COSC3000 - REPORT
Visualisation

Teanlouise

April 21, 2020

Contents

1 Introduction

The topic of this project is the Modern Olympics. The games have been a global competition since 1896 with both Summer and Winter sports. The goal is to analyse the patterns of a medal winner depending on their physical characteristics (weight, height, age, sex), home country (athleticism, GDP, population) and the games in which they compete (location). The Olympics are supposed to be a celebration of peace, inclusion and human persistence. It is an opportunity for people to be proud of their country, and be in awe of the feats of athletes. By exploring the above topics it may be possible to determine whether there is a fair representation at the Olympics, and whether the winners are too predictable. If this is the case than the Olympics are no longer serving their purpose.

2 About the data

To explore and understand how the Olympics has changed over time, a variety of data was collected from numerous sources. There are three main sources broken up over five datasets.

2.1 Athlete Information

The first set of data that needs to be collected relates to the Athlete's information. This includes their physical characteristics (height, weight, age), their role in the Olympics (sport, medal, country) and when they competed (season, year). This information is available for public download on Kaggle under the title '120 years of Olympic history (1896 - 2018)'. This dataset was created by scraping from www.sport-reference.com. The data is broken down into two files:

1. Athlete and Events - This file contains all of the information recorded about the athlete from all Modern Olympics. The variables of interest are ID, Sex, Age, Height, Weight, NOC, Year, Season and Medal.

ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event	Medal
1	A Dijiang	M	24	180	80	China	CHN	1992 Summer	1992	Summer	Barcelona	Basketball	Basketball Men's Basketball	NA
2	A Lamusi	M	23	170	60	China	CHN	2012 Summer	2012	Summer	London	Judo	Judo Men's Extra-Lightweigh t	NA
3	Gunnar Nielsen Aaby	M	24	NA	NA	Denmark	DEN	1920 Summer	1920	Summer	Antwerpen	Football	Football Men's Football	NA
4	Edgar Lindenau Aabye	M	34	NA	NA	Denmark/Sw eden	DEN	1900 Summer	1900	Summer	Paris	Tug-Of-War	Tug-Of-War Men's Tug-Of-War	Gold
5	Christine Jacoba Aaftink	F	21	185	82	Netherland s	NED	1988 Winter	1988	Winter	Calgary	Speed Skating	Speed Skating Women's 500 metres	NA
5	Christine Jacoba Aaftink	F	21	185	82	Netherland s	NED	1988 Winter	1988	Winter	Calgary	Speed Skating	Speed Skating Women's 1,000 metres	NA
5	Christine Jacoba Aaftink	F	25	185	82	Netherland s	NED	1992 Winter	1992	Winter	Albertvill e	Speed Skating	Speed Skating Women's 500 metres	NA
5	Christine	F	25	185	82	Netherland	NED	1992	1992	Winter	Albertvill	Speed	Speed	NA

Figure 1: athlete_events.csv

2. NOC regions - A list of the countries and their NOC code. It is important to note some

countries changed their code in the data. This is noted in Appendix A.

1	AFG	Afghanistan	
2	AHO	Curacao	Netherlands Antilles
3	ALB	Albania	
4	ALG	Algeria	
5	AND	Andorra	
6	ANG	Angola	
7	ANT	Antigua	Antigua and Barbuda
8	ANZ	Australia	Australasia
9	ARG	Argentina	
10	ARM	Armenia	
11	ARU	Aruba	

Figure 2: noc_regions.csv

2.2 Country Information

The second set of data relates to the information about each country, including their GDP and population. The most trustworthy source for this data publicly available from World Bank national accounts data, and OECD National Accounts data files. The data is available from 1960 to present, and is accessed as separate files.

