Overview of Data Cleaning and Analysis

Kimia Akhavein, Nicholas Harp, Wendy Huynh, Isabel Kees Kuebler

Overview

- Data Cleaning (in excel)
- Data importing (into R)
- Graphing (R)
- Analyses (R)
- Disseminating Data (GitHub & Open Science Framework)

Methods & Data Collection (osf.io/b2trn)

Participants rated words in a 2 alternative forced choice task (positive vs. negative ratings).

- Positive word: brave
- Negative word: crook
- Ambiguous word: break

Calculated percent negative ratings as percentage of trials rated as negative for each word type (positive, negative, ambiguous).

Data Cleaning

Data Cleaning

- Renaming Variables
- Duplicates
- Dummy-coding
- Missing Data

Variables

Remove variables you don't want/need, keep variables you do want/need, and rename variables in ways that make sense.

- Reduces the amount of information to retain
- Allows for ease of data sharing
- Small investment in the beginning = time saved in the future

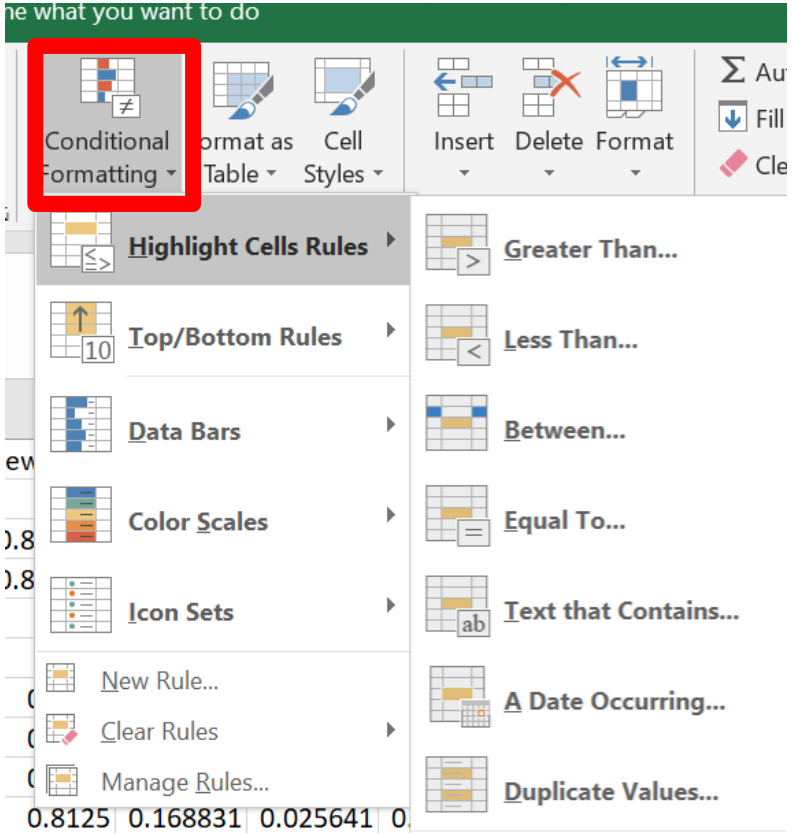
	A
1	Participant.Public.ID
2	1
3	2
4	3
5	4

