CAPSTONE PROJECT

MILESTONE 4

JUAN SÁNCHEZ

DESCRIPTION

- My project consists in analyzing the data from all the olympics events since 1896 for summer and 1924 for winter, until summer of 2016.
- From this data I expect to find relationships between winning medals and sex, height, weight, which countries have better athletes for specific sports, which sports are more popular among other things.

MY AUDIENCE

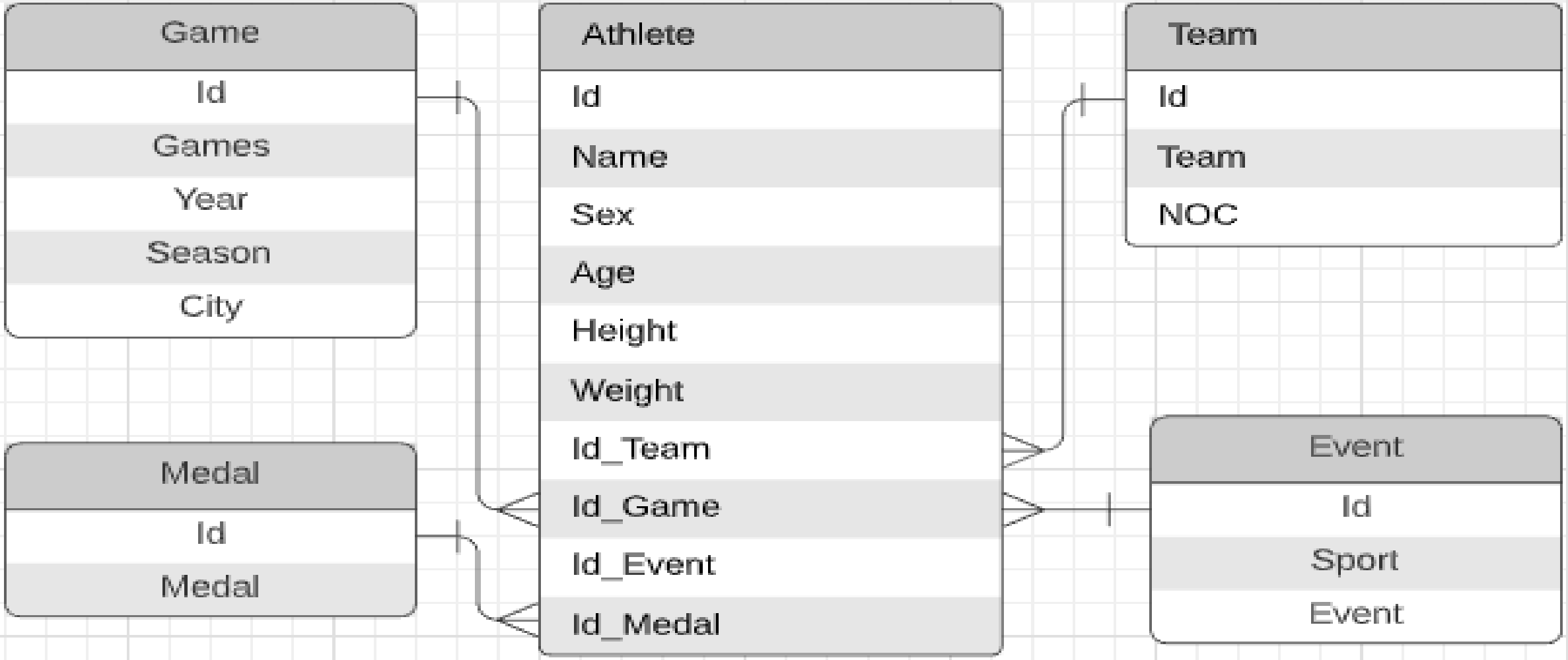
- Current athletes
- Sports endorsement companies
- People interested in sports
- Companies that make sports clothings and other related products.

