Milestone 1: Project Proposal and Data Selection/Preparation

Step 1: Preparing for Your Proposal

Which client/dataset did you select and why?

I have chosen 'Client 3: SportsStats' with the Olympics Dataset for 120 years of data. The fact that I have chosen this client is that I spend a lot of my free time practicing sports, and I would like to get interesting insights from the dataset. in addition, the .csv files are not large and can be easily handled.

Describe the steps you took to import and clean the data.

First, the data was downloaded and stored locally since the volume of files is not big and does not require Databricks or several clusters to work with. I have used my local Juypter notebook text editor for coding and querying since I am used to it. I have also used Excel to check the integrity of the data and both datasets appear to be OK.

Second, I have used pandas from Python to read the .csv files, and used pandas methods info and describe to know more about the dataset

Third, I checked the amount of NaN or NULL values to know how to deal with them and remove them or not.

Perform an initial exploration of data and provide some screenshots or display some stats of the data you are looking at.

In my Python notebook, I imported a CSV file for my initial data analysis. To enhance the data cleanup process, I focused on handling missing values and eliminating duplicates from the dataset, which contained a total of 271,116 data points. Among these, 9,474 entries lacked age information, 60,171 lacked height data, and 62,875 lacked weight data. However, I ensured that essential attributes such as gender, games, and their associated details (such as year and city) remained complete. To improve efficiency and consistency, I have performed a quick EDA with pandas. The athlete_events.csv contains 271116 entries. Some columns can be dropped, since are not relevant to this analysis, for example, the 'Team' column contains some character in the string that should be removed (e.g.: Poland-1). This is more tedious than just using the 'NOC' since it gives us the same information.

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event	Medal
0	1	A Dijiang	M	24.0	180.0	80.0	China	CHN	1992 Summer	1992	Summer	Barcelona	Basketball	Basketball Men's Basketball	NaN
1	2	A Lamusi	M	23.0	170.0	60.0	China	CHN	2012 Summer	2012	Summer	London	Judo	Judo Men's Extra-Lightweight	NaN
2	3	Gunnar Nielsen Aaby	M	24.0	NaN	NaN	Denmark	DEN	1920 Summer	1920	Summer	Antwerpen	Football	Football Men's Football	NaN
3	4	Edgar Lindenau Aabye	M	34.0	NaN	NaN	Denmark/Sweden	DEN	1900 Summer	1900	Summer	Paris	Tug-Of-War	Tug-Of-War Men's Tug-Of-War	Gold
4	5	Christine Jacoba Aaftink	F	21.0	185.0	82.0	Netherlands	NED	1988 Winter	1988	Winter	Calgary	Speed Skating	Speed Skating Women's 500 metres	NaN
5	5	Christine Jacoba Aaftink	F	21.0	185.0	82.0	Netherlands	NED	1988 Winter	1988	Winter	Calgary	Speed Skating	Speed Skating Women's 1,000 metres	NaN
6	5	Christine Jacoba Aaftink	F	25.0	185.0	82.0	Netherlands	NED	1992 Winter	1992	Winter	Albertville	Speed Skating	Speed Skating Women's 500 metres	NaN
7	5	Christine Jacoba Aaftink	F	25.0	185.0	82.0	Netherlands	NED	1992 Winter	1992	Winter	Albertville	Speed Skating	Speed Skating Women's 1,000 metres	NaN
8	5	Christine Jacoba Aaftink	F	27.0	185.0	82.0	Netherlands	NED	1994 Winter	1994	Winter	Lillehammer	Speed Skating	Speed Skating Women's 500 metres	NaN

[5]:
```
region_data.head()
```

[5]:

	NOC	region	notes
0	AFG	Afghanistan	NaN
1	AHO	Curacao	Netherlands Antilles
2	ALB	Albania	NaN
3	ALG	Algeria	NaN
4	AND	Andorra	NaN

[6]:
```
athlete_data.info()
```

```
<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
```

[19]:
```
nan_count_per_column = athlete_data.isna().sum()
print("Number of NaN values in each column:")
print(nan_count_per_column)
```

