Olympic Games Analysis

Project Overview

This project explores Olympic Games data to uncover trends, performance patterns, and key insights across various domains such as medal distribution, sports evolution, and athlete characteristics.
By leveraging Python and powerful visualization libraries, the analysis addresses ten critical research questions, providing a comprehensive understanding of Olympic history and performance.

Research Questions

1. Which are the top 10 countries with the most gold medals?
2. Which nations have consistently excelled in gymnastics over the years?
3. What is the gender distribution of Olympic participants, and how has it evolved over time?
4. How has the number of sports held in each Olympic year changed over time?
5. What is the age distribution of gold medalists, and what does it reveal about peak performance?
6. How has the number of Summer Olympic sports evolved over the years?
7. How has the number of Winter Olympic sports evolved over the years?
8. How has the average height of male and female athletes changed over time?
9. Who are the top-performing athletes in Ice Hockey over the past 10 Olympic Games?
10. How has India performed across all Olympic Games?

Datasets Used

- athlete_events.csv: Contains detailed data on athletes, their events, and outcomes.

- noc_regions.csv: Maps NOC codes to countries/regions for comprehensive analysis.

- olympics2024.csv: Aggregates data on country performances for ease of visualization.

Key Findings

1. The United States leads in gold medal counts, followed by Russia and China, showcasing a consistent dominance in global sports.

2. Nations like the USA, Russia, and China have demonstrated unparalleled excellence in gymnastics over decades.

3. A steady rise in female participation reflects the growing focus on gender equality in sports.

4. An increasing number of events indicate a significant diversification and inclusion of sports in the Olympics.

5. Gold medalists primarily belong to the 20-30 age range, highlighting this as the peak performance period for athletes.

6. Both Summer and Winter sports have expanded, with Winter sports growing at a more gradual pace.

7. While male athletes remain taller on average, the heights of both male and female athletes show a slight upward trend over time.

8. Athletes from Canada and the USA dominate Ice Hockey, excelling in both male and female categories.

9. India's medal tally has shown a positive trend, with significant achievements in field hockey, wrestling, and shooting in recent years.

Visualizations

Visualizations for all research questions are available in the olympic_visualizations folder.

These include:

- Medal counts by country.

- Trends in gender participation.

- Evolution of sports events over time.

- Athlete characteristics (age and height distributions).

(Screenshots of visualizations are embedded below.)

Tools and Libraries

- Python: For data processing and analysis.

- Pandas: For efficient data manipulation.

- Matplotlib and Seaborn: For creating insightful visualizations.

References

- Data provided by the user.

- Analysis and code generation supported by ChatGPT.

