Lab 03 - Nobel laureates

Tina Huynh

Contents

In January 2017, Buzzfeed published an article on why Nobel laureates show immigration is so important for American science. You can read the article here. In the article they show that while most living Nobel laureates in the sciences are based in the US, many of them were born in other countries. This is one reason why scientific leaders say that immigration is vital for progress. In this lab we will work with the data from this article to recreate some of their visualizations as well as explore new questions.

1 Learning goals

- Replicating published results
- Data wrangling and visualisation

2 Lab prep

Read the Buzzfeed article titled *These Nobel Prize Winners Show Why Immigration Is So Important For American Science.* We will replicate this analysis in the workshop so it's crucial that you're familiar with it ahead of time.

3 Getting started

Go to the course GitHub organization and locate your lab repo, which should be named `lab-03-nobel-laureates-YOUR_GITH`
Grab the URL of the repo, and clone it in RStudio. First, open the R Markdown document `lab-03.Rmd` and Knit it. Make sure it compiles without errors. The output will be in the file markdown `.md` file with the same name.

3.1 Warm up

Before we introduce the data, let's warm up with some simple exercises.

- Update the YAML, changing the author name to your name, and **knit** the document.
- Commit your changes with a meaningful commit message.
- Push your changes to GitHub.
- Go to your repo on GitHub and confirm that your changes are visible in your Rmd **and** md files. If anything is missing, commit and push again.

3.2 Packages

We'll use the **tidyverse** package for much of the data wrangling. This package is already installed for you. You can load them by running the following in your Console:

```
library(tidyverse)
```

3.3 Data

The dataset for this assignment can be found as a CSV (comma separated values) file in the `data` folder of your repository. You can read it in using the following.

```
nobel <- read_csv("data/nobel.csv")
```

The variable descriptions are as follows:

- `id`: ID number
- `firstname`: First name of laureate
- `surname`: Surname
- `year`: Year prize won
- `category`: Category of prize
- `affiliation`: Affiliation of laureate
- `city`: City of laureate in prize year
- `country`: Country of laureate in prize year
- `born_date`: Birth date of laureate
- `died_date`: Death date of laureate
- `gender`: Gender of laureate
- `born_city`: City where laureate was born
- `born_country`: Country where laureate was born
- `born_country_code`: Code of country where laureate was born
- `died_city`: City where laureate died
- `died_country`: Country where laureate died
- `died_country_code`: Code of country where laureate died

- `overall_motivation`: Overall motivation for recognition
- `share`: Number of other winners award is shared with
- `motivation`: Motivation for recognition

In a few cases the name of the city/country changed after laureate was given (e.g. in 1975 Bosnia and Herzegovina was called the Socialist Federative Republic of Yugoslavia). In these cases the variables below reflect a different name than their counterparts without the suffix '_original'.

- `born_country_original`: Original country where laureate was born
- `born_city_original`: Original city where laureate was born
- `died_country_original`: Original country where laureate died
- `died_city_original`: Original city where laureate died
- `city_original`: Original city where laureate lived at the time of winning the award
- `country_original`: Original country where laureate lived at the time of winning the award

4 Exercises

4.1 Get to know your data

1. How many observations and how many variables are in the dataset? Use inline code to answer this question. What does each row represent?

There are 935 observations and 26 variables in the dataset. Each row represents a Nobel laureate.

There are some observations in this dataset that we will exclude from our analysis to match the Buzzfeed results.

2. Create a new data frame called `nobel_living` that filters for

- laureates for whom `country` is available
- laureates who are people as opposed to organizations (organizations are denoted with `"org"` as their `gender`)
- laureates who are still alive (their `died_date` is `NA`)

```
nobel_living <- nobel %>%
  filter(
    !is.na(country),
    gender != "org",
    is.na(died_date)
  )
```

Confirm that once you have filtered for these characteristics you are left with a data frame with 228 observations, once again using inline code.

Knit, *commit, and push your changes to GitHub with an appropriate commit message. Make sure to commit and push all changed files so that your Git pane is cleared up afterwards.*

4.2 Most living Nobel laureates were based in the US when they won their prizes

... says the Buzzfeed article. Let's see if that's true.

