Data-Driven Insights on Olympic Sports Participation and Performance

INTRODUCTION:

The Olympic Games stand as a pinnacle of athletic achievement, uniting nations in the spirit of competition, sportsmanship, and human potential. For over a century, this global event has captivated the hearts and minds of people around the world, drawing athletes from diverse backgrounds and disciplines to showcase their talents on the world stage. The Olympics are not just a celebration of athletic prowess; they are also a reflection of the changing dynamics in the world of sports. Understanding the trends, participation, and performances in Olympic sports is crucial for both enthusiasts and stakeholders. This project delves into the world of Olympic sports and participation, employing data-driven methodologies to gain deeper insights into this iconic sporting event.

1.1 PROJECT OVERVIEW:

The "Data-Driven Insights on Olympic Sports Participation and Performance " project aims to provide a comprehensive understanding of the intricate detailsof the Olympic Games. This research venture employs a multifaceted approach, leveraging a rich dataset spanning several decades of Olympic history. Here's an overview of the key aspects of this project:

1.2 PURPOSE:

The purpose of the " Data-Driven Insights on Olympic Sports Participation and Performance " project is to:

a) Provide insights into the historical trends and developments in Olympic sports and athlete participation.
b) Support informed decision-making by Olympic committees, sports organizations, and policymakers.
c) Promote inclusivity and diversity in sports by analyzing gender and geographic representation.
d) Offer an educational resource for students, educators, and the general public interested in the Olympics.
e) Create predictive models for entertainment and engagement of sports enthusiasts.
f) Serve as a resource for academic research in fields related to sports, data analysis, and culture, contributing to the academic understanding of the Olympics.

2. LITRATURE SURVEY:

2.1 EXISTING PROBLEM:

The analysis of Olympic sports and participation encounters challenges such as inconsistent data quality, data privacy concerns, biases in the data, complex data integration, the difficulty of accurate predictive modeling, and effective communication of findings while addressing ethical and cultural considerations.

2.2 REFERENCE:

Shoval (2012) divides the history of mega-events such as the Olympic Games into four periods. During the first two, respectively from 1851 to 1939 and from 1948 to 1984, there was the rise and subsequent fall of the World's Fair company, the most important mega-event organizers.The feature of the third phase, which began with the Los Angeles Olympic Games in 1984 and ended in 2000 with the Sydney Games.
the Olympics have also experienced an accelerated increase since the 1960 Games in Rome, which saw the participation of about 6,000 athletes, almost all men (women were only slightly more than 600). Gender catchup (with an increase of more than 4,000 number of women) explains almost all the raise of the number of participants, which constantly exceeded 10,000 athletes since the 1999 Games in Atlanta. Participation rates (ratio between tickets issued and sold), on the other hand, went first down from more than 80% in the 1960s to about 72% in Athens 2004, and then soared topractically 100% in Beijing 2008 and London 2012.

2.3 PROBLEM STATEMENT DEFINITION:

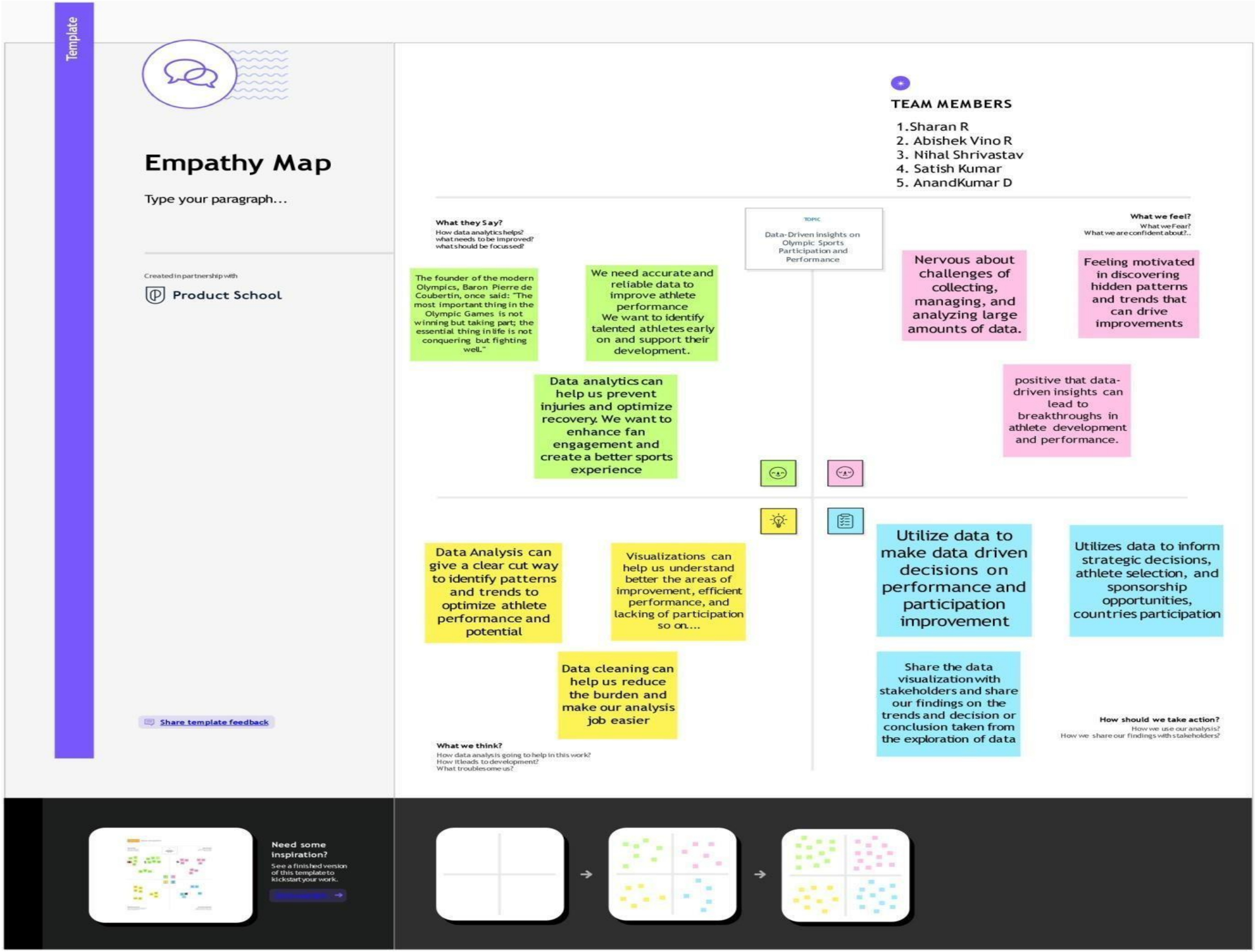
The challenge of "Olympic Sports and Participation" encompasses a wide range of crucial aspects related to organizing and facilitating the Olympic Games, a globally celebrated sporting event held every four years. It involves selecting which sports and disciplines should be included in the Olympic program, ensuring diverse and representative athlete participation, planning and building the necessary venues and infrastructure, managing complex logistics and operations, securing and managing finances, promoting and marketing the Games, enforcing rules and regulations, assessing the cultural and social impact, and planning for a lasting legacy and sustainability in the host city. All of these elements are critical to delivering a successful and meaningful Olympic Games experience while upholding the principles of the Olympic movement.

Problem Statement (PS)	**I am (Customer)**	**I'm trying to**	**But**	**Because**	**Which makes me feel**
PS-1	coach	Optimize training strategies	struggle to effectively track and analyze athlete performance data in Olympic sports	I lack in data	Confused and sad
PS-2	stakeholders	Identify opportunities for increased diversity and incusivity	lack comprehensive and easily accessible data-driven insights on Olympic sports participation trends	Data analyzers	low
PS-3	Athlete	Trying to integrate wearable technology and data	Its difficult for me to choose	Limiting my ability to leverage technological	Confused and lost

		analytics to monitor and improve performance		advancements in olympic sports	

3.IDEATION AND PROPOSED SYSTEM

3.1 EMPATHY MAP CANVAS:



Empathy Map

Type your paragraph...

