Project Proposal

DATA ANALYSIS PROJECT PROPOSAL FOR SPORTSSTATS

JULY 2025

MIA TROIANO

Proposal Description

This project will analyze data from 100+ years of the Olympic games. This analysis will include various trends on the data such as the teams with the most participating athletes, teams with the most medals, mean athlete age and more. This type of data may be interesting for those on the Olympic Committees, as it shows these trends over many years and can inform them on which countries are top performers. This data may also be used by sports commentators as it gives history of the games to discuss with the fans.

Section 1: Questions to Answer

1. Which teams have the most representation and how does this change over time?
2. Which teams have the most medals and how does this change over time?
3. What are the mean demographics in each year of the games? How does the age and gender distribution change over time and by team?
4. Which teams are the top performers in each event?

Section 2: Initial Hypotheses

1. I think that European teams will have the most athlete representation per games in the early years of the games (early 1900s), however the United States, Russia and China will have more athlete representation in the later years and into the 2000s.
2. Similarly to before, I believe that European teams will have more medals in the early games with the Unites States and China having more medals in the later years. I think China will have the most medals overall within the years from this dataset.
3. I think the age demographics will show a gradual decline in age of the athletes over the years, with an increase in height and a decrease in weight as athlete body type norms have evolved. I think this will stay pretty consistent by team with no noticeable differences.
4. I think that teams such as Russia, Canada and Scandanavian countries will dominate the events in the winter Olympics with a higher total medal count. I think teams such as the United States and China will dominate in summer Olympic events with a higher total medal count.

Section 3: Data Analysis Approach

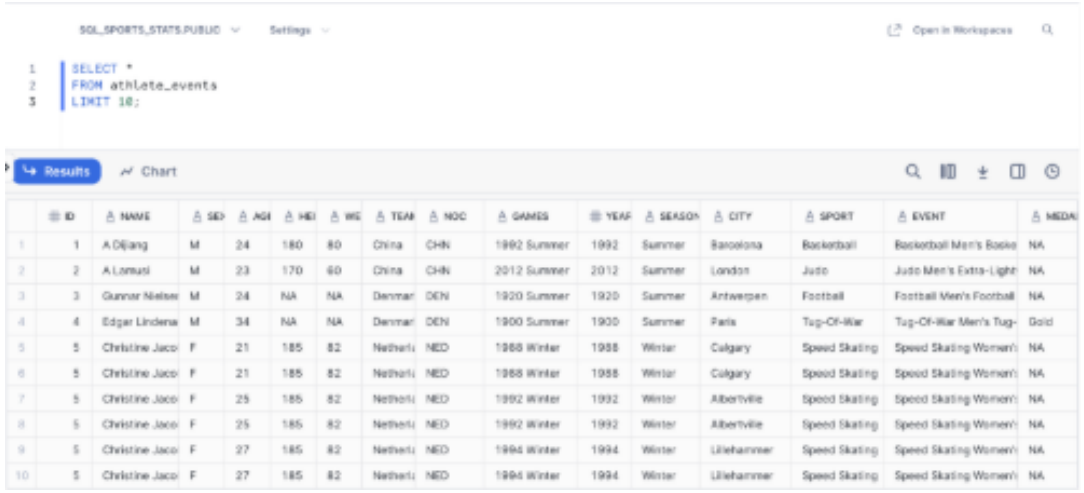
- I will be mostly examining the teams, games, medal and events columns.
- I will be using aggregates to examine descriptive statistics of the demographics of the athletes.
- I will be using aggregates to count various variables, such as athletes and medals.
- I will utilize the chart feature on Snowflake to create a simple visualization with the year of the games on the Y-axis to demonstrate how a variable has changed over the years.

