Olympic Prediction

Janelle Anderson (square), Billy Cartwright (X), Kristina Castro (X), Matthew Price (triangle), Ricardo Robles (circle)

Project Overview and Purpose

Question to answer

Why do some countries outperform others in the Olympics?

What factors influence olympic performance at the country and athlete level?

What countries and athletes are most likely to have a stronger performance at the next olympics?

How we will answer

Using a dataset from the 2016 olympics, we seek to determine the most influential factors that relate to medal performance of olympic athletes, and build a predictive model that can forecast athlete performance for the next summer olympics.

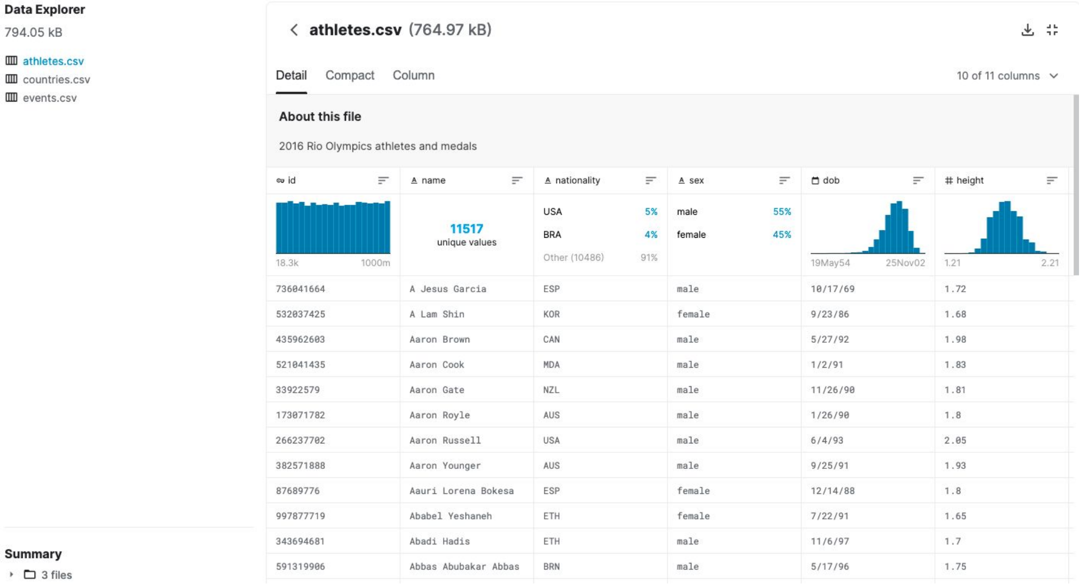
Why we ask...

This analysis could provide multiple investment opportunities related to sports/olympic consumer market sizing, athlete endorsements, sports analytics, and sports betting, to name a few possibilities.

Data Set

We used the official dataset outlining olympic events, athletes, participating countries, and medal counts for the 2016 Rio Olympics.

