DATA SCIENCE CAPSTONE PROJECT

SPORTS STATS

INSIGHTS INTO THE OLYMPICS

Author - Bhavesh Khamesra

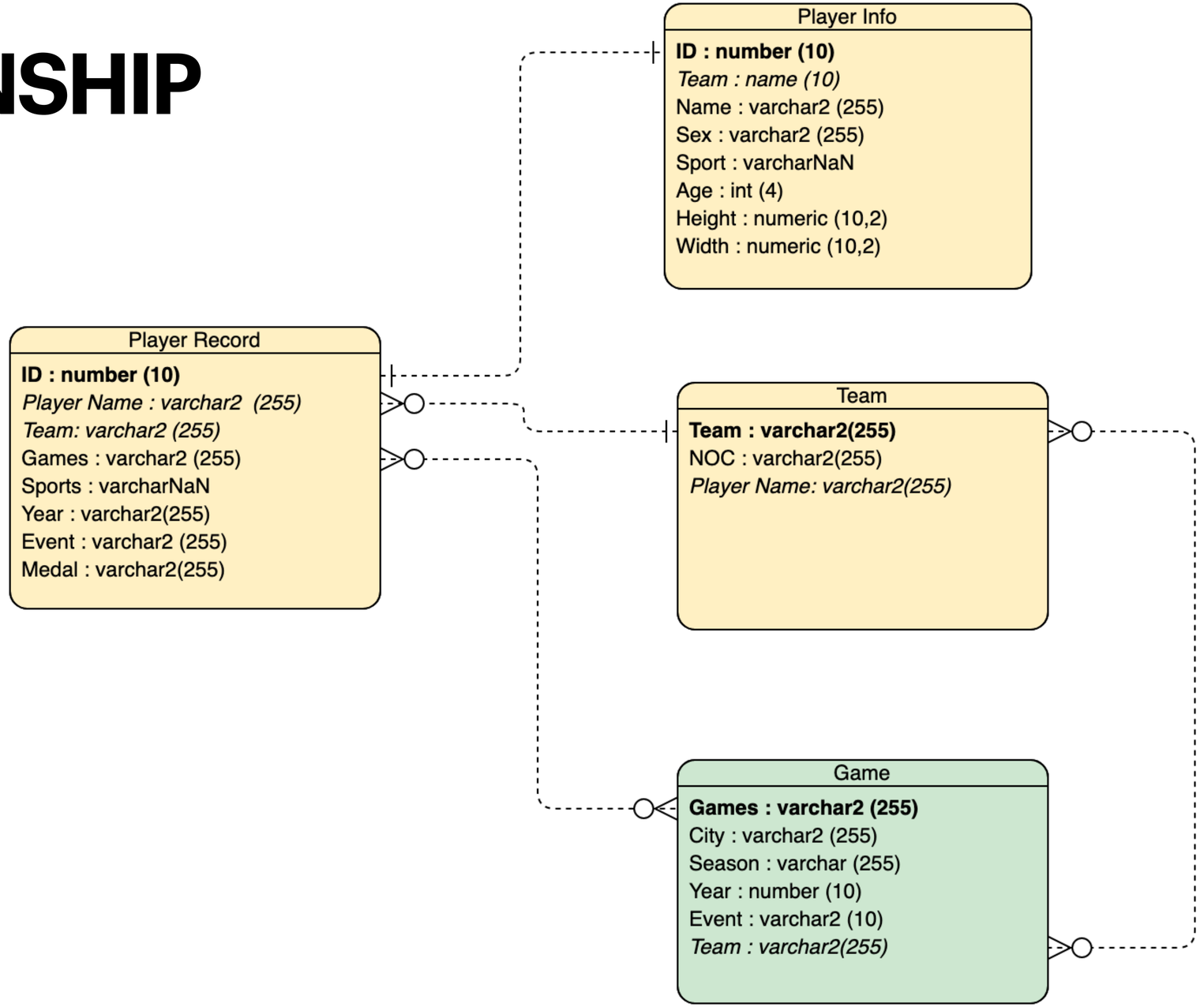
CONTENT

- Introduction
- Proposal - Questions and Hypothesis
- Analysis Methods
- Results
- Final Remarks

INTRODUCTION

- **Dataset** - Provided by SportStats - a sport analysis firm partnering with local news and elite personal trainers.
- **Goal** - Analyse the data for interesting patterns/trends which could be useful to partners in areas of journalism/personal training etc.
- **Dataset Details - Two separate datasets provided.**
- **Dataset 1 contains the following fields**
  - **Player's Personal Details - ID, Name, Age, Sex, Height, Width**
  - **Player's affiliation - Team, NOC**
  - **Player's Sports Records - Games (participated), Sport, Event, Medal**
  - **Games Details - Year, Season, City**
- **Dataset 2 contains list of NOC and corresponding Region.**

Entity relationship Diagram



DEVELOPMENT - Questions & Assumptions

Questions -

1. Does there exist a gender inequality in Olympics?
2. Does there exists an age bias in different sports categories?
3. Are all games dominated by a few countries or the distribution varies depending on the sports?

