Milestone 1: Project Proposal and Data Selection/Preparation

Client 3: SportsStats (Olympics Dataset - 120 years of data)

SportsStats 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.

Step 1: Preparing for Your Proposal

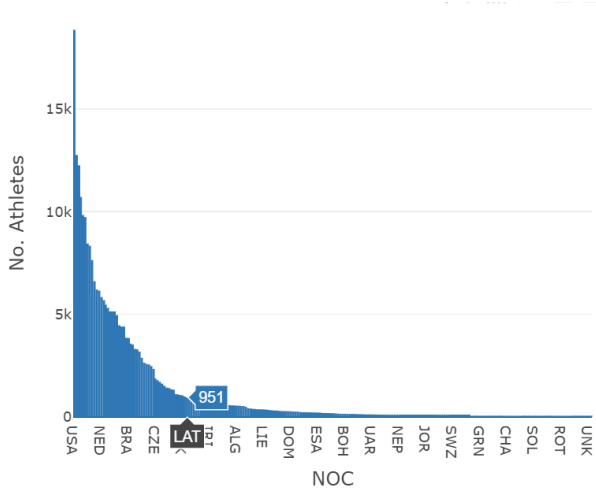
1. I have chosen to use the Olympics dataset due to problems importing the lobbyists dataset into Databricks.
2. I uploaded the two CSV files into Databricks Community Edition using the create table API. I allowed Databricks to infer the schema, manually correcting three of the columns from string to integer during the table creation process.
3. Data exploration:

```
-- Display number of athletes per region
SELECT
  NOC,
  COUNT(ID) AS `No. Athletes`
FROM
  athlete_events
GROUP BY
   NOC
ORDER BY
  `No. Athletes` DESC
```

▸ (2) Spark Jobs

	NOC	No. Athletes
1	USA	18853
2	FRA	12758
3	GBR	12256
4	ITA	10715
5	GER	9830
6	CAN	9733
7	JPN	8444

Showing all 230 rows.

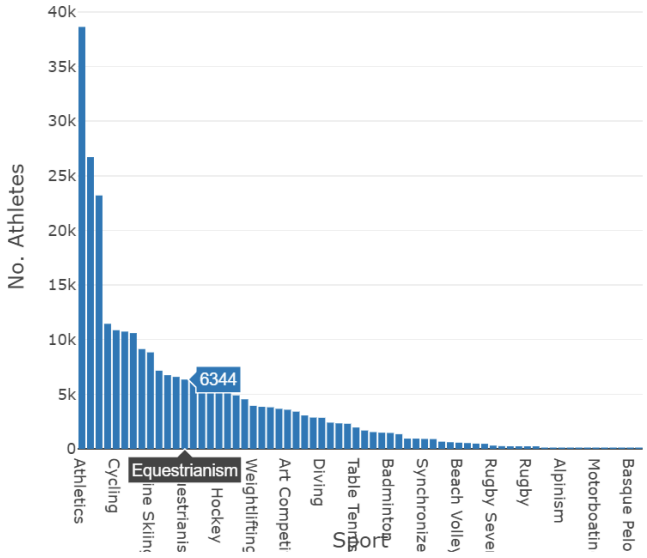



```
-- Display number of athletes per sport
SELECT
  Sport,
  COUNT(ID) AS `No. Athletes`
FROM
  athlete_events
GROUP BY
   Sport
ORDER BY
  `No. Athletes` DESC
```

▸ (2) Spark Jobs

	Sport	No. Athletes
1	Athletics	38624
2	Gymnastics	26707
3	Swimming	23195
4	Shooting	11448
5	Cycling	10859
6	Fencing	10735
7	Rowing	10595

Showing all 66 rows.




```sql
-- Find earliest and latest dates
-- and total number of years covered
SELECT
  MIN(Year) AS `Earliest Year`,
  MAX(Year) AS `Latest Year`,
  MAX(Year) - MIN(Year) AS `Total Years`
FROM athlete_events
```

▸ (2) Spark Jobs

	Earliest Year	Latest Year	Total Years
1	1896	2016	120

```sql
-- Total number of men and women
SELECT
  Sex,
  COUNT(ID) AS `No. Athletes`
FROM
  athlete_events
GROUP BY
   Sex
ORDER BY
  `No. Athletes` DESC
```

▸ (2) Spark Jobs

	Sex	No. Athletes
1	M	196594
2	F	74522

```sql
-- Age, height and weight spread
SELECT
  MIN(Age) AS `Youngest Athlete`,
  MAX(Age) AS `Oldest Athlete`,
  MIN(Height) AS `Shortest Athlete`,
  MAX(Height) AS `Tallest Athlete`,
  MIN(Weight) AS `Lightest Athlete`,
  MAX(Weight) AS `Heaviest Athlete`
FROM athlete_events
```

▸ (2) Spark Jobs

	Youngest Athlete	Oldest Athlete	Shortest Athlete	Tallest Athlete	Lightest Athlete	Heaviest Athlete
1	10	97	127	226	25	214

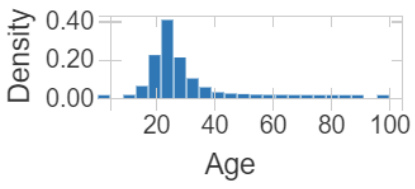
```sql
--Average age, height and weight
SELECT
  ROUND(AVG(Age),2) AS `Average Age`,
  ROUND(AVG(Height),2) AS `Average Height`,
  ROUND(AVG(Weight),2) AS `Average Weight`
FROM athlete_events
```

