Olympic medals (1896 – 2012)

We carried out an analysis of medal trends by country for the Summer Olympic Games between 1896 and 2012. We looked at three questions in particular:

1) Who were the top performing nations across time?
2) Who performed best at new events introduced to the Games?
3) Does being a host make a statistical difference to the medal count?

Datasets

Our primary dataset was 'Olympic Sports and Medals 1896 – 2014':
https://www.kaggle.com/the-guardian/olympic-games

We used a secondary dataset of host cities and countries,
https://www.kaggle.com/piterfm/olympic-games-hosts

We used a secondary dataset of country names and IOC country codes. This was obtained by copying data from a website providing the information.

Question 1

For this question, we have visualised the data set to see the top ten performing countries for the olympic over time, by using the medals counts, with this we saw the following countries top the rank as below:

USA 4585

URS 2049

GBR 1720

FRA 1396

GER 1305

ITA 1296

AUS 1189

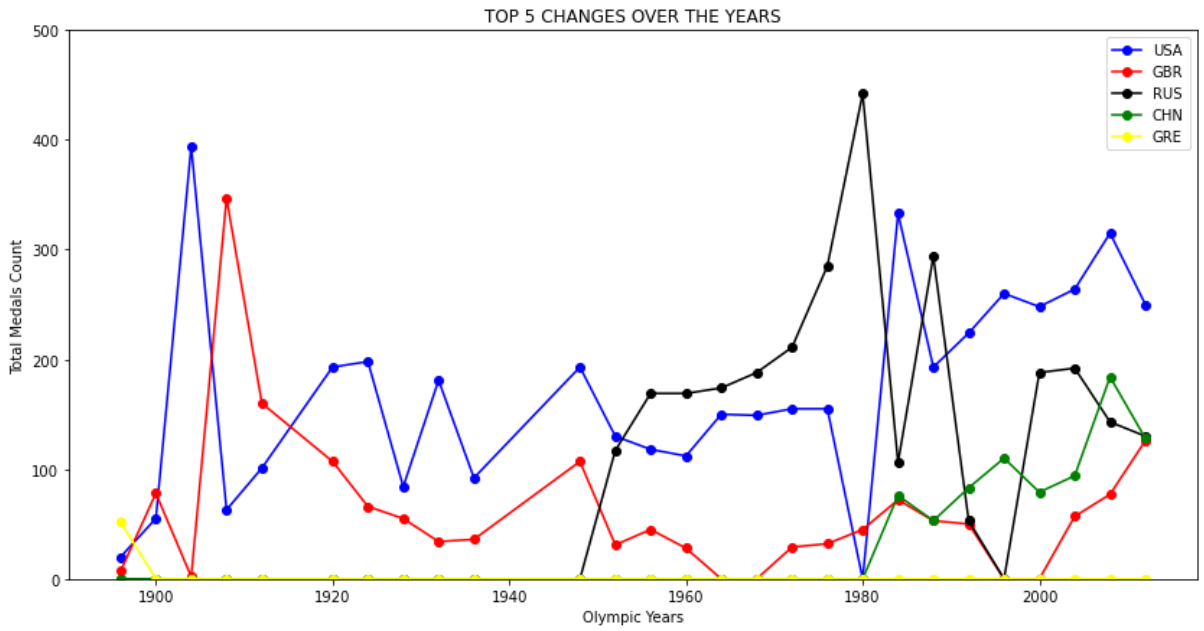
HUN 1079

SWE 1044

NED 851

The idea was to analyse the data of the Olympic games and find the top 10 according to the total medals won for each Olympic year. According to the data we found that there were about 25 different countries who appeared in the top 10 since record began (1896) to 2012. Some countries changed names, some

countries appeared once (especially) Greece, while other countries like GBR appeared all the time. From the line graph we wanted to find how the top 5 performed over time.



Looking at the correlation between medal success and time, we used the Pearson correlation statistic. This showed that there was a correlation for Russia and China, but not for the other countries.

