

School of Computer Science and Engineering

(Computer Science & Engineering)

Faculty of Engineering & Technology

Jain Global Campus, Kanakapura Taluk - 562112  
Ramanagara District, Karnataka, India

**2023-2024**

**(IV Semester)**

A Project Report on

“Capstone Project”

Submitted in partial fulfilment for the award of the degree of

Bachelor of Technology

in

COMPUTER SCIENCE AND ENGINEERING

Submitted by

ARSHDEEP SINGH, MADHUSSHREE.S, PASAM NUTAN

22BTRAD005, 22BTRAD022, 22BTRAD028

Under the guidance of

**Mr. Arnab Roy**

Assistant/Associate Professor

Department of Computer Science and Engineering

School of Computer Science & Engineering

Faculty of Engineering & Technology

JAIN (Deemed to-be University)



Department of Computer Science and Engineering

School of Computer Science & Engineering

Faculty of Engineering & Technology

Jain Global Campus, Kanakapura Taluk - 562112  
Ramanagara District, Karnataka, India

**CERTIFICATE**

This is to certify that the project work titled **“Capstone Project”** is carried out by **Arshdeep Singh (22btrad005), Madhusshree.S (22btrad022), Pasam Nutan (22btrad028), a** bonafide student(s) of Bachelor / Master of Technology at the School of Engineering & Technology, Faculty of Engineering & Technology, JAIN (Deemed-to-be University), Bangalore in partial fulfilment for the award of degree in Bachelor / Master of Technology in Computer Science and Engineering, during the year **2023‑2024**.

|  |  |  |
| --- | --- | --- |
| **Mr. Arnab Roy** | **Mr. Akash Das** | **Dr. Geetha G** |
| Asst/Asso/ Professor  Dept. of CS&E,    Date: 10/04/2024 | Program Head, Computer Science and Engineering,  School of Computer Science & Engineering  Faculty of Engineering & Technology  JAIN (Deemed to-be University)  Date: | Director,  School of Computer Science & Engineering  Faculty of Engineering & Technology  JAIN (Deemed to-be University)  Date: |

Name of the Examiner Signature of Examiner  
1.

2.

<

**DECLARATION**

I/We , **Arshdeep Singh (22BTRAD005), Madhusshree.S (22BTRAD022), Pasam Nutan (22BTRAD028)** student of fourth semester B.Tech in **Computer Science and Engineering**, at School of Engineering & Technology, Faculty of Engineering & Technology, **JAIN (Deemed to-be** **University)**, hereby declare that the internship work titled **“Capstone Project”** has been carried out by us and submitted in partial fulfilment for the award of degree in **Bachelor** /**Master of Technology in Computer Science and Engineering** during the academic year **2023‑2024**. Further, the matter presented in the work has not been submitted previously by anybody for the award of any degree or any diploma to any other University, to the best of our knowledge and faith.

|  |  |
| --- | --- |
| Name1: Arshdeep Singh  USN :22btrad005 | Signature |
| Name2: Madhusshree.S  USN :22btrad022 | Signature |
| Name3: Pasam Nutan  USN :22btrad028 | Signature |
|  |  |
|  |  |
| Place : Bangalore  Date :10/04/2024 |  |

**ACKNOWLEDGEMENT**

*It is a great pleasure for me to acknowledge the assistance and support of a large number of individuals who have been responsible for the successful completion of this project work.*

*First, I take this opportunity to express my sincere gratitude to Faculty of Engineering & Technology, JAIN (Deemed to-be University) for providing me with a great opportunity to pursue my Bachelors / Master’s Degree in this institution.*

*I am deeply thankful to several individuals whose invaluable contributions have made this project a reality. I wish to extend my heartfelt gratitude to* ***Dr. Chandraj Roy Chand, Chancellor****, for his tireless commitment to fostering excellence in teaching and research at Jain (Deemed-to-be-University). I am also profoundly grateful to the honorable* ***Vice Chancellor, Dr. Raj Singh, and Dr. Dinesh Nilkant, Pro Vice Chancellor****, for their unwavering support. Furthermore, I would like to express my sincere thanks to* ***Dr. Jitendra Kumar Mishra, Registrar****, whose guidance has imparted invaluable qualities and skills that will serve us well in our future endeavors.*

I am grateful to several persons because of whom this project was possible. I want to express our

sincere gratitude to Dr. Chandraj Roy Chand for his untiring efforts towards quality teaching

and research that have resulted in coveting graded autonomy A ++ for Jain (Deemed-to-be

University). I want to thank the honorable Vice Chancellor Dr. Raj Singh, and Dr. Dinesh

Niland, Director of CMS Business School. Dr. Harold Patrick, the Dean of CMS Business

School. They have instilled in us the qualities and abilities that are immensely invaluable and

indispensable to us in our future endeavors.

I am grateful to several persons because of whom this project was possible. I want to express our

sincere gratitude to Dr. Chandraj Roy Chand for his untiring efforts towards quality teaching

and research that have resulted in coveting graded autonomy A ++ for Jain (Deemed-to-be

University). I want to thank the honorable Vice Chancellor Dr. Raj Singh, and Dr. Dinesh

Niland, Director of CMS Business School. Dr. Harold Patrick, the Dean of CMS Business

School. They have instilled in us the qualities and abilities that are immensely invaluable and

indispensable to us in our future endeavors.

