How do body mass, height, body mass index, and age influence a professional athlete during the Olympic games?

Step 1: Preparing for Your Proposal

Client: SportsStats has a partner in the field of Sports and Conditioning that is responsible to develop training routines to the American Olympic teams. They are interested in knowing how weight, height, and age can influence the changes of their athletes on winning a medal in the next Olympic game.

This analysis can also be interesting to sports fans and the audience of those events. The result can be also applied to select talents in high-school and college-level events to engage in a more systematic and professional training process.

Database: SportsStats (Olympics Dataset - 120 years of data)

The database was downloaded in my machine and then I used Jupyter Notebooks to import in a Pandas DataFrame.

```
In [3]: import pandas as pd
        import numpy as np
        import matplotlib.pyplot as plt
        import seaborn as sns
```

```
In [4]: olympics = pd.read_csv('/Users/luciomuramatsu/Google Drive/Code/python/DataScience/dataOlympics/athlete_events.csv')
```

After the data was imported I extract basic stats to get familiar with the columns and datatypes. Also, I plotted some charts to see how the data was distributed along the rows.

```
In [5]: olympics.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
```

In [37]:
```
olympics.head(10)
```

Out[37]:

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event	Medal
0	1	A Dijiang	M	24.0	180.0	80.0	China	CHN	1992 Summer	1992	Summer	Barcelona	Basketball	Basketball Men's Basketball	NaN
1	2	A Lamusi	M	23.0	170.0	60.0	China	CHN	2012 Summer	2012	Summer	London	Judo	Judo Men's Extra-Lightweight	NaN
2	3	Gunnar Nielsen Aaby	M	24.0	NaN	NaN	Denmark	DEN	1920 Summer	1920	Summer	Antwerpen	Football	Football Men's Football	NaN
3	4	Edgar Lindenau Aabye	M	34.0	NaN	NaN	Denmark/Sweden	DEN	1900 Summer	1900	Summer	Paris	Tug-Of-War	Tug-Of-War Men's Tug-Of-War	Gold
4	5	Christine Jacoba Aaftink	F	21.0	185.0	82.0	Netherlands	NED	1988 Winter	1988	Winter	Calgary	Speed Skating	Speed Skating Women's 500 metres	NaN
5	5	Christine Jacoba Aaftink	F	21.0	185.0	82.0	Netherlands	NED	1988 Winter	1988	Winter	Calgary	Speed Skating	Speed Skating Women's 1,000 metres	NaN
6	5	Christine Jacoba Aaftink	F	25.0	185.0	82.0	Netherlands	NED	1992 Winter	1992	Winter	Albertville	Speed Skating	Speed Skating Women's 500 metres	NaN
7	5	Christine Jacoba Aaftink	F	25.0	185.0	82.0	Netherlands	NED	1992 Winter	1992	Winter	Albertville	Speed Skating	Speed Skating Women's 1,000 metres	NaN
8	5	Christine Jacoba Aaftink	F	27.0	185.0	82.0	Netherlands	NED	1994 Winter	1994	Winter	Lillehammer	Speed Skating	Speed Skating Women's 500 metres	NaN
9	5	Christine Jacoba Aaftink	F	27.0	185.0	82.0	Netherlands	NED	1994 Winter	1994	Winter	Lillehammer	Speed Skating	Speed Skating Women's 1,000 metres	NaN

In [8]:
```
from pandasql import sqldf
pysqldf = lambda q: sqldf(q, globals())
```

In [9]:
```
pysqldf('''
SELECT
    Sport,
    COUNT(Sport) Count
FROM olympics
WHERE
    Height IS NOT NULL
    AND
    Weight IS NOT NULL
    AND
    Age IS NOT NULL
GROUP BY Sport
ORDER BY Count DESC
;
''')
```

Out[9]:

	Sport	Count
0	Athletics	32374
1	Swimming	18776
2	Gymnastics	18271
3	Rowing	7790

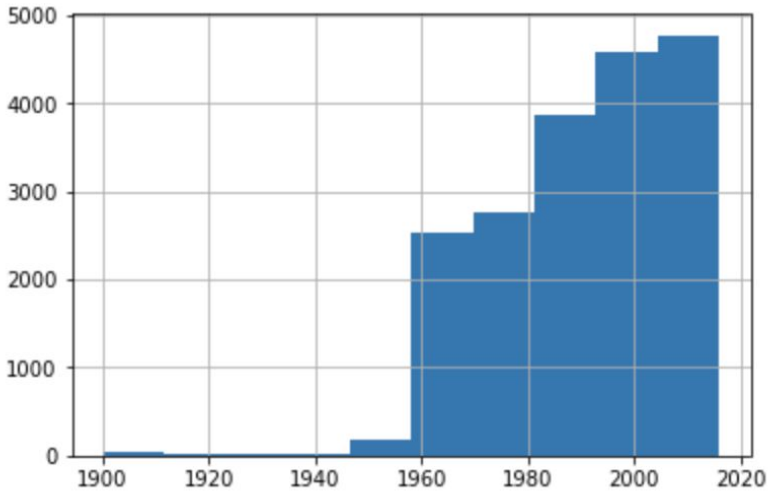
In [11]:
```
pysqldf('''
SELECT
    *
FROM swimming
WHERE
    Medal > 0
;
''')
```

