2020 Summer Olympics Medal Count

Introduction

'Sports has the power to change the world.' – Nelson Mandela

The Olympics is the largest sporting event and is presented on a global scale. With it comes excitement, determination, unity, and a lot of data. Surely this data could shed insight on such a facsinating competition. Questions you may ask yourself are why do some countries perform better than others? Is it possible to predict how succesful a given country will be? To frame a very specific question, what will be the overall country medal count for the 2020 summer Olympics in Tokyo, Japan? This is the very question we set out to answer with this analysis.

Understanding the driving forces behind the success of a country at the Olympics is a well studied question. As outlined in a wikipedia article on the Olympic medal count (Wikipedia), there has been several articles and research papers tackling the problem of predicting a country's performance at the Olympics. This past work has noted relevant driving forces to be the population, GDP, and past performance of a country. Another interesting claim is the relationship on a country's performance if they are the host country. To investigate further and tackle this problem lets first get our hands dirty with some data.

The Data

We have relied on data sources collected by the kind open source community, primarily on kaggle.com. Our data sources capture information on all of the athletes who performed in the olympics for the last 120 years. Relevant variables include the country they performed for, the year, the sport they competed in, the city the olympics took place in, and if they won a medal. A glimpse of this raw data is as follows:

Table 1. Raw Data of the Individual Athlete Performance Overtime

ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event	Medal
1	A Dijiang	M	24	180	80	China	CHN	1992 Summer	1992	Summer	Barcelona	Basketball	Basketball Men's Basketball	NA
2	A Lamusi	M	23	170	60	China	CHN	2012 Summer	2012	Summer	London	Judo	Judo Men's Extra-Lightweight	NA
3	Gunnar Nielsen Aaby	M	24	NA	NA	Denmark	DEN	1920 Summer	1920	Summer	Antwerpen	Football	Football Men's Football	NA
4	Edgar Lindenau Aabye	M	34	NA	NA	Denmark/Sweden	DEN	1900 Summer	1900	Summer	Paris	Tug-Of-War	Tug-Of-War Men's Tug-Of-War	Gold
5	Christine Jacoba Aaftink	F	21	185	82	Netherlands	NED	1988 Winter	1988	Winter	Calgary	Speed Skating	Speed Skating Women's 500 metres	NA
5	Christine Jacoba Aaftink	F	21	185	82	Netherlands	NED	1988 Winter	1988	Winter	Calgary	Speed Skating	Speed Skating Women's 1,000 metres	NA

The total number of medals for each country by year was wrangled from the raw data. The team variable was not always related to country so we had to reference the NOC (National Olympic Committee) codes to get the corresponding region each athlete competed for. The regions were converted to the standard ISO3 country code which was used as a primary key through out the analysis. Additional challenges included adjusting the medal count for team events which should only account as 1 medal on the country medal table yet counted as several medals in the raw data depending on the number of athletes on a team. After taking this into account we were able to join the GDP and Population data from our other data sources. The past performance feature was manufactured using two lag variables on the total medal counts for each country. Lastly, a home court advantage feature was calculated using the city variable in the raw data. For your veiwing pleasure, the cleaned data is tabulated below.

Table 2. Clean Data of the Olympic Medal Count by Country and Year

City	ISO3	country	country_num	year	population	GDP_USD	gdp_per_capita	home_adv	tot_gold	tot_silver	tot_bronze	tot_medals	last_medals
Mexico City	AFG	Afghanistan	1	1968	10604346	1373333367	129	0	0	0	0	0	0
Munich	AFG	Afghanistan	1	1972	11721940	1595555476	136	0	0	0	0	0	0
Moskva	AFG	Afghanistan	1	1980	13248370	3641723322	274	0	0	0	0	0	0
Athina	AFG	Afghanistan	1	2004	24118979	5285465686	219	0	0	0	0	0	0
Beijing	AFG	Afghanistan	1	2008	27294031	10190529882	373	0	0	0	1	1	0
London	AFG	Afghanistan	1	2012	30696958	20536542737	669	0	0	0	1	1	1

Data Exploration

Using the cleaned data, we were able to explore how effective certain features were in predicting the medal count. It is quite evident from Figure 1 and Figure 2 that for the best performing countries in 2016 they need to have a large population and a large GDP.

