PROJECT PROPOSAL FOR SPORTSSTAS

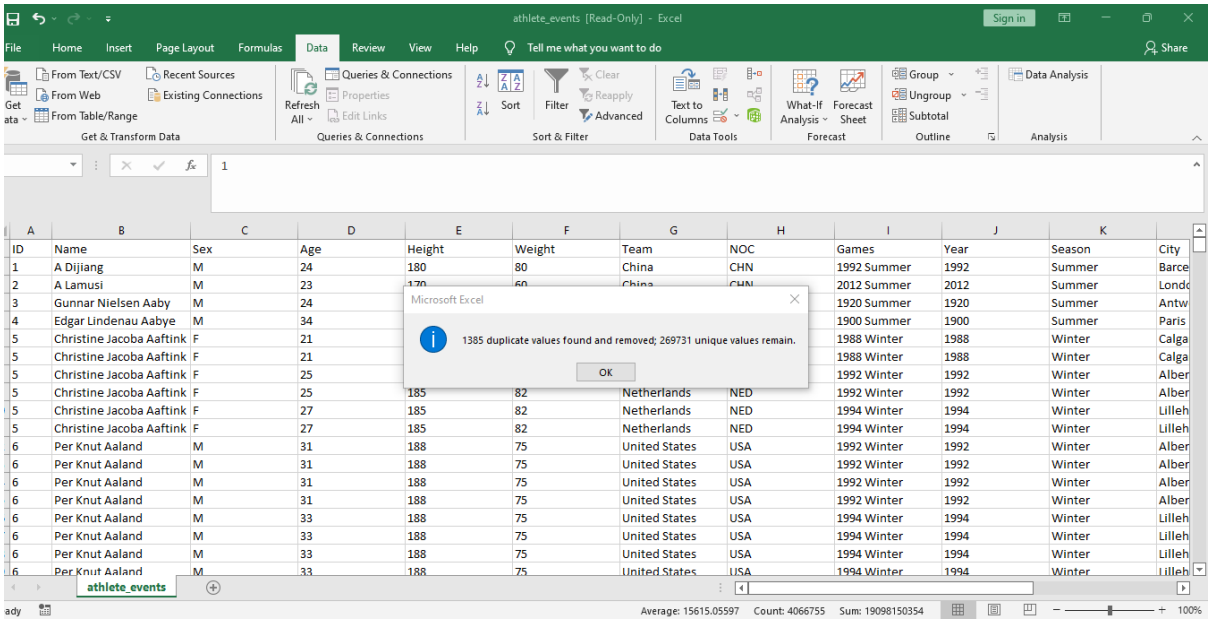
(Olympics Dataset - 120 years of data)

--Which Client and why

SportsStats is a sports analysis firm partnering with local news and elite personal trainers to provide “interesting” insights to help their partners.
As a data scientist I could easily identify patterns/trends highlighting certain groups/events/countries. I could develop insight and analysis for the purpose of developing a news story or discovering key health insights.

--Steps to import and clean the data

I imported the data into an excel sheet to explore and understand my data. A total of 271,116 records. I aligned all my data, removed duplicates (1385 records) leaving a total of 269,731 when importing it into Databricks.



--Data Exploration

Databricks analysis

```
--Total count after removing duplicates
SELECT count(*)
FROM `sportsstats_1_csv`
```

	count(1)
1	269731

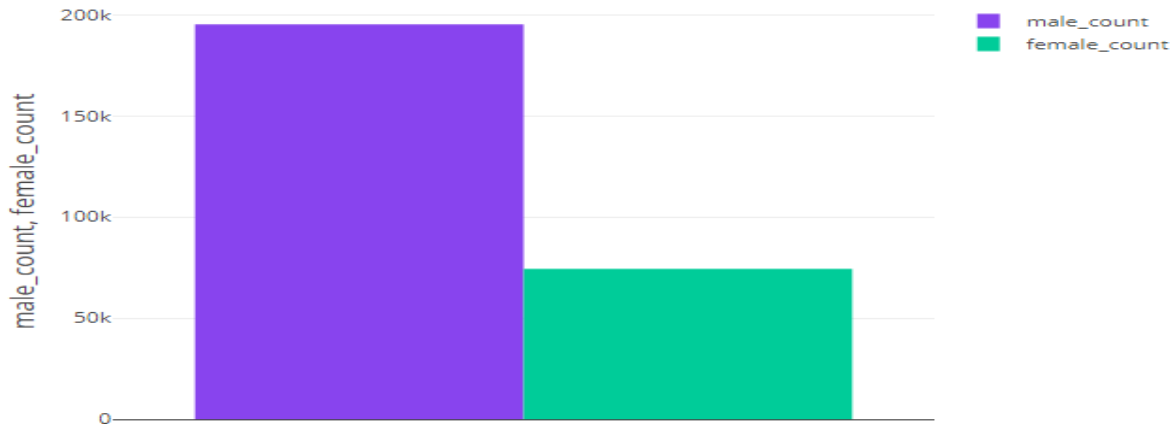
Showing all 1 rows.