Assumptions -

- The provided dataset holds no bias against players of any countries, sex, age, sports or any other catagories. The data is not misrepresentated or false.
- The sports itself provides equal opportunities to players of all age, sex and nationality.

HYPOTHESIS

1.Gender Disparity - The representation of males and females may be more fair in developed nations compared to developing nations, where opportunities are not equally available to females. The disparity is also expected to be higher in early years than present.

2. Sports and Optimal Age - Performance of player depends on both agility and experience. Hence, for every sport, there would be an optimal age at which player's peak performance can be expected which can very depending on nature of sports.

3**. Sport domination** - More developed nations would have higher dominance in majority of sports due to better infrastructure and opportunities provided to their players.

Analysis Method

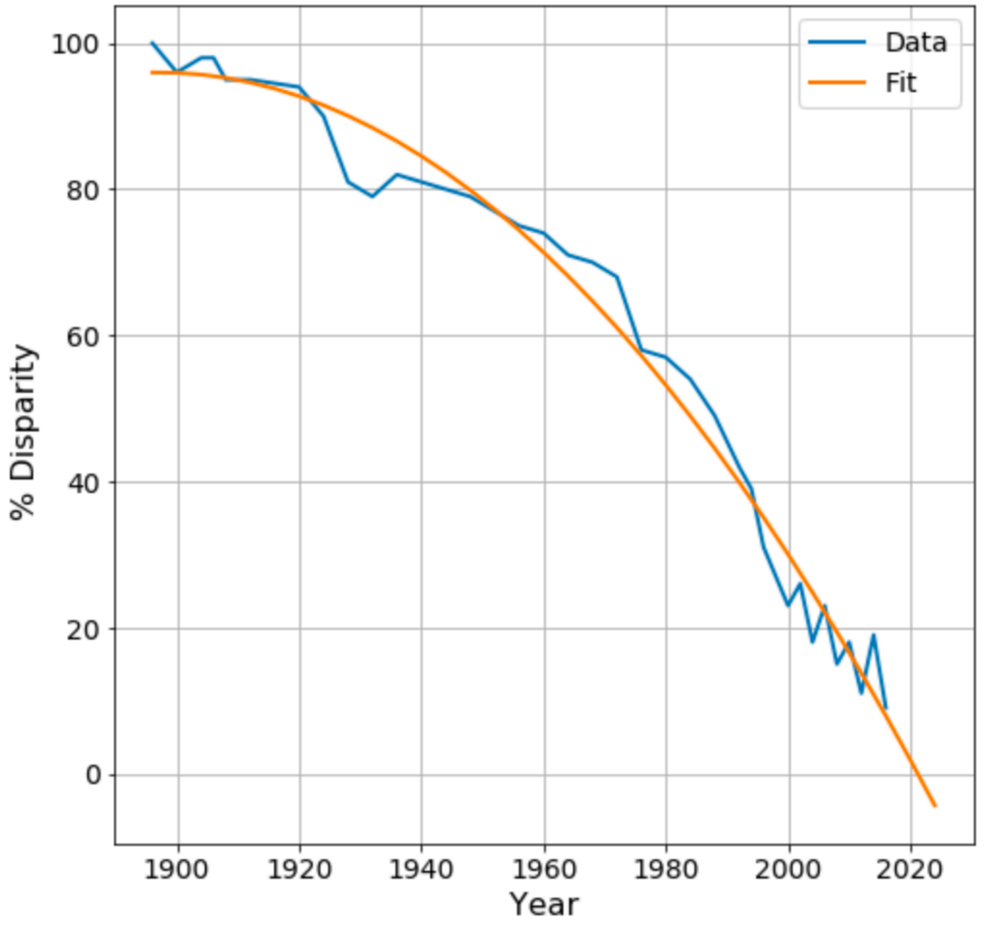
1. Analysis methods vary based on the questions.
2. General Approach includes -
    1. Looking at new features
    2. Creating new metrics
    3. Exploring the Distributions and Relationships via mean, median, variations and correlations
    4. Prediction using Least square fitting

Analysis Method

- **Gender Disparity -**
  - M - Number of Male Players, F - Number of Female Players
  - Metric - We look at “Disparity (%)” Metric=100 x (M - N)/(M+N) for each region and for each Olympic game.
  - Relationship Analysis - Pearson Correlation coefficients to understand the relationships and least square fitting to get the dependence and make future predictions.
- **Sports and Player’s Age -**
  - Metric - We look at Average age of players for each Sport category
  - Analysis tools - Mean, median and Standard Deviation.
- **Sport Domination -**
  - Metric 1 - Total Events won by each country
  - Metric 2 - Number of times each country was listed in top ten places of events tally for each game.

