120 years of Olympic History Data Analysis

Yuyang Han, Lin Tao

This document is the project report for the IE 7275 Data Mining course at Northeastern University. The project aims to quantify the qualitative description of Olympic history and athlete data using data mining techniques and use it as an effective tool for identifying olympic performance of athletes and countries/regions

Introduction

The modern Olympic Games are leading international sporting events featuring summer and winter sports competitions in which thousands of athletes from around the world participate. During the past 120 years, some countries perform better while others perform worse than before. It might be because the overall performance improved or an outstanding athlete emerged. Therefore, we decided to do some data analysis about 120 years of olympics to see the performances of countries at the olympics and what makes them better or worse.

The main objective is to discover the performances of athletes based on sports types, countries and year. Are there any relationships between GDP and the number of medals? Does hosting the Olympics improve the performances? Based on the existing data, which countries will be the top five of the next Olympic games in the future? Let's Explore.

Dataset

Our Data is retrieved from Kaggle "120 years of Olympic history: athletes and results"[1]. This is a historical dataset on the modern Olympic Games, including all the Games from Athens 1896 to Rio 2016. The Winter and Summer Games were held in the same year up until 1992. After that, they staggered.The file athlete_events.csv contains 271116 rows and 15 columns. Each row corresponds to an individual athlete competing in an individual Olympic event (athlete-events).

Our additional data is retrieved from the world bank, world development indicators. We selected all 266 countries' dataset, with features including country name, country code, series name, series code, current GDP, GDP growth, GDP per capita, population, population growth and land area. The selection is based on 'The Olympics and Economics 2012' by Goldman Sachs

Data Description

1. **Athlete dataset**

Column Names	**Description**	**Other**
ID	Unique number for each athlete	
Name	Athlete's name	
Sex	M or F	Sexual instinct of the athlete
Age	Integer	The athlete’s age
Height	In centimeters	The athlete’s height
Weight	In kilograms	The athlete’s weight
Team	Team Name	
NOC	National Olympic Committee 3-letter code	
Games	Year and season	
Year	Integer	
Season	Summer or Winter	
City	Host City	
Sport	Sport	Sport Type
Event	Event	Specific event of the Sport
Medal	Gold, Silver, Bronze, or NA	

2. **Region dataset**

Column Names	**Description**	**Other**
NOC	National Olympic Committee 3 letter code	
region	Country name (matches with regions in map_data("world"))	

notes	Notes	

3. **World Bank dataset**

Column Names	**Description**	**Other**
Country name	Annual population growth rate.	Population is based on the de facto definition of population, which counts all residents regardless of legal status or citizenship.
Country code	Country code	
Series name	Name of series	
Series code	Represented code	
Current GDP	GDP at purchaser's prices is the sum of gross value added by all resident producers in the economy plus any product taxes and minus any subsidies not included in the value of the products.	It is calculated without making deductions for depreciation of fabricated assets or for depletion and degradation of natural resources. Data are in current U.S. dollars. Dollar figures for GDP are converted from domestic currencies using single year official exchange rates. For a few countries where the official exchange rate does not reflect the rate effectively applied to actual foreign exchange transactions, an alternative conversion factor is used.
GDP growth	Annual percentage growth rate of GDP at market prices based on constant local currency.	Aggregates are based on constant 2015 prices, expressed in U.S. dollars. GDP is the sum of gross value added by all resident

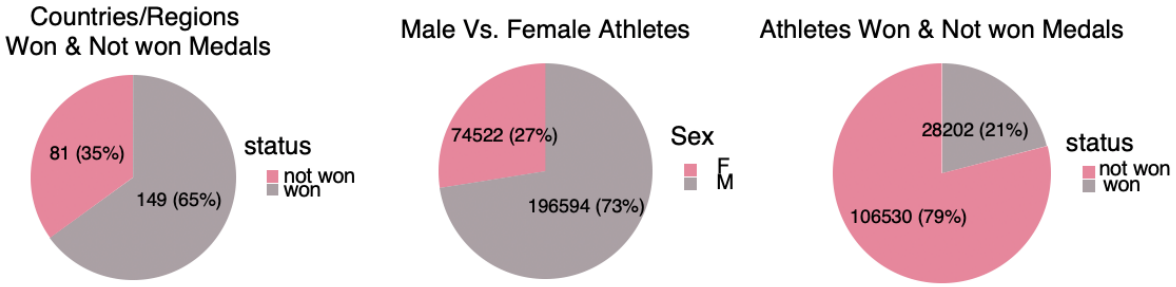
		producers in the economy plus any product taxes and minus any subsidies not included in the value of the products. It is calculated without making deductions for depreciation of fabricated assets or for depletion and degradation of natural resources.
GDP per capita	GDP per capita is gross domestic product divided by midyear population.	GDP is the sum of gross value added by all resident producers in the economy plus any product taxes and minus any subsidies not included in the value of the products. It is calculated without making deductions for depreciation of fabricated assets or for depletion and degradation of natural resources. Data are in current U.S. dollars
Population	Total population is based on the de facto definition of population, which counts all residents regardless of legal status or citizenship.	The values shown are mid-year estimates.
Population growth	Annual population growth rate.	Population is based on the de facto definition of population, which counts all residents regardless of legal status or citizenship.
Land area	Land area is a country's total area, excluding area under inland water bodies, national claims to continental shelf, and exclusive economic zones.	In most cases the definition of inland water bodies includes major rivers and lakes.

