ORCAE 4500: Foundations of data science
Project Report

Rui Wen (rw2793)

Xin Ou (xo2119)

Introduction

The 2020 Summer Olympics, commonly known as Tokyo 2020, is an upcoming international multi-sport event that is scheduled to take place from 24 July to 9 August 2020 in Tokyo, Japan. The Olympic Games has always received extensive attention from people. In addition to the massive promotion of sports in recent years, the national participation and attention are increasingly high. We think it will be a very meaningful topic to effectively predict the final medal table of Tokyo 2020 Olympic Games. The changing tendency of women's participation rate and awards in recent years is also presented. Finally, by combining the GDP of each country in recent years with our prediction, we try to find the connection between the three.

Background

The Olympic Games is the world's largest comprehensive games hosted by the international Olympic committee. It is held every four years and lasts no more than 16 days. It is the most influential sports event in the world. In the Olympic Games, countries use sports to exchange cultures and skills with each other. The purpose is to encourage people to keep playing sports. Games of the Tokyo Olympic is the 32nd Summer Olympic Games. The Games will feature 339 events in 33 different sports, encompassing 50 disciplines. Alongside the five new sports that will be introduced in Tokyo, there will be fifteen new events within existing sports, including 3x3 basketball, freestyle BMX and Madison cycling, and new mixed events in several sports. 206 countries are expected to participate in the games.

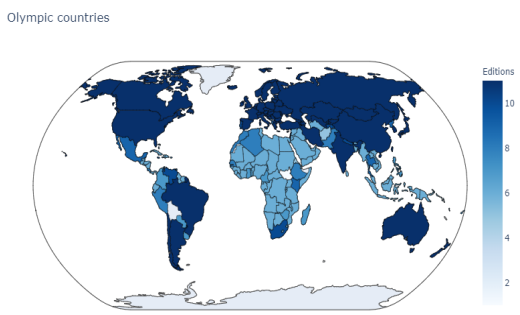
General datasets determination and improvement

Based on the two aspects of background information, we searched about the participation and awards of the Olympic Games in the past 120 years. *(Dataset one: "120 years of Olympic history: athletes and results"--basic bio data on athletes and medal results from Athens 1896 to Rio 2016)* In order to improve the validity and analyzability of the data, we conducted a preliminary screening of the data we found, retained the data of the past 30 years, Then ten representative countries were selected for the study: China, the United States, Germany, Russia, Italy, France, Netherlands, Japan, Canada and the United Kingdom. Find the relationship between population, GDP per Capita and the number of medals. *(Dataset two: "Olympic Sports and Medals, 1896-2014")* Research topics can only be developed after narrowing down the scope in the huge database.

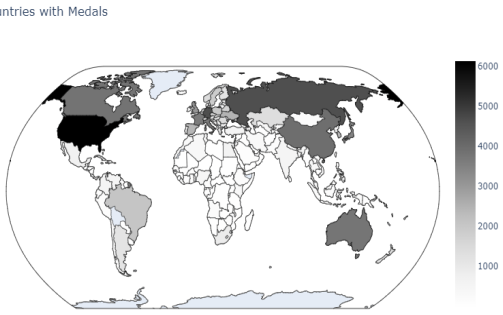
Data visualization and inspiration

Data visualization is a robust tool that we use to do feasibility analysis for our questions. Followed by online tutorial, several graphs to analysis multi-dimension data were made.

The figure on the left tells us how many Olympic games editions that each country has joined. The darker color the country has, the more edition it has joined. But there was only slight difference for each country. Which indicts that it can hardly become a proper "separator" for each country. Then, the right Figure tells us the medal numbers that each country has since 1996. It shows a clear difference with each country, which gave us the idea that we can try to find the model to predict the medal number for the next year's games.



Number of Olympic editions for each country



Number of medals for each country

Therefore, we could analysis following three problems:

1. ***What is the relationship between GDP and medal table?***

   We want to explore the relationship between GDP and medals. We search the GDP data of ten countries and make a new column with the 'total country medal distribution' table. And the hypothesis test will be necessary. We may set the null hypothesis of there is a positive relationship with country's GDP and medals. And may set the alternative hypothesis of there isn't a positive relation between them.

2. ***How to get a more general prediction of Tokyo Olympic Games' medal table?***

   We can approach the changes clearly by using many bar charts and connected them together continuously. We collected data from ten countries, but ultimately predicted the top five on the medal list. For the prediction, regression, hypothesis test and simulation must be the powerful skills.

3. ***What are the trends in women's participation?***

   With the growth of GDP, the participation rate of various countries has gradually increased, and the participation rate and awards of women have also changed year by year. 'Female' was screened using the 'group' function and researched, using various charts and plots to show its changing trends and choosing the best one which can present the tendency perfectly.

Model assumption

Before the prediction, there are some model assumptions:

1. Experts predict that during this Olympic cycle, the British team will continue to be depressed, so we take 50% of the final forecast.

2. At the Rio Olympics, Japan won a record 41 medals. Experts believe that as the host,

the number of medals at the Tokyo Olympics in Japan will increase by about 50%, traditional advantages (judo, wrestling) and newly added items (karate, etc.). Is the largest source of medals. So, as the host country, Japan has a 3% higher medal ratio than other countries.

3. The information provided by the IOC is true and reliable. (e.g. total medals to be awarded in 2020)

Question analysis

Question 1. What is the relationship between GDP and medal table?

1.1 Single year model

First, I tried to locate which factor will have the effect on the medal rank in previous year. What I have at that time was the latest year GDP per capita and the total population for each country. Basing on it, I tried to figure out whether those two parameters can be such factors. However, both covariances are only near 0.33 which shows no significant correlation with medal rank. Then, I thought whether the total GDP can be this factor. So, I timed those two columns up and got the total GDP for the single year. The covariance between GDP and medal number reached 0.83, so I made the one-year table with the 2014-year medal number and the latest total GDP, and got the single year model.

Code	Country	Population	GDP per Capita	count
AFG	Afghanistan	3.25266e+07	594.323	1
ALG	Algeria	3.96665e+07	4206.03	1
ARG	Argentina	4.34168e+07	13431.9	20
ARM	Armenia	3.01771e+06	3489.13	2
AUS	Australia	2.37812e+07	56311	114
AZE	Azerbaijan	9.65135e+06	5496.34	10
BAH	Bahamas	388019	22817.2	4
BEL	Belgium	1.12857e+07	40324	3
BLR	Belarus	9.513e+06	5740.46	21
BOT	Botswana	2.26248e+06	6360.14	1

Corr(pop, medal)	Corr (GDP per captia, medal)
0.3195	0.3364644635461069

Code	Country	Population	GDP per Capita	count	pop*gdp
AFG	Afghanistan	3.25266e+07	594.323	1	1.93313e+10
ALG	Algeria	3.96665e+07	4206.03	1	1.66839e+11
ARG	Argentina	4.34168e+07	13431.9	20	5.83169e+11
ARM	Armenia	3.01771e+06	3489.13	2	1.05292e+10
AUS	Australia	2.37812e+07	56311	114	1.33914e+12
AZE	Azerbaijan	9.65135e+06	5496.34	10	5.30471e+10
BAH	Bahamas	388019	22817.2	4	8.85352e+09
BEL	Belgium	1.12857e+07	40324	3	4.55086e+11
BLR	Belarus	9.513e+06	5740.46	21	5.4609e+10
BOT	Botswana	2.26248e+06	6360.14	1	1.43897e+10

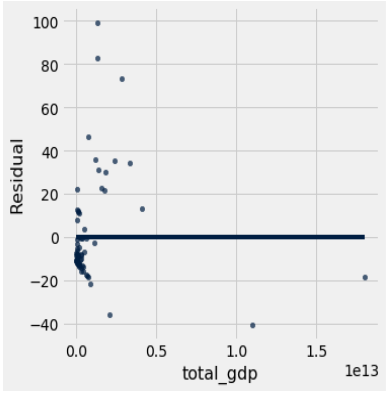
Corr (pop*gdp,medal)
0.8291630330280012

Dep. Variable:	medal_number	R-squared:	0.688
Model:	OLS	Adj. R-squared:	0.683
Method:	Least Squares	F-statistic:	158.4
Date:	Fri, 29 Nov 2019	Prob (F-statistic):	7.32e-20
Time:	12:42:06	Log-Likelihood:	-339.32
No. Observations:	74	AIC:	682.6
Df Residuals:	72	BIC:	687.3
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	12.1050	2.976	4.067	0.000	6.172	18.038
total_gdp	1.42e-11	1.13e-12	12.586	0.000	1.19e-11	1.64e-11

Omnibus:	46.623	Durbin-Watson:	2.235
Prob(Omnibus):	0.000	Jarque-Bera (JB):	135.565
Skew:	2.097	Prob(JB):	3.65e-30
Kurtosis:	8.135	Cond. No.	2.81e+12

Single-year model table



Residual scatter

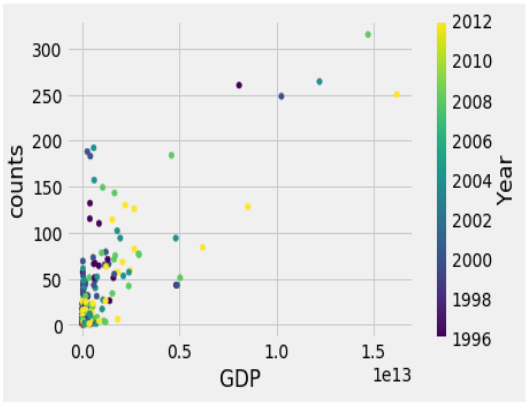
1.2 Multiply year model

1.2.1 Data processing

However, the single year model only contains near 80 countries data which is far away sufficient to create a reliable data. Therefore, we collected more year data for total GDP in the world bank website, and did a lot table operation, then merged with our medal number table for each year and each country. After data processing, the table and scatter plot are shown as follow.

Country	Year	GDP	counts
AFG	2008.0	1.010922e+10	1
AFG	2012.0	2.000162e+10	1
ARG	1996.0	2.721498e+11	20
ARG	2000.0	2.842038e+11	20
ARG	2004.0	1.646579e+11	47
ARG	2008.0	3.615580e+11	51
ARG	2012.0	5.459824e+11	20
ARM	1996.0	1.596969e+09	2
ARM	2000.0	1.911564e+09	1
ARM	2008.0	1.166204e+10	6
ARM	2012.0	1.061932e+10	2
AUS	1996.0	4.003027e+11	132
AUS	2000.0	4.152226e+11	183
AUS	2004.0	6.124904e+11	157

Multiply year table



Scatter plot

1.2.2 Module building

Then, we got the multiple year model. The results are shown as follow. From the residual plot, we can tell that the model is quilt reliable since scatter around 0. Then we check the summary table for the model, since the adj R-squared is larger than 0.5, we can draw a rough conclusion that the multiple year model is reliable. In order to test our model, we can select just one-year data and get the predict result. Then, we can compare the real medal numbers at that year. We chose the data in 2014 and drew the comparison graph.

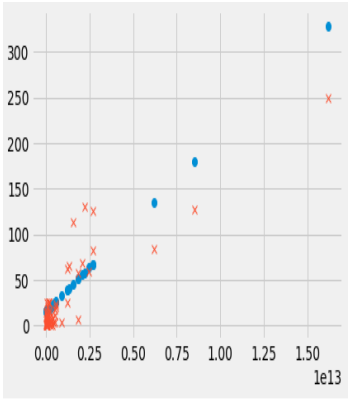
Dep. Variable:	counts	R-squared:	0.578
Model:	OLS	Adj. R-squared:	0.577
Method:	Least Squares	F-statistic:	373.3
Date:	Mon, 02 Dec 2019	Prob (F-statistic):	5.96e-53
Time:	16:57:22	Log-Likelihood:	-1336.2
No. Observations:	274	AIC:	2676.
Df Residuals:	272	BIC:	2684.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	15.5022	2.058	7.533	0.000	11.451	19.554
GDP	1.935e-11	1e-12	19.321	0.000	1.74e-11	2.13e-11

Omnibus:	165.655	Durbin-Watson:	0.677
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1162.099
Skew:	2.442	Prob(JB):	4.50e-253
Kurtosis:	11.829	Cond. No.	2.20e+12

Multi-year model table

Residual scatter



Test on a certain year

1.3 Result and discussion

We use the latest GDP of each country combine with our model to make our medal rank prediction. The top rank table shown as follow. However, there is a significant difference between our prediction with some table made by other experts. This is because the GDP is not the only parameter for the prediction model. What's more, we used the linear model to make the prediction, while the Poisson regression could work better, since the outcome is count data.

Country	predicted medal number
USA	129.0
CHN	88.0
JPN	35.0
GBR	22.0
FRA	22.0
IND	21.0
ITA	17.0
BRA	16.0
KOR	15.0
RUS	15.0
CAN	15.0
ESP	14.0
AUS	14.0
MEX	12.0

Medal rank prediction using our model

gracenote A NIELSEN COMPANY | **Virtual Medal Table** 1-10

#	NOC	G	S	B	T
1	United States	53	30	41	124
2	China	40	21	23	84
3	Japan	30	27	14	71
4	Russia	21	21	24	66
5	Great Britain	15	10	18	43
6	Germany	10	14	17	41
7	Australia	14	14	10	38
8	France	8	16	13	37
9	Netherlands	12	10	14	36
10	Italy	6	13	17	36

Last update: 27 August 2019

Prediction table made by experts

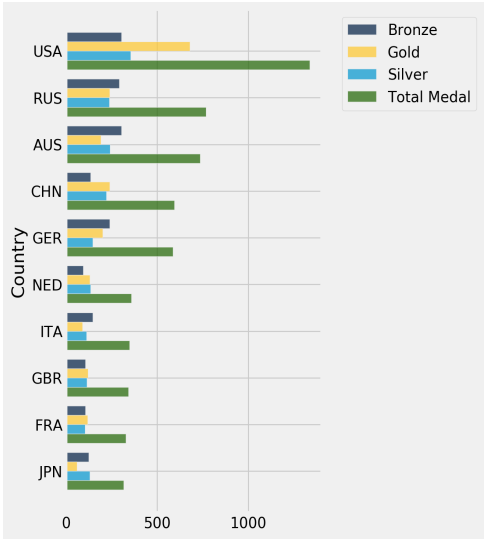
Question 2. How to get a more general prediction of Tokyo Olympic Games' medal table?

2.1 General trend of total medal table in previous years

Using the 'pivot' function of 'Medal' and 'Country' to get the table where we can see the ranking of the approximate total number of medals in the ten countries surveyed.

Country	Bronze	Gold	Silver	Total Medal
USA	304	678	355	1337
RUS	291	239	238	768
AUS	303	191	241	735
CHN	134	240	221	595
GER	239	201	147	587
NED	94	130	135	359
ITA	146	89	113	348
GBR	106	120	115	341
FRA	106	119	103	328
JPN	125	59	131	315

Total number of medal



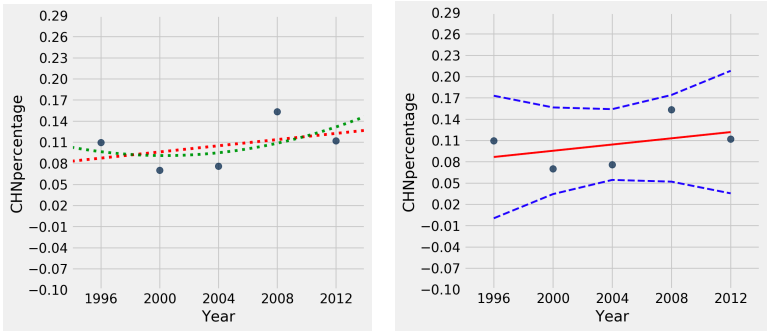
Bar chart of medal number

But the trend in the total number of medals in previous years can't fully predict the medal table in 2020. The total medal number can't present the changing tendency. We should consider every year separately and build each country's prediction model.

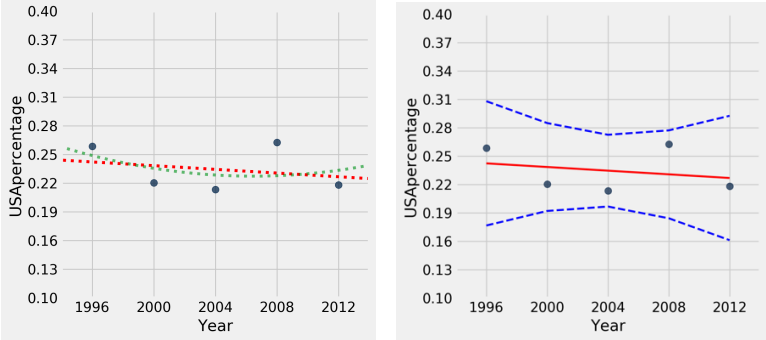
2.2 Prediction model for each country

For each country, we will calculate the country's medal percentage in that year which means the country's total medal number in that year divided by the all countries' total medal number in that year. Because the total number of medals awarded each year changes, forecasting with the percentage of medals is more accurate than forecasting with the number of medals. Finally multiply the predicted medal percentage by the total number of medals in 2020 to get our final predicted medal list.

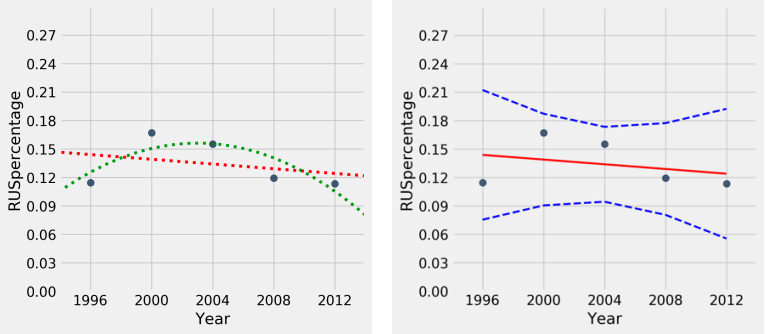
Each country uses two prediction models, one linear and one non-linear. Establish a hypothesis test. The null hypothesis is that the two models do not have much difference in model prediction. If the null hypothesis is accepted, it shows that both models can achieve the prediction effect well, then we take the linear model, which is simpler. If we reject the null hypothesis, indicating that the nonlinear model can better achieve the prediction effect, then we take the nonlinear model. We also use the prediction method to present the upper prediction level and the lower prediction level. There are the final prediction models:



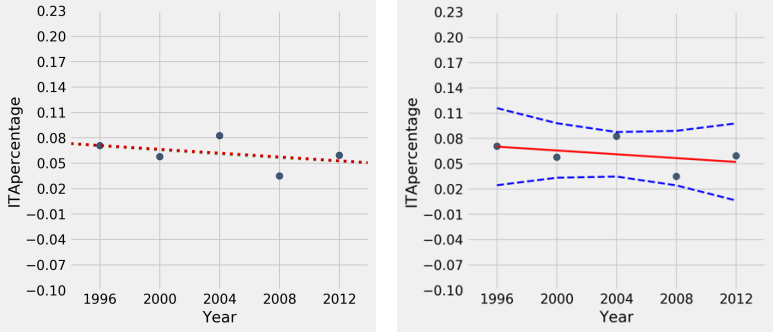
CHN prediction model



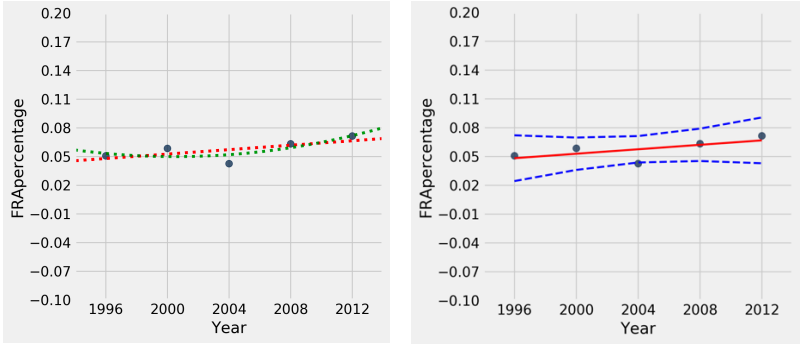
USA prediction model



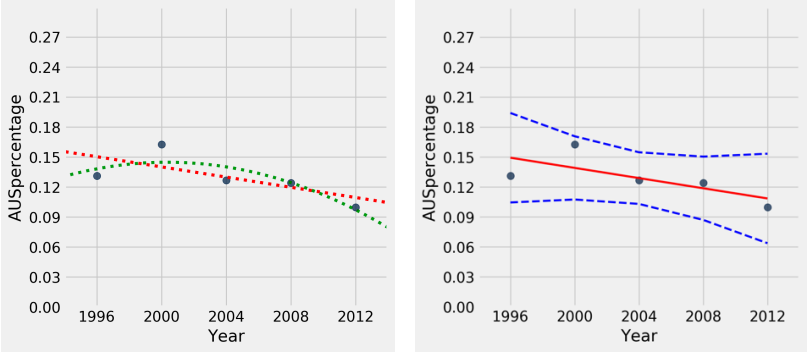
RUS prediction model



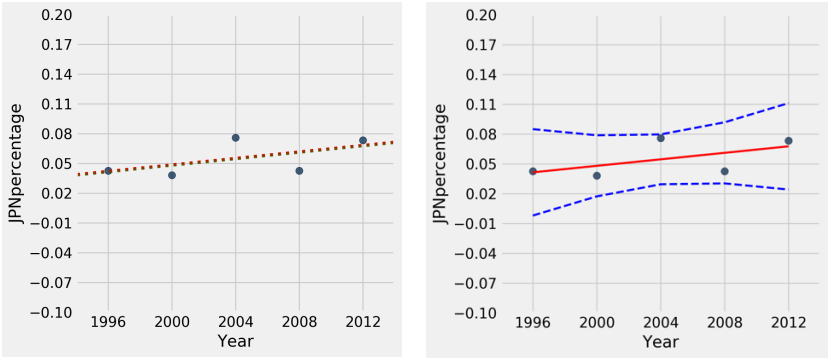
ITA prediction model



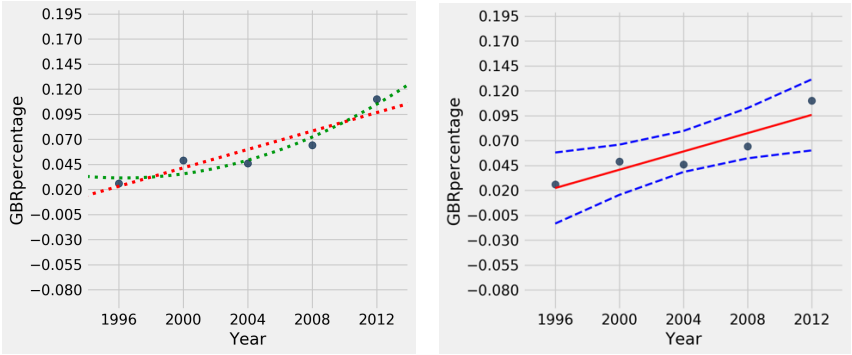
FRA prediction model



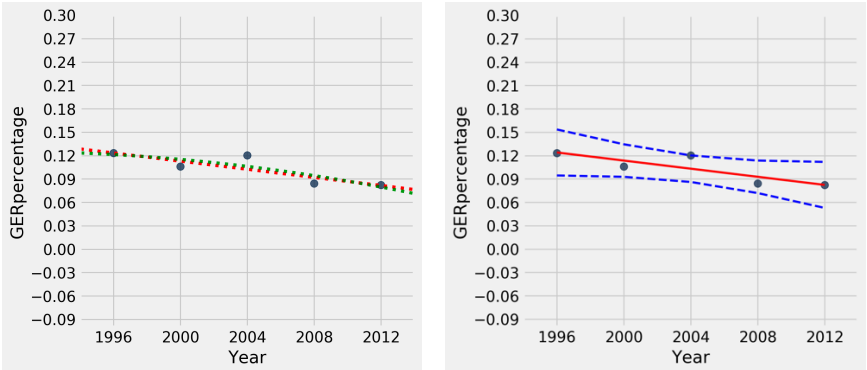
AUS prediction model



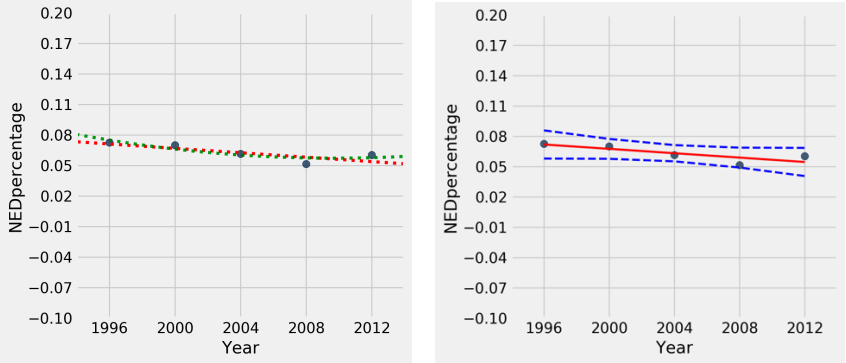
JPN prediction model



GBR prediction model



GER prediction model



NED prediction model

2.3 Result and discussion

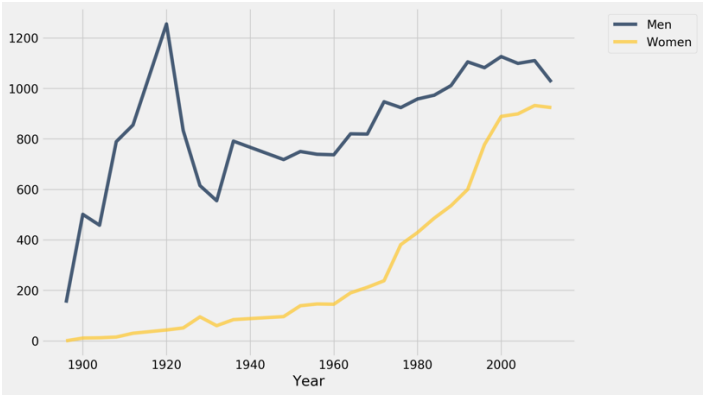
Our final medal list based on the percentages of awards won by countries in previous years is almost the same as that predicted by experts, indicating that our prediction model is reasonable.

Country	2020 Medal Predict
USA	127
CHN	81
RUS	66
JPN	64
AUS	51
FRA	43
GBR	39
GER	35
UKR	26
ITA	25

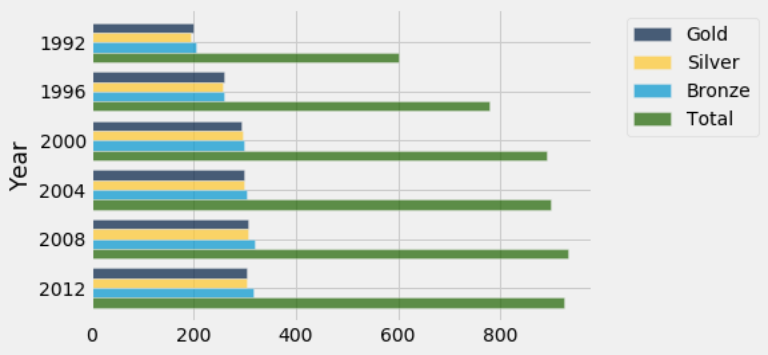
Medal rank prediction using our model

Prediction table made by experts

Question 3. What are the trends in women's participation?



Men and women's participation graph



Women's winning situation

With the rise of the women's liberation movement, women have begun to move out of the family and into society, which has laid a social foundation for women to participate in the Olympic Games. Advances in medicine and related research have also provided scientific support for women's participation in sports. In the 1900 Olympic Games, women participated in golf and tennis competitions. Since then, female athletes have participated in each Olympic Games. The number of participating events and participation has generally increased. However, at that time, the number of women participating in sports was small and the impact was not large. Therefore, in terms of the overall proportion, the proportion of female athletes in the Olympic Games was still low, and it was not even officially recognized. It was not until 1924 that the International Olympic Committee officially allowed women athletes to participate, thereby helping women athletes to enter the Olympic Games. Beginning in 1928, the number of women's events and participation in the Olympic Games began to increase. From the figure we can find that women's participation rate has increased year by year, and they can get more and more medals.

Conclusion

In practical problems, the nature of different countries and the level of economic development have greatly affected the Olympic medal table ranking. A country's total national economic income (GDP), total population (POP), The proportion of young people in the total population (Gen) host countries (Gen) and the political system (Sys) and other factors to conduct a comprehensive analysis of the forecast. If you want to make a complete prediction, we need to use the following formula to obtain a more accurate prediction result

$$M = \beta_0 + \beta_1 \bullet GDP_S + \beta_2 \bullet POP_S + \beta_3 \bullet Gen + \beta_4 \bullet Host + \beta_5 \bullet Sys + \varepsilon$$

Task distribution

We divided our work almost to 50% and 50% for all the parts including coding, reporting and presentation.