Individual homework 3

Please push your assignment to GitHub by 10:00pm (Central) Sunday, Apr 14.

You are currently in the GitHub repository (repo) for `hw3-username`. The assignment prompt is shown below (i.e. in `README.Rmd`). You can view this online in your homework 3 GitHub repository as a Markdown file(`README.md`) or a pdf.

Please **use `hw3.Rmd` to complete this assignment**. Be sure to **knit your file to PDF before your final push to GitHub**.

Homework process

For help on the homework process, review

- Assignments in Stat 220 for content/formatting questions.
- GitHub Guide for Students in Stat 220 for Git and Github instructions.

When you are done with your homework, **don't forget to push your changes to ALL files back to GitHub!** This means you should commit and push all related files, not just your final PDF. Additionally, ensure you post the link to your **GitHub repository to Gradescope for the final submission and grading**. This step is crucial as it allows for a comprehensive review of both your code and the rendered output.

Assignment prompt

Problem 1: babynames

Use the `babynames` dataset and `dplyr` to answer the following:

a.

Create a data set `babynames_hailey` that contains data on female babies named `Hailey` and some other variations including `Hailee`, `Haleigh`, `Haley` or `Hayleigh`. Create one graph that shows the how the number of babies with each of these five names varies over time. Summarize what this graph shows in 1-2 sentences.

b.

Historically, which name is most balanced between the number of males and females with that name aggregated over all years? To answer this, aggregate `babyname` name counts for each name over all years by sex, then filter to names that have at least 10,000 occurrences by sex. For each name, compute the proportion of male and female names and find which name (or names) have male/female proportions closest to 0.50.

Note: Filtering to "common" names avoids finding rare names that, for example, have 1 instance of a male and 1 of a female (which would be 50% male and 50% female).

c.

For the names identified in part (b) as most balanced between males and females, find the time period during which these names were the most popular. Plot the number of babies with these names by year and sex.

d.

Find the names that have consistently increased in popularity from 1950 to 2012. Create a graph that shows the trends of these names over time.

Hint: you may need to use `lag()` function that allows you to access and manipulate previous observations in a sequence. `lag()` function shifts the data by one position, padding the beginning with an `NA`. So, when manipulating data using `lag()`, it's advisable to use `tidyr::drop_na()` post-lag operation to get rid of the `NA`'s.

Problem 2: Consumption

The data set `food_consumption.csv` was compiled from the website nu3 and contains the following measurements from 2018:

variable	class	description
country	character	Country Name
food_category	character	Food Category
consumption	double	Consumption (kg/person/year)
co2_emmission	double	Co2 Emission (Kg CO2/person/year)

```
food_consumption <- read.csv("https://raw.githubusercontent.com/deepbas/statdatasets/main/foodconsumpti
```

Use `dplyr` to answer the following questions.

a.

Create a data frame that shows the 3 food category types with the highest total food consumption (per person) in 2018.

b.

For each country, compute the percentage of consumption for each food category out of that country's total food consumption (per person). Then create a data frame that shows the 5 countries with the highest proportion of total consumption in the category `Milk - inc. cheese`.

c.

For each country, compute total food consumption and total CO2 (per person) across all categories, then create a scatterplot of total CO2 vs. total consumption. Label points by country name using the `geom_text_repel` from the `ggrepel` package (which you may need to install). Describe the trend you see.

Problem 3: restaurant violations

The data set `Violations` contains info about the outcome of health inspections of restaurants in New York City. See `?mdsr::Violations` for more details.

```
data("Violations", package = "mdsr")
```

a.

Use `dplyr` to construct a data frame with the following:

- only cases from Manhattan (`boro`)
- includes the median violation score (`score`) by zip code (`zipcode`)
- includes the number of inspections by zip code
- only includes zip codes with with 50 or more inspections

Note: The first line of your pipe should be `tidyr::drop_na(score)` to remove rows with missing values for `score`. `drop_na` is from the `tidyr` package.

b.

Create and interpret a `ggplot2` that shows the relationship between the number of inspections (x) and the median score (y). Use both the `point` and `smooth` geometries in your graph and add an information title and axes labels to your plot.

c.

Now, let's calculate the mean and standard error for the violation scores across all the zip codes of Manhattan. Then, create and interpret a `ggplot2` plot that shows the relationship between the zip code (x) and the ordered mean violation scores (y), along with error bars representing one standard error. Also based on this plot, identify the neighborhood with the highest and lowest mean violation scores.

Problem 4: joins

The data set below called `Students` contains information on five students with their ID number, first name and computer preference.

Id	Name	Computer
1	Arya	m
2	Gregor	m
3	Cersei	w
4	Jon	m
5	Jon	w

The data set below called `Classes` contains the roster information (student first name and ID) for two classes.

Class	Student	Stud_Id
CS	Jon	4
CS	Arya	1
CS	Cersei	3
Stats	Gregor	2
Stats	Jon	4
Stats	Jon	5
Stats	Arya	1

What data set will be produced by the following commands? Describe the data set in words and show what it looks like using an R Markdown table to display the new data set.

a.

```
left_join(Classes, Students, by = c("Stud_Id" = "Id"))
```

b.

```
CS <- Classes %>% filter(Class == "CS")
Stats <- Classes %>% filter(Class == "Stats")
semi_join(Stats, CS, by = "Stud_Id")
```

c.

```
anti_join(Stats, CS, by = "Stud_Id")
```

Problem 5: restructure

Consider the `Lakes_wide` data set below that records lake clarity (in meters) for 2012 through 2014.

LakeId	2012	2013	2014
1	6.5	5.8	5.8
2	2.1	3.4	2.8

What data set will be produced by the following commands? Describe the data set in words and show what it looks like using an R Markdown table to display the new data set.

a.

```
Lakes_wide %>%
  pivot_longer(
    cols = 2:4,
    names_to = "Year",
    values_to = "Clarity"
  )
```

b.

```
Lakes_wide %>%
  pivot_longer(
    cols = 2:4,
    names_to = "Year",
    values_to = "Clarity"
  ) %>%
  group_by(LakeId) %>%
  arrange(Year) %>%
  mutate(Change_in_Clarity = Clarity - lag(Clarity))
```