Findings presentation - Data Analysis of Olympic games data

for the Client **SportsStats**

Hassan RHANIMI
August, 2021

Milestone 1

Step 1: Preparing for the Proposal

Client :

"SportsStats" is a sports analysis firm partnering with **local news** and elite personal **trainers** to provide "interesting" **insights** to help their partners.

The Client is chosen because of the size of data files witch can be handled easily and also because I am a fun of sport in general.

Inputs:

We have two csv files as input.

- The "athlete_events.csv" file contains information on the athletes and the events where they participated as well as their result.
- The "noc_regions.csv" contains the country and the National Olympic Committee code

Step 1: Preparing for the Proposal

Figures:

This figures depict the data uploaded as dataframes.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 230 entries, 0 to 229
Data columns (total 3 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   NOC     230 non-null    object
 1   region  227 non-null    object
 2   notes   21 non-null     object
dtypes: object(3)
memory usage: 5.5+ KB
```

```
region.head(3)
```

	NOC	region	notes
0	AFG	Afghanistan	NaN
1	AHO	Curacao	Netherlands Antilles
2	ALB	Albania	NaN

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 271116 entries, 0 to 271115
Data columns (total 15 columns):
 #   Column  Non-Null Count   Dtype
---  ------  --------------   -----
 0   ID      271116 non-null  int64
 1   Name    271116 non-null  object
 2   Sex     271116 non-null  object
 3   Age     261642 non-null  float64
 4   Height  210945 non-null  float64
 5   Weight  208241 non-null  float64
 6   Team    271116 non-null  object
 7   NOC     271116 non-null  object
 8   Games   271116 non-null  object
 9   Year    271116 non-null  int64
 10  Season  271116 non-null  object
 11  City    271116 non-null  object
 12  Sport   271116 non-null  object
 13  Event   271116 non-null  object
 14  Medal   39783 non-null   object
dtypes: float64(3), int64(2), object(10)
memory usage: 31.0+ MB
```

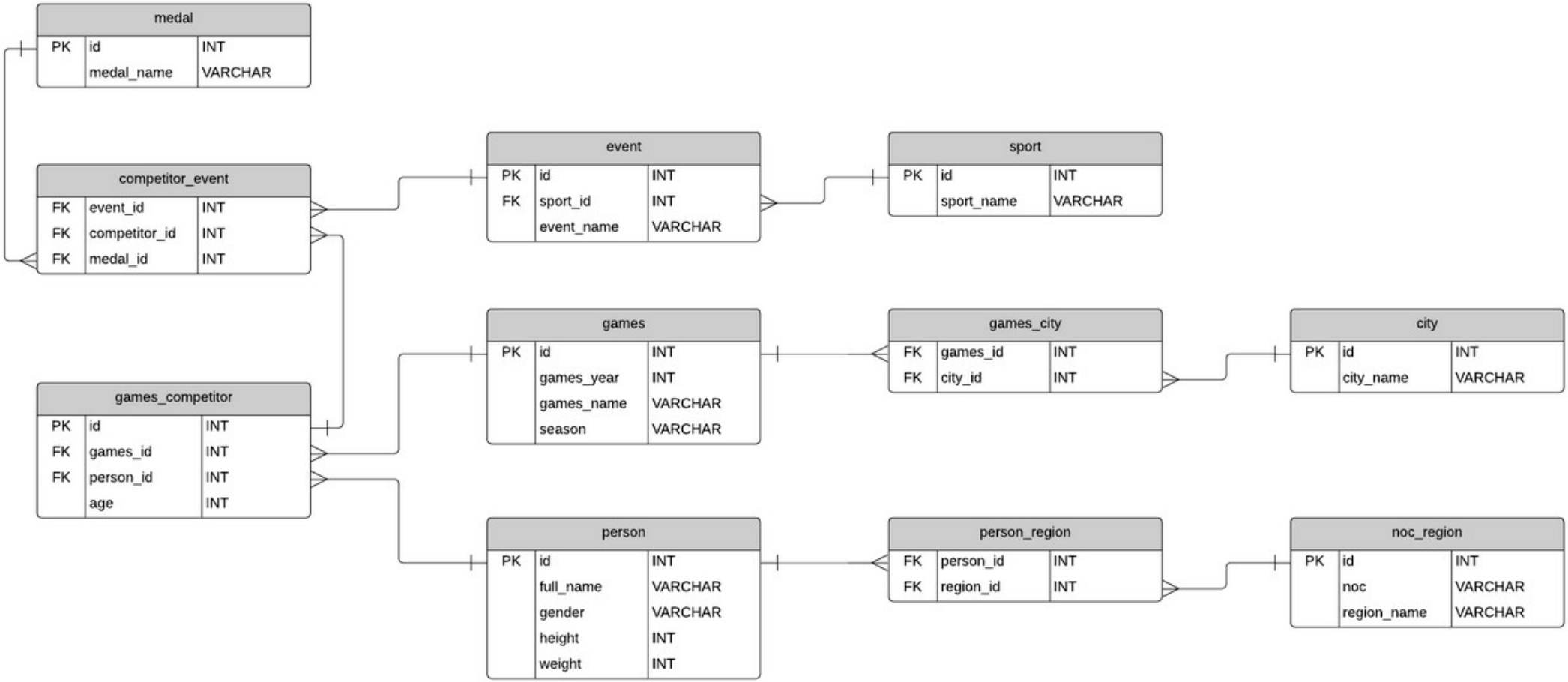
```
data.head(3)
```

	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

Step 1: Preparing for the Proposal

ERD:

After a first exploration of data, I suggest the modelization shown by the Entity-Relationship Diagram bellow (an olympics game could be in more than one city, an athlete could have been participating in different games thus different age)



Step 2: Develop Project Proposal

Description

This project has a goal of exploring the data about Olympic events and competitors and try to find trends or some knowledge. The findings might interested fans of Olympic games, sport commentators, trainers and news firms.

