Project 1: Olmypic History Trend Analytics

Group 9 - Thomas George Thomas, Yang Liu, Pratyush Pothuneedi

10/28/2021

1. Introduction

Data Description

We are considering 120 years of Olympic history where we find some interesting trends after analysis. There are two files in our data set:

1.The file *athlete_events.csv* contains 271116 rows and 15 columns. Each row corresponds to an individual athlete competing in an individual Olympic event (athlete-events). The columns are:

- ID : Unique number for each athlete
- Name : Athlete's name
- Sex : M or 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

2.The file *noc_regions.csv* contains 230 rows and 3 columns. Each row shows the special NOC code that denotes a region/country along with notes. The columns are:

- NOC : National Olympic Committee 3 letter code
- Country name : matches with regions in map_data("world")
- Notes : Special notes if any

We take a lot at 9 business questions that we want answered.

Data Acquisition

We acquire the data set from Kaggle: https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results?select=athlete_events.csv

Looking at the structure of Data set 1 - Athletes Data. Previewing only a few columns for PDF formatting

Name	Sex	Age	Height	Weight	Medal	Sport
A Dijiang	M	24	180	80	NA	Basketball
A Lamusi	M	23	170	60	NA	Judo
Gunnar Nielsen Aaby	M	24	NA	NA	NA	Football
Edgar Lindenau Aabye	M	34	NA	NA	Gold	Tug-Of-War
Christine Jacoba Aaftink	F	21	185	82	NA	Speed Skating

Looking at the structure of Data set 2 - Region Data

NOC	region	notes
AFG	Afghanistan	
AHO	Curacao	Netherlands Antilles
ALB	Albania	
ALG	Algeria	
AND	Andorra	

2. Data Wrangling

2.1 Data Discovery

A. Summary Statistics

Computing summary of the athletes data (Data set 1)

```
##        ID              Name               Sex                 Age
##  Min.   :     1   Length:271116      Length:271116      Min.   :10.00
##  1st Qu.: 34643   Class :character   Class :character   1st Qu.:21.00
##  Median : 68205   Mode  :character   Mode  :character   Median :24.00
##  Mean   : 68249                                        Mean   :25.56
##  3rd Qu.:102097                                        3rd Qu.:28.00
##  Max.   :135571                                        Max.   :97.00
##                                                        NA's   :9474
##      Height          Weight          Team               NOC
##  Min.   :127.0   Min.   : 25.0   Length:271116      Length:271116
##  1st Qu.:168.0   1st Qu.: 60.0   Class :character   Class :character
##  Median :175.0   Median : 70.0   Mode  :character   Mode  :character
##  Mean   :175.3   Mean   : 70.7
##  3rd Qu.:183.0   3rd Qu.: 79.0
##  Max.   :226.0   Max.   :214.0
##  NA's   :60171   NA's   :62875
##     Games                Year         Season              City
##  Length:271116      Min.   :1896   Length:271116      Length:271116
##  Class :character   1st Qu.:1960   Class :character   Class :character
##  Mode  :character   Median :1988   Mode  :character   Mode  :character
##                     Mean   :1978
##                     3rd Qu.:2002
##                     Max.   :2016
##
```

```
##     Sport              Event              Medal
##  Length:271116      Length:271116      Length:271116
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
```

We can see that there are NA's in the numerical fields of Age, Height, Weight which we will handle

Computing summary of region data (Data set 2)

```
##      NOC              region             notes
##  Length:230         Length:230         Length:230
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
```

B. Discovering Discrete Data

Table 3: Total Records for Athletes Dataframe

total_records
271116

Table 4: Total Records in Regions Dataframe

total_records
230

Looking for NA's in all the columns

```
## [1] "Age"    "Height" "Weight" "Medal"
```

We have NA's for numerical data: Age, Height & Weight and for categorical data: Medal.

