Big Data Analytics

Olympics Sport and Medals

The modern Olympic Games or Olympics are the most important international sporting events in which thousands of athletes from around the world participate in a variety of competitions. More than 200 nations participate in it. The Olympic Games are normally held every four years, alternating between the Summer and Winter Olympics every two years.

The modern Olympic Games have origin from Greece, where they were held during the 8th century BC and to the 4th century AD. In 1894 was founded the International Olympic Committee (IOC) and in 1896 the first modern Games took place in Athens. The IOC is the governing body of the Olympics Movement, with the Olympics Charter defining its structure and authority.

The goal of my analysis is to answer to these questions:

- Which are the Countries that have won the most and have they changed over the years?
- Can we find correlation between the victories and the GDP per capita? And between the victories and the population?
- Who are the athletes who have won the most?
- How has female participation in games evolved over the years?
- Are men's and women's victories balanced in the nations that have won the most?

To realize my analysis, I used two different datasets obtained from Kaggle. One dataset, ***Olympic Sport and Medal, 1896-2014***, contains more than 35,000 rows and 22 columns. The other dataset***, Summer Olympics Medals (1976-2008)***, contains 15,000 rows and 11 columns including basically the same information - for a shorter period of time and excluding winter games-but it was more complete on information regarding the codes and names of countries dissolved or unified over the years.

Lets' see the details of them. The first dataset contains three tables: 'Summer', 'Winter', 'Dictionary'.

Summer contains information about the victories in the Summer Olympic Games from 1896 to 2014. It is composed by 9 columns, which are:

- 'Year', year in which the Olympic Games were held.
- 'City', city in which the Olympic Games were held.
- 'Sport', sport category (eg. Aquatics, Athletics etc.).
- 'Discipline', discipline within the sport (eg. Diving, Swimming etc.).
- 'Event', event name within the discipline (eg. 3m springboard, 10m platform etc.).
- 'Athlete', winner name in that event.
- 'Country', 3-character country code.
- 'Gender', gender of the athlete.
- 'Medal', type of medal won (eg. Bronze, Silver, Gold).

Winter contains the same 9 columns with the same features, but it regards the victories in the Winter Olympic Games from 1896 to 2014.

Dictionary contains information about IOC country codes and population/GDP estimates. It is composed by 4 columns:

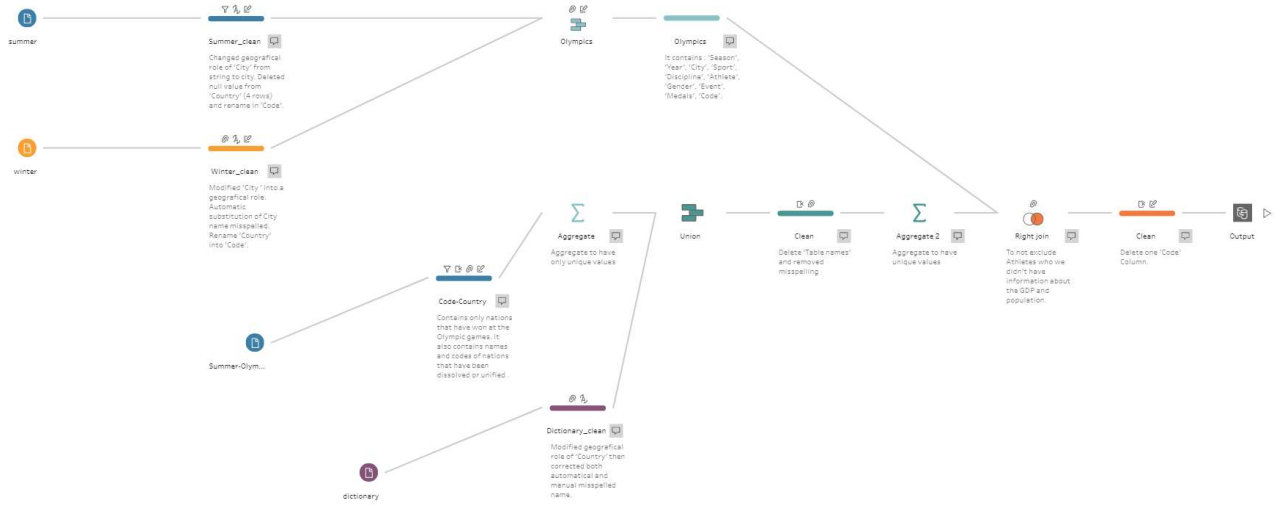
- ‘Country’, country to which the winning athlete belongs to.
- ‘Code’, 3-character country code.
- ‘Population’, discrete numerical value of the inhabitants.
- ‘GDP per capita’, continuous numerical value of the GPD per capita.

The second dataset contains one tables: *‘Summer-Olympic-medals-1976-to-2008’*. It is composed by 11 columns:

- ‘Year’, year in which the Olympic Games were held.
- ‘City’, city in which the Olympic Games were held.
- ‘Sport’, sport category (eg. Aquatics, Athletics etc.).
- ‘Discipline’, discipline within the sport (eg. Diving, Swimming etc.).
- ‘Event’, event name within the discipline (eg. 3m springboard, 10m platform etc.).
- ‘Athlete’, winner name in that event.
- ‘Country_code’, 3-character country code.
- ‘Country’, Country to which the winning athlete belongs to.
- ‘Gender’, gender of the athlete.
- ‘Medal’, type of medal won (eg. Bronze, Silver, Gold).
- ‘Event_Gender’, genders which participated in the event (Male, Female or Common Event).

Data Preparation

This is a general vision on all my preparation data:

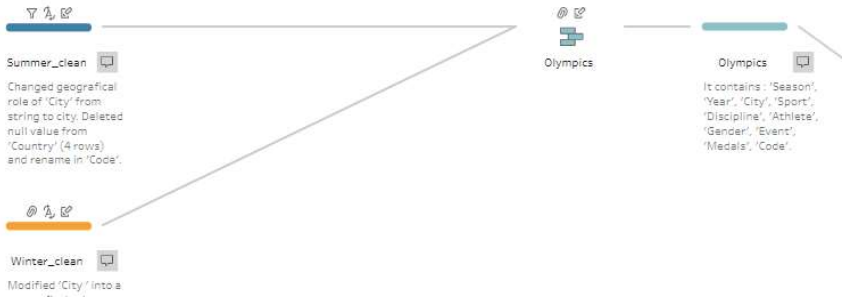


Let us see step by step:

I started *cleaning* the two table 'Summer' and 'Winter', that have the same features so I proceed in the same way.

First I modified 'City' geographical role from string to city, and renamed the columns 'Country' into 'Code'.

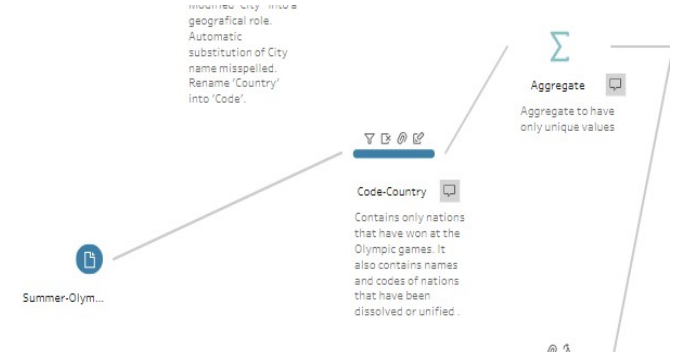
Then in the Summer table I deleted null value from 'Code' and in Winter table I used the automatic substitution to rename a City name that was mispelled.



The second step was an *union* of the two table, in which I renamed two values in the 'Table Names' and then renamed the column into 'Season'.

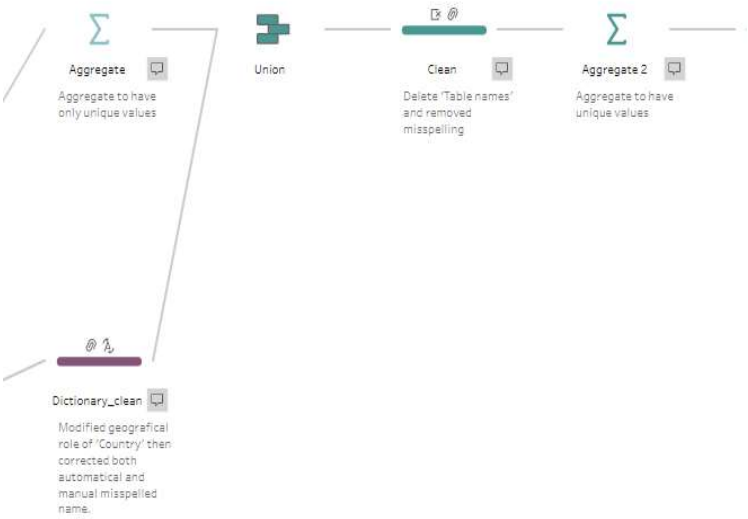
Initially I did not plan to use the dataset '***Summer Olympics Medals (1976-2008)'***, but during my data preparation I realized it was necessary. In fact, with the first dataset, during the join of the union table -made with summer and winter tables- with the dictionary table, there was a large number of mismatched values and this was due to the fact that the dictionary table only contained information of the existing country and left out the country now dissolved or unified, of which we knew only the code-country.

