PROJECT REPORT

Project Title: Paris Olympics 2024 Data Analysis

Introduction:

The Olympic Games are one of the most significant sporting events globally, uniting athletes from around the world in a celebration of sportsmanship and competition. The 2024 Summer Olympics, set to be held in Paris, promise to continue this rich tradition. With the rise of data analytics, there is an increasing opportunity to gain deep insights into the performance, trends, and outcomes of the games.

This project explores the potential of data analysis tools like Python and Power BI to analyze the extensive datasets generated by the Olympics. Python will be used for data extraction, cleaning, and preliminary analysis, while Power BI will create dynamic and insightful visual representations. The goal is to provide stakeholders, such as analysts, sports organizations, and fans, with comprehensive insights into the 2024 Paris Olympics, enhancing the overall understanding and appreciation of the event.

Objectives:

1. Athlete Demographics and Participation

2. Medal Distribution and Country Performance

3. Performance Trends for Various Activities.

4. Gender parity in event attendance

Scope of Work:

The project will involve the following tasks**:**

1.Data Collection and Cleaning:

- Gathering data related to the Paris 2024 Olympics from various sources (e.g., official websites, datasets).
- Cleaning and preprocessing the data using Python libraries (e.g., Pandas, NumPy, Matplotlib and Seaborn).

2.Data Analysis:

- Performing exploratory data analysis (EDA) to understand the data distribution and relationships.
- Clean and preprocess the data, ensuring that it is structured for modelling (e.g., handling missing values, normalizing data).

3.Data Visualization:

- A Power BI dashboard can effectively display insights into athlete demographics and medal distributions, offering interactive visuals for deeper exploration.
- Users can interact with the map to drill down into specific countries or regions, and compare medal counts across different Olympic Games.

4.Reporting and Documentation:

- Develop reports summarizing analysis findings, methodologies, and insights gained from the data. Include visual aids, tables, and charts to effectively communicate results.
- Document the data sources, analysis methods, and tools used throughout the project for transparency and reproducibility.

5.Final Documentation and Handover:

- Compile all documentation, including methodologies, findings, and visualizations, into a final report.

Methodology

The Paris 2024 Olympics data analysis project will follow a structured approach to ensure comprehensive and accurate insights. The methodology is outlined below:

1.Data Collection

- The dataset will be sourced from official Olympic databases, APIs, and other relevant data sources.
- Data will be collected on various aspects of the Olympics, including athlete performance, medal counts, ticket sales, and event schedules.

[58]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

Loading the Dataset

[59]:
```
df = pd.read_csv('C:\\users\\mohan\\Downloads\\athletedata.csv')
medal = pd.read_csv('C:\\users\\mohan\\Downloads\\archive (1)\\medals_total.csv')
medalist_data = pd.read_csv('C:\\users\\mohan\\Downloads\\archive (1)\\medallists.csv')
```

2.Data Preprocessing

- Handle missing data using forward fill techniques or imputation methods suitable for the dataset.
- Detect and remove duplicate entries and outliers to ensure data quality.
- Normalize data using techniques like Min-Max scaling or Standardization to prepare it for analysis and modelling.
- Calculate additional features like athlete performance metrics, event attendance rates, and revenue generated from ticket sales.

Data Cleaning

Handling Missing Values

[70]:
```
# Check for missing values
df.isnull().sum()
```

[70]:
```
code                  0
name                  0
name_short            3
name_tv               3
gender                0
country_code          0
country               0
nationality           3
nationality_long      3
nationality_code      3
height                3
weight                5
disciplines           0
events                0
birth_date            0
birth_place        2386
birth_country      1638
occupation         1529
education          5575
lang                508
coach              2891
dtype: int64
```

Handling missing data: Drop missing values or fill them with appropriate values (mean, median, etc.).

