CP5 – The Olympic Games And How Being The Host Affects The Difference Of Medalists

António Filipe
92425
a.angeja.filipe@tecnico.ulisboa.pt

Maria Ribeiro
93735
maria.f.ribeiro@tecnico.ulisboa.pt

Catarina Sousa
93695
catarinasousa2000@tecnico.ulisboa.pt

INTRODUCTION

Our problem domain consists in the Olympic Games. We have all data from all the athletes and hosts from 1896 until 2016. The Olympic Games are very famous worldwide, every country wants to win medals and be well-represented in the world. Every athlete wants to represent his country in the Olympics, and they fight for it in the 4 years leading to the Olympic Games. It's every athlete's dream to participate in this competition. A lot of people don't watch the "normal" competitions, but they watch the Olympics. This competition moves a lot of athletes and watchers throughout the world. All these facts made us believe that this topic was very interesting to approach, and our objective was to make a visualization that shows everyone the performance of the countries and their national teams, giving the users insights about the number of participants and their gender, the number of medalists and the influence that being the host country has in succeeding in the Olympic Games.

The questions we wanted to answer were:

1. How does the number of medalists in the Olympic Games (Summer and Winter) vary according to the participants, per NOC (National Olympic Committee)?
2. How does the gender of medalists in the Olympic Games (Summer and Winter) vary, per NOC (National Olympic Committee)?
3. How being the host country of the Olympic Games affects the number of medalists in that country?
4. How does the number of participants in the Olympic Games (Summer) evolve throughout the years?
5. How does the number of women participants in the Olympic Games (Summer) evolve throughout the years?

We decided to change our project only for Summer Olympic Games. Firstly, we did it for Summer and Winter, but then we realized that there was a big discrepancy between the south countries and the north countries. For example, the difference between South Africa and Norway would be much bigger, because South Africa doesn't participate in the Winter Olympic Games. And, since the Summer Olympics are the most important, we thought that our decision was appropriate.

For our 4th and 5th question, we decided to expand it to all the countries, so that is possible to see the evolution not only in a general way but also per country.

With this visualization, we discovered many things about the Olympics that we didn't know. For example, by looking at our data we realized that there were no values in 1916, 1940 and 1944. We started to think about this and learnt that the Olympic Games didn't occur in these years because of World War I and World War II. Another thing that we searched for was the explanation for the decrease of the number in participants in some years, and we learnt that due to political issues, some countries didn't participate in protest.

RELATED WORK

We searched for many graphs in d3-graph-gallery to easily understand how to implement them.

We looked at the projects in the Hall of Fame and we enjoyed the one implemented by Afonso Luís, Joana Sesinando, and Tiago Delgado (*Visualizing Personality Tests*). The idea of creating a sidebar was based on this project. We also got inspiration in this project to create the aesthetics of ours.

THE DATA

We downloaded the datasets from Kaggle. We found one dataset that had information about all the athletes that participated in the Olympic Games from 1896 until 2016. This dataset was very complete and allowed us to answer every question we wanted to answer, except for the 3rd one. To answer the 3rd question, we found a dataset, also on Kaggle, that had all the hosts from the Summer Olympics. The first dataset had the NOC and the name of the team of each athlete, but the name of the team isn't always the name of the country, so we needed to download another dataset from Kaggle to correlate the NOC and the country it represents. We needed to do this to be able to connect the choropleth map with the other idioms. We ended up having all the datasets we thought we needed in the beginning, making it easier for us to use this data.

In terms of scalability, we had some problems developing the two Cleveland Dot Plots due to the number of NOCs that exist. In the beginning, we wanted to have all the NOCs in the overview, but since there are so many NOCs, it was impossible to do this, so we decided to show in the overview the most relevant ones, which means that only the countries with more than 5000 participants in all the Olympics appear.

In order to answer our questions, we had to calculate some derived measures. We needed to sum the participants generally and per country (female + male, only female and only male), the medalists generally and per country (female + male, only female and only male) and the percentage of medalists, per NOC and per gender. In our Checkpoint II we ended up removing the columns that consisted in the number of participants per country (they were only intermediate calculus to reach the percentage), but since we decided to see the evolution of participants also per country, as we mentioned before, we needed to revisit our data to get back these calculi.

