Data Visualization

Final Project

Hugo Morvan

Track and Field Olympic Results analysis.

Description of the Data Set:

The dataset for this project was found and imported from Kaggle.com. Its author, Jay Ravaliya, describes it as “Olympic Track & Field Results, Results from all Olympic Track & Field Events, 1896 – 2016”. This dataset was directly scraped from <https://olympic.org/athletics> , the official websites of the Olympic committee, and it is formatted as a CSV file. The file is composed of 9 different columns:

* Gender – the gender of the athlete (M for man or W for women),
* Event - indicating the name of the event,
* Location - indicating the name of the city where the game took place,
* Year – the year of the edition,
* Medal – the color of the medal (G for gold, S for silver, B for bronze),
* Name – the name of the athlete
* Nationality – the nationality of athlete (3 letter country code),
* Result – the performance (time, distance, or height) registered for the event. The time can be display using different format depending on the event.
* Wind – the wind registered during the performance. Very few wind measurements are present in the dataset, partially due to the recent implementation in history of this measurement.

figure 1: a preview of the dataset


*Figure 1: a preview of the dataset*

Goals

In a first part, the goal is to analyze this dataset using Python and look at trends in the performance of different events and see if it is possible to predict the results from the 2021 Tokyo Olympic games solely based on the trends from the previous editions.

In a second part, an analysis will be made using tableau to make more advanced statistics and visualization about this dataset, including geographical analysis and correlations between the location of the Olympic games and the success of the hosting country.

Part 1: Python analysis

The first step to do for analyzing this data using Python was to understand what data type was present in this dataset. Using commands such as .info() and .describe() helps to do so.

Text

Description automatically generated with low confidenceTable

Description automatically generated

*Figure 2 : info() and describe() commands*

Since only the year and the wind are stored as numbers (float64), the .describe() function does not give a lot of information. However, it still provides an insight for the few (10) measurements of wind recorded. With a minimum of -0.9 m/s and a maximum of 0.6 m/s of wind, all the measurements are within the wind regulations set by the World Athletics federation. These regulations stipulate that if a wind recording is outside of a +/- 2.0 m/s range, the performance cannot be validated and therefore cannot count as a record of any kind. Without diving to much into the technicalities of track and field, what these result means for the dataset is that no performance present in the dataset is invalidated by an excessive amount of wind, therefore any performance is a valid performance.

Before further in-depth analysis, it is also important to look at how the dataset is distributed. Using the command showed on Figure 3, it is possible to see that some events have more datapoints than other. For example, the Marathon and the 100m sprint are the two most documented events. This is partially due to the fact that those two disciplines have been present at the Olympic games for the longest time and are also among the most popular disciplines.

Table

Description automatically generated

*Figure 3 shows the number of data points for the documented events*

Based on that information I chose to make a more in-depth analysis of the 100m event. I therefore start by isolating the data for the 100m event:

Table

Description automatically generated

*Figure 4 shows the commands used to isolate the 100m event*

And then I used this isolated data to graph the evolution of the 100m podium times over the years:

Chart, scatter chart

Description automatically generated

*Figure 5*

In this graph, each medal (gold, silver, bronze) is represented with a corresponding color. This graph shows a clear progression in the 100m times since 1896. However, it is interesting to notice that some data are missing, in the year 1916, 1940, 1944 due to the two World Wars, but also in the year 1988 but because of imperfections in the dataset .

Same graphs can be made for other events:

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

*Figure 6, 7, 8*

*Side note*: As a track and field athlete myself, I was interested to see where I would stand on this graph. I therefore added an extra point (in green) to see. The result shows that running this time at any of the Olympic games before World War II would have earned me a gold medal at the Olympic games. In comparison, when I ran this time last year, it only allowed me to qualify to the French Olympic trials.

Predictions

Using trendlines and regression tools, it is possible to make predictions on the results from the next editions of the Olympic games. This is a question that is particularly interesting from an athletic point of view because it can help answer the question “How fast can humans run?”.

Different models were used to make predictions for the result of the 2021 Olympic 100m race and the results were compared with the actual value. For reference, the 2021 100m race was won by the Italian sprinter Lamont Marcell Jacobs with a time of 9.80 seconds.

