Introduction:

The following case presents insights from data collected on the Olympics over time. Visualizations were created to show insights surrounding 2 distinct tasks; showing country participation and performance across time, sex, and seasons, and how player body types may impact their performance in different sports.

The data utilized includes attributes of specific athletes such as their name, age, height, and weights. It also includes attributes describing their country, their event, the season, year, and location of the events.

When first looking at the data, some major issues could be found. Specifically, there were many instances of data being very sparse. Especially in the player height and weight columns, many of the heights and weights were found to be missing. In visualizations where player height and weight was required, these records were dropped. This unfortunately might skew the results, as the players with no data might be higher or lower than the average for unknown reasons. For visualizations that didn't focus on player data, there was no need to remove these rows. In the 'Medals' column, null values indicated that the player didn't win any trophies. The null values were replaced with the value 'no_medal'. A breakdown of the different features and their meaning can be seen in Appendix A

Visualizations:

Visualization 1: Country Participation Over Time

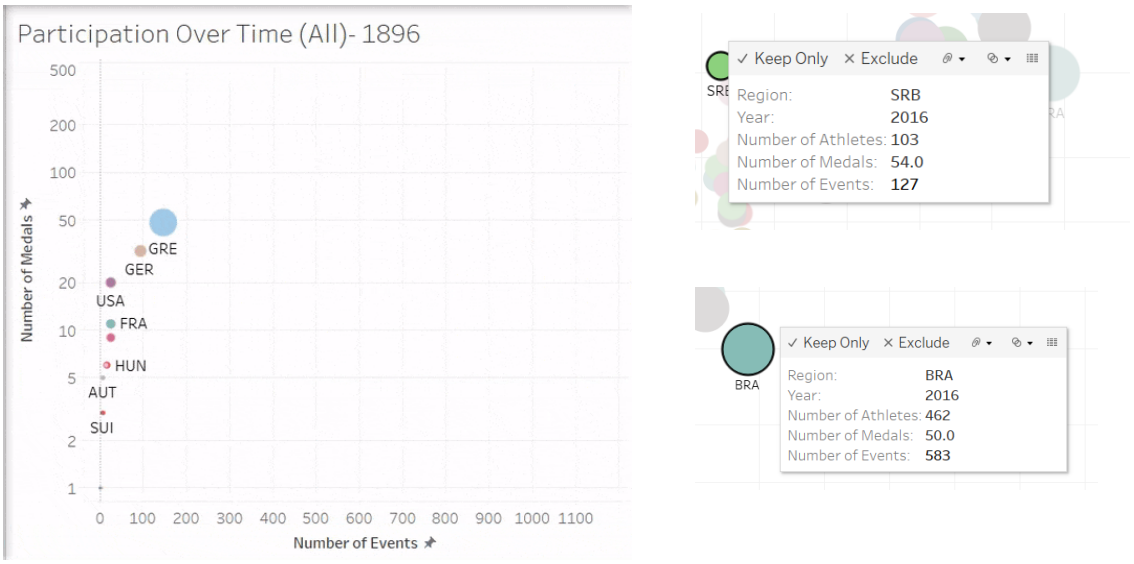


Figure 1: Animation of country participation over time (Link)

Visual 1 provides an overview of country participation over time. In order to display participation metrics cleanly and concisely, an animated bubble graph was developed. With

this approach, the bubbles represented the countries, with their size reflecting the number of athletes participating. The bubbles were then displayed on a coordinate axis to show the number of medals and events over time. The user can statically flip through each year, or play the progression automatically. The animation highlights a few interesting patterns and trends. For instance, for most of the earlier Olympics there was a clear frontrunner in both participation and number of medals awarded. In the later years, this discrepancy begins to even out as circle sizes and location become more homogenous. Another trend is the increase in participation from more countries with a significant range in athlete participation. The visual displays an increase in the number of events, and the number of athletes. Pausing the visual and looking at certain years also provides interesting insights. For example in the 2016 Summer Olympics, Serbia and Brazil were rewarded almost the same number of medals. However, Brazil participated in almost 400 more events and had nearly three times as many athletes. Another interesting outlier the visual caught was the United States not participating in the 1980 Olympics. These insights are important because they give a historical perspective on current participation. For example, in 1980 the United States boycotted the Olympics, a historical event displayed in the data. The Olympics has been around for 100s of years, seeing these patterns over time tell a story of how a world comes together. Future analysis can further explore these underdogs like Serbia, how were they able to compete with much larger countries like Brazil.

