Exploring Olympic History through Data Visualization

Data visualization serves as a pivotal aspect of data analysis and communication, leveraging graphical representations to distill complex datasets into accessible insights. Through the use of various visual elements such as charts, graphs, and maps, data visualization enhances the interpretability of information, enabling both experts and laypersons to discern patterns and trends at a glance. The selection of appropriate visualization types, thoughtful use of color and design, and a focus on clarity contribute to creating compelling visual narratives. Additionally, interactive features empower users to dynamically engage with the data, promoting a deeper understanding and facilitating more informed decision-making.

In the context of exploring Olympic history, the dataset obtained from Kaggle offers an exciting opportunity to delve into 120 years of Olympic stories, triumphs, participating nations, and the global unity celebrated through the Games. Comprising two primary CSV files, 'athlete_events.csv' and 'noc_regions.csv,' this dataset forms the foundation for our project. Our primary goal is to explore and visualize this dataset using the R programming language, aiming to extract valuable insights and narratives that capture the essence of the Olympic Games.

Data Set Selection and Source

The chosen dataset, the Olympics Dataset, is sourced from Kaggle, a prominent platform for data science and analysis. This meticulously curated dataset aims to encapsulate the essence of Olympic history by providing comprehensive data on athletes, their performances, and participating nations. The dataset consists of two central CSV files,

1. 'athlete_events.csv': This file is substantial, with 271,116 rows and 15 columns. Key variables include:

- 'ID': A unique identifier for each athlete. The data type is Numeric / Integer.
- 'Name': The athlete's name. The data type is Text / String.
- 'Sex': The athlete's gender. The data type is Text / Character.
- 'Age': The athlete's age during the event. The data type is Numeric / Integer.
- 'Height': The athlete's height (recorded in centimeters). The data type is Numeric / Integer.
- 'Weight': The athlete's weight (recorded in kilograms). The data type is Numeric / Integer.
- ‘Team’: The athlete’s country name. The data type is Text / String.
- 'NOC': The National Olympic Committee's short-form code. The data type is Text / String.
- 'Games': The Olympic Games in which the athlete participated. The data type is Text / String.
- 'Year': The year of the event. The data type is Numeric / Integer.
- 'Season': The season of the event (Summer/Winter). The data type is Text / String.
- 'City': The host city of the event. The data type is Text / String.
- 'Sport': The sport in which the athlete competed. The data type is Text / String.
- 'Event': The specific event within the sport. The data type is Text / String.
- 'Medal': The medal won by the athlete (Gold/Silver/Bronze/NA). The data type is Text / String.

** The missing values in columns like height, weight etc. are represented by NA.

2. **'noc_regions.csv'**: This file comprises 230 rows and 3 columns. Key variables include:

- 'NOC': National Olympic Committee's 3-letter code. The data type is Text / String.
- 'Country name': The country's name. The data type is Text / String.
- 'Notes': Additional notes about the committee's participation.

** The missing values in Notes column is represented by blank space.

Data Wrangling

Data wrangling, often referred to as data munging or data cleaning, is a crucial step in the data preparation process that focuses on transforming and cleaning raw data into a format suitable for analysis. It involves several key tasks to enhance the quality, consistency, and usability of the data, ensuring that it is well-structured and ready for exploration. Some generic aspects of data wrangling include:

1. Handling Missing Values: Identifying and addressing missing values is a fundamental aspect of data wrangling. Strategies may include imputing missing values based on statistical measures (such as mean or median) or removing incomplete records. This process is essential to maintain data integrity and prevent biases in subsequent analyses.
2. Dealing with Duplicates: Detecting and removing duplicate records helps eliminate redundancy and ensures that each observation in the dataset is unique. Duplicate values can skew analytical results and lead to inaccurate interpretations, making their identification and removal a critical part of the data wrangling process.

3. Data Type Conversion: Ensuring that data types are consistent with their intended use is vital for analysis. Data wrangling often involves converting variables to the appropriate data types, such as converting strings to numeric values or handling date and time formats for standardized analysis.
4. Handling Outliers: Identification and treatment of outliers are essential for maintaining the robustness of analyses. Extreme values can significantly impact statistical measures, and data wrangling may involve methods like trimming, winsorizing, or transforming data to mitigate the influence of outliers.
5. Data Standardization: Standardizing data formats and units across variables facilitate comparisons and analyses. Data wrangling may include converting units, aligning date formats, or harmonizing categorical variables to ensure consistency throughout the dataset.
6. Text Cleaning and Parsing: In datasets containing textual information, data wrangling involves cleaning and parsing text data to extract relevant information. This may include removing special characters, standardizing text formats, or tokenizing text for further analysis.
7. Creating Derived Variables: Data wrangling allows for the creation of new variables derived from existing ones. This includes generating features that provide additional insights or aggregating data to a more manageable level for analysis. Derived variables often enhance the dataset's richness and analytical potential.
8. Data Integration: Combining data from multiple sources is a common data wrangling task. This involves merging datasets based on common identifiers or keys, ensuring that relevant information is consolidated for a comprehensive analysis.

Data wrangling is an iterative and often creative process that requires a deep understanding of both the data and the analytical goals. Its successful execution lays the groundwork for meaningful and reliable data analyses and visualizations.

Data Discovery and Cleaning

The data discovery and cleaning process reflects a meticulous approach to handling missing values and ensuring data integrity. The identification and replacement of missing values demonstrate a commitment to maintaining the dataset's accuracy.

athlete_events Data:

- Identified 'NA' values in height, weight, age, and medal columns.
- Handled 'NA' values in the medal column (categorical data) by replacing with 'None.'
- Filled 'NA' values for numerical columns (Age, Height, Weight) with median values.
- Removed duplicate values.
- Validated with boundary cases.

noc_regions Data:

- Identified 'NA' values in the region values.
- Filled 'NA' values for the 'Notes' column with 'None.'

Data Integration

Joined both datasets using the 'NOC' code present in both files to create a comprehensive dataset for analysis. This integration allows for a holistic examination of Olympic history, incorporating details about athletes and their respective

The process of working with the dataset involves a multi-step approach, including data exploration, cleaning, and integration. The thorough validation with boundary cases and the removal of duplicate values further contribute to the reliability of the dataset.

The chosen dataset, coupled with the meticulous data processing steps, lays a solid foundation for a comprehensive exploration of Olympic history through data visualization. The project's findings have the potential to contribute significantly to our understanding of the Olympic Games, capturing the essence of global unity, athletic achievements, and the evolving landscape of this historic sporting event.

Project Objectives

Main objectives for working with this dataset are as follows:

1. Demographic Analysis: To examine the demographic characteristics of athletes over the years, including age, gender, height, and weight.

2. Country Performance: To explore trends in Olympic participation and medal distribution among countries, identifying consistent top-performing nations and analyzing their performance evolution.

3. Sports Evolution: To investigate the evolution of Olympic sports and events, highlighting changes in popularity and the emergence of new disciplines.

