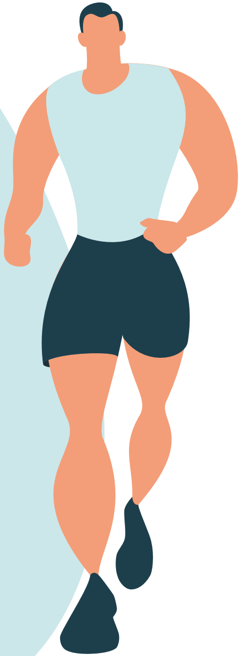
Sports Stats



Sports Stats

(Olympics Dataset - 120 years of data)

Coursera :**SQL for Data Science Capstone Project**
University of California, Davis

-Sethuraman B

TABLE OF CONTENTS

01 Review of Questions: Answer/Hypotheses/Approach

02

- Technical Challenges;
- Entity Relationship Diagram (ERD)

03

- Initial Findings
- Deeper Analysis

04 Hypotheses Results

#EXPLORE 1: Questions to Answer

1. Is there any correlation between the performance of a country in Winter Olympics and that in Summer Olympics?
2. Does country performance by year change more in Winter Olympics or Summer Olympics?
3. How has the Male: Female ratio evolved through time?

#EXPLORE 2 Initial Hypotheses

- Hypotheses 01: Yes;
- Hypotheses 02: Winter Olympics;
- Hypotheses 03: Decreased.

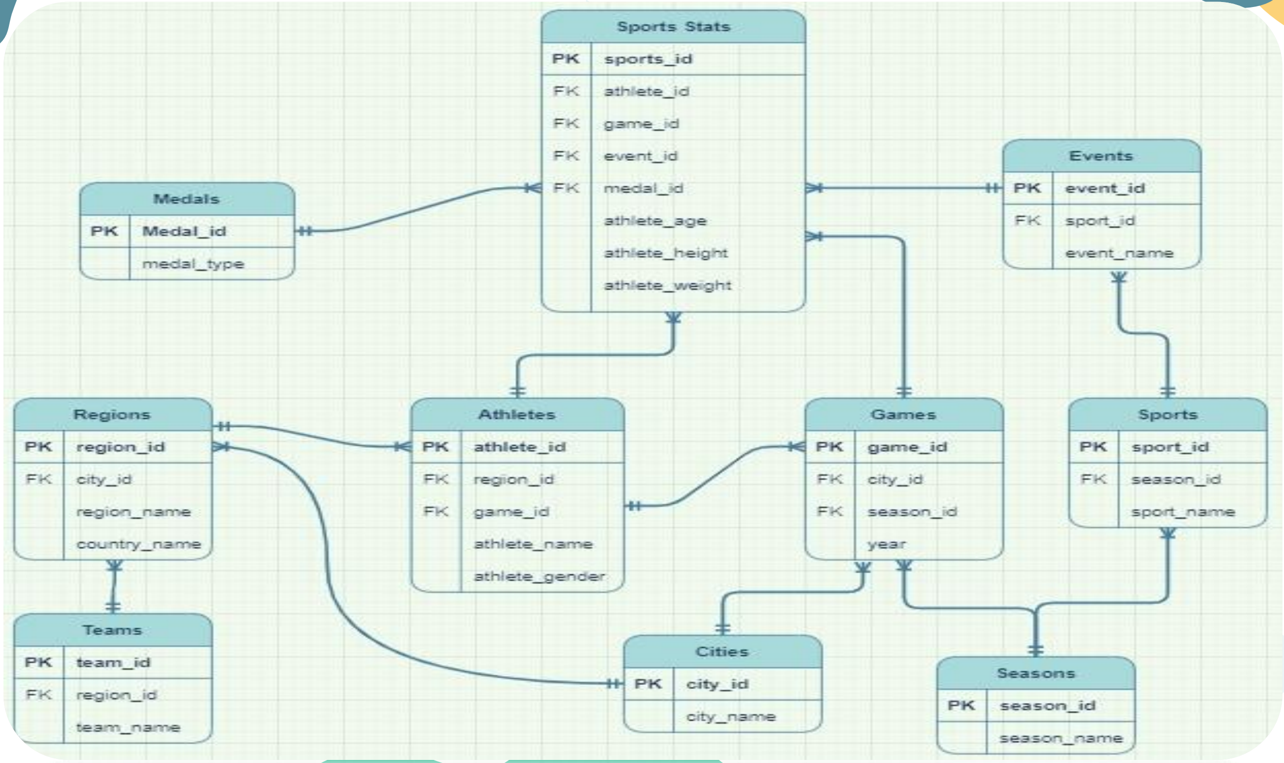
#EXPLORE 3: Data Analysis Approach

- To calculate the Pearson correlation coefficient.
- To calculate the standard deviation in country performance through years. A Comparison between average std of Winter and that of Summer Olympics will help.
- To draw a simple histogram.

