'SHED'-ing Light on Survey Data

Using R to Automate Production Processes for the Survey of Household Economics and Decisionmaking

Kim Kreiss @KimberlyKreiss

Federal Reserve Board

1. Background
2. Survey, data, and output
3. R and our production process
4. How it works
5. R

Disclaimer

The views expressed in this presentation are those of the authors and do not reflect the position of the Federal Reserve Board of Governors or its staff.

Background

SHED-ing Light on Survey Data

SHED-ing Light on Survey Data

- Each year the Federal Reserve conducts the Survey of Household Economics and Decisionmaking (SHED)

SHED-ing Light on Survey Data

- Each year the Federal Reserve conducts the Survey of Household Economics and Decisionmaking (SHED)
- Nationally representative survey, focusing on the financial lives and experiences of U.S. individuals and households

Our survey is really cool!

- Ask people questions on a range of topics:

Our survey is really cool!

- Ask people questions on a range of topics:
  - Economic Wellbeing
  - Financial Fragility
  - Student loans and education
  - Income and employment
  - Credit and banking experiences
  - Housing, neighborhoods, and living situations
  - Retirement

You may have heard of us!

Snapshot of the data

- The public use dataset has 11,316 observations and 379 variables
- Metadata on the respondent
- Demographic data
- Survey questions
  - Exclusive multiple choice questions
  - Select all that apply/grid multiple choice questions

Snapshot of the data

	CaseID	duration	weight1	weight1b	weight2	weight2b	xsflag	xlaptop	L0_a	L0_b	L0_c	L0_d
1	8931	14	0.7668	20235.199	1.0771	23707.611	Main	non laptop member	Yes	Yes	No	No
2	2323	34	0.4111	10847.140	0.4160	9156.446	Main	non laptop member	No	No	No	No
3	8416	76	1.0188	26884.654	1.1085	24399.799	Main	non laptop member	No	Yes	No	No
4	5838	21	0.7393	19509.830	0.7149	15736.872	Main	non laptop member	No	No	No	No
5	3981	13	1.0096	26640.295	1.1039	24298.135	Main	non laptop member	Yes	No	No	No
6	12096	44	*NA*	*NA*	1.1320	24915.920	LMI	non laptop member	No	Yes	No	No
7	10519	76	*NA*	*NA*	0.7100	15626.964	LMI	non laptop member	No	Yes	No	No
8	5477	18	*NA*	*NA*	0.8245	18147.954	LMI	non laptop member	Yes	No	No	No
9	964	35	0.2394	6318.087	0.1995	4391.896	Main	non laptop member	No	No	No	No
10	192	24	0.7065	18643.273	0.7111	15651.669	Main	non laptop member	No	No	Yes	No
11	11736	17	1.0517	27752.368	0.4530	9970.226	Main	non laptop member	No	No	No	Yes
12	2952	25	*NA*	*NA*	0.6281	13825.679	LMI	non laptop member	Refused	Refused	Refused	Refused

Survey, data, and output

Economic wellbeing

General Well-Being Section

Base: All respondents

[S]

B2. Overall, which one of the following best describes how well you are managing financially these days:

4. Living comfortably
3. Doing okay
2. Just getting by
1. Finding it difficult to get by

Economic wellbeing

CaseID	B2	duration	weight1	weight1b	weight2	weight2b	xsflag
8931	Doing okay	14	0.7668	20235.199	1.0771	23707.611	Main
2323	Doing okay	34	0.4111	10847.140	0.4160	9156.446	Main
8416	Just getting by	76	1.0188	26884.654	1.1085	24399.799	Main
5838	Finding it difficult to get by	21	0.7393	19509.830	0.7149	15736.872	Main
3981	Living comfortably	13	1.0096	26640.295	1.1039	24298.135	Main
12096	Finding it difficult to get by	44	*NA*	*NA*	1.1320	24915.920	LMI
10519	Doing okay	76	*NA*	*NA*	0.7100	15626.964	LMI
5477	Living comfortably	18	*NA*	*NA*	0.8245	18147.954	LMI
964	Just getting by	35	0.2394	6318.087	0.1995	4391.896	Main
192	Doing okay	24	0.7065	18643.273	0.7111	15651.669	Main
11736	Doing okay	17	1.0517	27752.368	0.4530	9970.226	Main

$400 expense

Base: All respondents

[M][SUPPRESS DEFAULT INSTRUCTION]

EF3. Suppose that you have an emergency expense that costs $400. **Based on your current financial situation**, how would you pay for this expense?