So, after a research I found the second dataset which instead contained the country code of dissolved or unified nations combined with the name of the nation. I proceed in this way:

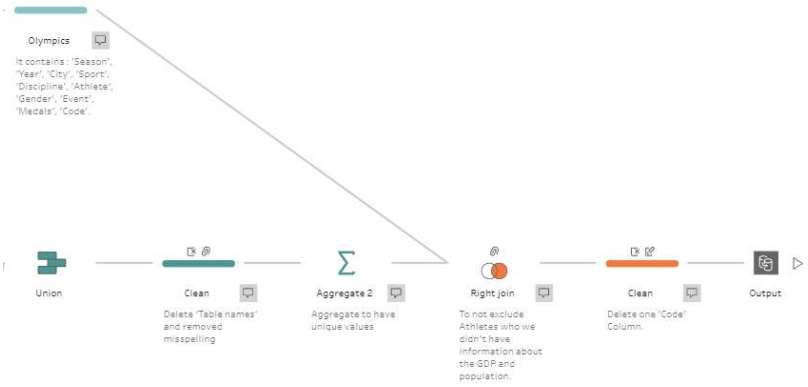


From the 'Summer Olympics Medals' table I deleted all the columns that I didn't need, in particular: 'City' 'Year', 'Sport', 'Discipline', 'Event', 'Athlete', 'Gender', 'Event Gender', 'Medal'. Then I excluded the Null values, substituted one misspelled Country name and renominated 'Country_code' into 'Code'. At the end I had two columns: 'Code' and 'Country'. Then I *aggregated* both Code and Country to have only unique values.

From the Dictionary table, I modified the geographical role of 'Country' from string into region. Then I substituted 15 misspelled Country names.



Eventually, I made a *union* from the aggregate and the dictionary clean, between the columns 'Code' and 'Country'.
So at the end I had: 'Population', 'GDP per capita', 'Code', 'Country'.
After that I *cleaned* the table removing the column 'table names' and I proceeded with a second *aggregation,* again between 'Code' and 'Country'.



Next, I created a *right join* with the table 'Olympics' and 'Aggregate 2', seen before. The join was between 'Code' of both tables. I've chosen the right join because I didn't want to exclude from my analysis the victories from 'Athletes' which we didn't have informations about the GDP and population. I have instead excluded Countries in which there had never been an Olympic victory. I also replaced two codes that were related to the same country but were different from table to table (TTO into TRI that refers to Trinidad and Tobago, and SGP into SIN that refers to Singapore).

Finally, I *cleaned* the table removing one 'Code' column and renominated the other into 'Country-code and created the *output*.

Dimensional Fact Model

Dimensional Fact Model is a graphical conceptual model for data marts, designed to support the conceptual design, create an environment in which user queries may be formulated intuitively, make communication possible between designers and end users to refine requirement specifications, build

a stable platform from which starting the logical design, provide clear and expressive design documentation.

The conceptual representation generated by the DFM consists of a set of fact schemata. Fact schemata basically model facts, measures, dimensions, and hierarchies.

A ***Fact*** is a concept relevant to decision-making processes. A fact is represented by a box that displays the fact name along with the measure names.

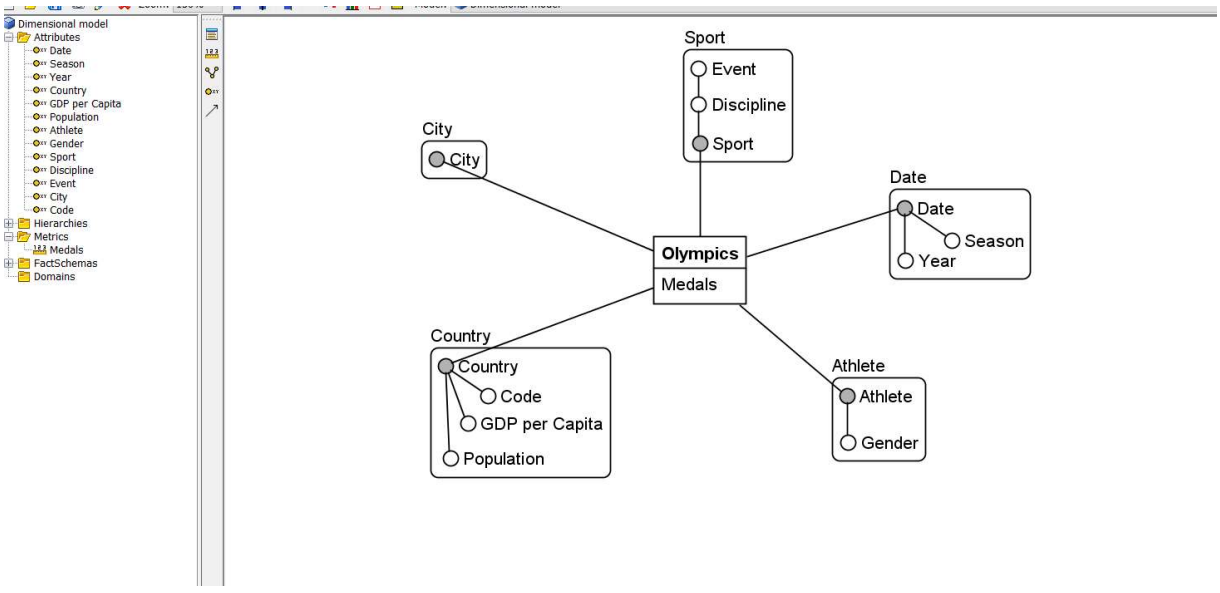
A ***Measure*** is a numerical property of a fact that describes a quantitative attribute that is relevant to analysis. A fact can also have no measures, as in the case when you might be interested in recording only the occurrence of an event.

A ***Dimension*** is a property, with a finite domain, that describes an analysis coordinate of the fact. A fact generally has multiple dimensions that define its minimum representation granularity. Small circles represent the dimensions, which are linked to the fact by straight lines.

A ***Dimensional attribute*** is a property, with a finite domain, of a dimension. Like dimensions, a dimensional attribute is represented by a circle. The relationships among the dimensional attributes are expressed by hierarchies.

A ***Hierarchy*** is a directed tree whose nodes, represented by circles, are dimensional attributes and whose arcs represent relationships between pairs of attributes. A hierarchy includes a dimension, positioned at the tree's root, and all the dimensional attributes that describe it. Hierarchies define the way elemental business events can be selected and aggregated for decision-making processes.

Using My Business Intelligence Modeler software I was capable to create my DFM using the already descripted variables.



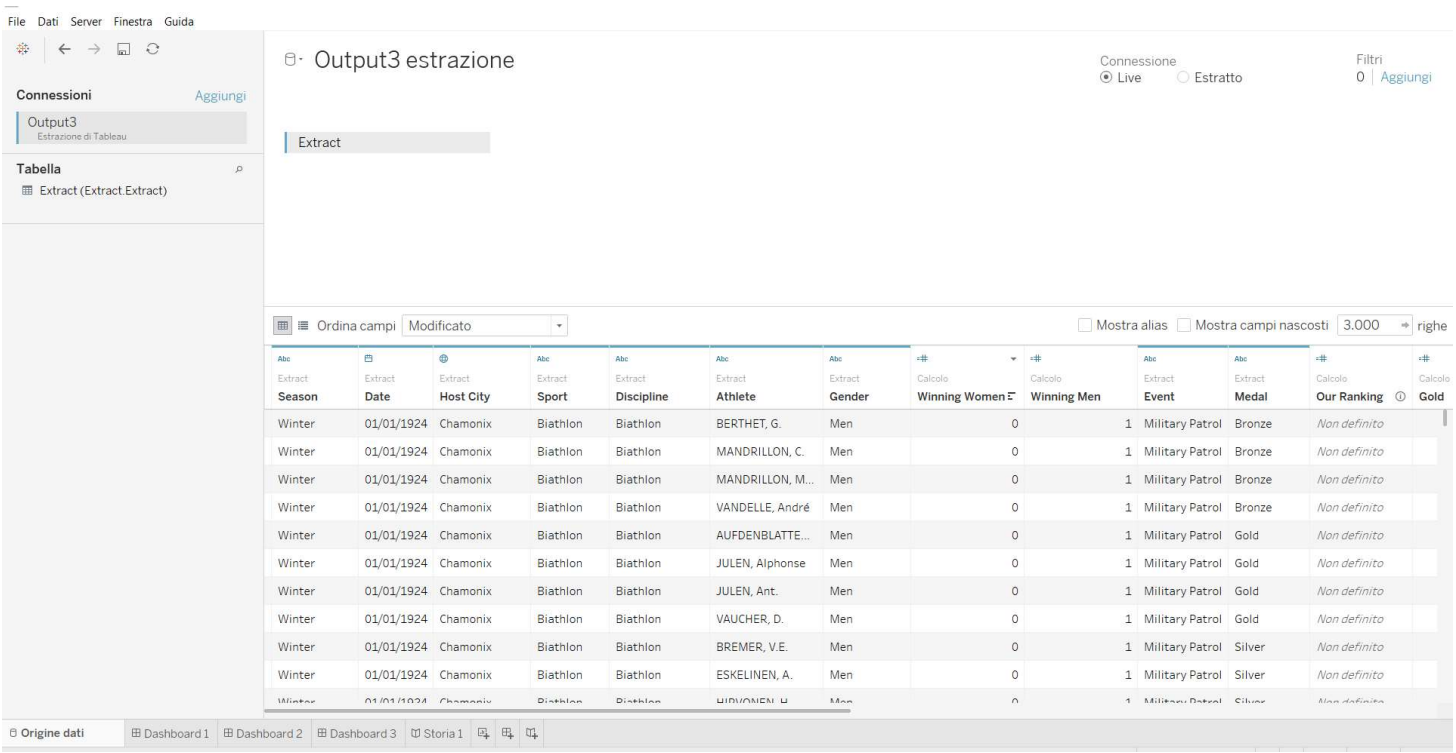
As a fact I used Olympics, the focus of my research, and as its measures I used ‘Medals’ because the count of it is the numerical attribute more relevant in my study.

