Unit 1 Homework

DKU Stats 101 Spring 2025 Session 3

AUTHOR
Sophia Caslin

PUBLISHED
January 17, 2025

Introduction

```
str(artdata_cleaned)
```

```
'data.frame':   40070 obs. of  23 variables:
 $ artist            : chr  "Loredana Raciti" "Loredana Raciti" "W'Angalevia" "Adrian
Paci" ...
 $ country           : chr  "African" "African" "African" "Albanian" ...
 $ yearOfBirth       : int  1959 1959 NA 1969 1971 1971 1971 1971 1971 1971 ...
 $ yearOfDeath       : int  NA NA NA NA NA NA NA NA NA NA ...
 $ name              : chr  "Il marito" "Stop everything, I had enough, I want to go"
"Les braconniers" "Turn On" ...
 $ year              : int  NA NA NA 2004 2014 2006 2007 2012 2006 2012 ...
 $ ageOfPainting     : int  NA NA NA 21 11 19 18 13 19 13 ...
 $ price             : int  205 639 252 16866 274351 50284 332279 4081 24219 5001 ...
 $ material          : chr  "mixed_media_and_collage_on_panel"
"oil_and_collage_on_board" "oil_on_unstretched_canvas" "chromogenic_print" ...
 $ height            : num  37.4 31 19.7 40.5 NA ...
 $ width             : num  29.5 31 32.7 48.4 NA ...
 $ link              : chr  "http://artsalesindex.artinfo.com/asi/lots/5281904"
"http://artsalesindex.artinfo.com/asi/lots/460981" "http://artsalesindex.artinfo.com/
asi/lots/5439052" "http://artsalesindex.artinfo.com/asi/lots/5196580" ...
 $ source            : chr  "http://artinfo-images-350.s3.amazonaws.com/
asi2-114410/97.jpg" "http://artinfo-images-350.s3.amazonaws.com/0611961/182.jpg"
"http://artinfo-images-350.s3.amazonaws.com/asi2-115409/97.jpg" "http://artinfo-
images-350.s3.amazonaws.com/asi2-111628/114.jpg" ...
 $ dominantColor     : chr  "reds" "grays" "grays" "whites" ...
 $ brightness        : int  102 105 125 221 206 241 207 68 180 62 ...
 $ ratioUniqueColors : num  0.1 0.32 0.6 0 0.58 0 0.1 0.23 0 0.22 ...
 $ thresholdBlackPerc: num  90.81 73.21 48.19 0.36 4.83 ...
 $ highbrightnessPerc: num  0.9 0 0 0 0 ...
 $ lowbrightnessPerc : num  0 7.69 7.56 0.03 0.8 ...
 $ CornerPer         : num  0.19 0.88 7.38 0.73 3.3 0.84 0.27 0.25 0.36 0.27 ...
 $ EdgePer           : num  3.11 4.71 23.77 3.98 25.27 ...
 $ FaceCount         : int  0 0 0 0 0 0 0 0 0 0 ...
 $ soldtime          : chr  NA NA NA NA ...
```

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?

Five Ws and How

Who?

The individual cases in this dataset represent artworks sold at auction. Each row in the dataset represents an individual artwork, with a total of 37,638 art pieces. This consists of work by 7 famous artists as well as 7,399 less-known artists.

What?

The dataset records various characteristics about each artwork (e.g., title, material, dimensions, and dominant color), including information about the artist (e.g., name, country of origin, birth year, and death year), and auction (e.g., sold price, sold time).

When?

The dataset captures information about artworks sold over multiple years, but the specific sold times are incomplete, as 50% of this data is missing. This makes it difficult to establish precise timeframes for many entries.

Where?

The data was collected through web scraping from various online auction platforms, likely including major auction houses such as Christie's.

Why?

The purpose of collecting this data is to understand how the commercial art auction market evaluates pieces of art. The goal is to classify whether an artwork will sell for more than $20,000 (for famous artists) or $2,000 (for less-known artists).

How?

The dataset was constructed by web scraping auction data and combining it with image features extracted using tools like OpenCV and Python Image Library (PIL). Various features such as dominant color, brightness, face count, and edge percentages were quantified to provide additional information about the artwork.

Possible Concerns

Who: this dataset may suffer from representation bias. Focusing on 7 famous artists and 7,399 less-known artists may not fully capture the diversity of the global art market. The criteria used to select these artists and artworks are also unclear, potentially introducing sampling bias. Additionally, the dataset lacks information about buyers, auction houses, or other stakeholders, which limits its

contextual depth.

What: Missing sold time data makes temporal analysis unreliable, and missing birth and death year data for artists reduces demographic completeness. Data quality features like brightness or corner percentage, analyzed through algorithms, might not adequately capture subjective qualities of art. The dataset also lacks critical contextual information, such as cultural significance, provenance, or historical importance, which are key factors in valuing art. Finally, combining lesser-known artists into a single category might obscure important nuances in trends within that group.

1b. What are the variable types?

- `artist`: Character (chr)
- `country`: Character (chr)
- `yearOfBirth`: Integer (int)
- `name`: Character (chr)
- `year`: Integer (int)
- `ageOfPaiting`: Integer (int)
- `price`: Integer (int)
- `material`: Character (chr)
- `height`: Integer (int)
- `dominantColor`: Character (chr)

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.

```
filter_art <-artdata_cleaned %>%
  filter(country == "Chinese")
```

