Group Name: Grid Lines

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

Project Deliverable 2

Part I: Group Meeting

A) Discussed about the dataset: This is a dataset on the Olympic Game. The file athlete_events.csv contains 271116 rows and 15 columns. Each row corresponds to an individual athlete competing in an individual Olympic event.

   The columns 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
   - NOC - National Olympic Committee 3-letter code
   - Games - Year and season
   - Year - Year
   - Season - Summer or Winter
   - City - Host city
   - Sport - Sport
   - Event - Event
   - Medal - Gold, Silver, Bronze, or NA

B) Group Liaison: Elizabeth Nieto

C) Distributions among subgroups:

   We will be dividing the dataset based on season of the game such as Winter and Summer Games. For better understanding of the overall dataset through visualizations. Following are lists of visualizations which are suggested by the team:

   - Male vs. female ratio on sports
   - Number of medals Won by athletes
   - Average age of the athletes
   - Host cities and total medals earned
   - Performance of specific players
   - Time series of the number of medals won by each country and the number of participating athletes.
   - 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)?

Key Insights

a. Summer Olympics attract more than 4 times athlete participation than Winter Olympics
b. Growth in the number of athletes in Summer Olympics have become stagnant in the recent games
c. 34 sports competitions is the maximum in a summer olympic game
d. Number of events have come to a saturation point in the last 5 summer games
e. Number of athletes in Winter Olympics keeps growing
f. 1916, 1940 & 1944 summer Olympic games and 1940 & 1944 winter Olympic games didn't happen due to the world wars. Neither of the world war breaks affected the athlete participation in the succeeding years.
g. Only the first edition of Olympics didn't have any female athletes. There has been a great improvement in female representation ever since. Late 1950s and late 1980s saw significant increase in female representation.
h. In Winter Olympic games, the average age of medal winners is mostly higher than the non-medal winners.

D) Previous Work:

a. Predict number of medals a country will win. What factors are important?

Link: https://medium.com/predictive-analytics-with-olympics-data-set/executive-summary-2f9796617a3a

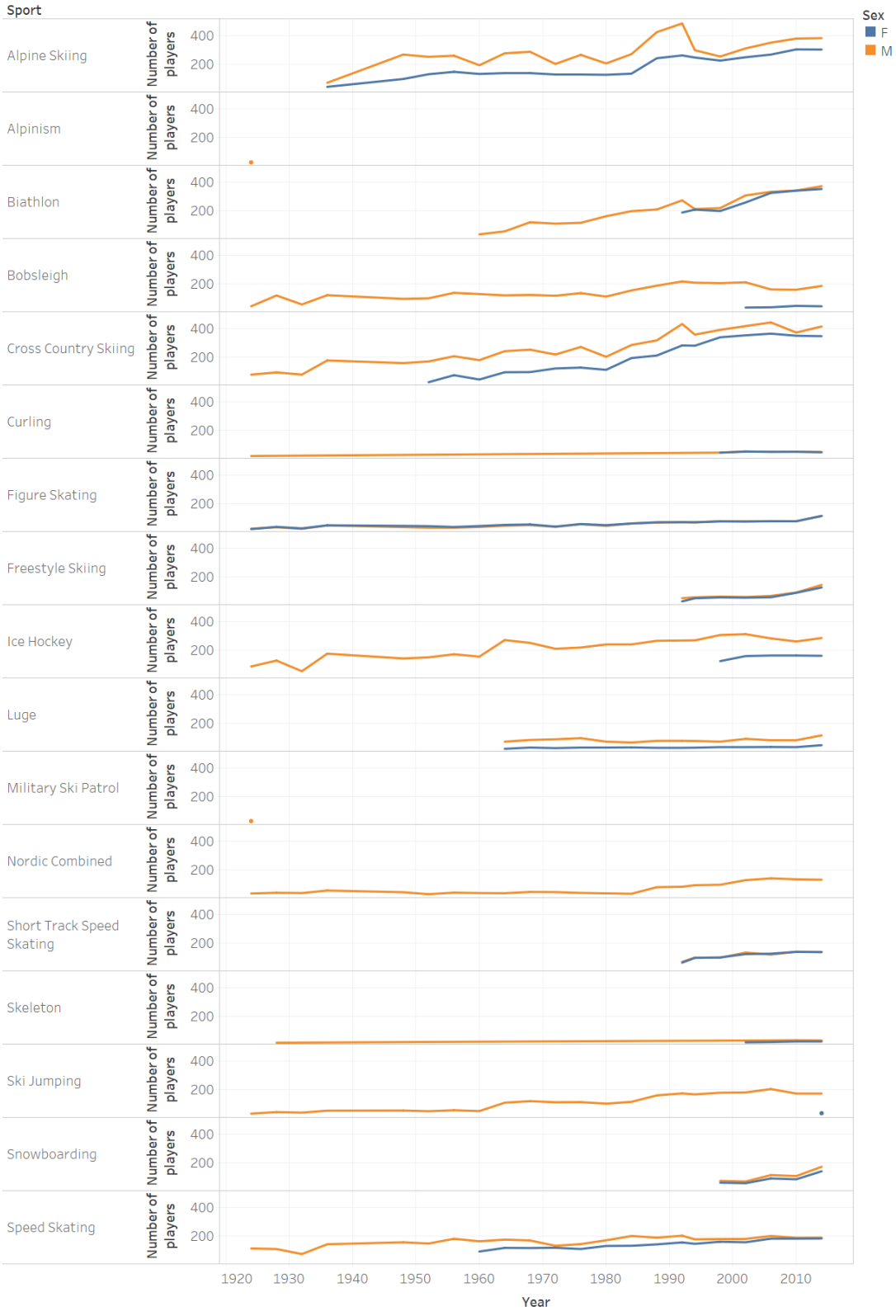
Part II: In-depth Visualization

Subgroup 1: Winter Games

Winter Game: Basic Exploratory Visualization by Gender

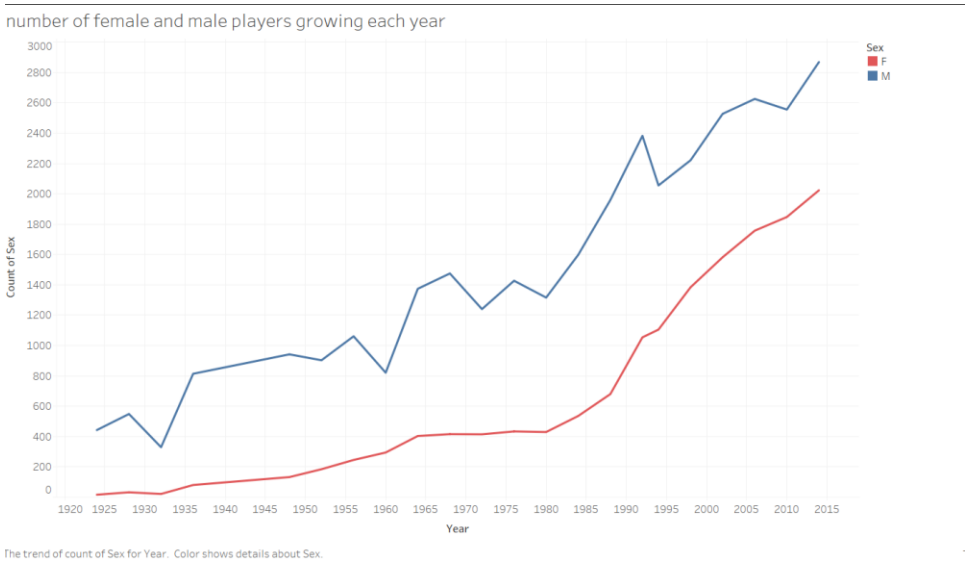
The graph below shows the number of players over the years for each sport played in winter based on genders.

