Managing Data Analysis in RStudio using project-oriented workflow

Andriy Koval, Ph.D.

Health System Impact Fellowship

3rd Annual Training Retreat

2019-11-26-Tuesday

Toronto, Ontario

UNIVERSITY OF CENTRAL FLORIDA

CIHR IRSC

Canadian Institutes of Health Research | Instituts de recherche en santé du Canada

https://github.com/andkov/hsif-2019-data-analysis

About me

- Ph.D. in Quantitative Methods, Psychology (2014)
- Reproducible research enthusiast since 2012
- Graph maker
- See work at https://github.com/andkov
- These slides and more at http://andriy.rbind.io

MIDDLE TENNESSEE STATE UNIVERSITY

The UNIVERSITY of OKLAHOMA

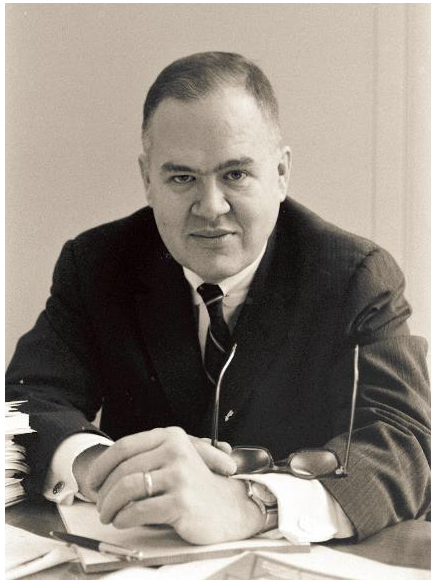
VANDERBILT

University of Victoria

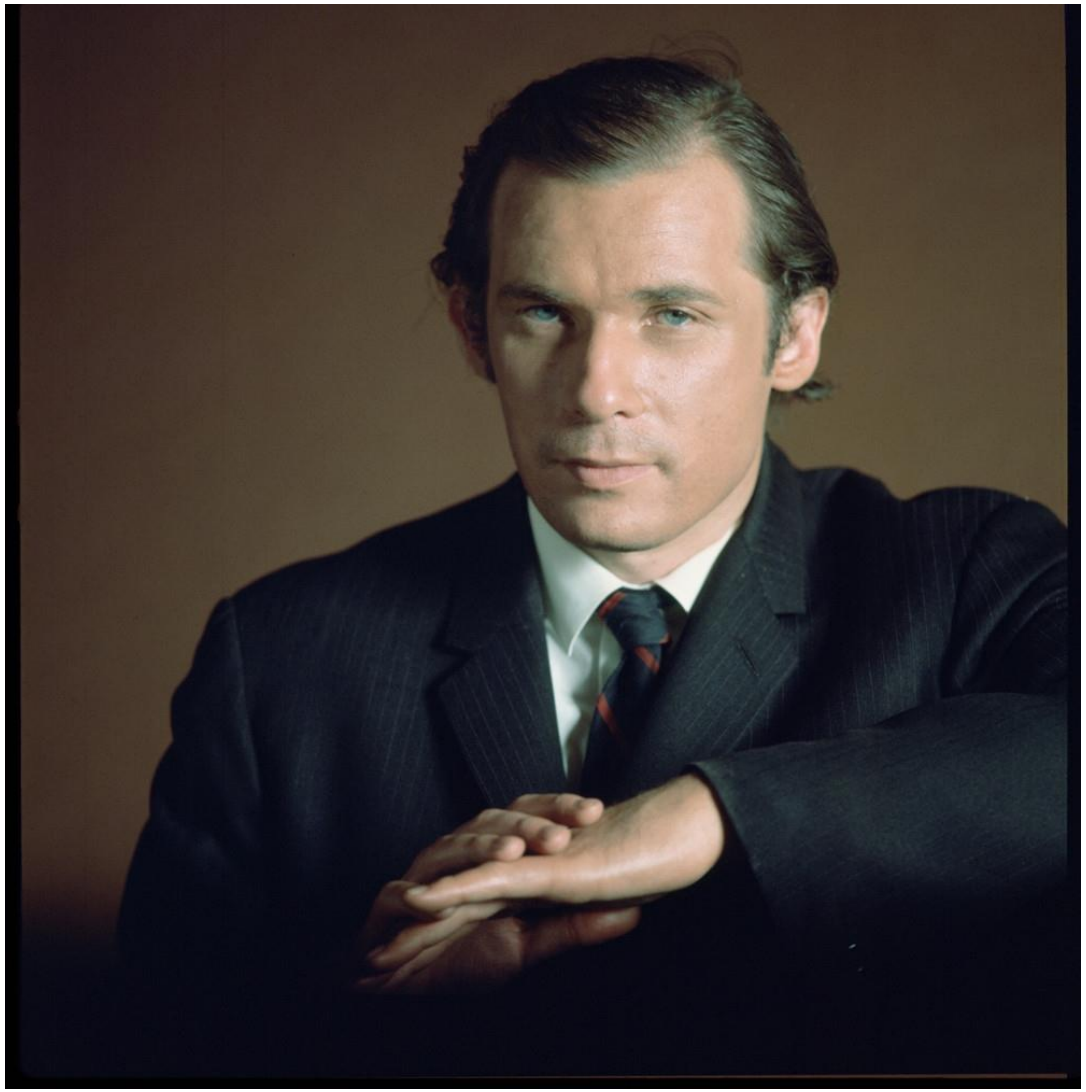
THE UNIVERSITY OF BRITISH COLUMBIA

UNIVERSITY OF CENTRAL FLORIDA

Andreas Vesalius



John Tukey



Glenn Gould

Hadley Wickham

Edward Tufte

Dialects of data expression

Tabular

id	time	attend	model
1	0	1	2.788
1	1	6	2.732
1	2	2	2.675
1	3	1	2.618
1	4	1	2.562
1	5	1	2.505
1	6	1	2.449
1	7	1	2.392
1	8	1	2.335
1	9	1	2.279
1	10	1	2.222
1	11	1	2.166
4	0	2	2.788
4	1	1	2.732

Algebraic

$$y_{it} = \beta_0 + \beta_1 time_t + \varepsilon_{it}$$

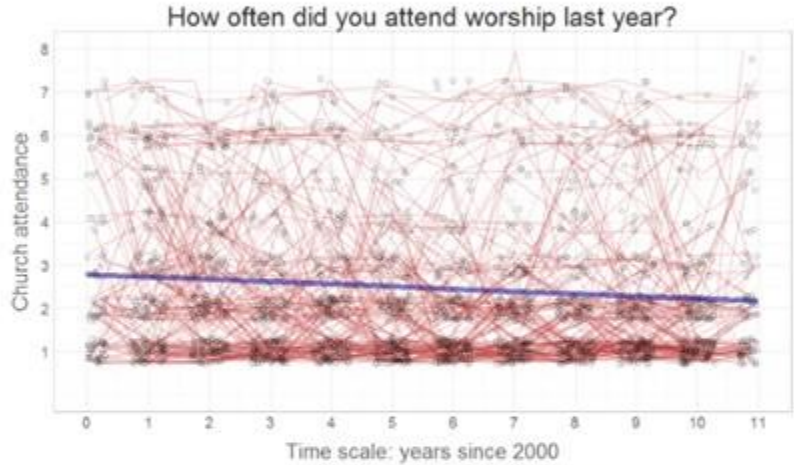
$$\beta_0 = \gamma_{00}$$

$$\beta_1 = \gamma_{10}$$

Semantic

In 2000 respondents attended church less than once a month (2.79) and gradually declined in their attendance since (.06 per year).

Graphical



Syntactic

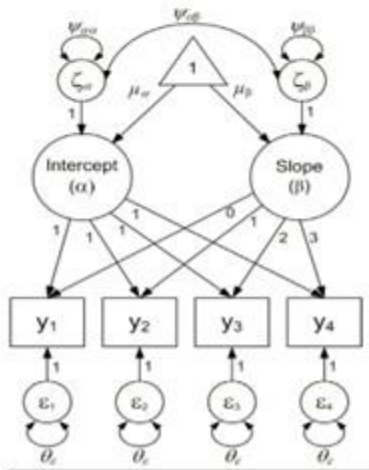
```
nlme::gls(attend ~ 1 + time, data=dsM)
```

Numeric

```
Coefficients:
               Value Std.Error t-value p-value
(Intercept)  2.7882   0.07774   35.86       0
time        -0.0566   0.01197   -4.73       0
```

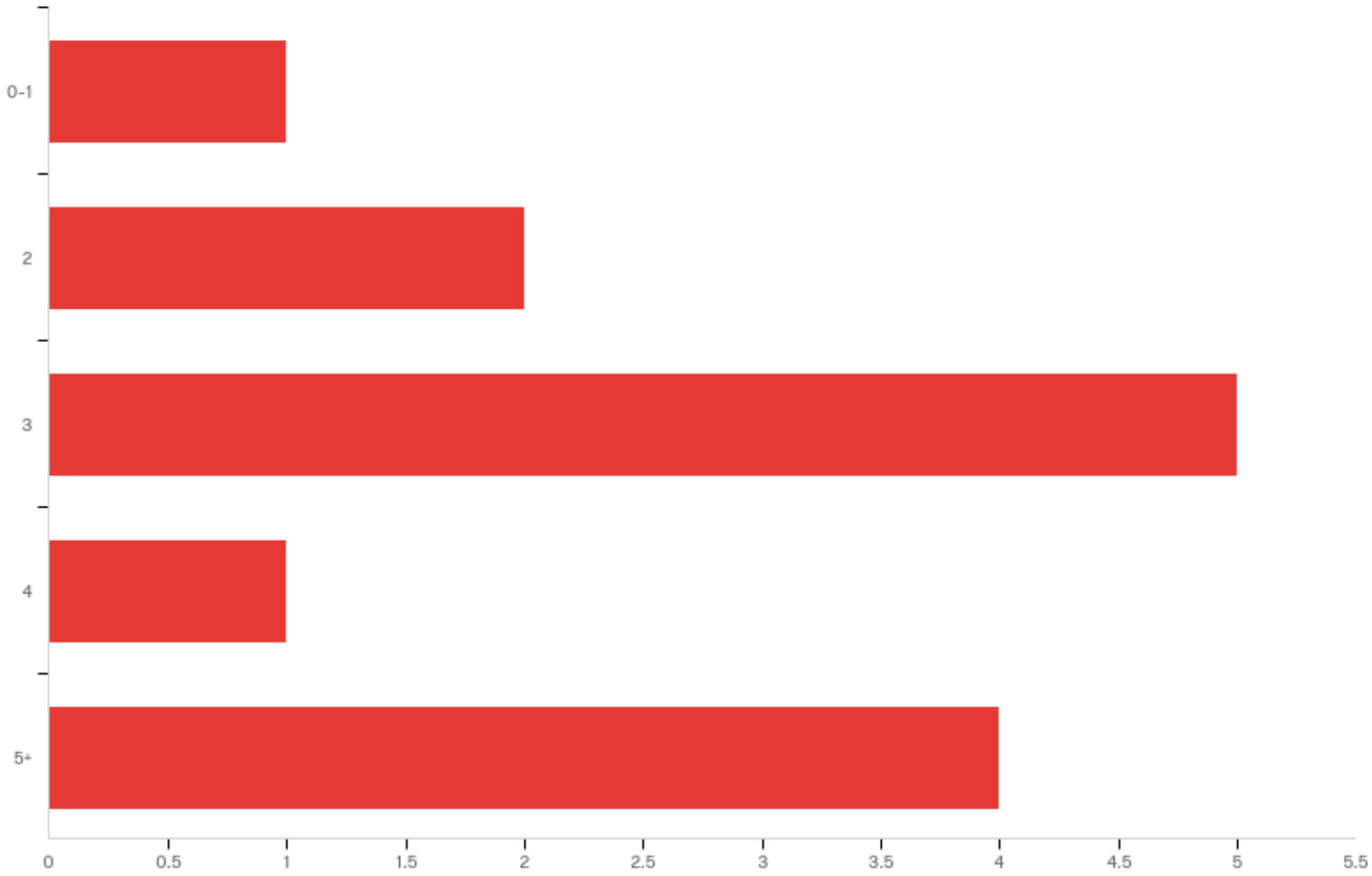
	modelB
logLik	-3719
deviance	7438
AIC	7444
BIC	7461
df.resid	1858
N	1860
p	2
ids	155

