# Olympic Games


A model to predict the chance of winning

# Agenda

What you'll learn about today


The Summer Olympic Games, also known as the Games of the Olympiad, are a major international multi-sport event normally held once every four years. The inaugural Games took place in 1896 in Athens, Greece, and most recently the postponed 2020 Summer Olympics were celebrated in 2021 in Tokyo, Japan.

# External Resources

**Olympics | Olympic Games, Medals, Results & Latest News**
Olympic Games - Wikipedia
Auxiliary dataframe - population
Auxiliary dataframe - GDP

# MECHINE LEARNING

## 01

### Predict the number of medals of a country

By compettor, event, year city, country, name, gdp world population and total athlete

## 02

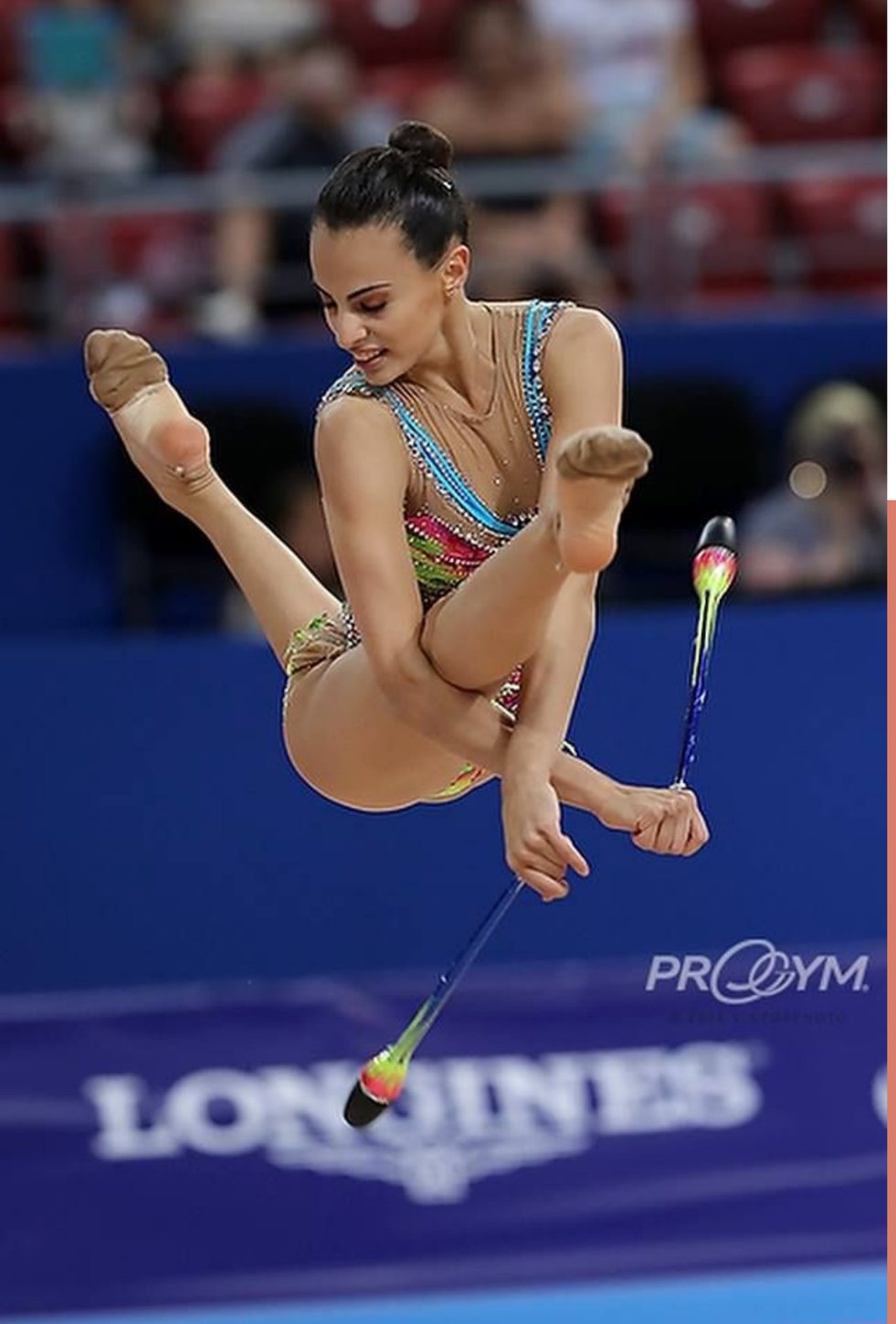
### Prediction of winning by parameters

By age, weight and height

## 03

### Predicting the amount of women in the next Olympics

By Year, total athletes, population, gdp, Log GDP, Log Population




# Stage 1: data acquisition

## Crawling

### using BeautifulSoup & Requests & Pandas & numpy

Extract all olympic information from 1896 until now

Extract :

- Season (always summer)
- Year , Country, The host city
- Game type (event), Medal Type
- Competitor name, sex, year of birth, height, age, Weight, Pos (Place in the competition)

# OUR AUXILIARY DATA FRAME

Helps us answer prediction questions by adding more parameters

## Population

We assume that the larger the population, the more participants there are and the more likely it is that more talented athletes will succeed in the Olympics.

Q 01 & 03

## GDP

We assumed that the more resources a country has to invest in its athletes and the Olympics, the higher the results.

Q 01 & 03

# Stage 2 : Data cleanup

- Remove duplicated
- Binary classification by sex
- Categorical variable by medal
- Categorical features - Ordinal Encoder by Country and Competitor
- Replace NaN / null / None Respectively

# Union between the primary and secondary data frame

Actually, we added to the main data frame the columns that interest us from the secondary data frame and according to the forecast question we used the relevant columns

# STAGE 3 : VISUALIZATION


## 01

A visualization that will help us predict the number of future medals of a country

