Sportstats Capstone Project

PREPARING PROPOSAL

CLIENT/DATASET

I selected this dataset from SportsStats, a sports analysis firm collaborating with local news and elite personal trainers. The goal is to extract valuable insights and determine which physical characteristics enhance the chances of winning a medal. By analyzing the data, we can uncover patterns and trends related to specific groups, events, countries, and more. These insights can be used to develop news stories or discover important health insights

DATA IMPORT

I imported the data into a pandas DataFrame in a Jupyter Notebook. The data consisted of two csv files - athlete_events.csv and noc_regions.csv. I noticed that there were null values in each column. To ensure data integrity, I decided to remove these null values from these columns.

However, I chose not to remove the null values in the "Medals" column as they were necessary for determining whether an athlete had won a medal or not. Similarly, I retained the null values in the "Region" column of the NOC table, as they were required for further analysis.

The null values in the "Notes" column were also preserved, as they contained additional information that was deemed important.

In summary, I removed the null values from the "Age", "Height", and "Weight" columns, while retaining the null values in the "Medals" and "Notes" columns, as well as the "Region" column in the NOC table.

INITIAL DATA EXPLORATION

Basic Information of the dataset:

```
athlete_events.describe()
```

	ID	Age	Height	Weight	Year
count	271116.000000	261642.000000	210945.000000	208241.000000	271116.000000
mean	68248.954396	25.556898	175.338970	70.702393	1978.378480
std	39022.286345	6.393561	10.518462	14.348020	29.877632
min	1.000000	10.000000	127.000000	25.000000	1896.000000
25%	34643.000000	21.000000	168.000000	60.000000	1960.000000
50%	68205.000000	24.000000	175.000000	70.000000	1988.000000
75%	102097.250000	28.000000	183.000000	79.000000	2002.000000
max	135571.000000	97.000000	226.000000	214.000000	2016.000000

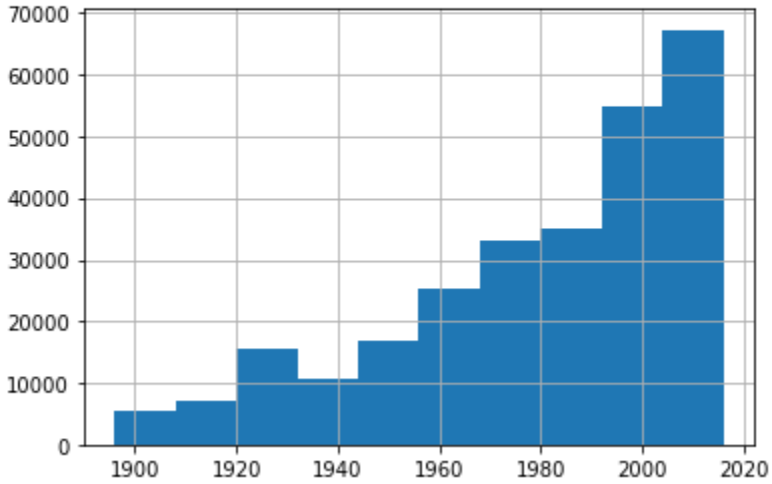
```
athlete_events.info()
```

```
<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
```

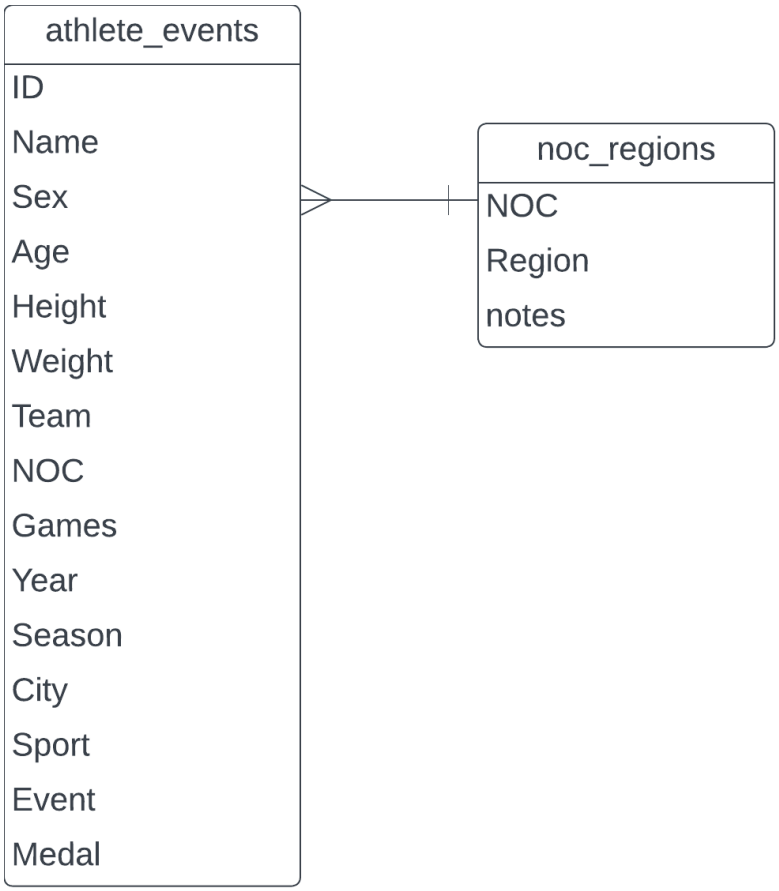
A Histogram of the years of the dataset:

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

```
<matplotlib.axes._subplots.AxesSubplot at 0x7efefc735050>
```



Proposed ERD:



DEVELOP PROPOSAL

DESCRIPTION

To determine the physical body characteristics required for winning a medal, an analysis of various factors should be conducted. These factors may include but are not limited to height, weight, muscle mass, body fat percentage, endurance, strength, agility, and flexibility. By studying successful medalists in different sports, patterns and correlations between these characteristics and medal-winning performances can be identified. Additionally, considering the specific requirements and demands of each sport or event would be crucial in determining the necessary

QUESTIONS

To find the answers to the following questions in the data, I would like to know:

1. The average height and weight characteristics for medal winners in each game, which will help me discover the ideal body type for a certain game.
2. The average age of medal winners in each game, which will help me discover the ideal age to win a game.
3. Which country has won the most medals in each game.

HYPOTHESIS

1. The BMI values of the medal-winning athletes correspond to the ideal values on average.
2. Developed countries have achieved more medals due to their advanced sports infrastructure

APPROACH

To find the answer, I will apply SQL queries to the given dataset. By using the Where and GROUP BY clauses, I can narrow down the results and analyze the data.