Module 3: R

Introduction

Instructor: Anjali Silva, PhD

TA: Tia Harrison, MSc

Data Sciences Institute, University of Toronto

27 June 2022

Welcome!

Course Documents

- Visit: https://github.com/anjalisilva/IntroductionToR
- All course material will be available via IntroductionToR GitHub repository (https://github.com/anjalisilva/IntroductionToR). Folder structure is as follows:
    - Lessons - All files: This folder contains all files.
    - Lessons - Data only: This folder contains data only.
    - Lessons - Lesson Plans only: This folder contains lesson plans only.
    - Lessons - PDF only: This folder contains slide PDFs only.
    - README - README file
    - .gitignore - Files to ignore specified by instructor

Welcome!

- Instructor: Anjali Silva, PhD
- Researcher and Lecturer, Department of Cell & Systems Biology, U of T
- Data Analyst, University of Toronto Libraries
- Pronouns: she/her
- Name Phonetic: Un-j-li Sil-va
- Email: a.silva@utoronto.ca (Must use the subject line DSI-IntroR. E.g., DSI-IntroR: Inquiry about Lecture I.)

Welcome!

- Teaching Assistant: Tia Harrison, MSc; PhD Candidate
- Department of Ecology and Evolutionary Biology
- Pronouns: she/her
- Name Phonetic: T-ee-ah
- Email: tia.harrison@mail.utoronto.ca

Description

This course is designed for students who have a degree in something other than Computer Science/Statistics who are looking to enhance their data science skills for their career. The first part of this course teaches R with a focus on manipulating and visualizing data. Students will get set up with a functional RStudio workflow, use different file types, transform data tables, import and manipulate data, use functions and loops, create data visualizations, make a Shiny app, and learn how to solve problems with their programming. Both base R and tidyverse methods are taught. To work reproducibly, students will create R Projects.

Learning Outcomes

- Setting up and using R and RStudio.
- Manipulating and visualizing data.
- Fixing errors.
- Understanding consent in data-based studies.
- Making presentations and managing projects.

Delivery Instructions

The course will be held over a period of 2 weeks, with classes taking place 3 days a week. Format will be online - synchronous via Zoom. Students must have internet connection and a computer with a microphone and required sofware implemented in order to participate. Keep microphones muted, unless you need to speak. Please indicate your name before speaking. Keeping your video on is optional, however, if you choose to leave it on, be mindful of what your peers can see. Course communications will take place via email. All course material will be available via IntroductionToR GitHub repository (https://github.com/anjalisilva/IntroductionToR). Folder structure is as follows:

- Lessons - All files: This folder contains all files.
- Lessons - Data only: This folder contains data only.
- Lessons - Lesson Plans only: This folder contains lesson plans only.
- Lessons - PDF only: This folder contains slide PDFs only.
- README - README file
- .gitignore - Files to ignore specified by instructor

Prerequisite knowledge

- The parts of a data table/spreadsheet
- Basics of file folder structure
- Summary statistics (mean, median, proportion, etc.)
- Basic data visualization types (bar charts, histograms, scatter plots)
- GitHub account

Submodules

- Hello World!
- Errors
- Reproducibility
- Data in R
- Manipulation
- Wrangling
- Programming
- Visualization
- Shiny (optional)

Key Texts

General reference:

R for Data Science by Wickham and Grolemund (2017)
https://r4ds.had.co.nz/index.html

DoSS Toolkit (2021) https://rohanalexander.github.io/doss_toolkit_book/.

Key Texts

For specific topics:

- Alexander, 2022, *Telling Stories with Data*, CRC Press. https://www.tellingstorieswithdata.com/
  - Alexander (eds), 2021, *DoSS Toolkit*, https://rohanalexander.github.io/doss_toolkit_book/.
  - de Graaf, 2019, *Managing Your Data Science Projects: Learn Salesmanship, Presentation, and Maintenance of Completed Models*, Apress.
  - Healy, 2018, *Data Visualization*: A Practical Introduction, Princeton University Press
  - Timbers et al., 2021. *Data Science: A First Introduction*. https://ubc-dsci.github.io/introduction-to-datascience/
  - Wickham and Grolemund, 2017, *R for Data Science*, O'Reilly. https://r4ds.had.co.nz/
  - Wickham, 2021, *Mastering Shiny*, O'Reilly. https://mastering-shiny.org/
  - Wiley, Matt, Wiley, Joshua F., 2020, *Advanced R 4 Data Programming and the Cloud*
  - *Using PostgreSQL, AWS, and Shiny*, Apress.

