SportsStats

March 14, 2024

[2]:
```
import pandas as pd
```

[4]:
```
import os
```

[5]:
```
athlete_events = pd.read_csv('/home/jovyan/work/athlete_events.csv')
```

[30]:
```
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
```

[31]:
```
athlete_events.head(25)
```

[31]:

	ID	Name	Sex	Age	Height	Weight	Team	\
0	1	A Dijiang	M	24.0	180.0	80.0	China	
1	2	A Lamusi	M	23.0	170.0	60.0	China	
2	3	Gunnar Nielsen Aaby	M	24.0	NaN	NaN	Denmark	
3	4	Edgar Lindenau Aabye	M	34.0	NaN	NaN	Denmark/Sweden	

4	5	Christine Jacoba Aaftink	F	21.0	185.0	82.0	Netherlands
5	5	Christine Jacoba Aaftink	F	21.0	185.0	82.0	Netherlands
6	5	Christine Jacoba Aaftink	F	25.0	185.0	82.0	Netherlands
7	5	Christine Jacoba Aaftink	F	25.0	185.0	82.0	Netherlands
8	5	Christine Jacoba Aaftink	F	27.0	185.0	82.0	Netherlands
9	5	Christine Jacoba Aaftink	F	27.0	185.0	82.0	Netherlands
10	6	Per Knut Aaland	M	31.0	188.0	75.0	United States
11	6	Per Knut Aaland	M	31.0	188.0	75.0	United States
12	6	Per Knut Aaland	M	31.0	188.0	75.0	United States
13	6	Per Knut Aaland	M	31.0	188.0	75.0	United States
14	6	Per Knut Aaland	M	33.0	188.0	75.0	United States
15	6	Per Knut Aaland	M	33.0	188.0	75.0	United States
16	6	Per Knut Aaland	M	33.0	188.0	75.0	United States
17	6	Per Knut Aaland	M	33.0	188.0	75.0	United States
18	7	John Aalberg	M	31.0	183.0	72.0	United States
19	7	John Aalberg	M	31.0	183.0	72.0	United States
20	7	John Aalberg	M	31.0	183.0	72.0	United States
21	7	John Aalberg	M	31.0	183.0	72.0	United States
22	7	John Aalberg	M	33.0	183.0	72.0	United States
23	7	John Aalberg	M	33.0	183.0	72.0	United States
24	7	John Aalberg	M	33.0	183.0	72.0	United States

	NOC	Games	Year	Season	City	Sport	\
0	CHN	1992 Summer	1992	Summer	Barcelona	Basketball	
1	CHN	2012 Summer	2012	Summer	London	Judo	
2	DEN	1920 Summer	1920	Summer	Antwerpen	Football	
3	DEN	1900 Summer	1900	Summer	Paris	Tug-Of-War	
4	NED	1988 Winter	1988	Winter	Calgary	Speed Skating	
5	NED	1988 Winter	1988	Winter	Calgary	Speed Skating	
6	NED	1992 Winter	1992	Winter	Albertville	Speed Skating	
7	NED	1992 Winter	1992	Winter	Albertville	Speed Skating	
8	NED	1994 Winter	1994	Winter	Lillehammer	Speed Skating	
9	NED	1994 Winter	1994	Winter	Lillehammer	Speed Skating	
10	USA	1992 Winter	1992	Winter	Albertville	Cross Country Skiing	
11	USA	1992 Winter	1992	Winter	Albertville	Cross Country Skiing	
12	USA	1992 Winter	1992	Winter	Albertville	Cross Country Skiing	
13	USA	1992 Winter	1992	Winter	Albertville	Cross Country Skiing	
14	USA	1994 Winter	1994	Winter	Lillehammer	Cross Country Skiing	
15	USA	1994 Winter	1994	Winter	Lillehammer	Cross Country Skiing	
16	USA	1994 Winter	1994	Winter	Lillehammer	Cross Country Skiing	
17	USA	1994 Winter	1994	Winter	Lillehammer	Cross Country Skiing	
18	USA	1992 Winter	1992	Winter	Albertville	Cross Country Skiing	
19	USA	1992 Winter	1992	Winter	Albertville	Cross Country Skiing	
20	USA	1992 Winter	1992	Winter	Albertville	Cross Country Skiing	
21	USA	1992 Winter	1992	Winter	Albertville	Cross Country Skiing	
22	USA	1994 Winter	1994	Winter	Lillehammer	Cross Country Skiing	
23	USA	1994 Winter	1994	Winter	Lillehammer	Cross Country Skiing	

```
24  USA  1994 Winter  1994  Winter  Lillehammer  Cross Country Skiing
```

	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
5	Speed Skating Women's 1,000 metres	NaN
6	Speed Skating Women's 500 metres	NaN
7	Speed Skating Women's 1,000 metres	NaN
8	Speed Skating Women's 500 metres	NaN
9	Speed Skating Women's 1,000 metres	NaN
10	Cross Country Skiing Men's 10 kilometres	NaN
11	Cross Country Skiing Men's 50 kilometres	NaN
12	Cross Country Skiing Men's 10/15 kilometres Pu…	NaN
13	Cross Country Skiing Men's 4 x 10 kilometres R…	NaN
14	Cross Country Skiing Men's 10 kilometres	NaN
15	Cross Country Skiing Men's 30 kilometres	NaN
16	Cross Country Skiing Men's 10/15 kilometres Pu…	NaN
17	Cross Country Skiing Men's 4 x 10 kilometres R…	NaN
18	Cross Country Skiing Men's 10 kilometres	NaN
19	Cross Country Skiing Men's 50 kilometres	NaN
20	Cross Country Skiing Men's 10/15 kilometres Pu…	NaN
21	Cross Country Skiing Men's 4 x 10 kilometres R…	NaN
22	Cross Country Skiing Men's 10 kilometres	NaN
23	Cross Country Skiing Men's 30 kilometres	NaN
24	Cross Country Skiing Men's 10/15 kilometres Pu…	NaN

