SportsStats Olympic Dataset Analysis

Osman Cem YILMAZ

July 2022

Table of Contents

Review of Questions to Answers/Hypotheses/Approach

Questions to Answer

1. Did other major world events have an impact on the Olympic Games?
   a. Did the Olympic Games show us insights about historical events in global society?
   b. Are the Olympic sports events devoid of politics?
2. What age is considered the peak age for athletic performance?
   a. Were the participating athletes from similar age groups? Were the medal winners from different age groups than others?
   b. Did it change based on the sport?
   c. Did it change over time?
3. Did participating with more athletes provide more winners?
   a. Is there a correlation between the number of medal-winner athletes and the number of participating athletes at the national base?

Review of Questions to Answers/Hypotheses/Approach

Initial Hypotheses

1. The Olympic Games were affected by other global social events.
   a. There were economic and political events that affected The Olympic Games.
2. The peak of athletic performance is between the ages of 20 and 30.
   a. The Summer Olympic Games or the Winter Olympic Games occur every 4 years. The athletes who can participate in the games are the best athletes in their nations. Because of that, medal winners and other athletes should be from similar age groups.
   b. Some sports groups require less speed and power. For those sport groups, the athletes' age group differs.
   c. Improvements in medicine, sports science, innovations in sports equipment and training regimes have affected the peak performance age positively.
3. More athletes competing in the Olympic Games meant more medals for the participating countries.

Review of Questions to Answers/Hypotheses/Approach

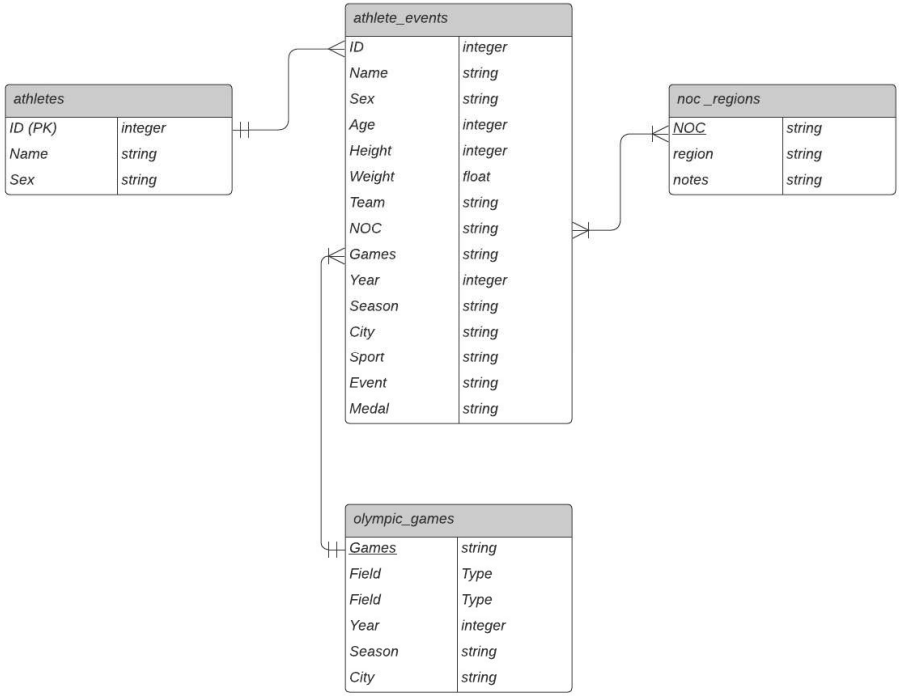
Data Analysis Approach

1. Examine the numbers of athletes over time
   a. Look trends for the numbers of athletes over time
   b. For unusual changes, look for the global events
2. Look for median and mean ages of participating athletes and medal winner athletes
   a. Look for changes according to the sports
   b. Look trends for changes over time
3. Look for the numbers of participating athletes and the numbers of medal winner athletes for each nation

Discuss Technical Challenges

- The duplicated rows
- CSV files for SQL queries, xlsx files for Tableau Public
- OFFSET statement is not supported by Databricks

Detail:Entity Relationship Diagram (ERD)



