Data Visualization Project

The evolution of the Olympic Games

Danilo Arfelli | m20211296
Diogo Tomás Peixoto | m20210993
Gabriel Avezum | m20210663
João Morais Costa | m20211005

https://olympic-evolution.herokuapp.com/

https://github.com/gavezum/Data-Visualization

Dataset description

Our group elected the Olympic Games dataset (available in Kaggle [1]), as the base dataset for the Data Visualisation project. The dataset gathers information about all athletes, and not just medal winners, that participated in all modern Olympic Games, from 1896 until 2016, in both *Summer* and *Winter* games. There were considered different scopes, but the final choice was to use a visualization to help understand the *Evolution of the Olympic Games*, due to the historic perspective inherent to the dataset. As a complement to the main dataset, the original folder also contained an auxiliary dataset ("NOC regions") that contained the National Olympic Committee (NOC) codes and its corresponding countries' names.

Variable	Type	Description
ID	Int64	Listing ID
Name	object	Athlete's name
Sex	object	Athlete's gender – "M" or "F"
Age	float64	Athlete's age
Height	float64	Athlete's height *(in centimeters))*
Weight	float64	Athlete's weight *(in Kilograms)*
Team	object	Athlete's team
NOC	object	National Olympic Committee (3 letters code)
Games	object	Corresponding Olympic Game
Year	Int64	Year of the corresponding Game
Season	object	Season of the corresponding Game ("Summer" or "Winter)
City	object	City where the Game was released
Sport	object	Olympic Game sport
Event	object	Olympic different inside a given sport
Medal	object	Either "Gold", "Silver", "Bronze" or NaN.

Table 1 - Dataset metadata

Visualizations and user interactions

The first interaction the user can experience is the choice of the games' season, by using the corresponding *radio button* on the right top corner of the dash border, Summer, or Winter. The decision to include this filter, was based on the different nature that each season has on the Olympic experience. The activation of this button not only filters the data fed into the visualization, but also changes the background colour and some of the plots channels. This allows the user to experience a layout environment that enhances the type of Olympic season portrayed. Other interactions were created specifically for each plot, and will be detailed on further steps.

Visualization 1 - Line plot

To get a first sense of the evolution of the games during the last decades, we wanted to understand how this evolution has been in terms of the number of athletes, nationalities present, number of sports and the number of events. Worth mentions that according to the International Olympic Committee, a *sport* is what is governed by an International Federation, and an event is a competition in a sport [2]. So, for each unique sport there are several different events. To visualize this information, we considered that the simplest and most effective way to address this step was through a line and dot plot. To avoid a clustered visualization and different data scales being displayed simultaneously, a dropdown menu was introduced to allow the user to choose between the four variables referred before. Together with the chosen line chart, extra information regarding the present chart is displayed on the left side of the box.

Structurally, the *x-axis* of the plot is fixed with the ordered years of Olympics games attribute, whilst the *y-axis* varies according with the chosen variable. One example is displayed below. On this plot the visual marks are the lines and dots, while the visual channels are the X and Y position. To complete the visualization, was also included a *hover box* with text that displays the year and the city where the corresponding game happened. To add extra contextual information, different non-interactive annotations were added to the plots in the form of lines, colour bars or colour bubbles. These annotations have a corresponding reference in the side text.

Visualization 2 - Stacked Bar plot

Olympic games are all about winning or losing. However, there isn't an official ranking system to determine which nation is the winner of each game. While some use the number of gold medals as the criteria, others determine the winner by the total number of medals won.

We wanted to understand which nations are the biggest winners along the Olympic history. In this project we used as criteria for the game winner, the nation with the highest number of gold medals. Taking this into account, the process of determining the most frequent winners started by calculating the nation's podium at each Olympic game, based only on the corresponding number of gold medals. Subsequently, the number of podiums of each nation was added up following each Olympic game, with the number of nations ordered by the total number of podiums. The visualization chosen to answer this question was the stacked bar plot. It encompasses two categorical attributes: the nations and the podiums positions. As well, one quantitative attribute: the number of podiums.

