Exploring Data Science Education: From Tutorials to Assessment

Duke Statistical Science | Graduation with Distinction

Evan Dragich supervised by Mine Çetinkaya-Rundel, PhD.

April 11, 2023

https://evandragich.github.io/thesis-work/

About Me

- Statistical Science B.S. & Psychology B.S.
- Combination of previous experiences and interest
  - STEM education research
  - Started StatSci sophomore spring
  - TAing Intro Data Science (STA199)

Thesis TOC/Agenda

- Thesis divided into 2 strands:
    - Building a introductory data science concept inventory-style assessment
    - Building `dsbox`, an introductory data science tutorial package
- Agenda
    - Background
    - Initial Steps
    - Interview Process
    - Item Case Studies
    - Package Construction + Examples
    - Discussion
    - Q&A

Building a Data Science Assessment

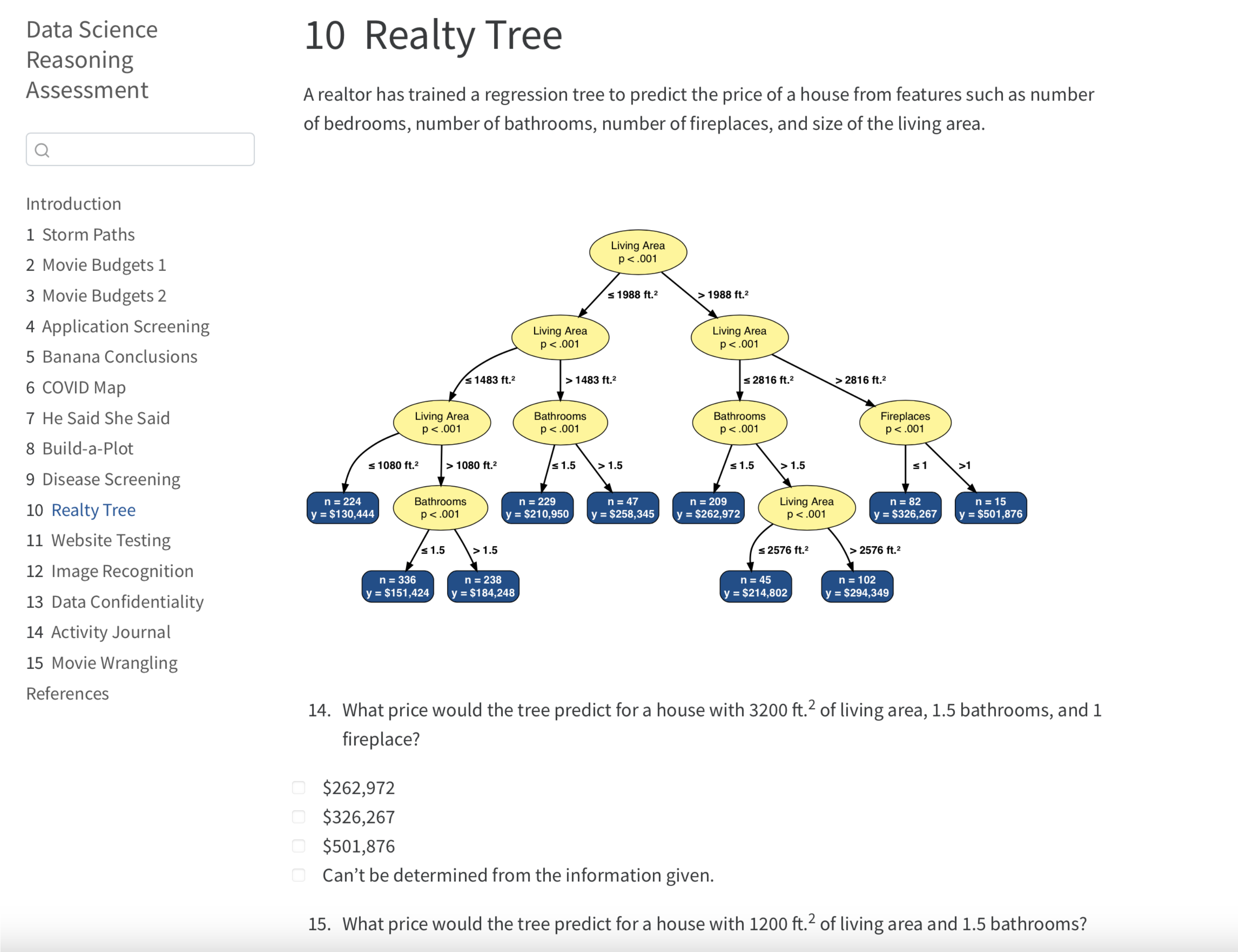
Background

- Concept inventories for educational research
  - CAOS for statistics
- Data science (DS) as it emerges as a field–what is it, exactly?
- How exactly do people: (1) make, (2) pilot, (3) validate new concept inventories or scales?

Initial cleaning

- Combine questions into single set of passages and items
- Draft into Quarto Book for easy browsing

Initial cleaning



Data Science Reasoning Assessment

10 Realty Tree

A realtor has trained a regression tree to predict the price of a house from features such as number of bedrooms, number of bathrooms, number of fireplaces, and size of the living area.

14. What price would the tree predict for a house with 3200 ft.^{2} of living area, 1.5 bathrooms, and 1 fireplace?

☐ $262,972
☐ $326,267
☐ $501,876
☐ Can't be determined from the information given.

15. What price would the tree predict for a house with 1200 ft.^{2} of living area and 1.5 bathrooms?

https://evandragich.github.io/thesis-work/

Interviews

- Two rounds of interviews:
  - 3 faculty
  - 3 intro DS teaching assistants

Interviews: Faculty

- Flow:
  - What topics *must* be in an introductory data science course?
  - What topics are *nice to have* in an introductory data science course?
  - Think-aloud thought process
  - Additional comments or suggestions
  - What are the strengths of the current assessment?
  - What topics are missing from the current assessment?
