Exploring 120 Years of Olympics

Group: The Five Olympians
By: Alex Rosenblum, Cassandra Steffey, Jaimi Patel, Matt Morais, Samuel Prasad Chinta

I. Introduction

Our dataset is about Olympic athletes starting from the 1896 Summer Games up through the 2016 Summer Games. The dataset is sourced from Kaggle, where the uploader reports that they scraped the data from sports-reference.com. There are 271,116 rows, each representing an athlete's participation in a single Olympic event. The variables are:

ID - Unique number for each athlete

Name - Athlete's name

Sex - M or F

Age - Integer

Height - In centimeters

Weight - In kilograms

Team - Team name

Medal - Gold, Silver, Bronze, or NA

NOC - National Olympic Committee 3-letter code

Games - Year and season

Year - Integer

Season - Summer or Winter

City - Host city

Sport - Sport

Event - Event

II. Exploratory Analysis

The team began by reviewing multiple distributions and visualizations prepared by each group member. From that discussion, there were two common themes that were of collective interest whilst being broad enough to garner real insights from the selected data set. We began by looking at the structure of the olympics in its simplest form; number of events per season per year. This can be seen in figure 1 below.

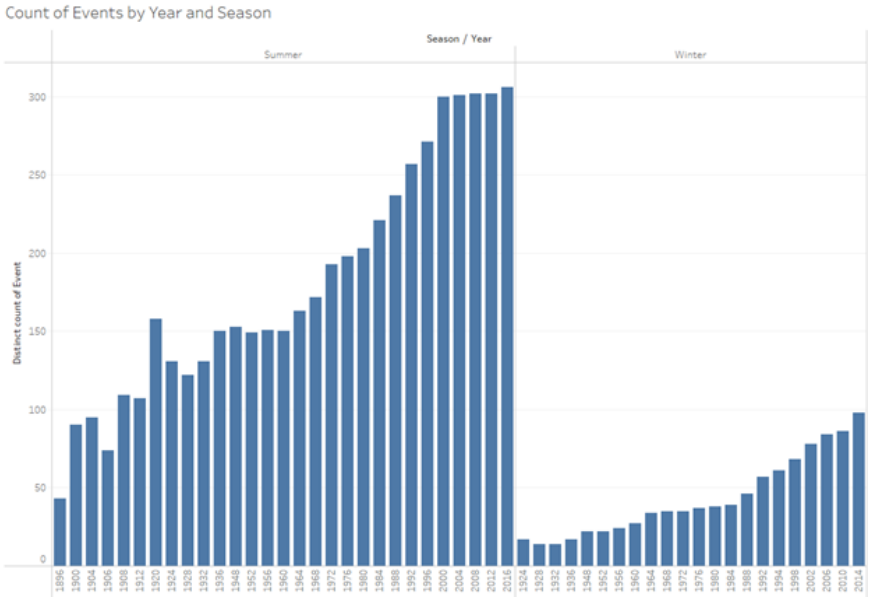


Fig 1: The count of events by year and season.

From this, the group could see that the summer and winter games were very different and wanted to learn more about why there are so many summer games in comparison to the winter games, and why the structure of the games is getting larger over time. The increase in size of the olympics followed the same relative trend displayed in another visualization that looked closer at gender; separating female participation in the games. It appears from the cursory analysis, that the increase in games and change in structure could be related to the addition of more female events as cultural norms shift in favor of gender equity.

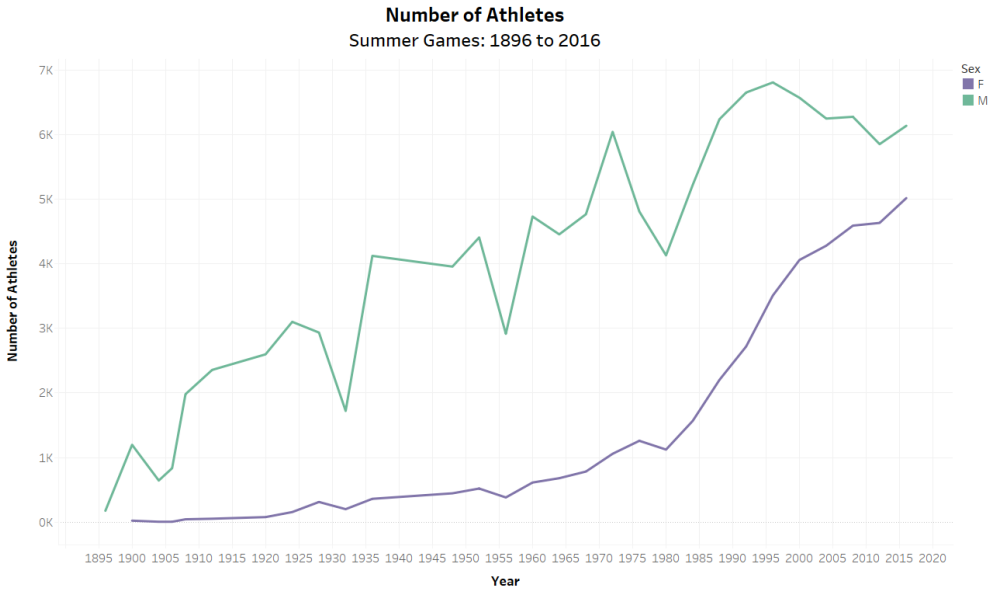


Fig 2: The number of athletes competing in the summer games separated by gender.

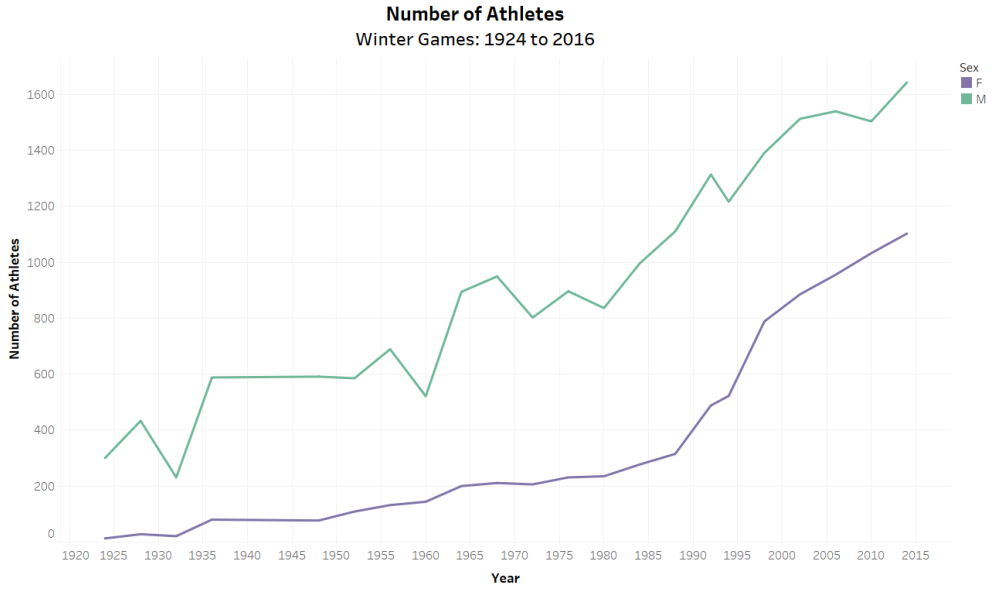


Fig 3: The number of athletes competing in the winter games separated by gender.

These two points of the differences between summer and winter olympic structure, along with how gender impacts the organization of the olympic games, brings us to our first direction for analysis; the olympic game Structure and Organization.

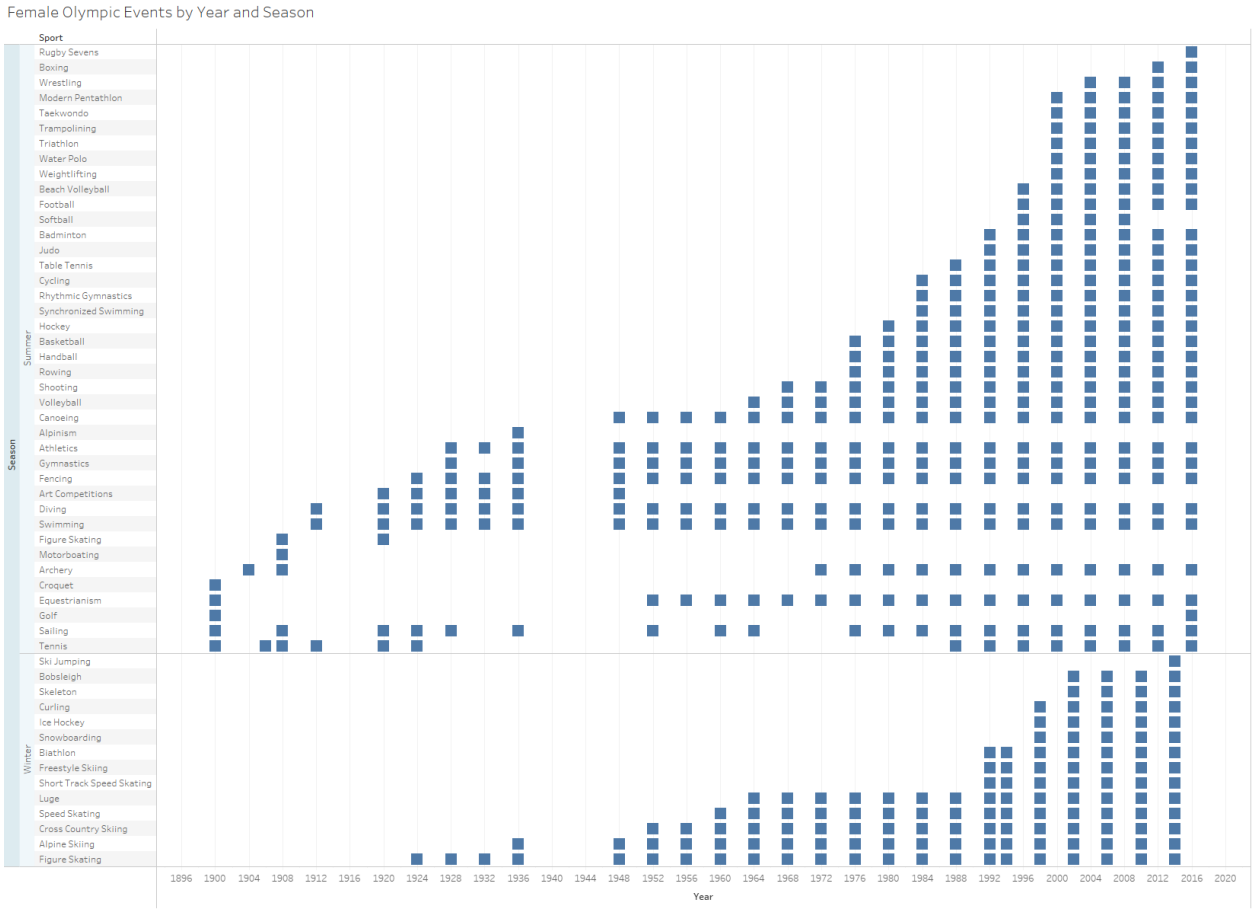


Fig 4: The Olympic sports by the years they were introduced to female athletes.

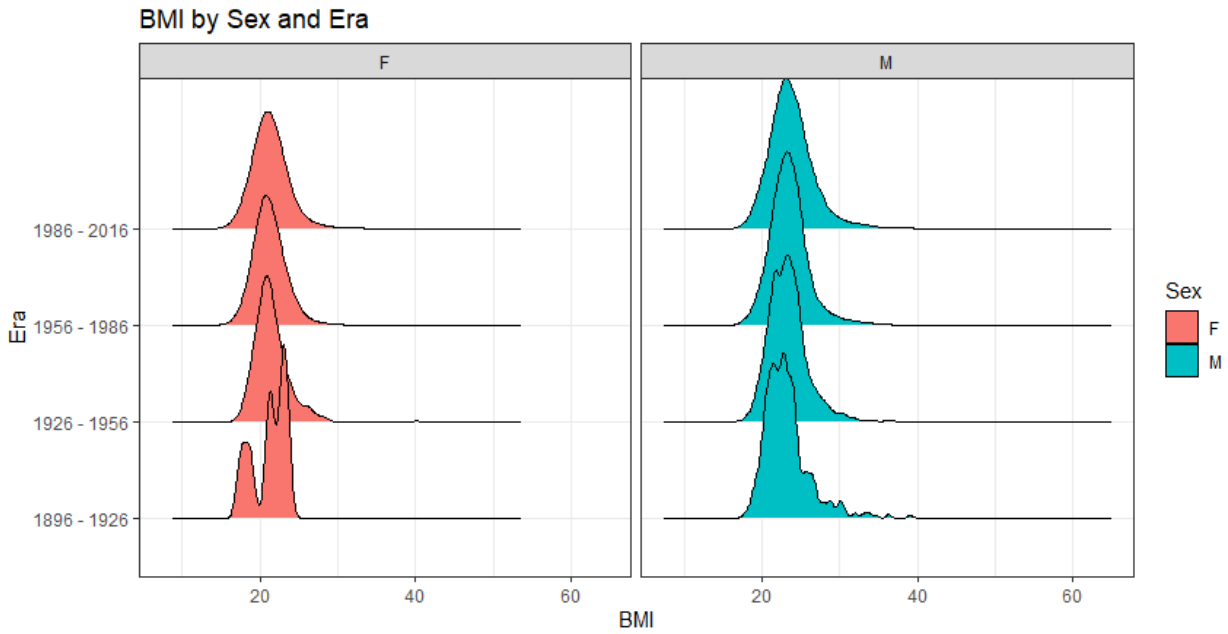


Fig. 5: Body Mass Index by Gender and Era

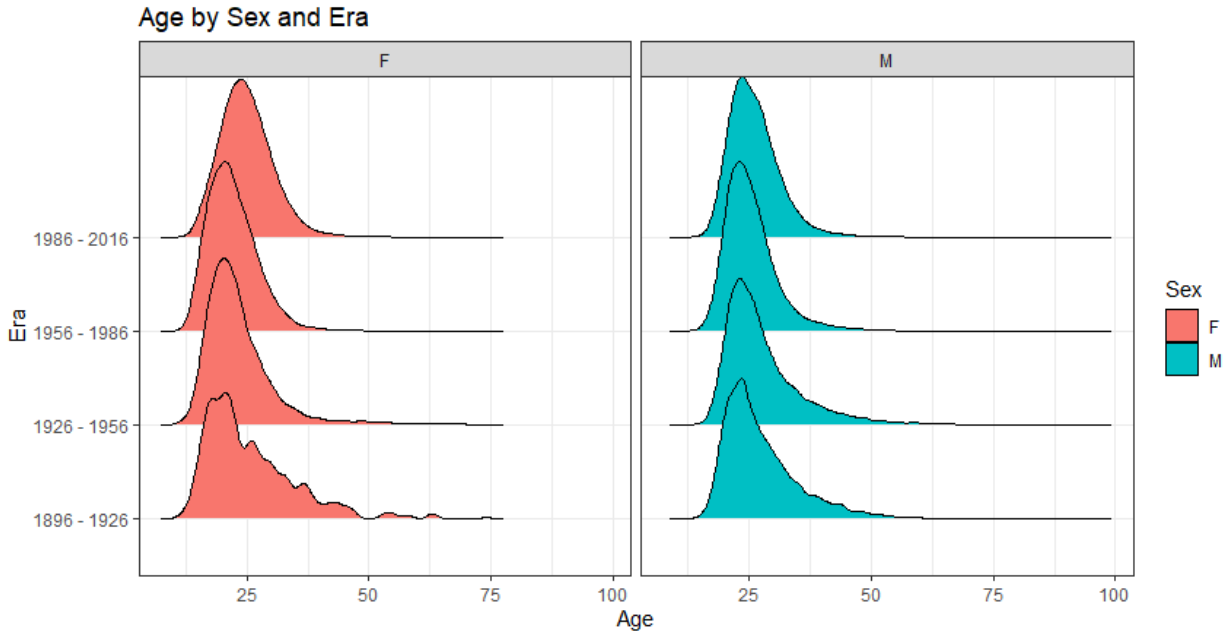


Fig 6: Age by Sex and Era

In our exploration into demographic changes over the years, we discovered that there wasn't much to learn by comparing Age and Body Mass Index (BMI) of Men and Women over the years. This brought to light the fact that our approach to this line of investigation may be overly specific, and a broader, less granular approach may be required. Distribution of Age over time will be discussed in a later visualization.

