APIs and Web Scraping Lab

David Gerard

2019-10-30

Learning Objectives

- Obtain data from an API.
- Scrape data from the web.
- Intense data cleaning.

NFL Arrests

Consider the API for arrests of NFL players: http://nflarrest.com/api/

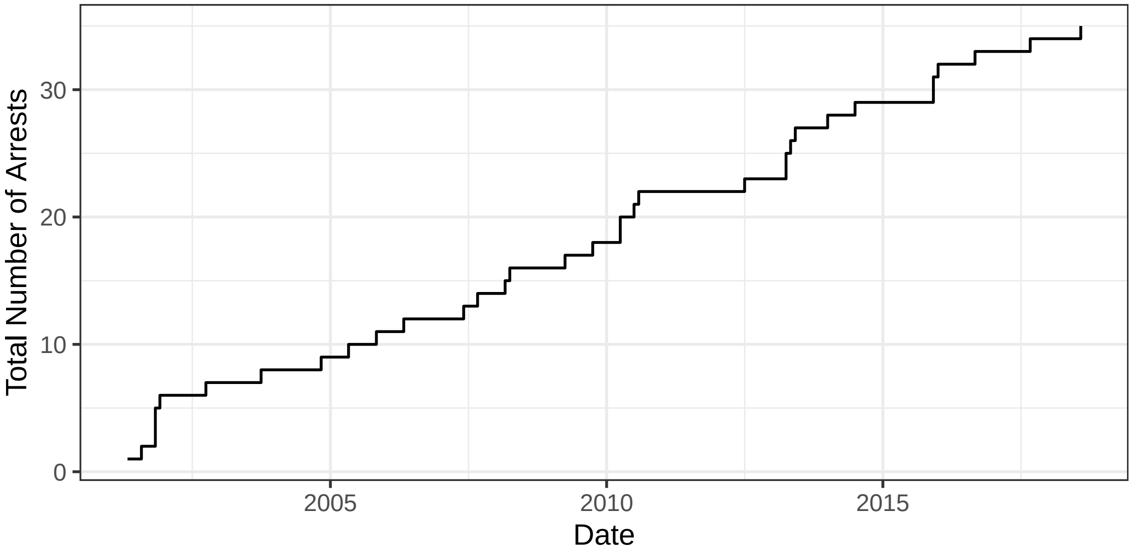
Hint: Don't forget that the `url` argument should start with http. Use the urls in the example queries in the help-page as guides.

1. Download the arrest counts for the Browns from 2001 to 2018.

2. Clean the data. Your data frame should look like this (I just made the day in the `Date` variable be on the first of the month for all dates):

```
## # A tibble: 30 x 5
##    Month  Year arrest_count Month_name Date      
##    <dbl> <int>        <int> <ord>      <date>    
##  1     5  2001            1 May        2001-05-01
##  2     8  2001            1 August     2001-08-01
##  3    11  2001            3 November   2001-11-01
##  4    12  2001            1 December   2001-12-01
##  5    10  2002            1 October    2002-10-01
##  6    10  2003            1 October    2003-10-01
##  7    11  2004            1 November   2004-11-01
##  8     5  2005            1 May        2005-05-01
##  9    11  2005            1 November   2005-11-01
## 10     5  2006            1 May        2006-05-01
## # ... with 20 more rows
```

3. Plot the cumulative sum by date. Your plot should look like this (use `geom_step()`):



4. There have been 29 players with at least 3 arrests since 2000. Get their names. You should get:

```
##  [1] "Kenny Britt"          "Adam Jones"           "Chris Henry"
##  [4] "Aldon Smith"          "Bryant McKinnie"      "Adam Jones"
##  [7] "Leroy Hill"           "Terry Johnson"        "Leonardo Carson"
## [10] "Fred Davis"           "Brandon Marshall"     "Larry Johnson"
## [13] "Eric Warfield"        "Chris McAlister"      "Bryan Robinson"
## [16] "Gerald Sensabaugh"    "David Terrell"        "Andre Rison"
## [19] "Joseph Jefferson"     "Sam Brandon"          "Reuben Foster"
## [22] "Kenny Mixon"          "Santonio Holmes"      "Albert Haynesworth"
## [25] "Jarrod Cooper"        "Johnny Jolly"         "Sebastian Janikowski"
## [28] "Vincent Jackson"      "Ray McDonald"
```