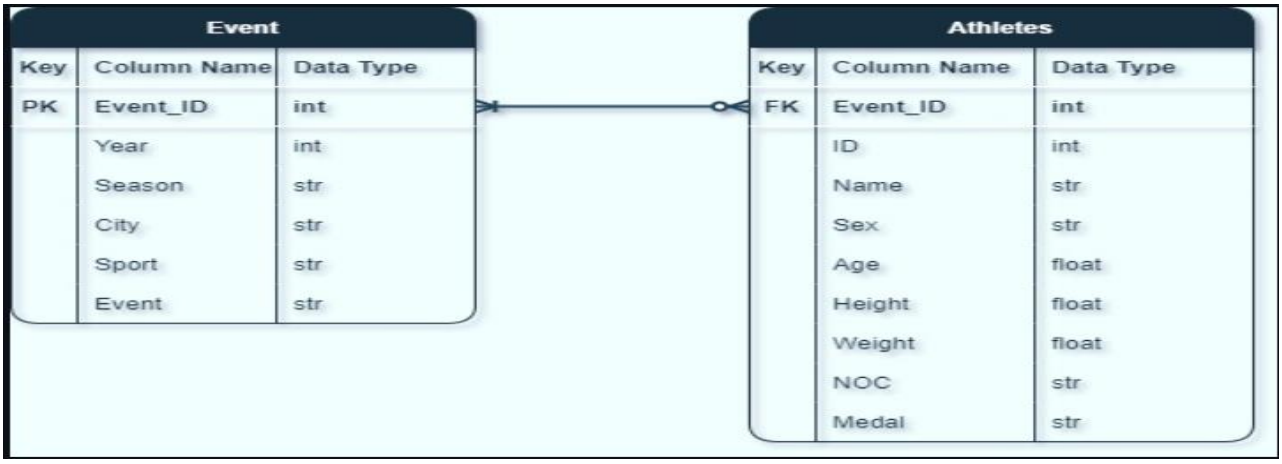
```
Number of NaN values in each column:
ID              0
Name            0
Sex             0
Age          9474
Height      60171
Weight      62875
Team            0
NOC             0
Games           0
Year            0
Season          0
City            0
Sport           0
Event           0
Medal      231333
dtype: int64
```

```python
# Merge the athletes' data with NOC data based on the 'NOC' column
merged_data = pd.merge(athlete_data, region_data, on='NOC', how='left')

# Count the number of medals per region
region_medal_counts = merged_data.groupby('region')['Medal'].count().reset_index()

# Sort the list by medal count in descending order
sorted_region_medal_list = region_medal_counts.sort_values(by='Medal', ascending=False).head(25)

# Print the list
for index, row in sorted_region_medal_list.iterrows():
    print(f"Region: {row['region']}, Medals: {row['Medal']}")
```

```
Region: USA, Medals: 5637
Region: Russia, Medals: 3947
Region: Germany, Medals: 3756
Region: UK, Medals: 2068
Region: France, Medals: 1777
Region: Italy, Medals: 1637
Region: Sweden, Medals: 1536
Region: Canada, Medals: 1352
Region: Australia, Medals: 1349
Region: Hungary, Medals: 1135
Region: Netherlands, Medals: 1040
Region: Norway, Medals: 1033
Region: China, Medals: 993
Region: Japan, Medals: 913
Region: Finland, Medals: 900
Region: Switzerland, Medals: 691
Region: Romania, Medals: 653
Region: Czech Republic, Medals: 644
Region: South Korea, Medals: 638
Region: Denmark, Medals: 597
Region: Poland, Medals: 565
Region: Serbia, Medals: 539
Region: Spain, Medals: 489
Region: Brazil, Medals: 475
Region: Belgium, Medals: 468
```

Create an ERD or proposed ERD to show the relationships of the data you are exploring.

The ERD shown below was intended for a small relational database, splitting them into two tables, the athletes, and the event. Some modifications have been needed, for example, the column 'ID' had no unique values, so it could not be used as a primary key (PK), so a new column "Event_ID" in the 'Event' table has been added as a PK, and as a FK in the 'Atheletes' Table.



Step 2: Develop a Project Proposal

This project aims to extract valuable insights from data pertaining to athletes across various Olympic events spanning the past 120 years. The intended audience for these insights includes sports enthusiasts, devoted followers of athletics, as well as coaches and trainers who can potentially benefit from this information. Moreover, this dataset could be of relevance to sports media outlets and channels of communication that cater to curiosity-driven audiences.

Questions:

1- To what extent does an athlete's age , height and weight influence their likelihood of winning a medal in each event?
2- Which countries, regardless of their resources, have a higher probability of securing?
3- medals in sports, both in terms of investment in early years and available resources?
4- What is the distribution of medals earned during a specific Olympic season, How has the gender balance in athlete participation evolved over the years, and has there been a trend towards greater gender equality in recent decades?

Hypotheses:

1- Countries situated at higher latitudes tend to perform better and win more medals in Winter Sports.
2- Over the years, there has been a trend towards achieving a more balanced participation of male and female athletes.
3- Developed countries are more likely to have a higher count of medals in their records.
4- The optimal age for winning medals tends to be around 25 years.

