SI 618 PROJECT 2 REPORT
Analysis of medal wins in Olympics

Hareeshwar Karthikeyan
hkarthik@umich.edu
October 27, 2021

Motivation:

Olympics have a long running history of being the most coveted global sporting event. Only the most elite class of athletes qualify to participate in Olympics and the medal winners are undisputedly the best in the world every year. The medal winning status of the athletes are influence by a variety of factors based on their background.

The overall goal of this project is to look into how the Olympic medals wins of a country are affected by the physical characteristics of their athletes and by their previous Olympic participation.

Main Research Questions:

1. Does the medal winning status of athletes depend on their physical features?
2. Does previously participating or winning a medal in an Olympic event increase the chances of winning a medal in the current Olympic season?
3. Does having more participants mean winning more medals for a country?

Dataset:

The 120 years of Olympic History: Athletes and results dataset from Kaggle is used for this purpose. It contains the bio data of athletes and their medal results for all the events in the modern Olympic games from Athens 1896 to Rio 2016 Olympics.
The file *athlete_events.csv* contains 271116 rows and 15 columns. Each row corresponds to an individual athlete competing in an Olympic event (athlete-events) with these columns:

1. ID - Unique number for each athlete
2. Name - Athlete's name
3. Sex - M or F
4. Age - Integer
5. Height - In centimeters
6. Weight - In kilograms
7. Team - Team name
8. NOC - National Olympic Committee 3-letter code *(Taken as the country for analysis)*
9. Games - Year and season
10. Year - Integer
11. Season - Summer or Winter
12. City - Host city
13. Sport - Sport
14. Event - Event
15. Medal - Gold, Silver, Bronze, or NA

It allows for analyses to be done on the medal history of athletes from various countries and various physical backgrounds(by age, height, weight and sex).

Data Manipulation Methods:

Selected data for analysis:

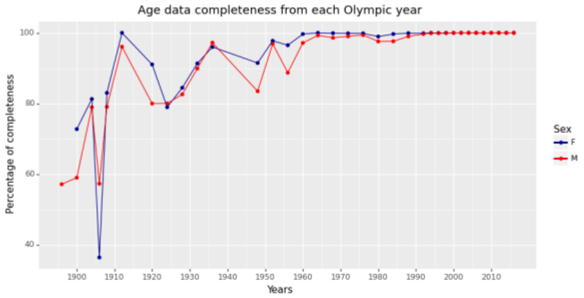
In the dataset, the *'athlete_events.csv'* was the data file considered for this project. The other *'noc_regions.csv'* was not of interest for any of the analyses considered in this project.

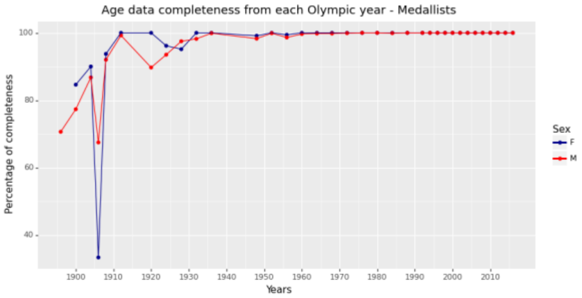
Handling missing, incomplete and noisy data:

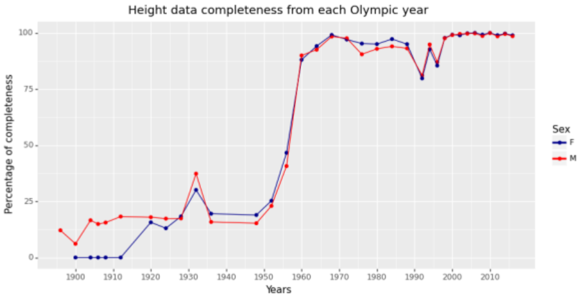
Data Completeness for the missing variables - Age, Height and Weight