Sheet 1

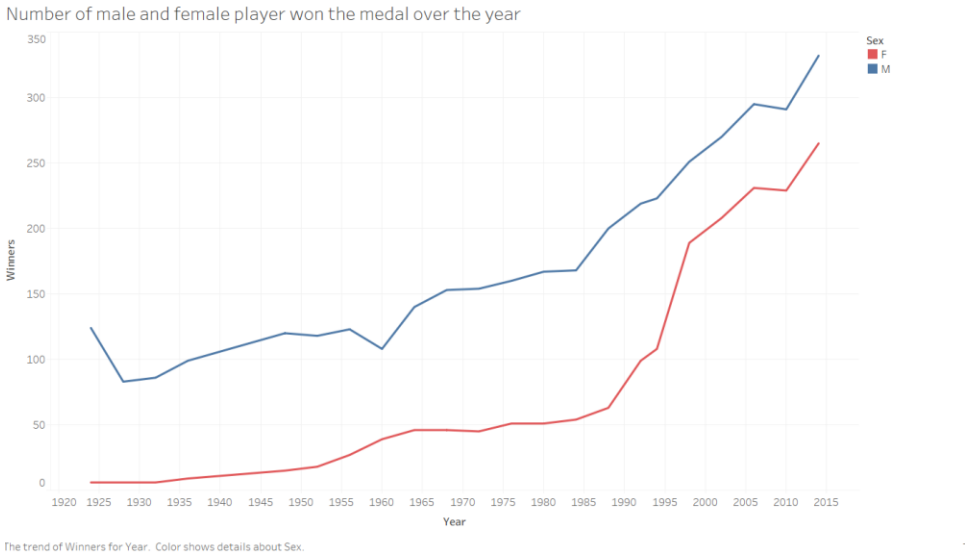


The trend of count of ID for Year broken down by Sport. Color shows details about Sex. The data is filtered on Season, which keeps Winter.

The below graph shows the overall increase in the number of male and female players in the Olympics game over the given period of year.

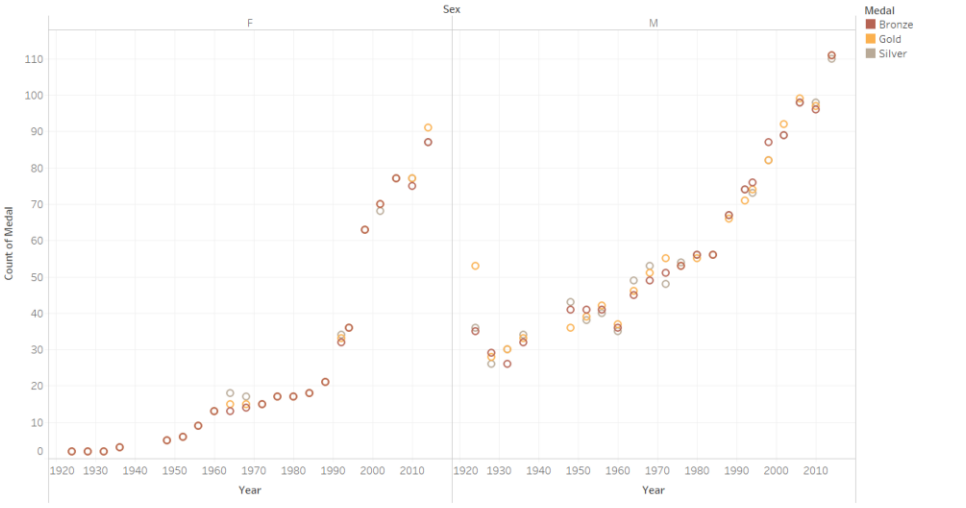


The below graph shows the number of male and female players won (Medal= Gold, Silver and bronze) in the Olympics game over the given period of year.



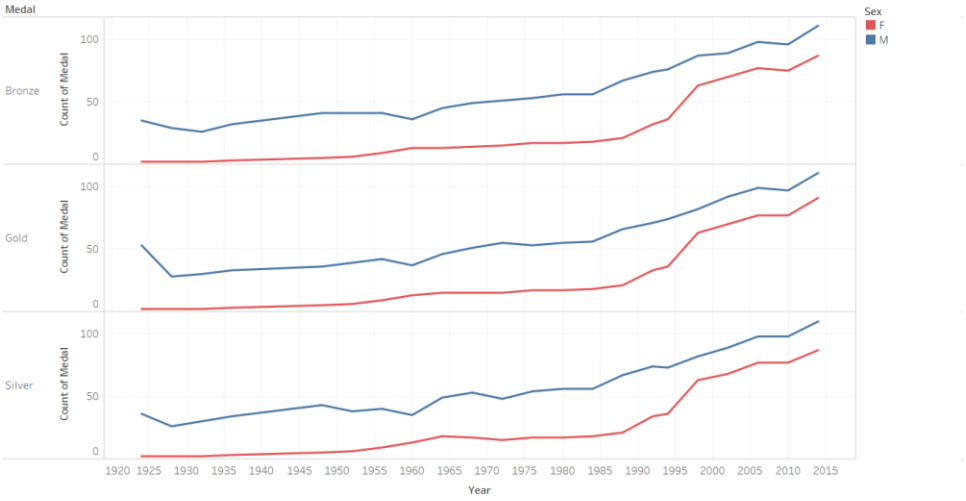
The below graph shows the number of male and female players won (Medal = Gold, Silver and bronze) in the Olympics game over the given period of year.

Male Vs Female wining medals in Olympic over the year



The plot of count of Medal for Year broken down by Sex. Color shows details about Medal. The view is filtered on Medal, which keeps Bronze, Gold and Silver.

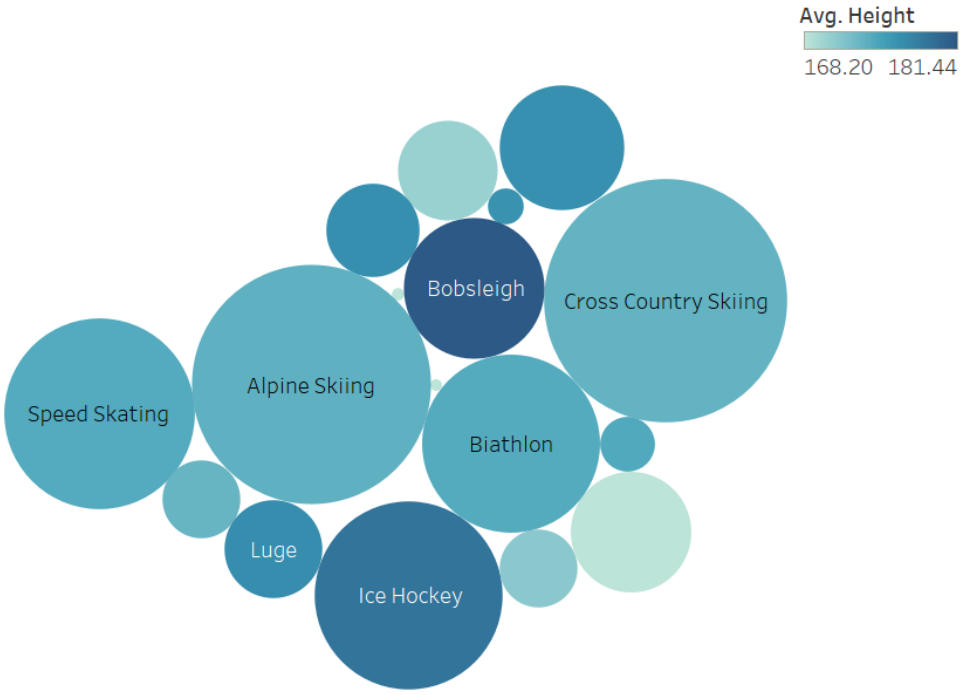
Male Vs Female wining medals in Olympic over the year



The trend of count of Medal for Year broken down by Medal. Color shows details about Sex. The view is filtered on Medal, which keeps Bronze, Gold and Silver.

The graph below shows the games played in winter, the size of the bubbles in the chart is adjusted based on the number of players for the respective sport and the color intensity is adjusted by the height of the players.

Sheet 1



Sport. Color shows average of Height. Size shows count of ID. The marks are labeled by Sport. The data is filtered on Season, which keeps Winter.

