EDAV Fall 2020 PSet 2

Jiayin Lin, Yifei Zhang

Read *Graphical Data Analysis with R*, Ch. 4, 5

Grading is based both on your graphs and verbal explanations. Follow all best practices as discussed in class. If calculations are involved, your scripts should be written so they would still work if the data values in the datasets you're working with were altered. For example:

Good

```
plot_df <- mtcars %>% group_by(cyl) %>% summarize(mean_mpg = mean(mpg))
```

Bad

```
plot_df <- tibble(cyl = c(4, 6, 8), mean_mpg <- c(26.7, 19.7, 15.1))
```

Hints: Pay attention to bar order. Coordinate fill colors and legends *across* graphs.

```
library("tidyverse")

library("rvest")
library("robotstxt")
library("tibble")

library("EDAWR")
library("hexbin")

library("alr4")
```

1. Water Taste Test

Data: `WaterTaste` dataset in the **Lock5withR** package (available on CRAN)

```
library("Lock5withR")
wt <- WaterTaste
```

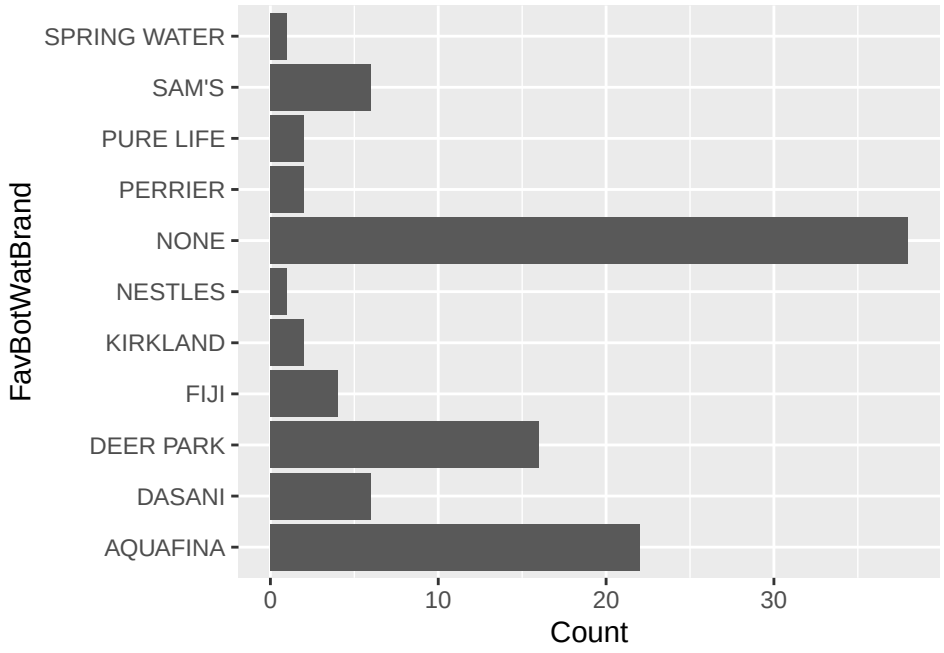
(a) Recode the `Class` and `Sex` columns using the human readable values listed in the help file. Display the first six rows of these two columns.

```
wt <- WaterTaste
wt$Class <- recode_factor(wt$Class, "F" = "First Year", "SO" = "Sophomore", "J" = "Junior", "SR" = "Seni
wt$Gender <- recode_factor(wt$Gender, "F" = "Female", "M" = "Male", .ordered = TRUE)
head(wt[c("Class", "Gender")], 6)
```

```
##        Class Gender
## 1 First Year Female
## 2 First Year Female
## 3 First Year Female
## 4 First Year Female
## 5     Junior   Male
## 6     Junior   Male
```

