Lab 5.1 - Nobel Laureates

Adham Rishmawi

Spring 2022

```
library(tidyverse)
library(tidyr)
```

This analysis explores the data used by P. Aldhous in this Buzzfeed article. The article claims that one key factor in the US's leadership in science and technology is immigration because while most living Nobel laureates in the sciences are based in the US, many of them were born in other countries.

The Dataset

You'll need to get the nobel dataset, install it in a `data` folder, and load it.

```
nobel <- read_csv("~/data202/lab05/nobel.csv")
```

```
## Rows: 935 Columns: 26
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (21): firstname, surname, category, affiliation, city, country, gender,...
## dbl   (3): id, year, share
## date  (2): born_date, died_date
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Because there is no formal webpage for this dataset, study the dataset and give a short summary here on what it contains. How many observations and how many variables are in the dataset? What does each observation represent? Use inline code to answer this question. Please include this data dictionary.

```
obs <- nrow(nobel)
col <- ncol(nobel)
```

there is 935 observations and 26 variables in this data set

- `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 part of 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

Cleansing the Data

Create a new data frame called `nobel_living` that includes only:

```
nobel_living <-
  nobel %>%

  filter( !is.na(country)) %>%

  filter( gender != "org") %>%

  filter( is.na(died_date))

nobel_living
```

```
## # A tibble: 228 x 26
##       id firstname   surname  year category affiliation city  country born_date
##    <dbl> <chr>       <chr>   <dbl> <chr>    <chr>       <chr> <chr>   <date>
##  1    68 Chen Ning   Yang     1957 Physics  Institute ~ Prin~ USA     1922-09-22
##  2    69 Tsung-Dao   Lee      1957 Physics  Columbia U~ New ~ USA     1926-11-24
##  3    95 Leon N.     Cooper   1972 Physics  Brown Univ~ Prov~ USA     1930-02-28
##  4    97 Leo         Esaki    1973 Physics  IBM Thomas~ York~ USA     1925-03-12
##  5    98 Ivar        Giaever  1973 Physics  General El~ Sche~ USA     1929-04-05
##  6    99 Brian D.    Joseph~  1973 Physics  University~ Camb~ United~ 1940-01-04
##  7   101 Antony      Hewish   1974 Physics  University~ Camb~ United~ 1924-05-11
##  8   103 Ben R.      Mottel~  1975 Physics  Nordita     Cope~ Denmark 1926-07-09
##  9   106 Samuel C.C. Ting     1976 Physics  Massachuse~ Camb~ USA     1936-01-27
## 10   107 Philip W.   Anders~  1977 Physics  Bell Telep~ Murr~ USA     1923-12-13
## # ... with 218 more rows, and 17 more variables: died_date <date>,
## #   gender <chr>, born_city <chr>, born_country <chr>, born_country_code <chr>,
## #   died_city <chr>, died_country <chr>, died_country_code <chr>,
## #   overall_motivation <chr>, share <dbl>, motivation <chr>,
## #   born_country_original <chr>, born_city_original <chr>,
## #   died_country_original <chr>, died_city_original <chr>, city_original <chr>,
## #   country_original <chr>
```

```
nrow(nobel_living)
```

```
## [1] 228
```

```
ncol(nobel_living)
```

```
## [1] 26
```

- laureates for whom country is available (i.e., it isn't `NA` – remember to use the `is.na()` function)
- 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`)

Confirm that once you have filtered for these characteristics you are left with a data frame with 228 observations.

Determining Where Nobel Laureates Lived

The Buzzfeed article claims that most living Nobel laureates were based in the US when they won their prizes. First, we'll create a new variable to identify whether the laureate was in the US when they won their prize.

We include a `mutate()` function that uses a functional variant of the classic "if" statement, called `if_else()`, to create this variable. The arguments to this new function, to be covered in more detail later in the course, are:

- the condition for which we're testing (e.g., is the country the USA?)
- the value to use if the condition is true (e.g., if `country` is equal to `"USA"`, it gives us `"USA"`)
- the value to use otherwise (e.g., if the country isn't `"USA"`, we get `"Other"`).

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

  mutate( country_us = if_else(country == "USA", "USA", "Other") )|>

    mutate (born_country_us = if_else(born_country == "USA", "USA", "Other"))

nobel_living_science
```

