

Olympics Data Analysis — Final Report

1) Basic Details

- Name: Vaibhav Parmar
- Email: vaibhavparmar.180708@gmail.com
- Institution: Atmiya University
- Project Title: Olympics Data Analysis

2) Dataset

- Rows after cleaning: 15316 (removed duplicates: 117)
- Time range: 1976–2008

3) What I Did

- Loaded the CSV and standardized the column names.
- Removed exact duplicate rows and trimmed text fields.
- Summarized medals by year, country, sport, gender, and athletes.
- Created simple charts to visualize patterns.
- Saved cleaned data and helper tables for reporting.

4) Insights

- Frequent top countries (based on this file): United States, Soviet Union, Australia, Germany, China.
- Sports with many medals: Aquatics, Athletics, Rowing, Hockey, Gymnastics.
- Medal counts across years show variation between 1976 and 2008.
- Gender breakdown is captured and shows distribution across categories present in the file.
- Medal type distribution (e.g., Gold/Silver/Bronze) is summarized.

5) Challenges

A few columns used different names across sources (e.g., 'discipline' vs 'sport', 'team' vs 'country'), so I normalized them. Some files are not UTF-8, so I used a safe encoding fallback to load the data. Light cleaning was enough to make the analysis clear and consistent.

6) Conclusion

The analysis highlights medal patterns by year, leading countries, popular sports, and active athletes in the 1976–2008 period. The simple visuals make it easy to understand where medals concentrate and how totals change over time.

7) How to Run

- Open and run the notebook: Olympics_EDA_Full.ipynb
- Cleaned data, tables, and figures are inside: outputs_olympics/