SportsStats Capstone Project

CUZMIN SIMION

Preparing for Your Project Proposal

Which client/dataset did you select and why?

After an analysis of datasets proposed, i have choosen the SportsStats dataset, as i were a professional waterpolo player and i have strong knowledge in sports, as well i am more passioned about the events in the world of sports as other topics proposed.

Description

The aim of this project is to extract valuable insights and compile various statistics related to athlete performances in Olympic events over the past 120 years. The intended recipients of this information are sports fans and aficionados, as well as coaches and trainers who might benefit from the data. Additionally, the insights could be of interest to sports media outlets and platforms dedicated to disseminating intriguing sports-related content.

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

- Firstly, the dataset was acquired and saved on a local drive, as the file sizes were manageable and didn't necessitate the use of Databricks or multiple clusters for processing. My preferred coding and querying tool is a tailored version of the VSCode text editor, which I'm accustomed to using.

- Secondly, for reading the .csv files, I employed pandas from Python, and to transfer the data into a MySQL database, I used its native to_sql() function.

```
df_athletes = pd.read_csv('athlete_events.csv')
df_regions = pd.read_csv('noc_regions.csv')
df_athletes.head(5)
```

```
engine = connect(':memory:')

Athletes.to_sql('AthletesTable', con=engine)
Event.to_sql('EventTable', con=engine)
```

- Given that the dataset contains NaN (Not a Number) values, I opted not to clean these out, as removing or altering them would compromise the authenticity of the data.

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

- General info

```
df_athletes.info()
```

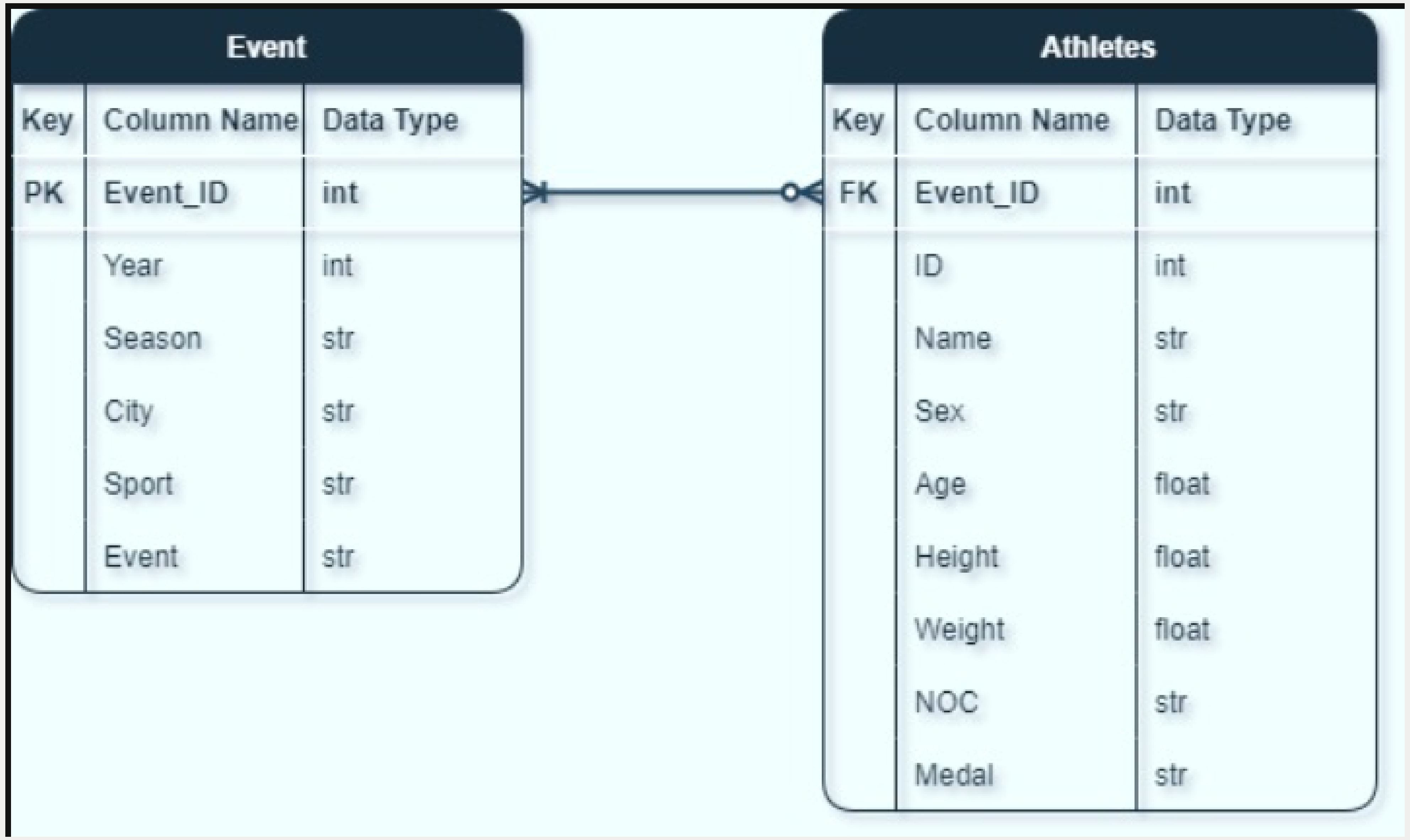
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 271116 entries, 0 to 271115
Data columns (total 15 columns):
 #   Column  Non-Null Count   Dtype  
---  ------  --------------   -----  
 0   ID      271116 non-null  int64  
 1   Name    271116 non-null  object 
 2   Sex     271116 non-null  object 
 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
```