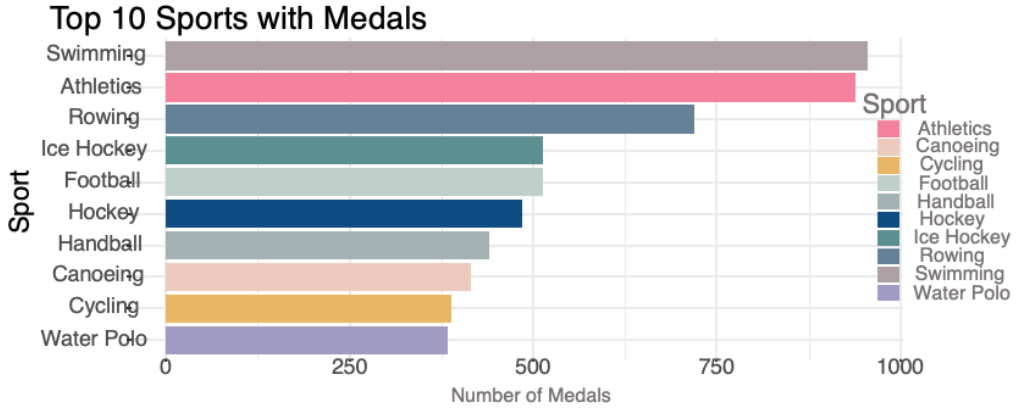
Data preprocessing and transformation

After exploring the 15 attributes of our athlete dataset, we found 23133 NAs in the Medal attribute. This may be because many athletes don't earn medals, we replaced these NAs with 'No medals'. Since neither the athlete dataset nor the region dataset has NAs in the NOC, we merged these two datasets using 'NOC'. We then noticed that 98% of the notes had missing data, therefore we dropped the 'notes'. Our 'Medal' variable includes 'Gold', 'Silver', 'Bronze' and 'Nomedal'. In order to explore and predict easier, we changed these to dummies. That is, we replaced the 'Medal' with 4 columns: 'Gold', 'Silver', 'Bronze' and 'Nomedal', containing only 0s and 1s.

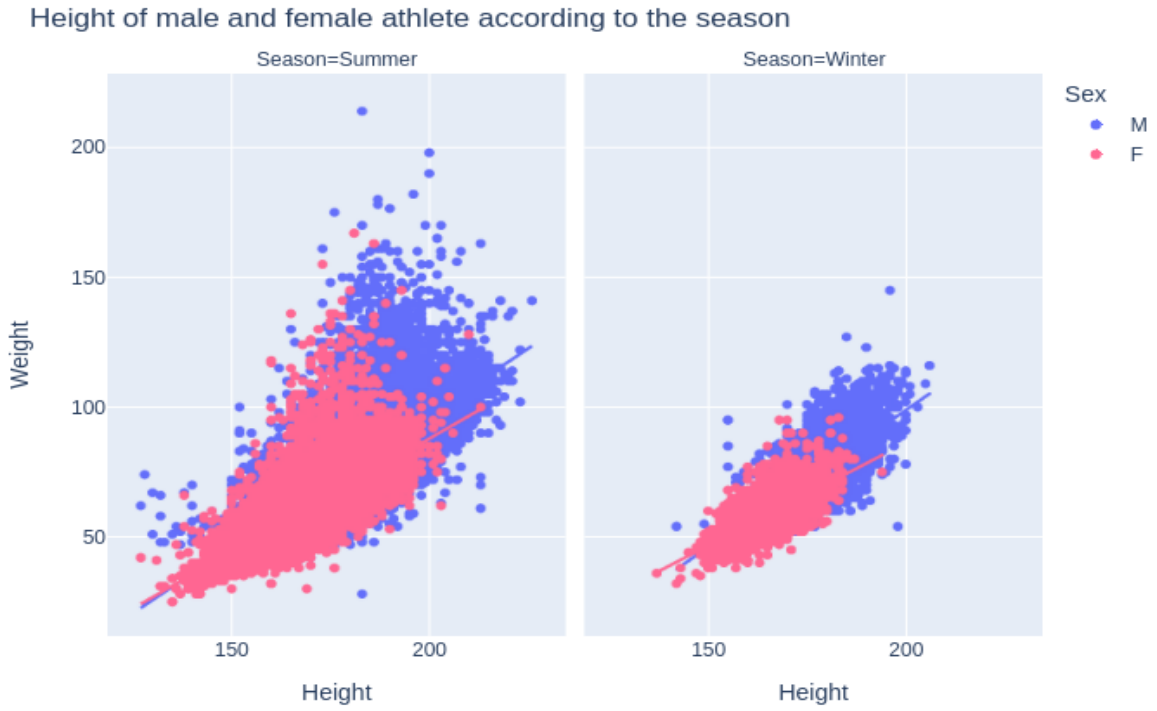
Exploratory Data Analysis

First, we are interested in overall information about athletes and medals. The pie chart shows that 35% of countries have never won any medals. There are more than twice as many male athletes as female athletes. And 79% of athletes have competed but never won any medals. In other words, the distribution of medal-winning and non-medal-winning countries is unbalanced, and the probability that we randomly select an athlete from the dataset and he or she never wins any medal is very high. We also have a general look of our top ten sports with medals. Swimming accounts for the most medals, this could be because of the team competition or the large number of events.

