

# Sport Stats

Preparation For Project Proposal

# Preparing of Project Proposal

## 1.) Which Clint did I select and Why?

- ▶ I choose the Sports Stats client.
- ▶ I selected this project because I have a lot of knowledge in sports both as a player and as a viewer. Also I might be interested some day to work in sports analytic reports.
- ▶ With an analysis of the dataset I can find hidden patterns and insights for players, teams and even the sports themselves.

# Preparing of Project Proposal

## 2.) Importing and Cleaning

- ▶ First of all I am working with Spark Databricks community edition.
- ▶ For importing I downloaded the .csv files and created the tables, using the first line as header and the infer schema options for the table UI.
- ▶ While scrolling through the data I saw some NaN values, usually in the weight and height columns so I decided to not “clean” the data since that would be a falsification of them.

# Preparing of Project Proposal

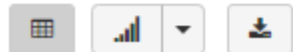
## 3.)Initial Exploration

```
Select count(distinct Name) as different_Players,  
       count(distinct Team) as different_Countries,  
       Min(Year) as First_Year,  
       Max(Year) as Latest_Year,  
       count(distinct City) as different_Cities,  
       count(distinct Sport) as different_Sports  
From athlete_events_csv
```

► (3) Spark Jobs

	different_Players ▲	different_Countries ▲	First_Year ▲	Latest_Year ▲	different_Cities ▲	different_Sports ▲
1	134732	1184	1896	2016	42	66

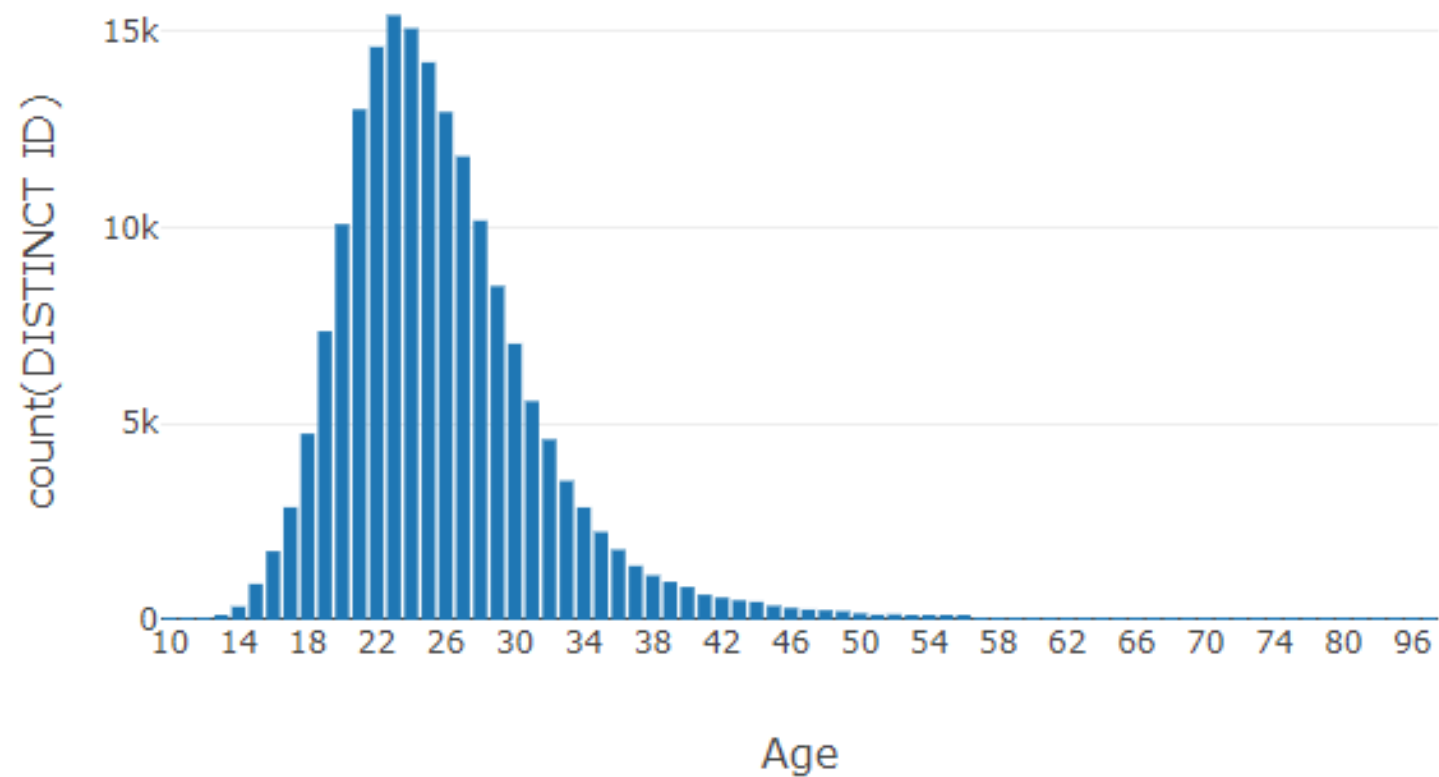
Showing all 1 rows.



```
Select `Age`,count(Distinct `ID`)  
From athlete_events_csv  
Where `Age` <> "NA"  
Group by `Age`  
order By `Age` ASC
```