https://www.kaggle.com/rio2016/olympic-games

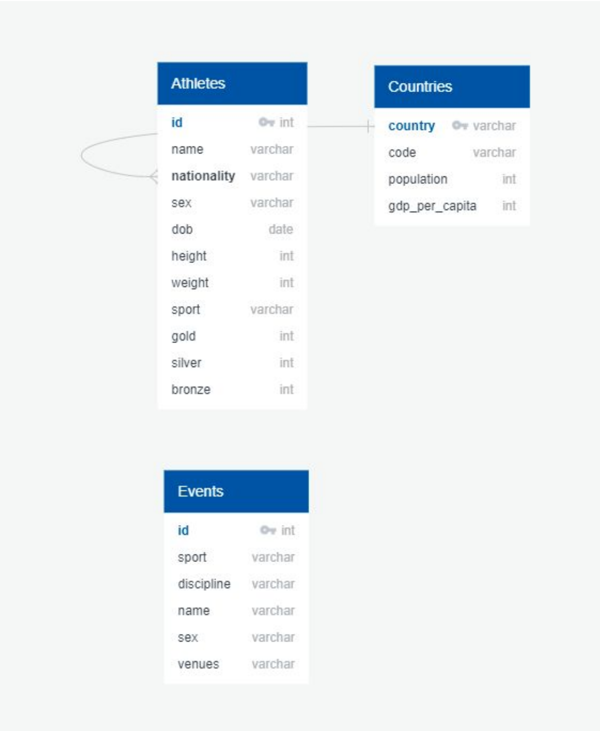


id	name	nationality	sex	dob	height
736041664	A Jesus Garcia	ESP	male	10/17/69	1.72
532037425	A Lam Shin	KOR	female	9/23/86	1.68
435962603	Aaron Brown	CAN	male	5/27/92	1.98
521041435	Aaron Cook	MDA	male	1/2/91	1.83
33922579	Aaron Gate	NZL	male	11/26/90	1.81
173071782	Aaron Royle	AUS	male	1/26/90	1.8
266237702	Aaron Russell	USA	male	6/4/93	2.05
382571888	Aaron Younger	AUS	male	9/25/91	1.93
87689776	Aauri Lorena Bokesa	ESP	female	12/14/88	1.8
997877719	Ababel Yeshaneh	ETH	female	7/22/91	1.65
343694681	Abadi Hadis	ETH	male	11/6/97	1.7
591319906	Abbas Abubakar Abbas	BRN	male	5/17/96	1.75

Data Cleaning and Analysis

- Python was the the main language used for data cleaning and analysis. Several packages utilized, including Pandas, NumPy, SciKitLearn, ImbalancedLearn, and DateTime functions.

Database & Exploration



Data Output Explain Messages Notifications

	id integer	name text	nationality text	sex text	dob text	height double precision	weight integer	sport text	gold integer	silver integer	bronze integer	population integer	gdp_per_capita double precision
1	736041664	A Jesu...	ESP	male	10/17/...	1.72	64	athletics	0	0	0	46418269	25831.5823052954
2	532037425	A Lam ...	KOR	female	9/23/86	1.68	56	fencing	0	0	0	50617045	27221.5240509661
3	435962603	Aaron ...	CAN	male	5/27/92	1.98	79	athletics	0	0	1	35851774	43248.529909341
4	521041435	Aaron ...	MDA	male	1/2/91	1.83	80	taekwon...	0	0	0	3554150	1848.06180430428
5	33922579	Aaron ...	NZL	male	11/26/...	1.81	71	cycling	0	0	0	4595700	37807.9672760442
6	173071782	Aaron ...	AUS	male	1/26/90	1.8	67	triathlon	0	0	0	23781169	56310.9629933721
7	266237702	Aaron ...	USA	male	6/4/93	2.05	98	volleyball	0	0	1	321418820	56115.7184261955
8	382571888	Aaron ...	AUS	male	9/25/91	1.93	100	aquatics	0	0	0	23781169	56310.9629933721
9	87689776	Aauri L...	ESP	female	12/14/...	1.8	62	athletics	0	0	0	46418269	25831.5823052954
10	997877719	Ababel ...	ETH	female	7/22/91	1.65	54	athletics	0	0	0	99390750	619.169406475891
