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

A Look Into The Implications of Americans' Incomes Reveals Information On The Veracity of American Dream.

Adam Labas Krupali Bhavsar

21 March 2022

Abstract

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

Contents

1 Introduction **2**

2 Data **2**

2.1 Data Collection 2

2.2 Problem With Data Upload into R Studio 2

2.3 Data Processing and Variable Selection 3

3 Methodology **4**

3.1 Methodology Overview 4

3.2 Key features & Pros And Cons of the Methodology and Survey 4

4 Visualizing the Data and The Implications **5**

4.1 Visual 1 5

4.2 Implications 1 6

4.3 Visual 2 7

4.4 Implications 2 8

4.5 Visuals 3 8

5 Limitations **8**

5.1 Uneven Bin width for Income Groups. 8

5.2 Why cut off income at $25,000? 8

5.3 Next steps 8

*Code and data are available at: https://github.com/adam-labas/The-US-GSS-And-The-American-Dream

Table 1: GSS Cleaned Data

Age	Gender	Born	Education	Birth Year	Grandparents	Income at 16	Current Income
20	Male	No	12	2001	Four	$90,000 TO $109,999	$25,000 OR MORE
76	Male	Yes	13	1945	Four	$60,000 TO $74,999	$25,000 OR MORE
61	Female	No	16	1960	Four	$8,000 TO $9,999	$8,000 TO $9,999
37	Male	Yes	11	1984	Four	$75,000 TO $89,999	$25,000 OR MORE
23	Male	No	15	1998	Four	$40,000 TO $49,999	$25,000 OR MORE
20	Female	Yes	14	2001	Three	$110,000 TO $129,999	$25,000 OR MORE
65	Female	No	12	1956	Four	$20,000 TO $22,499	$20,000 TO $24,999

Appendix **10**

References **11**

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 **??**.

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.

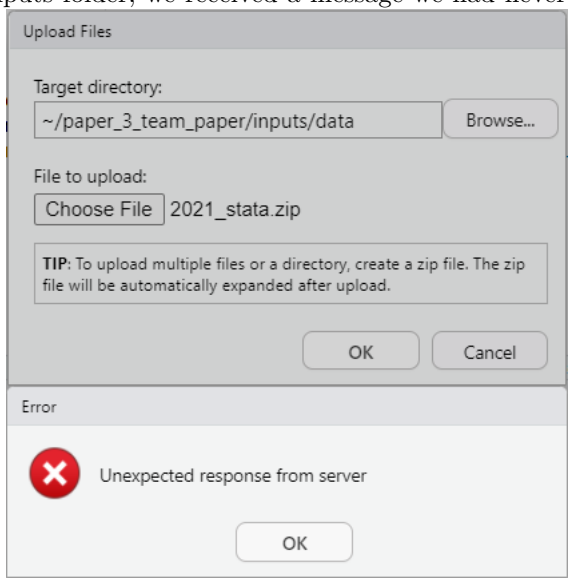
2.1 Data Collection

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 titled *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.

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), patchwork (Pedersen 2020), readr (Wickham, Hester, and Bryan 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 subsets 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 new columns called *income_cat* and *income16_cat* which are new variables which offer us the liberty of using the *income* and *income16* variables 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 code book 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 responses 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 individual's "interview" would be conducted online via a self administered questionnaire. The limitations of such will later be discussed in Section 5