## 02

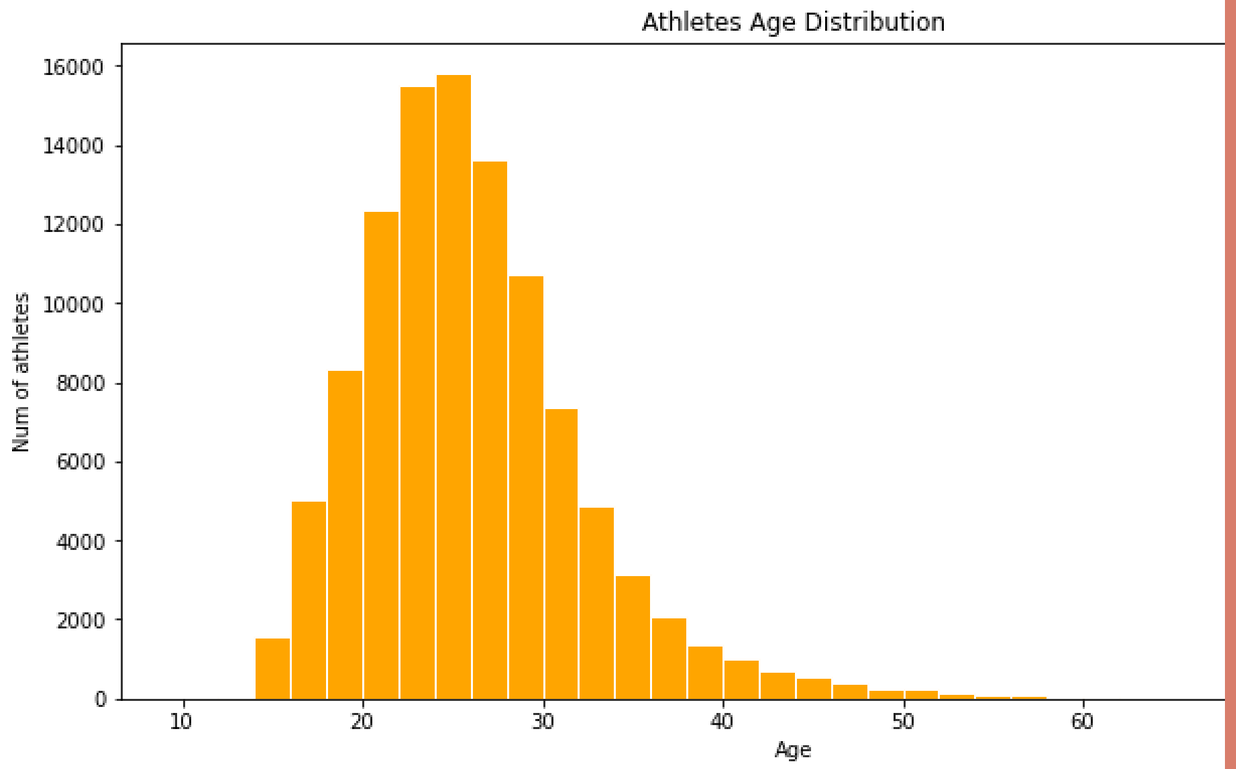
Visualization that will help us predict winning by parameters (weight, age and height)

## 03

A visualization that will help us predict the amount of women in the next Olympics

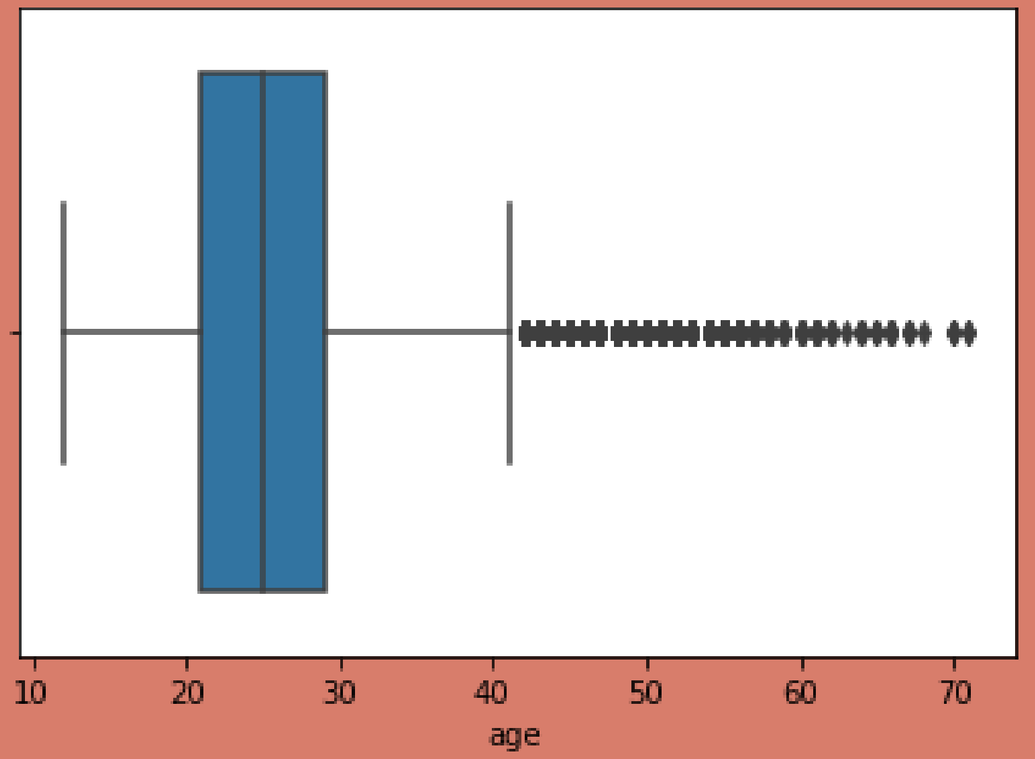
# M.L 02

Now let's see the age distribution to see what is the most common age to compete in.



