Nobel Prize Winners Analysis

Jonah Maroszek

2/16/2022

Introduction

I chose to analyze data about Nobel Prize winners because I have always been interested in science. I thought it would be interesting to explore some characteristics of my scientific heroes. I used three data sets to help me answer some questions about Nobel Prize winners. They are named nobel, world, and country_codes in my code. nobel provides general data about the Nobel Prize winners, world is used to help create the map of the world which is used in the first visual, and country_codes is used to link these two data sets together.

Load and Process Data

```
nobel = read_csv("https://tinyurl.com/yctm5rz5", show_col_types = FALSE)

world = ne_countries(scale = "medium", returnclass = "sf")

country_codes = read_csv("https://tinyurl.com/y9lfy35w", show_col_types = FALSE) %>%
                select(c(starts_with("Alpha-2"), starts_with("English"))) %>%
                rename("birth_country" = "English short name lower case")

nobel_born_count = left_join(nobel, country_codes, by = c("born_country_code" = "Alpha-2 code")) %>%
                   left_join(world, by = c("birth_country" = "name")) %>%
                   group_by(born_country_code) %>%
                   summarize(nobel_born_count = n())

final_data = left_join(world, country_codes, by = c("name" = "birth_country")) %>%
             left_join(nobel_born_count, by = c("Alpha-2 code" = "born_country_code")) %>%
             select(c(name, nobel_born_count, pop_est, "Alpha-2 code")) %>%
             mutate(nobel_per_million = nobel_born_count / pop_est * 1000000)
```

Where are most Nobel Prize winners born?

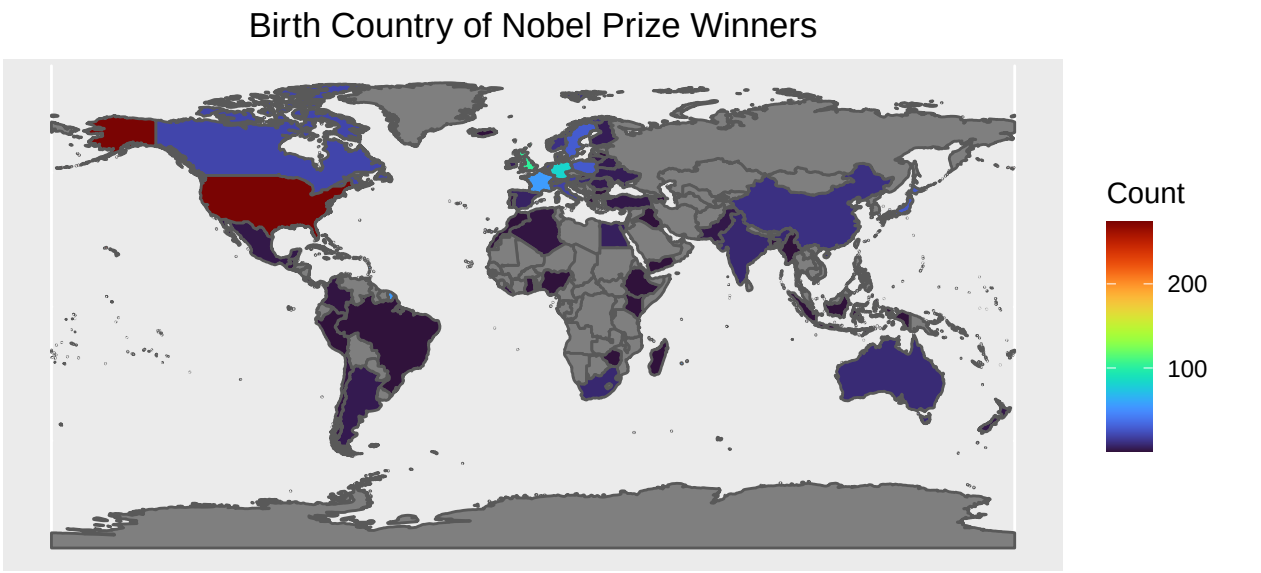
Through making this visualization, I have found that a vast majority of Nobel Prize winners are born in the United States. The top 10 countries can be seen in a table accompanying this figure. I thought that the US might be ahead because of its relative size, so I also calculated the number of Nobel prizes won per one million people as well for all the countries. The US is ranked 11 by this standard.

Although most winners are born in the US or Europe, I was surprised to find that there are many Nobel Prize winners born in South America and Africa too.

I have chosen the turbo color scheme because most other color schemes were very hard to discern the values between countries other than the United States.

```
ggplot(final_data, aes(fill = nobel_born_count)) +
  geom_sf() +
```

```
  ggtitle("Birth Country of Nobel Prize Winners") +
  labs(fill = "Count") +
  scale_fill_viridis(option="turbo") +
  theme(plot.title = element_text(hjust = 0.5))
```




```
final_data %>%
  arrange(desc(nobel_born_count)) %>%
  select(c(name, nobel_born_count, nobel_per_million)) %>%
  tibble() %>%
  head(10)
```

```
## # A tibble: 10 x 4
##    name           nobel_born_count nobel_per_million                  geometry
##    <chr>                     <int>             <dbl>        <MULTIPOLYGON [°]>
##  1 United States               274             0.873 (((-155.5813 19.01201, -15~
##  2 United Kingdom              103             1.65  (((-1.065576 50.69023, -1.~
##  3 Germany                      82             0.996 (((14.19824 53.91904, 14.2~
##  4 France                       56             0.874 (((55.79736 -21.33936, 55.~
##  5 Sweden                       29             3.20  (((16.52852 56.29053, 16.4~
##  6 Poland                       28             0.728 (((19.60439 54.45918, 19.6~
##  7 Japan                        27             0.212 (((123.8887 24.28013, 123.~
##  8 Canada                       20             0.597 (((-59.7876 43.9396, -59.9~
##  9 Switzerland                  19             2.50  (((9.524023 47.52422, 9.55~
## 10 Italy                        19             0.327 (((12.05127 36.75703, 12.0~
```

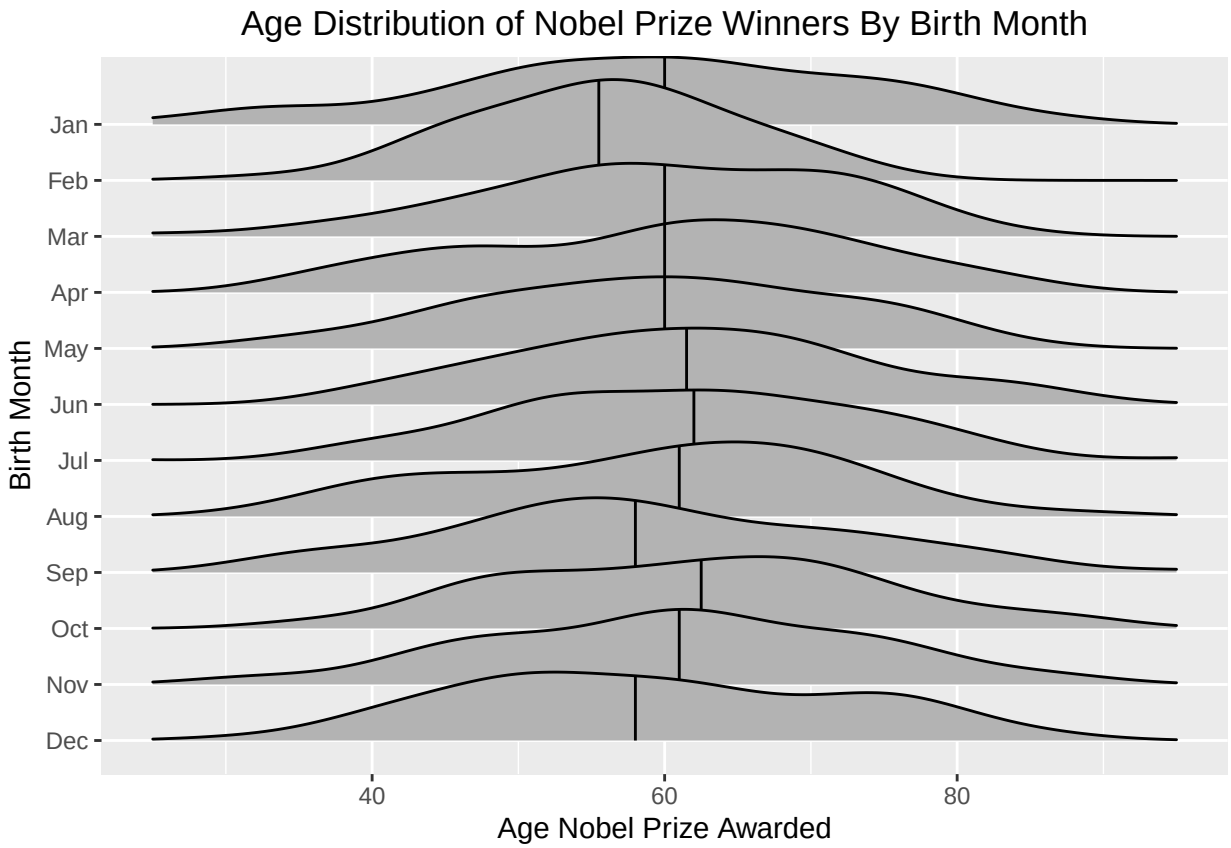
What is the age distribution of Nobel Prize winners by birth month?

I was wondering if the age distribution would be different for any birth month for some reason. I thought a ridge plot would be a good way to compare the age distributions across all months. I added a median line for each distribution as well to make it easier to spot deviations from the norm. I also reordered the months chronologically.

I was surprised that February Nobel Prize winners seemed younger than the other months based on this visual. I conducted a t-test which revealed that February winners were indeed younger than the other months at a statistically significant level (one-sided t-test, p-value = .0002642).

```
ggplot(nobel, aes(x = age_get_prize, y = born_month)) +
  stat_density_ridges(from = 25, to = 95, quantile_lines = TRUE, quantiles = 2) +
  scale_y_discrete(limits = rev(month.abb)) +
  ylab("Birth Month") +
  xlab("Age Nobel Prize Awarded") +
  ggtitle("Age Distribution of Nobel Prize Winners By Birth Month") +
  theme(plot.title = element_text(hjust = 0.5))
```

```
## Picking joint bandwidth of 4.65
```




```
feb_winners = nobel %>% filter(born_month == "Feb")
all_others = nobel %>% filter(born_month != "Feb")

t_test = t.test(feb_winners$age_get_prize, all_others$age_get_prize, alternative = "less")
```