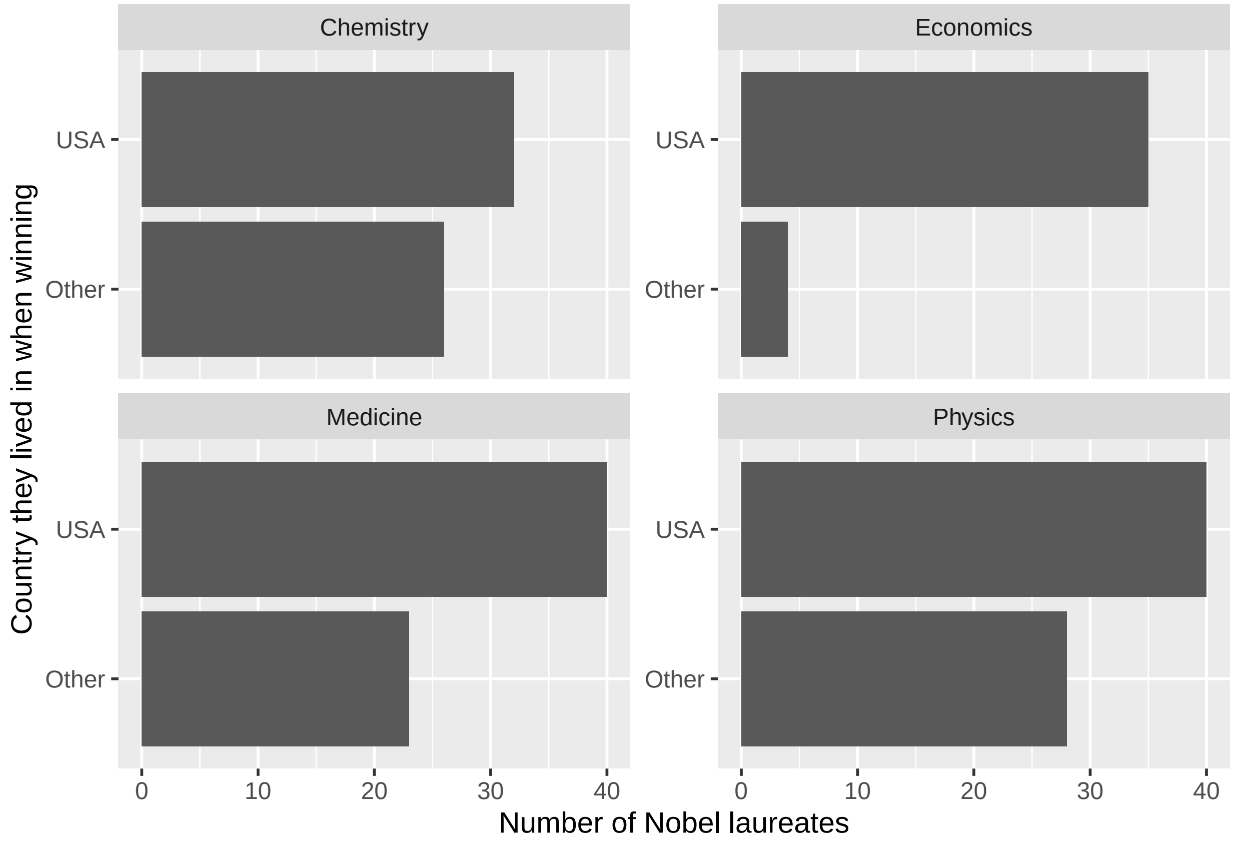
```
## # A tibble: 228 x 28
##       id firstname   surname  year category affiliation city  country born_date
##    <dbl> <chr>       <chr>   <dbl> <chr>    <chr>       <chr> <chr>   <date>
##  1    68 Chen Ning   Yang     1957 Physics  Institute ~ Prin~ USA     1922-09-22
##  2    69 Tsung-Dao   Lee      1957 Physics  Columbia U~ New ~ USA     1926-11-24
##  3    95 Leon N.     Cooper   1972 Physics  Brown Univ~ Prov~ USA     1930-02-28
##  4    97 Leo         Esaki    1973 Physics  IBM Thomas~ York~ USA     1925-03-12
##  5    98 Ivar        Giaever  1973 Physics  General El~ Sche~ USA     1929-04-05
##  6    99 Brian D.    Joseph~  1973 Physics  University~ Camb~ United~ 1940-01-04
##  7   101 Antony      Hewish   1974 Physics  University~ Camb~ United~ 1924-05-11
##  8   103 Ben R.      Mottel~  1975 Physics  Nordita     Cope~ Denmark 1926-07-09
##  9   106 Samuel C.C. Ting     1976 Physics  Massachuse~ Camb~ USA     1936-01-27
## 10   107 Philip W.   Anders~  1977 Physics  Bell Telep~ Murr~ USA     1923-12-13
## # ... with 218 more rows, and 19 more variables: died_date <date>,
## #   gender <chr>, born_city <chr>, born_country <chr>, born_country_code <chr>,
## #   died_city <chr>, died_country <chr>, died_country_code <chr>,
## #   overall_motivation <chr>, share <dbl>, motivation <chr>,
## #   born_country_original <chr>, born_city_original <chr>,
## #   died_country_original <chr>, died_city_original <chr>, city_original <chr>,
## #   country_original <chr>, country_us <chr>, born_country_us <chr>
```

Add a code chunk that creates a data frame called `nobel_living_science` by combining the two transformations above into a pipeline: use the `mutate()` with the `if_else` discussed above to create the a `country_us` variable; and use `filter()` to limit the results to include only categories with values `%in%` “Physics”, “Medicine”, “Chemistry”, “Economics”.

