STA130H1S – Winter 2024

Week 2 Problem Set - Sample Answers

N. Moon, J. Speagle, and [ADD YOUR NAME HERE]

Instructions

How do I hand in my solutions and how do I check my work

You will submit your solutions (.Rmd and .pdf) on MarkUs at the following link: https://markus4.teach.cs.toronto.edu/2024-01/courses/1 Submissions are due at 5pm on Thursdays; see Quercus page for the specific deadline for each problem set.

Usually when you do an assignment, you don't find out whether your answers are correct until *after* the deadline, when you get your grade back. However, using MarkUs, you can submit your work before the deadline and run tests to check your solutions!

Note: Some parts of some questions may not be covered by tests in MarkUs, but you're still responsible for reviewing the posted solutions and make sure you understand them. Some of these parts will also be graded in some weeks.

What to do if a test fails on MarkUs

- Take a deep breath! Your work won't really be graded until the deadline, so start early to make sure you have lots of time to resolve issues before the deadline.
- Read the message to get hints about what the problem is. For example "variable X not present" means that you may have a typo in the name of the variable we're looking for - re-read the question carefully and make sure you're following the instructions.
- Search on Piazza to see if other classmates have encountered a similar error (and if not, consider posting a screenshot of the error message).
- Come to TA or instructor office hours with your issue.

[Question 1]

The `artwork_sample.csv` file contains data for a sample of 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 `artwork_sample.csv` file (see below). Save this data object in an R object called `artwork`.

- `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
artwork <- read_csv("artwork_sample.csv")  # read in the data
```

(a) Use the `glimpse()` function to view properties of the `artwork` data set. How many observations does it include? How many variables are measured for each observation? Replace `NULL` with your answers in the code chunk below to save your answers in `Q1a_number_of_variables` and `Q1a_number_of_observations`.

```
glimpse(artwork)
```

```
## Rows: 1,000
## Columns: 10
## $ id              <dbl> 52628, 8756, 63026, 43485, 52118, 52634, 52755, 54884,~
## $ artist          <chr> "Turner, Joseph Mallord William", "LeWitt, Sol", "Turn~
## $ title           <chr> "The Bridge", "A Square Divided Horizontally and Verti~
## $ type            <chr> "Watercolour", "Watercolour", "Watercolour", "Watercol~
## $ year            <dbl> 1820, 1982, 1830, 1819, 1830, 1831, 1820, 1835, 1969, ~
## $ acquisitionYear <dbl> 1856, 1984, 1856, 1856, 1856, 1856, 1856, 1856, 2009, ~
## $ width           <dbl> 300, 607, 242, 259, 140, 307, 152, 181, 364, 163, 219,~
## $ height          <dbl> 485, 607, 303, 406, 192, 489, 243, 229, 376, 243, 152,~
## $ 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, 36~
```

