

# Managing Data Analysis in RStudio using project-oriented workflow

Andriy Koval, Ph.D.

Health System Impact Fellowship  
3<sup>rd</sup> Annual Training Retreat  
2019-11-26-Tuesday  
Toronto, Ontario



UNIVERSITY OF  
CENTRAL FLORIDA



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

# About me



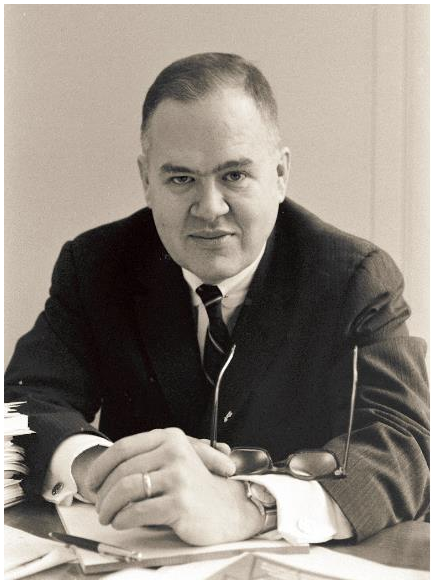
- Ph.D. in Quantitative Methods, Psychology (2014)
- Reproducible research enthusiast since 2012
- Graph maker
- See work at <https://github.com/andkov>
- These slides and more at <http://andriy.rbind.io>