```
df_regions.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 230 entries, 0 to 229
Data columns (total 3 columns):
 #   Column  Non-Null Count  Dtype 
---  ------  --------------  ----- 
 0   NOC     230 non-null    object
 1   region  227 non-null    object
 2   notes   21 non-null     object
dtypes: object(3)
memory usage: 5.5+ KB
```

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

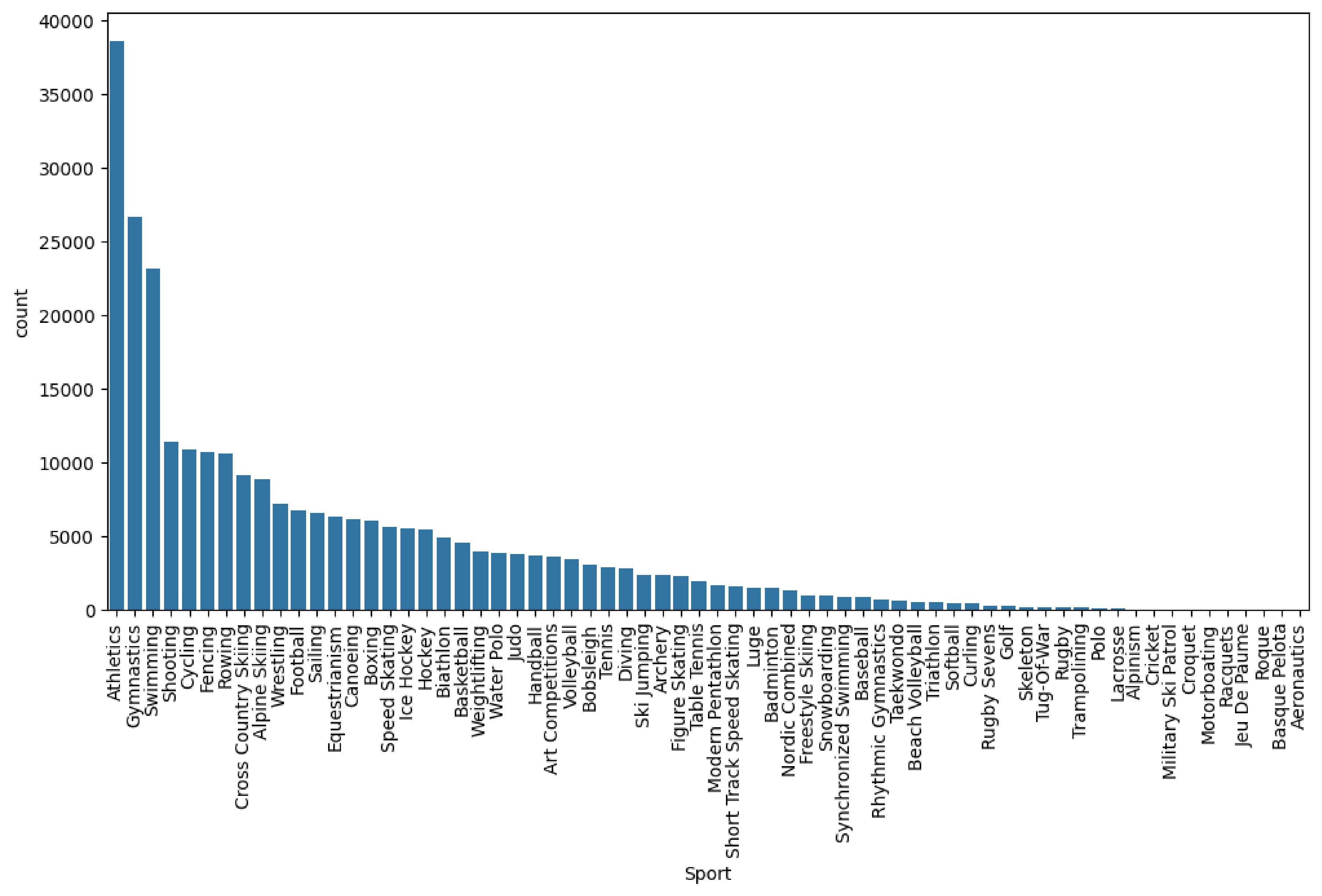
