Group Name: Grid Lines

Elizabeth Nieto, Chaitrali Panchal, Sarita Patel, Sean Sullivan, Chloe Wang

Olympic Data: Through the Ages

Introduction

In a world with so much turmoil, the Olympics serves as an event that momentarily slows us all down and connects the world for a common cause: competition. When it comes to the Olympics, the thing that we all care most about from each competition is who won. Which countries won the most medals? Which athletes broke records or upset the leading contenders? Has a newly participating country dethroned the previous champion? Or perhaps an unassuming athlete has cleared the house. The details about the winner and the number of medals awarded to a country are the details people remember. We used this to guide our analyses and wanted to work in the theme of how the Olympics have changed over time. The dataset begins in 1896 and since then, our world has experienced immense change. Specifically, our project used a choropleth map, a contingency table, a line graph, a parallel coordinate plot, and network graph on an Olympic dataset to explore characteristics of athletes and each country's participation statistics. The following is a list of topics and questions our team sough to answer.

Male vs. female ratio for every sport
Number of medals won by athlete
Average age of the athletes
Total medals earned by host cities
Performance of medaling athletes
Time series of countries' medal and participant counts per game.
How many athletes, sports, and nations are there?
Where do most athletes come from?
What is the characteristic of the athletes (e.g., gender and physical size)?

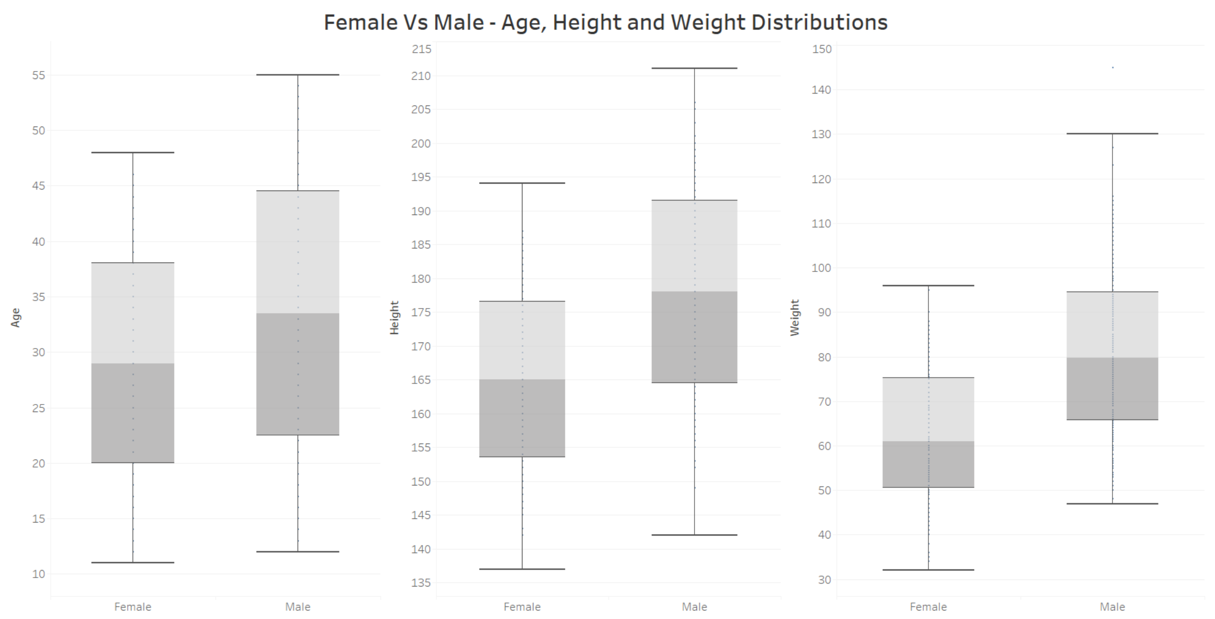
The data set was retrieved from Kaggle as the file, athlete_events.csv. It contains 271,116 data points of Olympic data for every summer (222,553) and winter (48,565) Olympic game held between 1896 and 2016. Each data point is an athlete participation record for a specific sporting event and contains 15 variables:

ID	Unique number assigned to each athlete
Name	Athlete's name
Sex	Athlete's Sex (M/F)
Age	Athlete's Age
Height	Athlete's height(cm)
Weight	Athlete's weight(kg)
Team	Athletes country
NOC	National Olympic Committee 3-letter country code
Games	Year and season
Year	Year Olympic games were held (1896 - 2016)
Season	Season of event (Summer/Winter)
City	Host city
Sport	Olympic sport category
Event	Competitions held within a sport
Medal	Medal type received (Gold, Silver, Bronze, or NA)

Exploratory Analysis

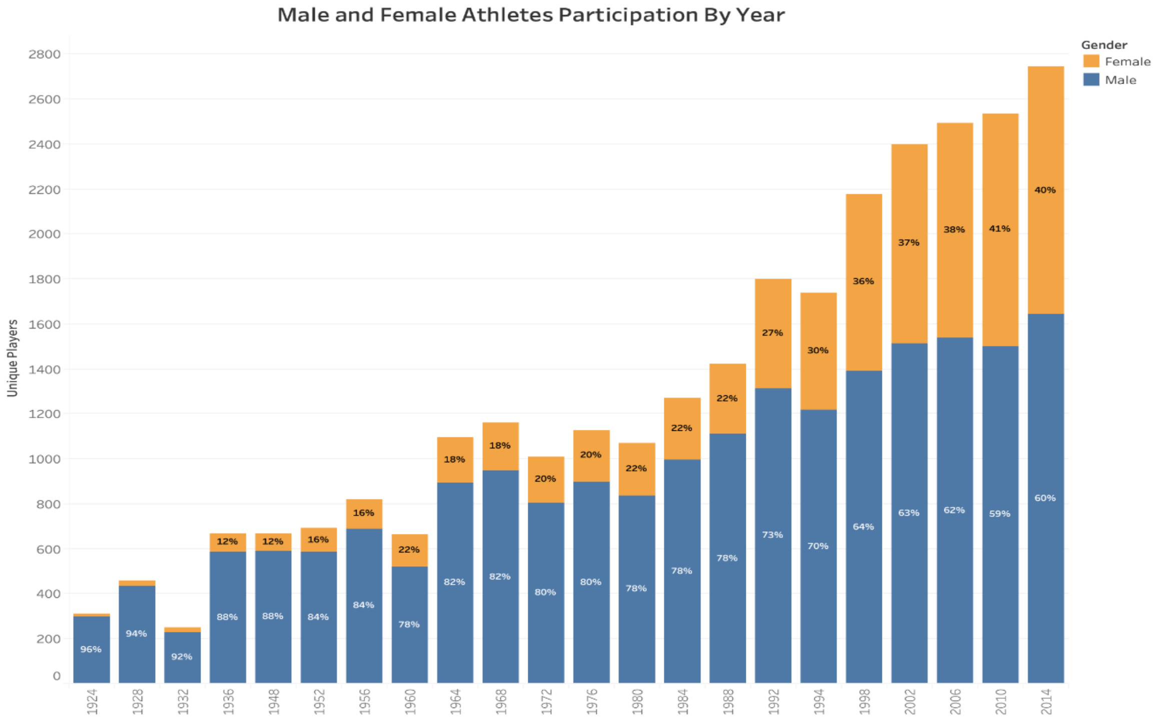
Before diving into different tracts of analysis, we needed to gain a basic understanding of the characteristics of our dataset. We explored the distribution of age, height, and weight of athletes that participated in the Winter Olympics. We also evaluated the proportion of male and female participation in the Winter Olympics – which proved to be a point of emphasis for one of the core visualizations. Another group member explored height and weight of athletes participating in the Summer Olympics and plotted the distribution of weight for athletes from 1896 to 2016. These three exploratory visualizations are merely a representation of the total number of visualizations that were created and the remainder of these can be found in the Appendix section of the report.

One of the key aspects in the Winter Olympics that the team analyzed and compared is the Male and Female Athletes participation and other criteria in the Winter Olympics. We firstly looked at the distribution of age, height and weight of Male and Female athletes that participated in the Winter Olympics. Below are the box plots that compares the distribution of age, height and weight for Male and Female athletes.

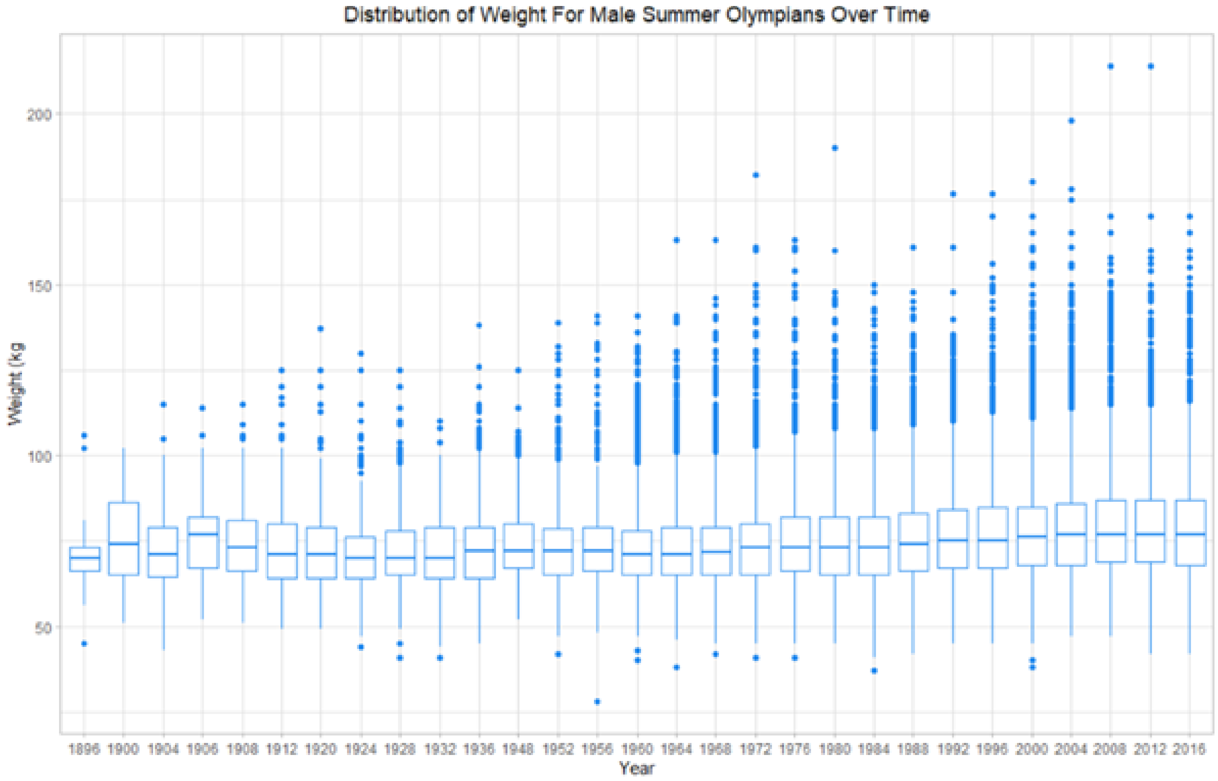


The above distributions suggest that the overall Female athletes are younger than the Male athletes, are shorter in height than the Male athletes and weight significantly less than the Male athletes.

Next, we look at the Male and Female Athletes participation by year. For this we created a bar plot of number of unique athletes participating by Year and divided by gender.



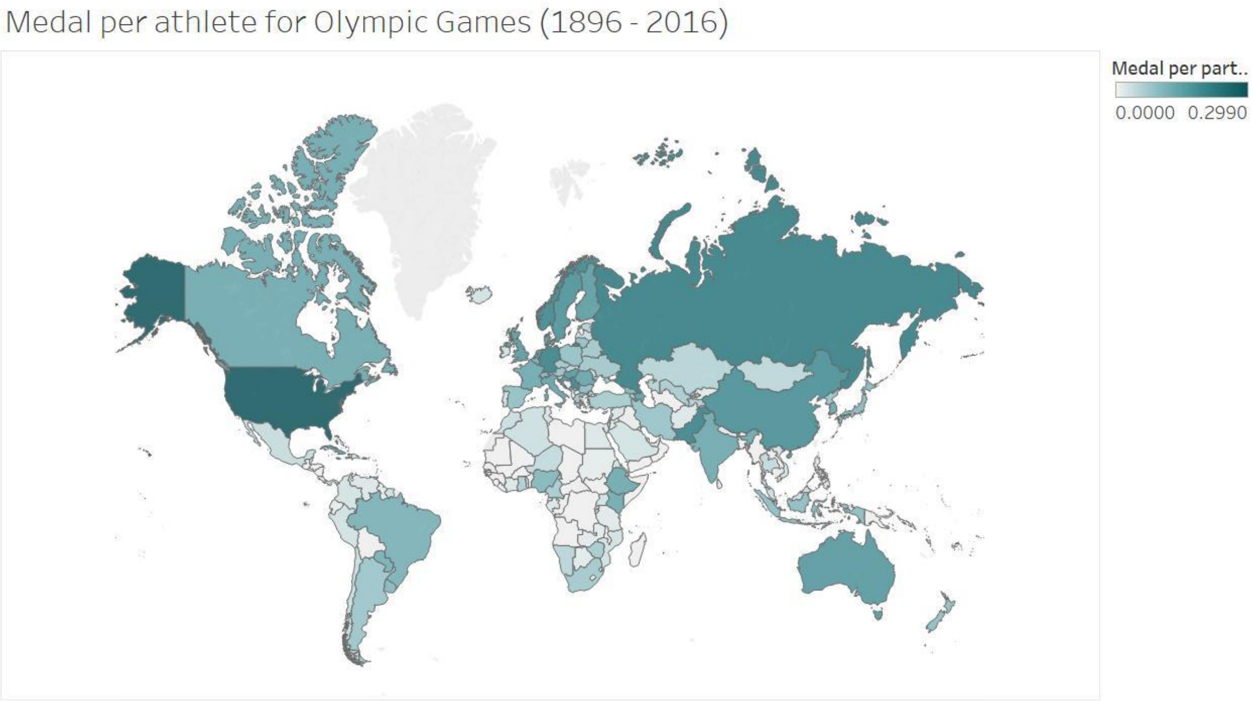
The above bar plot tells us that the participation in the Winter Olympics has increased yearly for both Male and Female athletes. During the initial Winter Olympics, the participation was almost entirely Male athletes. Since then that trend has changed and in some of the recent Winter Olympics, we see women athletes represent around 40% of the total participating athletes. The overall number of participating athletes has steadily increased with two major declines in participation in 1932 and 1960.



