STA130 Rstudio Homework

Problem Set 2

[Xuanqi Wei] ([1009353209]), with Josh Speagle & Scott Schwartz

Instructions

Complete the exercises in this `.Rmd` file and submit your `.Rmd` and knitted `.pdf` output through Quercus by 11:59 pm E.T. on Thursday, January 26.

Question 1: Data Types

Problem Set 1 included the following code:

```
my_answers <- c(r1,r2,c1,c2)
square_answers <- c(10,-1,3,12)
```

For the first three questions below, choose the correct answer from the following statements:

(1) A single value counting how many correct rows and columns you calculated.
(2) A numeric vector of the differences between the math square answers and your answers (should be all 0s if you got them all right).
(3) A character vector of 'TRUE' and 'FALSE', 'TRUE' for each answer that matches and 'FALSE' for any that don't.
(4) A logical vector of `TRUE` and `FALSE`, `TRUE` for each answer that matches and `FALSE` for any that don't.
(5) A single logical value `TRUE` or `FALSE`, `TRUE` if all the values match, `FALSE` if any of the values don't match.

(a) Which of the above best describes what `my_answers == square_answers` is?

> 4

(b) Which of the above best describes what `sum(my_answers == square_answers)` is?

> 1

(c) Which of the above best describes what `all(my_answers == square_answers)` is?

> 5

(d) What is the sequence of steps involved in getting the answer for `sum(c(TRUE,FALSE))`? What additional step is required to get the answer for `sum(my_answers == square_answers)`?

> For getting the answer for `sum(c(TRUE,FALSE))`, firstly, the TRUE and FALSE logial values are turned into the numbers 1 and 0, known as coercion. Secondly, the numbers are summed.
>
> For getting the answer for `sum(my_answers == square_answers)`, it doesn't include the summation and cercion. The vector are compared in an elementwise manner and entries that match are assigned TRUE and those don't match are assigned FALSE.

*Hint: The `sum` function works only on `numeric` data types and does not itself directly know anything about `logical` data types. How might this relate to the concept of **`coercion`**?*

Question 2: Super Bowl Ads

The data for this question will be based on a sample of Super Bowl ads. This is stored in the file `superbowl_ads.csv` in the same directory as this file and includes the following variables:

- `year` (double) Superbowl year
- `brand` (character) Brand for commercial
- `funny` (logical) Contains humor
- `show_product_quickly` (logical)) Shows product quickly
- `celebrity` (logical) Contains celebrity
- `danger` (logical) Contains danger
- `view_count` (double) Youtube view count
- `like_count` (double) Youtube like count
- `dislike_count` (double) Youtube dislike count
- `superbowl_ads_dot_com_url` (character) Superbowl ad URL

This data was posted on GitHub by the data-oriented reporting outlet FiveThirtyEight and subsequently featured on Tidy Tuesday. For more information, see the above links.

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

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

```
superbowl %>% glimpse()
```

```
## Rows: 211
## Columns: 11
## $ ID                        <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 1~
## $ year                      <dbl> 2018, 2020, 2006, 2018, 2003, 2020, 2020, 20~
## $ brand                     <chr> "Toyota", "Bud Light", "Bud Light", "Hynudai~
## $ funny                     <lgl> FALSE, TRUE, TRUE, FALSE, TRUE, TRUE, TRUE, ~
## $ show_product_quickly      <lgl> FALSE, TRUE, FALSE, TRUE, TRUE, TRUE, FALSE,~
## $ danger                    <lgl> FALSE, TRUE, TRUE, FALSE, TRUE, TRUE, FALSE,~
## $ celebrity                 <lgl> FALSE, TRUE, FALSE, FALSE, FALSE, TRUE, TRUE~
## $ view_count                <dbl> 173929, 47752, 142310, 198, 13741, 23636, 30~
## $ like_count                <dbl> 1233, 485, 129, 2, 20, 115, 1470, 78, 342, 7~
## $ dislike_count             <dbl> 38, 14, 15, 0, 3, 11, 384, 6, 7, 0, 14, 0, 2~
## $ superbowl_ads_dot_com_url <chr> "https://superbowl-ads.com/good-odds-toyota/~
```

There are 211 rows and 11 columns. Therefore, it includes 211 observations and 11 variables.

