Big Data Project:

Diversity of Olympic Games

Ilhame Hadouch, Abderrazzak Mahraye, Hicham Skiker

INSA TOULOUSE ISS B1

You can find all the source code, and the dataset on this github:
https://github.com/mahraye/DataProcessingMahrayeHedoucheSkiker_5ISSB1

Contents

I Introduction

During this semester, we worked on the processing of Big data. After some practical tutorials, we were asked to work on a project with R language. The aim of this project was to analyse and show some results of a dataset. In this report we will present our work.

The aim of this analytic work is to show how Olympic Games are diversified. In fact, we will analyse the winter Olympic Games and the Summer Olympic games in order to compare them.

II Dataset Description

For this project, we chose to analyse a dataset based on players of Olympic Games. This dataset is a table of 271116 "samples". For each sample, we have the following informations:

- ID which is an integer identifier, Name, Sex, Age, Height, Weight, Team, NOC, Games, Year, Season, City, Sport, Event

All information is separated by a coma and given in quotes unless the ID. Here we have two examples of what we call samples:

```
1,"A Dijiang","M",24,180,80,"China","CHN","1992 Summer",1992,"Summer","Barcelona","Basketball","Basketball Men's Basketball",NA
2,"A Lamusi","M",23,170,60,"China","CHN","2012 Summer",2012,"Summer","London","Judo","Judo Men's Extra-Lightweight",NA
```

III Result's Presentation

First of all, we decided to focus on the winter Olympic Games from 1980 to nowadays. To do that, we used the function filter from the *dplr* library.

Then, we did the same work for the summer Olympic Games from 1980 to nowadays in order to compare our results between.

III.1 The five sports with the bigger number of medals won

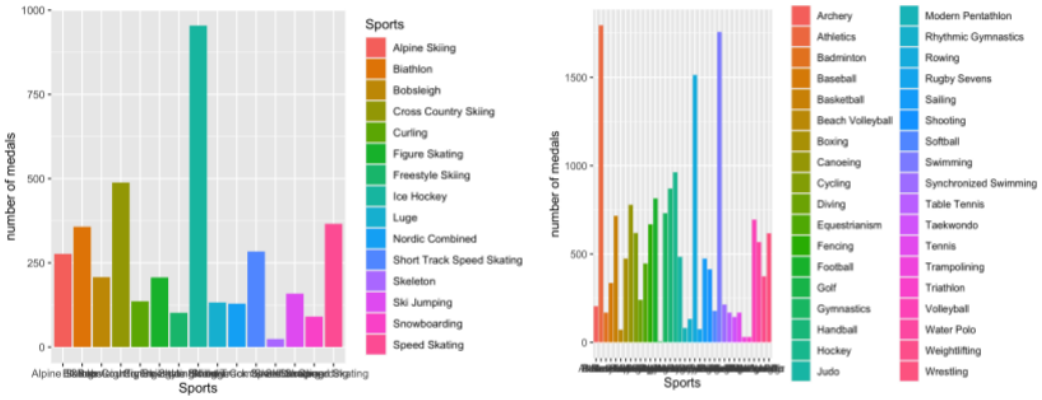


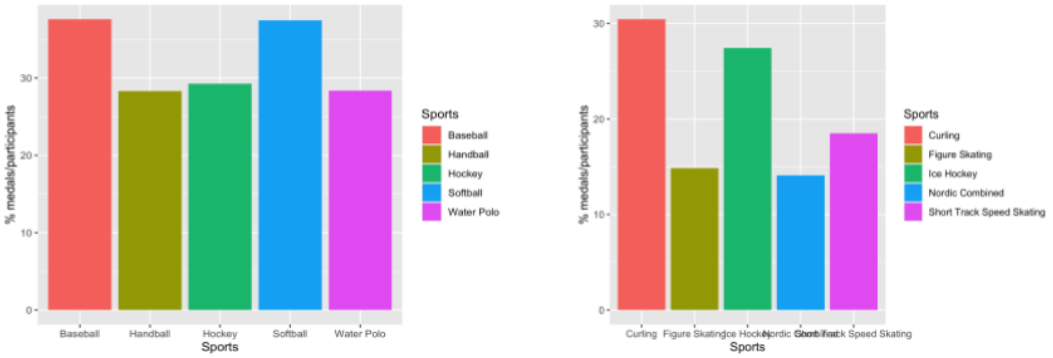
Figure 1: Medals wins by sport (Winter) VS Medals wins by sport (Summer)

