Unit 1 homework instructions

DKU Stats 101 Spring 2025 Session 3

AUTHOR
Anonymous

PUBLISHED
January 19, 2025

Scoring guide

Content

- Getting the right answer is only a small part of the grade
- Good quality interpretation of your results is the name of the game
- If you see something that looks unusual in your data (outlier, some unusual distribution type) - investigate it!
- When explaining your results, say something interesting about them. Did it match your expectations? Why or why not?
- Brief explanations that simply repeat what I can visually see myself will not receive a good score
- On the other hand, filling the homework with pages of not very interesting description is not valuable either. The goal isn't to write the most words, but find the most interesting things in the data.
- You do not need to be an expert in art for a good score, but I will expect you to look up basic information, such as "what does art typically sell for?" and "what is a standard size for paintings"? 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 HTML 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 HTML file.
- Convert your HTML file to PDF using the Microsoft Print to PDF option in the Print menu (PC) or the PDF button option from the Print menu (Mac)
- Make sure to accurately mark each page a question answer appears on when submitting on GradeScope.

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?

1b. What are the variable types?

For the following variables, please list the variable type as defined in the textbook Chapter 1.3:

- `artist`
- `country`
- `yearOfBirth`
- `name`
- `year`
- `ageOfPaiting`
- `price`
- `material`
- `height`
- `dominantColor`

Question 2: Displaying and describing the data (15 points)

For the moment, we are going to focus on paintings by Chinese artists. You can create a subset of your data using the `filter()` verb as you learned in the DataCamp lab.

2a. Filtering your data

Using the `filter()` verb as described in the DataCamp lab, make a subset of your data that only includes art from Chinese artists. Show the code you used to make the subset using the `#| echo: true` code block option.

2b. Investigating height

Using the Think-Show-Tell framework from the textbook (example on page 71), investigate the distribution of the height of the Chinese paintings

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

> Show

> Tell

2c. Investigating width

Using the Think-Show-Tell framework from the textbook, investigate the distribution of the width of the Chinese paintings

> Think

> Show

> Tell

2d. Thinking about your results

Consider the results of 2b. and 2c. together. What can we understand about Chinese art from viewing the distribution of these two variables?

Question 3: Relationships between categorical variables - American and Chinese artists and oil vs. ink. (15 points)

3a. Recoding your data

Using the `mutate()` verb and the `case_when()` verb combined with `grepl()`, create two new variables. The first is `material.type` and the second is `us.china`. The first variable should recode material to be either `Oil`, `Ink`, or `Other`, depending on whether the original values of `material` contained either the words `oil` or `ink`. The second variable should make a similar transformation to `country` where you recode the variable to be either `American`, `Chinese`, or `Other`. Show the code you used to make the new variables using the `#| echo:` true code block option.

Hint 1: you can see some examples of `case_when()` and `grepl()` hereand here .

Hint 2: make sure to use the `ignore.case=TRUE` option in `grepl()`

3b. Investigating the categorical relationship between `us.china` and `material.type`

Investigate the relationship between `us.china` and `material.type`

Hint 3: you can see an example of some ways to display this information here

> Think

> Show

> Tell

3c.Thinking about your results

Think carefully about why you have observed this result and provide some additional information about what this investigation means for understanding this dataset and art in general.

Question 4: Comparing groups (15 points)

4a. Recoding your data

Similar to the previous question, create a new variable called `famous.countries` that recodes country to be either `American`, `French`, `Italian` and `Spanish`. Mark art from all other countries as `NA` (the code that stands for missing or not available in R). Additionally, create a new variable called `area` that is a calculation of the area of the art (height times width). Show the code you used to make the new variables using the `#| echo:` true code block option.

4b. Compare the groups of countries on the variable `price`

> Think

> Show

> Tell

4c. Compare the groups of countries on the variable `area`

> Think

> Show

> Tell

4d. Thinking about your results

Consider the results of 4b. and 4c. together. What can we learn about the differences in art between the countries? What do you think causes these differences or similarities? How would you confirm your guess as to the cause of the differences/similarities?

Question 5: Considering deviations (10 points)

5a. Selecting your data

Pick three years of paintings to investigate whether the brightness of paintings has changed over time. You are free to pick any three years but you should pick years that correspond to different periods in art history. State the three years and justify your selection.

5b. Finding the average

Calculate the average brightness for each of the three years. Show your code using the `#| echo:` true code block option.

5c. Normalizing the data

Find how many \(z\) units each of the averages for the years are away from the overall mean of brightness and interpret your results.

> Think

> Show

> Tell

5d. Thinking about your results

What are some of the implications of your findings with regard to the motivation of this question? What are some of the limitations of this analysis? What other kind of analysis would you like to do to answer this question?

Question 6: Your own investigation (15 points)

6a. Selecting your own question

Similar to the previous questions, think of your own question that you would like to ask of the data. Use the Think-Show-Tell procedure to conduct your investigation. Think deeply about what your result means.

> Think

> Show

> Tell

6b. In summary

Sum up everything that you have learned in this investigation. 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.