Alfred Custodio

Alisha Hill-McElroy

Kidist Gebremedhin

Lucas da Silva

Introduction

With everything our society has experienced in the last 18 months, from global pandemic to social distancing and loss of loved ones, our group wanted to study a subject that would be just the opposite of that. We wanted a subject that focused on bringing communities together by celebrating accomplishments of other individuals and their countries and also brought a sense of unity back to our daily lives. We couldn't think of any other activity that embraces these aspects more than the Olympic Games. It is even part of the Olympic Charter "…promoting a peaceful society concerned with the preservation of human dignity." But, how do we explore data science in the Olympics?

We believe that everyone, when thinking of the Olympic Games, will immediately try to predict how their favorite athlete, country of origin, or favorite team will perform in comparison to the previous year or in comparison to their main rival. So, is there an analytical way for a person to try to predict how a country will perform? Outside of sporting technique, strategy and training, which can go in infinite details that we can't comprehend. Is there data generally available to the public that can help us predict how successful a country will be in the Olympic Games? Does the size of a population in a country contribute to the success? Does a richer country perform better than a poorer country? These are the questions our team is trying to answer.

Hypotheses

Hypothesis 1: Population will increase the chance of achieving medals in Olympic Games.

Null Hypothesis 1: Population will not increase the chance of achieving medals in the Olympic Games.

Hypothesis 2: GDP will increase the chance of achieving medals in Olympic Games.

Null Hypothesis 2: GDP will not increase the chance of achieving medals in the Olympic Games.

Sources of Data

We looked for data sets that contained: number of medals, country, Olympic year, and type of Olympic event because we felt these data points would be sufficient for us to understand the performance of the most successful countries in the Olympics throughout history. Also, we looked for population size and GDP (gross domestic product), a common measure of a country's wealth, so that we could cross reference with Olympic data.

Our team chose three data sets: Olympic Athlete Event Analysis, Olympic Games Medals 1896 2018, Country Wise GDP from 1994 to 2017 (Kaggle.com).

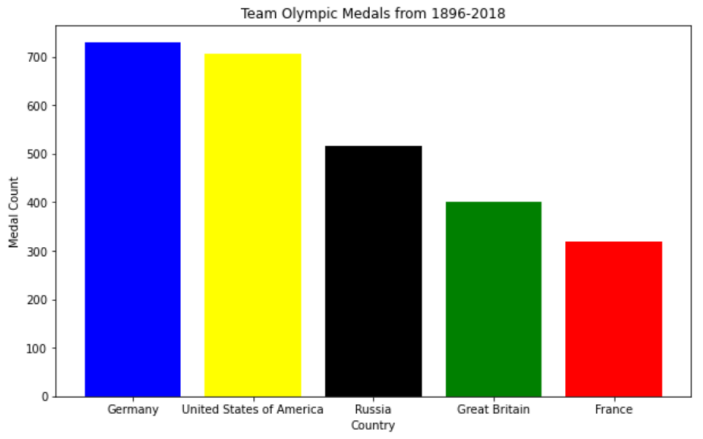
We found that there was a lot of redundant information with two of the three data sets (Olympic Athlete Event Analysis, Olympic Games Medals 1896 2018). This made it very difficult for us to clean and merge the data into one cohesive source. So we decided to omit data set Olympic Athlete Event Analysis.

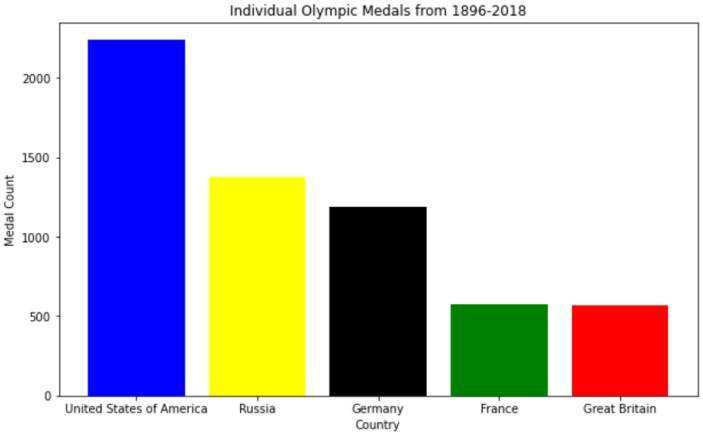
After that, we standardized country names because there were many variances between both GDP and Olympic Games Medals data sets. The team thought it would be better to follow the pattern seen in the Country Wise GDP data set.