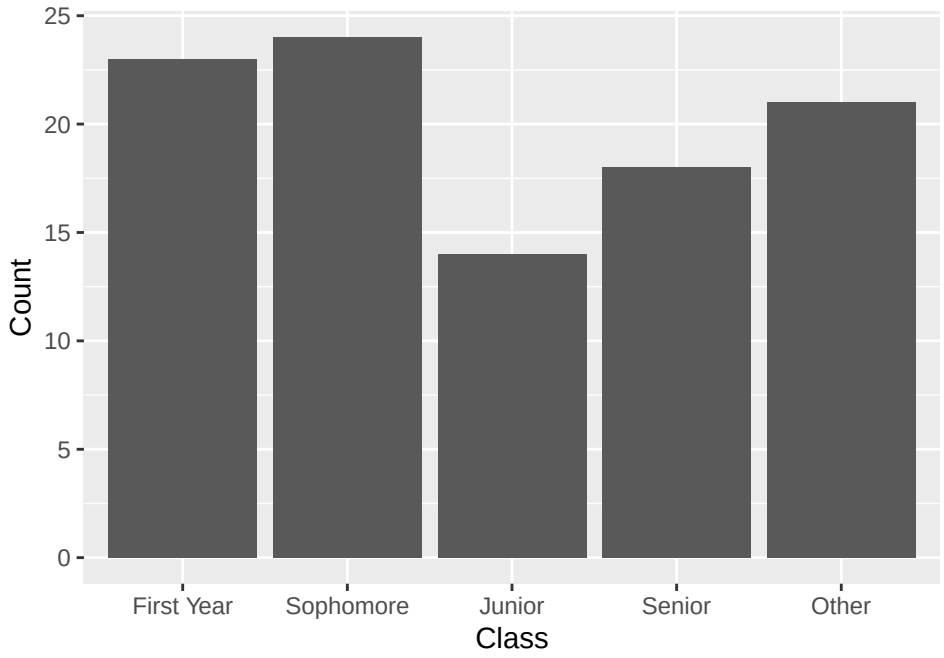
(b) Create a horizontal bar chart of `FavBotWatBrand` counts.

```
ggplot(wt, aes(x = FavBotWatBrand)) + geom_bar() + ylab("Count") + coord_flip()
```



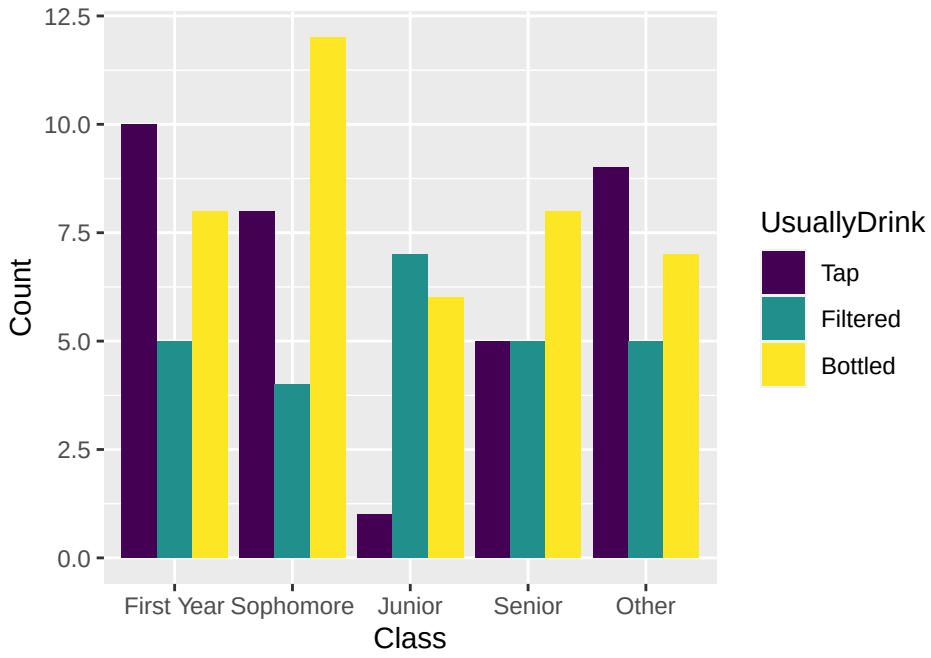
(c) Create a vertical bar chart of `Class` counts.

```
ggplot(wt, aes(x = Class)) + geom_bar() + ylab("Count")
```