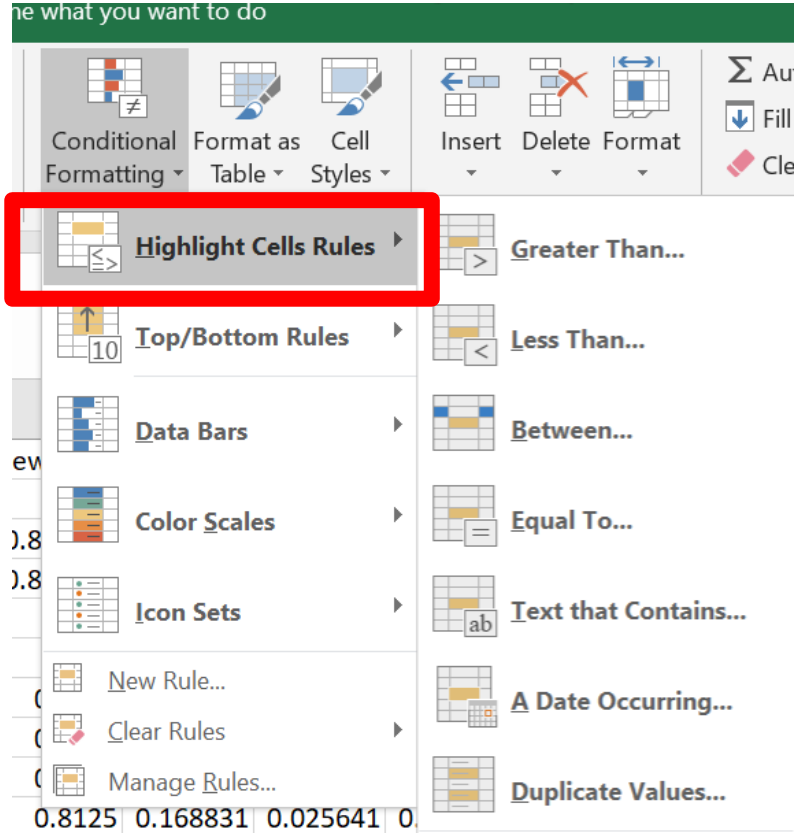
	A
1	ID
2	1
3	2
4	3
5	4

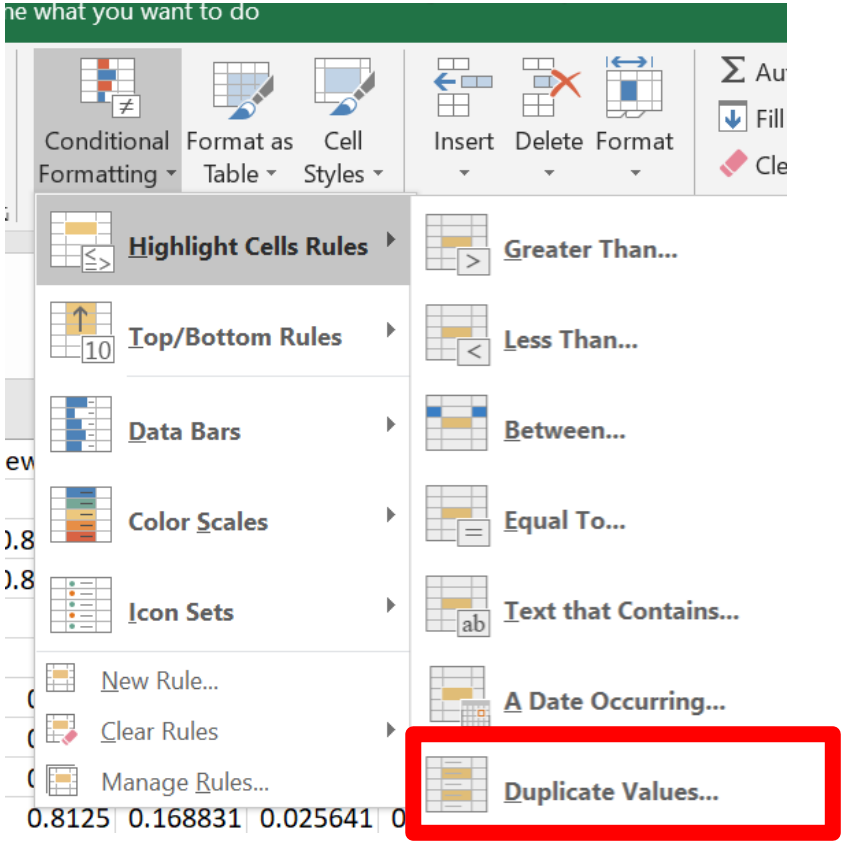
Duplicates

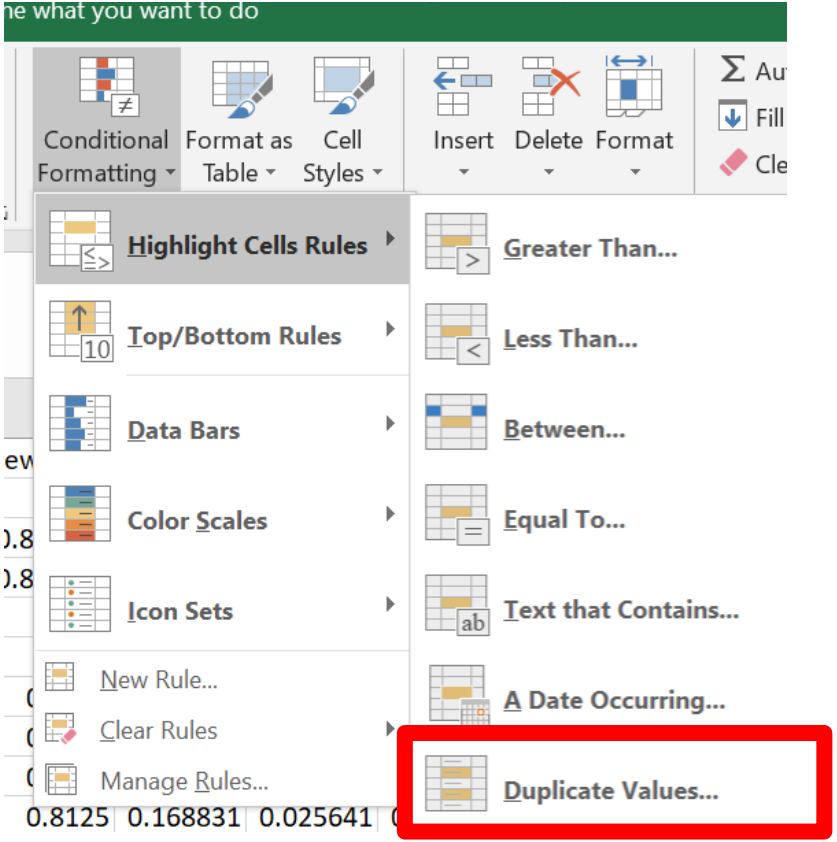
You should know exactly how many participants are supposed to be in your data set and ensure that the number stays constant.

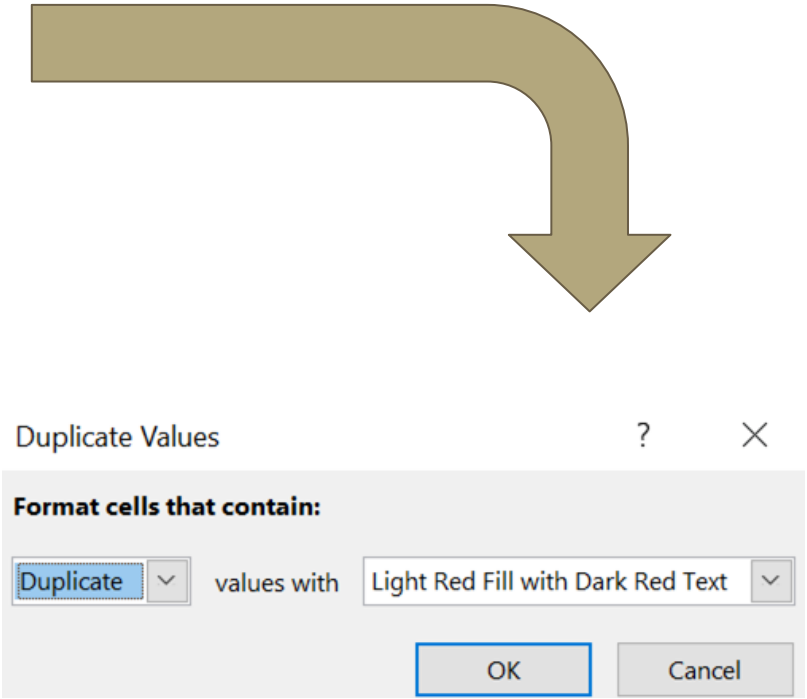
- Data entry error
- Participant could fill out data twice or more
- Tracking why participant or observation numbers may change


he what you want to do
Conditional Formatting
ormat as Table
Cell Styles
Insert
Delete
Format
Highlight Cells Rules
Top/Bottom Rules
Data Bars
Color Scales
Icon Sets
New Rule...
Clear Rules
Manage Rules...
Greater Than...
Less Than...
Between...
Equal To...
Text that Contains...
A Date Occurring...
Duplicate Values...
0.8125
0.168831
0.025641


he what you want to do
Conditional Formatting
Format as Table
Cell Styles
Insert
Delete
Format
Highlight Cells Rules
Top/Bottom Rules
Data Bars
Color Scales
Icon Sets
New Rule...
Clear Rules
Manage Rules...
Greater Than...
Less Than...
Between...
Equal To...
Text that Contains...
A Date Occurring...
Duplicate Values...
0.8125
0.168831
0.025641


he what you want to do
Conditional Formatting
Format as Table
Cell Styles
Insert
Delete
Format
Au
Fill
Cle
Highlight Cells Rules
Top/Bottom Rules
Data Bars
Color Scales
Icon Sets
New Rule...
Clear Rules
Manage Rules...
Greater Than...
Less Than...
Between...
Equal To...
Text that Contains...
A Date Occurring...
Duplicate Values...
0.8125 0.168831 0.025641


he what you want to do
Conditional Formatting
Format as Table
Cell Styles
Insert
Delete
Format
Highlight Cells Rules
Top/Bottom Rules
Data Bars
Color Scales
Icon Sets
New Rule...
Clear Rules
Manage Rules...
Greater Than...
Less Than...
Between...
Equal To...
Text that Contains...
A Date Occurring...
Duplicate Values...
0.8125
0.168831
0.025641