[6]:
```
regions = pd.read_csv('/home/jovyan/work/noc_regions.csv')
```

[33]:
```
regions.head(25)
```

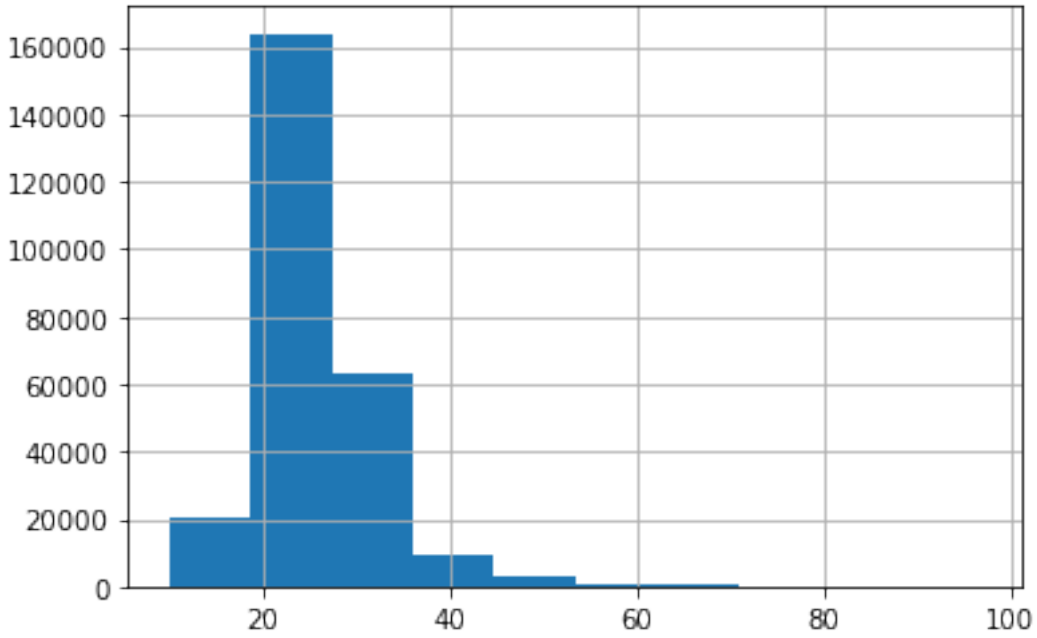
[33]:

	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
20  BEN                   Benin                 NaN
21  BER                 Bermuda                 NaN
22  BHU                  Bhutan                 NaN
23  BIH  Bosnia and Herzegovina                 NaN
24  BIZ                  Belize                 NaN
```

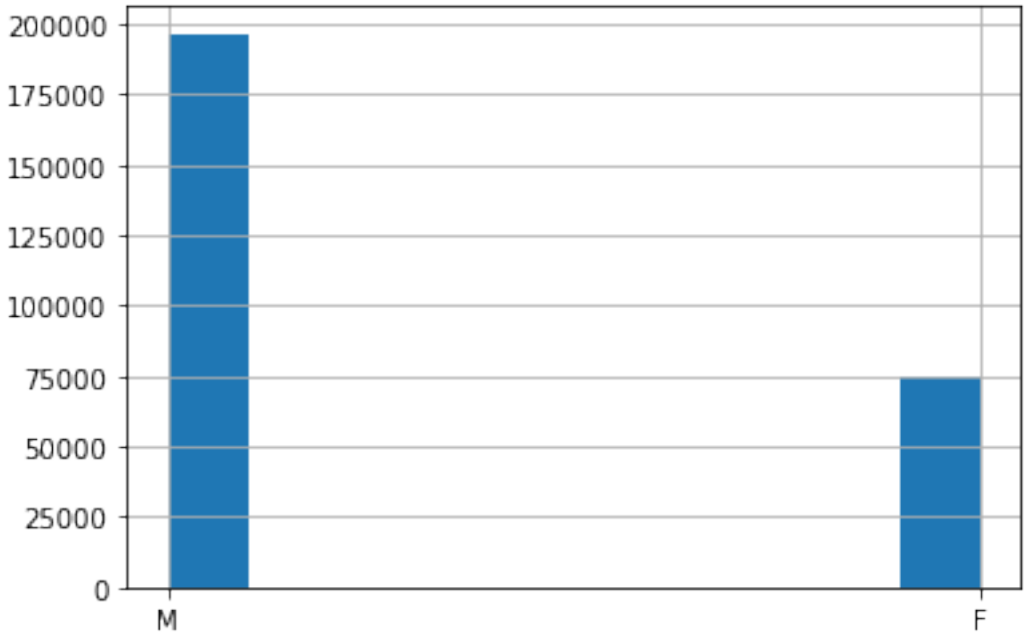
[34]:
```
athlete_events.Age.hist()
```

[34]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f4b8e532190>
```



