Project

- Data Analysis Project for Sports-Stats
    - Shivam Chaturvedi

Client/dataset

1. Sports-Stats/Olympic dataset

Data

1. I imported the data in jupyter notebook using read_csv of pandas library and I am also using pandasql library to run SQL queries on it
2. The data was of both Summer and Winter olympic but I filtered it for only Summer Olympic
3. There are NaN values in age,height,weight,medal column of data but we don’t need to clean them
4. I also scraped GDP per capita data for countries for world bank website to analyse some question

SECTION 1: Questions to Answer

1. How has the number of participating countries changed over time ?
2. How has the number of Sports changed over time?
3. Which countries have won the most medals in the Summer Olympics ?
4. How has the number of male and female athletes changed over time ?
5. Which athletes have won the most medals in olympic history?
6. Which sports are dominated by which countries ?
7. How does athletes age, height,weight affect performance ?
8. How does the number of medals won correlate with country’s GDP per capita ?
9. How does the athlete count of countries correlate with its performance(medal count)?
10. How does the number of sports countries participate correlate with its performance(medal count)?

SECTION 2: Initial Hypothesis

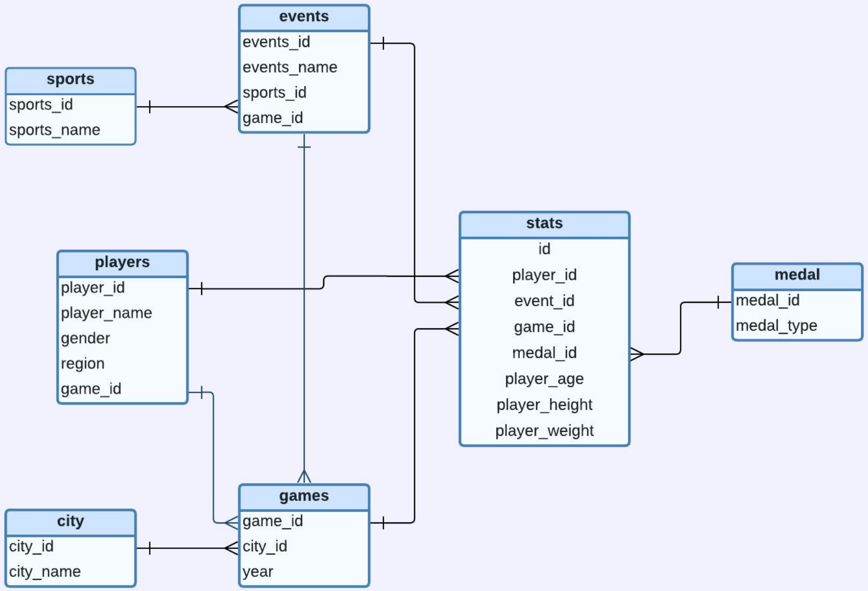
1. I believe the number of participating countries will increase over time
2. I believe most medal won will be either by CHINA or USA
3. I think number of female athletes will increase over time
4. I think number of medals won by countries will positively correlate with gdp per capita
5. Michael phelps probably has won the most medals
6. I think table tennis is always dominated by china, USA for basketball etc

SECTION 3: Initial approach

1) I will looking at these metrics to judge performance of countries and athletes
    a) Medal count
    b) Gender count
    c) Sports count
    d) Athlete count
    e) Medal - Gold, Silver ,Bronze
2) I will be looking at pearson coefficient for correlation to justify If a correlation exists
3) I will also be looking at yearwise trends

Entity Relationship Diagram

Entity Relationship Diagram
(olympic dataset)



events
events_id
events_name
sports_id
game_id

sports
sports_id
sports_name

stats
id
player_id
event_id
game_id
medal_id
player_age
player_height
player_weight

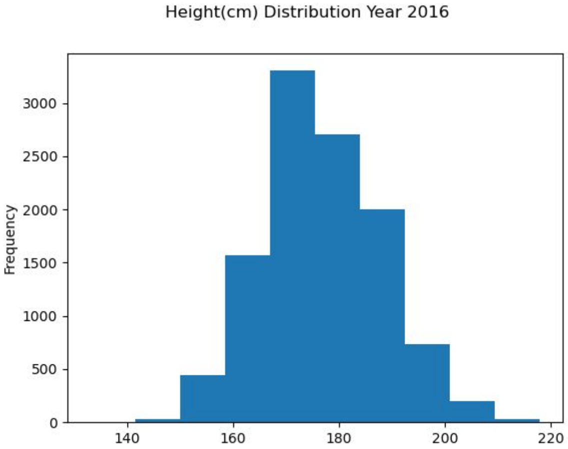
medal
medal_id
medal_type

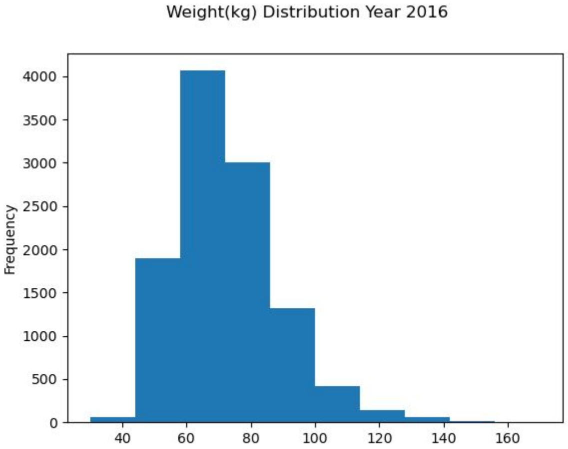
players
player_id
player_name
gender
region
game_id

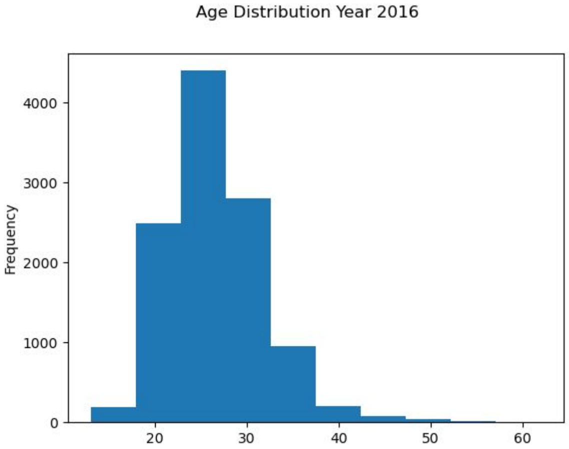
city
city_id
city_name

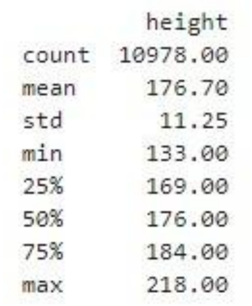
games
game_id
city_id
year

Initial Findings

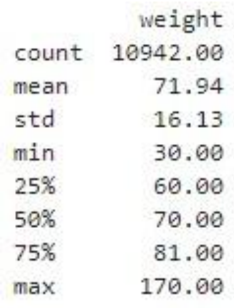




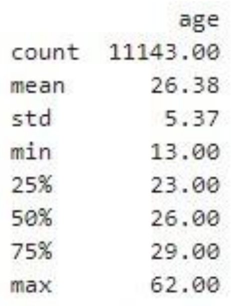




```
         height
count  10978.00
mean     176.70
std       11.25
min      133.00
25%      169.00
50%      176.00
75%      184.00
max      218.00
```



