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?

Who: 37638 artworks are the who of the data.

What: 37638 artworks sold at $9.47 billion. Maximum $119.92 million and minimum $3. The problem is the gap may be too large.

Where: Artist's country of origin

When: Sold time

Why: The goal of the project is to classify if an artwork by 7 famous artists will be sold for more than $20,000, and if an artwork from less known artists will be sold for more than $2,000.

How: Use OpenCV and Python Image Library (PIL) to quantify aesthetics and extract features like dominant color, mean brightness, face count, etc.

The original dataset can be found here.

1b. What are the variable types?

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

- `artist` (categorical)
- `country` (categorical)
- `yearOfBirth` (quantitative when, a particular year)
- `name` (categorical)
- `year` (quantitative when, a particular year)
- `ageOfPaiting` (quantitative when, a particular year)
- `price` (quantitative, $)
- `material` (categorical)
- `height` (quantitative, inches/centimeters)
- `dominantColor` (categorical)

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.

```
chinese_art <- read.csv("artdata.cleaned.csv") %>%
  filter(country == "Chinese")

head(chinese_art)
```

```
    artist country yearOfBirth yearOfDeath                          name year
1    A Lao Chinese        1920          NA                  Dance People 1990
2    A Lao Chinese        1920          NA                  Dance People 1981
3    A Lao Chinese        1920          NA                  Dance People 1985
4    A Lao Chinese        1920          NA                  Dance People 1977
5  A'erbai Chinese          NA          NA                          Lion   NA
6 Ba Weizu Chinese        1744        1793 Calligraphy in Clerical Script 1793
  ageOfPainting price                                  material height width
1            35  1499                      ink_and_color_on_paper  26.38 18.90
2            44  4686                      Ink_and_Color_on_Paper  25.98 16.93
3            40  4070                      ink_and_color_on_paper  26.38 17.72
4            48  2184                      ink_and_color_on_paper  17.72 15.75
5            NA  5814                      ink_and_color_on_paper  23.62 15.75
6           232  6410 A_set_of_four_hanging_scrolls,_ink_on_paper  52.17 10.24
                                        link
1 http://artsalesindex.artinfo.com/asi/lots/5025640
2 http://artsalesindex.artinfo.com/asi/lots/5026919
3 http://artsalesindex.artinfo.com/asi/lots/4935428
```

```
4 http://artsalesindex.artinfo.com/asi/lots/4745570
5 http://artsalesindex.artinfo.com/asi/lots/4754593
6 http://artsalesindex.artinfo.com/asi/lots/3630030
                                                                  source
1            http://artinfo-images-350.s3.amazonaws.com/asi2-110264/395.jpg
2            http://artinfo-images-350.s3.amazonaws.com/asi2-110288/113.jpg
3            http://artinfo-images-350.s3.amazonaws.com/asi2-109321/173.jpg
4            http://artinfo-images-350.s3.amazonaws.com/asi2-106927/686.jpg
5           http://artinfo-images-350.s3.amazonaws.com/asi2-106932/1693.jpg
6 http://artinfo-images-350.s3.amazonaws.com/missingImages/0886555/1324.jpg
  dominantColor brightness ratioUniqueColors thresholdBlackPerc
1        whites        222              0.16               5.68
2        whites        219              0.12               6.21
3        whites        221              0.17               7.99
4        whites        209              0.18               9.64
5       yellows        170              0.05               9.59
6         grays        171              0.13              18.57
  highbrightnessPerc lowbrightnessPerc CornerPer EdgePer FaceCount soldtime
1                  0              4.29      1.18    6.44         0     <NA>
2                  0              4.93      1.30    7.96         0     <NA>
3                  0              5.91      0.88    6.96         0     <NA>
4                  0              8.16      1.27    7.92         0     <NA>
5                  0              3.17      3.30    6.91         0     <NA>
6                  0             13.05      8.81   16.31         0     <NA>
```

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)

