STA130H1S – Winter 2021

Week 2 Problem Set

N. Moon and S. Caetano Yixing Xu

Instructions

How do I hand in these problems for the January 21st deadline ?

Your complete .Rmd file that you create for these practice problems AND the resulting pdf (i.e., the one you 'Knit to PDF' from your .Rmd file) must be uploaded into a Quercus assignment (link: https://q.utoronto.ca/courses/206597/assignments/533381) by 11:59AM ET, on Thursday, January 21st. Late problem sets are not accepted.

Problem set grading

There are two parts to your problem set. One is largely R-based with short written answers and the other is more focused on writing. We recommend you use a word processing software like Microsoft Word to check for grammar errors in your written work. Note: there can be issues copying from Word to R Markdown so it may be easier to write in this file first and then copy the text to Word. Then you can make any changes flagged in Word directly in this file.

Part 1

[Question 1] The `artwork500.csv` file contains data for a sample of 500 pieces of art owned by the Tate Art Museum. While more variables are available on the Tate Art Museum's site (github.com/tategallery/collection), you will only be working with the variables featured in the `artwork500.csv` file:

- id: Unique ID for each piece of artwork
- artist: Name of the artist
- title: Title of the artwork
- type: Medium used
- year: Year the artwork was created
- width: width of the artwork, in mm
- height: height of the artwork, in mm
- units: measurement units for width and height of the artwork
- area: surface area (in squared cm)

```
library(tidyverse) # Load the tidyverse package so it is available to use
artwork500 <- read_csv("artwork500.csv")
```

(a) Use the `glimpse()` function to view properties of the `artwork500` data set. How many observations does it include? How many variables are measured for each observation? How many rows and columns does the `artwork500` data frame have?

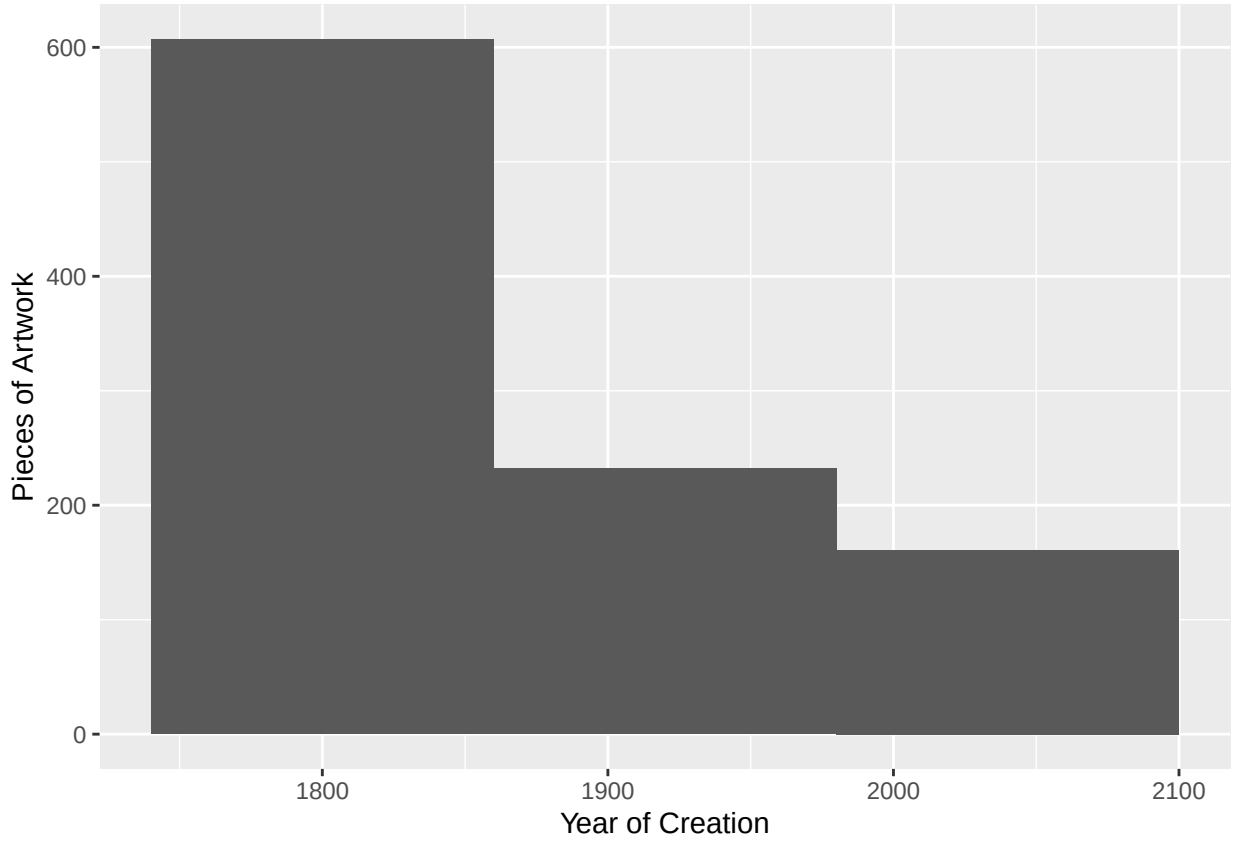
```
artwork500 %>% glimpse()
```

```
## Rows: 1,000
## Columns: 10
## $ id              <dbl> 52628, 8756, 63026, 43485, 52118, 52634, 52755, 548...
## $ artist          <chr> "Turner, Joseph Mallord William", "LeWitt, Sol", "T...
## $ title           <chr> "The Bridge", "A Square Divided Horizontally and Ve...
## $ type            <chr> "Watercolour", "Watercolour", "Watercolour", "Water...
## $ year            <dbl> 1820, 1982, 1830, 1819, 1830, 1831, 1820, 1835, 196...
## $ acquisitionYear <dbl> 1856, 1984, 1856, 1856, 1856, 1856, 1856, 1856, 200...
## $ width           <dbl> 300, 607, 242, 259, 140, 307, 152, 181, 364, 163, 2...
## $ height          <dbl> 485, 607, 303, 406, 192, 489, 243, 229, 376, 243, 1...
## $ units           <chr> "mm", "mm", "mm", "mm", "mm", "mm", "mm", "mm", "mm...
## $ area            <dbl> 1455.00, 3684.49, 733.26, 1051.54, 268.80, 1501.23,...
```