When initially exploring the data, I wanted to know the number of male and female that has participated in the Olympics over the years.

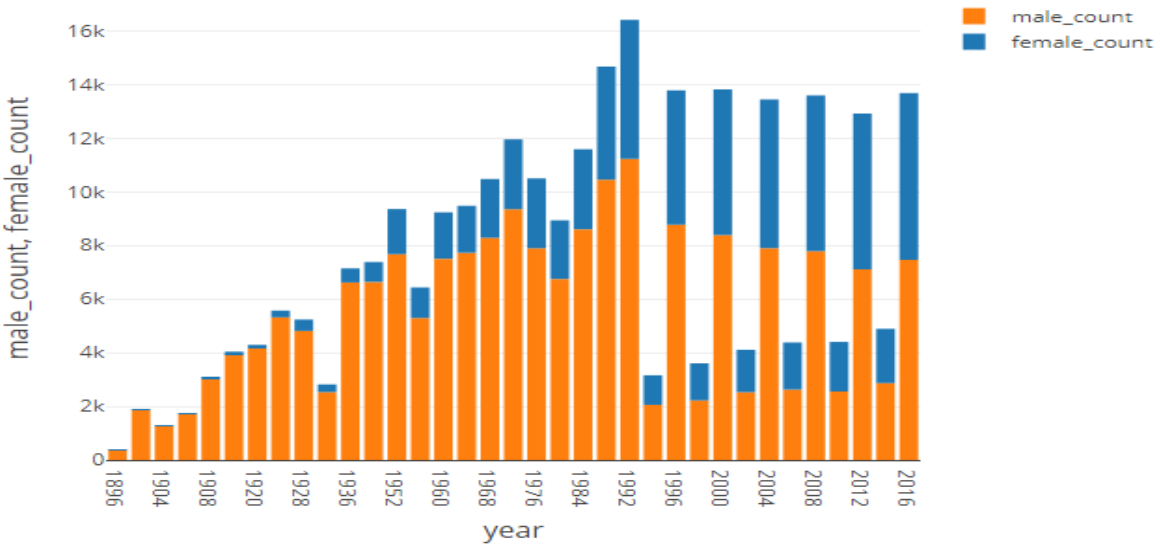
PROJECT PROPOSAL FOR SPORTSSTAS

(Olympics Dataset - 120 years of data)

```
--To count male and female that have ever participated
SELECT
  count(case when Sex = 'M' then 1 end) as male_count,
  count(case when Sex = 'F' then 1 end) as female_count
 FROM `sportsstats_1_csv`
```



I had to look at the number of male and female that has participated in the Olympics per year.

