Olympics Data Analysis

Aim

For a project based on Olympics data analysis, the primary focus will be on exploring and understanding the dataset, performing exploratory data analysis (EDA), and uncovering trends and insights related to athletes, countries, and sports over the years.

Problem Statement

The primary goal is to:

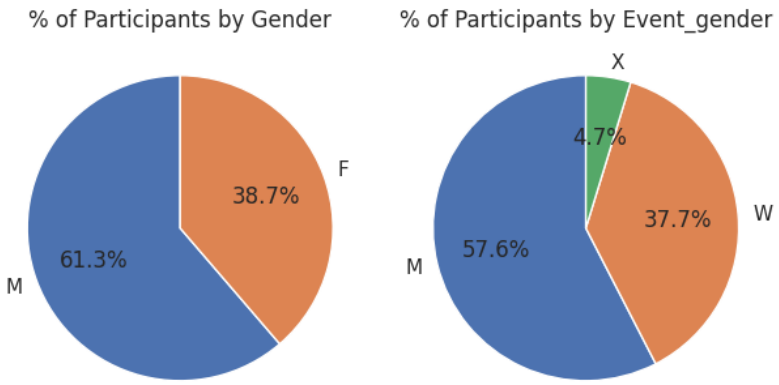
1. Analyze the dataset to understand trends in medal distribution.
2. Identify the top-performing countries and athletes.
3. Study the gender distribution of events and medals.
4. Visualize the data using Python.

1. Data Preparation

- Import libraries.
- Load the dataset.
- Clean the dataset (handling missing values).
- Standardize categorical values
- Check consistency between related columns
- Dropping unnecessary columns
- Normalize text values

2. Exploratory Data Analysis (EDA):

- Summary statistics of the dataset.
- Plot and analyze trends of medals across years.
- Identify the top-performing athletes and countries.

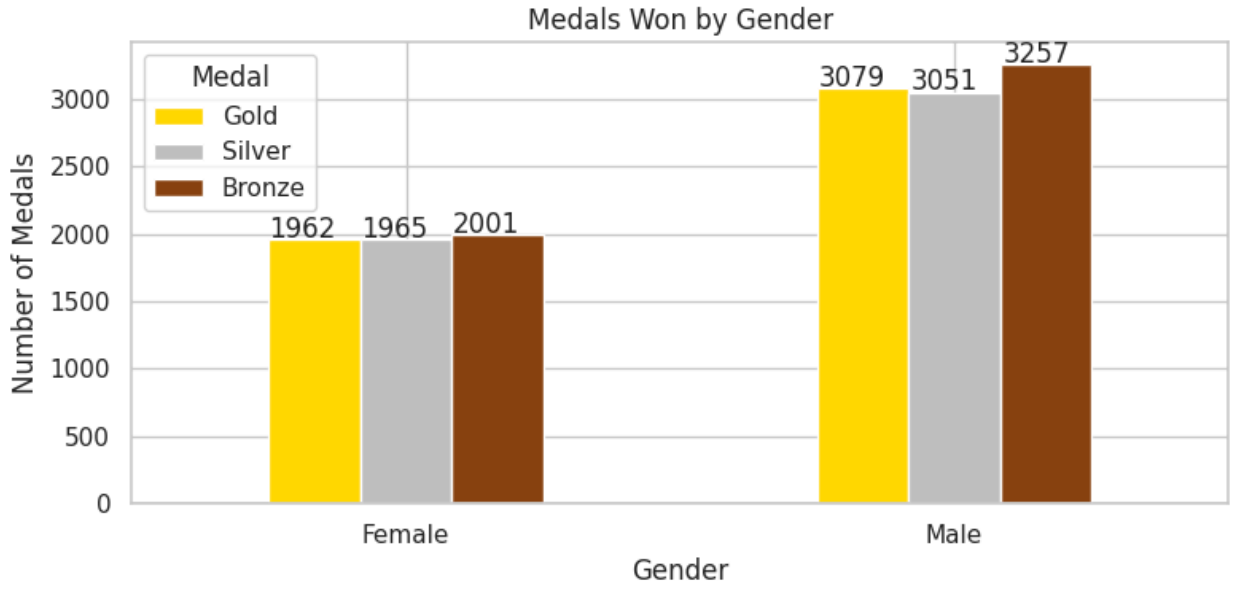


Gender Distribution (Left Pie Chart)

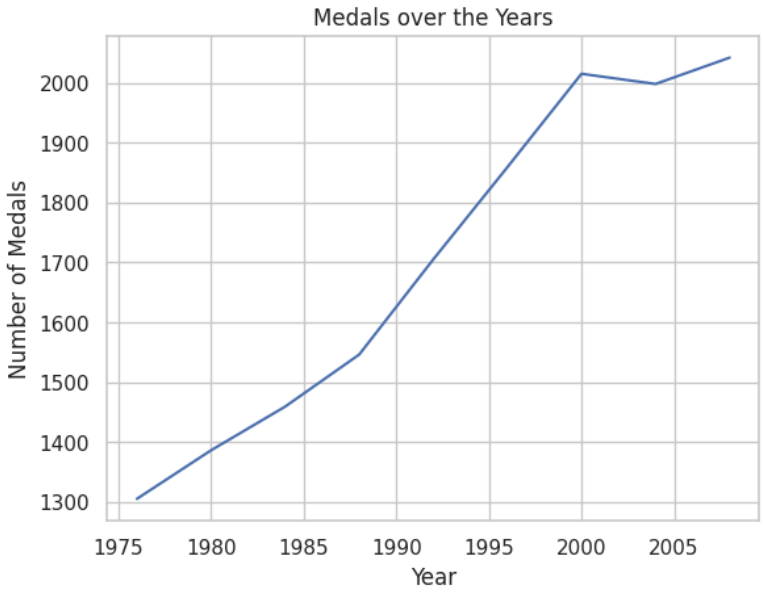
- Males (61.3%) outnumber Females (38.7%).
- This shows a gender gap in Olympic participation during 1976–2008.
- Over time, women's participation has been increasing, but in this dataset, men still dominate.

