DATA SCIENCE

PROJECT REPORT

(Project Semester January-April 2025)

Exploratory Data Analysis on Olympic Dataset

Submitted by

Priyanshu Raj Chauhan

Registration No: 12322487

Programme and Section: DATA SCIENCE TOOLBOX: PYTHON PROGRAMMING KM005

Course Code: INT375

Under the Guidance of

Maneet Kaur

Discipline of CSE/IT

Lovely School of Computer Science and Engineering

Lovely Professional University, Phagwara

DECLARATION

I, Priyanshu Raj Chauhan, student of DATA SCIENCE TOOLBOX: PYTHON PROGRAMMING under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: 2025-04-11

Signature Priyanshu Raj Chauhan
Registration No. 12325159

Name of the student: Priyanshu Raj Chauhan

CERTIFICATE

This is to certify that Mr. Priyanshu Raj Chauhan bearing Registration No. 12322487 has completed INT375 project titled, “Exploratory Data Analysis on Olympic Dataset” under my guidance and supervision. To the best of my knowledge, the present work is the result of his original development, effort and study.

Signature and Name of the Supervisor
Maneet Kaur

Designation of the Supervisor
School of Computer Science and Engineering
Lovely Professional University
Phagwara, Punjab.

Date: 2025-04-11

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to my mentor Ms. Maneet Kaur for her valuable guidance throughout this project. I also thank Lovely Professional University for providing the necessary resources and support.

TABLE OF CONTENTS

Description

This project focuses on analyzing Olympic athlete data using Python and core data science techniques. The primary objective was to explore the dataset to understand trends and patterns across various dimensions such as medal distribution, athlete demographics, sports performance, and country participation. By performing a thorough **Exploratory Data Analysis (EDA)**, we aimed to uncover insights that could be valuable for understanding global athletic achievements and the factors influencing Olympic success.

We specifically chose this dataset because the Olympic Games are one of the largest and most significant global events, with a rich history and a diverse set of data that spans across many years and countries. This made it an ideal candidate for data analysis, with numerous angles to explore, including medal counts, athletes' physical attributes (age, height, weight), sports participation, and trends over time.

Challenges Faced:

Like any data analysis project, this one had its challenges. First, we had to address **missing values** in the dataset, especially in columns like Age, Height, and Weight. Deciding on how to handle those missing values took considerable effort to ensure the data remained useful for analysis. Another challenge was analyzing data across multiple years and sports, which required proper handling of time-related fields and understanding the context of how different sports and countries performed over the decades.

Choosing the **right visualizations** was another hurdle. Representing trends in Olympic performance effectively through graphs like **medal distribution, top-performing countries,** or **sport-wise medal counts** took some experimentation. Furthermore, performing a **simple regression analysis** to predict medal outcomes was difficult due to limited numerical features in the dataset.

Lastly, ensuring that the findings were clearly **presented** was an important task. Proper labeling, choosing appropriate chart types, and making the visualizations intuitive for a broader audience was crucial to making the results meaningful and easy to interpret.

1. Introduction

In today's data-driven world, understanding raw data before jumping into conclusions is really important. That's exactly what Exploratory Data Analysis (EDA) helps us do — it allows us to explore datasets, spot patterns, find relationships, and make sense of things that aren't always obvious at first glance. For this project, I focused on a crime dataset from a district, aiming to uncover how and when crimes happen, what kind of offenses are most common, and whether things like time of day or location have any effect.

Using Python and some of its powerful data science libraries, I went through the dataset step by step — cleaned it up, explored the features, and visualized the findings. The idea was to get a clearer picture of the crime trends and figure out how different aspects like method of crime, reporting shifts, and types of offenses are connected. This project gave me a great hands-on experience in working with real-world data and helped me understand the practical side of data analysis

2. Source of Dataset

The dataset used in this project contains detailed records of all athletes who participated in the Summer Olympics from 1896 to 2016. It includes vital information such as the athlete's name, age, height, weight, country (NOC), sport, and medal achieved (if any). Additionally, it provides historical data like the year of the event and the athlete's performance in each Olympic Games.

This dataset was provided in **CSV** format, making it easy to import into Python for further analysis. Using **pandas** and **seaborn**, I was able to clean the dataset, explore its features, and visualize the results. It offered a rich source of data for analyzing patterns in Olympic performance over the years.

3. EDA Process

Exploratory Data Analysis (EDA) is a crucial step in any data science project. It's where you get to know your data inside and out before jumping into more advanced analysis or predictive modeling. For this project, I followed a structured approach:

- **Loading the Dataset**: I imported the dataset using pandas.read_csv(), making it easy to load and work with.
- **Handling Missing Data**: The dataset contained some missing values in columns like Age, Height, Weight, and Medal. I handled this by:

    - Filling missing **numerical values** (like Age) with the median to ensure the data stayed representative.
    - For **categorical values** (like Medal), I replaced missing values with "No Medal", assuming that the missing data meant the athlete didn’t win a medal.

- **Understanding the Structure**: I used .info(), .shape, and .describe() to quickly assess the dataset's structure, number of rows and columns, and to gain insights into the distribution of numerical values.
- **Cleaning and Formatting**: Several columns, such as Year, Age, Height, and Weight, required some cleaning. I converted the Year column into a datetime format and made sure other fields, like Age and Height, were in their proper data types for easy manipulation.
- **Visualization**: To better understand trends and patterns in the data, I used **Seaborn** and **Matplotlib** to create various plots, such as:
    - Medal distribution by year
    - Top sports and countries by medal count
    - Age and height distributions of athletes
- **Regression Analysis**: I applied a simple linear regression model to analyze the relationship between **Age** and **MedalValue** (encoded as numeric values). This helped me understand how athlete age might influence the likelihood of winning medals in the Olympics.

