STA130 Rstudio Homework

Problem Set 2

Daniel Sun (1008992609), 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?

(2)

(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)`?

1. We first combine the TRUE and FALSE into a logical vector 2. Convert the logical vector into their numerical values, TRUE = 1 and FALSE = 1 3. sum their values, which equals 1 The additional step required to get the answer for sum(my_answers == square_answers) to check if every value is the same in both vectors

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?

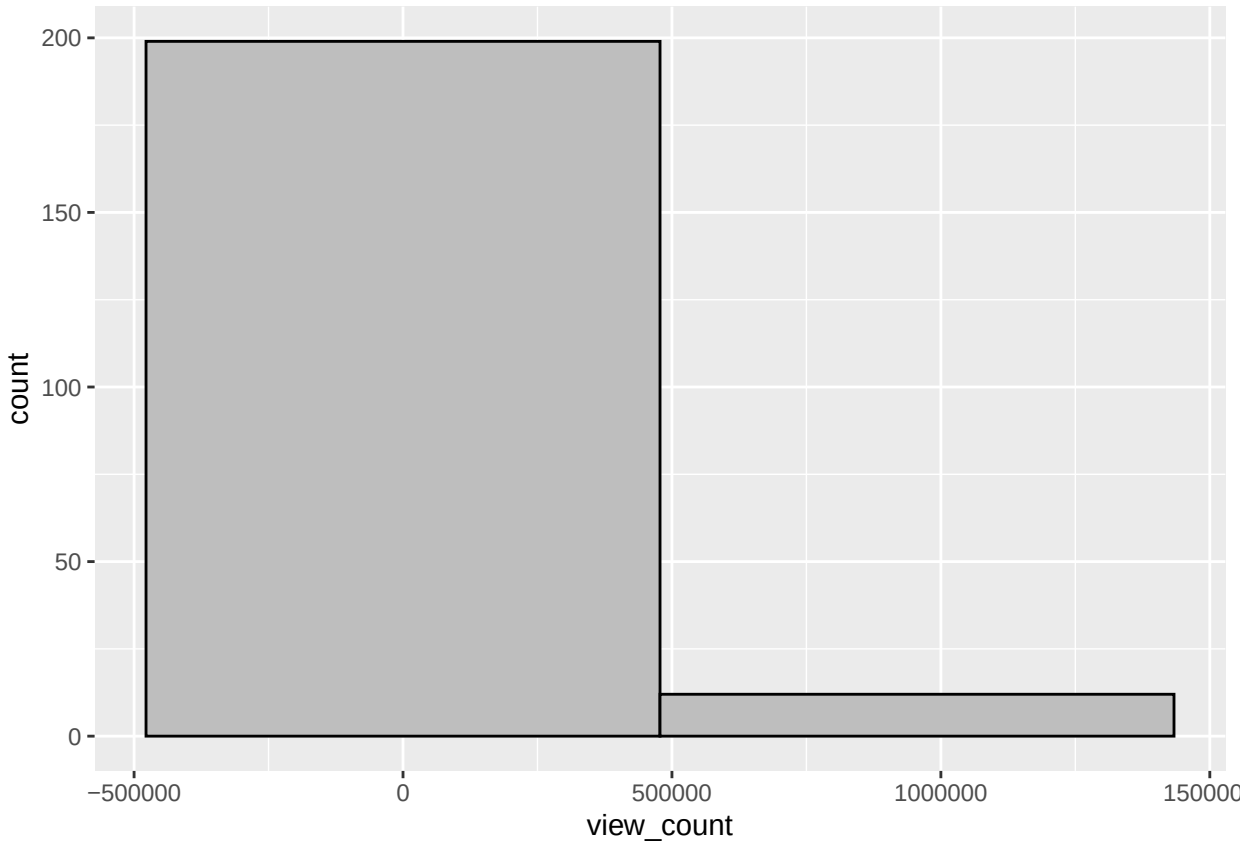
```
glimpse(superbowl)
```

```
## 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/~
```