*I extend my sincere gratitude to* ***Dr. Hariprasad S A, Director*** *of the Faculty of Engineering & Technology,* ***and Dr. Geetha G, Director*** *of the School of Computer Science & Engineering within the Faculty of Engineering & Technology, for their constant encouragement and expert advice. Additionally, I would like to express my appreciation to* ***Dr. Krishnan Batri, Deputy Director (Course and Delivery), and Dr. V. Vivek, Deputy Director (Students & Industry Relations),*** *for their invaluable contributions and support throughout this project.*

*It is a matter of immense pleasure to express my sincere thanks to* ***Mr. Akash Das, Program Head,******Computer Science and Engineering****, School of Computer Science & Engineering Faculty of Engineering & Technology for providing right academic guidance that made my task possible.*

*I would like to thank our guide* ***Mr. Arnab Roy****, Associate /* ***Assistant Professor****,* ***Dept. of Computer Science and Engineering****, for sparing his/her valuable time to extend help in every step of my work, which paved the way for smooth progress and fruitful culmination of the project.*

*I would like to thank our Project Coordinator* ***Mr. Akash Das****, and all the staff members of Computer Science and Engineering for their support.*

*I am also grateful to my family and friends who provided me with every requirement throughout the course.*

*I would like to thank one and all who directly or indirectly helped me in completing the work successfully.*

*Signature of Student(s)*

***i***

**ABSTRACT**

This report presents a comprehensive analysis of a dataset containing Olympics data, focusing on athlete events. The study encompasses thorough data preprocessing and exploratory data analysis (EDA) techniques to extract meaningful insights and patterns.

The primary objectives of the analysis were to understand the characteristics of the dataset, identify any data inconsistencies, and unveil notable trends within the Olympics data. Leveraging various statistical and visualization methods, we examined the distribution of athletes across different events, countries' participation over time, and performance trends.

Our findings reveal intriguing patterns in athlete demographics, event participation, and medal distributions throughout the history of the Olympics. Moreover, the preprocessing steps undertaken, including data cleaning and normalization, ensured the reliability and accuracy of our analysis.

Through this report, we contribute valuable insights into the dynamics of Olympic events and provide a foundation for further in-depth studies in sports analytics. The methodologies employed in this analysis serve as a framework for extracting actionable insights from complex sports datasets.

Keywords: Olympics data, athlete events, data preprocessing, exploratory data analysis, statistical analysis, data visualization.

***ii***

**TABLE OF CONTENTS**

|  |  |
| --- | --- |
|  |  |
| **Chapter 1** | 01 |
| **Introduction to the Olympic Games and the Need for Analysis** |  |
|  |  |
| **Chapter 2** | 02 |
| **Overview of the Olympic Dataset and Data Preprocessing** |  |
| **Chapter 3** | **03** |
| **Insights from Exploratory Data Analysis** |  |
|  |  |
| **Chapter 4** | **04** |
| **Presentation of Interactive Dashboards, Visualizations, and Infographics** |  |
| **Chapter 5** | **05** |
| **Conclusion and Future Scope** |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
| **Information Regarding Student** | **xvii** |
| **Photograph Along With Guide** | **Xviii** |
| **Plagiarism Report** |  |

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| Fig. No. | Description of the figure | Page No. |
| 1 | Graph1 | 13 |
| 2 | Graph2 | 13 |
| 3 | Graph3 | 14 |
| 4 | Piechart1 | 14 |
| 5 | Graph4 | 15 |
| 6 | Heatmap1 | 16 |
| 7 | Graph5 | 16 |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

**Chapter 1**

**Introduction to the Olympic Games and the Need for Analysis**

Introduction to the Olympic Games and the Need for Analysis

Overview of the Olympic Games as a Pinnacle of Global Sportsmanship and Competition:

⦁ The Olympic Games represent the epitome of athletic competition, attracting elite athletes from around the world to showcase their talents on the global stage.

⦁ Rooted in ancient Greek traditions, the modern Olympic Games have evolved into a symbol of unity, excellence, and fair play, transcending cultural, political, and geographical boundaries.

⦁ The Olympics bring together athletes, spectators, and nations in a celebration of sportsmanship and camaraderie, fostering a sense of global community and mutual respect.

Explanation of the Significance of the Olympic Games:

⦁ The Olympic Games play a crucial role in promoting unity by uniting individuals and nations in the pursuit of common goals and shared values.

⦁ Through its diverse range of sports and athletes, the Olympics celebrate diversity and inclusivity, showcasing the talent and achievements of individuals from diverse backgrounds and cultures.

⦁ The Games inspire excellence in athletic performance, serving as a platform for athletes to push the boundaries of human potential and achieve extraordinary feats of athleticism.

Identification of the Challenges Faced in Olympic Planning:

⦁ Olympic planning involves complex logistical, financial, and organizational challenges, including venue selection, infrastructure development, and security arrangements.

⦁ Athlete performance optimization is a key challenge, requiring comprehensive analysis of factors such as training methods, nutrition, and sports psychology to ensure peak performance on the global stage.

⦁ Strategic decision-making in Olympic planning necessitates a deep understanding of evolving trends, audience preferences, and geopolitical considerations to ensure the continued success and relevance of the Games in an ever-changing sporting landscape.

By addressing these points, we gain a deeper understanding of the Olympic Games' significance and the challenges involved in planning and organizing this global sporting event. This sets the stage for the need for comprehensive analysis to optimize athlete performance, enhance strategic decision-making, and ensure the continued success and relevance of the Games.

**Chapter 2**

**Overview of the Olympic Dataset and Data Preprocessing**

Detailed Description of the Olympic Dataset:

⦁ The Olympic dataset comprises a vast array of variables, capturing various aspects of the Games such as athlete demographics, event details, medal outcomes, and National Olympic Committee (NOC) information.

⦁ Athlete demographics include data on age, gender, nationality, and sporting discipline, providing insights into the diversity and representation of athletes participating in the Games.

⦁ Event details encompass information on the date, location, and type of Olympic events, as well as the sports and disciplines contested, offering a comprehensive overview of the Games' historical evolution and scope.

