SportStats

02/08/2020

—

Sambangi Naveen

Table of Contents

Introduction

Olympics is considered as the most important event worldwide, which provides a common platform to players from various nations to show their talents. The 'modern Olympics' comprises all the Games from Athens 1986 to Rio 2016. The Olympics is more than just a quadrennial multi-sport world championship. It is a lens through which to understand global history, including shifting geopolitical power dynamics, women's empowerment, and the evolving values of society.

In this report, my goal is to shed light on major patterns in Olympic history. How many athletes, sports, and nations are there? Where do most athletes come from? Who wins medals? What are the characteristic of the athletes (e.g., gender and physical size)?

I also zoom in on some particularly interesting aspects of Olympic history that you might not know about. Did you know that Nazi Germany hosted the 1936 Olympics and they totally kicked everyone's asses? These are the sort of tidbits I like to sprinkle in.

Build on Project Proposal

Basic data analysis description

1. **Which client/dataset did you select and why?**

   I had selected Olympics Dataset - 120 years of data for analysing key insights in the data. My client was SportsStats which is a sports analysis firm. Insights could be patterns/trends highlighting certain groups/events/countries, etc. for the purpose of developing nations to perform well in upcoming Olympics events and the dataset is of compact size with the necessary information for drawing important information.

2. **Describe the steps you took to import and clean the data.**

   I have used pandas library to import the Olympics dataset files into jupyter notebook environment, then checked for the basic info about null values in the dataset and used mean and most frequent methods to fill the null values.

   There are some null values for height, weight, age and medals are filled accordingly with mean for height, weight and mode values for age. We have assumed a basic condition of whose filled with np.nan values in medals column was taken as they haven't won any medal.

```
df['Age'].fillna(value=df['Age'].mode()[0],inplace=True)
df['Height'].fillna(value=df['Height'].mean(),inplace=True)
df['Weight'].fillna(value=df['Weight'].mean(),inplace=True)
df.isnull().sum()
```

3. **Perform initial exploration of data and provide some screenshots or display some stats of the data you are looking at.**

I have used pandasql library to derive some initial insights of the data like number of medals of each category are won till now, the top 10 players who won more medals, the number of participants in each year of Olympics.

```
medal_count=pysqldf('''
SELECT Medal
    ,count(*) AS Medal_count
FROM df
Where Medal != "None"
GROUP BY Medal
''')
medal_count
```

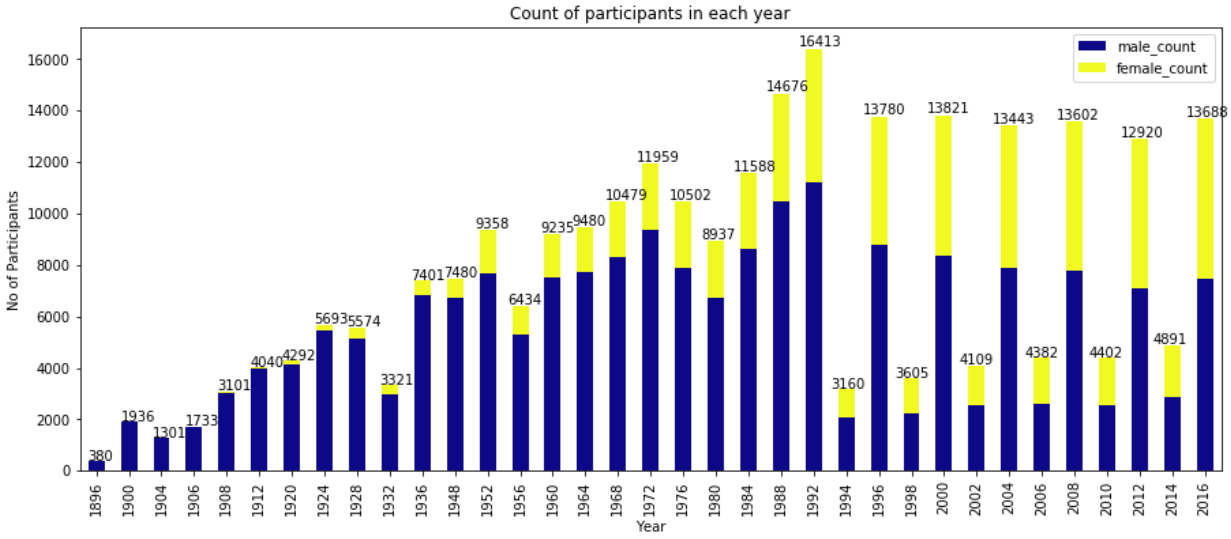
	Medal	Medal_count
0	Bronze	13295
1	Gold	13372
2	Silver	13116

To see the player with more number of medals using the query given below.

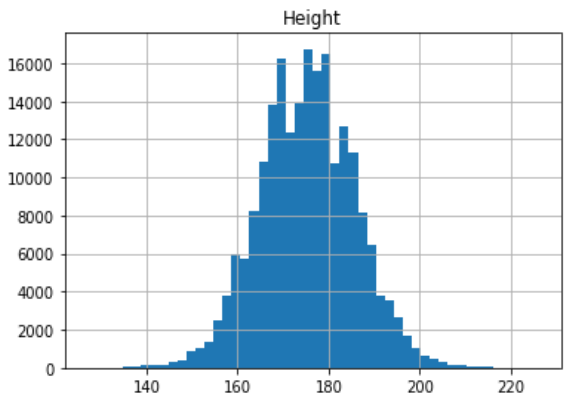
```
top_10=pysqldf('''
SELECT ID
    ,Name
    ,COUNT(*) as Count
FROM df
WHERE Medal IS NOT NULL
GROUP BY ID
ORDER BY Count DESC
LIMIT 10
''')
top_10
```

