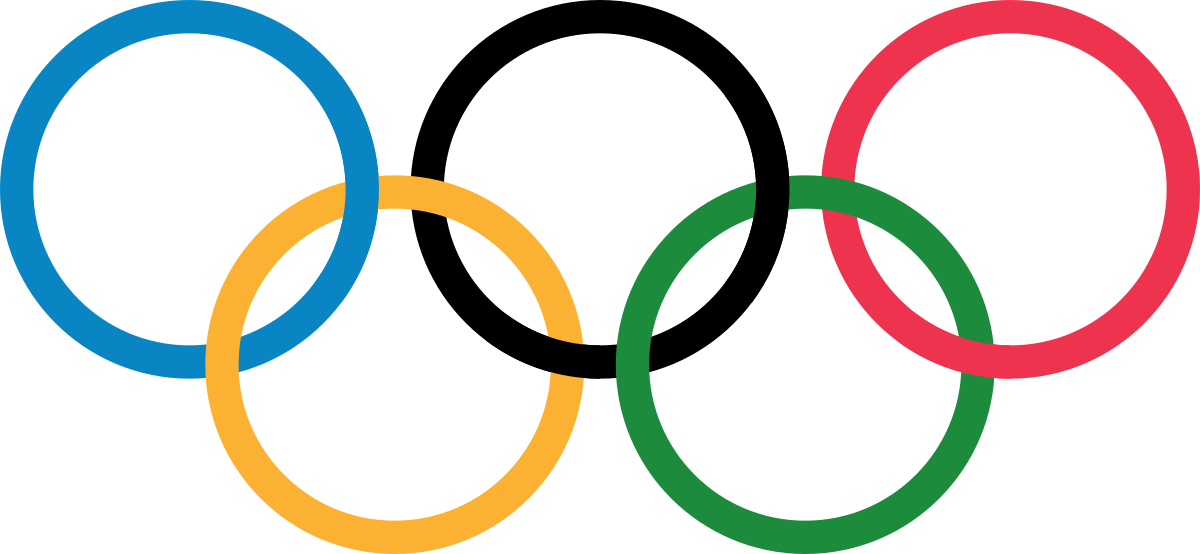
**Final Project**

**Data Science and Big Data Analytics**

**2019-SPR-IG-ITS836-21**

**University of The Cumberlands**



**By**

**Manoj Kumar Mukka,**

**Nilesh Narayan Patil,**

**Sai Kishore Reddy Tangella**

**Table of Contents**

[***Project Introduction: 2***](#_gjdgxs)

[**Project Choice Selected: 2**](#_30j0zll)

[**Dataset Selected: 2**](#_1fob9te)

[**Project Goal: 2**](#_3znysh7)

[**Data Abstract: 2**](#_2et92p0)

[***Description of Dataset: 3***](#_tyjcwt)

[**Columns: 3**](#_3dy6vkm)

[**Sample data: 3**](#_1t3h5sf)

[***Description of tools used: 4***](#_4d34og8)

[***Data Analysis: 4***](#_2s8eyo1)

[**Gender Ratio: 4**](#_17dp8vu)

[**﻿List of Athletes with the most appearances at Olympic Games: 5**](#_3rdcrjn)

[**Medal density by Age: 5**](#_26in1rg)

[**Youngest Athlete in Olympics who won medal: 6**](#_lnxbz9)

[**Oldest Athlete in Olympics who won medal: 7**](#_35nkun2)

[**Data Analysis based on the Weight and Height: 7**](#_1ksv4uv)

[**Observation on Height and Weight for individual sports: 9**](#_44sinio)

[***Top teams according to Participation and Medals: 9***](#_2jxsxqh)

# Project Introduction:

## Project Choice Selected:

Option II to implement analytics on an a publicly available data of significant size using the data analytics processes and modern tools.

## Dataset Selected:

120 years of Olympic history including all the games from Athens 1986 to Rio 2016

<https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results>

## Project Goal:

The goal of this project is to learn the best practices for big data analytics. By the end of this project each team member should have good experience on different stages of Data Analytics Life Cycle. By participating in this project, the team members learned most recognized technologies like Python including libraries like Pandas, NumPy, Matplotlib and Tableau.

## Data Abstract:

The project goal is to explore, analyze and interpret the dataset which consists data of all the records Olympics games from 1896 (Athens) to 2016(Rio). Right from 1896 Olympic games had great response from the athletes and audience. In the first year there were 14 participating countries which grew to 203 countries now, which shows how people encouraged it and Olympic game became prestige to international sporting festival and brought peace between countries. There is lot of scope for improvement, encouragement and prediction for the upcoming games with this available data.