III. Visualizations

For the structure and organization, we have two final visualizations. The first is about the trends of the sports colored based on gender and separated by summer and winter seasons.

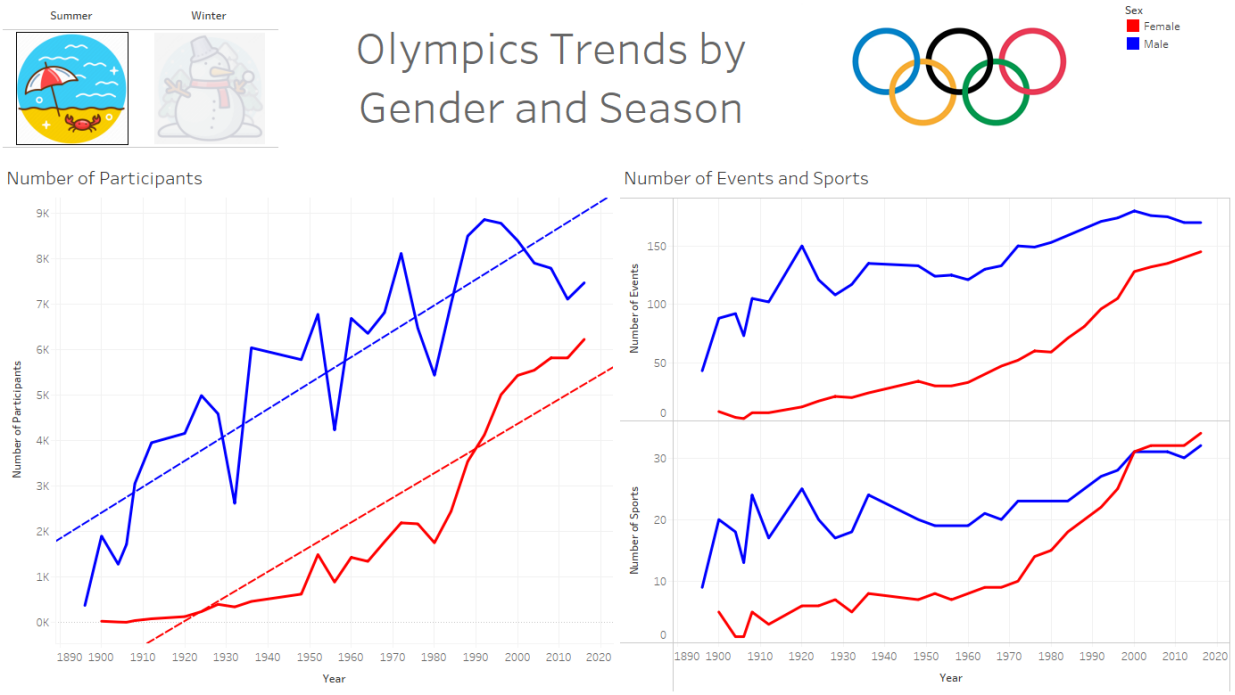


Fig 7: A dashboard showing different trends in the structure of the Summer Olympics

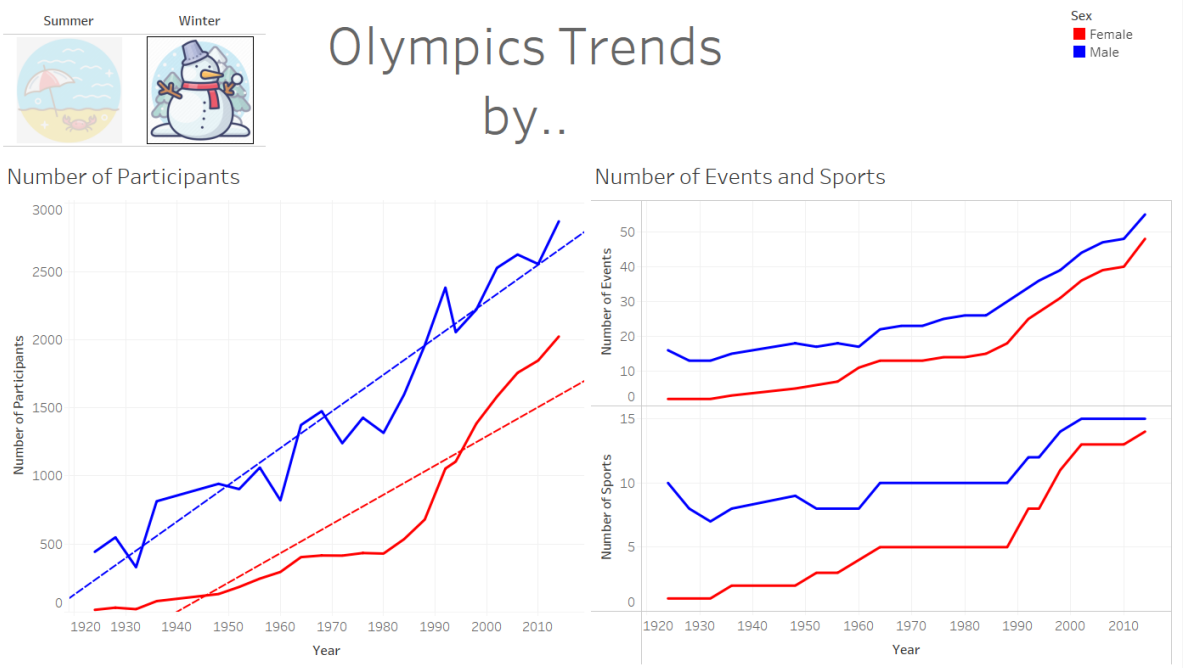


Fig 8: A dashboard showing different trends in the structure of the Winter Olympics

The second is to look closer at the specific sports added over time and compare them across genders. As can be seen, in both the illustrations above and the one below, the introduction of sports for women had a large impact on the structure of the olympics by greatly increasing the number of athletes.

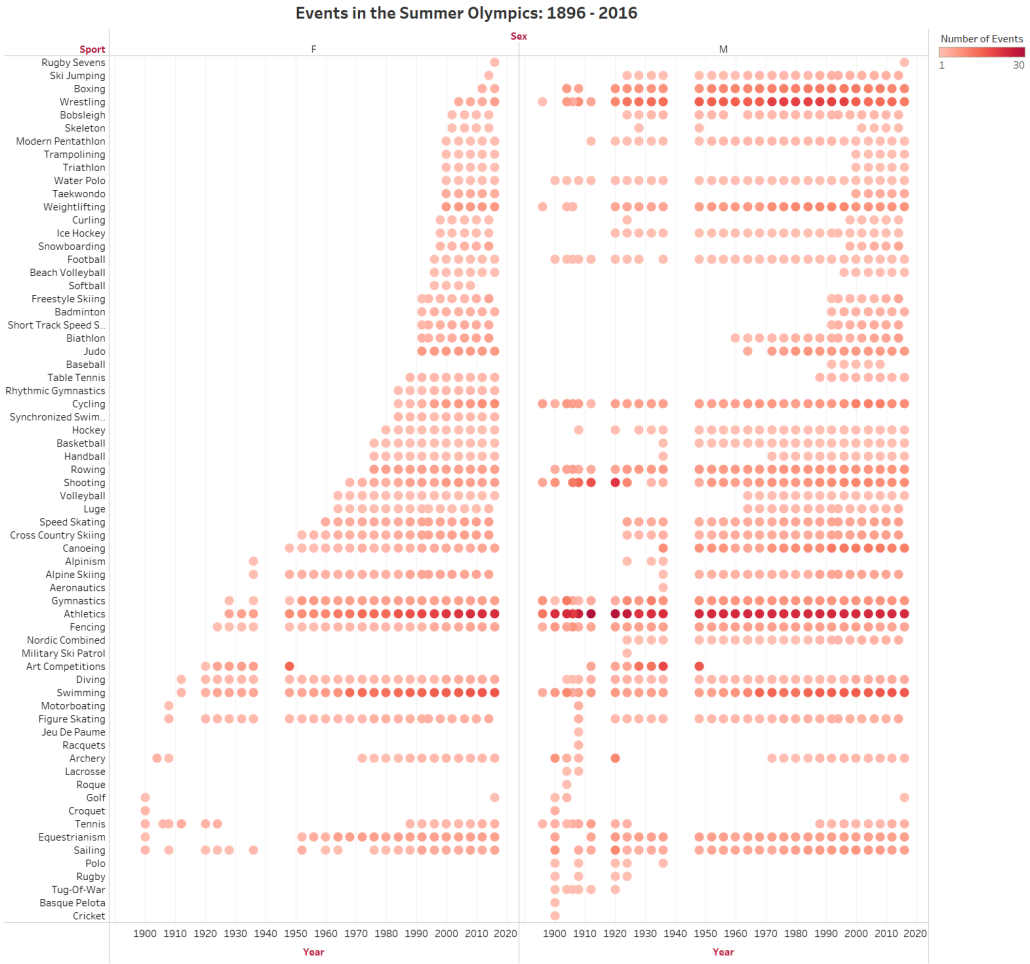


Fig 9: A univariate plot of the sports introduced by year for each gender and colored by the number of events in that sport to better understand its contribution to the overall Olympic Games structure.

For the second direction we have three visualizations that showcase different aspects of the Olympics on an athlete level as opposed to the overall structure.

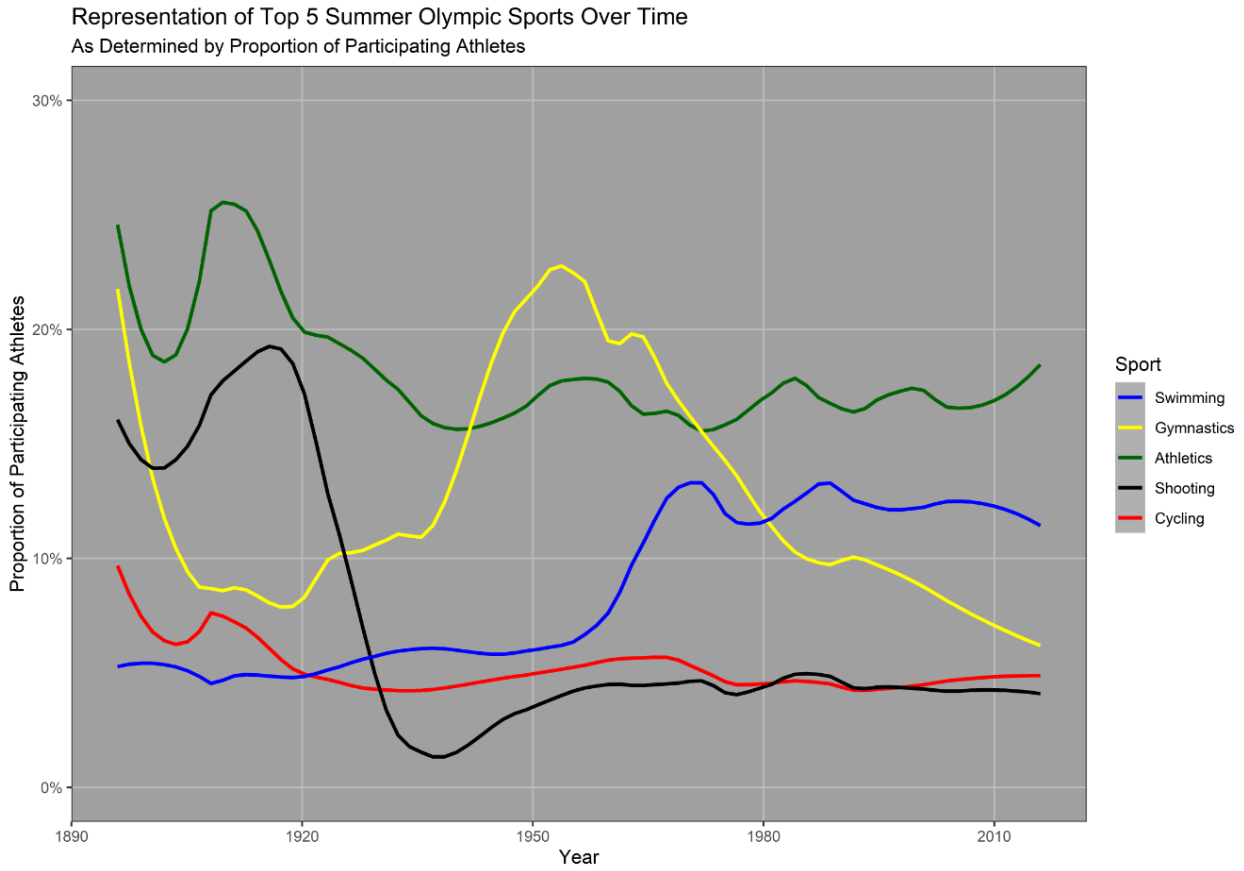


Fig 10: A line graph of the top five Olympic sports over time.

This visualization is a line graph representing the number of participants in each sport over time as a proportion of the total participants in the Olympics that year. Data is limited to the 5 sports with the highest participation across all 120 years of data, where each sport is a different color. It was made using ggplot2 in R, with smoothing applied via loess regression.

Early drafts of this visualization included more sports (10 – 12) than are seen here and was a pure line-graph with no smoothing applied. It was obvious that the number of sports would have to be small in order for the audience to distinguish between them easily, and once it was noticed that 4 of the top 5 sports went through a noteworthy journey, it was an easy enough decision to include only those (the exception is Cycling, which seemed to have maintained it's relative place while declining steadily as more sports are added over the years). The application of smoothing cleaned up the graph quite a bit, and made the relationships between the different sports much clearer. The color palette was decided based on the colors of the 5 rings in the Olympics logo. These colors are easily distinguished between for an audience with full-color vision, but more could be done to make this visualization accessible for those with colorblindness. In particular, the red used for Cycling and the green used for Athletics may be easily confused by a colorblind audience.

There are a few noteworthy insights to be drawn from this visualization that tell a compelling story. The rise and fall of Gymnastics is the most obvious and dramatic of these, peaking in the mid-20th century with participation above 20% of the entire Games and then dropping steadily until the 2016 Games, last in the dataset, where it ends around 7%. Meanwhile, Swimming had steady representation of about 5% – 7% up through 1950 before having a sharp, sudden increase in representation between about

1960 – 1976 up to 12% where it remains for the remainder of the time shown. This sharp increase follows the decline in Gymnastics by about 10 years. Athletics starts and ends at the top, being overtaken only briefly by Gymnastics during its peak. As it is the all-time most represented Olympic sport (most likely due to it's huge breadth of diverse events), this highlights just how dramatic a journey Gymnastics had gone on. Shooting started off very popular but fell off sharply early on. Cycling has been steadily represented throughout modern olympic history, declining slightly as more sports are added to the Summer Games.