Approach:

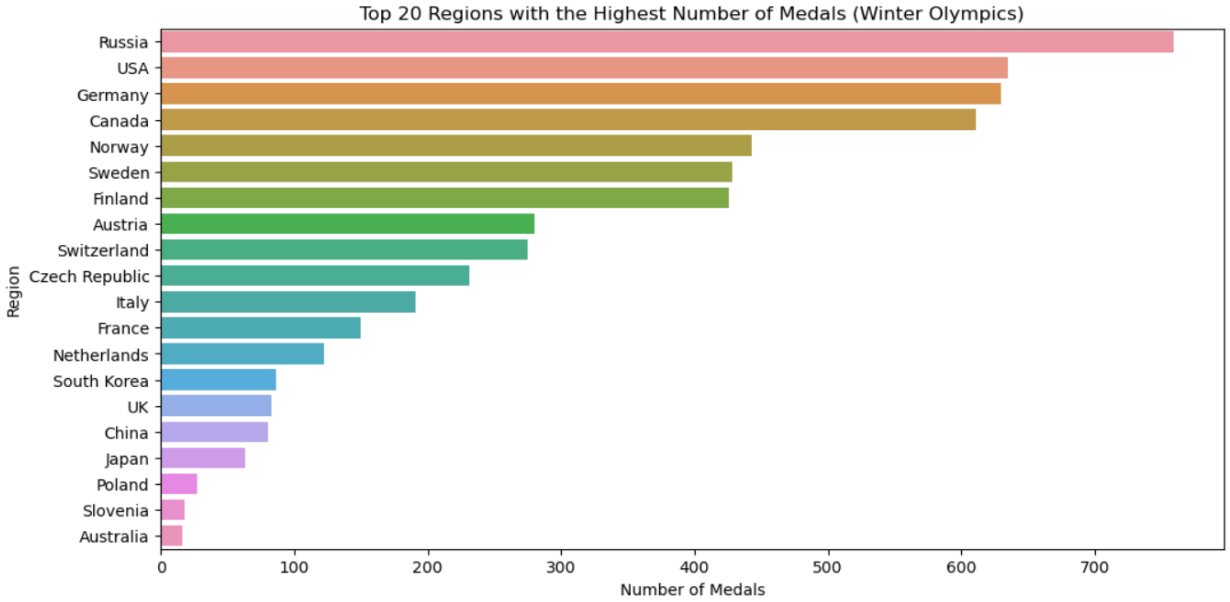
1- Analyze the distribution of athletes' ages , heights and weights and the corresponding medal counts.
2- Investigate the distribution of medals across different countries and assess any relationship with early investment in sports and available resources.
3- Examine the distribution of medals in various seasons and explore whether northern countries demonstrate a higher likelihood of success in Winter Seasons.
4- Evaluate the historical data to determine if the participation of male and female athletes has approached greater equality over the years.

Milestone 2: Descriptive Stats

1. Provide a summary of the different descriptive statistics you looked at and WHY.

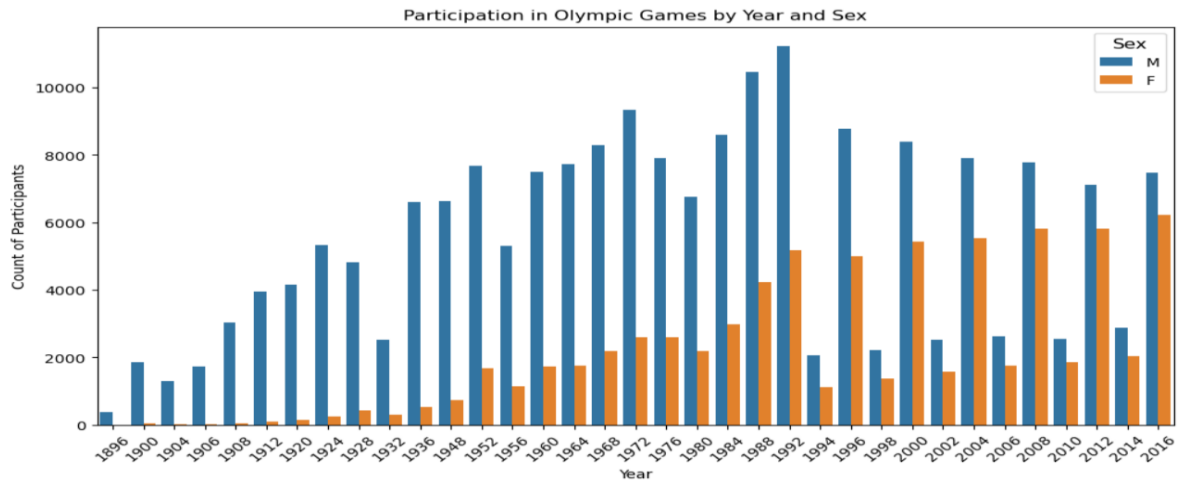
To achieve the business objectives, I would try to answer each of the hypotheses questions presented in Milestone 1. Even if the hypotheses are not true, it does not matter since it would help to get some insights from the data.

