

# PREPARING OF THE PROJECT PROPOSAL

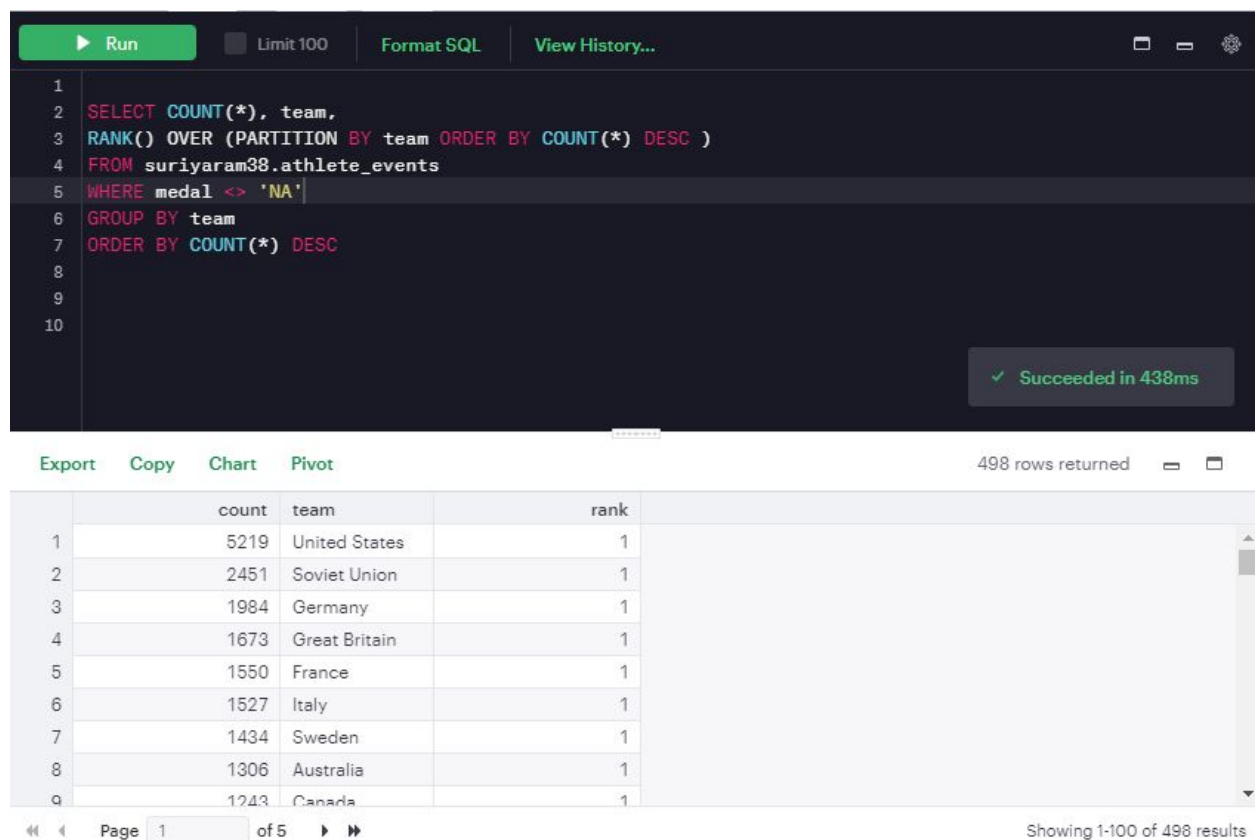
The dataset chosen for the project is the SportStats (120 years of Olympics Dataset). The reason that I chose this dataset is that as a sports buff, I am really interested in working with a sports-based dataset. The data has a lot of potentials to bring in revenue for news stations and media. It also provides a lot of opportunities for health experts and personal trainers of athletes to understand what can be improved in the fitness routines of athletes. I am really looking forward to working thoroughly with the dataset and gather some useful insights out of it.