```
Q1a_number_of_variables <- 10
Q1a_number_of_observations <- 1000
```

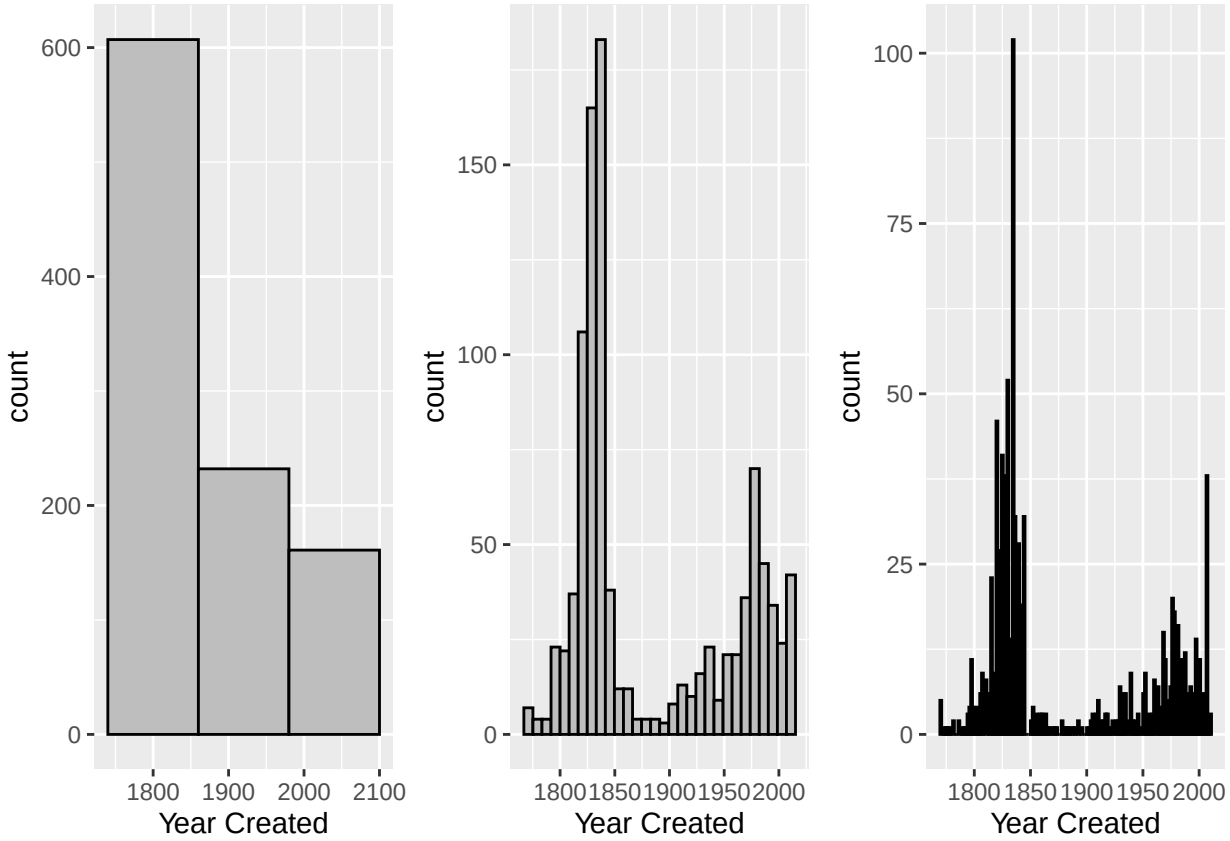
Write 1-2 sentences describing the sample using this information and the context.

There are 1000 different art pieces with 10 different observations for each.

(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 30 bins, and (iii) one with 150 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.

```
# Create your artwork below
hist1 <- artwork %>% ggplot(aes(x = year)) +
  geom_histogram(color = "black",
```

```
                 fill = "gray",
                 bins = 3) +
  labs(x = "Year Created")
hist2 <- artwork %>% ggplot(aes(x = year)) +
  geom_histogram(color = "black",
                 fill = "gray",
                 bins = 30) +
  labs(x = "Year Created")
hist3 <- artwork %>% ggplot(aes(x = year)) +
  geom_histogram(color = "black",
                 fill = "gray",
                 bins = 150) +
  labs(x = "Year Created")