Questions to answer :

Q1 : is the country known for a sport or where this last is so popular, the most medal wining?

- Football : What country is most medal-winning? is it Brazil or German?
- Basketball : What country is most medal-winning? is it the USA?

Q2 : Is there any relation between organizing the Olympic and being the country most medal-wining?

Q3 : Does the age of the competitor affect the result?

Step 2: Develop Project Proposal

Hypothesis

H1 : The country where a sport is popular must be within the most-medal-winning countries
- Brazil, German in football events
- USA in basketball
- Canada and Russia in Hockey

H2 : The country organizing the Olympic is the he most medal-wining

H3 : The gold winners must be around the average age of all competitors

Approach

A1 : Rank countries wining the olympics for two sports : football, Basketball and Hockey
- Consider the total number of medals not just Gold which means the country is at least in the 3 finalists

A2 : See if yes or no the country holding the olympics is within the 3 most medal winning countries

A3 : See if yes or no the age of winner is related to the average of all competitors in the same event
(I'll neglect the missing values)

Milestone 2

Data Understanding (first questions):

For this phase, to understand data and clean the issues, I will proceed by answering some pop-up questions and discovering in the same time.

How many Summer/Winter Seasons ? What years?

```
print(f"""
There is {len(data[data['Season']=='Summer'].Year.unique())} summer olympic Games.
\nThey are hold in this yers \n{sorted(data[data['Season']=='Summer'].Year.unique())}
      """)
```

```
There is 29 summer olympic Games.

They are hold in this yers
[1896, 1900, 1904, 1906, 1908, 1912, 1920, 1924, 1928, 1932, 1936, 1948, 1952, 1956, 1960, 1964, 1968, 1972, 1976,
1980, 1984, 1988, 1992, 1996, 2000, 2004, 2008, 2012, 2016]
```

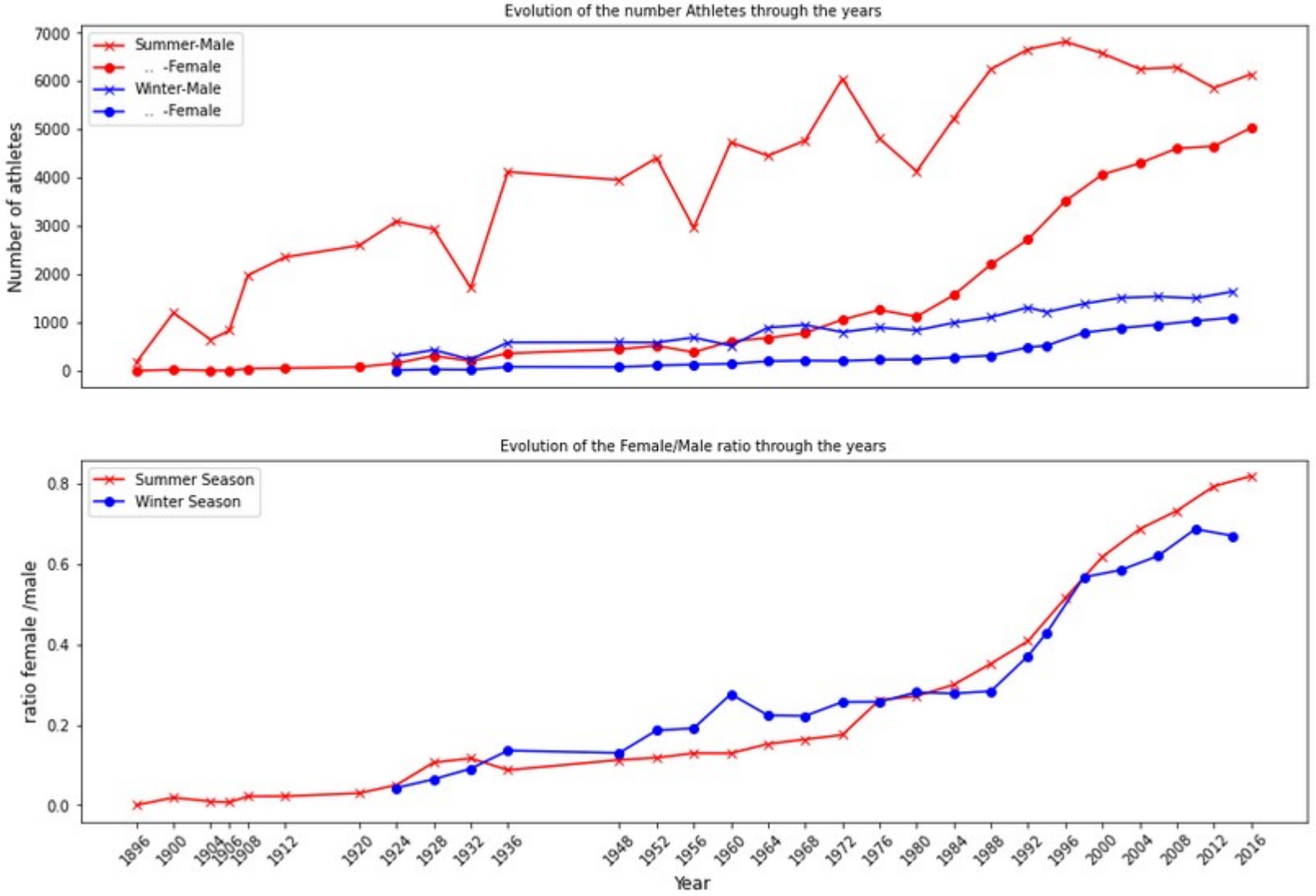
```
print(f"""
There is {len(data[data['Season']=='Winter'].Year.unique())} winter olympic Games.
\nThey are hold in this yers \n{sorted(data[data['Season']=='Winter'].Year.unique())}
      """)
```

```
There is 22 winter olympic Games.

They are hold in this yers
[1924, 1928, 1932, 1936, 1948, 1952, 1956, 1960, 1964, 1968, 1972, 1976, 1980, 1984, 1988, 1992, 1994, 1998, 2002,
2006, 2010, 2014]
```

