Unit 2 homework instructions

DKU Stats 101 Fall 2025 Session 2

Anonymous

2025-11-05

Figure 1: Bookshelf

Scoring guide

Content

- The information requested in the question prompts are only a starting point, if you find other interesting information along the way, please report that. You don't need to look at the data forever but if there is obviously something else interesting in the data you should report it.
- You must have up to Question 2 completed for the homework check on November 9

- You do not need to be an expert on books for a good score, but I will expect you to look up basic information, such as "what is a good review score on Amazon?", "what is an ISBN number?" and "what are the most popular genres of books on Amazon?" and so on to help you understand and set expectations your data.
- The information requested in the question prompts are only a starting point, if you find other interesting information along the way, please report that. You don't need to look at the data forever but if there is obviously something else interesting in the data you should report it.

Technical

- Make sure your graphs are produced using `ggplot()`, are well labeled, and are easy to read.
- Make sure your tables are produced with the `kable()` function from the `knitr` package, are well labeled, and are easy to read. You can make your tables prettier with the `kableExtra` package.
- Make sure you do not have anything rendered in your PDF file besides your results and, when asked for by a question, your code. That means no warnings, messages, or other output should appear in your final rendered PDF file.
- Make sure to accurately mark each page a question answer appears on when submitting on GradeScope.

Books on Amazon.

Introduction

Question 1: Describing your data (10 points)

1a. Where is this data from?

For this dataset, describe the data according to the five Ws & *how* defined in the textbook Chapter 1.2. What are some possible problems with the *who* and *what* of the dataset?

The dataset you are using for this assignment is a subset of the original dataset that can be found here.

1b. What are the variable types?

For the following variables, please make a table.

One column should be the variable name, the second should be the variable type as defined in the textbook Chapter 1.3, and the third the units of the variable (if applicable).

- `title`
- `published_date`
- `format`
- `page_count`
- `isbn_10`
- `category`
- `average_rating`
- `price`
- `features_text`

Question 2: Association (20 points)

Using the `mutate()` verb as described in the DataCamp lab, make a new variable called `age` that subtracts the year 2025 from the book's `publish_year`. Please display your code using `#| echo: true` code block option.

2a. Investigating `age` vs. `average_rating`

Using the Think-Show-Tell framework from the textbook (example on page 213), please examine the relationship *in association terms* between `age` and `average_rating`. How strongly are they associated?

Note: for this question and all other Think sections in the homework, you do not need to report the W's of the data (you have already completed this in Q1)

Think

For this section, please write down your expectations, why you expect it, the variable meaning, and, given the variable type, the best way to display the data

Show

For this section, please make an appropriate graph or table and briefly describe what you observe

Tell

Please interpret the meaning of your finding here, especially with respect to your expectation

2b. Investigating `age` vs. `rating_number`

Using the Think-Show-Tell framework from the textbook, please examine the relationship *in association terms* between `age` and 'rating_number". How strongly are they associated?

Think

For this section, please write down your expectations, why you expect it, the variable meaning, and, given the variable type, the best way to display the data

Show

For this section, please make an appropriate graph or table and briefly describe what you observe

Tell

Please interpret the meaning of your finding here, especially with respect to your expectation

2c. Thinking about your results

Consider the results of 2a. and 2b. together. What can we understand about how books are rated on Amazon from this information? What do you think explains the relationships you have identified?

Complete up to here for Homework Check - due November 9th at 11:59 pm.

Question 3: Simple regression (20 points)

3a. Investigating `price` vs. `average_rating`

Using the Think-Show-Tell framework from the textbook, please examine how the price of a book is related to the rating of a book.

Think

Show

Tell

3b. Checking model fit

Make use of all the tools described in the textbook to assess model fit in the `Think again` section - if it is necessary to revise your model, do it in the `Think again` section. Then state any updated conclusions in the `Revising conclusions` section.

Think again

Revising conclusions

3c. Investigating `price` vs. `item_weight`

Similar to 3a. and 3b., fully analyze the relationship between price and the weight of the book.

Think

Show

Tell

Think again

Revising conclusions

3d. Thinking about your results

What can we learn about how price is determined in these two investigations? Do the results surprise you? What lurking variables do you think could be at work here, if any?

Question 4: Multiple regression (30 points)

4a. Investigating `average_rating` vs. `price`, `rating_number`, and `age`

Using the Think-Show-Tell framework from the textbook, please examine, using a multiple regression model, how `average_rating` relate to `price` and `rating_number`. Make use of all the tools described in the textbook to assess model fit in the `Think again` section - if it is necessary to revise your model, do it in the `Think again` section. Then state any updated conclusions in the `Revising conclusions` section.

Think

Show

Tell

Think again

Revising conclusions

4b. Interpreting coefficients of 4a. model

Carefully interpret your coefficients from 4a. What do they mean? Are there any lurking variables here?

Think

Show

Tell

4c. Add the variable `category`

Now add the variable `category` to your model and analyze the relationship similar to what you did in 4a.

Think

Show

Tell

Think again

Revising conclusions

4d. Reinterpret your coefficients

Carefully re-interpret your coefficients from 4c and compare them to 4b. What do they mean? Any new lurking variables to consider?

Think

Show

Tell

4e. Thinking about your results

Consider the results of 4a.-4d. together. What can we learn about the ratings of books on Amazon? How did your conclusions change from 3d.? Why do you think they changed?

Question 5: Your own investigation (20 points)

5a. Selecting your own question

Develop your own model of `average_rating`. Use the Think-Show-Tell procedure to conduct your investigation. Think deeply about what your result means and interpret your coefficients carefully.

Think

Show

Tell

Think again

Revising conclusions

5b. In summary

Sum up everything that you have learned from questions 1-5. Do not simply repeat/rephrase your previous results but try to say something larger that synthesizes the results together to draw a more meaningful general conclusion.