**SQL for Data Science**

**Capstone**

**Milestone 1: Project Proposal and Data Selection/Preparation**

**Emi Bode**

**Step 1. Preparing for your Proposal**

* *Which client/dataset did you select and you?*

In this project we will analyse the SportsStats client in order to build sports analytics reports. Sports analytics can help coaches and players, as well as corporations, score touchdowns, prevent injuries and improve fan and sponsorship involvement.

* *Describe the steps you took to import and clean the data*

For data modelling, data analysis and data manipulation we’ll be using Pandas. After reading the CSV files we’ll use the in\_built pandas to store the data in MySQL.

* *Perform initial exploration of data and provide some screenshots or display some stats of the data you are looking at.*

First we will import the data using Pandas:

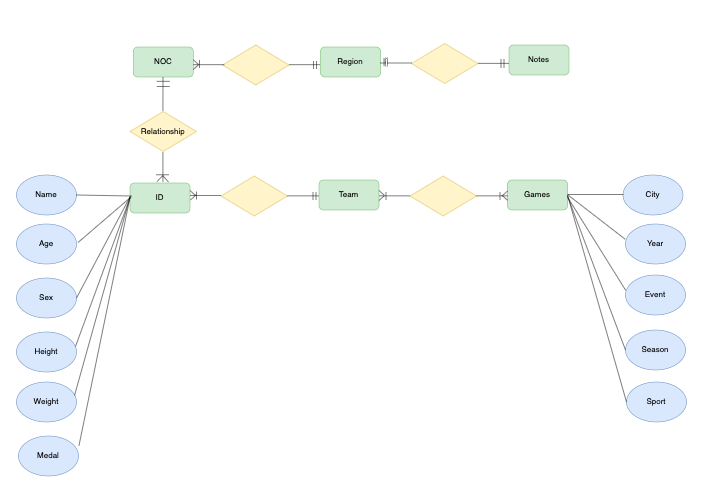


Now we will divide the data into two tables, one for the summer events and another one for the winter events. We will select ID, Name, Sex, Age, Height, Weight, NOC, Year, Sport, Event, Medal from our athlete\_events dataset.



After running the code we can see that from 271116 data ports, 9474 lack age value, 60171 lack height value. Sex values, as well as Year, City, etc. are complete.

* *Create an ERD or proposed ERD to show the relationships of the data you are exploring.*



Step 2. Develop Project Proposal

Using the SportsStats database we will be able to analyse different sports metrics. For this project, we will analyse both summer and winter Olympic games in terms of diversity and performance. We will be able to analyse the team performance and understand the demographic of Olympic games.

* *Questions*

1. Does playing in summer or winter have any impact on a team’s performance?
2. Does the performance of a team by year change more in Winter or in Summer games?
3. How has the male-female ratio changed over time?

* *Hypothesis*

1. Yes, there is a correlation between the performance of a team and the type of game - summer/winter Olympic games
2. The performance of a team changes more in Winter Olympics
3. The male-female ratio has decreased.

* *Approach*

In order to prove our hypothesis first we will calculate the Pearson correlation coefficient. Then we will calculate the standard deviation in country performance through years.

We will draw a histogram to better understand our data.