EDAV Fall 2020 PSet 2

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.

1. Water Taste Test

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

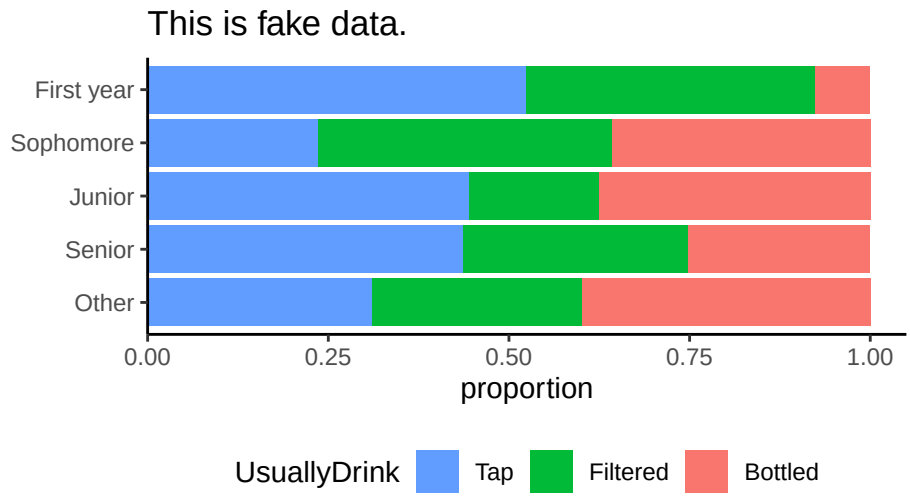
(a) Recode the `Class` and `Sex` columns using the human readable values listed in the help file. Display the first six rows of these two columns.

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

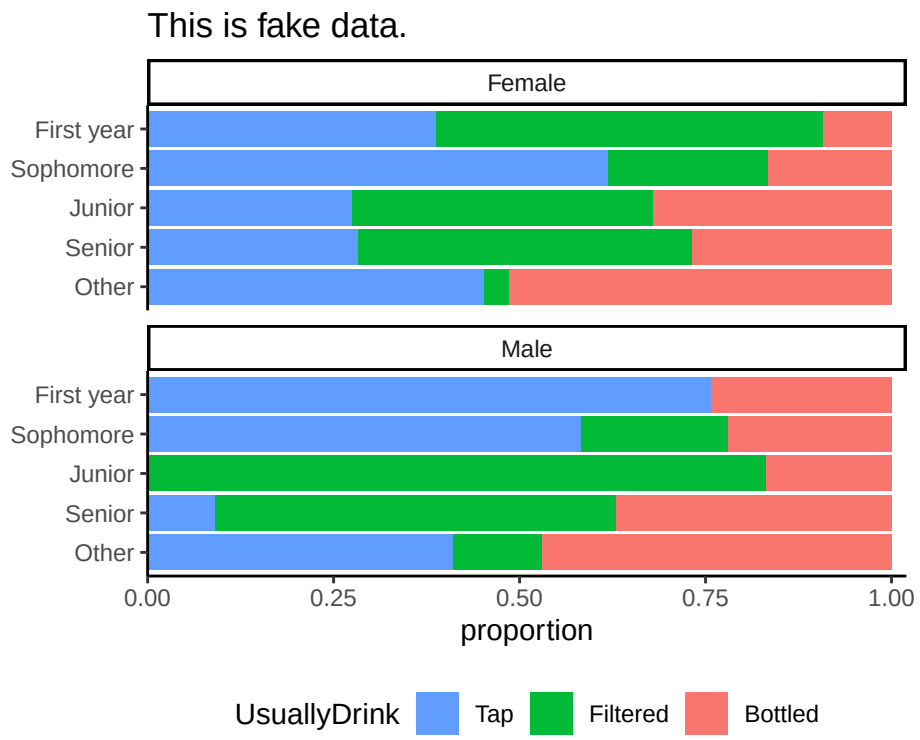
(c) Create a vertical bar chart of `Class` counts.

(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`.

(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 below. (Note that the order of the fill colors of the bars match the order of the fill colors in the legend.)



(d) Create a horizontal stacked bar chart showing the proportional breakdown of `Class` for each level of `UsuallyDrink`, faceted on `Gender`. Use a descriptive title. The order of the levels of the categories and the legend should look like this:



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?

(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.)

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

(d) Create a Cleveland dot plot of metascores.

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

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

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

(c) Hexagonal heatmap of bin counts

(d) Square heatmap of bin counts

(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?

(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?

4. Australian Institute of Sport data

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

(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?

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