Schematic

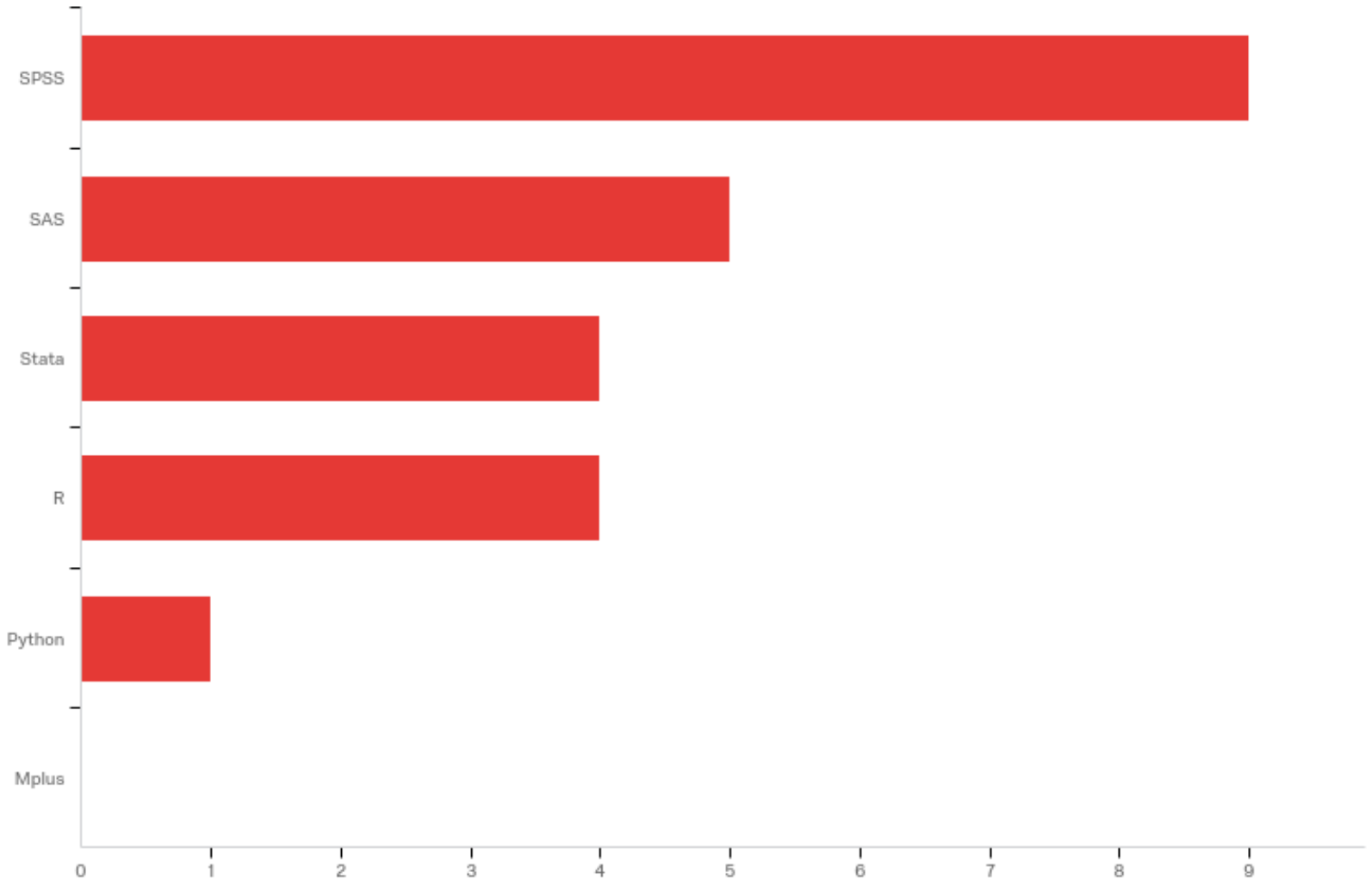


About you

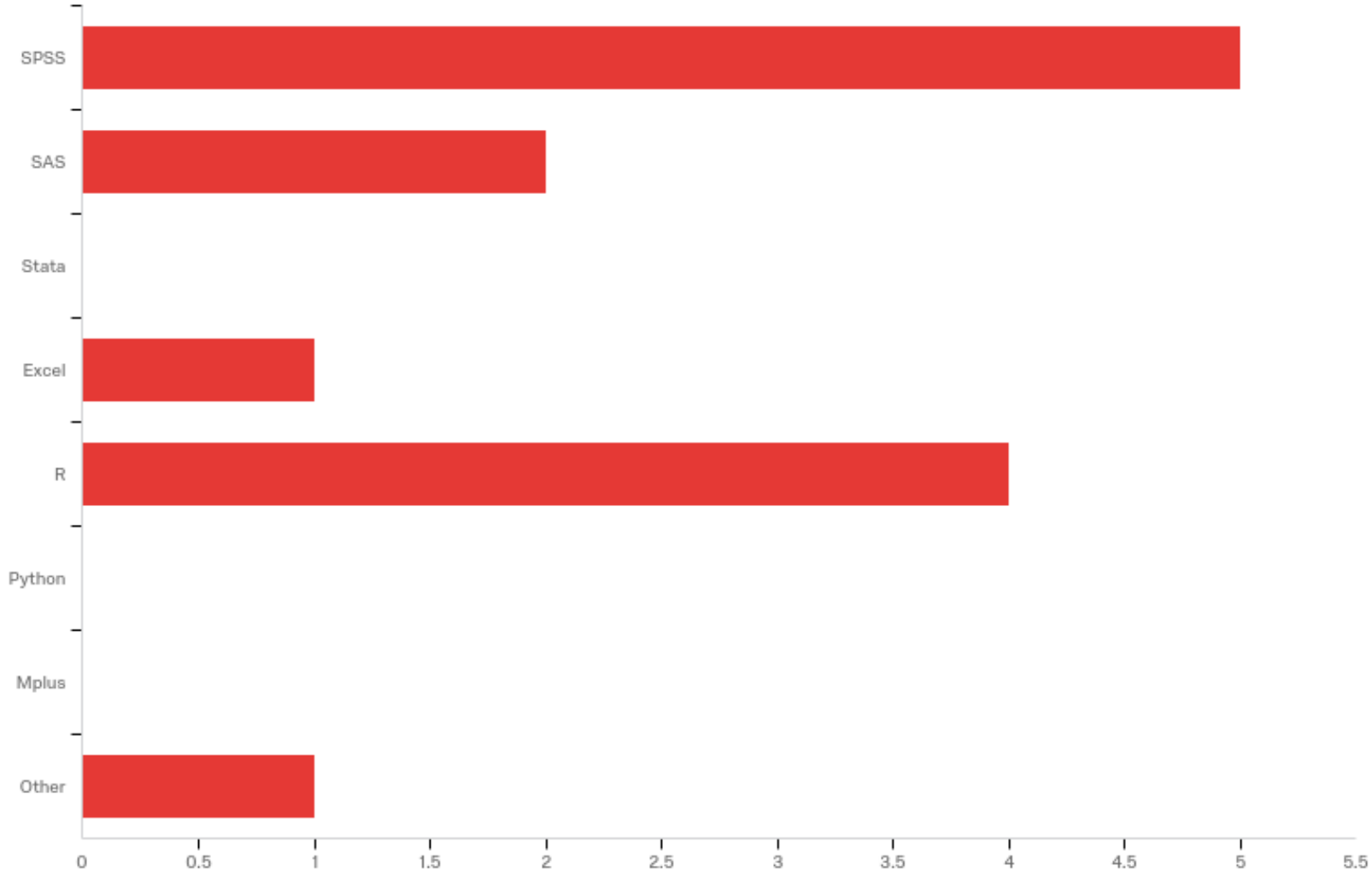
Q1.1 - How many undergraduate or graduate courses in statistics (or related field) have you taken so far?



