Reproducible Research in R

RLadiesSTL

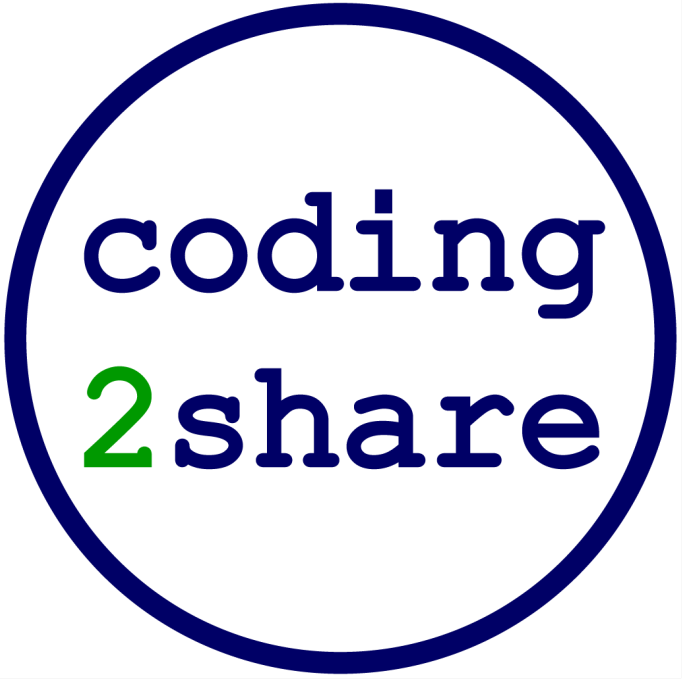
Bobbi J. Carothers

October 3, 2018

Introduction

- Ever inherit a project and have *no* idea how the data were managed and/or analyses performed?
- Ever come back to your own project after a few months and have no recollection of what you did or why you did it, even after looking at your own code?

- Ways to help others and future-you
  - Literate Programming
  - Code Formatting
  - Data Cleaning
  - Data Documentation



Literate Programming

Communicating what you did so that others (or future you) can replicate it

Literate Programming

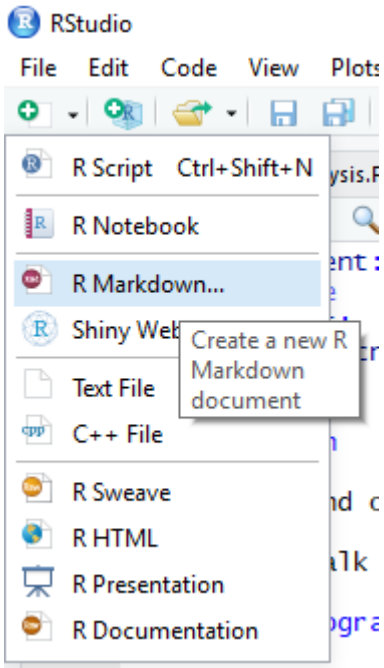
R Markdown

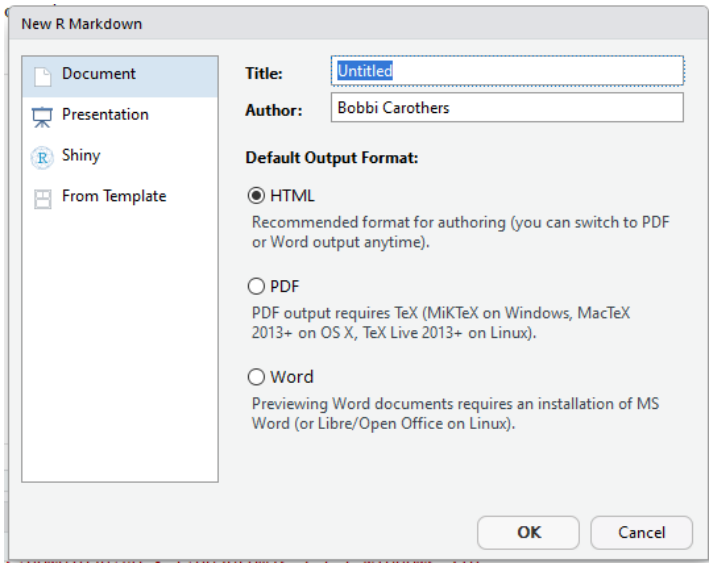
- Create polished documents detailing how data are managed and analyses are conducted
- Merge prose, code, analyses all in one document
- Final document is in a non-proprietary format (.pdf or .html) suitable to share publicly
- More attractive and more easily shared than syntax/code files from SPSS, SAS, Stata, etc
- Even these slides were written in R Markdown!

