**Olympic Athletes Data Analysis Project Report**

**Project Title:** Olympic Athletes Data Analysis  
**Student Name:** P.Kanagalakshmi  
**Course/Batch:** DataAnalysis  
**Date:**

**Table of Contents**

1. Introduction
2. Objectives
3. Dataset Description
4. System Design
   * 4.1 Python Data Analysis
   * 4.2 Visualization Techniques
5. Implementation

* 5.1 Data Cleaning & Preprocessing
* 5.2 Analysis & Insights

1. Results
2. Conclusion
3. Future Enhancements

**1. Introduction**

This project analyzes **Olympic athletes’ data** to gain insights into trends over the years. Using Python libraries such as **pandas, NumPy, and matplotlib,** the project explores:

* Medal distribution among countries
* Gender participation trends
* Age distribution
* Sport-specific statistics

The analysis helps understand patterns in Olympic participation, performance, and demographics.

**2. Objectives**

* Analyze Olympic medal distribution by country.
* Examine gender participation trends over time.
* Study the age distribution of athletes.
* Compute average height and weight by sport.
* Visualize athlete participation trends over the years.

**3. Dataset Description**

**Datasets used:**

**1. athlete\_events.csv** – Contains athlete details, Olympic events, results, and medals.

**2. noc\_regions.csv** – Maps National Olympic Committees (NOCs) to countries/regions.

**Key Columns:**

* **ID** – Athlete unique ID
* **Name** – Athlete name
* **Sex** – Gender
* **Age, Height, Weight** – Physical attributes
* **Team, NOC, Games, Year, Sport, Event** – Olympic info
* **Medal** – Medal won (Gold, Silver, Bronze)

**4. System Design**

**4.1 Python Data Analysis**

* **pandas**: Load, clean, merge, and group datasets.
* **NumPy:** Compute averages and statistics.
* **matplotlib:** Visualize trends and distributions with bar charts, line plots, and histograms.

**4.2 Visualization Techniques**

* Bar charts for top countries by medals
* Line charts for gender participation trends
* Histograms for age distribution
* Plots for athlete participation over time

**5. Implementation**

**5.1 Data Cleaning & Preprocessing**

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

athletes = pd.read\_csv("athlete\_events.csv")

regions = pd.read\_csv("noc\_regions.csv")

df = athletes.merge(regions, on="NOC", how="left")

* Handle missing values in Age, Medal, Height, Weight.
* Merge NOC regions for country-level insights.

**5.2 Analysis & Insights**

1. **Top 10 Countries by Total Medals**

medals = df.dropna(subset=["Medal"])

top\_countries = medals["region"].value\_counts().head(10)

top\_countries.plot(kind="bar", color="orange")

plt.title("Top 10 Countries by Total Medals")

plt.show()

1. **Male vs Female Participation Over Time**

gender\_trend = df.groupby(["Year", "Sex"])["ID"].nunique().unstack()

gender\_trend.plot(marker='o', figsize=(12,6))

plt.title("Male vs Female Participation Over Years")

plt.show()

1. **Age Distribution of Athletes**

df["Age"].dropna().hist(bins=30, edgecolor="black", color="skyblue")

plt.title("Distribution of Athlete Ages")

plt.show()

1. **Average Height & Weight by Sport**

sport\_stats = df.groupby("Sport")[["Height","Weight"]].mean().round(2)

print(sport\_stats.head(10))

1. **Athletes per Year Trend**

athletes\_per\_year = df.groupby("Year")["ID"].nunique()

plt.plot(athletes\_per\_year.index, athletes\_per\_year.values, marker='o', color='green')

plt.title("Number of Athletes per Year")

plt.show()

**6. Results**

* **Top countries** with maximum medals identified.
* **Gender trends** reveal increasing female participation.
* **Age distribution** indicates most athletes are in 20–30 years age group.
* **Sport-wise averages** provide insight into physical attributes.
* **Athlete participation** over the years shows growth and popularity trends.

**7. Future Enhancements**

* Include **interactive dashboards** using Plotly or Dash.
* Integrate with **real-time Olympic data APIs.**
* Predict medal winners using **machine learning models.**
* Provide **country-wise and sport-specific predictions.**

**8. Conclusion**

This project demonstrates the use of **Python data analysis and visualization** to extract meaningful insights from Olympic datasets. The analysis helps in understanding performance, participation trends, and demographics of athletes over time.