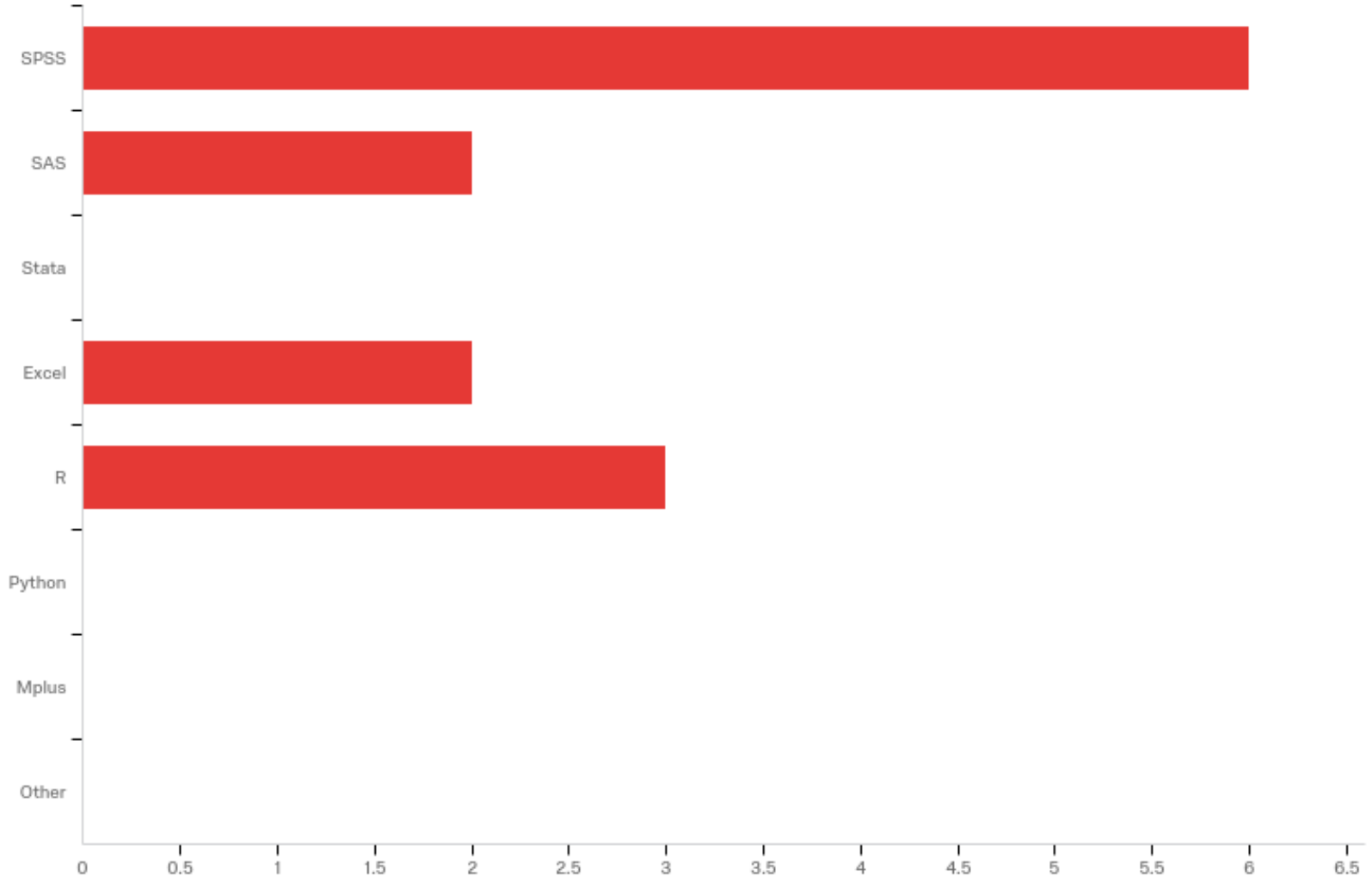
Q1.2 - What statistical software have you used AT LEAST ONCE in the last 3 years? (check all that apply)



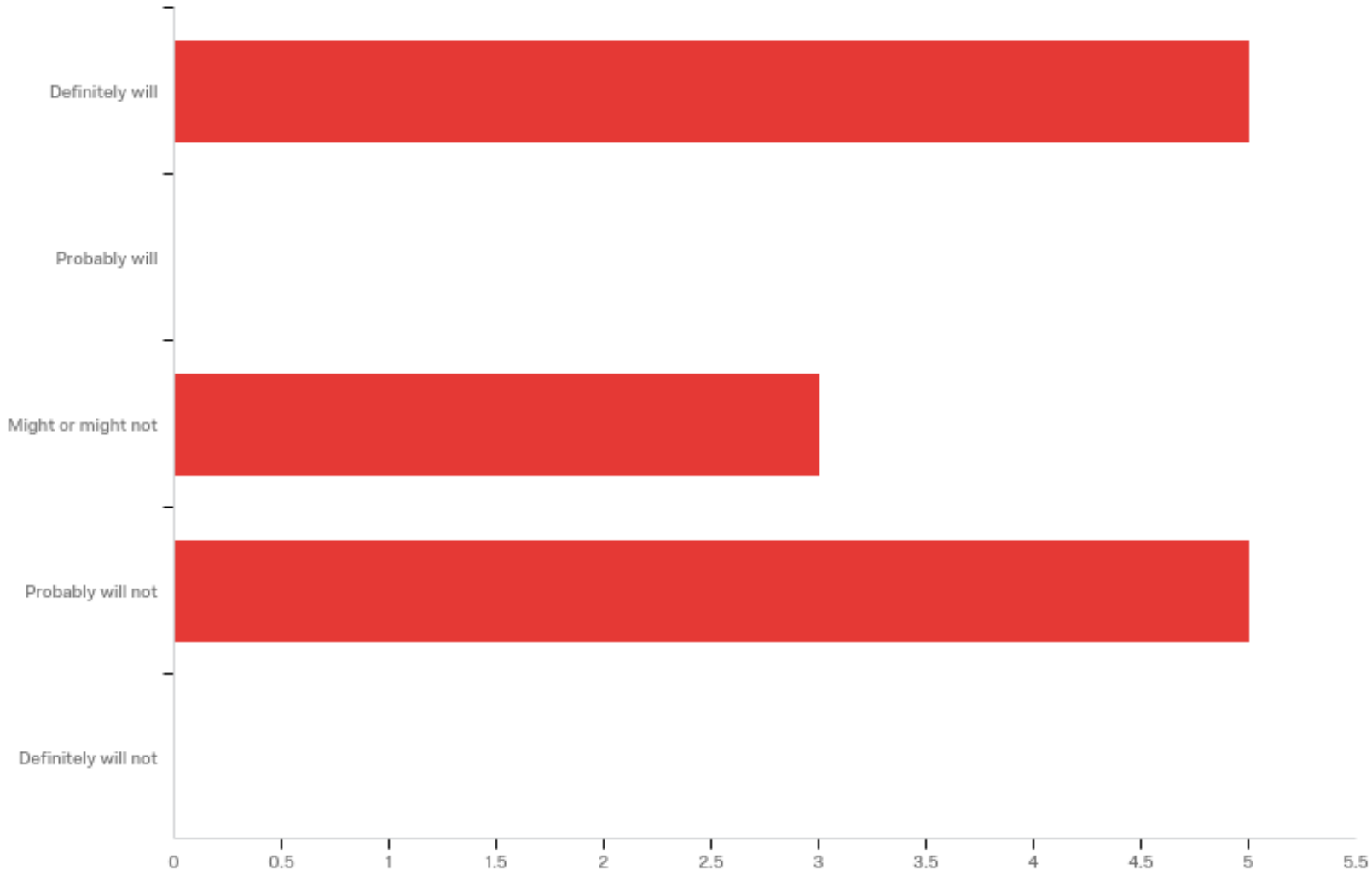
Q1.3 - You are asked to RUN A REGRESSION on an analysis-ready data set. With what software would you be most comfortable performing this task?



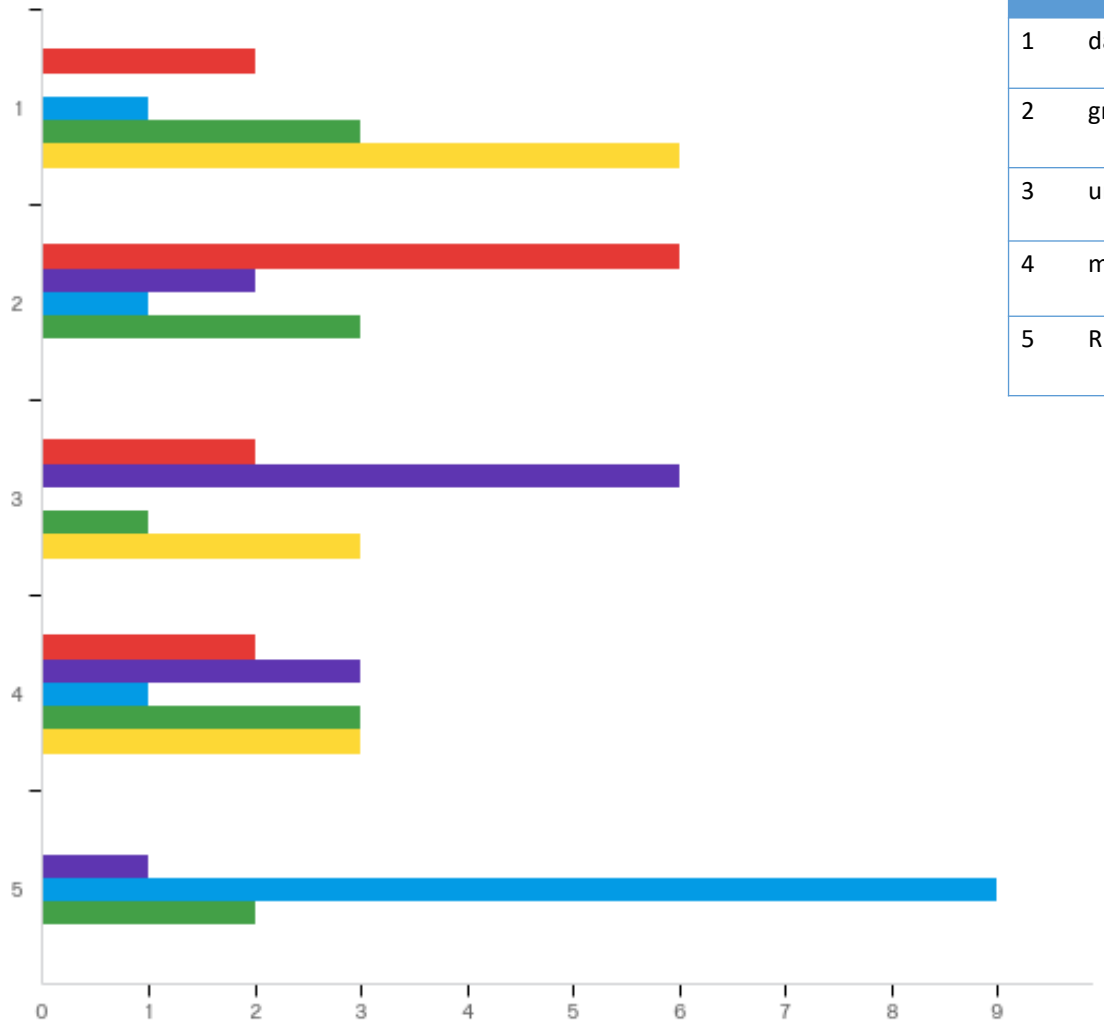
Q1.4 - You are asked to MAKE A GRAPH of a frequency distribution using an analysis-ready data set. With what software would you be most comfortable performing this task?