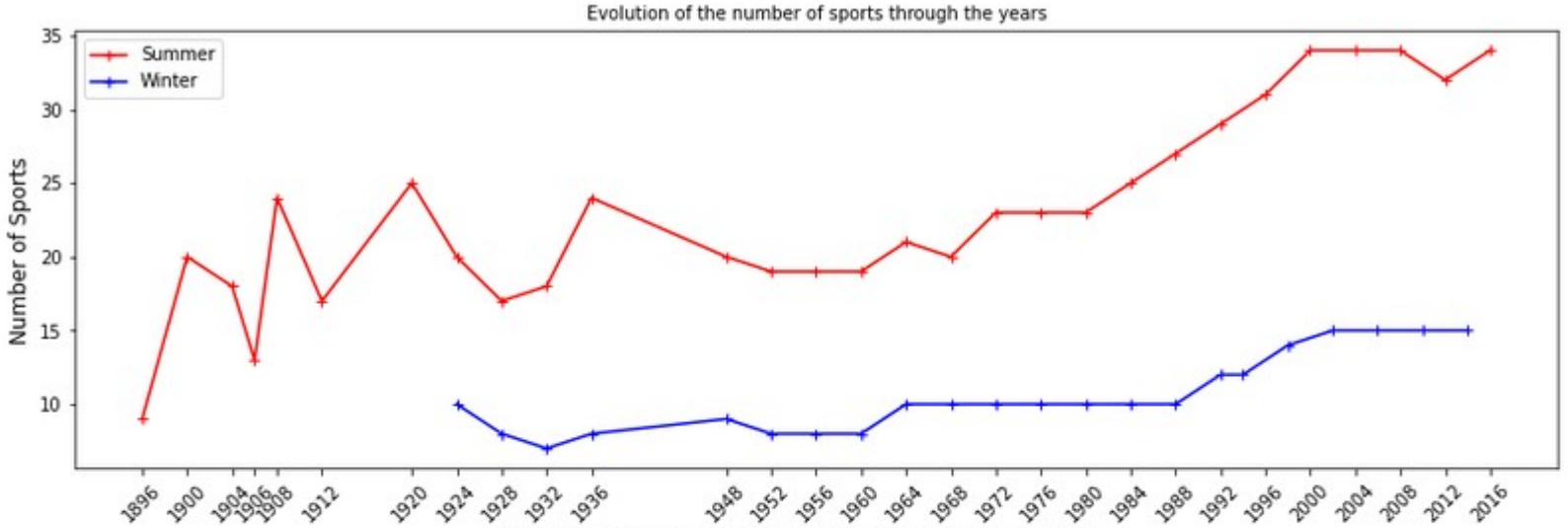
Data Understanding (first questions):

