HW 3: Data Visualization

Stat 133, Spring 2023

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 `hw3-first-last.Rmd`, where `first` and `last` are your first and last names (e.g. `hw3-gaston-sanchez.Rmd`).
- You will have to knit your `Rmd` file into an HTML file.
- Submit both your Rmd and HTML files to the corresponding assignment submission in bCourses.
- Please note that submitting only one of the files will result in an automatic 20% deduction.
- Also, if you submit the incorrect files you will receive no credit.

Problems start in next page.

1) Look for a Graphic

Look for a graphic from one of the websites listed below. To be clear: you have to find only one graphic from any of the listed websites.

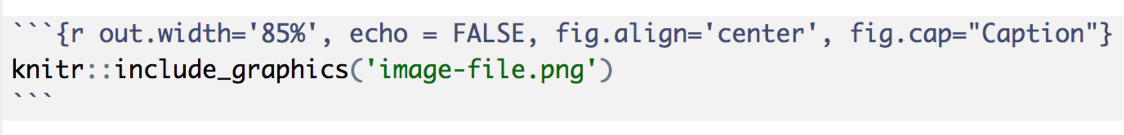
- **FiveThirtyEight**: https://fivethirtyeight.com/
- **Vox:** https://www.vox.com/
- **Politico**: https://www.politico.com/

Look for one graphic that catches your attention **except** the ones from FiveThirtyEight discussed in lecture by Prof. Sanchez (i.e. those in the lecture slides about data visualization).

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

a) Include a screenshot of the graphic in your report. To do this, 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 the 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 a caption
   - `echo = FALSE` prevents the code chunk to be displayed in the HTML doc



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

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

c) 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)?
   - Is the graphic static? Or is it dynamic/interactive?

d) What **color scheme** or **color palette** (if any) is being used in the graphic?

e) Is the graph **"colorblind friendly"**? You can take a look at the following resources that give you guidelines to choose colorblind safe colors:

- https://www.datylon.com/blog/data-visualization-for-colorblind-readers#:~:text=The%20first%20rule%20of%20making,out%20of%20these%20two%20hues.
- https://davidmathlogic.com/colorblind/#%23D81B60-%231E88E5-%23FFC107-%23004D40
- https://colorbrewer2.org/#type=sequential&scheme=BuGn&n=3

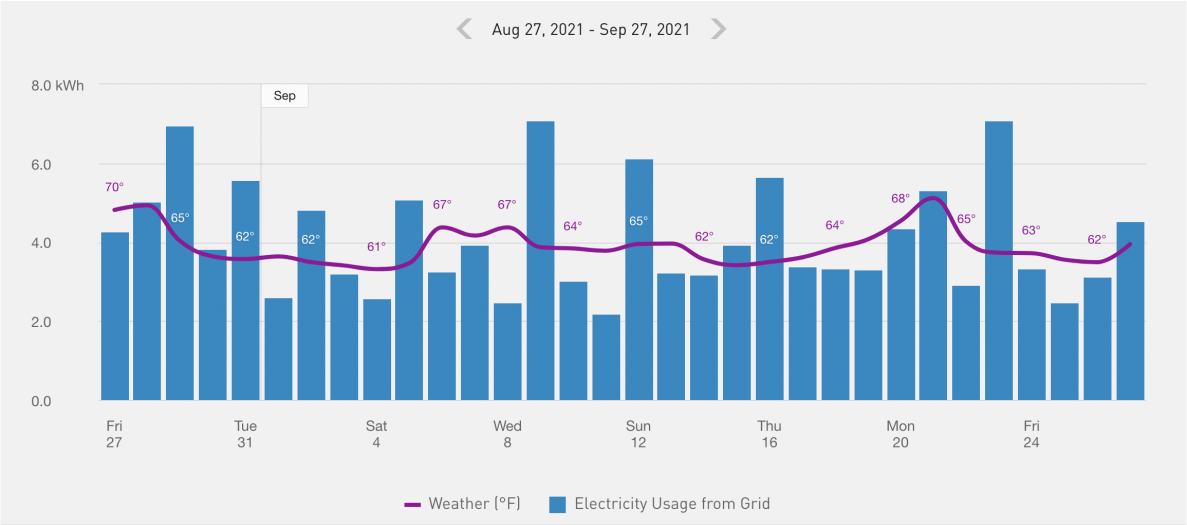
f) Taking into account the so-called **Data-Ink ratio**, explain whether the graphic seems to be maximizing this ratio or not.

g) 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?

