Content

1- Client/Hypotheses/Approach, Project Proposal and Data Selection/Preparation

- About client
- About data
- Data cleaning
- Data sample
- Some general statistics
- ERD
- Goal of this study
- Questions to answer
- Some hypothesis
- Approach
- Metric
- Correlations

2- Discovered insights

- Influence of being the host
- Influence of the globalization

3- Conclusions and recommendations

1- Client/Hypotheses/Approach, Project Proposal and Data Selection/Preparation

About client :

I chose to select the SportsStats dataset because It's always funny to try to predict the hierarchy of countries at the Olympics. The clients are gamblers and bet companies.

About data :

I downloaded two files from the following link :
https://www.dropbox.com/sh/0wqw8fmiwrzr8ef/AABQijjQM522INXX1FCdamzma?dl=0

- athlete_events.csv
- noc_regions.csv

which contains respectively the data about the athletes with the medals they won, and the list of the NOC (trigram of the National Olympic Committees) from the athletes come under.

Data cleaning

- Concerning the NOC data, I cleaned it by replacing some wrong country names, for example « United States of America » instead of USA, « United Kingdom » for UK, Bolivia for Boliva and so on.

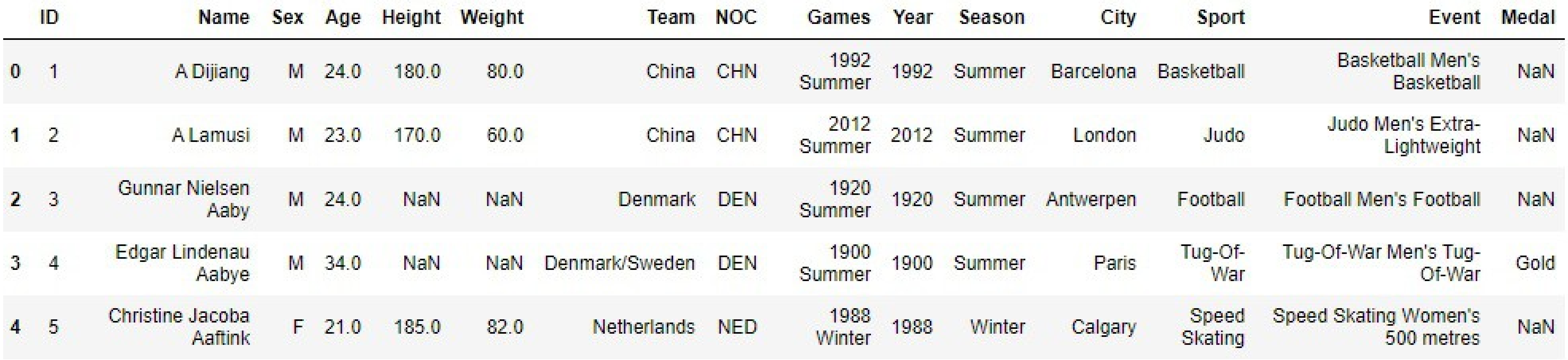
- I also replaced « NA » by Tuvalu on the TUV line, « NA » by « Refugee Olympic Team » on the ROT line and « NA » by « N/A » for the unknown so that NA only corresponds to an absence of Information.

- For the athletes, there's a lot of NaN data, 'Age' : 9474, 'Height' : 60171, 'Weight', 62875, but for these data, we can't replace them by 0 or by the mean, that wouldn't make sense.

- There is a country that has athletes in the athlete file but whose NOC does not exist in the committee file, it's Singapore. According to Wikipedia, the committee currently exists but only from 1947 and it merged for 2 years with that of Malaysia. But since all the Singaporean athletes, on the list, participated in the Olympics where the committee existed as Singapore, we can add the SGP / Singapore entry in the regions file.