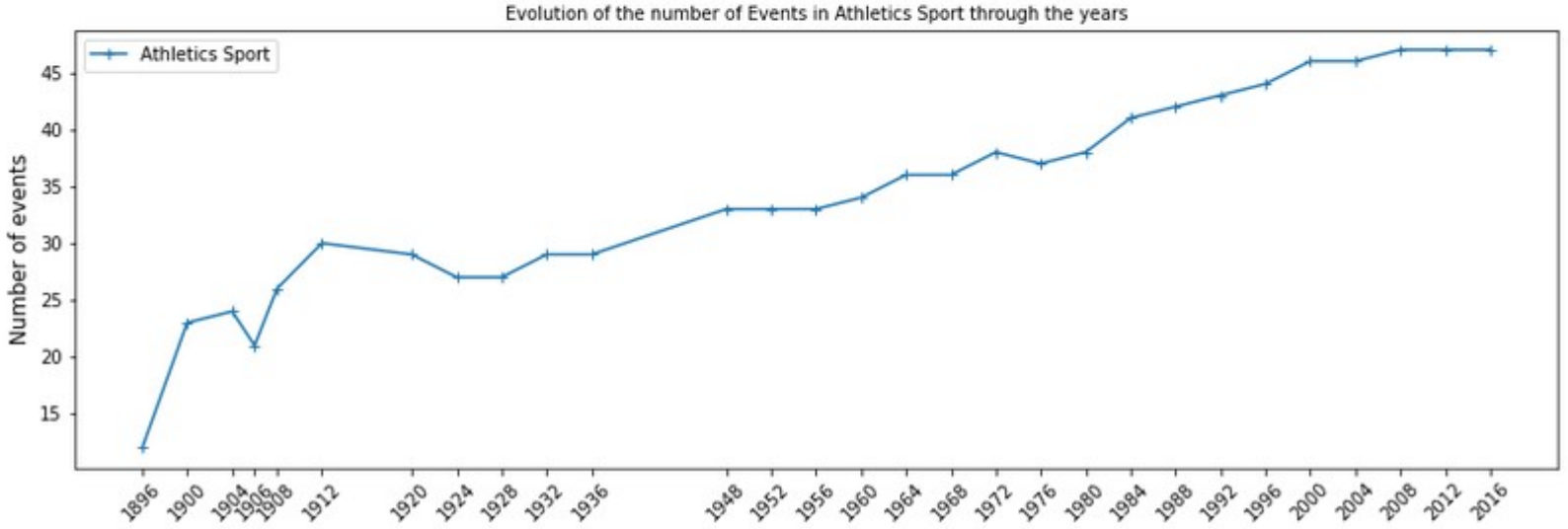
Evolution number of athletes, female/male ratio over the years for each season :



Data Understanding (first questions):

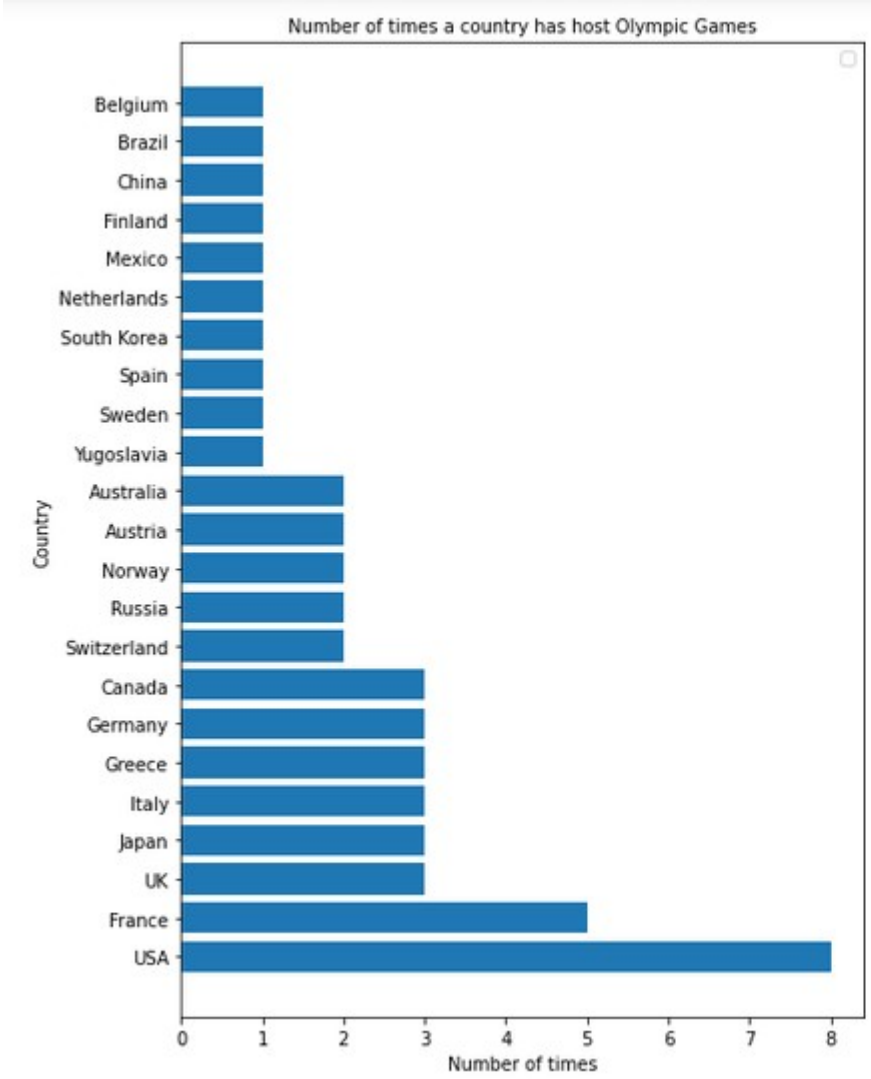
Evolution of the number of sports And the number of events in the “Athletics” sport :

