INFO 5100

Project 1 - Written Description

Junrui Deng (jd952), Xu Jing (xj88), Weiyi Hou (wh436)

Olympic Trivia

Part 1.

Introduction

After the data searching section, we chose the dataset of *120 years of Olympic history: athletes and results* [1] from the Kaggle website as the foundation of our data visualization. This dataset contains 15 types of data including personal information of the athletes (name, age, sex, height, weight, etc.) and the information of the Olympic game (years, seasons, cities, sports, events, medals, etc.). The abundant dataset provided us with various opportunities to create data visualization based on different combinations of the datasets. We came up with five different types of visualization during the brainstorming sketch section. With the helpful feedback from TA during the office hour, we finally chose two data visualization plans based on the variables we had in the dataset.

The dataset for the first visualization

The first visualization is to show the comparison of the number of participants of a country for the earliest and the latest Summer Olympic games that the country participated. From the "World.geojson", we copied the data to the *"Country"* variable and extract the *"coordinates"* and *"id"* features to construct the map. We calculated the center points of each country based on the coordinate data from the *"Country"* and stored the results in the variable *"list"*.

Later, we stored the data from the "athlete_events.csv" to the variable AthletesData. The first participation year and the last participation year of each country were filtered base on the *"Year"* features from the variable *"AthletesData" and* were stored in the new variables *"Min"* and *"Max".* After that, athlete amounts of the first participation year and the last participation year were calculated and stored in the variables *"Mincount"* and *"Maxcount"*. Recalling the *"Mincount"* and *"Maxcount"* by using the country *"id"s*, we were able to map the circle radius with those variables.

Importantly, based on the variables in the dataset, we could only display the data visualization for those countries which the three-letter abbreviation country code for Olympic Games (IOC) matches with the three-letter abbreviation country code for the global map (ISO). We got the data of IOC and ISO from Wikipedia [4], downloading and saving the dataset as Trans.csv file for visualization. The file of World.geojson and popular.geojson came from the D3v5 map example [5]. In order to show the comparison circles clearly, we self-calculated the center of the circle and adjusted the color to form an obvious contrast.

The dataset for the second visualization

The second visualization is to analyze if there is a difference in medal awards of a country when that country is or is not holding a Summer Olympics. During the programming section, we found that we only needed the variables of the host city of Olympic, medals of the host country for each Olympic game, the total number of each Olympic game, and the datasets we were using was too large for data processing. Therefore, we started to search datasets again to see if there is any dataset more targeted on our project with less additional data. We found two useful datasets on the Kaggle website [2] [3] related to our project so we combined two new datasets to form a new dataset document, summer.csv for our data visualization.
After parsing the .csv file, we went on to process the data and calculate the two kinds of numbers we need by using the following formulas:

$$\text{hpr} = \frac{\Sigma \frac{number\ of\ the\ medals\ won\ by\ the\ host\ country}{Total\ number\ of\ medals\ the\ Olympic\ awarded}}{Total\ Olympic\ Games\ the\ country\ participated}$$

$$\text{nonhpr} = \frac{\Sigma \frac{number\ of\ the\ medals\ won\ by\ the\ host\ country}{Total\ number\ of\ medals\ the\ Olympic\ awarded}}{Total\ Olympic\ Games\ the\ country\ participated}$$

For the equation above, hpr is the average winning medal percentage when the country is the host and nonhpr is the average winning medal percentage when the country is not hosting Olympic Games.

Part 2.

Design rationale for the first visualization

Since we want to show the comparison of the number of participants of a country for the earliest and the latest Olympic games that the country participated, we start the visualization based on the world map. We use two circles to represent the number of participants of a country for the first Olympic and the number of participants of that country for the latest Olympic. We made the center of the two circles for each country at the same position and used the area of the circles to show the contrast between participant numbers. We chose Tiffany blue as the background color of our website to provide a novel and energetic impression in general. In order to reveal the data visualization clearly and neatly, we chose purple and orange, which are the contrast color and secondary color for the circles and sector bars. We wish the lively color combination could bring a thriving feeling, corresponding to the Olympic Spirit.

