Analyzing 120 years (1896-2016) Olympic Dataset

Team Members :

1. Rajath John Bosco
2. Manshi Shah
3. Chaitya Mehta

Agenda:

- Data Cleaning
- Does team enjoy Home Advantage?
- Top 5 performing events in a Country?
- Analyzing number of athletes in Summer and Winter Olympics over the years
- Analyzing the number of events in Summer and Winter Olympics over the years
- Calculating top performing countries in summer vs winter
- BMI analysis over the years
- Distribution of Age vs Height from 1896 to 2016
- Does GDP play a part to decide if a country would win medal or not
- Calculating the medals won by country in summer and winter olympics

Datasets

Olympics Dataset - https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results

GDP Dataset - https://datahub.io/core/gdp

Population Dataset - https://vizhub.com/celtic660/datasets/world_pop

Data Cleaning:

Cleaning data of the medal column by replacing NA with DNW(Did not Win)

Categorizing events into Individual and Team events

- We awarded 1 medal per team instead of awarding medal to each player in the team.
- If a Country has won medal in soccer, it should count one single medal instead of 11 medals (since 11 players in a team)

Removing events where age of athlete is more than 80 years

- After verifying data from internet, there were players who were dead, but their age was still updating. So, we removed events whose name started from “Arts”

Does Team enjoy Home Advantage?

	Year	Host Country	Team	Season	Medal_Won_Prev_Year	Medal_Won_Host_Year	Medal_Won_Next_Year
0	1900	France	France	Summer	NaN	66.0	1.0
1	1904	United States	United States	Summer	45.0	193.0	46.0
2	1912	Sweden	Sweden	Summer	24.0	57.0	NaN
3	1924	France	France	Summer	NaN	39.0	24.0
4	1924	France	France	Winter	NaN	39.0	24.0
5	1928	Netherlands	Netherlands	Summer	10.0	22.0	8.0
6	1932	United States	United States	Summer	56.0	114.0	59.0
7	1932	United States	United States	Winter	56.0	114.0	59.0
8	1936	Germany	Germany	Winter	25.0	103.0	NaN
9	1936	Germany	Germany	Summer	25.0	103.0	NaN
10	1952	Norway	Norway	Winter	6.0	18.0	3.0
11	1952	Finland	Finland	Summer	22.0	30.0	15.0
12	1956	Sweden	Sweden	Summer	37.0	17.0	7.0
13	1956	Australia	Australia	Summer	11.0	33.0	0.0

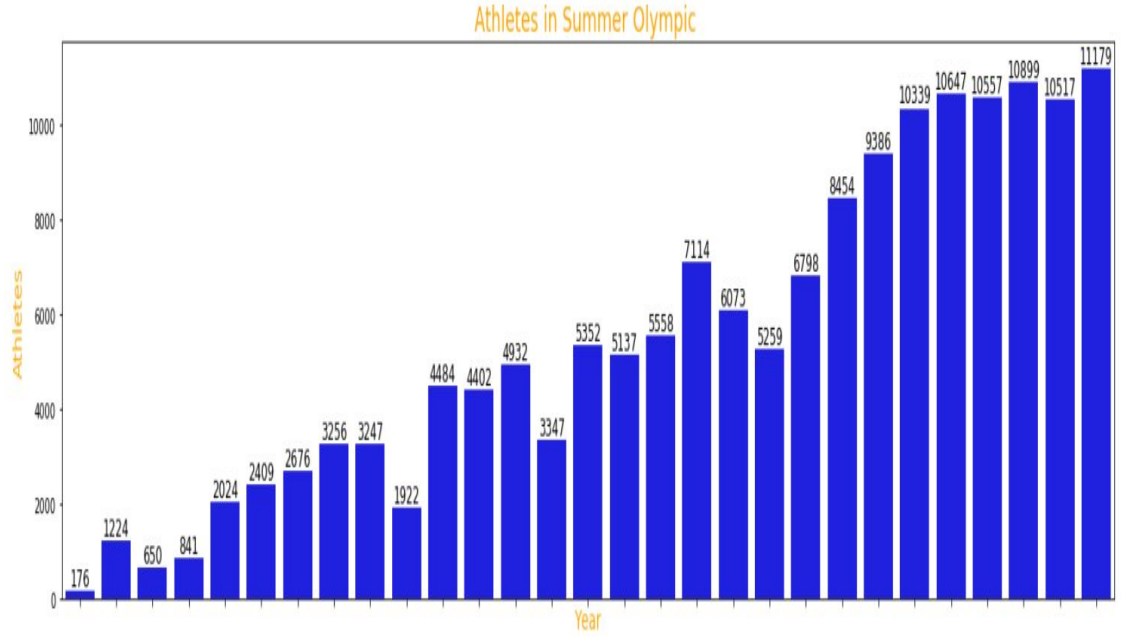
Top 5 performances by Germany, Russia, USA

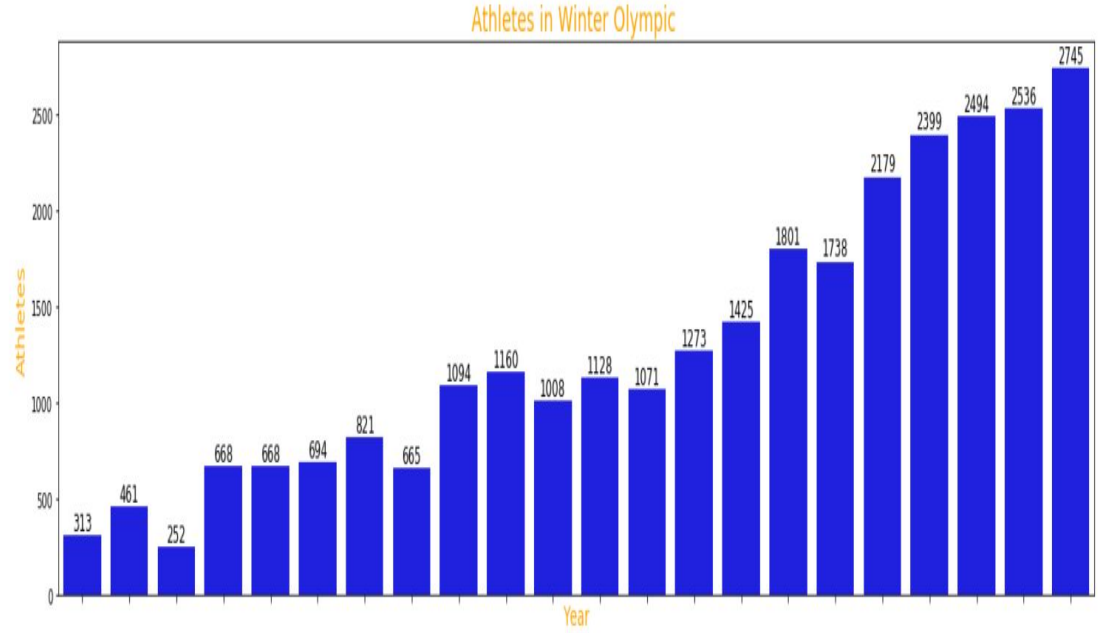
Team	Event	Gold_Medal_Count
Germany	Luge Men's Singles	6
Germany	Equestrianism Mixed Dressage, Team	5
Germany	Luge Women's Singles	5
Germany	Alpine Skiing Women's Combined	4
Germany	Canoeing Men's Canadian Doubles, 1,000 metres	4
Russia	Gymnastics Women's Uneven Bars	4
Russia	Rhythmic Gymnastics Women's Group	4
Russia	Rhythmic Gymnastics Women's Individual	4
Russia	Synchronized Swimming Women's Duet	4
Russia	Synchronized Swimming Women's Team	4
United States	Athletics Men's Long Jump	19
United States	Athletics Men's 110 metres Hurdles	17
United States	Athletics Men's 400 metres Hurdles	17
United States	Athletics Men's 400 metres	16

Changing Trends in women participation in Olympics

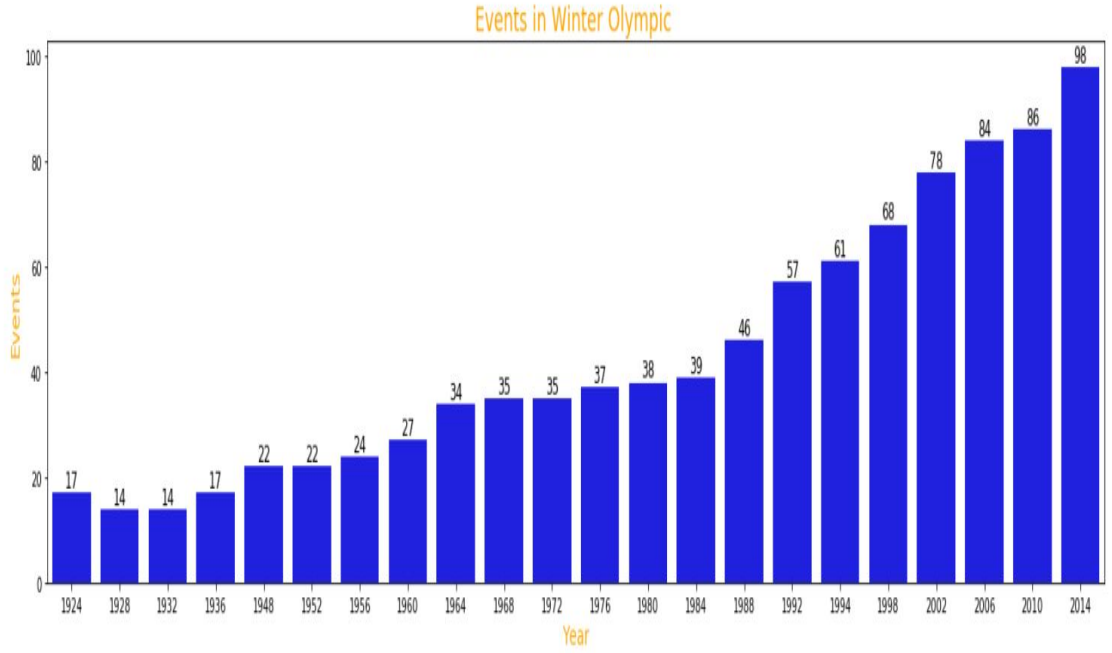
	Year	Women Athlete Count
0	1900	23
1	1904	6
2	1906	6
3	1908	44
4	1912	53
5	1920	78
6	1924	169
7	1928	340
8	1932	222
9	1936	441
10	1948	523
11	1952	629
12	1956	516
13	1960	757
14	1964	880
15	1968	994
16	1972	1266
16	1972	1266
17	1976	1490
18	1980	1358
19	1984	1845
20	1988	2511
21	1992	3205
22	1994	522
23	1996	3511
24	1998	788
25	2000	4065
26	2002	885
27	2004	4288
28	2006	955
29	2008	4598
30	2010	1032
31	2012	4645
32	2014	1102
33	2016	5031

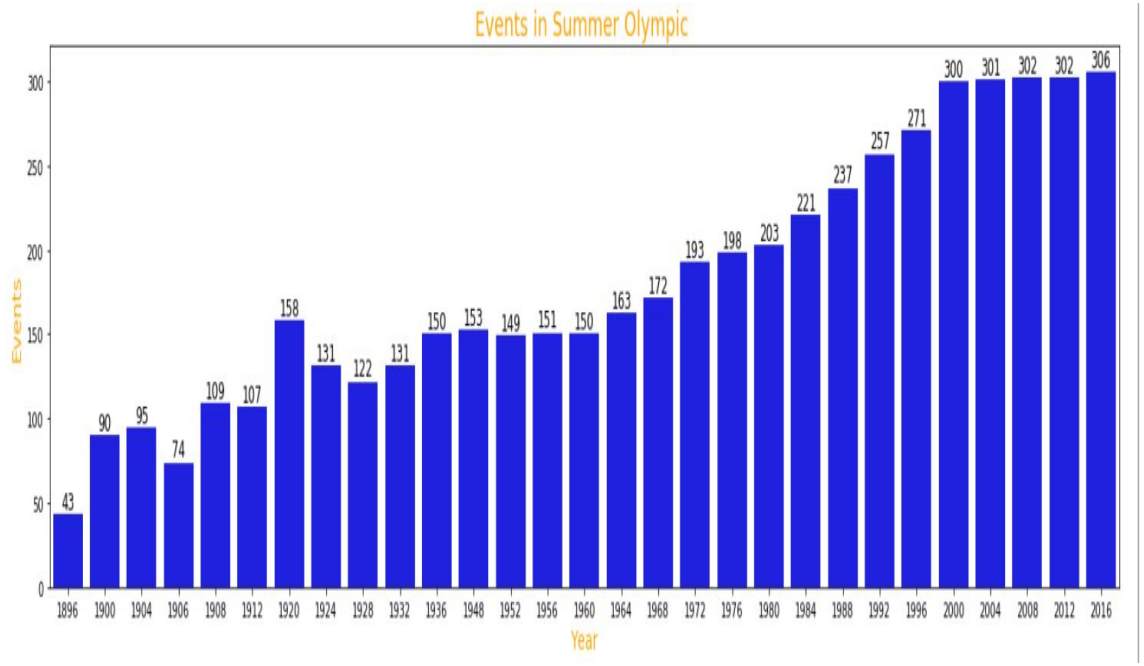
Analyzing number of athletes in summer vs winter over the years



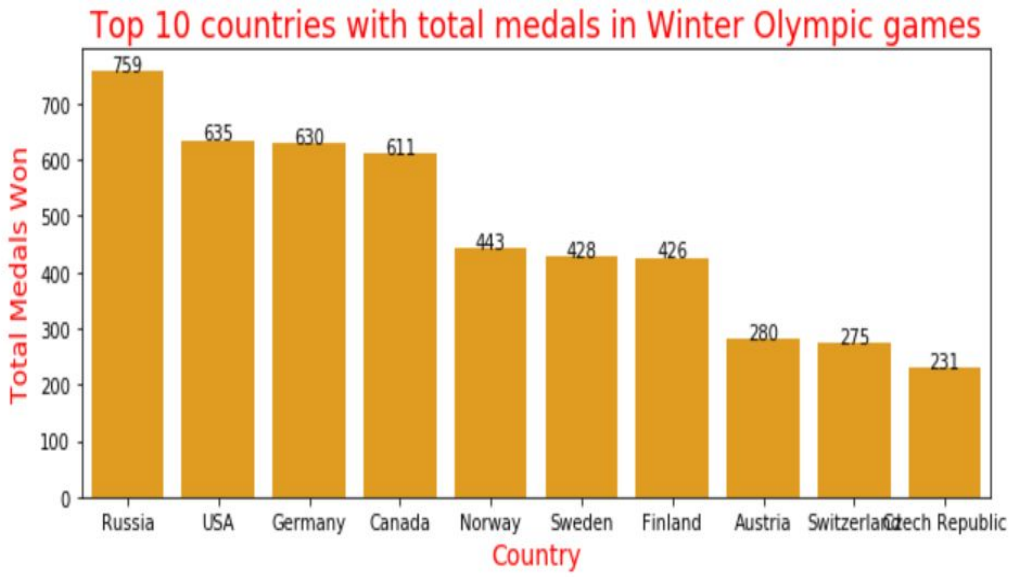


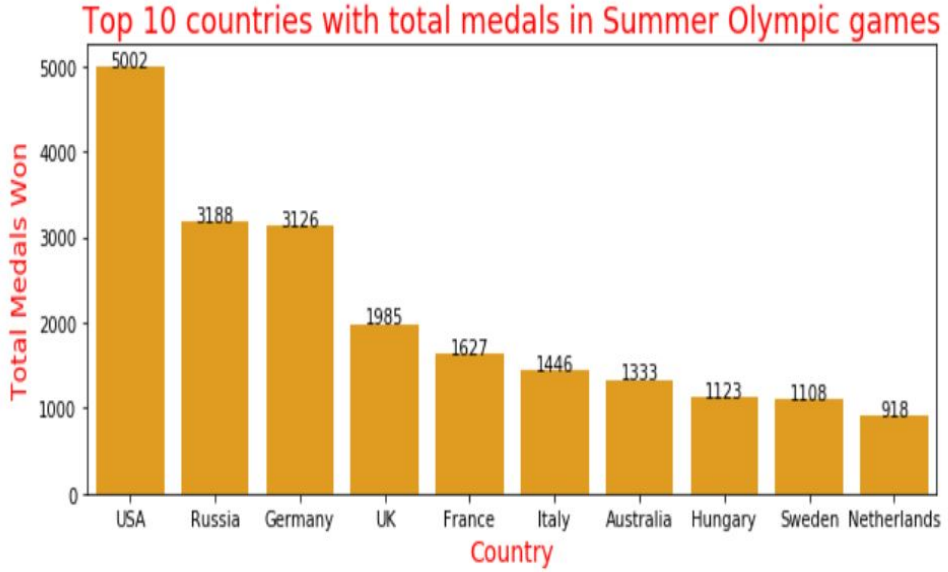
Analyzing number of events in summer vs winter over the years



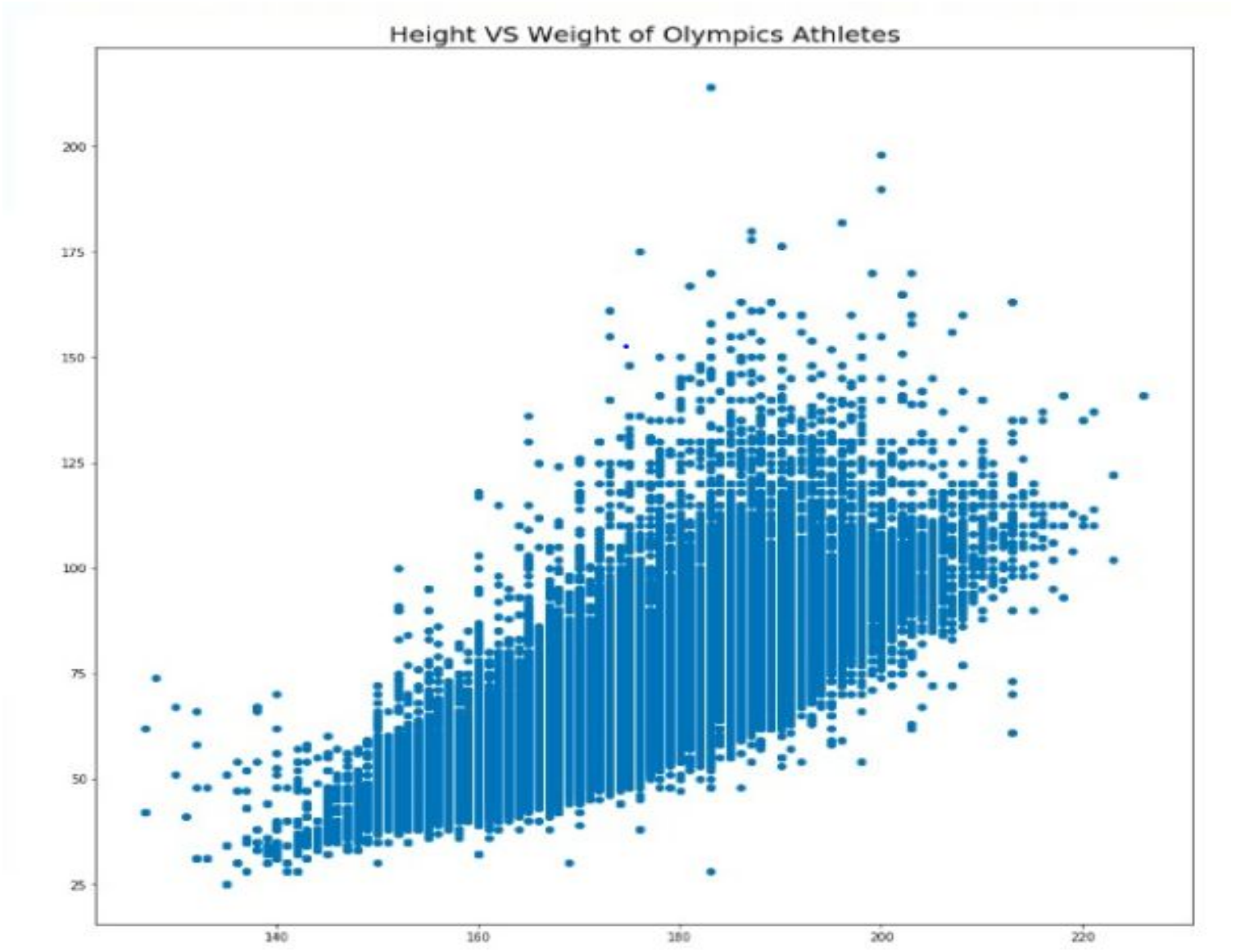


Calculating top performing countries in summer vs winter

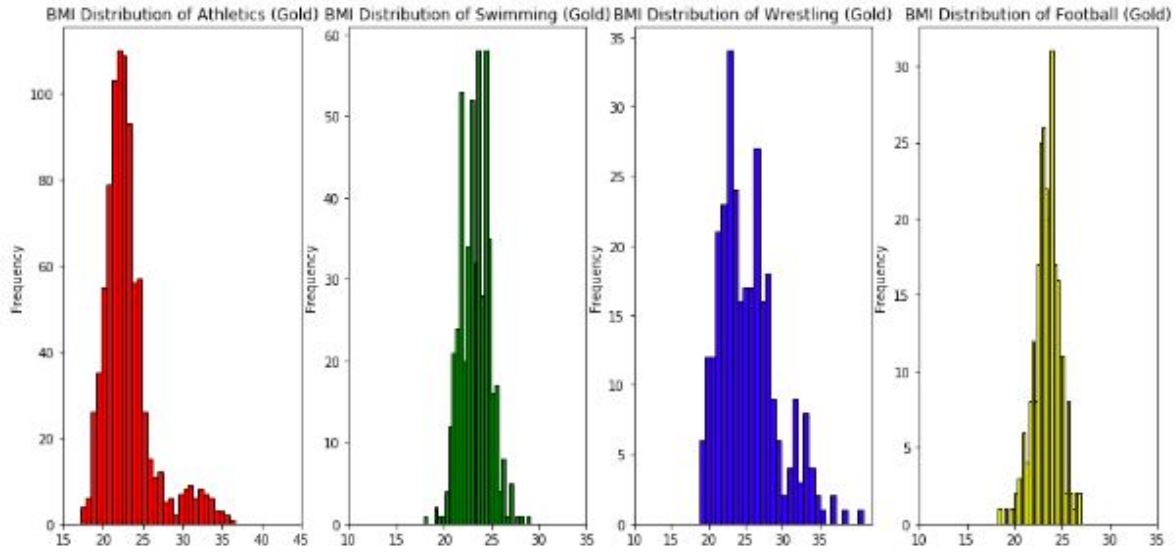




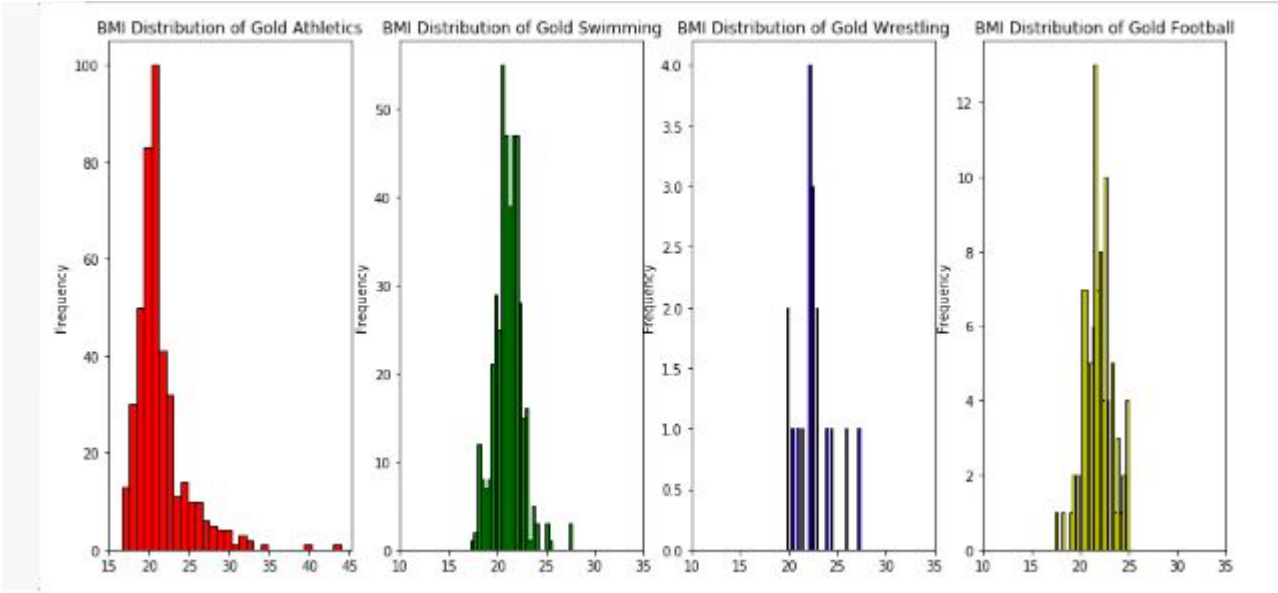
Distribution of Age vs Height from 1896 to 2016



BMI distribution of male vs female

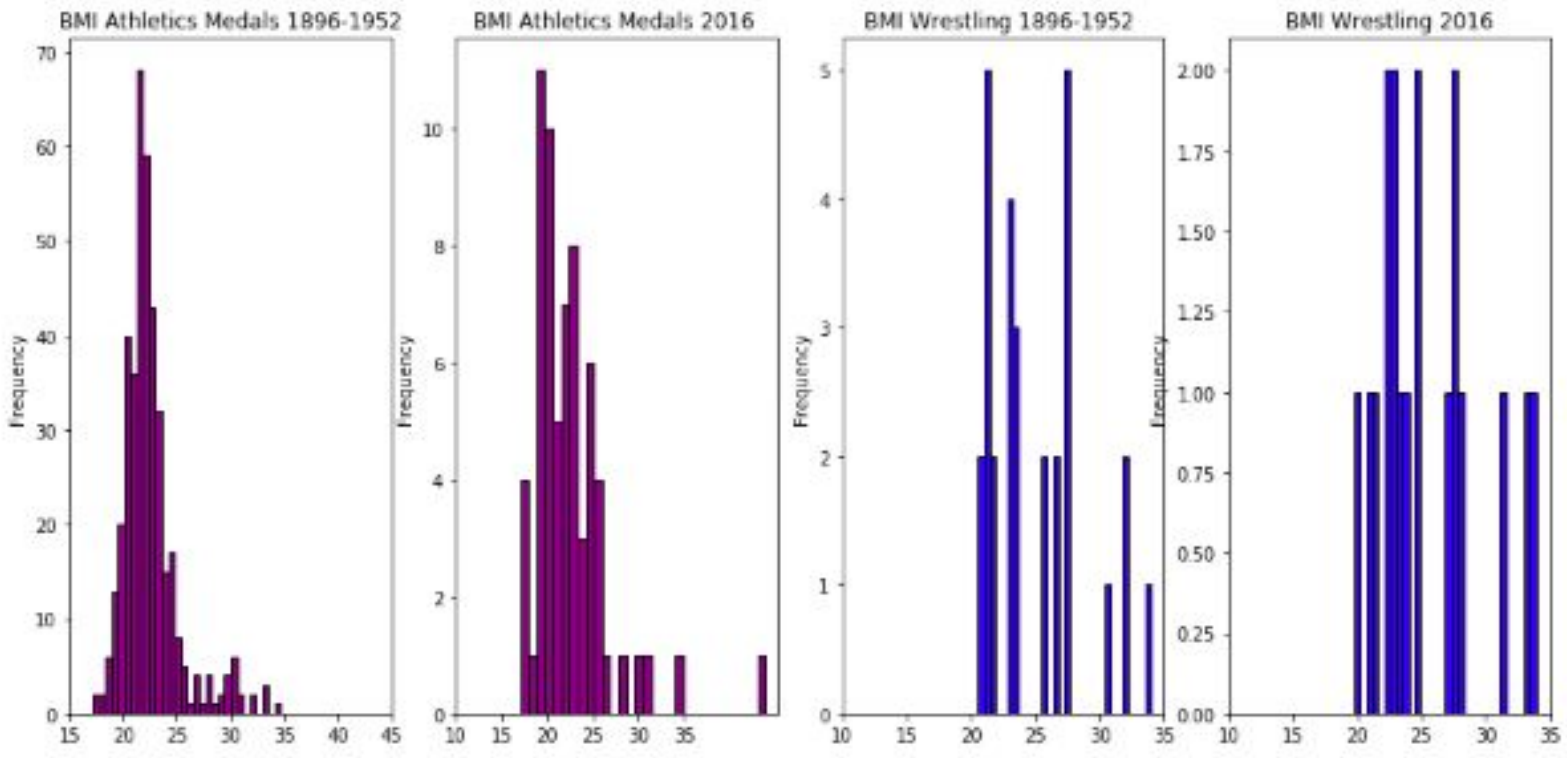


Males



Females

BMI analysis over the years:



1. BMI of Athletes (1896 – 1928)
2. BMI of Athletes Gold Medalists (2016)
3. BMI of Wrestling Events (1896 – 1928)
4. BMI of Wrestling Events Gold Medalists 2016

Physiological Advantage?

Out[49]:

Year	Height
1896	176.285714
1900	175.714286
1904	176.333333
1906	177.181818
1908	176.500000
1912	177.787879
1920	178.000000
1924	172.615385
1928	174.500000
1932	175.541667
1936	178.172414
1948	177.057143
1952	175.200000
1956	175.633333
1960	176.101695
1964	175.575342
1968	175.384615
1972	176.148148
1976	176.225806
1980	176.288136
1984	176.632911
1988	177.232323
1992	178.246753
1996	178.280000
2000	177.817204
2004	177.682927
2008	178.378378
2012	177.492958
2016	177.064103

Name: Height, dtype: float64

Average Height

Year	Height
1896	183.0
1900	167.0
1904	167.0
1906	167.0
1908	170.0
1912	182.0
1920	172.0
1924	183.0
1928	170.0
1932	170.0
1936	178.0
1948	178.0
1952	173.0
1956	186.0
1960	182.0
1964	183.0
1968	183.0
1972	183.0
1976	187.0
1980	183.0
1984	188.0
1988	188.0
1992	189.0
1996	183.0
2000	176.0
2004	186.0
2008	196.0
2012	196.0
2016	196.0

Name: Height, dtype: float64

Gold Medalists

Year	Height
1896	173.333333
1900	176.000000
1904	174.500000
1906	171.000000
1908	174.000000
1912	177.666667
1920	175.000000
1924	175.333333
1928	177.000000
1932	176.333333
1936	178.000000
1948	181.000000
1952	179.333333
1956	179.666667
1960	183.666667
1964	176.666667
1968	179.666667
1972	183.000000
1976	181.666667
1980	178.666667
1984	181.000000
1988	185.000000
1992	181.333333
1996	179.666667
2000	178.000000
2004	185.666667
2008	186.333333
2012	187.333333
2016	186.000000

Name: Height, dtype: float64

All Medalists

Height Comparison of Men's 100m Runners

Standards on age, height and weight in Olympic running events for men.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1859632/?page=2

Physiological Advantage?

Year	
1896	70.857143
1900	67.571429
1904	68.833333
1906	70.090909
1908	69.368421
1912	69.200000
1920	71.076923
1924	67.195652
1928	67.360000
1932	70.117647
1936	72.571429
1948	69.457143
1952	68.933333
1956	69.733333
1960	70.186441
1964	70.273973
1968	70.630769
1972	69.666667
1976	71.354839
1980	69.864407
1984	70.607595
1988	72.292929
1992	74.052632
1996	74.260000
2000	73.771739
2004	74.475610
2008	77.000000
2012	74.084507
2016	74.423077

Average Weight

Year	
1896	66.0
1900	58.0
1904	64.0
1906	64.0
1908	61.0
1912	73.0
1920	75.0
1924	75.0
1928	56.0
1932	65.0
1936	71.0
1948	69.0
1952	67.0
1956	75.0
1960	71.0
1964	86.0
1968	81.0
1972	80.0
1976	90.0
1980	86.0
1984	80.0
1988	80.0
1992	94.0
1996	82.0
2000	80.0
2004	80.0
2008	95.0
2012	95.0
2016	95.0

Name: Weight, dtype: float64

Gold Medalists

Year	
1896	63.666667
1900	64.333333
1904	67.500000
1906	67.000000
1908	66.333333
1912	70.000000
1920	76.000000
1924	66.666667
1928	71.666667
1932	73.333333
1936	77.666667
1948	71.000000
1952	68.000000
1956	70.666667
1960	71.000000
1964	76.666667
1968	76.333333
1972	79.666667
1976	80.000000
1980	76.666667
1984	75.000000
1988	81.000000
1992	79.000000
1996	76.666667
2000	77.666667
2004	79.666667
2008	85.000000
2012	85.000000
2016	81.666667

All Medalists

Weight Comparison of Men's 100m Runners

Standards on age, height and weight in Olympic running events for men.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1859632/?page=2

Can we predict the medal tally of a country ?

Dimension	Dependent Variable:	Indicator	Hypothesized Sign
Olympic Performance	Medal Count (M_t)	The number of medal won by a country in a particular Olympics	N/A
	Independent Variables:		
Demographic Environment	Population (N_t)	The population size of a country at a particular Olympic year	+
Economic Development	GDP per capita (Y_t/N_t)	The per capita GDP (measured in PPP current international dollars) of a country at a particular Olympic year	+
Social Development	Human Development Index HDI_t	The Human Deployment Index of a country at a particular Olympic year	+
Political Environment	Political System (Pol)	1 if the country is or used to be a socialist country or 0 otherwise	+
Geographic Environment	Hosting country (H_t)	1 if the country is the hosting country of the year or 0 otherwise	+

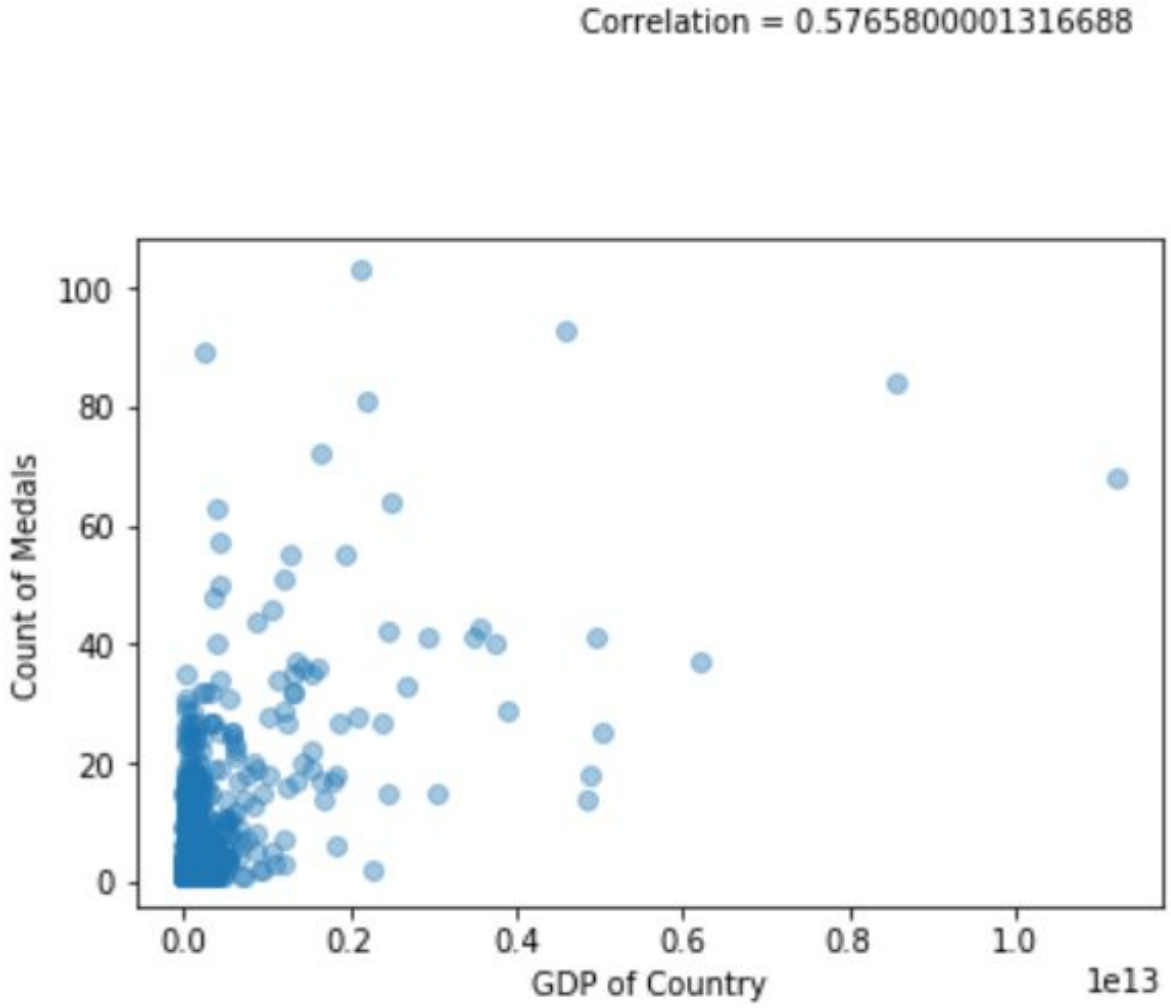
Model - Linear Regression

$$M_t = \beta_0 + \beta_1 (N_t) + \beta_2 (Y_t/N_t) + \beta_3 HDI_t + \beta_4 (Pol) + \beta_5 H_t + E$$

Predictions -
R-Squared ~ 0.75
RMSE ~ 7

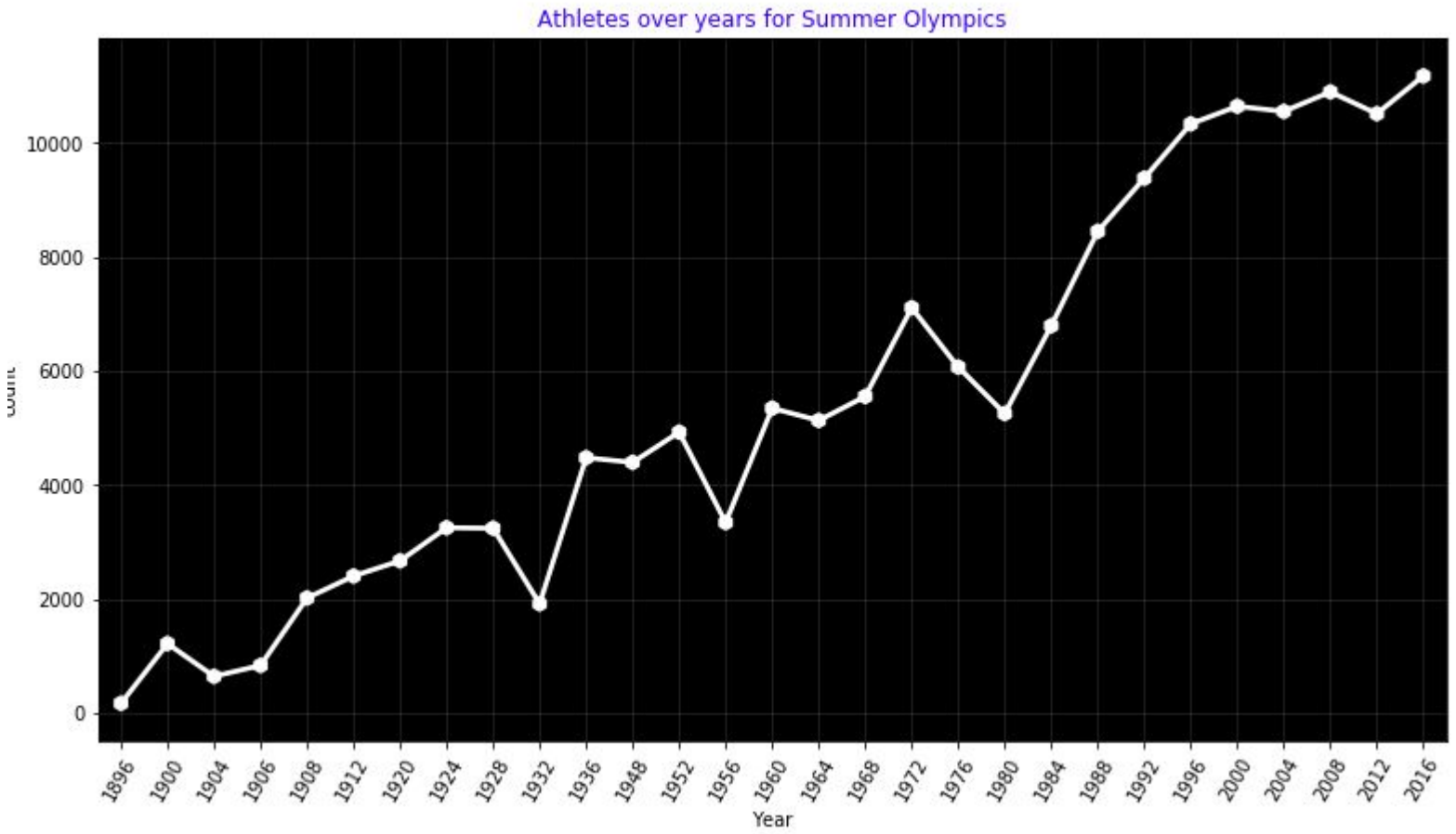
Reference - “Going for Gold Medals: Factors affecting Olympic Performance”
“https://www.researchgate.net/publication/278585677_Going_for_Gold_Medals_Factors_affecting_Olympic_Performance

How does GDP affect one of the most popular country ie USA, Germany, China ?



Dataset - https://datahub.io/core/gdp

Does History Impact the olympics?



Does History Impact the olympics?

1936 Summer Olympics (Berlin, Germany) -

Hitler’s rise to power, less countries were invited , Racism .

1940 and 1944 Summer Olympics (not held due to World War II)

1956 Summer Olympics (Melbourne, Australia) -

Suez Crisis (Egypt, Iraq, and Lebanon) , Hungarian revolution (Netherlands, Spain, and Switzerland) , Republic of China (Formosa) .

1980 Summer Olympics (Moscow, Soviet Union) -

Soviet invasion of Afghanistan (USA and its allies boycotted)

1984 Summer Olympics (Los Angeles, United States) -

Russia Boycotted (Security Reasons)

Optimizations

1. **Index Optimalization (merge tables by index)**
2. **Use vectorized operations (.iloc , .apply , lambda expressions)**
3. **Numba (Optimize Numpy)**
4. **Modin (read csv)**

Reference -

1. https://towardsdatascience.com/understanding-the-need-for-optimization-when-using-pandas-8ce23b83330c
2. https://towardsdatascience.com/understanding-the-need-for-optimization-when-using-pandas-8ce23b83330c

THANK YOU!