📘 Internship Project Report

Title: Olympic Data Analysis & Performance Insights
Internship Organization: Unified Mentor Pvt. Ltd.
Intern Name: Himanshu Kaushik
Domain: Data Analyst Intern
Duration: 3 Months
Tools & Technologies: Python, Pandas, NumPy, Matplotlib, Seaborn, Plotly, Streamlit, Excel, SQL
Dataset Source: Unified mentor

🔍 1. Objective

- Examine Olympic trends over time: participation (by sex, age), medals, and country representation.
- Explore travel's influence on performance via geospatial analysis.
- Create interactive dashboards and predictive models for performance forecasting.

📊 2. Dataset Overview

- Athlete records from **1896–2016**, with ~270,000 entries.
- Key features: ID, Name, Sex, Age, Height, Weight, Team, NOC, Year, Season, City, Sport, Event, Medal.
- Additional region data merged from NOC and country metadata for geographical context.

🧠 3. Data Cleaning & Feature Engineering

- Filled missing Medal entries ("NaN" → "None").
- Merged with NOC for athlete region; removed ambiguous cases like refugees.
- Added host city and capital city metadata (lat/long & altitude).
- Computed **DistanceTravelled** (geodesic distance between athlete's home capital and host city).
- Encoded:
    - Age, Height, Weight (missing handled via median).
    - Sex, Season, Sport, Medal (one-hot encoding).
- Final cleaned dataset included ~270,147 records with 18 key attributes.

📈 4. Exploratory Data Analysis (EDA)

A. Athlete Demographics

- **Sex ratio** has steadily balanced over 120 years; some modern instances show parity or female dominance.
- **Age distribution**: most athletes cluster in early 20s; Winter Games show slightly older median age .

B. Country-Level Participation

- Examined UK, Germany, Sweden, Japan: host nations (e.g., UK 2012, Germany 1972) saw participation spikes.

---

5. Distance vs. Performance Analysis

- Hypothesis: greater travel distance negatively impacts medal counts; effect diminishes post-1980.
- **Pre-1980** correlation: **–0.1465** between DistanceTravelled and total medals.
- **Post-1980** correlation: **–0.0615**—indicating reduced travel disadvantage.

---

6. Interactive Dashboard

- Built with Streamlit:
    - Visuals include gender participation trends, age distributions, medal counts by country/sport.
    - Maps show medal-per-capita and travel effects.
    - Sliders and filters allow users to explore temporal and regional aspects.

---

7. Predictive Modeling

- *Potential scope:*
    - Predict medal-winning likelihood from features like DistanceTravelled, HostCityAltitude, AthleteCountry, etc.
    - Suggested models: logistic regression or Random Forest, calibrated via cross-validation and evaluated with accuracy/F1/AUC metrics.

✅ 8. Insights & Conclusions

- **Gender parity** in participation has steadily improved; anomalies in small countries like Sweden recently.
- Athletes primarily aged in early 20s with some sport-specific variations.
- **Host country effect** visible in participation spikes.
- **Travel distance** modestly hinders performance; modern logistics reduce this effect.
- The dashboard enables data-driven exploration for researchers, fans, and sports administrators.

---

9. Challenges & Limitations

- **Incomplete demographic data**: many missing height/weight entries, particularly pre-Union.
- **Merging inconsistencies**: country naming discrepancies required fuzzy matching; some entries removed.
- **Simplified distance proxy**: using capital cities may not reflect actual athlete origins.

---

10. Future Work & Recommendations

- **Data enrichment**: include GDP, population, Olympic funding, sport/event counts.
- **Modeling**: implement predictive models for medal tally per country/athlete.
- **Clustering**: segment countries/athletes by performance, geography, demographics.
- **Dashboard enhancements**: add predictive tools, year-over-year scenario simulations.