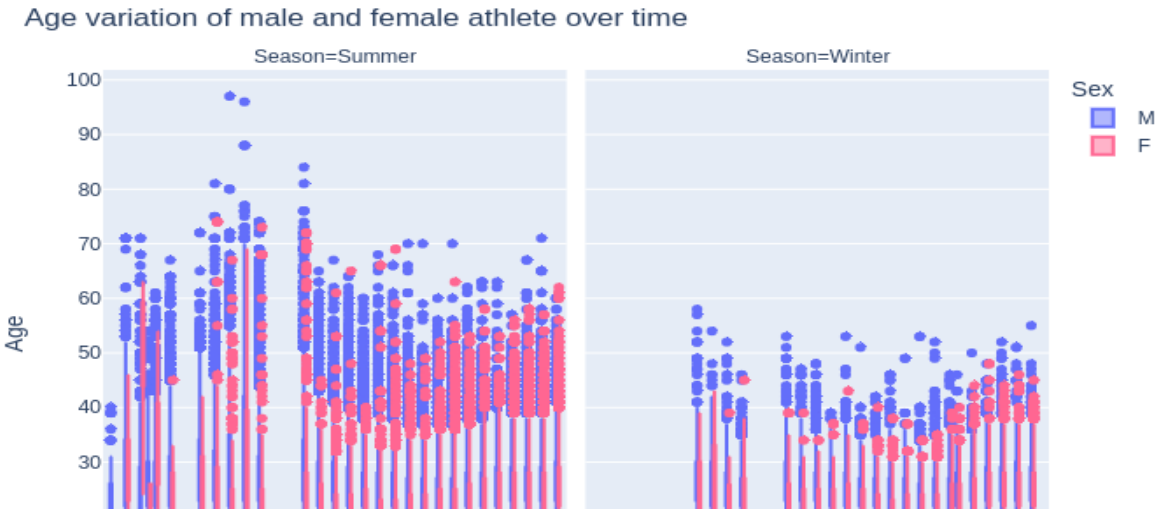




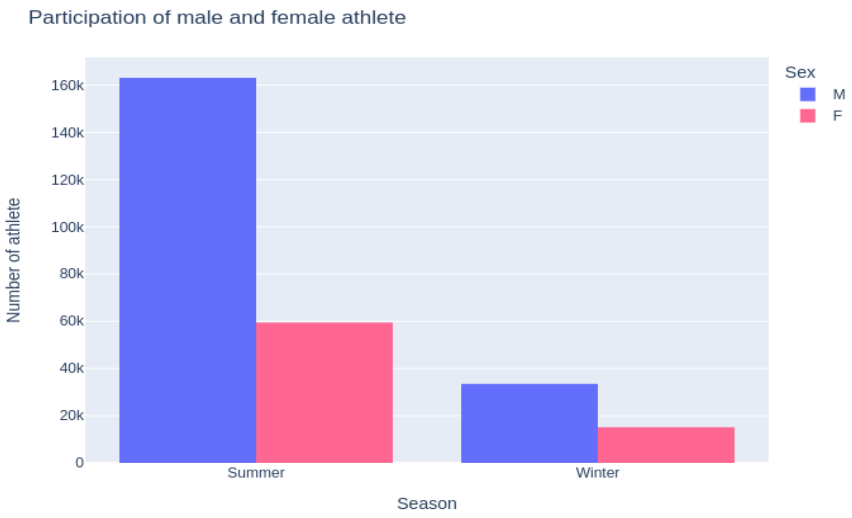
Now we are going to dive deep into the athletes' information, including the female and male. According to the figure below, compared to winter olympics, summer olympics had more athletes involved during the past 150 years. And summer olympics athletes tend to have a larger range of weights and heights, this probably because of the diverse number of sports types and the number of athletes who attended the event. The trend of different genders looks similar. In both winter and summer olympics, males weigh more and have greater height than females and summer olympics's athletes seem to have a wider distribution in weights and heights. Let's then explore the age information for both summer and winter olympics.



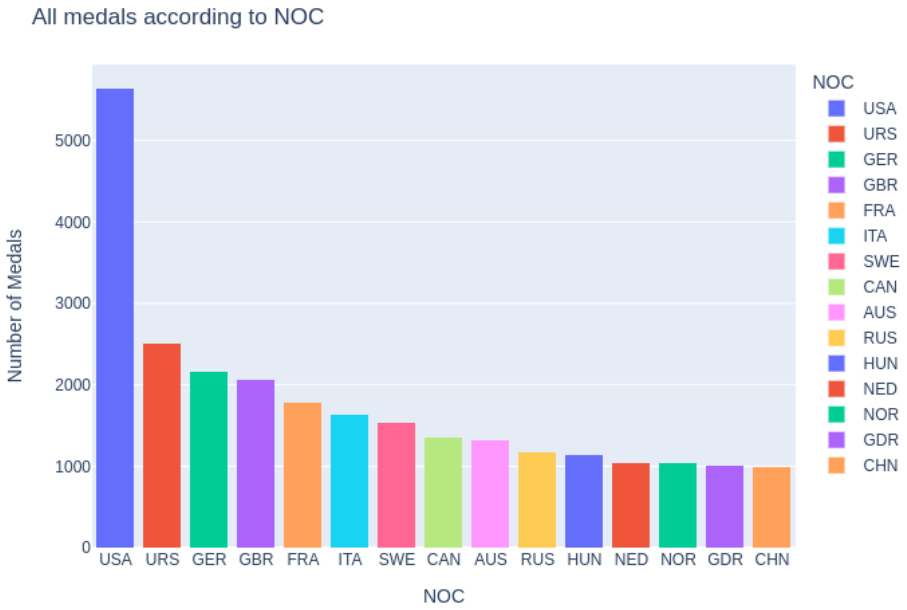
According to the graph below, it shows that the average age of male athletes is greater than female athletes. And the athletes who attended Summer olympics have more outliers in age.