Results - Gender Disparity Over Time

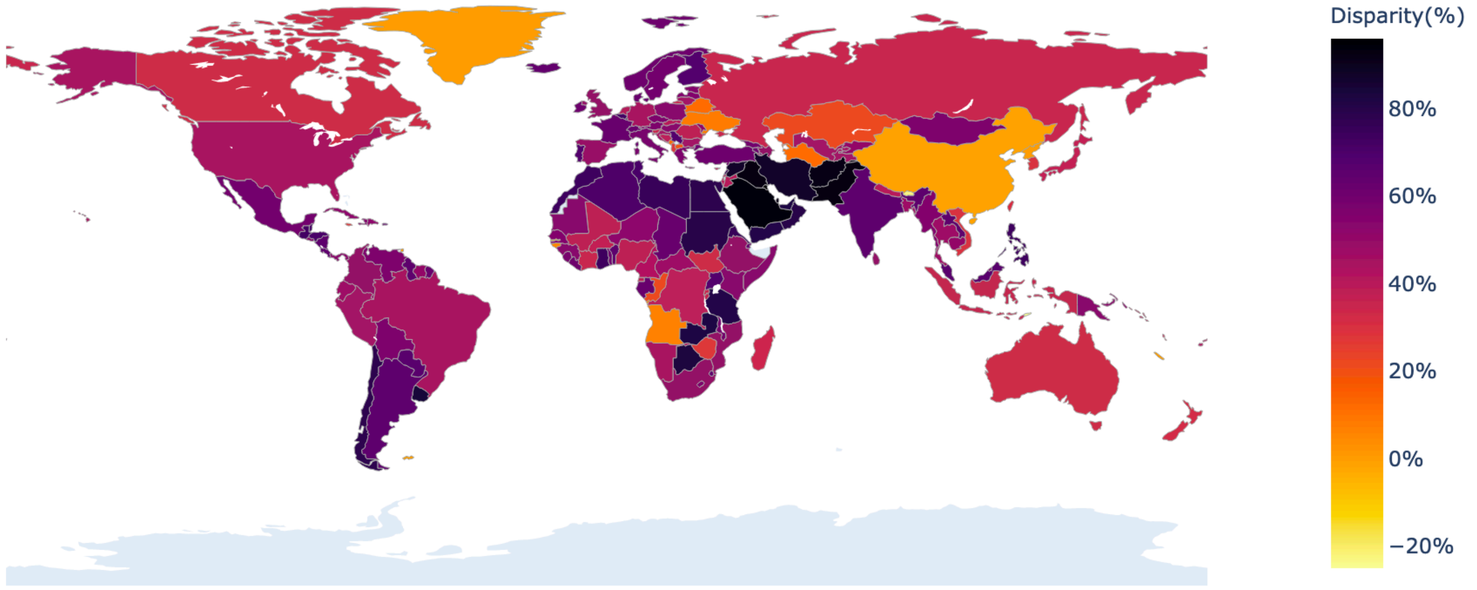
1. Figure - Disparity (%) as a function of time
2. Disparity is negatively correlated with time with Pearson correlation coefficient of -0.96
3. A simple extrapolation of data shows gender disparity should decreases significantly by 2020.
4. This plot considers players from all the regions and hence misses out on regional disparity. Let's consider that now.