Materials

- Learners must have internet connection and a computer with a microphone in order to participate in online activities.
- Learners must have R (http://www.r-project.org/).
- Learners must have RStudio (http://www.rstudio.com/).
- Screen space can be a limitation during online learning since you'll want to see the instructor's screen and have your RStudio open so that you can type along. If you have access to a second monitor or a larger tablet to attend the course while keeping your laptop screen available for coding - this would be great! If not - don't worry, we'll manage!

Course Expectations

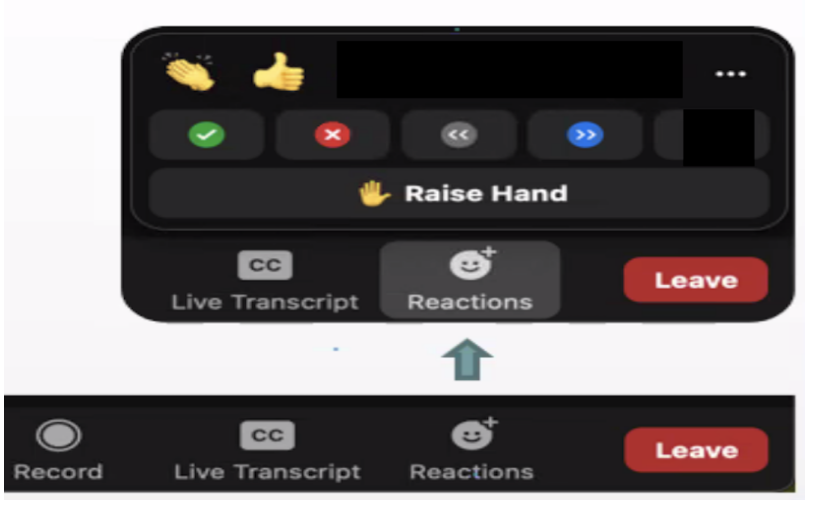


Figure: Zoom ‘Reactions’ that you may use.