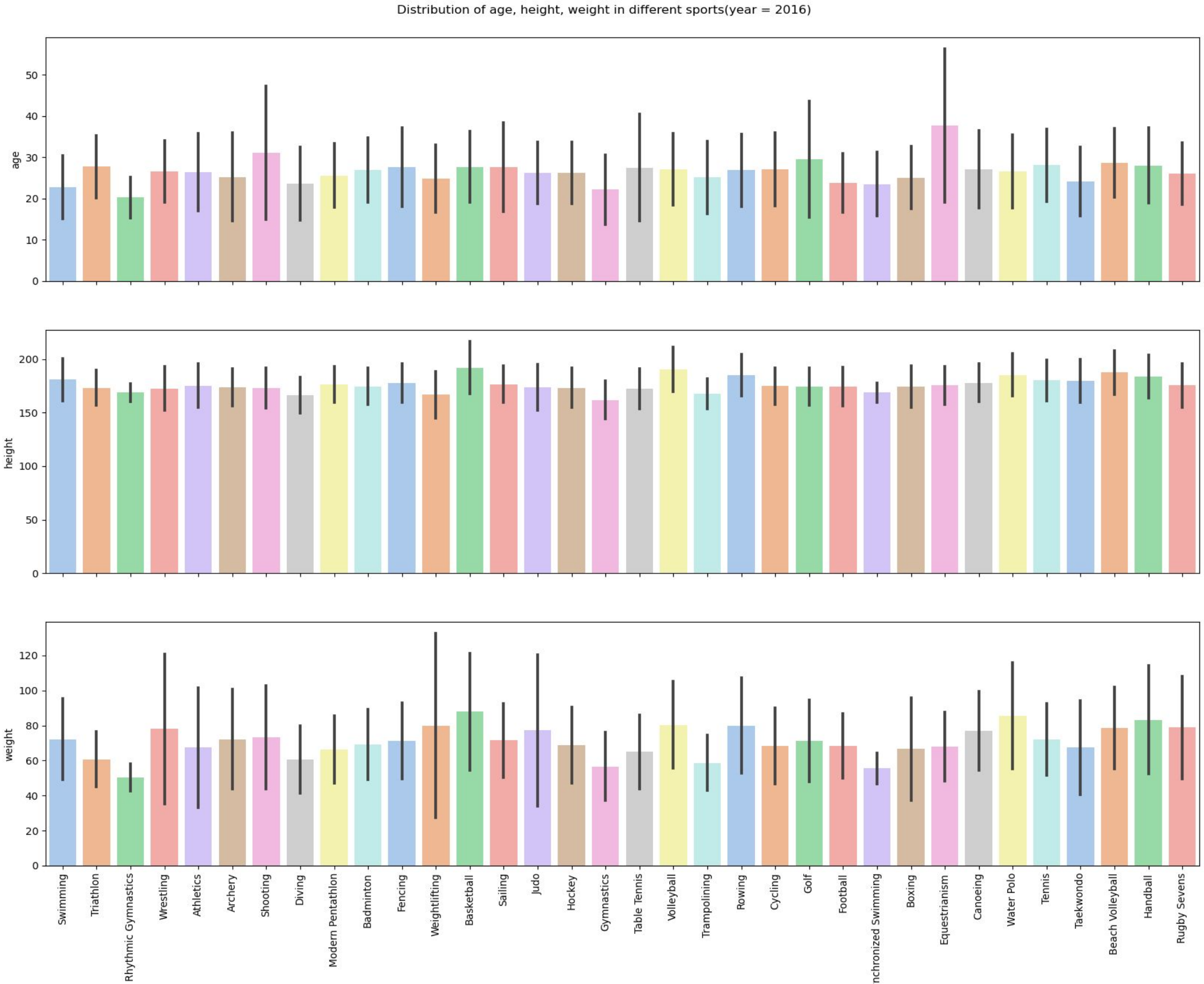
```
         weight
count  10942.00
mean      71.94
std       16.13
min       30.00
25%       60.00
50%       70.00
75%       81.00
max      170.00
```



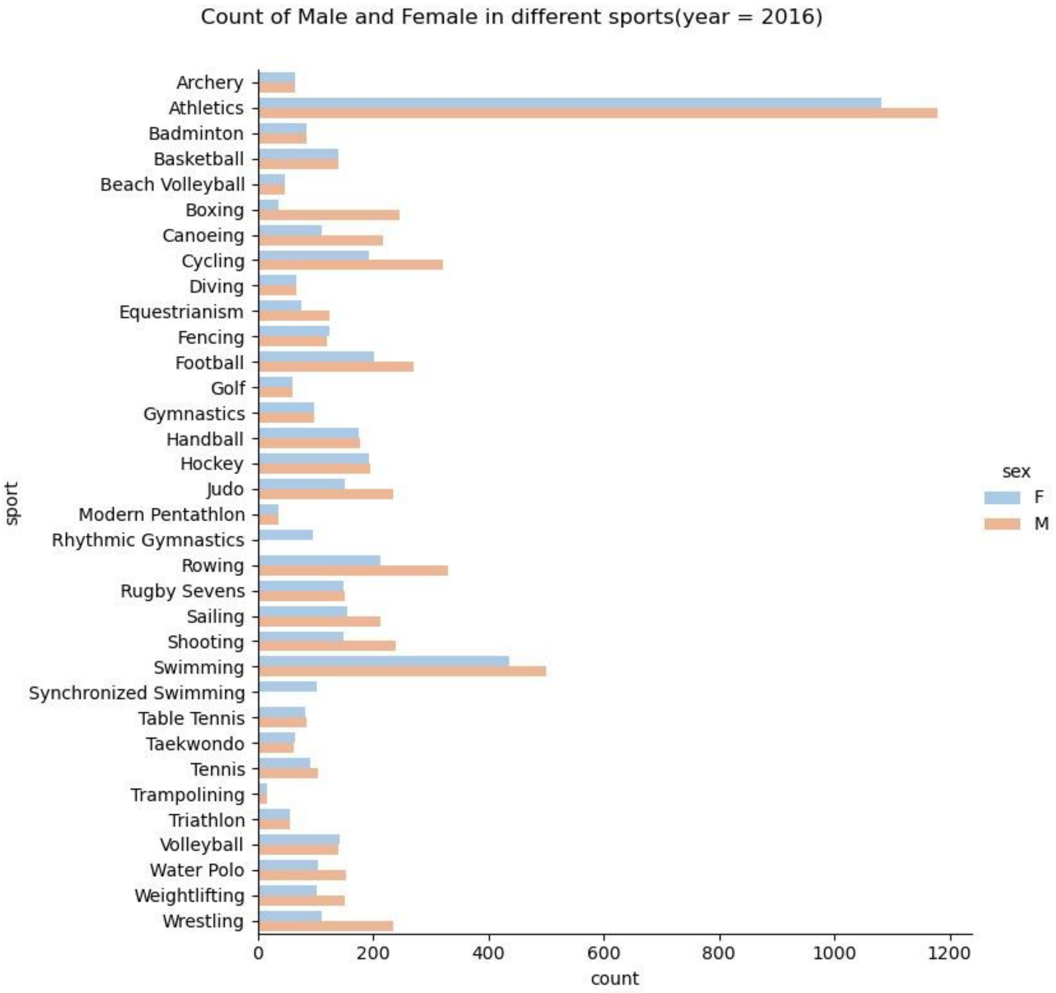
```
            age
count  11143.00
mean      26.38
std        5.37
min       13.00
25%       23.00
50%       26.00
75%       29.00
max       62.00
```

The graph is showing mean ± 2 SD of height, age weight in different sports in Olympic 2016

- In height you can notice basketball mean is greater than every other sports
- In equestrianism apparently a lot 35 + age athletes are there
- Mean Age, weight of athletes in rhythmic gymnastics is minimum with less variability