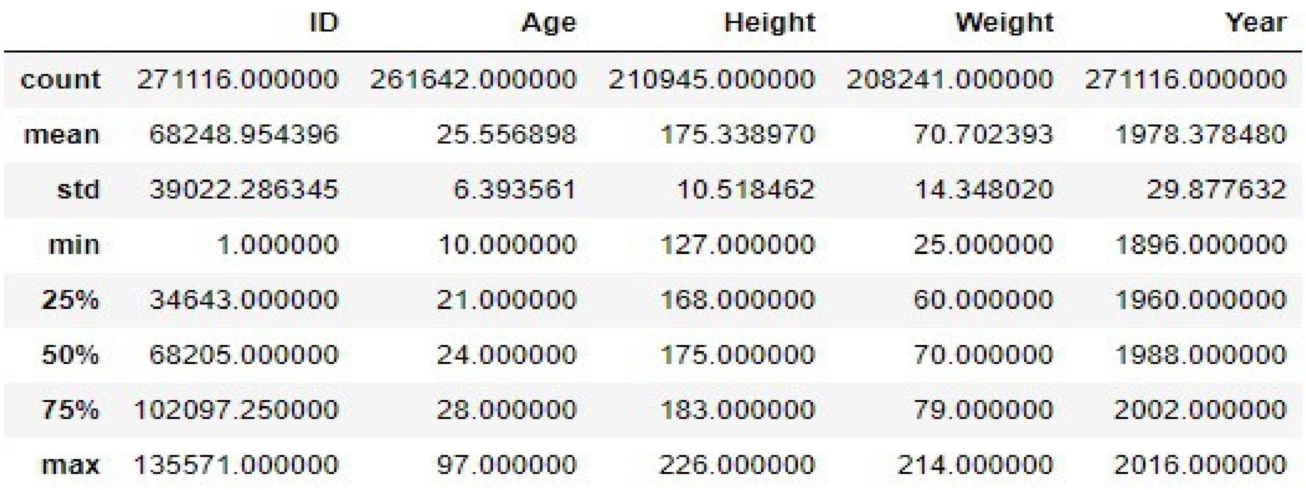
Data sample

Athletes dataframe :



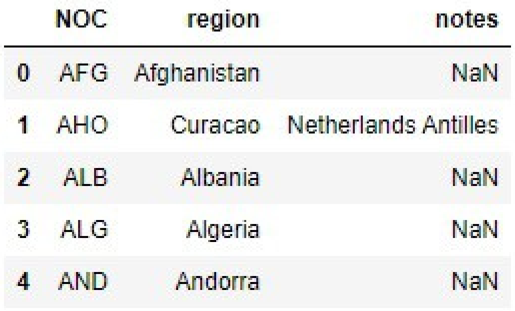
	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event	Medal
0	1	A Dijiang	M	24.0	180.0	80.0	China	CHN	1992 Summer	1992	Summer	Barcelona	Basketball	Basketball Men's Basketball	NaN
1	2	A Lamusi	M	23.0	170.0	60.0	China	CHN	2012 Summer	2012	Summer	London	Judo	Judo Men's Extra-Lightweight	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

Athletes statistics :

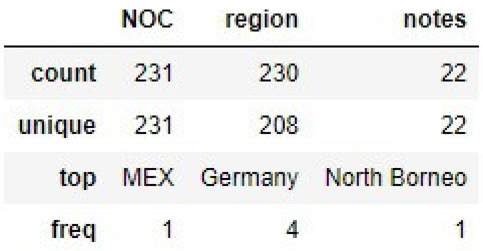


	ID	Age	Height	Weight	Year
count	271116.000000	261642.000000	210945.000000	208241.000000	271116.000000
mean	68248.954396	25.556898	175.338970	70.702393	1978.378480
std	39022.286345	6.393561	10.518462	14.348020	29.877632
min	1.000000	10.000000	127.000000	25.000000	1896.000000
25%	34643.000000	21.000000	168.000000	60.000000	1960.000000
50%	68205.000000	24.000000	175.000000	70.000000	1988.000000
75%	102097.250000	28.000000	183.000000	79.000000	2002.000000
max	135571.000000	97.000000	226.000000	214.000000	2016.000000

NOC dataframe and statistics :