Initial Findings

- The numbers of participating athletes for seasons differs. The Summer Olympic Games more popular than The Winter Olympics Games

```sql
SELECT Season, COUNT(Year) as Number_of_events, MIN(Number_of_Athletes) AS min_athletes, MAX(Number_of_Athletes) AS max_athletes,
AVG(Number_of_Athletes) AS avg_athletes
FROM
(
  SELECT COUNT(DISTINCT ID) as Number_of_Athletes, Year, Season
  FROM AthleteEventsBronze
  GROUP BY Year, Season
) count_ath
GROUP BY Season
```

(4) Spark Jobs

	Season	Number_of_events	min_athletes	max_athletes	avg_athletes
1	Summer	29	176	11179	5477.862068965517
2	Winter	22	252	2745	1299.6818181818182

Initial Findings

- The median age of athletes between the ages 20 and 30. There are older ages in the Summer Games.

```sql
SELECT Season, median_age
FROM
  (
  SELECT Season, Age as median_age,
  row_number() over (partition by Season order by Age) as row_num
  FROM
    (
      SELECT DISTINCT ID, Name, Sex, Age, Year, Season
      FROM AthleteEventsBronze
      WHERE AGE IS NOT NULL
    ) age
    ) age_row
WHERE (Season="Summer" and row_num = 76171 ) or (Season="Winter" and row_num= 14194)
```

(3) Spark Jobs

	Season	median_age
1	Summer	25
2	Winter	25

```sql
SELECT DISTINCT LAST_VALUE(Age) OVER (Partition by Season, prcnt) as last_val, Season,
(CASE WHEN prcnt=1 THEN '%25' WHEN prcnt=2 THEN '%50' WHEN prcnt=3 THEN '%75' ELSE '%100' END) as pct
FROM
  (
  SELECT Season, Year, Age, ntile(4) OVER(Partition by Season Order by Age) as prcnt
  FROM
    (
    SELECT
    DISTINCT ID,
    Name,
    Sex,
    Age,
    Year,
    Season
    FROM AthleteEventsBronze
    WHERE AGE IS NOT NULL
    ) as percent_temp
  ) as ntile_temp
```

(3) Spark Jobs

	last_val	Season	pct
1	22	Summer	%25
2	25	Summer	%50
3	29	Summer	%75
4	97	Summer	%100
5	22	Winter	%25
6	25	Winter	%50
7	28	Winter	%75
8	58	Winter	%100

