IST 707 Final Project

Jocelyn Pitman & Kelly Schillacci

March 19, 2019

## Introduction

The first recorded Olympic Games were held at Olympia in the Greek city-state of Elis in 776 B.C., but it is generally accepted that the Olympics were at least 500 years old at that time. The ancient Olympics were held every four years and occurred during a religious festival honoring Zeus. Olympic competitions were originally limited to foot races. The subsequent added events included wrestling, boxing, horse and chariot racing, and military competitions. The pentathlon, introduced in 708 B.C., consisted of a foot race, the long jump, discus and javelin throws, and wrestling. With the rise of Rome, the Olympics declined, and in 393 A.D. the Roman Emperor Theodosius I abolished the Games.

1,500 years after the ban, the Modern Olympic Games began in 1896 and were held in Athens. While forming the International Olympic Committee (IOC) in 1894, Baron Pierre de Coubertin shared his philosophy. He wrote “The important thing in life is not the triumph but the struggle, the essential thing is not to have conquered but to have fought well.” A beacon of hope in an ever-changing political climate, the Olympic Games have been held nearly consistently over the last 123 years. Amidst the chaos of World Wars I & II, the Olympic Games were missed in 1916, 1940, and 1944. Although countries have decided not to participate in certain instances of the Games, they have been held as scheduled like clockwork aside from those three years. As the Games have aged, the number, size, and composition of competing teams has grown at a staggering pace. 241 participants from 14 nations competed in 43 events, track-and-field, swimming, gymnastics, cycling, wrestling, weightlifting, fencing, shooting, and tennis. All the competitors were men, a few of the participants were visiting tourists and could sign up. The track-and-field events were held at the Panathenaic Stadium, which was originally built in 330 B.C. Americans won nine out of 12 of these events. The 1896 Olympics featured the first marathon competition, which followed the 25-mile route run by the Greek soldier who brought news of a victory over the Persians from Marathon to Athens in 490 B.C.

The Olympic Games currently consist of biennial sporting events that include winter and summer-based competitions. The Olympic Movement has expanded its influence to create the Paralympic, Special Olympics, and Deaf Olympics for those with physical disabilities, cognitive disabilities, or hearing impaired. Initially, the Olympic Games only featured amateur athletes who took pride in competing against other athletes. Due to the emerging technological advances the International Olympic Committee shifted its perspective by including more professional athletes in competitions, especially after the Games started to receive media coverage and corporate sponsorship. The Games originally hosted both Summer and Winter events at one time. In the early 1990s, however, the Winter Olympics were split off and became a competition in their own right. Following the 1992 Games held in France, Winter athletes again convened just two years later in Lillehammer Norway for the first solely Winter Olympic Games (officially known as the 27th Olympic Winter Games).

## Analysis

### Data

#### Dataset

Two datasets were used for this analysis. The first, athlete\_events, contains 271,116 observations of 15 columns. Each row in this set represents one athlete’s participation in one event at one Olympic Games. The second dataset, called noc\_regions, contains 230 observations of 3 columns. The NOC (National Olympic Committee) dataset contains each unique 3-letter NOC name, the region represented, and a notes column containing additional pertinent information. Both datasets are read in and then joined on the NOC column, resulting in a raw data frame containing 271,116 observations of 17 columns. Example rows and the data frame structure are shown below.

## id name sex age height weight team noc  
## 1 1 A Dijiang M 24 180 80 China CHN  
## 2 2 A Lamusi M 23 170 60 China CHN  
## 3 3 Gunnar Nielsen Aaby M 24 NA NA Denmark DEN  
## 4 4 Edgar Lindenau Aabye M 34 NA NA Denmark/Sweden DEN  
## 5 5 Christine Jacoba Aaftink F 21 185 82 Netherlands NED  
## 6 5 Christine Jacoba Aaftink F 21 185 82 Netherlands NED  
## games year season city sport  
## 1 1992 Summer 1992 Summer Barcelona Basketball  
## 2 2012 Summer 2012 Summer London Judo  
## 3 1920 Summer 1920 Summer Antwerpen Football  
## 4 1900 Summer 1900 Summer Paris Tug-Of-War  
## 5 1988 Winter 1988 Winter Calgary Speed Skating  
## 6 1988 Winter 1988 Winter Calgary Speed Skating  
## event medal region notes  
## 1 Basketball Men's Basketball <NA> China   
## 2 Judo Men's Extra-Lightweight <NA> China   
## 3 Football Men's Football <NA> Denmark   
## 4 Tug-Of-War Men's Tug-Of-War Gold Denmark   
## 5 Speed Skating Women's 500 metres <NA> Netherlands   
## 6 Speed Skating Women's 1,000 metres <NA> Netherlands

## 'data.frame': 271116 obs. of 17 variables:  
## $ id : int 1 2 3 4 5 5 5 5 5 5 ...  
## $ name : chr "A Dijiang" "A Lamusi" "Gunnar Nielsen Aaby" "Edgar Lindenau Aabye" ...  
## $ sex : chr "M" "M" "M" "M" ...  
## $ age : int 24 23 24 34 21 21 25 25 27 27 ...  
## $ height: int 180 170 NA NA 185 185 185 185 185 185 ...  
## $ weight: num 80 60 NA NA 82 82 82 82 82 82 ...  
## $ team : chr "China" "China" "Denmark" "Denmark/Sweden" ...  
## $ noc : chr "CHN" "CHN" "DEN" "DEN" ...  
## $ games : chr "1992 Summer" "2012 Summer" "1920 Summer" "1900 Summer" ...  
## $ year : int 1992 2012 1920 1900 1988 1988 1992 1992 1994 1994 ...  
## $ season: chr "Summer" "Summer" "Summer" "Summer" ...  
## $ city : chr "Barcelona" "London" "Antwerpen" "Paris" ...  
## $ sport : chr "Basketball" "Judo" "Football" "Tug-Of-War" ...  
## $ event : chr "Basketball Men's Basketball" "Judo Men's Extra-Lightweight" "Football Men's Football" "Tug-Of-War Men's Tug-Of-War" ...  
## $ medal : chr NA NA NA "Gold" ...  
## $ region: chr "China" "China" "Denmark" "Denmark" ...  
## $ notes : chr "" "" "" "" ...

