Warmup 2: Data Visualization

Stat 133, Spring 2022

Associated readings

We are assuming that you have read the papers listed below. These are part of the reading materials for weeks 5 and 6. The corresponding pdf files are in bCourses, section **Files**, inside folder **readings**:

- *Effectively Communicating Numbers*, by Stephen Few.
- *How to Display Data Badly*, by Howard Wainer.

General Instructions

- Write your narrative and code in an `Rmd` (R markdown) file.
- Name this file as `warmup02-first-last.Rmd`, where `first` and `last` are your first and last names (e.g. `warmup02-gaston-sanchez.Rmd`).
- Please do not use code chunk options such as: `echo = FALSE`, `eval = FALSE`, `results = 'hide'`. All chunks must be visible and evaluated.

1) Graphic from FiveThirtyEight

FiveThirtyEight is an American website that focuses on opinion poll analysis, politics, economics, and sports blogging.

https://fivethirtyeight.com/

Visit this website and look for one graphic that catches your attention. It can be almost any graphic **except** the ones discussed in lecture and/or the ones used in some of the slides/videos about data visualization (week 6).

You will have to provide an assessment of the chosen graphic based on the following aspects:

a) Description/explanation of its context:
    - What is the data—e.g. individuals & variable(s)—behind the graph?
    - Is there a time period associated to it?
    - What is the type of graphic (e.g. barchart, piechart, timeline, histogram, map, heatmap, etc)?

b) What color scheme (if any) is being used in the graphic?

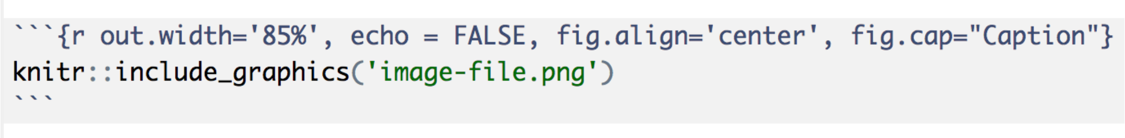
c) Taking into account the so-called “Data-Ink ratio”, explain whether the graphic seems to be maximizing this ratio or not.

d) Describe the things that you find interesting about the chosen graphic

- Is it the colors?
- Is it the visual appearance?
- Is it the way in which data has been encoded graphically?
- Is there anything that catches your attention?

e) To include a screenshot of the graphic in your report, we suggest using the function `include_graphics()` from `"knitr"`. This function gives you more control on the appearance of the graphics in your html document. See figure below with a hypothetical example with the following code-chunk options:

- `out.width='85%'` allows you to control the width of the figure with respect to the html output. There’s also `out.height`
- `fig.align` allows you to control the figure alignment (left, right, etc)
- `fig.cap` lets you include captions



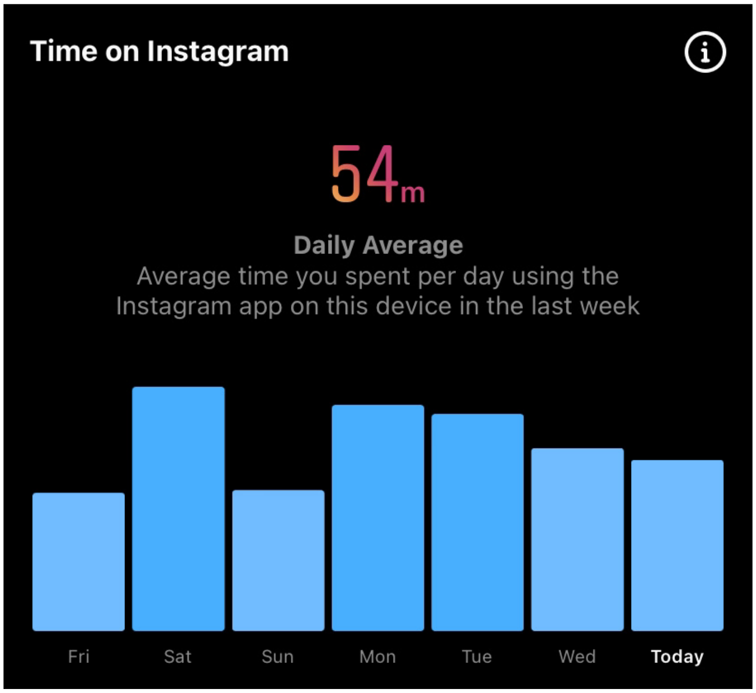
~~~
```{r out.width='85%', echo = FALSE, fig.align='center', fig.cap="Caption"}
knitr::include_graphics('image-file.png')
```
~~~

f) Also, include a link of the graphic’s webpage, and the names of the authors/designers of the related article.

Instructions continue in next page.

2) Instagram Graph

Consider the following graphic from the Instagram app (in a user's smartphone):



Time on Instagram

54m

Daily Average

Average time you spent per day using the Instagram app on this device in the last week

This graphic is a bar-chart from Instagram presenting the app's usage over a given week. Each bar is supposed to represent the number of minutes spent on Instagram in that day.

2.1) Graphic's Assessment

Provide an assessment of the above bar-chart, describing the *good* and the *bad* about this data visualization.

2.2) Replicate Instagram's Bar-chart

The following table contains the data of the bar-chart (minutes spent on Instagram's app by day):

Day	Time
Fri	40
Sat	70
Sun	40
Mon	65
Tue	62
Wed	52
Thu	48

Write code in R to replicate, as much as possible, the visual appearance of the Instagram barchart.

2.3) Improved Alternative Instagram Visualization

Write code to generate a graphic that improves the visualization of the original Instagram bar-chart.

Also, provide a description/explanation of how your proposal improves the original chart.

3) Cal's Football Game-by-Game Statistics

The data for this section is in the file `cal-games.csv` (available in bCourses, in the same folder containing this `pdf` file).

Download a copy of the CSV file, and place it in your computer in the same folder where you have your `Rmd` file for this assignment.

Here's a suggestion for importing the table in a data frame called `games`

```r
# import using 'read.csv()'
games = read.csv(
  file = "cal-games.csv",
  stringsAsFactors = FALSE,
  colClasses = c(
    "Date",
    "character",
    "character",
    "numeric",
    "numeric",
    "numeric",
    "numeric"
  ))
```

The data has to do with Cal's Football Team, and it contains game-by-game statistics from seasons 2010 to 2021. Statistics from 2020 were not included because of the Covid-19 pandemic effects on that season.

The data set contains seven variables:

- `date`: date of game
- `opponent`: name of opponent team
- `home_away`: whether the game was at-home or away
- `cal_score`: Cal's score
- `opp_score`: Opponent's score
- `duration`: Duration of the game (in minutes)
- `attendance`: Game's attendance (number of people attending a game)

Consider the following parts (a) to (c). Based on the provided data set, create a data visualization that allows you to address and answer each of these parts. Make sure to follow good practices/habits for creating data visualizations.

a) Does playing at home give Cal's football team an advantage over its opponents?

b) If we focus on those games in September, October, and November, which month(s) tend to be a winning month for Cal? *Hint*: we recommend using the `month()` function from package `"lubridate"` to extract the month of a date-time object

c) Create a data visualization that uses the following variables/information:

  - Cal's score
  - Opponent's score
  - Date (e.g. can be year, and/or month, or full date)
  - `home_away` status

Hint: we recommend using the `year()` function from package `"lubridate"` to extract the year of a date-time object