1. Which client/dataset did you select and why?
   I chose the athlete_events dataset because I was interested in exploring a new dataset and I find sports to be interesting. The athlete_events dataset contains information about athletes, events, and results from a variety of sports. This dataset could be used to answer questions about individual athletes, teams, or sports in general. For example, I could use the dataset to find out which athlete has won the most gold medals in the Olympics, or which team has won the most Super Bowls. I could also use the dataset to compare the performance of athletes from different countries or to track the progress of a particular athlete over time. I think this dataset has a lot of potential for interesting and informative analysis.

2. Describe the steps you took to import and clean the data.
1. Downloaded the dataset
2. Unzipped files
3. Looked into csv file
4. Copied location
5. Used panda read csv to import data
6. Cleaned Null values

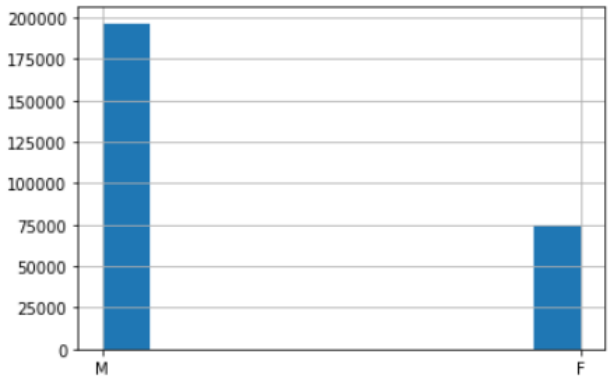
```
In [1]: import pandas as pd

In [3]: athlete_events=pd.read_csv("/Users/leno/Downloads/athlete_events.csv")

In [4]: athlete_events.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
```

3. Initial exploration and screenshots:

```
In [5]: athlete_events.Sex.hist()
```

```
Out[5]: <AxesSubplot:>
```




```
In [34]: pysqldf('''SELECT City,count(City) as count FROM athlete_events
                    GROUP BY City ORDER BY count Desc LIMIT 5 ''')
```

Out[34]:

	City	count
0	London	22426
1	Athina	15556
2	Sydney	13821
3	Atlanta	13780
4	Rio de Janeiro	13688

```
London Has the largest number of winning athletes
```

```
In [16]: pysqldf('''SELECT Sport,count(Sport) as count FROM athlete_events
                    GROUP BY Sport ORDER BY count Desc LIMIT 5''')
```

Out[16]:

	Sport	count
0	Athletics	38624
1	Gymnastics	26707
2	Swimming	23195
3	Shooting	11448
4	Cycling	10859

Most 5 common sports: 1- Athletics 2-Gymnastics 3-Swimming 4-Shooting 5- Cycling

In [20]:
```
pysqldf('''SELECT Name,count(Name) as counts,City,Medal FROM athlete_events
        WHERE Medal="Gold" GROUP BY Name ORDER BY counts desc''')
```

Out[20]:

	Name	counts	City	Medal
0	Michael Fred Phelps, II	23	Athina	Gold
1	Raymond Clarence "Ray" Ewry	10	Paris	Gold
2	Paavo Johannes Nurmi	9	Antwerpen	Gold
3	Mark Andrew Spitz	9	Mexico City	Gold
4	Larysa Semenivna Latynina (Diriy-)	9	Melbourne	Gold
...	...	...	...	...
10408	Aale Maria Tynni (-Pirinen, -Haavio)	1	London	Gold
10409	Aagje "Ada" Kok (-van der Linden)	1	Mexico City	Gold
10410	Aage Valdemar Harald Frandsen	1	Antwerpen	Gold
10411	Aage Jrgen Christian Andersen	1	Athina	Gold
10412	A. Albert	1	Paris	Gold

10413 rows × 4 columns

Most Winning Athlete was Michael Fred Phelps

In [24]:
```
pysqldf('''SELECT Name,count(Name) as counts,City,Medal FROM athlete_events
        WHERE Medal="Gold" GROUP BY Name HAVING counts>5 ORDER BY counts desc ''')
```

Out[24]:

	Name	counts	City	Medal
0	Michael Fred Phelps, II	23	Athina	Gold
1	Raymond Clarence "Ray" Ewry	10	Paris	Gold
2	Paavo Johannes Nurmi	9	Antwerpen	Gold
3	Mark Andrew Spitz	9	Mexico City	Gold
4	Larysa Semenivna Latynina (Diriy-)	9	Melbourne	Gold
5	Frederick Carlton "Carl" Lewis	9	Los Angeles	Gold
6	Usain St. Leo Bolt	8	Beijing	Gold
7	Sawao Kato	8	Mexico City	Gold
8	Ole Einar Bjrndalen	8	Nagano	Gold
9	Matthew Nicholas "Matt" Biondi	8	Los Angeles	Gold
10	Jennifer Elisabeth "Jenny" Thompson (-Cumpelik)	8	Barcelona	Gold
11	Birgit Fischer-Schmidt	8	Moskva	Gold
12	Vra slavsk (-Odloilov)	7	Tokyo	Gold
13	Viktor Ivanovych Chukarin	7	Helsinki	Gold
14	Nikolay Yefimovich Andrianov	7	Munich	Gold
15	Donald Arthur "Don" Schollander	7	Tokyo	Gold
16	Borys Anfiyanovych Shakhlin	7	Melbourne	Gold
17	Aladr Gerevich (-Gerei)	7	Los Angeles	Gold
18	Vitaly Venediktovich Shcherbo	6	Barcelona	Gold
19	Viktor An	6	Torino	Gold
20	Ryan Steven Lochte	6	Athina	Gold
21	Rudolf Krpti	6	London	Gold
22	Reiner Klimke	6	Tokyo	Gold

23	Pl dm Kovcs	6	Berlin	Gold
24	Nedo Nadi	6	Stockholm	Gold
25	Marit Bjrgen	6	Vancouver	Gold
26	Maria Valentina Vezzali	6	Atlanta	Gold
27	Lyubov Ivanovna Yegorova	6	Albertville	Gold
28	Lidiya Pavlovna Skoblikova (-Polozkova)	6	Squaw Valley	Gold
29	Kristin Otto	6	Seoul	Gold
30	Jason Francis Kenny	6	Beijing	Gold
31	Isabelle Regina Werth	6	Barcelona	Gold
32	Gert Fridolf Fredriksson	6	London	Gold
33	Gerard Theodor Hubert Van Innis	6	Paris	Gold
34	Edoardo Mangiarotti	6	Berlin	Gold
35	Christopher Andrew "Chris" Hoy	6	Athina	Gold
36	Amy Deloris Van Dyken (-Rouen)	6	Atlanta	Gold
37	Allyson Michelle Felix	6	Beijing	Gold
38	Akinori Nakayama	6	Mexico City	Gold

There are 38 athletes that have more than 5 medals

4. Write a 5-6 sentence paragraph describing your project; include who might be interested to learn about your findings. Who might be your audience?
Who might be interested in event_athelete findings areMy audience could be anyone who is interested in learning more about the findings of event_athlete. This includes people who are interested in sports, fitness, and health. It also includes people who are interested in data science and machine learning. My findings could be used by coaches, athletes, and trainers to improve their performance. They could also be used by researchers to learn more about human performance.

Here are some specific examples of people who might be interested in my findings:

- Coaches: My findings could help coaches to develop better training programs for their athletes.
- Athletes: My findings could help athletes to improve their performance by identifying areas where they can improve.
- Trainers: My findings could help trainers to develop better training programs for their clients.
- Researchers: My findings could help researchers to learn more about human performance and how it can be improved.

5. Question To Answer:

1-what are the most common sports?

2-What is the most common gender?

3- what is the most common city?