Course Expectations

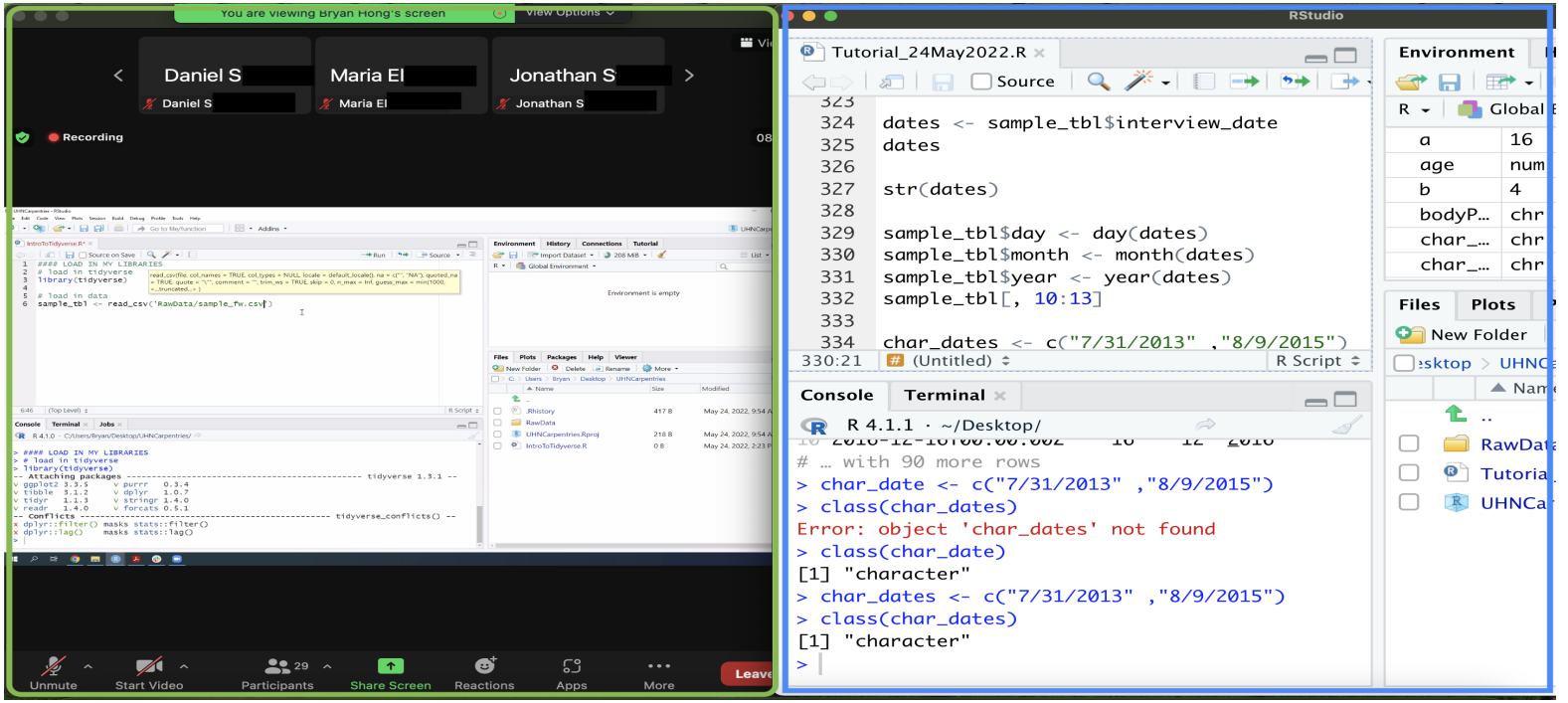


Figure: It is recommended that windows are resized so that both the user RStudio window and Instructor Zoom window (with RStudio) is visible at the same time. User may collapse panels of their RStudio not in current use.

Tentative Schedule

*Calendar may be modified as needed, and learners will be informed. Course will be taught using R version 4.2.0 and RStudio Desktop version 2022.02.2.

- Monday 27 June, 6pm-8pm EST
  - Introduction, Hello World! And Work practices
- Thursday 30 June, 6pm-8pm EST
  - Data in R
- Saturday 2 Jul, 9am-noon EST
  - Manipulation
- Monday 4 July, 6pm-8pm EST
  - Wrangling
- Thursday 7 July, 6pm-9pm EST
  - Programming
- Saturday 9 July, 9am-noon EST
  - Visualization (optional Shiny)

Course Policies

Course Expectations

- The course will include mainly live-coding classes. Students are expected to follow along with the coding. Be mindful of online fatigue. Be respectful and only one speaker at a time. Keep yourself on mute, unless you need to speak or ask a question. If you have a question, use raise hand feature. First say your name, then ask the question. If you have a question, you may type it to chat as well.
- Students with diverse learning styles and needs are welcome in this course. We are dedicated to providing an accessible learning environment for all. Please notify in advance of the course start date if you require any accommodations or if there is anything we can do to make this course more accessible to you.

Acknowledgements

- Slides covered in the lectures were originally developed by Amy Farrow under the supervision of Rohan Alexander, University of Toronto. Slides have been modified by Anjali Silva for 2022.

Any questions?

Data Science Tools

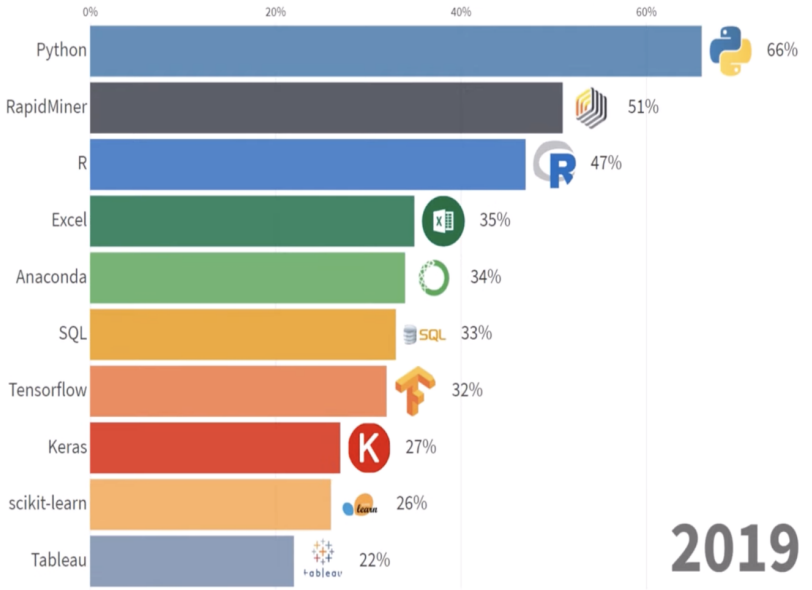
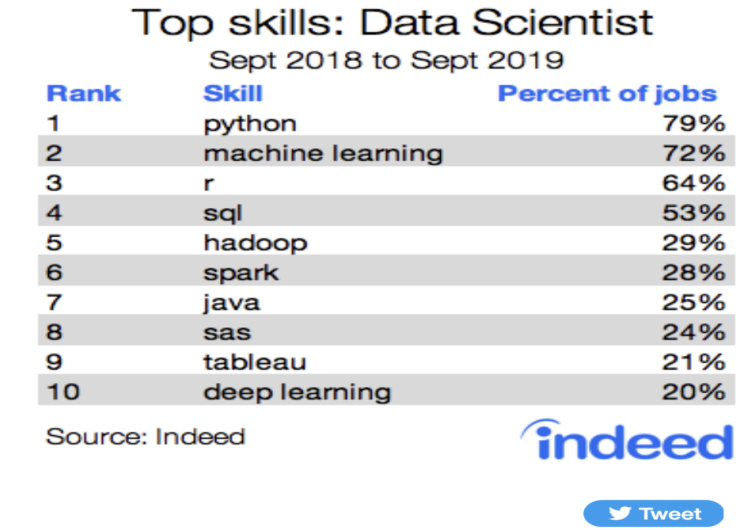