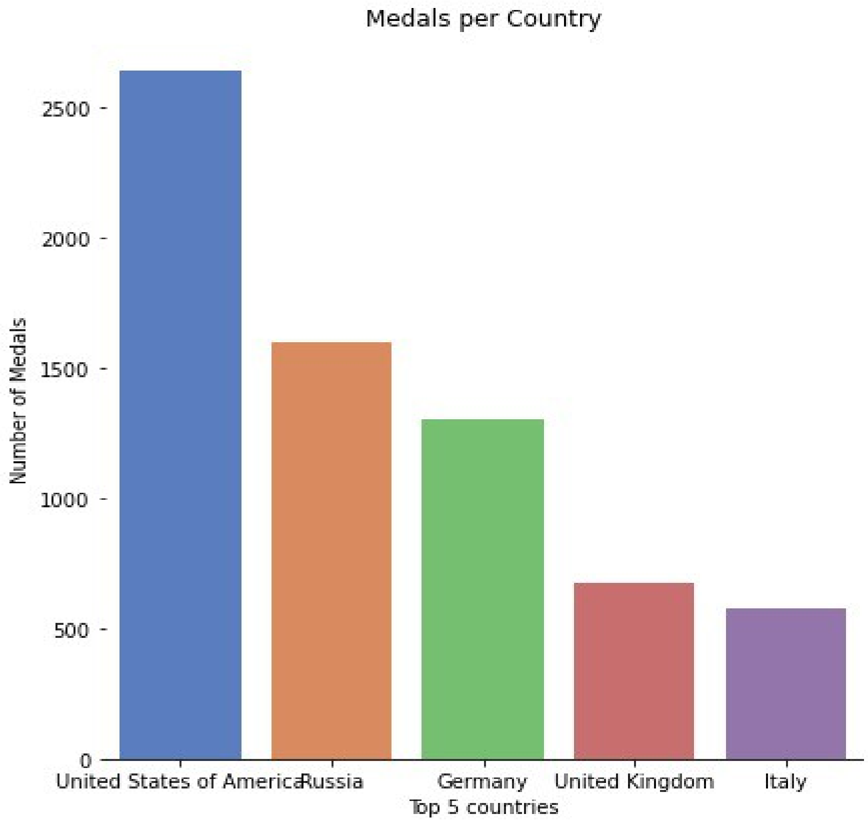
	NOC	region	notes
0	AFG	Afghanistan	NaN
1	AHO	Curacao	Netherlands Antilles
2	ALB	Albania	NaN
3	ALG	Algeria	NaN
4	AND	Andorra	NaN



	NOC	region	notes
count	231	230	22
unique	231	208	22
top	MEX	Germany	North Borneo
freq	1	4	1

Some general statistics :

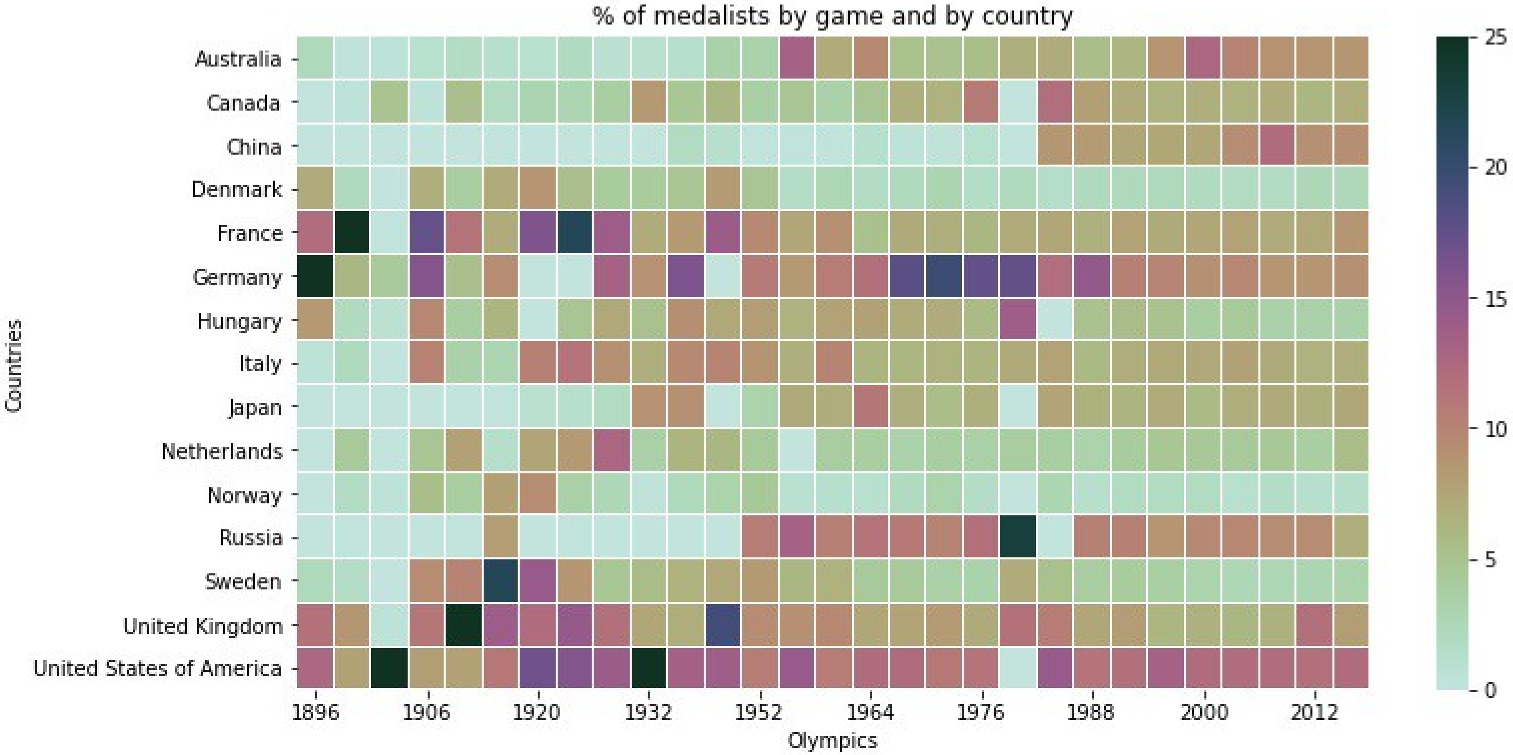
Overall number of medals for the top5 countries :



