Olympics Dataset Descriptive Stats

Jingying Pan

September 11, 2022

1 Olympic Dataset

This dataset contains two tables: *events* and *regions*. The table *regions* has three columns(NOC, region and notes) and 30 observations. The table *events* has 15 columns(ID, Name, Sex, Age, , Height, Weight, Team, NOC, Games, Year, Season, City, Sport, Event and Medal) and 271115 observations. This project would focus on trends of medal on age/countries/height/weight, etc.

The reason why I choose this dataset is because I like sports and watching Olympic games. I believe that the result of this project can give the sports teams some instructions on selecting the potential athletes. Also there may be some unreasonable rules in some kind of sports which can be found from the results.

2 Data Cleaning

First, I removed the duplicates. The clean dataset *events* has 269731 observations and *regions* has 30 observations.

Second, I filled the missing value of Height and Weight in *events* with the average of each athletes. Figure1 shows the first 5 records in the new table *events*. Some athletes don't have any records about their height and weight, so there are still some missing values in this table.

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event	Medal
0	1	A Dijiang	M	24.0	180.0	80.0	China	CHN	1992 Summer	1992	Summer	Barcelona	Basketball	Basketball Men's Basketball	NaN
1	2	A Lamusi	M	23.0	170.0	60.0	China	CHN	2012 Summer	2012	Summer	London	Judo	Judo Men's Extra-Lightweight	NaN
2	3	Gunnar Nielsen Aaby	M	24.0	NaN	NaN	Denmark	DEN	1920 Summer	1920	Summer	Antwerpen	Football	Football Men's Football	NaN
3	4	Edgar Lindenau Aabye	M	34.0	NaN	NaN	Denmark/Sweden	DEN	1900 Summer	1900	Summer	Paris	Tug-Of-War	Tug-Of-War Men's Tug-Of-War	Gold
4	5	Christine Jacoba Aaftink	F	21.0	185.0	82.0	Netherlands	NED	1988 Winter	1988	Winter	Calgary	Speed Skating	Speed Skating Women's 500 metres	NaN

Figure 1: First 5 records in events

3 Data Wrangling

In order to analyse the medal results, I quantified the competition result by: Gold=3, Silver=2, Bronze=1, Null=0. So I added a new column *score* to represent the medal results.

The column *NOC* in *events* cannot be used in Tableau directly. So I use left join to join *events* and *regions* together. Then I got a new table *joint* which contains every information I needed in the following analysis.

4 Data Exploration

I used histgram to indicate the trend of records on Year, Age. Then I used Tableau to draw a map to show the relationship between medal result and average height, average weight. Because these two method can quickly show us what we want in an obvious way.

Figure 2 shows the ERD diagram of the two tables. It is a one-to-many dataset. We would consider join the two tables and work on the regions of the events in the following analysis.

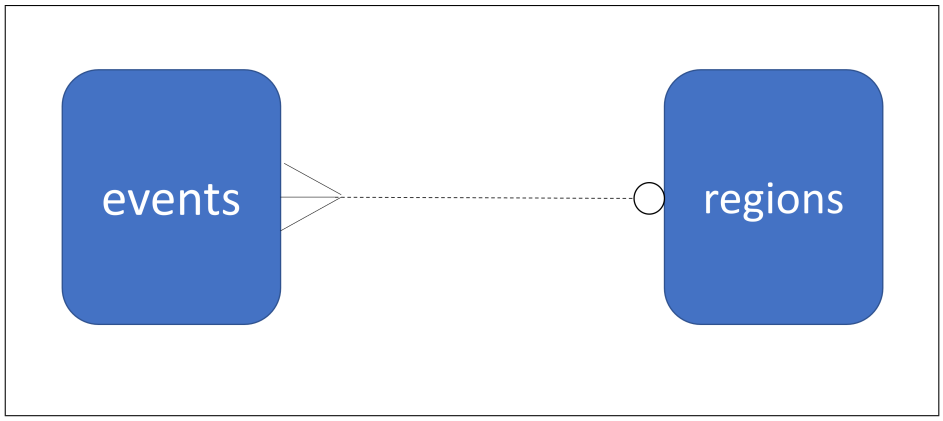

Figure 2: ERD diagram

Events has records range from 1896 to 2016. We made a histogram on the Year of the records in *events*. Figure 3 shows the trend of records on Year which indicates the records increased during history and has a bump during 1930s. Figure 4 shows the trend of records on Age. It indicates that most of the athletes are in age range of 20 and 30. It is reasonable and compatible with common sense.

