Client 3: SportsStats (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.

1. Provide a summary of the different descriptive statistics you looked at and WHY.

   Descriptive Statistics:

We look at the total number of participants (with Distinct names) and the total number of Good, Silver, and Bronze medals. The number of female participants is about a third of the male participants. However the medals per female participants is higher than the medals per male participants indicating that there are more male participants per corresponding event than female participants.

```
In [16]: pysqldf("SELECT Sex, Count(DISTINCT Name) AS Participants, Count(Medal) AS Total FROM df Group By Sex;")
```

Out[16]:

	Sex	Participants	Total
0	F	33808	11253
1	M	100979	28530

Trends in Participation and Medals by Year:

Female Athletes:

```
In [81]: df1 = pysqldf("SELECT Season, Year, Sex, Count(DISTINCT Name) AS Participant, Count(Medal) AS Total_Medals FROM df WHERE (Sex = 'F') Group By Sea
```

We can see distinct trends for summer olympics participants (top red curve) and winter olympics participants (bottom red curve) and the corresponding total medals awarded (blue curves). The trajectories of the medals and the participants are similar in growth.

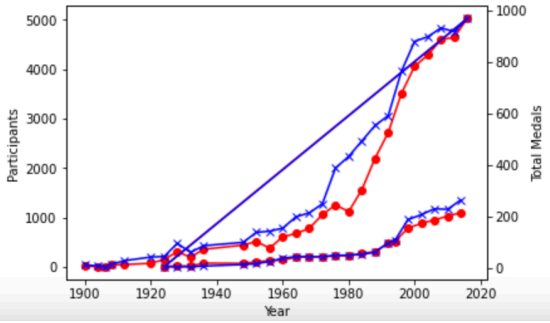
```
In [79]: import matplotlib.pyplot as plt

         %matplotlib inline
         x = df1['Year']
         y1 = df1['Participant']
         y2 = df1['Total']
         Fig, ax1 = plt.subplots()
         ax2 = ax1.twinx()

         curve1 = ax1.plot(x, y1, color ='r', marker ='o')
         curve2 = ax2.plot(x, y2, color ='b', marker ='x')

         ax1.set_xlabel('Year')
         ax1.set_ylabel('Participants')
         ax2.set_ylabel('Total Medals')

         plt.show()
```



Male Athletes:

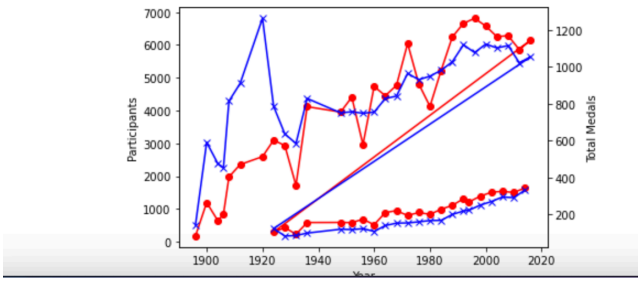
```
In [92]: import matplotlib.pyplot as plt

         %matplotlib inline
         x = df1['Year']
         y1 = df1['Participant']
         y2 = df1['Total_Medals']
         Fig, ax1 = plt.subplots()
         ax2 = ax1.twinx()

         curve1 = ax1.plot(x, y1, color ='r', marker ='o')
         curve2 = ax2.plot(x, y2, color ='b', marker ='x')

         ax1.set_xlabel('Year')
         ax1.set_ylabel('Participants')
         ax2.set_ylabel('Total Medals')

         plt.show()
```



We see significant spikes in medals for 1920, 1912, 1908 and will need to be analyzed further to identify root cause

Total medals by year:

As expected, the number of medals for both summer and winter Olympics mostly increase over a period of time

In [23]:
```
df_year = pysqldf("SELECT Season, Year, Count(Medal) AS Total_Medals FROM df Group By Season, Year")

df_year.head(70)
```

Out[23]:

	Season	Year	Total_Medals
0	Summer	1896	143
1	Summer	1900	604
2	Summer	1904	486
3	Summer	1906	458
4	Summer	1908	831

