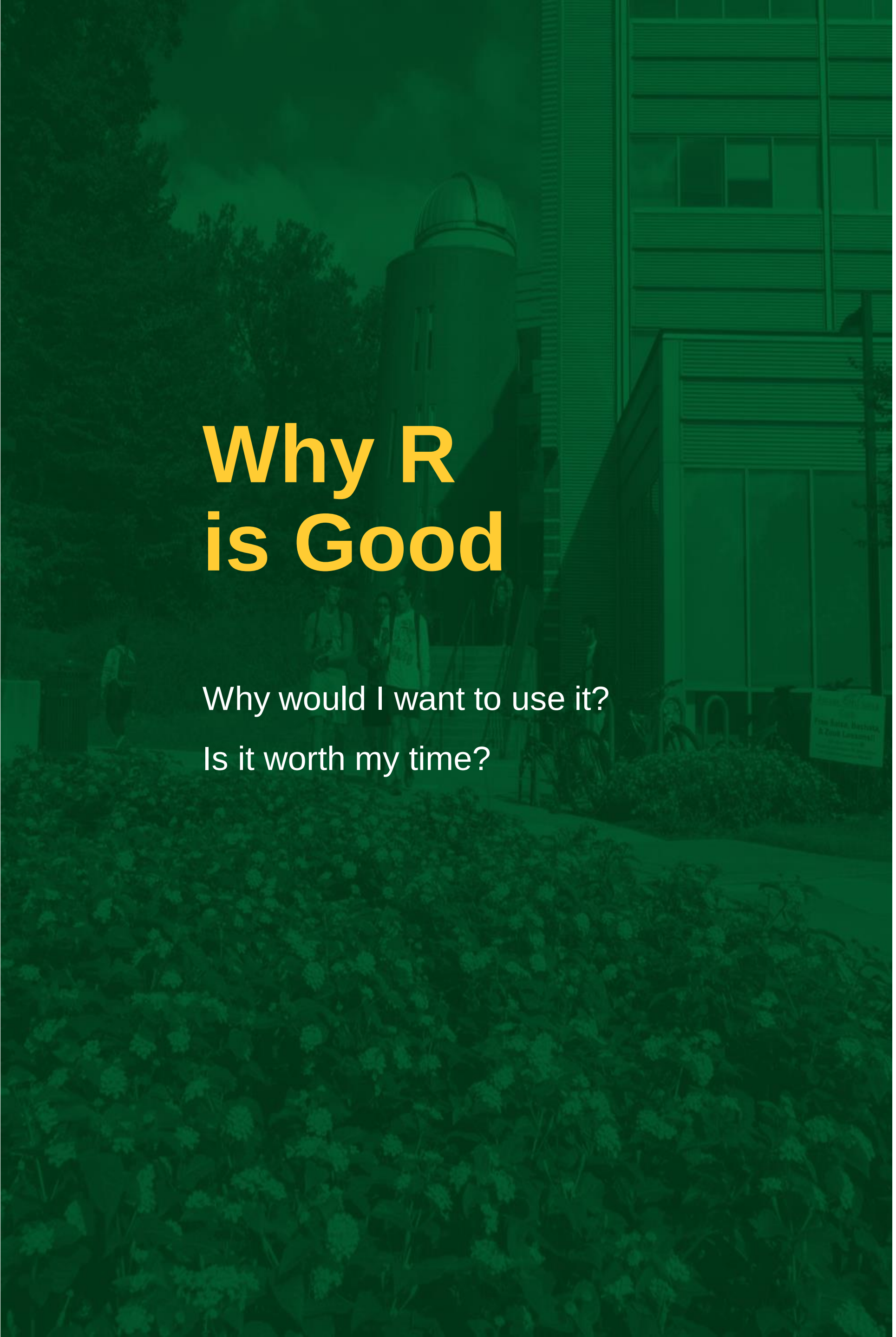
R for Data

How to Get Started

How do you currently use data?

What is your interest in R?

GEORGE MASON UNIVERSITY

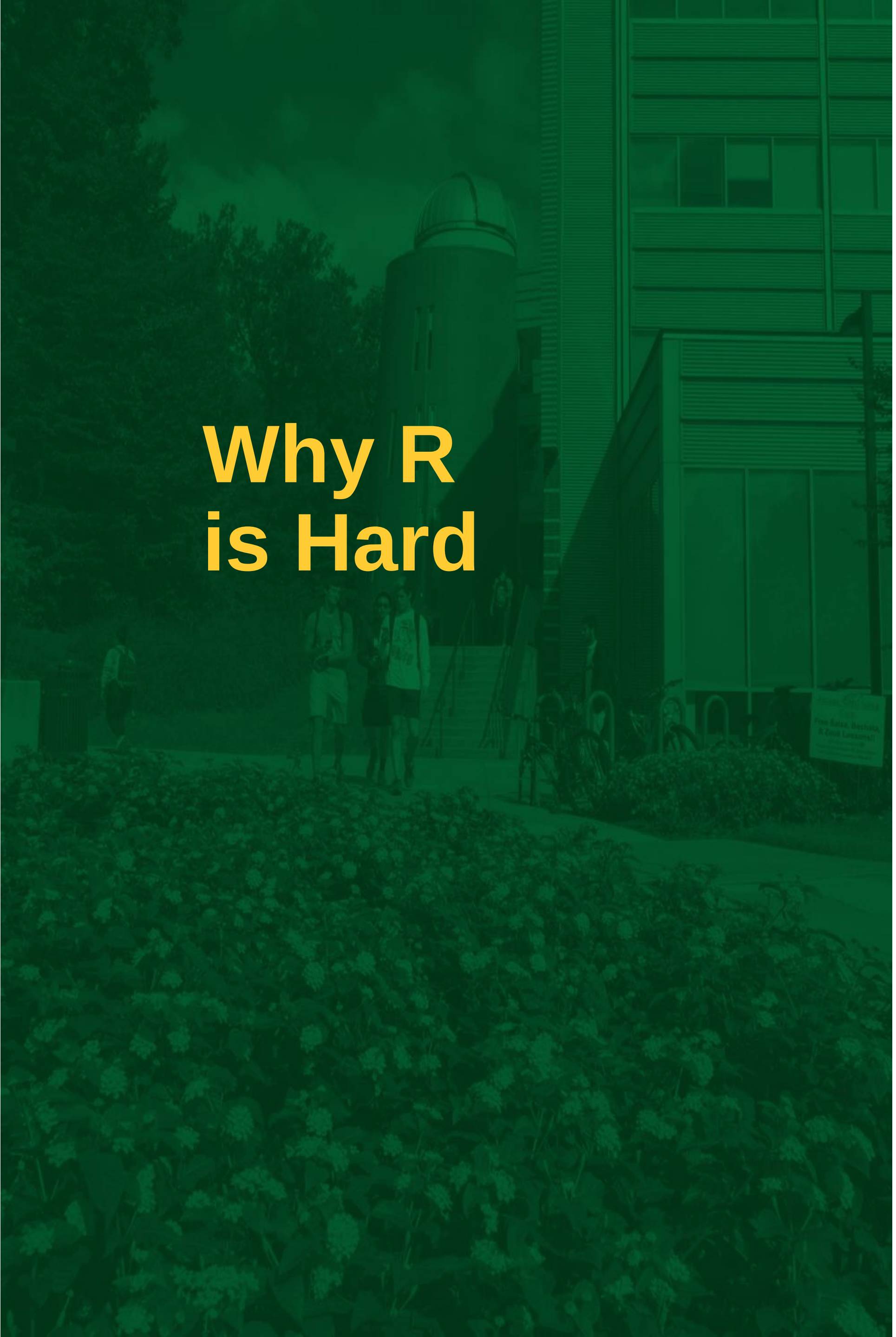


Why R is Good

Why would I want to use it?

Is it worth my time?

- Free
  - Can be used anywhere, for any purpose
- Syntax = Reproducibility
  - Statistical software also has syntax
- Data Notebook Capable
  - Easily write data documents
- Unlimited Capabilities
  - Easy to do complex things



Why R is Hard

- Text-based computer language
  - must know what you want
  - must use exact syntax

- Constantly changing, huge ecosystem
  - albeit improving

- Many ways to do the same thing
  - Many people add functionality
  - Too may tutorials

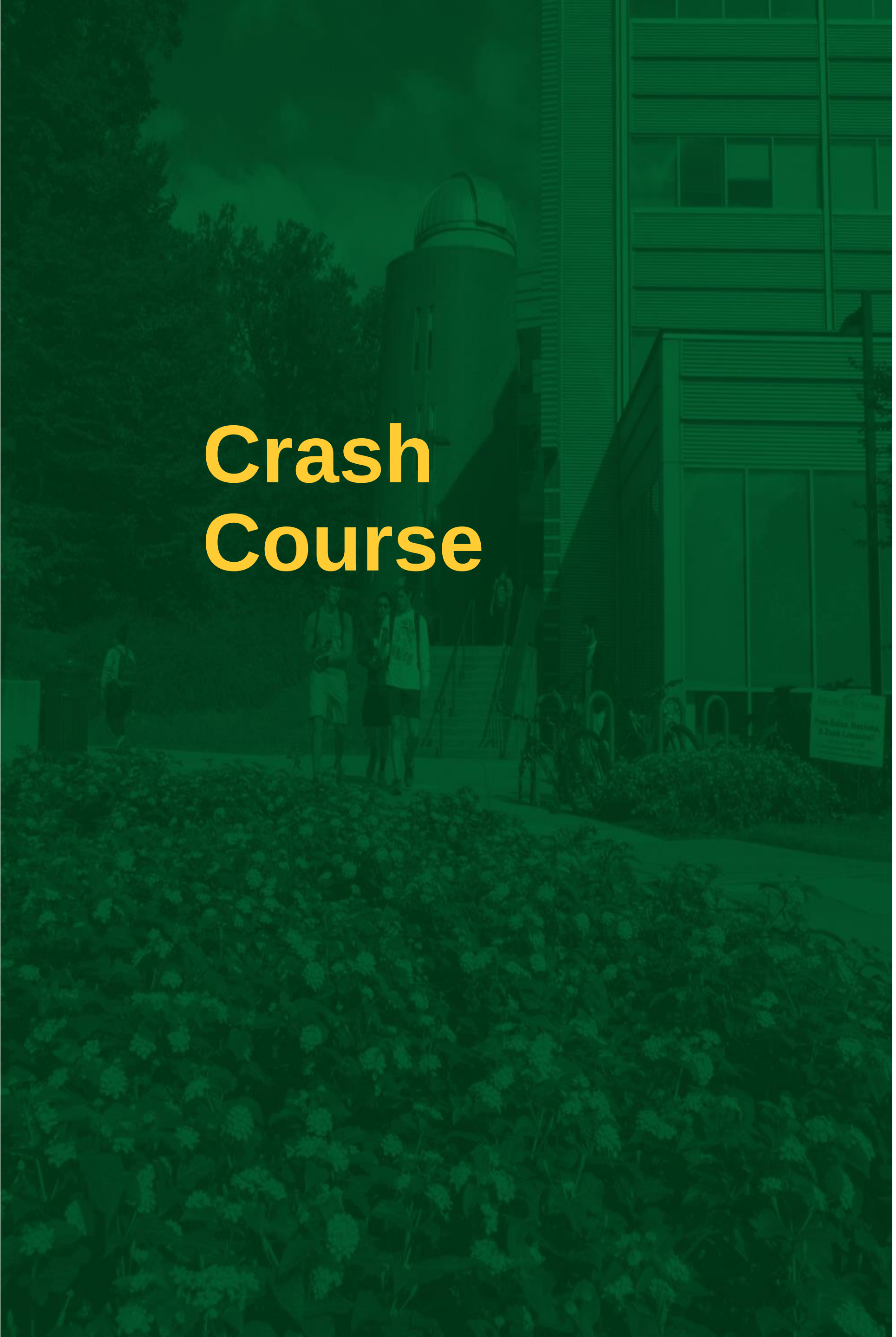
GEORGE MASON UNIVERSITY

Changes over time

- **RStudio → Posit**
- **R Markdown → Quarto Document**
- **%>% → |>**
- **plyr → dplyr**
- **reshape2 → tidyr**

Low on Time?

- Utilize Graphical User Interfaces
- Start learning basic programming concepts
- Use it just for specific tasks



Crash Course

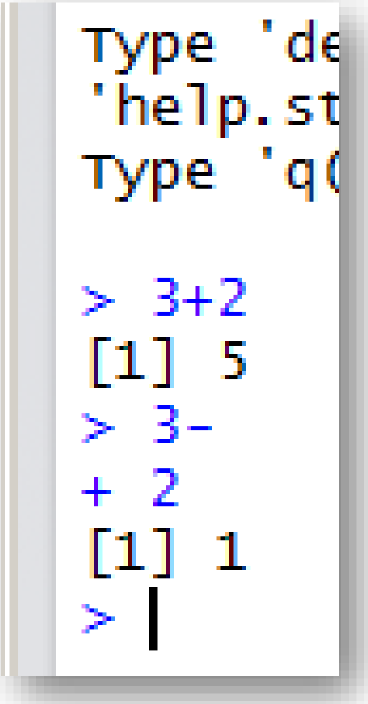
If you know these already, then you are in great shape.

