10/14/2024

120 Years Olympics Games

Data Analyst Project Report OCT 24

Prepared By

Youssef Medhat
Shimaa Mostafa
Moataz Adel
Abdelaziz Abdelfatah

Under Supervision of
Eng Amira Shahin

DEPI

Table Of Content

Introduction

Objective of this project:

In this project we are looking at 120 years of Olympics and Olympic achievements we aimed for a Dataset that has empathies on many athletes and teams that have participated through the years and how to gain a better understanding on the dynamic change of gender the history of sports and display the various medals obtained by different countries.

Track Relevance:

We look at this dataset and see it's high potential to allow us to put into practice the multiple skills we have gained during our track.

we are using different data analytical tools and sheets in order to help us better explore and understand data of the Olympic games and the different variables that were taken in consideration in creating such data.

By implement Data Cleaning and Preprocessing techniques in order to elevate the data from any problems and prevent miscalculations to guarantee the integrity of the insights we obtained by analyzing such dataset.

Using dynamic tools to display and visualize our data foe all to see and share our insights by telling a story.

Expected Outcomes:

We hoped to gain many insights of the Olympic games and their diverse record and some of the outcomes we hoped to achieve are:

1) To analyze the historical progression and milestones of the Olympic Games over 120 years.
2) The effect of increasing the number of participant females in the Olympics.
3) To explore the advancements and achievements in sports gained through Olympic history by Male and Female participants.
4) Explore and examine the different countries that both held and participated in the Olympic games.
5) Look at the changes brought to the Olympics during historical times.

Project Proposal

Project Title:
"From Athens to Tokyo": The Evolution of the Olympic Games Over 120-Years

Team Members & Their Student code:

Members	Student Code
Abdelaziz Abdelfatah	21007258
Youssef Medhat	21041503
Shimaa Mostafa	21015699
Moataz Adel	21003520

Problem Statement:

The Olympics is one of the Largest held events all around the globe, and of this year (2024) one of the most impactful Olympics were held, so our study aims to try and look at 120 years of Olympics and analyze the impact of the years on the Olympics, how it affected the participants? what has changed over the course of the years and what insights and knowledge can we leverage from 120 years of sportsmanship among hundreds of sports and investigate the historical events and patters across the Olympics aiming to answer many of the questions many asked themselves.

Problem Solution:

1-Technologies Used: Power BI and Excel

2- System Architecture:

       1. Data Collection:

▪ Gather historical data on the Olympic Games from reliable sources such as official Olympic records, sports databases, and academic journals.

▪ Conduct interviews with athletes, coaches, and officials to gain qualitative insights.

       2. Data Analysis:

▪ Use Python for data cleaning, preprocessing, and statistical analysis.

▪ Employ SQL queries to extract relevant information and perform complex data manipulations.

▪ Utilize Power BI for creating interactive visualizations and dashboards to present findings.

Scope and Limitations

The constraints meet for our project is to limit the insights to the existing data without generating the data using any artificial tools or other related datasets

Project Plans

Milestones & Task breakdown:

Milestone	Task breakdown	Durations
1. Identifying project idea		One day
2. Collecting datasets	ALL MEMEBERS	One day
3. Preparing project proposal		One day
4. Exploring the Dataset		Two days
5. Data cleaning and preprocessing		Five days
6. Modeling Data by power query		Two days
7. Creating Dashboard and Visualization of Data		Ten days
8. Extracting insights from Data		Two days
9. Final reporting		One day

Resources Needed:

Tools & resources	Type of tool
120 years of Olympic games	Data set
Excel spreadsheets	Microsoft app
Python	Programming Language used for Analytical tools
Power BI	Microsoft Analysis and dashboard app
Microsoft teams	Software app for meetings
Adobe color ballets	web application for coloring and visualization
Laptops	Hardware tools
Google Drive	Web app for sharing files and folders
Microsoft Project	Microsoft app for managing projects

Risk Management:

In such team projects some risks are bound to happen and in attempt to fight against any errors and secure ourselves from any confusion or errors:

1) We have the Dataset excel sheets that edit shared among us on a drive to allow easy access and fight against loss of data.
2) Each version of Dashboard was updated and named accordingly to help us keep up with updates undertook by other team members and shared them among us to stay updated

Data preparation & EDA:

Data collection:

The team went through multiple Dataset providing websites like Kaggle, MavenAnalytics, Data.gov and much more when selecting this topic, we first went through different challenges data and topics that were popular among these sites and before setting on a topic we had many other topics that we continued to search for similar data sets about these topics in the aforementioned sites before setting on 120 years of Olympics from MavenAnalytics.