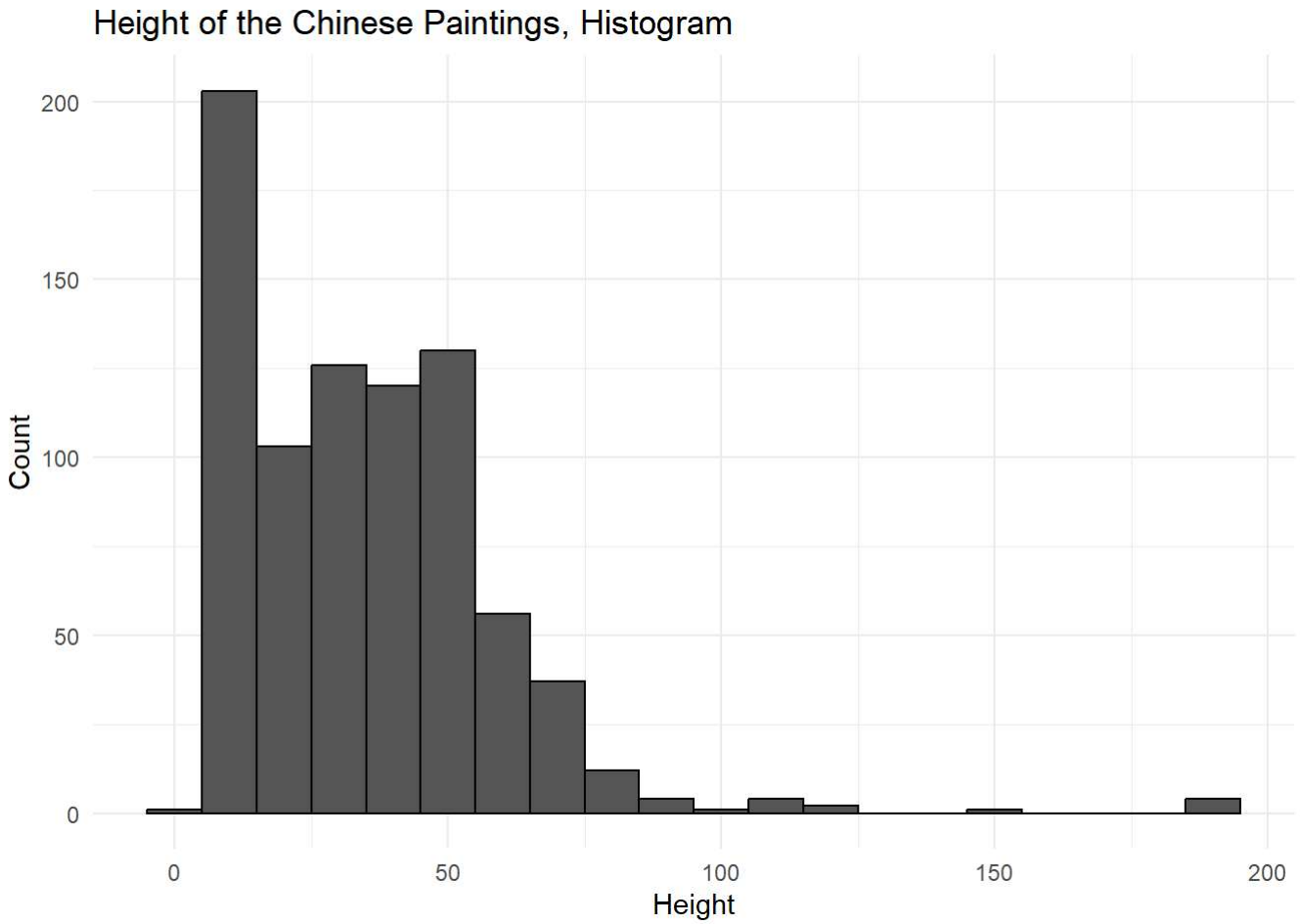
> Think: I want to investigate the distribution of the height of the Chinese paintings. The data are some of the heights of the Chinese paintings in history. The the heights of the Chinese paintings are quantitative.

