STA130H1S – Fall 2022

Problem Set 2

() and STA130 Professors

Instructions

Complete the exercises in this .Rmd file and submit your .Rmd and .pdf output through Quercus on September 22 by 5:00 p.m. ET.

Question 1

The Week 1 Problem Set included the following code.

```
my_answers <- c(r1,r2,c1,c2)
square_answers <- c(10,-1,3,12)
sum(my_answers == square_answers)
#sum(c(TRUE,FALSE))
#c("pizza","cheese")
#sum(c(1,2))
```

For the first three questions below choose the correct answer from the following.

(A) A single value counting how many correct rows and columns you calculated.
(B) A numeric vector of the differences between the math square answers and your answers (should be all 0s if you got them all right).
(C) A character vector of 'TRUE' and 'FALSE', 'TRUE' for each answer that matches and 'FALSE' for any that don't.
(D) A logical vector of `TRUE` and `FALSE`, `TRUE` for each answer that matches and `FALSE` for any that don't.
(E) 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?

A character vector of 'TRUE' and 'FALSE', 'TRUE' for each answer that matches and 'FALSE' for any that don't

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

A single value counting how many correct rows and columns you calculated

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

A single logical value `TRUE` or `FALSE`, `TRUE` if all the values match, `FALSE` if any of the values don't match

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)`?

Firstly the c() function or the combine function combines the arguments entered in the brackets and returns a vector of TRUE and FALSE and in this code it makes it into a vector let's call it A where A[1] is TRUE and

A[2] is false . When u do the sum function R uses coercion and interrupts TRUE as 1 and False as 0 so the sum is 0+1 =1 so you get a value 1. For the sum(my_answers==square_answers) the function first checks if the values from my_answers equal is equal to square_answers to return of vector of TRUE and FALSE and then due to coercion R interrupts TRUE as 1 and False as 0 because it has to sum the vector of TRUE and FALSE and it ads TRUE+TRUE+TRUE+TRUE=1+1+1+1=4 and 4 is the solution

Hints

- Your answer should include the word ***coercion***.
-

Question 2

The data for this question will be based on a sample of Superbowl ads.

- The data 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")
glimpse(superbowl)
```

```
## 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/~
```

(a) Use the `glimpse()` function to view 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?

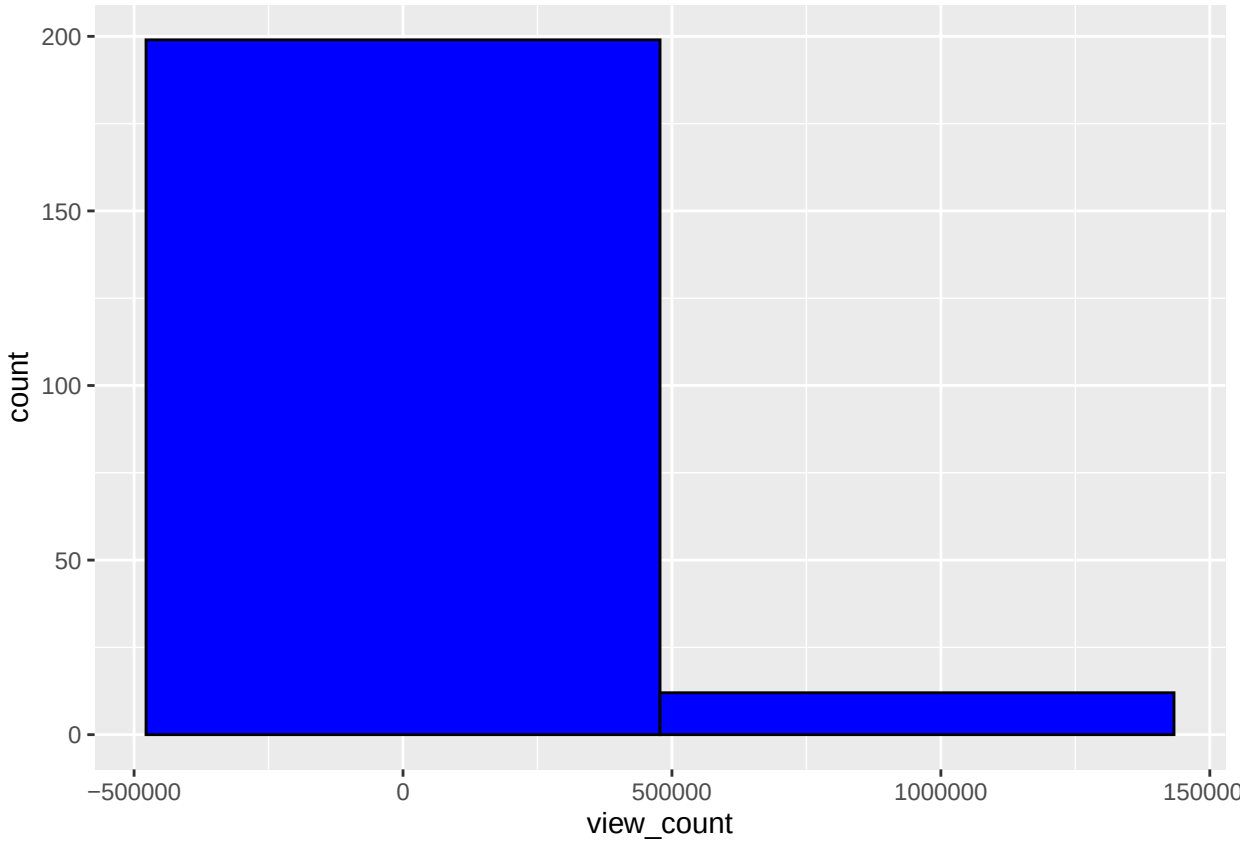
#ask prof if its fine if the glimpse function is on top