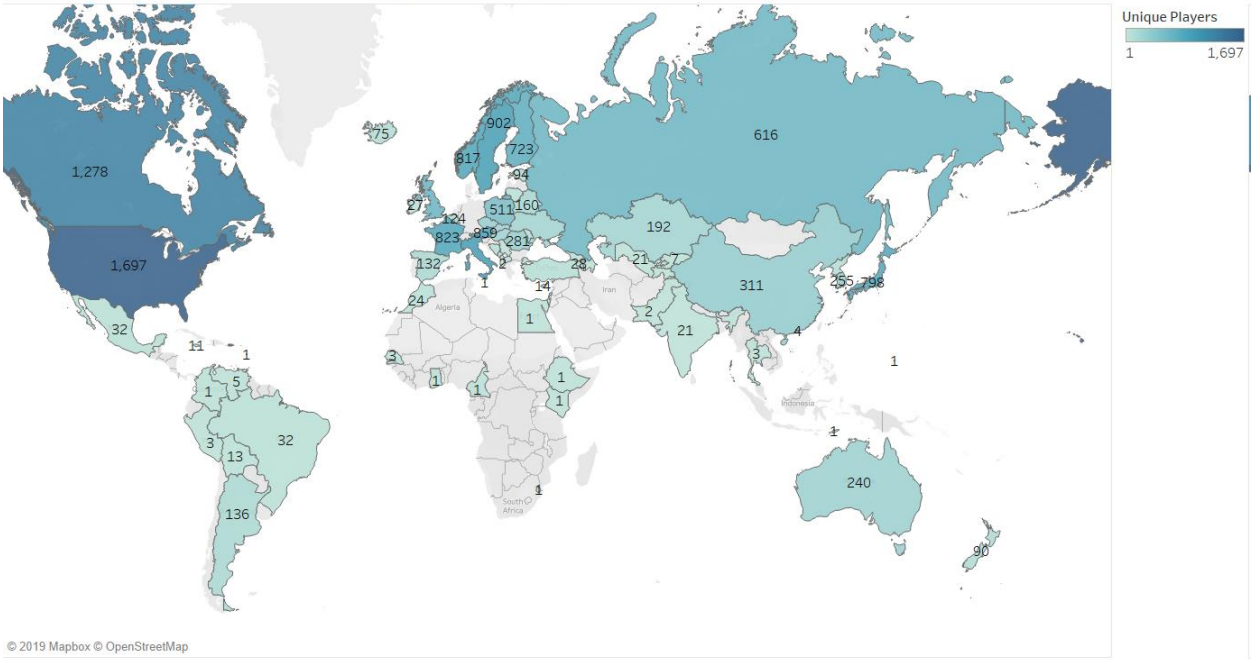
The graph below shows the height and weight data for players from sports played in winter. The color is adjusted based on the sex of the players.

Sheet 1

Weight vs. Height. Color shows details about Sex. The data is filtered on Season, which keeps Winter.

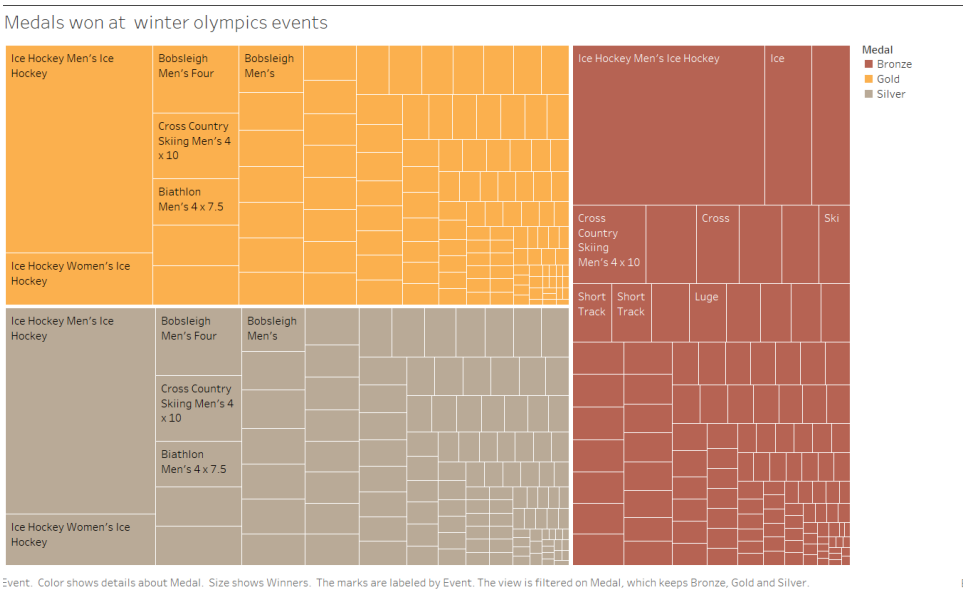
The below graph shows the countries which are participating in the Olympics game over the given period of year.

Nation with highest number of participation in olympics

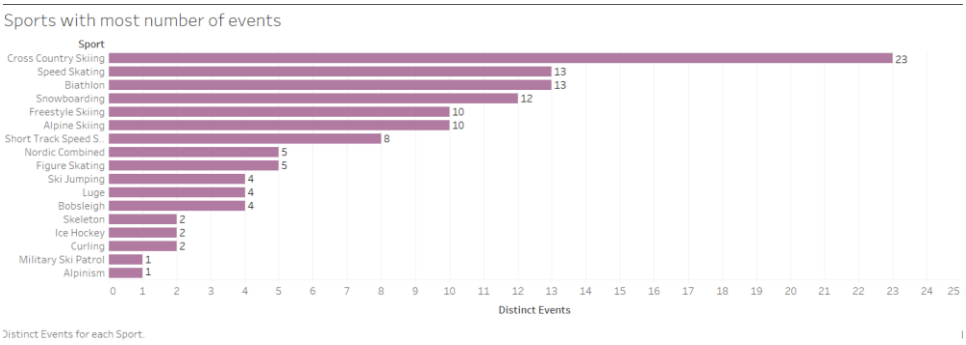


Map based on Longitude (generated) and Latitude (generated). Color shows Unique Players. The marks are labeled by Unique Players. Details are shown for NOC.

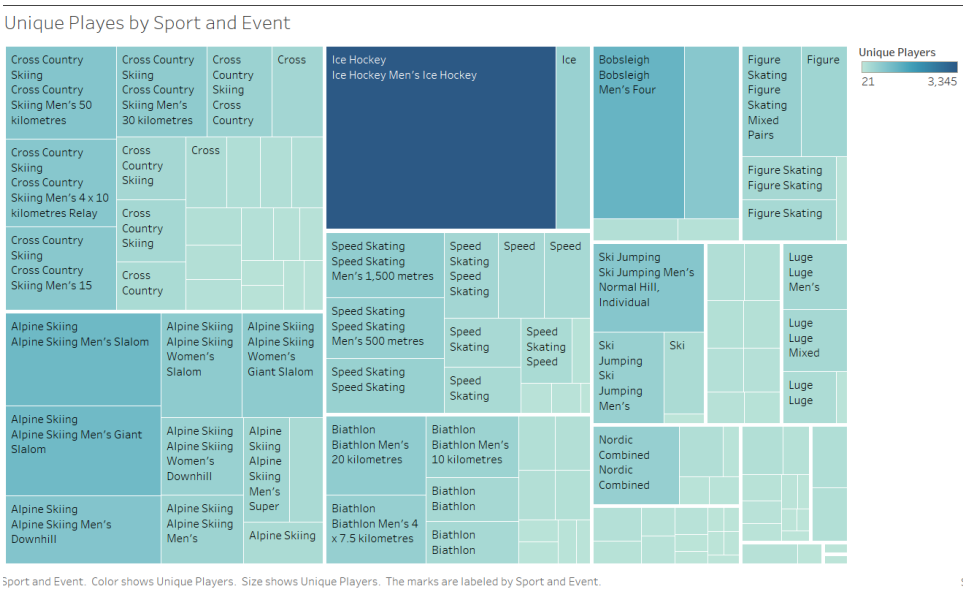
The below graph shows the most played game at winter Olympics and won the maximum number of medals.