As dimensions I used ‘City’, ‘Sport’, ‘Date’, ‘Athlete’ and ‘Country’. All of them except ‘City’ had two or more dimensional attributes, which contributed to characterize the respectively dimensions and the hierarchy within them is clear from the picture above.

Data Visualization

Before I start presenting my data analysis, let us start by describing the source of the data.

I originally had 13 columns, coming from the data preparation, but to create the charts that I imagined I needed to create new columns through what is called ‘calculated field’. In this way I created 9 more columns and I will describe them during the presentation of the specifical chart where I used them.



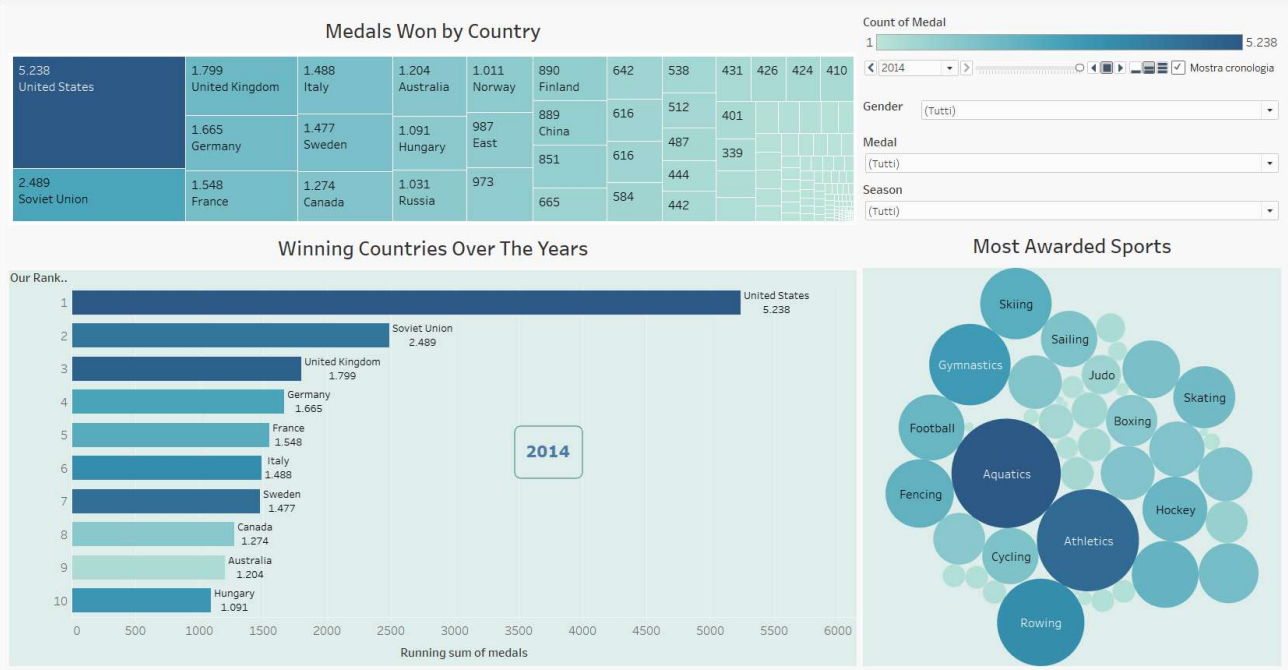
The extracted columns are:

- **Season**, string
- **Date**, date
- **Host City**, string with geographical role of ‘City’
- **Sport**, string
- **Discipline,** string
- **Athlete,** string
- **Gender,** string
- **Event,** string
- **Medal,** string
- **Country-Code**, string
- **Country**, string with geographical role of ‘Country’
- **GDP per Capita**, decimal number
- **Population**, decimal number

First dashboard

The first dashboard I created answer to the questions :

Which are the Countries that have won the most and have they changed over the years? And what are the most awarded sports?

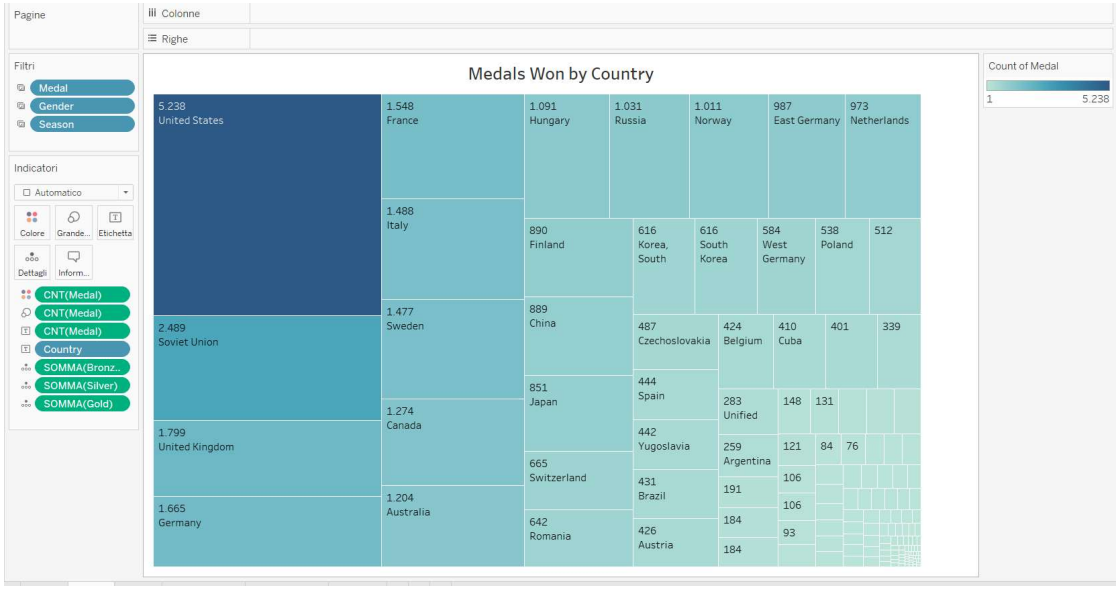


This is the first dashboard I made, it is composed by three different sheets: 'Medals won by Country', 'Winning Countries Over The Years' and 'Most Awarded Sports'. I have chosen for all of them to be colored of different shades of blue, to not have a blurred vision of the dashboard. In particular the color will go from light blue to dark blue, the more medals won, the darker the color.

We can also see the four *filters* and in particular: **Gender, Medal and Season**, applied to all the charts, and **Year** applied only to the chart 'Winning Countries Over the Year' that is needed for the race bar chart that I will explain later.

I also made the chart 'Medals Won by Country' became itself a filter therefore we could explore how a specific country evolved through the years and what are the most awarded sports for it.

The first chart we look at is on the top and it reports the count of medals won by every country, it is called **'Medal Won by Country'** :



A tree map is a visualization that nests rectangles in hierarchies so you can compare different dimension combinations across one or two measures (one for size; one for colour) and quickly interpret their respective contributions to the whole.

I used 6 measures and in particular:

- Count (Medal) to define the *colour* of the map
- Count (Medal) to define the *size* of the rectangles
- Count (Medal) as a *label* in the rectangles
- Sum (Bronze) for the *details* of the rectangles
- Sum (Silver) for the *details*
- Sum (Gold) for the *details*

For the last three measures it was necessary a calculate field, in this way I could have easily obtain a sum of them. To create it, I used this formula:

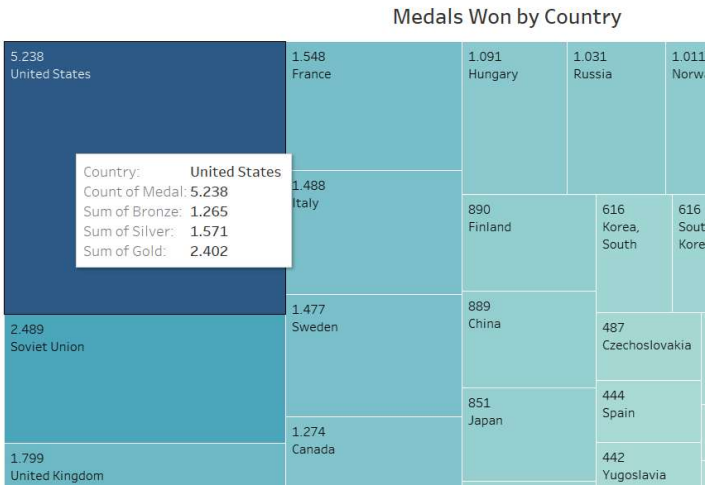
By changing three time the word in quotes in ‘Gold’, ‘Silver’ and ‘Bronze’.

With the function INT, I obtained a numerical value that was equal to 1 if the value contained in Medal was the same of the one contained in the quotes, otherwise it was 0.

