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?

The original dataset can be found here.

1. Who: What are the cases? These are the individual purchase records or rows in this dataset. In other words, each record might represent a different transaction by a consumer. If there are anonymous or unclear identifiers for the customers, such as the usage of customer IDs instead of names, then it may not be easy to link data with a particular customer or find patterns in individual behavior.

2. What: What's in the data? It would contain customer purchase records that may include variables such as customer ID, product information including the type and price of the product, and transaction details such as the quantity and time of purchase. This might be because some key features, like location of customer or product category, which can provide detailed insights into the analyses being made concerning customer preference, are missing in the dataset.

3. When For how long was this data collected? Probably continuously over time, since the startup of Amazon in 1995 to the present, though certain data points are possibly captured only at specific periods or years depending on what exactly one is analyzing.

4. Where Where was the dataset collected? Data are collected on the Amazon worldwide marketplace that operates in several countries around the world. The dataset can include all customers from different geographical regions, though with anonymized data, the precise location of each of the customers cannot be determined.

5. Why Why was the data collected? The data were collected to understand customer purchase behavior, track sales, and optimize business operations such as inventory management and marketing strategies.

6. How How was the data collected? Most probably, it was collected automatically on Amazon's online platform every time a customer made a purchase. It may include data from customers' interactions with the website and information from the back-end systems that process payments and track inventory.

Conclusion The "who" and "what" of the data are integral in making sense of the dataset-meaning and application, basically. In this respect, knowing what it holds in store or that it is complete and accurate is important, because if there is any incompleteness, biasedness, or poor contextualization in a dataset, several problems might occur. For example, missing data for a particular customer or product may lead to an incorrect conclusion about the trend of purchase. Similarly, if the dataset is not representative of the broader population, generalizing insights could be problematic. The proper definition of variables and collection of representative data are thus crucial to avoid misleading analysis.

1b. What are the variable types?

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

- `artist`: The variable artist classifies each painting or artwork into a category according to the name of the artist. The values are text and represent different categories or groups of artists.

- `country`: The variable country refers to the country of origin for the artist or painting. It is also categorical since it labels categories of geographic regions.

- `yearOfBirth`: yearOfBirth is a numeric variable that represents the year an artist was born. It measures an amount-the year-and so is quantitative because it represents a continuous scale.

- `name`: name is categorical because it states the name of an individual. While it is in text form, it doesn't have any inherent numerical value or scale for analysis.

- `year`: Year represents the variable for which year the painting or artwork was made. This numeric value variable is to measure and compare the temporal relationship among the artworks.

- `ageOfPaiting`: AgeOfPainting measures time in years that describe how old a painting is. It is a quantitative variable since it takes on numeric values, has units of measurement - years - and one can mathematically calculate means, differences, etc., with it.

- `price`: Price is a quantitative variable because it describes money. It can be measured in particular units, such as dollars, and reflects the amount paid for an artwork.

- `material`: Material is a categorical variable because it identifies the type of material used in the artwork, which may be oil paint, marble, or canvas.

- `height`: Height is the measurement of the piece of art in numerical variable measurement. That is quantitative since height can be measured in specific units, say inches or centimeters.
- `dominantColor`:

Categorical Variables: artist, country, name, material, and dominantColor. Quantitative Variables: yearOfBirth, year, ageOfPainting, price, and height.