2) Electricity Usage Graph

Consider the following graphic of *Electricity Usage* from a customer of PG&E:



This graph comes from PG&E and it shows the electricity usage per day from August 27th till September 27th, 2021. The data consists of time, weather (in Fahrenheit), and the electricity usage from grid (in kilo Watts per hour).

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) Graphic Replication

The following code contains the data behind the PG&E graphic.

```
# PG&E Data
Date = seq(as.Date("2021-08-27"), as.Date("2021-09-27"), by="days")

Usage = c(4.3, 5, 7, 3.8, 5.6, 2.6, 4.8, 3.2, 2.6, 5.1,
          3.3, 3.9, 2.5, 7.1, 3, 2.2, 6.1, 3.2, 3.2, 3.9,
          5.6, 3.4, 3.3, 3.3, 4.4, 5.3, 2.9, 7.1, 3.3, 2.5, 3.1, 4.5)

Weather = c(70, 71, 65, 63, 62, 63, 62, 61, 61, 62,
            67, 66, 67, 64, 64, 64, 65, 65, 62, 61,
            62, 63, 64, 65, 68, 72, 65, 63, 63, 62, 62, 65)

pge = data.frame(Date, Usage, Weather)
head(pge, n = 10)
```

```
##          Date Usage Weather
## 1  2021-08-27   4.3      70
## 2  2021-08-28   5.0      71
## 3  2021-08-29   7.0      65
## 4  2021-08-30   3.8      63
## 5  2021-08-31   5.6      62
## 6  2021-09-01   2.6      63
## 7  2021-09-02   4.8      62
## 8  2021-09-03   3.2      61
## 9  2021-09-04   2.6      61
## 10 2021-09-05   5.1      62
```

Write code in R to replicate, as much as possible, the visual appearance of the PG&E graphic.

3) Visualizing Household Expenditures

The data for this section has to do with the monthly expenditures of a family from a large city in the East Coast of the United States. The associated time-period goes from July 2018 to December 2022.

The actual data table is in the file `expenses.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 into a data frame called `dat`. You can use other approaches to import the data if you want.

```
dat = read_csv(
  file = "expenses.csv",
  col_types = list(
    concept = col_character(),
    amount = col_double(),
    date = col_date(format = "%m/%d/%y"),
    category = col_character(),
    mode = col_character()
  ))
```

There are 5 variables:

- `concept`: type of expenditure
- `amount`: amount (in dollars)
- `date`: date (mm/dd/yy)
- `category`: expenditure category
- `mode`: payment mode (debit, credit)

The categories in variable `category` can be divided into 2 buckets:

1) **Cost of living** categories:
    - `insurance`
    - `housing`
    - `utilities`
    - `food`
    - `transportation`
    - `healthcare`
    - `debt`
2) **Discretionary spending** categories:
    - `recreation`
    - `personal`
    - `gifts`
    - `vacation`

Some Comments:

- Most expenses are organized by month and year, although not necessarily by day. Also, there may be some expenses with the correct date but placed within expenses of another month.

3.1) Data Visualizations

Based on the provided data set:

a) Create a visual display that lets us see the total monthly expenditures over the provided months/years.

b) Create a visualization to display the top-5 annual expenditure categories in every year.

c) Create a plot that lets us compare, over time, the **cost-of-living** spending versus the **discretionary** spending.

Make sure to follow good practices/habits for creating data visualizations. In other words: apply what you've learned in lecture in weeks 4-6. For example (but not limited to):

- include a title
- include a subtitle
- label your axes
- make sure any legend labels are self-descriptive; even better try to avoid so-called "long-distance labeling"
- take advantage of pre-attentive features (e.g. help the viewer with their "early vision" as much as possible)
- make sure that the text of axis scales is viewer-friendly (e.g. avoid displaying large numbers in scientific notation such as `1e+06, 2e+06, 2e+06, ...`)
- choose an adequate color scheme or color palette, but don't overdo it
  - e.g. more often than not, a "rainbowy" palette is unnecessary
  - keep in mind that some viewers in your audience may have some type of color-blindness
- possibly, you may want to include annotations (specially if there's anything important to be highlighted)
- strive for creating graphics with a large data-to-ink ratio

Last but not least, include a description/interpretation of what's going on in each of your graphics.