#### Variable Examination

ID is unique for each athlete. Each ID is associated with exactly one athlete name and vice-versa. Athletes and IDs without a one-to-one relationship are removed from the analysis. M and F are the only valid values for sex, and it’s confirmed that no other values exist. The age range is checked and the values contained therein are confirmed to be valid. Although 73 may seem elderly for an Olympic athlete, this value is confirmed to be correct[[1]](#footnote-1). The weight values range from 25[[2]](#footnote-2) to 214[[3]](#footnote-3) in this data. Both extremes are confirmed to be valid. Next, the unique values for sport are checked to ensure there are no duplications due to misspellings or punctuation.

## [1] "Alpine Skiing" "Archery"   
## [3] "Art Competitions" "Athletics"   
## [5] "Badminton" "Baseball"   
## [7] "Basketball" "Beach Volleyball"   
## [9] "Biathlon" "Bobsleigh"   
## [11] "Boxing" "Canoeing"   
## [13] "Cross Country Skiing" "Curling"   
## [15] "Cycling" "Diving"   
## [17] "Equestrianism" "Fencing"   
## [19] "Figure Skating" "Football"   
## [21] "Freestyle Skiing" "Golf"   
## [23] "Gymnastics" "Handball"   
## [25] "Hockey" "Ice Hockey"   
## [27] "Judo" "Lacrosse"   
## [29] "Luge" "Modern Pentathlon"   
## [31] "Motorboating" "Nordic Combined"   
## [33] "Rhythmic Gymnastics" "Rowing"   
## [35] "Rugby" "Rugby Sevens"   
## [37] "Sailing" "Shooting"   
## [39] "Short Track Speed Skating" "Skeleton"   
## [41] "Ski Jumping" "Snowboarding"   
## [43] "Softball" "Speed Skating"   
## [45] "Swimming" "Synchronized Swimming"   
## [47] "Table Tennis" "Taekwondo"   
## [49] "Tennis" "Trampolining"   
## [51] "Triathlon" "Tug-Of-War"   
## [53] "Volleyball" "Water Polo"   
## [55] "Weightlifting" "Wrestling"

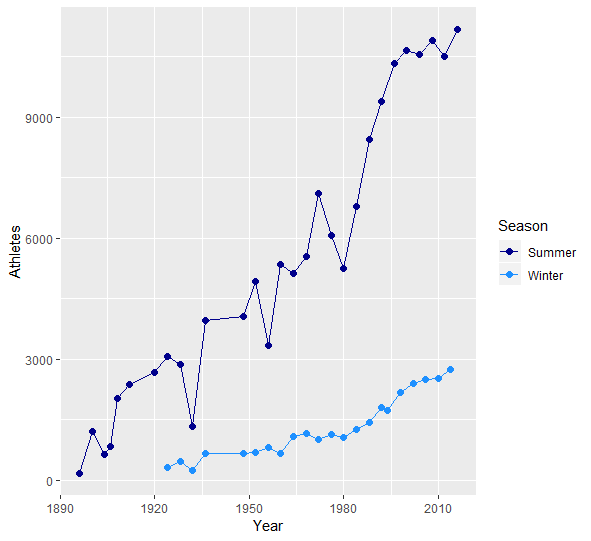
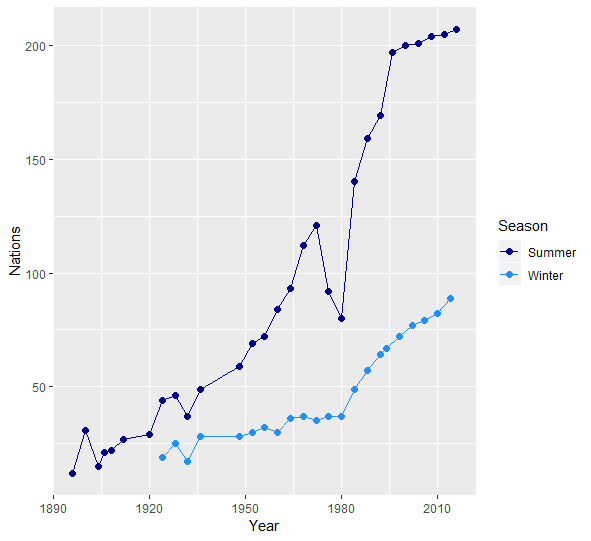
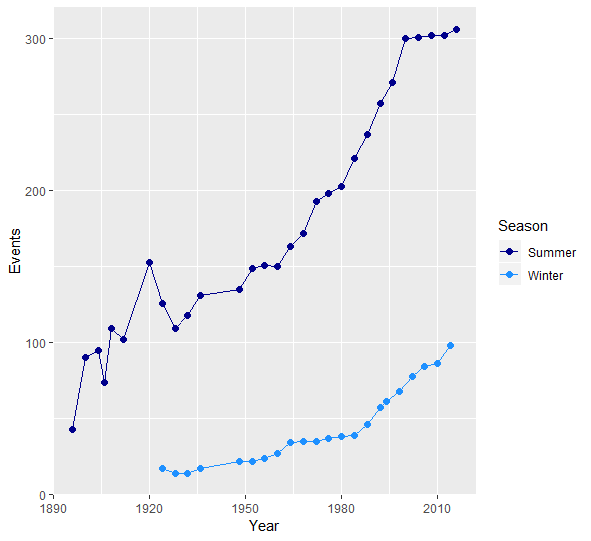
Medal values are checked for validity. An NA is allowed in this column as it indicates that an athlete did not medal in that particular event. Other allowable values are Gold, Silver, and Bronze. Each region should be associated with one NOC. The NOC value is used when region is NA and duplicates are subsequently corrected.

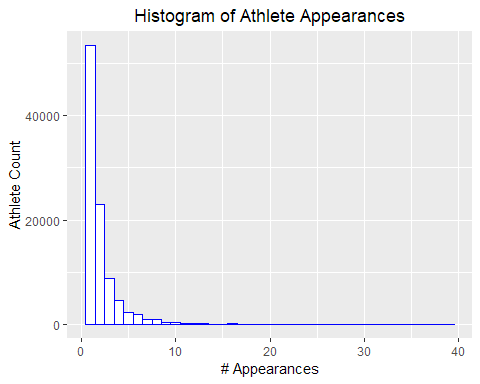
For classification, the data is limited to Men’s Swimming. This allows for a more valid comparison than trying to classify across sports, seasons, and genders.

For additional techniques, Winter and Summer games after 1992 are combined. The year is cast as an integer and any records with a sport of “Art Competition” are removed. Any NAs in the medal column are replaced with 0. An additional medal column is created with binary values, 1 for athletes who received a medal and 0 for athletes who did not. Two datasets of complete cases are created with Age, Height, Weight and Team; each including one form of the medal columns. The Teams and Medals columns are cast as factors and test & training sets are constructed from each. Label variables are created for each set, and the Age, Weight and Height data is normalized for clustering.

