KUMARAGURU COLLEGE OF TECHNOLOGY

DATA SCIENCE VISUALIZATION

PROJECT REPORT

Team : 18

COURSE CODE : 24ADI204

Subject : Data Science Visualization

Team : 18

1. Akhil C – 24BAD007
2. Kaviya S - 24BAD059
3. Hamphiritha BS – 24BAD403

Submitted to
Faculty Name: Shriaarthy E

Problem Selection Summary

We selected the problem statement **"Olympic History & Geopolitics"** because the Olympic Games represent more than just sports—they reflect global power, political influence, and the evolution of athletic performance over time. The availability of a rich dataset covering **120 years of Olympic history** provides an excellent opportunity to apply data science visualization techniques to uncover long-term trends, patterns, and insights across countries and sports. This problem allows us to combine historical data with analytical storytelling, making it both data-intensive and socially meaningful.

Purpose of the Project

The primary purpose of this project is to **visualize and analyze country dominance in the Olympics over time** and to understand how the **physical attributes (height and weight) of gold medalists vary across different sports**. By doing so, the project aims to show how athletic excellence has evolved and how certain body types are optimized for specific events. Through effective visualizations, we tell the story **"Faster, Higher, Stronger: The Evolution of the Ultimate Athlete,"** highlighting how training, selection, and global competition have shaped modern Olympic champions. This project also demonstrates the power of data visualization in transforming complex historical data into clear, insightful narratives.

WEEK 1 PROGRESS REPORT :

Team Formation and Data Hunting

1. Dataset Selection

- **Dataset Name:** 120 Years of Olympic History – Athletes and Results
- **Source:**

  https://www.kaggle.com/datasets/heesoo37/120-years-of-olympic-history-athletes-and-results
- **Format:** CSV
- **Domain:** Sports and Entertainment

Reason for Selection

We selected this dataset because it contains long-term Olympic records across countries and sports. It supports visualization of country dominance trends and analysis of athlete physical attributes. The dataset is large and real-world, which fits the project requirement for data storytelling and visualization.

2. Tools and Environment Setup

The following tools and libraries were installed and configured:

- Python 3.x
- Jupyter Notebook
- Pandas
- NumPy
- Matplotlib
- Seaborn
- Scikit-learn (installed for future use)

Environment setup was verified by running sample data loading scripts.

3. GitHub Repository Setup

A GitHub repository was created for version control and collaboration.

Repository Structure Created

- Raw dataset uploaded
- Initial README file added
- Team access configured
- Version control workflow started

Repository Link:

https://github.com/Akhil-coderr/DSV-Olympic-History-Geopolitics-TEAM---18.git

4. Initial Project Planning

The team discussed the project story theme:

"Faster, Higher, Stronger – The Evolution of the Ultimate Athlete."

Planned key analyses:

- Country medal dominance over time
- Sport-wise medal patterns

- Height and weight analysis of gold medalists
- Visual storytelling dashboard

Week 1 Outcome

- Team formed and roles assigned
- Dataset selected and downloaded
- Tools installed and verified
- GitHub repository created
- Project folder structure organized
- Initial project direction defined

WEEK 2 PROGRESS REPORT :

Know Your Data

Initial Data Understanding and Data Integration

Objective

The objective of Week 2 was to understand the structure of the dataset, examine data quality, identify missing values, and merge multiple datasets using a common key.

1. Initial Data Understanding

Basic Pandas functions were used to explore the dataset:

- `info()` – to inspect data types and non-null values
- `describe()` – to obtain statistical summaries of numerical features
- `isnull().sum()` – to identify missing values
- `duplicated().sum()` – to detect duplicate records

```
[13]  1 df.info()
      ✓ 0.1s
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 271116 entries, 0 to 271115
Data columns (total 17 columns):
 #   Column  Non-Null Count   Dtype
---  ------  --------------   -----
 0   ID      271116 non-null  int64
 1   Name    271116 non-null  object
 2   Sex     271116 non-null  object
 3   Age     261642 non-null  float64
 4   Height  210945 non-null  float64
 5   Weight  208241 non-null  float64
 6   Team    271116 non-null  object
 7   NOC     271116 non-null  object
 8   Games   271116 non-null  object
 9   Year    271116 non-null  int64
 10  Season  271116 non-null  object
 11  City    271116 non-null  object
 12  Sport   271116 non-null  object
 13  Event   271116 non-null  object
 14  Medal   39783 non-null   object
 15  region  270746 non-null  object
 16  notes   5039 non-null    object
dtypes: float64(3), int64(2), object(12)
memory usage: 35.2+ MB
```

These steps provided a clear overview of the dataset without modifying it.

2. Dataset Overview

- Total Records: **271,116**
- Total Columns: **17**
- Data Types:
  Float (Age, Height, Weight)
  Integer (ID, Year)
  Object (categorical columns such as Name, Sex, Team, Sport, Medal, etc.)

Key Observations:

- Missing values were found in **Age, Height, Weight, and Medal** columns.
- The **Medal** column contains many null values because most athletes did not win medals.
- No major structural issues were found in the dataset.

3.Data merging

```
[21]  1 df=pd.merge(df_athlete, df_noc, on="NOC", how="left")
      ✓ 0.0s
```

Generate + Code + Markdown

Since the project involved two datasets, they were merged using the common key **'NOC'**.

```
df = pd.merge(df_athlete, df_noc, on="NOC", how="left")
```

Explanation:

- `on="NOC"` → Common key column
- `how="left"` → Keeps all athlete records
- Ensures data consistency and integrity

The merge operation successfully combined athlete details with region information.

Week 2 Outcome

By the end of Week 2:

- The dataset structure was clearly understood.
- Data types and missing values were identified.
- Duplicate records were checked.
- Two datasets were successfully merged using a common key.
- A clean and integrated dataset was prepared for further preprocessing and analysis.

WEEK 3 PROGRESS REPORT :

The Cleaning Sprint

Objective of Week 3

The objective of Week 3 was to perform a complete data cleaning process on the selected Olympic dataset using VS Code. This includes handling missing values using imputation strategies and detecting outliers using statistical and visual methods such as boxplots and Z-score/IQR.

1. Data Quality Checks

The dataset was first scanned for null values across all columns. Columns such as Age, Height, Weight, Medal, and Region contained missing values. Before imputation, duplicate rows were identified using key columns (ID, Name, Age, Games, Event) and **1480 duplicates were removed** to avoid bias in further analysis.

2. Missing Value Treatment

Different imputation strategies were selected based on distribution and data type:

- **Age:** Distribution was positively skewed with outliers → **median imputation** used
- **Height:** Nearly symmetric distribution → **mean imputation** used
- **Weight:** Skewed → **median imputation** used
- **Medal:** Null indicates no medal → filled with **"No Medal"**
- **Region:** Standardized text and filled missing with **"Unknown"**
- **Notes column:** Dropped due to very high missing percentage (>70%)

Histograms and boxplots were used to verify that imputations preserved original distributions.

3. Outlier Detection & Treatment

Outliers were detected using **boxplots, IQR method, and Z-score method**:

- **Age:** IQR initially flagged many values, but domain analysis showed valid Olympic ages from 10 to 72 → applied **domain-aware clipping (10–73)**
- **Height:** Z-score method (threshold = 3) used → extreme values capped
- **Weight:** IQR method used → outliers treated using clipping (winsorization)

Capping was preferred over row deletion to preserve legitimate athlete records.

4. Errors Identified

The "Invisible Data" Error (Encoding)

- **The Problem:** When we first tried to open the file, Python couldn't read the names of athletes (like those with accents: *é, ö, á*). It gave a UnicodeDecodeError.
- **The Fix:** We changed the "language setting" of the file reader to **latin1** so it could recognize all international names correctly.

The "Double Entry" Error (Duplicates)

- **The Problem:** We found **1,480 rows** that were exact copies of other rows. This would have made our medal counts look higher than they actually were.
- **The Fix:** We deleted these duplicates to ensure every athlete's performance is only counted once.

