INTERSHIP REPORT

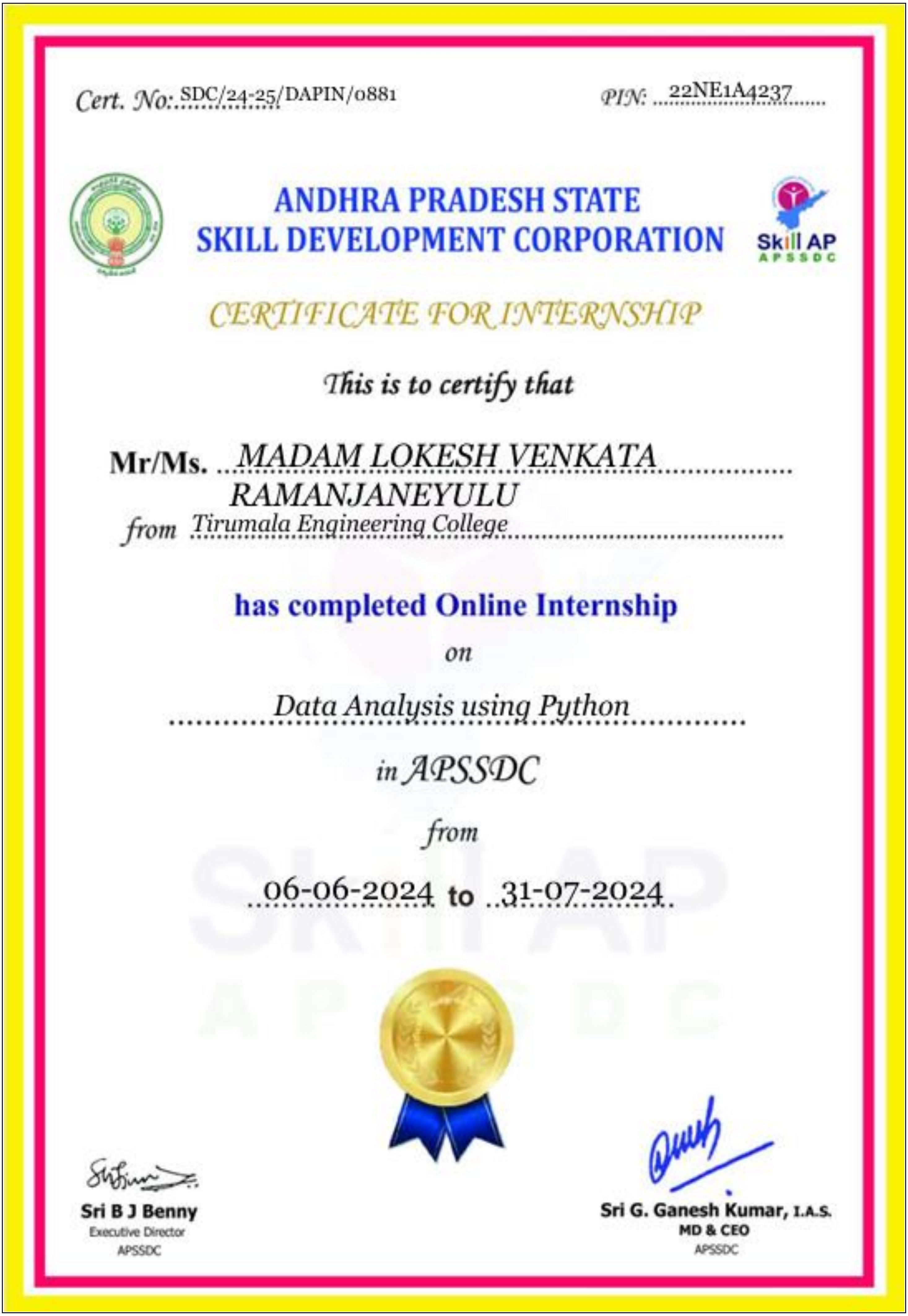
DATA ANALYSIS USING PYTHON

Skill AP

APSSDC

GOVERNMENT OF ANDHRA PRADESH

సత్యమేవ జయతే



Cert. No: SDC/24-25/DAPIN/0881

PIN: 22NE1A4237

ANDHRA PRADESH STATE SKILL DEVELOPMENT CORPORATION

CERTIFICATE FOR INTERNSHIP

This is to certify that

Mr/Ms. *MADAM LOKESH VENKATA RAMANJANEYULU*

from *Tirumala Engineering College*

has completed Online Internship

on

Data Analysis using Python

in APSSDC

from

06-06-2024 **to** 31-07-2024

Sri B J Benny
Executive Director
APSSDC

Sri G. Ganesh Kumar, I.A.S.
MD & CEO
APSSDC

INTERSHIP REPORT

A report submitted in partial fulfillment of the requirements for the Award of Degree of

BACHELOR OF TECHNOLOGY

IN

CSE-ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

BY

MADAM LOKESH VENKATA RAMANJANEYULU

Regd.No.:22NE1A4237

Under Supervision of Mr.B.J.Benny,Executive Director,APSSDC.

