

SQL for Data Science Capstone Project

SportsStats (Olympic Dataset – 120 years of data)

The scope of this project is to introduce information to marketing companies looking to direct their advertising and merchandise to specific audiences. This information is separated into 3 main buckets to help advertising efforts: Gender, Athlete and Sport, organized by Country.

With the understanding that performance and success drives popularity, by the end of this report marketing companies will have data delivering them popular sports and popular athletes in countries that have seen Olympic success.

My original hypothesis on this was that large countries like USA, Russia, and China would lead in medals for a large amount of sports and that smaller countries would have few medals and focus heavily on 1 or 2 sports. Also, I figured that the most successful athletes would be the most popular. So I instantly dug into which athletes were the most successful and grouped them by how many medals were won, as follows:

```
SELECT count(name) as #_of_Athletes,  
       medals as Medals_won  
FROM profile  
group by medals  
order by medals desc
```

The top 4 results of this returned:

| #_of_Ahletes | Medals_Won |
|--------------|------------|
| 3 | 8 |
| 8 | 7 |
| 25 | 6 |
| 57 | 5 |

I was then curious about breaking out medals won by NOC:

```
SELECT NOC,  
       count(medal) as Medals  
FROM NOC_Medals  
GROUP BY NOC  
ORDER BY count(medal) desc
```

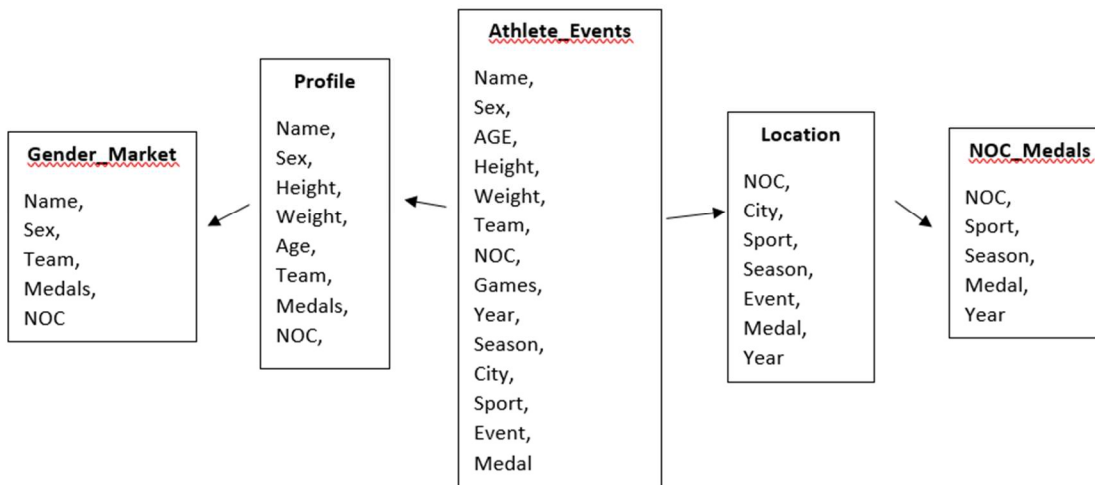
The top 4 results of this returned:

| NOC | Medals |
|-----|--------|
| USA | 5637 |
| URS | 2503 |
| GER | 2165 |
| GBR | 2068 |

The results to both had some surprising outcomes. I thought China would be in the top 4 NOCs. I also never imagined the number would be

so high for athletes who have won 5 and 6 medals. A stellar accomplishment!

I also created a rough ERD to help with the initial information grab, of which I expected to clean up later:



As I dug into the data, I was surprised to see smaller countries like GBR and GER toward the top of the board. A conclusion I have now drawn is that modernization plays a big role on vast Olympic success. From here my hypothesis evolved to include countries like Germany, Great Britain, USA and Russia to have a large marketing opportunity across the board, and not so technologically advanced countries like Tonga, Bermuda, and Macedonia to have very precise markets to advertise toward. I then filtered out medals won more than 20 years ago, to keep the data relevant.

To help me tell the story of successful marketing areas, I manipulated the 'medals' column into a count of medals earned by athletes, and by NOC.

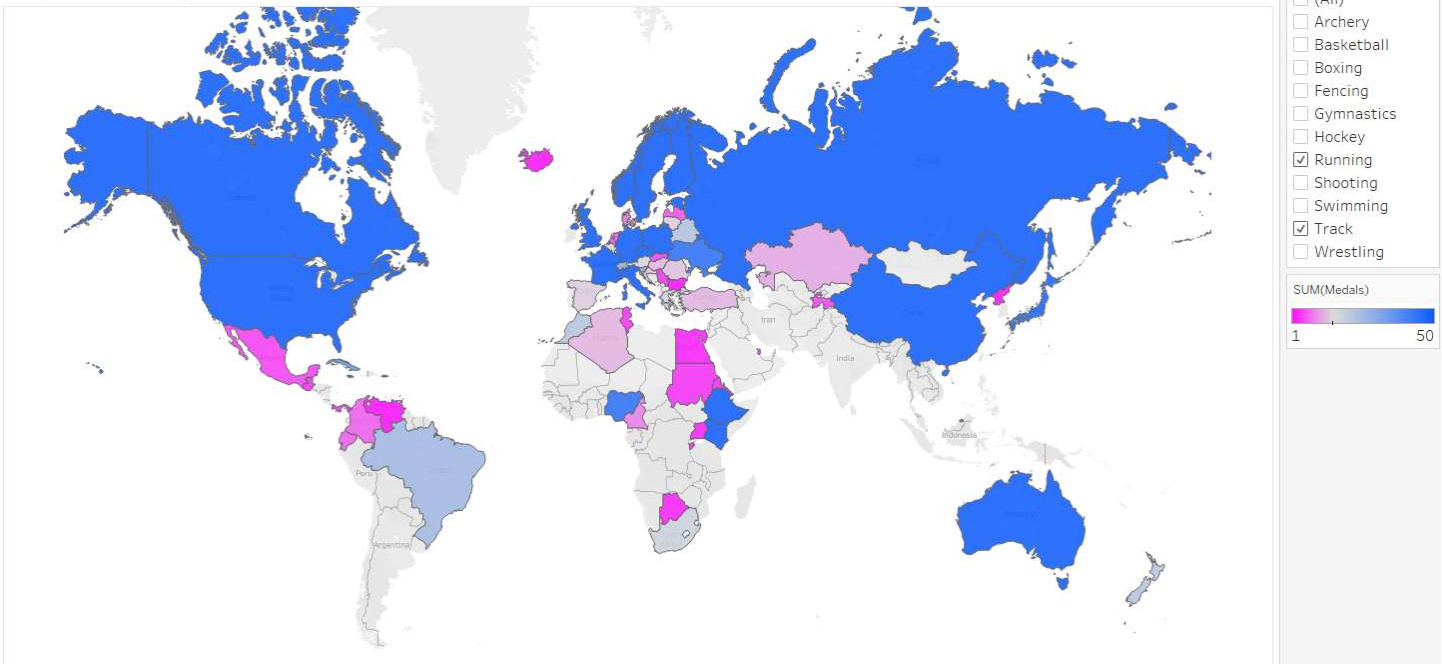
Then I sorted my data by NOC and plugged-in medals and athlete breakdowns to bring me to a final data set, of which I created interactive maps to get to the final presentation.

Presentation

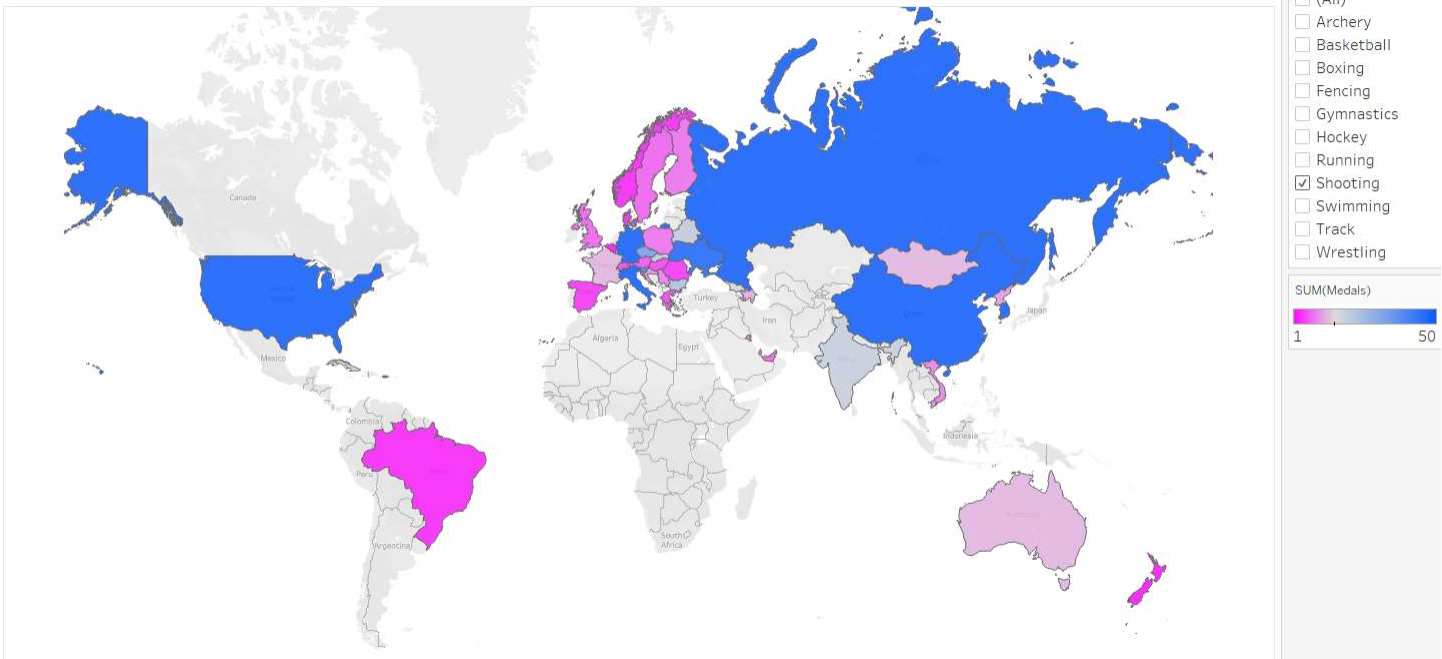
For my Presentation, I will be showing an interactive map to marketing agencies that would be interested in marketing merchandise to areas showing strong success. The greater the success in an event, the better audience to market products pertaining to that event to.

Here is a screenshot of the interactive map. This allows countries to be filtered by how much success they've had in a diverse array of Olympic sports. I have combined similar events, such as 200 meter dash and 100 meter dash into buckets. This cleans up the data to display the overall sport of track/ running, for marketing agencies to appeal to the sport in its entirety, not just 1 facet.

Map of Country Medals



Map of Country Medals



Recommendations

Use this map to determine popular sports by country. For example, it wouldn't make much sense to focus a marketing campaign in Mexico for Olympic Rifles for shooting events, delivering marketing efforts in Track and Running could make more sense.