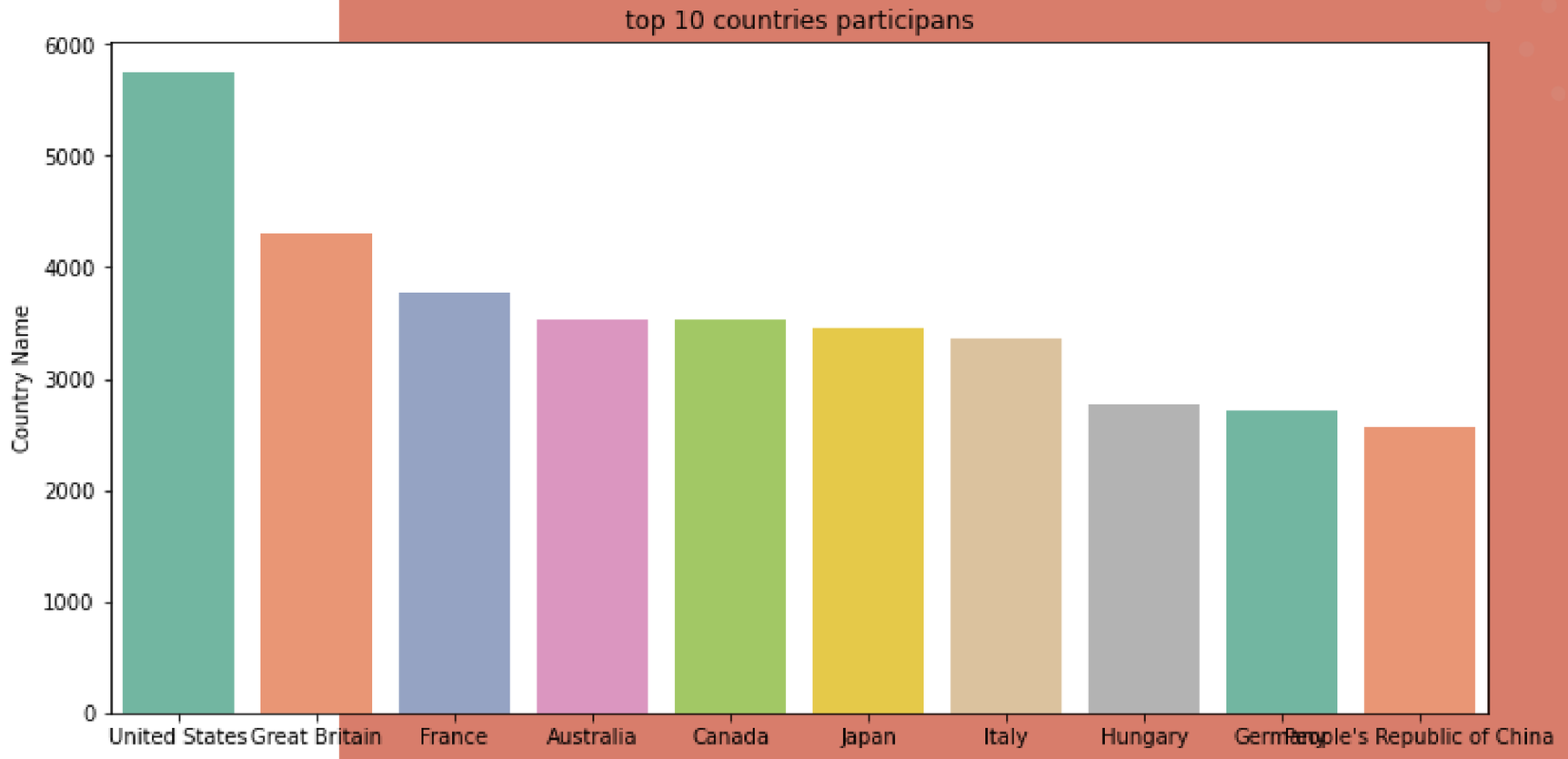

# M.L 01 + 02

In this graph we can see the most common age and the more unusual ages to compete in the Olympics