# the gridExtra package allows for plots to be
# arranged in a grid layout - we'll load it here
# you are NOT REQUIRED to know or use this function
# we're just showing you how we set things up
library(gridExtra)
grid.arrange(hist1, hist2, hist3, nrow=1, ncol=3)
```



Which histogram is most appropriate to visualize these data?

I think the histogram with 30 points is the most appropriate way to visualize the data.

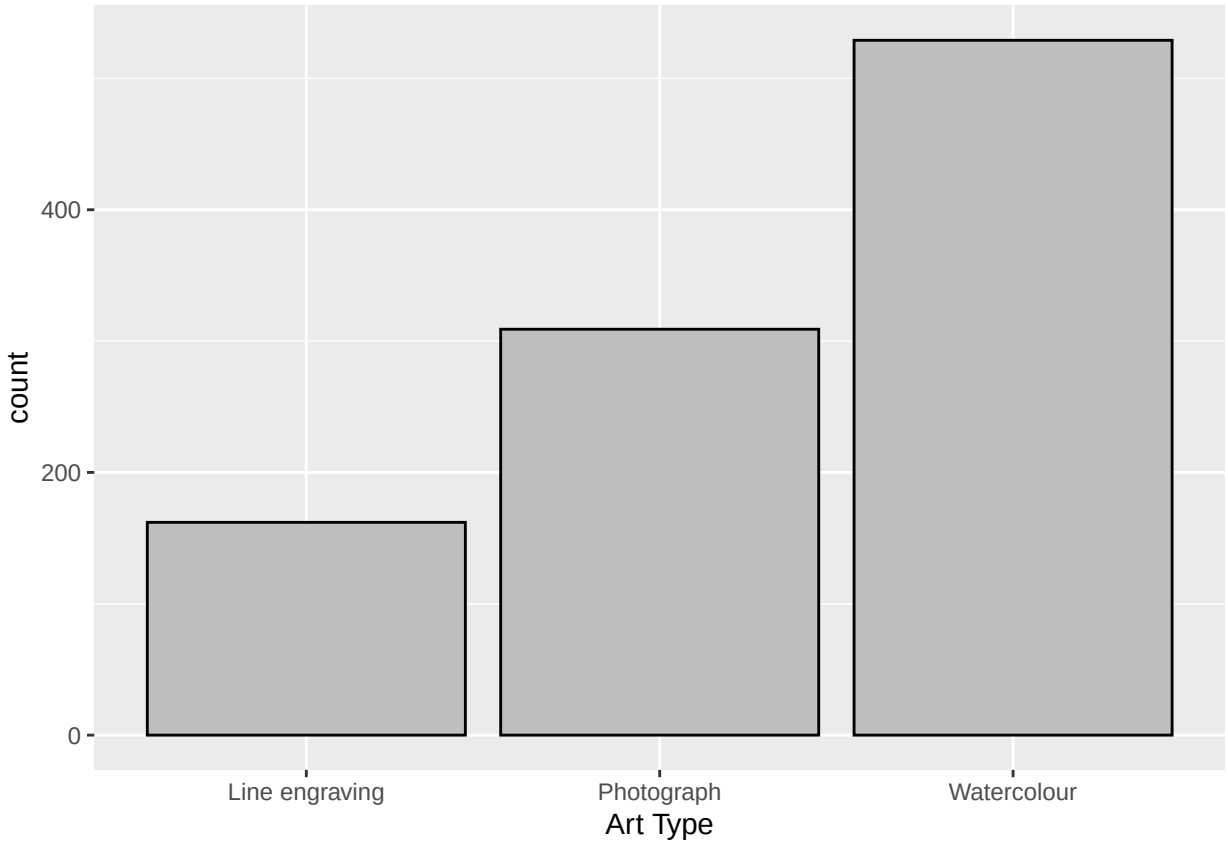
Justify your answer above in a few sentences. Hint: Don't forget to refer to the 3 aspects of quantitive distributions and comment on how each plot lets you visualize each aspect.

There are 3 aspects that we need to look at when judging these plots: shape, center, and spread. With only

3 bins, it is hard to determine the trend of the data and whether or not there's any peaks, and thus it's hard to see how concentrated/distributed data is from the center. With 150 bins, the chart looks like a series of jagged peaks. It is easier to see that there's 1 big peak on the left, but the (true) peak on the right is harder to see. With 30 bins however, it is quite easy to see the shape of the data and the number of peaks.

(c) Construct a plot to visualize the distribution of the `type` variable. 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}`). From the choices below, select the best description for the distribution of the `type` variable.

```
artwork %>% ggplot(aes(x=type)) +
  geom_bar(color = "black",
           fill = "gray") +
  labs(x = "Art Type")
```