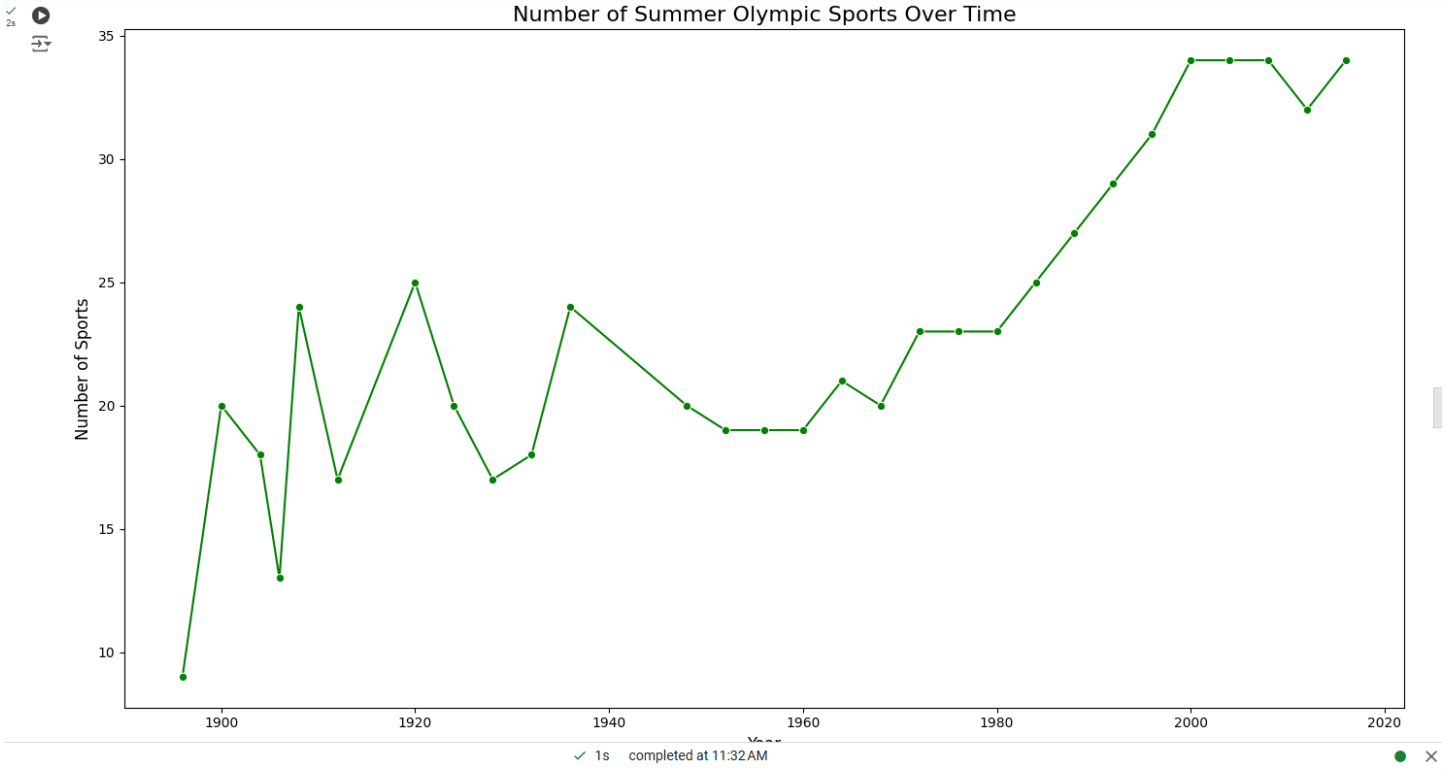
How to Use

1. Run the provided Python script (untitled0.py) to reproduce the analysis and visualizations.

2. Review the generated PNG files in the output directory for detailed insights.

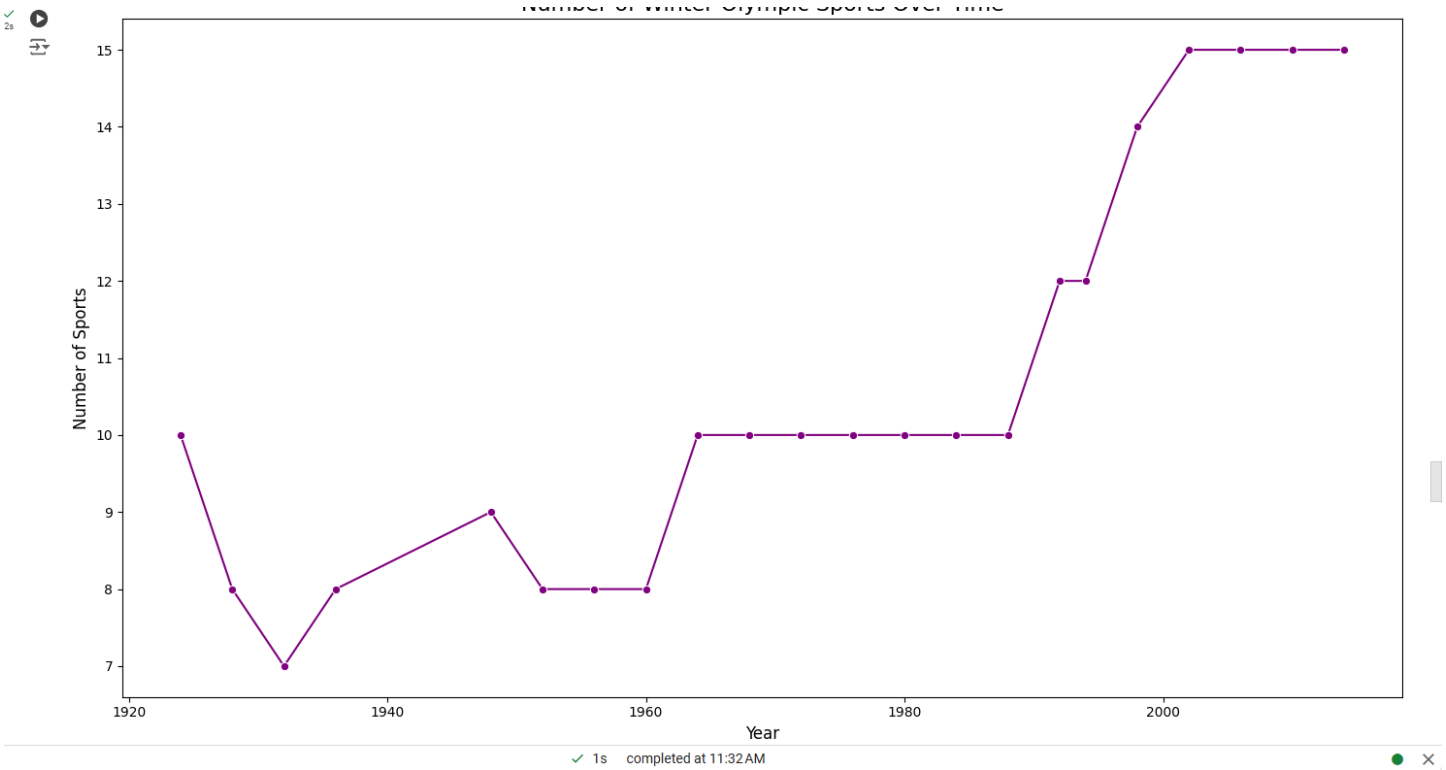
Screenshots of Visualizations


```
[45] # Load the datasets
     olympics_data = pd.read_csv('/content/drive/MyDrive/dataset/olympics2024.csv')
     noc_regions_data = pd.read_csv('/content/drive/MyDrive/dataset/noc_regions.csv')
     athlete_events_data = pd.read_csv('/content/drive/MyDrive/dataset/athlete_events.csv')

     # Task 6: Summer Sports Over Time
     def plot_summer_sports():
         # Filter data for Summer Olympics
         summer_sports = athlete_events_data[athlete_events_data['Season'] == 'Summer']

         # Count unique sports by year
         sports_count = summer_sports.groupby('Year')['Sport'].nunique().reset_index()

         # Plot
         plt.figure(figsize=(14, 8))
         sns.lineplot(data=sports_count, x='Year', y='Sport', marker="o", color='green')
         plt.title('Number of Summer Olympic Sports Over Time', fontsize=16)
