Name: Muhammad Ilham Siddiqqulhakim

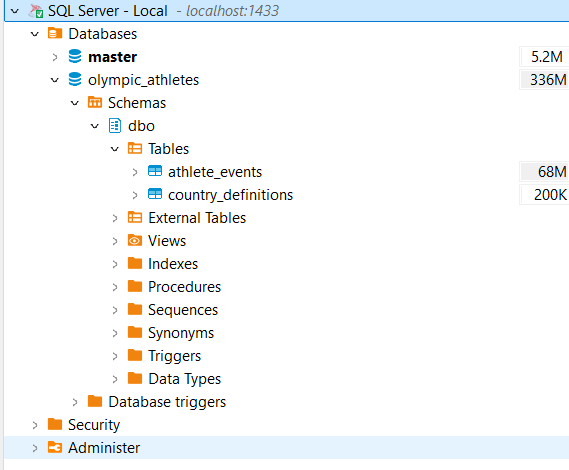
Role: Data Engineer

# Data Engineer Test

## Dataset

1. SQL Server

The tool that I use to manage the database are SQL Server Management Studio and DBaver



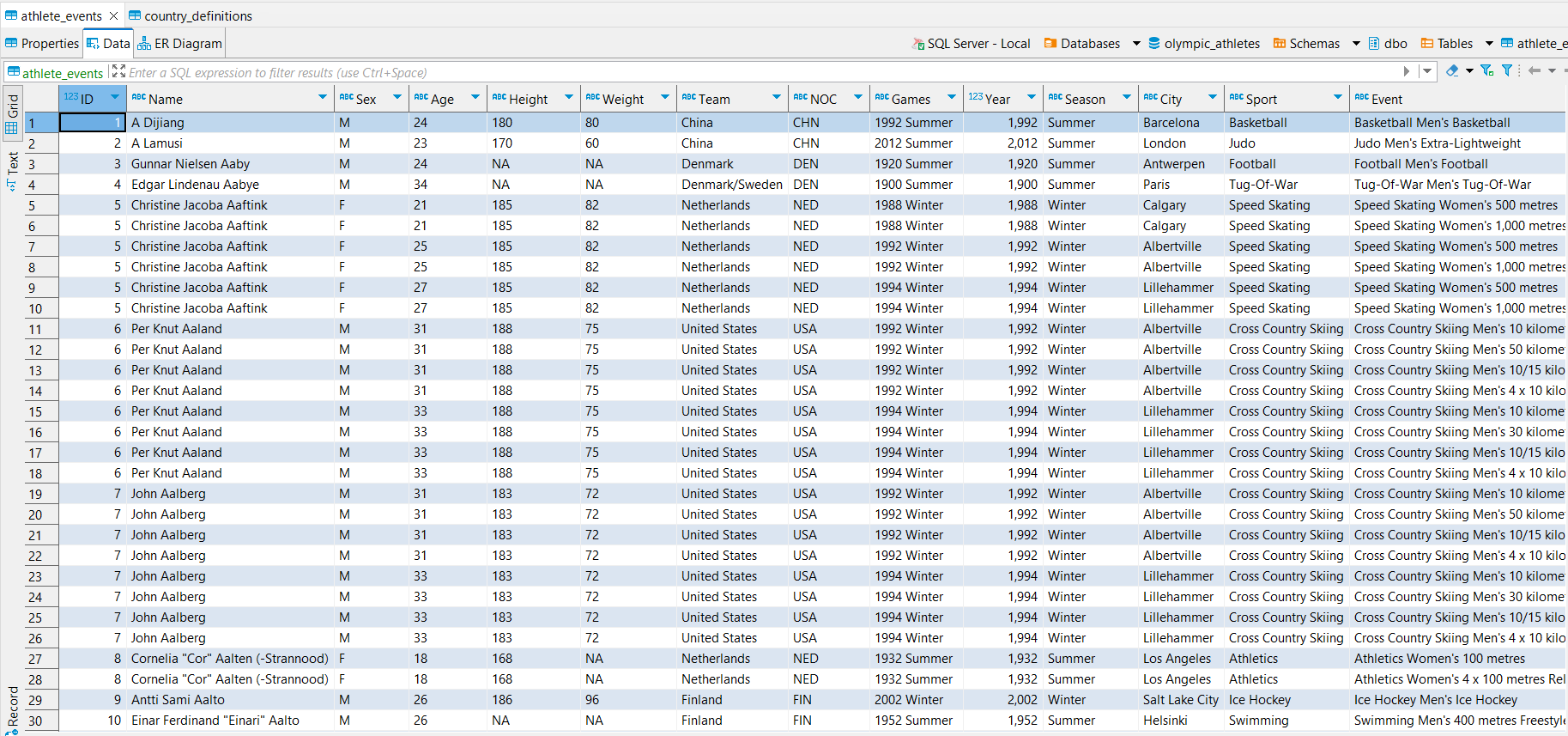
1. I use the dataset 120 Years of Olympic History from <https://mavenanalytics.io/data-playground>. This dataset talks about athletes who participated in the Olympics over 120 years (from 1896 to 2016). Below is the definition of each column:
2. athlete\_events table

