Analysis of 120 Years of Olympic Athletes and Results

Yiwen Jiang | Tianlin Lan | Lu Xu | Yu Yuan

Abstract— This final report for visual analytics system is to use dataset of 120 years of Olympic Games to analyze the history of it, such as the medals each countries won over time and the range and majority age, weight, height of the athletes. The dataset of the athletes and results from 120 years of the Olympic Games are available at Kaggle. There are two visualization charts. One is Medal Board for medal numbers based on the geolocation or the rank of each country. The other is Athlete Board, for the analysis of athletes' age, weight, and height. Users can type their age, weight, and height into our interaction interface to find a match. This final report mainly describes our team's goals, our intended visualization design, and our end result (the prototype). Some fun findings will also be mentioned in our report.

Index Terms— Olympic Games, history, sports, athletes visual analysis, statistical analysis

1 Motivation

The modern Olympic Games are leading international sporting events featuring summer and winter sports competitions, in which thousands of athletes from around the world participating in a variety of competitions. The Olympic Games are considered the world's foremost sports competition with more than 200 nations participating.

With a history of more than one hundred years and a large quantity of participants and results, the Olympic Games are a good example to analyze. It is a lense through which to understand global history and sport history, including shifting geopolitical power dynamics and women's empowerment, the evolving values of society, as well as the improvement of sport skills.

This project will analyse the data of Olympic Games, including the information of athletes participated and the results of each sports.

1.1 Target users

Our proposed visual analytics system is helping users who are interested in the Olympic Games and want to know more about the history, trend of the sports and participants of the Olympic Games. For this purpose, we derived two possible personas who would like to gather some information using our software. Firstly, sport brands operators could find out countries and sport events which are worth to sponsor and market. Secondly, Olympic fans interested in digging out the attributes of athletes, like age, height and weight, could use our software to find out the distribution of those attributes of athletes. They could also use the interface to interact with the software to find out the best choice of the sports that they can develop.

1.2 Questions to answer

There are so many questions that people could ask about Olympic Games. To narrow it down and make our proposed visual analytics system specific, we want to answer the following questions:

1. Questions related with the athletes: the trend of the total number of the athletes participated in the Olympic Games, and the ratio of male and female athletes over time.
2. Questions related with the nations participated in the Olympic Games: the total number of countries that participated in the Olympic Games over time, the metals each country won, and total number of athletes from each country over time.
3. Questions related with the host countries: the geolocation of the countries host the Olympic Games.

While considering the time and scope of our project, we decided to focus on two aspects of visualizations: the medal information and athlete information..

2 Data

For our proposed visual analytics system, We will need the following data:

1. The information about the athletes participated, such as their name, gender, age, height, weight, and nation.
2. The data related with the event of Olympic Games, such as the location where the event was held, the year and season of the event, and the sport name of the event.
3. The data associates the athletes and the events, such as in which year, which event the athlete participated, and what medal the athlete won.

To keep consistent with the scope of our project, we will adjust the original dataset correspondingly.

2.1 Available dataset

There is existing dataset in Kaggle of the results of Olympic Games and the athletes participated. The dataset can be downloaded as CSV file. This historical data set on modern Olympic Games includes all the Games from Athens 1896 to Rio 2016. The data file contains 271116 rows and 15 columns. Each row corresponds to an individual athlete competing in an individual Olympic event (athlete-events). The columns are:

ID - Unique number for each athlete
Name - Athlete's name
Sex – M (Male) or F (Female)
Age – Integer
Height - In centimeters
Weight - In kilograms
Team - Team name
NOC - National Olympic Committee 3-letter code
Games - Year and season
Year - Integer
Season - Summer or Winter
City - Host city for Olympic Games
Sport – Sport name
Event - Event
Medal - Gold, Silver, Bronze, or NA

2.2 Concerns for the dataset

There are more than 0.2 million rows in the historical dataset, which requires filtering and data analysis to make it easier to visualize. We used R to clean and wrangle the data.

We also raised concerns of our project during the discussion, mainly about the data selection and cleaning step.

