DATA VISUALIZATION

Report for Milestone 1

Group: Bugger
Liu Haowen, Sun Haoxin, Wu Yujie

Spring semester 2022

1 Dataset

1.1 Dataset Introduction

We decide to explore the dataset *Olympic Games, 1986-2021* [1] combined with the dataset *Beijing 2022 Olympic Winter Games* [2] to present the data of excellent athletes in all Winter Olympics since 1924 visually. Both of them are derived from *Kaggle*.

The first dataset contains the data of Olympics from Athens 1896 Olympics to the latest Tokyo 2020 Summer Olympics. Containing more than $150,000$ pieces of data, it consists of four following tables.

- **Athletes** contains detailed information about all athletes in each Olympics like their names and committees.
- **Game hosts** contains the host time and city of every Olympics.
- **Medals** lists the information of all medals in each Olympics.
- **Results** contains the details of event results in each Olympics like rankings.

To keep data up to date, we complement the above dataset with the second one, which contains the data about medals and athletes in Beijing 2022. It consists of ten tables, but only three tables — **Athletes**, **Events**, **Medals** — matter in our project.

1.2 Data Preprocessing

We first merge useful parts of two datasets. As we take medalists as excellent athletes, we focus on the data of medals, remove the data of athletes who got no medal, and use other tables to complement the information of medalists. During our preliminary pre-processing, we encountered several crucial problems listed as follows.

- **Winter and Summer.** The first dataset contains the data of both Summer and Winter Games, so we only reserved data for the Winter Olympics.
- **Country change.** From 1924 to 2022, athletes from the same area may have different committees, which makes geographical visualization difficult. So we use capitals to stand for countries to simplify visualization.
- **Name Format.** The name format of athletes is not unified, i.e., the first and last names are in different orders. We first figured out their first and last names separately, then unified the format.

2 Problematic

2.1 Motivation

The XXIV Olympic Winter Games in Beijing closed early this year, in which lots of brilliant athletes contributed their wonderful performances to the Game. However, not many people know details about them. Besides, elite athletes in past Games easily fade from people's memory. There are few websites to sort out those pieces of information and display them intuitively. So we decide to design a webpage to record the information and wonderful moments of elite athletes in all Olympic Winter Games visually. In this way, people can have a better understanding of this global event.

2.2 Overview

To do this, we plan to use a world map as the primary dimension of the visualization, embellished with information about medalists of all Olympic Winter Games. These pieces of information are in the form of dots spread on the motherlands of athletes, which creates the second dimension. When users interact with those dots, detailed information about corresponding athletes will be displayed as the third dimension. Besides, the intensity of dots reflects the strength of winter sports in each country, which is the fourth dimension. Furthermore, we plan to add time, discipline, and event filters to allow users to explore the trend of ice sports freely and intuitively. We believe that our project can be useful to the majority of ice sports enthusiasts.

3 Exploratory Data Analysis

Before visualizing our dataset, we first do processing using Python to find out its characteristics in several aspects.

3.1 Temporal Characteristics

The dataset provides the time when the medal was awarded, so we count the total number of medals for each Winter Olympics. As shown in Fig. 1, it is clear that the number has increased over time, especially from the 1920s to the 2000s, while from 2018 to 2022 there is a dramatic increase, reaching 700.