And 1 dimension:

- Country as a *label*

If we hover the mouse over a rectangle of the map, we will get this:



Name of the Country,

Total Count of Medals,

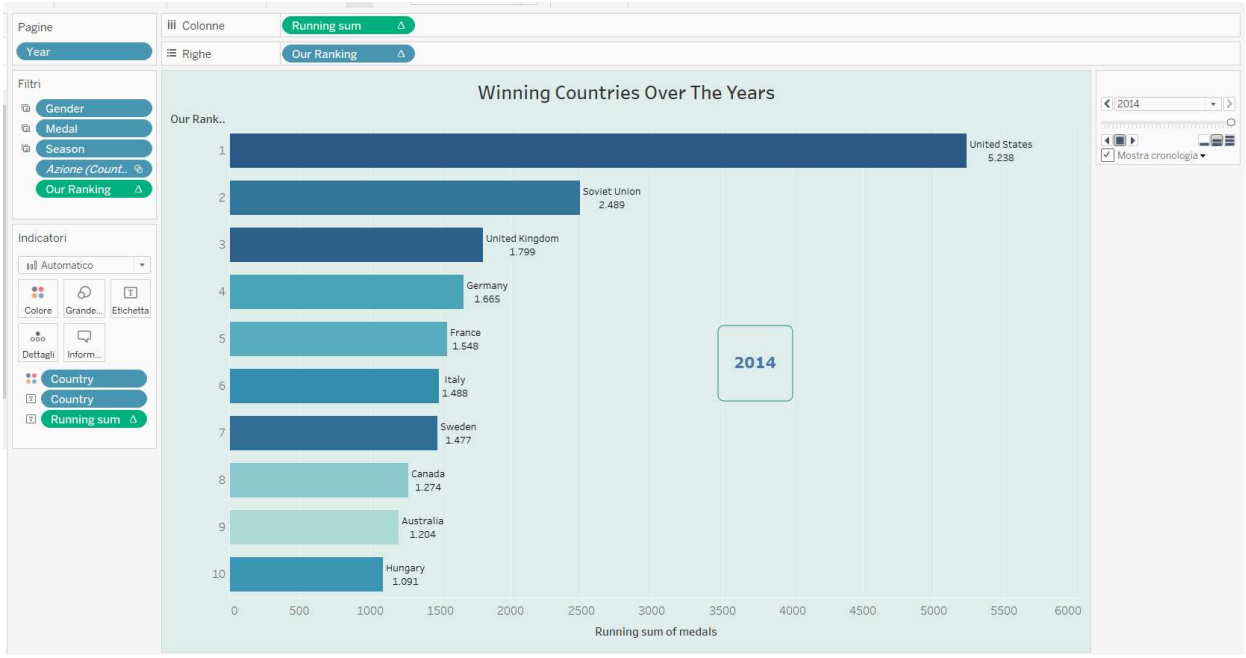
Sum of Bronze, Silver and Gold.

We can easily see by the dark colour and the dimension that the United States is the country that has won the most (5238 medals), followed by the Soviet Union (2489 medals), the United Kingdom (1799 medals) and so on.

The *filters* used in this map are:

- **Medal**, we can choose to show only one or two type of medal.
- **Gender**, we can choose to show only male victories or female victories.
- **Season**, we can choose between summer or winter Olympics.

Let us see now the second dashboard at the bottom left that I called **'Winning Countries over the Years':**



It is a race bar chart, so it means that it is animated. I wanted to show the evolution through the year of the first 10 countries for numbers of medals won.

To make it I had to create two calculated fields: ‘Running Sum’ and ‘Our Ranking’.

For the first I used the function Running sum of the count of medal.

For the second calculated field I used the function Rank_unique of the Running sum, created before.

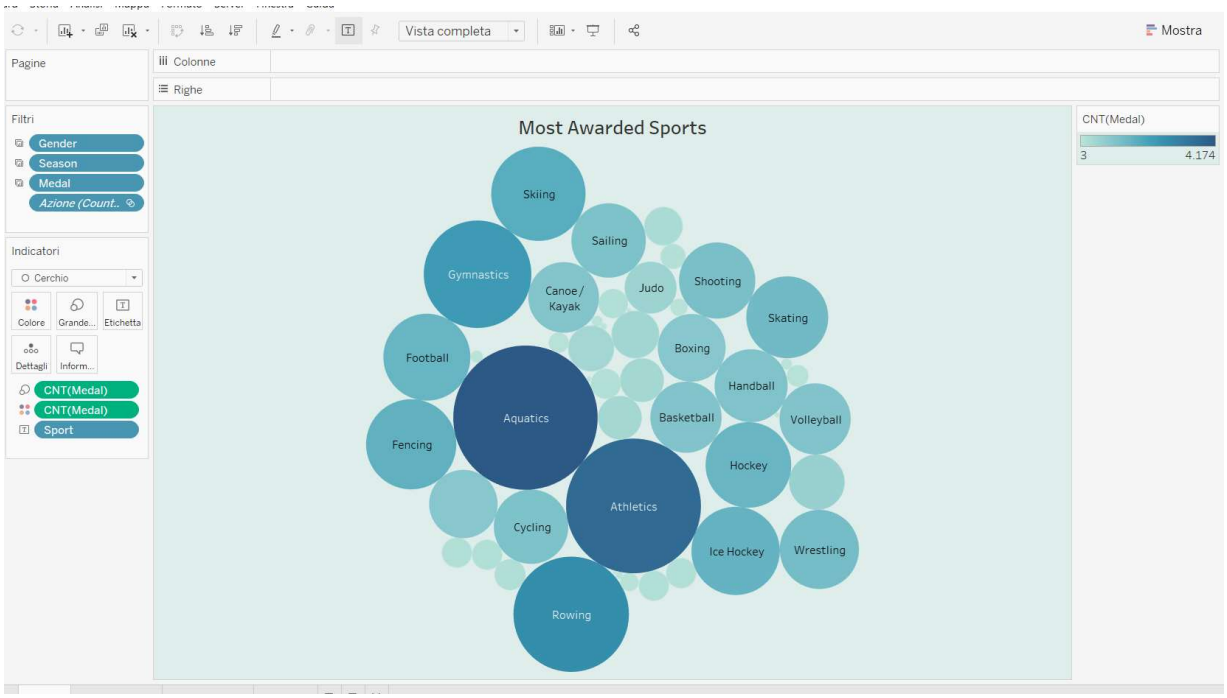
I put in column ‘Running_sum’, calculated through the ‘Year’ and in row ‘Our ranking’ calculated using ‘Year’ and ‘Country’ (at the level: Country, reboot every: Year).

As indicators I used ‘Country’ to indicate the *colours* of the bars and both ‘Country’ and ‘Running sum’ to be the *label*.

As *filters* we find again **Gender, Medal, Season**, the action of the **map** already described above, and in addition we find ‘**Our Ranking**’ because I decided to show only the top ten countries of the ranking.

To make the bar chart animated I added to the field ‘*Pages’* the calculated field ‘**Year**’ that just shows the year from ‘Date’, without months or days. We can see on the right of our sheet the possibility to show a particular year or to select the first year and then press the play button and see the animation going on.

The last chart of the first dashboard is **'Most Awarded Sports'**:



I wanted to show which sports were accountable of most medals, and to do this I chose the bubble chart that encodes data using size of circle to show comparisons.

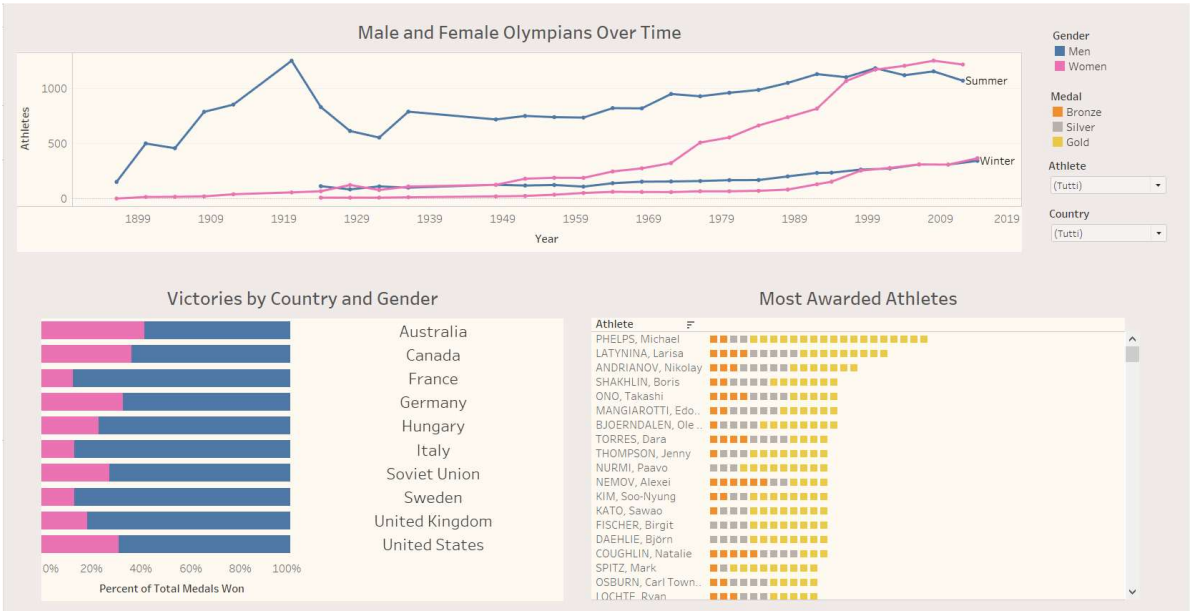
I used Count (Medals) to specify the *colours* and *size* of the bubbles, and Sport as the *label* of the bubbles.