Figure 1. Population Correlation in 2016

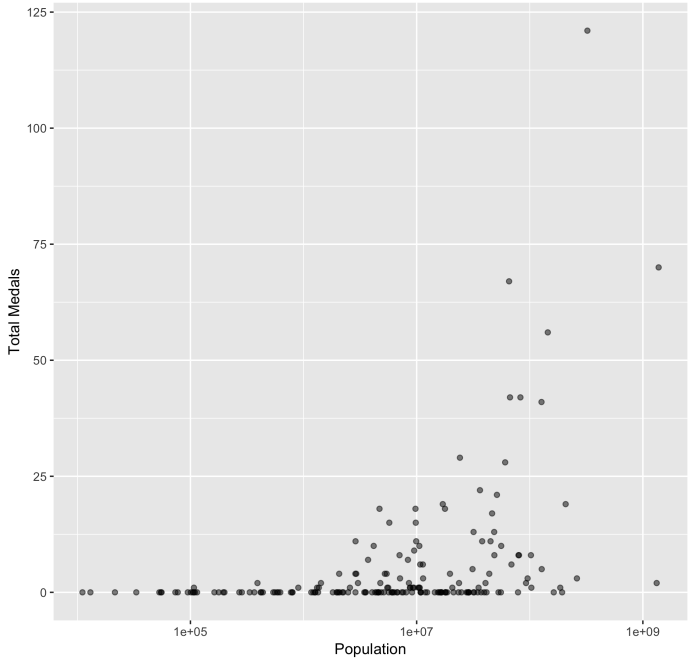
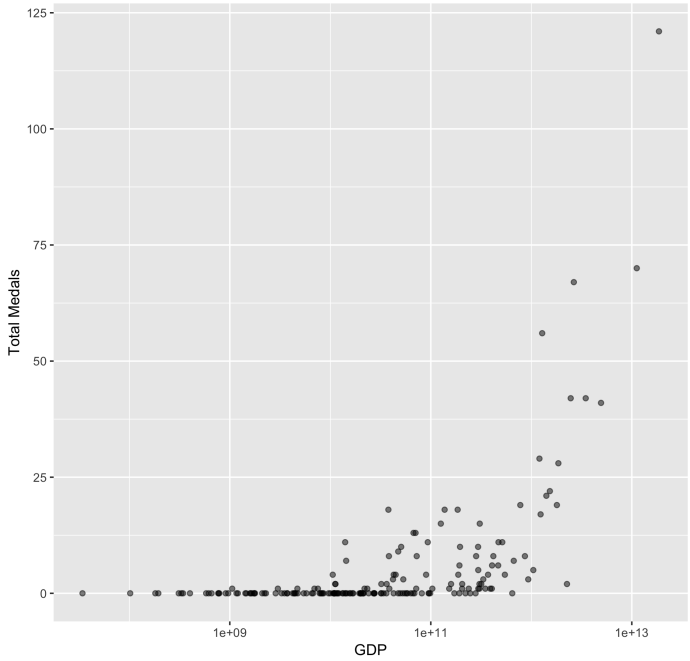
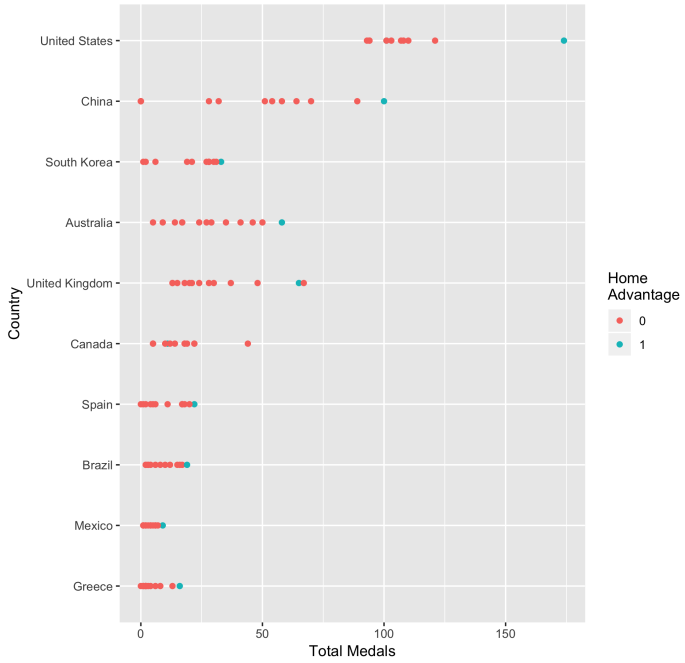


