25847333_Project_3_Report

Project 3: Olympics Report

Introduction

The Olympic Games have been a constant event that brings together some of the world's best athletes from across the globe and provides them with an opportunity to display their skills and talents. However, there are a number of complex dynamics which have influenced the Olympic success of both the athletes and their respective countries. The goal of this analysis is to investigate trends and relationships in Olympic performance by looking at countries, sports, age and gender. By analysing changes over time, visualizing the data, and discussing the key factors that influence performance, this analysis aims to provide the reader with a comprehensive understanding of the factors that influence Olympic success. The Olympic dataset that we will be analysing includes data about all athletes, medallists, hosts, and results from 1896 to 2022.

Furthermore, this analysis will be conducted using the following research questions:

- How have participation trends for male and female athletes in the Olympic Games evolved over time?
- What is the relationship between the distribution of medals won by the top 3 countries in individual events at the Winter Olympics and at the Summer Olympics?
- How do age demographics of athletes correlate with medal wins?
- What are the relationships between host countries and the number of medals their athletes win in individual events?
- How does the number of medals won by the top athletes from a specific country in their sport compared to the total number of medals the country has won in that sport for individual events?

Data Cleaning and Preparation

Tidying data is an important step to ensure that the user is able to easily read and work with the data. In order to tidy the Olympic datasets, we will be making use of the 'tidyverse' package because it contains several different packages that aid in tidying data.

According to Wickham, Çetinkaya-Rundel and Grolemund, there are 3 rules to ensure that a data set is tidy. The first rule being that each variable has its own column, and each column has its own variable. The second rule is that each observation is a row, and each row is an observation. Lastly, each value has a cell, and each cell has a single value (74).

Data Preparation Process for The Athlete Dataset

- Reading in csv files on an IOS device. First step is using the read_csv function. I then have to save my quarto document into the same folder where my project data is stored. I then had to click on the 'sessions' button and click on 'set working directory' and set it to the 'source file location'.
- The first step to tidying the data was to check if any athletes had only a first name or a last name. I used the str_detect which was able to check if any other the athletes did not have a space between their first and last name. If there was no space detected this meant that there was only one name provided.
- Secondly, I needed to separate 'athlete's full names' into their first and last name. This is essential because it definitely makes it much easier and convenient for me to analyse the data and the athletes clearly. I used the function known as 'separate wider delim' in order to separate the athlete_full_name column into two separate columns. This specific function allows you to identify a delimiter within the rows to be able to easily separate the values. For example, in my data I identified an empty space as the delimiter so when R came across an empty string (" ") it knew to separate them. Along with this I had to make use of the "too_few=align_end" function. When there are missing values, which we have previously noticed for one of our athletes, R is able to add an NA to our data and create an explicit missing value. This function allows us to control where the NA's go (Wickham, Çetinkaya-Rundel, Grolemund, 234). By adding NA to our athletes first name, it ensures that our code will not provide us with an error due to 'too few' values. In the line below this, we have "too many=merge". This essentially does the opposite and if there are more than two names in the dataset it merges them together to ensure that we only have two values.
- In order to ensure that I coded exactly what I wanted to see in my output, I tested my code using the filter function. Filter allows me to select which rows are present and I have asked R to only print out the rows which have a NA value in it (is.na) (Wickham, Çetinkaya-Rundel, Grolemund, 41).