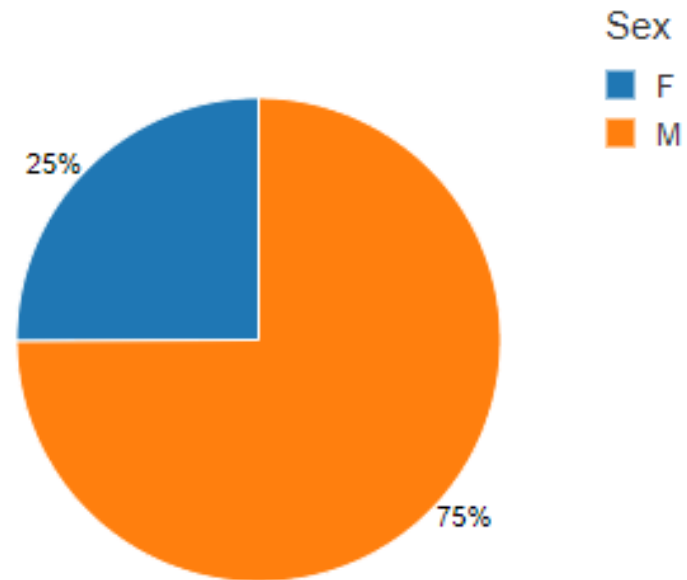
---





► (3) Spark Jobs



```
Select `Sex`, Count(Distinct Name)
From athlete_events_csv
Group By `Sex`
```

