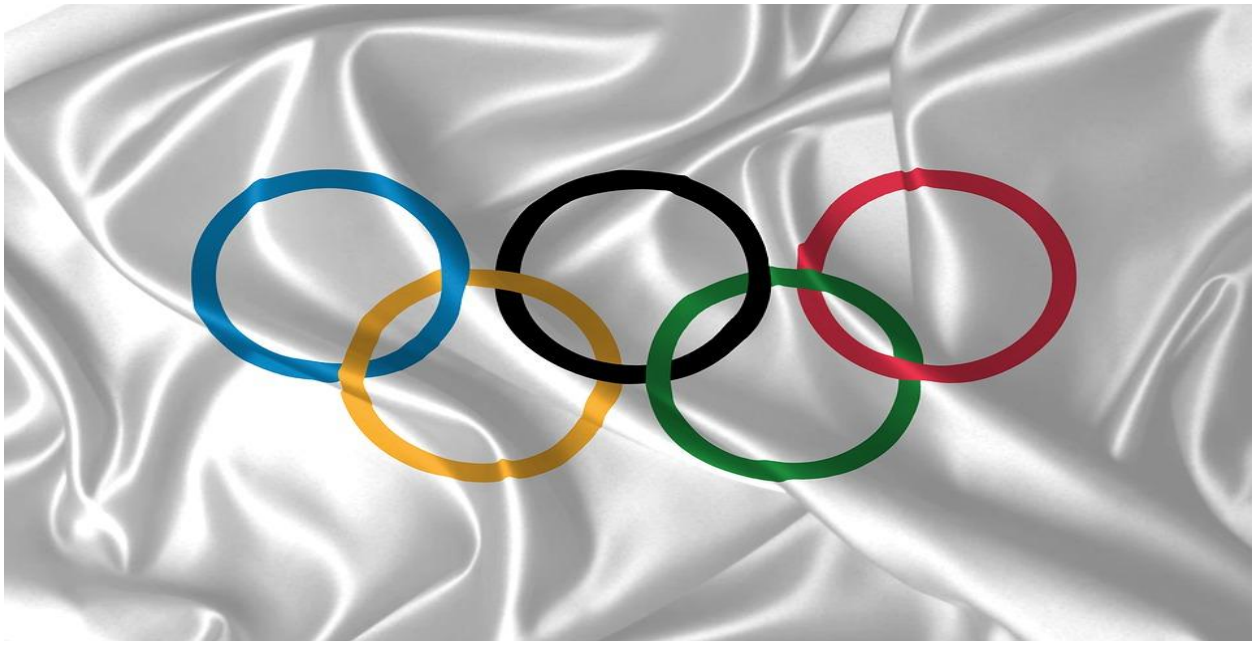


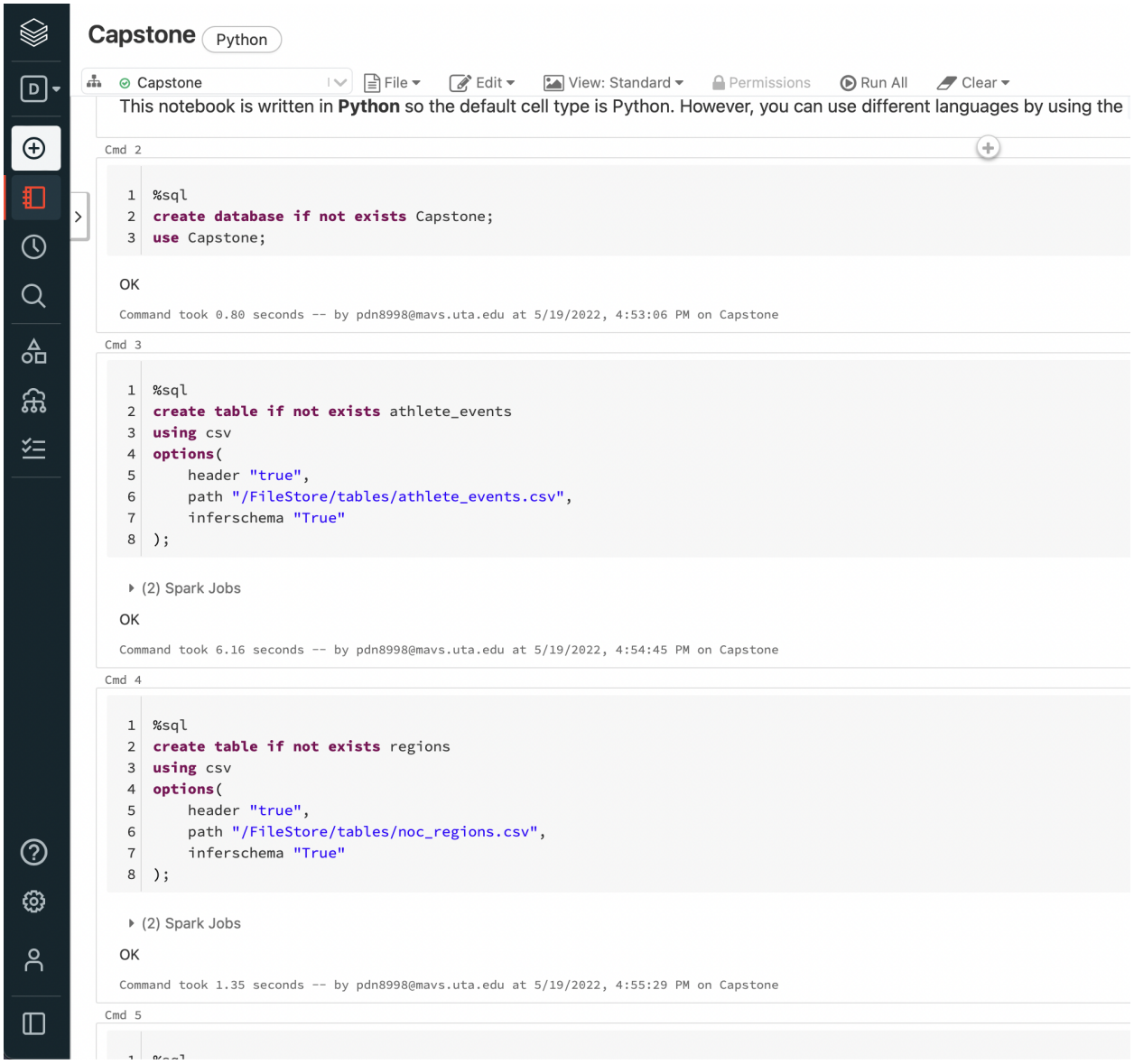
MILESTONE - 1
CAPSTONE PROJECT PROPOSAL - SPORTSTATS: PREPARING THE DATA

The dataset I selected for this project is the Sportstats (Olympics dataset); I chose this because it has a lot of potential throughout the sports industry impacting media, sports buffs, and organizations helping players excel in sports. Insights from this would help me understand trends in the Olympics for many decades. As someone who enjoys athletics, I found this dataset fun to use for my analysis.

- The dataset was in CSV format, easier to view, upload and work with.

I have imported the dataset on Databricks in the CSV file format and observed a certain number of records. I have uploaded the athletic_events dataset that consists of records pertaining to the player details and the regions, medals won, and the teams; the second table consists of the country and the region.

- The dataset also has a lot of ‘NA’ values that have not been eliminated as this would impact the dataset, but the ‘NA’ values have been taken care of while doing the analysis.
- I created a database on data bricks and imported both tables.