```
#done in the r chunk before
```

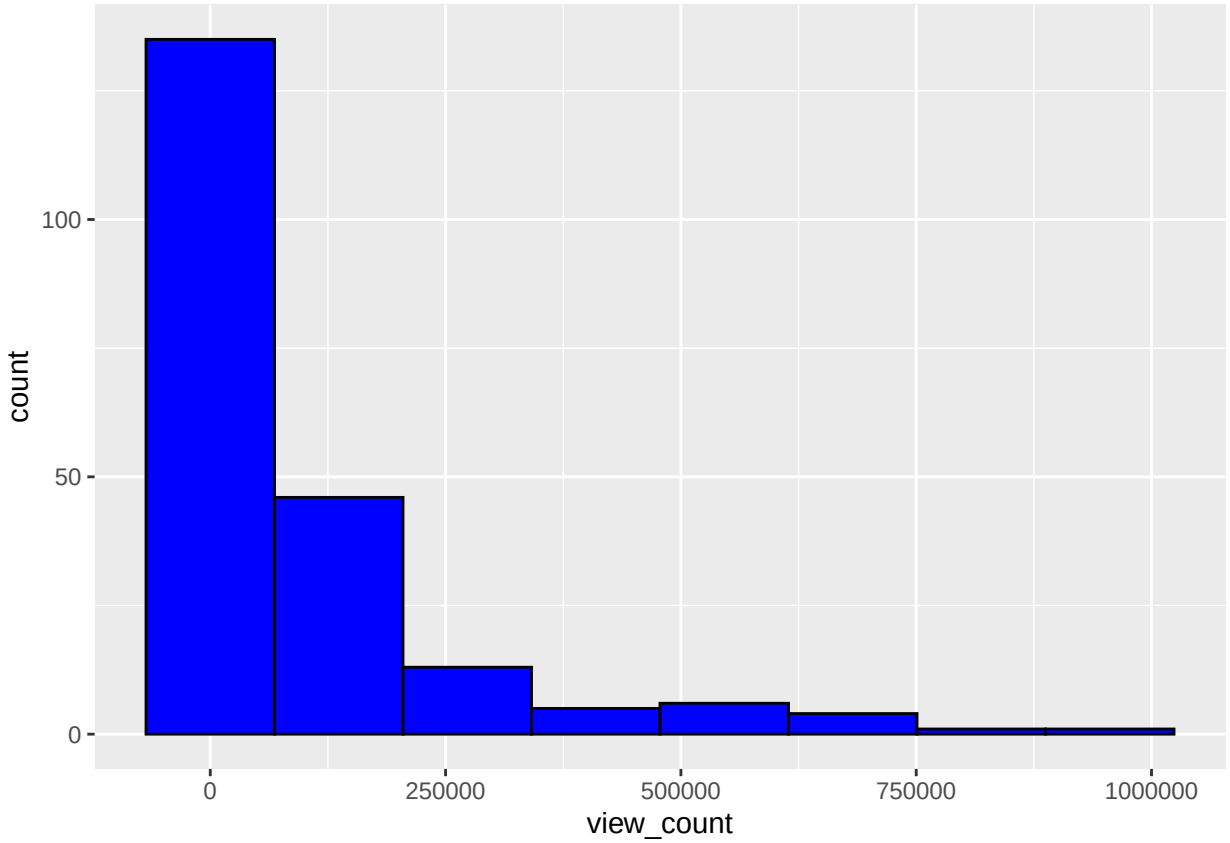
There are 211 rows and 11 columns and 211 observations

(b) Create 3 histograms to explore the distribution of `view_count`: (i) one with 2 bins, (ii) one with 8 bins, and (iii) one with 50 