SQL Project

Milestone 1: Preparing my proposal

- I decided to work with SportsStats and the Olympics dataset.
- I'm very interested in sports and what separates Olympic-level athletes from their competition.
- My goal will be to uncover any trends or patterns in the traits of successful Olympians.

Data preparation

To prepare this data I downloaded the .CSV files and read them into data frames in my python environment.

From there, I connected to my SQL database using the SQLite3 package and created the tables in the database.

```
conn = sql.connect(db)
conn
```

```
<sqlite3.Connection at 0x24f44c8f010>
```

```
df1.to_sql('athlete_events',conn,if_exists='replace')
```

```
271116 ...
```

```
df2.to_sql('countries',conn,if_exists='replace')
```

```
230 ...
```

Descriptive Statistics

Used the Pandas command describe(include='all') to generate the list of descriptive statistics for the main dataset, **athlete_events.**

The dataset features information about each athlete including what event they participated in and which country they represent.

```
df1.describe(include='all')
```

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event
count	271116.000000	271116	271116	261642.000000	210945.000000	208241.000000	271116	271116	271116	271116.000000	271116	271116	271116	271116
unique	NaN	134732	2	NaN	NaN	NaN	1184	230	51	NaN	2	42	66	765
top	NaN	Robert Tait McKenzie	M	NaN	NaN	NaN	United States	USA	2000 Summer	NaN	Summer	London	Athletics	Football Men's Football
freq	NaN	58	196594	NaN	NaN	NaN	17847	18853	13821	NaN	222552	22426	38624	5733
mean	68248.954396	NaN	NaN	25.556898	175.338970	70.702393	NaN	NaN	NaN	1978.378480	NaN	NaN	NaN	NaN
std	39022.286345	NaN	NaN	6.393561	10.518462	14.348020	NaN	NaN	NaN	29.877632	NaN	NaN	NaN	NaN
min	1.000000	NaN	NaN	10.000000	127.000000	25.000000	NaN	NaN	NaN	1896.000000	NaN	NaN	NaN	NaN
25%	34643.000000	NaN	NaN	21.000000	168.000000	60.000000	NaN	NaN	NaN	1960.000000	NaN	NaN	NaN	NaN
50%	68205.000000	NaN	NaN	24.000000	175.000000	70.000000	NaN	NaN	NaN	1988.000000	NaN	NaN	NaN	NaN
75%	102097.250000	NaN	NaN	28.000000	183.000000	79.000000	NaN	NaN	NaN	2002.000000	NaN	NaN	NaN	NaN
max	135571.000000	NaN	NaN	97.000000	226.000000	214.000000	NaN	NaN	NaN	2016.000000	NaN	NaN	NaN	NaN

Quick insights

A few quick insights from the initial data exploration:

- The average age of an Olympic athlete is 25.56 years old.
- There are 51 different Olympics represented in the dataset, along with 765 unique events.
- 42 different cities have hosted the Olympics over this time, with London having the most.

