4.0 Introduction to R: Introduction

Julia Gallucci

Data Science Institute, University of Toronto

5/29/23

Acknowledgements

Slides are adapted from Anjali Silva, originally from Amy Farrow under the supervision of Rohan Alexander, University of Toronto. Slides have been modified by Julia Gallucci, 2023.

About myself

Instructor: Julia Gallucci, HBSc, MSc

PhD student, Institute of Medical Science UofT and CAMH

Pronouns: She/Her

julia.gallucci@mail.utoronto.ca (Please use subject line: i.e., DSI-IntroR: Inquiry about Lecture 1)

Course Documents

Visit: https://github.com/UofT-DSI/04-intro_r

All course material will be available via IntroductionToR GitHub repository (https://github.com/UofT-DSI/04-intro_r). 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

Course Description

This course is designed for learners who have a degree in something other than Computer Science/Statistics who are looking to enhance their data science skills for their career.

1. Manipulating and visualizing data in R. Learners 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, learners will create R Projects.
2. Ethics of consent, Equity, Diversity & Inclusion (EDI) training, and professional skills including presentation, project management, and data security.
3. Industry case study.

Learning Outcomes

1. Setting up and using R and RStudio.
2. Manipulating and visualizing data.
3. Fixing errors.
4. Understanding consent in data-based studies.
5. Making presentations and managing project

Content Delivery

This course will be held over a 3-week period, with classes taking place 3 days per week (2 hours of lecture, 1 hour of tutorial).

Tutorials will be used for homework/assignment help, and support hours

Format: online synchronous via zoom

Further course communication will take place via email

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

Outline of course schedule

Monday 29 May 6pm-8pm	Hello World! And Work practices (R basics; file types; errors)
Thursday 1 June 6pm-8pm	Data in R (tibbles, strings, factors, times, missing values)
Saturday 3 June 9am-noon	Manipulation (filtering; arranging; selecting; mutating, pipe; grouping; summarize)
Monday 5 June 6pm-8pm	Wrangling (importing data; pivot, joining data; data.table)
Thursday 8 June 6pm-8pm	Programming (custom functions, loops, if/else logic, purr, simulations)
Saturday 10 June 9am-noon	Visualization (initialization, choosing chart types, ggplot, customizing)
Monday 12 June 6pm-8pm	Shiny applications; Ethics, inequity and professional skills
Thursday 15 June 6pm-8pm	Professional skills: Industry case study – Kamilah Ebrahim
Saturday 17 June 9am-noon	R: Review and Practice

Assessments

Class attendance**:** 10% of final grade.

Problem sets: 90% of final grade (2 assignments, 45% each).

1. Problem set 1 is based on R basics, navigating RStudio, data types and structures, R coercion rules, using built-in functions, working with missing values, use of external functions by downloading R packages and string manipulation. **Due: Sunday June 11, 2023 11:59 PM EST**
2. Problem set 2 is based on data reshaping techniques and tidyverse R package, including applications of data manipulating, wrangling, functional programming and data visualization. **Due: Sunday June 18, 2023 11:59 PM EST**

Some Resources

Key Texts

General references:

- 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/.

Some Resources

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

Some Resources

Key Texts

For specific topics:

- 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/).

Optional Try posit cloud to access RStudio right from your browser - no installation required! (https://posit.cloud/)