Programmers often find R confusing / harder.

Easier to teach R to non-programmers

Principles: Console vs Script

>	prompt
+	waiting for more
[1]	returned element

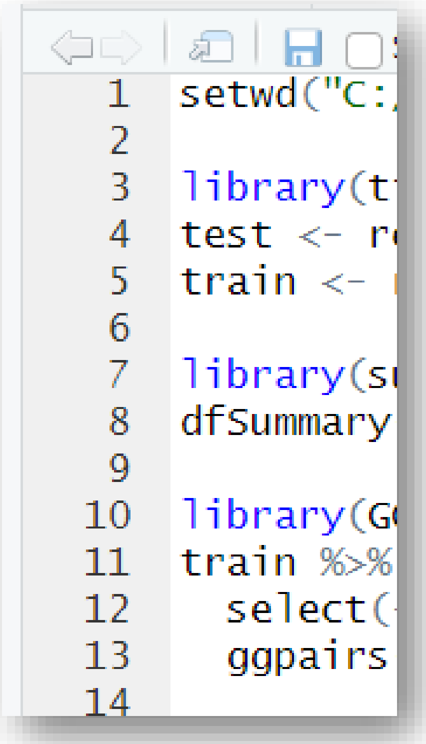


```
Type 'de
'help.st
Type 'q(

> 3+2
[1] 5
> 3-
+ 2
[1] 1
> 
```

Code separate, saves in file

"Run" or "Execute" code

# is a comment

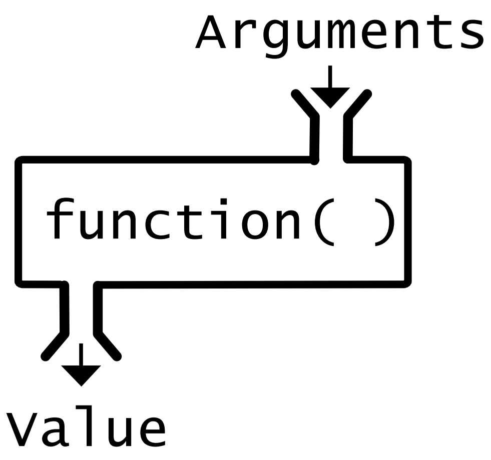


```
setwd("C:/

library(t
test <- r
train <- 

library(s
dfSummary

library(G
train %>%
  select(
  ggpairs

```

Principles

Programmers call these variables. But columns in a data table are variables to statisticians.

Functions



```
sum(2,4)
countif(A2:A50,">5")
read.table(mydata, header=TRUE, sep=",")
```

Objects

- Numbers
- Characters

Data Types

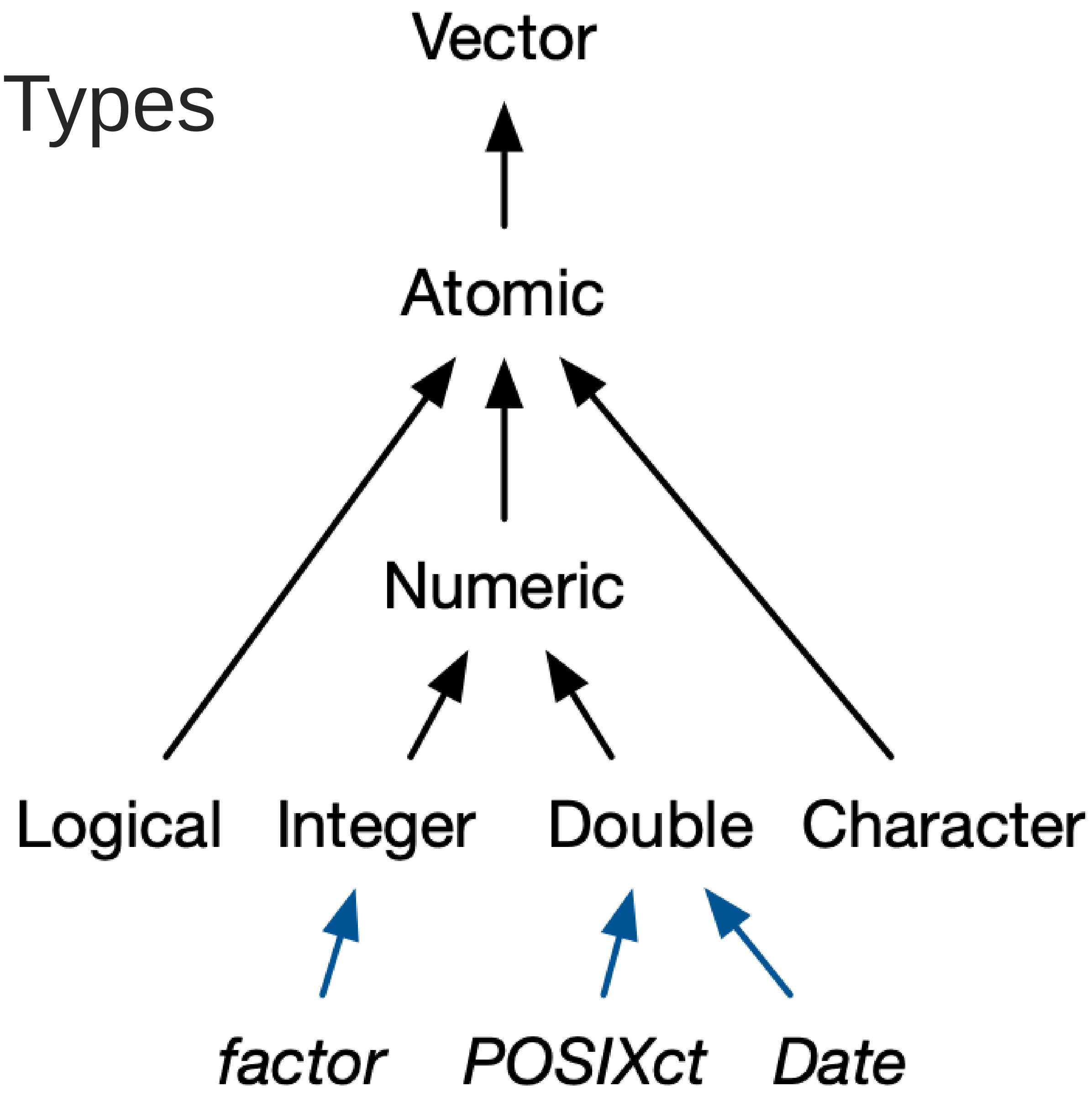
- Vectors
- Datasets

Data Structures

```
people <- 9
school <- "GMU"
```

Assignment Operator

Data Types



Vector

Atomic

Numeric

Logical Integer Double Character

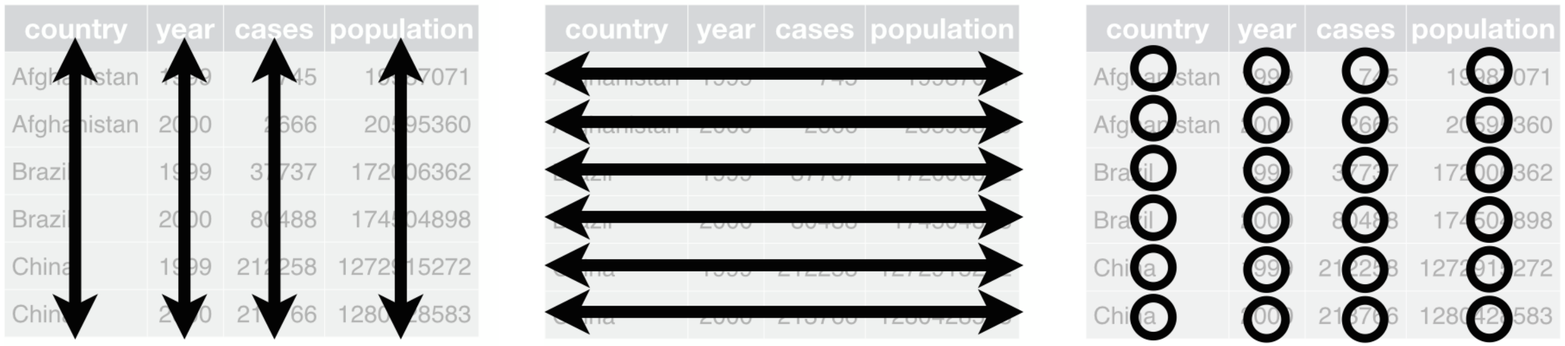
factor *POSIXct* *Date*

https://adv-r.hadley.nz/vectors-chap.html

Data Structures

	Homogeneous (same Type)	Heterogeneous (different Types)
1 Dimension	**Vector**	**List**
2 Dimensions	**Matrix**	**Data Frame**

Principles: Data Tables & Tidy Data



variables

observations

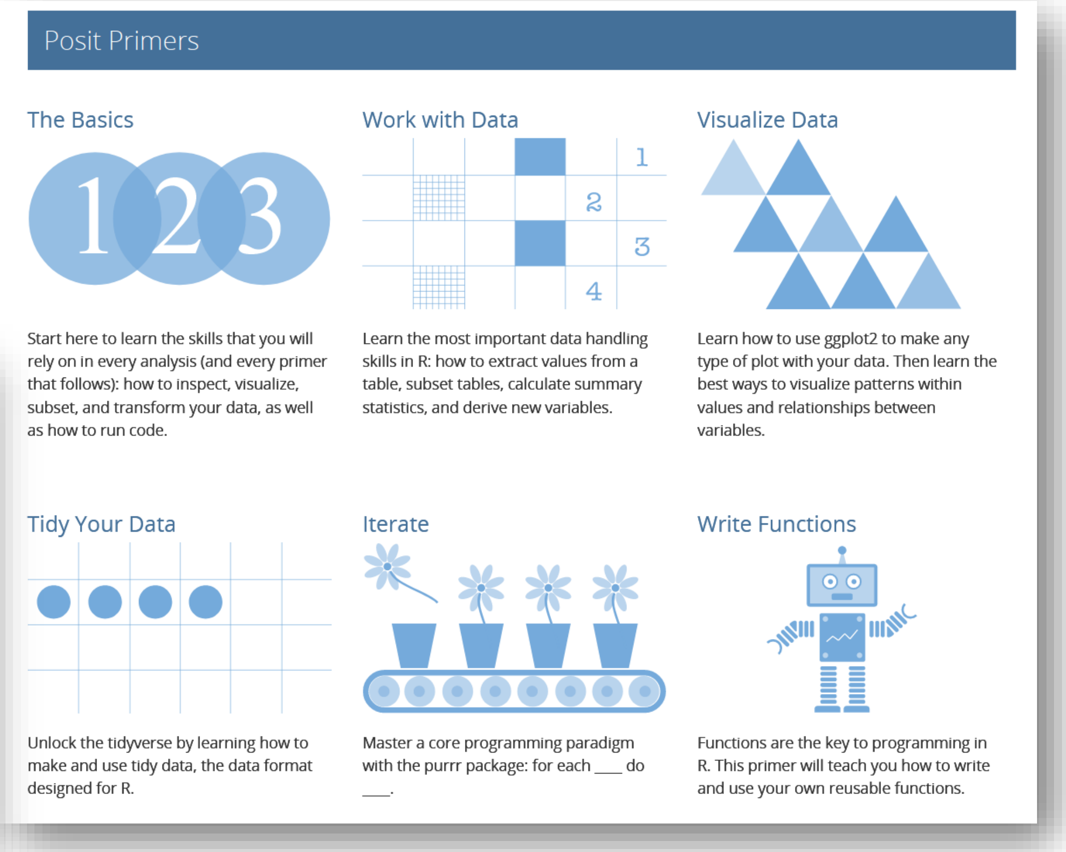
values

https://r4ds.had.co.nz/tidy-data.html

Programming Concepts that Help

- **Data Types** (e.g., Numeric, String)
- **Data Structures** (e.g., Lists, Dictionaries)
- **Variables** and **Objects**
- **Conditional** Statements (If Statements)
- **Functions, Methods**, and **Modules**
- **Working Directory** and **File Paths**

Quick Start: Posit Primers & learnr Tutorials



Posit Primers

The Basics

Start here to learn the skills that you will rely on in every analysis (and every primer that follows): how to inspect, visualize, subset, and transform your data, as well as how to run code.

Work with Data

Learn the most important data handling skills in R: how to extract values from a table, subset tables, calculate summary statistics, and derive new variables.

Visualize Data

Learn how to use ggplot2 to make any type of plot with your data. Then learn the best ways to visualize patterns within values and relationships between variables.