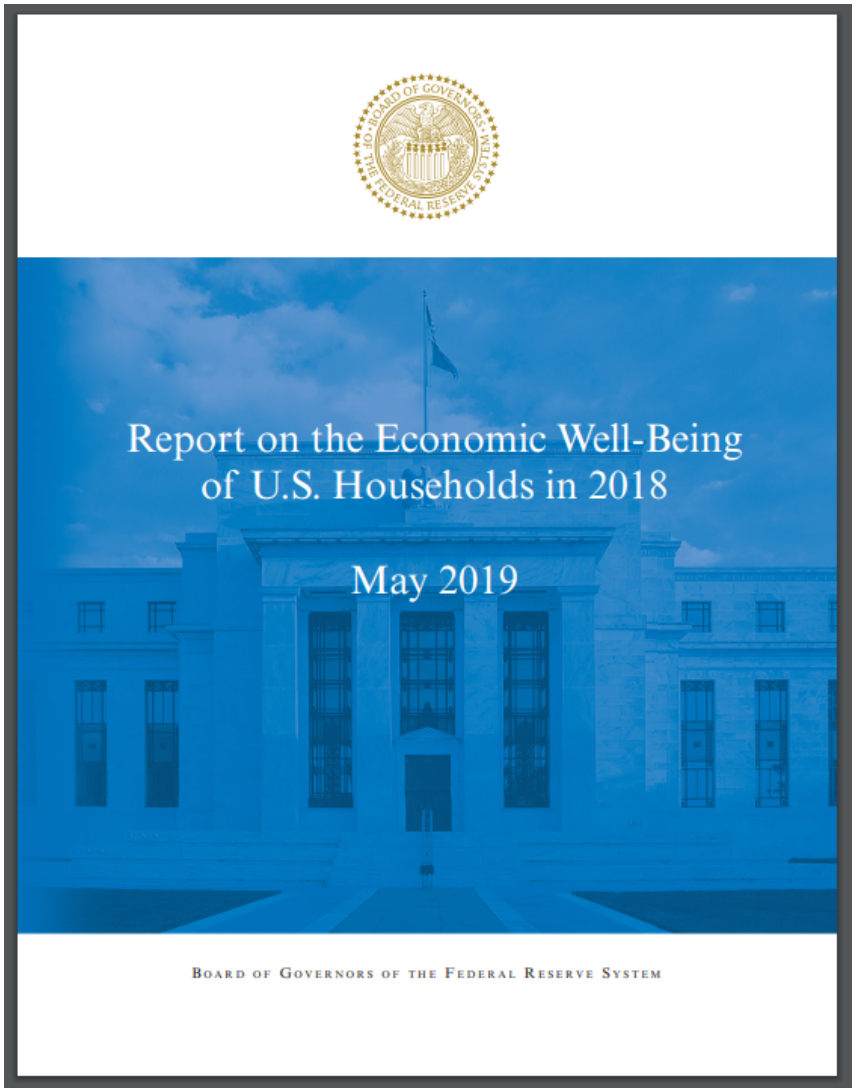
If you would use more than one method to cover this expense, please select all that apply.

a. Put it on my credit card and pay it off in full at the next statement
b. Put it on my credit card and pay it off over time
c. With the money currently in my checking/savings account or with cash
d. Using money from a bank loan or line of credit
e. By borrowing from a friend or family member
f. Using a payday loan, deposit advance, or overdraft
g. By selling something
h. I wouldn't be able to pay for the expense right now
i. Other (please specify): **[text box]**

$400 expense

CaseID	EF3_a	EF3_b	EF3_c	EF3_d	EF3_e	EF3_f	EF3_g	EF3_h	EF3_i	EF3_Refused
8931	No	No	Yes	No	No	No	No	No	No	No
2323	No	No	Yes	No	No	No	No	No	No	No
8416	No	No	No	No	Yes	Yes	Yes	No	No	No
5838	No	No	No	No	Yes	Yes	No	No	No	No
3981	Yes	No	No	No	No	No	No	No	No	No
12096	No	No	No	No	No	No	No	Yes	No	No
10519	No	No	Yes	No	No	No	No	No	No	No
5477	No	No	Yes	No	No	No	No	No	No	No
964	No	No	No	No	No	No	No	Yes	No	No
192	No	No	No	No	Yes	No	No	No	No	No
11736	No	No	No	No	No	No	No	Yes	No	No
2952	No	Yes	No	No	No	No	No	No	No	No

Annual Report



- Produce an annual report on findings
- 10 chapters (64 pages, 31 tables, 36 charts)
- 2 appendices: (84 pages, table for every question)

Annual Report

Income

Income is central to most people's economic well-being. The ability to meet current expenses and save for the future typically depends on income being sufficient and reliable. Some families also depend on financial support from, or provide such support to, their family or friends. Frequent changes in the level of family income, referred to here as "income volatility," can be a source of economic hardship.