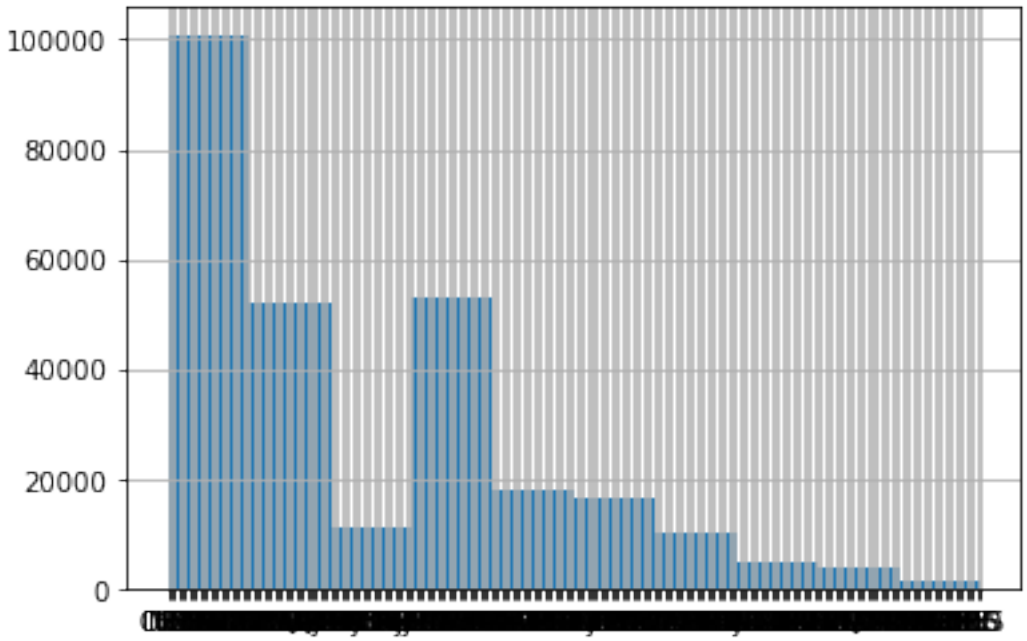
[35]:
```
athlete_events.Sex.hist()
```

[35]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f4b8e0aef10>
```



[36]:
```
athlete_events.NOC.hist()
```

[36]: <matplotlib.axes._subplots.AxesSubplot at 0x7f4b89fa9590>



[7]:
```
from pandasql import sqldf
pysqldf = lambda q: sqldf(q, globals())
```

[38]:
```
pysqldf("Select * from athlete_events;")
```

[38]:

	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]

[ ]:
```
# Count number of male and female athletes
```

[39]:
```
pysqldf("Select Count (*) from athlete_events where Sex = 'M';")
```

[39]:

	Count (*)
0	196594

[40]:
```
pysqldf("Select Count (*) from athlete_events where Sex = 'F';")
```

[40]:

	Count (*)
0	74522

[41]:
```
# stats for male athletes
male_stats= pysqldf("Select avg(Age) as AVG_Age, avg(Weight) as AVG_Weight,␣
 ↪avg(Height) as AVG_Height, Sport From athlete_events where Sex ='M' group by␣
 ↪Sport;")
```

[42]:
```
# stats for female athletes
female_stats= pysqldf("Select avg(Age) as AVG_Age, avg(Weight) as AVG_Weight,␣
 ↪avg(Height) as AVG_Height, Sport From athlete_events where Sex ='F' group by␣
 ↪Sport;")
```

[41]:
```
pysqldf("Select * from male_stats;")
```

[41]:

	AVG_Age	AVG_Weight	AVG_Height	Sport
0	26.000000	NaN	NaN	Aeronautics
1	23.758266	78.626035	177.891374	Alpine Skiing
2	38.533333	NaN	NaN	Alpinism
3	29.083267	77.066866	178.477842	Archery
4	46.062816	75.290909	174.896552	Art Competitions
..	…	…	…	…
58	29.309524	95.615385	182.480000	Tug-Of-War
59	25.810475	86.925926	193.265660	Volleyball
60	25.736542	87.706172	186.801739	Water Polo
61	25.710832	80.251796	169.153061	Weightlifting
62	25.821827	76.400640	172.870686	Wrestling

[63 rows x 4 columns]

[42]:
```
pysqldf("Select * from female_stats;")
```

[42]:

	AVG_Age	AVG_Weight	AVG_Height	Sport
0	22.334609	62.640307	167.221001	Alpine Skiing
1	43.000000	NaN	NaN	Alpinism
2	26.508458	62.013575	167.166483	Archery
3	44.411960	NaN	160.000000	Art Competitions
4	24.933574	60.239798	169.285714	Athletics
5	25.047297	61.586364	168.438519	Badminton
6	25.517516	73.685170	182.454836	Basketball

7	28.315217	68.350943	178.866667	Beach Volleyball
8	25.799249	57.306586	166.540541	Biathlon
9	27.832168	72.804196	173.181818	Bobsleigh
10	26.638889	61.836066	168.800000	Boxing
11	25.337446	64.878906	169.628995	Canoeing
12	39.500000	NaN	NaN	Croquet
13	25.655244	57.488018	166.650819	Cross Country Skiing
14	29.972973	62.888350	167.520362	Curling
15	26.899570	59.770553	168.018629	Cycling
16	21.517677	53.566563	161.173604	Diving
17	34.217846	58.601908	167.875755	Equestrianism
18	26.341105	60.656144	168.954292	Fencing
19	20.812554	49.861039	160.610849	Figure Skating
20	24.909091	60.925813	167.676142	Football
21	24.330254	58.352459	164.777262	Freestyle Skiing
22	27.338235	63.436364	168.733333	Golf
23	19.232751	47.791276	156.143325	Gymnastics
24	25.883922	68.876851	174.840278	Handball
25	25.332419	60.530935	166.125267	Hockey
26	24.055703	65.712865	168.209549	Ice Hockey
27	25.163769	67.067164	166.267000	Judo
28	23.649867	66.908832	169.065527	Luge
29	25.524390	58.310976	170.073171	Modern Pentathlon
30	26.000000	NaN	NaN	Motorboating
31	18.737082	48.760976	167.870253	Rhythmic Gymnastics
32	25.402645	70.102214	176.771079	Rowing
33	26.114865	66.628378	167.636986	Rugby Sevens
34	26.698378	62.777268	169.510808	Sailing
35	29.119048	61.143243	164.932934	Shooting
36	22.462549	57.123139	164.530447	Short Track Speed Skating
37	27.560606	61.000000	167.923077	Skeleton
38	21.266667	52.615385	164.600000	Ski Jumping
39	24.723558	60.444175	166.298077	Snowboarding
40	26.299163	67.471655	169.395089	Softball
41	23.746276	62.010046	167.420704	Speed Skating
42	19.487450	61.482748	171.468735	Swimming
43	22.366851	55.863529	168.481481	Synchronized Swimming
44	25.665964	58.027778	165.064965	Table Tennis
45	23.408027	61.136824	170.811644	Taekwondo
46	24.775457	62.094398	172.335736	Tennis
47	25.539474	52.893333	161.733333	Trampolining
48	27.923954	54.724138	166.996183	Triathlon
49	24.431627	69.333779	179.494983	Volleyball
50	25.161885	70.180328	175.563525	Water Polo
51	24.028078	67.724622	160.467391	Weightlifting
52	25.305921	60.554455	163.865132	Wrestling

[ ]:

[43]:
```
pysqldf("Select Count(*) from athlete_events where Sex IS NULL;")
```

[43]:
```
   Count(*)
0         0
```

[44]:
```
pysqldf("Select Count(*) from athlete_events where Age IS NULL;")
```

[44]:
```
   Count(*)
0      9474
```

[45]:
```
pysqldf("Select Count(*) from athlete_events where Weight IS NULL;")
```

[45]:
```
   Count(*)
0     62875
```

[48]:
```
pysqldf("Select Count(*) from athlete_events where Height IS NULL;")
```

[48]:
```
   Count(*)
0     60171
```

[46]:
```
pysqldf("Select Count(*) from athlete_events where Team IS NULL;")
```

[46]:
```
   Count(*)
0         0
```

[50]:
```
pysqldf("Select Count(*) from athlete_events where NOC IS NULL;")
```

[50]:
```
   Count(*)
0         0
```

[47]:
```
pysqldf("Select Count(*) from athlete_events where Games IS NULL;")
```

[47]:
```
   Count(*)
0         0
```

[52]:
```
pysqldf("Select Count(*) from athlete_events where Year IS NULL;")
```

[52]:
```
   Count(*)
0         0
```

[48]:
```
pysqldf("Select Count(*) from athlete_events where Medal IS NULL;")
```

[48]:
```
   Count(*)
0    231333
```

[54]:
```
pysqldf("Select Count(*) from athlete_events where Sport IS NULL;")
```

[54]:

	Count(*)
0	0

[61]:
```
Cities = pysqldf('''SELECT DISTINCT City as City_name
                    FROM athlete_events
                    Group by City;''' )
```

[62]:
```
pysqldf("SELECT * From Cities;")
```

[62]:

	City_name
0	Albertville
1	Amsterdam
2	Antwerpen
3	Athina
4	Atlanta
5	Barcelona
6	Beijing
7	Berlin
8	Calgary
9	Chamonix
10	Cortina d'Ampezzo
11	Garmisch-Partenkirchen
12	Grenoble
13	Helsinki
14	Innsbruck
15	Lake Placid
16	Lillehammer
17	London
18	Los Angeles
19	Melbourne
20	Mexico City
21	Montreal
22	Moskva
23	Munich
24	Nagano
25	Oslo
26	Paris
27	Rio de Janeiro
28	Roma
29	Salt Lake City
30	Sankt Moritz
31	Sapporo
32	Sarajevo
33	Seoul
34	Sochi