APPROACH

- For this Project the approach to the data start with the design of a ERD-Diagram
- The csv file which contains the data is modified in Excel in order to make smaller tables that correspond to the ERD-Diagram
- With SQLite I clean the data checking for missing values and querying it to find insights
- With Spark I run the querys made in SQLite and it also gives me the posibility to make graphics and use functions that SQLite cannot use to further my analysis

ERD OF SPORTSTATS

Juan Sanchez | May 13, 2020



FIRST HYPOTHESES

- For a specific sport, there is an optimal height/weight ratio in order to succeed in that sport
- Gender does not have a significance influence in medal achievement
- Team sports are less popular tan individual sports

Some of these hypotheses had to be modified in some form and other were discarded entirely, new hypotheses were later formulated

RESULTS

- First metric created: Height_Weight_ratio

Cmd 3

```
1  SELECT *,
2         round(avg_Height/avg_Weight,2) AS Height_Weight_Ratio
3  FROM
4  (SELECT   Sex as Gender
5          ,round(avg(age),2) as avg_age
6          ,round(avg(Height),2) as avg_Height
7          ,round(avg(Weight),2) as avg_Weight
8          ,COUNT(DISTINCT ID) AS athlete_count
9  FROM athlete
10 GROUP BY Sex)
```

▸ (5) Spark Jobs

Gender	avg_age	avg_Height	avg_Weight	athlete_count	Height_Weight_Ratio
F	23.73	167.84	60.02	33981	2.8
M	26.28	178.86	75.74	101590	2.36

Command took 19.09 seconds -- by js.sanchez130@uniandes.edu.co at 28/5/2020 11:21:54 on cluster

- Initial descriptive analysis, the first query is made to find mean values of the athletes characteristics

RESULTS

- The athletes which did not win medals were filtered out and then grouped by the medals won

Cmd 4

```
--Descriptive Analysis for athletes that won medals grouped by medal
SELECT *,
       round(avg_Height/avg_Weight,2) AS Height_Weight_Ratio
FROM
(SELECT  Sex as Gender
        ,Medal
        ,round(avg(age),2) as avg_age
        ,round(avg(Height),2) as avg_Height
        ,round(avg(Weight),2) as avg_Weight
        ,COUNT(DISTINCT ID) AS athlete_count
FROM athlete
LEFT JOIN medal ON athlete.Medalid=medal.Medalid
WHERE Medal IS NOT NULL
GROUP BY Medal,Gender)
ORDER BY athlete_count
```