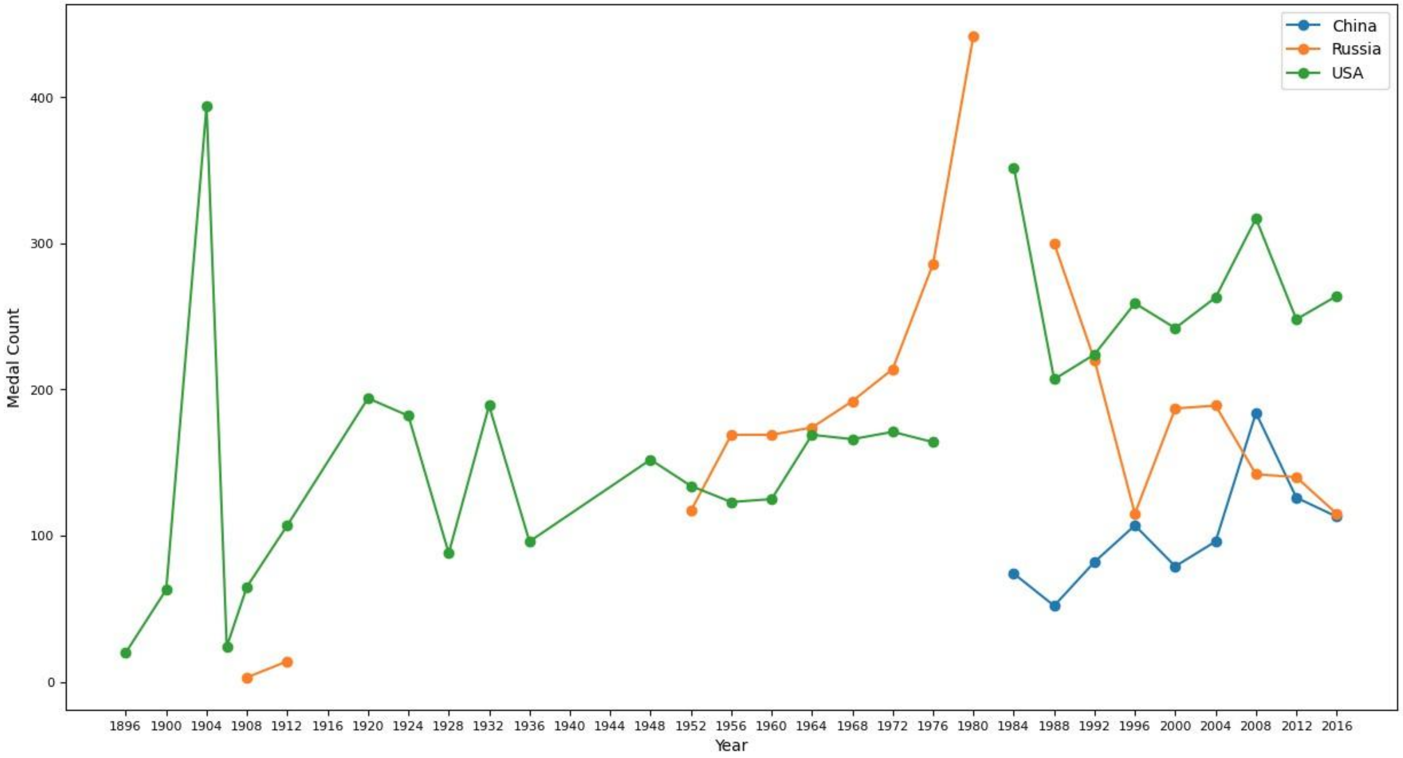


- Rhythmic Gymnastics is all female
- Synchronized Swimming is all female
- Maximum difference between male and female is in Boxing
- A lot of sports are very even in terms of gender



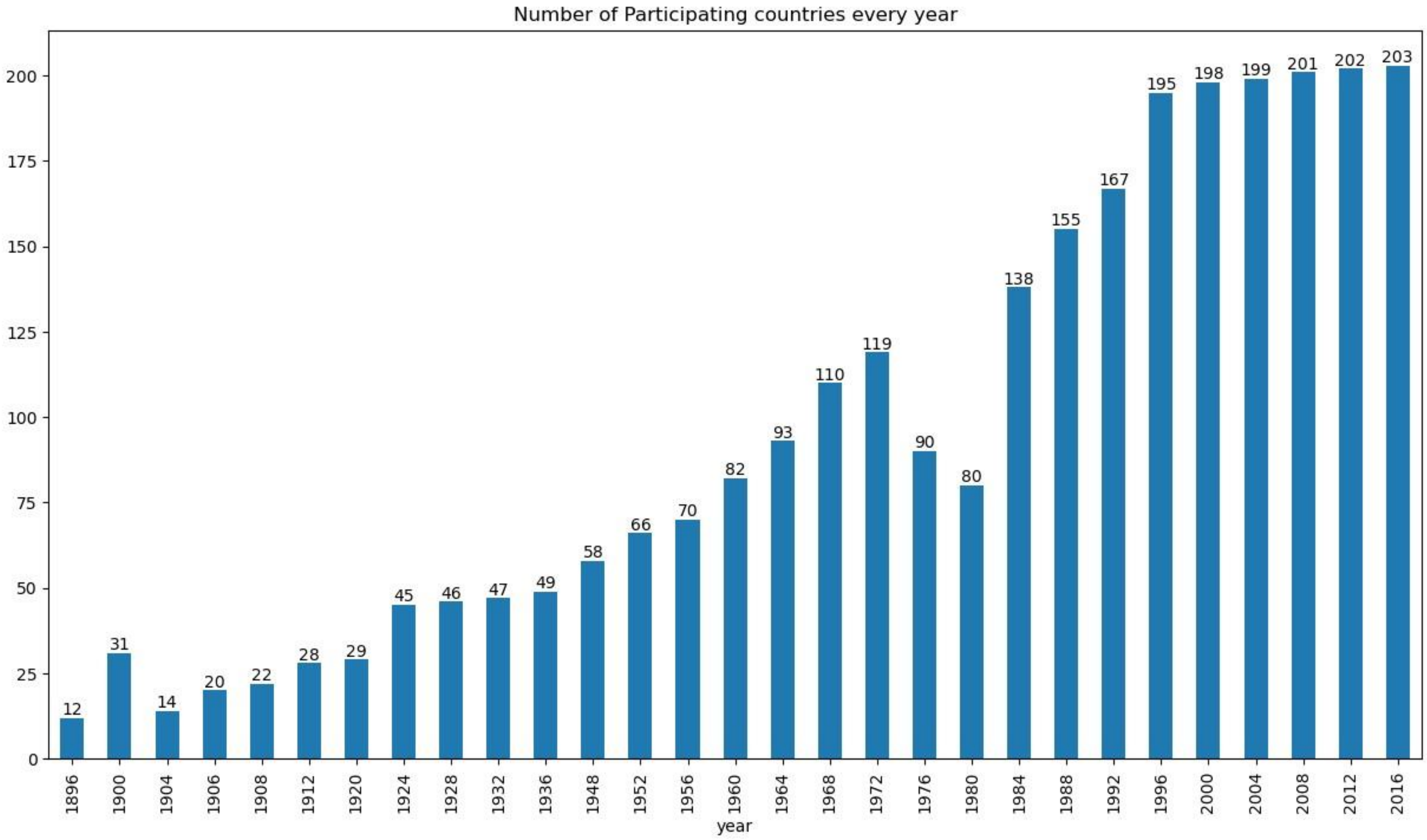
Medal count over the years for USA, Russia, China

- You can see China and Russia joined quite late
- US is still at the top as far as medal count is concerned

