Olympic Historical Dataset (1896 - 2022)

Min Pak • Flora Aka • Maria Ardila

Questions we found interesting

- Country medal count versus economy (GDP) & Population
- How age has changed and how age is a factor in different events
- Women and men’s participation throughout the years
- How height and weight have changed through the years

Data you used to answer these questions

- Kaggle: Olympic Historical Dataset (1896 - 2022)
  - Olympic_Athlete_Bio.csv
  - Olympic_Athlete_Event_Results.csv
  - Olympic_Games_Medal_Tally.csv
  - Olympic_Results.csv
  - Olympics_Country.csv
  - Olympics_Games.csv

https://www.kaggle.com/datasets/muhammadehsan000/olympic-historical-dataset-1896-2020?select=Olympic_Athlete_Event_Results.csv

Additional Data Sources

- World Bank Group
  - data.worldbank.org/indicator/SP.POP.TOTL
    - csv file with the population from different countries from 1960 to current
  - data.worldbank.org/indicator/NY.GDP.MKTP.CD
    - csv file with the GDP of different countries from 1960 to current

Data exploration and cleanup process

DataFrame with the countries that won medals needed alterations so the data in DataFrames would match

```
#1960 Olympics change from Soviet Union
medal_df.at[341, 'country'] = 'Russian Federation'

#1980 Olympics change from Soviet Union
medal_df.at[559, 'country'] = 'Russian Federation'

#2020 Olympics change from ROC
medal_df.at[1254, 'country'] = 'Russian Federation'
```

Additional Data Cleanup

```
#change East Germany to Germany for 1980
medal_df.at[560, 'country'] = 'Germany'
```

GDP DataFrame

```
#fix United Kingdom to Great Britain
gdp_df.at[81, 'Country Name'] = 'Great Britain'

#add GDP to Russia Federation (USSR) for 1980 and 1960
gdp_df.at[202, '1980'] = 1210000000000
gdp_df.at[202, '1960'] = 142400000000
```

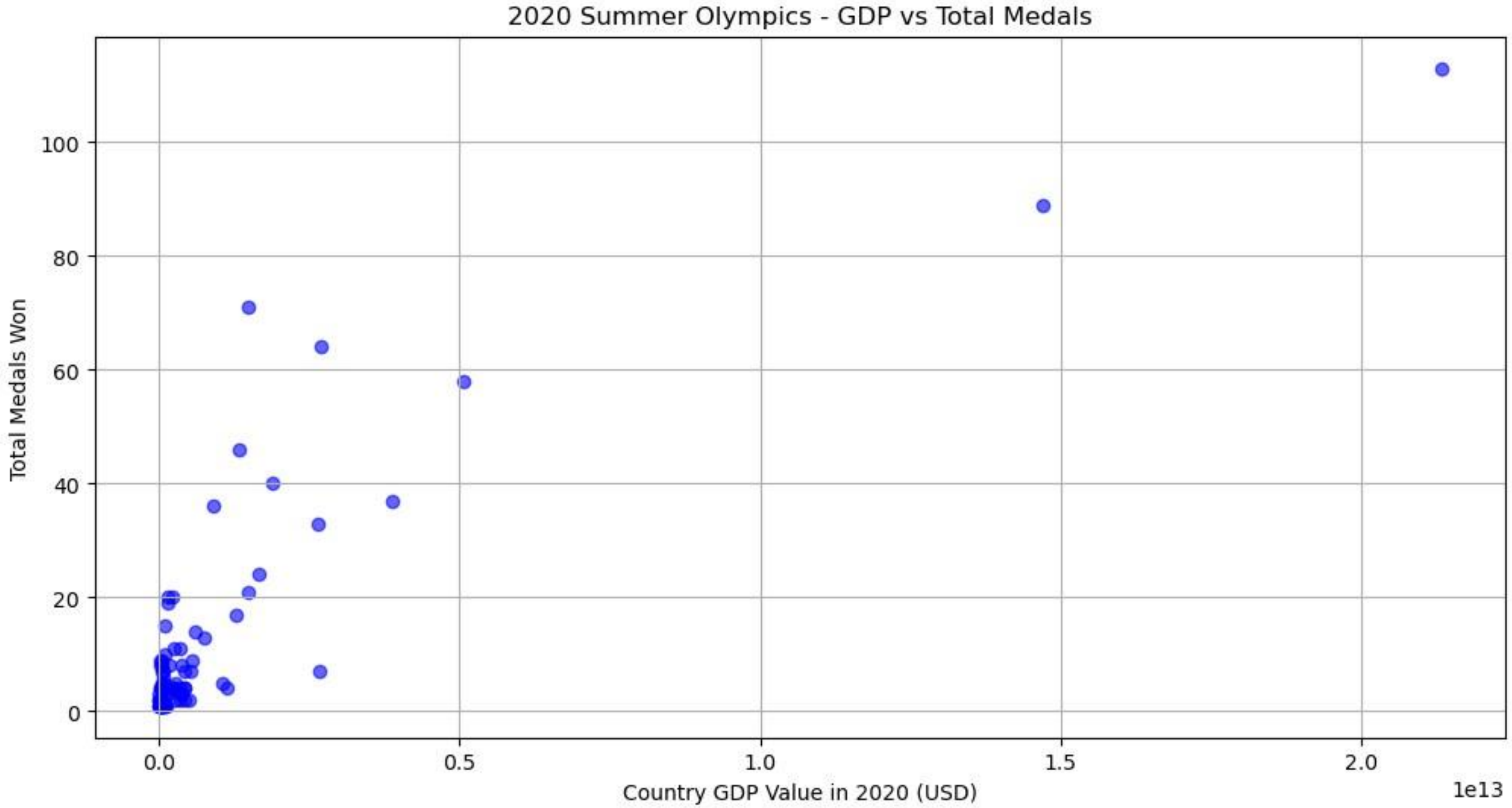
Analysis process

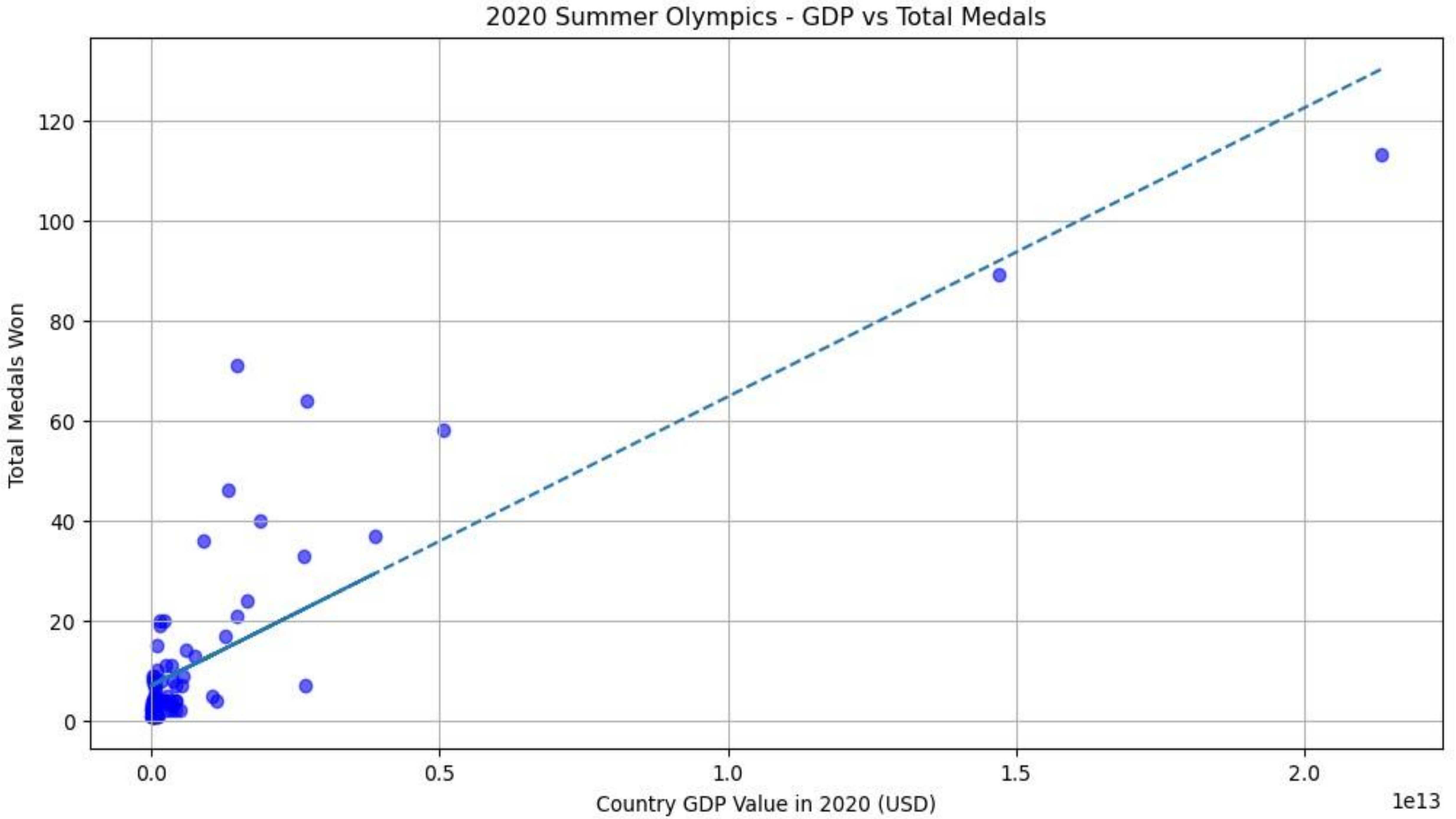
After merging GDP and Medals DataFrames and reorganzing the columns, I plotted the Data. I first used GDP per capita but did not like the results. I switched to Total GDP.