1-Countries at higher latitudes have better performance (medals) in Winter Sports.



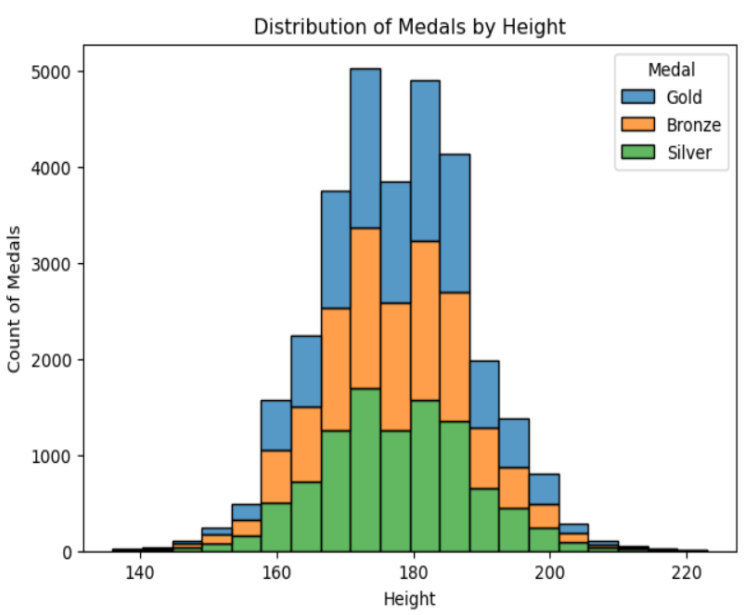
Countries located at higher latitudes tend to excel in Winter Sports, evident in the top 20 nations with the highest medal counts. These countries frequently experience prolonged and harsh winter conditions, which naturally fosters a greater familiarity and affinity for winter sports. Additionally, many of them allocate substantial investments towards the development and promotion of the Winter games. **Note: Russia medals here are combined with USSR medals.**

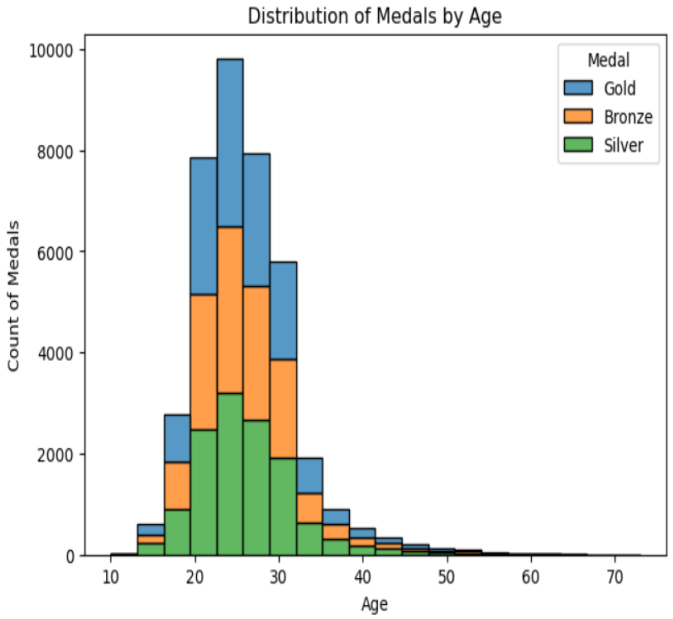
2-Female and Male participants tend to be equilibrated over the years.

