Using Data Mining to Predict the Winter Olympics Medal Counts in Sochi

By Dan Graettinger with Tim Graettinger

- Which nation will bring home the most medals at the upcoming Winter Olympics in Sochi, Russia?
- Will any nation from Africa, South America, or the Middle East finally break through and win a medal?[1]
- Why do some nations win a bundle of medals while others win only a few?
- Can data mining give us the answers to these questions?

This last question came into my mind four years ago after the Winter Games in Vancouver. As a data miner working with Discovery Corps, Inc., I use data about the past to predict the future all the time. We help businesses decide which potential customers are the most likely to want their product or service. We help non-profit organizations predict which small-dollar donors have the potential to become big-dollar donors. If an organization has data on the past, we can help them predict the future. So I knew that data mining techniques could give us an estimate of the number of medals each nation might win; but I wondered how close we could get to the actual outcomes.

It was a tantalizing project. My mind immediately began to analyze the problem. What is it about a nation that causes it to win medals at the Olympics – and would I be able to find data on those characteristics? Wealth had to play a part. A nation whose people are struggling to survive is not going to have many individuals with the leisure time for recreational pursuits like becoming world class in a sporting event. Also, geography might be part of the equation. I was going ***way*** out on a limb here, but I didn't think a nation like Western Sahara would probably bring home a lot of medals at the **WINTER** Olympics! The other thought that immediately struck me was that, in order to win a medal at a sport like downhill skiing, a nation has to have mountains. Clearly, I was going to need to start collecting data – as much as I could – about the nations of the world. (That is, after I got my boss's okay to pursue this project when we had some down time.)

What Kind of Data?

As data miners know, the data you expect to tell you the story isn't always the stuff that actually does the job[2], so I decided to cast my net as wide as I could, gathering as many different pieces of data as possible. I wanted all kinds of data on the nations of the world, even data that I didn't expect to be relevant to the outcome (See Appendix B for the list of data I eventually used.) And in fact, a column of data that I thought would be irrelevant and might easily have deleted turned out to be the single most useful variable in predicting the number of medals a nation would win! Fortunately, I was able to find data in many categories:

- Economic
- Population
- Human Development
- Geography
- Religion
- Politics and Freedom

Thankfully, there were some good sources out there[3], and I collected enough data that I felt I had a good chance to predict some meaningful outcomes. But would it be enough? There is more than one way to go about predicting the medal count at the Olympics, and the route before me was the "30,000 feet" approach. Far from having information on individual athletes in the various events, I would be working entirely from data about nations. Excellence in anything has a lot to do with individual motivation. Instead, I would be approaching the problem from perhaps the most aggregate viewpoint possible. Then again, what might I learn about nations while studying their ability to produce excellence? Yes, I could probably make better predictions if I had the resources of a news organization, gathering experts on every sport, predicting the winners in each, and summing them up into national totals. But that wouldn't tell me anything about the great questions – the 'Why?' questions. Why is a nation able to produce excellent individuals? What factors contribute to such success? If I found answers to these questions, perhaps those answers might cross over from athletic excellence to other areas of human endeavor: science and technology, the arts, theology, etc. Well … that was getting way beyond the original scope of the project. For the time being, I would just focus on predicting the nations' medal counts in Sochi.

Building the Models

Once I had married the data on the nations to their medal counts in the last two Winter Games, my team at Discovery Corps and I could begin exploring it and preparing to build a predictive model. We decided that we would first use a logistic regression to predict which nations would win at least one medal and which would come home empty-handed[4]. As we got the results from profiling each variable against our outcome (medals > 0), immediately the most useful variable of the bunch showed itself – and it was a real shock! I had dreamed up this project after watching the Winter Olympics, but I knew I'd have to wait four years for my chance to predict the outcome at the next Winter Games. So we decided in the interim to predict the medal counts at the London Summer Games of 2012[5]. When we picked up the data again this year to make our Winter predictions, my subconscious data miner's habit of not deleting data kept me from removing the column of medal counts from the summer games. To our shock, the medal count from the preceding summer games was the best variable for predicting a nation's medal count in the winter games! At the last two Winter Games, no nation won a medal without having won at least one medal in the preceding Summer Olympics. I never expected that! Our predictive model would ultimately fill in a zero for the anticipated medal count in Sochi if the nation did not win a medal in London. Also during the profiling stage, we saw other variables rise to the top: migration rate, doctors per thousand people, latitude of the capital city, value of the nation's exports, and some measures of gross domestic product. Ultimately, once we built our logistic model, it had a 96.5% correct rating. Not too shabby! (Correct predictions included those instances where we predicted the nation *would* win a medal *and it did* as well as instances where we predicted a nation *would not* win a medal *and it didn't*. All others outcomes were 'misses'.)