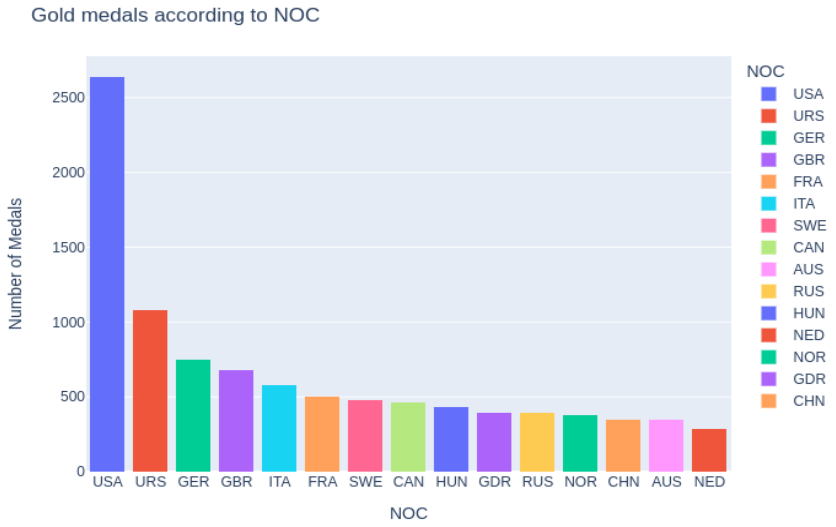
Next we can examine the participation of females and males during both summer and winter olympic games. During the winter olympics, female athletes' number is about a half of the male athletes. However, in the summer olympics, the number of male athletes is more than twice the number of the female athletes. It reminds us that we may think about separating the female and male dataset in future exploration.



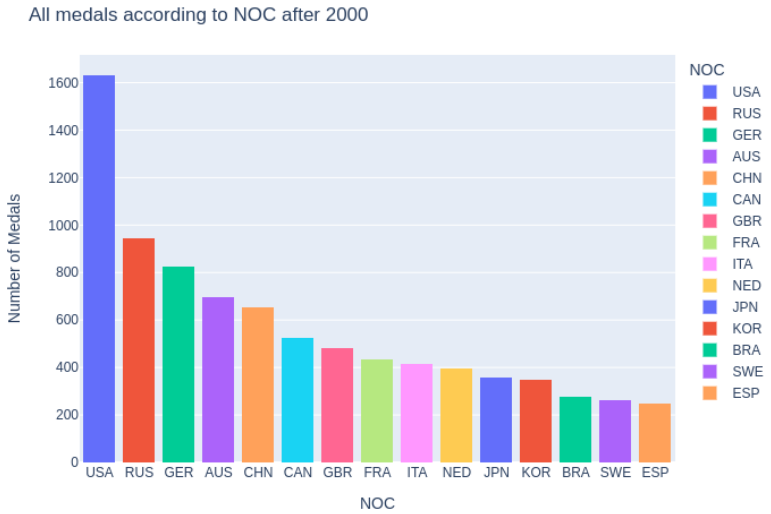
We have already seen a lot of information about the athletes. Then about the countries? Here we use NOC(National Olympic Committee) instead of countries/regions.We first take a look at the



overall medals of the 150 years. According to the graph below, the USA seems to be the big boss of the medalists, the number of medals is more than 5000, much more than the second, Soviet Union. Then what about the gold medal? Does it have the same distribution?



Well, the USA is still the big boss, and the number of gold medals is much more than the other countries. The top 3 stays the same as the number of medals.Therefore, the number of total medals (Includes gold, silver and bronze) that each top country achieved in the Olympics represent the number of gold medals in some way. However, these data are based on the history of 150 years. How about now? Let's go a little bit more to 2000 and see if there are any changes. The USA is still at the top while other countries are getting better. The gap between these countries is narrowing. We can also see China is making big progress, ranking from 15 to 5. distribution according to NOC after 2000. In future model prediction, we may consider only the total number of medals instead of analyzing the medals separately.



Then, we focus back to medals. From the analysis above, we already knew the top 15 countries of total medals. To be specific, we first look into these countries' medals' categories. USA not only has the maximum total medals, it also has the highest number of gold, silver and bronze medals respectively.

Medals by Country [Top 15]

NOC	Bronze	Silver	Gold	Total
USA	1358	1641	2638	5637
URS	689	732	1082	2503
GER	746	674	745	2165
GBR	651	739	678	2068
FRA	666	610	501	1777
ITA	531	531	575	1637
SWE	535	522	479	1536
CAN	451	438	463	1352
AUS	517	455	348	1320
RUS	408	367	390	1165
HUN	371	332	432	1135
NED	413	340	287	1040
NOR	294	361	378	1033
GDR	281	327	397	1005
CHN	292	347	350	989

Then, we look into the medals of these countries after 1960 to see if there is a difference between 120 years' records and recent years. The USA still has the top 1 medals and some Countries like China, have more medals after 1960.

Medals by Country after 1960 [Top 15]

NOC	Bronze	Silver	Gold	Total
USA	836	1042	1628	3506
URS	536	549	884	1969
GER	488	417	489	1394
RUS	399	360	389	1148
AUS	450	390	287	1127
GDR	281	327	397	1005
CHN	292	347	350	989
CAN	310	310	324	944
ITA	349	317	249	915
GBR	301	281	268	850
FRA	333	253	247	833
JPN	322	244	212	778
NED	242	249	222	713
HUN	209	190	237	636
KOR	180	231	221	632

We also look into the different season's games. We can first find that the total number of medals in Summer are much larger than that in Winter. The USA got the most medals not only in Summer but also in Winter. However, it is interesting that although the USA has the most medals in winter, the country that has the most Gold medals is Canada. Also, we can clearly find that most of the top 15 countries in Summer are totally different from those in Winter.