Other adjustments were also made to combine the historical records of ROC (Russian Olympic Committee) and the USSR into Russia. As well as, Taiwan and Hong Kong into the China historical data.

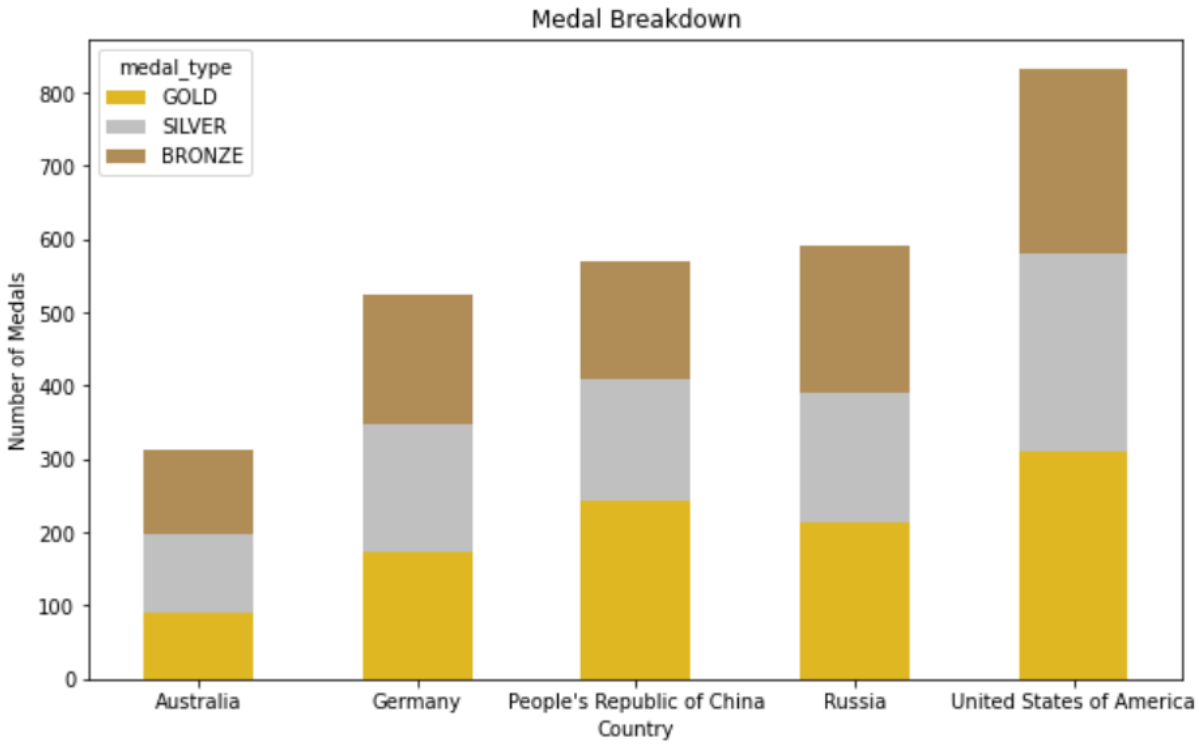
Top 5 Countries

To start understanding the data collected, the total medal count from 1896 to 2016 for all countries was tallied for individual events and team events. The top 5 countries are the same in both comparisons (with some placement variances among them), but mostly observing the USA, Germany and Russia having the best historical performances in these two categories.