(Duration:6th June,2024 to 31st July, 2024)

DEPARTMENT OF CSE-ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

TIRUMALA ENGINEERING COLLEGE

(An Autonomous Institution)

Approved by AICTE, Permanently affiliated to JNTU, Kakinada

JONNALAGADDA, NARASARAOPET-522601

(YEAR:2022-2026)

TIRUMALA ENGINEERING COLLEGE

(AUTONOMOUS)

An ISO 9001:2015 Certified Institution, Accredited by NAAC (A+) & NBA

(Approved by AICTE, New Delhi & Affilliated to JNTUK, Kakinada)

Jonnalagadda, Narasaraopet, Guntur Dist. - 522601web : www.tecnrt.org Email : tecnrt@gmail.com

UGC
University Grants Commissicn

CERTIFICATE

This is to certify that the "Internship Report” submitted by MADAM.LOKESH VENKATA RAMANJANEYULU (Regd. No:22NE1A4237) is work done by him and submitted during 2024 – 2025 academic year, in partial fulfillment of the requirements for the award of the degree of BACHELOR OF TECHNOLOGY in

CSE-ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING, at APSSDC.

HEAD OF THE DEPARTMENT EXTERNAL EXAMINER

STUDENT’S DECLARATION

I______________________a student of________________ program,Regd.no________________of department of________ here by declare that I have completed the mandatory intership from___________to__________in_____________under the department of_____________.

(Signature of the student and Date)

ACKNOWLEDGEMENT

First I would like to thank **Mr. B.J. Benny**, Executive Director of **APSSDC** for giving me the opportunity to do an internship within the organization.

I also would like all the people that worked along with me in **APSSDC** with their patience and openness they created an enjoyable working environment.

It is indeed with a great sense of pleasure and immense sense of gratitude that I acknowledge the help of these individuals.

I am highly indebted to Managing Director & CEO **Mr.G.Ganesh Kumar I.A.S** and Principal **Dr. Y.V.Narayana**, for the facilities provided to accomplish this internship.

I would like to thank my Head of the Department **Dr.M.Aparna** for his constructive criticism throughout my internship.

I would like to thank **Mr.M.Sambasiva Rao,** College internship coordinator **Mrs.Sk.Shahina** internship coordinator Department of CSE-AIML for their support and their advices to get and complete internship in above said organization.

I am extremely great full to my department staff members and friends who helped me in successful completion of this internship.

MADAM LOKESH VENKATA RAMANJANEYULU

[22NE1A4237]

ABSTRACT

- The APSSDC Data Analysis Using Python Internship Program is a 8-week program that provides participants with the opportunity to learn about Data Analysis and gain hands-on experience working on real-world AI projects.
- The program is open to students from all backgrounds, and no prior experience with Data Analysis is required.
- The Data Analysis Internship Program is divided into two parts .There are two different types of data analysis methods: qualitative data analysis and quantitative data analysis. Qualitative Data Analysis "In qualitative data analysis, data is obtained through words, symbols, pictures, and observations.
- Data analytics can transform other raw data into useful and valuable insights that companies can use to improve their operations, marketing strategies and business decisions.
- Benefits:
    - Enhanced Decision-Making and understanding of Data.
    - Gain hands-on experience working on real-world Data Analytical projects.
    - Network with other students and professionals in the Data Analysis field.
    - Build your resume and portfolio.
    - Get a head start on your career in Data Visualization.

APSSDC Programs and Opportunities:

- SkillsBuild
    - Online learning platform.
    - Promote Skill Development & Entrepreneurship.

Topics include:

- Artificial intelligence and machine learning
- AWS Cloud computing.
- Python Programming.
- Andriod Application development.
- Web Designing using React.
- PCB Designing.

❖ Advanced Diploma in IT

- Two-year, full-time diploma
- Teaches emerging technology and future skills

❖ Internship program

Participants work on industry-relevant projects that simulate workplace challenges.

❖ Teacher and trainer programs

- Provides training and resources for teachers and trainers
- Helps them integrate emerging technologies into their curriculum

❖ Contact Information

- Website: [http://engineering.apssdc.in/]
- Email: [compsecy-apssdc@ap.gov.in]
- Phone: [099888 53335]

❖ Social Media

- Facebook:[https://www.facebook.com/apssdcskilldevelopment/]
- LinkedIn:[https://www.linkedin.com/company/apstateskilldevelopment?originalSubdomain=in]
- Twitter:[https://twitter.com/AP_Skill?ref_src=twsrc%5Egoogle%7Ctwcamp%5Eserp%7Ctwgr%5Eauthor]

