SQL CAPSTONE PROJECT

WEEK 1

Step 1: Preparing for Your Proposal

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

   Client 3: Sports Stats (Olympics Dataset - 120 years of data)
   SportsStats is a sports analysis firm partnering with local news and elite personal trainers to provide “interesting” insights to help their partners. Insights could be patterns/trends highlighting certain groups/events/countries, etc. for the purpose of developing a news story or discovering key health insights.

   I chose this dataset to gain key insights from the data and which physical characteristics increase the likeliness to win a medal

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

a) I imported the data into a pandas Data Frame into a Jupyter Notebook. There were two csv files - athlete_events.csv and noc_regions.csv

b) I observed the null values present in each column. I saw that the null values of the following columns must be removed.

Age	9474
Height	60171
Weight	62875

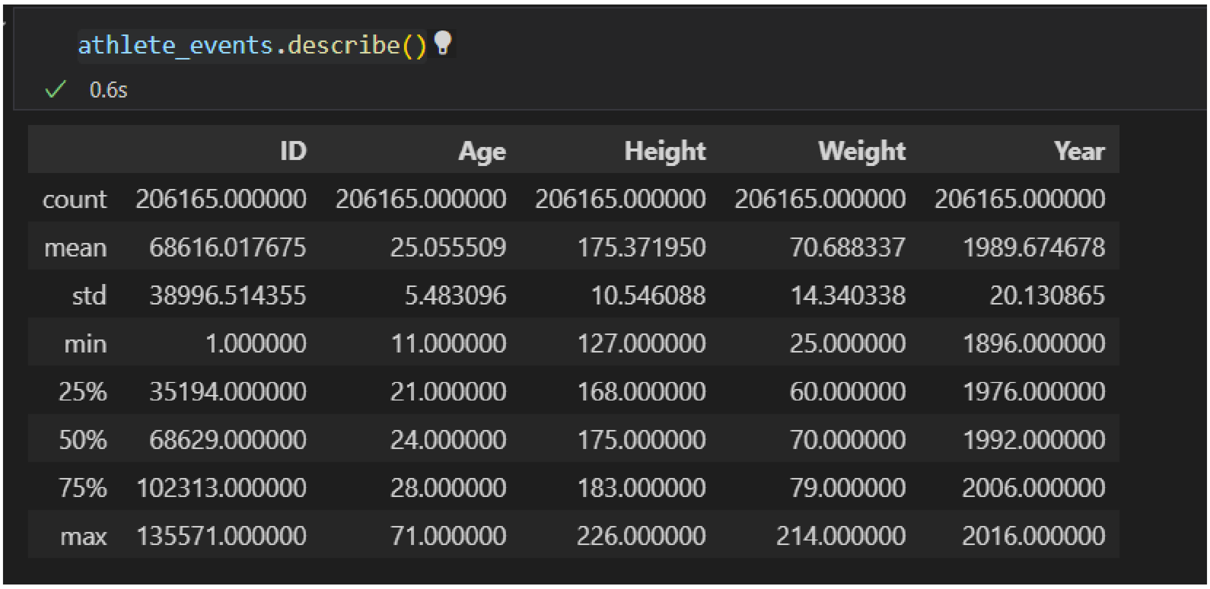
The null values of the medals column were required in order to judge whether an athlete has won a medal or not. So they were not removed

Similarly I removed the Null values of region in the NOC table.

The null values of the notes column were required since that was additional information so it was not removed.