Initial Findings

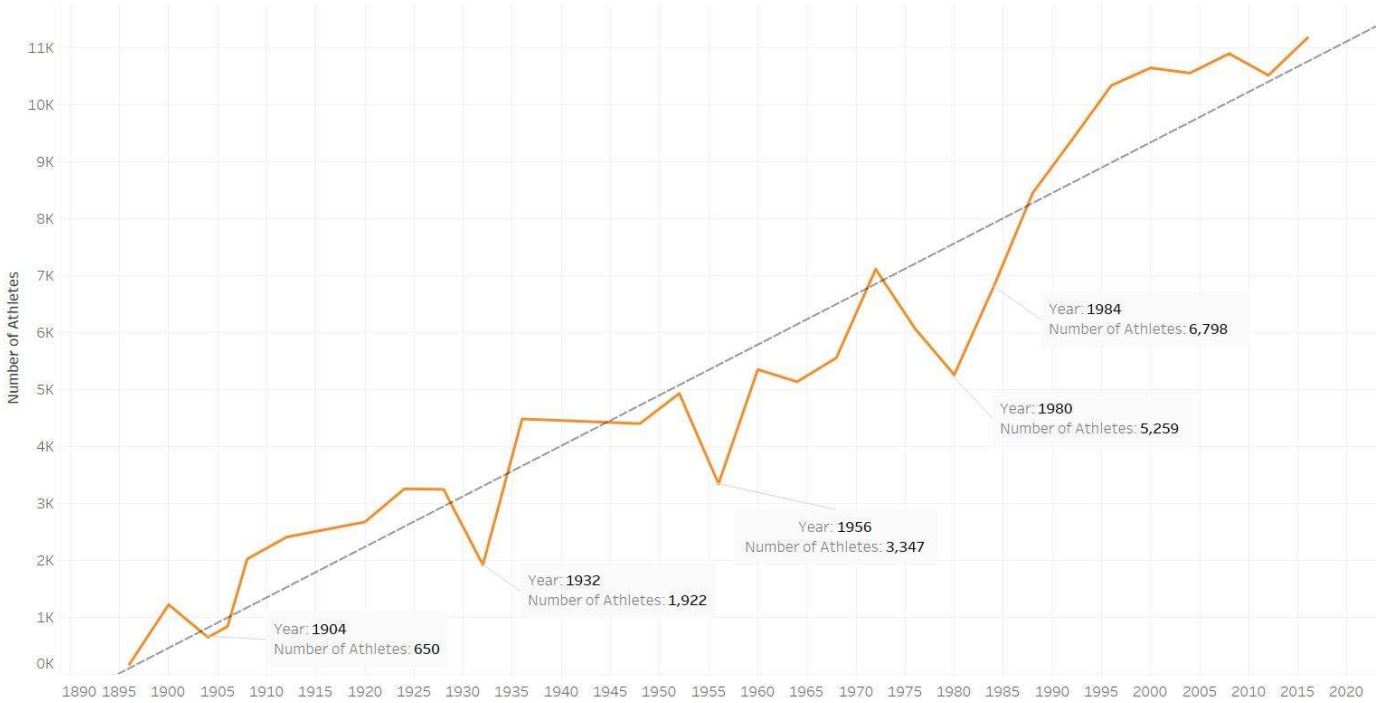
- There are data that did not comply with the hypothesis.

```sql
SELECT sum(temp_table.number_of_athletes) as num_athletes_region, COALESCE(sum(temp_win_table.number_of_winner_athletes),0) as num_winner_athletes_region , temp_table.region
FROM
  (
  SELECT COUNT(DISTINCT A.ID) as number_of_athletes, A.Season, A.Year, N.region
  FROM AthleteEventsBronze A
  LEFT JOIN noc_regions N
  ON A.NOC=N.NOC
  GROUP BY A.Season, A.Year, N.region
  ) as temp_table
LEFT JOIN
  (
  SELECT COUNT(DISTINCT A.ID) as number_of_winner_athletes, A.Season, A.Year, N.region
  FROM AthleteEventsBronze A
  LEFT JOIN noc_regions N
  ON A.NOC=N.NOC
  WHERE Medal <> "NA"
  GROUP BY A.Season, A.Year, N.region
  ) as temp_win_table ON temp_table.region=temp_win_table.region AND temp_table.year=temp_win_table.year AND temp_table.season=temp_win_table.season
GROUP BY temp_table.region
```

	num_athletes_region	num_winner_athletes_region	region
1	12853	4658	USA
2	10654	3261	Germany
3	8490	1877	UK
4	8280	1520	France
5	8100	3401	Russia
6	7226	1461	Italy
7	6467	1256	Canada
8	5534	1158	Australia
9	5458	788	Japan
10	5417	1380	Sweden
11	4239	538	Poland

Deeper Findings

- The numbers of participating athletes are on an upward trend through the years.
- There are some points in Summer Games that the numbers of participating athletes are significantly downed compared to the prior years.



Deeper Findings

- Why did these decreases happen?
  - In **1904** the Olympic Games were held outside Europe for the first time. Very few top-class athletes from outside the US and Canada took part in the games due to the difficulties in traveling to St.Louis and also tensions caused by the Russo-Japanese War.
  - There were no Olympic Games in **1916**, **1940** and **1944** due to war.
  - In **1932** the Olympic Games were held in the middle of the Great Depression in the relatively remote region of California that was difficult to transport for that time.

Deeper Findings

- In **1956** eight countries boycotted the Games for protest.
  - Egypt, Iraq, Cambodia, and Lebanon were in protest at the Suez Crisis when Egypt was invaded by Israel, the United Kingdom, and France.
  - The Netherlands, Spain, and Switzerland were in protest at the Soviet Union presence in light of their recent crushing of the Hungarian Revolution.
  - The People's Republic of China chose to boycott the event because Taiwan had been allowed to compete.

Deeper Findings

- In **1980** the United States led a boycott in response to the Soviet Invasion of Afghanistan. 65 nations refused to participate in the games held in Moscow.
- In **1984** the Soviet Union and its allies boycotted the Summer Olympics in Los Angeles.