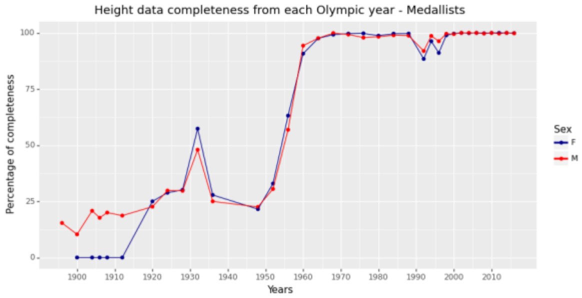
Overall

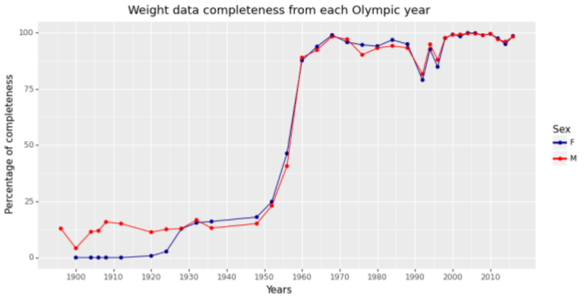
Medallists

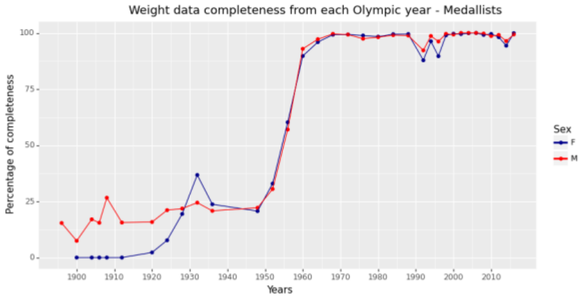












Missing values were mostly before 1960. After 1960, most of the values were present and especially for medallist, the missing data was very low.

- <u>Missing values for medals</u>
  The values for athletes that did not win any medals was not available. This was replaced by 'None'
- <u>Missing values for age, height and weight</u>
  The missing values for an athlete's age/height/weight entry was replaced by the average of the age/height/weight value of the athletes participating in the same event in the same year and season.
- <u>Missing values for ratio of medallists in final task</u>

This was a custom value calculated for the final task. The ratio of first timers/previous medallists or non medallists who won a medal in the current season was calculated and the missing values meant no athletes in that category. These were replaced by zero.

Challenges encountered:

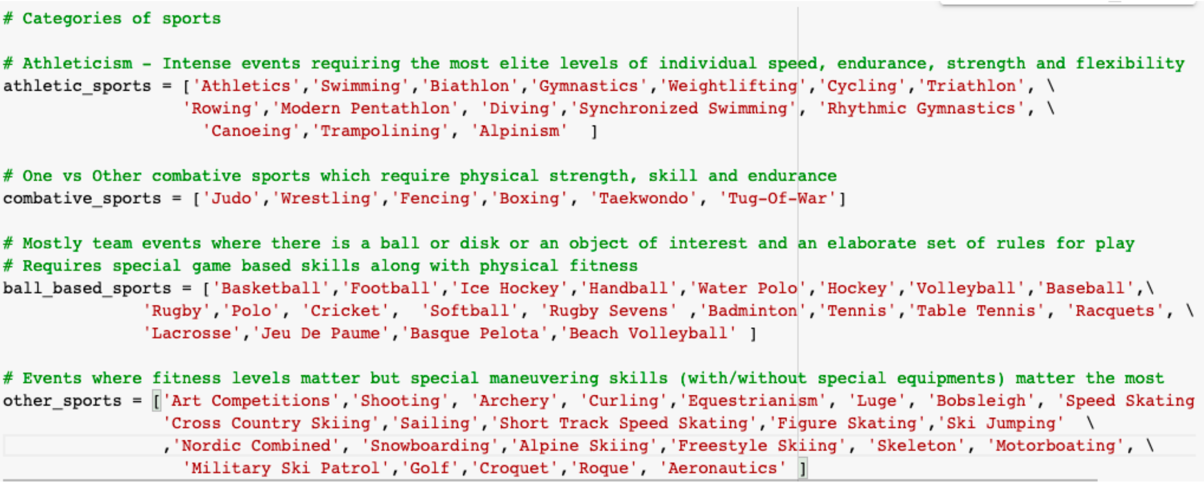
- <u>Incomplete data in all instances of an athlete</u>
  Initially, when considering filling the missing values for an athlete, all the other data rows for the same athlete ID were considered. The data was uniformly missing as the missing value like age was also missing in all the future records of the athlete corresponding to other Olympic seasons. Hence, I had to settle for an approach as mentioned above in the handling missing data section of the report.
- <u>Switching to colab for calculating the values</u>
  Calculating the averages to fill the missing data took a long amount of time and especially for the task 2, the calculation of new features that involved running through the records of the table to find if the athlete has previously participated or won a model took an absurdly long amount of time. I switched from my local machine to Google Colaboratory and there was some improvement in speed.
- <u>Categorizing the sports</u>
  Different sports have different physical demands but there was no specific particular way to measure this and categorize the sports. So I had to manually categorize them into four categories – athletic, ball-based, combative and other.

Analysis and Results:

Analysis 1:

Does the medal winning status of athletes depend on their physical features?