1. Summer Olympics and winter Olympics have completely different events. If visualize both, the number of sports will be doubled. Compared with Winter Olympics, Summer Olympics have more data, so we decided only visualize data for Summer Olympics.
2. We need to decide the countermeasures against the special cases of Olympics such as change of country names caused by breaking event like the collapse of the Soviet Union.
3. There are sports that held during the entire Olympic history, some sports that only held several years of Olympic Games, and some sports that start from recent years. Both two charts, the Medal Board and the Athlete Board, need to consider this problem. We decided to use all sports to calculate the medals for each country, but use some of the sports to analyze the athletes.
4. In our dataset, each row refers to an instance of an athlete and the medal information of this athlete. If the a sport is finished by team not individual, we will count the medal once, while in our dataset, the medals will appear multiple times. We need to manipulate the data to calculate the correct medal numbers.
5. There are many null values of weight, height and age of the athletes, we need to clean them before we calculate the average of them.

2.3 Data wrangle and clean

Correspond to the concerns listed above, we used Rstudio to solve them in the following ways:

1. Remove winter olympics data.
2. Remove the rows with N/A data in weight, height or age column. Considering the huge size of the dataset, the removed data will not affect the result too much.
3. As for the discontinued sports, we will use the principles:
    a. remove data of sports that are held less than 10 times in the whole Olympic history and meanwhile not held in recent 5 Olympics.
    b. remove data of sports that has less than 20 participant countries.
4. We used “tidyverse” package in R to do the data cleaning. Functions mainly used are group_by(), filter(), count(), drop_na() and inner_join().

After these, the dataset of this project was downsized from 271,116 to 164,913 rows with formatted structure for the processing and analysis of next stage. Detailed process of data cleaning can be found in the appendix B of in our Status Report.

During the further process of our visualization project, to facilitate coding of visual graphs, we will always use the Rstudio to generate formatted data frames (contain required columns) according to the visualization needs in advance. However, the calculation like mean, quantile of weight or height will still be made by Javascript.

3 Intended Visualization Design and Corresponding Tasks

Based on the two types of personas mentioned in 1.1 target users, sport brands operators and Olympic fans, we designed two charts for our proposed visual analytics system: Medal Board and Athlete Board.

3.1 Graph 1: Medal Board

The Medal Board is designed to display the change and the distribution of amount of medals won by different countries in previous Olympic Games.

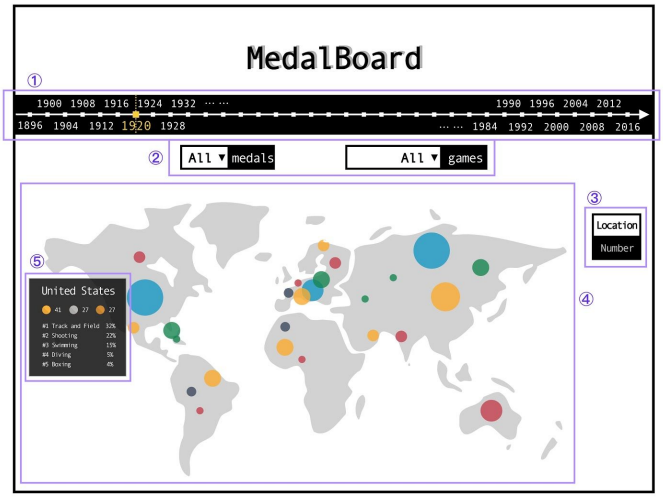

Figure. 1 Medal Board UI Sketch

As shown in the sketch, there are four main parts of the graph:

(1) the timeline bar;

The draggable and clickable timeline bar contains dots which refer to the year of the past Olympic Games.

(2) the filter menus;

The filter menus would let users to customize the display. One contains a list of specific games, which could let users to add filter to the graph if they would like to see the number of medals of a specific games. The other one which contains options for three kinds of medals: gold, silver and bronze, is designed for users who have a interest on a specific kind of medal.

(3) the tabs;

Two tabs are provided, which related to two themes of the display panel which will decide where the circles would locate: one based on geographical location and one based on the ranking of numbers.

(4) the display panel;

In the display panel, one country would have one correspond circle. We use the size of the circles to present the amount of medals won by each country. The label of the country would lie on the circle.

(5) the tooltip.