Please refer to the example below (figure 1), which displays an example of the cumulative results until the 2016 Summer Olympic game. For the sake of clarity, the United States was the country with more wins, with a total of 28 podiums in 28 different Olympic games. Out of all summer Olympics events that took place between 1996 and 2016, the United States had a total of 17 first places, 8 second places and 3 third places.

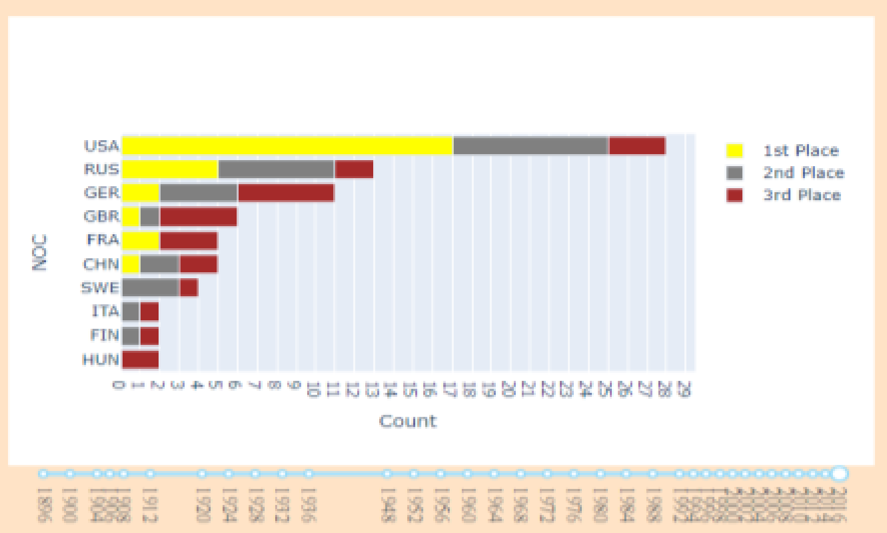

Figure 1 - Stacked bar plot example

The dataset type used to build the visualization was a table. As visual marks are lines and the visual channels used are the length/position of each bar that pinpoint the nations ranking order and different colours that point out the ranking podium.
The visualization also includes a hover box with text that displays the country name, corresponding number of podiums by position. The user can also experience interaction by using a *slider*, that choses the time boundaries for the plot' data, which will output the cumulative results inside the chosen temporal window.

Visualization 3 – Choropleth map plot

The Olympic games are fuelled by rivalries between countries, where the boundaries between sport, economy and/or politics are often blurry. To understand if the participation of the different countries and regions were similar or not, a choropleth map plot was built to tackle this question. The plots were built based on data processed from the original dataset, that capture the number of participations of a specific NOC in each game on the period from 1896 and 2016. To translate the NOC code to the corresponding country, the "NOC regions" auxiliary data frame was used.

The user experience is done by the traditional choropleth map zoom tools from *plotly*, and by using a *dropdown* menu. This latter which allows the user to focus the scope of the map to a specific region. The regional options are: "World", "Europe", "Asia", "Africa", "North America" and "South America". We believe that such a regional/continental scope view option allows a deeper understanding of possible differences between regions and even between countries within each region. The presence of a hover box completes the information available. This box can be accessed when the user passes the cursor over the country and displays the country's name together with the exact number of presences on the corresponding games' season.

As stated before, each visualization has a *"Summer"* and *"Winter"* version. In this case, we used a red colour scale for the Summer Olympic Games and a blue colour scale for the Winter Olympic Games. Both scales have the same core principle: the darker the colour, the higher is the number of participations that a specific country had in the history of the Olympic Games. The upper boundary of each scale varies according to the season chosen previously. In the case a country has never participate in the games, its absence on the map is represented by the absence of colour and frontier demarcation on the graph, therefore appearing solely the default terrain colour, grey.

As data attributes, were used the country name and number of presences, while the country area represents the marker. The channels used were:

- *Magnitude channels* that order attributes, such as area of a country and colour saturation of the colour scale used.
- *Identity channel*s such as the base colour of the colour scale. This identifies the season for which data refers to.

Visualization 4 – Parallel plot

Where the early Olympic athletes different from the current ones? Has the weight and height of the athletes on each sport changed over time? To tackle this and other questions we aim at building a visualization that allowed us to select different sports and different metrics and plot them with a *time-effect* perspective.

