Project Proposal

Data Science Project Proposal for SportStats
2023-06-10
Gurpreet Singh

Which client/dataset did you select and why?

I selected the SportStats dataset because, as an athlete myself, I have a profound knowledge and passion for various sports. This personal experience allows me to thoroughly analyze the data within the dataset and derive answers to a wide range of questions. With my expertise and the comprehensive SportStats dataset, I aim to uncover valuable insights that would be of great interest to major sports media outlets and enthusiasts. By leveraging this dataset, I am confident in my ability to provide meaningful analyses that can contribute to the understanding and advancement of the sporting community as a whole.

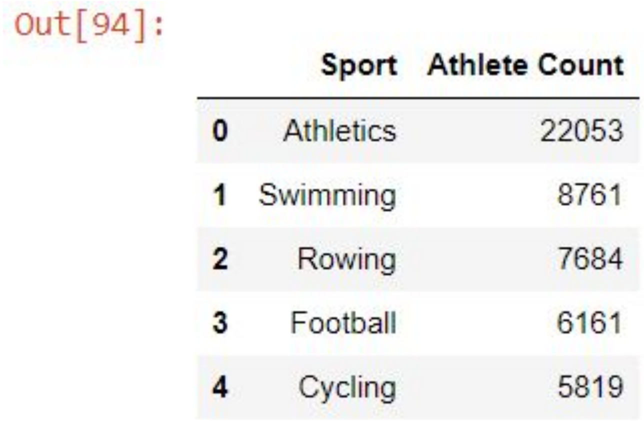
Describe the steps you took to import and clean the data

Steps I took to import are:

1. Download the data from the SportStats website
2. Use the pandas library to import the csv data into Jupyter notebook
3. Used the built in library called pandasql to analyze the dataset
4. Used built in Matplotlib library to visualize the data
5. Lastly, I did not clean the data yet because the dataset contains NaN values. Hence the data can be tested or falsified by others.

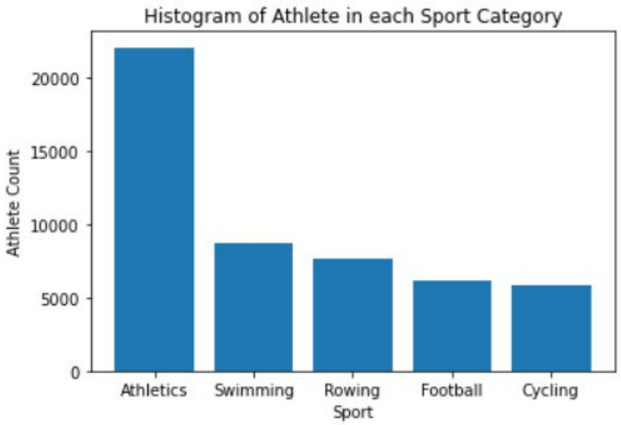
Perform initial exploration of data and provide some screenshots or display some stats of the data you are looking at

This table explores the total count of athletes in each sport categories. (Result are limited to 5 since the data was too large)



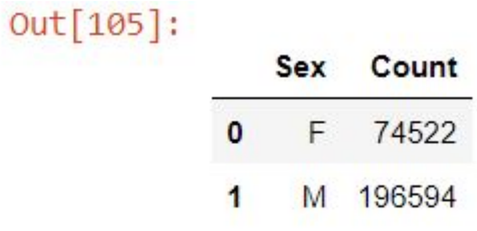
Out[94]:

	Sport	Athlete Count
0	Athletics	22053
1	Swimming	8761
2	Rowing	7684
3	Football	6161
4	Cycling	5819