Technical Challenges

- Encountered challenges with getting the starting year of the Summer Olympics different from that of the Winter Olympics;
- Limitation of Pandas for SQL (SQLite) made some SQL difficult to execute but manageable.

Entity Relationship Diagram (ERD)



Sports Stats	
PK	sports_id
FK	athlete_id
FK	game_id
FK	event_id
FK	medal_id
	athlete_age
	athlete_height
	athlete_weight

Medals	
PK	Medal_id
	medal_type

Events	
PK	event_id
FK	sport_id
	event_name

Regions	
PK	region_id
FK	city_id
	region_name
	country_name

Athletes	
PK	athlete_id
FK	region_id
FK	game_id
	athlete_name
	athlete_gender

Games	
PK	game_id
FK	city_id
FK	season_id
	year

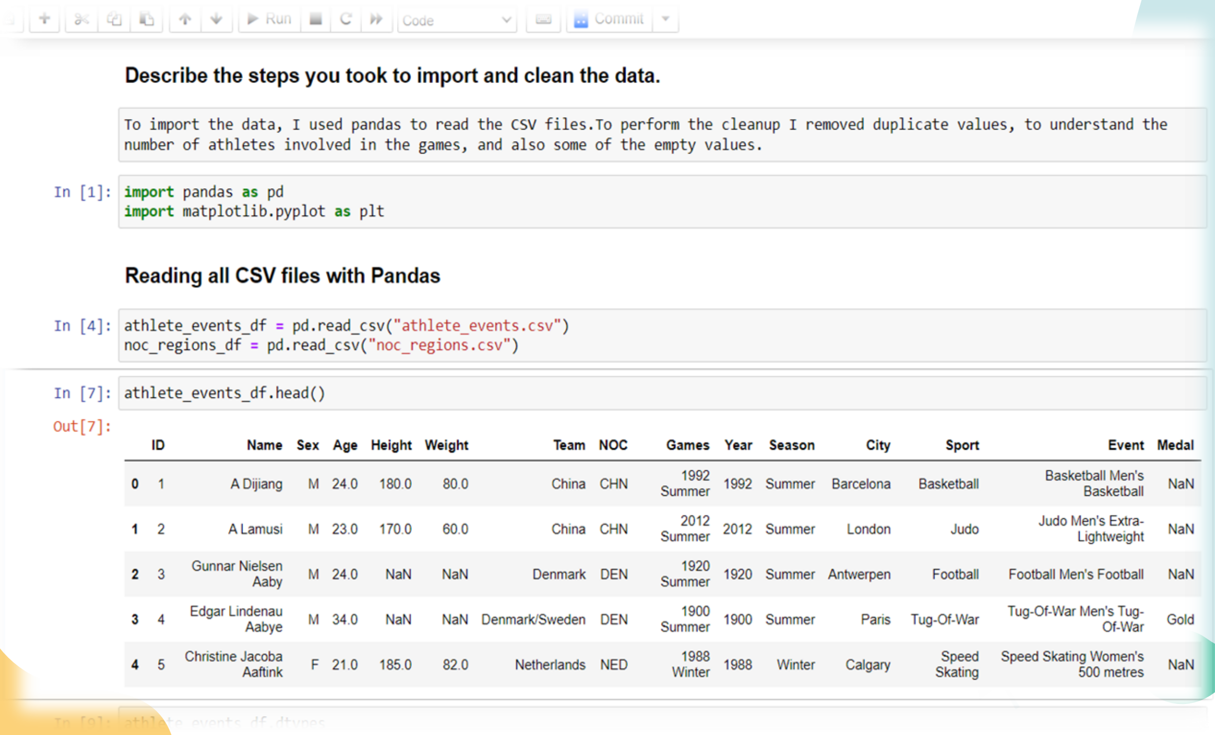
Sports	
PK	sport_id
FK	season_id
	sport_name