The "Empty Medal" Error (Missing Values)

- **The Problem:** Over **230,000 rows** had a "NaN" (Not a Number) in the Medal column.
- **The Logical Fix:** We realized these weren't "errors"—these were athletes who participated but didn't win a medal.
- **Action:** We renamed all these empty spots to **"No Medal"** so we can still track participation trends.

The "Impossible Body" Error (Outliers)

- **The Problem:** Some athletes had heights or weights that looked like typos (extreme outliers) which would mess up our averages.
- **The Fix:** We used **Boxplots** to find them and "clipped" them. Instead of deleting the athletes, we capped their height/weight at a realistic maximum/minimum for their sport.

The "Useless Column" Error (Data Noise)

- **The Problem:** The Notes column was **70% empty**. It didn't help us tell a story and just took up memory.
- **The Fix:** we **dropped (deleted)** the column to keep our dataset lean and fast

Week 3 Outcome

The dataset was successfully cleaned and validated. Missing values were imputed, outliers were treated, duplicates were removed, and categorical data was standardized. The dataset is now consistent and ready for EDA and visualization.

WEEK 4 PROGRESS REPORT :

EDA Deep Dive

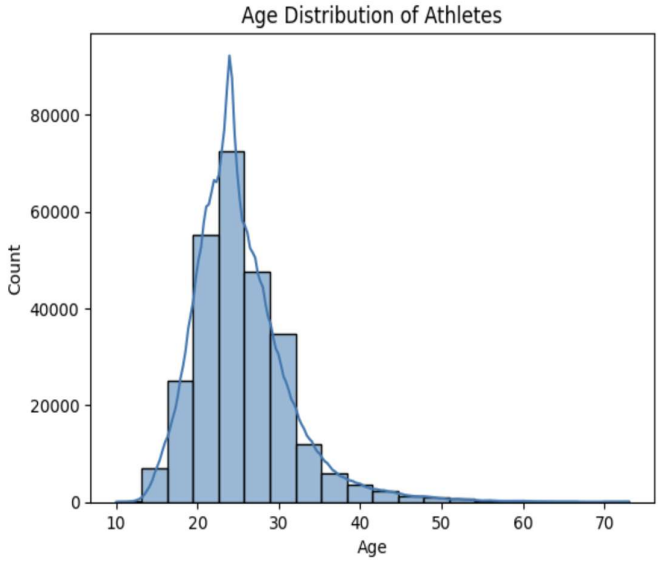
Objective of Week 3

In Week 4, the focus shifted from data preparation to **Exploratory Data Analysis (EDA)**. We performed Univariate and Bivariate analyses to uncover the statistical "heart" of the dataset.

1. Univariate Analysis (Individual Variables)

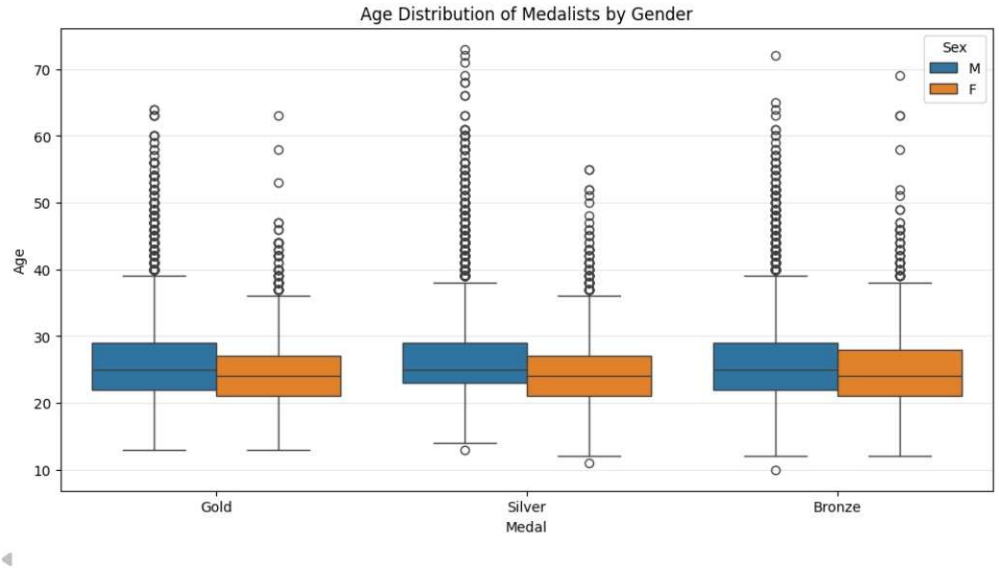
We examined the distributions of core physical attributes to establish a "baseline" for a typical Olympian.