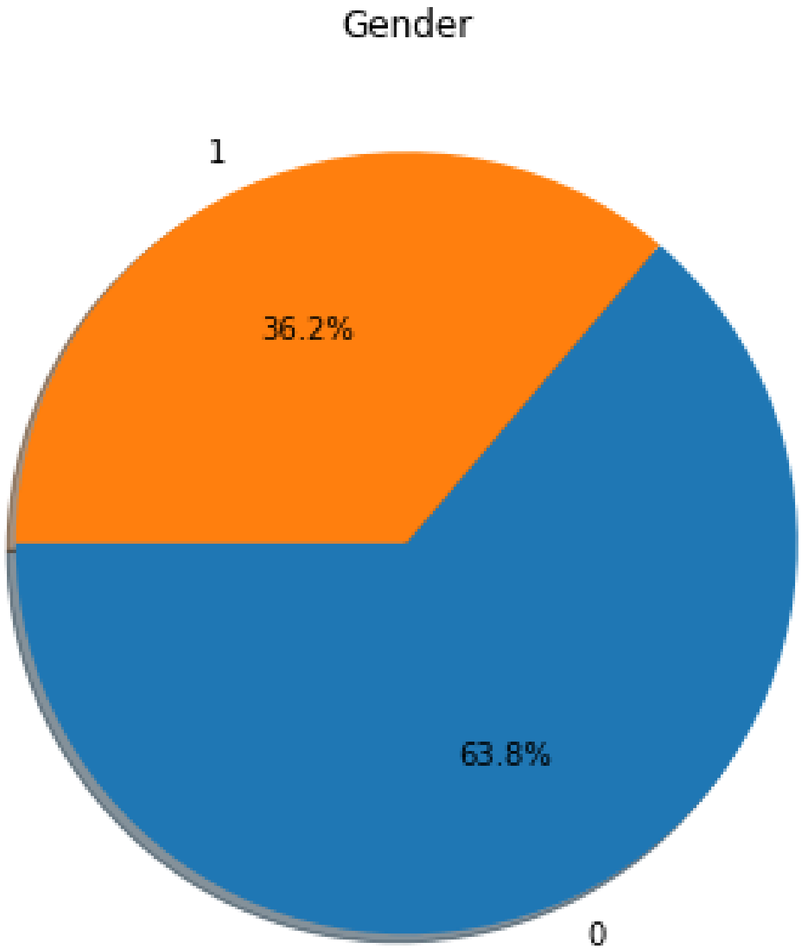
# M.L 01

Comparisons of top 10 countries
by amount of participants



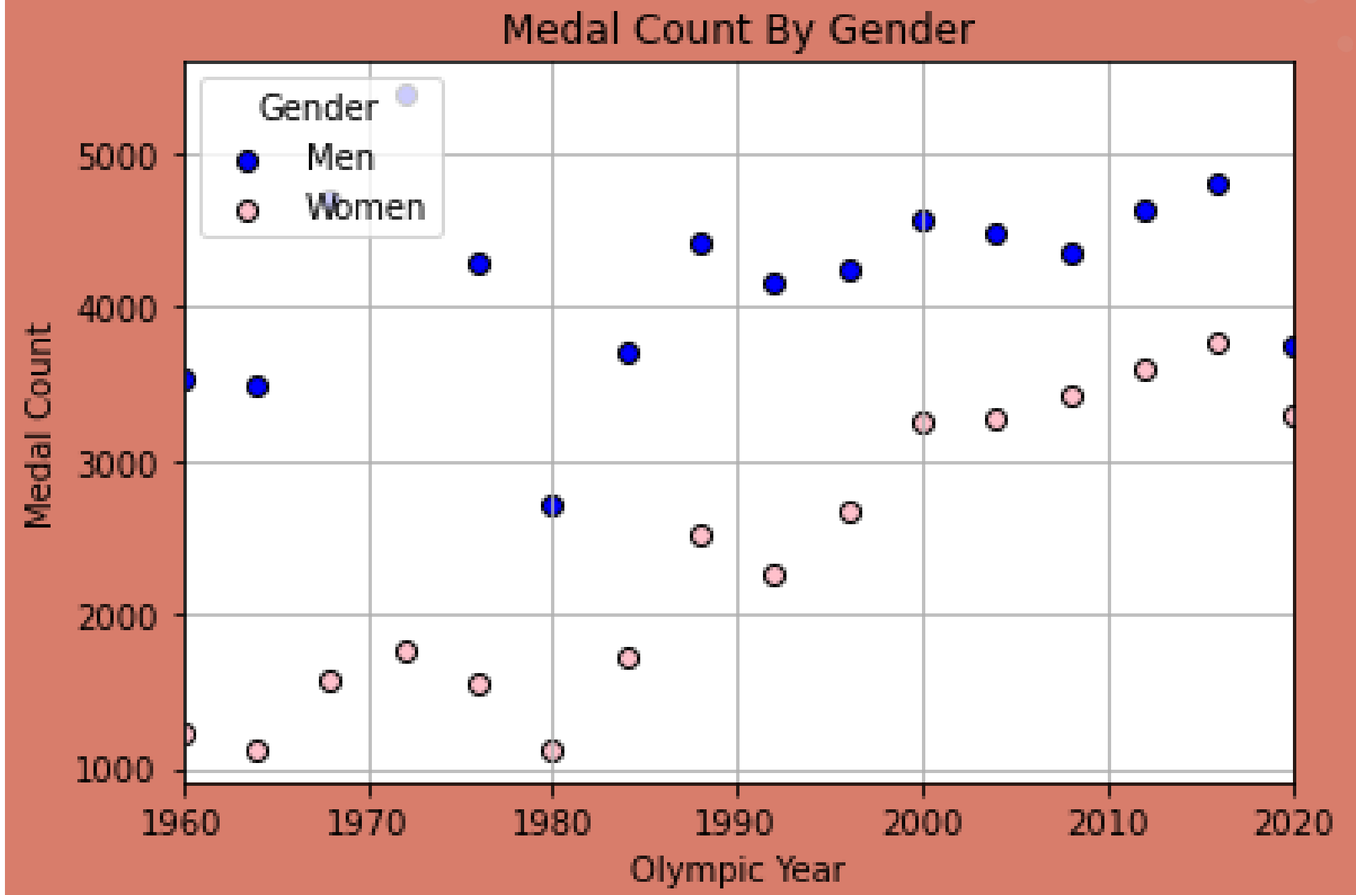
# M.L 02

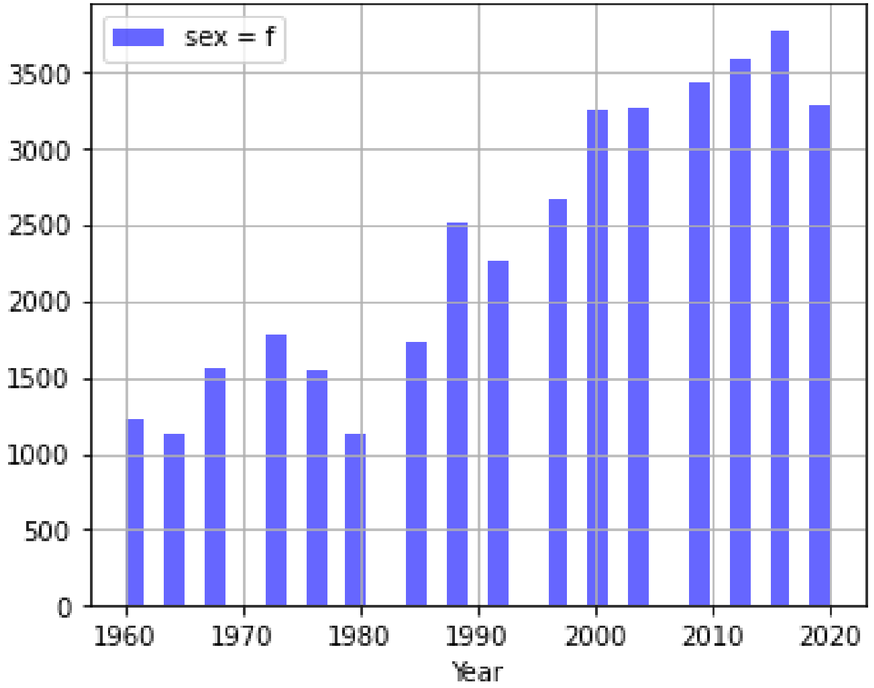
the number of females and males by years and comparison between them