The only one of these changes that was well explained when researched was Swimming, for which several technological advances took place including indoor pools and more dynamic material for swimwear. These changes led to a higher standard of competition and subsequently increased public interest in the sport as records were broken repeatedly, leading to the culture of superstar swimmers we see today. (Source: https://olympics.com/en/news/the-history-of-olympic-swimming)

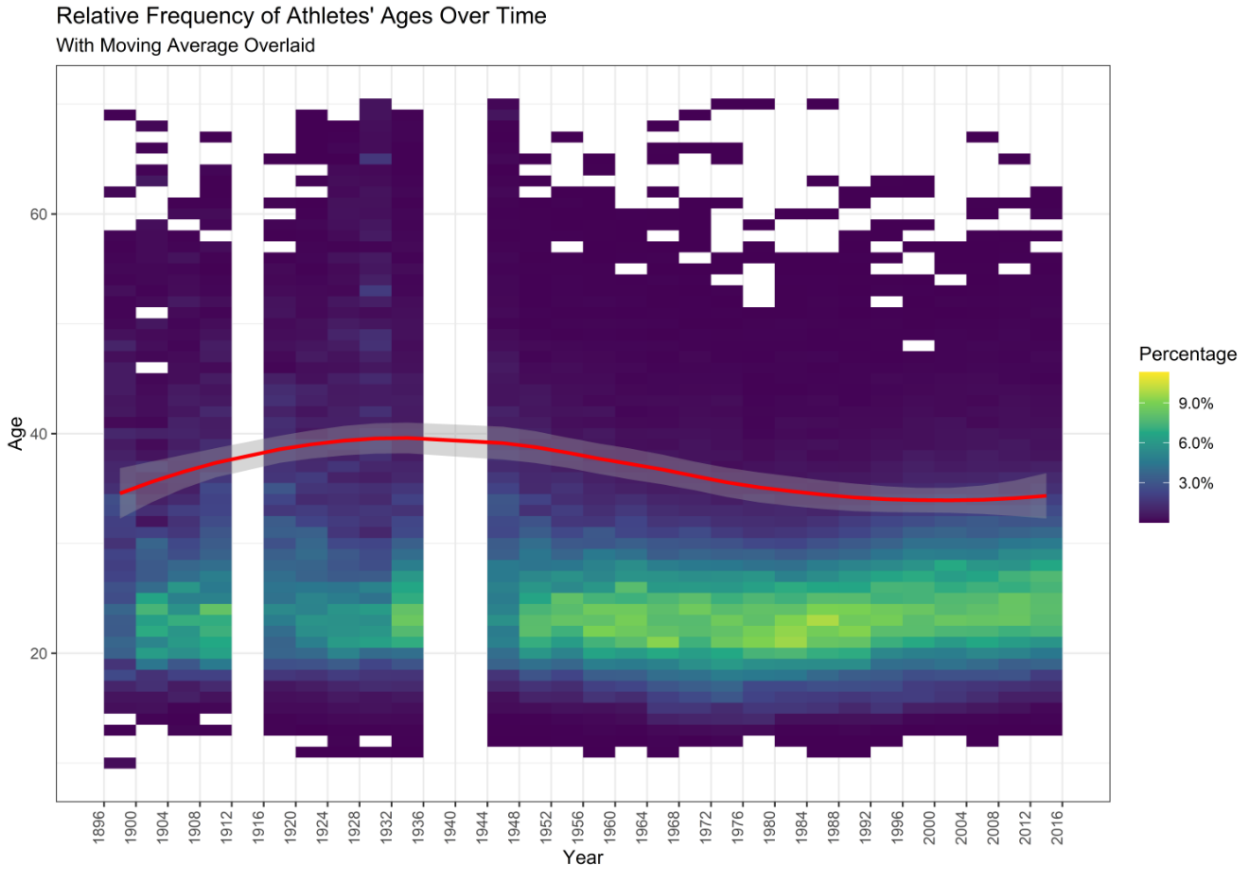


Fig 11: A heatmap of athletes' Age over time.

This visualization is a heatmap of athlete's ages over time, with a moving average calculated with loess regression overlaid. Color is mapped to the proportion of each age represented in a given year. It was made with ggplot2 in R. Originally, this visualization had Age tracked by Sport instead of Year. However, there wasn't really anything interesting to see there, and it proved hard to read. After that change was made, adding the moving average was an obvious next step to take to illustrate the trend more clearly. Additionally, there are several athletes over the age of 70 which had to be excluded from view in order to

better read the interesting part of the visualization. Finally, because color is mapped to a numerical variable, the viridis color palette was chosen for its readability.

The visualization shows that the highest proportion of athletes have been in their early-to-mid 20's throughout modern olympic history, with the distribution tightening over time until about the 1980's when it began to widen out again. Average age rose steadily from the 1896 Games, hit its high of about age 39 around 1936, then fell steadily until leveling off around the year 2000 at age 35.

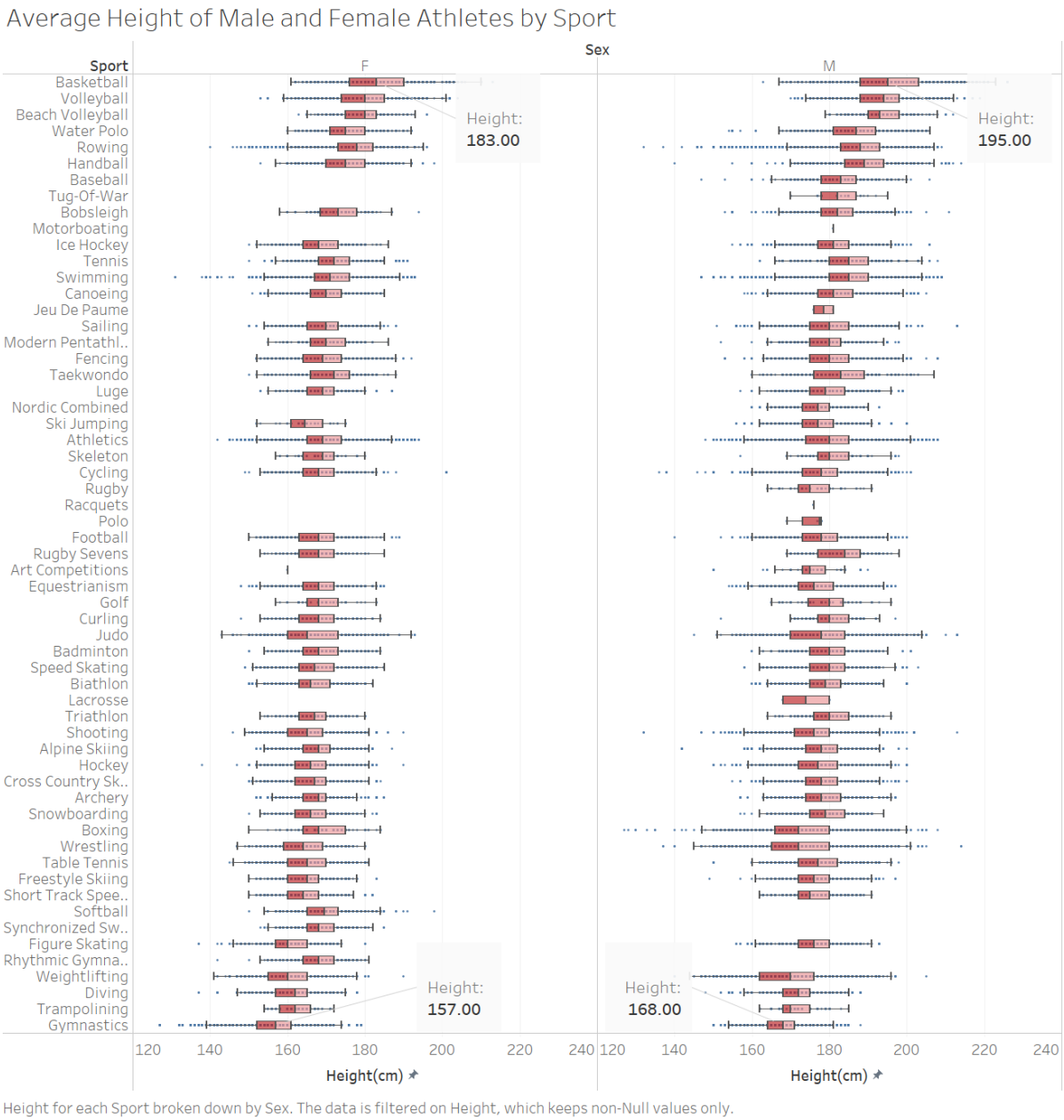


Fig 12: A boxplot to show the differences in athletes' heights by gender and sport.

This visualization is made using a box plot to show how the athletes' physical parameters matter in each sport by representing the average height(cm) of each athlete in each sport which is separated by gender. The box plot would be best fit to show the average kind of information. I realize from the data that there is a drastic difference between heights of female and male athletes in each sport. The basic idea behind this visualization is to compare the height male and female athletes in a particular sport and overall male or female height in each sport. From this graph We can get to know which sport does not have any athletes and what is the minimum height required for a particular sport.

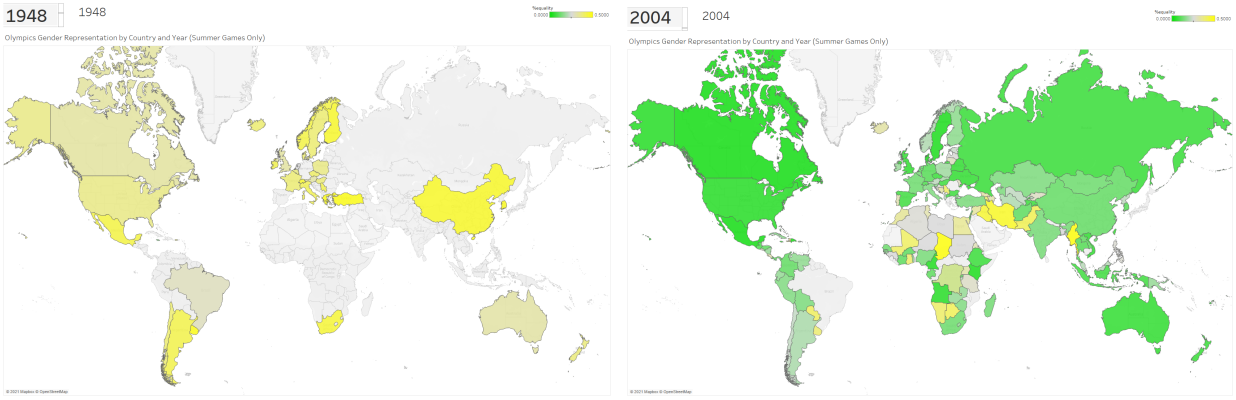


Fig 13: A geographic representation of gender participation in the Olympic games.

Gender participation. Here we look to see that male participation equals female participation, or a 50/50 split of gender representation per country. Green represents 50/50, and yellow represents an imbalance.

IV. Analysis and Discussion

Our visualizations all show unique conclusions that can be drawn from the data. For instance, we can see that over the years, the most common age for competitors has stayed pretty consistent, ranging between 20-30 years old. We can also see that the median height and weight for each sport varies greatly, and that basketball and gymnastics hold the tallest and shortest median competitors. This is expected because both of these sports require a specific body type in order to have the best chance at winning. From this, we can conclude that the Meta for the games has not changed dramatically since the start of the games. For the majority, we can see the Meta very much depends on the sport, as the distributions for height and weight vary between all sports.

The structure of the games has changed since the start 120 years ago. We see that the popularity of sports has been pretty consistent over the years, the top 5 sports have not changed, but their ranking in respect to total participants has. Cycling and shooting held the top spot for the first 30 years or so, until WWII in 1940 where we see a sharp decline in its popularity. At that point, gymnastics becomes the most popular sport to participate in, and we see it hold that spot until recently when swimming and athletics (track and field) take over as the most popular sports. The calculations for popularity were done as percentages of the total participants in the games, as this shows what countries are participating in the most. For instance, athletics, gymnastics, and swimming all have many different events to participate in, and some competitors may compete in more than one event. Our calculations exclude the possibility of double counting, or counting the same player twice for two different events.

Finally, we looked at the gender representation across all of the games since their beginning. Two years of interest were 1948 and 2004, where we can see that the majority of the world holds a gender imbalance that at this point was heavily male dominated. About 50 years later in 2004, we see the majority of countries now hold very close to a 50/50 split in gender representation. While this shows great progress towards equalizing gender participation in the Olympics, there is still some inequality in the games, like the number of events available to women, and the raising the popularity of new female events as they are added.

