).
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)?*
    - Field staff were trained in Baguio City, Manila, Cebu City and Davao City, some of them were interviewers and some are field editors ((NSO) and (MI) 1994).
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 was collected in 1993.
6. *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 the NDS of Philippines in 1993 report.

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.*
    - Some of the text are misidentified in the process of OCR, and I had to correct them mannually.
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.
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 programming language was used to preprocess the data (R Core Team 2020).

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*
    - The dataset is used in the analysis of male Filipinos education level.
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.*

- National Demographic Survey of Philippines in 1993 ((NSO) and (MI) 1994).

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.*
   - The dataset will be opensource on GitHub.

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

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*
   - Zihao Liu
2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*
   - zihaohans.liu@mail.utoronto.ca

The datasheet in the form of a separate pdf can be found in inputs/datasheet

References

Auguie, Baptiste. 2017. *gridExtra: Miscellaneous Functions for "Grid" Graphics*. https://CRAN.R-project.org/package=gridExtra.

Firke, Sam. 2021. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://CRAN.R-project.org/package=janitor.

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.

Henry, Lionel, and Hadley Wickham. 2020. *Purrr: Functional Programming Tools*. https://CRAN.R-project.org/package=purrr.

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.

Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

(NSO), National Statistics Office, and Macro International Inc. (MI). 1994. "National Demographic Survey of Philippines 1993."

Ooms, Jeroen. 2022. *Tesseract: Open Source OCR Engine*. https://CRAN.R-project.org/package=tesseract.

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. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.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.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2022. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC. http://www.crcpress.com/product/isbn/9781466561595.