Project – Summer Olympics

# Team B2:

* Damiso Hutchinson
* Diana Kennen
* Caroline Larry
* Andrew Miller
* Valerie Pippenger

# Data Sets Used from Kaggle:

* <https://www.kaggle.com/divyansh22/summer-olympics-medals> (CSV)
  + Data period 1976 through 2008
* <https://ourworldindata.org/grapher/gross-domestic-product> (CSV)

# Data Cleaning Challenges:

We read in both the summer\_olympics.csv and the gdp\_csv as dataframes to start. Originally, this did not work due to encoding errors, but the ISO-8859-1 encoding setting seemed to work. We next dropped any missing data rows from the olympics dataframe. In order that the merge point between the summer olympics and gdp be identical, we converted the year value to be an integer, then a string. There was also a large gap of blank rows in the summer olympics csv that caused some trouble, but once we discovered this, we could proceed. South Korea and North Korea had different name formats in both dataframes, so these had to be replaced in the olympics dataframe to match the format in the gdp dataframe. In addition, a function was created to create a new column with the host country from the name of the host city. This was needed for the GDP analysis.

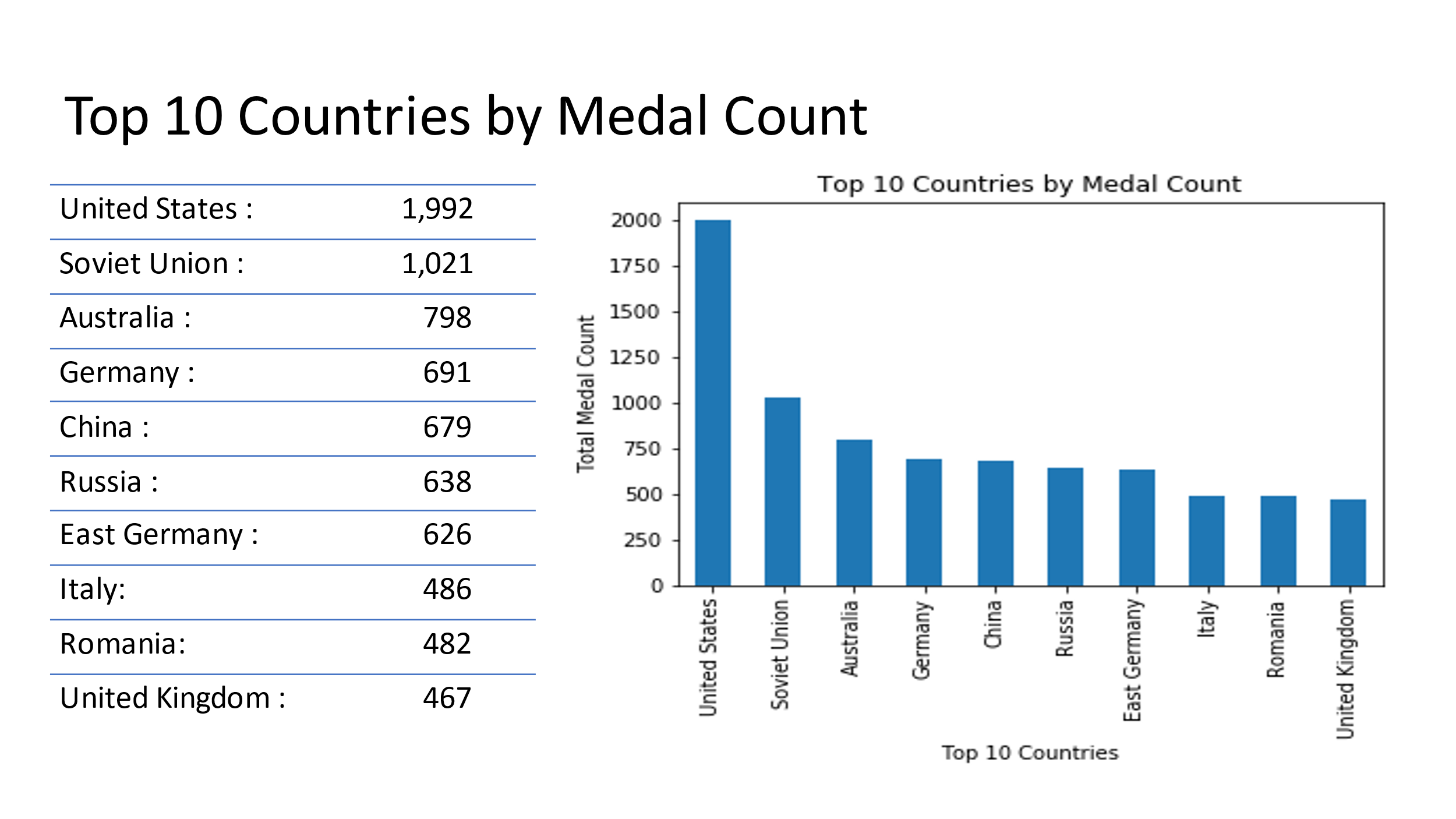
To merge the two dataframes, we created a new column with a concatenation of the country name and year in both dataframes. The two dataframes were merged on this combo column using a left merge. Our GDP dataframe was missing some values for Eastern Bloc countries prior to the dissolution of the USSR. The medal rows were left for these countries so that the medal data analysis would be based on a complete set of the medal data.

