

Project Report

Olympics Analysis using Data Science

Subtitle: *Data-Driven Insights into Olympic History*

1. Introduction

The Olympic Games are the world's foremost sporting event, featuring athletes from across the globe. With more than a century of historical data, the Olympics provide an excellent opportunity for data-driven analysis. This project leverages data science techniques to preprocess, analyze, and visualize Olympics datasets. The goal is to uncover performance trends, analyze medal tallies, and provide interactive dashboards for exploration.

2. Objectives

The primary objectives of this project are:

- To build an interactive data analysis platform for Olympic Games data.
 - To analyze medal tallies by country and year.
 - To study the growth of nations, athletes, and events over time.
 - To evaluate country-specific performances using heatmaps and top-athlete rankings.
 - To gain actionable insights into global sports performance trends.
-

3. Datasets Used

1. Athlete Events Dataset (`athlete_events.csv`):

- Contains details about athletes, their demographics, Olympic participation, and medals won.

2. NOC Regions Dataset (`noc_regions.csv`):

- Maps National Olympic Committees (NOCs) to their respective countries/regions.

Both datasets were merged and preprocessed for analysis.

4. Methodology

4.1 Data Preprocessing

- Filtered only **Summer Olympics** data.
- Merged athlete dataset with NOC regions for country mapping.
- Removed duplicates and inconsistencies.
- Applied **one-hot encoding** for medals (Gold, Silver, Bronze).

4.2 Application Development

The project was implemented using **Python** and **Streamlit**.

- **Modules Used:**

- `preprocessor.py`: Data cleaning and preprocessing.
 - `helper.py`: Functions for medal tally, country analysis, and most successful athletes.
 - `app.py`: Streamlit front-end for interactive dashboards.
-

5. Key Functionalities

5.1 Medal Tally Module

- Displays medal counts based on user-selected year and country.
- Provides overall tallies and detailed breakdowns.

5.2 Overall Analysis

- Key statistics: Editions, Nations, Sports, Events, Athletes.
- Trends over time for nations, events, and athletes using line charts.
- Heatmap showing event growth across sports.
- Most successful athletes identified for each sport.

5.3 Country-wise Analysis

- Year-wise medal tally for a selected country.
 - Heatmap showing performance across different sports.
 - Top 10 athletes from a chosen country.
-

6. Tools and Technologies

- **Programming Language:** Python
 - **Libraries:** Pandas, NumPy, Seaborn, Matplotlib, Plotly
 - **Framework:** Streamlit
 - **Data Processing:** One-hot encoding, groupby aggregation
 - **Visualization:** Interactive charts (Plotly) and heatmaps (Seaborn)
-

7. Results and Insights

- Participation in the Olympics has grown significantly over the years.
- The number of sports and events expanded, reflecting increasing diversity.
- A few countries consistently dominated medal tallies.

- Athlete-level analysis identified the most decorated Olympians.
 - Heatmaps highlighted country-specific strengths in certain sports.
-

8. Conclusion

This project successfully developed an **interactive Olympics Analysis platform**.

It demonstrated effective **data preprocessing, aggregation, and visualization** techniques, enabling meaningful insights into Olympic history.

The system allows users to analyze global sports trends, evaluate country performances, and explore athlete achievements interactively.

9. Future Scope

- Incorporate **Winter Olympics** data for a holistic analysis.
 - Add **predictive analytics** to forecast medal tallies.
 - Include **real-time updates** for upcoming Olympics.
 - Expand athlete-level analysis with machine learning models.
-

10. References

- Kaggle Olympics Dataset
- Streamlit Documentation – <https://streamlit.io>
- Python Libraries: Pandas, NumPy, Matplotlib, Seaborn, Plotly