Descriptive Statistics and Understanding the Data

In [188…
```
import pandas as pd
import matplotlib.pyplot as plt
from pandasql import sqldf
pysqldf = lambda q: sqldf(q, globals())

%store -r athlete_events
%store -r summer_games
%store -r winter_games
%store -r female
%store -r male
```

In [189…
```
athlete_events.head()
```

Out[189]:

	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

In [190…
```
summer_games.head()
```

Out[190]:

	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	None
1	2	A Lamusi	M	23.0	170.0	60.0	China	CHN	2012 Summer	2012	Summer	London	Judo	Judo Men's Extra-Lightweight	None
2	3	Gunnar Nielsen Aaby	M	24.0	NaN	NaN	Denmark	DEN	1920 Summer	1920	Summer	Antwerpen	Football	Football Men's Football	None
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	8	Cornelia "Cor" Aalten (-Strannood)	F	18.0	168.0	NaN	Netherlands	NED	1932 Summer	1932	Summer	Los Angeles	Athletics	Athletics Women's 100 metres	None

In [191…
```
winter_games.head()
```

Out[191]:

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event	Medal
0	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	None
1	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	None
2	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	None
3	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	None
4	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	None

In [192…
```
#Calculating the ratios for Men vs Women
# Total Ratio
total_ratio = pysqldf('''SELECT Sex,
                                COUNT(*) Total_Count,
                                COUNT(*) * 100.00 / SUM(COUNT(*)) OVER () AS ratio
                                FROM athlete_events
                                GROUP BY sex''')

#Summer games ratio
summer_ratio = pysqldf('''SELECT Sex,
                                COUNT(*) Total_Count,
                                COUNT(*) * 100.00 / SUM(COUNT(*)) OVER () AS ratio
                                FROM summer_games
                                GROUP BY sex''')

#Winter games ratio
winter_ratio = pysqldf('''SELECT Sex,
                                COUNT(*) Total_Count,
                                COUNT(*) * 100.00 / SUM(COUNT(*)) OVER () AS ratio
                                FROM winter_games
                                GROUP BY sex''')
```

In [193…
```
total_ratio.head()
```

Out[193]:

	Sex	Total_Count	ratio
0	F	74522	27.487127
1	M	196594	72.512873

In [194…
```
summer_ratio.head()
```

Out[194]:

	Sex	Total_Count	ratio
0	F	59443	26.709713
1	M	163109	73.290287

In [195…
```
winter_ratio.head()
```

Out[195]:

	Sex	Total_Count	ratio
0	F	15079	31.049749
1	M	33485	68.950251

In [196…
```
#Calculating the total average for Men and Women
total_averages = pysqldf('''SELECT Sex,
                                   AVG(Weight) Average_Weight,
                                   AVG(Height) Average_Height,
                                   AVG(Age) Average_Age
                            FROM athlete_events
                            GROUP BY Sex''')

#Calculating the Summer average for Men and Women
summer_averages = pysqldf('''SELECT Sex,
                                   AVG(Weight) Average_Weight,
                                   AVG(Height) Average_Height,
                                   AVG(Age) Average_Age
                            FROM summer_games
                            GROUP BY Sex''')

#Calculating the Winter average for Men and Women
winter_averages = pysqldf('''SELECT Sex,
                                   AVG(Weight) Average_Weight,
                                   AVG(Height) Average_Height,
                                   AVG(Age) Average_Age
                            FROM winter_games
                            GROUP BY Sex''')
```

In [197…
```
total_averages.head()
```

Out[197]:

	Sex	Average_Weight	Average_Height	Average_Age
0	F	60.021252	167.839740	23.732881
1	M	75.743677	178.858463	26.277562

In [198…
```
summer_averages.head()
```

Out[198]:

	Sex	Average_Weight	Average_Height	Average_Age
0	F	60.087644	168.169025	23.660997
1	M	75.604195	178.901874	26.443944

In [199…
```
winter_averages.head()
```

Out[199]:

	Sex	Average_Weight	Average_Height	Average_Age
0	F	59.755156	166.528250	24.014398
1	M	76.357058	178.668699	25.504261

Next we're are going to dive into the medals awarded for the olympians and I will divide it into total medals, summer medals and winter medals. I am interesting in seeing how the medals are divided out, is there a significant area where males are being awarded the most medals? and vice versa is there a specific area where females are dominating medals.