  - What is in the current assessment, but doesn’t belong?
- Themes from faculty interviews
  - CS vs. Statistics perspectives
  - Context concerns
  - Cognitive load

Interviews: Students

- Flow:
  - Think-aloud thought process
  - Additional comments or suggestions
  - Are the pacing and length appropriate?
  - Based on what you remember learning in intro data science, what topics are missing from the current assessment?
  - Based on what you remember learning in intro data science, what is in the current assessment, but doesn’t belong?
- Themes from student interviews
  - General agreement
  - Gradient of mastery

Current Prototype

- 15 passages, 26 items

Passage	Learning Objective(s)
Storm Paths	modeling; simulation; uncertainty
Movie Budgets 1	compare summary statistics visually
Movie Budgets 2	modeling; R^2; compare trends visually
Application Screening	ethics; modeling; proxy variable
Banana Conclusions	causation; statistical communication
COVID Map	complex visualization; spatial data; time series; sophisticated scales
He Said She Said	basic visualization; sophisticated scales
Build-a-Plot	data to visualization process
Disease Screening	compare classification diagnostics visually
Realty Tree	modeling; regression tree; variable selection
Website Testing	compare trends visually; uncertainty; modeling; time series; extrapolation
Image Recognition	ethics; modeling; representativeness of training data
Data Confidentiality	ethics; data deidentification; statistical communication
Activity Journal	structure data; store data
Movie Wrangling	data cleaning; data wrangling; column-wise string operations; pseudocode; joins

https://evandragich.github.io/thesis-work/

Case Study: Application Screening

You are working on a team that is making a deterministic model to quickly screen through applications for a new position at the company. Based on employment laws, your model may not include variables such as age, race, and gender, which could be potentially discriminatory.

Your colleague suggests including a rule that eliminates candidates with more than 20 years of previous work experience, because they may have high salary expectations. Why might using this variable be considered unethical? Explain your answer.