Level and Source

Family income in this survey is the income from all sources that the respondent and his or her spouse or partner received during the previous year. Income is reported in dollar ranges as opposed to exact amounts. One-quarter of adults had a family income of less than $25,000 during 2018, and 37 percent had less than $40,000 (figure 2).[4]

Wages and salaries are the most common source of family income: nearly 7 in 10 adults and their spouse or partner received wage income during 2018 (table 4). Yet, many families also receive non-wage income, and the sources of non-wage income vary substantially with age. Among young adults (ages 18 to 29), other paid activities—often referred to as

[4] The income distribution in the SHED is largely similar to the 2018 March Current Population Survey, although a higher fraction of adults in the SHED have family incomes above $40,000 and a lower fraction have incomes below $40,000. The higher income may partly reflect the fact that unmarried partners are treated as one family in the SHED, while the Current Population Survey treats them as two separate families.

Figure 2. Family income distribution

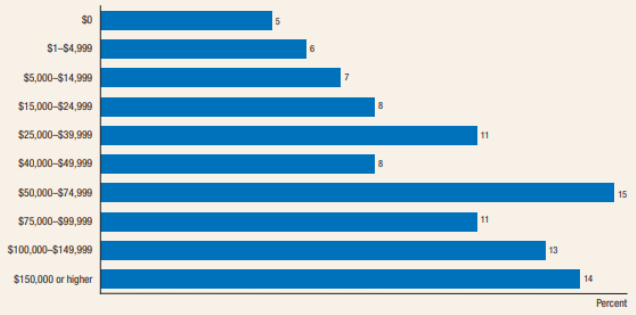
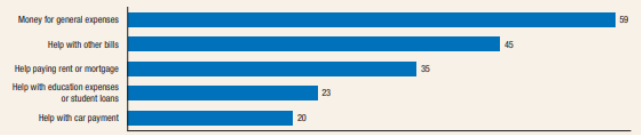
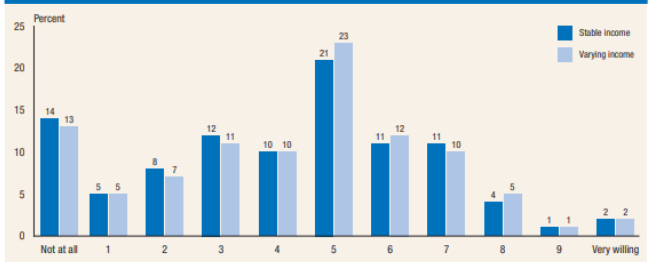


Figure 3. Forms of financial support received from someone outside of the home



Note: Among adults receiving any support from outside the home.

Figure 4. Willingness to take financial risks (by income volatility)



fident in their credit availability (table 6). (Access to credit is discussed further in the "Banking and Credit" section of this report.)

More risk-tolerant individuals may be willing to accept income that is more volatile. On a scale of zero to ten, with "zero" being unwilling to take risks and "ten" being very willing to take risks, more risk-tolerant individuals are somewhat more likely to have varying income than those who are less risk tolerant (figure 4). However, the difference in income volatility by risk tolerance is modest. This suggests that factors other than individual risk preferences likely drive income volatility.