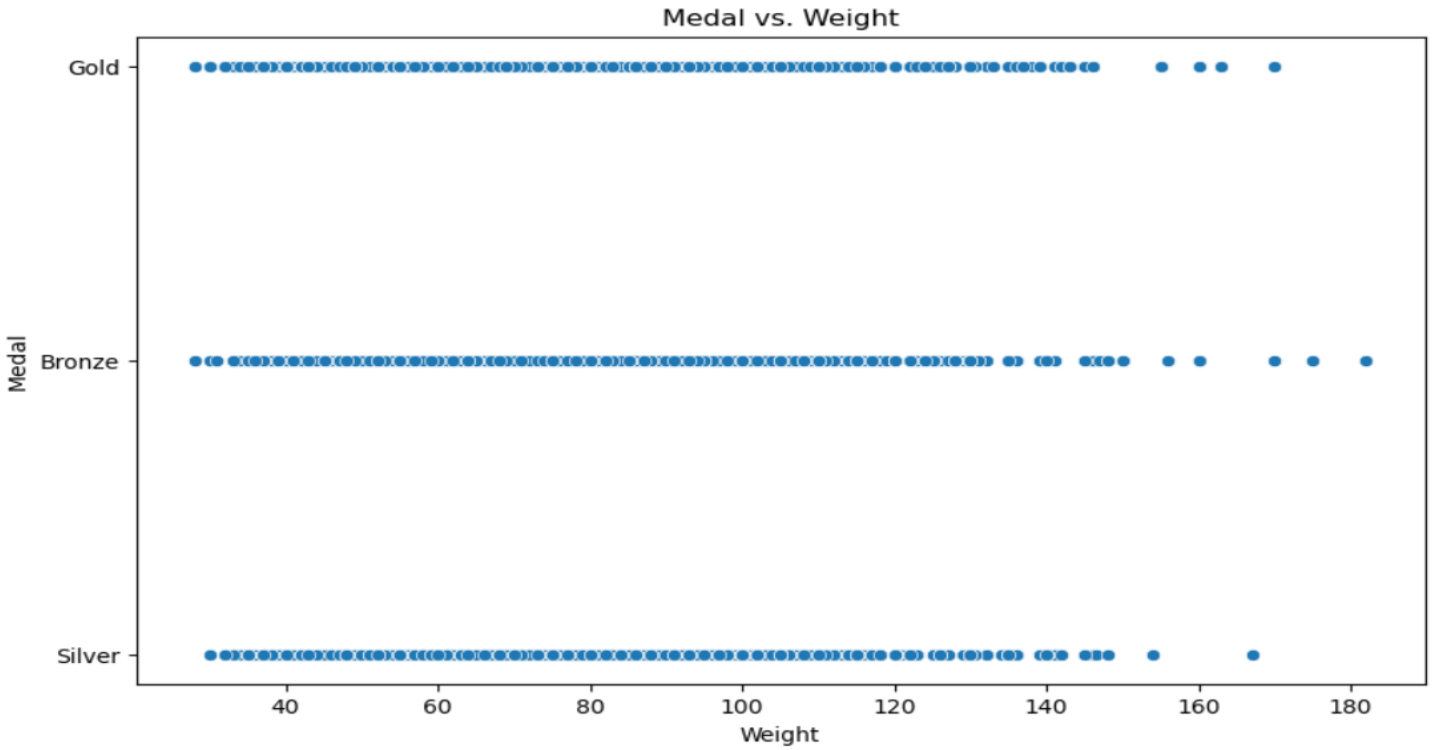


The plot illustrates a notable disparity in Olympic Games participation between males and females, particularly evident until around 1992. Over the past three decades, both genders have been on a similar trajectory, with a slight edge in participation still favoring males. However, it appears that this trend is poised to converge toward greater gender equality in the coming years, possibly within the next decade.

3- The optimal age for winning medals tends to be around 25 years.







Observing both genders, it becomes apparent that the optimal age for winning the most medals in the Olympics hovers around 25 years, give or take approximately five years.Additionally, athletes within the height range of 170 cm to 185 cm and with a weight below 160 kg tend to perform best in terms of winning medals in the Olympic Games.

1- Submit 2-3 key points you may have discovered about the data, e.g. new relationships? Aha's! Did you come up with additional ideas for other things to review?

- There is some drop in overall participation in the Olympics in some years, it could be interesting to find out why.
- Some countries do not exist anymore, or they have a new NOC, these medals should be added if the country has not changed geographically quite a lot.
- In the case of the USSR, the medal list country should be 'changed' to the actual country, but this could carry out some political issues.

2. Did you prove or disprove any of your initial hypotheses? If so, which one and what do you plan to do next?

I have confirmed that the initial hypotheses were largely accurate. Nonetheless, a more thorough examination of the data is required to gain deeper insights. Up to this point, the findings are somewhat superficial and should be regarded as a broad overview or general trend.