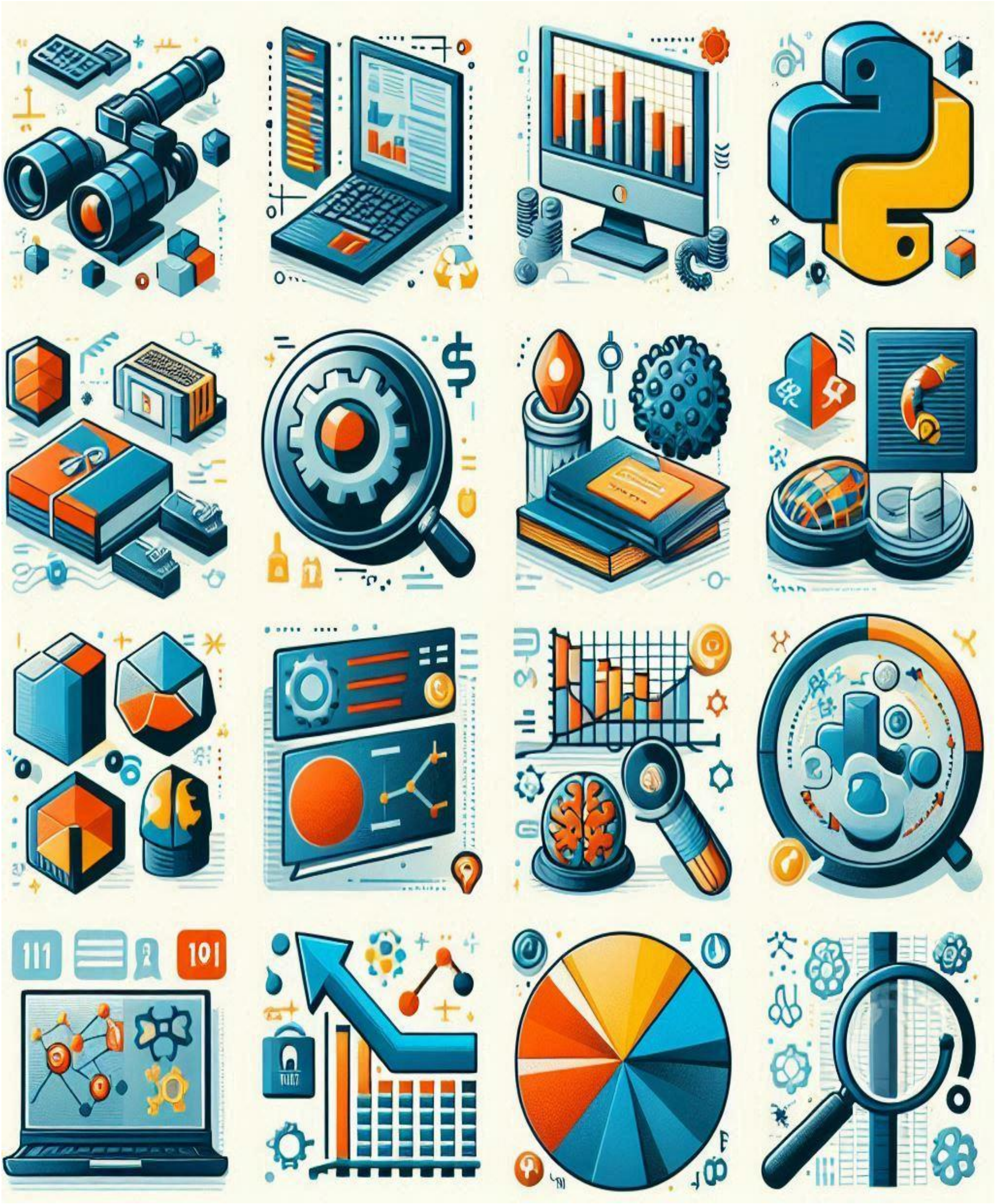
❖ How to Apply: To apply please visit the http://engineering.apssdc.in/ website.

INDEX

1.INTRODUCTION

Olympic data analysis involves examining historical data from the Olympic Games to uncover trends, patterns, and insights related to athletes, events, and participation. Here are some key points:

1. **Data Collection:**
   - Gather data on medal counts, athlete performance, and other relevant metrics from various Olympic Games.
   - Structured data (e.g., databases) and unstructured data (e.g., news articles) can be part of the dataset.
2. **Exploratory Data Analysis (EDA):**
   - Use EDA techniques to understand data distribution, relationships, and anomalies.
   - Visualize medal counts over time, explore athlete demographics, and identify trends.
3. **SQL and Python:**
   - SQL is crucial for data extraction, manipulation, and analysis.Combine SQL with Python to extract, clean, and analyze data effectivelyCombine SQL with Python to extract, clean, and analyze data effectively
4. **Challenges:**
   - Dealing with missing data, inconsistent records, and large datasets.
   - Crafting complex SQL queries to answer specific questions.
5. **Insights:**
   - Positive correlation between athletes' height and weight.
   - Increasing trend in female participation over time.


111
101

2.SYSTEM REQUIREMENTS

1. Functional Requirements:

- Python Libraries: You’ll need the following Python libraries for data analysis and visualization:
  - Numpy
  - Pandas
  - Plotly
  - Matplotlib
  - Seaborn
  - GeoPandas (optional, depending on your specific needs)

2. Development Environment:

- You can work with tools like:
  - MS Excel
  - PyCharm
  - Jupyter Notebook

3. External Interface:

- To create a user-friendly web application, consider using:
  - Streamlit
  - Google News API (if needed)

4. Operating Environment:

- You can choose between:
  - MacOS
  - Windows

3. ARCHITECTURE OF PROJECT

The architecture for Olympic data analysis can vary based on the specific project and tools used. Here is the approach of Anaconda and python

Anaconda and Python:

1. Data Collection:
   - Gather Olympic data from sources like Kaggle or official Olympic websites.
   - Obtain datasets containing information about athletes, events, medals, and other relevant details.
2. Data Cleaning and Formatting:
   - Import the datasets into a Pandas dataframe.
   - Clean and preprocess the data (handle missing values, standardize formats, etc.).
3. Exploratory Data Analysis (EDA):
   - Use Pandas, NumPy, and Matplotlib for EDA.
   - Explore trends, distributions, and correlations within the data.
4. Visualization:
   - Create visualizations using Matplotlib and Seaborn.
   - Examples:
     - Distribution of gold medals by age (countplot).

        - Medal distribution by country(bar chart or choropleth map).