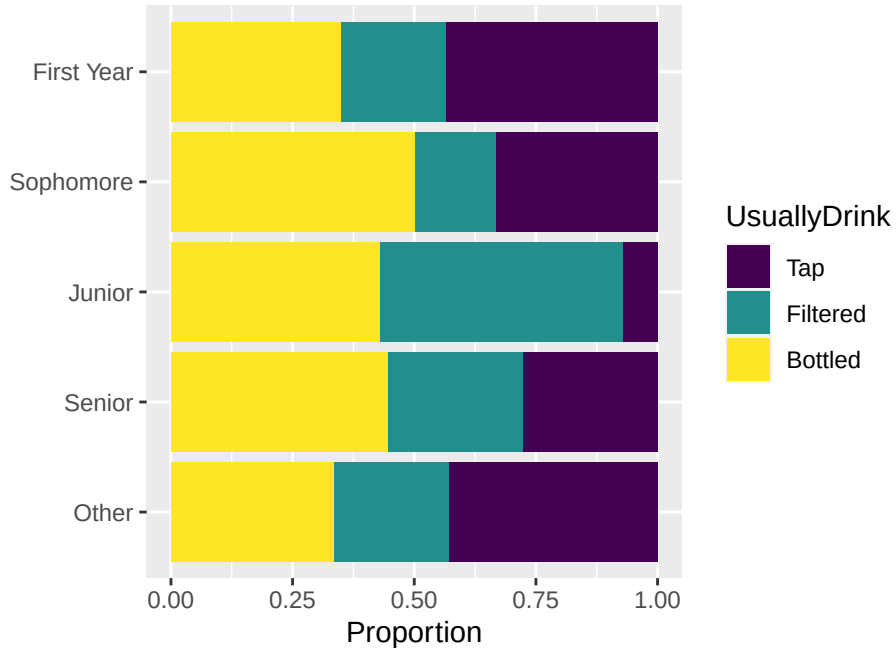
(d) Create a vertical grouped bar chart of `Class` and `UsuallyDrink` in which each level of `Class` forms one group containing three bars representing the three levels of `UsuallyDrink`.

```
wt$UsuallyDrink <- recode_factor(wt$UsuallyDrink, "Tap" = "Tap", "Filtered" = "Filtered", "Bottled" = "
ggplot(wt, aes(x = Class, fill = UsuallyDrink)) + geom_bar(position = "dodge") + ylab("Count")
```



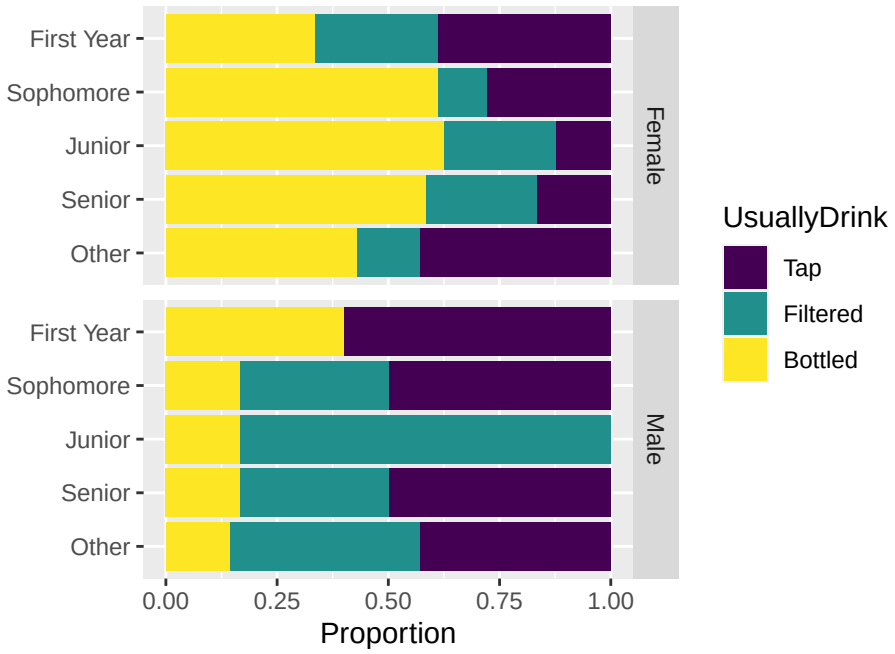
(e) Create a horizontal stacked bar chart of proportions showing the type of water respondents usually drink by Class. The order of the levels of both categorical variables should match what is shown in the assignment. (Note that the order of the fill colors of the bars match the order of the fill colors in the legend.)

```
wt$UsuallyDrink <- recode_factor(wt$UsuallyDrink, "Tap" = "Tap", "Filtered" = "Filtered", "Bottled" = "
ggplot(wt, aes(x = reorder(Class, desc(Class)), fill = UsuallyDrink)) + geom_bar(position = "fill") + x
```



