Women and Men in the Olympics

A Dive into the Data

Contents

1: Questions to Answer

1. What difference in level of participation in the Olympics is there between men and women?
    - Is there a seasonal difference?

2. How did this difference evolve over time?
    - Did relative participations grow or decline over time?

3. What are the physical differences in men and women in equal events?
    - Did any events select for a particular body type?

2: Initial Hypothesis

1. More Men compete than women

2. There will be a proportional increase in women participation over time.
    - Events might not been available to women in early days
    - Events may have began as male centric with growing interest by women

3. What are the physical differences in men and women in equal events?
    - Did any events select for a particular body type?
    - Men will generally be bigger

4. There will be no differences in seasonal relationship participation between the sexes.

3: Data Analysis Approach

1. Focus on modern events followed events with most women participation
    - Isolate events still in the Olympics in last Olympics in data set
    - 2014-2016
    -
2. Raw counts of participation by event and season for modern events
    - Line chart
    - Barchart / Histograms (year bins)

3. Statistical Analysis
    - Relationship between genders & participation
    - Deviations in relational participation
        - Pearlman
        - Spearman
        - Levene

Technical Challenges

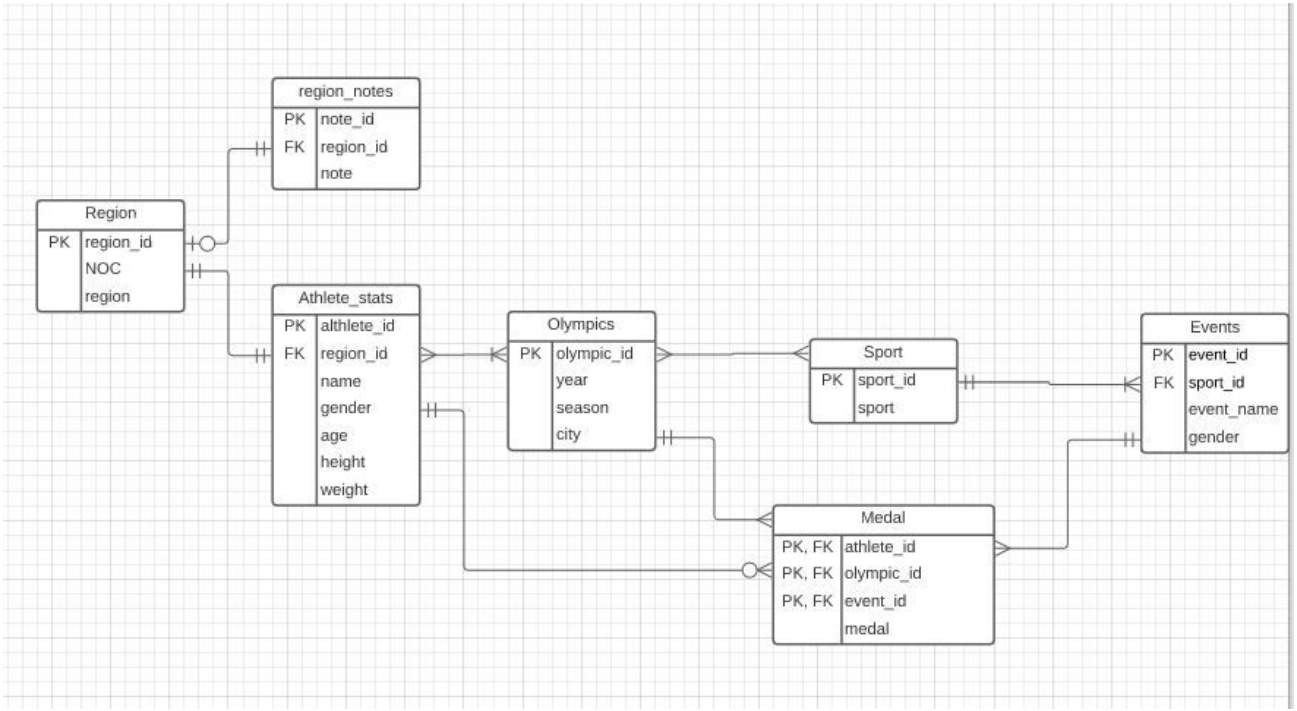
1. Limited amount of data restricting broader analysis
    - Missing data in series

2. Gender embedded in event name
    - Needed text cleaning for comparing between events

3. pandasql did present some challenges
    - More complex SQL expressions did not function properly
        - SQL not fully supported by pandas
        - Pandas is more efficient at transforming data than SQL
    - Used pandas to manipulate data in place of SQL

4. Jupyter notebook limitations
    - Mass graphing blocked or broke down
    - Loops used to bypass

Entity Relationship Diagram



Initial Findings

Findings generally supported hypotheses but not completely:

✔ More men than women participate in the olympics

✔ Women participation grew over time but
    X Not necessarily faster than men's participation.