3.2 Key features & Pros And Cons of the Methodology and Survey

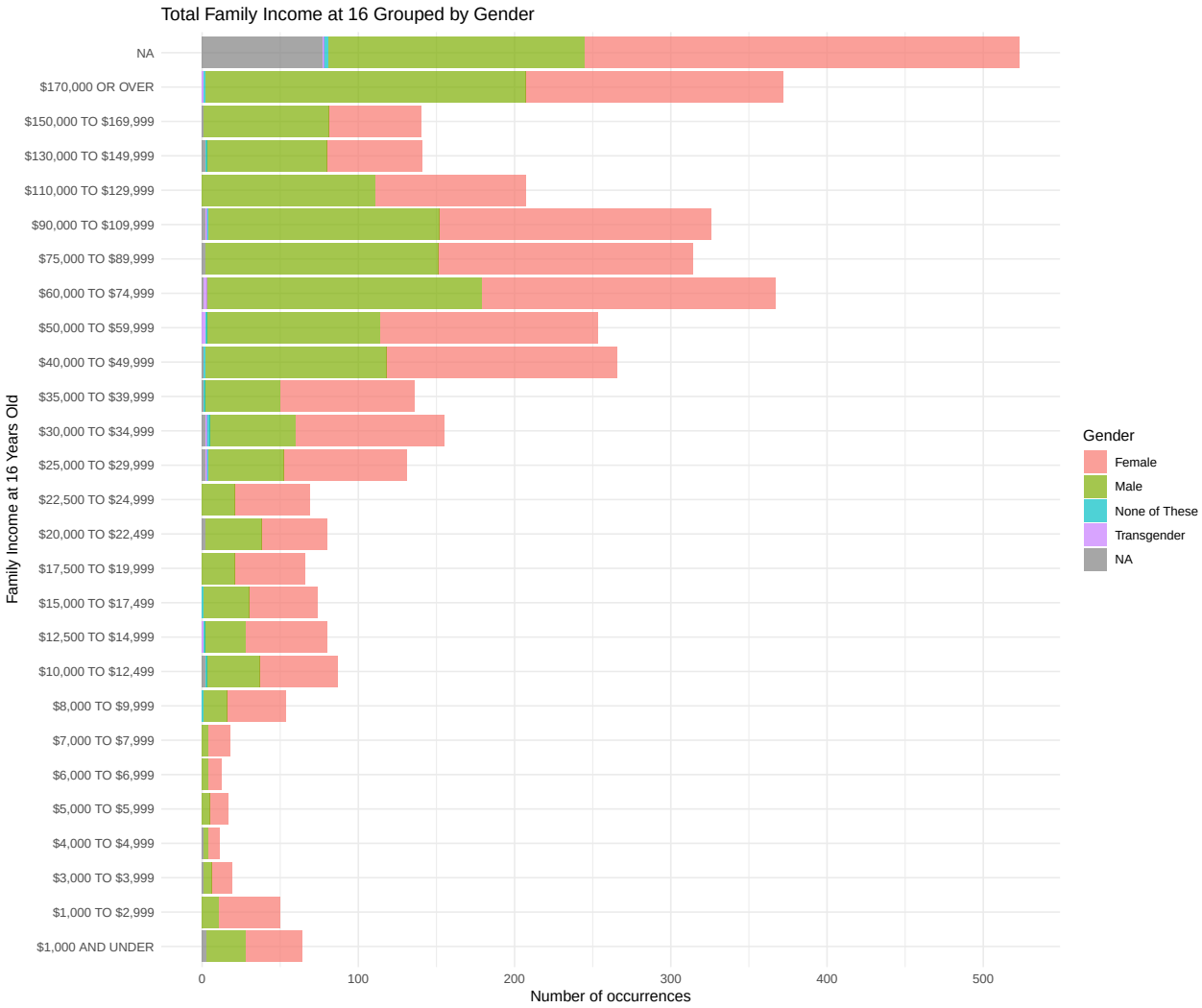
It is important to discuss the reach of our data. We know from the provided documentation that the population group eligible for taking the survey are adults 18 or older in the United States who live in non institutional housing at the time of interviewing. Also, the 2021 GSS uses the last birthday method to decide which individual in the household will take the survey. This was done by having "a professional phone interviewer team conducted phone outreach to complete screeners, answer inbound phone calls from sampled households, identify respondents, and complete interviews by phone and/or prompt them to complete the web survey." As far as the invitation method goes for data collection, materials that provide a web link to invite people to participate in the survey was mailed out and an option to conduct the interview via the phone instead of a self-administered questionnaire was also provided. Out of the 27,591 lines of sample (people who had been prompted to answer the questionnaire), 4,032 responses were collected. This has implications on the finding of our analysis in Section 5 as we will have to consider if we are able to generalize our finding from the relatively small sample that we have to the population of all US citizens over 18 years old. The small sample size and extremely long survey are both weaknesses of the survey which we will discuss in more depth in Section 5.

In addition, the 2021 GSS used census data as well as publicly available information to try and identify areas and specific household with Spanish speakers whereby the survey was then offered in English and in Spanish. This is a big strength of the survey at it shows the attempt of the data analyst to be as mature and ethical as possible and offer the survey in a language that is more suitable for a large portion of US citizens.

4 Visualizing the Data and The Implications

In this analysis, we am interested at looking at how real the American Dream is and if it is found to be possible, whether or not there is any difference in attainment amoung genders. As such, in this Section 4, we will create various visuals which will enable us to extract as much information as we can and we will speak about the implications that they make.