Data cleaning and Preprocessing:

We used multiple tools in this section to help us prepare the data, we went through multiple steps of data cleaning, such Tools used in cleaning:

Excel spreadsheets:

It is a program that allows users to organize, format, and calculate data in a spreadsheet. We used this program to explore our data further and gain an understanding of what it included and the role of each column, the data type of columns and to create and manipulate data.

Structure of our Data it consists of Three spreadsheets

1. Athelete_events: this is the main spread sheet that consisted of all a lot of our data

Fields	Number of nulls	Data Type
ID	none	General
Name	none	General
Sex	none	General
Age	+9,000	Number
Height	+60,000	Number
Weight	+62,000	Number
Team	none	General
NOC	none	General
Games	none	General
Year	none	number

Season	none	General
City	none	General
Sport	none	General
Event	none	General
Medals	+53,000	General

No. of Fields: 15   No. of records: 271,11

2. NOC_region: this spread sheet contains all the regions of athletes with their National Olympic Committee (NOC)

Fields	**Number of nulls**	**Data Type**
NOC	none	General
region	none	General
notes	none	General

No. of Fields: 3   No. of records: 230

3. Flags_iso: this spread sheet was added as process of Preprocessing that adds Images for Flags to countries for better visualization.

Fields	**Number of nulls**	**Data Type**
Country	none	General
Alpha-2	none	General
Alpha-3	none	General
URL	none	General

Such things we used are Excel formulas functions:

Table function: we transformed our Data set to a large group of tables to be able to use Excel functions more efficiently and navigate through the table more accurately.

Power Query & Data Model:

Power Query is a data connection technology that enables users to discover, connect, and combine data from various sources. It offers a user-friendly interface to perform data transformation and preparation tasks without needing extensive programming knowledge.
Data Model, on the other hand, is a feature within Power BI that allows users to define relationships between different tables, enabling complex data analysis and visualization. The Data Model stores data in a structured format, facilitating efficient querying and reporting.

What Do We Use Them For?
Power Query and the Data Model are used primarily for:

- **Data Transformation:** Cleaning and reshaping data for analysis.
- **Data Integration:** Combining data from multiple sources to create a cohesive dataset.
- **Visualization Preparation:** Streamlining data for easier and more effective reporting in Power BI.
- **Automating Data Refresh:** Setting up connections that automatically update when data sources change.

Utilizing Power Query to Import Data from Excel Sheets to Power BI

Using Power Query to import data from Excel sheets into Power BI has significantly streamlined my workflow. The steps I followed included:

- **Connecting to Excel Files:** Power Query allows me to easily connect to multiple Excel files stored locally or on cloud services.
- **Data Transformation:** I could transform the data as needed—removing unnecessary columns, filtering rows, and changing data types—directly within the Power Query editor.
- **Appending Data:** When dealing with multiple sheets or files, Power Query's append feature made it simple to consolidate them into a single table.
- **Loading Data:** Once the data was clean and ready, I could load it directly into Power BI for further analysis.

Applying Data Model Feature to Find and Established Relations Between Existing Data Tables.

The Data Model feature allowed me to create a robust relational structure between my existing data tables:

- **Identifying Relationships:** Power BI automatically detected potential relationships based on matching column names and data types. This helped me quickly understand how my tables interacted.
- **Creating Custom Relationships:** I had the option to manually define relationships, specifying cardinality and cross-filter direction, which is essential for accurate data analysis.
- **Enabling Complex Analysis:** With established relationships, I could create calculated measures and run cross-table analyses, facilitating deeper insights and improved decision-making.
- **User-Friendly Interface:** Both Power Query and the Data Model have intuitive interfaces that lower the learning curve for new users, making it accessible to a wider audience.
- **Performance Optimization:** The Data Model is designed for high performance, handling large datasets efficiently. By using in-memory storage and DAX (Data Analysis Expressions) for calculations, it enables fast query responses.
- **Reusability:** Once queries and models are set up, they can be reused across different reports and dashboards, promoting consistency and saving time in future projects.
- **Collaboration Features:** Integration with Power BI Service allows for easy sharing and collaboration on reports, enabling teams to work together more effectively.

Python:

Python is a high-level, general-purpose programming language renowned for its readability and ease of use. Its clean syntax and extensive standard library, Python has improved the process of Data cleaning.

Why Python for Data Cleaning and EDA?

