T412006: Applied Methods

Alice Iannantuoni*

31 October 2022

Assignment #2

This is Assignment #2 for T412006: Applied Methods in Fall 2022 at the University of Geneva. You are asked to produce an integrated report document in RMarkdown to answer a set of questions, based on `R` code previously used in class as well as resources linked on the course website, and particularly the *R for Data Science* (henceforth, R4DS) online textbook.

Instructions

- Submissions are due at **10:00 am Geneva time on Friday, November 18, 2022**
- Submit your answers in either a `.pdf` or `.html` document created with RMarkdown
- Submit the `.Rmd` file that produces your report
- For both documents, follow the file naming pattern: `T412006-12345678-assignment2.Rmd`, where you substitute `12345678` with your UNIGE number and `.Rmd` with the relevant file extension
- Use your UNIGE number as the `author` of your report; your name should **not** appear in your file names or within your files in order to facilitate anonymous grading
- Unless otherwise specified, each question must be answered with one or more lines of `R` code
- It is up to you whether you want your code to be visible in your final report document (`.pdf` or `.html`), or whether to leave it hidden in a code chunk within your `.Rmd` document; however, one must be able to read your report document and find answers to all of the questions
- You can work with classmates, but **you are expected to write your code on your own**

Grading

- This assignment is worth up to **120 points** (20% of the final class grade)
  - **25 points** for turning in an `.Rmd` file and the report document that it generates (in `.pdf` or `.html`)
    - Every file should follow the naming pattern: `T412006-12345678-assignment2.Rmd`, where you substitute `12345678` with your UNIGE number and `.Rmd` with the relevant file extension
    - Should you find yourself entirely unable to compile your answers into an RMarkdown report, you may turn in an `.R` script instead—but you will lose out on these first 25 points[1]
  - **25 points** if your `.Rmd` file compiles without errors into the same report as what you submit
  - **70 points** points for Questions 1–25
  - Up to **5 extra credit points** for the bonus tasks
  - Up to **3 extra credit points** for following the style guidelines below
  - Up to **2 extra credit points** kindly telling your instructor about any typos or errors :)
- Late submissions will cost you some of your **Punctuality Points** (PPs):
  - You start the semester with 30 PPs (5% of the final class grade)
  - Late submissions cost 2 PPs for each 24 hours of delay (late work submitted before 10:00 am on Saturday, November 19, costs 2 PPs; and before 10:00 am on Monday, November 21, costs 6 PPs)

*Postdoctoral Researcher, Political Science and International Relations, University of Geneva (alice.iannantuoni@unige.ch)

[1] Don't give up too soon! Follow the advice in my *Getting Started with RMarkdown* tutorial, available on the course website as `2022-10-21-T412006-rmd-template.pdf`, and start with its corresponding `.Rmd` file as a template.

Style

Generally, we want to follow the `tidyverse` Style Guide. In particular, I will check for the following:

- Clearly structure your documents with headers and subheaders
- Name your objects in `R` with only lowercase letters, numbers, and underscores (`_`)
- Always put a space after a comma, but never before
- Infix operators (e.g., `==`, `+`, `-`, `<-`, `=`, `&`, etc.) should have spaces on both sides
- Pipe operators (`%>%`) should always have a space before them, and usually be followed by a new line

Strive to produce pretty plots with `ggplot2`! This includes making sure that your plots have informative captions and/or titles; clearly-labeled axes; nicely-formatted legends... For help, start with the "Graphics for Communication" chapter in R4DS. *Data Visualization* by Kieran Healy is another good resource.

Set-Up

In this assignment, you will work with data from the Quality of Government Standard Dataset.[2]

Start by downloading the `2022-10-31-T412006-assignment2-data.csv` data file from the course website and save it to the same folder where you intend to set your working directory for this assignment. Next, load your package(s); set your working directory; and upload the data into RStudio, saving it as `data_qog`.

This first code chunk in your `.Rmd` file should look something like this:

```
# load the tidyverse package from your library
library(tidyverse)

# set your working directory (your path will be different from mine!)
setwd("~/OneDrive - unige.ch/2022-teaching-applied-methods/classwork")

# upload the data
data_qog <- read_csv("2022-10-31-T412006-assignment2-data.csv")
```

Take a first look at the data. You can use command `names()` to print a vector of the variable names. These variable names probably do not mean much to you yet. Let's change that!

Familiarize yourself with the source of these data: the Quality of Government Institute and, more specifically, their Quality of Government (henceforth, QoG) Standard Dataset. The key resource for you to understand the variables in this dataset will be their official codebook, which you can find online at https://www.qogdata.pol.gu.se/data/codebook_std_jan22.pdf as well as on our course website as `qog_codebook_std_jan22.pdf`.

In particular, I recommend searching for each of the variable names in the codebook `.pdf` document. You will find information on each variable, including a brief description, the primary source it comes from, and the number of years and countries that it is available for.