**MIDDLE  
TENNESSEE**  
STATE UNIVERSITY

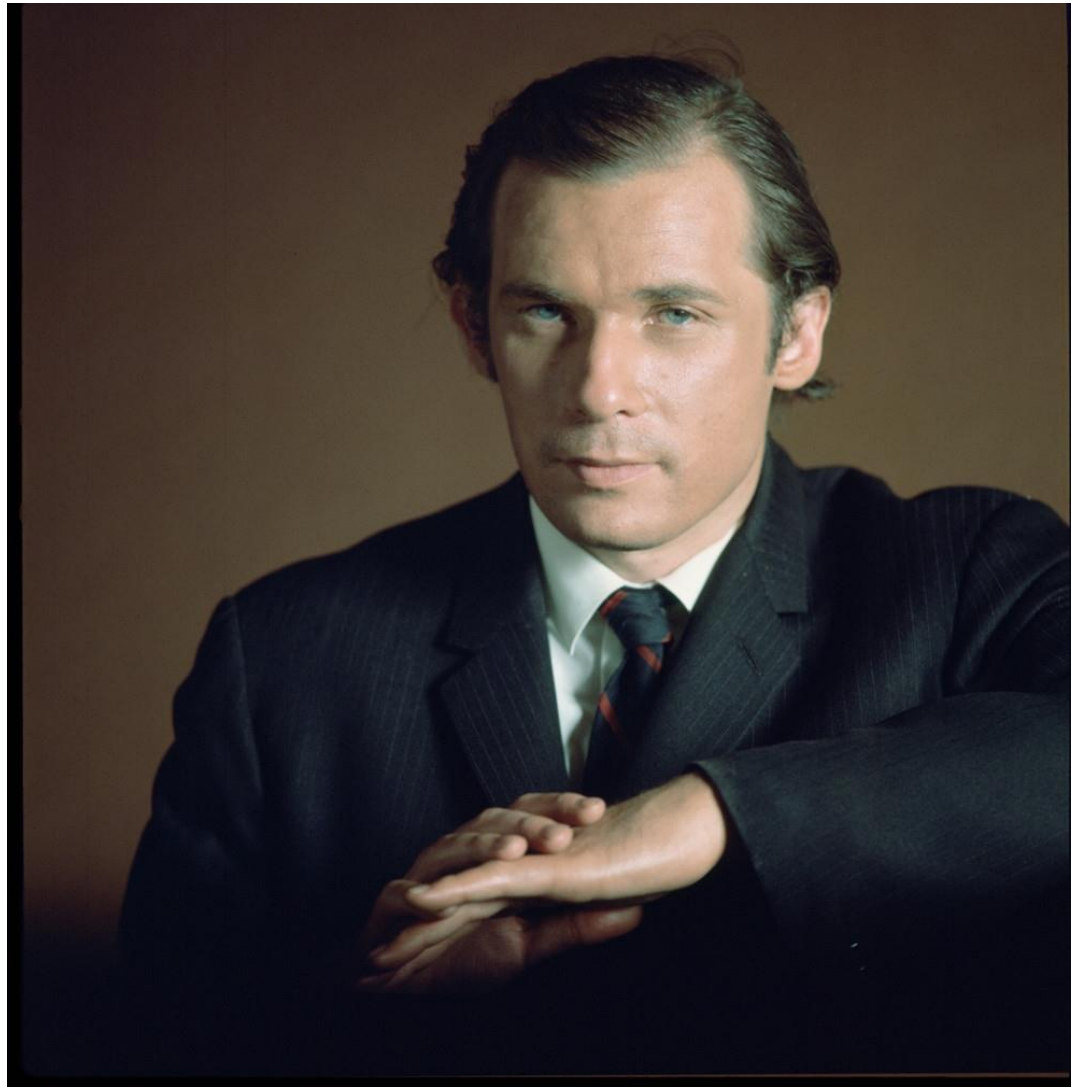




Andreas Vesalius



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

$$y_{it} = \beta_0 + \beta_1 \text{time}_t + \varepsilon_{it}$$

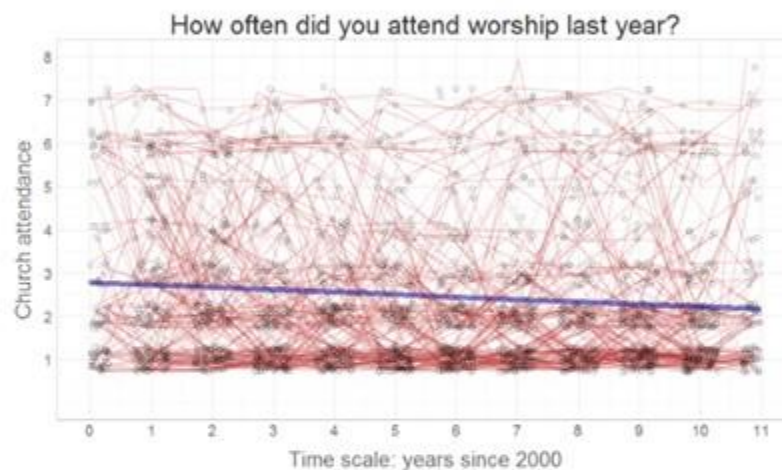
$$\beta_0 = \gamma_{00}$$

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

## 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)
```

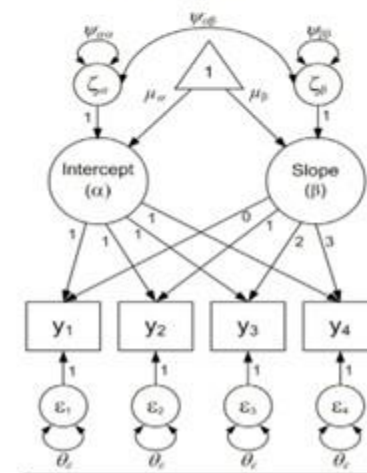
## Numeric

Coefficients:

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

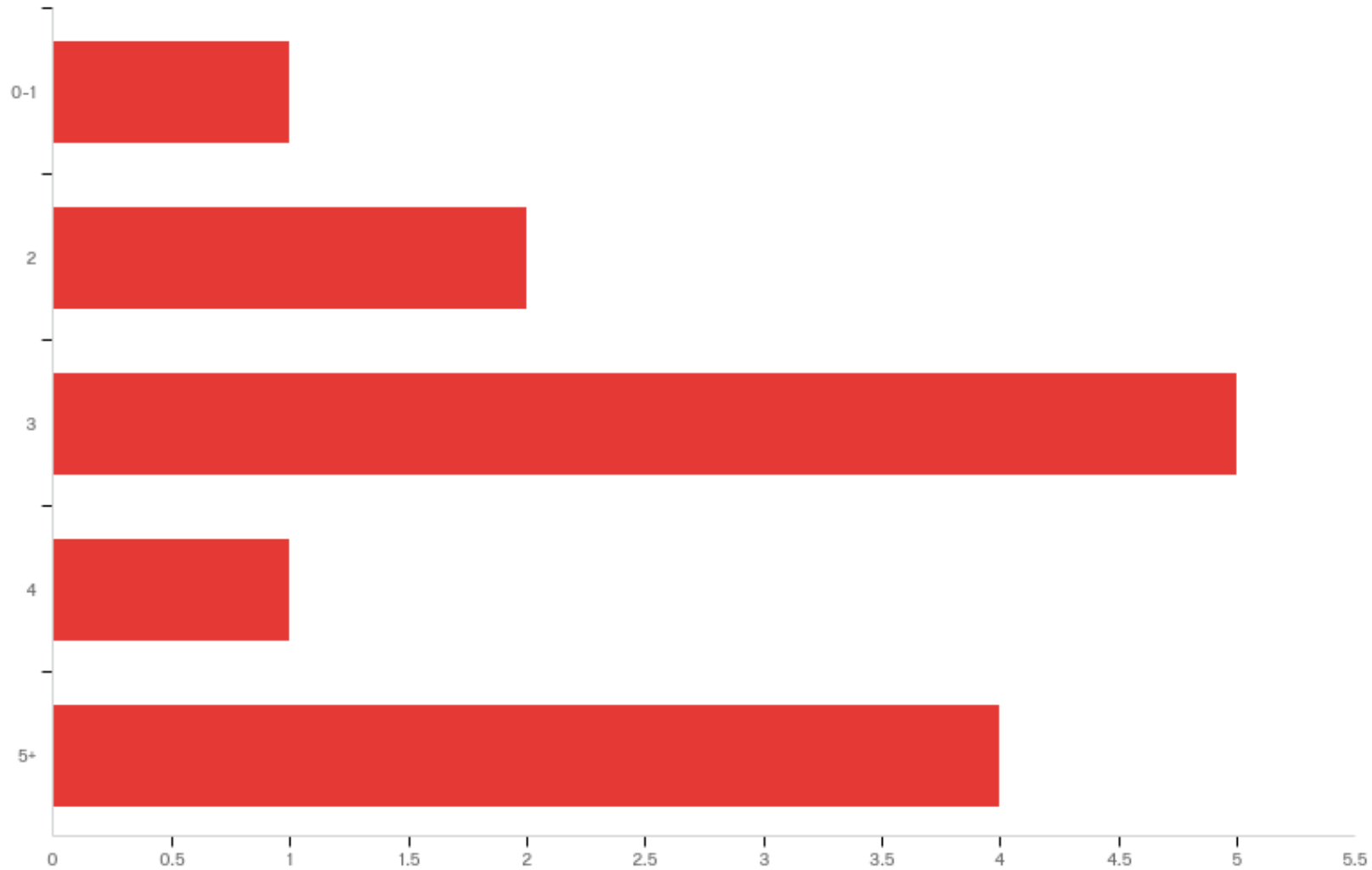
	modelB
logLik	-3719
