Unit 1 homework instructions

DKU Stats 101 Spring 2025 Session 3

AUTHOR
Runpeng Cao

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?

The original dataset can be found here.

Who: This dataset contains artists, which created the paintings. The dataset consists of 37,638 art pieces sold at a total valuation of $9.47 billion. What: This dataset includes characteristics of the paintings such as artists, countries, year of birth, material, price, etc. When: Inspired on last November, until 2025. Where: Mostly focus on Chinese and the U.S. artists. Why: The dataset creater have been curious about how the commercial art auction market evaluates art pieces, so the creater turned his curiosity and passion for art into this data science project. How: The creater divided dataset into two subsets: 7 famous artists and 7399 less known artists. And built machine learning models and used different combinations of features for each subset. For this demonstration, the creater primarily focus on the 7 famous artists model.

1b. What are the variable types?

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

- `artist` It's categorical by names of artists, and nominal
- `country` It's categorical by countries of origin for the artist, and nominal
- `yearOfBirth` It's numeric by integer values representing the year of birth, and discrete

- `name` It's categorical by name of the artwork, and nominal
- `year` It's numeric by integer representing the year the painting was created, and discrete
- `ageOfPaiting` It's numeric by typically calculated as the difference between the year of creation and the current year, and continuous
- `price` It's numeric by representing the price of the artwork, which can have decimal values, and continuous
- `material` It't categorical by different materials used in creating the artwork, and nominal
- `height` It's numeric by measured in some unit, and continuous
- `dominantColor` It's categorical by color classification, and nominal

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.

```
library(dplyr)
chinese_art <- art_data %>%
 filter(country == "China")
head(chinese_art)
```

```
 [1] artist             country            yearOfBirth        yearOfDeath
 [5] name               year               ageOfPainting      price
 [9] material           height             width              link
[13] source             dominantColor      brightness         ratioUniqueColors
[17] thresholdBlackPerc highbrightnessPerc lowbrightnessPerc  CornerPer
[21] EdgePer            FaceCount          soldtime
<0 rows> (or 0-length row.names)
```

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) We are interested in the data of height of Chinese paintings, which might have huge differences depending on tthe artist, the genre, or the historical time periods. The size of the paintings might dominate different art styles. Therefore, it's important to consider whether there's a typical size range of the Chinese artwork. One typical way we've learnt in class to demonstrate this distribution visually would be with a histogram or boxplot. Accordint to the class activities and learning experiences, we oftenly use ggplot2 to visualize the height distribution.

```
library(ggplot2)
ggplot(chinese_art, aes(ex = height)) +
 geom_histogram(bins = 20, fill = "red", color = "black") +
 labs(title = "Distribution of Painting Heights", x = "Height (cm)", y = "Frequency")
```

2c. Investigating width

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

We are interested in the data of height of Chinese paintings, which might have huge differences depending on tthe artist, the genre, or the historical time periods. The size of the paintings might dominate different art styles. Therefore, it's important to consider whether there's a typical size range of the Chinese artwork. One typical way we've learnt in class to demonstrate this distribution visually would be with a histogram or boxplot. Accordint to the class activities and learning experiences, we oftenly use ggplot2 to visualize the height distribution.

```
library(ggplot2)
ggplot(chinese_art, aes(ex = height)) +
 geom_histogram(bins = 20, fill = "red", color = "black") +
 labs(title = "Distribution of Painting Heights", x = "Height (cm)", y = "Frequency")
```

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?

The heights and widths of the paintings are mostly distributed in certain ranges, this result may suggest that Chinese art, no matter of the artists or the produced time period, has typical size preferences.

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()`

```
library(dplyr)
art_data <- art_data %>%
mutate(
 material.type = case_when(
      grepl("oil", material, ignore.case = TRUE) ~ "Oil",
      grepl("ink", material, ignore.case = TRUE) ~ "Ink",
      TRUE ~ "Other"),
us.china = case_when(
      grepl("USA", country, ignore.case = TRUE) ~ "American",
      grepl("China", country, ignore.case = TRUE) ~ "Chinese",
      TRUE ~ "Other"))
```

```
Warning: There were 4 warnings in `mutate()`.
The first warning was:
ℹ In argument: `material.type = case_when(...)`.
Caused by warning in `grepl()`:
! unable to translate 'etching_on_Arches_<c3>' to a wide string
ℹ Run `dplyr::last_dplyr_warnings()` to see the 3 remaining warnings.
```