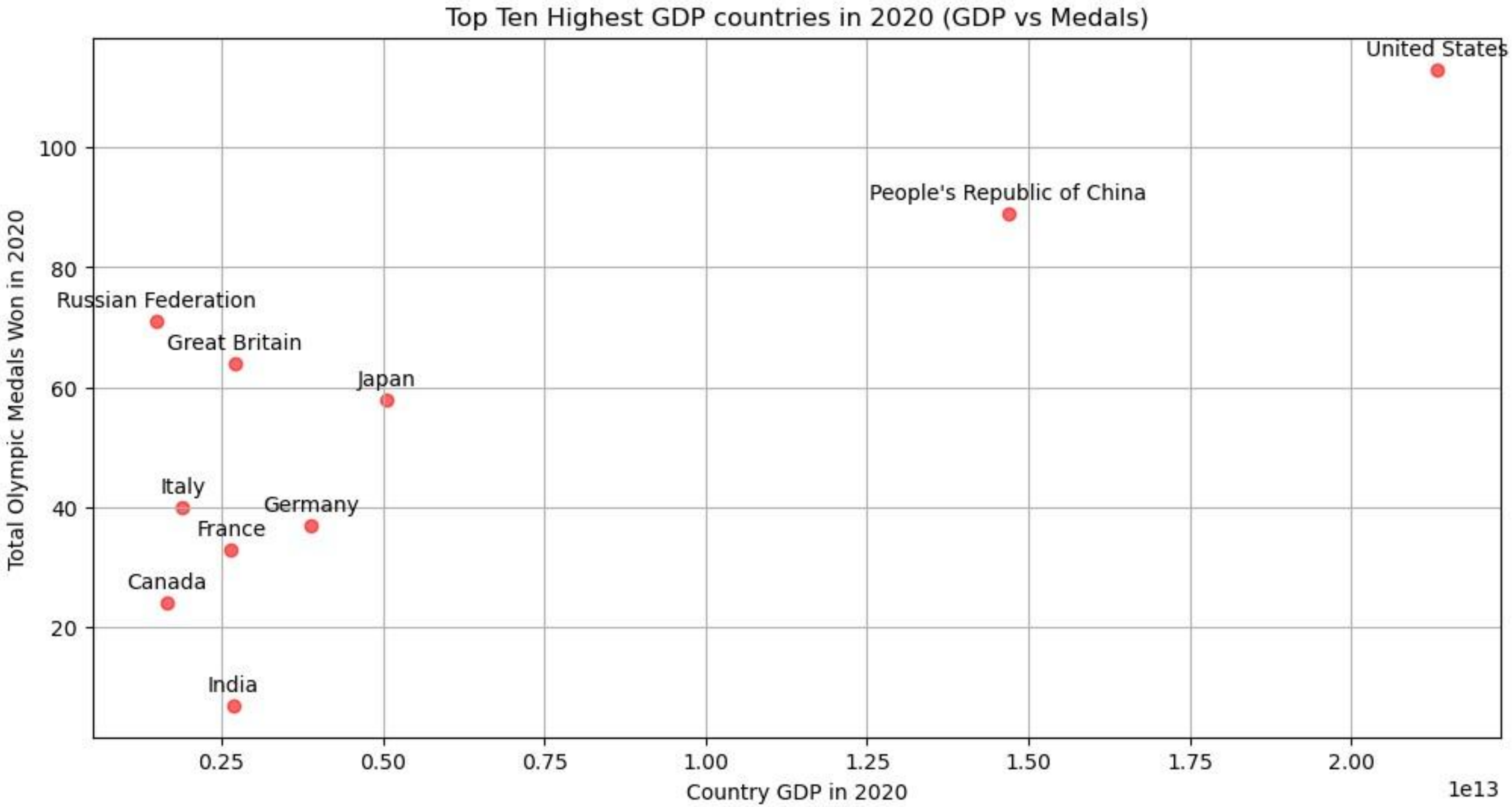
I made sample plots of 2020, 2016, 2012, 2008, and 2004 Summer Olympics and I found them to be very similar.

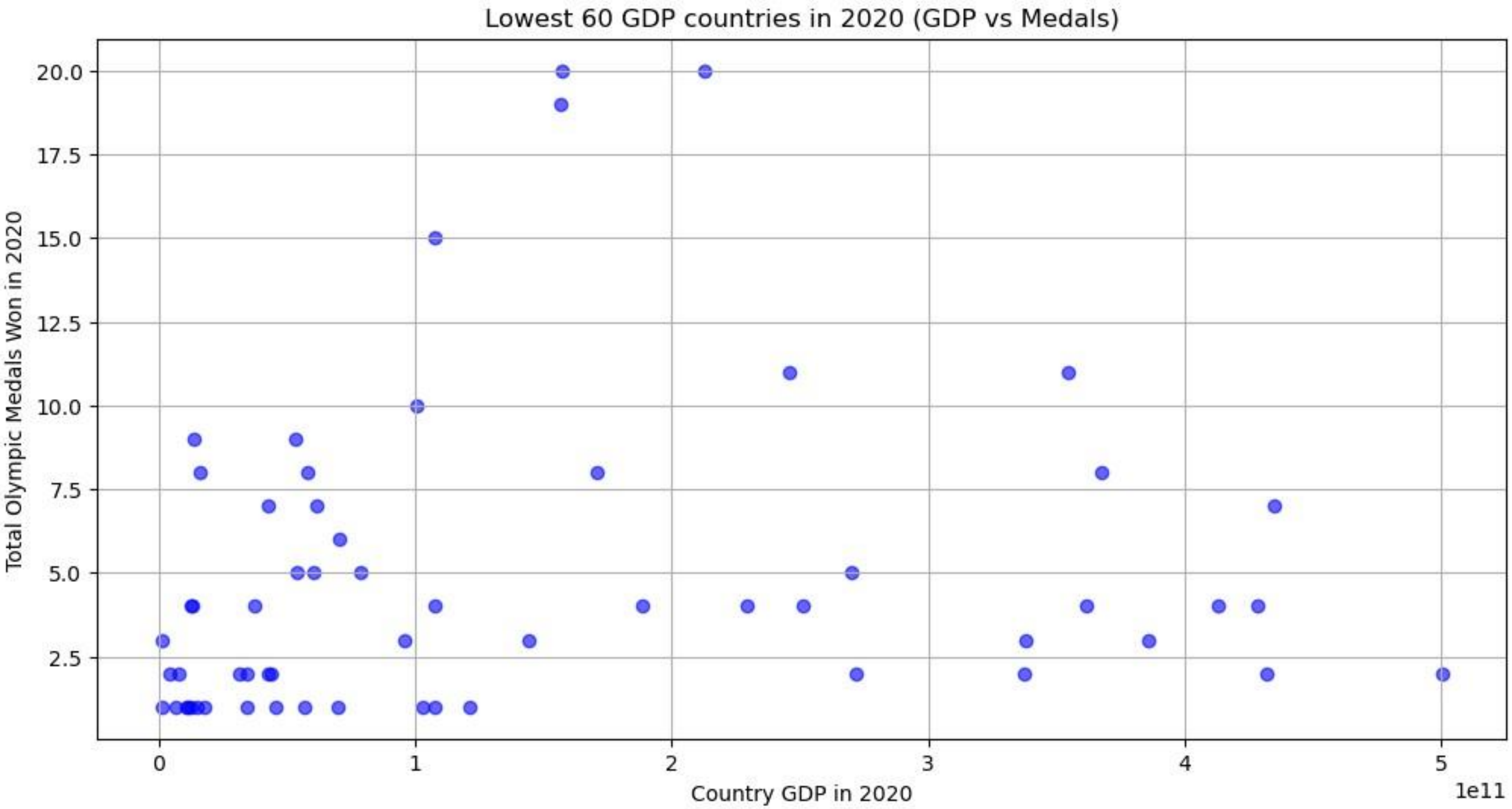
After seeing this, I decided looking at a wider range of years would provide more interesting data. So I reorganized the Data for 1960, 1980, 2000, and 2020 Summer Olympics.