In [200…
```
#Total view of each year showing Participants and the number of medals awarded
pysqldf('''SELECT Year,
                  COUNT(*) Participants,
                  SUM(CASE WHEN Medal IS NOT NULL THEN 1 ELSE 0 END) AS Total_Medals,
                  SUM(CASE WHEN Medal = 'Gold' THEN 1 ELSE 0 END) AS Gold_Medals,
                  SUM(CASE WHEN Medal = 'Silver' THEN 1 ELSE 0 END) AS Silver_Medals,
                  SUM(CASE WHEN Medal = 'Bronze' THEN 1 ELSE 0 END) AS Bronze_Medals
           FROM athlete_events
           GROUP BY Year''')
```

Out[200]:

	Year	Participants	Total_Medals	Gold_Medals	Silver_Medals	Bronze_Medals
0	1896	380	143	62	43	38
1	1900	1936	604	201	228	175
2	1904	1301	486	173	163	150
3	1906	1733	458	157	156	145
4	1908	3101	831	294	281	256
5	1912	4040	941	326	315	300
6	1920	4292	1308	493	448	367
7	1924	5693	962	332	319	311
8	1928	5574	823	275	267	281
9	1932	3321	739	261	246	232
10	1936	7401	1025	348	347	330
11	1948	7480	987	330	332	325
12	1952	9358	1033	351	335	347
13	1956	6434	1043	353	342	348
14	1960	9235	1058	359	342	357
15	1964	9480	1215	408	406	401
16	1968	10479	1256	425	410	421
17	1972	11959	1414	474	455	485
18	1976	10502	1531	508	505	518
19	1980	8937	1602	529	531	542
20	1984	11588	1698	571	551	576
21	1988	14676	1845	607	601	637
22	1992	16413	2030	663	657	710
23	1994	3160	331	110	109	112
24	1996	13780	1842	608	605	629
25	1998	3605	440	145	145	150
26	2000	13821	2004	663	661	680
27	2002	4109	478	162	157	159
28	2004	13443	2001	664	660	677
29	2006	4382	526	176	175	175
30	2008	13602	2048	671	667	710
31	2010	4402	520	174	175	171
32	2012	12920	1941	632	630	679
33	2014	4891	597	202	197	198
34	2016	13688	2023	665	655	703

Check the amount of medals won as a % this time

In [201…
```
#Shows the % of medals won compared to the total participants e.g
# 2016 - 14.77% of participants received a medal
# 2016 - 4.85% of participants received a Gold medal
# 2016 - 4.78% of participants received a Silver medal
# 2016 - 5.13% of participants received a Bronze medal
total_ratio_medals = pysqldf('''SELECT Year,
                                       CAST(Total_Medals AS FLOAT) * 100 / Participants AS Total_Medals,
                                       CAST(Gold_Medals AS FLOAT) * 100 / Participants AS Gold_Medals,
                                       CAST(Silver_Medals AS FLOAT) * 100 / Participants AS Silver_Medals,
                                       CAST(Bronze_Medals AS FLOAT) * 100 / Participants AS Bronze_Medals
                               FROM
                                  (
                                  SELECT Year,
                                         COUNT(*) Participants,
                                         SUM(CASE WHEN Medal IS NOT NULL THEN 1 ELSE 0 END) AS Total_Medals,
                                         SUM(CASE WHEN Medal = 'Gold' THEN 1 ELSE 0 END) AS Gold_Medals,
                                         SUM(CASE WHEN Medal = 'Silver' THEN 1 ELSE 0 END) AS Silver_Medals,
                                         SUM(CASE WHEN Medal = 'Bronze' THEN 1 ELSE 0 END) AS Bronze_Medals
                                  FROM athlete_events
                                  GROUP BY Year) Medals_Table
                                  ''')
```

In [202…
```
total_ratio_medals.tail()
```

Out[202]:

	Year	Total_Medals	Gold_Medals	Silver_Medals	Bronze_Medals
30	2008	15.056609	4.933098	4.903691	5.219821
31	2010	11.812812	3.952749	3.975466	3.884598
32	2012	15.023220	4.891641	4.876161	5.255418
33	2014	12.206093	4.130035	4.027806	4.048252
34	2016	14.779369	4.858270	4.785213	5.135885

Calculating the % of medals won for the Summer Olympic Games

In [203…
```
#Calculating the % of medals won for the Summer Olympic Games
summer_ratio_medals = pysqldf('''SELECT Year,
                                       CAST(Total_Medals AS FLOAT) * 100 / Participants AS Total_Medals,
                                       CAST(Gold_Medals AS FLOAT) * 100 / Participants AS Gold_Medals,
                                       CAST(Silver_Medals AS FLOAT) * 100 / Participants AS Silver_Medals,
                                       CAST(Bronze_Medals AS FLOAT) * 100 / Participants AS Bronze_Medals
                               FROM
                                  (
                                  SELECT Year,
                                         COUNT(*) Participants,
                                         SUM(CASE WHEN Medal IS NOT NULL THEN 1 ELSE 0 END) AS Total_Medals,
                                         SUM(CASE WHEN Medal = 'Gold' THEN 1 ELSE 0 END) AS Gold_Medals,
                                         SUM(CASE WHEN Medal = 'Silver' THEN 1 ELSE 0 END) AS Silver_Medals,
                                         SUM(CASE WHEN Medal = 'Bronze' THEN 1 ELSE 0 END) AS Bronze_Medals
                                  FROM summer_games
                                  GROUP BY Year) Medals_Table
                                  ''')
```

In [204…
```
summer_ratio_medals.tail()
```

Out[204]:

	Year	Total_Medals	Gold_Medals	Silver_Medals	Bronze_Medals
24	2000	14.499674	4.797048	4.782577	4.920049
25	2004	14.885070	4.939374	4.909618	5.036078
26	2008	15.056609	4.933098	4.903691	5.219821
27	2012	15.023220	4.891641	4.876161	5.255418
28	2016	14.779369	4.858270	4.785213	5.135885

Calculating the % of medals won for the Winter Olympic Games

In [205…
```
#Calculating the % of medals won for the Winter Olympic Games
winter_ratio_medals = pysqldf('''SELECT Year,
                                       CAST(Total_Medals AS FLOAT) * 100 / Participants AS Total_Medals,
                                       CAST(Gold_Medals AS FLOAT) * 100 / Participants AS Gold_Medals,
                                       CAST(Silver_Medals AS FLOAT) * 100 / Participants AS Silver_Medals,
                                       CAST(Bronze_Medals AS FLOAT) * 100 / Participants AS Bronze_Medals
                               FROM
                                  (
                                  SELECT Year,
                                         COUNT(*) Participants,
                                         SUM(CASE WHEN Medal IS NOT NULL THEN 1 ELSE 0 END) AS Total_Medals,
                                         SUM(CASE WHEN Medal = 'Gold' THEN 1 ELSE 0 END) AS Gold_Medals,
                                         SUM(CASE WHEN Medal = 'Silver' THEN 1 ELSE 0 END) AS Silver_Medals,
                                         SUM(CASE WHEN Medal = 'Bronze' THEN 1 ELSE 0 END) AS Bronze_Medals
                                  FROM winter_games
                                  GROUP BY Year) Medals_Table
                                  ''')
```