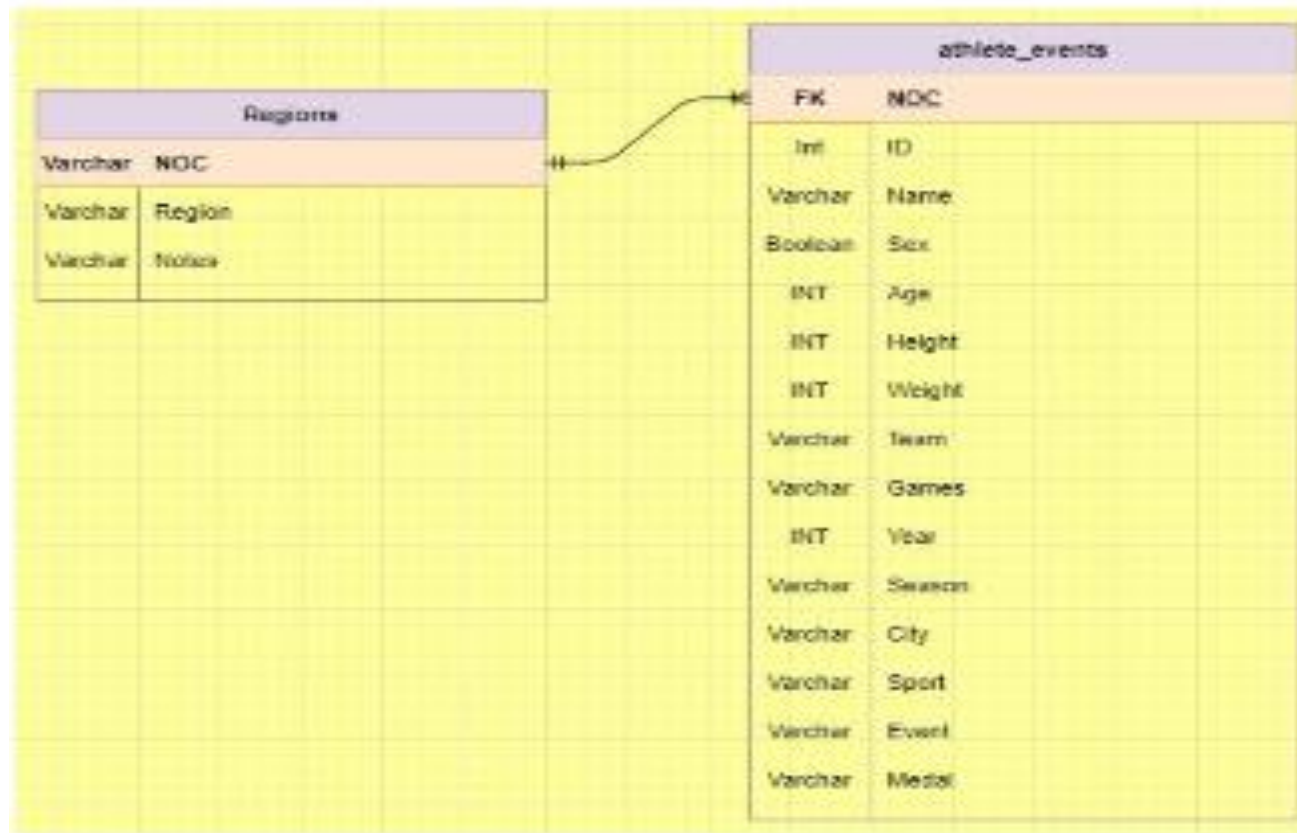
► (3) Spark Jobs



   Plot Options... 

# Preparing of Project Proposal

## 4.)Proposed ERD



# Develop Proposal

## 1.)Description

- ▶ My Project targets on discovering a relationship between the data and hidden patterns.
- ▶ Getting to know more insights on past Sports, such as when and where was the first event.
- ▶ This analysis will help both Coaches and the SportsStats firm to aid in their clients decisions with data-driven results.
- ▶ The audience for the project would be Coaches, Trainers and players who will be able to see for themselves some past “trend” or change.



# Develop Proposal

## 2.) Questions

- ▶ When was the first event and where?
- ▶ What is the age distribution?
- ▶ Is there a connection between age and medals?
- ▶ When were Sports added in the games?
- ▶ Which country has the most medals?
- ▶ Which country had the highest number of players?
- ▶ Is there a correlation among the previous two?

# Develop Proposal

## 3.) Hypothesis

- ▶ Probably the ages will follow a normal distribution with the max near 24.
- ▶ There is a connection between medals and age.
- ▶ Probably the more players a country has the more medals it gets.

# Develop Proposal

## 4.) Approach

- ▶ The features I am mostly interested in is Age, Medals and Year.
- ▶ I described the relations I will dig into.
- ▶ The evaluation I will use is mostly by Graphs and some statistical norms like means and A/B testing.