Since our goal was to predict how many medals each country would win, we needed to go beyond the binary outcome the logistic regression used (simply whether the nation would win a medal or not). So we decided to create a linear regression model that would predict actual medal counts. And for readers who are interested in the nitty gritty details, we also had to scale the results of our linear regression to the correct number of medals being awarded this year. (Every four years the number of events changes, as some new events are added to the program and occasionally some are removed. Thus the total number of medals ebbs and flows.) So we put together the linear regression, scaled it, and got our results!

The Survey Says …

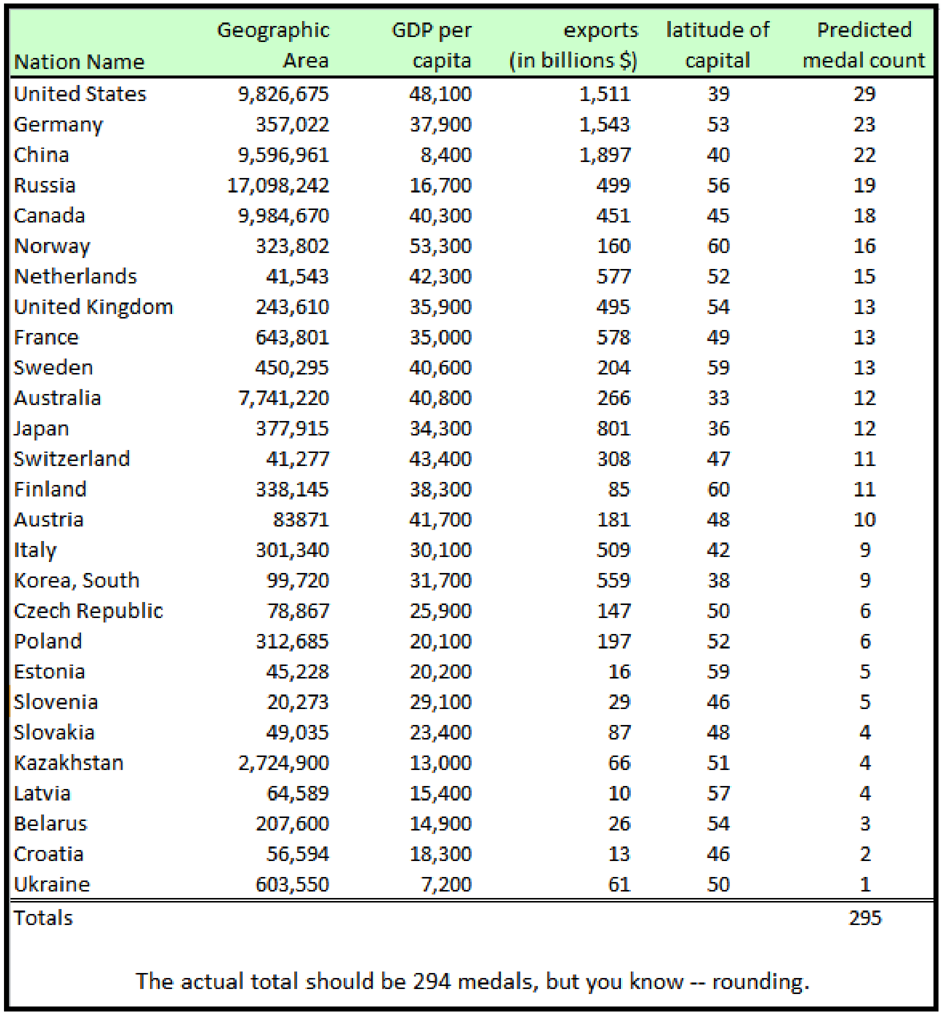
The table at right shows our predictions. (For all nations not shown, we are predicting a medal count of zero.) The four variables the linear model uses to make these predictions are as follows:

- Geographic area - We are a little perplexed to find this variable in the model. Our best guess is that it may reflect the nation's population and/or the genetic diversity in the population and/or the presence of mountain ranges on which to ski and snowboard. Also, it does separate the relatively larger nations of the world from the many small (geographically and population-wise) island nations in the Caribbean and the Pacific.