```
35            Squaw Valley
36               St. Louis
37               Stockholm
38                  Sydney
39                   Tokyo
40                  Torino
41               Vancouver
```

[59]:
```
Medals = pysqldf('''SELECT DISTINCT Medal as Medal_type
                    FROM athlete_events
                    Group by Medal;''' )
```

[60]:
```
pysqldf("SELECT * From Medals;")
```

[60]:
```
  Medal_type
0       None
1     Bronze
2       Gold
3     Silver
```

[57]:
```
pysqldf("SELECT * From athlete_events Where ID = 15;")
```

[57]:
```
   ID                  Name Sex   Age Height Weight     Team  NOC  \
0  15  Arvo Ossian Aaltonen   M  22.0   None   None  Finland  FIN
1  15  Arvo Ossian Aaltonen   M  22.0   None   None  Finland  FIN
2  15  Arvo Ossian Aaltonen   M  30.0   None   None  Finland  FIN
3  15  Arvo Ossian Aaltonen   M  30.0   None   None  Finland  FIN
4  15  Arvo Ossian Aaltonen   M  34.0   None   None  Finland  FIN

          Games  Year  Season       City     Sport  \
0  1912 Summer  1912  Summer  Stockholm  Swimming
1  1912 Summer  1912  Summer  Stockholm  Swimming
2  1920 Summer  1920  Summer  Antwerpen  Swimming
3  1920 Summer  1920  Summer  Antwerpen  Swimming
4  1924 Summer  1924  Summer      Paris  Swimming

                                    Event   Medal
0  Swimming Men's 200 metres Breaststroke    None
1  Swimming Men's 400 metres Breaststroke    None
2  Swimming Men's 200 metres Breaststroke  Bronze
3  Swimming Men's 400 metres Breaststroke  Bronze
4  Swimming Men's 200 metres Breaststroke    None
```

[58]:
```
pysqldf("SELECT * From athlete_events Where Medal = 'Bronze';")
```

[58]:
```
           ID                              Name Sex   Age  Height  Weight  \
0          15  Arvo Ossian Aaltonen   M  30.0     NaN     NaN
```