# Sport Stats

Project Final Results After Analysis

# Contents

- ▶ A deeper look into the data
- ▶ Relationships discovered

# A deeper look into the data.

## First things that pop out

- ▶ First was at Athens at 1896 where 9 Sports were documented and 12 countries participated
- ▶ First Winter documented was at 1924 at Chamonix where 19 countries participated and 10 Sports played
- ▶ There are too many missing values about height and weight to have good educated results and be able to predict anything based on those.

# A deeper look into the data.

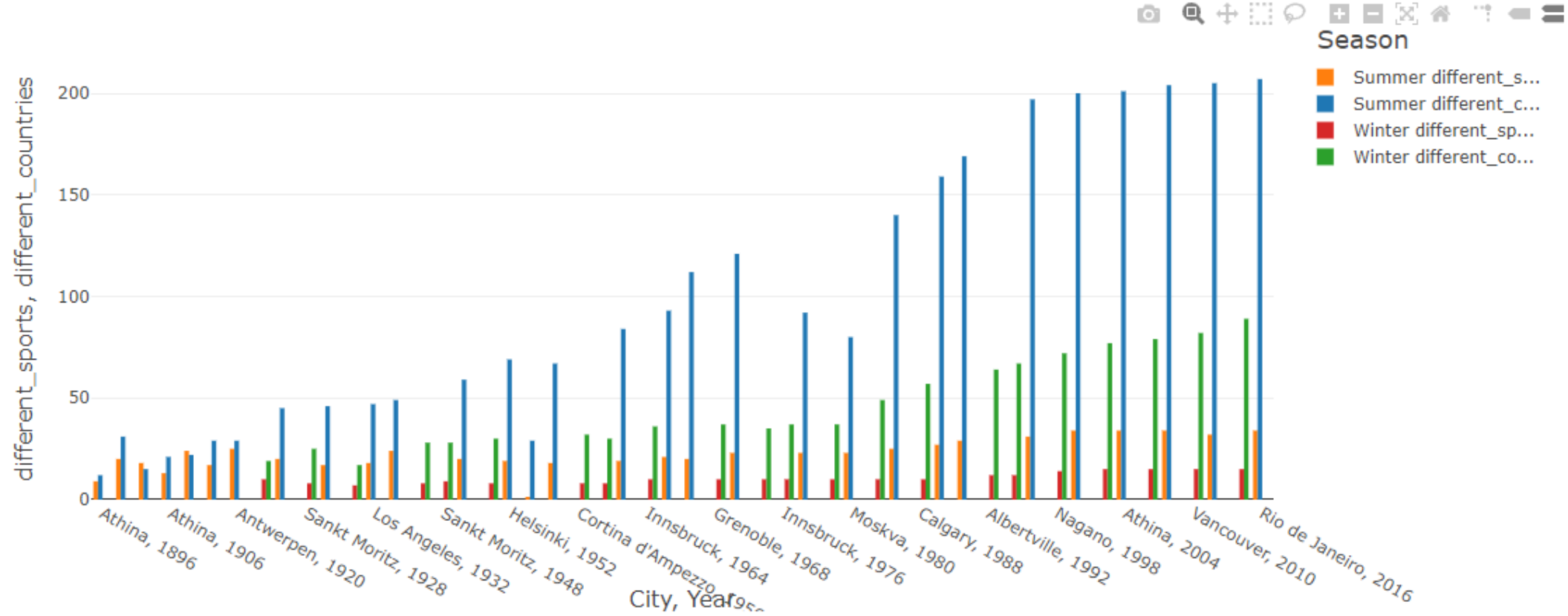
## Some observations

- ▶ Most of the Athletes are male (75%) and due to the 3:1 ratio advantage in relation to the women population I believe it would not be fair to compare any stats.
- ▶ Most of the athletes is between 18 kai 36 years old. But there have been athlete both 10 and 97 years old.
- ▶ There are some athletes who have been awarded 2,3 even 4 times in the same event.