The dataset consists records of each athlete with fields such as name, gender, age, height, weight, country, corresponding country code for the National Olympic Committee, the season, the hosting city, sport and awards which is scraped from [www.sports-reference.com](http://www.sports-reference.com) in May 2018. The dataset has 27k records with 15 different variables. There are 203 National Olympic Committee codes for all the countries participated in the Olympic games. This data gives access to find many patterns and analyze how Olympics has changed its face.

# Description of Dataset:

The dataset has 27k records with 15 different columns giving bio data of the athletes participated in Olympics games from 896 (Athens) to 2016(Rio), which is to be joined with a National Olympic Committee code map for 203 countries.

## Columns:

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Description** | **Number of Null values** |
| ID | Unique number for each athlete | 0 |
| Name | Athlete's name | 0 |
| Sex | M or F | 0 |
| Age | Integer | 9474 |
| Height | In centimeters | 60171 |
| Weight | In kilograms | 62875 |
| Team | Team name | 0 |
| NOC | National Olympic Committee code | 0 |
| Games | Year and season | 0 |
| Year | Integer | 0 |
| Season | Summer or Winter | 0 |
| City | Host city | 0 |
| Sport | Sport | 0 |
| Event | Event | 0 |
| Medal | Gold, Silver, Bronze, or NA | 231333 |

## Sample data:

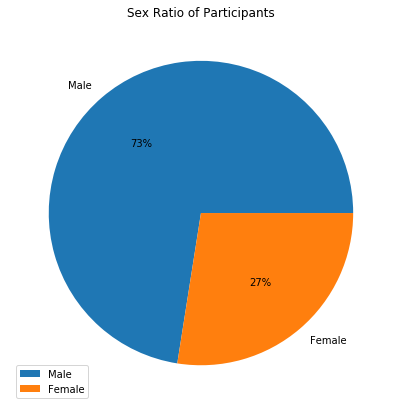


# Description of tools used:

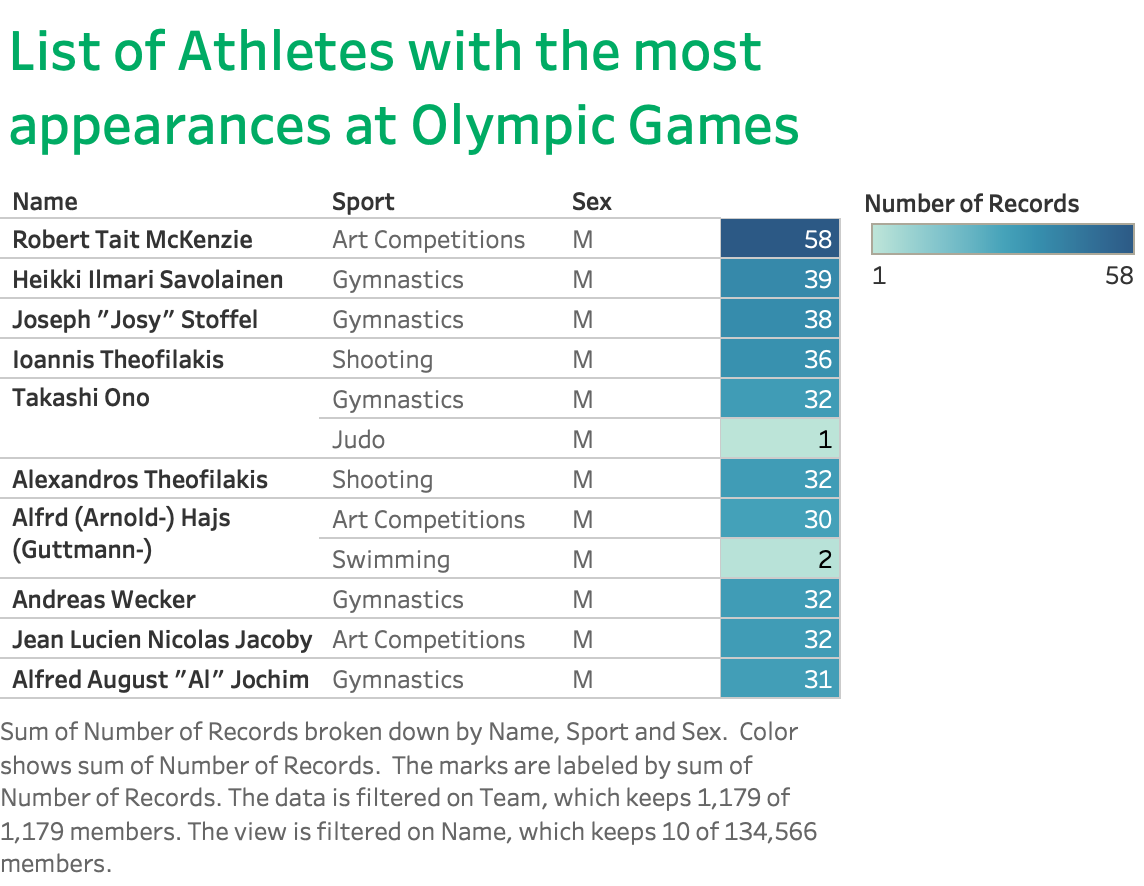
|  |  |  |
| --- | --- | --- |
| **Tool** | **Purpose** | **Libraries** |
| Python | To import, analyze and generate graphs from the data | Pandas, NumPy, Matplotlib, seaborn |
| Excel | Quick Data View |  |
| Jupyter Notebook | To generate documentation |  |
| Tableau | To import, analyze and generate graphs from the data |  |