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

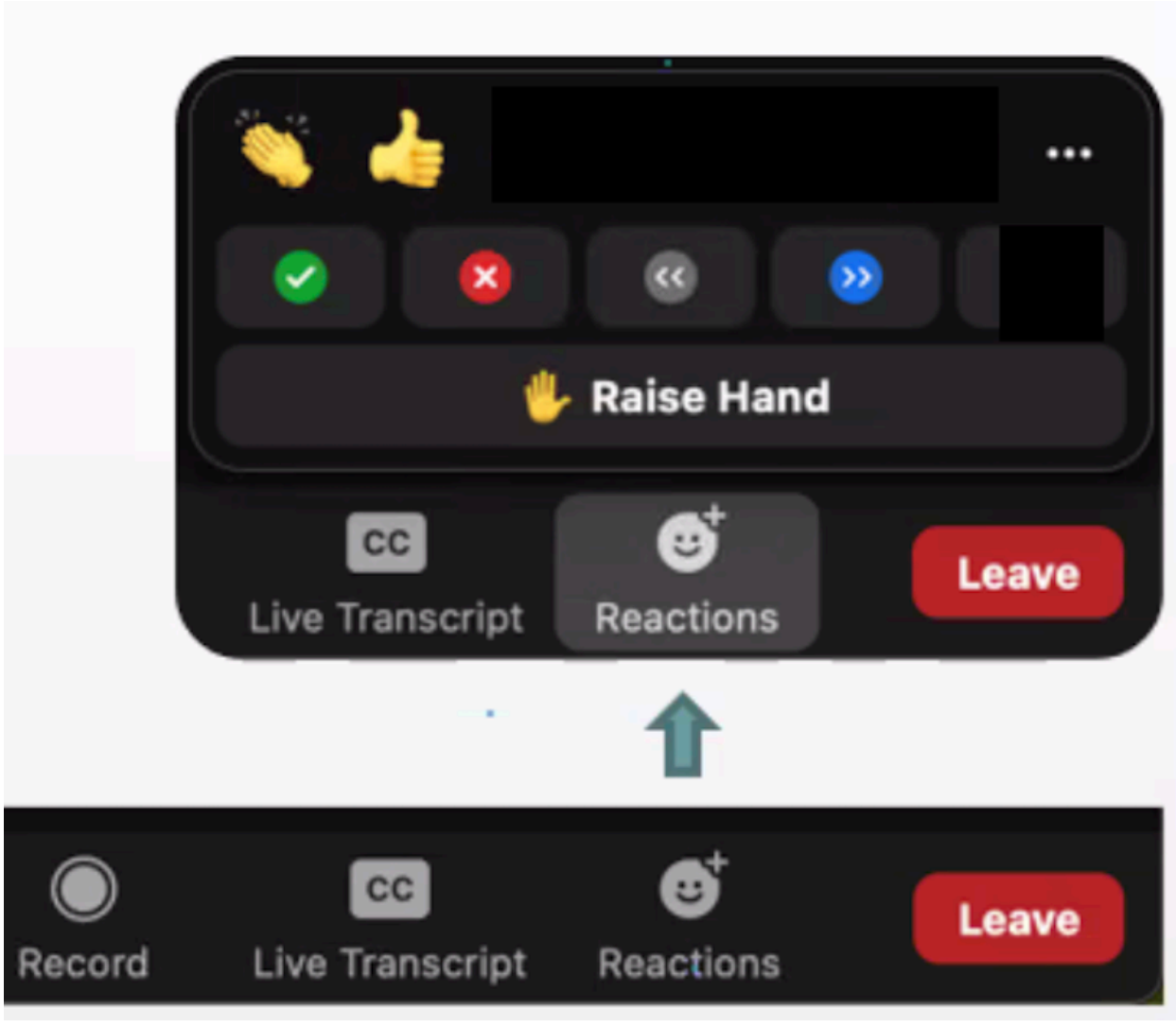


Figure 1: Zoom 'Reactions' that you may use.

