Module 3: R

Introduction

Data Sciences Institute, University of Toronto

2022-06-22

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

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
```

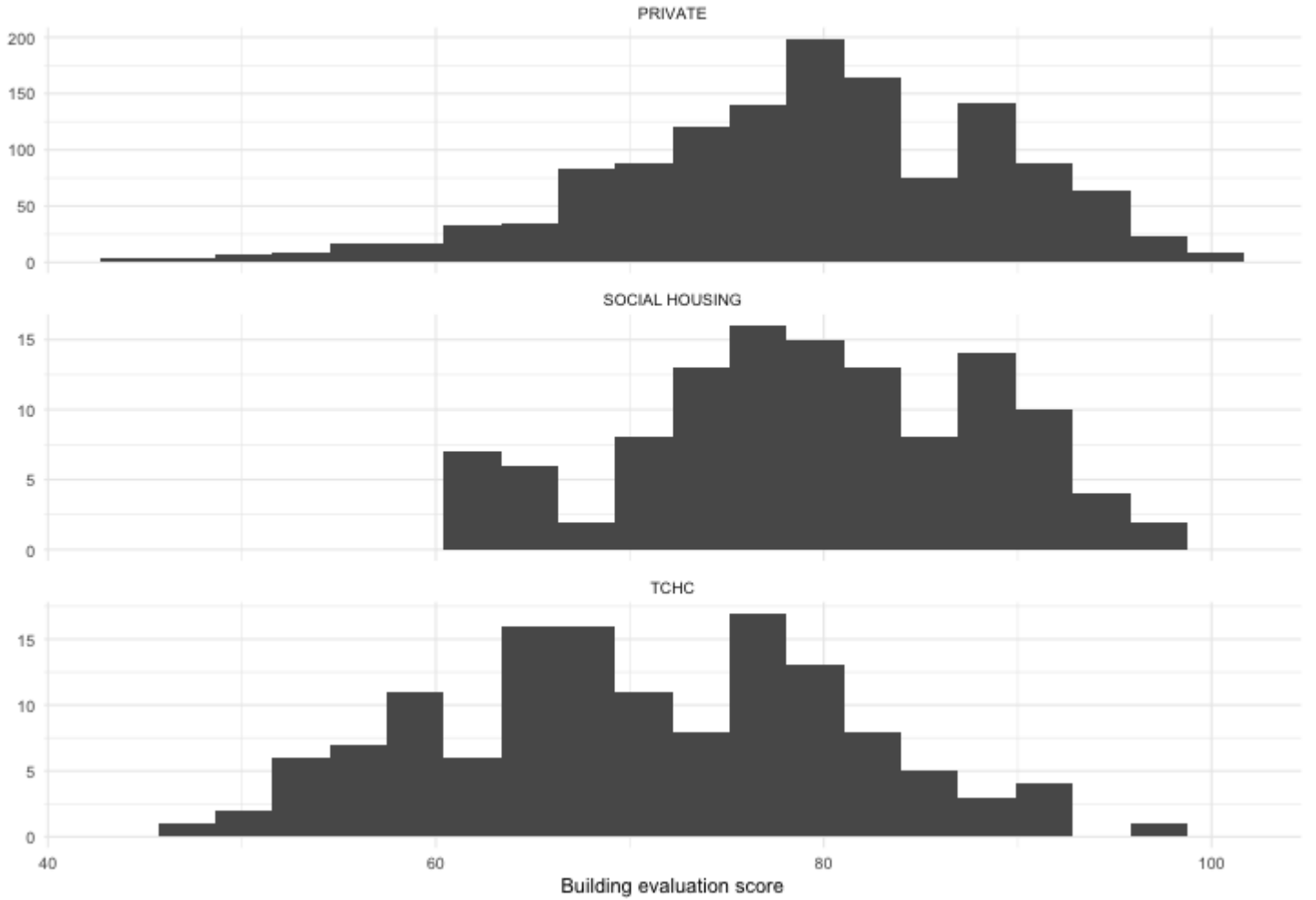
Manipulate and combine data