I used the following steps to import and clean the data before it could be made ready for analysis:

- Downloaded the .gz file from DropBox.
- Extracted and obtained the .CSV files containing tabular data.
- Uploaded the CSV files on the Mode Analytics platform and obtained the tables.
- The noc\_regions table contained a column titled notes which had a combination of NULL values and string type values which represented the region/city from which the athletes were a part of. I replaced these NULL values with 'NA' to make it more presentable.

While initially exploring the data, I came across the following:

If we had to look at countries who have won most of the medals in the history of the Olympics, they would be:



The screenshot shows a SQL query execution interface. At the top, there are buttons for 'Run', 'Limit 100', 'Format SQL', and 'View History...'. The SQL query is as follows:

```
1 SELECT COUNT(*), team,
2 RANK() OVER (PARTITION BY team ORDER BY COUNT(*) DESC )
3 FROM suriyaram38.athlete_events
4 WHERE medal <> 'NA'
5 GROUP BY team
6 ORDER BY COUNT(*) DESC
7
```

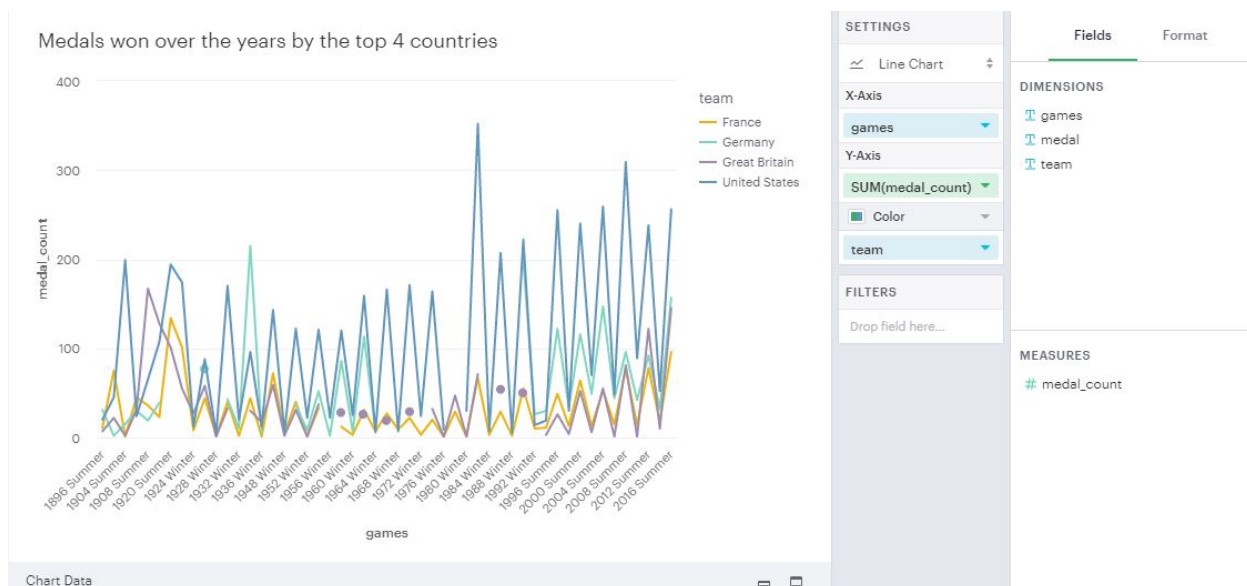
Below the query, a green status bar indicates 'Succeeded in 438ms'. At the bottom, there are tabs for 'Export', 'Copy', 'Chart', and 'Pivot'. The results table shows 498 rows returned. The first few rows are:

	count	team	rank
1	5219	United States	1
2	2451	Soviet Union	1
3	1984	Germany	1
4	1673	Great Britain	1
5	1550	France	1
6	1527	Italy	1
7	1434	Sweden	1
8	1306	Australia	1
9	1243	Canada	1

At the bottom, there are navigation controls for 'Page 1 of 5' and a status bar indicating 'Showing 1-100 of 498 results'.

