Capstone Project

STEP 1 :

- Choice of client: SportsStats. The dataset is divided into two basic CSV files that are easy to work with.

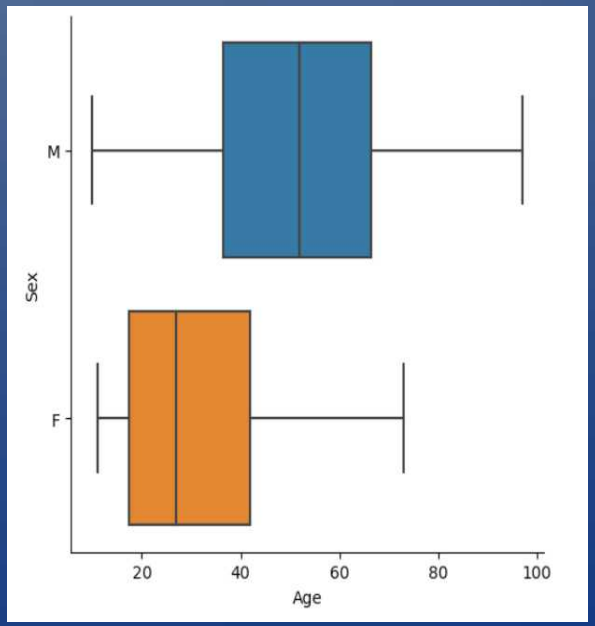
- Importing Data: Data analysis is being done through the use of Jupyter Notebooks. CSV files were uploaded to the notebooks for easy work, followed by pandas reading, initially as csv in python language for better analysis. Nulls were kept not to infere with the data.

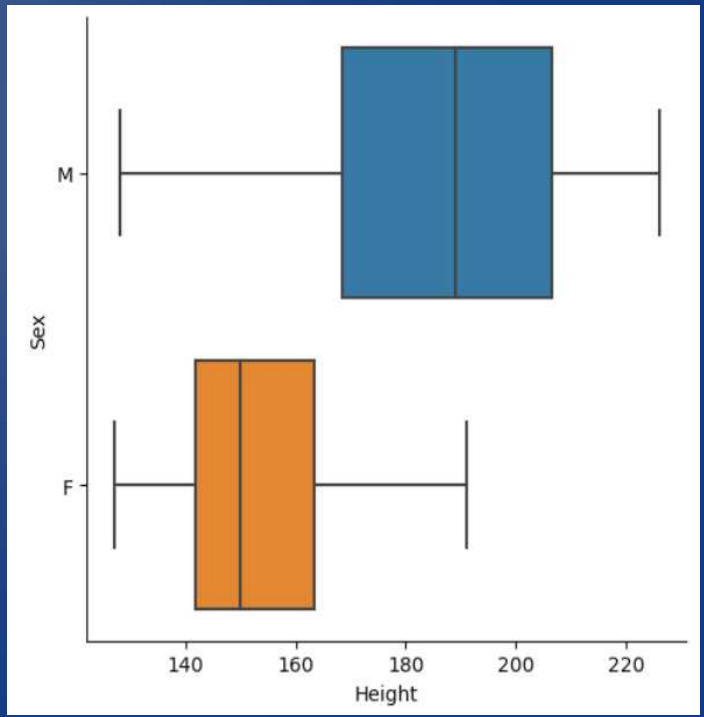
STEP 1:
Screenshots of EDA:

	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
5	5	Christine Jacoba Aaftink	F	21.0	185.0	82.0	Netherlands	NED	1988 Winter	1988	Winter	Calgary	Speed Skating	Speed Skating Women's 1,000 metres	NaN
6	5	Christine Jacoba Aaftink	F	25.0	185.0	82.0	Netherlands	NED	1992 Winter	1992	Winter	Albertville	Speed Skating	Speed Skating Women's 500 metres	NaN
7	5	Christine Jacoba Aaftink	F	25.0	185.0	82.0	Netherlands	NED	1992 Winter	1992	Winter	Albertville	Speed Skating	Speed Skating Women's 1,000 metres	NaN
8	5	Christine Jacoba Aaftink	F	27.0	185.0	82.0	Netherlands	NED	1994 Winter	1994	Winter	Lillehammer	Speed Skating	Speed Skating Women's 500 metres	NaN
9	5	Christine Jacoba Aaftink	F	27.0	185.0	82.0	Netherlands	NED	1994 Winter	1994	Winter	Lillehammer	Speed Skating	Speed Skating Women's 1,000 metres	NaN