The detailed information about medals of that correspond country will be shown on a tooltip near the circle, including the numbers of each kind of medals(gold, silver and bronze) and top 5 sports of that country.

Task 1: Users want to view the amount change of medals won by countries.

This task could basically involved the following (sample) questions emerged from users:

(1) For one Olympic Game(one year), what are the differences of the numbers of medals between different countries?

(2) Through the past Olympic Games(dragging the timeline bar), what is the trend of changes of one specific country about the number of medals?

(3) Which country won the most/least medals in one Olympic Games?
(4) What are the sortings of the number of medals won by each country for each Olympic Games?

To find out the answers, users can change the year of Olympic Games by dragging the cursor on the timeline bar or by clicking on the year label directly, the size of the circles will change correspondingly. Users could also change the view by choosing "Location" tab or "Numbers" tab. When "Location" is selected, the circles would locate based on the geographic location, which means each country's circle would lie on the correspond location on the background world map. When "Number" is selected, the circles would be sorted on their size from largest to smallest. Thus, users could observe the change, distribution and sorting of the amount of medals won by each country directly.

Task 2: Users want to view the trend of medal numbers of a specific game or a specific kind of medal(gold, silver, bronze).
This task could basically involve the following (sample) questions emerged from users:

(1) For one game category in one Olympic Game, which country won the most/least medals?
(2) Through the past Olympic Games, is there a change in the country that won the most/least medals?
(3) Which country won the most gold medals?

In addition to view the total number of medals won by each country in each Olympic Game, users could add filters to the graph in order to see the numbers of a specific category: a specific game(balls, swim, etc) or a specific kind of medals(gold, silver, bronze). Users could also apply the filter into the sorting view to see the rankings..

Task 3: Users want to view the detailed information about medals won by one specific country.
This task could basically involve the following (sample) questions emerged from users:

(1) For one specific country, what is the exact number of gold/silver/bronze medals that country won in that year of Olympic Game?
(2) For one specific country, what are top 5 sports that this country is most good at and how many medals the country won of those top 5 sports?

Users can view the detailed information about the medals one country won in one Olympic Game and their strengths, they can hover their mouses on the circle of that country and then the tooltip would show up near the circle which contains such information.

3.2 Graph 2: Athlete Board

The athlete board is designed to let users find out the age, height and weight of athletes participated in the past Olympic Games and explore their possible sports to develop based on their own data by interacting with the graph.

As shown in the sketch, the graph contains four main parts:
(1) the gender toggle buttons;
Three toggle buttons are provided for users to filter data they want to observe: all, female and male.
(2) the customized input boxes;
The most attractive part of this graph is to let Olympic fans find out their possible games to participate using their own data. The three customized input boxes would gather users input data(their own age/height/weight) and then the backend code will transfer the data into the visualization code to show related data in the display panel.
(3) the display panel;
All the data about athletes' age, height and weight would be shown in the three coordinate graphs of the display panel in boxplot format. The x-axis of each coordinate refers to intervals of age/height/weight. The three coordinate graphs share the same y-axis, which refers to the games. Those labels of the games are sorted by the numbers of participated athletes.

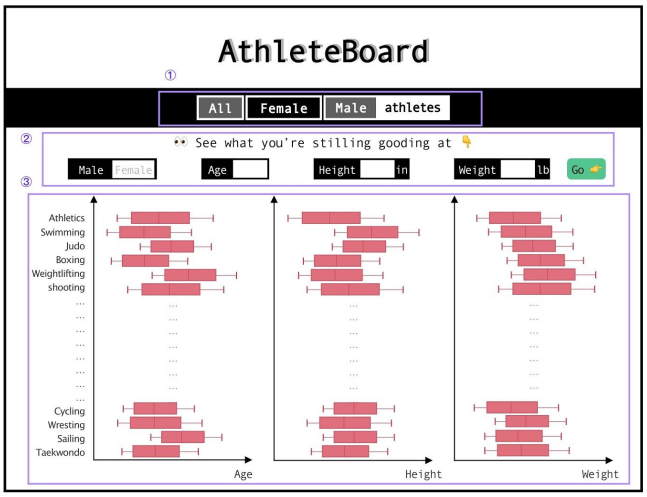

Figure. 2 Athlete Board UI Sketch

(4) the tooltip for boxplot.
The detailed information about the boxplot of a specific game would be shown on the tooltip when users hover their mouses on the boxplot, which includes the median value, maximum, minimum and the scope between the first quarter to the third quarter of athlete's age/height/weight.