[72]:
```
df['height'].fillna(df['height'].mean(), inplace=True)
df['weight'].fillna(df['weight'].mean(), inplace=True)
```

[73]:
```
#Handle Irrelevant or Redundant Columns:

df.drop(columns=['name_short', 'name_tv'], inplace=True)
```

[74]:
```
df.drop(columns=['birth_place', 'birth_country','lang'], inplace=True)
```

[75]:
```
df.drop(columns=['nationality_long', 'nationality_code'], inplace=True)
```

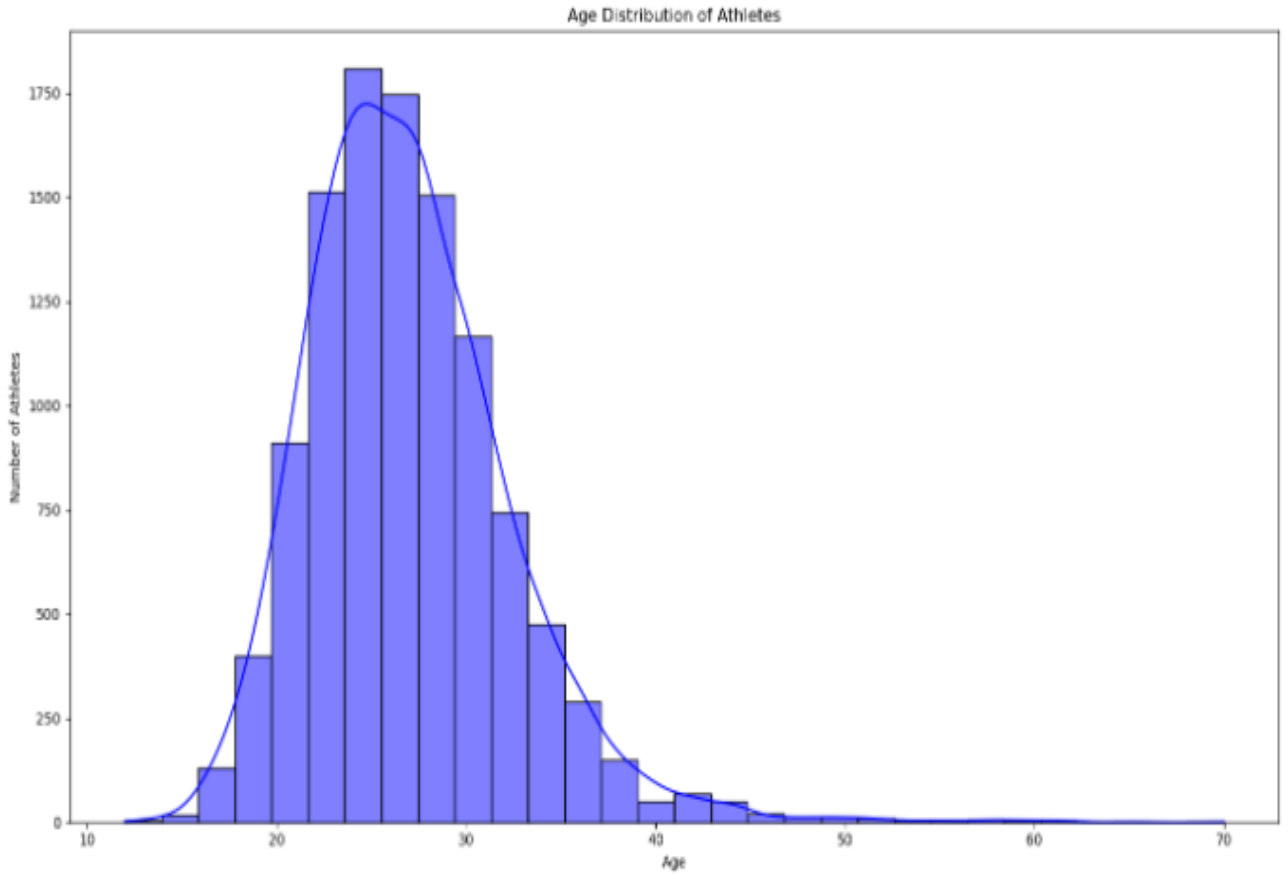
[76]:
```
# Fill missing values for categorical columns with 'Unknown' or the mode
df['education'].fillna(df['education'].mode()[0], inplace=True)
df['occupation'].fillna(df['occupation'].mode()[0], inplace=True)
df['coach'].fillna('Unknown', inplace=True)

# Fill missing nationality-related fields
df['nationality'].fillna('Unknown', inplace=True)
```