Design rationale for the second visualization

Inspired by *Grouped Radial Bar Chart* [6], we displayed the proportion comparison using sector with the different colors on the big circle. We put circles with dash lines corresponding to the winning medal percentage to make it easier for the viewers to see the exact proportion for each sector. We put the country code for each sector along the perimeter of the big circle, trying to follow the mind flow of the viewer as they are relating the color bars with the country code.

Instead of showing the comparison of the two rates directly, we chose to show the ratio between the two for each country since the difference between some countries' winning rate is too huge.

We chose to leave a vacuum hole in the center because in this way country with small proportions can be viewed clearly instead of being clustered at the center of the circle. We put annotations on the top right corner as explanations for the color sector.

Part 3.

The story for the first visualization

The first visualization is based on a worldwide map. We compared the number of participants in a country for the earliest and the latest Olympic games that the country participated. With a general glance, it is easy to spot some big circles on the map, introducing the countries which send a large number of athletes to the Olympic games, including the United States, Brazil, Russia, China, Korea, the United Kingdom, France, etc. This is a clear trend that initially the number of athletes participating in the Olympic Games is small but for some countries, the numbers considerably. From the data visualization, we can see that for most of the countries in Africa, West Asia, and South Asia, the participation number of athletes are relatively small and do not change much during these years. In comparison to this situation, we can see an obvious cluster in Europe. Although the area of the countries in Europe is relatively small on the world map, the participation number of athletes are much larger.

The story for the second visualization

The second visualization is to analyze if there is a difference in medal awards of a country if that country is or is not holding an Olympic game. The light orange sectors, which stands for the winning medal percentage of the country when it is the host of the Olympic Game, are very impressive in comparison to the light purple sectors when the country is not the host. Generally, the countries tend to win much more medals when they are the host country. However, Canada is the exception. We guess that maybe Canada performs better in the Winter Olympic rather than the Summer Olympic. Greece, with the country code of GRE, is another eye-catching sector in the chart. Since it held the first Summer Olympics in 1896, the percentage of the winning medal is so large that it even is not influenced by the percentage in 2004.

Fun facts about the Olympics

The following are some findings from the visualization and some fun facts we discovered during the data analyzing process.

- Since Athens holds the first Summer Olympic game in 1896, it won the most medals overall, 46.
- The United States held the Summer Olympics four times (in 1904, 1932, 1984 and 1996). The United Kingdom held the Summer Olympics three times (in 1908, 1948, and 2012).
- Greece (1896, 2004), France (1900, 1924), Germany (1936, 1972), and Australia (1956, 2000) held the Summer Olympics two times.

Part 4.

Task Distribution:

- Junrui Deng (jd952)

  Mainly responsible for programming, project management, and communication, including set up meeting schedules for group meetings and getting feedback from TA. Write and check the progress report and documents.

- Xu Jing (xj88)

  Mainly responsible for programming, including data processing and data visualization design. Write and check the progress report and documents. Team leader on programming section.

- Weiyi Hou (wh436)

  Mainly responsible for writing a PDF file containing the written description of the project and programming. Write and wrap up the progress report and documents for final submission.