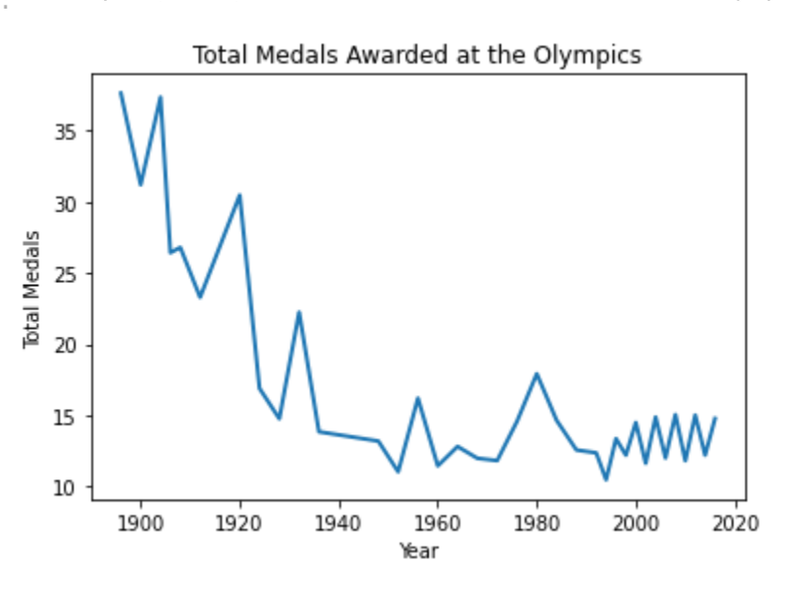
In [206…
```
winter_ratio_medals.tail()
```

Out[206]:

	Year	Total_Medals	Gold_Medals	Silver_Medals	Bronze_Medals
17	1998	12.205270	4.022191	4.022191	4.160888
18	2002	11.633001	3.942565	3.820881	3.869555
19	2006	12.003651	4.016431	3.993610	3.993610
20	2010	11.812812	3.952749	3.975466	3.884598
21	2014	12.206093	4.130035	4.027806	4.048252

In [207…
```
#Visualizing the total amount of medals awarded throughout the years
plt.plot(total_ratio_medals.Year, total_ratio_medals.Total_Medals, marker='', linewidth = 2, label='Total Medals')
plt.xlabel("Year")
plt.ylabel("Total Medals")
plt.title("Total Medals Awarded at the Olympics")
```

Out[207]: Text(0.5, 1.0, 'Total Medals Awarded at the Olympics')



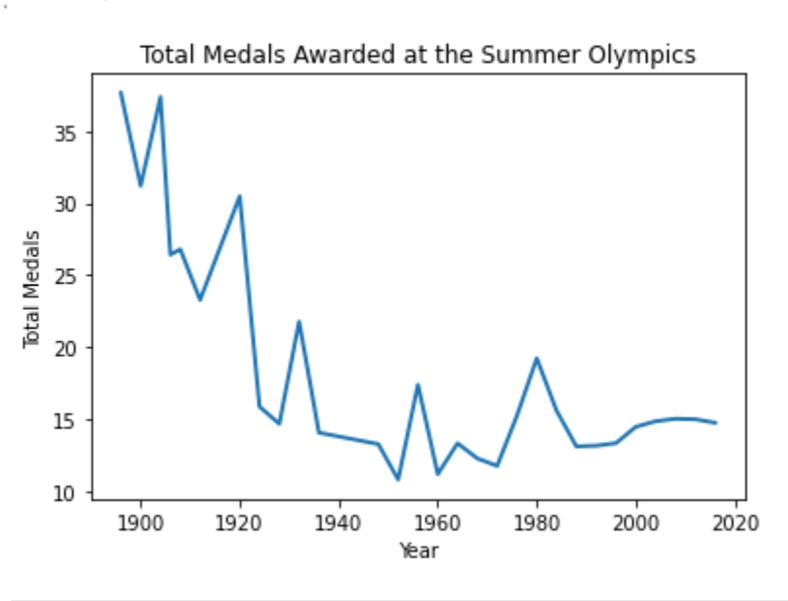
This is very interesting seeing how many medals were awarded in the early years compared to now. It does make sense though considering how many participants competed in the early years(for example in 1900-1904 there were less than 2,000 people competing compared to current times we are looking at 13,000 olympians).

Similar to my intial research in my proposal there are aggressive spikes throughout the chart which relates to the switch between Summer Games and Winter Games every 2 years. To get a better view I will break this out into two seperate charts, Summer and Winter to give us a better visual representation of the data.

Theres an interesting spike in medals around 1980 which I will need to look into further to understand what this spike relates to.