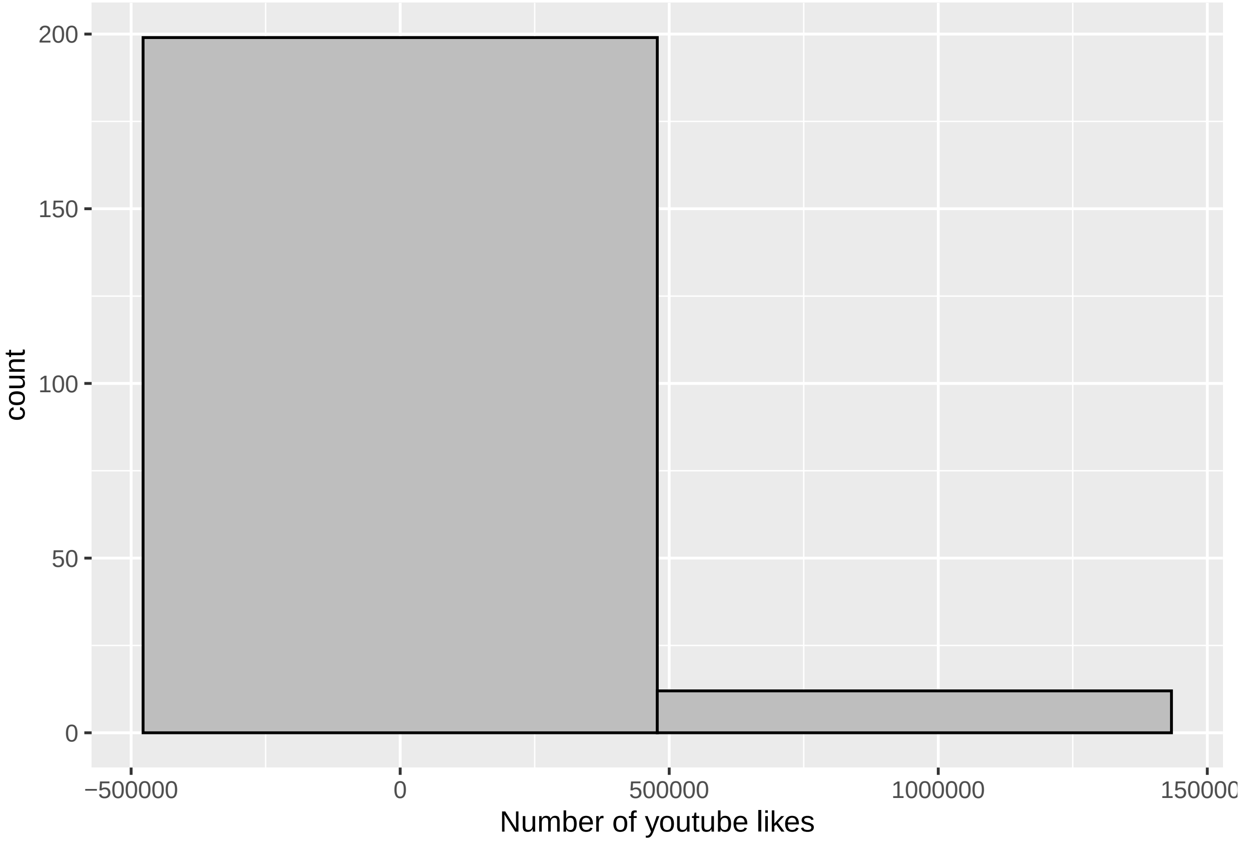
(b) Explore the distribution of `view_count` using a histogram with 2 bins, a histogram with 8 bins, and a histogram with 50 bins (3 histograms total). Make sure to specify meaningful axis labels where appropriate.

```
hist1 <- ggplot(data=superbowl, aes(x=view_count)) +
  geom_histogram(color='black', fill='gray', bins=2) +
  labs(x='Number of youtube likes')
```

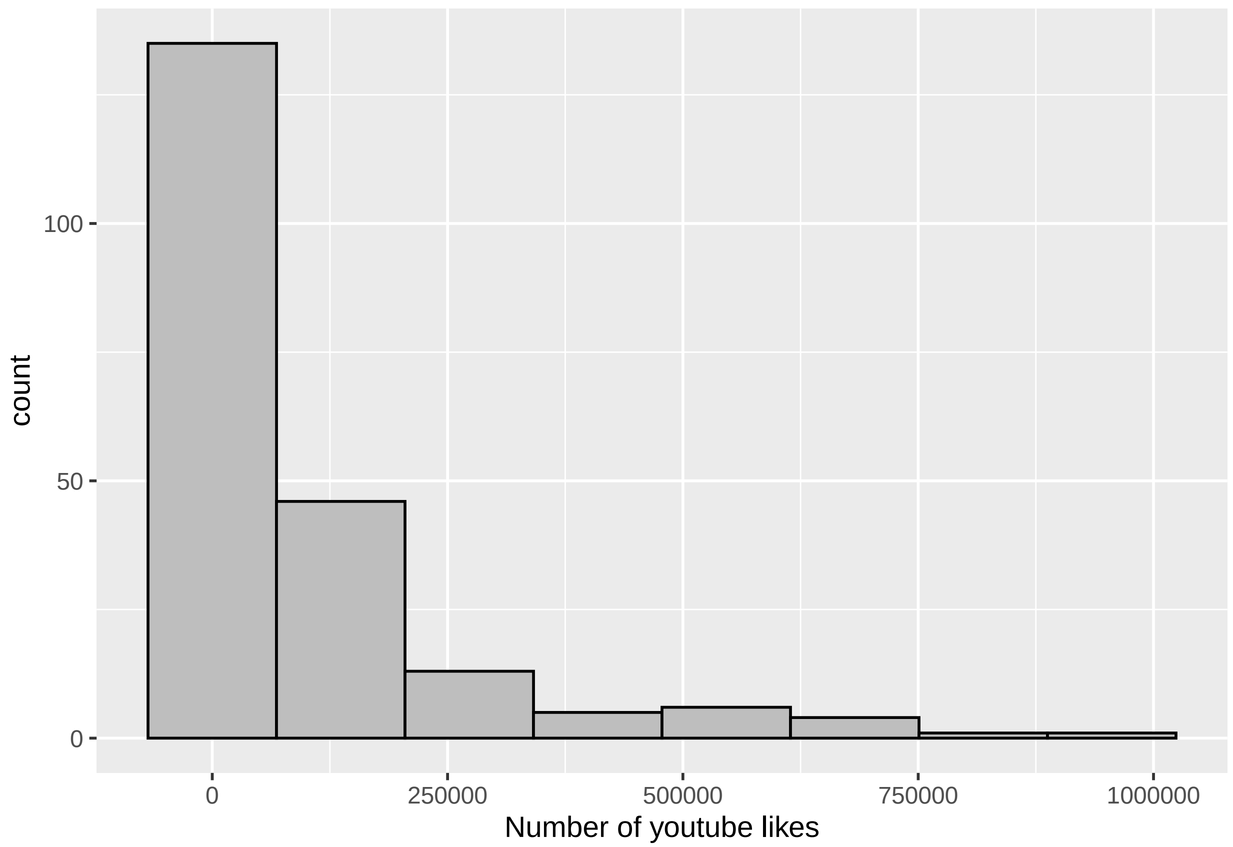
```
# Or your can put different plots in separate code chunks
hist2 <- ggplot(data=superbowl, aes(x=view_count)) +
  geom_histogram(color='black', fill='gray', bins=8) +
  labs(x='Number of youtube likes')
```

```
# Feel free to add or remove code chunks as desired
hist3 <- ggplot(data=superbowl, aes(x=view_count)) +
  geom_histogram(color='black', fill='gray', bins=50) +
  labs(x='Number of youtube likes')
```