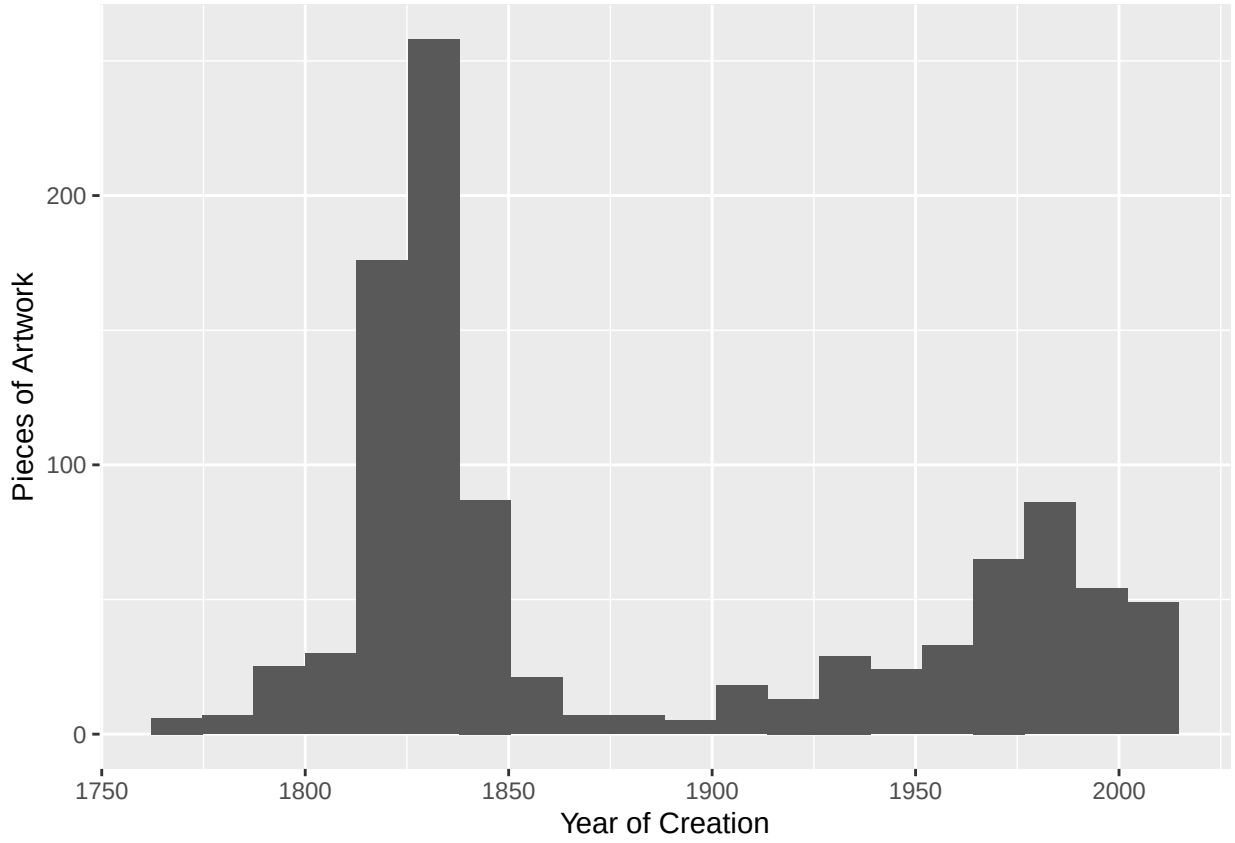
<There are 1000 rows/observations, and there are 10 columns/variables for each observation.>