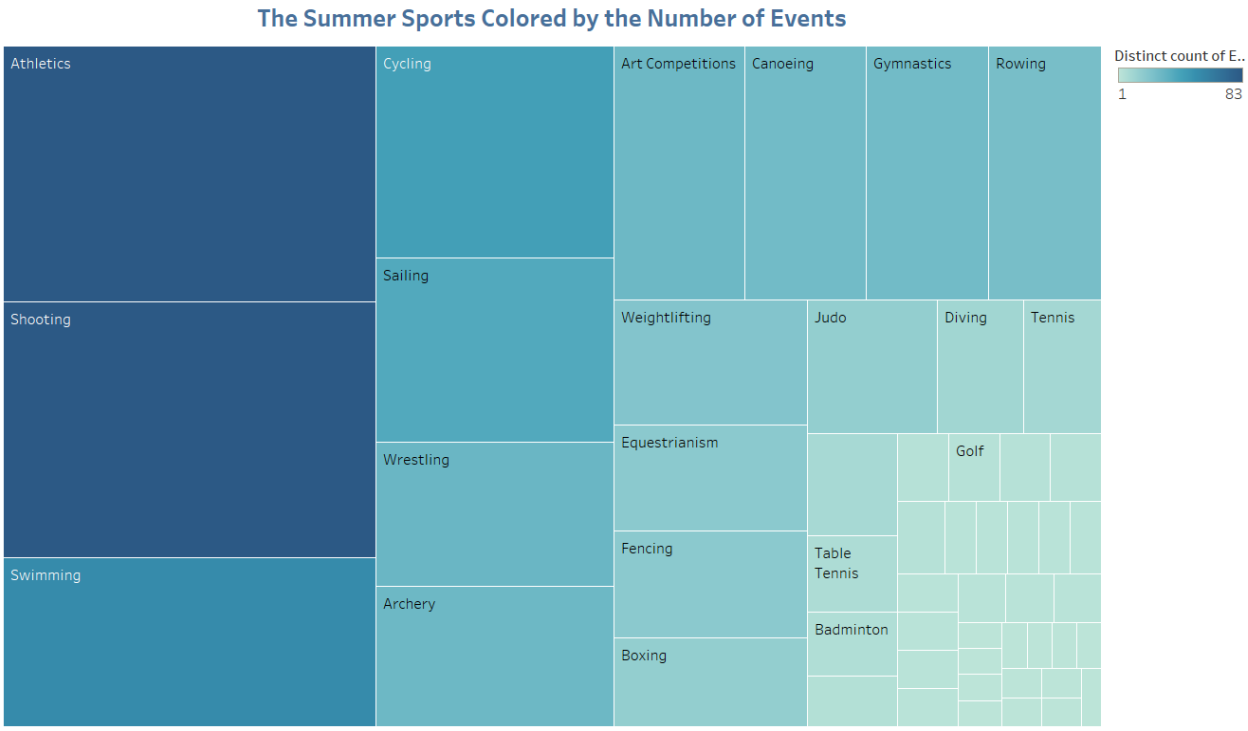
V. Appendix

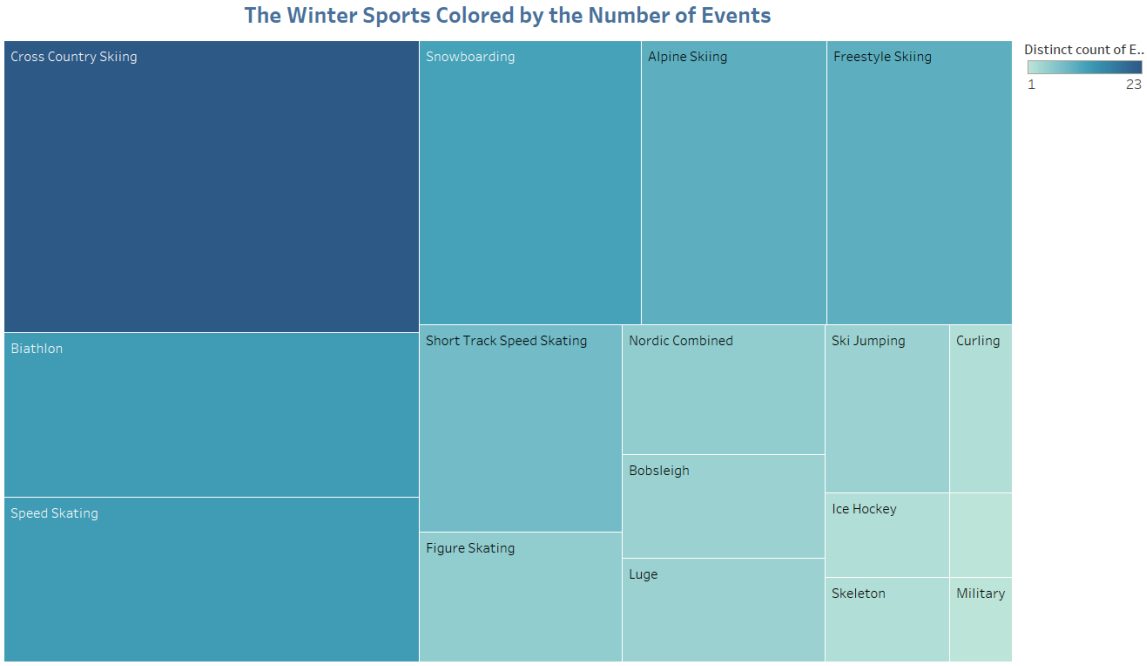
Cassandra Steffey

In this project, I started by looking at the breakdown of the Olympics and its structure over time. Initially, I looked at the number of athletes in the Summer and Winter games separately and separated by gender. This line graph was interesting because there was a large increase in the number of female athletes around 1990 and it made me curious about what that looked like on the sport level. This led to looking at the years in which each sport was introduced for Men and Women as a univariate plot. Originally, I looked at the genders separately by each season, but it was hard to compare that way. It was decided that I should try to do them as a quadrant where the male sports were next to the female sports and then the top was summer, and the bottom was winter. The problem that was found with this was that the Olympics did not used to be separate seasons and so while trying to sort the data I was having trouble with the years for the winter sports. So instead, I created a plot with all sports together separated by gender side by side. I then sorted by females because I was most interested in, and I colored the points based on the number of events in that sport. This version of the plot was one of the ones demonstrated above in the final chosen visualizations.

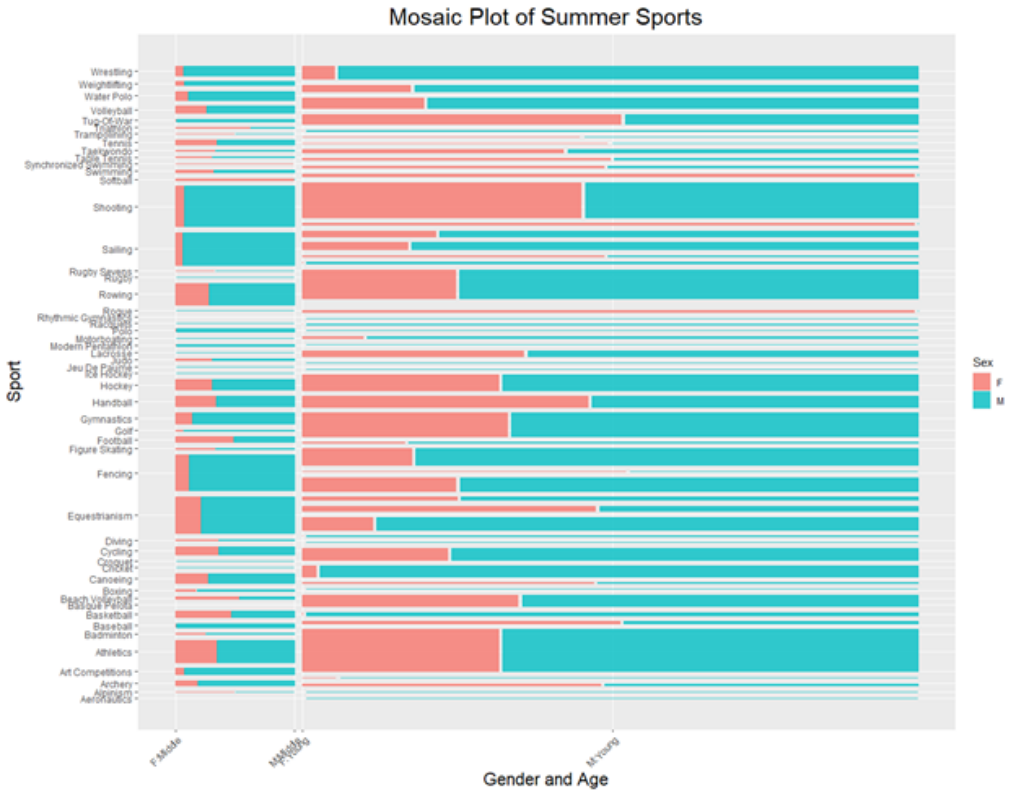
To visualize the same information in a different way, I created a tree map and a mosaic plot. In the tree map, I looked at the sports and the size was based on the number of events in each sport. This was separated by season but not by gender. The tree map for the summer Olympics was hard to read due to the large number of sports available but the winter Olympics tree map was a very good representation of

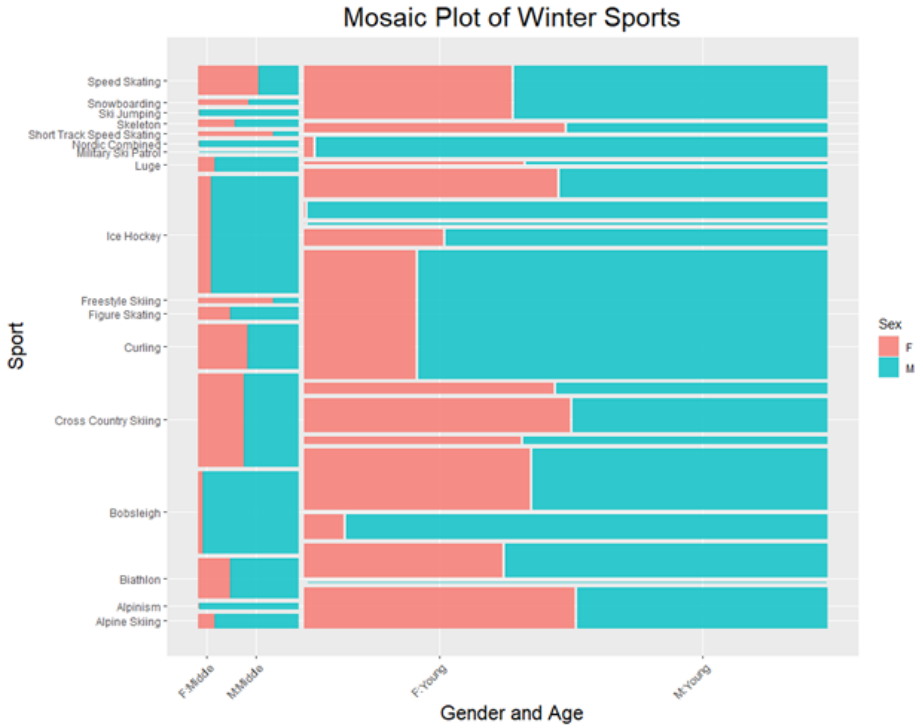
the breakdown of the sports.



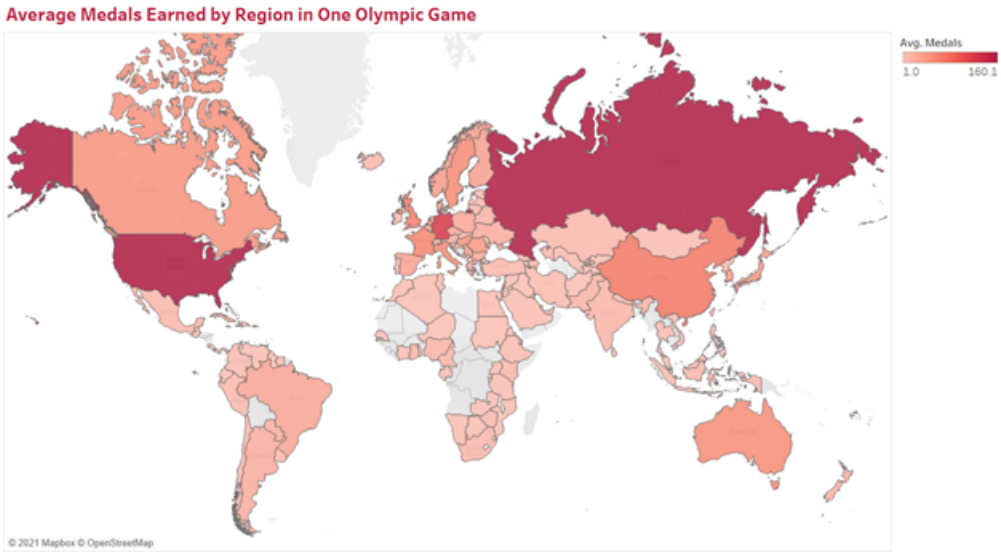


For the mosaic plot, I looked at the breakdown of age and gender with the sports. In the mosaic plot we see the same issue as the tree map, that the number of sports is too much to visualize as well as the number of ages by gender. To help with the gender and age, I separated the age into four equal bins. However, the two oldest groups are very small, so I ended up only comparing the two younger groups. These are still hard to read so they were not considered for the final visualizations.


Mosaic Plot of Summer Sports
Sport
Wrestling
Weightlifting
Water Polo
Volleyball
Tug-Of-War
Tennis
Taekwondo
Synchronized Swimming
Swimming
Softball
Shooting
Sailing
Rugby Sevens
Rugby
Rowing
Modern Pentathlon
Lacrosse
Judo
Jeu De Paume
Ice Hockey
Hockey
Handball
Gymnastics
Golf
Football
Figure Skating
Fencing
Equestrianism
Diving
Cycling
Cricket
Canoeing
Boxing
Basque Pelota
Basketball
Baseball
Badminton
Athletics
Art Competitions
Archery
Alpinism
Aeronautics
F:Midde
M:Young
Gender and Age
Sex
F
M


Mosaic Plot of Winter Sports
Sport
Speed Skating
Snowboarding
Ski Jumping
Skeleton
Short Track Speed Skating
Nordic Combined
Military Ski Patrol
Luge
Ice Hockey
Freestyle Skiing
Figure Skating
Curling
Cross Country Skiing
Bobsleigh
Biathlon
Alpinism
Alpine Skiing
F:Midde
M:Midde
F:Young
M:Young
Gender and Age
Sex
F
M

The last visualization I worked on was a world map. This took a little bit more working with and manipulating the data. I was focusing on the performance of the different countries, so I took the number of medals (gold, silver, and bronze) for each region for each year. I then saved this new dataset and worked in tableau to plot the map. For the map, the coloring was done based on the average medals earned by each region in a year to try and limit the weight that some countries like the US and Russia have for participating for more years. That being said, they still average higher than most regions in the number of Medals they earn in a year. Looking at this plot below, we can see that this gave a range of one to one hundred and sixty medals earned in a year.



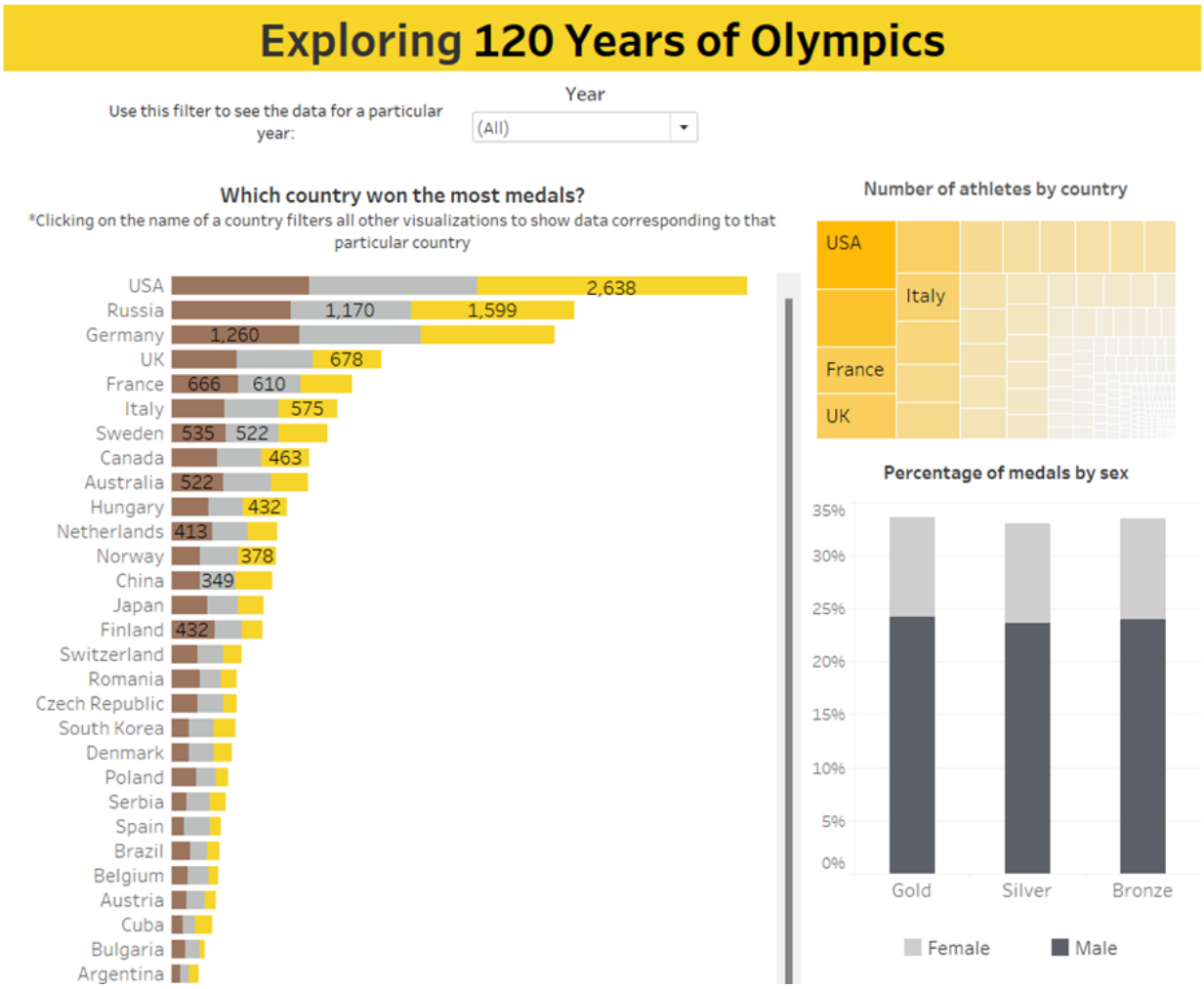
Due to the nature of an average, the later years with more events and consequently more levels have more weight. To try and counter this, I looked at using the frequency instead of the average, but I could not get the map to work in the way I wanted. For that reason, these were not included in the final visualizations.

