

## Step 1: Preparing for Your Proposal

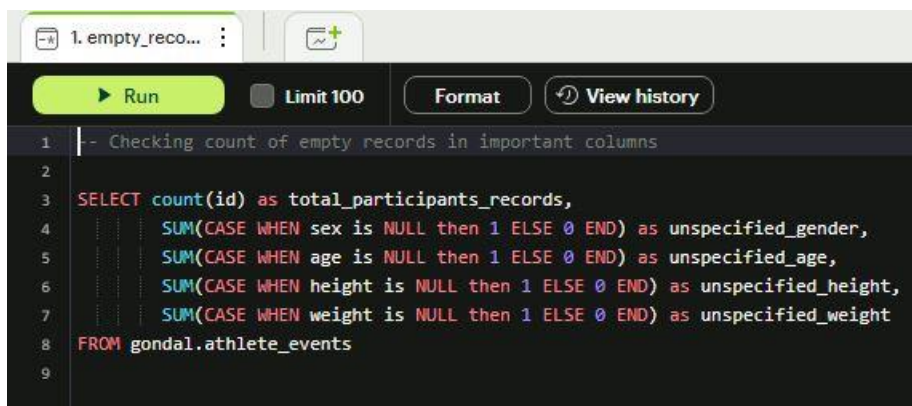
The dataset chosen for the project is the Sport Stats (120 years of Olympics Dataset). The reason why I chose is that for some reason I couldn't access Yelp dataset and political tweets dataset was larger than 1gb and I haven't learned python (memory optimization) yet. So, it was impossible to read data as my notebooks kept running out of memory and online platforms don't allow uploading large datasets. After spending some time on this dataset, I decided to go with Sports Stats.

Data import and cleaning involved following steps:

- Download dataset from the link provided on Coursera week 1 section.
- Unzip the folder with WinRAR.
- Mode analytics kept showing errors when importing CSV file to mode(.)com. Errors were related to datatype of weight, height and age because their datatype was Integer and empty cells had "NA" instead of just being empty or zero.
- I opened CSV file in excel and used 'Find and Replace' on these 3 columns to replace all "NA" with empty spaces (which is zeros).
- Now CSVs were uploaded successfully and tables generated.
- When events table was merged with region table, NOC column had a duplicate. Also, notes column was not required so I used a subquery to filter these columns out and used the updated query in rest of the analysis.

## Exploratory Analysis:

I started with checking for columns with null values.