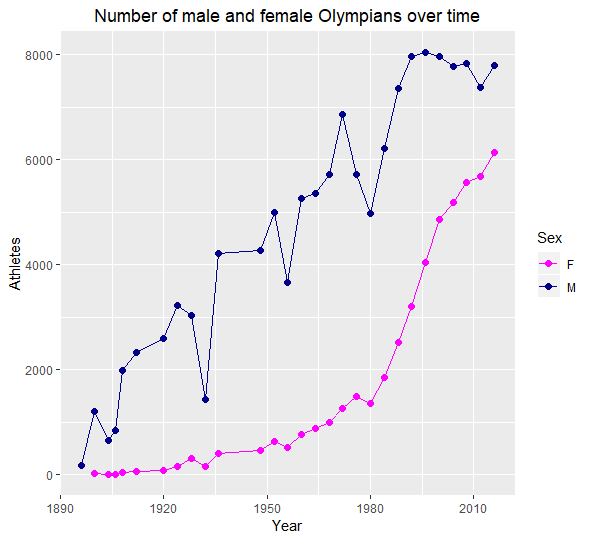
### Exploratory Data Analysis

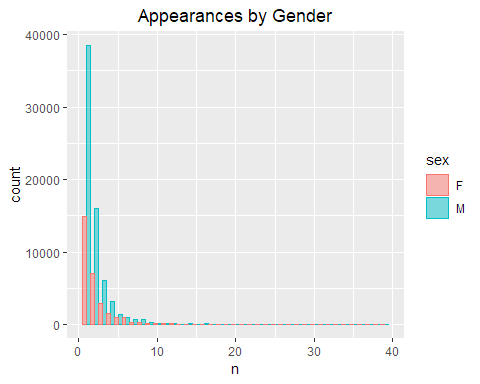
Athletes, nations, and events are examined by season over time: Dips were due to economic hardship/depression and several boycotts by different nations throughout the years.

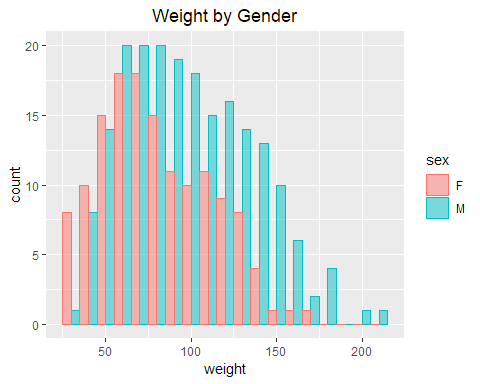
  



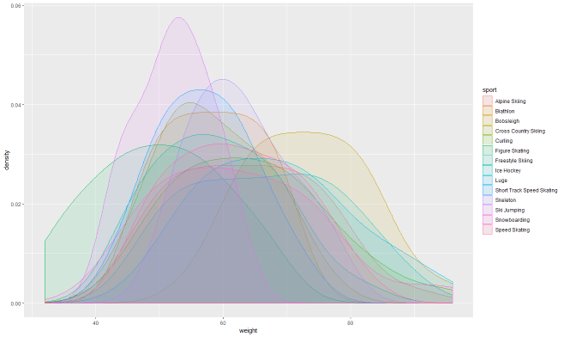
Most athletes appear in only one event in one Olympic Games, but one athlete has appeared in 39 events! Although there have been more male athletes than female in Olympic history, the distribution of events is not significantly different across genders.



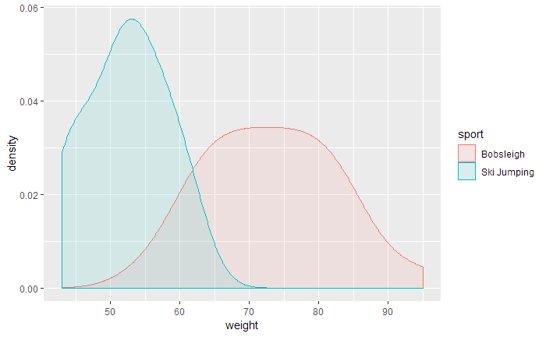


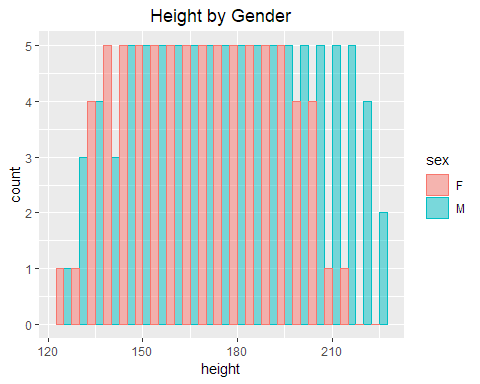


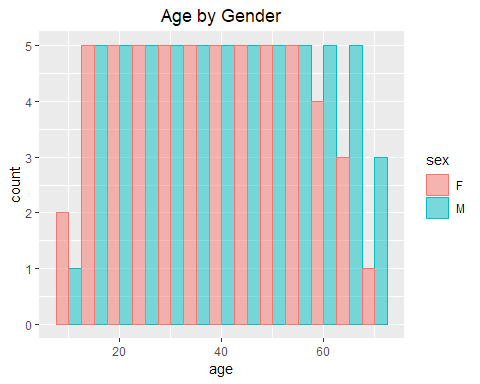
**Weight Distribution for Female Winter Sports**



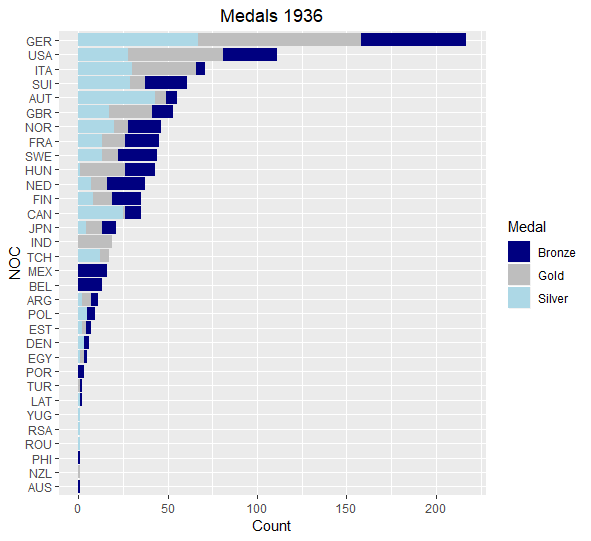
**Weight Distribution for Female Bobsleigh & Ski Jumping**