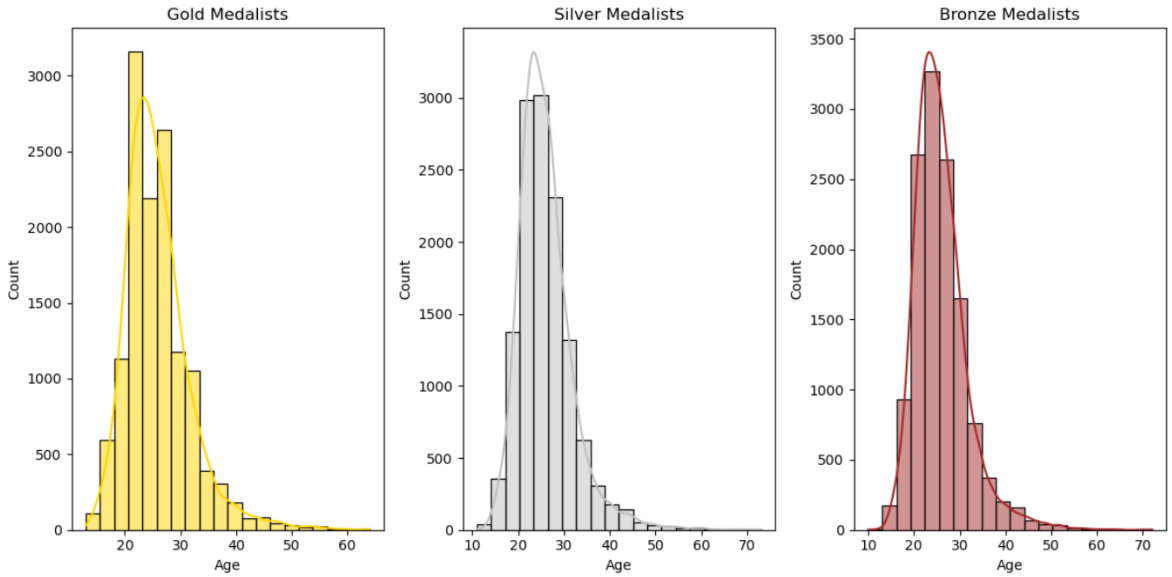
3. What additional questions are you seeking to answer?

1. The distribution of age and Medal type (Gold, Silver, Bronze)
2. The ratio between age and year by Medal type (e.g: if the average age of winning Silver is 30 years old, how has this evolved with the years)
3. Are there athletes who have won medals in multiple Olympics, and if so, what are their characteristics and achievements?
4. Does the host city or host country tend to have an advantage in terms of medal-winning during the Olympics they host?