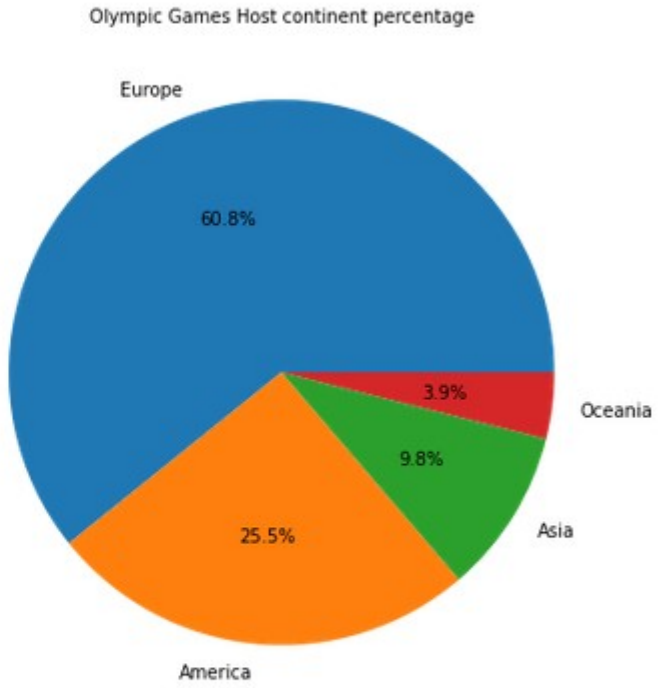




Data Understanding (first questions):

What countries have host the most Olympic Games?

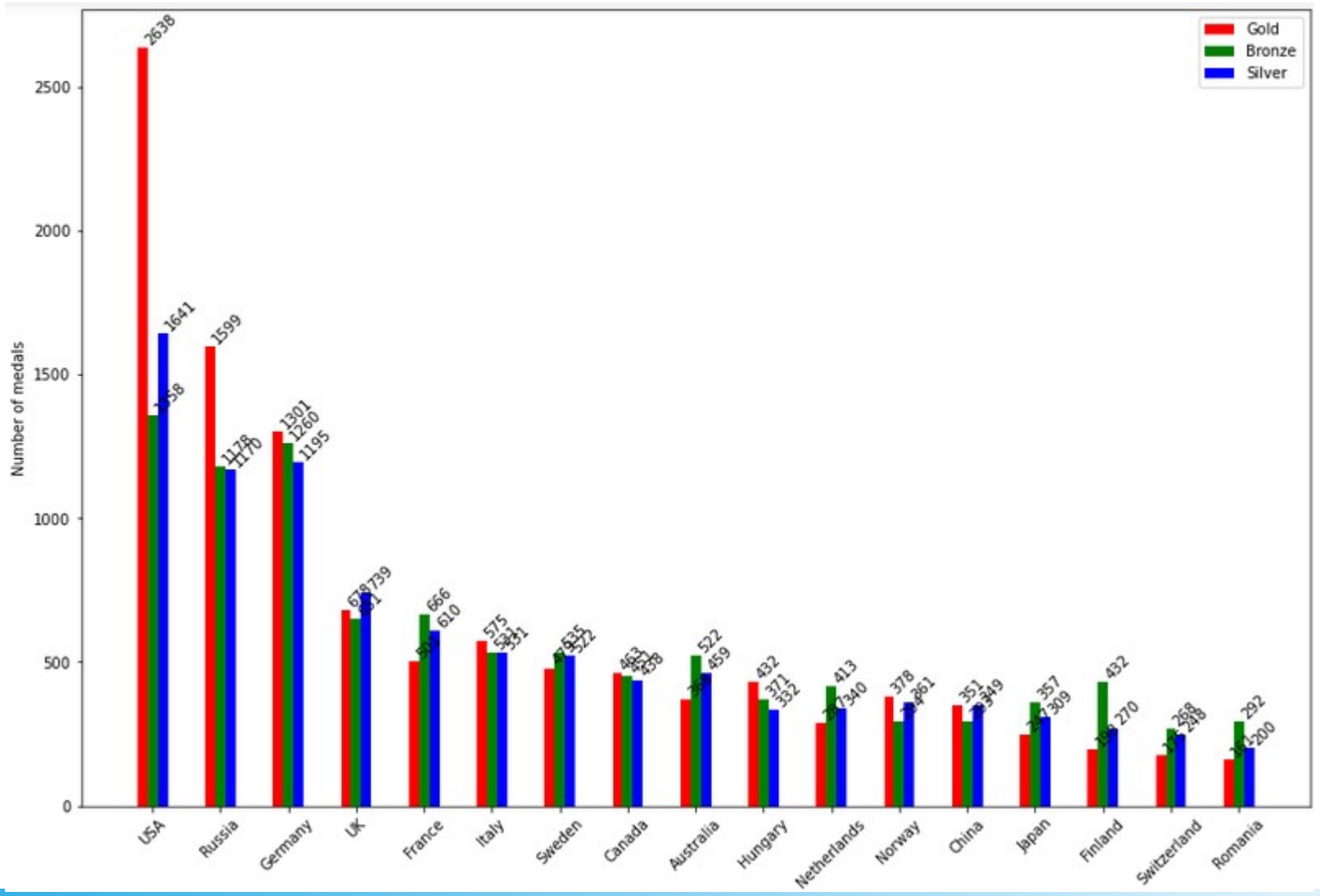




Data Understanding (first questions):