Two important variables that we worked with were height and weight of the participating athletes. To gain an understanding of the changes to both variables over time, we plotted the distribution of weight for male athletes participating in the Summer Olympics from 1896 to 2016. The visualization above quickly shows a steady, yet slight, increase in median weights of athletes through time, but the interesting observation lies in the number of outliers that are present as the data moves toward 2016. Similar exploratory visualizations were created for female athletes and height and are in the appendix.

Visualizations (Use these visualizations for grading purposes)

Visualization C1

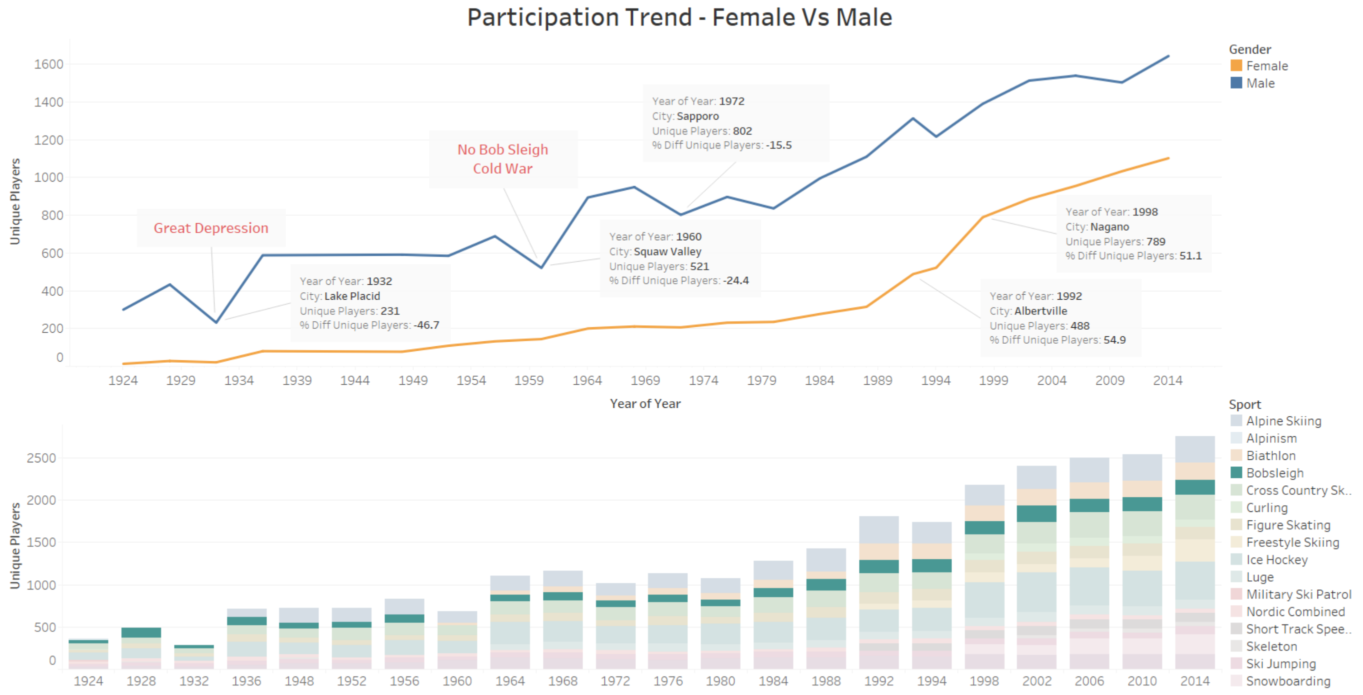

Medal per athlete for Olympic Games (1896 - 2016)

In the graph above, we can see two important things when it comes to Olympic games; Which countries won the most medals and what is the participation level from each country from the year 1896 through 2016. The color scheme is based on the medal per participant rate, which is calculated by dividing the number of total medals won by the total number of participants, for each country. From the visual above, we can observe medal per participation rate, in the Olympics is the highest for USA at .29, which is the highest, followed by Russia with a rate of 0.22. This means that out of total number of athletes participating from US there are 29% of participants winning the medals in game. We have some of the countries in South Africa where we have 0 participation in any of the games held in Olympic therefore, we have the medal rate for them as 0.0.

The visual keeps filters for years in which the Olympic Games were organized and Seasons as winter and summer, as well, which can help in customizing the results.

Overall, we see a lot of countries participating in these games every year. We have some countries participating which have a larger population rate with a larger number of athletes sent to the Olympic games like Russia, United States, and Canada but that does not imply that higher population results in higher medal rate. The number of medals won are higher as we can see some smaller area and population wise countries like Germany, Norway and Serbia with the medal rate as 0.22, 0.20 and 0.21 respectively which shows a notable performance from the athletes from these countries.

Visualization C2



As we noticed in Exploratory analysis the overall participation dipped in the year 1932 and 1960. We further researched to find the reason behind and looked at the participation trend of Male and Female Athletes in the Winter Olympics. Now, we explore this a bit further. The above image contains two graphs. The first graph is a line plot that gives the number of unique Male and Female athletes that participated in the Winter Olympics by year with some important annotations for key events. The second graph gives the number of athletes that participated in the Winter Olympics by Year and Sport with a highlight on the Bob Sleigh event. The second graph help explains a finding in the first plot. In the first line graph we can see that there is a drastic dip in male participation in year 1932 and 1960.

Due to great depression the participation from male athletes was low in 1932.

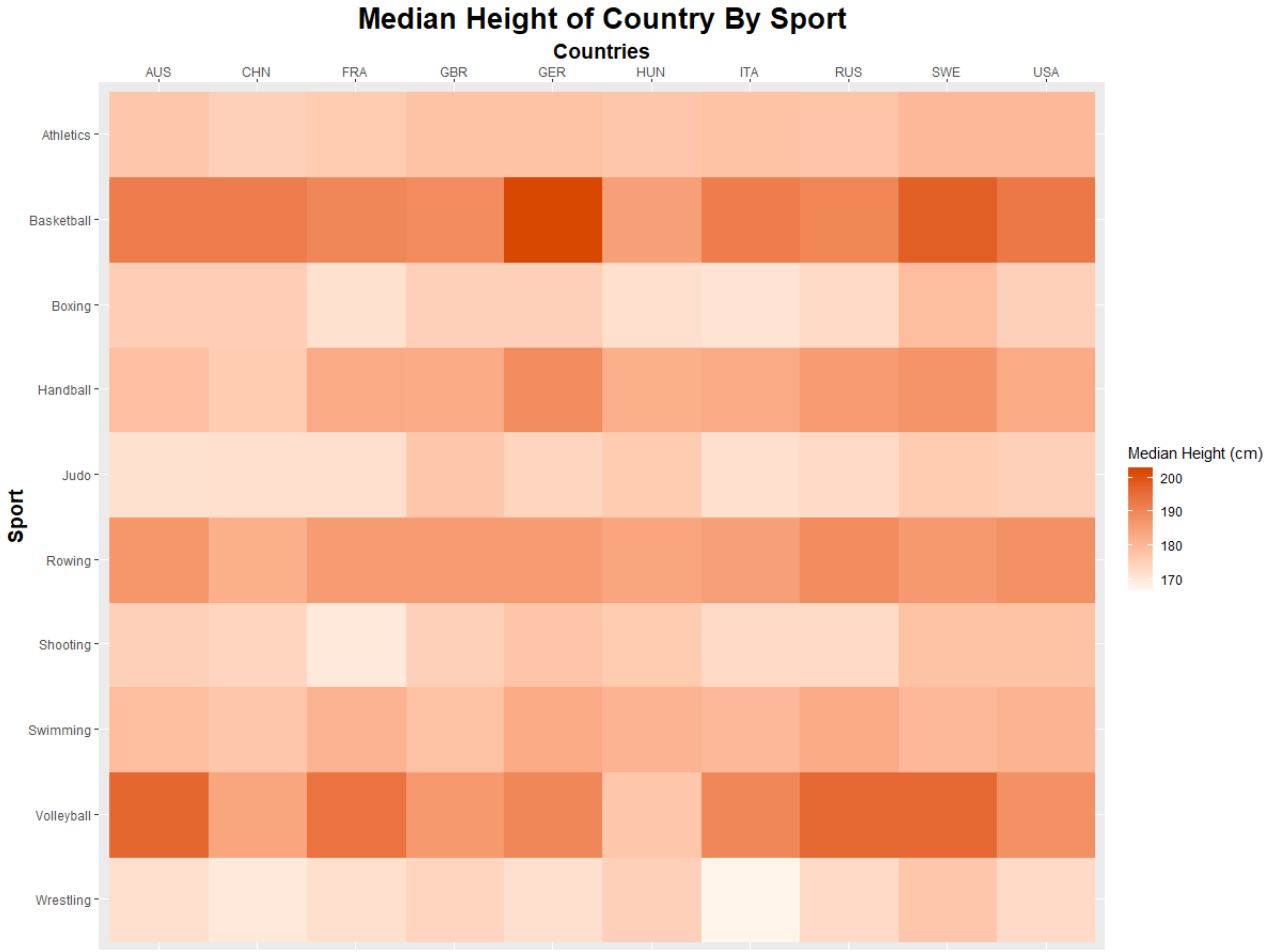
And in the year 1960 the male participation went down because of two reason:

1. No Bob sleigh- In history of winter Olympics it was the first time that bob sleigh was not happening due to some budget constraints (second part of the graph shows histogram for same).

2. Cold war – Many of the countries did not sent their athletes due to the ongoing cold war situation.

And Overall participation from the female athletes seems improving.

Visualization C3

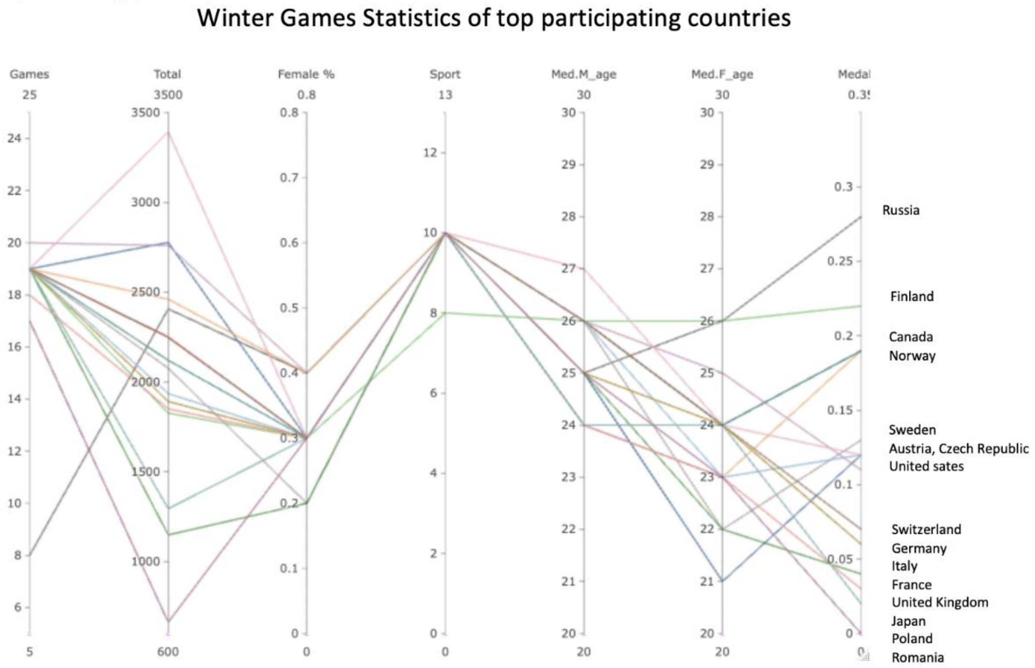


One aspect of the dataset that warranted further exploration involved the physical attributes of the Olympians. For this, we had the heights and weights of Olympians available for analysis. Initially, we began the exploration by looking at the distribution of heights and weights for each sport and quickly observed that some sports had wider distributions, meaning that there were Olympians who were at either end of the height or weight spectrum for that sport. The objective of this analysis was to further explore how the perception of a sport and the ideal physical attributes it takes to successfully compete, compare with the actual attributes of the athletes competing. To cut down on noise, we decided to only focus on the Summer Olympics and created a methodology for further honing the dataset to effectively tell a story. We included the ten countries who have won the most medals in the Summer Olympics (Australia, China, France, Great Britain, Germany, Hungary, Italy, Russia, Sweden, and the United States) and ten sports that had the largest difference between the maximum and minimum heights and weights of competing athletes (Athletics, Basketball, Boxing, Handball, Judo, Rowing, Shooting, Swimming, Volleyball, and Wrestling).

The result of the analysis is the visualization you see above, which is a contingency table. The visualization has two categorical variables (Sport and Country) and a numerical variable (median height of athletes), where the numerical variable is encoded to a sequential color scheme. (Please note that a similar visualization for this information, but plotting median weight is available in the appendix: C3 Figure 5). The color scheme allows an observer to quickly understand how the median height of a country's athletes compare to other countries, by comparing the lightness of the color in the tile associated with that country. For example, Germany's basketball players are on average, taller than the other countries included in the visualization. We know this because the darker the cell, the taller the median height is and the lighter the cell, the shorter the median height is.

For the most part, the countries included in the visualization had similar median heights for each sport, but it was interesting to observe subtle differences made visible by the sequential color scheme. Through this, we can evaluate each sport and compare the results to our pre-conceived notions about the attributes of athletes for a sport. Volleyball is a great example of this because we tend to associate volleyball with taller athletes and while that remains true in the findings in the visualization, Hungary and China's median heights are quite shorter than the other countries present, namely Australia, Russia, and Sweden. Scaling this visualization for additional sports and countries, or even using this as a template for other sports and country combinations would certainly yield interesting results and aid in continuing to improve our understanding of how the attributes of athletes compare across countries for the various sports included in the Summer Olympics.