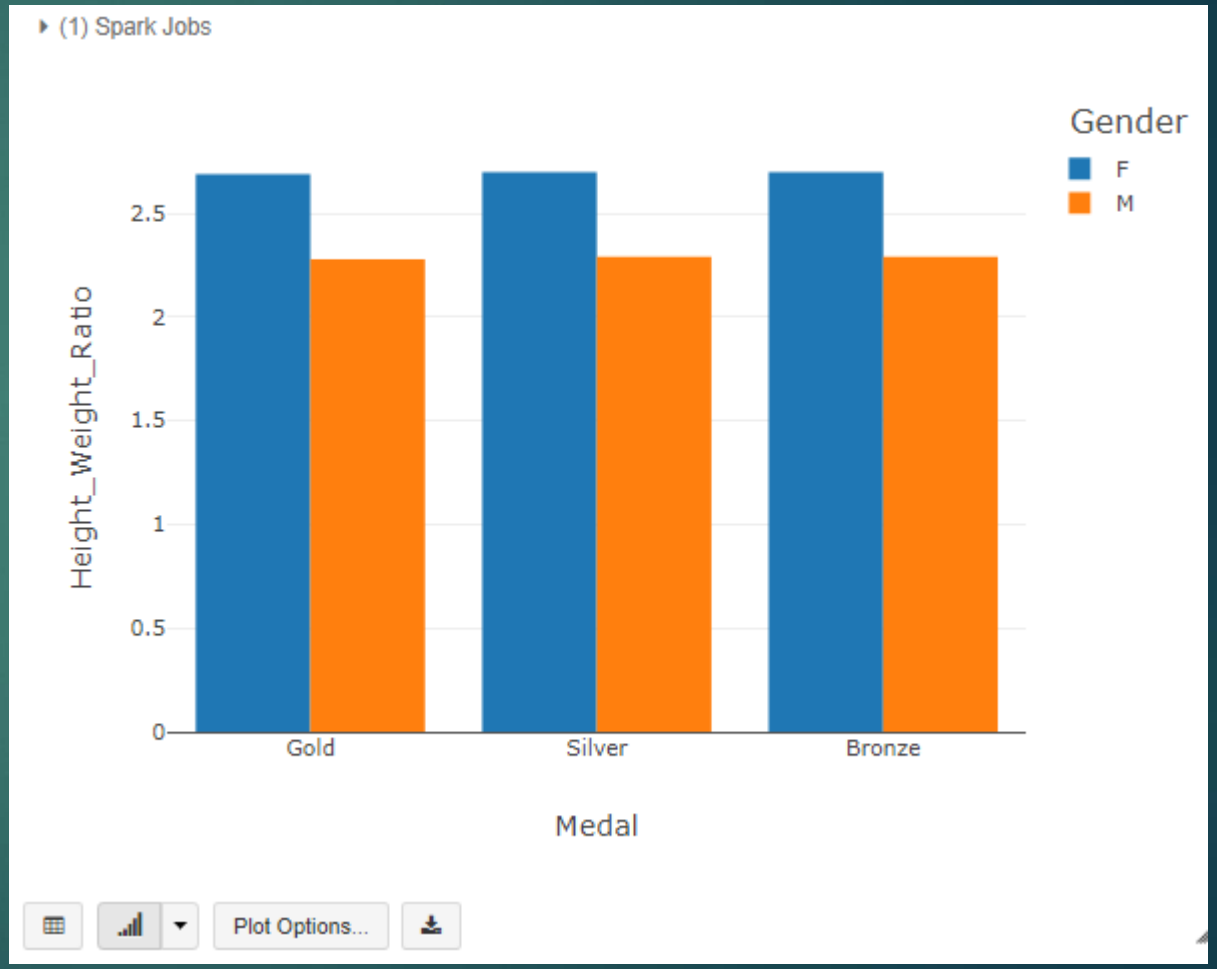
(1) Spark Jobs

Gender	Medal	avg_age	avg_Height	avg_Weight	athlete_count	Height_Weight_Ratio
F	Gold	24.37	170.69	63.46	2706	2.69
F	Silver	24.44	170.5	63.17	3147	2.7
F	Bronze	24.71	170.23	63.03	3325	2.7
M	Gold	26.5	181.47	79.59	7719	2.28
M	Silver	26.63	181.07	79.16	8294	2.29
M	Bronze	26.36	180.92	78.86	8554	2.29

Command took 10.35 seconds -- by js.sanchez130@uniandes.edu.co at 28/5/2020 11:40:51 on cluster

