Exam 1 Rubric

02/05/22

First, please load the `tidyverse`, `ggridges`, and `viridis` packages.

```
library(tidyverse)
library(viridis)
library(ggridges)
```

Load the data here:

```
nutrition <- read_csv("starbucks_nutrition.csv")
menu <- read_csv("starbucks_menu.csv")
```

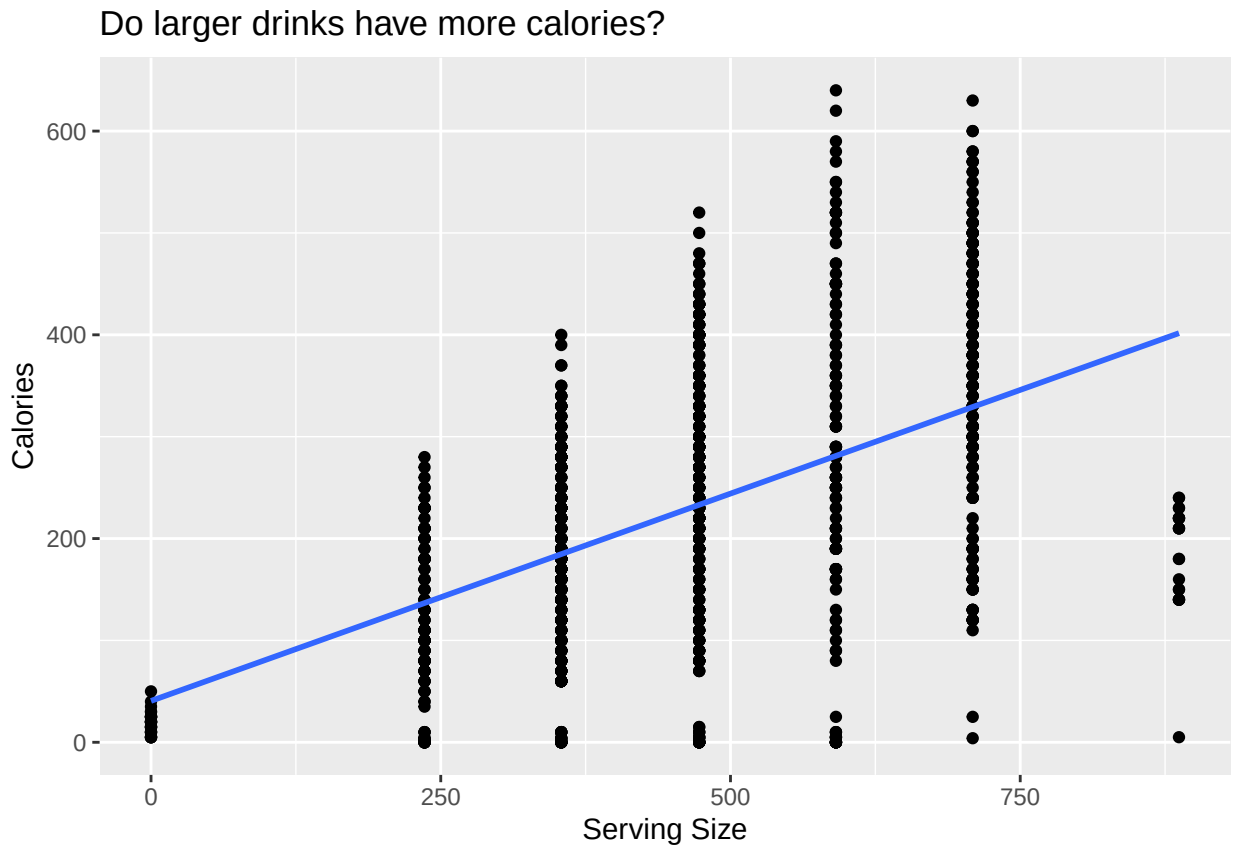
Exercise 1:

```
starbucks <- menu %>%
  left_join(nutrition, by = c("product_id"))
```