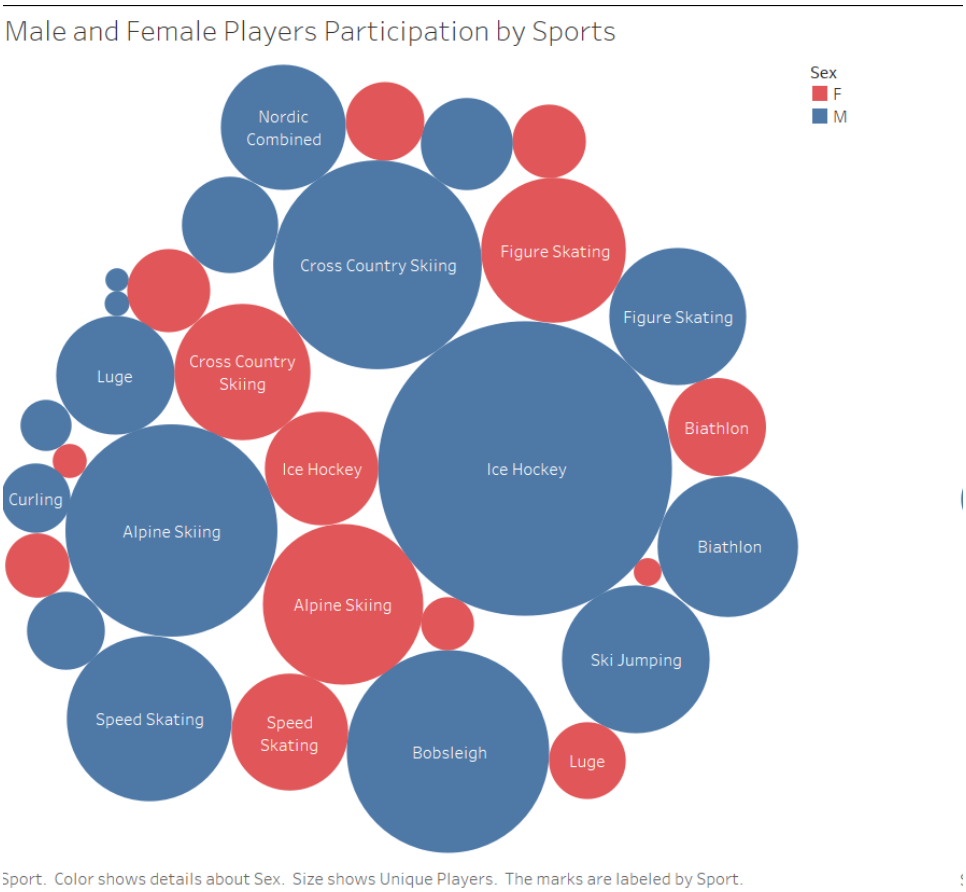
The below graph shows histogram of sports which are having a maximum number of winter games.



The heat map is showing the split of unique players by sports and games they play.



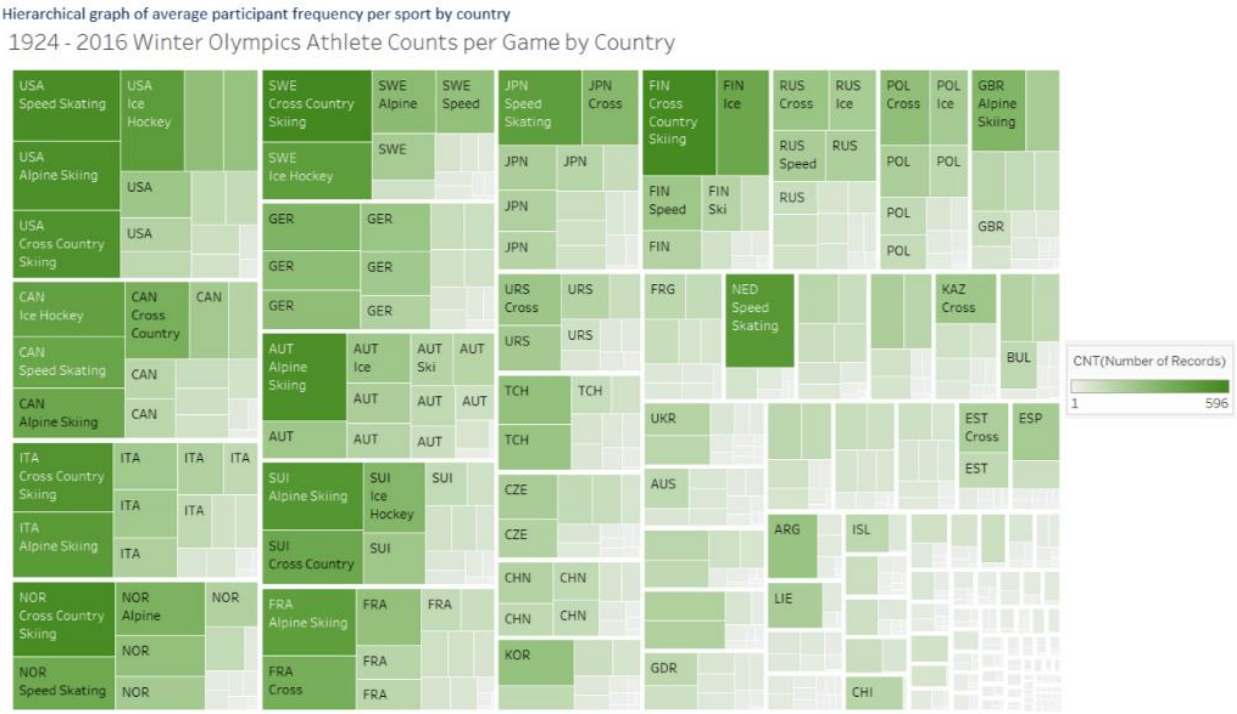
The bubble chart shows the famous sports among male and female athletes.



Elizabeth Nieto – Winter Graphs

Graph Description: The three graphs below examine the overall average number of participants for winter games across all years split up by country and sport, another split up only by country and the last showing the games played at each winter Olympics.

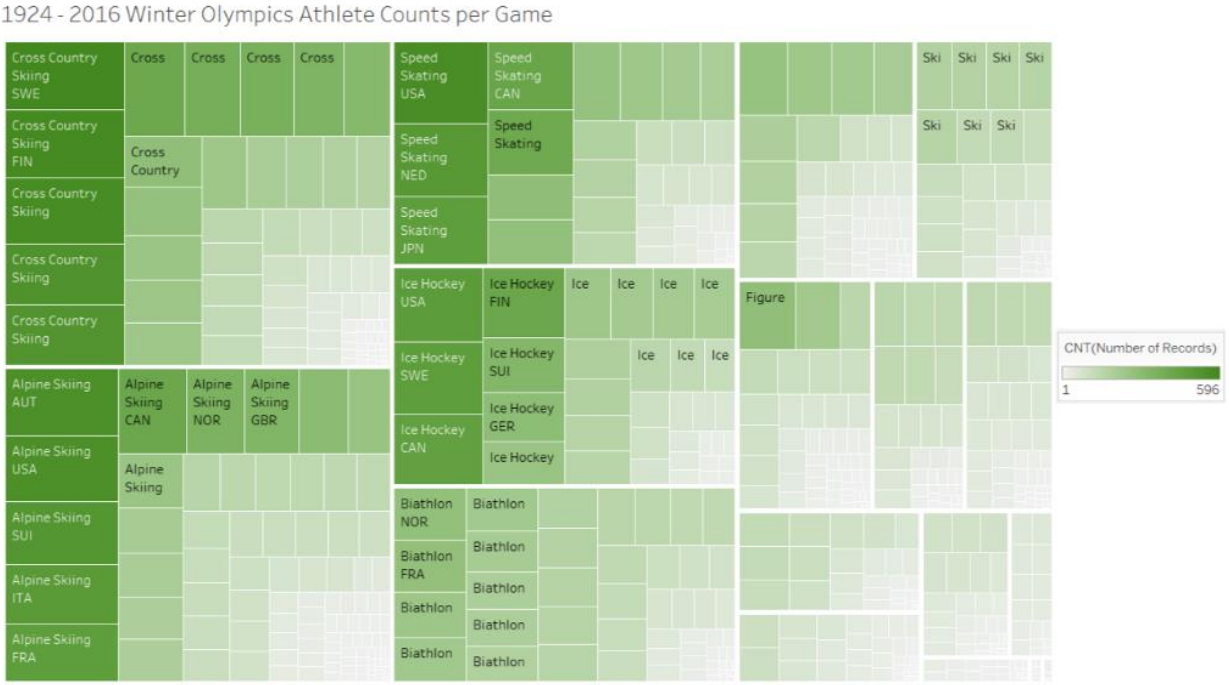
Hierarchical graph of average participant frequency per sport by country



Steps: To create this graph on tableau I added country and sport variables to the columns section and added the number of records to be encoded as color and size of the rectangles. All summer seasons games were filtered out.

