What affects the Athletes' Performance in the Olympic Games?

Yilan Ye

Info640

11.20.2019

Introduction

The modern Olympic Games are the leading international sporting events held every four years. Thousands of athletes from more than 200 countries around the world participate in a variety of sporting events. Every country desires an excellent athlete performance and wins more medals in the Olympic Games to show the nation's spirit and power. As an audience, I love to watch the Olympic Games and to see how athletes break records and human limits. However, looking at the results of the national metal performance in the Olympics, I found that the individual strength may create a miracle in one competition. Still, there are more external factors that contribute to the overall athletes' performance of a country. Inspired by the research paper "Going for Gold Medals: Factors affecting Olympic Performance" by Jayantha K and Ubayachandra E. G, I would like to check out whether the social factors such as population, GDP per capita, host city advantage will influence the athletes' performance in the Olympic Games.

Hypothesis

1. The population of a country is positively correlated with the athletes' performance.
2. GDP per capita is positive affects the Olympic performance of a county.
3. The host city advantage has a positive influence on the athlete's performance.

If all of the three hypotheses are true, we can receive a solid conclusion that the Olympics performance of the athletes and the comprehensive national factors are tightly related. A country with a larger population, higher GDP per capita, and the host city right could result in a better Olympics performance.

Data Analysis

This data analysis will be the predictive data analysis, focusing on whether there is a correlation between the athletes' performance of a country in the Olympic Games and its population, GDP per capita, or the host city advantage.

Dataset

To find out the result, I used a public dataset "Olympic Sports and Medals, 1896-2014", which was from The Guardian, a user from Kaggle. The original data was collected by the IOC Research and Reference Service and published by The Guardian's Datablog. In this dataset, there were three individual data, the data of the Summer Olympic Games, the Winter Olympic Games, and the countries' data information. Overall, there were 11 variables: year, city, sport, discipline, athlete, country, gender, event, medal, population, GDP per capita. The limitations of this dataset were that the information of the 2016 Olympic Games in Rio was not included, and the population and GDP per capita have no source date.

To reach the Olympics data of 2016 Rio, I found another dataset, "120 years of Olympic history: athletes and results", provided by rgriffin, another user of Kaggle. The source was from www.sports-reference.com in May 2018 and was collected by an incredible amount of research by a group of Olympic history enthusiasts and self-proclaimed 'statisticians'.

Besides, I searched for when the population of China was 1371220000 as recorded in the dataset, and then ensure the year of the population and GDP per capita should be 2015.

Method

With this dataset, I will use exploratory data analysis and visualization to see the general athletes and medal performance of different countries during the years with either scatter points or line chart. Then using the hypothesis test based on the test statistic to find the confidence level with p-values. Also, I will try to find out whether there is a correlation between the athlete performance or the three factors and whether there exist regression lines of them. I can evaluate the model by checking out the root mean squared error or residual standard error. To test on the linear model of the three hypotheses, I may use it to predict the medal earned by Brazilian athletes in the 2016 Rio Olympic Games and then compared the predicted value and the real value to see the validity of the predictive model.

In the hypothesis, the dependent variable will be the total medals earned by the athletes in a country, which shows the athletes' performance as general. For the first hypothesis, I would choose the variables of the population in 2015 and the total medals earned by each country in the 2012 Summer Games and 2014 Winter Games. For the second hypothesis, I would also choose the variables of the total medals earned by each country in the 2012 Summer Games and 2014 Winter Games with the GDP per capita collected in 2015. The variables for the third hypothesis will be the total medals earned by the British athletes during the years when the Games are held in their own country or not.

For the first and second hypotheses, the target is to seek the correlation between the variables and to see which is the better variable to predict the total medals earned by a country. If there exists the best-fit line with a positive correlation, it shows the hypothesis is true. In the third hypothesis, I will make a scatter points chart and draw a linear line with the total medal earned by the British against years. And to predict the total medal earned by the British athletes when the host city is in Britain. Then, I can calculate the difference between the calculated ones and the actual ones. If the actual medal number is higher than the predicted one, it can imply that the hypothesis is true.

Future Direction

It is really interesting to find out the triggering factors of the medal performance of athletes, and for extended research, the following topics will be intriguing to seek:

1. Find the trends of the female participants in the Olympic Games, based on the year and nation factors.
2. How anthropometric factors such as age, weight, and height will affect the athletes' performance.

Reference

1. Jayantha K, Ubayachandra E. G. Going for Gold Medals: Factors affecting Olympic, International Journal of Scientific and Research Publications, Volume 5, Issue 6, June 2015.
2. The Guardian, "Olympic Sports and Medals, 1896-2014", https://www.kaggle.com/the-guardian/olympic-games.

Appendix

```
sum_metal<- read.csv(file = "Desktop/olympic-games/summer.csv")

head(sum_metal)

summary(sum_metal)

glimpse(sum_metal)

unique(sum_metal$Sport)


sum_london<-sum_metal %>% filter(City=="London") %>% filter(Year==2012)

summary(sum_london)

sum_london_GBR<-sum_metal %>% filter(City=="London") %>% filter(Country=="GBR")
%>% filter(Year==2012)

summary(sum_london_GBR)
```

```
ggplot(sum_london,
aes(x=sum_london$Country,y=sum_london$Event,color=sum_london$Metal))+geom_point(
)

sum_london<-sum_metal %>% filter(City=="London") %>% filter(Gender=="Women")


summary(sum_metal$Gender)


nation<- read_csv(file = "Desktop/olympic-games/dictionary.csv")

head(nation)

summary(nation)

gdp<-nation %>% arrange(desc(GDPperCapita))

topgdp<-head(gdp,n=10)

topgdp

ggplot(topgdp,aes(x=topgdp$Country,y=topgdp$GDPperCapita))+geom_point() + labs(title =
"Top Ten GDP per Capita ")


pop<-nation %>% arrange(desc(Population))

toppop<-head(pop,n=10)

toppop


ggplot(toppop,aes(x=toppop$Country,y=toppop$Population))+geom_point()  + labs(title =
"Top Ten Populations ")
```