NOVA SCHOOL OF BUSINESS & ECONOMICS

120 years of Olympic History

Data Visualization for Business Analytics

Fall 2023

Group 23

Sara Favita | 60141

Jasmin Hamrouni | 61773

Guilherme Raimundo | 58053

NOVA SCHOOL OF BUSINESS & ECONOMICS

Index

NOVA SCHOOL OF
BUSINESS & ECONOMICS

Abstract

The following paper presents a comprehensive examination of the 120-year trajectory of the Olympic Games, employing Tableau visualizations as a principal analytical tool. Leveraging an array of datasets, including Athlete_Events, GDP, Country_Definitions, and Population, our study seeks to offer a perspective on the historical evolution of the Olympics, transcending mere chronological documentation.

The primary aim of our work involves exploring the correlation between athletes' ages and the medals they have attained. Our goal is to disclose patterns and trends, thereby elucidating the nuanced interplay between age and athletic achievement.

A pivotal facet of our analytical framework involves the creation of Tableau visualizations that articulate the longitudinal narrative of medals secured by each participating country over the course of the Olympics.

In summary, this proposal amalgamates methodological rigor with sophisticated visualization techniques, aiming to provide a scholarly contribution to the understanding of the complex dynamics underpinning the Olympic Games.

NOVA SCHOOL OF BUSINESS & ECONOMICS

Introduction

The creation of the dashboard will be based on historical data of the modern Olympic Games, ranging from Athens 1896 to Rio 2016. The dashboard included rows corresponding to individual athletes competing in individual events. This data is firstly separated in athlete_events, country_definition. Additionally, we are utilizing a gdp and population dataset.

The variables included in the athlete_events dataset contains 15 columns, each of them referencing specific information. A few important variables, which will be frequently referenced are the Medals, containing the type of medals won, being “Gold”, “Silver”, “Bronze” or “NA”. Moreover, the variable sport is also in the forefront of the analysis, describing the type of sport executed as well as the variable season provide information regarding Summer or Winter sports. It is important to note that the dataset also includes the variable NOC, displaying the National Olympic Committee 3-letter code.

The dataset country-definition contains information about the NOC, the Region, this being the country name used in geospatial mapping as well the Notes. This column shows the “Real country name” if the “Region” column doesn’t make an exact match.

The gdp dataset contains four columns, these being the name of the country, the country code in 3 letters, the year of the data as well as the gpd value.

The dataset populations contain 4 columns. The column GEO describes the country code in 3 letters, the country name, the year of the associated data and most importantly the population of that specific country.

Dataset and data cleaning

We start the cleaning process by reading the dataset. After that we print the missing data, which will be filled in with “None” instead. This step is essential for later visualization purposes, to generate informative visualizations. After that we are merging two datasets – the athlete dataset as well as the country dataset. But this isn’t enough to detect whether the NOCs codes still don’t have matching countries, which is why we drop duplicates and replace the missing Team values with their respective NOC codes. Additionally, we are dropping the column “notes” since we don’t need it for the analysis or the visualization.

After that, we move on to the gdp file. The purpose of this dataset is to bring more background information about the country’s economic situation. We start by reading the data, followed by changing the name of the column value to GDP to make the merging possible. In addition, we are converting the year column to numerical values. Next we are checking if the NOCs in the Olympic data match with those in the Country Code. After doing this we find out that 111 countries don’t have a NOC match. To fix this problem we added a country code for each Team in the Olympic dataset first. This is followed by merging the country code.

The population set is the last dataset which will be added. This will also be read at the beginning. After that we convert the Codes in the column GEO to uppercase letters. We then check if the NOCs in the Olympics data match the Country Code. Since a lot of the data doesn’t match we want to detect if the country data is a better way to merge. Next, we drop unnecessary columns.