Adjusting the data to look at years with active data for Russia (since 1952) and China (since 1984), showed no correlation for Russia, but still a correlation for China.

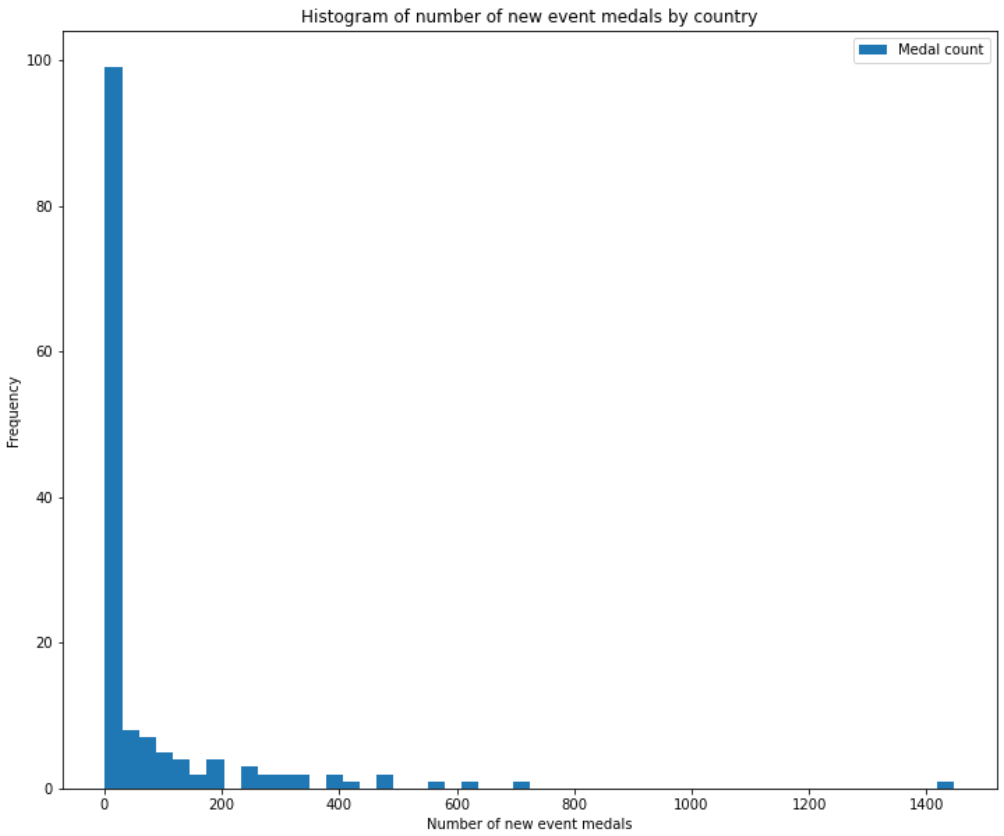
From this, we conclude that China's performance has increased over time, but the other countries examined have had no statistically significant increase or decrease in performance over time.

Country	**Correlation statistic**	**P-value**
USA	0.440101	0.0216021
UK	-0.264854	0.1818472
USSR/Russia	0.620794	0.0005503
Russia 1952 onwards	-0.149600	0.580273
China	0.737385	0.0000114
China 1984 onwards	0.735035	0.0377546
Greece	-0.332852	0.0897965