So as the fourth visualization of our dashboard, we choose a parallel plot, with the option for the user to choose between three different metrics: *Age*, *Height* and *Weight*. The user can further interact by choosing one or more sports (and the corresponding events), allowing not only to understand the evolution of the select metrics across different decades, but also to compare the metric along different sports. One additional characteristic of this visualization is that it allows us to understand in which decade the sport or event was first introduced in the games and/or was removed from the games.

The parallel graph is grouped by decades represented by the parallel vertical lines, and the point in which the line crosses the decade bar corresponds to the average value of that metric on all the athletes that participated in that event in that decade. The boundaries, upper and lower, correspond to highest and lowest values of the average for that event, sport of all decades corrected with 5 extra added to highest and 5 extra subtracted to the lowest value. Every time the user adds an extra sport, and its corresponding events are added to the graph, all the lines corresponding to those events are grouped by the same colour. This helps distinguish different events from different sports when they are plotted together.

The marks of the plot are the vertical lines that correspond to the decades and the horizontal lines that correspond to the evolution of the metric along the years. As channels, the plot makes use of coloured lines that correspond to different events of a sport.

Technical aspects

All code used on this project was done using Python language. The original dataset has been pre-processed to build up the visualizations presented above. We have used *Jupyter notebook* during the processing and transforming phase, with the final tables exported as csv files which were added to the *Github* portfolio afterwards. The source code was developed in the IDE *PyCharm* using the libraries *Plotly* and *Dash* to build the interactive dashboard. The source code is composed of three main parts. The first is the definition of the interactive components, including, the dropdown, the slider, radio items, etc. The second part refers to the structure of the app page itself. The last part contains the *callbacks* together with the respective functions, which create the visualizations and respond to the user interactivity updating the charts.

Finally, the previously created app is uploaded in the cloud platform HEROKU, whereby becomes permanently available through an URL (https://olympic-evolution.herokuapp.com/)

Discussion

What we have accomplished

We have managed to build four different visualizations that together allow the user to understand the history of the Olympic games. We believe the dash and its visualization helps portray: the link between political worldwide events and its' consequences in participation of different countries and athletes' participation in the games; The most awarded countries in the history of the Olympics; The geographic discrepancy between country representations in the games for different globe regions; and evolution of the athletes on several metrics for different sports along the time.

We believe that the visual design is well structured, with the content properly split for the user to grasp it easily. The messaging is straightforward either through the use of visualizations solely or together with small text boxes attached to it. The app design is user driven with different and *easy-to-use* features that allow the user to engage in the app experience.

In addition, we have partly validated the interactive visualization by checking the app time responsiveness to different inputs. It reacts fast and so the algorithm might be considered well built.

Limitations found

Some limitations were found and recognized on this part of the report, namely concerning the choropleth map and parallel plot.

Regarding the choropleth map, two forms of geometric information were considered: the built-in geometry within *Plotly* and by using the *GeoJSON* file. While the latter allows better quality visualization using different *mapbox* styles, such as *open street map*, during our testing phase, the browser time response was considerably longer. Due to this fact, we decided to plot our choropleth map using the built-in geometry from *Plotly*.

Regarding the parallel plot, after implementation it was noticeable that the visualization became harder, if many sports are selected in the dropdown menu, namely sports with a large number of events, such as gymnastics. An alternative option would be to allow only a single sport to be chosen for each plot, however, it would remove the visualization possibility to compare different events from different sports.

Future Work

For the app to be commercially implemented, further validation levels would have to be checked. It could be accomplished by measuring the time each user remains in the app, the number of errors/bugs transmitted, through online questionnaires or even ultimately by performing interviews. The user's overall feedback would guide us to finetune the app.

References

1 – 120 years of Olympic history: athletes and results – Basic bio daa on athletes and medal resuls from Athens 1896 to Rio 2016. Available: https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results

2 – “The Olympic programme comprises sports, disciplines and events – what is the difference between the three?” – IOC: International Olympic Committee. 2021. Available: https://olympics.com/ioc/faq/sports-programme-and-results/the-olympic-programme-comprises-sports-disciplines-and-events-what-is-the-difference-between-the-three