Literate Programming

R Markdown

Open up RStudio, then open up a new R Markdown document:





Literate Programming

YAML (Yet Another Markup Language) Header

Sets up the title and other output options:

```
---
title: "Reproducible Research in R"
subtitle: "RLadiesSTL"
author: "Bobbi J. Carothers"
date: "October 3, 2018"
output:
  html_document:
    toc: true
    toc_float:
      smooth_scroll: true
---
```

Save to wherever you've saved the data we're working with today.

Literate Programming

Code Chunks

The code that you want R to run goes in "chunks" that are wrapped in backticks with `{r}` at the end of the top wrapper. You can set options for how you want R to handle each chunk.

Chunk options:

```
```{r, include=FALSE}
# Code will run and results can be used in other chunks, but neither will
# appear in the knitted file.
```
```

```
```{r, echo=FALSE}
# Code will run, code will not be displayed but the results will be
```
```

```
```{r, eval=FALSE}
# Code will not run, but will be displayed. Good for demonstration purposes.
```
```

```
```{r, message=FALSE}
# Some code will produce messages other than results; can turn this off to
# reduce clutter.
```
```

```
```{r, warning=FALSE}
# Turn off warnings.
```
```

Literate Programming

Text options

Everything written outside of the chunks will *not* be evaluated by R and will display as normal text.

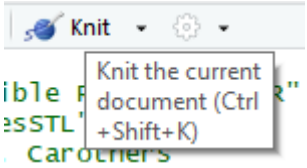
- `#Level 1 Heading`
- `##Level 2 Heading`
- `**Bold**`
- `*Italics*` or `_Italics_`

- # Level 1 Heading
- ## Level 2 Heading
- **Bold**
- *Italics*

Literate Programming

Output

When you're ready to create the output file, you "knit" it:



For more options:

- Download the R Markdown Cheatsheet
- Check out the R Markdown Website.

Code Formatting

Recommendations to make your life easier

Code Formatting: Using Space Wisely

Recommendation #1: Use white space to separate processes

Bad:

```
code_data_avail <- cbind(table(r$Q25_2_1),table(r$Q25_2_2))
colnames(code_data_avail) <- c("Did you make your data publicly available?","Did y
fig2 <- ggplot(code_data_avail, aes(x=data_or_code, y=number, fill=avail)) + geom_
fig2
```

Better:

```
code_data_avail <- cbind(table(r$Q25_2_1),table(r$Q25_2_2))
colnames(code_data_avail) <- c("Did you make your data publicly available?","Did y

fig2 <- ggplot(code_data_avail, aes(x=data_or_code, y=number, fill=avail)) + geom_

fig2
```

Code Formatting: Using Space Wisely

Recommendation #2: Limit line length to 80 characters

```
code_data_avail <- cbind(table(r$Q25_2_1),table(r$Q25_2_2))
colnames(code_data_avail) <- c("Did you make your data publicly available?",
"Did you make your code publicly available?")
code_data_avail <- melt(code_data_avail)
colnames(code_data_avail) <- c("avail","data_or_code","number")

fig2 <- ggplot(code_data_avail, aes(x=data_or_code, y=number, fill=avail)) +
geom_col(position="dodge") +  coord_flip() +
theme(legend.position = 'top') +
labs(y="Number of participants", x="", fill="") +
scale_fill_manual(values=fills)

fig2
```

Code Formatting: Using Space Wisely

Recommendation #3: Indent to group lines of code that belong together

```
code_data_avail <- cbind(table(r$Q25_2_1),table(r$Q25_2_2))
colnames(code_data_avail) <- c("Did you make your data publicly available?",
                               "Did you make your code publicly available?")
code_data_avail <- melt(code_data_avail)
colnames(code_data_avail) <- c("avail","data_or_code","number")

fig2 <- ggplot(code_data_avail, aes(x=data_or_code, y=number, fill=avail)) +
  geom_col(position="dodge") +  coord_flip() +
  theme(legend.position = 'top') +
  labs(y="Number of participants", x="", fill="") +
  scale_fill_manual(values=fills)

fig2
```