```
1          15                  Arvo Ossian Aaltonen   M  30.0     NaN     NaN
2          16               Juhamatti Tapio Aaltonen   M  28.0   184.0    85.0
3          17                Paavo Johannes Aaltonen   M  28.0   175.0    64.0
4          17                Paavo Johannes Aaltonen   M  32.0   175.0    64.0
…         …                                       …  ..   …     …      …
13290  135535             Claudia Antoinette Zwiers   F  22.0   181.0    78.0
13291  135545                        Henk Jan Zwolle   M  27.0   197.0    93.0
13292  135553  Galina Ivanovna Zybina (-Fyodorova)   F  33.0   168.0    80.0
13293  135554                           Bogusaw Zych   M  28.0   182.0    82.0
13294  135563             Olesya Nikolayevna Zykina   F  19.0   171.0    64.0

                Team  NOC        Games  Year  Season       City       Sport  \
0            Finland  FIN  1920 Summer  1920  Summer  Antwerpen    Swimming
1            Finland  FIN  1920 Summer  1920  Summer  Antwerpen    Swimming
2            Finland  FIN  2014 Winter  2014  Winter      Sochi  Ice Hockey
3            Finland  FIN  1948 Summer  1948  Summer     London  Gymnastics
4            Finland  FIN  1952 Summer  1952  Summer   Helsinki  Gymnastics
…              …  …          …   …      …         …          …
13290   Netherlands  NED  1996 Summer  1996  Summer    Atlanta        Judo
13291   Netherlands  NED  1992 Summer  1992  Summer  Barcelona      Rowing
13292  Soviet Union  URS  1964 Summer  1964  Summer      Tokyo   Athletics
13293        Poland  POL  1980 Summer  1980  Summer     Moskva     Fencing
13294        Russia  RUS  2000 Summer  2000  Summer     Sydney   Athletics

                                        Event   Medal
0      Swimming Men's 200 metres Breaststroke  Bronze
1      Swimming Men's 400 metres Breaststroke  Bronze
2                 Ice Hockey Men's Ice Hockey  Bronze
3      Gymnastics Men's Individual All-Around  Bronze
4            Gymnastics Men's Team All-Around  Bronze
…                                        …     …
13290              Judo Women's Middleweight  Bronze
13291             Rowing Men's Double Sculls  Bronze
13292             Athletics Women's Shot Put  Bronze
13293               Fencing Men's Foil, Team  Bronze
13294  Athletics Women's 4 x 400 metres Relay  Bronze

[13295 rows x 15 columns]
```

[8]:
```
Age = pysqldf('''SELECT Avg(age) as Average_age, Sex as Gender, Sport, NOC as␣
↪Country
                  FROM athlete_events
                  Group by Sex, Sport, Country;''' )
```

[65]:
```
pysqldf("SELECT * From Age;")
```

[65]:
```
      Average_age Gender          Sport Country
0       20.000000      F  Alpine Skiing     ALB
1       20.000000      F  Alpine Skiing     ALG
2       19.814815      F  Alpine Skiing     AND
3       20.666667      F  Alpine Skiing     ARG
4       23.000000      F  Alpine Skiing     ARM
…             …    …             …     …
6349    21.333333      M      Wrestling     VIE
6350    21.500000      M      Wrestling     YAR
6351    27.000000      M      Wrestling     YEM
6352    25.684211      M      Wrestling     YUG
6353    21.000000      M      Wrestling     ZAM