4. Host City Impact: To visualize the growth and impact of the Olympic Games on host cities, covering aspects such as infrastructure development, tourism, and long-term sports legacy.

To achieve these objectives, I wanted to find answers to specific research questions, including:

- How has the age and gender distribution of athletes evolved over the 120-year history of the Olympics?
- Which countries have consistently excelled in the Olympics, and how has their performance trended over time?
- What are the most popular sports and events in the Olympics, and how have they changed in terms of participation and prominence?

- Which cities have hosted the most successful Olympics or Olympics with highest participation?

In addition to these initially proposed questions, new inquiries arose during the project, such as:

- Country Performance in Olympics: Specifically, which countries tend to perform better in the Winter or Summer Olympics?
- Sport-specific Analysis: For specific sports, like Gymnastics, what is the longest duration between two consecutive medals won by countries?
- Exploring Least Played Sports: Which sports have the least participation in the Olympics?
- Top-performing Athletes: Who are the athletes with the highest number of medals?
- Olympic Participation Records: Who holds the record for participating in the highest number of Olympics?
- Cities with Multiple Olympic Hosting: In which cities were the Olympics held more than once?

  These additional questions further enhanced the depth and breadth of the analysis, providing a comprehensive understanding of various aspects of Olympic history.

Proposed Visualizations

To address these research questions effectively, I planned to create a range of visualizations using the R programming language. These visualizations included, but were not limited to:

1. Track changes in the age and gender distribution of athletes over the decades, revealing trends and shifts.

2. Visually represent the distribution of Olympic medals among countries, highlighting patterns and standout performers.

3. The historical growth of specific sports and events, allowing to discern changes in popularity over time.

4. The frequency of Olympics hosted by different cities, offering insights into the impact of hosting on various locations.

Over time, the project objectives or questions remained consistent, and visualizations were tailored to address these aspects. Various types of data visualizations, such as bar charts, histograms, pie charts, and area charts, were used to represent the above information.

In the following section, we will delve into each data visualization in detail, discussing the specific chart chosen for each use case, the rationale behind selecting that particular chart, and the insights derived from the data visualization implementation.

Data Visualizations

1. **Top 10 Regions with the Highest Number of Medals Over Time**

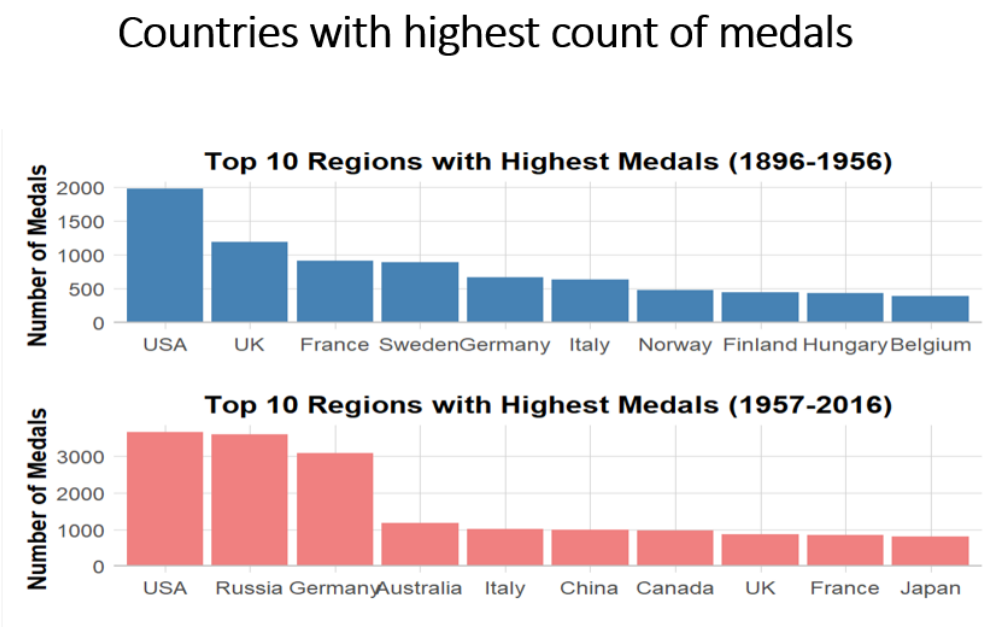

Chart Type: Grouped bar charts for two time periods (1896-1956 and 1957-2016).

Reasoning:

- **Comparison of Categories:** Grouped bar charts are effective for comparing multiple categories (regions) across different time periods. In this case, it allows for a clear visual comparison of medal counts for each region over two distinct time ranges.
- **Highlighting Trends:** The chart enables the audience to easily identify trends, shifts in dominance over the years.

Insights:

- USA consistently held the top spot across the entire 120 years.
- Russia gained prominence in the later period, securing the second position.
- Germany moved to the third position in the later years.
- The UK and France dropped from 2nd and 3rd to 8th and 9th places.

2. <u>Top 5 Countries with most and least medals</u>

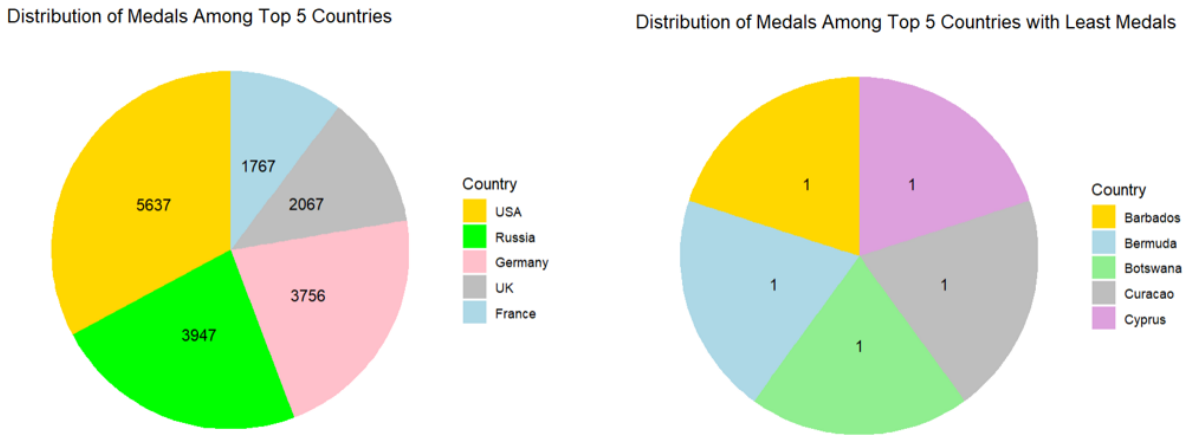

Chart Type: Two Pie Charts comparing the countries with most and least count of medals

Reasoning:

- **Visual Comparison of Proportions:** Pie charts are effective for visually comparing the proportions of different categories. In this case, they allow for a quick and intuitive assessment of how the medal counts are distributed among the top and bottom countries.