Percentage of medalists by game and by country for summer games :

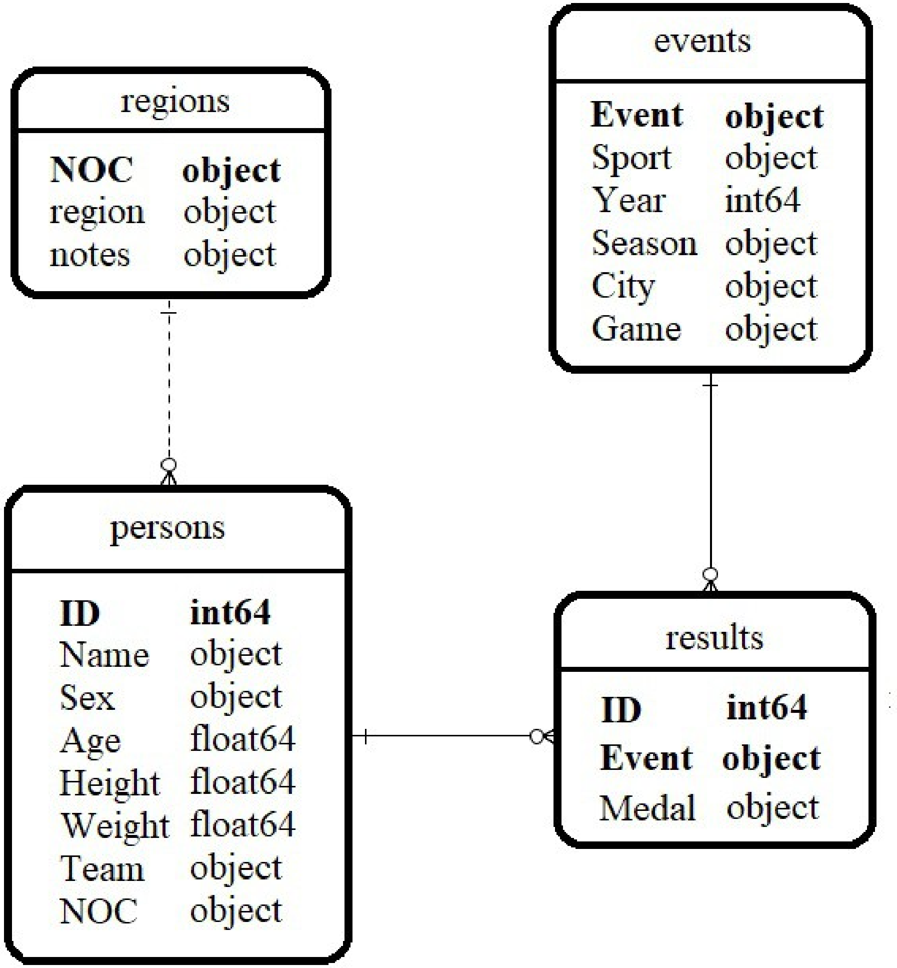
We can see with this chart, the % of medalists by game and by country. For the readability of the chart, I only displayed the top 15 countries of all time and I replaced 5 « big » data because if I don't replace them, we can't see the details of the others, they become almost all pale green. I only kept % under 25% (but I still put this « big » data to the maximum 25% to highlight them).