As shown in the example below, the calculations for the trend lines and the timing prediction were done using the NumPy library. The .polyfit() functions third parameter allow to choose the degree of the polynomial desired, in this example, the third parameter is the number 2, the trend line is therefore a 2nd order polynomial.

Chart

Description automatically generated

*Figure 10*

The results are compiled in the table below:

| Polynomial order | Predicted time | Time difference | Percent error |
| --- | --- | --- | --- |
| 1 | 9,431 | 0,369 | 3,77 |
| 2 | 9,896 | -0,096 | 0,98 |
| 3 | 9,395 | 0,405 | 4,13 |
| 4 | 9,927 | -0,127 | 1,30 |
| Logarithmic | 9,44 | 0,36 | 3,67 |

Part 2: Tableau analysis

Tableau is a great tool to use for data analysis because it allows to do some great visualization. Some analysis that are difficult to make using Python are geographical analysis. However, it is made simple by Tableau. The Olympic game is the most renowned and oldest sport event in the world. Hosting and edition of the Games is a privilege and a way for a country to demonstrate the extent of their culture as all of the cameras of the world will be pointing at the host-city for the span of a month every four years.

Looking at statistics about the results of the Olympic games can be an indicator of the geopolitics of the present and the past. It was already shown in the first part that the missing data in this dataset was indicating the periods of the two World Wars, but a lot more information can be retrieved form this data.

The first aspect that this data is showing is the democratization of sport in our society. From the first edition in 1896 where 37 countries received a medal to the 2016 where 141 countries received at least one medal, the is a clear trend in the augmentation of countries represented. This trend is shown in Figure 11 :Chart, line chart

Description automatically generated

*Figure 11*

*Side note:* The 2016 Rio de Janeiro Olympics were also the first edition held in a country that use to be classified as “third world country”. It is an indicator that the country is emerging as a global power and that it is now ready to show its greatness to the world.

The two following maps show the repartition of those medal between the year 1896 and 2016:

Map

Description automatically generated

*Figure 12 and 13*

An interactive version of this map is available on the Tableau file attached, or on this link: <https://public.tableau.com/app/profile/hugo.morvan3334/viz/FinalProject_16383163375640/Story1#1> . On top of showing that the distribution of the medal went from being only between the western countries to being worldwide, the second map also shows that track and Field is a sport that is blind to the economic status of who practices it. It shows that small countries such as Jamaica or poor countries such as Ethiopia or Kenya can compete and be better than more populated or richer countries such as the European countries or China and India.

The next map shows the distribution of the total medal that were distributed since the first edition: Map

Description automatically generated

*Figure 14*

This map shows that although the medals are widely distributed in the world, the United States are extremely dominant in Track and Field. In fact, it seems that the distribution of the medal follows a Pareto Distribution, as shown in the following graph:

Chart

Description automatically generated

*Figure 15*

The top 20% of the countries hold more than 80% of the total medals.

The parity of the medal distribution varies depending on the country. For the United States, is about 2 third men, 1 third women. The rest of the countries in the world can be seen and the following bar chart:Chart, pie chart

Description automatically generated

Chart, bar chart

Description automatically generated

*Figure 16 and 17*

It is importance to keep in mind that the parity in the event in Track and Field has not always been respected. While women have been allowed to participate in Track and Field since 1900, some events were only available for men, and the women were only able lately to participate in those events, such as the Marathon race, whose first women edition was in 1984.

Finally, it is interesting to notice the correlation between the location of the Olympic games in a country and the performances of this country during this edition. The “home-field” advantage is real and can be noticed for many countries that hosted the Olympic games, as shown in the following table: Chart

Description automatically generated

The United States, the USSR, Germany, Australia, Sweden, Greece, Spain all had their record amount of medal earned during the Olympic games that they hosted. This can be explained by many factors including the absence of jetlag, the knowledge of the stadium, having the crowd cheering for the local athlete, etc. It is a trend that was also observed during the Tokyo 2021 edition, as Japan earned a record number of medals.

Conclusion

Much information can be retrieved from this data and its analysis allows to make predictions about future Olympic results, analysis on the geography of Track and Field, statistics about parity and an interesting correlation between hosting the Olympic games and observing an increased number of medals won. However, this data did not possessed information on the temperature, the altitude, the weather during the different event, which could greatly influence the performances. Future analysis could therefore focus on crossing this data with data including those measurements and see if it is possible to make connections and find correlations.