▸ (2) Spark Jobs

	Average Age	Average Height	Average Weight
1	25.56	175.34	70.68

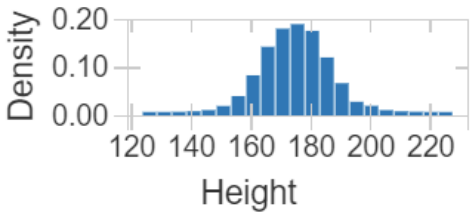
```sql
-- Histogram of ages
SELECT
  Age
FROM
  athlete_events
```

▸ (5) Spark Jobs



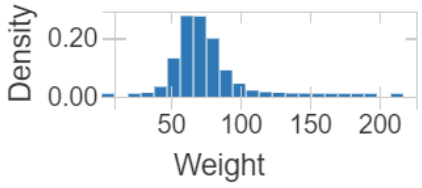
```sql
-- Histogram of heights
SELECT
  Height
FROM
  athlete_events
```

▸ (5) Spark Jobs




```sql
-- Histogram of weights
SELECT
  Weight
FROM
  athlete_events
```

▸ (5) Spark Jobs



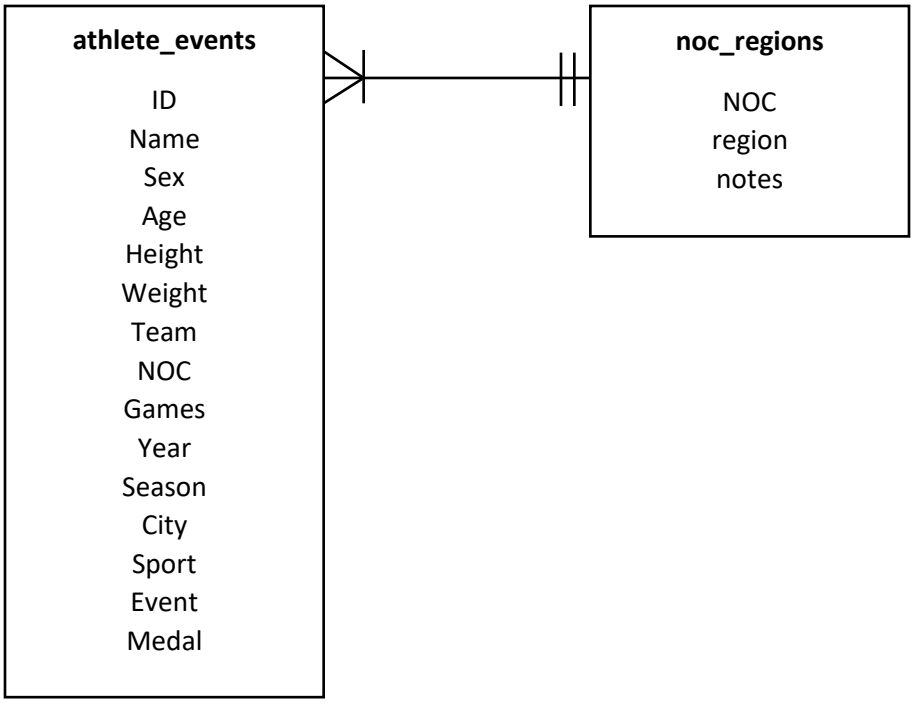
```sql
-- Number of medals by sex
SELECT
  Medal,
  Sex,
  COUNT(ID) AS `No. Athletes`
FROM
  athlete_events
GROUP BY
   Medal,
   Sex
ORDER BY CASE WHEN Medal = 'Gold' THEN '1'
        WHEN Medal = 'Silver' THEN '2'
        WHEN Medal = 'Bronze' THEN '3'
        ELSE Medal END,
        Sex
```

▸ (2) Spark Jobs

	Medal	Sex	No. Athletes
1	Gold	F	3747
2	Gold	M	9625
3	Silver	F	3735
4	Silver	M	9381
5	Bronze	F	3771
6	Bronze	M	9524
7	NA	F	63269

Showing all 8 rows.

4. ERD Diagram



Project Introduction

This project looks at the Olympics dataset, containing information on athletes competing at the Olympics over a 120-year period. The dataset includes demographic information such as age, sex, and country, information about the Olympics in question such as when it occurred and whether it was a summer or winter Olympics, and information about the performance of the athlete including which sport and what medal they achieved. The results of this project would be of interest to any companies in the sports industry, in order for them to gain insights into potential business opportunities.

Questions to Answer

1. Which sports are most popular in which countries?
2. Which sports are most popular today?
3. Which sports are most popular by gender?
4. Are there any nations that are underrepresented at the Olympics?
5. Does age affect performance?

Initial Hypothesis

1. Less wealthy nations will have fewer athletes competing at the Olympics
2. Wealthier nations will receive more medals than less wealthy nations
3. Younger athletes will receive more medals than older athletes

Approach

- I will be looking at counts of athletes by nation, counts of medals by nation, proportional number of medals per nation based on the number of athletes, average age by medal, spread of ages by medal, and more.