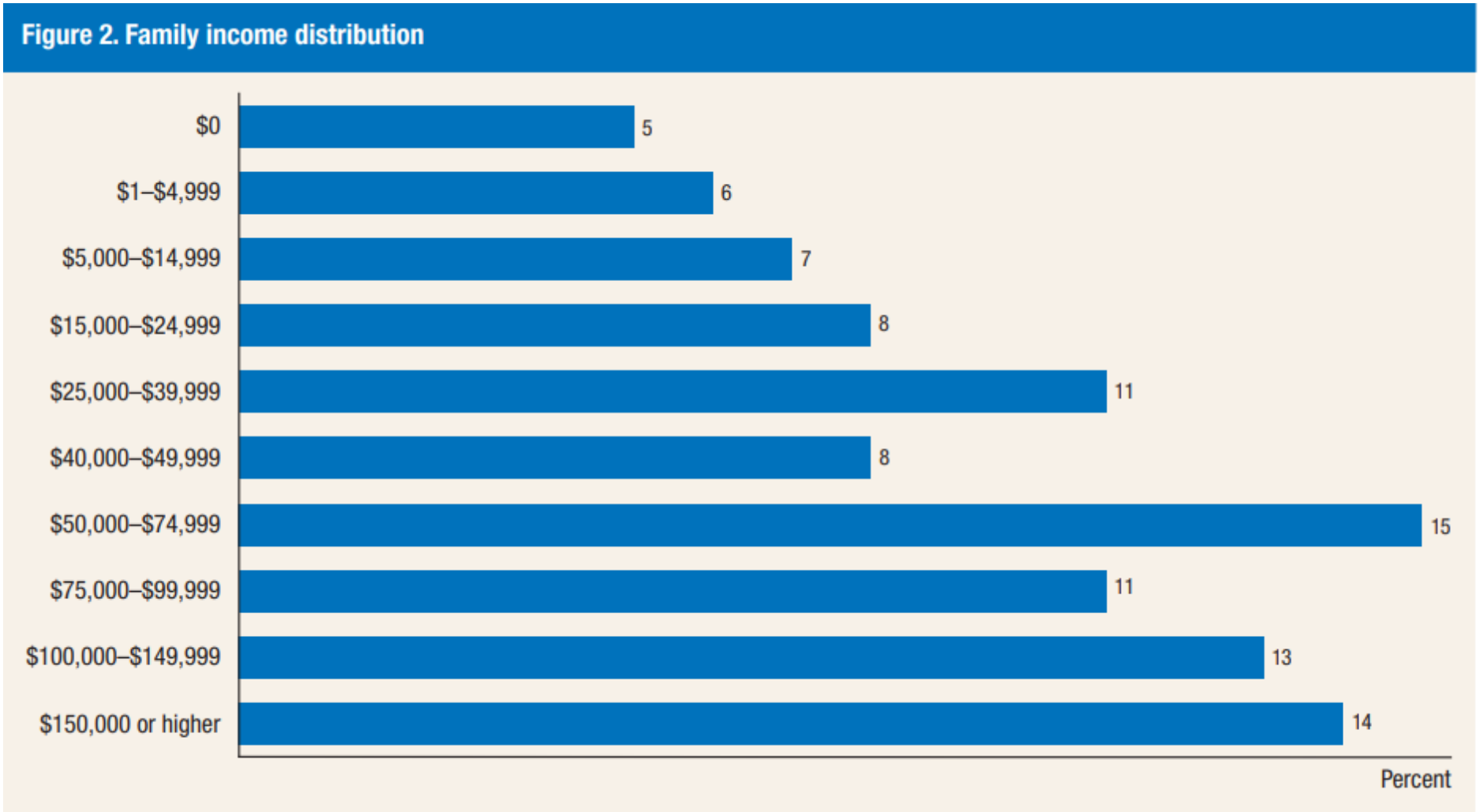
Table 6. Income volatility and related hardship (by credit confidence)
Percent

Expect credit card application would be approved	Stable income	Varying income	
		No hardship	Causes hardship
Confident	73	20	6
Not confident	64	9	26
Overall	**71**	**19**	**9**

Note: Among adults receiving any support from outside the home.