Visualization C4



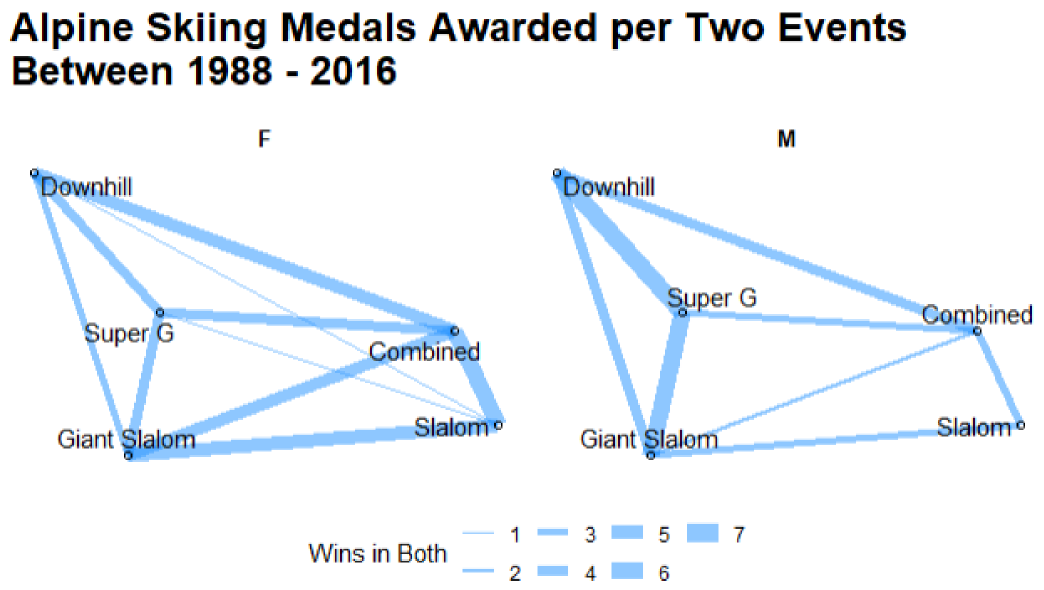
We created a parallel coordinates plot to visualize Olympic winter games census of countries. All the observations are aggregated by NOC (country). Sports that have discontinued were filtered out. The plot was constructed on preprocessed variables. Top 16 countries have the largest athletes' population is showing in the graph. It contains seven variables. (calculations are explained in appendix 2 C4--table 1)

1. Games: shows the number of games each country has participated throughout winter game history.
2. Total: is the sum of athletes each country has sent to Olympics winter games.
3. Female %: is the female athletes participation rate.

4. Sports: how many types of sports each country has played.
5. Med. M_age: median age of males
6. Med. F_age: median age of females
7. Medal: winning rate that is number of medals they won divided by the total number of athletes.

Each line represents the winter Olympic games statistics of a country. Lines are sorted, descending by Medal. By looking at Total, the U.S has been active in participating in the winter Olympics. It attended 19 games and the population of the athletes is the greatest. Female athletes have larger range of ages than that of male athletes. By comparing Medal, Russia is fruitful in winning medals. Romania on the other hand, has been participating in the games but the outcome is not as comparable to the countries listed. Finland only participated in 8 sports, but the medal rate is higher than most of the countries. It might suggest that they are good at the sports they participated.

Visualization C5



In addition to understanding the characteristics of top participating countries of winter Olympics in the last visual, another area of interest within the winter Olympic games is sporting event participation and medaling athlete participation within a sport. To narrow our focus and avoid sports with inconsistent event history, it was decided only the most popular winter sport would be explored; most popular in this case being the sport with the highest number of participants since the sports inception. For winter the two most popular sports were Cross Country Skiing with a total of 9,133 participation records and Alpine Skiing with 8,829 participation records (Appendix 2 C5 Figure 1). Exploring these two sports further revealed an inconsistent event history for cross country skiing with few events being played consistently on consecutive Olympics. Alpine Skiing, however, had a consistent event history for 10 events (5 for men and 5 for women) distinguishable between 1988 and 2016 (Appendix 2 C5 Figure 2). It was decided medaling athlete event participation within this sport would be explored to understand what combination of events within Alpine Skiing resulted in the most athletes winning a medal in both.

A network graph was selected to visualize these relationships. Each of the five nodes represents an event within alpine skiing hosted separately for men and women: Downhill, Super G, Giant Slalom, Slalom, Combined. Nodes are connected by lines of varying width representing the count of athletes who won both events in the same Olympic games. Disconnected nodes represent an absence of an athlete winning both events in the same Olympics.

The first draft of this visual consisted of one network graph with male and female records displayed on the same graph using the same color for edges. The amount of overlap between genders and events made it difficult to decode record counts and gender. Small multiples for gender were applied to create to separate networks with less gender overlap. The second draft of this network graph consisted of two networks one for men's events and one for women's events. Each network had a line edge width range of .2 to 2 representing the number of records in which an athlete had won in both sports. Since the line range was small differences between record counts were not clearly distinguishable. To change this the range was extended to from .2-2 to .2-4. This wider range allowed for clearly distinguishable line widths representing record counts.

This visual provides an in-depth view of the second most popular winter sport. From this visual it is clear women who participated and won a medal in the event Slalom were seven times more likely to also win a medal in the Combined event than Super G or Downhill. For men, those who won in Super G were more likely to have won in the Downhill and Gian Slalom event. For both genders athletes who won in Slalom also won medals in Combined and Giant slalom but were less likely to also win in the Downhill or Super G event.

Analysis and Discussion

For many athletes across the globe, reaching the Olympic games is the ultimate accomplishment. With such a robust dataset, our group wanted to approach the data from a wide scope. We knew that there were many avenues to pursue for an analysis, but we decided to focus on showing how the Olympic games have changed and subtleties that lie within the dataset that challenged our own notions of the Olympics.

One of the first realizations that stood out to group was how global the Olympics have become. Visualization C1 does a thorough job of displaying this, as it is easy to quickly observe how many countries have participated in the Olympics from 1896 through 2016. While each country doesn't necessarily win the same number of medals or send the same number of athletes to the games, we know that participation is now on a global scale.

We also saw the introduction of female athletes in the early half of the 1900s and wanted to further explore how the participation of women changed over time. Visualization C2 plots the participation trends of male and female athletes, in the Winter Olympics, from 1924 to 2016. Through this visualization, it's easy to see that women are beginning to make up a larger proportion of the athletes participating in the Winter Olympics. Additionally, the visualization shows dips in participation for specific years and calls out the reasons as to why this happened. In both cases, it was from major events whose impacts were felt on a global scale (Great Depression and Cold War). In many ways, this dataset mimicked many of the main events and accomplishments of the twentieth century.

As we got further into our analysis, we became interested in the height and weight of athletes and how they may vary depending on the sport or country participating. After exploring general trends in the data, we focused on showing the difference of median height across the ten most winning countries, that participated in the Summer Olympics, and the ten sports with the highest variance of height amongst its participants. We were able to learn that it really depends most on the sport and country at hand. While similar trends were observed across sports, there were countries that stood out as either being generally taller or shorter than the competition.

We also wanted to evaluate if we could identify trends across the top performing countries for the Winter Olympics, in hopes of illuminating a blueprint for future success. Unfortunately, we observed a variety of trends and the true takeaway is that it doesn't matter how many athletes a country sends, or how young the athletes are, or how many events they participate in; if the country focuses on sending the best athletes, its medal winning rate will prove to be strong. Lastly, the performance of athletes who competed in the second most popular winter sport, Alpine Skiing, was analyzed to uncover trends amongst medaling athletes. Through visualization C5 we were able to observe how winning in one event related to wining in other Alpine Skiing Events.

With the visualizations included in our report, we truly believe that we've only scratched the surface. For instance, it might be interesting to evaluate how the physical attributes of medaling athletes have changed over time or how the height and weight of medalists compare to each other depending on the sport. Other areas to pursue could include researching how countries with lower participation rates tended to perform, in terms of medals won, and which countries have made more of an effort to send equal numbers of male and female athletes to the Olympics. Regardless of the research topic pursued, there is a plethora of data at our disposal and there certainly are insights waiting to be discovered.

APPENDIX

APPENDIX 1 – Individual Reports

Chaitrali Panchal – C1

In this project, I particularly focused on showing the performance of athletes from different countries. Because we had some geographical attributes available in our dataset, I decided on doing a choropleth which shows which country has been winning the most medals and if has some correlation with the participation level from different countries for example, higher the number of athletes higher the chances of winning the medals.

To display this information, I thought of creating a bivariate choropleth map, but I wanted to keep it easier to look interpret. In the beginning I created a choropleth map and used color scheme to display the number of participations with number of medals won as the labels. But there were two problems with this visual; 1 – The map looked cluttered and 2 – The information I wanted to show wasn't conveying the message effectively like it was difficult to compare which country has the less participation level yet has the higher number of medals won. One of my team members suggested we can do a calculated field using normalization. Which I thought was a great idea to improve my map so using the calculated field where the total number of medals won was normalized by the total number of participants from each country which gave us a percent rate of how many athletes have been winning medals in the Olympic Game. I plotted this newly created field on the choropleth, and I thought there is still some room for improvising this visual, so I did a data source join.

The calculated field was from the Mpp dataset and we had another dataset named Athlete events which contained attributes like year, countries and seasons for each game. In these two datasets we had one common field named NOC which was an alias for the name of countries, so I did a complete outer join for mpp and athlete events dataset by NOC from both the dataset and used the filters for years and seasons from Athlete events dataset with a map created from the Mpp dataset, that way we can created a custom visual by selecting combination of these filters like medal rate of countries for games held in summer **and/or** winter in the year 2004 **and/or** 2016.

Sarita Patel- C2

For the final project analysis, I was interested in identifying the features of male and female athletes participating in winter Olympics and for this I plotted box plot to see overall variation among age, height and weight. The distributions suggest that the overall Female athletes are younger than the Male athletes, have a lesser height than the Male athletes and weights significantly lesser than the Male athletes.

Then focused on one of the key aspects in the Winter Olympics that the overall Male and Female Athletes participation by year. For this I created a bar plot of number of unique athletes participating by Year and divided by whether they are Male or Female. I noticed during the initial Winter Olympics, the participation was from Male Athletes majorly but over the period that trend has changed and in some of the recent Winter Olympics we see Women Athletes constitute ~40% of the total athletes participating. also noticed that the overall participation dipped in the year 1932 and 1960.

On further investigation found that in 1932: Due to great depression the participation from male athletes was low. And in the year 1960 the male participation went down because of two reason:

1. No Bob sleigh- In history of winter Olympics it was the first time that bob sleigh was not happening due to some budget constraints (second part of the graph shows histogram for same).

2. Cold war – Many of the countries did not sent their athletes due to the ongoing cold war situation.

And Overall participation from the female athletes seems improving.

I was also interested in understanding the medals won by male and female athletes for their respective countries. Hence Created a Mosaic plot that depicts the different type of Medals Won by Country and Gender (appendix (C2 figure 6)). This a great visualization to understand how countries Male and Female Athletes performed in the Winter Olympics (in terms of getting Gold, Silver and Bronze Medals). In the (appendix (C2 figure 6)) graph we can see that the United States and Canada has the highest number of medals in Winter Olympics.

Sean Sullivan – C3

As part of my role for the final project, I focused primarily on exploring the physical attributes of the Olympic athletes: height and weight. For this role, I created a multitude of visualizations, beginning with boxplots of the distribution of height or weight for the Summer Olympics from 1896 to 2016. From there, I created univariate scatterplots, where the color of the points encoded to the number of athletes who won a medal for that sport, to further visualize the distribution of heights and weights of athletes for the Summer Games. Finally, I narrowed my focus to plotting the median height for the ten countries with the most medals won at the Summer Games and for the ten sports that had the highest range of height amongst its participants. These visualizations are in Section C3 and Appendix 2, C3 Figure 5. Additionally, I assisted with pre-processing data for Visualization C1 and contributed to all group meetings and deadlines.

Over the course of the project, I can firmly state that I have increased my skill level and understanding of visualizations. There are so many ways to display your data and the story that you wish to convey, but it is important to have a thorough understanding of your data and the message that you wish to send before beginning the visualization stage of the project. It would be easy to simply throw my dataset into Tableau and cycle through the "Show Me" suggestions until a visually appealing graph was created. Obviously, this is not ideal, and I took this to heart when working through my own process to understand what exactly it was that I was trying to convey through a graph. With that said, I enjoyed learning how to code the different visualizations I created (many of which were ultimately scrapped) and relished the opportunity to work the class material, mainly the design guidelines, into my final work. Another important learning from this experience was that it's okay to take a more focused approach to the data. I initially was stuck on trying to compare height and weight for many sports and countries, which is difficult to keep from becoming too cluttered or impossible to decode. However, over time, I came to realize that I had the power to shape my story and took the opportunity to narrow the focus of my portion of the analysis.

Chloe Wang—C4

I created the parallel coordinates plot (C4) and the mosaic plot (C4 figure 2). Combined with exploratory graph of "sports that countries won medals" heatmap (C4 figure 1). Three visualizations tell a cohesive story of winter game sport types and countries. Before I decided to focus on sports and countries. I created visualizations in previous project submissions that compare gender distribution across countries and sports, age differences across sports. The basic exploratory graphs only compare at most four variables, however, the graphs become harder to decode the more variables included. Parallel coordinates plot is an optimal choice that allows many variables to be displayed at the same time for visualizing distinctions. However, the plotly package that was used to generate the graph only accepts numerical data types. Due to the volume of datapoints and the types of variables. Preprocessing was taken to reduce numerosity by aggregating rows that belongs to the same country. The preprocessed data not only preserved the same information contained in the original dataset but also generated visualizations that are concise and easy to interpret.

What I learned is that preprocessing and numerosity reduction are important when dealing with high volume of rows. Be flexible in choosing visualizing tools that convey the same story about the dataset more efficiently. It can prolong decoding time when there are too many symbols in the graph. Lastly, more advanced graph provides additional aspects of the data. Mosaic plot and heatmap were generated from the same dataset. Mosaic plot not only emphasizes on sports and countries relationships as encoded by size of the rectangular. The statistical inference of a country not good at playing certain sports is encoded by the colors.

