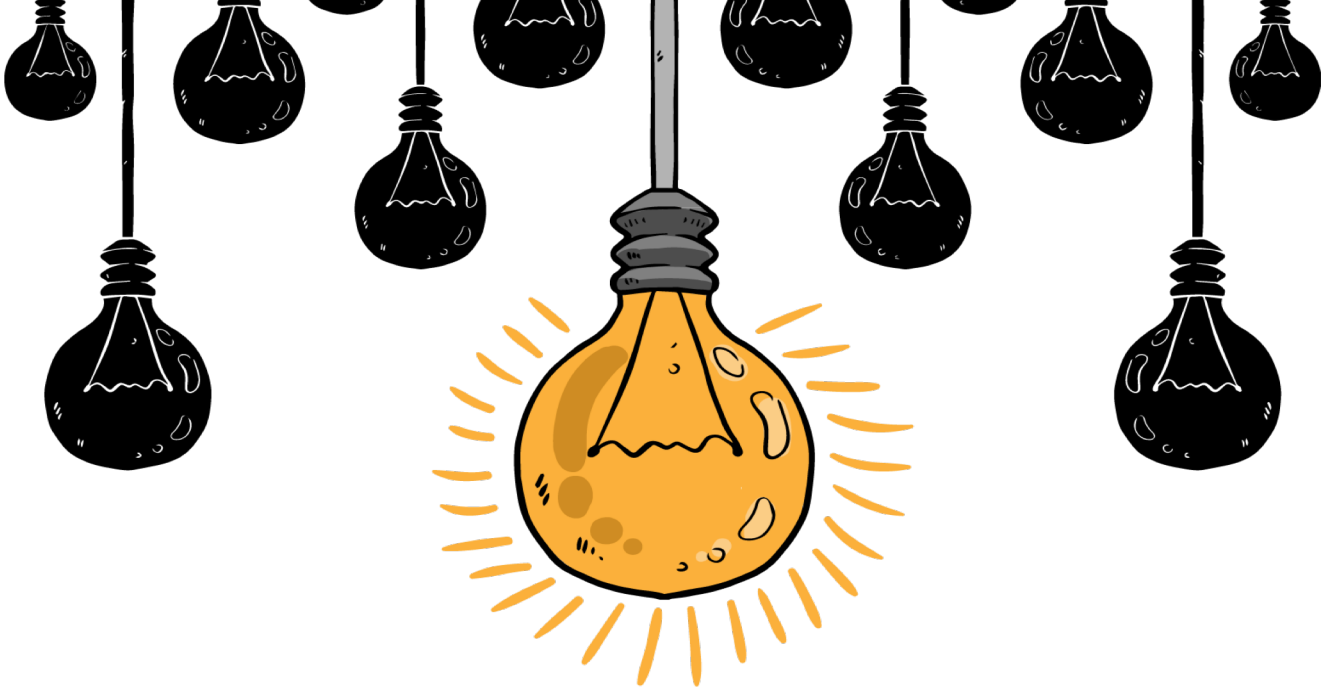
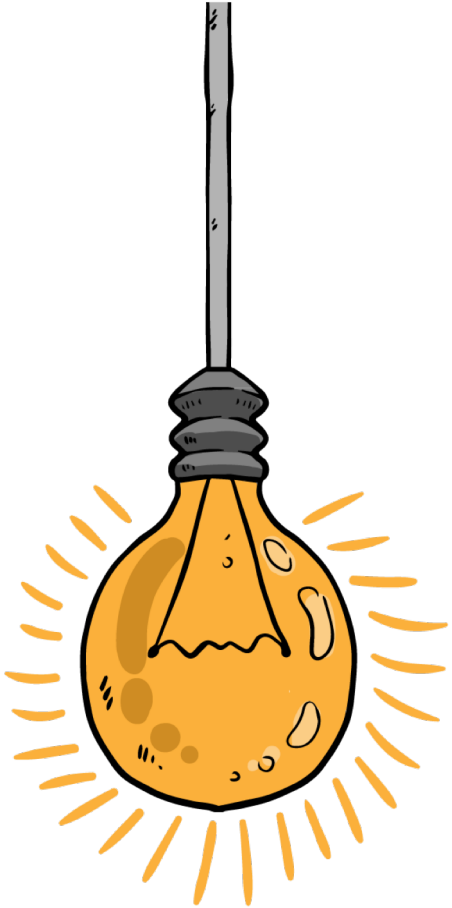


Olympics Medal Prediction

ORCA 4500 Foundation of Data science
Project Presentation

XIN OU (xo2119)
RUI WEN (rw2793)



Content

General View of Our Datasets

Olympic Sports and Medals, 1896-2014

Columns

- ID
- Name
- Sex
- Age
- Height
- Weight
- Team
- NOC
- Games

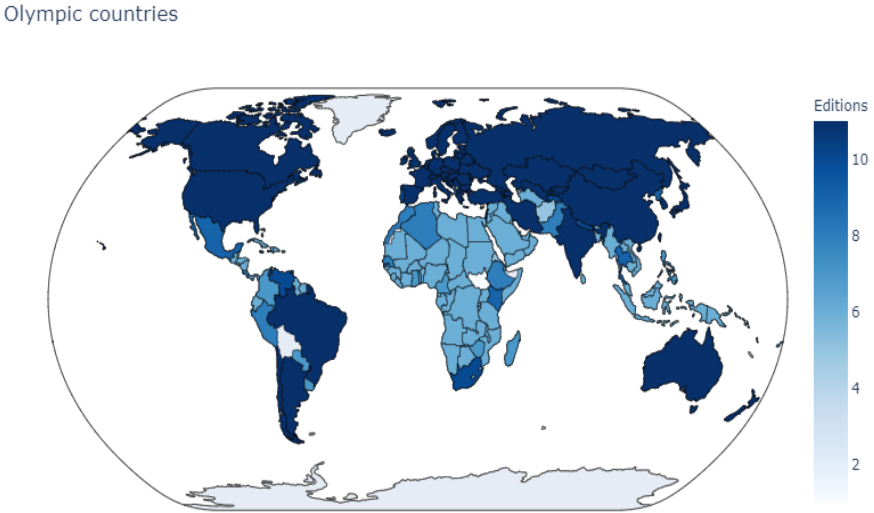
120 years of Olympic history: athletes and results