> Show:

```r
chinese_art <- read.csv("artdata.cleaned.csv") %>%
  filter(country == "Chinese")

ggplot(chinese_art, aes(x = height)) +
  geom_histogram(binwidth = 10,color = "black") +
  labs(
    title = "Height of the Chinese Paintings, Histogram",
    x = "Height",
    y = "Count"
```

```
  ) +
  theme_minimal()
```

```
Warning: Removed 71 rows containing non-finite outside the scale range
(`stat_bin()`).
```



> Tell: The mean an median are not close, so the outlier should be considered. Most height is under 100. The number of big paintings are relative small.

2c. Investigating width

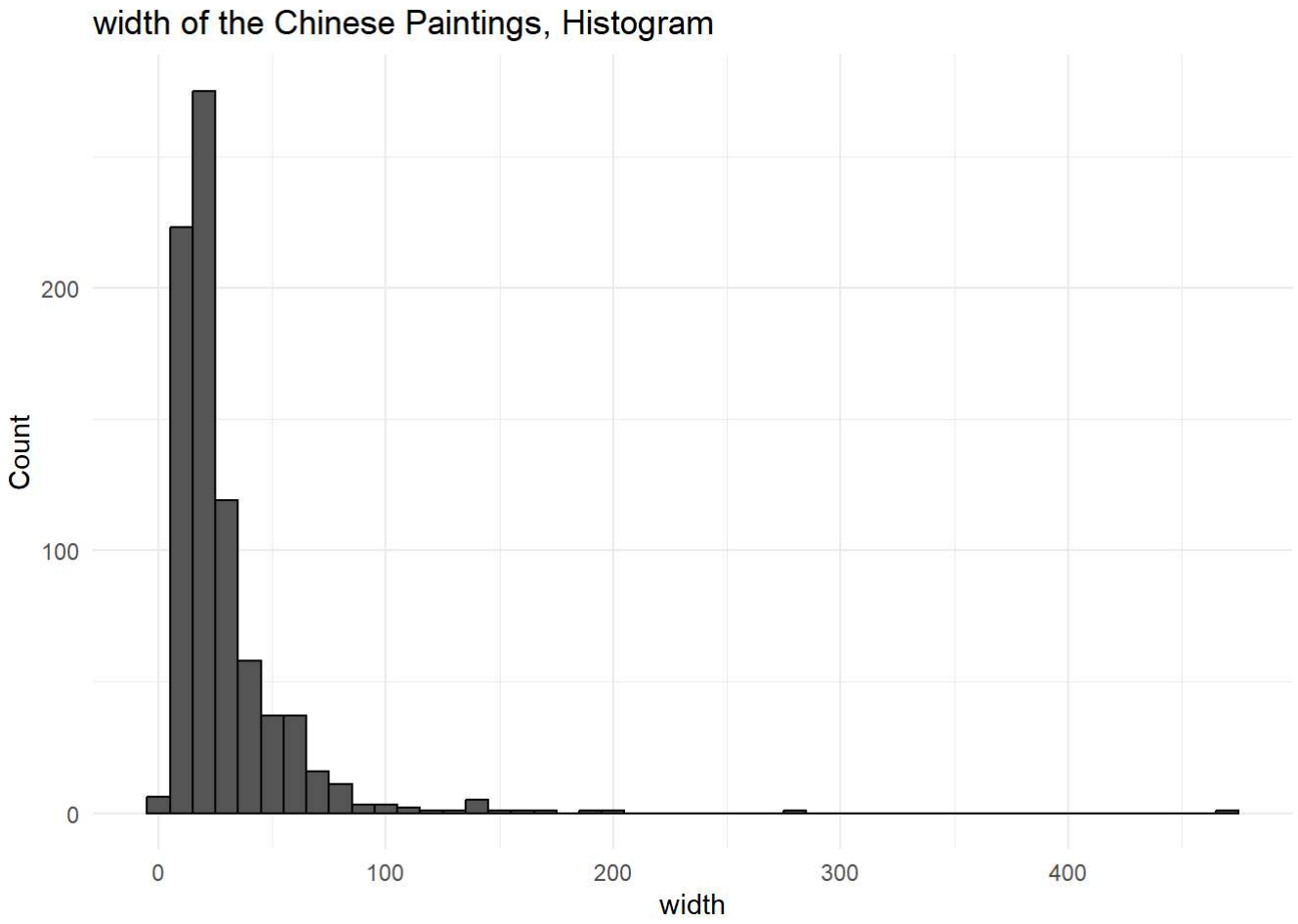
Using the Think-Show-Tell framework from the textbook, investigate the distribution of the width of the Chinese paintings