2b. Investigating height

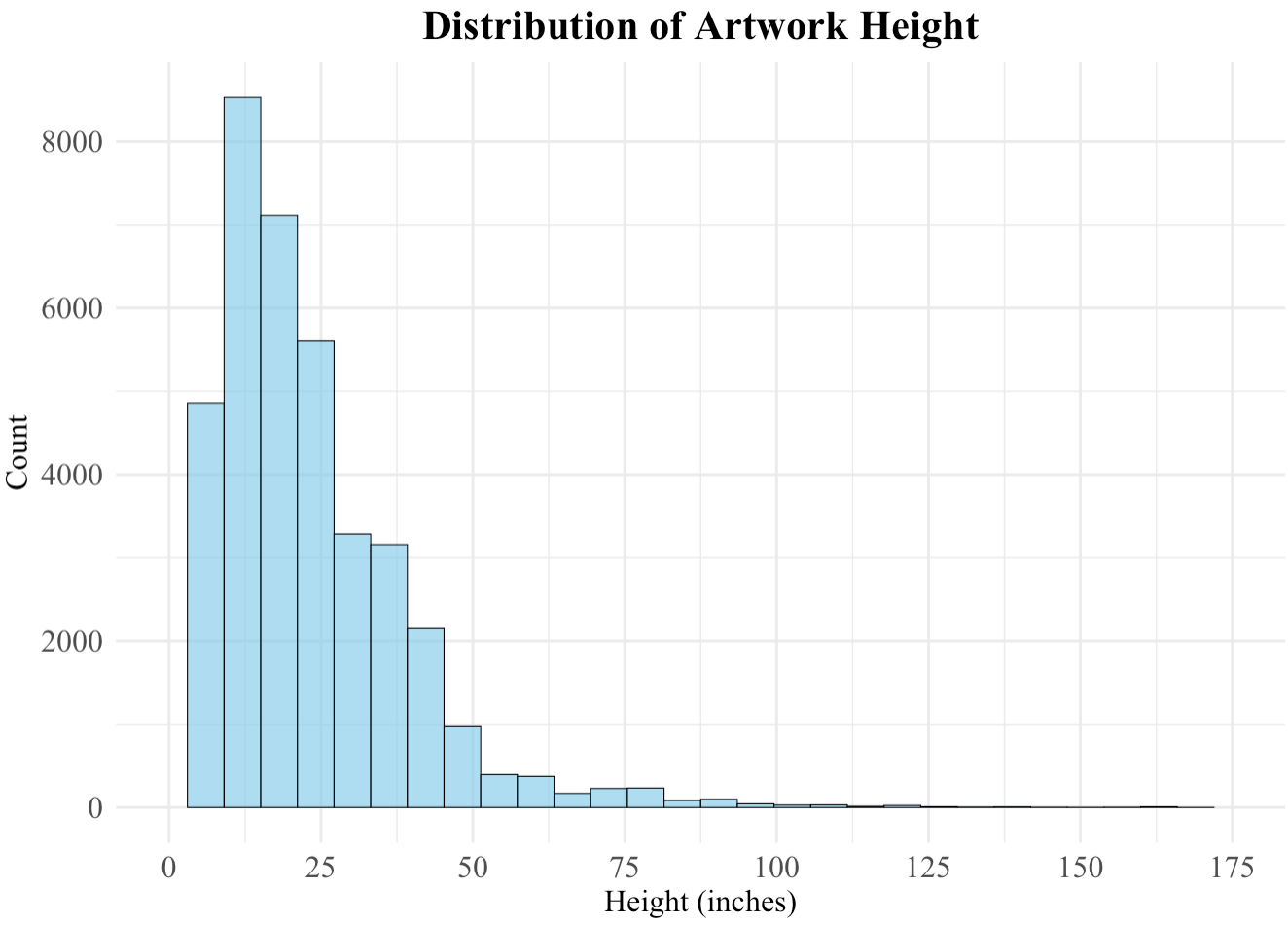
Using the Think-Show-Tell framework from the textbook (example on page 71), investigate the distribution of the height of the Chinese paintings

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 analyze the distribution of the height of Chinese paintings. The data are the heights (inches) of each artwork classified as "Chinese" in this dataset. Height is a

heights (inches) of each artwork classified as "Chinese" in this dataset. Height is a quantitative variable, so histogram is a useful tool to display this distribution.

Show: A histogram of the data shows a fairly symmetric distribution with a few high outliers.



Summary Statistics for Painting Heights

Minimum	Q1	Median	Mean	Q3	Maximum	SD
0.47	12.01	19.88	24.09	30.24	4055.12	28.98

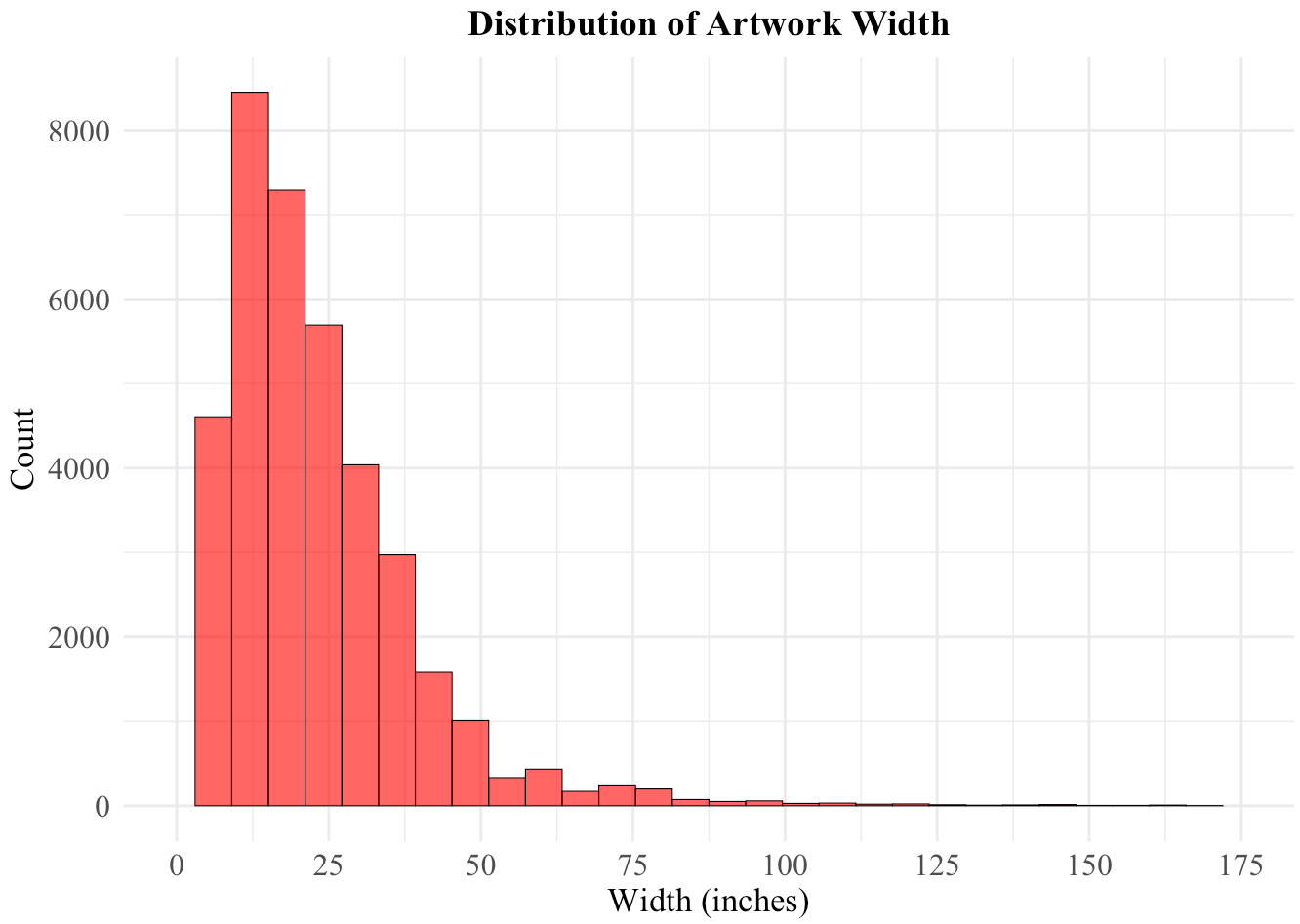
Tell: The distribution of heights is unimodal but exhibits a slight right skew, with the majority of paintings being smaller in size. However, a few large outliers, such as the extreme maximum value of 4055.12, cause the mean to be pulled upwards. The median (19.88) is lower than the mean (24.09), which reinforces the idea of right skewness, as the mean is influenced by the large outlier. The range of the data (0.47 to 4055.12) is quite extensive, and the standard deviation indicates significant variation from the mean. The unusually high maximum value (4055.12) warrants further investigation, as it could either represent an exceptionally large painting or a possible data entry error.

