Advanced Data Management

Analysing historical data pertaining to the Olympic Games

February 2019

4714800: Humberto Enrique Della Torre Escobar

4709433: Sanket Rajeev Sabharwal

The case at hand

A quick overview:

- Analyzing data related to the Olympic games
- **120 years of data:** Athens 1896 to Rio 2016
  - 2 dataset (athelete_events.csv & noc_regions.csv)
  - Consisting of player attributes such as:
    - Age
    - Height
    - Weight
    - Medal Won
    - Country
    - Region information; etc.

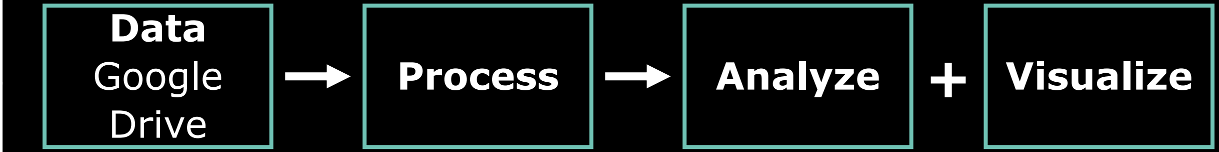
Objective:

- Understand global history of the Olympic games
- Observe major patterns in Olympic history
- Provide analytical observations and insights
- Research, and summarize, the findings

Our approach

System Particulars & Specifics	
Particulars	**Specifics**
Domain	Olympic Sports
Partitioning	Block-based
Replication	Single-leader
CAP	CA
Storage System	HDFS
Framework	Spark
APIs	PySpark and FindSpark
Programming Language	Python

Procedure:



Technique and Methodology:

- Pull data from a google drive link
- Process the data using Spark framework
- Individual funcs. called in the main func. that provide analytical results on choosing from an interactive list
- Comments to help understand the logic, and debugging if required
- Research on observations to better understand the underlying reasons for the cause of specific results

Estimate of effort:

• Project Plan:	5 Hours
• Conceptualization:	5 Hours
• Programming:	35 Hours
• Documentation:	10 Hours

A glimpse of our dataset

athlete_events
ID
Name
Sex
Age
Height
Weight
Team
NOC
Games
Year
Season
City
Sport
Event
Medal

Rows: 271116
Columns: 15

noc_regions
NOC
Region
Notes

Rows: 230
Columns: 3

Dataset 1	
Particulars	**Type**
ID	Number
Name	Text
Sex	Text
Age	Number
Height	Number
Weight	Number
Team	Text
NOC	Text
Games	Text
Year	Number
Season	Number
City	Number
Sport	Number
Event	Number
Medal	Number

Dataset 2	
Particulars	**Type**
NOC	Text
Region	Text
Notes	Text

Dataset 1

+

Dataset 2

↓ Datasets are merged using NOC as key to give a new column titled '**Country**' = **Region**

Dataset.csv:

- Rows: 27116
- Columns: 16

Processing the datasets

Dataset.csv		
Particulars	**Type**	**%Null**
ID	Number	0
Name	Text	0
Sex	Text	0
Age	Number	**3.5**
Height	Number	**22.2**
Weight	Number	**23.2**
Team	Text	0
NOC	Text	0
Country	Text	**0.1**
Games	Text	0
Year	Number	0
Season	Number	0
City	Number	0
Sport	Number	0
Event	Number	0
Medal	Number	**85.3**

Tackling Null values:

- Replacing ‘NA’ in appropriate format
- **Country**:
  - Singapore missing
    - Resolved with filling null values
- **Age, Height, Weight**:
  - Resolved with filling mean values
- **Medal:**
  - Resolved with filling null values with Participated

Workload – Empowering Insights

Observe major patterns and trends pertaining to:

History **Geopolitics** **Society**

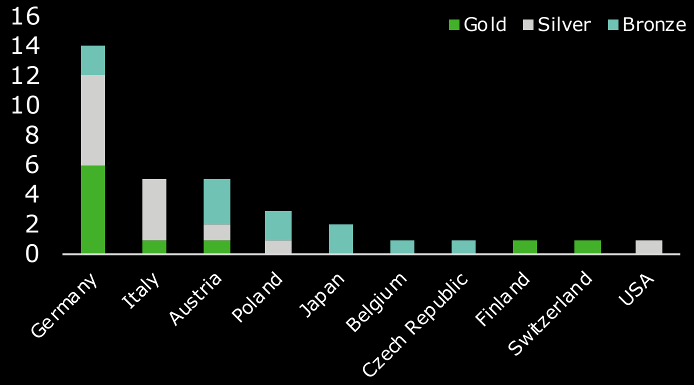
- **23 functions providing deep insights**
- **Examples of workload:**

1. Ratio of Males – Females by Country, Overall
2. Highest medals won so far by Country
3. Top 5/10 players:
   A. Won gold medals
   B. Youngest and Oldest players
4. "**Art Competitions**"
5. Number of sports/events over the years