Figure 2. GDP Correlation in 2016



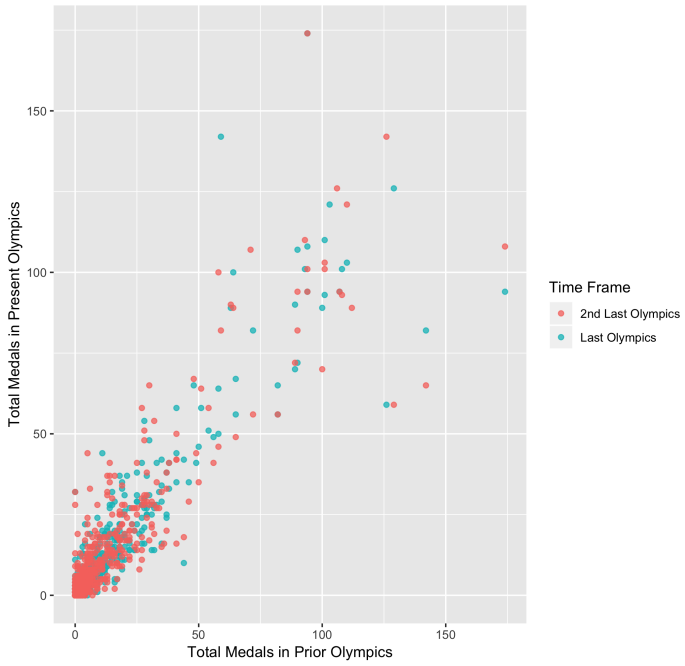
These features certainly carry some predictive power, but there is definitely more to the story. Let take a look at the performance of countries who hosted the olympics at home. Figure 3. illusrates exactly this. In almost all countries, if they were hosting the event they outperformed all the times that they competed not at home.

Figure 3. Summer Olympics Home Country Advantage



Lastly, to make predictions going forward there is no surprise that we will have to consider the historical performance. Figure 4 below shows a clear correlation between a country's performance compared to how they performed in the previous two Olympics.

Figure 4. Past Performance Correlation



The features explored all proove to carry predictive power. From here we can build our prediction model.

Analysis

Methodology

We have used DecissionTree from scikit learn , as our target data is numeric so we have used regressor.

We have worked with the whole set of data in three parts . Our first part of data we have used for training . The second part is been used for validation . The final part was only used for prediction.

Implementation

The implementation process has two stages :

1.Classiffier training and validation.
2.The prediction stage.

During the first stage the classiffier was trained on the training data . This was done in python script in Atom (titled : DecissionTree.py). The features which were chosen to train the data was extraploted from data exploration. Our features were last two years performances, home advantage , population and GDP of each country .

As we had three targets (count of Gold , count of Silver , count of Bronze) we had trained our model with the three specific targets and

predicted with three specific targets. Once we trained the data we we used our model to predict for only 2016 Summer Olympic to validate the performance of our model . We have produced a dashbord of Top 10 countries from our prediction and original record to show the accuracy .

After which we tested on 20% of the original test data which we kept separated to find the accuracy scores .