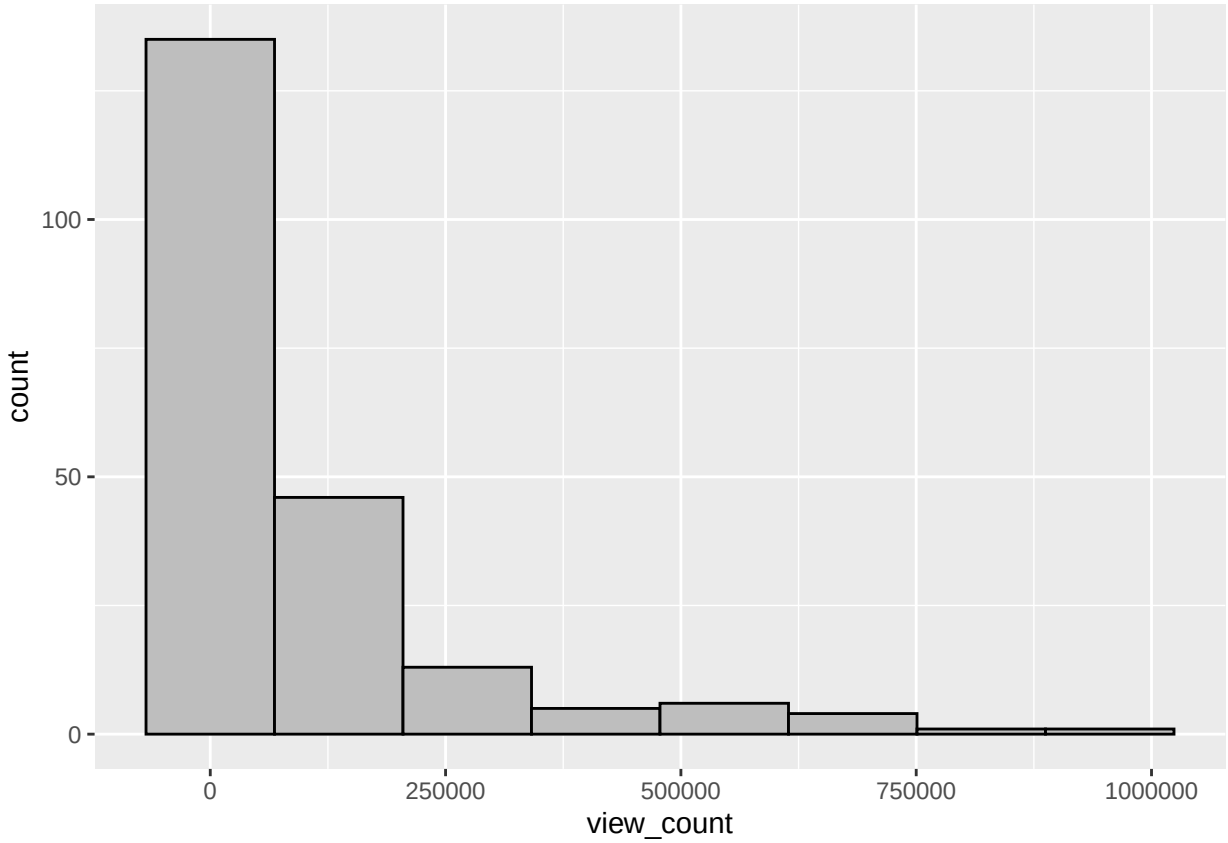
rows/observations: 211 columns/variables: 11

(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.