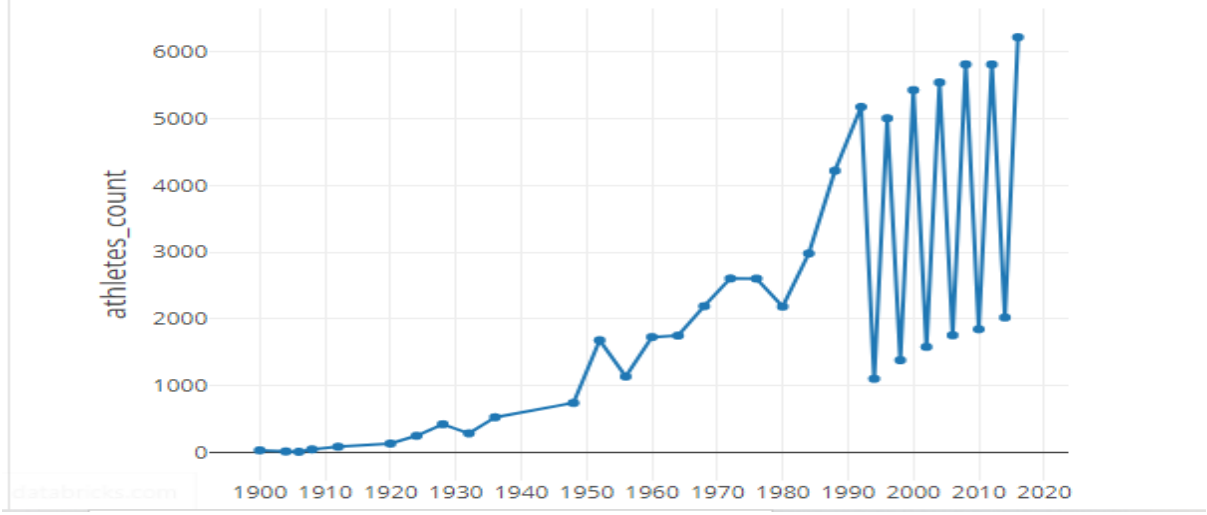


So, I had to take a look of only female athletes over the years and we can notice a trend of increase except from drastically changes in years 1994, 1998, 2002, 2006, 2010, 2014.

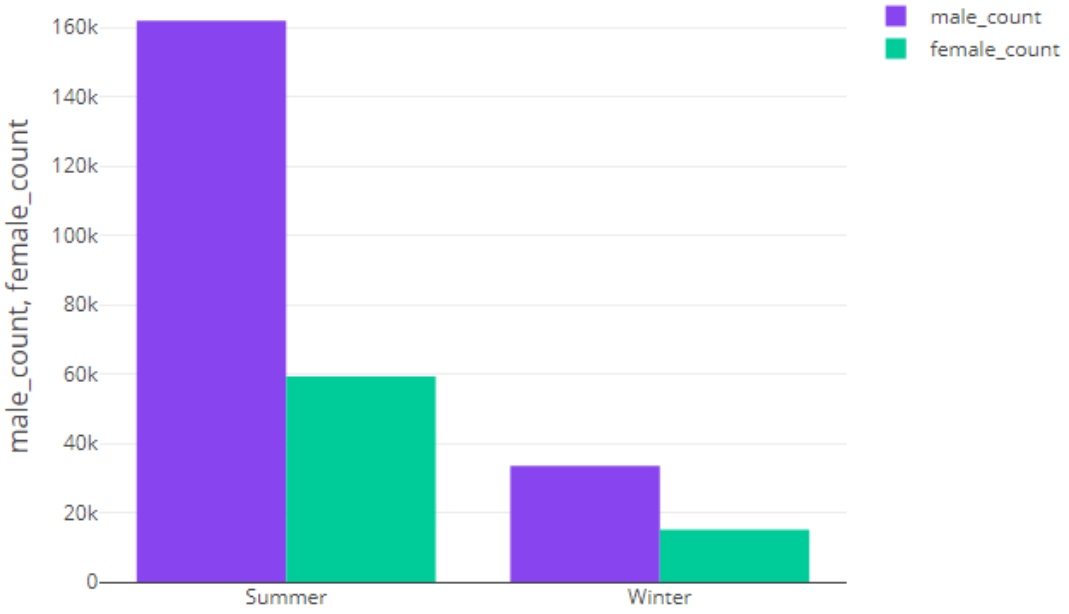
PROJECT PROPOSAL FOR SPORTSSTAS

(Olympics Dataset - 120 years of data)

```
--To view trends in female athelete per year
SELECT year, COUNT(*) AS athletes_count
FROM `sportsstats_1_csv`
GROUP BY year, sex
HAVING sex = 'F'
ORDER BY year
```



Next, I sought for possible options for the decline in female athletes and considered the seasons. Note these changes can be seen in male athletes also.
There are more athletes in the summer season than winter.

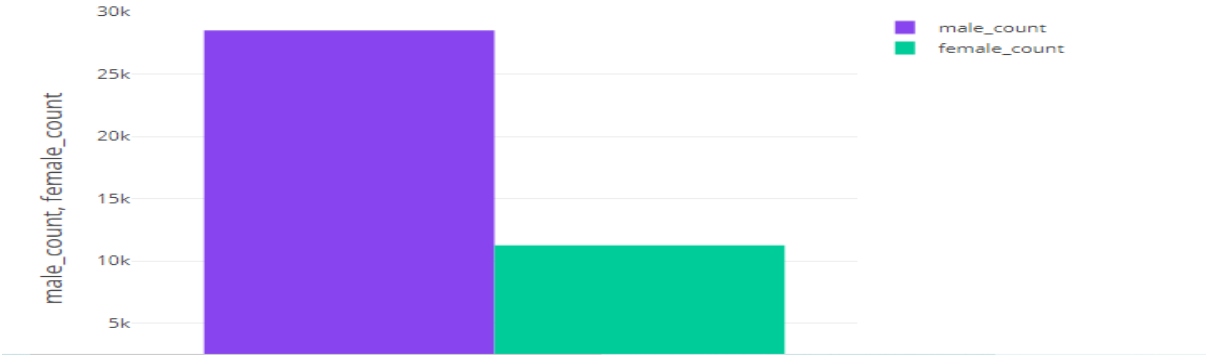


Since there are more male than female it makes more sense that more male had won more medals than females.