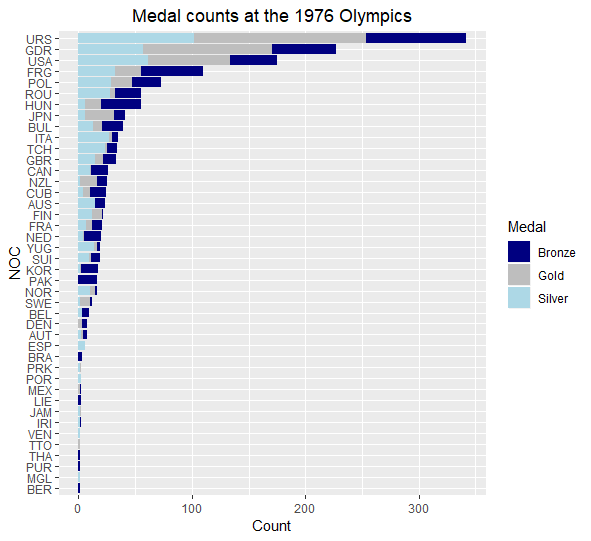




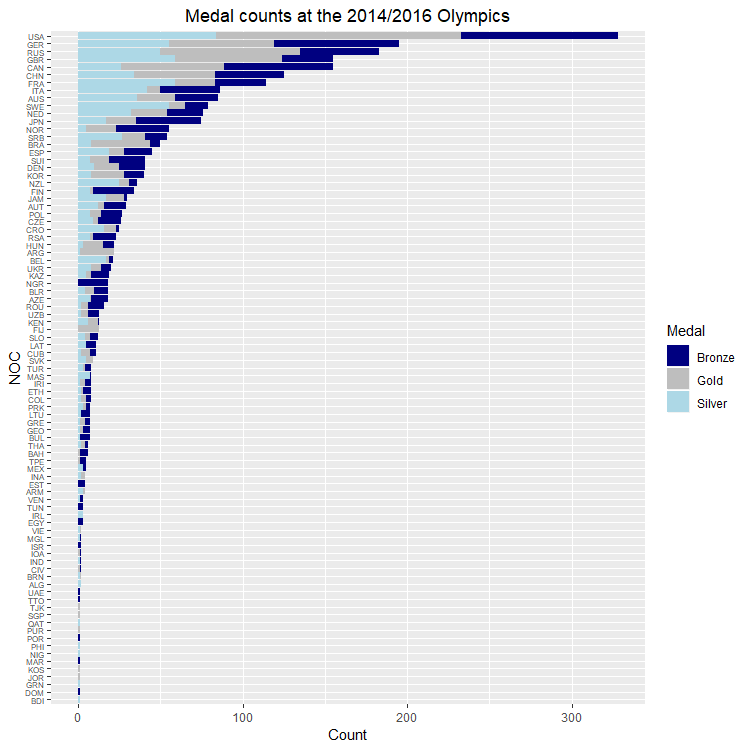
The evolution of medal counts over time is examined.



In 1936, Germany, US, Italy had the most medals. Germany and Italy were dominating in the World War and the Olympics



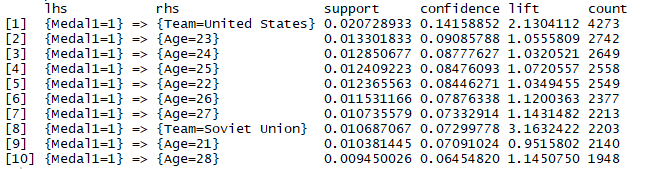
In 1976, Russia, Germany, and US topped the medal counts. During the Cold War era, Russia was doing well economically and women did particularly well in the games.



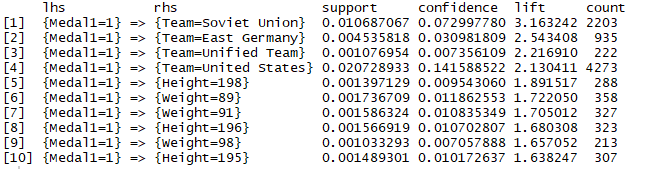
In 2016, US, Germany, and Russia continue to dominate the medal counts.

### Association Rule Mining

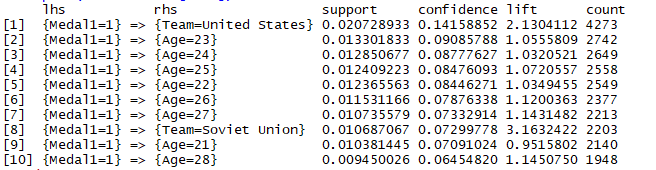
Association rules are discovered using the apriori algorithm. Below are the top ten rules sorted by lift.



Next are the top 10 rules sorted by support.

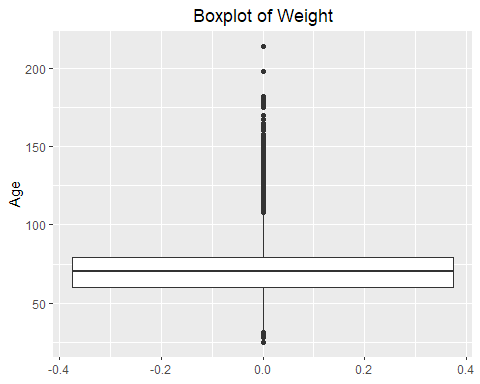
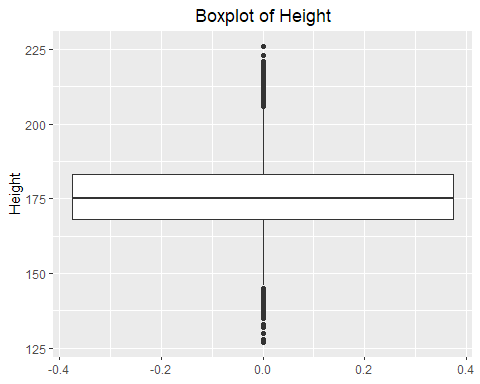
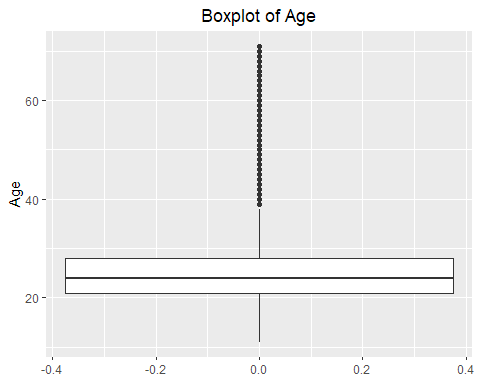


Finally, the top 10 rules sorted by confidence

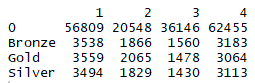


### Clustering

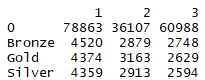
#### K-Means Clustering



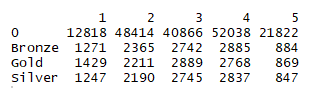
K-Means clustering is performed with k = 4.



