Project - José Martínez de la Flor

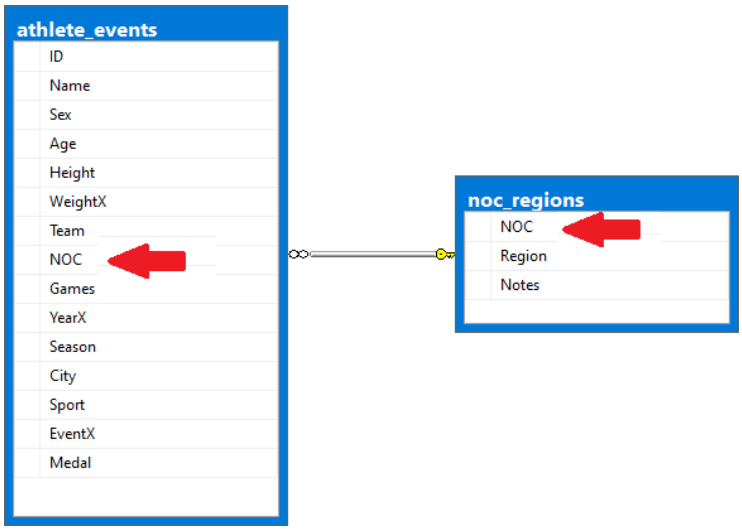
-Which client/dataset did you select and why?
SportsStats; I found it the most interesting at first glance in terms of its data.

-Describe the steps you took to import and clean the data
I had to remove some apostrophes and other symbols that had some names to facilitate the loading of data, I also had to change some values in weight and height from "NA" to "0" to define said column for integer values.

-Perform initial exploration of data and provide some screenshots or display some stats of the data you are looking at
As can be seen in the image (enlarge the view to see it better) the "ID" value seems to be being used as user data and not how this value would normally be used, since if it were, it would be unique and would not be repeated as seen in the data.

	ID	Name	Sex	Age	Height	WeightX	Team	NOC	Games	YearX	Season	City	Sport	EventX	Medal
1	1	A Dijiang	M	24	180	80	China	CHN	1992 Summer	1992	Summer	Barcelona	Basketball	Basketball Mens Basketball	NA
2	2	A Lamusi	M	23	170	60	China	CHN	2012 Summer	2012	Summer	London	Judo	Judo Mens Extra-Lightweight	NA
3	3	Gunnar Nielsen Aaby	M	24	0	0	Denmark	DEN	1920 Summer	1920	Summer	Antwerpen	Football	Football Mens Football	NA
4	4	Edgar Lindenau Aabye	M	34	0	0	Denmark/Sweden	DEN	1900 Summer	1900	Summer	Paris	Tug-Of-War	Tug-Of-War Mens Tug-Of-War	Gold
5	5	Christine Jacoba Aaftink	F	21	185	82	Netherlands	NED	1988 Winter	1988	Winter	Calgary	Speed Skating	Speed Skating Womens 500 metres	NA
6	5	Christine Jacoba Aaftink	F	21	185	82	Netherlands	NED	1988 Winter	1988	Winter	Calgary	Speed Skating	Speed Skating Womens 1,000 metres	NA
7	5	Christine Jacoba Aaftink	F	25	185	82	Netherlands	NED	1992 Winter	1992	Winter	Albertville	Speed Skating	Speed Skating Womens 500 metres	NA
8	5	Christine Jacoba Aaftink	F	25	185	82	Netherlands	NED	1992 Winter	1992	Winter	Albertville	Speed Skating	Speed Skating Womens 1,000 metres	NA
9	5	Christine Jacoba Aaftink	F	27	185	82	Netherlands	NED	1994 Winter	1994	Winter	Lillehammer	Speed Skating	Speed Skating Womens 500 metres	NA
10	5	Christine Jacoba Aaftink	F	27	185	82	Netherlands	NED	1994 Winter	1994	Winter	Lillehammer	Speed Skating	Speed Skating Womens 1,000 metres	NA
11	6	Per Knut Aaland	M	31	188	75	United States	USA	1992 Winter	1992	Winter	Albertville	Cross Country Skiing	Cross Country Skiing Mens 10 kilometres	NA
12	6	Per Knut Aaland	M	31	188	75	United States	USA	1992 Winter	1992	Winter	Albertville	Cross Country Skiing	Cross Country Skiing Mens 50 kilometres	NA
13	6	Per Knut Aaland	M	31	188	75	United States	USA	1992 Winter	1992	Winter	Albertville	Cross Country Skiing	Cross Country Skiing Mens 10/15 kilometres Pursuit	NA
14	6	Per Knut Aaland	M	31	188	75	United States	USA	1992 Winter	1992	Winter	Albertville	Cross Country Skiing	Cross Country Skiing Mens 4 x 10 kilometres Relay	NA
15	6	Per Knut Aaland	M	33	188	75	United States	USA	1994 Winter	1994	Winter	Lillehammer	Cross Country Skiing	Cross Country Skiing Mens 10 kilometres	NA
16	6	Per Knut Aaland	M	33	188	75	United States	USA	1994 Winter	1994	Winter	Lillehammer	Cross Country Skiing	Cross Country Skiing Mens 30 kilometres	NA
17	6	Per Knut Aaland	M	33	188	75	United States	USA	1994 Winter	1994	Winter	Lillehammer	Cross Country Skiing	Cross Country Skiing Mens 10/15 kilometres Pursuit	NA
18	6	Per Knut Aaland	M	33	188	75	United States	USA	1994 Winter	1994	Winter	Lillehammer	Cross Country Skiing	Cross Country Skiing Mens 4 x 10 kilometres Relay	NA
19	7	John Aalberg	M	31	183	72	United States	USA	1992 Winter	1992	Winter	Albertville	Cross Country Skiing	Cross Country Skiing Mens 10 kilometres	NA
20	7	John Aalberg	M	31	183	72	United States	USA	1992 Winter	1992	Winter	Albertville	Cross Country Skiing	Cross Country Skiing Mens 50 kilometres	NA

-Create an ERD or proposed ERD to show the relationships of the data you are exploring
The two tables that the databases would have been connected in the following way, in this case only the connection has been drawn to facilitate the sample process due to an error in the data that will be explained later.



-Description
In this project, a database is analyzed to demonstrate that writing repeated data and not using the "join" method can have results such as slowing down when calling data, thus affecting all the queries made by the organization that owns the data, which would mean in slowness in carrying out processes and increase in cost due to slow processes.

-Questions
*Could it be called the same cat but in a more efficient and faster way?
*What is the importance of correctly filling in information and not writing repeated information when filling in the columns?

-Hypothesis
*The "Games" column seems to be just repeated data that comes from the combination of the "Year" and "Season" columns.
*The data in the "Sport" column seems to be repeated at the beginning of the "Event" column.
*The data of the "NOC" and "City" column should not be in this table as it is, they should be as an integer value that would be used to obtain their respective data from the "noc_regions" table ("join" method), in the case of "City" should have another separate table for this, all in order to achieve data loading efficiency

-Approach
Use string verification methods to determine if I am correct or not with respect to the first two hypotheses, with respect to the last one, although it is obvious that I am correct, this will be demonstrated by analyzing the load times when trying to call the data before and after the changes that will be made to validate the hypothesis.