As *filters* we find again **Gender, Medal, Season**, and the action of the **tree map**.

The results show us that the most awarded sport is Aquatics, followed by Athletics, Rowing and Gymnastics.

Second dashboard

The second dashboard I created answer to these questions:
What is the gap between male and female victories? Who are the most awarded athletes?



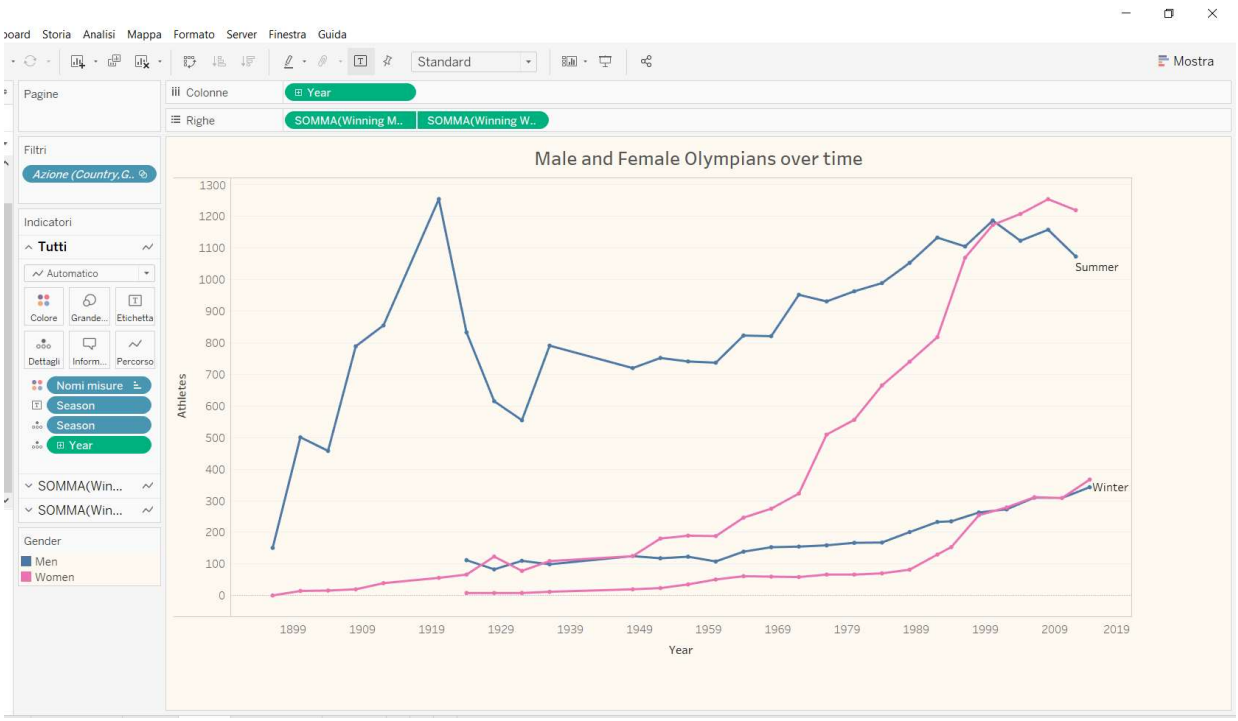
The second dashboard I made is composed by three sheets: ‘Male and Female Olympians Over Time’, ‘Victories by Country and Gender’ and ‘Most Awarded Athletes’.

In the first two plots I made a separation between male and female so I decided to color them with blue and pink. In the third chart I show each medals won by an Athlete and I decided to use the color of the three different kind of medal, bronze silver and gold, in their rispective colours.

On the top right of the dashboard we can see the two legends that explain what each colour stands for and two *filters* and in particular: **Athlete, and Country** that are applied only to the sheet ‘Most Awarded Athletes’.

I also made the chart ‘Victories by Country and Gender’ became itself a filter in this way we could explore the evolution of winnings by male and female olympians through the time of one of the top ten country or have a look on the medals won by athletes coming from that country.

The first chart we look at is on the top and it reports the evolution during the years of the winning of male and femal olympians during the summer and winter olympics and to do this I chose a Dual Lines Chart with Synchronized Axis, it is called **‘Male and Female Olympians Over Time’**:



We can see in 4 lines: blue ones are used to indicate male athletes, pink ones to indicate female athletes. Then you can see on the right of the chart the labels ‘Summer’ and ‘Winter’ that indicate the season of the olympics.

Summer Olympics game started in 1896 in Athene (Greece), at that time women were not allowed to participate. Women were able to participate only in 1900, but the competitions open to them were far lower than those of men, this also explains the lower number of medals won.

In 1920 we see a peak in the number of men's medals won. This is explained by the fact that in that year new competitions were added such as those of shooting swimming.

In 1924 the Winter Olympics were born and were held in Chamonix-Mont-Blanc (France).

During the following years we witness a progressive increase in the number of medals won by women athletes, which come to equal the male ones, both in the Summer and Winter Olympics.

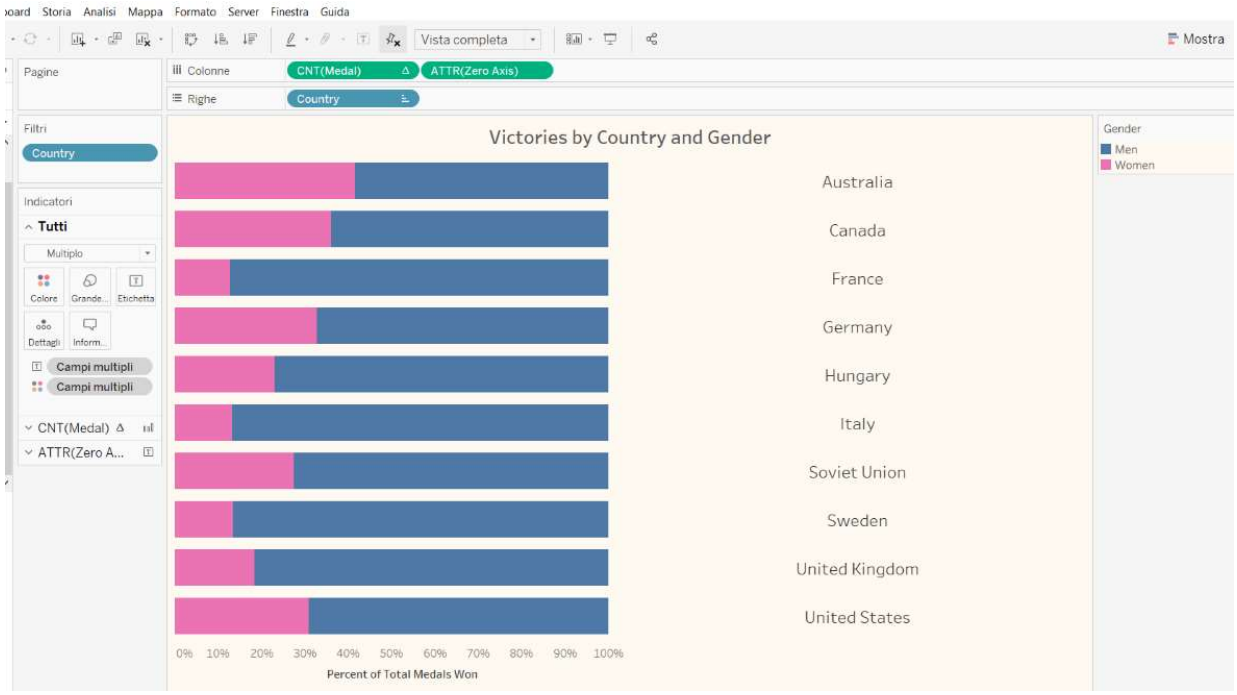
To create this chart, I used in columns ‘Year’, and in rows the sum of ‘Winning Women’ and the sum of ‘Winning Men’. These last two are calculated field, that I built in this way:

With the function INT, I obtained a numerical value that was equal to 1 if the value contained in Gender was the same of the one contained in the quotes, otherwise it was equal to 0.

The same formula was used to build Winning Men but changing the word in quotes in ‘Men’.

As filters I used only the action on the chart ‘Victories by Country and Gender’, to look at the evolution, of a particular one of the top 10 Countries, in the count of medals.

The second table I am going to show you is called **‘Victories by Country and Gender’**, it shows the percentage of medal won, respectively by women and men, for the first 10 Countries.

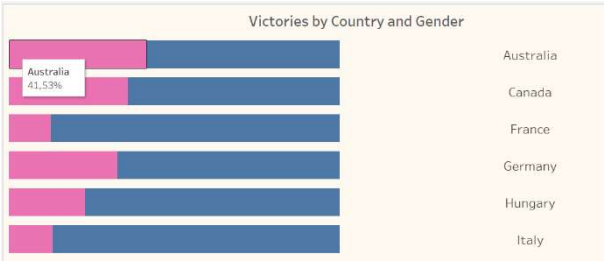


On the left of the chart we can see the bars that indicate the percentage of medal won. To distinguish between the count of medals of female and male I used the same colours of the plot above, pink and blue.