bins; make sure to specify meaningful axis labels where appropriate. 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.

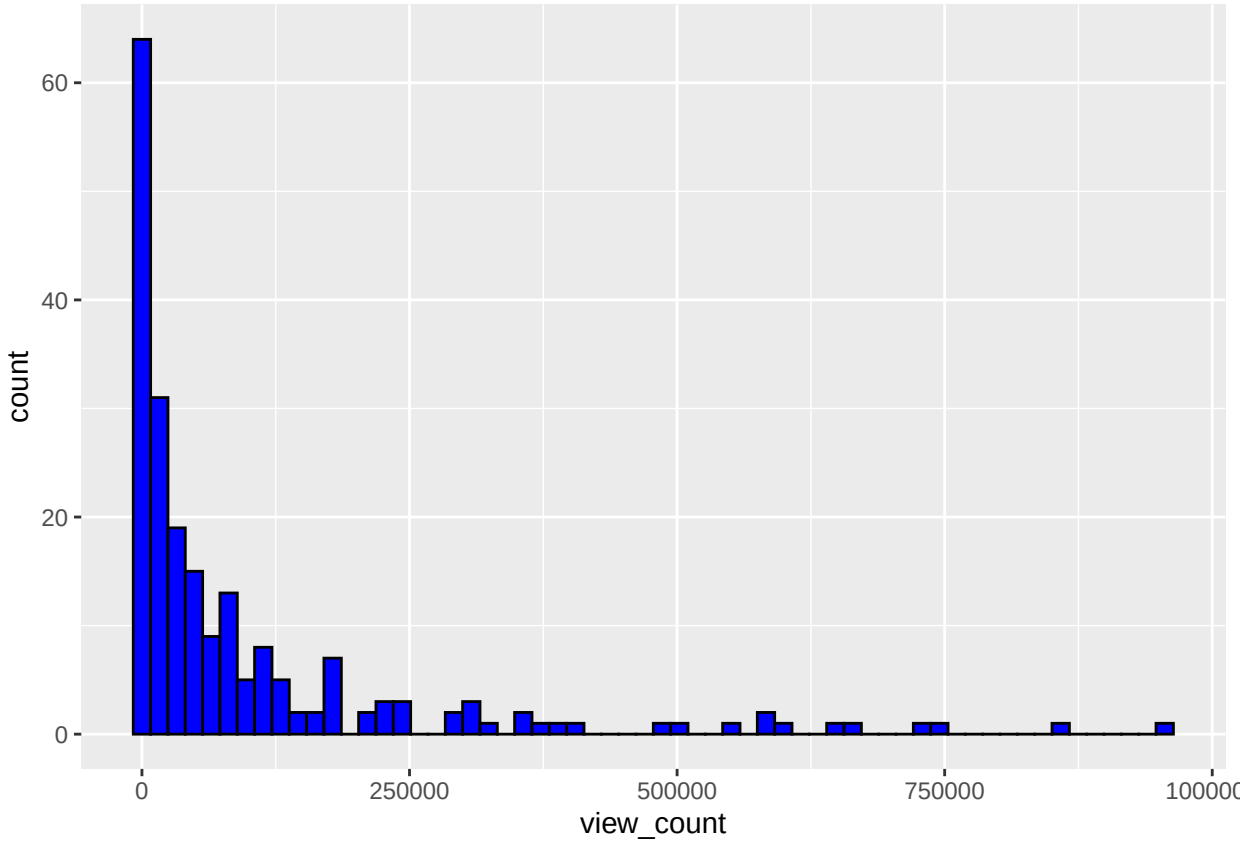
```
superbowl %>% ggplot(aes(x=view_count)) +
  geom_histogram(bins=2, color="black", fill="blue")
```




