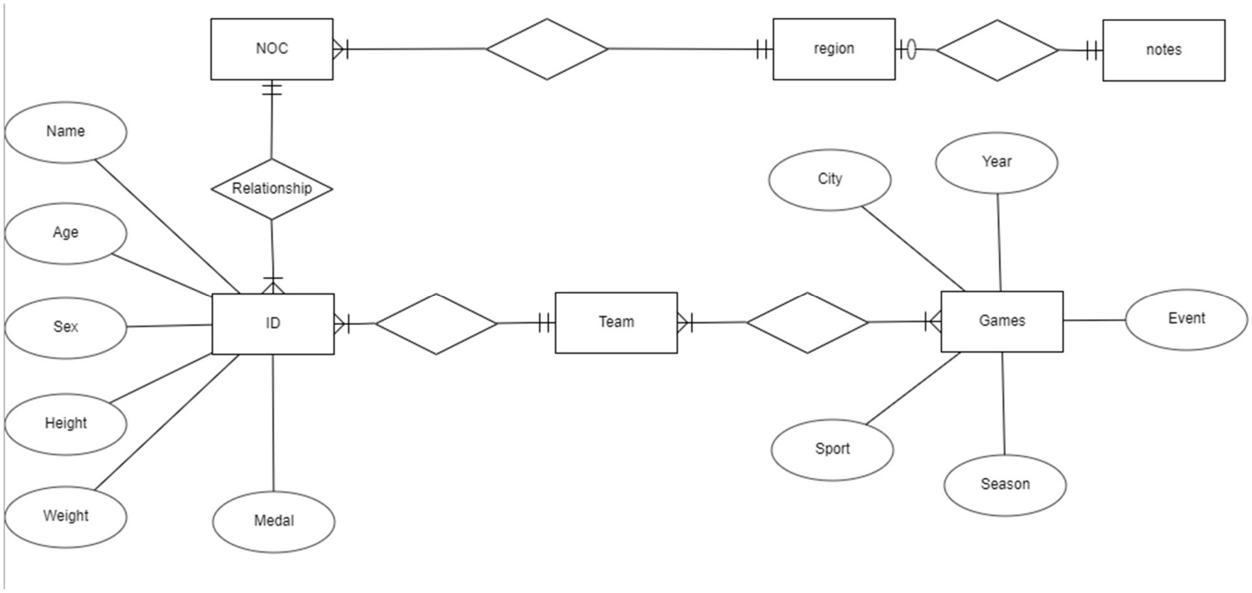
Week 1 Analysis

Importing Data Set

I chose the Sports Stats data set because I have a strong interest in sports analysis and looking at sports-related data. I used the pandas and pandasql to store the data as a MySQL dataset. Since the dataset contains NaN values, I decided to not clean it.



Description:

My project targets observing the performance in sports and analyzing them based on age and medals. Getting to know more insights on the data involves the range of age, as well as the events timelines and the medals athletes receive at certain ages. This analysis helps coaches and athletes understand for what events the suitable age range would be so that their country could provide more results and accomplishments as well as other physical conditions. My audience includes coaches, trainers, recruiters, and the players who will know when to start training and take notice of their own physical capacities and limits for their events. With this knowledge, athletes could advance in a more competitive, yet more energetic and improved performance which provide a lot more joy for many people.

Questions:

What is the youngest ages for each events?

What was the medal distribution for the ages?

What are all the athlete events conducted when the athletes receive their medals?

What were the average ages of each country’s athletes for men and women?

What was the medal distribution for the summer and winter sports?

Hypothesis:

Young men in their 30s with have the most medals

People with more age have more medals as they have more experience and attended more events likely

The age of medals being received is decreasing more than it was in the past

Approach:

I will be observing medal count as well as the Age field. Observing the number of medals earned in comparison to the events attended will also be important as it is more likely for athletes to have a chance at getting medals if they attend more events so it is important to also find a ratio for the total number of medals per total events attended for each athlete. I want to observe the number of medals owned and how old the athletes are along with their participation in the events. What is important is to also take note of when the athletes earned their medal which will be crucial in comparing around what age is the best time for men and women to perform their best in their individual sports. I will take an average from the number of event appearances as well as which seasonal events these athletes received their awards

In [6]:

```
import pandas as pd
from pandasql import sqldf
pysqldf = lambda q: sqldf(q, globals())

ath_csv = pd.read_csv('/Users/richa/SportsStats/athlete_events.csv/athlete_events.csv
noc_csv = pd.read_csv('/Users/richa/SportsStats/noc_regions.csv')
```

In [7]:

```
ath_csv = pd.read_csv('/Users/richa/SportsStats/athlete_events.csv/athlete_events.csv
noc_csv = pd.read_csv('/Users/richa/SportsStats/noc_regions.csv')
```