We tried to find the five sports which count the more medals. For that, we plotted an histogram showing the number of medals for each sport. We can easily see the five sports with the bigger number of medals and in the increasing order, we have: ice hockey, cross country skiing, speed skating, biathlon and short track speed skating for winter olympic games and Athletics, Swimming, Rowin, Hockey, Handball.

We can note, by the way, the diversity of sports that are present in the Olympic Games. Consequently, it was for that reason that we worked only on the first five sports.

III.2 The medals per participants ratio

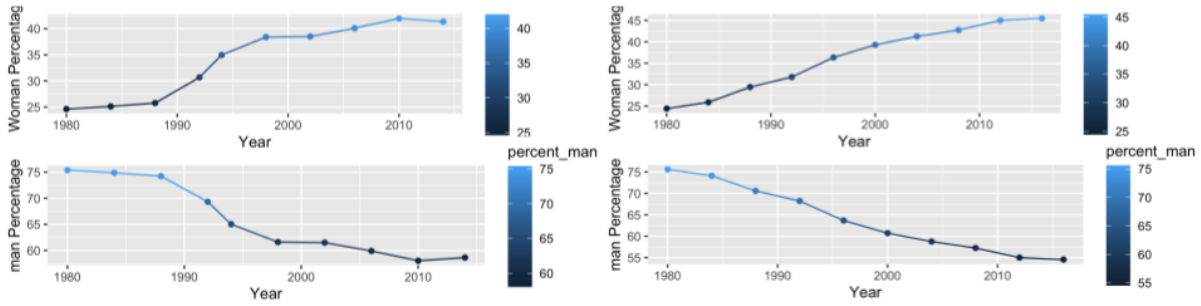
We got the top sports with the bigger number of medals won but , are the number of medals influenced by the number of participants ? And which sport got the best ratio of won medals/ participants in this sport ?



Medals per Participants ratio

As we can see, the top five changed. Actually, the participants influence the number of won medals however some sports got better medals won / participants: Curling in the winter OGs, and Baseball/ Softball won a lot of places and became the firsts. Thus, the sports with a low number of participants can have a lot of medals.

