PUBH 7462 Homework 2

Due 2/10/2022

General Expectations

Throughout the assignment, please:

- Use meaningful file names ("_" or "-" seperated)
- Use meaningful variable names
- 'Good' `R` style (white space, etc.)
- Consistent style (choose a style and stick to it)
- Appropriate titles, axes labels, and legend titles/group names
- Get into the habit of commenting your code chunks

- Use relative paths, ex. "./data"

With respect to the knitted .RMD:

- Omit extra output (anything from `R` with ## for example)
- Make sure your code chunks are visible with `echo = TRUE`
- Make sure your inline R works properly and `round()` digits

With respect to Github:

- Make sure your repository is public so the TA and myself can view it

- Keep the repository 'tidy' and well organized
- Use meaningful filenames such that a stranger who happened upon the repository could surmise what's going on

With respect to data visualizations in general:

- **Never output a tibble() or data.frame() as a table**
  - Please use `df %>% knitr::kable(caption = "")`
  - or `df %>% gt() %>% tab_header("")`
- Remember, a *good* data visualization should be self-explanatory
- This means that I shouldn't need to read your code to know what's going on in the plot
- I find it useful to imagine your audience knows little to nothing about what you're doing prior to seeing the plot (as is often the case)

Problem 1. Github repository (10pts)

- Please set up a repository named `pubh7462_hw2_`*your-email -handle*
- Connect to it Rstudio with an .Rproj
- Create a `/data` folder, add to the `.gitignore`
- Include all necessary .md figure files
- Keep the repository 'tidy', no extra files or folders

Problem 2. Best Practices and Consistent Style (20pts)

Problem 3. BRFSS SMART 2002-2010 (70pts)

Data from the Behavioral Risk Factors Surveillance System (BRFSS) for Selected Metropolitan Area Risk Trends (SMART) for 2002-2010 were were accessed from data.gov. The version we will be using can be found on Canvas here – **please download the brfss_smart_2010.csv and put it into the /data folder**

Background

2002-2010. BRFSS SMART County Prevalence land line only data. The Selected Metropolitan Area Risk Trends (SMART) project uses the Behavioral Risk Factor Surveillance System (BRFSS) to analyze the data of selected counties with 500 or more respondents. BRFSS data can be used to identify emerging health problems, establish and track health objectives, and develop and evaluate public health policies and programs. BRFSS is a continuous, state-based surveillance system that collects information about modifiable risk factors for chronic diseases and other leading causes of death. Data will be updated annually as it becomes available. Detailed information on sampling methodology and quality assurance can be found on the BRFSS website, and general Methodology. Glossary.

§ 3.1 Data Exploration & Cleaning (10pts)

The focus of this problem is going to be regarding **Overall Health** at the state and county level. Before doing anything else, recall that we all download the `DataExplorer` package. Check out this brief CRAN vignette and use a few of the EDA functions to familiarize yourself with the data.

- **Do not include any DataExplorer code or output in your knitted .RMD**
- This is is just good practice to familiarize yourself with new, raw data in an easy-to-read way
- Some of my go-to favourites are
  - `introduce()`
  - `plot_intro()`
  - `plot_str()`
  - `plot_missing()`
  - `plot_bar()` and `plot_histogram()`

After a brief exploration, please read and clean the data by –

- 'Tidying' up variable names
- Retaining only the *Overall Health* `topic`
- Retain only the **year, state, county, response, sample size, and prop. of responses (data_value)**
  - *Note* county needs to be extracted from another variable
- Renaming key variables as appropriate to ensure they are informative
- Transforming `numerics characters` or `strings` to `factors` *where it's appropriate*

§ 3.2 Data Description (10pts)

Please describe these data using `inline r`, including but not limited to

- Number of observations & variables
- Case definition of an observation
- What each variable describes (including units if applicable)
- Any other interesting information you think the audience should know

§ 3.3 Do Data Science (50pts)

Please answer the following questions using what we've learned about data wrangling in `tidyverse` and your existing `ggplot` plot knowledge (refer to lecture examples). Remember the **General Expectations** above when it comes to programming, plots and tables.

§ 3.3.1 In the year 2004, which states were observed at 6 locations? (10pts)

Hint `distinct()`

§ 3.3.2 Make a "spaghetti plot" that shows the number of observed locations in each state from 2002 to 2010. Which state has the highest mean number of locations over this period? (10pts)

- Please order the legend/states in a meaningful way (not alphabetically)
- Comment on the general temporal trends you observe

Hints
- A "spaghetti plot" is collection of line plots in a single frame (geom_line or geom_smooth)
- You could/should be able to tell which state has the highest by the visualization alone

§ 3.3.3 Make a table showing, for the years 2002, 2006, and 2010, the mean and standard deviation of *sample size* and *proportion* of *Excellent*, *Good*, and *Poor* responses across locations in MN. (15pts)

- Please use `knitr::kable(caption = "")` or `gt::gt() %>% tab_header("")`
- It's okay to leave the variable names in 'tidy' format (i.e. prop.mean or prop_mean)
- Comment on any trends you notice by year or response type.

Hints - `dplyr::across()`
- You may need to handle some missing proportion of responses with `na.rm`

§ 3.3.4 Create a `ggplot` that communicates the results/trends from the table above and *stands on its own* (15pts)

Hints
- Probably want to reshape the data first `tidyr`
- `stringr` (names) and `forcats` (order)