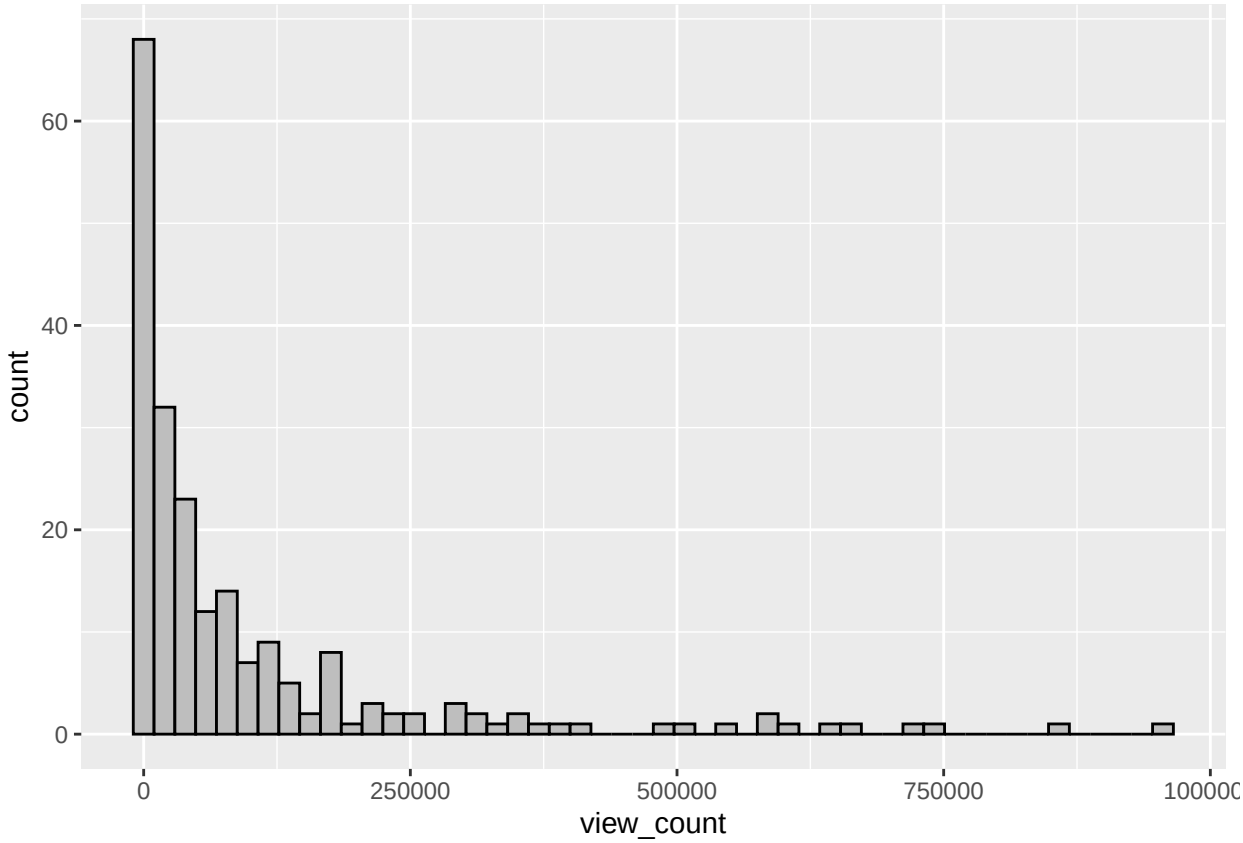
```
ggplot(data= superbowl) + aes(x = view_count) + geom_histogram(colour = "black", fill = "gray", bins = 1
```




```
ggplot(data= superbowl) + aes(x = view_count) + geom_histogram(colour = "black", fill = "gray", bins = 8
```




```
ggplot(data= superbowl) + aes(x = view_count) + geom_histogram(colour = "black", fill = "gray", bins = 5
```