(b) Create 3 histograms to explore the distribution of years of creation for this sample of pieces of art: (i) one with 3 bins, (ii) one with 20 bins, and (iii) one with 75 bins; make sure to specify meaningful axis labels where appropriate. Which of these histograms is most appropriate to describe the distribution of the artworks' years of creation? Why? Write a few sentences describing the distribution based on the histogram you chose as most appropriate.

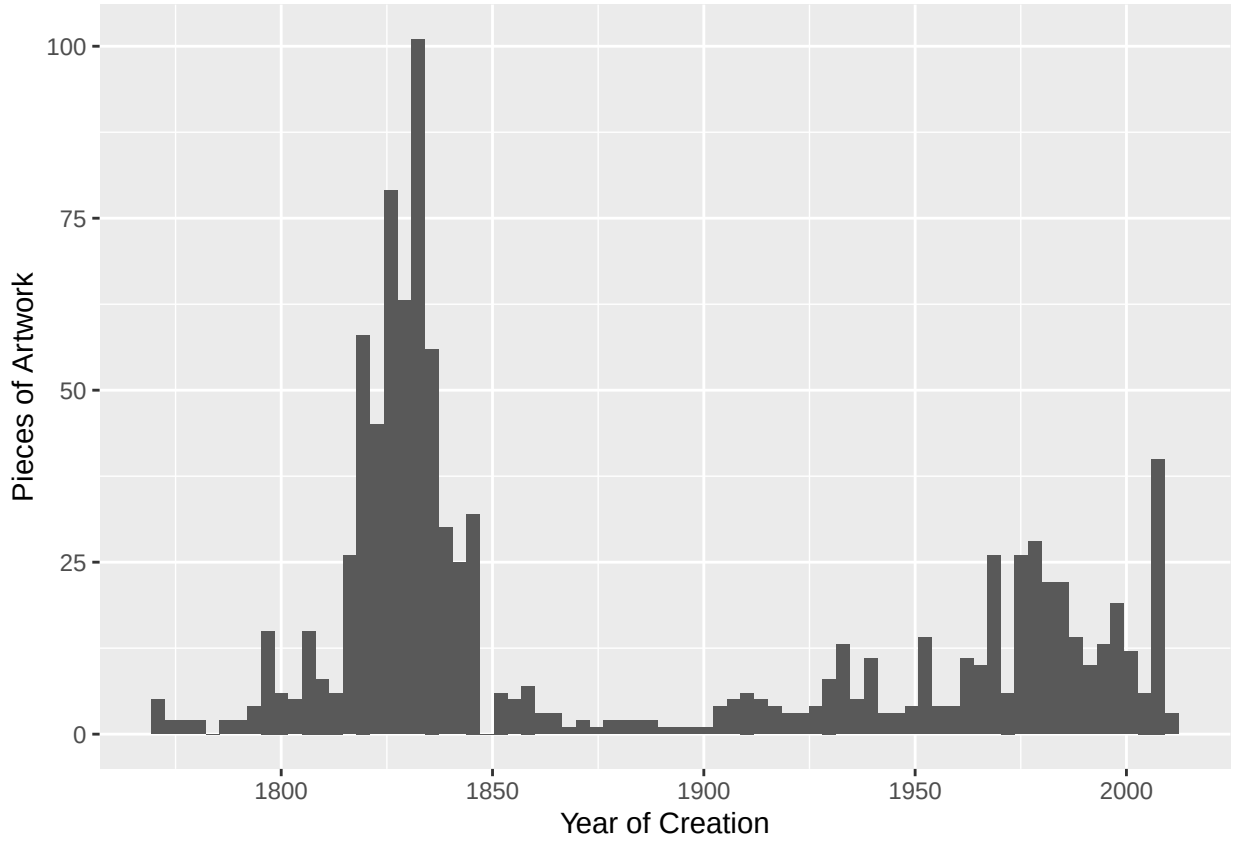
```
#(i)
artwork500 %>% ggplot(aes(x = year)) + geom_histogram(bins = 3) +
  labs(x = "Year of Creation", y = "Pieces of Artwork")
```




```
#(ii)
artwork500 %>% ggplot(aes(x = year)) + geom_histogram(bins = 20) +
  labs(x = "Year of Creation", y = "Pieces of Artwork")
```