Task 1: Users would like to find out more information about athlete's vital statistics
This task could basically involved the following (sample) questions emerged from users:

(1) Which sports skew young and which allow for more longevity?
(2) Which sports prefer tall people and which would be benefited from being shorter?
(3) Which sports require athletes to restrict their weight and which show more tolerance?
(4) Is there any gender difference in the age tolerance of specific sports?

User could view the boxplot describing specific game to find out their answers, in which it provides the median value, maximum, minimum and the scope between the first quarter to the third quarter of age/height/weight.. Instead of providing bar chart, boxplot would help users to figure out the outliers as well as the variation between this. By hovering it, user would see the specific value of each boxplot to know a detailed age scope of specific games. By selecting the toggle button of gender choices, users can see those statistical information only about female/male athletes.

Task 2: Users may be curious about their potentiality on Olympic Games

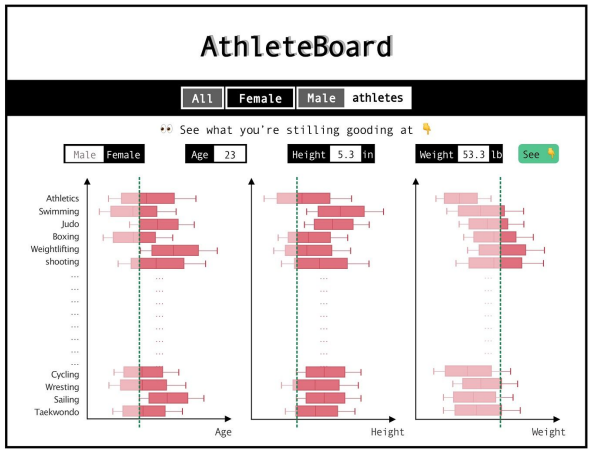

Figure. 3 Athlete Board UI Sketch

The feature corresponding to this task is initially designed for fun, with the aim of increase the user engagement with our visual design, we set this fun piece to allow users input their vital statistics to compare with those great athletes. By inputting their gender, age, height and weight information, list of games the events that they currently still have hope for. Besides, we also believe that this chart will inspire the youth to discover their interested sports and give them a big picture of the possible career length of specific sports.

4 Final Implementation

During our implementation, we applied all the features we intended to. While more detailed problems arose, we made several adjustments and improved our initial design.

4.1 Headers and introductions

To make users have a better understanding of our visualizations, we created a header for Medal Board and Athlete Board to introduce our goals and the features.

Figure. 4, 5 Introduction Screenshot

4.2 Graph 1: Medal Board

As shown in our screenshot of Medal Board, the timeline bar, the filter menus, the tabs, the display panels and the tooltip we intended are all in available in our final visualization chart.

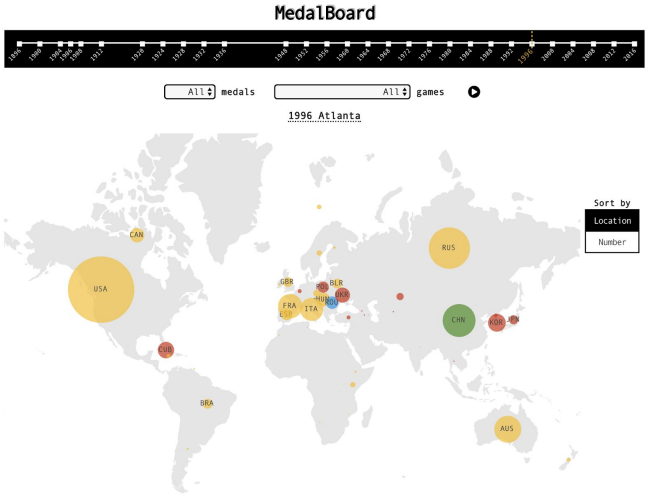

Figure. 6 Medal Board Screenshot

4.2.1 Improvement of the Dataset

Instead of using all the medals of sports, we filtered the dataset and only the selected sports left. We calculated the medal numbers of each country by these sports.