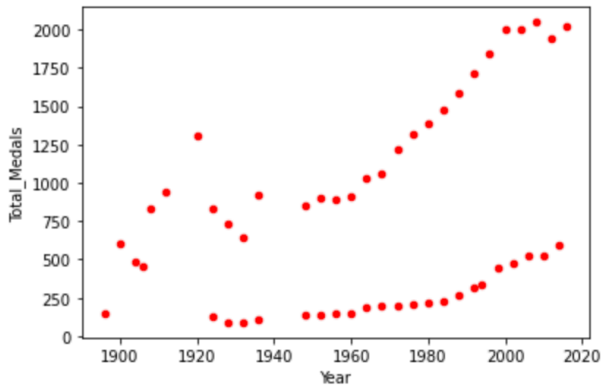
In [28]:
```
import matplotlib.pyplot as plt

%matplotlib inline
#x = df_year['Year']
#y = df_year['Total_Medals']

ax = df_year.plot.scatter(x='Year', y='Total_Medals', color ='r', marker ='o')

ax.set_xlabel('Year')
ax.set_ylabel('Total_Medals')

plt.show()
```



Total medals by sport:

Athletics, Swimming and Rowing have the most number of medals indicating the most number of events

```
In [22]: df_sport = pysqldf("SELECT Sport, Count(Medal) AS Total_Medals FROM df Group By Sport ORDER BY Total_Medals DESC")
         df_sport.head(50)
```

Out[22]:

	Sport	Total_Medals
0	Athletics	3969
1	Swimming	3048
2	Rowing	2945
3	Gymnastics	2256
4	Fencing	1743
5	Football	1571
6	Ice Hockey	1530
7	Hockey	1528
8	Wrestling	1296
9	Cycling	1263
10	Sailing	1232

Total number of medals by sport by country:

```
In [39]: df_CS = pysqldf("SELECT Sport, NOC, Count(Medal) AS Total_Medals FROM df Group By NOC, Sport ORDER BY Sport, Total_Medals DESC")
         df_CS.head(500)
```

Out[39]:

	Sport	NOC	Total_Medals
0	Aeronautics	SUI	1
1	Alpine Skiing	AUT	114
2	Alpine Skiing	SUI	59
3	Alpine Skiing	FRA	45
4	Alpine Skiing	USA	44
...	...	...	...
495	Badminton	ALG	0
496	Badminton	AUS	0
497	Badminton	AUT	0
498	Badminton	BEL	0
499	Badminton	BLR	0

Total number of medals by category by country over the time horizon:

```
In [*]: df_MCC = pysqldf("SELECT NOC, Team, Year, Season, Sport, (CASE WHEN Medal = 'Gold' then 1 else 0 end) AS GM, (CASE WHEN Medal = 'Silver' then 1 e
```

```
In [48]: df_MCC.head(50)
```

Out[48]:

	NOC	Team	Year	Season	Sport	GM	SM	BM
0	CHN	China	1992	Summer	Basketball	0	0	0
1	CHN	China	2012	Summer	Judo	0	0	0
2	DEN	Denmark	1920	Summer	Football	0	0	0
3	DEN	Denmark/Sweden	1900	Summer	Tug-Of-War	1	0	0
4	NED	Netherlands	1988	Winter	Speed Skating	0	0	0
5	NED	Netherlands	1988	Winter	Speed Skating	0	0	0
6	NED	Netherlands	1992	Winter	Speed Skating	0	0	0

```
In [49]: df_MC = pysqldf("SELECT NOC, Team, SUM(GM) AS GMT, SUM(SM) AS SMT, SUM(BM) AS BMT FROM df_MCC GROUP BY NOC ORDER BY GMT DESC")
         df_MC.head(25)
```

Out[49]:

	NOC	Team	GMT	SMT	BMT
0	USA	United States	2638	1641	1358
1	URS	Soviet Union	1082	732	689
2	GER	Germany	745	674	746
3	GBR	Great Britain	678	739	651
4	ITA	Italy	575	531	531
5	FRA	France	501	610	666
6	SWE	Sweden	479	522	535
7	CAN	Canada	463	438	451
8	HUN	Hungary	432	332	371
9	GDR	East Germany	397	327	281
10	RUS	Russia	390	367	408
11	NOR	Norway	378	361	294
12	CHN	China	350	347	292
13	AUS	Australia	348	455	517
14	NED	Netherlands	287	340	413

We observe that the maximum number of Gold medals (GMT) has gone to the US followed by the former USSR, Germany and Great Britain

Maximum number of gold medals over horizon for each country:

```
In [36]: df_CountryBestSport = pysqldf("SELECT NOC, Team, Sport, Max(GMT) AS Gold_Medals FROM df_MC1 GROUP BY NOC HAVING Gold_Medals > 0")
         df_CountryBestSport.head(50)
```