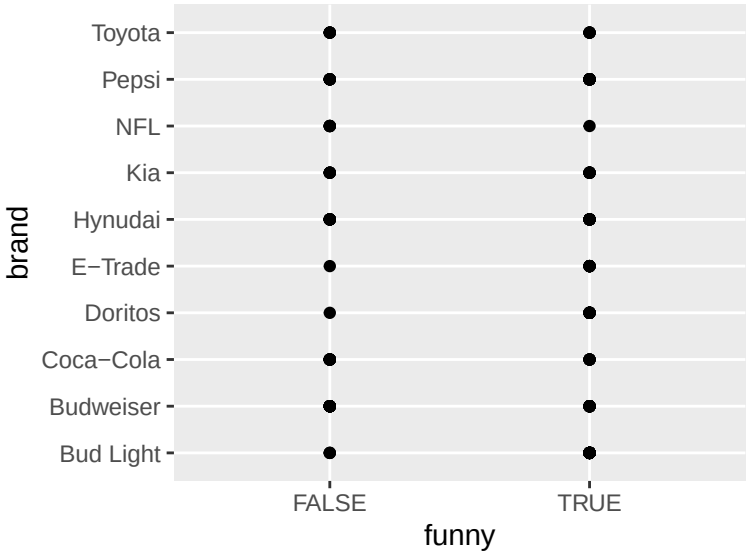
Which of these histograms is most appropriate to describe the distribution of `view_count`? Why? Write a few sentences describing the distribution based on the histogram you chose as most appropriate.

In this case I believe the one with 50 bins is more accurate as 2 and 11 are too little

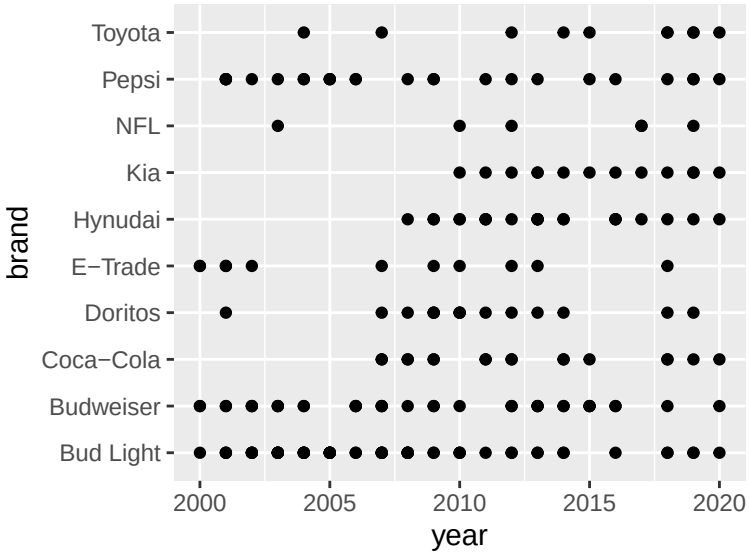
(c) Construct two plots (your choice) to visualize the distribution of `brand` and one of the following other categorical variables from the superbowl ads data: `funny`, `danger`, or `celebrity`. Make sure to specify meaningful axis labels where appropriate.

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}).

```
ggplot(data= superbowl) + aes(x = funny, y = brand) + geom_point()
```




```
ggplot(data= superbowl) + aes(x = year, y=brand) + geom_point()
```