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 analyze the distribution of the width of Chinese paintings. The data includes the width (in) of each artwork classified as "Chinese" in this dataset. Width is similarly a quantitative variable, so histogram is a useful tool to display this distribution.

> **Show:** A histogram of the data shows a similar symmetric distribution with a few high outliers.



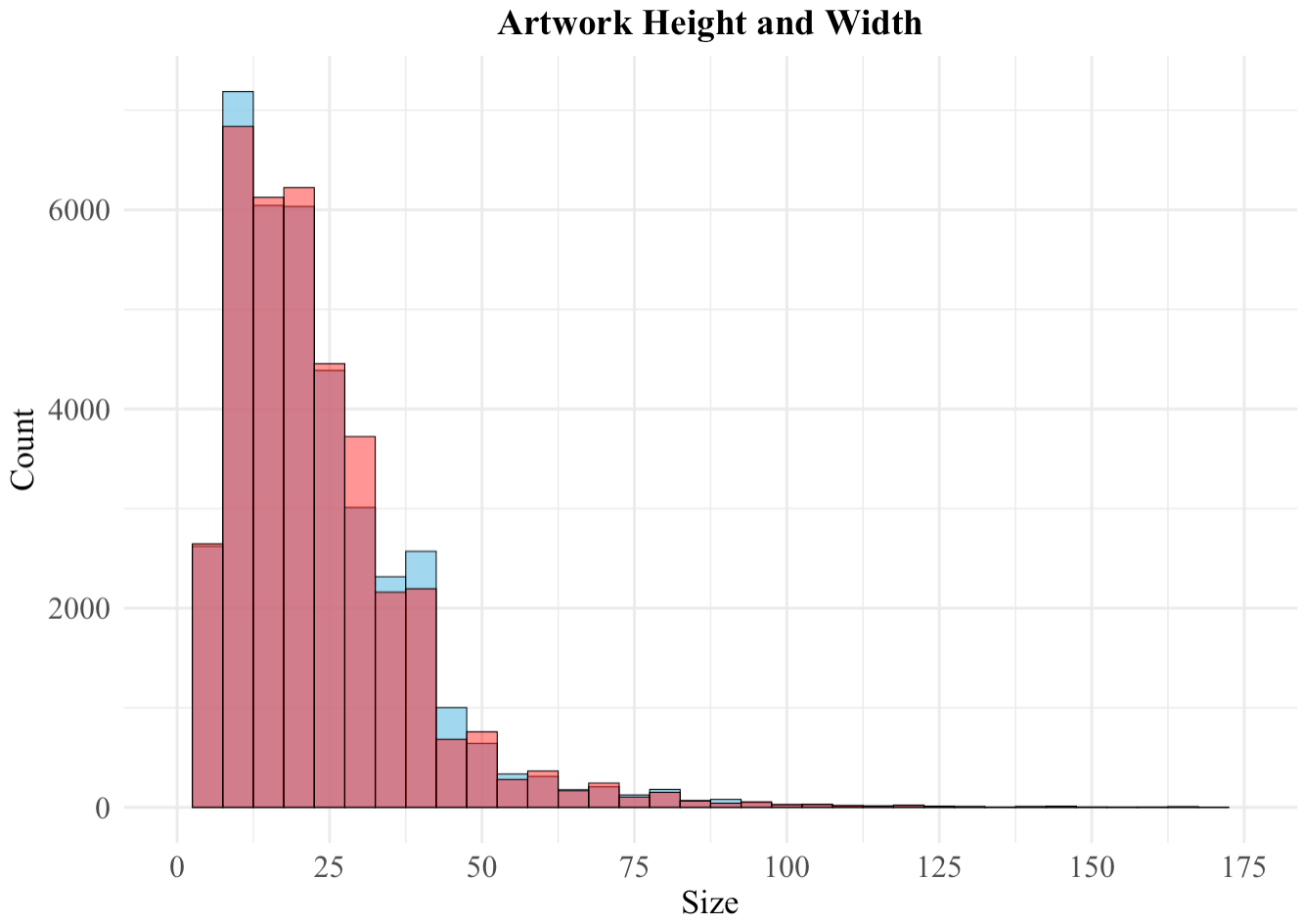
Summary Statistics for Painting Heights

Minimum	Q1	Median	Mean	Q3	Maximum	SD
0.08	12.36	20	24.05	30	3228.35	26.82

> **Tell:** The width distribution of Chinese paintings also appears to be right-skewed, with most paintings being smaller in size and a few large outliers that pull the mean upwards. The mean (24.05) exceeds the median (20.00), and the data spans a very wide range, from 0.08 to 3228.35. A few high outliers including the large maximum value could either represent exceptionally wide paintings or possibly a data entry error that warrants further investigation. Additionally, the dataset contains 2419 missing values, which could lead to inaccurate

> conclusions about the typical size of paintings. For example, the missing data may represent a subset of artworks that are either unusually large or small, affecting calculations of averages or medians.

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 overlay of the two histograms reveals a strong overlap, suggesting a close relationship between the height and width of Chinese paintings. This implies that Chinese paintings likely follow consistent proportions, with less variation in size. The similarity in distributions could also indicate a cultural preference for symmetry and balance in traditional Chinese art, where artists may have adhered to standard dimensions or materials, reflecting historical norms.

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

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()` here and here .

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

```
artdata_cleaned <- 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"
    )
  )
```

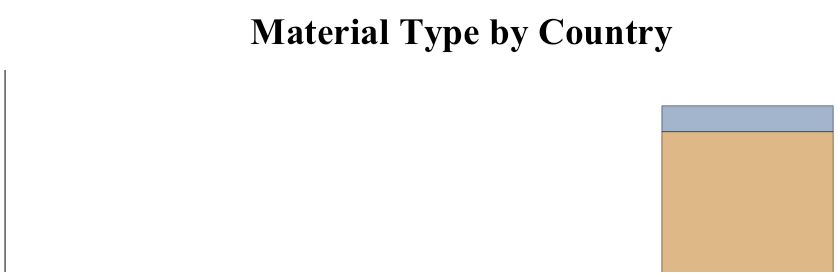
3b. Investigating the categorical relationship between `us.china` and `material.type`

