Winter Olympic Statistical Analysis

Kieran Billingham

17 January 2019

Abstract

Exploratory analysis is performed on a dataset populated with Winter Olympic medal records. It is found that the Games have been expanding despite political problems. The event is now at its largest with 500+ athletes competing in 2014. Gender representation of NOC's is equaling and there is no evidence for gold medal bias for gender.

Global population is growing linearly but not equally among countries. Global GDP is also growing, but individual GDP for nations is very broad. 'Wealthy' countries increasing the mean GDP so that most of the world sits below it.

An increase in GDP and population has a negative effect on a NOC's likelihood of an Olympic win.

Contents

Introduction

The modern Olympics are the pinnacle of international sporting events. The first modern Summer Games were held in Athens and they have been hosted as a quadrennial spectacle of excellence ever since. Winter sports had been added to the Olympic programme over the years but due to internal and external politics, they weren't added as official sports. However, in 1924 the first dedicated Winter Games were hosted in Chamonix, where 5 original sports were included (bobsleigh, curling, hockey, Nordic skiing and skating) [1].

The Games have suffered at the hands of world events, World War I and II caused cancellations of both the Summer and Winter Games, whilst the Cold War used the Olympics as a propaganda tool [2]. In terms of size, the Summer Olympics dwarfs the Winter Games, with 208 countries competing in London 2012 versus just 88 in Sochi 2014. [3] Almost 1000 Medals are awarded at each Summer Games whereas only around 100 are awarded at Winter Games (most recent) [4].

Total medal tables suggest that USA, URS, GBR, GER, and FRA are the major Olympic competitors, these countries have the highest total medal results [4]. However, since the Summer Games account for ten times the number of medals than the Winter Games, this does not give a fair representation of the top performing Winter Olympic competitors. This report looks to explore the history and evolution of Medallists at the Winter Games and to investigate some potential contributors to successful Olympians.

Dataset Selection

Olympic Medal Table

Historic Olympic success information is available from www.olympic.org through an interactive form. This source gives the full results including rank, timings, medals and scores, however, data is only available on an event per game basis. This would mean that the base file would have to be built by scraping, with much of the information (scores, timings, etc) being too in-depth for cross-games comparison.

The official success of each National Olympic Committee (NOC) is calculated through a total medal count, with higher importance placed on gold than on silver and bronze respectively. This means that only information on athletes who gained a podium position needed to be considered for this project.

The IOC Research and Reference service produce datasets that hold observations for every medal awarded at both the summer and winter games. A copy of the 2014 version was uploaded by The Guardian's Data blog and hosted by kaggle.com. This dataset had

enough variables for some exploratory analysis and had the potential to benefit from feature engineering [5].

Global Population by Country

The population of the country that each NOC represents gives an indication of how many elite athletes are produced by that nation. The dataset used for the global population was published by The World Bank Group in 2018 and gives a mid-year population estimate for each contributing country.
Different sources compute total population differently, in this case, the total population is based on the de facto definition of population, which counts all residents of a state regardless of legal status or citizenship [6].

Global Gross Domestic Product

Gross Domestic Product (GDP) is the sum of gross value added to a country's economy by all resident producers, plus any product taxes and minus any subsidies not included in the value of the product. It does not include adjustment for the degradation of natural resources.

GDP is considered a reasonable indicator of a state's wealth, which in turn indicates a greater spending power of a NOC to train an Olympic team. It could also be an indicator of better training facilities and more personal wealth to invest in sport through childhood.

The dataset used was curated by the Organization for Economic Co-operation and Development. GDP values are published in U.S. dollars which are produced through the conversion of domestic currency using single-year exchange rates [7].

Exploratory Analysis

Olympic Medal Table

The first imported dataset held the Olympic medal information. 5,770 observations across 9 variables (1 numeric (Year) and 8 factor). Information was held for Winter Olympics between the years of 1924 and 2014 with 7 sports, 15 disciplines, 83 events and 19 hosts. Athletes were recorded by surname and first initial and the podium placing countries were listed as their three-letter National Olympic Committee codes (NOC).

Between 1924 and 2014, 1,919 Bronze, 1,930 Silver and 1,921 Gold medals had been awarded (imbalance due to place-ties). 3,944 medals had been awarded to men and 1826 awarded to women, a disparity of later interest.

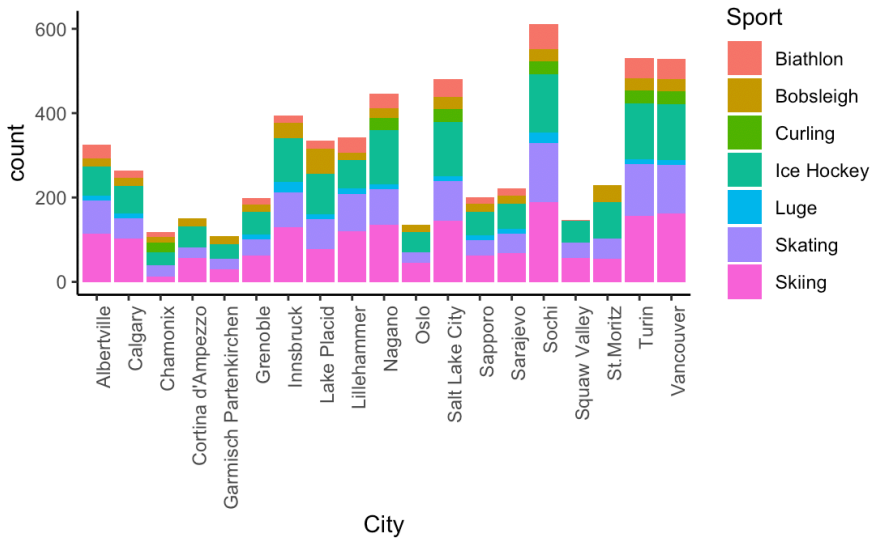

Figure 1, chart of total medals awarded at each Winter Olympics, grouped by sport.

The number of medals awarded was not consentient between Olympiads, the later games awarded more, as shown by Figure 1. There is also an inconsistency between sports. Indicating that some have a greater representation than others.

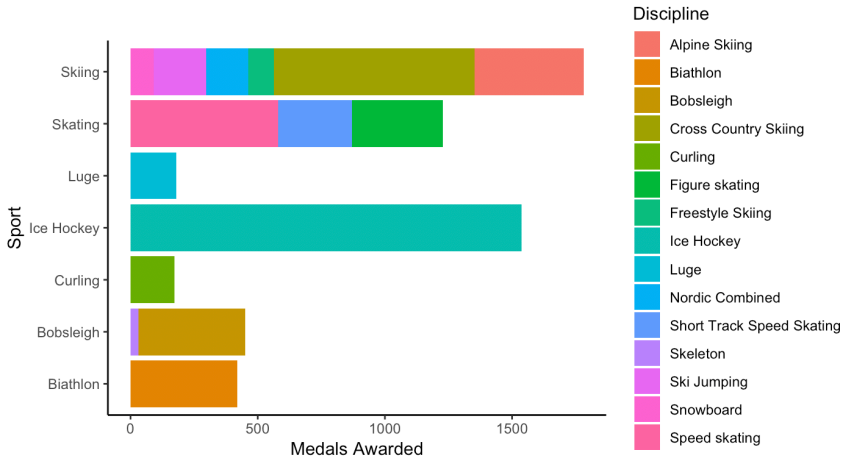

Figure 2, Chart of total medals awarded per sport grouped by discipline.

The total medal imbalance between sports is seen clearer in Figure 2, skiing represents the most disciplines and also awards the most medals. Hockey represents just one discipline but has the second highest medal count. This is an indication that within the discipline of hockey there are many events, the data agrees with this but with a total of 83 events over 15 disciplines, the plot becomes too unclear for this format.

The next series of figures explore the temporal changes in the dataset. Using the years of each Olympiad as the x-axis. In Figures 3-7 the missing information around the year 1940 is due to cancellations of the Games around WWII.

Figure 3 indicates the expansion of the Winter Games to include more sports. At Chamonix the luge was trialled but not accepted by the IOC, the luge was excluded for a number of years until in 1964 it was reincluded at Innsbruck.

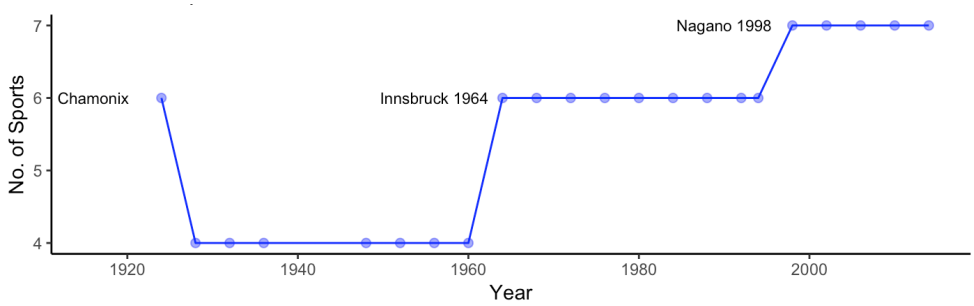

Figure 3, Point and line plot showing the increasing inclusion of sports in the Winter Games

The next change in sports included at the games happened at Nagano 1998, the biathlon (cross-country skiing and rifle shooting) was included for the first time.

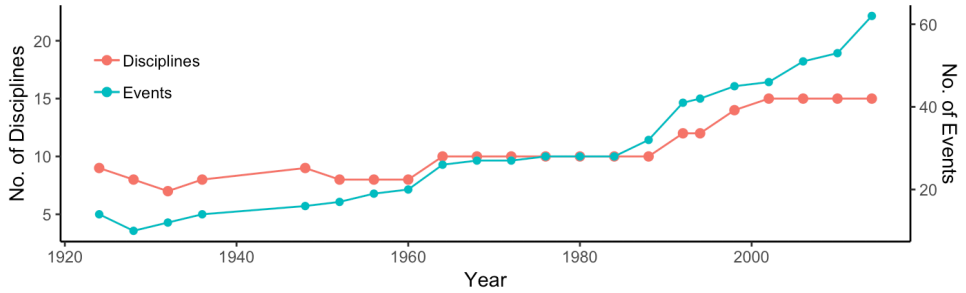

Figure 4, Point and line plot showing how the No. Events and No. Disciplines have increased over time.

The expansion of the sports at the games bought with it the inclusion of more disciplines and events. Figure 4 shows the increase of disciplines is similar to that of sports in Figure 3, note the stagnant period between 1960 and 1980. This is around the same time as the Cold War, where many of the large NOC's were boycotting the Games and so less investment took place. In recent years the number of disciplines has levelled off, but the number of events is quickly increasing.

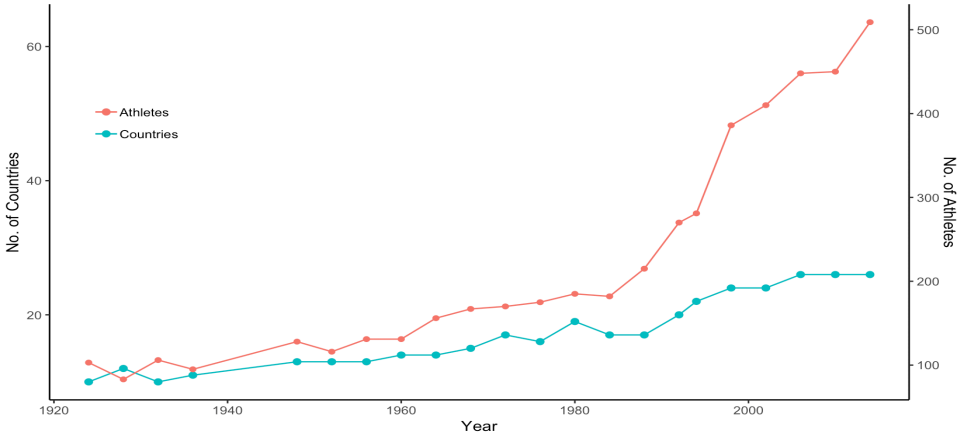

Figure 5, Point and line plot showing the increase of Countries and athletes at each consecutive Games.

Figure 5 shows the slow increase of countries attending the games since 1926 with just 10 NOC's to 2014 with 26 NOC's. It is also shown how the number of athletes has

increased at an even greater rate, NOC's are increasing the size of their Winter Olympic teams year on year.

In Figure 6 it can be seen that the increase in athletes is mainly due to the constant improvement to the gender balance within the Olympians. Women at the Olympics saw a slow but steady increase in numbers since the Games inception and since the 90's they have seen a period of rapid expansion. Male medalists have seen a slower growth more in line with the expansion of the Games.

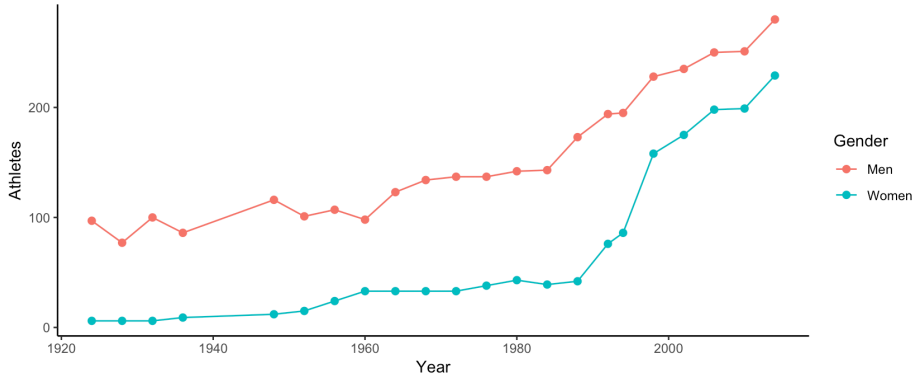

Figure 6, Point and line plot showing the amount of male and female athletes winning medals.

The proportion of male to female athletes from each NOC is something that has been discussed a lot within the Summer Olympics and so it has been plotted in Figure 7 for the Winter Games.

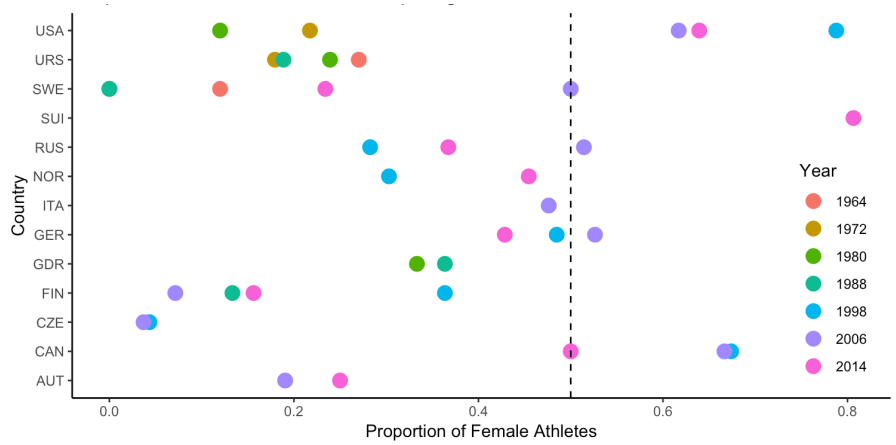

Figure 7, Scatter plot showing the proportion of women winning medals at each Olympic game, for each country.

It is clear from Figure 7 that as time advances the proportion of female athletes winning medals for each IOC has moved towards and even surpassed 0.5 (black line). This can be taken as an indication that a greater proportion of these IOC's teams are women.

Global Population by Country

A country's population indicates how many athletes they are likely to have to pick from for an Olympic team. As the Olympic games happen over decades, the populations at each Olympiad are going to be different.

The second imported dataset looked at worldwide population from between 1960 and 2017. It had dimensions of 264 rows (country codes) and 59 columns (years). In order to be better understood as a time series, the data frame was transposed to give each country code a column. It was then found that not all the county codes could be countries, some were continental sums, monetary unions and areas of economic importance. These were listed as invalid columns and removed where necessary (instances like RUS and SUN, for Russia and the Soviet Republic caused issues).

The next series of plots look at a few trends for the global and national population. Figure 8 shows the main trend of the global population between 1960 and 2017, the increase is almost perfectly linear. Notice that the total population exceed that of the 'known' value, this is because of the inclusion of some economic areas that link to IOC's in later analysis. In this scenario some people have been counted more than once, causing this inflation. The information given in this plot would suggest that each country's population is increasing at the same rate, this goes against popular knowledge and so was investigated further.

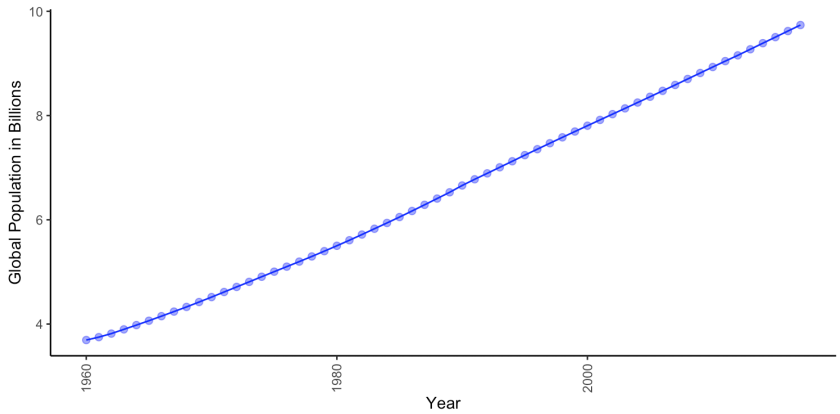

Figure 8, Total population calculated by summing Country data per year

The next plot, Figure 9, has two subplots that each show the population trends of four Countries. They are separated into very high population states and medium populated states. This gives some more in-depth information about the data that comes to produce Figure 8.

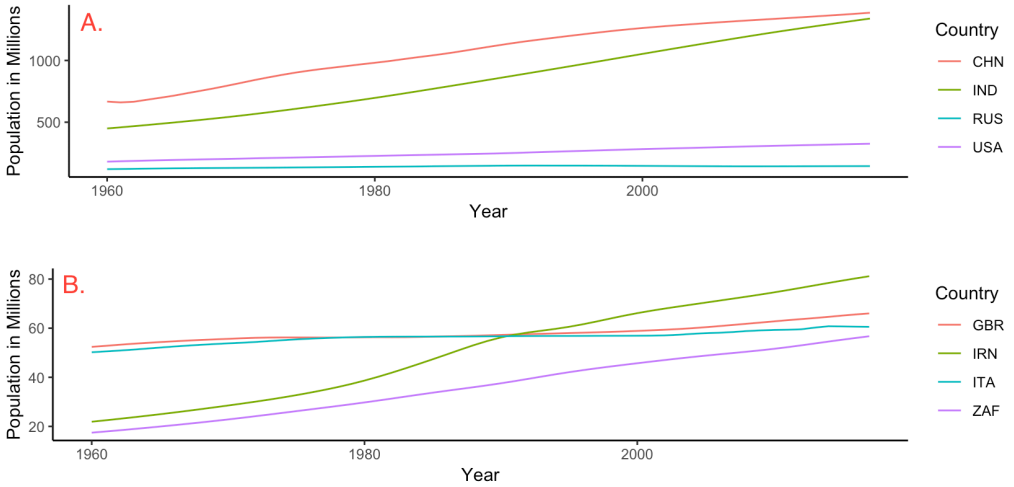

Figure 9, Population trends for sample Countries. A. Highest Populations, B. Medium Populations

Figure 9 A, shows the population development of 4 of the largest countries. Notice that India has the largest growth, followed by China. The United States has seen much slower growth while Russia seems to be fairly constant.

Figure 9 B has a more interesting shape. The most striking trend is the growth of Iran's population, overtaking both Italy and Great Britain.

Global Gross Domestic Product

The GDP of a NOC's state is being used here as an indicator of wealth. The larger the wealth of a state, the more funds available for facilities and training. As can be seen from Figure 10 the Global GDP has been on the incline since records began in 1960. They few troughs around the late 00's associated with the global economic crisis. Whilst the dip at 2017/2018 is a result of the drop in GDP of the United States and China having a knock-on effect globally.

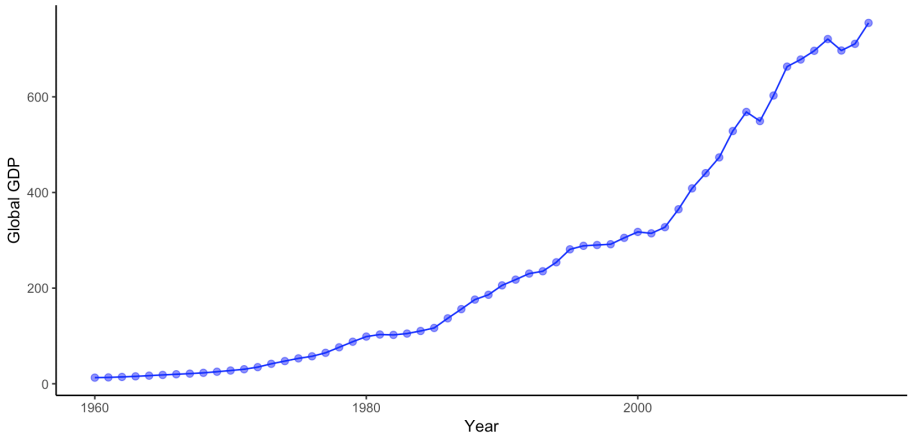

Figure 10, mean GDP of included countries over time

GGDP seems to be recovering and is starting to grow again. Figure 10 doesn't give any indication of wealth distribution, much like its equivalent for population (Figure 8), and so a breakdown of this GDP spread was plotted.

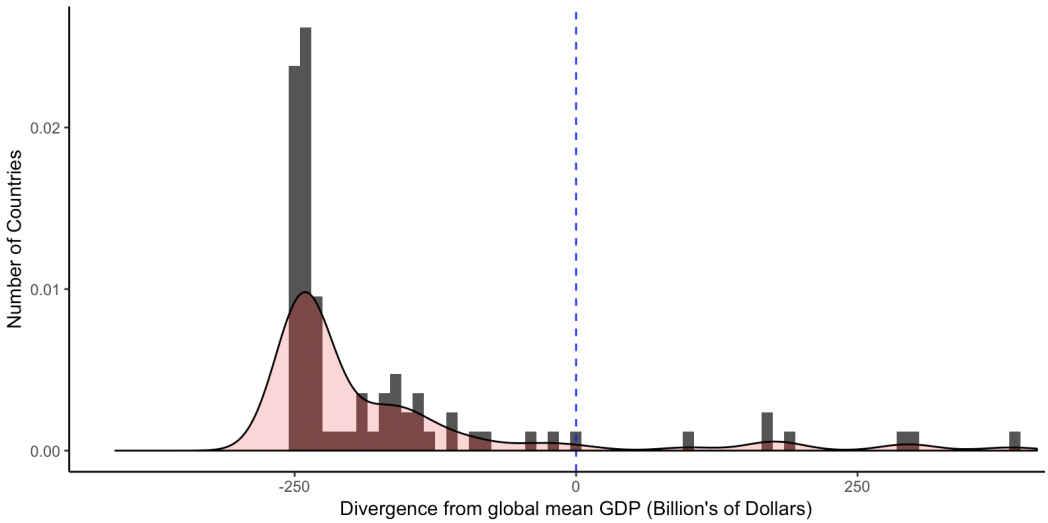

Figure 11, A plot showing the spread of GDP of countries from the Global mean. x-axis limited to ±0.5 s.d for clarity. Blue line = mean, red area = density of countries.

The divergence from the GGDP for countries that have a GDP within 0.5 standard deviations of the mean are presented in Figure 11. The choice was chosen to limit the plot to allow for better visualization. The harsh reality is that a few 'super economies'' lie at the far right of the x-axis and hold much of the world's wealth. It is clear that most of the countries lie below the global mean, with a large proportion at negative $250 billion dollars. The density overlay has a loose normal distribution to it as if the negative amount is the true mean and countries right of the blue line are outliers. This is investigated as part of the statistical analysis in the next section.

Statistical Analysis

In order to perform efficient and meaningful statistical analysis on the datasets, a smaller sample was taken. The medal information for the top performing (by total gold, silver and bronze count) NOC's at the last 5 Winter Olympics was chosen as the subject for investigation. The NOC's were then truth tested against valid ISO (International Standards Organization) codes of territories as listed in the GDP and population data. Through some manipulation and merging of the three independent datasets, a final data frame including all the relevant information was created.

The target for analysis will be whether or not a NOC was awarded a gold medal.

Testing for Normality

The data must be checked for an approximately normal distribution in order to meet the assumptions of parametric testing, so it is a sensible first statistical test to perform.

Kurtosis is the measure of the spread of data, which equates to the length of tails in normal distribution analyses. Skew is the measure of the symmetry of the data about the mean. The null hypothesis of these tests is that the distribution is normal. If they are found to be significant the distribution is not normal.

Functions for kurtosis and skew were assessed for both population and GDP in the joined dataset. And Z values produced using the associated errors in both.

Table 1, Table of Kurtosis and Skew Z scores for merged dataset

	Kurtosis	Skew	Accepted Range at 0.05 significance
Population	128.09	58.90	1.96 and 1.96
GDP	33.69	27.09	

Table 1, Table of Kurtosis and Skew Z scores for merged dataset the kurtosis and skew of the population data is far from the range and so the null hypothesis is broken. The data is not normally distributed at a 0.005 significance level. The GDP values are somewhat smaller but, the null hypothesis still cannot be accepted at this level. The visualization of GDP shown in Figure 11 does have a loose bell curve shape and so the Z figures do seem rather large.

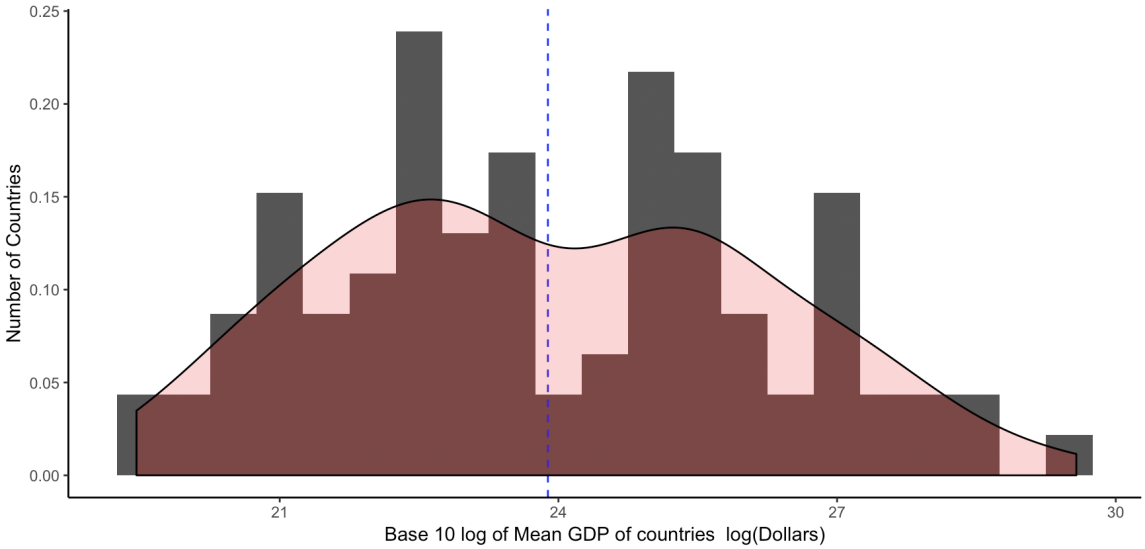

Figure 12, A log plot of the distribution of mean GDP. Bin size = 0.5. Notice the loose normal distribution of the data

In Figure 12, the mean GDP of countries have been transformed through the log function. This draws in outliers and gives a 'tighter' distribution view of the data. It can be seen that the shape of the density distribution looks like two merged normal distributions, leading to the 'dip' around the mean (blue line). This could represent the

distribution of more economically developed countries and less economically developed countries; two distinct economic environments. For the purpose of this report, if treated as an individual distribution and the kurtosis and skewness if now calculated again, a much better picture is formed.

Table 2, Kurtosis and skew Z scores for the log of the mean GDP. Skew is in the accepted range

	Kurtosis	Skew	Accepted Range at 0.05 significance
log GDP	4.66	0.7	-1.96 and 1.96

Table 2 shows the improvements to the kurtosis and skew made by producing the logarithmic. Kurtosis has reduced dramatically but it is still outside the accepted range for a normal distribution (likely due to the 'camel-hump' look of the data). Skew has been reduced enough so that the null hypothesis is true at 0.05 significance.

The logged GDP data could have parametric tests performed on it but the kurtosis of the data would have to be considered on any outcome. It is more favourable to use non-parametric tests.

Wilcoxon Signed Rank

The Wilcoxon Signed Rank test is a non-parametric version of a t-test. It is used in this section to assess the change in medal distribution between Olympic years.

The dataset is split into total medals won by each NOC each year. The count of medals won by each country is then separated into a vector that can be processed through the Wilcoxon test.

Each year was paired with every other year, giving 10 combinations of yearly results. These are shown in Table 3.

Table 3, The P-values of the Wilcoxon Signed Rank test between yearly medal results data

P-Values	1998	2002	2006	2010	2014
1998		0.683	0.563	0.454	0.272
2002	0.683		0.842	0.289	0.157
2006	0.563	0.842		0.157	0.327
2010	0.454	0.289	0.157		1.000

The null hypothesis for this test is that each year of the Olympics the same distribution of medals occurs, and therefore each year can be identified as part of an identical population. None of the *P-Values* in Table 3 are below the 0.005 threshold set for the

null hypothesis to be broken and so it can be said that the distribution of medals is from the same population. There has been no significant change in medal distribution for the top performing countries since 1998.

Pairings of interest include 2006 and 2010 (highlight blue), and 2010 and 2014 (highlighted yellow). The 06-10 pair has the highest probability of being independent samples, in 2010 Canada performed better than any host nation had before, and a number of planned events were ultimately excluded. This may be the cause of a differing trend is medal distribution.

The other pairing of 10-14, has a p-value of 1, indicating that they have the exact same distribution. This is strange because, in 2014, a number of new Nordic events were added to the programme, so a change in Medals was expected. It is a coincidence that the years have the same mean and sum, and so cannot be processed properly by the Wilcoxon-test. Exact pairings cause errors in the function.

Logistic Regression

The last statistical test performed on the data set was a binary logistic regression. A binary target of Gold=1 for observations having 'Medal = Gold' was created.

Logistic regression (LG) is a predictive analysis that is used to describe data and explain the interactions between an independent binary variable with one or more independent variables. LG was chosen because it can handle a mixture of nominal and interval variables, which this dataset is composed of. The logarithmic nature of the regression also prevents outliers (as seen in population and GDP data) from skewing the estimates.

To prepare the data for the regression analysis it was standardized, and variables of high correlation were removed.

The first classifier that was produced used all valid variables (Athlete, Medal and Year were removed). It was found that using the Events variable created sample sizes that were too small to be predictive. If Events were included, then a much larger sample would be needed (possibly include Summer Olympics information).

A second classifier was made with these adjustments and the statistically significant variables are shown in Table 4. They are displayed in descending order of significance (with the level set at 0.005). A third classifier was made including just the GDP and population data, this is also shown.

Table 4, statistically significant coefficients of two logistic regression models.

	Variable	Estimate	Std.Error	Pr(>|z|)	Significance
All Variables	Country: FIN	-1.969681	0.530377	0.000204	***
	GDP	-2.802258	0.989788	0.004638	**
	Country: NOR	0.935542	0.391997	0.017005	*
	City: Vancouver	0.642011	0.286396	0.024981	*
	Discipline:Nordic Combined	1.0086	0.474922	0.033694	*
	City: Turin	0.549338	0.278453	0.048516	*
	Population	12.872562	7.665875	0.093112	.
Pop & GDP	Intercept	-0.73069	-10.833	2.00E-16	***
	GDP	-0.25207	-2.974	0.00294	*
	Population	-0.19685	-2.169	0.0301	**

The regression model for *All Variables* had 7 statistically significant variables. Some of these made logical sense, such as population having a very positive impact (Athletes that have had to compete in a larger national pool are more likely to succeed). Other's did not make such logical sense.

Being from Finland had a large negative effect on the probability of an athlete winning gold, this was the most statistically significant variable and yet there is no plausible explanation for it.

GDP had the largest negative interaction with the gold target. This goes against logic that more wealthy countries can support athletes to succeed. It is important to consider the large standard error though, it is over 0.3 times the value of the interaction.

Population and GDP (the only continuous variables) were isolated and tested against the gold flag. Population saw an increase in significance, but GDP saw a decrease. Both are also seen to have a negative effect, this goes against the patterns that are linked to total medal tables. This may indicate a positive change in the bias towards wealth that the Summer Olympics has a reputation for. Winter Olympics favors lower population countries and/or those with smaller GDP's.

Table 5, In the 'All Variables' model, gender was not considered to be statistically significant.

Variable	Estimate	Std. Error z	Pr(>|z|)	Significance
Gender Women	-0.136826	0.157594	0.385276	

A positive interaction (or lack thereof) was found between gold medal flags and gender. Table 5 shows the interaction of Gender = Women, it slightly negative but the model has valued it to have no statistical significance.

Even though the number of women at the Olympics has fallen behind men until recent years, an Olympians gender has little to no effect of whether, when placing, they receive a gold medal or not. This acts as a small indicator that training and development opportunities for winter sports is mostly equal among genders.

The performance of the *All Variables* model was then assessed using a confusion matrix.

Table 6, Performance analysis of the All Variables logistic regression. At 0.5 cut off threshold

	All Variables Model
Accuracy	0.744
Error	0.256
Sensitivity	0.680
Specificity	0.761
Precision	0.431

Table 6 shows the performance of the logistic regression. The threshold for a positive prediction has been set at the probability depth of 0.5. Accuracy for the model is quite good, almost 75% successful classifications were achieved on the test set. The sensitivity for this model is also acceptable, with a maximum score of 1 and a minimum of 0, a score of 0.680 is respectable considering the very small dataset this model is trained on. A similar story is true for the specificity, it has a value of 0.761.

This model is better at predicting when the target is 0 than it is when the target is 1. So, it can predict athletes that will not receive a gold medal more accurately.

Conclusion

The explanatory analysis of the medal dataset showed that the games have been increasing in diversity ever since inception. The number of sports increased to include 7, whilst the number of disciplines and events grows year on year. This is expected to continue with the growing popularity of Snowboarding. The Wilcoxon tests did show that, although expanding, the distribution of medals was not changing from the original population.

The World Wars and the Cold War created tension at the Olympics, and so some stalling of progression was seen around those years. The events recovered quickly from these huge world events, signaling the importance to international relations that they hold.

The gender balance of Olympic teams is slowing becoming neural, traditionally more male athletes have been submitted to the Games but in recent years some NOC's have surpassed the 0.5 point and now have female dominated teams. The linear regression showed that there was no statistically significant gender bias in the awarding of gold medals, suggesting that selection for Olympic teams is happening on a performance basis rather than anything else. In a way of leading by example, the Winter Olympics' is setting a great example for the direction we as a society need to be moving forward by considering all genders as equal and testing people on their performance alone.

Global population is growing linearly but not equally among countries, some are seeing periods of rapid expansion (some recovering from population devastations) while others are seeing a slow decline. Global GDP is also growing, but individual GDP for nations is very broad and disproportionate. The few 'Wealthy' countries with vastly higher GDP's increase the mean GDP so that most of the world sits below it. This huge imbalance in wealth is shown in the figures, even when put through a log function the spread of data is still too large to be considered 'normal'. This is confirmed by the high Z values achieve for the kurtosis of the data.

The regression models found a negative interaction between GDP and population on the probabilities of being awarded a gold medal. This doesn't necessarily agree with logic and so to investigate this further, the dataset would have to be merged with the summer Olympic information. This would give a much larger sample size and reduce overfitting errors.

Other additions to the dataset to improve predictive modelling could be the temperature of the NOC's home states, giving an indication of natural 'winter' facilities. Another could be school information; curriculums for some countries with high altitude areas cover some Nordic sports, this is likely to have an effect on the number of adults with a passion and skill for the events. The list of additions could be endless and there is no doubt that with more information the results of the models and statistical fits could be improved.

References

[1] "CBC Archive," 18 December 2009. [Online]. Available: https://web.archive.org/web/20100302162004/http://www.cbc.ca/olympics/history/story/2009/11/25/sp-1924-chamonix.html. [Accessed 01 2019].

[2] M. Lund, "The First Four Olympics," *Skiing Heritage,* no. Fourth, p. 18, 2001.

[3] Diffen, "Summer Vs Winter Olympics," [Online]. Available: https://www.diffen.com/difference/Summer_Olympics_vs_Winter_Olympics. [Accessed January 2019].

[4] "Olympics," 06 2018. [Online]. Available: https://www.olympic.org/olympic-games. [Accessed 12 2018].

[5] [Online]. Available: https://www.kaggle.com/the-guardian/olympic-games. [Accessed Dec 2018].

[6] [Online]. Available: https://data.worldbank.org/indicator/SP.POP.TOTL. [Accessed Dec 2018].

[7] Dec 2018. [Online]. Available: https://data.worldbank.org/indicator/NY.GDP.MKTP.CD?view=map.