EDAV Fall 2019 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. Data manipulation should not be hard coded. That is, your scripts should be written to work for new data.

1. useR2016! survey

[18 points]

Data: `useR2016` dataset in the **forwards** package (available on CRAN)

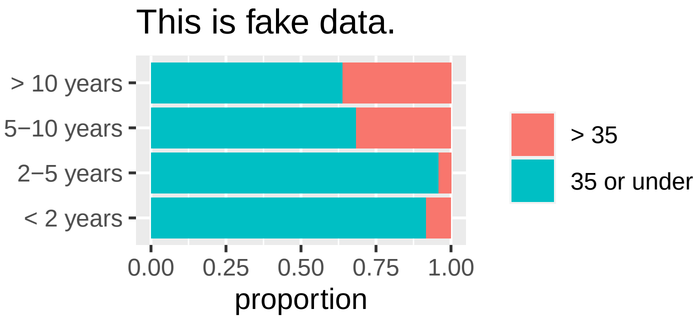
For parts (a) and (b):

- Do not toss NAs.
- Do some research to find the wording of the questions asked as relevant and include them in the titles of your graphs.
- Include the dataset name, package name, and link to the question wording source in the graph caption.

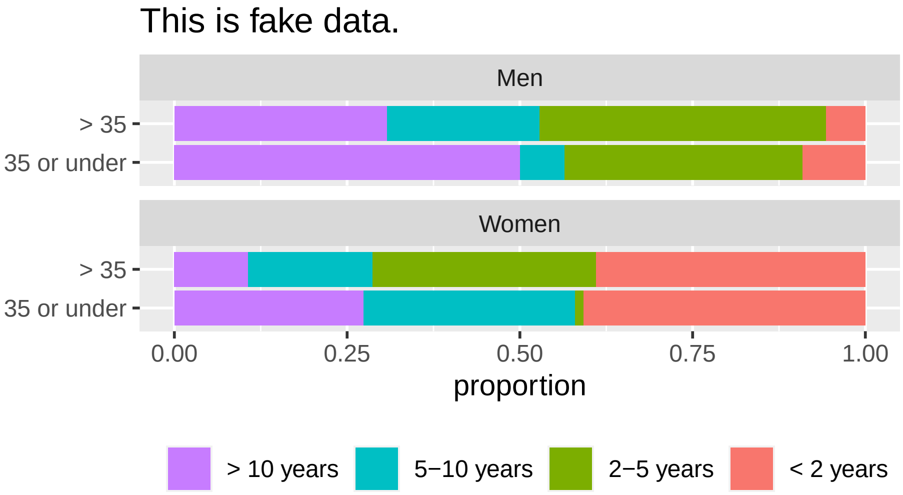
(a) Create a horizontal bar chart of the responses to Q20.

(b) Create a vertical bar chart of the responses to Q11.

(c) Create a horizontal stacked bar chart showing the proportion of respondents for each level of Q11 who are over 35 vs. 35 or under. Use a descriptive title. It should look like this:



(d) Create a horizontal stacked bar chart showing the proportional breakdown of Q11 for each level of Q3, faceted on Q2. Use a descriptive title. It should look like this:



(e) For the next part, we will need to be able to add line breaks (`\n`) to long tick mark labels. Write a function that takes a character string and a desired approximate line length in number of characters and substitutes a line break for the first space after every multiple of the specified line length. For example:

```
> x <- "We hold these truths to be self-evident, that all men are
created equal, that they are endowed by their Creator with certain
unalienable Rights, that among these are Life, Liberty and the pursuit
of Happiness."
> add_line_breaks(x, 50)
[1] "We hold these truths to be self-evident, that all men\nare created
equal, that they are endowed by their\nCreator with certain unalienable
Rights, that among\nthese are Life, Liberty and the pursuit of Happiness."
```

(f) Create a horizontal bar chart that shows the percentage of positive responses for `Q13 - Q13_F`. Use your function from part (e) to add line breaks to the responses. Your graph should have one bar each for `Q13 - Q13_F`.

2. Rotten Tomatoes

[18 points]

To get the data for this problem, we'll use the **robotstxt** package to check that it's ok to scrape data from Rotten Tomatoes and then use the **rvest** package to get data from the web site.

(a) Use the `paths_allowed()` function from **robotstxt** to make sure it's ok to scrape https://www.rottentomatoes.com/browse/box-office/. Then use **rvest** functions to find relative links to individual movies listed on this page. Finally, paste the base URL to each to create a character vector of URLs:

```
> head(links, 3)
[1] "https://www.rottentomatoes.com/m/it_chapter_two/"
[2] "https://www.rottentomatoes.com/m/hustlers_2019/"
[3] "https://www.rottentomatoes.com/m/angel_has_fallen/"
```

Display the first six lines of the vector.

(b) Write a function to read the content of one page and pull out the title, tomatometer score and audience score of the film. Then iterate over the vector of all movies using `do.call() / rbind() / lapply()` or `dplyr::bind_rows() / purrr::map()` to create a three column data frame (or tibble):

```
> head(df, 3)
# A tibble: 3 x 3
 title            tomatometer audience_score
 <chr>                  <dbl>          <dbl>
1 It Chapter Two            63             79
2 Hustlers                  88             67
3 Angel Has Fallen          39             93
```

Display the first six lines of your data frame.

(Results will vary depending on when you pull the data.)

For help, see this SO post: https://stackoverflow.com/questions/36709184/build-data-frame-from-multiple-rvest-elements

Write your data to file so you don't need to scrape the site each time you need to access it.

(c) Create a Cleveland dot plot of tomatometer scores.

(d) Create a Cleveland dot plot of tomatometer *and* audience scores on the same graph, one color for each. Sort by audience score.

(e) Run your code again for the weekend of July 5 - July 7, 2019. Use **plotly** to create a scatterplot of audience score vs. tomatometer score with the ability to hover over the point to see the film title.

3. Weather

[14 points]

Data: `weather` dataset in **nycflights13** package (available on CRAN)

For parts (a) - (d) draw four plots of `wind_dir` vs. `humid` 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 data, using the "Movie ratings" example on page 82 (last page of Section 5.3) as a guide.

(f) Draw a scatterplot of `humid` vs. `temp`. Why does the plot have diagonal lines?

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

(h) Color the points by `origin`. Do any new patterns emerge?