# Relationships discovered

## Year and Popularity

► (3) Spark Jobs





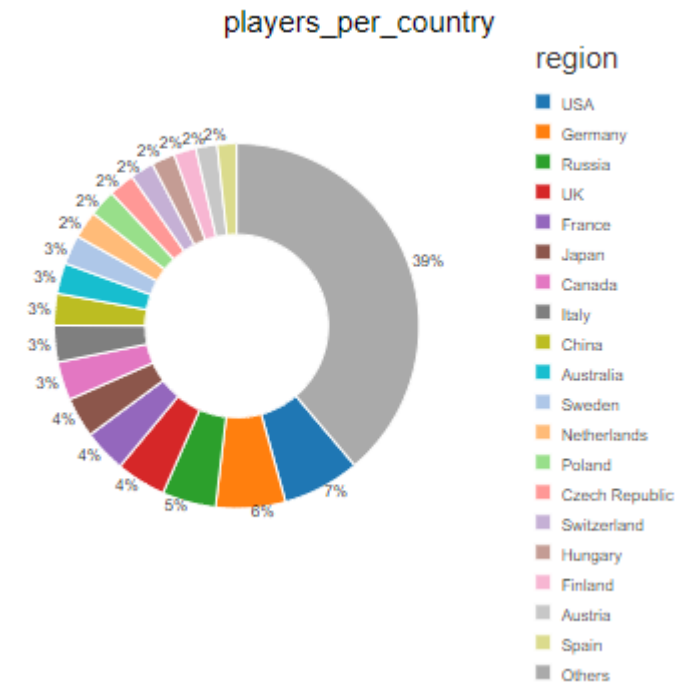
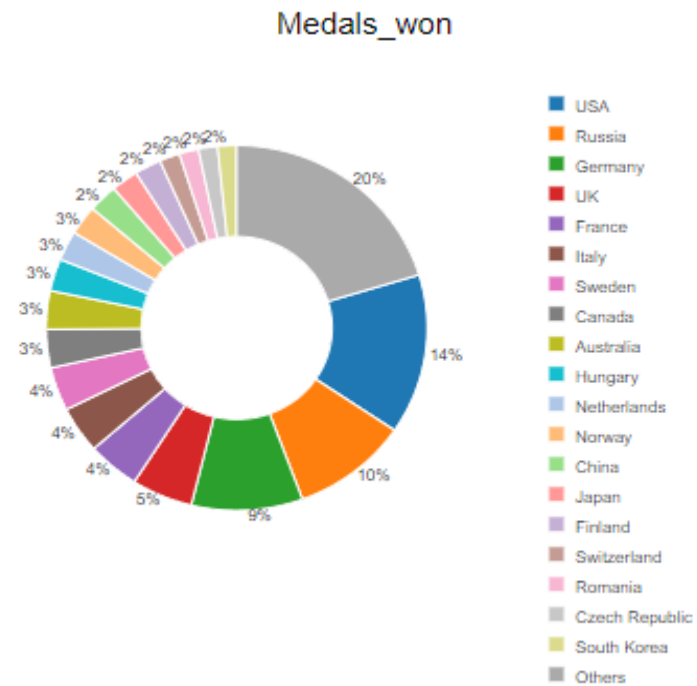
# Relationships discovered

## Year and Popularity

- ▶ This graph shows the by chronological order the cities in which the events took place and their popularity in terms of the following two questions
  1. How many countries participated in the event?
  2. How many different sports where played?
- ▶ It is very clear that as time progressed from the first event at Athens to the latest in Rio a lot more countries are joining and many sports are being added every time.