I have learned a lot in this course about the appropriate types of graphs as well as the tools necessary to create them. It was so interesting to learn about how people perceive information and the different types of ways that people can be deceived. Especially when it comes to the color schemes or the aspect ratio of the visualization. It was great to learn how to create the plots in R and how to make it personalized. The most important and useful part of this course was learning which types of visualizations work best for the different data types. It can be hard to decide what visualizations to use but now I know so many different types that I can experiment and compare for the best option.

Samuel Prasad Chinta:

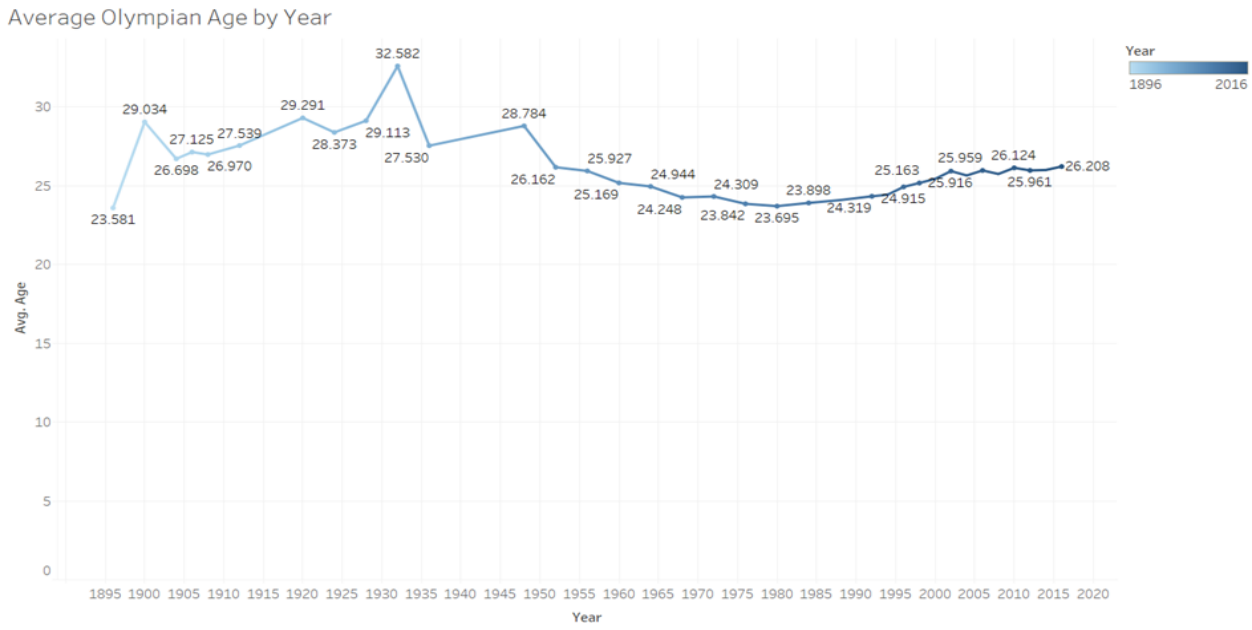
In this project, I looked at the number of athletes who won medals (Gold, Silver and Bronze) with respect to their countries in both the Summer and Winter games. The stacked bar graph was interesting because there was a large increase in athletes from 1860 to 2016 and it made me curious about what that looked like and how many women and men won the medals. This led to looking at the years in which both Men

and Women won the medals on percentage using the bar chart. Originally, I used tree map for the visualization where we can see that number of medals won by countries and created a dashboard with respect to medals won by the countries and with respect to percentage of medals won by the gender, with together I created a dashboard together, with the filter option where we can count the number medals won by the athlete with respect to the year.

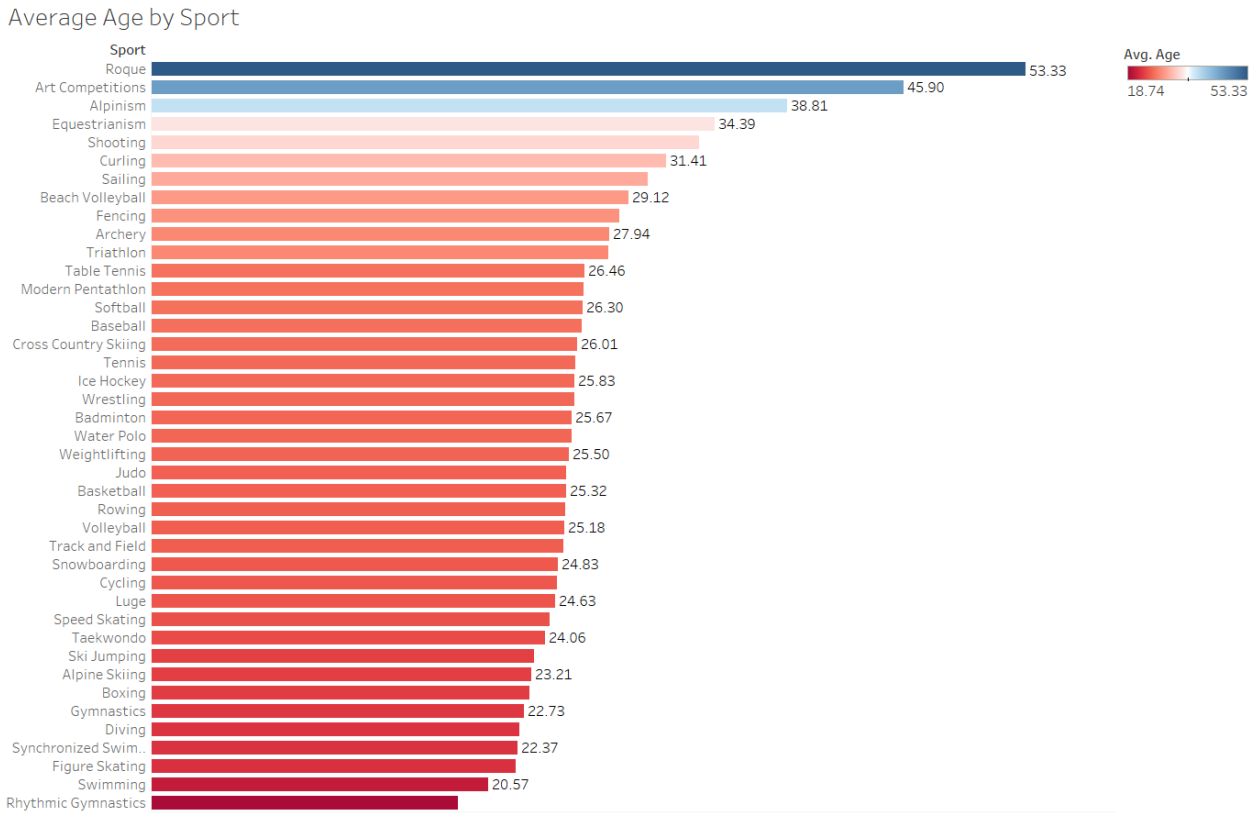


later I have created another visualization looking at the average age of all the athletes over the years, but it was hard to compare that way. So, I created a line graph which shows the average age of the athletes over the years, where we can observe that the minimum average age of the athlete was 23.58 in the year 1896 and maximum average age of the athlete was 32.58 in the 1932. And gradually it became an average of 26 by the year 2016. And another visualization using bar chart were we can see average age

by the sport

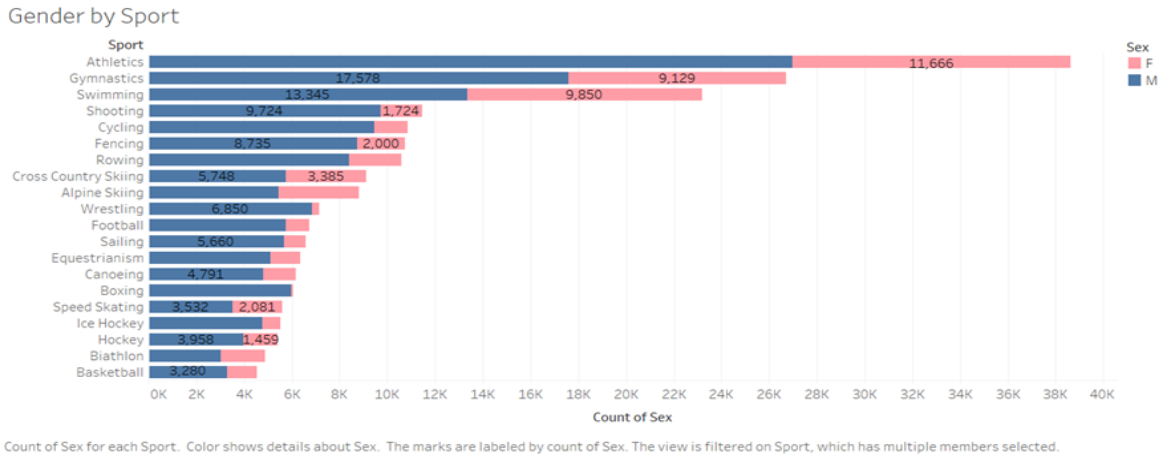


The trend of average of Age for Year. Color shows details about Year. The marks are labeled by average of Age.



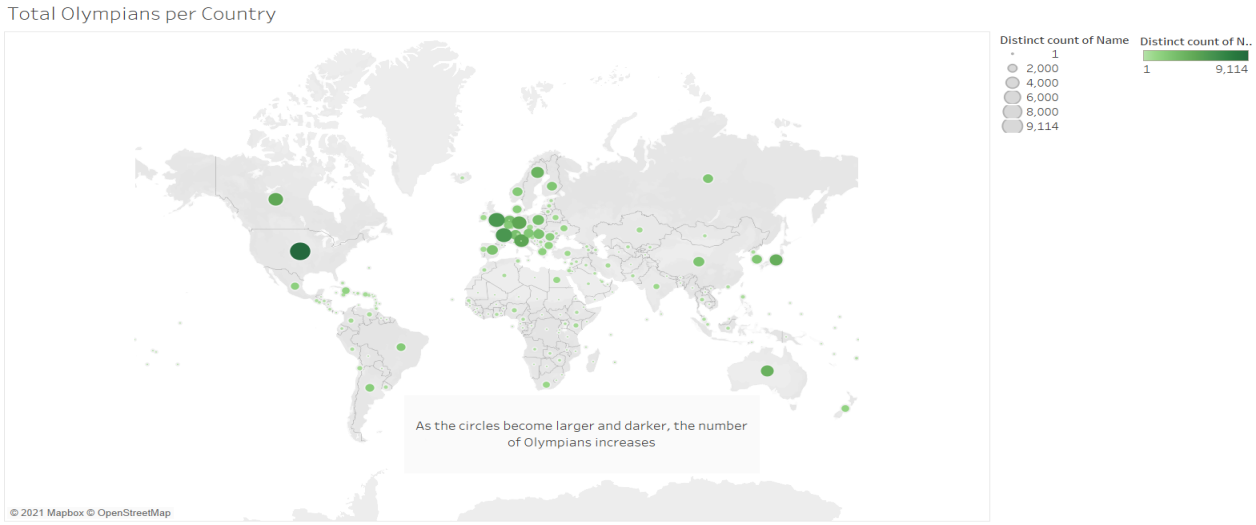
Average of Age for each Sport. Color shows average of Age. The marks are labeled by average of Age.

I have created another visualization where I have created a visualization on a stacked bar graph where the number of athletes participated with respect to gender in various sports over the years.



Count of Sex for each Sport. Color shows details about Sex. The marks are labeled by count of Sex. The view is filtered on Sport, which has multiple members selected.

I have created another visualization where I have created a visualization on geographical location where we can observe the integrity of olympic athletes over the years.



Map based on Longitude (generated) and Latitude (generated). Color shows distinct count of Name. Size shows distinct count of Name. Details are shown for Team. The data is filtered on Season and Games. The Season filter keeps Summer and Winter. The Games filter keeps 51 of 51 members.

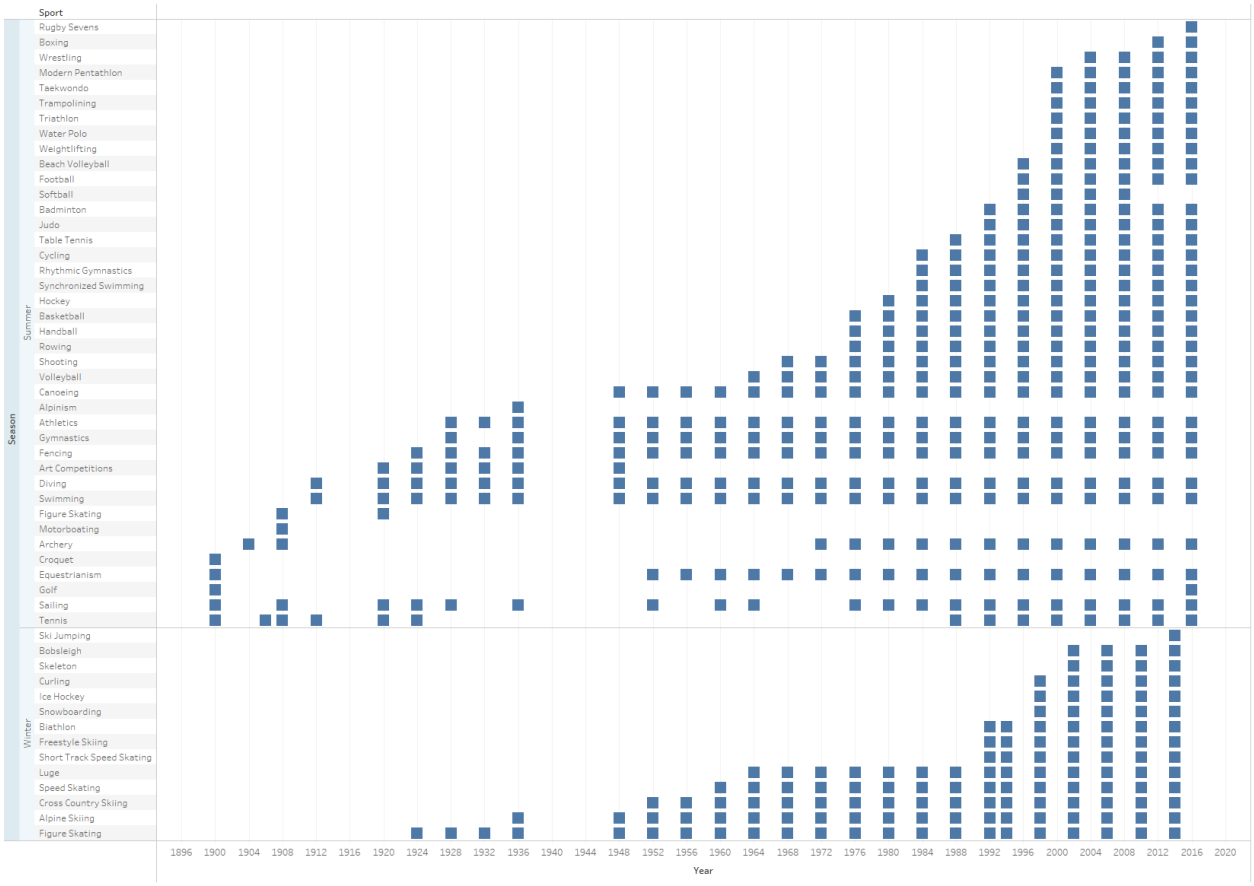
In this course I have learned a lot about the visualization techniques such as what type of variables to be used for visualizations and about appropriate types of graphs as well as the tools necessary as to how to create them. It was on the group project where I have learned even more, where using my teammates ideas and guidance I was able to learn even more and apply them on the project, especially it was very much interesting to learn about how people perceive information and the different types of ways that people can be deceived. Especially when it comes to the colour choices, scale choices or the aspect ratio of the visualization. Tableau was very much helpful for visualization, but it was R where we can create plots and make it even more personalized visualization. The most important and useful part of this course was learning which types of visualizations work best for the different data types. It can be hard to decide

what visualizations to use but now I know so many different types that I can experiment and compare for the best option. With a lot of practice on R and Tableau, I can create even more beautiful visualizations.