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

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

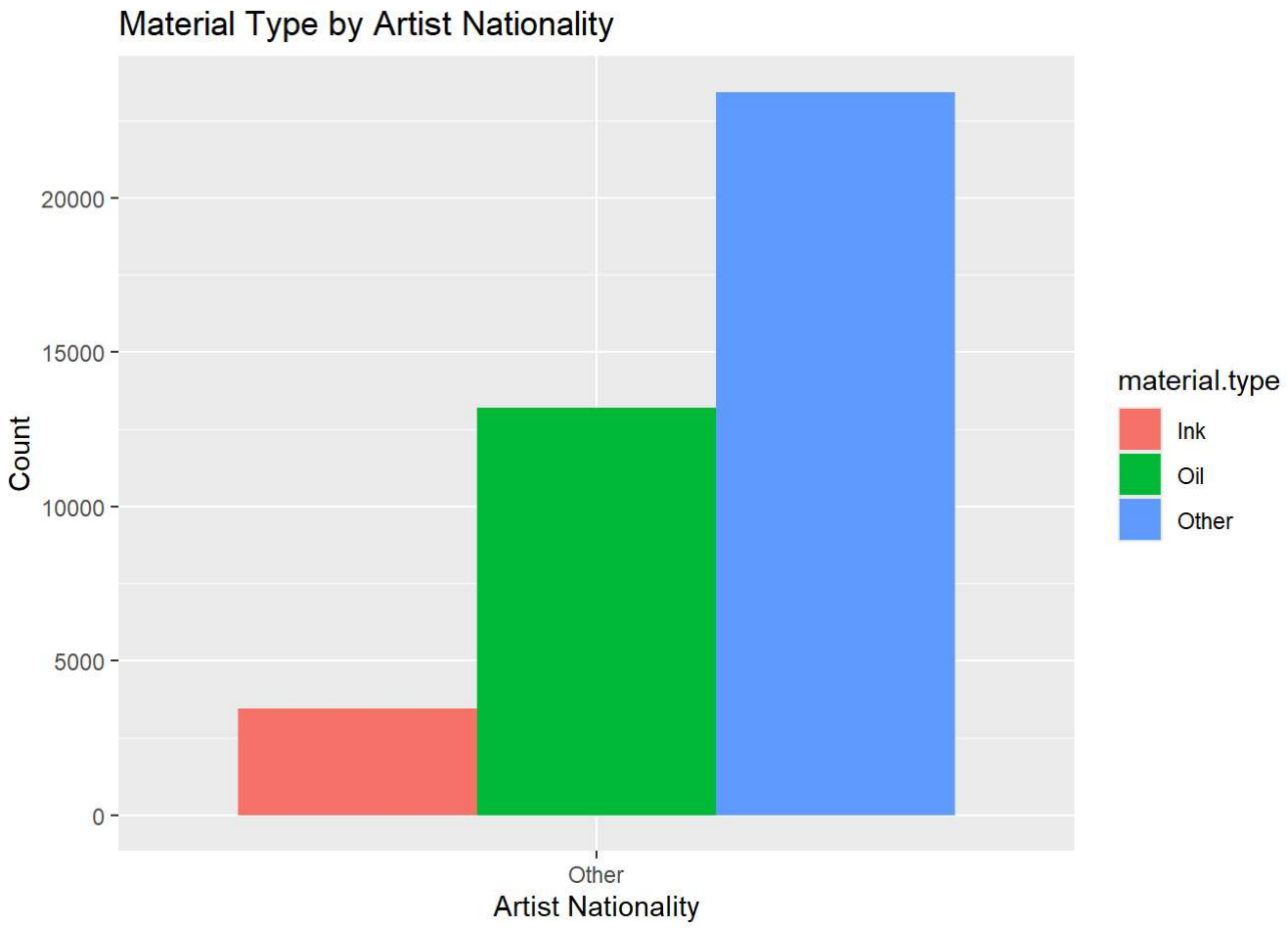
Hint 3: you can see an example of some ways to display this information here