- Removing columns that will not be necessary for my data analysis is essential to ensure my data looks clean and tidy. I did this by using the select function with a minus(-) symbol to ensure that it removes the specified columns from my data.
- After this, I arranged the games participation column in descending order to easily see which athletes have participated in the most Olympic games. Arranging changes the order of the rows and in this case, we have told it to arrange in descending order (Wickham, Çetinkaya-Rundel, Grolemund, 41). I also make use of the relocate function to change the position of the columns. In order to make the athletes details clearer, it would be easier to place their year of birth after their last name. By default, it moves the columns to the front, so in order to place it after the last name column we use the .after argument.
- In the next code, I have used the filter function which only selects the rows in the dataset where the athlete_medals contains a digit that is followed by a space and the letter "G", "S" or "B". The mutate function then creates 3 separate columns called Gold, Silver and Bronze which has the number of medals for each category. Str_extract allows us to effectively extract the digit that appears before the space and the letter.
- Once again removing columns that are not necessary helps to tidy data. I removed the athlete medals column because in the previous step we separated the data into their individual categories. I made use of the select function with a minus(-) symbol to ensure that it removes the specified columns from my data.
- Furthermore, I used the mutate function to change the existing gold, silver and bronze columns to numeric values. This was done in order to perform calculations at a later stage using these values.
- The relocate function allows us to relocate the Gold column to appear before the 'first_game' column to make the data easier to see.
- The relocate function allows us to relocate the Silver column to appear after the 'Gold' column to make the data easier to see.
- The relocate function allows us to relocate the Bronze column to appear after the 'Silver' column to make the data easier to see.
- I have made use of the 'separate' function. This allows us to separate the first_game character column into multiple columns with a regular expression. The data goes into 2 columns which we have named "host" and "year". We separate the values with a regular expression. The regular expression looks ahead to identify a space followed by any four digits which represent the year. If there are any additional values, we merge them together to prevent any errors from occurring.
- In order to make the column names look similar, more descriptive, cleaner and easier to read I have used the rename function to change them manually. I have provided the new column name on the left followed by the old column name on the right.

- Lastly, I have used the mutate function. Mutate replaces the old column with the new column which in this case converts everything to lowercase letters. After this, I made the column title case, to capitalize the first letter of their last names.

Data Preparation Process for The Hosts Dataset

- My first step to tidying this dataset is to modify the game_slug column. I have made use of the mutate function and have replaced all the dashes in the dataset with a space to make it cleaner and easier to read. I have then converted it to title case in order to capitalize the first letter of each country name.
- Secondly, I have made use of the 'separate' function. This allows us to separate the game_slug column into multiple columns with a regular expression. The data goes into 2 columns which we have named "host" and "year". We separate the values with a regular expression. The regular expression looks ahead to identify a space followed by any four digits which represent the year. If there are any additional values, we merge them together to prevent any errors from occurring.
- Thirdly, I separated 'game_end_date' into different columns for the date and the time. I used the function known as 'separate wider delim' which allows you to separate the values using a delimiter. The delimiter is an empty space so when R came across an empty string (" ") it knows to separate the values. Along with this I had to make use of the "too_few=align_end" function to add an NA to our data and create an explicit missing value. This function allows us to control where the NA's go (Wickham, Çetinkaya-Rundel, Grolemund, 234). By adding an explicit missing value, it ensures that our code will not provide us with an error due to 'too few' values. In the line below this, we have "too many=merge". This essentially does the opposite and if there are more than two names in the dataset it merges them together to ensure that we only have two values.
- I then needed to repeat the exact same step for the game_start_date column. Using the separate_wider_delim function I separated the date and time into their own columns when a " " space appeared.
- In addition to this, I have had to convert all our clean data to their respective formats. I used mutate to change the year column to numeric, the date columns to a date format using year, month, day formats and change the time columns to a time format using the hour, minute and second function.
- I have used the select minus(-) function to remove the game_name column because it is unnecessary for my analysis and I have also removed the game_year rows because it is a duplicate of the year column.

- In order to make the column names look similar, more descriptive, cleaner and easier to read I have used the rename function to change them manually. I have provided the new column name on the left followed by the old column name on the right.