Question 2

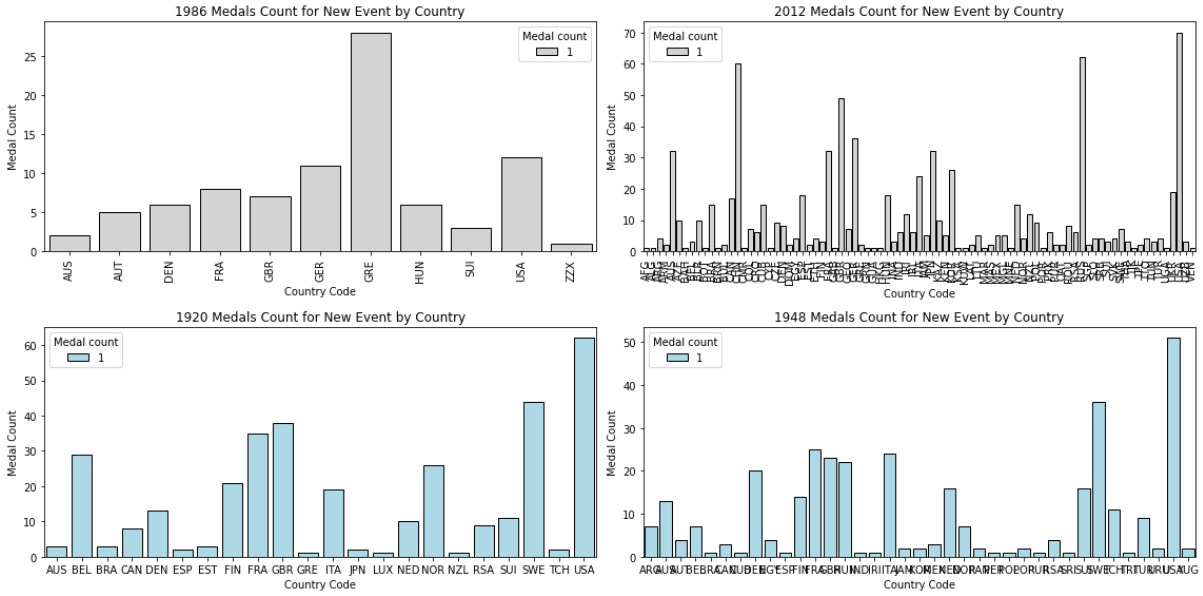
The second question seeks to investigate the performance of countries on the introduction of new events. In exploring the data it was observed that, some of the new events introduced to the Olympic games were team sports and for this reason we needed to clean the data by doing a groupby and dropping all repeated events and gender for Athletes based on keeping first event.

A Statistical analysis is carried out to look into the distribution of the medal counts by country based on new events added in each year. The histogram showed a significant skew towards small numbers of medals per country, with the mean (81 medals) being greater than the 75th percentile (75 medals), and median value being only 11 medals per country. The data shows that the USA is by far the top performer, with 1447 medals across all new events, with the next best performer, USSR/Russia having less than half that total with 706 medals. Third place, Great Britain, has 633 medals in new events.



In order to obtain a good presentation of the work, we picked specific years of interest that will help to drill down on our goal. 1986 was of particular interest since it was the first ever Olympic games, also of great interest was the years after both world wars to see if those bad times had any impact on Medal count. Finally, the maximum point of the data 2012.

The graph below shows a representation of country performance on new events. These show that, over time, more countries have won medals in new events.



Question 3

For this question, we needed the medal count for the host country in the year they hosted the Games, and for three Games before and three Games after.

This analysis was complicated because the host was given as a city, but the medals were listed by IOC country code. Use of the two secondary datasets provided the host country code for each Games.

Through manipulation of the data and dataframe merging, a complete dataset was calculated giving the number of medals awarded to host countries for their host year and three years either side. Due to country names not always matching across datasets and countries changing names (particularly Germany) some manual manipulation was required in specific cases to ensure the data was as complete as possible.