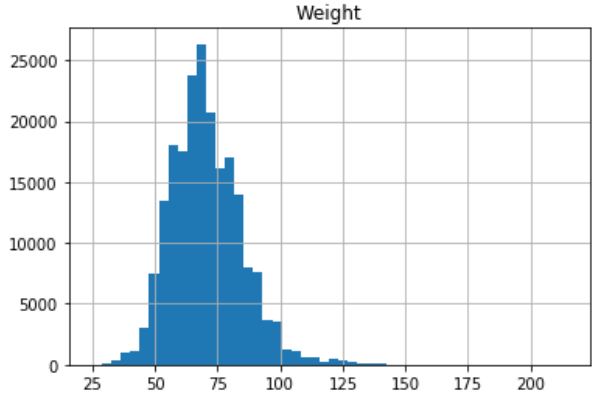
	ID	Name	Count
0	94406	Michael Fred Phelps, II	28
1	67046	Larysa Semenivna Latynina (Diriy-)	18
2	4198	Nikolay Yefimovich Andrianov	15
3	11951	Ole Einar Bjrndalen	13
4	74420	Edoardo Mangiarotti	13
5	89187	Takashi Ono	13
6	109161	Borys Anfiyanovych Shakhlin	13
7	23426	Natalie Anne Coughlin (-Hall)	12
8	35550	Birgit Fischer-Schmidt	12
9	57998	Sawao Kato	12

The below bar plot shows the number of participants participate over the years, where the bar in blue depicts the male count and the bar in yellow depicts female count.



The height and age distribution of various participants was shown below.





Audience

Write a 5-6 sentence paragraph describing your project; include who might be interested to learn about your findings. Who might be your audience?

The Olympic games are international sports events with more than 200 nations participating in various competitions. The Sportspersons from various countries participate in competitions and make their countries proud of their excellence in sports. Despite the massive population, many most populous countries fail to grab many medals at the Olympic games. The primary objective is to analyse the Olympic dataset using python to compare the overall performance of countries and to evaluate the contribution of each country in the Olympics. I

would like to work on men and women contribute to the Olympics, the sports that most of the people willing to participate and how to win more medals for the countries by analysing the winners. I would like to analyze the contribution of India in the Olympics to the country that won the most number of medals. Research analysts, sports analysts and newspapers are the most interested people who could go through my findings.

Questions to answer

1. What is the contribution of men and women in the Olympics games?
2. Which country has won number of medals?
3. Which sport has more willingness to participate in the Olympics games?
4. What was the contribution of India in the Olympics compared to the country that won most medals?
5. What was the best-performed sport for each country?
6. Is there any differences between the sports happening in summer to winter olympics?

Hypothesis

A hypothesis is a suggested solution for an unexplained occurrence that does not fit into current accepted scientific theory. The basic idea of a hypothesis is that there is no pre-determined outcome.

What is your initial hypotheses about the data?

I think men are the most participants in the Olympics compared to women as women were least encouraged during that time. But the ratio of winners to gender would be more for women to men as the women who participated in the Olympics during that difficult times could be more passionate and want to prove themselves. The country that won most medals could be USA or Russia those are the countries with serious competition between them. I think the number of participants are more likely to increase over the years as it became more and more outreach to the people by time. The indian contribution was could be around 10% as India was a vast country with huge population and many diversities.Our hypothesis Cricket and Football (Cricket and Football are the most played sport in recent years) being the top sports in the Olympics. I think both summer and winter sports could be the same with small differences.

Approach

I would be looking at the name (ID) to see the persons with more winning contributions, then the ROC column to compare the country with more winnings. Most important column to find the contributions of men and women to the Olympics is Gender. To check the relation between the summer and winter Olympics. I would like to evaluate the measure on simple comparisons and test my hypothesis using AB metric for statistical significance and normal average and total measures to get an insight about the data. I would like to join region data with sports data to get the country name and using different case statements to extract relevant data.

Technical Challenges

Interfacing with pandassql with minimal features making it difficult to use as the whole query is a string without any highlightings it would be difficult to manage, and find the misspelt words.

Entity-relationship diagram (ER diagram)

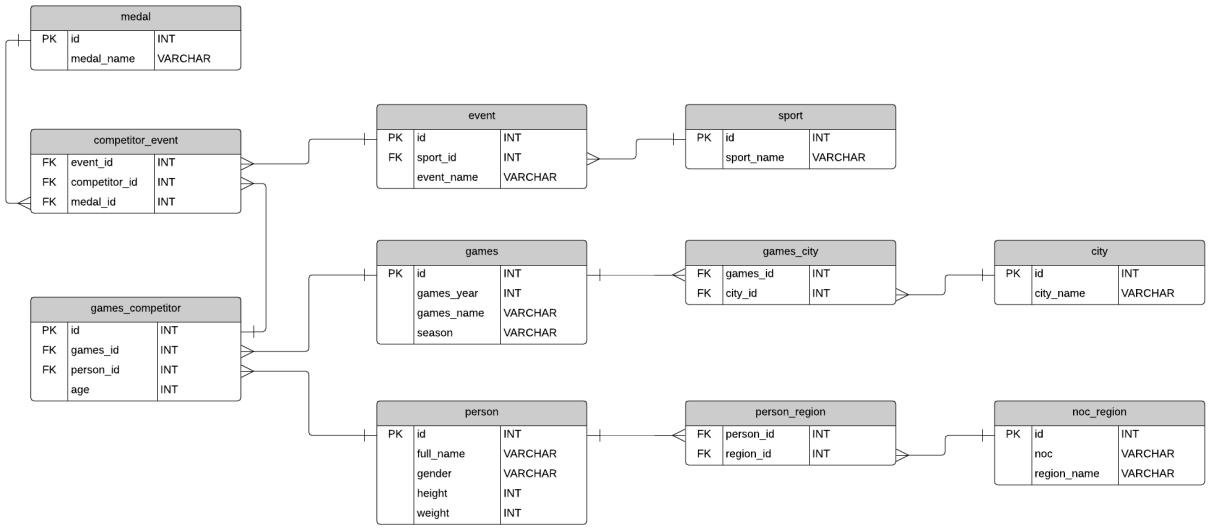


Table Explanations

Sport

This table contains a list of all sports in the Olympics: both summer sports and winter sports.