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

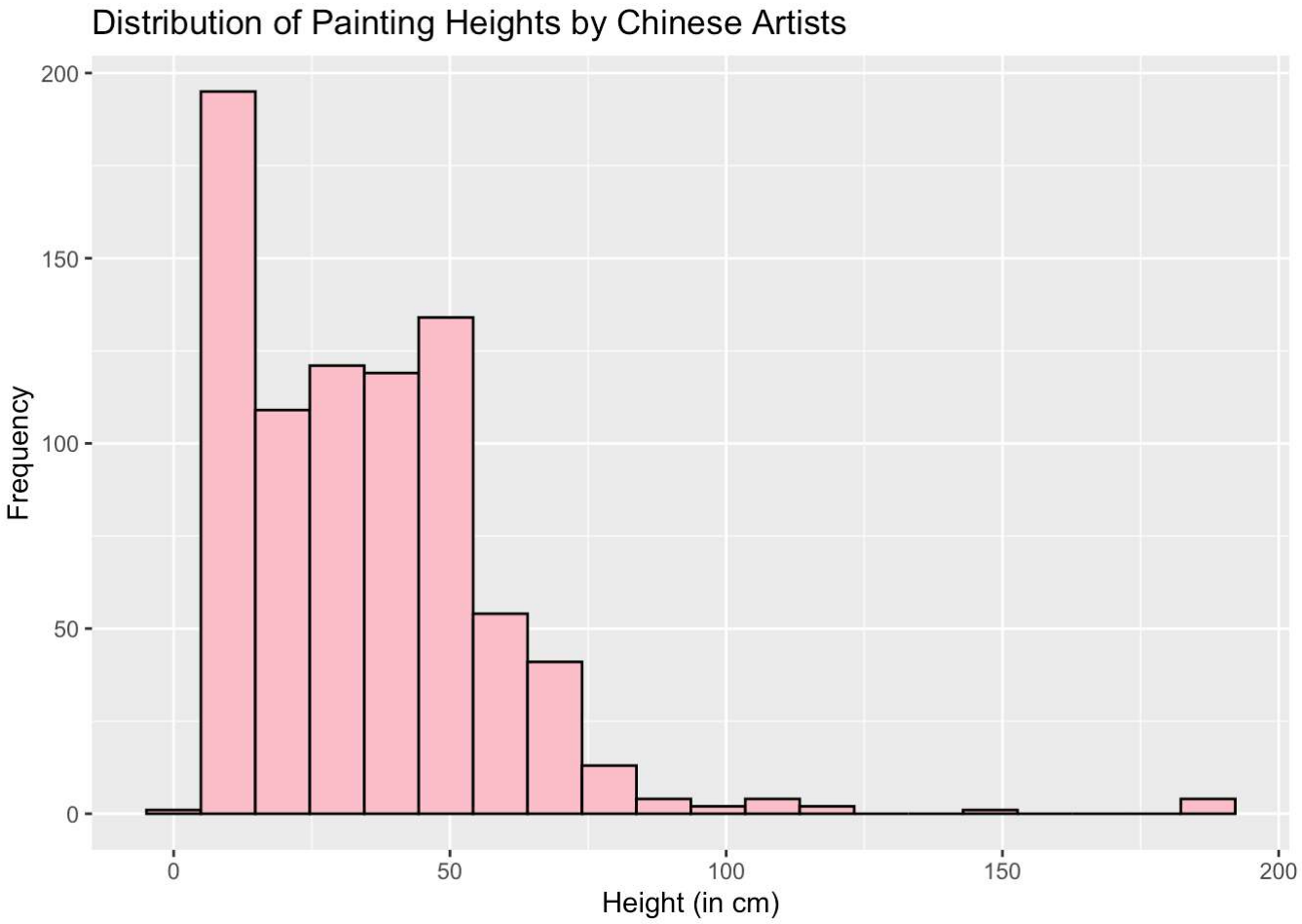
> Show

> Tell

```
library(ggplot2)

ggplot(chinese_artists_data, aes(x = height)) +
  geom_histogram(bins = 20, fill = "pink", color = "black") +
  labs(title = "Distribution of Painting Heights by Chinese Artists", x = "Height (in cm)", y = "
```

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



The x-axis represents the height of the paintings (in centimeters), divided into 20 bins.

The y-axis represents the frequency (count) of paintings that fall into each bin (or range of heights).

The pink bars show how many paintings have a height that falls within each of the 20 intervals. The height of each bar reflects the number of paintings in that range.

The black borders around the bars help make the histogram clearer.

If the histogram is skewed to the right, it means there are more paintings with shorter heights than taller ones.

If the histogram has one peak (unimodal), it suggests that most paintings have heights around a specific value.

If the bars are evenly spread out, it indicates that there is little variation in painting heights