Exercise 2:

```
ggplot(data = starbucks,
       mapping = aes(x = serv_size_m_l, y = calories)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Do larger drinks have more calories?",
       x = "Serving Size",
       y = "Calories")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Answer: We can see that the smaller drinks have a smaller calorie range and are usually lower in calories than the larger drinks. The line of best fit also has a positive slope which supports this. However, all larger drinks are not necessarily higher in calories : although some of them are, the data visualization is not sufficient to prove there is a relation between serving size and calories. This is supported by the fact that most points are not close to the line of best fit

Exercise 3:

```
starbucks %>%
  group_by(size) %>%
  summarise(meancals = mean(calories), number = n()) %>%
  arrange(desc(meancals))
```

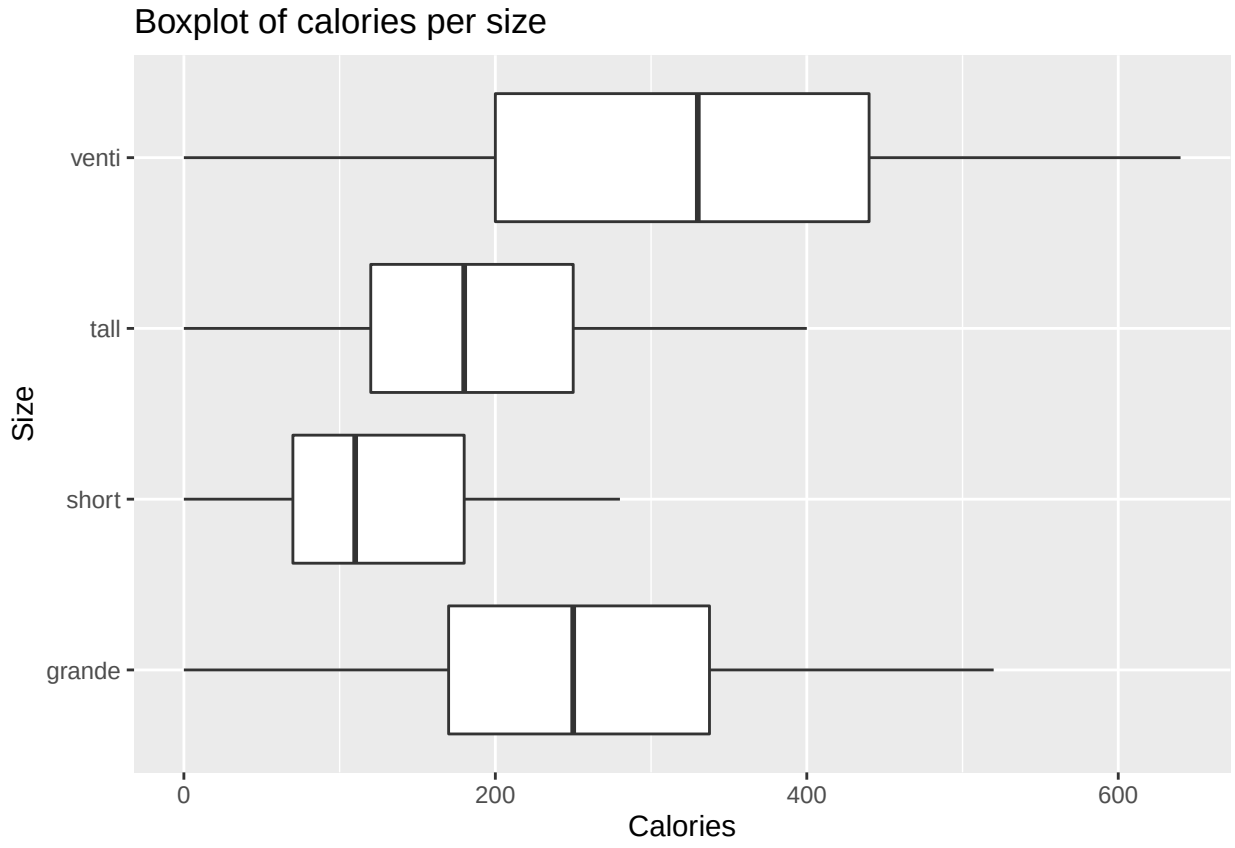
```
## # A tibble: 11 x 3
##    size    meancals number
##    <chr>      <dbl>  <int>
##  1 venti      320.     320
##  2 grande     248.     334
##  3 trenta     183.      21
##  4 tall       182.     318
##  5 short      116.     123
##  6 quad        27.9      7
##  7 1 scoop     27.5      2
##  8 triple      22.1      7
```

