Lab 03 - Nobel laureates

Alison Yao, Oscar Bray

7/10/21

Load packages and data

```
library(tidyverse)
```

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

Exercises

Exercise 1

There are **935** observations (ie Nobel laureates) and **26** variables (ie categories of information) in this data set. Each row represents a different Nobel laureate.

```
library(skimr)
```

```
skim(nobel)
```

Data summary

Name	nobel
Number of rows	935
Number of columns	26
________________________	
Column type frequency:	
character	21
Date	2
numeric	3
________________________	
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
firstname	0	1.00	2	59	0	720	0
surname	29	0.97	2	26	0	851	0
category	0	1.00	5	10	0	6	0

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
affiliation	250	0.73	4	110	0	303	0
city	255	0.73	4	27	0	185	0
country	254	0.73	3	14	0	27	0
gender	0	1.00	3	6	0	3	0
born_city	28	0.97	3	29	0	613	0
born_country	28	0.97	3	28	0	80	0
born_country_code	28	0.97	2	2	0	77	0
died_city	327	0.65	4	29	0	303	0
died_country	321	0.66	3	16	0	48	0
died_country_code	321	0.66	2	2	0	46	0
overall_motivation	918	0.02	55	114	0	7	0
motivation	0	1.00	24	337	0	656	0
born_country_original	28	0.97	3	52	0	122	0
born_city_original	28	0.97	3	36	0	616	0
died_country_original	321	0.66	3	35	0	52	0
died_city_original	327	0.65	4	29	0	303	0
city_original	255	0.73	4	27	0	185	0
country_original	254	0.73	3	35	0	29	0

Variable type: Date

skim_variable	n_missing	complete_rate	min	max	median	n_unique
born_date	33	0.96	1817-11-30	1997-07-12	1916-06-28	885
died_date	308	0.67	1903-11-01	2019-08-07	1983-03-09	616

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
id	0	1	475.12	277.83	1	234.5	470	716.5	969	▇▇▇▇▇
year	0	1	1970.44	33.30	1901	1947.0	1976	1999.0	2018	▃▃▆▆▇
share	0	1	1.99	0.94	1	1.0	2	3.0	4	▇▇▁▅▂

The `skim` function might be an overkill, `nrow()` and `ncol()` also works:

```
nrow(nobel)
```

```
## [1] 935
```

```
ncol(nobel)
```

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

Exercise 2

Filtering the original dataframe using 3 conditions, then save the new dataframe to `nobel_living`.