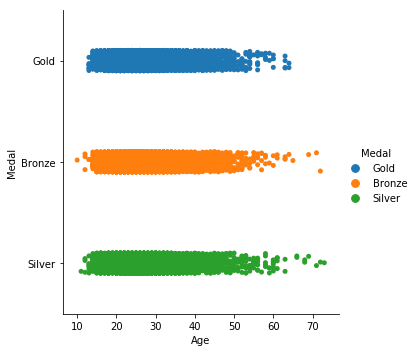
# Data Analysis:

## Gender Ratio:



## ﻿List of Athletes with the most appearances at Olympic Games:

Medal density by Age:



It is pretty interesting that there are Athletes who are older than 60 years winning Medals. Let's get deeper into Athletes with age>60

Sports Played by Athletes with Age>60.

['Shooting' 'Art Competitions' 'Archery' 'Roque' 'Sailing' 'Equestrianism']

Sports Played by Female Athletes with Age>60.

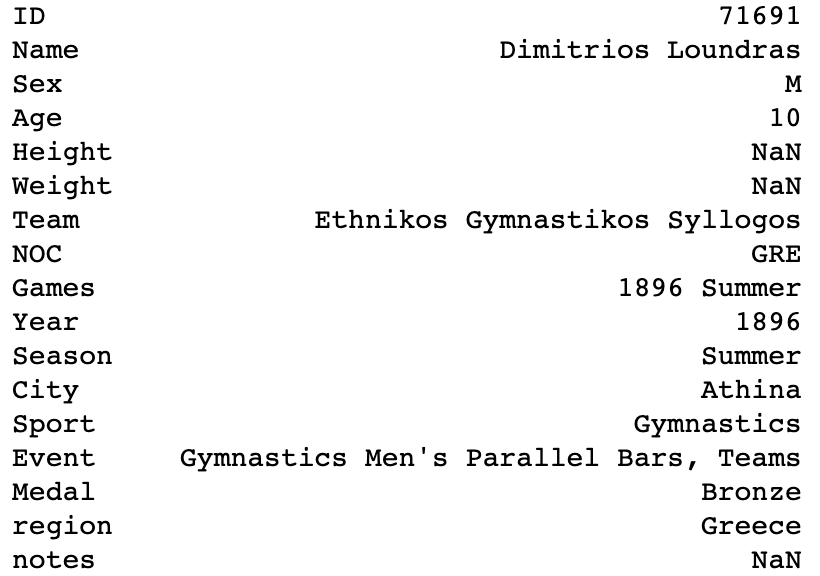
['Art Competitions' 'Archery']

Sports Played by Male Athletes with Age>60.

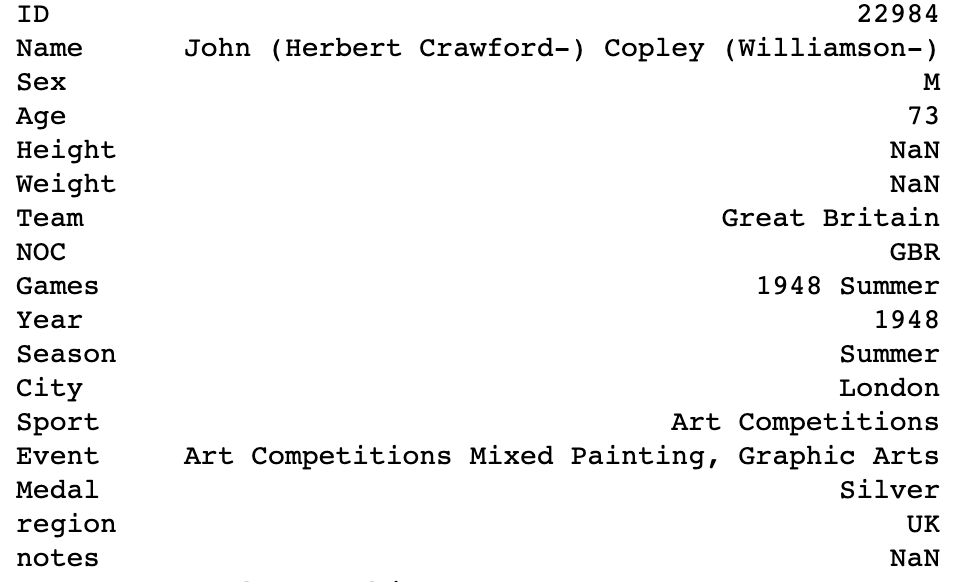
['Shooting' 'Art Competitions' 'Archery' 'Roque' 'Sailing' 'Equestrianism']

It is clear that all the sports with winning athletes older than 60 are about having keen eye sight.