Data Preparation Process for The Medals Dataset

- My first step to tidying this data set was to get rid of columns that would not be useful for my analysis. This was done using the select function and the minus (-) sign before each column name to indicate that these columns should be removed from the dataset.
- Secondly, I have made use of the 'separate' function. This allows us to separate the slug_game column into multiple columns with a regular expression. The data moves into 2 columns which have been named "host" and "year". We separate the values with a regular expression. The regular expression looks ahead to identify a dash followed by any four digits which represent the year. If there are any additional values, we merge them together to prevent any errors from occurring.
- I have then used the mutate function to change the existing year column to a numeric column. This was done in order to perform calculations at a later stage using these values.
- I have then used the mutate function to change the existing host column to title case. Title case makes each word in the host column start with a capital letter.
- Additionally, I separated 'athlete_full_name' into different columns for the first name and the last name. I used the function known as 'separate wider delim' which allows you to separate the values using a delimiter. The delimiter is an empty space so when R comes across an empty string (" ") it knows to separate the values. Along with this I had to make use of the "too_few=align_end" function to add an NA to our data and create an explicit missing value. This function allows us to control where the NA's go (Wickham, Çetinkaya-Rundel, Grolemund, 234). By adding an explicit missing value, it ensures that our code will not provide us with an error due to 'too few' values. In the line below this, we have "too many=merge". This essentially does the opposite and if there are more than two names in the dataset it merges them together.
- In the next step, I have used the mutate function. Mutate replaces the old column with the new column which in this case converts everything in the athlete_last_names column to lowercase letters. After this, I made the column title case, to capitalize the first letter of their last names.
- In order to make the column names look similar, more descriptive, cleaner and easier to read I have used the rename function to change them manually. I have provided the new column name on the left followed by the old column name on the right.

- Furthermore, I have used the fct_recode function to standardize and clean the values in the column. Fct_recode changes the values that appear in the medal type column which has allowed me to change the word “Gold”, “Silver” and “Bronze” to just the letter “G”, “S” and “B”.
- Lastly, I have arranged the Olympic_sport column in alphabetical order to be able to easily find and sort through specific sports.

Data Preparation Process for The Results Dataset

- I have used the ‘separate’ function to separate the slug_game column into multiple columns with a regular expression. The data moves into 2 columns named “host” and “year”. We separate the values with a regular expression. The regular expression looks ahead to identify a dash followed by any four digits which represent the year. If there are any additional values, we merge them together to prevent any errors from occurring.
- I then changed the year column to a numeric column by using the mutate function
- In order to only have events that individual athletes competed in we used the filter function. It filtered where the participant type was equal to athlete and therefore did not include the team events.
- I separated ‘athlete_full_name’ into different columns for the first name and the last name. I used the function known as ‘separate wider delim’ which allows you to separate the values using a delimiter. The delimiter is an empty space so when R comes across an empty string (“ ”) it knows to separate the values. Along with this I had to make use of the “too_few=align_end” function to add an NA to our data and create an explicit missing value. This function allows us to control where the NA’s go (Wickham, Çetinkaya-Rundel, Grolemund, 234). By adding an explicit missing value, it ensures that our code will not provide us with an error due to ‘too few’ values. In the line below this, we have “too many=merge”. This essentially does the opposite and if there are more than two names in the dataset it merges them together.
- In the next step, I have used the mutate function. Mutate replaces the old column with the new column which in this case converts everything in the athlete_last_names column to lowercase letters. After this, I made both the athlete_last_names column and the host column title case, to capitalize the first letter of each word.
- I then made use of the select function to select which columns to keep in our dataset using their names. I left out the column names that I did not want to appear in my dataset.
- Relocate allows us to move the columns around. I have used relocate to move the athlete_first_name column to appear after the year column. I then relocated athlete_last_name to appear after the athlete_first_name column. I also relocated the

participant_type column to appear before the country_name column. Finally, I arranged the rank_position column in ascending order.

- In order to make the column names look similar, more descriptive, cleaner and easier to read I have used the rename function to change them manually. I have provided the new column name on the left followed by the old column name on the right.
- Lastly, I have used the fct_recode function to standardize and clean the values in the column. Fct_recode changes the values that appear in the medal type column which has allowed me to change the word "Gold", "Silver" and "Bronze" to just the letter "G", "S" and "B".