Case Study: Application Screening

You are working on a team that is making a deterministic model to quickly screen through applications for a new position at the company. Based on employment laws, your model may not include variables such as age, race, and gender, which could be potentially discriminatory.

Your colleague suggests including a rule that eliminates candidates with more than 20 years of previous work experience, because they may have high salary expectations. ***Are there ethical implications of using this variable to select candidates?*** *Explain your answer.*

Case Study: Data Confidentiality

A newspaper reports on the results of a survey from a small (<2000 student) college. The college agrees to have the data released to the public so long as the students' identities and academic standing information are kept confidential. Which of the following combinations of variables is less likely to unintentionally identify any students? Explain.

a. Year, major, sports played

b. Year, major

Case Study: Data Confidentiality

A newspaper reports on the results of a survey from a small (<2000 student) ***university.*** *The* ***university*** *agrees to have the data released to the public so long as the students' identities and academic standing information are kept confidential. Which of the following combinations of variables is less likely to unintentionally identify any students? Explain.*

a. Year, major, sports played

b. Year, major

Case Study: Data Confidentiality

A newspaper reports on the results of a survey from a small (<2000 student) university. The university agrees to have the data released to the public so long as the students' identities and academic standing information are kept confidential. Which of the following combinations of variables is less likely to unintentionally identify any students? Explain.