RESULTS - REGIONAL GENDER DISPARITY

YEARS : 1896-2016

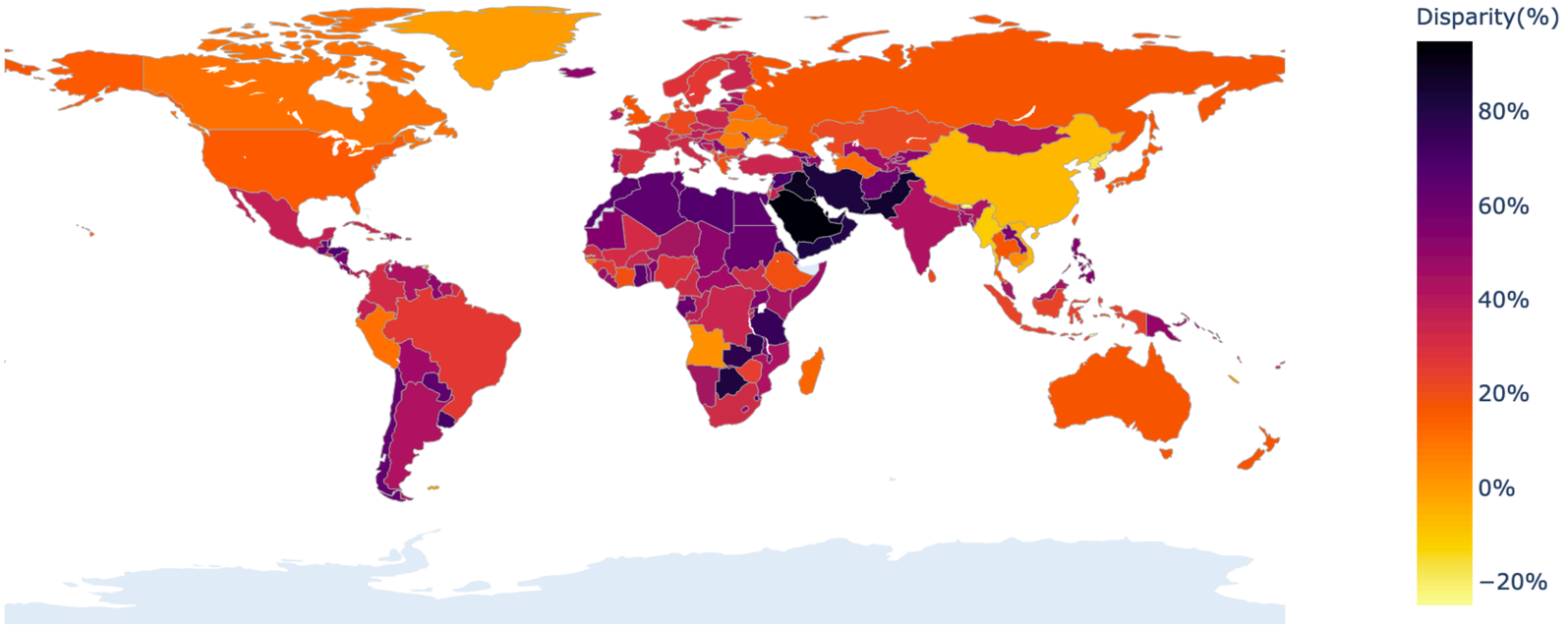
1. Disparity in almost all the countries >30%
2. Disparity (%)
    1. Developed nations - 30-70%
    2. Developing Nations - >50% - Includes many Asian, African and South American countries
    3. Gulf and Islamic Nations - >90%
    4. Negative Disparity - Bhutan, China



