Income Has Serious Implications On Many Facets of American Life in 2021*

A Look Into The Implications of Americans' Incomes Reveals Information On Disparity of The Gender Pay and the American Dream.

Adam Labas Krupali Bhavsar

21 March 2022

Abstract

First sentence. Second sentence. Third sentence. Fourth sentence.

1 Introduction

You can and should cross-reference sections and sub-sections. For instance, Section 2. R Markdown automatically makes the sections lower case and adds a dash to spaces to generate labels, for instance, Section 5.1.

2 Data

In this Data Section 2, we will provide a look into the data acquisition and processing methodology as well as a deep dive into the contents of the data. First of, (Table **??**) give us a glimpse of the data.

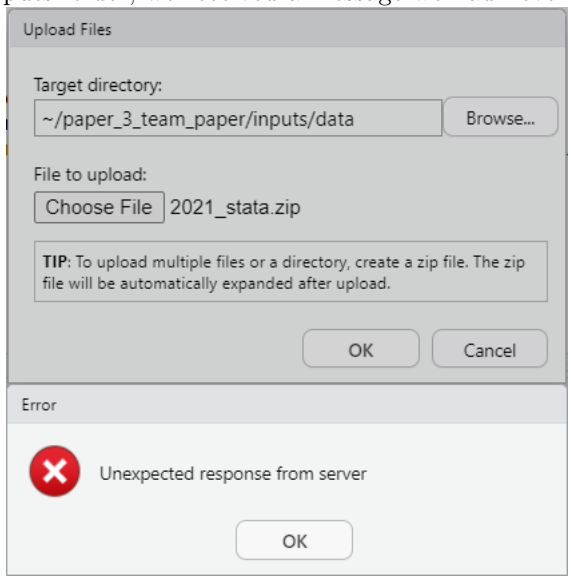
2.1 Data Collection

All of the data used in this analysis was directly collected from the United States 2021 General Social Survey, available here: https://gss.norc.org/ (link and additional information available in the README.md file.) We downloaded the desired data in the STATA format and loaded it into R (R Core Team 2020) for manipulation. The zip file acquired included 4 files: a read me text file, the raw data as a dta file, as well as two pdf files. *GSS 2021 Codebook R1b* is essentially the instruction manual for the data set. It goes over many topics like the objectives and intents of the survey, a brief overview of the methodology and an in depth explanation of the 565 variables in the raw data. With 565 variables collected in the survey, the data file titles *gss2021.dta* has a size of 2.6 MB. Although this isn't too large of a data file, we did encounter some problems which may have implications as it pertains to reproducibility of the workflow. The following is an explanation of the problems encountered as well as an explanation of the workaround/solution we implemented to get over the obstacle.

*Code and data are available at: XXXXXXXXXXXXXXXXXXXXXXX

2.2 Problem With Data Upload into R Studio

We know that having a reproducible workflow is crucial if we want to produce mature and ethical statistical papers which any data scientist down the line can reproduce for any purpose from verifying results or modifying methodology. As previously stated, the raw data provided by the United States 2021 General Social Survey was in a zip folder and in dta format. To conduct the development of this paper, we worked in RStudio {R Core Team (2020)) in a server developed, maintained an administered by one of the TAs for the STA304 Course at the University of Toronto (Winter 2022), Mauricio Vargas Sepúlveda, better known as Pachá. This server works just like any other server for RStudio like JupyterHub but the admin user, Pachá, has the ability to overview all activity on the server and access users accounts to aid with trouble shooting or data sharing: this will proved to be decisive in just a moment. When we tried to follow typical procedure to upload the 2021_stata.zip file into under data in the inputs folder, we received a message we had never received before. Figure 1 is a screen capture of that error.



This was hard for us to wrap our heads around as we were able to upload other files independently, and crucially we were able to upload the other file in the zip independently aside from the dta file. We tried to replicate this problem working in RStudio locally but surprisingly it was not a problem there. As such, we contacted the server admin, shared the zip folder with him, and using his administrator permissions, he uploaded the unzipped folder into a shared folder that the entire class has access to. We then copped this folder into the data folder under inputs. It is important to note that throughout this process, none of the information from the zip folder had been altered or manipulated and that all the information, and only the information, which had started in it, ended up in our instance of RStudio.