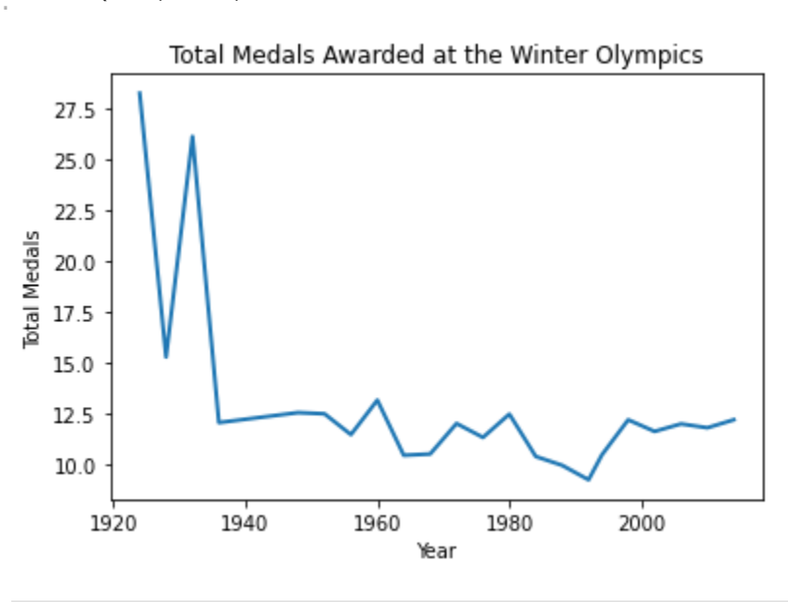
In [208…
```
#Visualizing the total amount of medals awarded throughout the years for the Summer Olympic Games

plt.plot(summer_ratio_medals.Year, summer_ratio_medals.Total_Medals, marker='', linewidth = 2, label='Total Medals')
plt.xlabel("Year")
plt.ylabel("Total Medals")
plt.title("Total Medals Awarded at the Summer Olympics")
```

Out[208]: Text(0.5, 1.0, 'Total Medals Awarded at the Summer Olympics')



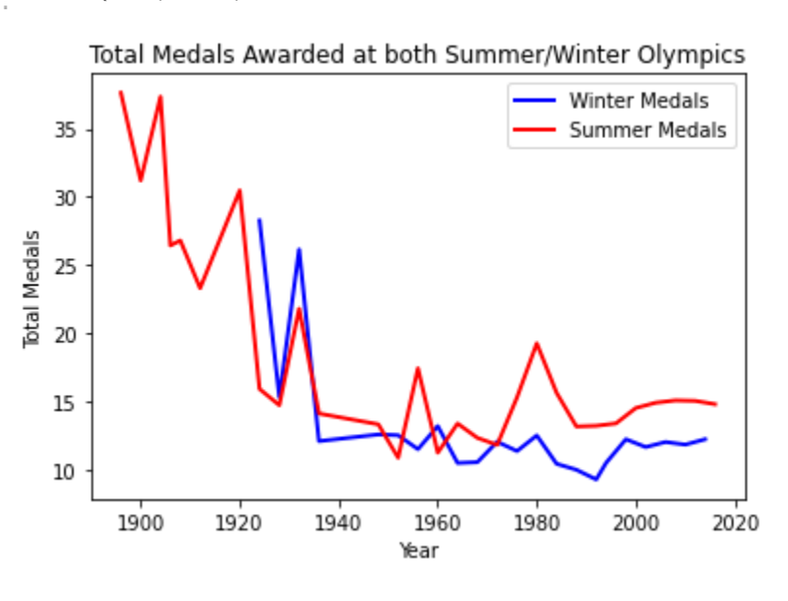
In [209…
```
#Visualizing the total amount of medals awarded throughout the years for the Winter Olympic Games

plt.plot(winter_ratio_medals.Year, winter_ratio_medals.Total_Medals, marker='', linewidth = 2, label='Total Medals')
plt.xlabel("Year")
plt.ylabel("Total Medals")
plt.title("Total Medals Awarded at the Winter Olympics")
```

Out[209]: Text(0.5, 1.0, 'Total Medals Awarded at the Winter Olympics')