Created in partnership with

Product School

Share template feedback

TEAM MEMBERS

1. Sharan R
2. Abishek Vino R
3. Nihal Shrivastav
4. Satish Kumar
5. AnandKumar D

TOPIC

Data-Driven insights on Olympic Sports Participation and Performance

What they Say?

How data analytics helps?
what needs to be improved?
what should be focussed?

The founder of the modern Olympics, Baron Pierre de Coubertin, once said: "The most important thing in the Olympic Games is not winning but taking part; the essential thing in life is not conquering but fighting well."

We need accurate and reliable data to improve athlete performance
We want to identify talented athletes early on and support their development.

Data analytics can help us prevent injuries and optimize recovery. We want to enhance fan engagement and create a better sports experience

What we feel?

What we Fear?
What we are confident about?..

Nervous about challenges of collecting, managing, and analyzing large amounts of data.

Feeling motivated in discovering hidden patterns and trends that can drive improvements

positive that data-driven insights can lead to breakthroughs in athlete development and performance.

Data Analysis can give a clear cut way to identify patterns and trends to optimize athlete performance and potential

Visualizations can help us understand better the areas of improvement, efficient performance, and lacking of participation so on....

Data cleaning can help us reduce the burden and make our analysis job easier

What we think?

How data analysis is going to help in this work?
How it leads to development?
What troublesome us?

Utilize data to make data driven decisions on performance and participation improvement

Utilizes data to inform strategic decisions, athlete selection, and sponsorship opportunities, countries participation

Share the data visualization with stakeholders and share our findings on the trends and decision or conclusion taken from the exploration of data

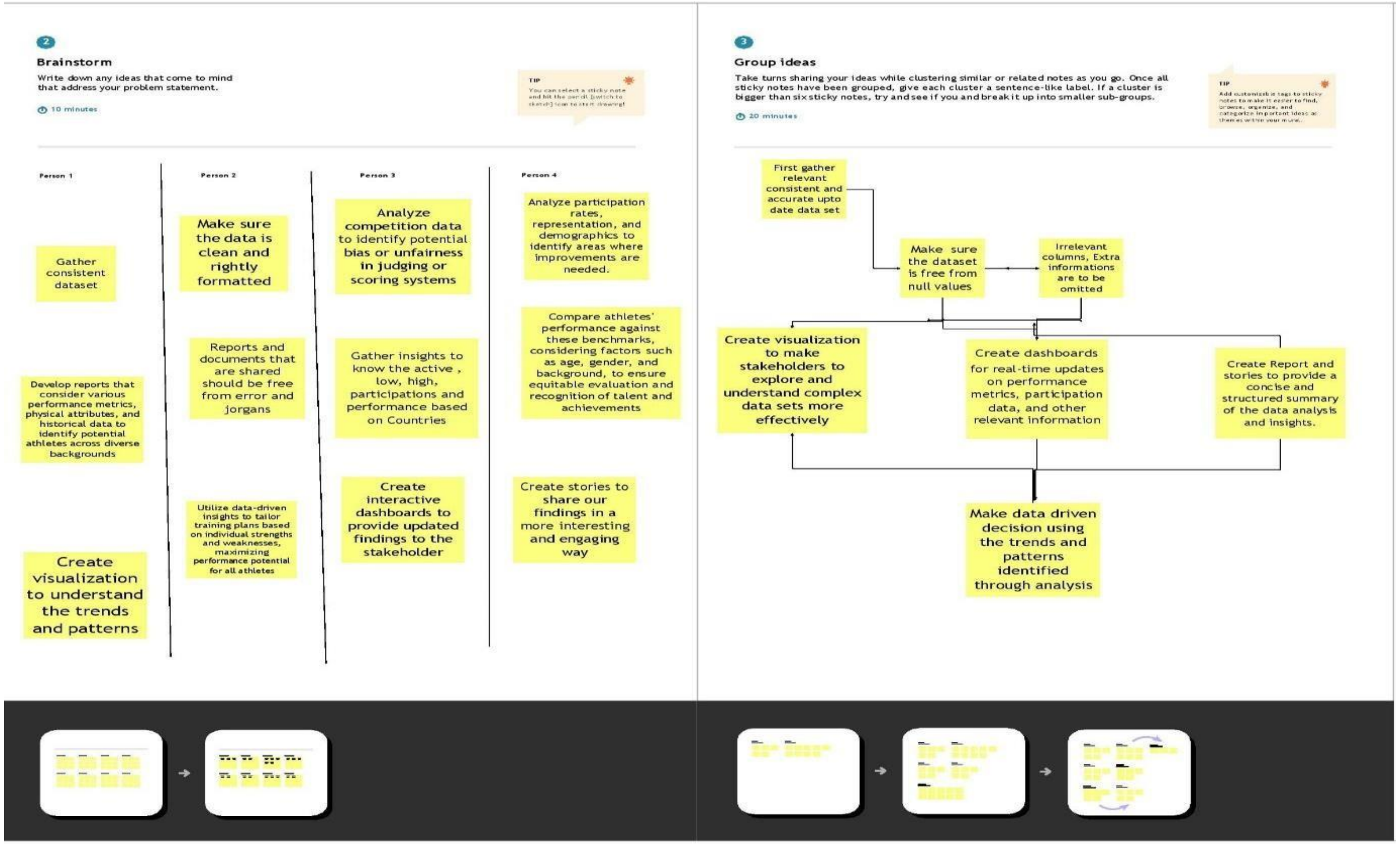
How should we take action?

How we use our analysis?
How we share our findings with stakeholders?

Need some inspiration?

See a finished version of this template to kickstart your work..

3.2 INDENTATION AND BRAIN STORMING:



2

Brainstorm

Write down any ideas that come to mind that address your problem statement.

10 minutes

TIP

You can select a sticky note and hit the pencil [switch to sketch] icon to start drawing!

Person 1

Gather consistent dataset

Develop reports that consider various performance metrics, physical attributes, and historical data to identify potential athletes across diverse backgrounds

Create visualization to understand the trends and patterns

Person 2

Make sure the data is clean and rightly formatted

Reports and documents that are shared should be free from error and jorgans

Utilize data-driven insights to tailor training plans based on individual strengths and weaknesses, maximizing performance potential for all athletes

Person 3

Analyze competition data to identify potential bias or unfairness in judging or scoring systems

Gather insights to know the active , low, high, participations and performance based on Countries

Create interactive dashboards to provide updated findings to the stakeholder

Person 4

Analyze participation rates, representation, and demographics to identify areas where improvements are needed.

Compare athletes' performance against these benchmarks, considering factors such as age, gender, and background, to ensure equitable evaluation and recognition of talent and achievements

Create stories to share our findings in a more interesting and engaging way

3

Group ideas

Take turns sharing your ideas while clustering similar or related notes as you go. Once all sticky notes have been grouped, give each cluster a sentence-like label. If a cluster is bigger than six sticky notes, try and see if you and break it up into smaller sub-groups.

20 minutes

TIP

Add customizable tags to sticky notes to make it easier to find, browse, organize, and categorize important ideas as themes within your mural.

First gather relevant consistent and accurate upto date data set

Make sure the dataset is free from null values

Irrelevant columns, Extra informations are to be omitted

Create visualization to make stakeholders to explore and understand complex data sets more effectively

Create dashboards for real-time updates on performance metrics, participation data, and other relevant information

Create Report and stories to provide a concise and structured summary of the data analysis and insights.

Make data driven decision using the trends and patterns identified through analysis

:

4. REQUIREMENT ANALYSIS:

4.1 FUNCTIONAL REQUIREMENTS:

Following are the functional requirements of the proposed solution.

FR No.	Functional Requirement (Epic)	Sub Requirement (Story / Sub-Task)
FR-1	Data collection	Retrieve data from various sources such as official Olympic records, sports federation.
FR-2	Data cleaning	Handle missing data, outliners and inconsistencies in the data.
FR-3	Performance analysis	Analyze athlete, team and country performance in different sports.
FR-4	Reporting and Dashboards	Generate reports, summaries and dashboards to communicate insights to stakeholders.