Visualization 2: Summer vs Winter Olympic Games

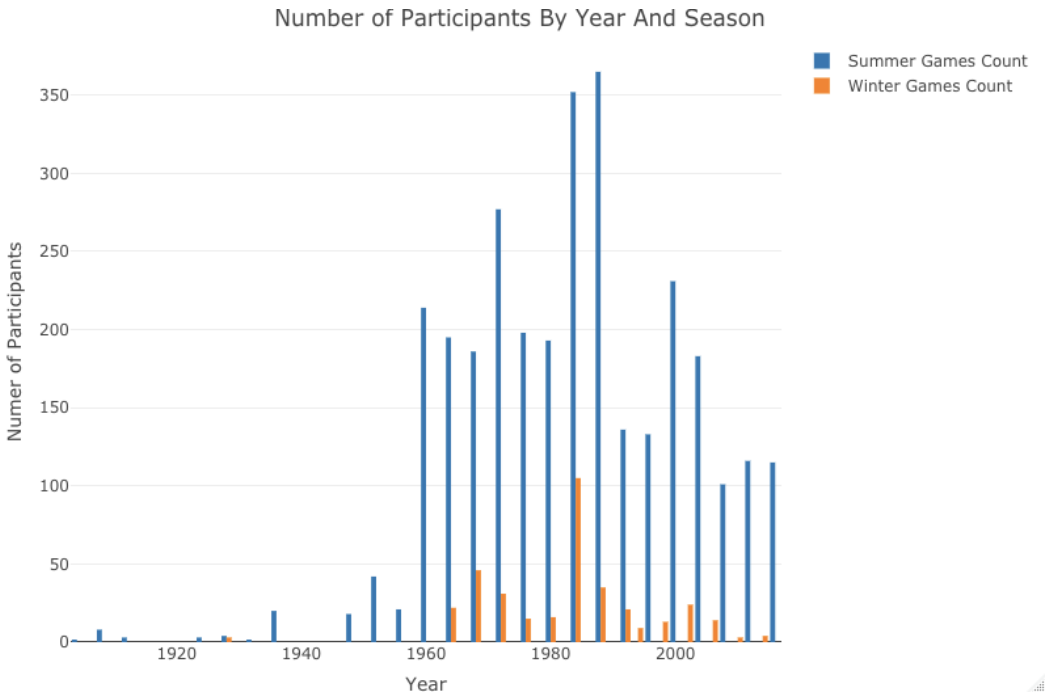


Figure 2: Bar Chart of Winter vs Summer Olympic Games

Visual 2 displays a bar chart of Winter Olympic games over Summer Olympic Games. This approach aggregated total participation over the years grouped by the different seasons and displayed on the graph. This visual clearly displays the disparities between participation in the summer and winter games. The Summer games by far has many more

participants. But over time, the amount of participants in the winter Olympics seems to slowly increase until it peaks in 1984, and then it declines. Further analysis can explore trends in the winter games and perhaps audience viewing patterns to try to understand why the winter games are not nearly as popular, and why it has been declining in recent years. Insights into participation grouped by season could be important decision factors to certain stakeholders. For example, sponsors may not want to spend their money on the winter season if participation is low.

Visualization 3: Player BMI Impact on Performance

In order to understand what body types are most valuable to different sports, player Body Mass Index (BMI) was calculated for those who had height and weight data included. BMI is calculated by using the athlete's weight (kg) divided by the square of height (meters converted from centimeters), and is then dropped into one of four different categories (the international standard for adults). If an athlete's body mass index is below 18.5, he/she will be defined as underweight, if body mass index is between 18.5 and 25, defined as normal/healthy, if body mass index is between 25 and 30, he/she is defined as Overweight, if above 30, he/she is defined as obese. A BMI dummy variable is created based on this data.