In [8]:

```
ath_csv.info()
```

```
<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
```

In [9]:

```
print("Youngest Age: ", ath_csv.Age.min())
print("Oldest Age: ", ath_csv.Age.max())
```

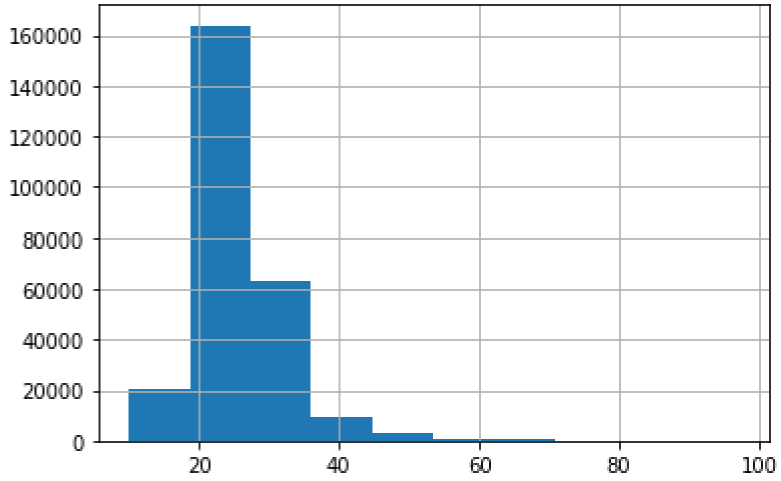
```
Youngest Age:  10.0
Oldest Age:  97.0
```

In [10]:

```
ath_csv.Age.hist()
```

Out[10]:

```
<AxesSubplot:>
```



In [11]:

```
ath_csv.head(25)
```

Out[11]:

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	
0	1	A Dijiang	M	24.0	180.0	80.0	China	CHN	1992 Summer	1992	Summer	Bar
1	2	A Lamusi	M	23.0	170.0	60.0	China	CHN	2012 Summer	2012	Summer	L
2	3	Gunnar Nielsen Aaby	M	24.0	NaN	NaN	Denmark	DEN	1920 Summer	1920	Summer	Antw
3	4	Edgar Lindenau Aabye	M	34.0	NaN	NaN	Denmark/Sweden	DEN	1900 Summer	1900	Summer	
4	5	Christine Jacoba Aaftink	F	21.0	185.0	82.0	Netherlands	NED	1988 Winter	1988	Winter	C
5	5	Christine Jacoba Aaftink	F	21.0	185.0	82.0	Netherlands	NED	1988 Winter	1988	Winter	C
6	5	Christine Jacoba Aaftink	F	25.0	185.0	82.0	Netherlands	NED	1992 Winter	1992	Winter	Alb
7	5	Christine Jacoba Aaftink	F	25.0	185.0	82.0	Netherlands	NED	1992 Winter	1992	Winter	Alb
8	5	Christine Jacoba Aaftink	F	27.0	185.0	82.0	Netherlands	NED	1994 Winter	1994	Winter	Lilleha
9	5	Christine Jacoba Aaftink	F	27.0	185.0	82.0	Netherlands	NED	1994 Winter	1994	Winter	Lilleha
10	6	Per Knut Aaland	M	31.0	188.0	75.0	United States	USA	1992 Winter	1992	Winter	Alb
11	6	Per Knut Aaland	M	31.0	188.0	75.0	United States	USA	1992 Winter	1992	Winter	Alb

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	
12	6	Per Knut Aaland	M	31.0	188.0	75.0	United States	USA	1992 Winter	1992	Winter	Alb
13	6	Per Knut Aaland	M	31.0	188.0	75.0	United States	USA	1992 Winter	1992	Winter	Alb
14	6	Per Knut Aaland	M	33.0	188.0	75.0	United States	USA	1994 Winter	1994	Winter	Lilleha
15	6	Per Knut Aaland	M	33.0	188.0	75.0	United States	USA	1994 Winter	1994	Winter	Lilleha
16	6	Per Knut Aaland	M	33.0	188.0	75.0	United States	USA	1994 Winter	1994	Winter	Lilleha
17	6	Per Knut Aaland	M	33.0	188.0	75.0	United States	USA	1994 Winter	1994	Winter	Lilleha
18	7	John Aalberg	M	31.0	183.0	72.0	United States	USA	1992 Winter	1992	Winter	Alb
19	7	John Aalberg	M	31.0	183.0	72.0	United States	USA	1992 Winter	1992	Winter	Alb