Medals by Country: Summer Olympics Game [Top 15]

NOC	Bronze	Silver	Gold	Total
USA	1197	1333	2472	5002
URS	596	635	832	2063
GBR	620	729	636	1985
GER	649	538	592	1779
FRA	587	575	465	1627
ITA	454	474	518	1446
AUS	510	452	342	1304
HUN	363	328	432	1123
SWE	358	396	354	1108
NED	371	302	245	918
CHN	258	317	334	909
RUS	331	278	296	905
JPN	333	287	230	850
GDR	227	277	339	843
CAN	344	239	158	741

Medals by Country: Winter Olympics Game [Top 15]

NOC	Bronze	Silver	Gold	Total
USA	161	308	166	635
CAN	107	199	305	611
NOR	127	165	151	443
URS	93	97	250	440
SWE	177	126	125	428
FIN	215	145	66	426
GER	97	136	153	386
AUT	103	98	79	280
SUI	129	70	76	275
RUS	77	89	94	260
ITA	77	57	57	191
GDR	54	50	58	162
TCH	75	81	2	158
FRA	79	35	36	150
NED	42	38	42	122

Feature Selection

We have explored two topics to learn about the Athletes data, first is to predict which athlete would get the medals during the Olympics, second is to predict the number of medals for the countries who had earned medals.

Feature Selection for Topic I

According to the number of medals analysis, we found that China has made great progress in recent years compared with other countries, especially in the olympics. Therefore, we decided to dive deep into the China Olympic data to predict which athlete would get the medals during the Olympics, In order to better fit the first experiment model, we did the encoding of our data, we encoded summer Olympics to 0 and winter olympics to 1, male athletes to 1 and female athletes to 0. We encoded those who got the medal during the Olympics to 1 and those with no medals to 0. We initially used one hot encoding method to translate the 46 types of sport features in order to apply PCA for the feature Selection. However, the cumulative proportion of variance reached 90% until the PC43, the result didn't help much with the selection.

To solve this problem, our first method was to divide our sport types into four groups named Four Sport Disciplines based on Isometric, Isotonic and Cardiac Remodeling. Our second

method was to fit the random forest model for both the original sports type dataset and the four sports dataset to calculate the feature importance. According to our result (see appendix.2), we think that it is hard to drop this column because of practical significance except the season.

Feature Selection for Topic II

Our goal for the second experiment was to predict the number of medals for the countries who had earned medals during the summer olympics. Therefore, we filtered all the summer olympics data, and now the personal information of the athlete was not useful according to existing literature, so we decided to drop Height and Weight of the athletes. Every team member in the team events will get a medal, this would cause the inaccuracy of the medal number prediction. We removed the duplicates to make sure in each event, unique medals were achieved. Then we found the number of athletes and the number of medals earned for each country by year of the olympics. Based on the above data, we also added the normalized athlete numbers for each country and the medals they earned in the last games to provide better prediction.

	1	2	3	4	5	6	7	8	9	10	11
Standard deviation	2.0271	1.4224	1.1211	0.9777	0.9364	0.8058	0.7132	0.6253	0.4093	0.2571	0.0564
Proportion of variance	0.3733	0.1838	0.1142	0.0868	0.0797	0.0590	0.0462	0.0355	0.0152	0.0060	0.0003
Cumulative proportion	0.3733	0.5571	0.6713	0.7581	0.8378	0.8968	0.9430	0.9785	0.9937	0.9997	1.0000

We can see the Principle Component Analysis output using all variables we add after normalization.. Note that the first seven components capture 94.3% of the total variation. According to the table below, GDP Growth and population growth features seem less important but not obvious enough, therefore we think all the features are essential to fit the model at this moment.

	1	2	3	4	5	6	7
Year	0.036027	-0.674724	0.147069	-0.087501	-0.042709	-0.081353	-0.003326
Athletes	0.462153	0.026905	-0.107469	0.036806	0.112175	-0.140788	-0.107404
Athletes_Normalized	0.458769	0.054160	-0.114729	0.040407	0.118072	-0.139929	-0.119016
Medals_Last_Games	0.451617	0.041077	0.000365	0.032781	0.003911	-0.220304	0.012073
Total_Medals_Year	0.035388	-0.670601	0.152879	-0.121594	-0.023330	-0.107656	-0.015618
GDP	0.377217	-0.026120	0.011898	0.198757	-0.122884	-0.108670	0.743688
GDP_Growth	-0.024101	0.008855	0.539864	0.411780	0.729803	0.042899	-0.012488
GDP_Per_Capita	0.131401	-0.248145	-0.368626	0.658055	-0.134577	0.513185	-0.236300
Area	0.352333	0.091704	0.302809	-0.153707	-0.182227	-0.046819	-0.574146
Pop_Growth	-0.144264	0.115069	0.472693	0.478430	-0.584842	-0.289484	-0.036755
Pop	0.253667	0.075571	0.433215	-0.280840	-0.173267	0.728598	0.183956

Topic Experiments I

In order to identify whether the Chinese athlete will win the medal in the Olympics, we decided to apply the Logistic Regression, Decision Tree and Random Forest model to our data with selected features. Note that the season feature was removed only in the logistic regression model. We found that the removal of the season doesn't improve the overall performance and according to our PCA result, season still plays an important role in the model fitting process.

We splitted both two datasets into training and validation datasets with testing size equal to 0.2. We first fitted a logistic regression for four sports dataset. Below we can see the summary of this regression. And then we calculated the accuracy of our regression and plotted the confusion matrix. The accuracy on the training dataset is 0.8 and on the validation is 0.79 which seems that our model performs well. However, when we looked into the confusion matrix, it seems that our model classified all target values into one category. Thus, we decided to change our threshold.