Then we need to prepare our data for the visualization. We add a column which captures the information about winning a Medal or not. Additionally, we fill the rows with a zero (not winning a Medal) with NAN. After that we are creating masks, categorizing Events and creating “Event Categories”. Additionally, we are creating a new column called Medals teams won, while also replacing errors with zero as well as adding columns about the type of medals won, these being Gold, Silver and Bronze. We are then renaming the column Medal Won to Medal Sum. Now, its time to calculate medal tally agnostic of the team size - one gold means one gold for an event. To do this we divide the number of

medals by the count of winning team members. Lastly, we make final cleaning steps on Tableau Prep before we start with the visualisations.

Exploratory data analysis

The exploratory data analysis showed that we will mostly focus on the US as a country, as there is a lot of data available. Moreover, we found that the number of medals won by the US is relatively high, which enables us to dig deeper into the reasoning behind the finding and relationships between variables. Selecting data in the dataset from 1960 for our analysis was strategic, considering the wealth of information available from those years. This choice allowed us to embark on a comprehensive exploration of historical Olympic data.

In our pursuit of meaningful insights, we employed descriptive statistics worksheets in Tableau to visualize and interpret the data effectively. These visualizations became highly important in finding which three central questions we will be guiding our analysis: 'Have the top-3 medal-winning countries maintained their dominance, or has there been a change in the pattern of medal distribution over the years?'; 'Does age influence medal achievements in US Athletics and Swimming?'; and 'Is there a relationship between Gender and won Medals in the Swimming Categories of 100 and 400 meters freestyle?'

The choice to focus primarily on the US stems not only from the abundance of available data but also from the patterns that emerged when examining the country's historical medal performances. Our EDA journey has not only affirmed the relatively high number of medals won by the US but has also unearthed deeper insights into the dynamics shaping these victories. This abundance of data not only enhances the granularity of our analysis but also allows for a nuanced exploration of relationships between variables.

As we navigate through the layers of Olympic history, our analysis of the top-3 medal-winning countries (Germany, Russia, and USA) revealed both enduring dominance and shifting patterns, emphasizing the evolving nature of Olympic competition. Delving into the influence of age on medal achievements in US Athletics and Swimming provided valuable insights into the interplay between age groups and athletic success. Additionally, our examination of the relationship between gender and medals in specific swimming categories uncovered nuanced connections that add a layer of complexity to the narrative.

In conclusion, our focused exploration of US Olympic data, coupled with the strategic choice of the dataset from 1960, has given us three questions that have interest to the audience and can provide valuable and specific insights.

NOVA SCHOOL OF BUSINESS & ECONOMICS

Analytical Questions

Have the top-3 medal-winning countries maintained their dominance, or has there been a change in the pattern of medal distribution over the years?

Looking at the data with summer and winter seasons of the ranking top 3 countries with more medals we can see that in the last years the USA has being staying at the top as the country with the highest amount of medals won. This pattern has been seen since 2000 until 2014, except in 2010 in which they took the second place.

If we are going to filter data only including the summer season, we detect that we don't have data from the USA in 1980. This is due to the fact that the games were held in Moscow in 1980 and as a form of protest against the Soviet Union's invasion of Afghanistan in 1979 the USA decided not to participate. Then 1984 the Russia didn't participate in the Olympics games because the games were held in Los Angeles, which was a response to the boycott from the USA in the last games. If we look at the last years, we can see that the pattern of the classification of the USA staying in the first position as the country with the most medals won is the same as the conclusions found when looking at both seasons together. If we look at the evolution of the medals won, we can clearly see that in the last year the USA has been the one with the most medals won, but before that we cannot detect a pattern.

In the winter Olympic games, we can see that the pattern in the ranking evolution is different from the pattern from previous observations, since Germany is the country that stays more frequently on time and USA only stay 3 consecutive years since 2002, then in 2014 it moves to the second position. Now looking at the evolution of medals we can see that the USA's success has been growing over the years reaching the best results in 2014 with 124 medals staying behind Russia that had 127 medals.