## Youngest Athlete in Olympics who won medal:

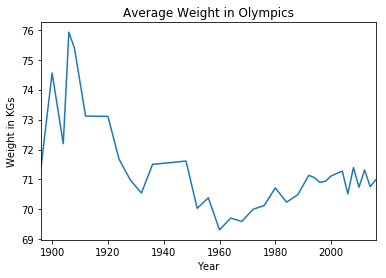


## Oldest Athlete in Olympics who won medal:

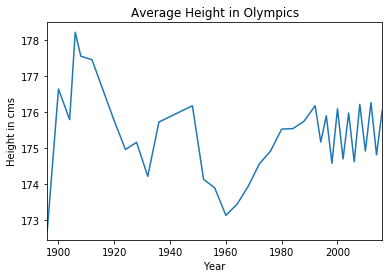


## Data Analysis based on the Weight and Height:

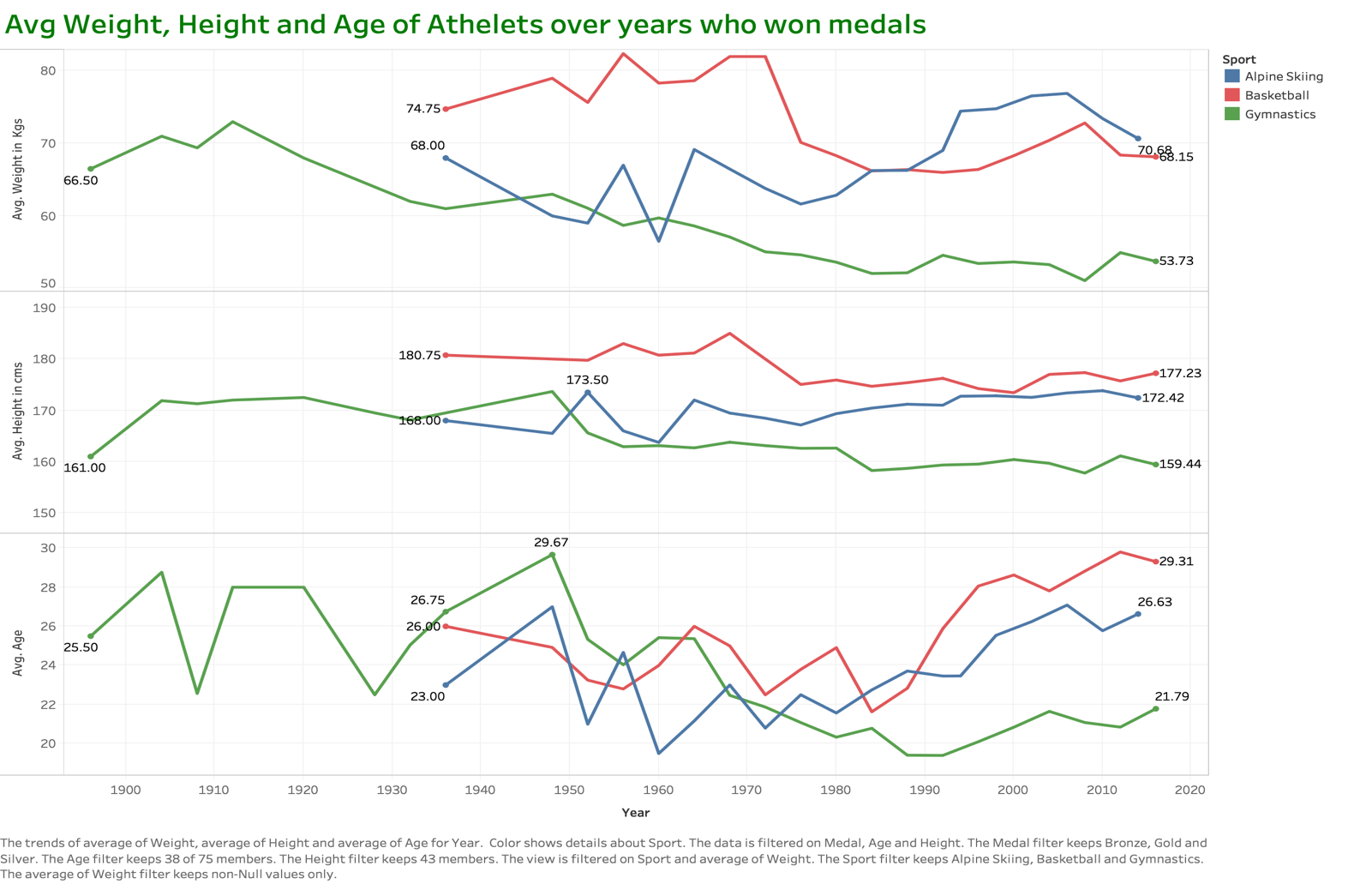
Average Weight of Athletes over Years:



Average Height of Athletes over Years:



Here we can observe that the Average weight and Height are pretty much same let’s analyze more deeper considering individual sports.

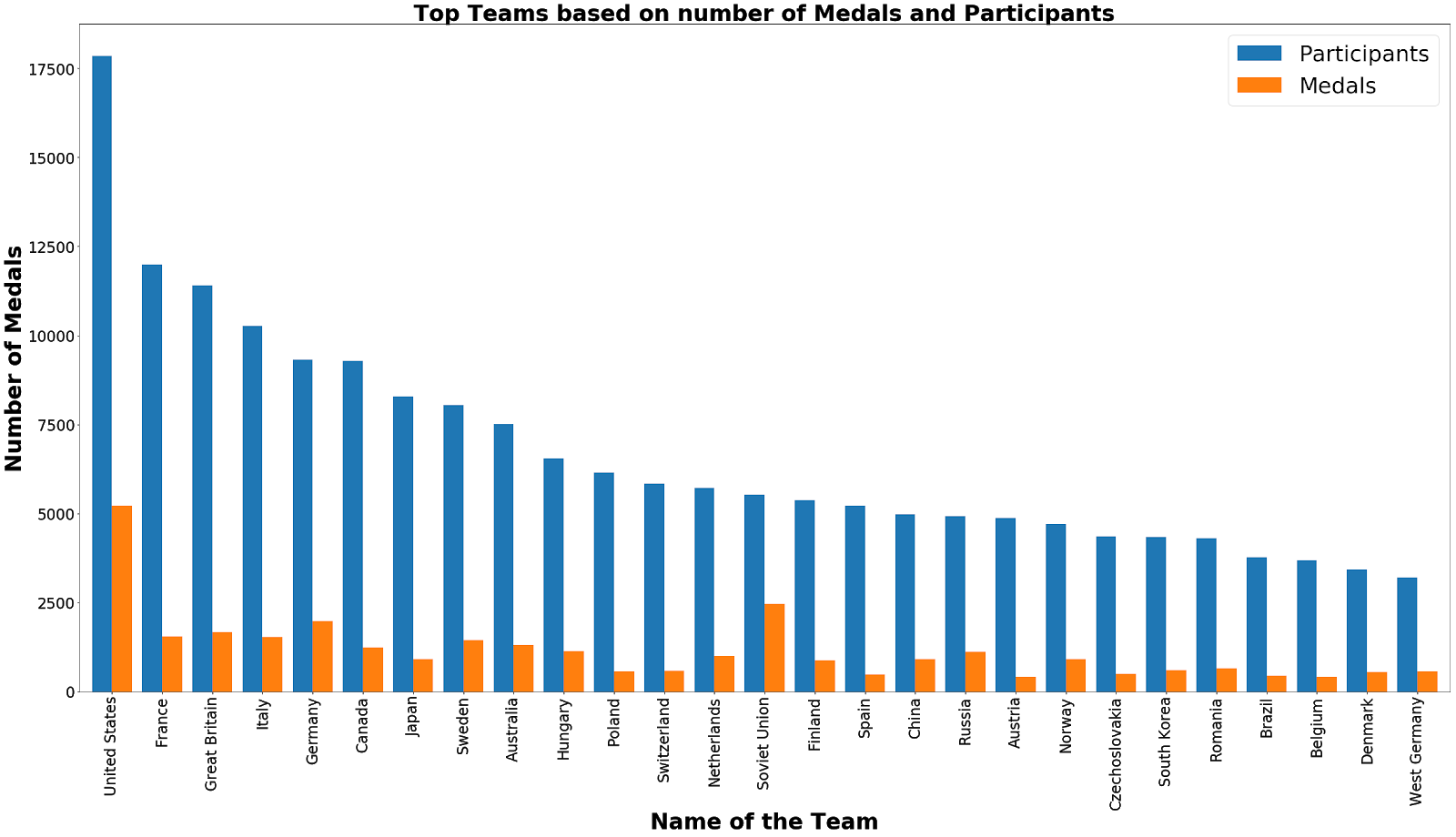


## Observation on Height and Weight for individual sports:

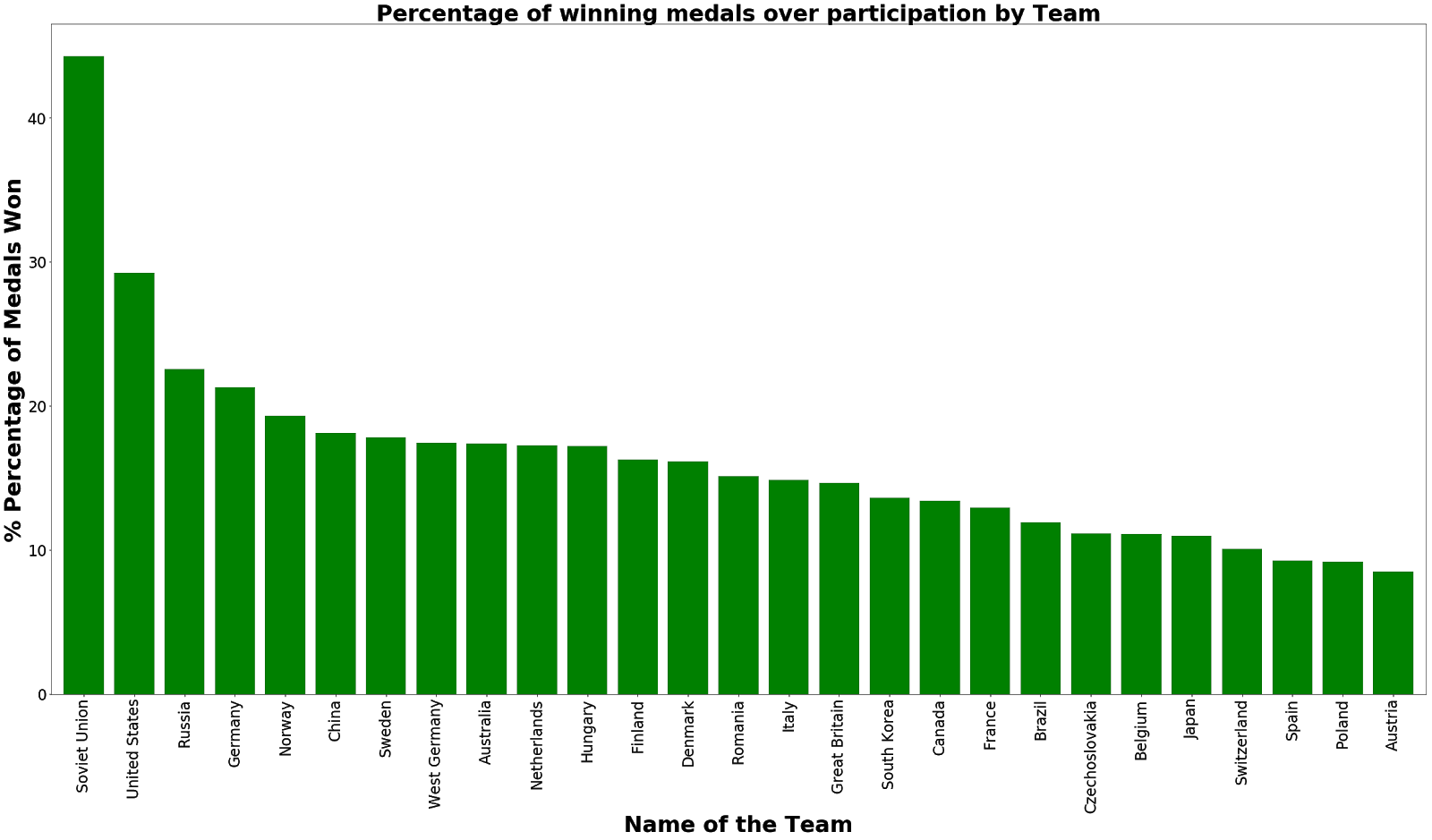
1. The average weight seen decreasing for the sports gymnastics and basketball while it is increasing for skiing this is because skiing needs more weight for speed.
2. Similarly, the average height seen decreasing for gymnastics and increasing for skiing, this is because gymnastics need stability which can be achieved by having less height.
3. The age pattern can be little unjudged because people may become more experienced by time can also become weak by age.