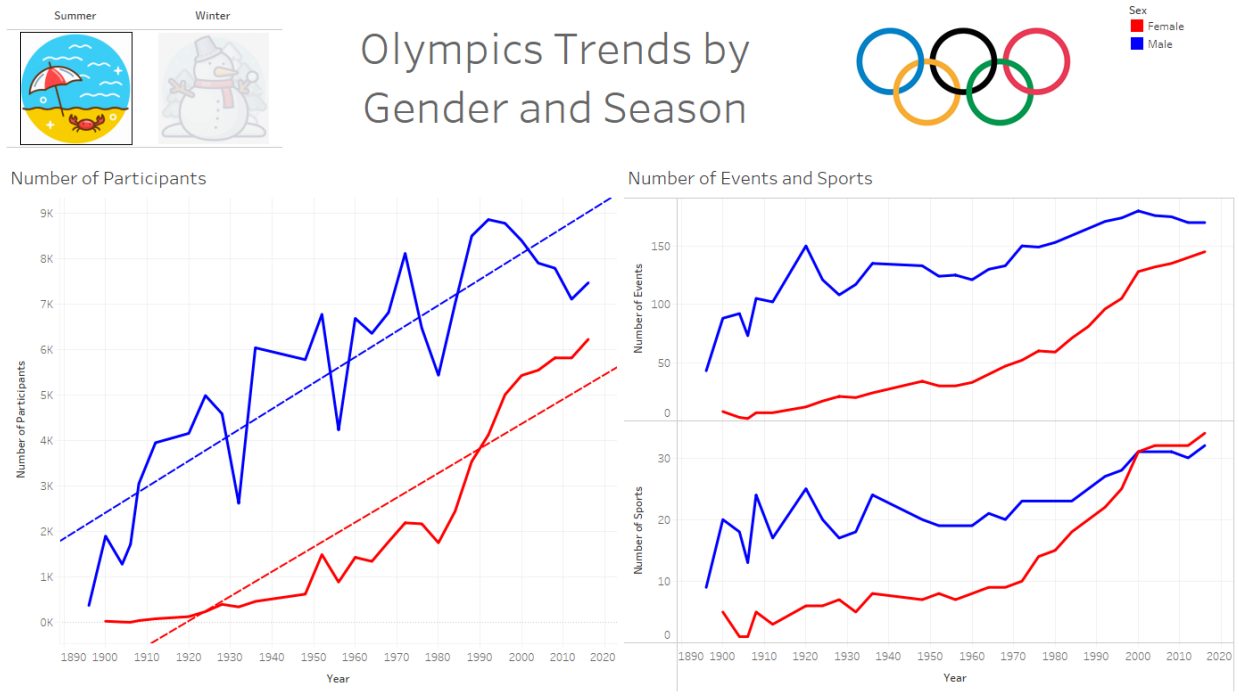
Matt Morais

For my first visualization, I worked with Cassie to look closer at what sports were held in which years, so we can both see how the size of the games changed as the Olympics grew from year to year. The visual below shows the top part as the summer games, and the lower as the winter. Cassie had the initial design, and I added to it by making the markers a bit larger, and inverting the y axis so that it looks like growth rather than decay. Some may say it is tough to see when wrestling started in the summer olympics, because the label is far from the actual data points. While this is true, the intent of the graph is more to show how the games have grown in size, and the rate at which they are doing so. With that as the intent, the goal of this visualization is captured. As for color schemes, it did not make sense to try and encode another variable to this mapping. The chart would have become too cluttered and the meaning behind it would have been increasingly difficult to see. I considered trying this with a line graph, but felt that seeing just two lines, that showed total games per season, did not fully embody the total number of events held in the Olympics each year. Perhaps an area line graph could capture the same message, but we would lose the sights of what sports were added most recently. This visual does an excellent job at both showing how the games have grown, as well as what sports were added in what years.

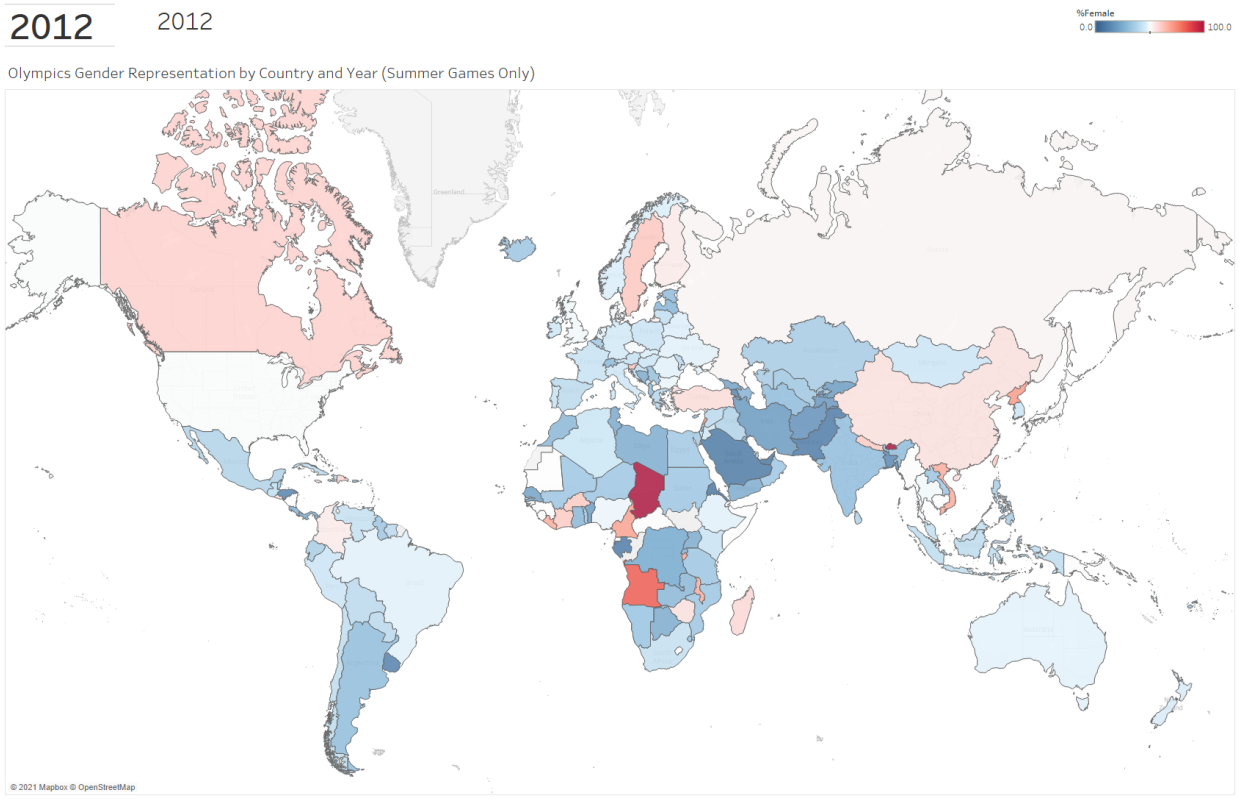
Female Olympic Events by Year and Season



My next visualization I began experimenting with interactivity, as this is something I have never had the opportunity to work with before in my daily work. I started simple, with a toggle between summer and winter, and we can see that for summer sports, the number of female sports exceeds the number of male sports. That said, there are still fewer female participants, but we can see that there has been a large increase in female participation starting in the 1990s. The colors I chose align with what people typically see as male (blue) and female (red) so that they can be easily encoded by the viewer. For my interactive buttons, I watched a tutorial on how to create your own icons in Tableau, and used some pngs for summer and winter to allow the toggleable filter. I wanted to look further at gender participation, while maintaining the interactivity of the dashboard. My next visualization dives deeper into that.



My next visualization is close to the final one I produced. I wanted to look at global participation of females by country, and how that has changed over time. I started thinking using the same color scheme would work, red for countries with majority female, and blue for countries that were majority male. I submitted this for the final presentation and was not satisfied. I thought immediately afterward that this is not showing the message I wanted to convey. The point of identifying gender gaps is not to highlight and praise those that send more female participants. The point is to see how we are progressing towards equal representation between genders in the olympics. The divergent color scheme and scale does not accomplish this.



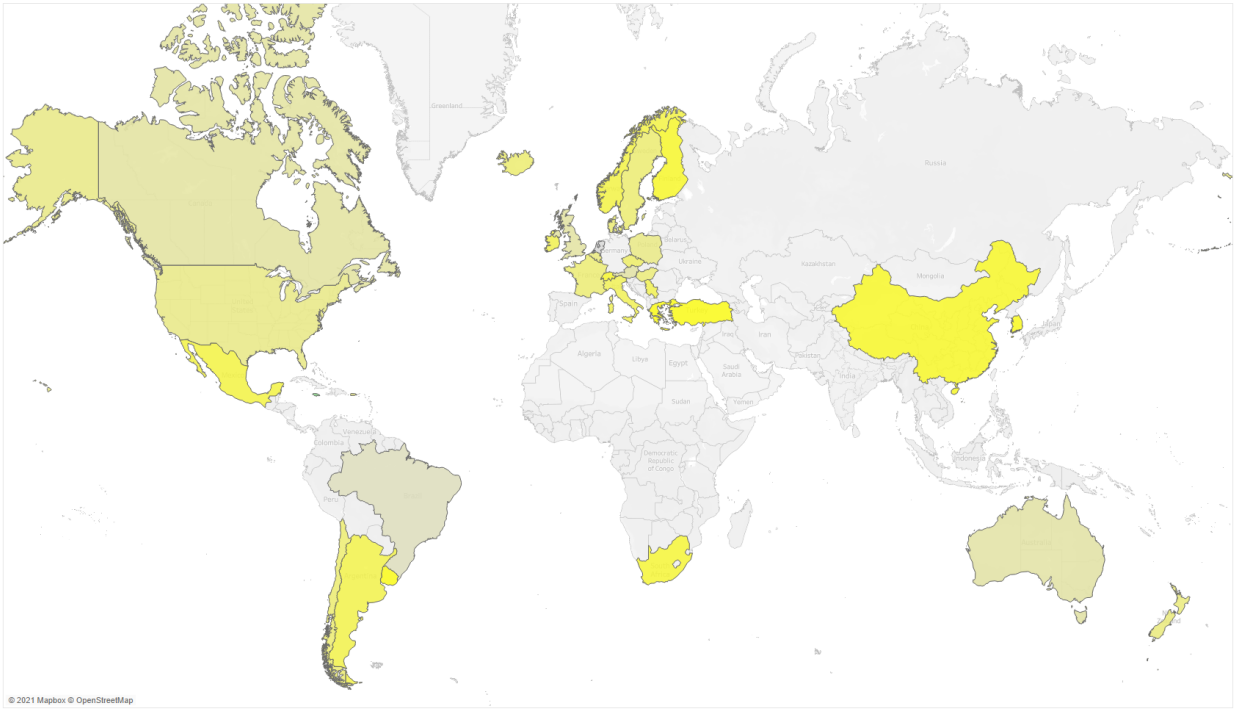
In my final graphic, I adjusted the scale to instead go from 0-0.5, where 0 represents a 50/50 representation of participants. To eliminate confusion, I also changed the colors to yellow (poor equality) and green (perfect 50/50 equality). You can see in the graph below, that the world has come a long way towards equality in the Olympics since 1948. One drawback to this representation, is that larger countries at the north and south appear to over-exaggerate how this looks. Ideally, we could use a diffusion cartogram so that more countries are visible, and size/distortion does not play such a large role. This task I have yet to see done on a global scale, and all the references I could find seem to target smaller locations for this technique. For that reason, I kept the global image as a choropleth, and feel that when viewed as the dynamic dashboard, the viewer can more easily see how countries change, as well as zoom in on areas they are interested in. This is where interactivity can be extremely helpful in ensuring no content is lost.

1948

1948

%equality
0.0000 0.5000

Olympics Gender Representation by Country and Year (Summer Games Only)

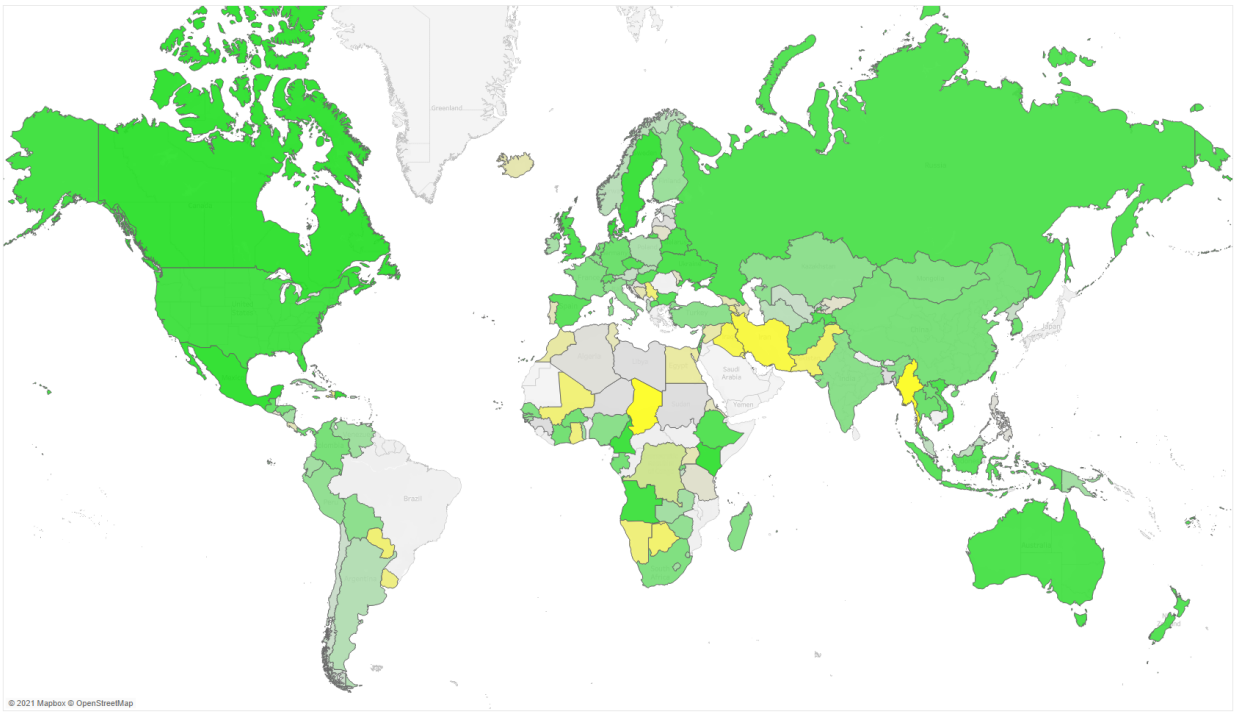


2004

2004

%equality
0.0000 0.5000

Olympics Gender Representation by Country and Year (Summer Games Only)