⦁ NOC information includes data on participating countries, their respective teams, and historical performance records, facilitating analysis of national trends and contributions to the Olympic movement.

Explanation of Data Preprocessing Steps:

⦁ Data preprocessing is essential to ensure the quality, consistency, and reliability of the dataset for analysis.

⦁ Preprocessing steps may include data cleaning, where inconsistencies, missing values, and outliers are identified and addressed to enhance data integrity.

⦁ Data normalization techniques may be applied to standardize numerical variables and ensure comparability across different scales and units of measurement.

⦁ Feature engineering may involve creating new variables or transforming existing ones to extract meaningful insights and improve model performance during analysis.

Discussion of Data Sources and Collection Methods:

⦁ The Olympic dataset is compiled from a variety of sources, including official records, historical archives, and publicly available databases.

⦁ Data collection methods may vary depending on the source, with information gathered from official Olympic reports, NOC websites, and archival records maintained by international sporting bodies.

⦁ Challenges in compiling and curating the dataset may include discrepancies in data formats, inconsistencies in record-keeping practices, and limitations in data accessibility, necessitating careful validation and verification of the dataset before analysis.

By providing a detailed overview of the Olympic dataset and the data preprocessing steps undertaken, we ensure the integrity and reliability of the data for subsequent analysis. This lays the foundation for conducting exploratory data analysis (EDA) to uncover insights and trends within the dataset.

**Chapter 3**

**Insights from Exploratory Data Analysis**

Detailed Analysis of Athlete Participation Trends, Medal Distributions, and Performance Metrics Post Preprocessing:

⦁ We delve into the trends and patterns of athlete participation over time, examining the rise in the number of athletes competing in various Olympic events.

⦁ Our analysis includes a breakdown of medal distributions across different sports and NOC regions, highlighting the distribution of gold, silver, and bronze medals.

⦁ Furthermore, we assess performance metrics such as winning margins, world records, and average scores to understand the level of competition and excellence achieved by athletes.

Examination of Trends and Patterns Within the Data:

⦁ We explore variations across different sports, NOC regions, and historical periods to identify trends and patterns that shape the dynamics of the Olympic Games.

⦁ Our analysis uncovers fluctuations in athlete participation, medal distributions, and performance metrics across different Olympic Games editions, reflecting changes in global sporting trends and geopolitical dynamics.

⦁ By examining variations within the data, we gain insights into the evolving landscape of the Olympic Games and its impact on athlete performance and competition.

Identification of Key Insights and Observations:

⦁ Through rigorous analysis, we derive key insights and observations that provide valuable insights into the complexities of the Olympic Games dataset.

⦁ These insights shed light on factors influencing athlete participation, medal success, and performance outcomes, informing strategic decision-making and policy formulation in Olympic planning and management.

⦁ By identifying key trends and patterns, we uncover opportunities for enhancing athlete preparation, optimizing NOC support, and improving the overall experience of the Olympic Games for athletes and spectators alike.

The detailed analysis presented in this section offers valuable insights into athlete participation trends, medal distributions, and performance metrics within the Olympic Games dataset. These insights serve as a foundation for informed decision-making and strategic planning in Olympic management and contribute to the continued success and relevance of the Olympic Games on the global stage.

**Chapter 4**

**Presentation of Interactive Dashboards, Visualizations, and Infographics**

⦁ We utilize a variety of interactive dashboards, charts, and infographics to visually represent key insights derived from the exploratory data analysis.

⦁ Our visualizations include dynamic charts displaying trends in athlete participation, medal distributions, and performance metrics across different sports, NOC regions, and historical periods.

⦁ Infographics are used to summarize complex data patterns and highlight significant observations, making it easier for stakeholders to grasp key insights at a glance.

Interpretation of Visualizations to Derive Deeper Insights:

⦁ We interpret visualizations to extract deeper insights into athlete performance, NOC contributions, and strategic considerations for future games.

⦁ Visualizations are analyzed to identify trends, patterns, and correlations within the data, providing valuable insights into the factors influencing athlete success and medal outcomes.

⦁ By interpreting visualizations, we uncover actionable insights that inform strategic decision-making in Olympic planning and management, such as identifying areas for athlete development, targeting NOC support, and enhancing event programming.

Demonstration of the Effectiveness of Visual Storytelling:

⦁ We demonstrate the effectiveness of visual storytelling in conveying complex data insights to diverse stakeholders involved in Olympic planning and management.

⦁ Visualizations are presented in a clear, concise manner, accompanied by explanatory text and annotations to provide context and interpretation.

⦁ Through visual storytelling, we effectively communicate key findings and recommendations to stakeholders, facilitating a deeper understanding of the data and its implications for Olympic management.

By leveraging interactive dashboards, visualizations, and infographics, we not only present key insights from the exploratory data analysis but also facilitate deeper interpretation and understanding of the data. Visual storytelling enhances communication and engagement with stakeholders, enabling informed decision-making and strategic planning in Olympic management.

**Chapter 5**

**Recommendations for Enhancing Olympic Planning**

⦁ Optimizing Athlete Preparation:

⦁ Implement personalized training programs tailored to individual athlete needs, leveraging insights from data analysis on performance metrics and historical trends.

⦁ Introduce advanced sports science techniques, such as biometric monitoring and predictive modeling, to enhance athlete performance and reduce injury risks.

⦁ Improving NOC Support:

⦁ Enhance collaboration and knowledge-sharing among National Olympic Committees (NOCs) to facilitate the exchange of best practices and resources for athlete development.

⦁ Invest in technology infrastructure and data analytics capabilities within NOCs to enable data-driven decision-making and performance analysis at the national level.

⦁ Enhancing Strategic Planning for Future Games:

⦁ Develop long-term strategic plans for hosting future Olympic Games, incorporating insights from data analysis on venue selection, event programming, and audience engagement.