2c. Investigating width

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

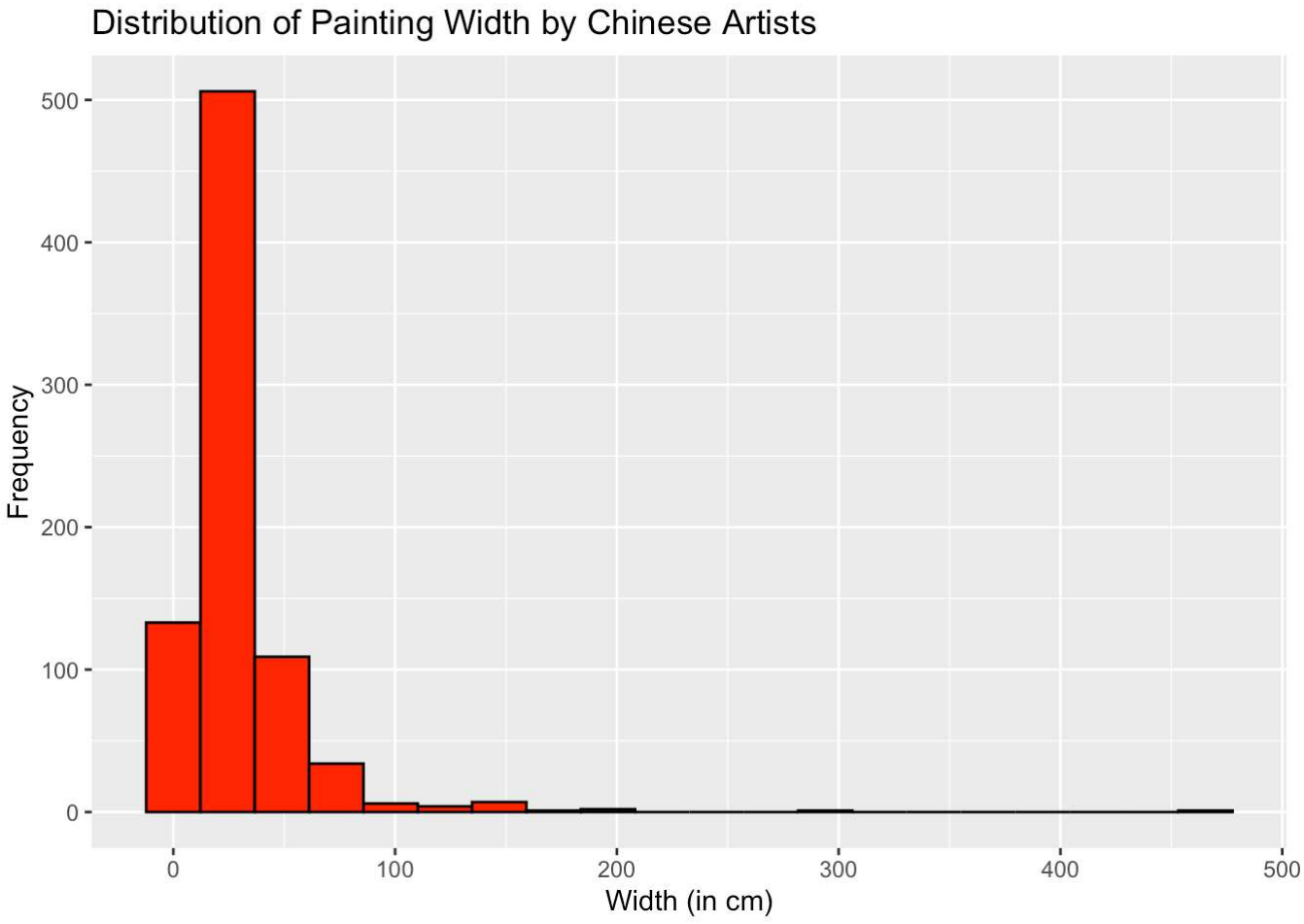
> Think

> Show

> Tell

```
ggplot(chinese_artists_data, aes(x = width)) +
  geom_histogram(bins = 20, fill = "red", color = "black") +
  labs(title = "Distribution of Painting Width by Chinese Artists", x = "Width (in cm)", y = "Fre
```

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



The x-axis represents the width of the paintings (in centimeters), divided into 20 bins.

The y-axis represents the frequency (count) of paintings that fall within each bin (or width range).

The red bars represent the frequency of paintings with particular widths. The height of each bar corresponds to the number of paintings within that range of widths.

The black borders around the bars help make the individual bars more distinguishable.

If the histogram is skewed to the right, it indicates that most paintings have narrower widths, and a few paintings have wider dimensions.

If the histogram shows a single peak (unimodal), this suggests that most paintings have widths around a central value.

If the bars are more evenly distributed across the width range, it means there is less variation in widths across the dataset.

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?

From the results of the first code, the histogram of painting heights, and the second code, the histogram of painting widths, we are able to obtain an even more profound understanding of Chinese art with respect to size preferences, composition styles, and perhaps historical or cultural trends.

A view on the distribution of both height and width allows us to contextualize our analysis in the field of form preferences, historic influences, aesthetic trends that have defined shape in Chinese art-to include, say, how or why vertical scrolls or horizontal landscapes predominate in Chinese art.

Cultural context: How the format of paintings might reflect traditional practices-calligraphy, scrolls-versus modern innovations.

Scale preferences: Whether smaller, intimate works are more common, or larger, more expansive paintings dominate.

Proportion and variety: the amount of standardization in proportions or variety of size that different periods have used by artists.

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()`

```r
artdata_cleaned_mut <- artdata_cleaned %>%
  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("American", country, ignore.case = TRUE) ~ "American",
      grepl("Chinese", 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`

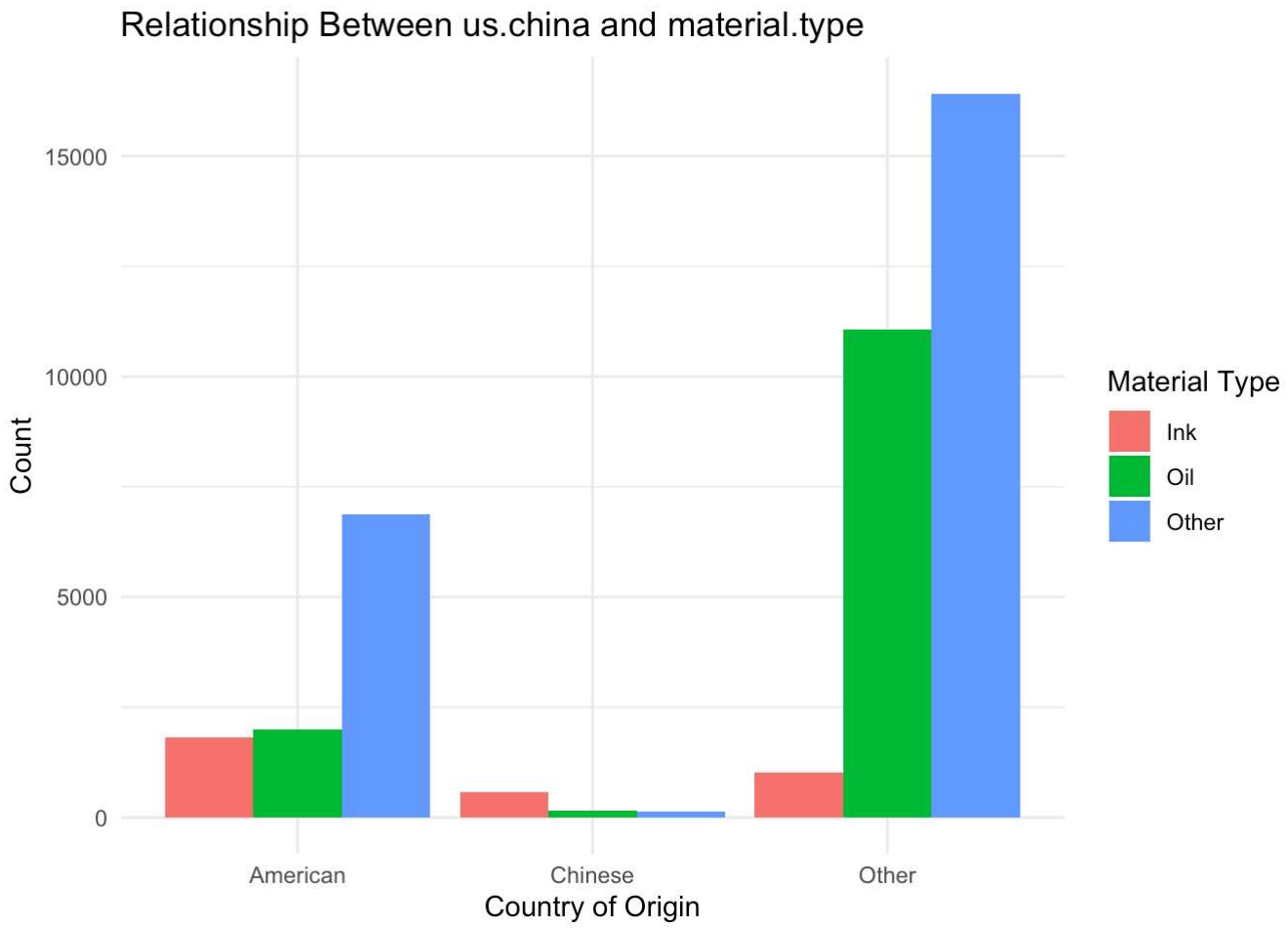
Hint 3: you can see an example of some ways to display this information here

> Think

> Show

> Tell