Germany, 1896, 43.51%
France, 1900, 62.70%
United States of America, 1904, 88.29%
United Kingdom, 1908, 35.10%
United States of America, 1932, 32.70%



With this chart alone, we can see the rise in sporting power of China since 1984, of Australia since 1956, the omnipresence of the USA from the beginning, and the decline of certain European countries.

ERD of the initial datasets



regions	
NOC	**object**
region	object
notes	object

events	
Event	**object**
Sport	object
Year	int64
Season	object
City	object
Game	object

persons	
ID	**int64**
Name	object
Sex	object
Age	float64
Height	float64
Weight	float64
Team	object
NOC	object

results	
ID	**int64**
Event	**object**
Medal	object

Goal of this study

The goal is to find a way to predict how many medals, gold, silver, bronze, total, countries will win. In which discipline, discipline group, man, woman. Which country that has never won medals will perhaps win one? As this is a global event, it could be of interest to gamblers around the world. I don't know this world, but it could also be of interest to betting organizations if ever algorithms were able to predict certain results with almost certainty. If everyone won, it would cause problems ...

Questions to answer :

- What will be the podium of the countries with the most overall medals?
- For a given country, will it win more medals if it is hosting the Olympics?
- Can we use globalization to visualize arriving countries and better predict recent trends?

Somes hypothesis :

What are your initial hypotheses about the data?

- Countries with large populations can get a lot of medals, although that's not decisive.
- The countries which are used, historically, to have many medals, must continue to have them (cultural or geographical specialties of a sport).
- Countries which tend to win more and more medals in certain disciplines, must continue to win more.

Etc.

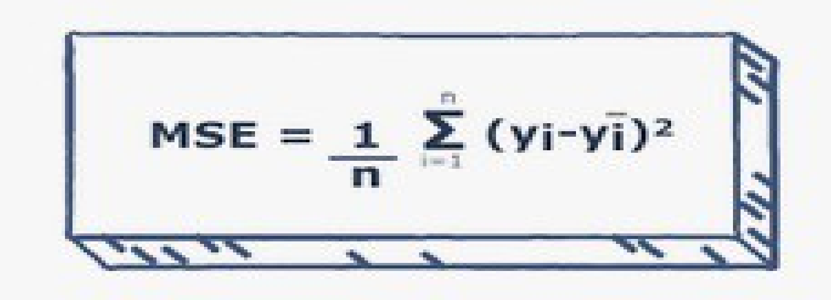
Approach :

It would be easier if we knew which athletes (their distribution) will be competing in the next Olympics, but in this period, it's hard to know and to predict ... So I can only use the trends of previous Olympics.
- For the relationship between the population and the number of medals won, I have to create another dataset containing the populations of the countries.
- For the historical development of countries with a lot of medals, the Year, Season, ID, Medal and NOC fields should suffice.
- For countries which win a lot of medals in certain disciplines and not others, the Event field must obviously be added.

The relationships risk not being linear, so perhaps make them linear by categorical partitions, or use ensemblist algorithms, we will see.

Metric :

Concerning the metric, I think I will choose the mean square error in order to penalize large deviations.



$$MSE = \frac{1}{n} \sum_{i=1}^{n} (yi - y\bar{i})^2$$

Correlations

- Correlation between the size of the population of the countries and the number of medals :

Why : because potentially, in very populated countries there can be a lot of choices of good athletes.
Results : between 1960 and 2020, the correlations vary between 0,23 and 0,43.
So the size of the population is therefore not a good indicator in general by itself but it can be used with other features.

- Correlation between obtaining a medal and the physical data of the athletes (age, height, weight) for individual competitions :

Results :
Age : -0.003
Height : 0.087
Weight : 0.088
No good indicators either because there is too much diversity of age, height and weight in sport to find a general correlation. The thing to do is to look for correlations sport by sport, event by event !

- Correlation between the number of athletes present and the number of medals obtained :

Result : 0.868
Finally an interesting correlation !