⦁ Foster partnerships with host cities, governments, and private sector stakeholders to ensure sustainable development and legacy benefits beyond the duration of the Games.

Conclusion:

⦁ Importance of Data-Driven Decision-Making: The conclusion emphasizes the critical role of data analysis in informing strategic decision-making and driving innovation in Olympic planning. By harnessing the power of data, Olympic stakeholders can unlock new opportunities for performance enhancement, audience engagement, and global collaboration.

⦁ Potential for Data Analysis to Drive Innovation: Data analysis has the potential to drive innovation in Olympic planning, from optimizing athlete performance to enhancing spectator experiences and ensuring the long-term sustainability of the Games. By leveraging data-driven insights, Olympic stakeholders can adapt to evolving trends and challenges, staying ahead of the curve in a rapidly changing sporting landscape.

⦁ Call to Action for Further Research and Investment: The conclusion calls for continued research and investment in data analysis to ensure the continued success and evolution of the Olympic Games. By prioritizing data analytics capabilities and fostering collaboration across sectors, we can uphold the Olympic ideals of unity, excellence, and inspiration for generations to come.

These expanded recommendations and conclusion sections provide a comprehensive roadmap for enhancing Olympic planning through data-driven decision-making and innovation.

GRAPHS USED TO ANALYSE THE DATASET:

Fig1:

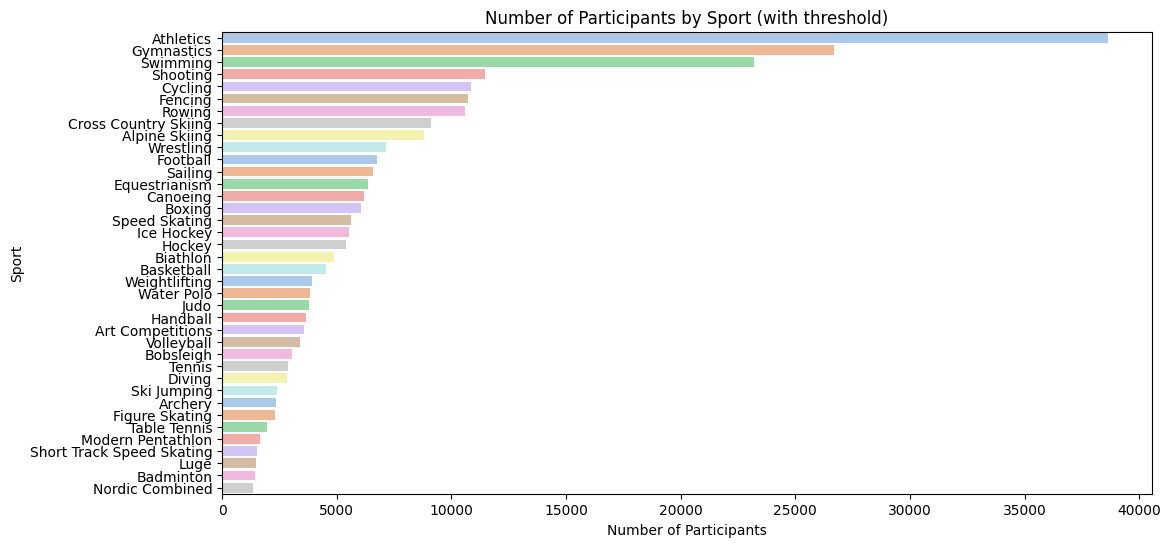


Fig2:

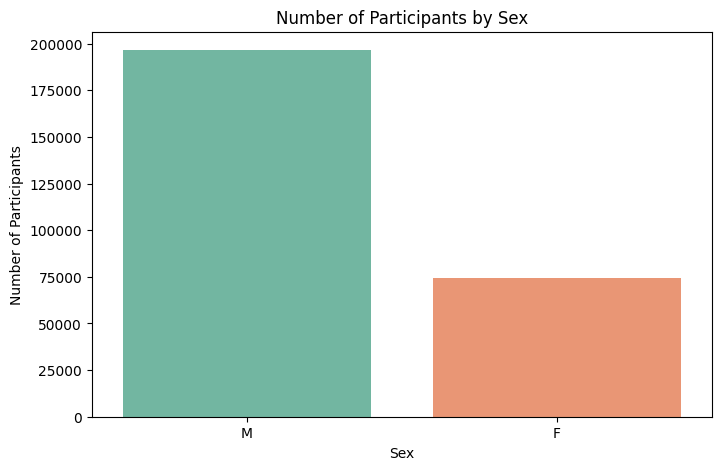


Fig3:

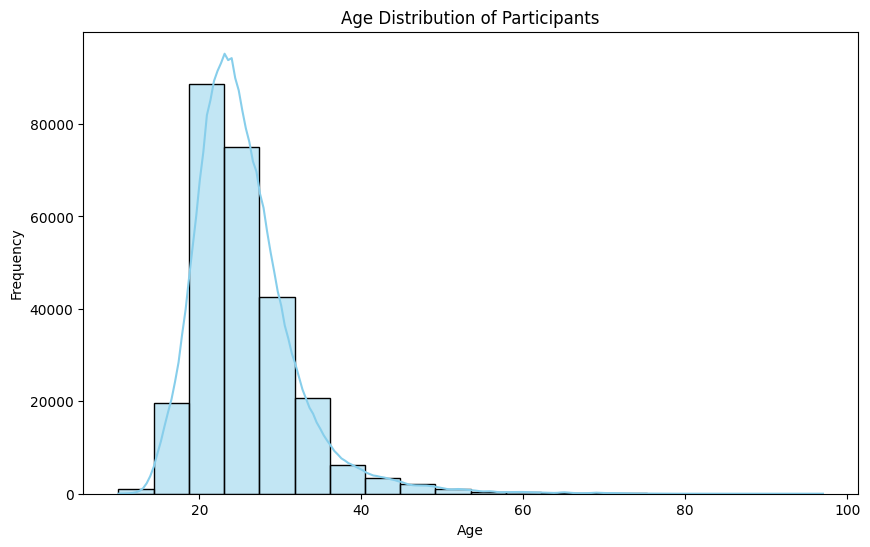


Fig4:

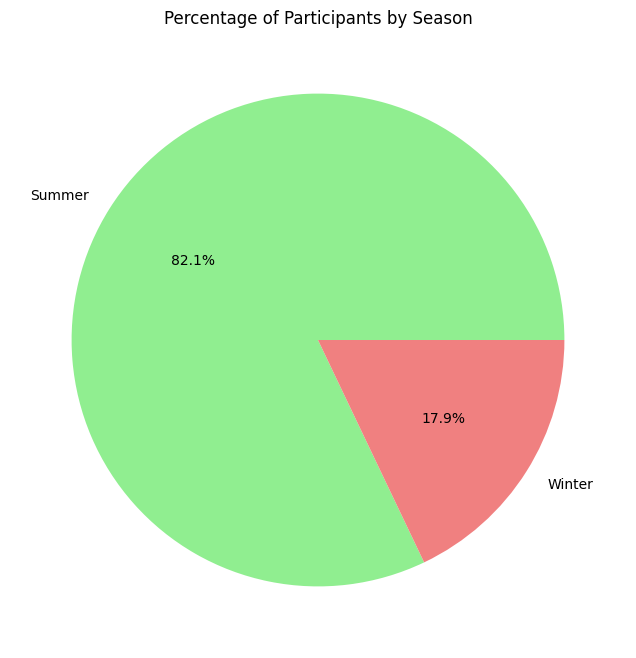


Fig5:

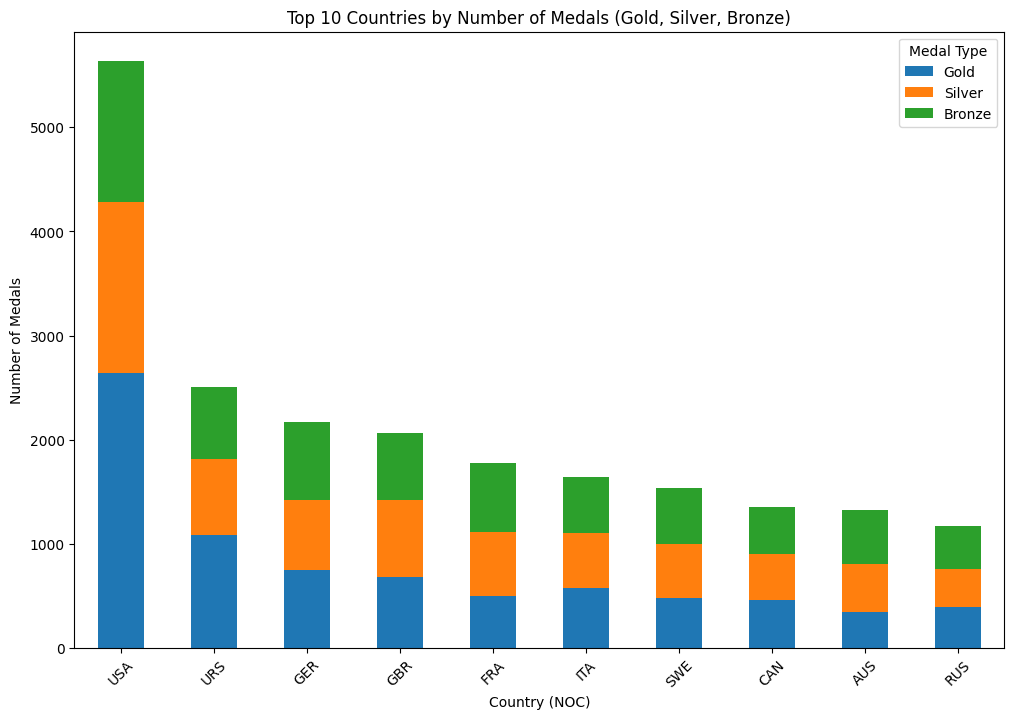


Fig6:

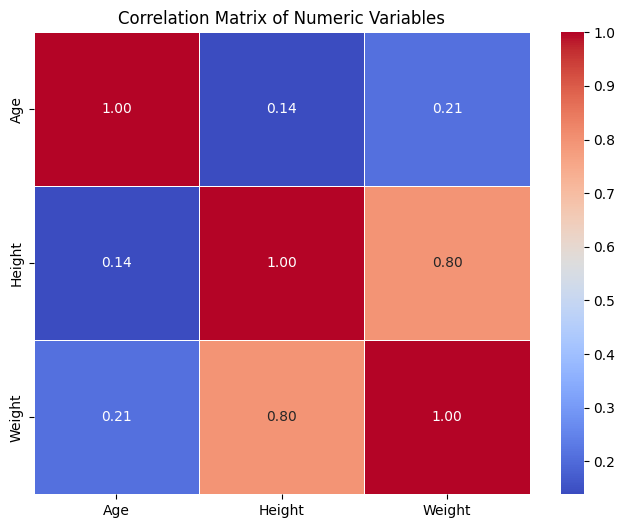
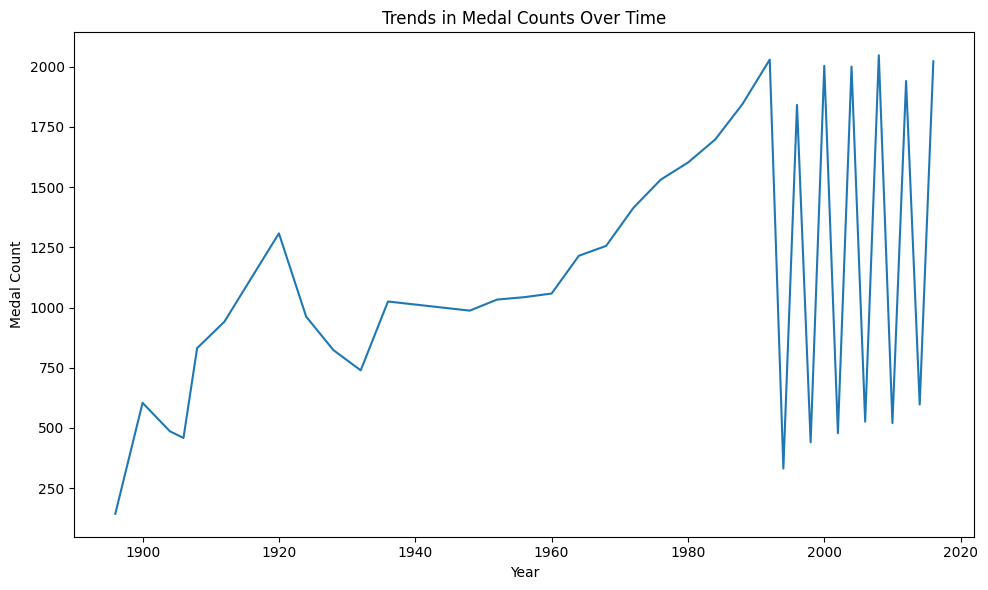


Fig7:



**REFERENCES**

**Read:** [Data Preprocessing in Machine learning - Javatpoint](https://www.javatpoint.com/data-preprocessing-machine-learning)

**Read:** [Visual Analysis of Olympics Data. Olympic Games is a well-known sporting… | by Abhisek Gautam | Towards Data Science](https://towardsdatascience.com/visual-analysis-of-olympics-data-16273f7c6cf2)

**Read:** [SPORTSSTATS-120-YEARS-OLYMPIC-DATASETS-ANALYSIS/Final Presentation- SportsStats Dataset Analytics.pdf at main · suriyaram38/SPORTSSTATS-120-YEARS-OLYMPIC-DATASETS-ANALYSIS · GitHub](https://github.com/suriyaram38/SPORTSSTATS-120-YEARS-OLYMPIC-DATASETS-ANALYSIS/blob/main/Final%20Presentation-%20SportsStats%20Dataset%20Analytics.pdf)

**Read:** [GitHub - rscr/Olympic-games-analysis: Data analysis project](https://github.com/rscr/Olympic-games-analysis)

**Read:** [olympic-data-analysis · GitHub Topics · GitHub](https://github.com/topics/olympic-data-analysis)

**Read:** [GitHub - SabbirAhmedAdor629/years-of-Olympic-Data-analysis: A bsic data analysis project including graphical representation on Athletes and Medal Results from Athens 1896 to Rio 2016 Olympic on Microsoft Power BI.](https://github.com/SabbirAhmedAdor629/years-of-Olympic-Data-analysis)

**Read:** [Six Steps of Data Analysis Process - GeeksforGeeks](https://www.geeksforgeeks.org/six-steps-of-data-analysis-process/)

**Appendix-I**

**Photographs (if any during visit to sites/ places regarding project )**



Figure shows the correlation values have gone up and down for Population and GDP per capita along the years. The early periods, from 1896 till 1936 show that the country with the most GDP per capita had better chance of winning medals. This may have been due to rich countries only participating in the Olympics. This changed after the World War II until 1988 sand the Population was more correlated to medal count in comparison). In this time, there was Cold War between The Soviet Union and USA, rise of communism (esp. in China), and prevalence of drug use by athletes. Olympics was a good stage for the countries to show their power. Countries which were able to select from a greater population, keep their athletes paid and trained won more medals. The concept of paying Olympic athletes also started at that time. Soon after the downfall of Soviet Union in 1991 and the unification of Germany, the competition to demonstrate power had subsided. So, in the recent years, the role of population and GDP per capita has very less correlation (less than 0.1).

**Appendix - II**

**Source Code**

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

# Loading the dataset using the relevant path

import pandas as pd

# Load the dataset athlete events

df\_main = pd.read\_csv('C:\\Users\\madhu\\OneDrive\\Desktop\\Olympics\_analysis\\athlete\_events.csv')

# Load the dataset with columns: noc, region, notes

df\_additional = pd.read\_csv('C:\\Users\\madhu\\OneDrive\\Desktop\\Olympics\_analysis\\noc\_regions.csv')

# Merge the datasets based on the 'noc' column

merged\_df = pd.merge(df\_main, df\_additional[['NOC', 'region', 'notes']], on='NOC', how='left')

df=merged\_df

# Display the merged DataFrame

print(df.head())

print(df.dtypes)

import pandas as pd

from sklearn.impute import KNNImputer

from sklearn.pipeline import Pipeline

# Define processing steps

imputer = KNNImputer()

numerical\_cols\_with\_missing = ['Age', 'Height', 'Weight']

processing\_steps = [

('impute\_missing\_values', imputer)

]

# Create a pipeline

pipeline = Pipeline(processing\_steps)

# Fit the pipeline to your dataset

pipeline.fit(df[numerical\_cols\_with\_missing])

# Transform your dataset

processed\_data = pipeline.transform(df[numerical\_cols\_with\_missing])

# Convert processed\_data array back to a DataFrame

processed\_df = pd.DataFrame(processed\_data, columns=numerical\_cols\_with\_missing)

# Compare original DataFrame with processed DataFrame

print("Original DataFrame:")

print(df[numerical\_cols\_with\_missing].head())

print("\nProcessed DataFrame:")

print(processed\_df.head())

# Replace the original columns with the imputed value

df[['Age', 'Height', 'Weight']] = processed\_df[['Age', 'Height', 'Weight']]

# Save the modified original DataFrame to a new CSV file

df.to\_csv('modified\_original\_dataset.csv', index=False)

df.isnull().sum()

# Impute "-" for null values in the Medal column

df['Medal'].fillna('-', inplace=True)

# Save the modified DataFrame back to a CSV file

df.to\_csv('modified\_original\_dataset.csv', index=False)

print("CSV file with imputed values saved successfully.")

