Data Visualization Report

Interactive visualization of the Olympic Games stats

Group elements identification Group_A			
Name:	Student number	Name:	Student number
Catarina Moreira	m20201034	Luísa Barral	m20201045
Madalena Valério	m20200657	Tomás de Sá	m20200630

Dataset Description

The dataset used was the ***athlete_events.csv*** from the link ***https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results?select=athlete_events.csv*** . The dataset contains all the participants in the olympic games from the first game to the last one in 2016. Each row of the dataset tells us the name of the athlete, their sex, age at that time, height, weight, the national team they are representing, the games they were playing in, the year, the season (Summer or Winter games), the city the games took place, the sport this athlete was competing in, the event the athlete was competing in, (Ex: "Judo, Judo Men's Extra-Lightweight) and what type of medal he/she received, if any.

This dataset needed some work so the group made some featuring engineering in order to obtain the best visualizations possible.

First we started by analyzing the dataset and we verified that the 'Medal' column had Nan values. We then proceeded to replace these values with the string 'No Medal'. Then, we also verified that both the 'Height', 'Weight' and 'Age' columns contained missing values. In this way, we chose to use KNN both in the 'Height' and 'Weight' column. Since 'Age' had an approximate 3% percentage of missing values, we ended up eliminating these values. Then, we decided to merge the obtained dataframe with the ***noc-regions.csv*** dataset, which was also in Kaggle. Thus, it was possible to obtain a new column with the name of each region / country. Finally, we created the following columns: 'Sem_Medalha', 'Gold', 'Silver', 'Bronze' and 'Total_Medals'. These new columns are binary and show us which athletes did not win a medal, who won a gold, silver or bronze medal and those who simply obtained a medal. An 'Interval_Age' column was also created that indicates the age range of the athlete.

In this way, we completed the cleaning and feature engineering of our dataset and the final dataset used was result.csv.

Visualization & Interaction Choices

The group took inspiration from two graphs available in the following two links, http://data-visualization-project-maa.herokuapp.com/ and http://spotify-dash-project.herokuapp.com/ . The group tried to merge both ideas and came up with a dashboard with 3 separators called "World", "Teams" and "Athletes" in which we implement different visualizations to show the viewer the most interesting information extracted from our selected dataset.

The interactive visualizations implemented by the group were done so in order to help the viewer get a clear understanding of the Olympic Games data such as the most victorious countries across the world by medal type or the most winning athlete given a certain sport and year scale. In our dashboard we implemented visualizations like a choropleth map, bar plots, scatter plots and line plots, all done so to understand the information extracted from our dataset.

Starting with the first separator, the user begins by choosing the type of medal that they want to view. They can choose between viewing the total of gold medals, the total of silver medals, the total of bronze medals, the total of medals without making this distinction and the total of athletes who have not won a medal. After making their choice, a choropleth map is presented showing the different countries in the world and in each country the number of medals is displayed. This map has a scale so that it is possible to compare this same number.

In the second separator, the interface starts by presenting two dropdowns where the user can choose one/several countries and one/several sports. Both countries and sports are presented in alphabetical order. Then the user can, using the keys on the computer or with the mouse, choose a year. The number of athletes, the total number of medals, the total number of gold medals, the total number of silver medals, the total number of bronze medals and the total number of medals are shown without making this distinction. These values are being changed as we introduce new countries, new sports and different years. When choosing two or more countries or two or more sports, the values shown are the sum of the values for each country or/and each sport. It also shows the city where the Olympic Games of the year that the user chose took place. If there are no Olympic Games in the chosen year, the phrase: 'Year without Olympic Games' is displayed.

Subsequently, two graphs are presented. The first is interactive and depends on the chosen country/countries, the chosen sport/sports, the chosen year (we used a slider so that it was easy and intuitive for the user) and the type of medals chosen. (The type of medals can be: Gold, Silver, Bronze, No_Medal or Total Medals)(radioItems was used to choose the type of medals). This graphic then shows the number of medals of the type chosen by the user by age and sex . That is, athletes are grouped by age group and by sex. This age group was chosen by us so that there was a good visualization. In the second one, the number of gold, silver and bronze medals by country, sport and year chosen by the user is displayed. These two graphs are bar plots.

Next, more general views are presented of the world Olympic records. Then, 3 graphs are presented. These 3 graphs depend only on the year chosen by the user. In the first, the 10 countries that obtained the highest number of medals in that year are presented. In the second, the 10 sports that had the largest number of participants are presented. Finally, a graph is presented showing the number of participants in each sport that year but grouped by season, that is, by Summer Olympic games and Winter Olympic games. These last 3 graphs are scatter plots.