# Relationships discovered

## Players per country and Medals



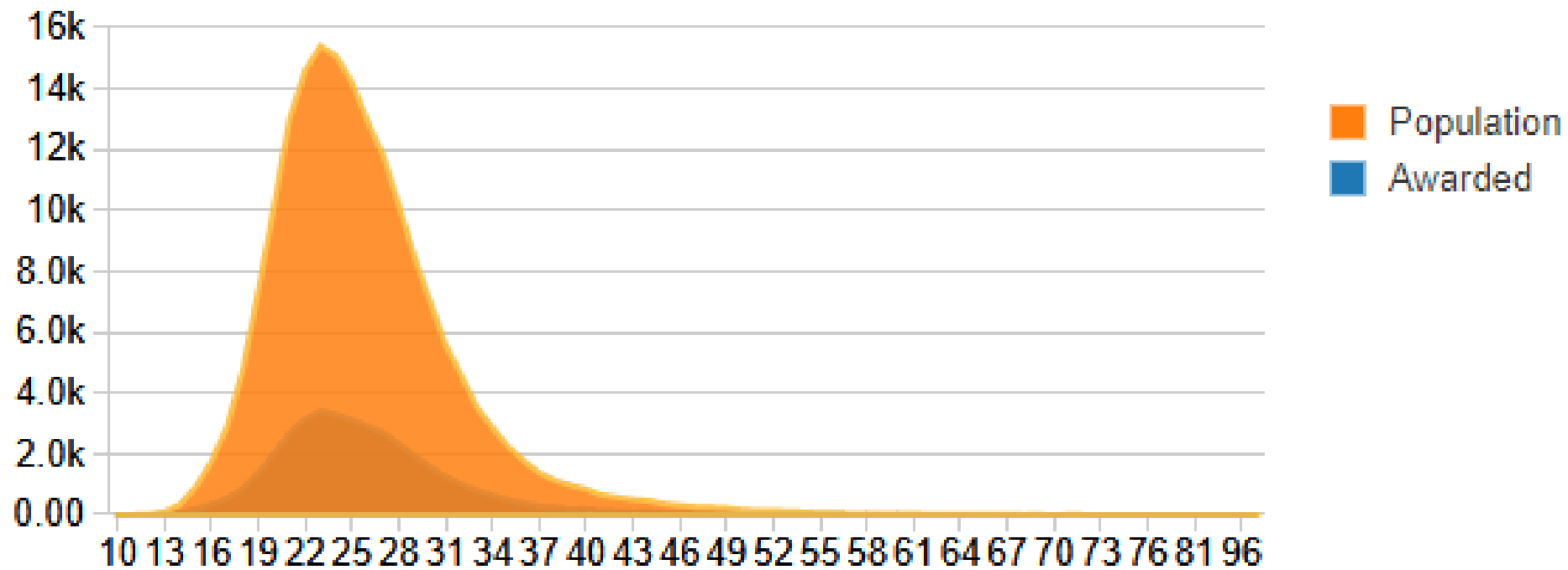
# Relationships discovered

## Players per country and Medals

- ▶ The pie chart on the left shows us the percentage of medals that each country has won over the whole history of the table.
- ▶ The pie chart on the right shows the percentage of the players it has sent to participate in these sports and events.
- ▶ By looking at these it is clear enough that the countries that have send overall more athletes have won more medals than the countries that send less.

# Relationships discovered

## Age and Medals



# Relationships discovered

## Age and Medals

- ▶ The area graph above shows the age distribution in the large orange area and the medals won by people in that age with the darker area.
- ▶ It agrees with my original hypothesis that there age would be a detrimental factor on medal winners and that the age around 24 would have the bigger success.

# Recommendations

- ▶ As for coaches and trainers who want to prepare athletes to be medal winners in sport my data driven recommendation would be to start training them from a young age because they are more likely to thrive in ages ranging between 18 and 30 so they must be ready by then.
- ▶ As for any company who wants to advertise any product it there is a big opportunity in these events and especially in summer events because more and more people are getting into them and a lot of countries are showing progressively more interested in joining.

The end!