```r
ggplot(artdata_cleaned_mut, aes(x = us.china, fill = material.type)) +
  geom_bar(position = "dodge") +
  labs(
    title = "Relationship Between us.china and material.type",
    x = "Country of Origin",
    y = "Count",
    fill = "Material Type"
  ) +
  theme_minimal()
```



3c.Thinking about your results

The x-axis represents the country of origin-us. china, which implies whether the art is from the US or China. Aesthetics for the fill are by material type, so the bars are divided for each of the different types of materials used in the various artworks.

Each bar shows the count of artworks from either the US or China, by material type. Bar Groups: There are two groups of bars for each country of origin, namely, US and China.

The height of a bar is count or the number of the artworks for material type and country of origin.

Side-by-Side Comparison: Because the position = “dodge” argument was used, the bars for each material type from the US and China will be positioned side by side to easily compare the count between the two countries.

Conclusion: This bar plot compares the type of materials that US and Chinese artists are using while showing how these two countries differ in practice. The side-by-side bars make a comparison of their preferences for materials transparent, while by color legend can be seen which materials each country used. This sets up a clear visual understanding of what your data look like with country of origin paired with material type.

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.

```
artdata_cleaned_mut2 <- artdata_cleaned %>%
mutate(
    famous.countries = case_when(
      grepl("American", country, ignore.case = TRUE) ~ "American",
      grepl("French", country, ignore.case = TRUE) ~ "French",
      grepl("Italian", country, ignore.case = TRUE) ~ "Italian",
      grepl("Spanish", country, ignore.case = TRUE) ~ "Spanish",
      TRUE ~ "NA"
    ),
    area = height * width
  )
```

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

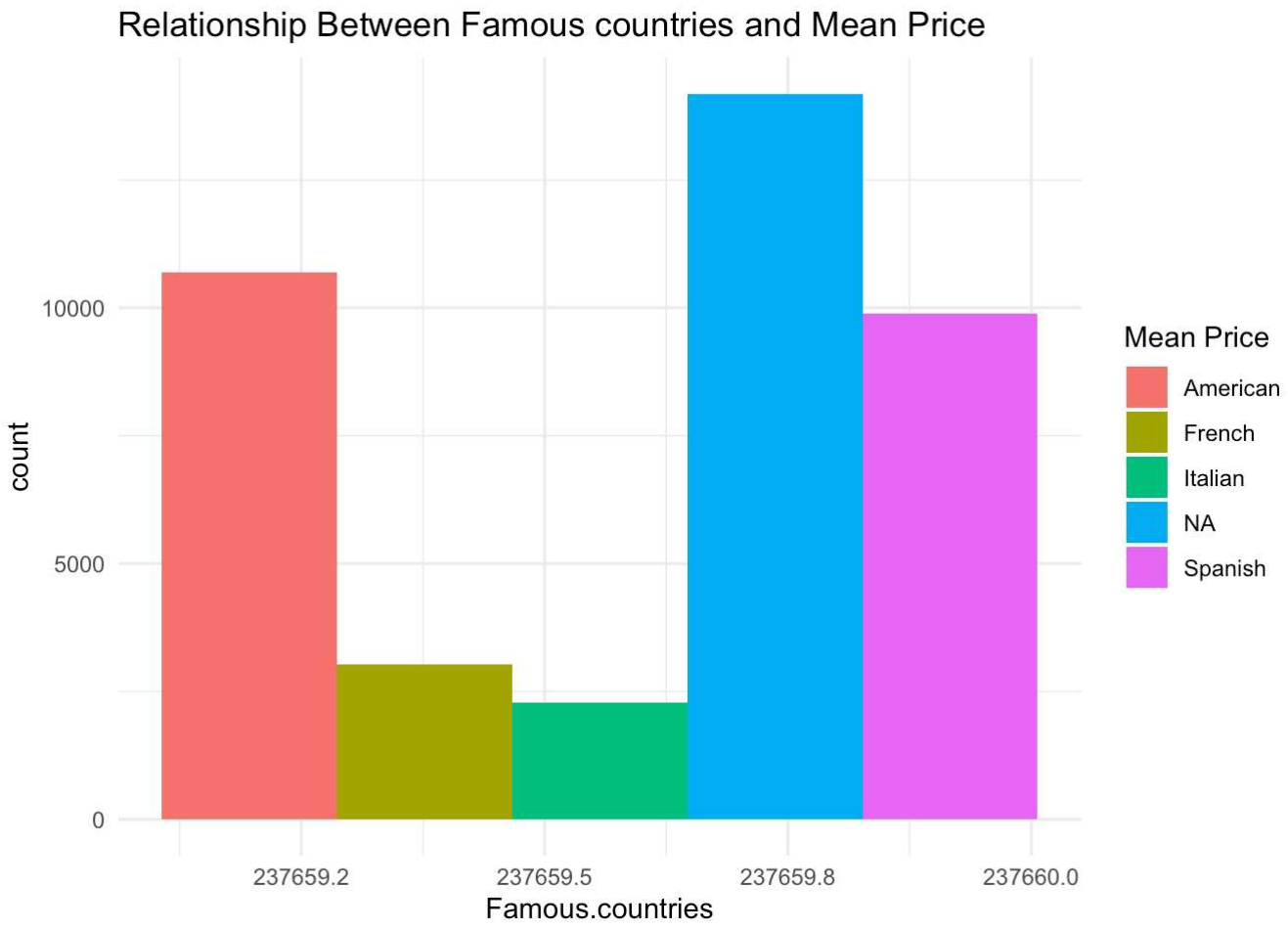
> Think

> Show

> Tell