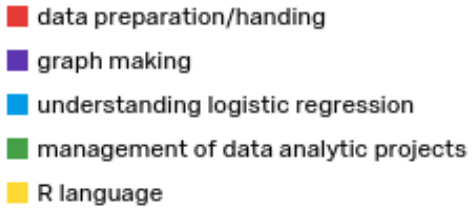
Q1.5 - Please complete the following sentence "I _____ use R for data analysis in my HSI Fellowship"



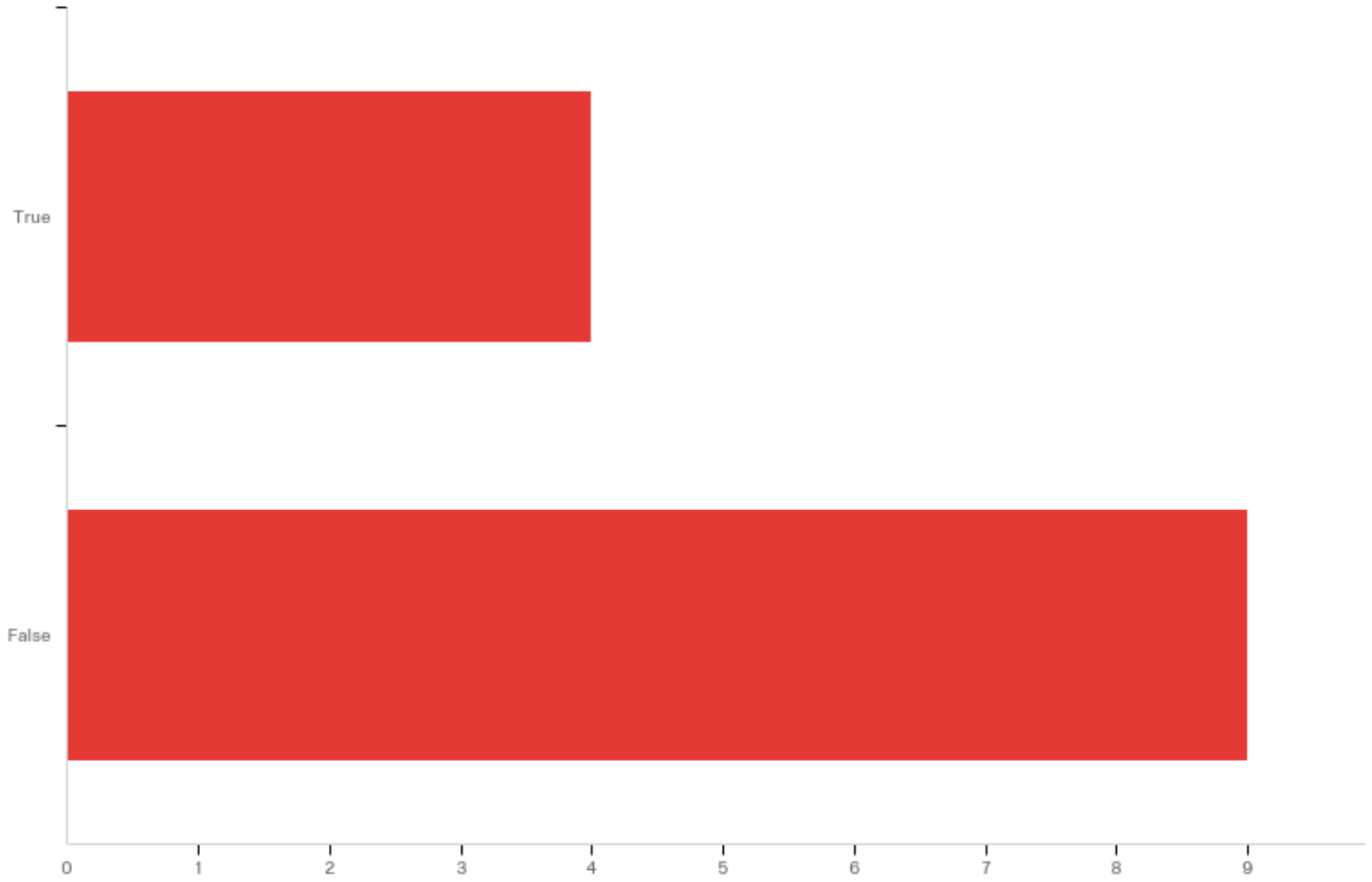
Q1.6 - With respect to learning more about data science with R, please rank your interests in the following learning objectives of this workshop.



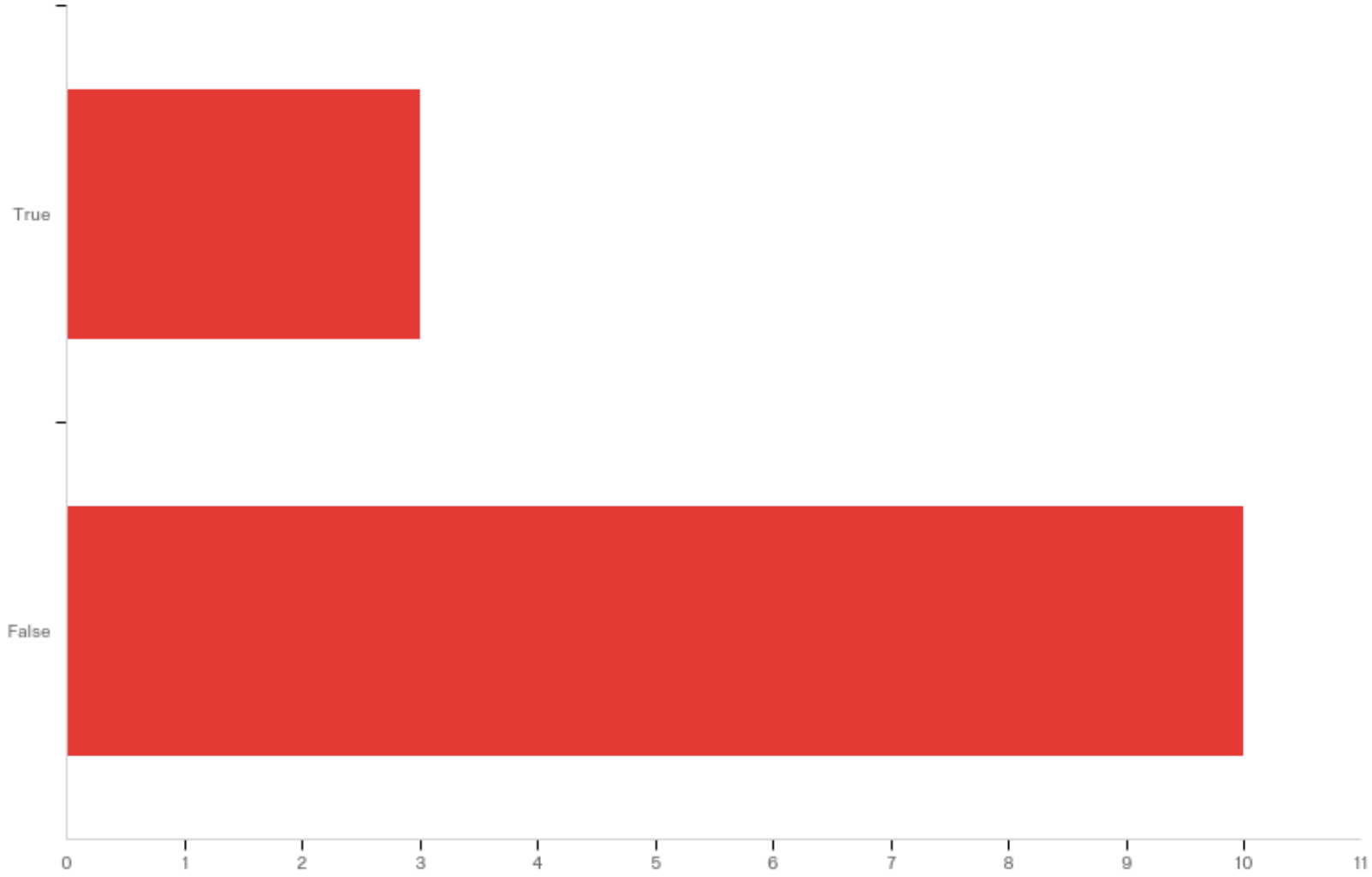
#	Field	Mean
1	data preparation/handing	2.33
2	graph making	3.25
3	understanding logistic regression	4.33
4	management of data analytic projects	2.83
5	R language	2.25