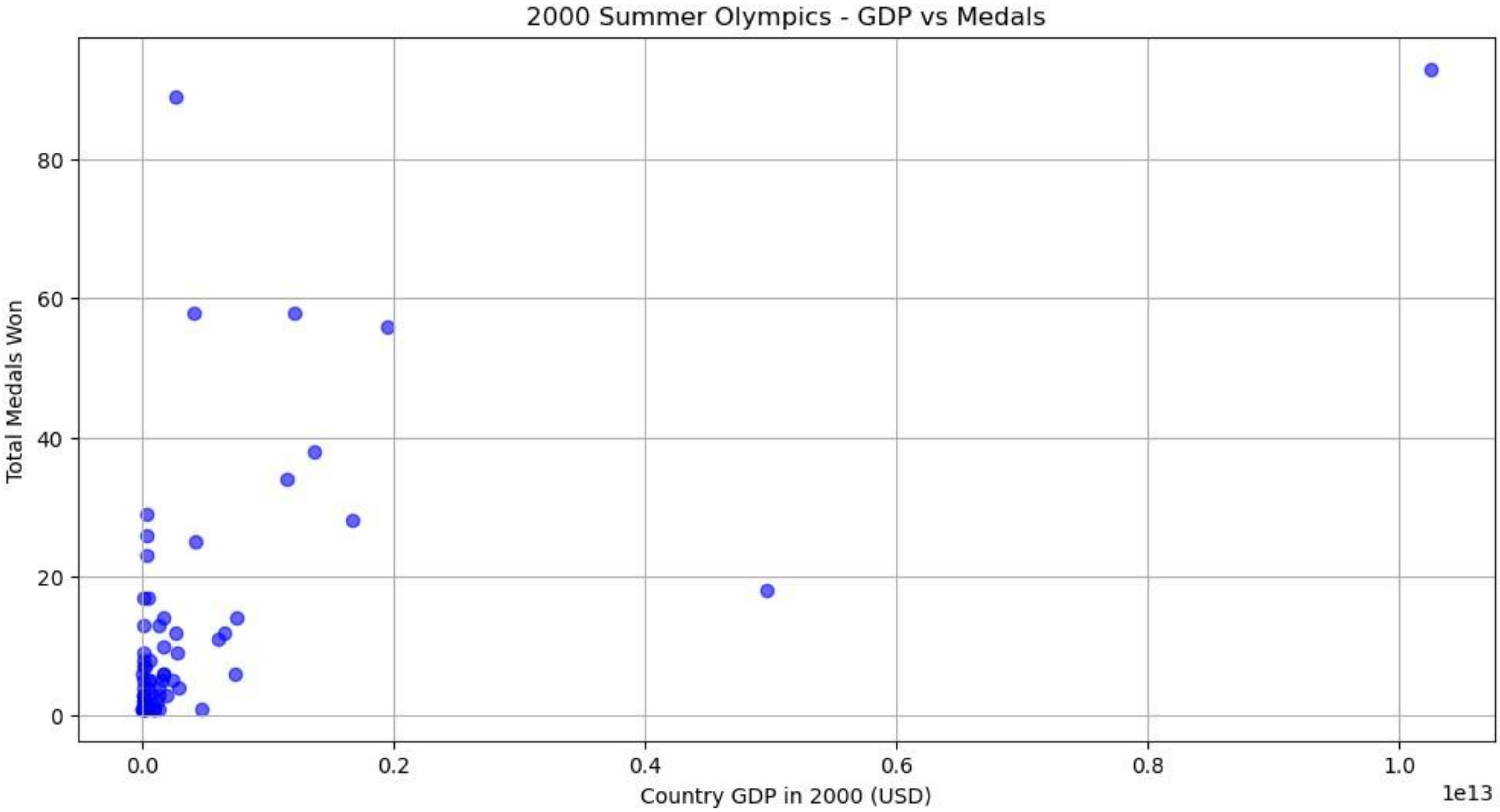
2020 Summer Olympics

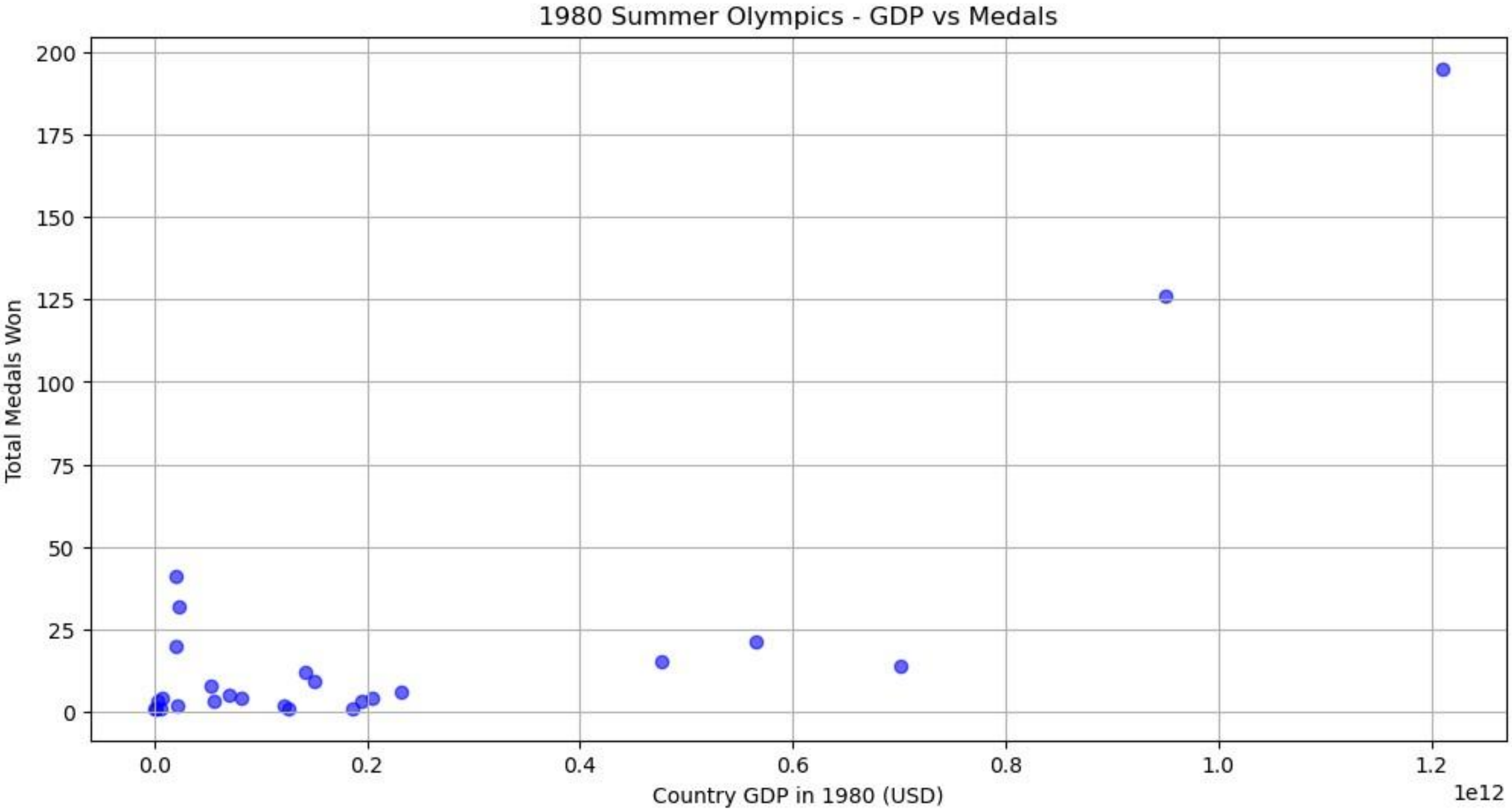


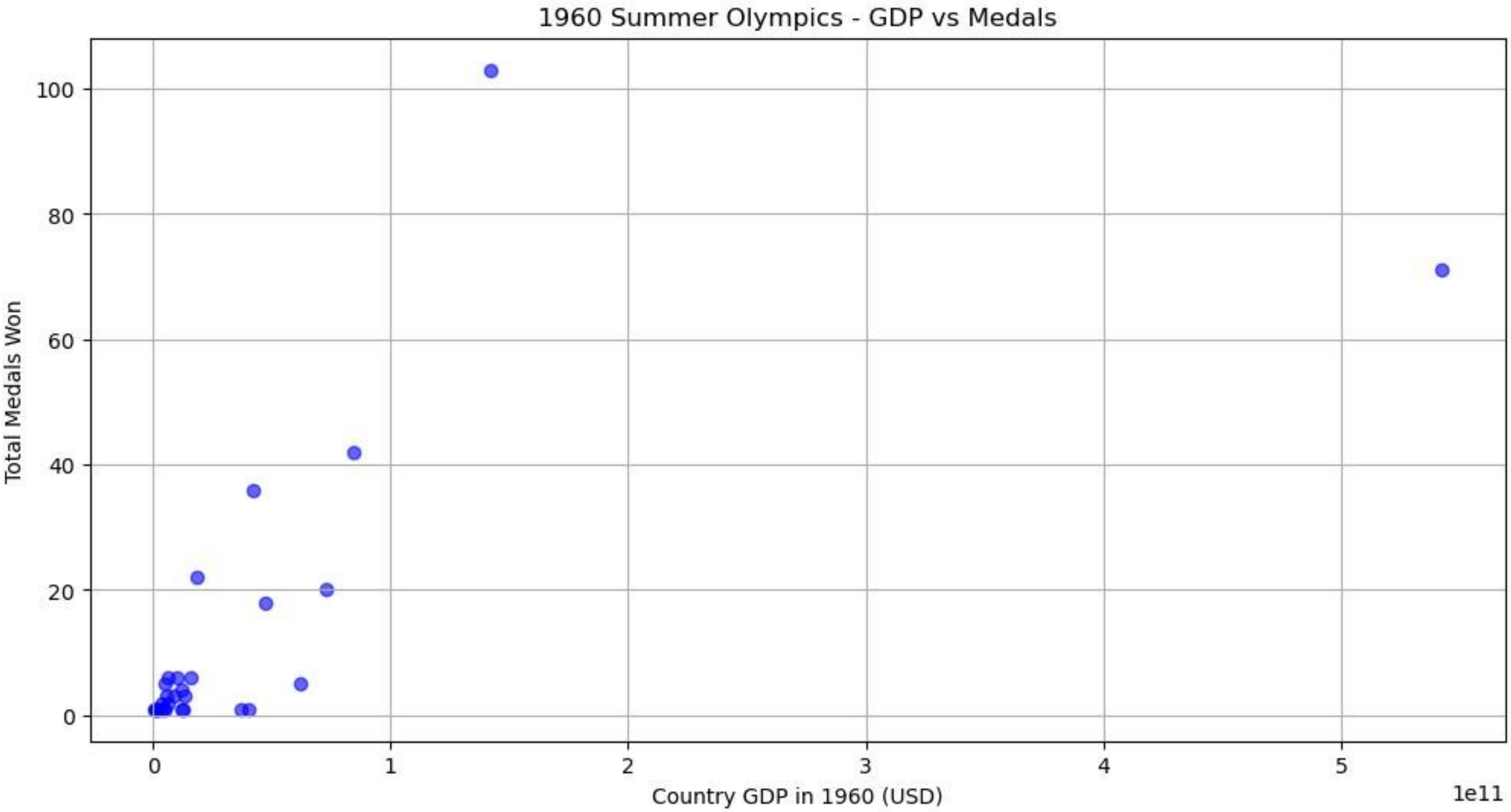

2020 Summer Olympics - GDP vs Total Medals
Total Medals Won
0
20
40
60
80
100
120
0.0
0.5
1.0
1.5
2.0
1e13
Country GDP Value in 2020 (USD)

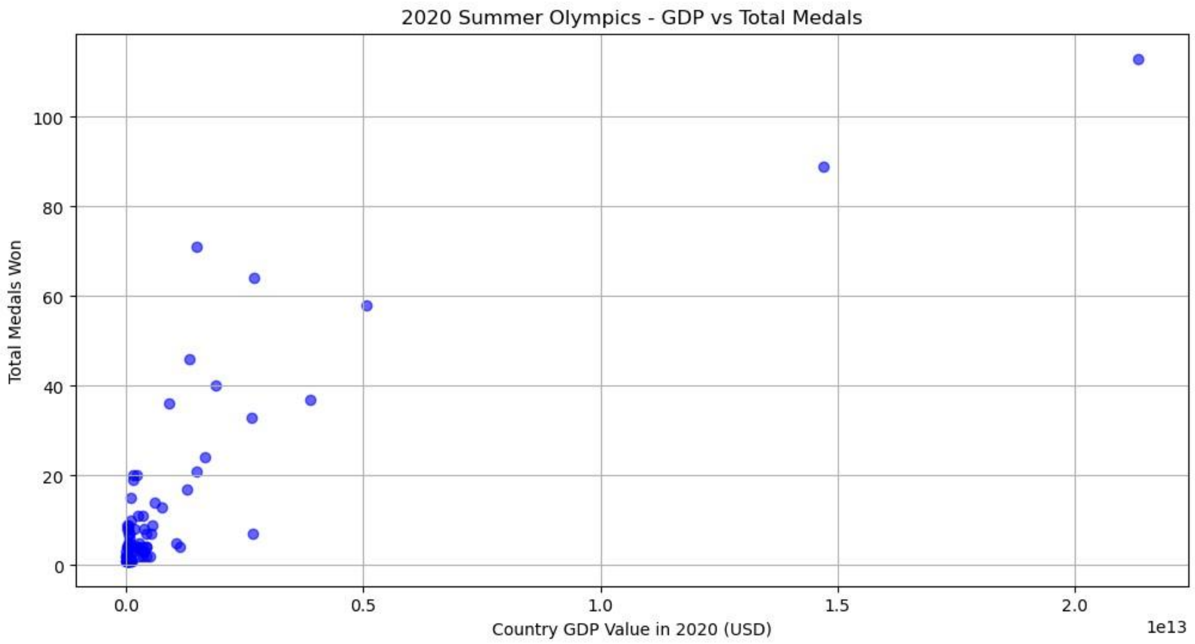

Top Ten Highest GDP countries in 2020 (GDP vs Medals)
United States
People's Republic of China
Russian Federation
Great Britain
Japan
Italy
Germany
France
Canada
India
100
80
60
40
20
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
1e13
Total Olympic Medals Won in 2020
Country GDP in 2020