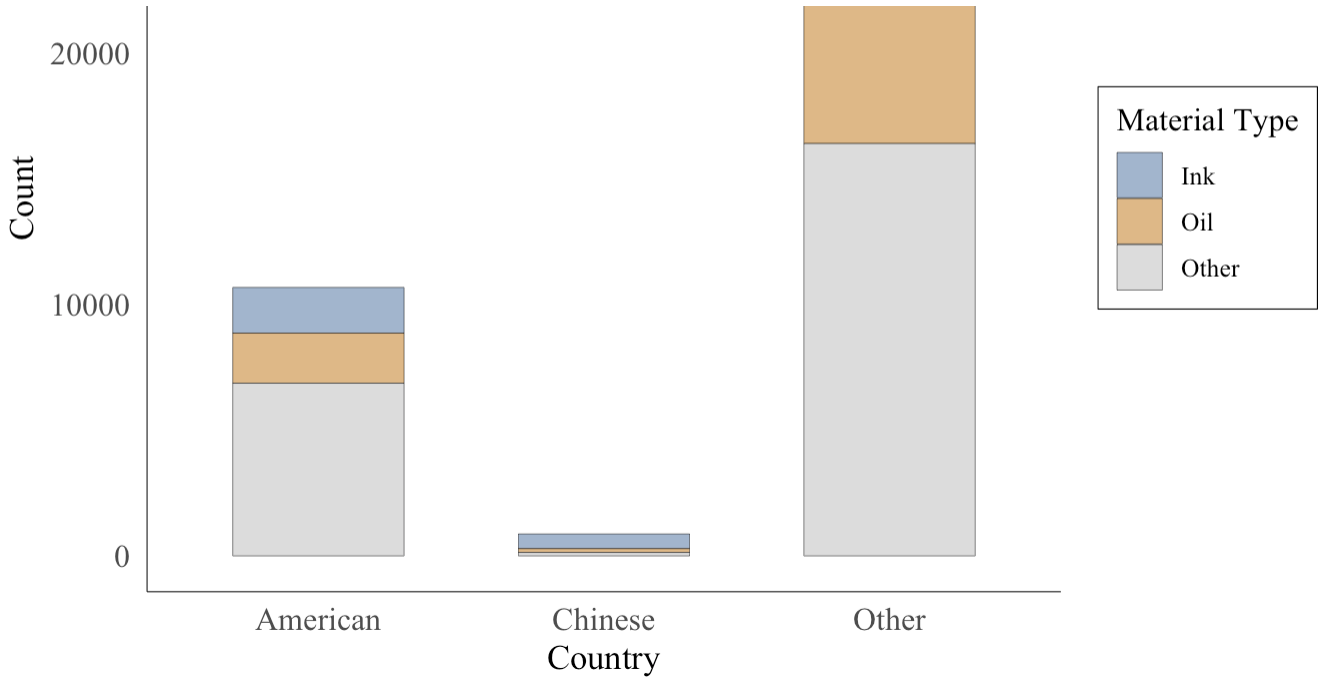
Investigate the relationship between `us.china` and `material.type`

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

> **Think:** I want to explore the distribution of material types (Oil, Ink, Other) in artworks based on the country of origin (American, Chinese, Other) to understand how material types differ across these countries. Both variables (material type, country) are categorical, meaning a bar chart is optimal for visualization.

> **Show:** A stacked bar chart to visualize the distribution of material types (Oil, Ink, Other) by country (American, Chinese, Other) in the dataset.





> **Tell:** The stacked bar chart shows that material types are distributed differently across countries. Chinese artworks are more likely to use ink, as reflected by the higher proportion of ink paintings in this category. In contrast, American artworks have a more balanced distribution between oil and ink, but with a much larger proportion of artworks using "Other" materials. The "Other" category is predominantly made up of non-oil and non-ink materials, and it exceeds the amount of ink used in American artworks. Although the total number of Chinese artworks in the dataset is smaller, a higher proportion of these artworks use ink compared to American and "Other" artworks.

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.

This result provides valuable insight into the distinct artistic practices and traditions that shape the use of materials in Chinese and American artworks. The higher prevalence of ink in Chinese paintings likely reflects centuries of tradition where ink has been central to the creation of classical Chinese art, such as calligraphy and landscape painting. This material is not only a medium of choice but also deeply tied to cultural and philosophical ideals, such as simplicity, fluidity, and the pursuit of balance in artistic expression.

The significant use of oil in American artworks may be indicative of Western artistic traditions, where oil paint became the dominant medium following the Renaissance. Oil painting allows for rich texture, depth, and versatility, which may explain its prominence in American art, particularly in portraiture, landscape, and still life. Additionally, the high proportion of "Other" materials in American artworks suggests that contemporary or modern American artists may experiment more freely with different materials and mixed media, reflecting a broader trend of innovation and

diversity in modern art movements.

In terms of understanding the dataset, this analysis highlights the importance of material choice as a cultural and historical indicator, which can offer context when interpreting the artworks themselves. The preference for certain materials might reveal how different artistic practices evolve based on geography, tradition, and access to resources.

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.