deviance	7438
AIC	7444
BIC	7461
df.resid	1858
N	1860
p	2
ids	155

## Schematic

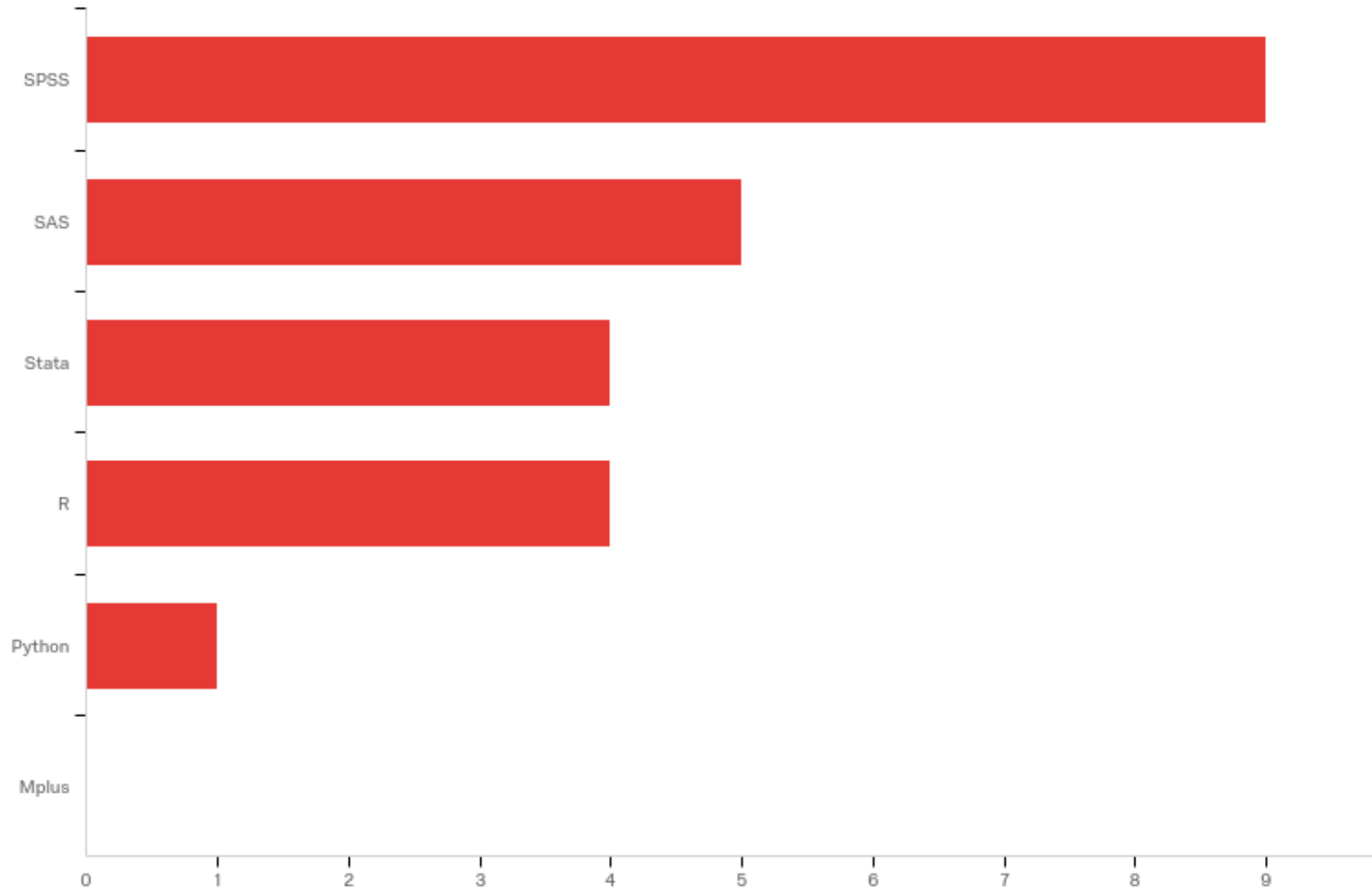


About you

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

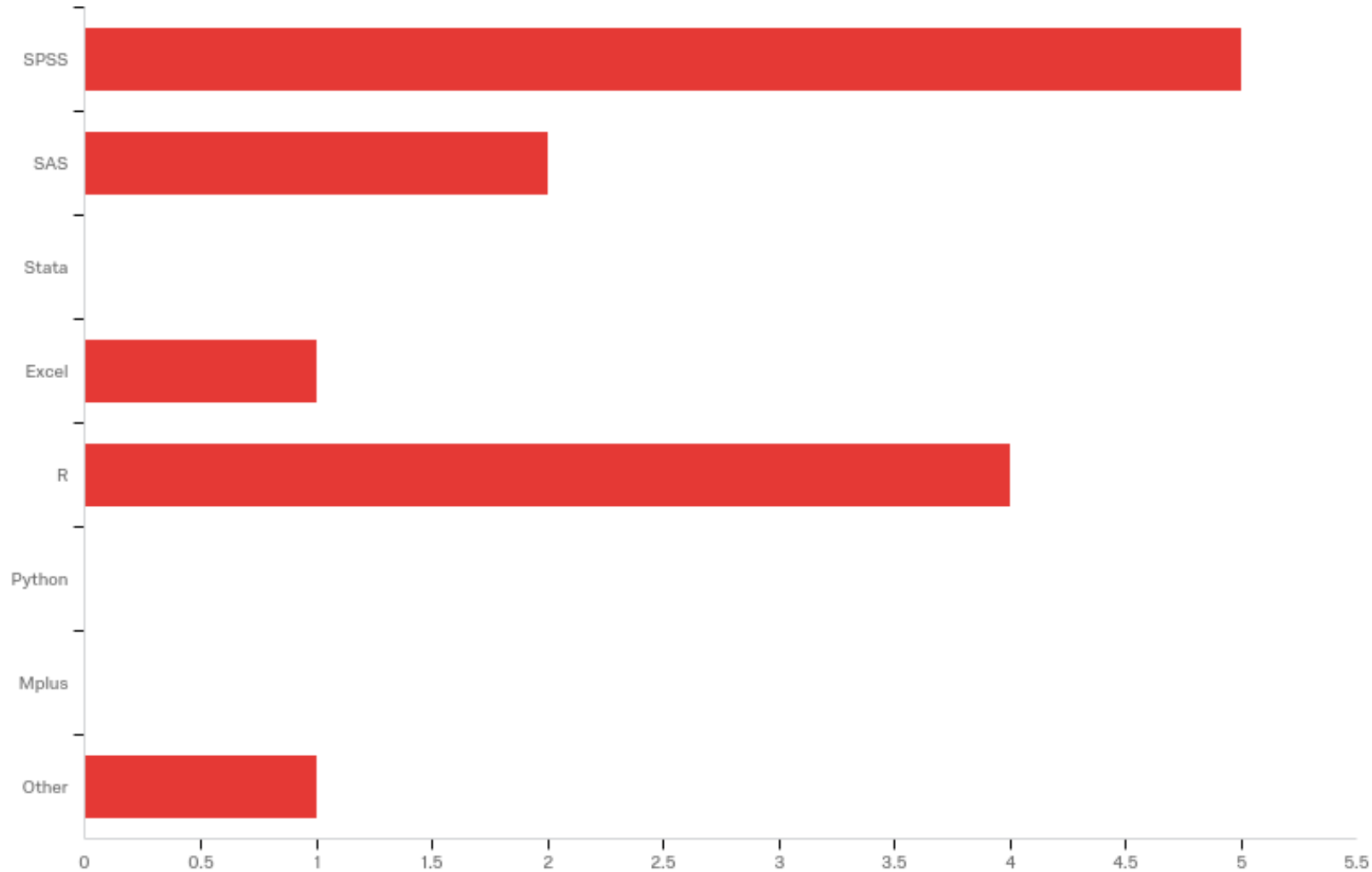


Q1.2 - What statistical software have you used AT LEAST ONCE in the last 3 years? (check all that apply)

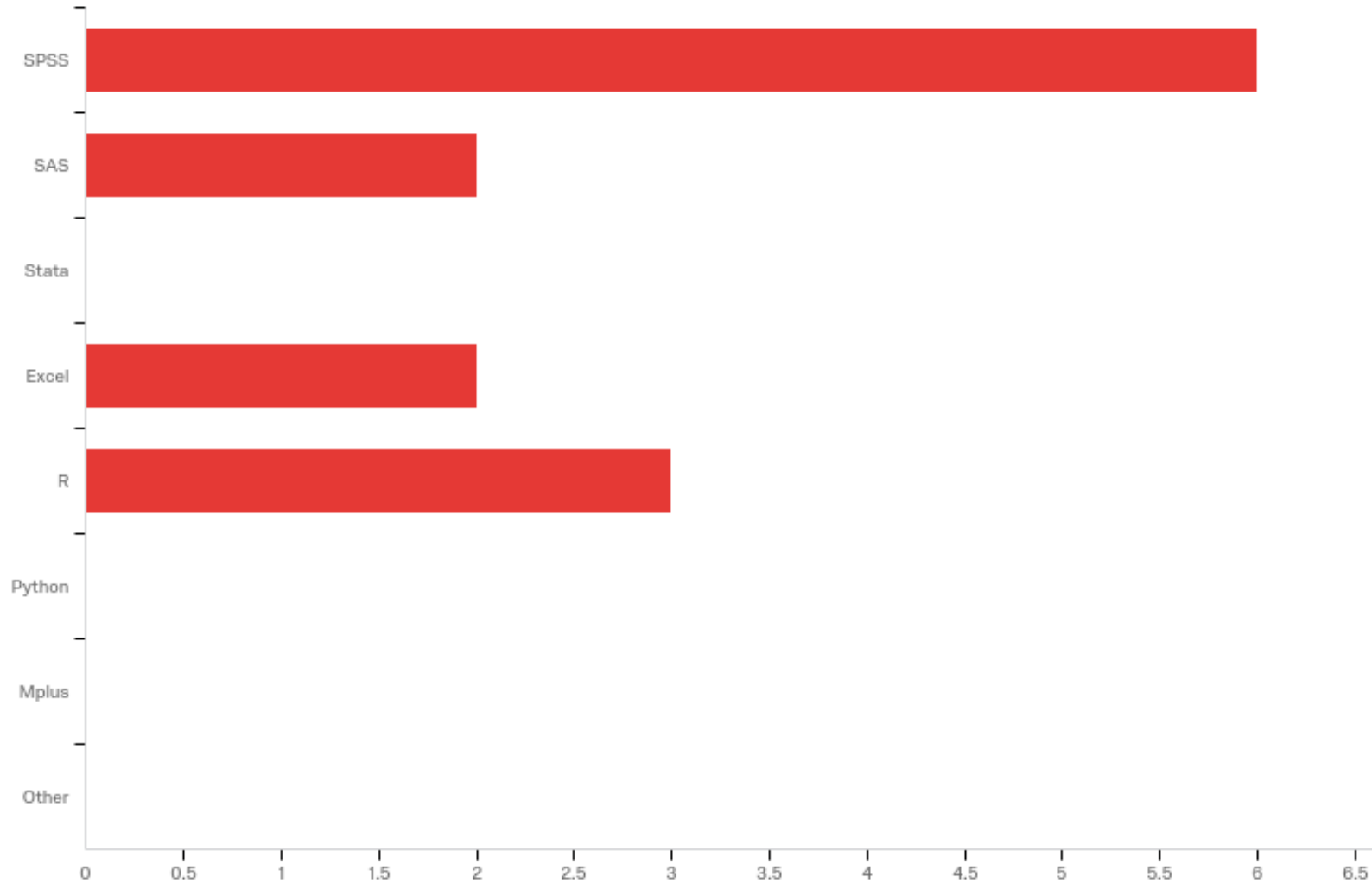




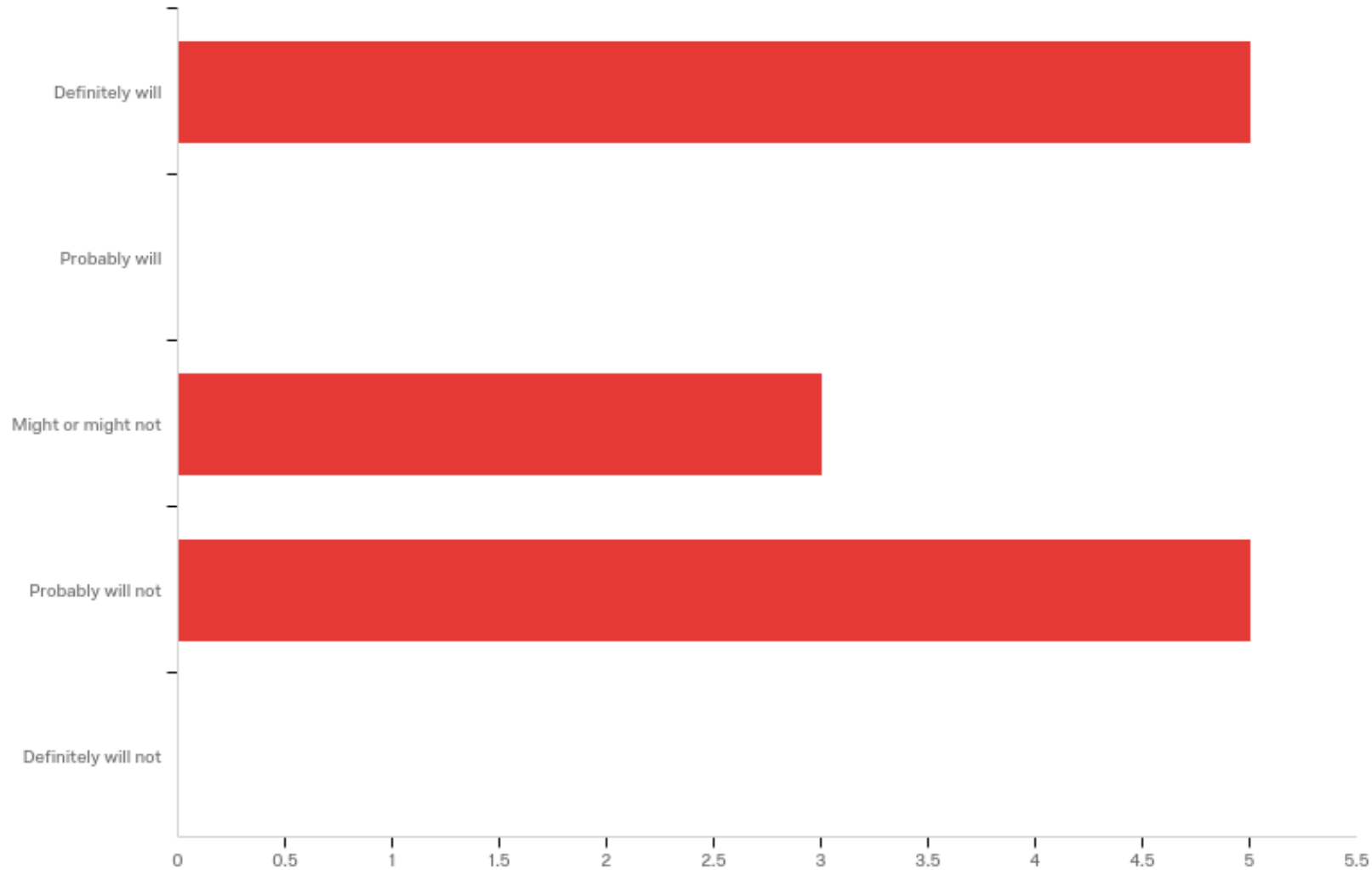
Q1.3 - You are asked to RUN A REGRESSION on an analysis-ready data set. With what software would you be most comfortable performing this task?



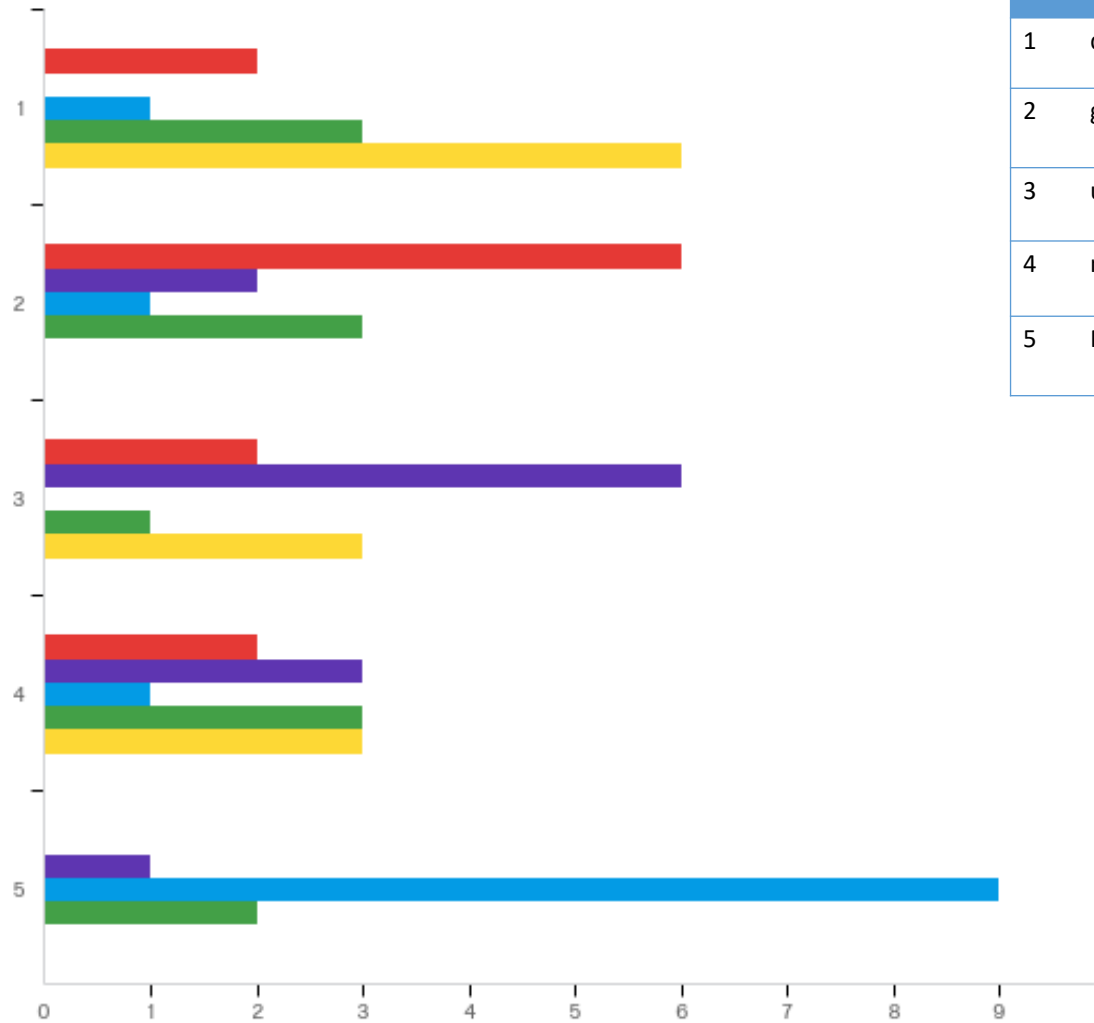