Out[11]:

	Name	Sex	Age	Height	Weight	Year	Medal
0	Reema Abdo	F	21.0	173.0	59.0	1984	1
1	Viktor Andreyevich Aboimov	M	22.0	190.0	78.0	1972	2
2	Viktor Andreyevich Aboimov	M	22.0	190.0	78.0	1972	1
3	Matthew "Matt" Abood	M	30.0	197.0	92.0	2016	1
4	Gary Abraham	M	21.0	175.0	64.0	1980	1
...	...	...	...	...	...	...	...
2481	Iris Zscherpe	F	17.0	174.0	55.0	1984	1
2482	Martijn Hendrik Zuijdweg	M	23.0	186.0	83.0	2000	1
2483	Robertas ulpa	M	20.0	193.0	82.0	1980	3
2484	Anastasiya Valeryevna Zuyeva-Fesikova	F	22.0	182.0	71.0	2012	2
2485	Klaas Erik "Klaas-Erik" Zwering	M	23.0	189.0	80.0	2004	2

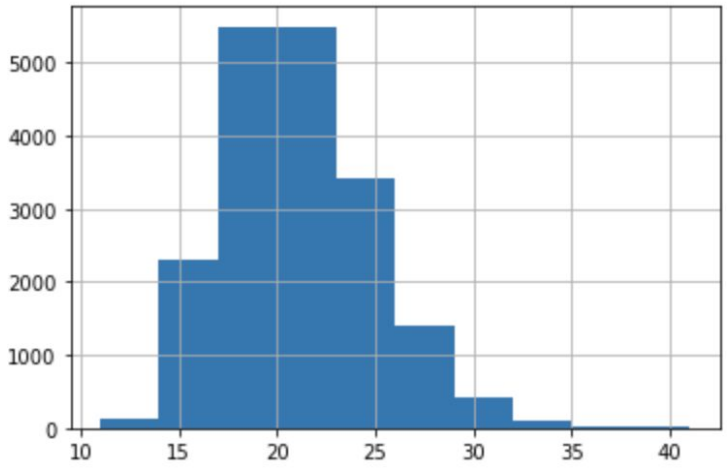
In [12]:
```
swimming.Year.hist()
```

Out[12]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f924b2a7940>
```



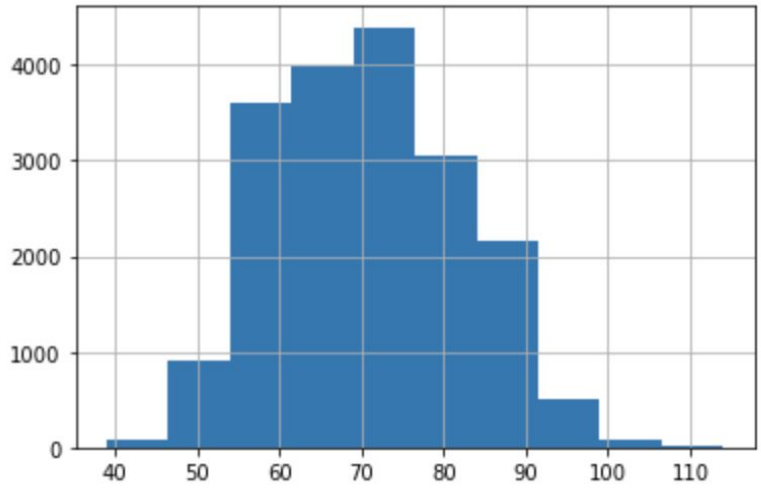
In [13]:
```
swimming.Age.hist()
```

Out[13]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f9235ae0fd0>
```