| Field | Description |
| --- | --- |
| ID | Unique number for each athlete |
| Name | Athlete's name |
| Sex | Male (M) or Female (F) |
| Age | Integer |
| Height | In centimeters |
| Weight | In kilograms |
| Team | Team name |
| NOC | National Olympic Committee 3-letter code |
| Games | Year and season |
| Year | Integer |
| Season | Summer or Winter |
| City | Host city |
| Sport | Sport |
| Event | Event |
| Medal | Gold, Silver, Bronze, or NA |

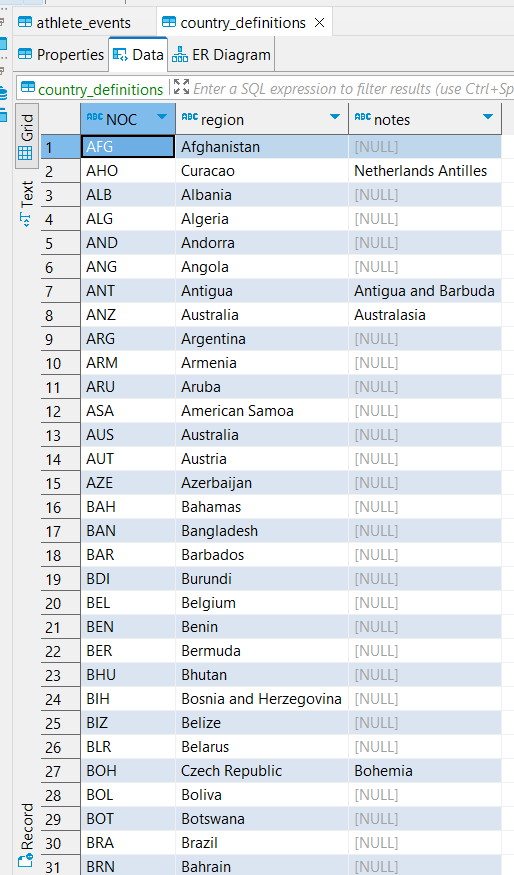
1. country\_definitions

| Field | Description |
| --- | --- |
| NOC | National Olympic Committee 3 letter code |
| region | Country name used for geospatial mapping |
| notes | Real country name if "region" isn't an exact match |

1. The process to import it the dataset into sql server is:
   1. Download dataset
   2. Extract it
   3. Open/Create database
   4. Using SSMS, right-click on the database, choose Task, then select Import Flat File
   5. The SSMS will guide you for the next process
2. Dataset
   1. athlete\_events



* 1. country\_definitions



## ETL Code

Below is the ETL code (I also give comments in the Python file).