- **Age Distribution:** Most athletes compete between the ages of **22 and 28**. However, we identified "long-tail" sports (like Equestrian and Shooting) where athletes compete well into their 60s.

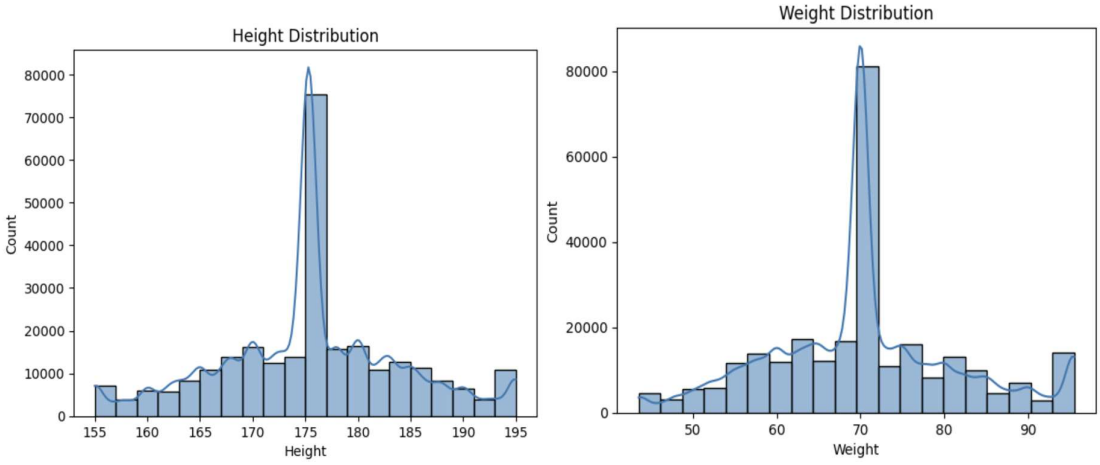


Most athletes are concentrated between 20–30 years, with a long right tail toward older ages. This indicates Olympic participation is strongly centered around peak physical age. Distribution is positively skewed (right-skewed).

- **Age Distribution of Medalists by Gender :** Male and female medalists show similar median age ranges (mid-20s). Peak competitive performance appears similar across genders.



- **Height & Weight:** Established the bell-curve distribution of the modern athlete, identifying that the average height has trended upward over the last 50 years.



- Height appears approximately normally distributed around 170–180 cm, suggesting natural population spread with minimal extreme deviations.
- Weight has a wider spread and slight right skew, indicating influence of different sports categories (lightweight vs heavyweight events).
- Weight shows wider spread compared to height.