```
superbowl %>% ggplot(aes(x=view_count)) +
  geom_histogram(bins=8, color="black", fill="blue")
```




```
superbowl %>% ggplot(aes(x=view_count)) +
  geom_histogram(bins=60, color="black", fill="blue")
```

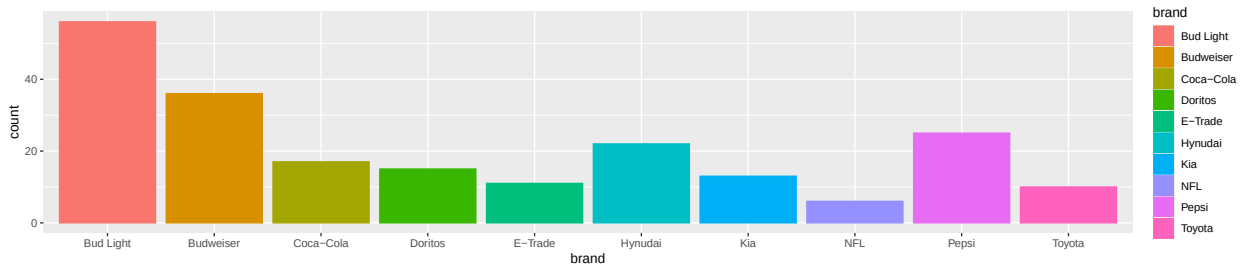


#Refine

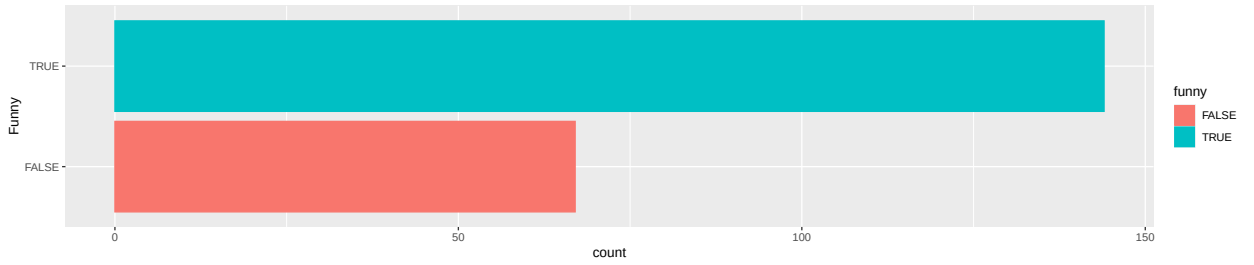
I believe that the most appropriate distribution is the one with 8 because it conveys the information needed about each data range without loosing its affect. The 60 bin histogram shows a precision which is to great which makes the graphics in specific areas empty which is ineffective and the one with 2 just not give misses to many of the trends and potential missing information. We can observe the fact that there histogram is skewed to the right and a majority of data is from the from right bar so we can say that the mode is on the right from -500000 to 500000. We can also say that the spread of this graph is high die to the range going from 0 to 1000000. We can also state using the one with 8 bins that the variance is present however other bin sizes may exagerate this fact or make it seem insignificant

(c) Construct two plots to visualize the distribution of `brand` and one of these other categorical variables: `funny`, `danger` or `celebrity` from the superbowl ads data 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})