Figure: KDnuggets Survey of Machine Learning Software that asked respondents which data science tools they had used for projects within the past year. The x-axis shows the proportion of users who used a particular data science tool within the past year. Figure from https://www.kdnuggets.com/2020/06/ data-science-tools-popularity-animated.html

Data Science Skills



Top skills: Data Scientist

Sept 2018 to Sept 2019

Rank	Skill	Percent of jobs
1	python	79%
2	machine learning	72%
3	r	64%
4	sql	53%
5	hadoop	29%
6	spark	28%
7	java	25%
8	sas	24%
9	tableau	21%
10	deep learning	20%

Source: Indeed

indeed

Tweet

Table titled "Top skills: Data Scientist." Indeed ranked the top skills in data scientist job postings from September 2018 to September 2019, comparing the percent of jobs for each skill. Results vary. Caption added post-publication.

R

What is R?

- A language and environment for statistical computing and graphics.
- R was initially written by Ross Ihaka and Robert Gentleman.
- Since mid-1997, the R Core Team modify the R source.
- R runs on a wide variety of UNIX platforms, Windows and MacOS.
- R is designed with interactive data exploration in mind.
- A version of R is released each year. Current release is 4.2.0.

Further Reading

- Ihaka,R. and Gentleman,R. (1996) R: a language for data analysis and graphics. *J. Comput. Graph. Statist.*, 5, 299–314.

Why R?

- R is open source and free.
- R has a community.
- With R, you can share your data analysis methods in a reproducible way.
- Packages (more than 18 thousand on CRAN!) extend R's capabilities to provide easy ways to accomplish a wide variety of tasks.
- R is one of the standard language recommendations for data science.
- RStudio makes it easier to do more with R.

RStudio

RStudio

- RStudio is an integrated development environment R.
- Not only for R, but can also use Python.
- Has:
    - A console
    - Syntax-highlighting editor for code execution
    - Tools for plotting, viewing history, debugging and workspace management
- RStudio contains many features that make the development process easier and faster.

