Assignment 1: Using ggplot2 for visualization

Scenario

Imagine you are the data scientist at a respected media outlet – say the "New York Times". Your editor-in-chief wants to write a feature on the government practices in registering populations. He points you to a dataset from the WorldBank and asks you for a memo and some data visualization in which you outline the main patterns around which to base the story. Since there is no way that **all** features of the data can be represented in such a memo, feel free to pick and choose some patterns that would make for a good story – outlining important patterns and presenting them in a visually pleasing way. The full background and text of the story will be researched by a writer of the magazine – your input should be based on the data and some common sense (i.e. no need to read up on this).

Memo Task

Provide a short memo (700 – 1000 words) in which you outline the main patterns of:

- How governments register people? At birth (civil registration), later in life (IDs), through occoasional censuses etc. How these patterns are related to features of the state/government, its location, its developmental status etc.? Which governments have been successful at these registration efforts?
- You can focus on the current state of affairs, or include some information about when these efforts were introduced (e.g. variable `CR Yr`, `NID Yr`).
- You can choose to zoom in to a few (or only one) of the subcategories the WorldBank uses (columns H–BJ in the 'ID4D' sheet): Civil Registration, Identification, e-ID, e-Passport, Legal.
- There is lots of information from other organizations included in the dataset that you can (but do not need to) rely on for your memo and graphs: e.g. UN population (BK-CF), UNICEF birth registration (CG-CM), elections (CV-CY), poverty (DY-EE), Democracy and Civil Liberty (GF-GH), etc.
- Provide 5-8 polished plots that are refined enough to include in the magazine with very little further manipulation (already include variable descriptions, titles, source, right color etc.) and are understandable to the average reader of the "New York Times".
- The design does not need to be NYTimes-like. Just be consistent.

Project Plan

Separate from the memo and the polished graphs, please include a brief description of the steps you took – including the data wrangling, some exploratory plots, perhaps even some pictures of some handwritten ideas of how you thought about your visualization if you like. This is for the graders to understand your thought process in getting to and including some plots and not others. This is also a good way to document a project – something we will use for the final project.

The Data

The data for the assignment comes from the Worldbank group "ID for Develepment" (ID4D). It is a global data set representing the institutional arrangements, practices, and systems for civil registration and identification, e-ID, and e-Passport in 198 economies. The current status of government practices are measured through 12 indicators that are updated annually.

Technical Details

The data comes in a reasonably clean Excel dataset from here: http://data.worldbank.org/data-catalog/id4d-dataset. Only the sheet "ID4D" should be of interest to you (and perhaps the metadata describing the variables in detail).

Part of the your task will be transforming the dataset into a shape that allows you to plot what you want in ggplot2. You will necessarily need to be selective in what to include and what to leave out.

Make sure to use at least three different types of graphs, e.g. line graphs, scatter, histograms, bar chats, dot plots, etc.

Assessment

The assessment will be largely based on the output you provide in the written memo (25%) and the associated polished plots (60%). We are looking for a cohesive whole of a well-written short memo and some plots that support the written outline. The small remaining part of the grade will be a description of the process, the associated exploratory graphs and the code (please include) you used to get the results (15%).

Submission

You should submit a single file with the following title: "Assignment1_Your_Name". The file should contain the following main sections:

1. Memo + polished graphs
2. Project book – a description of the process, including the code

We encourage submission via GitHub. If you choose that route, please go to this link: https://classroom.github.com/assignment-invitations/7c16492918b7370151744a086d552b5e

It will create a private repository in which you (and the teaching team) have admin rights. Please add your submission here. When you are fully done, please do a pull request with the title "Final submission". That way we know you are done. If you did include any images etc., make sure to upload them as well, so that a .md file renders

If you choose not to submit in Github, please submit a PDF file to courseworks.columbia.edu

Due Date

The assignment is due on Tuesday, February 14 at 6pm.

Please stay honest!

As far as we can tell, we have found little existing visual analysis of the this data yet (part of why we chose it). If you do come across something, please no wholesale copying of other ideas. We aretrying to evaluate your abilities in using ggplot2 and data visualization not the ability to do internet searches.