Results: Logit

Model:	Logit	Pseudo R-squared:	0.067
Dependent Variable:	Medal	AIC:	3667.4778
Date:	2022-03-28 00:08	BIC:	3723.9010
No. Observations:	3902	Log-Likelihood:	-1824.7
Df Model:	8	LL-Null:	-1954.8
Df Residuals:	3893	LLR p-value:	1.2455e-51
Converged:	1.0000	Scale:	1.0000
No. Iterations:	7.0000		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
Sex	-0.6040	0.1014	-5.9549	0.0000	-0.8028	-0.4052
Age	0.0020	0.0110	0.1857	0.8527	-0.0195	0.0235
Height	-0.0044	0.0077	-0.5735	0.5663	-0.0194	0.0106
Weight	-0.0025	0.0056	-0.4500	0.6527	-0.0136	0.0085
Year	0.0163	0.0044	3.7190	0.0002	0.0077	0.0248
Endurance	-33.7663	8.6151	-3.9194	0.0001	-50.6516	-16.8811
Mixed	-32.7959	8.6058	-3.8109	0.0001	-49.6630	-15.9288
Power	-32.7112	8.6122	-3.7983	0.0001	-49.5908	-15.8317
Skill	-32.0276	8.6083	-3.7206	0.0002	-48.8995	-15.1557

Figure. Logistic Regression Summary (threshold=0.5)

We first changed the threshold to 0.4. This time, the accuracy on the training dataset is 0.92 and on the validation is 0.93. It performed much better than the model that threshold equal to 0.5. And had a more reasonable confusion matrix. Then, we used the decision tree to fit the original sports dataset. We started with an initial guess of the parameters, getting an initial score 0.8413. We adapted the grid according to the result below and fitted it again. We got an improved score, 0.8442.We calculated the accuracy of our model and plotted the confusion matrix. The accuracy on the training dataset is 0.863 and on the validation is 0.85, the Area under curve is 0.804, the Decision Tree model is overall a good fit.

Lastly, we applied a random forest model to both our original dataset and four sports dataset. After applying to our original dataset, we can see our Mean Squared Error for the train dataset is 0.037 and the Mean Squared Error for the test dataset is 0.115. We found a little bit of overfitting here. Let's look at the report of our training and testing dataset. According to the precision and recall, the model fits well with the training dataset while the testing dataset is not as good as the training. But the model is fine in general.We then tried the Four Sports Discipline dataset for better fitting. Again, we applied the random forest model to the dataset. And we can see from the table that the training dataset performs better than using the original dataset, but the testing dataset performs worse. The overfitting problem still exists.

From our previously trained models, it is difficult to interpret the coefficients and implications of these models, although some of them have good fit and performance. We trained our model by athletes, but our data set of athletes was not very adequate. We only have the weight and height of the athletes, but both height and weight have NA values. In addition, different sports have different athlete criteria, and our sports discipline cannot fully explain the differences in these

sports. Therefore, we decided to change the dimensions of the model, train our model by country and predict the number of medals for each country.

Topic Experiments II

In this experiment, we would like to identify how many medals each country will earn in the next Summer Olympics. After considering all the existing and newly imported data, we decided to apply a commonly used type of predictive analysis, that is, linear regression with our selected features. We wanted to find out if these new selected features did a good job in predicting the medals and which features are significant in predicting the outcome medals.

Based on the R squared value in our OLS regression results, our model fits well; 89.2% variation in the number of medals can be explained by those features. Our Prob(Fstatistics) tells about the overall significance of our model, here the zero shows that our regression is meaningful. The Durbin -Watson 1.8 also implies that the regression results are reliable from the interpretation side of this metric. Although the linear regression model fits well, there is one issue that needs attention.

OLS Regression Results

Dep. Variable:	Medals	R-squared:	0.892
Model:	OLS	Adj. R-squared:	0.891
Method:	Least Squares	F-statistic:	772.7
Date:	Mon, 18 Apr 2022	Prob (F-statistic):	0.00
Time:	00:49:05	Log-Likelihood:	-3041.6
No. Observations:	1036	AIC:	6107.
Df Residuals:	1024	BIC:	6166.
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	137.6520	156.479	0.880	0.379	-169.405	444.709
Year	-0.0736	0.081	-0.904	0.366	-0.233	0.086
Athletes	0.0204	0.019	1.048	0.295	-0.018	0.059
Athletes_Normalized	2.9476	1.801	1.637	0.102	-0.586	6.481
Medals_Last_Games	0.5261	0.020	26.070	0.000	0.486	0.566
Total_Medals_Year	0.0098	0.008	1.304	0.193	-0.005	0.025
GDP	1.4129	0.251	5.636	0.000	0.921	1.905
GDP_Growth	0.0052	0.023	0.231	0.818	-0.039	0.050
GDP_Per_Capita	-0.0204	0.011	-1.790	0.074	-0.043	0.002
Area	0.4322	0.114	3.799	0.000	0.209	0.655
Pop_Growth	0.0379	0.115	0.330	0.742	-0.188	0.264
Pop	3.6097	1.363	2.648	0.008	0.935	6.284

Omnibus:	1257.781	Durbin-Watson:	1.863
Prob(Omnibus):	0.000	Jarque-Bera (JB):	522352.653
Skew:	5.617	Prob(JB):	0.00
Kurtosis:	112.429	Cond. No.	2.40e+06

count	1036.000000	**count**	193.000000
mean	4.716216	**mean**	4.683728
std	13.138669	**std**	13.459447
min	-2.950336	**min**	-4.477466
25%	-0.472288	**25%**	-0.722397
50%	0.147731	**50%**	-0.054519
75%	3.648837	**75%**	4.267740
max	127.601884	**max**	103.625586

If we take a closer look at our prediction table, our minimum prediction and the 25 percent confidence interval numbers of both dataset are not positive. Most of our predictions are less than one that means most countries are not predicted to win a medal. It empirically shows that our prediction for the number of medals can suffer from the negative number of medals, which degrades the model performance.