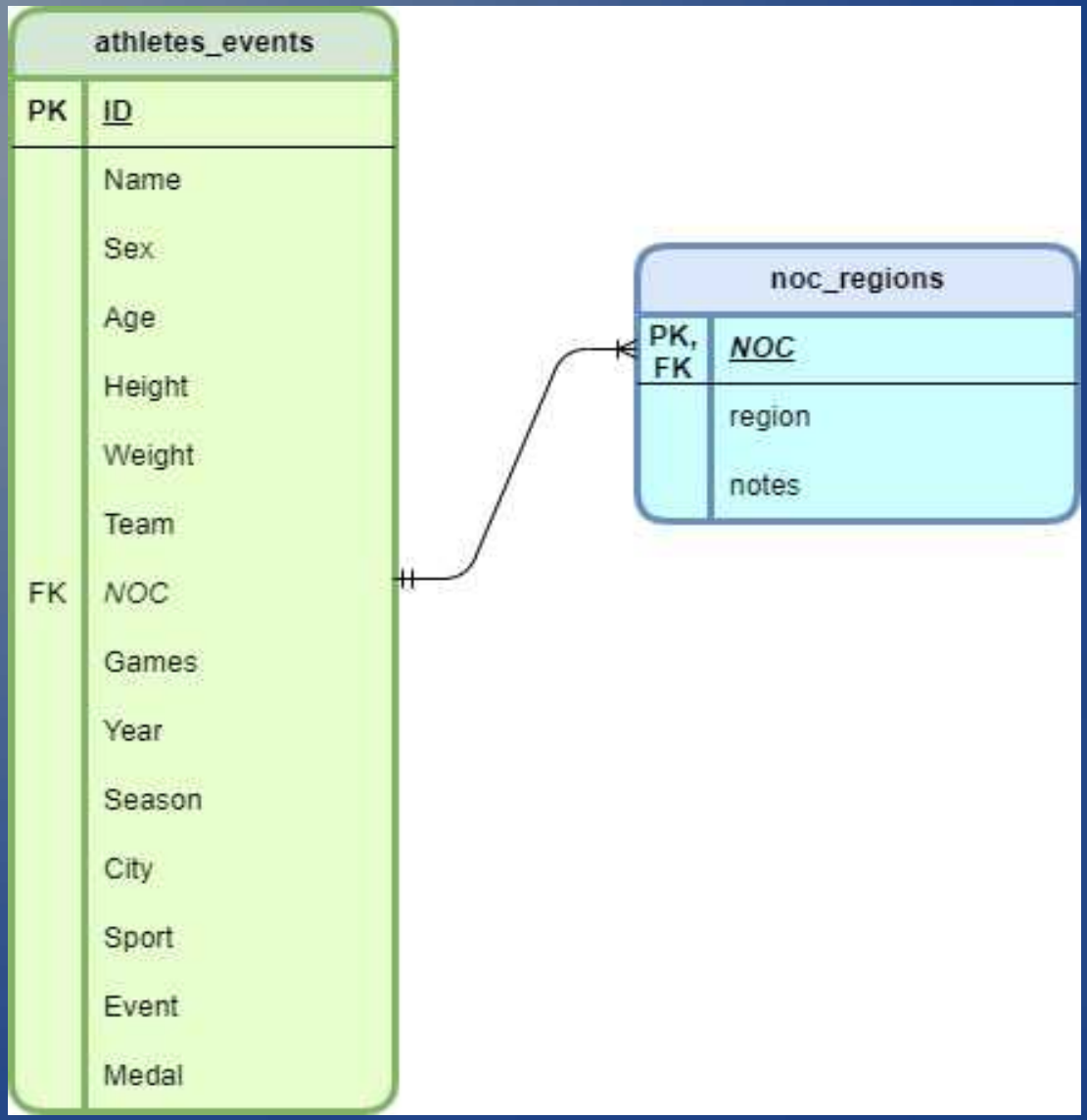
	ID	Age	Height	Weight	Year
count	271116.000000	261642.000000	210945.000000	208241.000000	271116.000000
mean	68248.954396	25.556898	175.338970	70.702393	1978.378480
std	39022.286345	6.393561	10.518462	14.348020	29.877632
min	1.000000	10.000000	127.000000	25.000000	1896.000000
25%	34643.000000	21.000000	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

```
<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
```





ERD PROPOSAL:



STEP 2:

Description:

This initial step of the analysis consisted in understanding better the consistency of the data. A lot more EDA was done looking for both numeric and categorical values and patterns in them. It was possible to see the variation in age, weight and height, with max, min, mean or median values. A far superior count of male athletes is shown compared to female athletes, accompanied by a surprisingly different profile by gender. The data can provide interesting information regarding the profile of athletes, including by sport as well as the number of medals earned by athletes or countries in each game. Anyone interesting in this information can look into this data.

STEP 2:

Questions:

Which games are being represented?

What is the time range of the data?

What are the sports best represented?

How does gender affect other values?

How does representation by a country increase the probability of winning?

Hypotheses and Approach:

So far the approach is to investigate correlation between the variables and see how one can affect the other. Due to the large size of the data, aggregations have been done for the purpose of better understanding the distribution, followed by several different attempts of plotting looking for insights. Investigation should proceed taking into account that athletic women have shown to be shorter and thinner than average, while men are much taller with a greater body mass than average. Pointers should also move forward as the relationship between their profiles and the sport they play become more evident. It should also be interesting to find out how representation increases the chance of winning, that is, more athletes by a country means the country wins more.