### Dependencies

# import dependencies

import pyodbc

import pandas as pd

### Load data

# Database connection

conn = pyodbc.connect(

'DRIVER={ODBC Driver 17 for SQL Server};'

'SERVER=localhost,1433;'

'DATABASE=olympic\_athletes;'

'Trusted\_Connection=yes;'

)

# Query to fetch the data

athlete\_query = "SELECT \* FROM athlete\_events"

country\_query = "SELECT \* FROM country\_definitions"

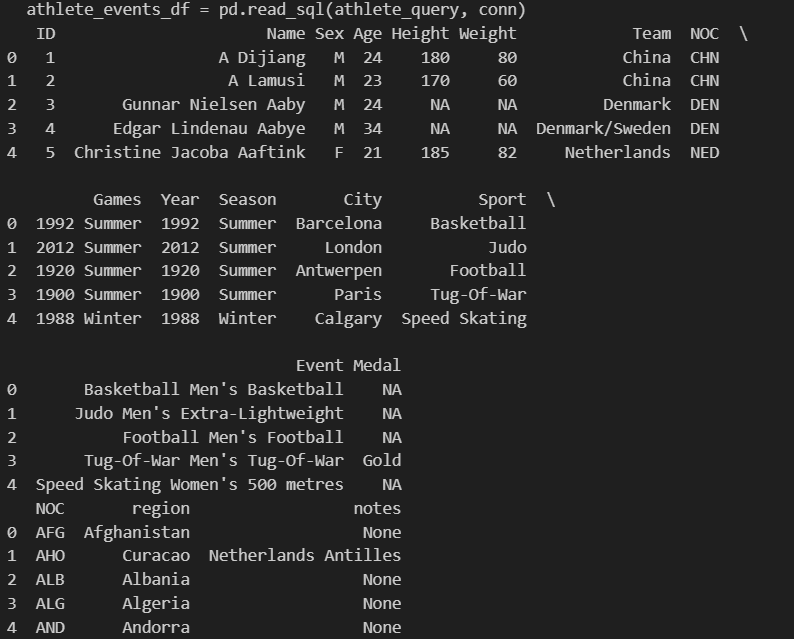
# Load data into pandas DataFrame

athlete\_events\_df = pd.read\_sql(athlete\_query, conn)

country\_def\_df = pd.read\_sql(country\_query, conn)

print(athlete\_events\_df.head())

print(country\_def\_df.head())



### Cleansing/Preparation

# 1. Handle Missing Values

# Replace 'NA' strings with actual NaN values for proper handling

athlete\_events\_df.replace('NA', pd.NA, inplace=True)

# 2. Data Type Conversion

# Convert numerical columns to the appropriate data types

athlete\_events\_df['Age'] = pd.to\_numeric(athlete\_events\_df['Age'], errors='coerce')

athlete\_events\_df['Height'] = pd.to\_numeric(athlete\_events\_df['Height'], errors='coerce')

athlete\_events\_df['Weight'] = pd.to\_numeric(athlete\_events\_df['Weight'], errors='coerce')

# 3. Remove Duplicates

# Drop duplicate records

athlete\_events\_df.drop\_duplicates(inplace=True)

# 4. Join with country definition data

# Merge the two datasets to get the region information

merged\_df = pd.merge(athlete\_events\_df, country\_def\_df, on='NOC', how='left')

merged\_df.head()



Since the data is already "good", so I just need to do a simple cleansing/transformation

### Data Summarization

#### Metric 1: Medal Summary by Country and Type

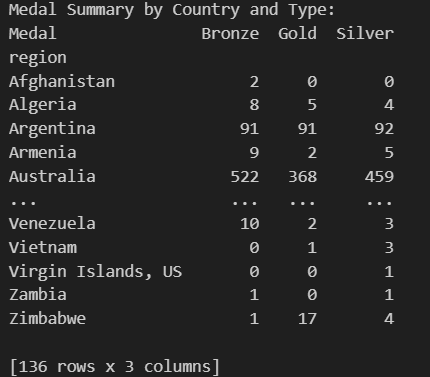
This metric shows medal achievement for each country by all types of medal

# Metric 1: Medal Summary by Country and Type

medal\_summary = merged\_df.groupby(['region', 'Medal']).size().unstack(fill\_value=0)

print("Medal Summary by Country and Type:")

print(medal\_summary)



#### Metric 2: Average Age, Height, and Weight by Country

This metric shows the average age, height, and weight of the athletes for each country