Columns

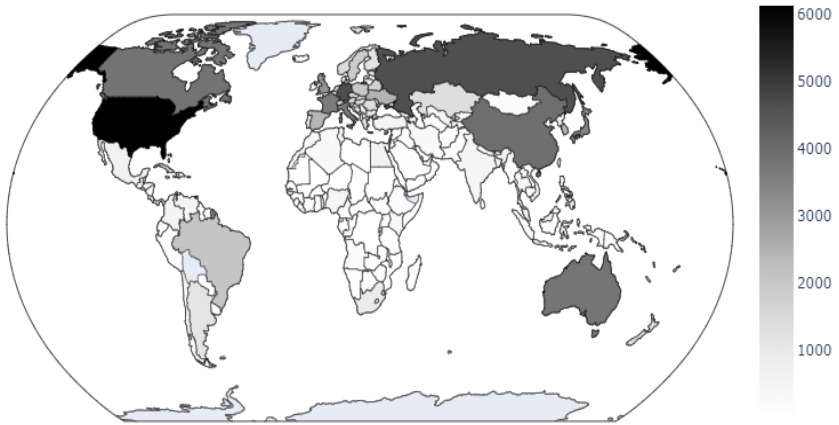
- # Year
- City
- Sport
- Discipline
- Athlete
- Country
- Gender
- Event
- Medal

General View of Our Datasets

Number of Olympic edition for each country



Number of medals for each country



Q1: What is the relationship between GDP and medal table

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

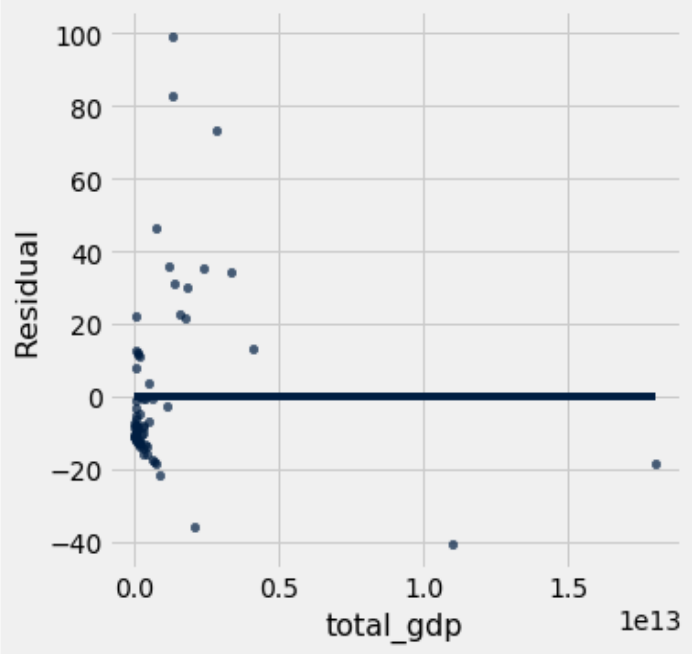
Corr (pop,medal)	Corr (GDP per captia,medal)
0.3195	0.3364644635461069

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

Q1: What is the relationship between GDP and medal table

Residual scatter graph



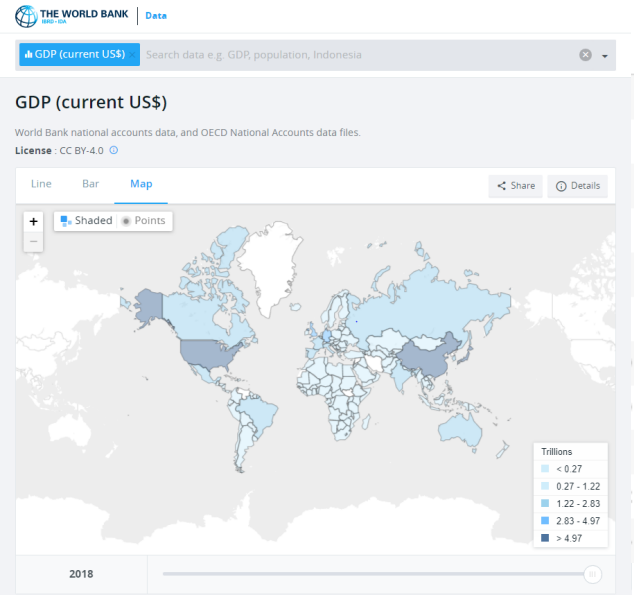
Single-year model table

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

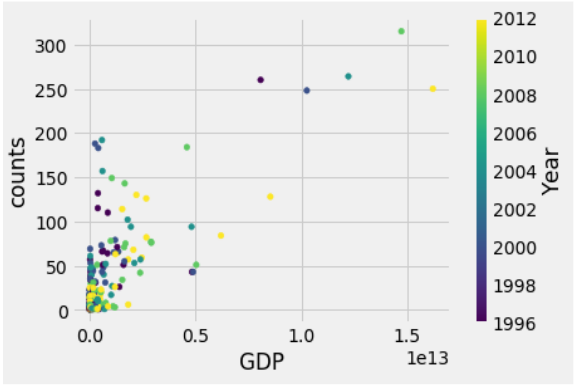
Q1: What is the relationship between GDP and medal table