Tidy Your Data

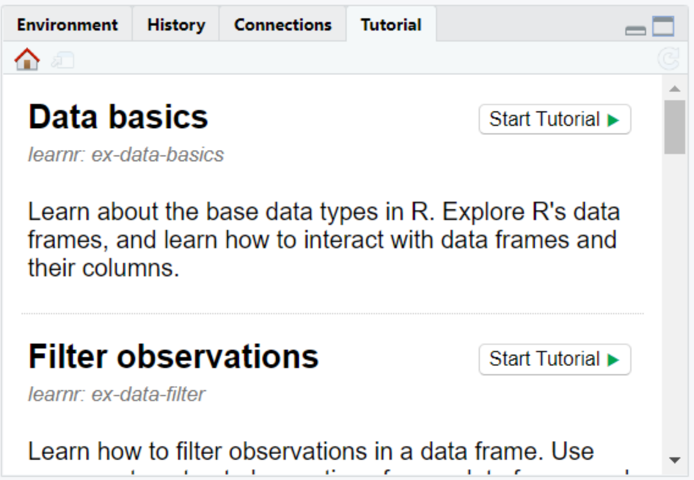
Unlock the tidyverse by learning how to make and use tidy data, the data format designed for R.

Iterate

Master a core programming paradigm with the purrr package: for each ___ do ___.

Write Functions

Functions are the key to programming in R. This primer will teach you how to write and use your own reusable functions.



Data basics

learnr: ex-data-basics

Learn about the base data types in R. Explore R's data frames, and learn how to interact with data frames and their columns.

Filter observations

learnr: ex-data-filter

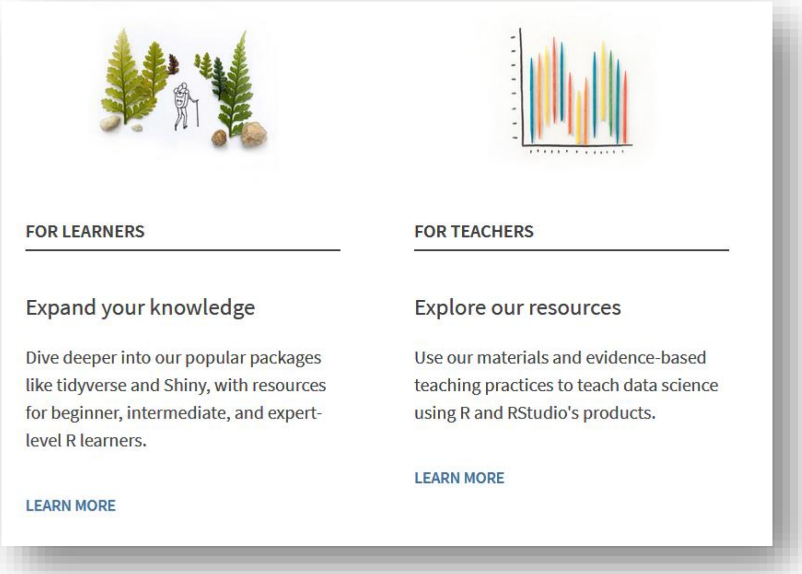
Learn how to filter observations in a data frame. Use

https://posit.cloud/learn/primers

https://rstudio.github.io/learnr/articles/examples.html

RStudio Education

https://education.rstudio.com/



FOR LEARNERS

Expand your knowledge

Dive deeper into our popular packages like tidyverse and Shiny, with resources for beginner, intermediate, and expert-level R learners.

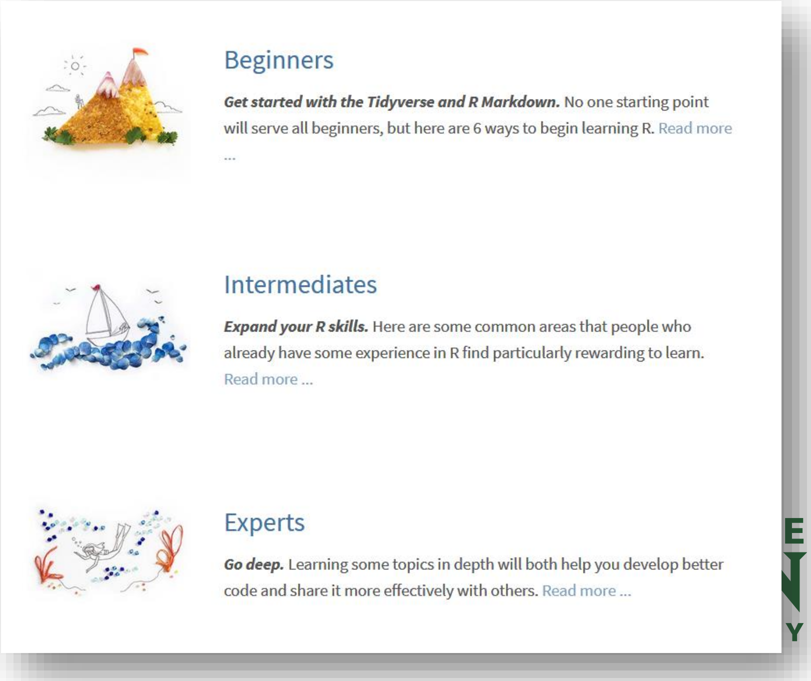
LEARN MORE

FOR TEACHERS

Explore our resources

Use our materials and evidence-based teaching practices to teach data science using R and RStudio's products.

LEARN MORE



Beginners

Get started with the Tidyverse and R Markdown. No one starting point will serve all beginners, but here are 6 ways to begin learning R. Read more ...

Intermediates

Expand your R skills. Here are some common areas that people who already have some experience in R find particularly rewarding to learn. Read more ...

Experts

Go deep. Learning some topics in depth will both help you develop better code and share it more effectively with others. Read more ...

What do I need to get started?

THE SOFTWARE

AN INTERFACE

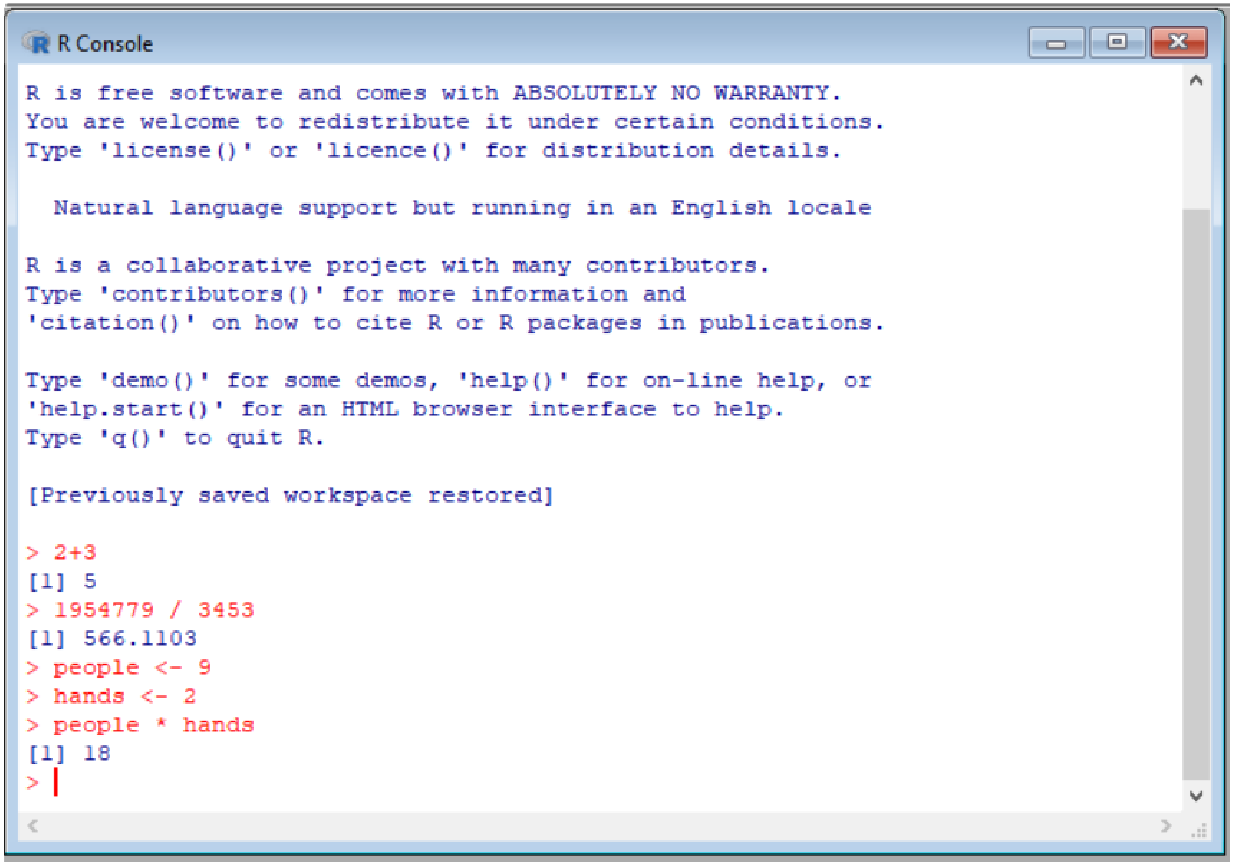
PACKAGES & FUNCTIONS

https://learnr-examples.shinyapps.io/ex-setup-r/

GEORGE MASON UNIVERSITY

The Software

- https://cran.r-project.org/

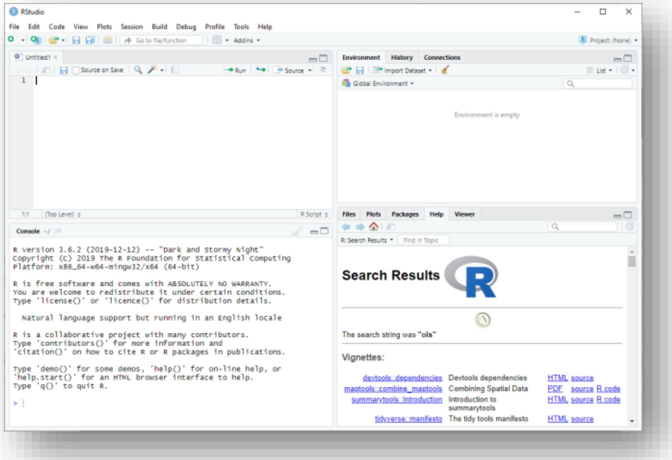


```
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Previously saved workspace restored]

> 2+3
[1] 5
> 1954779 / 3453
[1] 566.1103
> people <- 9
> hands <- 2
> people * hands
[1] 18
> 
```

An Interface – RStudio

Desktop

https://dsc.gmu.edu/files/Installing-R-RStudio.pdf



```
R version 3.6.2 (2019-12-12) -- "Dark and Stormy Night"
Copyright (C) 2019 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> 
```

Search Results

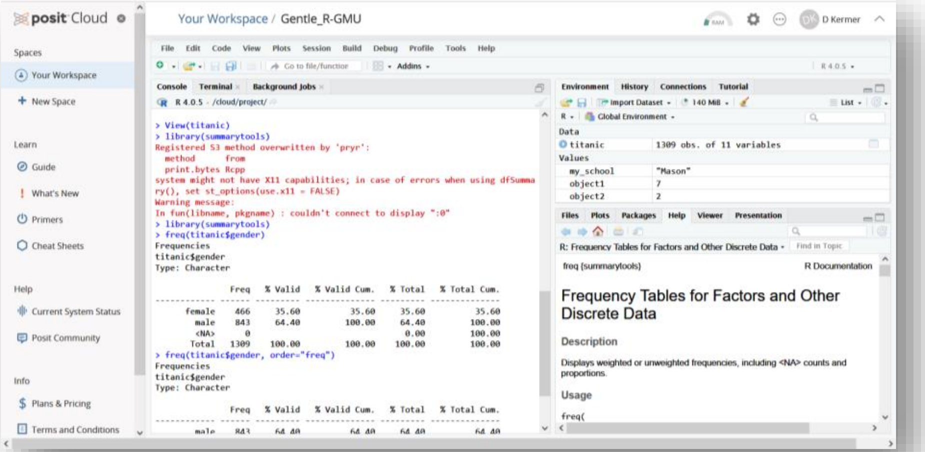
The search string was "ols"

Vignettes:

devtools::dependencies	Devtools dependencies	HTML	source	
maptools::combine_maptools	Combining Spatial Data	PDF	source	R code
summarytools::Introduction	Introduction to summarytools	HTML	source	R code
tidyverse::manifesto	The tidy tools manifesto	HTML	source	

Cloud

https://posit.cloud


posit Cloud
Spaces
Your Workspace
New Space
Learn
Guide
What's New
Primers
Cheat Sheets
Help
Current System Status
Posit Community
Info
Plans & Pricing
Terms and Conditions
Your Workspace / Gentle_R-GMU
RAM
DK D Kermer
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
R 4.0.5
Console Terminal Background Jobs
R 4.0.5 · /cloud/project/
Environment History Connections Tutorial
Import Dataset
140 MiB
List
R
Global Environment
Files Plots Packages Help Viewer Presentation
R: Frequency Tables for Factors and Other Discrete Data
Find in Topic

```
> View(titanic)
> library(summarytools)
Registered S3 method overwritten by 'pryr':
  method      from
  print.bytes Rcpp
system might not have X11 capabilities; in case of errors when using dfSumma
ry(), set st_options(use.x11 = FALSE)
Warning message:
In fun(libname, pkgname) : couldn't connect to display ":0"
> library(summarytools)
> freq(titanic$gender)
Frequencies
titanic$gender
Type: Character

              Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
------------ ------ --------- -------------- --------- --------------
      female    466     35.60          35.60     35.60          35.60
        male    843     64.40         100.00     64.40         100.00
        <NA>      0                               0.00         100.00
       Total   1309    100.00         100.00    100.00         100.00
> freq(titanic$gender, order="freq")
Frequencies
titanic$gender
Type: Character

              Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
------------ ------ --------- -------------- --------- --------------
        male    843     64.40          64.40     64.40          64.40
```