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	
20	7	John Aalberg	M	31.0	183.0	72.0	United States	USA	1992 Winter	1992	Winter	Alb
21	7	John Aalberg	M	31.0	183.0	72.0	United States	USA	1992 Winter	1992	Winter	Alb
22	7	John Aalberg	M	33.0	183.0	72.0	United States	USA	1994 Winter	1994	Winter	Lilleha
23	7	John Aalberg	M	33.0	183.0	72.0	United States	USA	1994 Winter	1994	Winter	Lilleha
24	7	John Aalberg	M	33.0	183.0	72.0	United States	USA	1994 Winter	1994	Winter	Lilleha

In [12]:

```
ath_csv.tail(25)
```

Out[12]:

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Se
271091	135558	ukasz Tomasz ygado	M	32.0	200.0	89.0	Poland	POL	2012 Summer	2012	Su
271092	135559	Pawe Jan Zygmunt	M	21.0	182.0	79.0	Poland	POL	1994 Winter	1994	W
271093	135559	Pawe Jan Zygmunt	M	21.0	182.0	79.0	Poland	POL	1994 Winter	1994	W
271094	135559	Pawe Jan Zygmunt	M	25.0	182.0	79.0	Poland	POL	1998 Winter	1998	W
271095	135559	Pawe Jan Zygmunt	M	25.0	182.0	79.0	Poland	POL	1998 Winter	1998	W
271096	135559	Pawe Jan Zygmunt	M	29.0	182.0	79.0	Poland	POL	2002 Winter	2002	W
271097	135559	Pawe Jan Zygmunt	M	29.0	182.0	79.0	Poland	POL	2002 Winter	2002	W
271098	135559	Pawe Jan Zygmunt	M	33.0	182.0	79.0	Poland	POL	2006 Winter	2006	W
271099	135560	Stavroula Zygouri	F	36.0	171.0	63.0	Greece	GRE	2004 Summer	2004	Su
271100	135561	Frantiek Zyka	M	26.0	NaN	NaN	Czechoslovakia	TCH	1928 Summer	1928	Su
271101	135562	Milan Zyka	M	24.0	173.0	68.0	Czechoslovakia	TCH	1972 Summer	1972	Su
271102	135563	Olesya Nikolayevna Zykina	F	19.0	171.0	64.0	Russia	RUS	2000 Summer	2000	Su
271103	135563	Olesya Nikolayevna Zykina	F	23.0	171.0	64.0	Russia	RUS	2004 Summer	2004	Su
271104	135564	Yevgeny Aleksandrovich Zykov	M	22.0	172.0	65.0	Russia-1	RUS	2002 Winter	2002	W
271105	135565	Fernando scar Zylberberg	M	23.0	168.0	76.0	Argentina	ARG	2000 Summer	2000	Su

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Se
271106	135565	Fernando scar Zylberberg	M	27.0	168.0	76.0	Argentina	ARG	2004 Summer	2004	Su
271107	135566	James Francis "Jim" Zylker	M	21.0	175.0	75.0	United States	USA	1972 Summer	1972	Su
271108	135567	Aleksandr Viktorovich Zyuzin	M	24.0	183.0	72.0	Russia	RUS	2000 Summer	2000	Su
271109	135567	Aleksandr Viktorovich Zyuzin	M	28.0	183.0	72.0	Russia	RUS	2004 Summer	2004	Su
271110	135568	Olga Igorevna Zyuzkova	F	33.0	171.0	69.0	Belarus	BLR	2016 Summer	2016	Su
271111	135569	Andrzej ya	M	29.0	179.0	89.0	Poland-1	POL	1976 Winter	1976	W
271112	135570	Piotr ya	M	27.0	176.0	59.0	Poland	POL	2014 Winter	2014	W
271113	135570	Piotr ya	M	27.0	176.0	59.0	Poland	POL	2014 Winter	2014	W
271114	135571	Tomasz Ireneusz ya	M	30.0	185.0	96.0	Poland	POL	1998 Winter	1998	W
271115	135571	Tomasz Ireneusz ya	M	34.0	185.0	96.0	Poland	POL	2002 Winter	2002	W

In [13]:

```
pd.set_option("display.max_colwidth", None)
ath_csv.head(25)
```

Out[13]:

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	
0	1	A Dijiang	M	24.0	180.0	80.0	China	CHN	1992 Summer	1992	Summer	Bar
1	2	A Lamusi	M	23.0	170.0	60.0	China	CHN	2012 Summer	2012	Summer	L
2	3	Gunnar Nielsen Aaby	M	24.0	NaN	NaN	Denmark	DEN	1920 Summer	1920	Summer	Antw
3	4	Edgar Lindenau Aabye	M	34.0	NaN	NaN	Denmark/Sweden	DEN	1900 Summer	1900	Summer	
4	5	Christine Jacoba Aaftink	F	21.0	185.0	82.0	Netherlands	NED	1988 Winter	1988	Winter	C
5	5	Christine Jacoba Aaftink	F	21.0	185.0	82.0	Netherlands	NED	1988 Winter	1988	Winter	C
6	5	Christine Jacoba Aaftink	F	25.0	185.0	82.0	Netherlands	NED	1992 Winter	1992	Winter	Alb
7	5	Christine Jacoba Aaftink	F	25.0	185.0	82.0	Netherlands	NED	1992 Winter	1992	Winter	Alb
8	5	Christine Jacoba Aaftink	F	27.0	185.0	82.0	Netherlands	NED	1994 Winter	1994	Winter	Lilleha
9	5	Christine Jacoba Aaftink	F	27.0	185.0	82.0	Netherlands	NED	1994 Winter	1994	Winter	Lilleha
10	6	Per Knut Aaland	M	31.0	188.0	75.0	United States	USA	1992 Winter	1992	Winter	Alb
11	6	Per Knut Aaland	M	31.0	188.0	75.0	United States	USA	1992 Winter	1992	Winter	Alb