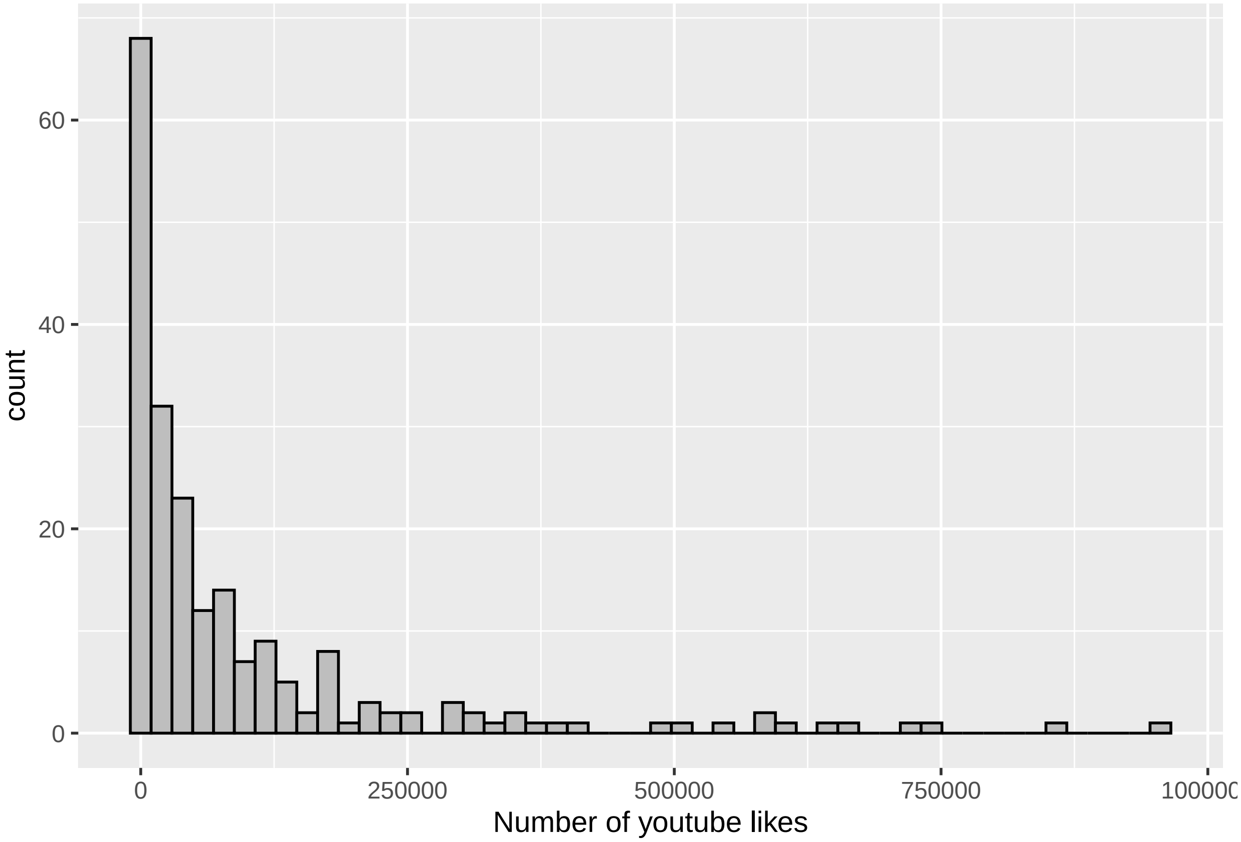
```
hist1
```



```
hist2
```


100
50
0
count
0
250000
500000
750000
1000000
Number of youtube likes

```
hist3
```



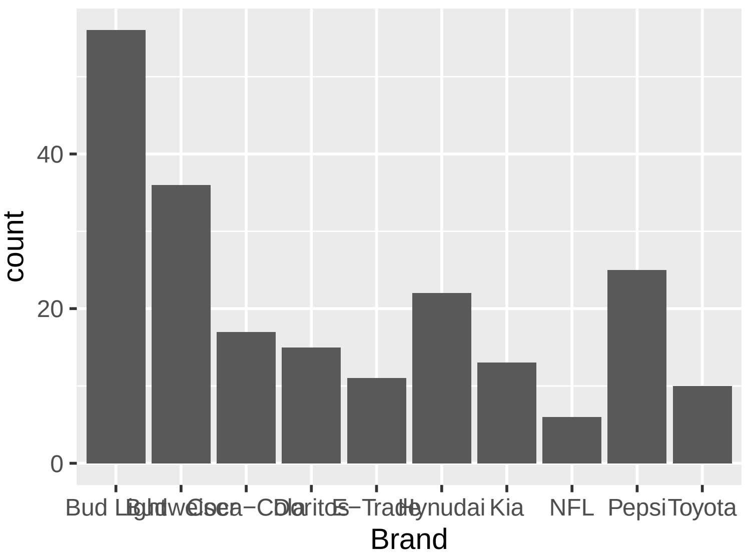
Which of these histograms is most appropriate to describe the distribution of `view_count`? Why? Write a few sentences describing the distribution based on the histogram you chose as most appropriate.

The histogram with 8 bins is most appropriate to describe the distribution of 'view_count' Firstly, based on this histogram, we can get the range of number of youtube likes is between 0 and 1,000,000. Those two numbers are eaily to be obtained by readers. Secondly, it's obvious that this graph is positive skewed which means the mean is larger than the median, poviding an important information about the number of youtube likes.