```
superbowl %>% ggplot(aes(x=brand,color=brand,fill=brand)) +
  geom_bar()
```



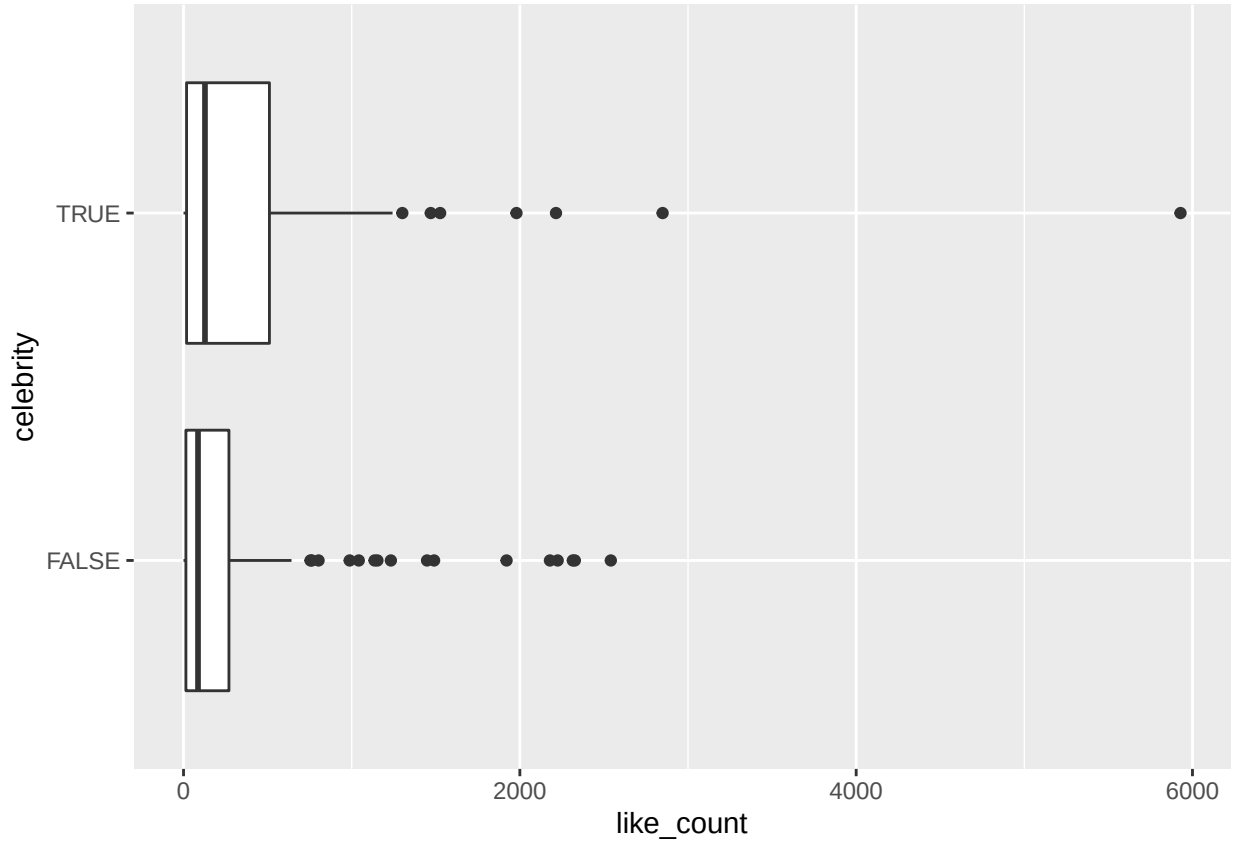

```
superbowl %>% ggplot(aes(x=funny,fill=funny,color=funny)) +
  geom_bar()+labs(x="Funny")+coord_flip()
```



The graph shows that the mode of the count is at Bud-light. This graph displays categorical data and quantifies the life count among different to display what brand is more liked and the second graph describes whether it was funny or not and counts it and appears to be skewed to the true side. The second graph may also talk more about the preferences of the count on how data is more from one side than the other

(d) Construct a set of two boxplots showing visual summaries of the distribution of number of likes (`like_count`) for whether ads included a celebrity or not (`celebrity`); make sure to specify meaningful axis labels where appropriate. Write 3-4 sentences comparing these distributions.