Since certain materials might be more common among American artists, while other materials might be more common among Chinese artists, due to cultural and historical preferences for mediums like oil or ink. Therefore, I want to explore the pattern of artists choice of the painting material. We can display the relationship between us.china and material.type using a contingency table or a bar plot.

```
table(art_data$us.china, art_data$material.type)
```

```
         Ink   Oil Other
  Other  3432 13208 23430
```

```
library(ggplot2)
ggplot(art_data, aes(x = us.china, fill = material.type)) +
 geom_bar(position = "dodge") +
 labs(title = "Material Type by Artist Nationality" , x = "Artist Nationality", y = "Count")
```




```
theme_minimal
```

```
function (base_size = 11, base_family = "", base_line_size = base_size/22,
    base_rect_size = base_size/22)
{
    theme_bw(base_size = base_size, base_family = base_family,
        base_line_size = base_line_size, base_rect_size = base_rect_size) %+replace%
        theme(axis.ticks = element_blank(), legend.background = element_blank(),
            legend.key = element_blank(), panel.background = element_blank(),
            panel.border = element_blank(), strip.background = element_blank(),
            plot.background = element_blank(), complete = TRUE)
}
<bytecode: 0x0000021acd7bf948>
<environment: namespace:ggplot2>
```

The plot I've generated shows that most American artists in this dataset use oil as their material, while a significant portion of Chinese artists use ink. Other materials are less common across both groups.

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.

The use of oil painting being more dominant in Western art, while the ink are more oftenly used for a traditional medium in Chinese art. These differences might varies for the culture and the ablity of local sourcing. Chinese art is much older than American art, especially traditional ink painting. The time of the fastest development of oil painting art in the United States basically coincides with the Western Renaissance.

Question 4: Comparing groups (15 points)

4a. Recoding your data

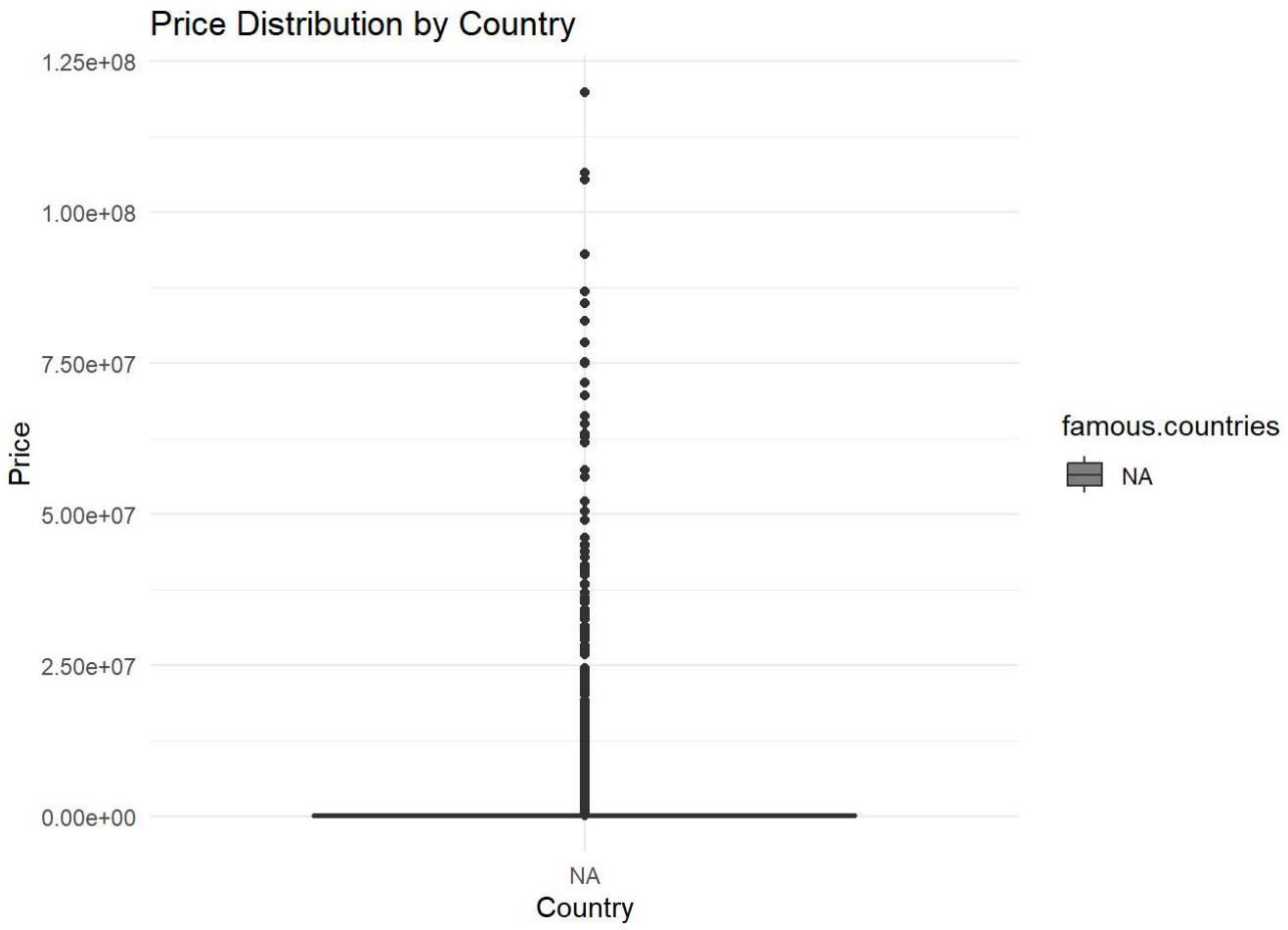
Similar to the previous question, create a new variable called `famous.countries` that recodes country to be either `American`, `French`, `Italian` and `Spanish`. Mark art from all other countries as `NA` (the code that stands for missing or not available in R). Additionally, create a new variable called `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.

```
art_data <- art_data %>%
  mutate(
    famous.countries = case_when(
      grepl("USA", country, ignore.case = TRUE) ~ "American",
      grepl("France", country, ignore.case = TRUE) ~ "French",
      grepl("Italy", country, ignore.case = TRUE) ~ "Italian",
      grepl("Spain", country, ignore.case = TRUE) ~ "Spanish",
      TRUE ~ NA),
    area = height * width)
```