Once the BMI is calculated, the athletes are grouped together in a multitude of ways. They are firstly grouped together based on their BMI. Within these groups, the proportion of athletes in each BMI group are found for different countries, different sports, and different medals.

Based on above data preprocessing, two subsets of table are generated that are the proportion of athlete body mass grouped by region and the proportion of athlete body mass grouped by sport, both of them provide many interesting insights about the Olympic History data.

The following tables are the three countries with the most athletes in each BMI category. The proportion is calculated as the amount of athletes in a country divided by the total athletes of that respective BMI.

region	body_mass_Underweight	Count	proportion
Ethiopia	109.0	352	0.309659
Burundi	10.0	38	0.263158
Namibia	16.0	71	0.225352

region	body_mass_Normal	Count	proportion
Timor-Leste	6.0	6	1.000000
South Sudan	3.0	3	1.000000
Saint Vincent	22.0	23	0.956522

region	body_mass_Overweight	Count	proportion
Montenegro	40.0	94	0.425532
Micronesia	10.0	25	0.400000
Tonga	14.0	36	0.388889

region	body_mass_Obese	Count	proportion
Nauru	6.0	11	0.545455
American Samoa	8.0	21	0.380952
Kiribati	3.0	11	0.272727

EDA Insights:

The above table shows a strong relationship between the location of the athlete country and the body mass of the athlete. Among all the countries, Ethiopia, Burundi, and Namibia have the most number of slim athletes. All of the top five countries of the underweight table are located in Africa, and are developing countries. Most of athletics in Nauru, Micronesia,

American Samoa, Kiribati, Tonga are more like to be overweight or obese. All of those countries are Pacific Islands.

The following tables are the top three rows selected from the proportion of athlete's BMI categories grouped by sport based on four different levels of body mass.

Sport	body_mass_Underweight	Count	proportion
Rhythmic Gymnastics	526.0	615	0.855285
Synchronized Swimming	129.0	849	0.151943
Figure Skating	217.0	1512	0.143519

Sport	body_mass_Normal	Count	proportion
Motorboating	1.0	1	1.000000
Lacrosse	2.0	2	1.000000
Nordic Combined	1047.0	1066	0.982176

Sport	body_mass_Overweight	Count	proportion
Bobsleigh	1527.0	2206	0.692203
Rugby	19.0	30	0.633333
Baseball	481.0	846	0.568558

Sport	body_mass_Obese	Count	proportion
Weightlifting	823.0	2995	0.274791
Tug-Of-War	5.0	22	0.227273
Judo	456.0	3382	0.134831

The above tables show there is a huge difference in athlete body mass between different sports. Over 85 percent of players in rhythmic gymnastics are underweight while the second sport is only 15 percent , which means if a person wants to become slim, rhythmic gymnastics would be the best choice. On the contrary, if that person wants to stay healthy and normal, motorboating and lacrosse are recommended, because 100 percent of the athletes in those two sports have the normal or healthy physique. Anyway, if a bigger physique is the goal, bobsleigh, rugby, and weightlifting would be the right decision.

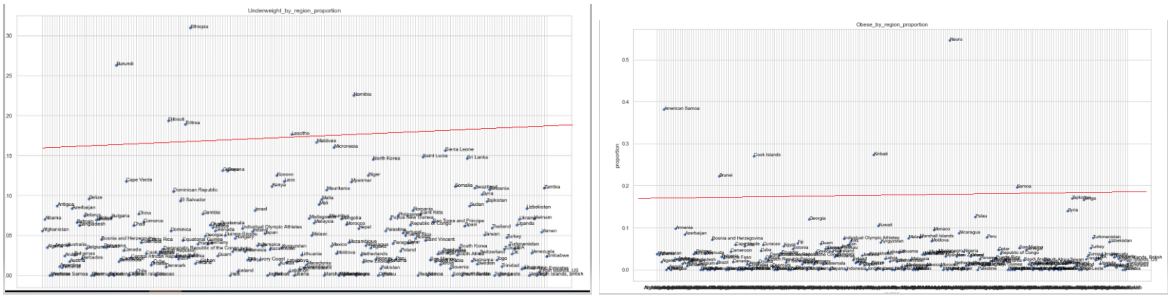


Figure 3: Scatterplot of the proportions of BMI groups by region. The plot on the left is for underweight athletes, and the plot on the right is for overweight athletes.