4.2.2 Improvement of the Color

Originally, the colors of the Olympic rings represent the five continents. But we rephrased the meaning of the colors as five categories for sports and divided the sports into one of them.
After the class presentation, we decided that the color of each country should be normalized, otherwise if the color category has less sports and medals, countries has a very low chance to be categorized into it. The color of countries are based on the percentage of each category instead of the absolute medal numbers.

4.2.3 Bubble hover and click

For each circle, a tooltip will show up when hover on as we designed. Besides, The tooltip contains the detail information of each country or region in this year olympics. This is to provide users with a very brief summary and feeling of a country or region, such as the number of different kinds of medals, gold, silver or bronze. Then we will list the top 5 sports which have the largest number of medals.

We also added the click function to allow users have more detailed view of the medal numbers of the country. When the bubble of a country is clicked, below the bubble chart, a list of medal numbers of each sport will be displayed. The list in the right of medal list is the category of sports that each color represents.

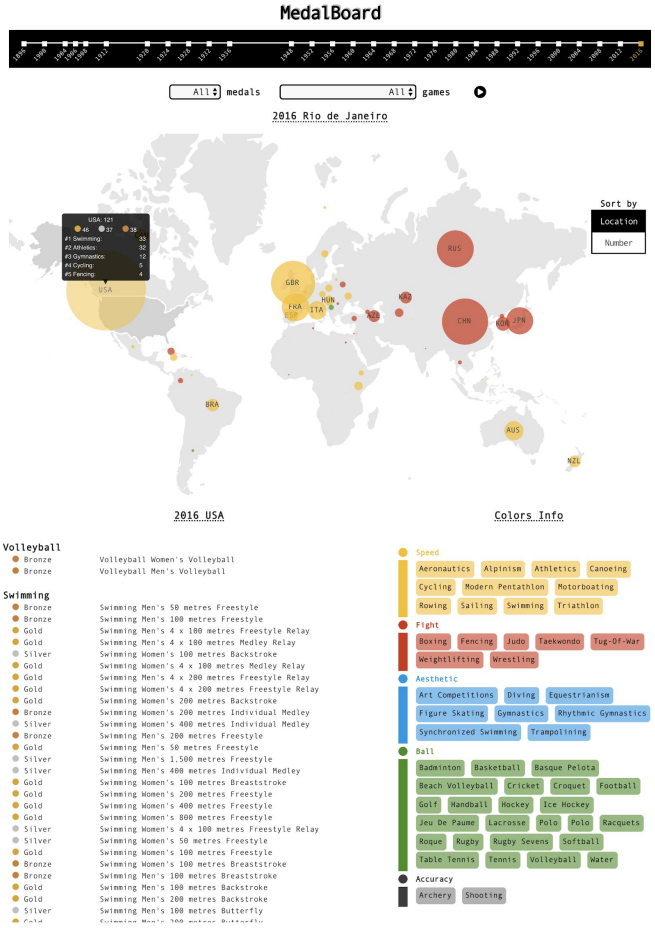

Figure. 7, 8 Medal Board Hover and Click Screenshot

4.2.4 Play Function

If the user clicks the play button, the bubble chart will automatically display the bubbles of each Olympic Games in the history. This play function shows how the medals of each country change over time.

4.3 Graph 2: Athlete Board

For Athlete Board, we implemented the gender toggle buttons, the customized input boxes, and the display panel. The dataset we used for Athlete Board are The same with Medal Board, instead of using all sports, we filtered the dataset and only the selected sports are displayed. We calculated the medal numbers of each country by these categories.

4.3.1 Improvement of the Display

1. Instead of using three y-axes in our mockup figure, we used one y-axes to represent the sports and aline the boxplots in the same y-axes.

