zyBooks
A Wiley Brand

Exploring Student Struggle in Introductory Data Science Courses

Aimee Schwab-McCoy

Content Manager, Data Science and Statistics

Electronic Conference on Teaching Statistics

June 2024

Data Science @ zyBooks

zyBooks:

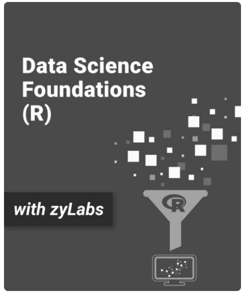
- Fully-interactive, web-native textbooks designed to encourage active reading
- Frequent, meaningful feedback

Data Science Foundations:

- Three versions: **Python, R,** non-programming
- Concepts first, programming second

Data Science Foundations with Python

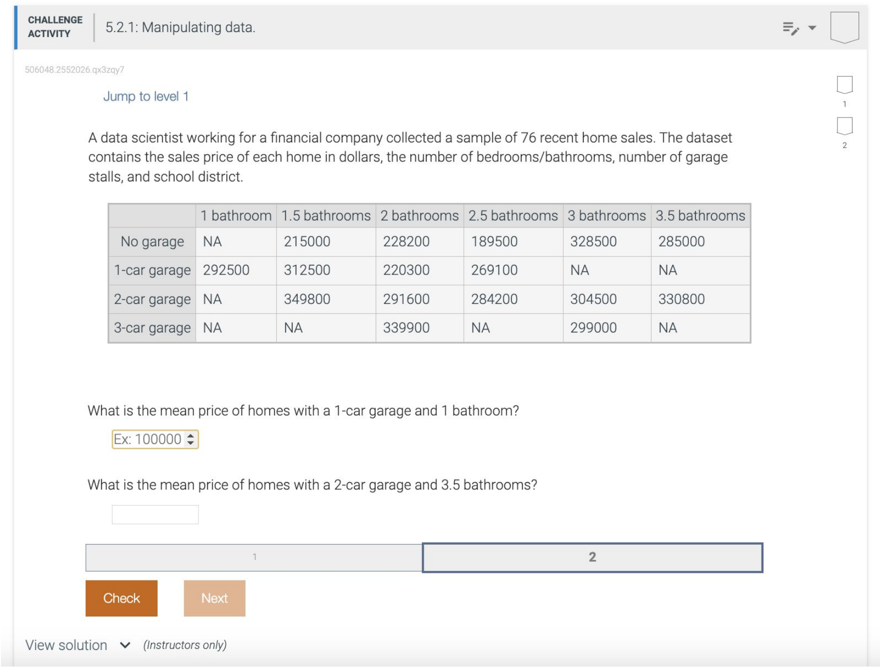
with zyLabs


Data Science Foundations (R)

with zyLabs

Data Science Foundations

Challenge activities



5.2.1: Manipulating data.

506048.2552026.qx3zqy7

Jump to level 1

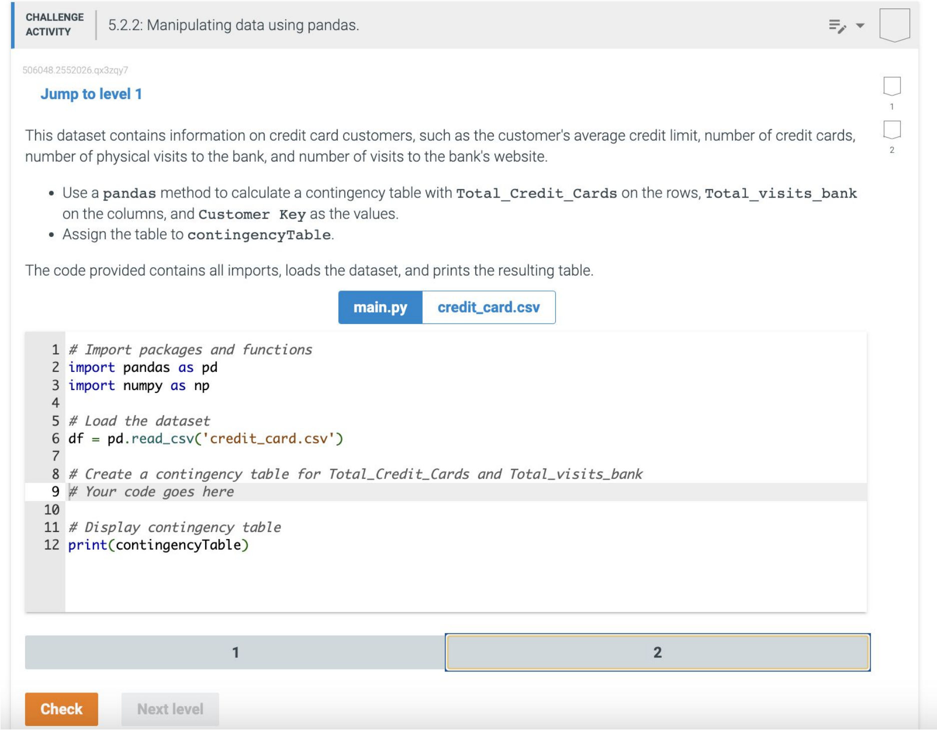
A data scientist working for a financial company collected a sample of 76 recent home sales. The dataset contains the sales price of each home in dollars, the number of bedrooms/bathrooms, number of garage stalls, and school district.