- GDP per capita - This was no surprise. It seems to confirm my hunch that nations whose people are affluent can afford to spend time pursuing excellence in sports, while poorer nations cannot.

- Value of Exports – A measure of a nation’s total economic power that seems to complement per capita GDP.

- Latitude of Nation's Capital - No surprise here. The further your country is from the equator, the more snow and ice you'll have – and the more medals you'll win at sports contested on snow and ice!

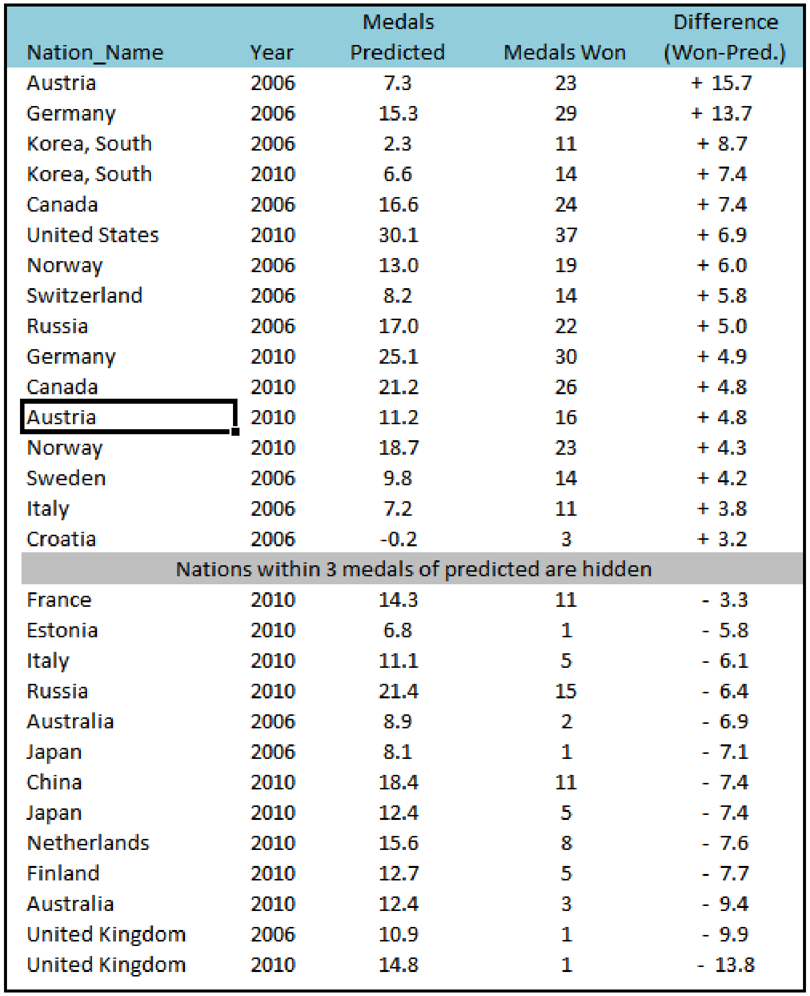


	Geographic	GDP per	exports	latitude of	Predicted
Nation Name	Area	capita	(in billions $)	capital	medal count
United States	9,826,675	48,100	1,511	39	29
Germany	357,022	37,900	1,543	53	23
China	9,596,961	8,400	1,897	40	22
Russia	17,098,242	16,700	499	56	19
Canada	9,984,670	40,300	451	45	18
Norway	323,802	53,300	160	60	16
Netherlands	41,543	42,300	577	52	15
United Kingdom	243,610	35,900	495	54	13
France	643,801	35,000	578	49	13
Sweden	450,295	40,600	204	59	13
Australia	7,741,220	40,800	266	33	12
Japan	377,915	34,300	801	36	12
Switzerland	41,277	43,400	308	47	11
Finland	338,145	38,300	85	60	11
Austria	83871	41,700	181	48	10
Italy	301,340	30,100	509	42	9
Korea, South	99,720	31,700	559	38	9
Czech Republic	78,867	25,900	147	50	6
Poland	312,685	20,100	197	52	6
Estonia	45,228	20,200	16	59	5
Slovenia	20,273	29,100	29	46	5
Slovakia	49,035	23,400	87	48	4
Kazakhstan	2,724,900	13,000	66	51	4
Latvia	64,589	15,400	10	57	4
Belarus	207,600	14,900	26	54	3
Croatia	56,594	18,300	13	46	2
Ukraine	603,550	7,200	61	50	1
Totals					295

The actual total should be 294 medals, but you know -- rounding.