Elizabeth Nieto – C5

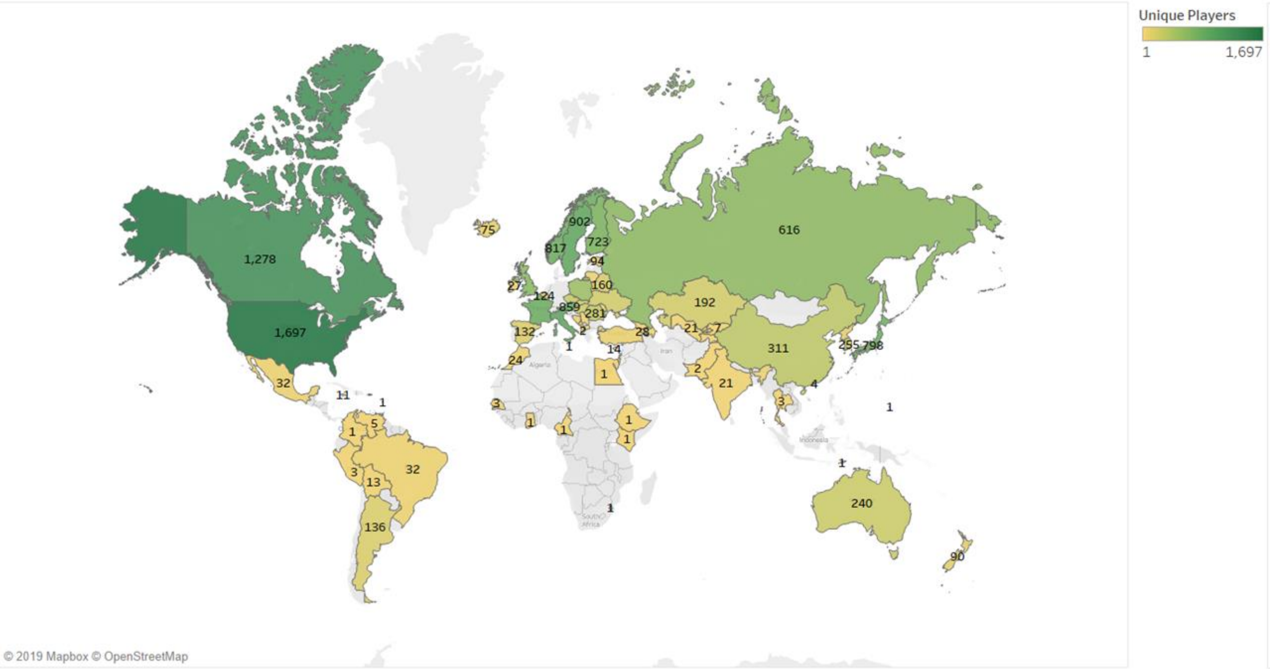
One of my contributions to this team was of liaison. As part of this role I was tasked with meeting with the professor to relay group concerns and questions and communicate the responses back to the group. I was also tasked with submitting deliverables and turned in PD1, PD2, and PD4. Aside from this I also facilitated group formation by asking different members early on if they wanted to join the group. Once the group was formed, I contributed a section of exploratory analysis for PD2 and three exploratory visualizations regarding country participation and medal count. For the final report I contributed to the introduction, and the visualization C5 about inter sport medaling by one athlete.

Prior to this class I had very little exposure to visualizing data using an explanatory analysis where I've had to process the data more extensively than a typical EDA analysis prior to visualization. For this project creating a graph that provided depth to a large dataset was very important to me and when I saw the network graph in class I realized that visualization would be the perfect visual for displaying what two events within a sport were athletes more likely to have been awarded a medal in the same Olympic games(year). What I didn’t realize is unlike an exploratory visual this visual would require the creation of two new datasets, one for the nodes and one for the lines connecting the nodes. To create the nodes, list the data needed to be subset to only include Alpine Skiing events from 1988 and onwards and a unique indexing sequence was added to those events. The dataset for the lines took a few more steps. First was selecting only alpine skiing athletes since 1988 who had won in more than one event for any Olympic Games. Once I had this subset, I then pivoted the data, so each event was a column header, not a value and the medal value were the value for each new event column. Finally, I created two event combinations and count the number of athletes who had won in both at one Olympic game. Using this information, I created a new dataset with three columns: from, to, weight. From and two and the node combinations and weight are the number of instances found. Once I had these two datasets, I could create a network graph. The code for this was relatively short and truly helped me reflect on all the data preprocessing work necessary to produce a network graph. Before when I wanted to create a visualization (typically exploratory visualizations) I relied on r cook book to provide an easy straight forward amendable code which I could easily apply to my data and produce a visualization. These exploratory visualizations typically did not require the data be restructured beyond a filter. Now I realize each visualization requires data be restructured prior to visualization and the restructuring is different for every visualization (and, of course, for every message and audience.)

APPENDIX 2 – Results of formative analysis

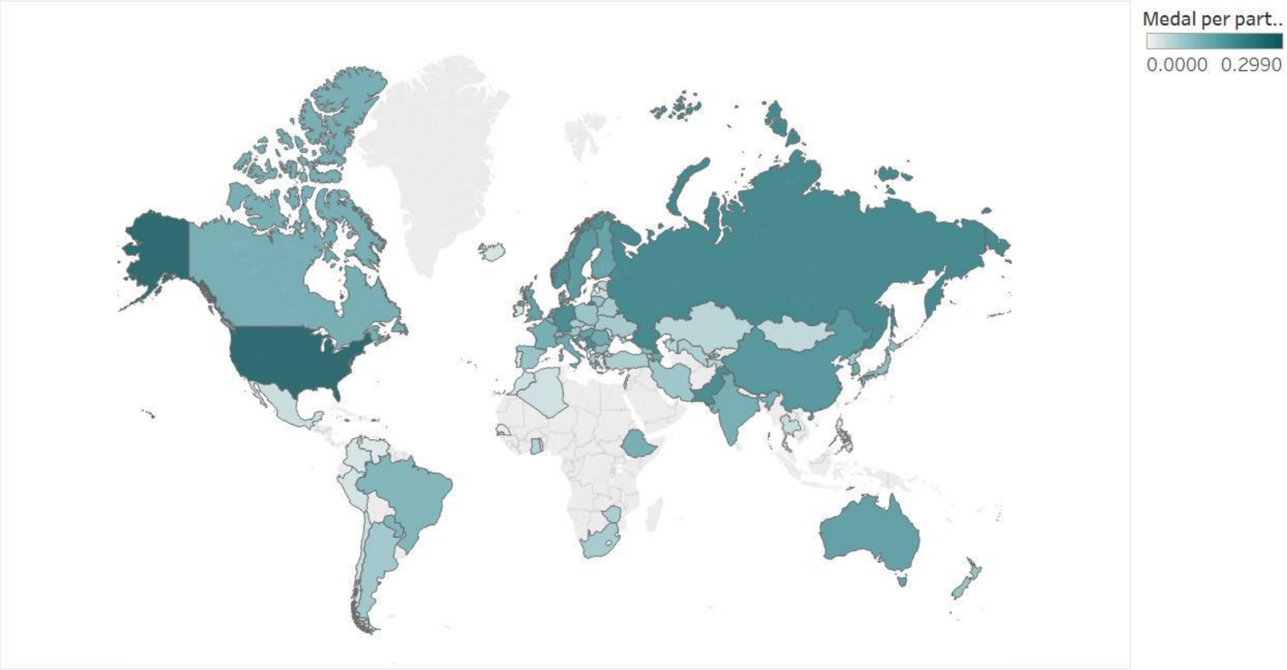
C1 figure 1:

Nation with highest number of participation in olympics

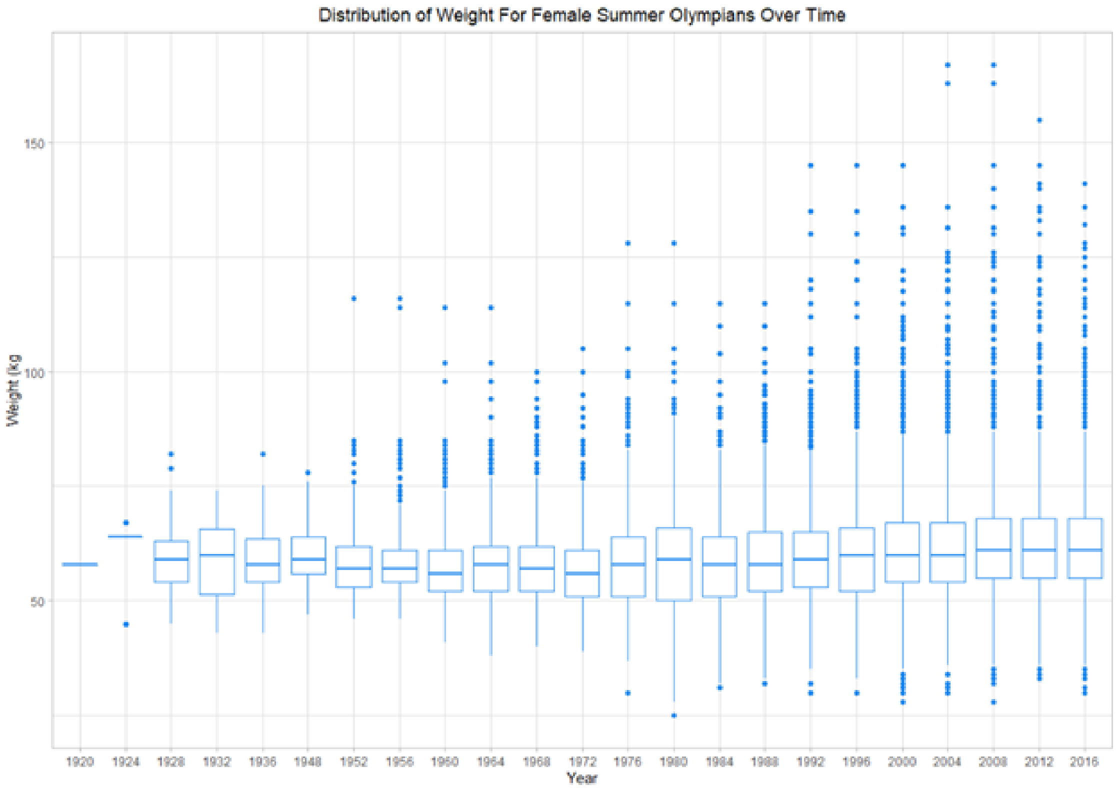


C1 Figure 2: Medal winning rate for each country for Olympic games held only in winter season from 2010-2016

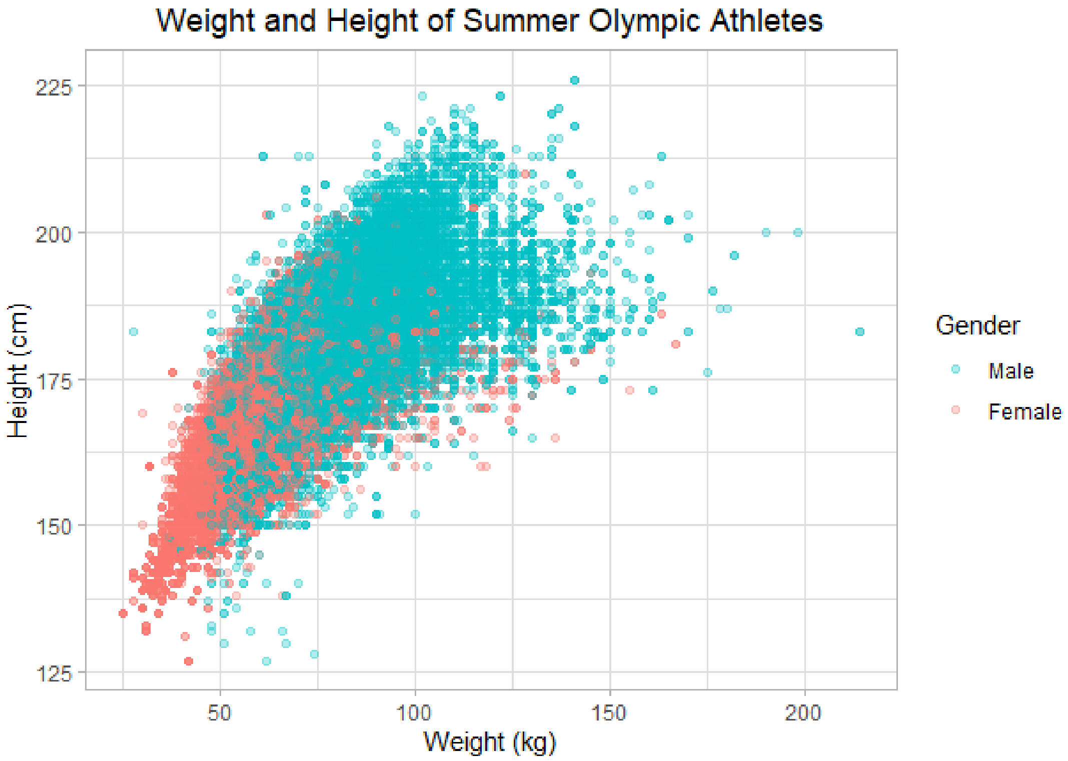
Medal per athlete rate for Olympic Games held in Winter (2010 - 2016)



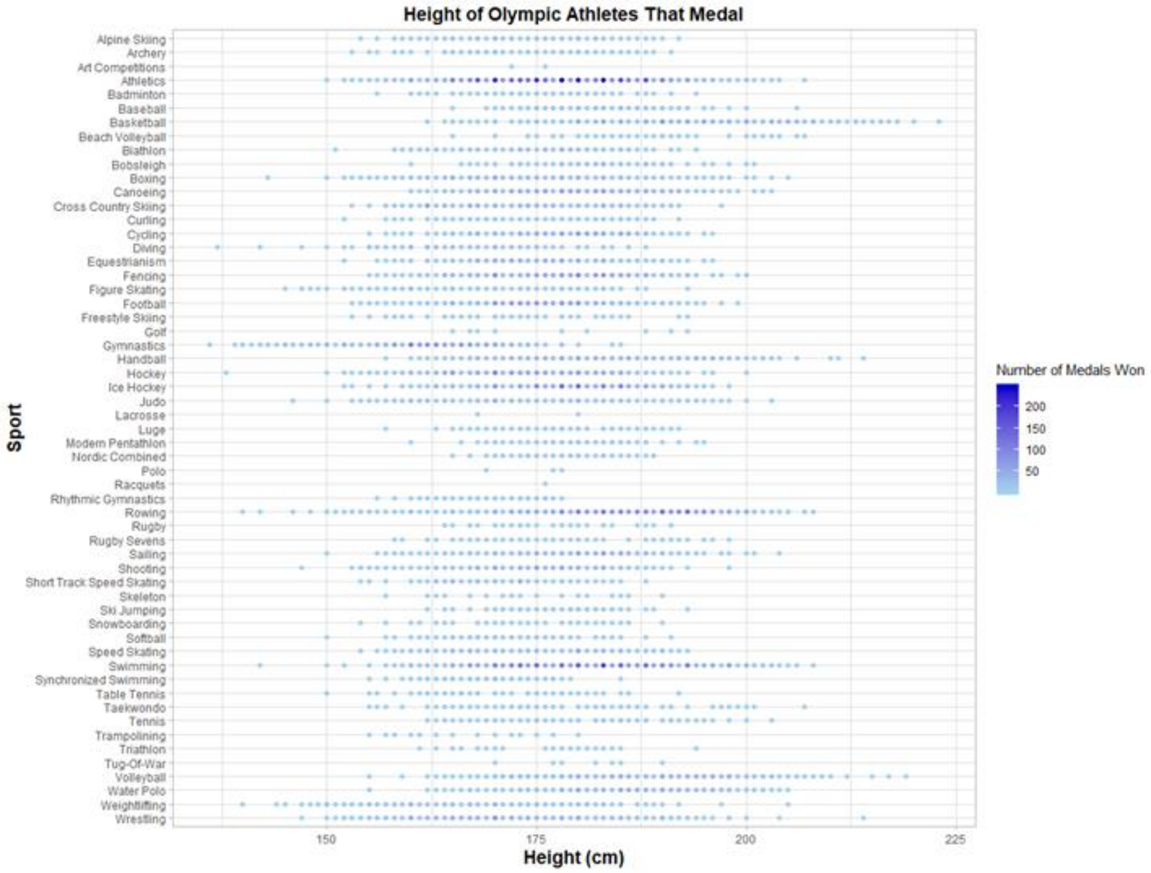
C3 Figure 1: Weight Distribution over Time for Female Summer Olympic Athletes