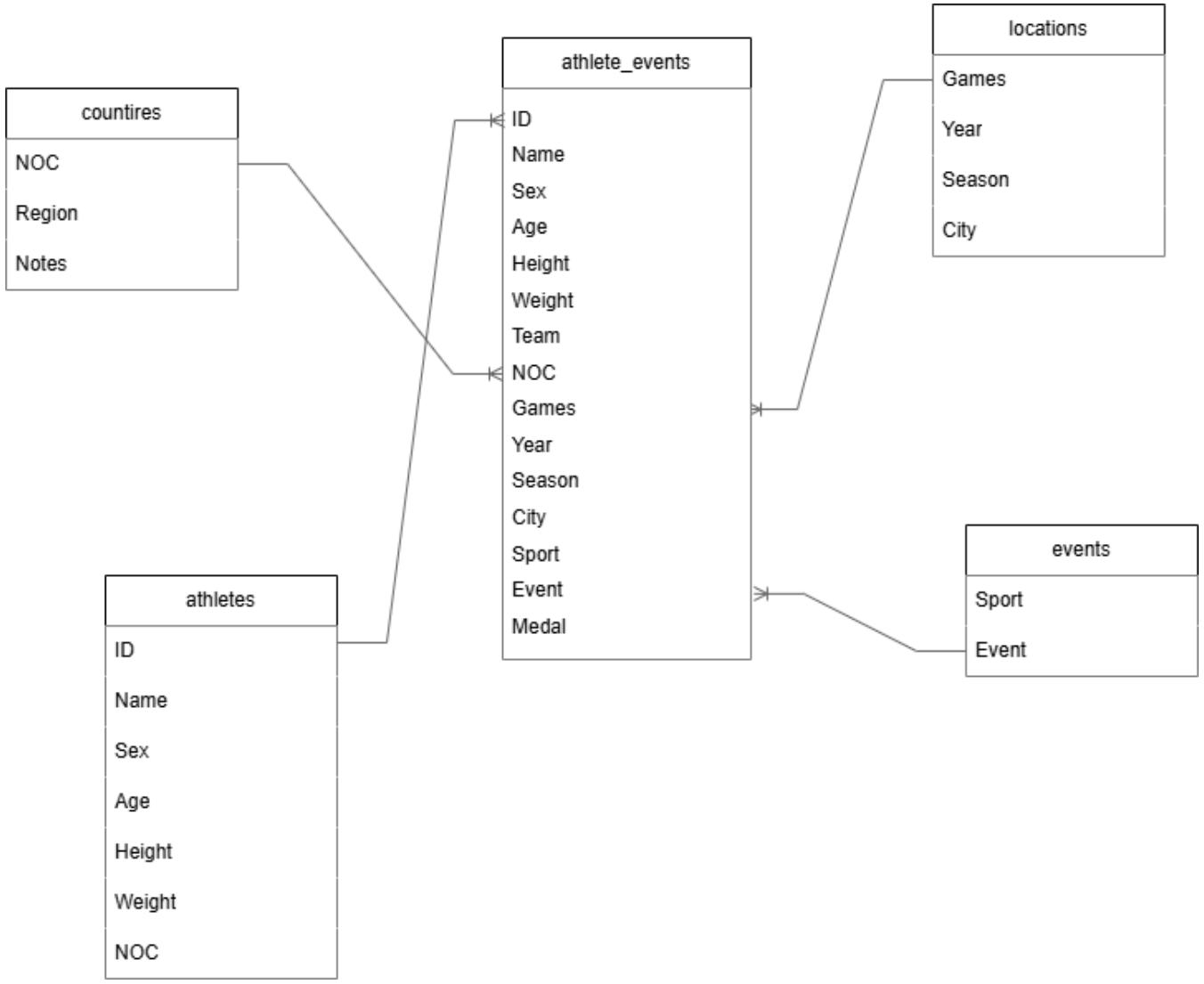
Data relationships and ERD diagram

- Two datasets have been provided, **athlete_events** and **countries**.
- I decided to create three additional tables for the database, named **athletes, locations,** and **events.**
- These auxiliary tables will link back to the main table in one-to-many relationships.
- Example:

```
cursor.execute("""
                CREATE TABLE locations
                as
                SELECT games, year, season, city
                FROM athlete_events
                group by games, year, season, city
                order by year ASC
                """)
```

ERD Diagram

All four auxiliary tables connect to **athlete_events** in a one-to-many relationship.



Project Proposal

As a partner of SportsStats, I will be investigating which attributes make a successful Olympian. These athletes are the best at what they do in the entire world, and need to find the smallest of advantages to edge out their competition in the chase for the coveted Olympic medals.

My audience for this project may be trainers, coaches, or athletes themselves. Trainers and athletes may be interested in hitting a target weight for peak performance. Coaches may be interested in what height and age their players should be when selecting their roster.