PROJECT PROPOSAL FOR SPORTSSTAS

(Olympics Dataset - 120 years of data)

```
--To count male and female that have ever won a medal
SELECT
  count(case when Sex = 'M' then 1 end) as male_count,
  count(case when Sex = 'F' then 1 end) as female_count
FROM `sportsstats_1_csv`
WHERE Medal <> 'NA'
```



To find a list of the youngest female athletes.

```
--Who are the youngest female athletes to ever compete?

SELECT DISTINCT name,
  age, sport, event, noc, games, city, medal
FROM `sportsstats_1_csv`
WHERE age <> 'NA' AND sex ='F'
ORDER BY age
```

	name	age	sport	event
1	Sonja Henie (-Topping, -Gardiner, -Onstad)	11	Figure Skating	Figure Skating Women's Singles
2	Liana Vicens	11	Swimming	Swimming Women's 200 metres Individual Medley
3	Luigina Giavotti	11	Gymnastics	Gymnastics Women's Team All-Around
4	Marcelle Matthews	11	Figure Skating	Figure Skating Mixed Pairs
5	Megan Olwen Devenish Taylor (-Mandeville-Ellis)	11	Figure Skating	Figure Skating Women's Singles
6	Liu Luyang	11	Figure Skating	Figure Skating Mixed Ice Dancing
7	Liana Vicens	11	Swimming	Swimming Women's 100 metres Breaststroke

To find a list of the oldest female athletes to ever compete.

```
--Who are the top 5 oldest female athletes to ever compete?
SELECT DISTINCT name,
  age, sport, event, noc, games, city, medal
FROM `sportsstats_1_csv`
WHERE age <> 'NA' AND sex ='F'
ORDER BY age DESC
LIMIT 5
```

	name	age	sport	event
1	Ernestine Lonie Ernesta Robert-Mrignac	74	Art Competitions	Art Competitions Mixed Sculpturing
2	Anne Marie Carl-Nielsen (Brodersen-)	73	Art Competitions	Art Competitions Mixed Sculpturing, Unknown Event
3	Winifred Marie Louise Austen (-Frick)	72	Art Competitions	Art Competitions Mixed Painting, Graphic Arts
4	Winifred Marie Louise Austen (-Frick)	72	Art Competitions	Art Competitions Mixed Painting, Unknown Event
5	Laura Knight (Johnson-)	70	Art Competitions	Art Competitions Mixed Painting, Graphic Arts

To find a list of the oldest female athletes to ever compete where event was not art competition.

PROJECT PROPOSAL FOR SPORTSSTAS

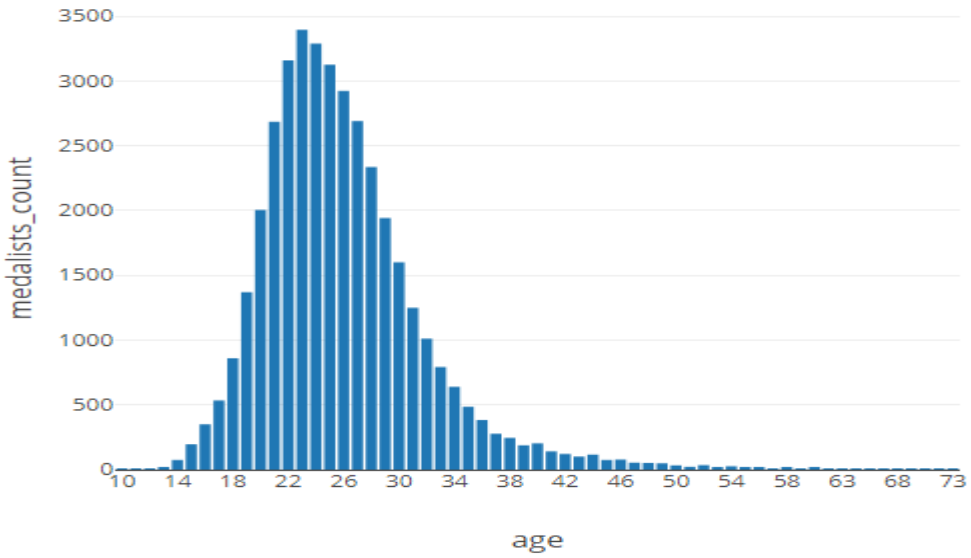
(Olympics Dataset - 120 years of data)

```
--Who are the oldest female athletes to ever compete and sport was not equal to Art?
SELECT DISTINCT name,
  age, sport, event, noc, games, city, medal
FROM `sportsstats_1_csv`
WHERE age <> 'NA' AND sport <> 'Art Competitions' AND sex ='F'
ORDER BY age DESC
```