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

There's no doubt that different countries have different value recognition for artwork. In my persipective, the more well-known, developed, and influential countries tend to charge higher prices for works of art, due to factors such as demand, market size, and historical significance. Therefore, we can compare prices across countries using a box plot.

```
ggplot(art_data, aes(x = famous.countries, y = price, fill = famous.countries)) +
  geom_boxplot() +
  labs(title = "Price Distribution by Country", x = "Country", y = "Price") +
  theme_minimal()
```

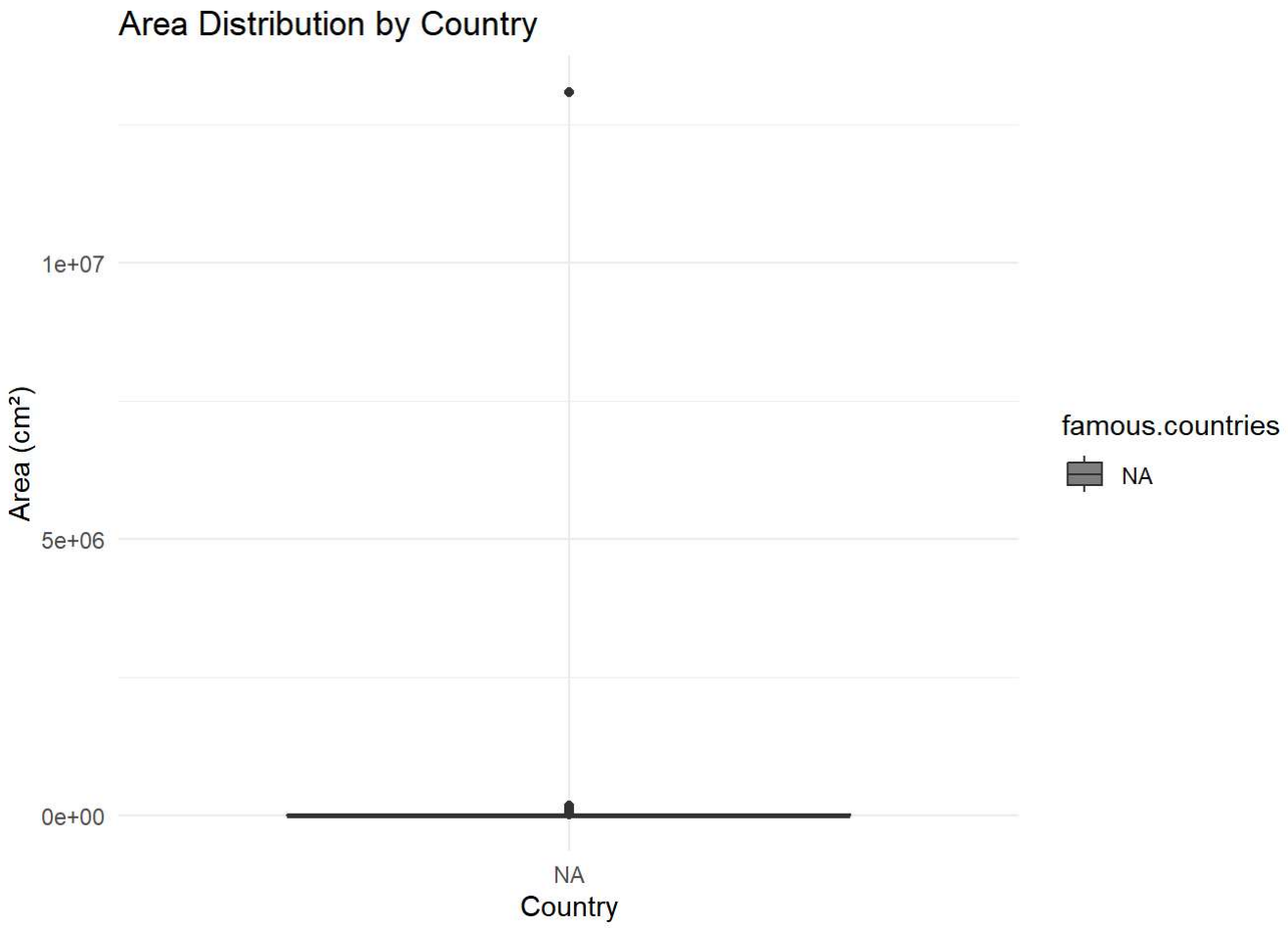


4c. Compare the groups of countries on the variable `area`

The canvas size of the artwork may also have different preferences depending on the country. This may be due to social factors, the personal preferences of artists, as well as the complexity of preserving paintings, and so on. Therefore, I want to look for patterns related to artistic traditions, so I will also use a box plot.

```
ggplot(art_data, aes(x = famous.countries, y = area, fill = famous.countries)) +
  geom_boxplot() +
  labs(title = "Area Distribution by Country", x = "Country", y = "Area (cm²)") +
  theme_minimal()
```

```
Warning: Removed 2419 rows containing non-finite outside the scale range
(`stat_boxplot()`).
```



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?

The differences between the prices of the artworks might related with economic factors, for example, the U.S. has a well economic market, so the price of the artworks is more likely to become higher. On the other hand, the prices might related with historical value as well, though it's subjective. For example, an artwork in China that has longer history and has been well protected is more likely to has a higher price. In order to figure out other connections between prices, area, cultural preferences, historical values, and other factors, further investigation on the data is necessary.