- **Development Environment**:
    - Anaconda Notebooks or Jupyter Notebook for interactive coding.
- **Deployment**:

    No specific deployment architecture, as this approach is

DATA COLLECION
DATA EXPLORATION
YOUR TEXT

Fig: Architecture Diagram

4. Learning Objectives/Internship Objectives

➢ Internships are generally thought of to be reserved for college students looking to gain experience in a particular field. However, a wide array of people can benefit from Training Internships in order to receive real world experience and develop their skills.⍰

➢

➢ An objective for this position should emphasize the skills you already possess in the area and your interest in learning more⍰

➢ Internships are utilized in a number of different career fields, including architecture, engineering, healthcare, economics, advertising and many more.⍰

➢ Some internship is used to allow individuals to perform scientific research while others are specifically designed to allow people to gain first-hand experience working.⍰

➢ Utilizing internships is a great way to build your resume and develop skills that can be emphasized in your resume for future jobs. When you are applying for a Training Internship, make sure to highlight any special skills or talents that can make you stand apart from the rest of the applicants so that you have an improved chance of landing the position.

5.WEEKLY OVERVIEW OF INTERNSHIP ACTIVITIES
(DATA ANALYSIS USING PYTHON)

Week-I

Day	**Name of Topic / Module Completed**
Monday	Project setup and planning
Tuesday	Research on Python data analysis tools
Wednesday	Dataset selection and initial exploration
Thursday	Literature review on data analysis techniques
Friday	Data preparation and validation
Saturday	Summary of findings and next steps

Week-II

Day	**Name of Topic / Module Completed**
Monday	Understanding project requirements
Tuesday	Basics of Python (pandas, numpy)
Wednesday	Data collection and sourcing
Thursday	Data cleaning and preprocessing
Friday	Handling missing data
Saturday	Exploratory Data Analysis (EDA) basics

Week-III

Day	**Name of Topic / Module Completed**
Monday	Advanced EDA with Python
Tuesday	Visualizing trends and patterns
Wednesday	Implementing basic statistical models
Thursday	Model testing and validation
Friday	Introduction to regression analysis
Saturday	Building and evaluating regression models

Week-IV

Day	Name of Topic / Module Completed
Monday	Introduction to machine learning
Tuesday	Train/test data splitting
Wednesday	Implementing classification models
Thursday	Evaluating model performance
Friday	Introduction to clustering algorithms
Saturday	Implementing K-Means clustering

Week-V

Day	Name of Topic / Module Completed
Monday	Random Forests for data analysis
Tuesday	Model comparison and optimization
Wednesday	Introduction to Python dashboards (Plotly)
Thursday	Creating interactive visualizations
Friday	Pipeline creation for data analysis
Saturday	Reviewing and refining the analysis workflow

Week-VI

Day	Name of Topic / Module Completed
Monday	Model evaluation metrics
Tuesday	Hyperparameter tuning
Wednesday	Visualization of results
Thursday	Finalizing optimized models
Friday	Integration of visual tools
Saturday	Preparing final analysis report

Week-VII

Day	**Name of Topic / Module Completed**
Monday	Deployment strategies
Tuesday	Building user interfaces (Flask/Streamlit)
Wednesday	Testing with real-world data
Thursday	Debugging and refinement
Friday	Final presentation preparation
Saturday	Rehearsal of the project presentation