Options To Work With R

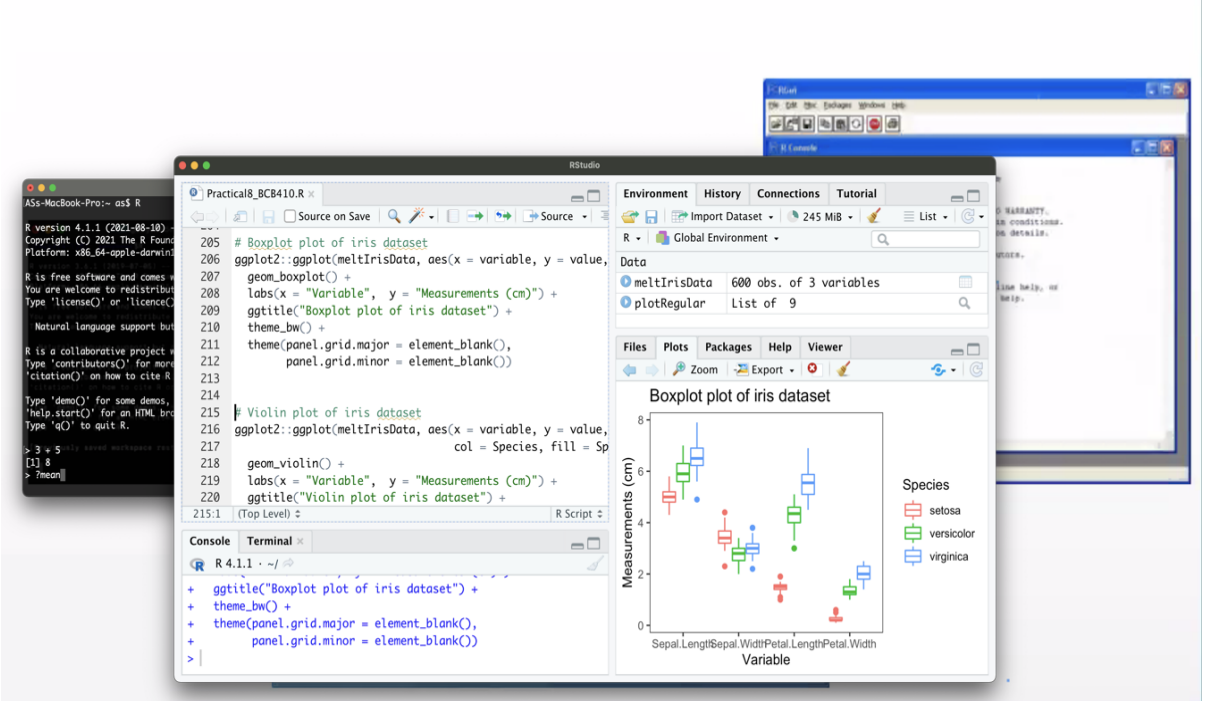


Figure: Some options to work with R. Several other options are present including the Jupyter Notebook.

Options To Work With R

Any questions?

What can you do with R?

Load data

```
## # A tibble: 9,113 × 5
##    YEAR_BUILT YEAR_EVALUATED LONGITUDE LATITUDE SCORE
##         <dbl>          <dbl>     <dbl>    <dbl> <dbl>
##  1       1950           2021     -79.5     43.7    64
##  2       1960           2021     -79.5     43.7    60
##  3       1969           2021     -79.4     43.7    64
##  4       1960           2021     -79.5     43.7    91
##  5       1973           2021     -79.5     43.7    91
##  6       1960           2021     -79.3     43.7    88
##  7       1962           2021     -79.5     43.6    84
##  8       1993           2021     -79.4     43.7    83
##  9       1995           2021     -79.3     43.7    89
## 10       1964           2021     -79.3     43.7    74
## # … with 9,103 more rows
```

Clean data

```
## # A tibble: 9,113 × 5
##    year_built year_evaluated longitude latitude score
##         <dbl>          <dbl>     <dbl>    <dbl> <dbl>
##  1       1950           2021     -79.5     43.7    64
##  2       1960           2021     -79.5     43.7    60
##  3       1969           2021     -79.4     43.7    64
##  4       1960           2021     -79.5     43.7    91
##  5       1973           2021     -79.5     43.7    91
##  6       1960           2021     -79.3     43.7    88
##  7       1962           2021     -79.5     43.6    84
##  8       1993           2021     -79.4     43.7    83
##  9       1995           2021     -79.3     43.7    89
## 10       1964           2021     -79.3     43.7    74
## # … with 9,103 more rows
```