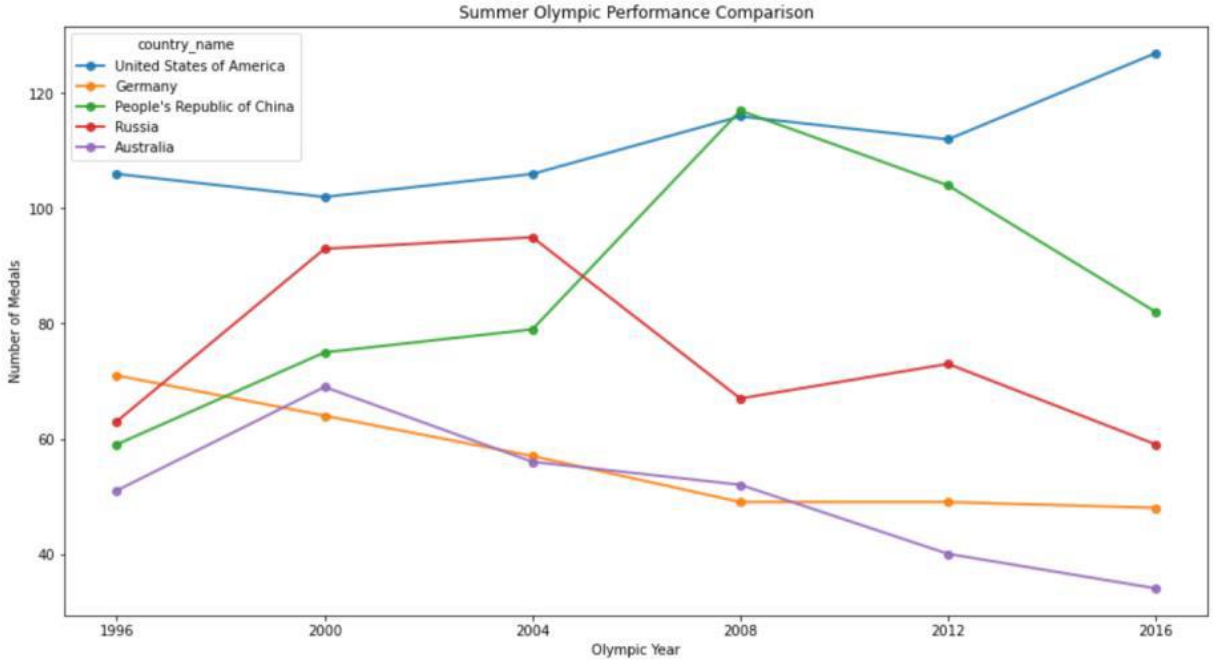


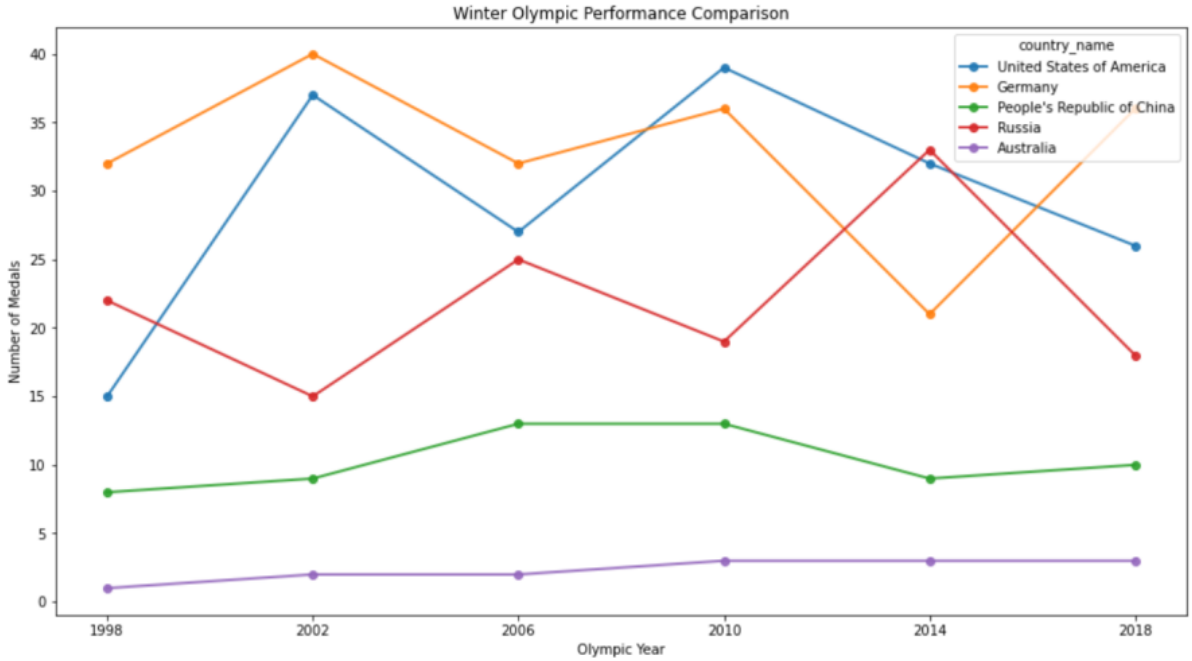


For a breakdown of the historical medal count for these countries, please reference the graph below:



The second step was to differentiate between summer and winter Olympic Games. During this analysis, we focused on the last 5 summer and winter Olympic Games. A trend in performance over that 22-year period can be better analyzed in the graph below.