First, we'll create a new variable to identify whether the laureate was in the US when they won their prize. We'll use the `mutate()` function for this. The following pipeline mutates the `nobel_living` data frame by adding a new variable called `country_us`. We use an if statement to create this variable. The first argument in the `if_else()` function we're using to write this if statement is the condition we're testing for. If `country` is equal to `"USA"`, we set `country_us` to `"USA"`. If not, we set the `country_us` to `"Other"`.

```
Note that we can achieve the same result using the `fct_other()` function we've seen before (i.e. with
```

```
nobel_living <- nobel_living %>%
  mutate(
    country_us = if_else(country == "USA", "USA", "Other")
  )
```

Next, we will limit our analysis to only the following categories: Physics, Medicine, Chemistry, and Economics.

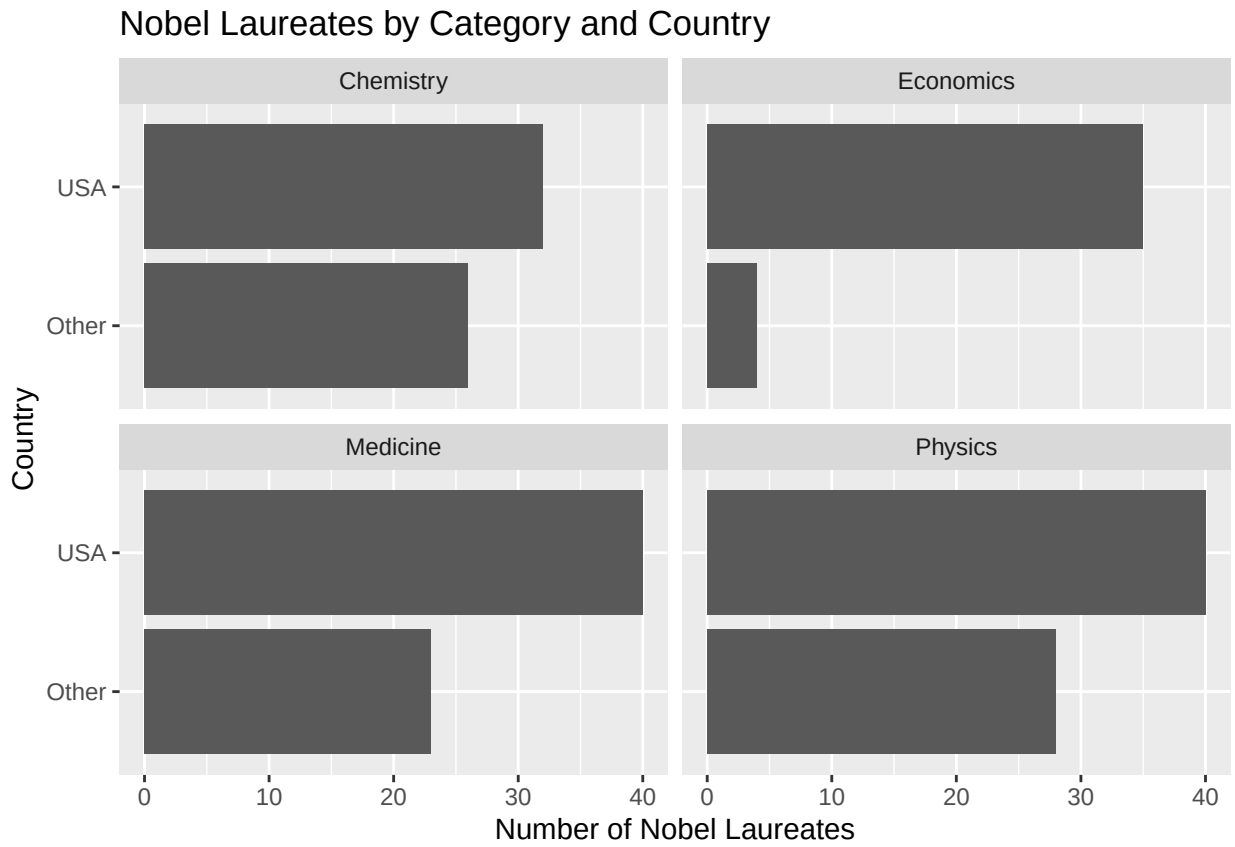
```
nobel_living_science <- nobel_living %>%
  filter(category %in% c("Physics", "Medicine", "Chemistry", "Economics"))
```

For the next exercise work with the `nobel_living_science` data frame you created above. This means you'll need to define this data frame in your R Markdown document, even though the next exercise doesn't explicitly ask you to do so.