Firstly, depending on the style of sport, the sports were categorized as follows:



```
# Categories of sports

# Athleticism - Intense events requiring the most elite levels of individual speed, endurance, strength and flexibility
athletic_sports = ['Athletics','Swimming','Biathlon','Gymnastics','Weightlifting','Cycling','Triathlon', \
                  'Rowing','Modern Pentathlon', 'Diving','Synchronized Swimming', 'Rhythmic Gymnastics', \
                    'Canoeing','Trampolining', 'Alpinism'  ]

# One vs Other combative sports which require physical strength, skill and endurance
combative_sports = ['Judo','Wrestling','Fencing','Boxing', 'Taekwondo', 'Tug-Of-War']

# Mostly team events where there is a ball or disk or an object of interest and an elaborate set of rules for play
# Requires special game based skills along with physical fitness
ball_based_sports = ['Basketball','Football','Ice Hockey','Handball','Water Polo','Hockey','Volleyball','Baseball',\
              'Rugby','Polo', 'Cricket',  'Softball', 'Rugby Sevens' ,'Badminton','Tennis','Table Tennis', 'Racquets', \
              'Lacrosse','Jeu De Paume','Basque Pelota','Beach Volleyball' ]

# Events where fitness levels matter but special maneuvering skills (with/without special equipments) matter the most
other_sports = ['Art Competitions','Shooting', 'Archery', 'Curling','Equestrianism', 'Luge', 'Bobsleigh', 'Speed Skating
                'Cross Country Skiing','Sailing','Short Track Speed Skating','Figure Skating','Ski Jumping'  \
                ,'Nordic Combined', 'Snowboarding','Alpine Skiing','Freestyle Skiing', 'Skeleton', 'Motorboating', \
                  'Military Ski Patrol','Golf','Croquet','Roque', 'Aeronautics' ]
```

For a measure of fitness/physical size, BMI was calculated using the following formula:

$$\text{BMI} = \frac{\text{weight (kg)}}{\text{height (m}^2\text{)}}$$

BMI was then categorized into the five standard categories[1] as follows

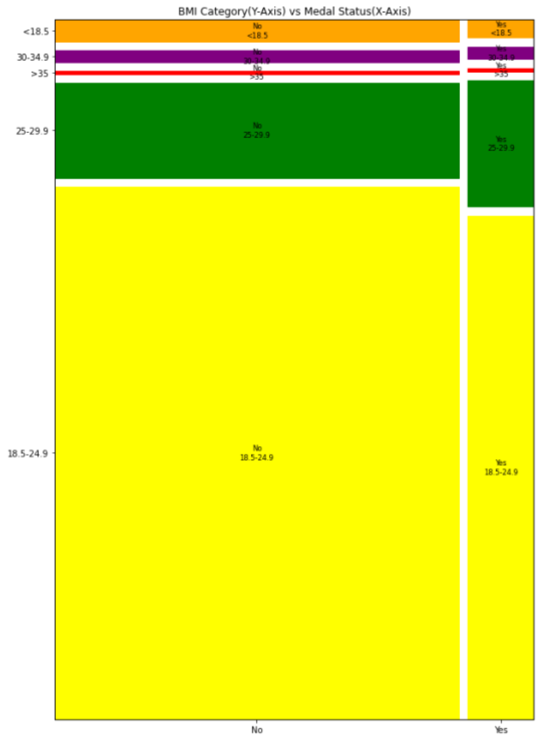
- <18.5, 18.5-24.9, 25-29.9, 30-34.9, >35

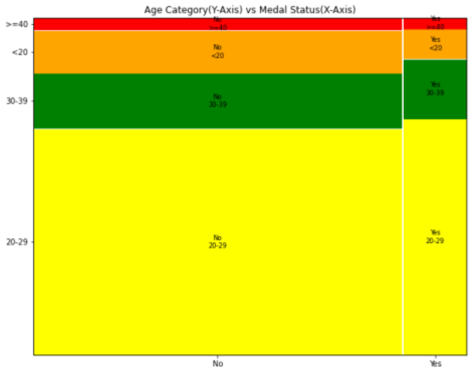
Then, the following plots/tests were done.