3.Exploratory Data Analysis (EDA)

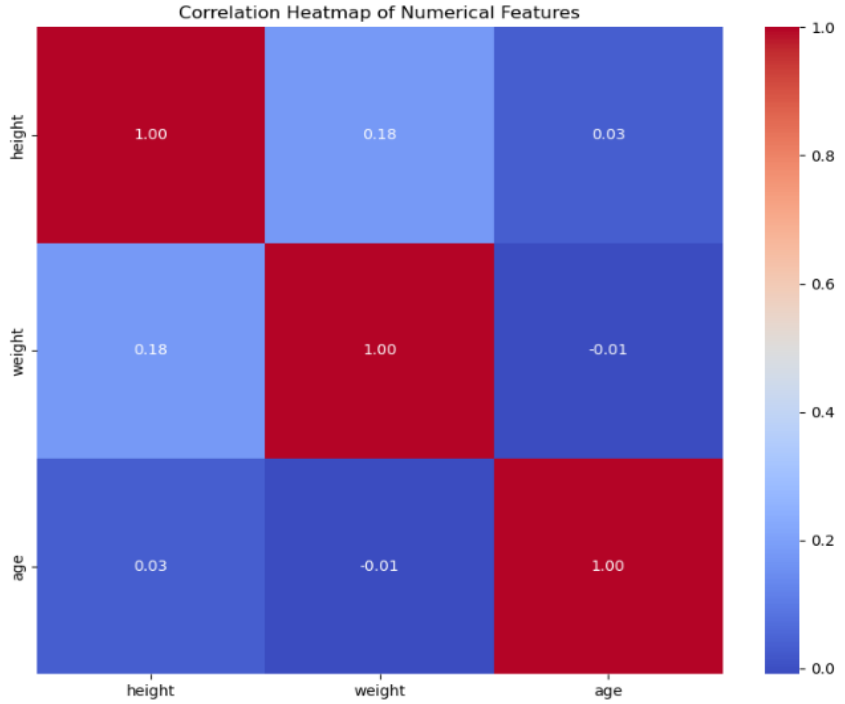
- Use descriptive statistics to summarize athlete performance, medal counts, and other relevant metrics.
- Visualize trends using line plots, bar charts, and heatmaps to explore patterns in the data.
- Generate pair plots and correlation matrices to examine relationships between different variables, such as athlete performance and medal counts.

EDA Visualizations

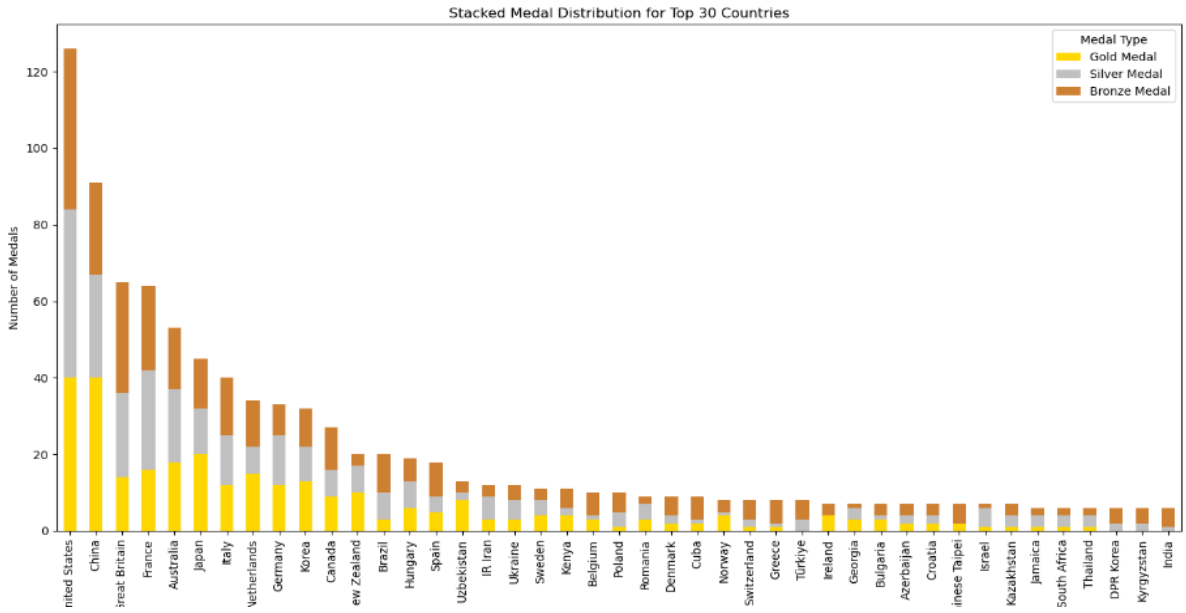
[95]:
```
# Create a histogram for age distribution
plt.figure(figsize=(14, 8))
sns.histplot(df['age'], bins=30, kde=True, color='blue', edgecolor='black')

# Add titles and labels
plt.title('Age Distribution of Athletes')
plt.xlabel('Age')
plt.ylabel('Number of Athletes')

plt.tight_layout()  # Adjust layout
plt.show()
```