From this dataset three analyses were undertaken:

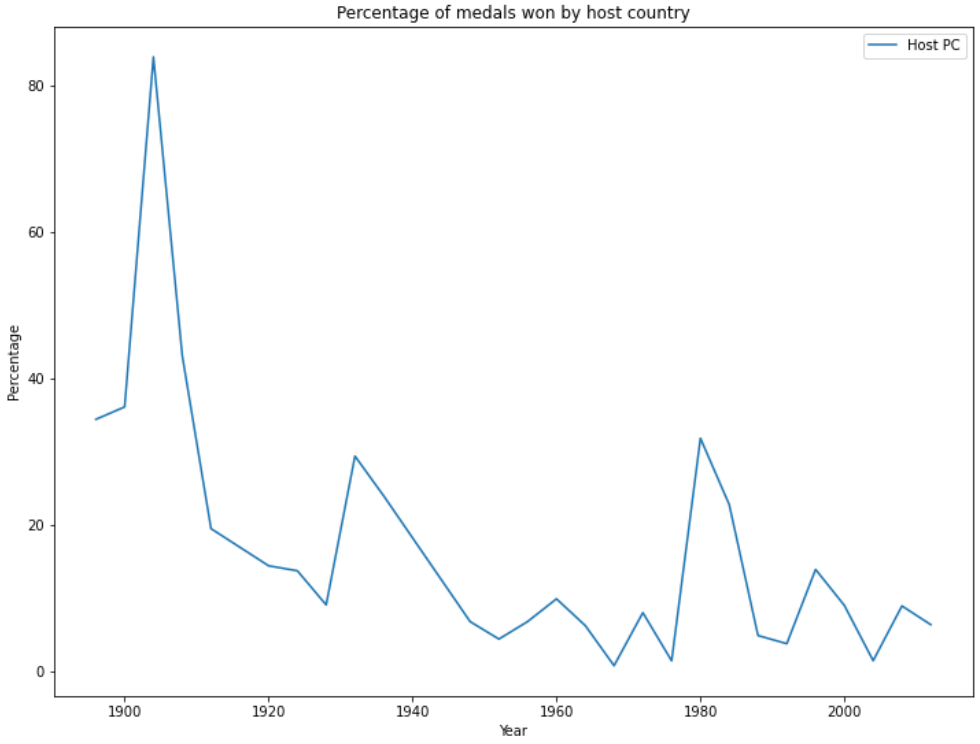
1) Visualisation of the medal counts
2) Visualisation of the percentage of medals won by the host nation in each games
3) Statistical analysis to test the null hypothesis that being a host has no impact on medal count