Event Gender Distribution (Right Pie Chart)

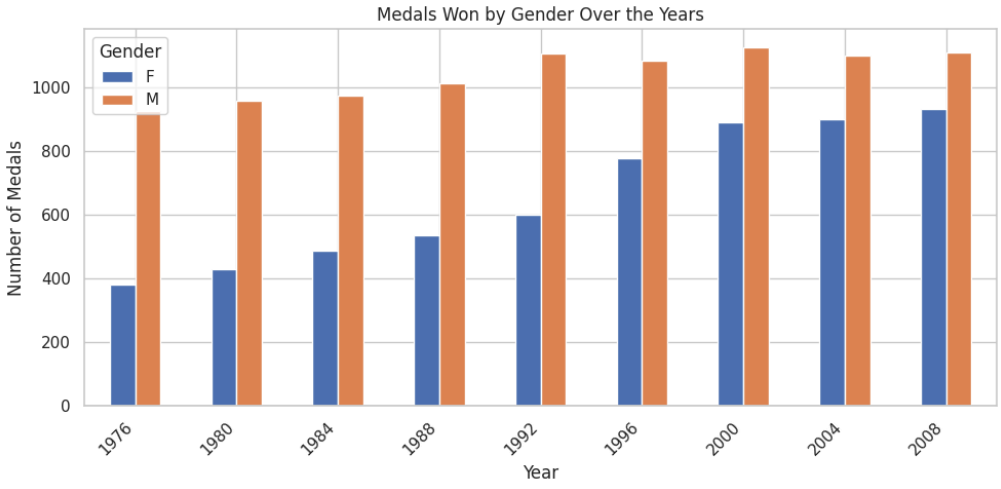
- Male-only events: 57.6%
- Female-only events: 37.7%
- Mixed events (X): 4.7%
- Mixed events are rare (~5%), but they play an important role



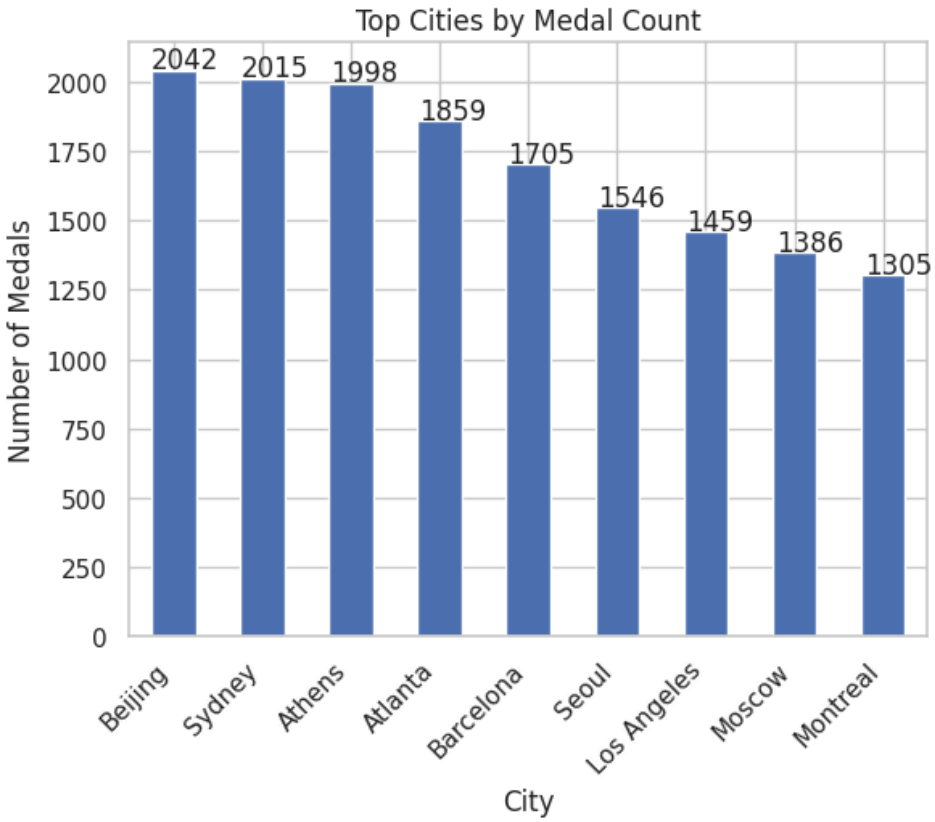
- Males dominate across all medal types
- The difference is consistent across Gold, Silver, Bronze, not just one medal type.
- However, women still secured ~40% of medals overall, which is a significant share compared to their historical underrepresentation.



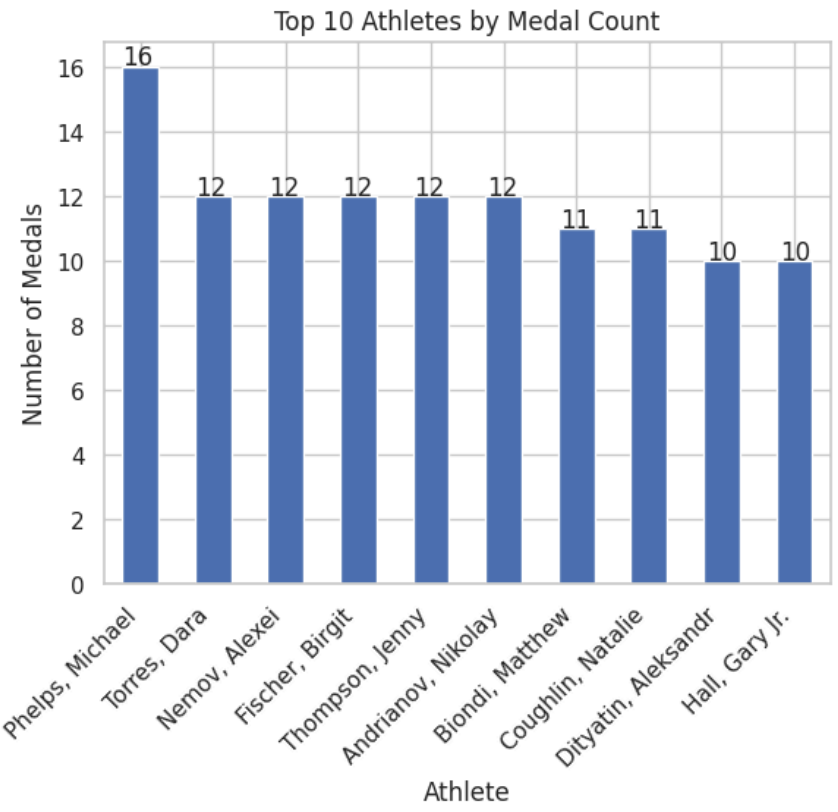
- The total number of medals has steadily increased over the years
- The biggest jump occurs between 1988 => 2000, reflecting:
    - Expansion of Olympic events.
    - Inclusion of more women's categories
- After 2000, medal counts stabilized but remained high.



