Project Proposal

Data Analysis for SportsStats

By: Kay Royo

I. Project Proposal Preparation

I.I Client and Dataset

This project aims to provide an extensive data analysis for the client, SportsStats. The Olympics Games dataset, which contains 120 years of records, will be analyzed to provide useful insights to the client's partners, local news and personal trainers. Specifically, this dataset will be used to identify patterns/trends that highlight certain groups, events, countries, and more for the purpose of developing a news story or discovering key health insights.

I.II Importing and Cleaning Data

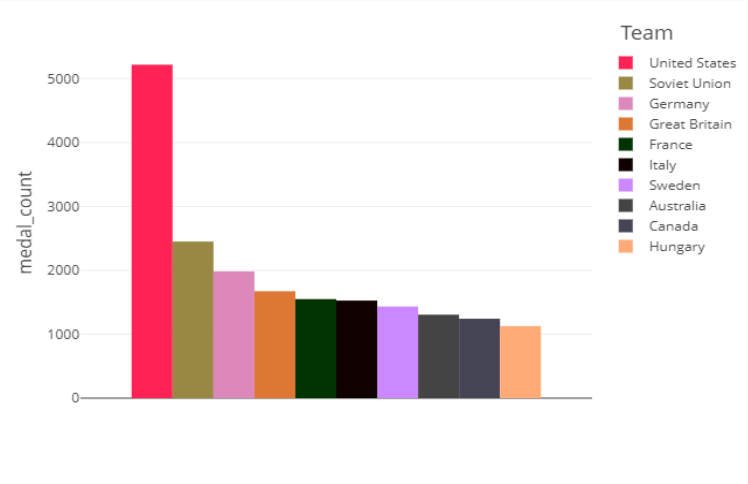
A medallion architecture (bronze, silver, gold) with Delta Lake is utilized for the datasets. The following specific tasks are performed while importing and cleaning the data:

1. Load original datasets in csv file format into the databricks environment under the Data tab and create them as tables with UI, instead of creating using a notebook, where the infer schema option is used.
2. Check the total number of records and data types in each table
3. Remove unwanted columns from the *host_countries* table and write it to Data Silver.
4. Join three tables (*athletic_events*, *noc_regions,* and *host_countries*) using the feature NOC and City.
5. Check the number of null values in the final aggregated table and make sure there are no missing values in Country.
6. Replace null values in the Region column.
7. Change literal 'NA' strings to null values in Medal.

I.III Data Exploration

The following summary results are obtained from the exploratory data analysis task performed:

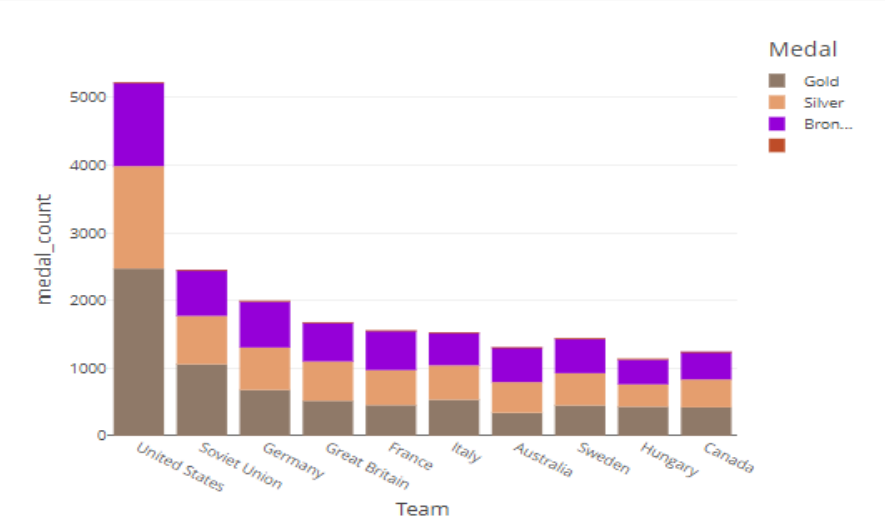
1. Top 10 Teams with the highest number of Medals



	Team	medal_count
1	United States	5219
2	Soviet Union	2451
3	Germany	1984
4	Great Britain	1673
5	France	1550
6	Italy	1527
7	Sweden	1434
8	Australia	1306
9	Canada	1243
10	Hungary	1127

Truncated results, showing first 1000 rows.

2. Distribution of medals within each team in the top 10 list:



3. Number of distinct values in each column

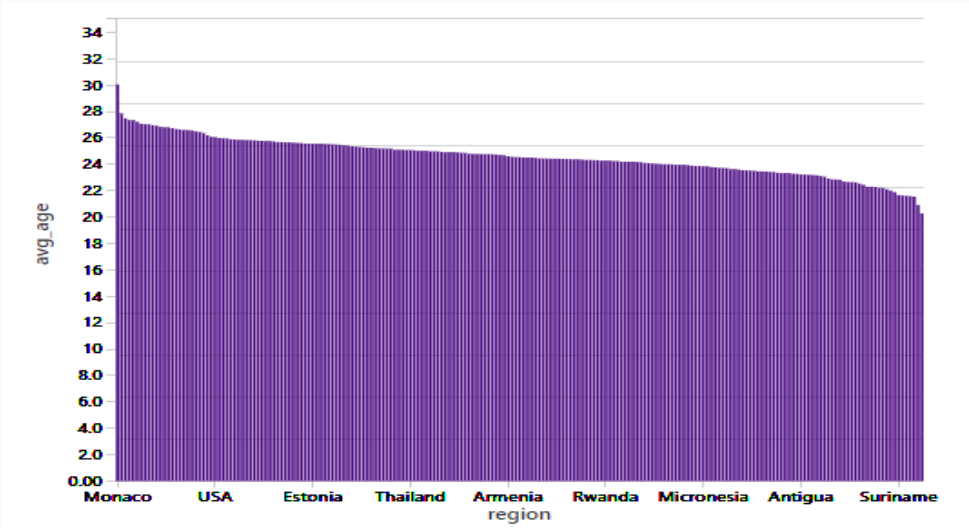
	IDs	Names	Sex	Teams	NOCs	Regions	Sports	Events	Games	Cities	Countries	Years	Season	Medals
1	135571	134732	2	1184	230	207	66	765	51	42	23	35	2	3