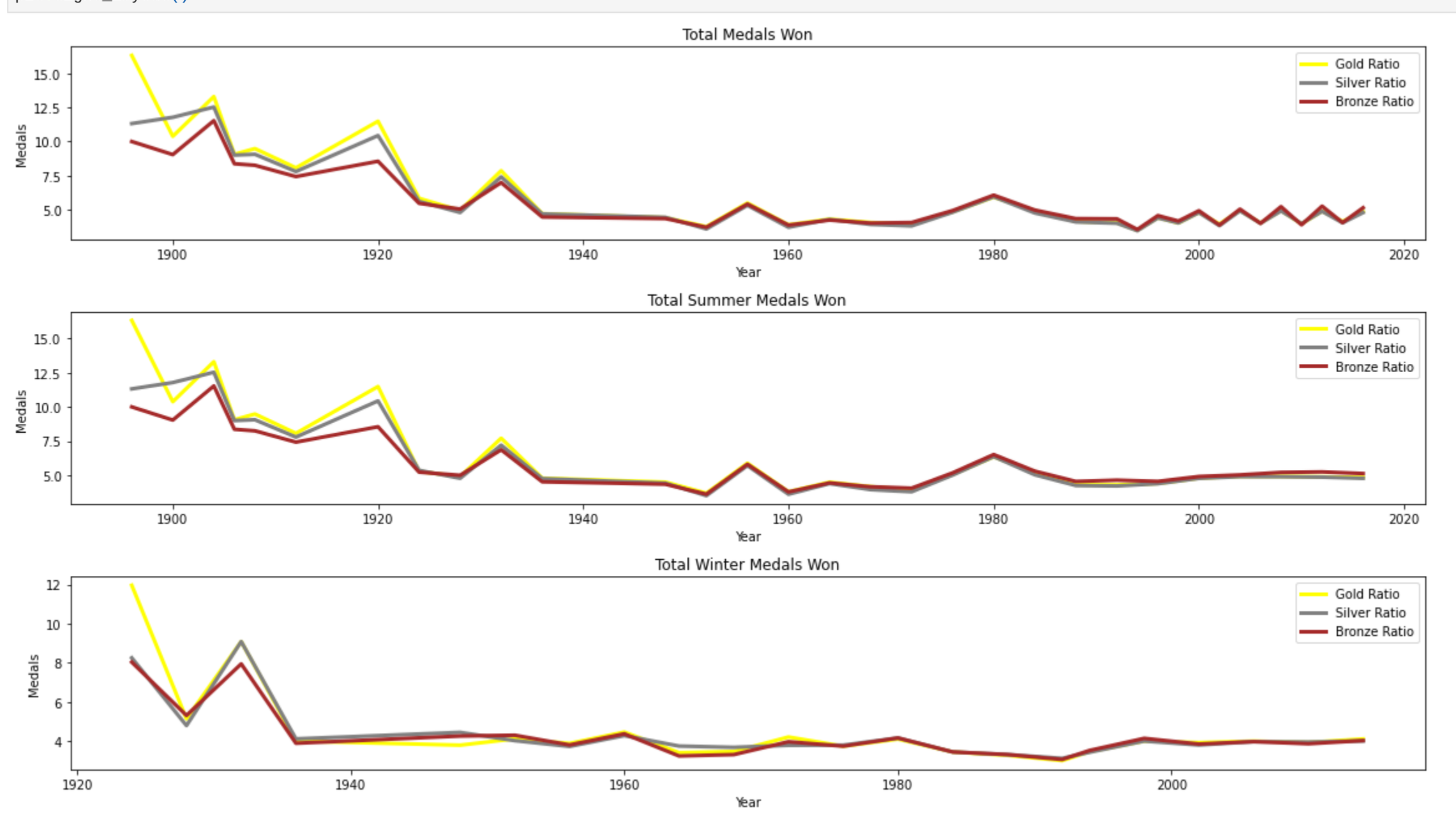
In [210…
```
#Comparing the medals awarded from the Summer games to Winter games

plt.plot(winter_ratio_medals.Year, winter_ratio_medals.Total_Medals, marker='', linewidth = 2, label='Winter Medals', color='Blue')
plt.plot(summer_ratio_medals.Year, summer_ratio_medals.Total_Medals, marker='', linewidth = 2, label='Summer Medals', color='Red')
plt.xlabel("Year")
plt.ylabel("Total Medals")
plt.legend(loc=1)
plt.title("Total Medals Awarded at both Summer/Winter Olympics")
```