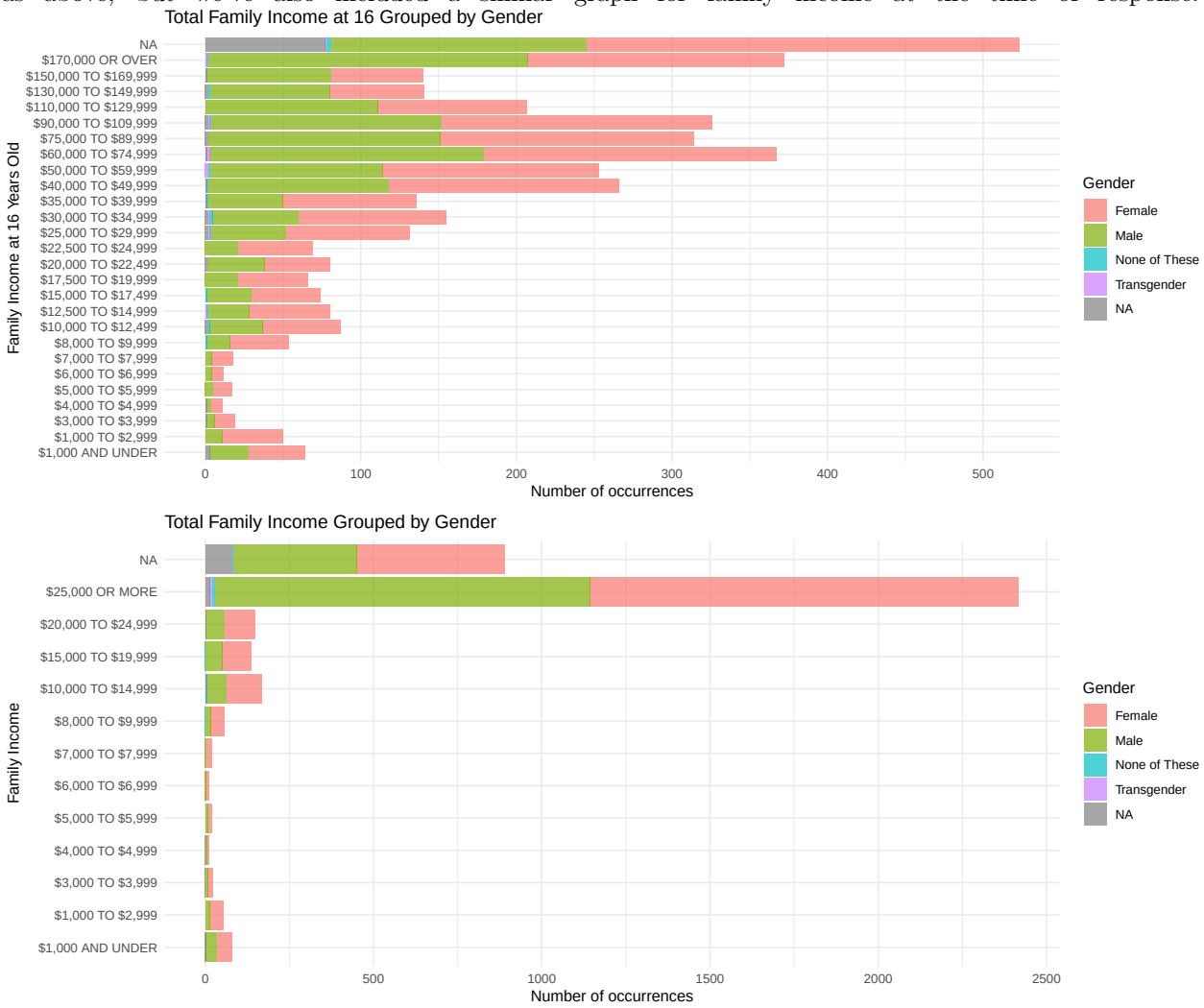
4.1 Visual 1



In this graph (Figure **??**), we can see the proportions of family income when the responded was 16 years of age. We also see that an over-whelming majority of people in our sample had an income of over $40,000. We also see that the single largest individual group is the *NA* group. We will discuss the implications of this in Section 5. As we've previously discussed, when the questioned is prompted with a question, in every situation, they have the option to refuse to answer the question or say that they do not know. The both of

these are listed as NA responses. We also see that of the 5 possible genders (including the NA gender, not to be confused with the NA response to income16), the overwhelming majority of respondents identified as male or female as expected.

