Analyzing Olympic Games Dataset

Data Analysis Project for SportStats

12/2023

Rostyslav Husaruk

Initial Overview

About Client

I chose **"SportsStats"** as my client. SportsStats is a sports analysis firm that collaborates with local news outlets and elite personal trainers. Their mission is to deliver captivating insights to assist their partners. These insights encompass patterns and trends, shedding light on specific groups, events, countries, and more. The primary objective is to facilitate the development of engaging news stories or uncover essential health insights.

Why I Chose This Client?

I chose "SportsStats" as my client due to my personal interest in the realm of professional competition within the sports field. I always wanted to participate in athletic contests. I want to find some insights that can be helpful to understand the topic better and maybe someday it will possibly benefit my own athletic pursuits.

Importing and Instruments

Describing the steps I undertook for importing and cleaning the data

- Download dataset
- Import it in jupyter lab
- Cleaning
- Explore dataset

Instruments that I will need:

- Jupyterlab
- Python
- Pandas
- Numpy
- Matplotlib
- Seaborn

Project Proposal

There is a lot of useful information that is hidden from our sight. I want to try to uncover it and show my findings to my audience, so they can achieve better results.

Becoming a winner requires a lot of work. With my project, I aim to gather valuable information on athletes who have achieved great results to identify commonalities among them. The focal audience for my findings is individuals with a keen interest in professional sports.

About the dataset: It contains information spanning approximately 100 years, starting from 1896, with almost 300,000 rows of data.

Questions

Years and Seasons

- Are there any sports considered irrelevant?
- What are the differences between the Summer and Winter seasons?
- How has the number of events changed over the years?"

Athlete Statistics

- What is the distribution of athletes across different age groups?
- What is the distribution of wins based on height and weight?

Athlete Statistics

- How are wins distributed among different age groups?
- Is the distribution of wins aligned with the win rate for each age?

More deeper analysis of age

- Which events show the highest number of wins among athletes aged 26-31?
- Which events have the highest number of wins among athletes aged 40+?
- What differences can be observed between these two categories of athletes?

Data Analysis Approach

Data Analysis Approach

Regarding my approach, I will be using columns with athletes' statistics and medals. Age is a crucial factor as it indicates a person's physical condition. Exploring height and weight will provide additional insights into athletes. The medal is significant, serving as both the final result and an indicator of success.

DataFrames

Creating dataframes for my exploration

	Age	Gold	Silver	Bronze	Medal_sum
3	13.0	7	7	2	16
4	14.0	27	29	18	74
5	15.0	73	67	54	194
6	16.0	116	129	105	350
7	17.0	199	163	170	532

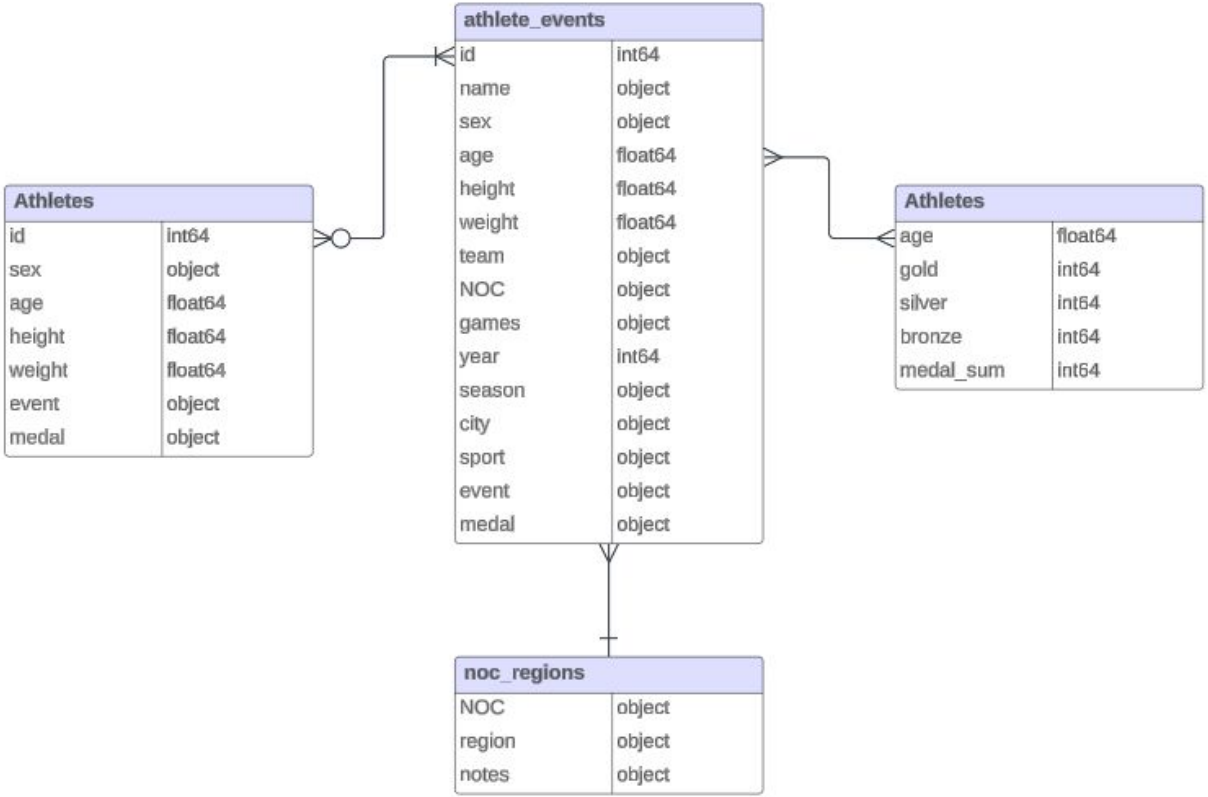
	Sex	Age	Height	Weight	Event	Medal
0	M	24.0	180.0	80.0	Basketball Men's Basketball	NaN
1	M	23.0	170.0	60.0	Judo Men's Extra-Lightweight	NaN
2	M	24.0	NaN	NaN	Football Men's Football	NaN
4	F	21.0	185.0	82.0	Speed Skating Women's 500 metres	NaN
5	F	21.0	185.0	82.0	Speed Skating Women's 1,000 metres	NaN
...	...	...	...	...	...	...
271111	M	29.0	179.0	89.0	Luge Mixed (Men)'s Doubles	NaN
271112	M	27.0	176.0	59.0	Ski Jumping Men's Large Hill, Individual	NaN
271113	M	27.0	176.0	59.0	Ski Jumping Men's Large Hill, Team	NaN
271114	M	30.0	185.0	96.0	Bobsleigh Men's Four	NaN
271115	M	34.0	185.0	96.0	Bobsleigh Men's Four	NaN

266842 rows × 6 columns

age_medal_df - to convert medals into numerical values.

df_stats - to work with the most useful columns in my opinion from this dataset.