1. **Readability:** Python's code is often described as "pseudocode-like," making it easy to understand and write. This is especially important when working with complex datasets.
2. **Large Ecosystem:** Python boasts a vast ecosystem of libraries and tools specifically designed for data science. These libraries, such as NumPy, Pandas, Matplotlib, and Scikit-learn, provide powerful functionalities for data manipulation, analysis, visualization, and machine learning.
3. **Community Support:** A large and active community of data scientists and developers contributes to Python's growth and development. This means you can find abundant resources, tutorials, and support online.

NumPy and Pandas: Essential Libraries

- **NumPy:**
  - Provides efficient multi-dimensional arrays and matrices.
  - Offers a wide range of mathematical functions for numerical computations.
  - Is the foundation for many other data science libraries.

- **Pandas:**
  - Builds upon NumPy to provide high-performance data structures and data analysis tools.
  - Introduces DataFrames, which are tabular data structures similar to spreadsheets.
  - Offers functions for data cleaning, manipulation, and analysis.

Strengths of NumPy and Pandas in Data Exploration and Cleaning

1. **Data Manipulation:**
   - Both libraries excel at performing various data manipulations, such as filtering, sorting, grouping, and merging.
   - Pandas' Data Frames provide a convenient way to work with structured data.

2. **Data Cleaning:**
   - NumPy and Pandas offer functions for handling missing values, outliers, and inconsistencies.
   - Tasks like removing duplicates, standardizing data formats, and imputing missing values can be efficiently performed using these libraries.

3. **Data Exploration:**
   - You can easily calculate mean, median, standard deviation, and other metrics, as well as create plots and charts to explore your data.

In summary, Python, combined with NumPy and Pandas, offers a powerful and efficient toolkit for data cleaning, exploration, and analysis. Its readability, versatility, and extensive ecosystem make it a preferred choice for many data scientists.

Data Analysis expression (DAX):

Data Analysis Expressions (DAX) is a formula language that is designed to perform advanced calculations on data models, allowing users to create custom calculations and aggregations based on their data.

Strengths of DAX:

1. **Powerful for Aggregation and Filtering**: DAX enables complex data manipulation, allowing users to aggregate data, filter it dynamically, and create sophisticated calculated columns and measures.

2. **Time Intelligence**: DAX has built-in time intelligence functions, which allow users to compare data across different time periods (like month-over-month or year-over-year comparisons) without writing complicated code.

3. **Efficient for Large Datasets**: DAX works efficiently with large datasets because it's optimized for high-performance queries. It's designed to work with columnar databases like the ones used in Power BI.

4. **Integration with Other Data Sources**: DAX can interact with multiple data sources, transforming raw data from different systems into unified, analyzed reports.

Why Use DAX in Data Cleaning and Exploratory Data Analysis (EDA):

- **Data Transformation**: DAX allows for easy transformation of raw data into more meaningful, clean datasets by creating calculated columns, measures, and virtual tables that can fill gaps or correct inconsistencies.
- **Data Aggregation and Filtering**: During EDA, it's common to need dynamic calculations and custom metrics that DAX facilitates, such as generating specific summary statistics or exploring patterns in data subsets.
- **Creating Custom Metrics**: In EDA, DAX allows you to explore relationships and trends in your data by creating new calculated fields or measures based on your business needs.
- **Efficiency in Handling Missing or Inconsistent Data**: Through DAX formulas, you can create rules to handle missing data, identify outliers, or fix data inconsistencies, which is essential during the data cleaning phase.

DAX we used in our Dashboard:

We used to create multiple filters and measures to better visualize our Data and provide a more interactive and visually appealing board.

1. **CALCULATE** Filter: It allows you to modify the context in which data is evaluated by applying filters to your calculations.
    a. Creating Filter to count all gold, silver and bronze medal winners each.
    b. Total count of medals that were achieved.
    c. The count of medals won by Teams each.
    d. Filter the number of Female athletes in Summer/Winter Olympics.
    e. Filter the Number of Male athletes in Summer/Winter Olympics.
    f. Filter any Null Athlete with Medals
    g. Filter any Null Teams with medals

2. **DIVIDE**: it is specifically designed for division operations. It is a safer way to perform division compared to using the standard division operator
    a. Finding % of all Female athletes across years of Olympics
    b. Finding % of all Female Athletes across Summer/Winter games
    c. Finding % of all Male Athletes across Summer/Winter games
    d. Finding % of all Male Athletes across Summer/Winter games
    e. Finding % of all Athletes Medal winners across Summer or Winter games.