Duplicate Values
Format cells that contain:
Duplicate
values with
Light Red Fill with Dark Red Text
OK
Cancel

2	10
3	11
4	12
5	13
6	14
7	15
8	15
9	16
0	17
1	18
2	19
3	20
4	21

12	10
13	11
14	12
15	13
16	14
17	15
18	16
19	17
20	18
21	19
22	20
23	21

Dummy coding

Categorical variables are common, but are not always useable in their current state.

- Common examples: sex, race/ethnicity, treatment/control group
- Coded as 0/1
- Split dummy variables into equal number of categories that the variable contains

=IF(CELL=“Female”,1,0)

=IF(AG2=“Female”,1,0)

AG
sex
Female
Female
Female
Female
Male
Female
Female
Male
Female

AG	AJ
sex	female
Female	1
Female	1
Female	1
Female	1
Male	0
Female	1
Female	1
Male	0
Female	1

AI	AK	AL	AM
race	Hispanic or Latino	White - not of Hisp	Asian
Hispanic or Latino	1	0	0
White - not of Hispanic Origin	0	1	0
White - not of Hispanic Origin	0	1	0
Asian	0	0	1
White - not of Hispanic Origin	0	1	0
White - not of Hispanic Origin	0	1	0
Hispanic or Latino	1	0	0
White - not of Hispanic Origin	0	1	0
White - not of Hispanic Origin	0	1	0
White - not of Hispanic Origin	0	1	0
Asian	0	0	1
White - not of Hispanic Origin	0	1	0
White - not of Hispanic Origin	0	1	0
White - not of Hispanic Origin	0	1	0
Asian	0	0	1

AI	AK	AL	AM
race	Hispanic or Latino	White - not of Hisp	Asian
Hispanic or Latino	1	0	0
White - not of Hispanic Origin	0	1	0
White - not of Hispanic Origin	0	1	0
Asian	0	0	1
White - not of Hispanic Origin	0	1	0
White - not of Hispanic Origin	0	1	0
Hispanic or Latino	1	0	0
White - not of Hispanic Origin	0	1	0
White - not of Hispanic Origin	0	1	0
White - not of Hispanic Origin	0	1	0
Asian	0	0	1
White - not of Hispanic Origin	0	1	0
White - not of Hispanic Origin	0	1	0
White - not of Hispanic Origin	0	1	0
Asian	0	0	1

AI	AK	AL	AM
race	Hispanic or Latino	White - not of Hisp	Asian
Hispanic or Latino	1	0	0
White - not of Hispanic Origin	0	1	0
White - not of Hispanic Origin	0	1	0
Asian	0	0	1
White - not of Hispanic Origin	0	1	0
White - not of Hispanic Origin	0	1	0
Hispanic or Latino	1	0	0
White - not of Hispanic Origin	0	1	0
White - not of Hispanic Origin	0	1	0
White - not of Hispanic Origin	0	1	0
Asian	0	0	1
White - not of Hispanic Origin	0	1	0
White - not of Hispanic Origin	0	1	0
White - not of Hispanic Origin	0	1	0
Asian	0	0	1

AI	AK	AL	AM
race	Hispanic or Latino	White - not of Hisp	Asian
Hispanic or Latino	1	0	0
White - not of Hispanic Origin	0	1	0
White - not of Hispanic Origin	0	1	0
Asian	0	0	1
White - not of Hispanic Origin	0	1	0
White - not of Hispanic Origin	0	1	0
Hispanic or Latino	1	0	0
White - not of Hispanic Origin	0	1	0
White - not of Hispanic Origin	0	1	0
White - not of Hispanic Origin	0	1	0
Asian	0	0	1
White - not of Hispanic Origin	0	1	0
White - not of Hispanic Origin	0	1	0
White - not of Hispanic Origin	0	1	0
Asian	0	0	1

AI	AK	AL	AM
race	Hispanic or Latino	White - not of Hisp	Asian
Hispanic or Latino	1	0	0
White - not of Hispanic Origin	0	1	0
White - not of Hispanic Origin	0	1	0
Asian	0	0	1
White - not of Hispanic Origin	0	1	0
White - not of Hispanic Origin	0	1	0
Hispanic or Latino	1	0	0
White - not of Hispanic Origin	0	1	0
White - not of Hispanic Origin	0	1	0
White - not of Hispanic Origin	0	1	0
Asian	0	0	1
White - not of Hispanic Origin	0	1	0
White - not of Hispanic Origin	0	1	0
White - not of Hispanic Origin	0	1	0
Asian	0	0	1

AI	AK	AL	AM
race	Hispanic or Latino	White - not of Hisp	Asian
Hispanic or Latino	1	0	0
White - not of Hispanic Origin	0	1	0
White - not of Hispanic Origin	0	1	0
Asian	0	0	1
White - not of Hispanic Origin	0	1	0
White - not of Hispanic Origin	0	1	0
Hispanic or Latino	1	0	0
White - not of Hispanic Origin	0	1	0
White - not of Hispanic Origin	0	1	0
White - not of Hispanic Origin	0	1	0
Asian	0	0	1
White - not of Hispanic Origin	0	1	0
White - not of Hispanic Origin	0	1	0
White - not of Hispanic Origin	0	1	0
Asian	0	0	1

AI	AK	AL	AM
race	Hispanic or Latino	White - not of Hisp	Asian
Hispanic or Latino	1	0	0
White - not of Hispanic Origin	0	1	0
White - not of Hispanic Origin	0	1	0
Asian	0	0	1
White - not of Hispanic Origin	0	1	0
White - not of Hispanic Origin	0	1	0
Hispanic or Latino	1	0	0
White - not of Hispanic Origin	0	1	0
White - not of Hispanic Origin	0	1	0
White - not of Hispanic Origin	0	1	0
Asian	0	0	1
White - not of Hispanic Origin	0	1	0
White - not of Hispanic Origin	0	1	0
White - not of Hispanic Origin	0	1	0
Asian	0	0	1

Missing Data

There are many ways to treat missing data. Pick the one that best suits your data.