END OF PROPOSAL

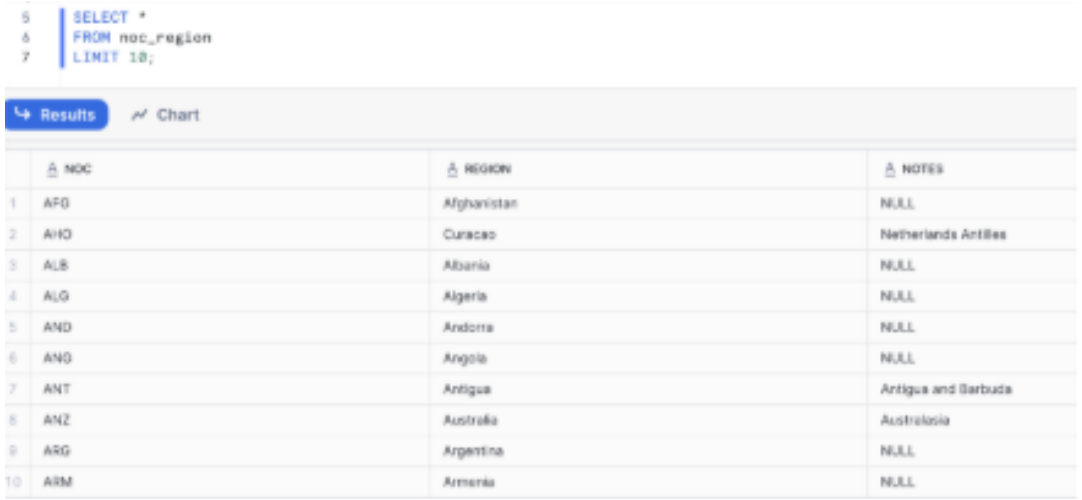
NEXT: PREPARATION BACKGROUND AND ERD

Preparation Questions

- I chose client 3 with SportsStats as my proposal. I chose this client because that is a data set that I find interesting and want to be able to sort through and visualize the data to answer a variety of questions.
- I used Snowflake as the primary server. I initially created a new database and inputted the two tables accompanying my dataset. Once I created the new database, I was able to add a new table and input the data as a CSV. I was able to ensure the schema was correct with the correct headers prior to finalizing the table. There were two tables: an "athlete_events" table and a "noc-region" table. For the "noc_region" region, I had to clean the data fill in NULL variable to allow for proper formatting of the columns with the appropriate header, as many rows were missing this value and the formatting was incorrect. Once the data was loaded into the database, I created a new worksheet and linked it to my new database, where I was now able to run queries.