4.2 NON FUNCTIONAL REQUIREMENTS:

Following are the non-functional requirements of the proposed solution.

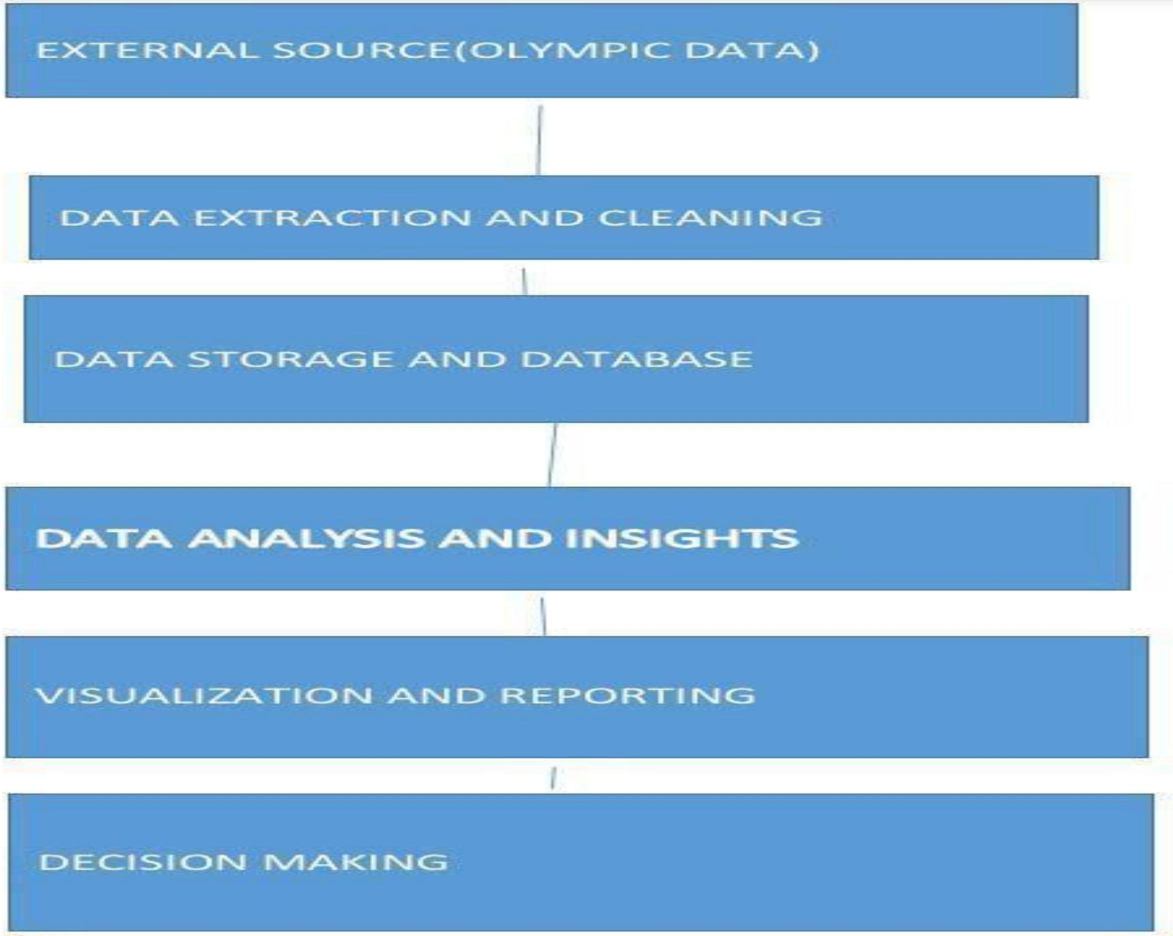
FR No.	Non-Functional Requirement	Description
NFR-1	**Usability**	The usability aspect of a data-driven insights system on Olympics sports and participation is crucial to ensure that users can interact with the system effortlessly and achieve their goals efficiently
NFR-2	**Security**	Security is a critical aspect of a data-driven insights system on Olympics sports and participation. It focuses on protecting the system, its data, and the privacy of users.

NFR-3	**Reliability**	It focuses on ensuring that the system performs its intended functions accurately and consistently, without failures or disruptions.
NFR-4	**Performance**	It focuses on ensuring that the system operates efficiently and provides timely responses to user interactions.
NFR-5	**Availability**	It focuses on ensuring that the system is operational and accessible to users whenever they need it.
NFR-6	**Scalability**	Scalability ensures that the system can accommodate growth and maintain its performance as usage and data requirements increase over time.

5. PROJECT DESIGN:

5.1 Data Flow Diagrams & User stories:

A Data Flow Diagram (DFD) is a traditional visual representation of the information flows within a system. A neat and clear DFD can depict the right amount of the system requirement graphically. It shows how data enters and leaves the system, what changes the information, and where data is stored.

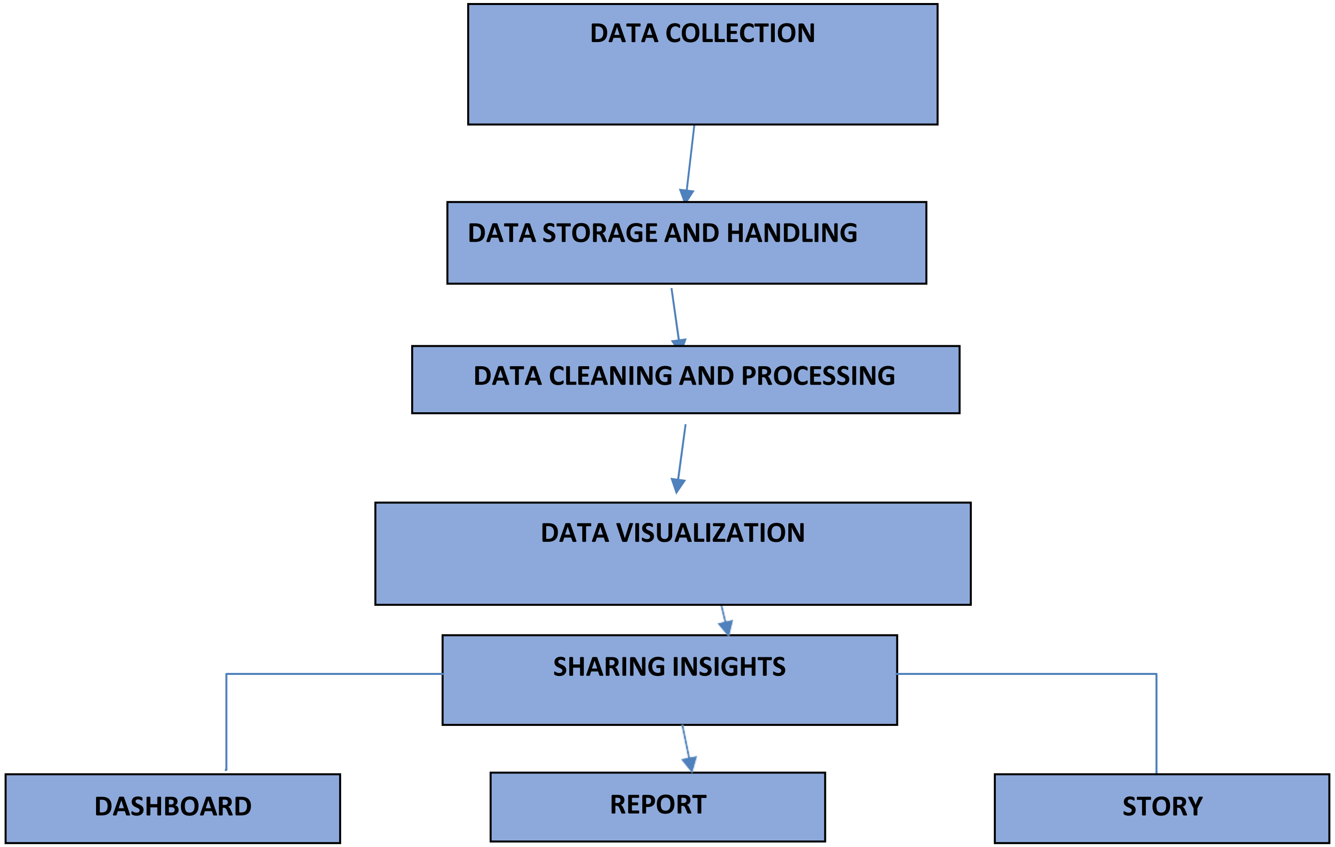

EXTERNAL SOURCE(OLYMPIC DATA)
DATA EXTRACTION AND CLEANING
DATA STORAGE AND DATABASE
DATA ANALYSIS AND INSIGHTS
VISUALIZATION AND REPORTING
DECISION MAKING