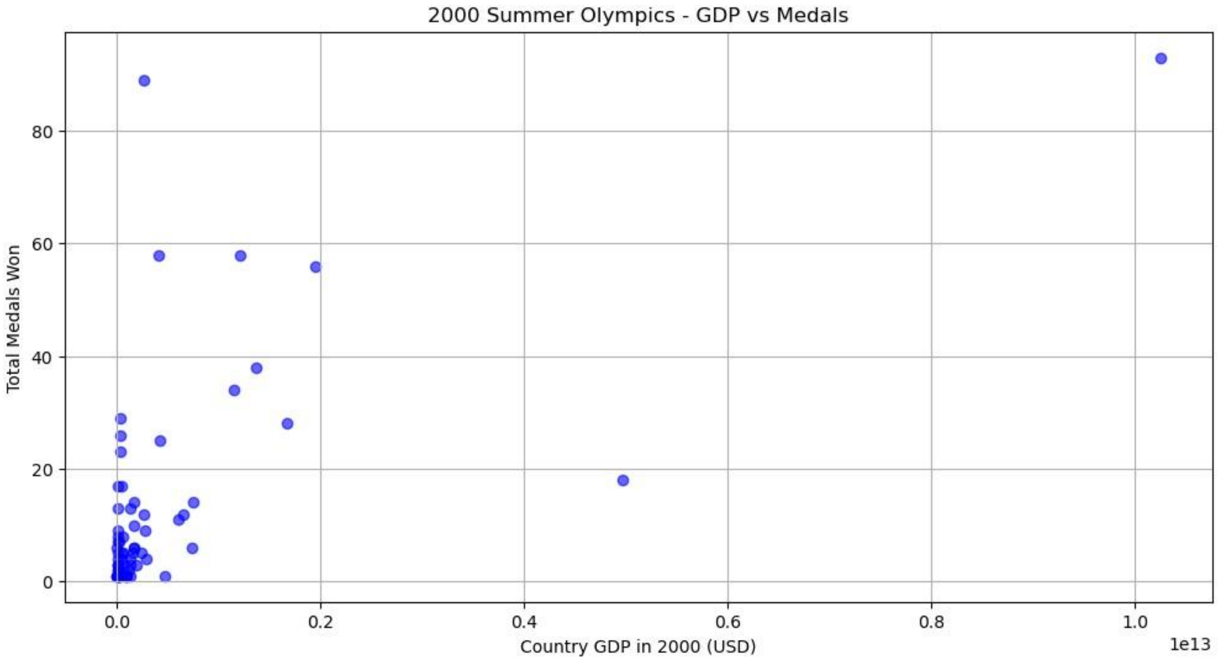

Lowest 60 GDP countries in 2020 (GDP vs Medals)
Total Olympic Medals Won in 2020
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0
1
2
3
4
5
1e11
Country GDP in 2020

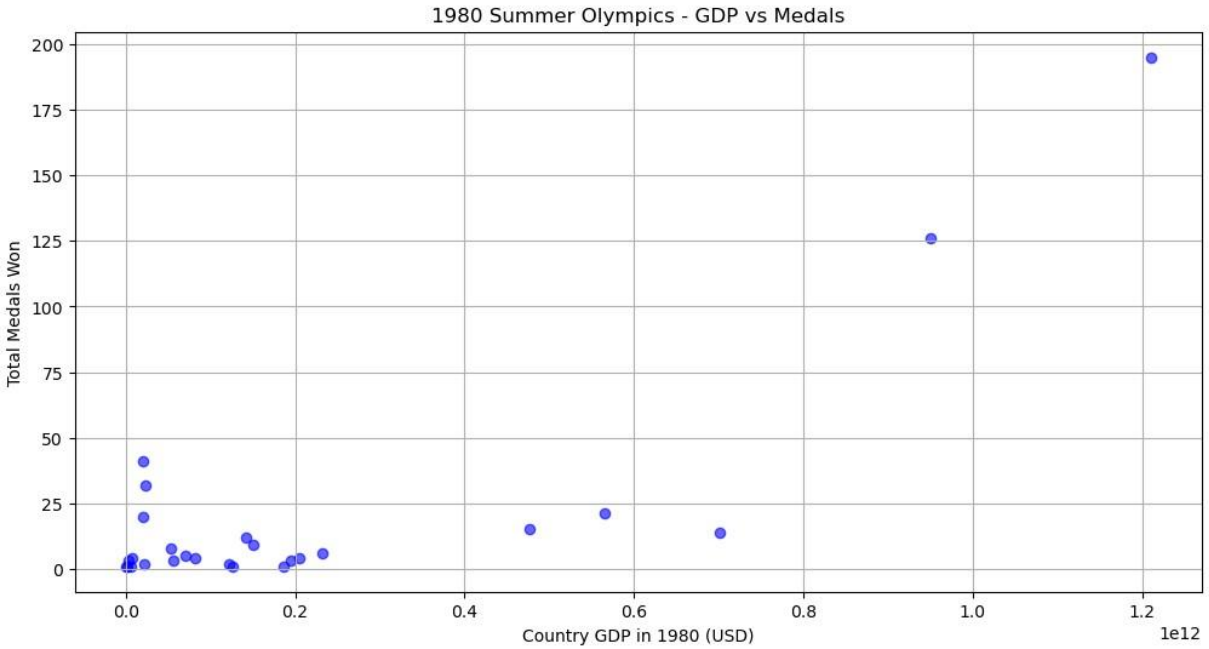

2000 Summer Olympics - GDP vs Medals
Total Medals Won
0
20
40
60
80
0.0
0.2
0.4
0.6
0.8
1.0
1e13
Country GDP in 2000 (USD)

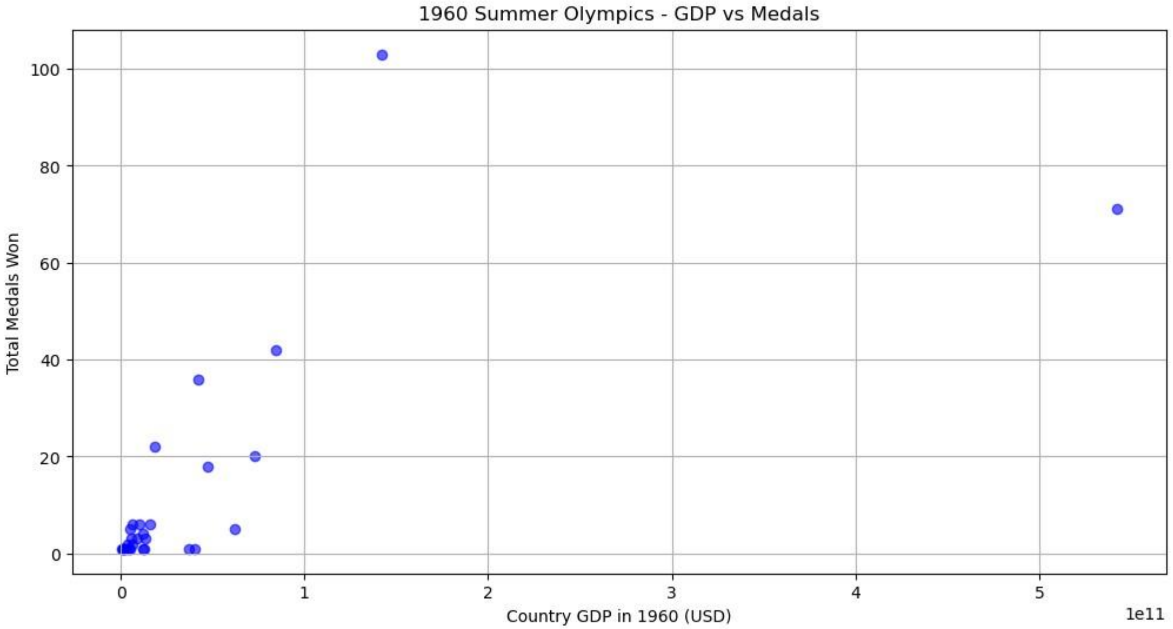

1980 Summer Olympics - GDP vs Medals
Total Medals Won
200
175
150
125
100
75
50
25
0
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1e12
Country GDP in 1980 (USD)

