# Untitled1

July 30, 2023

## 1 Milestone 1: Project Proposal and Data Selection/Preparation By Aiman

## 2 Step 1: Preparing for Your Proposal

## 3 1. Which client/dataset did you select and why?

For my capstone project, I selected the sports dataset provided by SportsStats. The reason behind choosing this dataset is twofold: its relevance and potential for impactful insights.

Relevance: As a sports analysis firm partnering with local news agencies and elite personal trainers, SportsStats' dataset encompasses a vast range of sports-related information. This dataset aligns perfectly with the focus of my capstone project, which revolves around utilizing SQL and data science techniques to extract meaningful insights from sports data.

Potential for Impactful Insights: Sports datasets are incredibly rich and diverse, providing an opportunity to uncover interesting patterns, trends, and correlations across various sports, events, countries, and more. The insights derived from this dataset can be leveraged to create compelling news stories, develop in-depth sports analyses, and even discover key health-related findings. By working with this dataset, I aim to provide valuable and engaging insights that cater to the interests of SportsStats' partners and empower elite personal trainers to optimize their training programs for clients.

In conclusion, I selected the sports dataset from SportsStats for my capstone project due to its relevance to the specialization's focus on SQL for data science and its potential to yield impactful and interesting insights that can benefit the sports industry and health-related endeavors.

## 4 2. Describe the steps you took to import and clean the data.

First, the data was downloaded and stored locally since the volume of files is not big, and does not require Databricks or several clusters to work with. I have used my own jupyter notebooks for this purpose. Second, I have used pandas from Python to read the .csv files, and the built-in to_sql() function to store the data in a MySQL dataset.

[1]:
```
import pandas as pd
import matplotlib.pyplot as plt
```

```
from pandasql import sqldf
pysqldf = lambda q: sqldf(q, globals())

athlete_events = pd.read_csv("athlete_events.csv")
noc_regions = pd.read_csv("noc_regions.csv")
```

Third, and Most Importantly, I did not cleaned the dataset, because the dataset has NaN or Null values, meaning it did not need to be cleaned

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

This preliminary or initial EDA has been carried out with Pandas and Pandas SQL libraries to query the data. Other libraries like Matplotlib and numpy has been used to help the EDA

[3]:
```
athlete_events.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 271116 entries, 0 to 271115
Data columns (total 15 columns):
 #   Column  Non-Null Count   Dtype
---  ------  --------------   -----
 0   ID      271116 non-null  int64
 1   Name    271116 non-null  object
 2   Sex     271116 non-null  object
 3   Age     261642 non-null  float64
 4   Height  210945 non-null  float64
 5   Weight  208241 non-null  float64
 6   Team    271116 non-null  object
 7   NOC     271116 non-null  object
 8   Games   271116 non-null  object
 9   Year    271116 non-null  int64
 10  Season  271116 non-null  object
 11  City    271116 non-null  object
 12  Sport   271116 non-null  object
 13  Event   271116 non-null  object
 14  Medal   39783 non-null   object
dtypes: float64(3), int64(2), object(10)
memory usage: 31.0+ MB
```

[4]:
```
noc_regions.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 230 entries, 0 to 229
Data columns (total 3 columns):
 #   Column  Non-Null Count  Dtype
```

```
---  ------  --------------  -----
 0   NOC     230 non-null    object
 1   region  227 non-null    object
 2   notes   21 non-null     object
dtypes: object(3)
memory usage: 5.5+ KB
```

[5]:
```
athlete_events.describe()
```

[5]:
```
                ID            Age         Height         Weight  \
count  271116.000000  261642.000000  210945.000000  208241.000000
mean    68248.954396      25.556898     175.338970      70.702393
std     39022.286345       6.393561      10.518462      14.348020
min         1.000000      10.000000     127.000000      25.000000
25%     34643.000000      21.000000     168.000000      60.000000
50%     68205.000000      24.000000     175.000000      70.000000
75%    102097.250000      28.000000     183.000000      79.000000
max    135571.000000      97.000000     226.000000     214.000000

                Year
count  271116.000000
mean     1978.378480
std        29.877632
min      1896.000000
25%      1960.000000
50%      1988.000000
75%      2002.000000
max      2016.000000
```