Results - Regional Gender Disparity

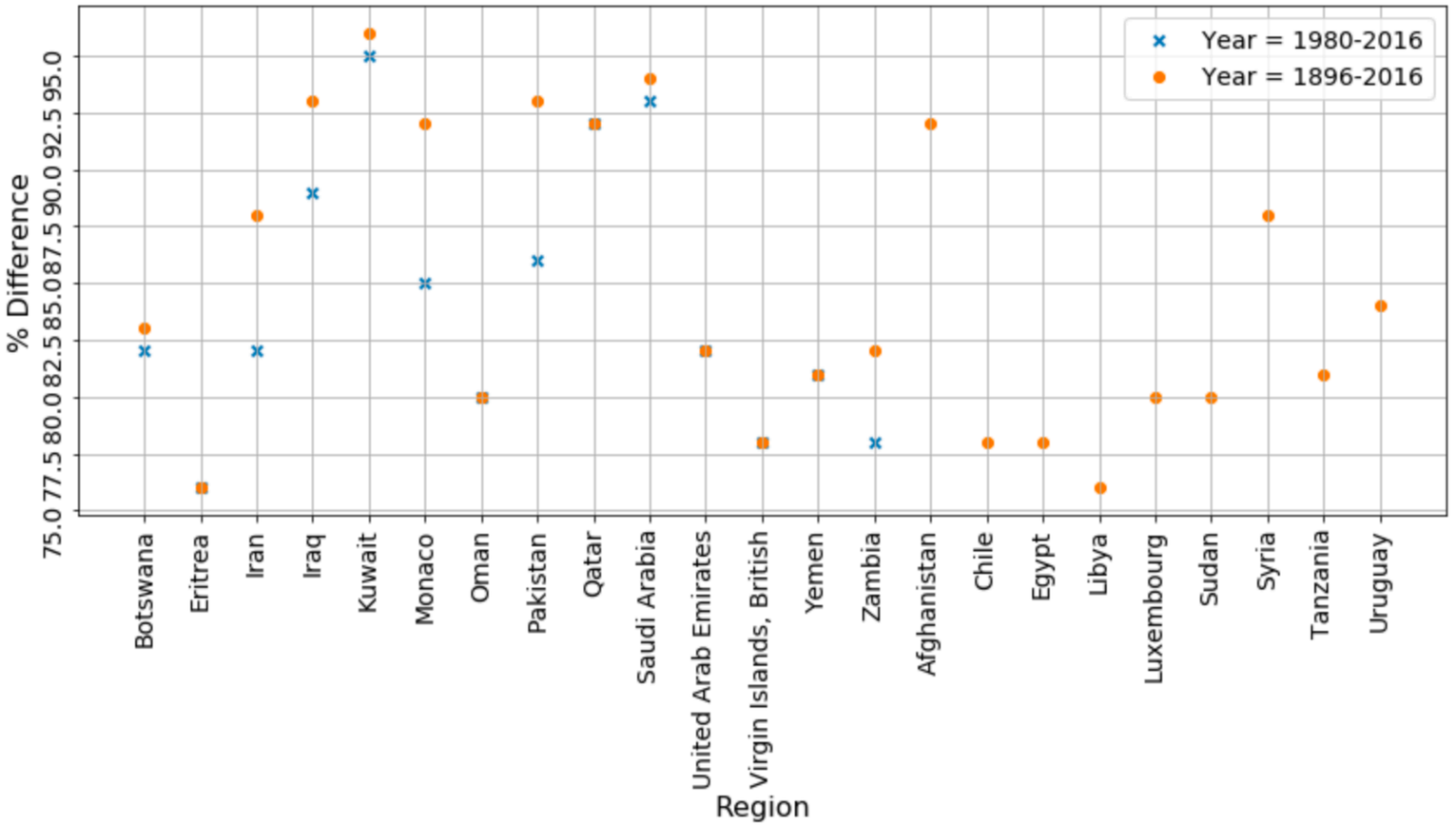
Years : 1980-2016

1. Disparity significantly reduced in recent years consistent with previous slide.
2. Disparity (%)
    1. Developed nations <40%. - **Reduced**
    2. Most Developing Nations - <60% - Includes many Asian, African and South American countries **- Reduced**
    3. Gulf and Islamic Nations - >85% - **Still quite high.**
    4. Negative Disparity - Bhutan, China, Myanmar, Vietnam **- Improved**