a. Class year and sports played

b. Student ID and dorm zip code

c. GPA and major

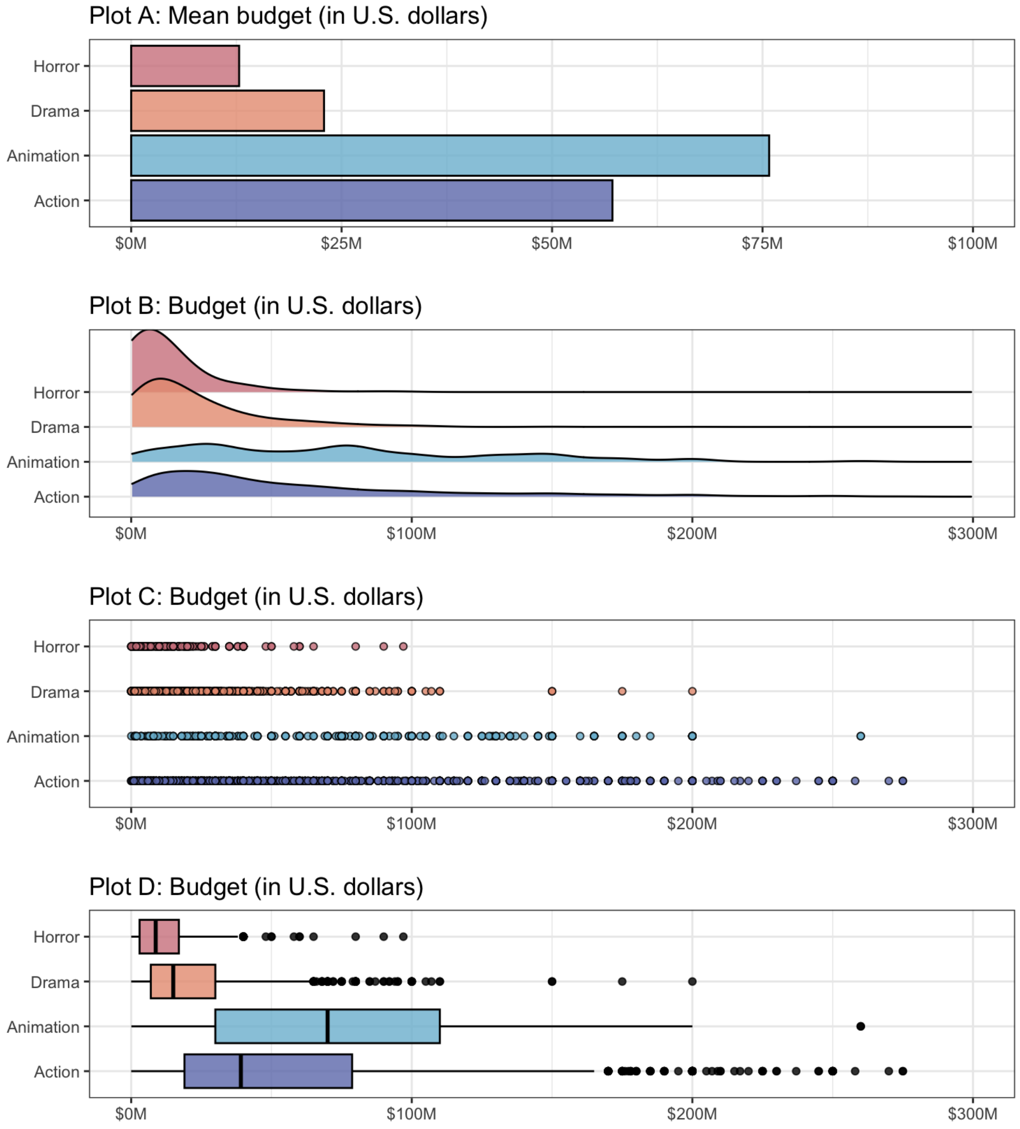
d. Birth date and phone number

e. None of the above

Case Study: Movie Budgets 1

A data scientist at IMDb has been given a dataset comprised of the revenues and budgets for 2,349 movies made between 1986 and 2016.

Suppose they want to compare several distributional features of the budgets among four different genres—Horror, Drama, Action, and Animation. To do this, they create the following plots.



https://evandragich.github.io/thesis-work/

Case Study: Movie Budgets 1

Fill in the following table by placing a check mark in the cells corresponding to the attributes of the data that can be determined by examining each of the plots.

	Plot A	Plot B	Plot C	Plot D
Mean	□	□	□	□
Median	□	□	□	□
IQR	□	□	□	□
Shape	□	□	□	□

Case Study: Movie Wrangling

The table below provides data about 10 movies released in the United States. It provides data on the movie's title , the movie's director, the date the movie was released, the season the movie was released, the worldwide gross intake in U.S. dollars, the cleaned version of the worldwide gross intake in U.S. dollars, and whether or not the movie won the Best Picture Oscar.