Annual Report

Figure 2. Family income distribution



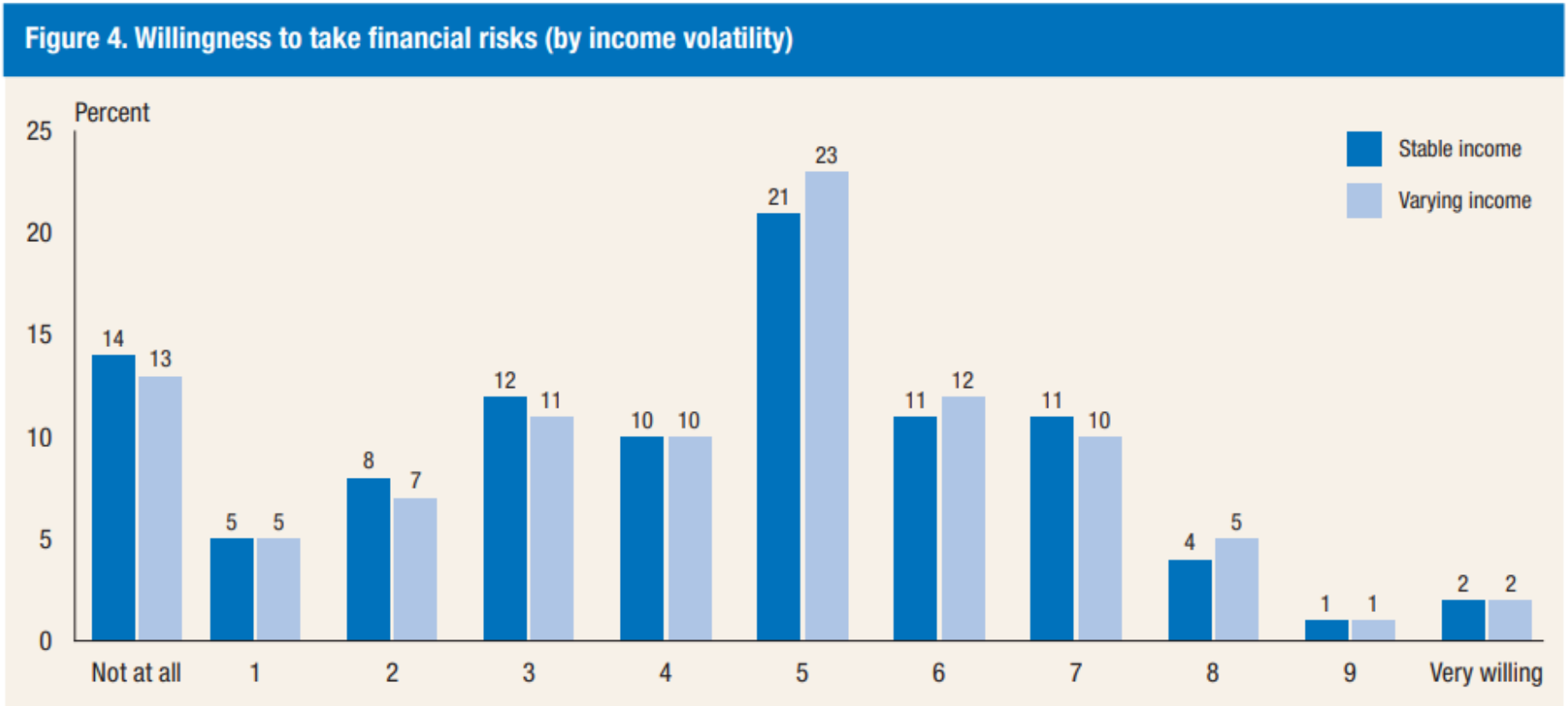
Annual Report

Table 4. Family income sources (by age)
Percent

Income source	18–29	30–44	45–59	60+	Overall
Wages or salaries	77	83	80	38	68
Self-employment	14	19	19	14	16
Other paid activities	19	13	9	7	12
Interest, dividends, or rental income	15	21	29	44	28
Social Security (including old age, SSI, and DI)	4	7	14	76	28
Unemployment income	3	3	3	2	3
Pension	1	2	9	51	18
Any other income	7	6	7	15	9

Note: Respondents can select multiple answers.

Annual Report



R and our production process

Different parts of the production process

1. Generating numbers for the report
2. **Generating tables and charts for the report**
3. Quality control

Production process

- It takes a lot of work to turn the raw data into analysis for the report–especially the generation of tables and charts

Production process

- It takes a lot of work to turn the raw data into analysis for the report–especially the generation of tables and charts
- We have several different types of tables and charts we want to produce

Production process

- It takes a lot of work to turn the raw data into analysis for the report–especially the generation of tables and charts
- We have several different types of tables and charts we want to produce
- We want to be flexible in changing our analysis if we need to

Production process

- It takes a lot of work to turn the raw data into analysis for the report–especially the generation of tables and charts
- We have several different types of tables and charts we want to produce
- We want to be flexible in changing our analysis if we need to
- Need to produce excel workbooks of underlying data of each table/chart and a rendering of charts

Our process with R

- Use R:
  - to eliminate any manual data entry such as copy and pasting
  - eliminate any manual reformatting of data such as editing format of table in Excel workbook

Our process with R

- Use R:
  - to eliminate any manual data entry such as copy and pasting
  - eliminate any manual reformatting of data such as editing format of table in Excel workbook
  - to generate tables and charts for the report in a pdf document and excel file of the underlying data

Our process with R

- Use R:
  - to eliminate any manual data entry such as copy and pasting
  - eliminate any manual reformatting of data such as editing format of table in Excel workbook
  - to generate tables and charts for the report in a pdf document and excel file of the underlying data
- Use git for version control

Our process with R

- Efficient

Our process with R

- Efficient
- Flexible

Our process with R

- Efficient
- Flexible
- Minimizes human error from tasks such as data entry

Our process with R

- Efficient
- Flexible
- Minimizes human error from tasks such as data entry
- Neatly tracks workflow and manages files using git

How it works

How it works

```
library(tidyverse)
library(rmarkdown)
library(kableExtra)
library(questionr)
library(data.table)
library(scales)
library(plotly)
library(ggalt)
library(reshape2)
```

How it works

- Receive and do basic cleaning on data

How it works

- Receive and do basic cleaning on data
- Write custom functions that take in basic information such as variable(s), chart type, and titles and produce the figures