Q1.4 - You are asked to MAKE A GRAPH of a frequency distribution using an analysis-ready data set. With what software would you be most comfortable performing this task?



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



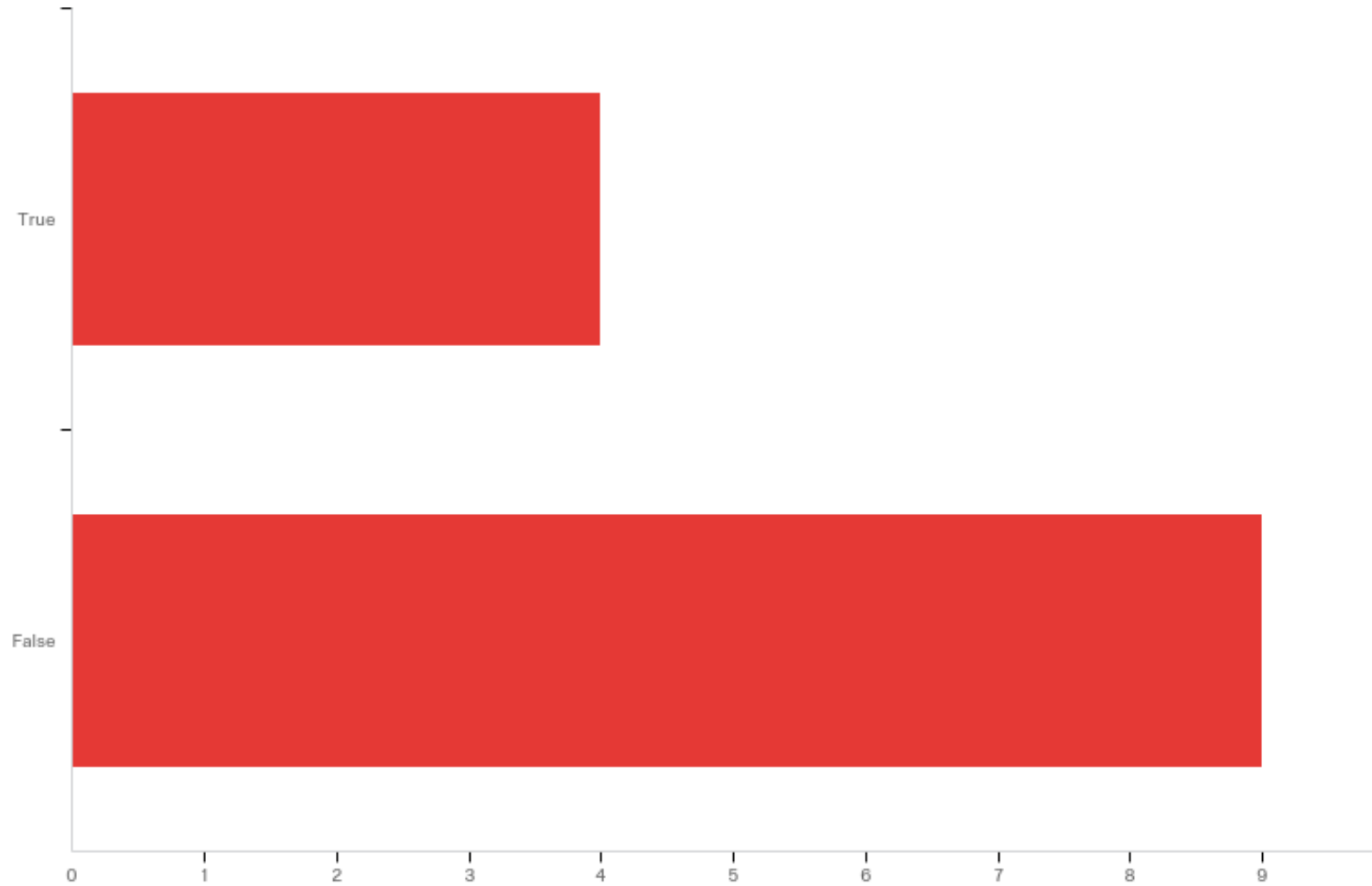
Q1.6 - With respect to learning more about data science with R, please rank your interests in the following learning objectives of this workshop.



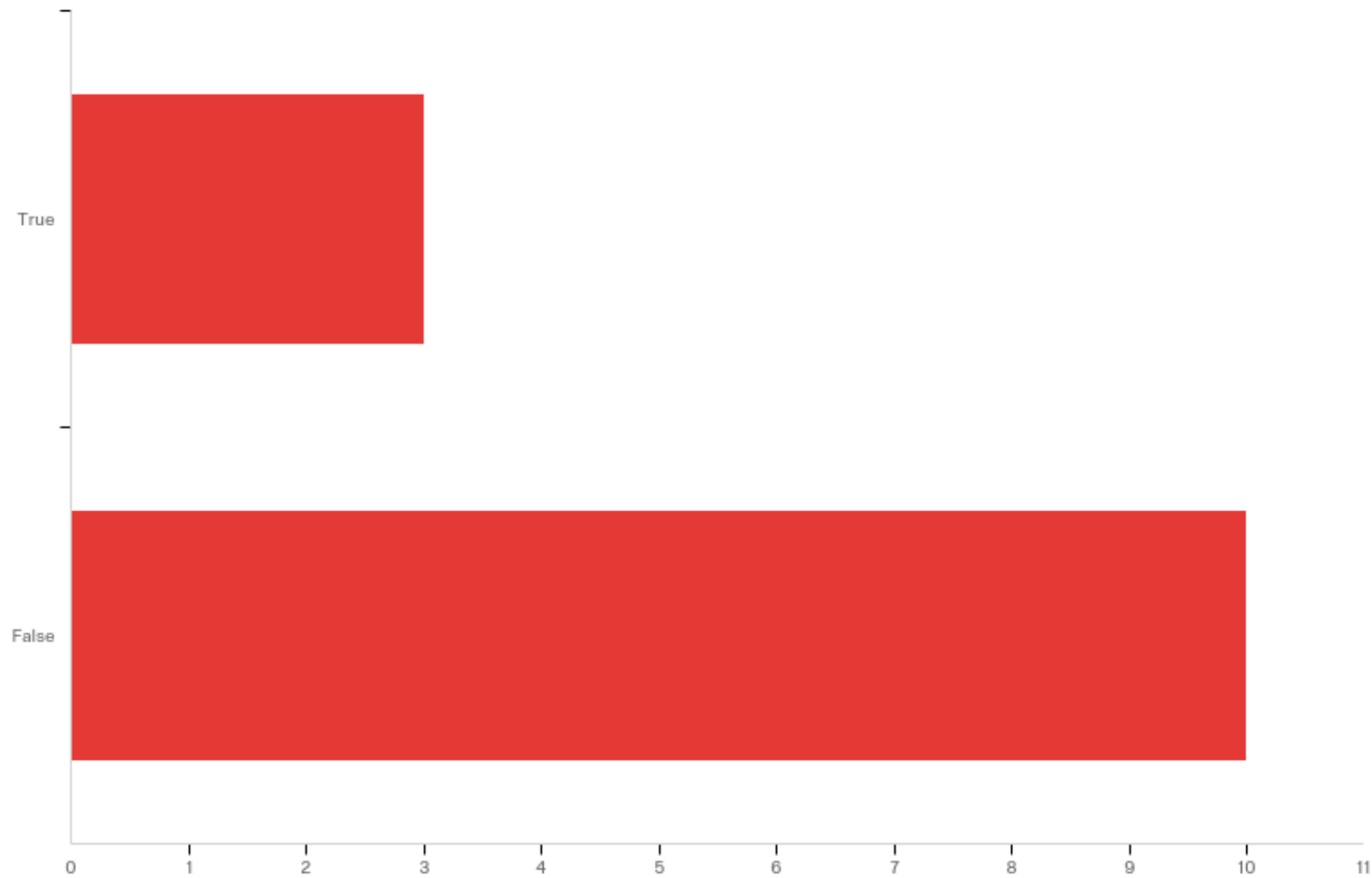
#	Field	Mean
1	data preparation/handling	2.33
2	graph making	3.25
3	understanding logistic regression	4.33
4	management of data analytic projects	2.83