# Top teams according to Participation and Medals:

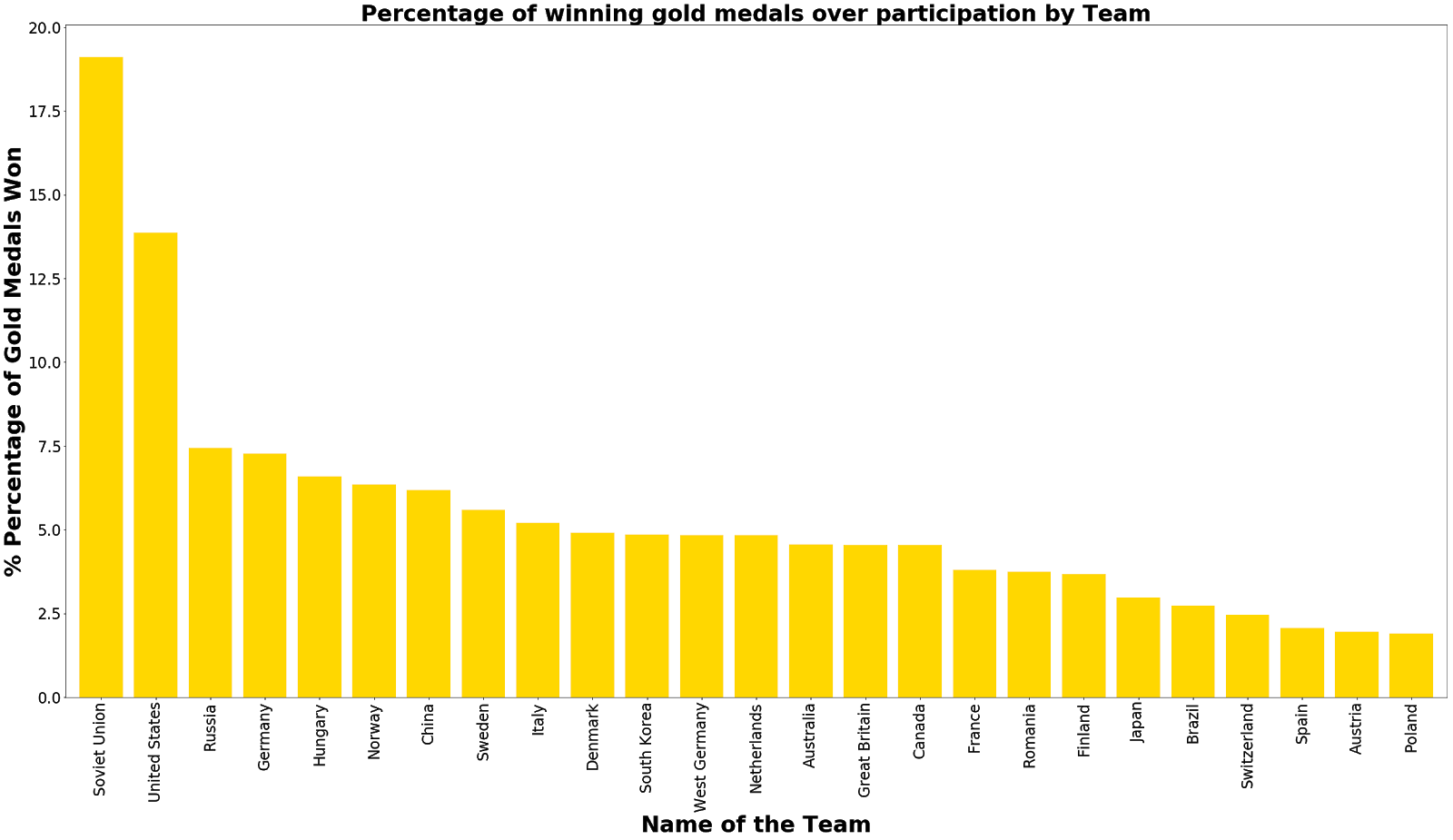
Top 30 teams by number of Medals and participations.



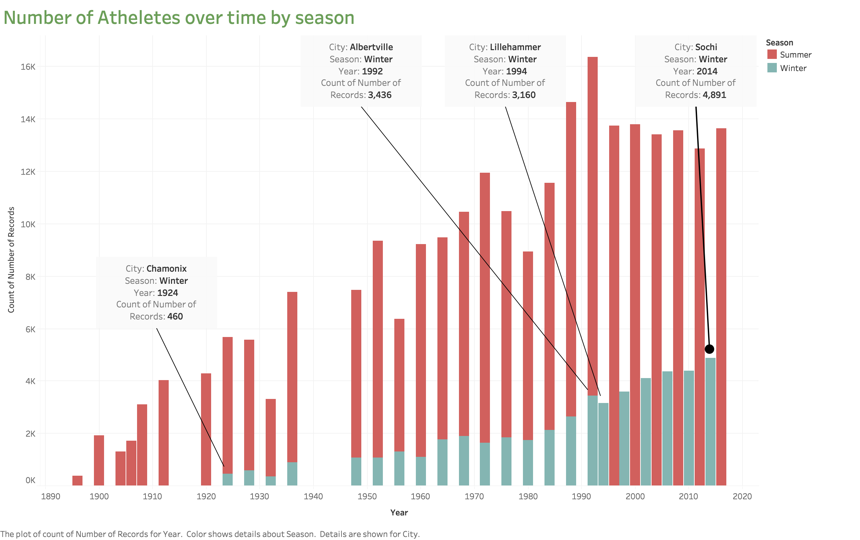
When we see the ratio of participation vs winning, Soviet Union is in the first place which has very less participation compared to USA but has more medals.