On the right of the chart we can see the name of the respectively Countries, put in alphabetical order.

To create this I used in columns 'Count(Medal)' and calculated a *percentage of the total* using 'Gender', and I created a calculated field called 'Zero Axis' in which I could insert 'Country' as label. In row I used 'Country'. As filter I used 'Country' to show only the first 10 countries having by the count of medals.

Hovering the mouse over the pink or blue part of the bar, we will get the percentage of that country calculated for female and male athletes:

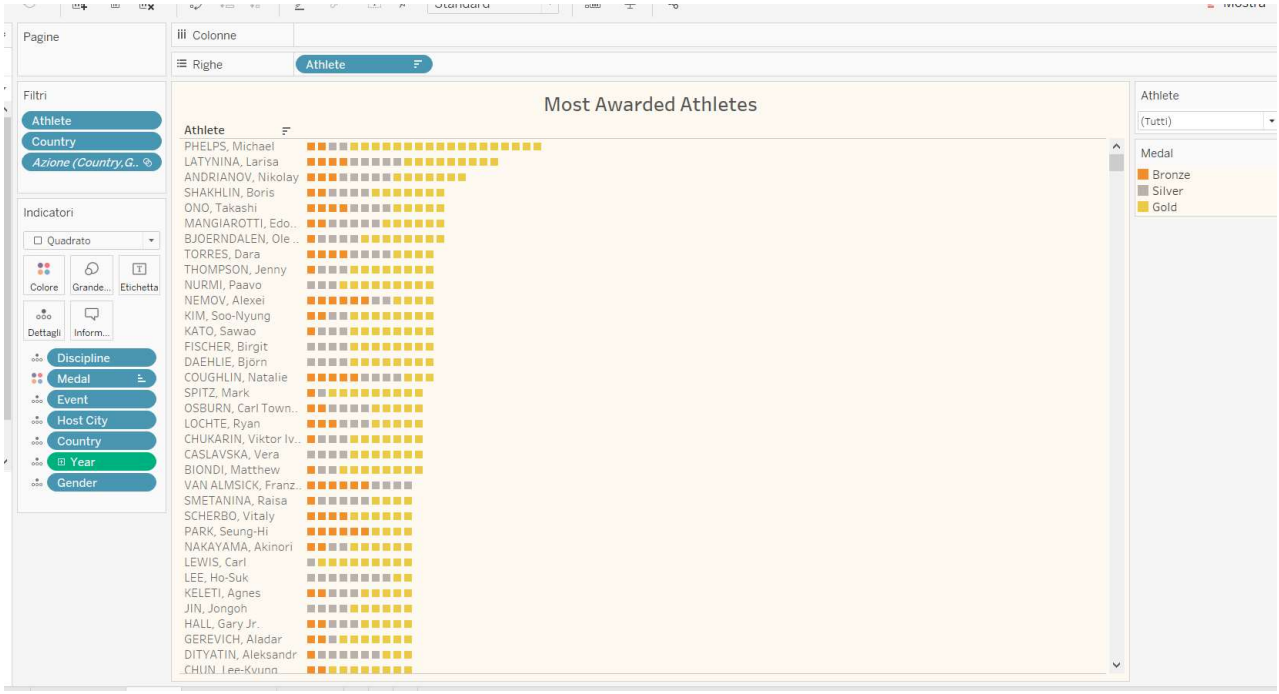


Having a general look on all the chart we can see a discrepancy between the count of medals won by female and male. In particular the ones won by femal are very often a small part of the percentage, just three countries differ : Australia, Canada and Germany.

We can presume that this discrepancy is due to the fact that during the first Olympic games only a small number of disciplines was open to women and that has contribuited in having a smaller count of medals. But, after having seen the positive trend of the last fifty years of the number of medals won by women increased, we can hope that this discrepancy will be reduced or, better, eliminated entirely.

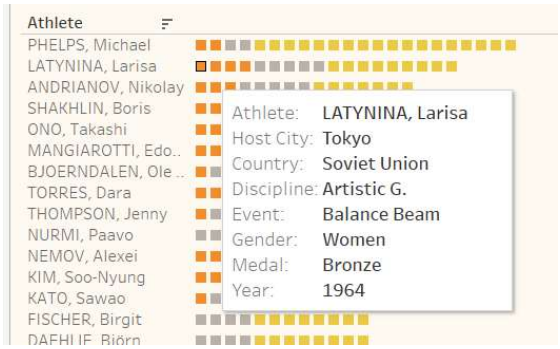
The last chart of the dashboard is '**Most Awarded Athletes**'.

I wanted to create a ranking with all the medals won by an athlete, differentiate for the type of medal (Bronze, Silver and Gold), and that included all the information about the discipline, the event, the year or the host city of where that particular medal was won.



We can see on the left the name of the athletes and next to them squares that reminds the idea of medals, all of them coloured by the specific type of medal.

To build it I put in row 'Athletes', then as indicators I used 'Medal' for the colour of the squares, 'Discipline', 'Event', 'Host City', 'Country', 'Year' and 'Gender' as details.



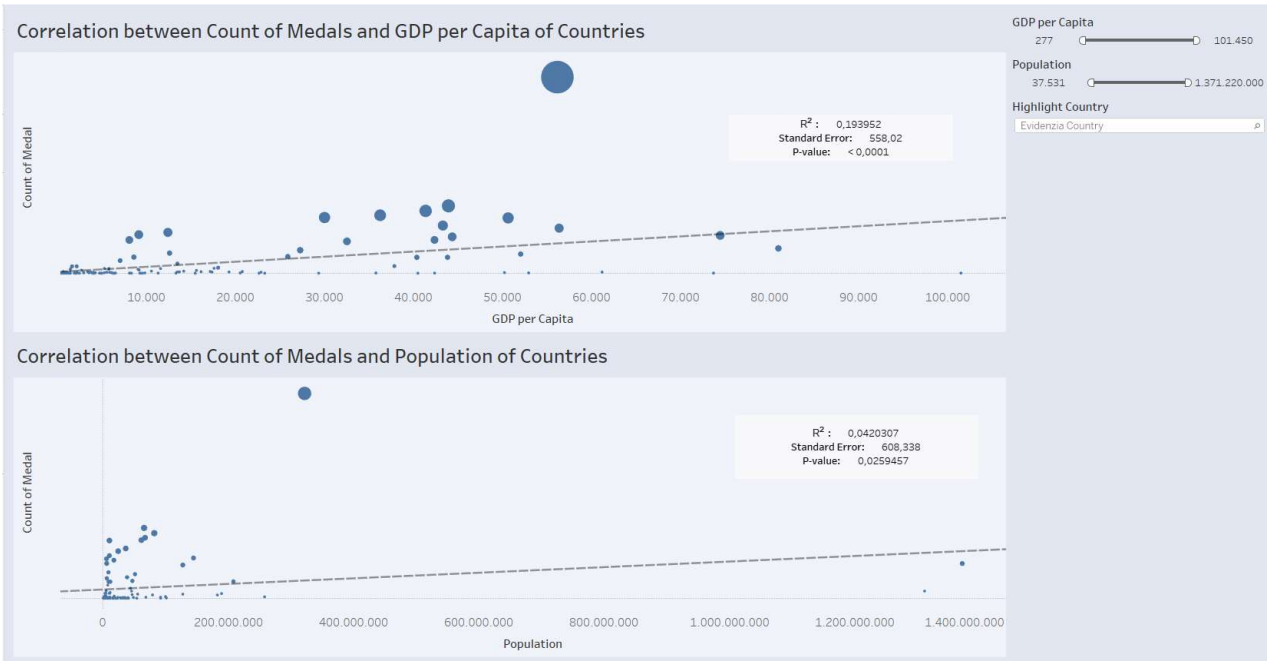
This is what we get if we hover the mouse over a square.

Athletes are shown in a descending nested order obtained by the count of medals. Also, medals are ordered and showed from bronze to gold.

The filters used in this sheet are Athletes, Country and the action we make on the chart 'Victories by Country and Gender' when we see them in the dashboard.