---

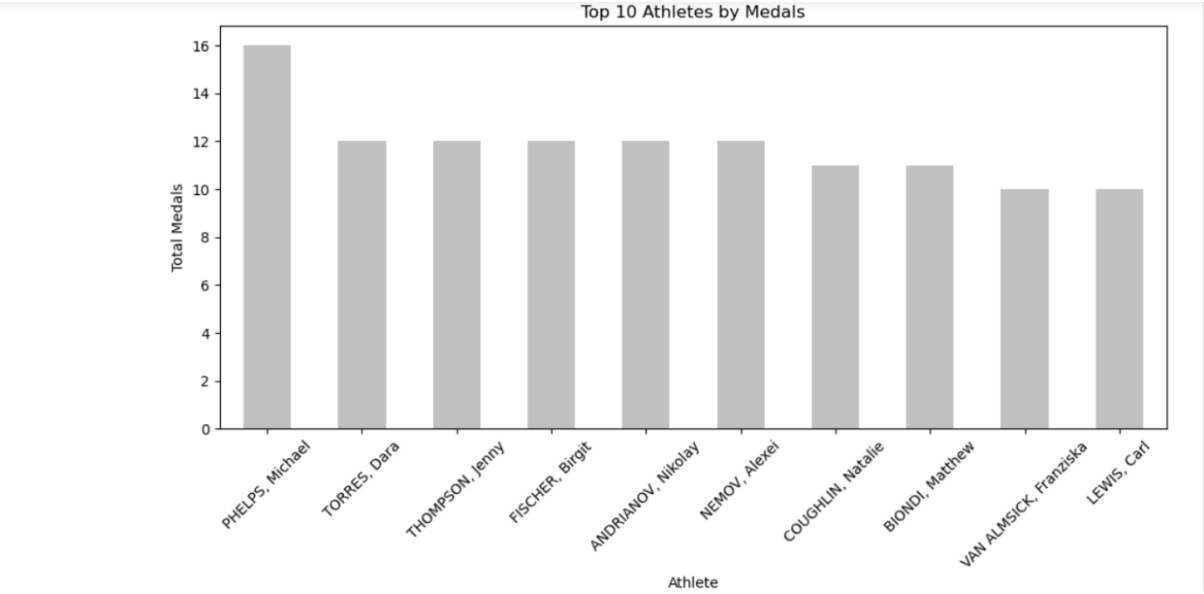
11. Appendices & Visualizations

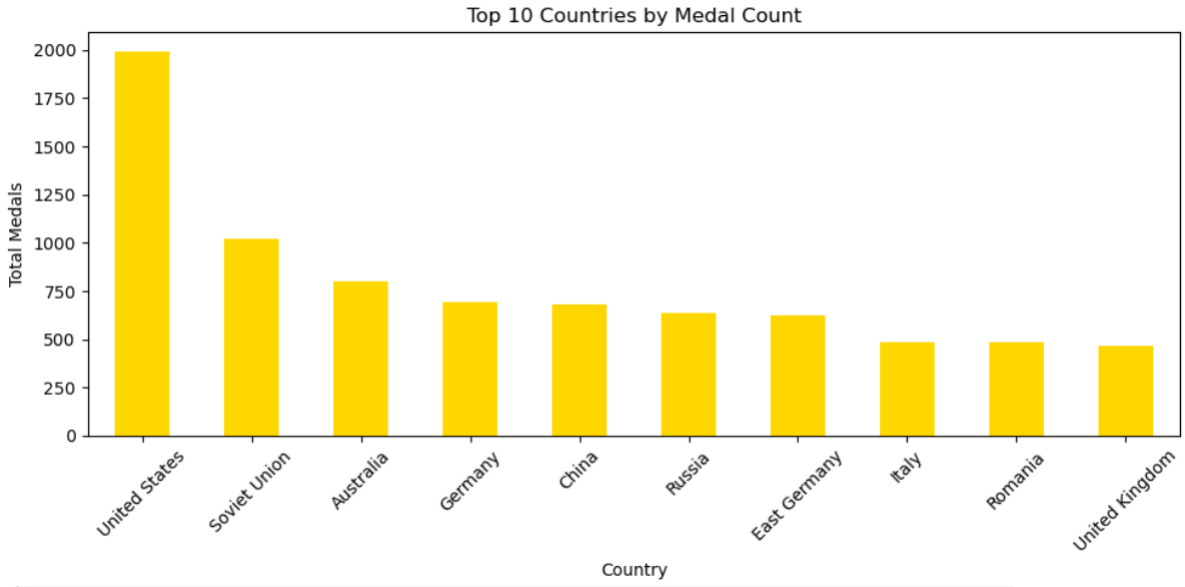
Include relevant:

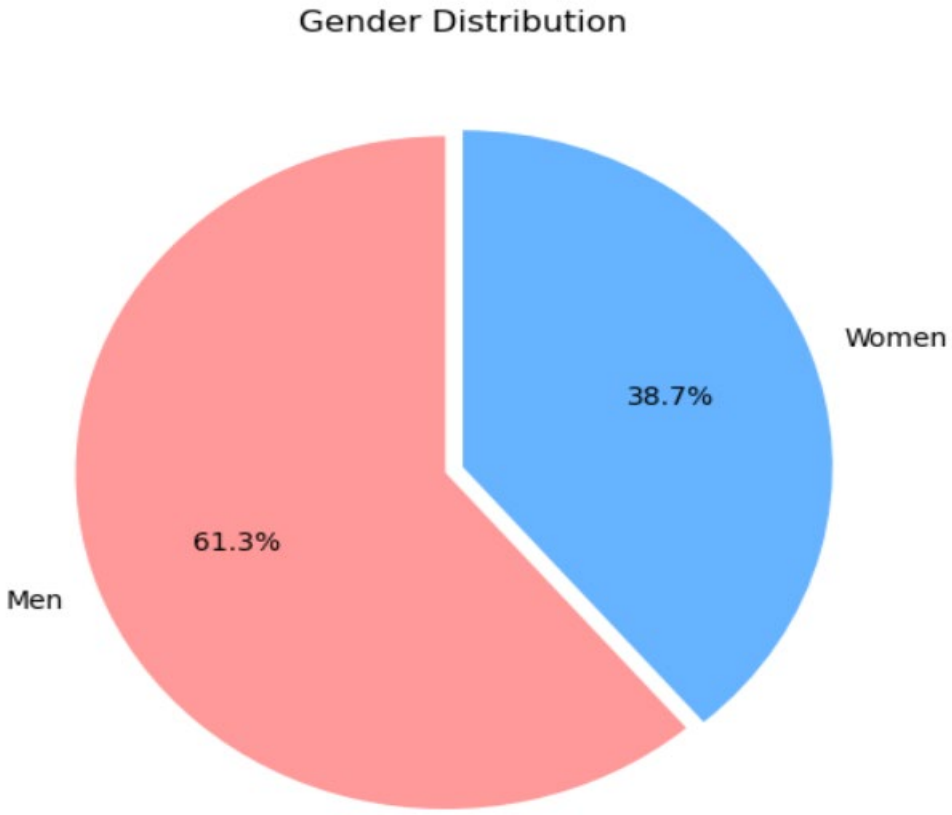
- Gender ratio and age distribution plots
- Host country participation spikes
- Distance vs. medals scatterplots
- Streamlit dashboard screenshots
- Fuzzy match logs & distance calculation tests

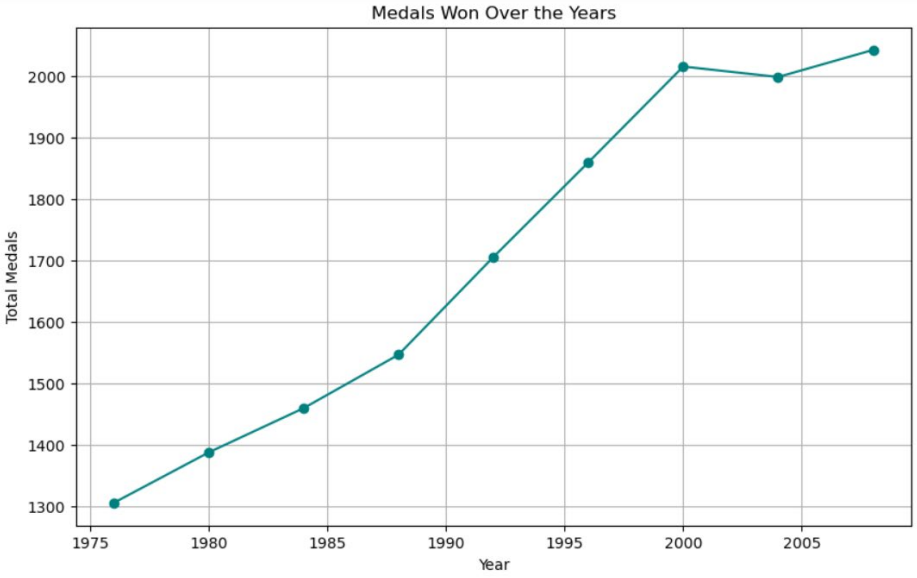
---

12. Snapshots:




Top 10 Countries by Medal Count
Total Medals
2000
1750
1500
1250
1000
750
500
250
0
United States
Soviet Union
Australia
Germany
China
Russia
East Germany
Italy
Romania
United Kingdom
Country


Gender Distribution
Women
38.7%
61.3%
Men


Medals Won Over the Years
Total Medals
Year
1300
1400
1500
1600
1700
1800
1900
2000
1975
1980
1985
1990
1995
2000
2005