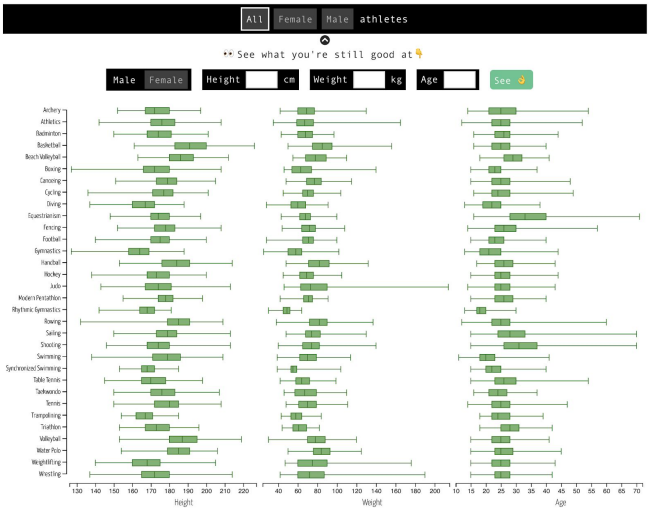

Figure. 9 Athlete Board Screenshot

2. We highlight each row when hovered over for three boxplots of the same sport, so it is easier for user to see which sport is corresponding to this boxplot, since there are over 30 lines in the graph.

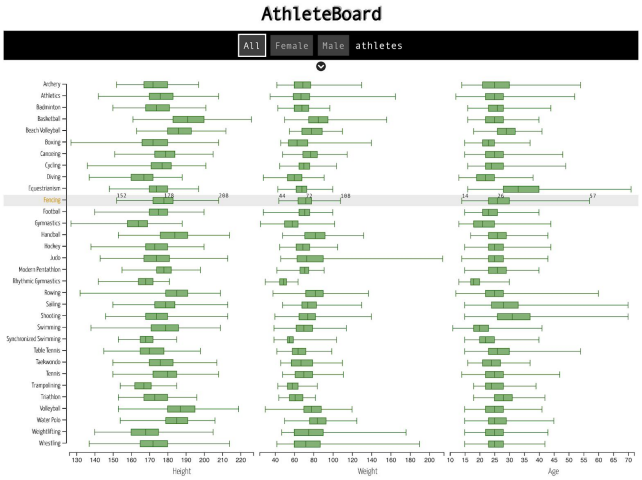

Figure. 10 Athlete Board Highlighted Line Screenshot

3. What is more, when user hover over the boxplot, the numbers of median, the first and third quantiles will be displayed.
4. To better distinguish the box plots for different gender, we used three different colors to fill the boxplots. When the user click ‘All’/ ‘Female’ / ‘Male’, the boxplots will be rendered in the corresponding color.