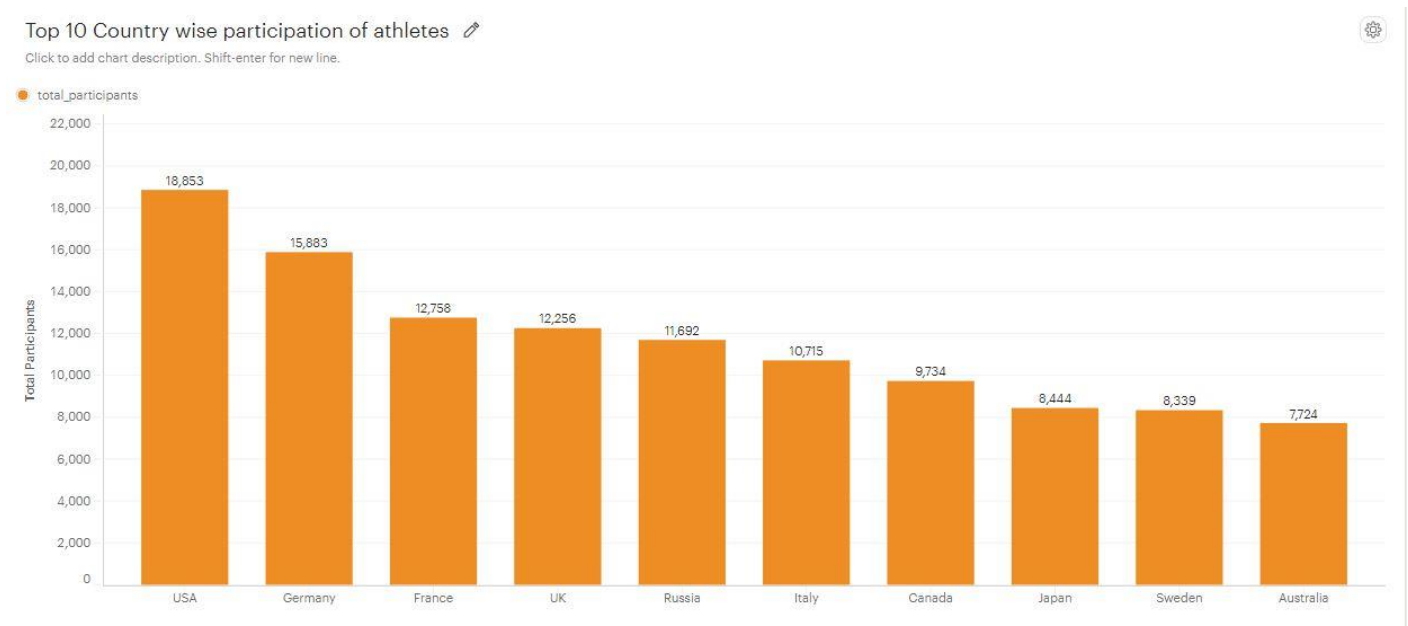


```
1  - Checking count of empty records in important columns
2
3  SELECT count(id) as total_participants_records,
4         SUM(CASE WHEN sex is NULL then 1 ELSE 0 END) as unspecified_gender,
5         SUM(CASE WHEN age is NULL then 1 ELSE 0 END) as unspecified_age,
6         SUM(CASE WHEN height is NULL then 1 ELSE 0 END) as unspecified_height,
7         SUM(CASE WHEN weight is NULL then 1 ELSE 0 END) as unspecified_weight
8  FROM gondal.athlete_events
9
```

Data	Fields	Source
	total_participants_records	unspecified_gender
	unspecified_age	unspecified_height
	unspecified_weight	
1	271116	0
	9474	60171
		62875

Nearly 22.22% of records don't have weight and height data. For now, I didn't deal with it because I'm not sure how it'll impact my analysis.

In Initial phase of data exploration, I checked Country wise participation. There are many reasons for success of Olympic Games but it all starts with registered participants. Screenshot below shows top 10 participating countries in Olympic Games from 1896 to 2016.