Course expectations

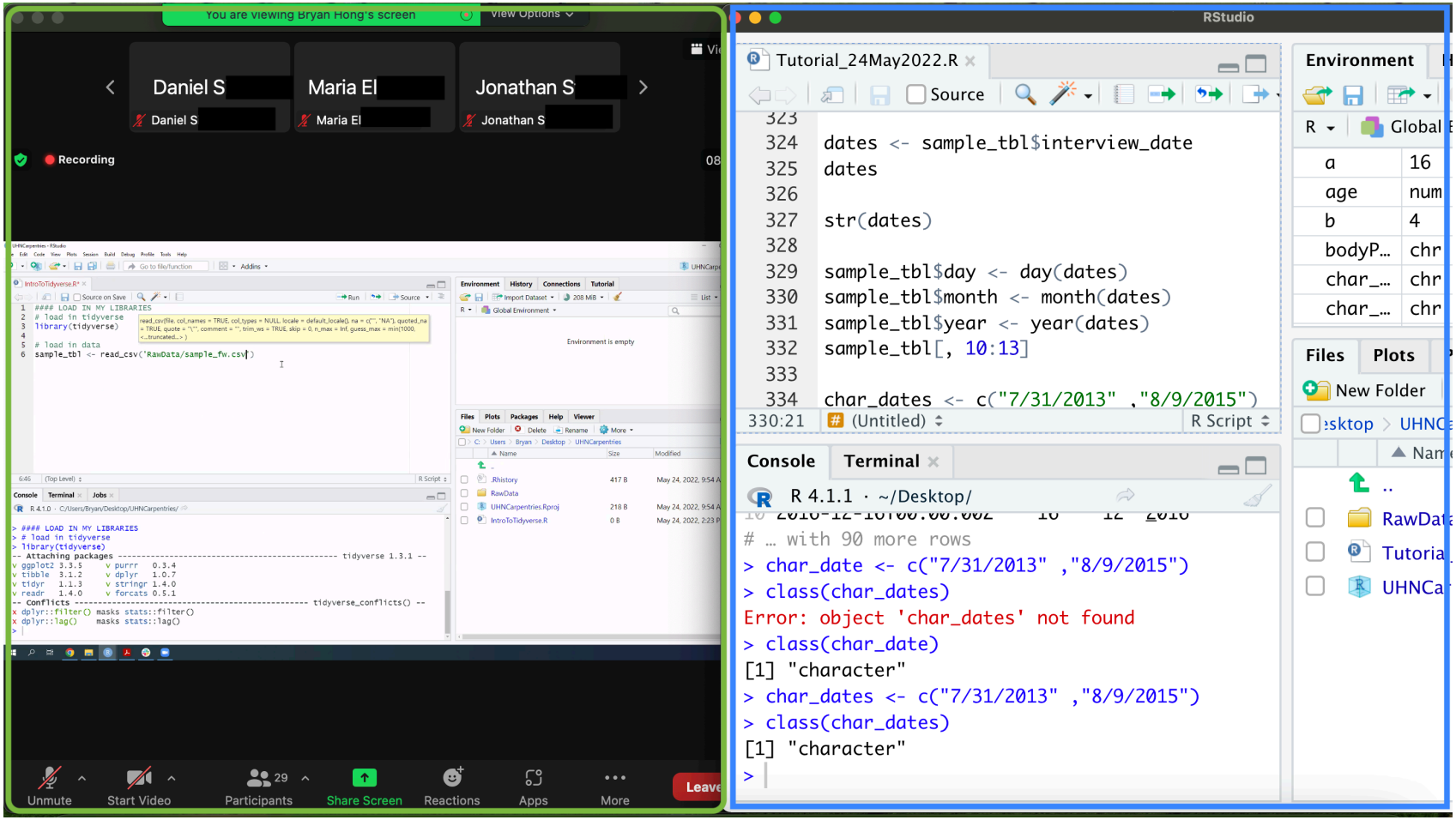


Figure 2: It is recommended that windows are resized so that both the user RStudio window and Instructor Zoom window (with RStudio) is visible at the same time. User may collapse panels of their RStudio not in current use

Course expectations

This course will mainly be a live-coding class, learners are expected to follow along

Students with diverse learning styles and needs are welcome to this course. I aim to provide an accessible learning environment. Please notify me in advance via email if you require accommodations or if there is anything that can be done to make the course more accessible to you.

Questions?