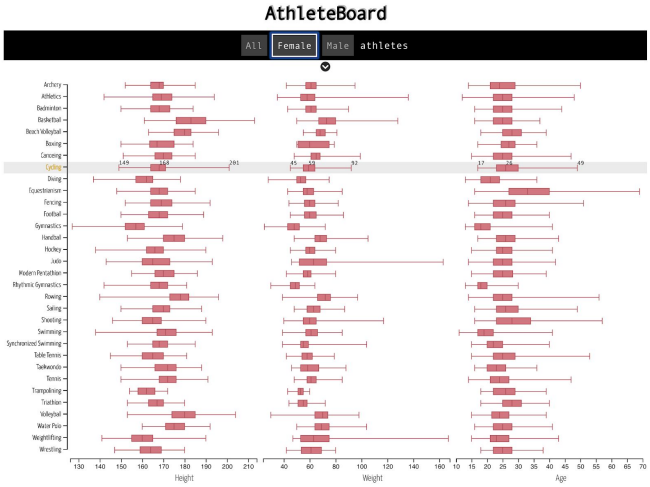

Figure. 11 Athlete Board Female Athlete Screenshot

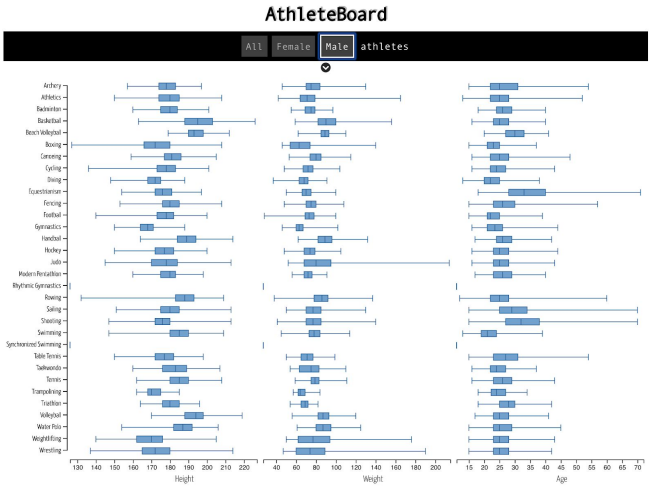

Figure. 12 Athlete Board Male Athlete Screenshot

4.3.2 Sports Recommandation

In our original design, after user's input of gender, age, weight, and height, lines will show up in Height, Weight, and Age boxplots to let users know how their vitals are when compared to the Olympic Athletes.

We highlighted the boxplot if the user's input is between its first and third quantile. Meantime, the sports meet the range of Height, Weight, and Age boxplots are selected and recommended to the user.

Figure. 13 Athlete Board Recommandation Screenshot

5 Fun Findings

After our implementation and exploration of the visualizations, we found some fun facts.

5.1 The Olympic hosting countries tend to won more medals than other years

Below are examples of the countries that won more medals than it usually to be in the Olympic Games.