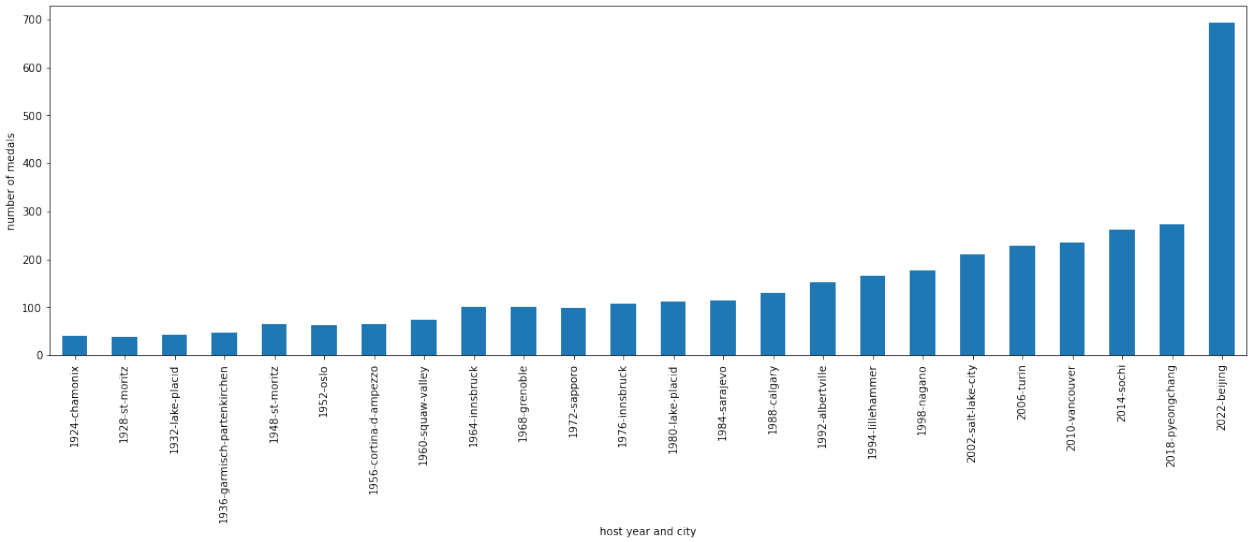

Figure 1: The number of medals at each Olympics.

3.2 Categorical Characteristics

The discipline and event information is also available. There are 19 disciplines in these Winter Olympics, and the number of medals of each discipline varies a lot, as shown in Fig. 2.

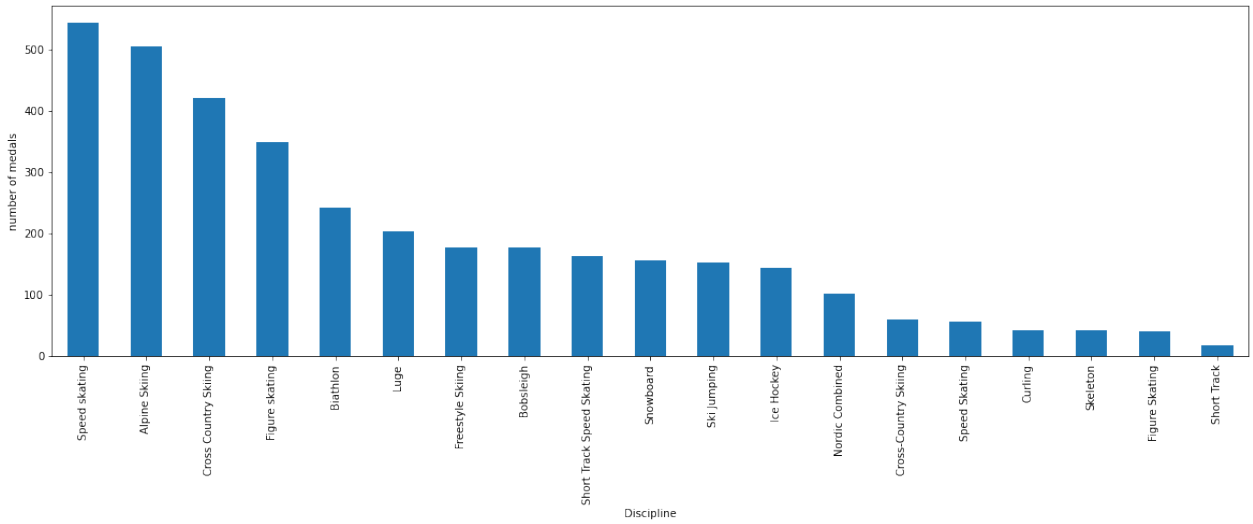

Figure 2: The number of medals at each discipline.

