STAT 201: Problem Set 3 (R)

Your name

3/3/25

In January 2017, Buzzfeed published an article titled "These Nobel Prize Winners Show Why Immigration Is So Important For American Science". In the article they explore where many Nobel laureates in the sciences were born and where they lived when they won their prize.

In this homework we will work with the data about Nobel laureates to recreate/update some of their visualizations with new data as well as explore new questions.

```
library(readr)
# add more packages here as necessary


url_file <- "https://raw.githubusercontent.com/midd-stat201-spring2025/midd-stat201-spring
nobel <- read_csv(url_file, name_repair = "unique_quiet")
```

0. Change your name in the YAML and add the package(s) necessary for creating ggplots and wrangling data in the code chunk above. Then run the code chunk to load in the data.

A description of the variables in the `nobel` data are as follows:

- `id`: unique identifier of laureate
- `firstname`: first name (and possible middle initial) of laureate
- `surname`: last name/surname
- `year`: the year the prize was awarded
- `category`: category of prize (Chemistry, Economics, Literature, Peace, Physics, or Medicine)
- `born_year`: year laureate was born
- `died_year`: year laureate died
- `affiliation`: affiliation of laureate at time of winning
- `city`: city of laureate in prize year
- `country`: country where laureate was based in prize year

- `gender`: gender or laureate (male, female, or org, where org represents an organization)
- `share`: reciprocal of the portion of prize awarded to the laureate
- `motivation`: motivation for recognition

1. Display a summary table of the sample average and standard deviation of the ages of Nobel laureates at the time of receiving the prize. Do this in a single pipeline by:

- Creating a new variable that represents the age of the laureate when they one their prize, calculated as the year they received the award minus the year they were born
- Filtering to only retain observations for which your newly calculated age variable is available
- Writing code to actually create the summary statistics

Be sure to explicitly set/define the column titles of your summary table. Then interpret these statistics (particularly the standard deviation) in context.

Answer:

2. Create a new data frame called `nobel_living` that only retains cases from the original data frame that meet the following criteria:

- laureates for whom `country` is available
- laureates who are people as opposed to organizations
- laureates who are still alive

Use code to confirm that you have 21 female and 222 male laureates in your new data frame:

3. **Buzzfeed's Claim #1: "Most living Nobel laureates were based in the US when they won their prizes".** Let's see if that's true.

Modify (i.e. store/assign over) your `nobel_living` data frame by with a new version that as an additional variable called `country_base`. The variable should equal:

- "USA" if the laureate was based in the USA when they won
- "Other" if the laureate's was based in the USA when they won

You will have to use the `if_else()` function. Take a look at its Help file (and in particular, its examples).

Now would be a good time to render your work to save the progress and make sure everything is working!

4. Create a new data frame called `nobel_living_science` that only retains observations with laureates from the Physics, Chemistry, Medicine, and Economics categories from the `nobel_living` data frame.

5. Using the data frame `nobel_living_science`, create a faceted bar plot with horizontal bars that visualizes the relationship between 1) the category of prize and 2) whether the laureate was in the US when they won the Nobel prize. Note: Your visualization should be faceted by category. For each facet you should have two bars, one for winners based in the US and one for Other.

Interpret your visualization, and say a few words about whether the Buzzfeed Claim #1 is supported by the data.

Answer:

Now would be a good time to rebder your work to save the progress and make sure everything is working!

6. **Buzzfeed's Claim #2: "But of those US-based Nobel laureates, many were born in other countries."** Let's investigate this second claim!

The following code reads in another dataset that has information about the birthplaces of these Nobel laureates. We store the resulting data frame as `born_info`. Run the following code chunk (no need to do anything here). The variables in the data frame are:

- `ID`: unique identifier of laureate
- `born_city`: city where laureate was born
- `born_country`: country where laureate was born
- `born_country_code`: two-letter country code of `born_country`

```
url2 <- "https://raw.githubusercontent.com/midd-stat201-spring2025/midd-stat201-spring2025
born_info <- read_csv(url2)
```

Now, add the information from `born_info` into the `nobel_living_science` data frame, retaining only observations from `nobel_living_science`. Store the resulting data frame as a new one called `nobel_living_science2`.

Then create a new variable called `born_country_us` that has the value "USA" if the laureate is born in the US, and "Other" if not. Be sure to save the variable to the data frame by storing the output back into `nobel_living_science2`.

7. Let's improve on our previous visualization here. Visualize the relationship between where the laureate was based when they won the Nobel Prize and where they were born, split by category. Your final visualization should:

- contain a facet for each category
- within each facet, bars for whether the laureate won the award in the US or not
- within each bar, display whether the laureate was born in the US or not

Based on your visualization, do the data appear to support Buzzfeed's Claim #2? Explain your reasoning in a few sentences.

Answer:

8. We will explore where the non-US laureates were born. In a single pipeline starting with `nobel_living_science2`, filter for laureates who were living in the US when they won their prize but where born outside of the US. Then create a frequency table for their birth country, only displaying the countries with at least three laureates. Display your order in descending order of number of laureates.

Once you're finished, render once more.