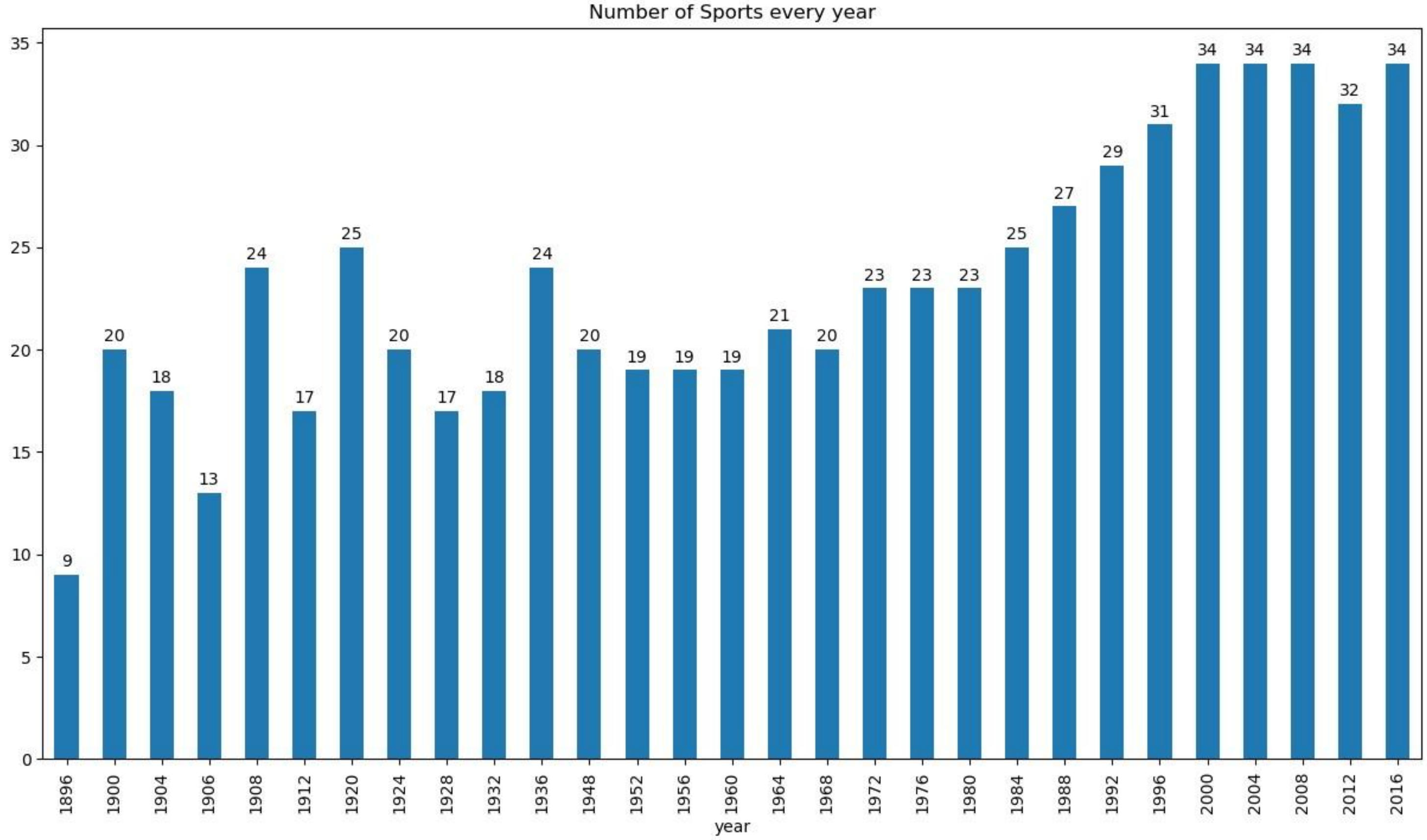


(Check out the source code to see every country trend)

How has the number of participating countries changed over time ?

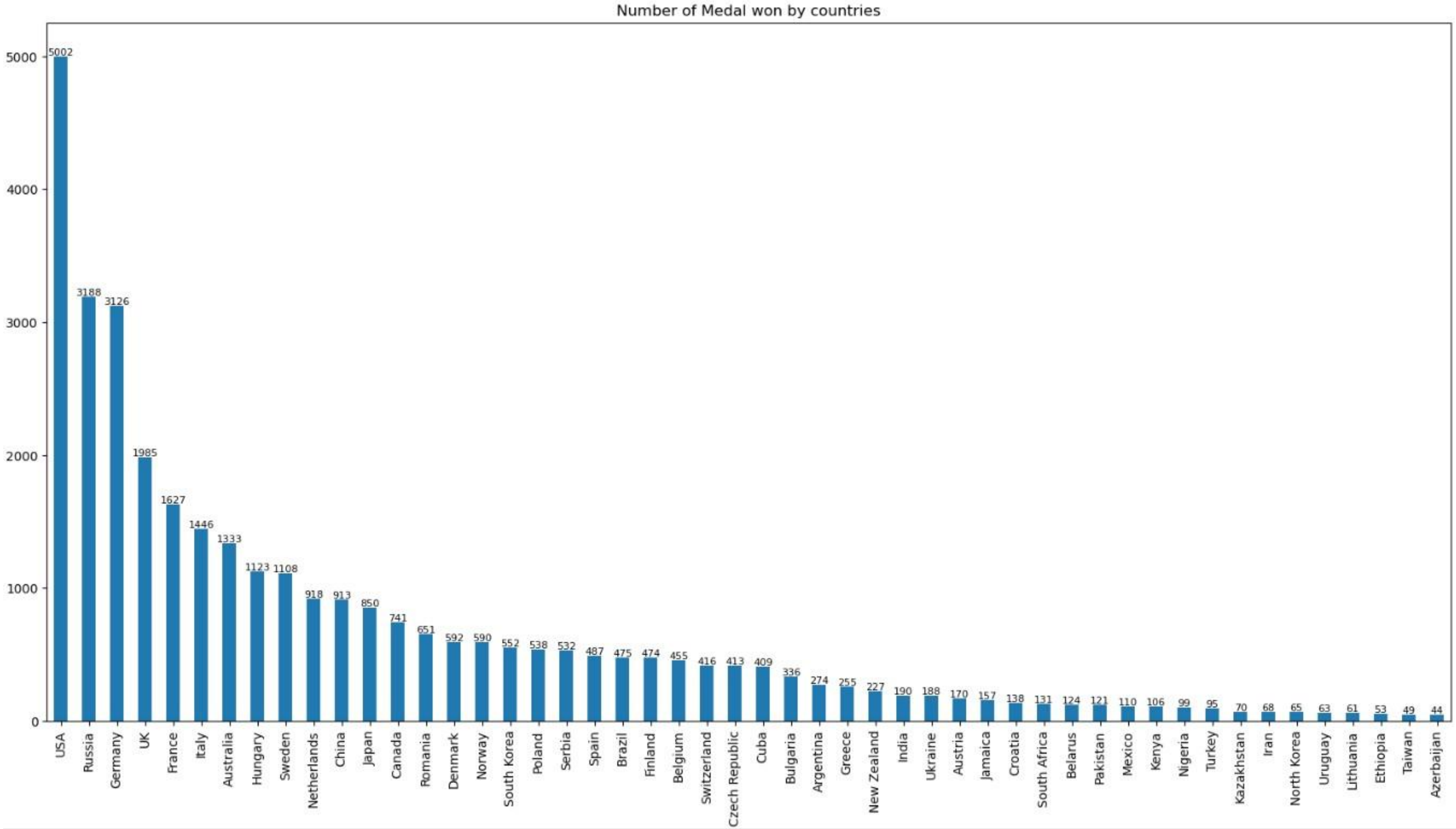


How has the number of Sports changed over time?