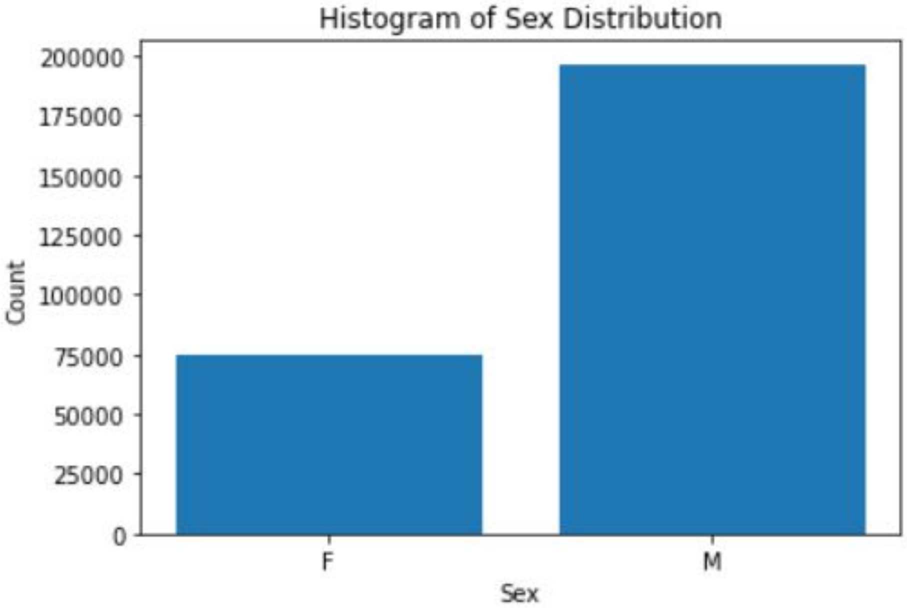
Continued

This table explores the sex distributions of athletes in all sports

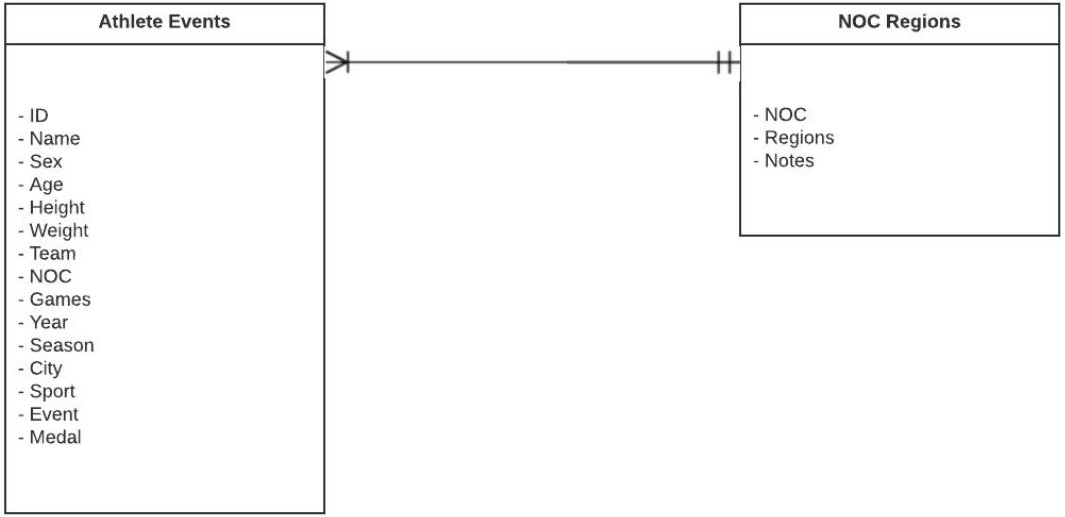


Out[105]:

	Sex	Count
0	F	74522
1	M	196594



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



Athlete Events

- ID
- Name
- Sex
- Age
- Height
- Weight
- Team
- NOC
- Games
- Year
- Season
- City
- Sport
- Event
- Medal

NOC Regions

- NOC
- Regions
- Notes

Description

SportsStats is a sports analysis firm partnering with local news and elite personal trainers to provide “interesting” insights to help their partners. Insights could be patterns/trends highlighting certain groups/events/countries, etc. for the purpose of developing a news story or discovering key health insights.