We would like to contrast this with the current income of responds. Below is the same graph as above, but we've also included a similar graph for family income at the time of response.



In figure **??**, we see the overlapping of the two graphs, the first of which is the same as figure **??** and the bottom one is the same in essence but with the information for current family income instead of at age 16. The first and most striking part of this visual is that for the current income, the highest threshold is $25,000 or more. We see that the both of our graphs are heavily skewed to wards higher incomes.wesuspect that this is mainly a production of the bin widths decided on for the analysis.

4.2 Implications 1

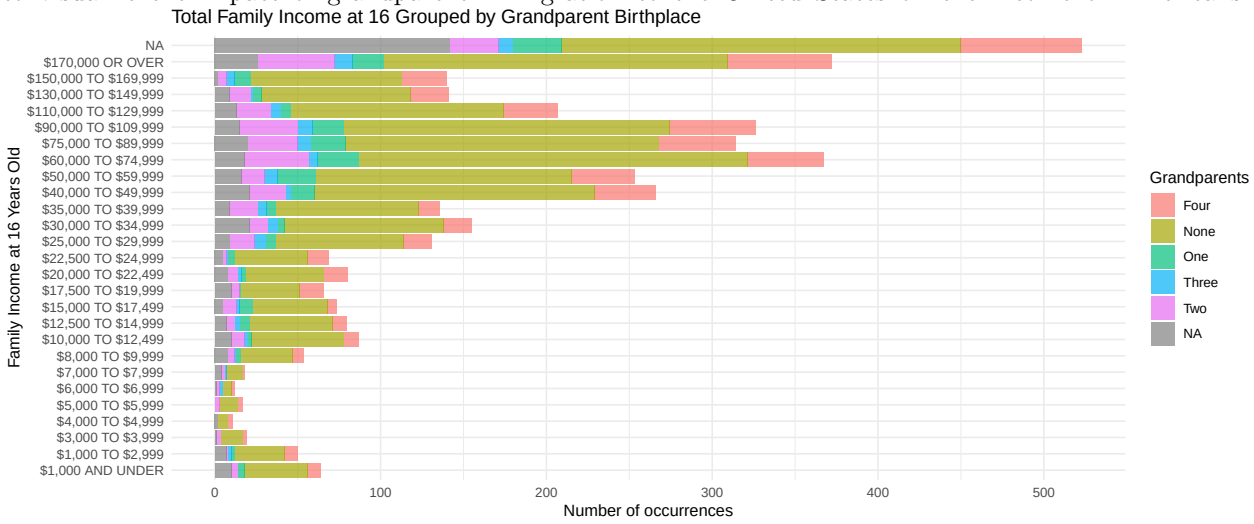
In Visuals 4.1 we can see that when we look into income16 the majority of people claimed that at the age of 16, their family income was over 40 thousand dollars. When we look over at the income variable, we are truly perplexed to see that the largest bin is 25 thousand or more. We will discuss the implications of the latter in Section 5. We can also see that the number of NA responses rises from about 530 in the case of the income at age 16 variable to approximately 900 in the income variable. This seems counter intuitive as it would appear that less people are uncertain about their family income at age 16 than they are now. We suspect however that this is not true. We suspect that there could be a possible shame or embarrassment associated