Other interesting correlations :

- Correlation between the number of medal won and the GDP:

I had to add a field myself, the GDP indicating the wealth of the countries.
Because countries with great wealth can offer better sport facilities.
Result: there is indeed a good correlation between the wealth of the countries and the number of medals obtained : 0.739

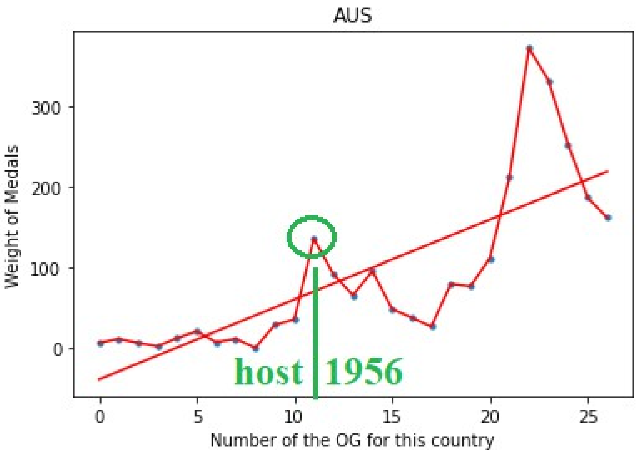
- Correlation between the number of medal won and the host status of the country:

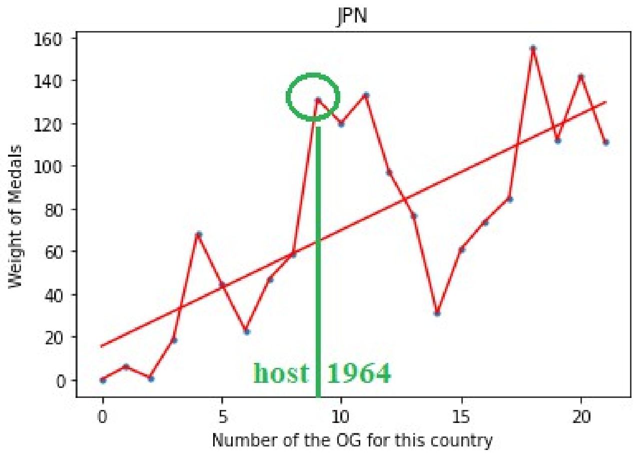
We can observe that for most of countries hosting the OG, they are obtaining much more medals during the Olympic Games but also during the last preceding one and the nexxt one following !!! So when a NOC is hosting, we must increase the number of medals predicted by a significant positive coefficient !

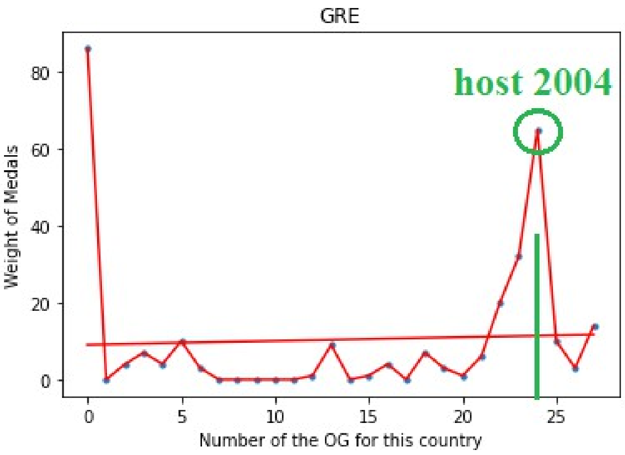
2- discovered insights

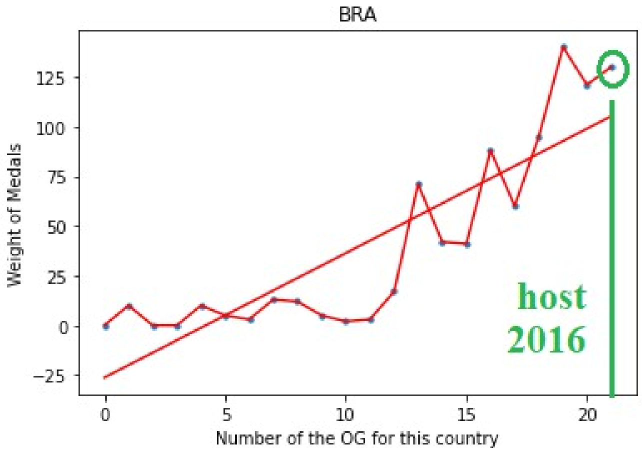
A- First insight, most of the time the host wins much more !

As we can see for Australia in 1956, Japon in 1964, Greece in 2004 and many more, there is a peak when hosting OG and often just before, an increase, and then a small plateau over one or two OG after.