Case Study: Movie Wrangling

Movies Table

title	director	release_date	season	gross	gross_clean	best_picture
Almost Famous	Cameron Crowe	22 September 2000	Fall	$47.39M	47.39	No
CODA	Sian Heder	13 August 2021	Summer	$1.61M	1.61	Yes
E.T. the Extra-Terrestrial	Steven Spielberg	11 June 1982	Summer	$792.91M	792.91	No
Luca	Enrico Casarosa	18 June 2021	Summer	$49.75M	49.75	No
Middle of Nowhere	Ava DuVernay	1 September 2014	Fall	$0.24M	0.24	No
Moonlight	Barry Jenkins	18 November 2016	Fall	$65.34M	65.34	Yes
Parasite	Bong Joon Ho	8 November 2019	Fall	$262.69M	262.69	Yes
Say Anything	Cameron Crowe	14 April 1989	Spring	$21.52M	21.52	No
Selma	Ava DuVernay	9 January 2015	Winter	$66.79M	66.79	No
We Bought a Zoo	Cameron Crowe	23 December 2011	Winter	$120.08M	120.08	No

Case Study: Movie Wrangling

The table below provides data about 10 movie directors. It provides data on the director’s name, the number of Oscars the movie’s director has been nominated for, and the number of Oscars the director has won.