- Deletion Methods: Listwise or Pairwise deletion
- Imputation Methods

Graphing in R

Preparing Data

- Prepare your data FIRST
  - Data cleaning
  - Missing values
  - Outliers
  - Transformations
  - Generating summary values (relevant means, standard error, etc.)

Generate Summary Data

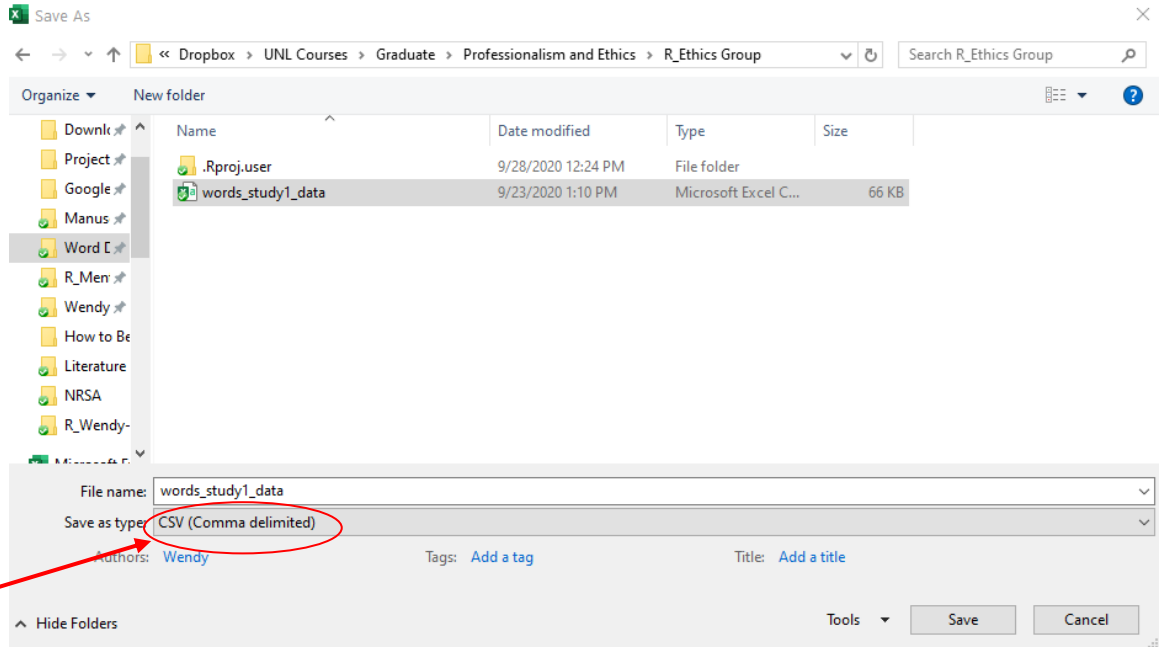
Determine what you want to graph and prepare those values in Excel.

ParticipantID	Ambiguous	Negative	Positive
1	0.531	1	0.0625
2	0.452	0.857	0
3	0.469	1	0

Valence	Mean	StdErr
Ambiguous	0.438	0.0115
Negative	0.927	0.00632
Positive	0.0475	0.00511

Dataset

- Save as **.csv** format
  - words_study1_data.csv



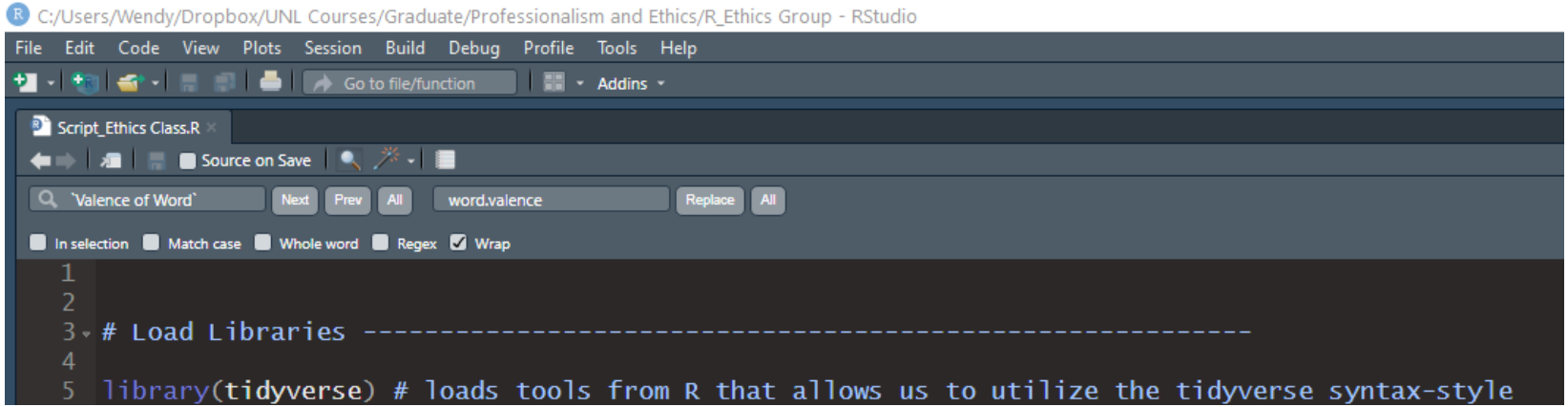
Setting up R/RStudio

- Download
  - R Statistics
  - R Studio
- Create
  - R Project
  - R script
- Install
  - R Packages (i.e., tidyverse)

- https://bookdown.org/yih_huynh/Guide-to-R-Book/getting-started.html

Loading Packages

- R packages contain tools that help programmers with specific tasks



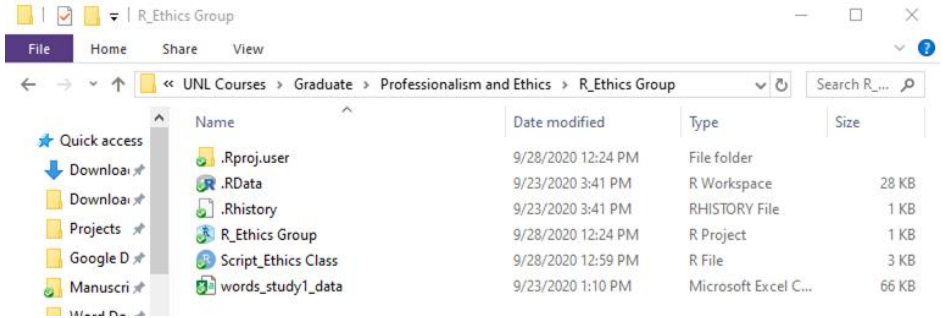
Working Directory

- A working directory tells R where to pull files from (file location/address)

```
# Check Working Directory ---------------------------------------------------

getwd() # retrieve information about the current working directory
```