4. Average age, height, weight of players in each region

	region	avg_age
1	Monaco	30.047120418848166
2	Namibia	27.857142857142858
3	Ireland	27.473727422003282
4	Denmark	27.351487928130265
5	Virgin Islands, US	27.33676975945017
6	American Samoa	27.216216216216218
7	Montenegro	27.074468085106382
8	Portugal	27.03942895989123
9	Belgium	27.03159851301115
10	Austria	26.95985547972702

	region	avg_age
197	Sao Tome and Principe	22.2
198	North Korea	22.099378881987576
199	Somalia	22
200	Niger	21.892857142857142
201	Suriname	21.671641791044777
202	Kiribati	21.636363636363637
203	Yemen	21.604166666666668
204	Laos	21.574074074074073
205	Saint Vincent	21.547619047619047
206	Maldives	20.918367346938776
207	Marshall Islands	20.285714285714285

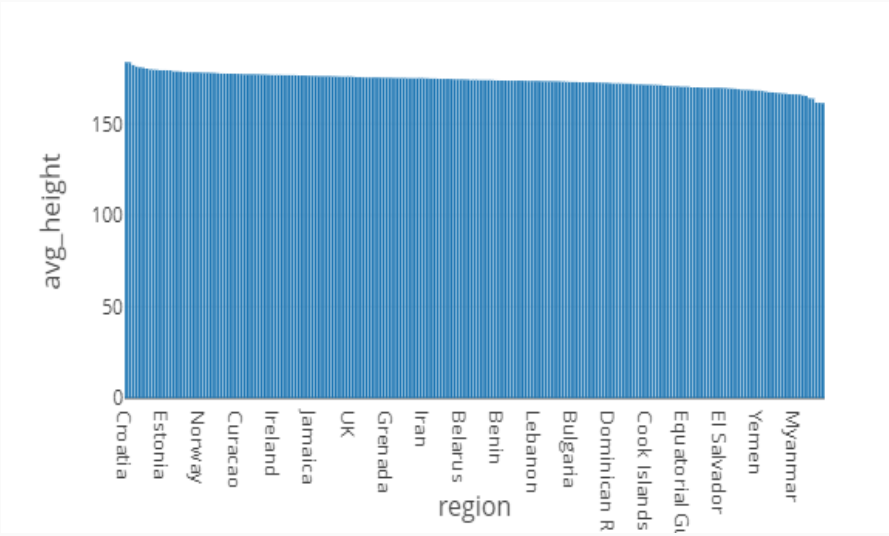
Showing all 207 rows.



	region	avg_height
1	Croatia	183.6131736526946
2	Montenegro	183.56382978723406
3	Lithuania	182.09950248756218
4	Iceland	181.2454128440367
5	Serbia	180.80417198808004
6	Mali	180.60655737704917
7	Senegal	180.17888563049854
8	Latvia	179.64450127877237
9	Cayman Islands	179.5068493150685
10	Netherlands	179.48890649762282

	region	avg_height
197	Bangladesh	166.1875
198	Brunei	166.11111111111111
199	Myanmar	166.10638297872342
200	Malawi	165.95
201	Palau	165.45833333333334
202	Vietnam	165.2663043478261
203	Nepal	163.90588235294118
204	Maldives	163.76190476190476
205	North Korea	161.5579598145286
206	Timor-Leste	161.44444444444446
207	Micronesia	161.24

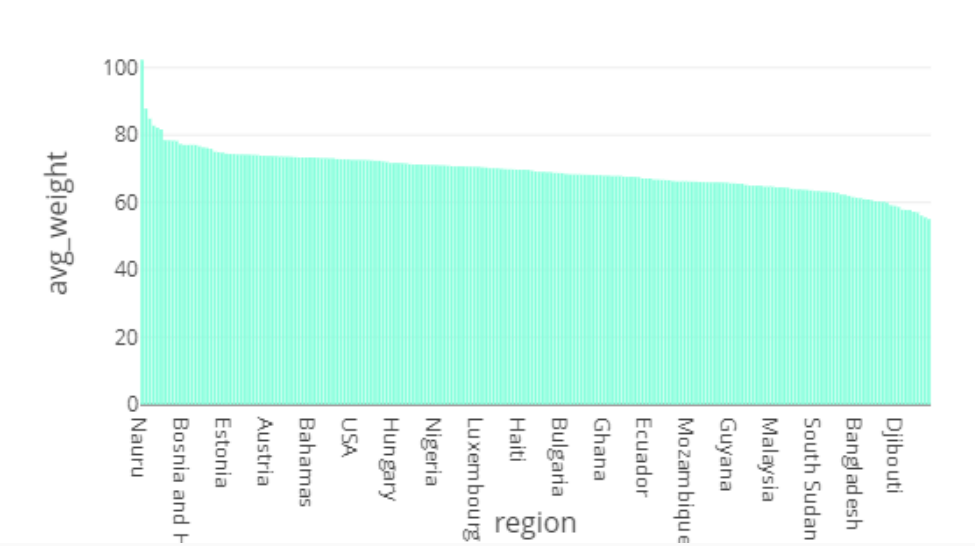
Showing all 207 rows.



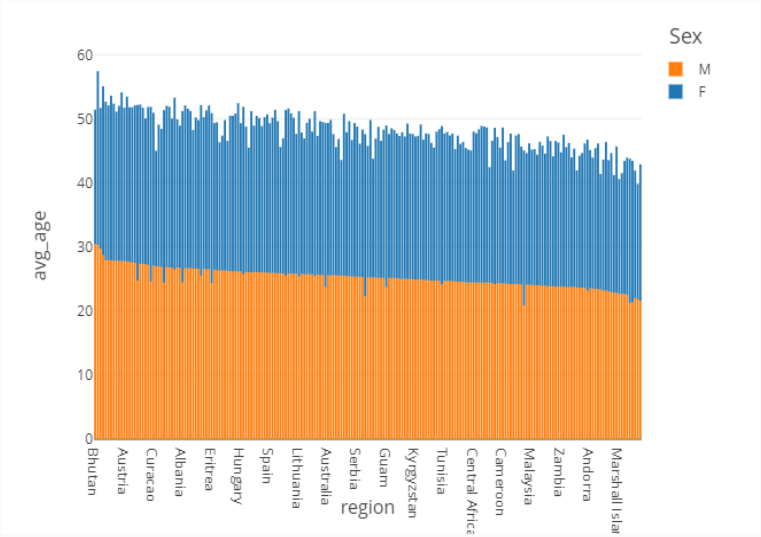
	region	avg_weight
1	Nauru	102.27272727272727
2	American Samoa	87.81818181818181
3	Montenegro	84.82978723404256
4	Tonga	82.75555555555556
5	Samoa	82.12244897959184
6	Croatia	81.62634730538922
7	Iceland	78.3892773892774
8	Kiribati	78.36363636363636
9	Latvia	78.26913265306122
10	Lithuania	78.18013468013469