So as we look at the table, we see nations far from the equator, with modern economies, with relatively high wealth, and which are relatively large geographically. Some other interesting facts pop out. Of the 27 nations listed, only seven are outside of Europe. China, Japan, South Korea, and Kazakhstan represent Asia, while the United States and Canada are in North America. The only other nation – and the only one located in the southern hemisphere! – is Australia. It will be interesting to see how close the prediction for the U.S. will be. In 2010, the U.S. team set a new record with 37 total medals, only their second time winning the total medal count.

Outliers



Nation_Name	Year	Medals Predicted	Medals Won	Difference (Won-Pred.)
Austria	2006	7.3	23	+ 15.7
Germany	2006	15.3	29	+ 13.7
Korea, South	2006	2.3	11	+ 8.7
Korea, South	2010	6.6	14	+ 7.4
Canada	2006	16.6	24	+ 7.4
United States	2010	30.1	37	+ 6.9
Norway	2006	13.0	19	+ 6.0
Switzerland	2006	8.2	14	+ 5.8
Russia	2006	17.0	22	+ 5.0
Germany	2010	25.1	30	+ 4.9
Canada	2010	21.2	26	+ 4.8
Austria	2010	11.2	16	+ 4.8
Norway	2010	18.7	23	+ 4.3
Sweden	2006	9.8	14	+ 4.2
Italy	2006	7.2	11	+ 3.8
Croatia	2006	-0.2	3	+ 3.2
Nations within 3 medals of predicted are hidden				
France	2010	14.3	11	- 3.3
Estonia	2010	6.8	1	- 5.8
Italy	2010	11.1	5	- 6.1
Russia	2010	21.4	15	- 6.4
Australia	2006	8.9	2	- 6.9
Japan	2006	8.1	1	- 7.1
China	2010	18.4	11	- 7.4
Japan	2010	12.4	5	- 7.4
Netherlands	2010	15.6	8	- 7.6
Finland	2010	12.7	5	- 7.7
Australia	2010	12.4	3	- 9.4
United Kingdom	2006	10.9	1	- 9.9
United Kingdom	2010	14.8	1	- 13.8

Of course, we know that our model won't be perfect. The chart at left shows medals won versus medals predicted for the last two Winter Olympiads. There are some nations which consistently over-perform and others which regularly under-perform, at least according to our model and the data on which it is based. Looking at the outliers is interesting and also points to the kind of data that we would need to improve the model.

South Korea - This nation over-performed by about 8 medals in both 2006 and 2010. How do you account for the fact that short-track speed skating is hugely popular there, and they routinely win lots of Olympic medals in that discipline?

Germany - In 2006 they over-performed our prediction by 14 medals and in 2010 by five. Is it their work ethic or a love for competition? Whatever it is, they've got it.

Austria, Norway, and Canada - These countries *always* perform well at the Winter Olympics. In 2006, Austria outpaced our model by a full 16 medals! Now that is getting it done! (Take a look at Appendix A that shows the all-time medal counts and prepare to be astonished at the all-time leaders.)

The UK – Our best guess at why our friends across the pond generally underperform (at least according to our model) is that the UK's geographic location causes their winters to be milder and filled with much more rain than snow and ice. Perhaps the Winter Olympics sports have never really caught on there. Historically, they are the third highest medal winners at the *Summer* Olympics. The winter sports simply must not be their cup of tea.

Australia – Our model predicted bigger medal counts for them in both of the last two Winter Olympiads. Of all the nations we predicted to get medals, they are closest to the equator. Perhaps a milder climate is again the culprit.

What to watch for

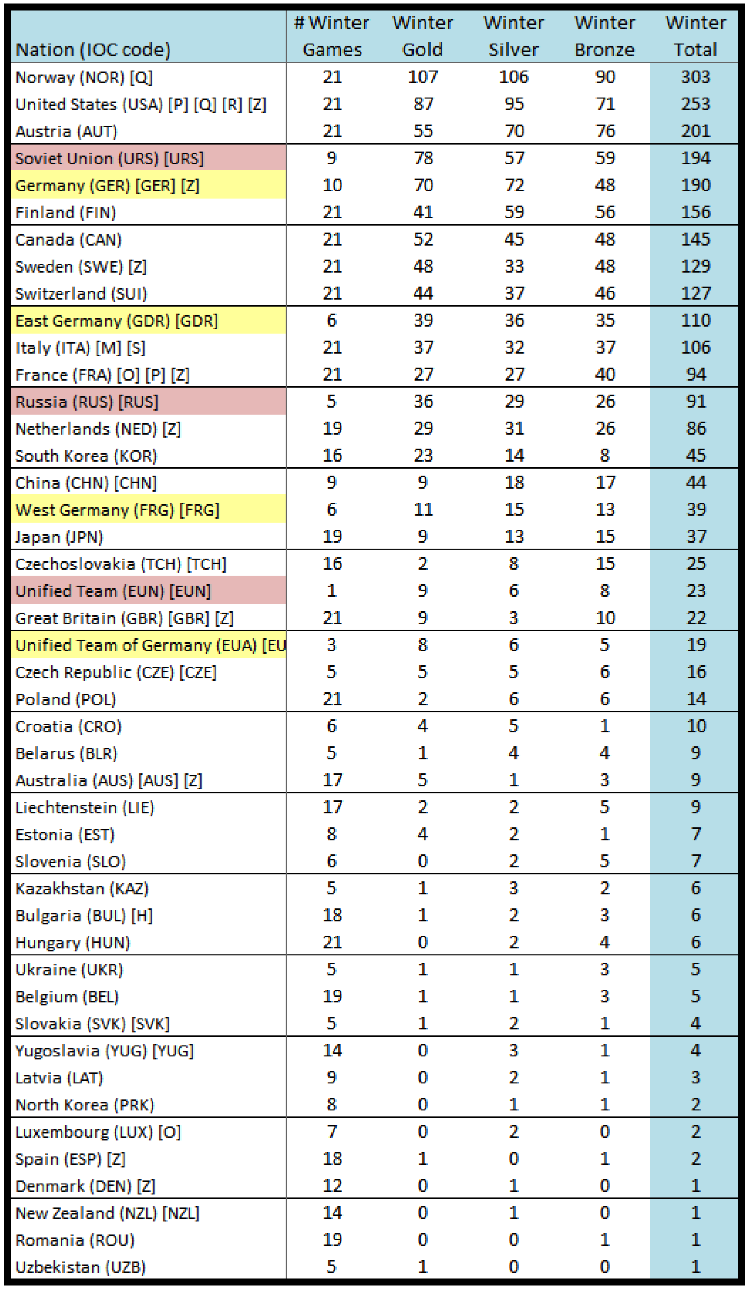
- *Home Team success* – History has shown us that the nation hosting the games often over-performs. From the outliers table above, notice that both Italy and Canada over-performed in 2006 and 2010, respectively, as the Winter Games were hosted in Torino and Vancouver. Although not shown above, Canada's gold medal count while hosting the games in 2010 was especially impressive: with 14 gold medals, they broke the all-time record in that category. Some readers may remember Canada's "Own the Podium" effort for those games, which obviously paid off. Will Russia pull off a similar feat this year?

- *Breakthrough nations* – Will a nation from Africa or South America break through and finally win a Winter medal? Also, some nations like the former Soviet republics of Georgia, Kyrgyzstan, and Tajikistan as well as Serbia and Israel seem to have many of our predictive factors in place and are on the bubble to end their medal droughts.

- *Reappearances* – Nations like Bulgaria and New Zealand have won Winter Games medals before, but not in the last Olympiad. Bulgaria was the nation who scored closest in our model to winning a medal without actually reaching the mark. As for New Zealand, we're rooting for them. But it's tough to practice skiing and skating when your country is overrun with elves and dwarves and orcs[6].

- *Mankind striving for athletic excellence* – It's a sure bet that we'll see what ABC's *Wide World of Sports* used to call, "the thrill of victory and the agony of defeat!"

Appendix A – All Time Medal Counts for the Winter Olympics



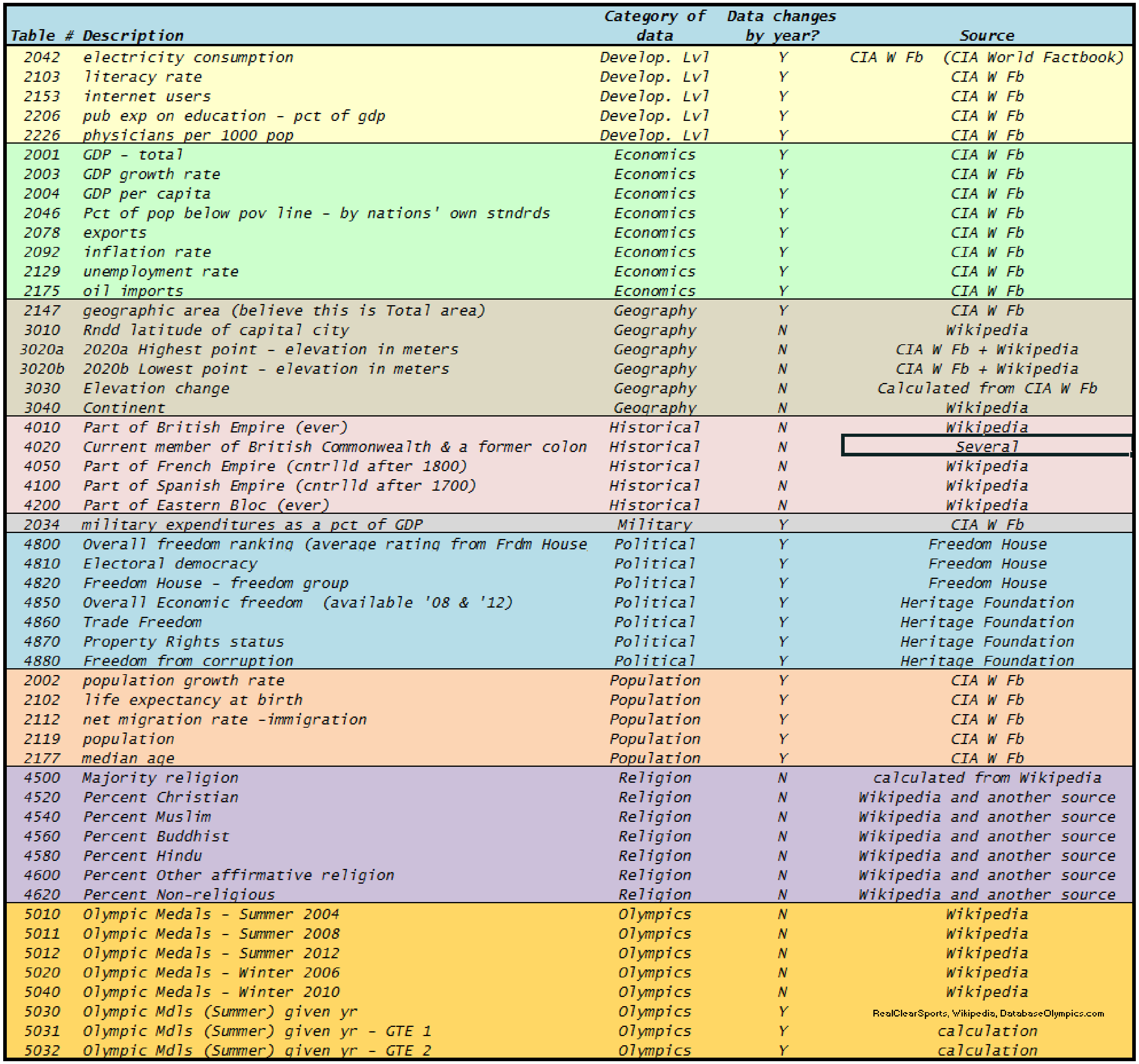
Nation (IOC code)	# Winter Games	Winter Gold	Winter Silver	Winter Bronze	Winter Total
Norway (NOR) [Q]	21	107	106	90	303
United States (USA) [P] [Q] [R] [Z]	21	87	95	71	253
Austria (AUT)	21	55	70	76	201
Soviet Union (URS) [URS]	9	78	57	59	194
Germany (GER) [GER] [Z]	10	70	72	48	190
Finland (FIN)	21	41	59	56	156
Canada (CAN)	21	52	45	48	145
Sweden (SWE) [Z]	21	48	33	48	129
Switzerland (SUI)	21	44	37	46	127
East Germany (GDR) [GDR]	6	39	36	35	110
Italy (ITA) [M] [S]	21	37	32	37	106
France (FRA) [O] [P] [Z]	21	27	27	40	94
Russia (RUS) [RUS]	5	36	29	26	91
Netherlands (NED) [Z]	19	29	31	26	86
South Korea (KOR)	16	23	14	8	45
China (CHN) [CHN]	9	9	18	17	44
West Germany (FRG) [FRG]	6	11	15	13	39
Japan (JPN)	19	9	13	15	37
Czechoslovakia (TCH) [TCH]	16	2	8	15	25
Unified Team (EUN) [EUN]	1	9	6	8	23
Great Britain (GBR) [GBR] [Z]	21	9	3	10	22
Unified Team of Germany (EUA) [EU	3	8	6	5	19
Czech Republic (CZE) [CZE]	5	5	5	6	16
Poland (POL)	21	2	6	6	14
Croatia (CRO)	6	4	5	1	10
Belarus (BLR)	5	1	4	4	9
Australia (AUS) [AUS] [Z]	17	5	1	3	9
Liechtenstein (LIE)	17	2	2	5	9
Estonia (EST)	8	4	2	1	7
Slovenia (SLO)	6	0	2	5	7
Kazakhstan (KAZ)	5	1	3	2	6
Bulgaria (BUL) [H]	18	1	2	3	6
Hungary (HUN)	21	0	2	4	6
Ukraine (UKR)	5	1	1	3	5
Belgium (BEL)	19	1	1	3	5
Slovakia (SVK) [SVK]	5	1	2	1	4
Yugoslavia (YUG) [YUG]	14	0	3	1	4
Latvia (LAT)	9	0	2	1	3
North Korea (PRK)	8	0	1	1	2
Luxembourg (LUX) [O]	7	0	2	0	2
Spain (ESP) [Z]	18	1	0	1	2
Denmark (DEN) [Z]	12	0	1	0	1
New Zealand (NZL) [NZL]	14	0	1	0	1
Romania (ROU)	19	0	0	1	1
Uzbekistan (UZB)	5	1	0	0	1

As you look at this table, you'll see I've color-coded the rows for Germany and Russia in their various historical incarnations. If Germany had been unified for the full 21 Olympiads, their tally would be 358 and they would have the top spot.

Aside from the obviously interesting counts for the nations in the table, one of the most astounding things for me was the nations which do not appear in the table:

- Greece – As the originators of the Olympic Games, I didn't expect to see them having been shut out at the Winter Olympics.
- Iceland! – How can Iceland never have won a Winter medal??? They're all about ice and snow! It can't be! (It's all the more mind-boggling to know they've won four Summer Games medals.)
- Argentina and Chile – They've each got mountains and some cold climate zones, but no Winter medals.
- The former Soviet republics of Georgia, Kyrgyzstan, Moldova, and Tajikistan – During the Cold War era, we know that the Soviets and the West at times approached the Olympics as a propaganda tool, each attempting to show that their economic and social system was superior by winning the most medals at the Olympics. With that history, I would have expected that the Soviet athlete development system would have set the stage for each of the current republics to produce medal-winning athletes. Only Belarus, Kazakhstan, Ukraine, and Uzbekistan have so far reached the podium.

Appendix B – List of Data Elements used in this project



		Category of	Data changes	
Table #	Description	data	by year?	Source
2042	electricity consumption	Develop. Lvl	Y	CIA W Fb (CIA World Factbook)
2103	literacy rate	Develop. Lvl	Y	CIA W Fb
2153	internet users	Develop. Lvl	Y	CIA W Fb
2206	pub exp on education - pct of gdp	Develop. Lvl	Y	CIA W Fb
2226	physicians per 1000 pop	Develop. Lvl	Y	CIA W Fb
2001	GDP - total	Economics	Y	CIA W Fb
2003	GDP growth rate	Economics	Y	CIA W Fb
2004	GDP per capita	Economics	Y	CIA W Fb
2046	Pct of pop below pov line - by nations' own stndrds	Economics	Y	CIA W Fb
2078	exports	Economics	Y	CIA W Fb
2092	inflation rate	Economics	Y	CIA W Fb
2129	unemployment rate	Economics	Y	CIA W Fb
2175	oil imports	Economics	Y	CIA W Fb
2147	geographic area (believe this is Total area)	Geography	Y	CIA W Fb
3010	Rndd latitude of capital city	Geography	N	Wikipedia
3020a	2020a Highest point - elevation in meters	Geography	N	CIA W Fb + Wikipedia
3020b	2020b Lowest point - elevation in meters	Geography	N	CIA W Fb + Wikipedia
3030	Elevation change	Geography	N	Calculated from CIA W Fb
3040	Continent	Geography	N	Wikipedia
4010	Part of British Empire (ever)	Historical	N	Wikipedia
4020	Current member of British Commonwealth & a former colon	Historical	N	Several
4050	Part of French Empire (cntrlld after 1800)	Historical	N	Wikipedia
4100	Part of Spanish Empire (cntrlld after 1700)	Historical	N	Wikipedia
4200	Part of Eastern Bloc (ever)	Historical	N	Wikipedia
2034	military expenditures as a pct of GDP	Military	Y	CIA W Fb
4800	Overall freedom ranking (average rating from Frdm House	Political	Y	Freedom House
4810	Electoral democracy	Political	Y	Freedom House
4820	Freedom House - freedom group	Political	Y	Freedom House
4850	Overall Economic freedom (available '08 & '12)	Political	Y	Heritage Foundation
4860	Trade Freedom	Political	Y	Heritage Foundation
4870	Property Rights status	Political	Y	Heritage Foundation
4880	Freedom from corruption	Political	Y	Heritage Foundation
2002	population growth rate	Population	Y	CIA W Fb
2102	life expectancy at birth	Population	Y	CIA W Fb
2112	net migration rate -immigration	Population	Y	CIA W Fb
2119	population	Population	Y	CIA W Fb
2177	median age	Population	Y	CIA W Fb
4500	Majority religion	Religion	N	calculated from Wikipedia
4520	Percent Christian	Religion	N	Wikipedia and another source
4540	Percent Muslim	Religion	N	Wikipedia and another source
4560	Percent Buddhist	Religion	N	Wikipedia and another source
4580	Percent Hindu	Religion	N	Wikipedia and another source
4600	Percent Other affirmative religion	Religion	N	Wikipedia and another source
4620	Percent Non-religious	Religion	N	Wikipedia and another source
5010	Olympic Medals - Summer 2004	Olympics	N	Wikipedia
5011	Olympic Medals - Summer 2008	Olympics	N	Wikipedia
5012	Olympic Medals - Summer 2012	Olympics	N	Wikipedia
5020	Olympic Medals - Winter 2006	Olympics	N	Wikipedia
5040	Olympic Medals - Winter 2010	Olympics	N	Wikipedia
5030	Olympic Mdls (Summer) given yr	Olympics	Y	RealClearSports, Wikipedia, DatabaseOlympics.com
5031	Olympic Mdls (Summer) given yr - GTE 1	Olympics	Y	calculation
5032	Olympic Mdls (Summer) given yr - GTE 2	Olympics	Y	calculation

[1] Yes, that's right – no nation from Africa, South America, or the Middle East has ever won a medal at the Winter Olympic Games. No nation from the Caribbean has either, despite the worthy efforts of the Jamaican bobsled team!

[2] See our article, "Using Data Mining for a Reality Check" for more on this subject. http://www.discoverycorpsinc.com/data-mining-reality-check

[3] The CIA World Factbook was an excellent resource, as was Wikipedia.

[4] Logistic regressions lend themselves well to yes/no questions, but not very well to 'scale' questions like "how many?"

[5] See our article "Predicting the London Olympics Medal Count". http://www.discoverycorpsinc.com/predicting-the-olympic-medal-c/

[6] New Zealand has been the backdrop for the filming of "The Lord of the Rings" and "The Hobbit" movies.