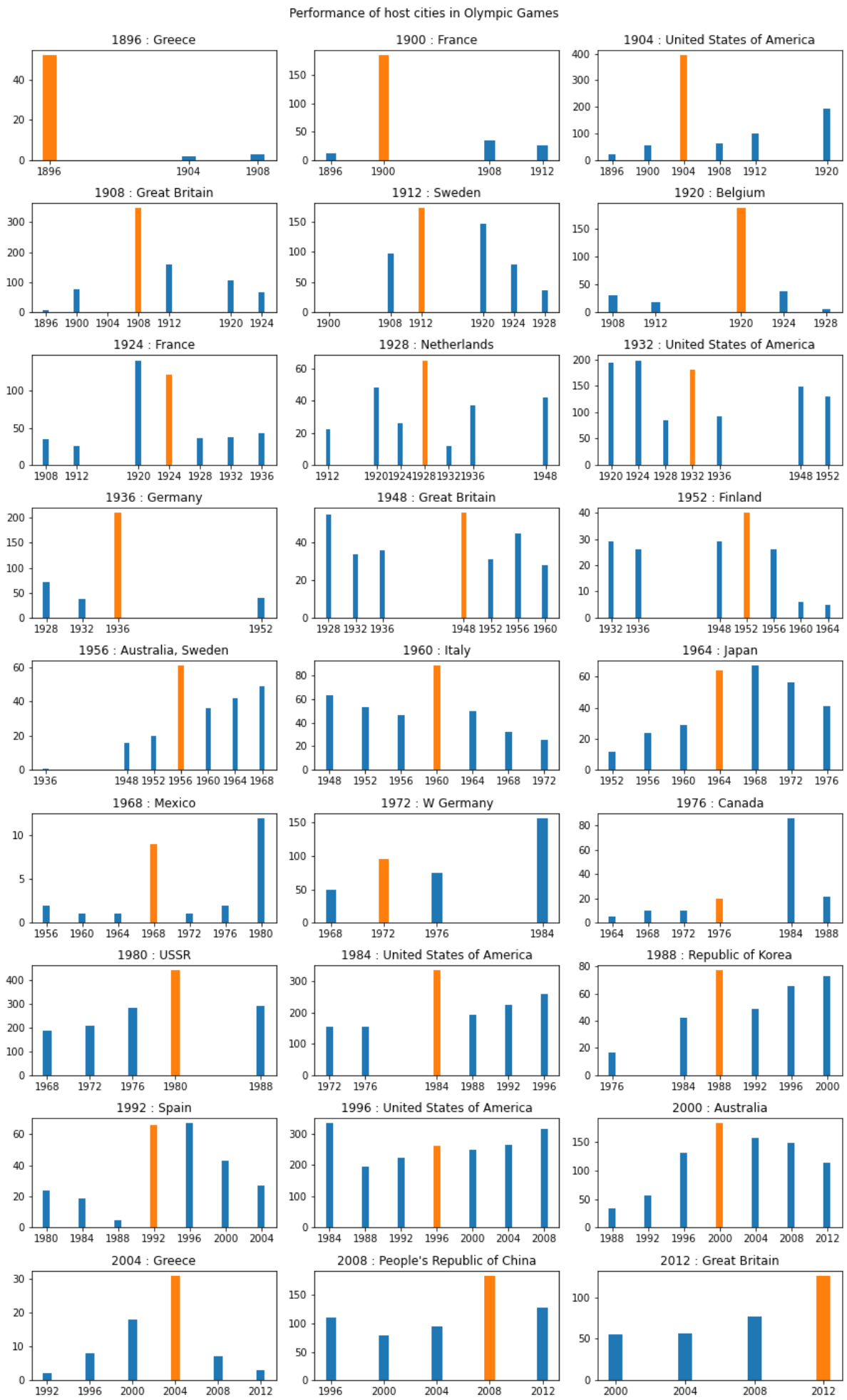
The visualisations show that, in most cases, the host nation performs better in the year it hosted than in other years, in some cases dramatically. In the first few Games, the host country performed significantly better than in other years, but this effect has decreased over time. Notably the USA (1996) had very little bonus from hosting.

Since the second world war, the percentage of medals won by the host nation has been relatively low (with the exception of USSR in 1980 when some countries did not attend the Games). This could be

interpreted as a ‘host effect’ with the ‘host bonus’ passing between countries; however more statistical analysis is required before this can be confirmed.

A Chi Squared analysis was carried out with a null hypothesis that being host made no difference to the medal count. The statistical analysis showed that for all countries, the null hypothesis was rejected (p-value < 0.05), meaning that being a host country does provide a boost to medal performance.




Performance of host cities in Olympic Games
1896 : Greece
1900 : France
1904 : United States of America
1908 : Great Britain
1912 : Sweden
1920 : Belgium
1924 : France
1928 : Netherlands
1932 : United States of America
1936 : Germany
1948 : Great Britain
1952 : Finland
1956 : Australia, Sweden
1960 : Italy
1964 : Japan
1968 : Mexico
1972 : W Germany
1976 : Canada
1980 : USSR
1984 : United States of America
1988 : Republic of Korea
1992 : Spain
1996 : United States of America
2000 : Australia
2004 : Greece
2008 : People's Republic of China
2012 : Great Britain

Conclusions

Our conclusions on the three questions posed are:

1) The top five countries represented in each Games have changed many times, but we found very little evidence of trends in time
2) Many countries won medals for new events, but the USA have by far the best performance
3) Hosting the Games does have a positive impact on medal performance in the year of hosting

Further work could be carried out to investigate further the trends in best performing countries, even when they are not host nations, and the way other countries perform below the top 5. There could be more research done on the impact of GDP growth, population size and distance travelled on the performance at each Games.