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	
12	6	Per Knut Aaland	M	31.0	188.0	75.0	United States	USA	1992 Winter	1992	Winter	Alb
13	6	Per Knut Aaland	M	31.0	188.0	75.0	United States	USA	1992 Winter	1992	Winter	Alb
14	6	Per Knut Aaland	M	33.0	188.0	75.0	United States	USA	1994 Winter	1994	Winter	Lilleha
15	6	Per Knut Aaland	M	33.0	188.0	75.0	United States	USA	1994 Winter	1994	Winter	Lilleha
16	6	Per Knut Aaland	M	33.0	188.0	75.0	United States	USA	1994 Winter	1994	Winter	Lilleha
17	6	Per Knut Aaland	M	33.0	188.0	75.0	United States	USA	1994 Winter	1994	Winter	Lilleha
18	7	John Aalberg	M	31.0	183.0	72.0	United States	USA	1992 Winter	1992	Winter	Alb
19	7	John Aalberg	M	31.0	183.0	72.0	United States	USA	1992 Winter	1992	Winter	Alb

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	
20	7	John Aalberg	M	31.0	183.0	72.0	United States	USA	1992 Winter	1992	Winter	Alb
21	7	John Aalberg	M	31.0	183.0	72.0	United States	USA	1992 Winter	1992	Winter	Alb
22	7	John Aalberg	M	33.0	183.0	72.0	United States	USA	1994 Winter	1994	Winter	Lilleha
23	7	John Aalberg	M	33.0	183.0	72.0	United States	USA	1994 Winter	1994	Winter	Lilleha
24	7	John Aalberg	M	33.0	183.0	72.0	United States	USA	1994 Winter	1994	Winter	Lilleha

In [14]:

```
from pandasql import sqldf
pysqldf = lambda q: sqldf(q, globals())
```

In [15]:

```
pysqldf("SELECT * FROM ath_csv;")
```

Out[15]:

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Seasc
0	1	A Dijiang	M	24.0	180.0	80.0	China	CHN	1992 Summer	1992	Summ
1	2	A Lamusi	M	23.0	170.0	60.0	China	CHN	2012 Summer	2012	Summ
2	3	Gunnar Nielsen Aaby	M	24.0	NaN	NaN	Denmark	DEN	1920 Summer	1920	Summ
3	4	Edgar Lindenau Aabye	M	34.0	NaN	NaN	Denmark/Sweden	DEN	1900 Summer	1900	Summ
4	5	Christine Jacoba Aaftink	F	21.0	185.0	82.0	Netherlands	NED	1988 Winter	1988	Wint
...	...	...	...	...	...	...	...	...	...	...	
271111	135569	Andrzej ya	M	29.0	179.0	89.0	Poland-1	POL	1976 Winter	1976	Wint
271112	135570	Piotr ya	M	27.0	176.0	59.0	Poland	POL	2014 Winter	2014	Wint
271113	135570	Piotr ya	M	27.0	176.0	59.0	Poland	POL	2014 Winter	2014	Wint
271114	135571	Tomasz Ireneusz ya	M	30.0	185.0	96.0	Poland	POL	1998 Winter	1998	Wint
271115	135571	Tomasz Ireneusz ya	M	34.0	185.0	96.0	Poland	POL	2002 Winter	2002	Wint

271116 rows × 15 columns

In [16]:

```
summer_events = pysqldf('''SELECT
                             ID,
                             Name,
                             Sex,
                             Age,
                             Height,
                             Weight,
                             NOC,
                             Year,
                             Sport,
                             Event,
                             Medal
                           FROM
                             ath_csv
                           WHERE
                             Season = "Summer"''')

winter_events = pysqldf('''SELECT
                             ID,
                             Name,
                             Sex,
                             Age,
                             Height,
                             Weight,
                             NOC,
                             Year,
                             Sport,
                             Event,
                             Medal
                           FROM
                             ath_csv
                           WHERE
                             Season = "Winter"''')
```

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]: