Assignment 1: Using ggplot2 for visualization

Thomas Brambor

2019-02-10

Summer Olympics Medals over Time

Scenario

Imagine you are the data scientist at a respected media outlet – say the “New York Times”. For the upcoming Olympics coverage next year, your editor-in-chief asks you to analyze some data on the history of `Summer Olympics Medals by Year, Country, Event and Gender` and prepare some data visualizations in which you outline the main patterns around which to base the story.

Since there is **no way that all features of the data can be represented** in such a memo, feel free to pick and choose some patterns that would make for a good story – outlining important patterns and presenting them in a visually pleasing way.

The full background and text of the story will be researched by a writer of the magazine – your input should be based on the data and some common sense (i.e. no need to read up on this).

Provide **polished plots** that are refined enough to include in the magazine with very little further manipulation (already include variable descriptions [if necessary for understanding], titles, source [e.g. “International Olympic Committee”], right color etc.) and are understandable to the average reader of the “New York Times”. The design does not need to be NYTimes-like. Just be consistent.

Data

The main data is provided as an excel sheet, containing the following variables on all participating athletes in all olympics from 1896 to 2016:

- `ID` : a unique indentifier of the entry
- `Name` : name of the athlete
- `Sex` : sex of the athlete
- `Age` : age of the athlete
- `Height` : height of the athlete
- `Weight` : weight of the athlete
- `Team` : usually the country team of the athlete, with the exception of political accomodations, e.g. the “Refugee Olympic Athletes” team.
- `NOC` : national olympic comittee abbreviation.
- `Games` : year and season of games.
- `Year` : year of games
- `Season` : season of games.
- `City` : host city
- `Sport` : a grouping of disciplines
- `Event` : the particular event / competition
- `Medal` : the particular event / competition

For example, an `event` is a competition in a sport or discipline that gives rise to a ranking. Thus `Athletics` is the discipline, and the `Athletics Men's 100 metres` is a particular event.

In addition, you are provided with some additional information about the countries in a separate spreadsheet, including the `IOC Country    Code`, `Population`, and `GDP per capita`.

Tasks

1. Medal Counts over Time

a. Combine the information in the three spreadsheets `athletes_and_events.csv`, `noc_regions.csv`, and `gdp_pop.csv`. Note, that the `noc_regions.csv` is the set all NOC regions, while `gdp_pop.csv` only contains a snapshot of the current set of countries. You have to decide what to do with some countries that competed under different designations in the past (e.g. Germany and Russia) and some defunct countries and whether and how to combine their totals. Make sure to be clear about your decisions here, so that the editor (and potentially a user of your visualizations) understands what you did.

b. Calculate a summary of how many summer games each country competed in, and how many medals of each type the country won. Use that summary to provide a **visual comparison of medal count by country**.

Feel free to focus on smaller set of countries (say the top 10), highlight the United States or another country of your choice, consider gender of the medal winners etc. to make the visualization interesting.

Please provide one visualization showing an over time comparison and one in which a total medal count (across all Summer Olympics) is used. Briefly discuss which visualization you recommend to your editor and why.

Bonus Point: Currently, the medal data contains information on *each athlete* competing, including for team events. For example, in 2016 China received *12 gold medals for their women’s win in volleyball* alone. Since this is usually not how it is done in official medal statistics, try to wrangle the data so that *team events are counted as a single medal*.

2. Medal Counts adjusted by Population, GDP

There are different ways to calculate “success”. Consider the following variants and choose one (and make sure your choice is clear in the visualization):
- Just consider gold medals.
- Simply add up the number of medals of different types.
- Create an index in which medals are valued differently. (gold=3, silver=2, bronze=1).
- A reasonable other way that you prefer.

Now, adjust the ranking of medal success by (a) GDP per capita and (b) population. You have now three rankings: unadjusted ranking, adjusted by GDP per capita, and adjusted by population.

Visualize how these rankings differ. Feel free to highlight a specific pattern (e.g. “South Korea – specialization reaps benefits” or “The superpowers losing their grip”).

3. Host Country Advantage

Until the 2016 Rio Summer Olympics (our data ends here), there were 23 host cities. Calculate whether the host nation had an advantage. That is calculate whether the host country did win more medals when the Summer Olympics was in their country compared to other times.

Note, that the 23 host cities are noted in the data but not the countries they are located in. This happens commonly and often Wikipedia has the kind of additional data you want for the task. To save you some time, here is a quick way to get this kind of table from Wikipedia into R:

```r
library(rvest)
library(stringr)
wiki_hosts <- read_html("https://en.wikipedia.org/wiki/Summer_Olympic_Games")
hosts <- html_table(html_nodes(wiki_hosts, "table")[[8]], fill=TRUE)
hosts <- hosts[-1,1:3]
hosts$city <- str_split_fixed(hosts$Host, n=2, ",")[,1]
hosts$country <- str_split_fixed(hosts$Host, n=2, ",")[,2]
```

Provide a visualization of the host country advantage (or abscence thereof).

4. Most successful athletes

a. Now, let’s look at the most successful athletes. Provide a visual display of the most successful athletes of all time.

b. Choose one or two additional dimensions among gender, height, weight, sport, discipline, event, year, and country to highlight an interesting pattern in the data.

Interactivity

5. Make two plots interactive

Choose 2 of the plots you created above and add interactivity. Briefly describe to the editor why interactivity in these visualization is particularly helpful for a reader.

6. Data Table

Prepare a selected dataset and add a datatable to the output. Make sure the columns are clearly labelled. Select the appropriate options for the data table (e.g. search bar, sorting, column filters etc.). Suggest to the editor which kind of information you would like to provide in a data table in the online version of the article and why.

Technical Details

The data comes in a reasonably clean Excel dataset. If needed for your visualization, you can add visual drapery like flag icons, icons for sports, icons for medals etc. but your are certainly not obligated to do that.

Part of the your task will be transforming the dataset into a shape that allows you to plot what you want in ggplot2. For some plots, you will necessarily need to be selective in what to include and what to leave out.

Make sure to use at least three different types of graphs, e.g. line graphs, scatter, histograms, bar chats, dot plots, heat maps etc.

Submission

Please follow the instructions to submit your homework. The homework is due on Monday, February 25 at 5pm

Please stay honest!

Yes, the medal counts of the olympics have surely been analyzed before. If you do come across something, please no wholesale copying of other ideas. We are trying to evaluate your abilities in using ggplot2 and data visualization not the ability to do internet searches. Also, this is an individually assigned exercise – please keep your solution to yourself.