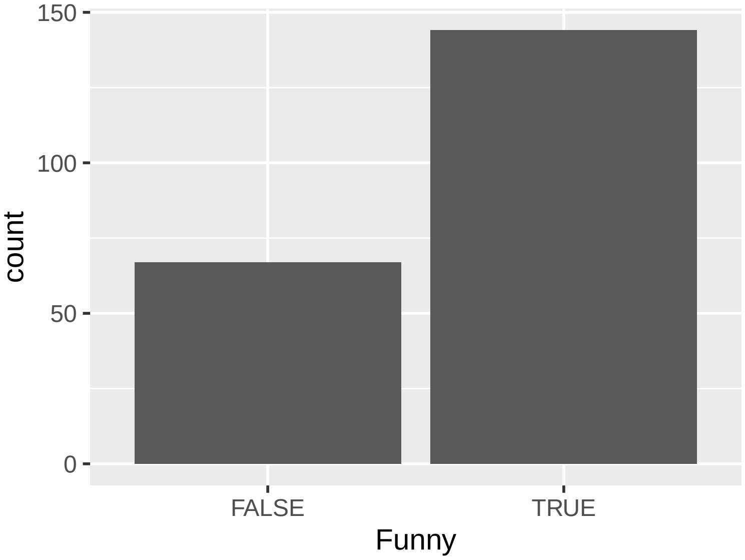
(c) Construct two plots (your choice) to visualize the distribution of `brand` and one of the following other categorical variables from the superbowl ads data: `funny`, `danger`, or `celebrity`. Make sure to specify meaningful axis labels where appropriate.

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 larger (e.g., {r, fig.height=15, fig.width=5}).

```
# One reason to use use different code chunks for different figures is
# to assign different figure aspect ratio controls to different figure
ggplot(data=superbowl, aes(x=brand)) +
  geom_bar() +
  labs(x='Brand')
```



```r
ggplot(data=superbowl, aes(x=funny)) +
  geom_bar() +
  labs(x='Funny')
```