Table 5: Records by Medal Count

Medal	total_records
Bronze	13295
Gold	13372
Silver	13116
NA	231333

There are 231333 NA's for Medals which is categorical data and we need to handle this in the cleaning part

Table 6: Records grouped by categories

NOC	region	notes	Total_records
ROT	NA	Refugee Olympic Team	1
TUV	NA	Tuvalu	1
UNK	NA	Unknown	1

There are 3 records with NA in region_df

2.2 Structuring

We can see that we don't need to do additional restructuring as columns like "Games" is already split and available as Year and Season

2.3 Cleaning

A. Handling Missing Data

We don't filter out the NA values since the columns that exhibit them are required for our analysis and we don't want to drop crucial data which would lead to data skewness. We will be filling the NA values for numerical columns like Age, Height, Weight with the **median** values since we require whole numbers and to reduce the degree of skewness while maintaining data integrity. The Medals are filled with **'None'** which would signify that the athletes simply didn't win any of the categories of Medals (Gold, Silver, Bronze).

B. Replacing NA's in Medals Calculating Missing Median for the missing values for Age, Height and Weight

```
##    Age Height Weight
##     24    175     70
```

C. Handling Missing data in Regions We replace the Na's values with region/notes for respective NOC's for the region data.

Table 7: Number of NA's in Region after fix

region	total_records

2.4 Enriching

A. Adding Attribute region

We will join regions_df and atheletes_df based on the NOC code to get the Region for enriching the data.

We don't have any other attribute to split or to create a new category since we believe that we have all the required columns for our analysis.

2.5 Validating

A. Check for any missing values

```
## [1] "Region"
```

Table 8: Null Records check by Medal Count

NOC	Region	total_records
SGP	NA	349

For NOC SGP, there are no records in our regions_df but is present in atheltes_df, as a result we are getting NA values after the join. We will add Singapore Region to the NOC in the joined data. Applying fix:

Table 9: Checking for NA records in Region after change

Region	total_records

B. Check for Duplicates

```
## [1] 1385
```

There **1385** duplicate records on the whole data set. Taking unique values to remove duplicates.

C. Checking boundary cases

Table 10: Age boundary cases

max_age	min_age	Average_Age
97	10	25.40454

Table 11: Height boundary cases

max_height	min_height	Average_height
226	127	175.265

Table 12: Weight boundary cases

max_weight	min_weigt	Average_weight
214	25	70.5417

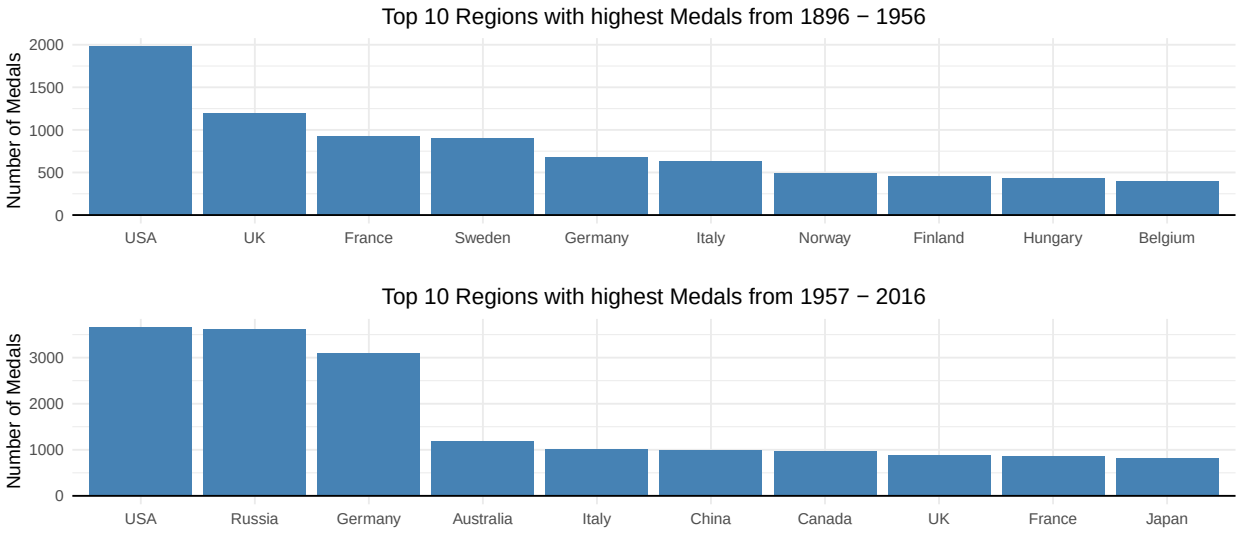
All our boundary cases looks reasonable and accurate after our wrangling.