Deeper Findings

- The mean ages of athletes for the Summer seasons and the Winter seasons are respectively 25.9 and 25.1.
- The mean ages have supported the hypothesis but there are some outlier data in the Summer games that need to be checked.

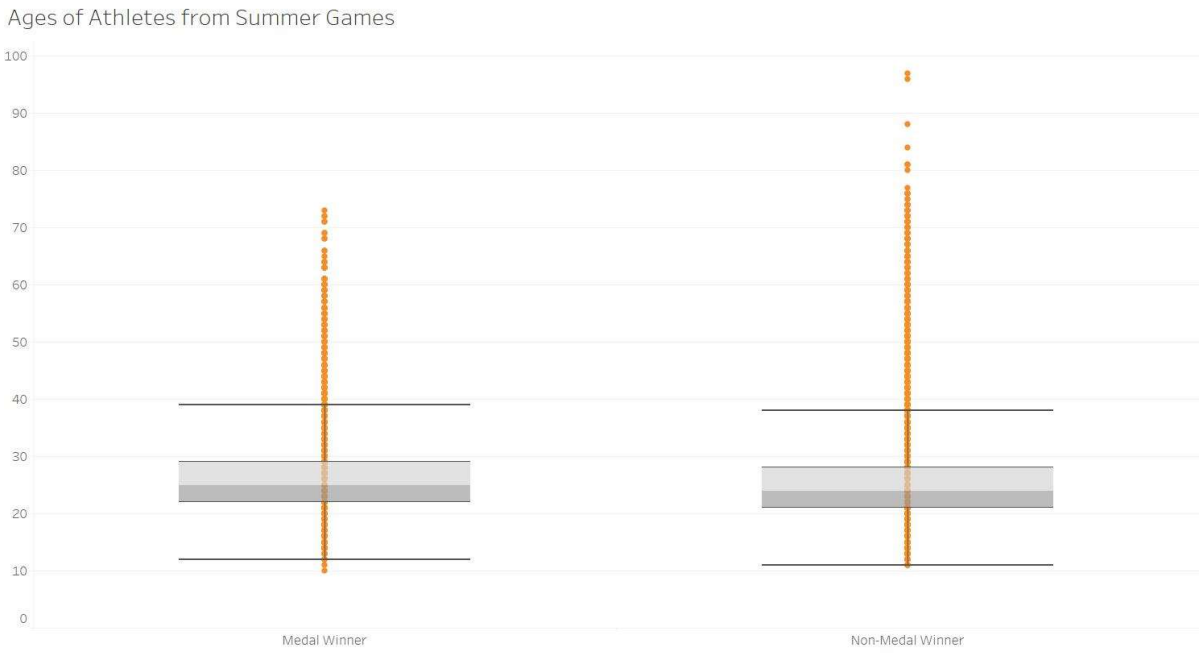
```sql
SELECT Season, COUNT(*) as Number_of_athletes, MIN(Age) AS min_age, MAX(age) AS max_age, AVG(age) AS avg_age
FROM (
SELECT DISTINCT ID, Name, Sex, Age, Year, Season
FROM AthleteEventsBronze
)
GROUP BY Season
```

(3) Spark Jobs

	Season	Number_of_athletes	min_age	max_age	avg_age
1	Summer	158858	10	97	25.907778653012997
2	Winter	28593	11	58	25.100503751717337