C:/Users/Wendy/Dropbox/UNL Courses/Graduate/Professionalism and Ethics/R_Ethics Group



Import Dataset

- Import the dataset using the function: **read_csv()**
- Save the dataset as an R Object under the name **“data”**

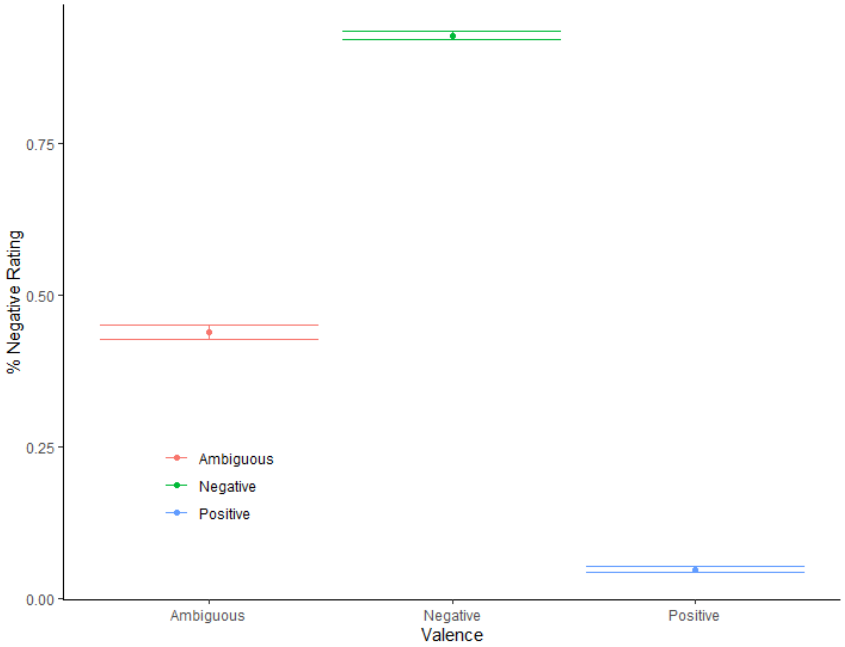
```
# Import Dataset ------------------------------------------------------------

data <- read_csv("words_study1_data.csv")
```

Graphing

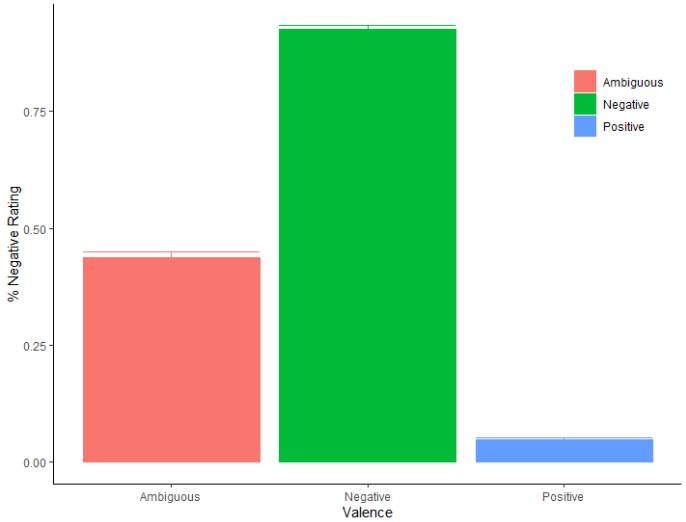
Using the ggplot2 package, we add each graph element.

```
data %>%
  ggplot(aes(x = Valence, y = Mean, color = Valence)) +
  geom_point() +
  geom_errorbar(aes(ymin = Mean - StdErr,
                    ymax = Mean + StdErr)) +
  theme_classic() +
  labs(x = "Valence",
       y = "% Negative Rating") +
  theme(legend.title = element_blank(),
        legend.position = c(.2, .2))
```

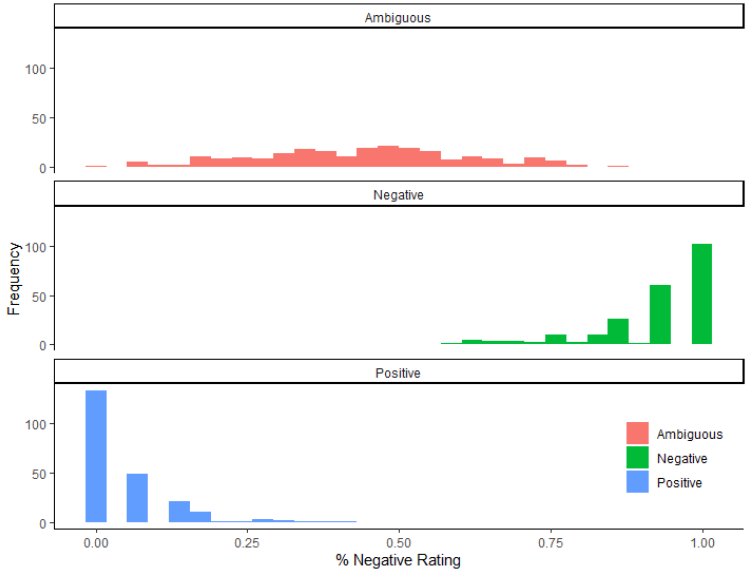


```r
data %>%
  ggplot(aes(x = Valence, y = m, color = Valence, fill = Valence)) +
  geom_bar(stat = "identity") +
  geom_errorbar(aes(ymin = m - s,
                    ymax = m + s)) +
  theme_classic() +
  labs(x = "Valence",
       y = "% Negative Rating") +
  theme(legend.title = element_blank(),
        legend.position = c(.9, .8))
```