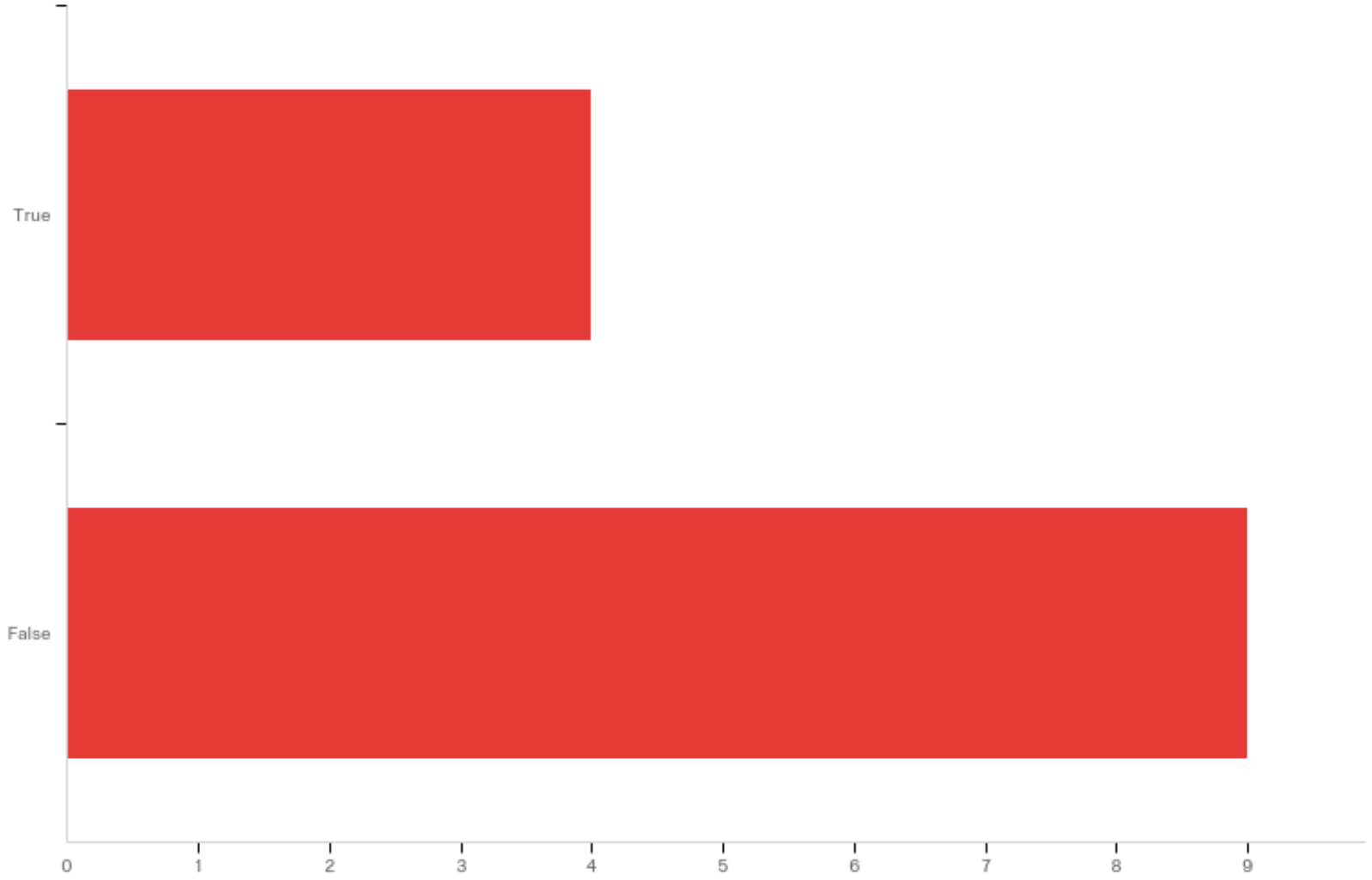
Q2.1 - I have created a graph using ggplot2 package before



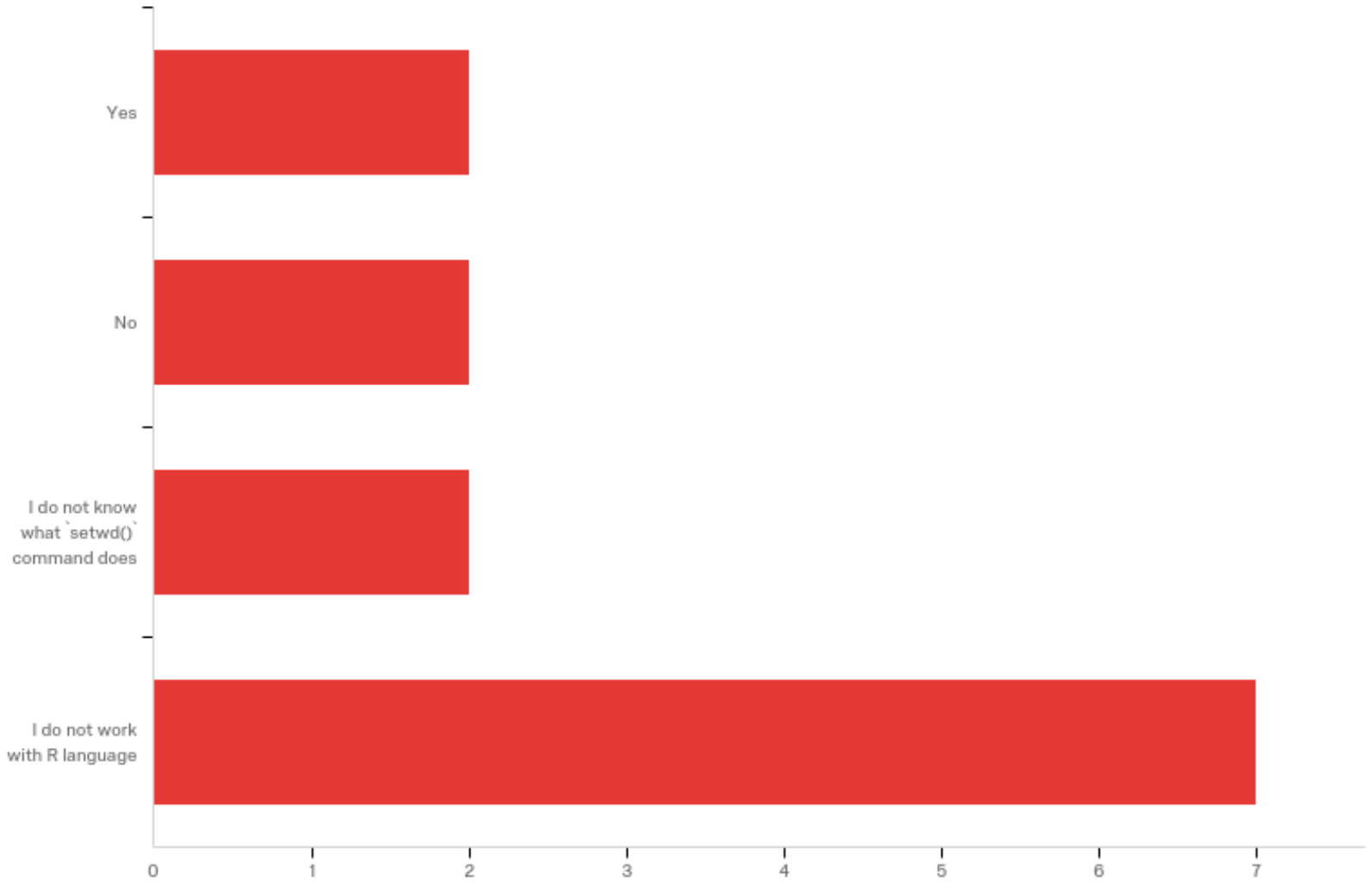
Q2.2 - I have worked with a Rmarkdown notebook before



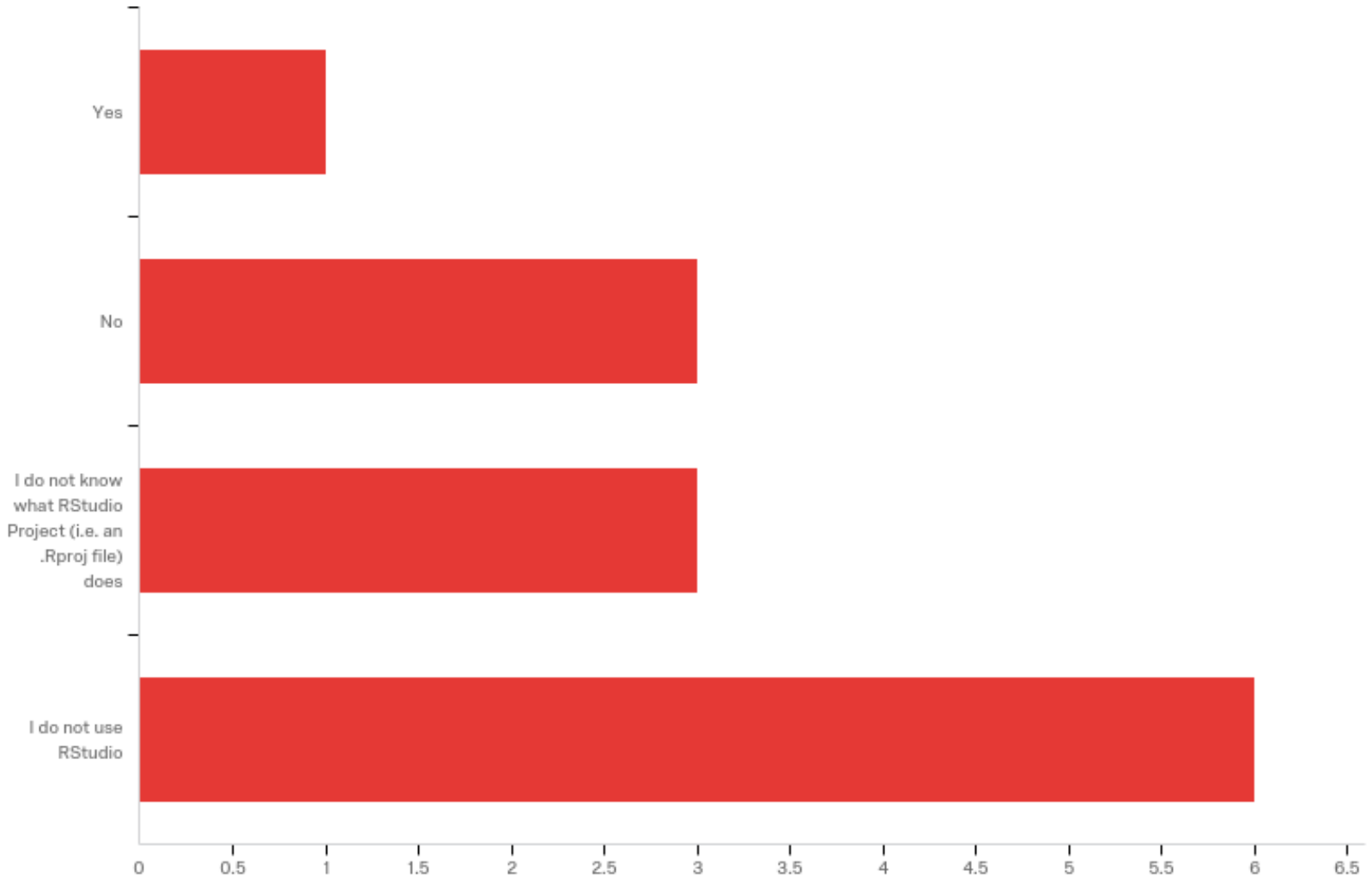
Q2.3 - I have written a custom function in R before



Q2.4 - When working with R language, I use `setdw()` command to establish a home directory



Q2.5 - When working in RStudio, I use Projects (i.e. create an .Rproj file)