[104]:
```
plt.figure(figsize=(10, 8))
correlation = df[['height', 'weight', 'age']].corr()
sns.heatmap(correlation, annot=True, cmap='coolwarm', fmt='.2f')
plt.title('Correlation Heatmap of Numerical Features')
plt.show()
```

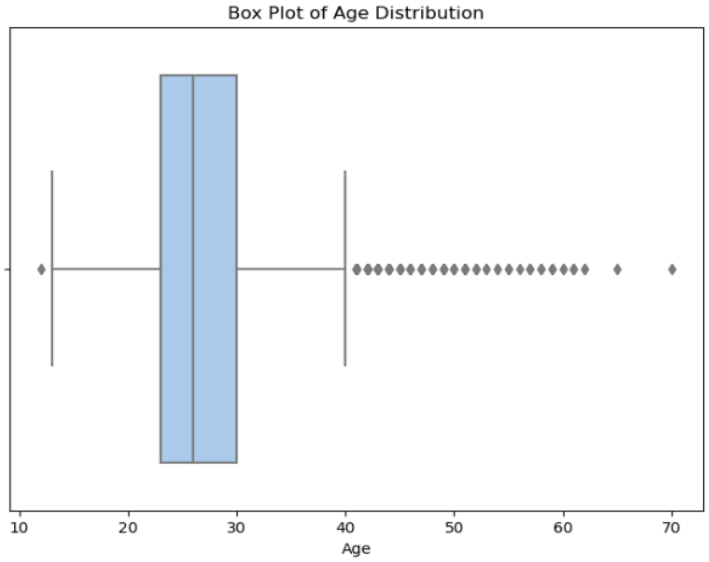


[105]:
```
top_countries = medal.nlargest(43, 'Total')  # Get the top 30 countries

# Set the country as the index for easier plotting
top_countries.set_index('country', inplace=True)

# Create a stacked bar chart
plt.figure(figsize=(14, 8))
top_countries[['Gold Medal', 'Silver Medal', 'Bronze Medal']].plot(kind='bar', stacked=True, color=['gold', 'silver', '#cd7f32'], ax=plt.gca())

# Add labels and title
plt.title('Stacked Medal Distribution for Top 30 Countries')
plt.xlabel('Countries')
plt.ylabel('Number of Medals')

plt.legend(title='Medal Type')

plt.tight_layout()  # Adjust layout
plt.show()
```