```
# Among the four descriptions below, decide which one is most accurate
# and replace NULL with "A", "B", "C" or "D" accordingly
Q1c <- "C"
```

A: There are about twice as many line engravings than photographs, and almost twice as many photographs than watercolours. The most common type of artwork in this sample of 1000 is watercolour.

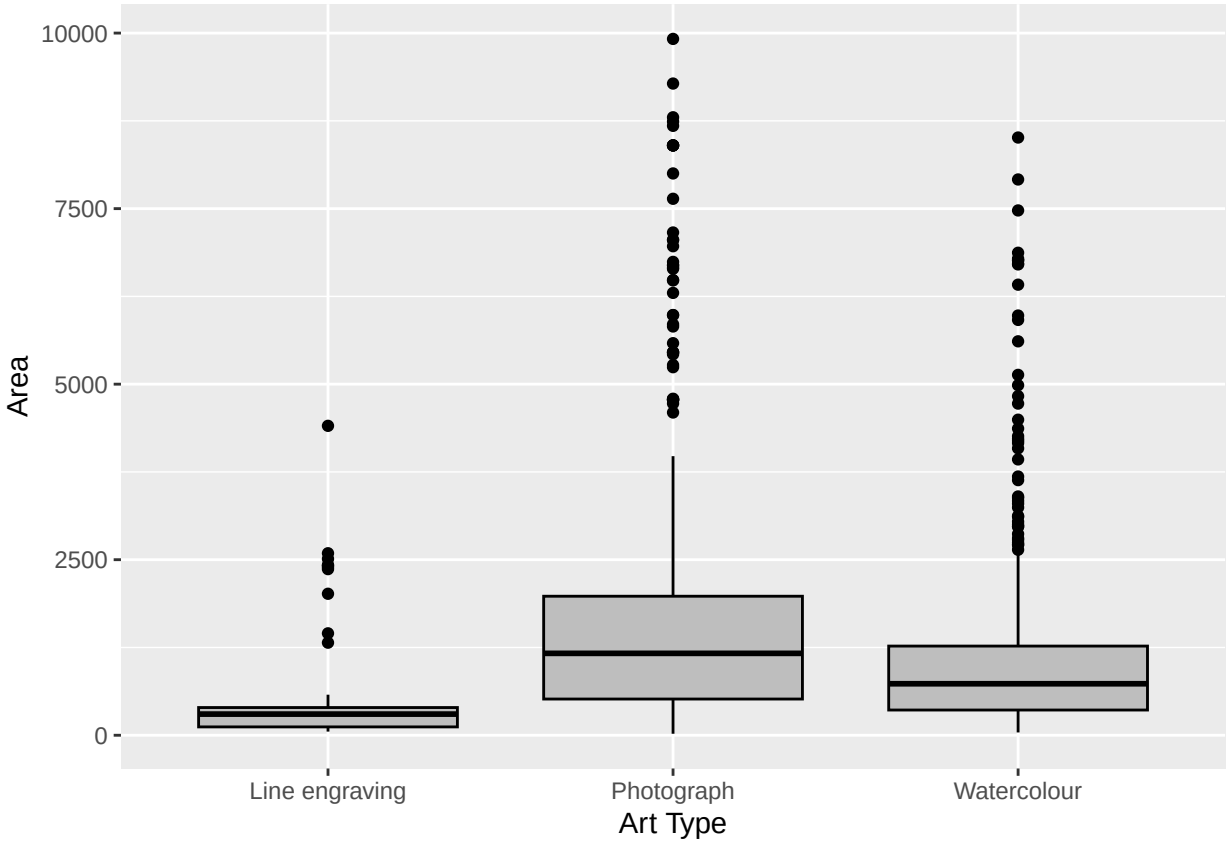
B: The distribution of artwork type is not symmetrical. The center of the distribution is photograph because it appears in the middle.

C: There are about twice as many photographs than line engravings, and almost twice as many watercolours than photographs. The most common type of artwork in this sample of 1000 is watercolour.

D: The distribution of artwork type is left skewed because there are fewer line engravings than photographs and watercolours. The most common type of artwork in this sample is waterclour.

(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.