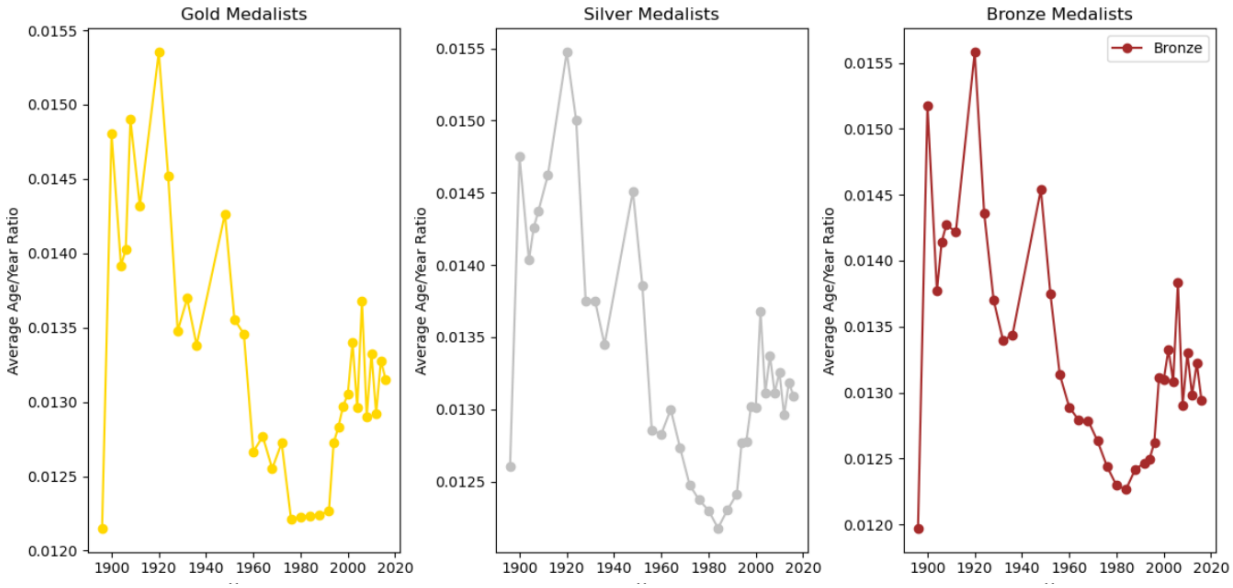
Milestone 3: Beyond Descriptive Stats

**Dive Deeper

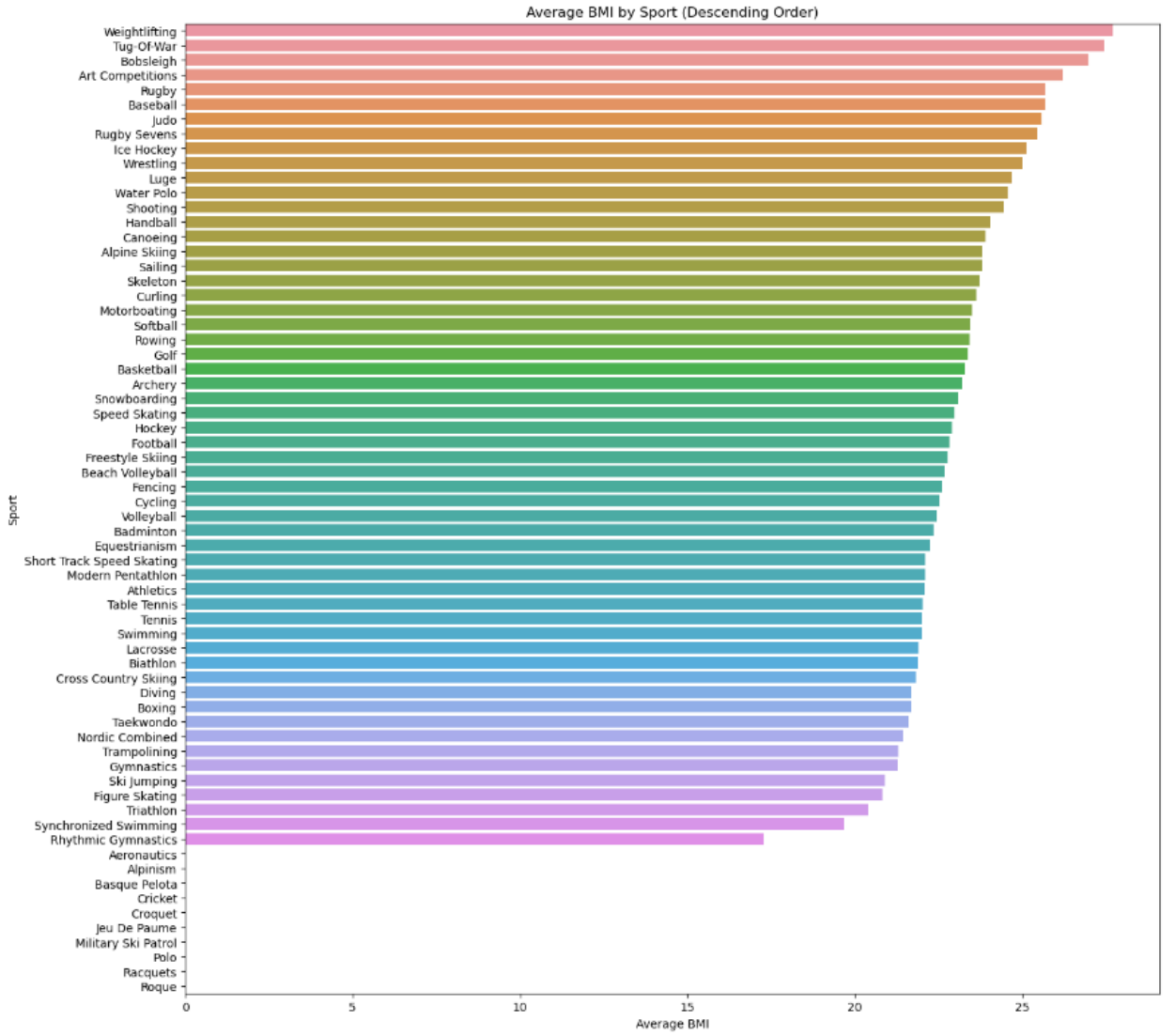
In this Milestone, I would look through Finding New metrics for my data form that form better understanding of the dataset.