         plt.xlabel('Year', fontsize=12)
         plt.ylabel('Number of Sports', fontsize=12)
         plt.tight_layout()
         plt.show()

     # Run the function
     plot_summer_sports()
```



```python
# Load the datasets
olympics_data = pd.read_csv('/content/drive/MyDrive/dataset/olympics2024.csv')
noc_regions_data = pd.read_csv('/content/drive/MyDrive/dataset/noc_regions.csv')
athlete_events_data = pd.read_csv('/content/drive/MyDrive/dataset/athlete_events.csv')

# Task 7: Winter Sports Over Time
def plot_winter_sports():
    # Filter data for Winter Olympics
    winter_sports = athlete_events_data[athlete_events_data['Season'] == 'Winter']

    # Count unique sports by year
    sports_count = winter_sports.groupby('Year')['Sport'].nunique().reset_index()

    # Plot
    plt.figure(figsize=(14, 8))
    sns.lineplot(data=sports_count, x='Year', y='Sport', marker="o", color='purple')
    plt.title('Number of Winter Olympic Sports Over Time', fontsize=16)
    plt.xlabel('Year', fontsize=12)
    plt.ylabel('Number of Sports', fontsize=12)
    plt.tight_layout()
    plt.show()

# Run the function
plot_winter_sports()
```