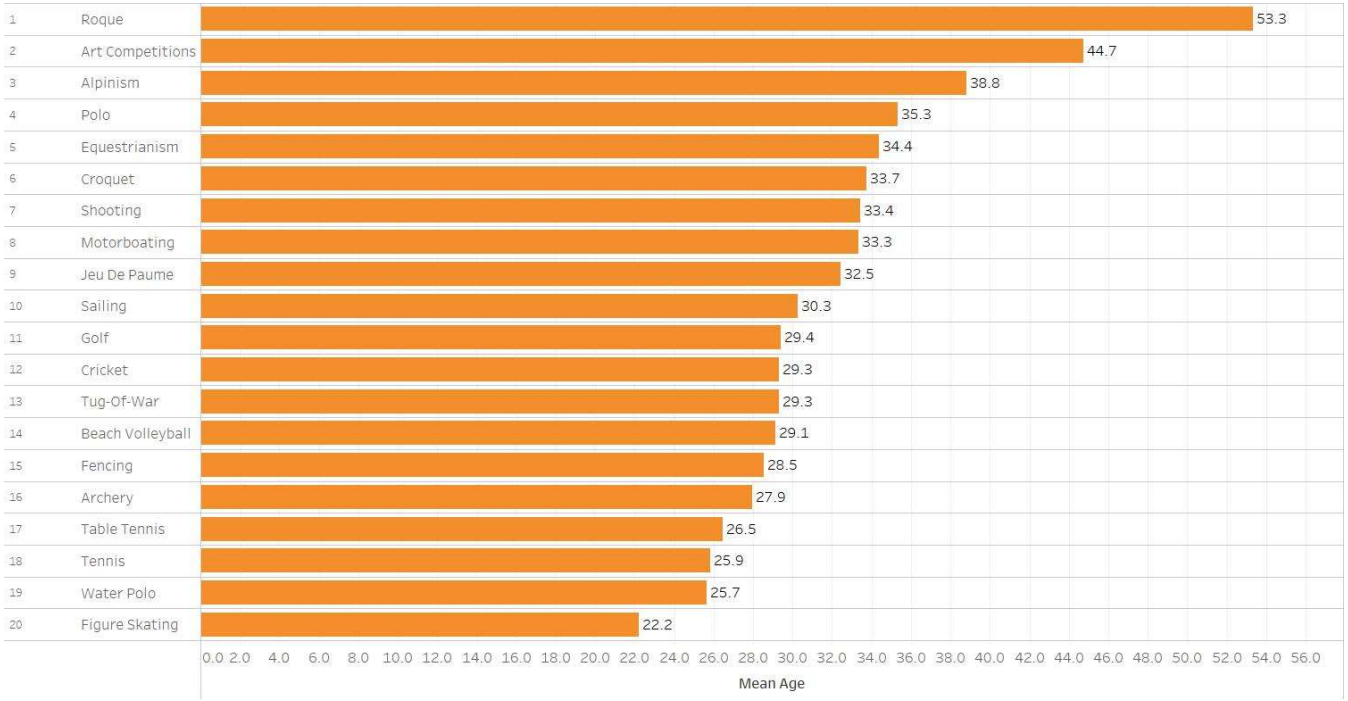
Deeper Findings

- The medal winner athletes' ages were not significantly different than others.
- There are many outliers at old aged athletes.
- The all outliers data is valid.