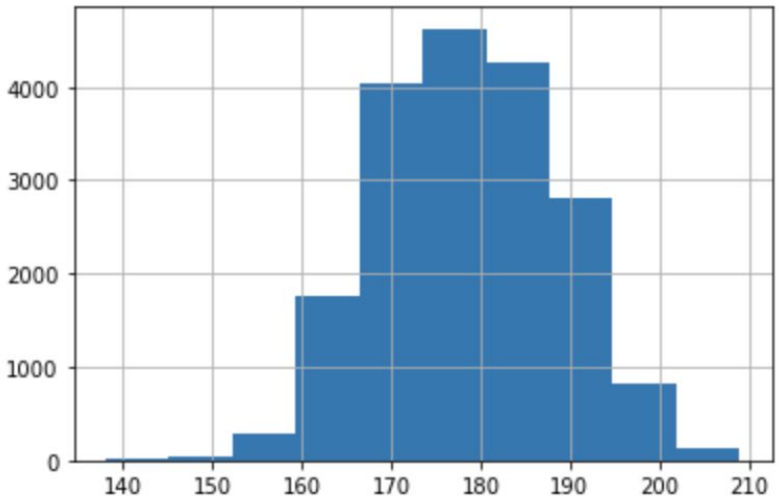
In [14]:
```
swimming.Weight.hist()
```

Out[14]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f9232ecf130>
```

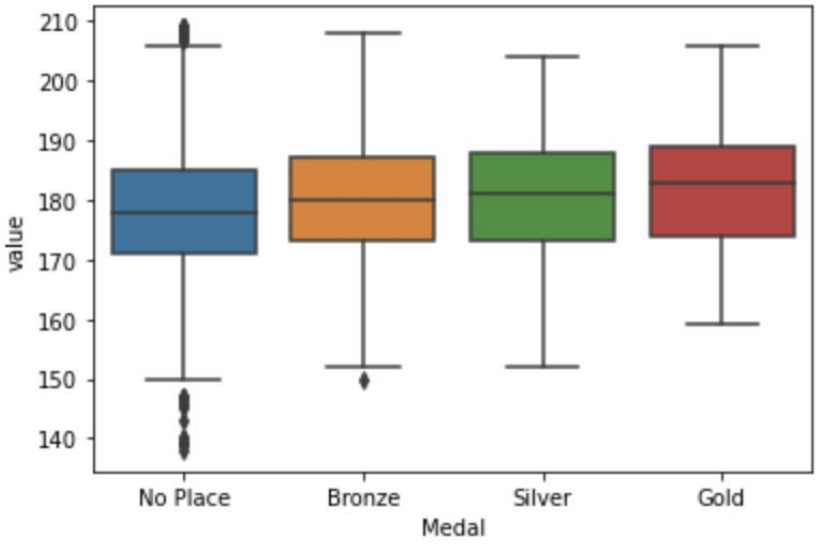


```
In [15]: swimming.Height.hist()
```

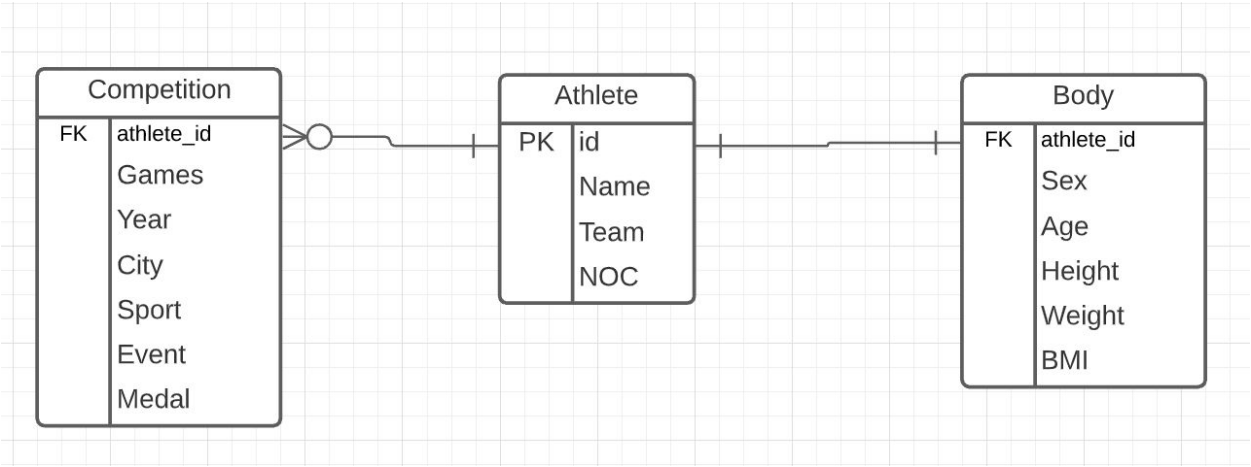
```
Out[15]: <matplotlib.axes._subplots.AxesSubplot at 0x7f9232ce3790>
```




```
In [36]: ax = sns.boxplot(x="Medal", y="value", data=mdf)
plt.show()
```



Entity Relationship Diagram:



Step 2: Develop the Project Proposal

Description

This project was requested by a company in the field of Sports and Conditioning that is responsible to develop training routines for the American Olympic teams. They are interested in knowing how weight, height, and age can influence the changes of their athletes on winning a medal in the next Olympic game.

This analysis can also be interesting to sports fans and the audience of those events. The result can be also applied to select talents in high-school and college-level events to engage in a more systematic and professional training process.

Questions

1. How do body height, body mass, body mass index (BMI), and age influence the success of an Olympic Athlete?
2. What are the sports that require a higher height to be successful?
3. Is BMI a predictor of success in certain sports? If yes, what sports are impacted?
4. Does gender play a role in the height component?

Hypothesis

1. Body height will be a predictor of success in sports that reach play a big role, such as, swimming, basketball, volleyball, and athletics.
2. BMI will be a predictor of success for the sports that are divided into weight divisions, such as, boxing, judo, wrestling, and karate.

3. Age will be a predictor of success in sports where the physical component of muscular power is more important than the technical and the endurance component, such as sprinting in athletics and weight lifting

Approach

1. Histograms to determine if the data is normal.
2. Boxplots and Anova to verify the difference between groups.
3. Machine learning techniques with linear and multiple regression to analyze how the variables (weight, height, BMI, and age) will impact the results during the Olympic games.