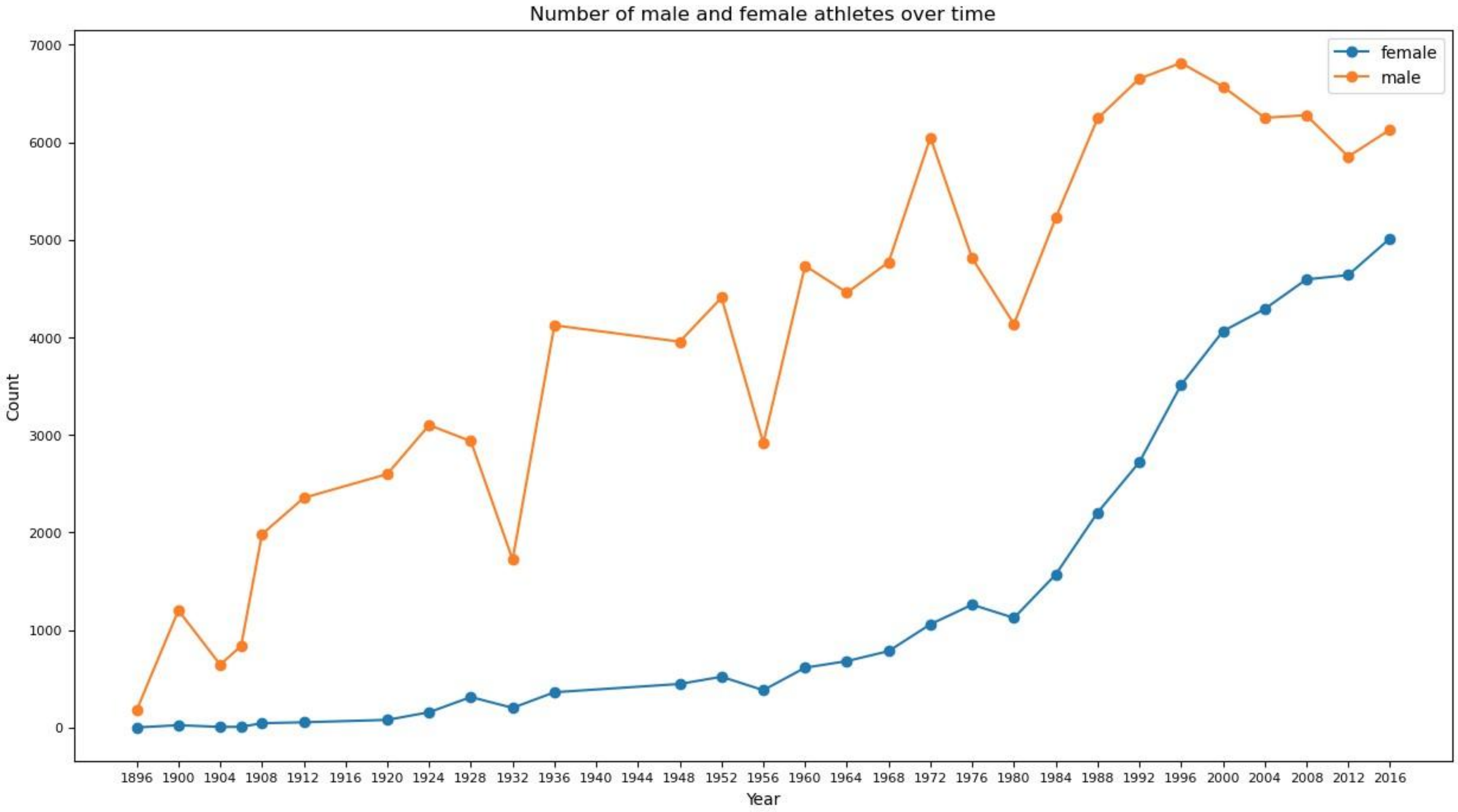
Which countries have won the most medals in the Summer Olympics ?

- USA has won the most medal in Olympic history(5002) ,Russia 2nd(3188), Germany 3rd(3126), UK 4th(1985), although USA and UK has been In the Olympics from very start while other countries came later.

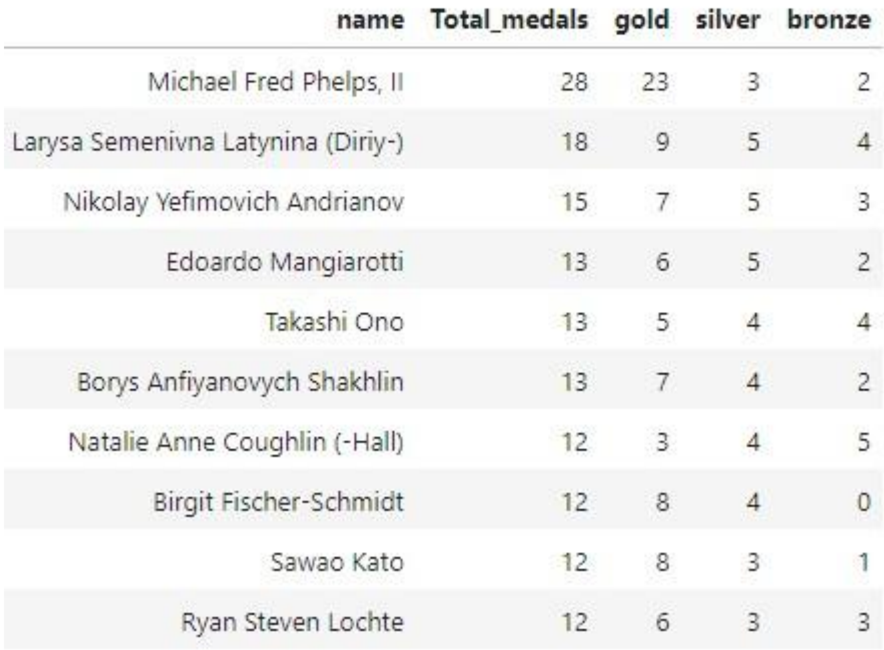


How has the number of male and female athletes changed over time ?

- Females are slowly catching up to male which is a good thing



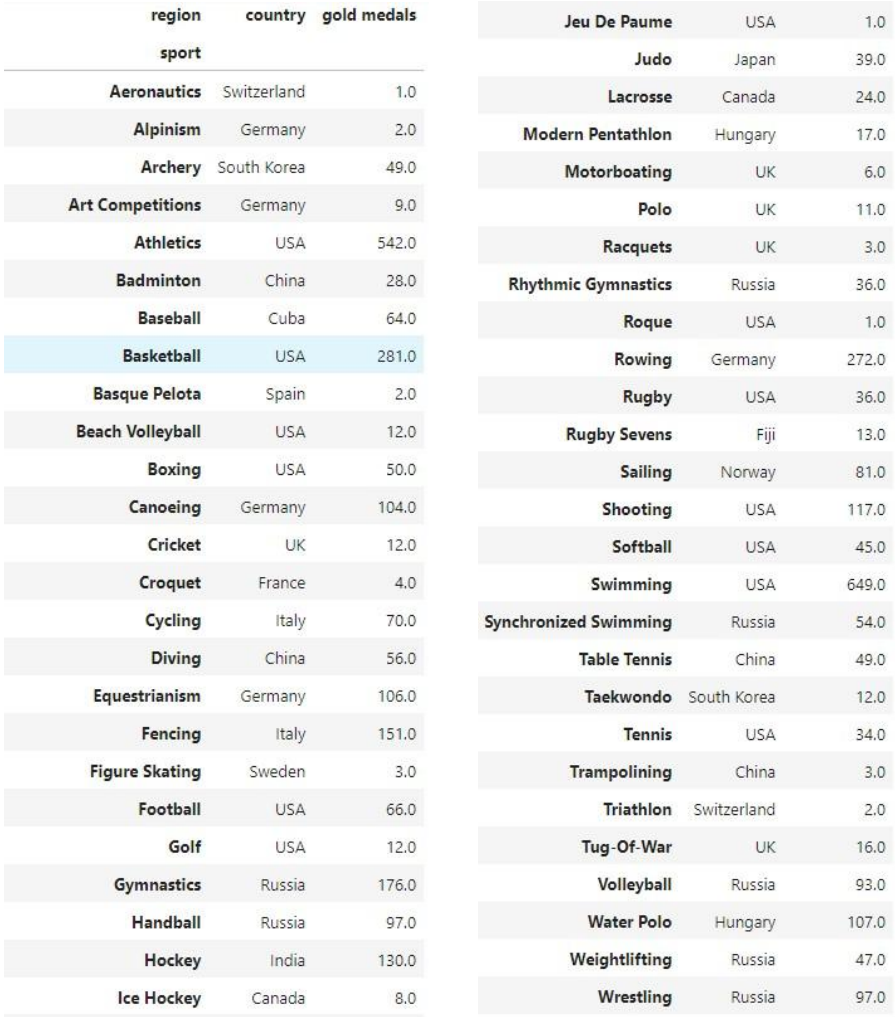
Which athletes have won the most medals in olympic history?