2.6 Publishing

The data is cleaned & wrangled and made available for the team to develop business cases.

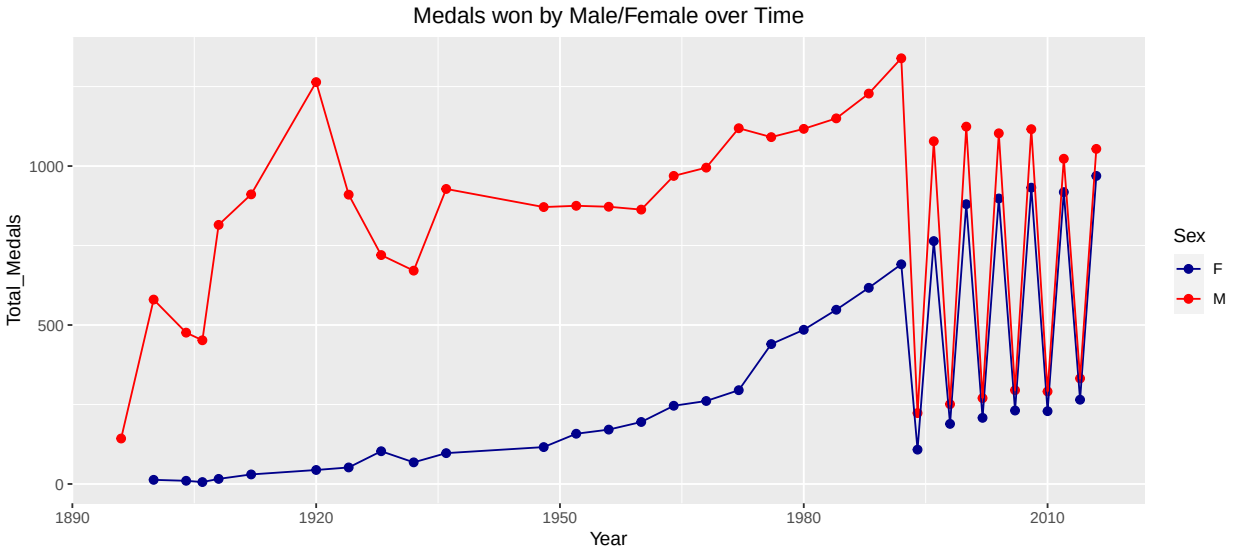
3. Analytical Questions

1. Trend analysis of Top 10 regions with the highest number of medals between 1896 - 1956 & 1957 - 2016



Observation: USA remains the region with the highest number of Medals in the combined history of 120 years in Olympics. Russia, a new inclusion in the top 10 took 2nd position in the later half.Germany moved into the 3rd position in the second half while UK and France slipped from 2nd and 3rd to 8th and 9th position. There are new countries in the later half such as Australia, Canada and Japan which where not in the top 10 for the earlier history of Olympics.

2. Medals won by Males/Females over Time



Observation: From the graph, we can see that there is a gradual increase in the number of medals won by female athletes over time. Male athletes tend to outnumber female athletes but their numbers also keep fluctuating over time. After the years 1994, the summer and winter Olympic games were split and held during seperate years, hence why the graph shows different points.