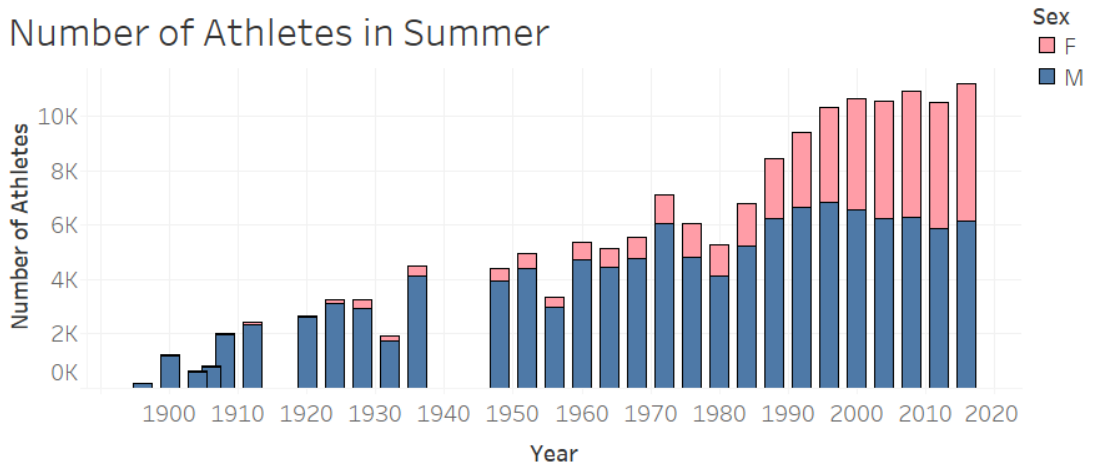


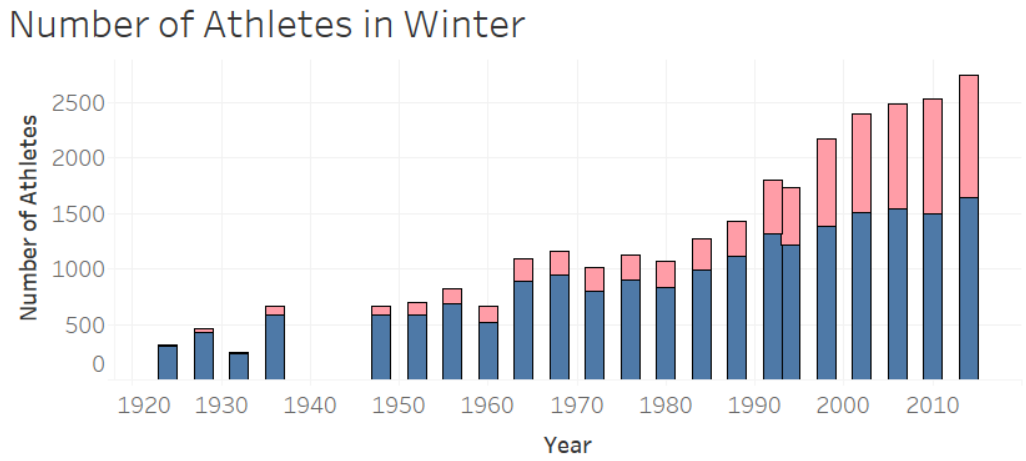
Conclusions that can be drawn from my visualizations:

1. There has been a large growth in the number of sports/events in the Olympics since the games began.
2. With that growth, the spark of female events has taken off in the 1990s, and continues to push the games towards equal gender representation.
3. Visually, the first world countries seem to send the most participants, and have balanced out their gender representations roughly 20 years faster than countries that send fewer athletes, or are under-developed. Part of this may be due to cultural norms / gender roles.

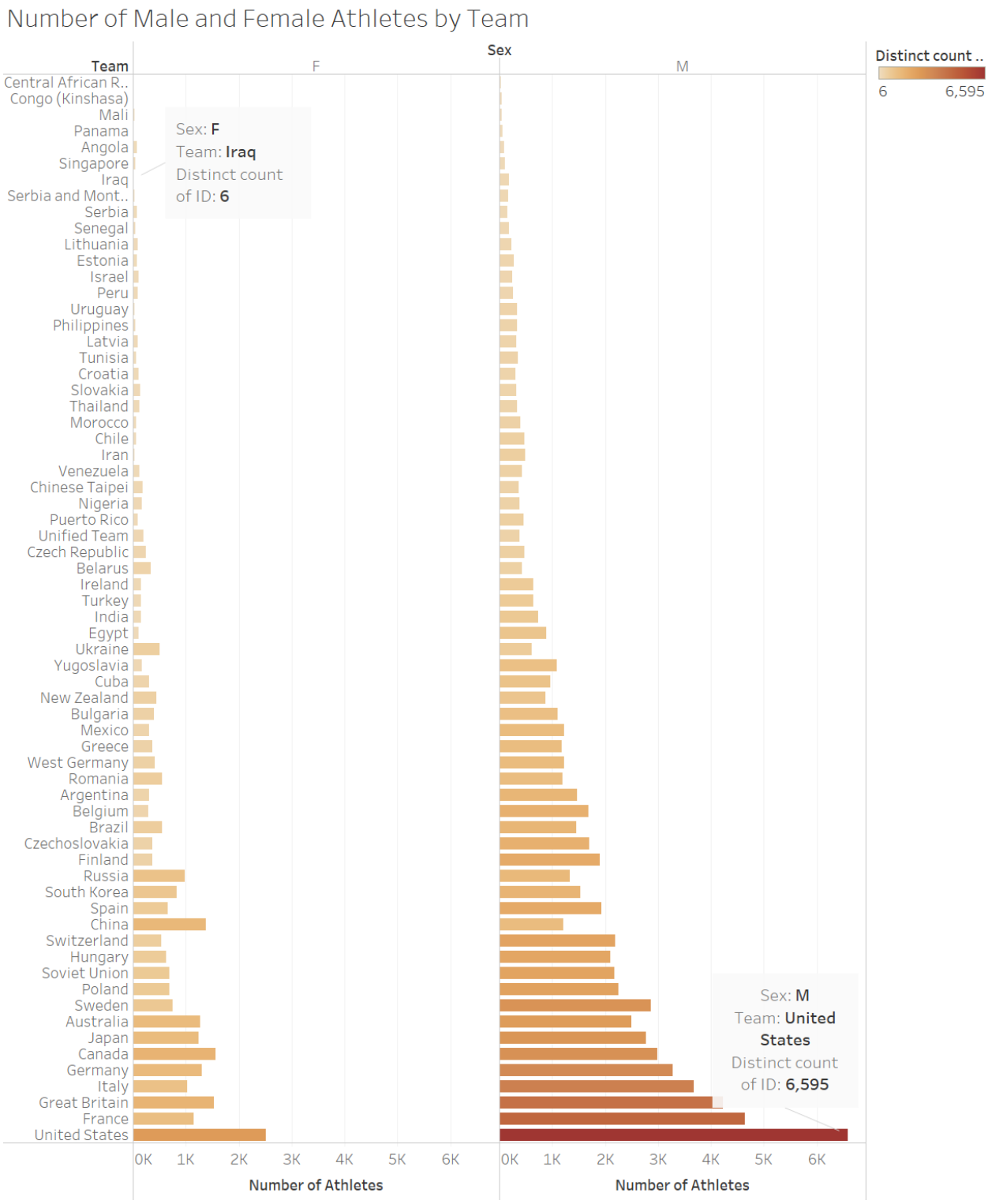
Jaimi Patel

I found the data from kaggle and the main reason to take this dataset is it has many entities that we can work on and make various visualizations. I was very much interested in the athlete physique meta. So Initially I started by comparing Number of Athletes per season and separated them by their gender. From the graph I get to know that no female athletes participated in summer Olympics until 1900 and in 1940 no athletes participated in the Olympics. Overall the number of male athletes is more than female but the trend is now changing over time by more and more females participating in the Olympics. The line graph conveys the comparison more clearly than the bar graph. That's why we considered earlier line graphs.



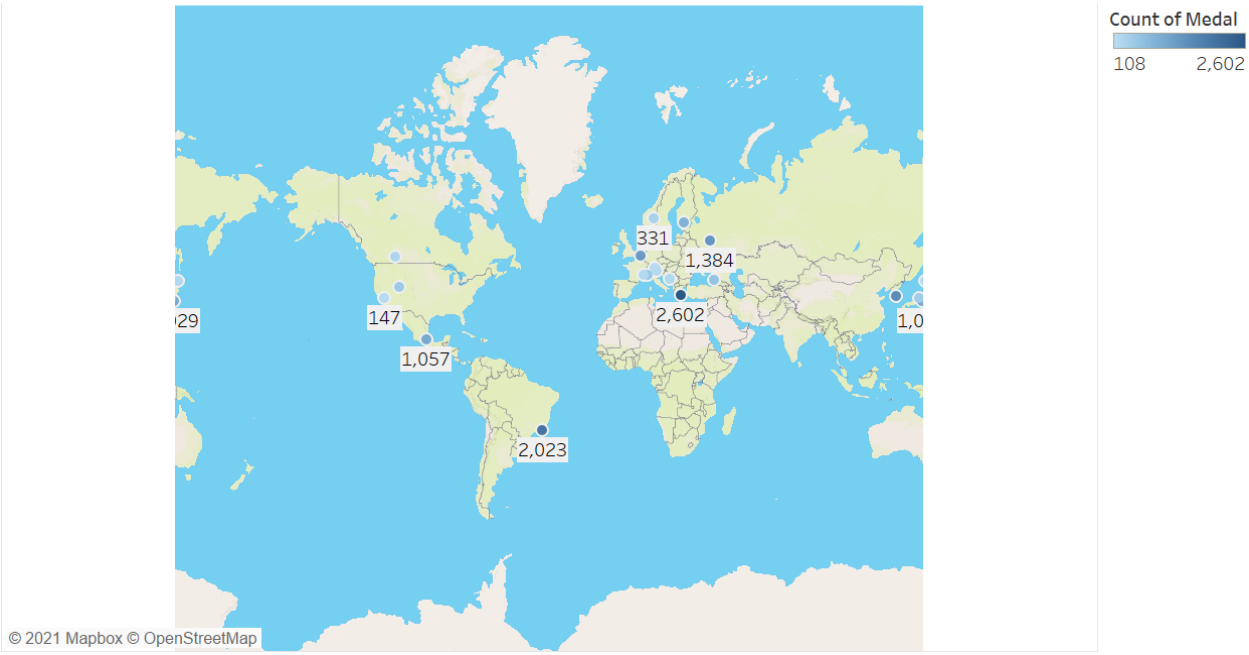


Then I started visualizing the number of female and male athletes in each team. The graph is separated by gender in a single graph so viewers can compare the male vs female athletes in each team by color frequency. I wanted to know which team has more athletes and which team has the least number of athletes. This basic visualization conveys the information that Iraq has
the least female athletes and the USA has the highest number of athletes. The simple bar graph represents the information more clearly.

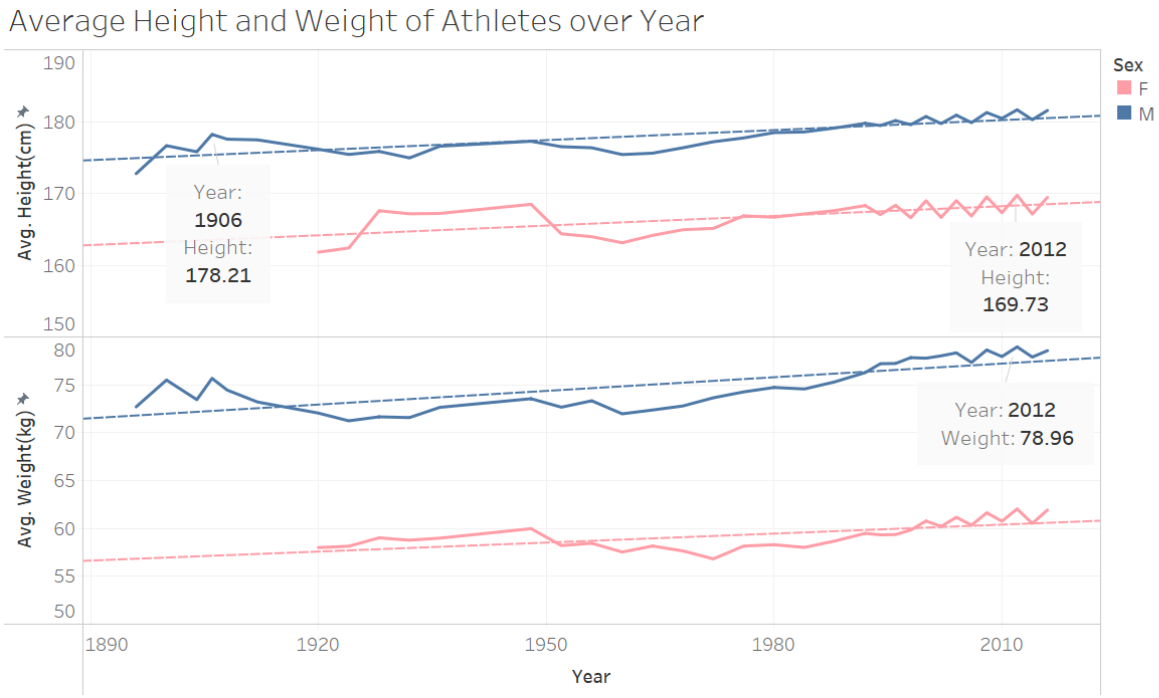


In this visualization, I tried a world map. I was concentrating on the different countries' performances, thus I counted the total medal won by each country. The dot color frequency represents the count of medals and also the number of medals labeled in each region. The data was not conveyed as it needed to be,