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

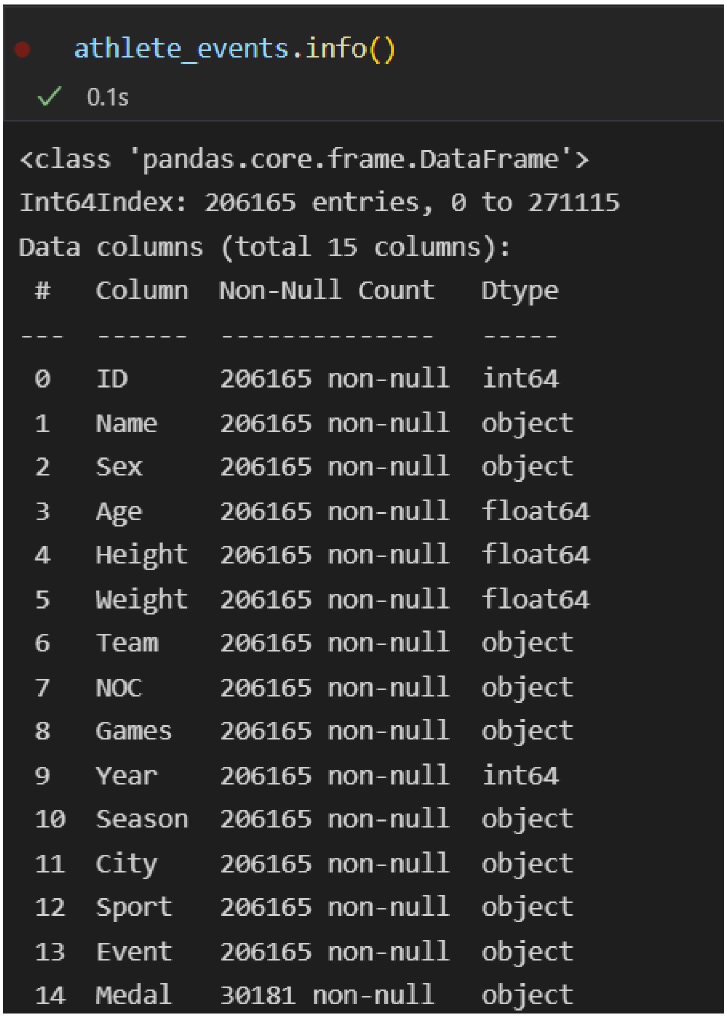
Basic Info of the Dataset:



```
athlete_events.describe()
```

✓ 0.6s

	ID	Age	Height	Weight	Year
count	206165.000000	206165.000000	206165.000000	206165.000000	206165.000000
mean	68616.017675	25.055509	175.371950	70.688337	1989.674678
std	38996.514355	5.483096	10.546088	14.340338	20.130865
min	1.000000	11.000000	127.000000	25.000000	1896.000000
25%	35194.000000	21.000000	168.000000	60.000000	1976.000000
50%	68629.000000	24.000000	175.000000	70.000000	1992.000000
75%	102313.000000	28.000000	183.000000	79.000000	2006.000000
max	135571.000000	71.000000	226.000000	214.000000	2016.000000