1. GDP - The GDP for all countries, represented in current US$.

	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
China	5,101,702,432,8	6,087,164,527,4	7,551,500,425,5	8,532,230,724,1	9,570,405,758,7	10,438,529,153	11,015,542,352	11,137,945,669	12,143,491,448	13,608,151,864
Netherlands	868,077,243,67	846,554,894,93	904,085,980,79	838,971,306,99	876,923,518,85	890,981,311,07	765,264,949,78	783,528,181,70	831,809,944,96	913,658,465,70
United States	14,448,933,025	14,992,052,727	15,542,581,104	16,197,007,349	16,784,849,190	17,521,746,534	18,219,297,584	18,707,188,235	19,485,393,853	20,544,343,456
Afghanistan	12,439,087,076	15,856,574,731	17,804,280,538	20,001,615,788	20,561,054,090	20,484,873,230	19,907,111,419	19,362,642,266	20,191,764,940	19,362,969,582
Albania	12,044,223,457	11,926,962,835	12,890,867,535	12,319,784,701	12,776,277,648	13,228,244,336	11,386,927,679	11,861,353,752	13,025,064,966	15,102,500,898
Algeria	137,214,821,17	161,205,065,46	200,015,355,52	209,062,886,91	209,754,763,86	213,808,808,74	165,978,425,16	160,032,930,35	167,390,266,10	173,757,952,82

Figure 3: gdp.csv

2. Population - The total population of all countries.

	2000	2010	2011	2012	2013	2014	2015	2016	2017	2018
Afghanistan	20,779,953.0	29,185,507.0	30,117,413.0	31,161,376.0	32,269,589.0	33,370,794.0	34,413,603.0	35,383,128.0	36,296,400.0	37,172,386.0
Albania	3,089,027.0	2,913,021.0	2,905,195.0	2,900,401.0	2,895,092.0	2,889,104.0	2,880,703.0	2,876,101.0	2,873,457.0	2,866,376.0
Algeria	31,042,235.0	35,977,455.0	36,661,444.0	37,383,887.0	38,140,132.0	38,923,687.0	39,728,025.0	40,551,404.0	41,389,198.0	42,228,429.0
American Samoa	57,821.0	56,079.0	55,759.0	55,667.0	55,713.0	55,791.0	55,812.0	55,741.0	55,620.0	55,465.0
Andorra	65,390.0	84,449.0	83,747.0	82,427.0	80,774.0	79,213.0	78,011.0	77,297.0	77,001.0	77,006.0
Angola	16,395,473.0	23,356,246.0	24,220,661.0	25,107,931.0	26,015,780.0	26,941,779.0	27,884,381.0	28,842,484.0	29,816,748.0	30,809,762.0

Figure 4: population.csv

2.3 Host Cities

The finally set of data is location of each of the games. The 'City' is included as a column in 'athlete_events.csv', however it is not paired with a country which is needed to compare an athlete's country with where they are competing. This data was not readily available as a data file but the information was found on https://architectureofthegames.net/olympic-host-cities/. The data was copied into two separate text files as is; summer and winter. Using python the

files were read, reformatted and combined to create a csv file. The file contains the year, city, country and season of each Olympic games. The code is in Appendix A.

Year	City	Country	Season
1896	Athens	Greece	Summer
1900	Paris	France	Summer
1904	St. Louis	United States	Summer
1908	London	England	Summer
1912	Stockholm	Sweden	Summer
1920	Antwerp	Belgium	Summer
1924	Paris	France	Summer
1924	Chamonix	France	Winter
1928	Amsterdam	Netherlands	Summer

Figure 5: host_city.csv

2.4 Final Dataset

From these files, a new dataset was created using pandas (and article research) to refine the data selection, remove redundancies, combine related variables and update incorrect data to ensure a cleaner dataset for the visualisations. These steps were taken to combine the following five datasets:

1. Remove 'Art Competitions' under 'Sport'
2. Remove 'Name', 'Team', 'Games', 'Event', 'Sport' from athlete_events.csv
   - 'Name' this is not relevant
   - 'Team' sometimes contradicts NOC/Country
   - 'Games' is duplicate, already split into 'Year' and 'Season'
   - 'Event' is not relevant and not consistently
   - 'Sport' is not relevant
3. Remove Years 1896-1920 from athlete_events.csv
   - Women didn't compete in 1896
   - Winter Games didn't commence until 1924
4. Make NOC codes consistent for countries that have changed.
   - Singapore (SIN): Stored as SGP in athlete_events
   - Russia (RUS): URS (1952-1988), EUN (1992), RUS (1994-2018)
   - Taiwan (TPE): ROC (1952-1976), TPE(1984-2018)
   - China (CHN): ROC (1924-1948), CHN (1980-2018)
   - Germany (GER): GER (1896-2018), EUA (1956-1964), FRG & GDR (1968-1988)
   - Czech Republic (CZE): CZE (1994-2018), TCH (1920-1992), BOH (1900-1912)
   - Serbia (SRB): SCG (2004-2006), SRB (1912, 2008-2018), YUG (1920-2002)
5. Add column COUNTRY by matching 'NOC' with the same from noc_regions.csv
6. Update CITY in athlete_events.csv to match more common names used in host_cities.csv
   - Athina to Athens
   - Roma to Rome
   - Antwerpen to Antwerp
   - Moskva to Moscow
   - Torino to Turin
   - Sankt Moritz to St Moritz

7. Add column HOST COUNTRY by matching 'country' with same from host_city.csv
8. Add column GDP by matching 'country' in gdp.csv
9. Add column POPULATION by matching 'country' in population.csv

ID	Sex	Age	Height	Weight	Year	Season	City	Host Country	Medal	NOC	Country	GDP	Population
1	M	24	180	80	1992	Summer	Barcelona	Spain	NA	CHN	China	426,915,712,711.10	1,164,970,000
2	M	23	170	60	2012	Summer	London	UK	NA	CHN	China	8,532,230,724,141.80	1,350,695,000
5	F	21	185	82	1988	Winter	Calgary	Canada	NA	NED	Netherlands	261,910,508,306.40	14,760,094
5	F	21	185	82	1988	Winter	Calgary	Canada	NA	NED	Netherlands	261,910,508,306.40	14,760,094
5	F	25	185	82	1992	Winter	Albertville	France	NA	NED	Netherlands	362,962,871,804.50	15,184,166
5	F	25	185	82	1992	Winter	Albertville	France	NA	NED	Netherlands	362,962,871,804.50	15,184,166
5	F	27	185	82	1994	Winter	Lillehammer	Norway	NA	NED	Netherlands	379,130,260,201.00	15,382,838
5	F	27	185	82	1994	Winter	Lillehammer	Norway	NA	NED	Netherlands	379,130,260,201.00	15,382,838
6	M	31	188	75	1992	Winter	Albertville	France	NA	USA	USA	6,520,327,000,000.00	256,514,000
6	M	31	188	75	1992	Winter	Albertville	France	NA	USA	USA	6,520,327,000,000.00	256,514,000
6	M	31	188	75	1992	Winter	Albertville	France	NA	USA	USA	6,520,327,000,000.00	256,514,000
6	M	31	188	75	1992	Winter	Albertville	France	NA	USA	USA	6,520,327,000,000.00	256,514,000
6	M	33	188	75	1994	Winter	Lillehammer	Norway	NA	USA	USA	7,287,236,000,000.00	263,126,000
6	M	33	188	75	1994	Winter	Lillehammer	Norway	NA	USA	USA	7,287,236,000,000.00	263,126,000
6	M	33	188	75	1994	Winter	Lillehammer	Norway	NA	USA	USA	7,287,236,000,000.00	263,126,000
6	M	33	188	75	1994	Winter	Lillehammer	Norway	NA	USA	USA	7,287,236,000,000.00	263,126,000
7	M	31	183	72	1992	Winter	Albertville	France	NA	USA	USA	6,520,327,000,000.00	256,514,000
7	M	31	183	72	1992	Winter	Albertville	France	NA	USA	USA	6,520,327,000,000.00	256,514,000
7	M	31	183	72	1992	Winter	Albertville	France	NA	USA	USA	6,520,327,000,000.00	256,514,000