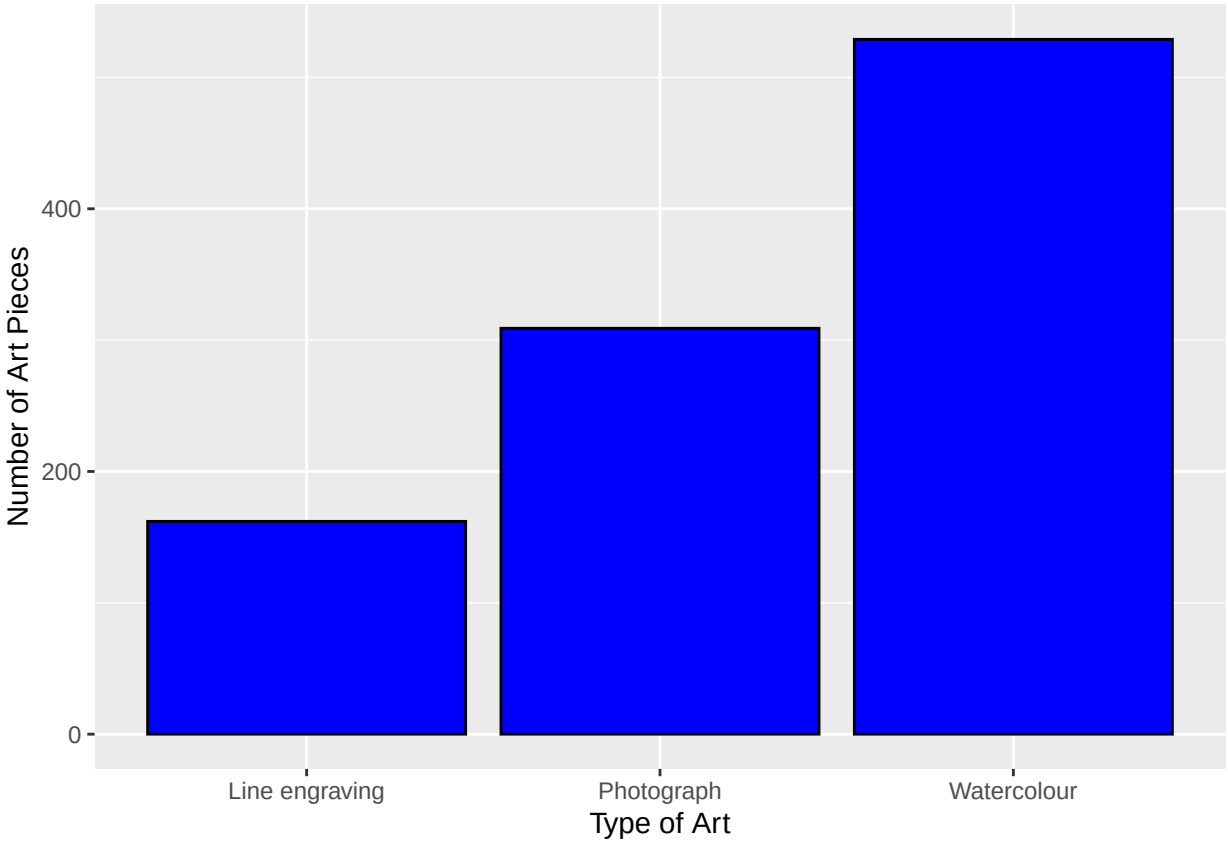

```
#(iii)
artwork500 %>% ggplot(aes(x = year)) + geom_histogram(bins = 75) +
  labs(x = "Year of Creation", y = "Pieces of Artwork")
```



The second histogram is the best for representing the distribution of the artworks' years of creation, because it is easier to make conclusions about its shape, centre, and spread when the data is not too spread out as it is in the third graph or when the data is not too clunky and concentrated as it is in the first graph. The second graph allows us to make conclusions about the era of the works based on the century or decade in which it occurs. The third graph gets too specific regarding the year of creation, and the first graph would make too many generalizations about each century because its bars/bins are too wide and encompasses over a 100 years.

