JOUR 2076 Data Journalism Individual Report

My Topic: In each kind of sports, what age of the players are more likely to get achievements?

Step One: The inspiration

Olympics Tokyo 2020 may be the one of the most special Olympic Games. Due to the pandemic, it was put off to this year 2021. Few months ago, when I was watching the Olympics Tokyo 2020, I always heard the narrator said, “someone is 27 years old now and it may be his last Olympic games because as the age grows, it is hard to keep his current race”. But I was wondering, what was the best age for player? In other words, what age of athletes were more likely to get achievements? My hypothesis is athletes aged between 22 to 25 are more likely to get achievements.

Step Two: Find the data

After deciding my topic, I began to find data. It is quite easy to find the medalist information on the official website (<https://olympics.com/en/olympic-games/tokyo-2020/athletes> ), while it is hard to scrape it.

As we learnt the web scraper in the class, I used web scraper to scrape data.

First, I searched and decided what data should I use. I listed them:

1. Name of persons who get prizes
2. Birth of the person (To get their age)
3. Gender
4. Participated sports
5. Their awarded medals in 2020

However, I met a problem when I scrape data. Before I present my progress in October, the website is like:

图形用户界面, 文本, 应用程序, 电子邮件

描述已自动生成

After the presentation, when I opened the website for the second time, it was found the website has changed to:

Graphical user interface, text

Description automatically generated

The previous website changed to:

图形用户界面, 应用程序

描述已自动生成On the new website, it doesn’t show the gender. Therefore, I would use the previous database mainly to present the topic. But I would show how I scrape the data on the current page.

In the process of data scraping, it takes my time. As the website has lots of information and need to scroll down to get full medalist information. Therefore, I learnt how to scroll down with web scraper to get accurate and full information.

Graphical user interface, text

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

The progress of waiting is torment, web scraper needs more than one hour to scrape all the data from the website. As I could not make sure the results have right formats, I tried many times and waited for several hours, sometimes I got only rows, sometimes I got wrong formats. But finally, I got all the correct data.

Graphical user interface, application, table

Description automatically generated

Step Three: Clean the data

After I scraped all the data I need, I found there are some “null” blanks or strange message inside. Therefore, I must clean the data. I used both Excel and Open-refine. When I was cleaning the data, I found Excel and Open-refine are complementary. It is convenient to separate various types of data with Excel, while Open-refine helps us to delete the useless data.

I deleted the useless information and re-organised the data in one blank:

Graphical user interface, application, table, Excel

Description automatically generated

I don’t want the “social media” blanks.

So, change it to null and delete it.

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, application, table, Excel

Description automatically generated

Then delete the “#Value!” blanks.

And extract useful information: Numbers of medals.

Graphical user interface, application, table, Excel

Description automatically generated

Graphical user interface, application, table

Description automatically generated

Delete the empty blanks: already change them to null.

图形用户界面, 文本, 应用程序, 表格

描述已自动生成

Graphical user interface, application, table, Excel

Description automatically generated

Make all related information in one row (same name in same row):