```
##  9 doppio      16.4       7
## 10 solo        10         7
## 11 1 shot       5         1
```

Answer: Venti is the size category with the most mean calories while 1 shot is the size category with the least. I am doubtful about the reliability of this because it is possible that most of the higher calorie drinks are only available in size venti or tall and not grande. I am also doubtful because only 1 drink is available in the 1 shot category, which means they are probably special drinks which are not the same as most of the other drinks on the menu

Exercise 4

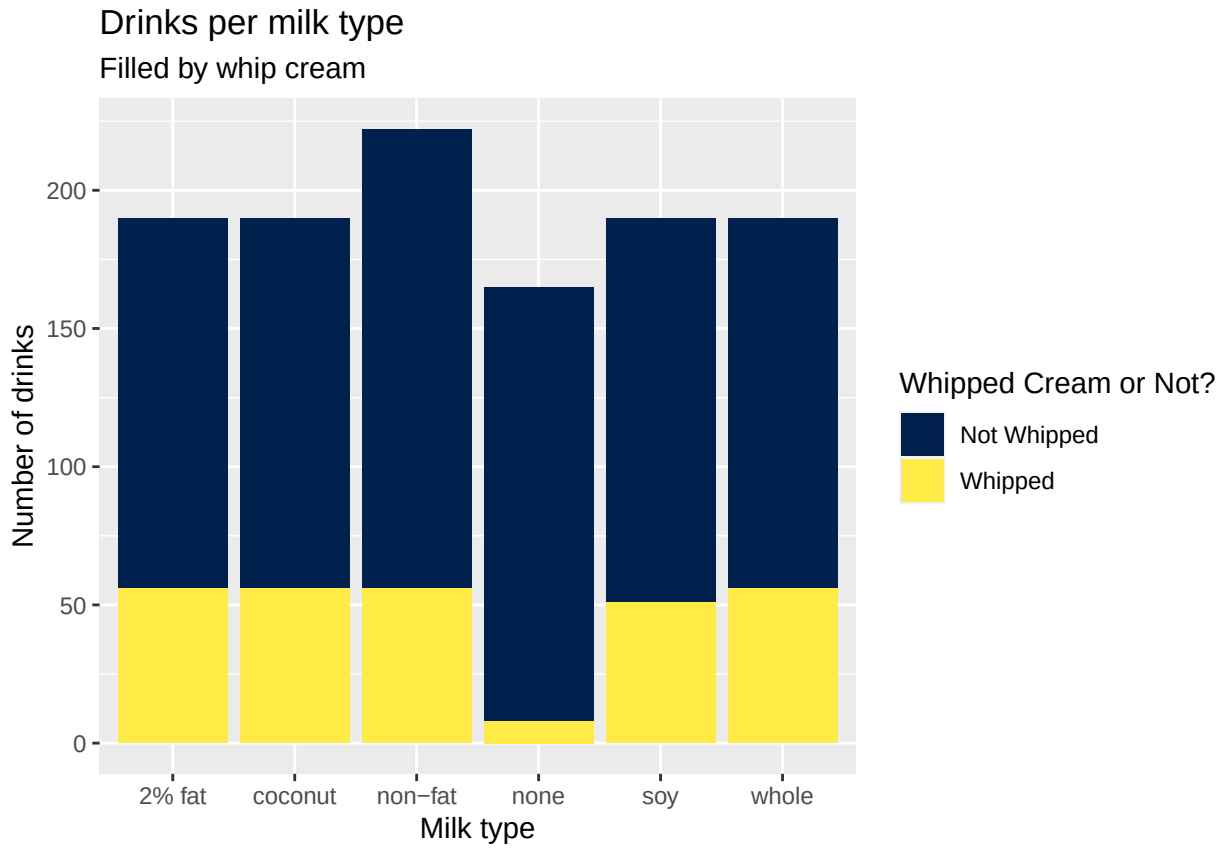
```
starbucks %>%
  group_by(size) %>%
  filter(n() > 100) %>%
  ggplot(mapping = aes(x = calories, y = size)) +
  geom_boxplot() +
  labs(title = "Boxplot of calories per size",
       x = "Calories",
       y = "Size")
```



Answer: We see that venti has the most mean calories followed by grande, then tall, then short. Venti also has the largest spread, and the mean is pretty centred. Grande has a smaler spread and is also pretty centred. However, for tall and short, the spreads are considerably smaller, and the means are centred more towards the lower quartiles, significantly so in drinks of the short size category

Exercise 5:

```
starbucks %>%
  mutate(milktype = case_when(milk==0 ~ 'none',
                              milk==1 ~ 'non-fat',
                              milk==2 ~ '2% fat',
                              milk==3 ~ 'soy',
                              milk==4 ~ 'coconut',
                              milk==5 ~ 'whole')) %>%
  mutate(whipona = if_else(whip == 1, "Whipped", "Not Whipped")) %>%
  ggplot(mapping = aes(x = milktype, fill = whipona)) +
  geom_bar() +
  scale_fill_viridis(discrete = TRUE, option = "E",
                     name = "Whipped Cream or Not?") +
  labs(title = "Drinks per milk type",
       subtitle = "Filled by whip cream",
       x = "Milk type",
       y = "Number of drinks")
```



Exercise 6:

```
starbucks %>%
  filter(size == "venti") %>%
```

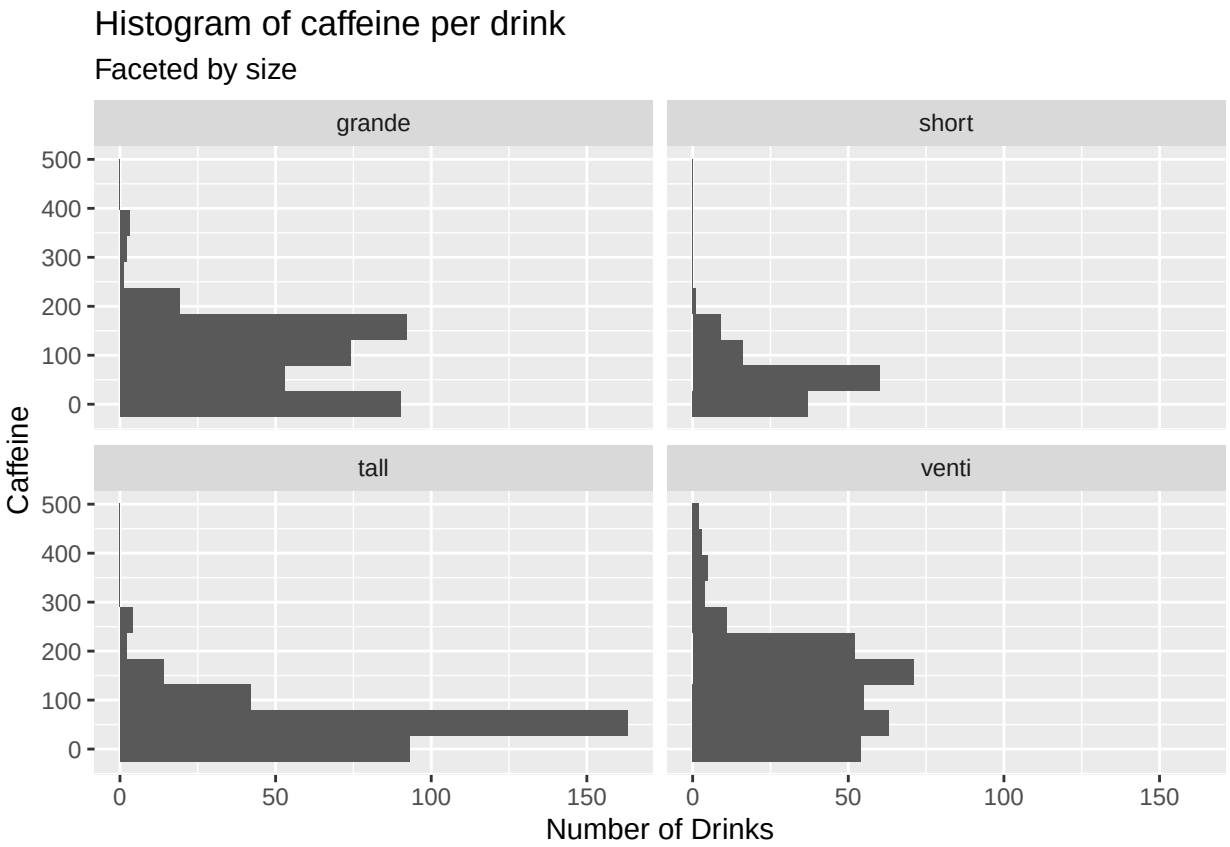
```
  arrange(desc(sodium_mg)) %>%
  slice(1:5) %>%
  summarise(product_name, sodium_mg)
```

```
## # A tibble: 5 x 2
##   product_name                                        sodium_mg
##   <chr>                                                   <dbl>
## 1 Double Chocolaty Chip Crème Frappuccino Blended       370
## 2 Strawberries & Crème Frappuccino Blended              370
## 3 Java Chip Frappuccino Blended                         360
## 4 Java Chip Frappuccino Blended                         360
## 5 Double Chocolaty Chip Crème Frappuccino Blended       360
```

Answer: The five drinks are: Double Chocolaty Chip Crème Frappuccino Blended Strawberries & Crème Frappuccino Blended
Java Chip Frappuccino Blended
Java Chip Frappuccino Blended
Double Chocolaty Chip Crème Frappuccino Blended

Exericse 7:

```
starbucks %>%
  group_by(size) %>%
  filter(n() > 100) %>%
  ggplot(mapping = aes(y = caffeine_mg)) +
  geom_histogram(bins = 10) +
  labs(title = "Histogram of caffeine per drink",
       subtitle = "Faceted by size",
       x = "Number of Drinks",
       y = "Caffeine") +
  facet_wrap(~size)
```



Answer: In the venti and grande category, the spread is pretty high and is mostly centred around 0-200mg. For tall and short, the spread is low and has a high peak around 50mg.

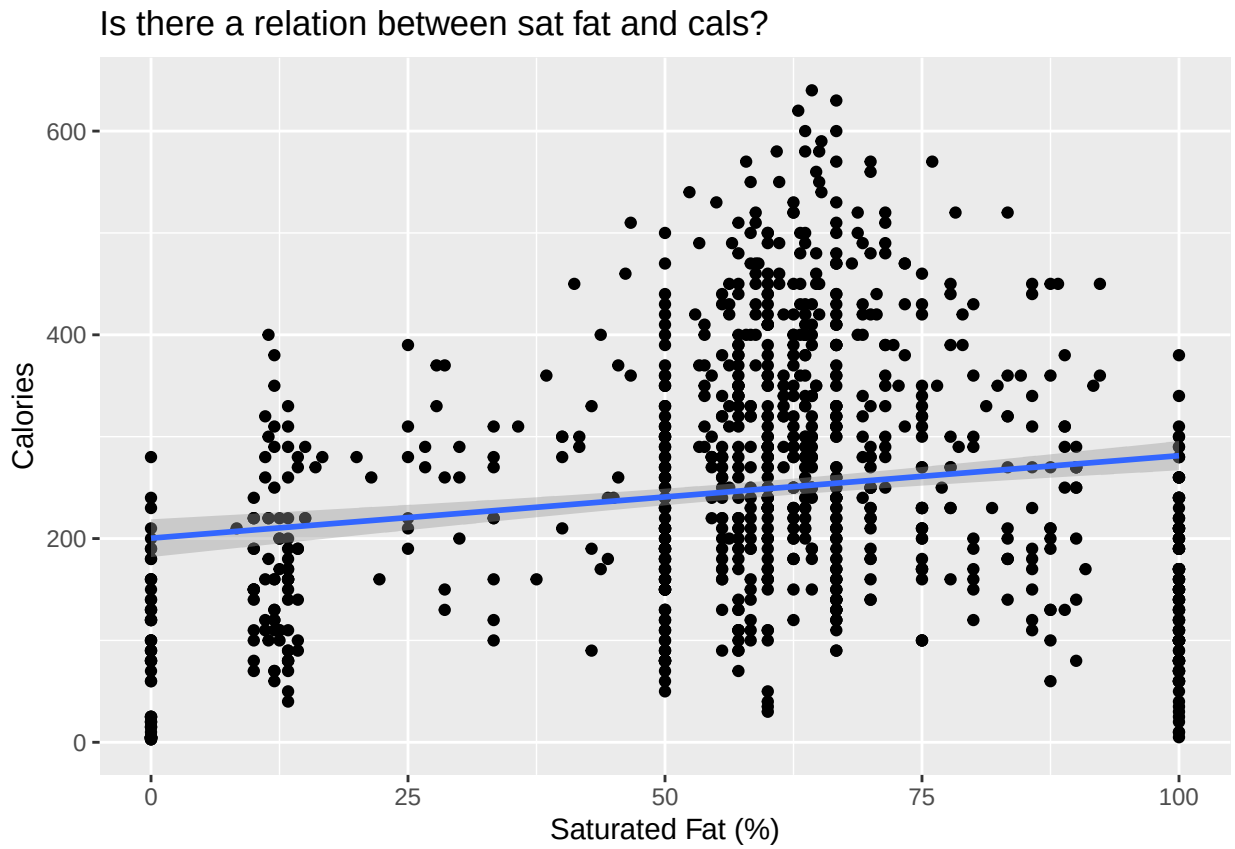
Exercise 8:

```
starbucks %>%
  mutate(satfatpercent = (saturated_fat_g / total_fat_g) * 100) %>%
  ggplot(mapping = aes(x = satfatpercent, y = calories)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(title = "Is there a relation between sat fat and cals?",
       x = "Saturated Fat (%)",
       y = "Calories")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 112 rows containing non-finite values (stat_smooth).
```

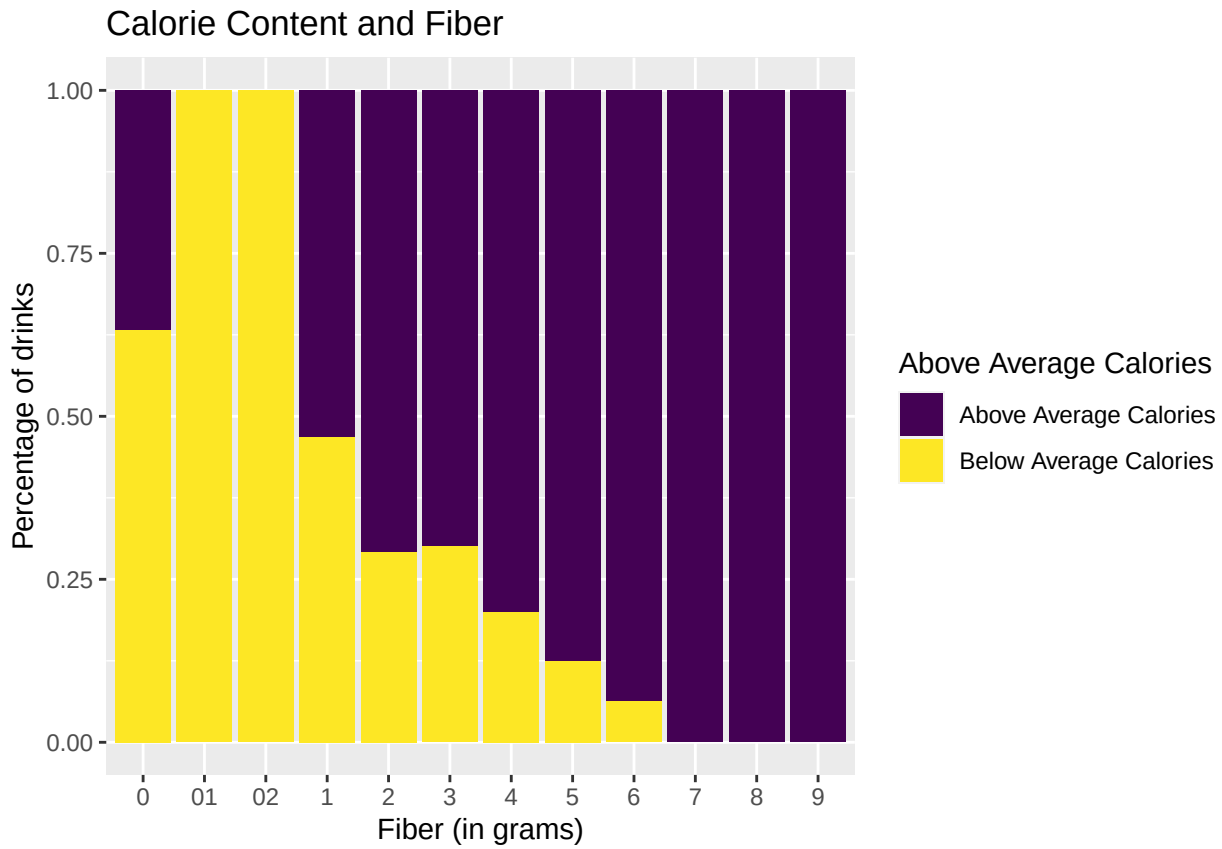
```
## Warning: Removed 112 rows containing missing values (geom_point).
```



Answer: Although the regression line shows a slight increase in calories with an increase in saturated fat, the large and random spread of the points show that there is no clear relation between them.

Exercise 9:

```
starbucks %>%
  mutate(calmeter = if_else(calories > mean(calories),
                            "Above Average Calories",
                            "Below Average Calories")) %>%
  group_by(fiber_g) %>%
  ggplot(mapping = aes(x = fiber_g,
                       fill = calmeter)) +
  geom_bar(position = "fill") +
  scale_fill_viridis(discrete = TRUE, option = "D",
                     name = "Above Average Calories") +
  labs(title = "Calorie Content and Fiber",
       x = "Fiber (in grams)",
       y = "Percentage of drinks")
```



Answer: Yes, you do need to consume drinks with above average calories to get more fiber. It was helpful to fill the bars as percentages because the numbers vary and don't necessarily give you an idea about the increase in % of drinks with increase in fiber, which is something the filled bars do. I think it is a bit strange though that some of the bars are completely full. It is likely that Starbucks is not being fully transparent with their data. Inaccurate data leads to inaccurate analyses, which is why we must strive for data transparency