Third dashboard

In the last dashboard I wanted to show the **'Correlation between the Count of Medals and GDP per capita and Population of Countries'.** I wanted to understand if a greater number of medals won could be explained by a greater wealth or greater inhabitants of a country, and this is what I found:

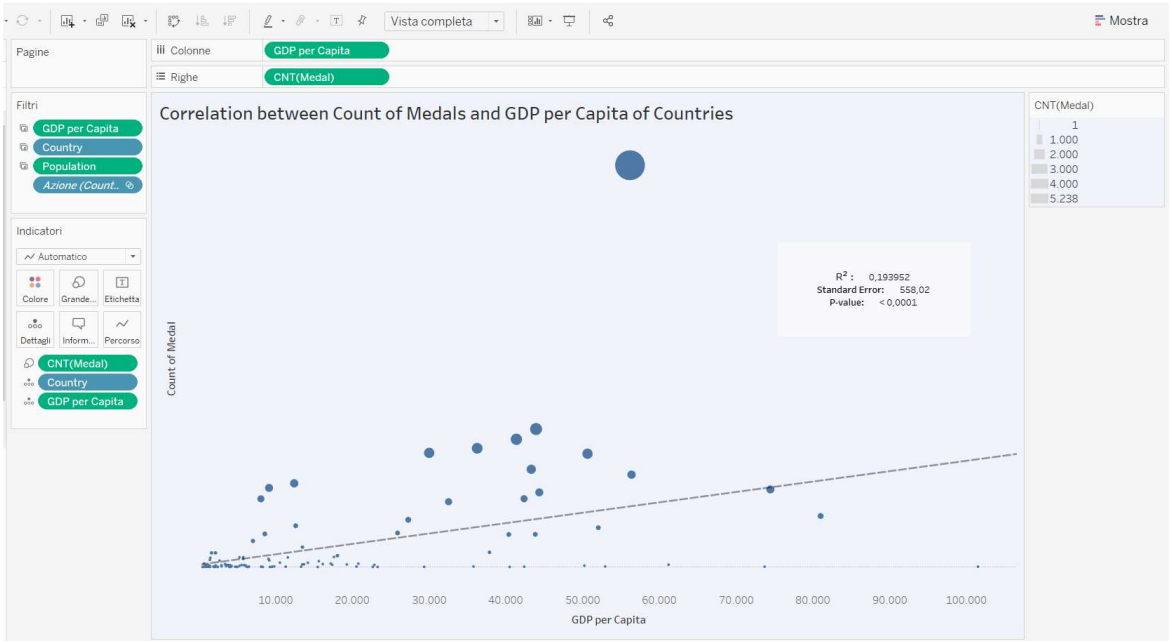


The first chart shows a positive correlation between the two variables, so this means that a greater number of medals won is associated with a greater GDP per Capita.

To build this plot is collocated in columns 'GDP per Capita' and in rows the count of 'Medal', then as indicator I used Count of medal to indicate the size of the circles, Country and GPD as details. As filters I used 'GDP per Capita', 'Country', 'Population' and the action on the other plot.

At that point I checked the option to be able to see the trend line in the chart, and thanks to 'describe trend pattern' I was able to get more precise information about the correlation.

Next I reported some of those informations in an area of the graph.

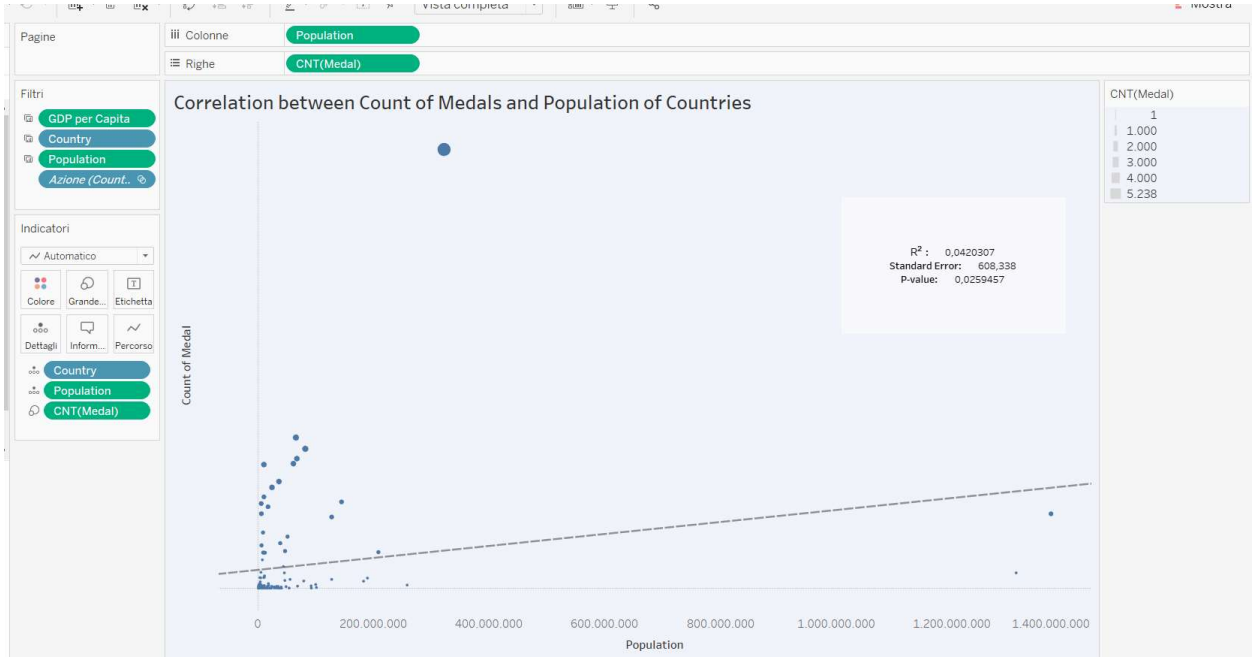


Those values refer to:

- **R^2**, statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable. R-squared values range from 0 to 1. An R-squared of 1 means that all movements of a dependent variable are completely explained by movements in the independent variable(s).
- **Standard Error**, measures how far the sample mean of the data is likely to be from the true population mean. When the SE is small, the data is said to be more representative of the true mean. In cases where the SE is large, the data may have some notable irregularities.
- **P-value**, the null hypothesis states that there is no relationship between the two variables being studied (one variable does not affect the other). It states the results are due to chance and are not significant in terms of supporting the idea being investigated. The alternative hypothesis states that the independent variable did affect the dependent variable, and the results are significant in terms of supporting the theory being investigated. The level of statistical significance is often expressed as a p-value between 0 and 1. The smaller the p-value, the stronger the evidence that you should reject the null hypothesis.

In our case the values were very lower in terms of p-value (<0.0001) meaning that there is a relation between the two variables. But our R^2 have a value close to zero (0.29) meaning that just a 20% of the variation of one variable is completely explained by the other. Also, the Standard Error does not have good value, so the analysed sample is not sufficiently representative to be able to study. I think it would be interesting to deepen and carry out more complete statistical studies.

The second plot of the dashboard is called ‘Correlation of Count of Medals and Population between Countries’. It is basically the same plot of the one above but in this case, I put on columns ‘Population’ instead of GDP. As indicators I used ‘Population’, ‘Country’ for details and count of ‘Medal’ for the size of the circles.



As well, I checked the option to be able to see the trend line in the chart, and obtain more precise informations about the correlation thanks to 'describe trend pattern' .

Also the second chart shows a positive correlation between the two variables, so this means that a greater number of medals won is associated with a greater Population.

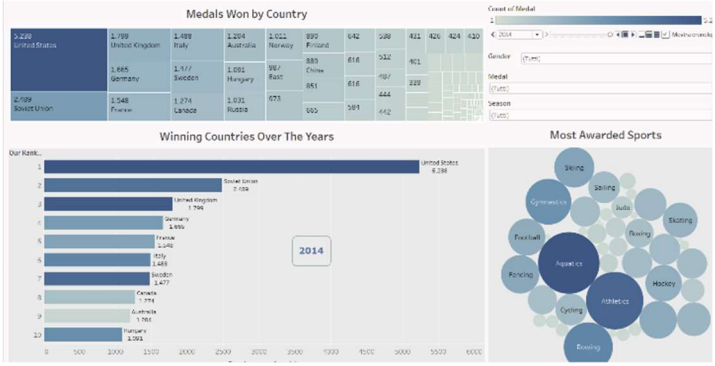
Again, the values were very lower in terms of p-value (<0.02) meaning that we reject the null hypothesis. R^2 have a value close to zero (0.04) meaning that the variation of one variable is not explained by the other. The Standard Error does not have good value, so the analysed sample is not sufficiently representative to be able to study.

Colour Blindless Simulation

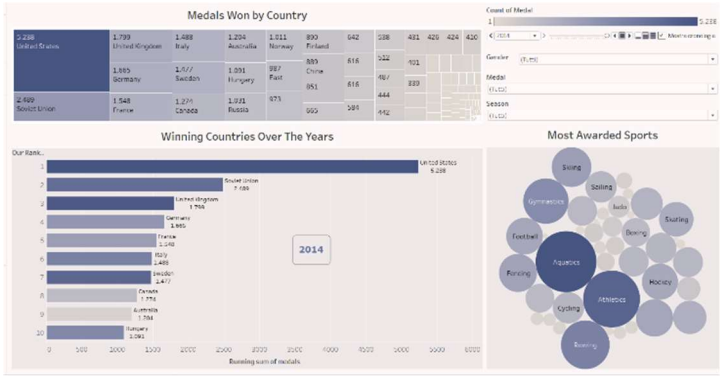
It is important to make sure that the dashboards can also be used by people with color blindness. Color blindness (color vision deficiency) is the decreased ability to see color or differences in color. Red–green color blindness is the most common form, followed by blue–yellow color blindness and total color blindness.

Using the tool provided by ColorBlindor, we are able to understand if the dashboard created can be visualized by color blind people.

