Olympic Medal Prediction and Geospatial Analysis

1. Problem Definition

The primary objective of this analysis is to conduct an in-depth exploration of the Summer Olympics dataset, focusing on various aspects such as gender participation, medal distributions, country performance, and specific sport-based achievements. Additionally, a predictive model is implemented to estimate the probability of winning a medal based on features such as the athlete's country, sport, and gender. The goal is to identify trends, patterns, and insights into Olympic achievements, and enhance the analysis through visualization and machine learning techniques.

Key Questions:

How has athlete participation changed over the years?

What are the top-performing countries and sports in terms of medals won?

What is the likelihood of a country winning a medal in a particular sport based on past data?

How does the performance of different regions (continents) compare over the years?

2. About the Dataset and Its Components

The dataset used for this analysis is the **Summer Olympics Medals (1896-2024)dataset**, which contains information on athletes, their participation, and the medals won over several Olympic games.

Key Columns in the Dataset:

player_id: Unique identifier for each athlete.

Name: Name of the athlete.

Sex/Gender: The gender of the athlete (Male/Female).

Team: The team or country the athlete represents.

NOC: National Olympic Committee country code.

Year: The year the athlete participated.

City: The city where the Olympics were held.

Sport: The sport in which the athlete competed.

Event: The specific event in which the athlete participated.

Medal: The type of medal won (Gold, Silver, Bronze, or No Medal).

3. About the Work Done

The analysis is divided into several steps:

Data Cleaning and Preparation:

- Duplicates were removed from the dataset to ensure data integrity.

- The 'Sex' column was renamed to 'gender' and mapped to more readable values ('Male' and 'Female') for better understanding.

- Univariate analysis was conducted to explore gender distribution, city participation, and sport popularity.

Exploratory Data Analysis (EDA):

1. Gender Distribution: A bar plot was used to explore the number of male and female athletes over the years.

2. Year-wise Athlete Participation: A line plot visualized how the number of athletes has changed across the Olympic Games, showing trends in participation.

3. Medal Distribution: A bar chart was created to showcase the distribution of medals (Gold, Silver, Bronze, and No medal).

4. Top 10 Sports by Participation: The most popular sports, in terms of athlete participation, were identified and visualized through a bar plot.

5. City-wise Medal Distribution: A stacked bar chart was used to show which cities hosted the most successful Olympic events based on the number of medals awarded.

Medal Winning Probability (Logistic Regression):

- A Logistic Regression model was implemented to predict the probability of winning a medal based on the country, sport, and gender.

- The categorical variables (NOC, Sport, and gender) were encoded, and the model was trained on these features to predict the likelihood of winning a medal.

- The model achieved reasonable accuracy, and predictions could be made, such as the probability of an athlete from the USA in Athletics (Female) winning a medal.

Region-Based Analysis:

- A region-based analysis was performed by mapping countries to their respective continents.

- A choropleth map was created to visualize the total number of medals won by different countries, highlighting the geographical distribution of Olympic success.

- This analysis provided a clear understanding of the dominance of various regions in the Olympics over the years.

4. Conclusions from the Results:

From the analysis, we can draw several meaningful conclusions:

1. Athlete Participation Over the Years :

- The participation of athletes has grown significantly over time, indicating the increasing scale and popularity of the Olympics.

- Spikes in athlete participation align with years where the Olympics were hosted in major cities, reflecting the global draw of these events.

2. Medal Distribution :

- Most athletes do not win a medal, which is expected given that only the top three performers in each event are awarded medals.

- Among the medals awarded, there is a nearly even distribution between Gold, Silver, and Bronze.

3. Top Performing Sports and Teams :

- Sports such as Athletics, Gymnastics, and Swimming dominate in terms of athlete participation and medal counts.

- Teams like the USA, China, and Great Britain consistently appear at the top of the medal tables.

4. Medal Winning Probability:

- The logistic regression model demonstrated that certain countries (such as the USA) have a higher probability of winning medals in certain sports (e.g., Athletics).

- Gender plays a role in the probability of winning a medal, and the model can be fine-tuned further with additional features.

5. Region-Based Performance:

- Countries from Europe and North America have traditionally dominated in medal counts, as shown on the choropleth map.

- Recent Olympic Games have seen growing success from countries in Asia and South America.

5. Final Summary

This analysis provided a comprehensive exploration of the Olympic Games dataset, uncovering key insights into athlete participation, medal distribution, and country performance over time. Through the use of logistic regression, we also modelled the probability of an athlete winning a medal based on their country, sport, and gender, offering predictive capabilities for future Olympic performances.

Additionally, region-based analysis helped visualize the geographical distribution of Olympic success, highlighting how different continents have performed over the years. The use of both static and interactive visualizations allowed for a rich understanding of the data, while the machine learning model added a predictive dimension to the analysis.

Overall, the study successfully captured the key dynamics of the Olympics, offering valuable insights for sports analysts, historians, and enthusiasts.