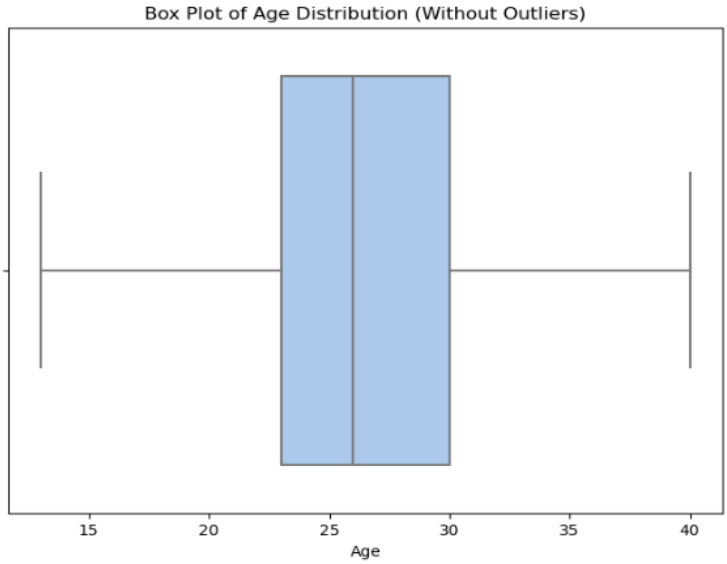
[96]:

```
plt.figure(figsize=(8, 6))
sns.boxplot(x='age', data=df, palette='pastel')
plt.title('Box Plot of Age Distribution')
plt.xlabel('Age')
plt.show()
```



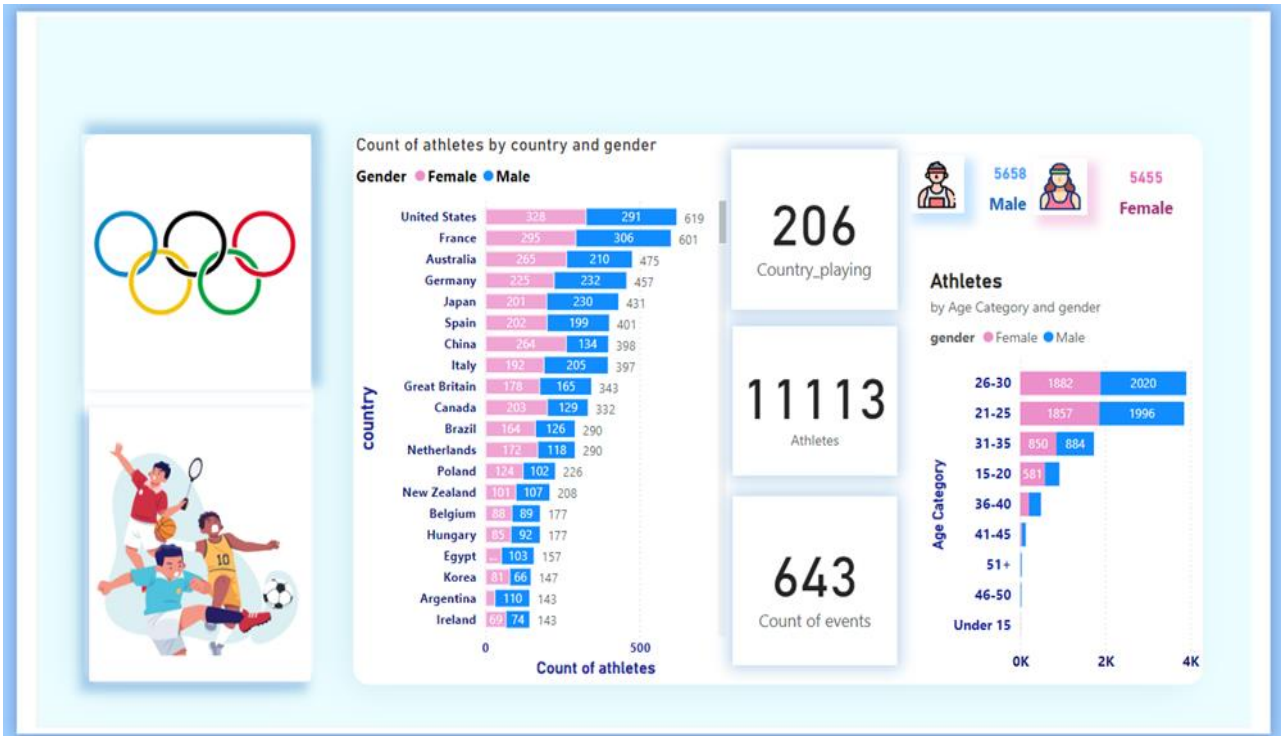
[97]:

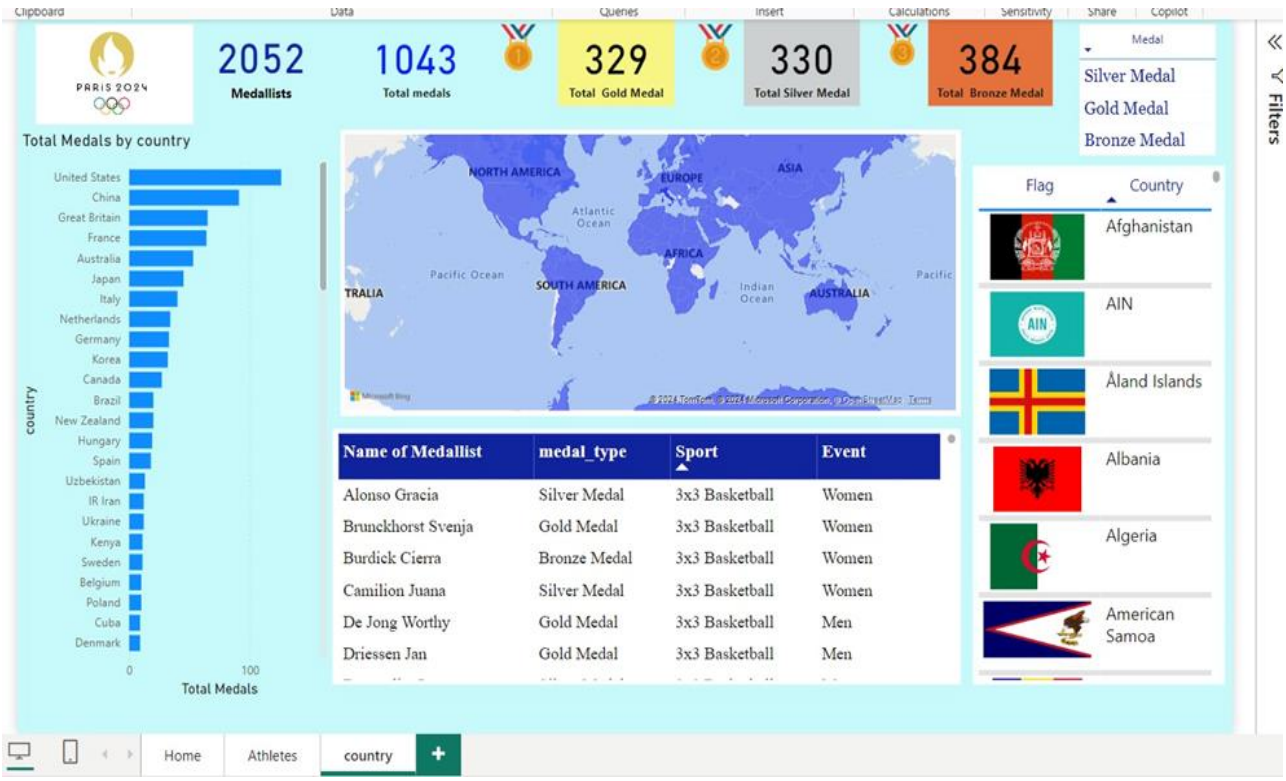
```
# Calculate Q1 (25th percentile) and Q3 (75th percentile)
Q1 = df['age'].quantile(0.25)
Q3 = df['age'].quantile(0.75)

# Calculate the IQR (Interquartile Range)
IQR = Q3 - Q1

# Define the bounds for outliers
lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR

# Remove outliers by keeping only the values within the IQR bounds
df_filtered = df[(df['age'] >= lower_bound) & (df['age'] <= upper_bound)]

# Plot the boxplot again without the outliers
plt.figure(figsize=(8, 6))
sns.boxplot(x='age', data=df_filtered, palette='pastel')
plt.title('Box Plot of Age Distribution (Without Outliers)')
plt.xlabel('Age')
plt.show()
```