VISUALIZATION

Overall description

Our visualization layout consists in dividing the screen, almost equally, into two rows, the top row is divided in two idioms, the Choropleth map, which occupies about 55% of the top row on the left, and the Progress Bars, which occupies about 42% of the top row on the right. The bottom row is divided into three idioms, a Line Chart on the left, which occupies about 44% of the bottom row, and two Cleveland Dot Plots, which each occupies about 20% of the bottom row on the right. Besides that, on the left we have a sidebar with the title "The Olympic Games And How Being The Host Affects The Difference Of Medalist", with the screen height, which contains a dropdown button, which allows the user to select a country, and 2 other buttons which allow the user to filter the data to a specific millennium: "Before 2000" and "After 2000" and another one that allows the user to see all of the years; this sidebar also has extra information about the many idioms represented and about the data.

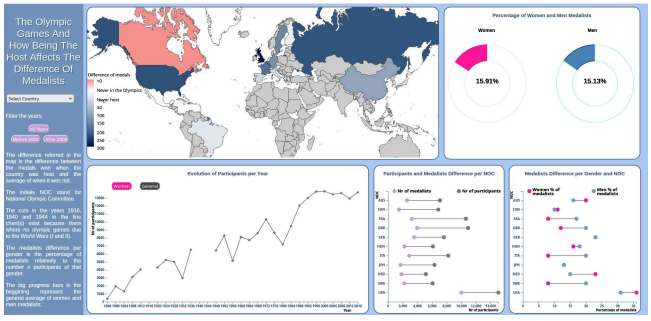

Figure 1 – overview of the visualization

This implementation allows the user to compare Countries/NOC's up until the limit of four NOCs simultaneously. The user can select a country or NOC by clicking it in one of the idioms or by selecting it in the dropdown button, the same logic applies to deselecting a country or NOC.

In the beginning, the visualization starts with no filter regarding the years, which is the equivalent to the "All Years" button. Besides that, the Line Chart starts with the data for all the countries summed representing all the participants, and the progress bars represent (separately) the general percentage of women and men medalists according to the number of participants of that gender. The Cleveland Dot Plots start showing the countries that have more than 5000 participants, that is, the most relevant NOC's (National Olympic Committees).

The Choropleth Map implemented encodes the difference between the medals a country won in the year(s) in which it was the host and the average of medals that the same country wins overall, using the color of the country to represent this value. The countries that never participated in the Olympics are filled with the color white, the countries that were never host, and therefore don't have this difference to be encoded, are filled with a light grey hue. For the countries that were somewhere in time, hosts of the Olympics, we chose a red hue to represent the ones with a negative difference of medals and a scale of blue tones, in which the lighter the blue the less the difference (starting at 0), and the darker the blue the bigger the difference (with a maximum difference encoded of 300 medals). This color encoding is easy to understand by the user, since there is a legend right on the Choropleth Map, that shows what each color means. This idiom allows the user to zoom-in and out, by using the touchpad or the scroll on the computer mouse, at the beginning the user gets a more overall and zoomed-out view of the map. When interacting with this idiom, if we hover over a country, the opacity of the rest of the map lowers, so the user can easier focus on the country he is hoovering; while hoovering a country, a tooltip appears, giving us information about that same country: it always shows the name of the country first, then if the country never participated in the Olympics the following message is shown: "This country was never in the Olympics in the interval chosen"; if the country was never host the message shown is "This country was never host in the interval chosen"; if the country hoovered has been host, in the message it is shown the exact value of the difference and the year(s) in which that country was host of the Olympics. On clicking a country in the Choropleth Map, if the country is not already selected, it gets selected for comparison, and to represent that it is selected the stroke of that country gets thicker in the map. This selection, triggers changes in the other idioms: in the Line chart is shown a line that represents the evolution of participants for that country only, and in the Cleveland Dot Plot's and Progress Bars are shown the values for each NOC of that country. If we click

a country and it was already selected, it gets deselected, its stroke goes back to normal and the other idioms update, ceasing to show information about that country and its NOCs.