Speed skating has the largest number of medals (545) while Short Track only has 18 medals. Two reasons lead to this difference. Firstly, some disciplines are newly developed compared to others which exist for a long time. Another reason is that the event number of different disciplines also varies. For instance, Speed skating has 25 events, while many others are only divided into Men and Women.

3.3 Spatial Characteristics

We also try to figure out the spatial characteristics of Olympic athletes. The dataset includes committee information. Thus, we first count how many medals each committee won, as shown in Fig. 3. As the figure indicates, only 45 committees once won medals in Winter Olympics. Besides, this number contains committees with different names in different games, e.g. Russia and ROC.

Then, we use GeoPy to get the corresponding longitude and latitude so that we can draw the map of athletes' nationality, as presented in Fig. 4. In Fig. 4, each circle corresponds to a country, and its size represents the number of medals won by its athletes. It is displayed that European countries have the most medals, while African countries have less.

4 Related Work

It is imperative for us to understand how should we demonstrate the data. Therefore, we look up the previous work on the datasets *Olympic Games, 1986-2021* and *Beijing 2022 Olympic Winter Games*. What's more, our approach and inspiration are analyzed in the following sector.

4.1 Previous Work

Previous work on this dataset focuses on categorizing different information. For example, hosts, medals, and results are extracted separately. The provider of the dataset also drew graphs depicting both timeline of disciplines contested at the Winter Olympics, 1924-2018.

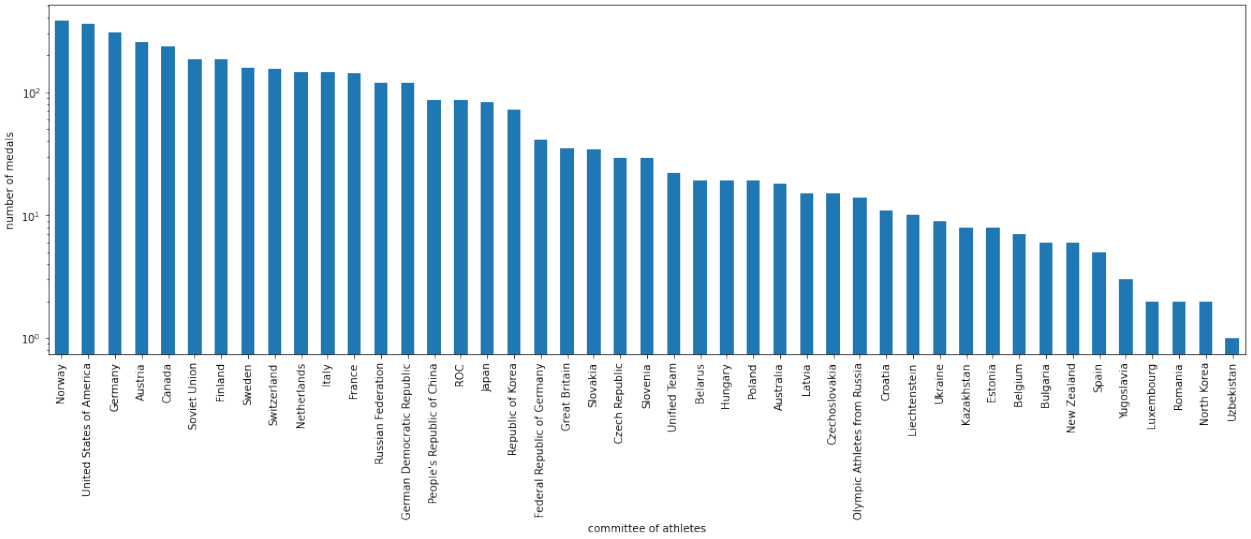

Figure 3: The number of medals in each committee.