```
artdata_cleaned_mut2 <- artdata_cleaned_mut2 %>%
  filter(!is.na(price))

ggplot(artdata_cleaned_mut2, aes(x = mean(price), fill = famous.countries)) +
  geom_bar(position = "dodge") +
  labs(
    title = "Relationship Between Famous countries and Mean Price",
    x = "Famous.countries",
    fill = "Mean Price"
  ) +
  theme_minimal()
```



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

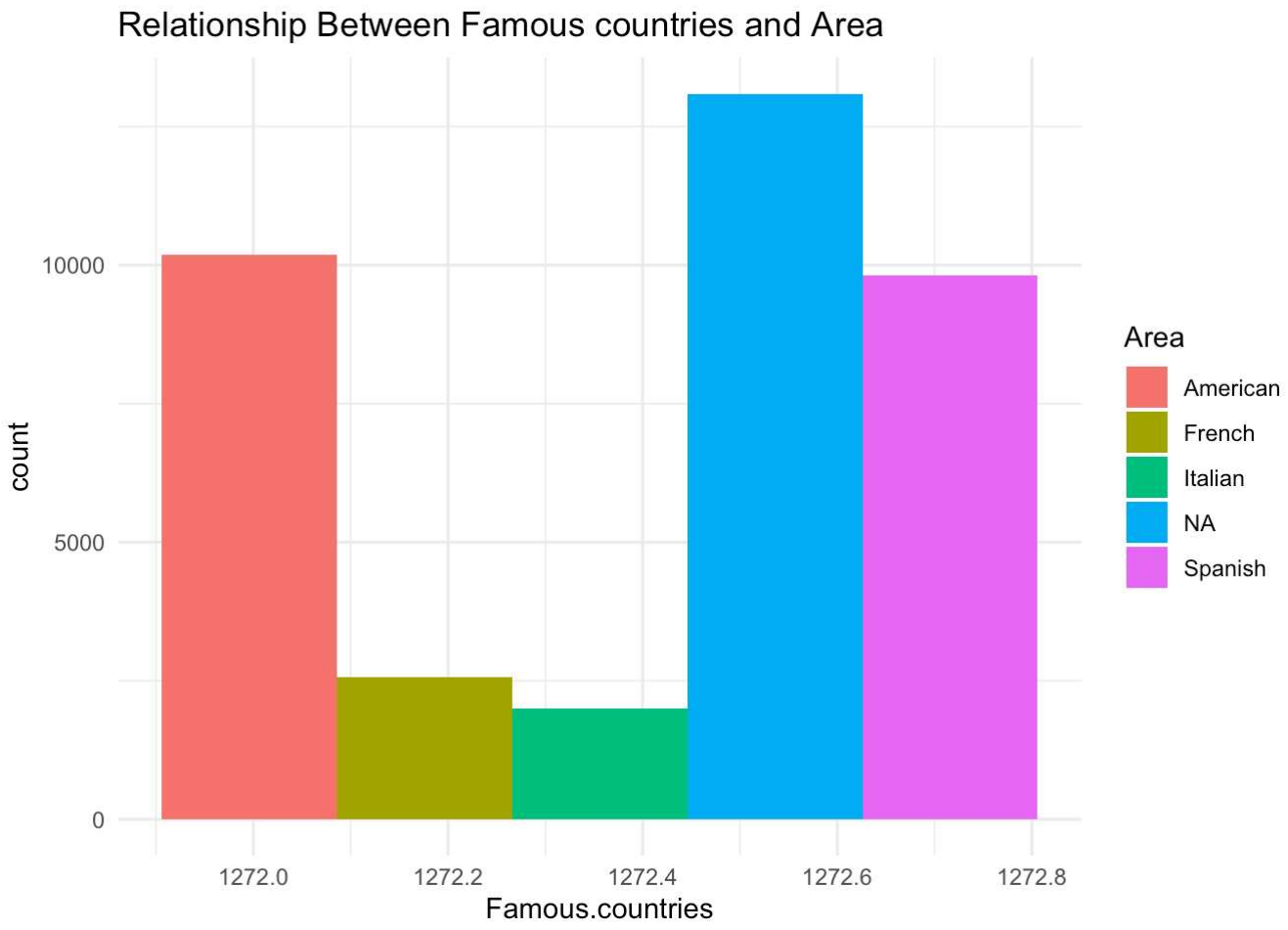
> Think

> Show

> Tell
::: {.cell}