III.3 The evolution of participants number in function of time



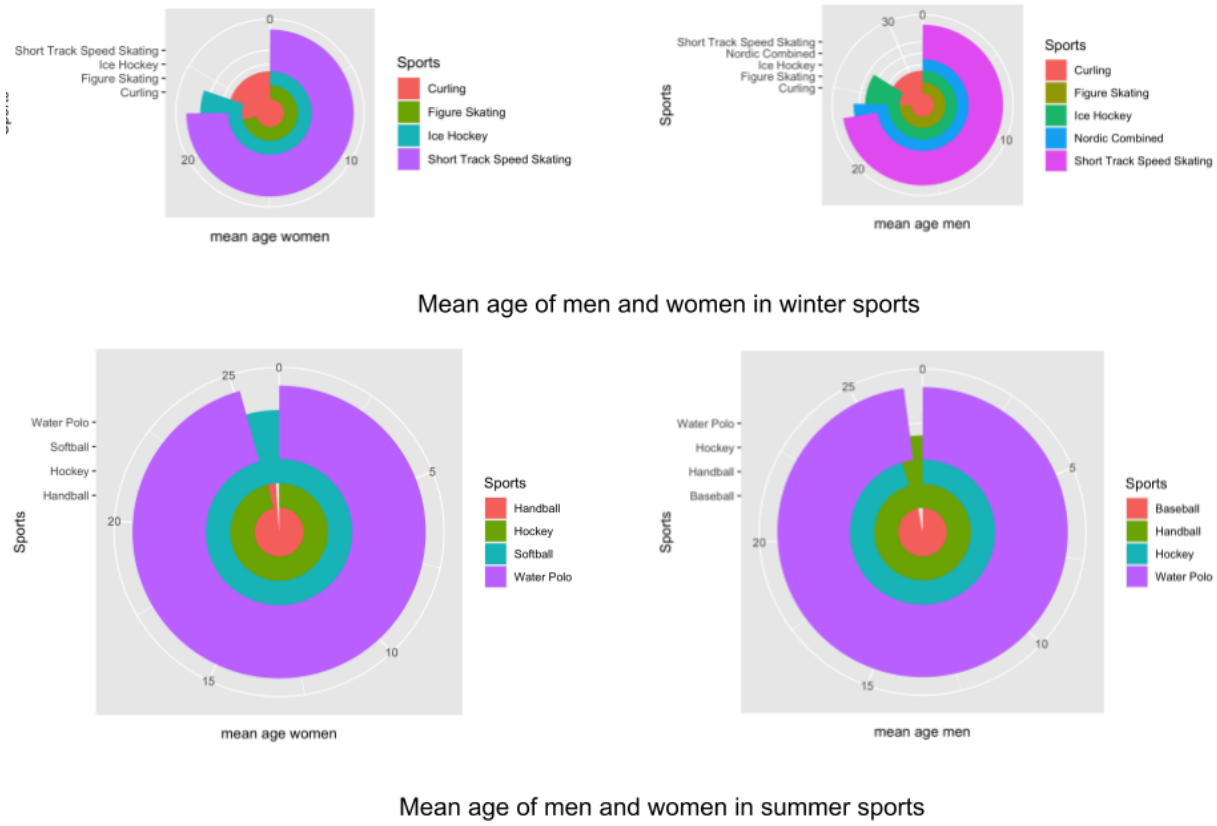
Evolution of the number of men and women between 1980 to nowadays
for winter sports (left) and summer sports (right)

In this part, we have studied the percentage of men and women players in function of the years. In the following graphs, we can see that two cases, i.e. in winter (left) and in summer (right), between 1980 and nowadays, the percentage of women has increased whereas it has decreased for men. Also, we can see that in the two situations we have exactly the reverse tendency of graphs. We can also note that in the summer sports the percentage of women and men are almost 50/50 (45/55) in 2014. We think that in the next decade we will approach this percentage.

For the winter OGs, in 2014 the percentage are 40/60. Thus, the women grow is a little bit lower than the summer one. In fact, the summer one is a decade ahead of the winter one: the 40/60 is already obtained at the year 2000.

That is very interesting regarding the man/woman equality in our societies. This graph shows how this equality evolved from the 80s to 2010s.

III.4 The mean age of the players



In this part, we wanted to study the age mean in the top sports in order to compare the results for men and women. The different following diagrams show that there is a variation between men and women in terms of age: the mean age of women is lower than the mean age of men. For example, for the winter ones, the age of woman is on average two years lower than the men ages (Curling 32 vs 30, Figure Skating 23 vs 21, Ice Hockey 26 vs 24). We can note that the participants are on average young ones. In fact, on average the interval is 21 to 32. On average, There aren't a lot of teenagers, and elderly people.

Moreover, we can see that on average, the ages of Winter Olympic Games are lower than the Summer's ones.

III Conclusion

These analytics showed how diversified are the olympics games from the 1980s to the 2010s. It was interesting to put the data into plots, and we were surprised by the figures. In fact, it showed us how important doing analysis of data is and how we can retrieve a lot of interesting information from big data.

Moreover, in this report, we have seen that we could use our dataset for different studies. This shows the importance of big data. Working on this project allowed us to understand better how to use and analyse datas, it gave us an overview of what it is possible to do with big data. It was also a good opportunity to work with R language which is used by some companies.