> Think: I want to investigate the distribution of the width of the Chinese paintings. The data are some of the width of the Chinese paintings in history. The the width of the Chinese paintings are quantitative.

> Show:

```r
chinese_art <- read.csv("artdata.cleaned.csv") %>%
  filter(country == "Chinese")

ggplot(chinese_art, aes(x = width)) +
  geom_histogram(binwidth = 10, color = "black") +
  labs(
    title = "width of the Chinese Paintings, Histogram",
    x = "width",
    y = "Count"
  ) +
  theme_minimal()
```

```
Warning: Removed 71 rows containing non-finite outside the scale range
(`stat_bin()`).
```



> Tell: The mean an median are not close, so the outlier should be considered. Most Width is under 100. The number of big paintings are relative small.

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? Both height and width are mainly under 100cm. From the graph the

median of height is larger than width. The maximum of width is larger than height. Maybe a big painting, whose length and width are larger than 100, its width is more likely longer than height.

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)

chinese_art <- read.csv("artdata.cleaned.csv") %>%
  filter(country == "Chinese")

chinese_art <- chinese_art %>%
  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("america", country, ignore.case = TRUE) ~ "American",
      grepl("china", country, ignore.case = TRUE) ~ "Chinese",
      TRUE ~ "Other"
    )
  )
```

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

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

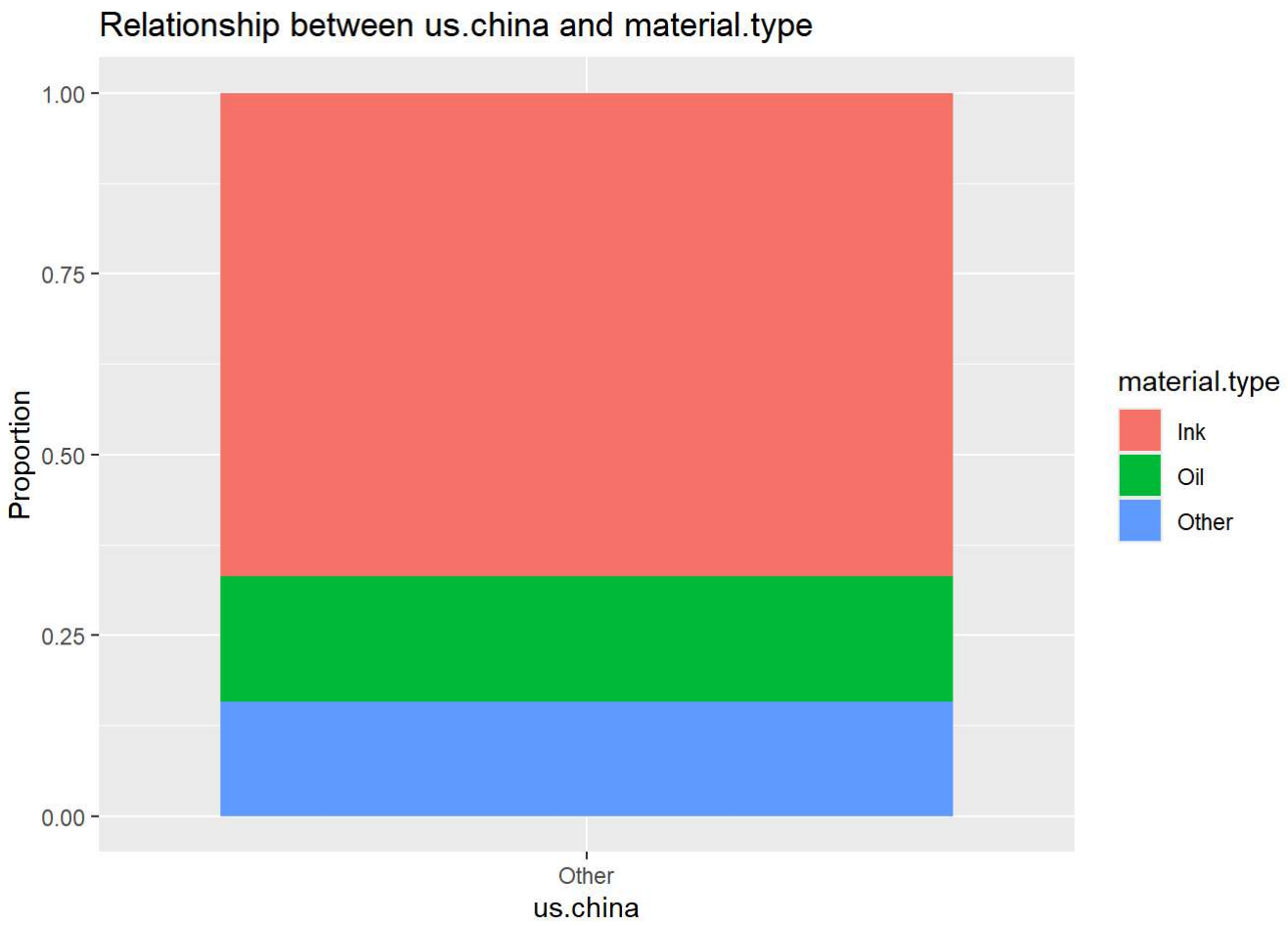
Hint 3: you can see an example of some ways to display this information here

> Think: I want to investigate the I want to investigate the distribution of the width of the Chinese

> paintings.

> Show:

```
ggplot(chinese_art, aes(x = us.china, fill = material.type)) +
  geom_bar(position = "fill") +
  labs(
    title = "Relationship between us.china and material.type",
    x = "us.china",
    y = "Proportion"
  )
```



> Tell: ink is more popular in paintings from the US and China. ink nearly 70%, more than a half. the proportion of oil and other are similar. oil is the second popular material.

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. Oil painting techniques began to spread to China and the United States around the 18th and 19th centuries. Traditional Chinese ink painting (water-ink painting) has a long history. By the 19th century, advancements in the production of oil

painting materials, such as oil paints and canvases, made oil painting more accessible to artists. ## 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. library(dplyr)

```
chinese_art <- read.csv("artdata.cleaned.csv")


chinese_art <- chinese_art %>%
  mutate(
    famous.countries = case_when(
      grepl("america", 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_character_
    ),
    area = height * width
  )
```

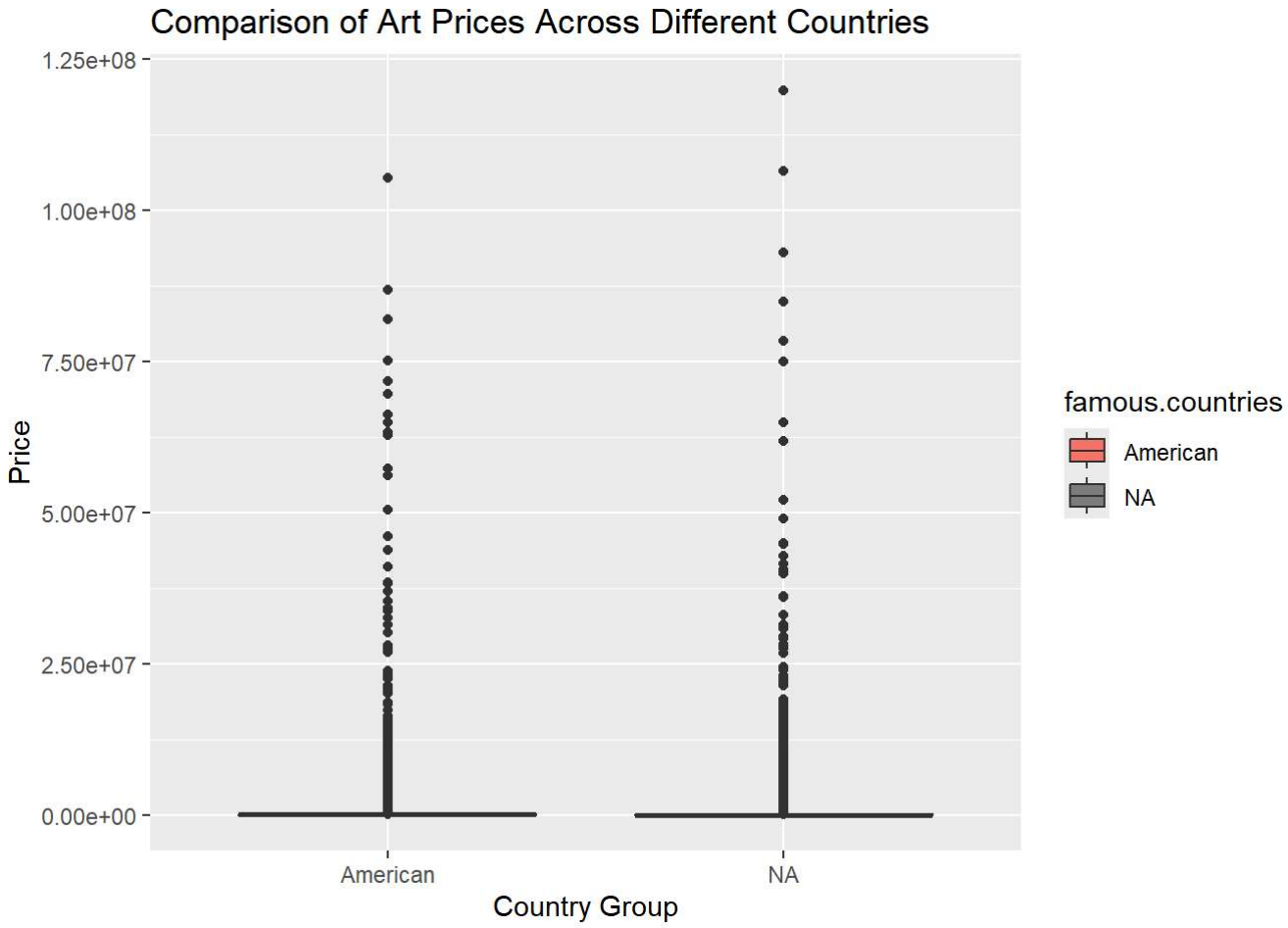
4b. Compare the groups of countries on the variable `price`

> Think I want to investigate the groups of countries on the variable `price` The data are price of paintings The the price of paintings paintings are quantitative.

> Show:

```
library(ggplot2)
library(dplyr)

chinese_art <- read.csv("artdata.cleaned.csv")


chinese_art <- chinese_art %>%
  mutate(
    famous.countries = case_when(
      grepl("america", 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_character_
```

```
    ),
    area = height * width
  )


chinese_art$price <- as.numeric(as.character(chinese_art$price))