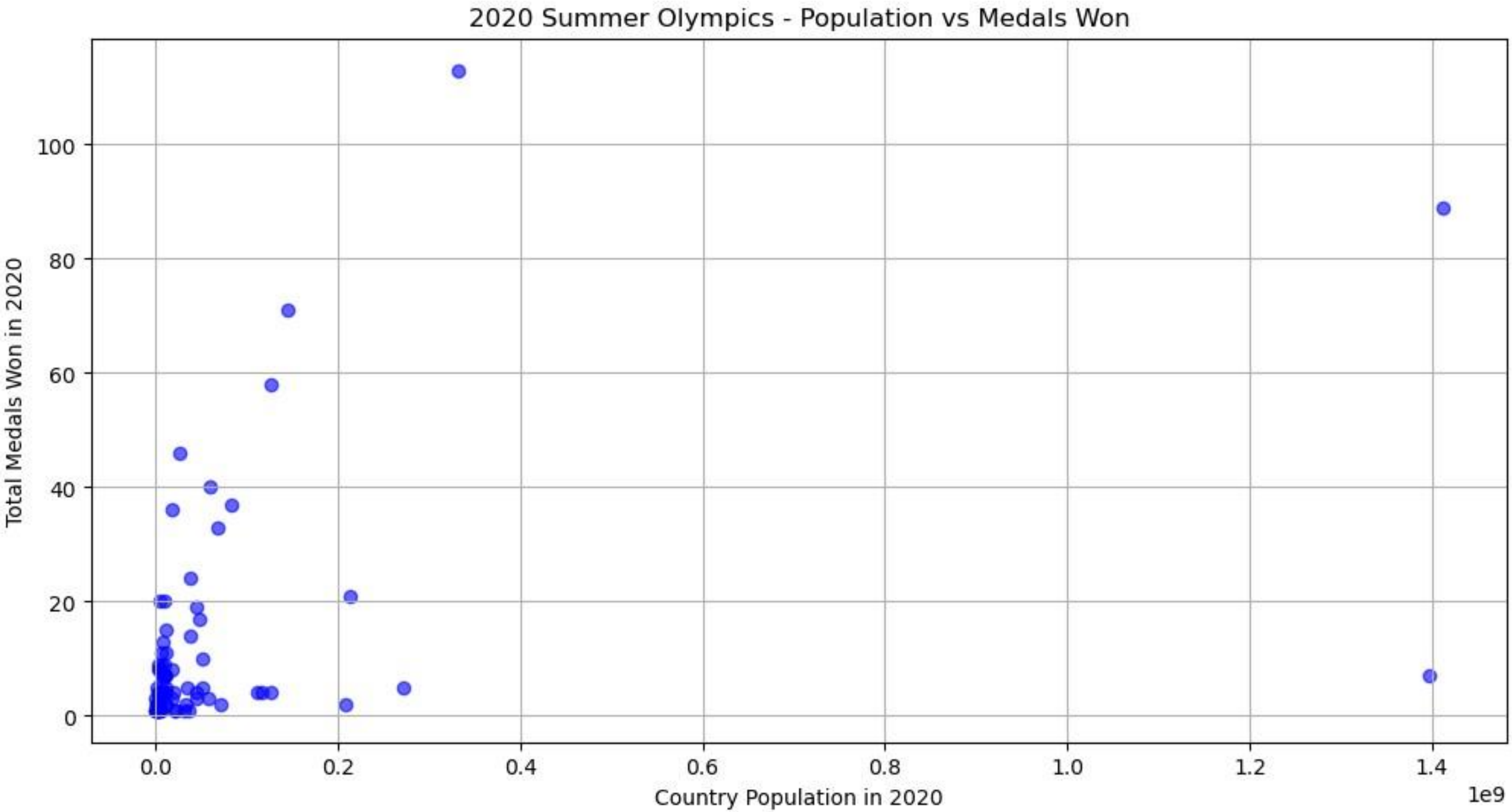

1960 Summer Olympics - GDP vs Medals
Total Medals Won
0
20
40
60
80
100
0
1
2
3
4
5
1e11
Country GDP in 1960 (USD)

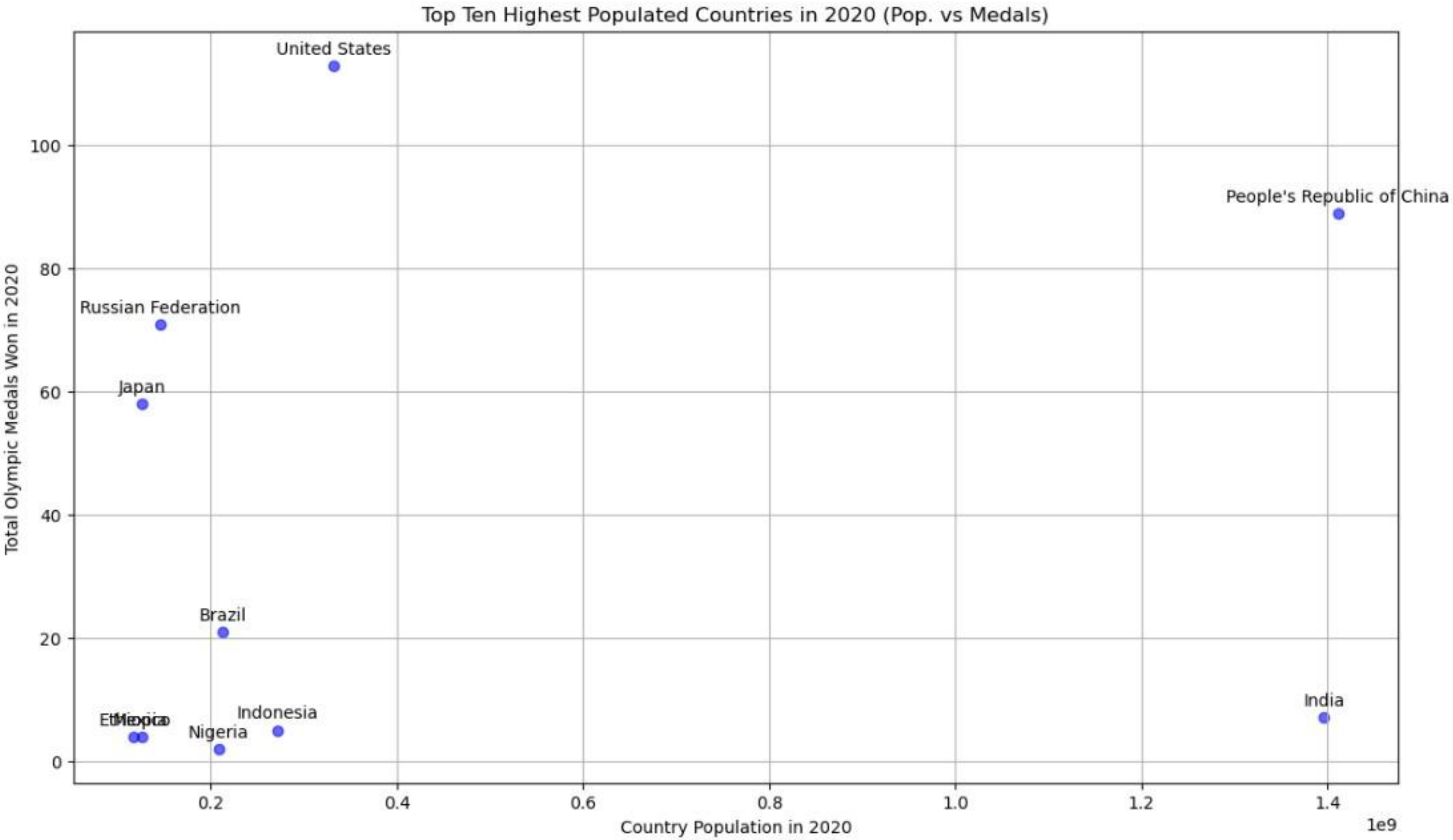

2020 Summer Olympics - GDP vs Total Medals
Total Medals Won
0
20
40
60
80
100
0.0
0.5
1.0
1.5
2.0
1e13
Country GDP Value in 2020 (USD)