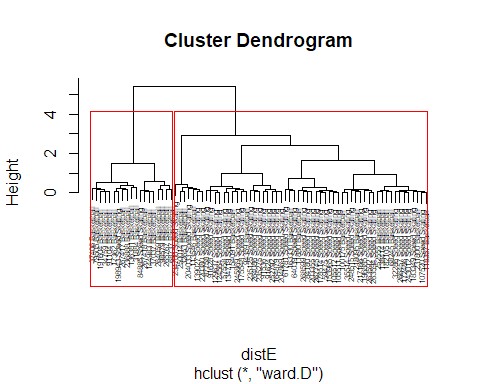
K-Means clustering is performed with k = 3.

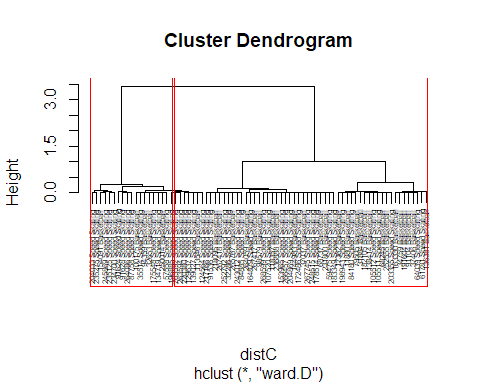


K-Means clustering is performed with k = 5.



#### Hierarchical Clustering





### Decision Tree

CART   
   
 2606 samples  
 3 predictor  
 2 classes: '0', '1'   
   
 No pre-processing  
 Resampling: Cross-Validated (3 fold)   
 Summary of sample sizes: 1737, 1738, 1737   
 Resampling results across tuning parameters:  
   
 cp Accuracy Kappa   
 0.0 0.6085945 0.2171651  
 0.1 0.5893999 0.1787994  
 0.2 0.5215058 0.0437788  
 0.3 0.4996164 0.0000000  
 0.4 0.4996164 0.0000000  
 0.5 0.4996164 0.0000000  
 0.6 0.4996164 0.0000000  
 0.7 0.4996164 0.0000000  
 0.8 0.4996164 0.0000000  
 0.9 0.4996164 0.0000000  
 1.0 0.4996164 0.0000000  
 Accuracy was used to select the optimal model using the largest value.  
 The final value used for the model was cp = 0.

### Random Forest

2606 samples  
 3 predictor  
 2 classes: '0', '1'   
   
 No pre-processing  
 Resampling: Cross-Validated (3 fold)   
 Summary of sample sizes: 1736, 1738, 1738   
  
 Accuracy was used to select the optimal model using the largest value.  
 The final value used for the model was mtry = 1.

### Naive Bayes

2606 samples  
 3 predictor  
 2 classes: '0', '1'   
   
 No pre-processing  
 Resampling: Cross-Validated (3 fold)   
 Summary of sample sizes: 1737, 1738, 1737   
   
 Accuracy was used to select the optimal model using the largest value.  
 The final values used for the model were fL = 0, usekernel = TRUE  
 and adjust = 1.

### SVM

#### Linear Kernel

Support Vector Machines with Linear Kernel   
   
 2606 samples  
 125 predictor  
 2 classes: '0', '1'   
   
 No pre-processing  
 Resampling: Cross-Validated (3 fold)

Accuracy was used to select the optimal model using the largest value.

#### Polynomial Kernel

Support Vector Machines with Polynomial Kernel   
   
 2606 samples  
 125 predictor  
 2 classes: '0', '1'   
   
 No pre-processing  
 Resampling: Cross-Validated (3 fold)

Accuracy was used to select the optimal model using the largest value.  
 The final values used for the model were degree = 2, scale = 0.1 and C= 0.75.

#### Radial Kernel

Support Vector Machines with Radial Basis Function Kernel   
   
 2606 samples  
 125 predictor  
 2 classes: '0', '1'   
   
 No pre-processing  
 Resampling: Cross-Validated (3 fold)

Accuracy was used to select the optimal model using the largest value.  
 The final values used for the model were sigma = 1 and C = 1.