- **Simplicity:** Pie charts are simple and easy to understand, making them accessible to a broad audience. The clear representation of each country's share of the total medal count provides a straightforward way for viewers to grasp the information without the need for complex data interpretation.

Insights:

- The USA is the country with the highest number of Olympic medals.
- The USA is the only country to win more than 5,000 medals in the Olympic Games.
- Russia and Germany stand at the 2nd and 3rd positions for winning the highest number of medals.
- Barbados, Bermuda, Botswana, Curacao, and Cyprus are countries with the least number of medals, i.e., 1.

3. Comparison of Medals Won by Regions in Summer & Winter

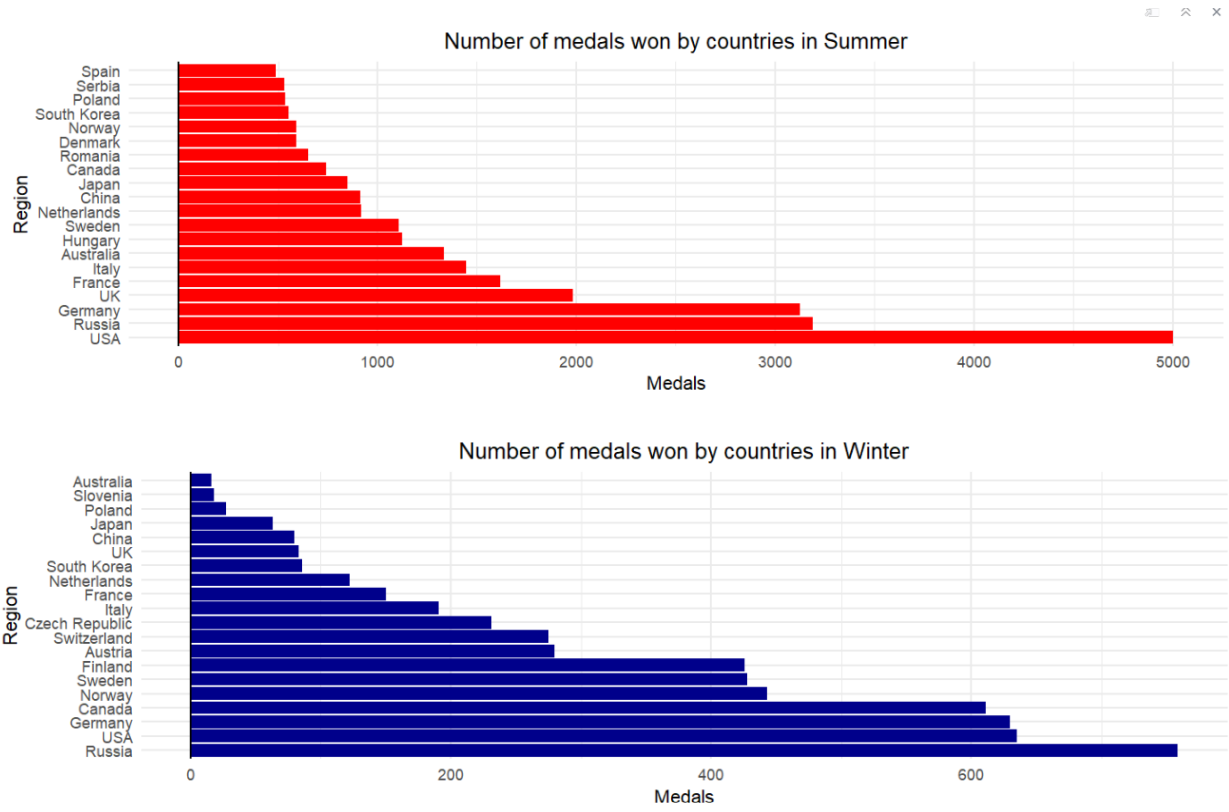

Chart Type: Two grouped bar charts comparing the number of medals won by regions in Summer and Winter.

Reasoning:

- **Comparison Across Categories:** Grouped bar charts are chosen to compare medal counts for different regions in both Summer and Winter Olympics. It facilitates a side-by-side visual analysis of the performance of regions in the two seasons.
- **Highlighting Differences:** The chart allows for the identification of regions that excel in either Summer or Winter, or those with consistent performances in both seasons.

Insights:

- USA dominates in Summer, while Russia excels in Winter.
- Germany maintains a consistent performance in both seasons.

4. <u>Gap Between Medals for Gymnastics in Different Countries</u>

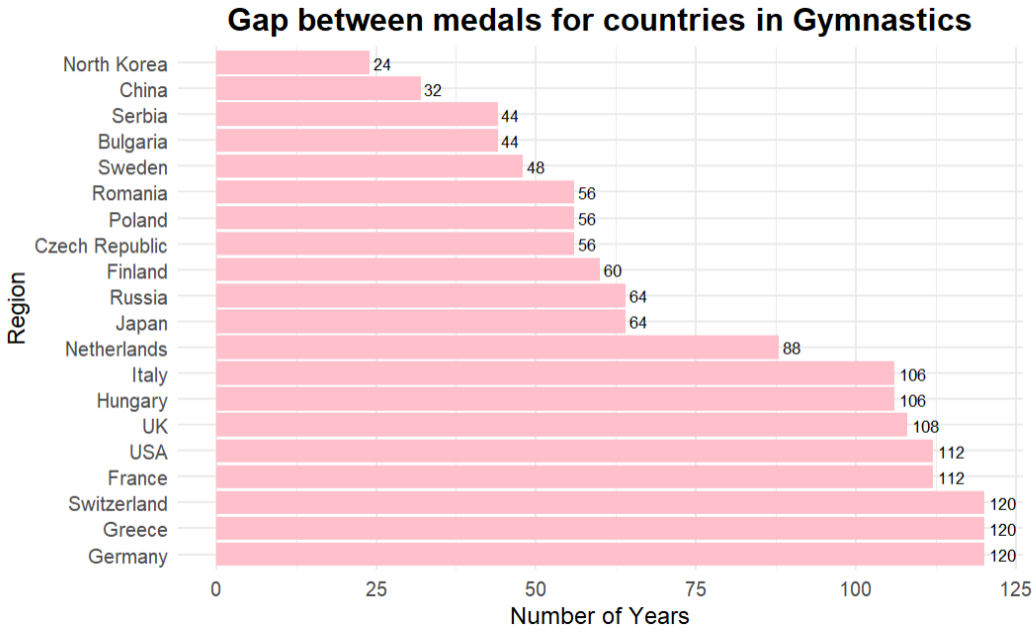

Chart Type: Bar chart showing the gap between the first and last medals for Gymnastics in different countries.

Reasoning:

- **Quantitative Comparison**: Bar charts are effective for comparing quantities. In this case, the chart visually represents the gap in years between the first and last medals won in Gymnastics by different countries.
- **Clear Visualization**: The bar chart provides a clear visual representation of the time gaps, making it easy to identify countries with significant intervals between their Gymnastics achievements.