[6]:
```
noc_regions.describe()
```

[6]:
```
        NOC   region           notes
count   230      227              21
unique  230      206              21
top     SKN  Germany  Virgin Islands
freq      1        4               1
```

[7]:
```
athlete_events.head(100)
```

[7]:
```
    ID                              Name Sex   Age  Height  Weight  \
0    1                         A Dijiang   M  24.0   180.0    80.0
1    2                          A Lamusi   M  23.0   170.0    60.0
2    3               Gunnar Nielsen Aaby   M  24.0     NaN     NaN
3    4              Edgar Lindenau Aabye   M  34.0     NaN     NaN
4    5          Christine Jacoba Aaftink   F  21.0   185.0    82.0
..  ..                               ...  ..   ...     ...     ...
95  32              Olav Augunson Aarnes   M  23.0     NaN     NaN
```

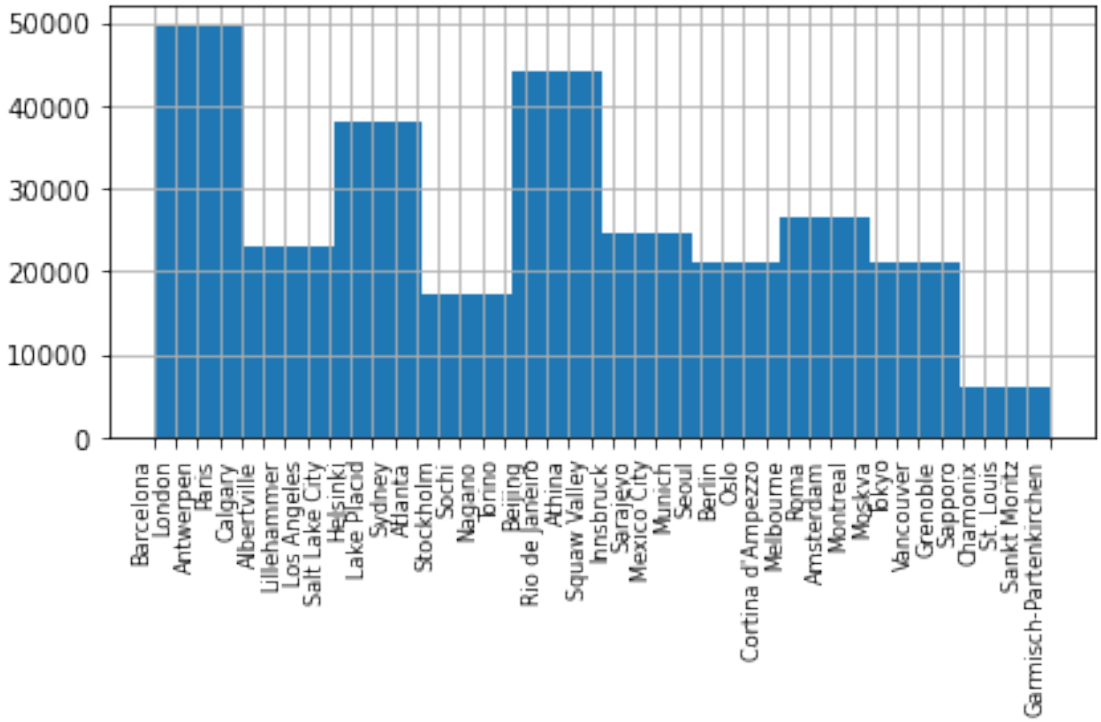
```
96  33                Mika Lauri Aarnikka   M  24.0   187.0    76.0
97  33                Mika Lauri Aarnikka   M  28.0   187.0    76.0
98  34  Jamale (Djamel-) Aarrass (Ahrass-)   M  30.0   187.0    76.0
99  35              Dagfinn Sverre Aarskog   M  24.0   190.0    98.0

              Team  NOC        Games  Year  Season       City          Sport  \
0            China  CHN  1992 Summer  1992  Summer  Barcelona     Basketball
1            China  CHN  2012 Summer  2012  Summer     London           Judo
2          Denmark  DEN  1920 Summer  1920  Summer  Antwerpen       Football
3   Denmark/Sweden  DEN  1900 Summer  1900  Summer      Paris     Tug-Of-War
4      Netherlands  NED  1988 Winter  1988  Winter    Calgary  Speed Skating
..             …  …          …   …     …       …            …
95          Norway  NOR  1912 Summer  1912  Summer  Stockholm      Athletics
96         Finland  FIN  1992 Summer  1992  Summer  Barcelona        Sailing
97         Finland  FIN  1996 Summer  1996  Summer    Atlanta        Sailing
98          France  FRA  2012 Summer  2012  Summer     London      Athletics
99          Norway  NOR  1998 Winter  1998  Winter     Nagano      Bobsleigh

                              Event Medal
0        Basketball Men's Basketball   NaN
1       Judo Men's Extra-Lightweight   NaN
2            Football Men's Football   NaN
3        Tug-Of-War Men's Tug-Of-War  Gold
4   Speed Skating Women's 500 metres   NaN
..                               …   …
95         Athletics Men's High Jump   NaN
96   Sailing Men's Two Person Dinghy   NaN
97   Sailing Men's Two Person Dinghy   NaN
98      Athletics Men's 1,500 metres   NaN
99              Bobsleigh Men's Four   NaN

[100 rows x 15 columns]
```