Data Science Tools

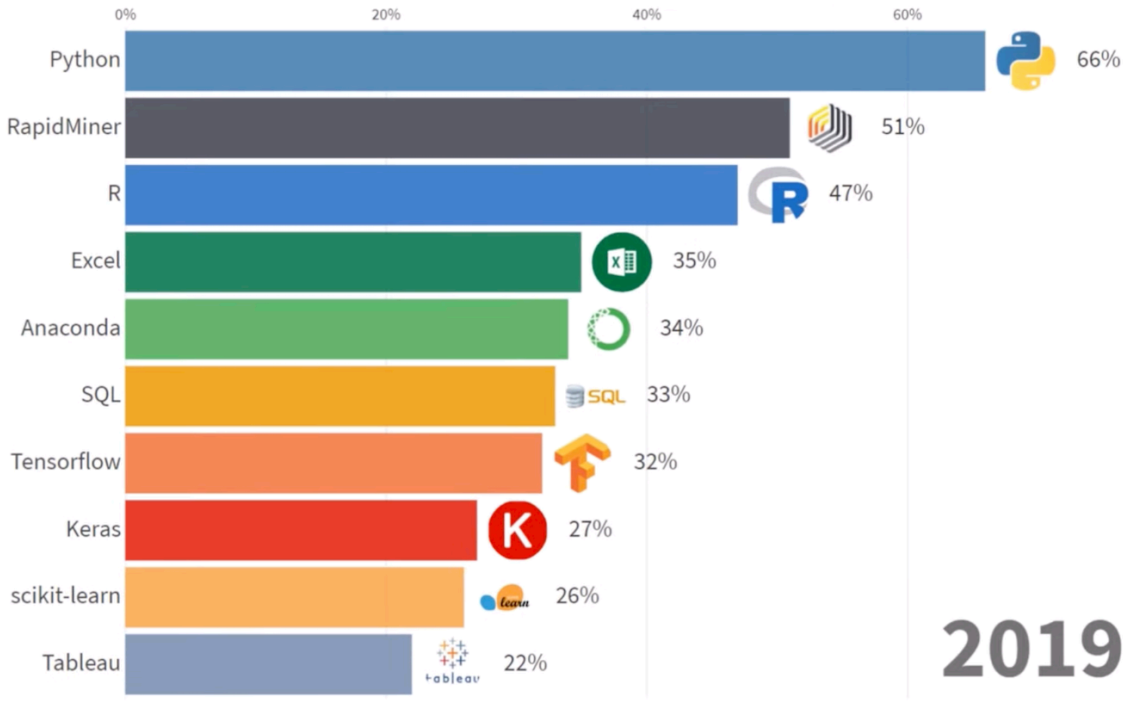
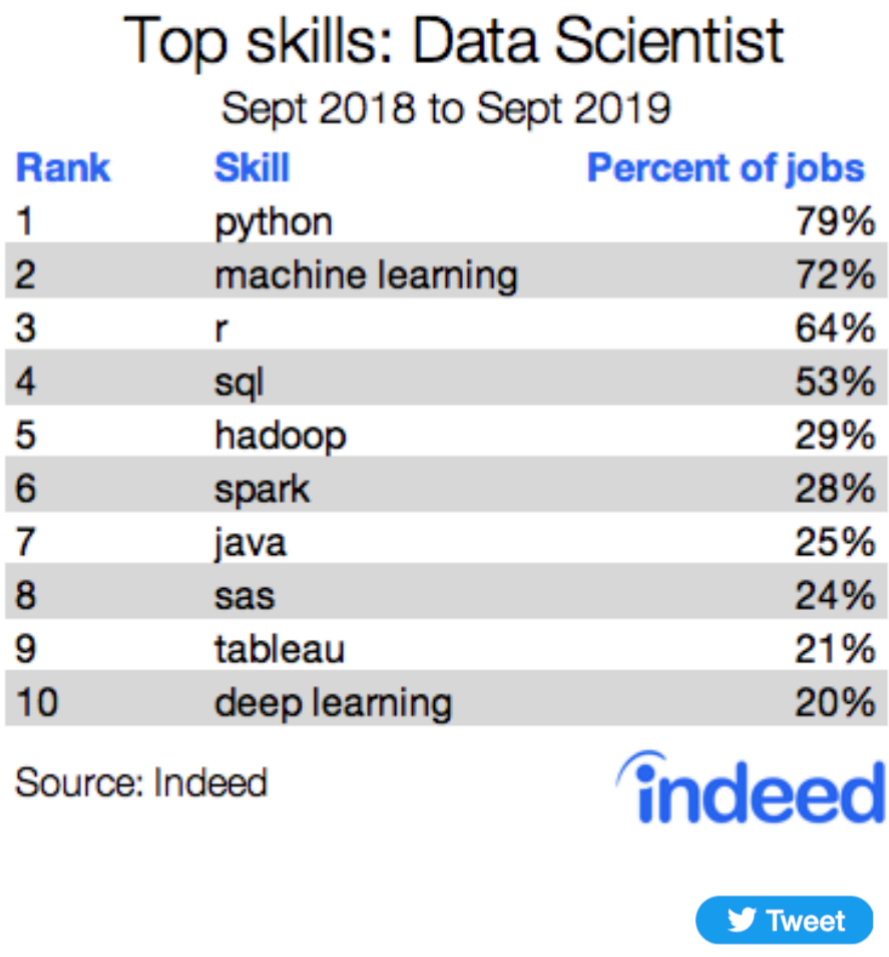


Figure 3: KDnuggets Survey of Machine Learning Software that asked respondents which data science tools they had used for projects within the past year. The x-axis shows the proportion of users who used a particular data science tool within the past year. Figure from https://www.kdnuggets.com/2020/06/data-science-tools-popularity-animated.html

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.

What is R?