3. Finding the most participated sport in Olympics every year

Table 13: Most participated Sport in Olympic Games every year

Year	Sport	Participation
1896	Athletics	106
1900	Fencing	317
1904	Gymnastics	458
1906	Athletics	470
1908	Athletics	778
1912	Athletics	962
1920	Athletics	849
1924	Athletics	1003
1928	Athletics	992
1932	Art Competitions	620
1936	Athletics	1007
1948	Gymnastics	1060
1952	Gymnastics	2391
1956	Athletics	1013
1960	Gymnastics	1746
1964	Gymnastics	1484
1968	Gymnastics	1496
1972	Athletics	1686
1976	Athletics	1297
1980	Athletics	1268
1984	Athletics	1674
1988	Athletics	2062

Year	Sport	Participation
1992	Athletics	2054
1994	Cross Country Skiing	639
1996	Athletics	2386
1998	Cross Country Skiing	733
2000	Athletics	2468
2002	Cross Country Skiing	774
2004	Athletics	2175
2006	Cross Country Skiing	812
2008	Athletics	2244
2010	Cross Country Skiing	725
2012	Athletics	2278
2014	Cross Country Skiing	765
2016	Athletics	2508

Observation: This table shows that Athletics has remained the most contested sport in 120 years of Olympics. Art Competitions were the highest participated Olympic Sport in 1932 before it was removed from the Olympics. As the Olympic Winter and Summer games were seperated into different years from 1994, Cros Country Skiing emerged as the most participated game held during the Winters.

4. In which Olympic year did a particular country win a medal for the first time for a particular sport

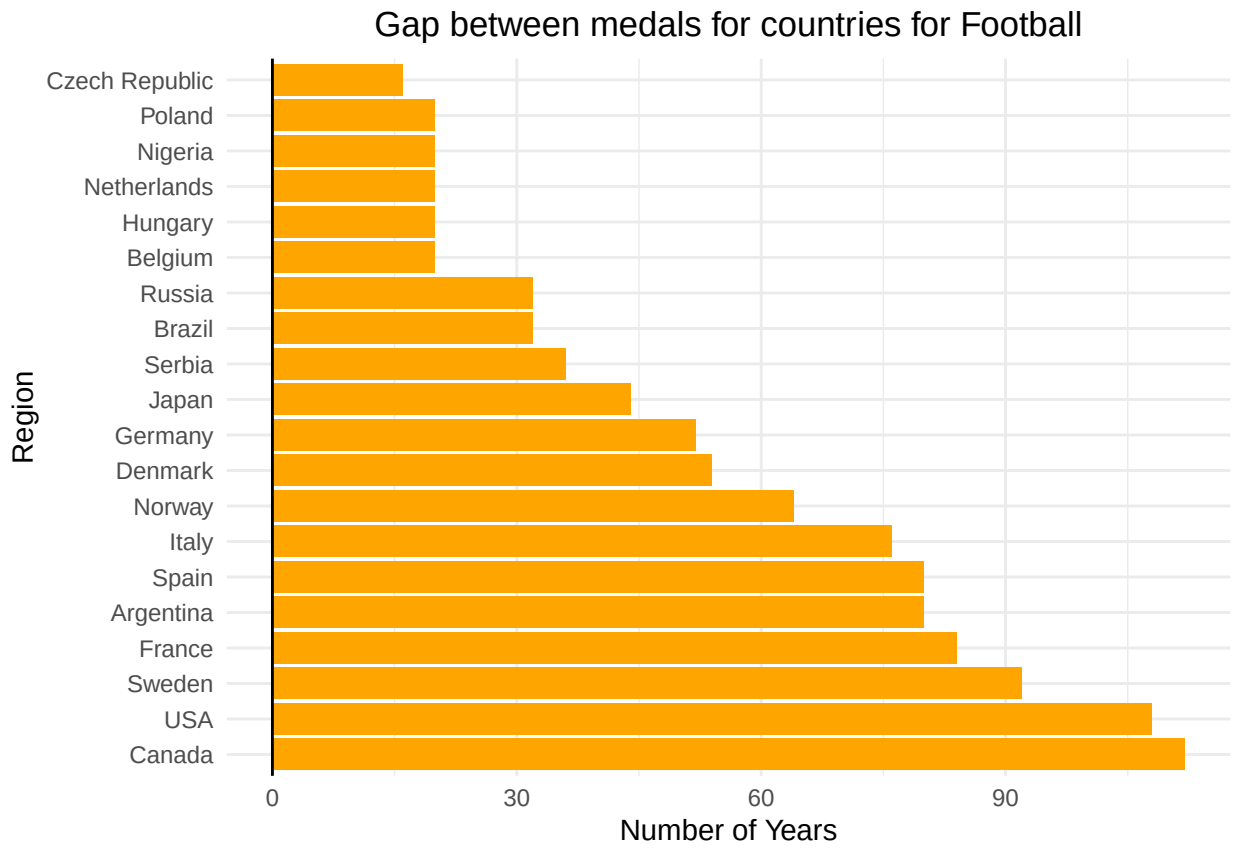
Table 14: First year in which countries won medal in Football