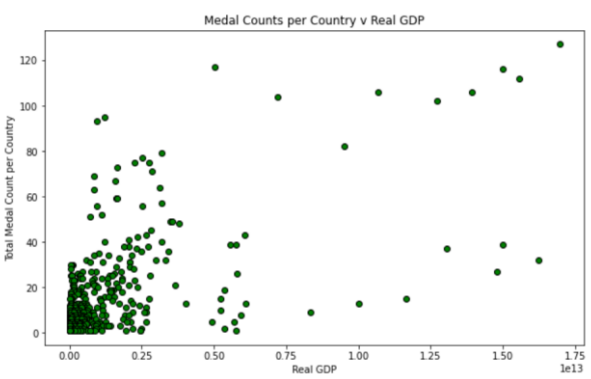


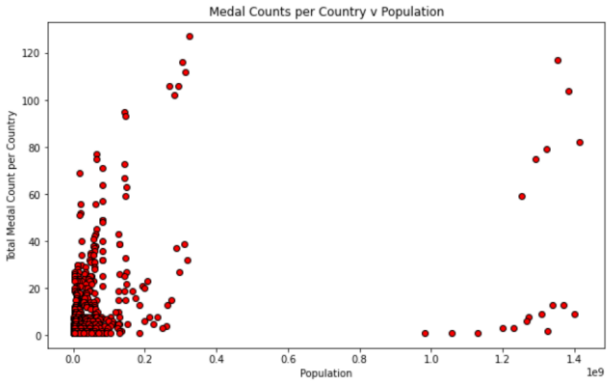


In this comparison, the USA remains mostly on top, with a second place during the 1998, 2002 and 2006 winter Olympics. USA continued to dominate every since. It is also interesting to see the entry of two new countries that historically have not stood out in the medal count. But those countries improved considerably over the last 5 Olympic Games: China and Australia.