ERD


athlete_events
id int64
name object
sex object
age float64
height float64
weight float64
team object
NOC object
games object
year int64
season object
city object
sport object
event object
medal object
Athletes
id int64
sex object
age float64
height float64
weight float64
event object
medal object
Athletes
age float64
gold int64
silver int64
bronze int64
medal_sum int64
noc_regions
NOC object
region object
notes object

Cleaning

- NaN values
- Deleting Columns
- Deleting Duplicates
- Irrelevant sport
- Cleaning weight column

NaN values

Identify regions with NaN values and replace them with the correct region names 'Tuvalu' and 'Refugee Olympic Team' respectively.

```
df2.loc[pd.isna(df2.region)]
```

	NOC	region	notes
168	ROT	NaN	Refugee Olympic Team
208	TUV	NaN	Tuvalu
213	UNK	NaN	Unknown

```
df2.loc[df2.NOC == 'TUV', ['region']] = 'Tuvalu'
df2.loc[df2.NOC == 'ROT', ['region']] = 'Refugee Olympic Team'
```

```
df1 = df1.drop('Games', axis=1)
```

```
df1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 266912 entries, 0 to 271115
Data columns (total 14 columns):
 #   Column  Non-Null Count   Dtype
---  ------  --------------   -----
 0   ID      266912 non-null  int64
 1   Name    266912 non-null  object
 2   Sex     266912 non-null  object
 3   Age     258093 non-null  float64
 4   Height  210788 non-null  float64
 5   Weight  208096 non-null  float64
 6   Team    266912 non-null  object
 7   NOC     266912 non-null  object
 8   Year    266912 non-null  int64
 9   Season  266912 non-null  object
 10  City    266912 non-null  object
 11  Sport   266912 non-null  object
 12  Event   266912 non-null  object
 13  Medal   39128 non-null   object
dtypes: float64(3), int64(2), object(9)
memory usage: 30.5+ MB
```

Deleting Columns. Cleaning Weight column

The ‘Games’ column duplicates information already present in ‘Year’ and ‘Season’.

The ‘Weight’ column contains float values. Round the values in this column for consistency.

```
df.Weight = df.Weight.round()
```

Deleting Duplicates

```
duplicates_all = df1[df1.duplicated(keep=False)]
duplicates_all.info()
```

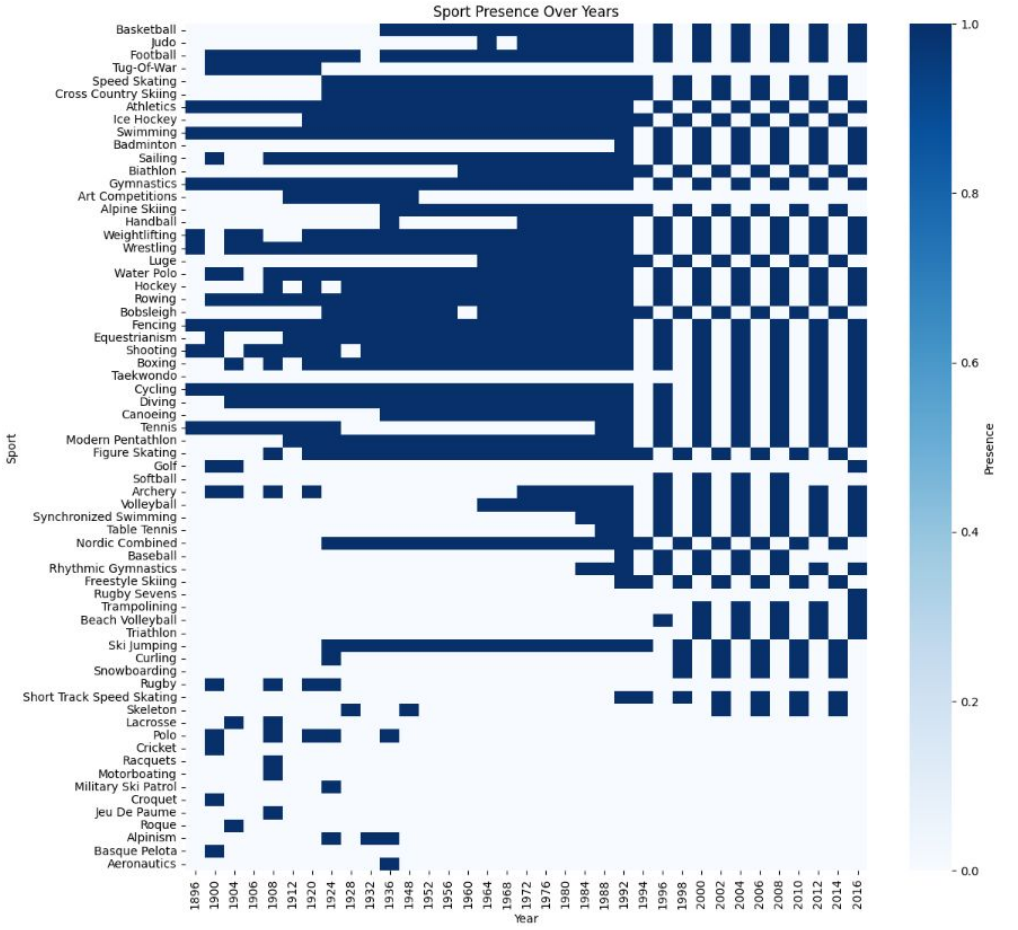
```
<class 'pandas.core.frame.DataFrame'>
Index: 140 entries, 10866 to 257121
Data columns (total 14 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   ID      140 non-null    int64
 1   Name    140 non-null    object
 2   Sex     140 non-null    object
 3   Age     76 non-null     float64
 4   Height  2 non-null      float64
 5   Weight  2 non-null      float64
 6   Team    140 non-null    object
 7   NOC     140 non-null    object
 8   Year    140 non-null    int64
 9   Season  140 non-null    object
 10  City    140 non-null    object
 11  Sport   140 non-null    object
 12  Event   140 non-null    object
 13  Medal   22 non-null     object
dtypes: float64(3), int64(2), object(9)
memory usage: 16.4+ KB
```

```
df1.drop_duplicates(inplace=True)
```

This dataset contains some duplicate data, 140 rows.

Irrelevant sport

As observed from the heat map, certain sports became irrelevant and were no longer present in the general list after 1940.



Irrelevant Sport. Deleting Rows

Deleting rows with data on irrelevant sports events, affecting approximately 5000 rows.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 271116 entries, 0 to 271115
Data columns (total 15 columns):
 #   Column  Non-Null Count   Dtype
---  ------  --------------   -----
 0   ID      271116 non-null  int64
 1   Name    271116 non-null  object
 2   Sex     271116 non-null  object
 3   Age     261642 non-null  float64
 4   Height  210945 non-null  float64
 5   Weight  208241 non-null  float64
 6   Team    271116 non-null  object
 7   NOC     271116 non-null  object
 8   Games   271116 non-null  object
 9   Year    271116 non-null  int64
 10  Season  271116 non-null  object
 11  City    271116 non-null  object
 12  Sport   271116 non-null  object
 13  Event   271116 non-null  object
 14  Medal   39783 non-null   object
dtypes: float64(3), int64(2), object(10)
memory usage: 31.0+ MB
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 266912 entries, 0 to 271115
Data columns (total 15 columns):
 #   Column  Non-Null Count   Dtype
---  ------  --------------   -----
 0   ID      266912 non-null  int64
 1   Name    266912 non-null  object
 2   Sex     266912 non-null  object
 3   Age     258093 non-null  float64
 4   Height  210788 non-null  float64
 5   Weight  208096 non-null  float64
 6   Team    266912 non-null  object
 7   NOC     266912 non-null  object
 8   Games   266912 non-null  object
 9   Year    266912 non-null  int64
 10  Season  266912 non-null  object
 11  City    266912 non-null  object
 12  Sport   266912 non-null  object
 13  Event   266912 non-null  object
 14  Medal   39128 non-null   object
dtypes: float64(3), int64(2), object(10)
memory usage: 32.6+ MB
```

Data Exploration

- Creating dataframes for my exploration
- Working on season table
- 1. On summer
- 2. On winter
- Analyzing Number of Events Each Year
- Exploring Athletes by Age
- Examining Wins by Weight
- Examining Wins by Height
- Comparing Prizes by Age
- Golden years for athletes
- Determining Sports with the Most Wins for Younger Athletes
- Identifying Sports with the Most Wins for Athletes 40+

Working on Sport, Season, and Year columns

Sorting the table by season, we can now observe the differences in data between the winter and summer seasons. Summer has four times more entries.