Workload – Empowering Insights – Examples

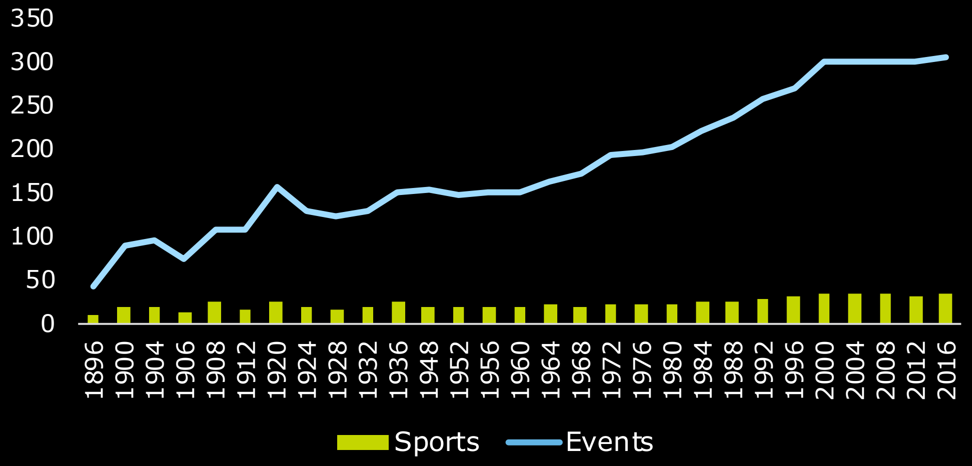
Art competitions were held in the following years: 1912, 1920, 1924, 1928, 1936, 1948

Top 10 countries by Medal Count in Art Competitions	
Germany	26
France	15
Italy	14
Austria	10
Switzerland	9
Denmark	9
UK	9
USA	9
Belgium	8
Poland	8



Top Gold Medal Winners	
Michael Fred Phelps, II	23
Raymond Clarence "Ray" Ewry	10
Paavo Johannes Nurmi	9
Mark Andrew Spitz	9
Larysa Semenivna Latynina (Diriy-)	9
Frederick Carlton "Carl" Lewis	9

Number of Sports and events over the years

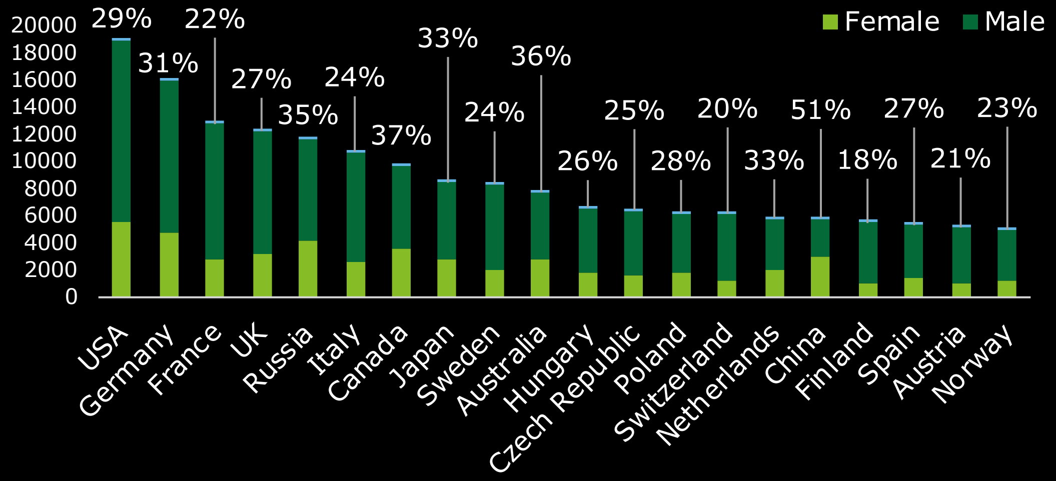


Workload – Empowering Insights – Examples (Contd.)

Top 10 Youngest Athletes

Name	Age	Year	Sport
Dimitrios Loundras	10	1896	Gymnastics
Carlos Barrera	11	1992	Rowing
Liana Vicens	11	1968	Swimming
Luigina Giavotti	11	1928	Gymnastics
Beatrice Hutiu	11	1968	Figure Skating
Marcelle Matthews	11	1960	Figure Skating
Liu Luyang	11	1988	Figure Skating
Etsuko Inada	11	1936	Figure Skating
Magdalena Colledge	11	1932	Figure Skating
Sonja Henie	11	1924	Figure Skating

Female to Male Ratio by Country



Top 10 Oldest Athletes

Name	Age	Year	Sport
Arthur Szent-Mikls	72	1936	Equestrianism
Oscar Swahn	72	1920	Shooting
Hiroshi Hoketsu	71	2012	Equestrianism
Thomas Scott	71	1904	Archery
Charles Martin	71	1900	Sailing
Krasimir Krastev	70	1980	Sailing
Owen D. Phillips	70	1976	Shooting
Durward Knowles	70	1988	Sailing
Hilda Johnstone	69	1972	Equestrianism
Louis Graville	69	1900	Equestrianism

Highest Medals won by Country

USA	5637
Russia	3947
Germany	3756
UK	2068
France	1777

Lowest Medals won by Country

Eritrea	1
Senegal	1
Guyana	1
Djibouti	1
Tonga	1

Key findings

- Number of participating nations, athletes – increased dramatically over the years.
- Although, the inclusion of most countries has resulted in a level off now
- Geographic representation has grown since 1986 – some countries in developing / under-developed parts of the world are yet under-represented
- Female representation has increased significantly
- With the pro-Nazi propaganda, during the 1936 games, Germany dominated
- During the years 1916, 1940, and 1944 – no Olympics were held due to the ongoing World Wars 1 & 2
- Art competitions were a part of Olympics from 1912 to 1948