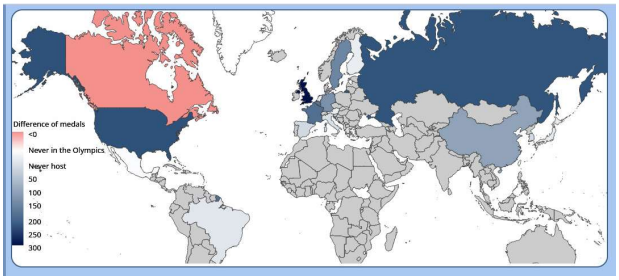

Figure 2 – Overview of the Choropleth Map

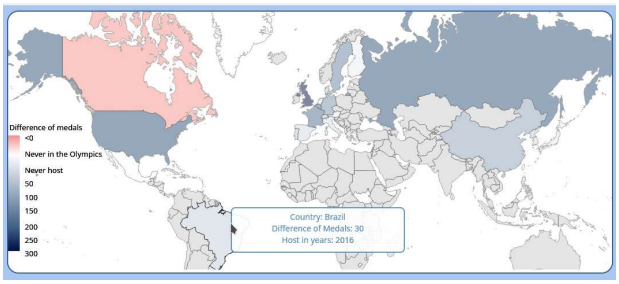

Figure 3 – Tooltip in the Choropleth Map

The Line Chart implemented encodes the evolution of the number of participants in the Olympics, representing the years (in which there were Olympics) in the x-axis and the number of participants in the y-axis. The x-axis updates when a millennium filter (referred before) is used. This idiom allows filtering within itself, allowing the user to choose between seeing the evolution of the general number of participants or only the evolution of women participants, using two buttons: "General" and "Women", the y-axis updates depending on which of the filters is being used. When a country is selected in one of the other idioms, the initial line with the information about all participants is removed, and a line will be shown in the line chart, encoding the evolution of participants for that country. When there are countries selected it is still possible to filter to only women participants and filter back to normal. If we click a line in the Line chart (except the general line that appears in the beginning), the country that line is representing gets deselected, the other idioms also update, ceasing to show information about that country, and the line disappears from the Line chart. For each line representing a country created in the line chart, it is attributed a color to it, and a dynamic legend is shown. Each line also has a dot in the x position of each year when there were Olympics. If we hover over a line, we get the name of the country that line represents, and the line gets thicker to indicate that we are hovering over it. If we hover over the dots in the Line chart, we get a tooltip with the exact number of participants in that year.

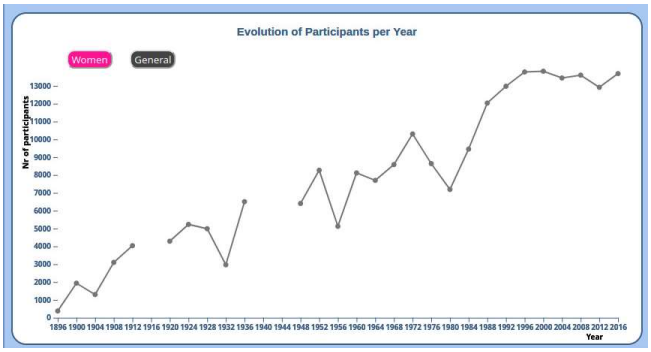

Figure 4 – Overview of the Line Chart in the General selection

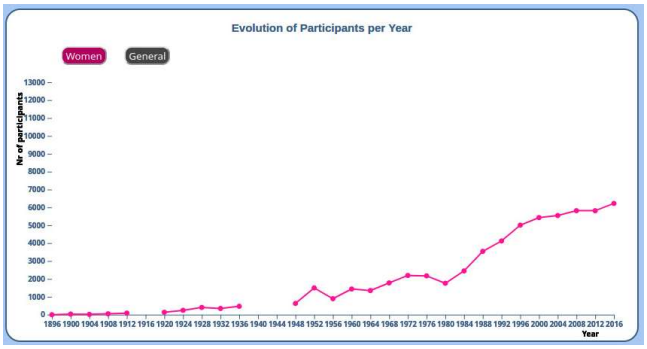

Figure 5 – Overview of the Line Chart in the Women selection

The Cleveland Dot Plot that represents the "Participants and Medalists Difference per NOC" encodes the values of the number of participants, with a grey dot, and the number of medalists, with a lilac dot, per NOC. The x-axis represents the number of participants/medalists, and the y-axis represents the NOCs. In the beginning it is possible to select a country in this idiom, by clicking one of the dots that represent a NOC of that country, when that happens, the other lines and dots that represent NOCs of different countries disappear, showing only the country selected. When there are many countries/NOCs selected, if we click a dot, the dots/line will disappear, and all the NOCs of that country will stop being encoded in the Cleveland Dot Plots and the country also gets deselected in the other idioms. If we hover over the grey dots, a tooltip shows the exact number of participants, and the same happens for the lilac dots, which shows the exact number of medalists. If we hover over either the dots, the dots and the rest of the elements of that NOC will now appear with a black stroke to indicate that we are hovering over it.