The displayed Entity-Relationship Diagram (ERD) was designed for a compact relational database, organizing the data into two tables: 'athletes' and 'event'. A few adjustments were necessary; for instance, the 'ID' column did not contain unique entries and therefore couldn't serve as a primary key (PK). As a solution, a new column named "Event_ID" was introduced in the 'Event' table to act as the PK and was also included in the 'Athletes' table as a foreign key (FK).



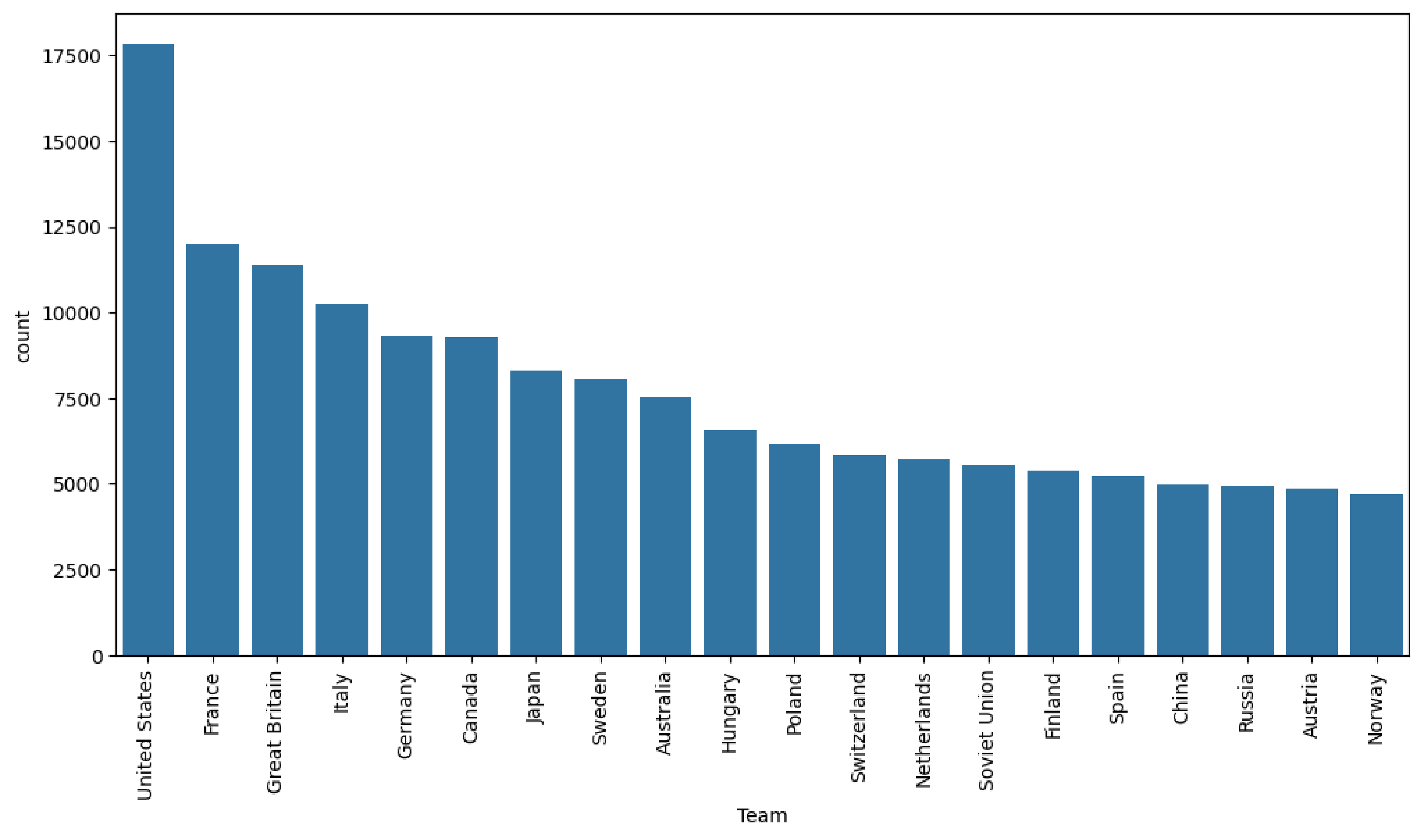
Hypothesis

1. Nations situated in higher latitudes may exhibit superior performance in winter sports, as evidenced by their medal tallies.