Directors Table

director	nominations	oscars
Ava DuVernay	1	0
Barry Jenkins	3	1
Bong Joon Ho	3	3
Cameron Crowe	3	1
Enrico Casarosa	2	0
Loveleen Tandan	0	0
Nora Ephron	3	0
Penny Marshall	0	0
Sian Heder	1	1
Steven Spielberg	19	3

https://evandragich.github.io/thesis-work/

Case Study: Movie Wrangling

```
start_with(the Movies table) then
  keep_rows_where(the season value is "Fall") then
  count(the number of rows)
```

```
start_with(the Movies table) then
  keep_rows_where(the season value is "Fall") then
  add_columns_from(the Director table) matching_by(the
director column) then
  count(the number of rows) where (oscars value is 3) and
(best_picture value is "No")
```

Assessment Next Steps

- 199 Pilot
- IRB Roadblocks
- NSF Grant

Working on the dsbox package

dsbox package

- Growing interest in DS requires scalability
- Data Science in a Box project
- Turning it into dsbox

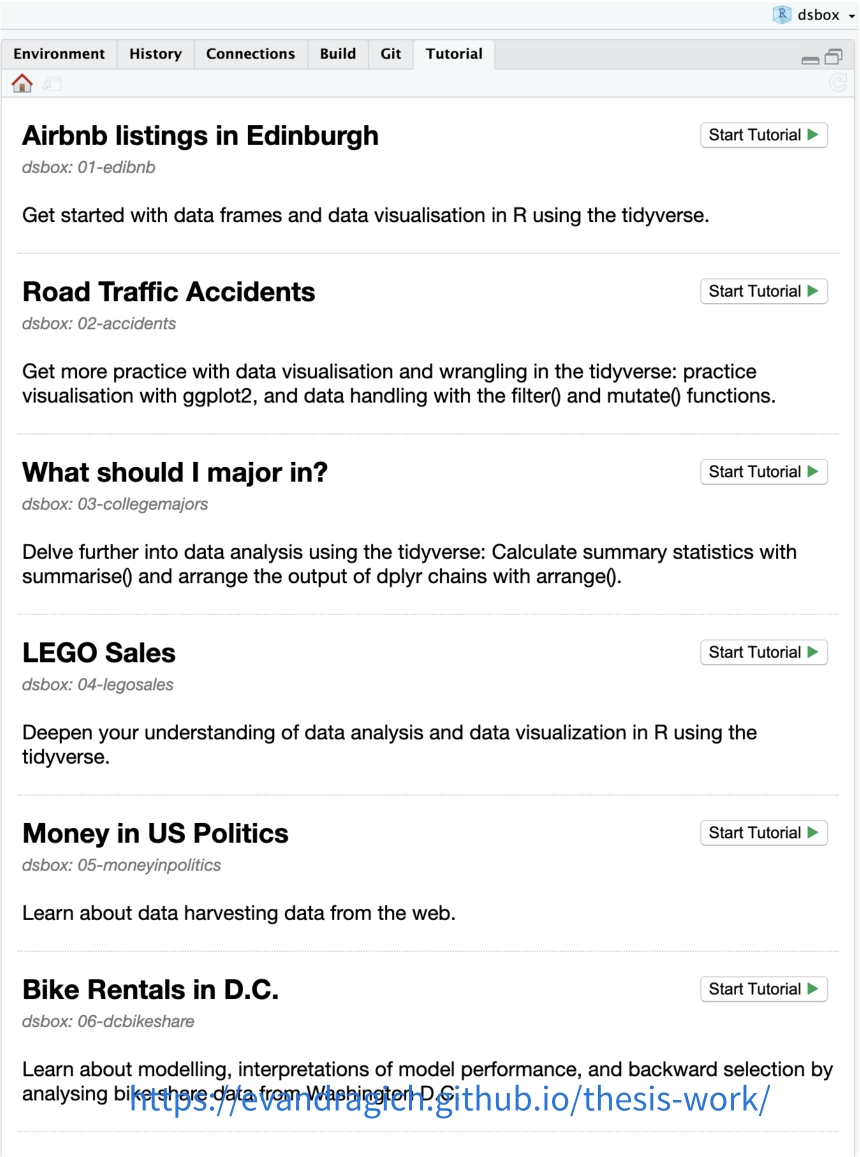
How does it work?