```{.r .cell-code}
artdata_cleaned_mut2 <- artdata_cleaned_mut2 %>%
  filter(!is.na(area))

ggplot(artdata_cleaned_mut2, aes(x = mean(area), fill = famous.countries)) +
  geom_bar(position = "dodge") +
  labs(
    title = "Relationship Between Famous countries and Area",
    x = "Famous.countries",
    fill = "Area"
  ) +
  theme_minimal()
```



:::

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?

This might suggest, if the two countries show similar patterns, that both countries have similarly valued and sized artworks because of a common global art market or similar trends in art.

By considering the two plots together, you will understand the differences and similarities in arts across countries. It is likely that the main causes of these differences are cultural and economic factors such as artistic tradition, market demand, material cost, and reputation of the artist. So you would go deeper to verify your hypotheses regarding the types of art material used, research into the market, and artist profiles.

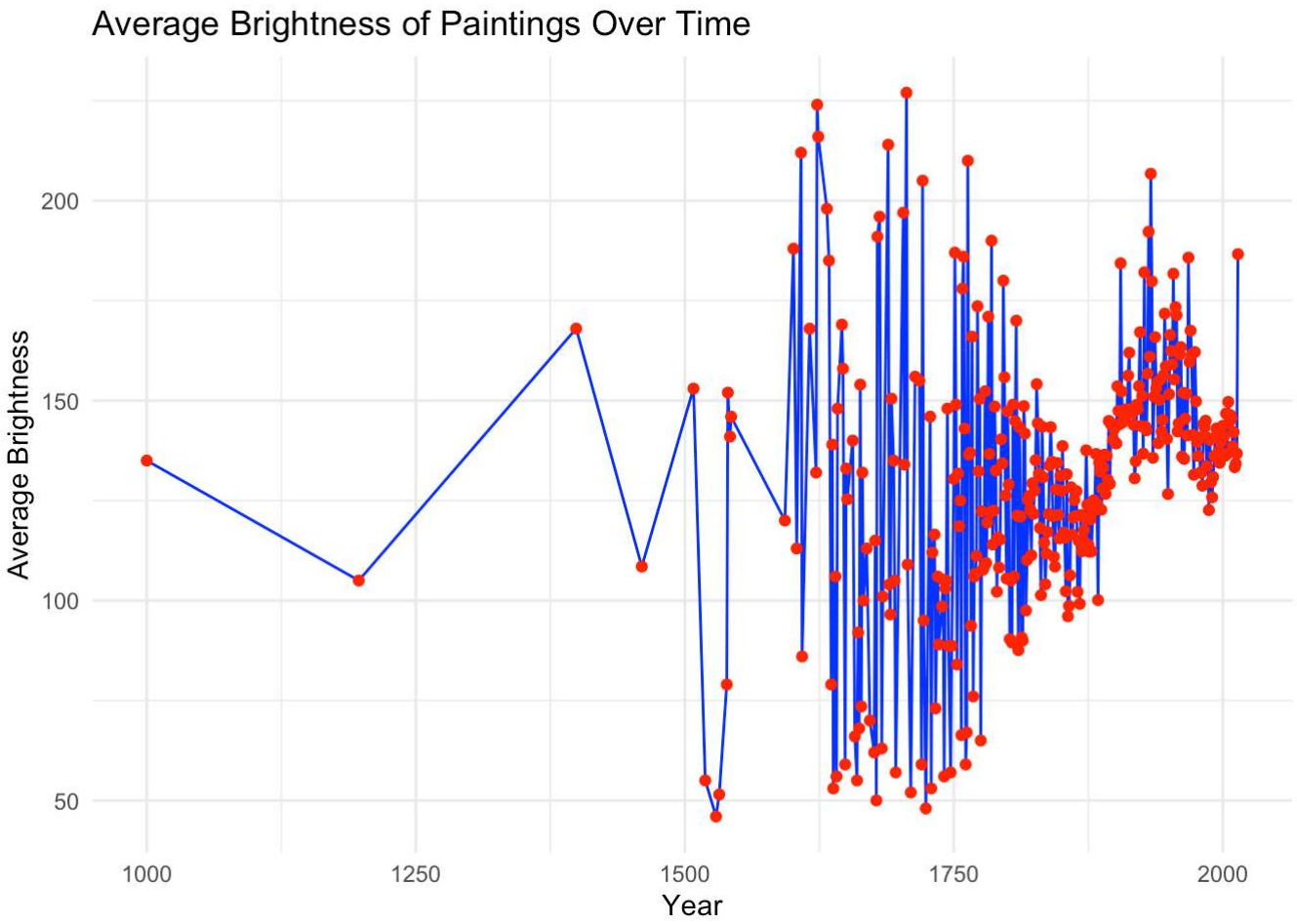
Question 5: Considering deviations (10 points)

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.
::: {.cell}

```{.r .cell-code}
average_brightness_by_year <- aggregate(brightness ~ year, data = artdata_cleaned, FUN = mean)

