Visualisation Proposal

What—what is the topic and goal of your project?

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).

Why—why is your project important and/or interesting?

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.

How—how will you go achieve the goal of your project?

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]

Data—what is your data source, how you will obtain it, and, if possible at this stage, provide a sample

1. 120 years of Olympic history (1896 - 2018) - Available for public download on Kaggle. Created by scraping www.sport-reference.com.
    - noc_regions.csv - A list of the countries and their code (at the time).

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	

- athlete_events.csv Relevant information of all athletes. The variables of interest are ID, Sex, Age, Height, Weight, NOC, Year, Season, City, 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

2. The World Bank (1960 - 2018) - Available to the public from World Bank national accounts data, and OECD National Accounts data files.
   - gdp.csv - 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

   - population.csv - 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

3. List of Olympic Host Cities - Manually create host_city.csv of each games with city and country from https://architectureofthegames.net/olympic-host-cities/

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

From these datasets a combined csv file will be created with all of the relevant variables.

- Remove 'Name, Team, Games, Event' & Years 1896-1920 from athlete_events.csv
- Add column COUNTRY by matching 'NOC' with the same from noc_regions.csv

- Add column HOST COUNTRY by matching 'country' with same from host_city.csv
- Add column GDP by matching 'country' in gdp.csv
- 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

Notes—some important things to keep in mind during analysis

1. athlete_events.csv - Possible factors that may affect results of each Olympics
    - 1924: Winter games commence 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)