STAT 201: Problem Set 3 (R)

Your name

2025-09-29

Unless explicitly stated, you do not need to store the result of data wrangling as new variables in `R`.

In this homework we will work with the data about Nobel laureates to learn about how the prizes are distributed across different variables.

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


url_file <- "https://raw.githubusercontent.com/midd-stat201-spring2025/midd-stat201-spring202
nobel <- read_csv(url_file, name_repair = "unique_quiet")
```

0. Change your name in the YAML and add *only* 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 won 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

Create a frequency table to confirm that you have 21 female and 222 male laureates in your new data frame:

3. **Buzzfeed published an article in 2017 claiming: "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 with a new version that has 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 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.

Interpret your visualization, and say a few words about whether Buzzfeed's claim is supported by the data.

Answer:

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

6. Let's return to the general data set `nobel` again. Display a data frame that shows the five institutions that have produced the most Nobel laureates, along with the number of winners. What do you notice about these institutions?

Answer:

7. Finally, let's do some EDA about the distribution of female Nobel Laureates.

In the first chunk below: Using the original data set and focusing on female Nobel Laureates, display a summary table that shows: for each prize category, 1) the total number and 2) the average age at winning (for whom we have information about age). *Hint: the function n() might be helpful. Take a look at its Help file!* Order your summary table from most to least winners.

In the second chunk below: find/display the number of unique female Nobel Laureates.

Comment briefly on some findings.

Answer:

Render one last time, and then submit the rendered PDF alongside the written portion to Gradescope!