- 2 key packages: `learnr` and `gradethis`.
- `learnr`: robust, broad framework.
- `gradethis`: sophisticated testing logic.

Creating a Tutorial

- 9 existing, 1 started
- Modifying for interactive tutorial
  - Scaffolding, clear section breaks, engaging flow

Sample Tutorial: Home Page



Airbnb listings in Edinburgh
Start Tutorial ▶

dsbox: 01-edibnb

Get started with data frames and data visualisation in R using the tidyverse.

Road Traffic Accidents
Start Tutorial ▶

dsbox: 02-accidents

Get more practice with data visualisation and wrangling in the tidyverse: practice visualisation with ggplot2, and data handling with the filter() and mutate() functions.

What should I major in?
Start Tutorial ▶

dsbox: 03-collegemajors

Delve further into data analysis using the tidyverse: Calculate summary statistics with summarise() and arrange the output of dplyr chains with arrange().

LEGO Sales
Start Tutorial ▶

dsbox: 04-legosales

Deepen your understanding of data analysis and data visualization in R using the tidyverse.

Money in US Politics
Start Tutorial ▶

dsbox: 05-moneyinpolitics

Learn about data harvesting data from the web.

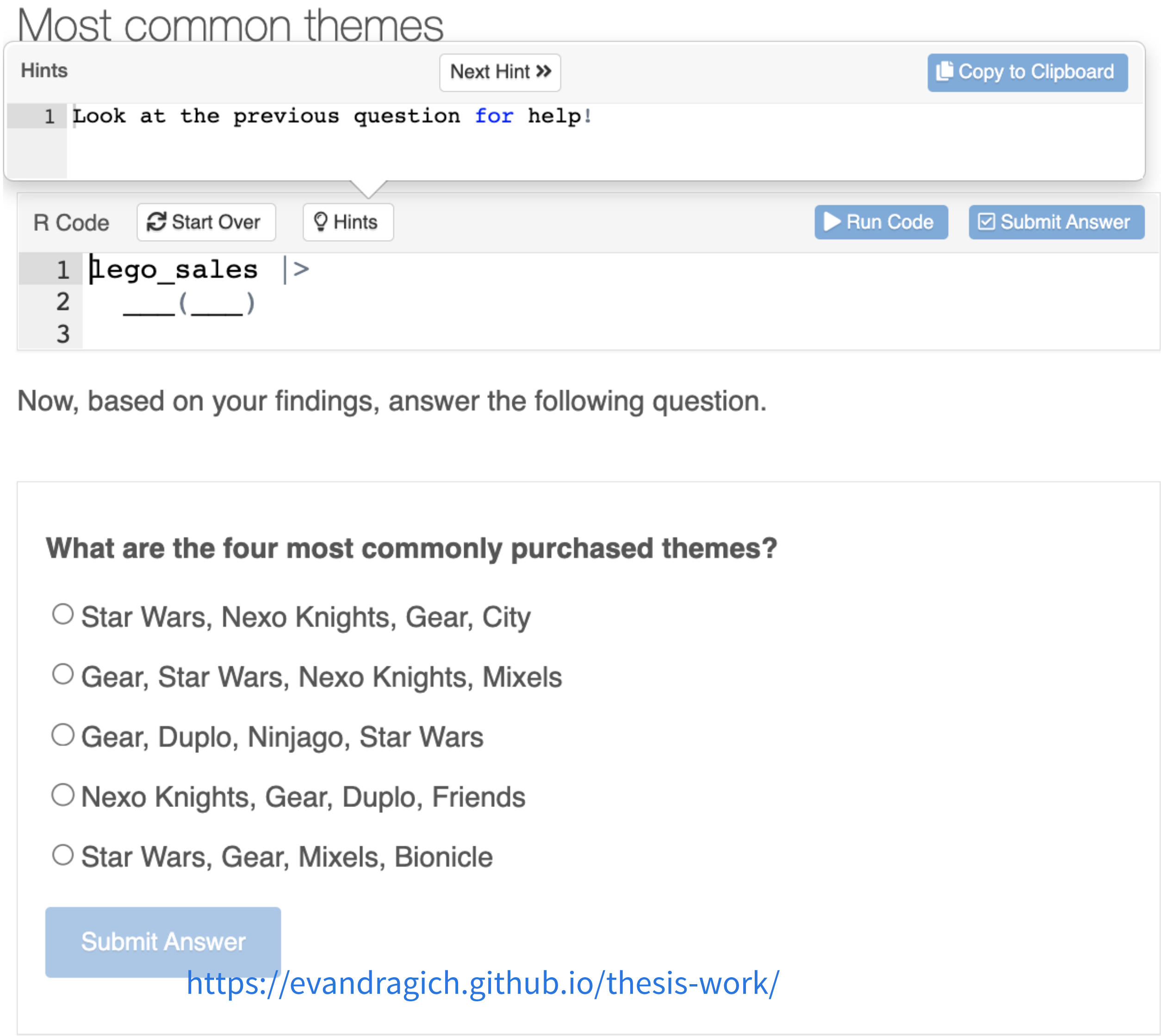
Bike Rentals in D.C.
Start Tutorial ▶

dsbox: 06-dcbikeshare

Learn about modelling, interpretations of model performance, and backward selection by analysing bike share data from Washington D.C.

https://evandragich.github.io/thesis-work/

Sample Tutorial: Code chunk with hint



Most common themes

```
Look at the previous question for help!
```

```
lego_sales |>
  ___(___)

```

Now, based on your findings, answer the following question.

What are the four most commonly purchased themes?