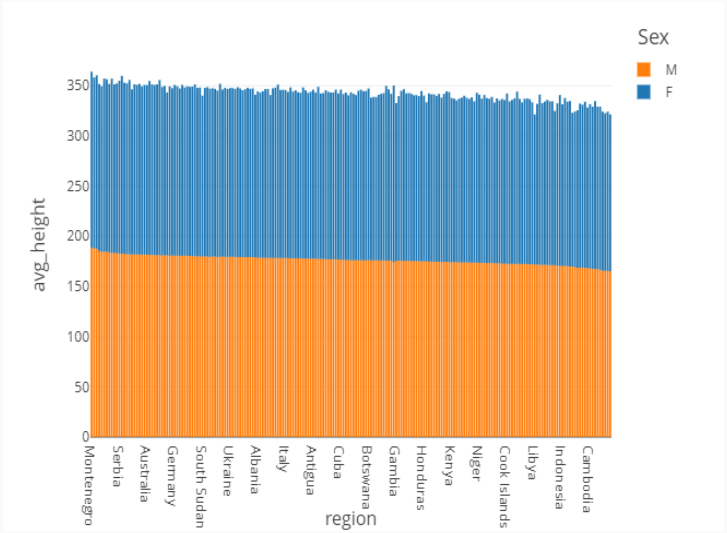
	region	avg_weight
197	Burundi	59.07692307692308
198	Myanmar	58.93814432989691
199	Djibouti	58.58064516129032
200	Laos	57.80392156862745
201	Vietnam	57.73513513513514
202	Nepal	57.666666666666664
203	North Korea	57.18055555555556
204	Ethiopia	56.940677966101696
205	Eritrea	56.054054054054056
206	Timor-Leste	55.5
207	Maldives	54.976190476190474

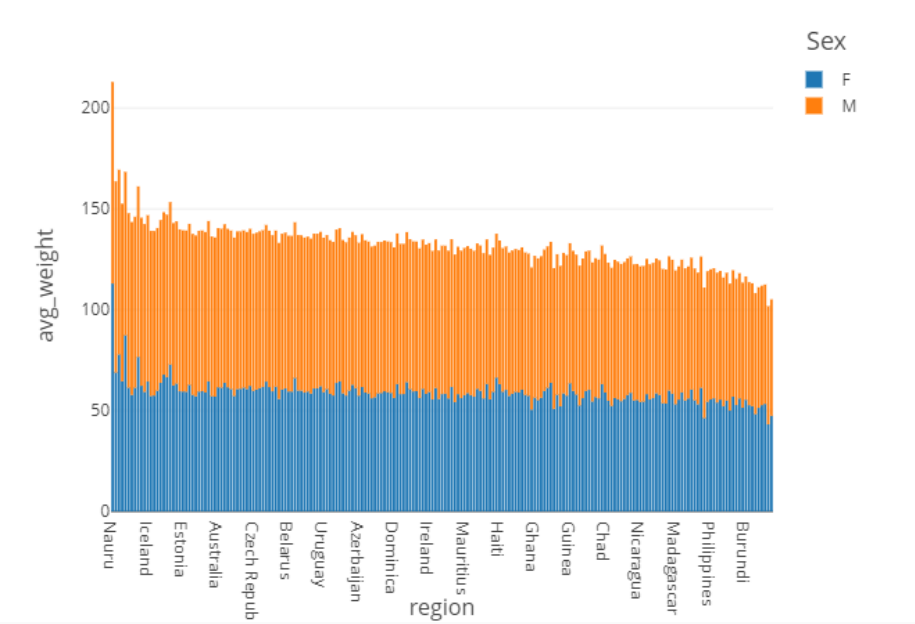
Showing all 207 rows.



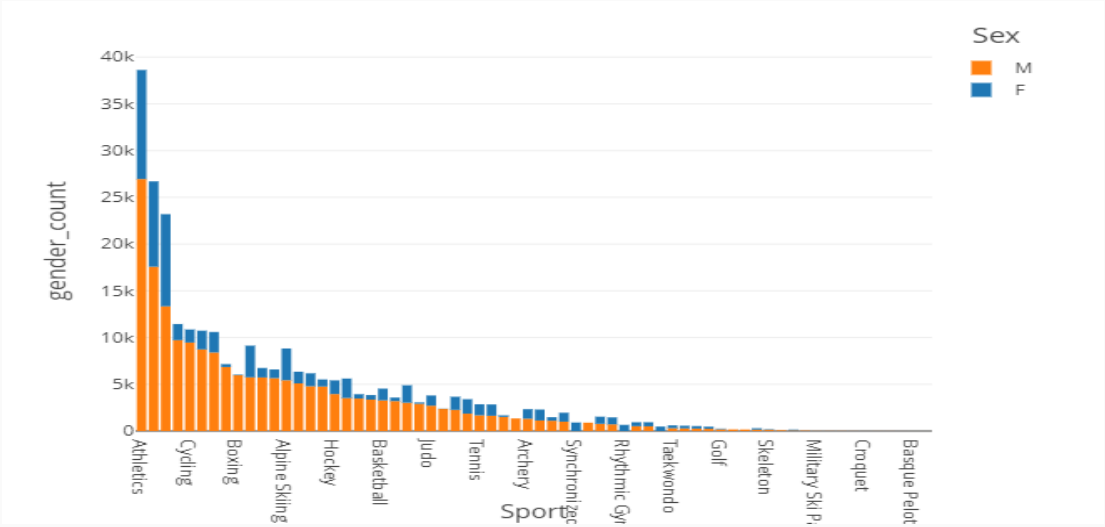
5. Average age, height, weight of men and women players in each region