Using a **Bar Graph**

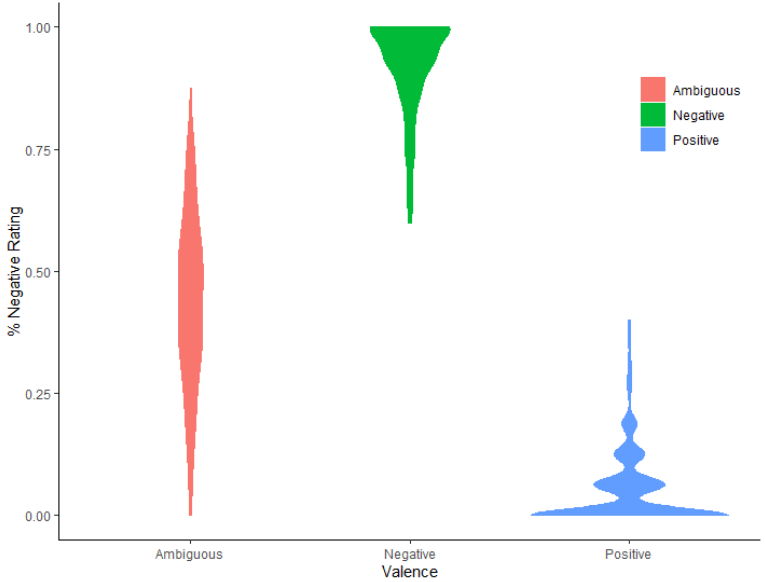


```r
data %>%
  ggplot(aes(x = percent.negative.rating, fill = Valence)) +
  geom_histogram() +
  theme_classic() +
  labs(x = "% Negative Rating",
       y = "Frequency") +
  theme(legend.title = element_blank(),
        legend.position = c(.9, .15))+
  facet_wrap(~Valence, ncol = 1)
```



```r
data %>%
  ggplot(aes(x = Valence, y = percent.negative.rating, fill = Valence)) +
  geom_violin() +
  theme_classic() +
  labs(x = "Valence",
       y = "% Negative Rating") +
  theme(legend.title = element_blank(),
        legend.position = c(.9, .8))
```

Violin Plot



Analysis in R

Analysis

- One-way ANOVA on negative rate for words
- Prepare data with only relevant columns
  - Participant.Public.ID
  - amw_rate
  - pow_rate
  - new_rate

Import Dataset

```
#Import words dataset
#make the dataset an object
  words <- (read.csv("words_neg_rate_study1_data.csv")
```

1.R* | wordsrt.long | wordsrt | words | words.long

Filter

	Participant.Public.ID	amw_rate	pow_rate	new_rate
1	1	0.53125000	0.06250000	1.0000000
2	2	0.45161290	0.00000000	0.8571429
3	3	0.46875000	0.00000000	1.0000000
4	4	0.54838710	0.00000000	1.0000000
5	5	0.32258064	0.12500000	0.9375000

Arrange dataset for ANOVA

```
words.long <- gather(words,
                     Condition,
                     PerNeg,
                     amw_rate,
                     pow_rate,
                     new_rate)
```

: 1.R* × | wordsrt.long × | wordsrt × | words × | words.long ×

Filter

	Participant.Public.ID	Condition	PerNeg
1	1	amw_rate	0.53125000
228	1	pow_rate	0.06250000
455	1	new_rate	1.00000000
2	2	amw_rate	0.45161290
229	2	pow_rate	0.00000000
456	2	new_rate	0.85714286

Perform one-way ANOVA

```
words.results <-aov(PerNeg~
                      Condition +
                      Error(Participant.Public.ID/
                              Condition),
                    data=words.long)
summary(words.results)
```

```
Error: Within
           Df   Sum Sq Mean Sq F value   Pr(>F)
Condition   2   429489  214744   9.615 7.65e-05 ***
Residuals 666 14874569   22334
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Post-hoc test