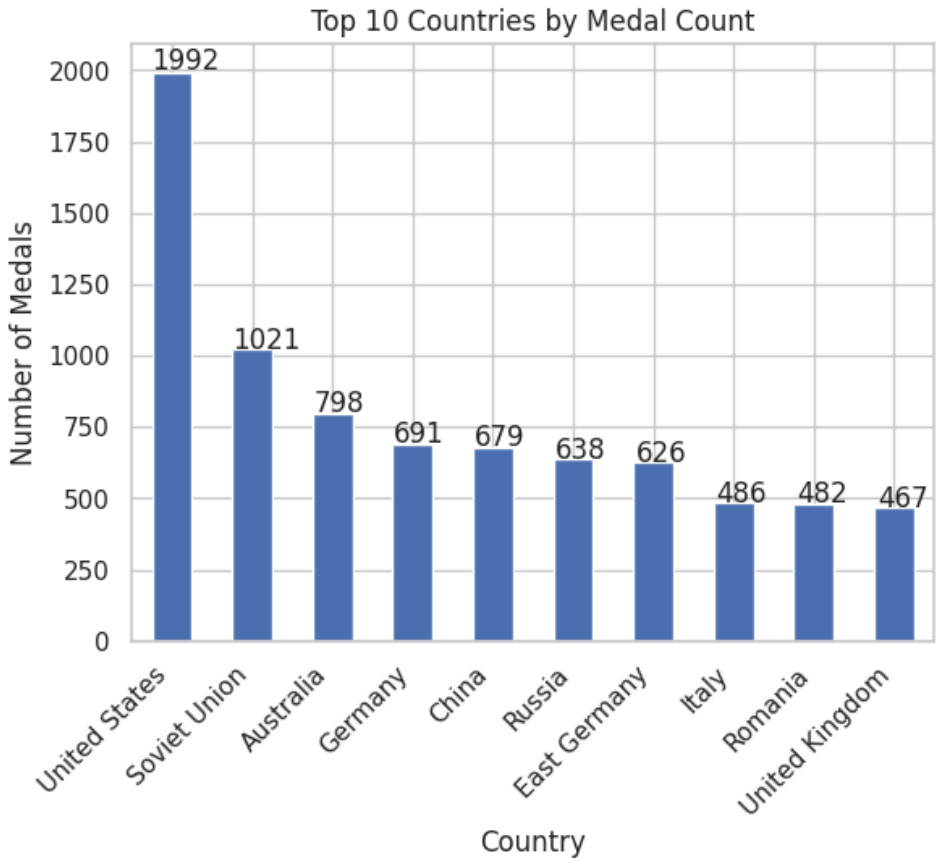
- 1976–1988: Large gap — men dominate, women's medals <50% of men's.
- 1992–2000: Female medals increase significantly, narrowing the gap.
- 2004–2008: Women's medals nearly reach parity (~900 vs ~1100).
- Trend shows steady rise in women's participation & medal success.



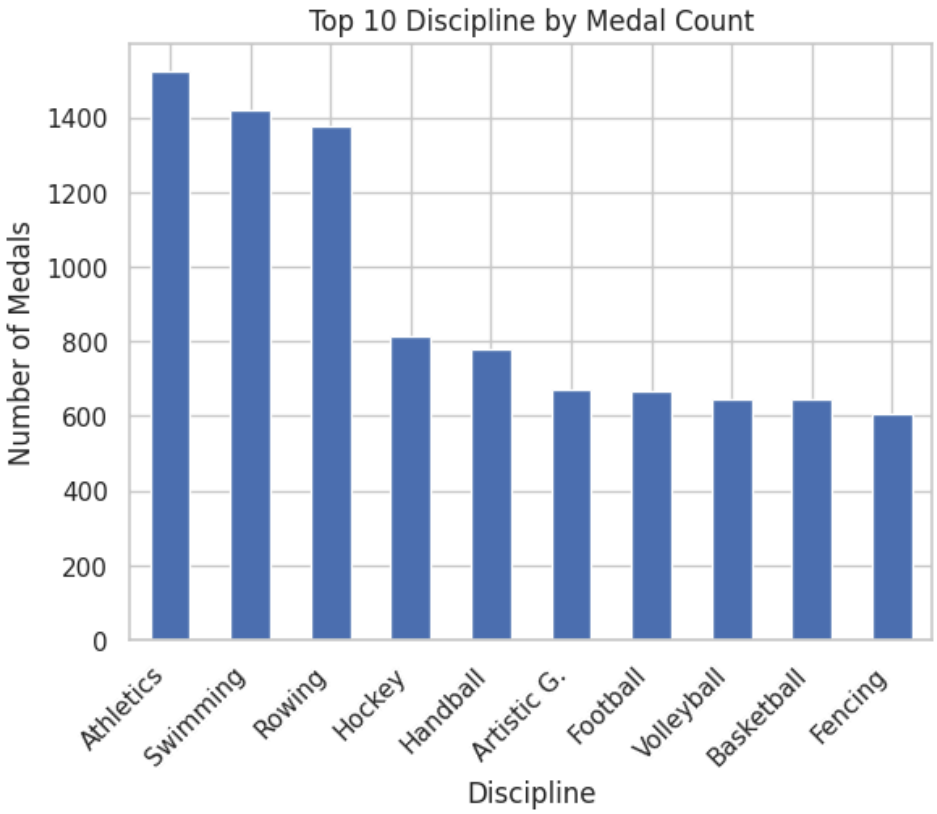
- Beijing (2042), Sydney (2015), Athens (1998) are the top 3 cities with the highest medal counts.
- Cities that hosted recent Olympics (Beijing 2008, Sydney 2000, Athens 2004) rank high,
- Earlier hosts like Montreal (1305) and Moscow (1386) are lower, showing the expansion of events in later years.



