SQL for Data Science Capstone Project

By Gulchachak

03/13/2022

Table of contents:

Questions

- 1. Which countries have shown the best results since 2000 year?
  A. How many medals did they earn?
  B. What kinds of sport were their sportsmen more successful?
- 2. What is the distribution by gender?
  A. Do women get more medals last years?
  B. Which sports women are more successful?
- 3. What is the distribution by season games?
  A. Is there correlation between success in winter and summer games?

Hypothesis

- 1. Well developed countries with big population get more medals in Olympics.
- 2. There are more men participating in Olympics than women.
- 3. If country successful in winter games, it is also successful in summer games.

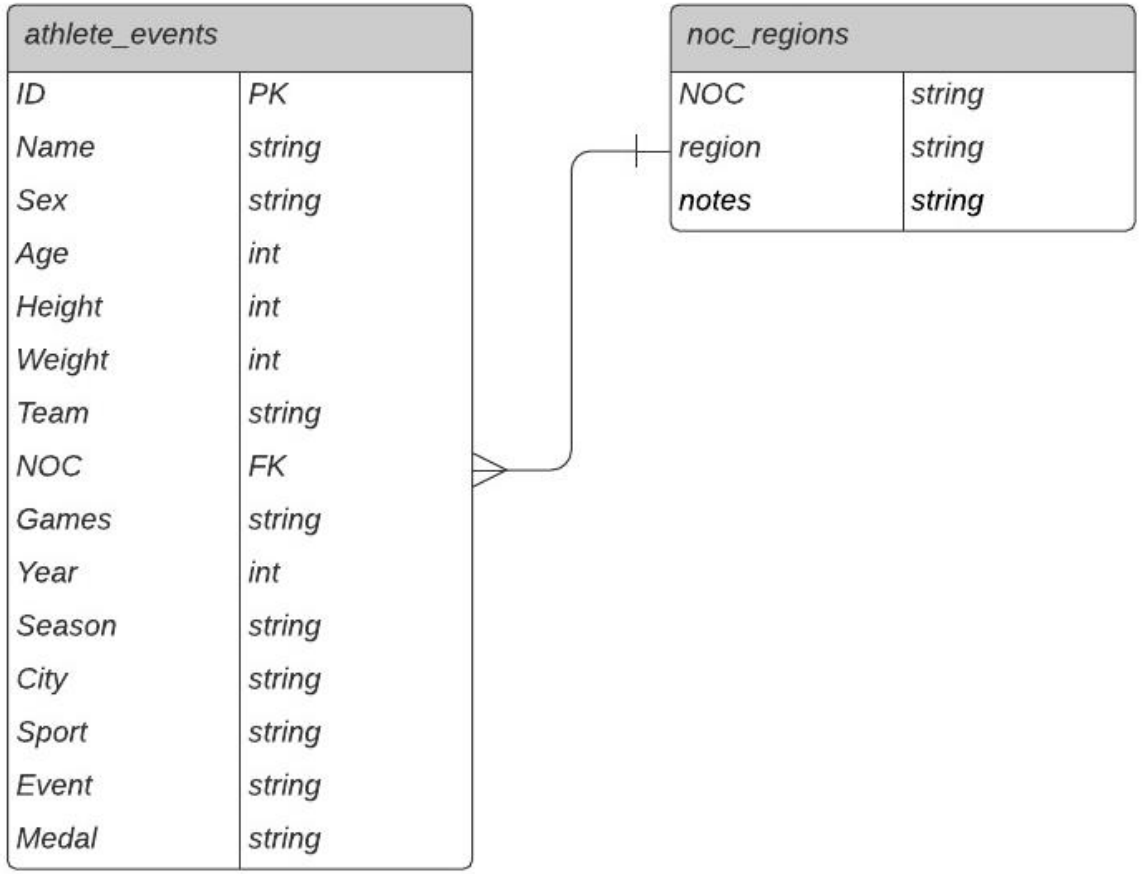
Approach

- I'll be looking primarily in distributions of different indicators (age, gender, etc).
- I'll use SQL aggregate functions, such as COUNT, SUM.
- Also, represent results in graphs or histograms.
- Look for patterns and insights distributed by year, age, gender.