```
emmeans(words.results, pairwise~Condition,
        adjust="Tukey")
```

Condition	emmean	SE	df	lower.CL	upper.CL
amw_rate	0.4384	0.0142	123	0.40408	0.4727
new_rate	0.9317	0.0142	123	0.89741	0.9660
pow_rate	0.0427	0.0142	123	0.00844	0.0771

contrast	estimate	SE	df	t.ratio	p.value
amw_rate - new_rate	-0.493	0.0235	666	-20.982	<.0001
amw_rate - pow_rate	0.396	0.0235	666	16.827	<.0001
new_rate - pow_rate	0.889	0.0235	666	37.808	<.0001

Disseminating Data
(GitHub & Open Science Framework)

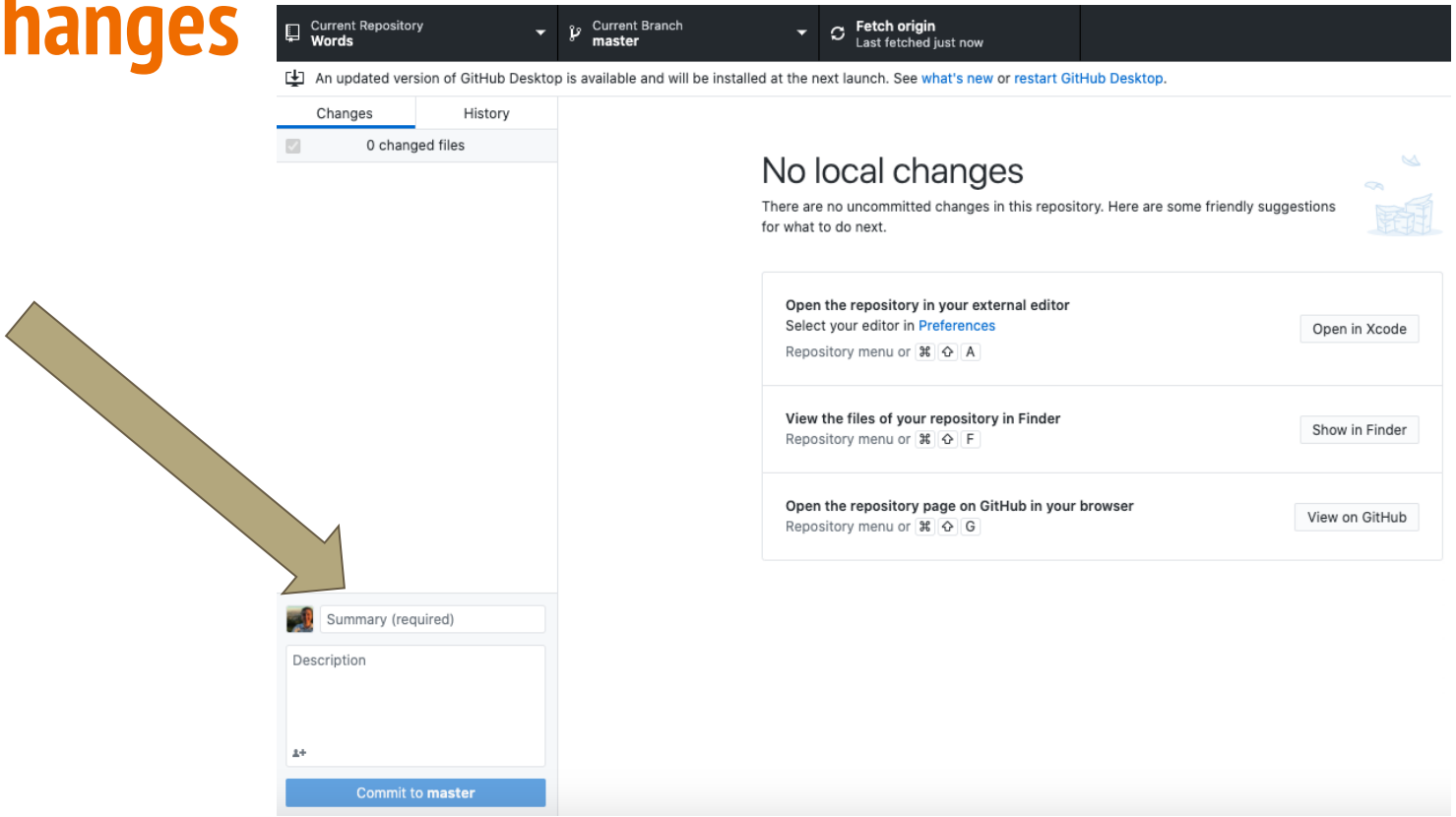
Git(Hub)

- https://git-scm.com/downloads

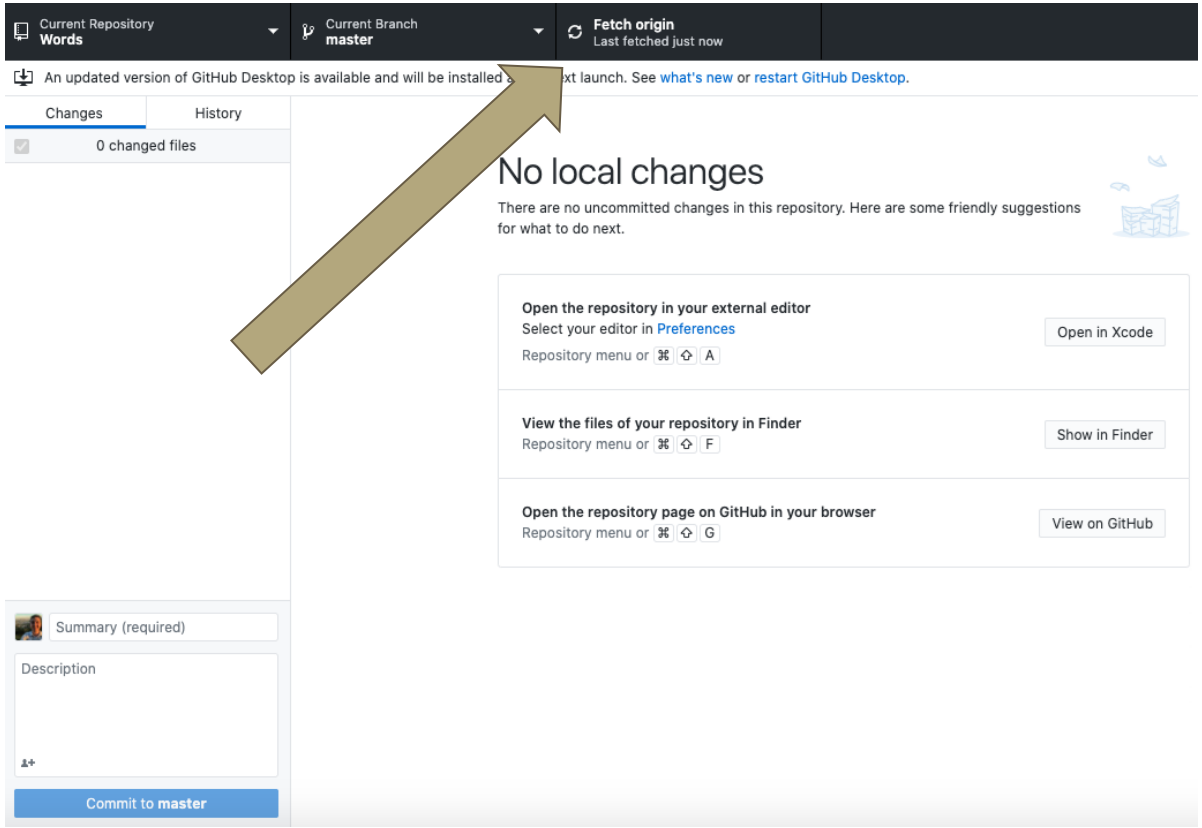
GitHub & Other GUIs

- GitHub: https://desktop.github.com/
- GitKraken: https://www.gitkraken.com/
- Many other options available @ https://git-scm.com/downloads/guis

Commit changes



Push changes



Changes | History

No Branches to Compare

Merge branch 'master' of github.com:nharp189/Words

SPPS R1
Nicholas Harp • 5d

SPPS R1 Updates
Nicholas Harp • Aug 14, 2020

Revisions R1 SPPS
Nicholas Harp • Jul 23, 2020

updates
Nicholas Harp • Jun 24, 2020

Do ANOVA correcty
Nicholas Harp • May 5, 2020

Making csv's for Maital to explore...
Nicholas Harp • May 5, 2020

Merge branch 'master' of github....
Nicholas Harp • Apr 27, 2020

My last changes... about to be ov...
Nicholas Harp • Apr 27, 2020

sync
Catie Brown • Apr 19, 2020

Update WordsManuscript.docx
Nicholas Harp • Mar 24, 2020

updated ms
Nicholas Harp • Mar 16, 2020

starting supplemental analyses
Nicholas Harp • Feb 26, 2020

SPPS R1 Updates

Nicholas Harp 55aaf9b 30 changed files Hide Whitespace

- .DS_Store
- Final_MS_Analyses.R
- README.rtf
- RT.png
- WordsOutline.docx
- audio.html
- data/study1_redo_data/RT_Matrix_2020.07.28.csv
- data/study1_redo_data/RT_Matrix_2020.08.07.csv
- data/study1_redo_data/RT_Matrix_2020.08.10.csv
- data/study1_redo_data/Resp_Matrix_2020.07.28.csv
- data/study1_redo_data/Resp_Matrix_2020.08.07.csv
- data/study1_redo_data/Resp_Matrix_2020.08.10.csv
- data/study1_red.../words_study1_data_2020.07.28....
- data/study1_red.../words_study1_data_2020.08.07....
- data/study1_red.../words_study1_data_2020.08.10....
- data_cleaning_nh.R
- data_cleaning_nh_2020.07.21.R
- WordsManuscript.R... → .../WordsManuscript.R...
- WordsManuscript.docx → old_files/...cript.docx
- WordsManuscript.log → o.../WordsManuscript.log

