Milestone 4: Presenting your findings (Storytelling)

Project Proposal

- My client/dataset is SportsStats. It is a sports analysis firm partnering with local news and elite personal trainers to provide “interesting” insights to help their partners. Insights could be patterns/trends highlighting certain groups/events/countries, etc. for the purpose of developing a news story or discovering key health insights.
- I downloaded the datasets from Coursera. The data is in a csv file. I used Databricks to import and clean my data. The data was already clean, so I didn’t have much to do.
- Determining the overall athlete ranking:

```
1  select Name, Medal, count(Medal)
2  from delta_athlete_events
3  where Medal != 'No medal'
4  group by Name, Medal
5  order by count(Medal) desc
```

▸ (2) Spark Jobs

Table +

	Name	Medal	count(Medal)
1	Michael Fred Phelps, II	Gold	23
2	Raymond Clarence "Ray" Ewry	Gold	10
3	Paavo Johannes Nurmi	Gold	9
4	Mark Andrew Spitz	Gold	9
5	Frederick Carlton "Carl" Lewis	Gold	9
6	Larysa Semenivna Latynina (Diriy-)	Gold	9
7	Ole Einar Birndalen	Gold	8

10,000 rows | Truncated data | 3.29 seconds runtime

- Determining the overall Team ranking:

```
1   select Team, count(Medal) as gold_medal
2   from delta_athlete_events
3   where Medal = 'Gold'
4   group by Team
5   order by gold_medal desc
```

(2) Spark Jobs

Table +

	Team	gold_medal
1	United States	2474
2	Soviet Union	1058
3	Germany	679
4	Italy	535
5	Great Britain	519
6	France	455
7	Sweden	451

242 rows | 2.60 seconds runtime

- Determining the number of athletes by Team:

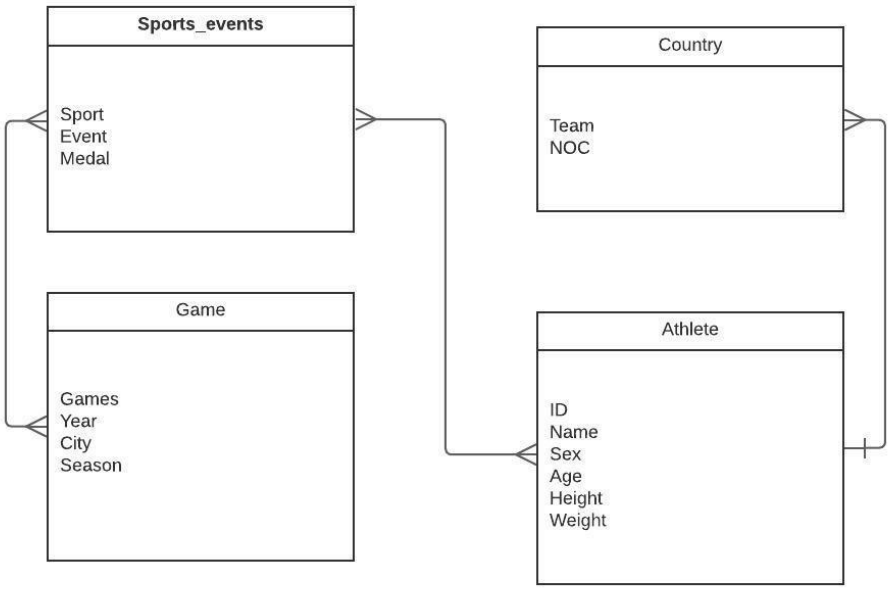
```
1   select Team, count(distinct Name) as athlete_count
2   from delta_athlete_events
3   group by Team
4   order by athlete_count desc
```

(3) Spark Jobs

Table +

	Team	athlete_count
1	United States	9114
2	France	5777
3	Great Britain	5758
4	Italy	4688
5	Germany	4569
6	Canada	4546
7	Japan	3981

- Here is the ER diagram related to my project that help us have a better understanding of the structure of the data:



Description

My project analyses Olympic games performances overtime by athlete. It gives a better understanding on what athlete, what country won a medal in a given Olympic game. The goal of this project is to identify key patterns on the winning element at the Olympics. The audience can be country sport federations, journalists, CIO, bookmakers, Olympics enthusiasts...