How it works

- Receive and do basic cleaning on data
- Write custom functions that take in basic information such as variable(s), chart type, and titles and produce the figures
- For each chapter, create an Rmarkdown file that produces a pdf of all the tables and charts in that chapter and excel file of the underlying data

R

Examples

- Use tables and charts from income chapter
- Show code for different functions

Figure 2: Income distribution

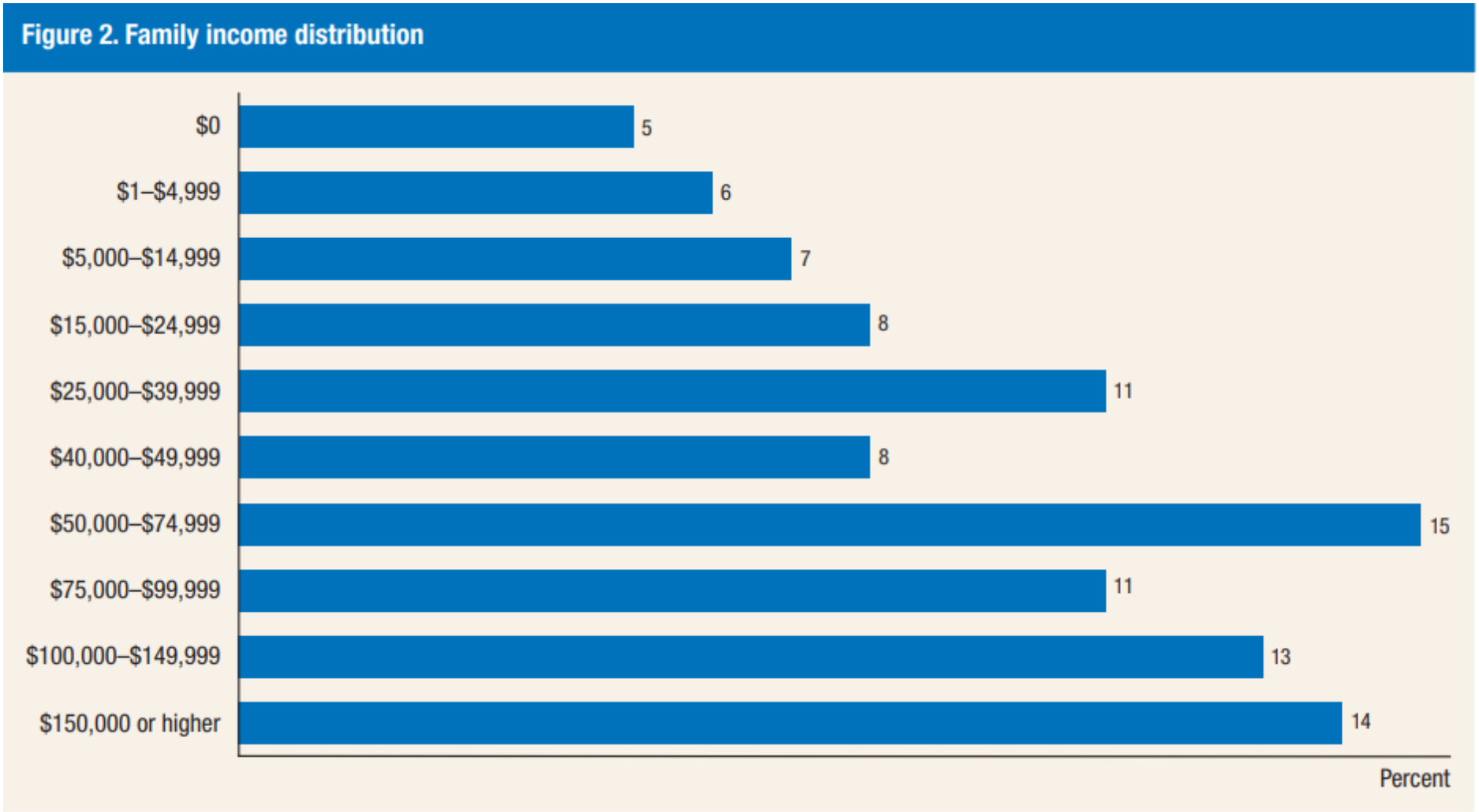


Figure 2: Income distribution

Figure 2: Income distribution

Base: I0=1 for any response OR I0A=1 or refused
(Report having any income)

[S]

[If refused, prompt once: "We ask for information about your income because it is extremely important for our understanding of household finances in the United States. We greatly appreciate your response and your answer will remain completely anonymous."]

I40. Which of the following categories best describes the total income that you [**IF PPMARIT=1, INSERT:** and your spouse / **IF PPMARIT=6, INSERT:** and your partner] received from all sources, before taxes and deductions, in the past 12 months?