Technical Challenges

- I used Databricks environment for analyzing Olympic Dataset.
- Firstly, I imported two tables athlete_events.csv and noc_regions.csv
- Then I checked data for missing information and converted some columns to integer data type (age, height, weight).

Create an ERD to show the relationships of the data you are exploring.



Initial findings

- Olympic dataset consists of a variety of data in different aspects, by year, gender, age, sport kind, etc.
- Found that initial hypotheses were evident in the data, however, then decided to broad analysis to more historical data.
- I was interested in analyzing this data in general, tried to find some interesting facts about Olympic games. For example, USA is the country that get maximum amount of medals during the history of Olympics from 1900 year (given data starts from this year), and also is getting maximum nowadays.

```
1  SELECT team, COUNT(Medal) AS counted
2  FROM athlete_events
3  WHERE NOT Medal = 'NA'
4  GROUP BY team
5  ORDER BY counted DESC
6  LIMIT 15
```

▸ (2) Spark Jobs

	team	counted
1	United States	5219
2	Soviet Union	2451
3	Germany	1984
4	Great Britain	1673
5	France	1550
6	Italy	1527
7	Sweden	1434

Showing all 15 rows.

```
1  SELECT team, COUNT(Medal) AS counted
2  FROM athlete_events
3  WHERE (year BETWEEN 2000 AND 2016) AND NOT Medal = 'NA'
4  GROUP BY team
5  ORDER BY counted DESC
```

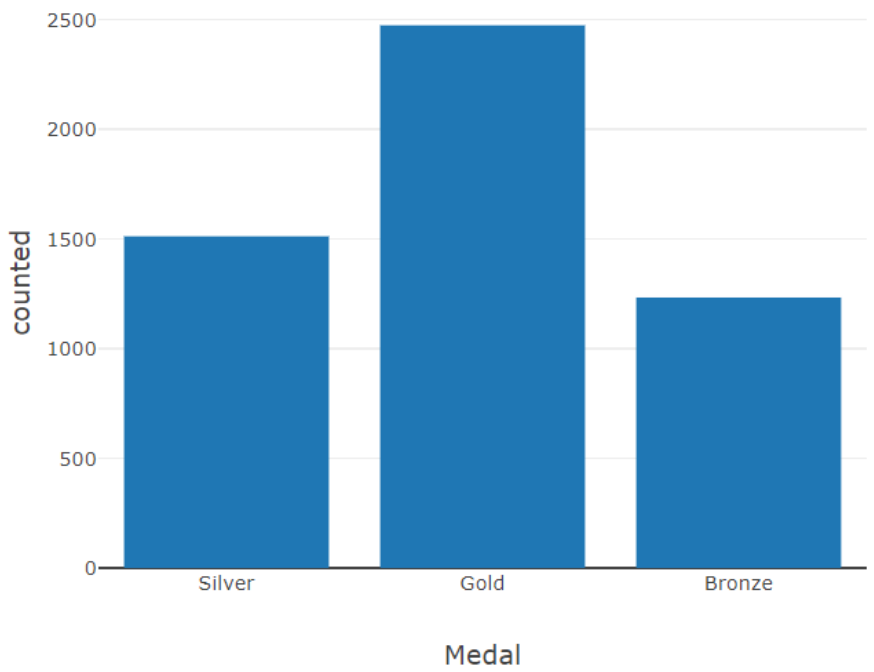
▸ (2) Spark Jobs

	team	counted
1	United States	1561
2	Russia	913
3	Germany	775
4	Australia	693
5	China	593
6	Canada	508
7	Great Britain	473

Showing all 164 rows.

Deeper analysis

- Best results was considered as the amount of medals during Olympics.