# Metric 2: Average Age, Height, and Weight by Country

athlete\_summary = merged\_df.groupby('region').agg({

'Age': 'mean',

'Height': 'mean',

'Weight': 'mean'

}).reset\_index()

print("Athlete Summary by Country (Average Age, Height, Weight):")

print(athlete\_summary)



#### Metric 3: Gender Distribution by Country

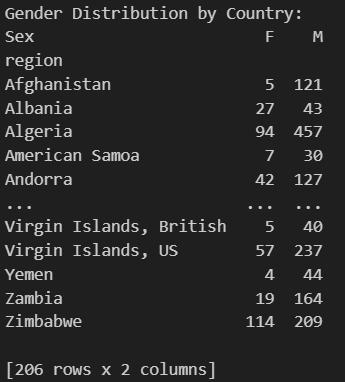
This metric shows how many Male and Female athletes that each country sent to Olympic

# Metric 3: Gender Distribution by Country

gender\_summary = merged\_df.groupby(['region', 'Sex']).size().unstack(fill\_value=0)

print("Gender Distribution by Country:")

print(gender\_summary)



#### Metric 4: Total Number of Athletes Participating in Each Sport by Country

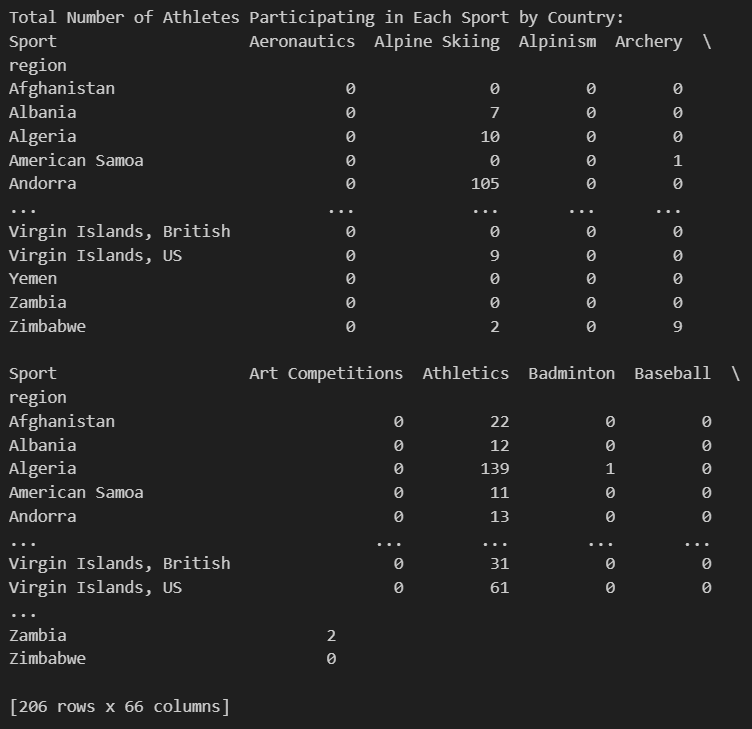
This metric shows how many athletes that participate in each sport by country

# Metric 4: Total Number of Athletes Participating in Each Sport by Country

sport\_summary = merged\_df.groupby(['region', 'Sport']).size().unstack(fill\_value=0)

print("Total Number of Athletes Participating in Each Sport by Country:")

print(sport\_summary)



### Finishing

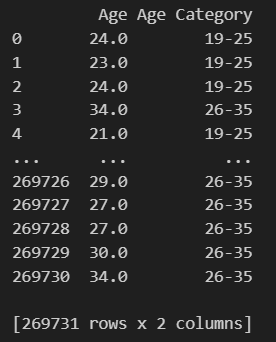
# Add Age Category For Dashboard

bins = [0, 18, 25, 35, 45, 100]

labels = ['0-18', '19-25', '26-35', '36-45', '46+']

merged\_df['Age Category'] = pd.cut(merged\_df['Age'], bins=bins, labels=labels, right=False)

print(merged\_df[['Age', 'Age Category']])