2. The representation of female and male athletes in competitions has become more balanced throughout the years.
3. More industrially developed nations tend to accumulate a greater number of medals.
4. Athletes around the age of 25 may have a higher likelihood of securing medals in competitive events.

The distributon of sports



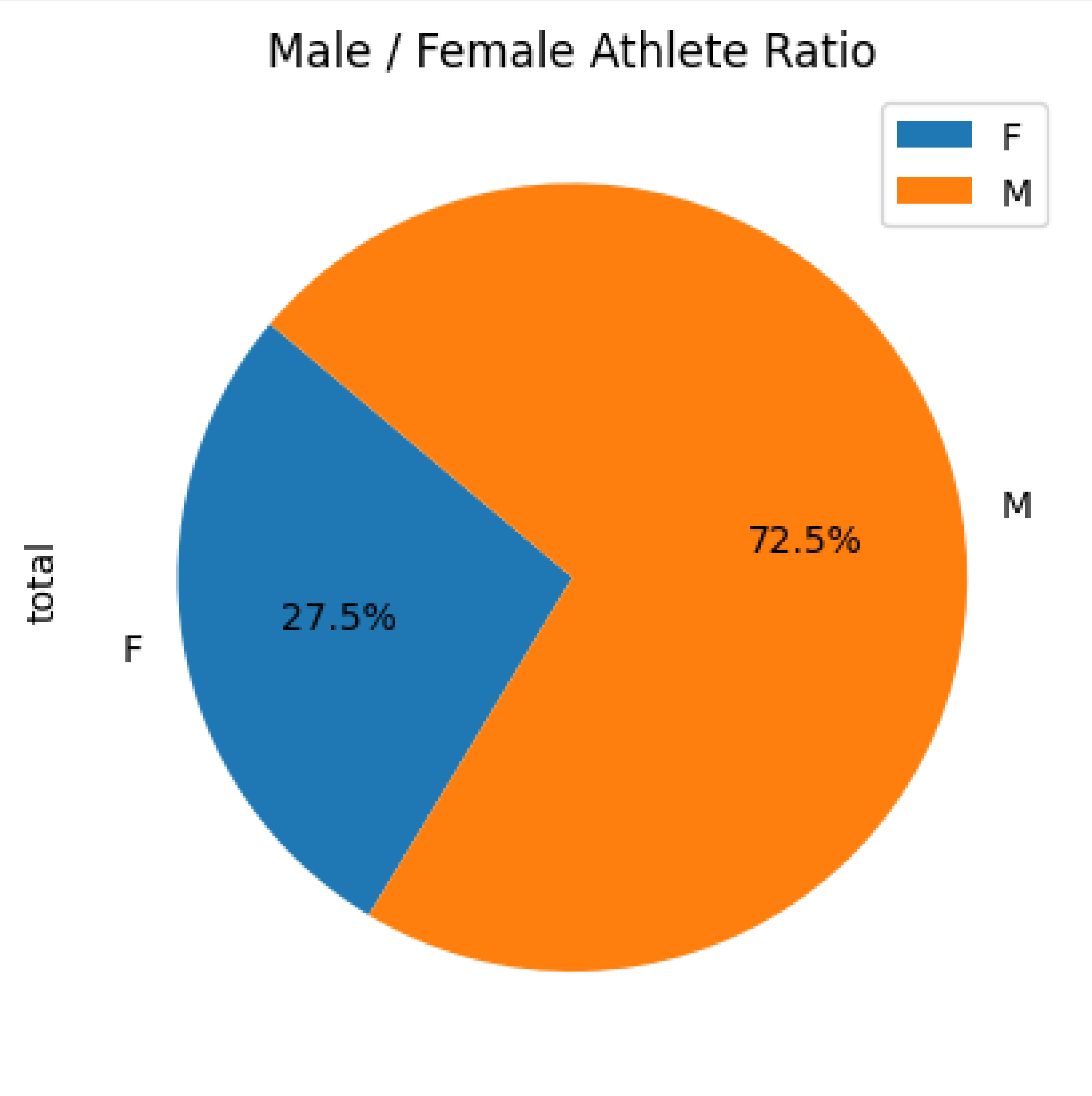
The distributon of teams



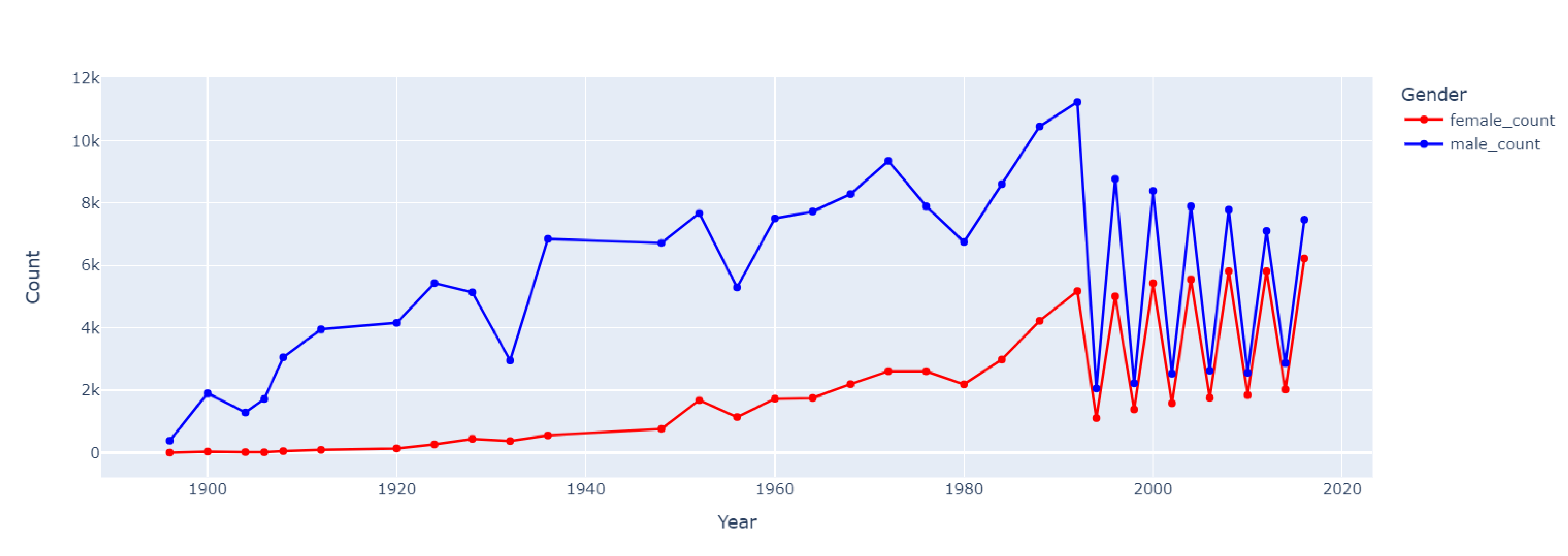
Gender distribution