3. And many more functions such as
    a. COUNTA: Counts the number of non-blank values in a column, ideal for tallying data entries where blanks may exist.
        Count number for Athletes for filtration

    b. SWITCH: Evaluates an expression against multiple conditions and returns a specified result based on the first matching condition, like a series of nested IF statements.
        Get Data related to condition between Summer or winter

    c. AVERAGE: Returns the arithmetic mean of all non-blank values in a column, providing a quick way to calculate average metrics.
        Find average of Ages of Athletic players

    d. SUM: Adds up all the numerical values in a column, providing a total or aggregate value for that dataset.
        Sum of many columns and total of Data for some filters

This section is dedicated to be a quick Manual to anyone who looks at the Dashboard of insights or Controls.

Split of Dashboard:
The Dashboard is split into Three sections a Header, Central and Left-side section, each section serves different purpose and goal, there are also TWO more dashboards for Summer and Winter Olympics each with the same section split and a way to access them.

Head Section:

This section contains 3 major parts:
The Left-Side has the ICON and Name of the Dashboard you are currently on, by default when opening Dashboard for the first time this part will show 120 YEARS of OLYMPICS Image and text for the MAIN dashboard.

The Middle-Section has an ICON that allows to apply universal filters to the Dashboard this section is also universal across all three dashboards for Main, Summer and Winter dashboards, by clicking the this ICON a drop downlist will show which will allow filtering with the following:

- First Filter allows to filter by Gender you can select Male or Female and all graphs and cards will change accordingly
- Second Filter allows to filter by the type of Medal (GOLD, SILVER or BRONZE) and all graphs and card will change accordingly
- Third Filter allows to filter by the sports which participated in the Olympics.
- Fourth Filter allows to filter by the years starting from 1896 to 2016 by simply sliding the bar to select the start and end time period you wish to gain insights on

The Right-Side section allows for navigation across the three Dashboards, by simply click selecting "" The Olympics Highlights "" for MAIN dashboard, "Summer Games" for summer Olympics dashboard and "Winter Games" for winter Olympics dashboard.

Left-side section:

This section contains a list of cards which allow for quick overview and insights, each card changes accordingly to the selected filters and switches for different dashboard you are currently navigating

Only for the MAIN dashboard on top of these cards there is an addition flag that shows the percentage of MALE to FEMALE athletes
The Cards going through them descending each show the following data:

1. Total Number of Competitors

2. Total Number of Medals obtained
3. Sum of editions of Olympics
4. Number of sports involved
5. Number of participating countries
6. Total number of teams

On their own these cards provide good insights but they improve when ever you select or filter by any of the years, countries or players from the Central part allowing for a Dynamic and insightful experience.

The Central section:

The central section itself is split into many graphs each we will go through.

THE OLYMPICS HIGHLIGHTS

The Olympic Edition & Athletics Details Graph.

This bar graph shows the Olympics played across the years and percentage of participant Males and Females across each of the Olympics.

It also is showing the incline/decline of Male and Female participants for the Olympics across the years by selecting a bar on the chart all Cards and graphs will adjust dynamically to the years selected to show insights related to this year in the Olympics

Attached Tool TIP: by hovering this graph a tool tip will show that contains additional information related to the selected year(s) of the Olympics, this tooltip shows (YEAR- HOST CITY - %MALE AND FEMALE PARTICPANTS – SPORTS WITH THE HIGHEST NUMBER OF ATHLETES – TOP MEDAL EARNING COUNTRIES – NUMBER OF MEDAL EARNED BY EACH COUNTRY)

Medals in Season:

This s a graph chart that shows the percentage of participants that obtained a medal across the summer and winter Olympics.

Top Athletes Medals:

This is a Table that shows the Name of the Athletes, Region number of Participation to the Olympics across the years and Number of Gold medals won.

Attached Tool TIP: by hovering players a too tip appears that contains addition details on the athlete such as the number of gold, silver and bronze medals obtained by the athletes, their specialized sport and The Flag of their country

Top Teams Medals:

This graph shows the top Teams that participated in the Olympics as part of a team sport ordering them by number of medals won for each team across the Olympic years

Attached Tool TIP: this tool tips shows the country that this team plays for and the number of medals won and the total number of athletes that played within these teams and total numbers of medals obtained.

Top Countries Medals:

This table shows the top countries with the highest number of medals and total number of participants that played for this country.

SUMMER GAMES & Winter Games

These two Dashboard share all graphs in central section

Each Medal by Gender:
This bar graph shows the % of Males to Females athletes that got gold, silver or bronze medals.

Highest Medals by sport:
This graph shows the number of medals given for each category of sports across the summer Olympics

These two share the Top Teams Medals & Top Countries Medals tables and all their tooltip and data.