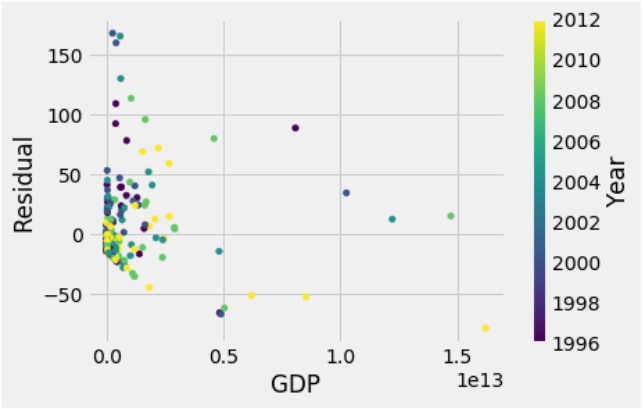


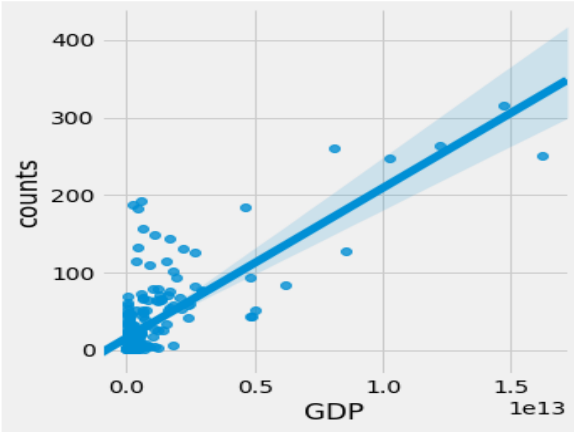
Country Code	1996	1997	1998	1999	2000
ABW	1.379961e+09	1.531944e+09	1.665101e+09	1.722799e+09	1.873453e+09
AFG	NaN	NaN	NaN	NaN	NaN
AGO	7.526447e+09	7.648377e+09	6.506230e+09	6.152923e+09	9.129595e+09
ALB	3.199643e+09	2.258516e+09	2.545967e+09	3.212119e+09	3.480355e+09
AND	1.223945e+09	1.180597e+09	1.211932e+09	1.239876e+09	1.434430e+09
...	...	...	...	...	...
XKX	NaN	NaN	NaN	NaN	1.849196e+09
YEM	5.785685e+09	6.838557e+09	6.325142e+09	7.641103e+09	9.652436e+09
ZAF	1.476063e+11	1.525874e+11	1.377748e+11	1.366323e+11	1.363613e+11
ZMB	3.597221e+09	4.303282e+09	3.537683e+09	3.404312e+09	3.600683e+09
ZWE	8.553147e+09	8.529572e+09	6.401968e+09	6.858013e+09	6.689958e+09

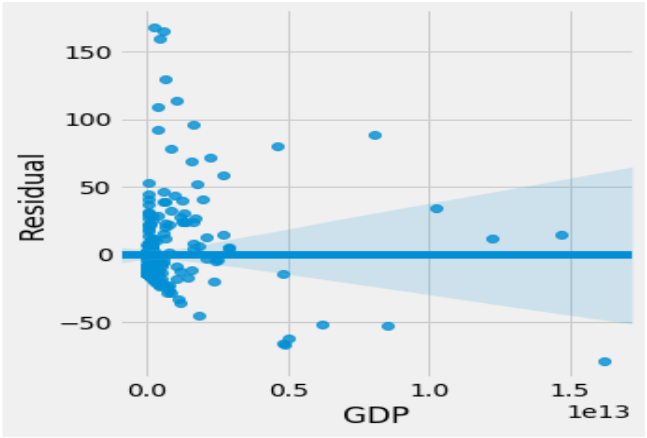
Country	Year	GDP	counts
AFG	2008.0	1.010922e+10	1
AFG	2012.0	2.000162e+10	1
ARG	1996.0	2.721498e+11	20
ARG	2000.0	2.842038e+11	20
ARG	2004.0	1.646579e+11	47
ARG	2008.0	3.615580e+11	51
ARG	2012.0	5.459824e+11	20
ARM	1996.0	1.596969e+09	2
ARM	2000.0	1.911564e+09	1
ARM	2008.0	1.166204e+10	6
ARM	2012.0	1.061932e+10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

Q1: What is the relationship between GDP and medal table









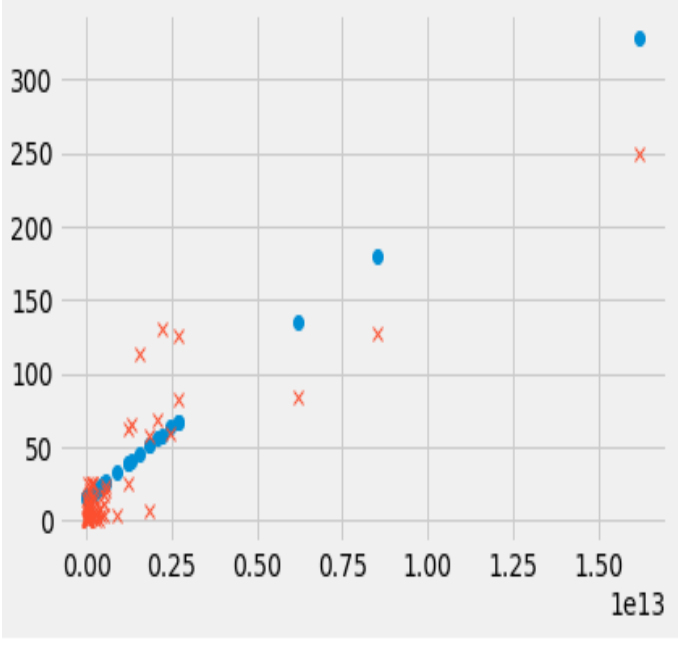
Multiple-year model table

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

Q1: What is the relationship between GDP and medal table



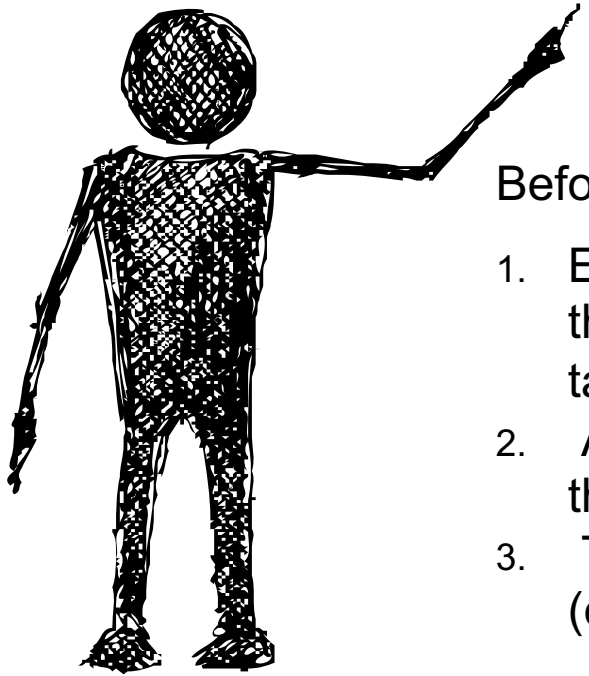
Country	predicted medal number
USA	129.0
CHN	88.0
JPN	35.0
GBR	22.0
FRA	22.0
IND	21.0
ITA	17.0
BRA	16.0
KOR	15.0
RUS	15.0
CAN	15.0
ESP	14.0
AUS	14.0
MEX	12.0