5	R language	2.25

- data preparation/handling
- graph making
- understanding logistic regression
- management of data analytic projects
- R language

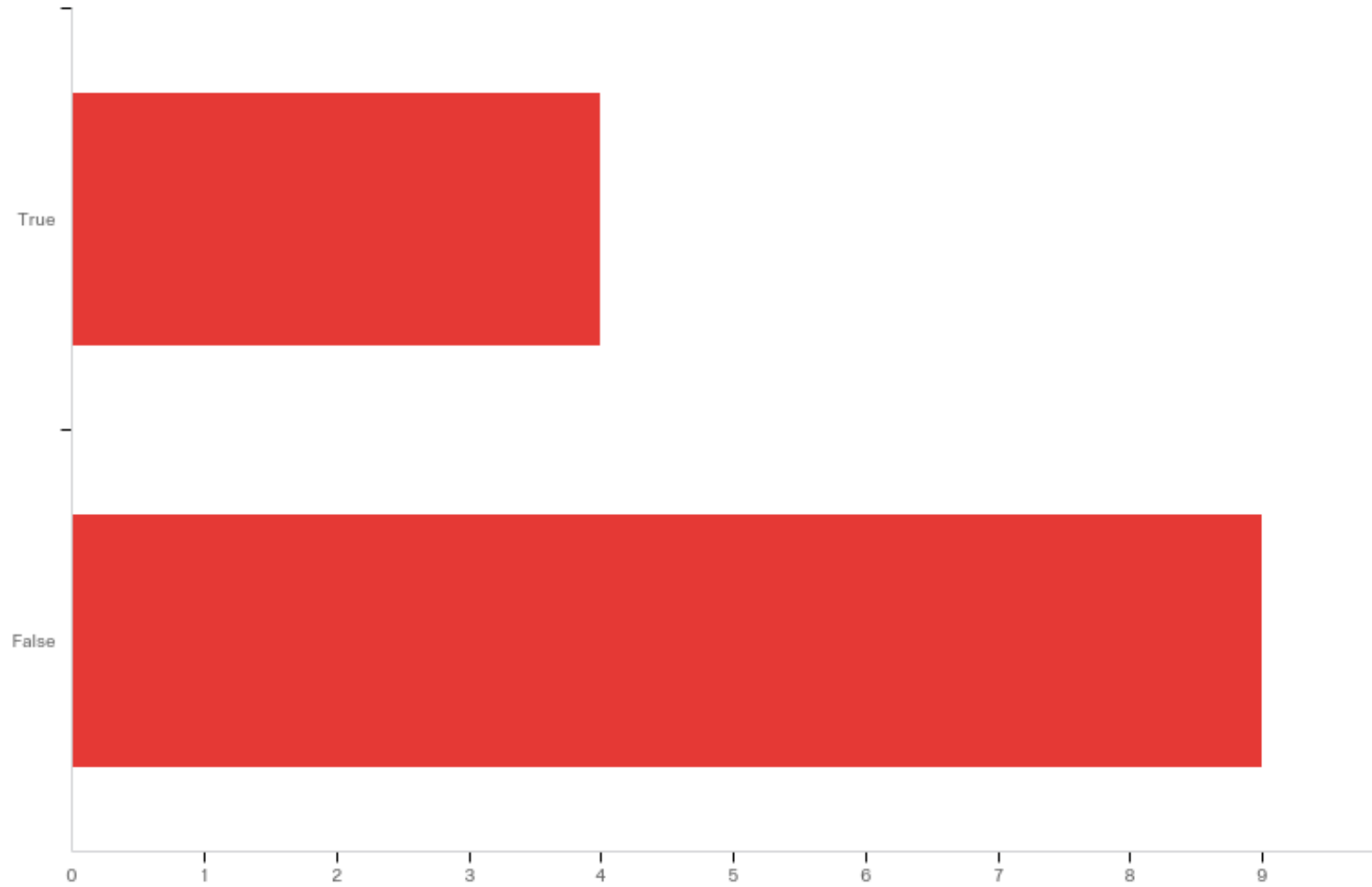
Q2.1 - I have created a graph using ggplot2 package before



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

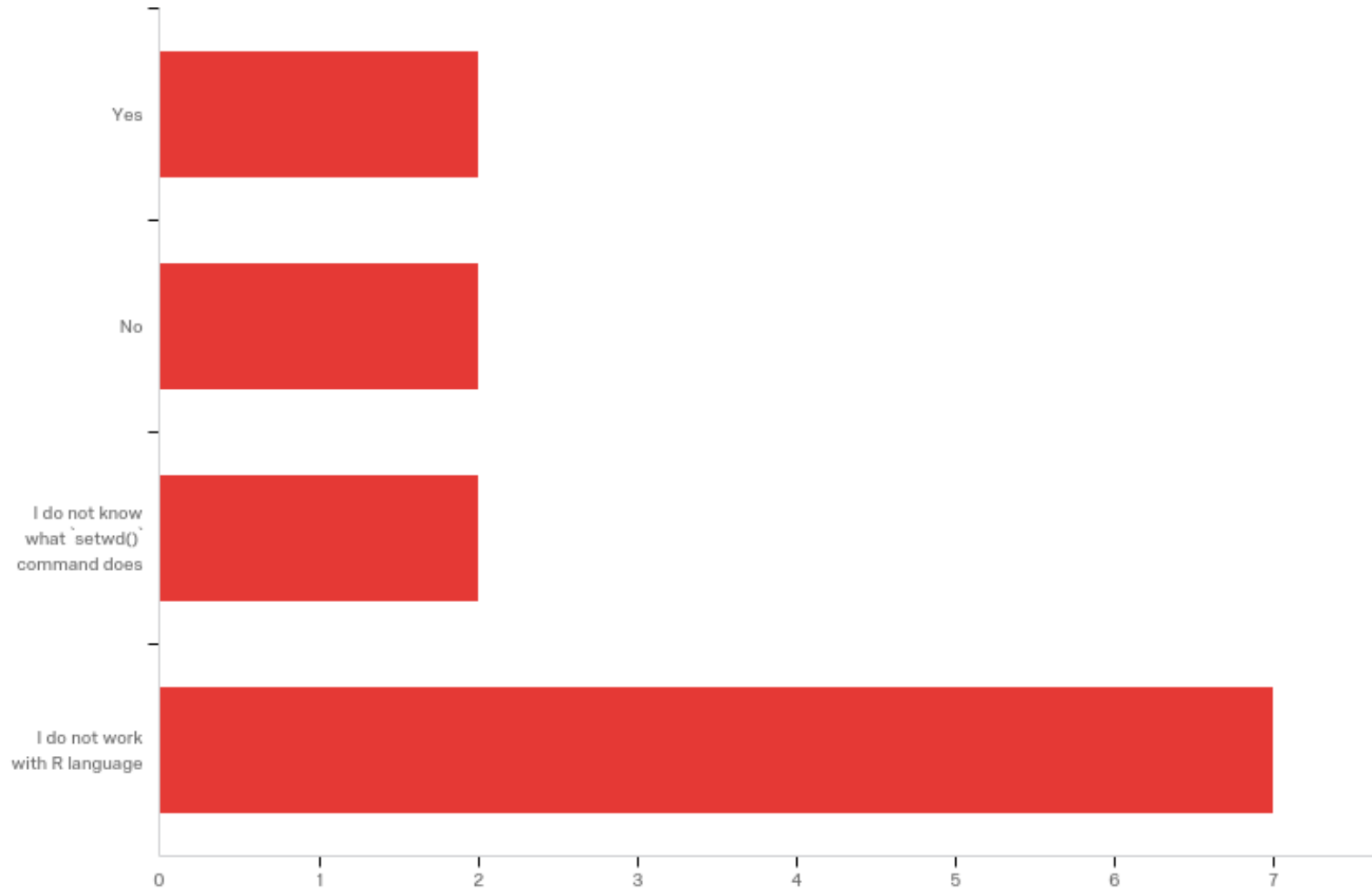


## Q2.3 - I have written a custom function in R before

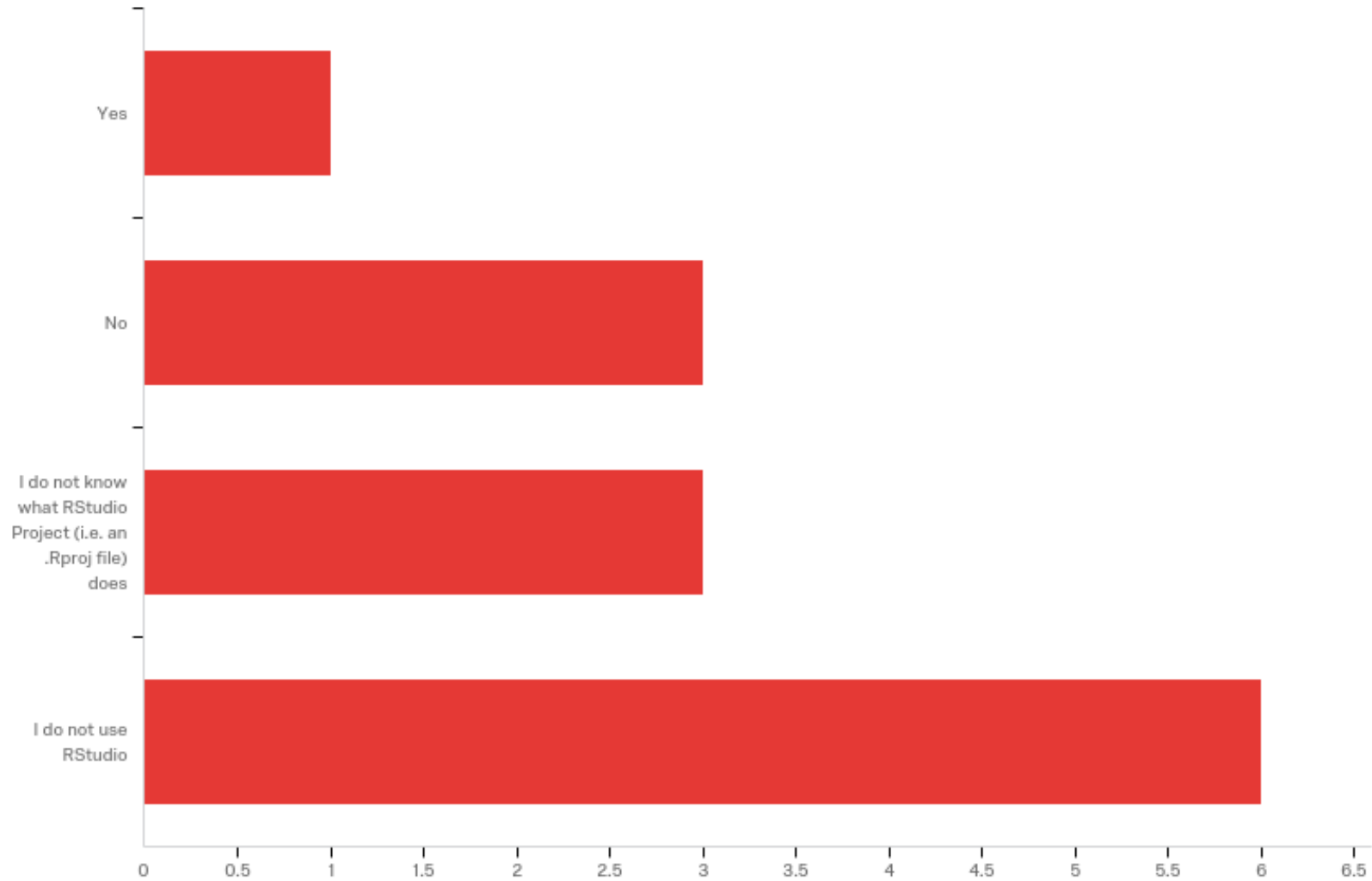




Q2.4 - When working with R language, I use ``setwd()`` command to establish a home directory