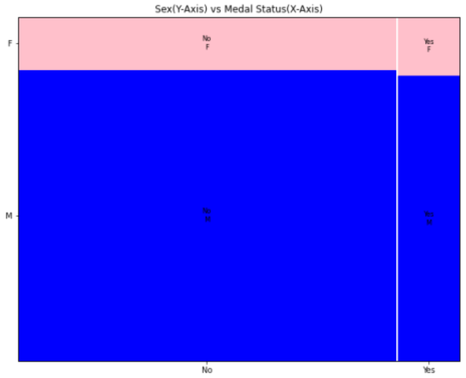
Mosaic Plots

Medal Status on X axis

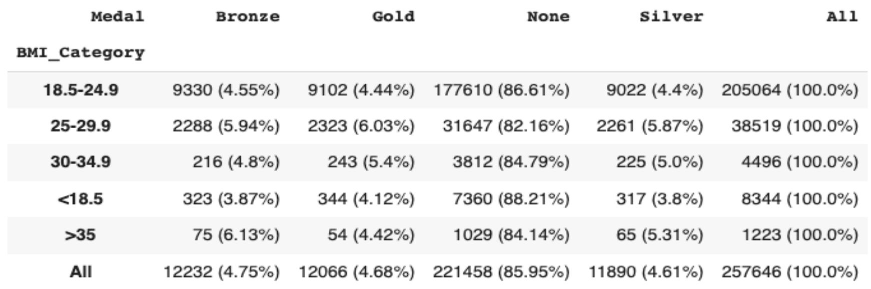
- Yes – Won Gold/Silver/Bronze medal
- No – Did not win any medals



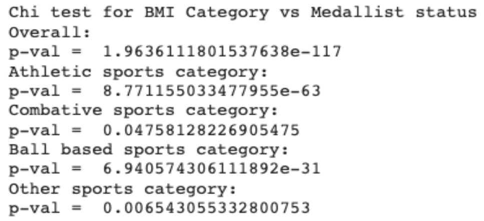




BMI Categories vs Medal Wins

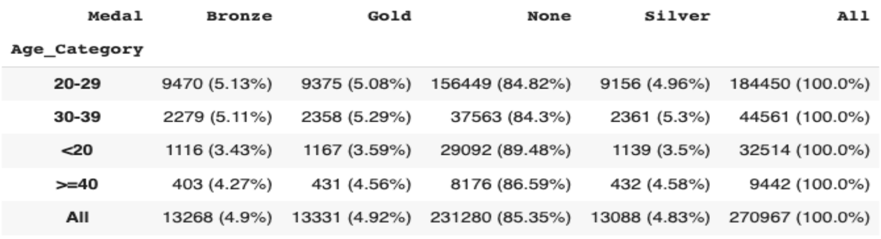


Medal BMI_Category	Bronze	Gold	None	Silver	All
18.5-24.9	9330 (4.55%)	9102 (4.44%)	177610 (86.61%)	9022 (4.4%)	205064 (100.0%)
25-29.9	2288 (5.94%)	2323 (6.03%)	31647 (82.16%)	2261 (5.87%)	38519 (100.0%)
30-34.9	216 (4.8%)	243 (5.4%)	3812 (84.79%)	225 (5.0%)	4496 (100.0%)
<18.5	323 (3.87%)	344 (4.12%)	7360 (88.21%)	317 (3.8%)	8344 (100.0%)
>35	75 (6.13%)	54 (4.42%)	1029 (84.14%)	65 (5.31%)	1223 (100.0%)
All	12232 (4.75%)	12066 (4.68%)	221458 (85.95%)	11890 (4.61%)	257646 (100.0%)

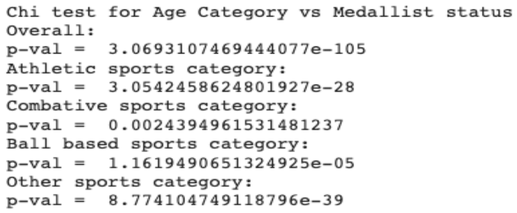


```
Chi test for BMI Category vs Medallist status
Overall:
p-val =  1.9636111801537638e-117
Athletic sports category:
p-val =  8.771155033477955e-63
Combative sports category:
p-val =  0.047581282269054 75
Ball based sports category:
p-val =  6.940574306111892e-31
Other sports category:
p-val =  0.006543055332800753
```

Age Categories vs Medal Wins



Medal Age_Category	Bronze	Gold	None	Silver	All
20-29	9470 (5.13%)	9375 (5.08%)	156449 (84.82%)	9156 (4.96%)	184450 (100.0%)
30-39	2279 (5.11%)	2358 (5.29%)	37563 (84.3%)	2361 (5.3%)	44561 (100.0%)
<20	1116 (3.43%)	1167 (3.59%)	29092 (89.48%)	1139 (3.5%)	32514 (100.0%)
>=40	403 (4.27%)	431 (4.56%)	8176 (86.59%)	432 (4.58%)	9442 (100.0%)
All	13268 (4.9%)	13331 (4.92%)	231280 (85.35%)	13088 (4.83%)	270967 (100.0%)