Event

The event table contains all of the different events for each sport. For example, if a sport is “Alpine Skiing”, then an event is “1KM Men’s Alpine Skiing”. There are multiple events per sport, and events are split into Men, Women, and Mixed.

City

This represents a list of many of the cities in the world.

Games

The games table lists all of the Olympic Games since 1896, the year they were held, and whether they were Summer or Olympic games.

Looking at this data, I learned that the Summer and Winter Olympics used to be in the same year. I've only ever known them to be in alternating years.

NOC Region

This contains a list of NOC (National Olympic Committee) codes and their names. This translates roughly to a country that competes in the Olympics.

Person

This table lists all people that have competed in Olympic games. It has their name, gender, height (in CM) and weight (in KG). The height and weight did not differ between Olympic games. If it did, it would have been stored in a different table.

Person Region

This is a joining table that lists all people and the NOC regions (countries) they competed for. This captures the fact that some people competed for more than one country, which is another scenario I didn't realise until I looked at the data.

Games Competitor

This table is a joining table that relates a person to an Olympic games, which shows who competed what each Olympic games.

Medal

This small table lists the different medals available: Gold, Silver, Bronze, and N/A.

Competitor Event

This table lists the combination of competitors (the people and the games they competed in), the event they competed in, and the medal (if any) they received. This is the largest table.

Correlations

Correlation is a term that is a measure of the strength of a linear relationship between two quantitative variables (e.g., height, weight). The most common measure of correlation is Pearson's product-moment correlation, which is commonly referred to simply as the correlation, the correlation coefficient, or just the letter r (always written in italics). The correlation coefficient r measures the strength and direction of a linear relationship, for instance:

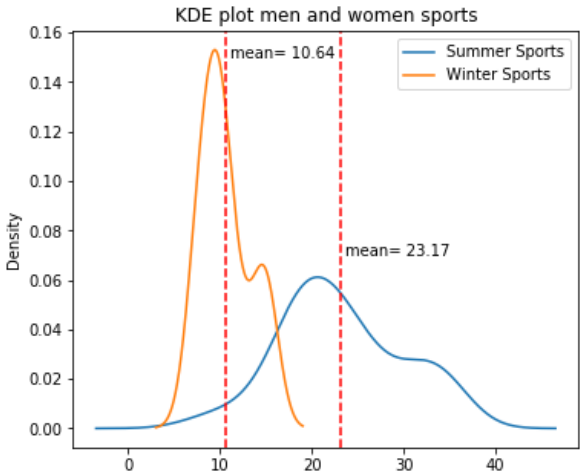
- 1 indicates a perfect positive correlation.

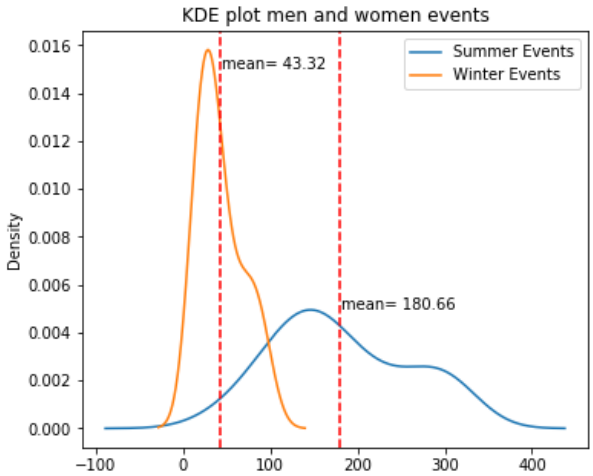
- -1 indicates a perfect negative correlation.
- 0 indicates that there is no relationship between the different variables.

Values between -1 and 1 denote the strength of the correlation

List the fields that you found to be correlated and describe what you learned from these correlations.

1. Correlation between the number of participants and the number of events each year
    a. Correlation between participants and events for men over each year is **0.9997512333242972**
    b. Correlation between participants and events for women over each year is **0.983446945772769**
    c. The above correlation values show that there is a heavy correlation between a number of participants and number of events as when people are more willing to participate, the event management committee adds more and more events to get the best out them. If there are more events and fewer participants, then many events are empty causes revenue loss for maintaining events and in the other case of more participants and fewer events, even though all events are full many participants didn't even get a chance to participate.
2. There is a moderate correlation between height and weight of a person (i.e0.7962130921162269) since we already know that height and weight grows progressively so there should be a positive correlation between height and weight.
3. The below plot shows the male and female distribution in sports and events. As hypothesized male has more sports and events compared to women. The mean value of sports for men is 23 and that of women is 10. The mean value of events for men is 180 and for women is 43.





Regression Analysis

Regression analysis is a powerful statistical method that allows you to examine the relationship between two or more variables of interest. While there are many types of regression analysis, at their core they all examine the influence of one or more independent variables on a dependent variable. I found a linear relationship between participants over years for both summer and winter seasons.