Data flow diagram

Data collection	Login	USN-1	As a data analyst, I want to gather comprehensive data on Olympics sports, athletes, countries, and competitions, so that I can perform thorough analysis and generate valuable insights	The system should provide mechanisms to collect data from official Olympic records, sports federations, athlete profiles, and historical records.	High	Sharan R
Data cleaning		USN-2	As a data analyst, I want to clean and preprocess the collected data to ensure its accuracy, consistency, and suitability for analysis.	Data variables should be standardized and formatted correctly to maintain consistency throughout the data set.	High	Abishek Vino R
Data analysis		USN-3	As a data analyst, I want to analyze the collected data to identify performance patterns, trends, and correlations between different variables.	It should support the generation of charts, graphs, and maps to visualize the data and present insights effectively.	Medium	Nihal Shrivastav
Performance analysis		USN-4	As a data analyst, I want to develop predictive models to forecast outcomes such as medal counts, performance rankings, and the likelihood of participation in specific sports.	The system should support the development and implementation of predictive algorithms, such as regression models or machine learning algorithms.	Medium	Satish Kumar
Reporting		USN-5	As a data analyst, I want to generate reports, summaries, and dashboards to communicate the insights and findings to stakeholders.	It should support the creation of interactive dashboards with drill-down capabilities to explore the data further.	High	Anand Kumar D

visualization	Dashboard					

5.2 Solution Architecture:



DATA COLLECTION:

At the core of the architecture is the data collection layer, responsible for gathering data from various sources. We use kaggle dataset to collect and gather our required data for analysis

DATA MANAGEMENT:

Technological advancements in data management, processing power, and analytics tools have significantly enhanced the feasibility of our solution. Modern data storage systems and cloud computing infrastructure allow for the efficient storage and processing of large volumes of data

We use IBM DB2 which support various data-driven applications and is available on Linux, UNIX and Windows operating systems It provides a wide range of features such as data security, scalability, high availability, and performance. It also supports various programming languages such as SQL, Java, C++, and others

DATA PROCESSING:

We use IBM COGNOS tools which provide a wide range of capabilities, including data cleansing, feature engineering, predictive modeling, and visualization. By utilizing these technologies, we can efficiently process and analyze complex Olympic sports datasets, thereby enhancing the feasibility of our solution**.**

DATA VISUALIZATION:

Visualization techniques help identify patterns and trends that might be hidden in raw data. By visually representing data over time or across different variables, stakeholders can identify correlations, outliers, and anomalies that might not be apparent in tabular or numerical form. This allows for the identification of factors that contribute to Olympic sports participation, performance improvement, or other relevant insights.

REPORTS, DASHBOARD AND STORIES:

Visualization tools can support real-time monitoring of performance metrics, training progress, and other key indicators. By providing live dashboards that update with new data, stakeholders can monitor performance trends, track training effectiveness, and make timely adjustments to training programs. Real-time monitoring allows for proactive decision-making and quick responses to emerging trends or issues. Data visualization helps in telling compelling stories with data. By combining visual elements with narrative techniques, stakeholders can create data-driven stories that engage and captivate the audience. This storytelling approach enables stakeholders to communicate the impact of their insights, drive behavior change, and inspire athletes and coaches to strive for better performance

Overall, the solution architecture combines data collection, storage, processing, analytics, visualization to enable comprehensive data-driven insights on Olympic sports participation and performance. This architecture facilitates the efficient and effective generation of valuable insights that empower stakeholders to make informed decisions and drive improvements in the Olympic sports ecosystem

6. PROJECT PLANNING AND SCHEDULING:

Product Backlog, Sprint Schedule, and Estimation

Sprint	Functional Requirement (Epic)	User Story Number	User Story / Task	Story Points	Priority	Team Members
Sprint-1	Registration	USN-1	As a user, I can register for the application and use this for future analysis	2	High	Nihal Shrivastav Member 1
Sprint-1		USN-2	As a user, I will receive many data visualization diagrams of the datasets	1	High	Sharan R Member 2
Sprint-2		USN-3	As a user, I can use this for references	2	Medium	Abishek Vino R Member 3
Sprint-1		USN-4	As a user, I can analyze and help others	1	Low	Anand Kumar D Member 4
Sprint-1	Login	USN-5	As a user, I can help others and create data	2	Medium	Satish Kumar Member 5
	Dashboard					

Project Tracker, Velocity & Burndown Chart:

Sprint	Total Story Points	Duration	Sprint Start Date	Sprint End Date (Planned)	Story Points Completed (as on Planned End Date)	Sprint Release Date (Actual)
Sprint-1	20	6 Days	24 sept	29 sept	20	29 Oct
Sprint-2	20	6 Days	31 sept	09 sept	18	09 sept
Sprint-3	20	6 Days	07 oct	12 oct	20	12 oct
Sprint-4	20	6 Days	14 oct	19 oct	19	19 oct

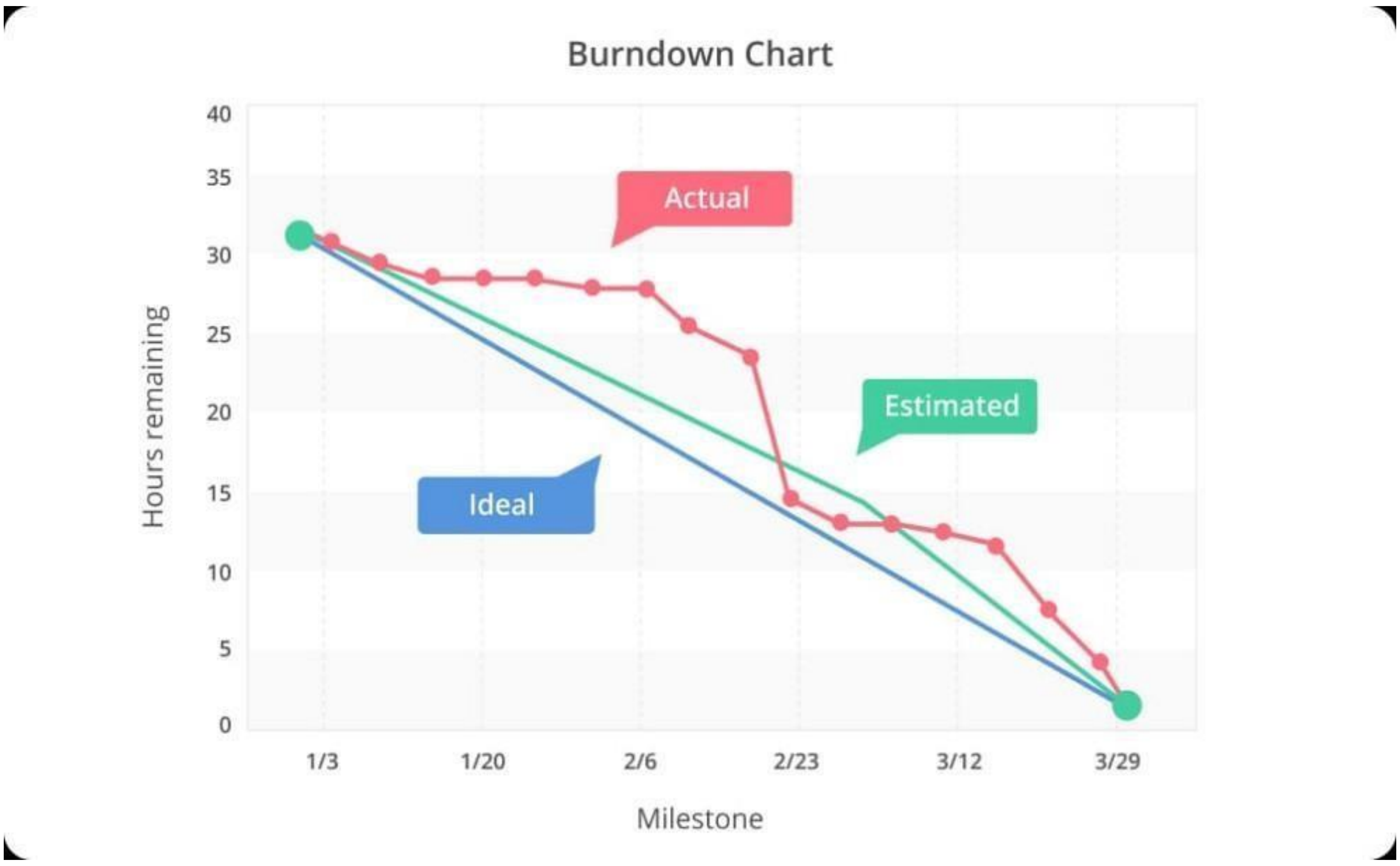
Velocity:

Imagine we have a 10-day sprint duration, and the velocity of the team is 20 (points per sprint). Let's calculate the team's average velocity (AV) per iteration unit (story points per day)

$$AV = \frac{sprint\ duration}{velocity} = \frac{20}{10} = 2$$

Burndown Chart:

A burn down chart is a graphical representation of work left to do versus time. It is often used in agile software development methodologies such as Scrum. However, burn down charts can be applied to any project containing measurable progress over time.



7. CODING AND SOLUTIONING:

```
!pip install pandas
!pip install seaborn
import pandas as pd
import numpy as np
from google.colab import files
uploaded = files.upload()
df = pd.read_csv('Automobile_data.csv')
df.head(10)
import seaborn as sns

"""UNIVARIATE ANALYSIS"""
sns.countplot(x=df['make'])
import matplotlib.pyplot as plt
sns.violinplot(x='wheel-base',data=df,color='yellow')


"""BIVARIATE ANALYSIS"""
sns.scatterplot(x='engine-type', y='engine-size', data=df)
sns.lineplot(x='body-style', y='length', data=df)
```

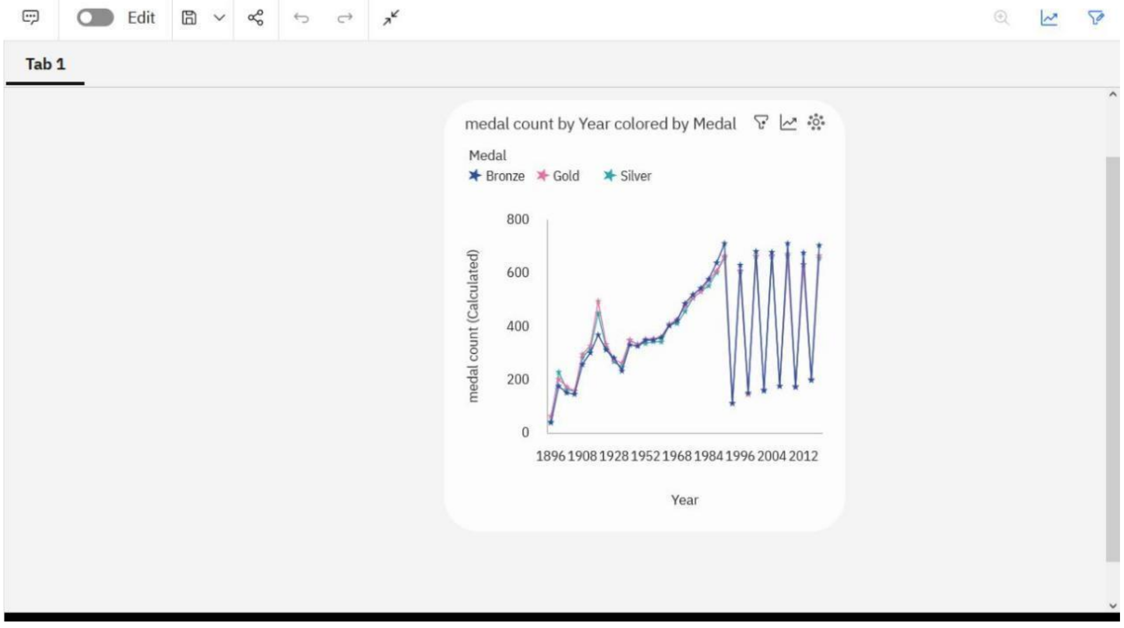
```
"""Multivariate Analysis"""
sns.heatmap(df.corr(), annot=True)
sns.jointplot(x='fuel-type', y='aspiration', data=df)
sns.set_style("whitegrid")
sns.pairplot(
 df[["length", "height", "width", "make"]],
   hue = "make",
    height = 3,
    palette = "Set1")

""" Handling missing values"""
df.replace("?", np.nan, inplace = True)
df.head()
df.isnull().sum()
#replacing missing values with average value
avgnorm = df["normalized-losses"].astype("float").mean(axis=0)
df["normalized-losses"].replace(np.nan, avgnorm, inplace=True)
avgbore = df["bore"].astype("float").mean(axis=0)
df["bore"].replace(np.nan, avgbore, inplace=True)
avgstroke = df["stroke"].astype("float").mean(axis=0)
df["stroke"].replace(np.nan, avgstroke, inplace=True)
avghp = df["horsepower"].astype("float").mean(axis=0)
df["horsepower"].replace(np.nan, avghp, inplace=True)
avgpeak = df["peak-rpm"].astype("float").mean(axis=0)
df["peak-rpm"].replace(np.nan, avgpeak, inplace=True)
avgprice = df["price"].astype("float").mean(axis=0)
df["price"].replace(np.nan, avgprice, inplace=True)
df.head()
x='unknown'
df["num-of-doors"].replace(np.nan,x, inplace=True)
df.isnull().sum()

""" Handling categorical variables(Encoding)"""
df = pd.read_csv('Automobile_data.csv')
from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()
df['make_1'] = le.fit_transform(df['make'])
df['fuel-type_1'] = le.fit_transform(df['fuel-type'])
df['engine-type_1'] = le.fit_transform(df['engine-type'])
df['body-style_1'] = le.fit_transform(df['body-style'])
df['drive-wheels_1'] = le.fit_transform(df['drive-wheels'])
```

```
print(df.head)

"""Performing Scaling"""
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
df=pd.DataFrame({'wheel-base':[88.6,99.8,99.4,105.8,101.2,88.4,93.7,96.5,113,102,98.8,102.7,93,96.3,99.2,93.3,95.7,102.4,104.5,97.3]})
df_standardized = scaler.fit_transform(df)
df_standardized = pd.DataFrame(df_standardized, columns=df.columns)
print(df_standardized.head())

"""Correlation and Descriptive Analysis"""
df=pd.DataFrame({'length':[168.8,171.2,176.6,192.7,178.2,176.8,189,197],
                 'width':[64.1,65.5,66.2,66.4,71.4,67.9,64.8,66.9],
                 'height':[48.8,52.4,54.3,55.7,52,56.3,47.8,50.2]})
corr_matrix = df.corr()
sns.heatmap(corr_matrix, annot=True, cmap='Purples')
df.describe()

"""Building Machine Learning Model"""
df=pd.DataFrame({'price':[13495,16500,13950,17450,17710,16430,24565],
                 'engine-size':[130,109,136,131,164,90,79]})
y=df['price']
X = df.drop("price", axis = 1)
df["engine-size"] = [float(str(i)) for i in df["engine-size"]]
df["price"] = [float(str(i)) for i in df["price"]]
from sklearn.model_selection import train_test_split
X_train,X_test,y_train,y_test=train_test_split(
    X,y,
    train_size = 0.50,
    random_state = 1)
from sklearn.linear_model import LinearRegression
lr = LinearRegression()
lr.fit(X_train,y_train)

"""Evaluating Machine Learning Model"""
from sklearn.metrics import mean_squared_error
import math
y_pred = lr.predict(X_test)
print("our model predicts with the deviation of ",math.sqrt(mean_squared_error(y_test, y_pred)))
```