Data

titanic	1309 obs. of 11 variables

Values

my_school	"Mason"
object1	7
object2	2

freq {summarytools} R Documentation

Frequency Tables for Factors and Other Discrete Data

Description

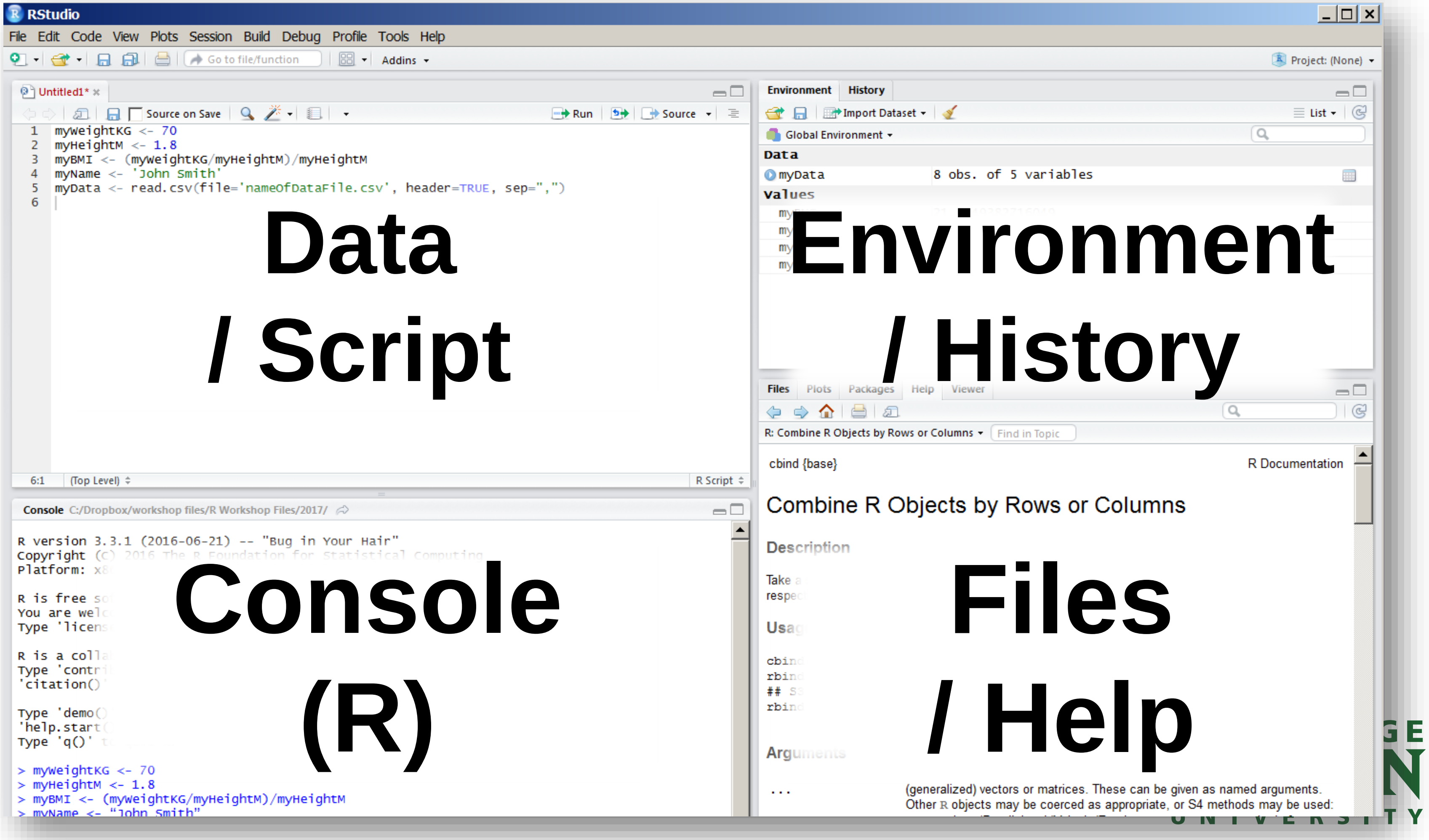
Displays weighted or unweighted frequencies, including <NA> counts and proportions.

Usage

freq(

RStudio → Posit

RStudio Sections



```
myWeightKG <- 70
myHeightM <- 1.8
myBMI <- (myWeightKG/myHeightM)/myHeightM
myName <- 'John Smith'
myData <- read.csv(file='nameOfDataFile.csv', header=TRUE, sep=",")
```

Data

/ Script

```
Data
myData                8 obs. of 5 variables
Values
my
my
my
my
```

Environment

/ History

```
R version 3.3.1 (2016-06-21) -- "Bug in Your Hair"
Copyright (C) 2016 The R Foundation for Statistical Computing
Platform: x8

R is free s
You are wel
Type 'licen

R is a coll
Type 'contr
'citation()

Type 'demo(
'help.start(
Type 'q()' t

> myWeightKG <- 70
> myHeightM <- 1.8
> myBMI <- (myWeightKG/myHeightM)/myHeightM
> myName <- "John Smith"
```

Console

(R)

cbind {base} R Documentation

Combine R Objects by Rows or Columns

Description

Take a
respe

Usag

```
cbin
rbin
## S3
rbin
```

Arguments

... (generalized) vectors or matrices. These can be given as named arguments. Other R objects may be coerced as appropriate, or S4 methods may be used:

Files

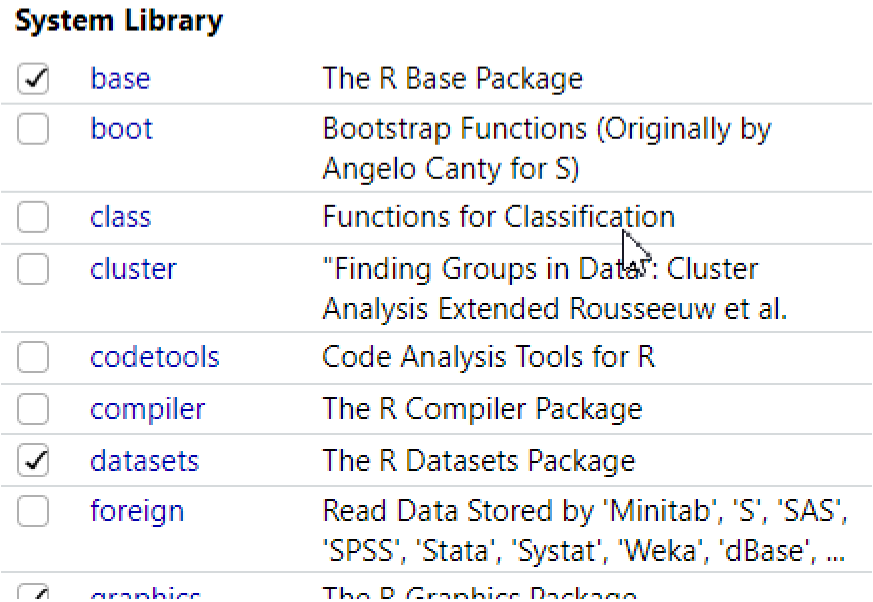
/ Help

Benefits of RStudio

- One window to contain everything
- Projects help you contain projects and use a working directory
- Point-and-click for simple tasks (import files, install packages)
- See objects in environment.
- Autocompletion for packages, functions, and objects
- Help documentation at your fingertips
- Easy to connect to a git repository.

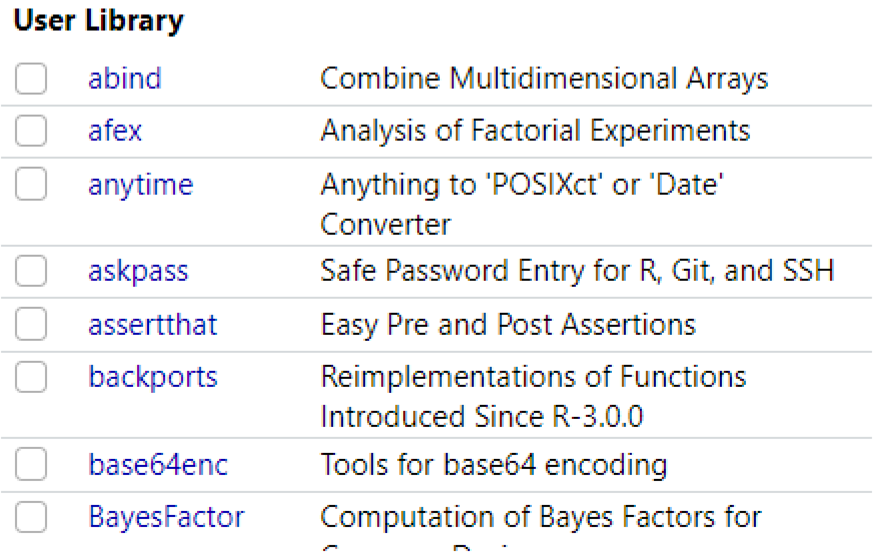
Packages

- Packages = Groups of functions
- Some are built-in to R
- Most are written by researchers
  - Make it easier to do something they want to do
  - Must be installed, just once
- Packages may have functions with the same name!
  - Must tell R which package to use.
  - `library(`**`package`**`)`
  - `require(`**`package`**`)`
  - **`package`**`::function()`



System Library

☑	base	The R Base Package
□	boot	Bootstrap Functions (Originally by Angelo Canty for S)
□	class	Functions for Classification
□	cluster	"Finding Groups in Data": Cluster Analysis Extended Rousseeuw et al.
□	codetools	Code Analysis Tools for R
□	compiler	The R Compiler Package
☑	datasets	The R Datasets Package
□	foreign	Read Data Stored by 'Minitab', 'S', 'SAS', 'SPSS', 'Stata', 'Systat', 'Weka', 'dBase', ...
☑	graphics	The R Graphics Package



User Library

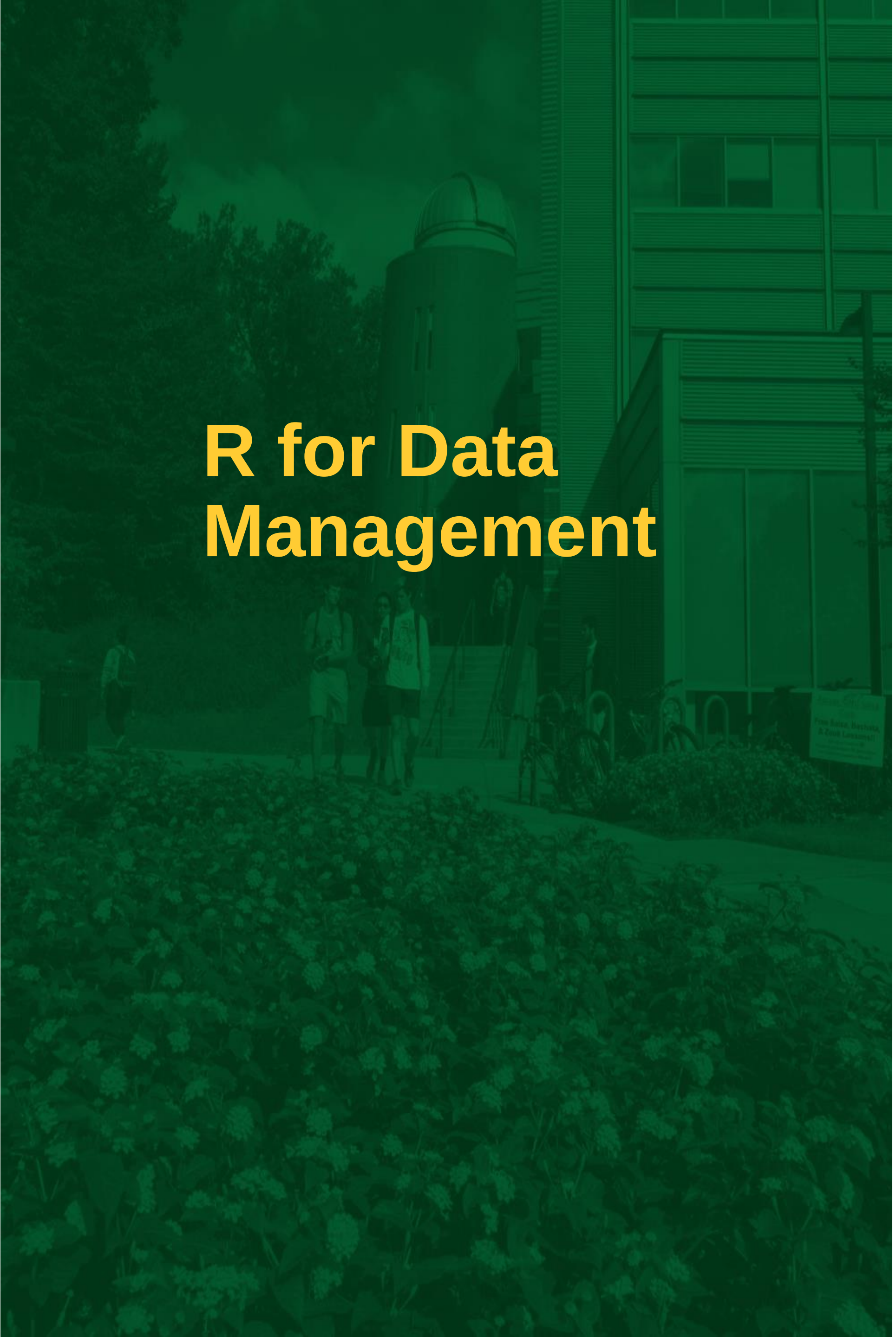
□	abind	Combine Multidimensional Arrays
□	afex	Analysis of Factorial Experiments
□	anytime	Anything to 'POSIXct' or 'Date' Converter
□	askpass	Safe Password Entry for R, Git, and SSH
□	assertthat	Easy Pre and Post Assertions
□	backports	Reimplementations of Functions Introduced Since R-3.0.0
□	base64enc	Tools for base64 encoding
□	BayesFactor	Computation of Bayes Factors for