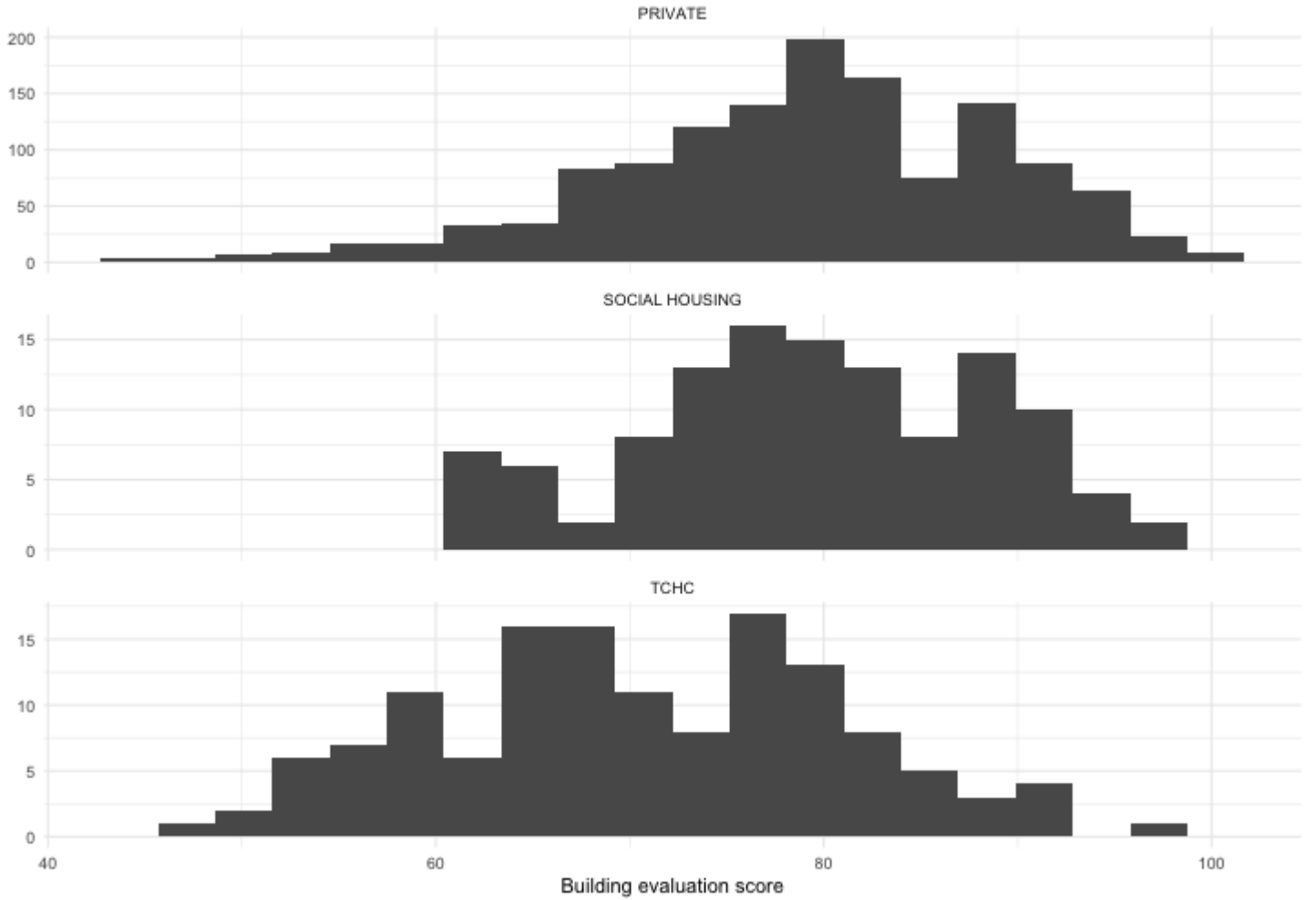
Manipulate and combine data

```
## # A tibble: 8,291 × 6
##    year_built property_type confirmed_units score  year count
##         <dbl> <chr>                   <dbl> <dbl> <dbl> <int>
##  1       1960 PRIVATE                    12    73  2020     3
##  2       1960 PRIVATE                    12    81  2020     3
##  3       1962 PRIVATE                    10    73  2020     9
##  4       1968 PRIVATE                   174    81  2020    48
##  5       1965 PRIVATE                    27    73  2020     9
##  6       1950 PRIVATE                    10    77  2020     9
##  7       1974 TCHC                      350    82  2020    36
##  8       1928 PRIVATE                    15    73  2020    19
##  9       1938 PRIVATE                    32    74  2020    16
## 10       1958 PRIVATE                    55    72  2020    36
## # … with 8,281 more rows
```

Summarize Data

ward	Count	Average Score	Median Year Built	Median Number of Storeys	Median Number of Units
1	221	69.28507	1967	7	97
2	336	71.46131	1965	7	68
3	597	70.47906	1957	4	32
4	483	68.05797	1960	5	42
5	597	69.00000	1960	4	37
6	581	70.80379	1960	4	39
7	277	68.07942	1970	11	135
8	617	71.26580	1958	4	31
9	210	68.00476	1959	4	27
10	93	74.16129	1987	7	103