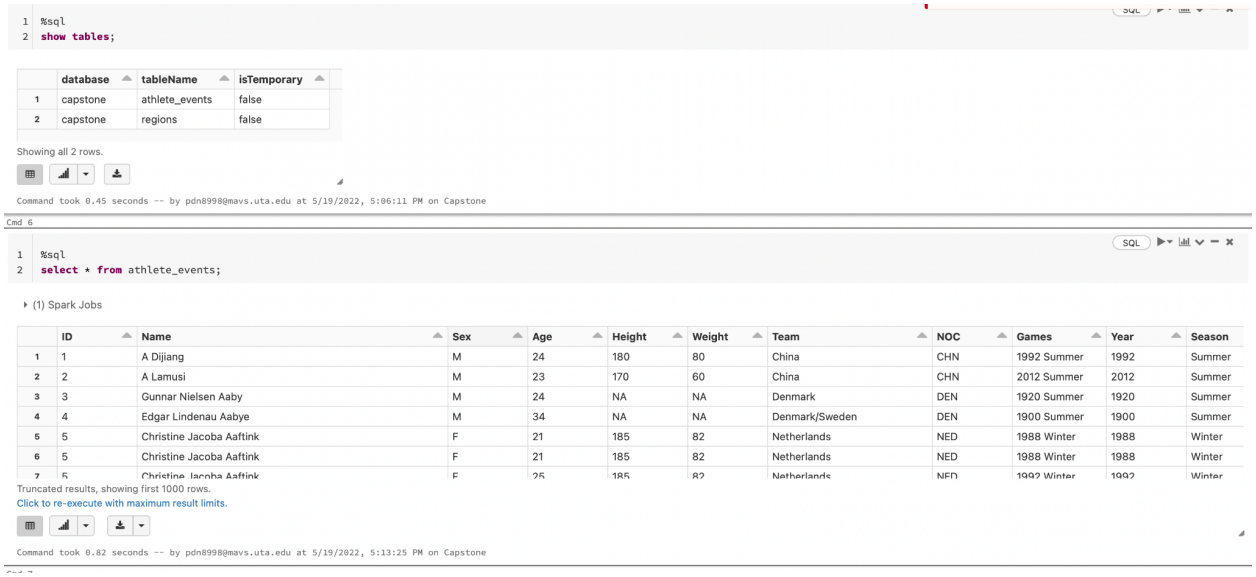


SUMMARY OF THE DATA

Sex: Females (74357), Males (196086)
Athletes: 271116 (Total)
Age : 80 Unique values
Teams : 1246 Unique values
NOC : 278 Unique values
Events : 811 events
Sports : 108 Values
Medals : 3 Unique medal (Gold, Silver, Bronze)

PREPARING FOR PROJECT PROPOSAL

Some observations have been made, such as follows :



```
%sql
show tables;
```

	database	tableName	isTemporary
1	capstone	athlete_events	false
2	capstone	regions	false

Showing all 2 rows.

Command took 0.45 seconds -- by pdn8998@mavs.uta.edu at 5/19/2022, 5:06:11 PM on Capstone

Cmd 6

```
%sql
select * from athlete_events;
```

(1) Spark Jobs

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season
1	1	A Dijiang	M	24	180	80	China	CHN	1992 Summer	1992	Summer
2	2	A Lamusi	M	23	170	60	China	CHN	2012 Summer	2012	Summer
3	3	Gunnar Nielsen Aaby	M	24	NA	NA	Denmark	DEN	1920 Summer	1920	Summer
4	4	Edgar Lindenau Aabye	M	34	NA	NA	Denmark/Sweden	DEN	1900 Summer	1900	Summer
5	5	Christine Jacoba Aaftink	F	21	185	82	Netherlands	NED	1988 Winter	1988	Winter
6	5	Christine Jacoba Aaftink	F	21	185	82	Netherlands	NED	1988 Winter	1988	Winter
7	5	Christine Jacoba Aaftink	F	25	185	82	Netherlands	NED	1992 Winter	1992	Winter

Truncated results, showing first 1000 rows.

Click to re-execute with maximum result limits.

Command took 0.82 seconds -- by pdn8998@mavs.uta.edu at 5/19/2022, 5:13:25 PM on Capstone

We can notice from the screenshot below that a total of 74,357 Females have taken part in the Olympics over the years and a total of 196086 males have taken part over the years.

The regions table has the country with the NOC string values and a lot of ‘NA’ values in the notes column.

```
%sql
--counting number of males and females in total
select Sex, count(Sex) as gender_count
from athlete_events
group by Sex;
```

▸ (2) Spark Jobs

	Sex	gender_count
1	-Campbell	1
2	F	74357
3	-Russell)"	3
4	-Knig)"	1
5	IV"	22
6	M	196086
7	-Clarke)"	1

Showing all 76 rows.

Command took 2.92 seconds -- by pdn8998@mavs.uta.edu at 5/19/2022, 7:41:22 PM on Capstone

Cmd 9

```
%sql
--Observing the datatypes of columns
desc table athlete_events;
```

	col_name	data_type	comment
1	ID	int	null
2	Name	string	null
3	Sex	string	null
4	Age	string	null
5	Height	string	null
6	Weight	string	null
7	Team	string	null

The below image represents the total number of medals won by teams over the years and its observed that United States has won the most number of medals, followed by Soviet Union, Germany and Great Britain.

```
%sql
--Observing the total number of medals won by a team
select Team,count(*)
from athlete_events
where Medal <> 'NA'
group by Team
order by count(*)desc;
```

▸ (2) Spark Jobs

	Team	count(1)
1	United States	5040
2	Soviet Union	2451
3	Germany	1981
4	Great Britain	1671
5	France	1550
6	Italy	1527
7	Sweden	1432

Showing all 561 rows.