Questions

- What sport categories gave out the most medals ?
- Who received the most medals in each sport categories ?
- What country produces the most athletes (male and female) ?

Hypothesis

- I believe that football gave out the most medals since is one of the most played sports in the world
- I believe that Lebron James received the most medals in Basketball, Lionel Messi received the most medals in Football, Michael Phelps received the most medals in swimming (Bias: I believe that these are the best players in each sport category !)
- I believe that China produces the most athletes because it has the largest population.

Approach

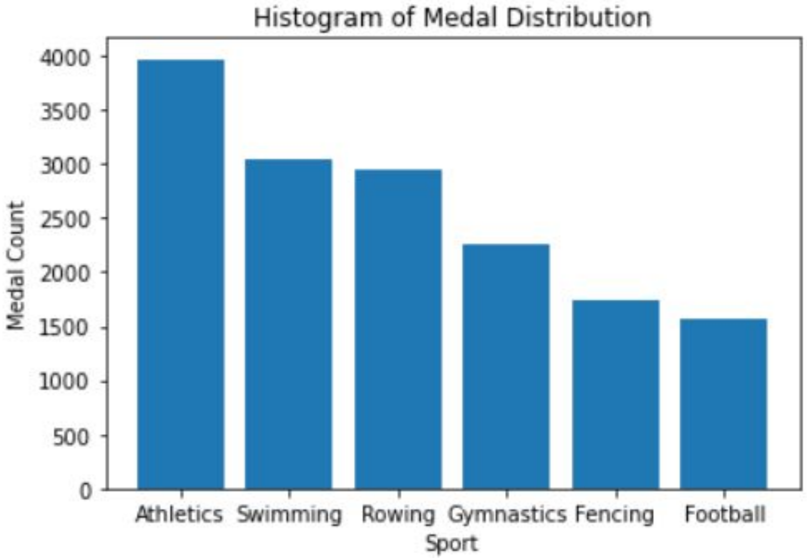
- Initially, I will focus on analyzing key features such as event type, location, and participant demographics.
- I will explore the relationships between these features and identify any correlations or trends.
- To evaluate my hypotheses, I will use appropriate metrics and evaluation measures, such as correlation coefficients, regression analysis, or statistical significance tests.
- By following this approach, we aim to provide valuable insights into sports events, their patterns, and their connections to health indicators. Our analysis will benefit news outlets and personal trainers, enabling them to make informed decisions and enhance their understanding of the sports landscape

Initial Findings

1. Swimming has awarded the most medals because it offers a wide range of individual events and multiple opportunities for athletes to compete
2. Teresa Edwards received the most medals in Basketball and Christie Patricia Pearce-Rampone in soccer. Michael Phelps received the most medals in swimming as predicted.
3. USA produced the most athletes.

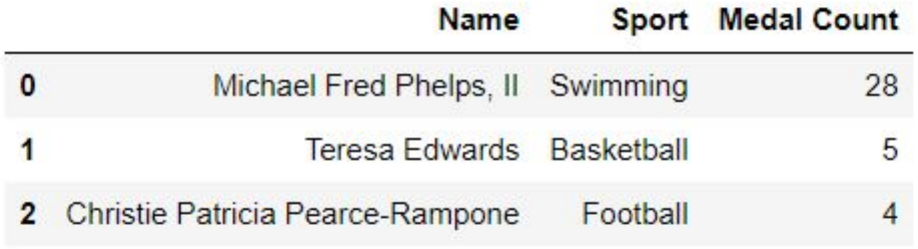
Initial Findings (1)

Swimming tends to award more medals than football due to the greater diversity of events and categories within swimming competitions. The wide range of swimming disciplines allows for a larger number of medal opportunities. In contrast, football typically offers fewer medal events, such as the FIFA World Cup or Olympic tournaments, which are limited in number.