4. Analysis on Dataset

i. General Description

This dataset is packed with valuable information, especially for understanding how Olympic athletes perform across different years, countries, and sports. It includes key details like athlete demographics (age, height, weight), event information (medal type, sport, year), and country-specific data (NOC). The goal was to explore this data to uncover trends — for example, identifying if certain countries perform better in specific sports, or how athlete characteristics like age and height might influence their chances of winning a medal.

Through this analysis, we aimed to understand the dynamics of Olympic performance, and to identify patterns in medal distribution, country participation, and athlete profiles over time.

ii. Specific Requirements, Functions and Formulas

To carry out the analysis of the Olympic dataset, I used several core Python libraries, each serving a specific purpose:

- **pandas** and **numpy** for data manipulation and handling
- **seaborn** and **matplotlib** for data visualization
- **scikit-learn** for simple regression modeling

Some important functions and techniques I used in this analysis include:

- .isnull().sum() to identify missing values across the dataset
- .fillna() to handle missing data by filling gaps with either the **mean** (for numerical values) or **mode** (for categorical variables)
- .value_counts() and .groupby() to get quick summaries of categorical variables, like counting medal types or grouping by country
- countplot(), histplot(), and heatmap() from **seaborn** to visualize the distribution of medals, athlete demographics, and correlations between variables
- LinearRegression() from **scikit-learn** to examine relationships between variables, such as the potential effect of an athlete's age or height on their chances of winning a medal

iii. Analysis Results

 Medal Distribution: The dataset revealed that certain countries consistently perform better than others, with the **USA**, **Soviet Union**, and **Germany** being top medal winners across different Olympic years.

 Sports Performance: Some sports, such as **Athletics**, **Swimming**, and **Gymnastics**, show a higher concentration of medals, whereas other sports have relatively fewer medals awarded, indicating a potential area for deeper exploration of how certain sports dominate the Olympic Games.

 Athlete Demographics: Athletes in the Olympics show a strong **age** correlation with medal success, where athletes typically fall between the ages of 20-30 years, suggesting that age plays a crucial role in performance.

 Country Participation: As expected, larger countries with more resources tend to have more athletes competing, and consequently, they have higher chances of winning medals. However, some smaller countries still manage to perform well, indicating the influence of sport specialization and athlete development programs.

 Correlations: There is a noticeable correlation between **athlete height** and **medal performance** in certain sports, with taller athletes excelling in sports like **Basketball** and **Volleyball**. The analysis also identified a **weak correlation** between **age** and **medal outcome**, where younger athletes (especially in events like swimming) tend to perform better.

iv. Visualizations

To better understand the data, I created several types of graphs:

- A **bar graph** to show which shifts had the most crimes
- A **histogram** for the distribution of latitude values
- A **pie chart** to visualize the proportion of each crime method
- A **boxplot** comparing longitude across different shifts
- A **heatmap** to check correlation between numeric values
- A **line graph** showing how crimes vary month by month
- A **regression plot** to visualize the relationship between latitude and longitude

These visuals made the analysis more intuitive and easier to explain.

5. Conclusion

Working on this project gave me a comprehensive experience in handling raw, unstructured data. Starting with a large and complex Olympic dataset, I went through the process of cleaning, transforming, and analyzing it to uncover meaningful insights. By utilizing visualizations and basic regression techniques, I was able to identify interesting patterns, such as which countries consistently perform better, which sports dominate medal counts, and how athlete demographics (like age and height) correlate with medal success.

The insights gained, such as understanding the relationship between athlete attributes and Olympic performance, not only deepened my understanding of sports analytics but also provided real-world relevance. This project allowed me to refine my skills in **Python**, **data wrangling**, **visualization**, and **regression modeling**, while also offering a deeper appreciation for how data science can uncover hidden trends in complex datasets. Ultimately, it was a rewarding experience that enhanced my analytical skills and gave me a better understanding of Olympic history and athlete performance.

6. Future Scope

There's significant potential to expand on this analysis and dive deeper into the Olympic dataset. Some ideas for future work include:

- **Predictive Models**: Building machine learning models to predict future Olympic medal distributions based on factors like country, athlete demographics, and historical performance.
- **Clustering Techniques**: Using clustering algorithms (like K-Means) to identify groups of countries or sports with similar performance patterns, which could help inform strategies for athlete development and training.
- **Advanced Visualizations**: Adding interactive **visualizations** using libraries like **Plotly** or **Dash** to allow users to explore trends across different years, sports, or countries in more detail.
- **Athlete Profiling**: Developing a model to classify the likelihood of an athlete winning a medal based on attributes such as age, height, weight, and sport, allowing for more personalized predictions.
- **Deep Dive into Specific Sports**: Analyzing specific sports in more detail (e.g., **track and field**, **swimming**) to understand what factors, such as training regimens, technology, and country-specific resources, influence success in those fields.

7. References

[1] J. VanderPlas, *Python Data Science Handbook*, O'Reilly Media, 2016.
[2] NumPy documentation: https://numpy.org/doc/
[3] Matplotlib Gallery: https://matplotlib.org/stable/gallery/index.html
[4] Towards Data Science – Exploratory Data Analysis: https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15

[5] Scikit-learn Regression Models Guide: https://scikit-learn.org/stable/supervised_learning.html