Insights: Germany, Greece, and Switzerland have the largest gap of 120 years, while the North Korea has the smallest gap of 24 years.

5. Top 10 Most Famous Sports at Olympics

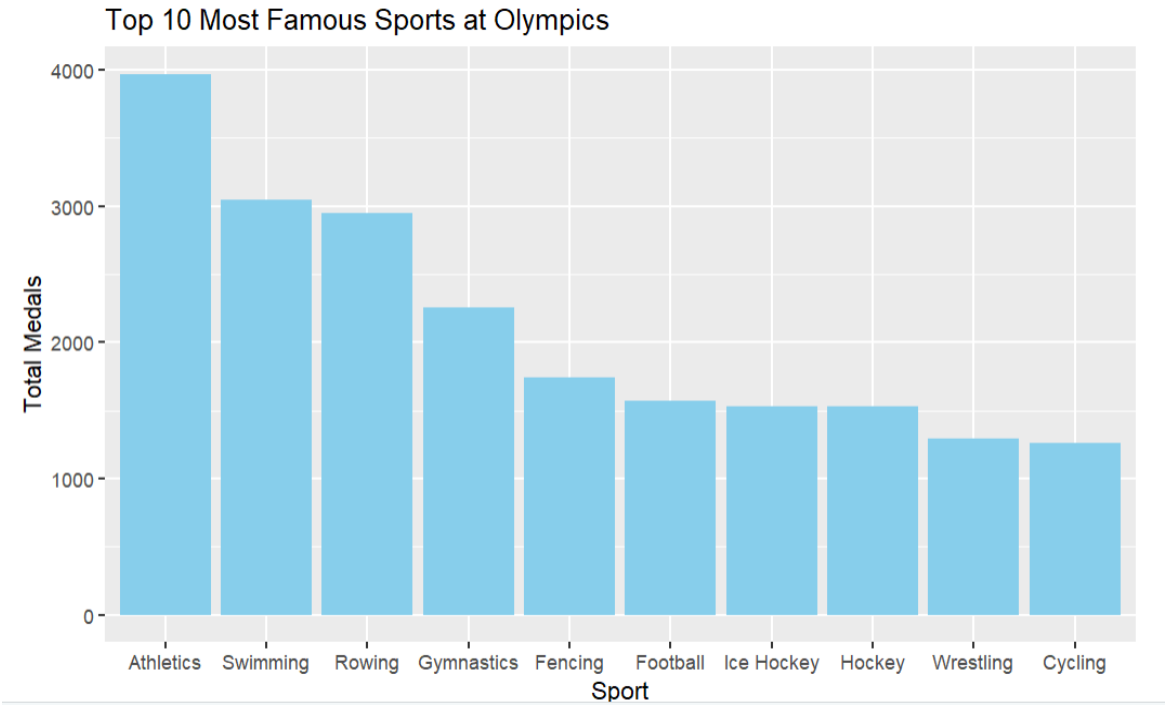

Chart Type: Bar Graph showing top 10 sports with highest number of medals

Reasoning:

- **Comparative Clarity**: Bar graphs allow for a clear visual comparison of the medal counts among the top 10 sports at the Olympics, aiding quick and intuitive analysis.
- **Ranked Presentation**: The inherent ranking order in bar graphs simplifies the identification of the most successful sports, offering a representation of their respective medal counts.

Insights:

- Athletics is the most renowned sport, boasting the highest number of won medals.
- Swimming and Rowing stand at 2nd and 3rd positions, respectively, while Wrestling and Cycling claim the 9th and 10th positions.

6.Top 10 Least Played Sports

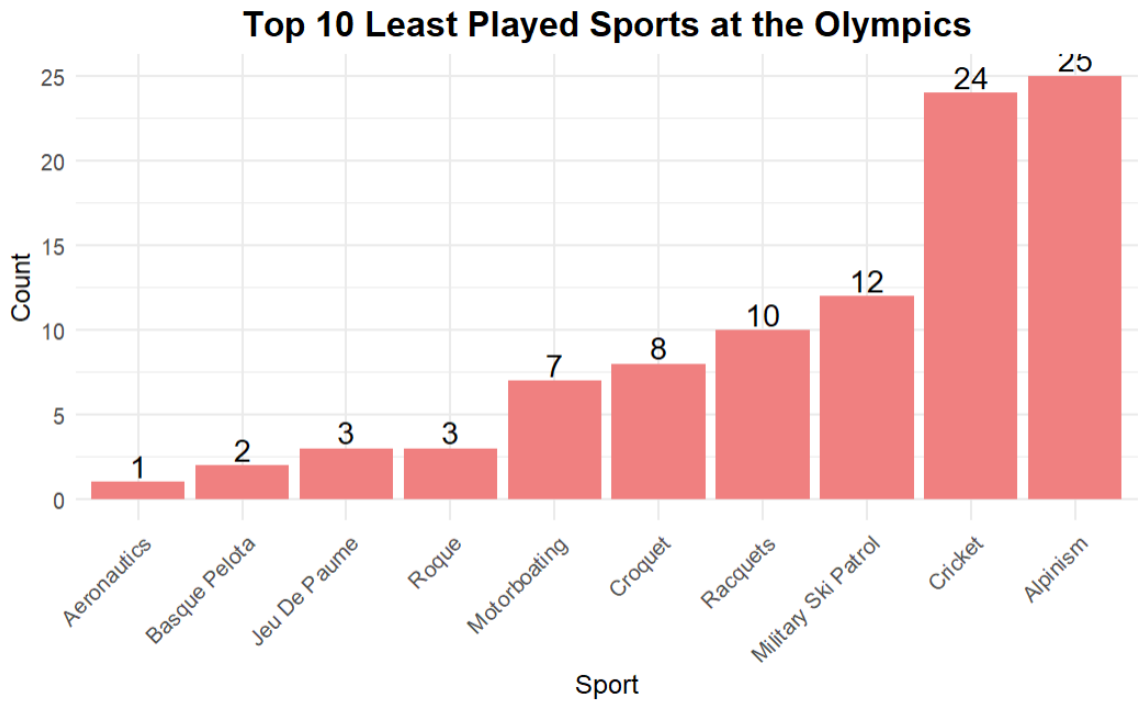

Chart Type: Bar Graph showing top 10 sports in least number of Olympics

Reasoning:

- **Comparative Analysis**: Bar graphs provide a clear visual comparison of the participation levels among the top 10 least played sports at the Olympics. The length of each bar directly represents the frequency of Olympic appearances, aiding quick and intuitive analysis.
- **Ranked Order Presentation**: Bar graphs inherently present data in a ranked order, allowing viewers to easily identify which sports have the lowest participation rates

Insights:

- Aeronautics is the least played sport in the Olympics, having appeared only once.
- It's evident that Cricket and Alpinism were each played approximately 25 times before being discontinued from the Olympic lineup.