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

Q2: An Prediction of Tokyo Olympic Games' Medal Table



Before the prediction, there are some model assumptions:

1. Experts comprehensively analyze the British team in this Olympic cycle will be weak in many projects, so we take the results of the forecast of 1/2.
2. 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)

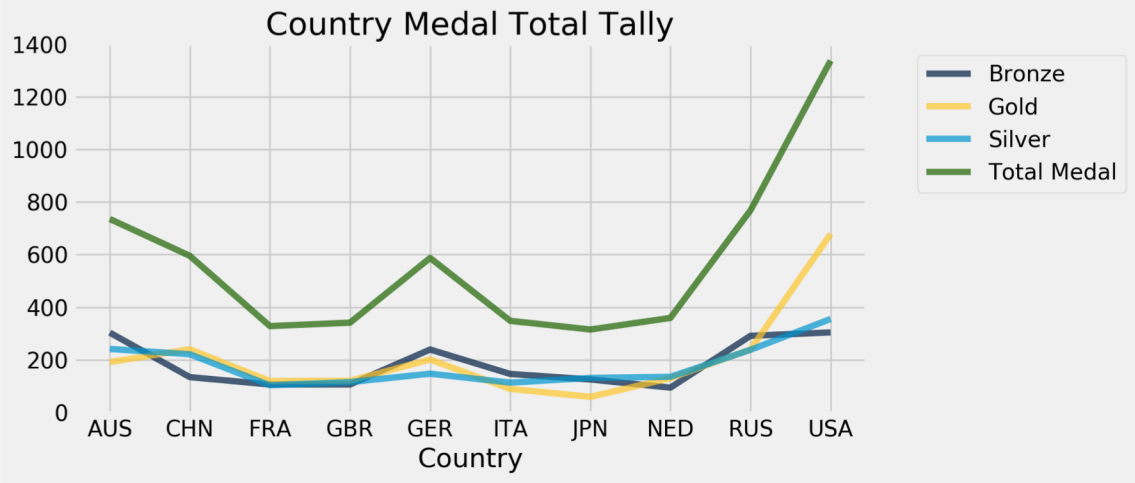
General trend of total medal table in previous years

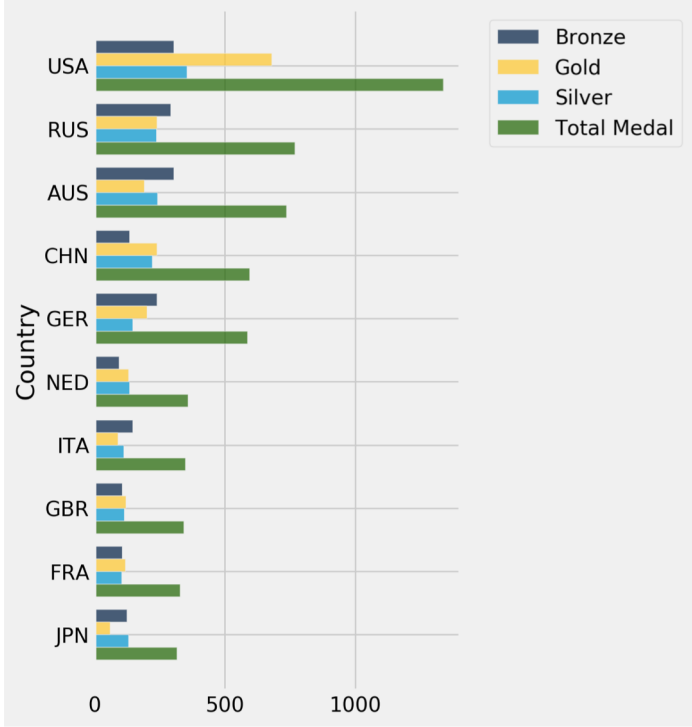
```
country_medal1 = medal2.pivot('Medal','Country')
#explore the ten countries' achievements based on three kinds of medal
country_medal2 = country_medal1.with_column('Total Medal', country_medal1.column('Bronze')
                                            + country_medal1.column('Gold')
                                            + country_medal1.column('Silver')).sort('Total Medal'
                                                                                    ,descending = True)
```

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

From this table we can see the ranking of the approximate total number of medals in the ten countries surveyed.

General trend of total medal table in previous years



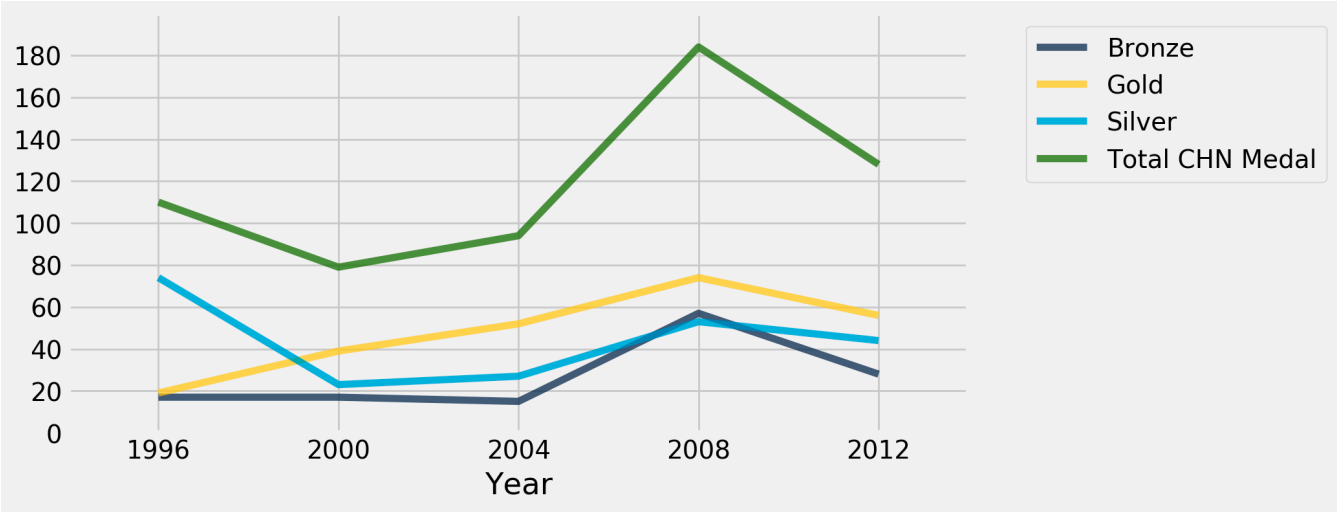


But the trend in the total number of medals in previous years does not fully predict the medal table in 2020. We should consider every year separately and build each country's prediction model.