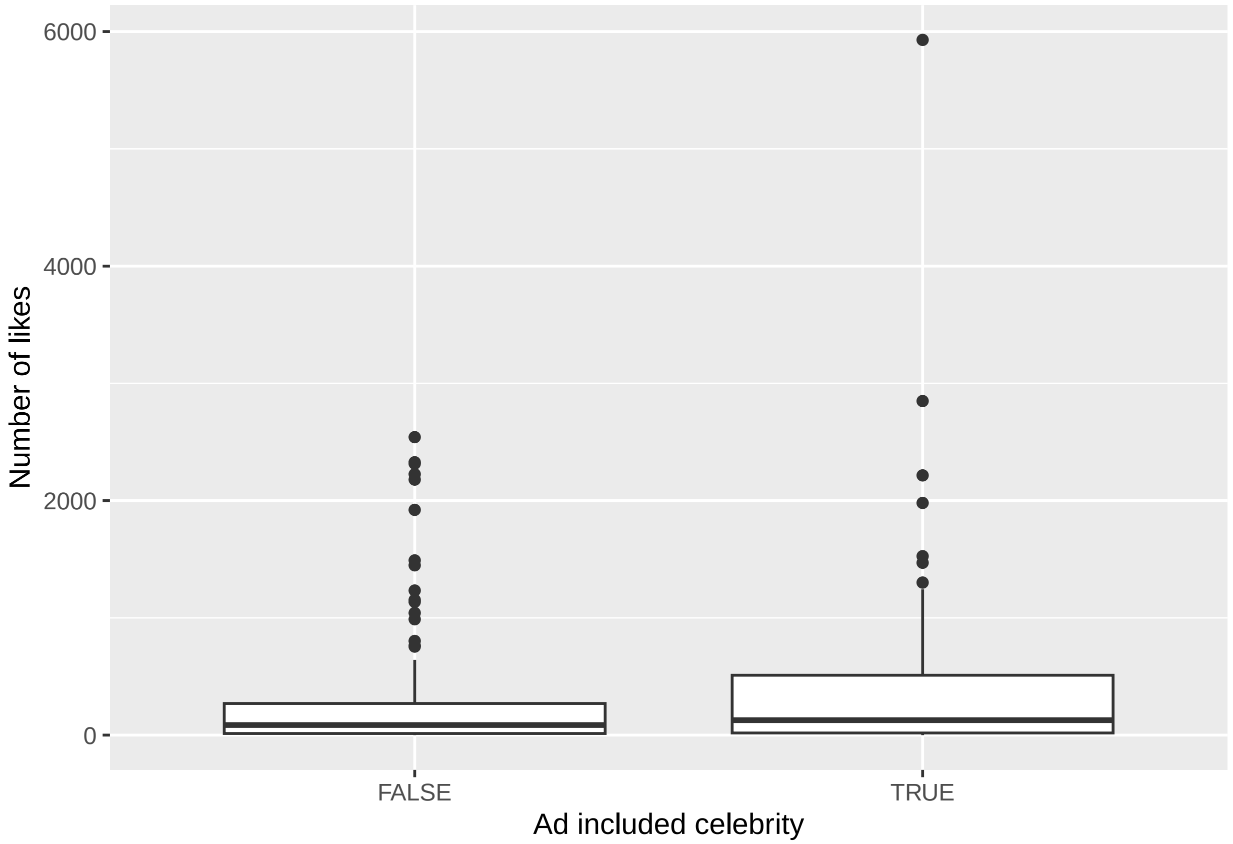


Describe the distribution in 1-2 sentences.

> According to the diagram, the most popular brand is BudLight, which knows aprroximately 57 ads. The second and third are Budweiser and Pepsi respectively. To be pointed out, NFL seems the most not popular brand. By the way, there are more 'Funny' superbowl ads than not 'Funny' ones. y

(d) Construct a joint set of two boxplots showing visual summaries of the distribution of number of likes (`like_count`) depending on whether ads included a celebrity or not (`celebrity`). Make sure to specify meaningful axis labels where appropriate.

```r
# This should be a single plot, NOT TWO!
# Boxplots can be put in the same plot!
superbowl %>% ggplot(aes(x=celebrity, y=like_count)) + geom_boxplot() +
  labs(x='Ad included celebrity', y='Number of likes')
```



Write 2-4 sentences comparing these distributions.

> It's apparent that when the ads include celebrity, the number of likes is much more than those don't include, which is acceptable due to most of the fans of every celebrity are happy to give a like to the ads. Moreover, from the graph, we can obtain that the distributino for the number of likes when the ads don't include a celebrity is within a much smaller range of number of likes than the ones that include the celebrity.

Question 3: Births and Smoking

The `births` data set is part of the `openintro` package. It consists of random sample of 100 births for babies in North Carolina where the mother was not a smoker and another 50 where the mother was a smoker. The code below loads the required libraries for this question and provides a glimpse of the `births` data frame.

Hint: Type `?births` in the R console for more information about the data.

```
## Loading required package: airports
```

```
## Loading required package: cherryblossom
```

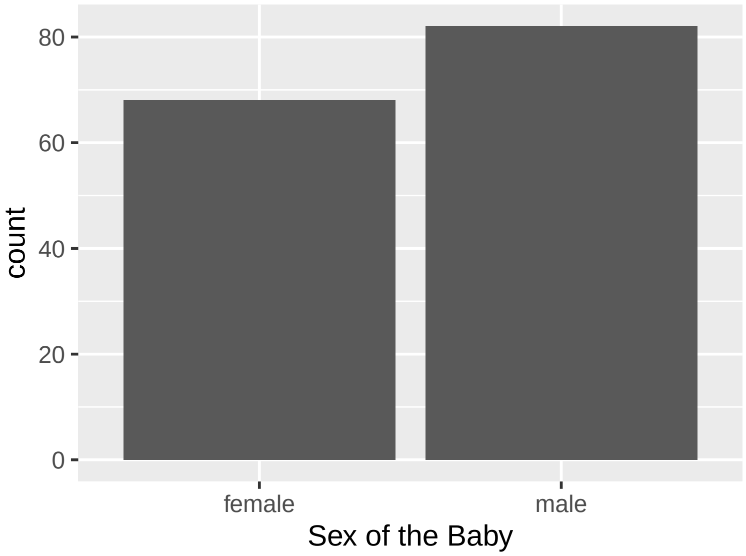
```
## Loading required package: usdata
```

```
births %>% glimpse()
```