R's Three Areas:

Data Management

Statistical Modeling

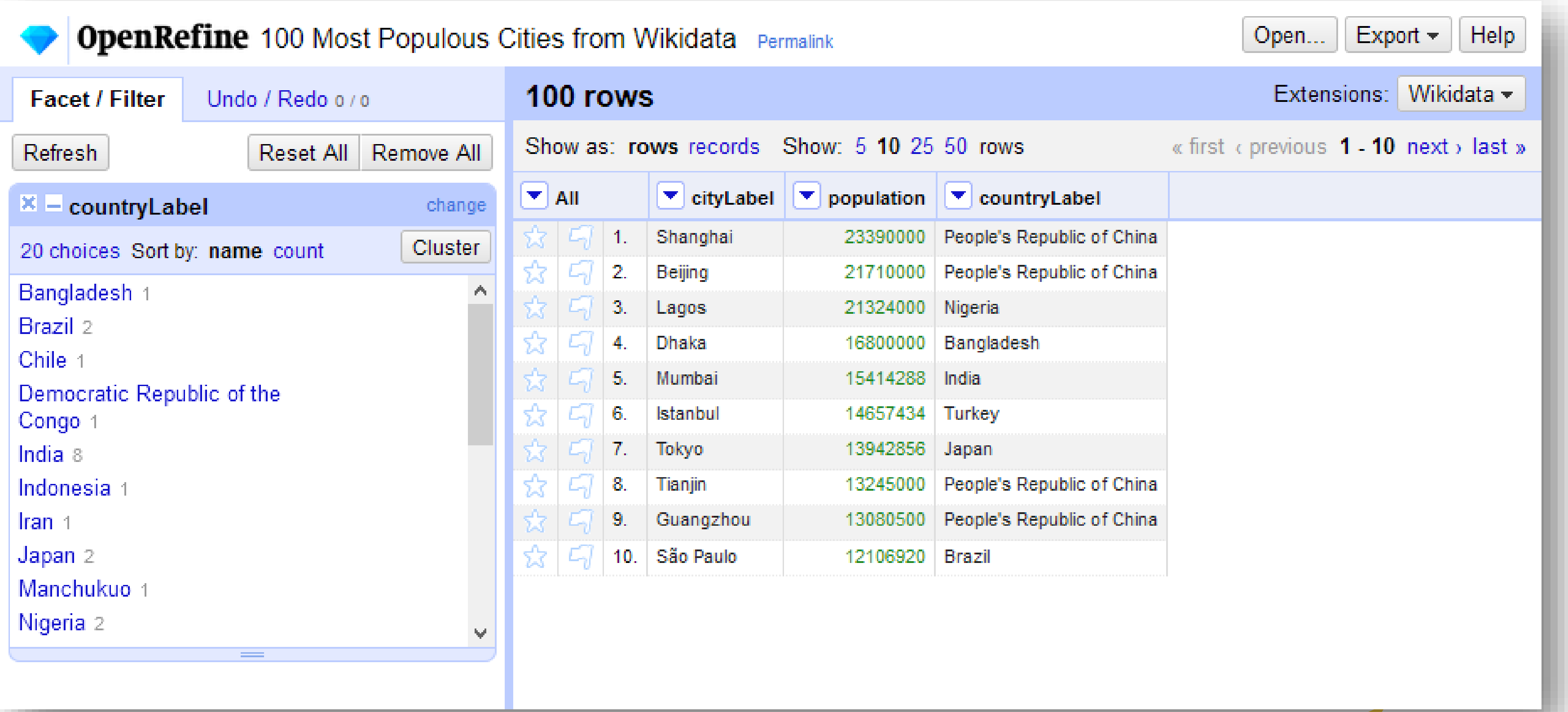
Visualization



R for Data Management

- Import Data
- Tidy Data
- Transform Data

Open Refine - https://openrefine.org/



OpenRefine 100 Most Populous Cities from Wikidata Permalink

100 rows

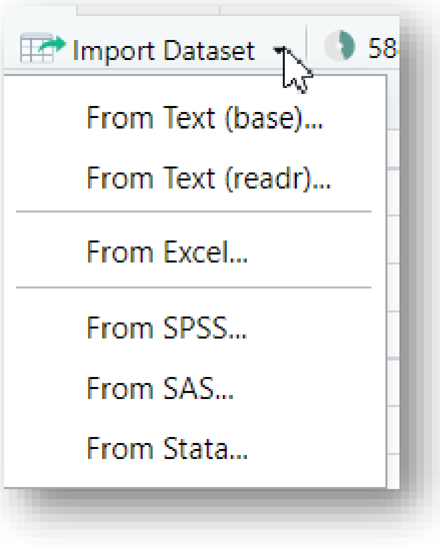
Country	Count
Bangladesh	1
Brazil	2
Chile	1
Democratic Republic of the Congo	1
India	8
Indonesia	1
Iran	1
Japan	2
Manchukuo	1
Nigeria	2

All	cityLabel	population	countryLabel
1.	Shanghai	23390000	People's Republic of China
2.	Beijing	21710000	People's Republic of China
3.	Lagos	21324000	Nigeria
4.	Dhaka	16800000	Bangladesh
5.	Mumbai	15414288	India
6.	Istanbul	14657434	Turkey
7.	Tokyo	13942856	Japan
8.	Tianjin	13245000	People's Republic of China
9.	Guangzhou	13080500	People's Republic of China
10.	São Paulo	12106920	Brazil

GEORGE MASON UNIVERSITY

Importing Data

- CSV / Delimited
- Excel
- SPSS, Stata, SAS
- json, xml
- packages



Import Dataset
58
From Text (base)...
From Text (readr)...
From Excel...
From SPSS...
From SAS...
From Stata...

see other sections

R for Data Management

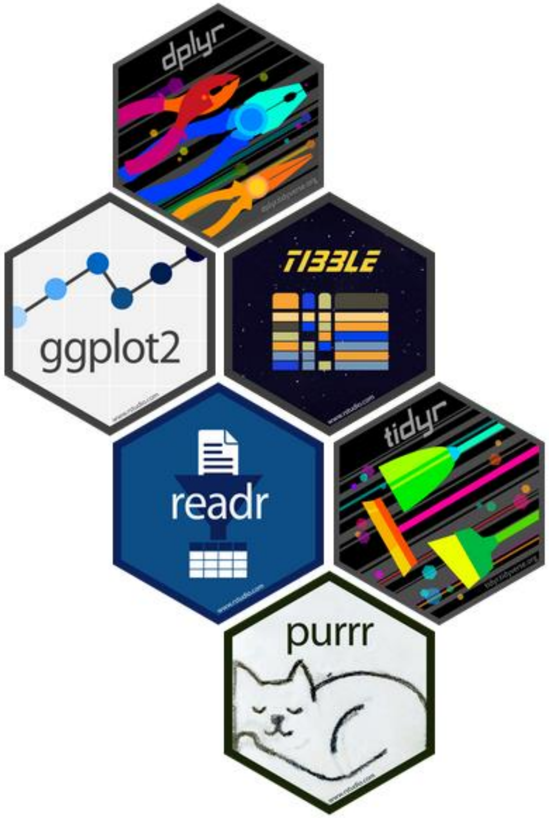
Base R	**Tidyverse**	**Data Table**
data.frame	tibble	data.table
df$word	%>%	dt[.., .., ..]
df[…]	Hadley Wickham	
	RStudio/Posit	

Tidyverse

https://www.tidyverse.org/

dplyr

- Select
- Filter
- Mutate
- Summarize
- Arrange



The pipe

Tidyverse functions take the data as the 1st argument.
This makes it possible to use the pipe.

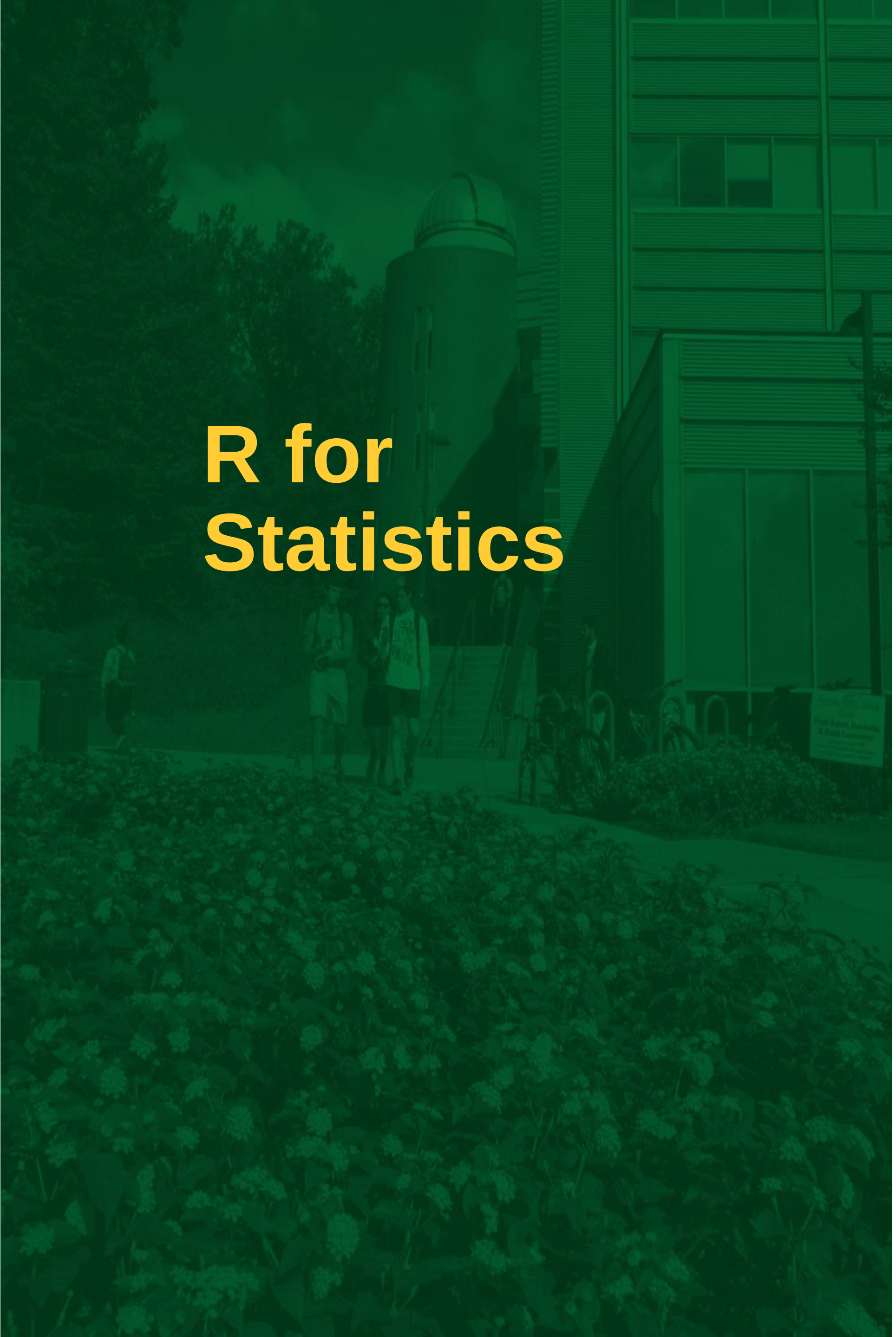
function(**data**, *arguments*)

data %>% *function*(*arguments*)

data >| *function*(*arguments*)