Below is the comparison of 2016 Predicted output vs Original record for top 10 countries.

Table 3: Top 10 countries of Predicted Olympic 2016 (Validation)

country	Gold	Silver	Bronze	Total
United States	46	24	35	105
China	38	39	24	101
United Kingdom	17	17	45	79
Russia	24	21	28	73
Germany	11	19	14	44
France	10	9	18	37
Australia	7	11	18	36
Italy	9	12	11	32
South Korea	9	8	10	27
Brazil	12	5	9	26

Table 4 : Top 10 countries of Original Olympic 2016

country	Gold	Silver	Bronze	Total
United States	46	37	38	121
China	26	18	26	70
United Kingdom	27	23	17	67
Russia	19	17	20	56
France	10	18	14	42
Germany	17	10	15	42
Japan	12	8	21	41
Australia	8	11	10	29
Italy	8	12	8	28
Canada	4	3	15	22

From validation we found that we are able to predict 8/10 countries in top 10 and also first four countries in same order . The number of medal count does not exactly match for all but the mismatch is not extreme . We calculated the score by using scikit learn model score and we found our model is giving 65% to 70% accuracy .

In our next stage , we took fresh data of 2020 Olympic and predict on that using our trained model. Our prediction cycle ran for three different targets on the 2020 data and we merged the three predictions into one data , showing which country will win how many Gold , Silver , Bronze , along with their population , home advantage , GDP.

The is our raw data output which we formatted to make the dashboard of Country and medal count for Gold, Silver , Bronze and Total by running Rscript .

Finally, we produce the final prediction chart of number of Gold, Silver , Bronze and Total medals for each country. Based on the chart we produced our visualisations .

Results

Below is the final prediction table for the 2020 Summer Olympics of the Top 10 countries and a visualization of these results.

Table

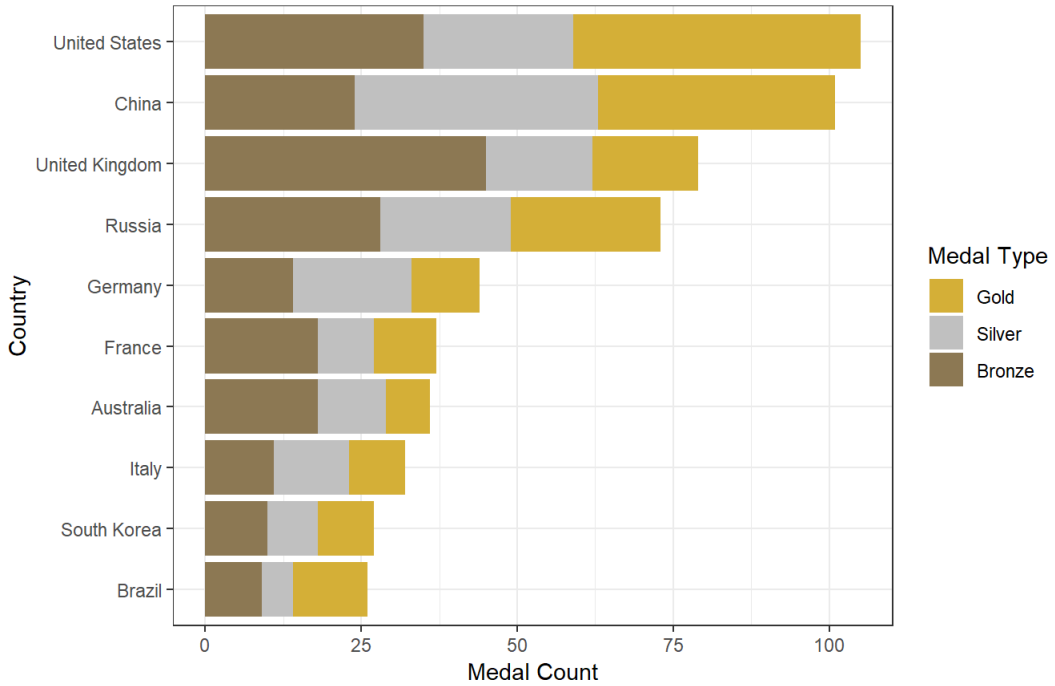
Table 5: Prediction 2020 Summer Olympic (Top 10).