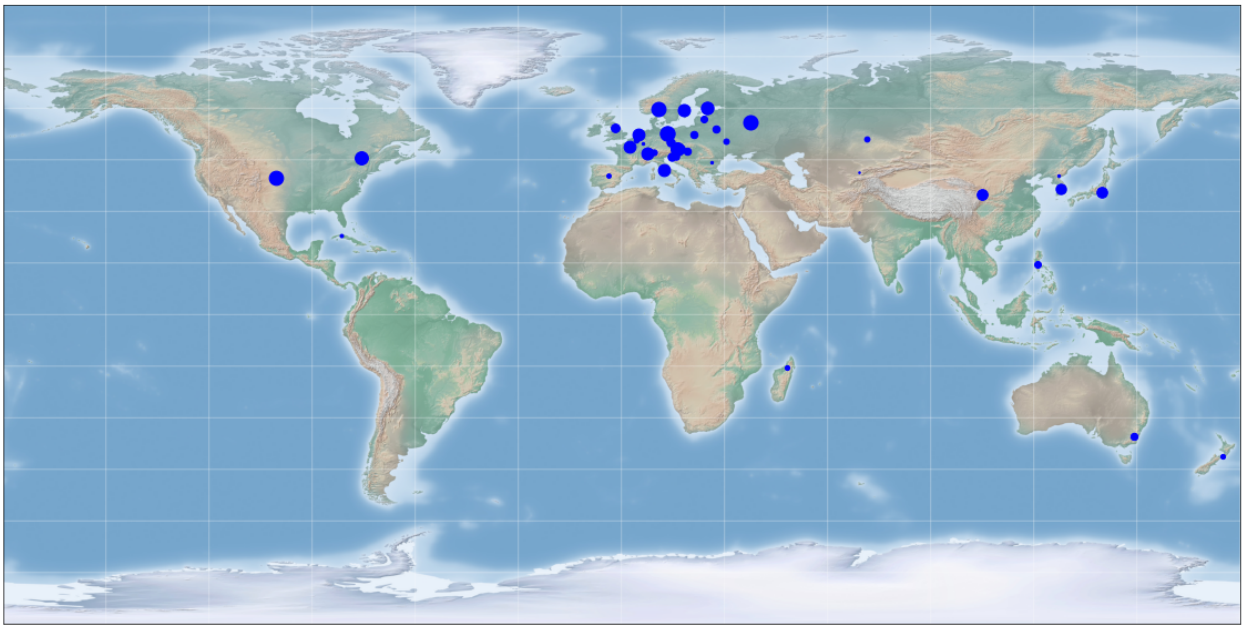

Figure 4: The number of medals in each country.

These graphs vividly show the sports held in each winter Olympics. He also concludes the medal heat map for each country. Besides, there are some works concerning athletes. For example, top medal-winning athletes and nations are demonstrated in bar graphs.

4.2 Originality

It is obvious that previous work mainly aims to categorize nations and competition items. So we decide to build up a website to show people athletes' information throughout history. Indeed, most of the previous work lacks athletes' information, which could be the birth date and place, gender, and so on. Nevertheless, the medal heat map considering each country is not intuitive enough, since there is little information about athletes shown on the map. We want to design a website that introduces past Winter Olympic Medalists and their profiles by a world map. Therefore, users could know athletes' information by clicking their country. The number of medals obtained by each country is also shown on the map.

4.3 Inspiration

Data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. According to [3], Geographic Information System is the most effective for problem identification and brainstorming. In this case, we decide to utilize a geographic map to illustrate data to improve the visualization effect in order to facilitate users' interpretation.

References

[1] Petro. Olympic games, 1986-2021 — results, medals, athletes and hosts from athens to tokyo, 2022. [Online; accessed 24-March-2022].

[2] Petro. Beijing 2022 olympic winter games — full data about medals and athletes. ice hockey and curling results, 2022. [Online; accessed 24-March-2022].

[3] Kheir Al-Kodmany. Using visualization techniques for enhancing public participation in planning and design: process, implementation, and evaluation. *Landscape and urban planning*, 45(1):37–45, 1999.