PUBH 7462 Homework 2

Due 2/3/2022

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
- In Problem 1, below, for example:
  - `indicatorTRUE` is *not* an appropriate name for a legend group, "Yes" or "True" are
  - `sum_indicator` is *not* an appropriate legend title, "X + Y > 0.5" is
  - `x`, `y` are *not* informative axes label,; "X ~ N(0, 1)" or "Random Sample N(0, 1)" are
  - *Hint:* In can be useful to use `mutate()` or `rename` to manipulate names/variables/categories prior to plotting
- Get into the habit of commenting your code chunks

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` (default in the setup)
- Make sure your inline R works properly and `round()` digits as appropriate to avoid things like "mean of X was observed to be 1.1234234634576982304 or 1.213451e-10"

With respect to data visualizations in general:

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

Problem 2.1 Independent Bivariate Normal Random Sample (25pts)

- Create a `tibble()` (data frame) with 3 variables:
  - `x =` a random sample of size `n = 1000` from $N(0, 1)$ (`rnorm()`)
  - `y =` a random sample of size `n = 1000` from $N(1, 2)$
  - `sum_indicator =` a logical variable, `TRUE` if `x + y > 0.5` (hint: `ifelse()`)
- Prior to generating the plot below, please `mutate` the `sum_indicator` variable to be a factor with levels `Yes` and `No`. In addition, utilize `forcats::fct_relevel()` to reorder the factor such that `Yes` comes before `No` (the default is alphabetical ordering).
- Use `ggplot` to create a scatter plot of $y \sim x$, coloured by the `sum_indicator` variable (`scale_colour_...`), with appropriate titles, axes labels, and legend title/category names

Problem 2.2 Penguin EDA (25pts)

In this problem, you will perform a brief, guided exploratory analysis of the `palmerpenguins` data set. These data were collected by Dr. Kristen Gorman and the Palmer Station, Antarctica LTER. The .RDS file can be downloaded here. Please create a /data folder and put `penguin.RDS` in there. We may now all load it with the same *relative path* below -

```
#Read data with relative path
penguin.df <- read_rds("./data/penguin.RDS")
```

Problem 2.2.1 Data Description

Using inline R, please describe these data including the following aspects:

- The "case definition", i.e. what each observation (row) details
  - Documentation may be found on CRAN here
  - We are working with the `penguins` data set, **not** the `penguins_raw` data set
- The number of observations (rows), variables/features (columns), and what each variable/feature describes about each observation.
- The mean flipper and bill length, respectively, and their associated standard deviation(s)
  - You may need to handle missing values with `na.rm = TRUE` option

Problem 2.2.2 Visualization

Using ggplot:

- Make a scatterplot of flipper length (y) by bill length (x), coloured by species
- Add on to this plot, and generate a new one, by separating the scatterplot into panels by `sex` with `+ facet_wrap()` or `+ facet_grid()`
- Briefly comment on any trend(s) you observe
- *Note:* make sure there's no extra output (i.e. you may need to utilize something like `message = FALSE` or `warning = FALSE` in the chunk options)