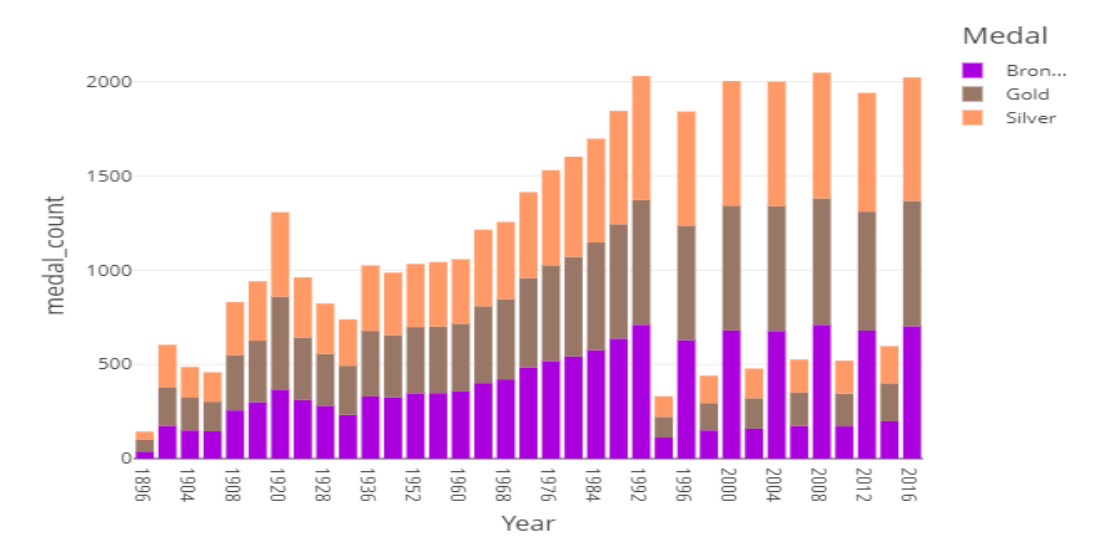




6. Gender distribution in sports



7. Number of medals per year

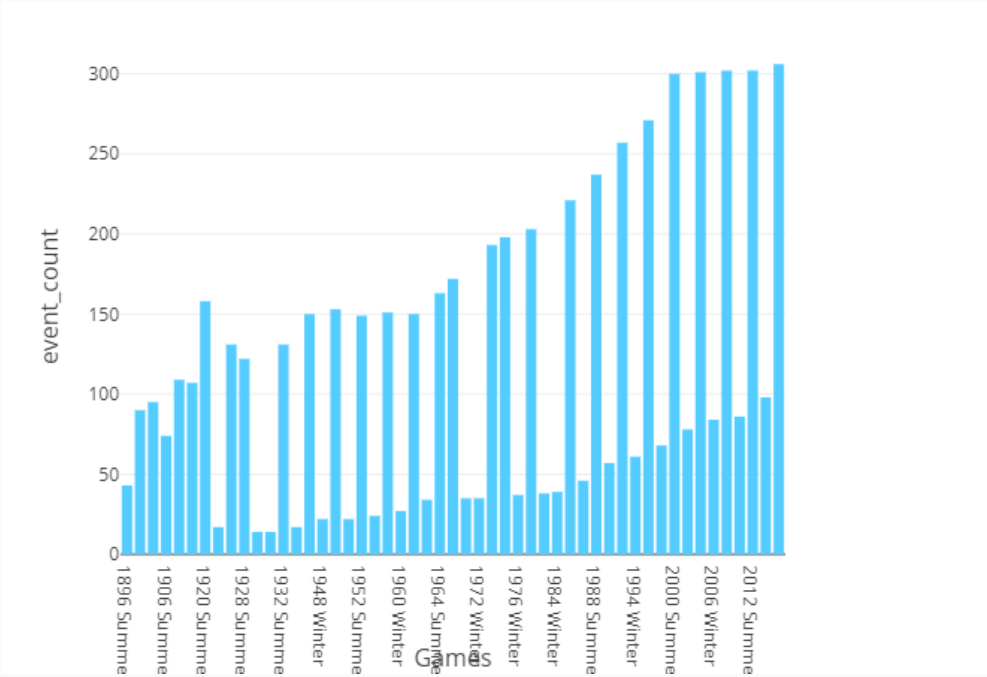


8. Top players (fix query)

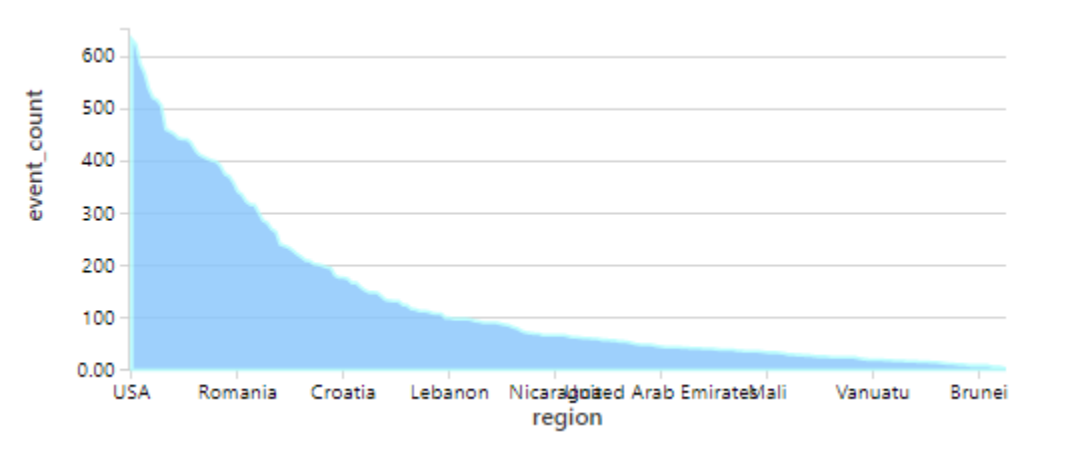
	Name	Medal	medal_count
1	Ole Einar Bjrndalen	Gold	8
2	Stefania Belmondo	Bronze	5
3	Thomas Alsgaard (Alsgrd-)	Gold	5
4	Ivar Eugen Ballangrud (Eriksen-)	Gold	4
5	Francina Elsje "Fanny" Blankers-Koen	Gold	4
6	Kjetil Andr Aamodt	Gold	4
7	David Cal Figueroa	Silver	4
8	Vra slavsk (-Odloilov)	Silver	4
9	Sergey Aleksandrovich Belov	Bronze	3
10	Regla Maritza Bell McKenzie	Gold	3
11	Ethelda Marguerite Bleibtrey (-Schlatke)	Gold	3
12	Albert Azaryan	Gold	3
13	Laura Kay Berg	Gold	3
14	Denis Mikhaylovich Ablyazin	Silver	3
15	George Thomas Breen	Bronze	3
16	Sebastian Brendel	Gold	3
17	Grgory Benot Baug	Silver	3
18	G. Alberto Braglia	Gold	3
19	Niccol Campriani	Gold	3
20	Brooke Marie Bennett (-Frioud)	Gold	3

Truncated results, showing first 1000 rows.