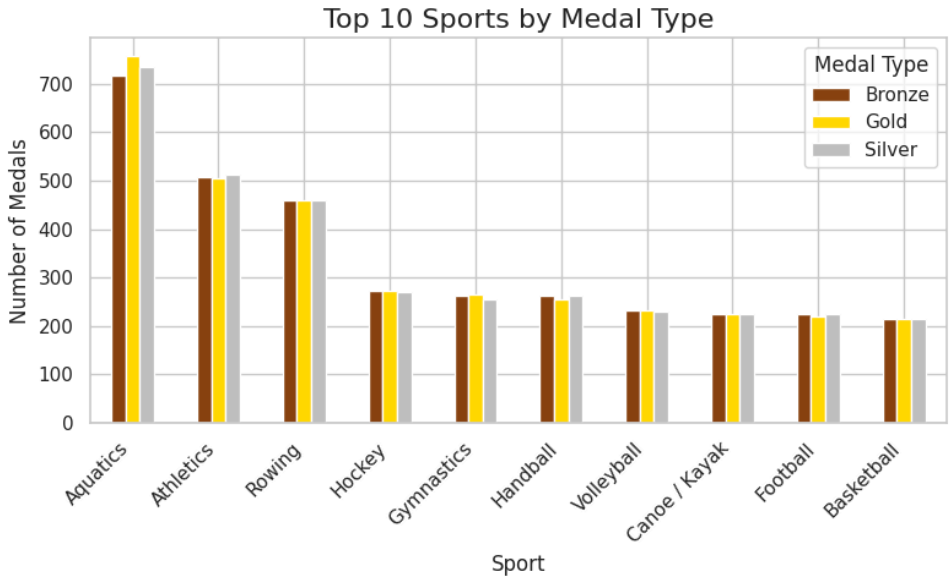
- Michael Phelps leads with 16 medals, far above others. He is the most decorated Olympian in this dataset (1976–2008).
- Several athletes with 12 medals: Dara Torres (USA, swimming), Alexei Nemov (Russia, gymnastics), Birgit Fischer (Germany, kayaking), Jenny Thompson (USA, swimming).
- Mix of sports: Swimming, Gymnastics, Rowing, Athletics, which offer multiple medal opportunities per athlete.



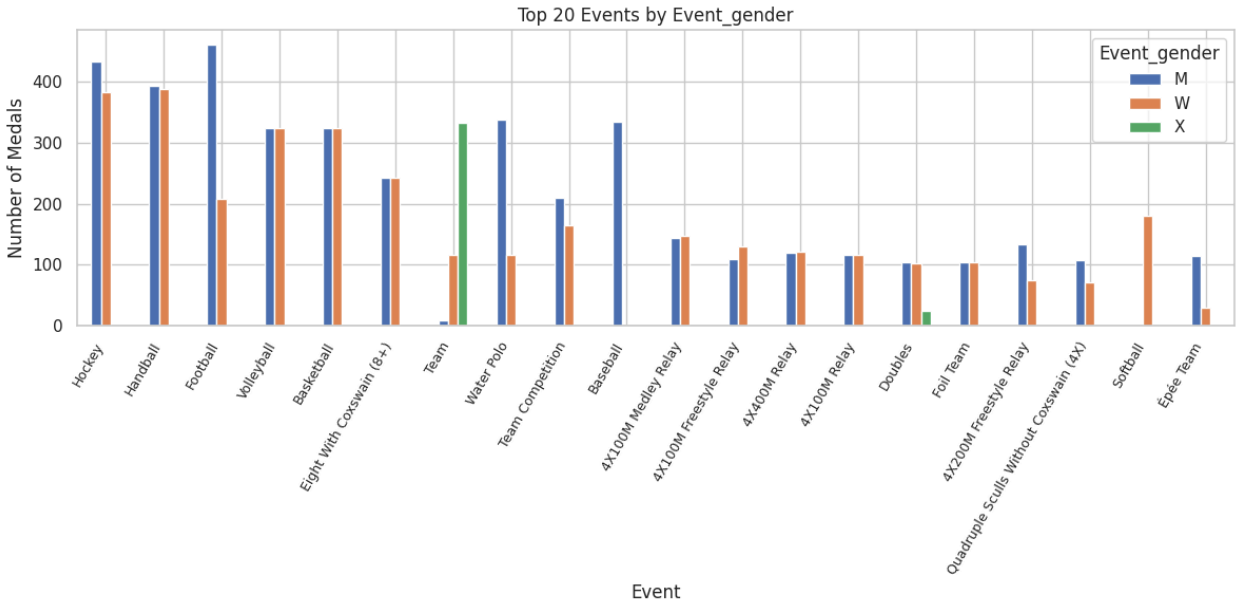
- USA dominates (1992 medals), far ahead of others.
- The Soviet Union (1021) was historically strong before dissolution.



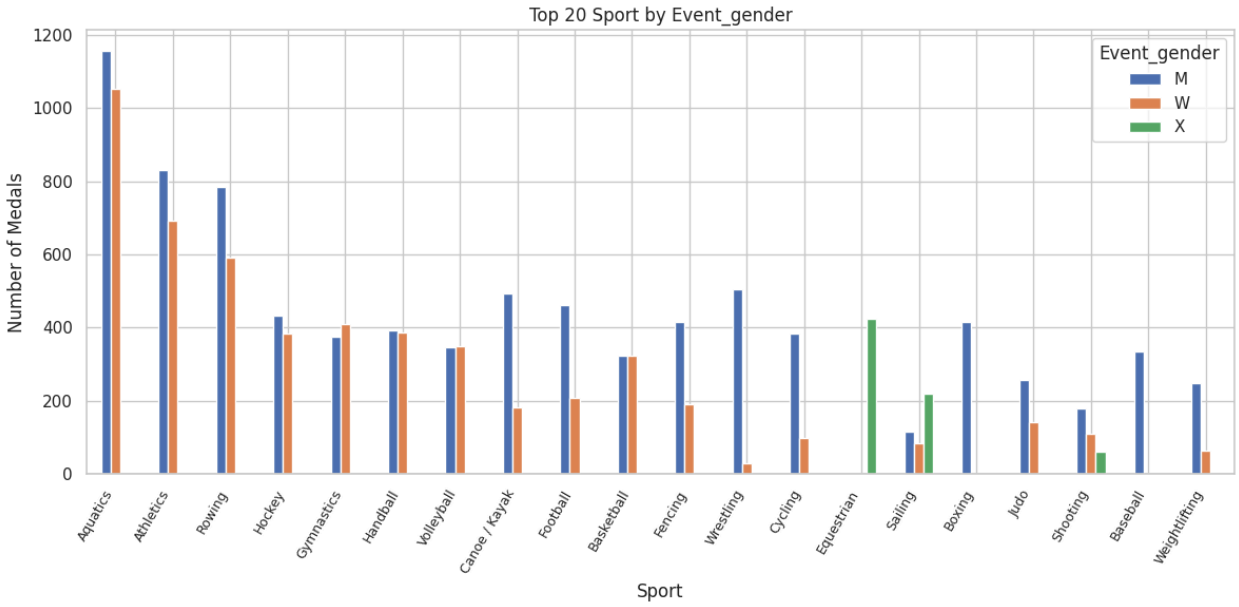
- Athletics and Swimming dominate, being core Olympic sports with many medal events.
- Rowing also has high medal counts due to multiple event categories.
- Team sports like Hockey, Handball, Volleyball, Basketball contribute, but with fewer medals since they award only 1 medal set per gender per edition.
- Artistic Gymnastics, Football, Fencing are long-standing traditional Olympic sports with steady medal contributions.