Figure 6: all_data.csv

3 Discussion

The prediction of medal winners will be explored through a number of topics. Each of these will be presented as a comparison between the winter and summer games. Data will be considered from 1924 (when the Winter Olympics were introduced), except for BMI, GDP and population which will be taken from 1960 (when data is collected for majority of points for these variables). The variables 'NOC', 'year' and 'season' are used for all analysis. From current knowledge these are possible visualisations that can be explored:

- The distribution of the number of events and medals [Histogram (#events, #medal)]
- The change in age for medal and non-medal winners [Turkey box (year v. age)]
- The BMI of medal and non-medal winners [q-q plot (medal v. non-medal BMI)]
- The proportion of medal winners to number athletes [scatter (#athletes v. #medals)]
- The proportion of men and women competing [scatter (#females v. #males)]
- The difference in % of medals when competing at host [scatter (% hosting v. visiting)]
- The effect of population and GDP on number of medals [multi pop, GDP, #medals]

Appendices

A Important notes about the data

1. athlete_events.csv - Possible factors that may affect results of each Olympics
    - 1924: Winter games commence
    - 1928: Women now compete in more than 2 sports
    - 1932: Low attendance due to Great Depression
    - 1940 & 1944: Cancelled due to WW2
    - 1948: Art sports (architecture, literature, music, painting, sculpture) removed
    - 1952: USSR/Russia starts competing, Republic of China (ROC) discontinued
    - 1956: Boycotts by 8 nations, including China
    - 1960: Height and Weight measured consistently from now
    - 1976: Boycotts by 25 nations (mostly from Africa)
    - 1980: Boycotts by 66 nations, including US
    - 2000: Summer Olympics capped at 28 sports, 300 events, 10,000 athletes
2. noc_regions.csv - The following countries are recorded under multiple codes:
    - Australia: AUS, ANZ (New Zealand, 19081912)
    - Russia: URS (19521988), EUN (1992), RUS (19942018)
    - China: ROC (19241948), CHN (19522018), HKG (Hong Kong, 19522018)
    - Germany: GER (18962018), EUA (19561964), FRG & GDR (19681988)
    - Czech Republic: CZE (19942018), TCH (19201992), BOH (19001912)
    - Serbia: SCG (20042006), SRB (1912, 20082018), YUG (19202002)

Appendices

A About the data

A.1 Host Cities

```
import pandas as pd

def get_season_df(file_name, season):
    host_data = []
    with open(file_name) as file_var:
        for line in file_var.readlines():
                year = line[:5]
                location = line[6:-1].split(', ')
                city = location[0]
                country = location[1]
                host_data.append([year, city, country, season])

    season_df = pd.DataFrame(host_data, columns=['Year',
                                                 'City',
                                                 'Host_Country',
                                                 'Season'])
    return season_df

# Read in the data for each seasons
summer_df = get_season_df(
    'F:/TEAN/Portfolio/olympics/code/host_summer.txt', 'Summer')
winter_df = get_season_df(
    'F:/TEAN/Portfolio/olympics/code/host_winter.txt', 'Winter')
# Combine to create 1 DF
host_df = pd.merge(summer_df, winter_df, how='outer').reset_index(drop = True

#Check if all cities accounted for:
athlete_df = pd.read_csv(
                'F:/TEAN/Portfolio/olympics/data/athlete_events.csv')

for host_city in host_df.City.unique():
    for athlete_city in athlete_df.City.unique():
        if (host_city not in host_df.City.unique()) \
                and (athlete_city not in host_df.City.unique()):
            print("Host City: ", host_city)
            print("Athlete City: ", athlete_city)

# Write to CSV
host_df.to_csv('F:/TEAN/Portfolio/olympics/data/host_countries.csv')
```