```
## # A tibble: 8,291 × 6
##    year_built property_type confirmed_units score  year
##         <dbl> <chr>                   <dbl> <dbl> <dbl>
##  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
## # … with 8,281 more rows, and 1 more variable: count <int>
```

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

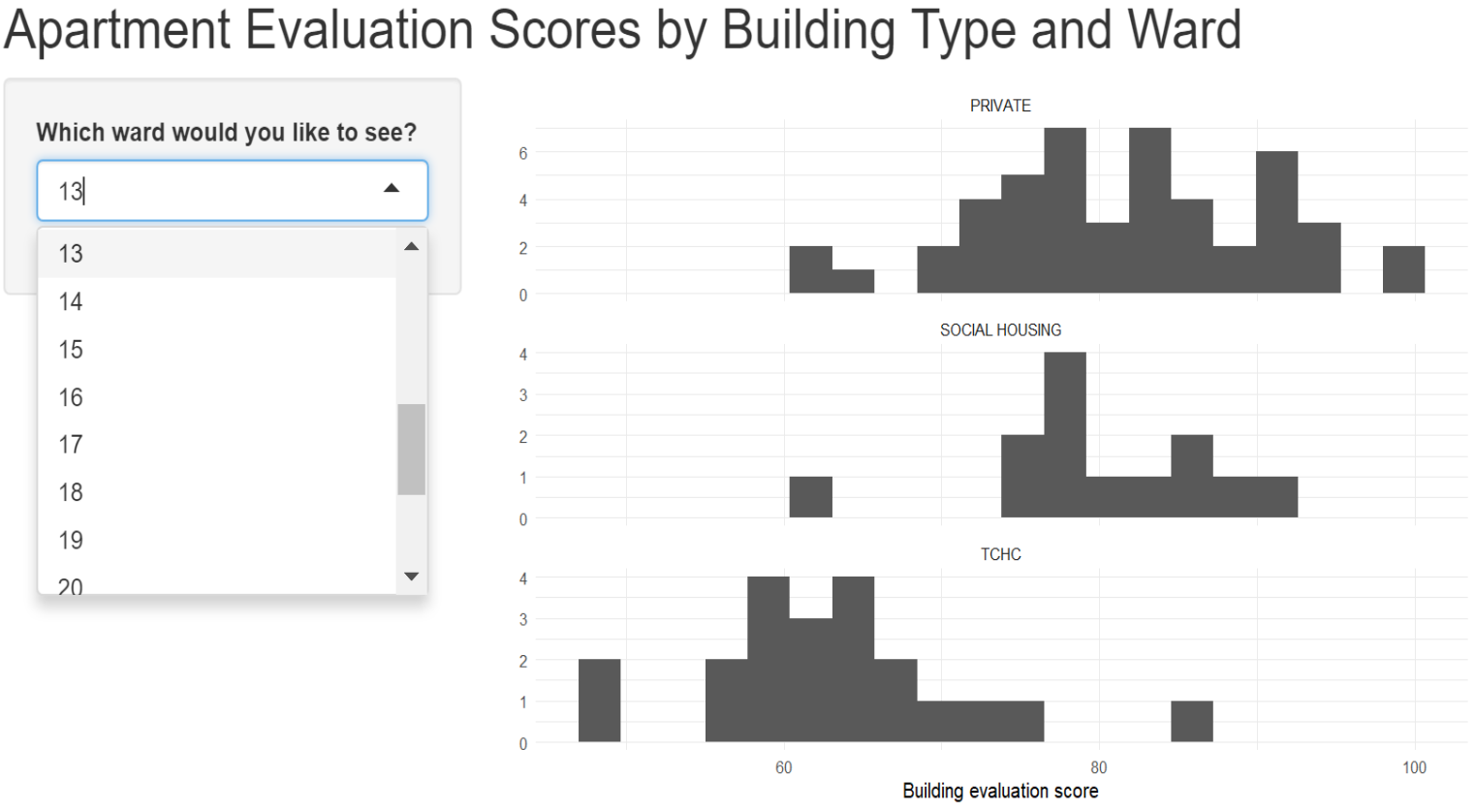
Abstract

An abstract

Contents

1 Introduction

Build Interactive Applications



And more:

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

Content

Learning Outcomes

By the end of this module, you will be able to:

- Set up and use R and RStudio.
- Manipulate and visualize data.
- Fix errors.
- Understand consent in data-based studies.
- Make presentations and manage projects.

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
- Ethics
- Inequity
- Professional Skills
- Industry Case Study

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/.

For specific topics:

Advanced R 4 Data Programming and the Cloud Using PostgreSQL, AWS, and Shiny by Wiley and Wiley (2020), Chapters 7 and 10

Data Visualization: A Practical Introduction by Healy (2018), Chapter 3

Managing Your Data Science Projects: Learn Salesmanship, Presentation, and Maintenance of Completed Models by de Graaf (2019), Chapters 2 and 6

Mastering Shiny by Wickham (2021) https://mastering-shiny.org/index.html, Chapter 1

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

(Advanced)

How can we make interactive applications using R?

Goal:

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

Shiny

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

Ethics

Why does consent matter in data-based studies?

Goal:

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

Ethics

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

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

Inequity

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

Goal:

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

Professional skills

Goals:

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

Industry case study

Delivery

For technical sections:

- Short lectures
- Examples

For non-technical sections:

- Readings
- Discussions

Assessment

Formative

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