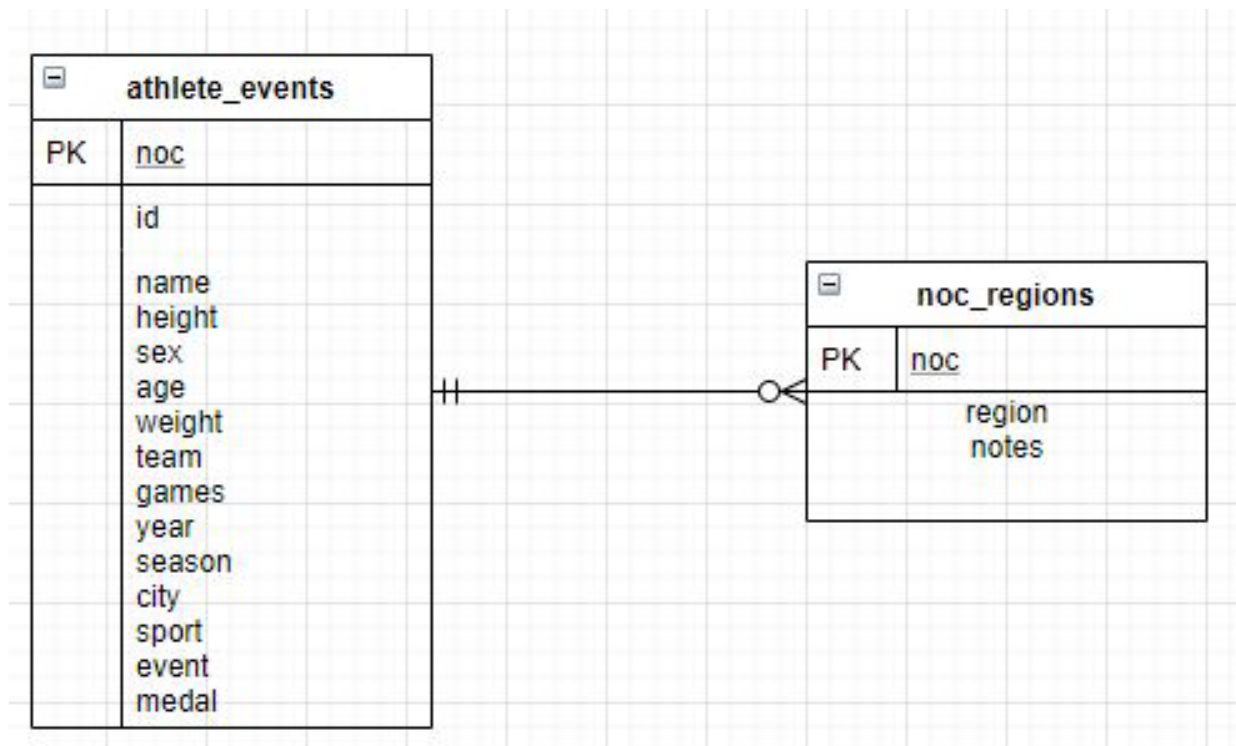


USA is on top with 18,853 athletes followed by Germany (15883), France till Australia on 10<sup>th</sup> spot.

This concludes exploratory phase of data, we found out that our dataset contains null values, which columns we'll need for further analysis and what kind of data they hold (datatypes and their format).

## Entity Relation Diagram

There are two tables in our dataset. Athlete\_events contains information about individual participation in a single event and all the details related to. Noc\_region contains region code and name. Noc is primary key in both tables which connects them and provides us name of countries whose teams and athletes participated in those events in first table. Noc\_region tables has one to many relationship with athlete\_events. One region can have many teams and participants but any team/athlete can't have more than one region.



## Step 2: Develop Project Proposal

### Description

My data analysis project is tailored to engage sports and history enthusiasts, particularly those who enjoy revisiting the rich tapestry of Olympic history. If you're someone who appreciates delving into the past, exploring the data related to height, weight, medals won, and more in Olympic sports events held every 4 years from 1896 to 2016, this analysis is designed for you.

Furthermore, I believe that this exploration can hold value not only for fellow enthusiasts but also for analysts and professionals in the sports industry. Coaches, strategists, and athletes can potentially glean insights from these historical findings to fine-tune their approaches for current and future Olympic competitions. By delving into the lessons of the past, we can strive for even greater excellence in the world of sports.

### Questions

1. How has the participation of countries in the Olympics evolved from 1896 to 2016, and what are the trends and patterns in terms of the number of countries represented each year?
2. What is the historical trend in the participation of female athletes in the Olympics, and how has it changed over the years?
3. How does the participation of female athletes differ between Summer and Winter Olympics, and are there any notable trends or differences between the two?
4. What can historical Olympic data tell us about the performance of athletes and countries over time? Are there any factors, such as height, weight, or other variables, that correlate with better performance?

### Hypothesis

1. The number of athletes participating in the Olympics has increased over time, with occasional fluctuations due to historical events and geopolitical changes.
2. The participation of female athletes in the Olympics has shown a positive upward trend since the inclusion of women's events, reflecting global progress in providing equal rights.

3. Female athlete participation in Summer and Winter Olympics differs, with a higher proportion of female athletes in the Summer Olympics, reflecting variations in the number and types of events offered.
4. Performance in the Olympics is influenced by a combination of factors, including athletes' age, height, and weight, with potential trends and correlations emerging over time.
5. Number of athletes directly influences the number of medals won by any country

## Approach

1. For first two hypotheses I will be looking at total number of participants of both genders separately who have attended Olympics since the beginning.
2. For third hypothesis I'll look at instances of events that involve female athletes grouped by summer and winter season.
3. For fourth hypothesis - performance metric, I'll need columns like weight, height, age and medal count.
4. For fifth hypothesis I'll compare the total number of athletes versus medals won, categorized by country.