Analysis

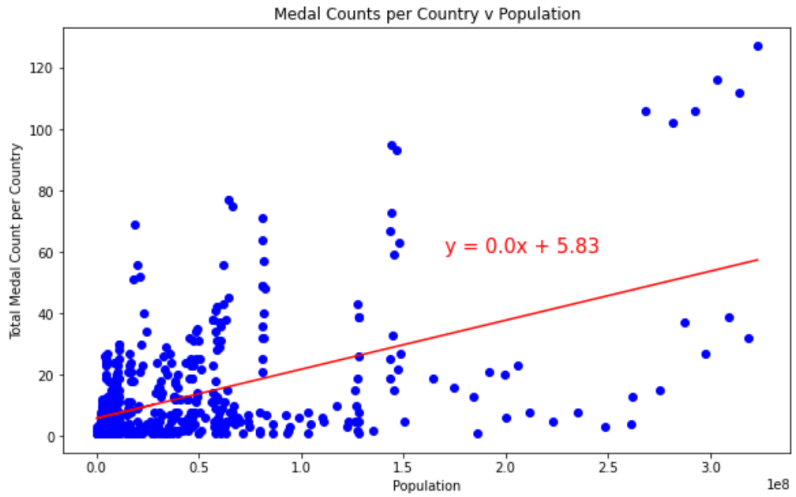
Our team plotted the initial findings and attempted to identify any trends of correlations.





When observing medal counts per country versus population, the regression model returns with an R-squared value of 0.278.

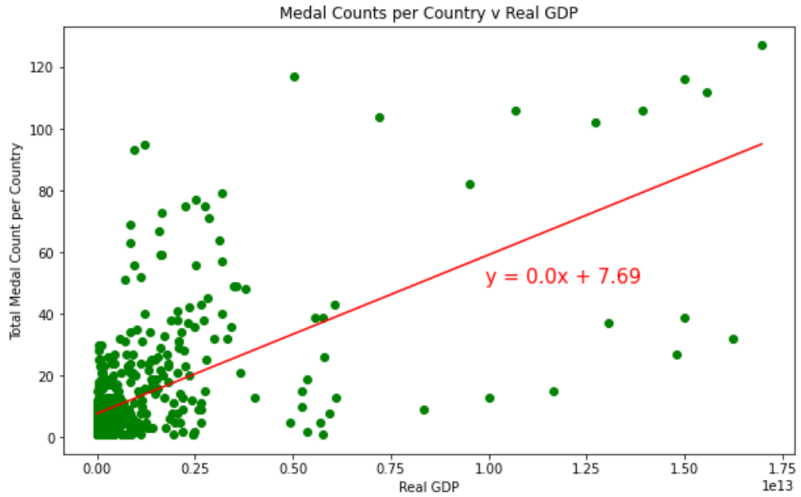
```
The r-squared is: 0.2782635297754487
```



Although there is a positive relationship, the low R-squared value indicates only a small amount of the variability can be explained. This was even after removing China and India (outliers) from the regression model since their population counts exceeded one billion and drove the R-squared value down to 0.10. There are many countries under one million population having medal counts of less than three and greater than ten, which their data points exist far from the regression line. The predictability of the medal count when considering a country's population cannot be determined due to the high unexplained variances in the model.

When observing medal counts per country versus GDP, the regression model returns with an R-squared value of 0.376.

```
The r-squared is: 0.3764755767113357
```



Although there is a positive relationship, the low R-squared value indicates only a small amount of the variability can be explained. Unlike the model in medal counts versus population, there are data points spread throughout the Real GDP range, so all data points were considered for this regression model. Again, most of the countries with their data points existing far from the regression line have a Real GDP of less than five trillion dollars. The predictability of the medal count when considering a country's Real GDP cannot be determined due to the high unexplained variances in the model.

T-Test with Population and GDP

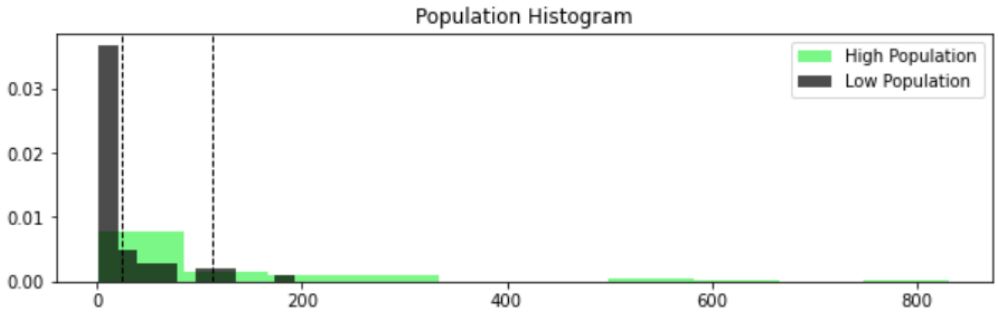
After a deeper dive into the data, it became clear that there are too many outliers in the data set and those outliers distort the statical measures the team tried to use. For example, the country of India that has the second largest population in the world normally performs poorly in the Olympics. Also, the country of Cuba that has a very small population and GDP normally performs very well in the Olympics.