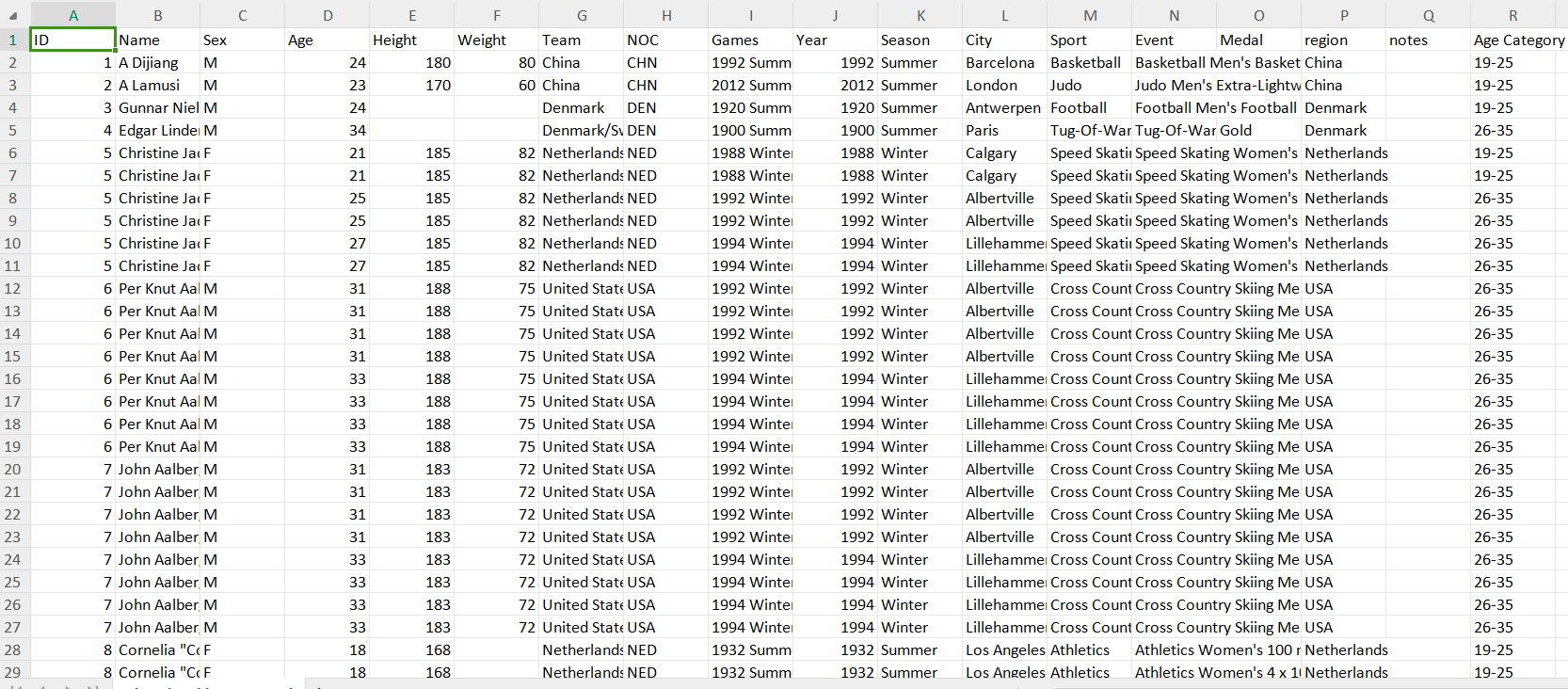
# Export to CSV for Power BI

merged\_df.to\_csv("olympic\_athletes\_merged.csv", index=False)

# Close the connection

conn.close()