## 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

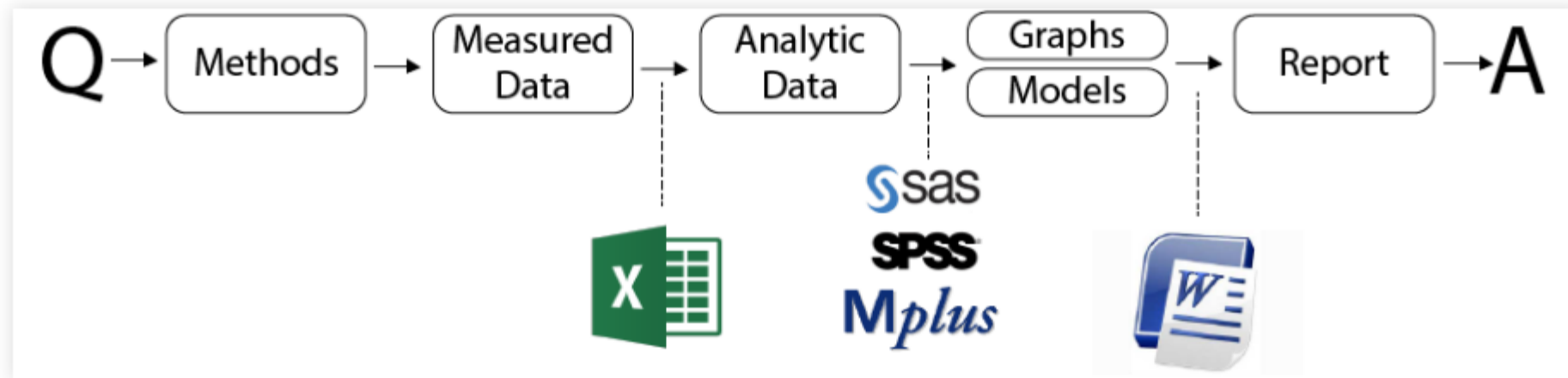
<https://github.com/andkov/hsif-2019-data-analysis>

# Things to keep in mind

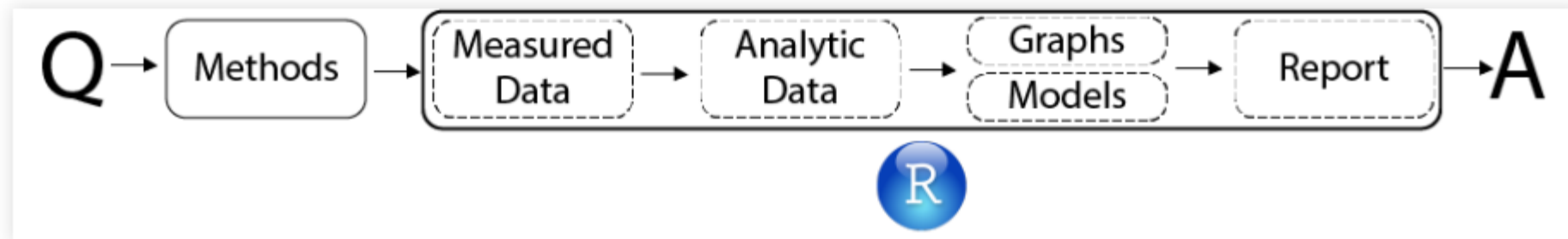
- What makes “data science” a science? **Reproducibility**
- Principles to keep in mind
  - **Scripts** are better than GUIs
  - **Notebooks** are better than scripts
  - **Projects** are better than Notebooks
- “*There are only two hard things in programming: cache validation and naming things*” – Unknown
  - Success in Data Science = Craft + **Imagination**

# Approaches to managing data analysis

## Traditional

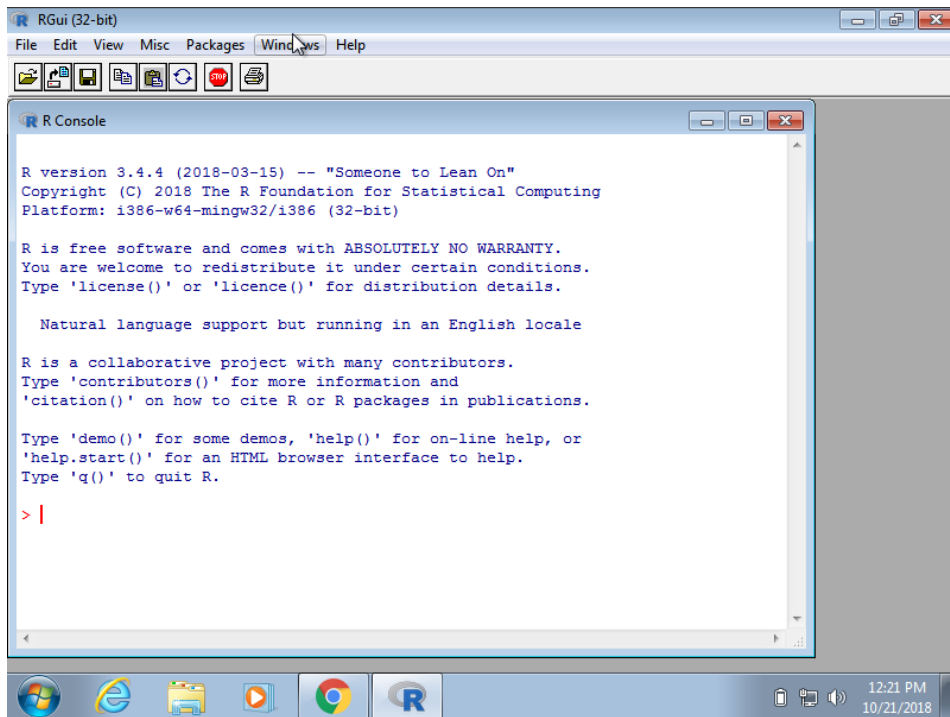


## Reproducible

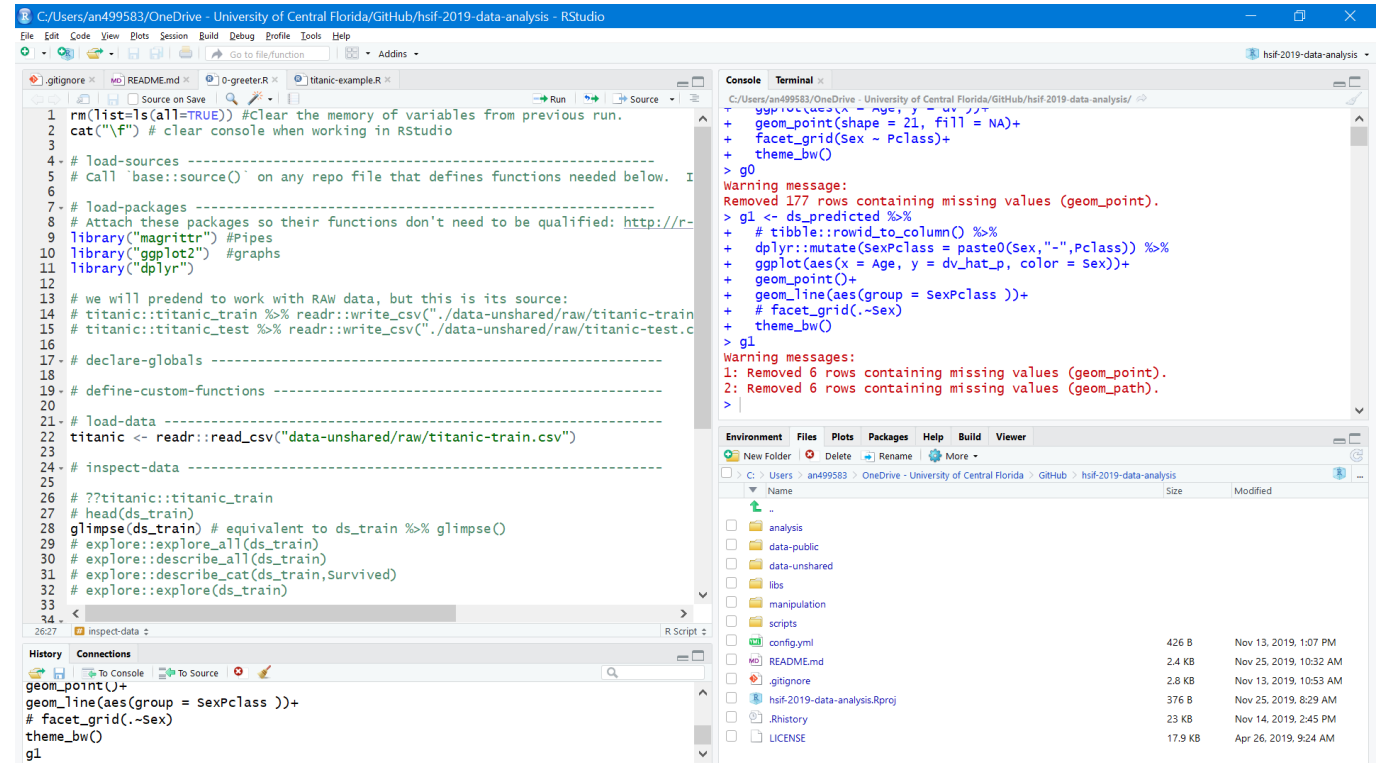




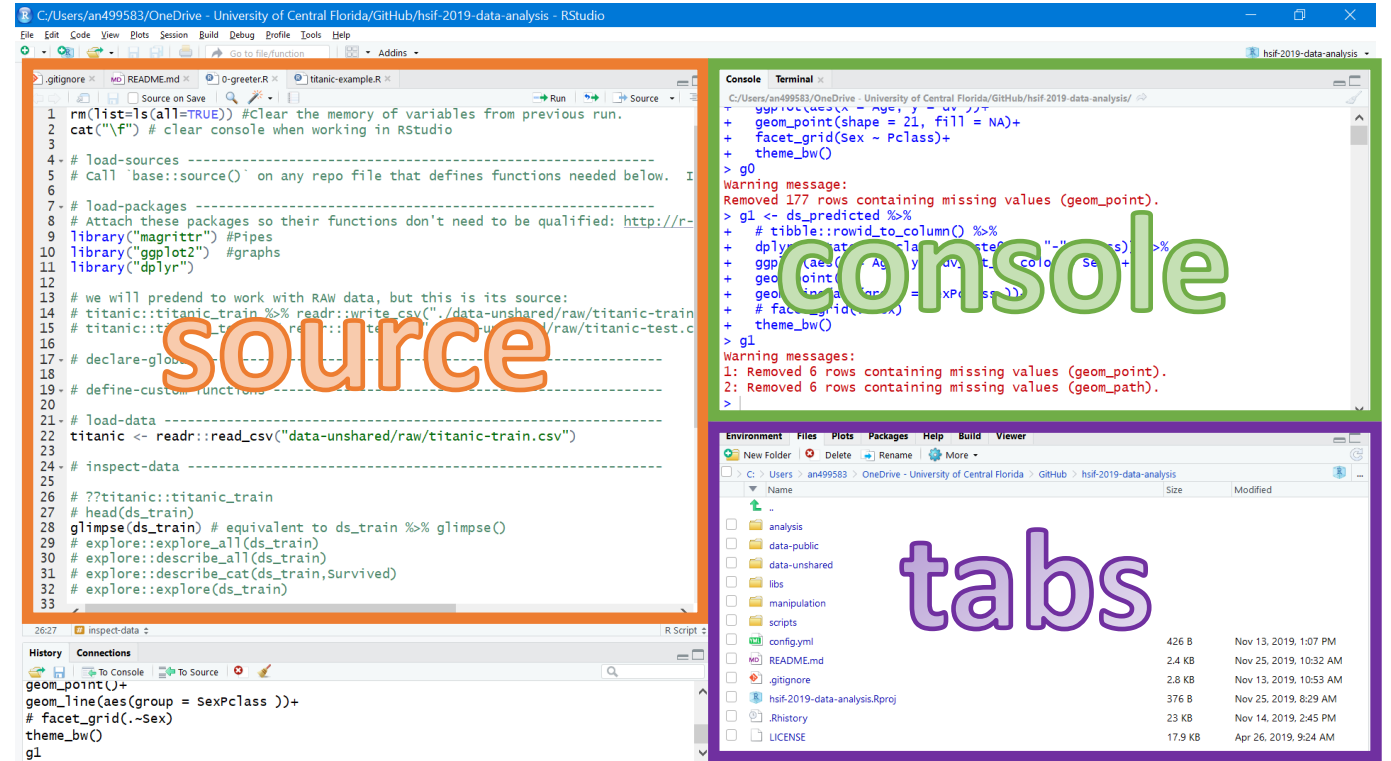
language



Integrated Development Environment (IDE)

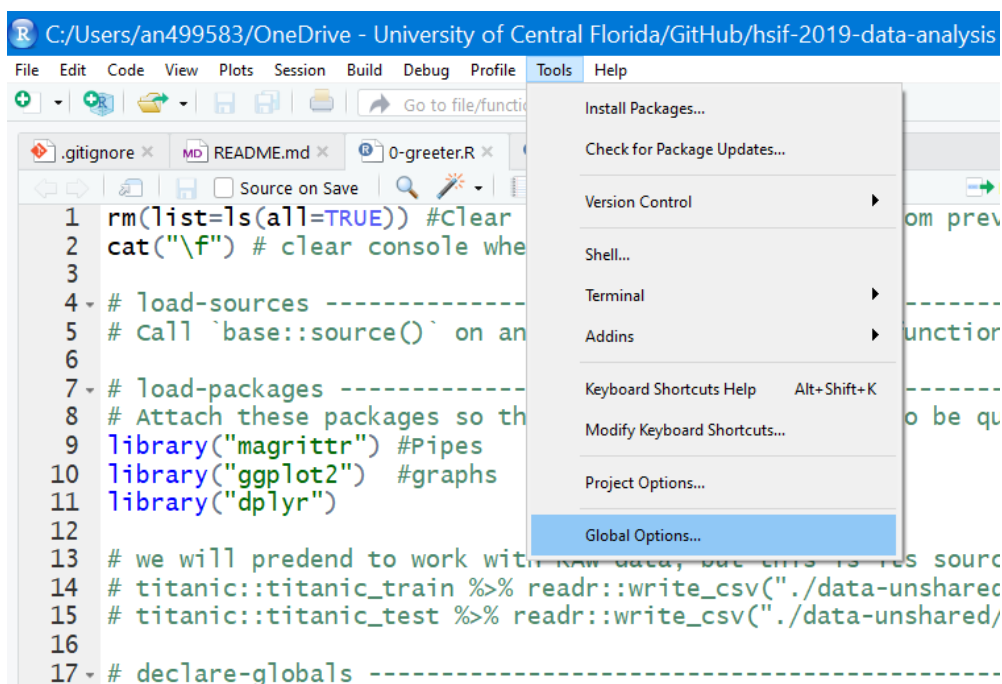






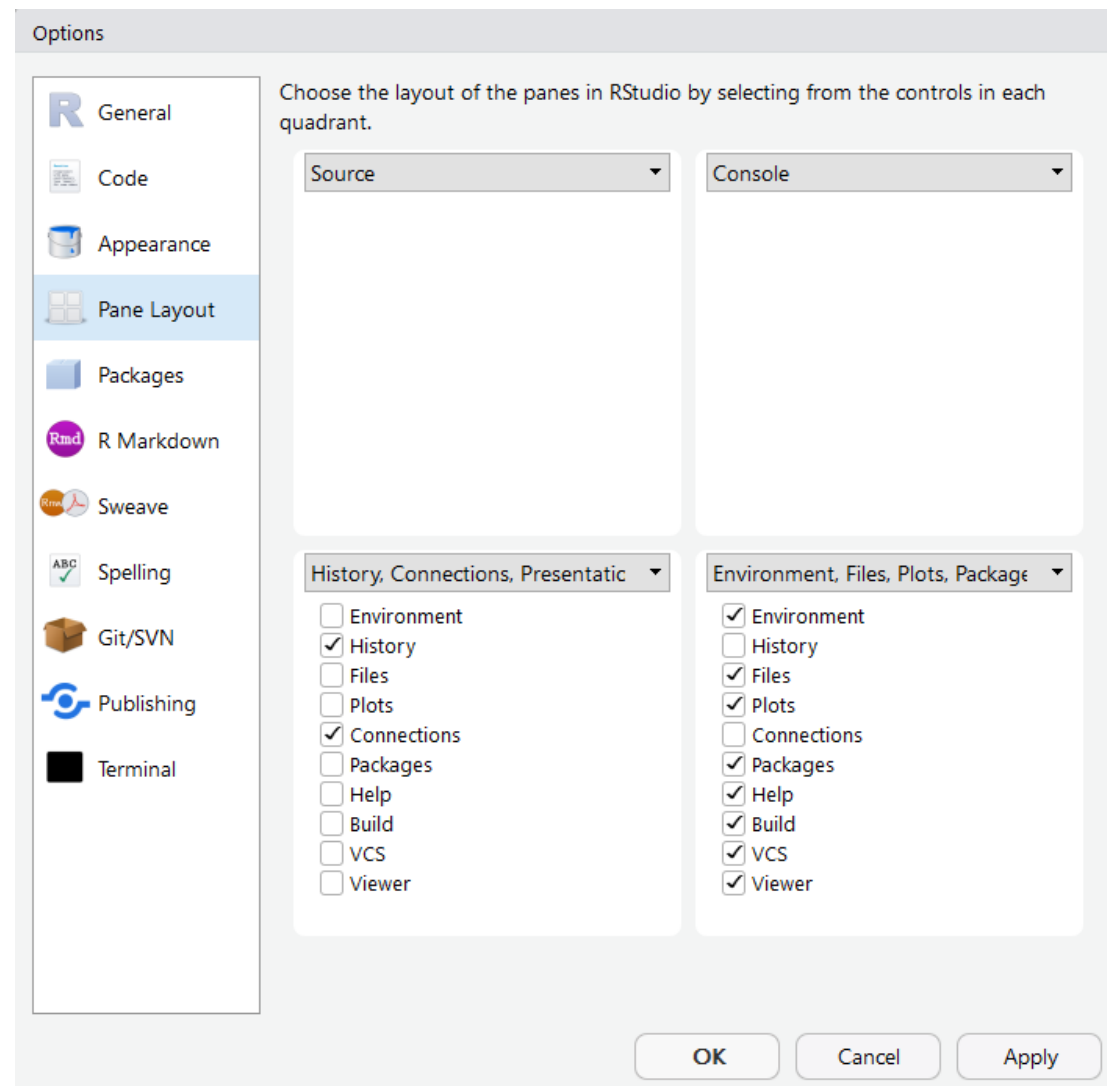
Let us begin!

# Setting up RStudio: Suggested Pane Layout



The screenshot shows the RStudio interface with the 'Tools' menu open. The 'Global Options...' option is highlighted at the bottom of the menu. The background shows a code editor with R code for clearing the environment and loading packages.

```
1 rm(list=ls(all=TRUE)) #clear
2 cat("\f") # clear console whe
3
4 # load-sources -----
5 # call `base::source()` on an
6
7 # load-packages -----
8 # Attach these packages so th
9 library("magrittr") #Pipes
10 library("ggplot2") #graphs
11 library("dplyr")
12
13 # we will pretend to work with raw data, but this is res source
14 # titanic::titanic_train %>% readr::write_csv("./data-unshared/
15 # titanic::titanic_test %>% readr::write_csv("./data-unshared/
16
17 # declare-globals -----
```



**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

andkov / hsif-2019-data-analysis

Unwatch 1 Star 0 Fork 0

Code Issues 0 Pull requests 0 Actions Projects 0 Wiki Security Insights Settings

Managing Data Analysis in RStudio using Project-Oriented workflow

Manage topics

9 commits 2 branches 0 packages 0 releases 1 contributor GPL-2.0

Branch: master New pull request

Create new file Upload files Find file Clone or download

andkov Update README.md

analysis	tabula rasa	
data-public	tabula rasa	
data-unshared	tabula rasa	
libs	tabula rasa	
manipulation	tabula rasa	12 days ago
scripts	tabula rasa	12 days ago
.gitignore	tabula rasa	12 days ago
LICENSE	tabula rasa	12 days ago
README.md	Update README.md	12 days ago
config.yml	tabula rasa	12 days ago
hsif-2019-data-analysis.Rproj	tabula rasa	12 days ago

Clone with HTTPS ⓘ Use SSH

Use Git or checkout with SVN using the web URL.

<https://github.com/andkov/hsif-2019-data>

Open in Desktop Download ZIP

# Folder Architecture

andkov / hsif-2019-data-analysis

Unwatch 1 Star 0 Fork 0

Code Issues 0 Pull requests 0 Actions Projects 0 Wiki Security Insights Settings

Managing Data Analysis in RStudio using Project-Oriented workflow [Edit](#)

[Manage topics](#)

9 commits 2 branches 0 packages 0 releases 1 contributor GPL-2.0

Branch: master New pull request

Create new file Upload files Find file Clone or download

andkov Update README.md

analysis	tabula rasa	
data-public	tabula rasa	
data-unshared	tabula rasa	
libs	tabula rasa	
manipulation	tabula rasa	12 days ago
scripts	tabula rasa	12 days ago
.gitignore	tabula rasa	12 days ago
LICENSE	tabula rasa	12 days ago
README.md	Update README.md	12 days ago
config.yml	tabula rasa	12 days ago
hsif-2019-data-analysis.Rproj	tabula rasa	12 days ago

Clone with HTTPS Use SSH

Use Git or checkout with SVN using the web URL.

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

Open in Desktop Download ZIP

- analysis
- data-public
- data-unshared
- libs
- manipulation
- scripts

R

config.yml

✓ R hsif-2019-data-analysis

LICENSE

R README



# Learning Resources

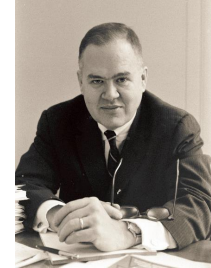
- Rmarkdown guide (<https://rmarkdown.rstudio.com/>)
- Logistic regression (Youtube: StatQuest + Logistic Regression)
- R4DS (<https://r4ds.had.co.nz/>) + swirl (<https://swirlstats.com/>)
- Introduction to ggplot2 (<http://www.cookbook-r.com/Graphs/>)

# Lessons & Metaphors

- Handle your data! (Vesalius)



- Look at your data! (Tukey)



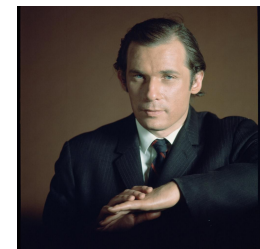
- Graph is art (Tufte)



- Graph is language (Wickham)



- Coding is music (Gould)



# Lessons & Metaphors

- Handle your data! (Vesalius)



- Look at your data! (Tukey)



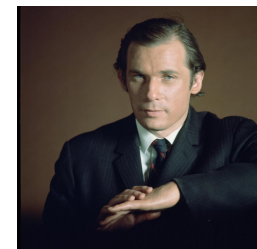
- Graph is art (Tufte)



- Graph is language (Wickham)



- Coding is music (Gould)



# Questions? Comments?



Andriy Koval

<https://github.com/andkov>

<http://andriy.rbind.io>