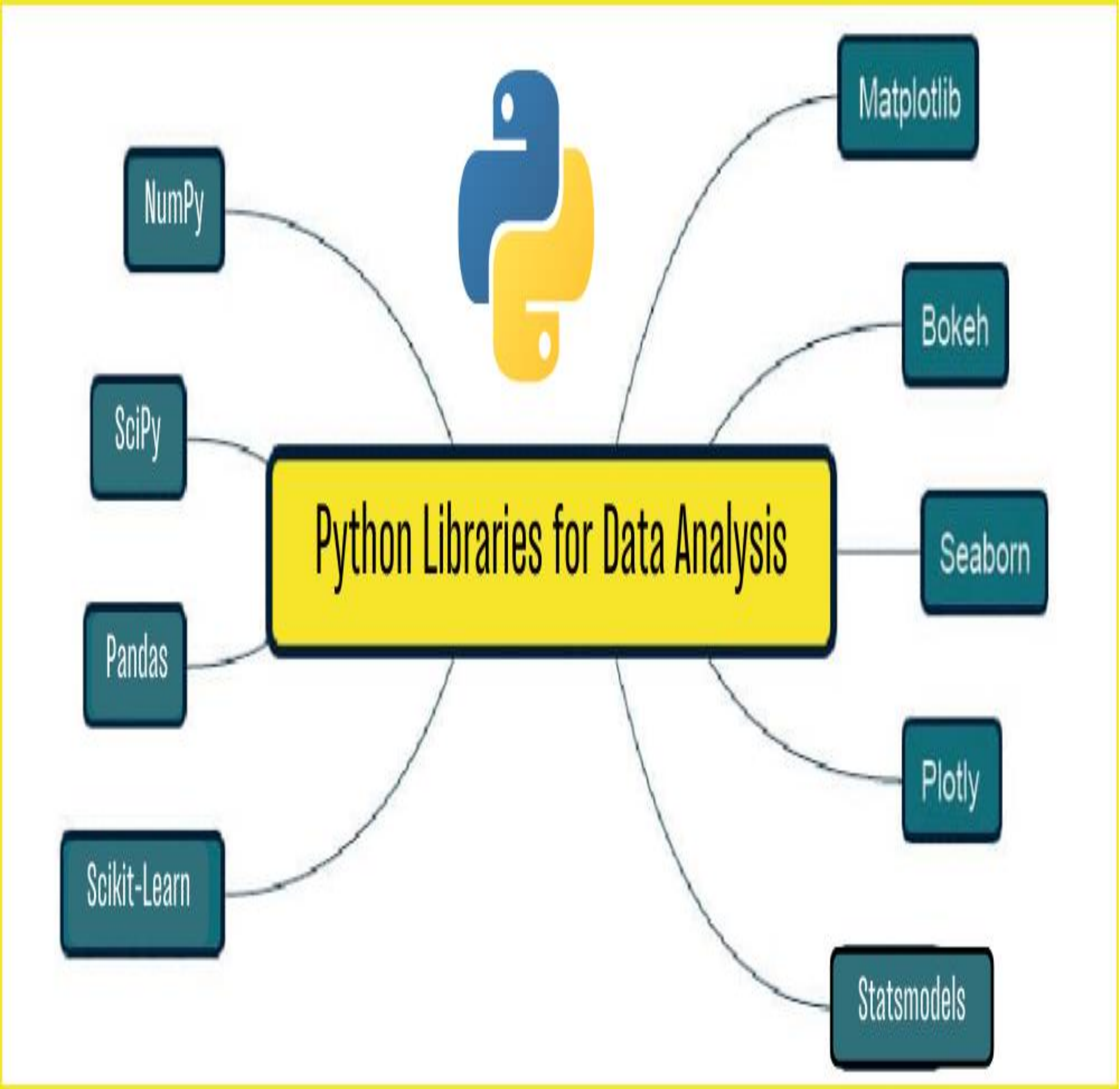
Week-VIII

Day	**Name of Topic / Module Completed**
Monday	Project documentation compilation
Tuesday	Writing the final report: methodology
Wednesday	Writing the final report: results analysis
Thursday	Finalizing conclusions and recommendations
Friday	Reviewing and refining the final report
Saturday	Submitting the final project

6.USES OF DATA ANALYSIS LIBRARY

In Olympic data analysis, various Python libraries play crucial roles. Let's explore how these libraries are used:

1. **Pandas**:
   - It is used for analyzing the data
   - **Data Manipulation**: Pandas is essential for loading, cleaning, and transforming data. It allows you to work with data frames efficiently.
   - **Example**: Loading Olympic datasets, handling missing values, and filtering relevant information.
2. **Matplotlib** and **Seaborn**:
   - Matplotlib is a numerical mathematics extension Numpy and Sea born is used for visualization statistical graphics plotting in python
   - **Data Visualization**: These libraries help create charts, plots, and graphs.
   - **Examples**:
     - Visualizing medal distributions by country (bar charts).
     - Plotting trends in athlete participation over time (line plots).
3. **NumPy**:
   - **Numerical Computations**: NumPy provides powerful array operations and mathematical functions.
   - **Example**: Calculating summary statistics(mean,median,etc.) for athlete attributes.


Python Libraries for Data Analysis
NumPy
SciPy
Pandas
Scikit-Learn
Matplotlib
Bokeh
Seaborn
Plotly
Statsmodels

7.PROJECT CODE

```
# Import libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

```
data = pd.read_csv('athlete_events.csv')
 # read file
# data.head() display first 5 entry
# data.describe  about model
# data.info give info about data
print(data.head(), data.describe(), data.info())
```

```
# regions and country noc data csv file

regions = pd.read_csv('datasets_31029_40943_noc_regions.csv')
print(regions.head())

# merging to data and regions frame

merged = pd.merge(data, regions, on='NOC', how='left')
print(merged.head())
```

```
goldMedals = merged[(merged.Medal == 'Gold')]
print(goldMedals.head())
```

```
plt.figure(figsize=(20,10))
plt.title('Distribution of Gold Medals')
sns.countplot(data=goldMedals,x='Age')
plt.show()
```

```
goldMedals = merged[(merged.Medal == 'Gold')]

print('The no of athletes is',goldMedals['ID'][goldMedals['Age'] > 50].count(),
'\n')

print(goldMedals[goldMedals['Age'] > 50])
```

```
masterDisciplines = goldMedals.loc[goldMedals['Age'] > 50]
plt.figure(figsize=(20, 10))
plt.tight_layout()
sns.countplot(data=masterDisciplines,x='Sport')
plt.title('Gold Medals for Athletes Over 50')
plt.show()
```

```
womenInOlympics = merged[(merged.Sex == 'F') &
                         (merged.Season == 'Summer')]
print(womenInOlympics.head(10))

sns.set(style="darkgrid")
plt.figure(figsize=(20, 10))
sns.countplot(x='Year', data=womenInOlympics)
plt.title('Women medals per edition of the Games')
plt.show()
```

```
print(goldMedals.region.value_counts().reset_index(name='Medal').head())
medals_by_region =
goldMedals['region'].value_counts().reset_index(name='Medal').head(5)
g = sns.catplot(x="region", y="Medal",data=medals_by_region,
                height=6, kind="bar", palette="muted")
g.despine(left=True)
g.set_xlabels("Countries")
g.set_ylabels("Number of Medals")
plt.title('Medals per Country')
plt.show()
```