Questions

- Which team/country has the most chance of getting a medal at a given sport?
- What kind of athlete has the most chance of getting a medal at a given sport?
- Do more athletes mean more medals?

Hypothesis

- The more athletes your team have, the higher chances you have to a win a medal
- Tall athletes have a higher of winning at sports like basketball, swimming...

Approach

I am going to filter and group by sports to see how the other columns are correlated or not. I will group by team as well to see if there are any key patterns. Also, I will look at the historical performances of teams and athletes to see if there are signs for future great performances.

Discuss Insights Discovered

I did some descriptive statistic to answer the following questions:

- What kind of athlete has the most chance of getting a medal?
    - To do so, I looked at the average values of Age, Height and Weight for the top winning countries in the Olympics to see if I can identify some key patterns.
- Do more athletes mean more medals?
    - To do so, I looked for any correlation between the number of athletes in a team and the number of medals won.

What kind of athlete has the most chance of getting a medal?

```
pysqldf('''SELECT NOC, Team, avg(Age), avg (Height), avg(Weight), count(Medal) as Total_Medal
           from df
           where Medal is not null
           group by NOC
           order by count(Medal) desc
           limit 10;''')
```

	NOC	Team	avg(Age)	avg (Height)	avg(Weight)	Total_Medal
0	USA	United States	24.896638	179.519904	75.091918	5637
1	URS	Soviet Union	25.268877	176.833992	75.069118	2503
2	GER	Germany	26.416510	178.682646	74.468635	2165
3	GBR	Great Britain	27.846310	178.585981	74.552347	2068
4	FRA	France	27.982153	178.591992	74.383300	1777
5	ITA	Italy	26.956195	178.426520	75.703008	1637
6	SWE	Sweden	28.066406	179.541457	76.492188	1536
7	CAN	Canada	25.826677	176.171004	73.541705	1352
8	AUS	Australia	24.819011	178.884459	74.095533	1320
9	RUS	Russia	25.661790	177.277729	72.109692	1165

Here, we can see from the Top 10 country with the most medals that an athlete has a higher winning chance when:

- He/she is between 24 and 27 years old.
- He/she is between 176 cm and 179 cm.
- He/she weighs between 74 kg and 76 kg

Obviously, these stats depend on what sports is considered. These stats give an overall for all sports.

Do more athletes mean more medals?

```
pysqldf('''SELECT Year, Team as Top_team, count(Medal) as Total_Medal, count(distinct Name) as Athlete_number
           from df
           where Medal is not null
           group by NOC
           order by count(Medal) desc
           limit 15;''')
```

	Year	Top_team	Total_Medal	Athlete_number
0	2004	United States	5637	3836
1	1972	Soviet Union	2503	1675
2	2002	Germany	2165	1518
3	1980	Great Britain	2068	1601
4	1956	France	1777	1277
5	2016	Italy	1637	1115
6	1912	Sweden	1536	1107
7	1984	Canada	1352	1064
8	2008	Australia	1320	863
9	2008	Russia	1165	826
10	1992	Hungary	1135	692
11	1988	Netherlands	1040	731
12	1992	Norway	1033	661
13	1972	East Germany	1005	694
14	1998	China	989	653

From here, we can the top country with most medals in a single competition. We can see that there is a correlation between the number of athlete and the number of medals won. The more athletes you have, the more medals you get.

Next, I can look at the influence of a given sport as you have individual sport (1 athlete per competition = 1 medal) and collective sport (X athletes per competition = X medals).