```
df.loc[df.Season == 'Winter'].count()
```

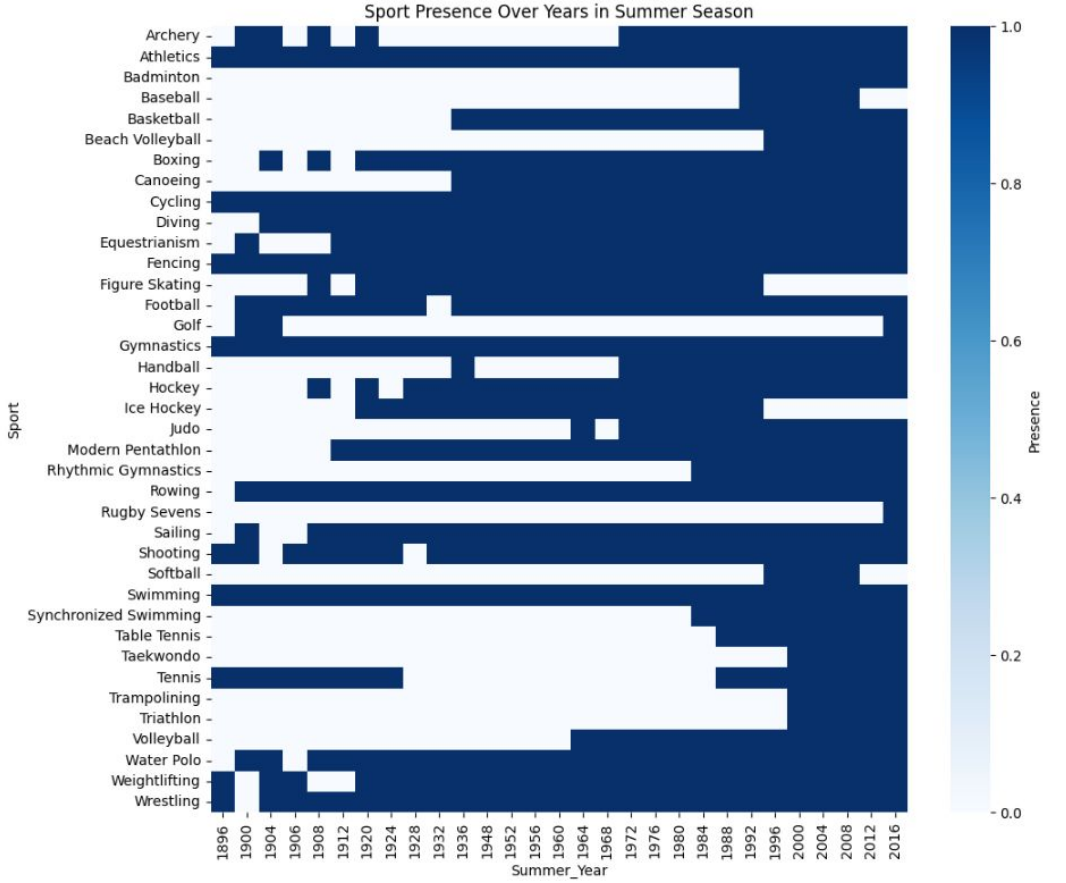
```
ID        48519
Name      48519
Sex       48519
Age       48248
Height    40250
Weight    39543
Team      48519
NOC       48519
Year      48519
Season    48519
City      48519
Sport     48519
Event     48519
Medal      5662
dtype: int64
```