United States won the most number of medals, followed by the Soviet Union, Germany, Great Britain, and France. For the purpose of analysis, the Soviet Union is not considered because it was dissolved and there is no data pertaining to it in recent years.

Now if we had to look at the trend of these 4 countries (United States, Germany, Great Britain, and France) in winning medals over time, a graph would look like this:



Now, if we look at the cleaned table of noc\_region, it would look like this:

**Run** ☒ Limit 100 **Format SQL** **View History...**

```

1 SELECT noc, region,
2 (CASE WHEN notes IS NULL THEN 'NA' ELSE notes END) AS notes_cleaned
3 FROM suriyaram38.noc_regions

```

**Export** **Copy** **Chart** **Pivot** 100 rows returned

	noc	region	notes_cleaned
1	AFG	Afghanistan	NA
2	AHO	Curacao	Netherlands Antilles
3	ALB	Albania	NA
4	ALG	Algeria	NA
5	AND	Andorra	NA
6	ANG	Angola	NA
7	ANT	Antigua	Antigua and Barbuda
8	ANZ	Australia	Australasia
9	ARG	Argentina	NA

Showing 1-100 of 100 results

The top gold, silver and bronze medal-winning teams are as follows, along with the total number of participants too:

Run

Limit 100

Format SQL

View History...

```
1
2 SELECT COUNT(*), team, medal,
3 RANK() OVER (PARTITION BY team ORDER BY COUNT(*) DESC )
4 FROM suriyaram38.athlete_events
5 WHERE medal <> 'NA' AND medal='Gold'
6 GROUP BY team, medal
7 ORDER BY COUNT(*) DESC
8
9
10
```

Succeeded in 568ms

Export Copy Chart Pivot

242 rows returned

	count	team	medal	rank
1	2474	United States	Gold	1
2	1058	Soviet Union	Gold	1
3	679	Germany	Gold	1
4	535	Italy	Gold	1
5	519	Great Britain	Gold	1
6	455	France	Gold	1
7	451	Sweden	Gold	1
8	432	Hungary	Gold	1
9	422	Canada	Gold	1

Page 1 of 3

Showing 1-100 of 242 results

Run

Limit 100

Format SQL

View History...

```
1
2 SELECT COUNT(*), team, medal,
3 RANK() OVER (PARTITION BY team ORDER BY COUNT(*) DESC )
4 FROM suriyaram38.athlete_events
5 WHERE medal <> 'NA' AND medal='Silver'
6 GROUP BY team, medal
7 ORDER BY COUNT(*) DESC
8
9
10
```

Succeeded in 806ms

Export Copy Chart Pivot

273 rows returned

Page 1 of 3

Showing 1-100 of 273 results

```

1
2 SELECT COUNT(*), team, medal,
3 RANK() OVER (PARTITION BY team ORDER BY COUNT(*) DESC )
4 FROM suriyaram38.athlete_events
5 WHERE medal <> 'NA' AND medal='Bronze'
6 GROUP BY team, medal
7 ORDER BY COUNT(*) DESC
8
9
10

```

✓ Succeeded in 441ms

Export Copy Chart Pivot 268 rows returned

	count	team	medal	rank
1	1233	United States	Bronze	1
2	678	Germany	Bronze	1
3	677	Soviet Union	Bronze	1
4	577	France	Bronze	1
5	572	Great Britain	Bronze	1
6	511	Australia	Bronze	1
7	507	Sweden	Bronze	1
8	484	Italy	Bronze	1
9	415	Finland	Bronze	1

Page 1 of 3 Showing 1-100 of 268 results

```

1 SELECT COUNT(*) AS total_medal_count, team
2 FROM suriyaram38.athlete_events
3 GROUP BY team
4 ORDER BY total_medal_count DESC
5