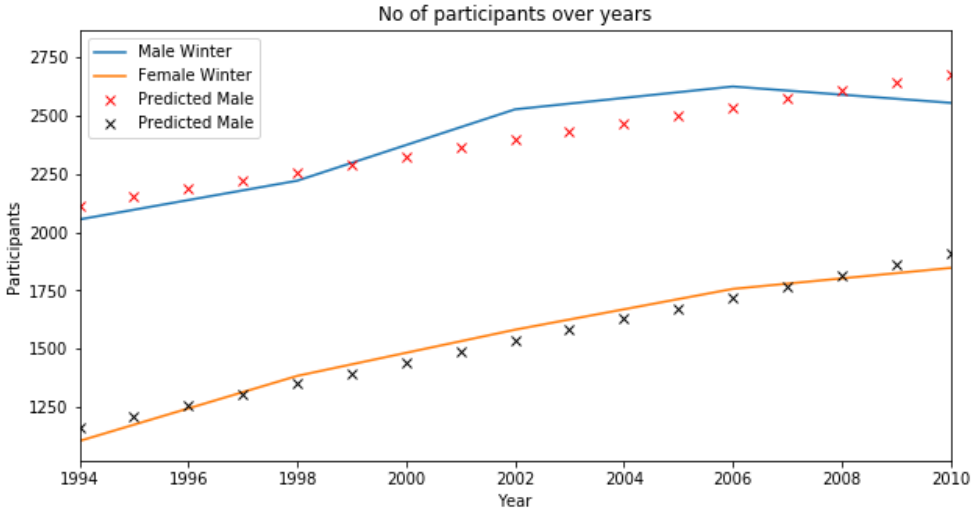
```
from sklearn.linear_model import LinearRegression
regressor1 = LinearRegression()
regressor1.fit(winter_after_1992['Year'].to_numpy().reshape(-1,1),winter_after_1992['Male Winter'].to_numpy().reshape(-1,1))
regressor2 = LinearRegression()
regressor2.fit(winter_after_1992['Year'].to_numpy().reshape(-1,1),winter_after_1992['Female Winter'].to_numpy().reshape(-1,1))
print('Number of male expected in 2014 winter olympics is ',regressor1.predict(np.array([[2016]]))[0][0])
print('Number of female expected in 2014 winter olympics is ',regressor2.predict(np.array([[2016]]))[0][0])
print('Number of male expected in 2018 winter olympics is ',regressor1.predict(np.array([[2018]]))[0][0])
print('Number of female expected in 2018 winter olympics is ',regressor2.predict(np.array([[2018]]))[0][0])
```

```
Number of male expected in 2014 winter olympics is  2888.0000000000146
Number of female expected in 2014 winter olympics is  2184.949999999997
Number of male expected in 2018 winter olympics is  2958.2000000000116
Number of female expected in 2018 winter olympics is  2277.800000000003
```

1. Predicting the number of participants in the year 2014 from the previous years.

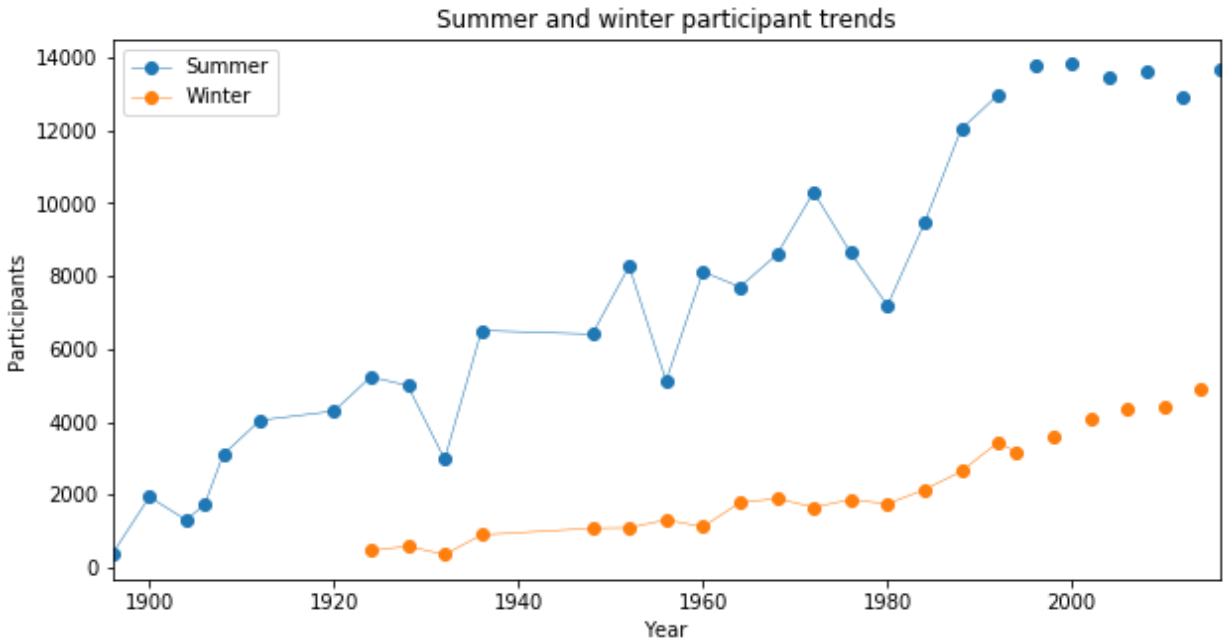
   Predicted number of male participants in 2014 = 2888
   Actual number of male participants in 2014 = 2868
   Error = *20*
   Predicted number of female participants in 2014 = 2184
   Actual number of female participants in 2014 = 2023
   Error = *161*
   Number of male expected in 2018 winter Olympics is 2958.2000000000116
   Number of female expected in 2018 winter Olympics is 2277.800000000003

By visualising the number of participants I have found a linear relationship between participants and years so used a simple linear regressor to fit the best line to predict future participants.