(f) Create a horizontal stacked bar chart showing the proportional breakdown of `Class` for each level of `UsuallyDrink`, faceted on `Gender`. Use a descriptive title. (See assignment for example.)

```
wt$UsuallyDrink <- recode_factor(wt$UsuallyDrink, "Tap" = "Tap", "Filtered" = "Filtered", "Bottled" = "
ggplot(wt, aes(x = reorder(Class, desc(Class)), fill = UsuallyDrink)) + geom_bar(position = "fill") + x
```



2. Metacritic

To get the data for this problem, we'll scrape data from `www.metacritic.com`. Important: you should only execute parts (a) and (b) *once*. Therefore, it should be clear to us that the code isn't being run each time you knit the document. You may either set `eval=FALSE` in these chunks or comment out the appropriate lines.

(a) Use the `paths_allowed()` function from **robotstxt** to make sure it's ok to scrape https://www.metacritic.com/publication/digital-trends. What is the result?

```
paths_allowed("https://www.metacritic.com/publication/digital-trends")
```

```
## [1] TRUE
```

(b) Use the **rvest** package to read the URL in part (a), and then find the title, metascore and critic score for each game listed. Create a data frame with these three columns and save it. (You may remove any rows with missing data.)

```
wp <- read_html("https://www.metacritic.com/publication/digital-trends")

title_text <- html_nodes(wp, 'div.review_product a') %>% html_text()
metascore_text <- html_nodes(wp, '.brief_metascore') %>% html_text() %>% str_trim() %>% substring(1, 2)
criticscore_text <- html_nodes(wp, '.brief_critscore') %>% html_text() %>% str_trim() %>% substring(1, 1

table <- tibble(title = title_text, metascore = metascore_text, criticscore = criticscore_text)
table <- table[complete.cases(table), ]

write.csv(table,'hw2q2.csv')
```

(c) Read your saved data back in and display the first six rows.

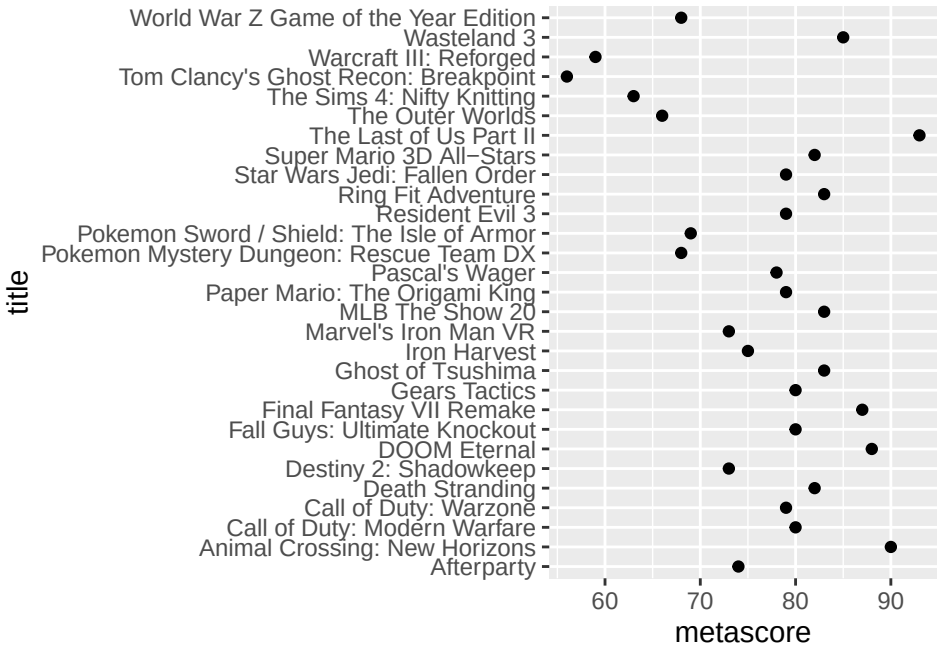
```
metacritic <- read.csv(file = 'hw2q2.csv')
head(metacritic, 6)
```

```
##   X                         title metascore criticscore
## 1 1      Super Mario 3D All-Stars        82          80
## 2 2                   Iron Harvest        75          60
```

```
## 3 3                   Wasteland 3        85          80
## 4 4  Fall Guys: Ultimate Knockout        80          90
## 5 5    The Sims 4: Nifty Knitting        63          80
## 6 6 Paper Mario: The Origami King        79          70
```