Overall ranking

	Country	Total	Gold	Bronze	Silver
0	USA	5637	2638	1358	1641
1	Russia	3947	1599	1178	1170
2	Germany	3756	1301	1260	1195
3	UK	2068	678	651	739
4	France	1777	501	666	610
5	Italy	1637	575	531	531
6	Sweden	1536	479	535	522
7	Canada	1352	463	451	438
8	Australia	1349	368	522	459
9	Hungary	1135	432	371	332
10	Netherlands	1040	287	413	340
11	Norway	1033	378	294	361
12	China	993	351	293	349
13	Japan	913	247	357	309
14	Finland	900	198	432	270
15	Switzerland	691	175	268	248
16	Romania	653	161	292	200



Data Understanding (Disc. issues):

```
region[region['region'].isnull()]
```

	NOC	region	notes
168	ROT	NaN	Refugee Olympic Team
208	TUV	NaN	Tuvalu
213	UNK	NaN	Unknown

```
region.loc[[168,208,213],'region'] = list(region.loc[[168,208,213],'notes'])
```

```
data[data['NOC']=='UNK']
```

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport
61080	31292	Fritz Eccard	M	nan	nan	nan	Unknown	UNK	1912 Summer	1912	Summer	Stockholm	Art Competitions
130721	65813	A. Laffen	M	nan	nan	nan	Unknown	UNK	1912 Summer	1912	Summer	Stockholm	Art Competitions

```
for frame, group in data.groupby('Games'):
    if len(group['City'].unique()) != 1:
        print(frame, group['City'].unique())
```

```
1956 Summer ['Melbourne' 'Stockholm']
```

```
data[(data['City']=='Stockholm') & (data['Year']==1956)].Sport.unique()
```

```
array(['Equestrianism'], dtype=object)
```

In 1956, Sweden hold just one event "Equestrianism". Then Olympic Games can be held in more than one city

```
data[(data['Event'] == 'Speed Skating Women\'s 500 metres') & (data['Year'] == 2010)].Age.unique()
```

```
array([17., 26., 18., 24., 20., 21., 27., 31., 23., 32., 38., 30., 25.,
       33., 34., 22.])
```

then the competetors in the same event at the same game may have diffrent age

```
# I noticed that for Singapore the NOC used in data is "SGP", wherase 'SIN' is used in the region file.
# So before the merge I'll change it in the region dataframe
region.loc[region[region['NOC']=='SIN'].index,'NOC']= 'SGP'
# I'll drop two rows where the National Olympic Code is UNK (unkown)
data = data.drop(data[data['NOC']=='UNK'].index)
```

Descriptive Stats :

Descriptive stats:

using describe()