Initial Findings (2)

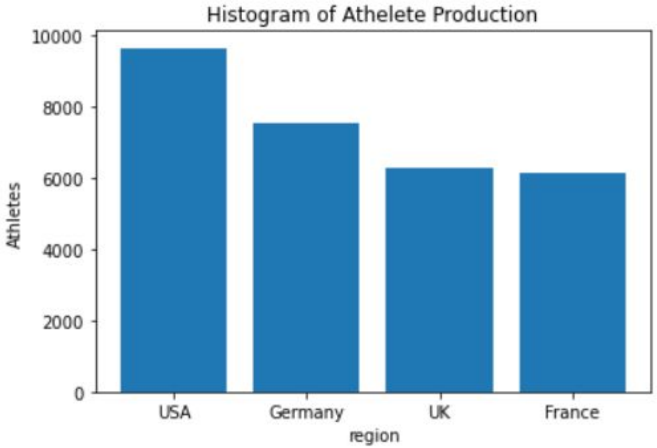
Teresa Edwards may have more basketball medals than LeBron James, despite his status as the best player in the world, due to the higher frequency of basketball tournaments and the potential for multiple medal opportunities. Similarly, Christie Patricia Pearce-Rampone could have more medals than Lionel Messi in football due to the greater number of international tournaments available for female players and the varying competitiveness of the tournaments they participated in.



	Name	Sport	Medal Count
0	Michael Fred Phelps, II	Swimming	28
1	Teresa Edwards	Basketball	5
2	Christie Patricia Pearce-Rampone	Football	4

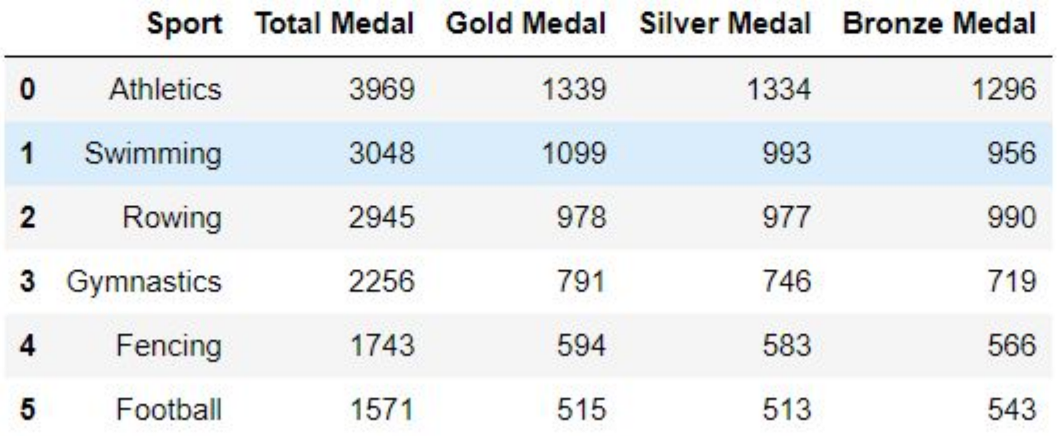
Initial Findings (3)

I initially believed that China, with its larger population, would produce more athletes than the USA. However, upon considering various factors such as differences in sporting culture, investment in sports infrastructure, development programs, and access to resources, it becomes clear that the USA has been able to produce a larger number of athletes. The USA's long-standing tradition of sports participation, coupled with significant investments in athletic development, has contributed to their success in cultivating a larger pool of athletes across a wide range of sports.