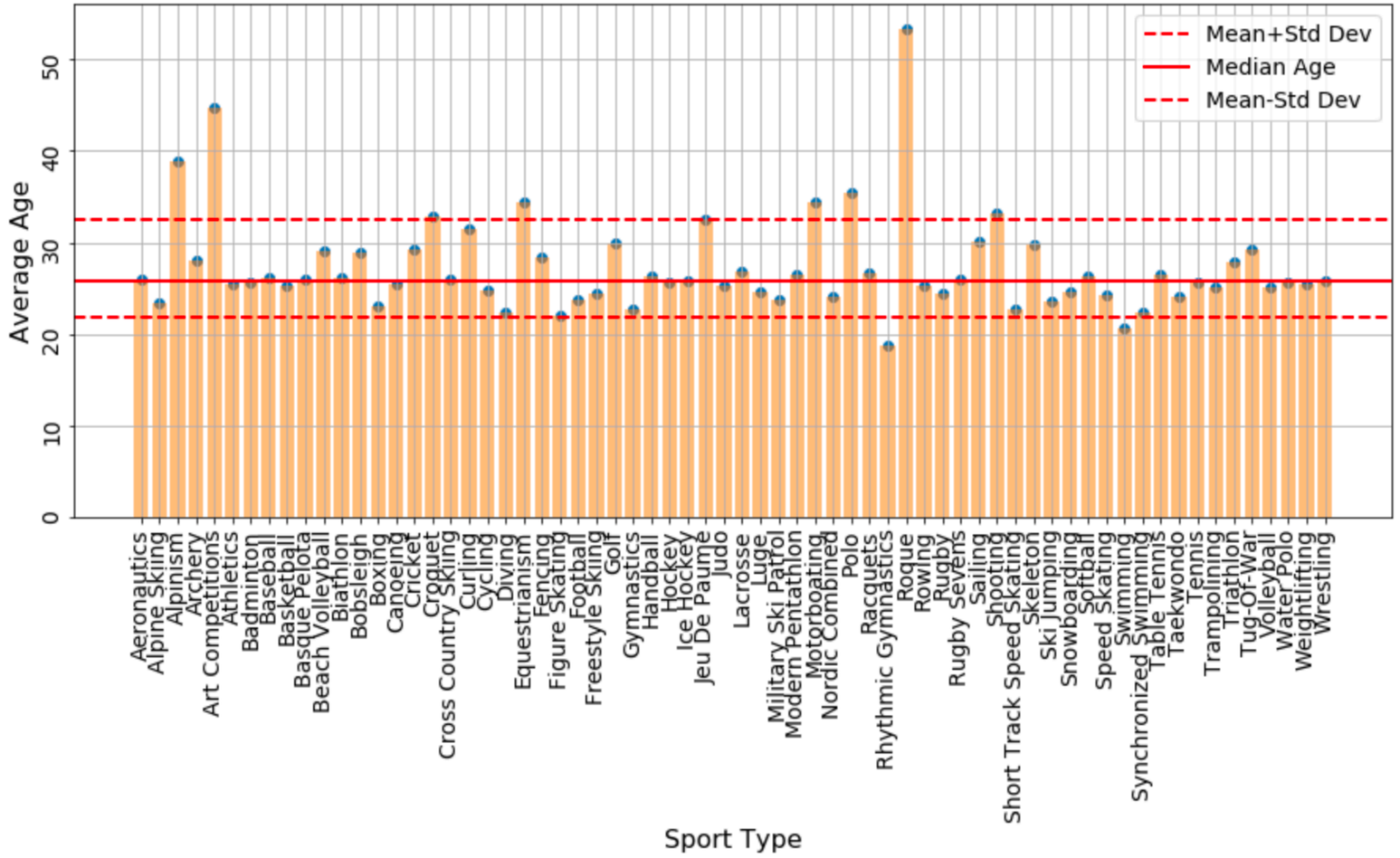


RESULTS - REGIONAL GENDER DISPARITY

1. Plot - Countries with disparity >75%.
2. Number have improved in recent years for many African and South American countries.
3. Disparity still staggeringly high for Islamic and gulf countries with no significant improvement over the years.