More Tidyverse Packages

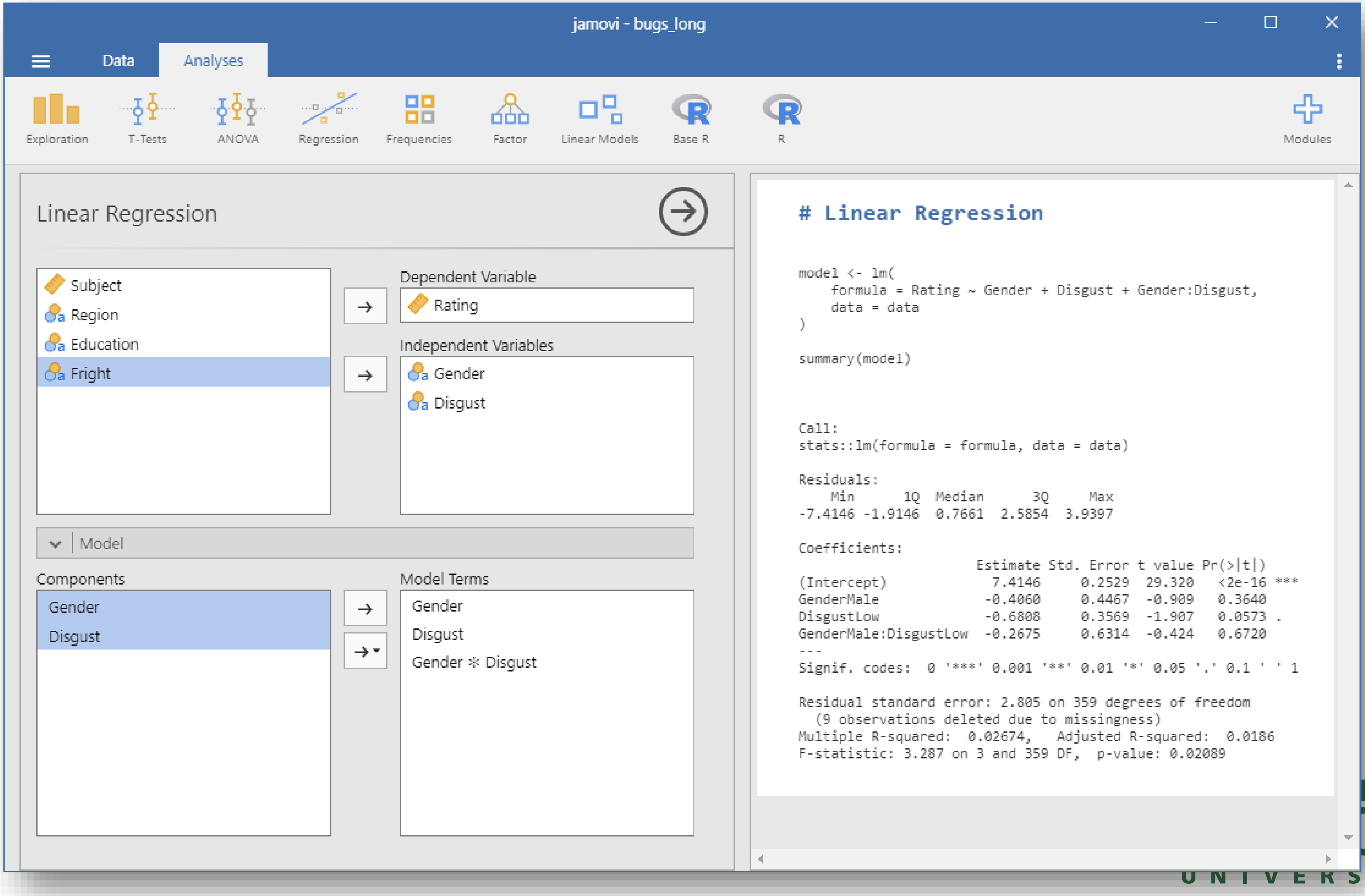
- For working with specific data types



R for Statistics

- Formula Notation
- Model Objects
- Output Formatting

jamovi – https://jamovi.org



Dependent Variable: Rating

Independent Variables: Gender, Disgust

Model Terms:
Gender
Disgust
Gender ✻ Disgust

```
# Linear Regression


model <- lm(
    formula = Rating ~ Gender + Disgust + Gender:Disgust,
    data = data
)

summary(model)



Call:
stats::lm(formula = formula, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.4146 -1.9146  0.7661  2.5854  3.9397 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)             7.4146     0.2529  29.320   <2e-16 ***
GenderMale             -0.4060     0.4467  -0.909   0.3640    
DisgustLow             -0.6808     0.3569  -1.907   0.0573 .  
GenderMale:DisgustLow  -0.2675     0.6314  -0.424   0.6720    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.805 on 359 degrees of freedom
  (9 observations deleted due to missingness)
Multiple R-squared:  0.02674,	Adjusted R-squared:  0.0186 
F-statistic: 3.287 on 3 and 359 DF,  p-value: 0.02089
```

General Linear Models

- Regression /ANOVA tools are part of base R.
- Uses formula notation (mostly the Wilkinson Notation)
- The name of the data table object is given as an argument.

```
myanova   <- aov(       fare ~ class + gender,     data = titanic   )

mylogistic <- glm(survived ~ class * gender,     data = titanic ,
                                                  family = binomial )
```

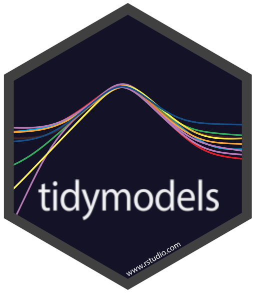
~	predicted from	:	interaction
+	include ("and")	*	factorial

Packages for Statistics

- summarytools – Descriptive Statistics
- afex – ANOVA (also mixed models with lme4)
- emmeans – Postestimation tests
- lme4 – Mixed Models
- lavaan – Structural Equation Modeling

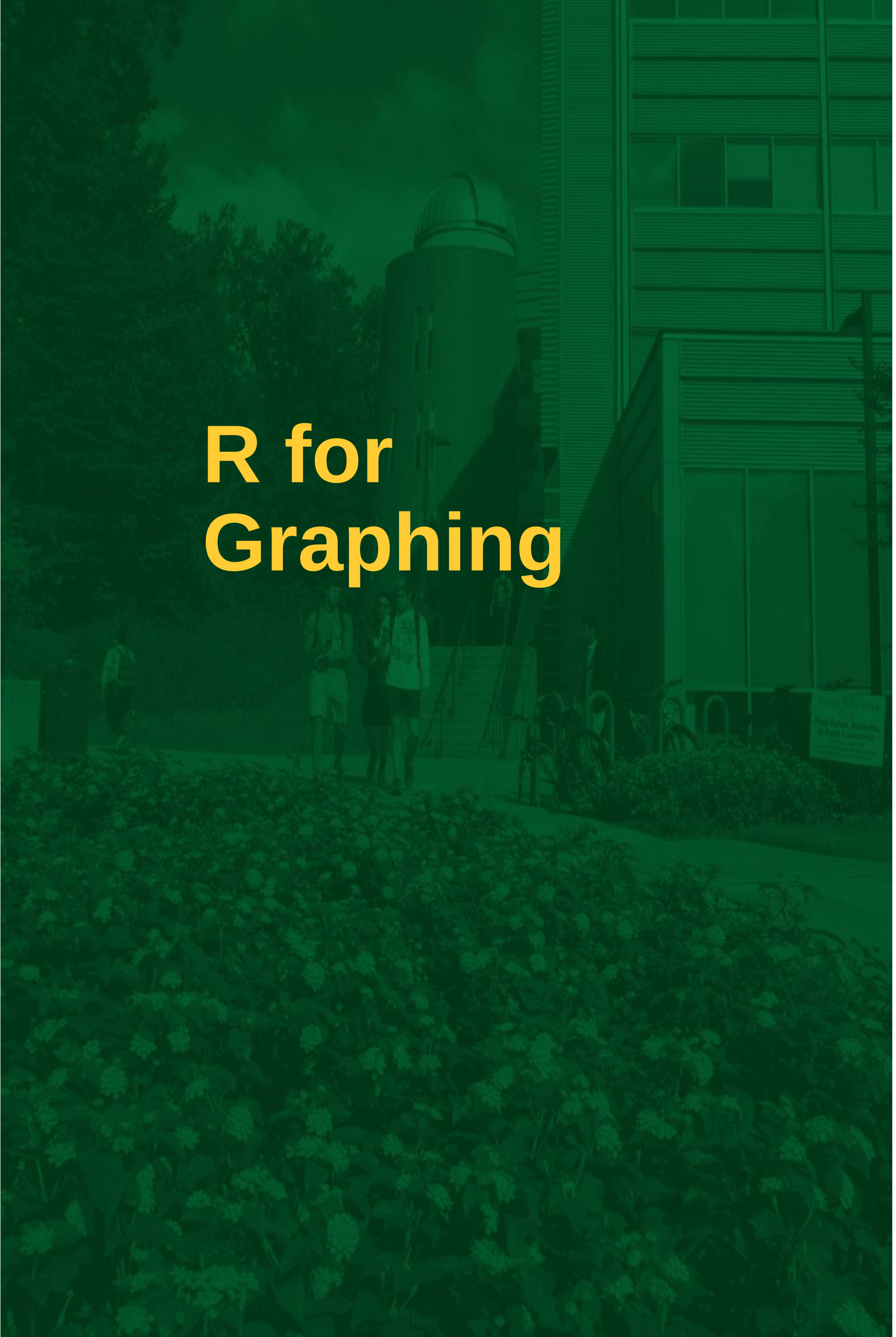
Output

- Functions create an analysis object, not useful by itself
- Use other functions to display tables, plots, etc
- e.g., summary() function for lm / glm / aov



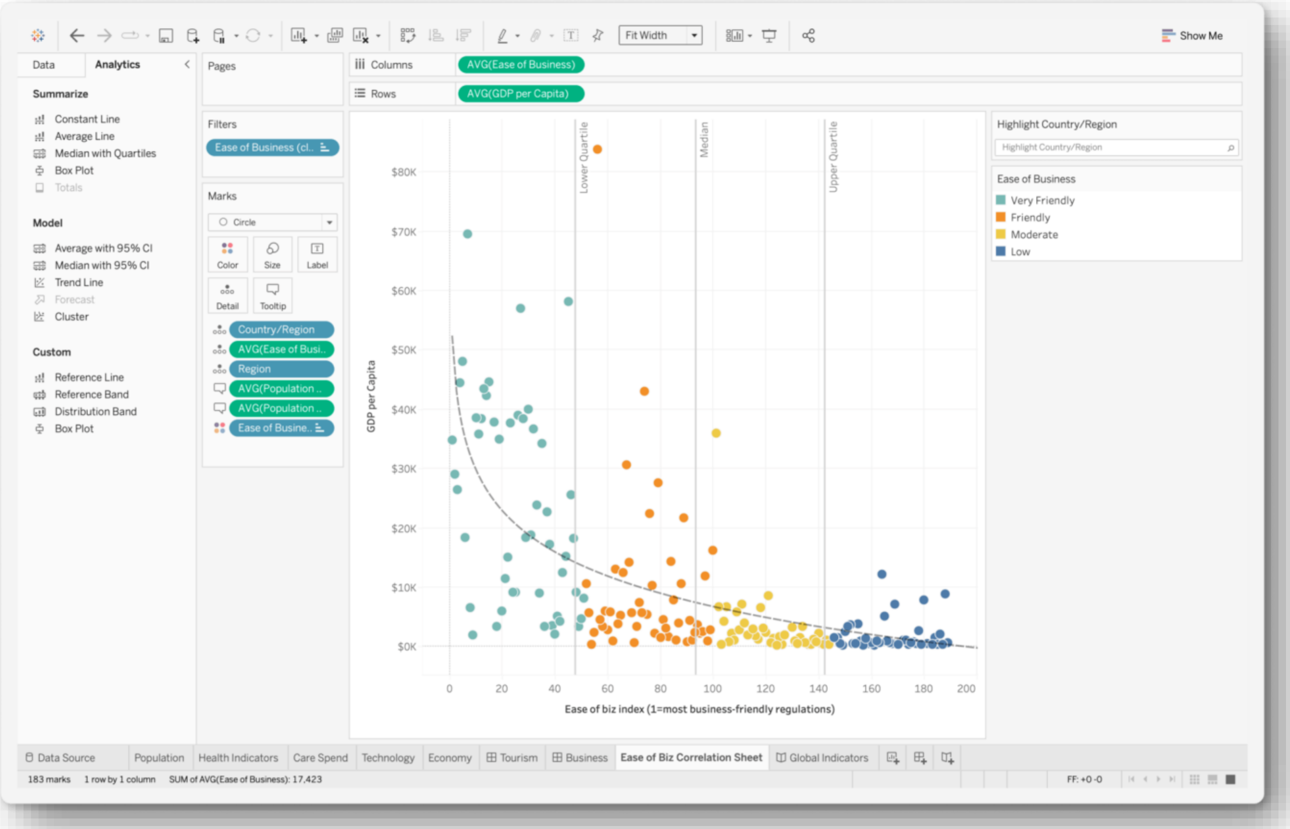
stargazer, gt / gtsummary, kable

R for Graphing



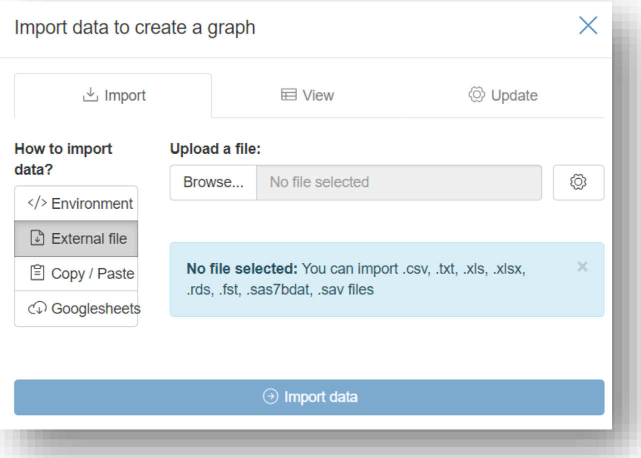
Tableau

Free for Academic Use: https://www.tableau.com/academic/students
Free for Any Use: https://public.tableau.com (can only save online)



esquisse - https://dreamrs.github.io/esquisse/

```
install.packages("esquisse")
esquisse::esquisser()
```