These top 3 countries have been maintaining their position in throughout the years on being the countries with more medals, being the trend to be in the first position the USA, then Russia and on the third position Germany. The distribution of medals won doesn't have a clear pattern since if we look at the two seasons together, but if we look separately, we can see that in the winter there is a pattern that the medals have been increasing over the years. But in the summer, we can see that in the beginner was increasing but then we cannot define a clear pattern for these 3 countries.

NOVA SCHOOL OF
BUSINESS & ECONOMICS

Does age influence medal achievements in US Athletics and Swimming?

The choice beyond this question lies in the 120-year history of the Olympics from 1960 in the context of US Athletics and Swimming that might reveal intriguing patterns regarding age and medal achievements.

Firstly, we have selected United States as a country to analyze because it is one of the most successful countries in terms of number of medals won in all sports across the 120 years. Swimming and Athletics are the sports where they achieved the highest number of medals.

Over the years, a trend emerges suggesting that age indeed plays a significant role in shaping success in these sports. In the dashboard, we included line graphs and scatterplots that describe the evolution of the number of medals won per year and per age in Swimming and Athletics. Also, there is a table with the main indicators of performance from the athletes and a bar chart regarding the number of medals won and a bubble chart to represent the importance of each age considering the number of medals achieved in each sport. The number of medals won in these two sports increased to even higher numbers over the years since 1960.

For example, in the realm of athletics, observations have substantiated the prevalence of younger athletes, specifically those aged between 22 and 25 years, in consistently excelling in events of this nature. This phenomenon is attributed to the propensity of individuals within this age range to attain their zenith of performance. It is imperative to underscore, however, that a discernible escalation in the acquisition of medals becomes apparent for athletes below the age of 22. Conversely, a contrary trend is observed in athletes aged 25 and above, where the frequency of medal attainment tends to decrease.

The exploration of the intricate relationship between age and medal achievements has yielded a nuanced understanding that influences success in US Athletics and Swimming throughout the extensive history of the Olympics. These insights contribute to a more holistic comprehension of the dynamics at play within these sports over time.

NOVA SCHOOL OF
BUSINESS & ECONOMICS

What is the relationship between Gender and won Medals in the Swimming Categories of 100- and 400-meters freestyle?

Since we saw that the US as a country is very dominant int the Olympics, especially in the Sport of swimming, we wanted to explore the relationship between gender and two of the categories winning the US the most amount of medals in the sport, these being 100 and 400 meters freestyle swimming. We focus on data starting in 1960

Even though the amount of Participants in both the categories is the same for both genders, we found that the amount of won medals is different in both categories. As we can see in the Packed Bubble chart we can see that female swimmers won 26 medals while male swimmers won 29 medals in the category of 100 meters freestyle. Looking at the bump chart, we can see that the pattern of male swimmers is very consistent. They won a medal in almost every year, except 1980 due to a boycott as In 2004 where they won 2 medals in each year. In comparison to that, we find that the pattern of females winning isn't that consistent. It fluctuates between the years leading to multiple spikes.

Comparing the amount of won medals in 400 meters freestyle swimming leads to the knowledge that there is an opposite effect to be seen. Females are here in the lead in compared to males, since they won 28 medals while males won 25 medals. The pattern which we saw in the other category isn't found here. Here the pattern of males is similar to females in the category of 100 meters swimming. When comparing females from both categories we found the opposite pattern of winning. Mostly, when females from 100 meters swimming won 2 medals, females from 400 meters swimming won 1 medal. The opposite is also found.

In conclusion it can be said that that it highly depends on the category that we are looking at. Female swimmers dominate the category of 400 meters swimming while male swimmers dominate 100 meter swimming. Consistency of performance is also a factor which should be considered when exploring this relationship as it also may give some more insights about the relationship. It is important to note that this gap is evident even though the amount participants is the same. Reasons for that could be explored in other papers, these being about the gender gap of funding in sports or a lack of support in many male dominated areas of sports.