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

And indeed, we do get 228 observations. Yay!

```
nrow(nobel_living)
```

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

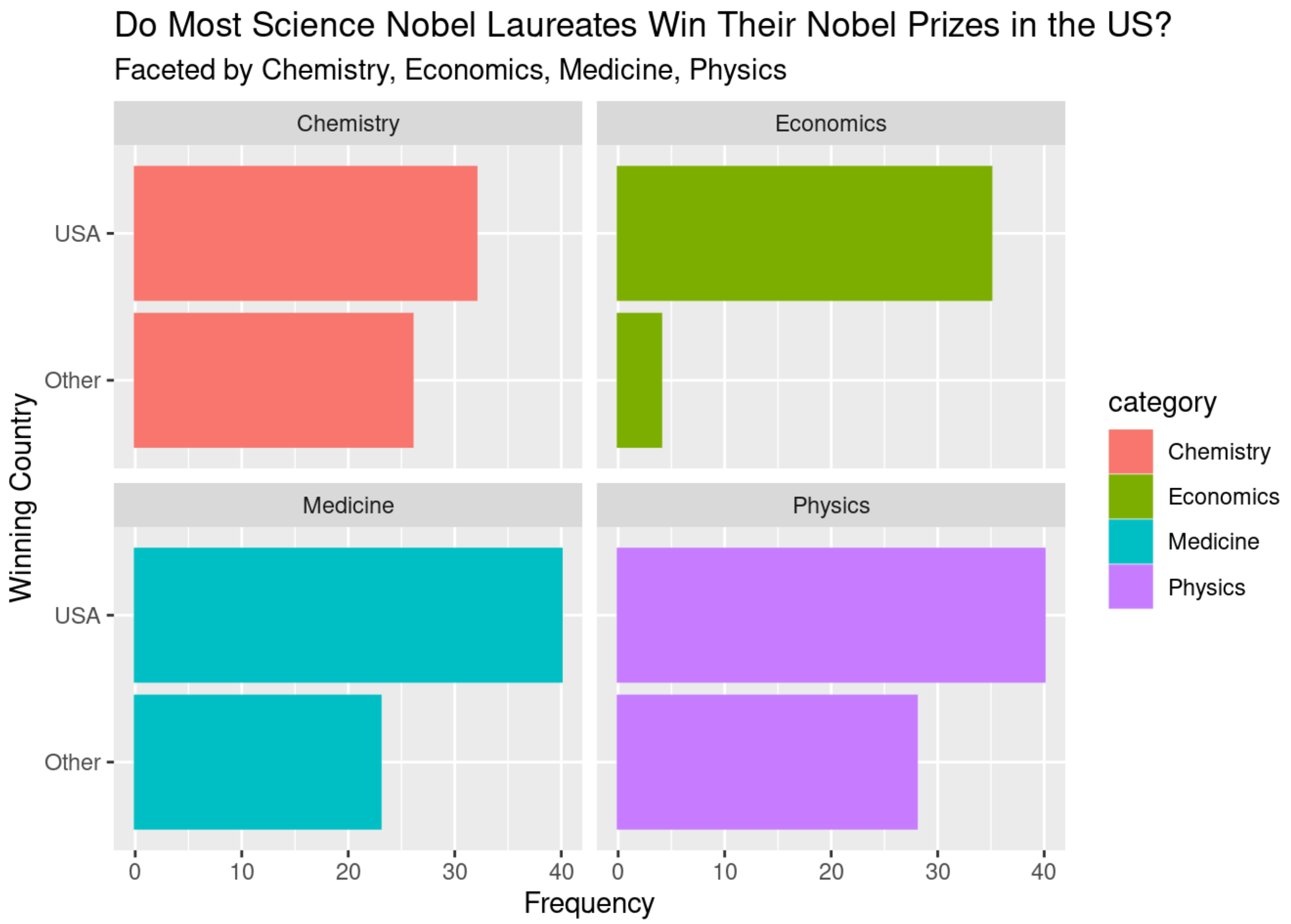
Before doing exercise3, following the instructions, let’s make a `nobel_living_science` dataframe.

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

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

Exercise 3

```
nobel_living_science %>%
  ggplot(aes(x = country_us,
             color = category,
             fill = category)) +
    geom_bar() +
    coord_flip() +
  facet_wrap(~ category) +
  labs(
    title = 'Do Most Science Nobel Laureates Win Their Nobel Prizes in the US?',
    subtitle = 'Faceted by Chemistry, Economics, Medicine, Physics',
    x = 'Winning Country',
    y = 'Frequency')
```



These plots show that there is a disparity in Nobel laureates in the US versus other countries, especially between economics laureates. This means that the Buzzfeed headline's proposal that 'immigration is important for American science' is supported because the fact that most laureates are from the US indicates a 'brain drain' in other countries. This means that the data could be used to support the idea that scientists are immigrating from other countries to the US and therefore winning more Nobel laureates for the US. However, to fully support this idea, we would need to demonstrate that the country of origin for most of the US scientists is outside the US, since then they would have immigrated thus supporting the argument.

Exercise 4

There are **105** winners who were born in the US.