Questions

Question 1: Are there any events where physical attributes (height, weight, and age) lead to more success?

Question 2: Does location play a factor in an athletes success at the Olympics?

Hypotheses

- I believe there will be certain sports where physical attributes do contribute to success, such as height leading to success in basketball. It is well known that basketball players need to be tall.
- Certain sports such as wrestling have weight classes, so it's possible other attributes may be important in those events.
- I also believe athletes may perform better when the Olympics are being held in their home country, due to fan support and familiarity.

Approach

- My first step will be to compare physical attributes across the different events. This will give me an idea of which events have notable physical attributes such as height, weight, or age.
- I'm also interested in the relationship between the athletes who medal and the athletes who don’t. For example, if basketball athletes are the tallest group on average, what other attributes, (if any), separate the winners from the losers?

Approach (continued)

- For locations I will be examining the change in success by country from year to year, and seeing how often performance increases or decreases based on the location.
- I can use the medal count for each country as an evaluation metric for this question.

Milestone 2: Descriptive Statistics

Two major points of focus:

1. Physical attributes (height, weight, and age) of athletes compared across sports.
2. Medal count by Olympic games per country – with a focus on host countries performance.

Code

```
avg_age = pd.read_sql("""SELECT
                         avg(age) as average_age, sport
                         FROM
                         athlete_events
                         GROUP BY sport
                         HAVING average_age > 0
                         ORDER BY average_age""",
                        conn)
```

Average age grouped by sport

```
medal_count = pd.read_sql("""SELECT
                                a.NOC,
                                c.region,
                                a.games,
                                a.year,
                                a.season,
                                a.city,
                                SUM(a.medal_bool) AS medal_count,
                                COUNT(a.id) as total_athletes
                                FROM
                                (
                                SELECT
                                *,
                                (CASE WHEN medal IS NOT NULL THEN 1
                                    ELSE 0
                                    END) AS medal_bool
                                FROM athlete_events) as a
                                LEFT JOIN
                                countries as c
                                ON
                                a.NOC = c.NOC
                                GROUP BY
                                a.NOC, a.games, a.year, a.season, a.city
                                ORDER BY medal_count DESC""",
                               conn)
```

Olympic medal count by year, with host city

Takeaways – physical attributes

	average_age	Sport
0	18.737082	Rhythmic Gymnastics
1	20.566803	Swimming
2	22.232190	Figure Skating
3	22.366851	Synchronized Swimming
4	22.481441	Diving
...	...	...
61	34.390831	Equestrianism
62	35.333333	Polo
63	38.812500	Alpinism
64	45.901009	Art Competitions
65	53.333333	Roque

- Some of the oldest sports included croquet, polo, and alpinism. This makes sense logically because physical fitness is not super important for success in those events.
- Average height seemed to have a relationship with both age and weight. For example, many sports with a shorter average height (e.g. gymnastics, figure skating, diving) also had a very low average age. Additionally, many sports with a taller average height (basketball, rowing, water polo) also had large average weights.

Takeaways – medal count

- Many of the highest medal counts did come from teams competing in their home country. In fact, the highest medal count in the dataset was Russia in 1980, when the event was held in Moscow.
- Great Britain, USA, France, and Germany also appear on this list of the most medals, all while competing in their home country.

	NOC	region	Games	Year	Season	City	medal_count	total_athletes
0	URS	Russia	1980 Summer	1980	Summer	Moskva	442	660
1	USA	USA	1904 Summer	1904	Summer	St. Louis	394	1109
2	GBR	UK	1908 Summer	1908	Summer	London	368	972
3	USA	USA	1984 Summer	1984	Summer	Los Angeles	352	693
4	USA	USA	2008 Summer	2008	Summer	Beijing	317	763
5	URS	Russia	1988 Summer	1988	Summer	Seoul	300	647
6	URS	Russia	1976 Summer	1976	Summer	Montreal	286	574
7	GDR	Germany	1980 Summer	1980	Summer	Moskva	264	495
8	USA	USA	2016 Summer	2016	Summer	Rio de Janeiro	264	719
9	USA	USA	2004 Summer	2004	Summer	Athina	263	726
10	USA	USA	1996 Summer	1996	Summer	Atlanta	259	839
11	USA	USA	2012 Summer	2012	Summer	London	248	689
12	USA	USA	2000 Summer	2000	Summer	Sydney	242	764
13	FRA	France	1900 Summer	1900	Summer	Paris	235	1071
14	GER	Germany	1936 Summer	1936	Summer	Berlin	224	581