Transforming Data

Games Participations Process

- The first step to transforming the data in order to answer my question regarding male and female participation in the Olympic games is to join tables together. I have inner joined the medal dataset and the athlete's dataset by their first_name and last_name.
- I have then filtered the dataset to include rows where the event_gender is equal to women and men which excludes any other event genders.
- The third step is to group the dataset by the Olympic_year and event_gender. After this, the summarise function applies to each column that has been specified in the group by function. It calculates the total number of participations for each Olympic year and gender.
- In order to tidy the data even more, I have arranged the Olympic_Year in ascending order so that it is in a chronological order.

The Winter vs Sumer Olympics Process

- The first step to transforming the data in order to answer my question regarding the distribution of medals won by the top 3 countries in individual events at the Winter Olympics versus at the Summer Olympics is to join 2 tables together. I have inner joined the medals dataset and the hosts dataset by the year of the Olympics. I have included many to many relationships to ensure that the join is able to handle many values for the same year.
- Secondly, I have filtered the game_season to only include winter Olympic games in the winter dataset and have filtered the game_season to only include summer Olympic games for the summer dataset.

- The third step for both datasets is to filter the data to only include events that have been completed by individual athletes. I am not interested in the team Olympic games and therefore want to exclude it from my data.
- I then group by the country and calculate the total number of medals for each country.
- Lastly, I arrange the total number of medals in descending order to identify who has the most medals.
- I only want to know the top 3 countries which is why I make use of the slice_head function. The slice_head function selects the top 3 countries with the highest total number of medals.

Age Demographics and The Number of Medals They Win

- The first step to transforming this data is to filter by values that are NOT NAs values. The & symbol ensures that it removes NA values from both the birth_year and the Olympic_year column.
- Secondly, I have used an inner join to join the medals and athlete datasets based on the columns for athlete_first_name and athlete_last_names.
- Mutate calculates a new column called age which does a mathematical equation to work out the age athletes were when they attended the Olympics.
- I then filtered the data to only include the athletes if their age is higher than 13 years old because this is the age that most countries allow their athletes to participate from.
- Lastly, I group by age and gender in order to summarise the total number of medals each age category has won. I have included gender in order to see which medals are won by males, females, mixed and open. I have used the .group = “drop” function to ungroup the data after summarisation is complete.

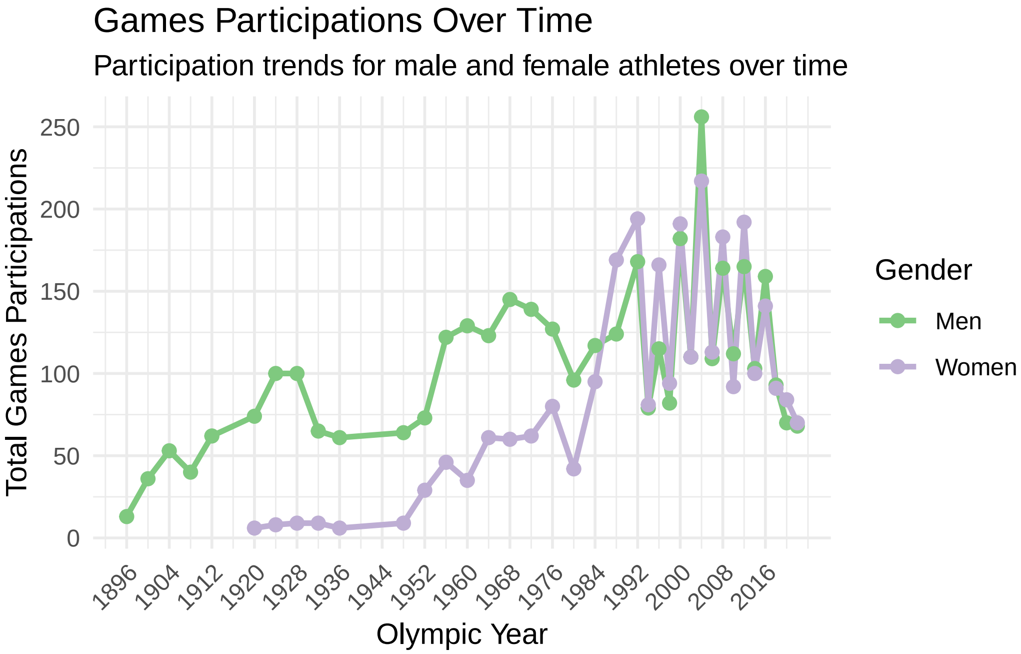
Hosts Versus Non-Hosts

- Firstly, I needed to calculate the total number of medals each country has by adding up the gold, silver and bronze columns. I then created a new database that included the total number of medals that each country has. I then filtered the data to only include countries that were not the Soviet Union because they did not host an Olympic event. After this, I arranged the total medals from biggest to smallest and only selected the top 3 countries using slice_head.