(c) Construct a plot to visualize the distribution of a categorical variable and describe the distribution in 1-2 sentences; make sure to specify meaningful axis labels where appropriate. Hint: If you choose a categorical variable with many different categories, you may find it useful to use coord_flip() to flip the bars horizontally and/or change the options in the R code chunk to make the plot large (ex: {r, fig.height=15, fig.width=5}).

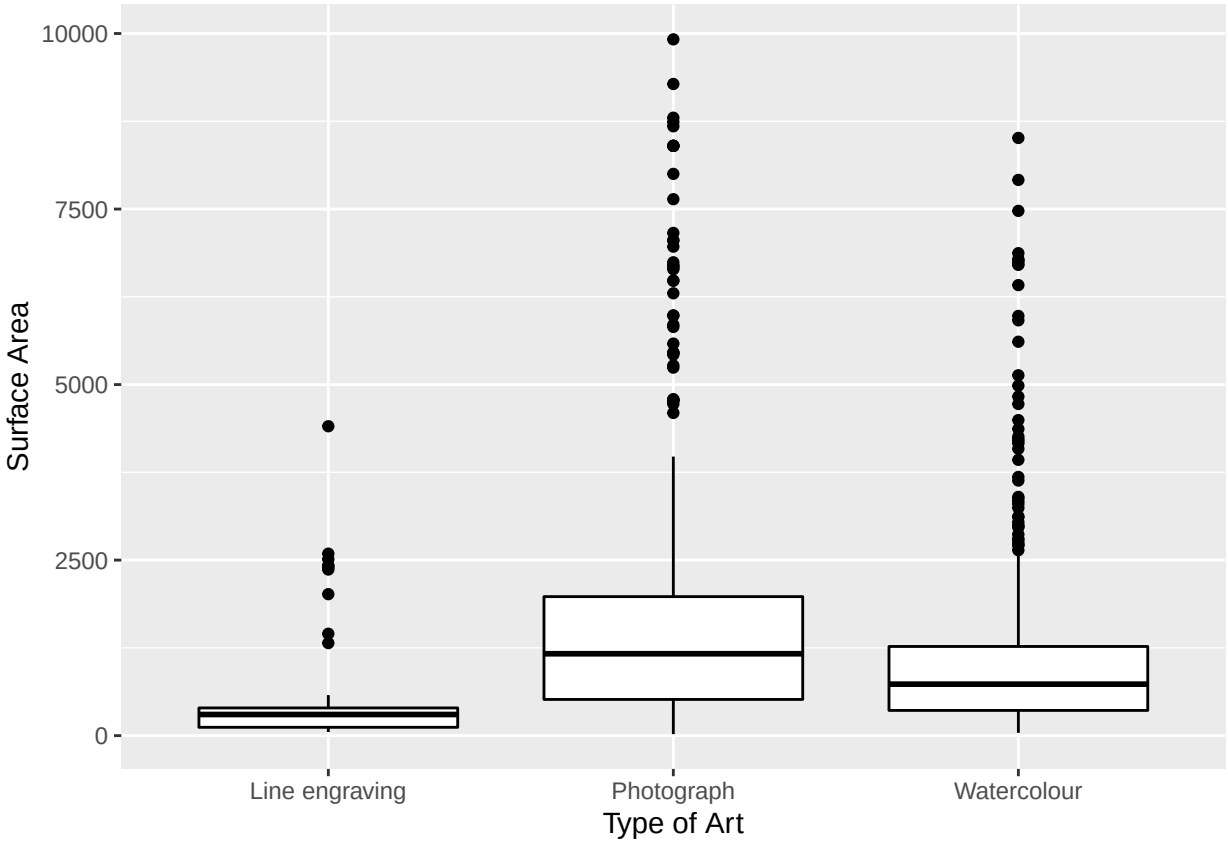
```
artwork500 %>% ggplot(aes(x = type)) +
  geom_bar(color = "black", fill = "blue") +
  labs(x="Type of Art", y = "Number of Art Pieces")
```



This is a barplot to compare the type of paintings to see which ones are most popular for artists to paint. Clearly, artists chose to produce Watercolour art the most frequently, while Line engraving art was the least popular form of art.