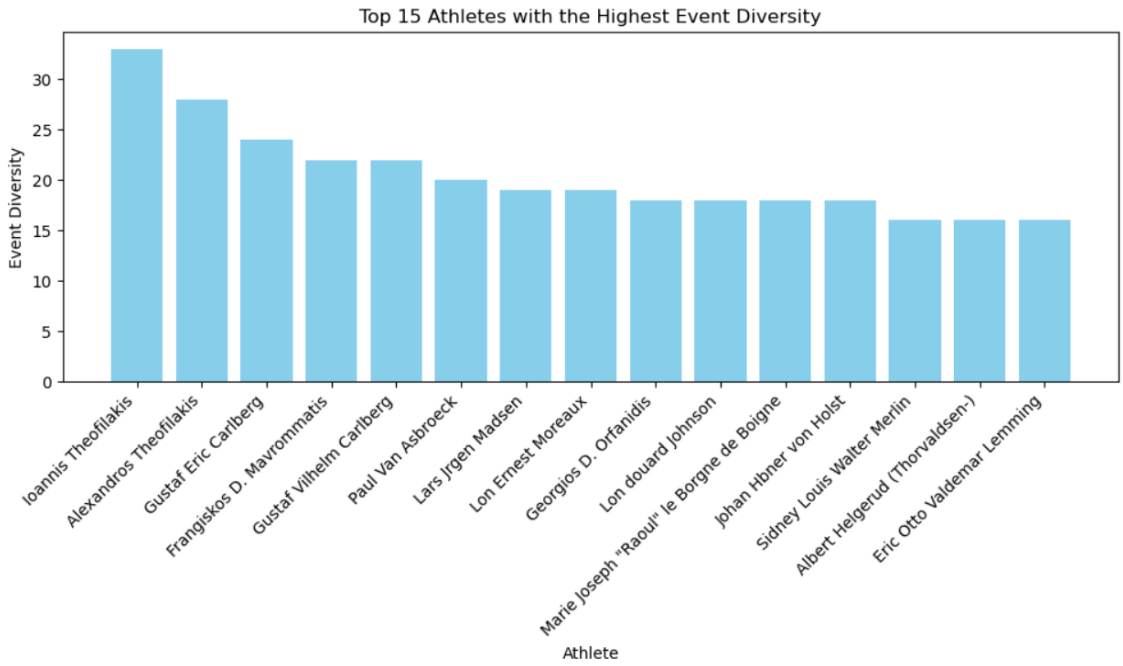
From the dataset, it appears that, on average, gold medal winners tend to be younger than both silver and bronze medalists. Additionally, the average age of silver medalists is lower than the average age of bronze medalists.



Here we Age/Year Ratio is a metric that can provide insights into the relationship between an athlete's age and the year of participation in a particular event or competition we could that in years between 1970 and 1990 that all medalists were younger than average age of participants in those years in 1920 all medalist were older average age of participants in that year.



Let's examine which is the average BMI for each sport for the entries that we have enough data. We can see that the sports with the largest BMI are Weightlifting and Tug-Of-War, while Rhythmic Gymnastics has the lowest BMI, followed by Triathlon. Surprisingly, Art Competitions is the 3rd sport with the largest BMI, above Ice Hockey or Rugby.



We could the who Athletes who joined most event in all Olympic history which is symbolled by the new metric Event Diversity.

New metrics:

1. BMI (Body Mass Index): A measure indicating an athlete's suitability for specific sports based on their body size and weight.

2. Medal Ratio: An indicator of the distribution of medals, revealing the competitive balance within a sport or event.

3. Event Diversity: Quantifies the range of different sports disciplines within a competition, showcasing athletic variety.

4. Age/Year Metric: Reflects the correlation between an athlete's age and the year of their sports participation, highlighting age-related trends over time.

Milestone 4: Delivering Your Insights (Storytelling)

For the presentation of my findings, I've opted for the Capstone Project Jupyter notebook saved as a .pdf format. My target audience consists of budding data scientists looking to both learn and collaborate on their discoveries, as well as anyone interested in exploring the dataset's general statistics.

The rationale behind choosing this format is rooted in transparency and interactivity. By sharing the Jupyter notebook in PDF form, the audience can delve into the Python and pandas code directly, allowing them to validate and gain a deeper understanding of the analytical process.

Within the notebook, I've comprehensively detailed the hypothesis, findings, and key insights derived from the SportsStats dataset. To enhance comprehension and engagement, I've complemented these explanations with relevant visualizations.

In structuring my presentation, I've adhered to the Milestones outlined in the Capstone Course. This ensures a logical and organized flow, facilitating a seamless journey through the dataset's exploration and analysis.