```
df.loc[df.Season == 'Summer'].count()
```

```
ID        218323
Name      218323
Sex       218323
Age       209807
Height    170537
Weight    168552
Team      218323
NOC       218323
Year      218323
Season    218323
City      218323
Sport     218323
Event     218323
Medal      33455
dtype: int64
```

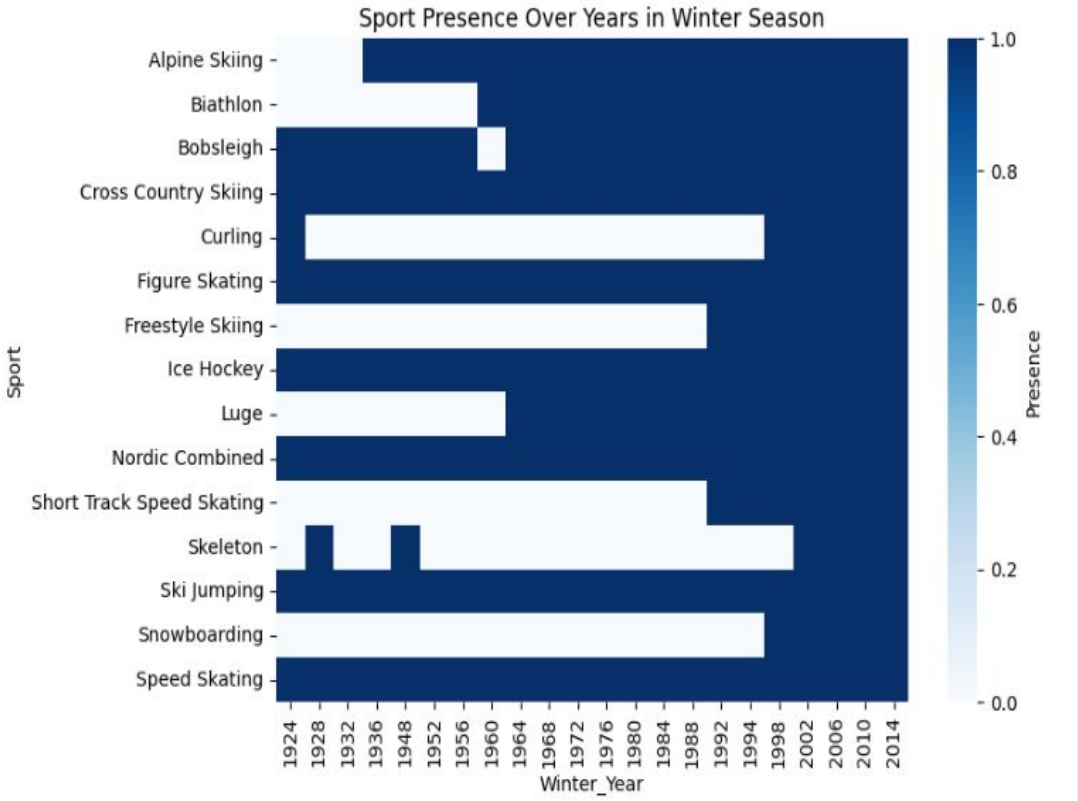
Summer

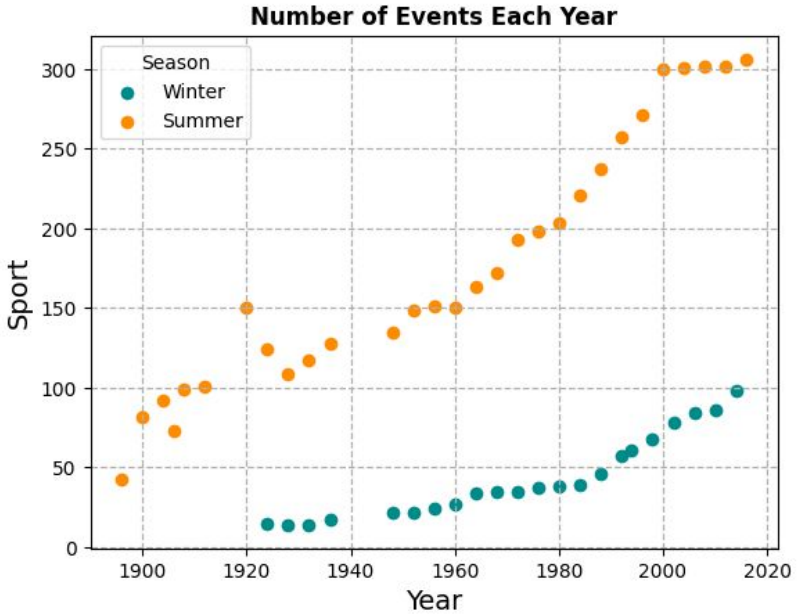
From this visualization, we can observe that only 10% of all Summer sports (Athletics, Cycling, Gymnastics, and Swimming) have been played since the start of the Olympic Games



Winter

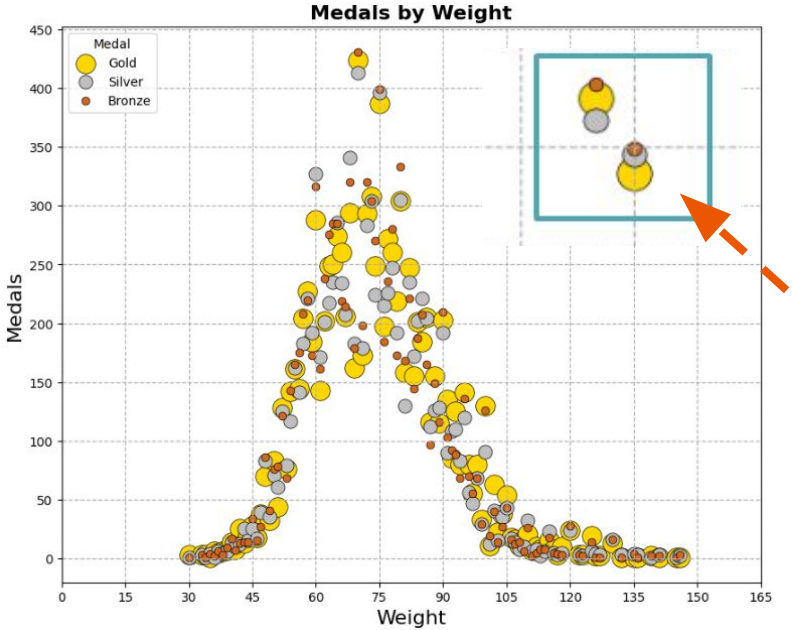
The visualization of sport presence in Winter shows that 40% of all Winter sports have been played since start of Winter Olympic games.





Analyzing the Number of Events Each Year

This visualization illustrates how the number of events has changed over time. It also highlights that the Winter Olympic Games have three times fewer events than the Summer Olympic Games.



Examining Wins by Weight

Starting from 45 kg, the number of medals per weight category rapidly increases until 70 kg. After that, we observe a slower decline until 105 kg.

Also, take note of these two outliers. After noticing them, I decided to investigate why these two weights have almost 100 more medals than the others.

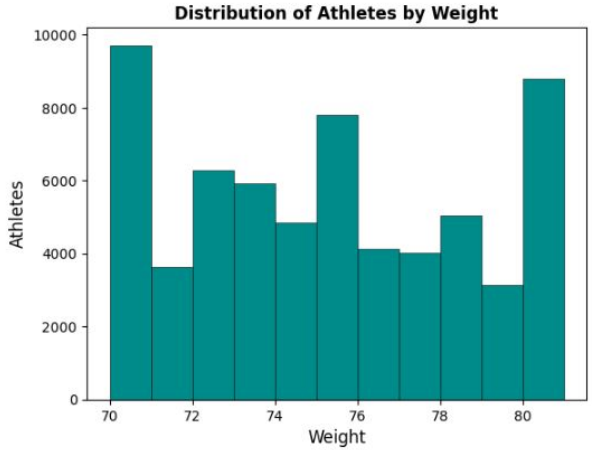
Steps I Will Take and Hypotheses for Analysis

Steps I Will Take to Find Answers to My Question:

1. Analyze the distribution of weights between 70-80 kg.
2. Check the number of rows with weights 70 and 71.
3. Examine martial arts events
4. Round numbers for better analysis.
5. Investigate the top 10 sports for each weight category.
6. Create a bar plot illustrating weights over the

I made a list of hypotheses that could explain such results:

- 1. Influence of martial arts on weight distribution.
- 2. The impact of round numbers and human tendencies.
- 3. Influence of trainers and diet.
- 4. The possibility of a specific weight being more favorable for winning."




