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.

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

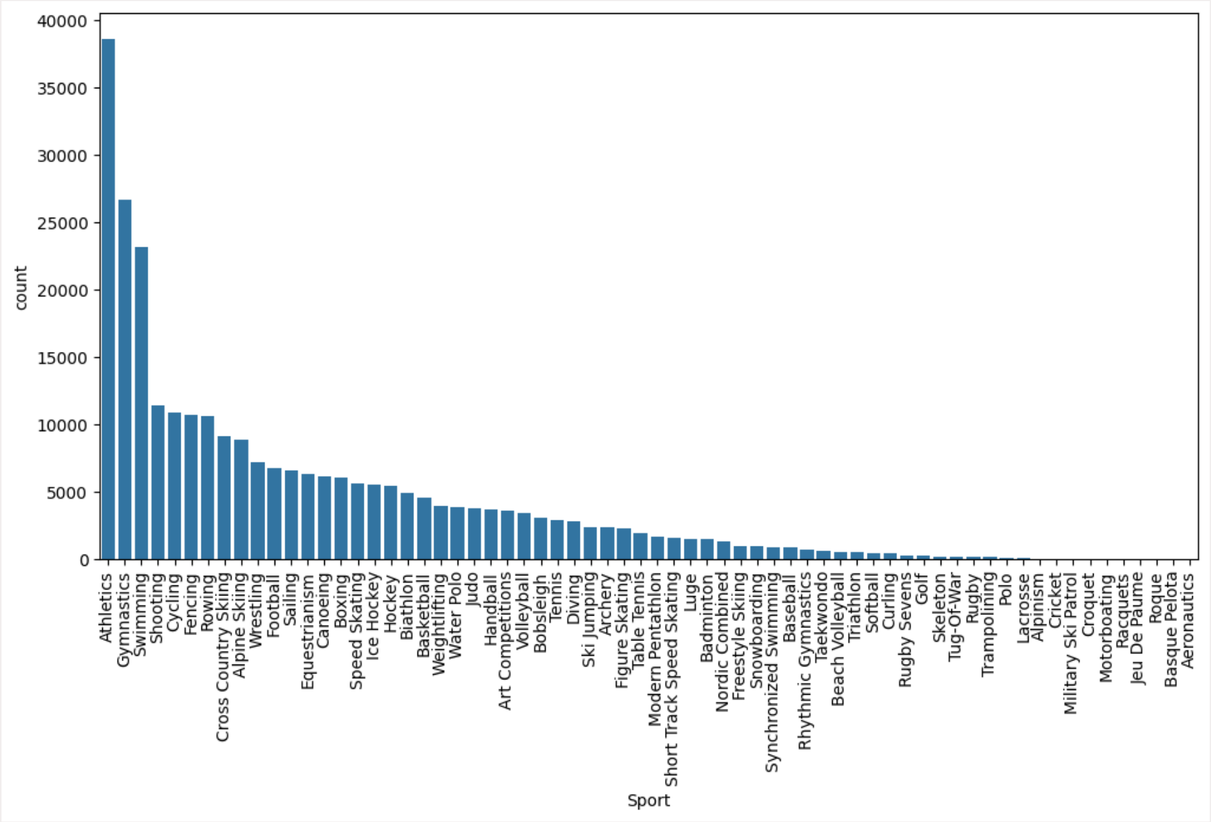
```
df_athletes.info()
```

```
<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
```