After that I created a new table *region_score* which contains *score*, *region*, average height of each region, average weight of each region. Then I imported this table to Tableau to draw a map to show the result, as shown in Figure 5. The size of the dots represents the scores in summary. The color of the dots represents the average height in the Olympic records of that region. As we can

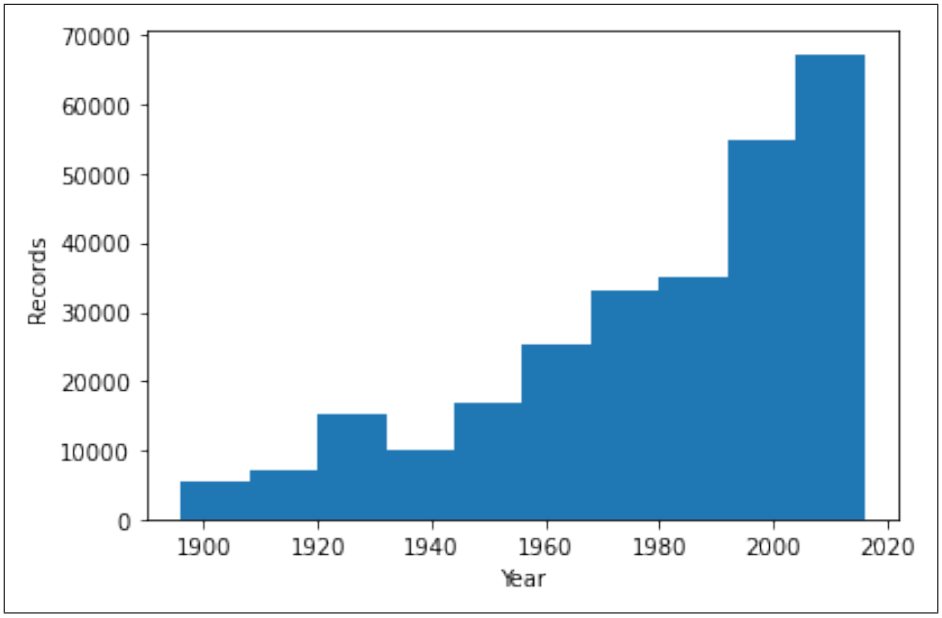

Figure 3: Trend of Records on Year

see, the European people got best result in Olympic Games. We can roughly conclude that the higher the athletes are, the better they will get.

Figure 6 shows the score and the average weight in each region. The size of the dots represents the scores in summary. The color of the dots represents the average weight in the Olympic records of that region. We can also roughly conclude: the heavier the athletes weights, the worse they will get.

5 Findings

From the initial analysis on the Olympic Games dataset, I can tell that most of the athletes are in age range of 20 and 30. The region of the athletes has considerable effect on the medal result. The height and the weight also has effect on the medal result. The higher the athletes are, the better they will get. The heavier the athletes weights, the worse they will get.

6 Future Analysis

In the future, I will calculate the relationship between the medal result and other factors with theoretical methods such as Pearson's Correlation and Spearman's Correlation. How about seperate the dataset according to sex? Is there any difference in the conclusions for different sex? They are my future questions.