- I then created individual datasets for the top 3 countries in the summer olympics. The first one being for the United States, the second one being Germany and the last one being Great Britain. I first checked when each country hosted the Olympics and then filtered each data set by the country's name. I looked at a period of 30 years for each country that surrounded the year that they had hosted. In order to ensure I was only looking at the summer Olympics, I had to filter the datasets by the game_season. Looking at one season ensures that we have more accurate comparisons throughout our analysis. Finally, I calculated the total number of medals that each country won every year over a period of 30 years. Within these 30 years, we can also see the number of medals they won when the hosted. United States hosted in 1996, and we are looking at the years from 1990 and 2020. Germany hosted in 2936, and we are looking between the years 1930 and 1960. Great Britain hosted in 2012, and we are looking at a period of 30 years between 1990 and 2020.

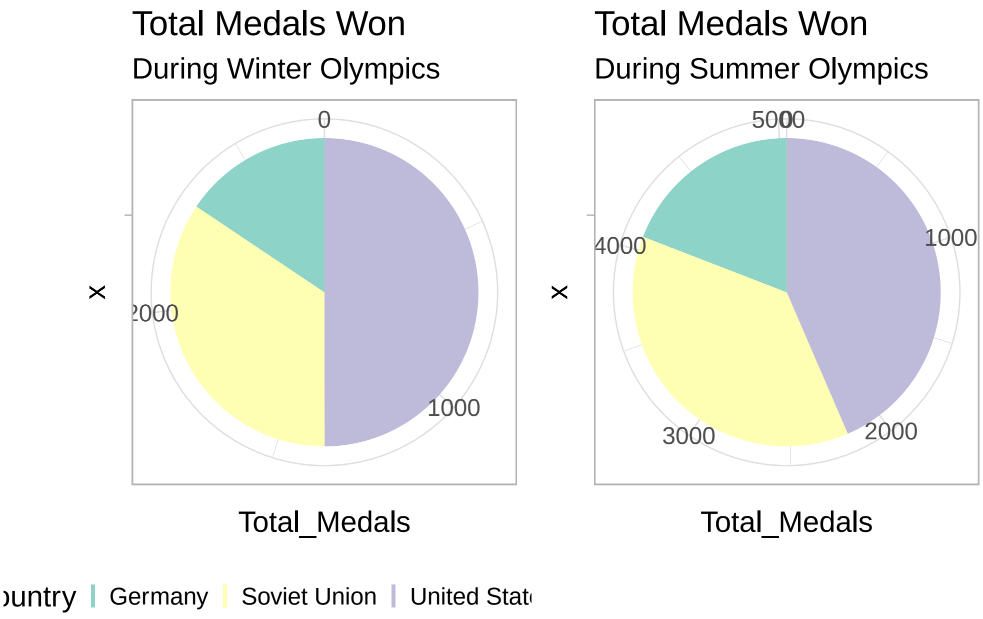
Top Athletes versus Their Country

- First step to transforming this data is to remove the rows that contain a NA value in the Country column using (! (is.na)) function. We then need to work out the sum of the total medals each country has received. After this, I organized the total number of medals in descending order and only selected the top 5 countries using slice_head.
- After filtering the dataset by the top 5 countries and ignoring NA values in the athlete_first_name and athlete_last_name column I moved onto grouping by the dataset. I have grouped by country, athlete and sport in order to count the total number of medals for each athlete in each sport using the n () function. Then the arrange function sorts the resulting dataset first by Country in alphabetical order and then by Total_Medals in descending order.
- The next step involves using the slice_head function in order to select the top athlete with the most medals from each country.
- Then we want to count the total medals by Sport for Each Country which we do by first grouping by the sport and country column. Then we use the n () function to count the total number of medals.
- Lastly, we join the best athlete data with the sport medals data using a left join. The rename function allows us to change the column names in the dataset to be more descriptive and easier to understand. It is essential to clean this new data set which is done using pivot_longer. Pivot longer takes the columns and joins them making it easier to work with because we have less columns.