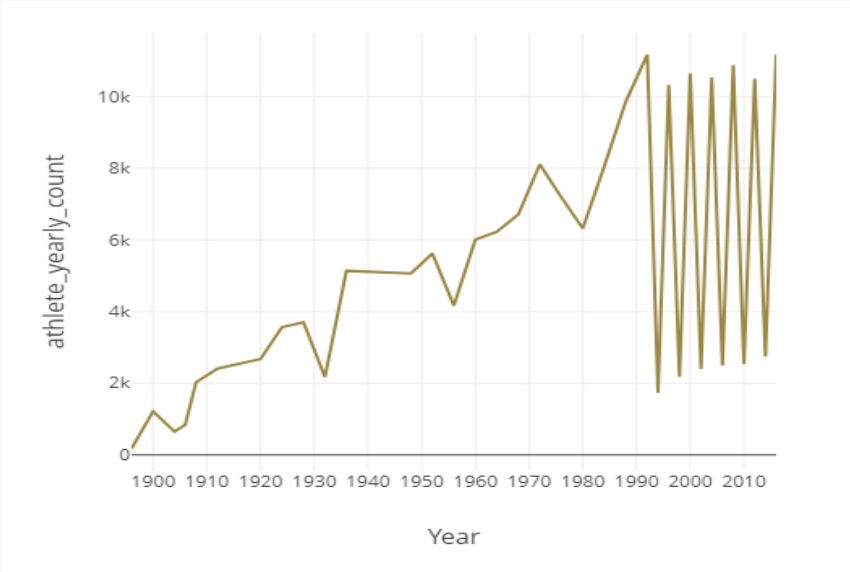
9. Number of events per game

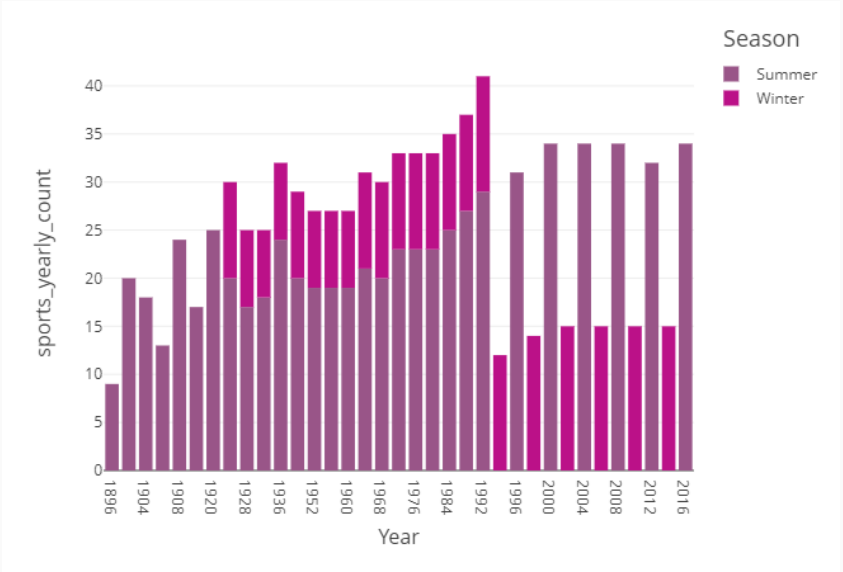


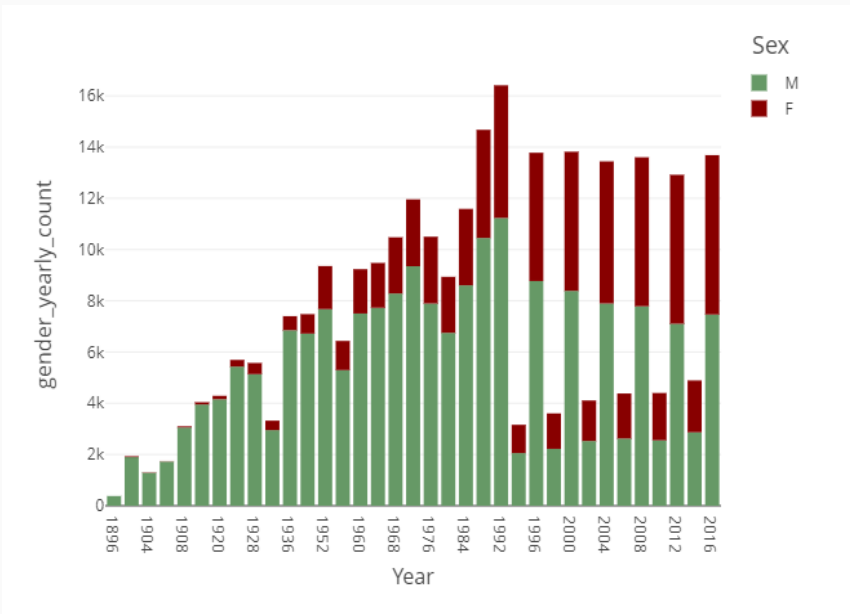
10. Number of events per NOC region

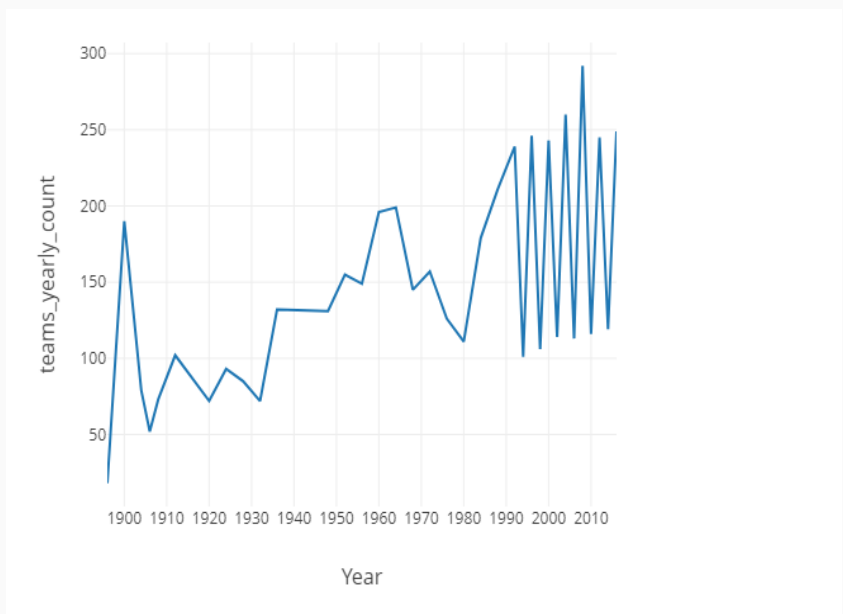


11. Number of athletes, sports, gender, and teams every year