Lastly on this matter, should you run into this problem while attempting to reproduce our findings or work, working locally in RStudio might fix the problem.

2.3 Data Processing and Variable Selection

Upon being able to successfully load the data into R (R Core Team (2020)) we began data processing and variable selction. This includes things like cleaning the data from missing values, choosing the variables of interest and producing new variables which will help us with our analysis. The following is an overview of the methods and events which took place in this phase of the analysis.

All the data processing, cleaning and variable selection takes place in an R script called *Data Acquisition and Processing.R* in the scripts folder. To process our data, we began by loading all the necessary libraries like haven (Wickham and Miller 2021) which was used to acess the information from the dta type file and we then load the data into R (R Core Team 2020). Additionally, we used a wide variety of libraries like knitr (Xie 2021), tidyverse (Wickham et al. 2019), tidyr (Wickham 2021), janitor (Firke 2021), dplyr (Wickham

2021) as well as tinytex (Xie 2022) at some points in the process for data processing and pdf document generation.

After loading the data into our local environment and saving its as a cvs called raw.csv, we began to create as subset of the data with the desired variables. There are 565 variables in the data and we obviously can't explain each of them, so here is a description of some of the variables we chose to work with. *sexnow1* is the current sex of the respondent and is a concept more similar to gender, *age*, *born*: born in America or not, *marital*: the marital status, *educ*: the highest attained level of education, *divorce*: divroced or not, whether or not they have a *degree*, *income*: their current family income, *income16*: their family income when they were 16 years old, *nathealy*: if they think enough is done for healthcare, *cohort*: year of birth, *agekdbrn*: age when first child was born and lastly, *granborn* which is a number from 0-4 signifying the number of grandparents of the respondent which were born in the united states.

Additionally, we created a new column called *income_cat* and *income16_cat* which are new variables which offer us the liberty of using the *income* and *income16* variables but as discreet, categorical variables instead of the coded numerical system used all throughout the data set and explained in the information code book titled *GSS 2021 Codebook R1b.pdf* available in the 2021_stata in the data folder in inputs.This was done by duplicating the original columns and then re-coding the values. In the case of *income* for instance, the codebook explains that a value of 1 for income means the response was that the annual income was less than $1000. As such, income_cat would read *$1,000 AND UNDER*. Similar mutations were done for the variables *sex*, *granborn*, *nathealy* and *born*.

We finished off by saving all the changed in a new data set: reduced_data.csv

3 Methodology

In matters of reproducibility, the methodology used to collect data is always important. This includes the various sampling techniques that were used, the way that non responses are dealt with and other key decisions taken by the data collection team and the survey itself. The following is an explanation of the methodology used in the 2021 GSS, as well as its key features, strengths, and weaknesses.

3.1 Methodology Overview

The 2021 GSS is in many ways different to the previous General Social Surveys in years prior to the ascension of the corona virus to international stardom. According to the GSS 2021 Codebook R1b, in previous iterations of the GSS, the survey was taken solely via in-person interview, where the interviewer had been adequately trained and prepared for various situations that would have arose like non responds or the person being interviewed not knowing an answer. However, for reasons related to health and safety protocols and "to safeguard the health of staff and respondents during the COVID-19 pandemic, the 2021 GSS data collection used a mail-to-web methodology instead of its traditional in-person interviews." This means that after being selected via address based sampling the selected individuals "interview" would be conducted online via a self administered questionnaire.

3.2 Key features

4 Results

5 Discussion

5.1 First discussion point

If my paper were 10 pages, then should be be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this.

5.2 Second discussion point

5.3 Third discussion point

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix

A Additional details

References

Firke, Sam. 2021. *Janitor: Simple Tools for Examining and Cleaning Dirty Data.* https://github.com/sfirke/janitor.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley. 2021. *Tidyr: Tidy Messy Data.*

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, and Evan Miller. 2021. *Haven: Import and Export 'SPSS', 'Stata' and 'SAS' Files.* https://CRAN.R-project.org/package=haven.

Xie, Yihui. 2021. *Knitr: A General-Purpose Package for Dynamic Report Generation in r.* https://yihui.org/knitr/.

———. 2022. *Tinytex: Helper Functions to Install and Maintain TeX Live, and Compile LaTeX Documents.*