

Lab 03 - Nobel laureates - results

Honorata Hurnik + Kamil Klecha

```
library(tidyverse)

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

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

```
nrow(nobel)

## [1] 935

ncol(nobel)

## [1] 26
```

Each row represents a nobel laureat.

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

## # A tibble: 6 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
## # ... with 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>
```

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.

```
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.

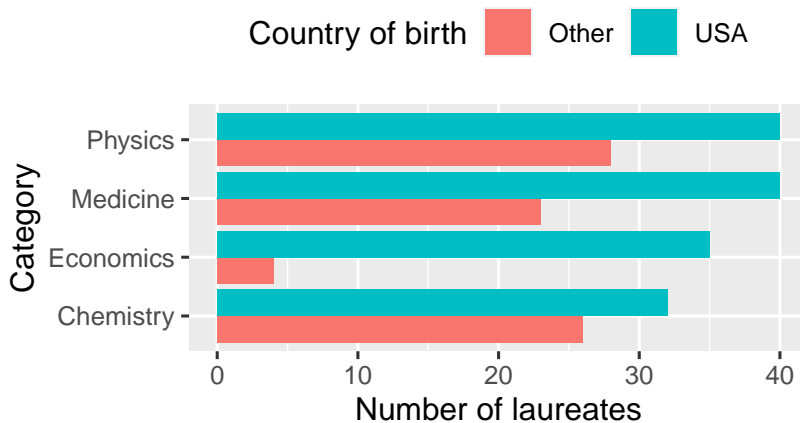
```
nobel_living_science <- nobel_living %>%
  filter(category %in% c("Physics", "Medicine", "Chemistry", "Economics"))
```

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.

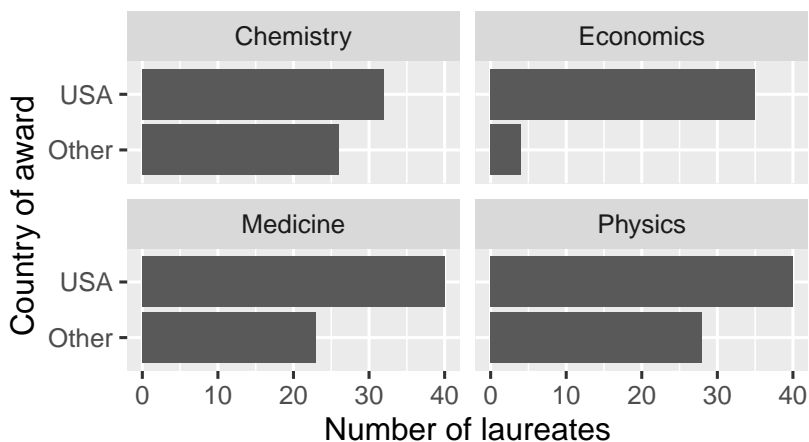
```
ggplot(nobel_living_science, aes(y = category, fill=country_us)) +
  geom_bar(position="dodge") +
  theme(legend.position="top") +
  ggtitle("On one graph") +
  labs(y="Category", fill="Country of birth", x="Number of laureates")
```

On one graph



```
ggplot(nobel_living_science, aes(y = country_us)) +
  geom_bar() +
  facet_wrap(~category) +
  ggtitle("Faceted version") +
  labs(y="Country of award", fill="Country of birth", x="Number of laureates")
```

Faceted version



```
print("It seems like the data supports the claim, as most of the Nobel lauerates were based in US when l
## [1] "It seems like the data supports the claim, as most of the Nobel lauerates were based in US when l
```

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?

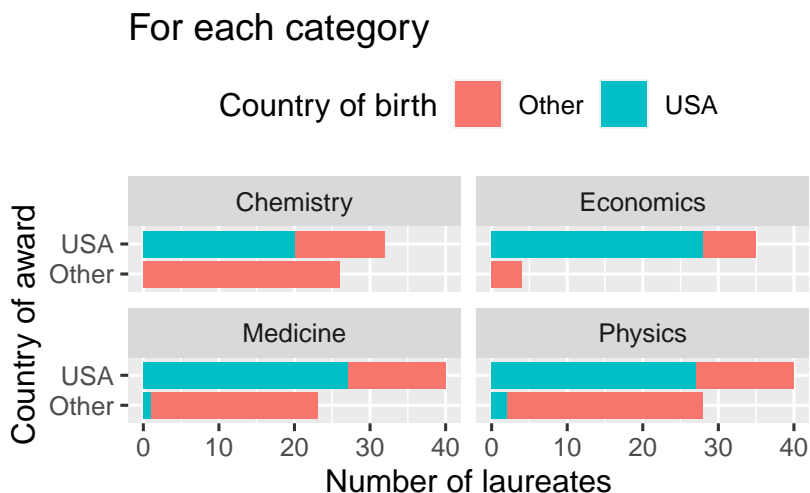
```
nobel_living_science <- nobel_living_science %>%
  mutate(
    born_country_us = if_else(born_country == "USA", "USA", "Other")
  )
sum(nobel_living_science$born_country_us == "USA")

## [1] 105
```

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(y = country_us, fill=born_country_us)) +
  facet_wrap(~category) +
  geom_bar(position="stack") +
  theme(legend.position="top") +
  ggtitle("For each category") +
  labs(y="Country of award", fill="Country of birth", x="Number of laureates")
```



As the bar chart shows, majority of Laureats who won their Nobel Prize in the USA were also born in the USA. However, the percentage of Laureats who won the Nobel Prize in the USA, but who were born in different countries is high, which leads to the conclusion that based

on the visualization, the data support Buzzfeed's claim.

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?

```
nobel_living_science_filtered <- nobel_living_science %>%
  filter(country_us == "USA") %>%
  filter(born_country_us == "Other")

count(nobel_living_science_filtered, born_country, sort = TRUE)

## # 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

print("The most common is Germany and United Kingdom.")

## [1] "The most common is Germany and United Kingdom."
```