ggplot(average_brightness_by_year, aes(x = year, y = brightness)) +
  geom_line(group = 1, color = "blue") +
  geom_point(color = "red") +
  ggtitle("Average Brightness of Paintings Over Time") +
  xlab("Year") +
  ylab("Average Brightness") +
  theme_minimal()
```



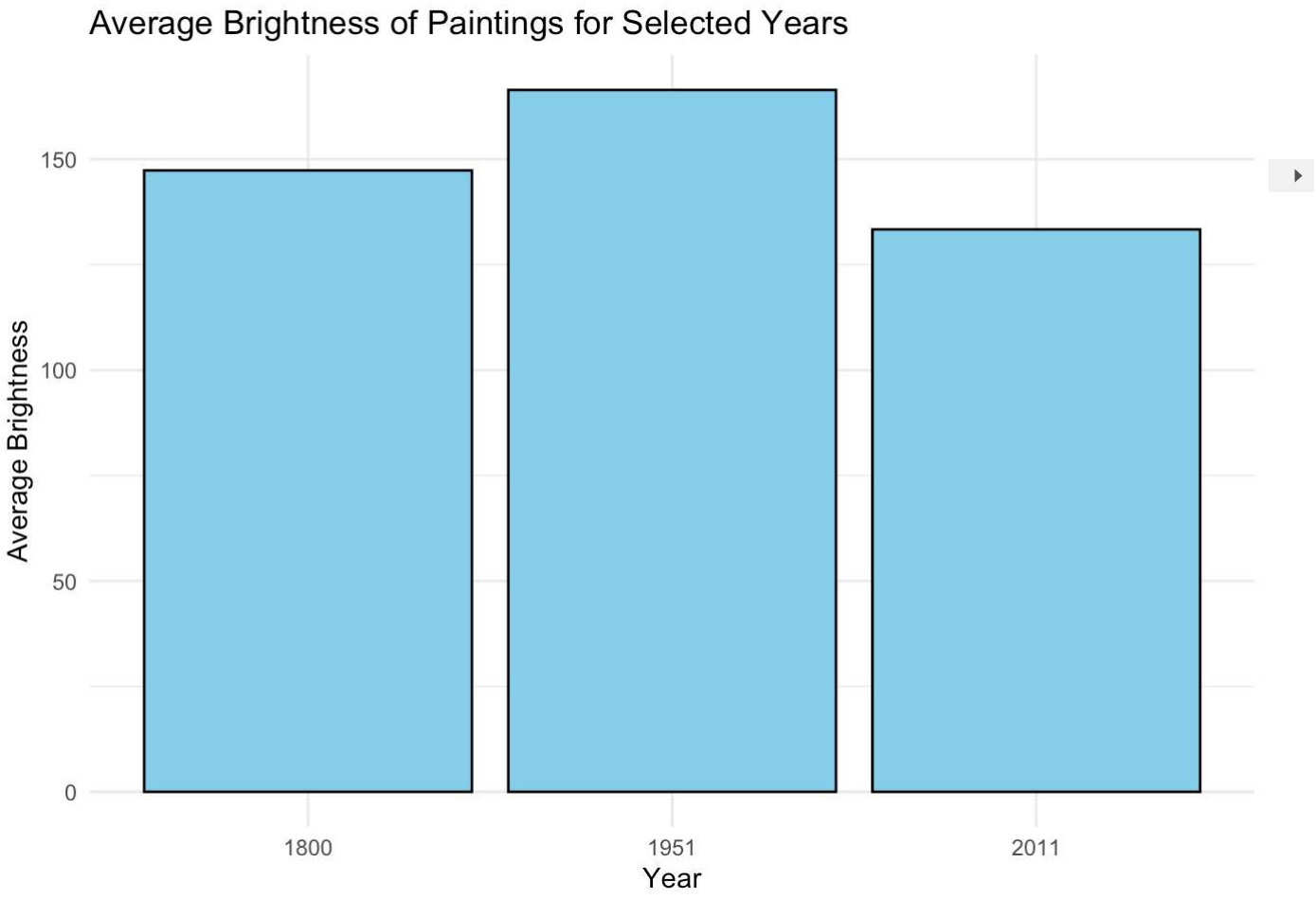
:::

```
selected_years <- c(1800, 1951, 2011)

filtered_data <- artdata_cleaned[artdata_cleaned$year %in% selected_years, ]

average_brightness_selected_years <- aggregate(brightness ~ year, data = filtered_data, FUN = mean
```

```r
ggplot(average_brightness_selected_years, aes(x = factor(year), y = brightness)) +
  geom_bar(stat = "identity", fill = "skyblue", color = "black") +
  ggtitle("Average Brightness of Paintings for Selected Years") +
  xlab("Year") +
  ylab("Average Brightness") +
  theme_minimal()
```




```r
print(average_brightness_selected_years)
```

```
  year brightness
1 1800   147.3333
2 1951   166.4211
3 2011   133.3482
```

5b. Finding the average

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

```r
selected_years <- c(1800, 1951, 2011)
```

```r
filtered_data <- artdata_cleaned[artdata_cleaned$year %in% selected_years, ]

average_brightness_selected_years <- aggregate(brightness ~ year, data = filtered_data, FUN = mean)

print(average_brightness_selected_years)
```

```
  year brightness
1 1800   147.3333
2 1951   166.4211
3 2011   133.3482
```

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

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

z_scores <- sapply(average_brightness_selected_years$brightness,
                   function(avg_brightness) (avg_brightness - overall_mean) / overall_sd)

z_scores_df <- data.frame(
  year = average_brightness_selected_years$year,
  avg_brightness = average_brightness_selected_years$brightness,
  z_score = z_scores
)

print(z_scores_df)
```

```
  year avg_brightness      z_score
1 1800       147.3333  0.009850112
2 1951       166.4211  0.383482954
3 2011       133.3482 -0.263901806
```

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?

The averaged brightness over the years 1800, 1951, and 2011 can serve as the basis of how bright the paintings were through time. Using the z-scores helps to show the relation between each year and the overall mean brightness. It therefore serves as an indication of possible shifts in styles and materials through the centuries and can be connected with the accessibility of brighter paints or other artistic currents. Yet the analysis is somewhat restricted to selecting just three years for sampling, while a broader understanding of art history might consider key periods like Impressionism or Baroque as influencing factors of brightness. Moreover, other relevant variables such as the type of materials and methods could be omitted, which determine the luminosity. Such an analysis could be further developed by expanding the timeframe, incorporating diverse art movements, and including even factors like the material used or the nationality of the artists to get a wider view of how brightness has evolved in art and hence deeper trends in the use of color and light by artists across various ages.