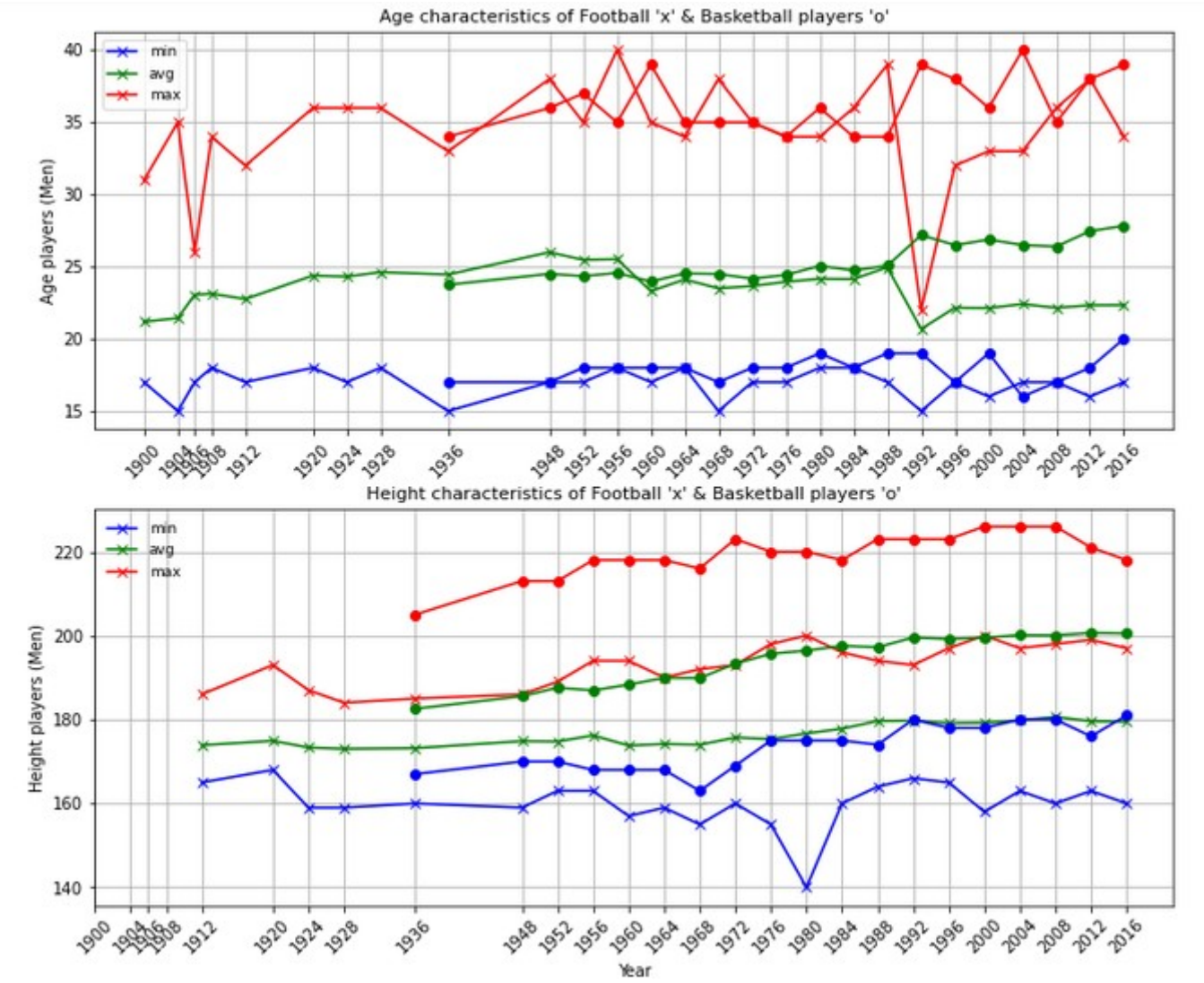
```
data_without_null.describe()
```

	ID	Age	Height	Weight	Year
count	206165.00	206165.00	206165.00	206165.00	206165.00
mean	68616.02	25.06	175.37	70.69	1989.67
std	38996.51	5.48	10.55	14.34	20.13
min	1.00	11.00	127.00	25.00	1896.00
25%	35194.00	21.00	168.00	60.00	1976.00
50%	68629.00	24.00	175.00	70.00	1992.00
75%	102313.00	28.00	183.00	79.00	2006.00
max	135571.00	71.00	226.00	214.00	2016.00

Descriptive Stats :

As a fun of Football, I am interested in a comparison between the age and Height between basketball and football players. I'll calculate the 'min', 'avg', 'max' for these two characteristics.

As we may think, the only difference is in the height and we can see that the average height of basketball players is almost equal to the maximum height we can find in football.



Milestone 3

Finding and Impact on hypothesis

Question 1 :
Does the country known for a sport is the most medal wining?
- Football : What country is most medal-winning? is it Brazil or German?
- Basketball : What country is most medal-winning? is it the USA?

=> I think the hypothesis made is correct. In fact, on three sport i.e. Football, Basketball and Hockey, I find that yes the countries known for these sports are actually the most medal wining (see pictures bellow)

Men's Football

	Country	Total	Gold	Bronze	Silver
0	Germany	6	1	3	2
1	Brazil	6	1	2	3
2	Serbia	5	1	1	3
3	Russia	5	2	3	0
4	Hungary	5	3	1	1
5	Denmark	5	1	1	3
6	UK	4	3	0	1
7	Netherlands	4	0	4	0
8	Argentina	4	2	0	2

Women's Football

Country	Total	Gold	Bronze	Silver
USA	5	4	0	1
Germany	4	1	3	0
Norway	2	1	1	0
Canada	2	0	2	0
Brazil	2	0	0	2
Sweden	1	0	0	1
Japan	1	0	0	1
China	1	0	0	1

Men's Basketball

	Country	Total	Gold	Bronze	Silver
0	USA	18	15	2	1
1	Russia	10	2	4	4
2	Serbia	7	1	1	5
3	Spain	4	0	1	3
4	Lithuania	3	0	3	0
5	Brazil	3	0	3	0
6	Uruguay	2	0	2	0
7	Italy	2	0	0	2
8	France	2	0	0	2
9	Argentina	2	1	1	0

Women's Basketball

Country	Total	Gold	Bronze	Silver
USA	10	8	1	1
Russia	6	3	3	0
Australia	5	0	2	3
Serbia	3	0	2	1
China	2	0	1	1
Bulgaria	2	0	1	1
Brazil	2	0	1	1
Spain	1	0	0	1
ɔuth Korea	1	0	0	1
France	1	0	0	1

Finding and Impact on hypothesis

Question 2:
Is there any relation between organizing the Olympic games and being within the top five countries most medal-wining?

First, I bring a table of the host countries from Wikipedia (see figure). After cleaning it, I joined it with the data on two columns (Year, City)

	City	Year	Host country	Host continent
0	Athina	1896	Greece	Europe
1	Paris	1900	France	Europe
2	St. Louis	1904	USA	America
3	London	1908	UK	Europe
4	Stockholm	1912	Sweden	Europe

```
link = "https://en.wikipedia.org/wiki/List_of_Olympic_Games_host_cities"

host = pd.read_html(link,header=0)[2].loc[0:54,['City.1','Year','Country','Continent']]

host.rename({'City.1':'City','Country':'Host country', 'Continent':'Host continent'}, axis=1, inplace=True)

host.drop([5,15,16,17,18], inplace=True)

host['Host country'].replace(['United States','United Kingdom','Australia\xa0Sweden','Soviet Union','West Germany']
                             ,['USA','UK','Australia','Russia','Germany'], inplace=True)
df=pd.DataFrame([{'City': 'Athens','Year': 1906, 'Host country': 'Greece', 'Host continent':'Europe'}
                ,{'City': 'Stockholm','Year': 1956, 'Host country': 'Sweden', 'Host continent':'Europe'}])

host=host.append(df)
def splitname(row):
    row['City'] = row['City'].split("[")[0]
    return row

host=host.apply(splitname, axis='columns')
host['City'].replace(['Athens','Rome', 'St. Moritz','Antwerp','Turin','MelbourneStockholm','Moscow']
                     ,['Athina','Roma','Sankt Moritz','Antwerpen','Torino','Melbourne','Moskva'], inplace=True)
host
```