	name	age	sport	event
1	Hilda Lorna Johnstone (Wailes-Fairbairn-)	69	Equestrianism	Equestrianism Mixed Dressage, Team
2	Hilda Lorna Johnstone (Wailes-Fairbairn-)	69	Equestrianism	Equestrianism Mixed Dressage, Individual
3	Hilda Lorna Johnstone (Wailes-Fairbairn-)	66	Equestrianism	Equestrianism Mixed Dressage, Team
4	Hilda Lorna Johnstone (Wailes-Fairbairn-)	66	Equestrianism	Equestrianism Mixed Dressage, Individual
5	Brenda Lilian Williams (Mander-, Hickman-)	65	Equestrianism	Equestrianism Mixed Dressage, Individual
6	Kikuko Inoue (Basugi-)	63	Equestrianism	Equestrianism Mixed Dressage, Individual
7	Lida Peyton "Eliza" Pollock (McMillen-)	63	Archery	Archery Women's Team Round

Truncated results, showing first 1000 rows

To find the age distribution with most medal won. Here we can see that most medals are won at the early twenties and early thirties, with the average age at 23.



Next, to find the top 5 sports in which female athletes have won the most medal.

```
--Top 5 sports in which female athletes have won the most medals.

SELECT sport, COUNT(medal) AS medal_count
FROM `sportsstats_1_csv`
WHERE sex = 'F' AND Medal <> 'NA'
GROUP BY sport
ORDER BY medal_count DESC
LIMIT 5
```

	sport	medal_count
1	Swimming	1374
2	Athletics	1275
3	Rowing	720
4	Gymnastics	701
5	Hockey	478

PROJECT PROPOSAL FOR SPORTSSTAS

(Olympics Dataset - 120 years of data)

```
--Number of athlete that has ever participated per country
SELECT team, count(Team)
FROM `sportsstats_1_csv`
GROUP BY team
ORDER BY count(team) DESC
```

	team	count(Team)
1	United States	17598
2	France	11817
3	Great Britain	11264
4	Italy	10213
5	Germany	9230
6	Canada	9226
7	Japan	8269

Truncated results, showing first 1000 rows.

```
--Number of female athlete that has participated per country
SELECT team, count(team) as number_of_female_participant
FROM `sportsstats_1_csv`
WHERE sex = 'F'
GROUP BY team
ORDER BY count(team) DESC
```

	team	number_of_female_participant
1	United States	5324
2	Canada	3473
3	Great Britain	3174
4	Germany	2906
5	Japan	2727
6	France	2723
7	Australia	2697

Showing all 374 rows.

```
--Top countries in terms of the total number of female medal winners
SELECT team, count(medal) as medal_count
FROM `sportsstats_1_csv`
WHERE medal <> 'NA' AND sex ='F'
GROUP BY team
ORDER BY count(medal) DESC
lIMIT 10
```

	team	medal_count
1	United States	1756
2	Soviet Union	667
3	Germany	657
4	China	600
5	Australia	543
6	Russia	533
7	Canada	505

Showing all 10 rows.