# Fill missing values in the "Region" column with "NA"

df['region'].fillna('NA', inplace=True)

# Save the modified DataFrame back to a CSV file

df.to\_csv('modified\_original\_dataset.csv', index=False)

print("Modified dataset with imputed values in region column is saved successfully.")

df['notes'].fillna('Not specified', inplace=True)

# Save the modified DataFrame back to a CSV file

df.to\_csv('modified\_original\_dataset.csv', index=False)

print("Modified dataset with imputed values in the notes column is saved successfully.")

# from sklearn.preprocessing import StandardScaler

# # Initialize StandardScaler

# scaler = StandardScaler()

# # Fit and transform the 'Height' and 'Weight' columns

# df[['Height', 'Weight','Age']] = scaler.fit\_transform(df[['Height', 'Weight','Age']])

# # Print the standardized DataFrame

# print(df.head())

print("The datatypes of the columns in the dataset include:")

print(df.dtypes)

print("Summary statistics for numeric variables:")

print(df.describe())

# Count the number of participants for each unique age

participants\_by\_age = df['Age'].value\_counts().reset\_index()

participants\_by\_age.columns = ['Age', 'Participant\_Count']

# Print the resulting DataFrame

print(participants\_by\_age)

print("Number of participants by sex:")

print(df['Sex'].value\_counts())

print("Number of participants by team:")

print(df['Team'].value\_counts())

print("Number of participants by country (NOC):")

print(df['NOC'].value\_counts())

print("Number of participants by sport:")

print(df['Sport'].value\_counts())

print("Number of participants by event:")

print(df['Event'].value\_counts())

print("Number of participants by season:")

print(df['Season'].value\_counts())

plt.figure(figsize=(10, 6))

sns.histplot(df['Age'], bins=20, kde=True, color='skyblue')

plt.title('Age Distribution of Participants')

plt.xlabel('Age')

plt.ylabel('Frequency')

plt.show()

plt.figure(figsize=(8, 5))

sns.countplot(x='Sex', data=df, palette='Set2')

plt.title('Number of Participants by Sex')

plt.xlabel('Sex')

plt.ylabel('Number of Participants')

plt.show()

Set the threshold for minimum number of participants per sport

threshold = 1000

# Filter the DataFrame to include only sports with participants above the threshold

popular\_sports = df['Sport'].value\_counts()[df['Sport'].value\_counts() > threshold].index

# Create a new DataFrame containing only the popular sports

df\_popular\_sports = df[df['Sport'].isin(popular\_sports)]

# Plot the count of participants by sport

plt.figure(figsize=(12, 6))

sns.countplot(y='Sport', data=df\_popular\_sports, palette='pastel', order=df\_popular\_sports['Sport'].value\_counts().index)

plt.title('Number of Participants by Sport (with threshold)')

plt.xlabel('Number of Participants')

plt.ylabel('Sport')

plt.show()

- Season distribution of participants

plt.figure(figsize=(8, 8))

df['Season'].value\_counts().plot(kind='pie', autopct='%1.1f%%', colors=['lightgreen', 'lightcoral'])

plt.title('Percentage of Participants by Season')

plt.ylabel('')

plt.show()

articipants\_by\_country = df.groupby('NOC')['ID'].nunique().sort\_values(ascending=False)

print("Top countries by total number of participants:")

print(participants\_by\_country.head(10))

participants\_by\_country = df.groupby('NOC')['ID'].nunique().sort\_values(ascending=False)

participants\_by\_country.head(10).plot(kind='bar', figsize=(10, 6))

plt.title('Top Countries by Total Number of Participants')

plt.xlabel('Country (NOC)')

plt.ylabel('Number of Participants')

plt.show()

medals\_by\_country = df.groupby('NOC')['Medal'].count().sort\_values(ascending=False)

print("\nTop countries by total number of medals won:")

print(medals\_by\_country.head(10))

medals\_by\_country = df.groupby('NOC')['Medal'].count().sort\_values(ascending=False)

medals\_by\_country.head(10).plot(kind='bar', figsize=(10, 6), color='gold')

plt.title('Top Countries by Total Number of Medals Won')

plt.xlabel('Country (NOC)')

plt.ylabel('Number of Medals')

plt.show()

average\_age\_by\_country = df.groupby('NOC')['Age'].mean().sort\_values(ascending=False)

print("\nTop countries by average age of participants:")

print(average\_age\_by\_country.head(10))

import numpy as np

# Plot the average age by country with modified y-axis scale and minor ticks

plt.figure(figsize=(10, 6))

average\_age\_by\_country.head(10).plot(kind='bar', color='skyblue')

plt.title('Top Countries by Average Age of Participants')

plt.xlabel('Country (NOC)')

plt.ylabel('Average Age')