Teams	
PK	team_id
FK	region_id
	team_name

Cities	
PK	city_id
	city_name

Seasons	
PK	season_id
	season_name

Initial Exploration of data



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

```
To import the data, I used pandas to read the CSV files.To perform the cleanup I removed duplicate values, to understand the
number of athletes involved in the games, and also some of the empty values.
```

```
In [1]: import pandas as pd
        import matplotlib.pyplot as plt
```

Reading all CSV files with Pandas

```
In [4]: athlete_events_df = pd.read_csv("athlete_events.csv")
        noc_regions_df = pd.read_csv("noc_regions.csv")
```

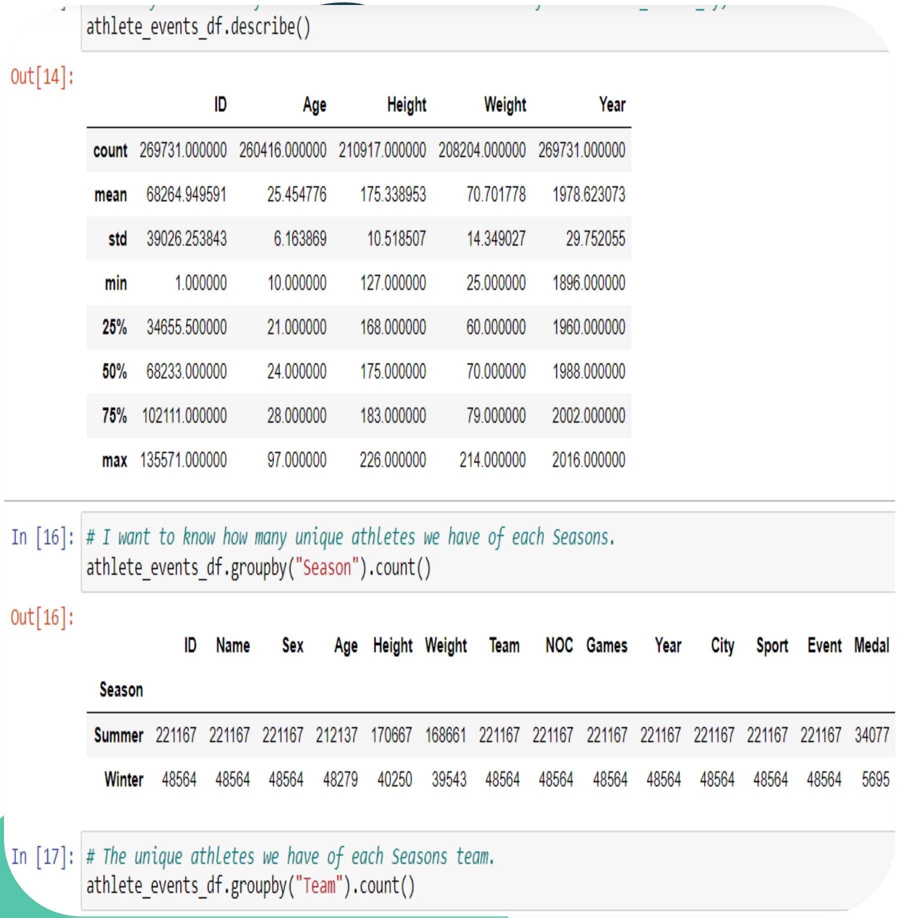
```
In [7]: athlete_events_df.head()
```

Out[7]:

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event	Medal
0	1	A Dijiang	M	24.0	180.0	80.0	China	CHN	1992 Summer	1992	Summer	Barcelona	Basketball	Basketball Men's Basketball	NaN
1	2	A Lamusi	M	23.0	170.0	60.0	China	CHN	2012 Summer	2012	Summer	London	Judo	Judo Men's Extra-Lightweight	NaN
2	3	Gunnar Nielsen Aaby	M	24.0	NaN	NaN	Denmark	DEN	1920 Summer	1920	Summer	Antwerpen	Football	Football Men's Football	NaN
3	4	Edgar Lindenau Aabye	M	34.0	NaN	NaN	Denmark/Sweden	DEN	1900 Summer	1900	Summer	Paris	Tug-Of-War	Tug-Of-War Men's Tug-Of-War	Gold
4	5	Christine Jacoba Aaftink	F	21.0	185.0	82.0	Netherlands	NED	1988 Winter	1988	Winter	Calgary	Speed Skating	Speed Skating Women's 500 metres	NaN

```
In [9]: athlete_events_df.dtypes
```