Revisiting Hypotheses

- From my original hypotheses, I believe the data mostly agrees with them. Basketball had the tallest average height which I predicted to be true.
- The data also supports host countries being more successful than usual. The USA's two highest medal counts happened in St. Louis in 1904 and Los Angeles in 1984.
    - Having more athletes competing may skew this data.

Further Questions

- First, I want to determine the difference in height, weight, and age between medalists and non-medalists. This may show some less obvious findings about which physical attributes drive success.
- Second, I want to compare the ratio of medals to athletes for these high medal count countries, to further determine if the host countries have more success.

Milestone 3: Insights and Analysis

```
medal_avg = pd.read_sql("""SELECT
                                NOC,
                                region,
                                season,
                                (AVG(medal_count) / AVG(total_athletes)) as avg_medal_pct
                                FROM medal_count
                                GROUP BY NOC, region, season
                                ORDER BY avg_medal_pct DESC""",
                           conn)
```

Average medal percent by country for summer and winter Olympics

New Metrics

1. Determined the average physical attributes of competitors split into medalists and non-medalists. This allows me to find sports with the biggest difference in stats for the more successful athletes.
2. For my other hypothesis I created a 'medal_percent' metric. This is a ratio of medal count divided by total athletes for each country. This can be used to determine if a country's percentage of medalists was higher or lower than their average performance.

Physical attributes

- For the majority of sports, there is NOT a relationship between physical attributes and success at the Olympics.
- While certain sports require specific attributes to qualify (height, weight, age), the difference between medalists and non-medalists is not noticeable. (Often less than one year).
- There are some outliers which will be noted on the next slide

Sport	no_medal	medal
Rhythmic Gymnastics	18.688213	18.931818
Swimming	20.511861	20.923356
Diving	22.521323	22.264775
Short Track Speed Skating	22.889600	22.429577
Boxing	23.054189	23.058002
Gymnastics	22.669937	23.406493
Taekwondo	24.244589	23.451389
Ski Jumping	23.308716	23.753623

```
age_diff = pd.read_sql("""SELECT
                            AVG(age) as average_age, sport, medal_bool
                            FROM
                            (
                             SELECT
                             *,
                             (CASE WHEN medal IS NOT NULL THEN 1
                                 ELSE 0
                                 END) AS medal_bool
                             FROM athlete_events)
                            GROUP BY sport, medal_bool
                            HAVING average_age > 0
                            ORDER BY sport, medal_bool
                            """, conn)
age_pivot = age_diff.pivot(index='Sport',columns='medal_bool',values='average_age')
```

Outliers

medal_bool	no_medal	medal	diff
Sport			
Archery	27.441860	30.927900	-3.486039
Curling	30.488746	33.302632	-2.813886
Jeu De Paume	33.750000	29.000000	4.750000
Croquet	36.714286	31.125000	5.589286

medal_bool	no_medal	medal	diff
Sport			
Golf	70.836538	80.500000	-9.663462
Softball	65.955390	69.843023	-3.887633
Tug-Of-War	97.478261	94.137931	3.340330
Skeleton	74.660000	71.083333	3.576667

medal_bool	no_medal	medal	diff
Sport			
Golf	173.991736	177.900000	-3.908264
Rhythmic Gymnastics	167.153386	170.638462	-3.485075
Gymnastics	163.030254	161.576860	1.453395
Tug-Of-War	183.933333	180.300000	3.633333

- For age, older athletes have an advantage in archery and curling. Younger competitors have an advantage in jeu de paume and croquet. This is interesting since croquet was one of the oldest sports on average.
- For weight, heavier competitors have an advantage in golf and softball. In both sports heavier players can likely hit the ball farther.
- For height, taller golfers once again have an advantage, likely for the same reason. Shorter athletes have an advantage in tug-of-war, possibly due to a lower center of gravity.