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.

The three year I select is 1900, 1950, and 2000. Because I want my selected data includes 100 years, for 100 years is a long and reasonable time interval in which a lot can happen and a lot can change. So I chose the year with the whole hundred, which are 1900 and 2000, which can both be founded in the original dataset. In order to better see the trend of change, I chose 1950 in the middle of the two selected years.

5b. Finding the average

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

```
brightness_data <- art_data %>%
  filter(year %in% c(1900, 1950, 2000)) %>%
  group_by(year) %>%
  summarise(average_brightness = mean(brightness, na.rm = TRUE))

brightness_data
```

```
# A tibble: 3 × 2
   year average_brightness
  <int>              <dbl>
1  1900               139.
2  1950               152.
3  2000               144.
```

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.

Sorry I really don't know the answer to this question so I used the ChatGPT as hint for this one and here's what I've got.

```
overall_mean <- mean(art_data$brightness, na.rm = TRUE)
overall_sd <- sd(art_data$brightness, na.rm = TRUE)

brightness_data <- brightness_data %>%
  mutate(
    z_score = (average_brightness - overall_mean) / overall_sd)

brightness_data
```

```
# A tibble: 3 × 3
   year average_brightness z_score
  <int>              <dbl>   <dbl>
1  1900               139. -0.146
2  1950               152.  0.0931
3  2000               144. -0.0611
```

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?

I didn't get the graph for this one... Some error happened and I can't figure out why. But I guess the brightness has changed significantly across the three years because of social art event during the 1900s.