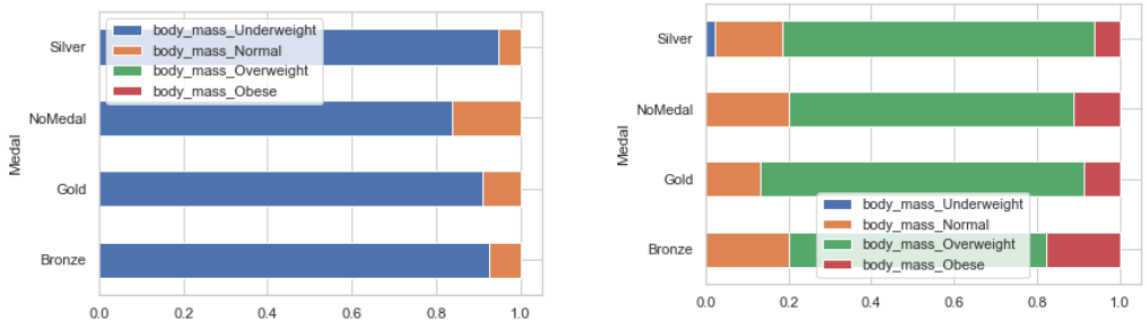


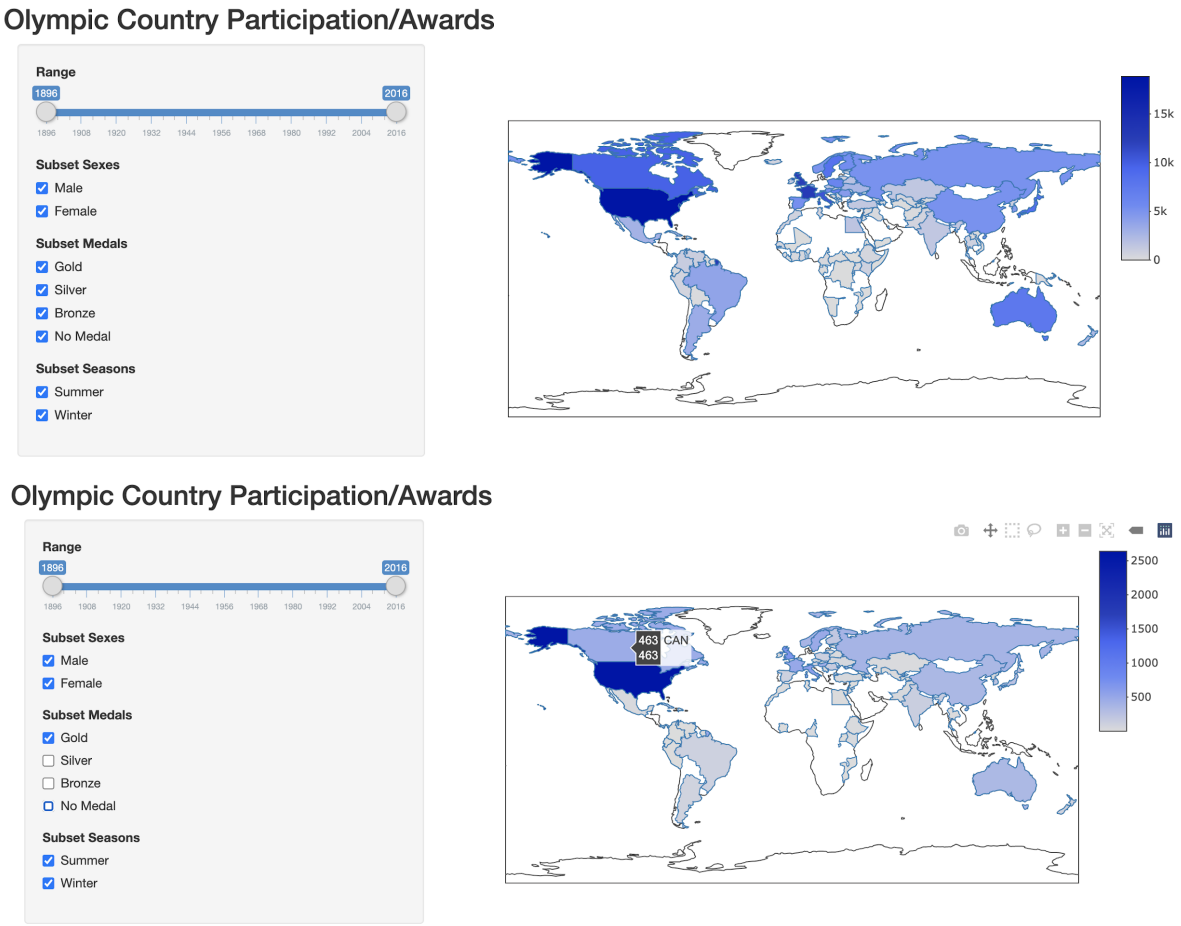
Figure 4: Stacked Bar Plot of proportion of athletes from each BMI type and how many medals they won. The left plot shows Rhythmic Gymnastics, the right plot shows Bobsleigh