Language and environment for statistical computing and graphics

R was initially written by Ross Ihaka and Robert Gentleman

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 annually. Current release is 4.2.3

Optional: Further reading, Ihaka R and Gentleman, R (1996) R: a language for data analysis and graphics. J. Comput. Grapj. Statist., 5, 299-314

Why use R?

R is open source and free

R has a community

With R, you can share your data analysis in a reproducible way

More than 18 thousand packages (on CRAN) that extend R's capabilities to provide easy ways to accomplish a wide variety of tasks

R is a standard language recommended for data science

RStudio makes it easier to use R

RStudio

RStudio is an integrated development environment R

Contains:

- Console
- Syntax-highlighting editor for code execution
- Tools for plotting, viewing history, debugging and workspace management

RStudio contains features that make development easier and faster!

Options to work with R

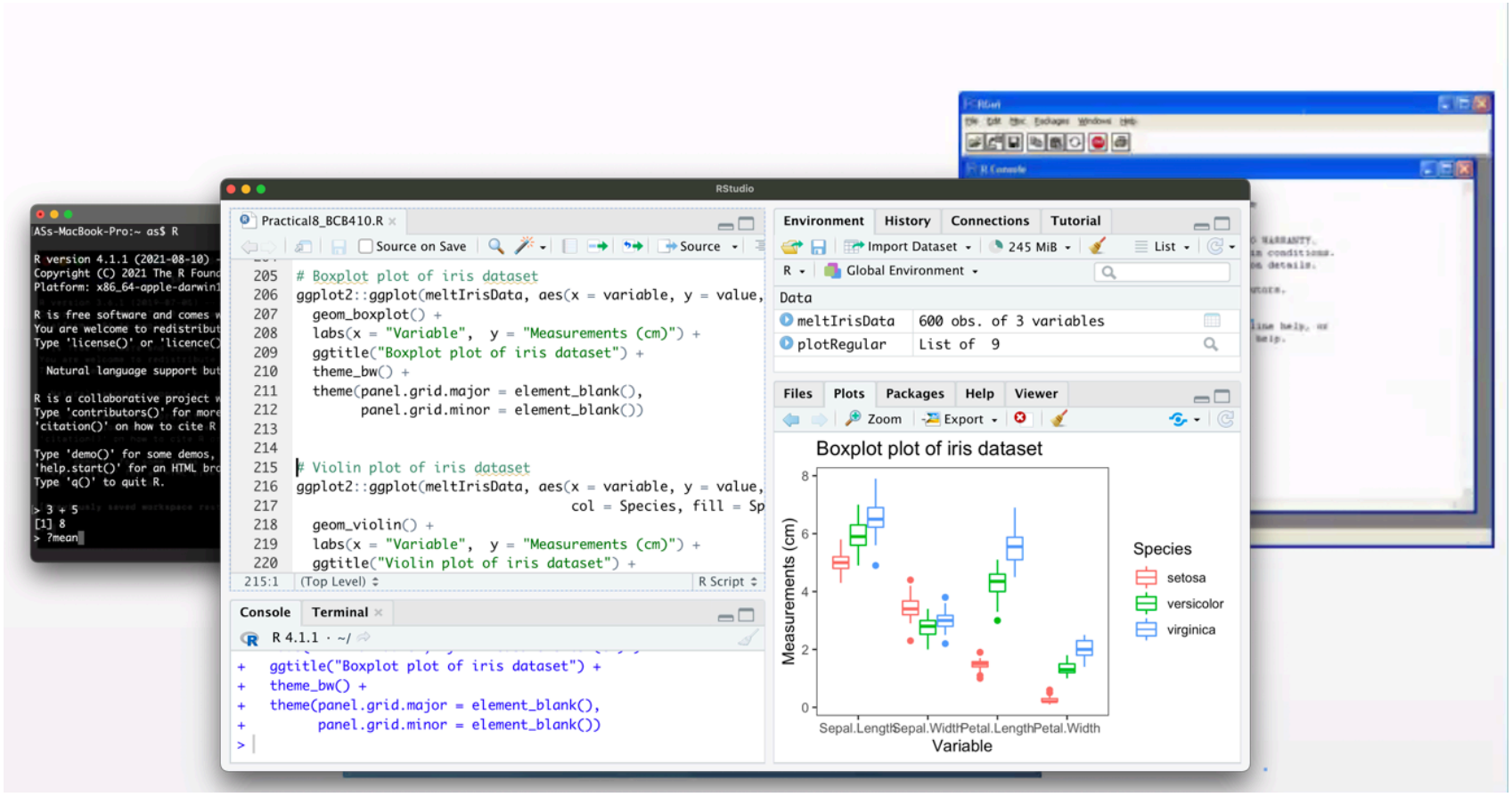


Figure 4: Several other options are present including the Jupyter Notebook and Posit cloud.

