A STUDY ON THE SUMMER OLYMPIC MEDAL TALLIES FROM THE VIEWPOINT OF ETHNIC AND GEOPOLITICAL REGIONS AND A PREDICTION OF THE NEXT EDITIONS

Author: MUDIT BHARTIA (16BCE1100)

Prof. RAMESH RAGALA | SCHOOL OF COMPUTER SCIENCE AND ENGINEERING (SCSE)

Introduction

The Olympics are a prestigious global competition, with athletes all over the world giving their all, and winning becomes a matter of national pride. But does it matter where the athlete is from? Will it affect the chances of winning a medal given that the sportsperson was from North America or South East Asia? What about future editions? Is it possible to predict or forecast if a region will score more medals than the other?

This is what we set forth to determine and we do so by the power of data visualization.

Dataset

The dataset used by us in this project was from the IOC Research and Reference Service, which maintains an open source database of all the medal winners of every Summer Olympics, from the very first Edition in 1896, to the 2008 Edition.

However, the world has changed from 1896, and many countries no longer exist, and a lot more have appeared over the years. Another problem which arose was that country codes changed over the years. We needed a method to use the data without any ambiguity. We also required a method to determine what country belonged to what region. We created an index of countries, linking them to regions. We arrived at 14 regions in all – North America, Central America, South America, Western Europe, Eastern Europe, Central Europe, Northern Europe, Africa, Australia, Middle East, Indian Subcontinent, East Asia, South East Asia, Minorities and Others

Methodology

The data had to first be grouped into regions. This was done by analyzing characteristics of each and every country based on the following criterion

1) Language or Etymology
2) Land borders
3) Economic Status

This worked well for our purposes, and another dataset was created. An inner join was done on both the datasets, and the regions were established. The medal distribution was then represented per region, as an aggregate of all years. After this, medal distributions of each region per year was plotted, and from there the ratio of medals won by region per year was calculated and plotted. This was the required graph. The important feature of this graph is that it is independent of number of medals award that particular year, only depending on ratios. This required the creation of two calculated fields:

```
[Percentage of Medals] = COUNT([Number of Records])/SUM({FIXED
[Edition]:SUM([Number of Records])})*100

[Total Medals this Edition] = {FIXED [Edition]:SUM([Number of Records])}
```

This data held the key for prediction for the next few editions. By performing specific operations on this data, we could tell what ratio of medals will be earned by which region.

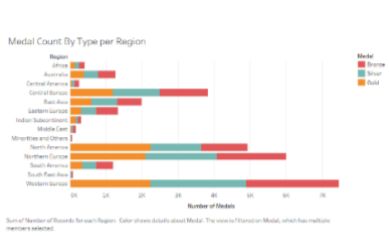
This was done in two approaches. The first was by a Python script we wrote. This script would first arrange the data into regions, determining what ratio of medals it earned in each edition. The average of fractions of medals was found, and then this represented the chances of winning a medal by that region.
The results were normalized to make them proper ratios and was plotted. This gave us the prediction of the next editions of the Olympics.

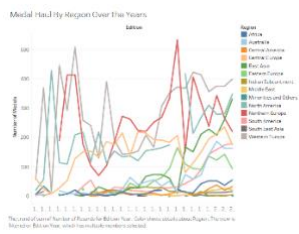
The second approach was to use the in-built forecasting feature of Tableau. This threw some errors, as it required continuous data. Thus, by using a calculated field, all intermediate years (the years between two Olympics editions) were grouped into the same edition.

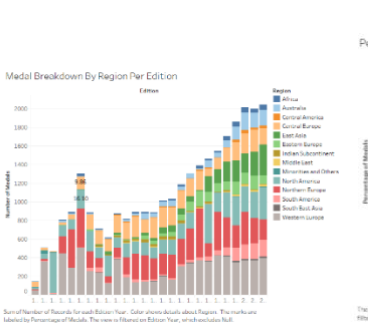
```
[Edition] = INT((year([Year])-1892)/4)
```

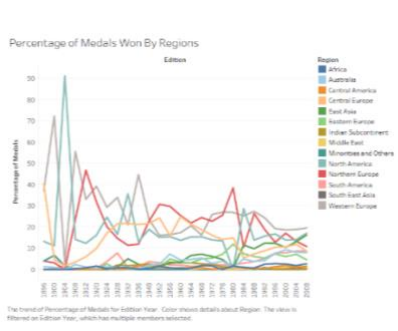
This allowed us to plot the data and use the forecasting feature.

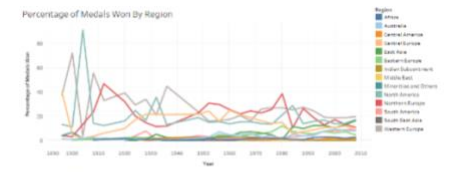
Data Visualization

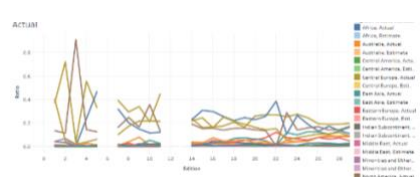


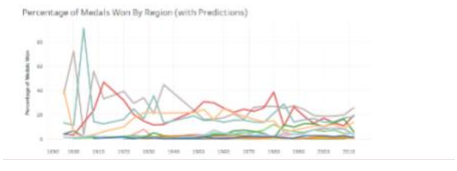


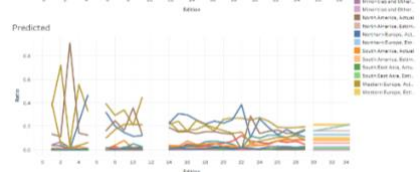












Analysis and Conclusions

First off, we notice that the distribution of medals has changed from higher ratio of medals (around 25-30%) in fewer regions to lower ratio of medals (around 10-15%) in more regions.
This means that some previously underperforming regions have been scoring medals in the past few years. This is true for regions such as Australia, East Asia, South America, Africa and Eastern Europe.
Another thing of note is the anomalies seen in the distribution of medals in 1900 and 1904.
In 1900, the Games were held in Paris, France. In those early days, countries did not send national teams. Rather, individuals would come to represent their country. Thus, the majority of players (700+) were from France itself (hence giving Western Europe a huge portion of the medals won).
The same happened in the 1904 Olympics, when the Games were held in USA. Of the 623 participants, only 34 weren't American. So, a whopping 92% of the medals were won by the Americans.

Regarding the predictions, when we performed our version of forecasting (via the Python script), it was observed that regions that were performing poorly consistently, would continue to do poorly. This included regions such as the Indian Subcontinent, South East Asia, the Middle East and Africa.
Regions which were doing poorly in the past but had performed well in the recent years were projected to bring in lower medals. This was seen for East Asia, South America and Australia.
Regions doing well in the past, and maintaining it, or slightly degrading in medal tallies were projected to gain more medals in the next edition.

The in-built forecasting option in Tableau presented a slightly different image. Of all the regions, only East Asia is predicted to make major gains in the medals tally, exceeding North America. All other regions were predicted to maintain or decrease slightly.
Given that the edition predicted was for 2012, comparing the prediction with the actual data, we can tell that the in-built function predicted more accurately than the python script we wrote.