***A logo of a university

Description automatically generated*****School of Mechanical & Manufacturing Engineering (SMME)**

**National University of Science and Technology (NUST),**

**Sector H-12, Islamabad**

# *Analysis of 120 Years of Olympics Data*

Bazla Tariq

460547

**Fundamentals of Programming-II**

**Lab Engineer Laiba Waheed**

**November 20th, 2024**

# Project Report: Analysis of 120 Years of Olympics Data

# Project Title:

**"Olympic Games Data Analysis: 120 Years of Insights Visualized on a Website"**

# 1. Introduction

The Olympic Games, with a history spanning over 120 years, offer a rich dataset for analysis. This project uses Python and web technologies to explore trends, patterns, and insights from the Olympics data and displays the results interactively on a website. By using a combination of libraries for data preprocessing, visualization, and interactivity, the project aims to make complex data accessible and engaging for users.

# 2. Objectives

* Analyze trends in Olympic participation, performance, and demographics over 120 years.
* Visualize data to uncover patterns in medal tallies, athlete characteristics, and country-wise participation.
* Create an interactive and user-friendly web application for data exploration.

# 3. Dataset

The project used a dataset comprising Olympic Games data spanning 120 years. The dataset included information about:

* Athlete names, gender, and nationality.
* Sports and events participated in.
* Medal details (Gold, Silver, Bronze).
* Years and locations of Olympic events.

# 4. Tools and Libraries

The project utilized the following Python libraries and tools:

1. **Streamlit**: For building an interactive web application to display analysis results.
2. **Pandas**: For data manipulation and preprocessing tasks.
3. **Preprocessor and Helper Modules**: Custom-built modules for cleaning and preprocessing the dataset.
4. **Plotly Express**: For creating interactive and aesthetically appealing visualizations.
5. **Matplotlib**: For static plots and charts.
6. **Seaborn**: For advanced statistical visualizations.
7. **Plotly Figure Factory**: For creating specialized plots like dendrograms and distribution plots.
8. **Scipy**: For statistical analysis and calculations.

# 5. Methodology

## Data Preprocessing

1. **Data Cleaning:** Handled missing values, removed duplicates, and formatted data types.
2. **Feature Engineering**: Added new columns for analysis (e.g., medal counts).
3. **Data Transformation**: Converted categorical data to usable formats for visualization.

## Exploratory Data Analysis (EDA)

* Trends in the number of events, participants, and countries over time.
* Medal tally comparisons among nations.
* Gender-wise distribution of participants and medallists.
* Popular sports and events by year and country.

## Visualization and Insights

* **Plotly Express**: Used for dynamic charts like:
  + Line charts for participation trends.
* **Matplotlib and Seaborn**:
  + Heatmaps for correlation analysis.
  + Scatterplots to study height vs. weight trends.
* **Figure Factory**:
  + Distribution plots for comparing age distributions of medalists.

# 6. Key Features of the Web Application

1. **Interactive Dashboard**:
   * Filters for year, country, and sport.
   * Dropdowns for customizing visualizations.
2. **Data Visualizations**:
   * Gender-wise trends in participation.
   * Athlete performance analysis by age and event.
3. **Insights Section**:
   * Highlights significant findings from the data (e.g., trends in female participation).
4. **Custom Visualizations**:
   * Use of advanced statistical plots for unique perspectives.

**7. Results and Insights**

* **Trend Analysis**: A consistent increase in the number of participating countries and athletes over time, with notable spikes after significant global events (e.g., post-WWII and post-Cold War).
* **Gender Representation**: Steady growth in female participation, particularly after the introduction of gender equality measures.
* **Country Performance**: Dominance of countries like the USA and USSR in the medal tally over decades.
* **Athlete Characteristics**: Correlation between physical attributes like height and weight with performance in specific sports.

# 8. Challenges

* Handling missing or incomplete data, especially for early years.
* Ensuring visualizations were intuitive and informative for all users.
* Optimizing the Streamlit app for performance with large datasets.

# 9. Conclusion

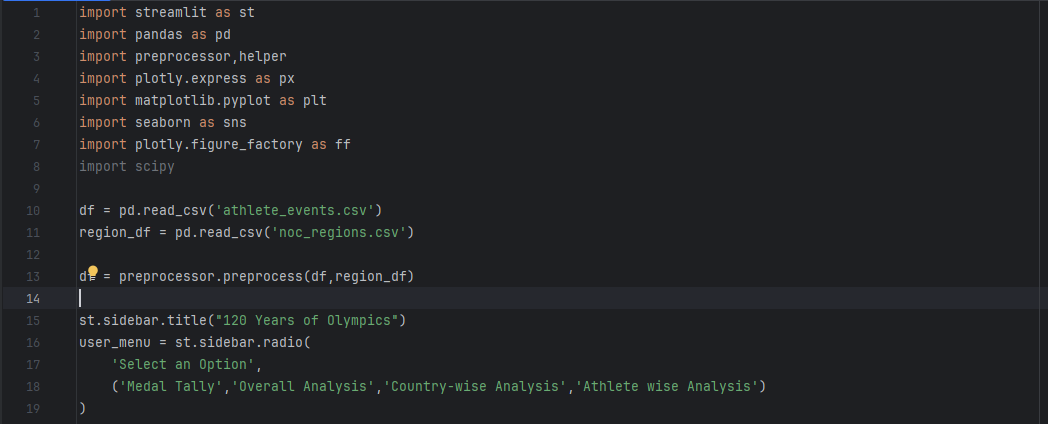
The project successfully analyzed and visualized 120 years of Olympic Games data, providing valuable insights into trends and patterns. By deploying the analysis on a user-friendly web application, the project achieved its goal of making data exploration accessible to a broader audience.

# 10. Future Work

* Expand the dataset to include Winter Olympics data.
* Incorporate machine learning to predict medal outcomes based on athlete profiles.
* Add real-time updates for future Olympic Games.

# 11. Code (Snippets):

## Main Code:



A computer screen shot of a program code

Description automatically generated

A screen shot of a computer

Description automatically generated

A screen shot of a computer code

Description automatically generated

A screen shot of a computer program

Description automatically generated

A computer screen shot of a program code

Description automatically generated

A computer screen shot of a program code

Description automatically generated

A screen shot of a computer code

Description automatically generated

A computer screen shot of text

Description automatically generated

## Helping Codes:

**a) helper.py:**

**A screen shot of a computer program

Description automatically generated**

**A computer screen shot of a program code

Description automatically generated**

****

**A computer screen with many colorful text

Description automatically generated**

**b) preprocessor.py:**

****