Create a faceted bar plot, with horizontal bars, 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.

```
nobel_living_science|>
  ggplot() +
  aes( y = country_us) +
  geom_bar()+
  labs( x ="Number of Nobel laureates", y = "Country they lived in when winning")+
  facet_wrap(~category, scales = "free_y")
```



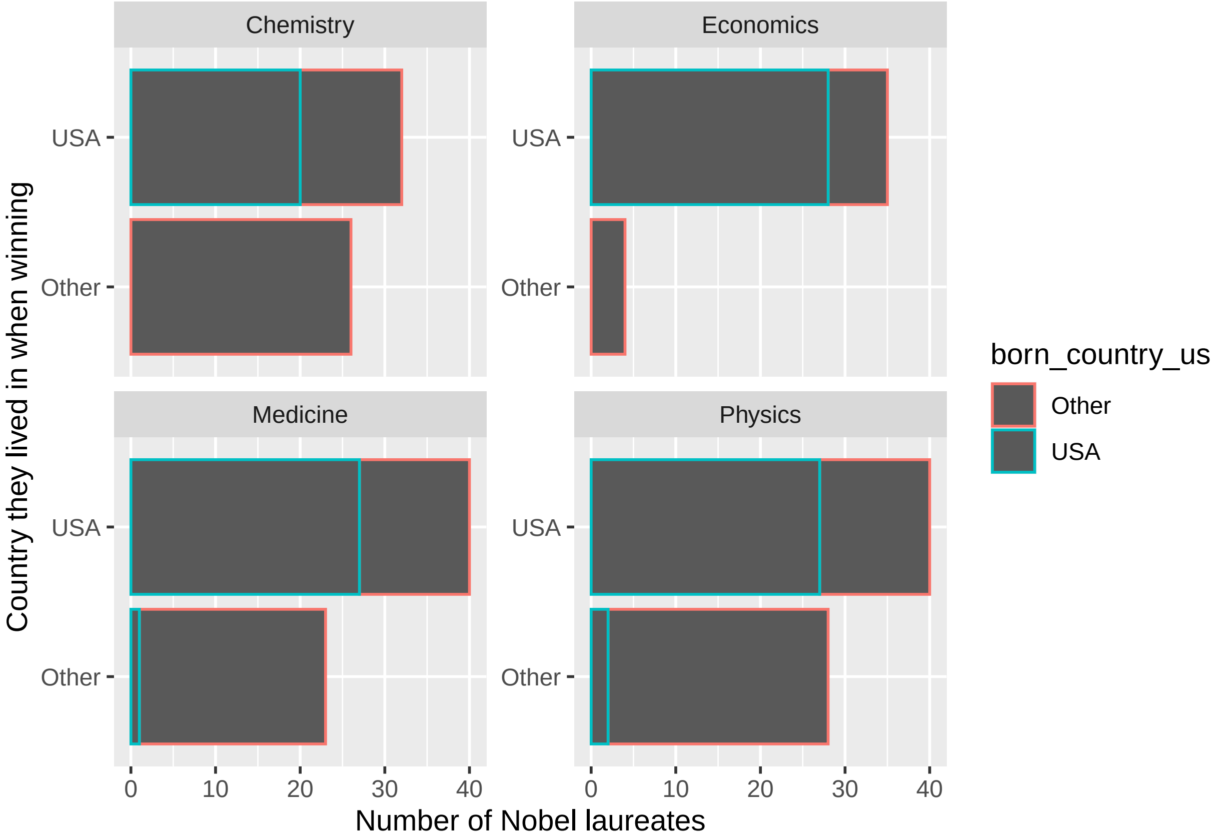
### THIS CLAIM ISNT SUPPORTED FROM OUR CONCLUSIONS

Determining Where Nobel Laureates Were Born

Go back to the code chunk that created `nobel_living_science` and add a new variable called `born_country_us` that has the value `"USA"` if the laureate is born in the US, and `"Other"` otherwise. Do this by modifying your earlier code chunk; you won’t add anything new here.

Remake your visualization and add a second variable: whether the laureate was born in the US or not. Your final visualization should contain a facet for each category, within each facet a bar for whether they won the award in the US or not, and within each bar whether they were born in the US or not. (Don't over-think this: you can do this by just adding another aesthetic mapping!) Based on your visualization, do the data appear to support Buzzfeed's claim? Explain your reasoning in 1-2 sentences.

```
nobel_living_science|>
  ggplot() +
  aes( y = country_us, color = born_country_us) +
  geom_bar()+
  labs( x ="Number of Nobel laureates", y = "Country they lived in when winning")+
  facet_wrap(~category, scales = "free_y")
```



This claim that buzzfeed has made was a blank claim and has resulted in miscommunication. This would have violated the first principle of data scientist! The data show that very few Nobel prize winners who won in other countries emmigrated there from the US. Conversely, however, the data show that many US prize winners were born in other countries, at least for fields other than Economics in which the majority of winners were not US-born.

Determining Where Immigrant Nobel Laureates Were Born

Make a table for where immigrant Nobelists were born, using a single pipeline:

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

```
## # 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
## # ... with 11 more rows
```

- filter for living STEM laureates who won their prize in the US, but were born outside of the US,
- then create a frequency table for their birth country, `born_country`,
- then sort the result in descending order of number of Nobelists for each country.

Recreating the Buzzfeed Visualizations [OPTIONAL]

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 welcome to do so for fun!