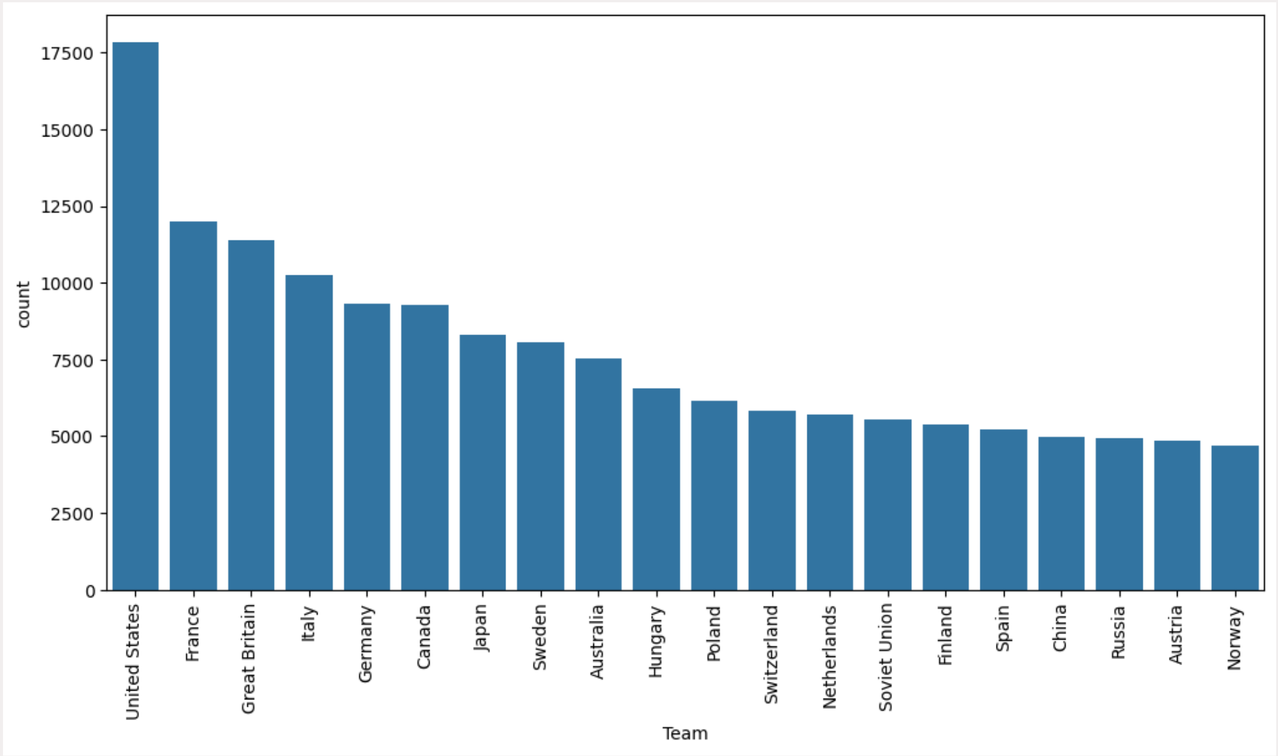
```
df_regions.info()
```

```
<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
```

