NAAN MUDHALVAN PROJECT

DATA DRIVEN INSIGHTS ON OLYMPICS SPORTS PARTICIPATION AND PERFORMANCE:

PROJECT OVERVIEW

A project analyzing **Olympic sports participation and performance** is an excellent way to showcase your data science and analytical skills. Below is a structured approach to conceptualizing and implementing this project:

1. Define the Project Scope

Objective:
Explore trends, patterns, and insights related to Olympic sports participation and performance. Examples include:

- Country-wise performance trends over time.
- Gender participation analysis.
- Correlation between GDP/population and medals won.
- Analysis of dominant sports per country.

Target Output:

- Interactive dashboards.
- Statistical models predicting medal counts.
- A comprehensive report/blog showcasing findings.

---

2. Data Collection

Datasets:

1. **Primary Source:**
    - Kaggle - Olympic Sports Dataset
    - The dataset typically includes:
        - Athlete details (name, age, gender, country, sport).
        - Event details (sports, results, medals).
        - Historical data (from 1896 to recent Olympics).
2. **Supplementary Data:**

- **World Bank Data:** GDP, population, and HDI for countries.
- **Weather Data:** Impact of host city conditions on performance.
- **Travel Restrictions:** Impact of COVID-19 (for recent Olympics).

---

3. Tools and Libraries

- **Programming Language:** Python or R.
- **Libraries for Python:**
  - Data Manipulation: `pandas`, `numpy`
  - Visualization: `matplotlib`, `seaborn`, `plotly`
  - Machine Learning: `scikit-learn`
  - Dashboard Creation: `dash`, `streamlit`
- **Environment:** VS Code or Jupyter Notebook.

---

4. Workflow Breakdown

Step 1: Data Cleaning and Preprocessing

- Handle missing values (e.g., missing athlete ages or results).
- Standardize country names and codes.
- Create derived fields (e.g., medals per capita, medals by GDP).

python
Copy code

```
import pandas as pd

# Load dataset
df = pd.read_csv("athlete_events.csv")

# Check for missing data
print(df.isnull().sum())

# Fill missing age values with the median
df['Age'] = df['Age'].fillna(df['Age'].median())

# Add a derived column: Medals per capita
population_data = pd.read_csv("population.csv")
df = df.merge(population_data, on="Country", how="left")
```

```
df["Medals_per_Capita"] = df["Medal_Count"] / df["Population"]
```

Step 2: Exploratory Data Analysis (EDA)

- **Trends Over Time:**
  - Total number of participating nations, athletes, and events.
  - Medal distribution across years and countries.

python
Copy code

```
import matplotlib.pyplot as plt
import seaborn as sns

# Participation over time
years = df['Year'].unique()
participants = df.groupby('Year')['Athlete_ID'].nunique()

plt.figure(figsize=(10, 6))
sns.lineplot(x=years, y=participants)
plt.title("Athlete Participation Over Time")
plt.xlabel("Year")
plt.ylabel("Number of Participants")
plt.show()
```

Top Performing Nations: Identify countries with the highest medal tallies.
python
Copy code

```
top_countries =
df.groupby('Country')['Medal'].count().sort_values(ascending=False).he
ad(10)
top_countries.plot(kind='bar', figsize=(10, 6), title="Top 10
Countries by Medal Count")
```

-

Step 3: Statistical Analysis

- **Factors Influencing Performance:**
  - Use correlation and regression analysis to find the relationship between GDP, population, and medal count.

python
Copy code

```
from scipy.stats import pearsonr

gdp_medal_corr = pearsonr(df['GDP'], df['Medal_Count'])
print(f"Correlation between GDP and Medal Count: {gdp_medal_corr[0]}")
```

-

Step 4: Machine Learning

Predict medal count based on socio-economic indicators (e.g., GDP, population, previous performance).

python
Copy code

```
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split

# Prepare data
X = df[['GDP', 'Population']]
y = df['Medal_Count']
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.2, random_state=42)

# Train model
model = LinearRegression()
model.fit(X_train, y_train)

print(f"Model R^2: {model.score(X_test, y_test)}")
```

-

Step 5: Visualization

- Create dashboards showing:
  - Medal distribution by country, gender, and sport.
  - Historical trends in participation and performance.
  - Predictive analysis (e.g., medals in future Olympics).

For example, using **Plotly Dash**:

python
Copy code

```
import dash
from dash import dcc, html

app = dash.Dash()

app.layout = html.Div([
    dcc.Graph(
        id='medal-chart',
        figure={
            'data': [
                {'x': years, 'y': participants, 'type': 'line',
'name': 'Participation'}
            ],
            'layout': {
                'title': 'Olympic Participation Over Time'
            }
        }
    )
])

if __name__ == '__main__':
    app.run_server(debug=True)
```

---

5. Expected Insights

1. **Participation Trends:**
    - Growth in participation from developing nations.
    - Increased female participation over the years.
2. **Country Dominance:**
    - Identification of sport-specific dominance (e.g., USA in swimming, Kenya in long-distance running).
3. **Economic Indicators:**
    - Strong correlation between GDP and medal count, with exceptions (e.g., Cuba, Ethiopia).
4. **Host Nation Advantage:**
    - Quantify the "home advantage" for host nations.

---

6. Presentation

- **Final Report/Presentation:**
  Include:
    - Overview of key findings.
    - Graphs and visualizations.
    - Recommendations based on insights.
- **Deployment:**
  Host dashboards or reports online using Streamlit, Flask, or platforms like Heroku.

Would you like help setting up any specific part of this project?

---

7. Conclusion

1. **Rising Participation**: Efforts to make the Olympics inclusive have led to increased global participation, especially from developing countries.
2. **Gender Parity Progress**: Female representation has significantly improved, but disparities still exist in some sports.
3. **Wealth Disparity**: Economic factors heavily influence performance, highlighting the need for equitable resource distribution.