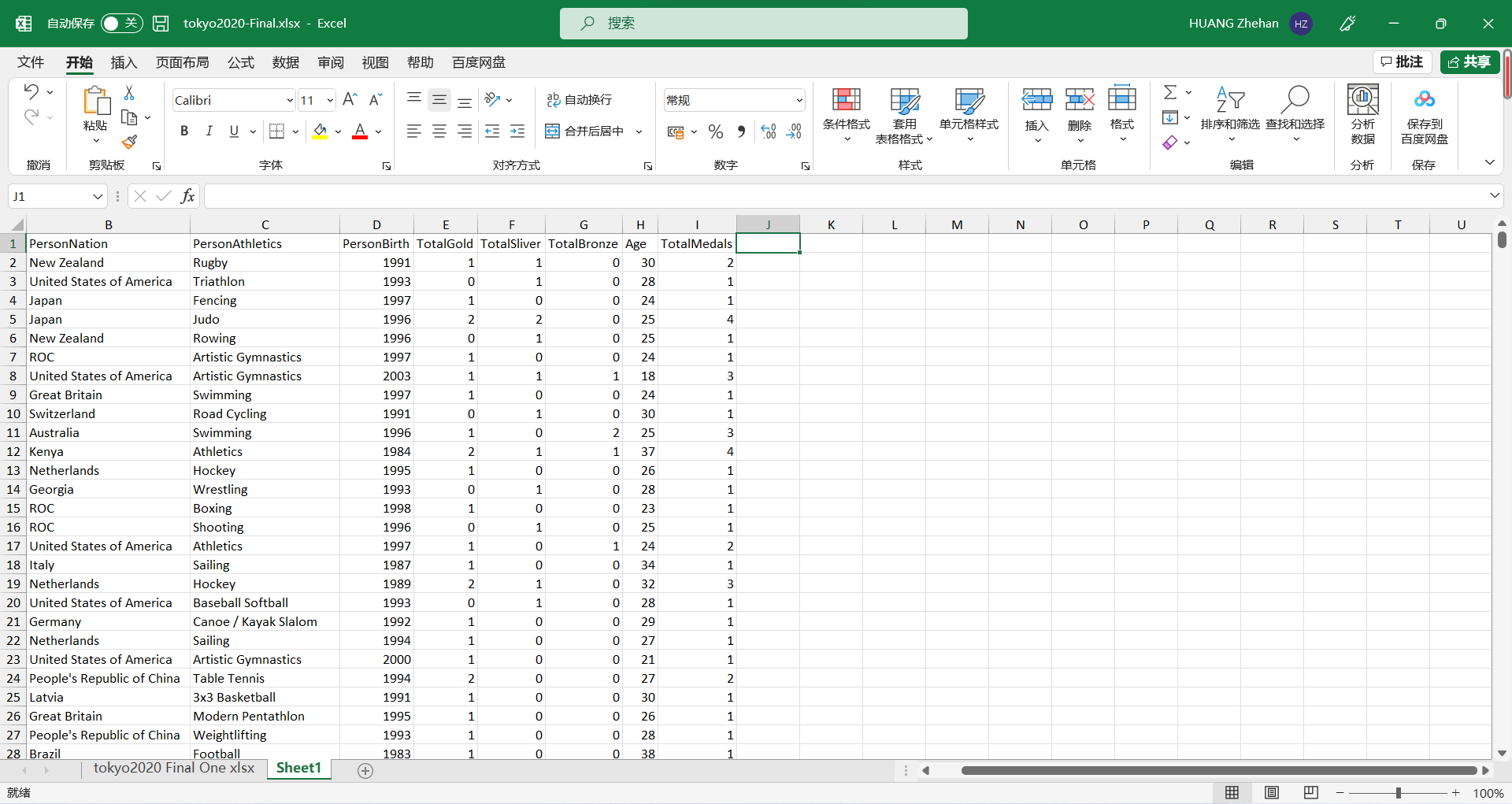
Graphical user interface, application, table, Excel

Description automatically generated

Graphical user interface, application, table, Excel

Description automatically generated

The final data table is like:



Step Four: Data Visualization

(As I said, the current data excludes the gender, so I use the pervious data set.)

Conclusion One: Female athletes are less than male athletes in Olympics Tokyo 2020.

图形用户界面, 文本, 应用程序

描述已自动生成

Conclusion Two: Athletics has the most players in Olympics Tokyo 2020.

图表, 树状图

描述已自动生成

Conclusion Three: Different athletics has different “best age”.

Detailed graph please go to:

<https://public.tableau.com/app/profile/huang.zhe.han/viz/1_16375362111470/1?publish=yes>

图形用户界面, 应用程序

描述已自动生成As showing in the graph, there is little data to make a judgement for some sports, for example, the Cycling BMX Freestyle or Golf.

For some sports, most medalists are between around 23-30 years old, for example, the Athletics, Rowing and Football.

For some sports, the medalists are quite young, for example, the Swimming, Artistic Gymnastics. It is hard to see medalists that are older than 30 years old.

However, not all sports obey the rule: the younger, the better. For example, in Equestrian, most of medalists are above 30 years old. The oldest medalist occurs in Equestrian in Olympics 2020.

The difference may cause by the different characteristics of sports. Artistic Gymnastics assesses the flexibility of each player, experienced horsemen are more likely to win the prize, while younger people with better stamina adapt to swimming for longer periods of time. Based on the different features, the conclusion may be different.

Conclusion Four: Players aged between 25 to 27 are most likely to get achievement in general.

图表, 散点图

描述已自动生成

Step Four: Limitations

Limitations:

1. The conclusion can be more precise if I scrape the previous Olympic Games data. I tried for RIO 2016, but the data are so large (In the appendix). I need to learn to build a database for it, but my technique skills limits.
2. As the website disappear suddenly, it doesn’t show how many pages in total, so I cannot make sure the data of the second time are complete.
3. Though the topic is interesting and meaningful, it is not so surprising when the results show. They are quite easy to understand.

Word Count:844 (Exclude the title and topic)