11	343694681	Abadi ...	ETH	male	11/6/97	1.7	63	athletics	0	0	0	99390750	619.169406475891
12	591319906	Abbas ...	BRN	male	5/17/96	1.75	66	athletics	0	0	0	1377237	22600.2140981035
13	376068084	Abbey ...	USA	female	5/25/92	1.61	49	athletics	0	0	0	321418820	56115.7184261955
14	162792594	Abbey ...	USA	female	12/3/96	1.78	68	aquatics	1	1	0	321418820	56115.7184261955
15	521036704	Abbie ...	GBR	female	4/10/96	1.76	71	rugby se...	0	0	0	65138232	43875.9696143686
16	149397772	Abbos ...	UZB	male	7/7/98	1.61	57	wrestling	0	0	0	31299500	2132.07036847857
17	256673338	Abbub...	RSA	male	2/18/94	1.75	64	football	0	0	0	54956920	5723.97335690212
18	337369662	Abby Er...	NZL	female	11/20/...	1.75	68	football	0	0	0	4595700	37807.9672760442
19	334169879	Abd Elh...	EGY	male	6/3/89	2.1	88	volleyball	0	0	0	91508084	3614.74676616271
20	215053268	Abdala...	MAR	male	3/25/87	1.73	57	athletics	0	0	0	34377511	2878.20134215919
21	763711985	Abdalel...	QAT	male	1/1/97	1.85	80	athletics	0	0	0	2235355	73653.3944346574
22	924593601	Abdalla...	SUD	male	9/28/96	1.77	65	athletics	0	0	0	40234882	2414.72360102858
23	578032534	Abdel ...	EGY	male	12/10/...	1.76	80	shooting	0	0	0	91508084	3614.74676616271
24	890222258	Abdela...	MAR	male	2/27/93	1.9	72	athletics	0	0	0	34377511	2878.20134215919
25	803161695	Abdela...	ESP	male	8/30/91	1.75	67	athletics	0	0	0	46418269	25831.5823052954
26	189931373	Abdela...	SUD	male	10/12/...	1.81	72	aquatics	0	0	0	40234882	2414.72360102858
27	677622742	Abdelg...	ALG	male	1/29/89	1.85	75	football	0	0	0	39666519	4206.03123244958
28	349871091	Abdelh...	ALG	male	9/26/86	1.86	[null]	boxing	0	0	0	39666519	4206.03123244958
29	904808208	Abdelh...	ALG	male	5/10/94	1.86	70	football	0	0	0	39666519	4206.03123244958
30	23564778	Abdelk...	ALG	male	12/12/...	1.78	[null]	boxing	0	0	0	39666519	4206.03123244958
31	133974151	Abdelk...	ALG	male	3/19/93	1.85	79	football	0	0	0	39666519	4206.03123244958
32	189886442	Abdelk...	MAR	male	7/15/62	1.74	67	equestrian	0	0	0	34[illegible]	[illegible]