6. Initial Hypothesis:

1- what are the most common sports: Athletics

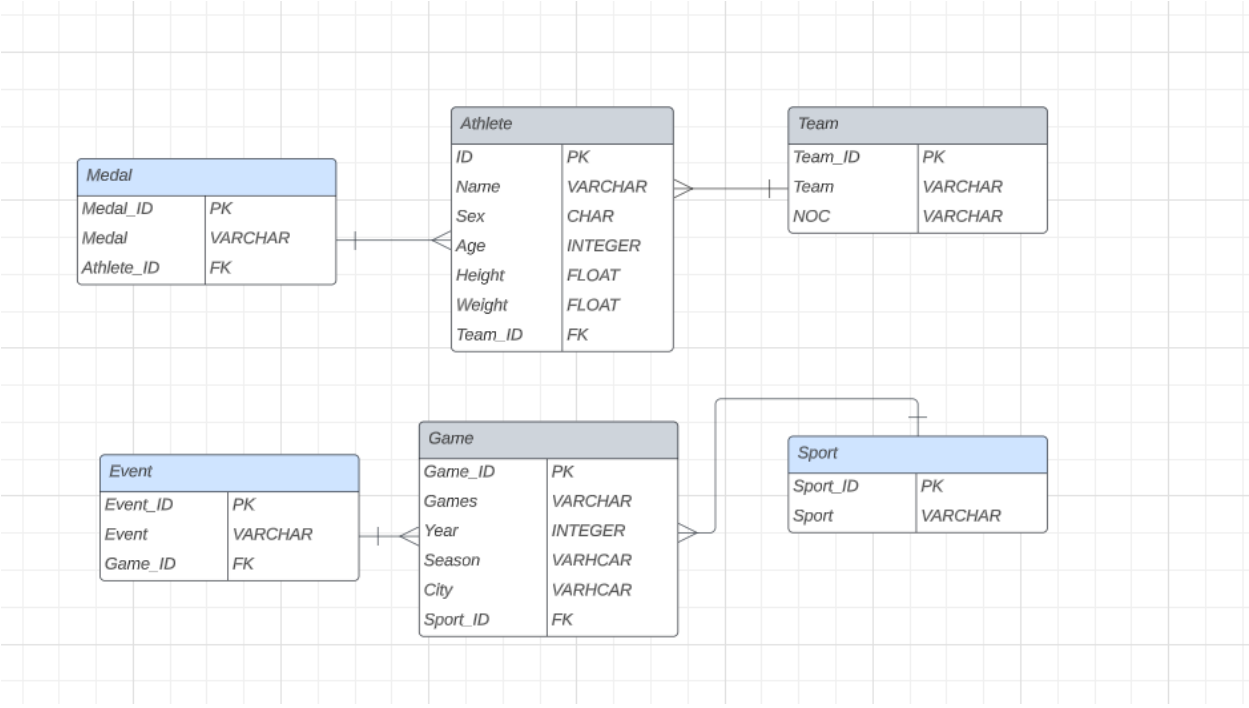
2- What is the most common gender:Males

3- what is the most common city:Barcelona

7. Approach:

- What features (fields/columns) are you going to look at first?
  City, Sport, Sex, Medal, Name
- Is there a relationship that exists that you want to explore?
  The participation of athletes in specific events during games
- What metric/ evaluation measure will you use?
  The metric I would use to explore participation of athletesThe metric I would use to explore participation of athletes is the number of athletes participating in a particular sport or event. This metric can be used to track the growth or decline of participation in a sport over time. It can also be used to compare participation rates between different sports or events. Additionally, this metric can be used to identify factors that may be influencing participation, such as the availability of facilities, the cost of participation, or the level of competition.

8. ERD:



9.Provide a summary of the different descriptive statistics you looked at and WHY.

I looked at Athletes Sex histogram Because I had initial hypothesis that there are more men than women and this was true,

I also looked at increasing athlete or sport events over the years

I looked at the average joined athletes age which was 25.5

10. Submit 2-3 key points you may have discovered about the data, e.g. new relationships. Aha's! Did you come up with additional ideas for other things to review?

- I found that london has the largest number of athletes
- I found that most common sport is athletics

11 . Did you prove or disprove any of your initial hypotheses? If so, which one and what do you plan to do next?

I approved my initial hypothesis that male athletes are more than woman And that most common sport is athletics

I disapproved of my initial hypothesis about the city with the largest number of athletes I thought that it was Barcelona but it turned out to be London.

And I plan to explore more about London sports and medals and names of athletes to know more.

12 . What additional questions are you seeking to answer?

The question is: What is the most popular sport among female participants?