5. Clean the player data from part 4. Your data frame should look like this:

```
## # A tibble: 29 x 6
##    Name            Team  Team_name Team_city     Position arrest_count
##    <chr>           <chr> <chr>     <chr>         <chr>    <chr>
##  1 Kenny Britt     TEN   Titans    Nashville     WR       7
##  2 Adam Jones      TEN   Titans    Nashville     CB       6
##  3 Chris Henry     CIN   Bengals   Cincinnati    WR       6
##  4 Aldon Smith     SF    49ers     San Francisco LB       5
##  5 Bryant McKinnie MIN   Vikings   Minneapolis   OT       4
##  6 Adam Jones      CIN   Bengals   Cincinnati    CB       4
##  7 Leroy Hill      SEA   Seahawks  Seattle       LB       4
##  8 Terry Johnson   CHI   Bears     Chicago       DT       4
##  9 Leonardo Carson LAC   Chargers  Los Angeles   DT       4
## 10 Fred Davis      WAS   Redskins  Washington DC TE       4
## # ... with 19 more rows
```

Film Remakes

Consider the list of film remakes from Wikipedia: https://en.wikipedia.org/wiki/List_of_film_remakes_(A-M) and https://en.wikipedia.org/wiki/List_of_film_remakes_(N-Z)

1. Download the html file and save it as a variable. You can also load the "remakes_1.html" and "remakes_2.html" files in the data folder.
2. Extract the "table.wikitable" elements from both files.
3. Now obtain a single list of all of the table elements.
4. Create a single data frame with two columns — `Remakes` and `Original version`.
5. Create a data frame that contains the year of the remake, the year of the original, and the name of the original. Note that there are many films with multiple remakes.

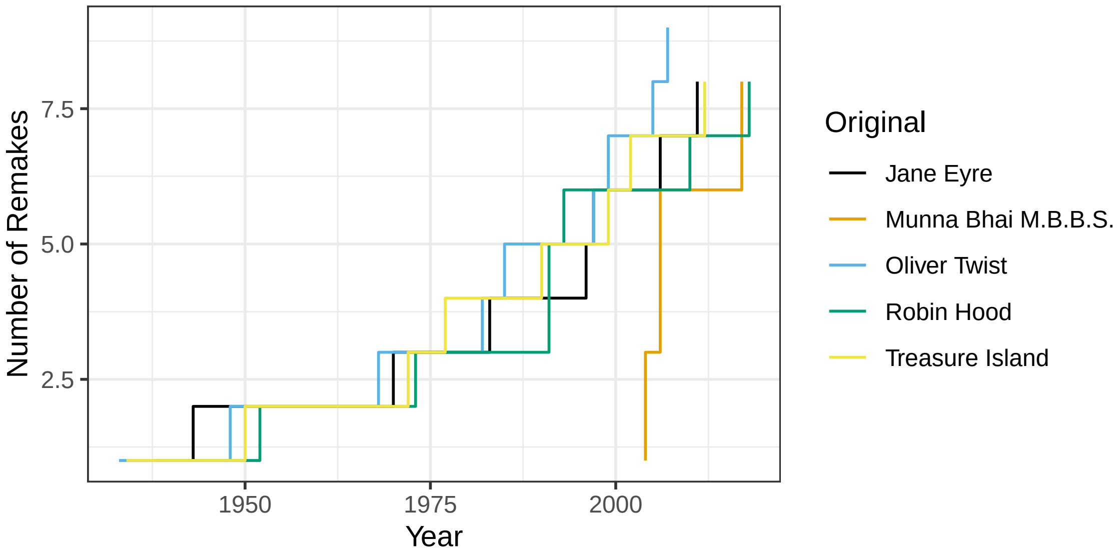
   Your final data frame should look like this:

```
## # A tibble: 973 x 3
##    year_rm name_ov                        year_ov
##      <dbl> <chr>                            <dbl>
##  1    2010 13 Tzameti                        2005
##  2    1951 Le Corbeau                        1943
##  3    1996 One Hundred and One Dalmatians    1961
##  4    2005 Two Thousand Maniacs!             1964
##  5    1948 The Three Godfathers              1916
##  6    1936 The Three Godfathers              1916
##  7    1930 The Three Godfathers              1916
##  8    1921 The Three Godfathers              1916
##  9    1919 The Three Godfathers              1916
## 10    2017 3 Idiots                          2009
## # ... with 963 more rows
```

6. There are two films remade in the same year as the original version. What were they?
7. Find the 5 movies that were remade the most number of times. You should get:

```
## # A tibble: 5 x 2
##   name_ov                 n
##   <chr>               <int>
## 1 Oliver Twist            9
## 2 Jane Eyre               8
## 3 Munna Bhai M.B.B.S.     8
## 4 Robin Hood              8
## 5 Treasure Island         8
```

8. Plot a step function for these movies by year. Your final plot should look like this:


Number of Remakes
Year
Original
Jane Eyre
Munna Bhai M.B.B.S.
Oliver Twist
Robin Hood
Treasure Island