(d) Construct a set of three boxplots showing visual summaries of the distribution of surface area (`area`) for each type of artwork (`type`); make sure to specify meaningful axis labels where appropriate. Write 3-4 sentences comparing these distributions.

```
artwork500 %>% ggplot(aes(x = type, y = area)) +
  geom_boxplot(color = "black") +
  labs(x = "Type of Art", y = "Surface Area")
```



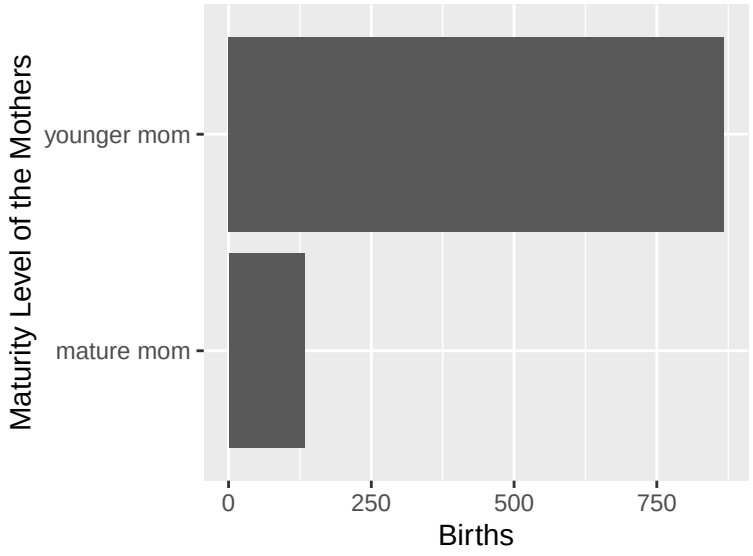
The distributions of all three types of art pieces are right skewed, which means each type's average surface areas are greater than their median surface areas. Also, half of the photograph artworks has greater surface area than 75% of the watercolour paintings, because the median mark for the boxplot for photographs is close to the 3rd quartile mark of the watercolour paintings' boxplot. We can also see that photographs has the greatest range of surface areas, while line engraving artworks had the least variance and range. What this tells us is that photographs are very versatile with surface areas.

[Question 2] The `ncbirths` data set is part of the `openintro` package. It consists of observations for a sample of 1000 births in North Carolina in 2004. Type `?ncbirths` in the R console for more information about the data and to see the definition of each variable. The code below loads the required libraries for this question and provides a glimpse of the `ncbirths` data frame.

```
?ncbirths
```

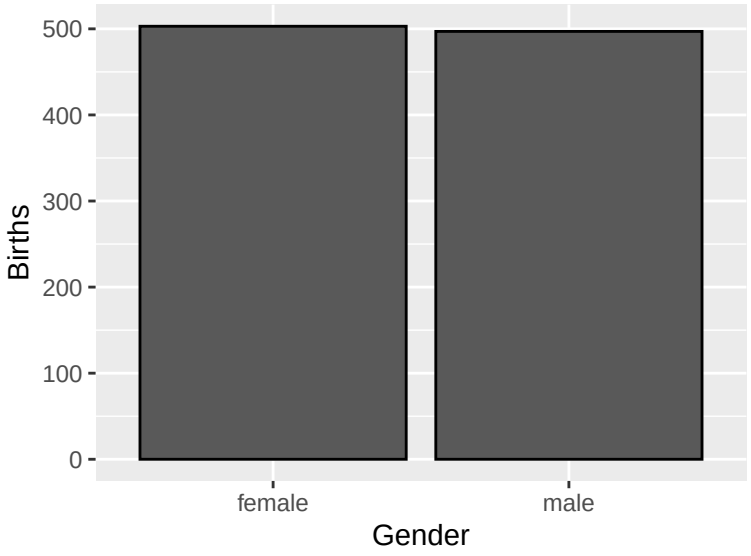
(a) Choose two categorical variables and plot their distributions. Identify whether each of these variables is a nominal or ordinal categorical variable. Write one or two sentences interpreting each plot.

```
# "Mature" is an ordinal categorical variable,
# because it can be ordered from least to most maturity level
# of the mothers or vice versa. What the barplot indicates is
# that there is a higher number of births for younger moms than for mature moms:
ncbirths %>% ggplot(aes(x = mature)) + geom_bar() +
  labs(x = "Maturity Level of the Mothers", y = "Births") +
  coord_flip()
```