### Result



## Dashboard/Report

### Overview Page

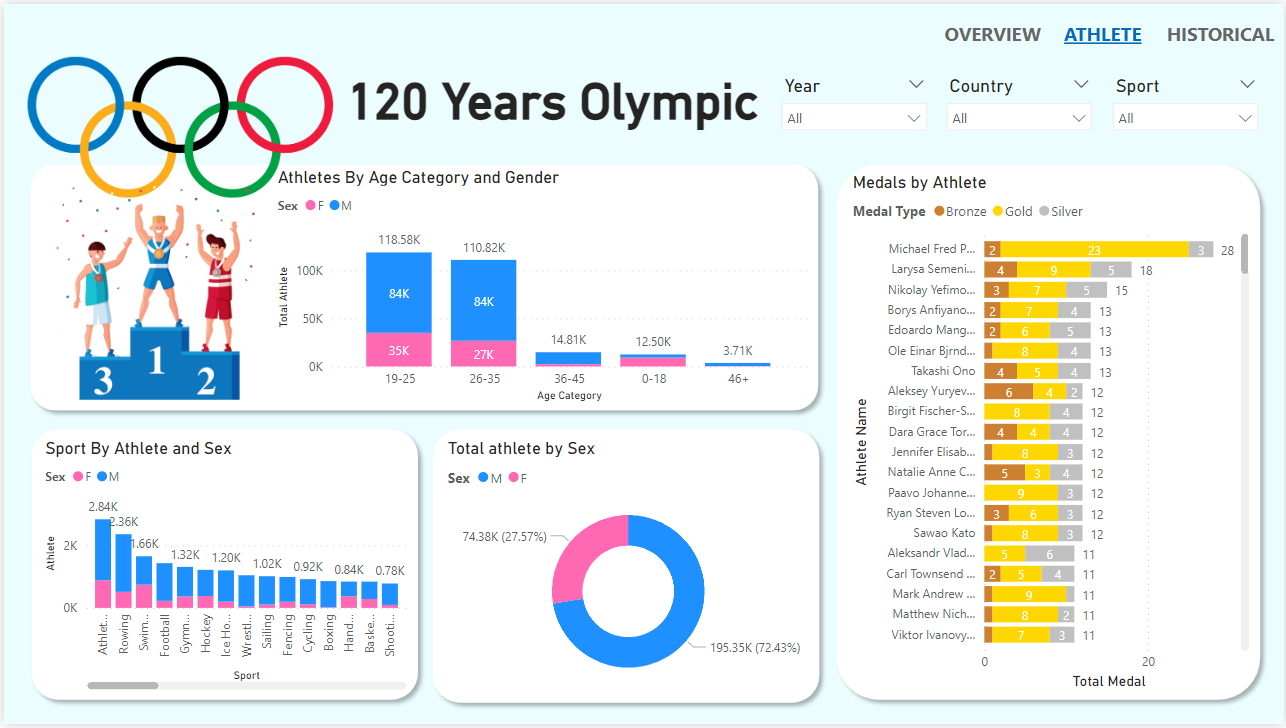


Over the past 120 years of the Olympic Games, a total of 134,731 athletes from 207 regions around the world have competed in 66 different sports, from 1896 to 2016. The USA stands as the highest-performing country in Olympic history, amassing an impressive total of 5,634 medals: 2,638 gold, 1,641 silver, and 1,358 bronze.

The dashboard reveals that most of the events have been held in the Summer Olympics, which account for 56.86%, while Winter is 43.14%. The timeline of Olympic events shows a significant increase over the years, from just 43 events in the early 1900s to over 300 events in recent decades, showcasing the growth and evolution of the Games.

You can filter this page by Year, Country, and also Sport using the slicer provided above the page.