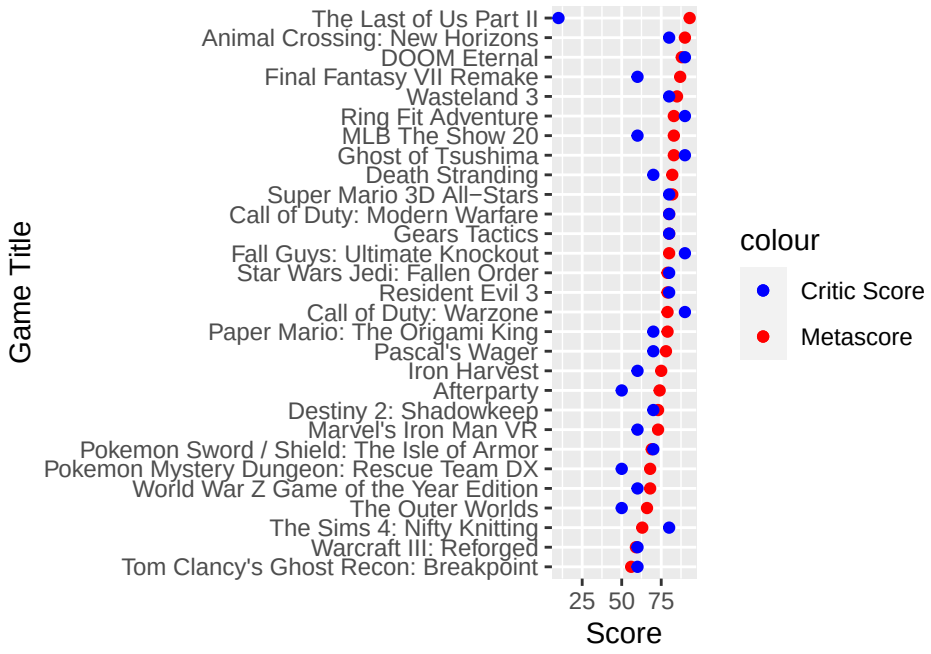
(d) Create a Cleveland dot plot of metascores.

```
ggplot(metacritic, aes(metascore, title)) + geom_point()
```



(e) Create a Cleveland dot plot of metascore *and* critic score on the same graph, one color for each. Sort by metascore.

```
metacritic$title <- factor(metacritic$title, levels = metacritic[order(metacritic$metascore),]$title)
colors = c("Metascore" = "red", "Critic Score" = "blue")
ggplot(metacritic) + geom_point(aes(metascore, title, color = "Metascore")) + geom_point(aes(criticscor
```



3. Nutrition

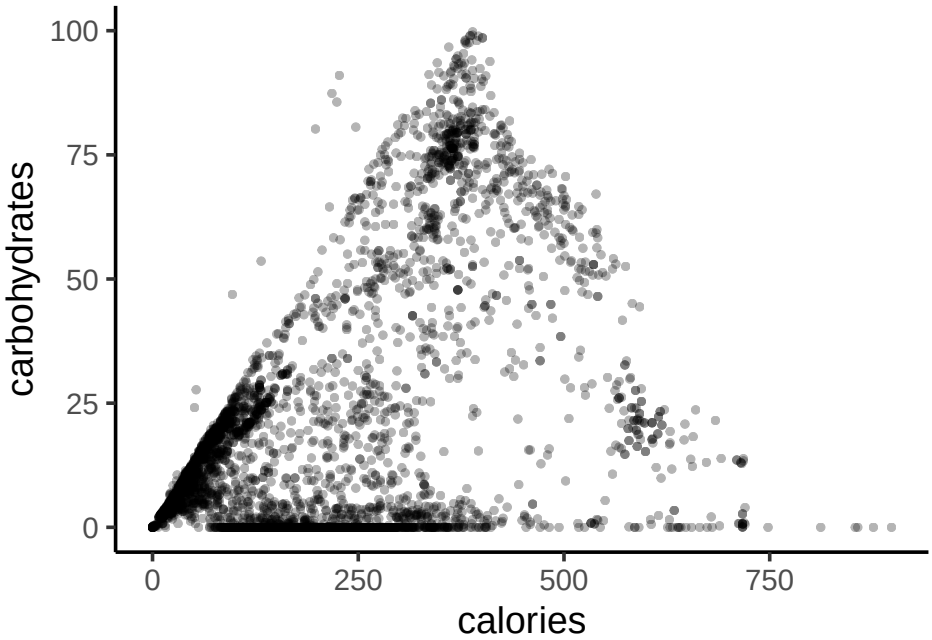
Data: `nutrition` dataset in **EDAWR** package, install from GitHub:

```
remotes::install_github("rstudio/EDAWR")
```

For parts (a) - (d) draw four plots of `calories` vs. `carbohydrates` as indicated. For all, adjust parameters to the levels that provide the best views of the data.