PROJECT PROPOSAL FOR SPORTSSTAS

(Olympics Dataset - 120 years of data)

```
--Highest number of medal won by a female athlete

SELECT name, noc, sport, count(medal) as medal_count
FROM `sportsstats_1_csv`
WHERE medal <> 'NA' AND Sex = 'F'
GROUP BY name, sex, noc, sport
ORDER BY medal_count DESC
LIMIT 10
```

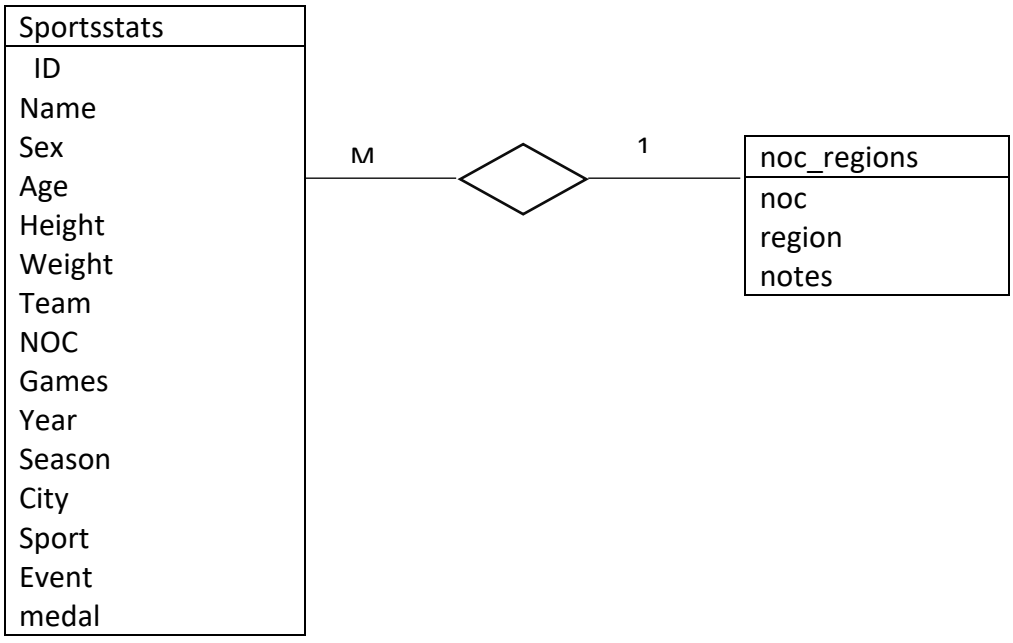
	name	noc	sport	medal_count
1	Larysa Semenivna Latynina (Diriy-)	URS	Gymnastics	18
2	Dara Grace Torres (-Hoffman, -Minas)	USA	Swimming	12
3	Natalie Anne Coughlin (-Hall)	USA	Swimming	12
4	Jennifer Elisabeth "Jenny" Thompson (-Cumpelik)	USA	Swimming	12
5	Vra slavsk (-Odloilov)	TCH	Gymnastics	11
6	Franziska van Almsick	GER	Swimming	10
7	Yang Yang	CHN	Short Track Speed Skating	10

Showing all 10 rows.

This concludes the exploratory data analysis. It is mainly to get an idea of top performing female team which will be used to answer subsequent questions over the course.

There are 2 tables in the dataset, the sportsstas and noc_region. The relationship of the datasets can be said as many to one. This is represented in the ER diagram below.

-- Develop an Entity Relationship Diagram (ERD)



PROJECT PROPOSAL FOR SPORTSSTAS

(Olympics Dataset - 120 years of data)

DEVELOPMENT OF PROJECT PROPOSAL

My analysis of the dataset is catered to the news media. My target audience is sports and history buffs like me who are interested in going down a memory lane and understanding how the top 4 medal-winning countries in history were able to win, what problems they faced, as well as any interesting patterns that emerge along the way. I believe this would make for an interesting read and quite possibly turn out to be something useful even for analysts who work for athletes, who could make use of the findings to finetune their strategies for current athletes who are competing to win a medal.

QUESTIONS:

1. While looking at the graph obtained during the exploratory data analysis, there is an up-down periodic rise and slump in the number of medals won by the top 4 teams. What is the reason for this trend? This is something that I am looking to answer.

2. Is there a relationship between sex and sport? For example, are men known to predominantly win in a specific sport and women in another sport? If yes, why the discrepancy? To be concise, is there any relation between the sport that is played and the more dominant sex in that game?

3. What impact does the season have on the medal-winning count on the country? Does summer mean more medals and winters mean less? Is there a trend associated with it?

HYPOTHESIS:

1. For the first hypothesis, I will be looking at the relationship between season and number of athletes.

2. For the second hypothesis, I will at the relationship between age and medal won.

APPROACH:

1. For the first hypothesis, I will be looking at the relationship between season and number of athletes. I was able to detect that there were more athlete participating in summer than winter.

2. For the second hypothesis, I will at the relationship between age and medal won. I was able to conclude that most medals were won between early twenties and late twenties.