China Prediction Model

Year	Bronze	Gold	Silver	Total CHN Medal	CHNpercentage
1996	17	19	74	110	0.109453
2000	17	39	23	79	0.0702222
2004	15	52	27	94	0.0759289
2008	57	74	53	184	0.153333
2012	28	56	44	128	0.11179

Percentage * Total Medal number in 2020



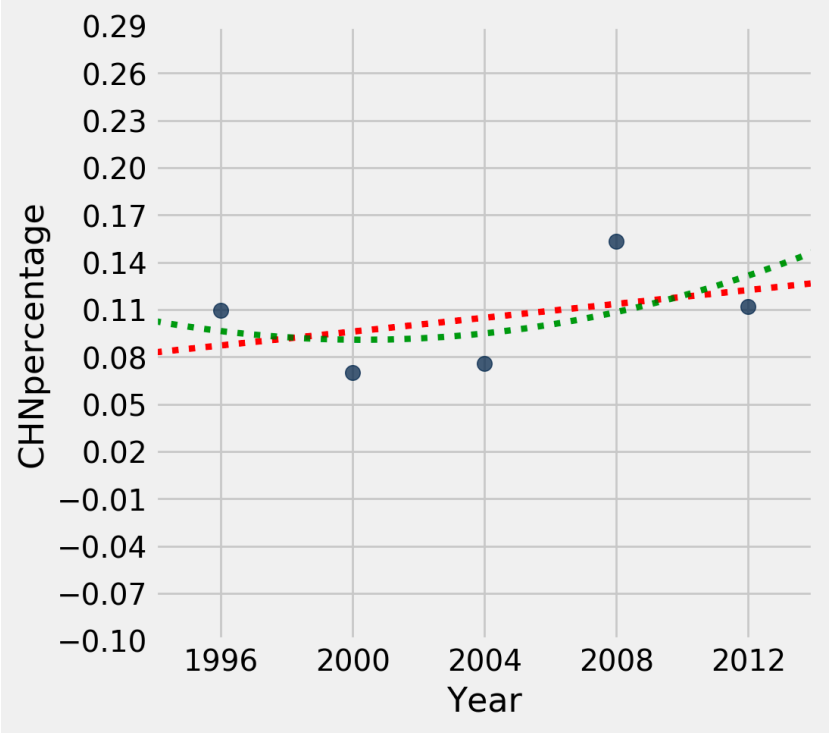
China Prediction Model

Linear model

```
CHN3_df = CHN3.to_df()
result1 = smf.ols(formula="CHNpercentage ~ Year", data=CHN3_df).fit()
result2 = smf.ols(formula="CHNpercentage ~ Year + I(Year**2) + I(Year**3)",
                  data=CHN3.to_df()).fit()
print(result1.params)
print(result2.params)
```

Cubic model (nonlinear)

China Prediction Model



H_0: Model of order 1 and order 3 fit the data equally well.

H_1: Model of order 1 and order 3 do not fit the data equally well.

F-test

```
from statsmodels.stats.anova import anova_lm
anova_lm(result1,result2)
```

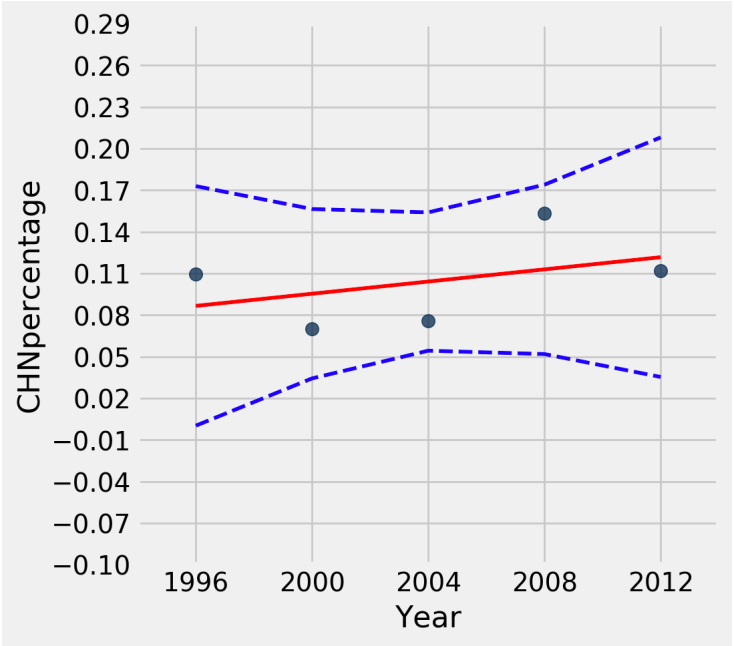
	df_resid	ssr	df_diff	ss_diff	F	Pr(>F)
0	3.0	0.003682	0.0	NaN	NaN	NaN
1	2.0	0.003365	1.0	0.000318	0.188781	0.706317

P-value > 0.05, can not reject the null hypothesis. Prefer the linear model with fewer parameters. (model of order 1)