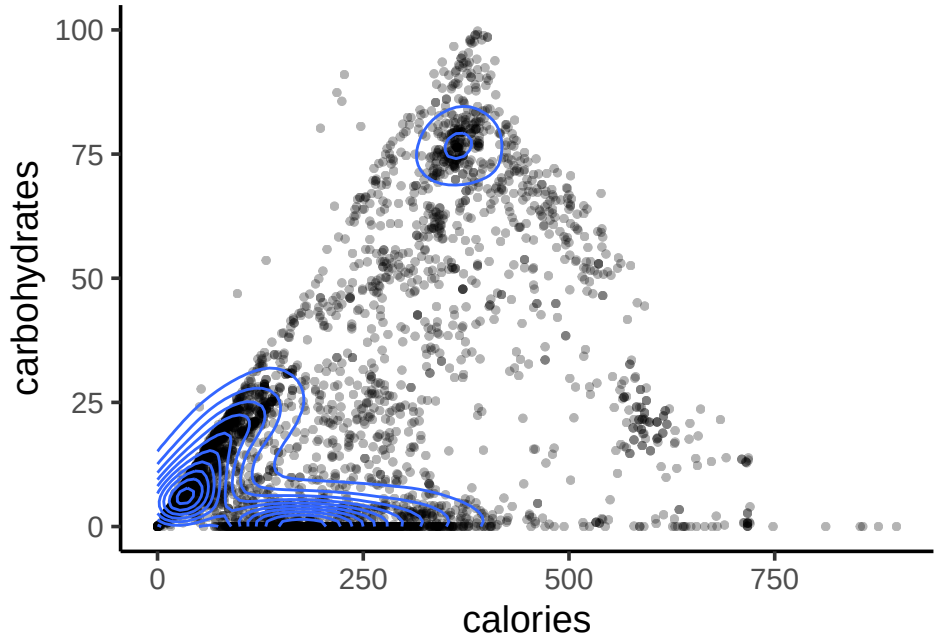
(a) Points with alpha blending

```
ggplot(nutrition, aes(x = calories, y = carbohydrates)) + geom_point(alpha = 0.3, stroke = 0) + theme_c
```



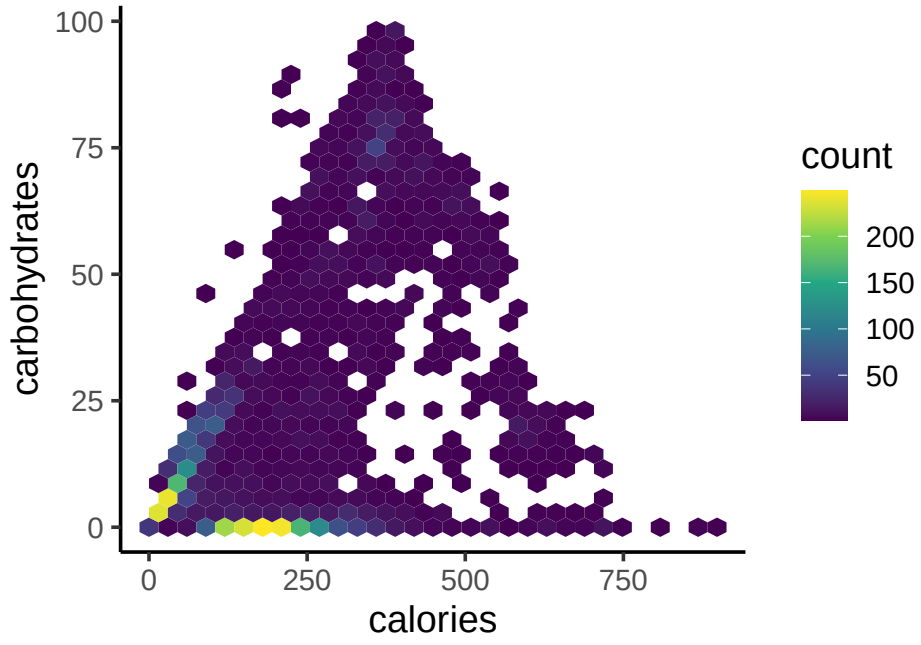
(b) Points with alpha blending + density estimate contour lines

```
ggplot(nutrition, aes(x = calories, y = carbohydrates)) + geom_point(alpha = 0.3, stroke = 0) + geom_de
```