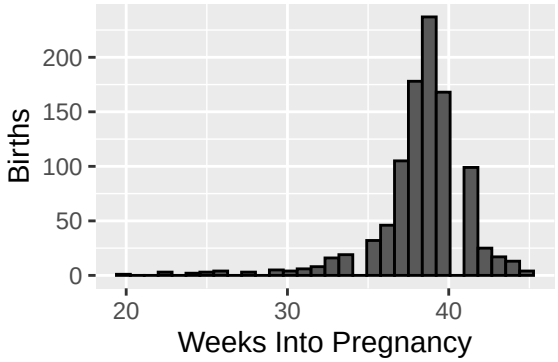

```
# "gender" is a nominal categorical variable, because it
# cannot be ordered/sorted. The barplot indicates that the
# birth of biologically female and male children are
# approximately equal.

ncbirths %>% ggplot(aes(x=gender)) +
  geom_bar(color = "black") +
  labs(x = "Gender", y = "Births")
```



(b) Choose a quantitative variable and plot its distribution. Identify whether the variable you selected is continuous or discrete, and write 2-3 sentences describing the distribution.

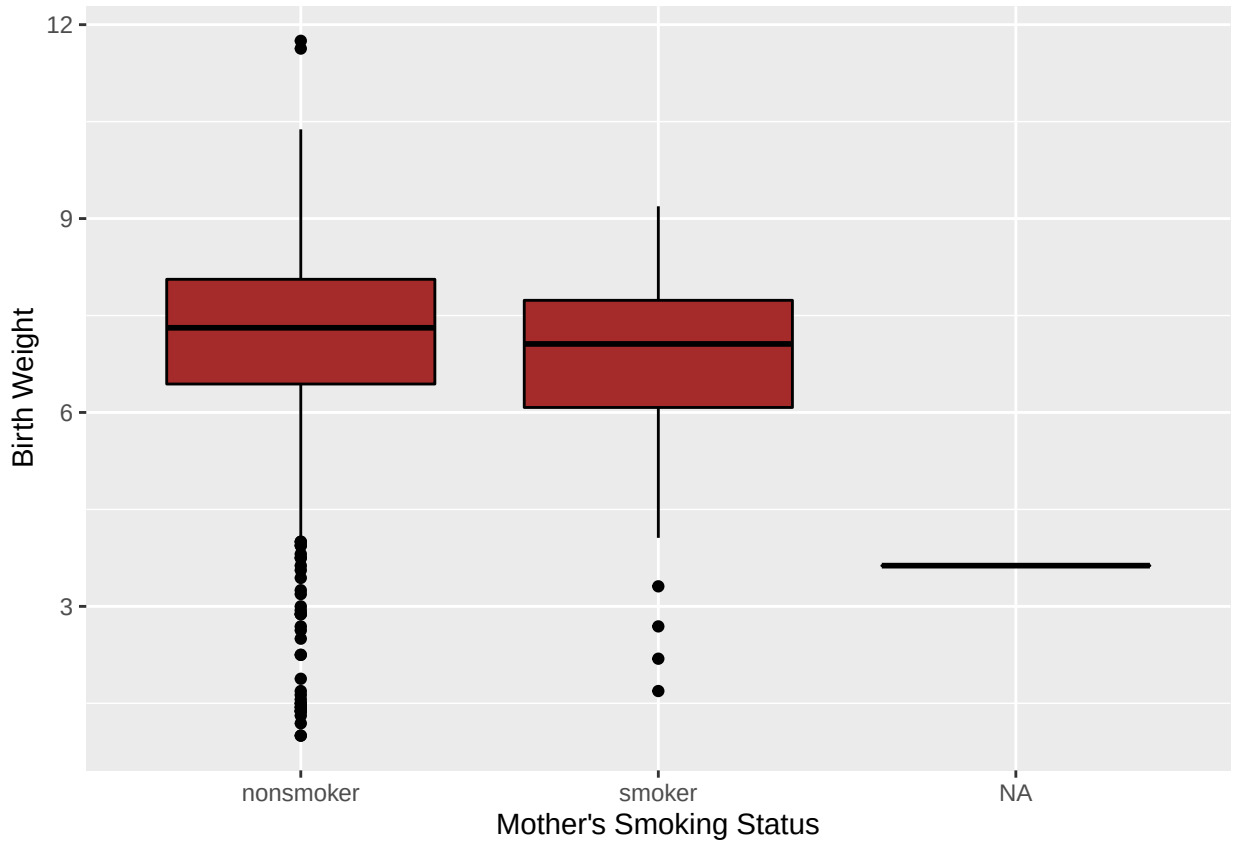
```
ncbirths %>% ggplot(aes(x = weeks)) +
  geom_histogram(color = "black") +
  labs(x = "Weeks Into Pregnancy", y = "Births")
```