3. Create a faceted bar plot visualizing the relationship between the category of prize and whether the laureate was in the US when they won the nobel prize. Interpret your visualization, and say a few words about whether the Buzzfeed headline is supported by the data.

    - Your visualization should be faceted by category.
    - For each facet you should have two bars, one for winners in the US and one for Other.
    - Flip the coordinates so the bars are horizontal, not vertical.

```
ggplot(nobel_living_science, aes(x = country_us)) +
  geom_bar() +
  facet_wrap(~ category) +
  coord_flip() +
  labs(
    x = "Country",
    y = "Number of Nobel Laureates",
    title = "Nobel Laureates by Category and Country"
  )
```



The visualization shows that the majority of living Nobel laureates in Physics, Chemistry, Medicine, and Economics were indeed based in the US when they won their prizes. This supports the Buzzfeed headline's claim.

Knit, *commit, and push your changes to GitHub with an appropriate commit message. Make sure to commit and push all changed files so that your Git pane is cleared up afterwards.d*

4.3 But of those US-based Nobel laureates, many were born in other countries

```
**Hint:** You should be able to borrow from code you used earlier to create the `country_us` variable.
```

4. Create a new variable called `born_country_us` that has the value `"USA"` if the laureate is born in the US, and `"Other"` otherwise. How many of the winners are born in the US?

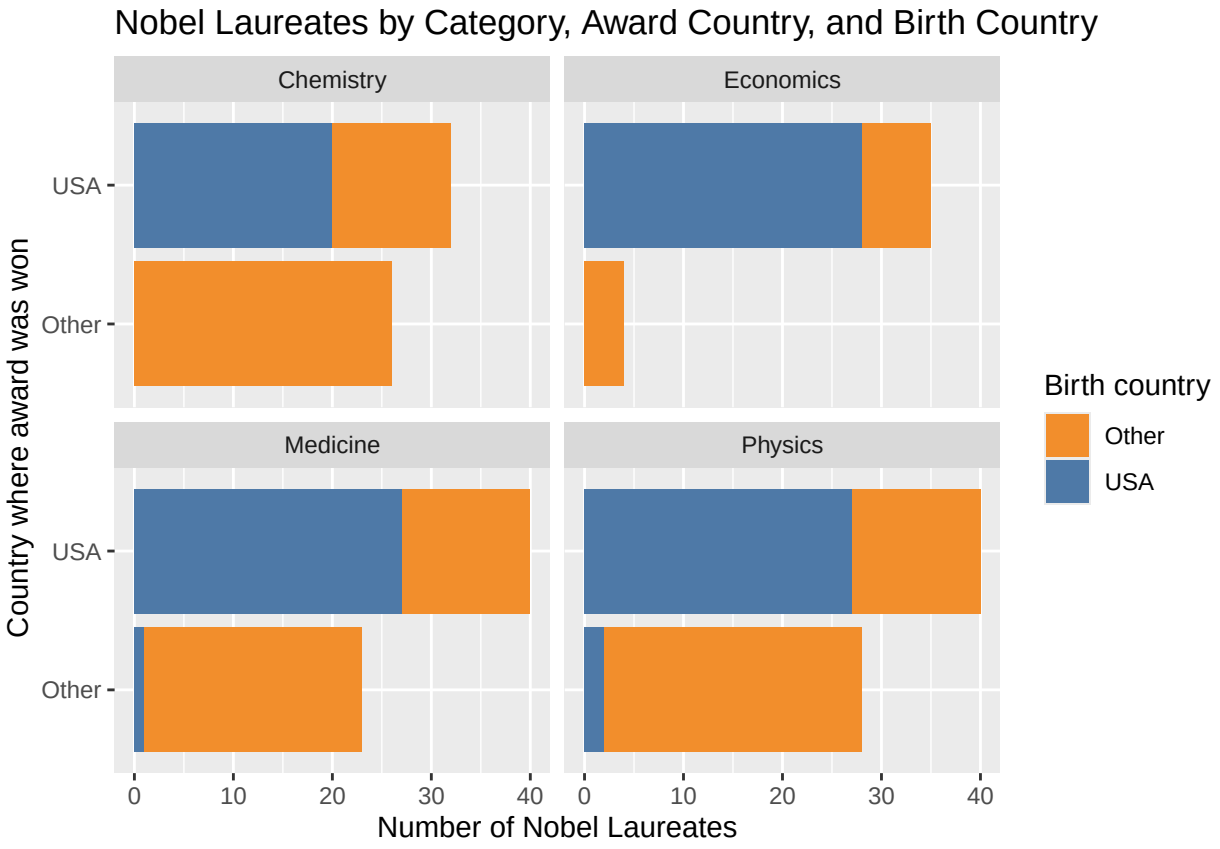
```
nobel_living_science <- nobel_living_science %>%
  mutate(
    born_country_us = if_else(born_country == "USA", "USA", "Other")
  )

# Count winners born in the US
us_born_count <- nobel_living_science %>%
  filter(born_country_us == "USA") %>%
  nrow()
```