The Cleveland Dot Plot that represents the "Medalists Difference per Gender and NOC" encodes the values of the percentage of medalists regarding the total number of participants, divided by gender, having pink dots that represent this percentage for women and blue dots that represent this percentage for men, by NOC. The x-axis represents the percentage referred before, and the y-axis represents the NOCs. In the beginning it is possible to select a country in this idiom, by clicking one of the dots that represent a NOC of that country, when that happens,

the other lines and dots that represent NOCs of different countries disappear, showing only the country selected. When there are many countries/NOCs selected, if we click a dot, that dots/line will disappear, and all the NOCs of that country will stop being encoded in the Cleveland Dot Plots and the country also gets deselected in the other idioms. If we hover over the pink dots, a tooltip shows the exact percentage of women medalists regarding all the women participants for that NOC, and the same happens for the blue dots, exact percentage of men medalists regarding all the men participants for that NOC. If we hover over either the dots, the dots and the rest of the elements of that NOC will now appear with a black stroke to indicate that we are hovering over it.

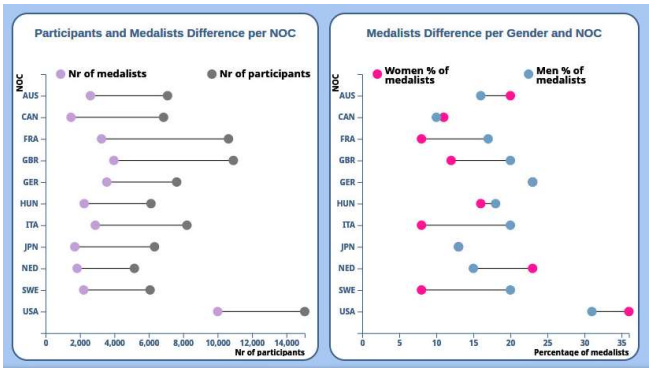

Figure 6 – Overview of the Cleveland Dot Plots

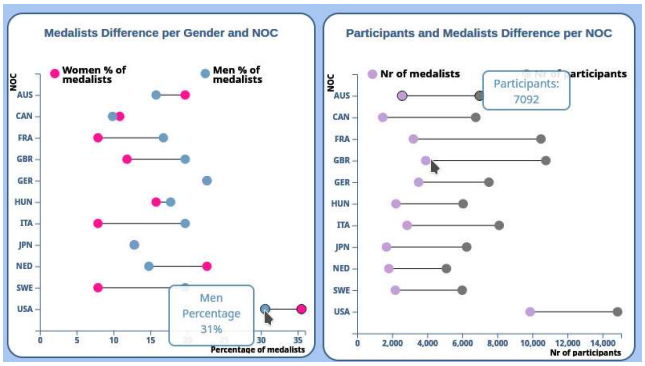

Figure 7 – Tooltips in the Cleveland Dot Plots

The Progress Bars, in the beginning, encode the total of the medalist's percentage regarding the total of participants for each gender. This idiom allows the user to have a different representation of the values encoded in the "Medalists Difference per Gender and NOC" Cleveland Dot Plot. For each NOC, of each selected country, this idiom shows two progress bars, one for women and one for men, encoding the value of the percentage for that NOC. In the middle of each progress bar, there is a text that shows the exact value encoded, and above the progress bars, the NOC name is shown. When clicking a progress bar, the country that NOC represents gets deselected and all the idioms update accordingly. If we hoover over a progress bar, a tooltip with the name of the country is shown.