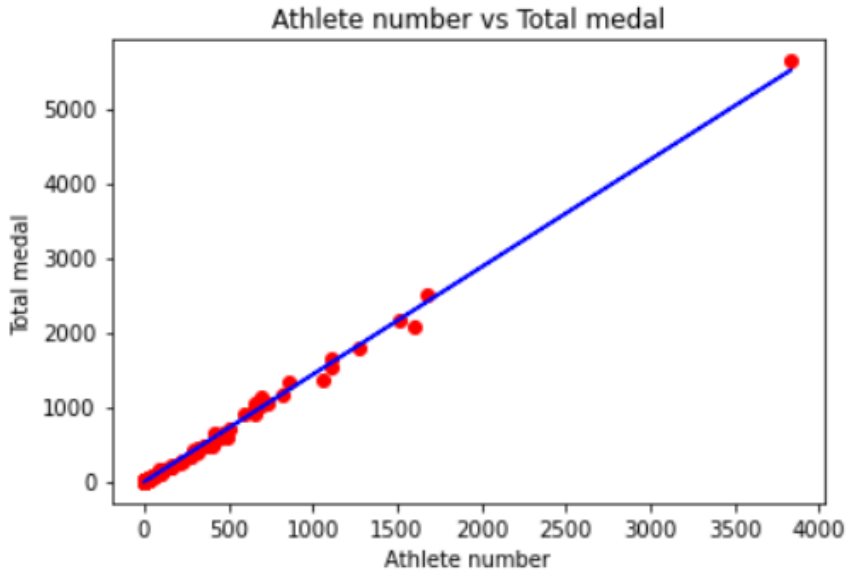
My hypotheses were:

- The more athletes your team have, the higher chances you have to a win a medal.
    - This hypothesis has been confirmed by my research. I planned to look deeper into it by filtering by sport.
- Young athletes in the prime of their career (around 25 years old) have a higher chance to win a medal.
    - This hypothesis has been confirmed by my research. I planned to look deeper into it by filtering by sport to see the influence of height and weight.

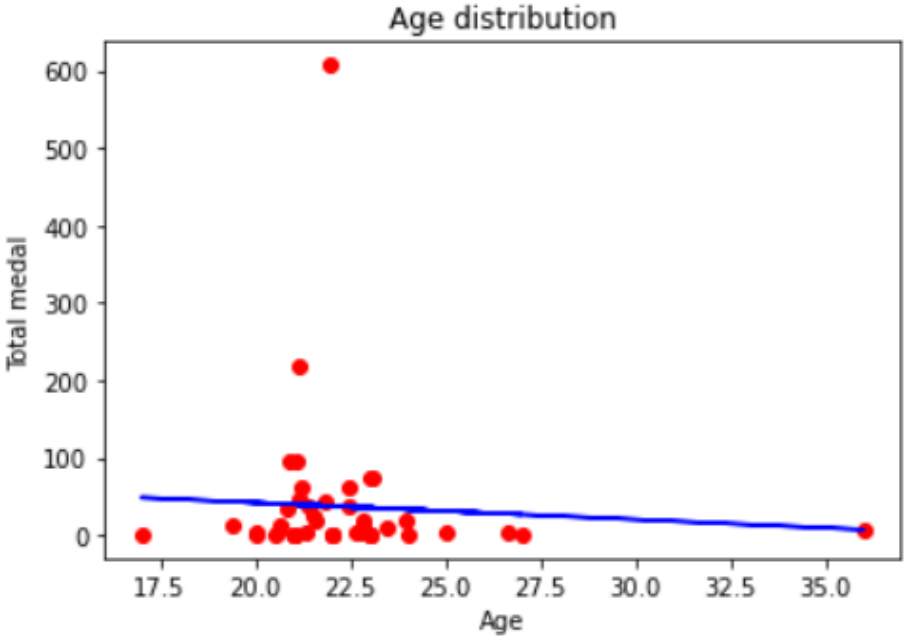
I need to find the results for a given sports to have a more precise understanding of my data. I will try to find the best athlete stats for swimming and basketball for example.

Go broader

I was able to find a positive correlation (linear regression) between the number of athletes per country and the number of medals won by that team: the more athletes you have in your team, the more medals you can win.



I deep dived into my data considering men swimmers and trying to find some correlation between Age / Height / Weight and Total medal won. Unfortunately, there is no correlation between these parameters.



I discovered that it is hard to determine any correlation between the number of medals won and athlete's physical characteristics. I can only consider average values (Age, Height and Weight) to give a global indication on the archetype of the winning athlete, but it is not something to rely on to predict future winner.

My cross-category metric is the **total number of medals** won. It helps me verify that any correlation, relationship between categories is valid or not.

Recommendations and Actions

- It is important to educate a lot of young talents in different sports so that you can constitute a huge team of talented athletes in order to increase your chances to win as much medals as you can.
- Furthermore, athletes in the prime of their career is when they are around 25 years old, regardless of the sport. This is the overall average age where athletes win a medal. Make sure that athletes reach their full potential, without any injuries, at the time of the Olympics.