China Prediction Model

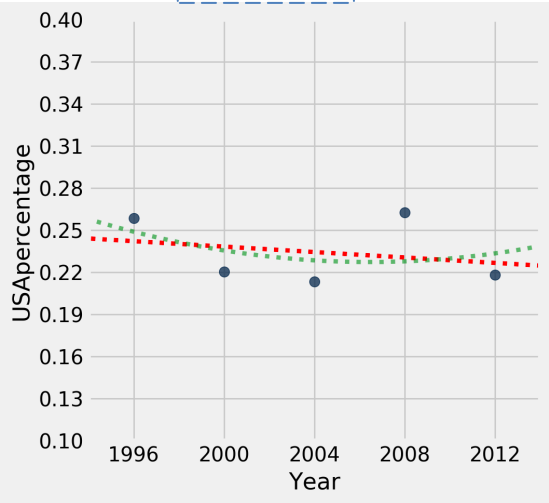
```
predictions = result1.get_prediction()
predictions.summary_frame(alpha=0.05).head()
frame = Table().from_df(predictions.summary_frame(alpha=0.05))
```

```
array([0.08658824, 0.09536688, 0.10414552, 0.11292416, 0.12170281])
```

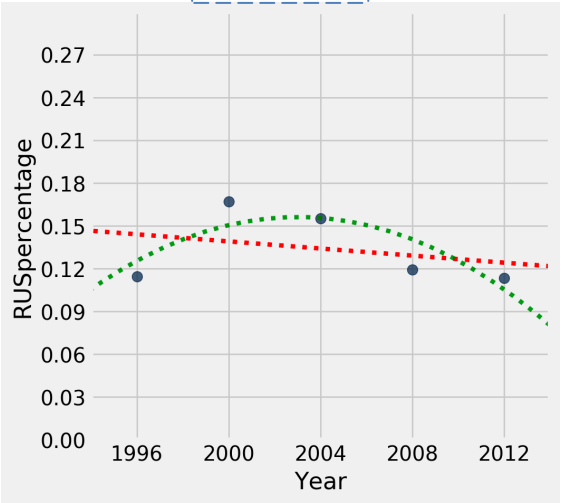


Other Prediction Model

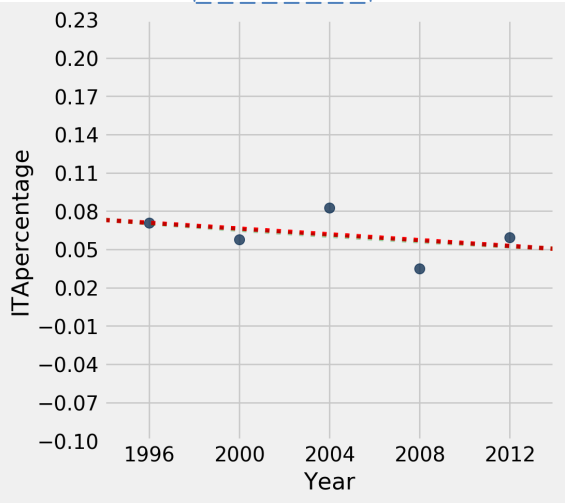
USA



RUS

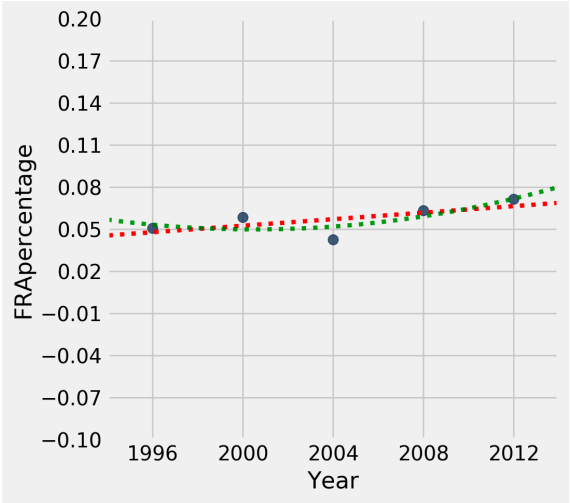


ITA

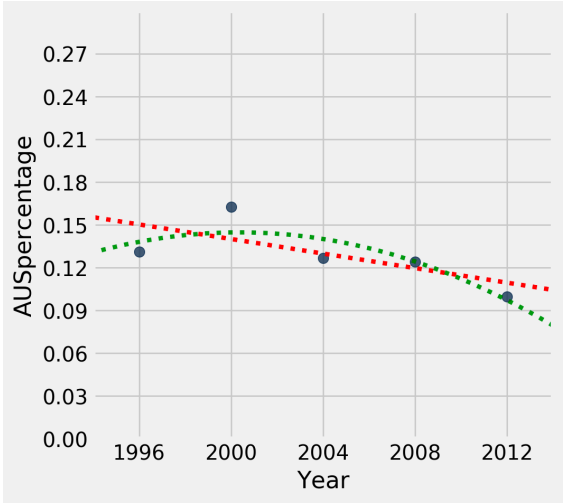


Other Prediction Model

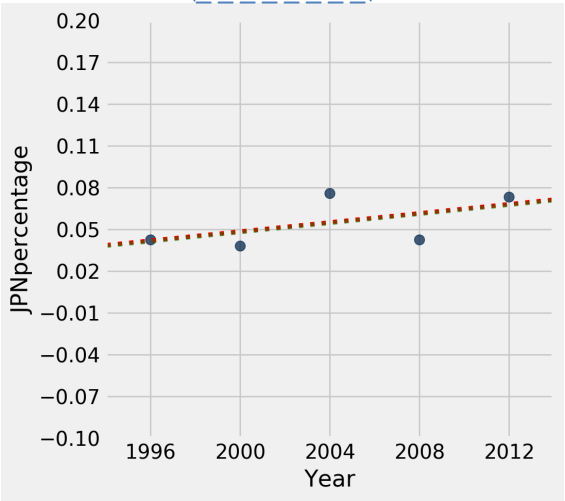
FRA



AUS

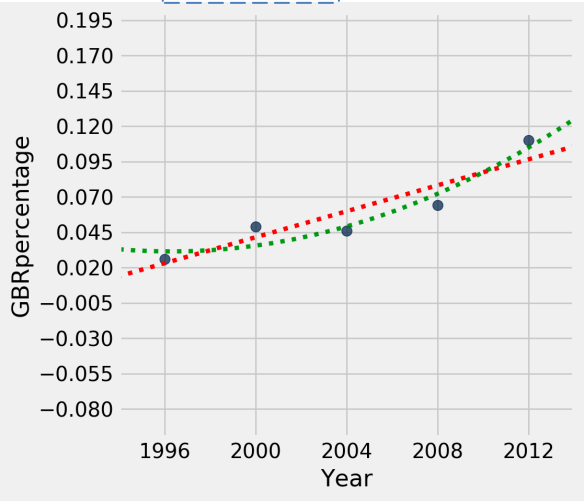


JPN