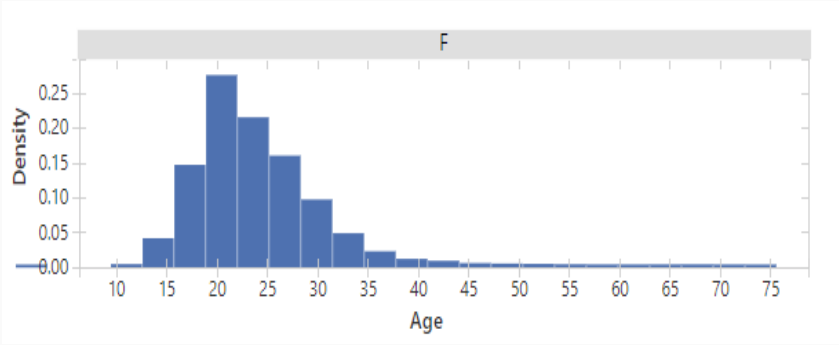


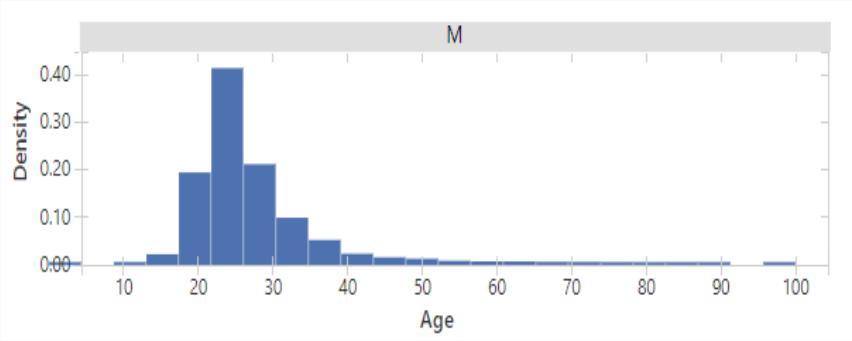


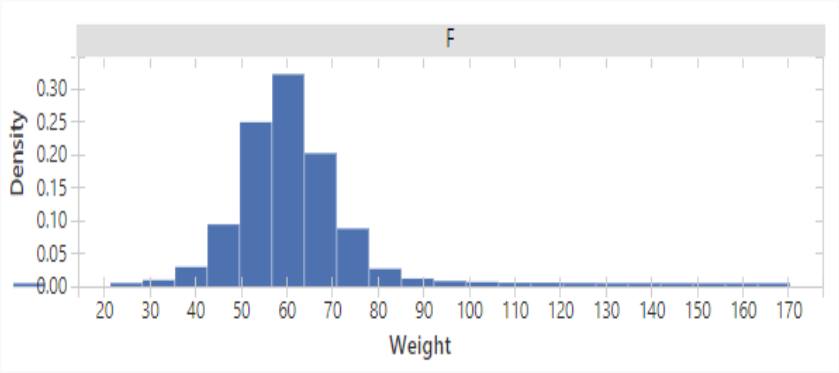


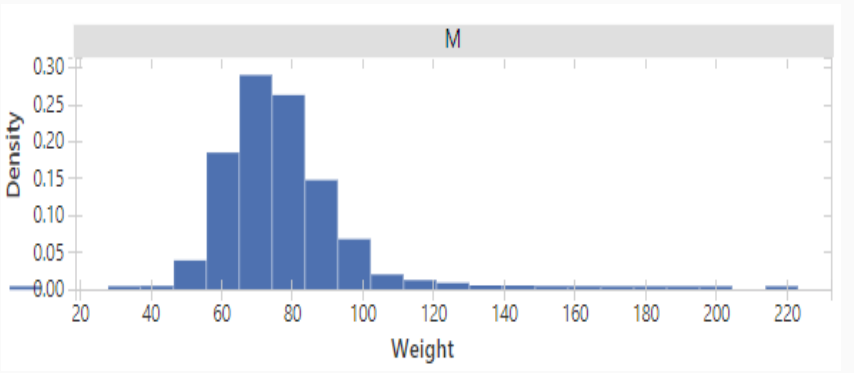


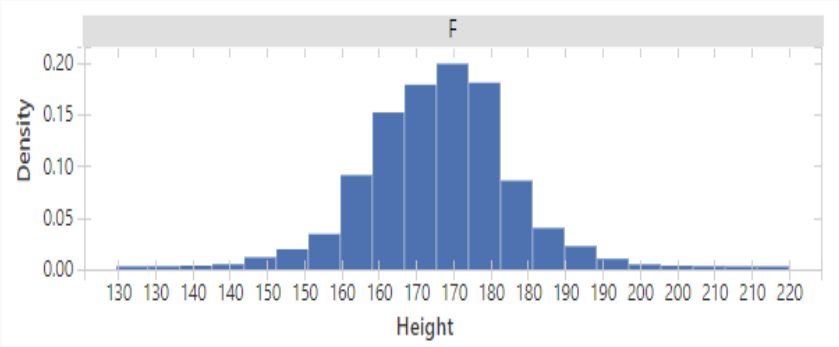
12. Age, Weight, and Height distribution for male and female athletes