NOVA SCHOOL OF BUSINESS & ECONOMICS

Data Visualization Proposals

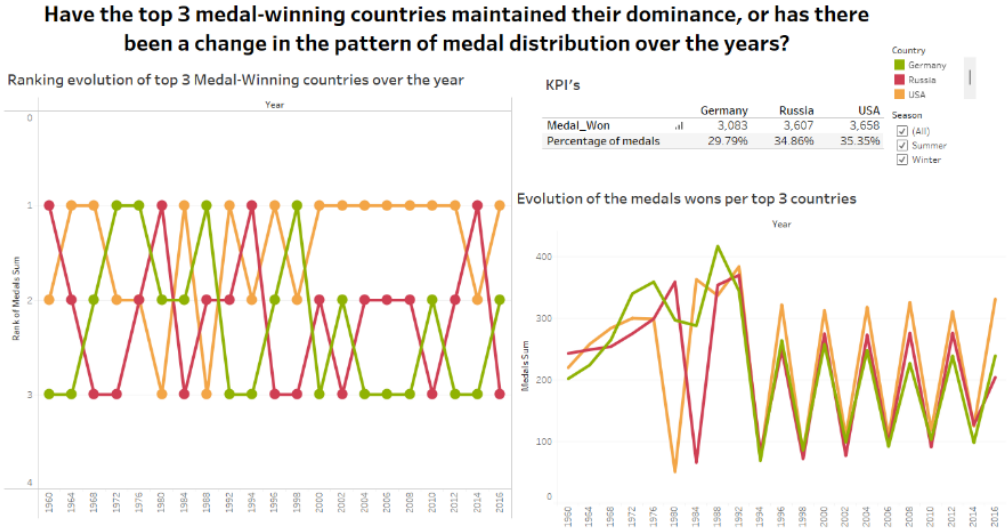

Figure 1: Dashboard 1 – Have the top 3 medal-winning countries maintained their dominance or has there been a change in the pattern of medal distribution over the years

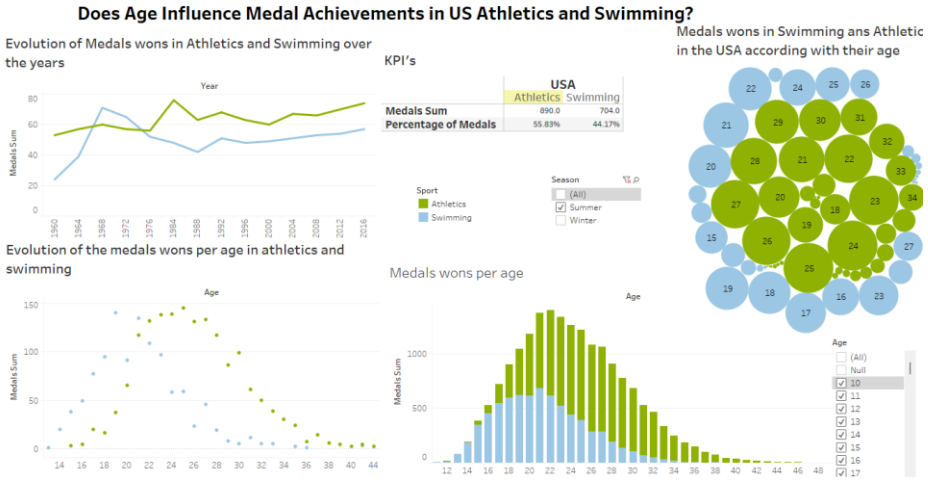

Figure 2: Dashboard 2 –Does Age influence Medal Achievements in US Athletics and Swimming