7. Most Participated Sport in Olympics Every Year

Chart Type: Table showing the most participated sport in each Olympic year.

Reasoning:

- **Categorical Display:** A table is chosen for its simplicity and effectiveness in displaying categorical information. It provides a straightforward representation of the most participated sport each year without unnecessary complexity.
- **Easy Reference:** A table format is easy to read and allows the audience to quickly reference the information for each Olympic year.

Insights:

- Athletics consistently remained the most contested sport.
- Art Competitions had the highest participation in 1932 before being removed from the Olympics.

Most participated Sport in Olympic Games every year

Year	Sport	Participation
1896	Athletics	106
1900	Fencing	317
1904	Gymnastics	458
1906	Athletics	470
1908	Athletics	778
1912	Athletics	962
1920	Athletics	849
1924	Athletics	1003
1928	Athletics	992
1932	Art Competitions	620
1936	Athletics	1007
1948	Gymnastics	1060
1952	Gymnastics	2391
1956	Athletics	1013
1960	Gymnastics	1746
1964	Gymnastics	1484
1968	Gymnastics	1496
1972	Athletics	1686
1976	Athletics	1297
1980	Athletics	1268
1984	Athletics	1674
1988	Athletics	2062
1992	Athletics	2054
1994	Cross Country Skiing	639
1996	Athletics	2386
1998	Cross Country Skiing	733
2000	Athletics	2468
2002	Cross Country Skiing	774
2004	Athletics	2175
2006	Cross Country Skiing	812
2008	Athletics	2244
2010	Cross Country Skiing	725
2012	Athletics	2278
2014	Cross Country Skiing	765
2016	Athletics	2508

8. **Top 10 Host Cities with the Highest Participation**

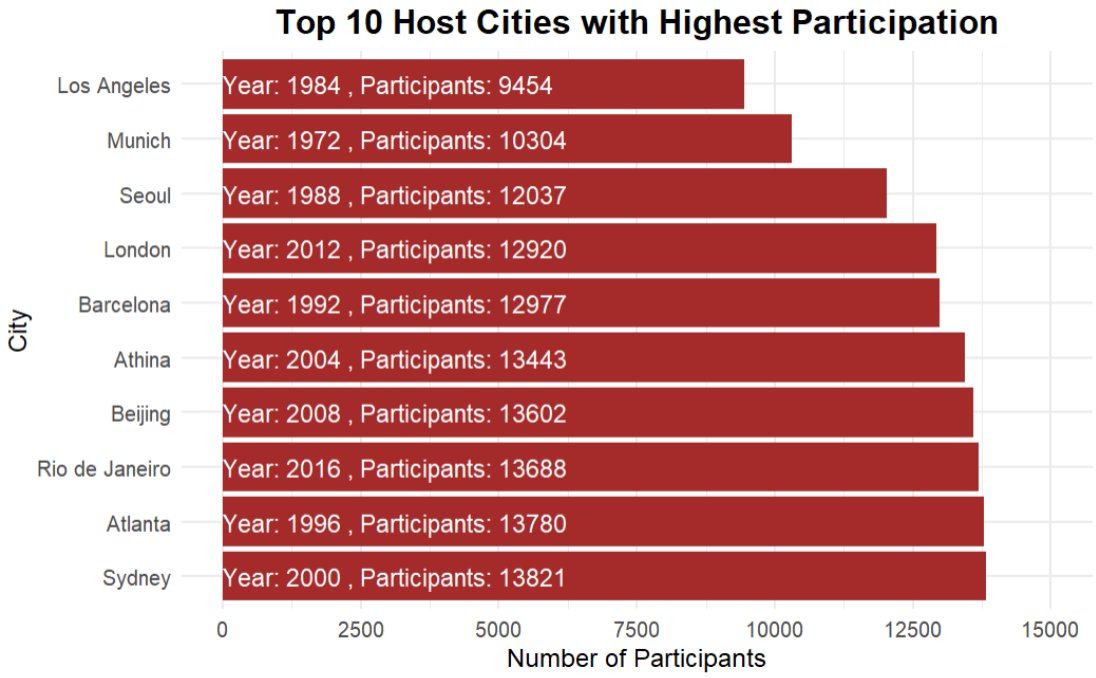

Chart Type: Horizontal bar chart displaying the top 10 host cities with the highest participation.

Reasoning:

- **Ranking and Comparison**: Horizontal bar charts are effective for ranking and comparing categories. In this case, it provides a clear ranking of host cities based on their participation, with the length of bars indicating the level of participation.
- **Focus on Top Performers**: The chart focuses on the top 10 host cities, offering a concise view of the cities with the highest Olympic participation.

Insights: Sydney, Atlanta, and Rio de Janeiro are among the top cities with the highest participation.

9. Cities with highest number of games hosted

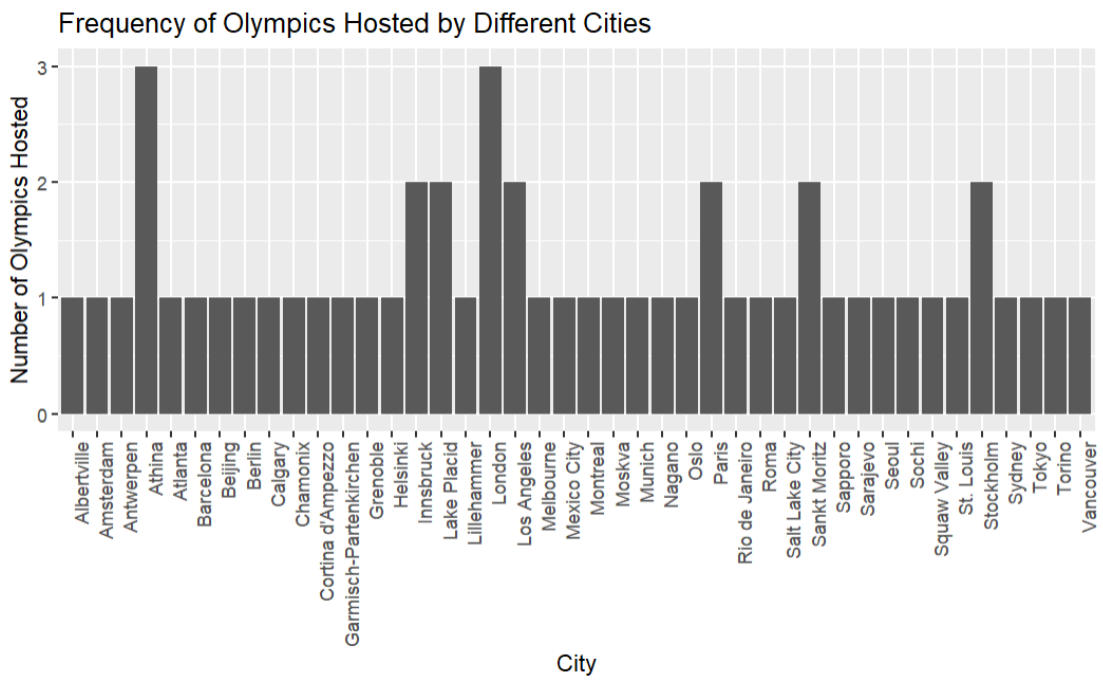

Chart Type: Bar chart depicting the frequency of Olympics hosted by different cities

Reasoning:

- **Comparative Analysis:** Bar graphs enable a clear visual comparison of the frequency of Olympics hosted by different cities. The length of each bar directly reflects the number of times a city has hosted the games, facilitating quick and intuitive comparisons.
- **Ranked Order Presentation:** Bar graphs inherently present data in a ranked order, allowing viewers to easily identify which cities have hosted the highest number of games. This natural ordering simplifies the interpretation of the information and helps convey the hierarchy of hosting frequency at a glance.

