Assignment 10B

James Chun

November 1, 2025

More JSON Practice: Nobel Prize

Load Libraries

```r
library(RCurl)
library(jsonlite)
library(purrr)
library(httr2)
library(tidyverse)
library(lubridate)
```

Import JSON

The data we'll be using is about the Nobel Prize Laureates. This information can be accessed virtually through the public API: https://app.swaggerhub.com/apis/NobelMedia/NobelMasterData/2.1 Furthermore, the actual database is found on http://api.nobelprize.org/2.1/laureates?limit=SETLIMIT Where the limit of how many individuals you want to observe can be set manually. This site will be the one used for the purposes of this markdown. All of these sites and endpoints are sourced from https://www.nobelprize.org/about/developer-zone-2/

```r
# Create url request
req <- request("http://api.nobelprize.org/2.1/laureates?limit=1050")
resp <- req_perform(req)

# Extract json
laureates_json <- resp %>%
  resp_body_json(flatten = TRUE)
```

The initial request has the limit set to 1050, it was an arbitrary number but does seem to collect all the laureates.

Tidy JSON

First tidy the JSON into a dataframe. Note that laureates with multiple prizes did not have their prizes properly represented in the dataframe.

```
laureates <- laureates_json$laureates %>%
  map_dfr(function(laureate) {
  tibble(
    name = if (!is.null(laureate$fullName$en)){
      laureate$fullName$en
      } else{
        laureate$orgName$en
        },

    country = if (!is.null(laureate$birth)) {
      laureate$birth$place$country$en
      } else{
        laureate$founded$place$country$en
        },

    birthday = if (!is.null(laureate$birth)){
      laureate$birth$date
      } else{
        laureate$founded$date
        },

    gender = laureate$gender,
    year.awarded = laureate$nobelPrizes[[1]]$awardYear,
    category = laureate$nobelPrizes[[1]]$category$en,
    prize.amount.adjusted = laureate$nobelPrizes[[1]]$prizeAmountAdjusted,
    affiliation = laureate$nobelPrizes[[1]]$affiliations[[1]]$name$en,
    affiliation.country = laureate$nobelPrizes[[1]]$affiliations[[1]]$country$en
  )
})
```

The final result of this importing and tidying is the dataframe ***laureates*** which contains the information: name, home country, birthday (or founding date), gender, category (of prize), prize amount, and affiliations, and affiliated country. Some of these values don't have data for natural reasons - such as if a laureate was an organization it won't have a gender. But, some other missing values are incomplete from the source.

Interesting Questions

1. Which country retained the most nobel laureates?

First, create new column to track if home country equals affiliated country, NA's in either column will result in NA.

```
# first ifelse is to filter NA's, the second is to compare the columns
laureates$country.retained <- ifelse(
  !is.na(laureates$country) &
    !is.na(laureates$affiliation.country),

  ifelse(
    laureates$country == laureates$affiliation.country,
    "yes",
    "no"
  ),
  NA
)
```

Then create summary

```
laureates %>%
  filter(!is.na(country)) %>%
  group_by(country) %>%
  summarise(
    laureates.born = n(),
    retained = sum(country.retained == "yes", na.rm = TRUE),
    ratio = retained / laureates.born
)
```

```
## # A tibble: 101 x 4
##    country          laureates.born retained ratio
##    <chr>                     <int>    <int> <dbl>
##  1 Argentina                     4        1 0.25
##  2 Australia                    11        3 0.273
##  3 Austria                      17        5 0.294
##  4 Austria-Hungary              13        0 0
##  5 Austrian Empire               4        0 0
##  6 Bangladesh                    1        0 0
##  7 Bavaria                       1        0 0
##  8 Belgian Congo                 1        0 0
##  9 Belgium                      10        4 0.4
## 10 Bosnia                        1        0 0
## # i 91 more rows
```

This shows that the USA has the highest retention rate for laureates, with 261 of their 302 naturally born laureates winning the nobel prize while still affiliated with the USA.

2. The second question to be answered will be the one provided in the assignment description.
Which country "lost" the most nobel laureates (who were born there but received their Nobel prize as a citizen of a different country)?

The code looks similar, except the ratio column will be replaced with the difference between laureates born and retained. This will show a better indicator of how many laureates a country lost as compared to ratios.

```
laureates %>%
  filter(!is.na(country)) %>%
  group_by(country) %>%
  summarise(
    laureates.born = n(),
    retained = sum(country.retained == "yes", na.rm = TRUE),
    difference = laureates.born - retained
)
```

```
## # A tibble: 101 x 4
##    country          laureates.born retained difference
##    <chr>                     <int>    <int>      <int>
##  1 Argentina                     4        1          3
##  2 Australia                    11        3          8
##   3 Austria                     17         5          12
##   4 Austria-Hungary             13         0          13
##   5 Austrian Empire              4         0           4
##   6 Bangladesh                   1         0           1
##   7 Bavaria                      1         0           1
##   8 Belgian Congo                1         0           1
##   9 Belgium                     10         4           6
##  10 Bosnia                       1         0           1
## # i 91 more rows
```

Surprisingly, the USA wins this again with having lost 41 laureates. Though Germany and the UK are tied for second place with 37 laureates lost each. Though, between the two the UK has the higher retention rate.

3. On average, which category of prize of awarded at the highest age?

First create column that indicates age of laureate at time of award.

```
laureates$age.awarded <- as.numeric(laureates$year.awarded) -
  as.numeric(substr(laureates$birthday, 1, 4))
```

Note that the ages of organizations were ignored, as the question intends to focus on individual laureates. This is achieved by filtering out cases where the gender is NA.

```
laureates %>%
  filter(!is.na(gender)) %>%
  group_by(category) %>%
  summarise(
    average.age = mean(age.awarded, na.rm = TRUE),
    youngest = min(age.awarded, na.rm = TRUE),
    oldest = max(age.awarded, na.rm = TRUE),
    total = n()
)
```

```
## # A tibble: 6 x 5
##   category               average.age youngest oldest total
##   <chr>                        <dbl>    <dbl>  <dbl> <int>
## 1 Chemistry                     59.2       35     97   197
## 2 Economic Sciences             67.0       47     90    99
## 3 Literature                    65.0       42     88   122
## 4 Peace                         60.8       17     87   111
## 5 Physics                       57.7       25     96   229
## 6 Physiology or Medicine        58.9       32     87   232
```

From this we can see that on average, Economics Science laureates tend to be older. This raises another interesting question as to why this trend is the case. Considering also Economic Sciences has the least number of awarded laureates.

4. On average, which gender receives the higher prize amount?

```
laureates %>%
  filter(!is.na(gender)) %>%
  group_by(gender) %>%
  summarise(
    average.amount = mean(prize.amount.adjusted, na.rm = TRUE),
    smallest = min(prize.amount.adjusted, na.rm = TRUE),
    largest = max(prize.amount.adjusted, na.rm = TRUE),
    total = n()
)
```

```
## # A tibble: 2 x 5
##   gender average.amount smallest  largest total
##   <chr>           <dbl>    <int>    <int> <int>
## 1 female       9892634.  3161325 14930730    67
## 2 male         7904780.  3006134 15547541   923
```

We can see that on average, female laureates are awarded higher amounts, but there is a significant difference between the total number of male and female laureates.