Visualization Insight:

Both of the scatter plots agree with the insights found in the exploratory data analysis, where the country above the red line in the underweight scatter plots are all located in Africa while the country above the red line in the overweight plots are island country located on pacific. This insight is interesting, because the athletes' group could be viewed as a sample of the total population of a region or type of region, in other words, it is reasonable to draw a rough conclusion of how geographical location can impact the body type of a person. The future study of this topic could involve more features, such as their income, living condition, and education level, etc.

The bar plots show us how does athlete's body mass impact their performance.In the Rhythmic Gymnastics sample, in those that won a medal there is 8 percent more underweight athletes than normal weight, which means lose weight would help rhythmic gymnastics' athletes perform well. On the contrary, bobsleigh requires high body mass, in the bar plot, the proportion of overweight athletes who have a medal is 10 percent higher than those who have not. For the next step, it may be interesting to study the reason why such sports like rhythmic gymnastics or bobsleigh require a slim or strong body.

Visualization 4: Country Performance and Participation *(link)*



Figures 5 and 6: Choropleth map of country participation. Participation can be filtered based on values shown on left (Sex, Medals, Season, Year Range)

In order to fully understand how many athletes participated for each country, and how many athletes won awards for their respective countries, a Dashboard was made with RShiny. The best way to display this data is through a choropleth map via Plotly, which is essentially a thematic world map where the different countries are shaded a different color based on the amount of participants (darker colors mean more participants). As can be seen, a plethora of interactive filters allow the user to properly define what subset of the data they want to visualize on the map. They can filter based on sex, the medals awarded, the season, and they can select the range of years using the slider on the top. The map itself is interactive, meaning the user can zoom in on specific countries, and if they want to see the exact amount of participants that fall into the predefined filters for a specific country, they can hover their cursor over the country of interest and it would produce the value. The data for this visual was created using Tidyverse in R. Within the dashboard, when the user set a filter, the code would update the filters within the dataframe to include only the data the viewer wants to see. A "participants" attribute is created to summarize the amount of participants included in the filter for each country. This is then used to update the graph.

Based on this visualization, quite a few insights could be gleaned. To begin with, it can be seen that no matter what the filter (sex, year range, season, and medals won) North America seems to consistently beat out all other areas. In the summer, the USA consistently beats out other countries. However in the winter, Canada seems to consistently hold the advantage in gold winners, but the US is able to keep up in participants, and in silver and bronze awards. France, Italy, and Great Britain were able to consistently keep up with the USA in terms of participants over time, but when it came to winning medals, the US edges them out.

