In [1]:

```
# firstly we have to import numpy and pandas
import numpy as np
import pandas as pd
```

In [2]:

```
# upload the file to read
df = pd.read_csv('athlete_events.csv')
```

In [3]:

```
df
```

Out[3]:

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year
0	1	A Dijiang	M	24.0	180.0	80.0	China	CHN	1992 Summer	1992
1	2	A Lamusi	M	23.0	170.0	60.0	China	CHN	2012 Summer	2012
2	3	Gunnar Nielsen Aaby	M	24.0	NaN	NaN	Denmark	DEN	1920 Summer	1920
3	4	Edgar Lindenau Aabye	M	34.0	NaN	NaN	Denmark/Sweden	DEN	1900 Summer	1900
4	5	Christine Jacoba Aaftink	F	21.0	185.0	82.0	Netherlands	NED	1988 Winter	1988
...	...	...	...	...	...	...	...	...	...	...
271111	135569	Andrzej ya	M	29.0	179.0	89.0	Poland-1	POL	1976 Winter	1976
271112	135570	Piotr ya	M	27.0	176.0	59.0	Poland	POL	2014 Winter	2014
271113	135570	Piotr ya	M	27.0	176.0	59.0	Poland	POL	2014 Winter	2014
271114	135571	Tomasz Ireneusz ya	M	30.0	185.0	96.0	Poland	POL	1998 Winter	1998
271115	135571	Tomasz Ireneusz ya	M	34.0	185.0	96.0	Poland	POL	2002 Winter	2002

271116 rows × 15 columns

In [4]:

```
# It gives first 6 row of the table
df.head(6)
```

Out[4]:

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season
0	1	A Dijiang	M	24.0	180.0	80.0	China	CHN	1992 Summer	1992	Summer
1	2	A Lamusi	M	23.0	170.0	60.0	China	CHN	2012 Summer	2012	Summer
2	3	Gunnar Nielsen Aaby	M	24.0	NaN	NaN	Denmark	DEN	1920 Summer	1920	Summer
3	4	Edgar Lindenau Aabye	M	34.0	NaN	NaN	Denmark/Sweden	DEN	1900 Summer	1900	Summer
4	5	Christine Jacoba Aaftink	F	21.0	185.0	82.0	Netherlands	NED	1988 Winter	1988	Winter
5	5	Christine Jacoba Aaftink	F	21.0	185.0	82.0	Netherlands	NED	1988 Winter	1988	Winter

In [5]:

```
# it gives last five row of the data
df.tail()
```

Out[5]:

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season
271111	135569	Andrzej ya	M	29.0	179.0	89.0	Poland-1	POL	1976 Winter	1976	Winter
271112	135570	Piotr ya	M	27.0	176.0	59.0	Poland	POL	2014 Winter	2014	Winter
271113	135570	Piotr ya	M	27.0	176.0	59.0	Poland	POL	2014 Winter	2014	Winter
271114	135571	Tomasz Ireneusz ya	M	30.0	185.0	96.0	Poland	POL	1998 Winter	1998	Winter
271115	135571	Tomasz Ireneusz ya	M	34.0	185.0	96.0	Poland	POL	2002 Winter	2002	Winter

In [6]:

```
# it give the information of the whole data regarding that how many null values are pres
df.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 [7]:

```
# it gives all the statistics result
df.describe()
```

Out[7]:

	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

In [8]:

```
# upload another file to read
regions = pd.read_csv("noc_regions (1).csv")
```

In [9]:

```
regions
```

Out[9]:

	NOC	region	notes
0	AFG	Afghanistan	NaN
1	AHO	Curacao	Netherlands Antilles
2	ALB	Albania	NaN
3	ALG	Algeria	NaN
4	AND	Andorra	NaN
...	...	...	...
225	YEM	Yemen	NaN
226	YMD	Yemen	South Yemen
227	YUG	Serbia	Yugoslavia
228	ZAM	Zambia	NaN
229	ZIM	Zimbabwe	NaN

230 rows × 3 columns

In [10]:

```
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
```

In [11]:

```
# change the name of dataframe from df to athletes
athletes = df
```

In [12]:

```
athletes
```

Out[12]:

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year
0	1	A Dijiang	M	24.0	180.0	80.0	China	CHN	1992 Summer	1992
1	2	A Lamusi	M	23.0	170.0	60.0	China	CHN	2012 Summer	2012
2	3	Gunnar Nielsen Aaby	M	24.0	NaN	NaN	Denmark	DEN	1920 Summer	1920
3	4	Edgar Lindenau Aabye	M	34.0	NaN	NaN	Denmark/Sweden	DEN	1900 Summer	1900
4	5	Christine Jacoba Aaftink	F	21.0	185.0	82.0	Netherlands	NED	1988 Winter	1988
...	...	...	...	...	...	...	...	...	...	...
271111	135569	Andrzej ya	M	29.0	179.0	89.0	Poland-1	POL	1976 Winter	1976
271112	135570	Piotr ya	M	27.0	176.0	59.0	Poland	POL	2014 Winter	2014
271113	135570	Piotr ya	M	27.0	176.0	59.0	Poland	POL	2014 Winter	2014
271114	135571	Tomasz Ireneusz ya	M	30.0	185.0	96.0	Poland	POL	1998 Winter	1998
271115	135571	Tomasz Ireneusz ya	M	34.0	185.0	96.0	Poland	POL	2002 Winter	2002

271116 rows × 15 columns

In [13]:

```
# join the dataframe
athletes_df = pd.merge(athletes,regions, how = 'left' , on = 'NOC')
```

In [14]:

```
athletes_df
```

Out[14]:

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year
0	1	A Dijiang	M	24.0	180.0	80.0	China	CHN	1992 Summer	1992
1	2	A Lamusi	M	23.0	170.0	60.0	China	CHN	2012 Summer	2012
2	3	Gunnar Nielsen Aaby	M	24.0	NaN	NaN	Denmark	DEN	1920 Summer	1920
3	4	Edgar Lindenau Aabye	M	34.0	NaN	NaN	Denmark/Sweden	DEN	1900 Summer	1900
4	5	Christine Jacoba Aaftink	F	21.0	185.0	82.0	Netherlands	NED	1988 Winter	1988
...	...	...	...	...	...	...	...	...	...	...
271111	135569	Andrzej ya	M	29.0	179.0	89.0	Poland-1	POL	1976 Winter	1976
271112	135570	Piotr ya	M	27.0	176.0	59.0	Poland	POL	2014 Winter	2014
271113	135570	Piotr ya	M	27.0	176.0	59.0	Poland	POL	2014 Winter	2014
271114	135571	Tomasz Ireneusz ya	M	30.0	185.0	96.0	Poland	POL	1998 Winter	1998
271115	135571	Tomasz Ireneusz ya	M	34.0	185.0	96.0	Poland	POL	2002 Winter	2002

271116 rows × 17 columns

In [15]:

```
# tell about how many row and columns are present in  the dataframe
athletes_df.shape
```

Out[15]:

```
(271116, 17)
```

In [16]:

```
athletes_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 271116 entries, 0 to 271115
Data columns (total 17 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 
 15  region  270746 non-null  object 
 16  notes   5039 non-null    object 
dtypes: float64(3), int64(2), object(12)
memory usage: 37.2+ MB
```

In [17]:

```
athletes_df.isnull().sum() # it gives you pandas series of column names along with the s
```

Out[17]:

```
ID             0
Name           0
Sex            0
Age         9474
Height     60171
Weight     62875
Team           0
NOC            0
Games          0
Year           0
Season         0
City           0
Sport          0
Event          0
Medal     231333
region       370
notes     266077
dtype: int64
```

In [18]:

```
# rename the columns
athletes_df.rename(columns={'region':'Region','notes':'Notes'},inplace = True)
```

In [19]:

```
athletes_df
```

Out[19]:

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year
0	1	A Dijiang	M	24.0	180.0	80.0	China	CHN	1992 Summer	1992
1	2	A Lamusi	M	23.0	170.0	60.0	China	CHN	2012 Summer	2012
2	3	Gunnar Nielsen Aaby	M	24.0	NaN	NaN	Denmark	DEN	1920 Summer	1920
3	4	Edgar Lindenau Aabye	M	34.0	NaN	NaN	Denmark/Sweden	DEN	1900 Summer	1900
4	5	Christine Jacoba Aaftink	F	21.0	185.0	82.0	Netherlands	NED	1988 Winter	1988
...	...	...	...	...	...	...	...	...	...	...
271111	135569	Andrzej ya	M	29.0	179.0	89.0	Poland-1	POL	1976 Winter	1976
271112	135570	Piotr ya	M	27.0	176.0	59.0	Poland	POL	2014 Winter	2014
271113	135570	Piotr ya	M	27.0	176.0	59.0	Poland	POL	2014 Winter	2014
271114	135571	Tomasz Ireneusz ya	M	30.0	185.0	96.0	Poland	POL	1998 Winter	1998
271115	135571	Tomasz Ireneusz ya	M	34.0	185.0	96.0	Poland	POL	2002 Winter	2002

271116 rows × 17 columns

In [20]:

```
 # check null values in each columns
nan_values = athletes_df.isna()
```

In [21]:

```
nan_values
```

Out[21]:

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	C
0	False	False	False	False	False	False	False	False	False	False	False	Fa
1	False	False	False	False	False	False	False	False	False	False	False	Fa
2	False	False	False	False	True	True	False	False	False	False	False	Fa
3	False	False	False	False	True	True	False	False	False	False	False	Fa
4	False	False	False	False	False	False	False	False	False	False	False	Fa
...	...	...	...	...	...	...	...	...	...	...	...	
271111	False	False	False	False	False	False	False	False	False	False	False	Fa
271112	False	False	False	False	False	False	False	False	False	False	False	Fa
271113	False	False	False	False	False	False	False	False	False	False	False	Fa
271114	False	False	False	False	False	False	False	False	False	False	False	Fa
271115	False	False	False	False	False	False	False	False	False	False	False	Fa

271116 rows × 17 columns

In [22]:

```
nan_values.any()
```

Out[22]:

```
ID         False
Name       False
Sex        False
Age         True
Height      True
Weight      True
Team       False
NOC        False
Games      False
Year       False
Season     False
City       False
Sport      False
Event      False
Medal       True
Region      True
Notes       True
dtype: bool
```

In [23]:

```
nan_values = athletes_df.isna().any() # we can also uses both function tagether
```

In [24]:

```
nan_values
```

Out[24]:

```
ID        False
Name      False
Sex       False
Age        True
Height     True
Weight     True
Team      False
NOC       False
Games     False
Year      False
Season    False
City      False
Sport     False
Event     False
Medal      True
Region     True
Notes      True
dtype: bool
```

In [25]:

```
# to find the first five row where team is equal to india
athletes_df.query('Team == "India"').head(5)
```

Out[25]:

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	
505	281	S. Abdul Hamid	M	NaN	NaN	NaN	India	IND	1928 Summer	1928	Summer	Amste
506	281	S. Abdul Hamid	M	NaN	NaN	NaN	India	IND	1928 Summer	1928	Summer	Amste
895	512	Shiny Kurisingal Abraham-Wilson	F	19.0	167.0	53.0	India	IND	1984 Summer	1984	Summer	An
896	512	Shiny Kurisingal Abraham-Wilson	F	19.0	167.0	53.0	India	IND	1984 Summer	1984	Summer	An
897	512	Shiny Kurisingal Abraham-Wilson	F	23.0	167.0	53.0	India	IND	1988 Summer	1988	Summer	

In [26]:

```
# to find the first five row where team is equal to pakistan
athletes_df.query('Team == "Pakistan"').head(5)
```

Out[26]:

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	
233	111	Aqarab Abbas	M	22.0	190.0	88.0	Pakistan	PAK	1996 Summer	1996	Summer	
237	115	Ghulam Abbas	M	24.0	181.0	74.0	Pakistan	PAK	1992 Summer	1992	Summer	E
245	121	Muhammad Abbas	M	23.0	168.0	55.0	Pakistan	PAK	2010 Winter	2010	Winter	V
247	123	Sohail Abbas	M	25.0	178.0	80.0	Pakistan	PAK	2000 Summer	2000	Summer	
248	123	Sohail Abbas	M	29.0	178.0	80.0	Pakistan	PAK	2004 Summer	2004	Summer	

In [27]:

```
#  to find the first five row where city is equal to sydney
athletes_df.query('City == "Sydney"').head(5)
```

Out[27]:

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	
31	12	Jyri Tapani Aalto	M	31.0	172.0	70.0	Finland	FIN	2000 Summer	2000	Summer	Sy
33	13	Minna Maarit Aalto	F	34.0	159.0	55.5	Finland	FIN	2000 Summer	2000	Summer	Sy
57	18	Timo Antero Aaltonen	M	31.0	189.0	130.0	Finland	FIN	2000 Summer	2000	Summer	Sy
81	23	Fritz Aanes	M	22.0	187.0	89.0	Norway	NOR	2000 Summer	2000	Summer	Sy
93	30	Pepijn Aardewijn	M	30.0	189.0	72.0	Netherlands	NED	2000 Summer	2000	Summer	Sy

In [28]:

```
#  to find the first five row where team is equal to india and sport is equal to badmint
athletes_df.query('Team == "India" and Sport == "Badminton" and Year == 2000').head(5)
```

Out[28]:

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season
82198	41744	Pulella Gopichand	M	26.0	182.0	63.0	India	IND	2000 Summer	2000	Summer
191419	96125	Lalji Aparna Popat	F	22.0	160.0	65.0	India	IND	2000 Summer	2000	Summer

In [29]:

```
athletes_df.query('Team == "India" or Sport == "Badminton" or Year == 2000').head(5)
```

Out[29]:

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	
31	12	Jyri Tapani Aalto	M	31.0	172.0	70.0	Finland	FIN	2000 Summer	2000	Summer	Sy
33	13	Minna Maarit Aalto	F	34.0	159.0	55.5	Finland	FIN	2000 Summer	2000	Summer	Sy
57	18	Timo Antero Aaltonen	M	31.0	189.0	130.0	Finland	FIN	2000 Summer	2000	Summer	Sy
81	23	Fritz Aanes	M	22.0	187.0	89.0	Norway	NOR	2000 Summer	2000	Summer	Sy
93	30	Pepijn Aardewijn	M	30.0	189.0	72.0	Netherlands	NED	2000 Summer	2000	Summer	Sy

In [30]:

```
athletes_male = athletes_df.query('Sex == "M"') # where sex is equal to male
```

In [31]:

```
athletes_male
```

Out[31]:

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year
0	1	A Dijiang	M	24.0	180.0	80.0	China	CHN	1992 Summer	1992
1	2	A Lamusi	M	23.0	170.0	60.0	China	CHN	2012 Summer	2012
2	3	Gunnar Nielsen Aaby	M	24.0	NaN	NaN	Denmark	DEN	1920 Summer	1920
3	4	Edgar Lindenau Aabye	M	34.0	NaN	NaN	Denmark/Sweden	DEN	1900 Summer	1900
10	6	Per Knut Aaland	M	31.0	188.0	75.0	United States	USA	1992 Winter	1992
...	...	...	...	...	...	...	...	...	...	...
271111	135569	Andrzej ya	M	29.0	179.0	89.0	Poland-1	POL	1976 Winter	1976
271112	135570	Piotr ya	M	27.0	176.0	59.0	Poland	POL	2014 Winter	2014
271113	135570	Piotr ya	M	27.0	176.0	59.0	Poland	POL	2014 Winter	2014
271114	135571	Tomasz Ireneusz ya	M	30.0	185.0	96.0	Poland	POL	1998 Winter	1998
271115	135571	Tomasz Ireneusz ya	M	34.0	185.0	96.0	Poland	POL	2002 Winter	2002

196594 rows × 17 columns

In [32]:

```
athletes_male.info()  # gives the information about total male candidates
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 196594 entries, 0 to 271115
Data columns (total 17 columns):
 #   Column  Non-Null Count   Dtype  
---  ------  --------------   -----  
 0   ID      196594 non-null  int64  
 1   Name    196594 non-null  object 
 2   Sex     196594 non-null  object 
 3   Age     187544 non-null  float64
 4   Height  143567 non-null  float64
 5   Weight  141470 non-null  float64
 6   Team    196594 non-null  object 
 7   NOC     196594 non-null  object 
 8   Games   196594 non-null  object 
 9   Year    196594 non-null  int64  
 10  Season  196594 non-null  object 
 11  City    196594 non-null  object 
 12  Sport   196594 non-null  object 
 13  Event   196594 non-null  object 
 14  Medal   28530 non-null   object 
 15  Region  196360 non-null  object 
 16  Notes   4138 non-null    object 
dtypes: float64(3), int64(2), object(12)
memory usage: 27.0+ MB
```

In [33]:

```
athletes_male.shape
```

Out[33]:

```
(196594, 17)
```

In [34]:

```
athletes_df.shape
```

Out[34]:

```
(271116, 17)
```

In [35]:

```
# selecting the each coloumns using []
athletes_df['Team']
```

Out[35]:

```
0                  China
1                  China
2                Denmark
3         Denmark/Sweden
4            Netherlands
                ...
271111          Poland-1
271112            Poland
271113            Poland
271114            Poland
271115            Poland
Name: Team, Length: 271116, dtype: object
```

In [36]:

```
athletes_df[['Team','City']] # Select multiple columns.
```

Out[36]:

	Team	City
0	China	Barcelona
1	China	London
2	Denmark	Antwerpen
3	Denmark/Sweden	Paris
4	Netherlands	Calgary
...	...	...
271111	Poland-1	Innsbruck
271112	Poland	Sochi
271113	Poland	Sochi
271114	Poland	Nagano
271115	Poland	Salt Lake City

271116 rows × 2 columns

In [37]:

```
athletes_df.iloc[2]                    # Select Rows by index in Pandas DataFrame using i
```

Out[37]:

```
ID                              3
Name          Gunnar Nielsen Aaby
Sex                             M
Age                          24.0
Height                        NaN
Weight                        NaN
Team                      Denmark
NOC                           DEN
Games                 1920 Summer
Year                         1920
Season                     Summer
City                    Antwerpen
Sport                    Football
Event     Football Men's Football
Medal                         NaN
Region                    Denmark
Notes                         NaN
Name: 2, dtype: object
```

In [38]:

```
athletes_df.iloc[2:6]
```

Out[38]:

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season
2	3	Gunnar Nielsen Aaby	M	24.0	NaN	NaN	Denmark	DEN	1920 Summer	1920	Summer
3	4	Edgar Lindenau Aabye	M	34.0	NaN	NaN	Denmark/Sweden	DEN	1900 Summer	1900	Summer
4	5	Christine Jacoba Aaftink	F	21.0	185.0	82.0	Netherlands	NED	1988 Winter	1988	Winter
5	5	Christine Jacoba Aaftink	F	21.0	185.0	82.0	Netherlands	NED	1988 Winter	1988	Winter

In [39]:

```
athletes_df.iloc[[2,3,4]]
```

Out[39]:

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season
2	3	Gunnar Nielsen Aaby	M	24.0	NaN	NaN	Denmark	DEN	1920 Summer	1920	Summer
3	4	Edgar Lindenau Aabye	M	34.0	NaN	NaN	Denmark/Sweden	DEN	1900 Summer	1900	Summer
4	5	Christine Jacoba Aaftink	F	21.0	185.0	82.0	Netherlands	NED	1988 Winter	1988	Winter

In [40]:

```
athletes_df.iloc[[2,3,4],[1,2,11]]  #Select multiple rows with some particular columns
```

Out[40]:

	Name	Sex	City
2	Gunnar Nielsen Aaby	M	Antwerpen
3	Edgar Lindenau Aabye	M	Paris
4	Christine Jacoba Aaftink	F	Calgary

In [41]:

```
athletes_df.iloc[[2],[3]]
```

Out[41]:

	Age
2	24.0

In [42]:

```
athletes_df.isnull().sum()
```

Out[42]:

```
ID             0
Name           0
Sex            0
Age         9474
Height     60171
Weight     62875
Team           0
NOC            0
Games          0
Year           0
Season         0
City           0
Sport          0
Event          0
Medal     231333
Region       370
Notes     266077
dtype: int64
```

In [43]:

```
 # drop a column
athletes_df.drop(['Notes'],axis = 1,inplace = True)
# athletes_df.drop(column = ['notes'])
```

In [44]:

```
athletes_df
```

Out[44]:

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year
0	1	A Dijiang	M	24.0	180.0	80.0	China	CHN	1992 Summer	1992
1	2	A Lamusi	M	23.0	170.0	60.0	China	CHN	2012 Summer	2012
2	3	Gunnar Nielsen Aaby	M	24.0	NaN	NaN	Denmark	DEN	1920 Summer	1920
3	4	Edgar Lindenau Aabye	M	34.0	NaN	NaN	Denmark/Sweden	DEN	1900 Summer	1900
4	5	Christine Jacoba Aaftink	F	21.0	185.0	82.0	Netherlands	NED	1988 Winter	1988
...	...	...	...	...	...	...	...	...	...	...
271111	135569	Andrzej ya	M	29.0	179.0	89.0	Poland-1	POL	1976 Winter	1976
271112	135570	Piotr ya	M	27.0	176.0	59.0	Poland	POL	2014 Winter	2014
271113	135570	Piotr ya	M	27.0	176.0	59.0	Poland	POL	2014 Winter	2014
271114	135571	Tomasz Ireneusz ya	M	30.0	185.0	96.0	Poland	POL	1998 Winter	1998
271115	135571	Tomasz Ireneusz ya	M	34.0	185.0	96.0	Poland	POL	2002 Winter	2002

271116 rows × 16 columns

In [45]:

```
# to find the mean
x = int(athletes_df["Height"].mean())
y = int(athletes_df["Weight"].mean())
```

In [46]:

```
x
```

Out[46]:

```
175
```

In [47]:

```
y
```

Out[47]:

```
70
```

In [48]:

```
x = 175.0
y = 70.0
```

In [49]:

```
athletes_df["Height"].fillna(x , inplace = True) # to fill null value with mean
```

In [50]:

```
athletes_df["Weight"].fillna(y , inplace = True) # to fill null value with mean
```

In [51]:

```
athletes_df
```

Out[51]:

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year
0	1	A Dijiang	M	24.0	180.0	80.0	China	CHN	1992 Summer	1992
1	2	A Lamusi	M	23.0	170.0	60.0	China	CHN	2012 Summer	2012
2	3	Gunnar Nielsen Aaby	M	24.0	175.0	70.0	Denmark	DEN	1920 Summer	1920
3	4	Edgar Lindenau Aabye	M	34.0	175.0	70.0	Denmark/Sweden	DEN	1900 Summer	1900
4	5	Christine Jacoba Aaftink	F	21.0	185.0	82.0	Netherlands	NED	1988 Winter	1988
...	...	...	...	...	...	...	...	...	...	...
271111	135569	Andrzej ya	M	29.0	179.0	89.0	Poland-1	POL	1976 Winter	1976
271112	135570	Piotr ya	M	27.0	176.0	59.0	Poland	POL	2014 Winter	2014
271113	135570	Piotr ya	M	27.0	176.0	59.0	Poland	POL	2014 Winter	2014
271114	135571	Tomasz Ireneusz ya	M	30.0	185.0	96.0	Poland	POL	1998 Winter	1998
271115	135571	Tomasz Ireneusz ya	M	34.0	185.0	96.0	Poland	POL	2002 Winter	2002

271116 rows × 16 columns

In [52]:

```
athletes_df.isnull().sum()
```

Out[52]:

```
ID             0
Name           0
Sex            0
Age         9474
Height         0
Weight         0
Team           0
NOC            0
Games          0
Year           0
Season         0
City           0
Sport          0
Event          0
Medal     231333
Region       370
dtype: int64
```

In [53]:

```
z = int(athletes_df["Age"].mode())
```

In [54]:

```
z =23.0
```

In [55]:

```
z
```

Out[55]:

```
23.0
```

In [56]:

```
athletes_df["Age"].fillna(z,inplace=True) # to fill null value with mode
```

In [57]:

```
athletes_df.isnull().sum()
```

Out[57]:

```
ID             0
Name           0
Sex            0
Age            0
Height         0
Weight         0
Team           0
NOC            0
Games          0
Year           0
Season         0
City           0
Sport          0
Event          0
Medal     231333
Region       370
dtype: int64
```

In [58]:

```
athletes_df["Medal"]
```

Out[58]:

```
0          NaN
1          NaN
2          NaN
3         Gold
4          NaN
          ...
271111     NaN
271112     NaN
271113     NaN
271114     NaN
271115     NaN
Name: Medal, Length: 271116, dtype: object
```

In [59]:

```
athletes_df["Medal"].fillna("No Medal", inplace =True) # fill null value of medal columr
```

In [60]:

```
athletes_df.isnull().sum()
```

Out[60]:

```
ID          0
Name        0
Sex         0
Age         0
Height      0
Weight      0
Team        0
NOC         0
Games       0
Year        0
Season      0
City        0
Sport       0
Event       0
Medal       0
Region    370
dtype: int64
```

In [61]:

```
athletes_df["Region"].fillna("Unknown", inplace =True) # fill null value of region colum
```

In [62]:

```
athletes_df.isnull().sum() # dataframe has zero null value
```

Out[62]:

```
ID        0
Name      0
Sex       0
Age       0
Height    0
Weight    0
Team      0
NOC       0
Games     0
Year      0
Season    0
City      0
Sport     0
Event     0
Medal     0
Region    0
dtype: int64
```

In [63]:

```
# upload the file in to csv
athletes_df.to_csv("athletes_dataset.csv")
```

In [64]:

```
df1 = pd.read_csv("athletes_dataset.csv") # check the file
```

In [65]:

```
df1
```

Out[65]:

	Unnamed: 0	ID	Name	Sex	Age	Height	Weight	Team	NOC	G
0	0	1	A Dijiang	M	24.0	180.0	80.0	China	CHN	Su
1	1	2	A Lamusi	M	23.0	170.0	60.0	China	CHN	Su
2	2	3	Gunnar Nielsen Aaby	M	24.0	175.0	70.0	Denmark	DEN	Su
3	3	4	Edgar Lindenau Aabye	M	34.0	175.0	70.0	Denmark/Sweden	DEN	Su
4	4	5	Christine Jacoba Aaftink	F	21.0	185.0	82.0	Netherlands	NED	W
...	...	...	...	...	...	...	...	...	...	
271111	271111	135569	Andrzej ya	M	29.0	179.0	89.0	Poland-1	POL	W
271112	271112	135570	Piotr ya	M	27.0	176.0	59.0	Poland	POL	W
271113	271113	135570	Piotr ya	M	27.0	176.0	59.0	Poland	POL	W
271114	271114	135571	Tomasz Ireneusz ya	M	30.0	185.0	96.0	Poland	POL	W
271115	271115	135571	Tomasz Ireneusz ya	M	34.0	185.0	96.0	Poland	POL	W

271116 rows × 17 columns

In [66]:

```
df1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 271116 entries, 0 to 271115
Data columns (total 17 columns):
 #   Column      Non-Null Count   Dtype  
---  ------      --------------   -----  
 0   Unnamed: 0  271116 non-null  int64  
 1   ID          271116 non-null  int64  
 2   Name        271116 non-null  object 
 3   Sex         271116 non-null  object 
 4   Age         271116 non-null  float64
 5   Height      271116 non-null  float64
 6   Weight      271116 non-null  float64
 7   Team        271116 non-null  object 
 8   NOC         271116 non-null  object 
 9   Games       271116 non-null  object 
 10  Year        271116 non-null  int64  
 11  Season      271116 non-null  object 
 12  City        271116 non-null  object 
 13  Sport       271116 non-null  object 
 14  Event       271116 non-null  object 
 15  Medal       271116 non-null  object 
 16  Region      271116 non-null  object 
dtypes: float64(3), int64(3), object(11)
memory usage: 35.2+ MB
```

In [72]:

```
 # we have to drop some more columns which are not useful
athletes_df.drop(columns=['ID','Games'],inplace = True)
```

In [73]:

```
athletes_df.to_csv("athletes_dataset_new.csv") # upload the file
```

In [74]:

```
df2 = pd.read_csv("athletes_dataset_new.csv")
```

In [75]:

```
df2
```

Out[75]:

	Unnamed: 0	Name	Sex	Age	Height	Weight	Team	NOC	Year	Seas
0	0	A Dijiang	M	24.0	180.0	80.0	China	CHN	1992	Sumn
1	1	A Lamusi	M	23.0	170.0	60.0	China	CHN	2012	Sumn
2	2	Gunnar Nielsen Aaby	M	24.0	175.0	70.0	Denmark	DEN	1920	Sumn
3	3	Edgar Lindenau Aabye	M	34.0	175.0	70.0	Denmark/Sweden	DEN	1900	Sumn
4	4	Christine Jacoba Aaftink	F	21.0	185.0	82.0	Netherlands	NED	1988	Wir
...	...	...	...	...	...	...	...	...	...	
271111	271111	Andrzej ya	M	29.0	179.0	89.0	Poland-1	POL	1976	Wir
271112	271112	Piotr ya	M	27.0	176.0	59.0	Poland	POL	2014	Wir
271113	271113	Piotr ya	M	27.0	176.0	59.0	Poland	POL	2014	Wir
271114	271114	Tomasz Ireneusz ya	M	30.0	185.0	96.0	Poland	POL	1998	Wir
271115	271115	Tomasz Ireneusz ya	M	34.0	185.0	96.0	Poland	POL	2002	Wir

271116 rows × 15 columns

In [76]:

```
df2.info()
```

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

In [ ]:

```

```