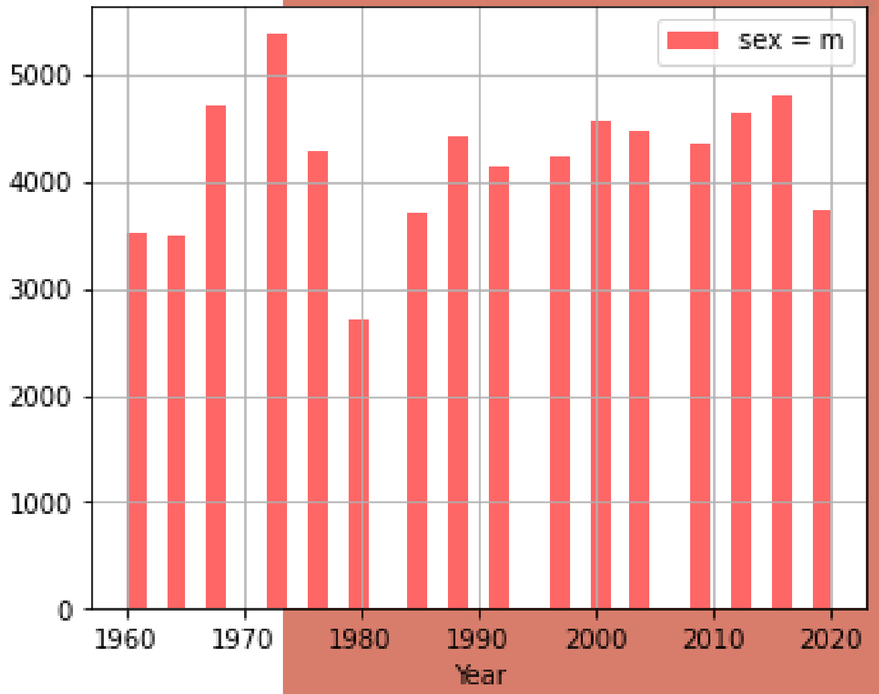


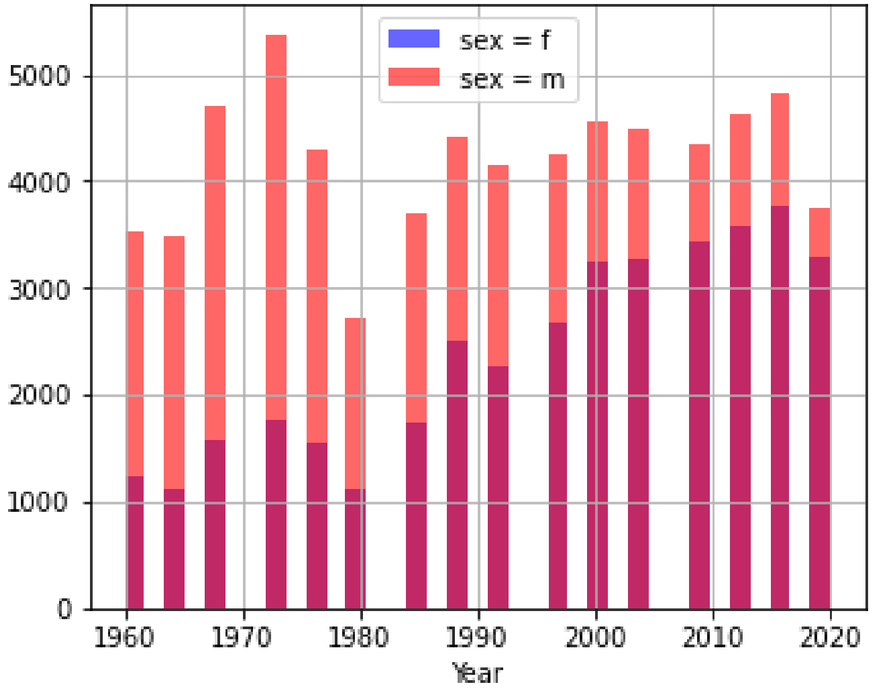

# M.L 03

Gender medal analysis





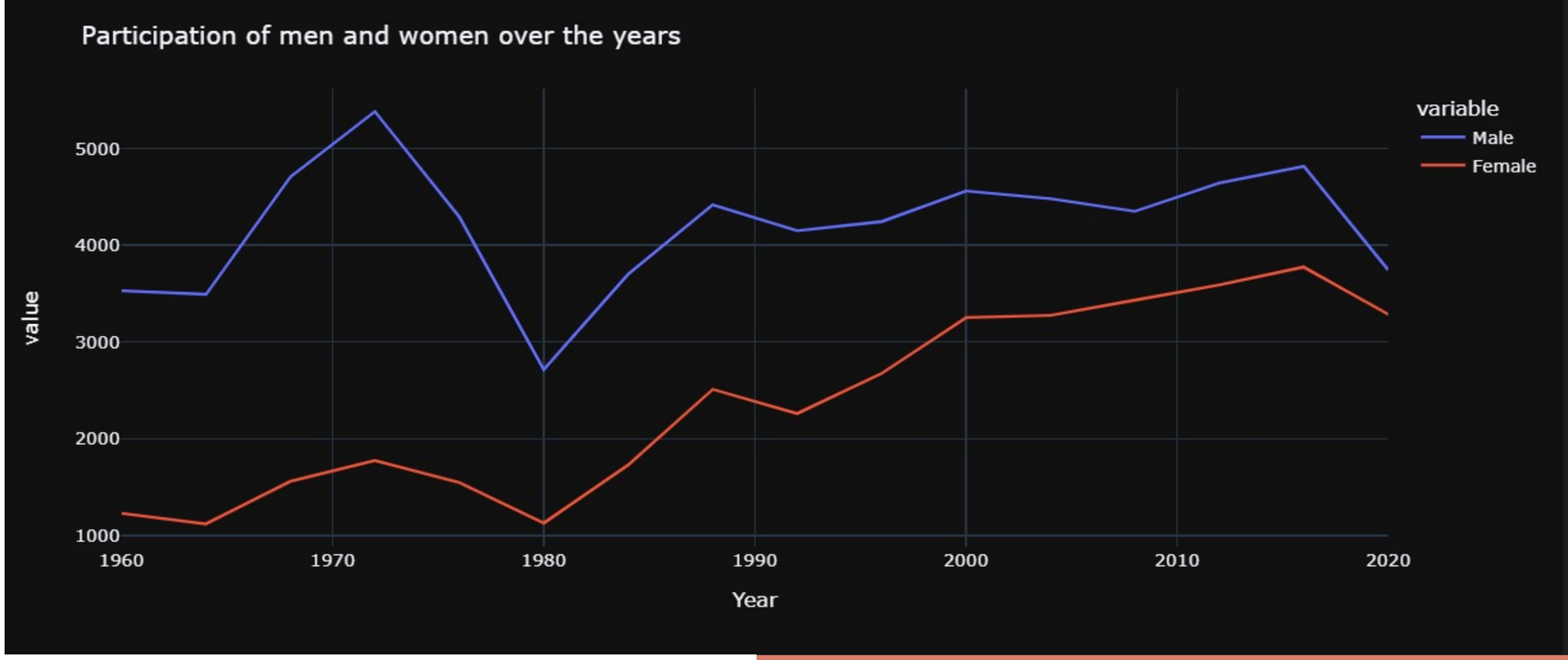



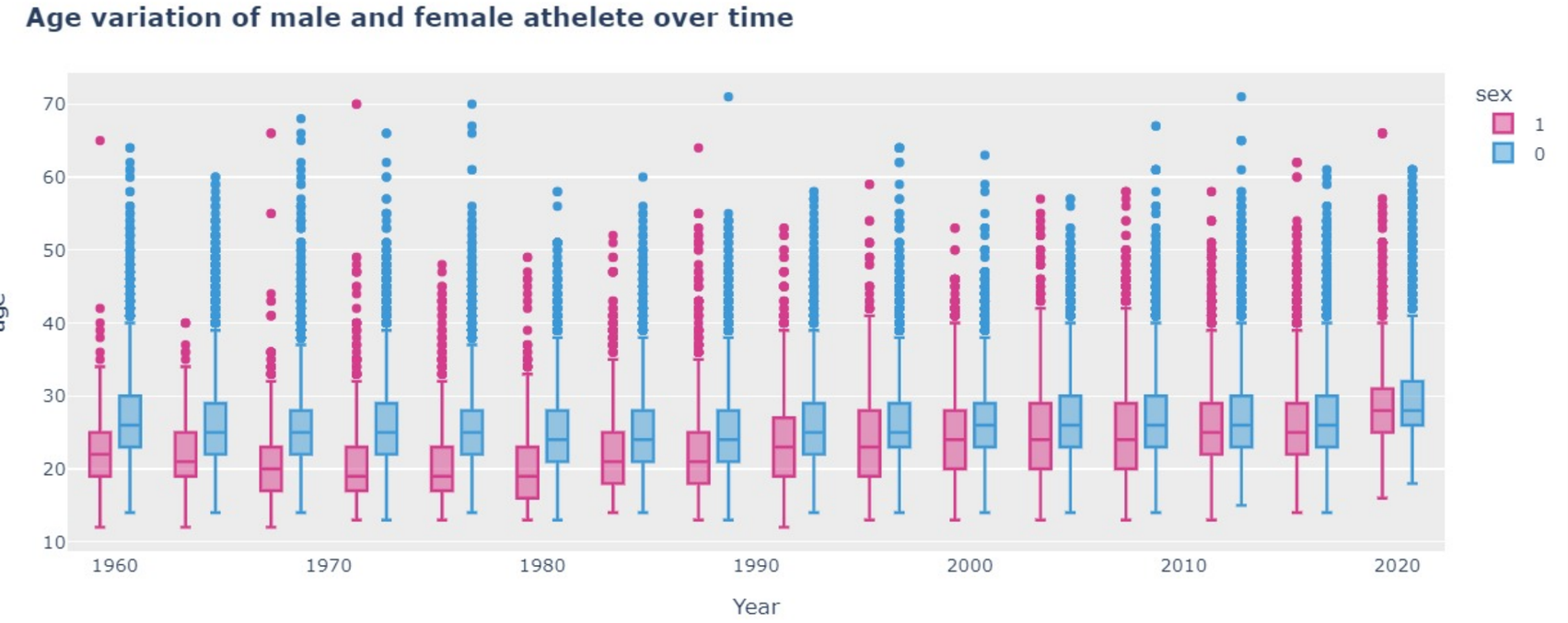




# M.L 03

gender split - how many men participated compared to the number of women.