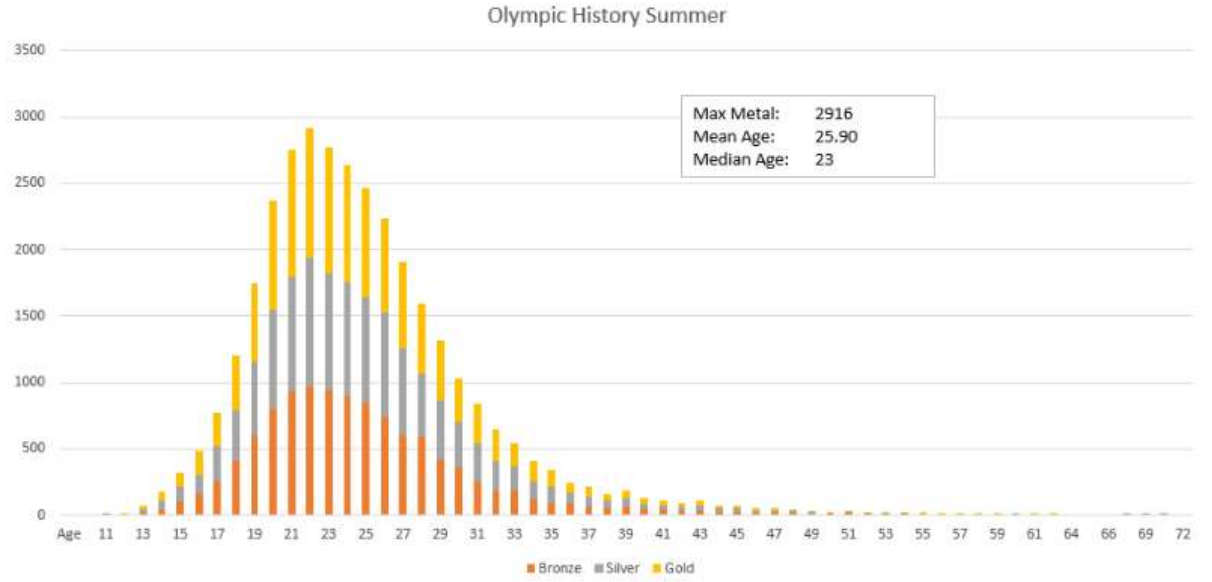
It can be seen that Africa and Asian countries in general weren't involved with the Olympics at the start, but over time their participation seemed to increase. Although with the exception of China from the 1970's onwards, the countries on these continents never really got close to rivalling Europe or North American countries in terms of medals. Across the Olympics history, China, Australia, and Russia were the only companies to really compete with the amount of medals being won in the European countries and North America.

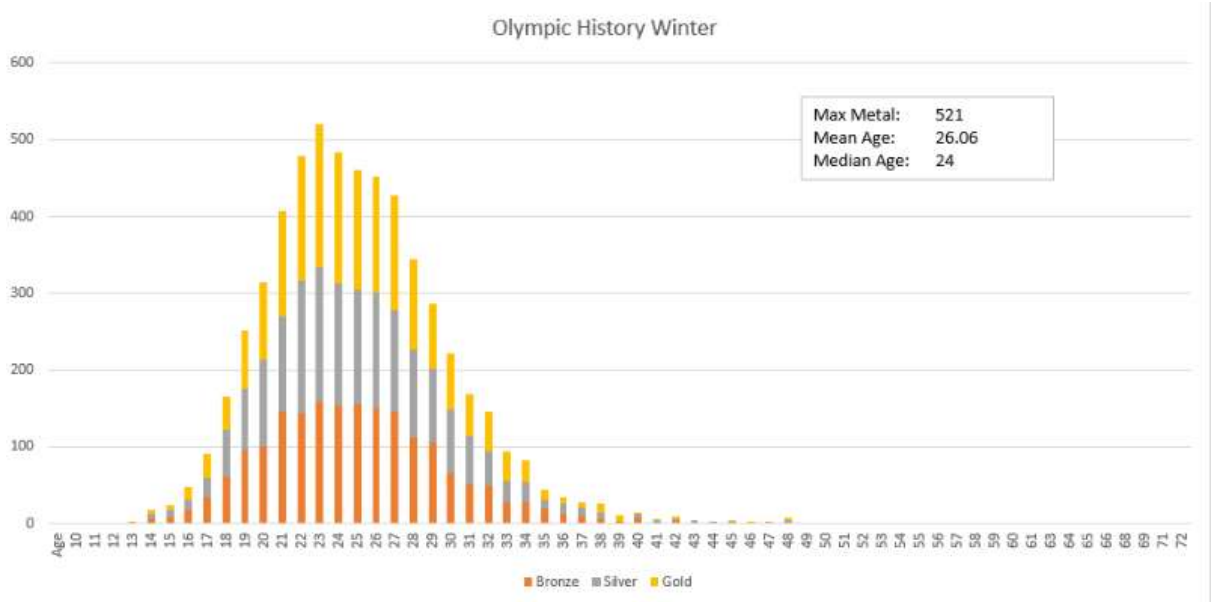
Women participation was not very high in the beginning; only a few European countries and the US seemed to have female athletes. This changed in the 1920's, and since then the amount of female participants has been drastically increasing. Although this might be inflated with the steady introduction of more sports over time.