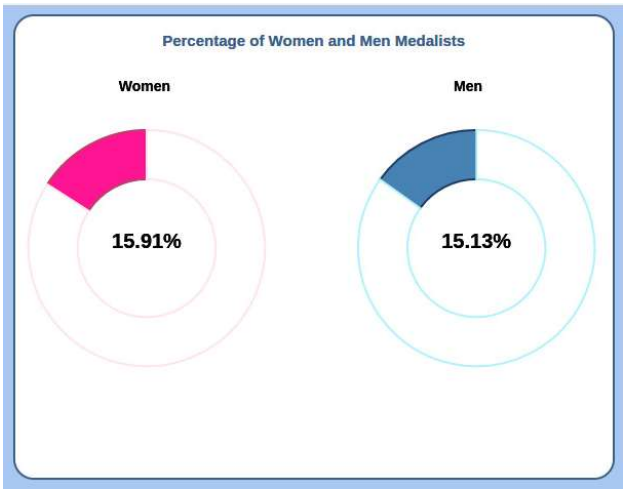

Figure 8 – Overview of the Progress Bars

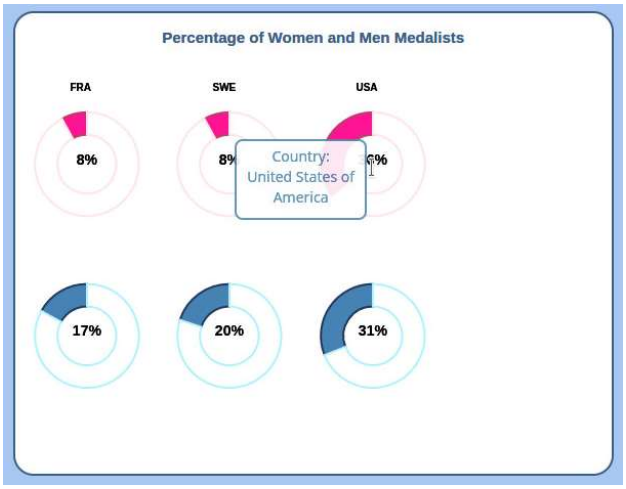

Figure 9 – Tooltip in the progress bars

When all the countries get deselected, the visualization goes back to its original state, which is described at the beginning of this section.

When filtering the data per millennium the countries that are already selected, continue selected when the filter is applied, keeping the visualization consistent.

Hovering over one of the idioms triggers the other idioms to change and highlight the country/NOCs which is being hoovered as if it was being hoovered in that idiom. For example, if we hover a country in Choropleth, the rest of the map appears with less opacity, the line of that country in the Line chart gets thicker and the dots of that country NOC's get a black stroke in the Cleveland Dot Plots.

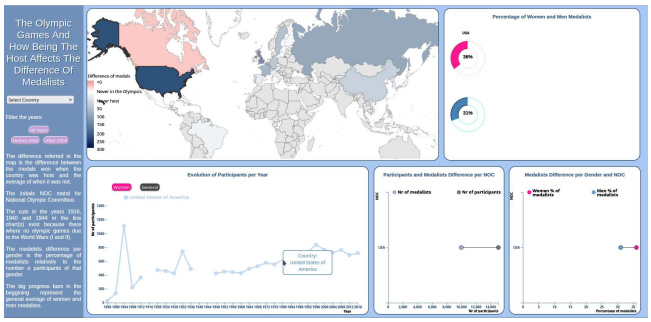

Figure 10 – Interaction with the idioms

There is a maximum of selected NOCs at the same time, which is four, and if the user tries to select more than four NOCs, a window alert pops up, letting the user know that he reached the maximum NOC's or is about to reach it. It's only possible to change the filter regarding the years when the sum of NOCs remains less or equal to 4. For example, if we are in the filter "After 2000" and we select Russia (after 2000 it has only one NOC - RUS), Portugal, and France we have 3 NOCs selected, but if we change to the filter "Before 2000" a window alert appears informing that the sum of NOCs is more than 4. This happens because Russia has 3 NOCs before 2000 (EUN, RUS and USR). To change the filter, we must deselect countries before.