A.2 Combined Data

```
import pandas as pd
import numpy as np


# COMPARE SUBSETS OF DATA WITH MAIN AS CHANGES
def check_item_not_in(df1, df2):
    item_list = []
    count = 0
    if df1.nunique() != df2.nunique():
        for item in df1.unique():
            if item not in df2.unique():
                item_list.append(item)
                count+=1
    return item_list,count

# PRINT CHECKS OF HOW DATA CHANGING
def checkpoint(action, all, bool=False, lost=None):
    print("{action}: \n Unique NOC: {num_noc} \
\n\n\n\n\n\n\n\n\n Unique Athletes: {num_athletes}."
          .format(action=action,
                  num_noc=all.NOC.nunique(),
                  num_athletes=all.ID.nunique()))

    if bool:
        print("\tLost NOC: {} \t Lost Athletes: {}"
                  .format(check_item_not_in(lost.NOC, all.NOC)[0],
                  check_item_not_in(lost.ID, all.ID)[1]))

# START
noc_df = pd.read_csv(
    'F:/TEAN/Portfolio/olympics/data/noc_regions.csv')
host_df = pd.read_csv(
    'F:/TEAN/Portfolio/olympics/data/host_countries.csv')
athlete_df = pd.read_csv(
    'F:/TEAN/Portfolio/olympics/data/athlete_events.csv')

# START
all_df = athlete_df
#checkpoint('START', all_df)

# 1. Remove art competitions
art_df = all_df[all_df['Sport'] == 'Art Competitions']
all_df = all_df.drop(art_df.index)
#checkpoint('REMOVE ART', all_df, True, art_df)

# 2. Remove irrelevant columns
extra_df = all_df
all_df = all_df.drop(["Name", "Team", "Games"], axis=1)
```

```
#checkpoint('REMOVE EXTRA', all_df, True, extra_df)

# 3. Remove years
years_df = all_df[all_df['Year'].isin(range(1896,1921))]
all_df = all_df.drop(years_df.index)
#checkpoint('REMOVE YEARS', all_df, True, years_df)

# 4. Make NOC codes consistent for countries that have changed.
noc_unique = all_df.NOC.unique()
all_df.loc[(all_df.NOC == 'TCH'),'NOC'] = 'CZE'
all_df.loc[(all_df.NOC == 'SGP'),'NOC'] = 'SIN'
all_df.loc[(all_df.NOC == 'EUN')
                | (all_df.NOC == 'URS'),'NOC'] = 'RUS'
all_df.loc[(all_df.NOC == 'FRG')
                | (all_df.NOC == 'GDR'),'NOC'] = 'GER'
all_df.loc[(all_df.NOC == 'SCG')
                | (all_df.NOC == 'YUG'),'NOC'] = 'SRB'
#checkpoint('UPDATE NOC', all_df)

# 5. Add Country Column to match NOC
all_df = all_df.merge(noc_df[['region', 'NOC']]
                        .rename(columns={'region':'Country'})) \
                        .reset_index(drop = True)
#checkpoint('ADD COUNTRY', all_df)

# 6. Update Host City names that don't match
all_df.loc[(all_df.City == 'Athina'),'City'] = 'Athens'
all_df.loc[(all_df.City == 'Roma'),'City'] = 'Rome'
all_df.loc[(all_df.City == 'Antwerpen'),'City'] = 'Antwerp'
all_df.loc[(all_df.City == 'Moskva'),'City'] = 'Moscow'
all_df.loc[(all_df.City == 'Torino'),'City'] = 'Turin'
all_df.loc[(all_df.City == 'Sankt_Moritz'),'City'] = 'St._Moritz'
#checkpoint('UPDATE HOST', all_df)

# 7. Add Host Country
all_df = all_df.merge(host_df[['Host_Country', 'City']]) \
                                .sort_values("Year") \
                                .reset_index(drop = True)
#checkpoint('ADD HOST COUNTRY', all_df)


# 8. Add GDP
# 9. Add Population

# WRITE TO FILE
all_df.to_csv('F:/TEAN/Portfolio/olympics/data/all_data.csv')
```