4.Data Visualization

- A Power BI dashboard can effectively display insights into athlete demographics and medal distributions, offering interactive visuals for deeper exploration.
- Gender Ratio by Country: A stacked bar or pie chart showing gender distribution of athletes for each country.
- Country Medal Distribution: A map visualization indicating the total medal count by country for quick geographic insights.
- Country Filter: Allow filtering by country to view athlete and medal data for a specific nation.


Count of athletes by country and gender
Gender Female Male
United States 328 291 619
France 295 306 601
Australia 265 210 475
Germany 225 232 457
Japan 201 230 431
Spain 202 199 401
China 264 134 398
Italy 192 205 397
Great Britain 178 165 343
Canada 203 129 332
Brazil 164 126 290
Netherlands 172 118 290
Poland 124 102 226
New Zealand 101 107 208
Belgium 88 89 177
Hungary 85 92 177
Egypt 103 157
Korea 81 66 147
Argentina 110 143
Ireland 69 74 143
country
0
500
Count of athletes
206
Country_playing
11113
Athletes
643
Count of events
5658
Male
5455
Female
Athletes
by Age Category and gender
gender Female Male
26-30 1882 2020
21-25 1857 1996
31-35 850 884
15-20 581
36-40
41-45
51+
46-50
Under 15
Age Category
0K
2K
4K


Clipboard
Data
Queries
Insert
Calculations
Sensitivity
Share
Copilot
PARIS 2024
2052
Medallists
1043
Total medals
329
Total Gold Medal
330
Total Silver Medal
384
Total Bronze Medal
Medal
Silver Medal
Gold Medal
Bronze Medal
Total Medals by country
United States
China
Great Britain
France
Australia
Japan
Italy
Netherlands
Germany
Korea
Canada
Brazil
New Zealand
Hungary
Spain
Uzbekistan
IR Iran
Ukraine
Kenya
Sweden
Belgium
Poland
Cuba
Denmark
country
0
100
Total Medals
NORTH AMERICA
EUROPE
ASIA
Atlantic Ocean
AFRICA
Pacific Ocean
SOUTH AMERICA
Indian Ocean
AUSTRALIA
Pacific
Flag
Country
Afghanistan
AIN
Åland Islands
Albania
Algeria
American Samoa
Name of Medallist
medal_type
Sport
Event
Alonso Gracia
Silver Medal
3x3 Basketball
Women
Brunckhorst Svenja
Gold Medal
3x3 Basketball
Women
Burdick Cierra
Bronze Medal
3x3 Basketball
Women
Camilion Juana
Silver Medal
3x3 Basketball
Women
De Jong Worthy
Gold Medal
3x3 Basketball
Men
Driessen Jan
Gold Medal
3x3 Basketball
Men
Filters
Home
Athletes
country

Tools and Technologies:

- **Programming Language**: Python
- **Libraries**: Pandas, NumPy, Matplotlib, Seaborn
- **IDE**: Jupyter Notebook, Google Colab.
- **Data Source**: Kaggle
- **Data Visualization**: Power BI
- **Version Control**: Git.

Expected Outcome:

- Enhanced Athlete Performance Analysis.
- Medal Distribution Insights.
- Dynamic Performance Insights Dashboard

Timeline:

- Week1: Data Collection and Preprocessing
- Week2: Exploratory Data Analysis and Feature Selection
- Week 3: Data Analysis and Insights Generation using Python
- Week4: Visualization, Reporting, and Final Submission

Conclusion

The data analysis project for the Paris Olympics 2024, utilizing Python and Power BI, has successfully integrated recent performance metrics and athlete statistics to provide valuable insights. Through data collection and preparation, high-quality datasets were established, enabling in-depth analyses that revealed trends in athlete performance and medal distributions. An Interactive Recent Data Dashboard was developed in Power BI, allowing stakeholders to visualize and engage with the data effectively.