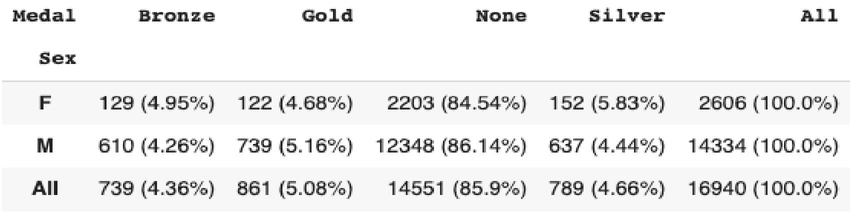


```
Chi test for Age Category vs Medallist status
Overall:
p-val =  3.0693107469444077e-105
Athletic sports category:
p-val =  3.0542458624801927e-28
Combative sports category:
p-val =  0.0024394961531481237
Ball based sports category:
p-val =  1.1619490651324925e-05
Other sports category:
p-val =  8.774104749118796e-39
```

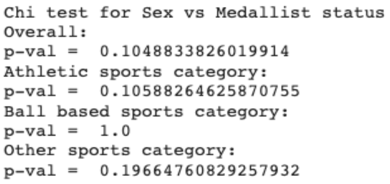
From the Mosaic plots, we can see that between the medallists and the non-medallists, there is a difference in the ratio of the participants from each category of BMI, Age and Sex, although not very pronounced in sex. In the Chi square test for significance, the p-values are very low(less than 0.05) for the Age and BMI Categories and so that indicates that BMI Category and Age Category have a significant relationship with the Medal wins.

Sex vs Medal Wins (For mixed gender events)

Only mixed gender events were considered as only in these events, both men and women would participate and all the other events were gender exclusive and this analysis would not be appropriate.



Medal	Bronze	Gold	None	Silver	All
Sex					
F	129 (4.95%)	122 (4.68%)	2203 (84.54%)	152 (5.83%)	2606 (100.0%)
M	610 (4.26%)	739 (5.16%)	12348 (86.14%)	637 (4.44%)	14334 (100.0%)
All	739 (4.36%)	861 (5.08%)	14551 (85.9%)	789 (4.66%)	16940 (100.0%)



```
Chi test for Sex vs Medallist status
Overall:
p-val =  0.1048833826019914
Athletic sports category:
p-val =  0.10588264625870755
Ball based sports category:
p-val =  1.0
Other sports category:
p-val =  0.19664760829257932
```

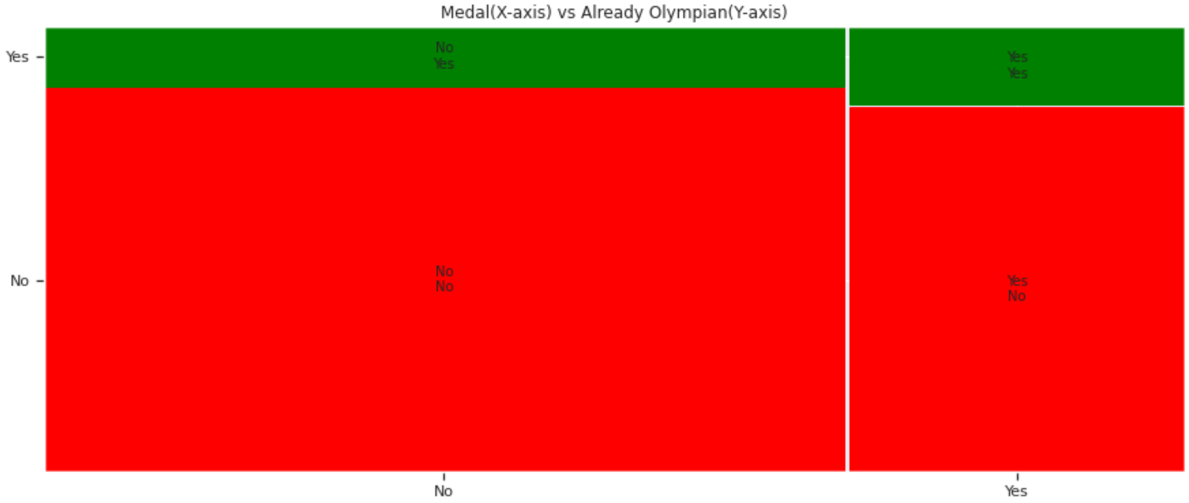
For Sex values however, the p-values are quite large(>0.05) and hence this indicates that the Sex does not play a significant role in the medal wins for the mixed gender events.

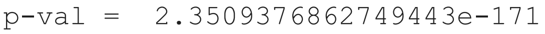
Analysis 2:

Does previously participating or winning a medal in an Olympic event increase the chances of winning a medal in the current Olympic season?

Firstly for each athlete in a certain year in a certain season, the status of whether he has already been an Olympian and whether he has already won a medal is computed. These are stored in the ‘Already Olympian’ and ‘Already Medallist’ columns respectively. Then the following plots/tests were done. The p-val is from the Chi square test for the two categorical variables in the respective plot.