country	Gold	Silver	Bronze	Total
United States	46	24	35	105
China	38	39	24	101
United Kingdom	17	17	45	79
Russia	24	21	28	73
Germany	11	19	14	44
France	10	9	18	37
Australia	7	11	18	36
Italy	9	12	11	32
South Korea	9	8	10	27
Brazil	12	5	9	26

Note: The full table of all countries participating in 2020 Olympic Summer is in the Appendix.

Visualisation

Figure 5.
Medal Count Prediction for the 2020 Summer Olympics



Conclusion

Predicting Olympic medal count is exciting and comes with many challenges. While doing our analysis we found that performance of athletes enhances if they play in home ground .Also the population and GDP of country matters in building athletes .Prediction of game results encompasses many dependencies , our EDA gave us the features which we used to train our model , but we believe there are other features which will make the prediction more accurate .

Our prediction accuracy is 65% to 70% , this can be improved if the multicollinearity of data is handled . We would also like to try other regression model instead of decission tree , like OSL .

Appendix

2020 Olympic Summer prediction (All countries):

Table 6: Prediction 2020 Summer Olympic.

country	Gold	Silver	Bronze	Total
United States	46	28	29	103
United Kingdom	27	16	45	88
China	26	10	35	71
Russia	13	17	32	62
Japan	17	10	19	46
Germany	17	10	15	42
Italy	19	12	8	39
France	11	11	14	36
Australia	9	11	10	30
Sweden	10	18	1	29
Poland	2	18	6	26
Spain	4	7	15	26
South Korea	5	5	12	22
Canada	4	3	12	19
Brazil	7	5	6	18
Netherlands	7	7	4	18
Azerbaijan	8	3	6	17
Denmark	9	2	5	16
Kenya	7	5	2	14
Ukraine	3	6	4	13
Georgia	2	4	4	10
Hungary	3	5	2	10
Finland	4	3	2	9
Malaysia	2	3	4	9
Uzbekistan	2	1	6	9
Ireland	4	4	0	8
Jamaica	4	2	2	8
Kazakhstan	0	5	3	8
Lithuania	0	1	7	8
Switzerland	0	6	2	8
Argentina	1	2	4	7

country	Gold	Silver	Bronze	Total
Belarus	2	4	1	7
Czechia	1	5	1	7
Iran	2	4	1	7
South Africa	1	2	4	7
Thailand	2	4	1	7
Turkey	0	3	4	7
Belgium	0	4	2	6
Colombia	1	3	2	6
Croatia	3	2	1	6
Ethiopia	1	2	3	6
Serbia	4	2	0	6
Singapore	4	0	2	6
Greece	1	1	3	5
Indonesia	0	2	3	5
Slovenia	2	1	2	5
Bulgaria	3	1	0	4
New Zealand	0	2	2	4
Mongolia	0	0	3	3
Trinidad & Tobago	1	1	1	3
Tunisia	0	0	3	3
Vietnam	0	0	3	3
Algeria	0	1	1	2
Angola	0	1	1	2
Armenia	0	1	1	2
Bahamas	0	1	1	2
Bahrain	1	0	1	2
Cyprus	0	0	2	2
Dominican Republic	2	0	0	2
India	1	1	0	2
Latvia	1	0	1	2
Mexico	1	0	1	2
Saudi Arabia	1	0	1	2
Bhutan	0	0	1	1
Bosnia & Herzegovina	0	1	0	1

