SQL CAPSTONE PROJECT
WEEK 2

Dataset:
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.

Step 1: Descriptive Statistics On The Dataset

1) Firstly, I found out the Average Height, Weight and Age of Medal Winners in each sport. This was in order to have better understanding of the data and information of the ideal body type for each sport.

SQL Query:

```
pysqldf("""
Select avgHeight, avgWeight, avgAge, (avgWeight/(avgHeight*avgHeight))*10000 AS BMI  ,Sport
FROM
    (Select AVG(Height) as avgHeight, AVG(Weight) as avgWeight, AVG(Age) as avgAge,Sport
    FROM athlete_events
    WHERE Medal IS NOT NULL
    GROUP BY Sport
    )
ORDER BY BMI
""")
```

✓ 2.2s

Output:

	avgHeight	avgWeight	avgAge	BMI	Sport
0	170.604839	49.306452	18.911290	16.940299	Rhythmic Gymnastics
1	168.864486	56.383178	23.817757	19.773015	Synchronized Swimming
2	177.262857	62.920000	23.725714	20.024110	Ski Jumping
3	174.933333	62.466667	28.200000	20.412829	Triathlon
4	178.343066	67.189781	24.905109	21.124711	Nordic Combined
5	161.254091	55.069337	21.810508	21.178167	Gymnastics
6	168.473896	60.236948	24.140562	21.222558	Figure Skating
7	165.874214	58.694969	22.468553	21.332567	Diving
8	166.310345	59.137931	23.965517	21.380980	Trampolining
9	178.555556	68.722222	23.451389	21.555121	Taekwondo
10	170.300613	63.380368	24.791411	21.853567	Table Tennis
11	173.311499	65.823144	27.685590	21.914111	Cross Country Skiing
12	174.000000	66.500000	28.000000	21.964592	Lacrosse
13	174.528351	66.929124	27.012887	21.972687	Biathlon
14	180.427746	72.057803	25.924855	22.134737	Tennis
15	175.446456	68.283560	35.315234	22.183341	Equestrianism
16	174.309322	67.445621	23.323446	22.197933	Boxing
17	179.522876	71.594771	27.156863	22.214764	Modern Pentathlon
18	174.590909	67.837662	25.493506	22.255001	Badminton
19	169.686131	64.094891	22.514599	22.260285	Short Track Speed Skating
20	181.111022	73.258045	21.040225	22.333951	Swimming
21	177.627467	71.513158	25.099507	22.665521	Athletics
22	178.025248	72.140667	27.153291	22.762343	Fencing
23	186.882046	79.600209	25.768267	22.791822	Volleyball
24	186.823529	79.632353	29.617647	22.815312	Beach Volleyball
25	173.657534	69.187500	25.749144	22.942481	Hockey
26	175.476015	70.736162	24.221402	22.972379	Football
27	175.708333	71.083333	28.958333	23.024121	Skeleton
28	173.561798	69.573034	24.719101	23.095781	Snowboarding
29	170.725490	67.362745	24.931373	23.111229	Freestyle Skiing
30	177.861538	73.169231	25.244970	23.129402	Cycling
31	175.209016	71.245902	25.325820	23.208495	Speed Skating
32	192.153000	86.692000	25.297000	23.479275	Basketball
33	173.349515	70.832524	24.791262	23.571513	Archery
34	184.631179	80.863593	25.879753	23.721544	Rowing
35	171.017442	69.843023	26.819767	23.880434	Softball
36	179.392910	77.509601	30.019202	24.084906	Sailing
37	173.658192	72.782486	24.508475	24.134390	Alpine Skiing
38	174.267176	73.389313	32.137405	24.165826	Curling