After browsing both the data and the codebook, you are ready to start answering questions:

- **Q1**: how many observations are in the tibble `data_qog`? And how many variables?
- **Q2**: print a vector of the variable names in `data_qog`.
- **Q3 (no `R` code required to answer)**: produce a table[3] listing each variable name in `data_qog` and its corresponding short description from the QoG codebook. If you wish, you can add one or more additional columns to your table with any other piece of information about the variables that you find relevant and want to have easy access to while you work on your assignment. See an incomplete example on the following page.

[2] Teorell, Jan, Aksel Sundström, Sören Holmberg, Bo Rothstein, Natalia Alvarado Pachon & Cem Mert Dalli. 2022. The Quality of Government Standard Dataset, version jan22. University of Gothenburg: The Quality of Government Institute, https://www.gu.se/en/quality-government, doi:10.18157/qogstdjan22

[3] Remember that you can use an online table generator like this one to help with RMarkdown table syntax.

Table 1: Codebook for data_qog

Variable Name	Short Description
ccode	Country Code
cname	Country Name
year	Year
ht_region	The Region of the Country
wdi_area	Land area (sq. km)

- **Q4**: What temporal range does the data_qog dataset cover?
- **Q5**: What countries are included in the data_qog dataset? List them alphabetically.

Data Transformations

Look up the variable ht_region in the QoG codebook to make sure you understand what its levels refer to.

- **Q6**: What type of variable is ht_region stored as? Does this make sense to you, or should you change it to a different type?
- **Q7**: Create and store a tibble called data_cntr_reg that includes (i) only observations from 2015 and (ii) only the variables cname and ht_region.
- **Q8**: Produce a table that neatly displays the data_cntr_reg tibble: a two-column table with country names and their corresponding region. List the countries by region and alphabetically within region.[4]
- **Q9**: What countries are categorized in level 5 of ht_region? List them alphabetically.
- **Q10**: Create a variable my_region with an alternative regional classification, as follows:

  1. Americas (including the Caribbean)
  2. Asia
  3. Europe
  4. North Africa & the Middle East (including Israel, Turkey & Cyprus)
  5. Sub-Saharan Africa
  6. The Pacific (including Australia & New Zealand)

  Note that my_region levels 4 and 5 have exact correspondents in ht_region levels, while the others will require you to combine or redistribute countries into different groupings. Make sure this new variable my_region is added to your data_cntr_reg tibble.
- **Q11**: Add your alternative regional classification my_region to the original data_qog tibble.[5]
- **Q12**: The data_qog tibble includes information on *per capita* gross national income (GNI PC) in variable wdi_gnicapcon2010. According to the QoG codebook, GNI PC is calculated as the GNI divided by midyear population. Create your own estimate of GNI PC by dividing GNI (wdi_gnicon2010) by population (wdi_pop) and store it in a new variable called my_gnicapcon2010.

  Your answer should come in the form of a new tibble called data_gnipc which starts from tibble data_qog but includes only the variables cname, wdi_pop, wdi_gnicon2010, wdi_gnicapcon2010, and my_gnicapcon2010.
- **Q13**: Let's check whether the GNI PC estimate you calculated resulted in the same value as the wdi_gnicapcon2010 variable! Create a new variable in your data_gnipc tibble called check. This

[4] Hint: try using function knitr::kable() on a tibble object to help with RMarkdown table syntax.

[5] Hint: check out the "Mutating Joins" section of R4DS on how to combine variables from two datasets or tibbles. Alternatively, the merge() function from base R is also an option—see how it compares to the tidyverse mutating joins here. Remember that you want to add the my_region classification for each country from the data_cntr_reg tibble to every data_qog row with that same country name, ensuring that the resulting tibble keeps all of the original rows of data_qog.

check variable should be equal to zero if wdi_gnicapcon2010 and my_gnicapcon2010 are the same, and equal to one if they are different.[6]

- **Q14**: How many observations have the same value for wdi_gnicapcon2010 and my_gnicapcon2010? What percentage of the total number of observations is that?
- **Q15**: Open your data_gnipc tibble in the RStudio viewer and look at the check variable; sort it in both increasing and decreasing order by clicking on the arrow on the top-right of the variable name. Does it look like, when check is equal to one, your calculated value my_gnicapcon2010 is indeed different from wdi_gnicapcon2010? Why or why not?

  Can you adjust the way in which you constructed check in order to have it correctly identify instances where my_gnicapcon2010 is meaningfully different from wdi_gnicapcon2010?[7]

  How many observations have meaningfully different values for my_gnicapcon2010 versus the original wdi_gnicapcon2010? What countries and years to they pertain to? Can you venture a guess as to why we see these discrepancies?

Describing Variation

Rejoice! You have already earned over half of your question-based points by completing the Set-Up and Data Transformation sections above. Take a break, have a snack, and come back ready to describe the variation in some of the variables in these QoG data.