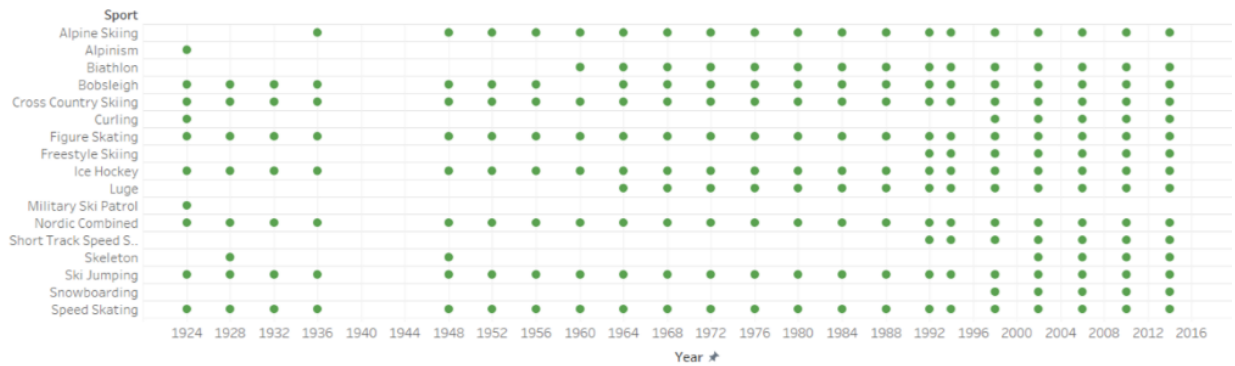
Description: On average, the US, Canada, Norway, Sweden, Germany, France, Italy, Japan, Finland, Netherlands, Australia have sent significantly more athletes per game and overall to winter Olympics from 1896-2016.

Hierarchical of per sport what country send the most players



Steps: To create this graph on tableau I added sport then country variables to the columns section and added the number of records to be encoded as color and size of the rectangles. All summer seasons games were filtered out.
Description: Cross Country and Alpinism feat(not an actual game, will be removed for subsequent visualizations) skiing has seen the most participants over time. Military Ski Control(Developed into Biathalon) and Skeleton Racing.

Winter Games played by year



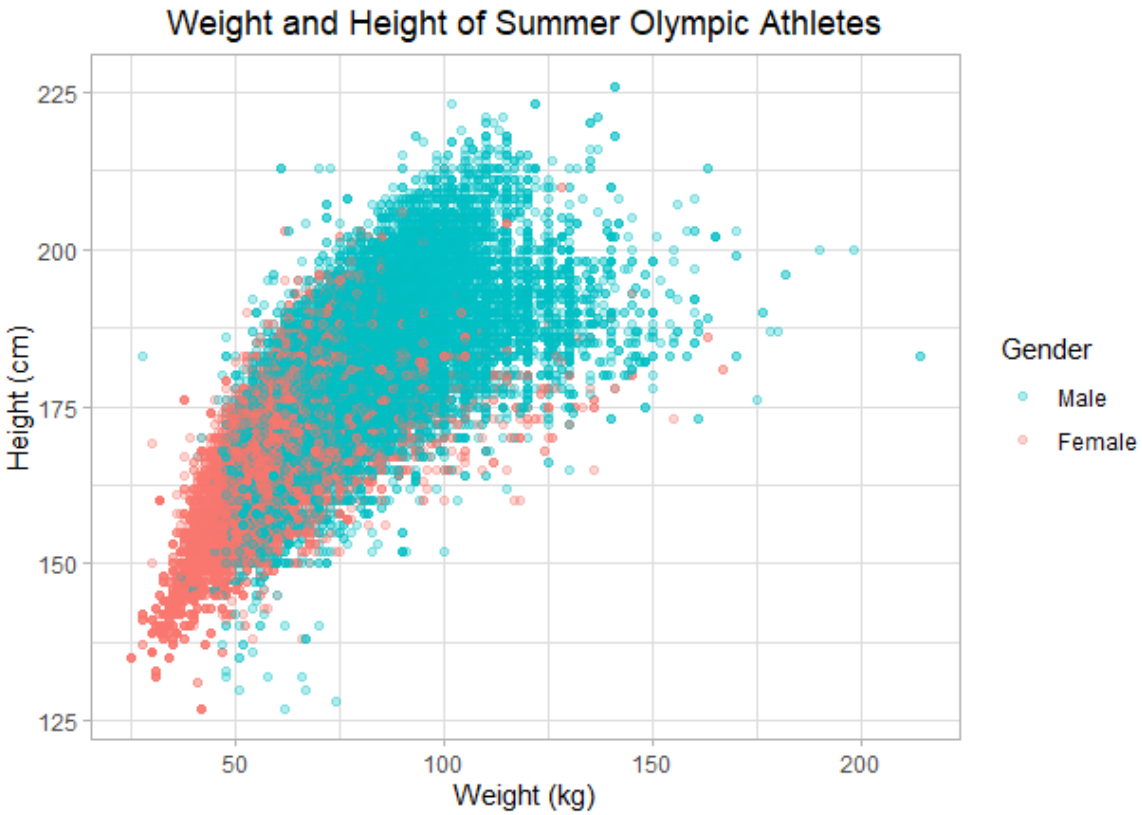
Steps: To create this graph the yea variable was added to the column and the sport added to the rows. Summer games were filtered out.
Description: There exist data for 17 winter sports. This graphics shows what sports were played at each Olympics. Some games like Alpinism was only awarded once at the 1924 inaugural winter games and not actually a competition. Rather an award for an the most impressive mountaineering accomplishment of the time.
Military Ski Patrol was another sport only present for the inaugural games. In 1960 is was reintroduced as the Biathalon and played at every subsequent game. Lunge was introduced at the following games in 1964 and Freestyle Skiing, Short Track Speed Skiing were both introduced in 1992.

Next Steps: Consider removing alpinism, merging Military Skii Control and Biathalon. Explore athletes participating in more than one sport and their medal count.