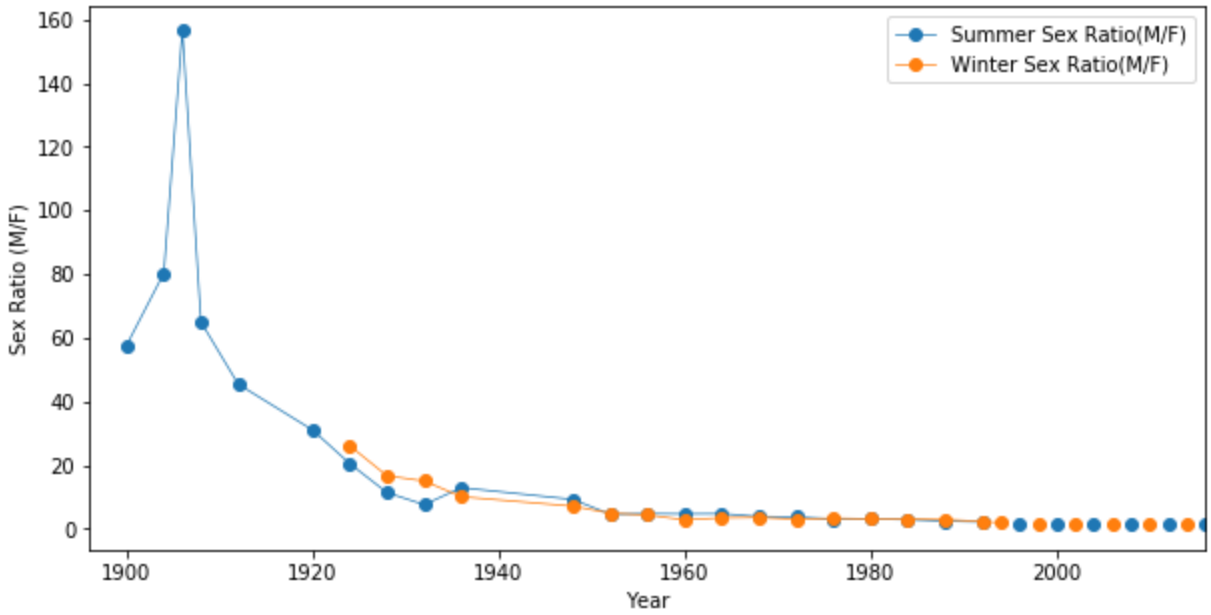


Visualizations

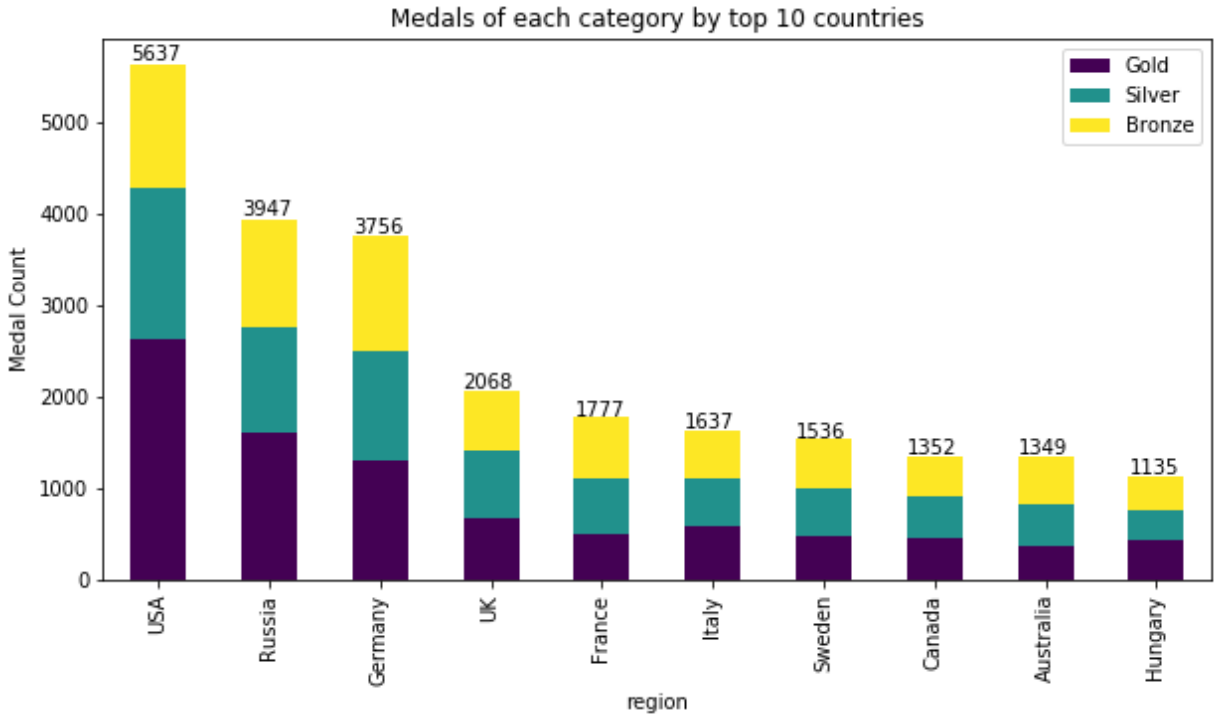
1. As the Olympics progressed over years more and more participants participated(i.e number of male and female participants grew.



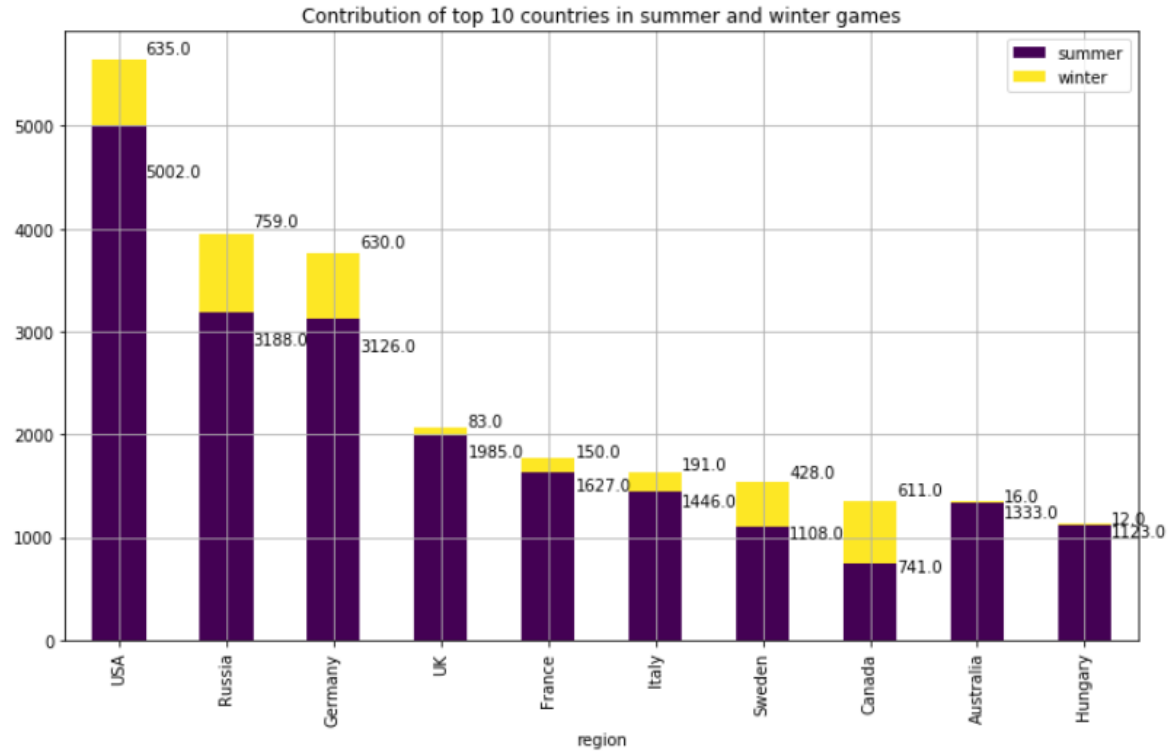
2. Sex Ratio of the male was very high during the initial year of Olympics, later the sex ratio became equal over the years(i.e close to 1)