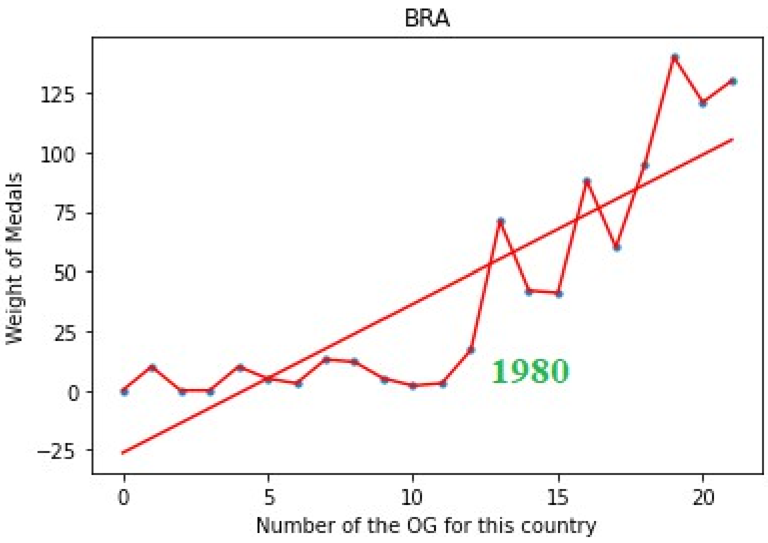
But that's not always the case. Sometimes, like in Brazil in last OG, the progression may have started a long time before the hosted games and then there is not necessarily a peak and only a very high result in the continuity of previous OGs.

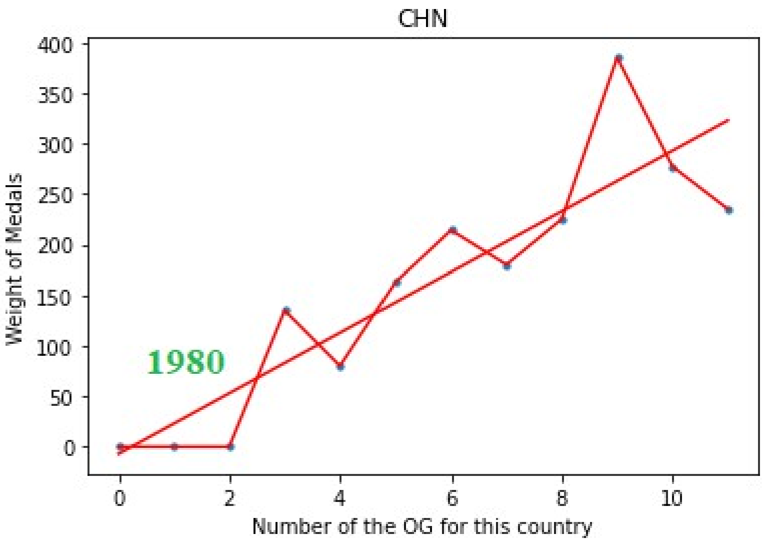
The coefficient of increase, to be applied, therefore also depends on the shape of the evolution curve.

B- second insight, globalization of the world, evolution of the countries.

Some countries have only appeared at the highest level for a few decades, so if we want to use linear regression for the prediction, the years where the countries were not dominant should not be used. The most obvious cases are China or Brazil, whose potentials have only been expressed for about ten games.

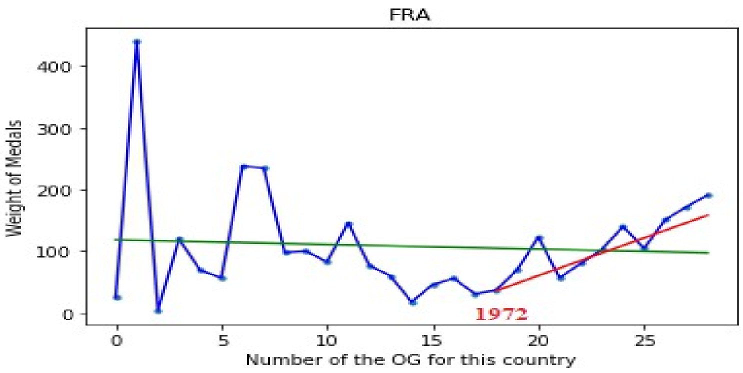
It is a way of incorporating in the linear regression, therefore at a global level in the country and not event by event, the correlations we highlighted, between the number of medals obtained and the emerging wealth of certain countries on the one hand and the number of athletes sent on the other hand.