```
artwork %>% ggplot(aes(x = type, y = area)) +
  geom_boxplot(color = "black",
               fill = "gray") +
  labs(y = "Area", x = "Art Type")
```

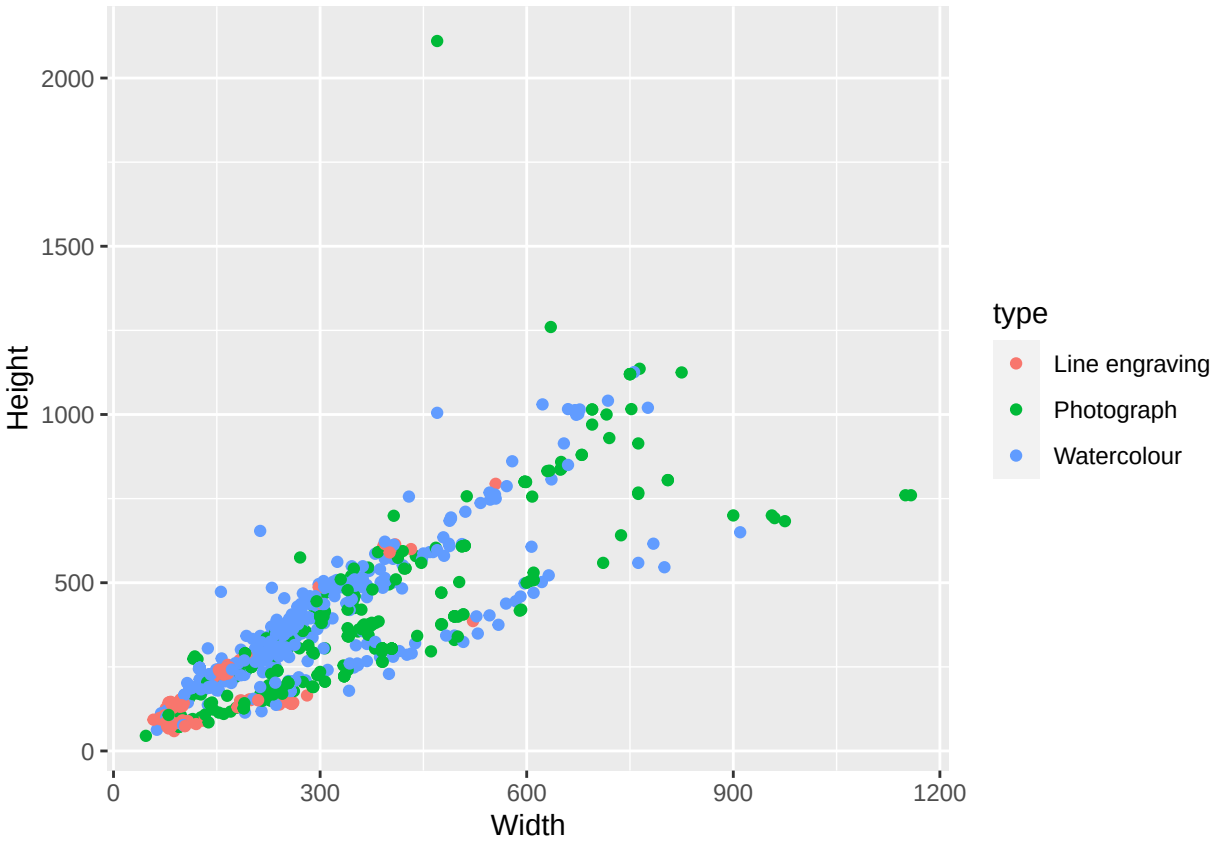


Write 3-4 sentences comparing these distributions.

The median area of the photograph is higher than the median area of watercolor paintings which is higher than the median area of line engraving art pieces. Similarly, the IQR follows the same pattern: photograph's IQR > watercolor's IQR > line engraving's IQR. For all three distributions, they skewed right with many outliers higher than the median.

(e) Make a scatterplot showing the `width` versus `height` for each type of artwork (`type`) using the `colour` and/or `facet_wrap()` option; make sure to specify meaningful axis labels where appropriate.

```
artwork %>% ggplot(aes(x = width, y = height, color=type)) +
  geom_point() +
  labs(x = "Width", y = "Height")
```



Write 1-3 sentences describing any trends you see.

There is a pretty strong positive correlation between width and height. The relation seems to be linear.

[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.

```
glimpse(births14)
```

```
## Rows: 1,000
## Columns: 13
## $ fage           <int> 34, 36, 37, NA, 32, 32, 37, 29, 30, 29, 30, 34, 28, 28,~
## $ mage           <dbl> 34, 31, 36, 16, 31, 26, 36, 24, 32, 26, 34, 27, 22, 31,~
## $ mature         <chr> "younger mom", "younger mom", "mature mom", "younger mo~
## $ weeks          <dbl> 37, 41, 37, 38, 36, 39, 36, 40, 39, 39, 42, 40, 40, 39,~
## $ premie         <chr> "full term", "full term", "full term", "full term", "pr~
## $ visits         <dbl> 14, 12, 10, NA, 12, 14, 10, 13, 15, 11, 14, 16, 20, 15,~
## $ gained         <dbl> 28, 41, 28, 29, 48, 45, 20, 65, 25, 22, 40, 30, 31, NA,~
## $ weight         <dbl> 6.96, 8.86, 7.51, 6.19, 6.75, 6.69, 6.13, 6.74, 8.94, 9~
## $ lowbirthweight <chr> "not low", "not low", "not low", "not low", "not low", ~
## $ sex            <chr> "male", "female", "female", "male", "female", "female",~
## $ habit          <chr> "nonsmoker", "nonsmoker", "nonsmoker", "nonsmoker", "no~
## $ marital        <chr> "married", "married", "married", "not married", "marrie~
## $ whitemom       <chr> "white", "white", "not white", "white", "white", "white~
```

(a) Type `?births14` in the R console to answer the questions below. Make sure not to change the variable names, and replace each `NULL` with your answer.

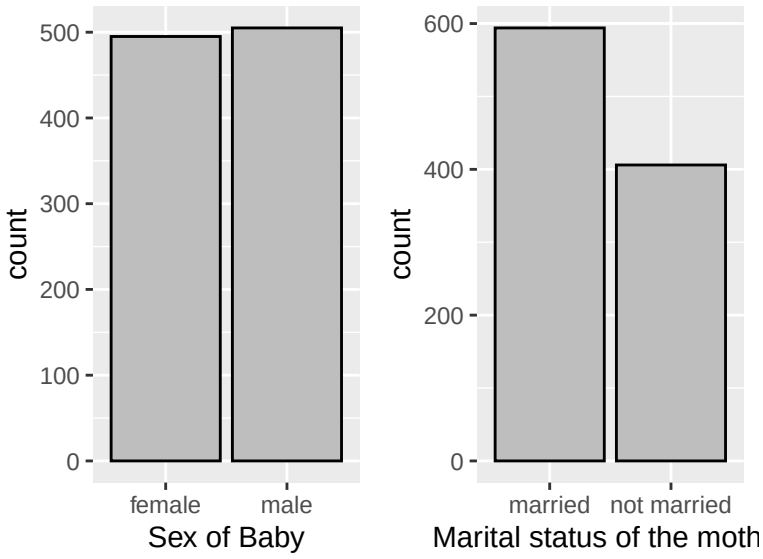
```
# In what year were these data collected?
Q2a_year <- 2014