Code Formatting: Naming Conventions

Recommendation #4: Use meaningful names for objects

Bad:

```
r$Q11_2[r$Q11_2==-99] <- NA
prop.table(table(r$Q11_2))
```

Better: replace `r` with `HEALTH_SURVEY` and `Q11_2` with `race`.

```
HEALTH_SURVEY$race[HEALTH_SURVEY$race==-99] <- NA
prop.table(table(HEALTH_SURVEY$race))
```

Code Formatting: Naming Conventions

Recommendation #5: Use dot.case, camelCase, or snake_case for multi-part names

- Variations include lower camelCase, upper CamelCase, UPPER_SNAKE_CASE and so on
- Consider using one format for functions, another for dataframes, and a third for variable names

```
find_mode(HEALTH_SURVEY$insure.status)
```

- Note that some formats might not work in certain software packages.

Code Formatting: Naming Conventions

Recommendation #6: Add meta-data to file names

- Include meta-data like the date and project name in file names
- Key principles:
    1. Machine readable
    2. Works with default ordering
    3. Human readable
- Examples:
    - 180130_raw_preProgram.csv
    - 180131_clean_preProgram.csv
    - 180228_raw_postProgram.csv
    - 180302_clean_postProgram.csv

Code Formatting: Explain

Recommendation #7: Write a prolog to introduce the code

```
# PROLOG   ################################################################

# PROJECT: NAME OF PROJECT HERE
# PURPOSE: MAJOR POINT(S) OF WHAT I AM DOING WITH THE DATA HERE
# DIR:     list directory(-ies) for files here
# DATA:    list dataset file names/availability here, e.g.,
#          filename.correctextention
#          somewebaddress.com
# AUTHOR:  AUTHOR NAME(S)
# CREATED: MONTH dd, YEAR
# LATEST:  MONTH dd, YEAR
# NOTES:   indent all additional lines under each heading,
#          & use the apostrophe hashmark bookends that appear
#          KEEP PURPOSE, AUTHOR, CREATED & LATEST ENTRIES IN UPPER CASE,
#          with appropriate case for DIR & DATA, lower case for notes
#          If multiple lines become too much,
#          simplify and write code book and readme.

# PROLOG   ################################################################
```

Code Formatting: Explain

Recommendation #8: Annotate to clarify code purpose

- Use comments to:
  - Explain the reason for the code (if needed)
  - Explain functionality or choices that are not obvious or are different from expected
  - Identify hacks or errors that should be fixed or rewritten
- Avoid using comments to:
  - Explain poorly named objects
  - Repeat things that can be easily understood from the code

```
#check normality assumption for age variable
histoAge <- hist(age)
histoAge
```

Data Cleaning

Format data for easy analysis, documentation, and sharing

Labeling variables

Recoding values

Labeling values

Data Cleaning

Code used to pull the example data:

```
library(RNHANES) # read data directly from NHANES site
NHANES <- nhanes_load_data("AUQ_G", "2011-2012", demographics=TRUE) # load auditor
save(NHANES, file="NHANES.Rdata") # save to working directory
```

Load libraries and import data:

```
library(dplyr) # data managment
library(labelled) # labeling variables
load("C:\\Wherever\\NHANES.Rdata")
```

Big pile of data, but we don't really know what the numbers mean. Luckily, they document it really well at their Data Documentation, Codebook, and Frequencies site. We'll use this as a model to aspire to.

Data Cleaning

Pull a subset of the data to work with to keep things manageable for this example.

```
NHclean <- subset(NHANES,
                  select=c(SEQN, RIAGENDR, RIDAGEYR, RIDRETH1,
                           AUQ054, AUQ060, AUQ100, AUQ144)
                  )
# Take a quick look
NHclean[1:5,]
```

```
##    SEQN RIAGENDR RIDAGEYR RIDRETH1 AUQ054 AUQ060 AUQ100 AUQ144
## 1 62161        1       22        3      2      1      5      4
## 2 62162        2        3        1      1     NA     NA     NA
## 3 62163        1       14        5      2     NA     NA     NA
## 4 62164        2       44        3      1     NA      4      4
## 5 62165        2       14        4      2     NA     NA     NA
```