country	Gold	Silver	Bronze	Total
Botswana	0	0	1	1
CÃÂ´te dÃ¢â¬â¢Ivoire	0	0	1	1
Chile	0	0	1	1
Comoros	1	0	0	1
Congo - Brazzaville	0	0	1	1
Equatorial Guinea	1	0	0	1
Estonia	0	1	0	1
Gabon	0	0	1	1
Ghana	0	1	0	1
Grenada	0	1	0	1
Israel	0	0	1	1
Laos	1	0	0	1
Lebanon	0	0	1	1
Lesotho	0	0	1	1
Moldova	1	0	0	1
Morocco	1	0	0	1
Mozambique	0	0	1	1
Nepal	0	1	0	1
Norway	0	0	1	1
Philippines	0	0	1	1
Portugal	0	0	1	1
Qatar	0	0	1	1
Slovakia	1	0	0	1
United Arab Emirates	0	0	1	1
Afghanistan	0	0	0	0
Albania	0	0	0	0
Antigua & Barbuda	0	0	0	0
Austria	0	0	0	0
Bangladesh	0	0	0	0
Barbados	0	0	0	0
Belize	0	0	0	0
Benin	0	0	0	0
Bolivia	0	0	0	0
Brunei	0	0	0	0

country	Gold	Silver	Bronze	Total
Burkina Faso	0	0	0	0
Burundi	0	0	0	0
Cambodia	0	0	0	0
Cameroon	0	0	0	0
Cape Verde	0	0	0	0
Central African Republic	0	0	0	0
Chad	0	0	0	0
Congo - Kinshasa	0	0	0	0
Costa Rica	0	0	0	0
Dominica	0	0	0	0
Ecuador	0	0	0	0
El Salvador	0	0	0	0
Fiji	0	0	0	0
Gambia	0	0	0	0
Guatemala	0	0	0	0
Guinea	0	0	0	0
Guinea-Bissau	0	0	0	0
Guyana	0	0	0	0
Haiti	0	0	0	0
Honduras	0	0	0	0
Iceland	0	0	0	0
Iraq	0	0	0	0
Jordan	0	0	0	0
Kiribati	0	0	0	0
Kyrgyzstan	0	0	0	0
Liberia	0	0	0	0
Luxembourg	0	0	0	0
Macedonia	0	0	0	0
Madagascar	0	0	0	0
Malawi	0	0	0	0
Maldives	0	0	0	0
Mali	0	0	0	0
Malta	0	0	0	0

country	Gold	Silver	Bronze	Total
Marshall Islands	0	0	0	0
Mauritania	0	0	0	0
Mauritius	0	0	0	0
Micronesia (Federated States of)	0	0	0	0
Montenegro	0	0	0	0
Myanmar (Burma)	0	0	0	0
Namibia	0	0	0	0
Nauru	0	0	0	0
Nicaragua	0	0	0	0
Niger	0	0	0	0
Nigeria	0	0	0	0
Oman	0	0	0	0
Palau	0	0	0	0
Panama	0	0	0	0
Papua New Guinea	0	0	0	0
Paraguay	0	0	0	0
Peru	0	0	0	0
Romania	0	0	0	0
Rwanda	0	0	0	0
SÃÂ£o TomÃÂ© & PrÃÂncipe	0	0	0	0
Samoa	0	0	0	0
San Marino	0	0	0	0
Senegal	0	0	0	0
Seychelles	0	0	0	0
Sierra Leone	0	0	0	0
Solomon Islands	0	0	0	0
Sri Lanka	0	0	0	0
St. Kitts & Nevis	0	0	0	0
St. Lucia	0	0	0	0
St. Vincent & Grenadines	0	0	0	0
Sudan	0	0	0	0
Suriname	0	0	0	0
Swaziland	0	0	0	0
Tajikistan	0	0	0	0

country	Gold	Silver	Bronze	Total
Tanzania	0	0	0	0
Timor-Leste	0	0	0	0
Togo	0	0	0	0
Tonga	0	0	0	0
Turkmenistan	0	0	0	0
Tuvalu	0	0	0	0
Uganda	0	0	0	0
Uruguay	0	0	0	0
Vanuatu	0	0	0	0
Yemen	0	0	0	0
Zambia	0	0	0	0
Zimbabwe	0	0	0	0