# What is the name of the variable with the following definition:
# "Weight gained by mother during pregnancy in pounds."
# Make sure that your answer is in quotation marks here, and
# remember that R is case sensitive!
Q2a_variable <- "gained"
```

(b) Before even doing any analysis, it is good to consider whether the data in question might have any strong bias that could impact any conclusions you may draw. Based on the information contained documentation (particularly in `Description` and `Source`), please write 3-4 sentences either (1) arguing the dataset should be generally unbiased and representative of all births in North Carolina or, if not, (2) what potential issues there might be with the data.

According to the source of this data set (for ?births14), the original set contains every birth recorded in 2014, and thus it is not merely representative, but rather is the source. Assuming that openintro picked 1000 names without replacement by random sample, then this set should be representative of all births in North Carolina in 2014. However if we want to generalize this to all births in North Carolina, we need to consider other factors, i.e. was there anything in 2014/2013 that discouraged people having children. In our case, an ebola outbreak happened in 2014 which could have discouraged children births, and thus makes this year irregular compared to other years.

(c) 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.

```
gg1 <- births14 %>% ggplot(aes(x=sex)) +
  geom_bar(color = "black",
           fill = "gray") +
  labs(x = "Sex of Baby")

gg2 <- births14 %>% ggplot(aes(x=marital)) +
  geom_bar(color = "black",
           fill = "gray") +
  labs(x = "Marital status of the mother")