Finding and Impact on hypothesis

Then, for each Olympic games I pulled out the first five countries and add a Yes/No column having 1 if the host country is within the 5 top ranked countries. And finally, I calculated the the percentage of countries hosting a the Games and ranked within the top 5.

It is 80% for the summer season and 54,5% for winter. The overall is 69%

	Year	City	Host country	Season	1st	2d	3d	4th	5th	Yes/No
0	1896	Athina	Greece	Summer	Greece	Germany	USA	France	UK	1
21	1956	Stockholm	Sweden	Summer	Germany	UK	Sweden	Italy	Switzerland	1
22	1960	Roma	Italy	Summer	Russia	USA	Germany	Italy	Hungary	1
27	1968	Mexico City	Mexico	Summer	Russia	USA	Germany	Hungary	Japan	0
28	1972	Munich	Germany	Summer	Germany	Russia	USA	Hungary	Japan	1
31	1976	Montreal	Canada	Summer	Russia	Germany	USA	Poland	Romania	0
33	1980	Moskva	Russia	Summer	Russia	Germany	Bulgaria	Romania	Hungary	1

I think those result confirm that the performance of a country is increased when it’s the host.

The different between Summer and winter percentage can be explained first by the fact that winter season are less than summers one and maybe also because the fans are less present in winter than summer.

Finding and Impact on hypothesis

Question 3:
Does the age of the competitor affect the result? I'll see the correlation between the gold winner Age and the Mean of all competitors in the event

Year	City	Host country	Season	Sport	Event	Winner gold age	Competitors avg age
1896	Athina	Greece	Summer	Athletics	Athletics Men's 1,500 metres	22.0	21.916667
1896	Athina	Greece	Summer	Athletics	Athletics Men's 100 metres	21.0	22.696429
1896	Athina	Greece	Summer	Athletics	Athletics Men's 110 metres Hurdles	23.0	22.500000
1896	Athina	Greece	Summer	Athletics	Athletics Men's 400 metres	21.0	21.833333
1896	Athina	Greece	Summer	Athletics	Athletics Men's 800 metres	22.0	21.500000
...	...	...	...	...	...	...	...
2016	Rio de Janeiro	Brazil	Summer	Wrestling	Wrestling Women's Flyweight, Freestyle	22.0	24.285714
2016	Rio de Janeiro	Brazil	Summer	Wrestling	Wrestling Women's Heavyweight, Freestyle	27.0	29.464286
2016	Rio de Janeiro	Brazil	Summer	Wrestling	Wrestling Women's Light-Heavyweight, Freestyle	21.0	24.321429
2016	Rio de Janeiro	Brazil	Summer	Wrestling	Wrestling Women's Lightweight, Freestyle	32.0	26.515625
2016	Rio de Janeiro	Brazil	Summer	Wrestling	Wrestling Women's Middleweight, Freestyle	21.0	24.683333

The Pearson correlation output is 0.6 with a p-value of 0. Meaning statistically highly significant that there is a positive correlation and that not due to hazard.

Finding and Impact on hypothesis

Question 4 :
Is the number of athletes sent by a country related to the number of Gold Medals?

	Year	Season	Country	Total participation	Number of distinct athletes	Gold	Bronze	Silver	None
0	1908	Summer	UK	972	735	147	90	131	604
1	1972	Summer	Germany	1041	721	74	96	83	788
2	1900	Summer	France	1071	720	52	82	101	836
3	1996	Summer	USA	839	648	159	52	48	580
4	2008	Summer	China	777	633	74	57	53	593
5	2000	Summer	Australia	788	617	60	54	69	605
6	1988	Summer	Germany	918	606	111	94	91	622

```
The pearson correlation between the Number of distinct athletes and the number of gold medals:

        -All Summer's Season data: (0.7494358949127844, 0.0)
        -All Winter's Season data: (0.6275118117588578, 1.9427884162486226e-112)

        -for USA in Summer's Season: (0.8446883438890789, 1.5856210123563883e-08)
        -for USA in winter's Season: (0.385038638767634, 0.07680623143865707)
        -for France in Summer's Season : (0.7067980802235567, 1.8205353211936 98e-05)
        -for France in winter's Season: (0.6120805763592412, 0.0024654333057815446)
```

Conclusion

Insight Discovered

- The country known for a sport is well ranked within the top winners
- Hosting the Olympic games increase the performance of the country
- the age of competitors affect the result

  the athletes around the average age of all competitors are highly likely to win Gold

Further actions:

- Continue analysis to answer questions related to sport businesses