	1 bathroom	1.5 bathrooms	2 bathrooms	2.5 bathrooms	3 bathrooms	3.5 bathrooms
No garage	NA	215000	228200	189500	328500	285000
1-car garage	292500	312500	220300	269100	NA	NA
2-car garage	NA	349800	291600	284200	304500	330800
3-car garage	NA	NA	339900	NA	299000	NA

What is the mean price of homes with a 1-car garage and 1 bathroom?

Ex: 100000

What is the mean price of homes with a 2-car garage and 3.5 bathrooms?



5.2.2: Manipulating data using pandas.

506048.2552026.qx3zqy7

Jump to level 1

This dataset contains information on credit card customers, such as the customer's average credit limit, number of credit cards, number of physical visits to the bank, and number of visits to the bank's website.

- Use a `pandas` method to calculate a contingency table with `Total_Credit_Cards` on the rows, `Total_visits_bank` on the columns, and `Customer Key` as the values.
- Assign the table to `contingencyTable`.

The code provided contains all imports, loads the dataset, and prints the resulting table.

```python
# Import packages and functions
import pandas as pd
import numpy as np

# Load the dataset
df = pd.read_csv('credit_card.csv')

# Create a contingency table for Total_Credit_Cards and Total_visits_bank
# Your code goes here

# Display contingency table
print(contingencyTable)
```

Research questions

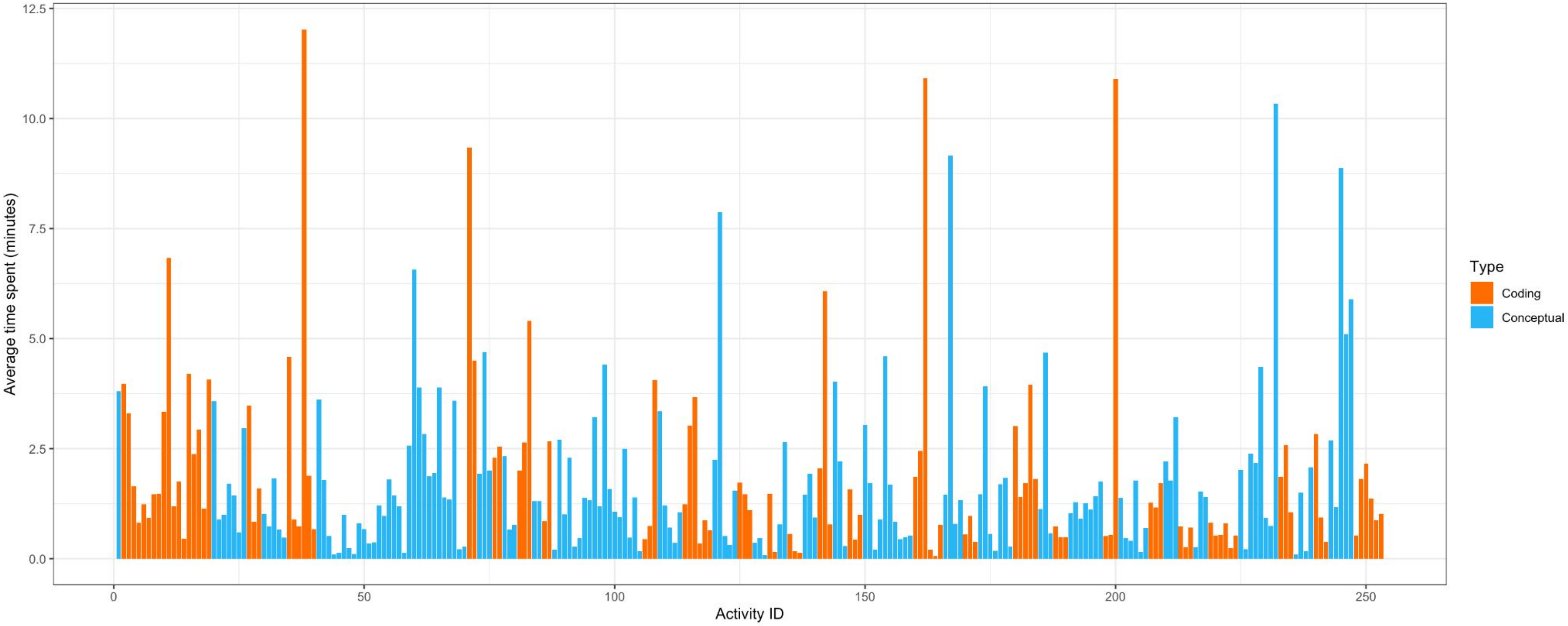
1. Which topics or sections do students struggle in?