Command took 2.45 seconds -- by pdn8998@mavs.uta.edu at 5/19/2022, 8:17:30 PM on Capstone

United States have had the most number of players over the 120 years.

```
%sql
--Country with most number of players
select Team, count(*)
FROM athlete_events
where Name <> 'NA'
group by Team
order by count(*) desc;
```

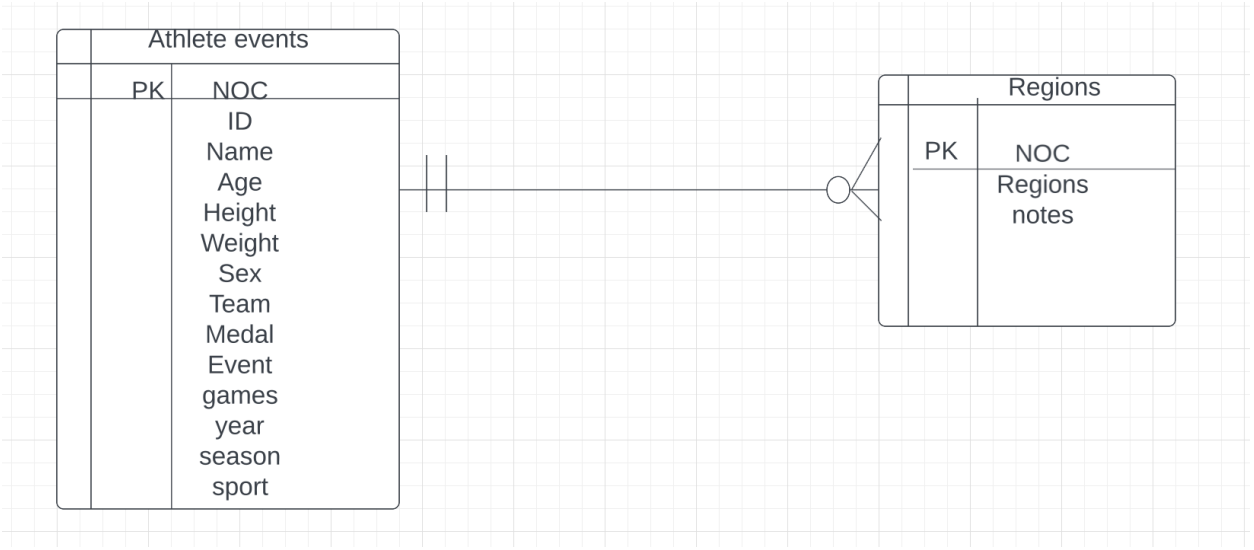
(2) Spark Jobs

	Team	count(1)
1	United States	17372
2	France	11983
3	Great Britain	11389
4	Italy	10260
5	Germany	9303
6	Canada	9246
7	Japan	8289

Truncated results, showing first 1000 rows.
Click to re-execute with maximum result limits.

Command took 3.78 seconds -- by pdn8998@mavs.uta.edu at 5/19/2022, 8:20:58 PM on Capstone

On observing the dataset, we can create an ER - Diagram to establish the relationship between the 2 tables. (athlete_events and regions) NOC is the Pk for joining the tables.



DEVELOPMENT OF PROJECT PROPOSAL

Project Description

My project involves a huge amount of data pertaining to the Olympics, and a lot of organizations, players, and sports media, benefit from it. The analysis done on this data would help in gaining an insight into the sports trends in different seasons of the years and where each country stands. It will also help in making decisions for future players and would help coaches have better knowledge for training and recommendations.

QUESTIONS

1. Since in the initial analysis it was observed that the US has won the most number of medals, I have chosen to select a period of years and observe if there is a relation to the number of players.
2. For all the 10 sports, which team has the most medals in each sport.
3. Have more Females won medals than Males in the US?

HYPOTHESIS

1. The US won more medals as it has sent a large number of participants.
2. Players below the age of 30 have won more medals
3. The season does not matter, an athlete has a chance of winning an equal number of medals in summer as won in winter.

APPROACH

Hypothesis 1 - Approach

I would be observing the total percentage of US participants among the top 5 teams that have played, and come to a decision on whether the hypothesis is proved or disproved based on the percentage of participants by the total number of participants from all 5 countries

Hypothesis 2 - Approach

Taking in categories of age groups below and above 30, I will observe the number of medals won.

Hypothesis 3 - Approach

Considering the number of medals won during summers and winters for each of the 10 sports, I will calculate which season has observed more winning trends, they can be the same or they could differ.

MILESTONE 2

REVIEW CRITERIA

We will be considering the top 3 teams for the hypothesis (US, Soviet Union, Germany)
HYPOTHESIS -1
Us won more number of medals because of sending large no of participants

United stated
On analysis I have observed that, the total number of participants from the US are 17,847.
The number of medals won by the US including gold, silver and bronze are 5,219.

Soviet Union
On analysis I have observed that, the total number of participants from Soviet Union are 5,535.
The number of medals won by the Soviet Union including gold, silver and bronze are 2,451.