Thus, we decided to apply a two-stage algorithm to our dataset. First, the classification algorithm was applied to the dataset to figure out which countries would get the medals in the Olympics and which countries would never. That is, we wanted to create a boundary between the training instances. Second, we trained the regression model on the samples that were predicted to have the potential to win a medal in the first stage.

Stage I. Classifier

We first applied a logistic regression model to our dataset. The models' confusion matrix and accuracy scores of training and validation dataset are shown below. The logistic model performs well, achieving 0.9 accuracy on training dataset and 0.89 accuracy on validation dataset. And from the confusion matrix, the number of each unit is reasonable. Most of the samples in both dataset were classified correctly. Thus, we set this model as the baseline model that other models we trained later would compare with it.

Logistic Regression		prediction	
Training(Accuracy **0.9025**)		0	1
Actual	0	612	25
	1	76	325
		prediction	
Validation(Accuracy **0.8860**)		0	1
Actual	0	104	8
	1	14	67

To better capture the division, leading to superior classification performance, we decided to generate decision boundaries in other ways by applying the tree model, with both bagging (Random Forest) and boosting method (Adaboost). The Randomforest model performs better than the AdaBoost model, with 0.93 accuracy in training data and 0.865 accuracy in the validation set. However, we can still observe some overfitting in both of these models. One may attempt to fix this problem by applying the Gussian Naive Bayes model. The graph below indicates that the overfitting issue has been improved while the Accuracy of the Naive Bayes model is less. Furthermore, the Naive Bayes model assumes that all the features are independent and this may not hold for our features since the number of medals won last year directly influences the number of medals won this year. We use Multilayer Perceptron, an artificial neural network model to improve, with its high degree of connectivity by the synapses of the network. The Accuracy score of the training and the testing dataset shows the performance is better than Gaussian Naive Bayes, and the overfitting issue is not obvious.

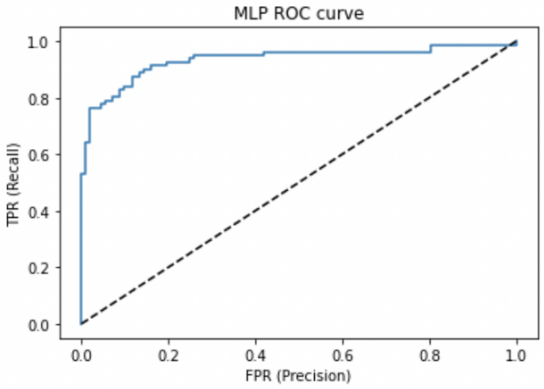
Random Forest		prediction	
Training(Accuracy **0.9344**)		0	1
Actual	0	621	16
	1	52	347
		prediction	
Validation(Accuracy **0.8653**)		0	1
Actual	0	101	11
	1	15	66

AdaBoost		prediction	
Training(Accuracy **0.9208**)		0	1
Actual	0	608	29
	1	53	346
		prediction	
Validation(Accuracy **0.8497**)		0	1
Actual	0	94	18
	1	11	70

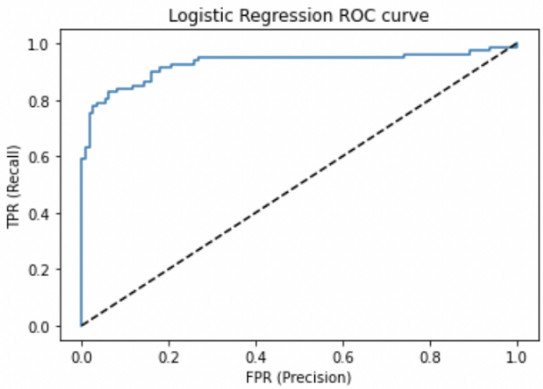
Multilayer Perceptron		prediction	
Training(Accuracy **0.8996**)		0	1
Actual	0	610	27
	1	77	322
		prediction	
Validation(Accuracy **0.8705**)		0	1
Actual	0	104	8
	1	17	64

Gaussian Naïve Bayes		prediction	
Training(Accuracy **0.8880**)		0	1
Actual	0	616	21
	1	95	204
		prediction	
Validation(Accuracy **0.8653**)		0	1
Actual	0	103	9
	1	17	64

Then we want to compare the Multilayer Perceptron model with the logistic regression model. Logistic Regression is better fit to our model based on several reasons: first, the accuracy of both the training and the testing dataset is higher. Second, MLP, as an artificial neural network, the duration is usually unknown and the cost is higher than the logistic regression model. Third, Logistic Regression is easier to implement, which is beneficial to our application of the next step's regression algorithm.



Area under curve (AUC): 0.9355158730158729



Area under curve (AUC): 0.9286816578483246

Stage II. Regressor

According to the models implemented on stage I, we decided to fit the logistic regression to our dataset which eliminated 118 countries out of 193 that would not win any medals in the Olympics. After that, we use regression models to predict the medals that the remaining countries would gain. These models include Linear Regression, Lasso and Ridge Regression, Support Vector regression, Poisson Regression and Random Forest.

We splitted both two datasets into training and validation datasets with Olympics data from 1988 to 2008 as training data, Olympics data of 2012 as validation set, we will then predict the number of medals of the top countries with medals and compare with the 2016 real data.

