Individual Assignment Report

Mathilda Ahr

January 16, 2020

Contents

1 Introduction

This report aims to answer the question: what was the total number of fatalities in Indonesia in year 2004 ? This will be done by using a the most aggregated data (Croicu Sundberg, 2015) from The Uppsala Conflict Data Program. This data set is in the json format and contains data for all years since 1995 for almost every country. Before conducting any coding and analysis, the data set will be explored through the shell.

After having filtered the file for Indonesia and the year 2004, the first big step will be to make it into a csv file with tidy data that would be much easier to work with in python and R. This step will be done through python. The csv file should have each line representing the information for one observation of violent event and each column representing one characteristic or variable about the events. When creating the csv file, only the relevant variables for the research will be kept: id, year, conflict_name, country, date_start. This tidy data set will then be opened in R which will serve to count the total number fatalities and make a plot displaying the fatalities for different conflicts.

2 Methods

2.1 Exploring data

This sections displays the first steps of the exploration of the data conducted through the shell.

The data set is in the json format and is a list of dictionaries where each of them is indicated by curly brackets within square brackets. There are 377 lines in the sample data (bash: 'wc UCDP_conflict_sample.json'). Reading the first few lines enabled me to find a few characteristics of the first event:

- "country": events took place in Turkey
- "best": the best estimate number of fatalities is 33
- "type_of_violence": type of violence 1 which is state-based conflict
- "year": took place in 1997

2.2 Choosing country and year

When the format of the data set was clear, I used the command 'grep' and 'wc' in order to pick a country and a year with a sufficient data quantity. The manual helped checking what the different numbers given by 'wc' represented.
I first selected the occurrences of violence in Indonesia recorded in the data set and counted them using this command: grep '"country": "Indonesia"' conflict_data_full_lined.json — wc. There are 1615 cases of violence for Indonesia since 1995 in the data set. 1615 entries for one country seems enough to answer the question.
I then added the condition that the events had taken place in the year 2004 using the command: grep '"country": "Indonesia"' conflict_data_full_lined.json — grep '"year": 2004' — wc. There are 315 entries for Indonesia for the year 2004. 315 cases of violence in a single year for the country Indonesia is enough to answer the research question.

2.3 Data pre-processing

The first step in the data pre-processing was to select the country Indonesia and the year 2004 through the shell and to redirect this sub-part of the file to another smaller json file. This new file therefore contains 315 dictionaries that represent the cases of violence for Indonesia in 2004.

The second step to be taken was to convert the smaller json file into a csv file that could be read and be interpreted by R to ultimately make a plot. Loading and reading the json file was done through python. Because it was best to keep this file intact, I created a csv file that would become the csv file to conduct the data analysis. In the python code I made it type the name of the columns that I wanted: id, year, conflict_name, country, date_start, best as the first line. Then, I created a fore-loop that would go through each dictionary and print as

a string the variables listed above. I also had to add "backslash n" symbol in order for the values of each dictionary to be printed on separate lines. These steps resulted into a tidy csv data set with 315 entries and 6 variables that was ready to be directed to R to be analysed.

2.4 Data analysis

In a R notebook in RStudio I first downloaded the tidyverse, read the csv file and put it into a variable. I then created a new variable in which I summarized the data set by summing all fatalities in one entry and calculating the average fatalities per day in the second entry. I then created a table with the number of fatalities for each different conflicts (with conflict name). Lastly, I created a bar graph that displayed the number of fatalities for each conflict.

Here is the link to my GitHub repository: https://github.com/TillyAhr/Individual-Assignment-GITHUB.git

3 Results

Total fatalities	Average fatalities per day
925	2,5342

Table 1: Total and average number of fatalities

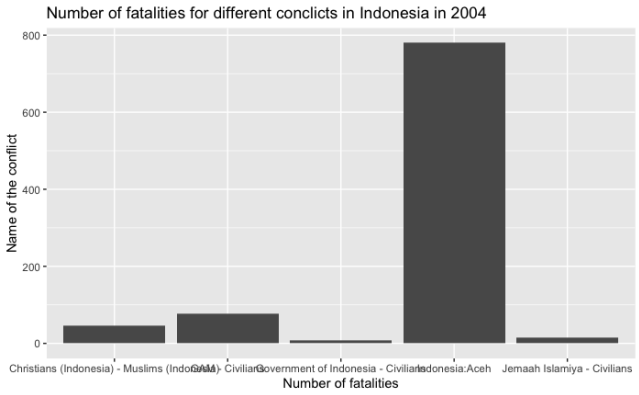

Figure 1: Figure 1

There has been 925 fatalities in Indonesia in the year 2004. The conflict that caused the majority of the fatalities was the Aceh conflict.

4 Bibliography

Croicu, M., Sundberg, R. (2015). UCDP georeferenced event dataset codebook version 4.0. Journal of Peace Research, 50(4), 523-532.