```
MenOverTime = merged[(merged.Sex == 'M') &(merged.Season == 'Summer')]

wlMenOverTime = MenOverTime.loc[MenOverTime['Sport'] == 'Weightlifting']

plt.figure(figsize=(20, 10))

sns.pointplot(x='Year',y='Weight', data=wlMenOverTime, palette='Set2')

plt.title('Weight over year for Male Lifters')
plt.xlabel('Year')
plt.ylabel('Weight')

plt.show()
```

8.GRAPHICAL REPRESENTATION OF OUTPUTS

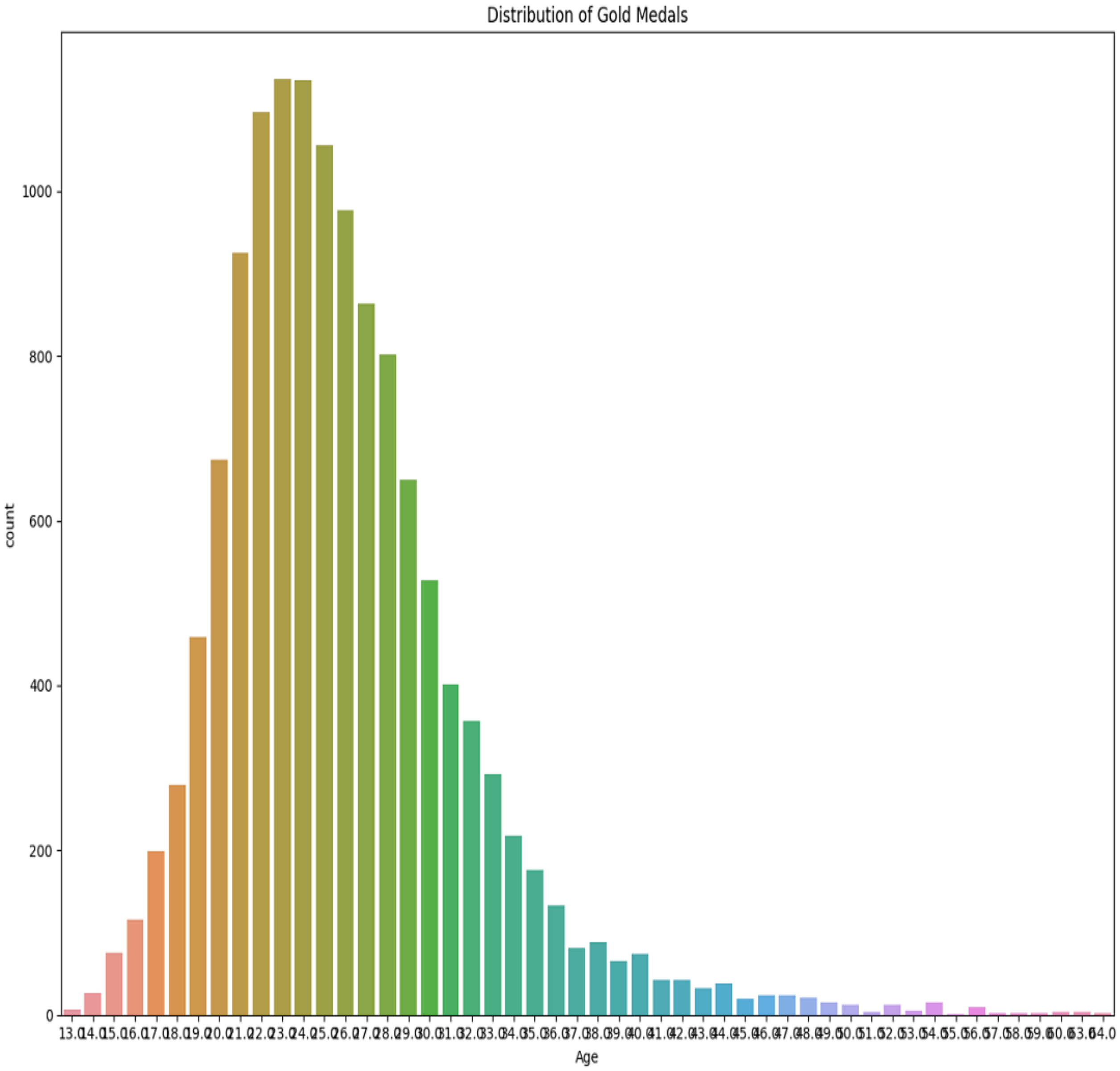


Fig: Gold Medalist of Respective Ages

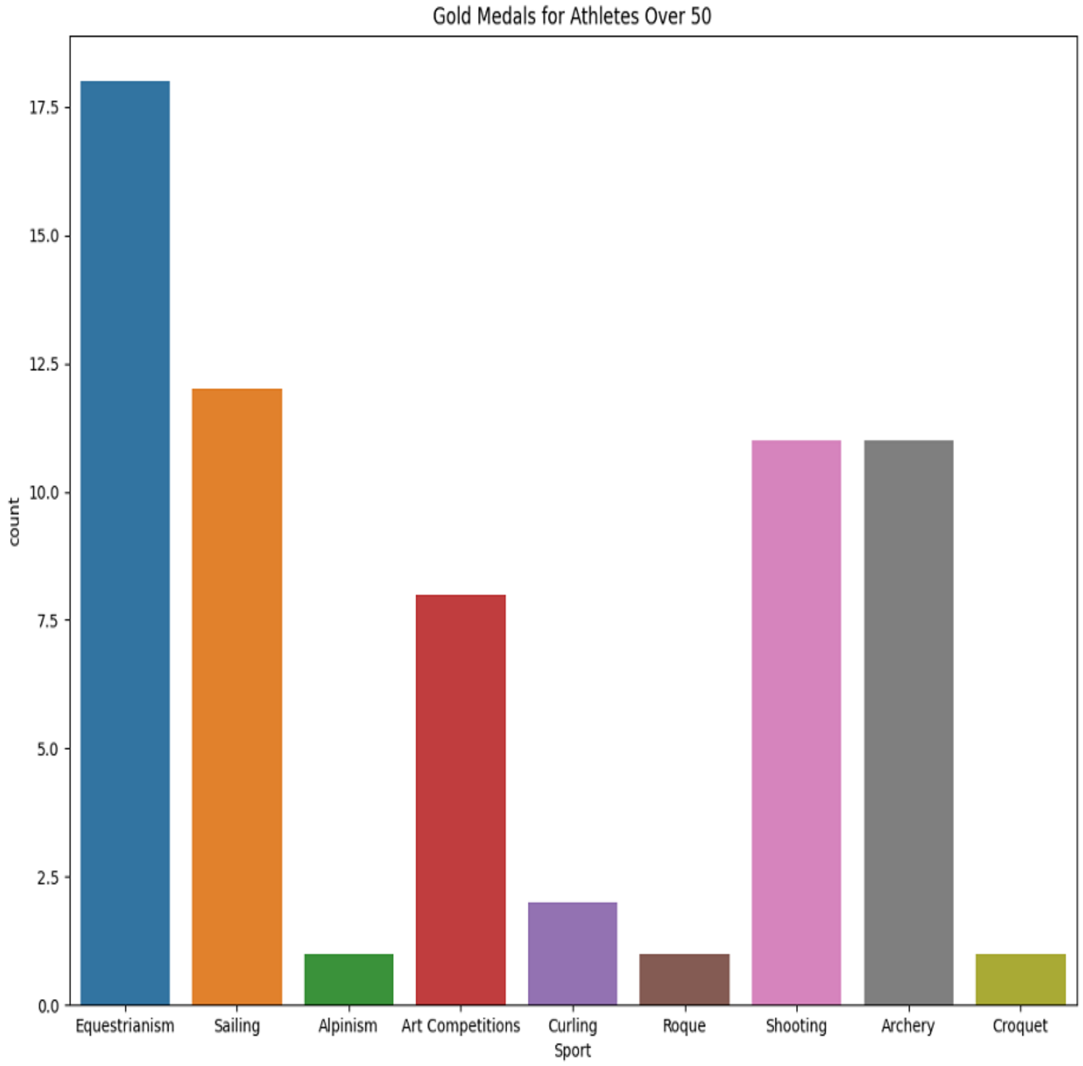


Fig: athletes who are gold medalists and whose age is greater than 50 by visualization

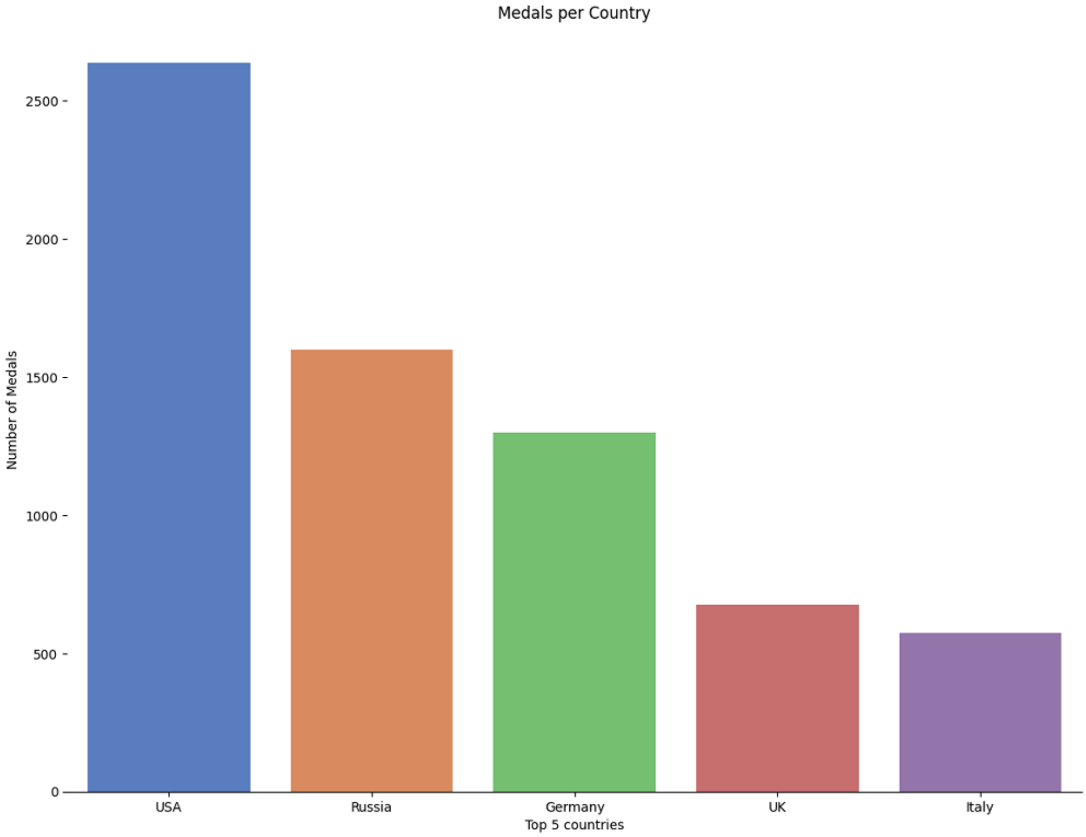


Fig: Top 5 countries by catplot

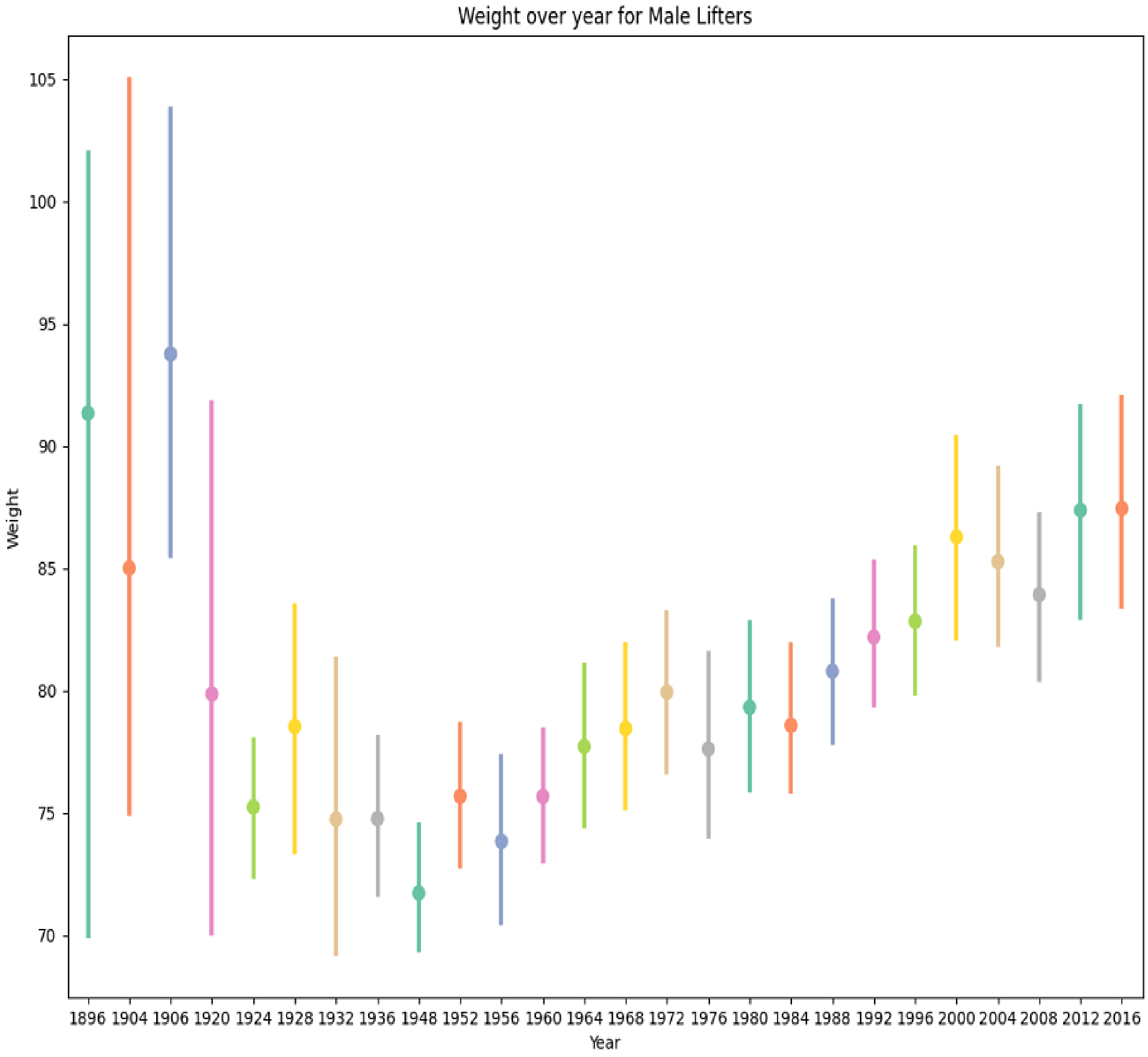


Fig: Male Lifters via graphical representation using pointplot.

9.ADVANATGES