- Sports like Aquatics and Athletics dominate the Olympic medal tally, making them key contributors to overall medal counts.
- These sports also offer balanced opportunities for athletes across medal types.



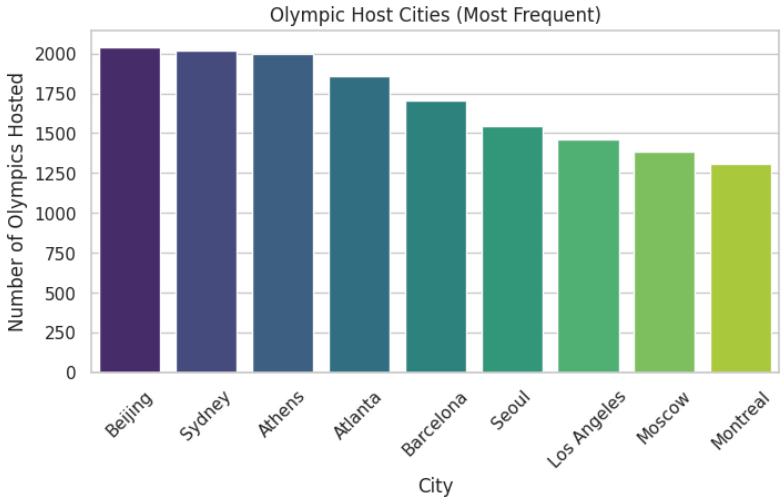
- Male athletes still dominate medal counts across most team sports, but women have strong participation in Hockey, Volleyball, Handball, and Basketball.
- Gender-exclusive events like Baseball (M) and Softball (W) create natural disparities. Mixed-gender events exist but contribute minimally.



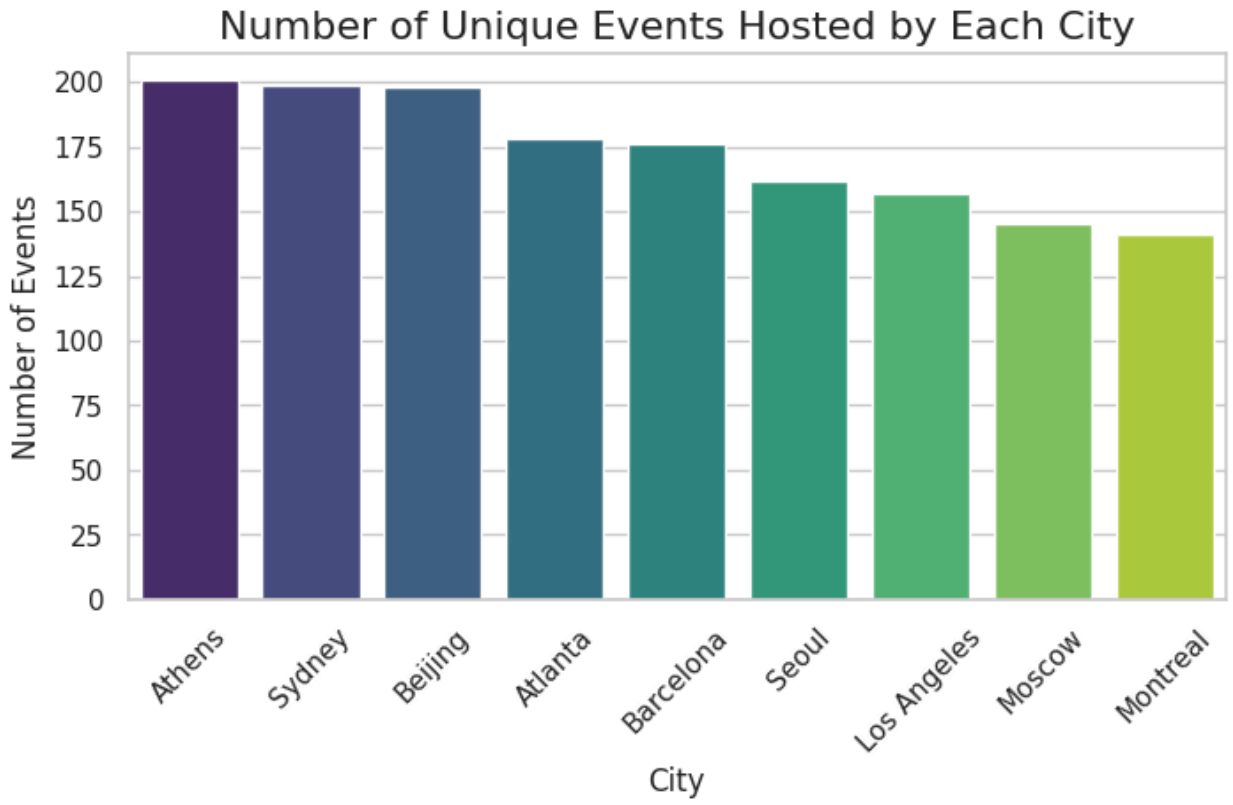
- In the most dominant sports, like Aquatics and Athletics, Male athletes have won a substantially greater number of medals.
- Some sports, such as Gymnastics and Handball, show a more balanced distribution of medals between genders.
- The number of medals from Mixed (X) events is relatively small across all sports, with the largest counts appearing in Equestrian and Sailing.

3. Questions

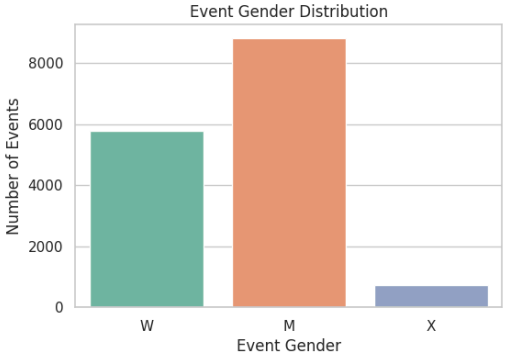
- Q1. Which city hosted maximum number of olympics

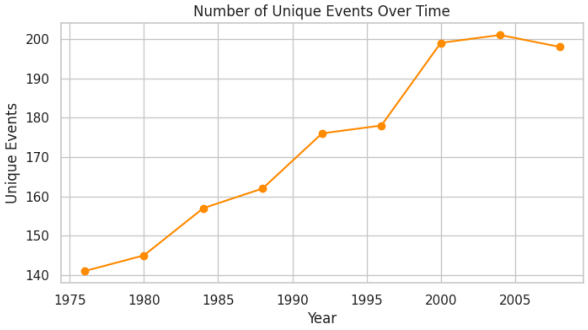


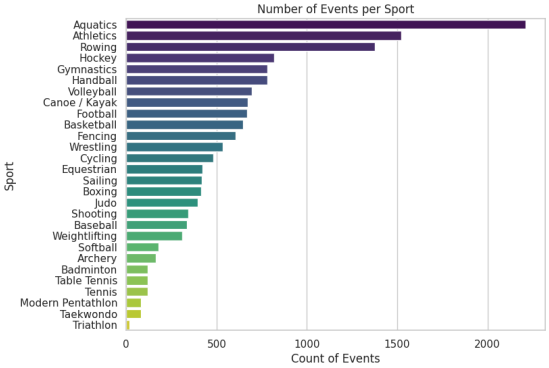
- Q2. Which city hosted the most events.

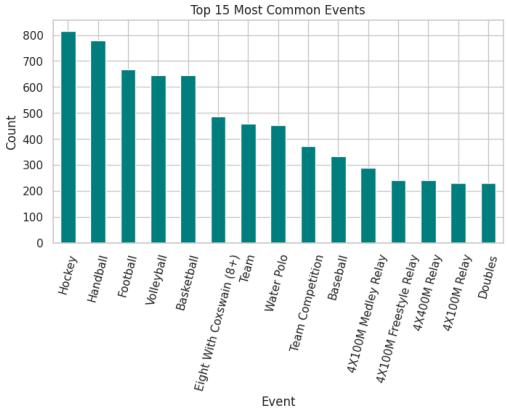


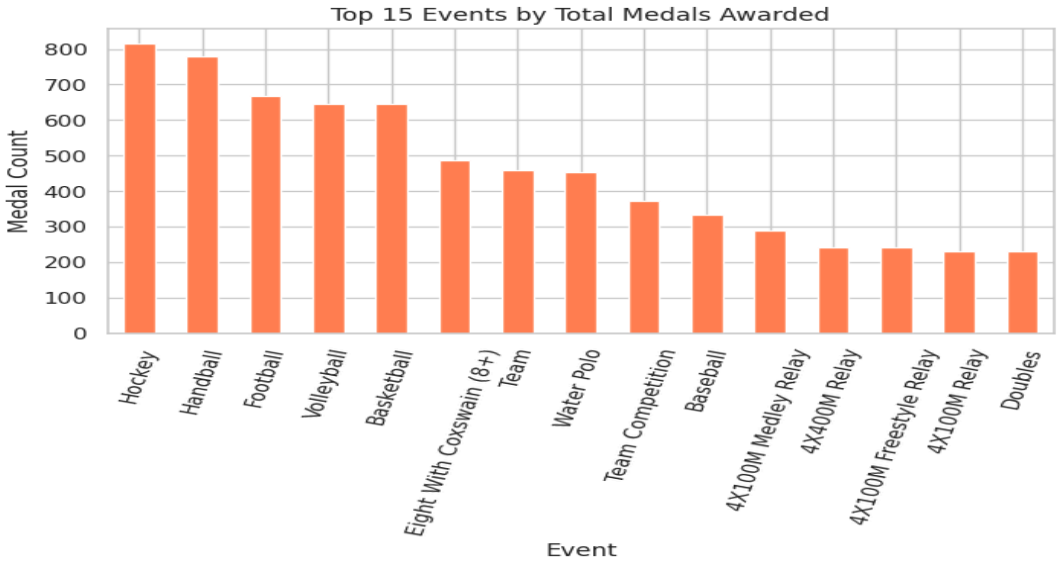
- Q3. Understand the events themselves.



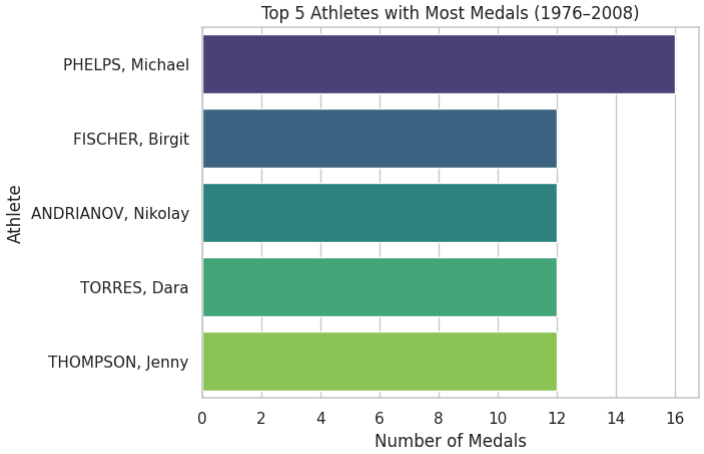




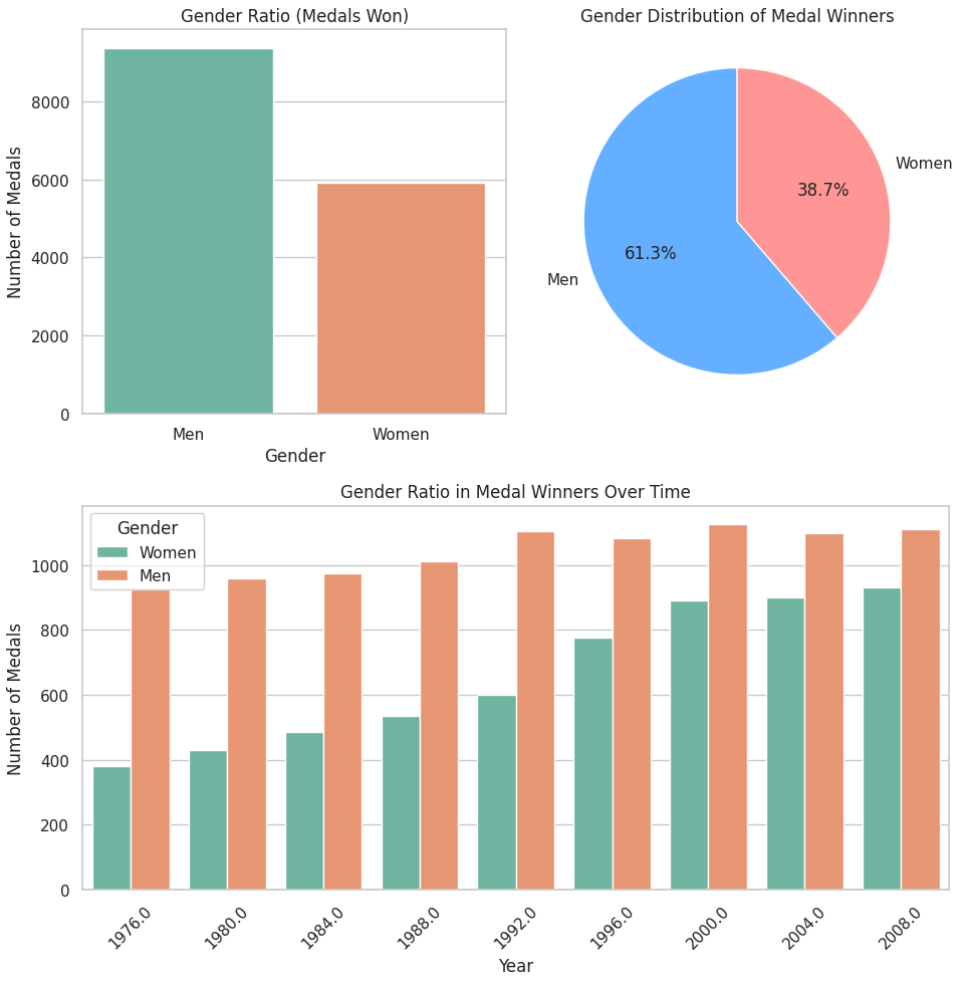




- Q4. Which Athlete has won the most medals from a given period?

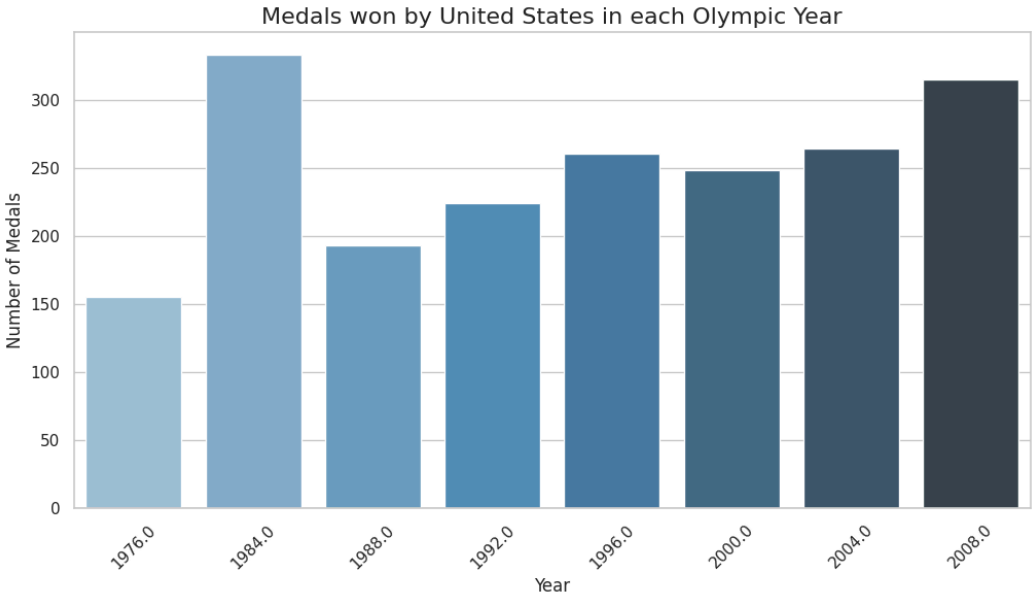


- Q5. Put some light on gender ratio in winning teams?

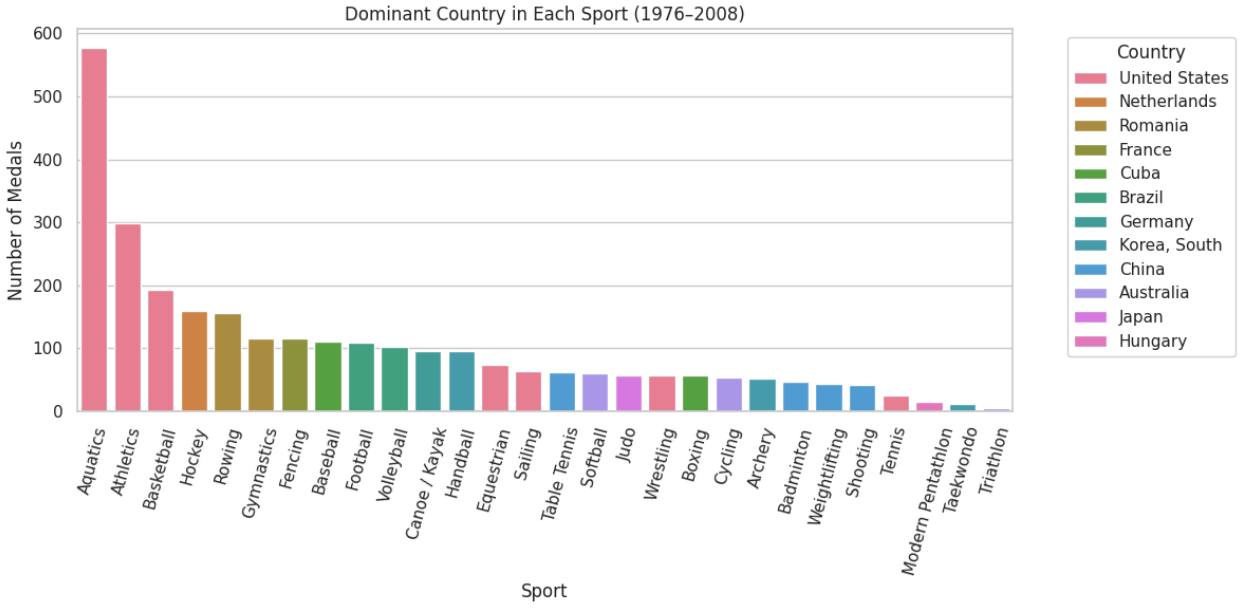


- Q6. Which country has won the most medals and how many in each year?

Top country: United States

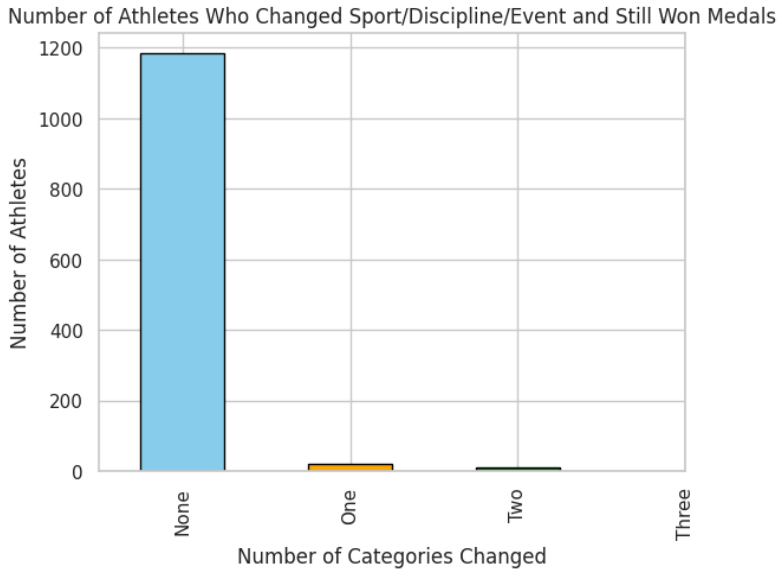


- Q7. Can you tell me which country has dominated any particular sport?



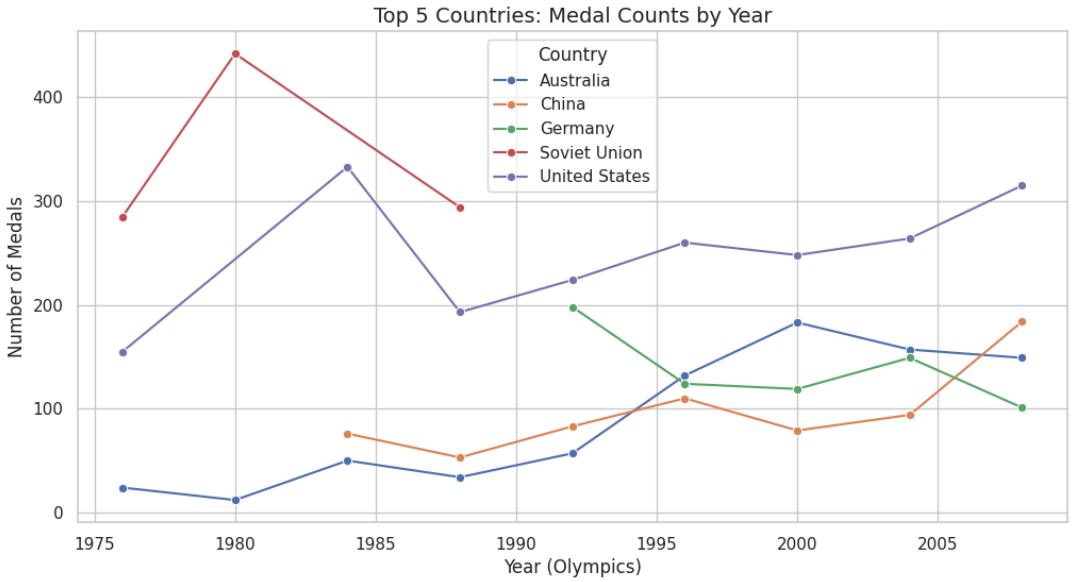
- Q8. Has any athlete changed his or her Event or Discipline or sport and still won the medal?

Athlete	Sport	Discipline	Event	Medal
ABBAGNALE, Agostino	1	1	2	3
ABDULLAYEV, Namig	1	1	2	2
ABEYLEGESSE, Elvan	1	1	2	2
ABRAHAM, Attila	1	1	2	3
ADAMS, Neil	1	1	2	2



- Q9. Which country has won the most medals and how many in each year? Elaborate the result and dive into details.(Pick any 5 country for this)

  United States has won the most medals (except in 1980)



4. Predictive Analysis:

Train a machine learning model to predict whether an athlete will win a medal based on their country, sport, and other attributes.

Steps:

1. Encode categorical variables

2. Feature Engineering
    1. Country-Level/City-Level Features
    2. Sport/Discipline-Level Features
    3. Athlete-Level Features
    4. Gender-Based Features
    5. Event-Level Features
    6. Temporal Features

3. Splitting and Training Dataset

    Training shape: (12252, 20) (12252,)

    Testing shape: (3063, 20) (3063,)

4. Choose a Model
    1. Logistic Regression
    2. Decision Tree
    3. Random Forest
    4. XGBoost
    5. LightGBM

5. Results

    Logistic Regression

Accuracy: 0.4101

	precision	recall	f1-score	support
0	0.40	0.55	0.47	1052
1	0.45	0.48	0.46	1008
2	0.35	0.19	0.25	1003

Decision Tree

Accuracy: 0.7107

	precision	recall	f1-score	support
0	0.71	0.71	0.71	1052
1	0.72	0.72	0.72	1008
2	0.70	0.70	0.70	1003

Random Forest

Accuracy: 0.7349

	precision	recall	f1-score	support
0	0.71	0.77	0.74	1052
1	0.76	0.74	0.75	1008
2	0.74	0.69	0.71	1003

XGBoost

Accuracy: 0.7173

	precision	recall	f1-score	support
0	0.70	0.75	0.72	1052
1	0.73	0.74	0.74	1008
2	0.72	0.66	0.69	1003

LightGBM

Accuracy: 0.6934

	precision	recall	f1-score	support
0	0.66	0.75	0.70	1052
1	0.71	0.73	0.72	1008
2	0.72	0.59	0.65	1003

Model Comparison:

Model	Accuracy(%)
Random Forest	73.49
XGBoost	71.73
Decision Tree	71.07
LightGBM	69.34
Logistic Regression	41.01

5. Conclusion

- Top Performing Country is : United States
- Top Athletes: We identified athletes who won the most medals are :
    1. Nemov, Alexei (12 medals)
    2. Andrianov, Nikolay (12 medals)
    3. Diabatin, Aleksandr (11 medals)
- Gender Participation:
    1. Male : 64%
    2. Female : 36%
- Trend of Medals Over Years:
    We visualized the trend of medal wins over the years, that shows the increasing graph and a significant increase, peaking at over 2000 medals in 2000.
- Predictive Analysis:
    Random forest performs best with 73.49% accuracy.