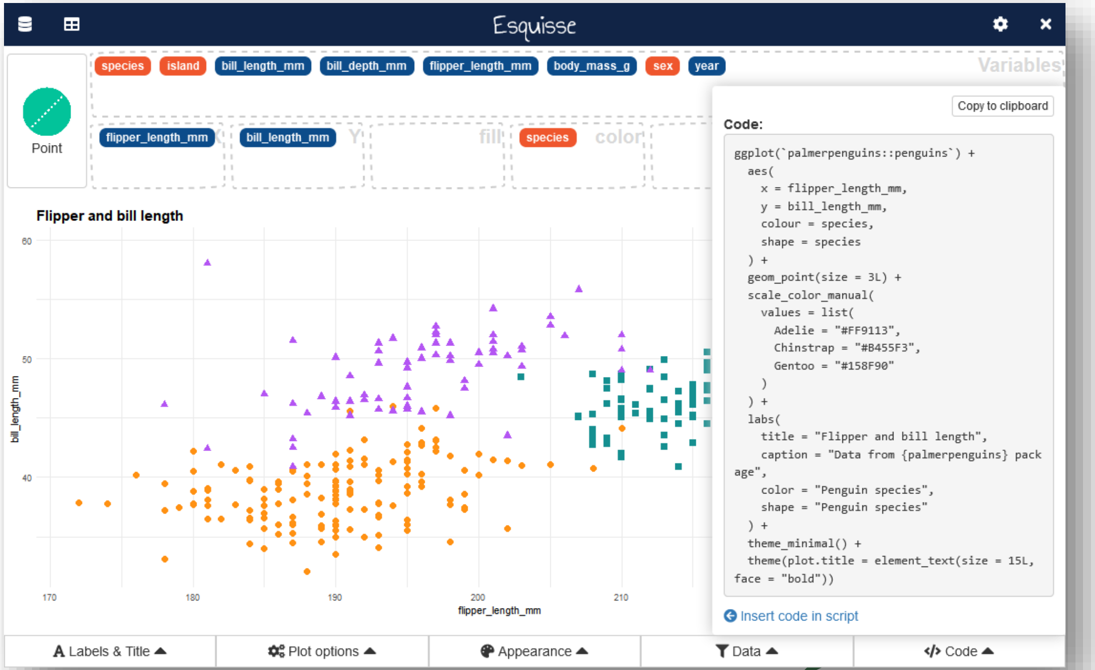


Import data to create a graph

How to import data?

Upload a file:

No file selected: You can import .csv, .txt, .xls, .xlsx, .rds, .fst, .sas7bdat, .sav files



Code:

```
ggplot(`palmerpenguins::penguins`) +
  aes(
    x = flipper_length_mm,
    y = bill_length_mm,
    colour = species,
    shape = species
  ) +
  geom_point(size = 3L) +
  scale_color_manual(
    values = list(
      Adelie = "#FF9113",
      Chinstrap = "#B455F3",
      Gentoo = "#158F90"
    )
  ) +
  labs(
    title = "Flipper and bill length",
    caption = "Data from {palmerpenguins} pack
age",
    color = "Penguin species",
    shape = "Penguin species"
  ) +
  theme_minimal() +
  theme(plot.title = element_text(size = 15L,
face = "bold"))
```

esquisse

www.dreamRs.fr

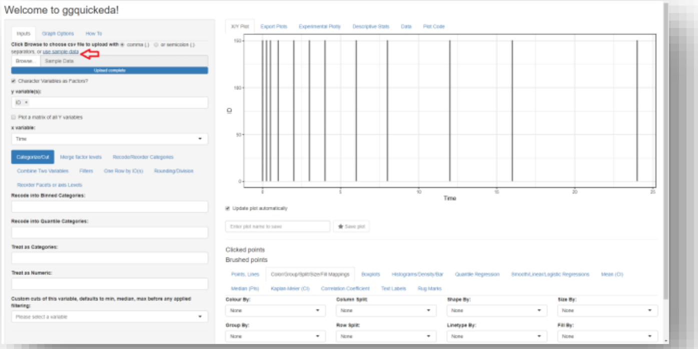
GEORGE MASON UNIVERSITY

ggplot (ggplot2)

- https://ggplot2.tidyverse.org
- https://r-graph-gallery.com/ggplot2-package.html

Easier Graphs

- ggquickeda – another Point and Click Interface



Welcome to ggquickeda!

Click Browse to choose csv file to upload with ● comma (,) ○ or semicolon (;) separators, or use sample data

☑ Character Variables as Factors?

y variable(s): ID

□ Plot a matrix of all Y variables

x variable: Time

Recode into Binned Categories:

Recode into Quantile Categories:

Treat as Categories:

Treat as Numeric:

Custom cuts of this variable, defaults to min, median, max before any applied filtering: Please select a variable

☑ Update plot automatically

Clicked points

Brushed points

Colour By: None

Column Split: None

Shape By: None

Size By: None

Group By: None

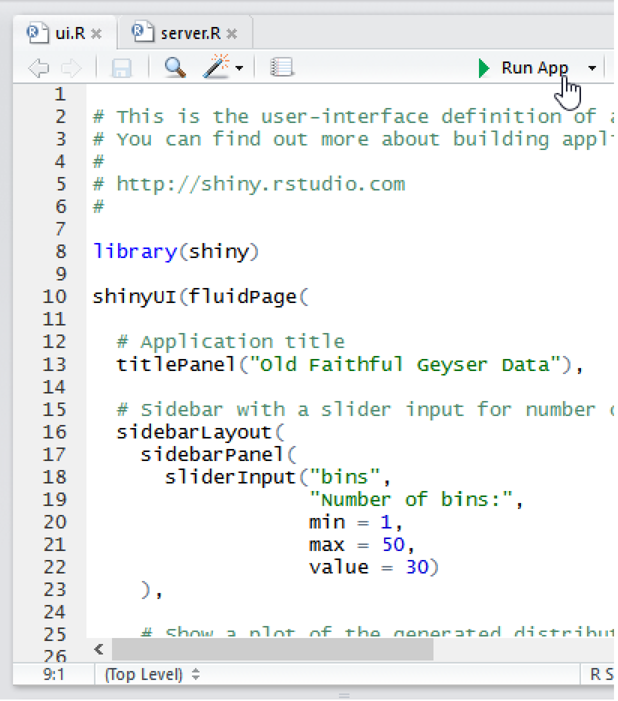
Row Split: None

Linetype By: None

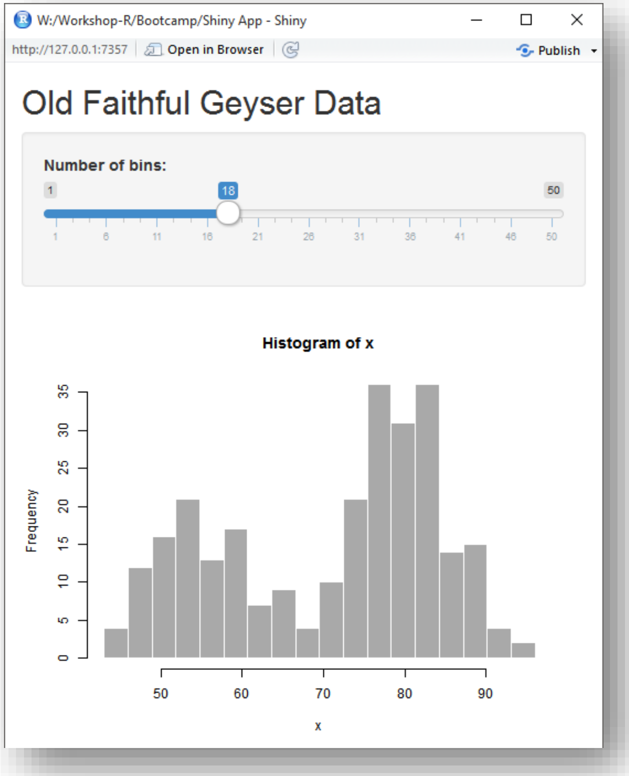
Fill By: None

- GGally - Good for bivariate graphing with functions like ggpairs()

Interactive Graphics (Shiny)

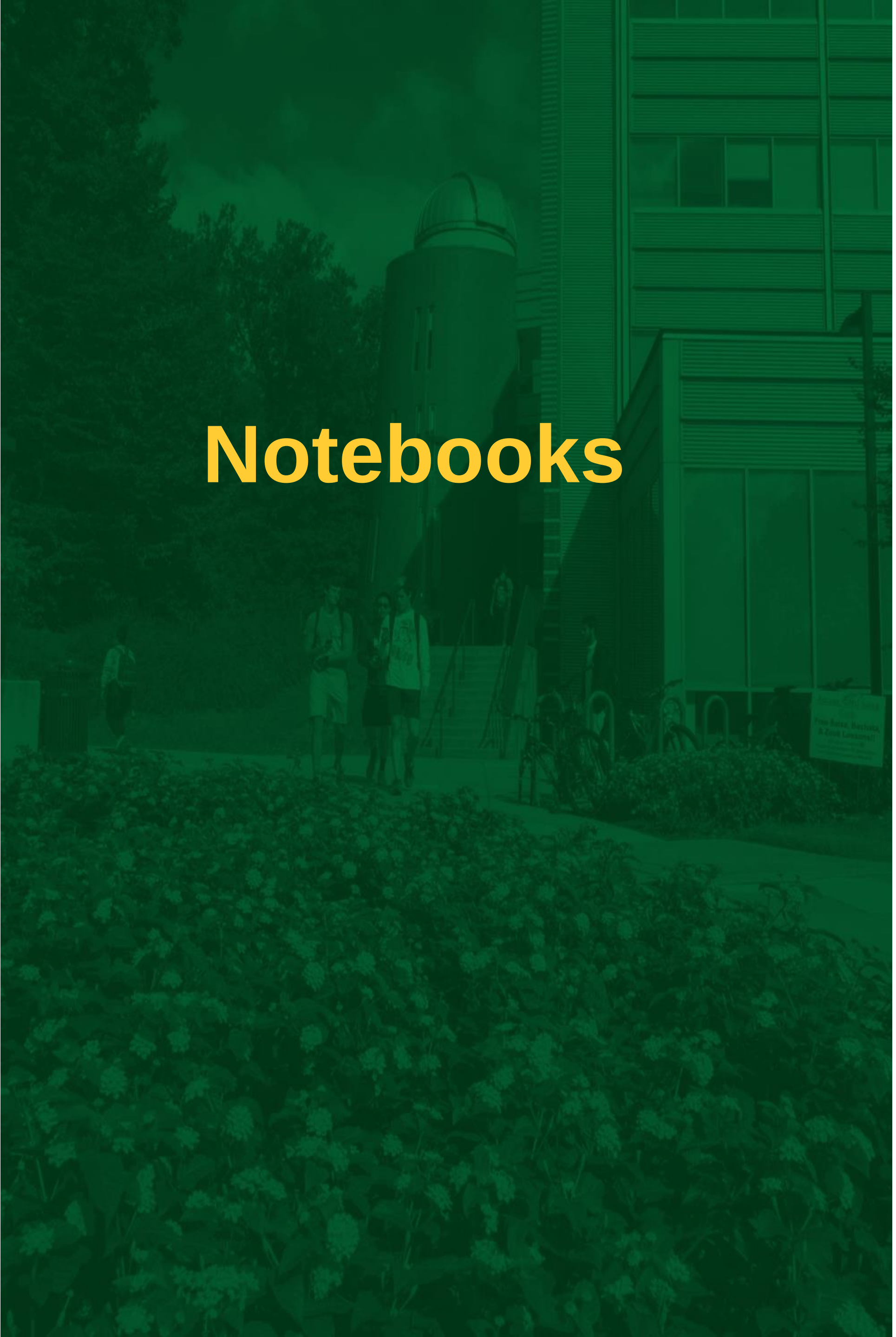


```

# This is the user-interface definition of a
# You can find out more about building appl
#
# http://shiny.rstudio.com
#

library(shiny)

shinyUI(fluidPage(

  # Application title
  titlePanel("Old Faithful Geyser Data"),

  # Sidebar with a slider input for number o
  sidebarLayout(
    sidebarPanel(
      sliderInput("bins",
                  "Number of bins:",
                  min = 1,
                  max = 50,
                  value = 30)
    ),

    # Show a plot of the generated distribut
```