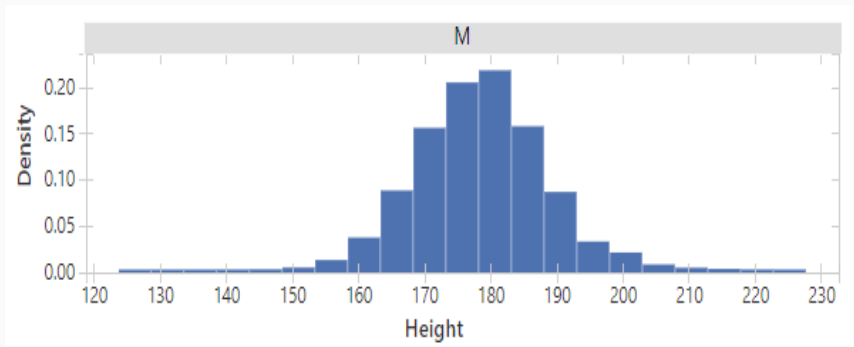








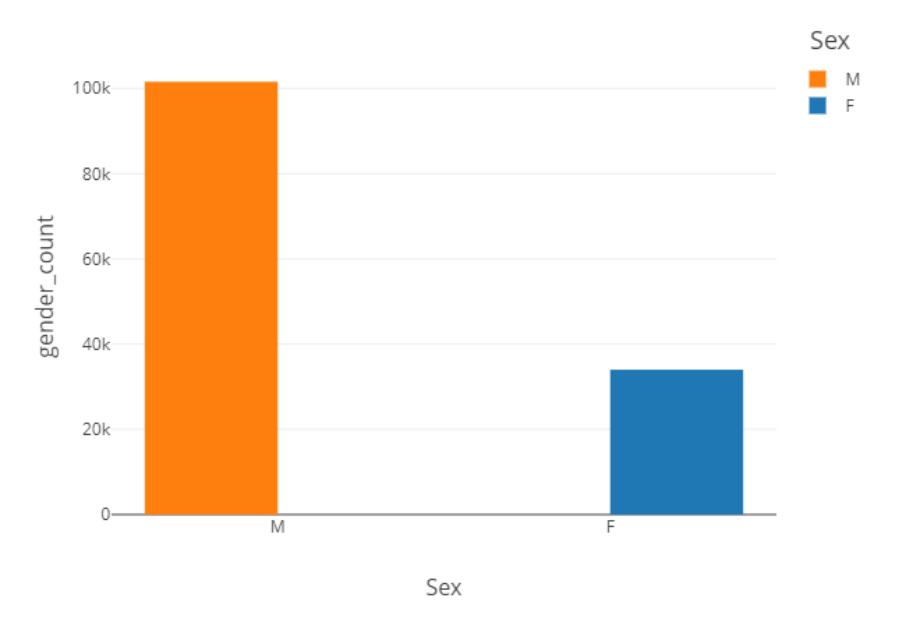




13. Names that belong to multiple IDs

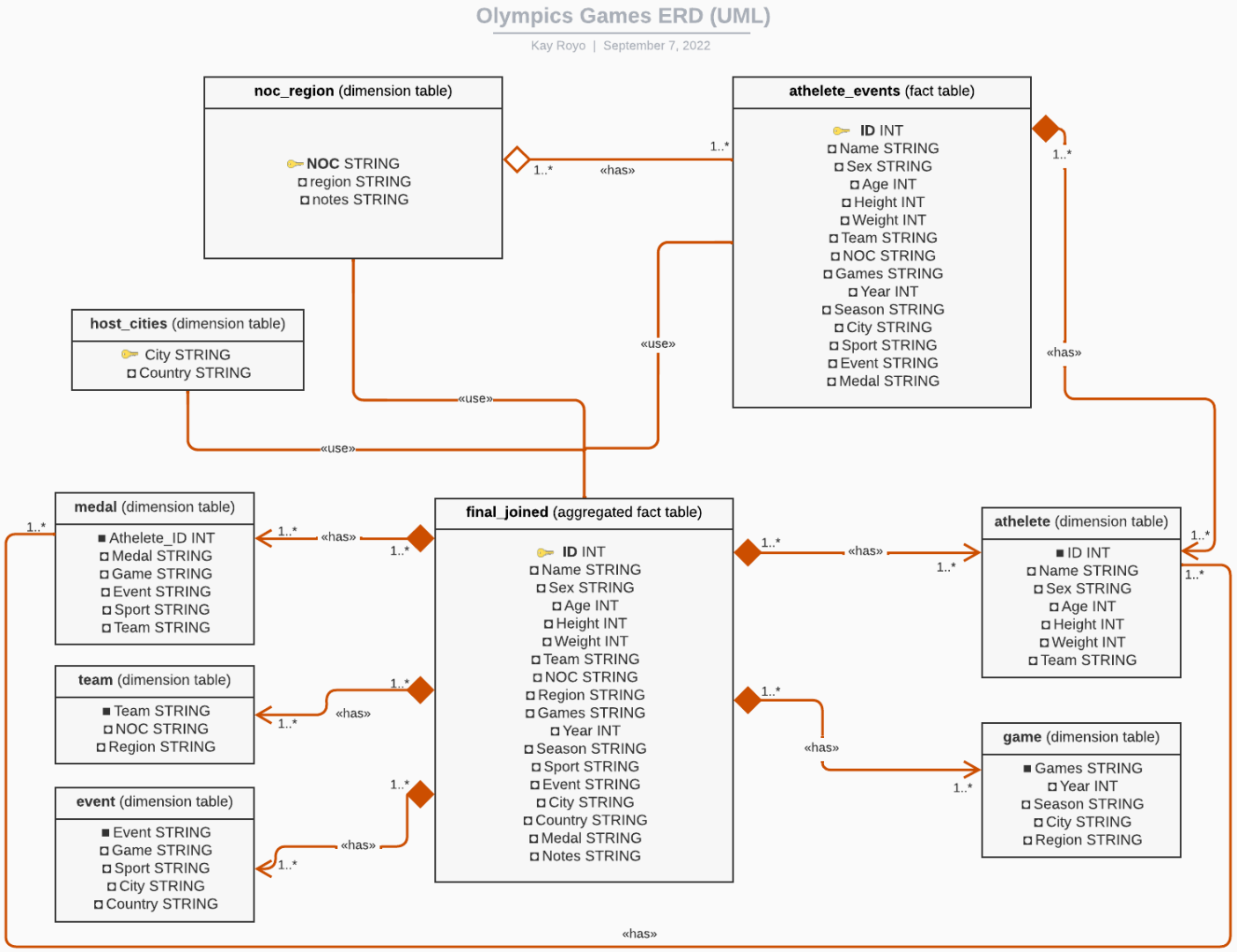
	Name	count
1	Wolfgang Mller	5
2	Wang Nan	5
3	Kim Seong-Eun	5
4	Lszl Szab	5
5	Li Jie	5
6	Ivan Ivanov	5
7	Zhang Li	5
8	Mohamed El-Sayed	4
9	Li Yang	4
10	Chen Jing	4