Analyzing Olympic data using Python can provide valuable insights into athlete performance, historical trends, and other factors influencing the outcome of the games. Here are some advantages of Olympics data analysis using Python:

1. **Rich Data Exploration**: The modern Olympic Games involve thousands of athletes from around the world participating in various sports competitions. By analyzing Olympic data, you can explore details such as athlete profiles, events, medals, and more.
2. **Data Cleaning and Formatting**: Python libraries like Pandas allow you to import and manipulate datasets efficiently. You can clean and format the data, handle missing values, and prepare it for analysis.
3. **Visualization with Seaborn and Matplotlib**: Seaborn and Matplotlib are powerful tools for creating visualizations. You can generate graphs, charts, and plots to visualize trends, distributions, and relationships within the data.

10.CONCLUSION

Conclusion For Olympic Data Analysis Project:

In this project, we explored Olympic data using Python. Here are the key takeaways:

1. Medal Distribution:
   - We analyzed medal distribution across countries and events.
   - Visualizations revealed trends in medal counts over time.
2. Athlete Profiles:
   - We examined athlete demographics (age, gender, nationality).
   - Identified patterns in medal-winning athletes.
3. Historical Trends:
   - Investigated changes in sports participation and medal distribution over decades.
   - Considered factors like host cities and geopolitical events.
4. Recommendations:
   - For future Olympics, focus on sports with low representation.
   - Explore correlations between training facilities, funding, and medal success.

Remember that data analysis is an ongoing process. Continue refining your insights and consider additional dimensions for a comprehensive view.

Reference links: https://www.geeksforgeeks.org/olympics-data-analysis-using-python/?ref=ml_lbp