2. How does struggle compare for Python vs. R?

Defining student struggle

- **Time spent:** Do students spend more time on a challenge activity?
  - *High time spent = more "struggle"*
- **Attempts:** Do students need more attempts to successfully complete a challenge activity?
  - *More attempts = more "struggle"*
- **Completion**: Do students eventually complete a challenge activity?
  - *Lower completion rate = more "struggle"*

Ex: Time spent (minutes)



Methods

Struggle: Data for students using latest releases (Aug. 23)

- "Top 10" coding and challenge activities flagged using each metric
- Activities with at least one struggle flag were reviewed

Python CAs	R CAs	Non-programming CAs
39 flagged/253 total levels	39 flagged/238 total levels	18 flagged/122 total levels

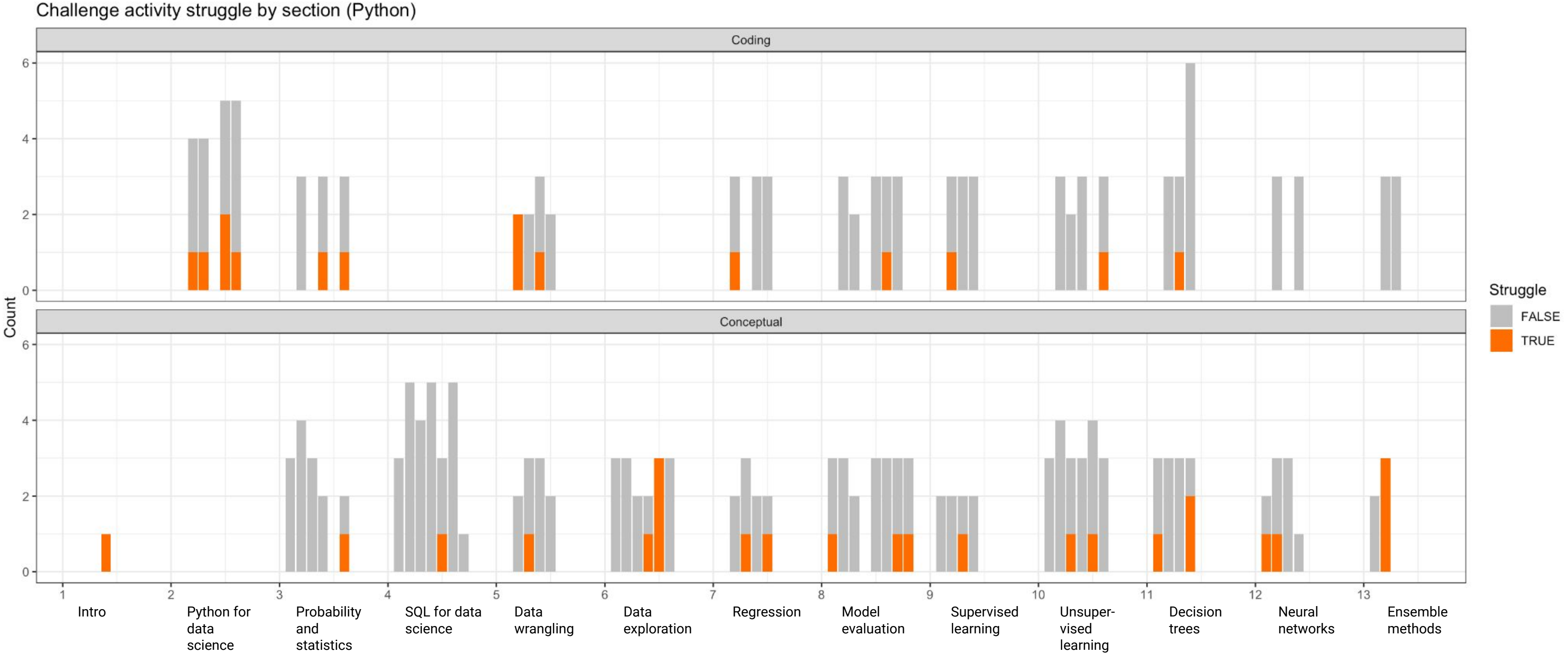
Disclaimers

Struggle is not necessarily bad.