1	ANZ	Australasia	Rugby	15
2	ARG	Argentina	Football	34
3	ARM	Armenia	Wrestling	2
4	AUS	Australia	Swimming	118
5	AUT	Austria	Alpine Skiing	34
6	AZE	Azerbaijan	Wrestling	4
7	BAH	Bahamas	Athletics	12
8	BDI	Burundi	Athletics	1
9	BEL	Belgium	Archery	35
10	BLR	Belarus	Canoeing	6
11	BRA	Brazil	Volleyball	60
12	BRN	Bahrain	Athletics	1

The above table tells us which sport may be the most played with a strong team by a country in the olympics as that's the sport in which the country has gotten most gold medals over the olympic history. For instance, Australia likely has participated in a significant number of events in Swimming with a strong team, as did Brazil in Volleyball.

Min, Max, Avg number of total medals by country over the time horizon:

```
In [51]: df_MinMaxAvgNOC = pysqldf("SELECT NOC, MIN(Temp.Total_M) AS Min_Medals, AVG(Temp.Total_M) AS Avg_Medals, MAX(Temp.Total_M) AS Max_Medals FROM (SE

         df_MinMaxAvgNOC.head(25)
```

Out[51]:

	NOC	Min_Medals	Avg_Medals	Max_Medals
0	EUN	279	279.000000	279
1	URS	56	250.300000	496
2	GDR	33	167.500000	303
3	USA	19	161.057143	394
4	FRG	8	97.666667	166
5	GER	16	83.269231	236
6	RUS	0	72.812500	189
7	GBR	1	59.085714	368
8	FRA	2	50.771429	235
9	CHN	0	49.450000	184
10	ITA	0	46.771429	104

2. 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?

Looking at trends in participation only by year, I observed that the numbers were going up and down and realized that the summer and winter olympics had to be separated to analyze trends.

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

The initial hypothesis was that the gender gap would be narrower over a period of time. Looking at the number of participants over time (male and female) and medals awarded by gender. Looking at the charts below, the number of female participants (in red) were a fraction of male participants in the early 1900s while by mid-2010's the gap had closed significantly. Similar trend in number of medals where in by mid-2010's the number of medals were trending closer to 1000 (per summer olympic event) for both genders though there is still a gap to be closed.

Female Athletes:

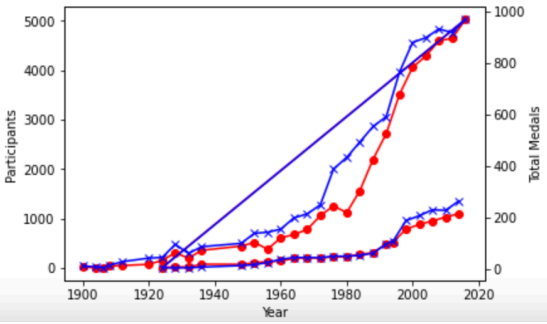
In [79]:
```
import matplotlib.pyplot as plt

%matplotlib inline
x = df1['Year']
y1 = df1['Participant']
y2 = df1['Total']
Fig, ax1 = plt.subplots()
ax2 = ax1.twinx()

curve1 = ax1.plot(x, y1, color ='r', marker ='o')
curve2 = ax2.plot(x, y2, color ='b', marker ='x')

ax1.set_xlabel('Year')
ax1.set_ylabel('Participants')
ax2.set_ylabel('Total Medals')

plt.show()
```



Male Athletes:

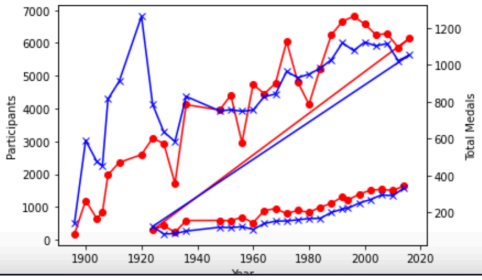
In [92]:
```
import matplotlib.pyplot as plt

%matplotlib inline
x = df1['Year']
y1 = df1['Participant']
y2 = df1['Total_Medals']
Fig, ax1 = plt.subplots()
ax2 = ax1.twinx()

curve1 = ax1.plot(x, y1, color ='r', marker ='o')
curve2 = ax2.plot(x, y2, color ='b', marker ='x')

ax1.set_xlabel('Year')
ax1.set_ylabel('Participants')
ax2.set_ylabel('Total Medals')

plt.show()
```



4. What additional questions are you seeking to answer?

   How has the medal tally by country changed over time?