Insights:

- Athina and London hosted the highest number of games, i.e., 3.
- In most of the cities, the Olympics were held only once.

10. Gender Distribution of Athletes over the years

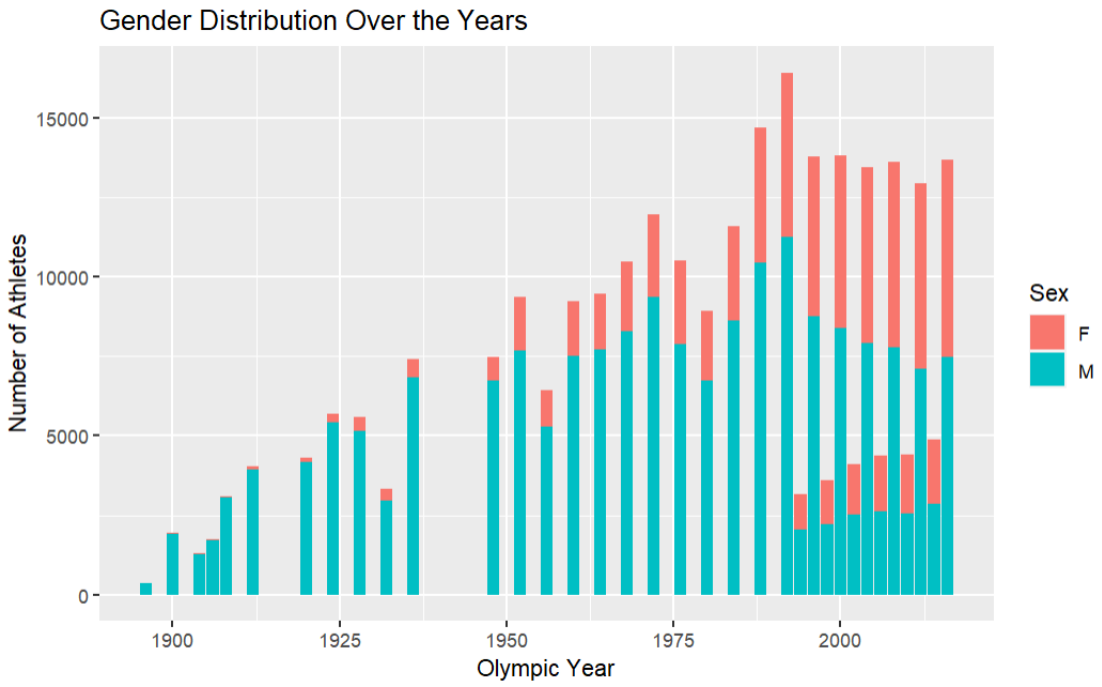

Chart Type: Stacked Bar Chart showing the number of male and female athlete frequencies in each Olympic game.

Reasoning:

- **Yearly Trend Comparison**: Stacked bar charts are optimal for comparing the gender distribution of athletes across multiple years, allowing for a clear visualization of evolving patterns.
- **Total and Gender-Specific Contributions**: Stacked bar charts effectively illustrate both the total athlete count and the contributions of male and female athletes in each Olympic game, providing a comprehensive view of gender distribution trends over the years.

Insights: Before 1950, the number of female athletes in the Olympics was very low compared to men. There was a significant increase in the frequency of female athletes in the Olympics after 1975, peaking in the 2000s.

11. Medals Won by Males/Females Over Time

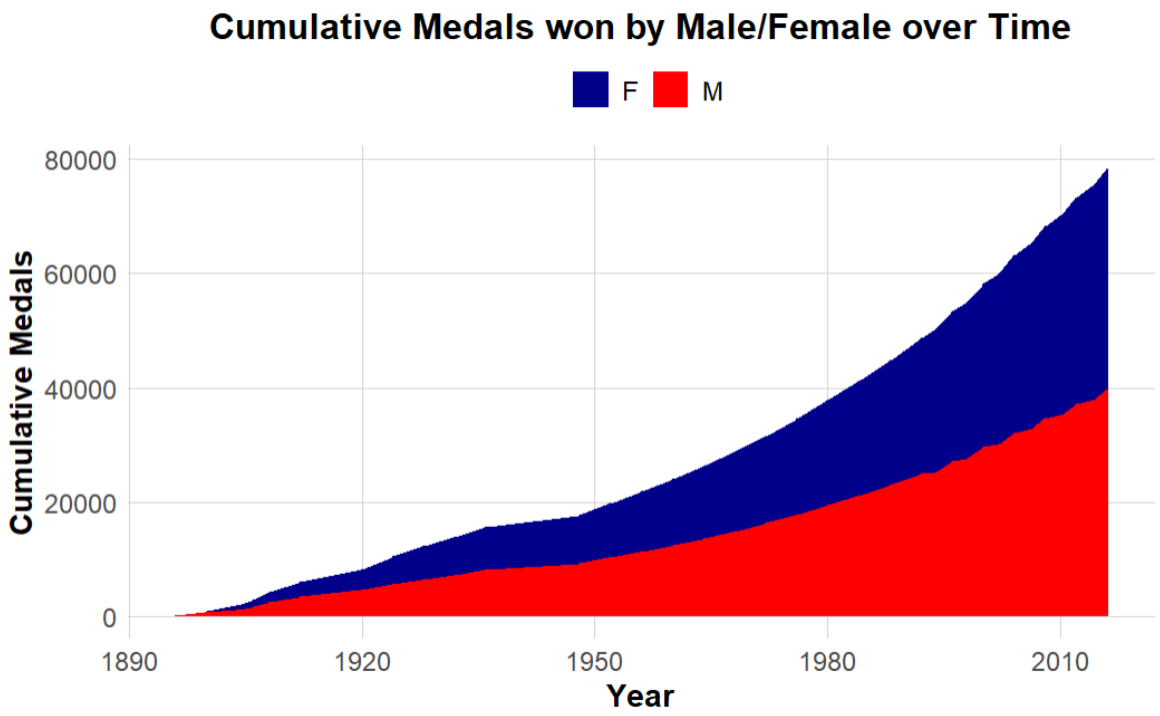

Chart Type: Area chart depicting cumulative medals won by male and female athletes over time.