Rationale

To represent the difference between the medals a country won in the year(s) in which it was the host and the average of medals that the same country wins overall, we needed a form of encoding that allowed us to represent very different values; with this specific attribute, three different things could happen: the country may have never participated in the Olympics, the country may have participated but never been host, or the country may have been host. So to encode these values we thought the best way to do it was to use colors, so we chose different hues for each: white, often associated with empty, for the countries who never participated; grey, often associated with neutral, for the countries that were never host; and for the countries that were host sometime, we chose a red hue for the ones with a negative difference of medals, since red is usually associated with "bad" so it made sense to use it for negative values and a scale of blue tones (lighter to darker) for the countries with a positive difference of values, since we decided that blue would be the main color of our visualization, and that this information was the one we were more excited to visualize. Since there were many colors to be represented and many countries to represent, we logical thought that a Choropleth Map would be a good choice, since it is easy and still understandable if we fill each country area with the color that encodes the difference, since the whole country has the same value (there are no different distributions throughout a country), it scales well, since it already contains all of the countries and we can zoom-in and zoom-out and the general population is already fairly acquainted with using maps and geographically know where some countries are.

To represent the evolution of the number of participants in the Olympics, the most intuitive idiom for us to use, was a Line chart, since the slopes on the lines allow the user to easily know if a value is ascending, descending, or constant. We also decided that it would be important to filter to just women, since their participation in the Olympics for many countries started later than men and with a very different quantity of participants. Regarding the colors, the general we wanted to keep it coherent with the visualization, so we used different tones of blue to represent the different countries and grey to represent the totals, since these are the colors, most used in our visualization. For the women lines, we decided to use different tones of pink, since this color is often associated with femininity and women. Another design choice we made, was to draw dots in the position of the years so it would be more intuitive for the user to see the years in which the Olympics happened and to check the exact values.

To represent the "Participants and Medalists Difference per NOC" and the "Medalists Difference per Gender and NOC", since they both encode differences, we decided that we wanted to do the same idiom for both, to keep the visualization consistent. At first, we considered doing a Slope Graph, but we ended up using two Cleveland Dot Plots, since in the Slope graph it may be confusing if many lines are intercepted between them, which is a problem that does not exist in the Cleveland Dot Plot, that also represents differences quite well. Regarding the colors chosen, for the "Participants and Medalists Difference per NOC" Cleveland Dot Plot the choice was purely made to keep the colors consistent with the rest, so we chose grey and lilac that mixed very well with the other blues and greys used. For the "Medalists Difference per Gender and NOC", we chose blue for the men, since it is a color often associated with males and we chose pink for women, since it is often associated with females. In the beginning, we thought about showing all the NOCs in the Cleveland Dot Plots when the visualization has started, but we soon found out that it wouldn't work out since there were many elements overlapped, so since it didn't really scale well, we decided to only show the NOCs of the most relevant countries, in terms of the number of participants, in the start.

We decided to also use progress bars to represent the "Medalists Difference per Gender and NOC", since we thought that would be nice to the user to have a different perception of these values, by using angles instead of a scale. The colors chosen are the same that are used in the Cleveland Dot Plot, since they represent the same values.

Another decision we made was to allow the user to filter the data in some way, since we thought it would be interesting to have different insights on the data, and not just a general view. What made more sense for us, since the Olympics only occur every four years, was to allow a time filter. We

have data from 1896 to 2016, so we chose to divide into two millenniums: before and after 2000, since with this division we still have data enough in each filter to make a valuable comparison.

We also felt the need to make this visualization more accessible for everyone, so that is why we decided to use tooltips in pretty much every important aspect of each visualization, so that it is easy to check the exact values of the variables encoded for anyone. By this exact motive, we also implemented a dropdown button with all the countries, so that people who don't know where a country is on the map, can easily select it anyway.

On top of that we also decided to add a sidebar with the title of our visualization and with some extra information about some derived measures we calculated that may not be obvious and some other curiosities and important aspects about our data and visualization. We chose to do this in the sidebar, because if it was in the middle of the idioms, they would get rather busy and harder to understand.

Overall, we didn't have many scalability issues, or at least issues to which we couldn't find a reasonable solution, but we did have to impose a limit to the selected NOC's, otherwise it would get really hard to understand all of the information, so we limited the selection of countries to 4 NOCs at the same time.

Potential

To answer our 4th and 5th questions, the ones related with the evolution of participants (female + male and only female) we implemented a line chart. In the beginning, the line chart starts with the general evolution.