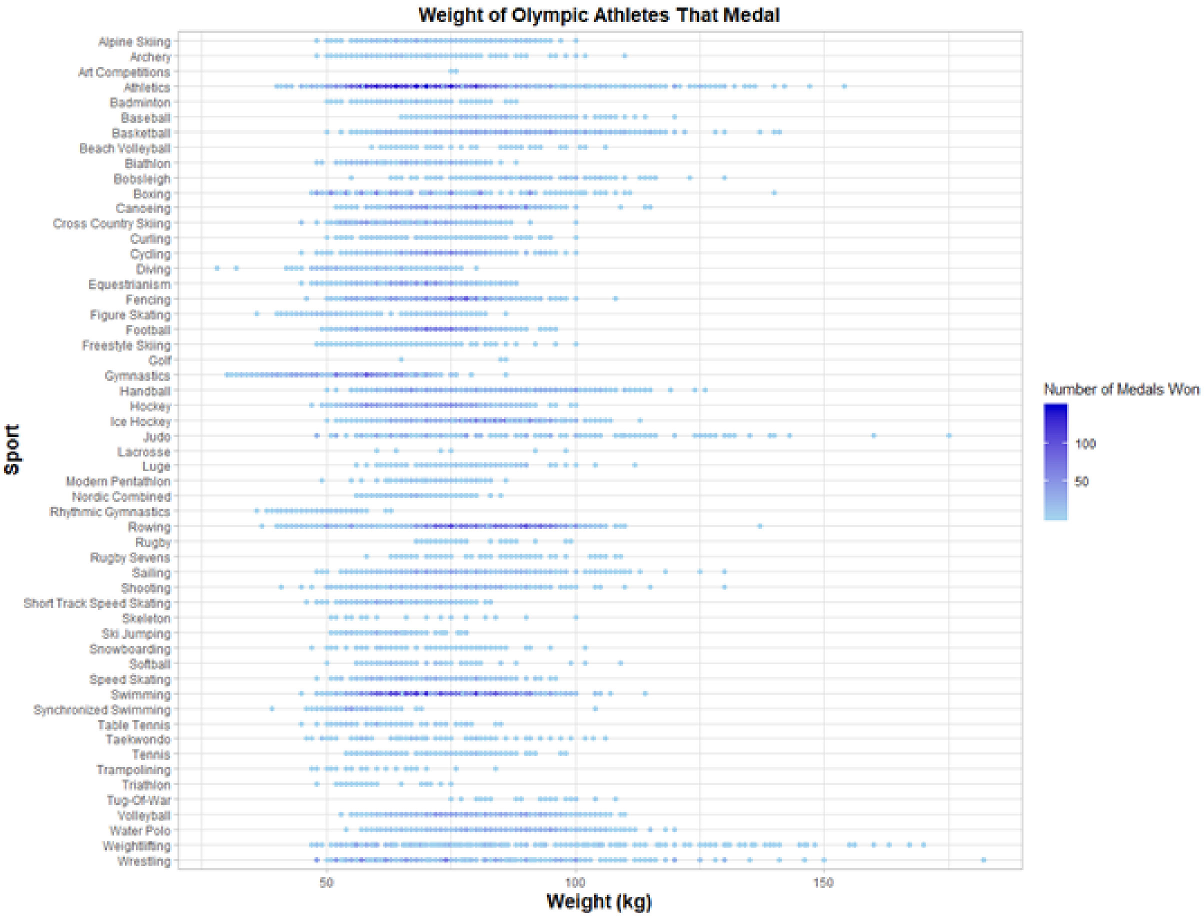
C3 Figure 2: Height and Weight of athletes plotted



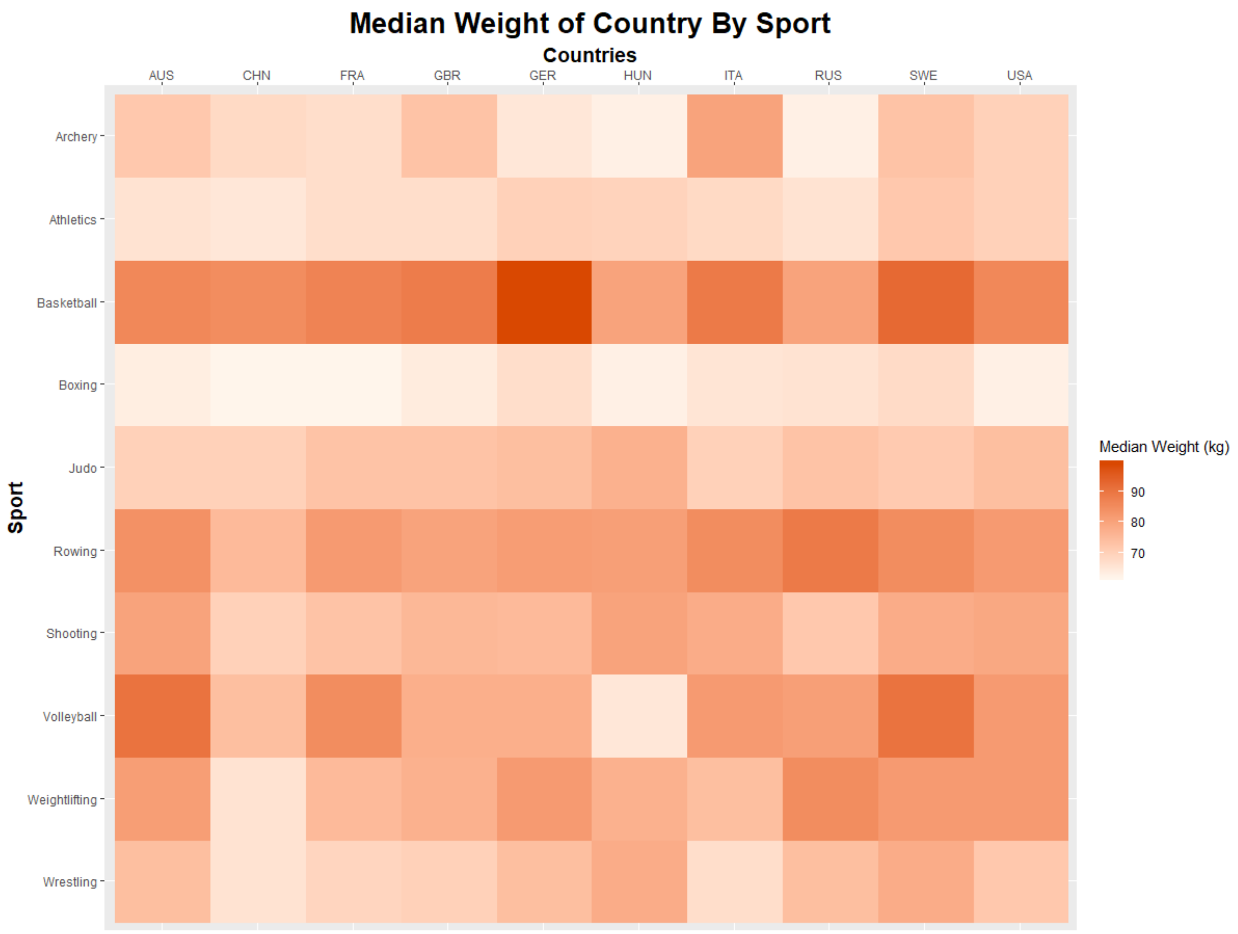
C3 Figure 3



C3 Figure 4


Weight of Olympic Athletes That Medal
Sport
Alpine Skiing
Archery
Art Competitions
Athletics
Badminton
Baseball
Basketball
Beach Volleyball
Biathlon
Bobsleigh
Boxing
Canoeing
Cross Country Skiing
Curling
Cycling
Diving
Equestrianism
Fencing
Figure Skating
Football
Freestyle Skiing
Golf
Gymnastics
Handball
Hockey
Ice Hockey
Judo
Lacrosse
Luge
Modern Pentathlon
Nordic Combined
Rhythmic Gymnastics
Rowing
Rugby
Rugby Sevens
Sailing
Shooting
Short Track Speed Skating
Skeleton
Ski Jumping
Snowboarding
Softball
Speed Skating
Swimming
Synchronized Swimming
Table Tennis
Taekwondo
Tennis
Trampolining
Triathlon
Tug-Of-War
Volleyball
Water Polo
Weightlifting
Wrestling
50
100
150
Weight (kg)
Number of Medals Won
100
50

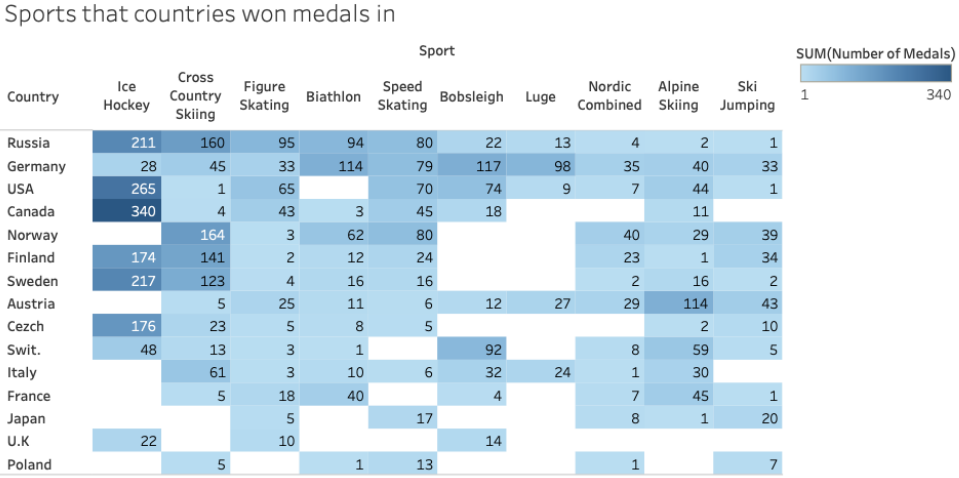
C3 Figure 5



C4 Table 1

Variables	Calculation
Games	Performed by counting distinct “year” for each country
Total	Count of rows for each country
Female %	Count of female/total
Sport	Distinct count of sports
Med. M_age	Median of male ages
Med.F_age	Median of female ages
Medal	number of medals they won/the total number of athletes

C4 Figure 1



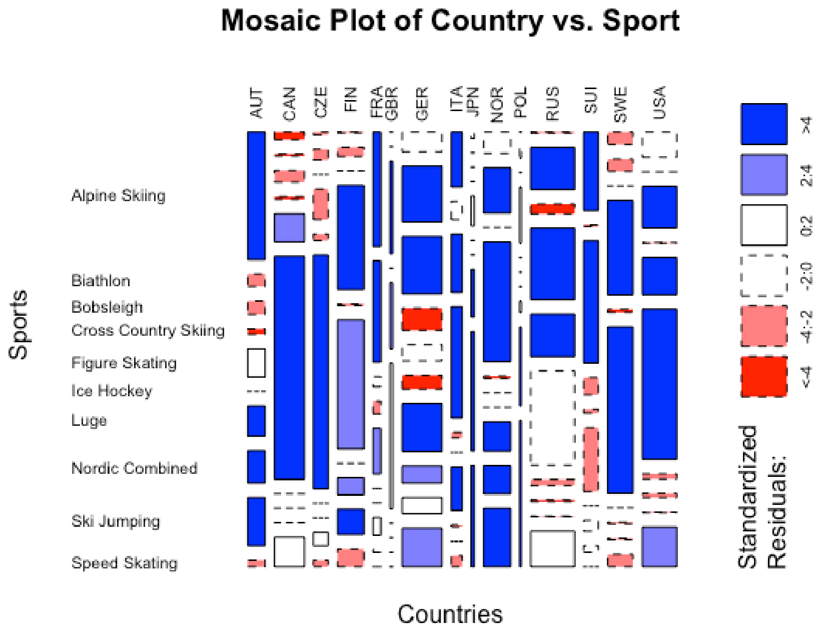
Sports that countries won medals in

SUM(Number of Medals)
1 340

	Sport									
Country	Ice Hockey	Cross Country Skiing	Figure Skating	Biathlon	Speed Skating	Bobsleigh	Luge	Nordic Combined	Alpine Skiing	Ski Jumping
Russia	211	160	95	94	80	22	13	4	2	1
Germany	28	45	33	114	79	117	98	35	40	33
USA	265	1	65		70	74	9	7	44	1
Canada	340	4	43	3	45	18			11	
Norway		164	3	62	80			40	29	39
Finland	174	141	2	12	24			23	1	34
Sweden	217	123	4	16	16			2	16	2
Austria		5	25	11	6	12	27	29	114	43
Cezch	176	23	5	8	5				2	10
Swit.	48	13	3	1		92		8	59	5
Italy		61	3	10	6	32	24	1	30	
France		5	18	40		4		7	45	1
Japan			5		17			8	1	20
U.K	22		10			14				
Poland		5		1	13			1		7

The heatmap shows the competing countries for major winter games sports. The color scheme encodes continuous values of medals that each country has won for the sport types. Rows are sorted by the total number of medals. Preprocessing was performed on the original data points. Observations that have a value as Gold, Silver and Bronze in the medal column were selected. 15 countries were selected because they have sent the greatest number of athletes to the winter games. Countries have different strength across sports types. Canada, the United States, Sweden and Russia are competing countries in Ice Hockey. The competition in cross country skiing is also tight. Russia and Germany are the only two countries that have won medals from all the sports.