The cleaned dataframe, olympics\_df, was then passed on for use in the analysis portion of the project.

# Analysis:

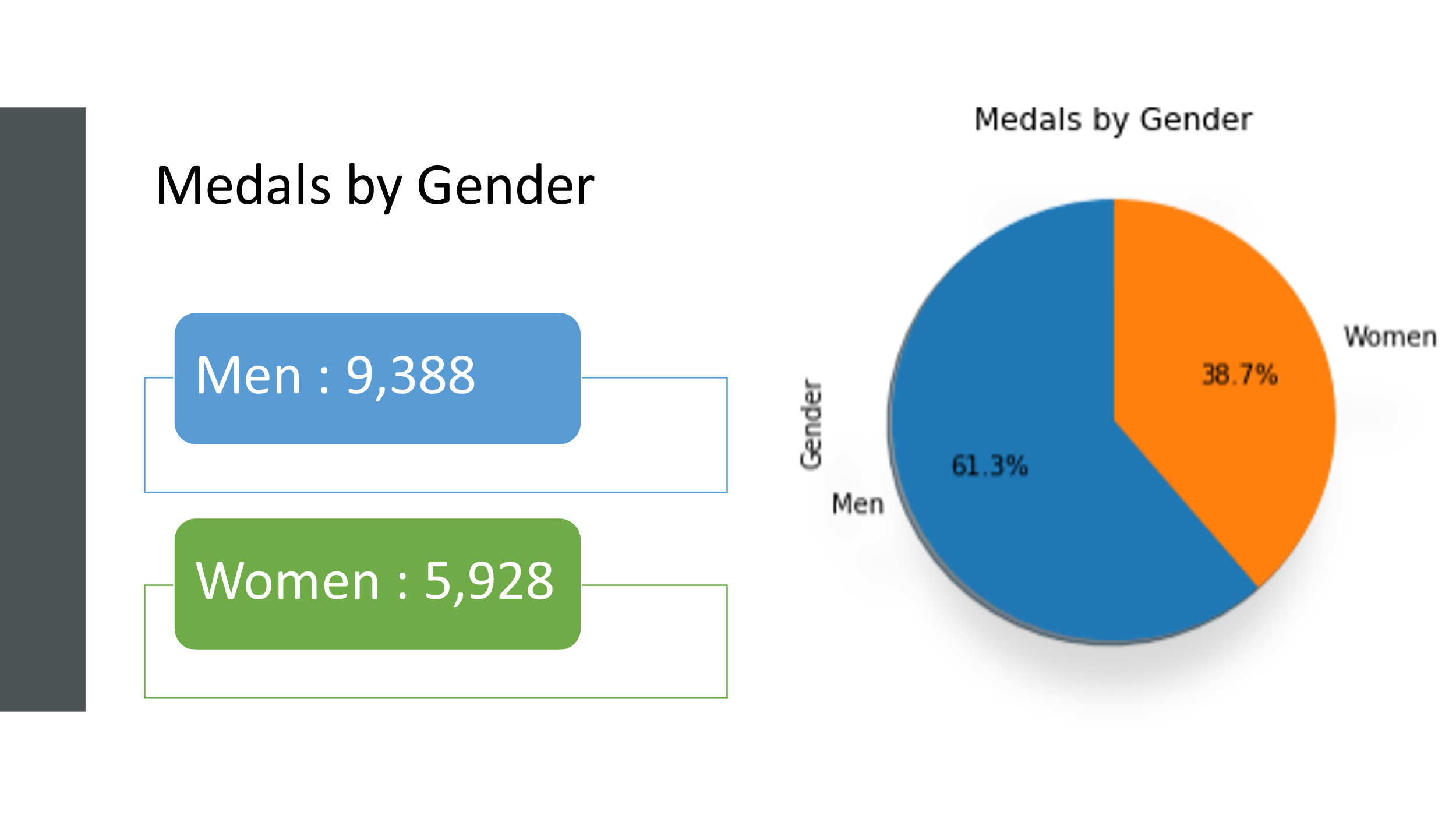
1. How many medals has each country earned?
   1. The data set included 127 countries that earned medals during the data period examined. The top medal earners are the United States and the Soviet Union.