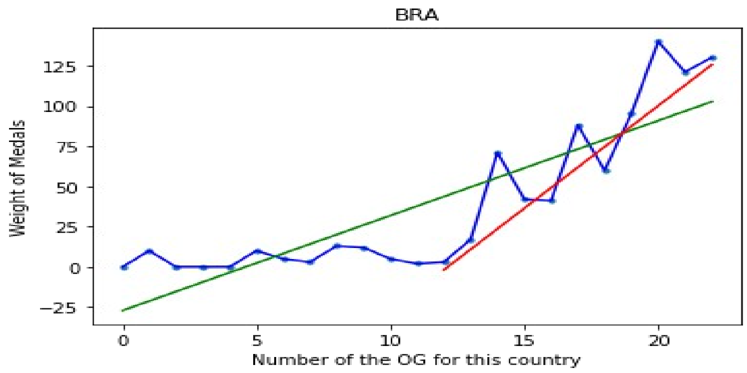




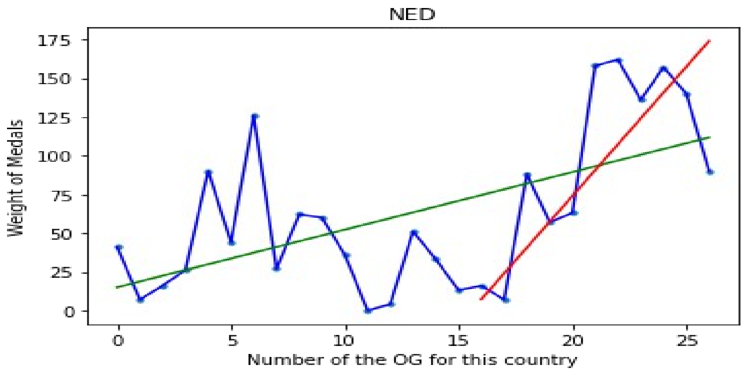
Indeed, as a hypothesis, we can no longer consider contemporary countries as they were a century ago when it comes to predicting their evolution for the next Olympics.

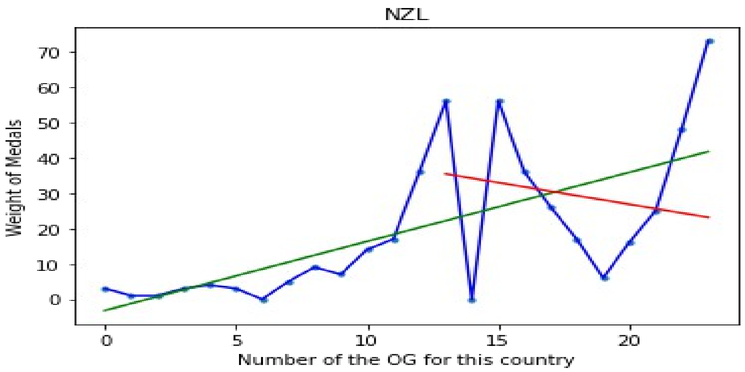
We can see that for some countries, France, Brazil and others, it's a good idea. But for the Netherlands, the and others, it doesn't work.





But for the Netherlands, New Zealand and others, it doesn't work.





$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y_i})^2$$

Overall, this is not a good solution.
MSE for all years : 683.94
MSE after 1972 : 734.73
The MSE for all years is lower thant the MSE after 1972.
We can conclude from that, that the choice to seek a general forecast of the number of medals per country must be made at a much finer level, sport by sport, event by event.

Conclusions and recommendations

1- We found that there were correlations in the data with the number of athletes present and thanks to the contribution of external datasets, such as the GDP and the fact of being the host of the Olympics.
We can conclude that we can compare two countries with each other, but not really predict the number of medals they will win.

2- On the other hand, there is no general correlation when it comes to athlete data.
On all the events, there is too much diversity in weight, height and age to find general correlations. We can conclude that we must look for correlations at a finer level than the set of events, so make event-by-event predictions and add everything together for the overall results.

3- The MSE metric also seems relevant for making event-by-event predictions.

4- I could have used the linear regression at the global level (on all the events) but it is obvious that we would find much better results by doing event-by-event regressions (or another algorythm).

5- I didn't have time to do it, but a way would have to be found to incorporate in the specific event medals time series and possible regressions on athletes (at event level), the more global correlations on the GDP and the number of athletes engaged.