name	Total_medals	gold	silver	bronze
Michael Fred Phelps, II	28	23	3	2
Larysa Semenivna Latynina (Diriy-)	18	9	5	4
Nikolay Yefimovich Andrianov	15	7	5	3
Edoardo Mangiarotti	13	6	5	2
Takashi Ono	13	5	4	4
Borys Anfiyanovych Shakhlin	13	7	4	2
Natalie Anne Coughlin (-Hall)	12	3	4	5
Birgit Fischer-Schmidt	12	8	4	0
Sawao Kato	12	8	3	1
Ryan Steven Lochte	12	6	3	3

Which sports are dominated by which countries ?

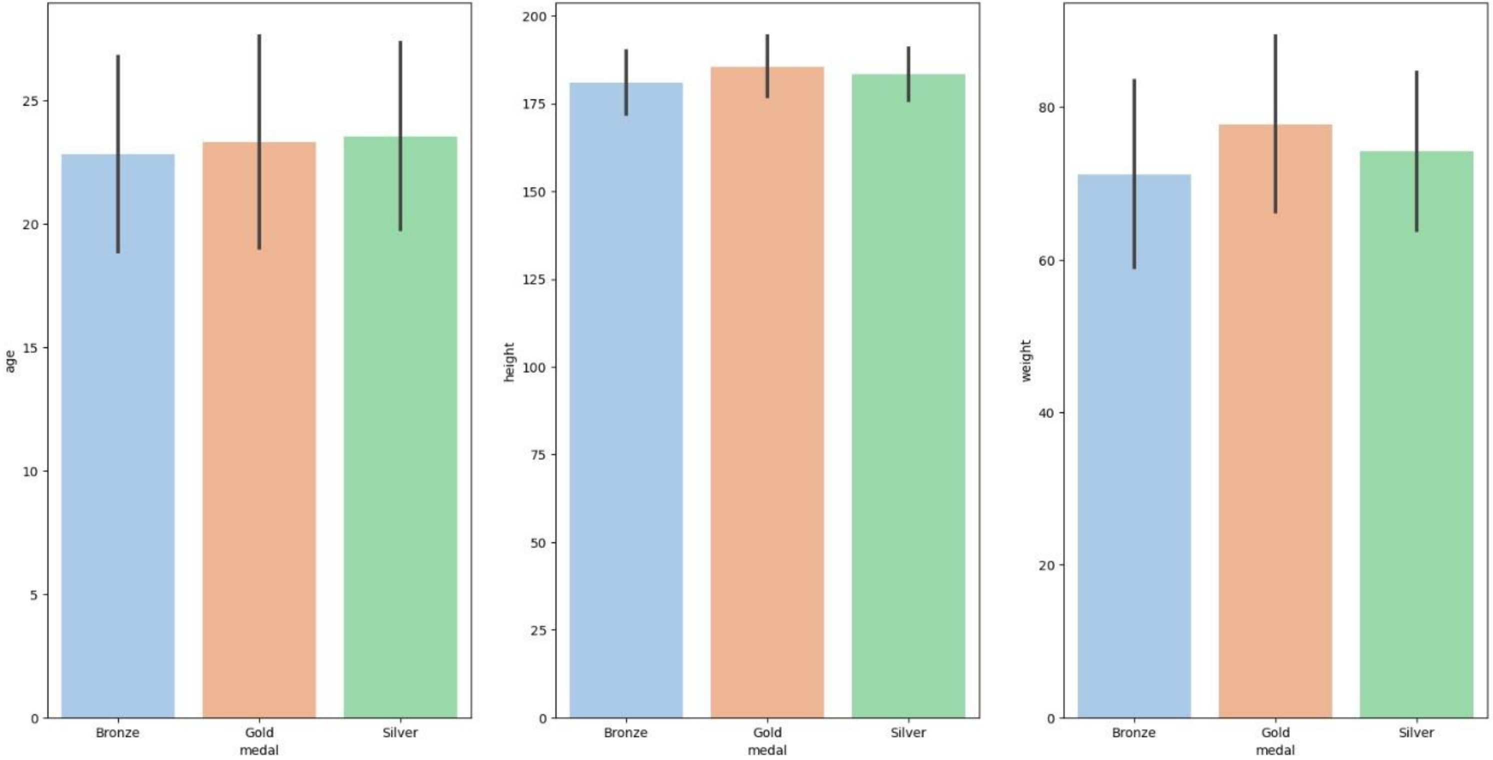
- USA dominates athletics, Basketball, swimming, Boxing etc
- China dominates table tennis, Badminton etc
- India has maximum golds in hockey but its recent performances has been not that good



region	country	gold medals
sport		
Aeronautics	Switzerland	1.0
Alpinism	Germany	2.0
Archery	South Korea	49.0
Art Competitions	Germany	9.0
Athletics	USA	542.0
Badminton	China	28.0
Baseball	Cuba	64.0
Basketball	USA	281.0
Basque Pelota	Spain	2.0
Beach Volleyball	USA	12.0
Boxing	USA	50.0
Canoeing	Germany	104.0
Cricket	UK	12.0
Croquet	France	4.0
Cycling	Italy	70.0
Diving	China	56.0
Equestrianism	Germany	106.0
Fencing	Italy	151.0
Figure Skating	Sweden	3.0
Football	USA	66.0
Golf	USA	12.0
Gymnastics	Russia	176.0
Handball	Russia	97.0
Hockey	India	130.0
Ice Hockey	Canada	8.0
Jeu De Paume	USA	1.0
Judo	Japan	39.0
Lacrosse	Canada	24.0
Modern Pentathlon	Hungary	17.0
Motorboating	UK	6.0
Polo	UK	11.0
Racquets	UK	3.0
Rhythmic Gymnastics	Russia	36.0
Roque	USA	1.0
Rowing	Germany	272.0
Rugby	USA	36.0
Rugby Sevens	Fiji	13.0
Sailing	Norway	81.0
Shooting	USA	117.0
Softball	USA	45.0
Swimming	USA	649.0
Synchronized Swimming	Russia	54.0
Table Tennis	China	49.0
Taekwondo	South Korea	12.0
Tennis	USA	34.0
Trampolining	China	3.0
Triathlon	Switzerland	2.0
Tug-Of-War	UK	16.0
Volleyball	Russia	93.0
Water Polo	Hungary	107.0
Weightlifting	Russia	47.0
Wrestling	Russia	97.0