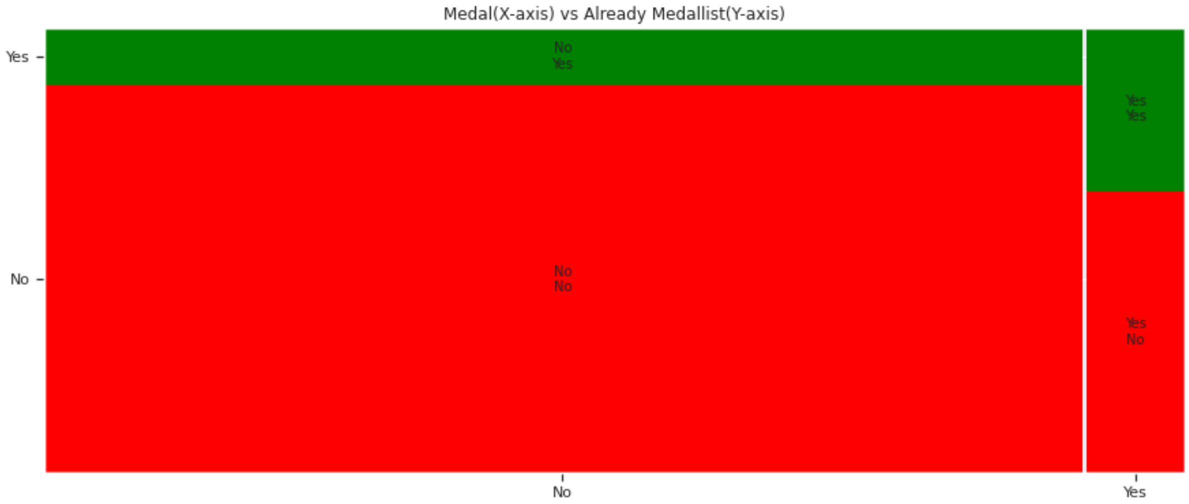
Mosaic Plot for Already Olympian status:





```
p-val =  2.3509376862749443e-171
```

Mosaic Plot for Already Medallist status:



```
p-val =  0.0
```

Both the low p-values and the Mosaic plots indicate that there is a significance in the previous participation and previous medal wins affecting the current medal winning status.

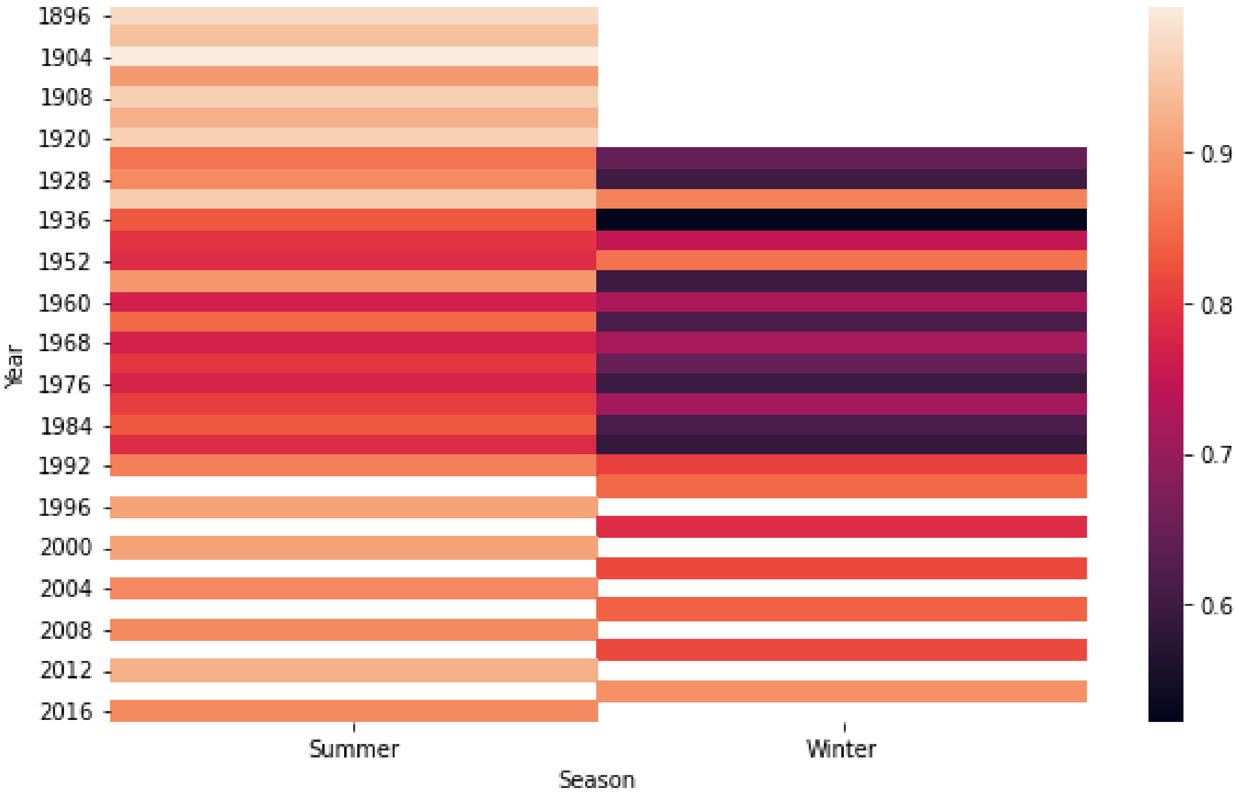
(Other analysis were made as well. But weren't relevant to the research question and are not included in the report owing to length restrictions. Please refer the python notebooks)