### Text Mining

1936



1976



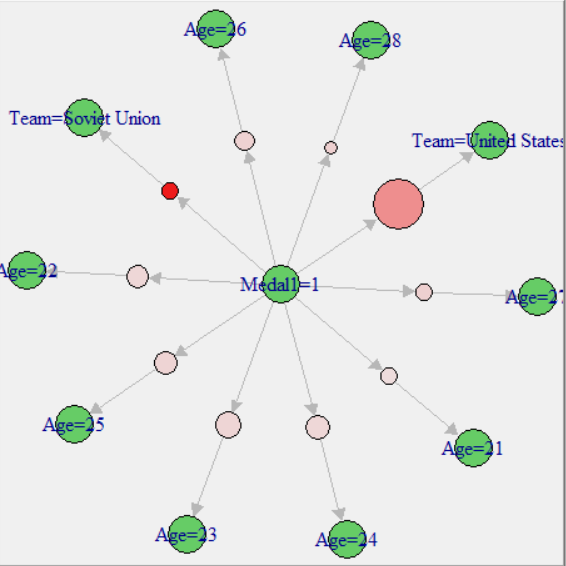
2016



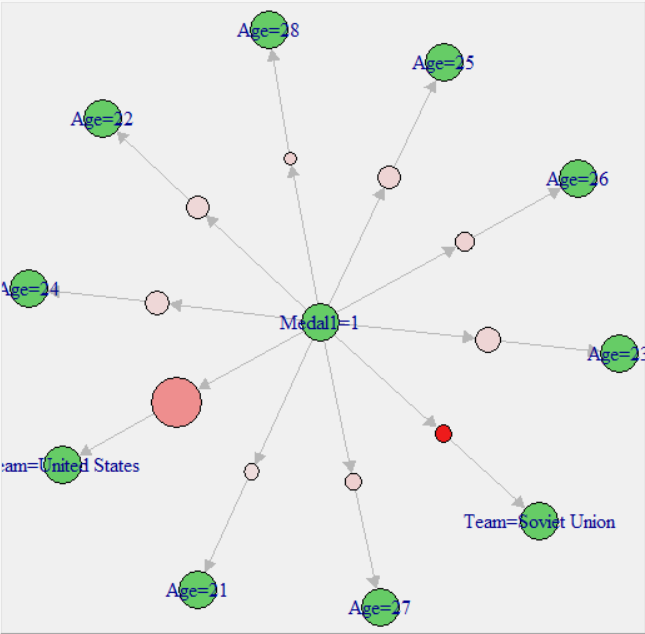
## Results

### Association Rule Mining

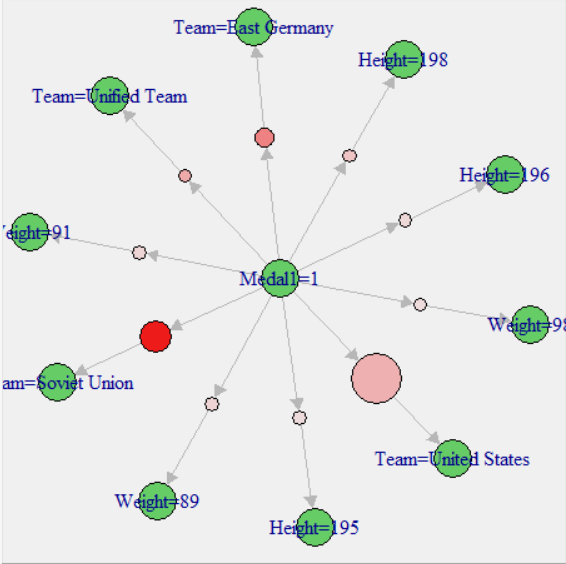
Top rules by lift:



Top rules by confidence:

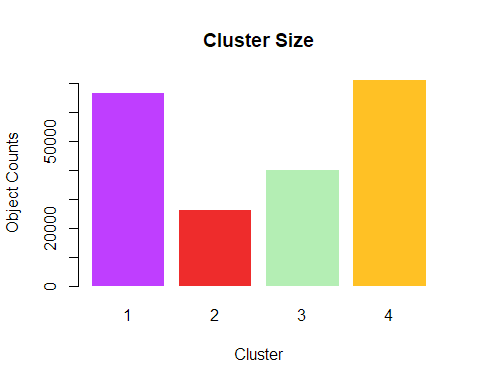


Top rules by support:

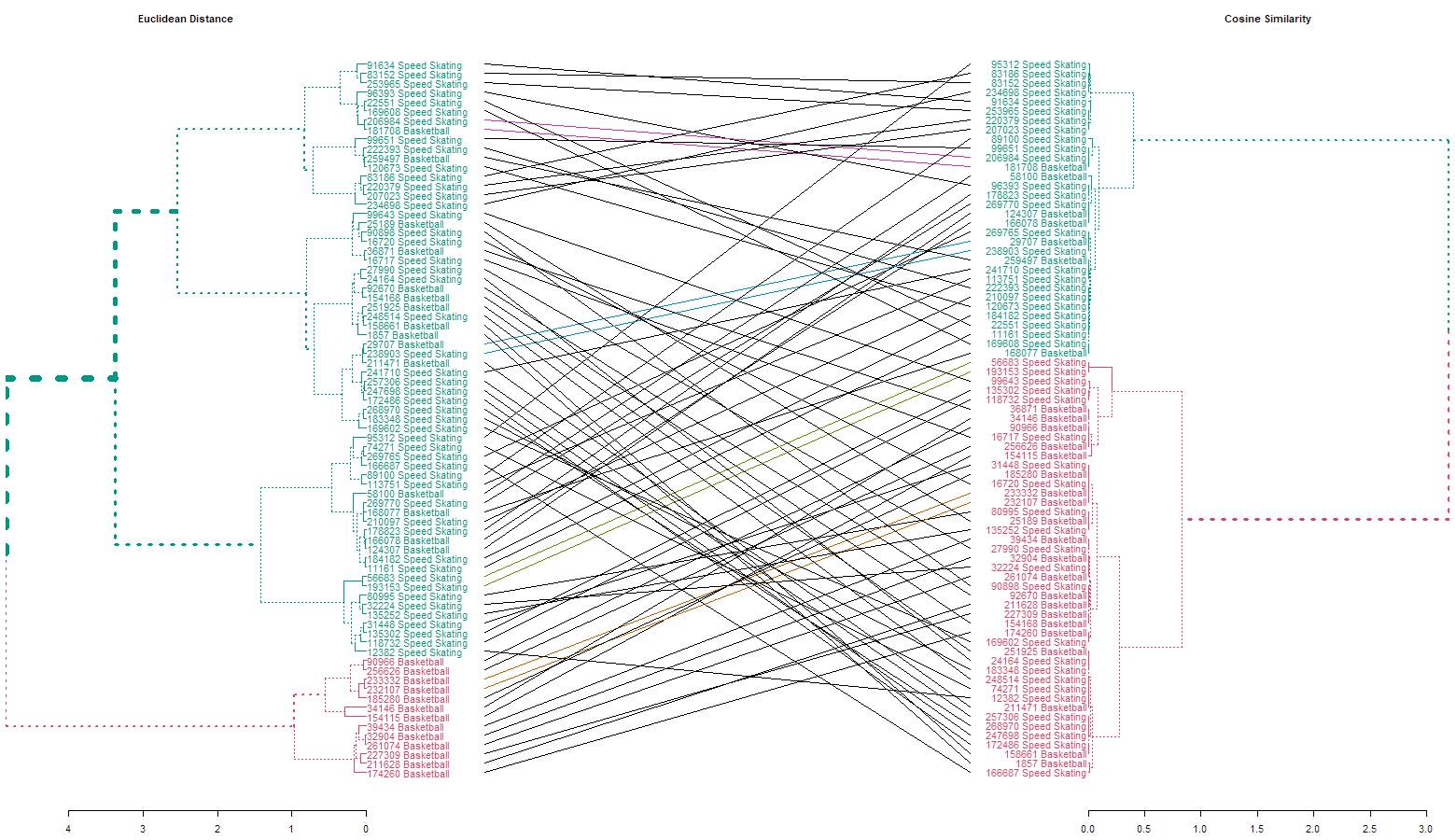


### Clustering

#### K-Means Clustering

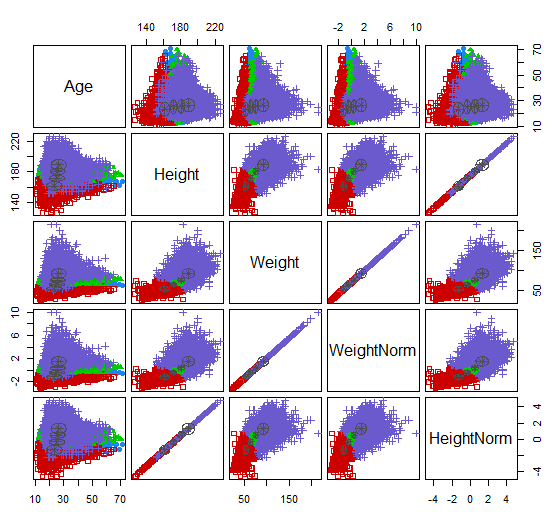


#### Hierarchical Clustering

****

An entanglement value of 0.56 confirms that the two dendrograms are not very similar.

#### Expectation Maximization Clustering



### Decision Tree

##   
## preds 0 1  
## 0 962 363  
## 1 341 940

### Random Forest

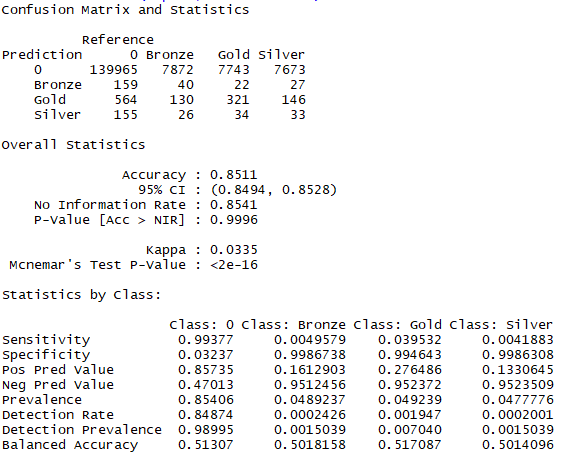
##   
## preds 0 1  
## 0 1078 187  
## 1 225 1116

### Naive Bayes

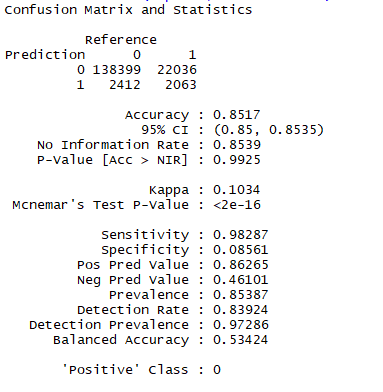
When predicting whether a male swimmer will medal, the following confusion matrix is generated:

##   
## preds 0 1  
## 0 723 488  
## 1 580 815

When predicting on all data whether an athlete will receive a Gold, Silver, Bronze, or no medal:



When predicting on all data whether an athlete will medal or not:



### SVM

#### Linear Kernel

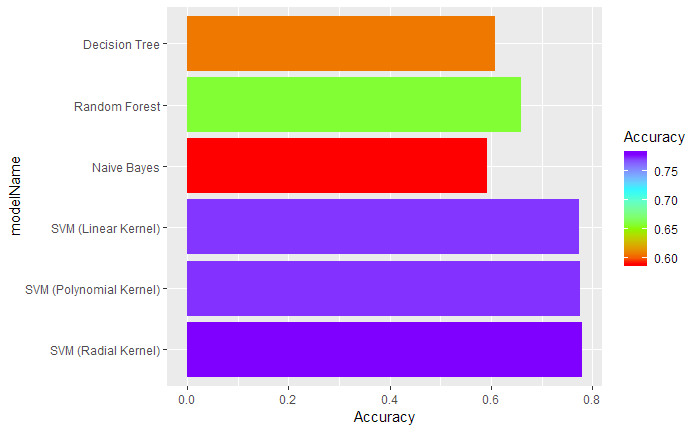
##   
## preds 0 1  
## 0 1072 335  
## 1 231 968

#### Polynomial Kernel

##   
## preds 0 1  
## 0 1008 202  
## 1 295 1101

#### Radial Kernel

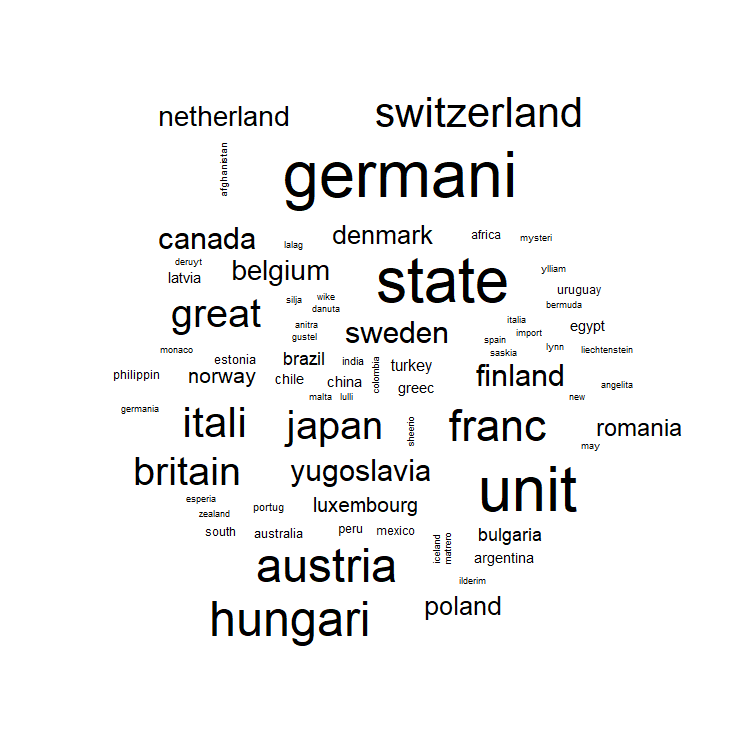
##   
## preds 0 1  
## 0 1105 176  
## 1 198 1127