✓ 1s completed at 11:32 AM

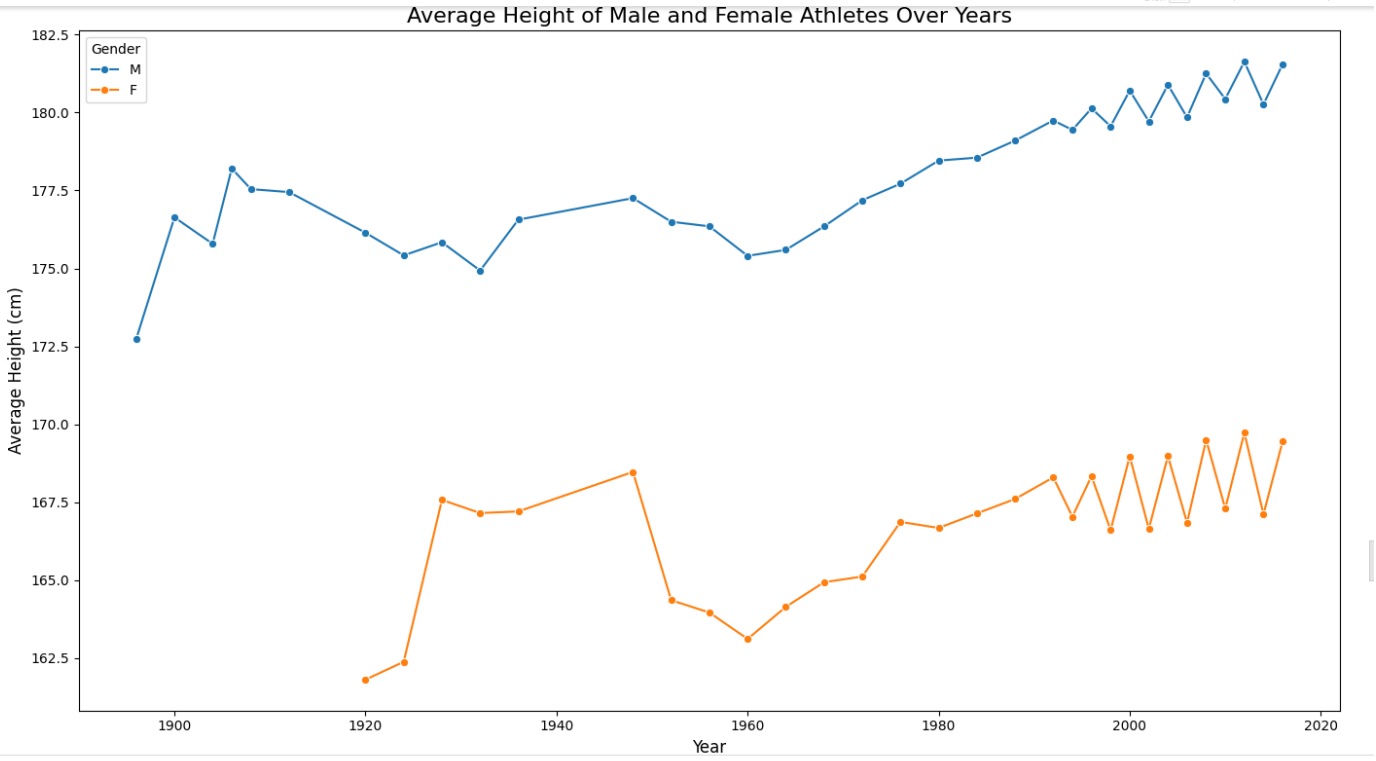


✓ 1s completed at 11:32 AM

```python
# Load the datasets
olympics_data = pd.read_csv('/content/drive/MyDrive/dataset/olympics2024.csv')
noc_regions_data = pd.read_csv('/content/drive/MyDrive/dataset/noc_regions.csv')
athlete_events_data = pd.read_csv('/content/drive/MyDrive/dataset/athlete_events.csv')

# Task 8: Average Height of Male and Female Athletes Over Years
def plot_avg_height():
    # Group by year and gender to calculate average height
    avg_height = athlete_events_data.groupby(['Year', 'Sex'])['Height'].mean().reset_index()

    # Plot
    plt.figure(figsize=(14, 8))
    sns.lineplot(data=avg_height, x='Year', y='Height', hue='Sex', marker="o")
    plt.title('Average Height of Male and Female Athletes Over Years', fontsize=16)
    plt.xlabel('Year', fontsize=12)
    plt.ylabel('Average Height (cm)', fontsize=12)
    plt.legend(title='Gender')
    plt.tight_layout()
    plt.show()

# Run the function
plot_avg_height()
```

