# Predicting and Analyzing Medal Performance Improvement Across Countries

AISHWARYA.S.M | Tools used (Python,Pandas,EDA, Scikit-learn,Power BI)

# Problem statement

The objective of this project is to identify countries that have shown improvement in their Olympic medal performance over time and to predict future medal outcomes using historical Olympic data. By analyzing long-term trends and patterns, the project aims to uncover insights into national performance growth and support data-driven forecasting.

## Dataset Overview

The dataset contains **2,144 rows** and **11 columns**, covering Olympic Games from **1964 to 2016**. It includes country-level historical performance metrics such as participation details and medal counts across multiple Olympic years.

## Target Variable

The target variable is **Medals**, which represents the total number of medals won by each country in a given Olympic year. This variable is used to measure performance trends over time and serves as the prediction target for the machine learning models.

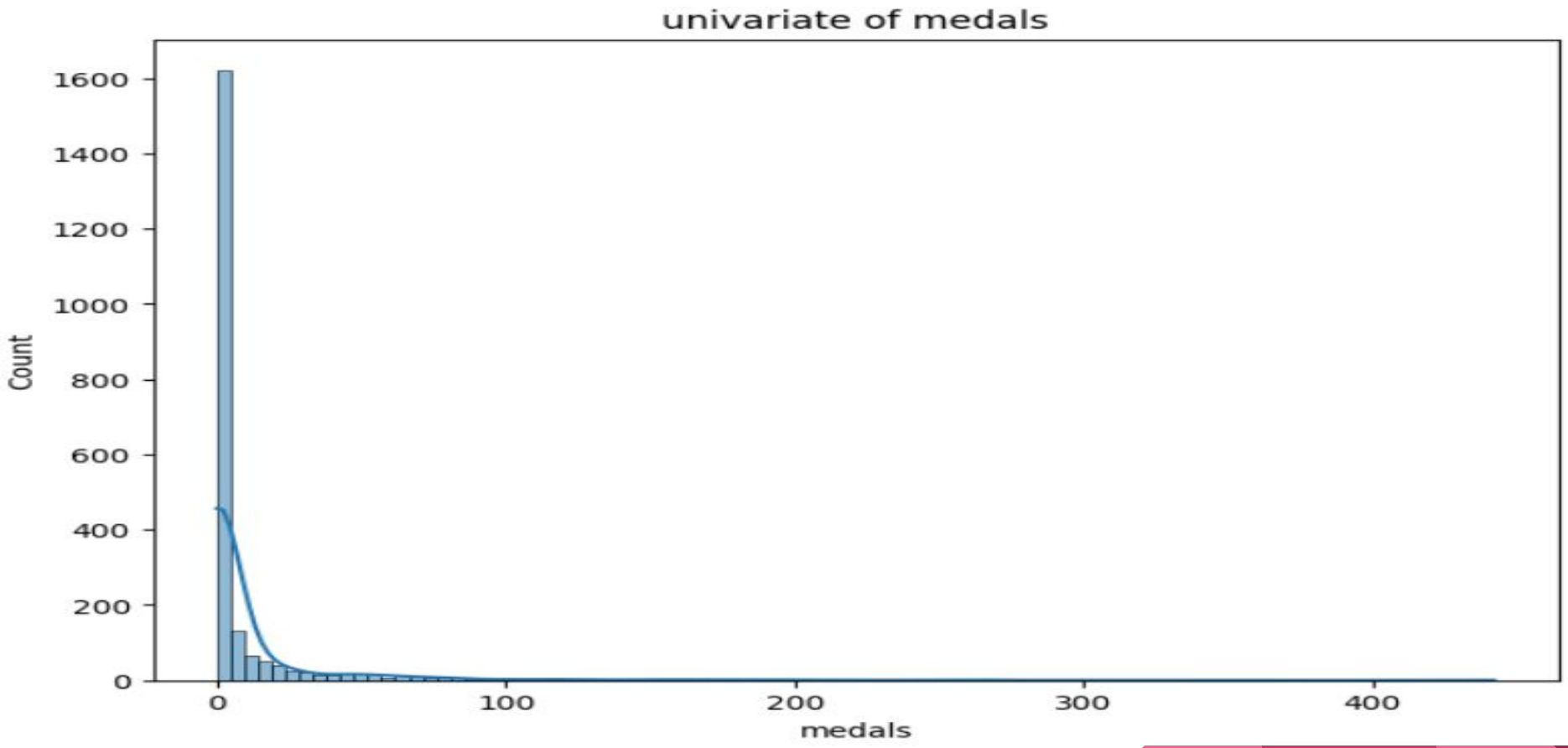
# Data Preprocessing

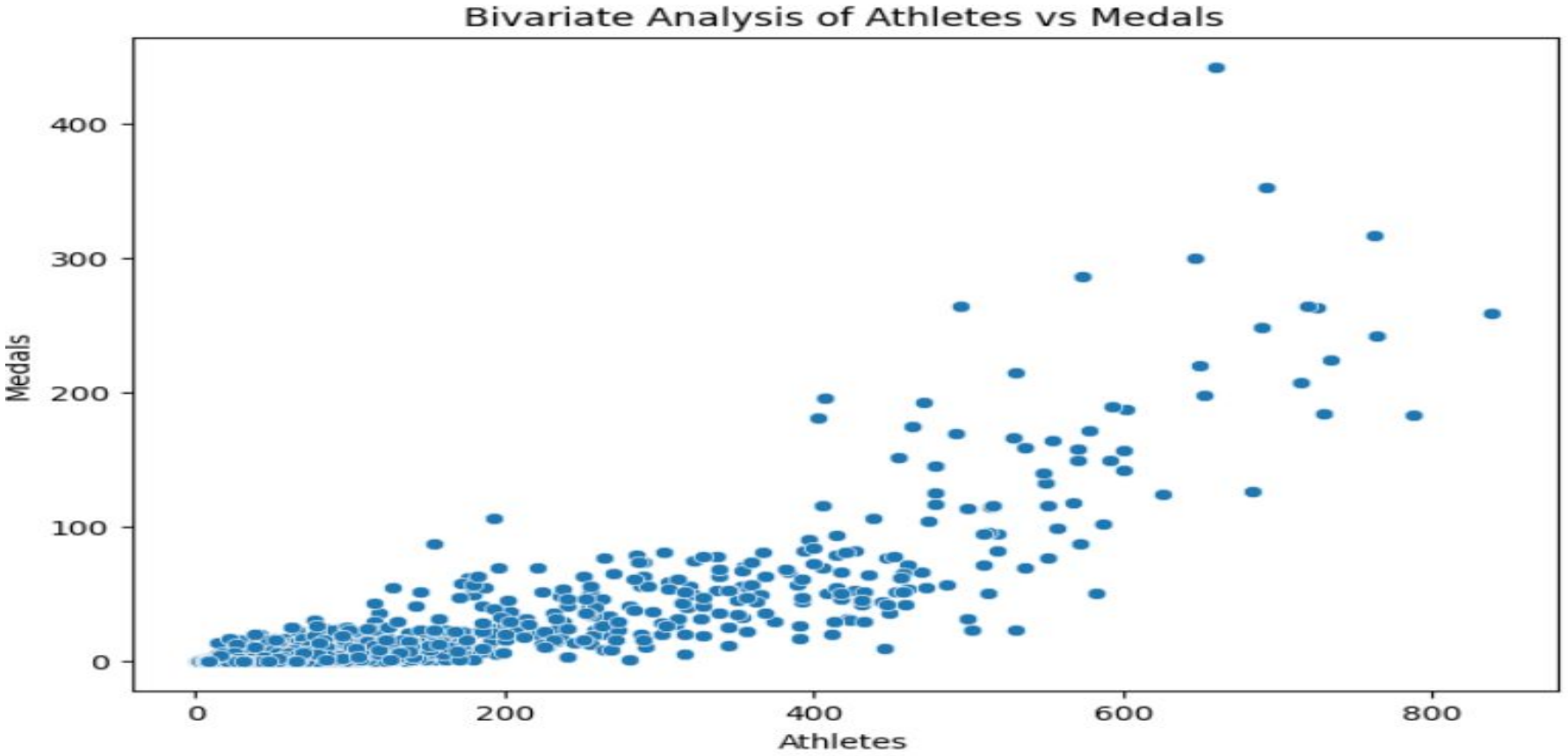
- The dataset was first examined for data quality issues by identifying **null values** and **duplicate records**.
- Appropriate data cleaning techniques were applied to handle missing values and remove duplicates to ensure data consistency and reliability.
- Outlier detection was performed using **descriptive statistical measures** such as **mean, median, and mode**, allowing for the identification of abnormal values that could potentially impact model performance.
- Encoding categorical variables

# Exploratory Data Analysis

Exploratory Data Analysis (EDA) was then conducted to understand the underlying patterns and relationships within the data. This included:

- **Univariate analysis** to examine the distribution of individual variables
- **Bivariate analysis** to explore relationships between pairs of variables
- **Multivariate analysis** to analyze interactions among multiple features and their impact on medal outcomes


univariate of medals
Count
1600
1400
1200
1000
800
600
400
200
0
0
100
200
300
400
medals


Bivariate Analysis of Athletes vs Medals
Medals
Athletes
0
100
200
300
400
0
200
400
600
800

# Feature Engineering

Before training the models, feature engineering techniques were applied to enhance the predictive power of the dataset. Relevant features were selected and transformed to better represent historical Olympic performance patterns.

New features were derived from existing variables to capture trends such as **country-wise performance over time**, **previous medal counts**, and **participation consistency**. Categorical variables were encoded into numerical formats, and numerical features were scaled where necessary to ensure uniformity across features.

These transformations helped reduce noise, improved model interpretability, and contributed to better regression performance.

# Model Training

Since the target variable **Medals** is a **continuous numerical value**, a **regression-based approach** was chosen for model training. Regression models are suitable for predicting the total number of medals won by a country based on historical and engineered features.

The dataset was divided into **training and testing sets** to evaluate generalization performance. Multiple regression algorithms were trained to capture both linear and non-linear relationships in the data.

Model performance was assessed using standard **regression evaluation metrics**, including **Mean Absolute Error (MAE)**, **Mean Squared Error (MSE)**, and **$R^2$ score**, to ensure accurate and reliable predictions.

# Metrics

| Model | $R^2$ Score | MAE | RMSE |
|---|---|---|---|
| Linear Regression | 0.9203 | 4.19 | 9.35 |
| Random Forest Regressor | 0.9237 | 3.67 | 9.15 |
| Decision tree Regressor | 0.7913 | 5.20 | 15.13 |
| Gradient Boosting  Regressor | 0.9135 | 3.75 | 9.74 |
| AdaBoost Regressor | 0.7215 | 3.75 | 9.74 |
| K-Nearest Neighbour [KNN] | 0.9018 | 4.01 | 10.38 |
| Supportive Vector Machine[SVM] | 0.2318 | 836.52 | 28.92 |

# Best Model

After evaluating multiple regression models including Linear Regression, Decision Tree, SVR, and Random Forest, the Random Forest Regressor achieved the best performance. It produced the highest $R^2$ score and the lowest MAE and RMSE values. This indicates strong predictive accuracy and good generalization on unseen data. Therefore, Random Forest was selected as the final model for predicting medal performance.

# Insights from powerbi

Medal trend over time

The line chart shows country-wise medal trends from the 1964 to 2016 Olympics. Each colored line represents a different country. Some countries display strong upward spikes, indicating periods of high medal performance, while others show lower or fluctuating trends, reflecting fewer medals won. Countries with consistently higher peaks demonstrate sustained Olympic success, whereas countries with lower or irregular spikes indicate limited or inconsistent performance over time. No medal trends are observed after 2016, as the dataset ends at that year

Distribution of medal growth across countries

The clustered bar chart shows countries ranked by medal growth. Countries at the top have significantly improved their medal counts compared to previous Olympics, indicating strong long-term sports development and increased participation.

## Events, Medals, Athletes by Country

The scatter plot shows a positive relationship between events participated and medals won, with each point representing a country. Countries like the **United States** stand out with the highest number of events, athletes, and medals, while most countries have lower participation and medal counts.

## Count of Country by Medal growth

The pie chart shows the distribution of countries based on their Medal Growth. Over half (57%) of the countries have zero medal growth, while smaller portions show positive or negative growth values, indicating varying trends in medal performance.

# Summary

When we look at the overall participation in the Olympics (scatter plot), we see a clear divide: a few countries like the United States lead with the highest number of events, athletes, and medals, while many others participate less and win fewer medals.

Over time (line chart), some countries show strong upward medal trends, reflecting sustained investment and success, whereas others have more variable results.

By examining medal growth (pie/bar chart), we find that a majority of countries show little to no growth, but a select group is rapidly improving their medal counts, signaling rising sports development.

These insights help us build predictive models that can forecast medal outcomes, enabling targeted strategies for countries to enhance their Olympic performance.