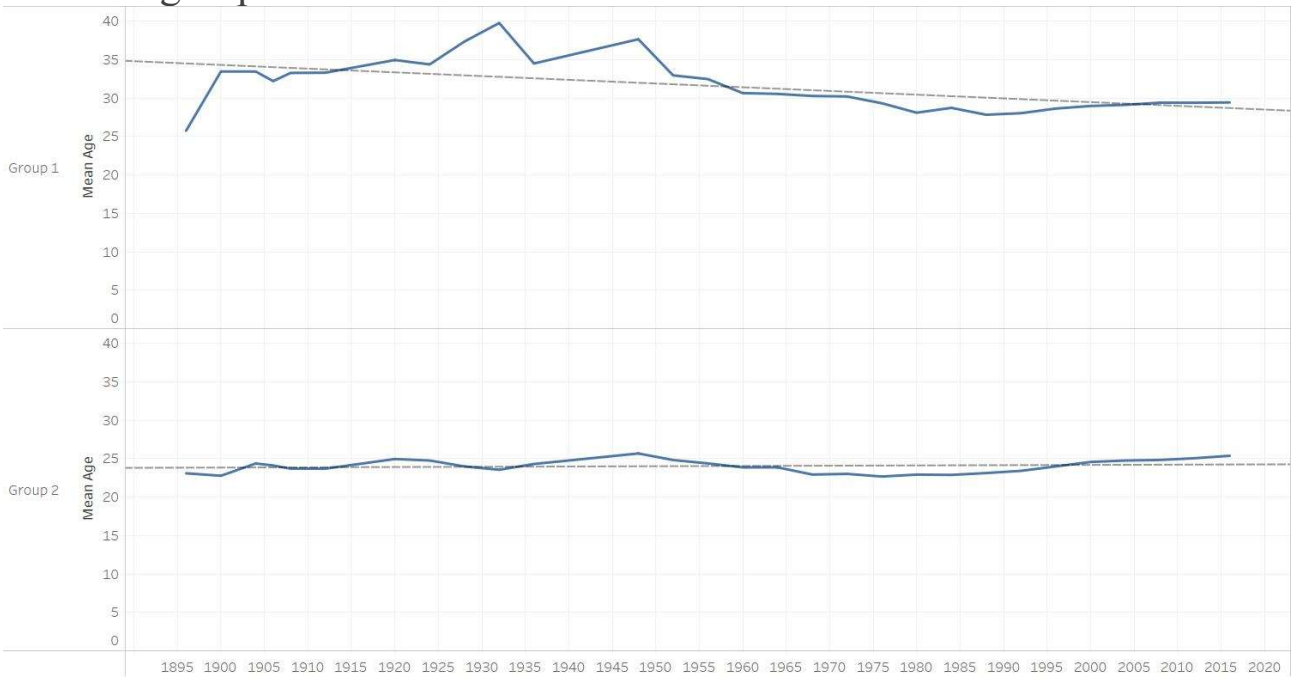
Deeper Findings

- The sports that age 38 is below upper whisker boundry:



Deeper Findings

- The sports that older-aged athletes are more common are added to group-1. The other summer sports are added to group-2.

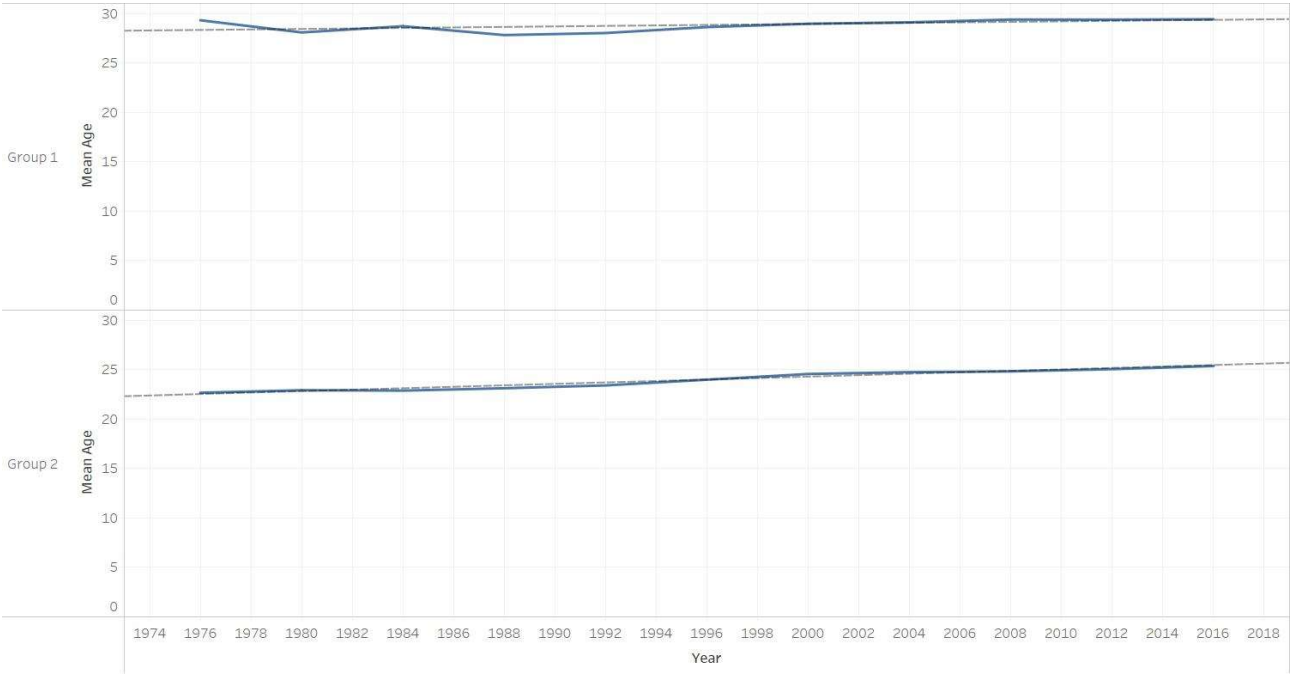


- There is a significant decrease trend in group 1 ages.