Visualize Data



Write Reports

Paper title*

Subtitle

Author

Date

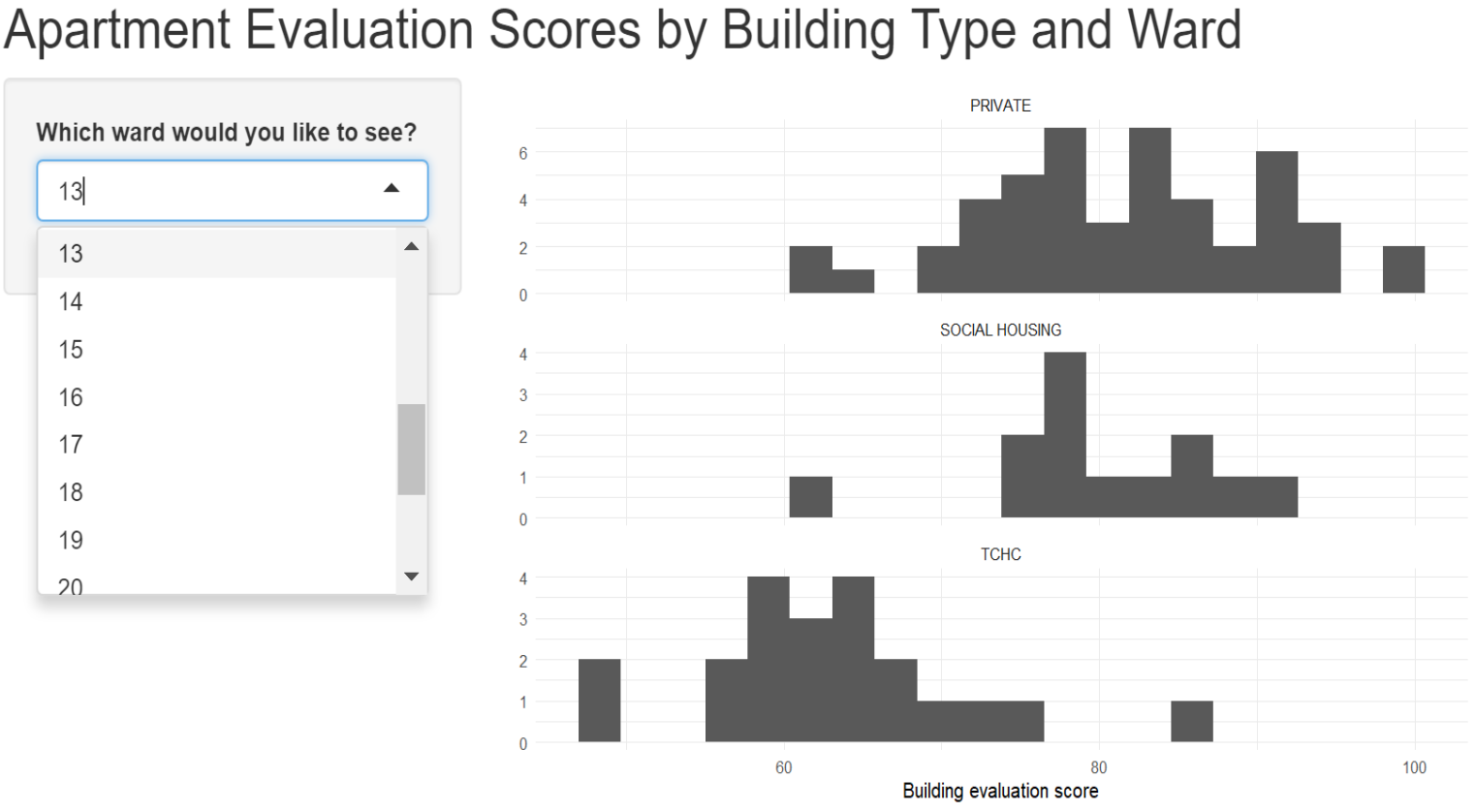
Abstract

An abstract

Contents

1 Introduction

Build Interactive Applications



And more:

- Data collection
- Statistical analysis
- Data modeling
- Presentations
- Websites

Hello World!

(Beginner)

How can we start using R?

Goals:

- a fully-functional R and RStudio setup
- understanding and using parts of the RStudio IDE
- run basic commands in R
- creating and using different R file types for different purposes

Hello World!