[8]:
```
noc_regions.head(20)
```

[8]:
```
    NOC          region                notes
0   AFG     Afghanistan                  NaN
1   AHO         Curacao  Netherlands Antilles
2   ALB         Albania                  NaN
3   ALG         Algeria                  NaN
4   AND         Andorra                  NaN
5   ANG          Angola                  NaN
6   ANT         Antigua   Antigua and Barbuda
7   ANZ       Australia           Australasia
8   ARG       Argentina                  NaN
9   ARM         Armenia                  NaN
10  ARU           Aruba                  NaN
```

```
11  ASA  American Samoa                NaN
12  AUS      Australia                 NaN
13  AUT        Austria                 NaN
14  AZE     Azerbaijan                 NaN
15  BAH        Bahamas                 NaN
16  BAN     Bangladesh                 NaN
17  BAR       Barbados                 NaN
18  BDI        Burundi                 NaN
19  BEL        Belgium                 NaN
```

[9]:
```
athlete_events['City'].hist()
plt.xticks(rotation=90, ha='right')
plt.xticks(fontsize=8)
plt.tight_layout()
plt.show()
```




## 6 Exploration of the data with SQL

[10]:
```
pysqldf('SELECT * FROM athlete_events')
```

[10]:
```
             ID                      Name Sex   Age  Height  Weight  \
0             1                 A Dijiang   M  24.0   180.0    80.0
1             2                  A Lamusi   M  23.0   170.0    60.0
2             3       Gunnar Nielsen Aaby   M  24.0     NaN     NaN
3             4      Edgar Lindenau Aabye   M  34.0     NaN     NaN
4             5  Christine Jacoba Aaftink   F  21.0   185.0    82.0
…           …                         …  ..   …     …       …
271111   135569                Andrzej ya   M  29.0   179.0    89.0
271112   135570                  Piotr ya   M  27.0   176.0    59.0
271113   135570                  Piotr ya   M  27.0   176.0    59.0
271114   135571        Tomasz Ireneusz ya   M  30.0   185.0    96.0
271115   135571        Tomasz Ireneusz ya   M  34.0   185.0    96.0

                    Team  NOC        Games  Year  Season            City  \
0                  China  CHN  1992 Summer  1992  Summer       Barcelona
1                  China  CHN  2012 Summer  2012  Summer          London