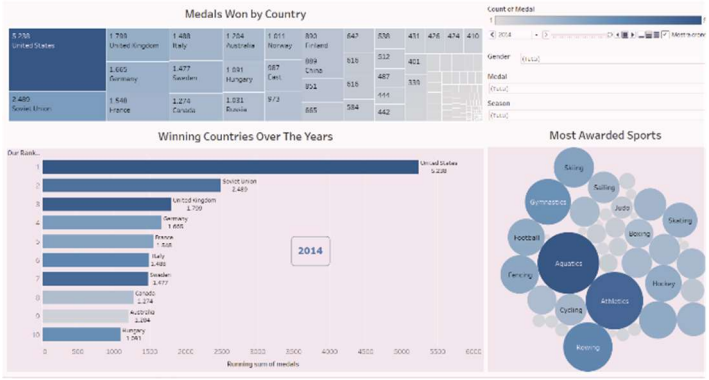
Let’s start with the first dashboard:



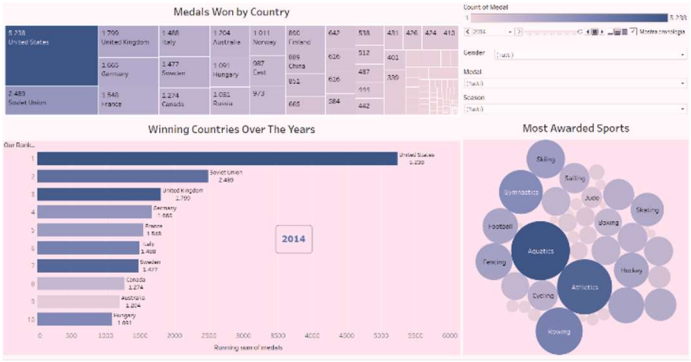
Red-Weak / Protanomaly



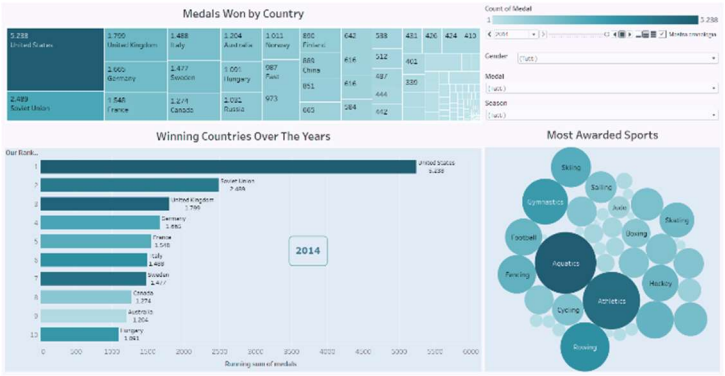
Red-Blind / Protanopia



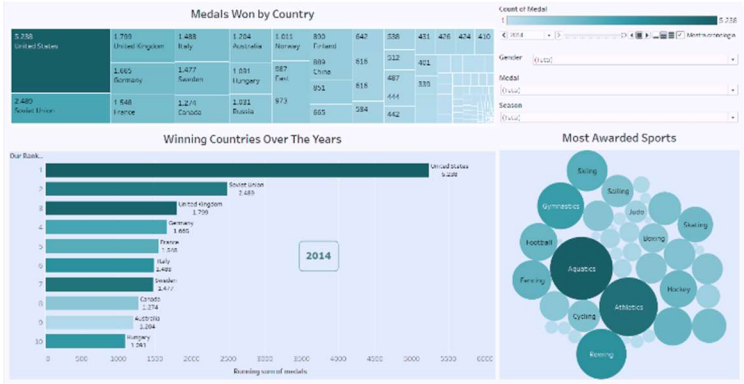
Green-Weak / Deuteranomaly



Green-Blind / Deuteranopia



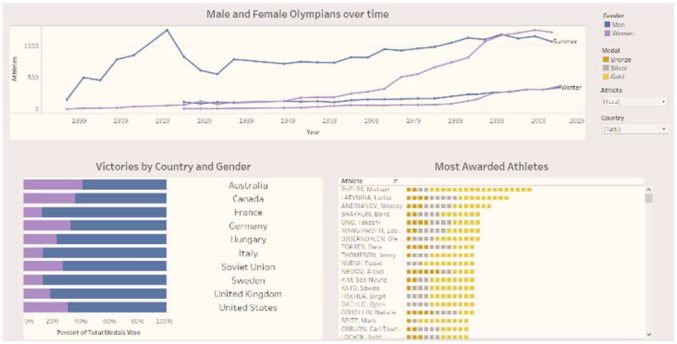
Blue-Weak / Tritanomaly



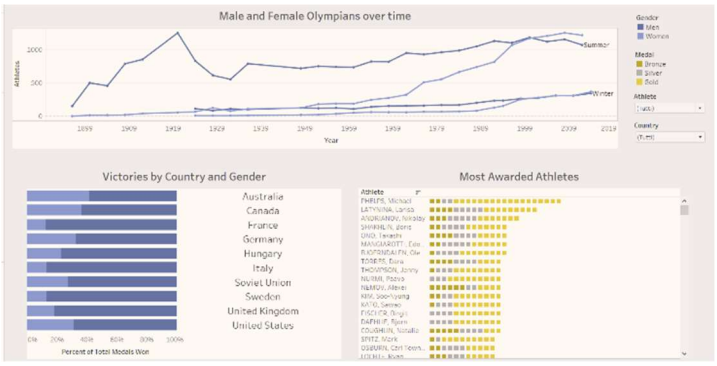
Blue-Blind / Tritanopia

We can affirm that the first dashboard is easily navigable even by those suffering from color blindness, as we can distinguish a more intense color that indicates a greater number of medals.

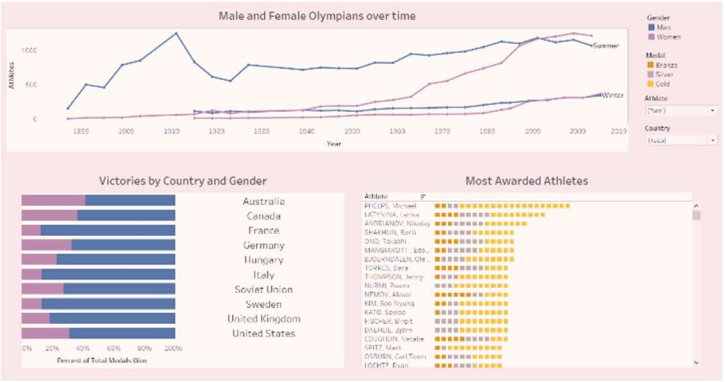
Test on the second dashboard:



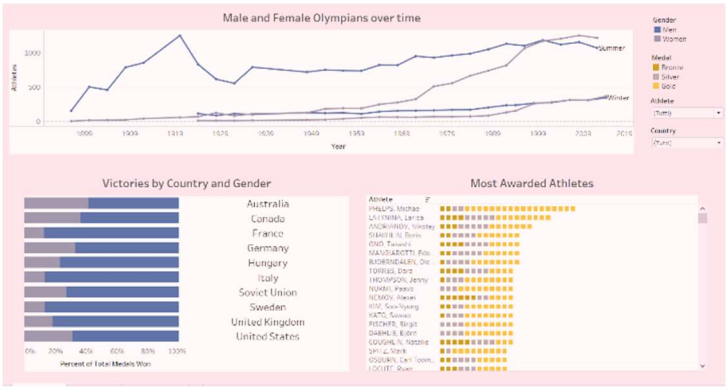
Red-Weak / Protanomaly



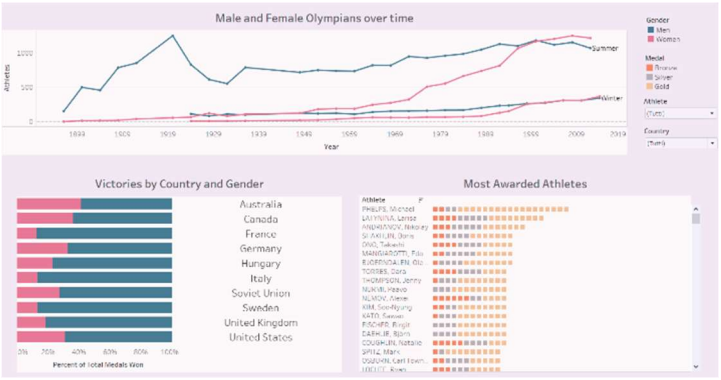
Red – Blind / Protanopia



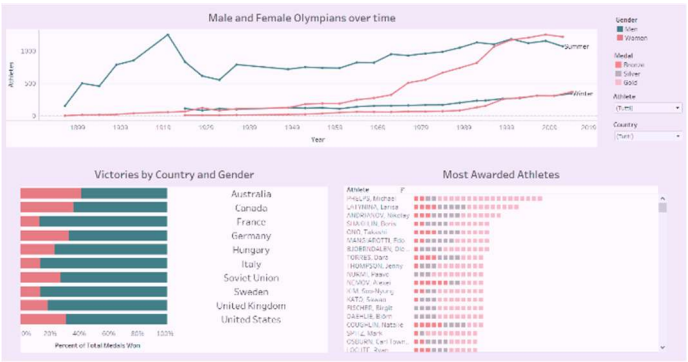
Green – Weak / Deuteranomaly



Green-Blind/ Deuteranopia



Blue-Weak/ Tritanomaly



Blue – Blind / Tritanopia

Also in this case we do not notice any particular problems in the vision of the dashboards except for those suffering from Red - Blind / Protanopia, who may have difficulty in distinguishing the pink and blue lines that differentiate male and female athletes.

The third dashboard has practically no colours that differentiate between countries or whose intensity depends on the medals won. So, since the colours are not essential in the dashboard display, I felt it was superfluous to include the test.

Conclusion

The goal of my report was to make both an exploratory and an explanatory analysis of the Olympics, to show its development over the centuries, considering major changes such as inclusion of women

and disciplines.
I tried to create an easy model to obtain any information needed. It was a stimulating project as while I was working on it, I discovered many curiosities about Olympics which made it more interesting.

Russo Federica

Matricola 1000011371