Individual Assignment: The Complete Journey

Lucie Kattenbroek & Teun van Gils

17th January 2019

With the advent of computers it has become possible to do quantitative analysis on topics that traditionally are only discussed from a qualitative perspective. These quantitative analyses can answer questions that could not be answered before, such as *Is there a visible trend in the decrease of violence?* or *Can we quantify the connection between political freedom and refugee return?*

For this assignment you will be doing such a quantitative analysis where you will have to use all of the tools you have learnt about these two weeks. You will be deciding your own research question, exploring the data, coming up with a solution and reporting your results.

1 The data

1.1 Uppsala Data Conflict Program

Conflict studies is one of those fields where a wide range of possibilities has opened thanks to computing power. The Uppsala Conflict Data Program (check: http://ucdp.uu.se/) has aimed to collect conflict data since the 1980s, and is considered to be one of the most reputable data collection programmes for conflict data. It is widely used in academic research, and has been prepared for that use, both in its definitions and in its reliability. The UCDP offers a wide variety of data sets, all ordered for different purposes. Their data sets can be used by anyone, as long as they are cited.

Today we are using their most aggregated data set (Croicu & Sundberg, 2015): every entry represents one violent incident. It is immensely informative: for every incident, there are over 40 properties given. The downside to such luxuriousness is that the file is big: over 100 MB. The data set has been updated until 2015. For more information about what is in the data, check the codebook belonging to the data set that is included with your data.

We accessed the data through their REST API (which you might learn about later this course) which means the data is in JSON format. You can download that data from the folder with data sets on drive (this is clickable text to that drive).

1.2 Population data

The UN keeps records of social statistics ('World Urbanization Prospects: The 2014 Revision', 2014). The department of Economic and Social Affairs allows you to load relevant data from their website: https://esa.un.org/unpd/wup/DataQuery/.

We requested global population data for the past few decades. The data aren't in a great format, and there are some issues. We've already adapted the records a bit but if you want to work with these data you will run into some issues.

These data are only necessary if you reach the bonus items.

1.3 Citations

Make sure you cite the data sets in your assignment. The citations in BibLaTeX style are included below.

```
@article{croicu2015ucdp,
  title={UCDP GED Codebook version 2.0},
```

```
  author={Croicu, M and Sundberg, R},
  journal={Department of Peace and Conflict Research, Uppsala University},
  year={2015}
}

@online{un2014pop,
    title={World Urbanization Prospects: The 2014 Revision},
    subtitle={custom data acquired via website},
    organisation = {United Nations, Department of Economic and Social Affairs,
    Population Division},
    year = {2014}
}
```

2 The assignment

For this assignment, you will have to pick a research question, answer it using the tools you have used the past week and a half, and write a report on it in LaTeX.

Your grade will consist of three parts: a part on your process (introduction and methods), a part on your code (methods and whatever is in your GIT repository) and a part on your representation (your results and discussion). All three constitute a significant part of your grade. The following section discusses the essential elements of the report.

2.1 Final report

Your final result needs to include:

- An introduction stating your research question, your hypothesis and how you plan to answer it using this data set (in non-technical terms).
- A methods section describing:
  - Your exploration process (e.g. manual inspection, CLI), including substantiation for why you tried certain things, what you expected the outcome to be, and what the actual outcome means and how it influenced your project.
  - Your data (pre-)processing (e.g. Python, R), including a substantiation for why you performed which processing steps.
  - Your analysis steps (e.g. R), including a substantiation for why you performed which analyses and/or why you wanted to plot certain maps or graphs.
  - *In all of the above, also include failed attempts and dead ends – the fact that something did not work out does not mean that the work does not count.*
  - Include a link to a GIT repository (on either GitHub or Bitbucket) with all your relevant Python and R code. Other code and/or information is optional.
- A results and discussion section, containing:
  - At least one plot or map, including captions.
  - At least one table with data (can be very simple; see the `xtable` package for LaTeX).
  - A conclusion paragraph where your research question is answered (or explicitly left open, if your results are inconclusive).
- Proper sourcing of your data sets.

2.2 How to go about this

We advise you to start by coming up with a research question and writing a draft for your introduction. Do not spend too much time on this, but do make sure you have properly thought about what you want to do before you start. Note that a big part of the grade is *not*, in fact, about your code, so hastily charging into battle with Python and R may not be the best strategy. For example, make sure to think about:

- How can you explore the data? Perhaps try some things out before definitively settling on a question.
- What question do you want to answer? *Is the question simple enough?*
- What is your general approach?
- Which language will you use for which task?

Use the command line and other provided information to guide your understanding of the data and design your question. Browse through the file with `less`, `grep` some things. Make sure that what you are interested in, is actually in the data set. Browse the code book.

After coming up with a plan and drafting an introduction, you will get to work with GIT, Python and R. Make sure to document your process for when you have to write it out later, possibly already in LaTeX. We advise that you try to finish your main process and analysis by the end of the first day, and spend Friday morning writing your report and, where necessary, polishing your plots. Work in small, incremental steps to avoid not having enough time left to write a proper report: you can always extend your project, but it's much harder to write a report about a half-finished project.

2.3 Teachers

Because you are delving into a project yourself that we have not completely prepared for you, you are bound to run into problems that you haven't seen before. This is great, and one of the beautiful parts of programming: pick your project, you'll learn along the way. If you run into any issues, we will happily brainstorm along. If you aren't sure if something is possible, do ask, we'll delve into it with you. Also make sure to ask questions when you are stuck on a problem; trying things out yourself is useful for about 10-15 minutes, but there is not really much room in this course for you to be stuck on a problem for a whole afternoon (which we can guarantee is a common occurrence in programming).

2.4 Project scope

First of all, two tips:

- The main data set is a .json file. You probably want to use Python to write it into a .csv format before attempting to open it in R. (We will help you with the specifics of writing your data to .csv)
- The main data set is a huge file, and your computers will probably not like loading it and working with it. Before you attempt to make it into a .csv, you may want to reduce the number of columns your final data set will have.

For the project, we suggest you to start by looking at differences between two countries or two years: this should provide a relatively simple starting point. It might well be that this task will turn out to take up all the time provided for this assignment, and this is not a problem: **we value quality over quantity**, and a well-answered simple question will obtain a higher grade than a sloppily executed complex question. However, if you do manage to answer this question to your satisfaction within the time-frame, there are some bonus goals you could consider:

- Correct for population size using the provided population data set.
- Compare both two (or more) countries and two (or more) years.

- Compare all available years for one (or more) countries, and try to perform a simple regression / fitting a simple trend-line.
- Add both a map and a plot, or simply more relevant plots.
- Include some (simple) statistical measures (note that you will not be graded on the statistical methods themselves, only on the related code, process and representation).

3 Grading

For the individual assignment, the focus will switch away from individual tools (GIT, Python, R, CLI, LaTeX) and move to how to deal with the data itself. We will still look at your code, and will take it into account when grading, but we are also very interested in whether you are able to use this code to effectively solve problems and answer questions involving large amounts of data. You have quite some freedom in what exactly you want to do and how: use this freedom wisely. You will receive a standard letter grade for this assignment. You are expected to hand in a single PDF file containing your report, conforming to all the requirements set out in subsection 2.1. This includes a link to your GIT repository, which should be contained *within* this PDF document.

References

Croicu, M. & Sundberg, R. (2015). Ucdp ged codebook version 2.0. *Department of Peace and Conflict Research, Uppsala University*.

World Urbanization Prospects: The 2014 Revision. (2014). Custom data acquired via website. (2014).