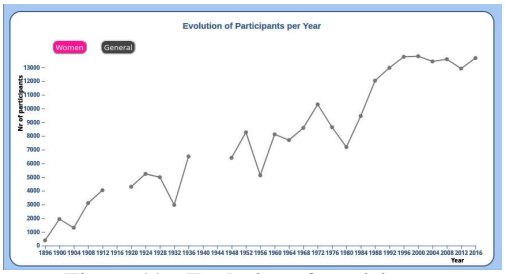

Figure 11 – Evolution of participants

To see the women evolution, the user only has to select the "Women" button.

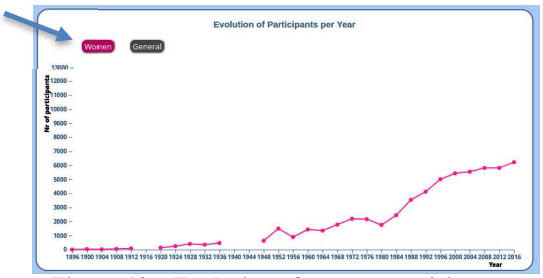

Figure 12 – Evolution of women participants

To see the evolution of participants per country the user can select the countries in the Choropleth Map or in the dropdown button (or in the Cleveland Dot Plots if we are on the overview page). We can also see the answer to the 3rd question "How being the host country of the Olympic Games affects the number of medalists in that country?" with the colors encoded on the map.

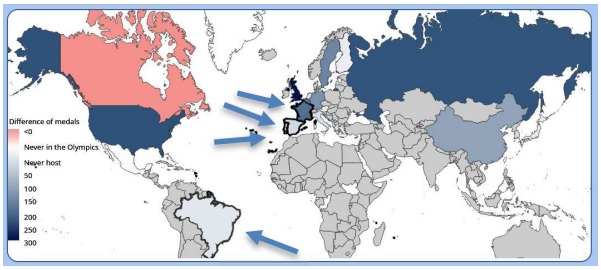

Figure 13 – Selection of countries in the Choropleth

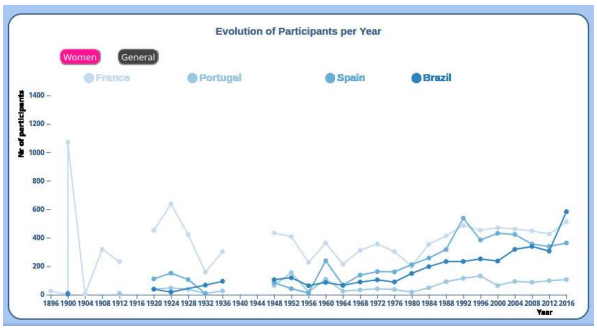

Figure 14 – Line Chart with selected countries in the General selection

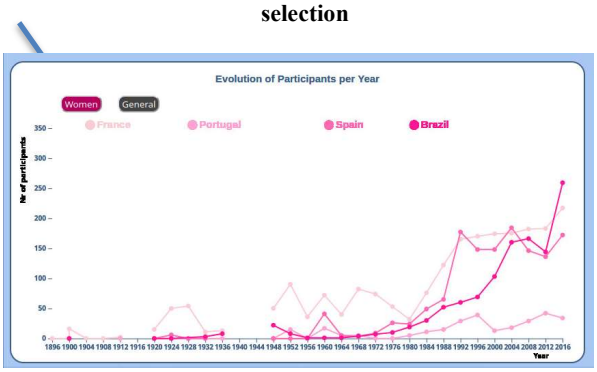

Figure 15 – Line Chart with selected countries in the Women selection

To answer our 1st and 2nd questions "How does the number of medalists in the Olympic Games vary according to the participants, per NOC (National Olympic Committee)?" and "How does the gender of medalists in the Olympic Games vary, per NOC (National Olympic Committee)?" we implemented two Cleveland Dot Plots that easily show the difference between the two variables.

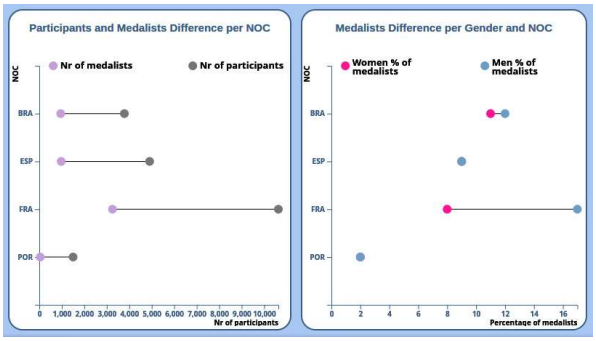

Figure 16 – Cleveland Dot Plots with selected countries

Firstly, we would like to comment on the fact that we didn't know that there were no Olympic Games in 1916, 1940, and 1944 due to World War I and World War II. We found these facts due to missing values in those years and found them really interesting. Although our dataset doesn't include data from Tokyo 2020, we would like to notice that there have been Olympics every 4 years except in the War and in 2020 due to COVID-19.