Region	Year
Argentina	1928
Austria	1936
Belgium	1900
Brazil	1984
Bulgaria	1956
Cameroon	2000
Canada	1904
Chile	2000
China	1996
Czech Republic	1964
Denmark	1906
France	1900
Germany	1964
Ghana	1992
Greece	1906
Hungary	1952
Italy	1928
Japan	1968
Mexico	2012
Netherlands	1900
Nigeria	1996
Norway	1936
Paraguay	2004
Poland	1972
Russia	1956

Region	Year
Serbia	1948
South Korea	2012
Spain	1920
Sweden	1924
Switzerland	1924
UK	1900
Uruguay	1924
USA	1904

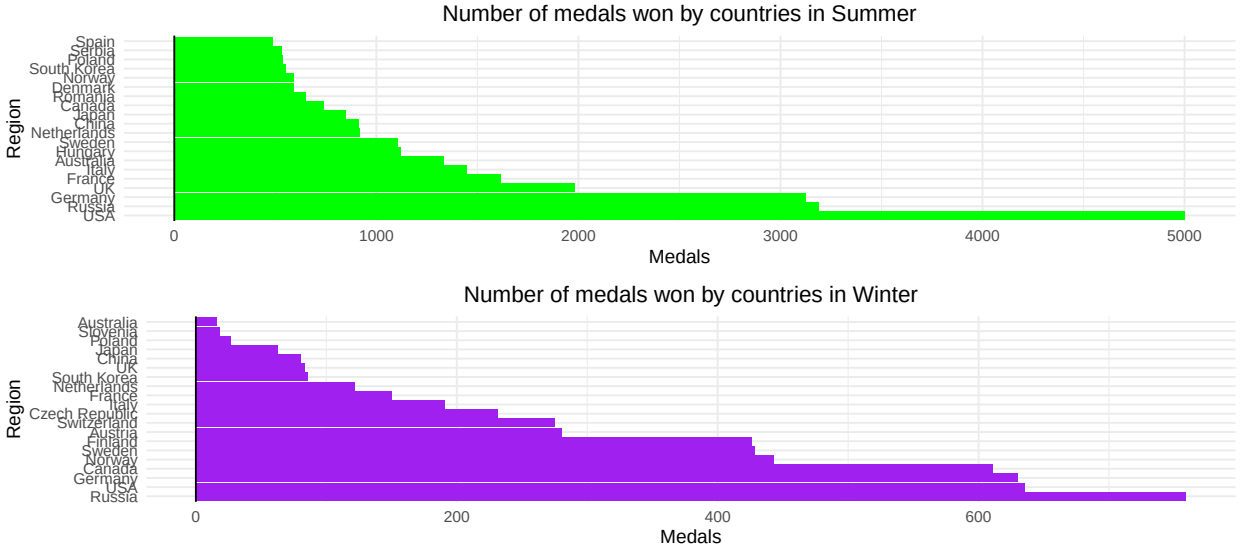
Observation: From the above table we can see the first year in which each country won a medal for football. The first countries to win medals for football are UK,Belgium,France and Netherlands and all these countries are from Europe.

5. Trend analysis per sport per country for the gap between medals for Football



Observation: The gap between first and last medals for football for each country is displayed. Canada has the largest gap of 112 years while Czech Republic has the smallest gap of 16 years.