library(gridExtra)
grid.arrange(gg1, gg2, nrow=1, ncol=2)
```

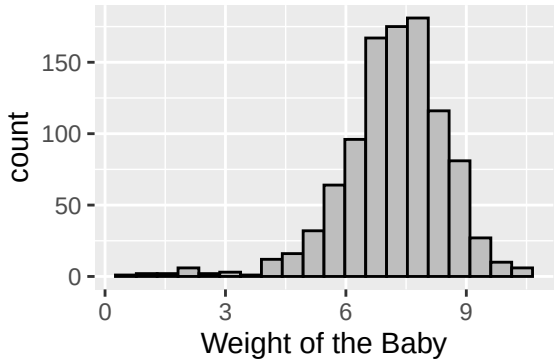


Baby sex variable is a nominal variable. There were slightly more males born than females, however this difference does not seem signifigant enough.

Marital status of the mother is also a nominal variable. There are significantly more married women giving birth than not married and its ratio is about 3:2.

(d) Consider the variable `weight`. Replace NULL below with the type of this variable (either "continuous numerical", "discrete numerical", "nominal categorical", "nominal ordinal", or "binary"). Create a plot to visualize the distribution and write 2-3 sentences describing the distribution.

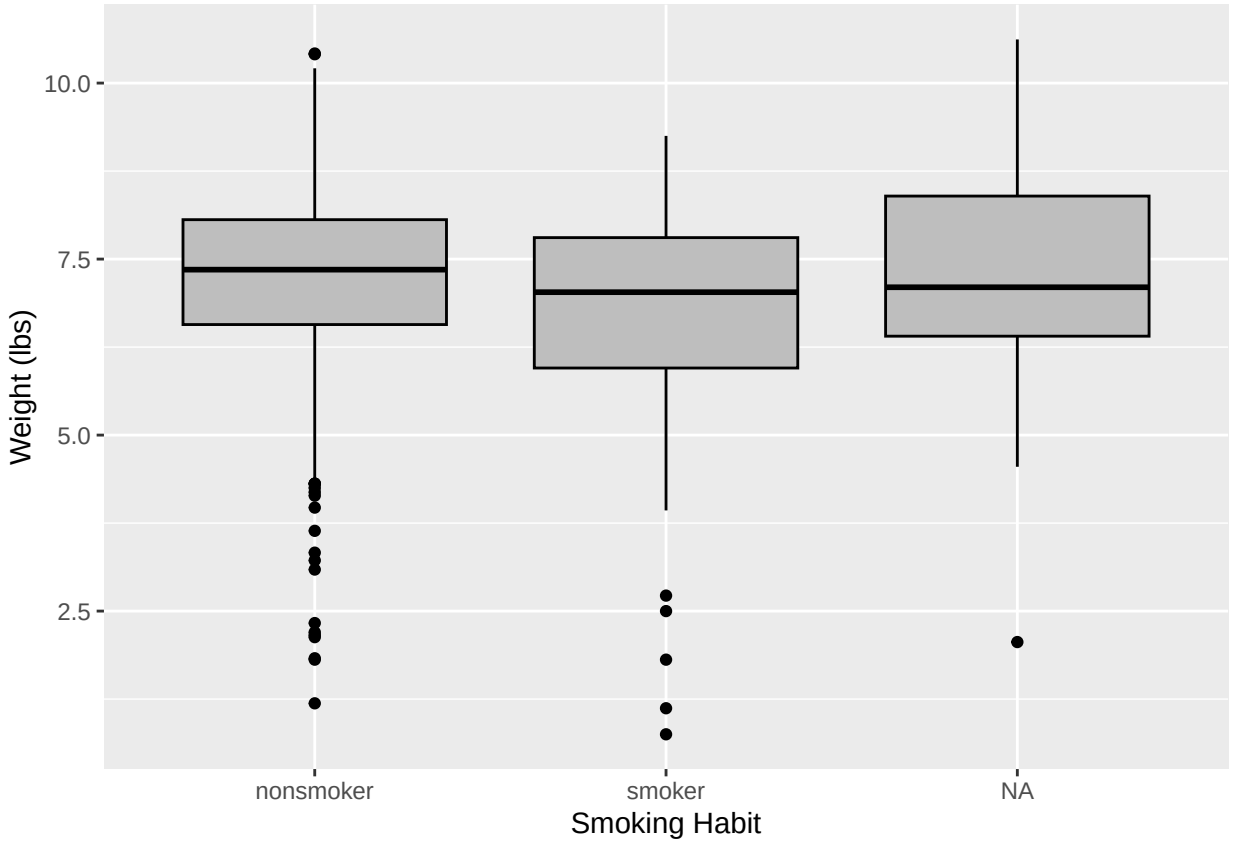
```
# What is the type of this variable?
# Hint: Replace NULL below with the correct option above, making sure to copy exactly and to include qu
Q2d <- "continuous numerical"

# Create your plot below
births14 %>% ggplot(aes(x = weight)) +
  geom_histogram(color = "black",
                 fill = "gray",
                 bins = 20) +
  labs(x = "Weight of the Baby")
```



This distribution is unimodal and approximately normal, with a center at around 7.5 lbs. There is a slight skew left however it isn't very major.

(e) 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.