RESULTS

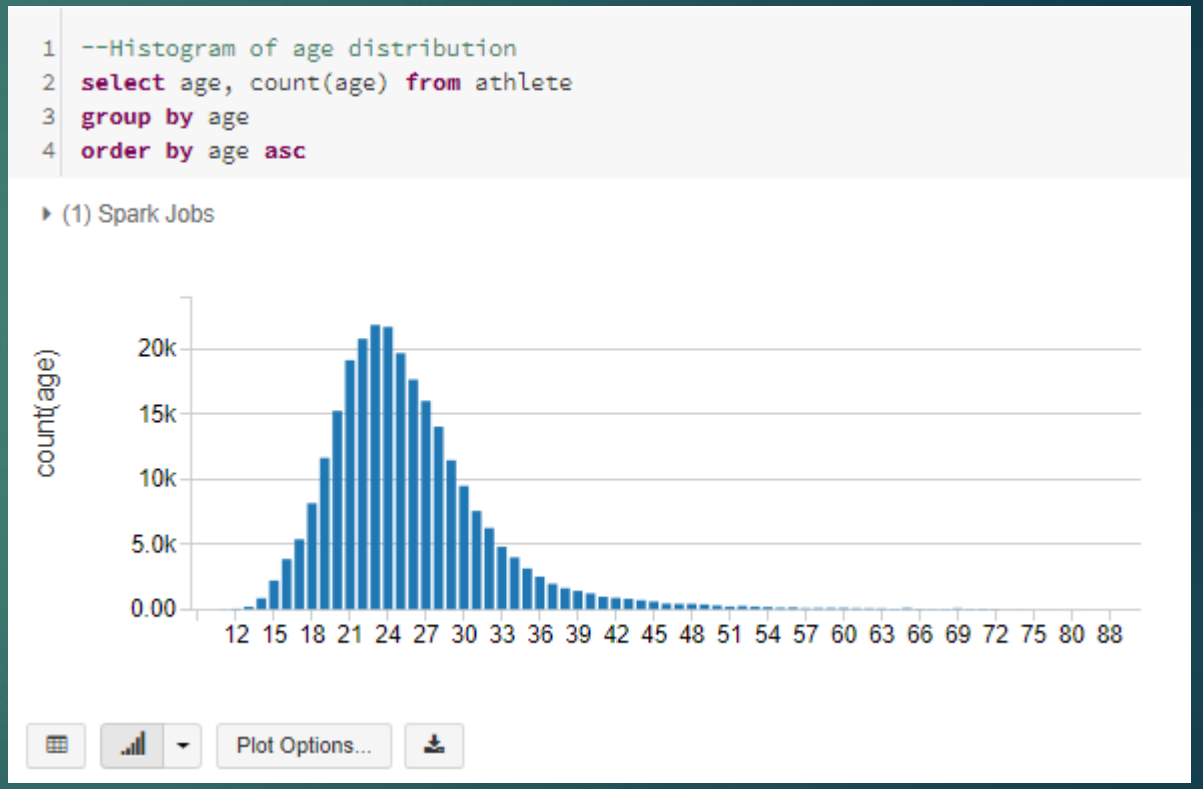
- Is noted that for men and women, the average values of age, height and weight slightly increases from winners of bronce medals to gold medals, yet the height-weight ratio the same



RESULTS

- Histograms of age, height and weight for all the athletes were made to see how is the distribution of data which seems to be a normal distribution in general and is noted that there have been athletes from age 10 to 97 which is very unespected and interesting

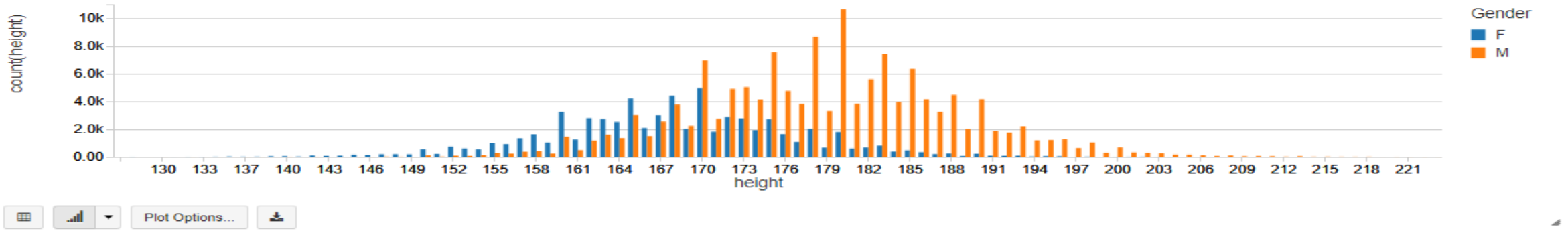
```
1 --Histogram of age distribution
2 select age, count(age) from athlete
3 group by age
4 order by age asc
```



RESULTS

```
1  --Histogram of height group by gender
2  select sex as Gender,height, count(height) from athlete
3  group by height, Gender
4  order by height asc
```