About today

Today we will learn to use R + RStudio for

- Wrangling
- Tabulating
- Modeling
- Graphing

We will re-create the analytic report posted on

https://github.com/andkov/hsif-2019-data-analysis

Things to keep in mind

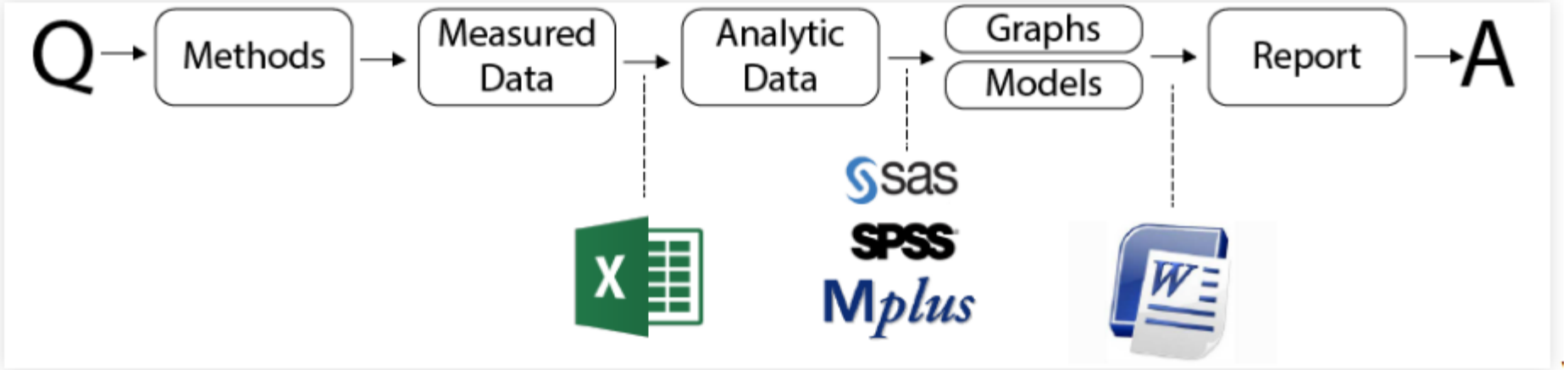
- What makes “data science” a science? Reproducibility

- Principles to keep in mind
  - Scripts are better than GUIs
  - Notebooks are better than scripts
  - Projects are better than Notebooks

- “*There are only two hard things in programming: cache validation and naming things*” – Unknown
  - Success in Data Science = Craft + Imagination

Approaches to managing data analysis

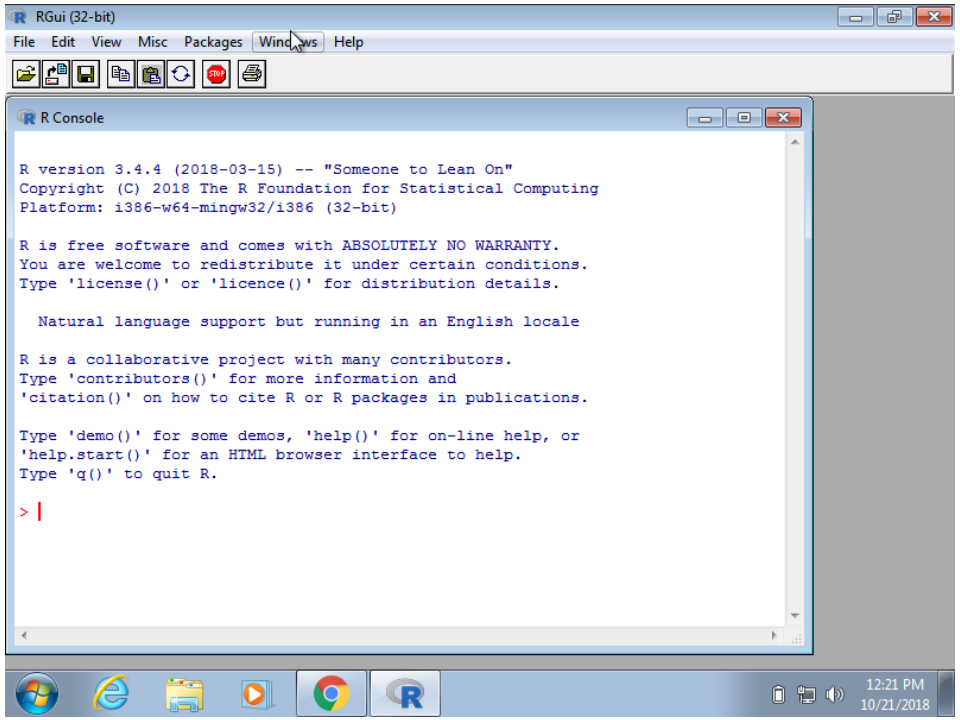
Traditional

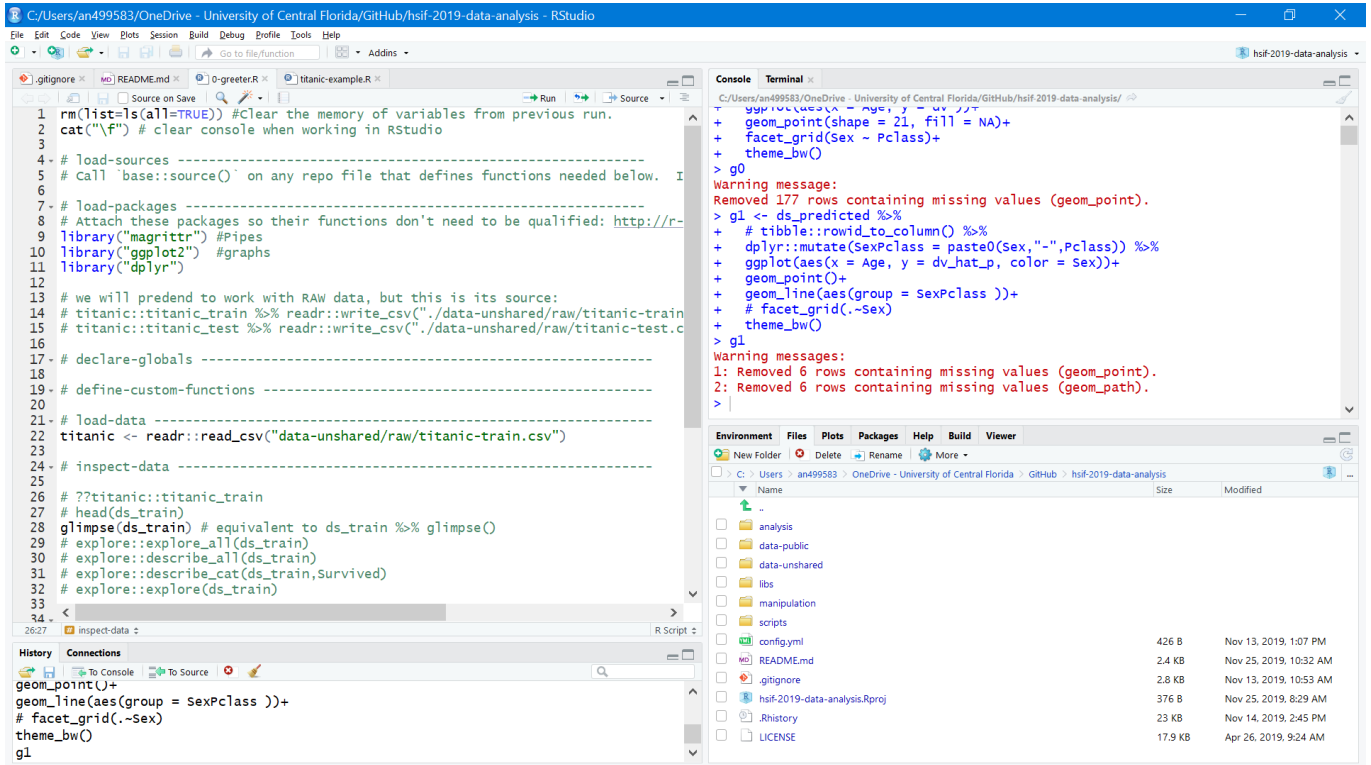


Reproducible

language

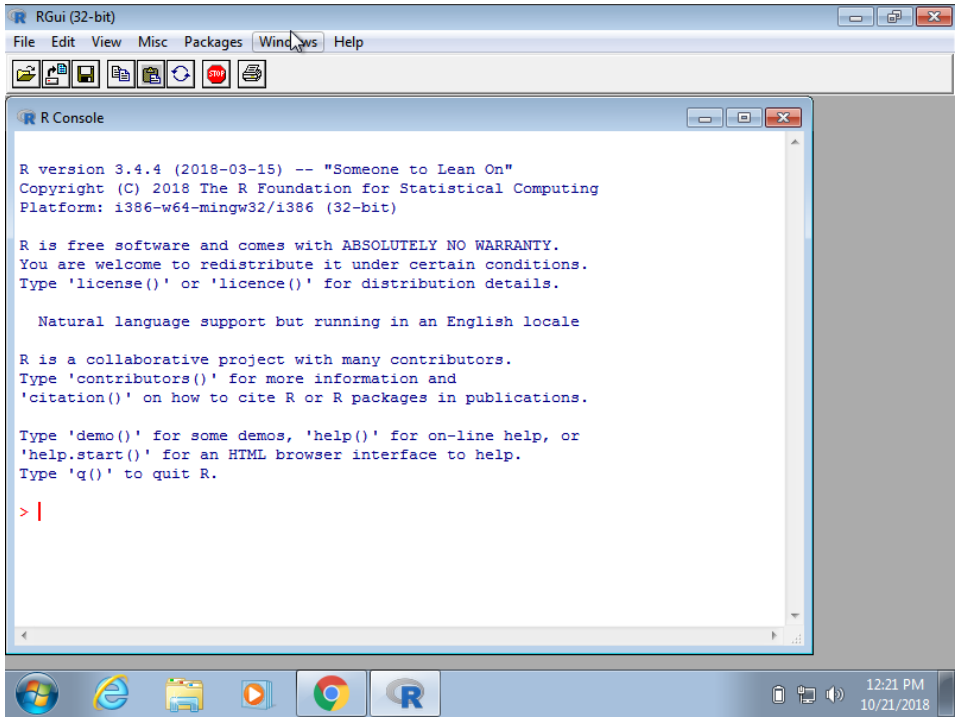
Integrated Development Environment (IDE)