Questions?

What can we do with R?

Load Data

```
# A tibble: 9,113 x 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
# i 9,103 more rows
```

What can we do with R?

Clean Data

```
# A tibble: 9,113 x 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
# i 9,103 more rows
```

What can we do with R?

Manipulate and Combine Data

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

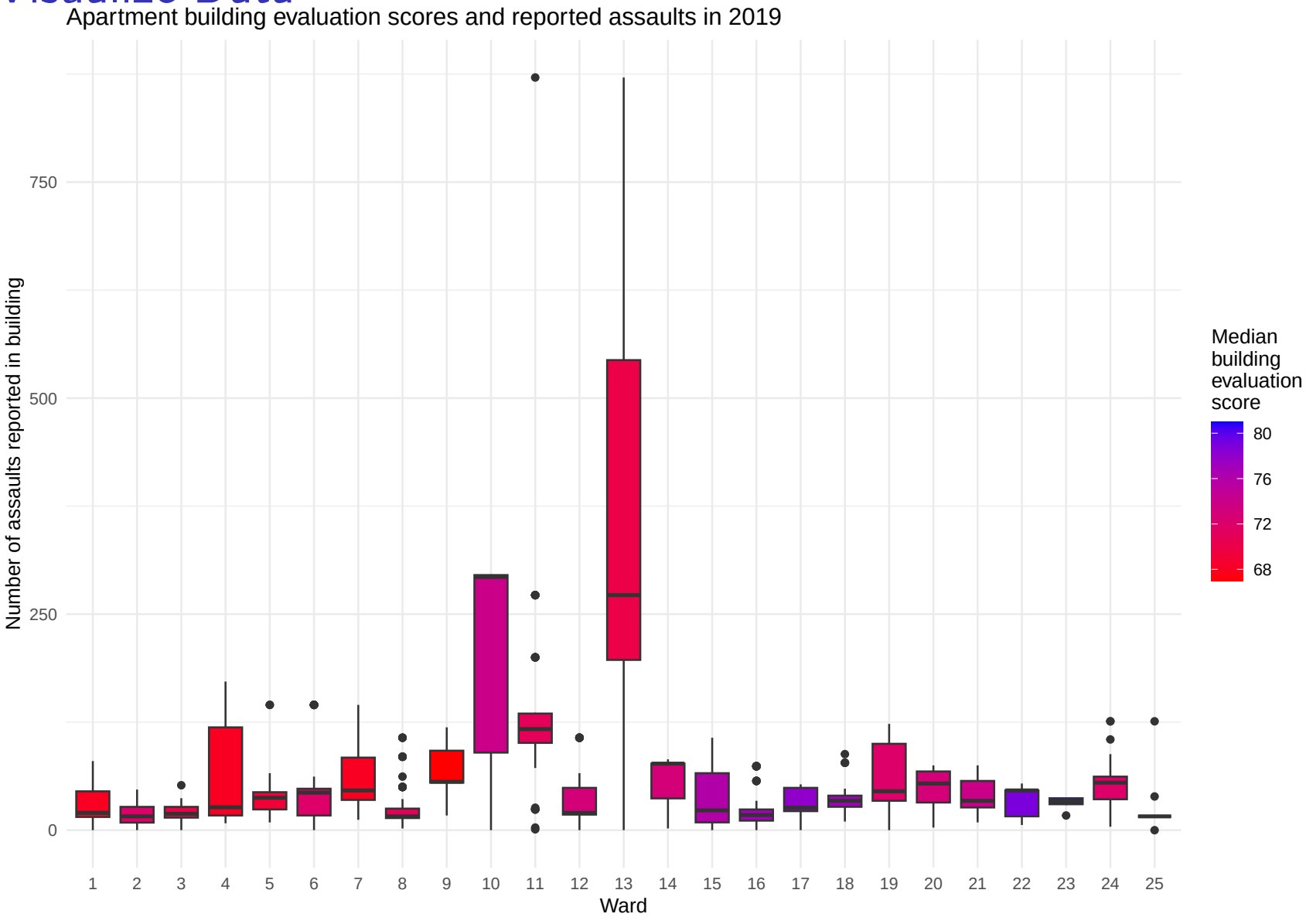
What can we do with R?