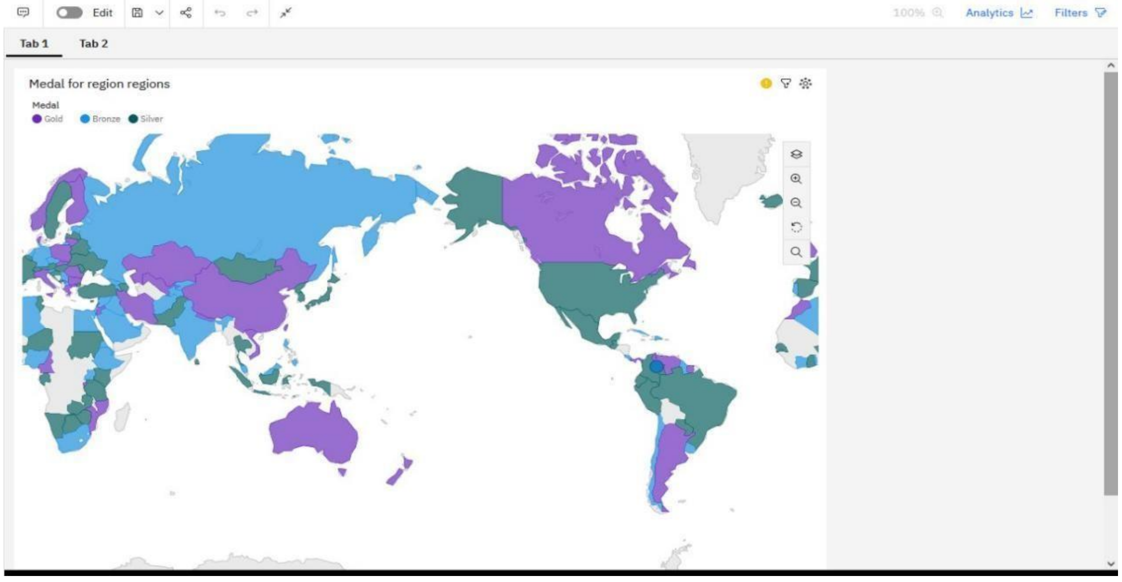
PERFORMANCE TESTING AND OUTPUT RESULTS :

9. RESULTS:

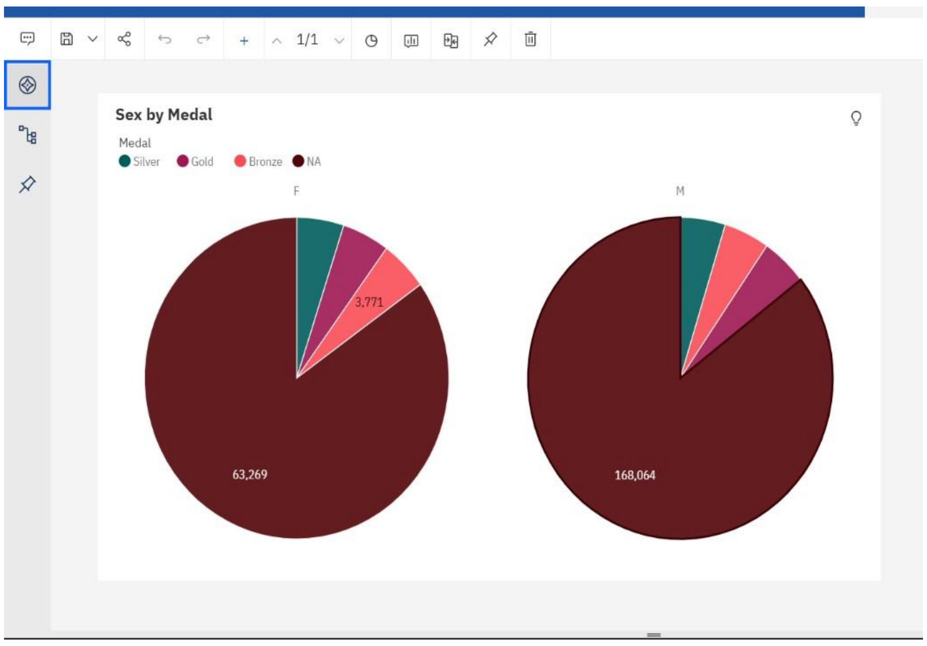
NO OF MEDALS WON BY YEAR:



NO OF MEDALS WON BY COUNTRIES:

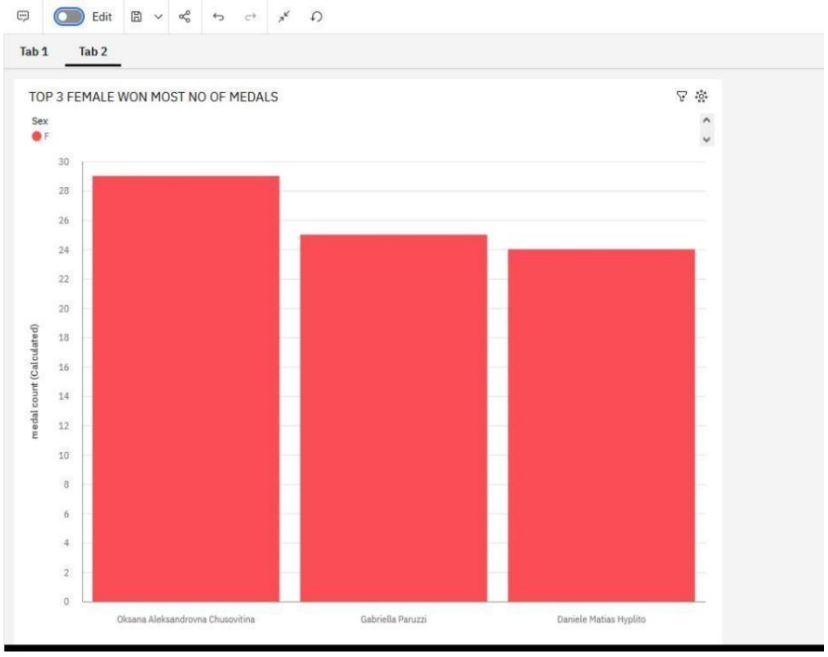


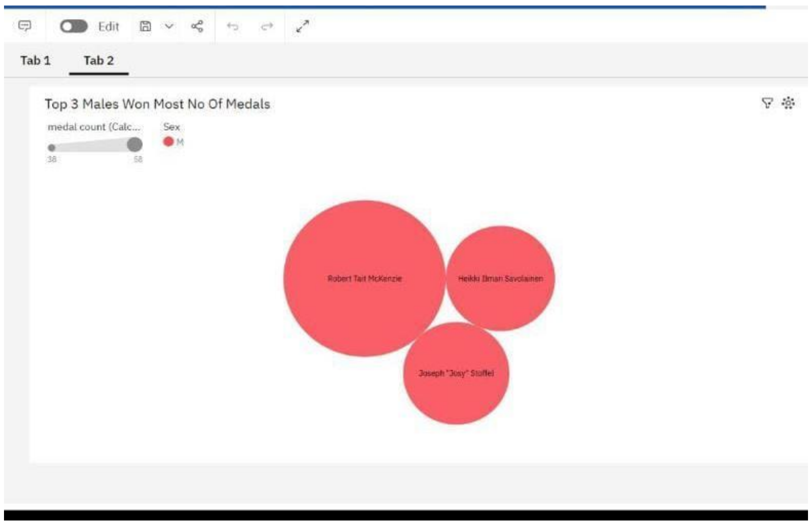
MALE AND FEMALE WON NO OF MEDALS:



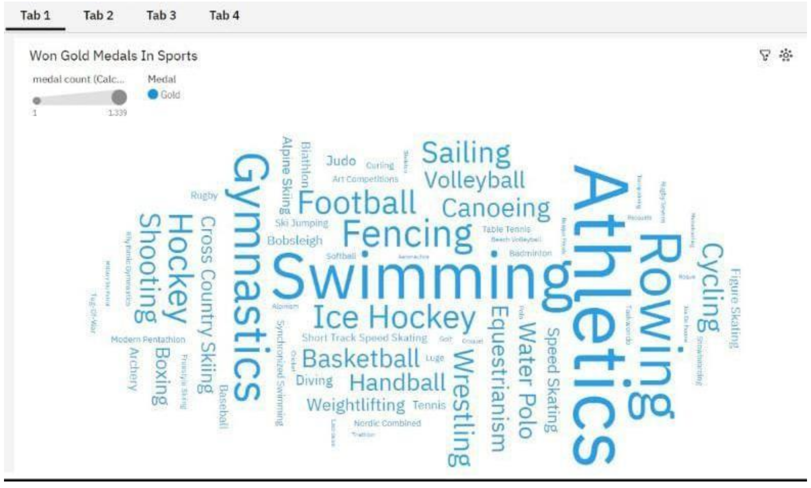
TOP 3 FEMALES WON MOST NO OF MEDALS:

TOP 3 MALE WON MOST NO OF MEDALS:

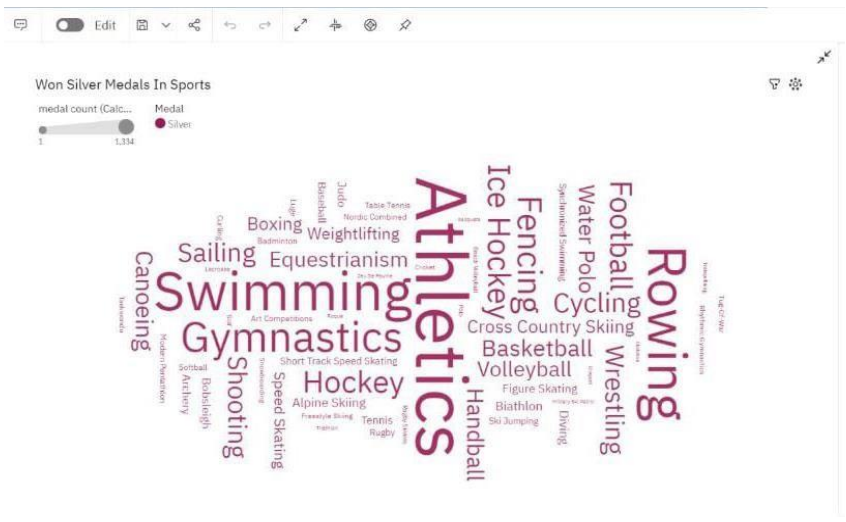




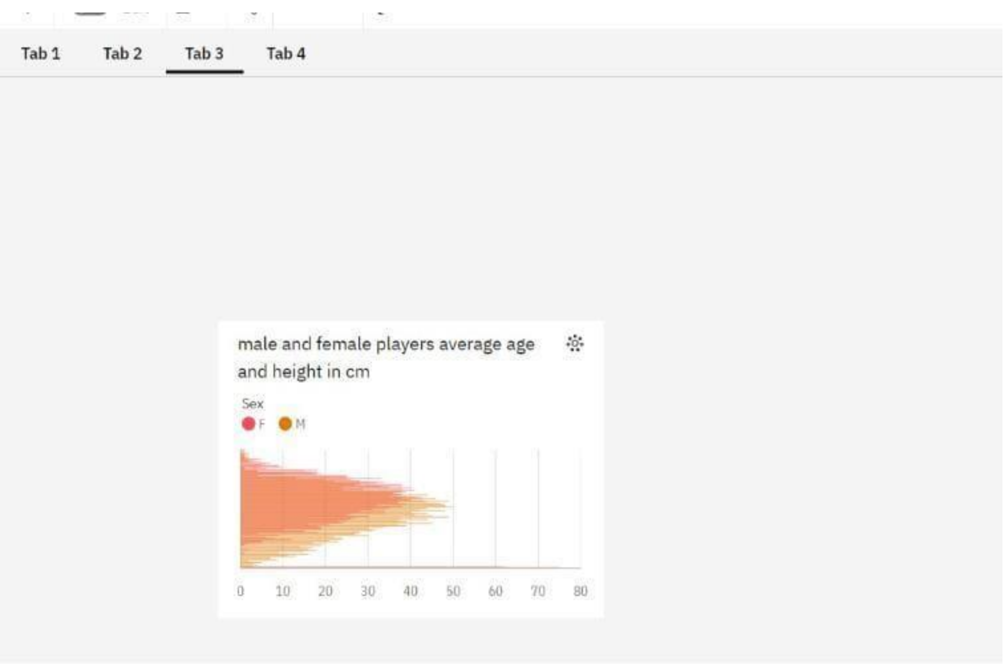
WON GOLD MEDALS IN SPORTS:



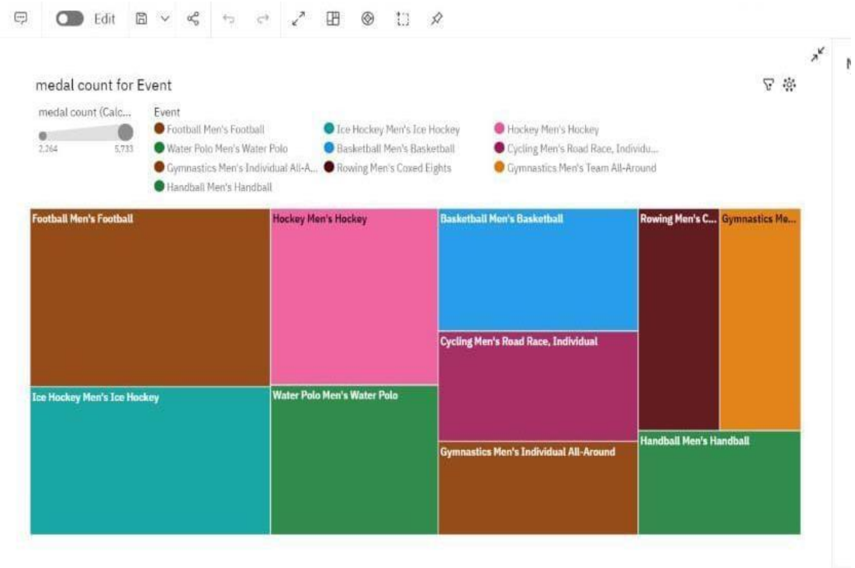
WON SILVER MEDALS IN SPORTS:



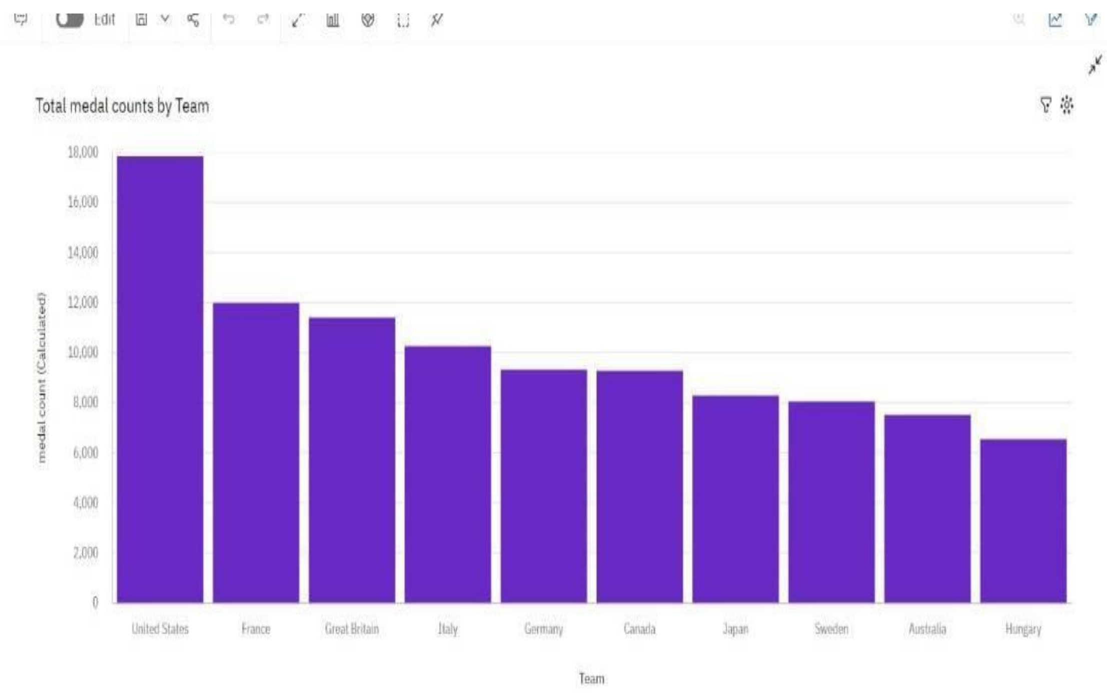
MALE AND FEMALES PLAYERS AVERAGE AGE AND HEIGHT IN CM:



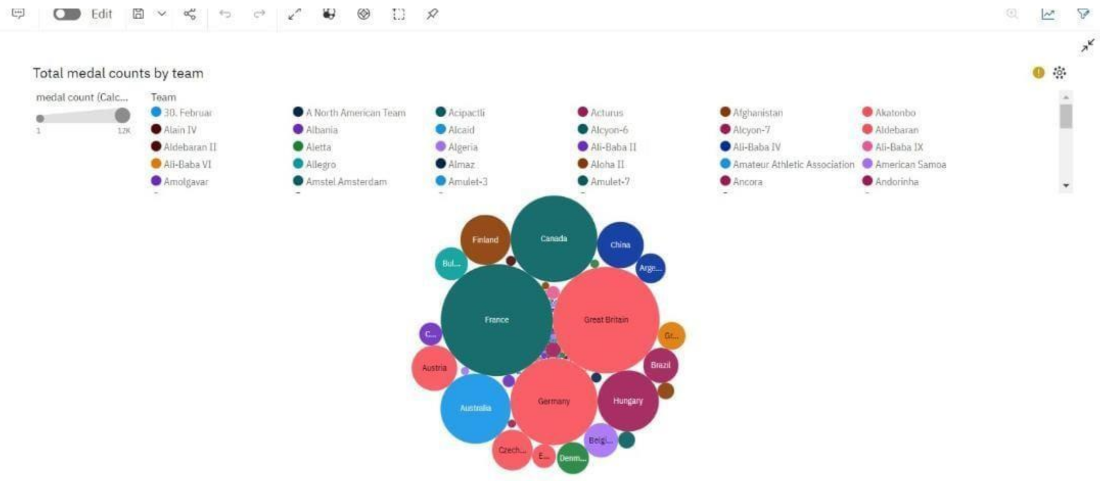
MEDAL COUNT BY EVENTS:


Edit
medal count for Event
medal count (Calc...
2,264
5,733
Event
Football Men's Football
Water Polo Men's Water Polo
Gymnastics Men's Individual All-A...
Handball Men's Handball
Ice Hockey Men's Ice Hockey
Basketball Men's Basketball
Rowing Men's Coxed Eights
Hockey Men's Hockey
Cycling Men's Road Race, Individu...
Gymnastics Men's Team All-Around
Football Men's Football
Hockey Men's Hockey
Basketball Men's Basketball
Rowing Men's C...
Gymnastics Me...
Cycling Men's Road Race, Individual
Ice Hockey Men's Ice Hockey
Water Polo Men's Water Polo
Gymnastics Men's Individual All-Around
Handball Men's Handball

GOLD,SILVER AND BRONZE MEDALS COUNT BY TEAM:



TOTAL MEDAL COUNTS BY TEAM:



10. ADVANTAGES AND DISADVANTAGES

Advantages:

1. Improved Performance Analysis: Data-driven insights provide coaches and athletes with detailed information about their performance. This can include metrics like speed, accuracy, endurance, and technique. Analysing this data helps identify strengths and weaknesses, enabling targeted training and performance improvement.
2. Injury Prevention: Data can be used to track an athlete's physical condition, workload, and fatigue levels. This information can help prevent overtraining and reduce the risk of injuries, which is crucial for athletes with rigorous training schedules.
3. Strategy Optimization: Coaches and teams can use data to develop and adjust game plans and strategies. This can involve studying opponent statistics, game simulations, and situational analysis to gain a competitive edge.
4. Talent Identification: Data analysis can assist in identifying talented athletes from a young age. Early recognition of potential can lead to more focused development and nurturing of future Olympic athletes.
5. Fan Engagement: Data-driven insights can enhance the viewer experience by providing real-time statistics, analysis, and visualizations, making Olympic sports more engaging for spectators.
6. Sponsorship and Revenue Generation: Access to data can attract sponsors and generate revenue through advertising, partnerships, and merchandise sales, which can be reinvested in sports development and participation.

Disadvantages:

1. Privacy Concerns: The collection and sharing of athlete data raise privacy issues. Athletes may be uncomfortable with the extent to which their personal and health information is collected and analysed.
2. Data Reliability: The accuracy and reliability of the data collected can be compromised by technical errors, equipment malfunctions, or human error. Relying on inaccurate data can lead to incorrect decisions.
3. Overemphasis on Metrics: Focusing too heavily on data-driven insights can shift the emphasis away from the human aspects of sports, such as passion, dedication, and teamwork. It can lead to a more mechanistic approach to training and competition.
4. Inequality: Access to sophisticated data analysis tools and technology can vary significantly between nations and teams. This can create disparities in training and competitive advantages, disadvantaging some athletes and nations.
5. Ethical Concerns: The use of data for talent identification from a young age can raise ethical concerns, such as pressuring children into highly competitive sports at an early age, potentially depriving them of a more well-rounded childhood.
6. Data Security: Data collected on athletes and teams can be vulnerable to security breaches and unauthorized access, potentially leading to leaks of sensitive information or tampering with performance data.

11. CONCLUSION

In conclusion, data-driven insights in Olympic sports are a double-edged sword, offering tremendous potential benefits while introducing complex challenges. Achieving the right balance between harnessing the power of data and safeguarding athlete rights, ethics, and the integrity of sports is an ongoing endeavor. As technology continues to advance, it is crucial to navigate these advantages and disadvantages thoughtfully to ensure that data-driven insights serve the best interests of athletes, coaches, and the broader Olympic community.

12. FUTURE SCOPE

The future of data-driven insights in Olympic sports is poised for significant growth. With advancements in analytics, wearable technology, and virtual reality, athletes will have access to more precise training methods and injury prevention. Fans can look forward to engaging experiences, and ethical considerations will play a crucial role in safeguarding athletes' rights and privacy. Equal access to technology and sustainability will be priorities, fostering a more inclusive and environmentally responsible Olympic movement. Collaboration and innovation will continue to drive the evolution of data-driven sports insights.

13. APPENDIX

SOURCE CODE GITHUB :

https://github.com/SharanSpidy/NM2023TMID03855?search=1

PROJECT DEMO LINK:

https://drive.google.com/file/d/1xBS245PdLIrj36uq_kAPb8jzWqidQyts/view?pli=1