```
df.Weight.loc[df.Weight == 70].count()
```

```
9713
```

```
df.Weight.loc[df.Weight == 71].count()
```

```
3636
```

Distribution of weight

Notably, weights 70 kg and 75 kg have more athletes than their neighboring weights. It's worth mentioning that 71 kg has three times fewer athletes compared to 70 kg.

What can be the reason for such difference?
Let’s find out

Martial arts

I know that athletes in martial arts sometimes aim for the highest weight in their category to enhance their chances of winning.

However, as we can see, the difference is very small. Therefore, this hypothesis was incorrect.

Weight	Event	Weight_sum
70.0	Boxing Men's Light-MiddleweightBoxing Men's Mi...	539
71.0	Boxing Men's FlyweightBoxing Men's Heavyweight...	455

Round Numbers and Humans

As humans, we often seek more understandable and straightforward goals. In the realm of professional sports, maintaining a physique conducive to better chances of winning is crucial. Setting a target weight of 70 kg or 75 kg before a tournament sounds like a practical goal.

However, consider the less common weights like 73 kg, 72 kg, and 71 kg – these might not be as popular targets due to their complexity. Creating tasks that are simple to understand is the initial step in the journey of achieving them.

Top 10 Events for Every Weight

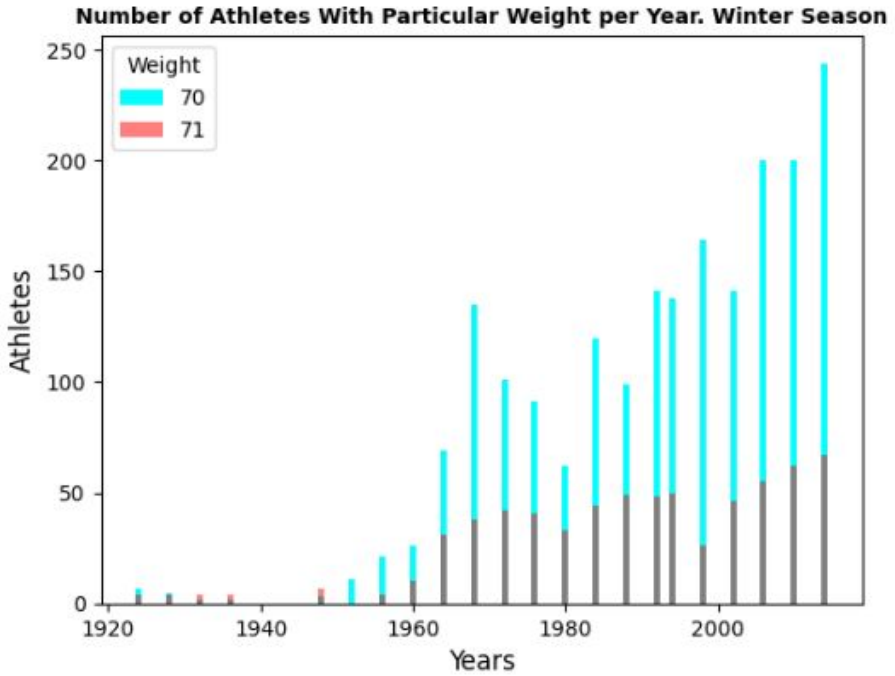
Interestingly, every weight category from the top places in the table with a weight of 70 kg surpasses 71 kg by a factor of two. Therefore, the theory that the observed difference is due to weight rules or better odds to win in specific events is proven incorrect.

	Weight	Event	Weight_sum
484	71.0	Boxing Men's Light-Middleweight	211
568	71.0	Football Men's Football	126
528	71.0	Cycling Men's Road Race, Individual	92
592	71.0	Judo Men's Lightweight	84
586	71.0	Hockey Men's Hockey	82
532	71.0	Cycling Men's Team Pursuit, 4,000 metres	54
431	71.0	Athletics Men's 4 x 400 metres Relay	53
585	71.0	Handball Women's Handball	50
730	71.0	Volleyball Women's Volleyball	49
432	71.0	Athletics Men's 400 metres	46

	Weight	Event	Weight_sum
193	70.0	Football Men's Football	324
214	70.0	Hockey Men's Hockey	191
146	70.0	Cycling Men's Road Race, Individual	190
25	70.0	Athletics Men's 4 x 100 metres Relay	142
26	70.0	Athletics Men's 4 x 400 metres Relay	129
18	70.0	Athletics Men's 100 metres	118
391	70.0	Volleyball Women's Volleyball	115
33	70.0	Athletics Men's 800 metres	107
120	70.0	Cross Country Skiing Men's 4 x 10 kilometres R...	100
213	70.0	Handball Women's Handball	100

Medals by Weight Over the Years

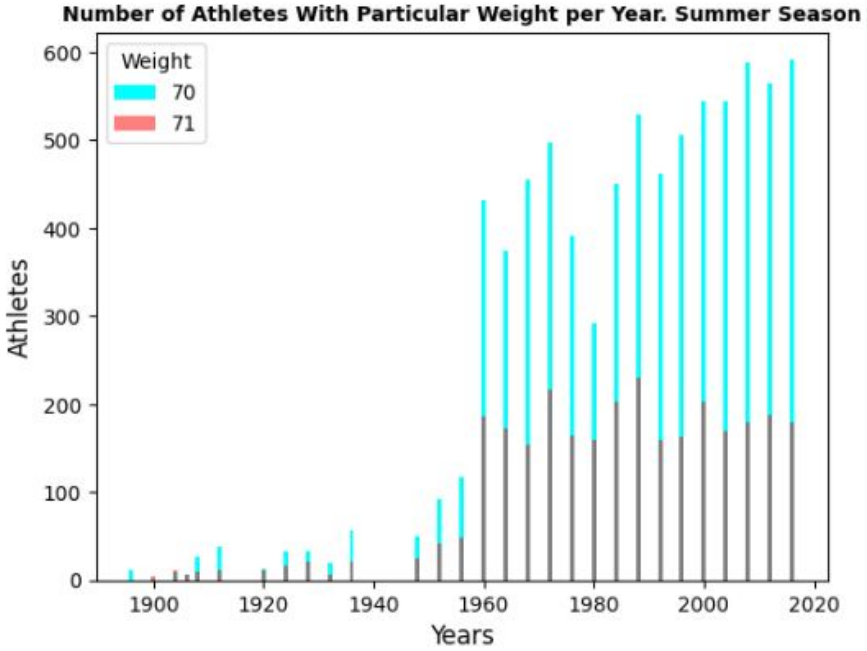
Upon analyzing across different years, it becomes apparent that the presence of athletes with a weight of 70 kg consistently surpasses those with 71 kg. This pattern persists regardless of the chosen year.