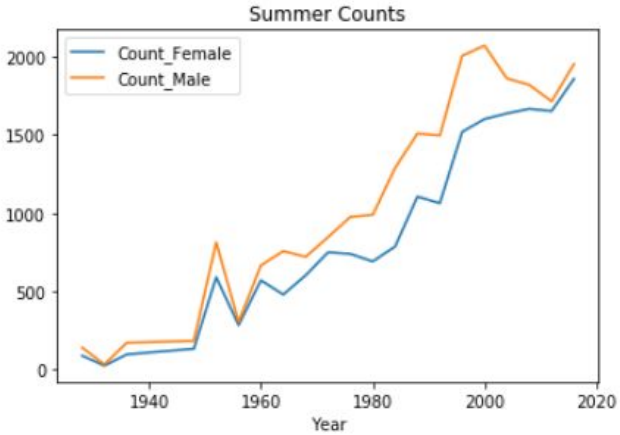
✔ Men were generally bigger than women

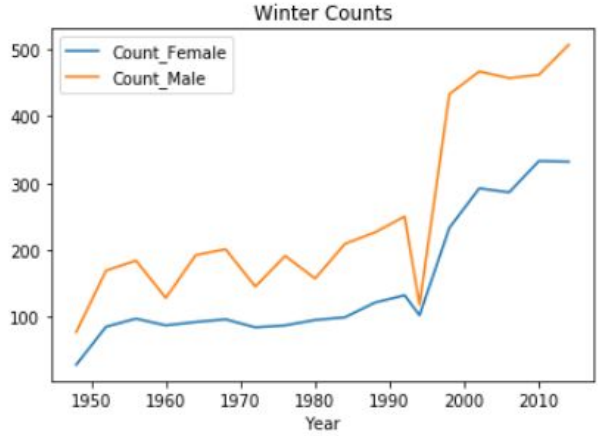
Seasonal participation:
    ✔ No real difference in summer participation
    X Clear difference in winter participation

Initial Findings Data

Gender	Count
Female	74522
Male	196594

Gender	Avg Height	Avg Weight
Female	167.8	60.0
Male	178.9	75.7

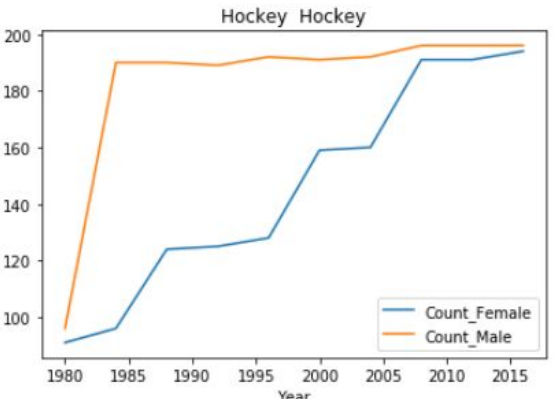


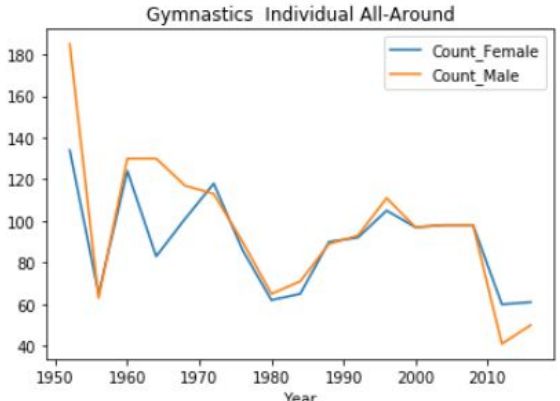


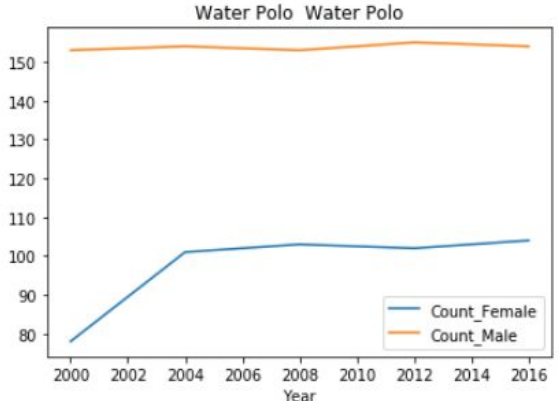
Deeper Analysis: Totals

Line charts were examined for modern events for the genders broken down by year

- Generally speaking, women participation was regularly lower but sometimes was only slightly lower and did cross
- Events had early strong participation by women
  - Gymnastics
  - Swimming
- Some didn't change much
  - Water Polo