[6354 rows x 4 columns]
```

[70]:
```
pysqldf('''SELECT Average_age, Gender From Age
        Group by Gender;''')
```

[70]:
```
   Average_age Gender
0         20.0      F
1         26.0      M
```

[9]:
```
sport_average_age = pysqldf('''SELECT Average_age, Sport From Age
        Group by Sport;''')
```

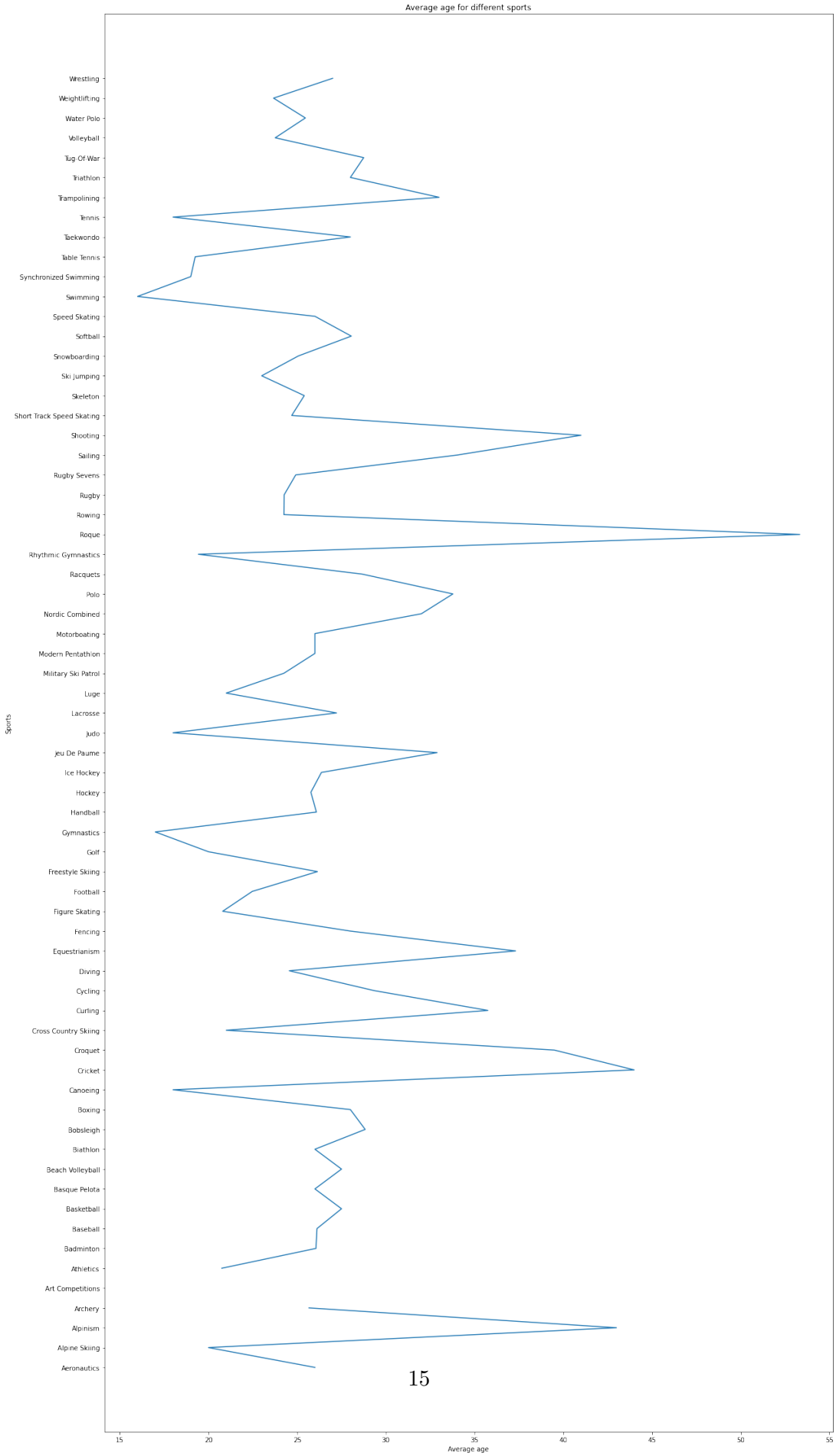
[84]:
```
pysqldf('''SELECT * From sport_average_age;''')
```

[84]:
```
    Average_age             Sport
0     26.000000       Aeronautics
1     20.000000     Alpine Skiing
2     43.000000          Alpinism
3     25.666667           Archery
4           NaN  Art Competitions
..          …               …
61    28.750000        Tug-Of-War
62    23.761905        Volleyball
63    25.461538        Water Polo
64    23.666667     Weightlifting
65    27.000000         Wrestling

[66 rows x 2 columns]
```

[10]:
```
from matplotlib import pyplot as plt
```

[100]:
```
plt.figure(figsize = (20,40))
plt.plot(sport_average_age.Average_age, sport_average_age.Sport)
plt.title("Average age for different sports")
```

```
plt.xlabel("Average age")
plt.ylabel("Sports")


plt.show()
```


Average age for different sports
Sports
Wrestling
Weightlifting
Water Polo
Volleyball
Tug-Of-War
Triathlon
Trampolining
Tennis
Taekwondo
Table Tennis
Synchronized Swimming
Swimming
Speed Skating
Softball
Snowboarding
Ski jumping
Skeleton
Short Track Speed Skating
Shooting
Sailing
Rugby Sevens
Rugby
Rowing
Roque
Rhythmic Gymnastics
Racquets
Polo
Nordic Combined
Motorboating
Modern Pentathlon
Military Ski Patrol
Luge
Lacrosse
Judo
jeu De Paume
Ice Hockey
Hockey
Handball
Gymnastics
Golf
Freestyle Skiing
Football
Figure Skating
Fencing
Equestrianism
Diving
Cycling
Curling
Cross Country Skiing
Croquet
Cricket
Canoeing
Boxing
Bobsleigh
Biathlon
Beach Volleyball
Basque Pelota
Basketball
Baseball
Badminton
Athletics
Art Competitions
Archery
Alpinism
Alpine Skiing
Aeronautics
15
20
25
30
35
40
45
50
55
Average age

[81]:
```
#Representation in Olympics
#Age distribution