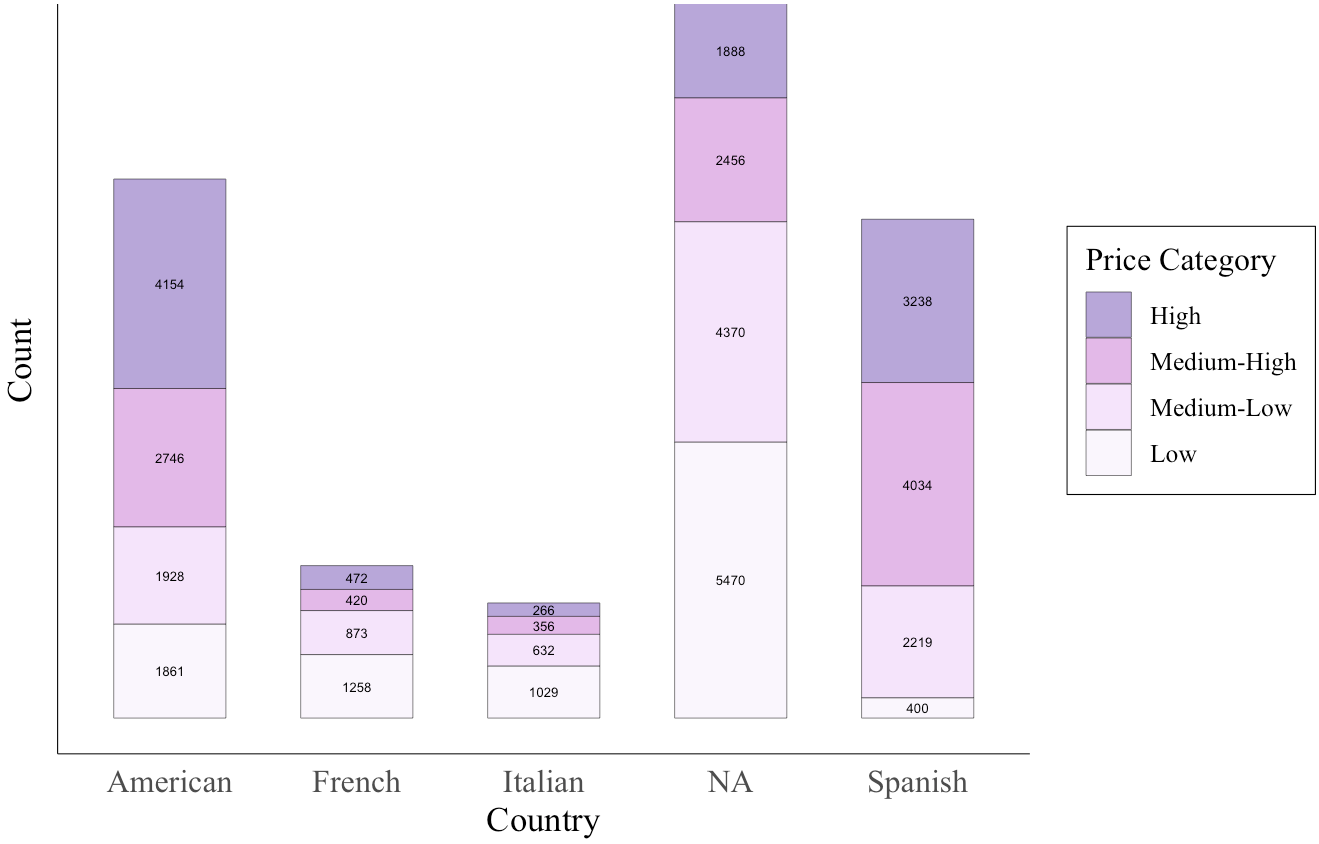
```
artdata_cleaned <- artdata_cleaned %>%
  mutate(
    famous.countries = case_when(
      grepl("American", country, ignore.case = TRUE) ~ "American",
      grepl("French", country, ignore.case = TRUE) ~ "French",
      grepl("Italian", country, ignore.case = TRUE) ~ "Italian",
      grepl("Spanish",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

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

> **Think:** I wanted to categorize the artworks based on their price ranges, grouping them into categories like Low, Medium-Low, Medium-High, and High. This will help in understanding how the pricing of artworks is distributed among different countries, and whether certain countries are more likely to have higher-priced artworks than others. I visualized the distribution with a stacked bar chart, where each country's bar represents the count of artworks in each price category, and I added the count labels to the bars to show the number of artworks in each category. The colors in the stacked bars correspond to the price categories.

> **Show:** A stacked bar chart to visualize the distribution of artworks across different price categories (Low, Medium-Low, Medium-High, and High) in various countries (American, French, Italian, Spanish, NA) in the dataset.

Price Category Distribution by Country

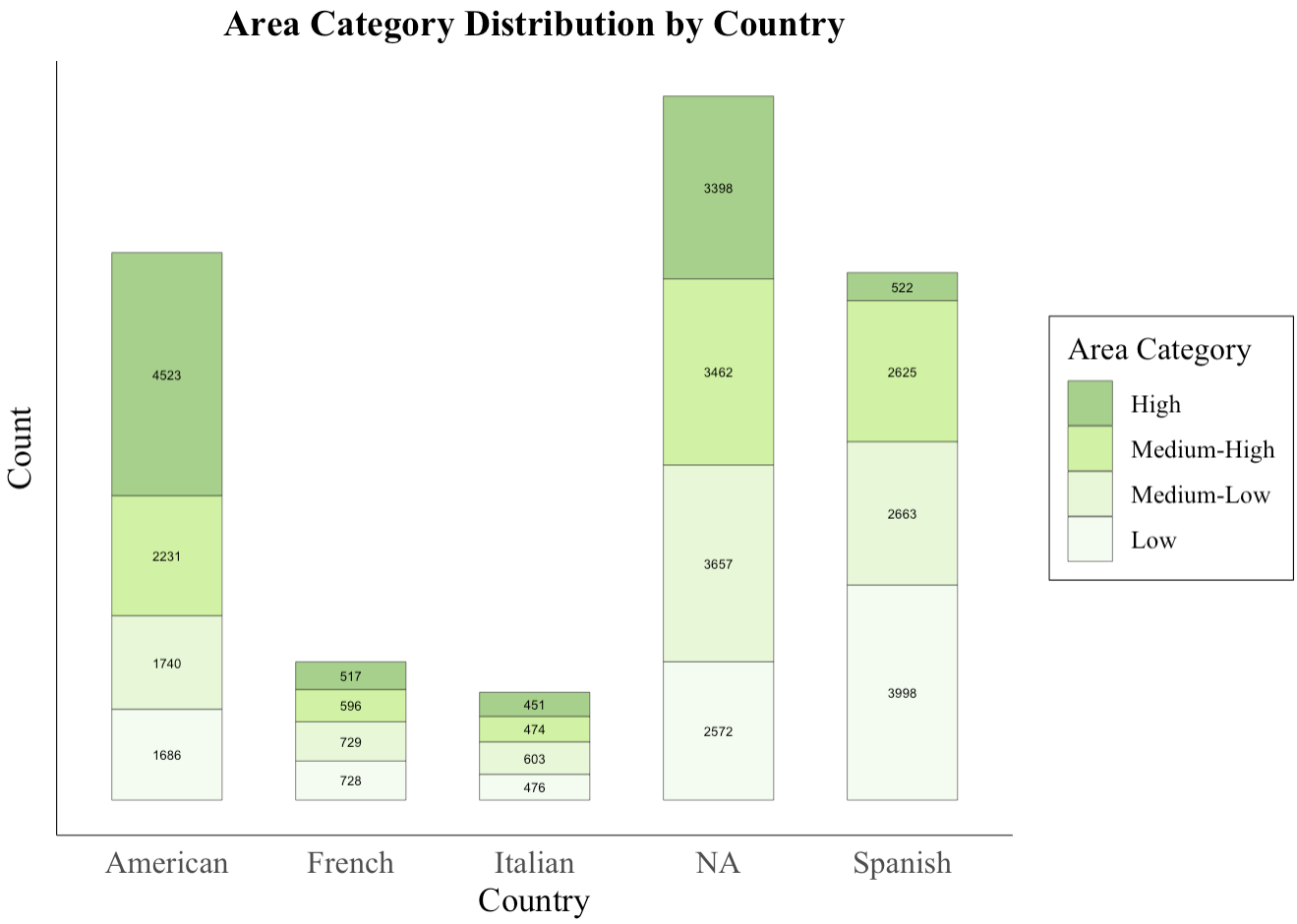


Tell: The distribution of artworks across price categories varies significantly by country. Generally, most countries have more artworks in the low price category than in the higher price categories. However, the United States stands out, with a notably higher number of artworks in the high and medium-high price categories compared to the low price category. In contrast, France and Italy show a trend of having more artworks in the low price categories and fewer in the high categories, which aligns with the overall pattern observed in the "NA" column (representing other countries). This is in direct opposition to the U.S. trend, where the lowest price category contains the fewest artworks, and the high price category contains the most. Spain also deviates from the general trend, with a minimal number of artworks in the low price category and a much larger proportion in the medium-high and high price categories. These differences in pricing distributions provide insight into how countries vary in terms of the value of their artworks, with the U.S. and Spain standing out for their preference for higher-priced artworks.

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

Think: This time, I wanted to categorize the artworks based on their areas, grouping them into categories like Low, Medium-Low, Medium-High, and High. Similar to before, this will help in understanding the distribution of area among different countries. I visualized this distribution with the same kind of stacked bar chart, where each country's bar represents the count of artworks in each area category, with count labels to show the number of artworks in each category. The colors in the stacked bars correspond to the area categories.

Show: A stacked bar chart to visualize the distribution of artworks across different area categories (Low, Medium-Low, Medium-High, and High) in various countries (American, French, Italian, Spanish, NA) in the dataset.



Tell: The area category distribution mirrors the price distribution for American, French, and Italian artworks, where larger artworks generally correspond to higher prices. The NA distribution is similar, though with slightly fewer artworks in the "low" area category, suggesting fewer smaller artworks. The most significant difference is in Spain, where a higher proportion of artworks fall into the "low" area category, and fewer artworks are in the "high" area category. This pattern suggests that countries like the U.S., France, and Italy tend to align in terms of larger artworks being more expensive. Spain, however, deviates by having more artworks of medium size and fewer large ones, which could reflect a preference for more accessible artwork sizes. Despite the smaller dimensions, these works may still be priced higher, suggesting a potential focus on quality or artistic value over size.

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?

confirm your guess as to the cause of the differences/similarities?

The observed differences in the price categories of artworks across countries can be attributed to various factors. In the United States, the higher prevalence of high and medium-high priced artworks likely reflects the thriving contemporary art market, where works by prominent American artists are highly valued. The U.S. art scene, with its significant presence of major auction houses and art fairs, contributes to the high prices of modern and contemporary artworks.