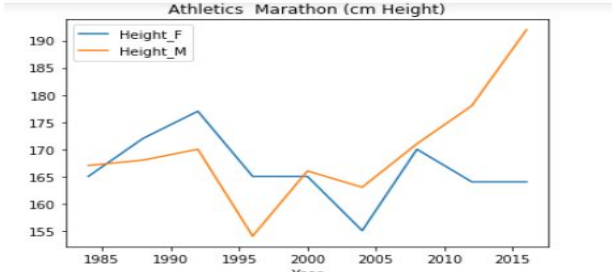


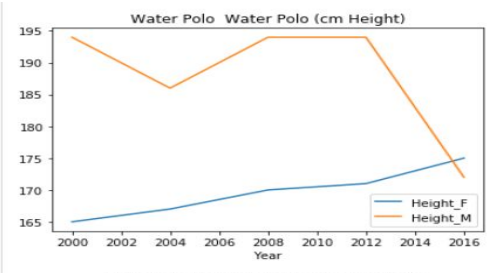


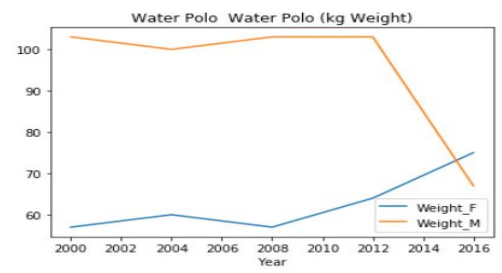
Deeper Analysis: Size (Height & Weight)

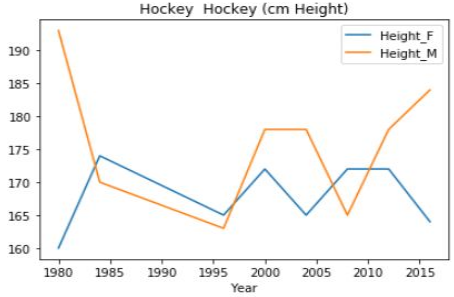
Line charts were examined for modern events for the genders broken down by year

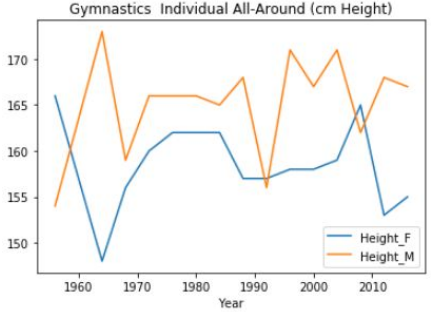
- Generally Men were bigger and heavier with exceptions and interesting notes
  - Some events showed no advantage to size such as handball & alpine skiing slalom
  - Athletics Marathon: Men sizes exploded while women's decreased marginally
  - Water polo converged by both height and weight
    - Hockey showed a similar convergence for height but women weighing less.
    - Several events had this trait

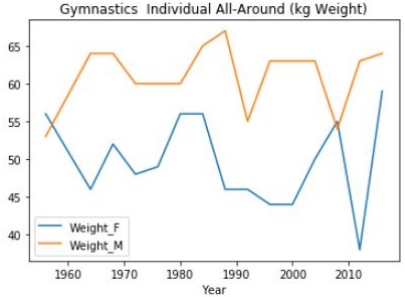












Deeper Analysis: Pearson and Spearman

Pearson correlation will give relative linearity to each other (men and women)

Season	Coef	p-value
Summer	0.98	6.3 e-14
Winter	0.98	5.9 e-12

Pearson correlation revealed no difference

Spearman correlation will show rank correlation and reveal breaks in ordering from year to year between women and men participation

Season	Coef	p-value
Summer	0.95	2.5 e-11
Winter	0.86	5.0 e-6

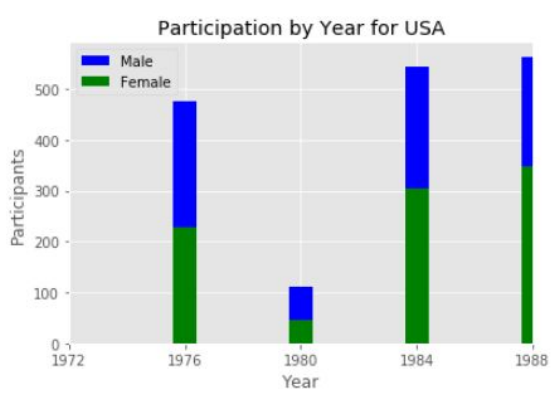
Spearman, although not significant, does show a strong trend for there being a difference. Breaking it down further may find variances. For that, Levene test can be used.