pysqldf('''SELECT Min(Age), Max(Age), Avg(Age) From athlete_events
            Where Sex = 'M';''')
```

[81]:

	Min(Age)	Max(Age)	Avg(Age)
0	10.0	97.0	26.277562

[51]:
```
agevsgender = pysqldf('''SELECT Min(Age), Max(Age), Avg(Age), Sex From␣
↪athlete_events
            group by Sex;''')
```

[52]:
```
agevsgender.to_csv('agevsgender.csv')
```

[82]:
```
pysqldf('''SELECT Min(Age), Max(Age), Avg(Age), Sex From athlete_events
            group by Sex;''')
```

[82]:

	Min(Age)	Max(Age)	Avg(Age)	Sex
0	11.0	74.0	23.732881	F
1	10.0	97.0	26.277562	M

[11]:
```
Sportvsage = pysqldf('''SELECT Min(Age), Max(Age), Avg(Age), Sport From␣
↪athlete_events
            group by Sport;''')
```

[103]:
```
Sportvsage.to_csv('Sportvsage.csv')
```

[104]:
```
pysqldf('''SELECT * From Sportvsage;''')
```

[104]:

	Min(Age)	Max(Age)	Avg(Age)	Sport
0	26.0	26.0	26.000000	Aeronautics
1	14.0	55.0	23.205462	Alpine Skiing
2	22.0	57.0	38.812500	Alpinism
3	14.0	71.0	27.935226	Archery
4	14.0	97.0	45.901009	Art Competitions
..	…	…	…	…
61	17.0	45.0	29.309524	Tug-Of-War
62	15.0	41.0	25.183800	Volleyball
63	14.0	45.0	25.659627	Water Polo
64	15.0	45.0	25.502010	Weightlifting
65	15.0	50.0	25.798289	Wrestling

[66 rows x 4 columns]

[ ]:

[118]:
```
pysqldf('''SELECT Sport, Sex, Count(Sex) as Paricipation
            From athlete_events
            group by Sport, Sex;''')
```

[118]:

	Sport	Sex	Paricipation
0	Aeronautics	M	1
1	Alpine Skiing	F	3398
2	Alpine Skiing	M	5431
3	Alpinism	F	1
4	Alpinism	M	24
..	…	..	…
111	Water Polo	M	3358
112	Weightlifting	F	463
113	Weightlifting	M	3474
114	Wrestling	F	304
115	Wrestling	M	6850

[116 rows x 3 columns]

[18]:
```
sportvsgenderF = pysqldf('''SELECT Sport, Count(*) as female_participants From␣
↪athlete_events
           Where Sex = 'F'
           Group by Sport;''')
```

[13]:
```
sportvsgenderM = pysqldf('''SELECT Sport, Count(*) as male_participants From␣
↪athlete_events
           Where Sex = 'M'
           Group by Sport;''')
```

[27]:
```
sportvsgender = pysqldf('''SELECT M.Sport, M.male_participants, F.
↪female_participants, (M.male_participants/F.female_participants) AS Ratio
            From sportvsgenderM As M Join sportvsgenderF As F
            On M.Sport = F.Sport;''')
```

[31]:
```
pysqldf('''SELECT * From sportvsgender;''')
```

[31]:

	Sport	male_participants	female_participants	Ratio
0	Alpine Skiing	5431	3398	1
1	Alpinism	24	1	24
2	Archery	1319	1015	1
3	Art Competitions	3201	377	8
4	Athletics	26958	11666	2
5	Badminton	717	740	0
6	Basketball	3280	1256	2
7	Beach Volleyball	288	276	1

8	Biathlon	3030	1863	1
9	Bobsleigh	2915	143	20
10	Boxing	5975	72	82
11	Canoeing	4791	1380	3
12	Croquet	13	6	2
13	Cross Country Skiing	5748	3385	1
14	Curling	241	222	1
15	Cycling	9465	1394	6
16	Diving	1632	1210	1
17	Equestrianism	5098	1246	4
18	Fencing	8735	2000	4
19	Figure Skating	1126	1172	0
20	Football	5733	1012	5
21	Freestyle Skiing	504	433	1
22	Golf	177	70	2
23	Gymnastics	17578	9129	1
24	Handball	2264	1401	1
25	Hockey	3958	1459	2
26	Ice Hockey	4762	754	6
27	Judo	2708	1093	2
28	Luge	1102	377	2
29	Modern Pentathlon	1513	164	9
30	Motorboating	16	1	16
31	Rowing	8402	2193	3
32	Rugby Sevens	151	148	1
33	Sailing	5660	926	6
34	Shooting	9724	1724	5
35	Short Track Speed Skating	773	761	1
36	Skeleton	133	66	2
37	Ski Jumping	2371	30	79
38	Snowboarding	520	416	1
39	Speed Skating	3532	2081	1
40	Swimming	13345	9850	1
41	Table Tennis	1002	953	1
42	Taekwondo	307	299	1
43	Tennis	1684	1178	1
44	Trampolining	76	76	1
45	Triathlon	266	263	1
46	Volleyball	1861	1543	1
47	Water Polo	3358	488	6
48	Weightlifting	3474	463	7
49	Wrestling	6850	304	22

[50]:
```
sportvsgender['Ratio'] = pd.to_numeric(sportvsgender['Ratio'],␣
↪downcast='float', errors='coerce')
```

[47]:
```
sportvsgender.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 50 entries, 0 to 49
Data columns (total 4 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   Sport                50 non-null     object
 1   male_participants    50 non-null     int64
 2   female_participants  50 non-null     int64
 3   Ratio                50 non-null     float32
dtypes: float32(1), int64(2), object(1)
memory usage: 1.5+ KB
```

[48]:
```
pysqldf('''SELECT * From sportvsgender;''')
```

[48]:

	Sport	male_participants	female_participants	Ratio
0	Alpine Skiing	5431	3398	1.0
1	Alpinism	24	1	24.0
2	Archery	1319	1015	1.0
3	Art Competitions	3201	377	8.0
4	Athletics	26958	11666	2.0
5	Badminton	717	740	0.0
6	Basketball	3280	1256	2.0
7	Beach Volleyball	288	276	1.0
8	Biathlon	3030	1863	1.0
9	Bobsleigh	2915	143	20.0
10	Boxing	5975	72	82.0
11	Canoeing	4791	1380	3.0
12	Croquet	13	6	2.0
13	Cross Country Skiing	5748	3385	1.0
14	Curling	241	222	1.0
15	Cycling	9465	1394	6.0
16	Diving	1632	1210	1.0
17	Equestrianism	5098	1246	4.0
18	Fencing	8735	2000	4.0
19	Figure Skating	1126	1172	0.0
20	Football	5733	1012	5.0
21	Freestyle Skiing	504	433	1.0
22	Golf	177	70	2.0
23	Gymnastics	17578	9129	1.0
24	Handball	2264	1401	1.0
25	Hockey	3958	1459	2.0
26	Ice Hockey	4762	754	6.0
27	Judo	2708	1093	2.0
28	Luge	1102	377	2.0
29	Modern Pentathlon	1513	164	9.0
30	Motorboating	16	1	16.0
31	Rowing	8402	2193	3.0