Out[210]: Text(0.5, 1.0, 'Total Medals Awarded at both Summer/Winter Olympics')



Lets take a closer look at the medal % awarded throughout the year using subplots

In [211…
```
fig, ax = plt.subplots(3, figsize=(16,9))
ax[0].plot(total_ratio_medals.Year, total_ratio_medals.Gold_Medals, marker='', color='Yellow', linewidth=3, label='Gold Ratio')
ax[0].plot(total_ratio_medals.Year, total_ratio_medals.Silver_Medals, marker='', color='Grey', linewidth=3, label='Silver Ratio')
ax[0].plot(total_ratio_medals.Year, total_ratio_medals.Bronze_Medals, marker='', color='Brown', linewidth=3, label='Bronze Ratio')
ax[0].legend(loc=1)
ax[0].set_xlabel("Year")
ax[0].set_ylabel("Medals")
ax[0].set_title("Total Medals Won")
ax[1].plot(summer_ratio_medals.Year, summer_ratio_medals.Gold_Medals, marker='', color='Yellow', linewidth=3, label='Gold Ratio')
ax[1].plot(summer_ratio_medals.Year, summer_ratio_medals.Silver_Medals, marker='', color='Grey', linewidth=3, label='Silver Ratio')
ax[1].plot(summer_ratio_medals.Year, summer_ratio_medals.Bronze_Medals, marker='', color='Brown', linewidth=3, label='Bronze Ratio')
ax[1].legend(loc=1)
ax[1].set_xlabel("Year")
ax[1].set_ylabel("Medals")
ax[1].set_title("Total Summer Medals Won")
ax[2].plot(winter_ratio_medals.Year, winter_ratio_medals.Gold_Medals, marker='', color='Yellow', linewidth=3, label='Gold Ratio')
ax[2].plot(winter_ratio_medals.Year, winter_ratio_medals.Silver_Medals, marker='', color='Grey', linewidth=3, label='Silver Ratio')
ax[2].plot(winter_ratio_medals.Year, winter_ratio_medals.Bronze_Medals, marker='', color='Brown', linewidth=3, label='Bronze Ratio')
ax[2].legend(loc=1)
ax[2].set_xlabel("Year")
ax[2].set_ylabel("Medals")
ax[2].set_title("Total Winter Medals Won")
plt.tight_layout()
```



As you can see from the above figures, in the early years there were different amounts of medals distributed for each class (gold, silver, bronze). Eventually this leveled out and it looks like the same amount of medals get distributed for gold, silver and bronze now but its interesting to see the difference compared to how it started.

In [ ]:

Submit 2-3 key points you may have discovered about the data, e.g. new relationships? Aha's! Did you come up with additional ideas for other things to review?

I discovered that there was a significant drop in the amount of medals awarded per the percentage of participants. If we look at the table below we can see that in the early years there was over 30% of participants awarded a medal of some kind whereas if we look at todays figures its dropped by around 50% with the amount of participants being awarded a medal ranging from 10% -15%.

In [212…
```
total_ratio_medals.head()
```

Out[212]:

	Year	Total_Medals	Gold_Medals	Silver_Medals	Bronze_Medals
0	1896	37.631579	16.315789	11.315789	10.000000
1	1900	31.198347	10.382231	11.776860	9.039256
2	1904	37.355880	13.297463	12.528824	11.529593
3	1906	26.428159	9.059435	9.001731	8.366994
4	1908	26.797807	9.480813	9.061593	8.255401

In [213…
```
total_ratio_medals.tail()
```

Out[213]:

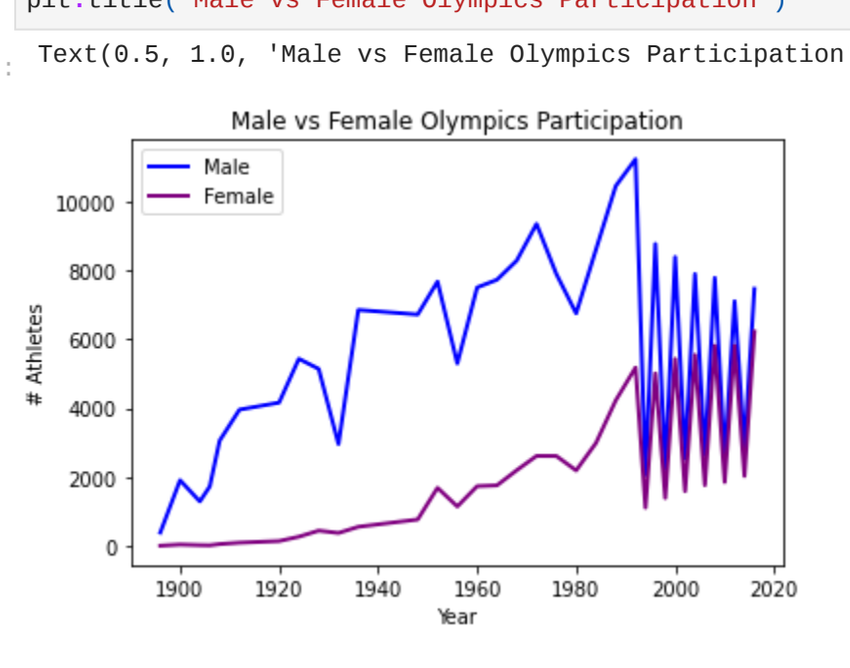
	Year	Total_Medals	Gold_Medals	Silver_Medals	Bronze_Medals
30	2008	15.056609	4.933098	4.903691	5.219821
31	2010	11.812812	3.952749	3.975466	3.884598
32	2012	15.023220	4.891641	4.876161	5.255418
33	2014	12.206093	4.130035	4.027806	4.048252
34	2016	14.779369	4.858270	4.785213	5.135885

The total count of males vs females has started to level out in recent years but in the early years it was a huge discrepency. The toal participants for the past 120 years included 74,522 (27.48%) females and a dominating 196,594 (72.51%) being males.

As you can see from the graph below the difference is starting to improve and hopefully in the coming years it will be around 50/50.

In [214…
```
plt.plot(male.Year, male.Male, marker='', color='blue', linewidth=2, label='Male')
plt.plot(female.Year, female.Female, marker='', color='purple', linewidth=2, label='Female')
plt.legend(loc=2)
plt.xlabel("Year")
plt.ylabel("# Athletes")
plt.title("Male vs Female Olympics Participation")
```

Out[214]: Text(0.5, 1.0, 'Male vs Female Olympics Participation')



The third key point I noticed was how in the early years there were different amounts of medals distributed for each class (gold, silver, bronze). Eventually this leveled out and it looks like the same amount of medals get distributed for gold, silver and bronze now but its interesting to see the difference compared to how it started.

Subplot with the 3 graphs above will help put this into perspective.

I had one 'Aha' moment while looking into the dataset.

Theres an interesting spike in medals awarded around the year 1980 which I will need to look into further to understand what this spike relates to.

Did you prove or disprove any of your initial hypotheses? If so, which one and what do you plan to do next?

1: I would think there are more men competing in the Olypmics currently but I am very interested in seeing how this has changed overtime and would think in current years it will be a lot more evenly divided. I was correct in thinking that there are more men competing currently in the olympics. I went into detail on this above but to summarize there was a huge gap in gender in the early years but it has been starting to level out ever since and more Women are competing from prior years. Eventually we will hopefully see this split level out completely and be 50/50.

2: Countries with the most medals are presuming to be the largest countries by population. e.g USA, China etc. I wasn't 100% correct on this assumption, U.S were leading by a significant amount but when it comes to countries behind them I was wrong which was why I wanted to review this dataset as I learned a lot. I then broke it down to see the Summer and Winter figures. Once again as expected for Summer games it was USA leading the pack. The winter games were led by Canada with USA coming in second and then followed by Scandavian countries such as Norway, Sweeden and Finland which is expected for the Winter games.

3: I would Imagine the age gap would be quite significant and will be interesting to see which sports they are most divided by. The age gap was very interesting to review as the youngest olympian to complete was only 10 years old which I was not aware of and the average age being from 20-30s.

What additional questions are you seeking to answer?

Theres an interesting spike in medals awarded around the year 1980 which I will need to look into further to understand what this spike relates to.