Germany
On analysis I have observed that, the total number of participants from Germany are 9,326..
The number of medals won by the Germany including gold, silver and bronze are 1,984.

We can observe that the US has won nearly 5k medals followed by Soviet union with 2k and germany with nearly 1k, by this we can say that US did win more number of medal because of sending more number of participants as compared to the other teams.
But, a point to be considered here is the ratio of number of participants who have been sent from each of the teams to the ratio of medals won.

US sending 17k participants has won only 5k medals but Soviet Union sending only 5k participants has won nearly 2k medals. We can infer from this, the training and quality of the players and methods of training.

HYPOTHESIS -2
Players wil age below 30 have won more medals than players above 30

United States
From the previous hypothesis we have noticed that the total number of players in the US are 17,847
Total number of players below 30 : 14,503
Number of medals won by players below 30 : 4,583

Soviet Union

From the previous hypothesis we have noticed that the total number of players in the Soviet Union are 5,535
Total number of players below 30 : 4,831
Number of medals won by players below 30 : 2,186

Germany
From the previous hypothesis we have noticed that the total number of players from Germany are 9,326
Total number of players below 30 : 7,430
Number of medals won by players below 30 : 1656

We can say that Soviet has more number of players below 30 who have won more medals as compared to US which has 14k players below 30 out of which they have won 4583 medals.

HYPOTHESIS -3
Season does not matter, an athlete has chances of winning equal number of medals in summer as in winter

United States
Medals won by the US in summer : 4,686
Medals won by the US in Winter : 533

Soviet Union
Medals won by the US in summer : 2,061
Medals won by the US in Winter : 390

Germany
Medals won by the US in summer : 1,687
Medals won by the US in Winter : 297

All the three top teams have won more medals during summer than during winter. Until further analysis we can assume that season matters for athletes having more chances of winning.

MILESTONE - 3

```
#Correlation between no of medals won to the age
athlete_events['Season']=athlete_events['Season'].astype('category').cat.codes
athlete_events['Event']=athlete_events['Event'].astype('category').cat.codes
athlete_events.corr()
```

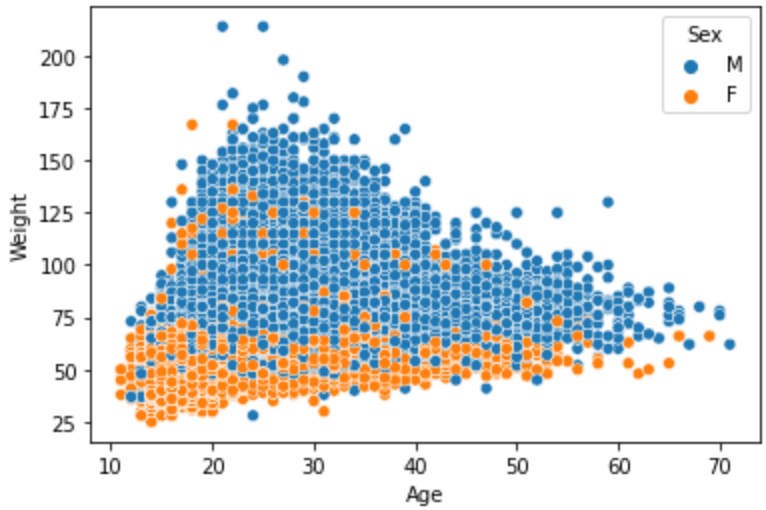
	ID	Age	Height	Weight	Year	Season	Event
ID	1.000000	-0.003631	-0.011141	-0.009176	0.011885	0.013716	0.014227
Age	-0.003631	1.000000	0.138246	0.212069	-0.115137	-0.038521	-0.092111
Height	-0.011141	0.138246	1.000000	0.796213	0.047578	-0.034572	0.000452
Weight	-0.009176	0.212069	0.796213	1.000000	0.019095	0.001919	0.047649
Year	0.011885	-0.115137	0.047578	0.019095	1.000000	0.147697	0.054565
Season	0.013716	-0.038521	-0.034572	0.001919	0.147697	1.000000	-0.142896
Event	0.014227	-0.092111	0.000452	0.047649	0.054565	-0.142896	1.000000

I have observed the correlations between the variables and could say that there is a positive correlation between (height, and weight).

There is a negative correlation between a lot of variables but I would like to take (Height, Age), (Season, and Event) under consideration.

I have decided to consider the following metric:

The team to the medals won by age category as this would help determine why a team is doing better than the others and we can relate it to the historical data and could even consider the age and wt to understand the body build and training strategies as a part of future analysis.


Sex
M
F
Weight
Age
200
175
150
125
100
75
50
25
10
20
30
40
50
60
70