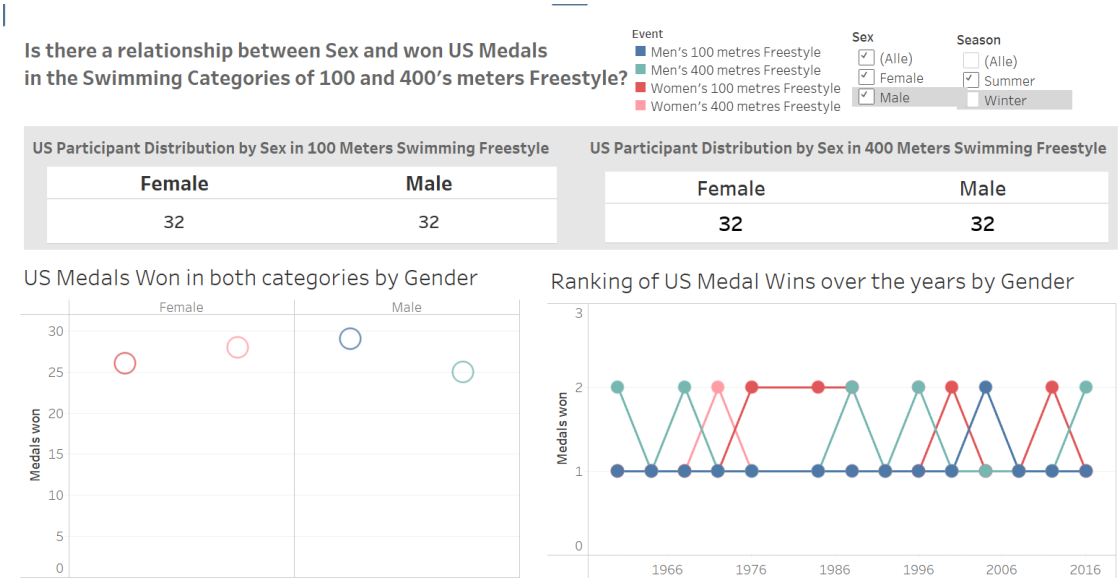

Figure 3: Dashboard 3 Is there a relationship between Gender and won Medals in the Swimming Categories of 100 and 400's meters freestyle?

Conclusions

Firstly, the exploratory data analysis (EDA) phase, executed primarily through Tableau for statistical analysis and visualization, deepened our understanding of the dataset. The Tableau dashboards provided a dynamic platform for descriptive statistics, allowing us to explore the distribution of key variables, unveil central tendencies, and identify potential outliers.

The dashboards provided insights into the evolution of medal distribution among the top-performing countries over the years. A thorough analysis of the data revealed that traditional powerhouses such as the United States, Germany, and Russia consistently secure top positions in the medal tally.

The data cleaning process in Python made before the construction of the dashboards, played a crucial part in ensuring the accuracy and reliability of the analysis. Cleaning the raw data allowed for a more robust exploration of trends and relationships, enhancing the overall quality of the findings.

The exploration of the relationship between age and medal achievements in US Athletics and Swimming yielded noteworthy findings. It was evident that certain age brackets exhibited higher performance levels, indicating that age does play a role in medal success. However, the nuances of this relationship require further investigation. Understanding the peak age for athletic achievement can be pivotal for talent identification, training program development, and athlete management strategies. Coaches, sports scientists, and administrators can leverage these insights to optimize training regimens and enhance the overall performance of athletes in both Athletics and Swimming disciplines.

The analysis of the relationship between gender and medals won in the Swimming Categories of 100- and 400-meters freestyle showcased interesting dynamics within the sport. Through the interactive Tableau dashboards, it became apparent that gender does influence medal outcomes, with variations in performance observed between male and female athletes. These insights can be instrumental in fostering gender equity initiatives within the swimming community, guiding sports federations, and informing coaching methodologies. Furthermore, understanding the nuances of gender-based performance

disparities can contribute to the broader conversation on inclusivity and representation in sports.