```
# This should be a single plot, NOT TWO... boxplots can be put in the same plot!
superbowl %>% ggplot(aes(x=like_count, y=celebrity)) + geom_boxplot()
```



The length of the box plot indicates sample variability so I can conclude that the variance of the true is greater than that of the false and due to the line in the middle of the box we can observe that both of the plots are centered to the left.The position of the box also indicates the skrew so the data is skrewed to the left. We may also state that the one on top had more outliers which may mean more data may be required to understand the outliers.

Question 3

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. Type `?births` 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 `births` data frame.

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

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

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

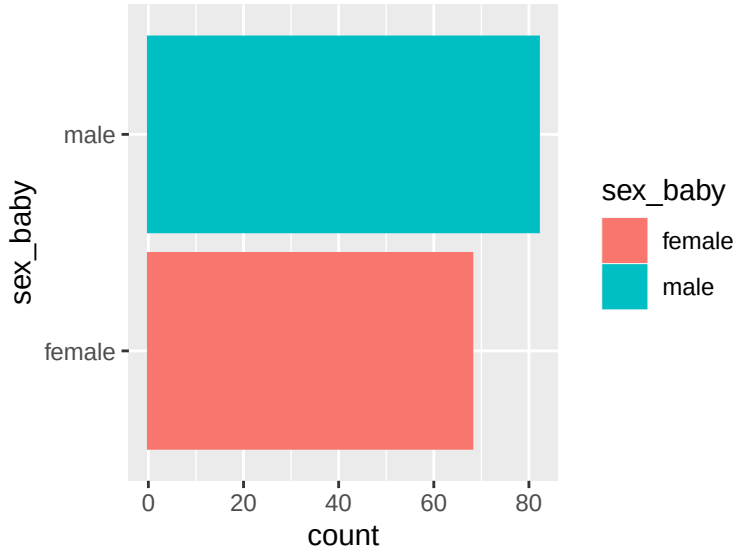
```
births %>% glimpse()
```

```
## 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~
```

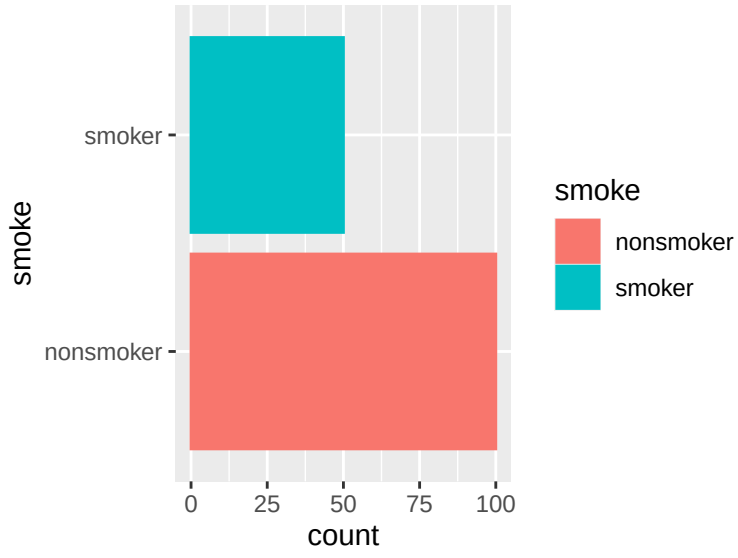
```
## $ weight    <dbl> 6.88, 7.69, 8.88, 9.00, 7.94, 8.25, 1.63, 5.50, 2.69, 8.75, ~
## $ sex_baby  <fct> male, male, male, female, male, male, female, female, male, ~
## $ smoke     <fct> smoker, nonsmoker, nonsmoker, nonsmoker, nonsmoker, smoker, ~
```

(a) Choose two categorical variables and plot the distribution of each one (in seperate plots). Identify whether each of these variables is a nominal or ordinal categorical variable. Write one or two sentences interpreting each plot.