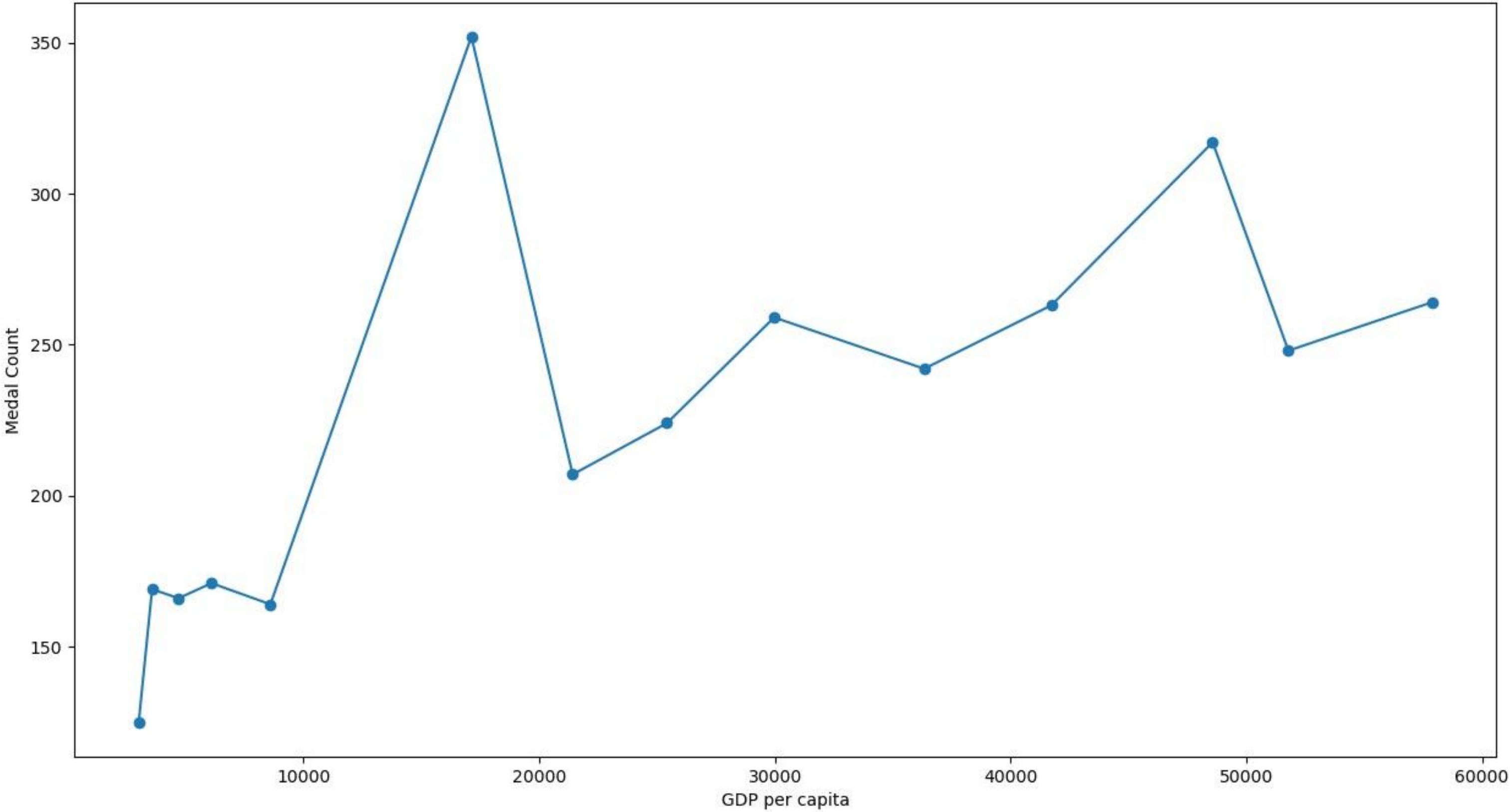
How does athletes age, height,weight affect performance(Swimming, 2016) ?

- Mean 土 SD is plotted for gold, silver, bronze medalist in swimming in year 2016
- There not a much difference and the variability overlaps but mean height and weight of gold medalists is greater than silver and bronze
- We will need further machine learning models to analyse this which I don’t know how to do



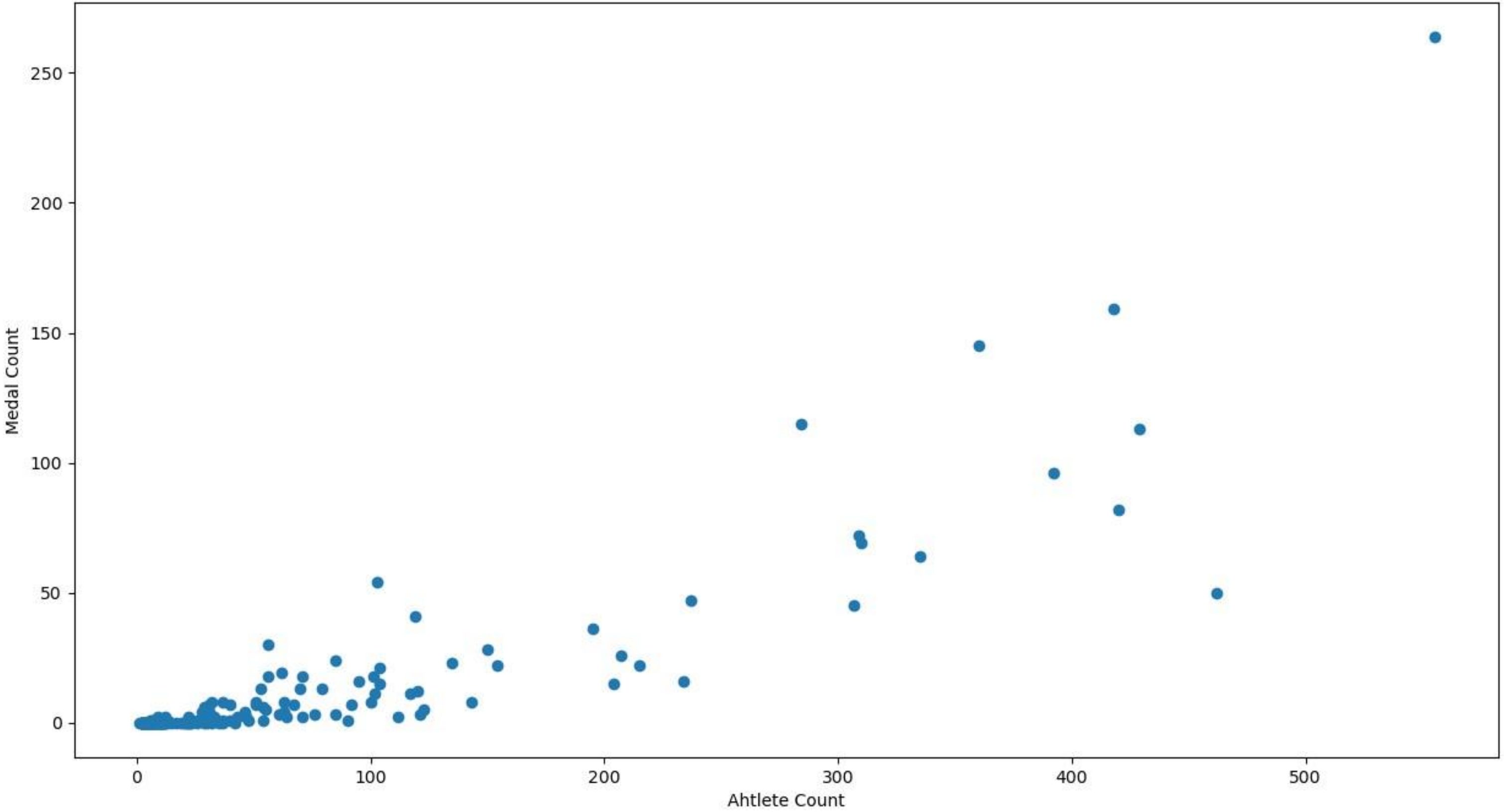
How does the number of medals won correlate with country’s GDP per capita(USA) ?