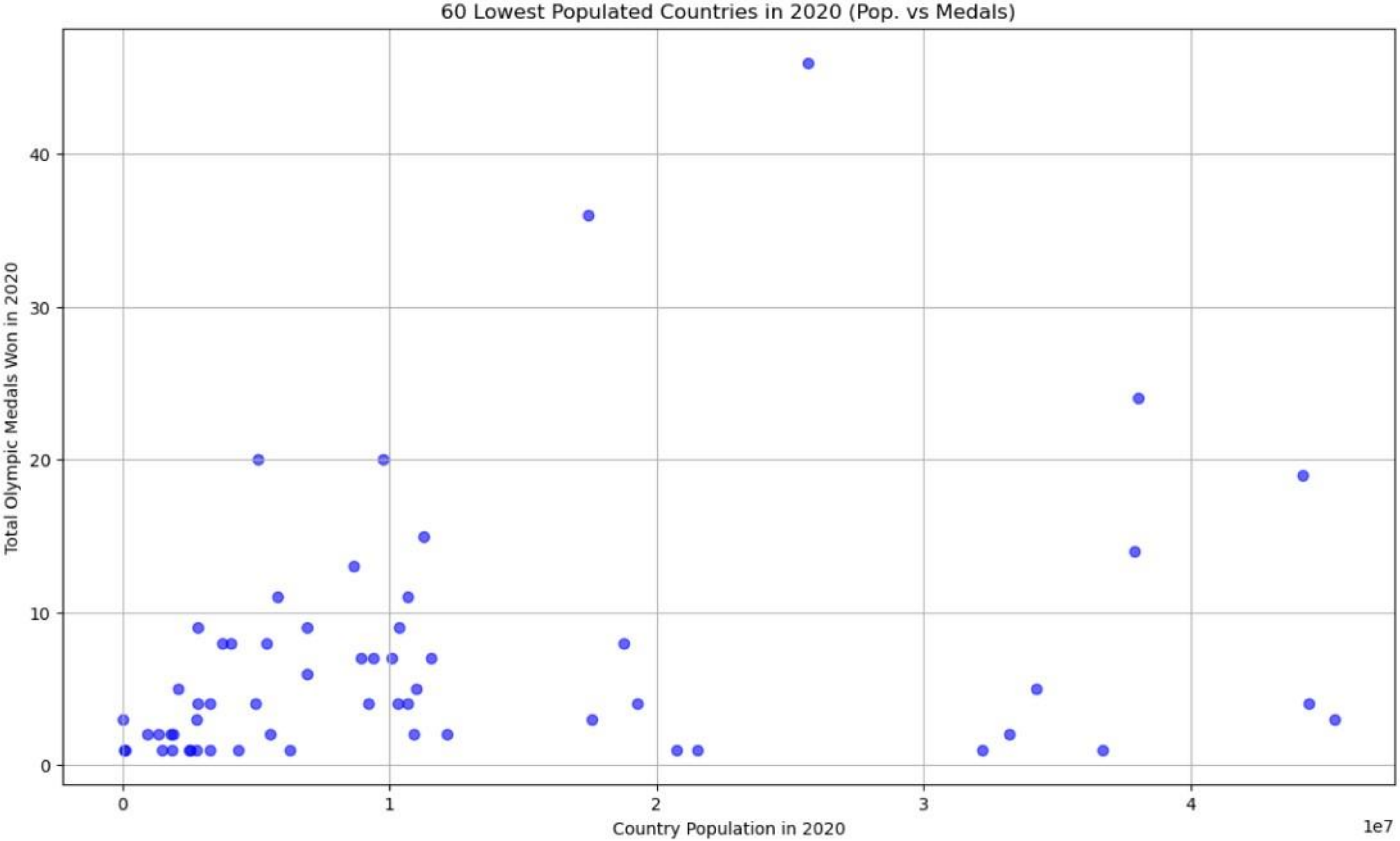

2000 Summer Olympics - GDP vs Medals
Total Medals Won
0
20
40
60
80
0.0
0.2
0.4
0.6
0.8
1.0
1e13
Country GDP in 2000 (USD)

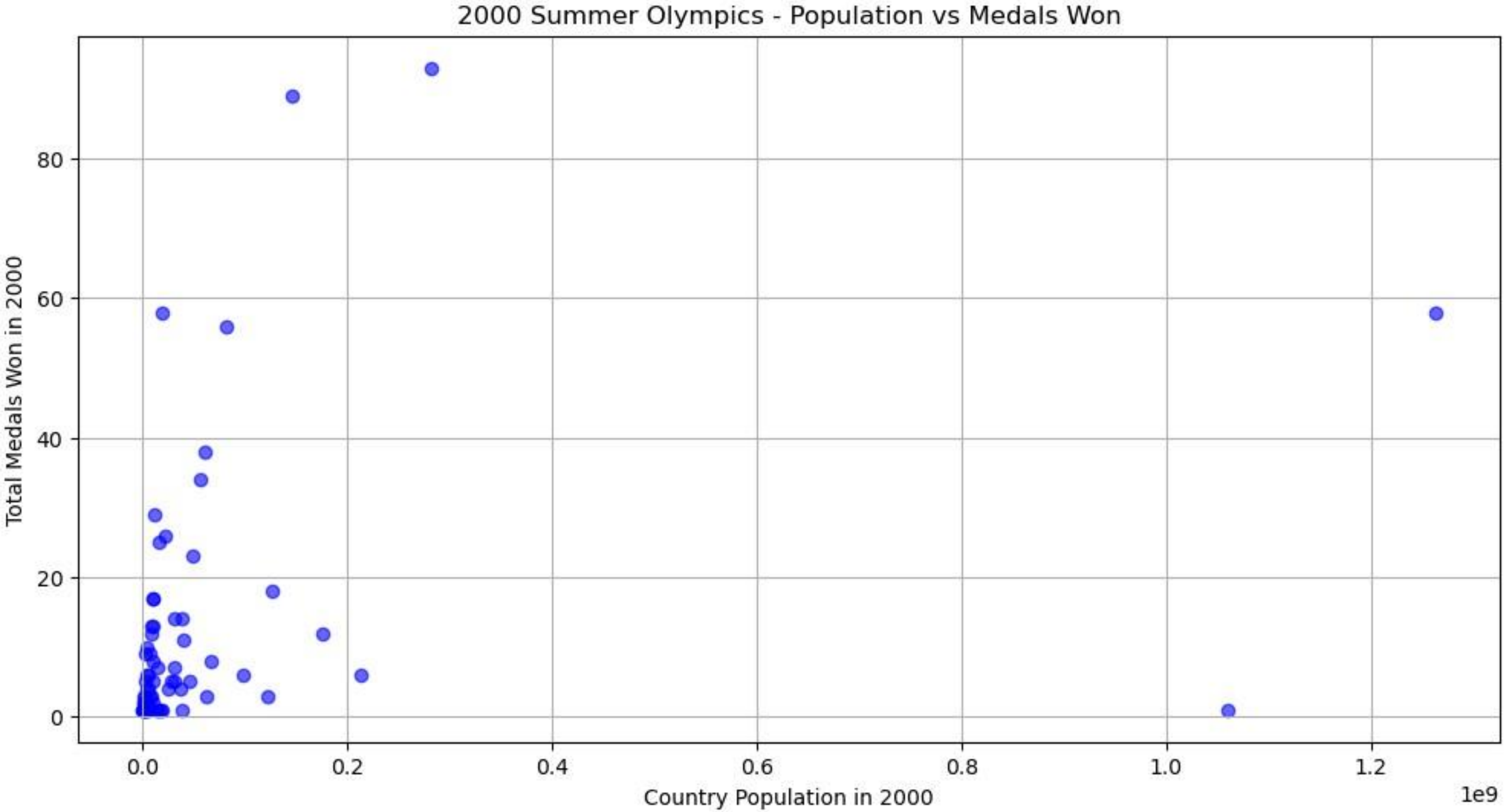

1980 Summer Olympics - GDP vs Medals
Total Medals Won
0
25
50
75
100
125
150
175
200
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1e12
Country GDP in 1980 (USD)