Question 6: Your own investigation (15 points)

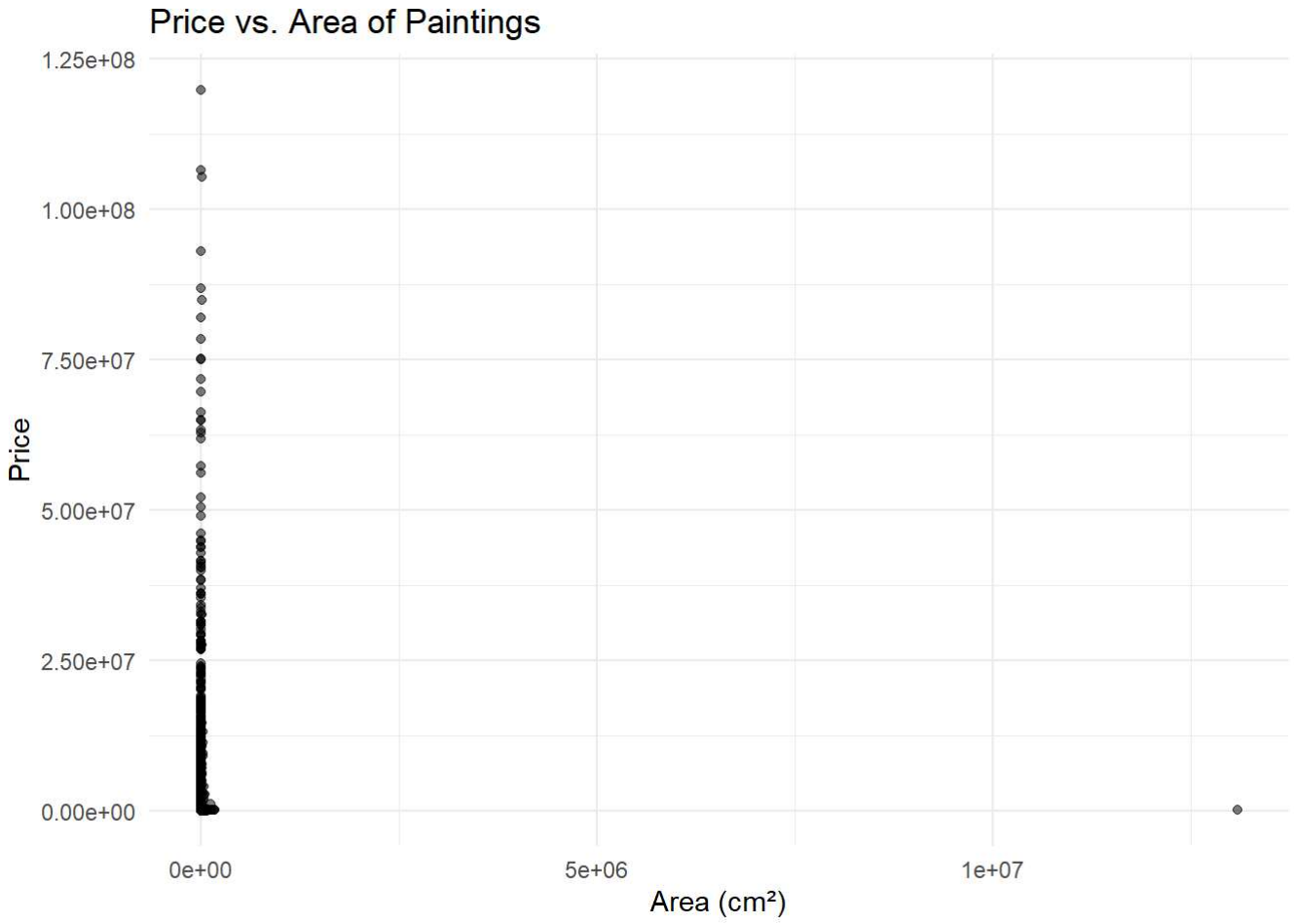
6a. Selecting your own question

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.

My research question: Is the size of the painting has correlation with the price of the painting?

```
ggplot(art_data, aes(x = area, y = price)) +
  geom_point(alpha = 0.5) +
  labs(title = "Price vs. Area of Paintings", x = "Area (cm²)", y = "Price") +
  theme_minimal()
```

```
Warning: Removed 2419 rows containing missing values or values outside the scale range
(`geom_point()`).
```



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.

1. I notices that different variables such as sizes, countries, and so on, was and is still having a strong impact on the artwork market.
2. American artists are more likely to use oil as a medium, while Chinese artists favor ink.
3. Brightness of the paintings varies from year to year, and the indicators are complicated.
4. Larger paintings are more likly to be expensive.
5. Historical and cultural values explain the price or artworks a lot.
6. Further research is necessary.