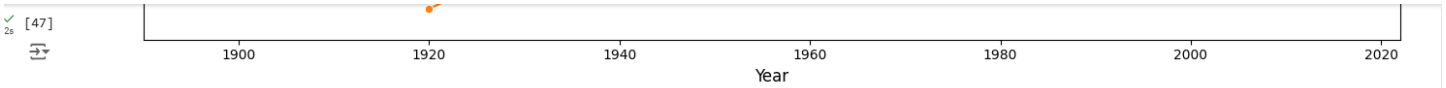
✓ 1s completed at 11:32 AM

+ Code + Text
RAM
Disk
Gemini

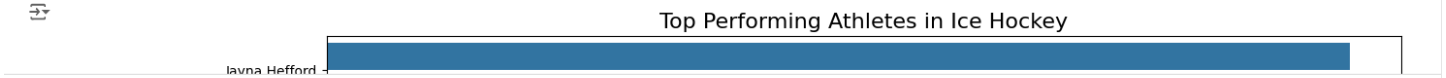


✓ 1s completed at 11:32 AM

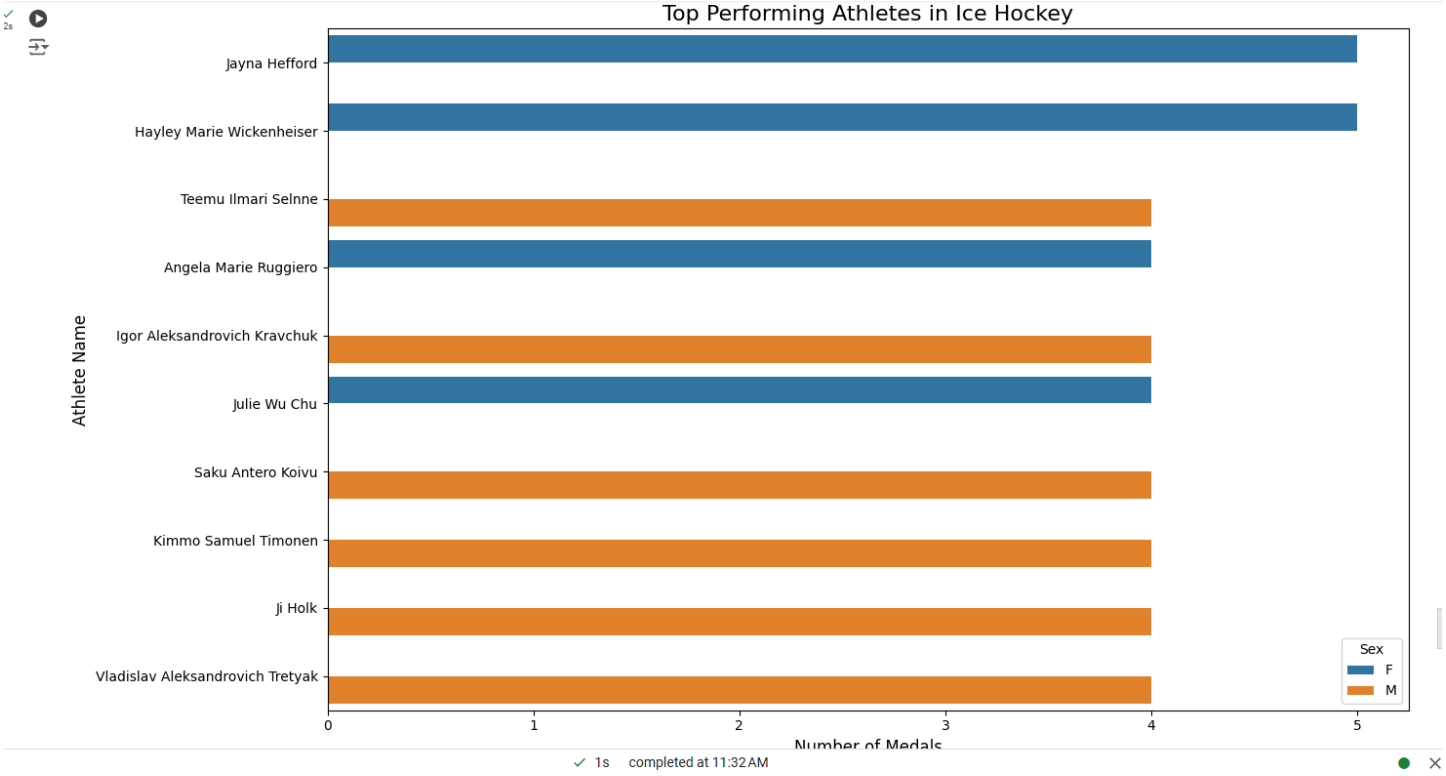
[47]




```python
# Load the datasets
olympics_data = pd.read_csv('/content/drive/MyDrive/dataset/olympics2024.csv')
noc_regions_data = pd.read_csv('/content/drive/MyDrive/dataset/noc_regions.csv')
athlete_events_data = pd.read_csv('/content/drive/MyDrive/dataset/athlete_events.csv')

# Task 9: Top Performing Athletes in Ice Hockey
def plot_top_ice_hockey():
    # Filter data for Ice Hockey
    ice_hockey = athlete_events_data[athlete_events_data['Sport'] == 'Ice Hockey']

    # Group by athlete name and gender to count medals
    top_athletes = (ice_hockey.groupby(['Name', 'Sex'])['Medal']
                    .count()
                    .sort_values(ascending=False)
                    .head(10)
                    .reset_index())

    # Plot
    plt.figure(figsize=(14, 8))
    sns.barplot(data=top_athletes, x='Medal', y='Name', hue='Sex', dodge=True)
    plt.title('Top Performing Athletes in Ice Hockey', fontsize=16)
    plt.xlabel('Number of Medals', fontsize=12)
    plt.ylabel('Athlete Name', fontsize=12)
    plt.tight_layout()
    plt.show()

# Run the function
plot_top_ice_hockey()
```