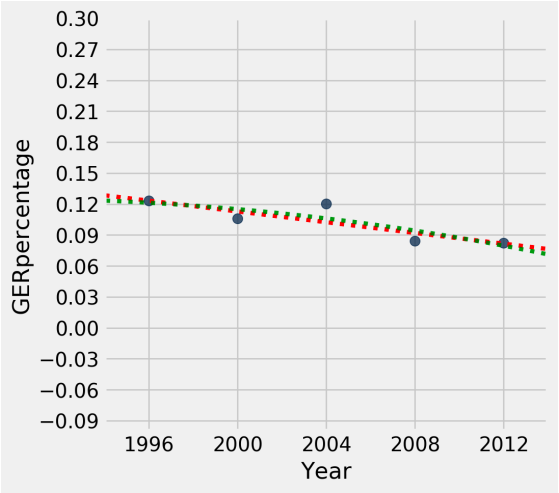


Other Prediction Model

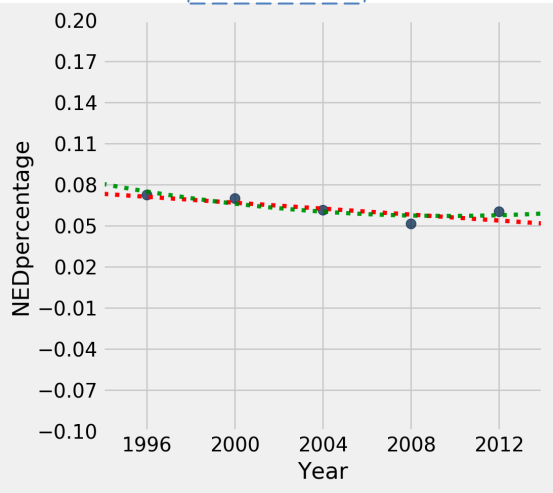
GBR



GER



NED



Final Prediction Table

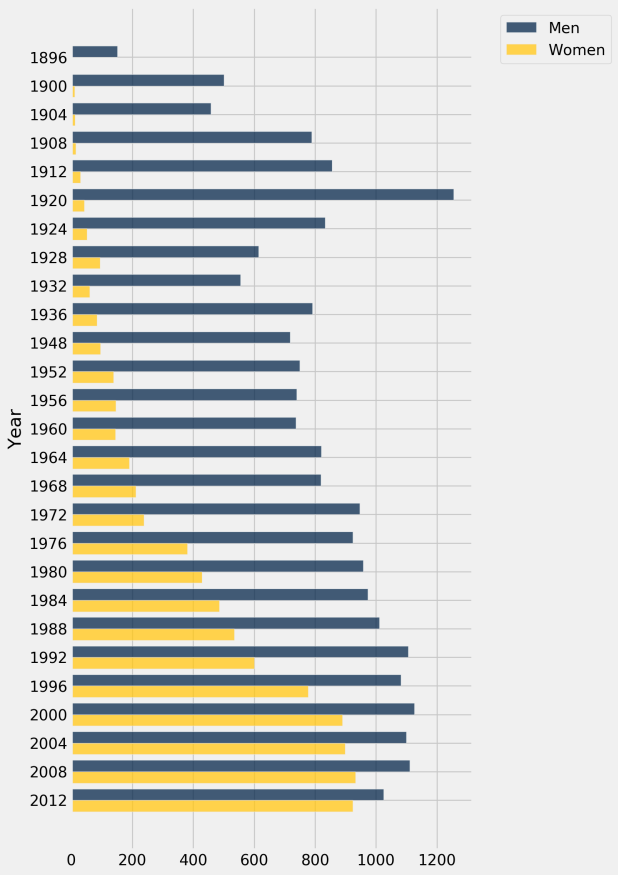
Country	2020 Medal Predict
USA	127
CHN	81
RUS	66
JPN	64
AUS	51
FRA	43
GBR	39
GER	35
UKR	26
ITA	25

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

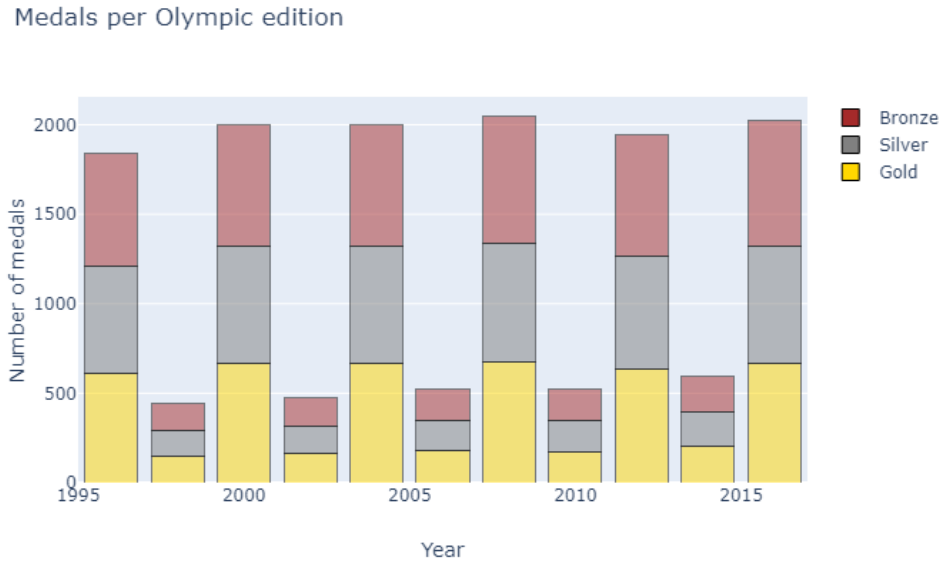
#	NOC	G	S	B	T
1	United States	53	30	41	124
2	China	40	21	23	84
3	Japan	30	27	14	71
4	Russia	21	21	24	66
5	Great Britain	15	10	18	43
6	Germany	10	14	17	41
7	Australia	14	14	10	38
8	France	8	16	13	37
9	Netherlands	12	10	14	36
10	Italy	6	13	17	36