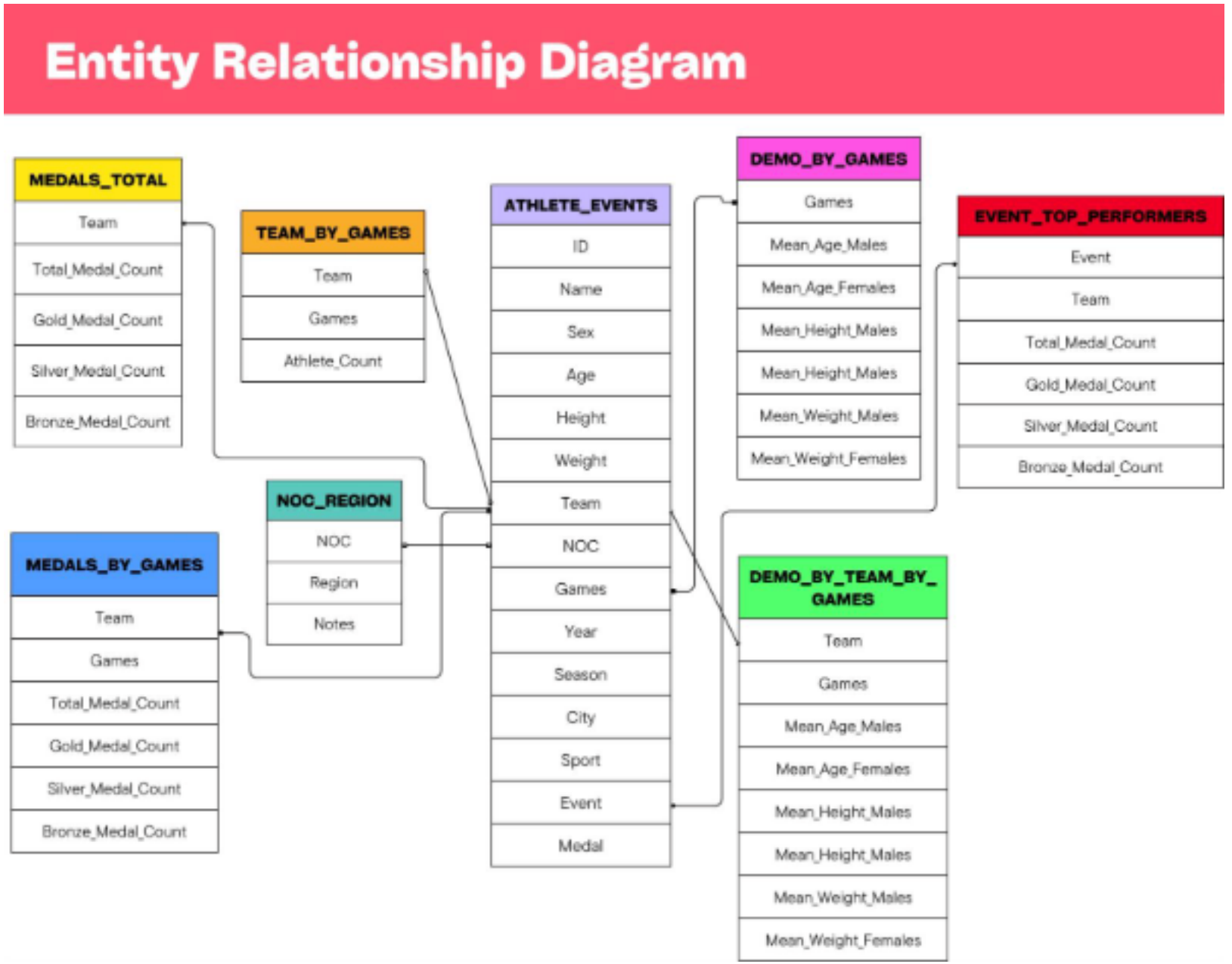


	ID	NAME	SEX	AGE	HEI	WE	TEAM	NOC	GAMES	YEAR	SEASON	CITY	SPORT	EVENT	MEDAL
1	1	A Dijiang	M	24	180	80	China	CHN	1992 Summer	1992	Summer	Barcelona	Basketball	Basketball Men's Baske	NA
2	2	A Lamusi	M	23	170	60	China	CHN	2012 Summer	2012	Summer	London	Judo	Judo Men's Extra-Light	NA
3	3	Gunnar Nielse	M	24	NA	NA	Denmar	DEN	1920 Summer	1920	Summer	Antwerpen	Football	Football Men's Football	NA
4	4	Edgar Lindena	M	34	NA	NA	Denmar	DEN	1900 Summer	1900	Summer	Paris	Tug-Of-War	Tug-Of-War Men's Tug-	Gold
5	5	Christine Jaco	F	21	185	82	Netherl	NED	1988 Winter	1988	Winter	Calgary	Speed Skating	Speed Skating Women's	NA
6	5	Christine Jaco	F	21	185	82	Netherl	NED	1988 Winter	1988	Winter	Calgary	Speed Skating	Speed Skating Women's	NA
7	5	Christine Jaco	F	25	185	82	Netherl	NED	1992 Winter	1992	Winter	Albertville	Speed Skating	Speed Skating Women's	NA
8	5	Christine Jaco	F	25	185	82	Netherl	NED	1992 Winter	1992	Winter	Albertville	Speed Skating	Speed Skating Women's	NA
9	5	Christine Jaco	F	27	185	82	Netherl	NED	1994 Winter	1994	Winter	Lillehammer	Speed Skating	Speed Skating Women's	NA
10	5	Christine Jaco	F	27	185	82	Netherl	NED	1994 Winter	1994	Winter	Lillehammer	Speed Skating	Speed Skating Women's	NA



	NOC	REGION	NOTES
1	AFG	Afghanistan	NULL
2	AHO	Curacao	Netherlands Antilles
3	ALB	Albania	NULL
4	ALG	Algeria	NULL
5	AND	Andorra	NULL
6	ANG	Angola	NULL
7	ANT	Antigua	Antigua and Barbuda
8	ANZ	Australia	Australasia
9	ARG	Argentina	NULL
10	ARM	Armenia	NULL

Data Tables


Entity Relationship Diagram
MEDALS_TOTAL
Team
Total_Medal_Count
Gold_Medal_Count
Silver_Medal_Count
Bronze_Medal_Count
TEAM_BY_GAMES
Team
Games
Athlete_Count
ATHLETE_EVENTS
ID
Name
Sex
Age
Height
Weight
Team
NOC
Games
Year
Season
City
Sport
Event
Medal
DEMO_BY_GAMES
Games
Mean_Age_Males
Mean_Age_Females
Mean_Height_Males
Mean_Height_Males
Mean_Weight_Males
Mean_Weight_Females
EVENT_TOP_PERFORMERS
Event
Team
Total_Medal_Count
Gold_Medal_Count
Silver_Medal_Count
Bronze_Medal_Count
NOC_REGION
NOC
Region
Notes
MEDALS_BY_GAMES
Team
Games
Total_Medal_Count
Gold_Medal_Count
Silver_Medal_Count
Bronze_Medal_Count
DEMO_BY_TEAM_BY_GAMES
Team
Games
Mean_Age_Males
Mean_Age_Females
Mean_Height_Males
Mean_Height_Males
Mean_Weight_Males
Mean_Weight_Females

ERD