# M.L 03

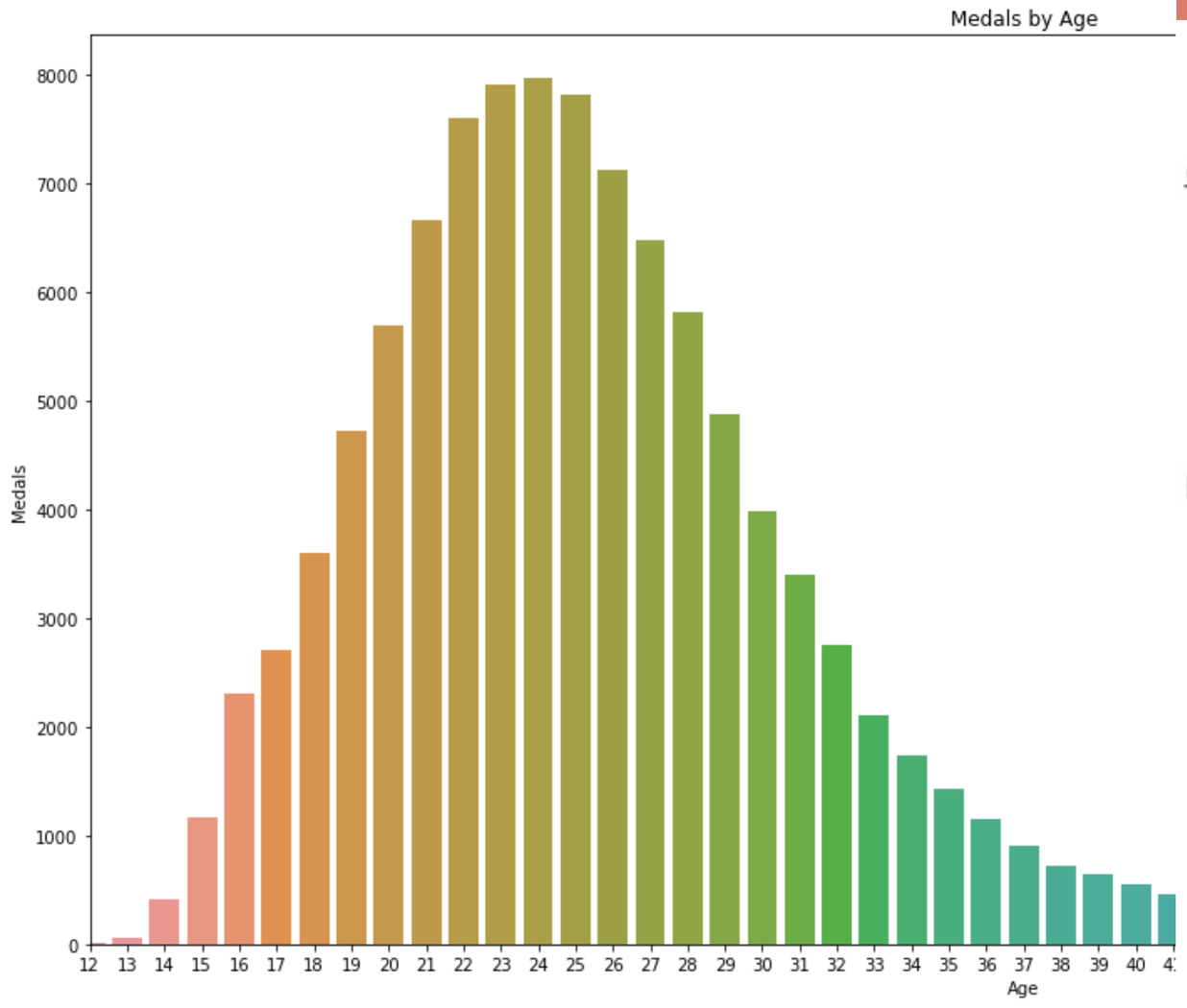



gender split - how many men participateion of men and women over the year by value.

