" The Olympics: A Data Odyssey- Exploring Performance and Trends "

Team: Innovation Pioneers

1. Ayushi Tawari
2. Mutta Datta Sai Vishnu Mohan
3. Niranjana J

Estimated Start Date: 04-04-2024

Estimated Completion Date: 10-04-2024

Details of the dataset:

- Athletes dataset
  Rows: 271116
  Columns: 15
  The dataset contains information about the athletes and their participation from 1896 to 2016.

- NOC regions dataset
  Rows: 271116
  Columns: 15
  The dataset contains information regarding different National Olympic Committees (NOC) and their respective regions.

Breakdown of the columns:

- ID: Unique identifier for each athlete.
- Name: Athlete's full name.
- Sex: Athlete's gender (male or female).
- Age: Athlete's age at the time of the competition.
- Height: Athlete's height (measurement in centimetres).
- Weight: Athlete's weight (measurement in kilograms).
- Team: Country the athlete represents.
- NOC: National Olympic Committee code (unique 3-letter code for each country).
- Games: Combined field indicating year and season (e.g., "2012 Summer").
- Year: Year of the Olympic Games.
- Season: Season of the Olympic Games (Summer or Winter).
- City: Host city of the Olympic Games.
- Sport: Specific sport the athlete competed in (e.g., gymnastics, swimming).

- Event: Specific event within the sport (e.g., 100m freestyle, men's high jump).
- Medal: Medal awarded to the athlete (gold, silver, bronze, or "NA" for no medal)

Overall Objective:

- Athlete Performance: Analyze factors influencing medal acquisition (e.g., age, height, weight, specific sports/events).
- Trends over Time: Identify trends in participation, medal distribution, and dominance in different sports across various Olympic Games.
- Demographic Analysis: Understand the demographics of Olympic athletes (age, sex, height, weight) across different sports and countries.
- National Performance: Compare performance of different countries based on medal count, participation rates, and athlete demographics.

By analyzing these different aspects, the dataset can reveal valuable information about the Olympics and the athletes who compete.

Inconsistencies in data:

The dataset is a Pubic Repository, which consists of

Missing Values:

```
notes     266077
Medal     231333
Weight     62875
Height     60171
Age         9474
region       370
```

Duplicates: 1385 rows

Ownership of each deliverable:

We've divided the work collaboratively. Each team member will analyze a specific aspect of the data, drawing their own conclusions and providing valuable insights. The PowerPoint presentation and final project report will be made the same way.

Here is a list of deliverables expected from each team member:

Ayushi Tawari:

- **Data Pre-processing : Data Cleaning**
  - Missing Value analysis: It is the process of identifying, understanding, and handling missing data in your dataset.
- **Time series analysis**
  - Analysis of different trends over the years.
- **Exploratory Data Analysis:**
  - Medal distribution analysis
  - Country performance analysis
  - Sports popularity analysis
  - Athlete performance analysis

Mutta Datta Sai Vishnu Mohan:

- **Data Pre-processing: Data Integration**
  - Integrating datasets: Combining multiple datasets and ensuring its accuracy and functionality since multiple datasets are required for Olympics data analysis.
- **Descriptive analysis**
  - Main characteristics of the data is analysed. Measure of central tendency, dispersion etc. is calculated.
- **Trend analysis**
  - Gender analysis
  - Summer and winter Olympics analysis
  - Participation analysis

Niranjana J:

- **Data Pre-processing: Data Cleaning**
  - Dealing with duplicates: It includes identifying, addressing the duplicates and deciding on a strategy on how to deal with the duplicate records.
- **Inferential statistics**
  - Pattern analysis in samples of dataset. It uses statistical methods to draw conclusions about a larger population based on a sample of data.
- **Correlation analysis**
  - Age analysis of the athletes
  - Physique analysis
  - Age – Physique – Performance correlation analysis