Initial Exploration of data



```
athlete_events_df.describe()
```

Out[14]:

	ID	Age	Height	Weight	Year
count	269731.000000	260416.000000	210917.000000	208204.000000	269731.000000
mean	68264.949591	25.454776	175.338953	70.701778	1978.623073
std	39026.253843	6.163869	10.518507	14.349027	29.752055
min	1.000000	10.000000	127.000000	25.000000	1896.000000
25%	34655.500000	21.000000	168.000000	60.000000	1960.000000
50%	68233.000000	24.000000	175.000000	70.000000	1988.000000
75%	102111.000000	28.000000	183.000000	79.000000	2002.000000
max	135571.000000	97.000000	226.000000	214.000000	2016.000000

In [16]:

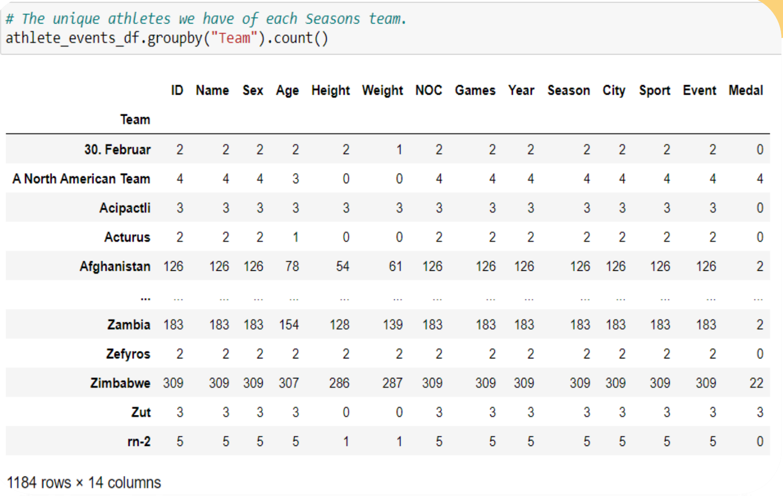
```
# I want to know how many unique athletes we have of each Seasons.
athlete_events_df.groupby("Season").count()
```

Out[16]:

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	City	Sport	Event	Medal
Season														
Summer	221167	221167	221167	212137	170667	168661	221167	221167	221167	221167	221167	221167	221167	34077
Winter	48564	48564	48564	48279	40250	39543	48564	48564	48564	48564	48564	48564	48564	5695

In [17]:

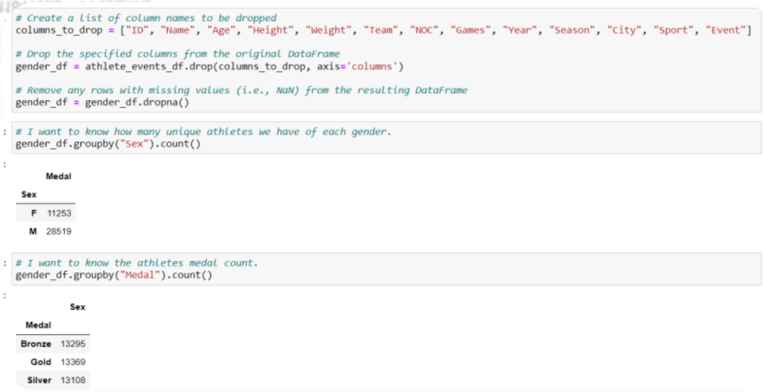
```
# The unique athletes we have of each Seasons team.
athlete_events_df.groupby("Team").count()
```



```
# The unique athletes we have of each Seasons team.
athlete_events_df.groupby("Team").count()
```

	ID	Name	Sex	Age	Height	Weight	NOC	Games	Year	Season	City	Sport	Event	Medal
Team														
30. Februar	2	2	2	2	2	1	2	2	2	2	2	2	2	0
A North American Team	4	4	4	3	0	0	4	4	4	4	4	4	4	4
Acipactli	3	3	3	3	3	3	3	3	3	3	3	3	3	0
Acturus	2	2	2	1	0	0	2	2	2	2	2	2	2	0
Afghanistan	126	126	126	78	54	61	126	126	126	126	126	126	126	2
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
Zambia	183	183	183	154	128	139	183	183	183	183	183	183	183	2
Zefyros	2	2	2	2	2	2	2	2	2	2	2	2	2	0
Zimbabwe	309	309	309	307	286	287	309	309	309	309	309	309	309	22
Zut	3	3	3	3	0	0	3	3	3	3	3	3	3	3
rn-2	5	5	5	5	1	1	5	5	5	5	5	5	5	0