- **Q16**: How much do people from different countries use the internet? We can look at the variation in the wdi_internet variable to get an idea. Across the entire dataset, provide the following descriptive statistics for the wdi_internet variable:
  - Mean
  - Median
  - Range
  - Standard Deviation
  - Interquartile Range

  Can you provide an example of a country-year observation where 100% of the population is estimated to have used the internet in the previous three months? What about an example where that is 0%?
- **Q17**: How much "missingness" is in the wdi_internet variable? Your goal is to create a plot that shows what percentage of the observations in each year are NAs for the wdi_internet variable. In other words, you want a plot with year on the x-axis and the percentage of NAs in that year on the y-axis.[8]

  What year has the best coverage for the wdi_internet variable?
- **Q18**: Subset the data to only include observations for the year with the best coverage of the wdi_internet variable.[9] Create a boxplot that shows the distribution of the wdi_internet variable in that year.

  Create a second boxplot that shows the distribution of the wdi_internet variable in that year, but this time for each of the levels in your my_region[10] variable.[11]

[6] In other words, check is an indicator of whether your calculated GNI PC is the same as the original GNI PC in the data.

[7] Hint: think about decimal points; consider the round() function.

[8] Hint: One way to go about this is to start by creating an indicator variable equal to one if a country-year observation is NA for the wdi_internet variable. From that, calculate what percentage of each year's observations are NA (this might come useful). Always remember to ungroup() every time you group_by()!

[9] If you were unable to answer **Q17**, feel free to use a year of your choice, from 1990 onward.

[10] If you were unable to answer **Q10–11**, feel free to use the regional classification in ht_region instead.

[11] In other words, the first plot should be akin to the one on Page 11 of the October 21 slides; and the second plot should be akin to the one on Page 14 of the same slides. You can find the 2022-10-21-T412006-slides.pdf document on the course website. Of course, I encourage to make all of your plots much prettier than those, using ggplot2.

- **Q19**: Describe the variation in life expectancy across countries and time. Start by computing measures of central tendency (mean and median) and measures of spread (range, variance, standard deviation, and interquartile range) for the `wdi_lifexp` measure across the entire dataset. What percentage of the observations are `NA`s?
- **Q20**: Create a new tibble with one observation per year and two variables: the year, and the average `wdi_lifexp` across all countries in that year. Visualize this variation in a plot with `year` on the x-axis and the average `wdi_lifexp` across all countries in that year on the y-axis.
- **Q21**: Create a new tibble with one observation per region-year and three variables: the year, the region, and the average `wdi_lifexp` across countries in that region in that year.[12] Visualize this variation in a plot with `year` on the x-axis and the average `wdi_lifexp` on the y-axis, using colors to differentiate the various regions.
- **Q22**: How does life expectancy correlate with levels of economic development? Create a scatterplot with each country-year level of gross domestic product *per capita* (GDP PC) on the x-axis and life expectancy on the y-axis. Do the two variables appear to be correlated? If so, is it a positive or negative correlation?
- **Q23**: Replicate the plot from **Q22**, but add colors to distinguish between country-year observations from different regions.[13]
  - What region(s) do observations with both low GDP PC and low life expectancy tend to be in?
  - What region(s) do observations with both high GDP PC and high life expectancy tend to be in?
  - What region(s) do observations with lower levels of life expectancy relative to their GDP PC tend to be in?
    - Why do you think these countries under-perform on the life expectancy metric, given their level of GDP PC? If you have a hypothesis about this sub-question and want to try to get descriptive evidence for it by adding to or modifying this plot to include information from additional variables, you can do so for some **extra credit points**!
- **Q24**: The QoG data includes different ways to classify regime types and measure democracy. Two well-known examples are the democracy-dictatorship measures proposed by Cheibub, Ghandi, and Vreeland (2010),[14] which is in our data set as `br_dem`; and the Revised Combined Polity Score last updated in Marshall and Gurr (2020),[15] in our data set as `p_polity2`.
  - What type of variable is `br_dem`? Does it make sense to take its mean? What information would you give to summarize its distribution?
  - What type of variable is `p_polity2`? What is its range? What information would you give to summarize its distribution?
- **Q25**: Create a plot that visualizes how these two measures of democracy (`br_dem` and `p_polity2`) relate to one another, and describe in words what the plot allows you to see. For example you might consider a boxplot of `p_polity2` by the two levels of `br_dem`—but this is only one possibility!
- **Bonus Task**: There are many more variables in the `data_qog` tibble that you learned about in **Q3** but did not get to use over the course of this assignment. If you feel inclined to go above and beyond and earn some **extra credit points**, write out a question for yourself that would allow you to play with one or more of these "unused" variables (inspired by the style of questions in this assignment), and answer it.

[12] Again, use the `my_region` variable if you were able to create it, and `ht_region` otherwise.

[13] Again, use the `my_region` variable if you were able to create it, and `ht_region` otherwise.

[14] You can read more here and here.

[15] You can read more here and here.