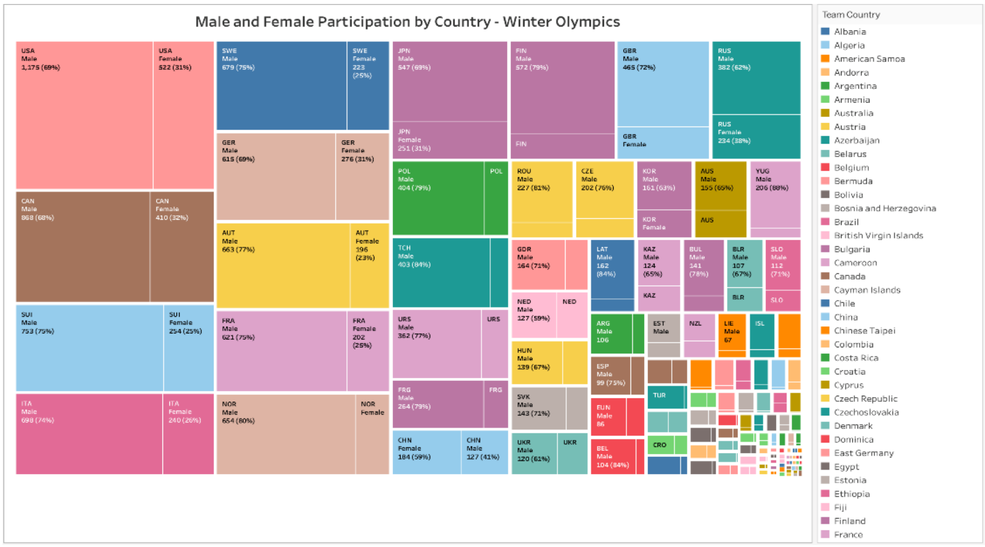
C4 Figure 2



The mosaic plot was generated from the same dataset from above. It shows both positive and negative associations between countries and sports.

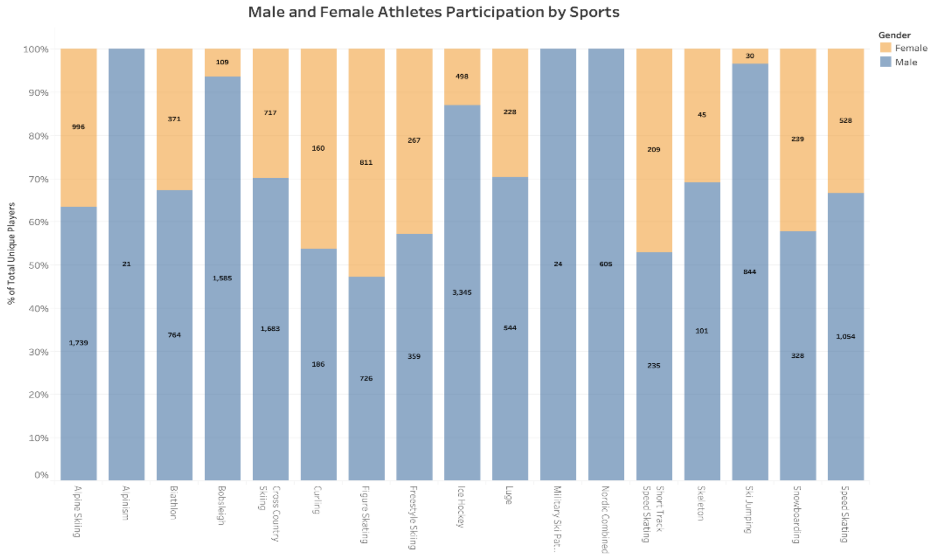
C2 Figure 1

The following graph is a heat map showing the Male and Female Athletes participation by Country for all Winter Olympics held until 2015.



The above graph helped us visualize the participation by various countries in the Winter Olympics as well as the participation of Male and Female Athletes by Country in the Winter Olympics. We see that most of countries that participated in the Winter Olympics have more Male athletes compared to Female athletes.

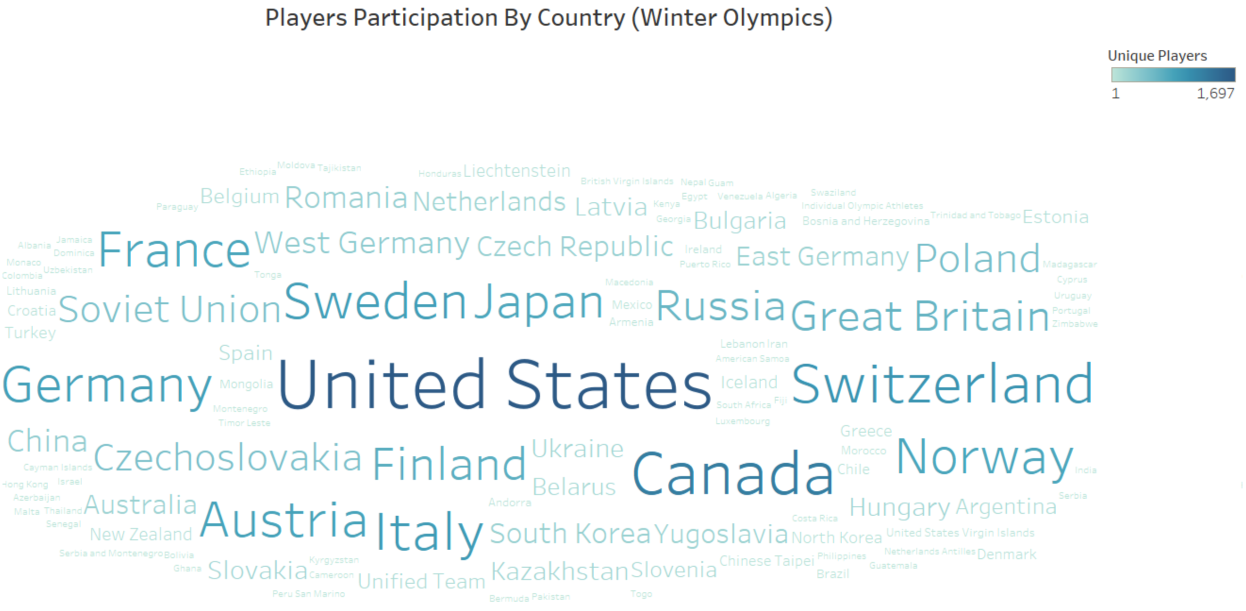
C2 Figure 2



The above graph shows the % of Male Vs Female Athletes Participation by various Sports in the Winter Olympics. Most of the games have more Male Athletes participating compared to Female Athletes. However, few sports like Figure Skating and Short Track Speed Skating have more participation from

Female Athletes. Like the previous analysis it would interesting to see the trend in participation of these sports over time.

C2 Figure 3



Players Participation By Country (Winter Olympics)

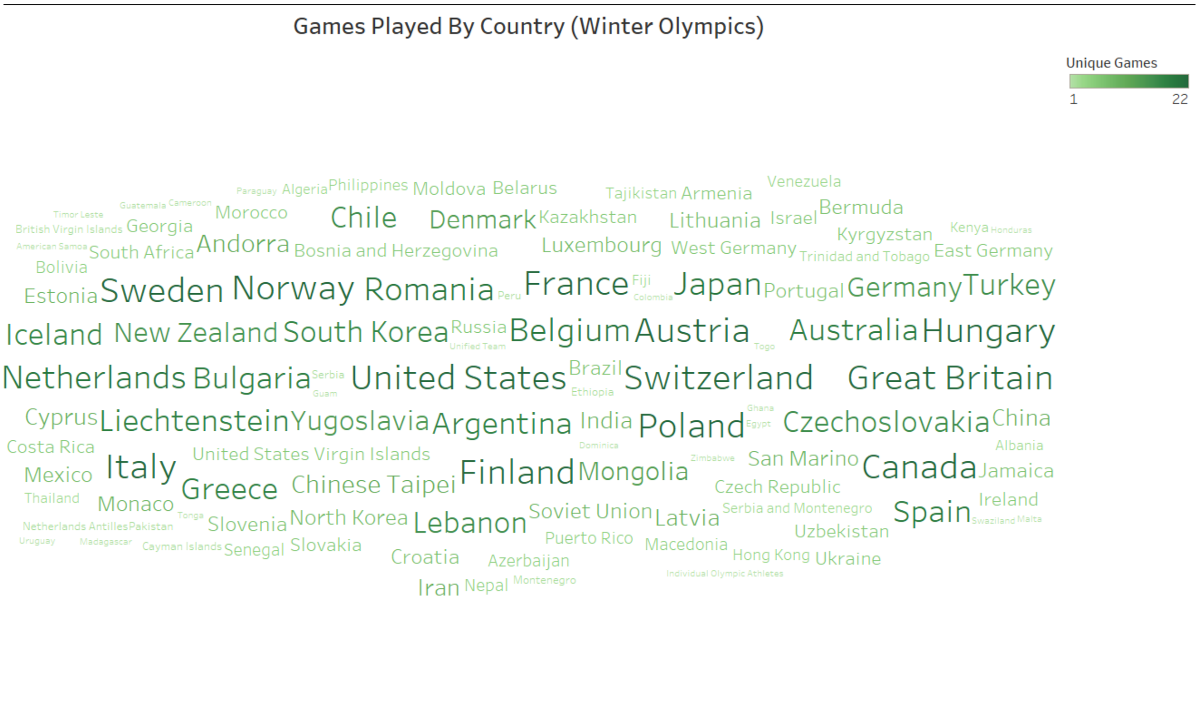
Team Country. Color shows Unique Players. Size shows Unique Players.

The above Word Cloud gives the information related to the Players Participation by Country in the Winter Olympics. The size and the color are used as an indicator for number of unique players participating from a given country in the Winter Olympics.

Steps: -

1. Create Unique Players calculated field as: COUNTD([ID])
2. Drag Team Country field to Text
3. Set the Mark to Text
4. Drag the Unique Players field to Color and Size

C2 Figure 4


Games Played By Country (Winter Olympics)

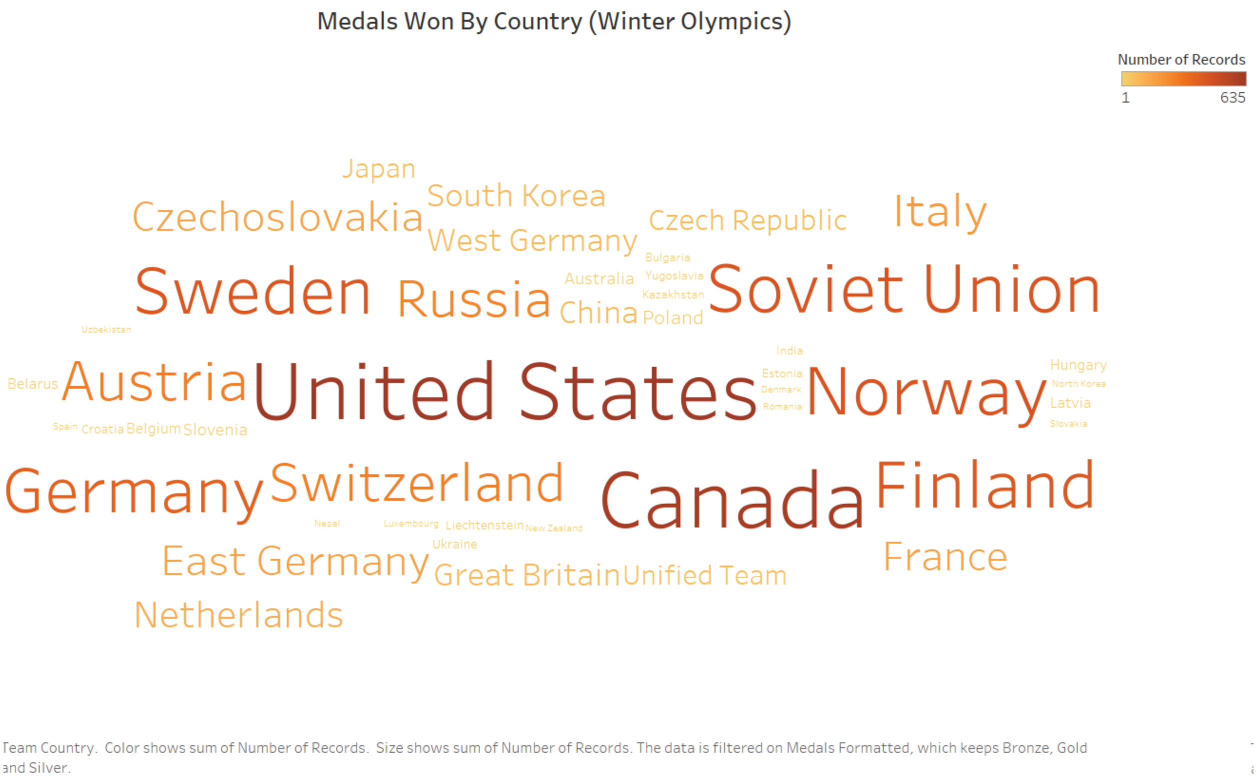
Ƭeam Country. Color shows Unique Games. Size shows Unique Games.

The above Word Cloud gives the information related to the Unique Games by Country in the Winter Olympics. The size and the color are used as an indicator for number of unique games played by a country.