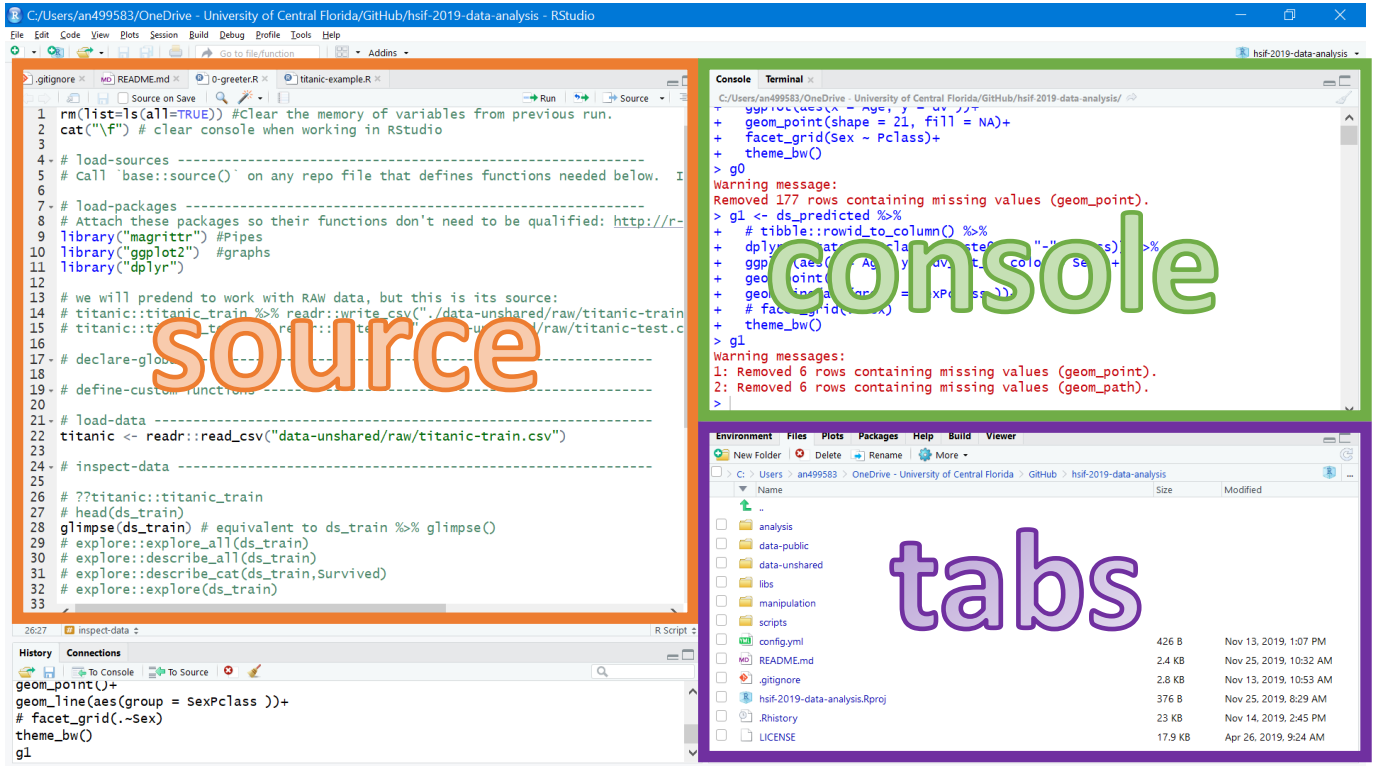




language

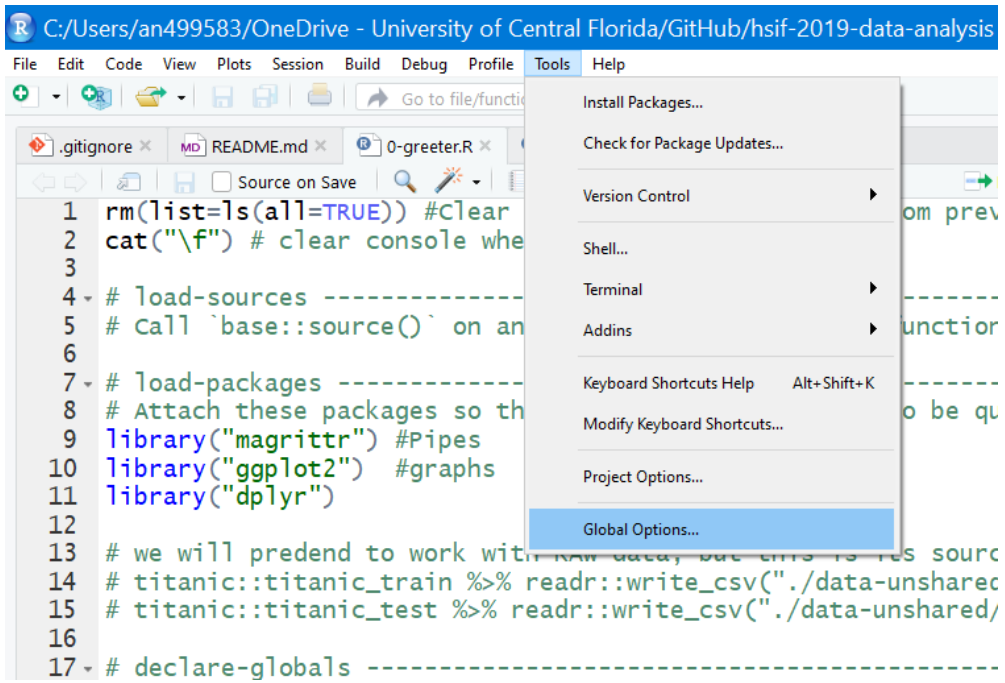
Integrated Development Environment (IDE)

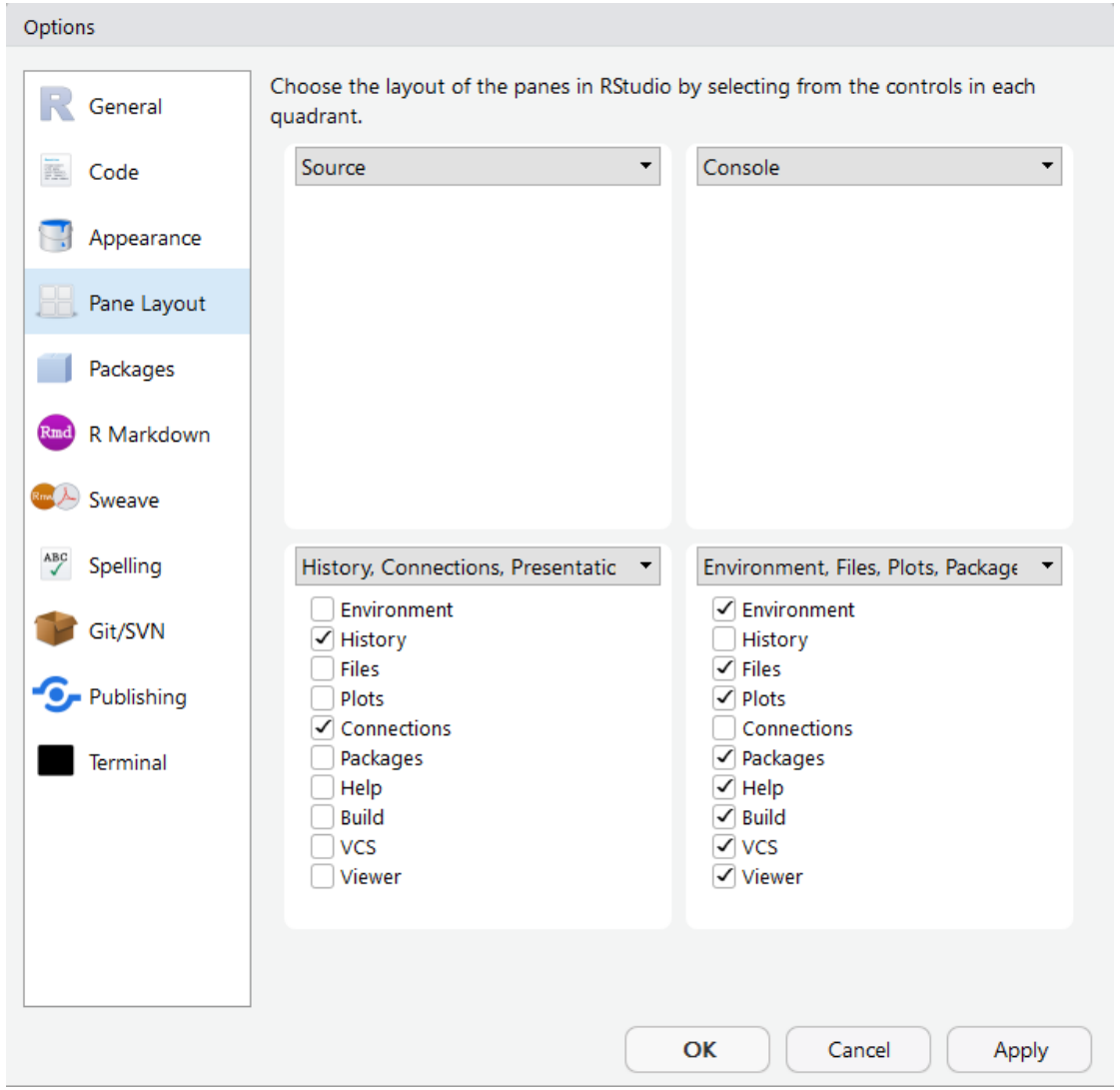




Let us begin!