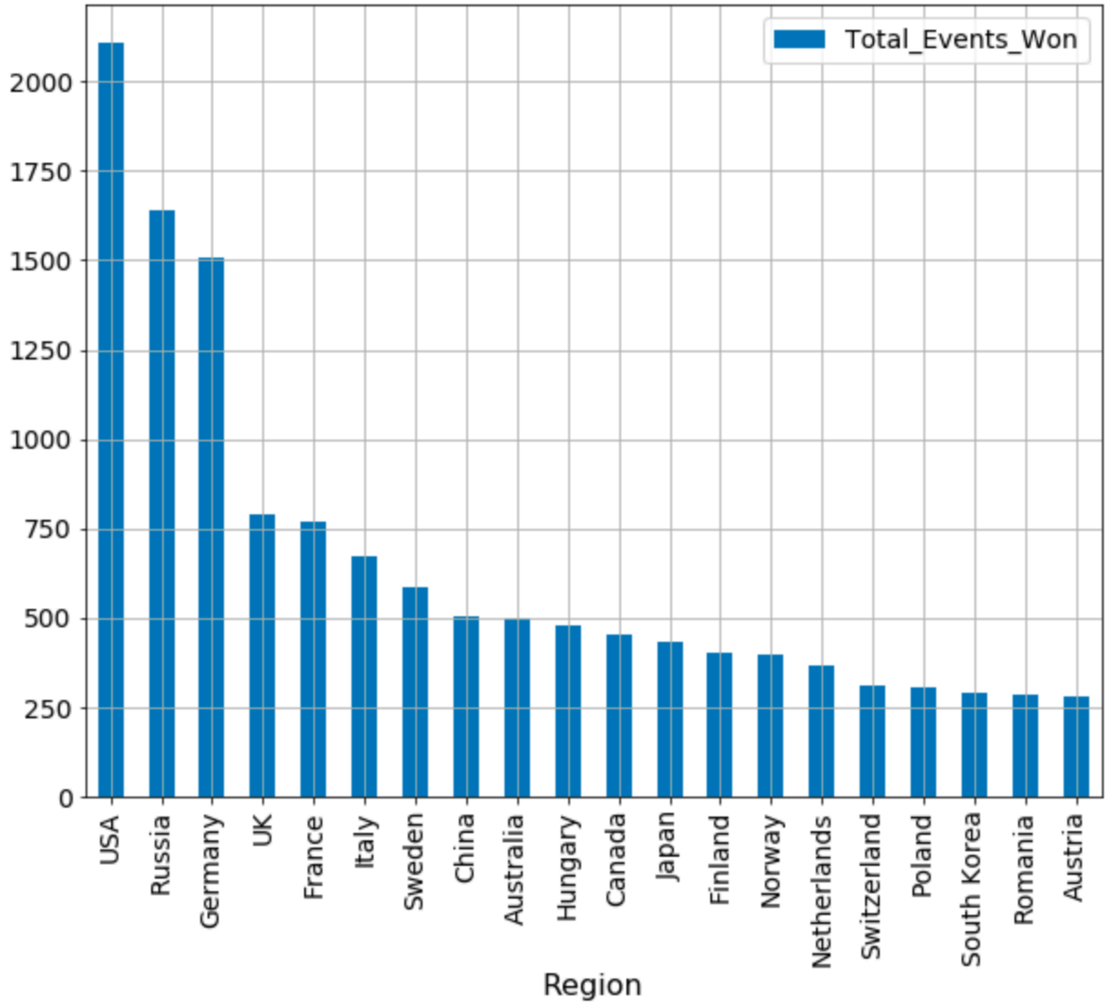
Results - Sports and Optimal Age



- **Plot -** Average age of all players for a sport category vs Sports Category
- **Basic Stats -**
  - Mean - 27.25
  - Median - 25.8
  - Standard Deviation - 5.34