Kind of medals distribution of US team during the 1900-2016.

```
1  SELECT Team, Sport, COUNT(Sport) AS counted_sport
2  FROM athlete_events
3  WHERE team = 'United States' AND NOT Medal = 'NA'
4  GROUP BY Sport, Team
5  ORDER BY counted_sport DESC
6  LIMIT 5
```

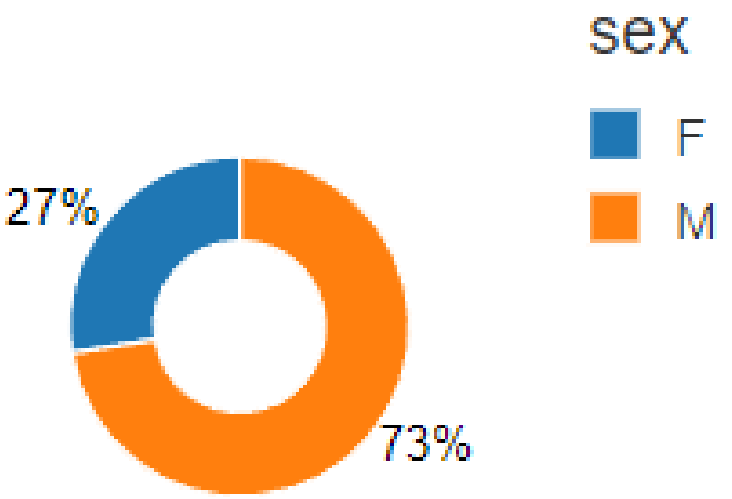
▸ (2) Spark Jobs

	team	Sport	counted_sport
1	United States	Athletics	1071
2	United States	Swimming	1066
3	United States	Basketball	341
4	United States	Rowing	333
5	United States	Ice Hockey	276

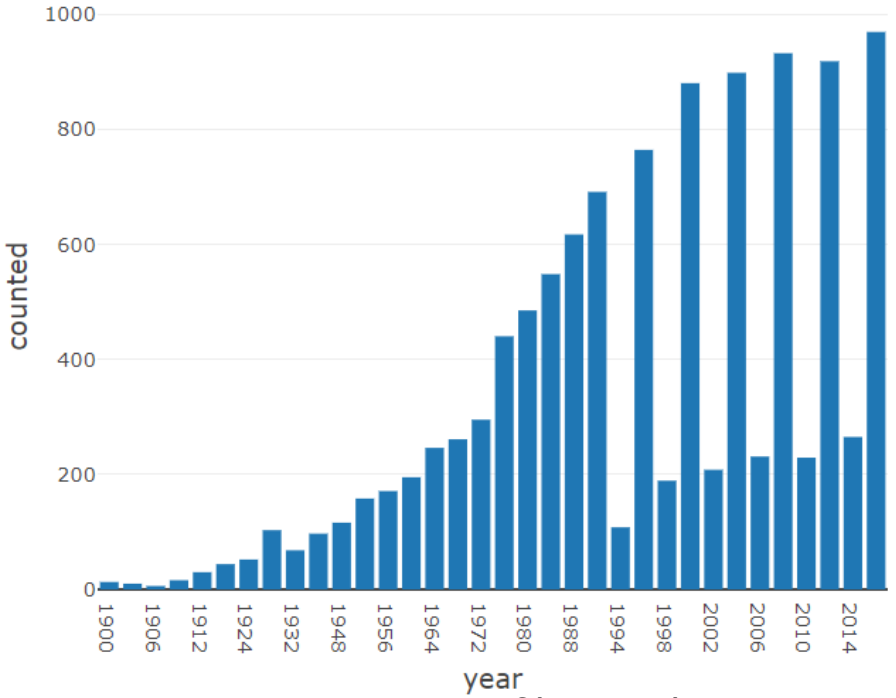
Successful kinds of sport of US team during the 1900-2016.

Deeper analysis

- What is the distribution by gender?