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

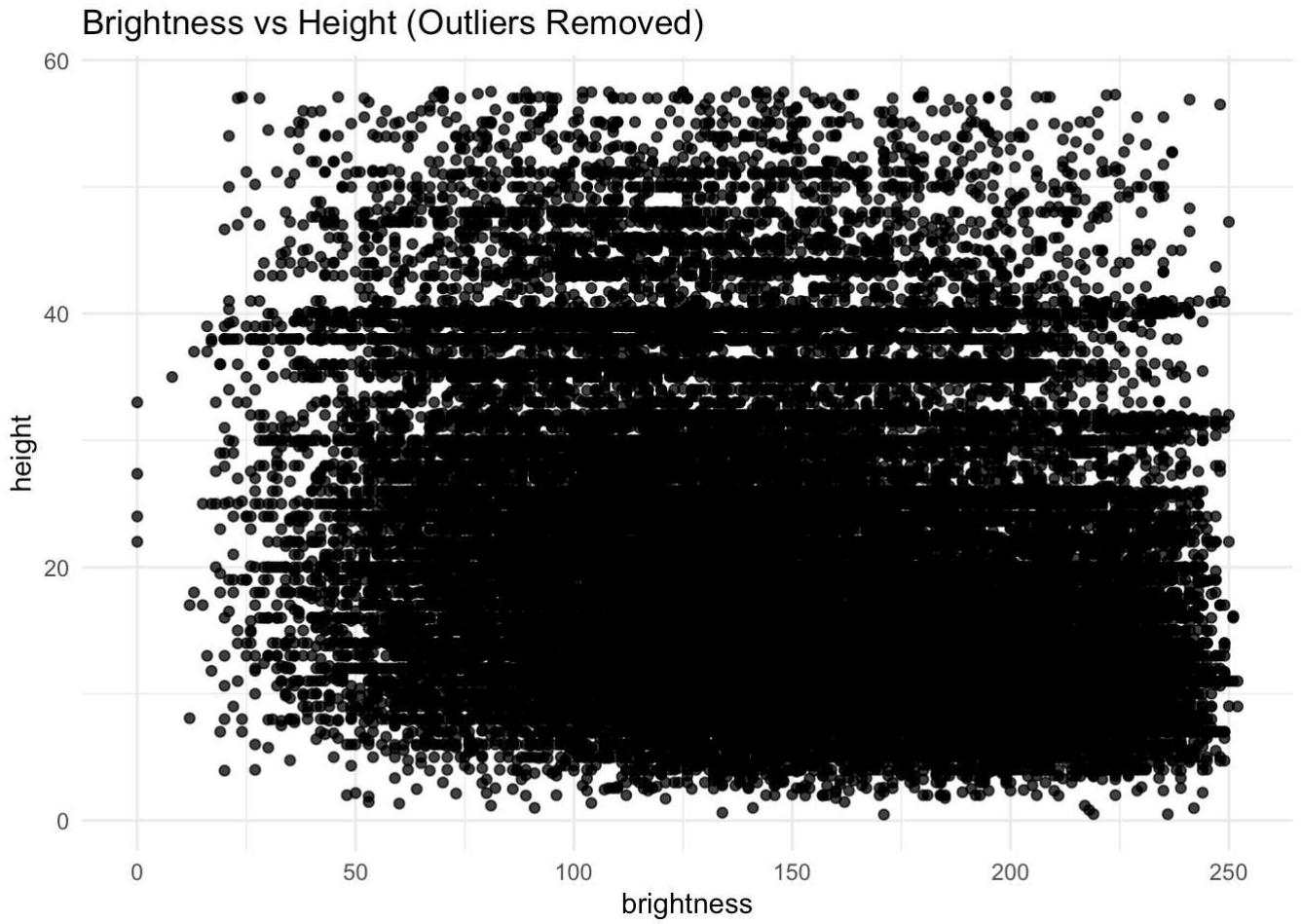
```
Q1 <- quantile(artdata_cleaned$height, 0.25, na.rm = TRUE)
Q3 <- quantile(artdata_cleaned$height, 0.75, na.rm = TRUE)
IQR <- Q3 - Q1

lower_bound <- Q1 - 1.5 * IQR
upper_bound <- Q3 + 1.5 * IQR

artdata_cleaned_no_outliers <- artdata_cleaned[artdata_cleaned$height >= lower_bound & artdata_cl

ggplot(artdata_cleaned_no_outliers, aes(x = brightness, y = height)) +
  geom_point(alpha = 0.8) +
  theme_minimal() +
  ggtitle("Brightness vs Height (Outliers Removed)")
```

```
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.

This investigation emphasizes the importance of data preprocessing, particularly the removal of outliers, to gain a more accurate understanding of the relationship between brightness and height in paintings. By identifying and excluding extreme values using the interquartile range (IQR) method, we were able to observe a clearer, more consistent pattern in the data, suggesting that outliers can distort the interpretation of relationships between variables. The analysis highlights that outliers and data quality play a critical role in shaping findings, and without addressing these issues, conclusions can be misleading. Overall, the investigation underscores the significance of data cleaning in providing more reliable insights, ultimately helping to uncover underlying trends in art-related datasets and making analyses more meaningful and accurate.