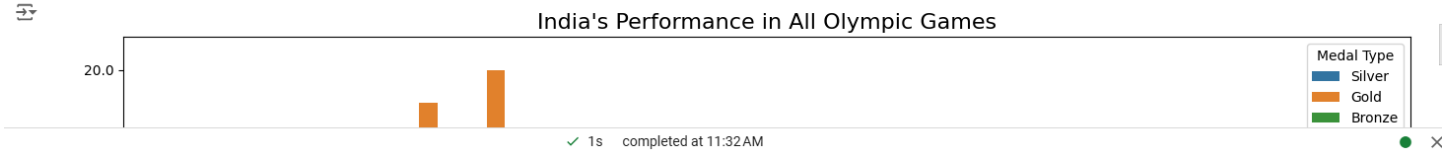


1s completed at 11:32 AM



1s completed at 11:32 AM


```python
# Load the datasets
olympics_data = pd.read_csv('/content/drive/MyDrive/dataset/olympics2024.csv')
noc_regions_data = pd.read_csv('/content/drive/MyDrive/dataset/noc_regions.csv')
athlete_events_data = pd.read_csv('/content/drive/MyDrive/dataset/athlete_events.csv')

# Task 10: Analyse India's Performance in All Olympic Games
def plot_india_performance():
    # Filter data for India
    india_data = athlete_events_data[athlete_events_data['NOC'] == 'IND']

    # Group by year and medal type
    india_medals = india_data.groupby(['Year', 'Medal'])['ID'].count().reset_index()
    india_medals.rename(columns={'ID': 'Count'}, inplace=True)

    # Plot
    plt.figure(figsize=(14, 8))
    sns.barplot(data=india_medals, x='Year', y='Count', hue='Medal')
    plt.title("India's Performance in All Olympic Games", fontsize=16)
    plt.xlabel('Year', fontsize=12)
    plt.ylabel('Number of Medals', fontsize=12)
    plt.legend(title='Medal Type')
    plt.tight_layout()
    plt.show()

# Run the function
plot_india_performance()
```



✓ 1s completed at 11:32 AM