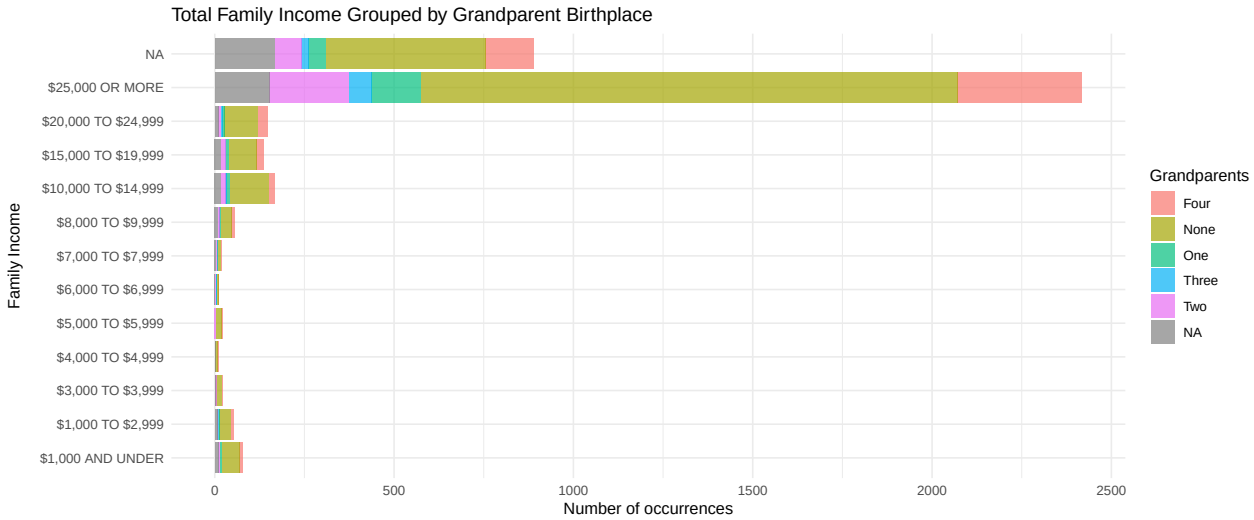
with sharing ones current family income as it represents the persons way of living or lifestyle today. When we speak of the American dream, we implicitly assume an increase in wealth over time and for someone who isn't as wealthy as they might want to be (likely an overwhelming large portion of the population), reluctance with sharing their current income is totally understandable. On the other hand, someones income at age 16 is psychologically less of importance to them and carries no burden in their life, as it is not an expected metric of success years down the line. We also recognize that among men and women, there is in general almost a 50:50 split at each level of income but we do notice that for income level NA, a lot of people chose not to identify their gender. It would be interesting in another analysis to analyse possible evidences of gender pay gap and how people who do not fit in the *traditional* gender roles are affected by the advent of the American Dream and whether or not their position in life hinders their ability to increase in wealth over time.

As we speak about an increase in wealth over time, we next want to see if the birth place of ones grandparents has any impact on income due to so called *generational wealth*.

4.3 Visual 2

As we have previously stated, the American dream is one which communicates the increase of wealth over time for hard working Americans, especially to the eyes of immigrants. As such, we will attempt to visualize the impact of grandparent immigration to the United States on the income of Americans.





4.4 Implications 2

In the above visuals, we can see that in all the income groupings for both age 16 and current income, the largest sub grouping within each is people who have no grandparents born in the United states. The second largest group are usually those with 2 grandparents born in the United States. It is pretty logical that 2 grandparents follows after zero grandparents as it is easy to recognize that most people do not go abroad to marry and marry a friend or someone that they know locally, who is most likely to be born in the United States. We would speculate that a similar trend can be observed in other *immigrant* countries like Canada or Sweden. What is the most fascinating however, is that the vast majority of people who answered the survey are first or second generation Americans, perhaps having grandparents or parents who immigrated to this country chasing the American Dream.

In a similar fashion to what was explained in Section 4.2, both of our graphs are heavily skewed towards higher income. In the case of current income, it is almost certainly because of the bin width decided on by the people who built the survey which we personally find many problems with. In Section 5, we will take a deep dive into this.

4.5 Visuals 3

5 Limitations

In this limitations section,wewant to focus on the data which is provided to us and have a look into why it might not be easy to take the implications mentioned above at face value.

5.1 Uneven Bin width for Income Groups.

The first major flaw I've found in the data is that the bins that both the income at age 16 and current income are recorded in are not all the same size. For example, in the case of the current income variable, the the first bin has a size of $1,000, the second bin has a width of $1,999, the third returns to about $1,000 and as we go higher up the echelon, we have seemingly random alterations in bin widths from $1,000 to $2,000 to $5,000. This can cause serious problems in statistics and especially for purposes of analysis.

We know in data science that when plotting bar charts, the bin width can have a big implications on the presentation of the data and the things that we can infer about its distribution. If the bin width is too small, too large or uneven, we might reduce or add characteristics to the data that aren't actually representative of it. This will cause us to over-fit or under-fit a model when the time comes to create them. The other problem with the bin widths is that they are different from the income at age 16 to the current income. This is the subject of the following Section 5.2.

5.2 Why cut off income at $25,000?

According to a news release published by the United States Bureau of Labor Statistics (see the document called wkyeng.pdf in the literature folder in inputs), the median weekly income of US citizens in the 4th quarter of 2021 (furthest from the start of the pandemic) is $1,010. This information is from that of 116.3 million full-time wage and salary workers. If we do some basic mathematics to approximate the annual earning, we get a figure of over $52,000 dollars per year. As such, we are unable to understand why the bins for current income have been designed as such. This is a major flaw in the technique used by the surveyers to collect the data. In section 5.3

5.3 Next steps

Weaknesses and next steps should also be included.