```

✓ Succeeded in 519ms

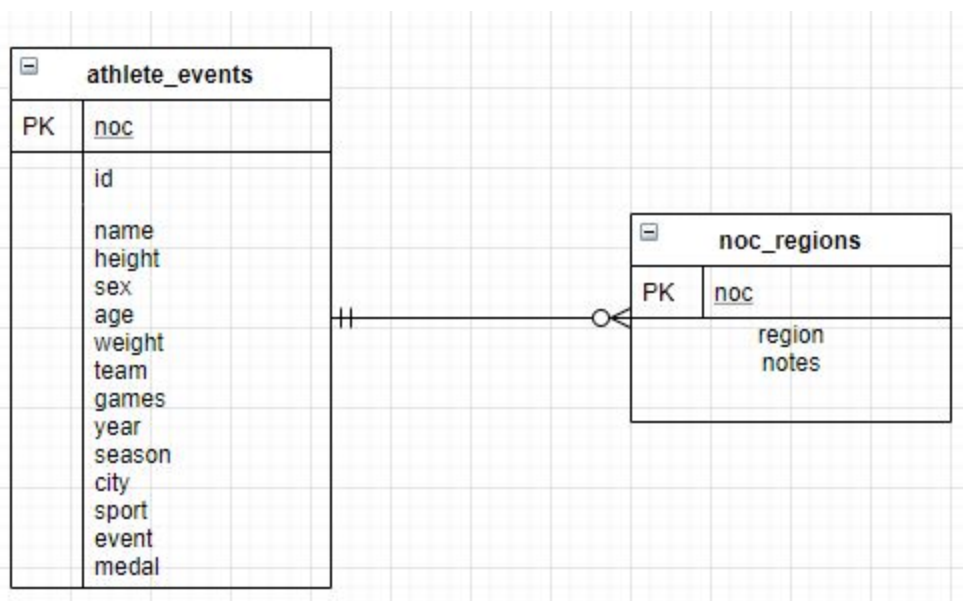
Export Copy Chart Pivot 100 rows returned

	total_medal_count	team
1	17847	United States
2	11988	France
3	11404	Great Britain
4	10260	Italy
5	9326	Germany
6	9279	Canada
7	8289	Japan
8	8052	Sweden
9	7513	Australia

Page 1 of 1 Showing 1-100 of 100 results

This concludes the exploratory data analysis. It is mainly done to get an idea on the top-performing team in the Olympics which will then be used to answer subsequent questions over the course of the project.

There are 2 tables in the dataset. The athletic\_events table contains all the details that we need to perform our data analysis. The noc\_region table contains the region code, country as well as what city/region the athletes are from. The primary key in the athletic\_events table is the noc column and this column is our primary key when it comes to the noc\_region table as well. The noc\_region table has a many-one relationship with our athletic\_events table. This is represented in the form of an ER diagram is plotted below:



## DEVELOPMENT OF THE PROJECT PROPOSAL

### DESCRIPTION:

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?

These are some of the questions that I am looking to answer with my analysis.

## **HYPOTHESIS:**

1. The United States has won so many medals only because they have sent the most number of participants.
2. The majority of a country's medals are concentrated in a certain number of sports. This is common for all countries.
3. The event, in general, does not matter compared to sports. For example, an athlete has an equal chance of winning a gold medal in any of the events that fall under athletics, and so on.

## **APPROACH:**

1. For the first hypothesis, I will be looking at the total number of participants who have attended the Olympics from the 4 countries. Then I will put up a percentage and see how much percentage of the people who participated have won medals. If this value is constant for 4 countries, then it means that the hypothesis is true. If not, I will look at other ways to explain the hypothesis, maybe even deny its veracity.
2. For the second hypothesis, I will make a list of the top 10 sports that have the most medals for these 4 countries. Then I will study how much percentage these sports contribute to the total medal tally. If this value is less, then my hypothesis is false. If not, it is true.
3. For the third hypothesis, I will look at these 10 sports and look at the names of the people who have won medals as well as the events that they have won in. If the same person has won medals in many events, then the hypothesis is true, otherwise, it is false.