1184 rows × 14 columns



```
# Create a list of column names to be dropped
columns_to_drop = ["ID", "Name", "Age", "Height", "Weight", "Team", "NOC", "Games", "Year", "Season", "City", "Sport", "Event"]

# Drop the specified columns from the original DataFrame
gender_df = athlete_events_df.drop(columns_to_drop, axis='columns')

# Remove any rows with missing values (i.e., NaN) from the resulting DataFrame
gender_df = gender_df.dropna()
```

```
# I want to know how many unique athletes we have of each gender.
gender_df.groupby("Sex").count()
```

	Medal
Sex	
F	11253
M	28519

```
# I want to know the athletes medal count.
gender_df.groupby("Medal").count()
```

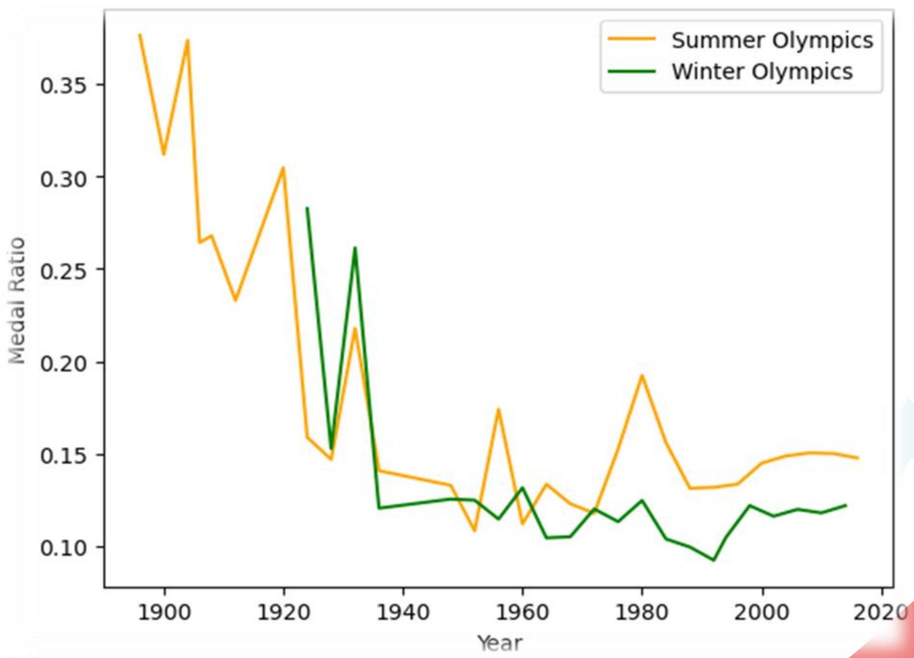
	Sex
Medal	
Bronze	13295
Gold	13369
Silver	13108

Initial Findings

- Although the ratio between the Summer Olympics and the Winter Olympics is indeed different, men happen to be dominant. My first assumption is that the ratio of women to men has increased over time. I began to dive into it.
- There are significant differences between male and female participants not only in terms of expected height and weight, but also in terms of age.
- The first two differences can be attributed to biology. Although the latter may require more than just: it is worth considering social factors at the same time.
- Another interesting fact is that the age gap in the Winter Olympics is much smaller (~2.8 years old and 1.5 years old)
- Another analysis of the number and ratio of medals is needed. I checked the ratio of total medal winners and the changes in the ratio of different medals:

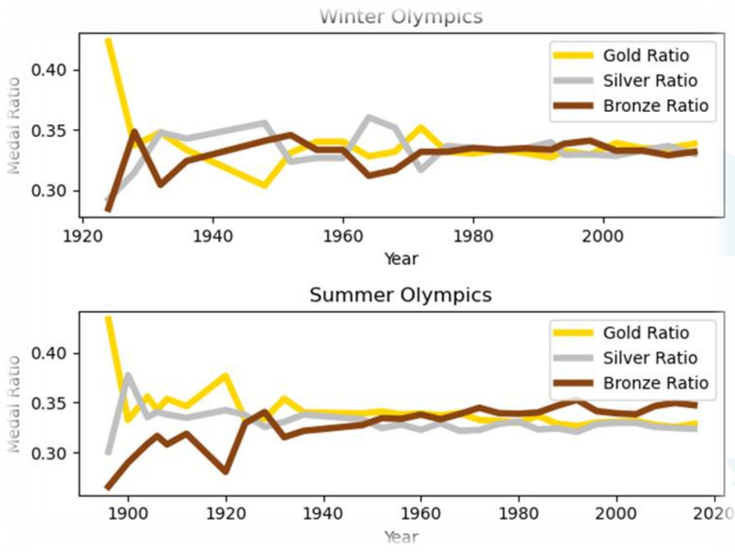
Findings

In the last century, the medal ratio fluctuated greatly in the two competitions, but eventually stabilized. This can be interpreted as establishing norms on these issues.



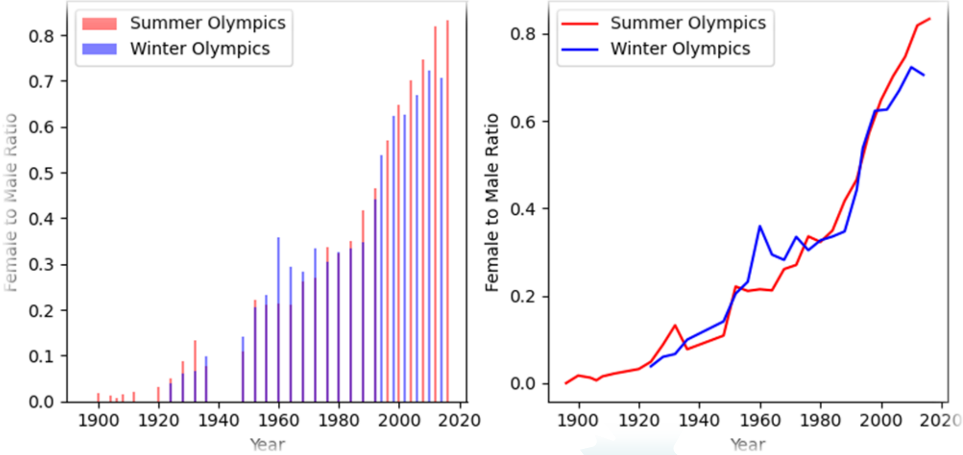
Findings

The relative percentages of gold, silver and bronze medals have also stabilized, which may be due to the reasons mentioned above.



Findings

This assumption seems to be correct. Over time, the ratio of women to men has indeed increased. However, there is an interesting detail: during the Second World War, the proportion of the Summer Olympics dropped sharply, but then it resumed its growth momentum. Without further analysis, I cannot explain this phenomenon.



Deeper Analysis

The length of the array of the number of medal count in the Winter Olympics and Summer Olympics are different because Winter Olympics started in 1924, but Summer Olympics started in 1896. Therefore, I have to create a new shortened table of the Summer Olympics started in 1924 to match the length of the Winter Olympics.

Deeper Analysis

The Pearson correlation coefficient between the total number of medals in the winter and Summer Olympics, from 1924 to 2016, is 0.94, which is highly positive. Therefore, the performance of a country in Winter Olympics is highly correlated to that in Summer Olympics

Then calculate the standard deviation in country performance through years. A Comparison between average std of Winter and that of Summer Olympics will help.

Deeper Analysis

- std_medal_count_summer_olympics = 475;
- std_medal_count_winter_olympics = 153;

From 1924 to 2016, as the standard deviation in the Summer Olympics is about 3 times that in the Winter Olympics, country performance by year change more in Summer Olympics.

Final Findings (Result of Hypotheses)

- Yes, the performance of a country in Winter Olympics is highly correlated to that in Summer Olympics;
- Yes, the country performance by year change more in Winter Olympics than that in Summer Olympics;
- The male: female ratio has decreased from 1896 to 2016.

Recommendations

1. The Olympiad Organizing Committee should devote more resource in the weather prediction to help organize the Olympics, as the weather affects the performance of athletes.
2. The Olympiad Organizing Committee should advocate the equality between male and female and keep encouraging more female to join the Olympics.

THE END!