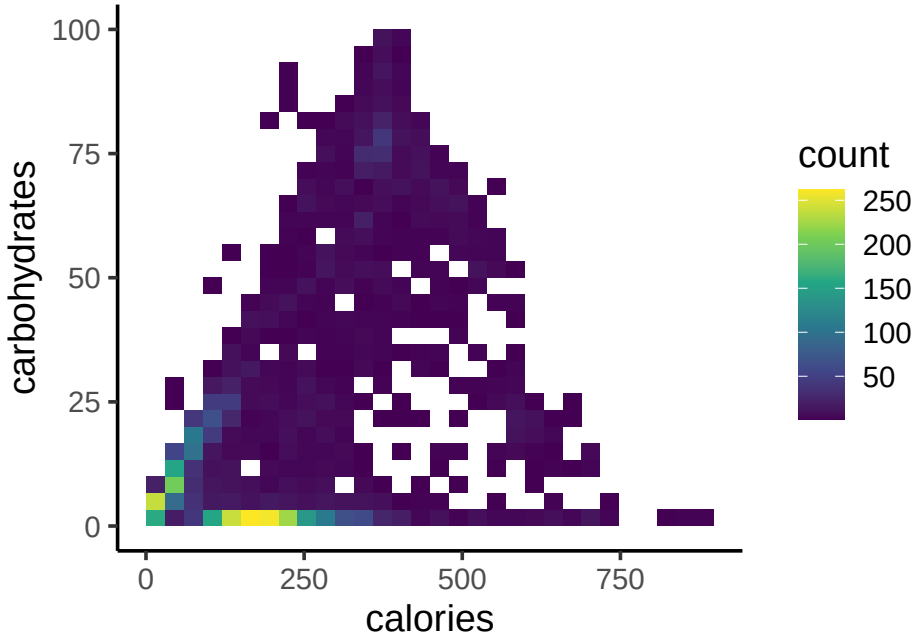
(c) Hexagonal heatmap of bin counts

```
ggplot(nutrition, aes(x = calories, y = carbohydrates)) + scale_fill_viridis_c() + geom_hex() + theme_c
```



(d) Square heatmap of bin counts

```
ggplot(nutrition, aes(x = calories, y = carbohydrates)) + scale_fill_viridis_c() + geom_bin2d() + theme
```



(e) Describe noteworthy features of the relationship between the variables based on your plots from parts (a)-(d), using the "Movie ratings" example on page 82 (last page of Section 5.3) as a guide. Which one do you think is most informative and why?

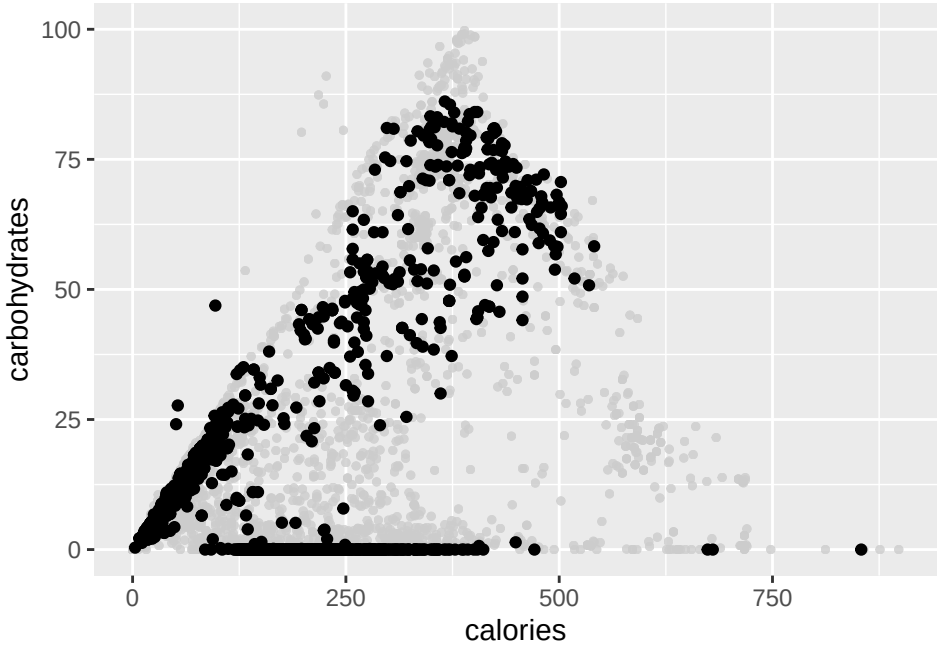
From (a): There is no food that comes with high carbonhydrates but low calories (or very high calories). Most food has calories under 400. From (b): There are many observations in the bottom left corner (calories from 0 to 400 and carbonhydrates from 0 to 35) and the top of "triangle" (calories arounnd 400 and carobonhydrates around 75). From (c) and (d): In addition to information from (a), we can see that most food falls in the bin where value of calories is around 0-250 and carbonhydrates is around 0-25.