```
df_sex_ratio = pd.read_sql(
    '''
    SELECT
        sex,
        COUNT(*) AS total
    FROM
        AthletesTable
    GROUP BY
        sex
    ''', con=engine
)
df_sex_ratio
```

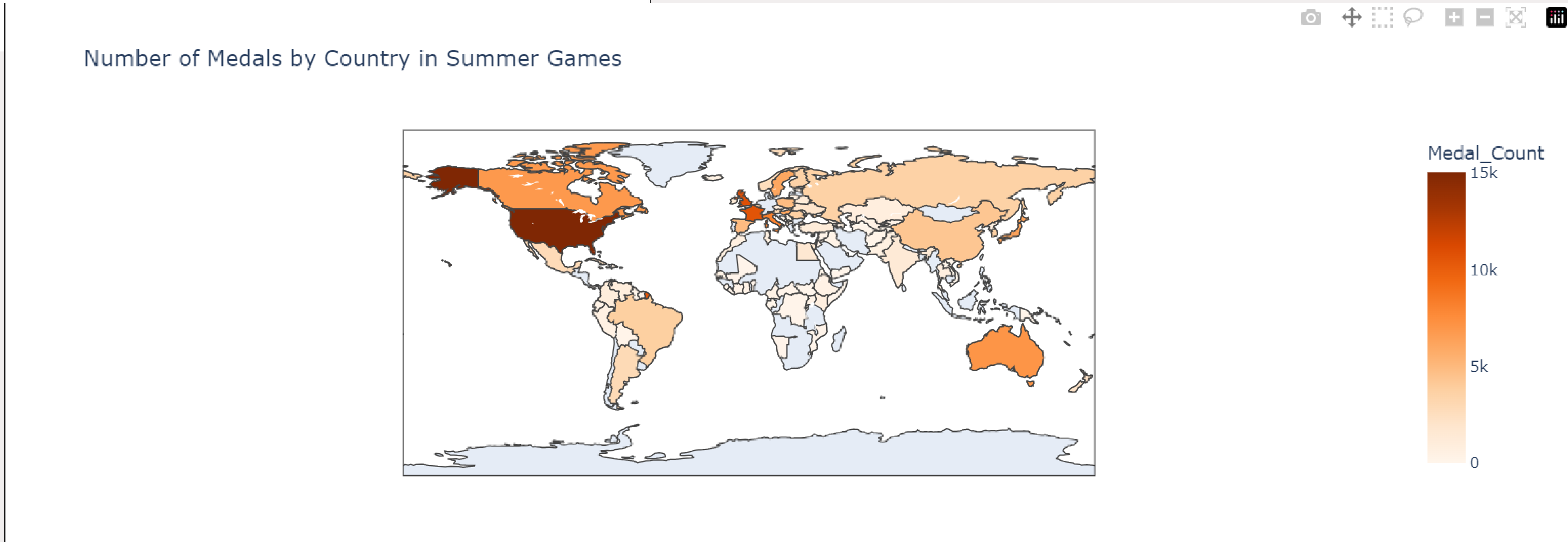
	Sex	total
0	F	74522
1	M	196594



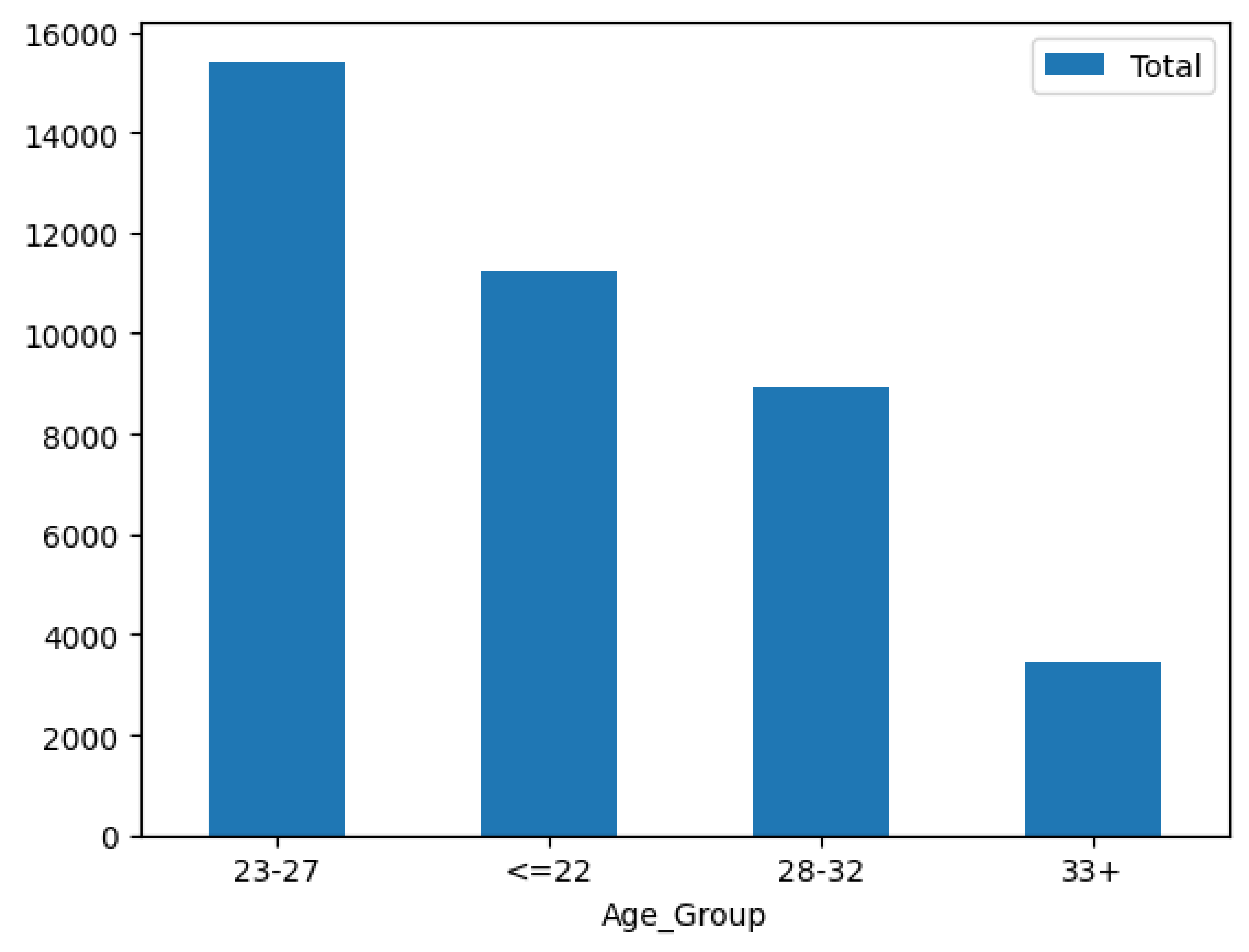
Gender distribution over years



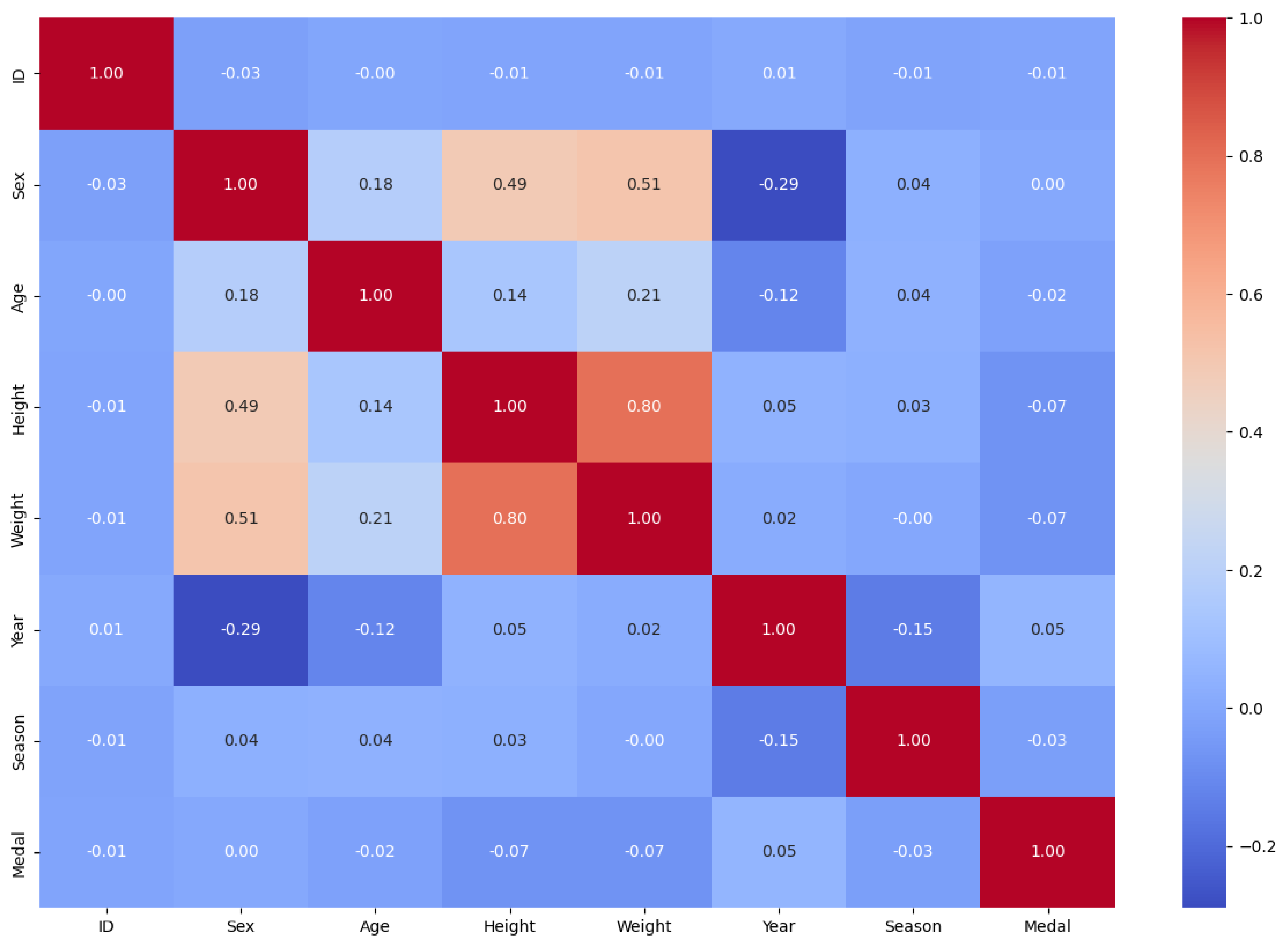
Distribution of medals



Distribution of medals by age groups



Correlation matrix



Creating new metrics

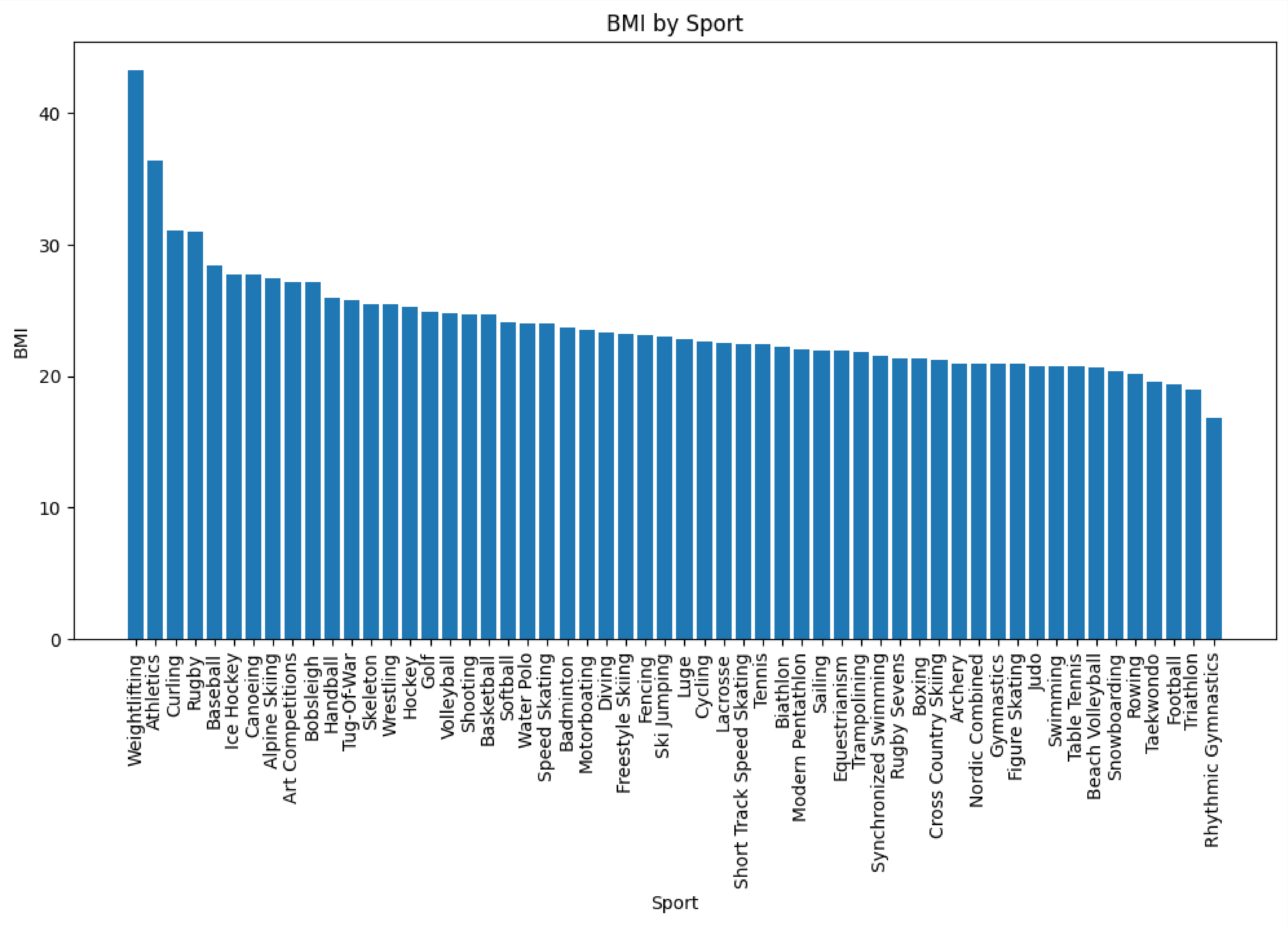
BMI

	Year	Age	Sex	Weight	Height	Season	Sport	BMI
0	2016	22.0	F	125.0	170.0	Summer	Weightlifting	43.252595
1	2000	31.0	M	130.0	189.0	Summer	Athletics	36.393158
2	2006	24.0	M	95.0	175.0	Winter	Curling	31.020408
3	1924	21.0	M	98.0	178.0	Summer	Rugby	30.930438
4	2000	21.0	M	91.0	179.0	Summer	Baseball	28.401111
5	2002	26.0	M	96.0	186.0	Winter	Ice Hockey	27.748873
6	1992	27.0	M	82.0	172.0	Summer	Canoeing	27.717685