Last update: 27 August 2019

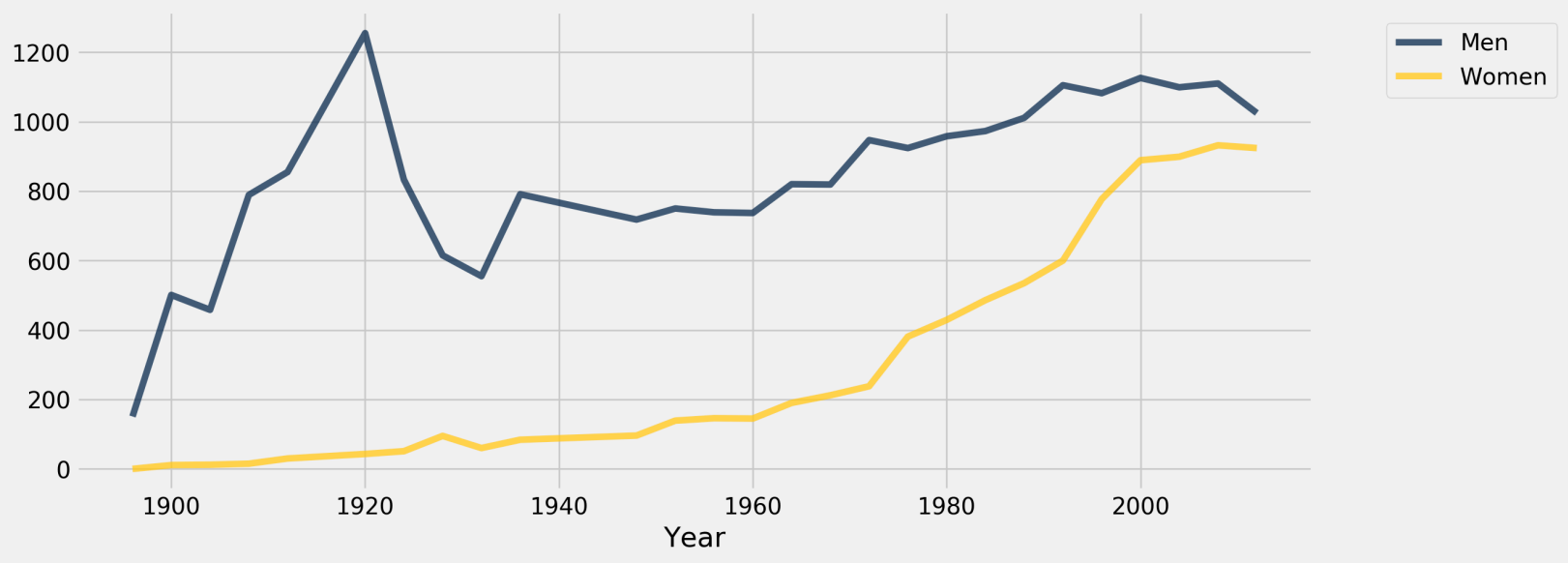
Q3: What is the trends in women's participation




```
female = medal.pivot( 'Gender','Year')
female.select(np.arange(0,3)).plot(0, width = 12, height = 5)
female.barh('Year')
female.scatter('Year', fit_line = True)
```

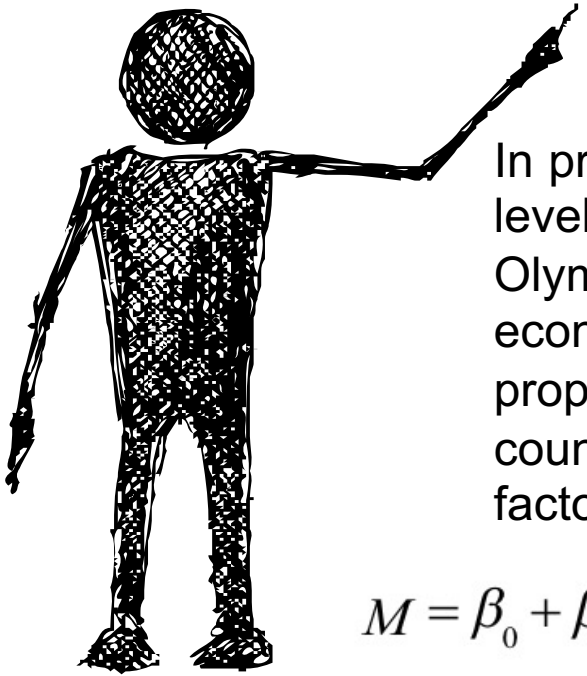


Q3: What is the trends in women's participation



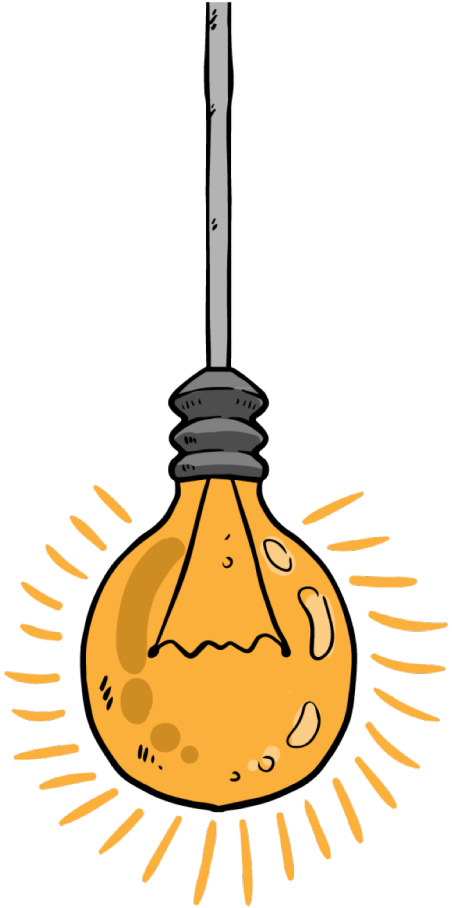
Increasing!!! More and more women will attend the sports contest and win a medal.

Summary



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.

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



Thank you for your listening!