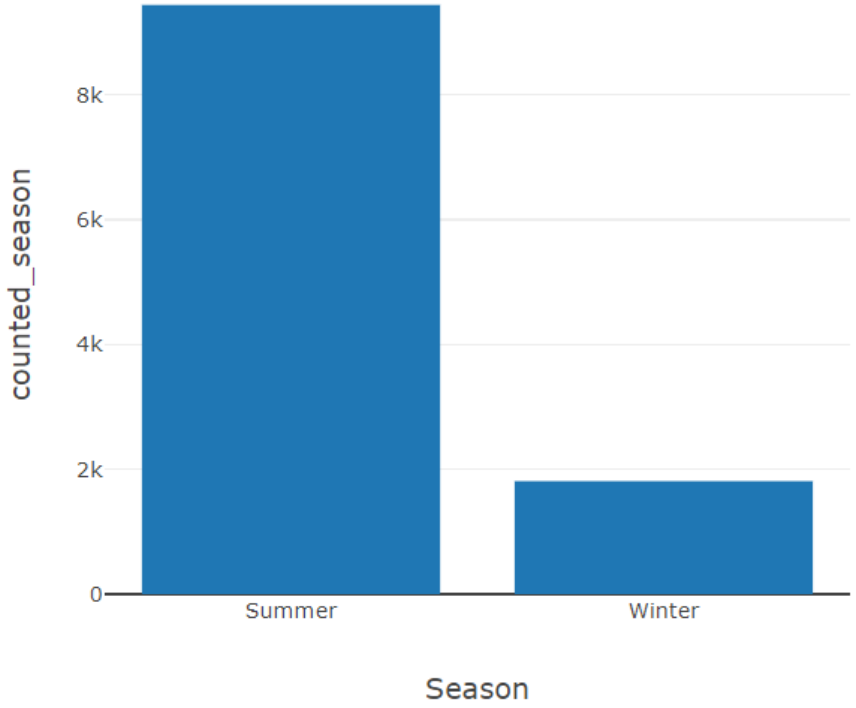
Female and male participants distribution in Olympics.

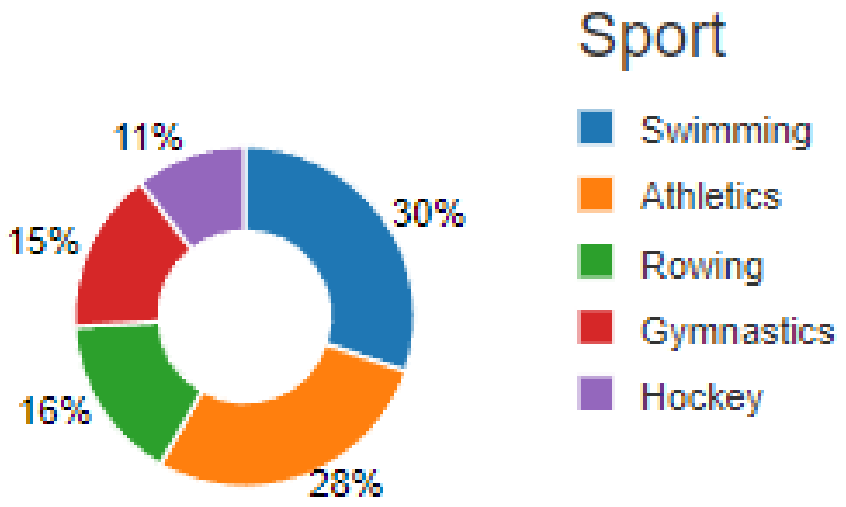


Women participation in Olympics by years.

Deeper analysis

- Women participation distribution by season and most successful sports.





Final findings (Results of Hypotheses)

- 1. Well developed countries with big population get more medals in Olympics.

United States get more medals in Olympics, it might be because they are able to invest more to sportsmen, sport equipment, and make training process more successful.

- 2. There are more men participating in Olympics than women.

This assumption proved, because there are more men's kinds of sports in Olympics, while women's sports still is being added every year.

- 3. Women more succeeded in summer games.

I found that the most popular sports in which women succeeded more: swimming, athletics, rowing, gymnastics and hockey, which are summer games.

Conclusion

- In this project I analyzed SportsStats data from Olympics Dataset.
- I think that people who are interested in sport, sportsmen, coaches might find insights and conclusions very interesting. As in Olympics there different kinds of sports presented, every person might know some new information. Sportsmen can observe how results are changing by time, by gender, new top results. Coaches might find how they could improve their effectiveness by analyzing competitive teams.