Deeper Analysis (1)

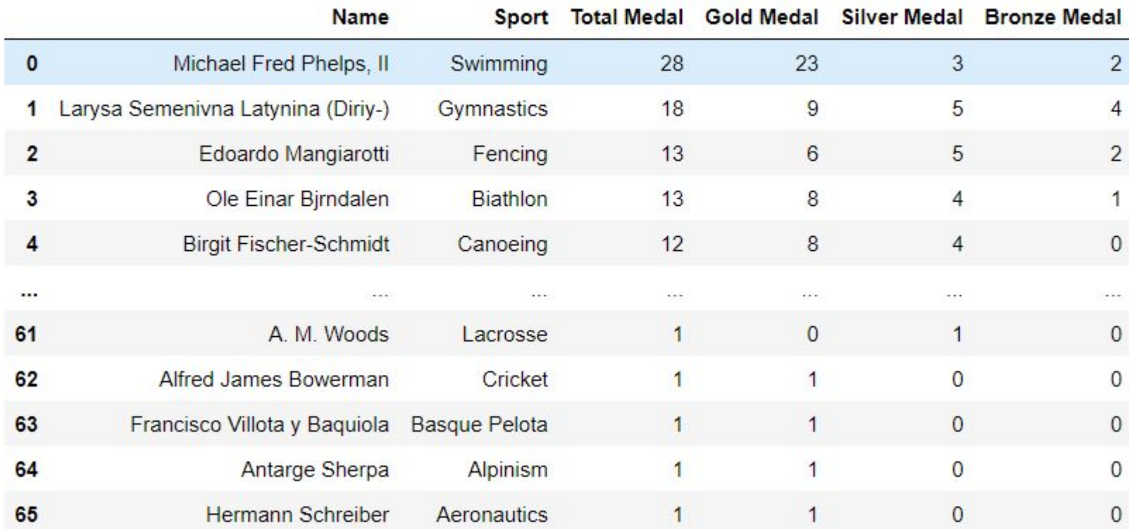
In my further analysis, I delved into the types of medals (Gold, Silver, Bronze) distributed in each sport category. This information can be highly useful in understanding the distribution of success and performance within different sports. By examining which sports tend to have a higher proportion of gold, silver, or bronze medals, we can gain insights into the level of competition, dominance of certain athletes or countries in specific sports, and the overall competitiveness of different sporting disciplines. This knowledge can guide strategic decisions, funding allocations, and resource distribution for athletes, coaches, and sporting organizations to maximize performance and focus on areas that offer the greatest potential for success.



	Sport	Total Medal	Gold Medal	Silver Medal	Bronze Medal
0	Athletics	3969	1339	1334	1296
1	Swimming	3048	1099	993	956
2	Rowing	2945	978	977	990
3	Gymnastics	2256	791	746	719
4	Fencing	1743	594	583	566
5	Football	1571	515	513	543

Deeper Analysis (2)

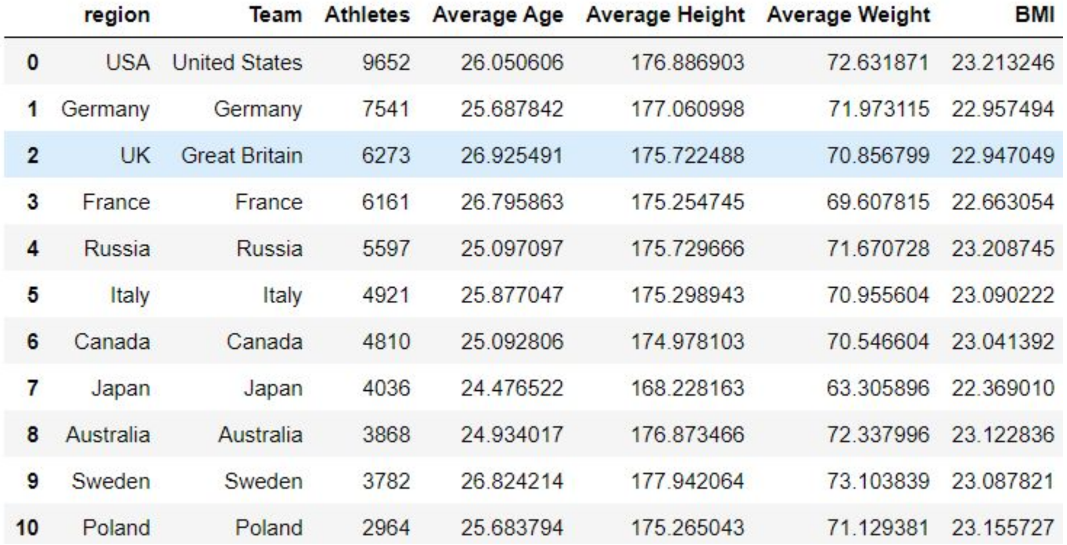
In my deeper analysis, I delved into the types of medals (Gold, Silver, Bronze) received by the highest award-winning athletes from each sport category. This information can be valuable as it provides insights into the distribution of medals among athletes in different sports. It allows for comparisons between sports in terms of their medal-winning patterns and can help identify sports where athletes tend to achieve higher levels of success, indicated by a higher number of gold medals. This analysis can be beneficial for sports organizations, researchers, and enthusiasts in understanding the dynamics and performance levels across various sports and their respective medal distributions.