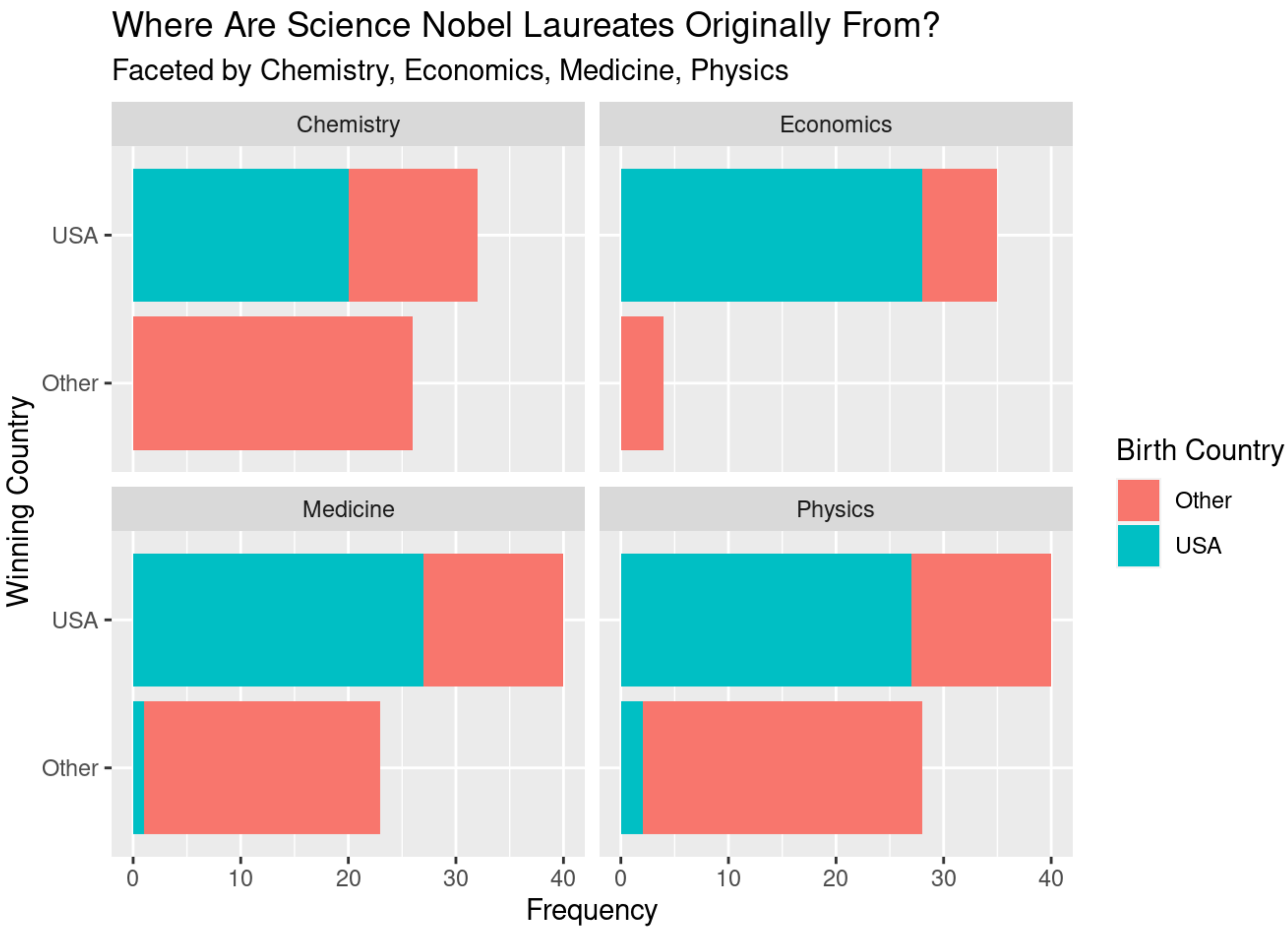
```
born_country_us <- nobel_living_science %>%
  mutate(
    born_country_us = if_else(born_country == "USA", "USA", "Other")
  )
```

```
filter(born_country_us, born_country_us == "USA") %>%
  count()
```

```
## # A tibble: 1 × 1
##       n
##   <int>
## 1   105
```

Exercise 5

```
born_country_us %>%
  ggplot(aes(x = country_us,
             fill = born_country_us)) +
    geom_bar() +
    coord_flip() +
  facet_wrap(~ category) +
  labs(
    title = 'Where Are Science Nobel Laureates Originally From?',
    subtitle = 'Faceted by Chemistry, Economics, Medicine, Physics',
    x = 'Winning Country',
    y = 'Frequency',
    fill = "Birth Country")
```



Now we can see that the data somewhat supports Buzzfeed's hypothesis about immigration correlating with more US-based Nobel laureates. While the majority of winners based in the US were from the US originally, the amount that came from other countries is not insignificant, constituting between a quarter and a third of all US-based laureates in every subject except economics. This means that immigrants to the US have contributed a significant amount of Nobel prize wins.

Exercise 6

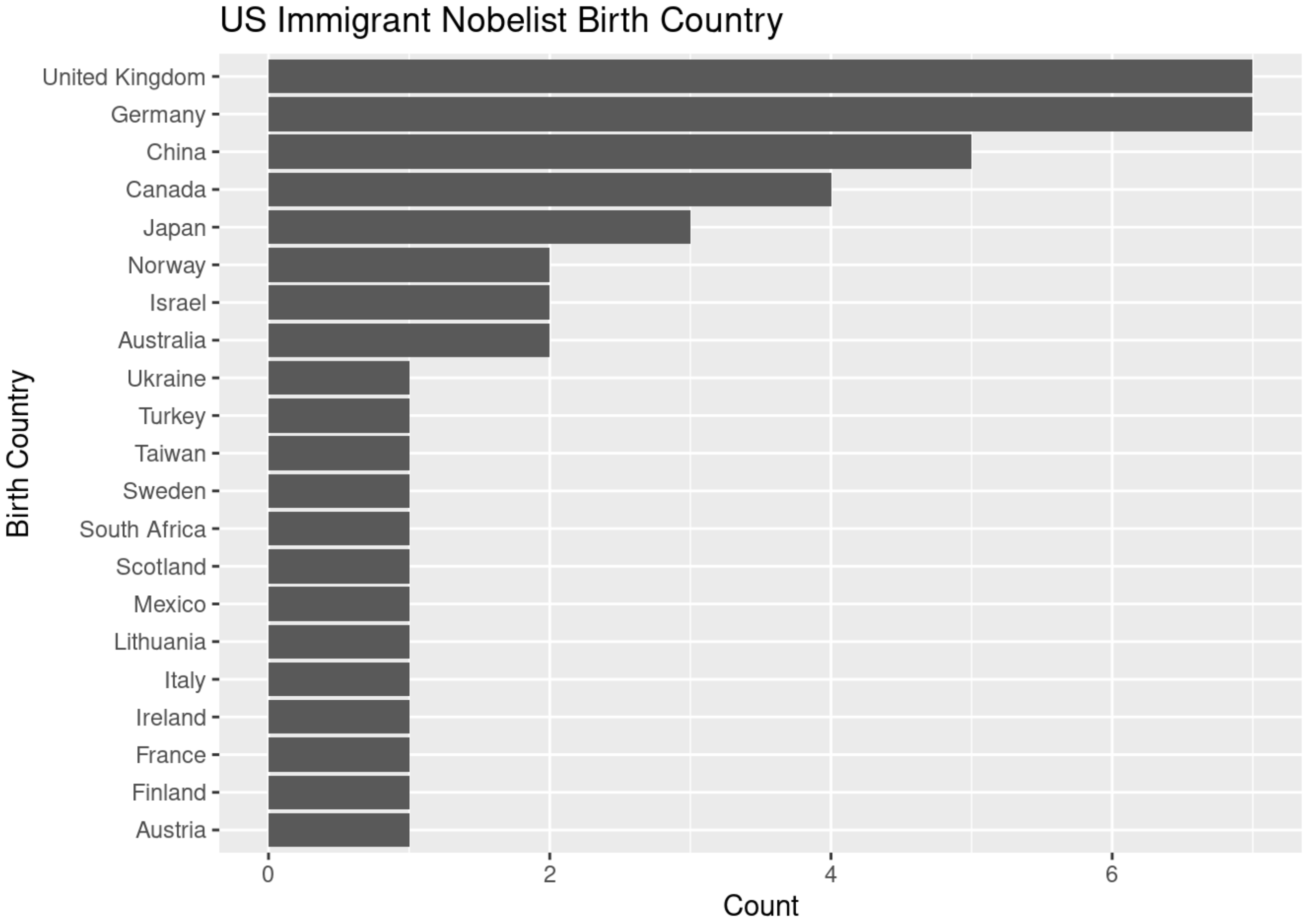
Germany and the **UK** are the most common.

```
born_country_us %>%
  filter(country_us == 'USA' & born_country != 'USA') %>%
  count(born_country) %>%
  arrange(desc(n))
```

```
## # A tibble: 21 × 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
```

We also used the data in this frequency table to create a bar plot, in case the HW want us to recreate the buzzfeed visualization.

```
born_country_us %>%
  filter(country_us == 'USA' & born_country != 'USA') %>%
  count(born_country) %>%
  ggplot(aes(x = reorder(born_country, n),
           y = n)) +
    geom_bar(stat="identity") +
    coord_flip() +
    labs(
      title = 'US Immigrant Nobelist Birth Country',
      x = 'Birth Country',
      y = 'Count'
    )
```


US Immigrant Nobelist Birth Country
Birth Country
United Kingdom
Germany
China
Canada
Japan
Norway
Israel
Australia
Ukraine
Turkey
Taiwan
Sweden
South Africa
Scotland
Mexico
Lithuania
Italy
Ireland
France
Finland
Austria
0
2
4
6
Count