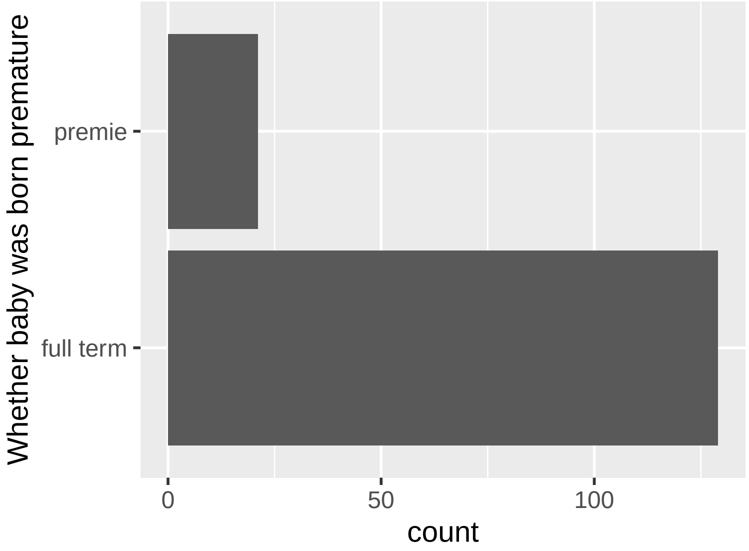
```
## Rows: 150
## Columns: 9
## $ f_age     <int> 31, 34, 36, 41, 42, 37, 35, 28, 22, 36, 27, 35, 25, 36, 27, ~
## $ m_age     <int> 30, 36, 35, 40, 37, 28, 35, 21, 20, 25, 19, 34, 19, 33, 27, ~
## $ weeks     <int> 39, 39, 40, 40, 40, 40, 28, 35, 32, 40, 32, 40, 41, 38, 39, ~
## $ premature <fct> full term, full term, full term, full term, full term, full ~
## $ visits    <int> 13, 5, 12, 13, NA, 12, 6, 9, 5, 13, 5, 15, 13, 10, 11, 13, 1~
## $ gained    <int> 1, 35, 29, 30, 10, 35, 29, 15, 40, 34, 32, 20, 47, 20, 5, 22~
## $ weight    <dbl> 6.88, 7.69, 8.88, 9.00, 7.94, 8.25, 1.63, 5.50, 2.69, 8.75, ~
## $ sex_baby  <fct> male, male, male, female, male, male, female, female, male, ~
## $ smoke     <fct> smoker, nonsmoker, nonsmoker, nonsmoker, nonsmoker, smoker, ~
```

(a) Choose two categorical variables from the `births` data set and plot the distribution of each one in separate plots using the visualization method of your choice.

```
ggplot(data=births) + aes(x=sex_baby) +
  geom_bar() + labs(x='Sex of the Baby')
```



```
ggplot(data = births) + aes(x=premature) +
  geom_bar() + coord_flip() +
  labs(x='Whether baby was born premature')
```

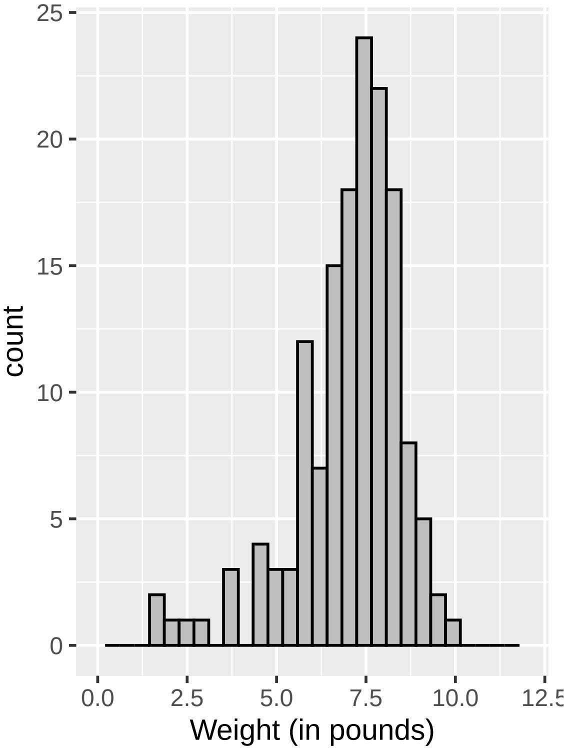


Identify whether each of these variables is a nominal or ordinal categorical variable and write one or two sentences interpreting each plot.

> Both of the variables, 'Sex of the baby' and 'Prematurity', are nominal categorical variable. The first graph reveals that the number of male and female babies are about to be the same. The second graph shows that most of the babies were born not prematurely.