7	1992	20.0	M	85.0	176.0	Winter	Alpine Skiing	27.440599
8	1932	44.0	M	91.0	183.0	Summer	Art Competitions	27.173102
9	1998	24.0	M	98.0	190.0	Winter	Bobsleigh	27.146814
10	2008	23.0	M	86.0	182.0	Summer	Handball	25.963048
11	1920	NaN	M	95.0	192.0	Summer	Tug-Of-War	25.770399
12	2002	24.0	M	78.0	175.0	Winter	Skeleton	25.469388
13	2000	22.0	M	89.0	187.0	Summer	Wrestling	25.451114
14	2000	25.0	M	80.0	178.0	Summer	Hockey	25.249337
15	2016	41.0	M	72.0	170.0	Summer	Golf	24.913495
16	2008	23.0	M	94.0	195.0	Summer	Volleyball	24.720579
17	1936	33.0	M	93.0	194.0	Summer	Shooting	24.710384
18	1992	24.0	M	80.0	180.0	Summer	Basketball	24.691358
19	2008	23.0	F	88.0	191.0	Summer	Softball	24.122146
20	1996	22.0	M	83.0	186.0	Summer	Water Polo	23.991213
21	1988	21.0	F	82.0	185.0	Winter	Speed Skating	23.959094

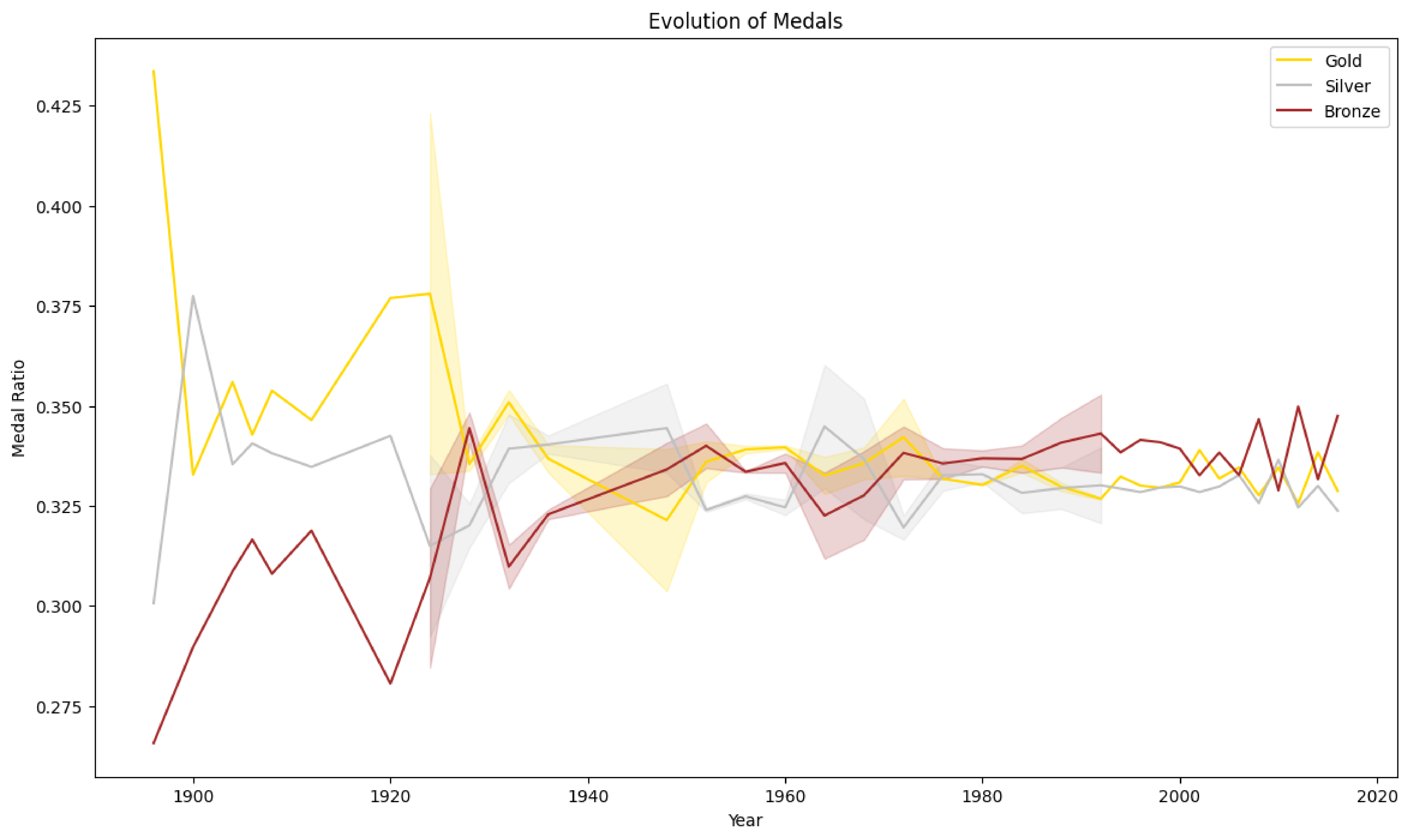
Medal Ratio

	year	season	medal_ratio	gold_ratio	silver_ratio	bronze_ratio
0	1896	Summer	1.0	0.433566	0.300699	0.265734