### Athlete Page

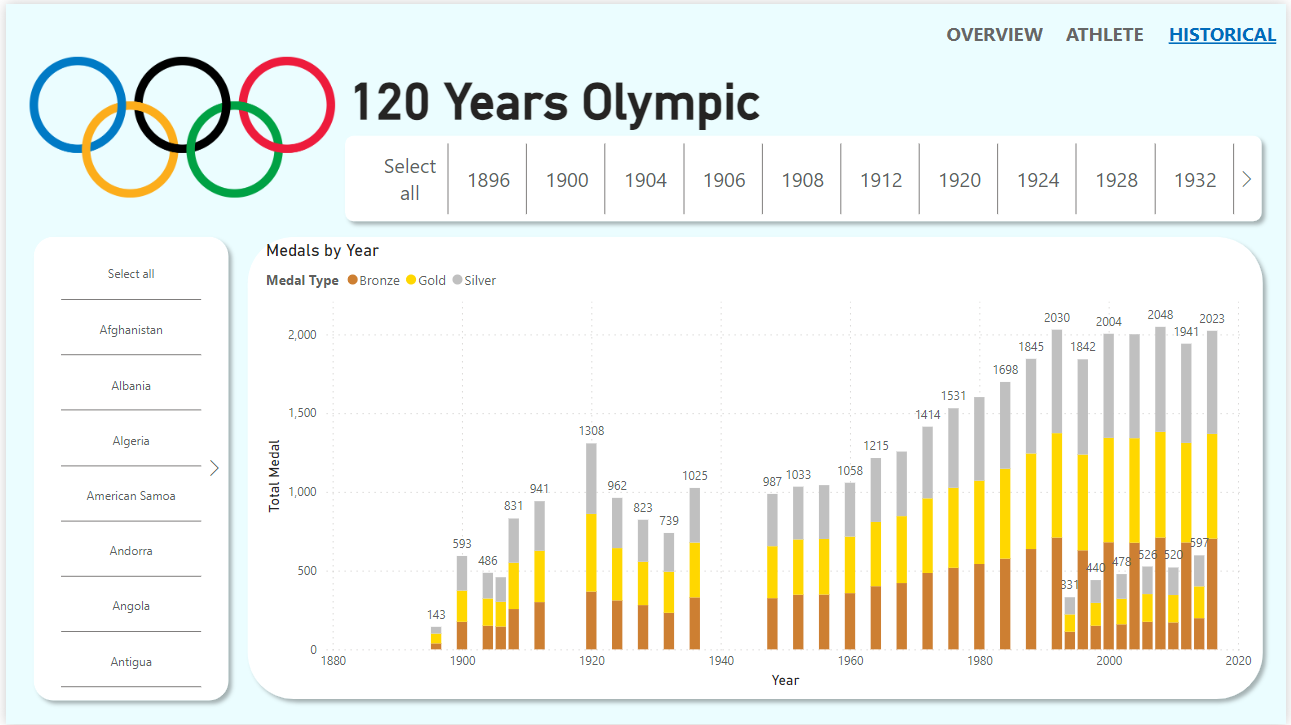


Age and Gender Distribution: The majority of Olympic athletes fall within the 19-25 age group, with over 118.58K participants. Interestingly, both the 19-25 and 26-35 age brackets show a strong representation of male athletes (84K in each category). Female athletes, while fewer in number, still contribute significantly, particularly in the 19-25 age group (35K). The 36-45 age group and those above 46 show a sharp decline in athlete participation, indicating that Olympic competition is primarily a young person's arena.

Total Athletes by Gender: The Olympics have traditionally seen more male participation, with 195.35K male athletes (72.43%) compared to 74.38K female athletes (27.57%). This difference highlights a historical gender imbalance, though recent years have seen strides toward greater gender equality in sports.

Sport by Athlete and Gender: Among the various sports, athletics leads with the highest participation, having 1959 male and 883 female athletes. Most of the sports have high levels of participation, with noticeable male dominance in these sports.

Medals by Athlete: The dashboard highlights some of the most decorated Olympians in history. Michael Fred Phelps. stands out with an incredible tally of 28 medals, including 23 golds, making him the most successful Olympian ever. Larysa Semenivna and Nikolay Yefimovich also make the top ranks with 18 and 15 medals, respectively. The visualization emphasizes not only the number of medals but also the distribution of gold, silver, and bronze.



Early Growth (1896-1940): The Olympics saw a gradual increase in the total number of medals, with noticeable spikes around 1920 and 1936. The interruptions during the World Wars are also evident, with lower medal counts during these periods.

Post-War Expansion (1948-1980): The post-war period brought a surge in Olympic participation, with a dramatic increase in the number of medals awarded, especially during the 1960s and 1970s.

Modern Olympics (1984-2023): The modern era of the Olympics shows a consistent rise in the total medals awarded, peaking in recent Games as the event became a truly global spectacle with widespread participation.

The distribution of gold, silver, and bronze medals remains consistent throughout, symbolizing the enduring competitive spirit and excellence that the Olympics represent.