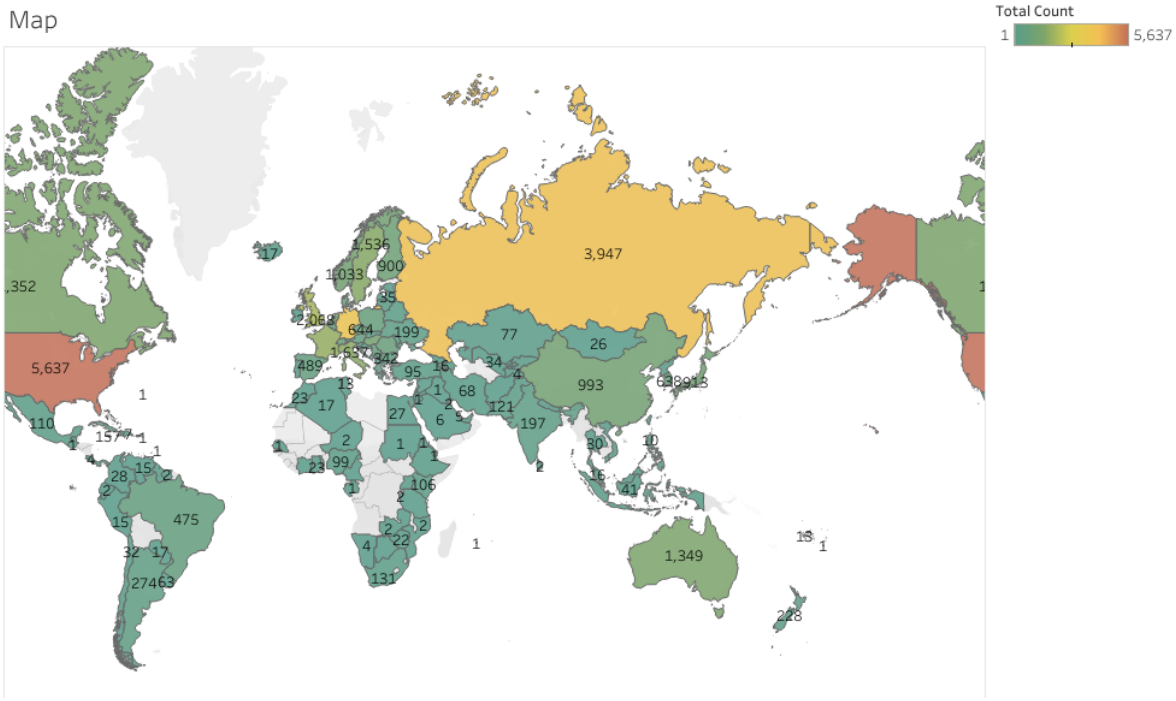
3. Below bar graph depicts the category of medals won by each country in all these years.