Further analysis can be done on the impact outside data may have on these trends. For example, we could correlate GDP to country participation and awards, or even employment rate. We could further understand women's rights in different countries based on their performances in the Olympics, and even though correlation does not necessitate causation, humanitarian organizations might be able to use country Olympic data to further understand the opportunities their people have, and the directions these countries are trending in.

These insights are also important to the stakeholders in the Olympic Games. They need to understand trends in participation rates over time, especially from different groups. They want to see which countries are winning more medals, because it then becomes more likely to get the people of those countries more excited about the games, bringing in more revenue.

Visualization 5: Medalist and Age





In order to understand the best performing age of athletes in both the winter game and the summer game. We made two stack bar charts to represent the amount of medals won by athletes of different ages stacking by the type of medal. Medalists in winter games are slightly older than the medalists in summer games. They are not only able to win medals after the mean age but also starting to earn medals older. However, we are able to see athletes still winning Olympic summer game medals even after their 50s while winter game athletes do not perform so well after they are old. We need further analysis about the data to fully understand which specific category of sports allows young and old athletes to out-perform those in the 20s. Insights into the age are important because we can understand more deeply into a human's physical ability throughout different stages of our life. This can be useful for the stakeholders in Olympic Games. They are able to find some athletes that are still performing well in their older age. This is good for targeted advertisement because they can advertise different sport events to people of different ages. Older people can be very excited

to see people of their age winning against people in the 20s, especially if they are sports icons, bringing in more revenue.

Conclusion (Future Work/Closing Remarks)

Overall the visualizations very adequately showed different insights surrounding the participation of different countries, and their performance within the Olympic Games as well as the performance of athletes based on their body types and the events they were taking part in. Animations, graphs, and maps were shown to give potential stakeholders insights on the countries that most actively took part in the Olympics, the countries that ended up winning, the countries that ended up losing, and different lurking variables behind these trends were discussed.

In the future, it would be insightful to stakeholders to further explore the reasoning behind the trends explored in this case study. It would help athletes further prepare for the games if they understood what body types to strive for to be successful and why. These insights would help investors understand which countries are going to be successful in the future as they can target their fan bases for support and advertisement revenue, and it could even help humanitarian organizations if these trends could be tied with country health and economic metrics, in order to understand which countries are trending in the right direction towards development, and which countries might be regressing socially.

Appendix: Data Feature Description:

"ID", A given ID for the dataset that increment automatically
"Name", The name of the athletes competed
"Sex", The Gender of the athletes competed
"Age", The age of the athletes competed
"Height", The height of the athletes competed
"Weight", The weight of the athletes competed
"Team", The team name(usually the name of the country/region the athletes represent)
"NOC", A three letter string that represent the Country/region
"Games", The year and(summer/winter) of the Olympic game
"Year", The year of the Olympic game
"Season", Summer or Winter Olympic game
"City", The city where the Olympic game took place
"Sport", The category of discipline that the athletes competed in.
"Event", The specific discipline that the athletes competed in.
"Medal" The Medal that the athletes won in the event.(‘NA’ is no medal)