Deeper Analysis: Bar chart overlays with Levene

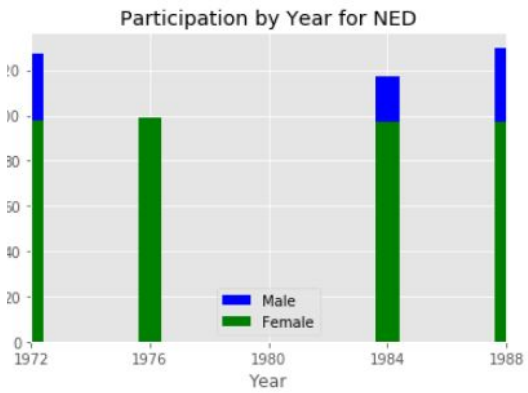
Since summer participation grew where winter participation appeared flatlined, overlaid bar charts were examined for modern events for the genders broken down by year with a focus on 1970-1990 to examine lull observed in winter olympic chart compared to the summer olympics.

- Focus was on countries with >100 athletes as the biggest influencers
- Levene was included to show significance between the male and female
    - $p < 0.05$ is considered different
        - differences in men and women participation has changed for some reason
        - NED for example approached $p<0.05$ with 0.070 showed a noticeable increase in men participation over women
    - $p > 0.05$ is generally considered the same
        - Differences in men and women participation cannot be separated as the number increases in size
        - USA for example with a $p=0.74$ has similar participation characteristics between the population
- Countries had different geopolitical influences for participation from on year to another
- Overall, the effect appears to be primarily from an increase in men participation
        - NED: 1976 had more women than men to grow by 1984 with lower p value than others evaluated.
        - JPN: visually had a significant growth in women compared to men but the Levene p-value was insiginficant at $p = 0.58$

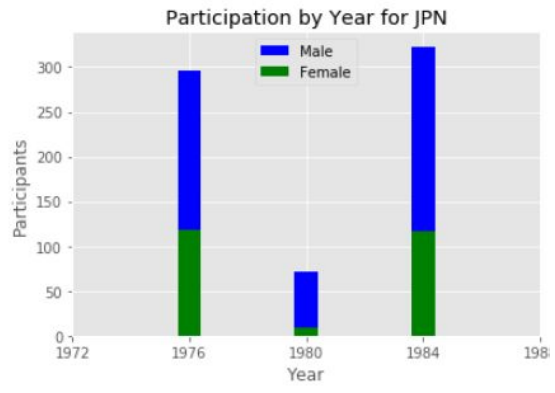
Deeper Analysis: Bar chart overlays with Levene



Levene: p = 0.7367406933921956



ne: p = 0.06997859974723263



Levene: p = 0.5831533044079136

Deeper Analysis:

Backing up and going Deeper with Levene

A closer look at all countries with a p-value <=0.05 revealed 4 countries

Country	p-value
BAR (Barbados)	0.038
GRE (Greece)	0.014
MGL (Mongolia)	0.0035
TUR (Turkey)	0.021

Deeper Analysis:

Backing up and going Deeper with Levene

- Date range were given for 199-2016, pre-1970, 1970-1990, and post 1990
- Instead of by country, the date ranges were used with all populations

Date Range	Levene test p-value
1900-2016	0.015
< 1970	0.0024
1972-1988	0.61
>1990	0.18

- The larger date range showed a significant variance but when broken down the largest contribution was prior to 1970 and to a lesser degree post 1990.

Deeper Analysis:

Backing up and going Deeper with Levene

Back to 1970-1990s,
A closer look at all countries with a p-value <=0.05 revealed 4 countries

Country	p-value	Female count	Male count
BAR (Barbados)	0.038	16	72
GRE (Greece)	0.014	19	240
MGL (Mongolia)	0.0035	26	167
TUR (Turkey)	0.021	13	227

Final Findings

Expanding on the previous hypothesis and exploring the hypotheses that were not confirmed.

- What was the reason for differences in growth?
  - It was found that it was largely an increase or decrease in male participation that drove the variance.

- What was the difference in winter participation?
  - It correlates with the same reason as above.

When the dates are checked and references, big changes happened with geo-political events such as Mongolia and China and political reforms. Boycotts were a big factor too.

The 1970-1990 range was revealing with Mongolia, Turkey and Greece having a bias against women participation which held the numbers low where variation in men's participation in other countries with significant women involvement.

Evaluations such as this for current periods can give meaning to social movements that need work and where they need to focus.

References:

Data sets used:
https://www.dropbox.com/sh/0wqw8fmiwrzr8ef/AABQijjQM522INXX1FCdamzma?dl=0

Primary work done in python with pandas libraries