	Name	Sport	Total Medal	Gold Medal	Silver Medal	Bronze Medal
0	Michael Fred Phelps, II	Swimming	28	23	3	2
1	Larysa Semenivna Latynina (Diriy-)	Gymnastics	18	9	5	4
2	Edoardo Mangiarotti	Fencing	13	6	5	2
3	Ole Einar Bjrndalen	Biathlon	13	8	4	1
4	Birgit Fischer-Schmidt	Canoeing	12	8	4	0
...	...	...	...	...	...	...
61	A. M. Woods	Lacrosse	1	0	1	0
62	Alfred James Bowerman	Cricket	1	1	0	0
63	Francisco Villota y Baquiola	Basque Pelota	1	1	0	0
64	Antarge Sherpa	Alpinism	1	1	0	0
65	Hermann Schreiber	Aeronautics	1	1	0	0

Deeper Analysis (3)

In my deeper analysis, I found the average height, weight, age, and BMI for athletes in each region. This information can be useful in several ways. It provides insights into the physical characteristics of athletes from different regions, allowing for comparisons and identification of any potential trends or patterns. Additionally, this data can be used by sports organizations, trainers, and coaches to understand the typical physique and age range of athletes in various regions, aiding in talent identification, team selection, and training program development. Notably, most athletes across regions had a BMI between 22-23, average age between 25-26 years old, average weight between 63-74 kg, and average height between 168 to 178 cm.



	region	Team	Athletes	Average Age	Average Height	Average Weight	BMI
0	USA	United States	9652	26.050606	176.886903	72.631871	23.213246
1	Germany	Germany	7541	25.687842	177.060998	71.973115	22.957494
2	UK	Great Britain	6273	26.925491	175.722488	70.856799	22.947049
3	France	France	6161	26.795863	175.254745	69.607815	22.663054
4	Russia	Russia	5597	25.097097	175.729666	71.670728	23.208745
5	Italy	Italy	4921	25.877047	175.298943	70.955604	23.090222
6	Canada	Canada	4810	25.092806	174.978103	70.546604	23.041392
7	Japan	Japan	4036	24.476522	168.228163	63.305896	22.369010
8	Australia	Australia	3868	24.934017	176.873466	72.337996	23.122836
9	Sweden	Sweden	3782	26.824214	177.942064	73.103839	23.087821
10	Poland	Poland	2964	25.683794	175.265043	71.129381	23.155727

Recommendations and Actions

- Sport committees to encourage more females to join sports and address the gender disparity.
- Tailor region-specific training programs considering the average height, weight, age, and BMI of athletes in each region. This customization can optimize performance by adjusting training techniques, nutrition plans, and fitness regimens accordingly. For instance, addressing slight underweight concerns observed in Japanese athletes through targeted interventions.