Medals by Weight Over the Years

Conclusion Part 1:

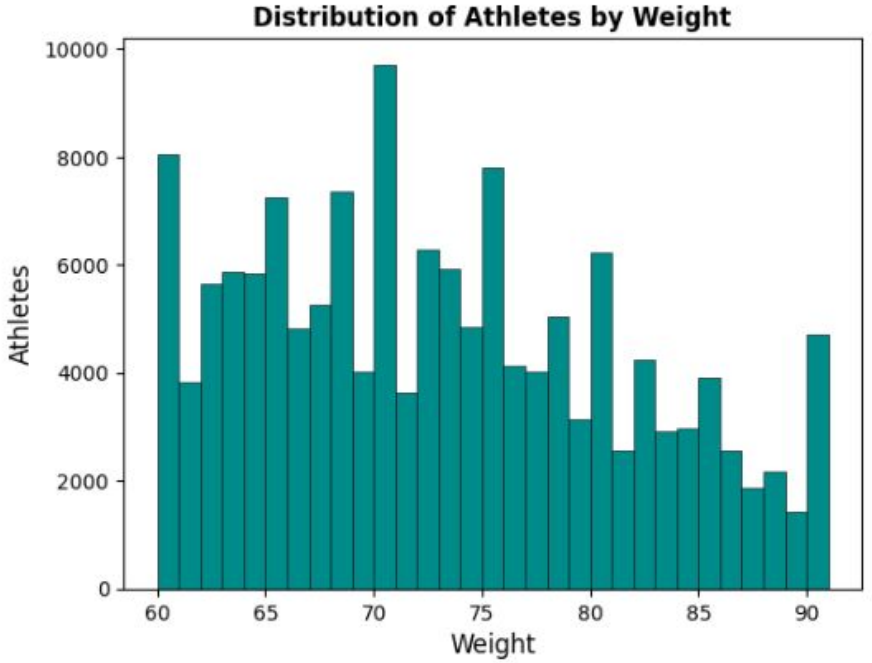
Athletes often train with coaches, and for precise dietary calculations, they tend to use round numbers. The simplicity of round weights facilitates easier adjustments to diet and training, potentially explaining the significant difference in weights observed.



The same situation with summer season

Conclusion Part 2:

I observed that there are weight peaks for every round number (60, 65, 70, 75, 80, 85, 90). This visualization provides evidence that supports my hypothesis regarding human tendencies to round numbers and as I mentioned before, coaches can also play some part in such a result.



Examining Wins by Height

The most successful results are observed in athletes with a height range of 165 - 190 cm. This is likely because it represents the average height for humans. Which is why it leads us to the conclusion:

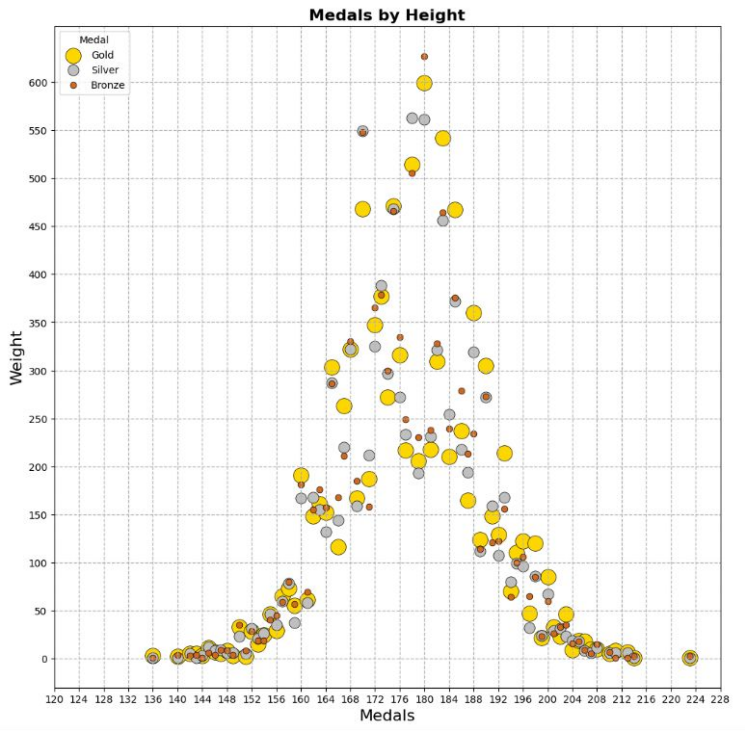
Average height

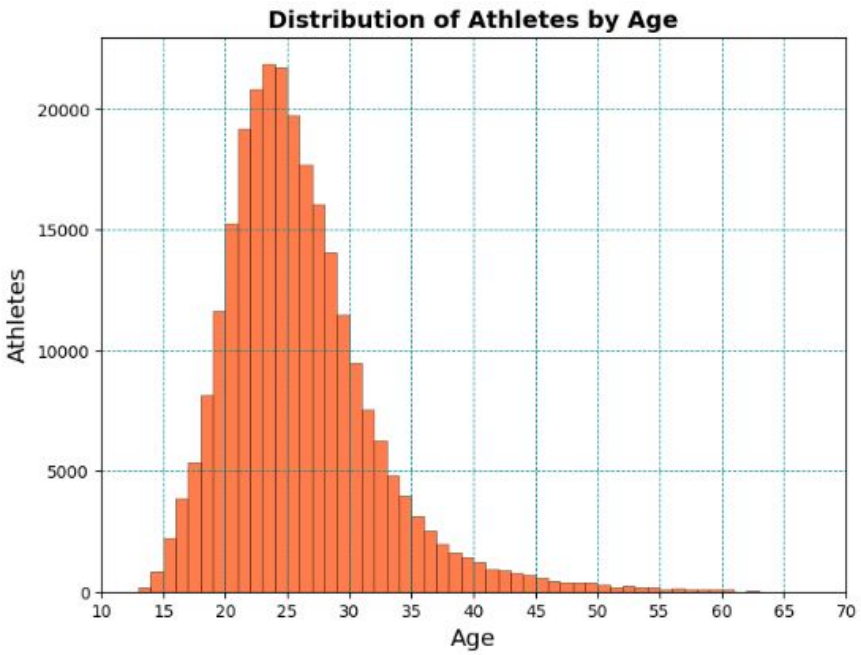
↓

More individuals with the same height

↓

More wins for that height





Each bin represents a specific age.

Exploring Athletes by Age

In this histogram, we can observe the number of participants categorized by age.

It's hard not to notice that the most common age for athletes is 21-27. Later, we will focus our analysis on these numbers and find something unexpected.

Golden years for athletes

The results, in my opinion, align with expectations. Naturally, the age when you have the best physical condition is optimal for participating in professional competitions and winning gold medals.

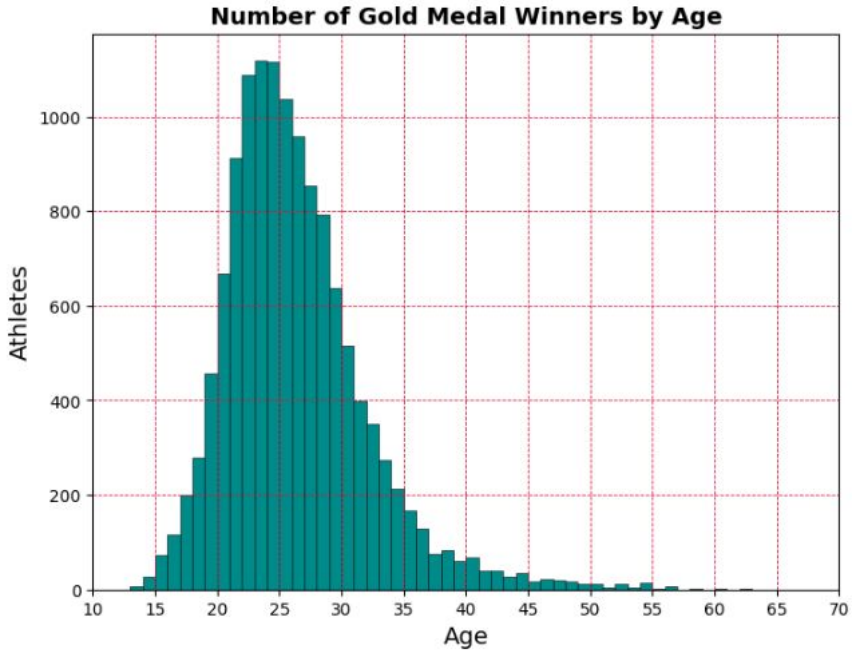
The peak physical condition

↓

More athletes with the same age

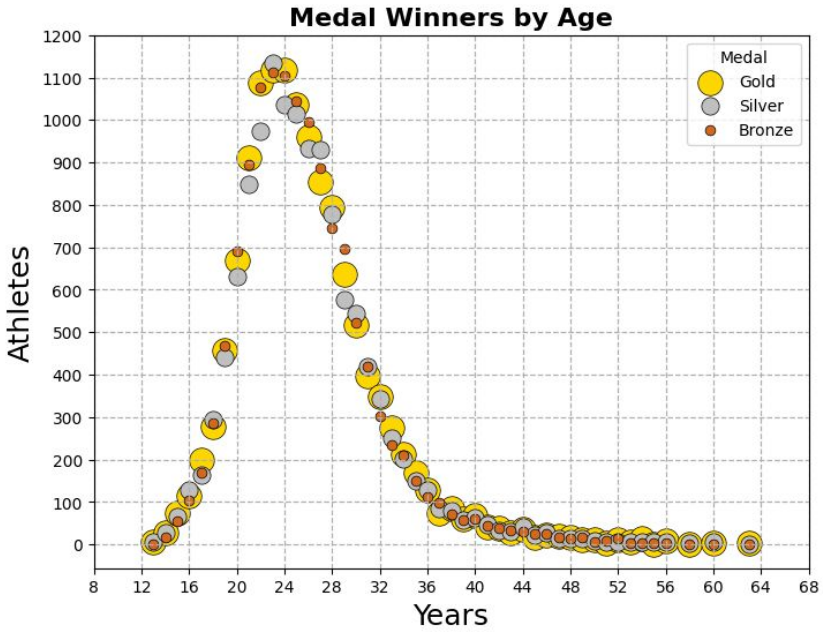
↓

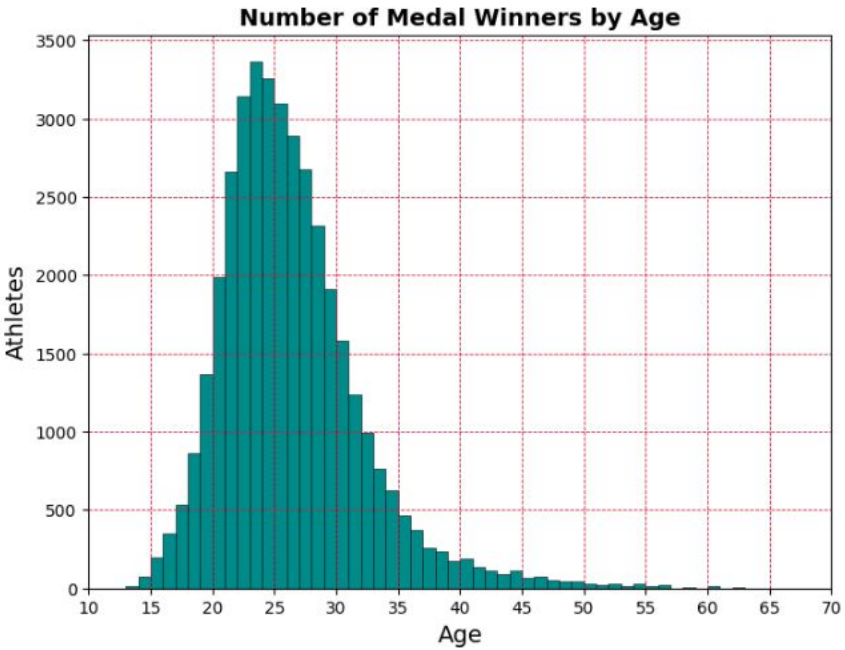
More wins in this age range



Comparing Prizes by Age

All the medals show a similar trend across different age groups. While there are variations in the number of medals per age, the overall patterns remain consistent.





Each bin represents a specific age.

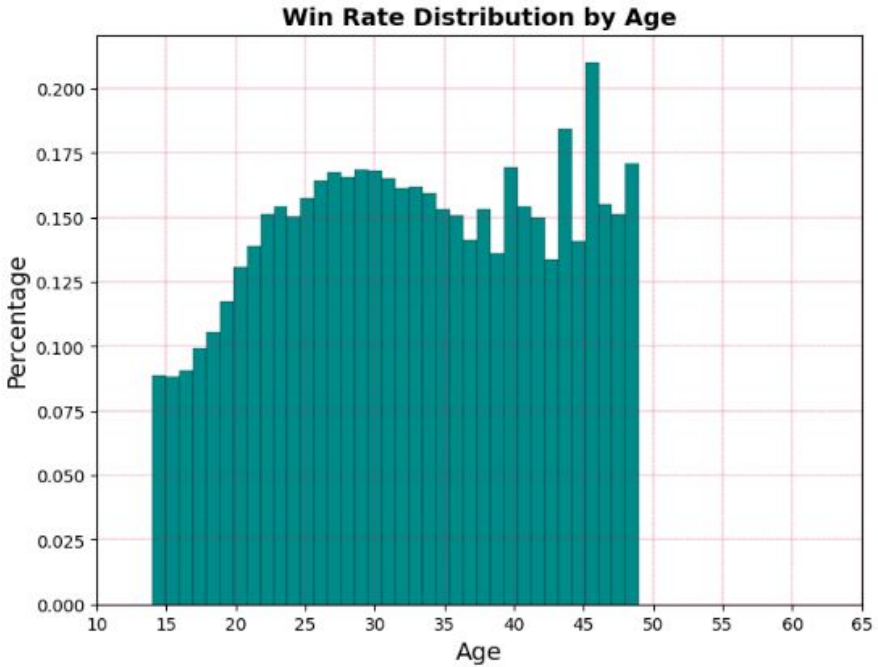
The “Golden years” for Athletes

As observed, this histogram mirrors the shape of the two previous ones. Naturally, the age at which individuals are in their peak physical condition is optimal for participating in professional competitions and achieving victory.

Initially, this was my intended interpretation, but through the course of the analysis, I came to the realization that it may not be the most effective approach to identify the “Golden years” for athletes.

Real “Golden years” for athletes

On this graph, you can observe that the actual range of the optimal age is distinct. Having more wins doesn't necessarily indicate superiority if there is a simultaneous increase in the number of athletes.