Reasoning:

- **Cumulative Trends**: Area charts are suitable for displaying cumulative values over time. In this case, it effectively illustrates the cumulative number of medals won by male and female athletes over the years.
- **Emphasis on Patterns**: The filled area between the two lines emphasizes the overall pattern of gender-based medal achievements, making it easy to observe trends and shifts.

Insights:

- Gradual increase in the number of medals won by female athletes over time.
- Male athletes tend to outnumber female athletes, with fluctuations over time.
- Clear distinction in points after 1994 when Summer and Winter Olympics were split.

12. **Distribution of Athlete Ages**

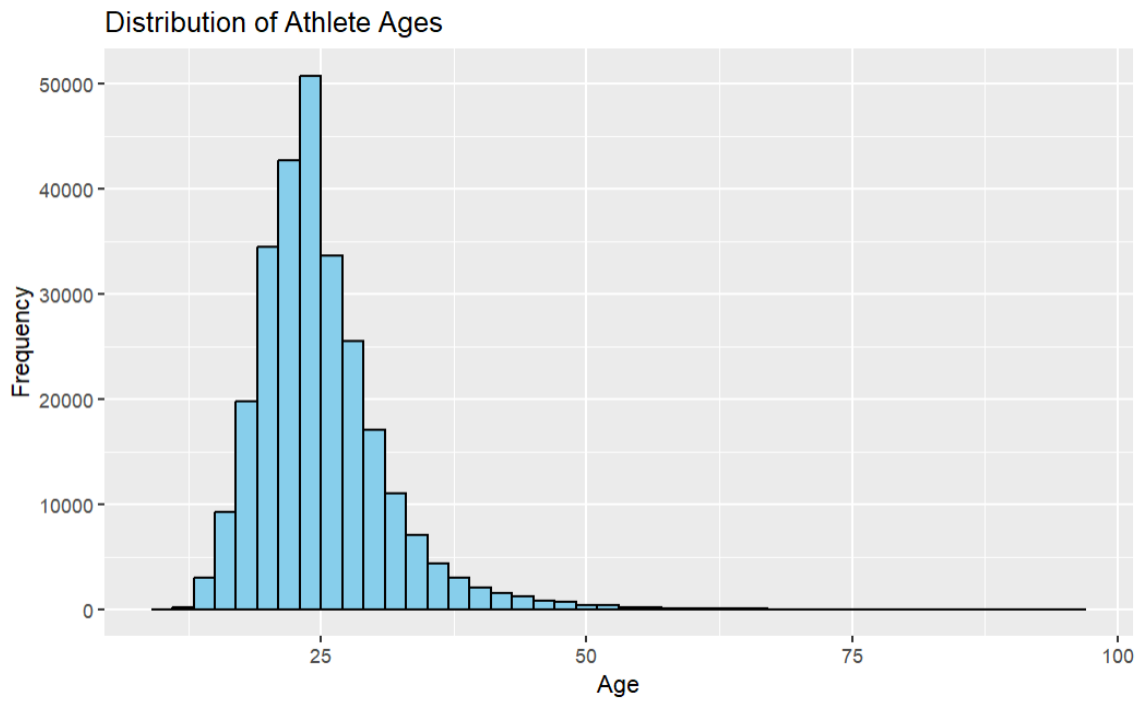

Chart Type: Histogram depicting frequency of athletes in an age group

Reasoning:

- **Frequency Display**: Histograms effectively present the frequency distribution of athlete ages, allowing for clear insights into the distribution pattern across various age groups.
- **Pattern Visualization**: Histograms are adept at visually representing patterns and trends in data, providing a quick and insightful overview of how athlete ages are distributed and any characteristic shapes or features in the distribution.

Insights: The athlete frequency steadily rises within the 0 to 25 age range, peaking at around 25 years. Subsequently, it gradually declines from 25 to 50 years, with minimal representation of athletes beyond the age of 50.

13. **Age Distribution of Male and Female Athletes**

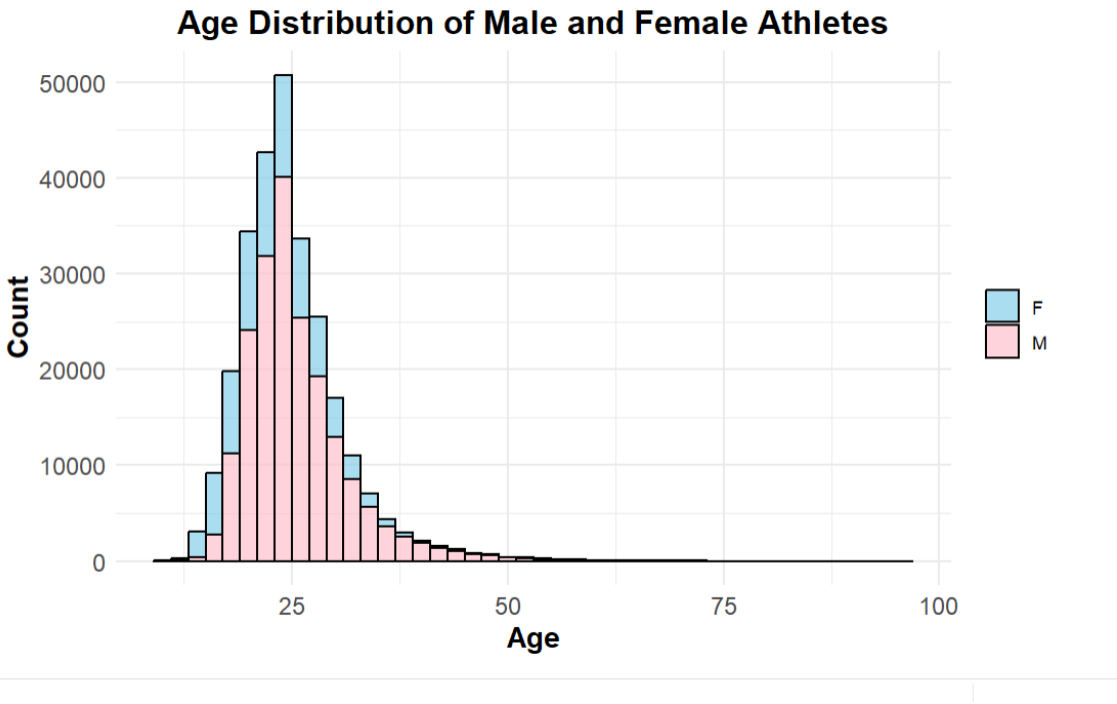

Chart Type: Stacked Histogram depicting frequency of male and female athletes in an age group

Reasoning:

- **Comparative Visualization**: A stacked histogram facilitates a direct visual comparison of age distributions between male and female athletes within each group.
- **Total Distribution Insight**: Stacked histograms provide a comprehensive view, illustrating both individual male and female age distributions as well as the overall composition of athletes across all age groups.