2                Denmark  DEN  1920 Summer  1920  Summer       Antwerpen
3         Denmark/Sweden  DEN  1900 Summer  1900  Summer           Paris
4            Netherlands  NED  1988 Winter  1988  Winter         Calgary
…                      …  …          …     …     …                …
271111          Poland-1  POL  1976 Winter  1976  Winter       Innsbruck
271112            Poland  POL  2014 Winter  2014  Winter           Sochi
271113            Poland  POL  2014 Winter  2014  Winter           Sochi
271114            Poland  POL  1998 Winter  1998  Winter          Nagano
271115            Poland  POL  2002 Winter  2002  Winter  Salt Lake City

                  Sport                                      Event Medal
0            Basketball              Basketball Men's Basketball  None
1                  Judo            Judo Men's Extra-Lightweight  None
2              Football                 Football Men's Football  None
3            Tug-Of-War              Tug-Of-War Men's Tug-Of-War  Gold
4         Speed Skating        Speed Skating Women's 500 metres  None
…                     …                                      …   …
271111            Luge               Luge Mixed (Men)'s Doubles  None
271112      Ski Jumping  Ski Jumping Men's Large Hill, Individual  None
271113      Ski Jumping       Ski Jumping Men's Large Hill, Team  None
271114        Bobsleigh                      Bobsleigh Men's Four  None
271115        Bobsleigh                      Bobsleigh Men's Four  None

[271116 rows x 15 columns]
```

[11]:
```
pysqldf("SELECT age, COUNT(1) FROM athlete_events WHERE age IS NULL")
```

[11]:
```
    Age  COUNT(1)
0  None      9474
```

[12]:
```
pysqldf("SELECT AVG(age) FROM athlete_events")
```

[12]:
```
    AVG(age)
0  25.556898
```

[13]:
```
pysqldf("SELECT AVG(weight) FROM athlete_events")
```

[13]:
```
    AVG(weight)
0    70.702393
```

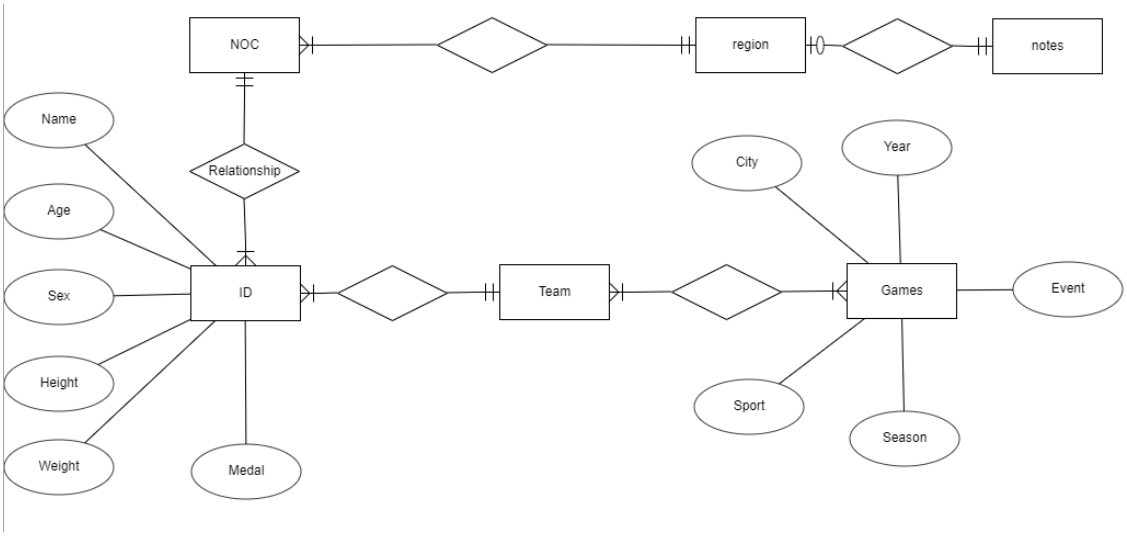
I have performed a quick EDA with simple queries. As seen above there are about 271116 data entries An initial EDA with basic queries shows that there are 271116 entries or Event_ID, while there are entries that are fully completed (Sex, years, season…) and do not contain missing values, there are some others like Age, Height, and Weight, that show missing values.