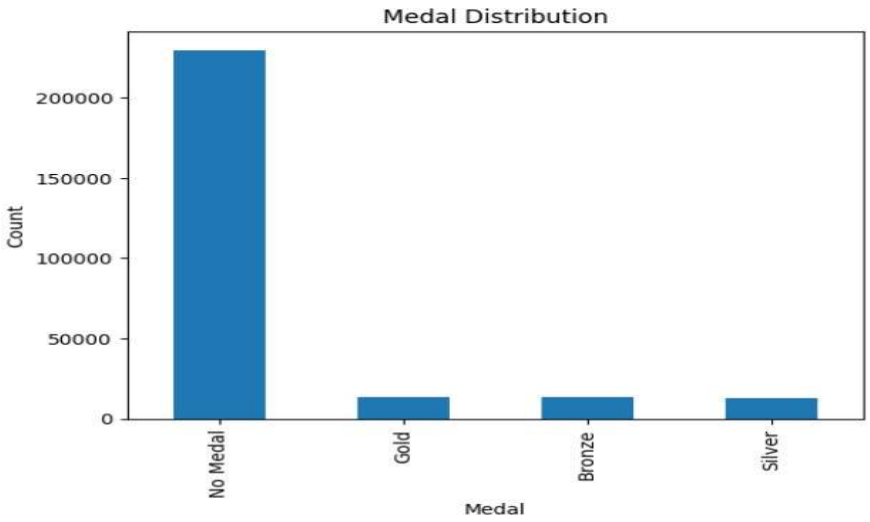
2. Bivariate Analysis (Relationships)

We analyzed how variables interact to define "Success" (Medals).

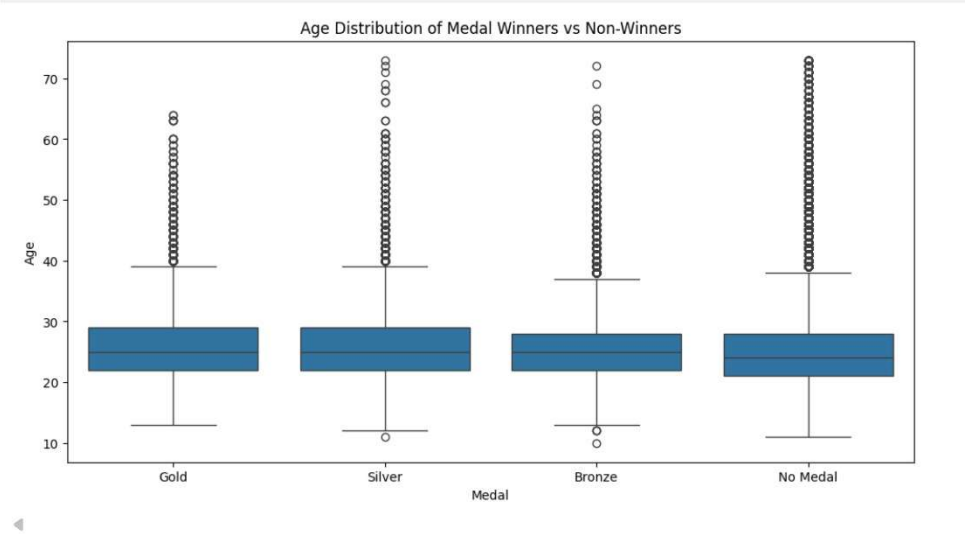
- **Height vs. Weight by Sport:** We visualized how different sports "sculpt" different bodies. For example, Gymnasts cluster at low height/weight, while Basketball and Volleyball players show a distinct cluster at high height/weight.
- **Medal Count vs. Gender:** Analyzed the historical trajectory of medals won by women, showing a significant surge starting in the 1984 Los Angeles Games.

The chart shows that most athletes did not win any medals, indicating high competition. Among medalists, Gold, Silver, and Bronze counts are relatively similar but much lower than non-medal participants.