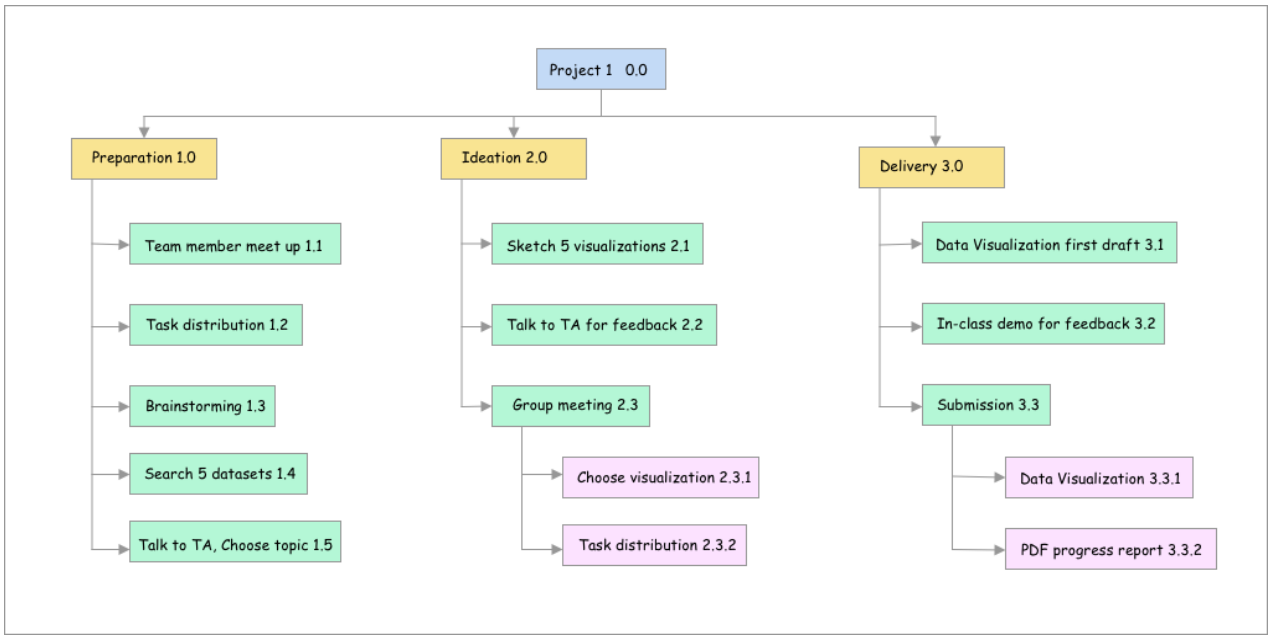
Hours for each task:

0.0 Project

- 1.0 Preparation (2.15 - 2.20)
  - 1.1 Team member meet up (2 hours)
  - 1.2 Task distribution (20 mins)
  - 1.3 Brainstorming (2 days)
  - 1.4 Search 5 datasets (2 days)
  - 1.5 Talk to TA, choose topic (30 mins)
- 2.0 Ideation (2.20 - 2.27)
  - 2.1 Sketch 5 visualizations (1 day)
  - 2.2 Talk to TA for feedback (20 min)
  - 2.3 Group meeting (2 hours)
    - 2.3.1 Choose visualization
    - 2.3.2 Task distribution
- 3.0 Delivery (2.27 - 3.5)
  - 3.1 Data visualization draft (2.28 - 3.4)
  - 3.2 In-class demo for feedback (15 min)
  - 3.3 Submission (3.5)
    - 3.3.1 PDF progress report
    - 3.3.2 Data Visualization

Creating data visualization is the most time-consuming task. The brainstorming process and ideation process are also time-demanding and are essential steps towards final deliverable.

Task Breakdown:



Citation

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

[2] 2016 Olympics in Rio de Janeiro
https://www.kaggle.com/rio2016/olympic-games#athletes.csv

[3] Olympic Sports and Medals, 1896-2014
https://www.kaggle.com/the-guardian/olympic-games

[4] Comparison of IOC, FIFA, AND ISO country codes
https://simple.wikipedia.org/wiki/Comparison_of_IOC,_FIFA,_and_ISO_3166_country_codes

[5] D3v5 map example
http://bl.ocks.org/almccon/1bcde7452450c153d8a0684085f249fd

[6] Grouped Radial Bar Chart
https://bl.ocks.org/bricedev/0a9bf537a64a55ab1fe8

[7] Olympic Logo
https://www.kisspng.com/png-summer-olympic-games-questagame-olympic-symbols-20-2587219/preview.html