6. Comparison of medals won by regions in Summer & Winter.



Observation: In Summer USA takes the top spot in the number of medals won and Russia stood second. When it comes to winter the positions are interchanged. Germany remains constant in both summer and winter. The graph has a uniform increase in the number of medals in summer but its not a uniform increase in winter, So this means that the winter games are more competitive while summer has a distinctive winner.

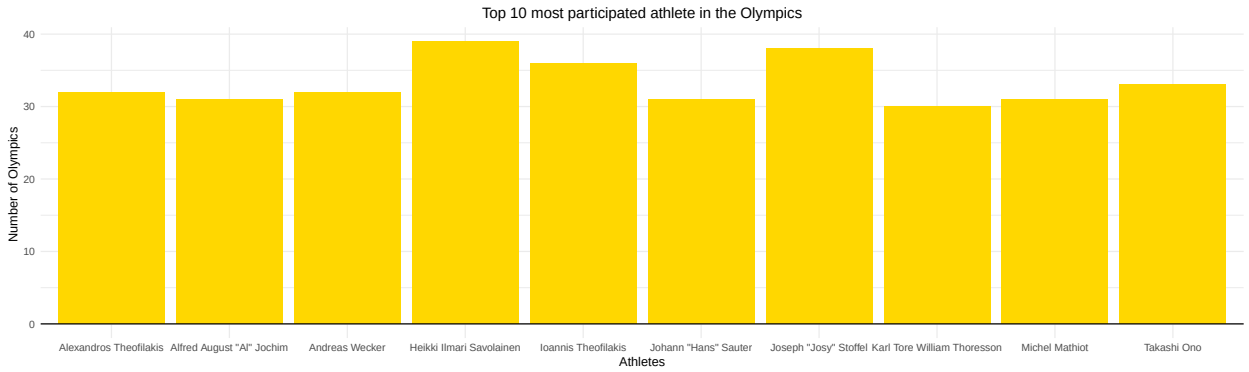
7. Top 10 Host cities with the highest participation.

Table 15: Host cities with highest participation

Year	City	number
2000	Sydney	13821
1996	Atlanta	13780
2016	Rio de Janeiro	13688
2008	Beijing	13602
2004	Athina	13443
1992	Barcelona	12977
2012	London	12920
1988	Seoul	12037
1972	Munich	10304
1984	Los Angeles	9454

Observation: The top ten cities with the highest participation are Sydney(2000), Atlanta(1996), Rio de Janeiro(2016),Beijing(2008), Athina(2004), Barcelona(1992), London(2012),Seoul(1988),Munich(1972),Los Angeles(1984)

8. Top 10 athletes with the highest participation in the Olympics



Observation: As we can see nearly all the top 10 athletes participated in the Olympics more than 30 times and the one with the most is Heikki Ilmari Savolainen with 39 times

9. Athletes with the most number of medals in each sport

Table 16: Athlete with the most number of medals per sport

Name	Sport	number
Michael Fred Phelps, II	Swimming	28
Larysa Semenivna Latynina (Diriy-)	Gymnastics	18
Edoardo Mangiarotti	Fencing	13
Ole Einar Bjrndalen	Biathlon	13
Birgit Fischer-Schmidt	Canoeing	12
Paavo Johannes Nurmi	Athletics	12
Carl Townsend Osburn	Shooting	11
Gerard Theodor Hubert Van Innis	Archery	10
Isabelle Regina Werth	Equestrianism	10
Marit Bjrgen	Cross Country Skiing	10
Yang Yang	Short Track Speed Skating	10

Observation: For Swimming, Michael Fred Phelps, II won the most number of medals(28) and there are 11 players with the number of medals equal or more than 10 in different sports.

Summary

After careful analysis of Olympic history worth 120 years. We were able to decipher lot of emerging patterns and visualize them. We were able to gain valuable insights about our business questions through various plots.

All students of Group 9 - Thomas George Thomas, Yang Liu, Pratyush Pothuneedi contributed equally to the project.