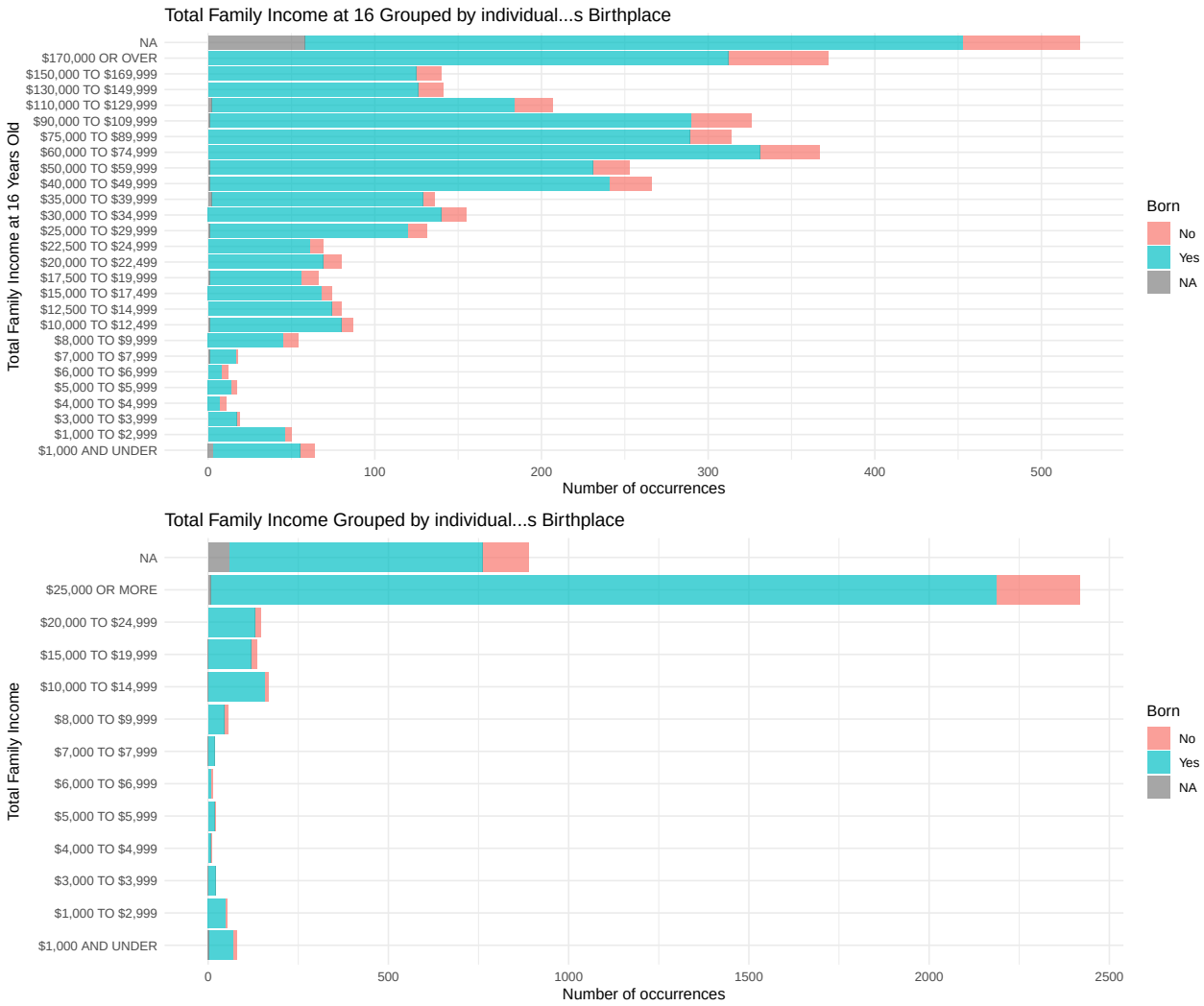


Figure 1: write a figure caption here

Appendix

References

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

Pedersen, Thomas Lin. 2020. *Patchwork: The Composer of Plots*.

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, Jim Hester, and Jennifer Bryan. 2021. *Readr: Read Rectangular Text Data*. https://CRAN.R-project.org/package=readr.

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*.