Data Cleaning

Rename demographic variables:

```
NHclean <- rename(NHclean, # dataset
                  ID=SEQN, # new varname goes first
                  Gender=RIAGENDR, Age=RIDAGEYR, Ethn=RIDRETH1)
NHclean[1:5,]
```

```
##      ID Gender Age Ethn AUQ054 AUQ060 AUQ100 AUQ144
## 1 62161      1  22    3      2      1      5      4
## 2 62162      2   3    1      1     NA     NA     NA
## 3 62163      1  14    5      2     NA     NA     NA
## 4 62164      2  44    3      1     NA      4      4
## 5 62165      2  14    4      2     NA     NA     NA
```

Data Cleaning

Change class from "integer" to "double" so things play well later on.

```
class(NHclean$ID)
```

```
## [1] "integer"
```

```
NHclean <- NHclean %>% # many dplyr functions use pipes for ease of use
  mutate(ID = as.double(ID),
         Gender = as.double(Gender),
         Age = as.double(Age),
         Ethn = as.double(Ethn),
         AUQ054 = as.double(AUQ054),
         AUQ060 = as.double(AUQ060),
         AUQ100 = as.double(AUQ100),
         AUQ144 = as.double(AUQ144)
         )
class(NHclean$ID)
```

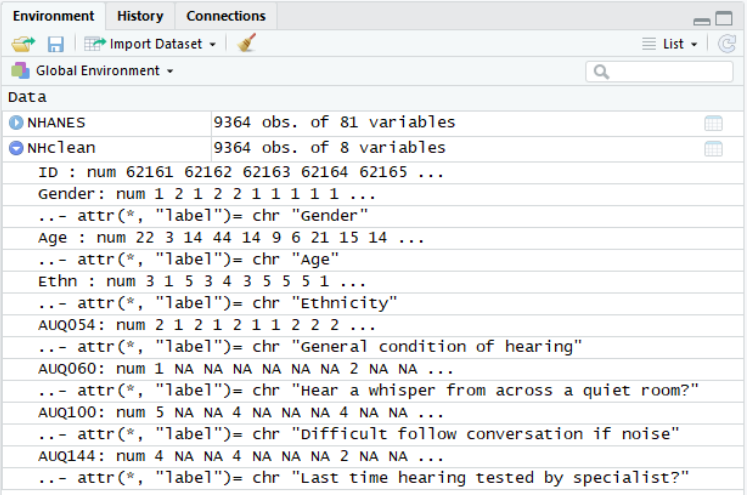
```
## [1] "numeric"
```

Shows up as "numeric," which is fine.

Data Cleaning

Add variable labels:

```
var_label(NHclean) <- list(Gender="Gender",
                           Age="Age",
                           Ethn="Ethnicity",
                           AUQ054="General condition of hearing",
                           AUQ060="Hear a whisper from across a quiet room?",
                           AUQ100="Difficult follow conversation if noise",
                           AUQ144="Last time hearing tested by specialist?")
```



Data Cleaning

Add value labels for categorical variables:

```
NHclean$Ethn <- labelled(NHclean$Ethn,
                         c("Mexican American" = 1,
                           "Other Hispanic" = 2,
                           "Non-Hispanic White" = 3,
                           "Non-Hispanic Black" = 4,
                           "Other Race - Including Multi-Racial" = 5
                           )
                         )
val_labels(NHclean$Ethn)
```

```
##                     Mexican American                     Other Hispanic
##                                    1                                  2
##                  Non-Hispanic White                  Non-Hispanic Black
##                                    3                                  4
## Other Race - Including Multi-Racial
##                                    5
```

Data Cleaning

Recode categories before applying labels when appropriate:

```
# Gender: male = 1 female = 2
table(NHclean$Gender)
```

```
##
##    1    2
## 4663 4701
```

```
# Recode to male = 0 female = 1
NHclean$Gender <- NHclean$Gender - 1
# Apply labels
NHclean$Gender <- labelled(NHclean$Gender,
                           c("Male" = 0,
                             "Female" = 1)
                           )
# Check value labels
val_labels(NHclean$Gender)
```

```
##   Male Female
##      0      1
```

Data Cleaning

Determine how to handle missing values with scale variables:

Variable Name: AUQ054

SAS Label: General condition of hearing

English Text: These next questions are about {your/SP's} hearing. Which statement best describes {your/SP's} hearing (without a hearing aid or other listening devices)? Would you say {your/his/her} hearing is excellent, good, that {you have/s/he has} a little trouble, moderate trouble, a lot of trouble, or {are you/is s/he} deaf?