We first decided to apply a commonly used Linear Regression again. After classification, less than 25% of the countries won negative medals and according to our results, 87.6% variation in the number of medals can be explained by those features. Our Prob(Fstatistics) tells about the overall significance of our model, here the number approximately to zero shows that our regression is meaningful. The Durbin -Watson 1.9 also implies that the regression results are reliable from the interpretation side of this metric. Although the overall fitting is good, we are still trying to find ways to improve the model performance. We want to introduce a little bias so that the variance can be reduced, which leads to lower MSE and RMSE. Here the Lasso and Ridge Regression is introduced.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 Medals   R-squared:                       0.876
Model:                            OLS   Adj. R-squared:                  0.873
Method:                 Least Squares   F-statistic:                     260.0
Date:                Mon, 18 Apr 2022   Prob (F-statistic):          2.56e-175
Time:                        00:52:06   Log-Likelihood:                -1399.2
No. Observations:                 415   AIC:                             2822.
Df Residuals:                     403   BIC:                             2871.
Df Model:                          11
Covariance Type:            nonrobust
===========================================================================================
                              coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------
const                     251.5252    401.639      0.626      0.532    -538.043    1041.094
Year                       -0.1359      0.209     -0.651      0.516      -0.547       0.275
Athletes                    0.0122      0.036      0.341      0.734      -0.058       0.082
Athletes_Normalized         4.6696      3.287      1.421      0.156      -1.792      11.131
Medals_Last_Games           0.4981      0.033     15.320      0.000       0.434       0.562
Total_Medals_Year           0.0214      0.019      1.108      0.268      -0.017       0.059
GDP                         1.6191      0.413      3.923      0.000       0.808       2.430
GDP_Growth                 -0.0119      0.072     -0.166      0.868      -0.153       0.129
GDP_Per_Capita             -0.0910      0.035     -2.565      0.011      -0.161      -0.021
Area                        0.4197      0.189      2.215      0.027       0.047       0.792
Pop_Growth                 -0.1276      0.374     -0.341      0.733      -0.864       0.608
Pop                         3.3166      2.218      1.495      0.136      -1.044       7.677
==============================================================================
Omnibus:                      372.840   Durbin-Watson:                   1.900
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            28266.875
Skew:                           3.352   Prob(JB):                         0.00
Kurtosis:                      42.872   Cond. No.                     2.50e+06
==============================================================================
```

However, after comparing the Lasso regression and the Ridge regression with our linear regression, the linear regression performs better at prediction of our training and testing data. We will keep the linear regression and sometimes a simple linear model informed by theory can trounce a Ridge Regression Model. While the linear Regression model minimizes the error between the actual and the predicted, the Support Vector Regression fits the best line within a threshold of values. Let's see how it performs.

Regressor	Training RMSE	Validation RMSE	R-squared
Linear Regression	7.04	4.28	0.8764
Lasso Regression	7.06	4.32	0.8765
Ridge Regression	7.19	4.45	0.8713
Support Vector Regression	18.28	18.03	0.1689
Poisson Regression	9.06	6.19	0.8021
Random Forest	6.09	4.73	0.9077

Based on the above result, we can see the performance of Support Vector regression is not good enough. It has extremely high RMSE in both training and validation dataset. If we want to improve this model's performance, it may take time to tune the parameters carefully. Therefore, we may not use this model after. Then, we applied poisson regression to our dataset because it is a classic method of modeling counts values, which is, in our cases, medals counts. From the performance above, the poisson regression is not that well compared with our baseline model, linear model. Both training and validation RMSE are higher than linear regression and its R-squared is lower than our baseline model.

At this point, we tried to reduce the overfitting and at the same time applied the logic that our underlying function is not truly linear and our selected features are not that many. Random Forest model was applied to our data set. As we can see from the result table, more than 90% variation in the number of medals can be explained by those features. The training MSE of the Random Forest model is less than that of the linear Regression. Besides, the overfitting problem has been improved. Therefore, we choose the RandomForest model at last.

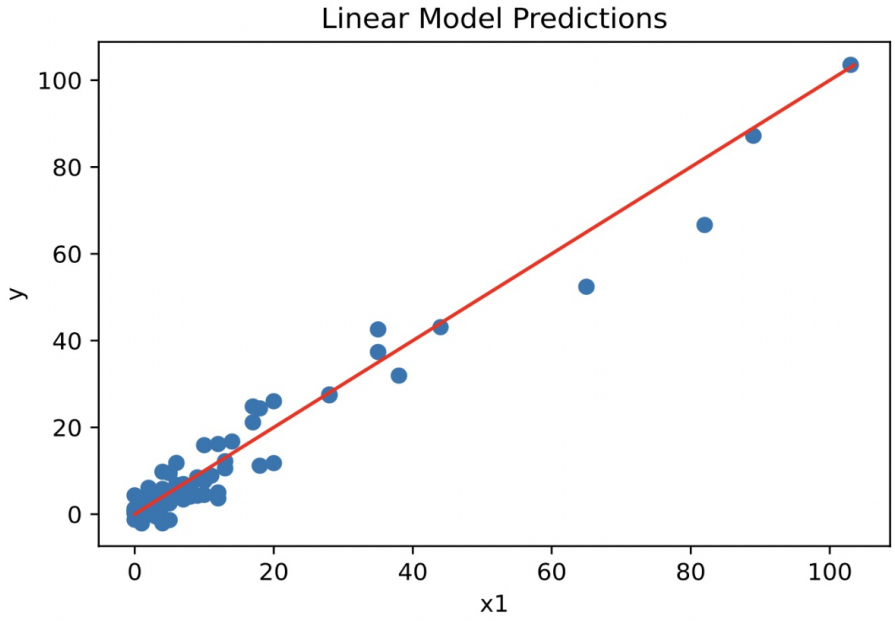


Figure. Baseline results for prediction of medals (y-axis) vs actual medals won (x-axis) in 2016 Summer Olympics Games

Let's now look at our predicted data with real world data, the Real world data for both models predicts really well.

Linear Regression Random Forest

Country	Actual	Predicted
France	35	37
China	89	86.7
Finland	3	3.2
Norway	4	5
Romania	9	8

Country	Actual	Predicted
France	35	34
China	89	87
Finland	3	4
Norway	4	6
Romania	9	10