Steps: -

1. Create Unique Games calculated the field as: COUNTD([Games])
2. Drag Team Country field to Text
3. Set the Mark to Text
4. Drag the Unique Games field to Color and Size

C2 Figure 5



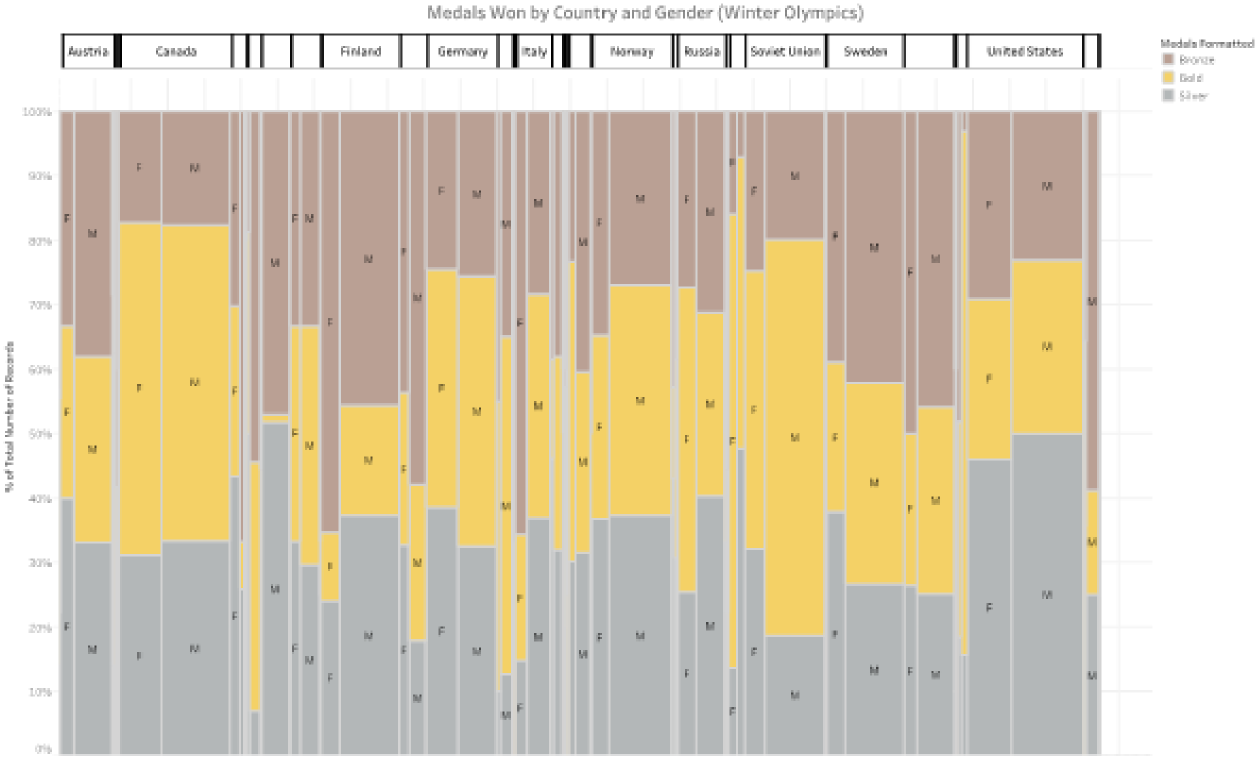
Team Country. Color shows sum of Number of Records. Size shows sum of Number of Records. The data is filtered on Medals Formatted, which keeps Bronze, Gold and Silver.

The above Word Cloud gives the information related to the Medals Won by Country in the Winter Olympics. The size and the color are used as an indicator for number of medals won by a country.

Steps: -

1. Drag Team Country field to Text
2. Set the Mark to Text
3. Drag the Number of Records fields to Color and Size
4. Filter the worksheet based on Medal Formatted; include only Bronze, Silver and Gold.

C2 Figure 6



The figure above is a Mosaic plot that depicts the different type of Medals Won by Country and Gender. This a great visualization to understand how countries Male and Female Athletes performed in the Winter Olympics (in terms of getting Gold, Silver and Bronze Medals). In the above graph we can see that the United States and Canada has highest number of medals in Winter Olympics.

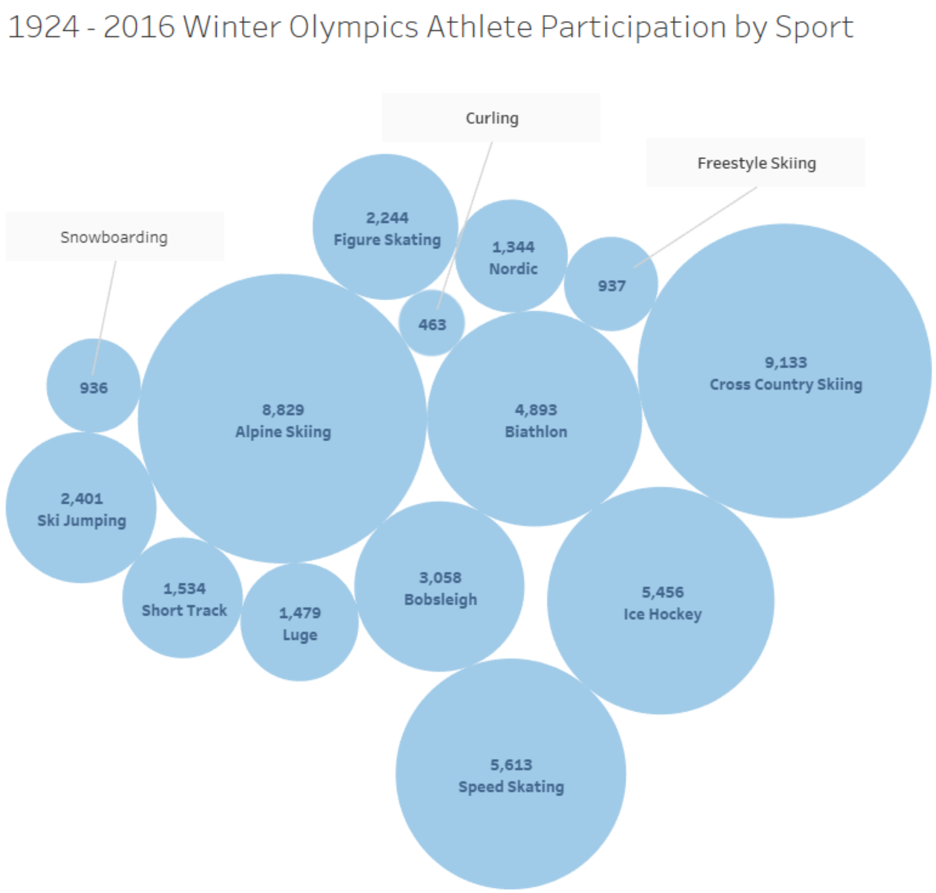
Steps: -

1. Create/Add Following Metrics to the Tableau Visualization
   a. Number of Records: Auto-generated
   b. % of Number of Records by Medals Won by each country and gender
   c. Records per Column: {EXCLUDE [Medals Formatted]: SUM([Number of Records])}
   d. # of Medals:

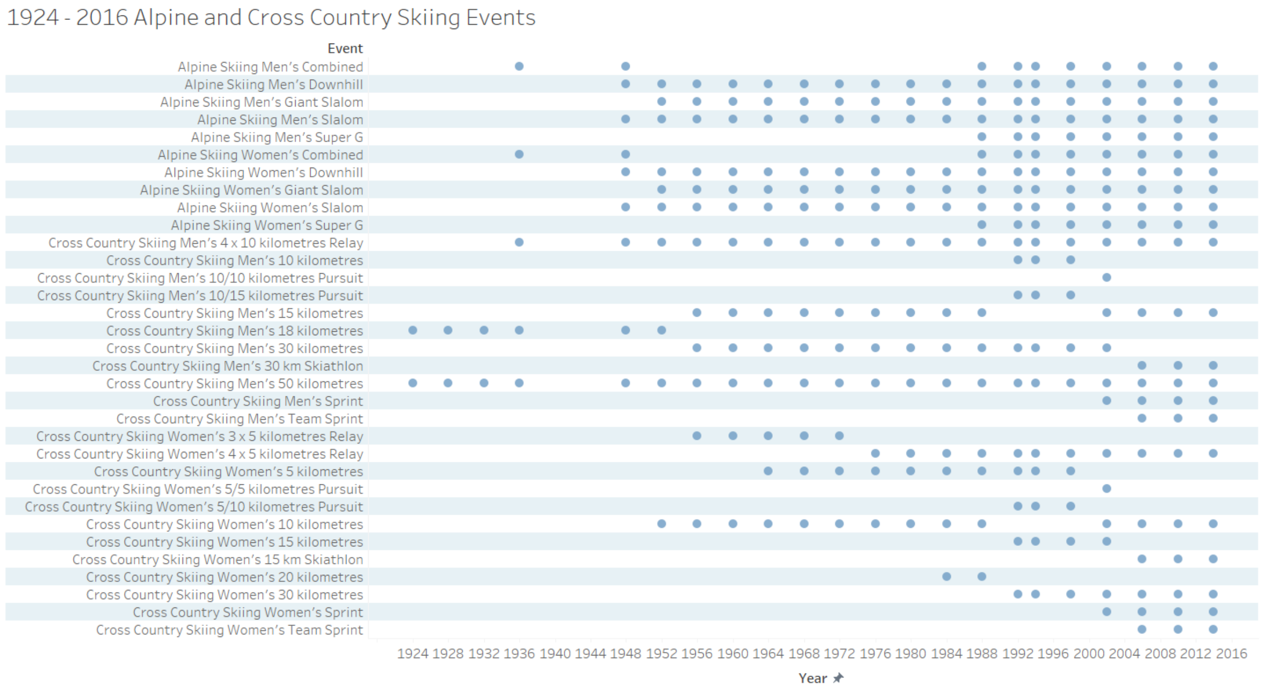
```
IF FIRST()==0 THEN

MIN([Records Per Column])

ELSEIF MIN([Team Country]) != LOOKUP(MIN([Team Country]),-1) THEN

PREVIOUS_VALUE(0) + MIN([Records Per Column])

ELSEIF MIN([Gender]) != LOOKUP(MIN([Gender]),-1) THEN

PREVIOUS_VALUE(0) + MIN([Records Per Column])

ELSE

PREVIOUS_VALUE(0)

END
```

2. Drag % of Number of Records by Medals Won by each country and gender to Rows.
3. Drag # of Medals to Columns
4. Set Marks to Bar
5. Add Medal formatted field to Color
6. Add Records per Column to Size
7. Add Gender Formatted to Label
8. Add Number of Records and Team Country to Detail
9. Create a separate visualization for Mosaic Plot Headers.
10. Combine both the visualization on Dashboard and add actions

C5 Figure 1



C5 Figure 2



APPENDIX 3 – Code

Chaitrali Panchal: C1 – Tableau Workbook submitted

Calculated field code in Python code using pandas:

```
import os
import pandas as pd
import numpy as npos.chdir('E://DePaul Classes//2019 - 2020//1. Fall 2019//DSC 465//Project')df = pd.read_csv('athlete_events.csv')for col in df.columns:
 print(col)df.groupby(['NOC'])['Medal'].value_counts()df['Medal'] = df['Medal'].replace(['Gold','Silver','Bronze'], 'Yes')
df2 = df.groupby(['NOC'])['Medal'].value_counts()
df2 = pd.DataFrame(df2)df3 = df.groupby(['NOC']).size()df3 = pd.DataFrame(df3)df4 = pd.merge(df2,df3, on='NOC', how='outer')df4 = df4.fillna(0)df4.to_csv('mpp.csv')
```

Sarita Patel: C2 – Tableau workbook submitted

Sean Sullivan: C3

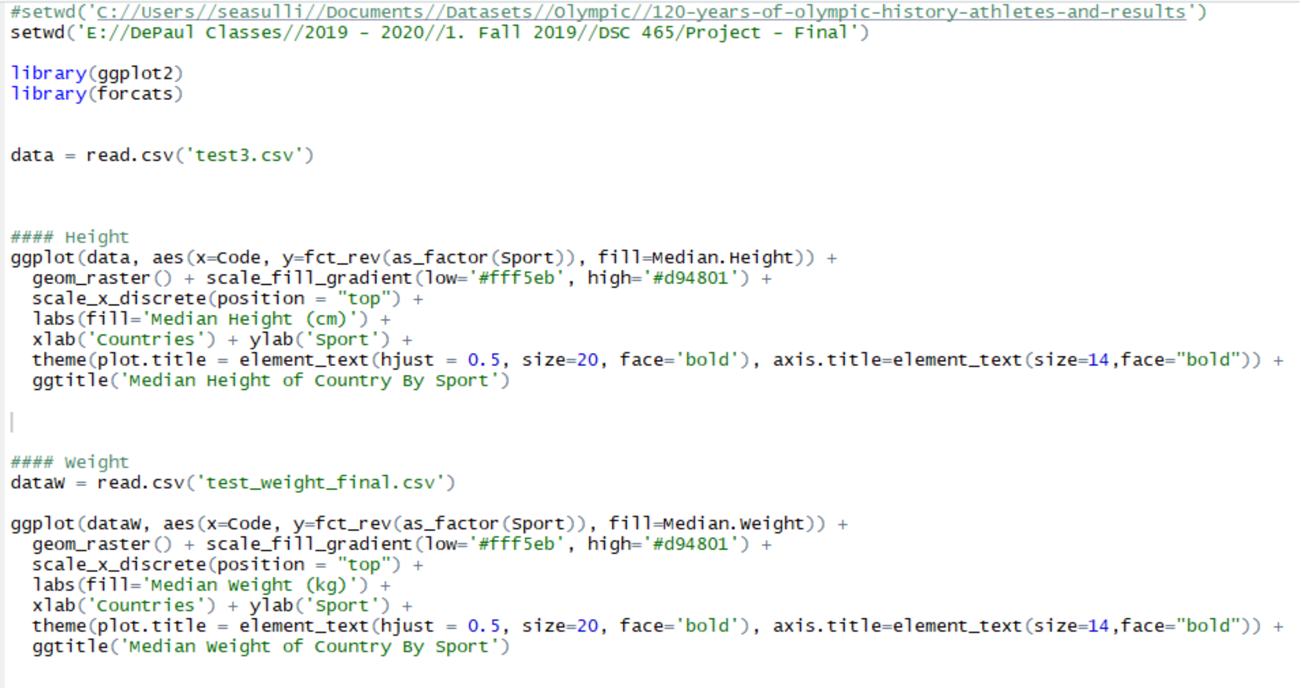
Data Cleaning for pre-processing of data (conducted in Python)

```
import os
import pandas as pd
import numpy as np
os.chdir('C://Users//seasulli//Documents//Datasets//Olympic//120-years-of-olympic-history-athletes-and-results')
```

```
df = pd.read_csv('athlete_events.csv')
df = df[np.isfinite(df['Height'])]


#### Largest Spread in Height per sport
df4 = df.groupby(['Sport'])['Height'].min()
df5 = df.groupby(['Sport'])['Height'].max()

df4.to_csv('sport_h_min.csv')
df5.to_csv('sport_h_max.csv')

#####

 filter = ['Boxing', 'Shooting', 'Swimming', 'Rowing', 'Wrestling', 'Handball',

         'Judo', 'Athletics', 'Volleyball', 'Basketball']

dfH = df.loc[df['Sport'].isin(filter)]
dfH.to_csv('test.csv')


dfHGrouped = dfH.groupby(['Sport','Team'])['Height'].median()
dfHGrouped.to_csv('test2.csv')

## Now get 10 countries

countries = ['United States', 'Russia', 'Great Britain', 'France', 'Germany',

            'Italy', 'China', 'Australia', 'Sweden', 'Hungary']

 df = pd.read_csv('test2.csv')
dfHC = df.loc[df['Country'].isin(countries)]
dfHC.to_csv('test3.csv')


#################### Weight
df = pd.read_csv('athlete_events.csv')
df = df[np.isfinite(df['Weight'])]


# Generate for Delta of Weight and pick top ten
df4 = df.groupby(['Sport'])['Weight'].min()
df5 = df.groupby(['Sport'])['Weight'].max()

 df4.to_csv('sport_w_min.csv')
df5.to_csv('sport_w_max.csv')

 sports = ['Judo', 'Wrestling', 'Athletics', 'Weightlifting', 'Basketball', 'Rowing', 'Shooting',

         'Boxing', 'Volleyball', 'Archery']
```

```
dfW = df.loc[df['Sport'].isin(sports)]
dfWGrouped = dfW.groupby(['Sport','Team', 'NOC'])['Weight'].median()
dfWGrouped.to_csv('test_weight.csv')

countries = ['United States', 'Russia', 'Great Britain', 'France', 'Germany',

        'Italy', 'China', 'Australia', 'Sweden', 'Hungary']


df = pd.read_csv('test_weight.csv')
dfWC = df.loc[df['Country'].isin(countries)]
dfWC.to_csv('test_weight_final.csv')
```