Average medal count

- There is a correlation between the number of athletes participating for a country and where the event was hosted. Many of the highest counts of participants occurred in years that country was hosting (for example Russia in Moscow or France in Paris).
- Determining the average medal count for a country throughout the entire dataset can tell us if countries performed better while hosting.

	NOC	region	Season	avg_medal_pct
0	URS	Russia	Summer	0.446344
1	URS	Russia	Winter	0.413923
2	GDR	Germany	Summer	0.400665
3	EUN	Russia	Summer	0.338462
4	ANZ	Australia	Summer	0.337209
5	USA	USA	Summer	0.332050
6	GDR	Germany	Winter	0.299445
7	EUN	Russia	Winter	0.275701
8	RUS	Russia	Summer	0.252582
9	WIF	Trinidad	Summer	0.250000
10	SRB	Serbia	Summer	0.236111
11	GER	Germany	Summer	0.233403

Conclusion

- The percent of medalists is higher for the majority of host countries when compared to their average medal percentage.
- For example, the USA's average medal percentage at Summer games is 33%. In Los Angeles they had 51% and in St. Louis they had 36%.
- This can be applied to the majority of host countries.

	NOC	region	Games	Year	Season	City	medal_count	total_athletes	medal_pct
0	URS	Russia	1980 Summer	1980	Summer	Moskva	442	660	0.669697
1	USA	USA	1904 Summer	1904	Summer	St. Louis	394	1109	0.355275
2	GBR	UK	1908 Summer	1908	Summer	London	368	972	0.378601
3	USA	USA	1984 Summer	1984	Summer	Los Angeles	352	693	0.507937
4	USA	USA	2008 Summer	2008	Summer	Beijing	317	763	0.415465
5	URS	Russia	1988 Summer	1988	Summer	Seoul	300	647	0.463679
6	URS	Russia	1976 Summer	1976	Summer	Montreal	286	574	0.498258
7	GDR	Germany	1980 Summer	1980	Summer	Moskva	264	495	0.533333
8	USA	USA	2016 Summer	2016	Summer	Rio de Janeiro	264	719	0.367177
9	USA	USA	2004 Summer	2004	Summer	Athina	263	726	0.362259
10	USA	USA	1996 Summer	1996	Summer	Atlanta	259	839	0.308701
11	USA	USA	2012 Summer	2012	Summer	London	248	689	0.359942
12	USA	USA	2000 Summer	2000	Summer	Sydney	242	764	0.316754
13	FRA	France	1900 Summer	1900	Summer	Paris	235	1071	0.219421
14	GER	Germany	1936 Summer	1936	Summer	Berlin	224	581	0.385542

Milestone 4: Recommendations

- My hypotheses on physical attributes have been proven to have some truth. Specific traits are required to qualify for the Olympics, such as height in basketball or volleyball. These traits should be considered by trainers and coaches when selecting their roster.
- Additionally there are some sports where physical attributes provide more or less success. These traits should be studied and targeted by the athletes and trainers alike.

Recommendations (cont.)

- In regards to athlete success in their home country, this was also proven to be true. The majority of countries were more successful in the years they hosted the Olympics.
- Athletes and trainers can learn from this by trying to create/foster a comfortable environment for themselves. Familiarity with the facilities or culture may also be worth considering.
- For sponsors, brands, or even gamblers, consider supporting the host nation athletes each year, as they may be more likely to succeed than others.

Thanks for your time!