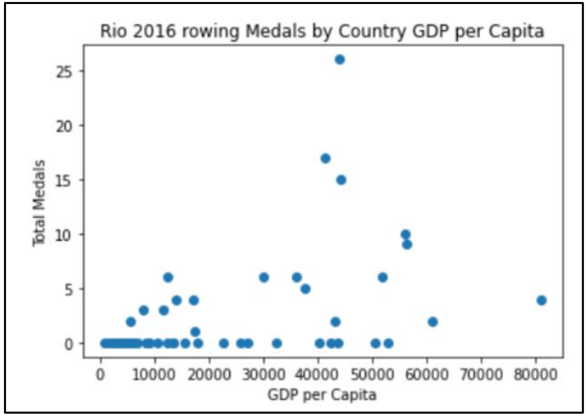
Successfully run. Total

We used a PostgreSQL relational database for data storage because it facilitates finding connections between different tables and answering questions with the results. Furthermore, the static database can be integrated with Pandas to perform ETL procedures.

Data Analysis

Overall, there appeared to be a small but significant correlation between a country's GDP per capita and their total medal counts, although the strength of the correlation varied between sports

Rowing
r = .47 --> moderate correlation



Overall
r = .37 --> low correlation

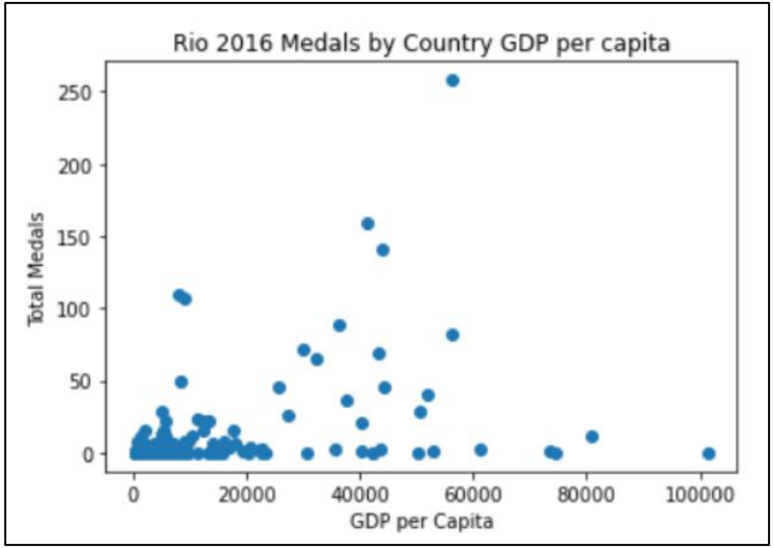
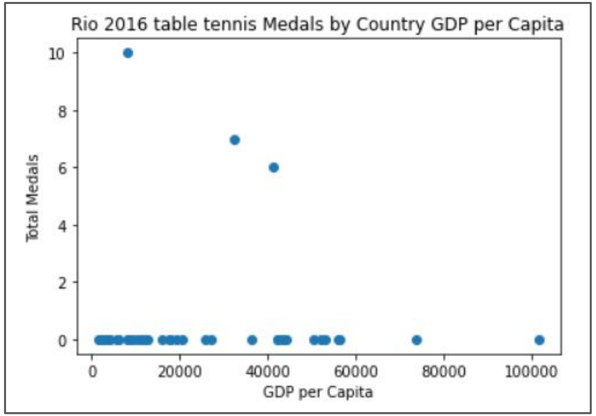


Table Tennis
r = -.002 --> no correlation



Note: Correlations calculated using the Pearson Correlation Coefficient (r)

Machine Learning Models

<u>**Approach 1**</u>

Initial approach using country GDP and athlete descriptors as independent variables, and a binary “placed” variable for the dependant variable

Independent Variables: Country GDP per Capita, Sport (dummy), Athlete Descriptors (Age, Weight, Height),
Dependant Variable: “Placed” Variable (binary output of 0 =didnt win a medal and 1= won a medal at Rio)
Observations: 10,109
Columns/Features: 33
Models: Logistic Regression and Random Forest

<u>**Approach 2**</u>

Same as Approach 1, but with added dummy variables for every country.

Independent Variables: Country (dummy variable), Country GDP per Capita, Sport (dummy), Athlete Descriptors (Age, Weight, Height)
Dependant Variable: “Placed” Variable (binary variable with 0 =didnt win a medal vs. 1= won a medal at Rio)
Observations: 10,109
Columns/Features: 198
Models: Logistic Regression and Random Forest

Approach 1

First pass without country dummy variables.

Model	Sampling Methods	Balanced Accuracy Score	Confusion Matrix	Precision	Recall	Specificity	F1
Logistic Regression	Naive Random Oversampling	0.5	[[2128 0] [400 0]]	0	0	1	0
Logistic Regression	SMOTE	0.6	[[1286 842] [162 238]]	0.22	0.6	0.6	0.32
Logistic Regression	Undersampling with Cluster Centroids	0.58	[[1204 924] [160 240]]	0.21	0.6	0.57	0.31
Logistic Regression	SMOTEENN	0.56	[[907 1221] [123 277]]	0.18	0.69	0.43	0.29
Random Forest	none	0.73	[[1541 587] [108 292]]	0.33	0.73	0.72	0.46
Random Forest	Easy Ensemble Classifier	0.64	[[1290 838] [128 272]]	0.25	0.68	0.61	0.36