## 7 4. Create an ERD or proposed ERD to show the relationships of the data you are exploring.

The ERD shown below was intended for a small relational database, splitting them into two tables, the athletes and the NOC Regions. Since teams can be changed, so NOC regions would be more consistent, and since it is also unique ID, and forms one-to-one relations to events table.

[14]:
```
from IPython.display import Image
Image(filename = "ERD.png", width = 800, height = 400)
```

[14]:



# 8 Step 2: Develop Project Proposal

# 9 Project Proposal Description:

My project revolves around analyzing a comprehensive sports dataset provided by SportsStats, a sports analysis firm in collaboration with local news agencies and elite personal trainers. Through the utilization of SQL and data science techniques, I aim to extract valuable insights, patterns, and trends from the sports data. The findings of this analysis can be of great interest to various stakeholders. Local news agencies might find the insights helpful in developing engaging news stories, highlighting intriguing sports-related phenomena, and enhancing their sports reporting. Elite personal trainers can benefit from the health-related discoveries to optimize training regimens tailored to their clients' needs and goals. Additionally, sports enthusiasts, sports analysts, and researchers in the field of sports and health might also be interested in exploring the project's findings to gain a deeper understanding of sports-related trends and their potential implications on athletes' performance and overall well-being.

# 10 Questions:

1.To what extent does the age of athletes influence their chances of obtaining a medal in a given event?

2.Which countries are more likely to achieve medal success: those with abundant resources invested in sports from an early age or those with fewer resources?

3.How is the distribution of medal-winning countries across different seasons? Do northern countries have a higher likelihood of winning medals in Winter Seasons?

4.Has the participation of male and female athletes achieved a state of equality over the years? Is the representation of both genders more balanced in recent decades?

# 11 Hypotheses:

The geographical location of countries, particularly those situated at higher latitudes, may have a significant impact on their Winter Sports performance, resulting in a higher likelihood of winning medals.

Over the years, there might have been a shift towards achieving a more equitable representation of both female and male athletes in sports competitions.

Countries classified as developed are expected to have accumulated a greater number of medals in their historical records compared to less developed nations.

Athletes around the age of 25 may demonstrate peak performance, which could lead to an increased probability of winning medals in their respective events.

## 12 Approach:

Age and Medals Distribution: To investigate the hypothesis related to athletes' age, I will analyze the distribution of age among medal winners. Utilizing graphical representations such as histograms or box plots, I will examine the average age of athletes who have won the most number of medals, providing insights into whether there is a peak age, possibly around 25 years, associated with higher medal-winning probabilities.

Medals and Countries Distribution: To assess the relationship between medals and a country's attributes, I will explore the distribution of medals across different countries. Employing visualizations like bar charts or heatmaps, I will examine if there is any correlation between a country's GDP (Gross Domestic Product) and its medal count, as well as whether countries situated at higher latitudes tend to perform better in Winter Olympics events.

Gender Distribution Over the Years: In order to examine the balance between female and male athletes' participation over time, I will analyze the historical distribution of men and women in sports events. Using line charts or stacked bar charts, I will explore the trends and assess whether both genders' representation has approached equilibrium in sports competitions.

By employing these approaches, I aim to gain valuable insights into the relationships between age and medal success, medals and country attributes, and the evolution of gender representation in sports over the years. The graphical nature of the analysis will facilitate clear and compelling visualizations of the data trends and patterns, leading to a comprehensive understanding of the various hypotheses under investigation.

[ ]: