Olympics Data Exploration

Intermediate Python for Data Science

Zihang.WANG
Liyang.FANG
Duytan.LE

Introduction

This project aims to discover the data on the modern Olympic Games, including all the games from Athens 1896 to Tokyo 2020 (or 2021). To gain insights into data, we will analyse the data from participation, medal distribution and other aspects. The datasets used for this project are from 120 years of Olympic history: athletes and results and Tokyo 2020 Olympic Summer Games from Kaggle.

Arrangement of Group Work

Data Preprocessing - Liyang FANG
Data Visualization - Zihang WANG & Duytan LE
Data Analysis - Zihang WANG & Duytan LE

Methods

There are 2 main parts in the process:

1. Data preprocessing:
    a. Data Import
    b. Data Cleaning
    c. Data Merging
    d. Data Selection
    e. Data Group
2. Data Visualization:
    a. Bar Plot
    b. Line Plot
    c. Pie Plot
    d. Histogram Plot
    e. Scatter Plot

Results

We select some interesting results here:

1. The United States is the country with the most participation and the most medals
2. Top 10 countries with the most athletes and medals are mostly developed countries.
3. Male athletes are almost 3 times more than female athletes but females are the larger medal winners relatively.
4. After 1980, both genders of athletes got a obvious increase in winning a medal
5. The medals distribution is correlated to the participations of countries
6. Some sports are competitive because more than one country is good at it
7. In general, there is a direct correlation between height and weight of athletes.

Discussion

The results highlight the dominance of certain countries in specific sports or events, whether it is relevant with the economy, culture and sports infrastructure is an interesting topic for future study. The analysis also revealed interesting trends in medal distribution over different Olympic years. For instance, there may be fluctuations in medal counts due to factors such as changes in competition formats, host country advantages, or geopolitical events. Furthermore, it should be pointed out the limitation of the two datasets. There is a lot of missing data and misaligned data granularity will obscure the accuracy and limit the dimensions we could explore.

During the process of working on analysis data, we found that it is hard to directly use the model into our dataset to make a prediction such as Linear Regression, Logistic Regression etc. Because for a special great event, its affect factors are diverse. It is impossible to make the prediction only based on the number of medals and countries. Although we could use a Linear Model to predict the number of medals, it is meaningless and inaccurate. Even if you can get 80% or 90% as your accuracy score, it only means the model is useful and appropriate to the datasets. It can not be used for real life, it is unrealistic.

To solve the problem mentioned above, we think Neural Networks would be a better method than Machine Learning or Deep Learning. Because for Neural Networks we could increase the number of hidden layers to make sure the output can be affected by almost all factors during the process. But it also needs to consider the activation function, the weight of each element and the resource we are using whether it can be satisfied.

Conclusion

This project shows the results of the Olympic games from 1896 to 2020. Despite the limited time, some data missing or not strongly supporting the prediction of the next game, it successfully shows the analysis of the medal distribution between countries vividly. Readers can easily find out the performance of genders, countries and weights of athletes from the visualisation part. Considering the future improvement, the more detailed data of athletes which can be predicted the individual performance or other datasets such as GDP, SportInfrastruce and so on can be used to analyse the correlation between them and medals.