```
births %>% ggplot(aes(x=sex_baby,color=sex_baby,fill=sex_baby)) +
  geom_bar()+coord_flip()
```




```
births %>% ggplot(aes(x=smoke,color=smoke,fill=smoke)) +
  geom_bar()+coord_flip()
```

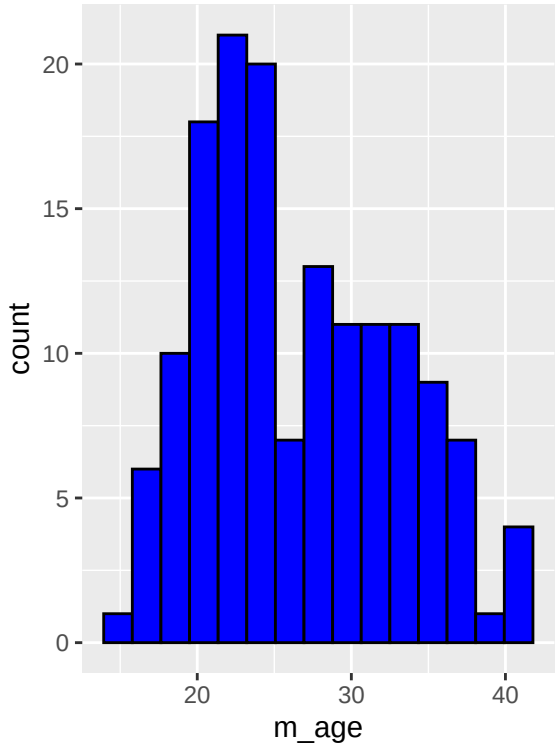


Both the variables in this case are nomimal.The first bar plot is skewed to the upper side in this case and the second plot is more skewed to the downwards direction.We can also concluded that the varraince of the first one is greater than that of the second one.This graph may also show that the data in relation to count

corresponds more with smokers in the second graph whereas the first graph is close in terms of distrution between men and woman.

(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.

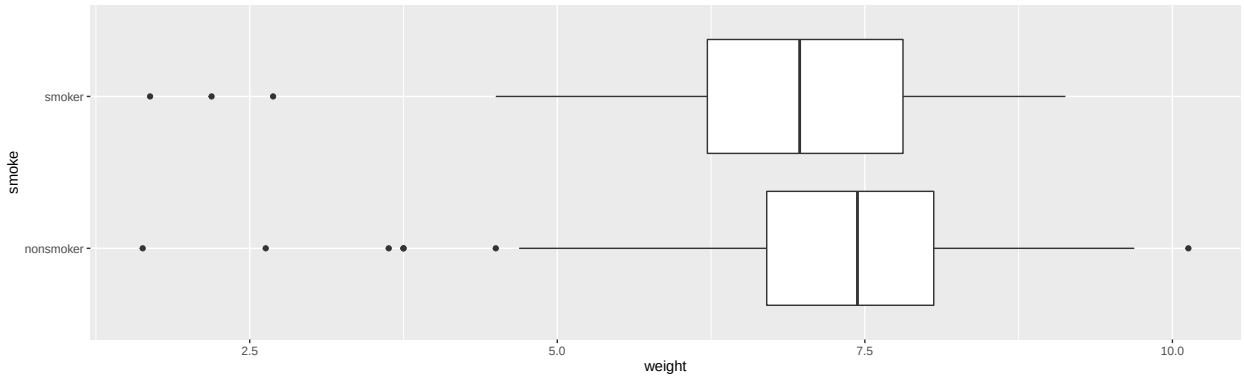
```
births %>% ggplot(aes(x=m_age)) +
  geom_histogram(bins=15, color="black", fill="blue")
```



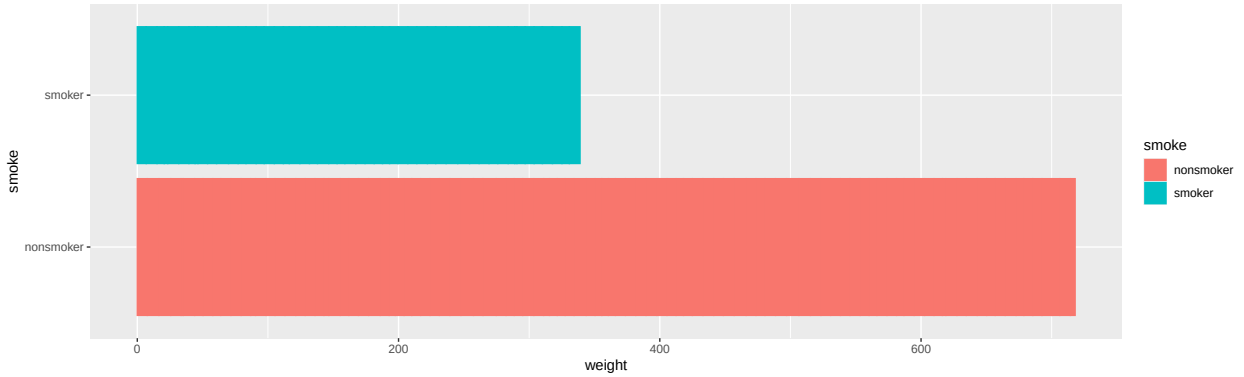
It is continuous.This graph showcases that it is bimodality.The graph also appears to be skewed to the right side. It has a range of 40 ,so we may also say that the spread is significant and thus the variance may be as well.

(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=weight, y=smoke)) + geom_boxplot()
```