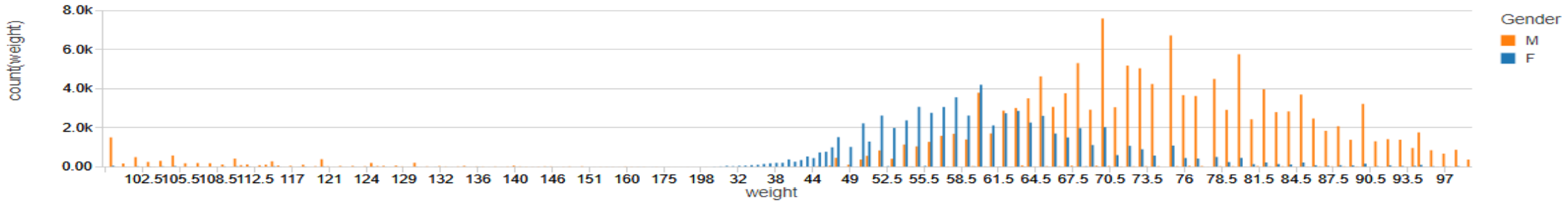
▸ (1) Spark Jobs



Plot Options...

```
1  --Histogram of weight group by gender
2  select sex as Gender,weight, count(weight) from athlete
3  group by weight, Gender
4  order by weight asc
```

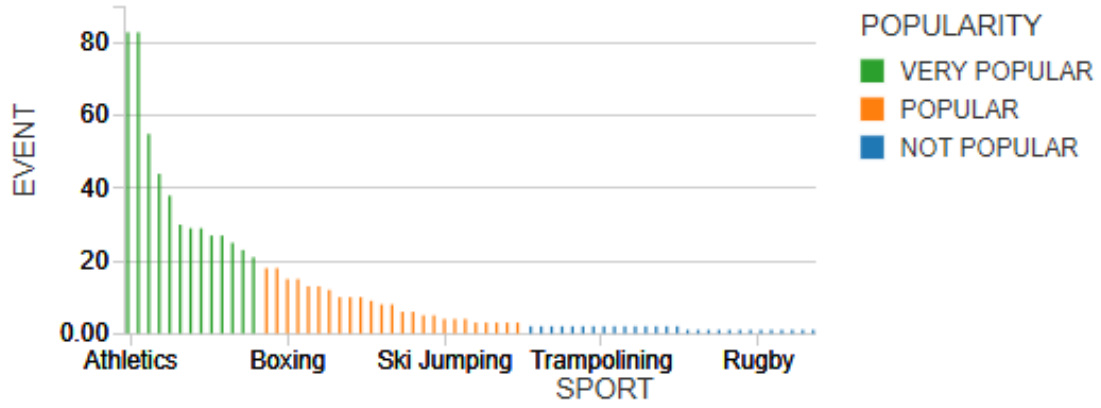
▸ (1) Spark Jobs



Plot Options...

```sql
--Popularity of sports based in number of events
SELECT *
    , CASE WHEN EVENT >=19 THEN 'VERY POPULAR'
           WHEN (EVENT<19 AND EVENT>=3) THEN 'POPULAR'
           ELSE 'NOT POPULAR' END AS POPULARITY
FROM
(SELECT SPORT, COUNT(DISTINCT EVENT) AS EVENT
FROM event
GROUP BY SPORT)
GROUP BY SPORT, EVENT
ORDER BY EVENT DESC
```

▸ (1) Spark Jobs



Plot Options...