Real “Golden years” for athletes

This is how I calculated the actual 'Golden years.'

I performed the following calculation:

$$\frac{\text{num of winners}}{\text{num of all participants}}$$

	Age	Percentage
16	26.0	0.164166
17	27.0	0.167513
18	28.0	0.165821
19	29.0	0.168399
20	30.0	0.167851
21	31.0	0.165020
30	40.0	0.169476
34	44.0	0.184122
36	46.0	0.210227
39	49.0	0.170732

Order by Age

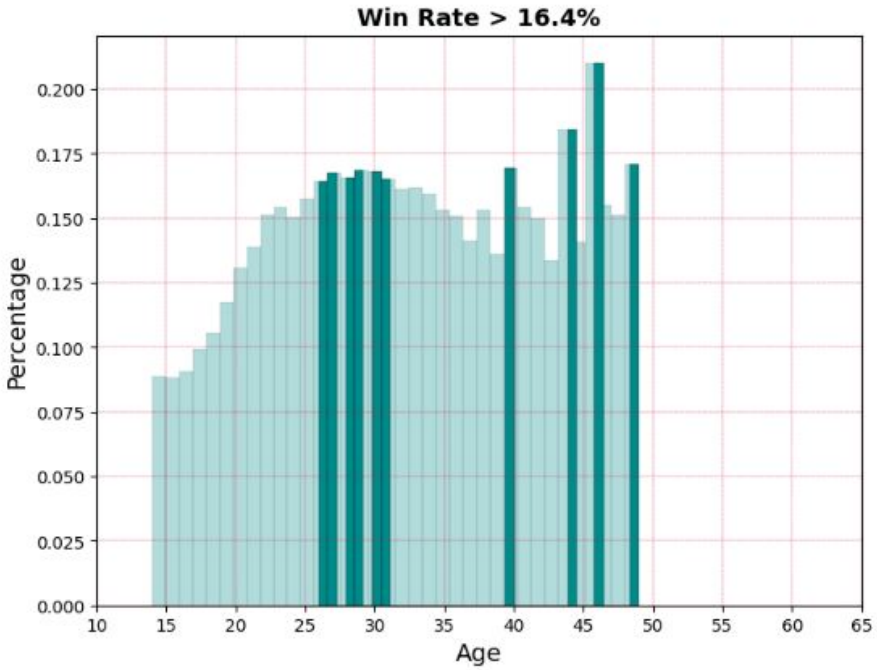
	Age	Percentage
36	46.0	0.210227
34	44.0	0.184122
39	49.0	0.170732
30	40.0	0.169476
19	29.0	0.168399
20	30.0	0.167851
17	27.0	0.167513
18	28.0	0.165821
21	31.0	0.165020
16	26.0	0.164166

Order by Percentage

Real “Golden years” for athletes

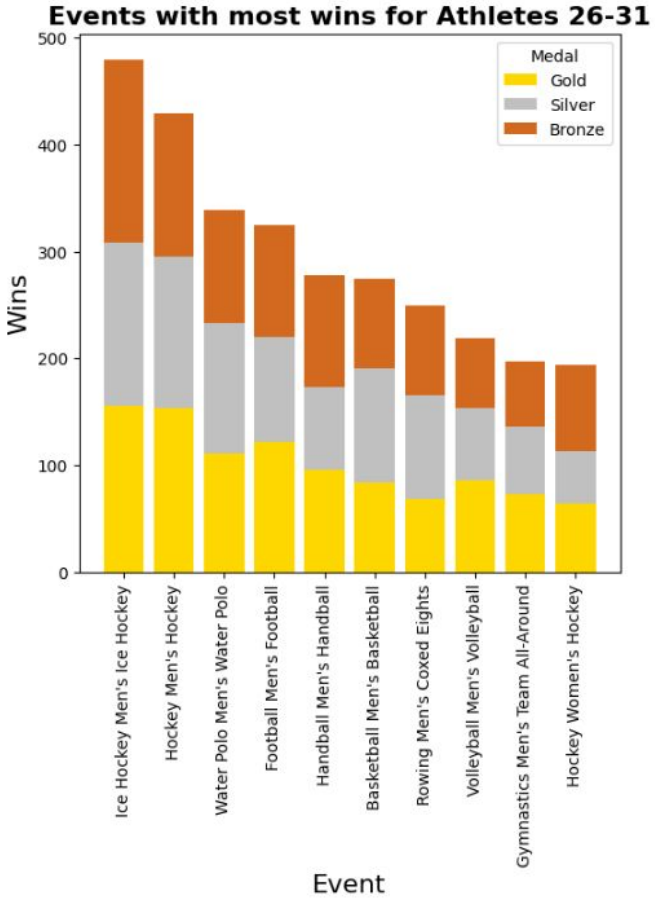
As observed in the previous slide, ages 26-31, 40, 44, 46, 49 have a winning rate exceeding 16.4%, which is 1.7% higher than the mean value.

count	36.000000
mean	0.147471
std	0.026909
min	0.088142
25%	0.138252
50%	0.153229
75%	0.164380
max	0.210227



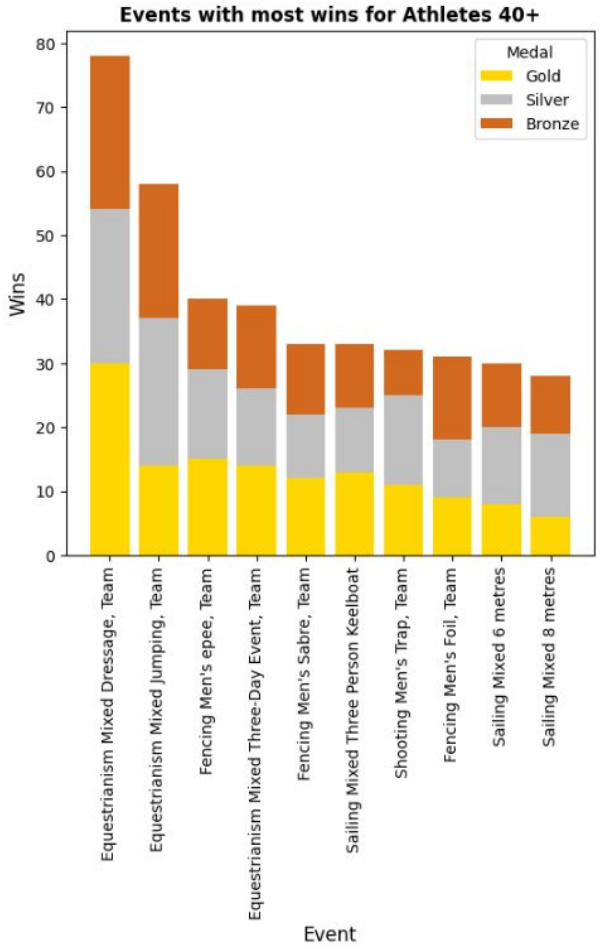
Determining Sports with the Most Wins for 26-31 Years Old Athletes

In this visualization, we can identify events with the best ratio of wins. As you may have already noticed, most of these events are very popular among young-mid aged people.



Identifying Sports with the Most Wins for Athletes 40+

This visualization, in contrast, features events that are less popular among young people. In my opinion, these events may require more financial resources to practice, such as Equestrianism or Sailing.



Differences between These Two Categories of Athletes

Reason #1

Young-mid aged people are more inclined to participate in popular sports, possibly due to affordability and accessibility, allowing anyone to engage in professional-level competition.

Reason #2

Individuals in this age range typically possess optimal physical fitness, enabling them to participate in events that demand high stamina and strength.

Reason #1

Athletes aged 40+ often have greater financial resources, allowing them to afford sports that may not be as accessible to younger individuals but are more affordable for the older demographic.

Reason #2

The events listed in the last visualization appear to require less physical ability and more knowledge, making them suitable for individuals with more experience and expertise, older age group.

Summary

Years and Seasons

- **Are there any sports considered irrelevant?**

  Yes, there were some irrelevant sports (15 out of 66).
- **What are the differences between the Summer and Winter seasons?**

  The primary distinction is that Winter Olympic Games started 28 years later than Summer Olympic Games.
- **How has the number of events changed over the years?**