If there are no values, or if the Olympic Games are not held in the chosen year, empty graphics are shown.

In the last separator named "Athletes", the user can choose one country, one sport and the Year Range they want to see the data from. The country Portugal and sport Swimming is already selected, as well as the year range slider from 1896 to 2016. There are three graphs. The bar plot, which shows the events that exist for the sport chosen, and the number of participations that existed for each event for the country and range of years selected. There is also a table with the names of the top 10 athletes with the most medals won. In some cases there are countries that have never won medals in a particular sport. Finally, there is an area chart that shows the percentage of participation for each gender per year, for the specific country, sport and year range, keeping in mind that some countries and sports don't have records for all the years the user chose. In case the country has never competed in the Olympics with a sport selected by the user, the output is three empty graphs.

Technical Aspects

All the code implement in order to obtain our visualizations and dashboard is available in the link **https://github.com/jtbpc-iscteiul/DataVisualizationDashboardGroupA**

The group implemented the visualizations using pycharm with plotly and dash. As you can see in the code presented in the github repository, the group implemented different graphs with different types of updates, that is, each graph focuses on different aspects of our dataset and provides the viewer with different *dash components* such as dropdowns, radioitems and sliders.

APP LINK -> https://olympicgamesdv2.herokuapp.com/

Discussion

Now that we have our full dashboard working, the group can be satisfied with what we have accomplished. We successfully implemented dropdowns, radioitems and range slider as well as correctly implemented the callbacks and graph updates given certain inputs.

Like every good school project this interactive visualization dashboard provided us with some challenges, firstly with the dataset itself, then with the way we were supposed to organize our final dataset in order to obtain the desired visualizations and lastly with "glueing all together" in a html page. The way the data was presented and the modifications necessary obliged us to, sometimes, take a step back and reorganize our line of thought. The classes code examples were really helpful, even though sometimes we required something that we could not find in the classes repository so the group had to investigate and explore some visualizations and techniques.

Lastly, as for future work, we would suggest and investment in the code implemented for the visualizations as the group thinks they are good but could be even better with some embellishments and more clear caption inside the graphs. As for the work relative to the information extracted and the dataset provided, the group doesn't feel like this dataset could get anymore information, since we can already extract a lot of information from the beginning of the Olympic Games until today .

References

1. Medium. 2021. *How To Visualize the Coronavirus Pandemic with Choropleth Maps*. [online] Available at: <https://towardsdatascience.com/visualizing-the-coronavirus-pandemic-with-choropleth-maps-7f30fccaecf5>
2. Dash, H. and Romero, M., 2021. *How to update choropleth map in Dash*. [online] Stack Overflow. Available at: <https://stackoverflow.com/questions/56086466/how-to-update-choropleth-map-in-dash>

3. Medium. 2021. *Area Charts with Plotly Express*. [online] Available at: <https://towardsdatascience.com/area-charts-with-plotly-express-510a1f12ac11>

4. Plotly.com. 2021. *Bar Charts*. [online] Available at: <https://plotly.com/python/bar-charts/>

5. Plotly.com. 2021. *Tables*. [online] Available at: <https://plotly.com/python/table/>

6. Plotly.com. 2021. *Setting the Font, Title, Legend Entries, and Axis Titles*. [online] Available at: <https://plotly.com/python/figure-labels/>

7. Iban.com. 2021. *List of country codes by alpha-2, alpha-3 code (ISO 3166)*. [online] Available at: <https://www.iban.com/country-codes>

8. Plotly.com. 2021. *Plotly Python Graphing Library*. [online] Available at: <https://plotly.com/python/#fundamentals>

9. Plotly.com. 2021. *Scatter Plots*. [online] Available at: <https://plotly.com/python/line-and-scatter/>

10. Plotly.com. 2021. *Styling Markers*. [online] Available at: <https://plotly.com/python/marker-style/>

11. Plotly.com. 2021. *Plotly Express*. [online] Available at: <https://plotly.com/python/plotly-express/>

12. Plotly.com. 2021. *Styling Markers*. [online] Available at: <https://plotly.com/python/marker-style/>

13. Plotly.com. 2021. *Sliders*. [online] Available at: <https://plotly.com/python/sliders/>

14. Plotly.com. 2021. *Horizontal Bar Charts*. [online] Available at: <https://plotly.com/python/horizontal-bar-charts/>