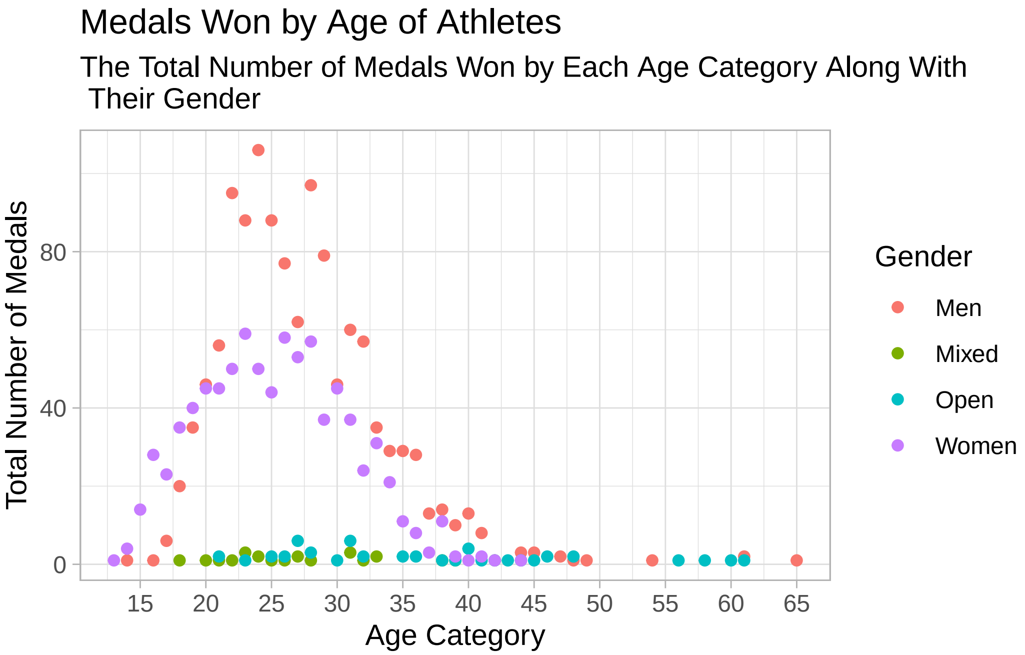


The above graph is a line graph that clearly shows how the male and female participation trends in the Olympics have evolved over time. The graph includes games from 1896 to 2022. While male participation remained dominant throughout most of the years, it is evident that female participation has drastically increased over time.

According to my statistical summary, the average male participation in 1920 was 2.64. This implies that on average, male athletes competed in just over 2 events during the Olympic Games. However, the statistical summaries show that the actual number of events varied from a minimum of 1 to a maximum of 5, indicating that there is some variation in the participation of individual athletes.