Target: Both males and females 1 YEARS - 150 YEARS

Code or Value	Value Description	Count	Cumulative	Skip to Item
1	Excellent	4244	4244	
2	Good	3744	7988	
3	A little trouble	869	8857	
4	Moderate hearing trouble	306	9163	
5	A lot of trouble	172	9335	
6	Deaf	12	9347	
77	Refused	1	9348	
99	Don't know	15	9363	
.	Missing	1	9364	

Are "Refused" and "Don't know" interesting answers? If not, recode as missing.

Data Cleaning

Remove "Refused" and "Don't know" from AUQ054

```
# check initial frequencies
table(NHclean$AUQ054)
```

```
##
##    1    2    3    4    5    6   77   99
## 4244 3744  869  306  172   12    1   15
```

```
# replace 77 and 99 with NA
NHclean$AUQ054 <- na_if(NHclean$AUQ054
NHclean$AUQ054 <- na_if(NHclean$AUQ054
# check frequencies again
table(NHclean$AUQ054)
```

```
##
##    1    2    3    4    5    6
## 4244 3744  869  306  172   12
```

Variable Name: AUQ054

SAS Label: General condition of hearing

English Text: These next questions are about {your/SP's} hearing. Which statement best describes {your/SP's} hearing (without a hearing aid or other listening devices)? Would you say {your/his/her} hearing is excellent, good, that {you have/s/he has} a little trouble, moderate trouble, a lot of trouble, or {are you/is s/he} deaf?

Target: Both males and females 1 YEARS - 150 YEARS

Code or Value	Value Description	Count	Cumulative	Skip to Item
1	Excellent	4244	4244	
2	Good	3744	7988	
3	A little trouble	869	8857	
4	Moderate hearing trouble	306	9163	
5	A lot of trouble	172	9335	
6	Deaf	12	9347	
77	Refused	1	9348	
99	Don't know	15	9363	
.	Missing	1	9364	

Data Cleaning

Remove "Refused" and "Don't know" from AUQ060

```
# check initial frequencies
table(NHclean$AUQ060)
```

```
##
##    1    2    9
## 2128  745   32
```

```
# replace 9 with NA
NHclean$AUQ060 <- na_if(NHclean$AUQ060
# check frequencies again
table(NHclean$AUQ060)
```

```
##
##    1    2
## 2128  745
```

Variable Name: AUQ060

SAS Label: Hear a whisper from across a quiet room?

English Text: These next questions refer to hearing without the use of a hearing aid or any other listening devices. If {you have/SP has} one ear that is better than the other, please answer the questions for the hearing in {your/SP's} better ear. Can {you/SP} usually hear and understand what a person says without seeing his or her face if that person whispers to {you/him/her} from across a quiet room?

Target: Both males and females 20 YEARS - 69 YEARS

Code or Value	Value Description	Count	Cumulative	Skip to Item
1	Yes	2128	2128	AUQ100
2	No	745	2873	
7	Refused	0	2873	
9	Don't know	32	2905	
.	Missing	6459	9364	

Data Cleaning

Remove "Refused" and "Don't know" from AUQ100

```
# check initial frequencies
table(NHclean$AUQ100)
```

```
##
##    1    2    3    4    5    9
##  173  356  551 1145 2448    2
```

```
# replace 9 with NA
NHclean$AUQ100 <- na_if(NHclean$AUQ100
# check frequencies again
table(NHclean$AUQ100)
```

```
##
##    1    2    3    4    5
##  173  356  551 1145 2448
```

Variable Name: AUQ100

SAS Label: Difficult follow conversation if noise

English Text: How often {do you/does SP} find it difficult to follow a conversation if there is background noise, for example, when other people are talking, TV or radio is on, or children are playing? Would you say...