Old Faithful Geyser Data

Number of bins: 18



Notebooks

- Quarto Document
- Jupyter Notebook

Reading Tutorials

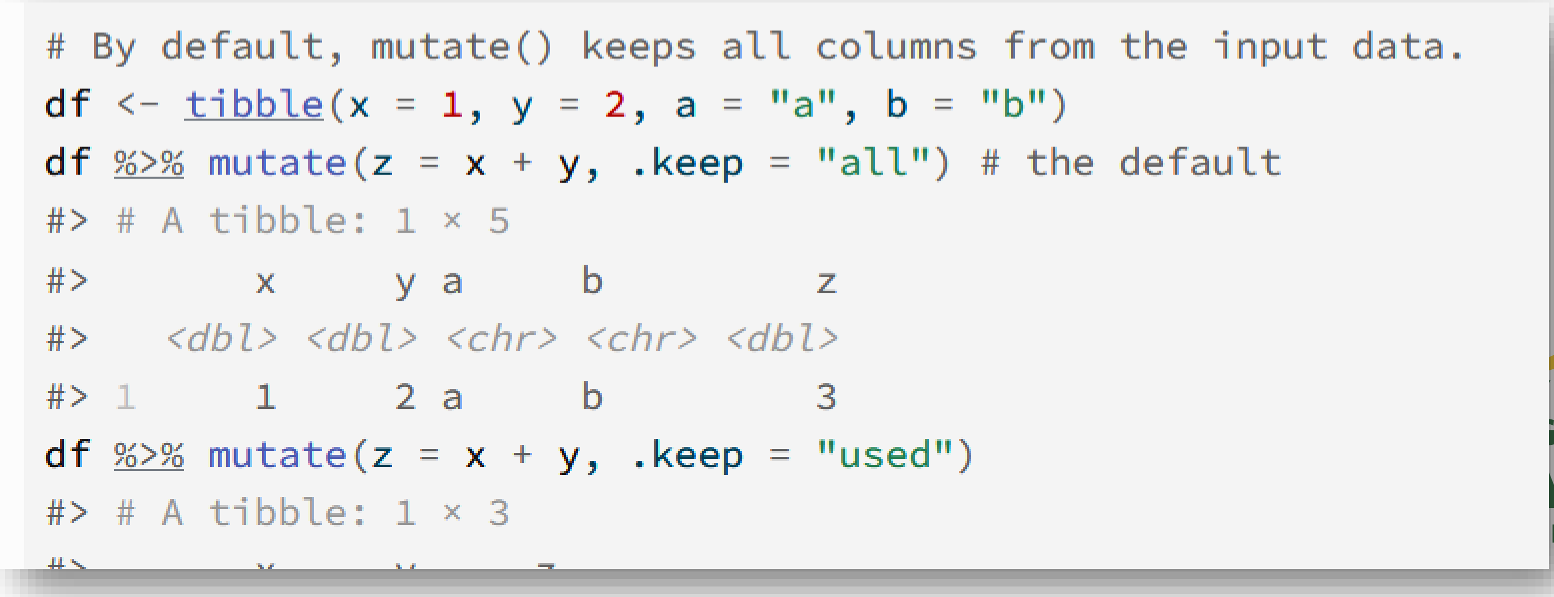
#   = A Comment

#>   = Commented output (what you should see)

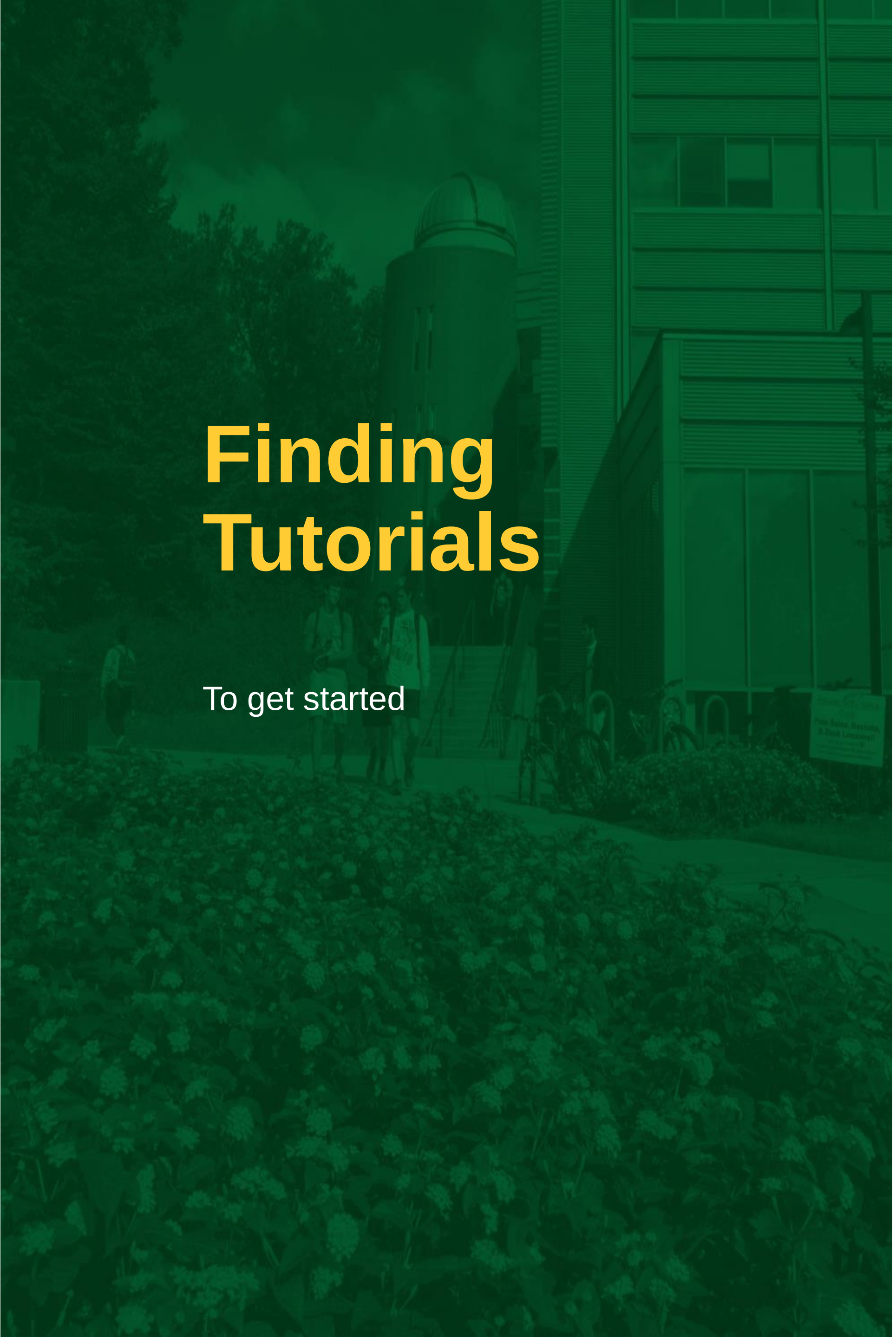
Does NOT save output in an object

df <- tibble(…)   = Creates a dataset

word   =   a link, click to learn more



```
# By default, mutate() keeps all columns from the input data.
df <- tibble(x = 1, y = 2, a = "a", b = "b")
df %>% mutate(z = x + y, .keep = "all") # the default
#> # A tibble: 1 × 5
#>       x     y a     b         z
#>   <dbl> <dbl> <chr> <chr> <dbl>
#> 1     1     2 a     b         3
df %>% mutate(z = x + y, .keep = "used")
#> # A tibble: 1 × 3
#>       x     y     z
```

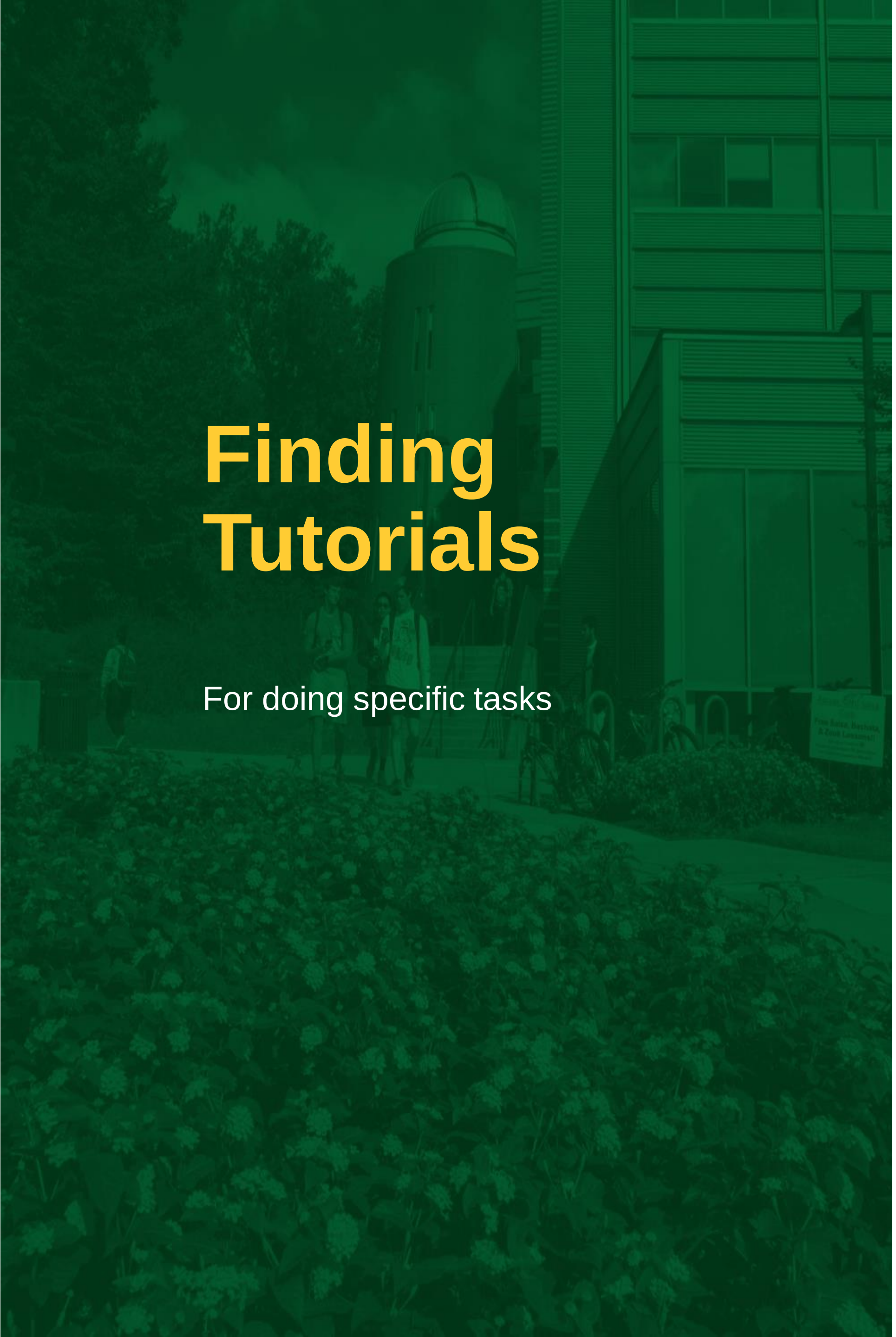


Finding Tutorials

To get started

- Look for ones that do what you need to do, preferably from someone in your field.
- Follow along!

- primers, learnr, swirl
- Many free online books
- Audit MOOCs
- Paid sites offer some free interactive tutorials



Finding Tutorials

For doing specific tasks

- Figure out what package you will want to use
  - Colleagues
  - In your field
- Look at "vignettes" and documentation for the package
- See recommended sites

Ignore tutorials that

- Don't use your data management scheme
  - e.g., if it uses the functions apply(), sapply(), tapply(), that is base R
- Look complicated
- Are more than 2 years old (generally)

Data for Follow-Along Tutorials

Use data loaded with base R or another package

Create data using functions or by hand

Provide a data file to download and read in

Teaching with R

- Posit Cloud
  - Free plan with limited sharing may be sufficient
  - Instructor plan with unlimited sharing is $15 + $5/month/student

- DataCamp
  - Free for students in a class, supports assignments
  - https://www.datacamp.com/groups/classrooms

- OER
  - Many OER textbooks (most online), more being added
  - Including https://bookdown.org
  - https://education.rstudio.com/teach/materials/