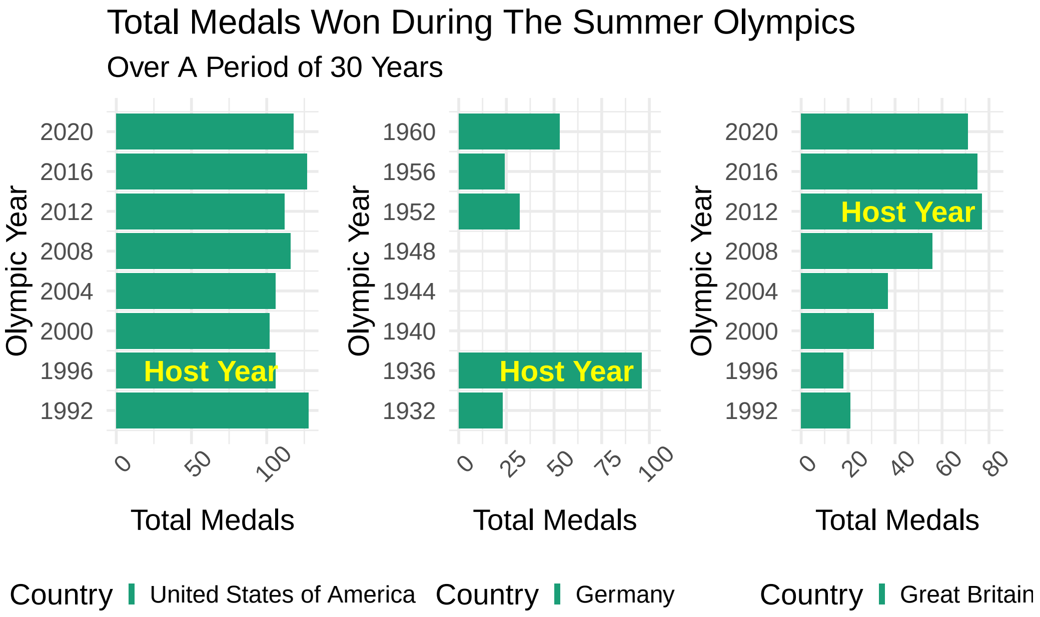


The above graph contains two different pie charts showing the relationship between the distribution of medals won by the top 3 countries in individual events at the Winter Olympics versus the number of medals won at the Summer Olympics. When examining the graphs, it is obvious that each country included in our analysis has experienced distinct differences in their performance during winter and summer.



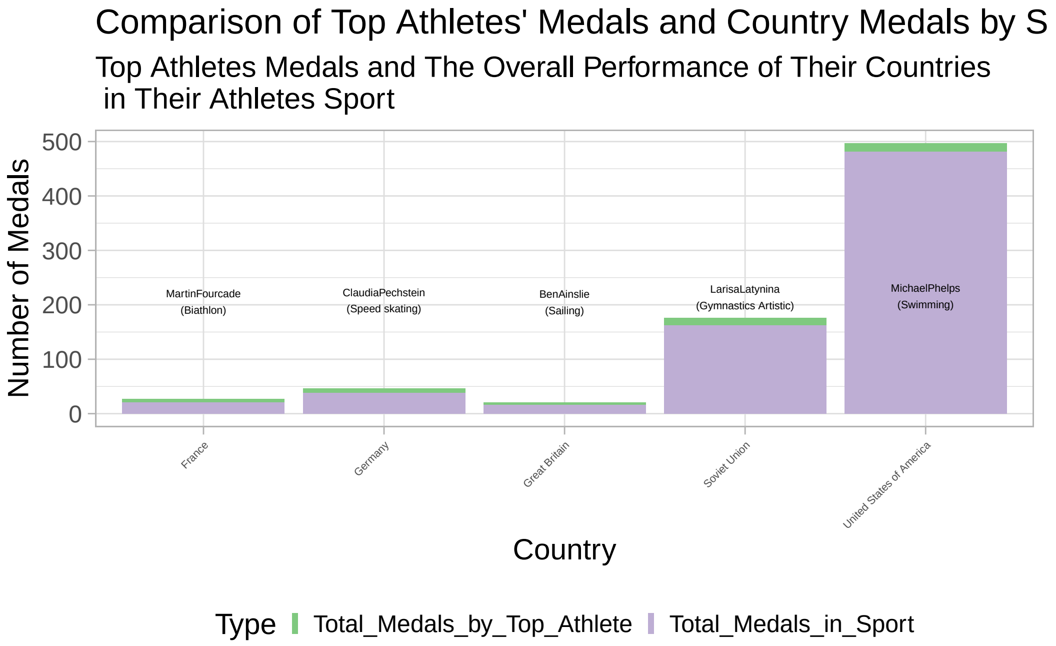
The above graph is a scatterplot graph that demonstrates how the age demographics correlate with the medal wins. This analysis has revealed some intriguing information regarding age, gender and medal success. We can visually see that younger male and women athletes tend to dominate the number of medals won.