- There is not a significant increase trend in group 2 ages.

Deeper Findings

- When we filter the data from 1976 to 2016.



- There is not a significant increase trend in group 1 ages (9 sports from the group 1 were excluded from games before that period).
- There is a significant increase trend in group 2 ages.

Beyond Descriptive Analysis of Number of Athletes for Regions

- Although there are contradictory data available, the number of participating athletes and the number of winner athletes are in positive correlation (r=0.93).

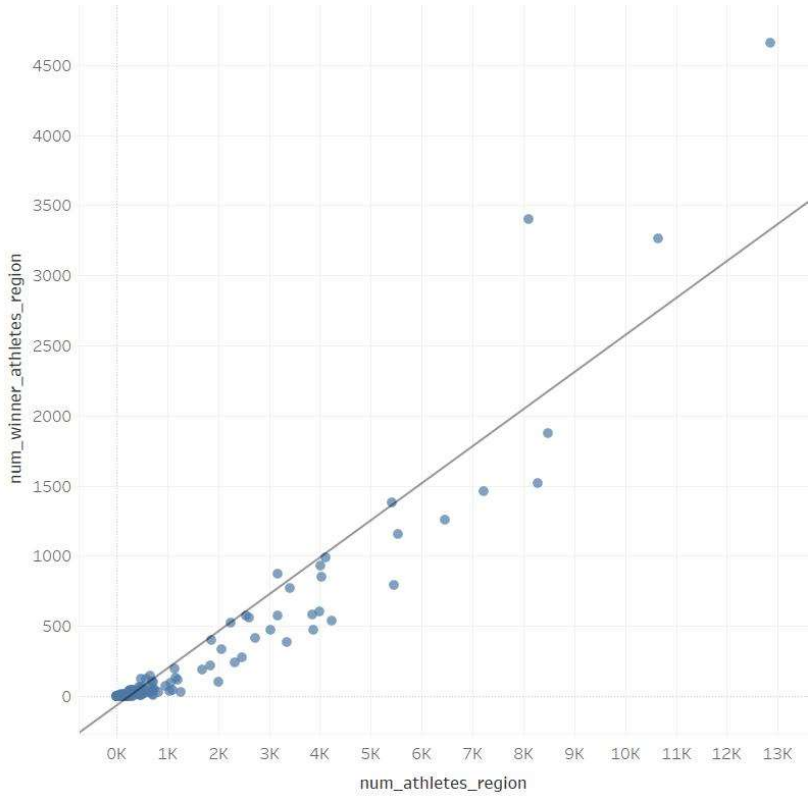
SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.937620921
R Square	0.879132992
Adjusted R Square	0.878543397
Standard Error	186.4263513
Observations	207

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	51822130.92	51822130.9	1491.07905	5.12584E-96
Residual	205	7124730.814	34754.7845		
Total	206	58946861.74			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-69.40716623	14.36264148	-4.8324792	2.6392E-06	-97.72460043	-41.08973204	-97.72460043	-41.08973204
num_athletes_region	0.264198555	0.006841953	38.6144927	5.1258E-96	0.250708936	0.277688174	0.250708936	0.277688174



Hypthoses Results

- The decreases at the numbers of participating athletes shown that the Olympic Games were affected by other global social events.
- The peak performance age of athletes is between 25 and 26 when considering all of the sports.
- The medal winner athletes' age and non-medal winner athletes' age didn't differ.
- The athletes' age groups for sports that require less speed and power are different than other sports. When the sports in Summer Games were grouped into two categories, the mean age for sports that require less speed and power is 30.8, and the mean age for the other group is 24.1.
- The athlete ages' for required speed and power is not significantly increased from 1896 to 2016 in Summer Games. When we filter the data from 1976 to 2016 there is a significant increase trend shown.
- There is a positive correlation between the numbers of participating athletes and the numbers of medal winner athletes for regions.