39	183.239061	81.157951	26.633938	24.171019	Handball
40	179.519693	78.027858	25.866475	24.211712	Canoeing
41	173.701107	74.885609	30.201107	24.819509	Shooting
42	185.564136	85.554974	25.968586	24.846040	Water Polo
43	179.750000	80.500000	33.000000	24.914839	Golf
44	178.252809	79.207865	25.882022	24.928465	Luge
45	174.000000	75.500000	32.000000	24.937244	Art Competitions
46	178.788624	80.880092	25.820907	25.302409	Ice Hockey
47	180.375000	83.000000	24.875000	25.510878	Tug-Of-War
48	172.745605	76.657704	25.924509	25.688680	Wrestling
49	173.666667	77.533333	24.100000	25.707244	Rugby
50	182.429429	85.777778	26.255255	25.774188	Baseball
51	176.594595	81.256757	25.337838	26.055816	Rugby Sevens
52	173.849810	80.109316	25.245247	26.505413	Judo
53	182.208481	90.077739	29.332155	27.131906	Bobsleigh
54	167.426692	81.218985	25.028195	28.973981	Weightlifting

This was the result of that query.

This helped me answer the following questions:

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

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

2) Secondly , I found out the total medals won by each country

```
pysqldf("""

  Select
  count(Medal) AS number_of_medals, region
  FROM athlete_events INNER JOIN noc_regions
  ON athlete_events.NOC  = noc_regions.NOC
  WHERE Medal IS NOT NULL
  GROUP BY region
  ORDER BY count(Medal) DESC
  """)
```

	number_of_medals	region
0	4383	USA
1	3610	Russia
2	3189	Germany
3	1210	Australia
4	1060	Italy
5	1060	Canada
6	1031	UK
7	989	China
8	987	France
9	843	Japan
10	791	Hungary
11	765	Sweden
12	724	Finland
13	708	Netherlands
14	597	Romania
15	561	South Korea
16	548	Poland
17	519	Czech Republic
18	514	Norway

This query helped me understand which countries have won the most medals in the games.

3) Thirdly, I found out which country has won the most medals in each game and how many medals.

```
pysqldf("""
Select MAX(number_of_medals) AS medals_won, Sport, region
FROM
    (Select
    count(Medal) AS number_of_medals, region, Sport
    FROM athlete_events INNER JOIN noc_regions
    ON athlete_events.NOC  = noc_regions.NOC
    WHERE Medal IS NOT NULL
    GROUP BY region, Sport
    ORDER BY count(Medal) DESC)
GROUP BY Sport
ORDER BY MAX(number_of_medals) DESC
""")
```

	medals_won	Sport	region
0	1009	Athletics	USA
1	918	Swimming	USA
2	395	Rowing	Germany
3	341	Basketball	USA
4	334	Gymnastics	Russia
5	281	Ice Hockey	Canada
6	218	Volleyball	Russia
7	211	Fencing	Russia
8	211	Canoeing	Germany
9	202	Hockey	Germany
10	173	Football	Germany
11	172	Wrestling	Russia
12	166	Equestrianism	Germany
13	157	Cross Country Skiing	Russia
14	155	Handball	Russia
15	134	Cycling	Germany
16	133	Water Polo	Russia
17	111	Baseball	Cuba

This helped me know which Country has won the most medals in each game

Step 2: Key Points

Key Points Discovered from The Data:

1. The BMI value of all athletes are ideal
2. Developed countries have won more medals in each sport
3. Developed countries have consistently won more total medals

Step 3: Hypothesis Testing

Hypothesis 1:

The average BMI value of the medal willing athletes matches with the ideal values

Answer – This hypothesis turned out to be mostly true. For almost all the sports the average BMI value was between 18.5 to 24.9

However for Gymnastics, the BMI was lower than the normal value.

For sports like Wrestling, Tug of War, Weightlifting was more than the normal value.

Hypothesis 2:

Developed countries have won more medals due to high infrastructure of sports.

Answer – This hypothesis also turned out to be true. The developed countries such as Russia, Germany, USA, Australia, Canada, UK have consistently scored more medals.

The reason for this might be the fact that these countries provide better infrastructure for the sports and thus produce more and better athletes.

Step 4: Additional Questions

How has the percentages of the genders playing each sport changed over the years?

Have more women started to play sports at the national level as the times progressed?