Setting up RStudio: Suggested Pane Layout





In conclusion

Verbs we have learned today

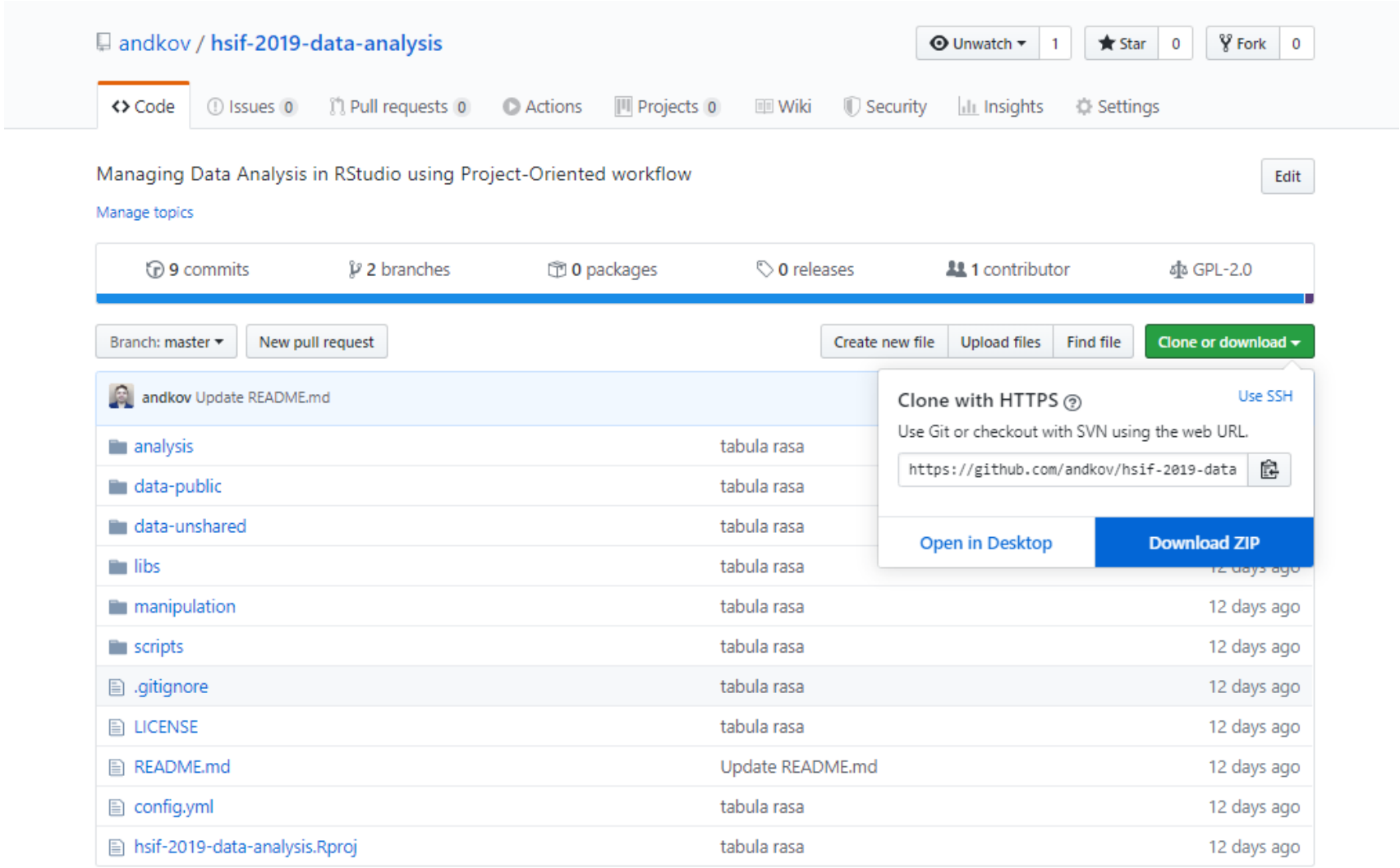
- head()
- dplyr::glimpse()
- explore::describe_all()
- names()
- dplyr::group_by()
- dplyr::summarize()
- tolower()
- dplyr::rename()
- dplyr::mutate()
- factor()

- stats::glm()
- summary()
- predict()
- plogis()

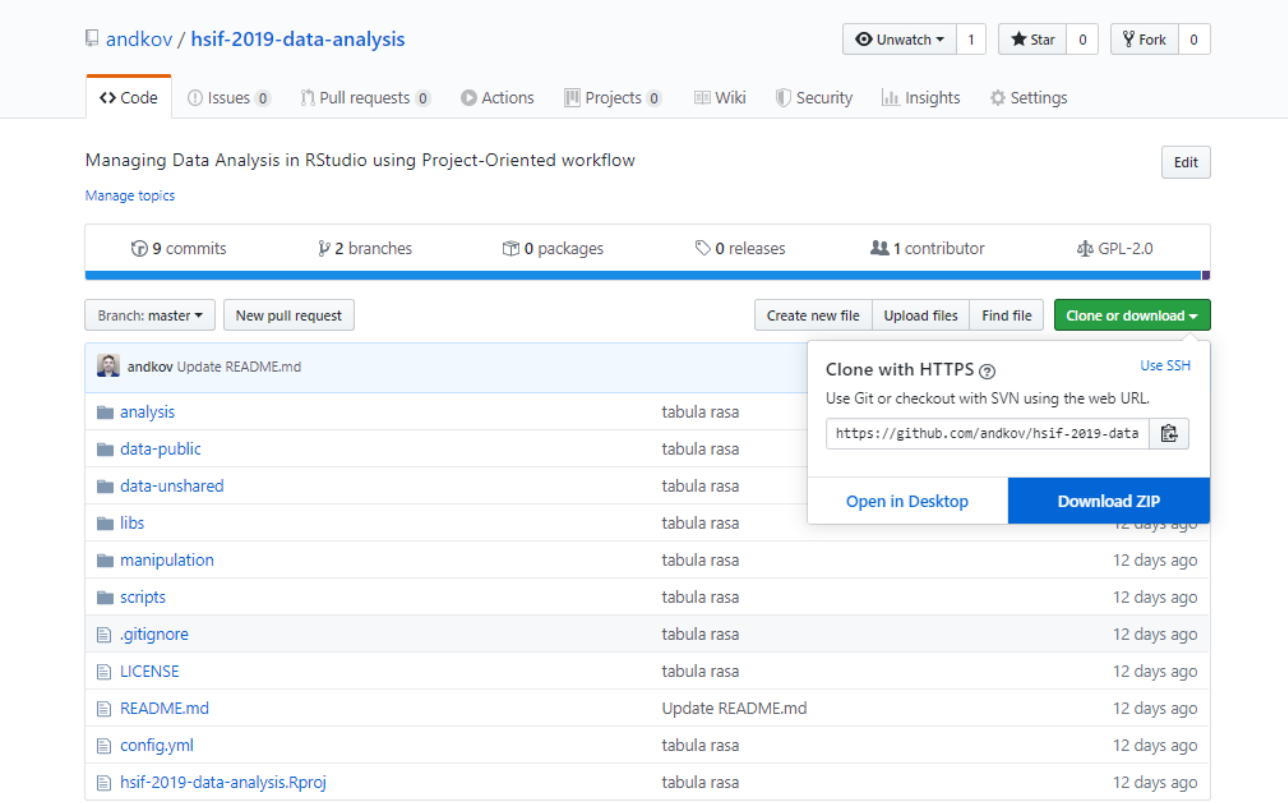
- ggplot()
- geom_bar()
- geom_point(

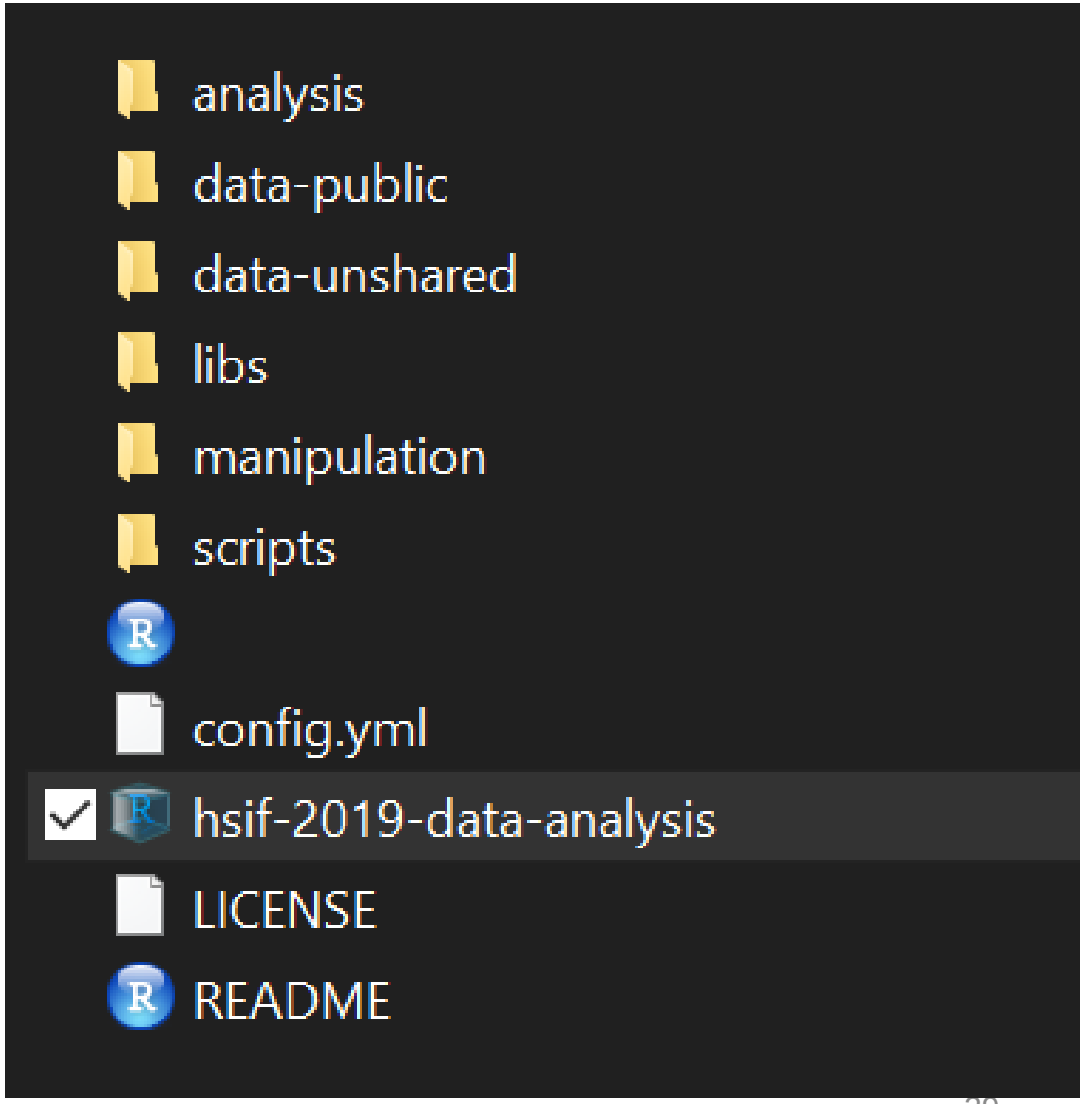
https://github.com/andkov/hsif-2019-data-analysis

Download repository to view all materials



Folder Architecture





Learning Resources

- Rmarkdown guide (https://rmarkdown.rstudio.com/)
- Logistic regression (Youtube: StatQuest + Logistic Regression)
- R4DS (https://r4ds.had.co.nz/) + swirl (https://swirlstats.com/)
- Introduction to ggplot2 (http://www.cookbook-r.com/Graphs/)

Lessons & Metaphors

- Handle your data! (Vesalius)

- Look at your data! (Tukey)

- Graph is art (Tufte)

- Graph is language (Wickham)

- Coding is music (Gould)

Questions? Comments?

Andriy Koval

https://github.com/andkov

http://andriy.rbind.io