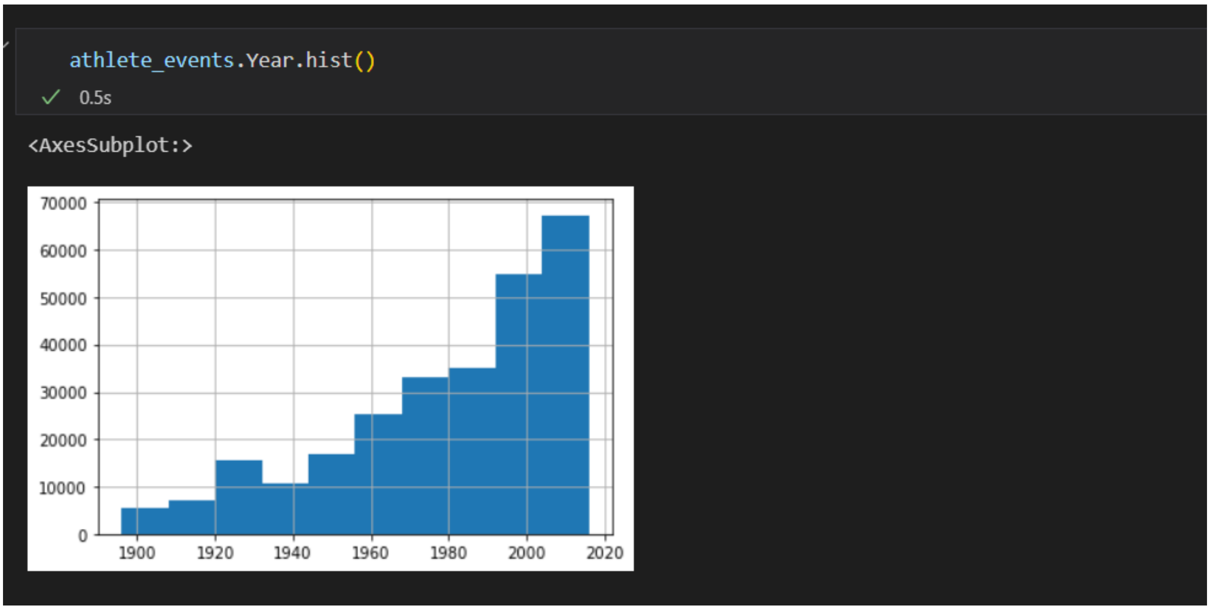


```
athlete_events.info()
```

✓ 0.1s

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 206165 entries, 0 to 271115
Data columns (total 15 columns):
 #   Column  Non-Null Count   Dtype
---  ------  --------------   -----
 0   ID      206165 non-null  int64
 1   Name    206165 non-null  object
 2   Sex     206165 non-null  object
 3   Age     206165 non-null  float64
 4   Height  206165 non-null  float64
 5   Weight  206165 non-null  float64
 6   Team    206165 non-null  object
 7   NOC     206165 non-null  object
 8   Games   206165 non-null  object
 9   Year    206165 non-null  int64
 10  Season  206165 non-null  object
 11  City    206165 non-null  object
 12  Sport   206165 non-null  object
 13  Event   206165 non-null  object
 14  Medal   30181 non-null   object
```

This is a histogram of the years of the dataset

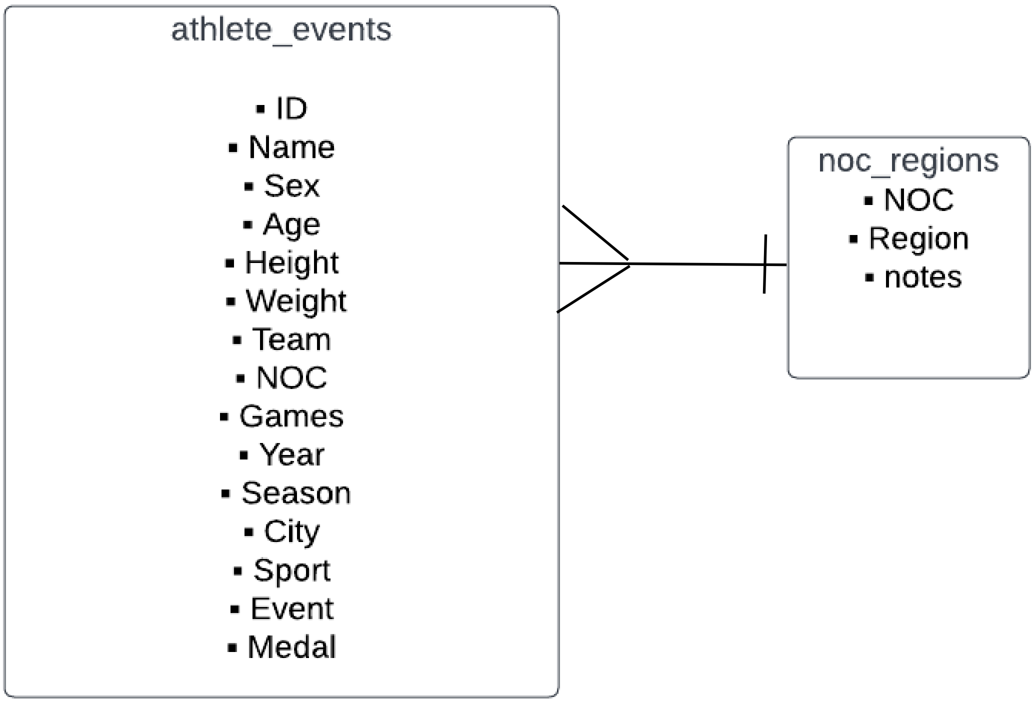


```
athlete_events.Year.hist()
```

✓ 0.5s

```
<AxesSubplot:>
```

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



Step 2: Develop Project Proposal

Description

I would like to find out which physical body characteristics are necessary for winning a medal

Questions

I would like to find the answers to the following questions in the data

1. The average height and weight characteristics for medal winners in each game

   This will help me discover the ideal body type for a certain game

2. What is the average age of medal winners in each game

   This will help me discover the ideal age in order to win a game

3. I wish to know which Country has won the most medals in each game

Hypothesis

1. The average BMI value of the medal willing athletes matches with the ideal values
2. Developed countries have won more medals due to high infrastructure of sports.

Approach

Applying SQL queries to the dataset.

Using Where and GROUP BY clauses will help me find the answer.