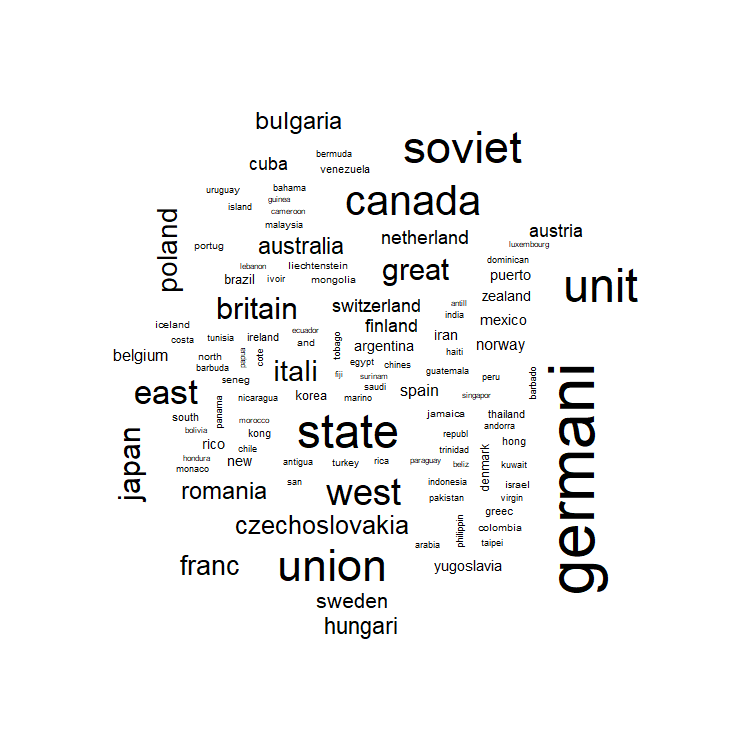
Overall, in predicting which male swimmers will medal, the SVM using a radial kernel performed best with 78% accuracy.

### Text Mining

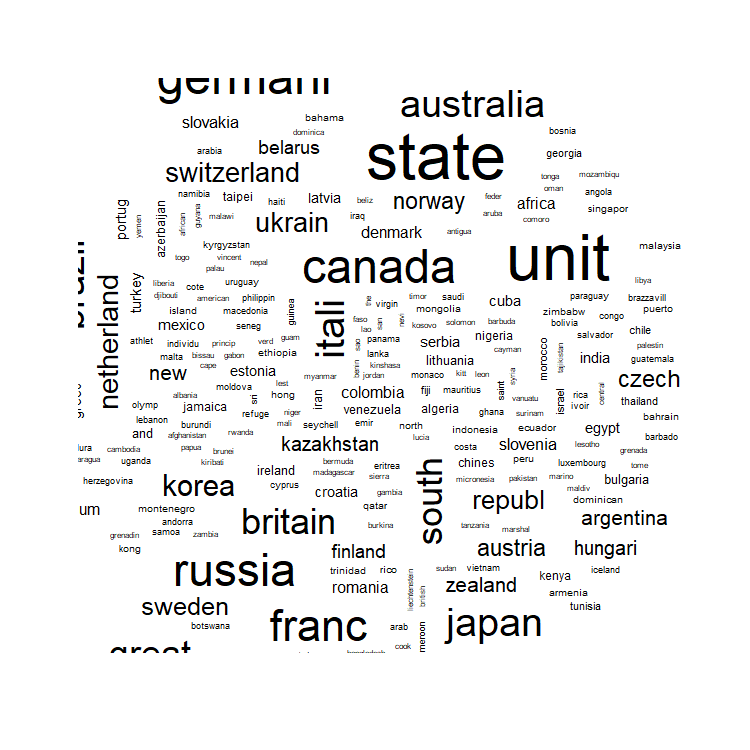
Word cloud for countries competing in 1936 Olympics:



Word cloud for countries competing in 1976 Olympics:



Word cloud for countries competing in 2016 Olympics:



## Conclusion

Examining the data associated with over 270,000 Olympic athlete participation records revealed some interesting trends. Overall, the size of the Olympic games has grown exponentially. Today, there are significantly more athletes, nations, and events at each Olympic games than could have been conceived of in 1896. Although the US continues to land near the top of the medal count, more athletes than ever before have the opportunity to strive for a medal every two years. In addition, women are competing in the Olympics at a rapidly increasing rate, closing the representation gap between the two genders. It is no longer beyond the realm of possibility that female athletes could one day outnumber male athletes at the Olympics. Just as competitors in 1896 could not imagine where the games would be today, the possibilities looking another 100 years in the future are innumerable.

In predicting the outcome for a certain athlete, country is one of the most significant factors. This makes sense, as unfortunately athletes from wealthy countries generally have access to more sophisticated training methodologies than those from less-wealthy countries. Additionally, while wealthy countries can support year-round training facilities for both Winter and Summer athletes, smaller nations may not be able to do so. Some equatorial nations rarely, if ever, produce athletes competing in Winter events because they don’t have the training conditions necessary to achieve excellence. Equipment cost can also be prohibitive for a country to produce Olympic-level athletes. While sports like distance running may only require special shoes (and sometimes no shoes at all), sports like swimming require expensive suits that often last for only 3 races. Team sports such as ice hockey even require multiple athletes to have full sets of professional equipment which can cost hundreds, if not thousands, of dollars.

An interesting follow-up analysis would be to monitor the performance of the countries that host the Olympics prior, during, and after they host to see the impact. The next Summer Olympics is set to take place in Japan in 2020. Olympic Legacy is the result of a vision. It encompasses all the tangible and intangible long-term benefits of hosting the Olympics for the people, territories, and the Olympic movement. Olympic Legacy is about the creation of positive changes that go beyond the 17 days of celebrations and competitions. Examples of these benefits are: training centers and facilities used to improve the performance of athletes, increased enthusiasm for less popular sports, new diplomatic relations and improved dialogue between countries, enhanced professional skills and career opportunities, new cultural heritage assets, growth of the volunteering movement, innovative design and visual arts, and increased global visibility for cities, upgraded urban parks and leisure areas, and national/regional cultures. For 17 days every two years, the world comes together to celebrate as one. Regardless of medal count, team size, and sponsorship the Olympics are the pinnacle of international cooperation and should continue to be celebrated as such.

1. <http://www.oldest.org/sports/olympians/> [↑](#footnote-ref-1)
2. <https://www.sports-reference.com/olympics/athletes/ch/choi-myong-hui-1.html> [↑](#footnote-ref-2)
3. <https://www.sports-reference.com/olympics/athletes/bl/ricardo-blas-jr-1.html> [↑](#footnote-ref-3)