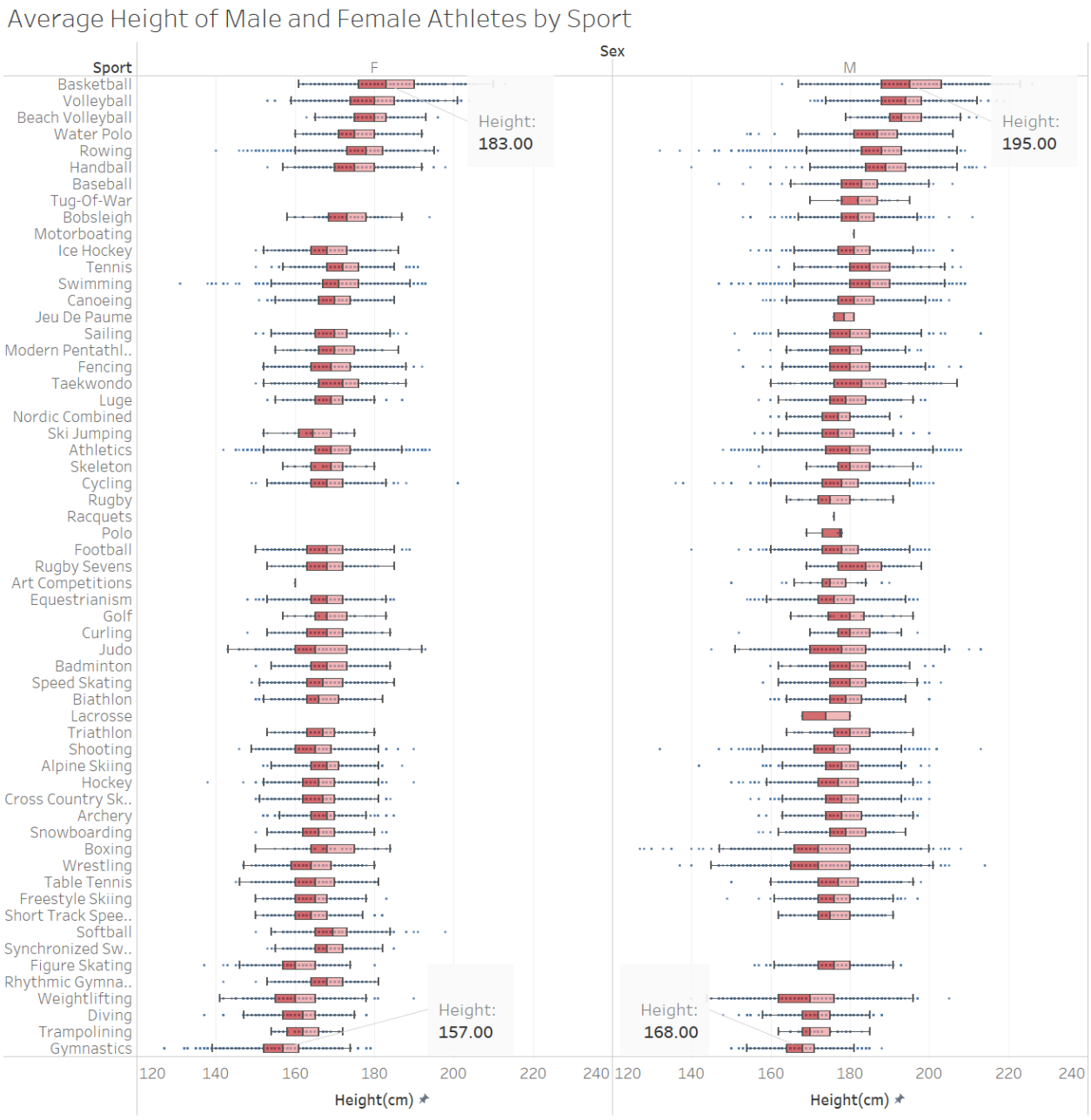
and we could do better with this graph so my other teammate worked on my idea and gave the final touch with giving color in each region by their count of medals won.



For the next visualization, I was concentrating on Height(cm) and Weight(kg) of the athletes to show viewers the athletes' BMI over the years. The line would be best to see a clear comparison of Height and Weight of the athletes each year. For the better comparison the whole graph is divided in two portions and separately shows data. The trend line is used for viewers to understand the flow of athletes height and weight over years. This graph conveys the information better as I needed. I also separated this visualization by gender to get seperate results. The idea is very similar to one of my teammates so we decided to consider his graph as it was in the dashboard.



My last visualization was very similar to the above one but I decided to dive into more details. I wanted the viewer to see how the athletes' physical parameters matter in each sport by representing the average height(cm) of each athlete in each sport which is separated by gender again. I thought the box plot would be best fit to show the average kind of information. I realize from the data that there is a drastic difference between heights of female and male athletes in each sport. The basic idea behind this visualization is to compare the height male and female athletes in a particular sport and overall male or female height in each sport. From this graph We can get to know which sport does not have any athletes and what is minimum height required for a particular sport.



Height for each Sport broken down by Sex. The data is filtered on Height, which keeps non-Null values only.

I learnt a lot about various visualization methods in this course such as what sorts of variables give the best visualizations and which types of technique to use for comparison. The best part was working with Tableau which makes visualization easy to make and I can play around with lots of techniques which suit my visualization best. Initially, I was not able to think more visualizations from entities but as I get involved myself in this course I am now able to think which type of visualization gives best information to the audience. Also the team project which gave me an opportunity to share ideas with others and see which directions can be considered to make visualizations.

Alex Rosenblum

I have been an active member of this project group, and occasionally took a leading role for certain key deliverables. Specifically, I concerned myself with final organization of several milestone submissions and facilitated decision-making for the final explanatory visualizations to be included in the presentation and report. Additionally, as the main group member working primarily in R, I assisted the rest of the team in some of the data manipulation tasks needed for their exploration.

During exploratory analysis, I took an interest in the distributions of the various numerical variables, like athlete's Age, Height, and Weight, comparing distributions from different categorical subsets like Sex, Sport, and Era (a variable I made by putting Year into bins). Some of these explorations were shown in the ridgeline plots in Part II, Figs 5 and 6. However, after trying many combinations of these categoricals and not finding anything terribly interesting, I pivoted my investigation to cover a wider scope. The area which I eventually found most interesting to investigate involved calculated variables, specifically analyzing the proportional makeup of various aspects of the Olympic Games events, year by year. My two submissions to the final report are those titled "Representation of Top 5 Summer Olympic Sports Over Time" and "Relative Frequency of Athletes' Ages Over Time", both of which I produced on my own using ggplot2 in RStudio.

VI. Appendix 2 Code

Cassandra Steffey - R Code

```
### Libraries ###
library(gcookbook)
library(ggplot2)
library(scales)
library(tidyverse)
library(lubridate)
library(dplyr)

### Exploratory Analysis ###
# Look at region data
noc_regions <- noc_regions %>% select(-notes)

# merge into new dataset
olympic <- merge(athlete_events, noc_regions)
# Summary Statistics
summary(olympic)

# Looking closer at the two games
ggplot(olympic, aes(x = Season)) + geom_bar()

# Look at the number of countries
n_distinct(olympic$NOC) - 299

# Filter by games season
summerGames <- filter(olympic, Season == "Summer")
winterGames <- filter(olympic, Season == "Winter")

# Look at the countries by games
n_distinct(summerGames$NOC) #Summer - 229
n_distinct(winterGames$NOC) #Winter - 119

# Look at the number of regions
n_distinct(olympic$region) #206

# Write to csv
write.csv(olympic, 'olympic.csv')
```

```
# OlympicMedal50
group_by(olympic, region)

#Remove count variable
olympic <- olympic %>% select(-X)

# Bar of male and female
ggplot(olympic, aes(y = Sex)) +
  geom_bar() +
  labs(title = 'Number of Male and Females in the Data',
       x = "Sex",
       y = "Count") +
  theme(
       panel.background = element_rect(fill = "white", linetype = "solid", size = 0.5, colour
= "grey50"),
       panel.grid.major = element_line(colour = "grey30"),
       panel.grid.minor = element_line(colour = "grey80"),
       axis.title.y = element_text(size = rel(1.5)),
       axis.title.x = element_text(size = rel(1.5)),
       plot.title = element_text(size = rel(2), hjust = 0.5)
  )

#Separate the data
olympicMales <- filter(olympic, Sex == "M")
olympicFemales <- filter(olympic, Sex == "F")

# line of male and female
ggplot(olympic, aes(x = Year)) +
  geom_line(data = olympicMales, aes(y = count(Sex)))
+
  labs(title = 'Number of Male and Females in the Data',
       x = "Sex",
       y = "Count") +
  theme(
       panel.background = element_rect(fill = "white", linetype = "solid", size = 0.5, colour
= "grey50"),
       panel.grid.major = element_line(colour = "grey30"),
       panel.grid.minor = element_line(colour = "grey80"),
       axis.title.y = element_text(size = rel(1.5)),
       axis.title.x = element_text(size = rel(1.5)),
       plot.title = element_text(size = rel(2), hjust = 0.5)
  )

### Visualizations ###
# Map
# remove x
participants_by_country <- participants_by_country %>% select(-X)
noc_regions <- noc_regions %>% select(-notes)

# merge to have region
participants <- merge(participants_by_country, noc_regions)
participants <- participants %>% select(-NOC.Factor, -NOC)

write.csv(participants, 'participants.csv')

### Libraries ###
library(gcookbook)
library(ggplot2)
library(scales)
library(tidyverse)
library(lubridate)
library(dplyr)
library(ggmosaic)

### Olympic with Medals ###
olympic_medals <- filter(olympic, Medal != 'NA')

### Female Data by Season ###
femaleSummer <- filter(summer, Sex == 'F')
femaleWinter <- filter(winter, Sex == 'F')
```

```
### Olympic Age and Medals ###
olympic_ageMedals <- filter(olympic, Medal != 'NA' & Age != 'NA')
olympic_ageMedals <-olympic_ageMedals %>% mutate(agegroup = case_when(Age >= 10  & Age <= 30 ~
'Young',
                                                                      Age >= 30  & Age <= 50 ~ 'Middle',
                                                                        Age >= 50  & Age <= 70 ~ 'Old',
                                                                      Age >= 70 ~ 'Very Old'))

### Olympic Data by Season and only Medal Winners###
summer <- filter(olympic_ageMedals, Season == 'Summer' & agegroup != 'Old' & agegroup != 'Very
Old')
winter <- filter(olympic_ageMedals, Season == 'Winter' & agegroup != 'Old' & agegroup != 'Very
Old')

### Mosaic Plot ###
ggplot(data = winter) +
  geom_mosaic(aes(x = product(Sport, agegroup), fill=Sex)) +
  labs(title = 'Mosaic Plot of Winter Sports', x = "Gender and Age", y = "Sport") +
  theme(
        axis.title.y = element_text(size = rel(1.5)),
        axis.title.x = element_text(size = rel(1.5)),
        axis.text.x = element_text(angle = 45, vjust = 1, hjust=1),
        plot.title = element_text(size = rel(2), hjust = 0.5)
  )

ggplot(data = summer) +
  geom_mosaic(aes(x = product(Sport, agegroup), fill=Sex, colour = 'YlGnBu')) +
  labs(title = 'Mosaic Plot of Summer Sports', x = "Gender and Age", y = "Sport") +
  theme(
        axis.title.y = element_text(size = rel(1.5)),
        axis.title.x = element_text(size = rel(1.5)),
        axis.text.x = element_text(angle = 45, vjust = 1, hjust=1),
        plot.title = element_text(size = rel(2), hjust = 0.5)
  )

### Grouping by Region ###
olympic_regions <- data.frame(select(olympic_medals, Medal, region, Year))

r <- olympic_regions %>% group_by(Year, region) %>% summarise(Medals = n())
y <- r %>% group_by(region) %>% summarise(Year = n(), Medals = sum(Medals))

### add a column for yearly average medals ###
y$Avg_Medals <- y$Medals / y$Year

### Export to csv ###
write.csv(y, 'olympic_geo.csv') # the data for mapping
```

Alex Rosenblum - R Code

```
### Exploratory ###
#setwd("~/School/DSC465 Data Visualization/Final Project")
setwd("~/School/Data Vis Final Proj")
library(ggplot2)
library(dplyr)
library(tidyr)

d = read.csv('olympics.csv')
summary(d)
```

```
aths = aggregate(d$Games, by = list(d$Name), FUN = mean)

aths = d[,-c(14,15)] %>% group_by(Name, Year)

aths = pivot_wider(d, names_from = Event, id_cols = c(Name, Sex, Year, Age, Weight, Height,
Sport, Medal))

aths$Era = cut(aths$Year, breaks = c(seq(1896, 2016, 30)))
levels(aths$Era) = c("1896 - 1926","1926 - 1956","1956 - 1986", "1986 - 2016")

aths$BMI = aths$Weight / ((aths$Height / 100)^2)

aths$Sport = as.factor(aths$Sport)

library(ggridges)

# How does BMI distribution change for Men and Women over the years?

aths %>% filter(!is.na(Era)) %>%

  ggplot(aes(y = Era, x = BMI, fill = Sex)) +

  geom_density_ridges() +

  scale_y_discrete(labels = c("1896 - 1926","1926 - 1956","1956 - 1986", "1986 - 2016")) +

  theme_bw() +

  facet_wrap(~ Sex) +

  labs(title = "BMI by Sex and Era")


# How does Age distribution change for Men and Women over the years?

aths %>% filter(!is.na(Era)) %>%
  ggplot(aes(y = Era, x = Age, fill = Sex)) +

  geom_density_ridges() +

  scale_y_discrete(labels = c("1896 - 1926","1926 - 1956","1956 - 1986", "1986 - 2016")) +

  facet_wrap(~ Sex) +

  theme_bw() +

  labs(title = "Age by Sex and Era")


```{r}
#setwd("~/School/DSC465 Data Visualization/Final Project")
setwd("~/School/Data Vis Final Proj")
#setwd("~/Resources")

############################
### Final Visualizations ###
library(ggplot2)
library(dplyr)
library(tidyr)
library(plotly)
```
```

```
library(scales)

d = read.csv('olympics.csv')
```

```{r}
#tiff("top-sports.tiff", units="in", width=10, height=7, res=900)

### Top Sports ###
proportion_events_summer = d %>%

  filter(Season == 'Summer') %>%
  group_by(Year, Sport) %>%
  summarize(n = n()) %>%
  mutate(freq = n / sum(n)) %>%
  mutate(rank = min_rank(desc(freq))) %>%

  arrange(Year, desc(n))

proportion_events_summer %>%
  filter(Sport %in% c('Swimming', 'Gymnastics', 'Athletics', 'Shooting', 'Cycling')) %>%

  ggplot(aes(x = Year, y = freq, color = Sport)) +
  geom_smooth(size = 1, method = 'loess', span = 0.25, alpha = 0) +

  theme_bw() +

  labs(
    title = 'Representation of Top 5 Summer Olympic Sports Over Time',
    subtitle = 'As Determined by Proportion of Participating Athletes',
    y = 'Proportion of Participating Athletes'
  ) +

  theme(
    panel.background = element_rect(fill = '#A0A0A0'),
    legend.key = element_rect(fill = '#B0B0B0'),
    panel.grid.minor = element_blank(),
    panel.grid.major = element_line(color = '#BBBBBB')) +

  scale_color_manual(values = c('blue', 'yellow', 'darkgreen', 'black', 'red'),
                     breaks = c('Swimming', 'Gymnastics', 'Athletics', 'Shooting', 'Cycling'))
+

  scale_y_continuous(
    labels = label_percent(),
    limits = c(0, 0.3))


#dev.off()

```

```{r}
### Athlete's Ages ###
aths_ages = d %>%
  mutate(Year.Factor = as.factor(Year)) %>%
  group_by(Year, Year.Factor, Age) %>%
  summarize(n = n()) %>%
```

```
  mutate(freq = as.numeric(format(n / sum(n), scientific = FALSE, digits = 2))) %>%
  filter(Age <= 70)

#tiff("age_over_time.tiff", units="in", width=10, height=7, res=1200)

aths_ages %>%
  ggplot(aes(x = Year, y = Age, fill = freq)) +
  geom_tile() +
  geom_smooth(color = 'red', method = 'loess') +
  scale_fill_viridis_c(labels = label_percent()) +
  scale_x_binned(breaks = seq(1896, 2020, 4)) +
  theme_bw() +

  theme(
    axis.text.x = element_text(angle = 90, vjust = 0)
  ) +

  labs(
    title = "Relative Frequency of Athletes' Ages Over Time",
    subtitle = 'With Moving Average Overlaid',
    fill = "Percentage"
   )


#dev.off()
```