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

What can we do with R?

Visualize Data



What can we do with R?

Write reports

Paper title*

Subtitle

Author

Date

Abstract

An abstract

Contents

1 Introduction 1

2 Literature review 2

3 Methodology 2

4 Data 2

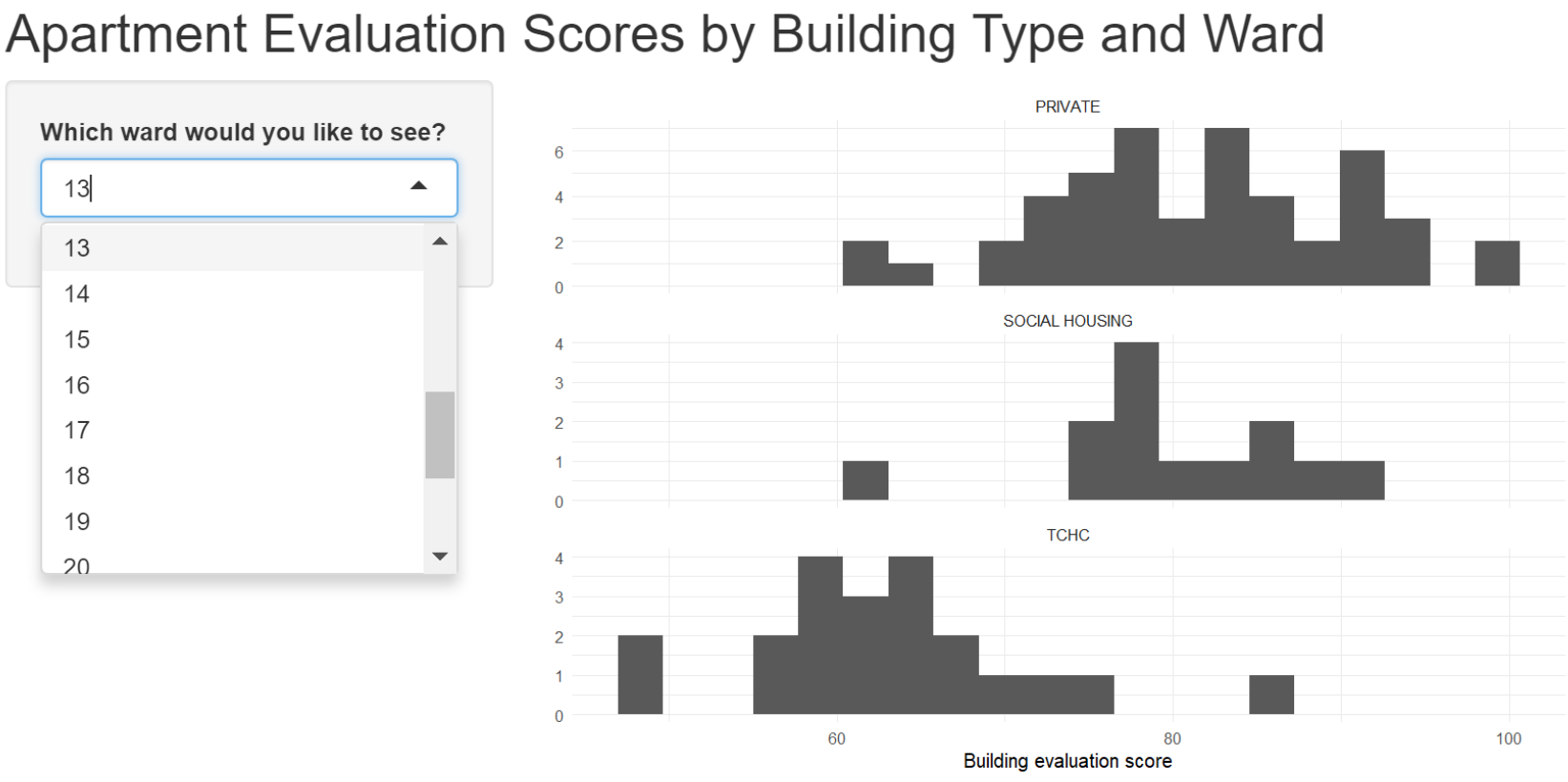
5 Model 2

Conclusion 2

1 Introduction

What can we do with R?

Built interactive applications



What can we do with R?

And more!

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

Questions?