(b) Choose a quantitative variable from the `births` data set and plot its distribution using the visualization method of your choice.

```
ggplot(data=births) + aes(x=weight) +
  geom_histogram(fill = 'grey', colour = 'black', bias = 25) + xlim(0, 12) +
  labs(x='Weight (in pounds)')
```

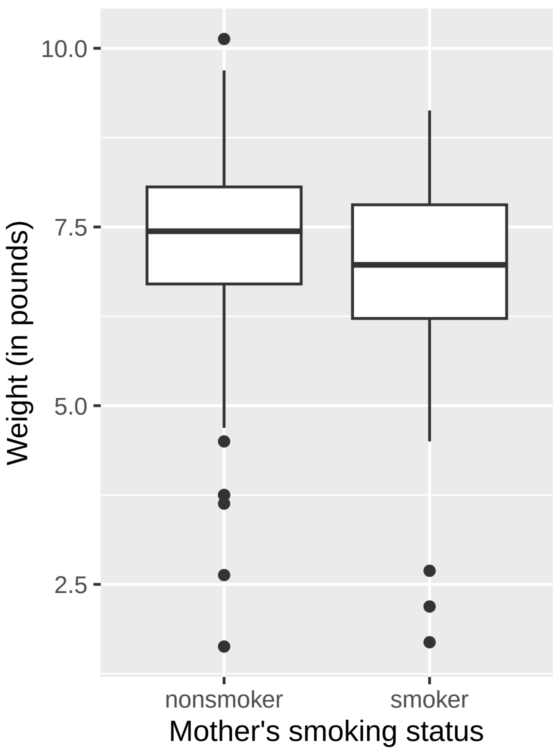


Identify whether the variable you selected is continuous or discrete, and write 2-3 sentences describing the distribution.

> The variable I selected is countinous. According to the graph, the distriution of babies' birth weight is a little bit left skewed from 1 pound to over 12 pounds but not very apparent. However, we can easily obtain the mode of babies' birth weight is 7.5 poinds.

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

```
births %>% ggplot(aes(x=smoke, y=weight)) +
  geom_boxplot() +
  labs(x="Mother's smoking status", y="Weight (in pounds)")
```

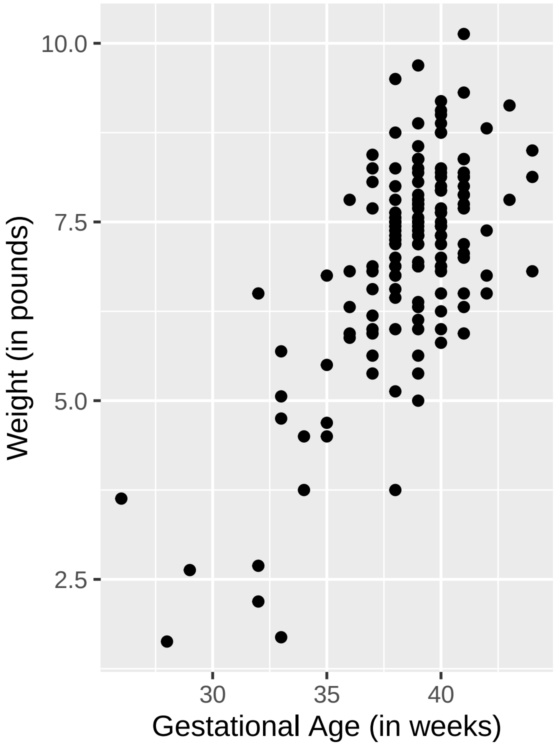


Write 2-3 sentences describing your observations and what this data might suggest (or not suggest) about the impact of smoking on birth weight.

> According to the graph, the majority of babies' birth weight of mother who is not a smoker is slightly higher than the mojority of babies's birth weight of mother who is a smoker. The distribution is about to be the same between these two categories. However, we can obtain that the top number occurs in the babies' birth weight of mother who is not a smoker which is over 10 pounds.

(d) Construct a plot that shows the relationship between birth weight (`weight`) and gestational age (`weeks`). Make sure to specify meaningful axis labels where appropriate.

```r
# To figure out how to do this google "ggplot2 scatter plot", or check out
# - https://ggplot2.tidyverse.org/#usage
# - https://ggplot2.tidyverse.org/#cheatsheet
# - https://github.com/rstudio/cheatsheets/blob/main/data-visualization-2.1.pdf
births %>% ggplot(aes(x=weeks, y=weight)) +
  geom_point() + labs(x='Gestational Age (in weeks)', y='Weight (in pounds)')
```



Write 2-3 sentences describing your observations and what this data might suggest (or not suggest) about the impact of gestational age on birth weight. Does this change the interpretation of your results above?

> According to the graph, if the Geostational Age is around 30 weeks, most of the babies' birth weight is range from 1.3 pounds to 3.75 pounds. If the geostatinoal age is around 40 weeks, the majority of the birth weight is around 7.5 pounds which is healthier.
>
> It doesn't change the interpretation of the results above. Most of the data are from the gestational age from 35 to 40 and the birth weight is around 7.5 which matches the result above. And both smoke or not smoke have the possibility that the baby is prematured.