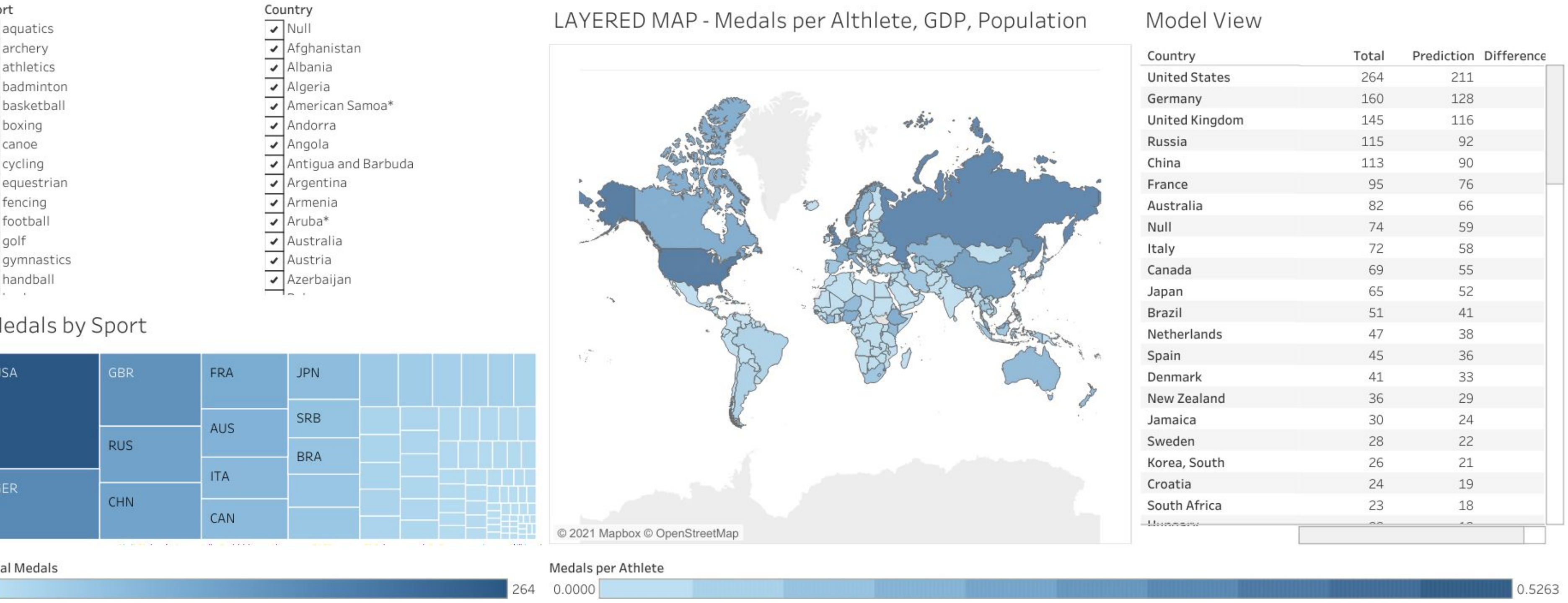
Approach 2

Added Country dummy variables improved the Random Forest accuracy scores

Model	Sampling Methods	Balanced Accuracy Score	Confusion Matrix	Precision	Recall	Specificity	F1
Logistic Regression	Naive Random Oversampling	0.5	[[2128 0] [400 0]]	0	0	1	0
Logistic Regression	SMOTE	0.6	[[1473 655] [199 201]]	0.23	0.5	0.69	0.32
Logistic Regression	Undersampling with Cluster Centroids	0.58	[[1204 924] [160 240]]	0.21	0.6	0.57	0.31
Logistic Regression	SMOTEENN	0.56	[[909 1219] [122 278]]	0.19	0.7	0.43	0.29
Random Forest	none	0.79	[[1676 452] [84 316]]	0.41	0.79	0.79	0.54
Random Forest	Easy Ensemble Classifier	0.68	[[1318 810] [104 296]]	0.27	0.74	0.62	0.39

Dashboard

The group created a Tableau dashboard to present the findings and allow users to sort and filter data based on their required analysis.



Recommendations

This analysis was done over a short period of time and with limited data.

- A major set back of this analysis is that we only had access to data for one year of Olympic games results. Access to additional years would have allowed us to further test predictions.
- The analysis would be stronger if we had time to build a more sophisticated model that looked at outliers (i.e. countries with low GDP per Capita but high performance.)

New question: could we use a similar model to predict number of new games being added to the Olympics