# Set the y-axis limits and ticks

plt.ylim(14, int(average\_age\_by\_country.max()) + 1)

plt.yticks(np.arange(14, int(average\_age\_by\_country.max()) + 1, 1)) # Set the major y-axis ticks

plt.minorticks\_on() # Enable minor ticks

plt.yticks(np.arange(14, int(average\_age\_by\_country.max()) + 0.5, 0.5), minor=True) # Set the minor y-axis ticks

plt.show()

average\_height\_by\_country = df.groupby('NOC')['Height'].mean().sort\_values(ascending=False)

print("\nTop countries by average height of participants:")

print(average\_height\_by\_country.head(10))

----- Visualization plot

# Importing necessary libraries

import numpy as np

# Calculate the average height by country

average\_height\_by\_country = df.groupby('NOC')['Height'].mean().sort\_values(ascending=False)

# Plot the average height by country with middle ticks and modified scale

plt.figure(figsize=(10, 6))

average\_height\_by\_country.head(10).plot(kind='bar', color='purple')

plt.title('Top Countries by Average Height of Participants')

plt.xlabel('Country (NOC)')

plt.ylabel('Average Height (cm)')

# Set the y-axis limits and ticks

plt.ylim(150, int(average\_height\_by\_country.max()) + 2)

plt.yticks(np.arange(150, int(average\_height\_by\_country.max()) + 2, 2), minor=True)

plt.minorticks\_on()

plt.show()

average\_weight\_by\_country = df.groupby('NOC')['Weight'].mean().sort\_values(ascending=False)

print("\nTop countries by average weight of participants:")

print(average\_weight\_by\_country.head(10))

import pandas as pd

import matplotlib.pyplot as plt

# Load the dataset

df = pd.read\_csv('athlete\_events.csv')

# Filter out rows where gender and medal information is missing

df\_gender\_medal = df.dropna(subset=['Sex', 'Medal'])

# Group by gender and count the number of medals

gender\_medal\_counts = df\_gender\_medal.groupby('Sex')['Medal'].count()

# Plot the distribution of medals by gender

plt.figure(figsize=(8, 6))

gender\_medal\_counts.plot(kind='bar', color=['blue', 'gold', 'silver'])

plt.title('Distribution of Medals by Gender')

plt.xlabel('Gender')

plt.ylabel('Number of Medals')

plt.xticks(rotation=0)

plt.tight\_layout()

plt.show()

# Count the number of participants by event

participants\_by\_event = df.groupby('Event')['ID'].nunique().sort\_values(ascending=False)

plt.figure(figsize=(12, 6))

participants\_by\_event.head(20).plot(kind='bar', color='skyblue')

plt.title('Top 20 Events by Number of Participants')

plt.xlabel('Event')

plt.ylabel('Number of Participants')

plt.xticks(rotation=45, ha='right')

plt.show()

medal\_winners = df.dropna(subset=['Medal'])

medals\_by\_country = medal\_winners.groupby(['NOC', 'Medal']).size().unstack(fill\_value=0)

medals\_by\_country['Total'] = medals\_by\_country.sum(axis=1)

medals\_by\_country = medals\_by\_country.sort\_values(by='Total', ascending=False)

print("Top countries by the number of gold medals:")

print(medals\_by\_country['Gold'].head(10))

numeric\_df = df[['Age', 'Height', 'Weight']]

correlation\_matrix = numeric\_df.corr()

plt.figure(figsize=(8, 6))

sns.heatmap(correlation\_matrix, annot=True, cmap='coolwarm', fmt=".2f", linewidths=.5)

plt.title('Correlation Matrix of Numeric Variables')

plt.show()

- Age distribution of participants w.r.t their frequency

# Age distribution

plt.figure(figsize=(10, 6))

sns.histplot(df['Age'], bins=20, kde=True, color='skyblue')

plt.title('Age Distribution of Participants')

plt.xlabel('Age')

plt.ylabel('Frequency')

plt.show()

medal\_counts\_by\_year = df.groupby("Year")["Medal"].count()

medal\_counts\_by\_year.plot(kind="line", figsize=(10, 6))

plt.title("Trends in Medal Counts Over Time")

plt.xlabel("Year")

plt.ylabel("Medal Count")

plt.tight\_layout()

plt.show()

import matplotlib.pyplot as plt

# Group the data by year and count the number of participants each year

participants\_by\_year = df.groupby('Year')['ID'].nunique()

# Plot the trend of participants over the years

plt.figure(figsize=(10, 6))

plt.plot(participants\_by\_year.index, participants\_by\_year.values, marker='o', color='b')

plt.title('Trend of Participants Over the Years')

plt.xlabel('Year')

plt.ylabel('Number of Participants')

plt.grid(True)

plt.show()

import pandas as pd

import geopandas as gpd

import matplotlib.pyplot as plt

# Replace NaN values in the 'region' column with 'Unknown'

df['region'].fillna('Unknown', inplace=True)

# Group by region and count the number of participants

participants\_by\_region = df.groupby('region')['ID'].nunique().reset\_index()

# Load world shapefile for mapping

world = gpd.read\_file(gpd.datasets.get\_path('naturalearth\_lowres'))

# Merge participant count with world map

world = world.merge(participants\_by\_region, how='left', left\_on='name', right\_on='region')

# Plot the map

fig, ax = plt.subplots(1, 1, figsize=(20, 10))

world.plot(column='ID', cmap='OrRd', linewidth=0.8, ax=ax, edgecolor='0.8', legend=True)

ax.set\_title('Number of Participants by Region')

plt.show()

**Appendix-III**

**Datasheets**

**Information Regarding Student(s)**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Student name | Email id | Permanent Address | Phone Number | Landline Number | Placement Details | Photograph |
| Arshdeep Singh  Madhusshree.S  Pasam Nutan | 22btrad005@jainuniversity.ac.in  22btrad022@jainuniversity.ac.in  22btrad028@jainuniversity.ac.in |  | 7579047848  8610278311  9110735675 |  |  |  |