RESULTS

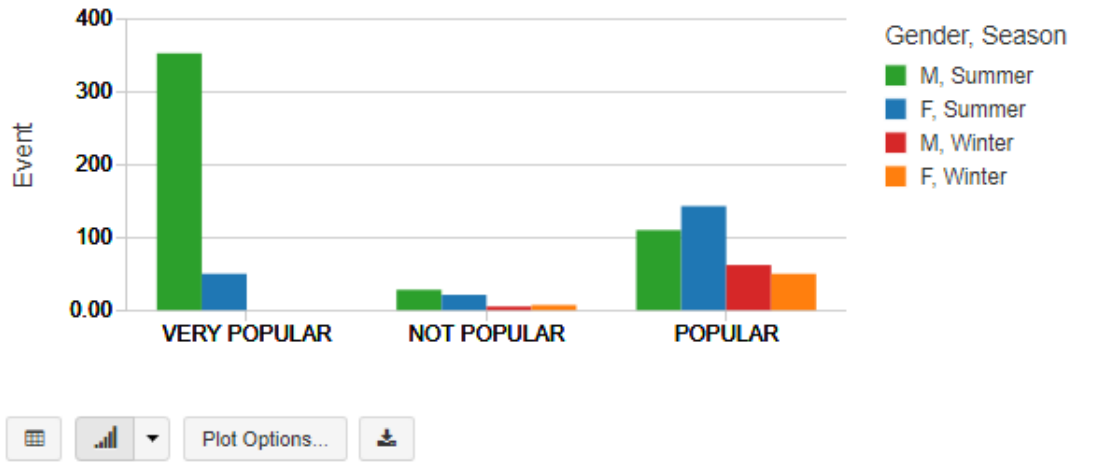
CLASSIFICATION OF SPORTS IN BASE OF THE NUMBER OF DIFFERENT EVENTS THEY HAVE

RESULTS

- There is a noticeable bias as the most popular games are from the summer olympics based on how many different events they have, and also a gender bias towards men for the very popular sports

```
--Count of events grouped by populatiry and gender and season
SELECT *
    , CASE WHEN EVENT >-19 THEN 'VERY POPULAR'
           WHEN (EVENT<19 AND EVENT>-3) THEN 'POPULAR'
           ELSE 'NOT POPULAR' END AS POPULARITY
FROM
(SELECT Sport,Sex AS Gender, Season, COUNT(DISTINCT Event) AS Event FROM athlete
LEFT JOIN game ON athlete.Gameid-game.Gameid
LEFT JOIN event ON athlete.Eventid-event.Eventid
GROUP BY Sport, Gender, Season)
```

▸ (5) Spark Jobs

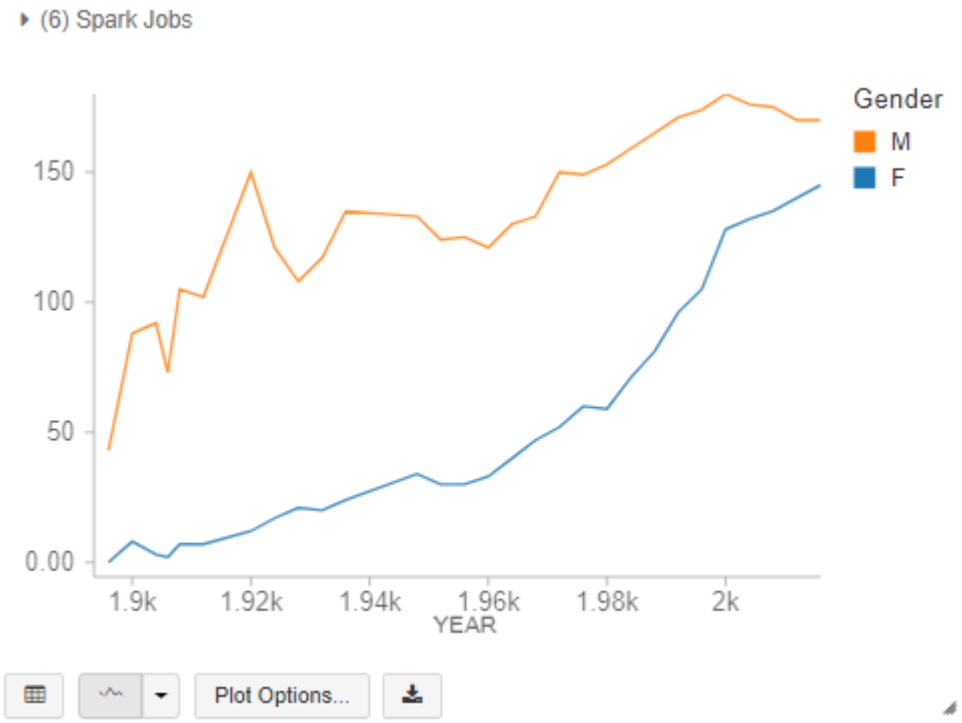


RESULTS

- A new interest arise to see if there is a gender bias in the data, for that reason, timelines of the amount of events for men and women is made to see if the participation of women and men is similar and the changes over the years
- The hypothesis of team sports vs individual sports was discarded because there was no practical method to group by those categories
- While analyzing the data it was found that there is something as a country winner of the games based in the amount of medals that country wins in each edition of the games, from this a new objective was formulated: calculate the correlation between people send by country and medals won