English Instructions: HAND CARD AUQ1

Target: Both males and females 20 YEARS - 69 YEARS

Code or Value	Value Description	Count	Cumulative	Skip to Item
1	Always	173	173	
2	Usually	356	529	
3	About half the time	551	1080	
4	Seldom	1145	2225	
5	Never	2448	4673	
7	Refused	0	4673	
9	Don't know	2	4675	
.	Missing	4689	9364	

Data Cleaning

Remove "Refused" and "Don't know" from AUQ144

```
# check initial frequencies
table(NHclean$AUQ144)
```

```
##
##    1    2    3    4    5    9
##  369  652  547 1381 1607  119
```

```
# replace 9 with NA
NHclean$AUQ144 <- na_if(NHclean$AUQ144
# check frequencies again
table(NHclean$AUQ144)
```

```
##
##    1    2    3    4    5
##  369  652  547 1381 1607
```

Variable Name: AUQ144

SAS Label: Last time hearing tested by specialist?

English Text: A hearing test by a specialist is one that is done in a sound proof booth or room, or with headphones. Hearing specialists include audiologists, ear nose and throat doctors, and trained technicians or occupational nurses. When was the last time {you had/SP had} {your/his/her} hearing tested by a hearing specialist?

English Instructions: READ CATEGORIES IF NECESSARY

Target: Both males and females 20 YEARS - 69 YEARS

Code or Value	Value Description	Count	Cumulative	Skip to Item
1	Less than a year ago	369	369	
2	1 year to 4 years ago	652	1021	
3	5 to 9 years ago	547	1568	
4	Ten or more years ago	1381	2949	
5	Never	1607	4556	
7	Refused	0	4556	
9	Don't know	119	4675	
.	Missing	4689	9364	

Data Cleaning

Reverse score where necessary, then add value labels

```
# Reverse score
NHclean$AUQ054 <- 7 - NHclean$AUQ054
# Add value labels
NHclean$AUQ054 <- labelled(NHclean$AUQ
                           c("Deaf" =
                             "A lot of
                             "Moderate
                             "A litttl
                             "Good" =
                             "Excellen
                             )
                           )
val_labels(NHclean$AUQ054)
```

Variable Name: AUQ054

SAS Label: General condition of hearing

English Text: These next questions are about {your/SP's} hearing. Which statement best describes {your/SP's} hearing (without a hearing aid or other listening devices)? Would you say {your/his/her} hearing is excellent, good, that {you have/s/he has} a little trouble, moderate trouble, a lot of trouble, or {are you/is s/he} deaf?

Target: Both males and females 1 YEARS - 150 YEARS

Code or Value	Value Description	Count	Cumulative	Skip to Item
1	Excellent	4244	4244	
2	Good	3744	7988	
3	A little trouble	869	8857	
4	Moderate hearing trouble	306	9163	
5	A lot of trouble	172	9335	
6	Deaf	12	9347	
77	Refused	1	9348	
99	Don't know	15	9363	
.	Missing	1	9364	

```
##                    Deaf         A lot of trouble Moderate hearing trouble
##                       1                        2                        3
##        A litttle trouble                     Good                Excellent
##                       4                        5                        6
```

Data Cleaning

Reverse score and add value labels to AUQ060

```
# Reverse score & set to binary
NHclean$AUQ060 <- 3 - NHclean$AUQ060
NHclean$AUQ060 <- NHclean$AUQ060 - 1
# Add value labels
NHclean$AUQ060 <- labelled(NHclean$AUQ
                           c("No" = 0,
                             "Yes" = 1
                             )
                           )
val_labels(NHclean$AUQ060)
```

```
##  No Yes
##   0   1
```

```
table(NHclean$AUQ060)
```

```
##
##    0    1
##  745 2128
```

Variable Name: AUQ060

SAS Label: Hear a whisper from across a quiet room?

English Text: These next questions refer to hearing without the use of a hearing aid or any other listening devices. If {you have/SP has} one ear that is better than the other, please answer the questions for the hearing in {your/SP's} better ear. Can {you/SP} usually hear and understand what a person says without seeing his or her face if that person whispers to {you/him/her} from across a quiet room?

Target: Both males and females 20 YEARS - 69 YEARS

Code or Value	Value Description	Count	Cumulative	Skip to Item
1	Yes	2128	2128	AUQ100
2	No	745	2873	
7	Refused	0	2873	
9	Don't know	32	2905	
.	Missing	6459	9364	

Data Cleaning

Reverse score and add value labels to AUQ100

```
# Reverse score
NHclean$AUQ100 <- 6 - NHclean$AUQ100
# Add value labels
NHclean$AUQ100 <- labelled(NHclean$AUQ
                           c("Never" =
                             "Seldom" 
                             "About ha
                             "Usually"
                             "Always" 
                             )
                           )
val_labels(NHclean$AUQ100)
```

Variable Name: AUQ100

SAS Label: Difficult follow conversation if noise

English Text: How often {do you/does SP} find it difficult to follow a conversation if there is background noise, for example, when other people are talking, TV or radio is on, or children are playing? Would you say...