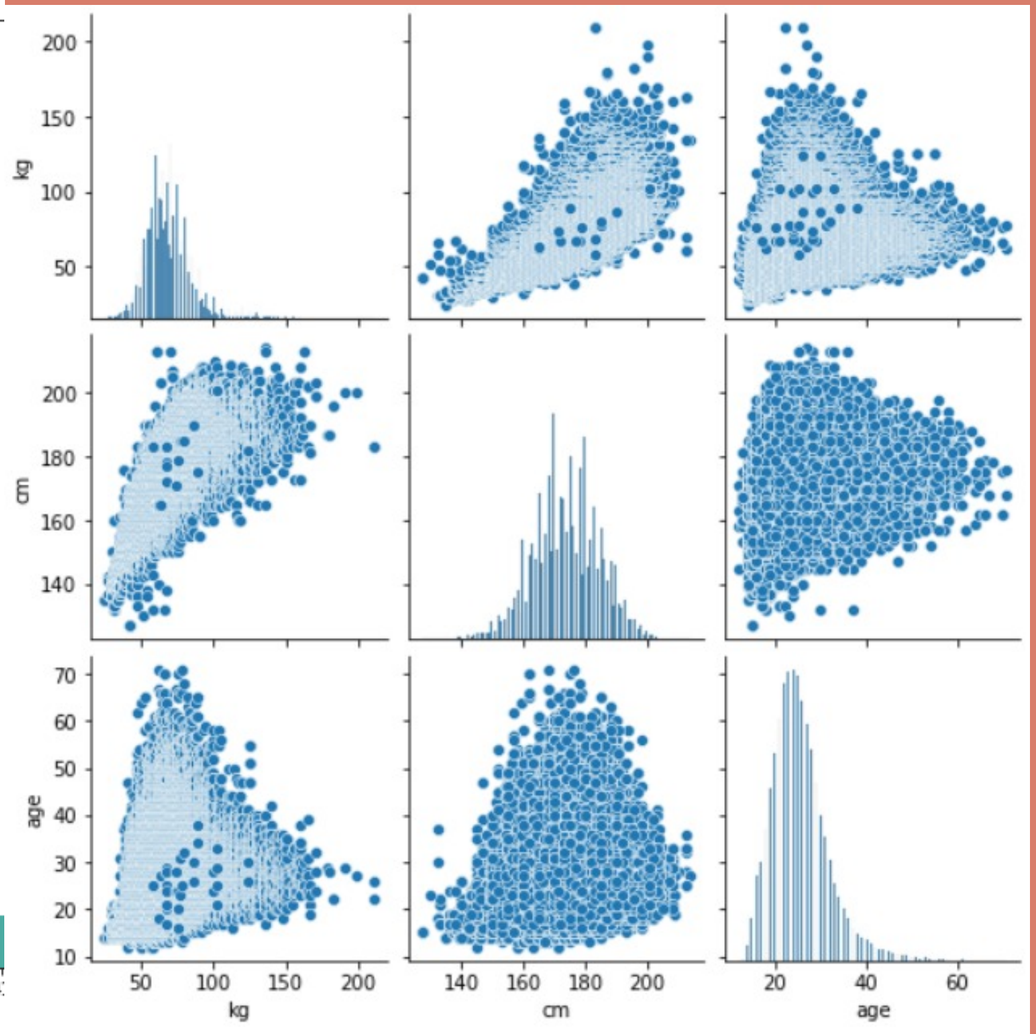
# M.L 03

Distribution of gold medal according to age

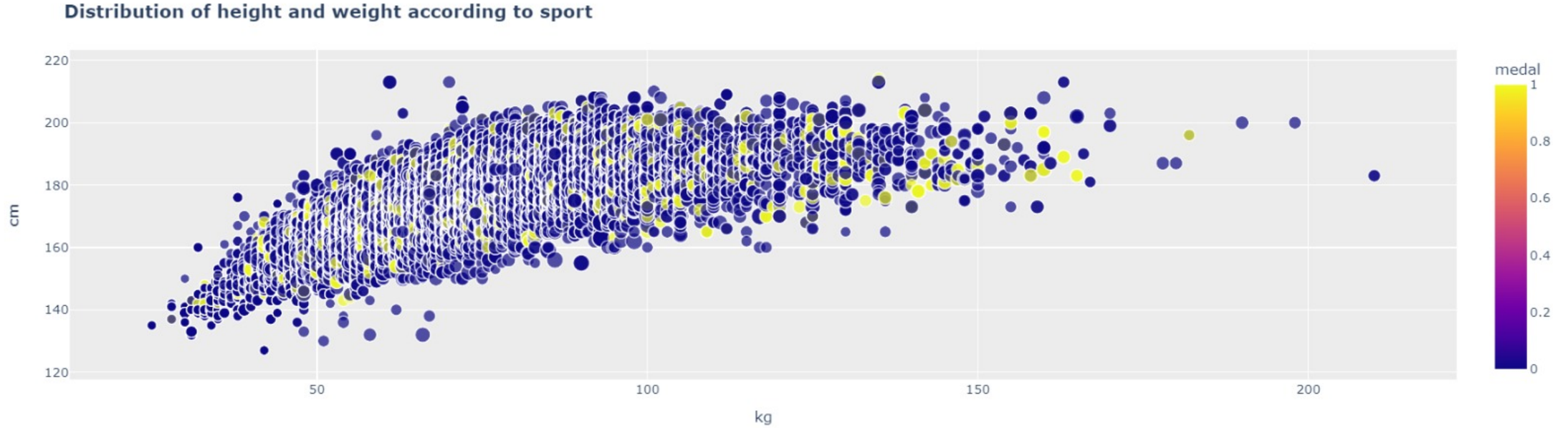


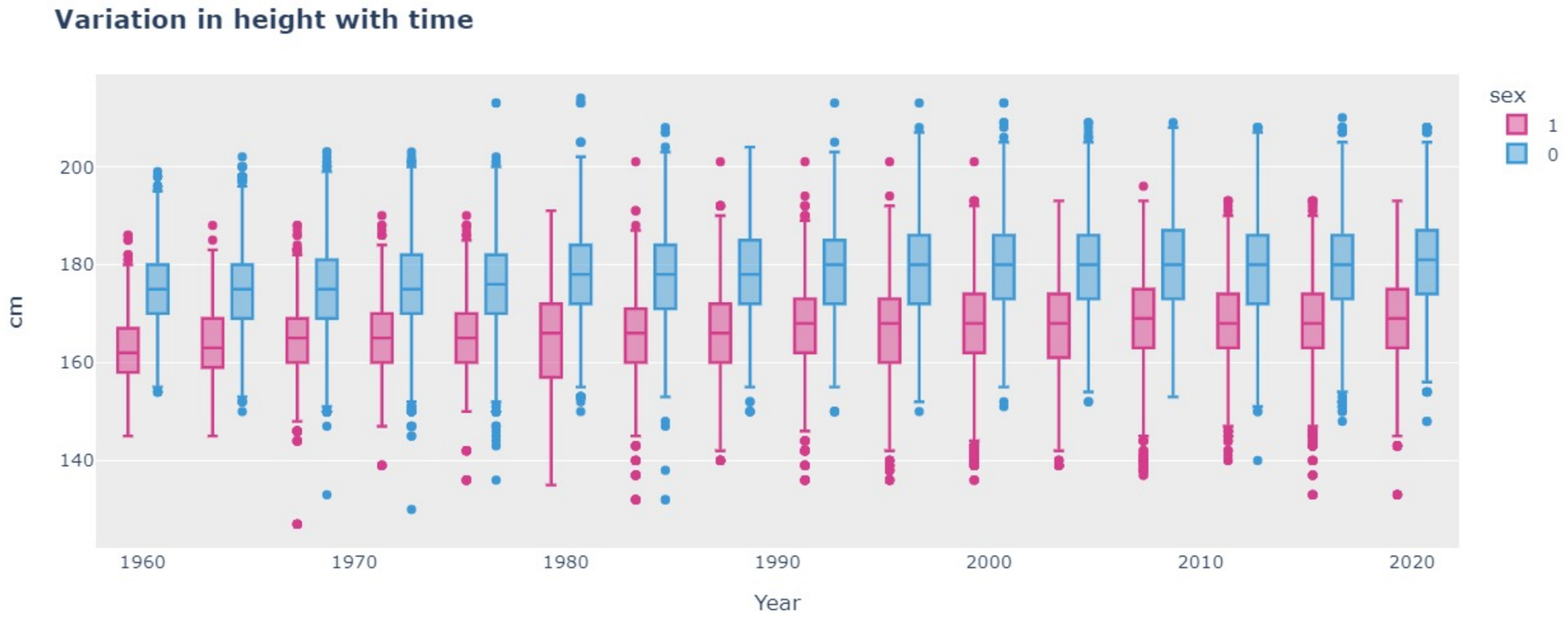

# M.L 02

The ratio between weight, height and age



# M.L 02


Distribution of height and weight according to sport
cm
kg
medal
220
200
180
160
140
120
50
100
150
200
1
0.8
0.6
0.4
0.2
0


Variation in height with time
cm
Year
sex
1
0
200
180
160
140
1960
1970
1980
1990
2000
2010
2020

# Stage 4 : Machine learning

Machine learning is the study of computer algorithms that can improve automatically through experience and by the use of data.

In order to best predict our questions we will adjust the best form of prediction

## Logistic regression **Q 02**

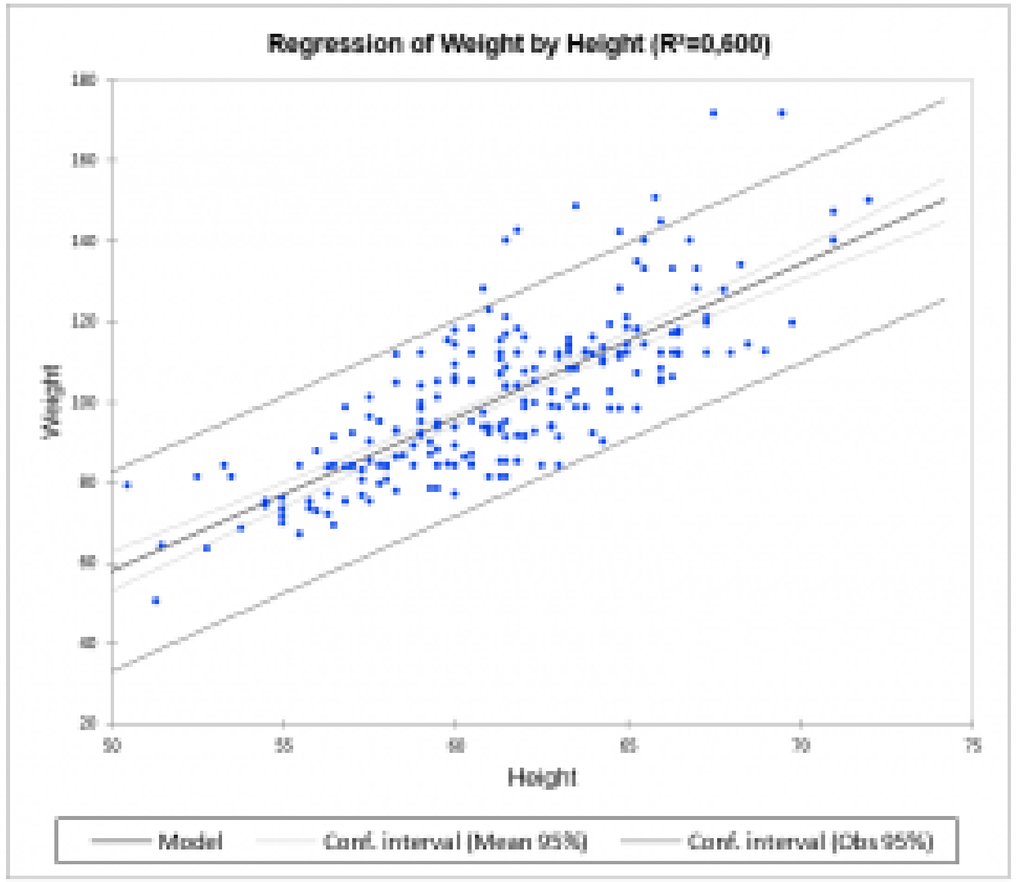
Logistic regression is a method to create classification between different object.