One important factor within our data was the total medal counts by country as we were looking to see if there is a correlation between medal count and a country’s GDP. Naturally this was a great starting point as we began to explore our data. Upon our initial analysis we discovered that there **may** be a positive correlation as we noticed some of the larger countries fell within the top 10 by medal count shown below..

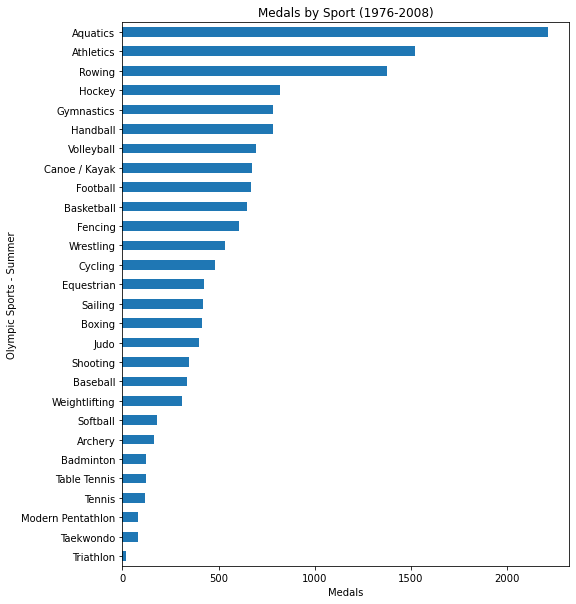


1. How many medals were won by gender:

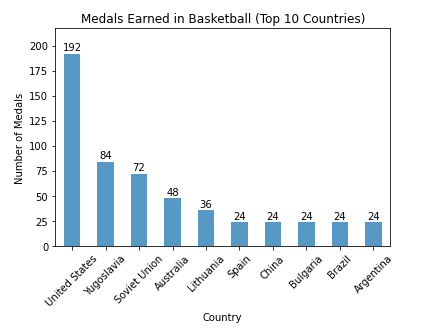
We continued to explore our data further answering other questions that relate to medal count such as total count my gender, and sporting events.



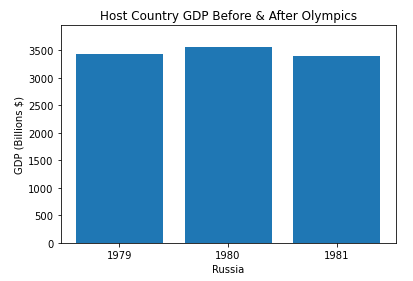
1. What are the top medal earning sports events?
   1. Aquatics by far earns the most medals due to the large number of events (26).



1. What sports do countries perform best in?
   1. As this varies across the different sports, the plot below is an example using basketball.



1. Does a country’s GDP change based on hosting the Olympics?
   1. Most countries did show some increase in GDP the year following the Olympics. Russia and Spain were exceptions. As an example, , see Russia below:



# Hypothesis & Correlation:

Our hypothesis was that a country’s GDP does positively correlate with the number of medals won. That hypothesis was proven true as you can see in the r-value of .61 and the linear regression plot below:

Chart, scatter chart

Description automatically generated