4. The below bar plot shows the contribution of medals in summer and winter sports by top ten countries.

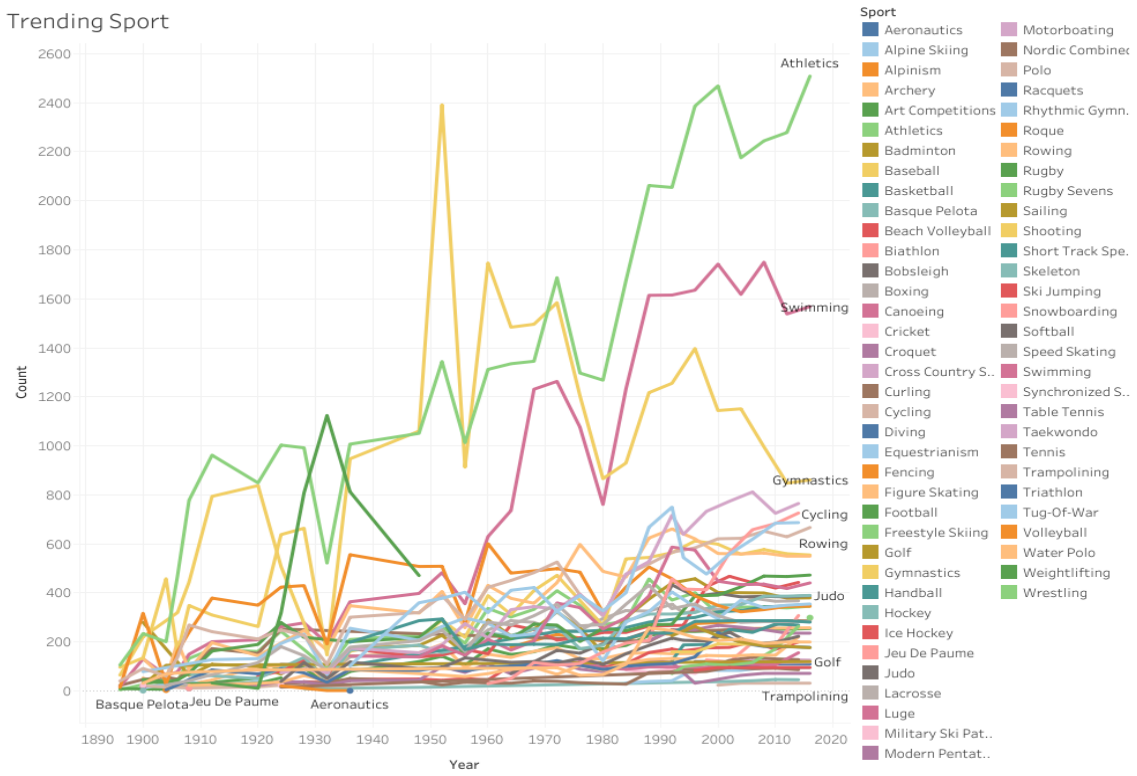


5. The below world map shows the colour legend of medals won by each country built using tableau online. As we thought earlier that, USA and Russia would be on the top with more number of medals was true. Here are the top 10 countries and their medals.

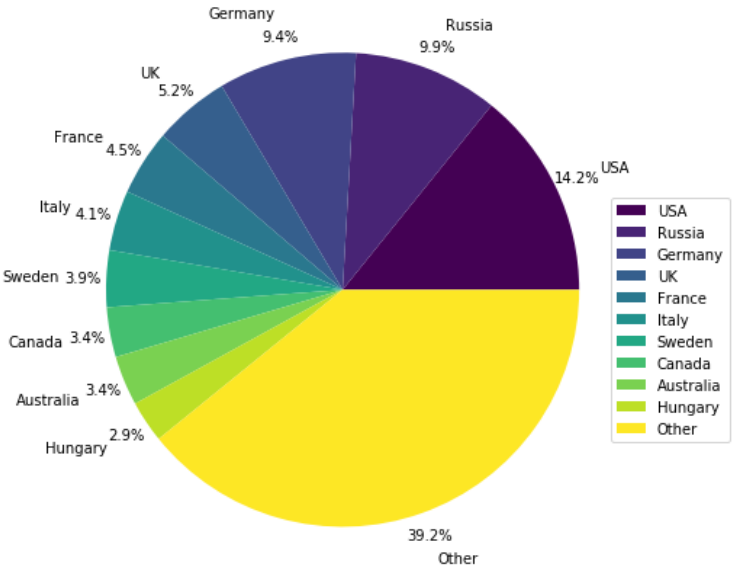


	Region	NOC	Gold	Silver	Bronze	Total Count
0	USA	USA	2638	1641	1358	5637
1	Russia	RUS	1599	1170	1178	3947
2	Germany	GDR	1301	1195	1260	3756
3	UK	GBR	678	739	651	2068
4	France	FRA	501	610	666	1777
5	Italy	ITA	575	531	531	1637
6	Sweden	SWE	479	522	535	1536
7	Canada	CAN	463	438	451	1352
8	Australia	AUS	368	459	522	1349
9	Hungary	HUN	432	332	371	1135

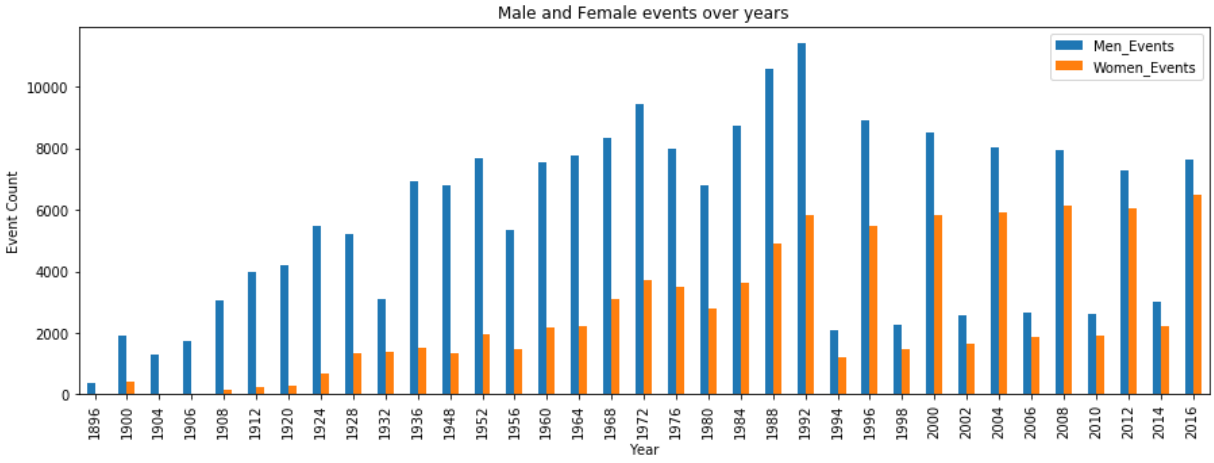
6. By grouping, each sport with year gave the number of participants participated each year for a particular sport to get the trends in the sport. By plotting the trends through line polt using tableau gave the top three sports that most people willing to participate.