1. $0 to $4,999
2. $5,000 to $14,999
3. $15,000 to $24,999
4. $25,000 to $39,999
5. $40,000 to $49,999
6. $50,000 to $74,999
7. $75,000 to $99,999
8. $100,000 to $149,999
9. $150,000 to $199,999
10. $200,000 or higher

Figure 2: Income distribution

```
# Read in the cleaned data
# Source the functions and required libraries
# path is defined here
source("readdata.R")
source("functions.R")
```

Figure 2: Income distribution

```
# Read in the cleaned data
# Source the functions and required libraries
# path is defined here
source("readdata.R")
source("functions.R")
```

```
wtd.table(x = data_input[["I40"]], weights = data_input[["weight2b"]]) %>%
  prop.table() %>%
  data.frame()
```

```
##                    Var1       Freq
## 1               Refused 0.01290644
## 2                   $0 0.05301767
## 3         $1 to $4,999 0.05881143
## 4     $5,000 to $14,999 0.07082167
## 5    $15,000 to $24,999 0.07779846
## 6    $25,000 to $39,999 0.11432883
## 7    $40,000 to $49,999 0.07993709
## 8    $50,000 to $74,999 0.14602770
## 9    $75,000 to $99,999 0.10901627
## 10 $100,000 to $149,999 0.13410922
## 11 $150,000 to $199,999 0.07853158
## 12   $200,000 or higher 0.06469364
```

Figure 2: Income distribution

```
# Specify the Excel file for this chapter's output
excel <- paste0(path, "income.xlsx")
wb <- openxlsx::createWorkbook()

#Figure 2: Income distribution
I40 <- prep_onevar_multicat(data_input, "i40_10cat") %>%
  rename(Income = question)
```

```
# prepare 1 variable with multiple categories so that it is ready to be charted
# typically, all categories should sum up to 100% or close to 100% (due to
# refusals)
prep_onevar_multicat <- function(df, var) {
  df %>%
    shed_table(var, var1_label = "question") %>%
    filter(question != "Refused") %>%
    mutate(question=factor(question,
                           levels = rev(labels[[var]]),
                           ordered = TRUE))
}
```