In contrast, countries like France and Italy, which are known for their historical and cultural ties to classical art, tend to have more artworks in the lower price range. Much of their art market centers around preserving and promoting traditional works, such as those from the Renaissance or Baroque periods, which are often more affordable than contemporary pieces. This focus on regional or traditional art likely explains the larger volume of lower-priced artworks in these countries.

Spain, on the other hand, shows a growing market for contemporary and postmodern Spanish artists, which is reflected in the higher concentration of medium-high and high-priced artworks. This trend may indicate a rising international demand for Spanish art, driving up prices for modern works. Overall, the variations in price categories across these countries highlight how cultural, historical, and market-driven dynamics influence the art market. While the U.S. benefits from a lucrative contemporary art market, countries like France and Italy focus more on preserving historical art traditions, and Spain's increasing emphasis on contemporary art reflects global market shifts.

Furthermore, in the U.S., there is a cultural emphasis on boldness, ambition, and grandeur, which may be reflected in the preference for larger artworks. This aligns with the country's overall appreciation for larger-than-life experiences, seen in architecture (e.g., skyscrapers), entertainment (e.g., blockbuster films), and consumer goods. Larger art pieces can symbolize wealth, prestige, and a sense of power, which resonates with American cultural values. Moreover, the U.S. has a market that highly values contemporary art, and large-scale installations or abstract works are common in galleries and exhibitions, which may contribute to the prevalence of larger, more expensive artworks.

In contrast, Spain has a strong tradition of art rooted in classical, religious, and historical contexts. Spanish art has been heavily influenced by intimate forms of expression, such as religious iconography and portraiture, where smaller works (e.g., altarpieces, still lifes) were more common. Additionally, Spain's art market may prioritize accessibility and regionalism. Smaller works are often more affordable, easier to display in various spaces, and could appeal to a broader audience, including local collectors and tourists. Furthermore, Spain has a deep historical connection to the "Cultural Revolution" during the 20th century, which valued a more personal, intimate form of art that may have been reflected in smaller, more detailed works.

To confirm how cultural and historical contexts affect price and area distributions, we can categorize artworks by era (classical, modern, contemporary) and then examine their price/area patterns. In countries like Italy and France, where classical art is dominant, lower-priced artworks should be more prevalent, reflecting a focus on traditional and regional works. In contrast, the U.S. and Spain, with a greater emphasis on contemporary art, should display a higher proportion of high-

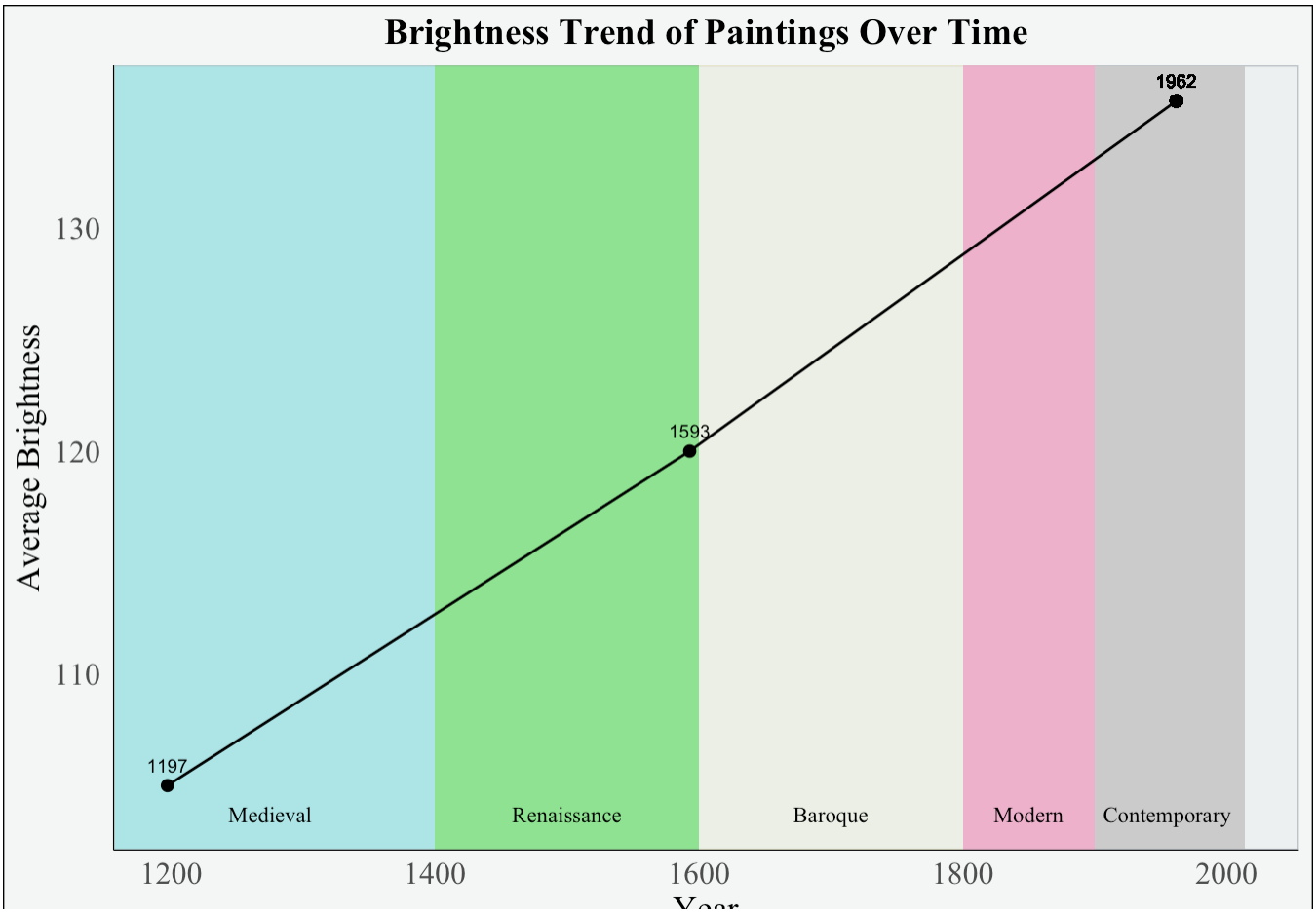
priced artworks, showing the influence of modern art's commercial appeal. By comparing these trends across countries, we can then determine if the price of artworks is shaped by the era in which they were created.

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.

To investigate how the brightness of paintings has changed over time, I would select the years 1197 (Medieval Art - Romanesque/Gothic), 1593 (Renaissance), and 1962 (Contemporary/Modern Art - Abstract Expressionism/Pop Art). The year 1197 represents the Medieval period, where art, often religious, used muted tones to convey spiritual themes, with brightness symbolically used in elements like stained glass. In contrast, by 1593 during the Renaissance, artists focused on naturalism, using light and shadow (chiaroscuro) to create realistic depictions, leading to brighter, more detailed artworks. By 1962, the shift to modern and contemporary art saw a dramatic use of bold, vibrant colors, as movements like Abstract Expressionism and Pop Art embraced artificial brightness, influenced by mass culture and commercialism. Examining these three periods reveals how brightness in art evolved alongside cultural, religious, and technological changes.



Year

5b. Finding the average

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