RESULTS

- Summer Games



- Winter Games

RESULTS

Summer Games

- There has been a big gap between the amount of events that men and women participate
- As the time goes by, the gap has been reduced nevertheless there still exists more events for men than for women

Winter Games

- There has been a gap between the amount of events that men and women participate
- As the time goes by the gap has been maintaned but it still is smaller than the one from the summer games

RESULTS

New hypothesis

- From the analysis made a new hypothesis was formulated:
  - There is a positive correlation between the amount of athletes send and the amount of medals won by country
- To follow through with the investigation a Pearson coefficient of correlation will be used to prove or disprove the hypothesis

Pearson Coefficient

- $r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$
- X = Número de personas enviadas
- Y = Número de medallas ganadas
- r > 0 → Correlación positiva
- r < 0 → Correlación negativa

Cmd 19

```
--POSITIVE CORRELATION BETWEEN PEOPLE SEND TO COMPETE AND MEDALS WON
SELECT NUM/(SQRT(DENOM1)*SQRT(DENOM2)) AS CORR
FROM
(SELECT *
    ,(n*EXY-(EX*EY)) AS NUM
    ,(n*EX2-E2X2) AS DENOM1
    ,(n*EY2-E2Y2) AS DENOM2
FROM
(SELECT
     SUM(X)         AS EX
    ,SUM(Y)         AS EY
    ,SUM(X*X)       AS EX2
    ,SUM(Y*Y)       AS EY2
    ,SUM(X*Y)       AS EXY
    ,SUM(X)*SUM(X)  AS E2X2
    ,SUM(Y)*SUM(Y)  AS E2Y2
    ,count(*) AS n
FROM
(SELECT
     People_send AS X
    ,Medals_Won  AS Y
FROM
(SELECT NOC, COUNT(ID) AS People_send, COALESCE(COUNT(Medal),0) as Medals_won FROM
athlete
LEFT JOIN team ON athlete.Teamid=team.Teamid
LEFT JOIN medal ON athlete.Medalid=medal.Medalid
GROUP BY NOC
ORDER BY COUNT(ID) DESC))))
```

