Olympics Data Analysis

Project Title: Olympics Data Analysis

Domain: Data Analytics & Predictive Modeling

Tools & Technologies Used: Python, Pandas, Matplotlib, Seaborn, NumPy, Scikit-learn, Jupyter Notebook

Dataset: Summer Olympics Medals Dataset (1976 - 2008) - Medal winners from Montreal 1976 to Beijing 2008

Project Objective

- Explore and analyze historical data of Olympic medalists.
- Identify key trends across countries, sports, athletes, and gender.
- Use data-driven analysis to uncover hidden insights.
- Enhance decision-making using visualization.
- Build a model to predict whether an athlete is likely to win a medal using machine learning techniques.

Data Cleaning and Preprocessing

- Dropped unnecessary columns like Event_gender and Country_Code.
- Removed fully null rows (117 entries).
- Converted Year to integer data type.
- Checked and handled null values.

Exploratory Data Analysis (EDA)

Q1. Cities hosting Olympics the most

No city hosted Olympics more than once between 1976 and 2008.

Q2. Cities with maximum events hosted

Beijing hosted the highest number of events, followed by Sydney and Athens.

Q3. Number of Unique Events

Total 334 unique events were conducted.

Sports with most events: Wrestling, Weightlifting, Judo.

Q4. Top Athletes by Medal Count

Michael Phelps won the most medals (16) in this period.

Q5. Gender Ratio in Winning

Male athletes dominated medal winnings. Certain events existed only for one gender.

Q6. Top Performing Countries by Year

USA, Soviet Union, Germany, Russia, and China consistently ranked high.

Q7. Sport-Wise Country Dominance

Example: Korea, South dominated Archery, Australia dominated Swimming.

Q8. Year-wise Country Performance Comparison

Merged results for East/West Germany into Germany and Soviet Union/Unified Team into Russia. - Observed rise and fall trends of medal dominance.

Predictive Analysis:

Model Used: Logistic Regression

Features Used:

- Country
- Sport
- Gender
- Event

Enhancements & Improvements:

- Dropped non-informative columns to improve model generalizability. - Label Encoded categorical features. - Combined ‘Gold’, ‘Silver’, ‘Bronze’ into a single binary target (1 = won a medal, 0 = no medal). - Trained and evaluated using train_test_split (70-30).
- Evaluation Metrics: - Accuracy Score - Confusion Matrix - Classification Report
- Model Outcome: - Reasonable accuracy in predicting medal winners based on limited features. - Can be further improved using athlete age, past records, country GDP, etc.

Insights & Conclusion

- USA and Russia were the most dominant countries overall.
- Certain sports/events are gender-exclusive or skewed.
- Michael Phelps stands out as the top-performing athlete.

- Visualizations helped uncover trends like city-wise hosting and sport-wise dominance.
- Predictive modeling demonstrated the possibility of anticipating medal wins, with scope for improvement.

Skills Demonstrated

- Data Cleaning & Transformation
- Exploratory Data Analysis (EDA)
- Data Visualization (Matplotlib, Seaborn)
- Machine Learning (Logistic Regression)
- Analytical Thinking & Interpretation

Challenges Faced & Overcome

- Incomplete data entries: resolved by dropping fully null rows.
- Non-uniform formatting: cleaned using pandas.
- Duplicated athlete names across events: noted and acknowledged.
- Modeling with limited features: simplified to demonstrate binary prediction.

Future Enhancements

- Integrate athlete physical metrics (age, height, weight).
- Incorporate country-wise sports infrastructure & funding.
- Build an interactive dashboard using Power BI or Tableau.