```
mean_1197 <- mean(brightness_1197$brightness)

mean_1593 <- mean(brightness_1593$brightness)

mean_1962 <- mean(brightness_1962$brightness)
```

The average brightness for the three years are as follows:

1197: 105

1593: 120

1962: 135.7083

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:** By calculating the z-scores for the years 1197, 1593, and 1962, we can compare how the brightness of paintings from these years deviates from the overall mean brightness. The smaller the absolute value of the z-score, the closer that value is to the overall mean. We will analyze the z-scores to see how the brightness of paintings changed over time in relation to the overall distribution of brightness. To calculate the z-scores for the years 1197, 1593, and 1962, we first need to find the total mean brightness (μ) and the standard deviation (σ) of the entire dataset. The mean is the average brightness across all artworks, which is calculated by summing all the brightness values and dividing by the number of data points. The standard deviation measures the spread of the brightness values, indicating how much the values deviate from the mean. It is calculated by taking the square root of the average squared differences from the mean.

> **Show:**

The z-scores for the three years are as follows:

1197: -0.8858569

1593: -0.5876851

1962: -0.2754337

> **Tell:** The z-scores indicate that the brightness of paintings from these years generally becomes closer to the overall mean as time progresses. In 1197, the brightness is the furthest below the average, while 1593 is closer, and 1962 is the closest, though still below the mean. This suggests that over time, art moved from darker, more subdued tones (as seen in Medieval art) toward brighter and more vibrant colors (as observed in the 1960s). The trend indicates a gradual increase in the brightness of paintings, which aligns with the cultural shifts in art from religious themes in the Medieval period to the realism of the Renaissance, and finally to the bold, vibrant colors of Modern and Contemporary art. However, since both 1197 and 1593 only had one artwork, the mean for these years is less reliable and may not accurately represent the broader trend. These individual artworks likely skew the mean, meaning the z-scores for these years might not fully reflect the broader historical context. This suggests caution when interpreting the early years due to the limited sample size.

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 z-scores suggest that brightness values from 1197 and 1593 are below the overall mean, while the 1962 brightness value is closer to the mean, although still slightly below it. However, the small sample size for 1197 and 1593, each containing only one artwork, significantly limits the reliability of these results. With only one data point per year, the mean brightness for those years is equal to the value of the single artwork, and there is no variation to provide a meaningful standard deviation. This makes the z-scores for these years less robust compared to the 1962 data, where there is a broader set of artworks to calculate more reliable mean and standard deviation values.

The limitations of this analysis stem from the sample size of the early years, which significantly impacts the accuracy and generalizability of the results. Since the z-scores are based on just one artwork for 1197 and 1593, these findings are not as indicative of the broader trend in those time periods. Additionally, the brightness value itself may be influenced by various factors such as lighting, materials, or the medium used in the artworks, which are not accounted for in this analysis.

To better investigate whether the brightness of paintings has changed over time, I would calculate the z-scores for artworks from different historical periods. First, I would categorize the artworks into distinct eras, such as Medieval, Renaissance, and Modern Art. For each era, I would calculate the mean brightness and the standard deviation of brightness of all artworks within that period. The z-score for each artwork would then be computed by subtracting the mean brightness of its respective era from the artwork's brightness and dividing by the standard deviation of the era's brightness. This would standardize the brightness values, allowing for direct comparisons across different periods. A high positive z-score would indicate that an artwork is brighter than the average for its era, while a negative z-score would indicate it is darker.

After calculating the z-scores, I would compare them across different periods to determine if there are consistent trends in brightness over time. If artworks from, for example, the Renaissance or Modern periods generally have higher z-scores, it would suggest that those periods tended to produce brighter works compared to earlier periods like Medieval art. Additionally, I would visualize the distribution of z-scores for each era using boxplots or histograms to examine the spread of brightness values and identify any outliers. This analysis would provide a clear, statistical view of how the brightness of paintings has evolved and whether certain periods favored brighter or darker artworks.