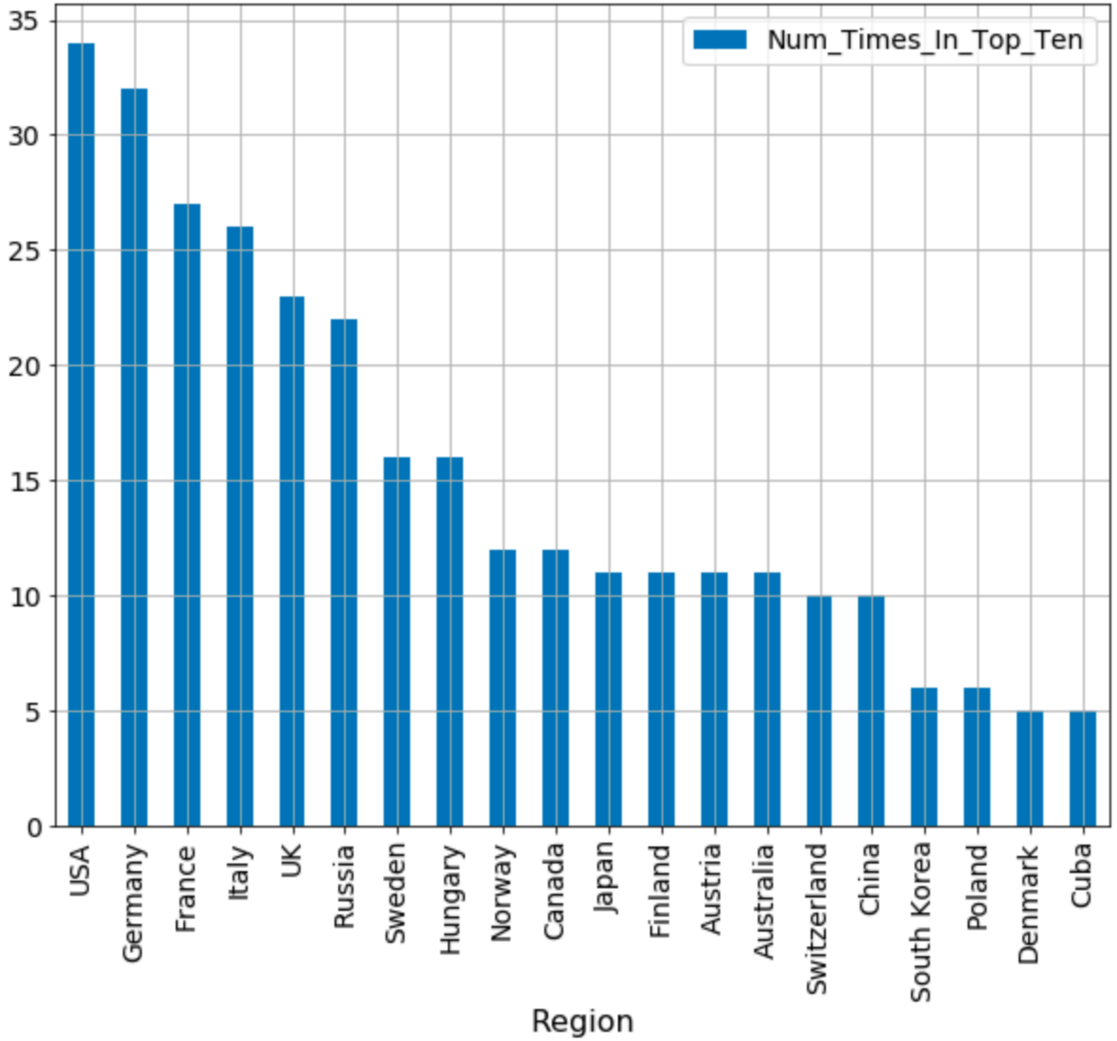
Dominance Of Countries - Metric I

- **Plot -** Top 20 countries with highest number awards since 1896.
- Most nations are developed countries in this list as our initial hypothesis.
- USA, Germany and Russia leading by high margin.
  - USA won more than 34% of total events.
  - Russia won 26.5% of total events.
  - Germany won 24.4% of total events.



DOMINANCE OF COUNTRIES - METRIC II

- **Plot -** Top 20 countries which was listed highest number of times in top ten places of events tally for each game.
- Again, most nations are developed countries with 17 nations same as in previous plot.
- USA, Germany rank highest with appearance in top ten winners in >30 Olympics.
- France and Italy follow with >25 times.
- Italy and UK next with >15 times.
- Digging into data, we find that Russia did not participate in Olympics between 1912-1952 and hence the lagging behind on 6th place.



SUMMARY AND CONCLUSION - QUERY 1

- **Gender Disparity -**
  - To understand gender disparity, we have looked at percent difference between number of male and female players for each game and for each country.
  - Time analysis showed that **gender disparity has decreased significantly over past 100 years**.
  - Though **regional disparity** is quite high for most countries when looked over all the years, things have **improved a lot in last 40 years**.
  - For most **developed nations**, the disparity is down to **<30%** which can be due to more equal opportunities and better laws and policies towards woman empowerment.
  - Some of the **developing nations** are still struggling with disparity >70% in some cases. But situation has improved a lot in many developing and underdeveloped countries with disparity in range of **20-60%**. The high numbers can be caused by lower economic and development index of the country.
  - **Situation is still concerning in many gulf countries and some Islamic nations with disparity >85%**. Note that the reason is not due to the religion of the country but due to the regional laws and policies towards woman. This can be seen by the fact that many other Islamic and Muslim majority countries have significantly lower disparity numbers.
  - To understand the differences in disparities, we need to look at the nation's policies towards people of different genders. This can helpful in identifying policies are cause of gender based inequality. Another possible factor towards high disparity can be the economic factors and another interesting investigation can be to look for the correlations between economic and development index of countries vs gender disparity.

SUMMARY AND CONCLUSION - QUERY 2

- **Sport Category vs Player's Age -**
  - Average age does not vary significantly for most sports and lies in the range of 21 to 31 years as initially predicted.
  - Few exceptions exists - Swimming and Gymnastics players have generally lower ages which suggests agility is more important than experience.
  - Certain other categories like Alpinism, Art Competitions and Roque have extremely high average age falling outside the 1-sigma confidence interval. This suggest higher preference for experience than agility in these games.
  - Instead of looking at all the players, I have also considered only players who won awards, but results do not change significantly.

SUMMARY AND CONCLUSION - QUERY 3

- **Domination of Olympics**
  - To find if Olympics is dominated by few countries, we have looked at two metrics.
    - First metric looks at the total events won by each country over last 100 years.
    - Second metric looks at number of times each country made it to top ten places in events tally for each game.
  - For both these metrics, we find that most countries are developed nations. Out of top 20 countries, 17 emerge to be the same though with difference in order.
  - This clearly suggests that better opportunities, training and facilities are awarded to players from almost all sports categories in these countries. However, sports are still lagging behind in developing countries and require better laws and policies, improved infrastructure and opportunities for their players.
  - With additional data about countries sports policies, sports budgets etc, one can learn more about the common factors in these leading nations which can help in designing new policies or changing current ones for other developing nations.

For Detailed Analysis -

https://github.com/bkhamesra/Olympics_SQL_DataAnalysis/blob/master/Codes/SQL4%20Assignment%202.ipynb


THE END