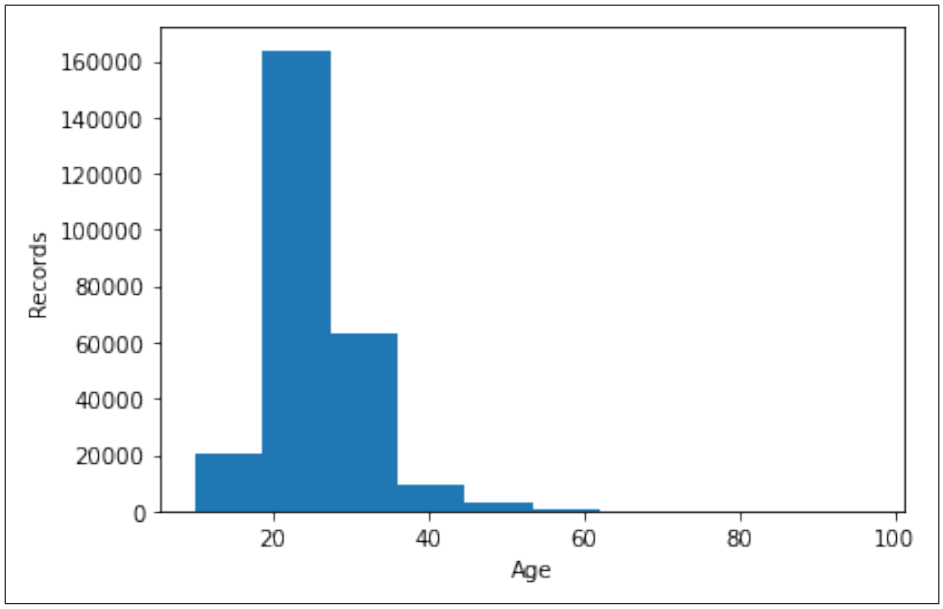

Figure 4: Trend of Records on Age

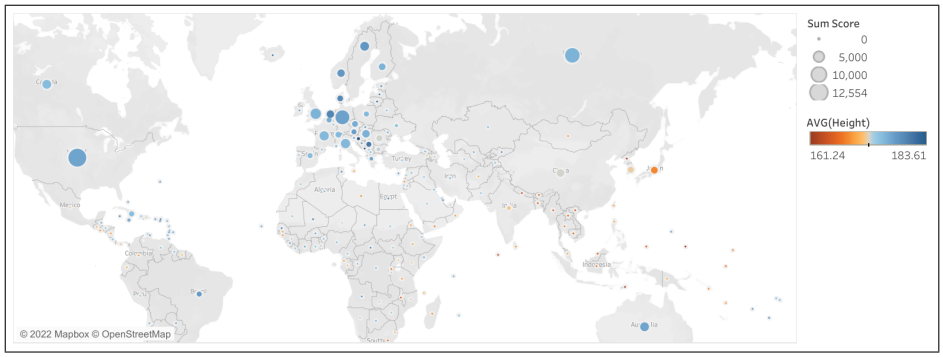

Figure 5: Map to show the score and average height

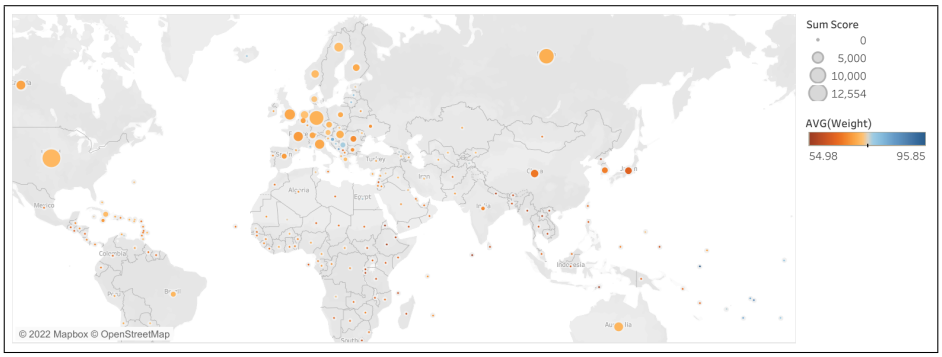

Figure 6: Map to show the score and average weight