1	1900	Summer	1.0	0.332781	0.377483	0.289735
2	1904	Summer	1.0	0.355967	0.335391	0.308642
3	1906	Summer	1.0	0.342795	0.340611	0.316594
4	1908	Summer	1.0	0.353791	0.338147	0.308063
5	1912	Summer	1.0	0.346440	0.334750	0.318810
6	1920	Summer	1.0	0.376911	0.342508	0.280581
7	1924	Summer	1.0	0.332933	0.337740	0.329327
8	1924	Winter	1.0	0.423077	0.292308	0.284615
9	1928	Summer	1.0	0.333787	0.325613	0.340599
10	1928	Winter	1.0	0.337079	0.314607	0.348315
11	1932	Summer	1.0	0.353941	0.330757	0.315301
12	1932	Winter	1.0	0.347826	0.347826	0.304348
13	1936	Summer	1.0	0.340240	0.338059	0.321701
14	1936	Winter	1.0	0.333333	0.342593	0.324074
15	1948	Summer	1.0	0.339202	0.333333	0.327465
16	1948	Winter	1.0	0.303704	0.355556	0.340741
17	1952	Summer	1.0	0.341137	0.324415	0.334448
18	1952	Winter	1.0	0.330882	0.323529	0.345588
19	1956	Summer	1.0	0.338186	0.328108	0.333707
20	1956	Winter	1.0	0.340000	0.326667	0.333333
21	1960	Summer	1.0	0.339188	0.322722	0.338090

Studying BMI impact



Studying Medal Ratio impact



Conclusion

Conclusion

In conclusion, the analysis has revealed insightful findings about Olympic sports dynamics. It confirmed initial hypotheses, showing geographical advantages in Winter Games for higher latitude countries, a trend towards gender balance in sports, and the influence of economic development on medal counts. Moving forward, exploring post-Soviet performance, diversity's impact on medals, and age-related trends is crucial. Additionally, new metrics like Body Mass Index (BMI) offer deeper insights into athletes' physical characteristics. Analysis of medal distribution ratios over time highlights the Olympics' enduring structure and evolving medal awards. In summary, this project offers a comprehensive understanding of Olympic data, suggesting potential for further analysis using advanced techniques and additional datasets.