```
@@ -5,6 +5,8 @@ setwd(path)
 ### Pilot Study ###
 {library(emmeans)
 library(papaja)
+library(car)
+library(ppcor)
 library(readr)
 library(tidyverse)
 library(utils)
@@ -197,7 +199,12 @@ sd(subset(final.words, final.words$wordlist %in% pos$wordlist | final.words$word
 ############################ Methods ##########################
 ## Demographic Questionnaire and Screener Questions
 #run study1_redo_data_cleaning.R to line 384 and then make v2_data.summary skipping some lines..
-full <- read_csv("data/study1_redo_data/words_study1_data_persubj2Trim.csv")
+full <- read_csv("data/study1_redo_data/words_study1_data_2020.08.10.csv")
+# temp <- combined %>% subset(Participant.Public.ID %in% full$Participant.Public.ID)
+# plyr::count(temp$perRetainedof160)
+# temp2 <- temp %>% subset(perRetainedof160 >= .95)
+sd(temp$perRetainedof160)
+
 full$age <- as.numeric(full$age)

 ### count sex ###
@@ -215,8 +222,8 @@ plyr::count(full$race)
 ### Manipulation Check ###
 ### Response Matrix ###
```

Open Science Framework (osf.io)

OSF HOME ▾ Search Support Donate Sign Up Sign In

Spring break or heart break? Extending ... Files Wiki Analytics Registrations

Public 0 ...

Spring break or heart break? Extending valence bias to emotional words

Contributors: Nicholas Harp, Catie Brown, Maital Neta

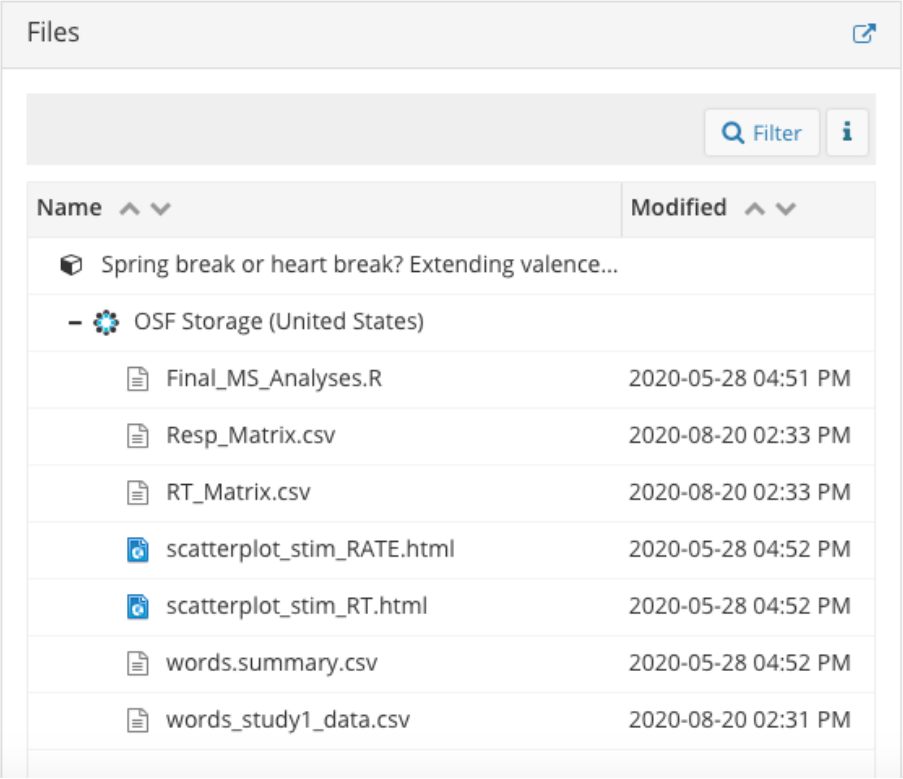
Date created: 2019-10-07 02:17 PM | Last Updated: 2020-08-20 02:33 PM

Category: Project

Has supplemental materials for Spring break or heart break? Extending valence bias to emotional words on PsyArXiv

Open Science Framework (osf.io)

- Store files
  - Analysis scripts
  - Data files
  - Others (e.g., html)



Resources

- Learning Statistics with R (https://learningstatisticswithr.com/)
- R for Graduate Students by Wendy (https://bookdown.org/yih_huynh/Guide-to-R-Book/)
- R for Data Science (https://r4ds.had.co.nz/index.html)
- SPSS Programming and Data Management, 3rd Edition - Raynald Levesque
- A Gentle Introduction to STATA, 6th Edition - Alan C. Acock

Questions & Comments?

Lab Manuals

- Helpful for keeping variables and scoring all in one place for studies

- Manuals can include, but not limited, to:
  - Description of survey or measure
  - Variables & Variable Names
  - Scoring
  - Reference/Citation for survey or measure

ANOVA Example 2

Compare means of response time

```
wordsrt <- (read.csv
            ("words_response_time_study1_data.csv"))

wordsrt.long <- gather(wordsrt,
                       Condition,
                       RTime,
                       amw_rt,
                       pow_rt,
                       new_rt)

wordsrt.results <-aov(RTime~
                       Condition +
                       Error(Participant.Public.ID/
                               Condition),
                     data=wordsrt.long)
summary(wordsrt.results)
```

Post-hoc test

```
emmeans(wordsrt.results, pairwise~Condition,
        adjust="Tukey")
```

Condition	emmean	SE	df	lower.CL	upper.CL
amw_rt	822	22.6	4.63	739	904
new_rt	733	22.6	4.63	651	816
pow_rt	700	22.6	4.63	618	783

contrast	estimate	SE	df	t.ratio	p.value
amw_rt - new_rt	88.3	28.6	666	3.084	0.0060
amw_rt - pow_rt	121.5	28.6	666	4.242	0.0001
new_rt - pow_rt	33.2	28.6	666	1.158	0.4787

Thank you!