A more appropriate approach to these comparisons seem to be analyzing the means of the countries with the highest populations versus the means of the lowest populations, same with thing with the GDP data, and see if any other conclusion could be found to help support our hypotheses, but even this analysis didn't seem to support our initial proposition.

<u>Population:</u>

HYPOTHESIS
These samples came from different distributions -THERE IS a SIGNIFICANT difference in means.

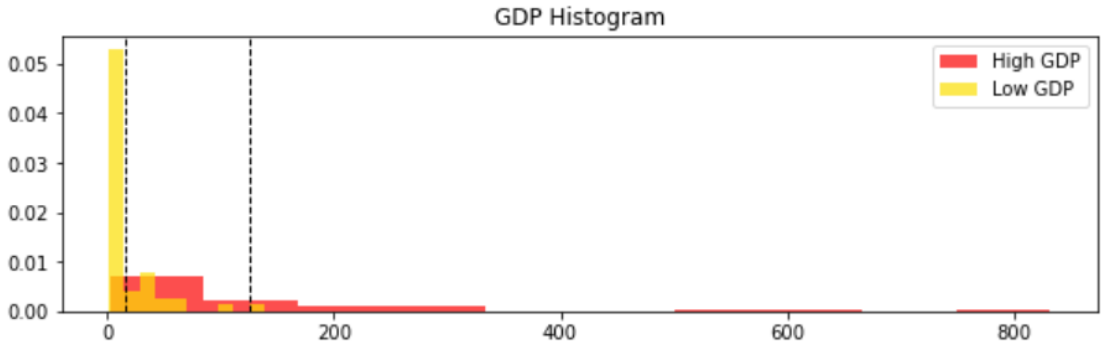
NULL HYPOTHESIS
These samples came from the SAME distribution -THERE IS NOT a difference in means.



<u>GDP:</u>

HYPOTHESIS
These samples came from different distributions -THERE IS a SIGNIFICANT difference in means.

NULL HYPOTHESIS
These samples came from the SAME distribution -THERE IS NOT a difference in means.

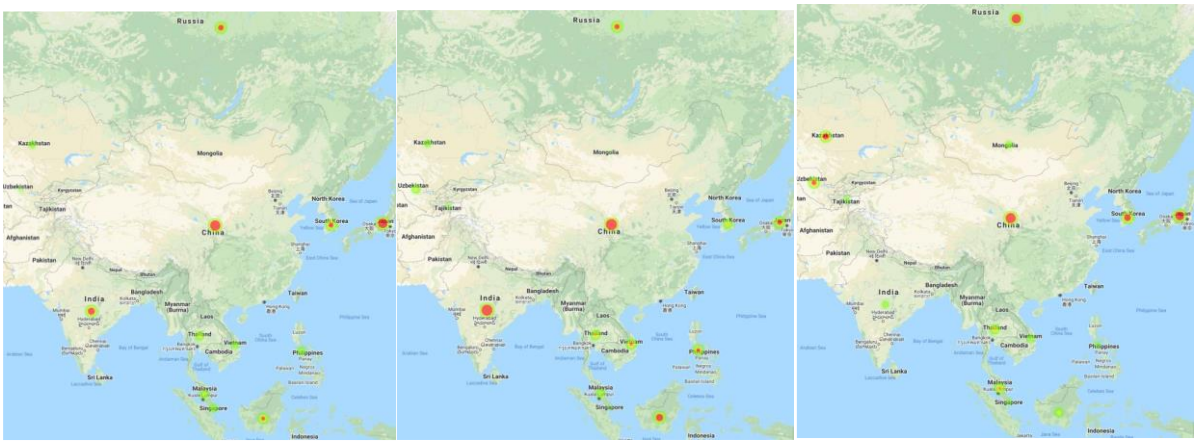


To better compare these countries' performance, GDP and population, a possible better approach would have been to sub-categorize them in their continents or geographical region, which would probably promote a more similar population size and GDP as the comparison using the heatmaps below lead us to understand:

Europe:

(Europe – GDP) (Europe – Population) (Europe – Medals Won)

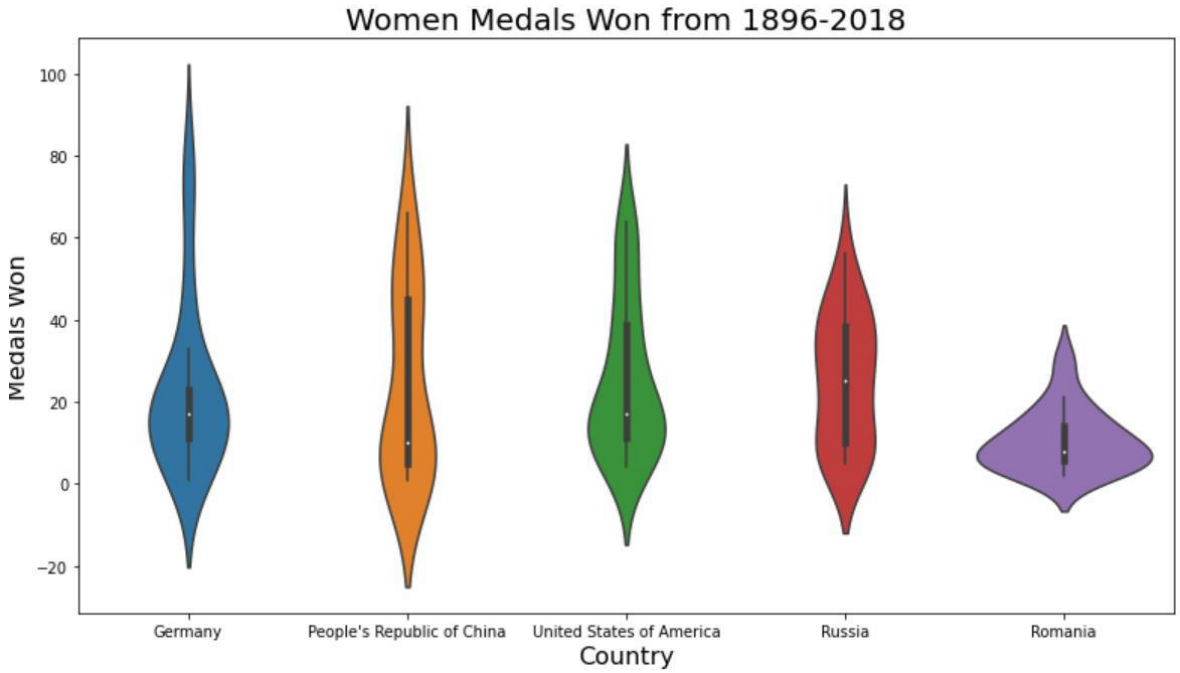
Russia/Asia:



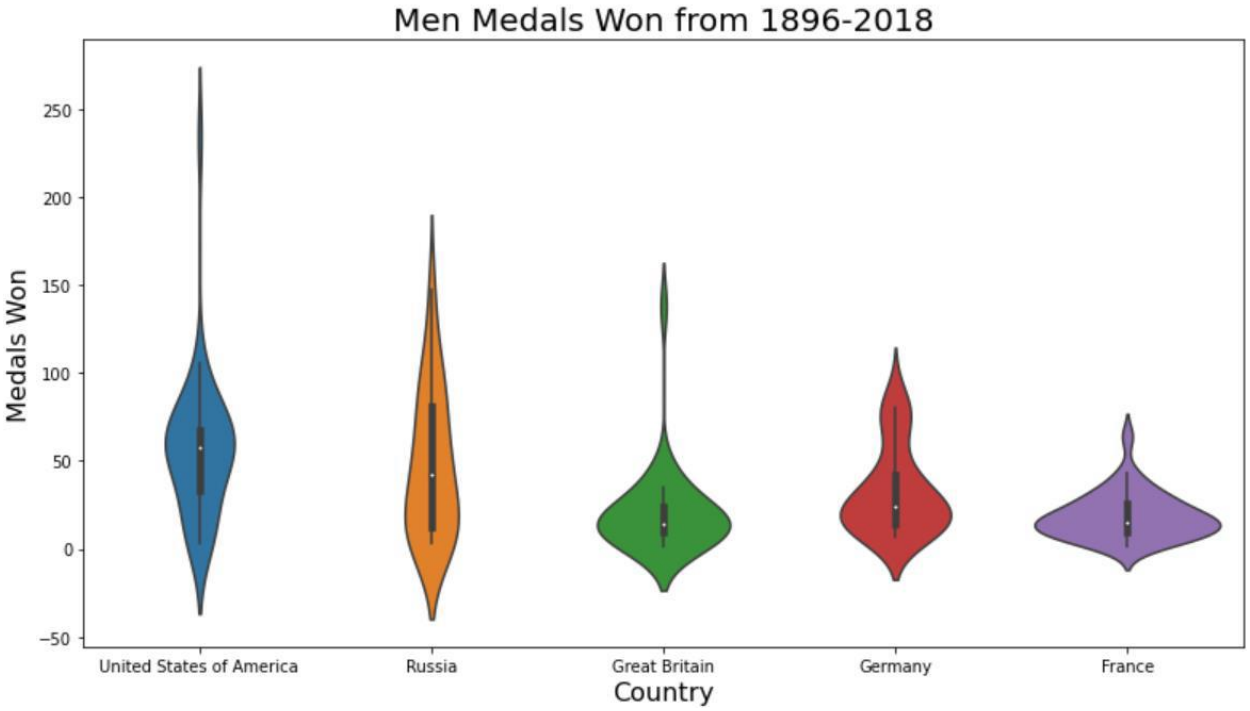
(Russia/Asia – GDP) (Russia/Asia – Population) (Russia/Asia – Medals Won)

Additional Observation

With all the data our team collected and the tools at our disposal, a comparison of countries' performance based on gender became an obvious question to analyze. Are the countries that generally perform in the top 5 of the Olympics overall also displaying the same level of performance when looking at only female athletes?



How is the performance by country if we looked at only male events?



After analyzing the gender breakdown, we observed the top 5 countries don't seem to vary much, except for the appearance of Romania that dominated gymnastics for decades. Germany takes the lead in performance by female athletes and the USA takes the lead in the performance of male athletes.

Conclusion

In conclusion, our hypothesis was proven incorrect that Olympics Games success cannot be determined or predicted by population or GDP due to the high variance and inability to explain those variances.

Limitations

If resources were unlimited and time was not an issue, the team would have looked for cleaner data. A lot of time was spent working out the discrepancies with names and medal count. The data set we chose listed group events as if they won individual medals. It also listed some double events (2-athlete teams) only won one medal. That polluted our data slightly. Additional exploration and analysis of the gender versus medals won. While attempting to figure out if wealthier countries are more likely to have women winning medals. We would have performed a comparison between North America and Europe. Lastly, a sub-group comparison of team sports versus individual sports.

References

Kaggle Notebooks

https://www.kaggle.com/themlphdstudent/country-wise-gdp-from-1994-to-2017

https://www.kaggle.com/piterfm/olympic-games-medals-19862018

https://www.kaggle.com/samruddhim/olympics-althlete-events-analysis

Google Maps API

https://could.google.com/maps-platform/