Analysis 3:

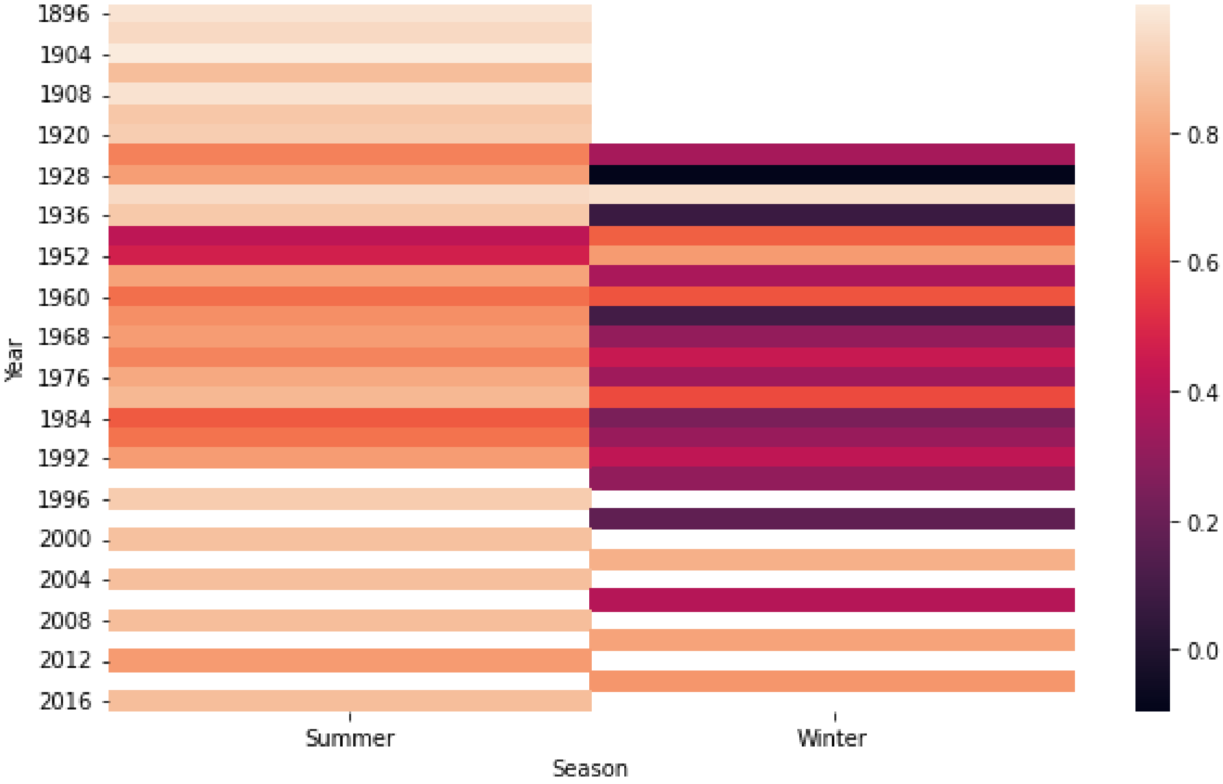
Does having more participants mean winning more medals for a country?

The number of participants and number of medals for each country in each season of the Olympics was calculated. Then two cases were considered – overall data and the data only for the top ten countries bagging the most medals.
For both these cases, the correlation between the overall number of participants and the total number of medals won by the country were calculated. For this, the heatmaps were plotted.