The variable I chose was "weeks", which was for the number of weeks into the pregnancy at which the birth occurred. The histogram shows a surprising range from approximately 19 to 45 weeks. However, most of the data is concentrated near 37 to 40 weeks in, as expected (9 months of pregnancy). We can also see that the distribution is unimodal and that the peak is at around 38-40 weeks into the pregnancy, which is also the normal expected length of pregnancy.

(c) Construct a plot that shows the relationship between birth weight (`weight`) and mother's smoking status (`habit`); make sure to specify meaningful axis labels where appropriate.

```
ncbirths %>% ggplot(aes(x = habit, y = weight)) +
  geom_boxplot(color = "black", fill = "brown") +
  labs(x = "Mother's Smoking Status", y = "Birth Weight")
```



Part 2

Writing prompt:

Pretend that you are on the phone with your friend, and you want to share some of the cool new data visualization techniques that you have been learning in STA130. Pick one of the above graphs and prepare a small paragraph on how you would describe the graph to your friend (keeping in mind that they cannot see the graph). Make sure you include at least 2 words/phrases from your vocabulary list. It is also important to keep in mind that the person you are talking to has not taken STA130, therefore they will not be as familiar with the statistical vocabulary as you are. Therefore, make sure the explain any terms you use in plain language.

When describing a figure, it is important to:

- Describe the data source
- State the type of graph
- Identify what are on the x- and y-axis (if appropriate)
- Describe the distribution
- Make note of potential outliers

Vocabulary List

- Where are the data centered (approximate values if available)
- How much spread (relative to what?)
- Shape: symmetric, left-skewed, right-skewed
- The tails of the distribution (heavy-tailed or thin-tailed)
- Modes: where, how many, unimodal, bimodal, multimodal, uniform
- Outliers, extreme values
- Frequency (which category occurred the most or least often; data concentrated near a particular value or category)
- Mean, median, mode
- Standard deviation, interquartile range

Some general reminders

- Try to not spend more than 20 minutes on the prompt.
- Aim for more than 200 words but less than 350 words.
- Use full sentences.
- Grammar is *not* the main focus of the assessment, but it is important that you communicate in a clear and professional manner. I.e., no slang or emojis should appear.
- Be specific. A good principle when responding to a writing prompt in STA130 is to assume that your audience is not aware of the subject matter.
- Remember to end with a conclusion. This means reiterating the key points from your writing sample.

Hi, Paul! In STA130, I just learned a few techniques to help visualize a graph, and it has been really helpful in helping me analyze data too. For example, I used a histogram to visualize numeric data like the number of births after a number of weeks into the pregnancy using the resulting histogram's shape, centre, and spread. The histogram's x-axis represented the length of the pregnancy and the y-axis of the histogram represented the number of births. Now I know those terms are weird and you're probably wondering what they are, so let's get into them! Shape is the general pattern of the values being presented on graph. An attribute of the shape is its skewness, which tells us how centered or off-centered the graph is. In the case of the

histogram for the number of births after a number of weeks into the pregnancy, it is off-centered and leaning to the right, so it is left-skewed. This told us that most of the data was populated towards the right of the histogram. This meant that there were a greater number of births that occurred between 37 to 42 weeks. As for the histogram's centre, which is the peak in the graph and highest occurring value, the highest number of occurring births or peak of the graph was at approximately 38 weeks of pregnancy. The spread of the graph indicates how concentrated/centralized or spread out the data is. The spread of the histogram for the length of pregnancy vs births was spread out from approximately 19 to 45 weeks, which indicates the range in which births may occur along a pregnancy. Of course, there may be some outliers for the shorter lengths of pregnancy due to premature births. However, this histogram overall helps us conclude that births generally occur in between 37 and 42 weeks into the pregnancy, that the greatest number of births occur in approximately 38 weeks, and that the births can occur in varying lengths of pregnancy between 19 to 45 weeks.