Question 6: Your own investigation (15 points)

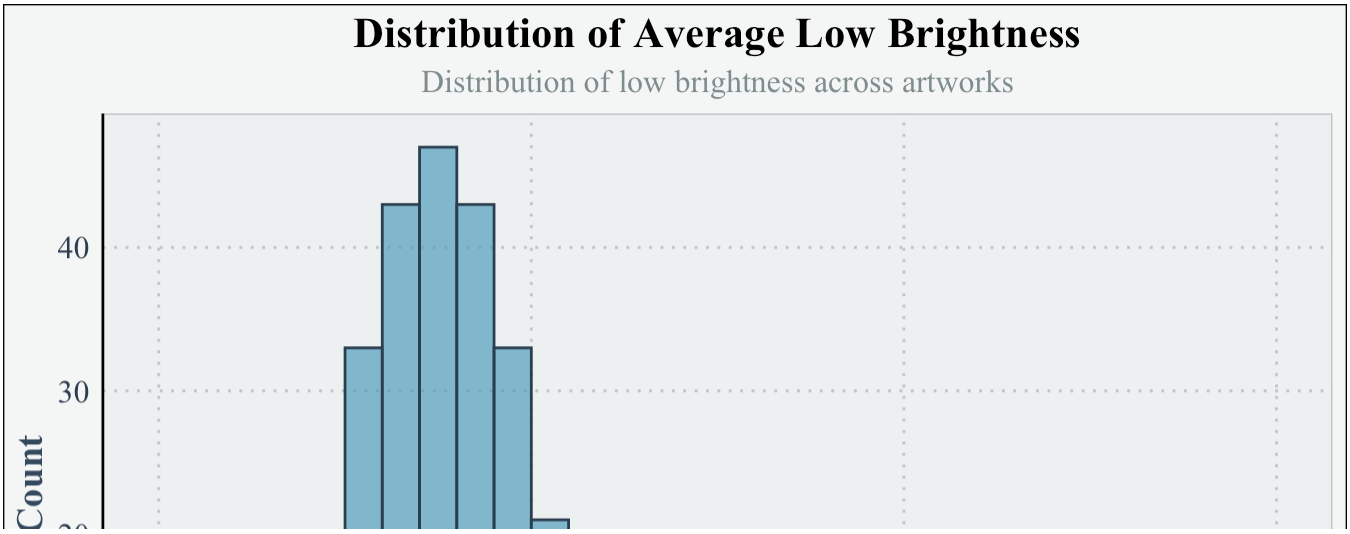
6a. Selecting your own question

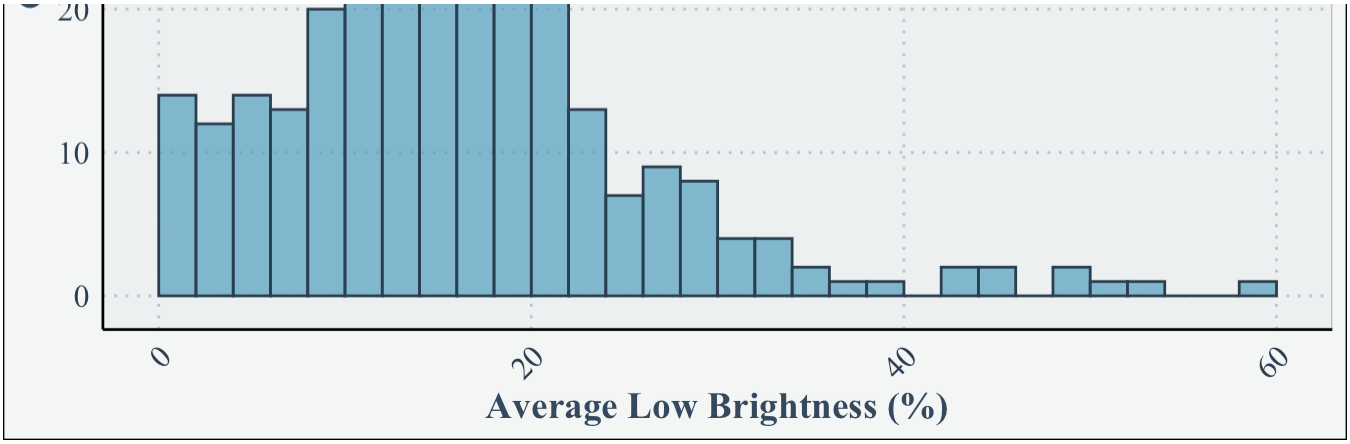
Similar to the previous questions, think of your own question that you would like to ask of the data. Use the Think-Show-Tell procedure to conduct your investigation. Think deeply about what your result means.

Question: "How is the distribution of average low brightness (percentage of low-light areas) across artworks?"

> **Think:** Following the prior question, I became curious about the general distribution of darkness levels across all the artworks in the dataset. If the low light percentage is measuring the proportion of an image that is dark or in shadow, we can think about how most artworks are likely to have a mix of light and dark areas, but some may have an overwhelming dominance of darker tones. Understanding these statistics will help us see whether the majority of artworks lean toward dark or light compositions, and if there are any extreme outliers that might suggest unusual cases. Similar to Question 2, Low Brightness % is a quantitative variable, so histogram is a useful tool to display this distribution.

> **Show:** A histogram of the data shows a fairly symmetric distribution with a slight right skew.





Summary Statistics for Painting Heights

Minimum	Q1	Median	Mean	Q3	Maximum	SD
0	2.79	10.67	14.35	21.94	90.47	13.91

Tell: The minimum value of 0.000 indicates that some artworks in the dataset contain no low-light or dark areas at all. This suggests that these artworks are either fully illuminated or feature very little darkness, resulting in a predominantly bright composition. Such pieces might be characterized by vibrant color palettes or an emphasis on light. Additionally, with 25% of the artworks having a lowlightpercentage below 2.953, it implies that a quarter of the artworks feature very little darkness, suggesting that many artworks are largely light or feature minimal low-light areas. The median value of 10.92 indicates that half of the artworks have a low light percentage less than this value. Since the median is relatively low, it suggests that for 50% of the artworks, the dark areas make up less than 11% of the total composition, meaning that most artworks in the dataset are quite light. The mean of 14.561, which is higher than the median, suggests a right-skewed distribution. This means that while most artworks contain minimal dark areas, a few artworks with much higher levels of darkness are pulling the average upward. These darker pieces might emphasize shadows or create strong contrasts. The third quartile of 22.11 indicates that 75% of the artworks have less than this percentage of low-light areas. This reinforces the idea that the majority of artworks remain relatively light, with low-light areas comprising less than a quarter of the total composition. However, the dataset still includes artworks that lean towards darker compositions. The maximum value of 91.75 points to some extreme outliers, where nearly 92% of the artwork is in shadow or dark areas. These artworks likely focus on high contrast or darker color schemes, standing in stark contrast to the majority of lighter artworks in the dataset.

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.

In summary, this investigation provided valuable insights into how different types of data

visualization techniques can be applied to reveal meaningful patterns in artwork data. I learned that histograms are particularly useful for visualizing general distributions of continuous quantitative variables, while bar plots are better suited for categorical comparisons, like country and material type. This process of trial and error helped me understand that while there are multiple approaches to solving a problem, choosing the right method can enhance the clarity of the results.

Moreover, I realized the importance of not rushing to interpret the results. For example, when analyzing price and area distributions by country, it's crucial to understand the cultural and historical context before drawing conclusions. For instance, larger artworks might be valued differently in various countries due to cultural preferences for size or tradition, and such context can heavily influence the data interpretation.

Furthermore, I now understand that while it's tempting to make things visually impressive right away, it's more important to first focus on accuracy and clarity of my analysis. It's better to take my time and ensure the plot is done right than to rush and make it fancy too soon. But at the same time, efficiency is also important, and that there's still so much I have to learn to make my data analysis more precise.

Ultimately, these lessons highlighted how data visualization techniques can tell us more than just numbers, but can uncover deeper, context-dependent stories. This homework reinforced the importance of considering both the data itself and the broader factors that could shape it when making interpretations.