- I chose USA because it has the most data points in terms of year
- **PearsonRResult(statistic=0.6698466328369534, pvalue=0.00877063273873163)**
- The coefficient shows positive correlation and is statistically significant
- We can correlate that 'USA' performance increases as its GDP per capita increases



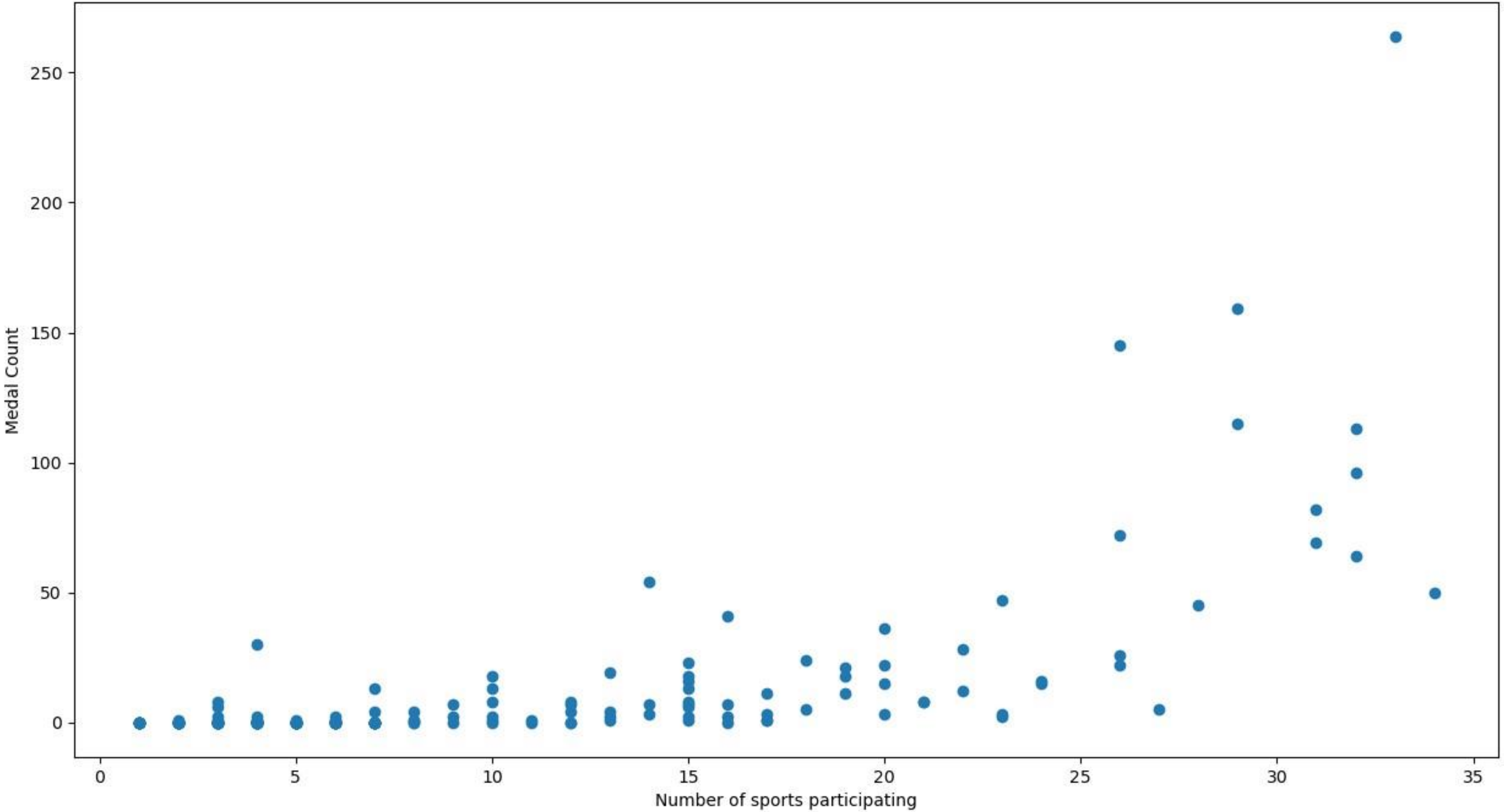
How does the athlete count of countries correlate with its performance(medal count)(Year 2016)?

- **PearsonRResult(statistic=0.8704132288491329, pvalue=8.942668717493107e-64)**
- It shows positive correlation and is statistically significant
- We can correlate that if country sends more athletes its performance also increases
- This could be used as a metric to train ML models to predict countries performance(which I don’t know how to do)



How does the number of sports countries participate correlate with its performance(medal count)(Year 2016)?

- **PearsonRResult(statistic=0.6664813706355885, pvalue=1.940407194827214e-27)**
- It shows positive correlation and is statistically significant
- We can correlate that if country participate in more sports its performance also increases
- This could used as a metric to train ML models to predict countries performance(which I don’t know how to do)



Insights Discovered

- Hypotheses
  - All my initial hypotheses seems correct. One of my hypotheses was that country performance will positively correlate with GDP per capita which is true for USA ,it give good insights on how a country development can in turn help athletes to perform well as they will get better resources to improve. I didn't change any initial hypothesis. I created a new one which correlated athlete count and number of sports participation of a country with performance which also turned out to be positively correlated .
- Metrics
  - Medal Count - To analyse performance of countries and athletes
  - Gold Count - To analyse performance of countries in Q6 which is much better than medal count
  - Athlete Count - To analyse correlation in Q9 and to identify if it can be further used to predict performance of countries using ML models
  - Sports Count - To analyse correlation in Q10 and to identify if it can be further used to predict performance of countries using ML models
  - Medals(Gold, Silver, Bronze) - To check if age,height,weight affect 1st,2nd,3rd positions in Q7

Insights Discovered(cont.)

- Data / themes discovered
    - Women are not behind in any sport now
        - A lot of sports are very even in terms of gender
        - Number of female athletes are increasing every year and slowly catching up with male athletes which is excellent
    - Participation of more and more countries are happening and not just developed countries
    - New sports are getting recognition and getting added in olympic every year
    - Michael Phelps has absurdly high number of gold medals which makes him a historical figure
    - We found correlation “USA performance increases as its GDP per capita increases” which shows good amount of resources when invested in any sports can help improve performance
    - Population and variety of athletes matters
        - We found correlation that if country sends more athletes its performance also increases
        - We found correlation that if country participate in more sports its performance also increases

Summary

- Summary
  - Height, age, weight almost follow a normal distribution(Olympic 2016)
  - The most variability in age is in sports shooting and Equestrianism(Olympic 2016)
  - The most variability in weight is in sport weightlifting(Olympic 2016)
  - Rhythmic Gymnastics is all female, Synchronized Swimming is all female, Maximum difference between male and female is in Boxing, A lot of sports are very even in terms of gender(Olympic 2016)
  - Number of female athletes are increasing every year and slowly catching up with male athletes which is excellent
  - USA dominates athletics, Basketball, swimming, Boxing etc, China dominates table tennis, Badminton etc
  - Number of countries and sports in olympic are increasing every year
  - Michael Phelps has absurdly high number of gold medals
  - We can correlate that 'USA' performance increases as its GDP per capita increases
  - We can correlate that if country sends more athletes its performance also increases
  - We can correlate that if country participate in more sports its performance also increases
- Next Steps
  - Some of the correlations discovered in questions 7-10 could be further analysed using ML models and make predictions