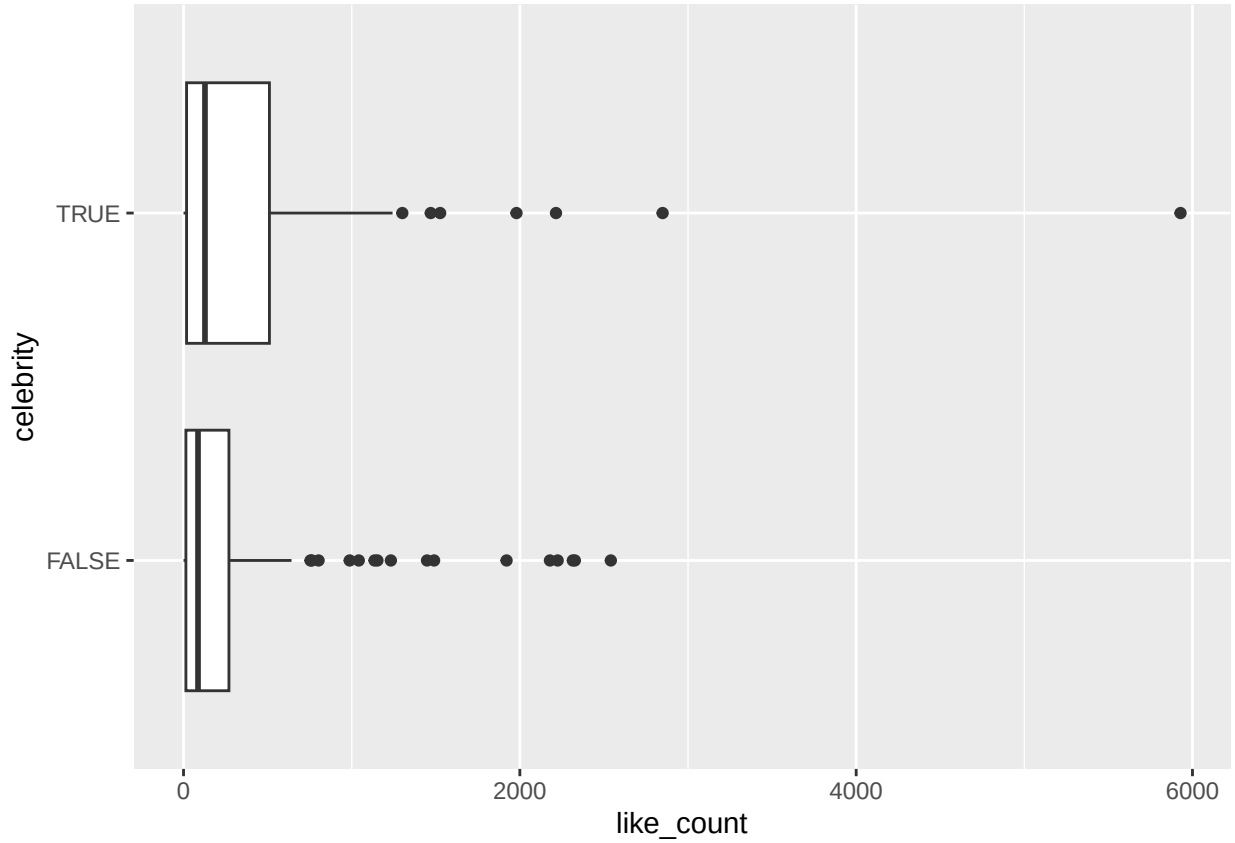


Describe the distribution in 1-2 sentences.

> *In the first graph we see all brands have done funny and non-funny ads. In the second graph we see some companies only began playing ads in the super bowl recently such as kia and they do so consistently which brands like bud light have been doing so pretty consistently for a while?*

(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.

```
ggplot(data=superbowl, aes(x = like_count, y = celebrity)) + geom_boxplot()
```



Write 2-4 sentences comparing these distributions.

> *We can see that while videos with celebrities do get more likes in general, it is extremely marginal as most videos with lots of likes with or without celebrities is near 0 on the graph. We can even see that for most of the outliers, having or not having a celebrity doesn't make a difference. However, the most liked video in both cases, does have a celebrity likely a very famous one to cause people to search for the ad on Youtube and like it.*

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
```

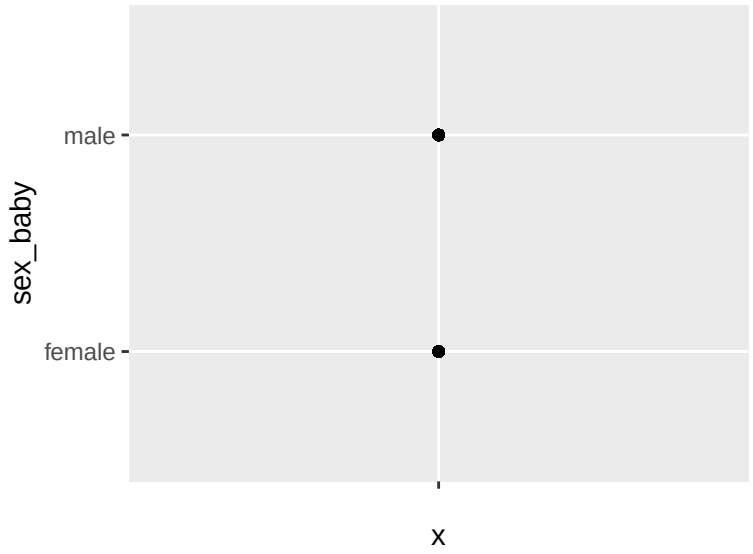
```
births %>% glimpse()
```

```
## 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, ~
```

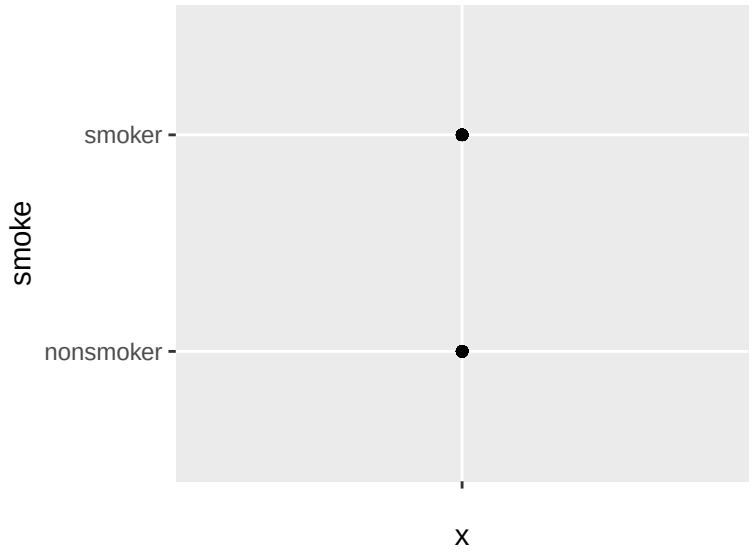
```
## $ premature <fct> full term, full term, full term, full term, full term, full ~
## $ visits    <int> 13, 5, 12, 13, NA, 12, 6, 9, 5, 13, 5, 15, 13, 10, 11, 13, 1~
## $ gained    <int> 1, 35, 29, 30, 10, 35, 29, 15, 40, 34, 32, 20, 47, 20, 5, 22~
## $ weight    <dbl> 6.88, 7.69, 8.88, 9.00, 7.94, 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 ="" , y=sex_baby) + geom_point()
```




```
ggplot(data= births) + aes(x = "", y=smoke) + geom_point()
```

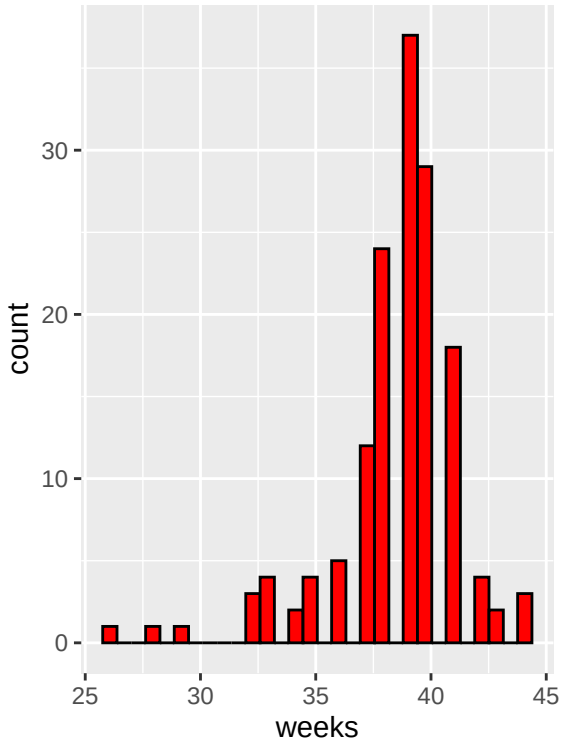


Identify whether each of these variables is a nominal or ordinal categorical variable and write one or two sentences interpreting each plot.

> *sex_baby: nominal, babies' are either male or female smoke: nominal, women are either smokers or non-smokers*

(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=weeks) + geom_histogram(colour = "black", fill = "red")
```