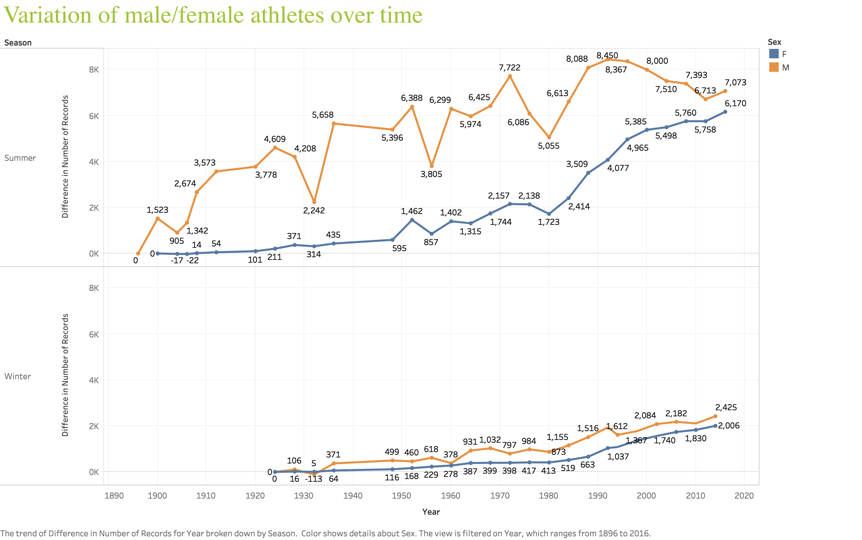
Let’s see how the ratio for gold medals is:

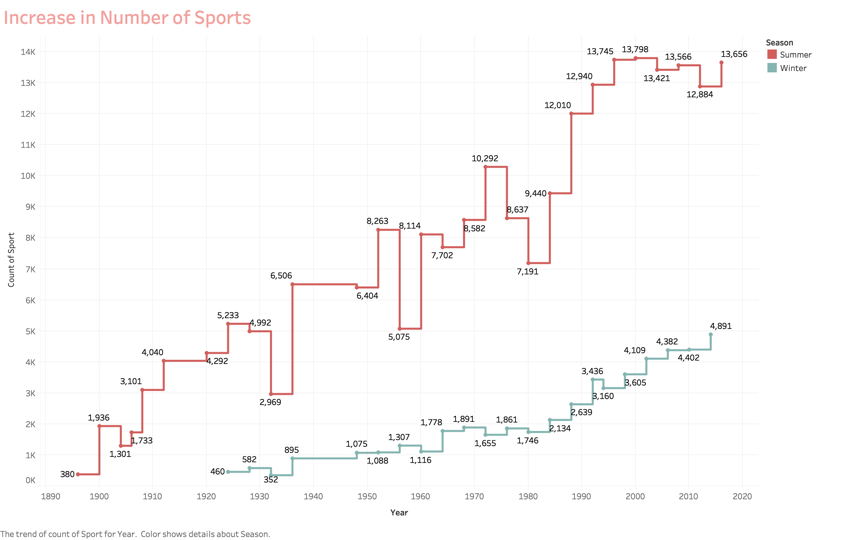


Analysis on number of Athletes over time by season:



1. Here we can observe that Winter games started from 1924 and got into separate event since 1994 , because in 1992 we had a huge number of hike in the number of participants.
2. There is a surge in number of participant in 1932 since the event was held in US and it is the same year when there was a worldwide recession which made very difficult for spending money on travelling to olympic games. After 1932 the number of participants grew gradually.





# Conclusion:

This dataset provides an opportunity to know about how the Olympics have evolved over time, including participation and performance of athletes, different nations, different sports and events. This project provided opportunity to gain high experience on the data analysis tools like Python, Tableau and Jupyter.

For more information and code please follow the below links

Code:

<https://github.com/mukkamanoj/Final-Project-2019-SPR-IG-ITS836-21>

Tableau Views:

<https://public.tableau.com/profile/manoj.kumar3772#!/vizhome/Analysis-on-Olympic-data/Variationofparticipation?publish=yes>

Presentation:

<https://docs.google.com/presentation/d/1NNqSX8XeepOhWxP834wblAmTvwOw2WvL9lDD1jSZI2w/edit?usp=sharing>