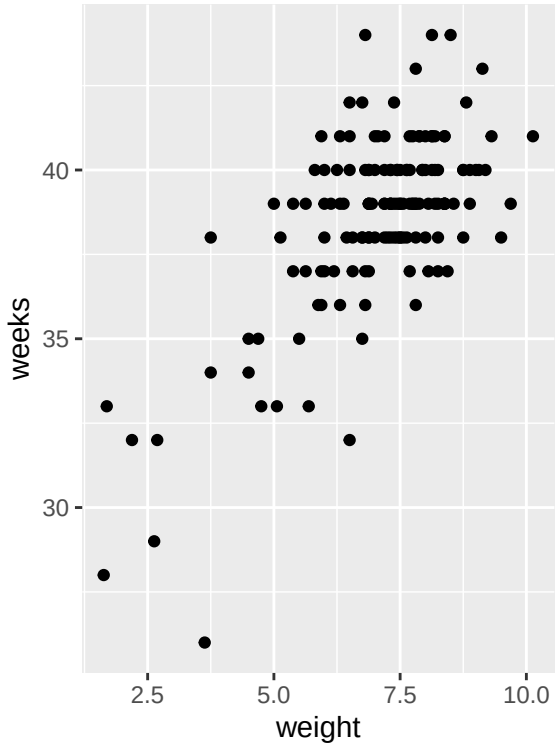



```
births %>% ggplot(aes(x=weight,y=smoke,color=smoke,fill=smoke)) +
 geom_bar(stat='identity')
```



(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.

```
ggplot(births, aes(x=weight, y=weeks)) +
geom_point()
```




```
# 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
```

Additional Recommended Study Material

Did you finish quickly? Do you still have an unused course study time allocation? Would you like to cover this material a little bit more?

ggplot2

- Official Cheatsheet
  - Finding Answers
- Learning Resources
  - Official Usage
  - R4DS Textbook
  - DoSS Toolkit

Markdown

Markdown supports efficiency and productivity, and it's needed for our class.

- RStudio Markdown Cheatsheet
  - R4DS Introduction
  - RStudio Introduction
- Markdown Tutorial

For Reference Only: *NOT a Reading Recommendation*

- knitr Documentation
- .Rmd Documentation