```
# Create your plot below
births14 %>% ggplot(aes(x = habit, y = weight)) +
  geom_boxplot(color = "black",
               fill = "gray") +
  labs(y = "Weight (lbs)", x = "Smoking Habit")
```

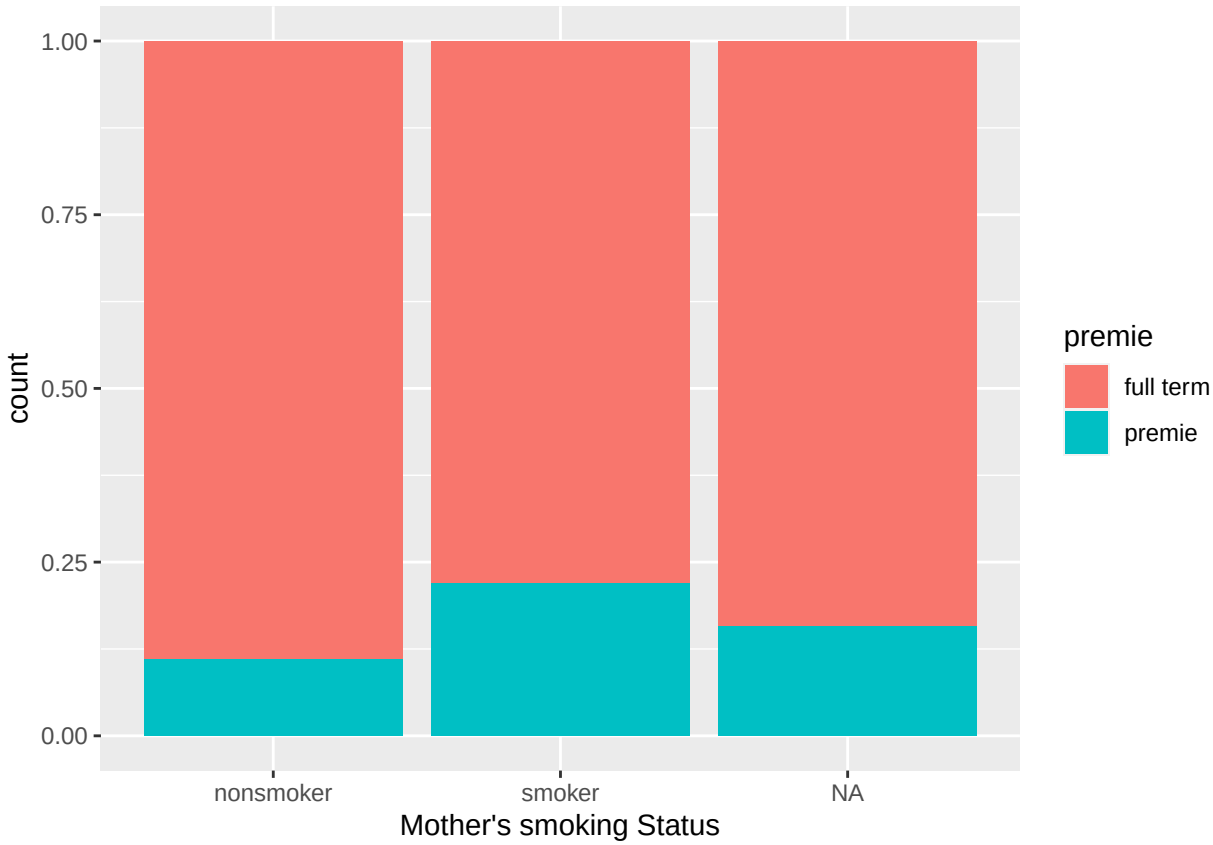


(f) Based on your plot, would you conclude that there is a strong association, weak association, or no association between birth weight (`weight`) and mother's smoking status (`habit`)? Why or why not?

There does not seem to be a strong association between birth weight and the mother's smoking status. The medians of each are roughly the same (at 7.5 lbs) and the IQR seems to be the same as well.

(g) Sometimes, a potential association could be due factors that affect both variables. These are called *confounding variables* (we will discuss these in more detail later in the course). Below, we have included a plot showing the *proportion* of births that were classified as premature or full-term (`premie`) versus smoking status (`habit`).

```
# Proportion of premature births vs mother's smoking status
births14 %>% ggplot(aes(x = habit, fill=premie)) +
  geom_bar(position="fill") +
  labs(x="Mother's smoking Status")
```



Does the above plot suggest that a mother's smoking habit could lead to premature births? Why or why not?

Yes, this plot suggests that the mother's smoking habit can lead to premature births. Approximately 25% of mothers who smoked had premie births while only about 15% of mothers who didn't smoke had premie births.

(h) In general, commomn medical advice for women who are pregnant or expect to become pregnant is to stop smoking to protect the health of their baby. Based on your plot and answer above, pleased write 2-4 sentences discussing whether this advice seems justified or not.

Although smoking may not directly impact the weight of the baby, it does seem to have an effect on whether or not the babie was born premie. As such, the baby may suffer health complications for that. Thus, I do think that this advice is justified.

We hope this exercise helped to emphasize the importance of the ethical considerations we discussed in class when it comes to making, interpreting, and communicating data visualizations. Don't forget to review sample solutions once they are posted on Quercus.