- Struggle can be productive! We don't want assessments to be too easy.
- Struggle metrics are sensitive to outliers and uncontrollable factors.

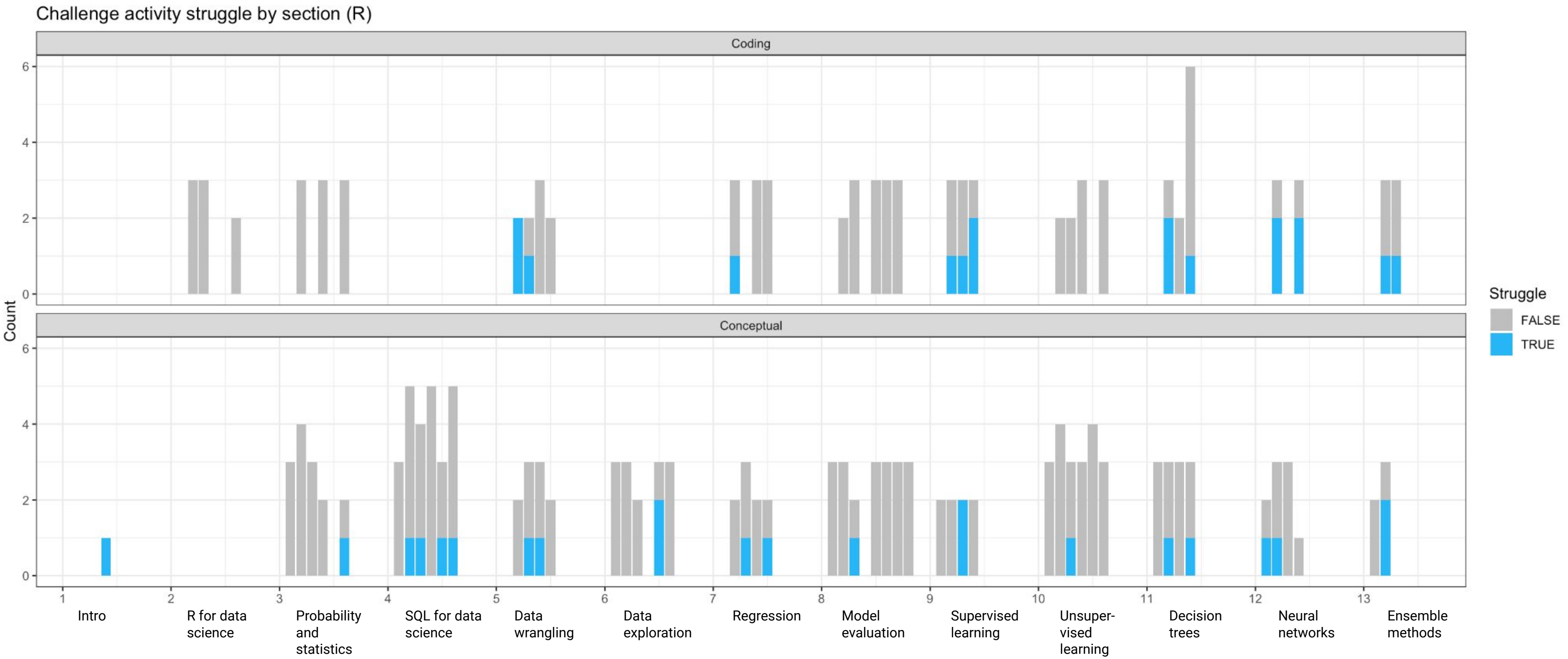
Our "struggle" is your "heads up".

Struggle by section



zyBooks
A Wiley Brand

Struggle by section



Conceptual tasks

- Writing null and alternative hypotheses
- Calculating standardized values
- Identifying outliers
- Describing histograms
- Logistic regression (log-odds and probabilities)
- Model selection (cross-validation)
- Fitted value vs. residual plots
- Interpreting dendrograms
- Interpreting decision trees

Coding tasks

Python

- Creating and manipulating arrays (NumPy)
- Frequency tables and contingency tables (pandas)
- Initializing models (scikit-learn)

R

- Frequency tables and contingency tables (dplyr)
- Most tasks using tidymodels

Limitations

- Website clicks are only a proxy for student struggle
  - Doesn't account for distractions, engagement, multitasking, etc.
- Certain activity types have more struggle by design
  - Ex: "Check all that apply" vs. multiple choice vs. calculations
- Challenge activities are not the only form of assessment
  - We know what's happening across classes, but not in class

zyBooks
A Wiley Brand

Questions?

Aimee Schwab-McCoy

Content Manager, Data Science and Statistics

aimee.schwab-mccoy@zybooks.com

aschwabmcc@wiley.com