When we started thinking about the theme for our project and got the idea of the Olympics, we started wondering if a country has more medals just for being the host. Now that we finished our project, we can conclude that we were right and almost every country wins more medals in the year that it was the host (only Canada has fewer medals). This is super interesting because we can see that the countries try harder when they are the host of the Games.

After implementing our Line Chart, we wanted to understand why there are some decreases in the evolution of participants. We searched for it and realized that the Games in 1932 were held during the worldwide Great Depression, with some nations not participating in the Olympics. In 1928, 46 nations participated and in 1932, only 37 nations did.

The 1956 Olympic Games was the first one to be held outside Europe or North America and 8 nations boycotted it for various reasons (Suez Crisis, Soviet Invasion of Hungary, and the presence of the Republic of China).

In 1980, there was the American Boycott, part of a package of actions to protest the December 1979 Soviet invasion of Afghanistan. Great Britain and Australia supported the boycott but allowed the athletes to decide for themselves whether to go to Moscow. The USA athletes were totally forbidden to participate. In the end, 67 nations did not participate.

IMPLEMENTATION

Implementing this project, we came across quite a few challenges, per example, in the Progress Bars it was kind of hard for us to understand how we would make them appear and disappear dynamically in the right positions, and how that connected with the css and html files, but after further investigation we finally understood how to use the divs and svgs necessary to implement them. Another idiom that caused a lot of issues was the line chart, it was really hard to create multiple lines and then remove them, but we then learned about the id attribute of the d3 objects, and after that it was easier to implement what we idealized.

To implement the links between the views when selecting a country, we used a global list, which contains the names of the countries selected. So, when the user selects a country/NOC the name of that country is added to this global list, and right after it is added, all the idioms are updated with this list, so that the country/NOC gets selected in all the idioms. When a country gets deselected, the inverse process occurs: the country is removed from the list and all the idioms get updated so that they no longer encode the values for that country.

To implement the links between the views when hovering over a country, we used the id's that are used in all the idioms, in this project, the id is the NOC. Since we have the id, when the user hovers a country in an idiom, it directly updates the other idioms, inside the mouse over function.

From the idioms we implemented, the one that needed the most adaptation from the examples we found online, was the Line chart, since we needed it to represent many lines at the same time and to switch from general to just women and back; we also had to do some adaptations to the Choropleth Map, since we wanted to represent four different hues, and the ones we found online usually had only one hue (with many tones).

The fact that the Progress bars and the legend in the Line chart change dynamically also made us think further from ourselves, since we didn't really find examples of this online.

CONCLUSION AND FUTURE WORK

By completing this project, since there is no better way to learn than to do it yourself, we learned a lot about the d3 JavaScript library and how to use it to create the visualizations we idealized, although sometimes frustrating, we ended up with a result that satisfies us. We also learned about how important the data is to make a good visualization, how to properly clean it and filter the important information from the datasets we found online. The topic we chose was actually great because there is a lot of information about it online, since many people are interested in it, allowing us to have enough data to create our visualization and much more regarding the Olympics, which made it fairly easy for us to answer the questions we proposed in the beginning, and maybe some others as well. During this project, we had to compromise and manage our expectations for what were the capabilities and tools of d3 and for the extent of the things we wanted to do initially, we didn't have time for everything. These little things we learnt during the project, turned out to be very valuable and the more we worked on the project, the more we learned, and it will for sure have a good impact on future projects. If we had one more month and a large budget to spend we

would do a much more complex dashboard with a lot more information that was able to answer a considerably larger amount of questions regarding past Olympic Games, such as questions about the performance of the countries in specific sports, we would probably also include information about the Winter games, etc... With the amount of data, there is about this theme, there are numerous paths to explore and with a whole month and a budget to do it, there is an unlimited number of ideas for visualizations and questions to answer.