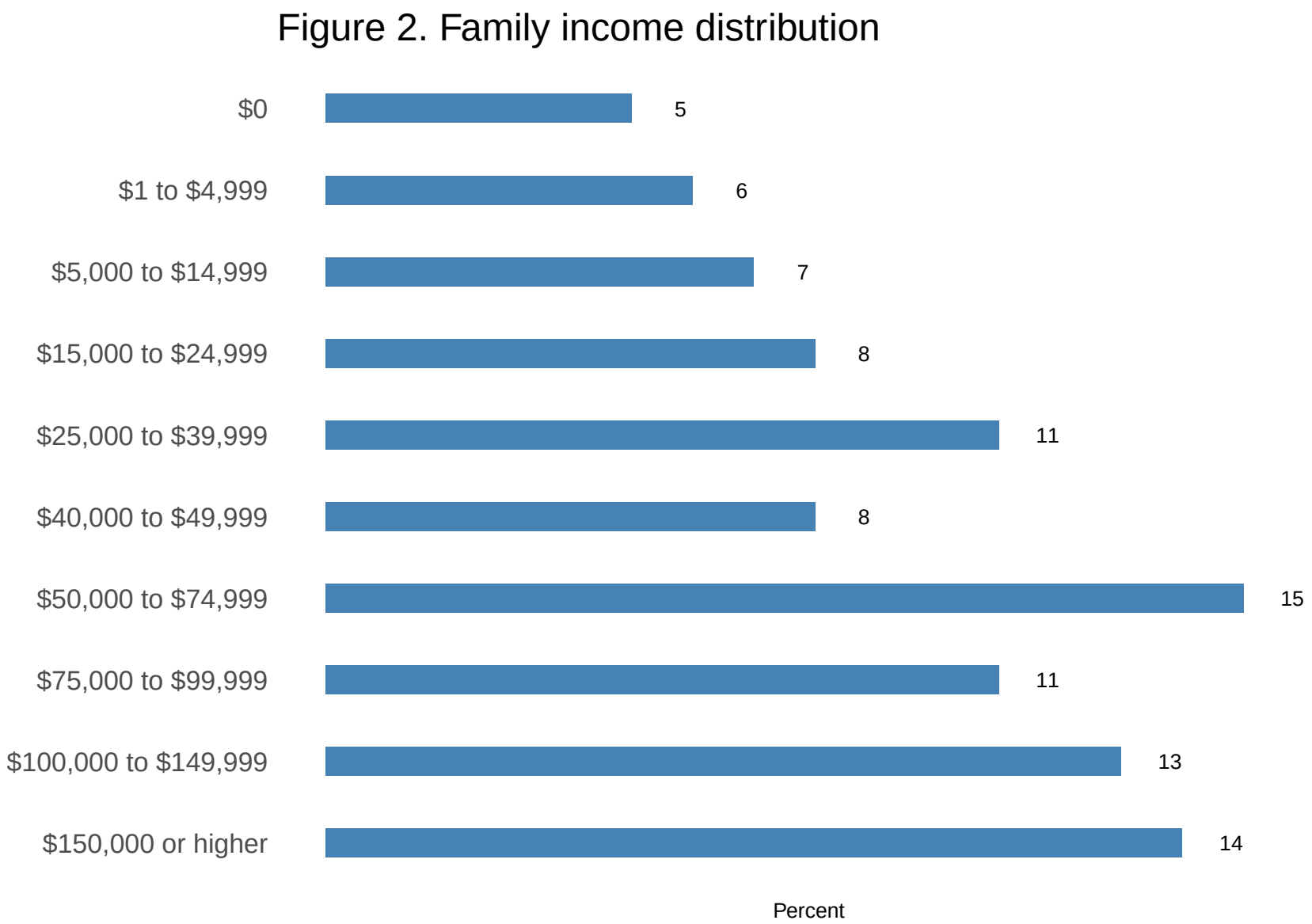
Underlying functions

```
shed_table <- function(df, var1, var2, var1_label, var2_label) {
  if (missing(var2)) {
    a <- wtd.table(df[[var1]],
                   weights = df[["weight2b"]],
                   na.show = TRUE) %>%
      prop.table() %>%
      as.data.frame() %>%
      as_tibble()
    colnames(a) <- c(var1, "Percent")
  } else {
    a <- wtd.table(df[[var1]],
                   df[[var2]],
                   weights = df[["weight2b"]],
                   na.show = TRUE) %>%
      prop.table(1) %>%
      as.data.frame() %>%
      as_tibble()
    colnames(a) <- c(var1, var2, "Percent")
  }

  if (missing(var2_label)) {
    if (missing(var1_label)) {
      var1_label <- attributes(df[[var1]])$label
    }
    colnames(a)[1] <- var1_label
  } else {
    colnames(a)[1:2] <- c(var1_label, var2_label)
  }
  a <- mutate(a,
              Percent = round(as.numeric(Percent) * 100,
                              digits = 0))
  return(a)
}
```

Figure 2: Income distribution

```
do_all(I40, sort = FALSE, reference = "Figure 2", Title = "Family income distribution",
       chart_type = "one_bar_chart", x_var = "Income", orientation = coord_flip())
```



Underlying functions

```
one_bar_chart <- function(df, orientation = NULL, special = NULL) {
  g <- ggplot(df,
              aes_string(y = "Percent",
                         x = attributes(df)$x_var)) +
    geom_bar(
      position = "dodge",
      fill = "steelblue",
      stat = "identity",
      width = .35
    ) +
    geom_text(aes(y = Percent + 0.8,
                  label = round(Percent, digits = 0))) +
    guides(fill = guide_legend()) +
    labs(y = "Percent",
         x = "",
         caption = paste(strwrap(attributes(df)$footnote,
                                 width = 80),
                         collapse = "\n")) +
    theme(
      panel.grid.major = element_blank(),
      panel.grid.minor = element_blank(),
      panel.background = element_blank(),
      axis.ticks = element_blank(),
      axis.text = element_text(size=14),
      plot.title = element_text(size=20, hjust = 0),
      plot.caption = element_text(size=14, hjust = 0)) +
    ggtitle(paste(strwrap(paste0(attributes(df)$reference,". ",
                                 attributes(df)$title),
                  width = 70),
            collapse = "\n")) +
    orientation +
    special
```

```
  if (is.null(orientation)) {
    g <- g + theme(axis.text.y = element_blank())
  } else {
    if (is.na(match("special", names(orientation)))) {
      g <- g + theme(axis.text.x = element_blank())
    }
  }
  g
}
```

Conclusion

- The format of our data can make it hard to easily produce tables and charts
- We can use R to account for that and automate production processes for our annual survey
- Still needs some work, but has greatly improved our process
- Eventually hope to make a package that can easily allow people to pull and use easy functions to work with SHED data

Thank you!

- More about SHED:
  www.federalreserve.gov/consumerscommunities/shed.htm
- More about me:
  - kimberlykreiss.github.io
  - github.com/kimberlykreiss
  - @KimberlyKreiss
- Questions, feedback, suggestions: kimberly.kreiss@frb.gov