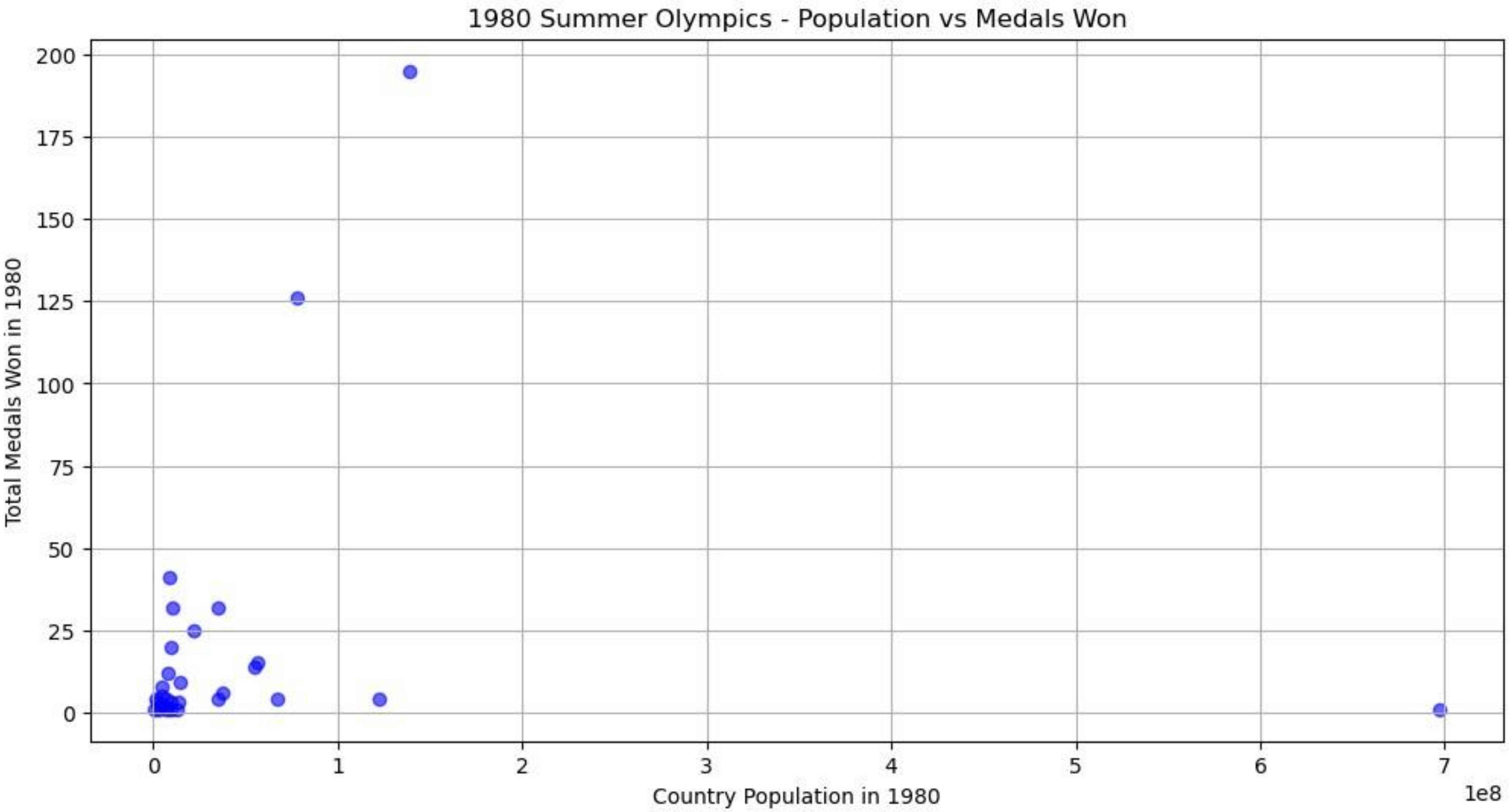

1960 Summer Olympics - GDP vs Medals
Total Medals Won
0
20
40
60
80
100
0
1
2
3
4
5
1e11
Country GDP in 1960 (USD)

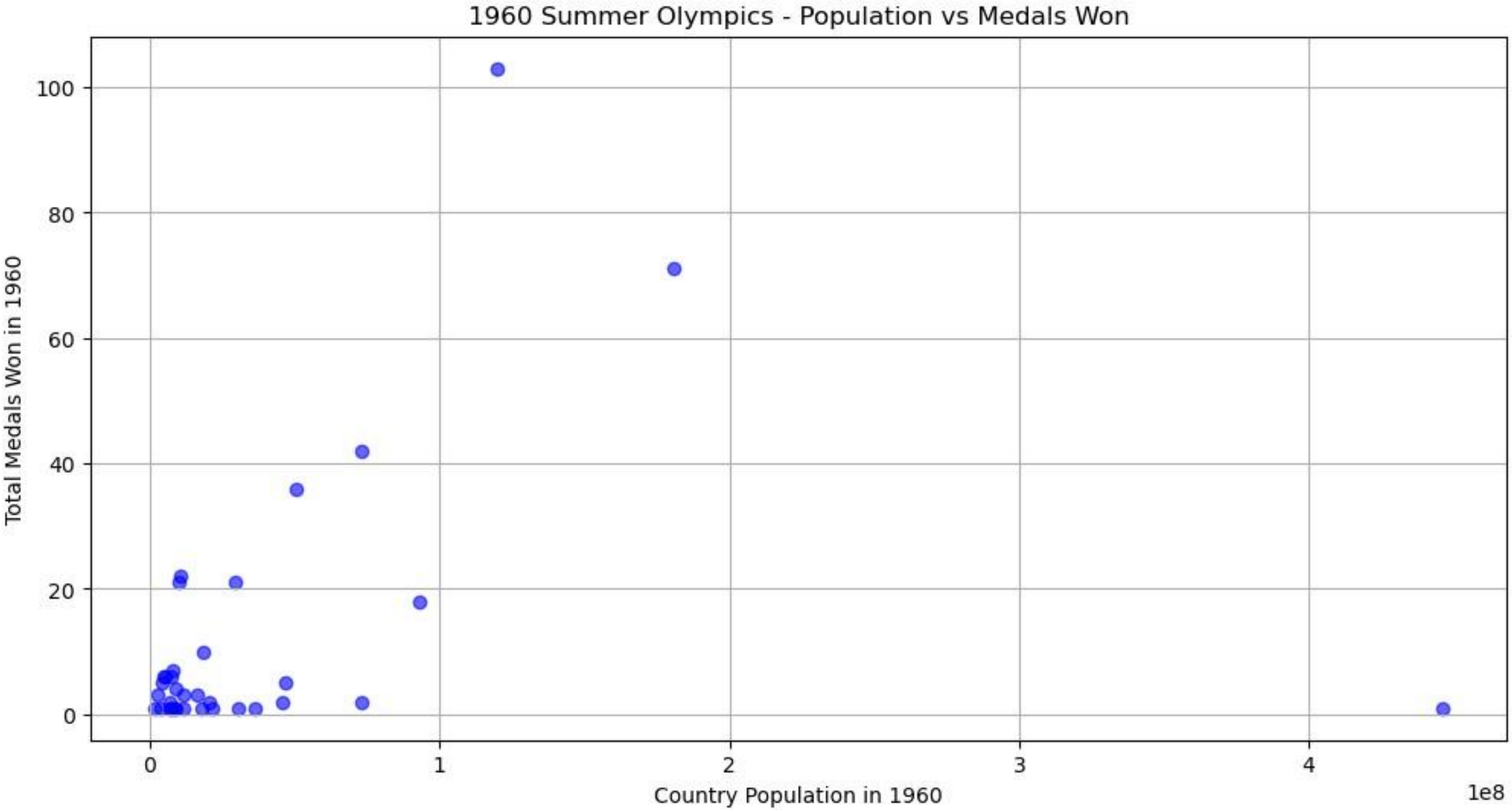

2020 Summer Olympics - Population vs Medals Won
Total Medals Won in 2020
0
20
40
60
80
100
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1e9
Country Population in 2020


Top Ten Highest Populated Countries in 2020 (Pop. vs Medals)
United States
People's Republic of China
Russian Federation
Japan
Brazil
Ethiopia
Mexico
Indonesia
Nigeria
India
Total Olympic Medals Won in 2020
0
20
40
60
80
100
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1e9
Country Population in 2020


60 Lowest Populated Countries in 2020 (Pop. vs Medals)
Total Olympic Medals Won in 2020
0
10
20
30
40
Country Population in 2020
0
1
2
3
4
1e7


2000 Summer Olympics - Population vs Medals Won
Total Medals Won in 2000
0
20
40
60
80
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1e9
Country Population in 2000


1980 Summer Olympics - Population vs Medals Won
Total Medals Won in 1980
200
175
150
125
100
75
50
25
0
0
1
2
3
4
5
6
7
1e8
Country Population in 1980


1960 Summer Olympics - Population vs Medals Won
Total Medals Won in 1960
0
20
40
60
80
100
Country Population in 1960
0
1
2
3
4
1e8

Limitations

- Assuming the GDP and Population numbers are correct in these CSV
- No Data charted for countries that did not win medals
- No Consideration for the Athletes who made it to the Finals
- How much a country invests into olympic training is probably a better indicator (India)
- Outside Factors of Cheating and Favoritism (Drugs, Biased Judges, Bribed Judges)

Conclusions

Based on this Data, GDP and Population are one of the positive factors for how many total medals a country wins at the Summer Olympics.