  Over time, events were progressively added to the list. In the Summer season, the number changed from approximately 50 to 300 events, while in the Winter season, it evolved from around 20 to 100 events.

Athlete Statistics

- **What is the distribution of athletes across different age groups?**

  The most common age for athletes is 21-27.

- **What is the distribution of wins based on height and weight?**

  From 45 kg to 70 kg, athletes have the highest number of medals. Round numbers serve as peaks, leading to several conclusions:

  1. Athletes often train with trainers and use round numbers.
  2. People prefer simple numbers or use them as goals.
  3. Certain weights might be more favorable for winning.

Athlete Statistics Part #2

- **How are wins distributed among different age groups?**

  On the visualization we saw copy of visualization ‘Number of Athletes by Age’, which leads us to the obvious conclusion the more athletes with the same age - the more wins this age will have.

  The best results are observed in athletes aged 21-27, aligning with the expectation that the age when you have the best physical condition is optimal for participating in professional competitions and winning gold medals.

- **Is the distribution of wins aligned with the win rate for each age?**

  Based on my analysis, the distribution of wins does not align perfectly with the win rate for each age.

  The win rate for each age group varies, and the age with the most wins doesn't necessarily coincide with the age having the highest win rate. This suggests that factors beyond the sheer number of wins contribute to the overall effectiveness or success rate within each age category.

More deeper analysis of age

- **Which events show the highest number of wins among athletes aged 26-31?**

  The events with the highest number of wins among athletes aged 26-31 are often popular young-mid aged people. Many of these events are practiced during school years and university years, providing participants with a strong foundation for various tournaments.

- **Which events have the highest number of wins among athletes aged 40+?**

  Athletes aged 40 and above tend to excel in events that are less popular among the younger demographic. These events often demand more skill and knowledge than sheer physical power.

- **What differences can be observed between these two categories of athletes?**

  Notably, there are distinct differences between athletes aged 26-31 and those aged 40+. The younger group tends to showcase strength and stamina, excelling in events popularized during school years and university years. In contrast, the 40+ category, while having fewer popular events, excels in activities requiring more skill and knowledge. In summary, the two age groups exhibit different strengths, with the younger group focusing on physical prowess and the older group emphasizing experience and expertise.