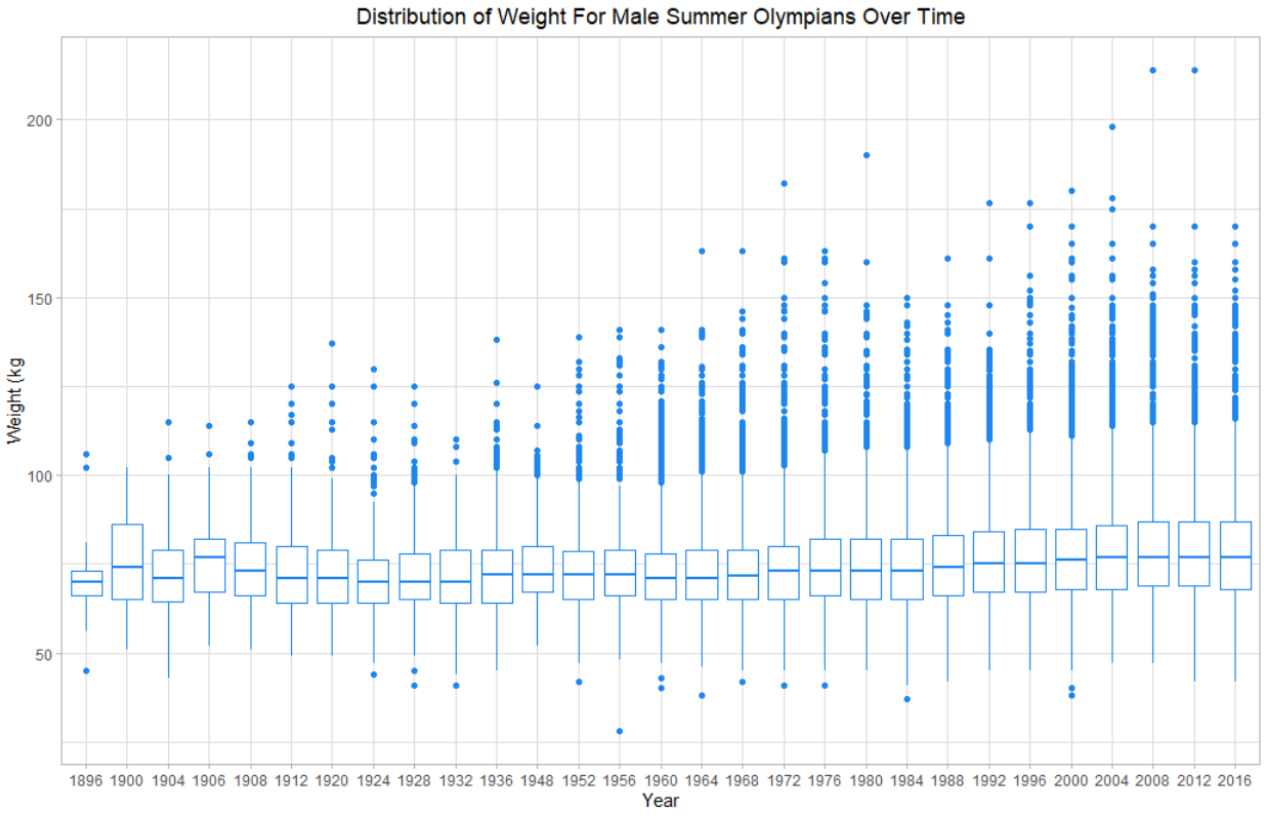
Subgroup 2: Summer Games

Sean Sullivan – Graphs

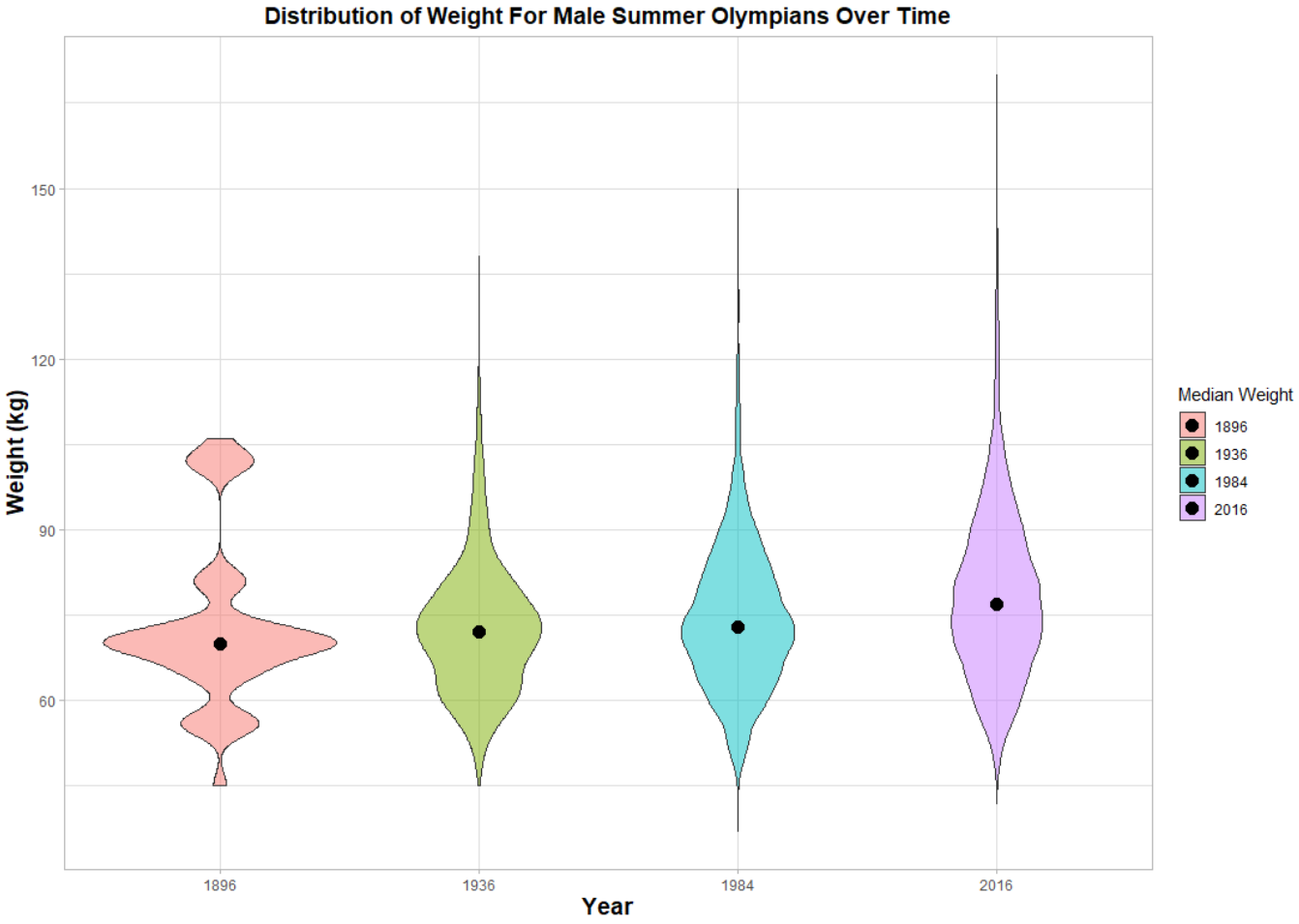
- I created five exploratory graphs for looking into the Weight and Height data.
    - Each graph is only for Summer Games data



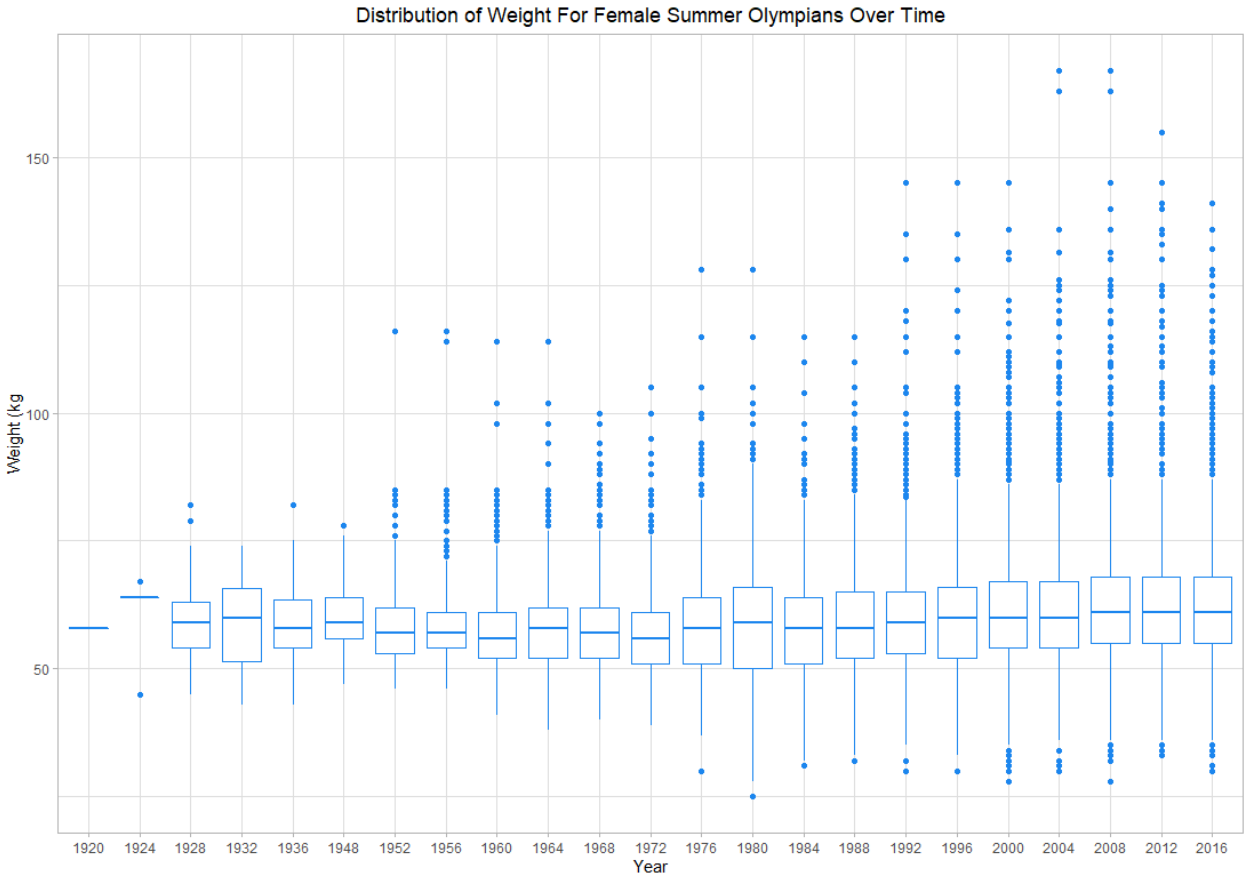
- I set a low alpha to really show where high concentrations of Weight/Height combinations are located. Obviously, you can see the split by gender quite easily.