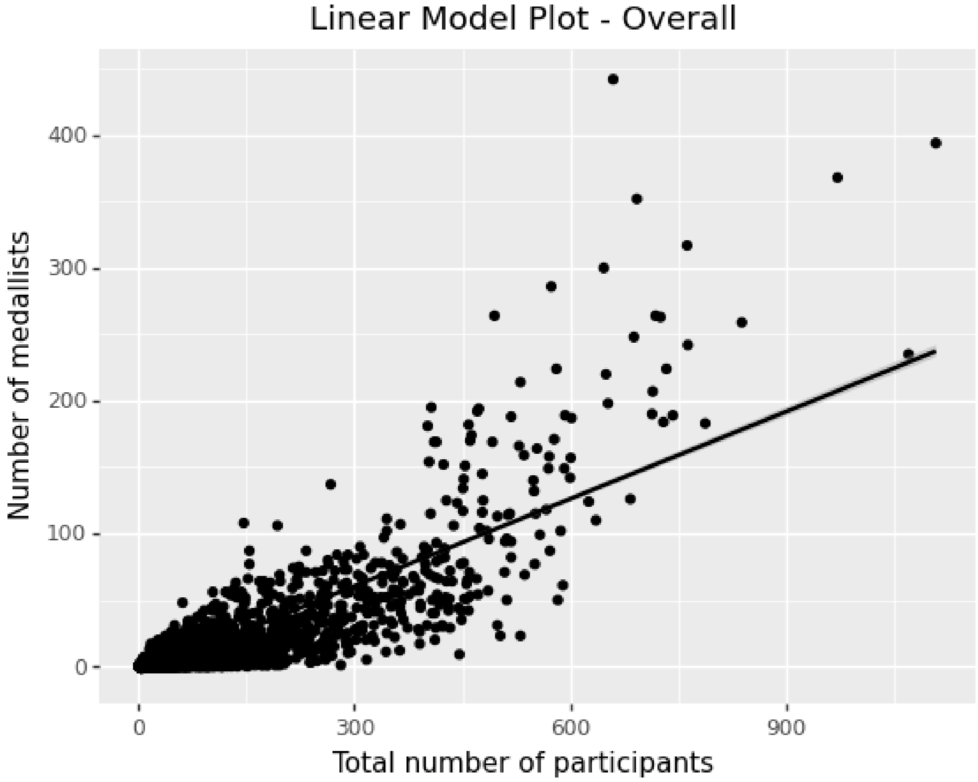
Correlation Heatmap (overall):

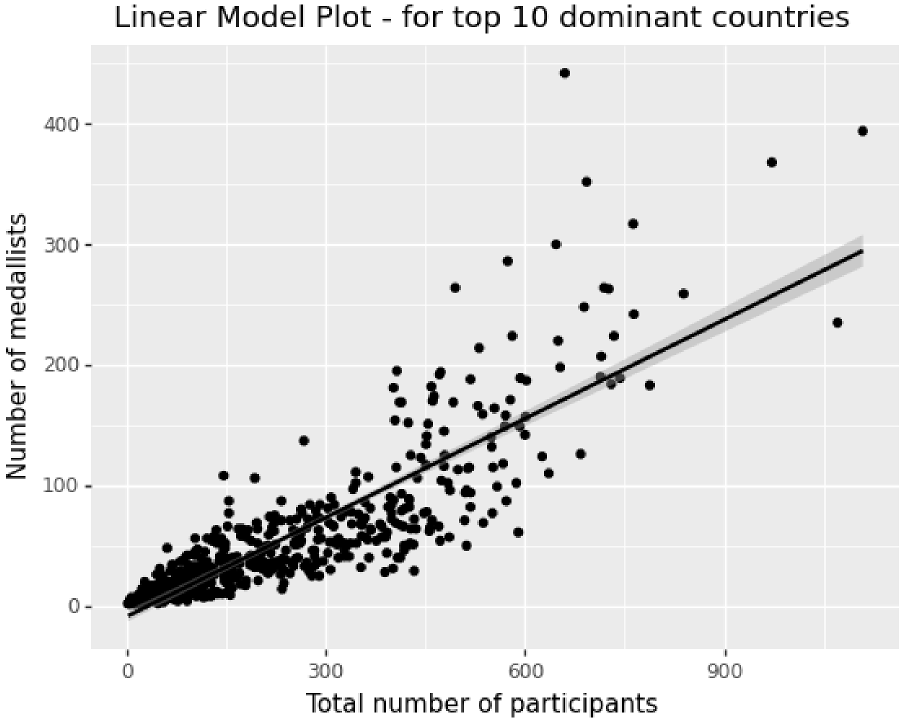


Correlation Heatmap (for the top 10 countries):



Visualising Linear Models:





The correlation is significantly high for both the cases.
The correlation is very high and the linear model also fits better for the case with the top 10 countries.
So in most cases, having high number of participants leads to high number of medals for the country.