Insights: The frequency of male and female athletes steadily increases in the 0 to 25 age range, reaching a peak around 25 years. Subsequently, it gradually declines from 25 to 50 years, with minimal representation of male athletes beyond the age of 50 and female athletes beyond the age of 40.

14. Athletes with the Most Number of Medals in Each Sport

Athletes with the Most Number of Medals in Each Sport

Sport	Athlete	Medals
Swimming	Michael Fred Phelps, II	28
Gymnastics	Larysa Semenivna Latynina (Diriy-)	18
Fencing	Edoardo Mangiarotti	13
Biathlon	Ole Einar Bjrndalen	13
Canoeing	Birgit Fischer-Schmidt	12
Athletics	Paavo Johannes Nurmi	12
Shooting	Carl Townsend Osburn	11
Archery	Gerard Theodor Hubert Van Innis	10
Equestrianism	Isabelle Regina Werth	10
Cross Country Skiing	Marit Bjrgen	10
Cross Country Skiing	Raisa Petrovna Smetanina	10
Cross Country Skiing	Stefania Belmondo	10
Short Track Speed Skating	Yang Yang	10

Chart Type: Table showing the athlete with the most medals in each sport.

Reasoning:

- **Structured Presentation**: A table is chosen for its structured presentation of information, listing the athlete with the most medals in each sport. It provides a concise overview without the need for a more complex visualization.
- **Easy Reference**: The table format allows for easy reference to identify top-performing athletes in each sport.

Insights:

- Michael Phelps holds the most medals (28) in Swimming.
- Several athletes have won 10 or more medals in different sports.

15. <u>Top 10 Athletes with the Highest Participation in the Olympics</u>

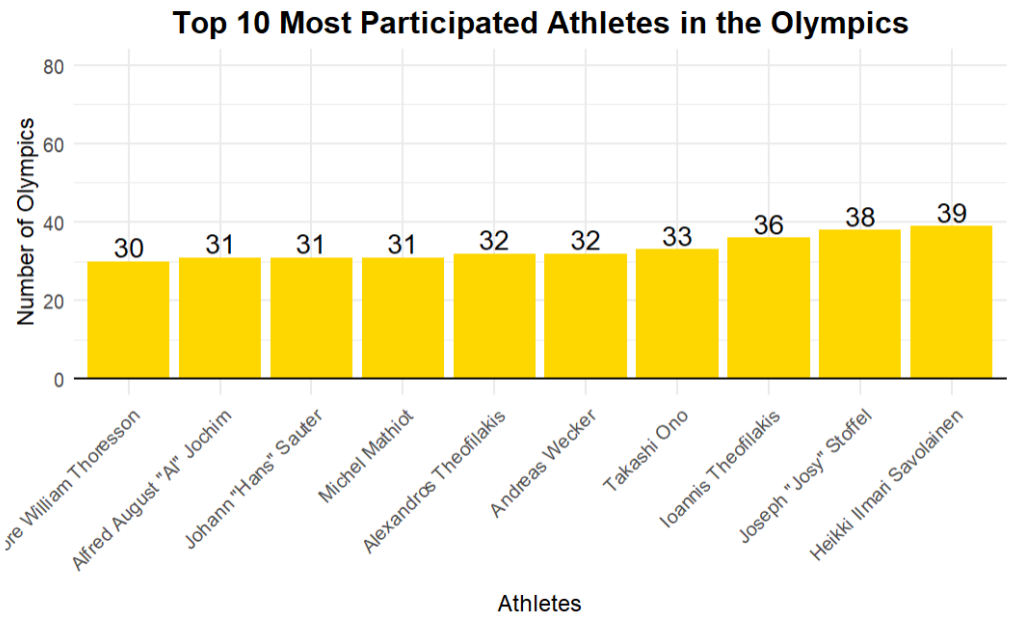

Chart Type: Horizontal bar chart showing the top 10 athletes with the highest participation.

Reasoning:

- **Ranking and Quantitative Comparison**: A bar chart is used to rank and quantitatively compare the participation of athletes. The length of bars directly represents the number of times an athlete participated.
- **Highlighting Extremes**: It effectively highlights the athletes with the highest participation, making it easy for the audience to identify the most prolific Olympic competitors.

Insights: Athletes participated more than 30 times, with Heikki Ilmari Savolainen having the most appearances (39 times).

These visualizations offer a comprehensive overview of Olympic trends, participation patterns, and outstanding performances over the years. the choice of visualization types is driven by the need to effectively convey specific types of information, whether it's trends over time, comparisons between categories, or the ranking of entities based on certain criteria. Each chart type serves its purpose in presenting data in a clear and understandable manner for the intended audience.

Overall Findings

Demographic Analysis:

- The age and gender distribution of athletes has evolved significantly over the 120-year history of the Olympics.
- Before 1950, female athlete participation was notably lower, but there was a significant increase after 1975, peaking in the 2000s.
- The frequency of male and female athletes steadily increases in the 0 to 25 age range, reaching a peak around 25 years.
- Gradual increase in the number of medals won by female athletes over time.
- Male athletes tend to outnumber female athletes, with fluctuations over time.

Country Performance:

- The USA consistently held the top spot in total medals over the entire period.
- Russia gained prominence in the later years, securing the second position, while Germany moved to the third position.
- The UK and France dropped from 2nd and 3rd to 8th and 9th places.

Sports Evolution:

- Athletics remains the most renowned sport, winning the highest number of medals.
- Swimming and Rowing are consistently strong, while Wrestling and Cycling claim the 9th and 10th positions in the most famous sports.
- Aeronautics is the least played sport in the Olympics, having appeared only once.
- Cricket and Alpinism were each played approximately 25 times before being discontinued from the Olympic lineup.

Host City Impact:

- Sydney, Atlanta, and Rio de Janeiro are among the top cities with the highest Olympic participation.
- Athina and London hosted the highest number of games, i.e., 3.

Top-performing Athletes:

- Michael Phelps holds the most medals (28) in Swimming.
- Several athletes have won 10 or more medals in different sports.
- Many athletes participated more than 30 times, with Heikki Ilmari Savolainen having the most appearances (39 times).

Conclusion

The exploration of Olympic history through data visualization has provided comprehensive insights into demographic trends, country performances, sports evolution, and the impact on host cities. The visualizations effectively addressed the research questions, offering a nuanced understanding of Olympic history. The project not only answered the initially proposed questions but also uncovered additional valuable insights, showcasing the richness and complexity of Olympic data. The choice of visualizations played a crucial role in conveying information clearly, and the structured data wrangling process ensured the reliability of the findings.

Overall, the project has successfully achieved its objectives, offering a compelling narrative of Olympic history through the lens of data visualization.