14. Number of players per gender

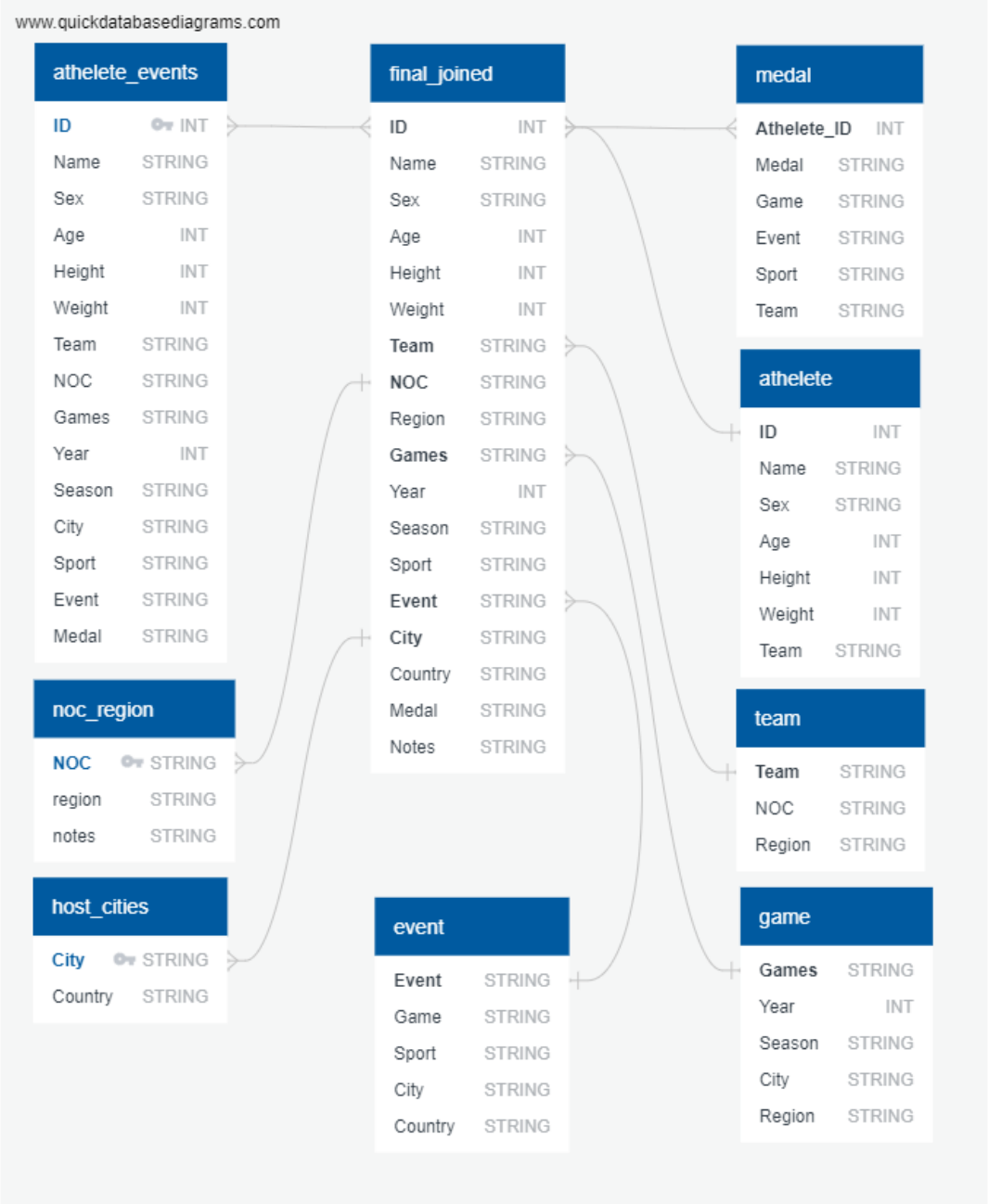


I.IV ERD

- UML notation



- Crow’ foot notation



II. Project Proposal Development

II.I Description

The main objective of this project is to extract valuable insights from the historical Olympics game dataset. This data analysis project uses the historical Olympic Games dataset containing 120 years of data from Athens 1896 to Rio 2016. This dataset contains information about the different Olympics Games held in over 200 National Olympics Committees (NOC) regions such as the professional athletes who participated in each game and the different events that were part of each game that are listed below. Another supplemental dataset that is used for this project contains information about the different NOC regions. A dataset containing information about the different host cities where the different Olympic Games were held in the past is also used. This project focuses on exploring how the Olympics Games have changed over the years, which can be used to predict its future outlook and outcome. The prediction of the future of the Olympics Games might be useful for the audiences that this project aims to target including media outlets, advertisers, athletes, prospective sponsors, and the general public interested in the Olympics games.

- athelete_events features:
    1. ID (unique number for each athlete)
    2. Name (Athlete's name)
    3. Sex (Male or Female)
    4. Age
    5. Height (In centimeters)
    6. Weight (In kilograms)
    7. Team (Name of team)
    8. NOC (National Olympic Committee 3-letter abbreviation)
    9. Games (Year and Season)
    10. Year
    11. Season (Winter or Summer
    12. City (Host City)
    13. Sport
    14. Event (Sport Event)
    15. Medal (Gold, Silver, Bronze, or Null)

- noc_regions features:
    1. NOC (National Olympic Committee 3-letter abbreviation)
    2. region (Full region/country name)
    3. notes (Current or former country/territory name)

- host_countries features:

1. City (Host city)
2. Country (Country where host city is located in)
3. Continent (Host country's continent)
4. Summer (Olympic game number)
5. Winter (Olympic game number)
6. Year
7. Opening Ceremony (Date)
8. Closing Ceremony (Date)

II.II Questions

The primary questions of interest this project aims to answer are listed as follows.

- Question 1: Does the location of the Olympics Games influence athlete performance? In other words, do hosting teams succeed more than visiting teams?
- Question 2: Do factors such as age, height, weight play an important role in succeeding in the Olympics Games?
- Question 3: Do male and female athletes' performance differ in different types of sports?

II.III Hypothesis

- Assumption 1: There is a significant difference between the performance of home and away teams.
- Assumption 2: Athlete's age, height, and weight play an important role in athlete performance.
- Assumption 3: Male perform better than female in some sports or vice versa.