STA205 Project - Exploring the GSS

Photo by Mauro Mora on Unsplash

In this project we analyze data from the GSS, which is a dataset gathered about contemporary American society in order to monitor and explain trends and constants in attitudes, behaviours, and attributes. Hundreds of trends have been tracked since 1972. In addition, since the GSS adopted questions from earlier surveys, trends can be followed for up to 70 years.

The GSS contains a standard core of demographic, behavioural, and attitudinal questions, plus topics of special interest. Among the topics covered are civil

liberties, crime and violence, intergroup tolerance, morality, national spending priorities, psychological well-being, social mobility, and stress and traumatic events.

In this project we use GSS to analyze values of population parameters of interest about US adults. We will first explore a small subset of the data `gss16` which has a portion of the dataset from 2016. [1] You will be working on this part independently; then in the second part of the project you will join a group and to expand your research scope and raise more interesting questions using the whole dataset.

Getting started

Packages

We'll use the **tidyverse** package for much of the data wrangling and visualization and the dataset `gss16` lives in the **dsbox** package. These packages are already installed for you. You can load them by running the following in your Console:

```
library(tidyverse)
library(dsbox)
```

Data

The data can be found in the **dsbox** package, and it's called `gss16`. Since the dataset is distributed with the package, we don't need to load it separately; it becomes available to us when we load the package. You can find out more about the dataset by inspecting its documentation, which you can access by running `?gss16` in the Console or using the Help menu in RStudio to search for `gss16`.

Project Part I (Individual work)

Section 1: Harassment at work

In 2016, the GSS added a new question on harassment at work. The question is phrased as the following.

> *Over the past five years, have you been harassed by your superiors or co-workers at your job, for example, have you experienced any bullying, physical or psychological abuse?*

Answers to this question are stored in the `harass5` variable in our dataset.

Question 1. What are the possible responses to this question and how many respondents chose each of these answers?

Question 2. What percent of the respondents for whom this question is applicable (i.e. excluding `NA`s and `Does not apply`s) have been harassed by their superiors or co-workers at their job.

Section 2: Time spent on email

The 2016 GSS also asked respondents how many hours and minutes they spend on email weekly. The responses to these questions are recorded in the `emailhr` and `emailmin` variables. For example, if the response is 2.5 hrs, this would be recorded as `emailhr = 2` and `emailmin = 30`.

Question 3. Create a new variable called `email` that combines these two variables to reports the number of minutes the respondents spend on email weekly.

Question 4. Visualize the distribution of this new variable. Find the mean and the median number of minutes respondents spend

on email weekly. Is the mean or the median a better measure of the typical among of time Americans spend on email weekly? Why?

Question 5. Create another new variable, `snap_insta` that is coded as "Yes" if the respondent reported using any of Snapchat (`snapchat`) or Instagram (`instagrm`), and "No" if not. If the recorded value was `NA` for both of these questions, the value in your new variable should also be `NA`.

Question 6. Calculate the percentage of Yes's for `snap_insta` among those who answered the question, i.e. excluding `NA`s.

Question 7. What are the possible responses to the question *Last week were you working full time, part time, going to school, keeping house, or what?* and how many respondents chose each of these answers? Note that this

information is stored in the `wrkstat` variable.

Question 8. Fit a model predicting `email` (number of minutes per week spent on email) from `educ` (number of years of education), `wrkstat`, and `snap_insta`. Interpret the slopes for each of these variables.

Question 9. Create a predicted values vs. residuals plot for this model. Are there any issues with the model? If yes, describe them.

Section 3: Political views and science research

The 2016 GSS also asked respondents whether they think of themselves as liberal or conservative (`polviews`) and whether they think science research is necessary and should be supported by the federal government (`advfront`).

- The question on science research is worded as follows:

> Even if it brings no immediate benefits, scientific research that advances the frontiers of knowledge is necessary and should be supported by the federal government.

And possible responses to this question are Strongly agree, Agree, Disagree, Strongly disagree, Don't know, No answer, Not applicable.

- The question on political views is worded as follows:

  > We hear a lot of talk these days about liberals and conservatives. I'm going to show you a seven-point scale on which the political views that people might hold are arranged from extremely liberal–point 1–to extremely conservative–point 7. Where would you place yourself on this scale?

⊕

And possible responses to this question are Extremely liberal, Liberal, Slightly liberal, Moderate, Slghtly conservative, Conservative, Extrmly conservative. Responses that were originally Don't

know, No answer and Not applicable are already mapped to `NA`s upon data import.

Question 10. In a new variable, recode `advfront` such that Strongly Agree and Agree are mapped to `"Yes"`, and Disagree and Strongly disagree are mapped to `"No"`. The remaining levels can be left as is. Don't overwrite the existing `advfront`, instead pick a different, informative name for your new variable.

Question 11. In a new variable, recode `polviews` such that Extremely liberal, Liberal, and Slightly liberal, are mapped to `"Liberal"`, and Slghtly conservative, Conservative, and Extrmly conservative disagree are mapped to `"Conservative"`. The remaining levels can be left as is. Make sure that the levels are in a reasonable order. Don't overwrite the existing `polviews`, instead pick a different, informative name for your new variable.

Question 12. Create a visualization that displays the relationship between these two new variables and interpret it.

Grading rubric (Part I)

In this part I of the project you are required to write a brief yet complete report about your questions and findings. You should explore at least one extra question of your own in addition to the 12 questions listed above, which must also be included in the report. Your report should include an introduction, questions and findings, as well as conclusion. The presentation of the document is also part of the rubric, which include the logical flow, cleanness and tidiness of the text and graphics.

Overall grade breakdown:

- Introduction: 10
- Questions and findings: 30
- Conclusion: 10
- Presentation: 15
- Code: 35

Submission

Submit your report as

- a pdf (text + plot) with code hidden and
- another file with code shown as an html knitted from an R notebook, as well as
- the Rmd file itself (technical wise you knitr first time to produce pdf with code hidden, then second time with code shown as html). Put this chunk in a code cell at top of Rmd will hide code in knitted doc:

```
opts_chunk$set(echo=FALSE)
```

Project Part II (Group Work)

- To be released separately.