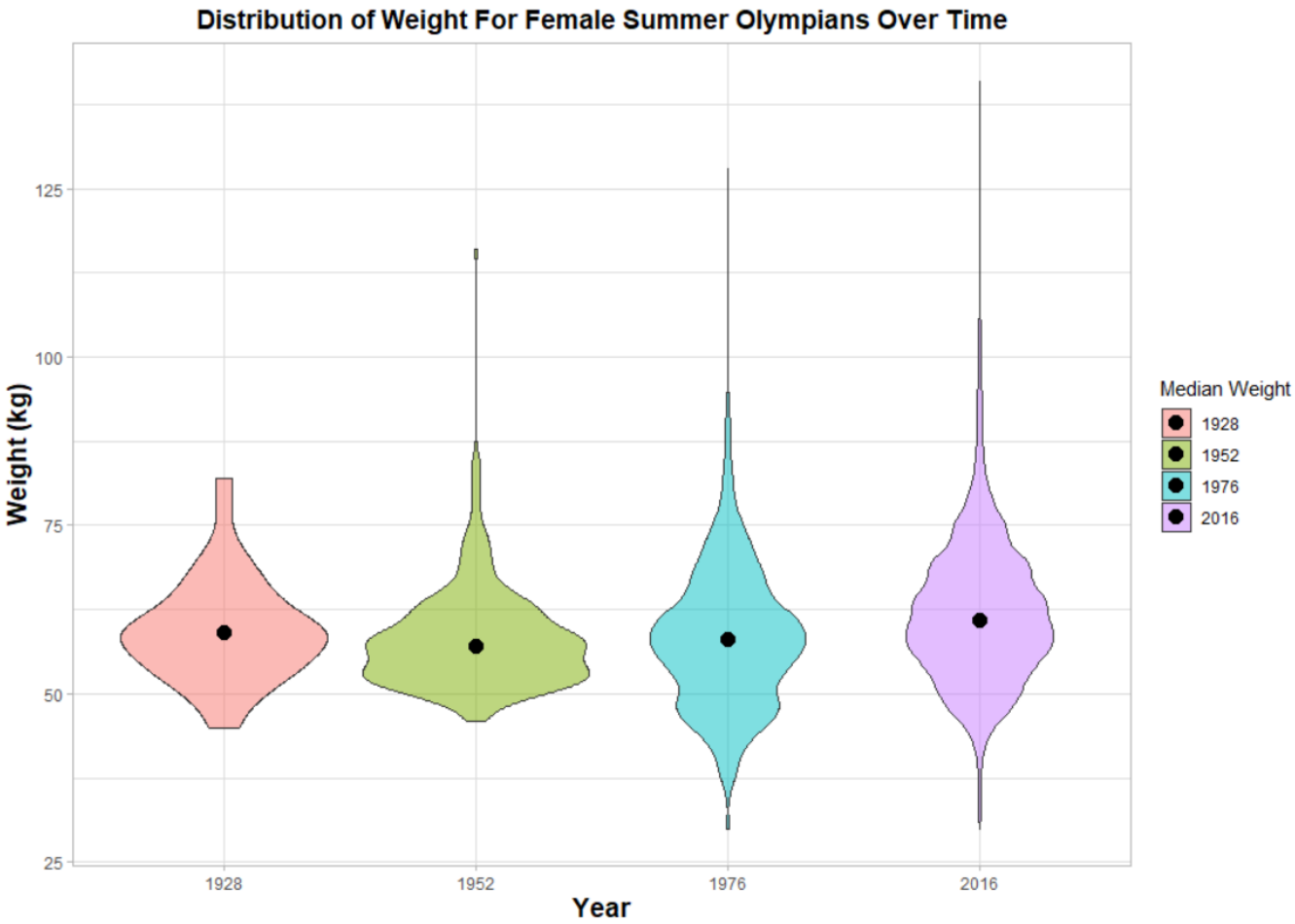
- General increasing trend in median weight...but lots of data...let's pick out four years to show trend (both extremes and two in the middle)



- Violin plot shows distribution of weights of male athletes for these four years. Chose 1896 and 2016 because they are oldest and most recent years in dataset. Chose two other years that fall roughly ½ and ¾ of the way through the set. Black dots represent median value for weight. You see that as time approaches modern day, outliers of weight increase. Will look into what sports began to draw heavier athletes as time went on.



- 1920 and 1924 did not have many female participants that had weights listed – hence crap data... I will adjust for cleaner version if we want to include this data in final presentation/report.



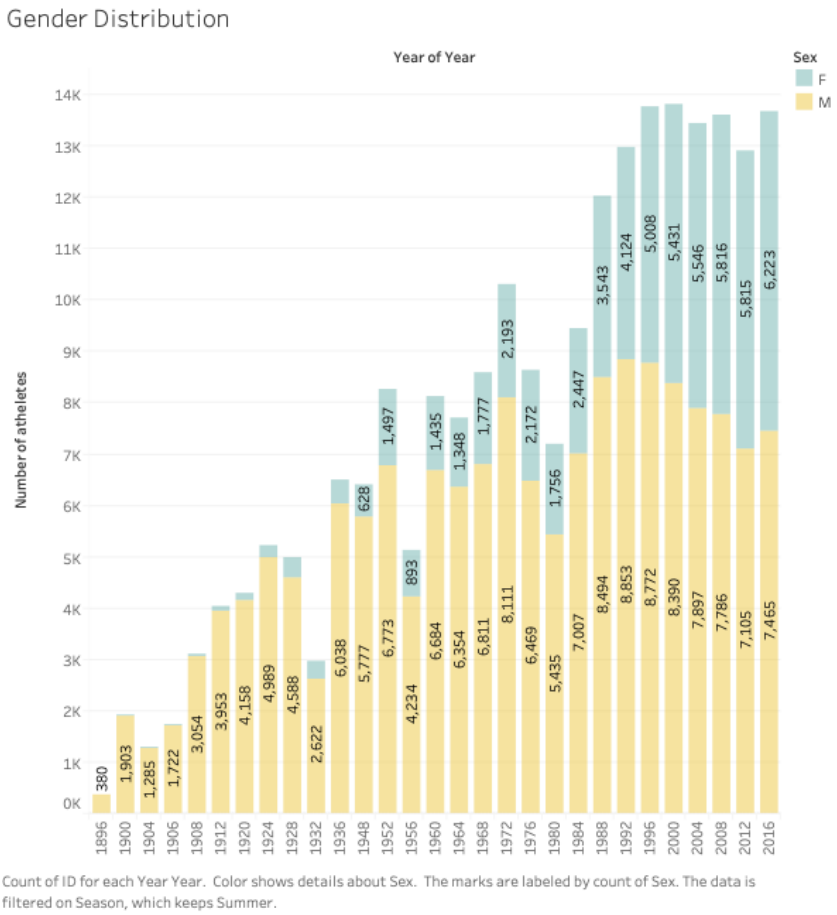
- Median weight slightly increases over time. Outliers become more extreme over time.