```
32                Rugby Sevens                151                   148    1.0
33                     Sailing               5660                   926    6.0
34                    Shooting               9724                  1724    5.0
35  Short Track Speed Skating                773                   761    1.0
36                    Skeleton                133                    66    2.0
37                 Ski Jumping               2371                    30   79.0
38                Snowboarding                520                   416    1.0
39               Speed Skating               3532                  2081    1.0
40                    Swimming              13345                  9850    1.0
41                Table Tennis               1002                   953    1.0
42                   Taekwondo                307                   299    1.0
43                      Tennis               1684                  1178    1.0
44                Trampolining                 76                    76    1.0
45                   Triathlon                266                   263    1.0
46                  Volleyball               1861                  1543    1.0
47                  Water Polo               3358                   488    6.0
48                Weightlifting               3474                   463    7.0
49                    Wrestling               6850                   304   22.0
```

[91]:
```
pysqldf('''SELECT * From sportvsgender Order by Ratio Desc Limit 10;''')
```

[91]:
```
                Sport  male_participants  female_participants  Ratio
0              Boxing               5975                   72   82.0
1         Ski Jumping               2371                   30   79.0
2            Alpinism                 24                    1   24.0
3           Wrestling               6850                  304   22.0
4           Bobsleigh               2915                  143   20.0
5        Motorboating                 16                    1   16.0
6  Modern Pentathlon               1513                  164    9.0
7   Art Competitions               3201                  377    8.0
8      Weightlifting               3474                  463    7.0
9             Cycling               9465                 1394    6.0
```

[53]:
```
sportvsgender.to_csv('sportvsgender.csv')
```

[54]:
```
pysqldf('''SELECT Medal, Sex, Count(Medal) as Winnings
            From athlete_events
            group by Medal, Sex;''')
```

[54]:
```
     Medal Sex  Winnings
0     None   F         0
1     None   M         0
2   Bronze   F      3771
3   Bronze   M      9524
4     Gold   F      3747
5     Gold   M      9625
6   Silver   F      3735
```

```
7  Silver   M      9381
```

[67]:
```
pysqldf('''SELECT Medal, Sex, Count(Medal) as Winnings_male
            From athlete_events
            Where Sex = 'M'
            group by Medal, Sex;''')
```

[67]:
```
     Medal Sex  Winnings_male
0     None   M              0
1   Bronze   M           9524
2     Gold   M           9625
3   Silver   M           9381
```

[68]:
```
winningsvsgenderM = pysqldf('''SELECT Medal, Sex, Count(Medal) as Winnings_male
            From athlete_events
            Where Sex = 'M'
            group by Medal, Sex;''')
```

[61]:
```
pysqldf('''SELECT Medal, Sex, Count(Medal) as Winnings_female
            From athlete_events
            Where Sex = 'F'
            group by Medal, Sex;''')
```

[61]:
```
     Medal Sex  Winnings_male
0     None   F              0
1   Bronze   F           3771
2     Gold   F           3747
3   Silver   F           3735
```

[70]:
```
winningsvsgenderF = pysqldf('''SELECT Medal, Sex, Count(Medal) as␣
↪Winnings_female
            From athlete_events
            Where Sex = 'F'
            group by Medal, Sex;''')
```

[71]:
```
pysqldf('''SELECT M.Medal, M.Winnings_male, F.Winnings_female, (M.Winnings_male/
↪F.Winnings_female)*1.0 AS "Winning Ratio"
            From winningsvsgenderM As M Join winningsvsgenderF As F
            On M.Medal = F.Medal;''')
```

[71]:
```
     Medal  Winnings_male  Winnings_female  Winning Ratio
0   Bronze           9524             3771            2.0
1     Gold           9625             3747            2.0
2   Silver           9381             3735            2.0
```

[72]:
```
winningsvsgender = pysqldf('''SELECT M.Medal, M.Winnings_male, F.
↪Winnings_female, (M.Winnings_male/F.Winnings_female)*1.0 AS "Winning Ratio"
```

```
            From winningsvsgenderM As M Join winningsvsgenderF As F
            On M.Medal = F.Medal;''')
```

[73]:
```
winningsvsgender.to_csv('winningsvsgender.csv')
```

[83]:
```
pysqldf('''SELECT NOC as Country, Medal, Count(Medal) as Winnings
            From athlete_events
            group by NOC, Medal
            Order by Winnings Desc, Medal Limit 15;''')
```

[83]:

	Country	Medal	Winnings
0	USA	Gold	2638
1	USA	Silver	1641
2	USA	Bronze	1358
3	URS	Gold	1082
4	GER	Bronze	746
5	GER	Gold	745
6	GBR	Silver	739
7	URS	Silver	732
8	URS	Bronze	689
9	GBR	Gold	678
10	GER	Silver	674
11	FRA	Bronze	666
12	GBR	Bronze	651
13	FRA	Silver	610
14	ITA	Gold	575

[84]:
```
Medalsbycountry = pysqldf('''SELECT NOC as Country, Medal, Count(Medal) as␣
↪Winnings
            From athlete_events
            group by NOC, Medal
            Order by Winnings Desc, Medal Limit 15;''')
```

[86]:
```
Medalsbycountry.to_csv('Medalsbycountry.csv')
```

[90]:
```
pysqldf('''SELECT Country, Sum(Winnings) as Total_Winnings
            From Medalsbycountry
            group by Country
            Order by Winnings Desc Limit 15;''')
```

[90]:

	Country	Total_Winnings
0	USA	5637
1	URS	2503
2	GER	2165
3	GBR	2068
4	FRA	1276
5	ITA	575