Milestone 1, COM-480 Data Visualization

Sylvain Lugeon Rodrigo Soares Granja
Benno Schneeberger

03.04.20

Contents

1 Dataset

The dataset we decided to use is "120 years of Olympic history athletes and results" which can be found on Kaggle [1]. It is a historical dataset on the modern Olympic Games which includes all the Games from Athens 1896 to Rio 2016. The data set has been scraped from sports-references.com [2] and is the result of an incredible amount of research by a group of Olympic enthusiasts.

The data which is stored in a CSV file contains 271116 rows and 15 columns. Each row corresponds to an individual athlete competing in an Olympic discipline. The columns are:

1. **ID** - Unique number for each athlete
2. **Name** - Athlete's name
3. **Sex** - M or F
4. **Age** - Integer
5. **Height** - In centimeters
6. **Weight** - In kilograms
7. **Team** - Team name
8. **NOC** - National Olympic Committee 3-letter code
9. **Games** - Year and season
10. **Year** - Integer
11. **Season** - Summer or Winter
12. **City** - Host city
13. **Sport** - Sport
14. **Event** - Discipline within the sport
15. **Medal** - Gold, Silver, Bronze, or NA

It will be the main dataset we are using to make our visualizations. We do not exclude using a second dataset to improve or add information on our visualizations, however, the data from our main dataset will likely be sufficient. By doing some exploratory data analysis (more on this in section 3), we were able to see that the overall quality of the dataset is very good and would require little or no preprocessing and data-cleaning. The only thing that deteriorates a bit the quality of the dataset is the missing values. We found out that about 25% of entries have a missing value in either the age, height or weight column. However, as the dataset is big enough and we are mainly interested in the big picture, it is still possible to create a relevant visualization. When thinking about the design, it will be important to keep in mind that some years contain more missing values than others.

2 Problematic

We choose the dataset "120 years of Olympic history athletes and results" because of our overall interest in sport and more particularly in the Olympic games which are arguably the biggest and most entertaining sports competitions. Moreover, we wanted to create some visualizations about an international topic and that could interest anyone. Finally, considering the actual situation with the Covid-19 crisis, we did not want to make a visualization using a dataset that is related to something negative such as disasters or pandemics.
We will make two main visualizations following two different axes:

- The first axis is a comparison of medals and participation of athletes in the past Olympic games. The goal of this visualization is to obtain a clear overview of the performances of a country, a region or a continent. Those performances can be compared with another country, region or continent at a specific Olympic game or over a given span of time. It will be aimed for the general public, at people that are not necessarily interested in sports but at least curious, for example, about a nation or a particular period of time.

- The second visualization is focused on the physical characteristics of athletes depending on the sport they are playing. The characteristics we are comparing are morphological ones such as height and weight as well as the sex and age. It will allow, for example, to see the morphology of medal-winning athletes over the years for a given sport. One could also observe the average characteristics of participating athletes of a particular country or region. This visualization is aimed at people that are more deeply interested in sport and are curious about the physical characteristics of athletes for different sports over the years. It is more oriented in a scientific way and will contain some specific data about the athletes that is not necessarily interesting for the general public.

3 Exploratory Data Analysis subsection

We first assessed the importance and distribution of missing values. We then explored the dataset to get some global insights as well as some in relation with the two axis of the problematic. The full EDA is done in the Jupyter notebook provided in the github repository. We'll summarize here the main points of the EDA.

Missing values

- The only features that contain missing values are what we will call *physical features*, those are the age, the weight and the height of the athletes.

- Over all the entries, 24% have at least one missing physical feature. The weight and the height are missing for 25% of the athletes, while the age is missing for 5% of them.
- Missing values are non-uniformly distributed along the years and the different sports.
- It is not possible to infer the missing values from the rest of the dataset.

Global insights

- In the past, Summer and winter Games were taking place the same year, every four years. Since 1994, the editions are organized every 2 years, alternating between summer and winter editions. No Games took place during both World Wars.
- There is a total of 66 different sports (e.g Athletics), each sports having then different disciplines (e.g Marathon or 4 x 100 metres Relay for Athletics).
- The number of competing athletes is globally increasing over the years.

Insights relative to our problematic

- The total number of athletes and medals by countries is very disproportionate, with the USA having far more athletes and medals than any other country.
- Within the USA, the number of athletes and medal is increasing. Nevertheless, the ratio medals/athletes (still within the USA) seems to be constant since 1950.
- If we take as example the Athletics, we can observe that the number of athletes is also growing. Averaging over all the athletes in that sport, we can see a evolution of the physical characteristics; age is increasing while height and weight are decreasing. However, if we look at specific disciplines inside Athletics (e.g Marathon or 4 x 100 metres Relay), the evolution of the physical characteristics can be different.

4 Related Work

Our data set is quite known on the web. There are many tutorials on exploratory data analysis that make use of it, such as [3][4][5][6]. Most of them aim to show simple graphs and general processing of the dataset, while others [7][8] actually created data visualizations.

The first example is a simple way of visualizing medal count per country. We find in the second example, which is more rich and interesting, both axis that

we seek to develop: physical characteristics and performance comparison. We can see that visualization of top performers (medals per countries/athletes) has already been done using simple graphs and world map. Graphical visualisation of physical characteristics in function of time is also present. As you can see, the examples provided are very simple but still a good starting point for our project.

Concerning the originality of our approach, as seen before, most of the work we found rely on simple data exploration and analysis using graphs and more static methods of visualisation, while we seek to provide a dynamic and interactive experience to users. We aim to use data visualization in a way that users can interact with the data set using a minimal and intuitive interface. Interacting with the data, as we see it, allows for the user to extract the information we seek to provide in a playful manner.

We looked for inspiration in simple data visualizations that were also original and easy to interact with. Since we are working with countries and cities we wanted to find ideas for interactive maps that were pleasant to look at. We came across [9] which contains examples of minimal interactive maps that we could further improve in our project. We also found an interesting way of implementing packed bubbles by giving it an interesting form, in our case we could approach that form to the logo of the Olympics in a similar way as is done in [10][11].

References

[1] dataset - 120 years of olympic history athletes and results. https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results.

[2] Website from where the data used in the dataset originates. https://www.sports-reference.com/olympics/.

[3] Exploratory data analysis 1. https://www.kaggle.com/gpreda/plotly-tutorial-120-years-of-olympic-games.

[4] Exploratory data analysis 2. https://towardsdatascience.com/120-years-of-olympic-history-analysis-cba815736ea9.

[5] Exploratory data analysis 3. https://www.kaggle.com/ahirulka/120-yrs-olympic-history-analysis-eda.

[6] Exploratory data analysis 4. https://rstudio-pubs-static.s3.amazonaws.com/536660_dac59a0816ca4c229f44ed3be2436873.html.

[7] Example of data visualization - olympic statistics. https://public.tableau.com/profile/stefano.maglietta#!/vizhome/OlympicGames-HowHasYourCountryPerformed/OlympicStats.

[8] Example of data visualization 2 - olympic statistics. https://nycdatascience.com/blog/student-works/olympic-games-data-visualization/.

[9] Example of minimalist data visualization with a map. http://www.puffpuffproject.com/languages.html.

[10] Example of data visualization with packed bubbles. https://www.pinterest.ch/pin/144396731786303034/.

[11] Example of data visualization with a graph. https://www.pinterest.ch/pin/529454499922501337/.