## OLS **Q 01**

common technique for estimating coefficients of linear regression equations which describe the relationship between one or more independent quantitative variables and a dependent variable

## Linear regression **Q 01 & 03**

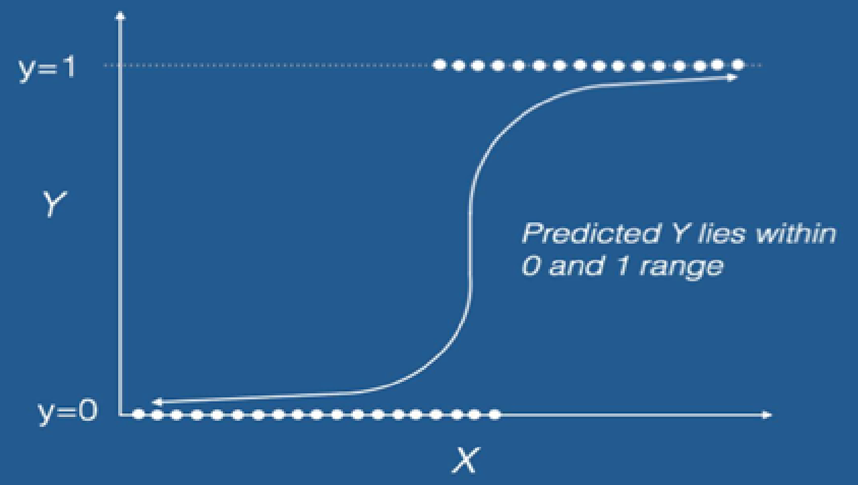
Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data




# M.L Q 01

We downloaded unnecessary variable (country code) and then We added host city and mapped the city to the state, we downloaded all NULL
and because rule variables (like population) can be biased it is best to pass to logarithms.
We turned on the OLS with - Log_GDP + Log_Population + total_athletes + Home_adv + GDP_per_person
- to predict the medal
The error associated with this prediction is 5
R^2 = 0.634

# M.L Q 02




We Splits each country into a column and uses binary classification and defining X as all the columns without the medal column using features with the highest importance - sex, age, height, weight and country name.

and Y as the medal column

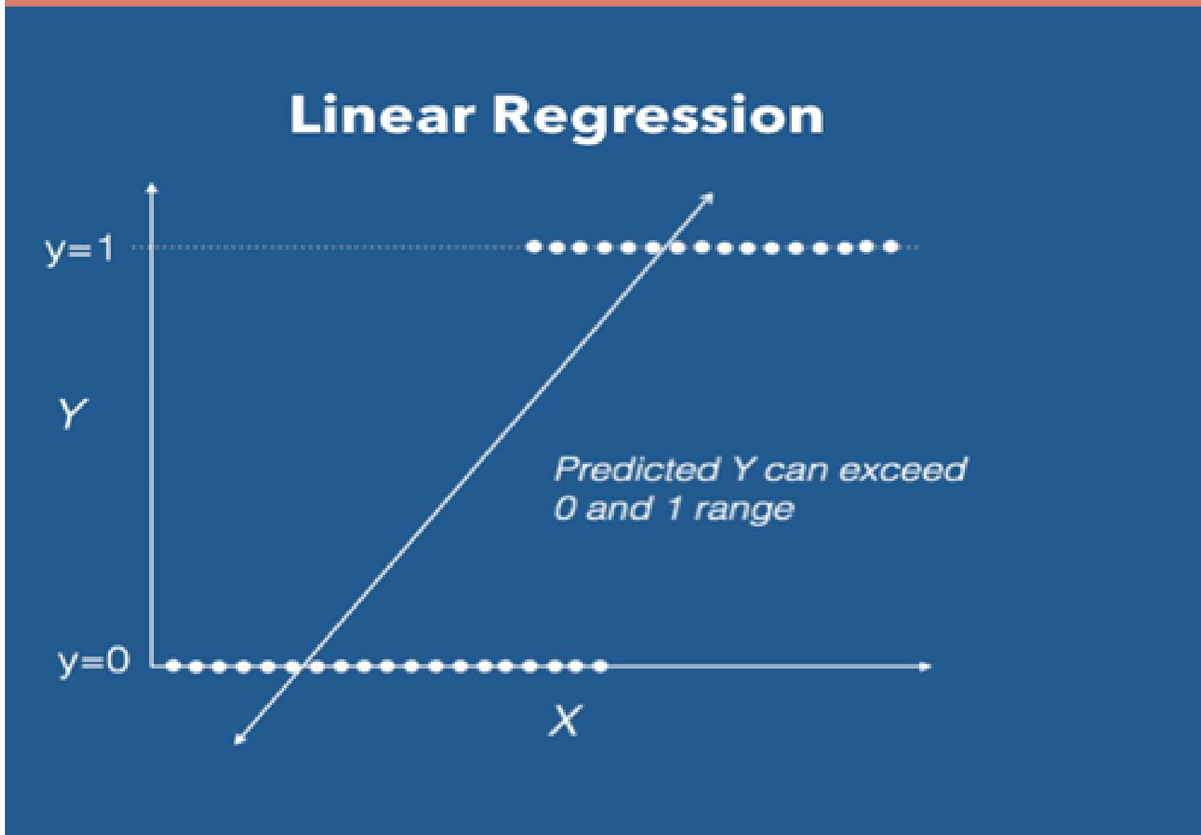
We turned on the Logistic Regression and train our X and Y.

Test - 30%

score = 92.5%

# M.L Q 03

In order to match the 2 data frames we used group-by (sex and year), We summed up the amount of women and men (separately) each year.
We added the quantities in one table
and created a linear regression object defining X as - 'Log_GDP','Log_Population', 'total_athletes','Male','Year' - and Y as Female - to predict the amount of future women
score = 95%



# Thank you for listening!

Got any questions? Feel free to call us, email us, or hit us up on social media.