There are 105 winners who were born in the US.

5. Add a second variable to your visualization from Exercise 3 based on whether the laureate was born in the US or not. Based on your visualization, do the data appear to support Buzzfeed's claim? Explain your reasoning in 1-2 sentences.

    - Your final visualization should contain a facet for each category.
    - Within each facet, there should be a bar for whether the laureate won the award in the US or not.
    - Each bar should have segments for whether the laureate was born in the US or not.

```
ggplot(nobel_living_science, aes(x = country_us, fill = born_country_us)) +
  geom_bar() +
  facet_wrap(~ category) +
  coord_flip() +
  labs(
    x = "Country where award was won",
    y = "Number of Nobel Laureates",
    fill = "Birth country",
    title = "Nobel Laureates by Category, Award Country, and Birth Country"
  ) +
  scale_fill_manual(values = c("USA" = "#4e79a7", "Other" = "#f28e2c"))
```



The data supports Buzzfeed's claim. While many Nobel laureates were based in the US when they won their prizes, a substantial portion of these US-based winners were actually born in other countries, highlighting the importance of immigration to American scientific achievement.

Knit, *commit, and push your changes to GitHub with an appropriate commit message. Make sure to commit and push all changed files so that your Git pane is cleared up afterwards.*

4.4 Here's where those immigrant Nobelists were born

```
Note that your bar plot won't exactly match the one from the Buzzfeed article. This is likely because th
```

6. In a single pipeline, filter for laureates who won their prize in the US, but were born outside of the US, and then create a frequency table (with the `count()` function) for their birth country (`born_country`) and arrange the resulting data frame in descending order of number of observations for each country. Which country is the most common?

```
immigrant_nobelists <- nobel_living_science %>%
  filter(country_us == "USA", born_country_us == "Other") %>%
  count(born_country) %>%
  arrange(desc(n))

# Display the table
immigrant_nobelists
```

```
## # A tibble: 21 x 2
##    born_country       n
##    <chr>          <int>
##  1 Germany            7
##  2 United Kingdom     7
##  3 China              5
##  4 Canada             4
##  5 Japan              3
##  6 Australia          2
##  7 Israel             2
##  8 Norway             2
##  9 Austria            1
## 10 Finland            1
## # i 11 more rows
```

The most common birth country for immigrant Nobel laureates who won their prize while in the US is Germany.

Knit, *commit, and push your changes to GitHub with an appropriate commit message. Make sure to commit and push all changed files so that your Git pane is cleared up afterwards and review the md document on GitHub to make sure you're happy with the final state of your work.*

Now go back through your write up to make sure you've answered all questions and all of your R chunks are properly labeled. Once you decide that you are done with the lab, choose the knit drop down and select `Knit to tufte_handout` to generate a pdf. Download and submit that pdf to Canvas.

5 Interested in how Buzzfeed made their visualizations?

The plots in the Buzzfeed article are called waffle plots. You can find the code used for making these plots in Buzzfeed's GitHub repo (yes, they have one!) here. You're not expected to recreate them as part of your assignment, but you're welcomed to do so for fun!