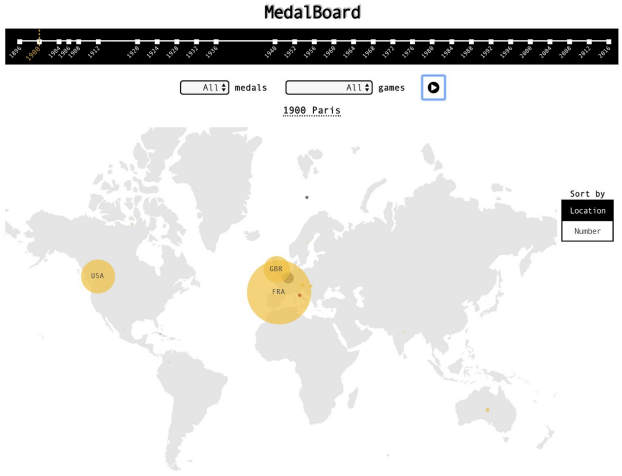

Figure. 14 1900, Paris

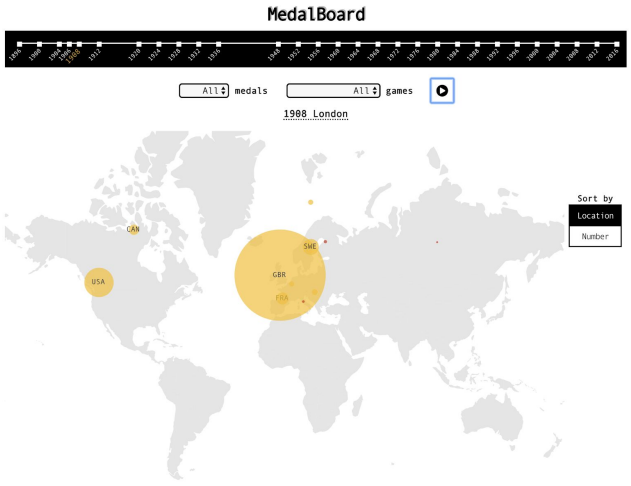

Figure. 15 1908, London

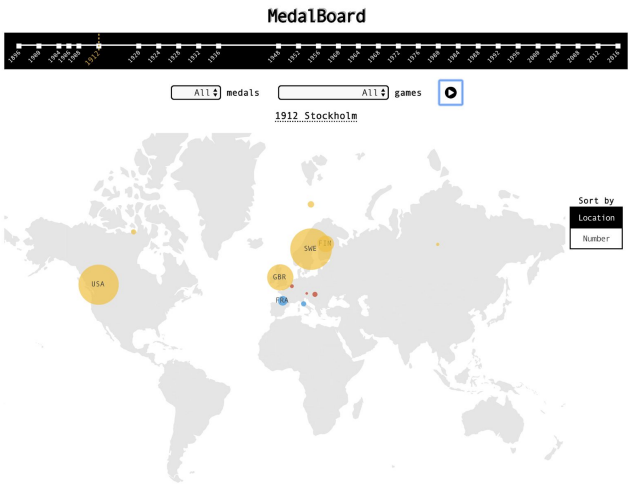

Figure. 16 1912, Stockholm

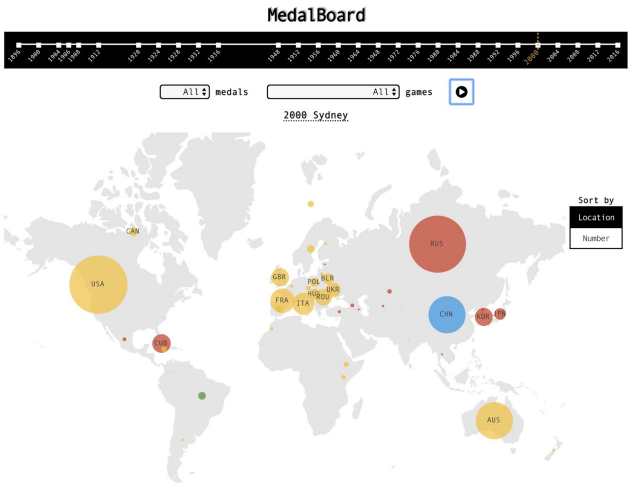

Figure. 17 2000, Sydney

5.2 Unusual Observations

We found in some Olympics, the countries participated are much less. Two example are in 1904, St. Louis, and in 1980, Moskva. We searched the internet and then found that

1. The 1904 Summer Olympics was the first time that the Olympic Games were held outside Europe. Tensions caused by the Russo–Japanese War and the difficulty of getting to St. Louis may have contributed to the fact that very few top ranked athletes from outside the US and Canada took part in these Games. Only 62 of the 651 athletes who competed came from outside North America, and only 12–15 nations were represented in all.
2. During The 1980 Summer Olympics, 80 nations were represented at the Moscow Games – the smallest number since 1956. Led by the United States at the insistence of U.S. President Jimmy Carter, 66 countries boycotted the games entirely because of the Soviet invasion of Afghanistan.

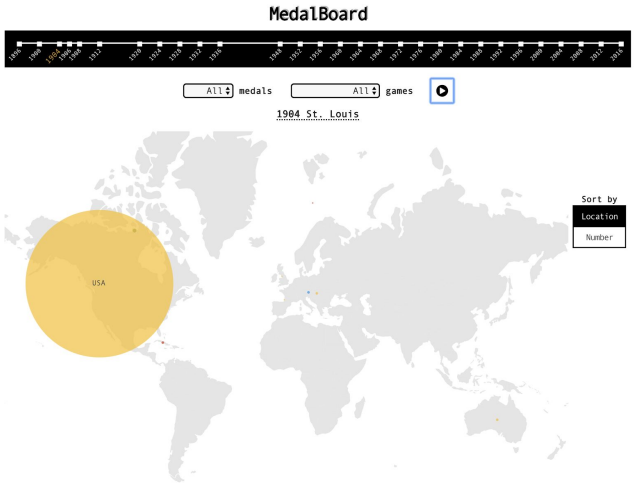

Figure. 18 1904, St. Louis

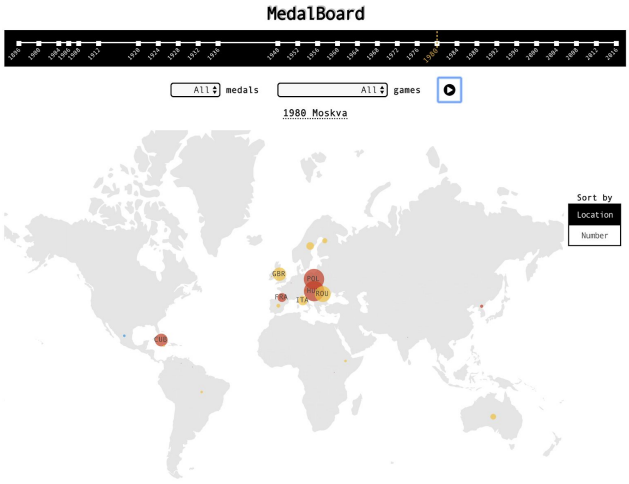

Figure. 19 1980, Moskva