Getting set up

- R
- RStudio

R basics

```
(27 / 52) * 100
object_name <- value
function(arguments)
```

File types

- scripts
- RMarkdown

Errors

(Beginner)

How can we avoid getting stuck on errors while using R?

Goal:

- Functional problem-solving abilities for learning and using R

Errors

Getting help

Using Stack Overflow

Making reproducible examples

Reproducibility

How does R help us work reproducibly?

Goal:

- Use an RProject and GitHub to make your data analysis project reproducible
- Understand coding conventions

Data in R

(Beginner)

What does data look like in R?

Goals:

- Know what data.frames, tibbles, and tidyverse are
- Understand key types of data, including strings, ordered factors, and dates and times
- Understand how R handles missing values

Data in R

Tidyverse

```
library(tidyverse)
```

Tibbles

```
tibble()
```

Strings

```
"This is a string"
```

Factors

```
factor(vector, levels)
```

Data in R

Dates and times

```
library(lubridate)
```

Missing values

```
NA
```

Manipulation

(Beginner)

How can we manipulate data tables in R?

Goals:

- View subsets of data tables
- Pick specific variables
- Create new variables
- Group observations by traits
- Summarise groups of observations
- Order data tables

Manipulation

Filtering

```
filter()
```

Arranging

```
arrange()
```

Selecting

```
select()
```

Mutating

```
mutate()
```

Manipulation

The pipe

```
%>%
```

Grouping

```
group_by()
```

Summarizing

```
summarise()
```

- Counting
- Proportions

Wrangling

(Intermediate)

How can we work with real data sets in R?

Goals:

- Load data tables into R
- Connect related but separate data tables
- Load data from an external database
- Work efficiently with larger data sets

Wrangling

Importing data

```
read_csv()
```

Interacting with databases

```
library(RPostgreSQL)
```

Cleaning

```
library(janitor)
```

Pivot

```
pivot_longer(), pivot_wider()
```

Wrangling

Joining data

```
left_join(), right_join(), full_join(), inner_join()
```

data.table

```
library(data.table)
```

Programming

(Intermediate)

How can we use programming concepts like iterators to enhance our work in R?

Goals:

- Write functions in R to perform custom operations
- Perform operations iteratively
- Perform operations given specific conditions
- Understand and use vectors in functions and loops
- Make data sets for simulation studies

Programming

Functions

```
name <- function(x) {

}
```

Vectors

```
c(), list()
```

Loops

```
for (i in 1:10) {

}

while (i < 10) {

}
```

Programming

If/else logic

```
if (x = 3) {

} else {

}
```

Simulation

```
set.seed(), runif(), rnorm(), sample()
```

Visualization

(Intermediate)

What kinds of visualizations can we make in R?

Goals:

- Make communicative and visually-pleasing bar graphs, histograms, and scatterplots

Visualization

Essentials

```
ggplot(aes())
```

Bar charts and histograms

```
geom_bar(), geom_histogram()
```

Scatter plots

```
geom_point(), geom_smooth()
```

Shiny (optional)

(Advanced)

How can we make interactive applications using R?

Goal:

- Make a basic functional Shiny application to display a data visualization

Shiny (optional)

```
library(shiny)
ui <- fluidPage(
  "Hello, world!"
)
server <- function(input, output, session) {
}
shinyApp(ui, server)
```

Examples:

Visit: https://shiny.rstudio.com/gallery/

Any questions?

Ethics (not covered)

Why does consent matter in data-based studies?

Goal:

- Understand the necessity and complexity of consent for data-based studies

Ethics (not covered)

James H. Ware, 1989, ‘Investigating Therapies of Potentially Great Benefit: ECMO’, Statistical Science.

Donald A. Berry, 1989, ‘Comment: Ethics and ECMO’, Statistical Science.

Inequity (not covered)

How can we undertake is Equity, Diversity, and Inclusion training?

Goal:

- Understand Equity, Diversity, and Inclusion (EDI) training

Professional skills (not covered)

Goals:

- Presenting data analysis results
- Managing data projects
- Data security

Industry case study (not covered)

Delivery (not covered)

For technical sections: (not covered)

- Short lectures
- Examples

For non-technical sections:

- Readings
- Discussions

Assessment (not covered)

Formative (not covered)

For technical sections:

- In-class independent exercises & solution discussion
- Problem solving exercises (individual solution and small group discussion)

For non-technical sections:

- Group activities

Summative

For technical sections:

- Multi-stage project using data sets chosen from a provided selection

For non-technical sections:

- Written reflections