When looking at the statistical summaries, the age group of 29 year old athletes have the highest average medals which is 58%, This indicates 29 year olds often perform exceptionally well in the Olympics in comparison to other age groups. We can see that the top 10 age categories with the most (max) number of medals are in order from 29 years old to 20 year old. This states that athletes in their 20s are at their peek success age.



The above graph displays three bar charts that examine the relationship between host countries and the number of medals their athletes have won over a 30-year period during the Summer Olympics, comparing it to the performance of non-host countries over the same timeframe. The graph reveals that host countries often experience a boost in performance leading to an increase in medals.

Our statistical summaries show us that overall United States of America had the max number of medals and Germany had the lowest (min) average of medals.



The above graph shows a stacked bar chart comparing the number of medals won by top athletes from a specific country in their sport to the total number of medals the country has won in that sport for individual events. The graph helps us to identify if the country's success is greatly driven by one athlete or if it is distributed amongst other athletes.

When looking at the statistical summaries we are provided with insightful perspectives. The comparison between the total number of medals earned by a top athlete (9.8) to the total number of medals won by a country in a particular sport (143.6) is shown in our data. A top athlete in a sport has an average of 9.8 medals which indicates the exact proportion that they have contributed to their country's overall performance in that sport.

Narrative:

The main objective of my analysis is to present a logical narrative that contains data about participation patterns, medal distribution, age demographics, host performance, and athlete accomplishments throughout the Olympic events. The goal in looking at these aspects is to draw attention to the important trends and factors that influence Olympic performance.

Over the years the Olympic Games have seen a significant shift in their participation trends with regards to the number of women participants. This shift is significant as it emphasises the ongoing societal challenge of overcoming gender inequality within sports. When looking at the seasons, there is a noticeable trend that countries often win substantially more medals during the summer Olympics compared to the winter Olympics. In addition to this, the age demographics reveal that athletes in their 20s tend to dominate the total number of medals won. This can possibly indicate that the prime years for athletic success is in your 20s. Host countries exhibit a notable increase in the number of medals won when hosting the Olympics. It suggests that hosting the Olympic games provides the host country with a clear competitive advantage. For example, being able to train in a familiar environment certainly gives you an advantage when competing against foreign countries. Top performing athletes play a significant role in their countries overall Olympic performance. It is evident that countries rely heavily on their top athletes due to them contributing a large share of their country's overall medals.

Key Takeaways:

- Female participation has increased significantly over time, especially after the 1980s.
- Countries generally achieve more medals during the Summer Olympic Games.
- Athletes in their 20s tend to dominate the most number of medals in the Olympics.
- Host countries tend to receive more medals in the Olympics due to their 'home advantage'.
- Top performing athletes often make up a significant portion of their country's total medal count in specific sports.

Implications:

- An increase in female participation can usually imply that there is progress in achieving gender equality within sports and society. It can further imply that women are increasingly being provided more opportunities.
- Athletes in their 20s tend to receive the most medals. This could imply that your 20s are the best years for athletic success. It can indicate that during your 20s athletes are in good physical condition and are strong. It could also imply that they have had enough time to build experience and develop skills.
- Host countries receiving more medals can imply that home advantage has a significant impact of their country's performance. For example, athletes training in their home country allows them to become familiar with the environment. Additionally, support from your country at home can also increase your performance levels.

Limitations:

- Missing values was one of my main limitations throughout my analysis such as athlete first and last names. Incomplete and missing values can lead to incorrect or skewed conclusions.

- The dataset only covers Olympic events from 1896 up until 2022. It does not include the recent Olympic events that were held this year (2024). It also excludes any previous Olympic events before 1906 which means there is a lack of data for my analysis.

Areas for Further Research:

- How technology impacts the overall performance of a country especially when it comes to the Paralympics. This is because better technology would allow athletes to be able to compete in certain events even though they have a disability.s