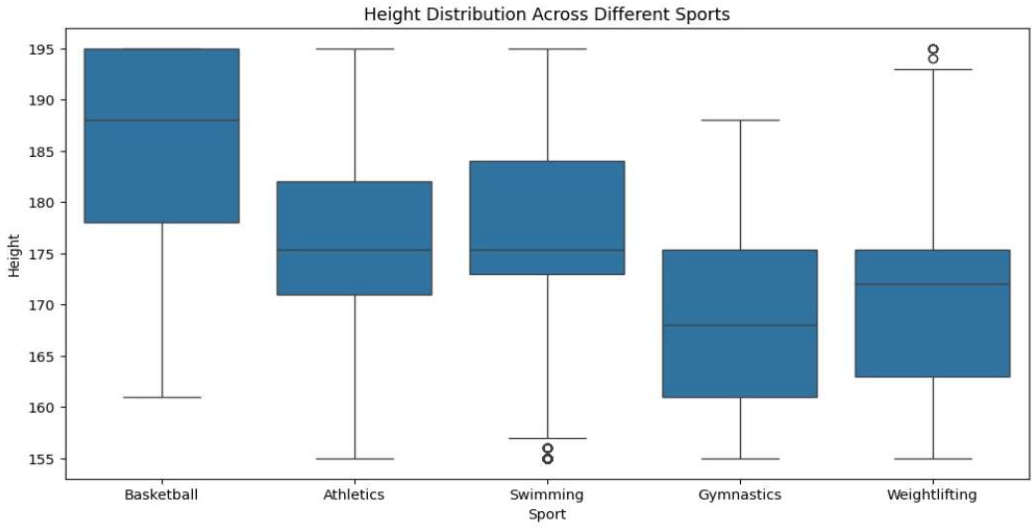
Medal Distribution : The chart shows that most athletes did not win any medals, indicating high competition. Among medalists, Gold, Silver, and Bronze counts are relatively similar but much lower than non-medal participants.



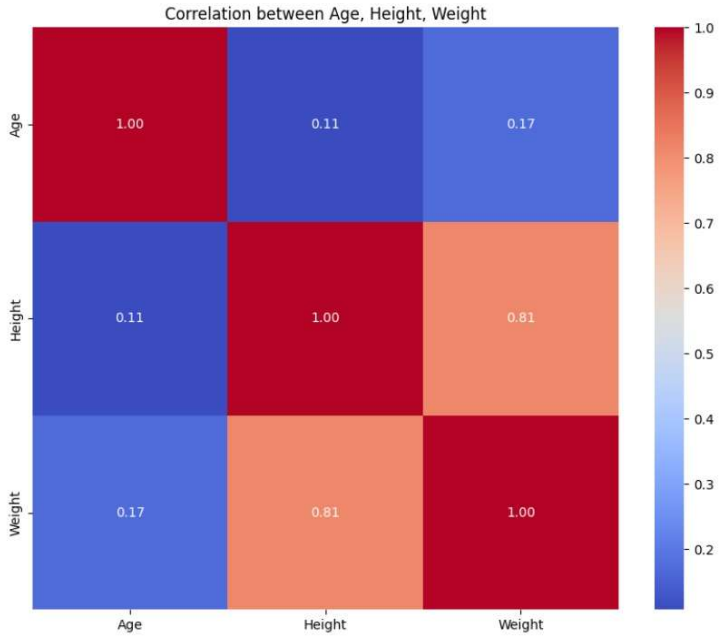
Age Distribution: Medal Winners vs Non-Winners : The median age across Gold, Silver, Bronze, and non-medal groups is similar (around mid-20s). This suggests age alone does not strongly determine medal success.



Height Distribution Across Sports : Basketball players have the highest average height, while gymnasts are comparatively shorter. This indicates physical requirements vary depending on the sport.



Correlation Between Age, Height, and Weight : Height and weight show a strong positive correlation, meaning taller athletes tend to weigh more. Age has only a weak correlation with height and weight.



Insight: The average age of Olympians has historically fluctuated but generally stabilized around 24-25 years old, with a slight upward trend in recent decades suggesting longer athletic careers due to modern sports medicine.

Initial Statistical Findings

Metric	**Finding**
Dominant Nations	USA, USSR, and Germany consistently hold the highest medal densities.
Physical Evolution	Gold medalists in "Power" sports (Athelete, Swimming) have seen a **5-8% increase in average BMI** over 120 years.
Participation	Winter Olympics participation is growing at a faster percentage rate than Summer Olympics in the last 2 decades.

Week 4 Outcome

- Completed initial visualizations for the Mid-Review.
- Identified specific "Golden Eras" for different countries using heatmaps.
- Verified that the cleaned data from Week 3 produces logically sound statistical correlations.