English Instructions: HAND CARD AUQ1

Target: Both males and females 20 YEARS - 69 YEARS

Code or Value	Value Description	Count	Cumulative	Skip to Item
1	Always	173	173	
2	Usually	356	529	
3	About half the time	551	1080	
4	Seldom	1145	2225	
5	Never	2448	4673	
7	Refused	0	4673	
9	Don't know	2	4675	
.	Missing	4689	9364	

```
##                Never              Seldom About half the time
##                    1                   2                   3
##              Usually              Always
##                    4                   5
```

Data Cleaning

What about AUQ144?

```
# Add value labels
NHclean$AUQ144 <- labelled(NHclean$AUQ
                           c("Less tha
                             "1 year t
                             "5 to 9 y
                             "10 or mo
                             "Never" =
                             )
                           )
val_labels(NHclean$AUQ144)
```

Variable Name: AUQ144

SAS Label: Last time hearing tested by specialist?

English Text: A hearing test by a specialist is one that is done in a sound proof booth or room, or with headphones. Hearing specialists include audiologists, ear nose and throat doctors, and trained technicians or occupational nurses. When was the last time {you had/SP had} {your/his/her} hearing tested by a hearing specialist?

English Instructions: READ CATEGORIES IF NECESSARY

Target: Both males and females 20 YEARS - 69 YEARS

Code or Value	Value Description	Count	Cumulative	Skip to Item
1	Less than a year ago	369	369	
2	1 year to 4 years ago	652	1021	
3	5 to 9 years ago	547	1568	
4	Ten or more years ago	1381	2949	
5	Never	1607	4556	
7	Refused	0	4556	
9	Don't know	119	4675	
.	Missing	4689	9364	

```
##  Less than a year ago 1 year to 4 years a
##                     1                    2                    3
##  10 or more years ago                Never
##                     4                    5
```

Data Cleaning

Wrap-up

Save out your clean data

- Rdata version for your analysis
- csv version if you plan on sharing with others who don't use R

```
# R version
save(NHclean, file="NHANESclean20180102.Rdata")
# csv version
write.csv(NHclean, file="NHANESclean20180102.csv", row.names=FALSE)
```

Save and knit your .Rmd so you have a nice record of what you did to clean your data.

Data Documentation

Codebooks to guide our way

Data Documentation

Automated option 1: dataMaid package

```
# Load library
library(dataMaid)
# Does what it says on the tin
makeCodebook(NHclean, output="html", codebook=TRUE)
```

- Creates a source .Rmd file and resulting HTML codebook
- Easy
- Kinda ugly
- Doesn't give us everything we want

Data Documentation

Automated option 2: codebook package

- Fire up a new .Rmd file with HTML output
- Set up the YAML so we get a nice TOC:

```
---
title: "NHANES Automatic Codebook Example"
output:
  html_document:
    toc: true
    toc_float:
      smooth_scroll: true
---
```

Data Documentation

Automated option 2: codebook package

- Write the following chunk (set echo=FALSE, warning=FALSE):

```
knitr::opts_chunk$set(echo=FALSE, warning=FALSE) # Don't show any additional
# chunks or warnings
load("NHANESclean20180102.Rdata")
library(codebook) # load codebook package
# Create codebook
cb <- codebook(NHclean, survey_repetition="single")
cb
```

- Knit

Data Documentation

Custom XML Options

Automated codebooks are quick and easy, but sometimes they're ugly, don't quite give you what you want, or give you too much. Being automated, they don't give you many options.

If you *really* want to make custom codebooks that provide both human- and machine-readable formats and are comfortable with writing raw HTML-like code (plus maybe a little css if you want to get really fancy), consider writing an XML codebook.

Check out the XML Codebook Example developed by the Coding2Share team.

GitHub

Collaborate and share

GitHub

- Online, free repository
- Share data, code, and programs
- Collaborate with team members
- Track versions

Coding2Share GitHub Page

Thanks!

Bobbi Carothers

bcarothers@wustl.edu

https://github.com/coding2share/ReproducibilityToolkit