○ Star Wars, Nexo Knights, Gear, City

○ Gear, Star Wars, Nexo Knights, Mixels

○ Gear, Duplo, Ninjago, Star Wars

○ Nexo Knights, Gear, Duplo, Friends

○ Star Wars, Gear, Mixels, Bionicle

https://evandragich.github.io/thesis-work/

Sample Tutorial: Opening the hood

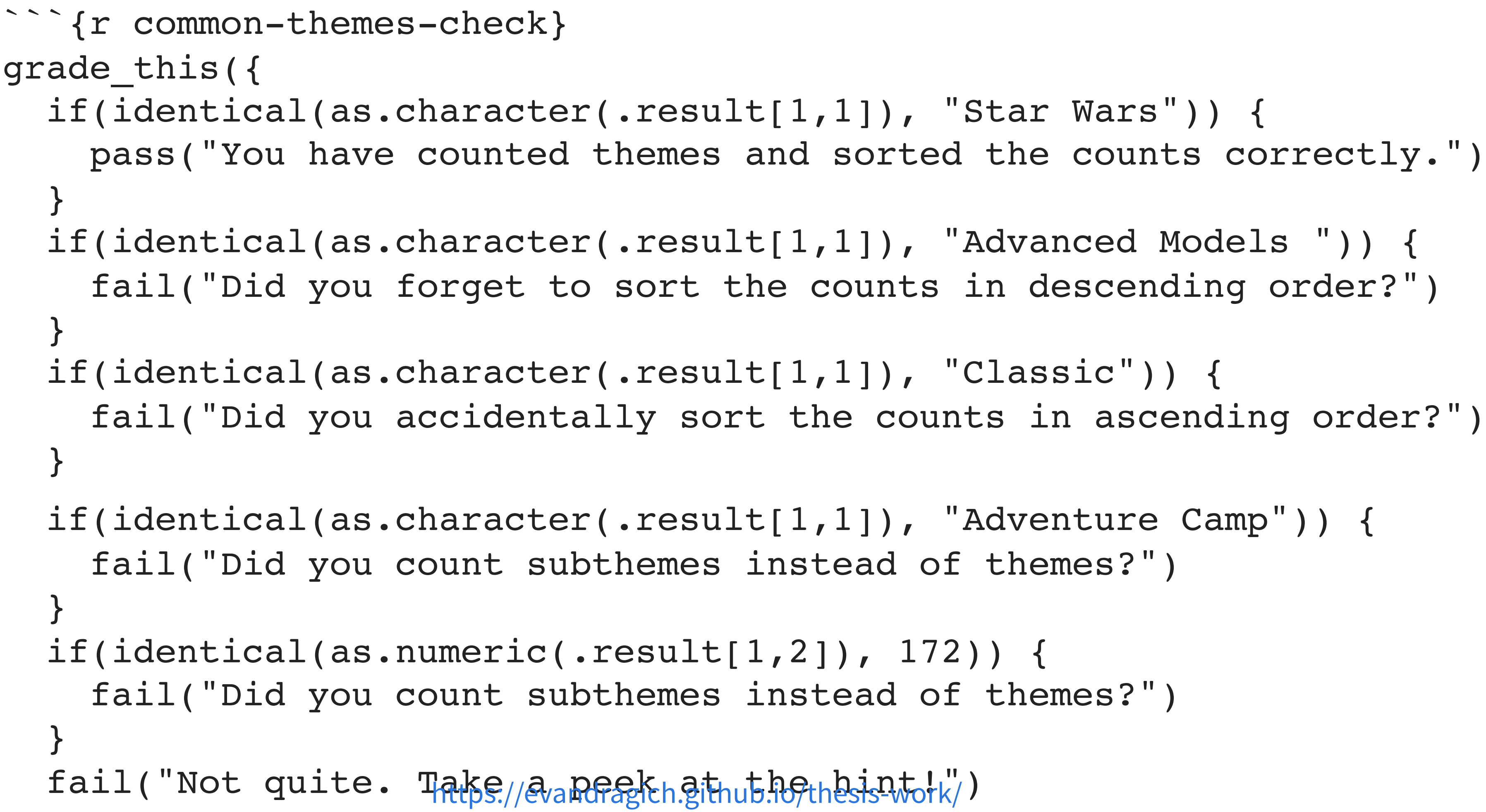
~~~
```{r common-themes, exercise = TRUE}
lego_sales |>
  ___(___)
```
~~~

~~~
```{r common-themes-hint-1}
Look at the previous question for help!
```
~~~

~~~
```{r common-themes-solution}
lego_sales |>
  count(theme, sort = TRUE)
```
~~~

https://evandragich.github.io/thesis-work/

Sample Tutorial: Opening the hood



~~~
```{r common-themes-check}
grade_this({
  if(identical(as.character(.result[1,1]), "Star Wars")) {
    pass("You have counted themes and sorted the counts correctly.")
  }
  if(identical(as.character(.result[1,1]), "Advanced Models ")) {
    fail("Did you forget to sort the counts in descending order?")
  }
  if(identical(as.character(.result[1,1]), "Classic")) {
    fail("Did you accidentally sort the counts in ascending order?")
  }
  if(identical(as.character(.result[1,1]), "Adventure Camp")) {
    fail("Did you count subthemes instead of themes?")
  }
  if(identical(as.numeric(.result[1,2]), 172)) {
    fail("Did you count subthemes instead of themes?")
  }
  fail("Not quite. Take a peek at the hint!")
~~~

Releasing to CRAN

- Comprehensive R Archive Network
- Package DESCRIPTION file
- `gradethis` still in development

Discussion

Learning Takeaways

- Advanced computing
- Interacting with others’ code

Reflections

- “Teaching material is only way to master it”
- New appreciation for existing educational materials and research
- Inspired me to continue interacting with the world of open source software

Q&A

- Browse at your own pace at https://evandragich.github.io/thesis-work/
- Email me at evandragich@gmail.com

https://evandragich.github.io/thesis-work/