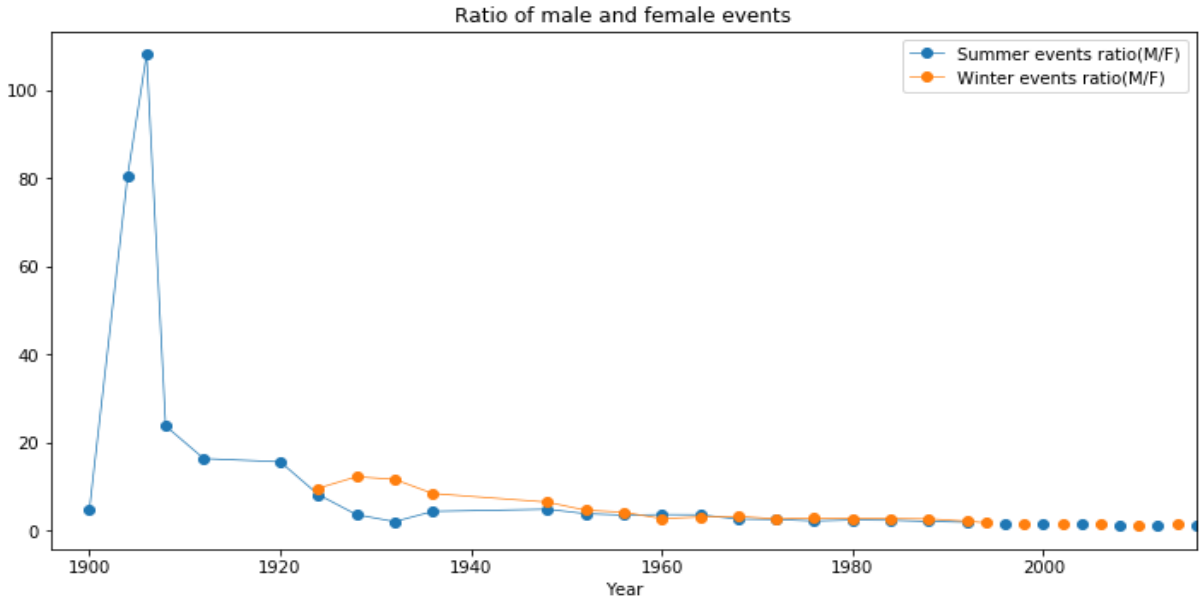


7. The below pie chart shows the contribution of medals in percentage by top countries. USA was leading with 14% of all the medals secured till now.



8. Male and female events over the years. The female events are very less during the initial years as time progressed both male and female events are comparable.





Discuss Insights Discovered

Hypothesis

- **Discuss your hypotheses and any direct outcomes from whether you were right or wrong. Did you change your hypotheses? Or create new ones?**
    1. Both men and women have an equal ratio of gold, silver, bronze medals. But the ratio of men with medals to total men participated was 0.145121. The ratio of women with medals to total women participated was 0.151002. It states that

women contribution was more than men even though they are less in number. So our initial hypothesis of women contribution high was proved.

2. The above world map shows the colour legend of medals won by each country built using tableau online. As we thought earlier that, USA and Russia would be on the top with more number of medals was true.
3. Our hypothesis Cricket and Football (Cricket and Football are the most played sport in recent years) being the top sports in the Olympics were wrong.
    I. Athletics
    II. Swimming
    III. Gymnastics
4.

	region	NOC	Gold	Silver	Bronze	Total Count
0	India	IND	138	19	40	197
1	USA	USA	2638	1641	1358	5637

The above table shows the number of medals won by USA and India. Medals won by USA are 28 times more than India. Our hypothesis of medals won by India to USA is 10 times was wrong and India being vast country with a huge population cannot able to compete with India.

5. The bar plot showing the contribution of medals in summer and winter sports by top ten countries. Our hypothesis of having an almost similar contribution in summer and winter sports was wrong and seems to be very less.

Metrics

Metrics are measures of quantitative assessment commonly used for comparing and tracking performance or production. Metrics can be used in a variety of scenarios. Metrics are heavily relied on in the financial analysis of companies by both internal managers and external stakeholders.

- **Discuss any metrics you created and why?**
    - **a.** I had used an average metric to check the percentage of events for men and women. Since we are thinking of equality in both men and women, I have used an average metric (i.e 50% of events belongs to men and 50% percentage events belong to women) to see the percentage of men and women events. Over the years the percentage of women years are gradually increasing and by 2016 both percentages are almost the same.

b. To calculate the percentage of medals by each country, I had used total medal count as a metric to compute the percentages for the top 10 countries and adding the medals of remaining countries as other.

Key Discoveries / Artifacts

- **Discuss discoveries about relationships in the data / themes discovered.**
  a. The number of athletes, events, and nations has grown dramatically since 1896, but growth levelled off around 2000 for the Summer Games.
  b. The Art Competitions were included from 1912 to 1948 and were dominated by Germany, France, and Italy. Nazi Germany was especially dominant in the 1936 Games.
  c. Geographic representation in the Games has grown since 1896, although Africa, Southeast Asia, the Middle East, and South America are still very under-represented.
  d. Female participation increased dramatically, and this trend started during the Cold War.
  e. Nazi women dominated the medals in 1936, East German and Soviet women dominated in 1976, and American women dominated in 2016.
  f. The size of Olympians has become more extreme over time. In most sports this means taller and heavier, but in a few sports such as gymnastics, athletes have become smaller.
  g. Every Olympics happened in only one city but the summer Olympics in 1956 happened in two different cities, Melbourne and Stockholm the reason I found in the internet was, It turns out that in 1956, the Summer Olympics was in Melbourne for most of the events, but due to Australia's strict horse quarantine rules, they couldn't bring the horses for the equestrian events. So they performed all of the Equestrian events in Stockholm earlier in 1956. So, the 1956 Olympics was in both Melbourne and Stockholm.

	Year	Season	City
0	1956	Summer	Melbourne
1	1956	Summer	Stockholm

Recommendations and Actions

Summarize the insights you found and make recommendations on what your client should do. What is the next steps or the action that should be taken as a result of your analysis?

1. **Winter vs Summer**: In all the above figures, we see that the values for winter games is comparatively lower than that of summer games. This is because of the less variety of winter games being available as well as the number of countries which participate in them being less, with the weather of Southern countries playing a part.
2. The size of Olympians has become more extreme over time. In most sports, this means taller and heavier, but in a few sports such as gymnastics, athletes have become smaller. The client should do some measures of low trending sports.
3. The number of countries participating in the Olympics has been increasing over the years.
4. Men events are more compared to women, and the participation of men is always more than women, so the sports committee should encourage more and more female participants.
5. The contribution of women in medals is more than men by comparing with medal count and participation count.