My next steps:

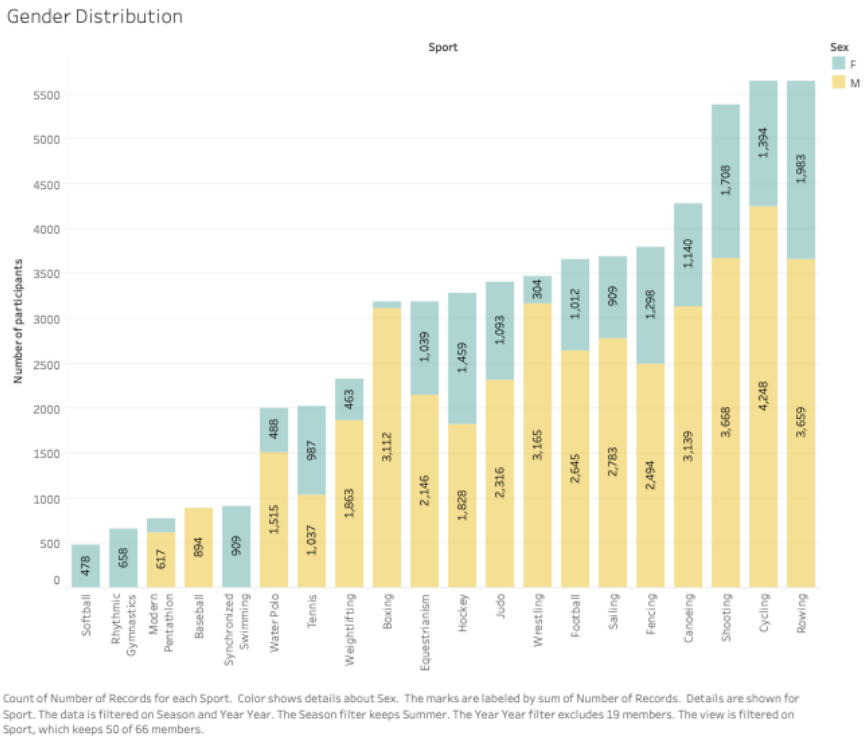
- Create same visuals but for height.
    - Boxplots for overall trend and violin for segmented approach
- Explore what sports have caused increase in H/W over time

Summer game—Chloe Wang

Stacked bar graph showing the gender distribution of numbers of athletes in each Olympics

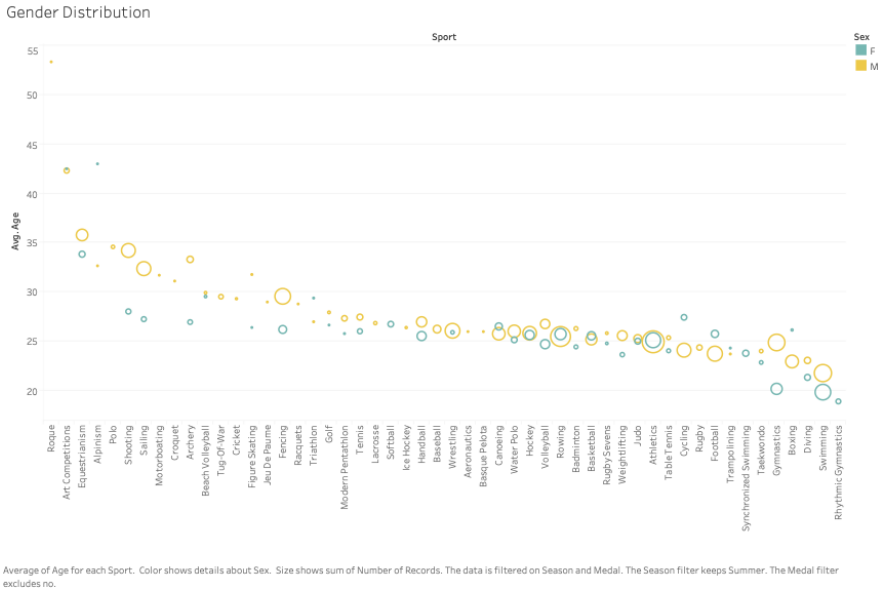


Stacked bar graph showing sports that have unbalanced gender distribution



Count of Number of Records for each Sport. Color shows details about Sex. The marks are labeled by sum of Number of Records. Details are shown for Sport. The data is filtered on Season and Year Year. The Season filter keeps Summer. The Year Year filter excludes 19 members. The view is filtered on Sport, which keeps 50 of 66 members.

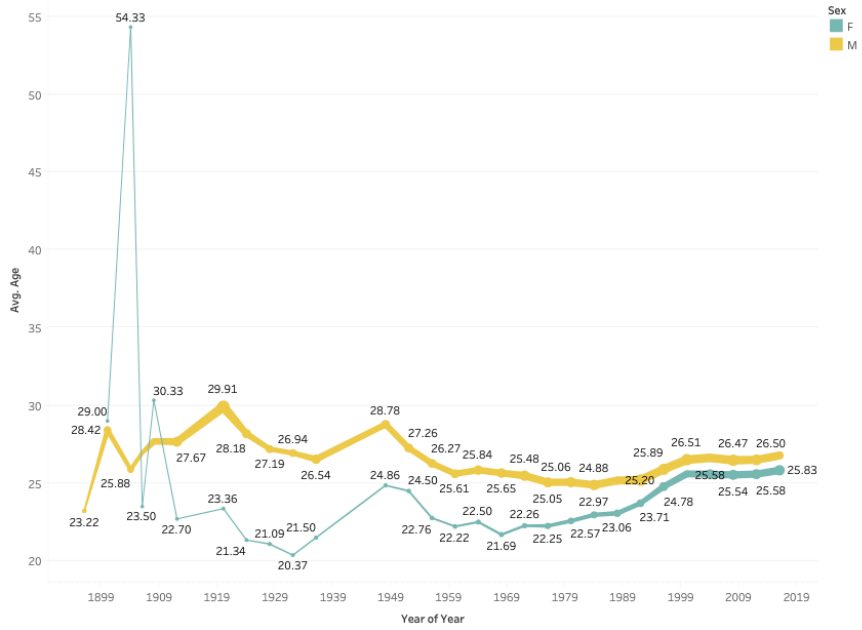
Average age of winning a medal per sport; size of circles represents a total number of athletes for that game.



Average of Age for each Sport. Color shows details about Sex. Size shows sum of Number of Records. The data is filtered on Season and Medal. The Season filter keeps Summer. The Medal filter excludes no.

Average age of winning a medal in 120 years. Line thickness represents the total number of athletes in the year.

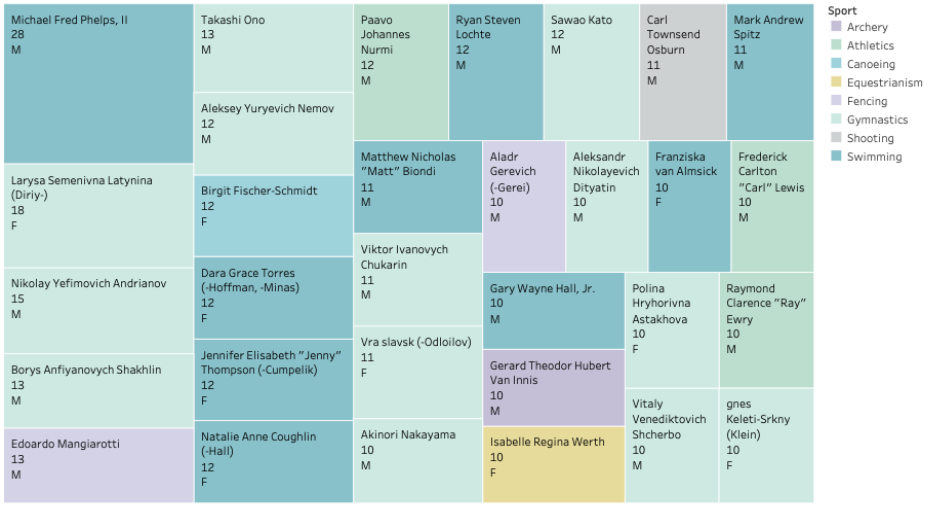
Gender Distribution



The trend of average of Age for Year Year. Color shows details about Sex. Size shows count of Sport. The marks are labeled by average of Age. The data is filtered on Season and Medal. The Season filter keeps Summer. The Medal filter excludes no.

Athletes who won more than ten medals. Colors represent sports

Top Winners



Name, sum of Number of Records and Sex. Color shows details about Sport. Size shows sum of Number of Records. The marks are labeled by Name, sum of Number of Records and Sex. The data is filtered on Season and Medal. The Season filter keeps Summer. The Medal filter keeps Bronze, Gold and Silver. The view is filtered on sum of Number of Records, which ranges from 10 to 28.

Next step: to explore variable country's relationships with other variables