R Code for Core Visualization(s)



```
#setwd('C://Users//seasulli//Documents//Datasets//Olympic//120-years-of-olympic-history-athletes-and-results')
setwd('E://DePaul Classes//2019 - 2020//1. Fall 2019//DSC 465/Project - Final')

library(ggplot2)
library(forcats)


data = read.csv('test3.csv')



#### Height
ggplot(data, aes(x=Code, y=fct_rev(as_factor(Sport)), fill=Median.Height)) +
  geom_raster() + scale_fill_gradient(low='#fff5eb', high='#d94801') +
  scale_x_discrete(position = "top") +
  labs(fill='Median Height (cm)') +
  xlab('Countries') + ylab('Sport') +
  theme(plot.title = element_text(hjust = 0.5, size=20, face='bold'), axis.title=element_text(size=14,face="bold")) +
  ggtitle('Median Height of Country By Sport')



#### Weight
dataW = read.csv('test_weight_final.csv')

ggplot(dataW, aes(x=Code, y=fct_rev(as_factor(Sport)), fill=Median.Weight)) +
  geom_raster() + scale_fill_gradient(low='#fff5eb', high='#d94801') +
  scale_x_discrete(position = "top") +
  labs(fill='Median Weight (kg)') +
  xlab('Countries') + ylab('Sport') +
  theme(plot.title = element_text(hjust = 0.5, size=20, face='bold'), axis.title=element_text(size=14,face="bold")) +
  ggtitle('Median Weight of Country By Sport')
```

Chloe Wang: C4

```
library(plotly)
df <- read.csv("imp.csv")
p <- df %>%
  plot_ly(width =800, height = 600) %>%
  add_trace(type = 'parcoords',
            line = list(color = ~color,
                        colorscale =list(c(0, '#6C9E12'), ## social services
                                         c(0.05,'#79706E'),
                                         c(0.1,'#4E79A7'),
                                         c(0.15,'#D4A6CB'),
                                         c(0.2,'#8CD17D'),
                                         c(0.25,'#FFBE7D'),
                                         c(0.3,'#E15759'),
                                         c(0.35,'#499894'),
                                         c(0.4,'#FABFD2'),
                                         c(0.45,'#BAB0AC'),
                                         c(0.5,'#A0C8EB'),
                                         c(0.55,'#9D7660'),
                                         c(0.6,'#B6992D'),
                                         c(0.65,'#59A14F'),
                                         c(0.7,'#FF9D9A'),
                                         c(0.75,'#86BCB6'),
                                         c(0.8,'#B07AA1'),
                                         c(0.85,'#D37295'),
                                         c(0.9,'#F28E2B'),
                                         c(0.95,'#D7B5A6'),
                                         c(1,'#F1CE63')),
                        showscale =FALSE,
                        reversescale = FALSE,
                        cmin =0,
                        cmax =20),
            dimensions = list(
              list(range = c(5,25),
                   label = 'Games', values = ~Games),
              list(range = c(600,3500),
                   label = 'Total', values = ~Total),
              list(range = c(0,0.8),
                   label = 'Female %', values = ~Female.pct),
              list(range = c(0,13),
                   label = 'Sport', values = ~Sport),
              list(range = c(20,30),
                   label = 'Med.M_age', values = ~Male.age),
              list(range = c(20,30),
                   label = 'Med.F_age', values = ~Female.age),
              list(range = c(0,0.35),
                   label = 'Medal%', values = ~win)
            )
  )
p
```

Mosaic Plot

```
df <- read.csv("ds.csv")
catable=table(df$NOC, df$Sport)
catable
margin.table(catable, 1)
margin.table(catable, 2)
round(prop.table(catable),4) # cell percentages
round(prop.table(catable, 1),4) # row percentages
round(prop.table(catable, 2),4) # column percentages
#correspondance fit
fitca <- ca(catable)
mosaicplot(catable,main='Mosaic Plot of Country vs.
Sport',las=2,off=30,xlab='Countries',ylab='Sports',shade=TRUE)
```

Elizabeth Nieto: C5

~~~~
####Steps to create a Network Graph
1. Set up environment
2. Get data for node and edge tibles
3. Create tidygraph
4. Create Network Graph


####Environment Setup
```{r}
#import libraries
library(tidygraph)
library(ggraph)
library(tidyr)
library(plyr)
library (ggrepel)

#set working directory
setwd("C:\\Users\\eniet\\Documents\\School\\Autumn19\\DSC465_Data_Visualization\\Project")
getwd()
```

####Get data for node and edge tibbles
```{r}

data <-read.csv("olympics.csv")
head(data,15)

#filter out athletes w/o medals
winners <- data[!is.na(data$Medal),]
#filter out all sports except Alpine Skiing
alpine_winners <- filter (winners, Sport == 'Alpine Skiing')
#filterout any case that was played before 1988 bc this is when all events started
alpine_winners_1988_2016 <- filter(alpine_winners, Year > 1987)
#select data set with only ID and events
id_events <- select(alpine_winners_1988_2016 ,ID,Year,Medal,Event)
#241 Rows
#this data sets has one row containing id and event per athlete who won in alpine skiing
id_events
#filter out athletes who only won one medal
athleteIDs <- select(id_events, ID)
ath_freq <- athleteIDs %>%
  count('ID')
multiple_wins <- filter(ath_freq, freq != '1')
#filter to only get unique events per id
~~~~

~~~~
#id_events %>%
#  distinct()

#join athletes w/ multiple wins to all athletes
multiplewins <- join(multiple_wins,id_events, by = 'ID', type = 'inner')
#pivot dataset to get each event to become a column containing a medal value
piv_event <- multiplewins %>%
  spread(Event, Medal)

#convert event values to numeric ids of alpine tibble evetn ids
renamed_events <- rename(piv_event, c("Alpine Skiing Men\'s Downhill" = '1',
                  "Alpine Skiing Men's Super G" = '2',
                  "Alpine Skiing Men's Giant Slalom" = '3',
                  "Alpine Skiing Men's Slalom" = '4',
                  "Alpine Skiing Men's Combined" = '5',

                  "Alpine Skiing Women's Downhill" = '6',
                  "Alpine Skiing Women's Super G" = '7',
                  "Alpine Skiing Women's Giant Slalom" = '8',
                  "Alpine Skiing Women's Slalom" = '9',
                  "Alpine Skiing Women's Combined" = '10'
                  ))
 ```
####GET- gender totals for athletes winning two events

```{r}
#MENS EVENTS
#athletes who won in sport 1 and 2
events_1_2 <-filter(renamed_events,!((is.na(`1`)| is.na(`2`))))
line_1_2 <- nrow(events_1_2)

#athletes who won in sport 1 and 3
events_1_3 <- filter(renamed_events,!((is.na(`1`)| is.na(`3`))))
line_1_3 <- nrow(events_1_3)

#athletes who won in sport 1 and 4
events_1_4 <- filter(renamed_events,!((is.na(`1`)| is.na(`4`))))
line_1_4 <- nrow(events_1_4)

#athletes who won in sport 1 and 5
events_1_5 <- filter(renamed_events,!((is.na(`1`)| is.na(`5`))))
line_1_5 <- nrow(events_1_5)

#athletes who won in sport 2 and 3
events_2_3 <- filter(renamed_events, !((is.na(`2`)| is.na(`3`))))
~~~~

```
line_2_3 <- nrow(events_2_3)

#athletes who won in sport 2 and 4
events_2_4 <- filter(renamed_events,!((is.na(`2`)| is.na(`4`))))
line_2_4 <- nrow(events_2_4)

#athletes who won in sport 2 and 5
events_2_5 <- filter(renamed_events,!((is.na(`2`)| is.na(`5`))))
line_2_5 <- nrow(events_2_5)

#athletes who won in sport 3 and 4
events_3_4 <- filter(renamed_events,!((is.na(`3`)| is.na(`4`))))
line_3_4 <- nrow(events_3_4)

#athletes who won in sport 3 and 5
events_3_5 <- filter(renamed_events,!((is.na(`3`)| is.na(`5`))))
line_3_5 <- nrow(events_3_5)

#athletes who won in sport 4 and 5
events_4_5 <- filter(renamed_events,!((is.na(`4`)| is.na(`5`))))
line_4_5 <- nrow(events_4_5)

#WOMENS EVENTS
#athletes who won in sport 6 and 7
events_6_7 <-filter(renamed_events,!((is.na(`6`)| is.na(`7`))))
line_6_7 <- nrow(events_6_7)

#athletes who won in sport 6 and 8
events_6_8 <- filter(renamed_events,!((is.na(`6`)| is.na(`8`))))
line_6_8 <- nrow(events_6_8)

#athletes who won in sport 6 and 9
events_6_9 <- filter(renamed_events,!((is.na(`6`)| is.na(`9`))))
line_6_9 <- nrow(events_6_9)

#athletes who won in sport 6 and 10
events_6_10 <- filter(renamed_events,!((is.na(`6`)| is.na(`10`))))
line_6_10 <- nrow(events_6_10)

#athletes who won in sport 7 and 8
events_7_8 <- filter(renamed_events,!((is.na(`7`)| is.na(`8`))))
line_7_8 <- nrow(events_7_8)

#athletes who won in sport 7 and 9
events_7_9 <- filter(renamed_events,!((is.na(`7`)| is.na(`9`))))
```

~~~
line_7_9 <- nrow(events_7_9)

#athletes who won in sport 7 and 10
events_7_10 <- filter(renamed_events,!((is.na(`7`)| is.na(`10`))))
line_7_10 <- nrow(events_7_10)

#athletes who won in sport 8 and 9
events_8_9 <- filter(renamed_events,!((is.na(`8`)| is.na(`9`))))
line_8_9 <-nrow(events_8_9)

#athletes who won in sport 8 and 10
events_8_10 <- filter(renamed_events,!((is.na(`8`)| is.na(`10`))))
line_8_10 <- nrow(events_8_10)

#athletes who won in sport 9 and 10
events_9_10 <- filter(renamed_events,!((is.na(`9`)| is.na(`10`))))
line_9_10<- nrow(events_9_10)
```
#CREATE new data frame for edges(event counts) using results from above
```{r}
from <- c(1,1,1,2,2,3,3,4,1,1,1,1,2,2,2,3,3,4)
to <- c(2,3,5,3,5,4,5,5,2,3,4,5,3,4,5,4,5,5)
weight <- c(7,4,4,7,3,3,2,3,4,3,1,5,4,1,4,5,5,7)
gender <- c('M','M','M','M','M','M','M','M', 'F', 'F', 'F', 'F', 'F', 'F', 'F', 'F', 'F', 'F')

ec_df2 <- data.frame(from, to, weight, gender)
names(ec_df2) <- c('from', 'to', 'weight','gender')

```
####CREATE edge tible
```{r}
ec_df2_tibble <- as_tibble(ec_df2)

ec_df2_tibble

```
####CREATE node dataset
#####Get unique events in Alpine Skiing
```{r}
events <- select(data, Event)
alpine <- filter (data, Sport == 'Alpine Skiing')
alpineevents <- unique(alpine$Event)
alpineevents
```
####Create new dataset using unique events for both men and women
~~~

~~~
```{r}
fiveevents <- c('Downhill', 'Super G', 'Giant Slalom', 'Slalom','Combined' )
fiveevent_ids <- c(1,2,3,4,5)

alpineevents <- data.frame(fiveevent_ids,fiveevents)
names(alpineevents) <- c('ID', 'event')
```
####CREATE Node tibble
```{r}
alpineevents_tibble <- as_tibble(alpineevents)

alpineevents_tibble

``
####CREATE tidygraph
```{r}
events_tidy <- tbl_graph(nodes = alpineevents_tibble, edges = ec_df2_tibble, directed = TRUE)

```
####CREATE NETWORK GRAPH
```{r}
p <-ggraph(events_tidy, layout = "drl") +

  geom_edge_link0(aes(width = weight), colour = 'dodgerblue', alpha = .5, show.legend = TRUE) +

  scale_edge_width(range =c(0.2,4.2)) +

  geom_node_point(shape = 21) +
  geom_node_text(aes(label = event), repel = TRUE)+

  labs(edge_width = 'Wins in Both') +

  theme_graph(background = 'white') +

  ggtitle('Alpine Skiing Medals Awarded per Two Events Between 1988 - 2016') +
  theme(
    legend.position = "bottom",

  )

p + facet_edges(~gender)

```
~~~