▸ (2) Spark Jobs

CORR
0.9153711778341583

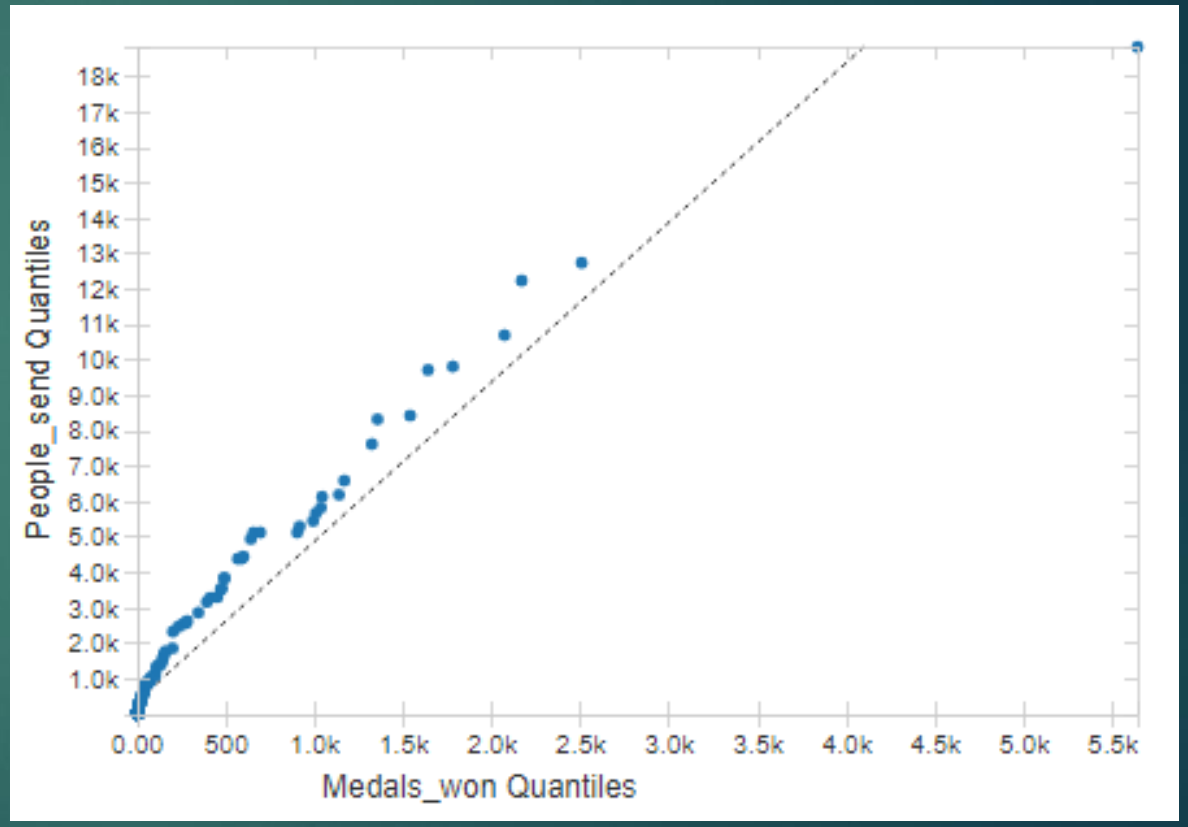
Command took 5.54 seconds -- by js.sanchez130@uniandes.edu.co at 20/5/2020 0:03:25 on Other another cluster

RESULTS

- The results show that there is a positive correlation between the amount of athletes send by country and the amount of medals won in the games
- Thus, the hypothesis is proven
- So, if a country wants to win more medals they should send more athletes
- For this analysis the fields used were a sum of the distinct athletes send by country for each game and a count of medals won by country for each game

CONCLUSIONS

- With the goal to verify if a bigger amount of athletes correspond to a bigger amount of medals won a coefficient of correlation was calculated with confirmed that in fact there is a strong correlation which can also be seen in the next graph.
- Every dot represents a different country and the amount of people send and the amount of medals won in all the games



CONCLUSIONS

- With the analysis some interesting insights have jumped out at me now. For instance how the gender breach is slowly reducing in modern times, by doing a series of graphs over time I found that women have had participated in some minor form in the olympic games since the seconth edition of the games and since then their role has increased to the point were the number of events for the summer are close to be the same as men (in the games of Rio 2016 there were 170 events for men and 145 for women and in the games of Paris in 1900 there were 88 for men and 8 for women) it is still a breach but it surely has decreased over time. On the other side, in the winter games the gap between events for men and women has been smaller (in average 8 since 1992) but it has remained almost constant since then so it very interesting to see that the gap is smaller than the one of the summer games but the tendency remains the same over time.

CONCLUSIONS

- In order to keep track of relationships between height and weight I created a metric called Height/Weight ratio to see if there are some relationship between those variables and the amount of medals won by an athlete in some sport and thus identify if there are some special height or weight that am athete should have to win medals
- The second metric I created is a popularity of the sport based on how long the sport has been an olympic sport and how many different events are for that sport. This was created because since there are 765 different events over 66 different sports its difficult to chose sports that have enough data in order to make meaningful analysis so with this metric is simpler to pick sports to make analysis.

NEXT STEPS

- Some descriptive stats are pointing me to look further into investigating if there is an specific height or weight that may correlate with getting more medals for example the mean, median and mode of those atributes but grouped by different sports, also grouped by the winners of some specific events over time. This in order to see if there is an optimal relationship for height and weight for an specific sport or even a specific evento and also I want to see if that changes over time
- Another aspect that might be interesting is to se if age matters into winning medals and a similar approach as the described above may be used to find out if there is relevant statistical evidence to prove some hypotheses