(f) Recreate your scatterplot from part (a) with `gray80` for the color, adding an additional `geom_point()` layer only containing points for foods in the top three food categories (`group` column) by count. What do you learn?

We learnt that there are two types of food for the top three food categories: the one whose carbonhydrate and calories are positively correlated, and the one that has no carbonhydrates and varing in calories. We do not see much food with high calories (above 500) but low cabonhydrates in these three groups. Additionally, the distribution closely follows the density contour lines we have in (b).

```
sorted <- nutrition %>% count(group, sort = TRUE)
top_three <- as.vector(sorted$group)[1:3]
top_three <- nutrition[nutrition$group %in% top_three,]

ggplot(NULL, aes(calories, carbohydrates)) + geom_point(data = nutrition, alpha = 0.8, color = 'gray80'
```



4. Australian Institute of Sport data

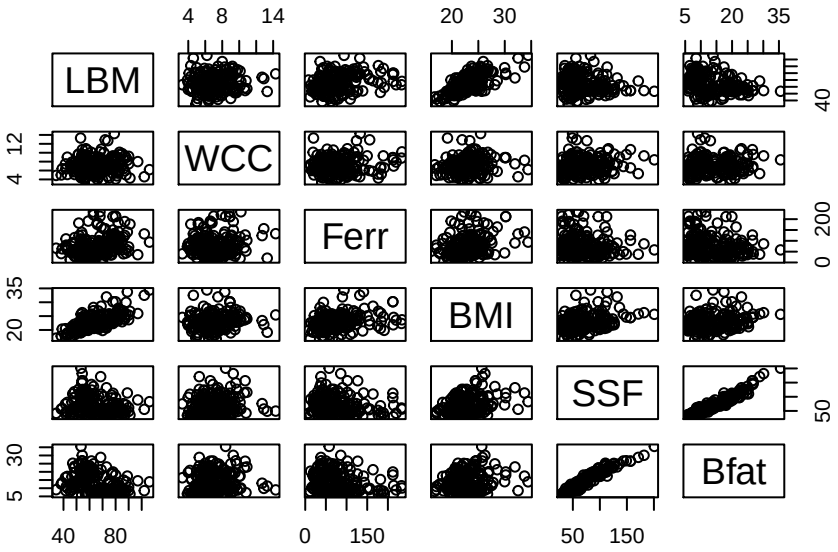
Data: `ais` dataset in **alr4** package (available on CRAN)

(For this question, we only plotted part of the columns to get a better visualization)

(a) Draw a scatterplot matrix of the continuous variables in the `ais` dataset. Which pairs of variables (if any) are strongly positively associated and which are strongly negatively associated?

We cannot see any pairs that are strongly negatively associated, but we can see that SSF and BFat are strongly positively correlated.

```
plot_matrix <- ais %>% dplyr::select("LBM", "WCC", "Ferr", "BMI", "SSF", "Bfat")
plot(plot_matrix)
```



(b) Color the points by `Sex`. Do new patterns emerge? Describe a few of the most prominent.

After dividing into two sex groups, we can see that within each group, LBM and BMI actually displays

a strong positive association, there is some positive association between Bfat and SSF, and there is some positive association between BMI and SSF.

```
plot_matrix <- ais %>% dplyr::select("LBM", "WCC", "Ferr", "BMI", "SSF", "Bfat")
ais$Sex <- as.factor(ais$Sex)
plot(plot_matrix, col = c("red", "blue")[ais$Sex])
```