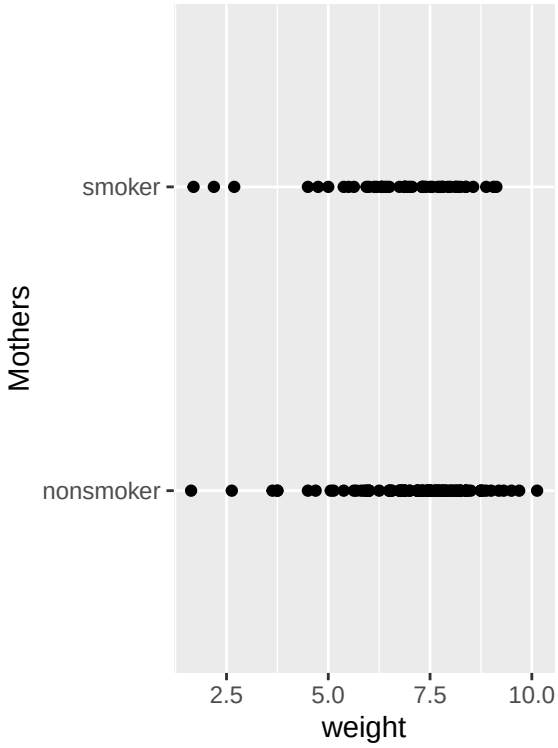


Identify whether the variable you selected is continuous or discrete, and write 2-3 sentences describing the distribution.

> *weeks: discrete, most women do give birth around the 8-10 month range some being born later and some being premature and maybe miscarriages*

(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.

```
ggplot(data= births) + aes(x = weight, y = smoke) + ylab("Mothers")+ geom_point()
```

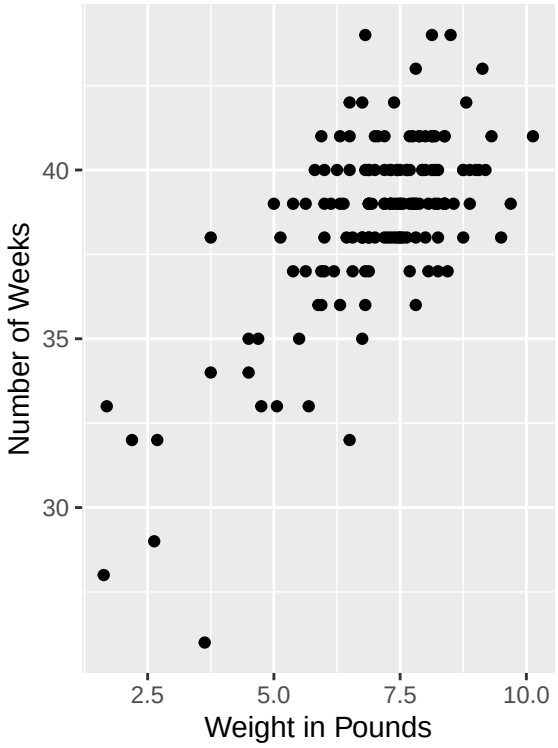


Write 2-3 sentences describing your observations and what this data might suggest (or not suggest) about the impact of smoking on birth weight.

> *We see that the majority of points for nonsmokers have heavier babies compared to smokers. Heavier babies isn't necessarily a good thing as a healthy weight for a baby is 5-8 pounds. However what we don't have data for is how heavy the baby should've been for the smokers due to effects of smoking of childbirth. We always want the natural weight for a myriad of reasons such as monitoring possible diseases.*

(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(data=births) + aes(x = weight, y = weeks) + ylab("Number of Weeks")+
   xlab("Weight in Pounds") + geom_point()
```



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?

> *There is a clear relationship between birth weight and gestational age; the more time in the womb, generally, the heavier the baby is. This definitely sheds new light on my interpretation as I would now like to see how long smokers are pregnant vs non-smokers and perhaps smoking is not directly responsible but indirectly responsible for the childs weight by making the gestational period shorter.*