ggplot(chinese_art, aes(x = famous.countries, y = price, fill = famous.countries)) +
  geom_boxplot() +
  labs(
    title = "Comparison of Art Prices Across Different Countries",
    x = "Country Group",
    y = "Price"
  )
```



> Tell: In American price of paintings are mainly under 3.75e+07. In NA price of paintings are mainly under 4.2e+07. Hishest price is not from American

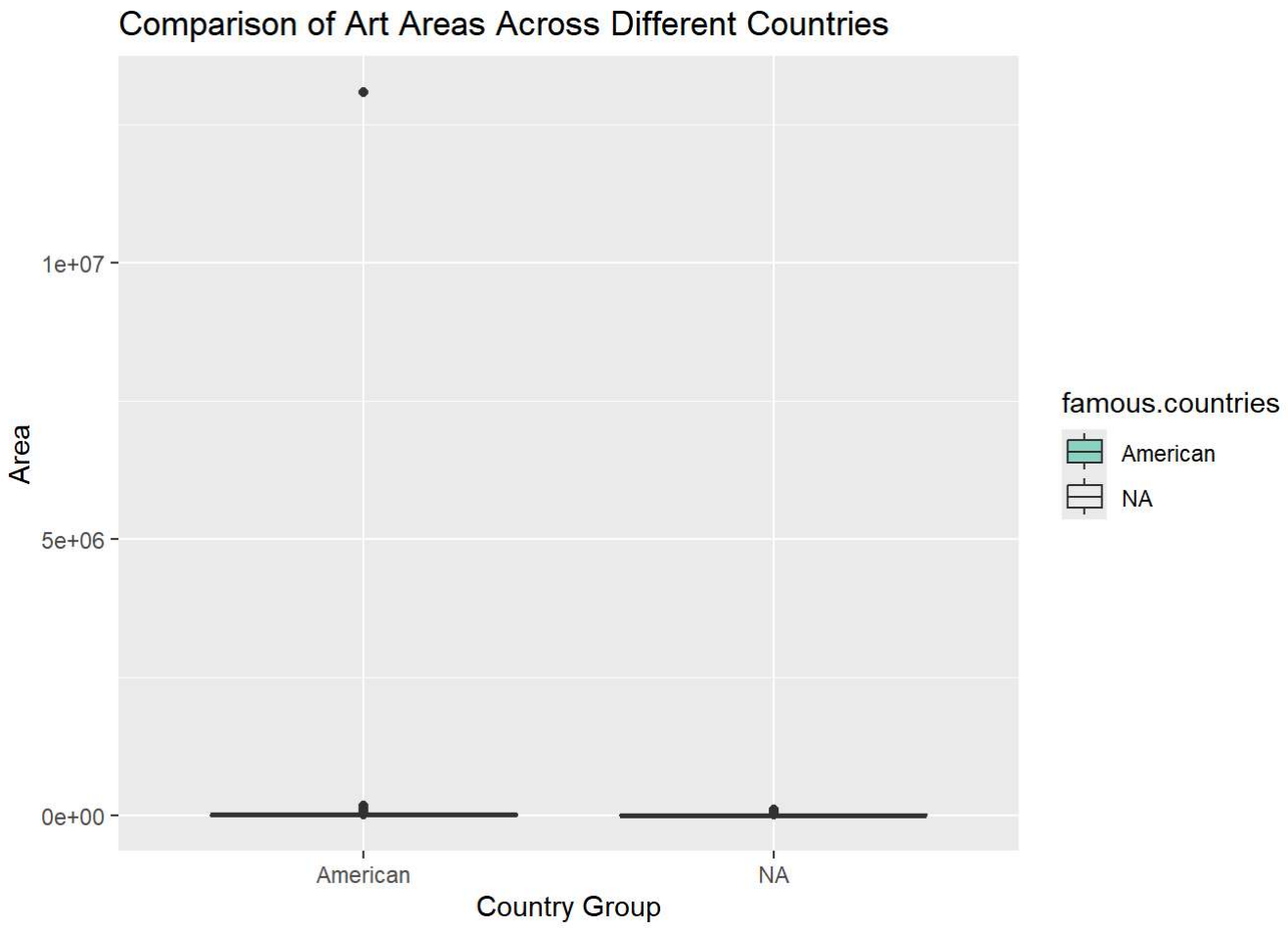
4c. Compare the groups of countries on the variable area

> Think : I want to investigate the groups of countries on the variable area The data are area of paintings The the area of paintings paintings are quantitative.

> Show:

```
library(ggplot2)
library(dplyr)

chinese_art <- read.csv("artdata.cleaned.csv")

chinese_art <- chinese_art %>%
  mutate(
    famous.countries = case_when(
      grepl("america", 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_character_
    ),
    area = height * width
  )

ggplot(chinese_art, aes(x = famous.countries, y = area, fill = famous.countries)) +
  geom_boxplot() +
  labs(
    title = "Comparison of Art Areas Across Different Countries",
    x = "Country Group",
    y = "Area"
  ) +

  scale_fill_brewer(palette = "Set3")
```

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



> Tell: The distribution of area is extreme. From the grahp, the paintings have really big area in American.

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? Paintings' price and area are different in different countries, but the difference is not big. The price of paintings are mainly under 4.2e+07. I think the similarities are beacuse art can be appreciate by everyone.

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. choose 1800 1850 1900

5b. Finding the average

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

```
library(dplyr)

selected_years <- c( 1800, 1850, 1900)

art_data <- read.csv("artdata.cleaned.csv")

average_brightness <- art_data %>%
  filter(year %in% selected_years) %>%
  group_by(year) %>%
  summarise(average_brightness = mean(brightness, na.rm = TRUE))

# Print the result
print(average_brightness)
```

```
# A tibble: 3 × 2
   year average_brightness
  <int>              <dbl>
1  1800               147.
2  1850               127.
3  1900               139.
```

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 :I want to find how many \(z\) units each of the averages for the years are away from the overall mean of brightness. use a table.

> Show:

```
library(dplyr)


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

selected_years <- art_data %>%
  filter(year %in% c( 1800, 1850, 1950)) %>%
  group_by(year) %>%
  summarise(average_brightness = mean(brightness, na.rm = TRUE))

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

```
# Print the result
print(selected_years)
```

```
# A tibble: 3 × 3
   year average_brightness  z_score
  <int>              <dbl>    <dbl>
1  1800               147.  0.00985
2  1850               127. -0.379
3  1950               152.  0.0931
```

> Tell: the absolute value of z_score is largest in 1850, is smallest in 1800. So in 1800 the observation is not far from the average value of the distribution. In 1850 the observation is further from the average value of the distribution.

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? Implications:artistic preferences for brightness have changed over time. If certain periods show consistently higher or lower brightness levels, it might indicate a preference for lighter or darker themes. Limitations: only choose there period. Other analysis: how brightness affect the price of a painting. ## 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 I want to find how many \(z\) units each of the averages for the years are away from the overall mean of highbrightness. use a table.

> Show:

```
library(dplyr)

overall_mean_highbrightnessPerc <- mean(art_data$highbrightnessPerc, na.rm = TRUE)
overall_sd_highbrightnessPerc <- sd(art_data$highbrightnessPerc, na.rm = TRUE)

selected_years <- art_data %>%
  filter(year %in% c( 1800,1850, 1950)) %>%
  group_by(year) %>%
  summarise(average_highbrightnessPerc = mean(highbrightnessPerc, na.rm = TRUE))

selected_years <- selected_years %>%
  mutate(z_score = (average_highbrightnessPerc - overall_mean_highbrightnessPerc) / overall_sd_hi
```

```
# Print the result
print(selected_years)
```

```
# A tibble: 3 × 3
   year average_highbrightnessPerc z_score
  <int>                      <dbl>   <dbl>
1  1800                       6.36  0.634
2  1850                       3.08  0.0804
3  1950                       3.32  0.121
```

> Tell: the absolute value of z_score is largest in 1800, is smallest in 1850. So in 1850 the observation is not far from the average value of the distribution. In 1800 the observation is further from the average value of the distribution.

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. Paintings have difference around the world. But the size and price do not have larger difference. Art have similarities in the world.

knitr::purl("my_code.R," documentation = 0)

tingtex::install_tingtex() tingtex::pdflatex("my_code.Rnw")
file.copy("my_code.pdf","path_to_destination/my_code.pdf")