The distributon of sports



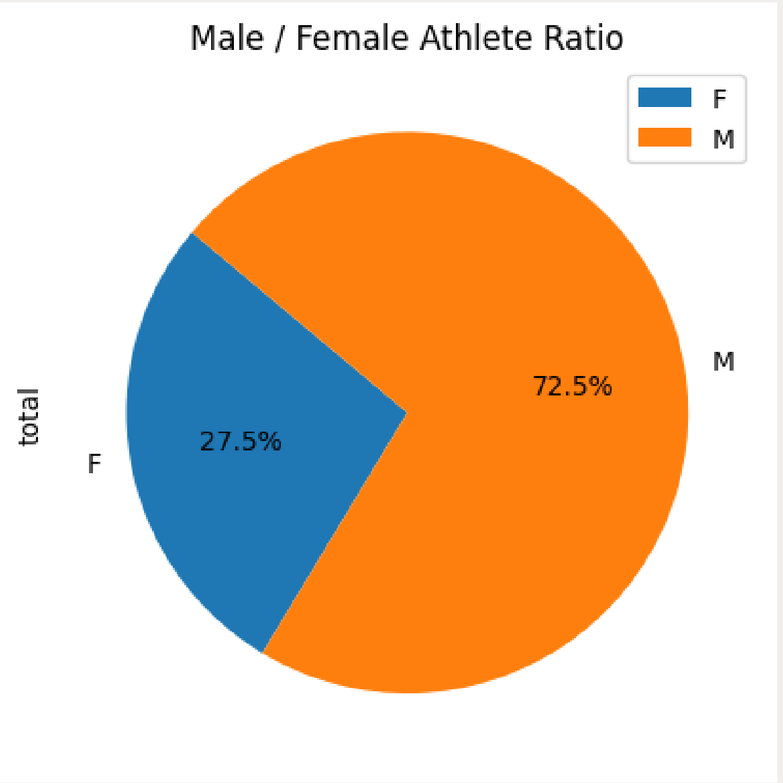
The distributon of teams



Gender distribution

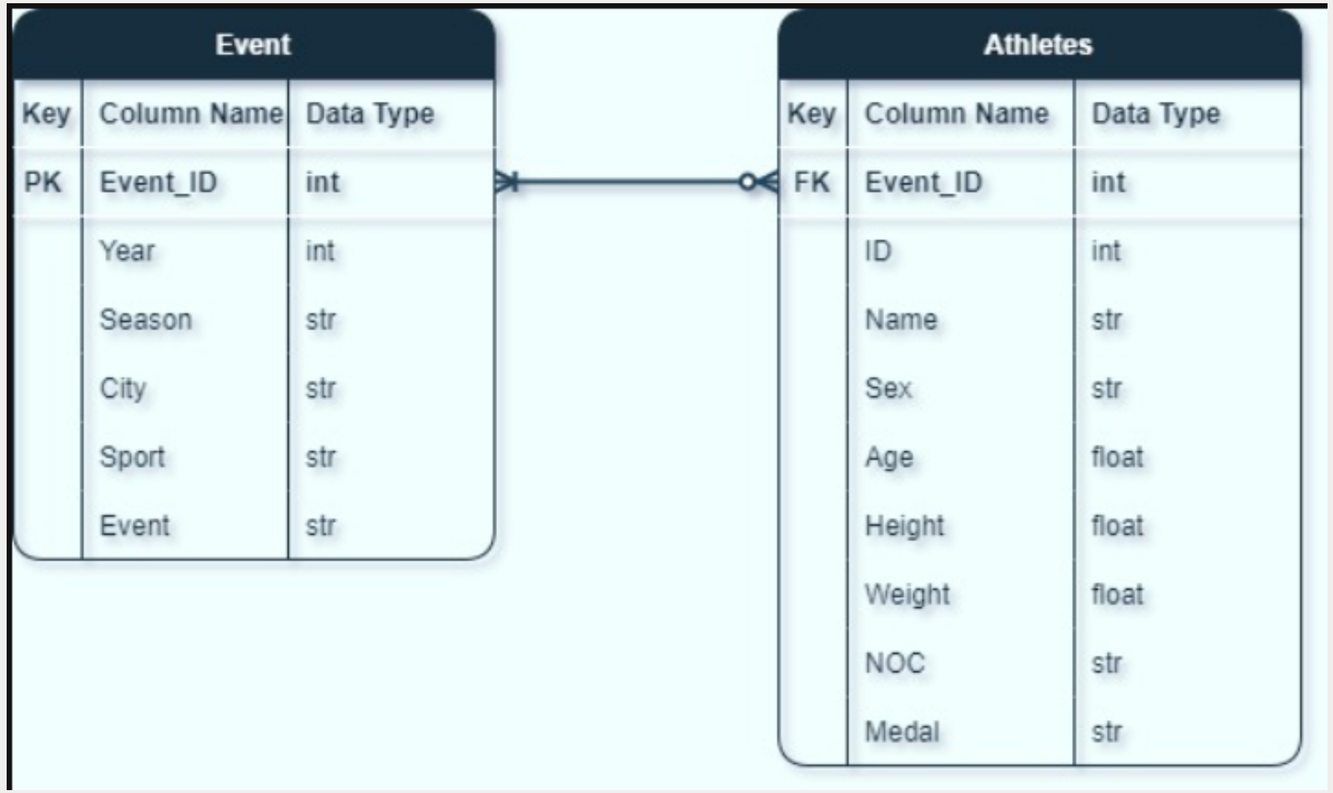
```
df_sex_ratio = pd.read_sql(
    '''
    SELECT
        sex,
        COUNT(*) AS total
    FROM
        AthletesTable
    GROUP BY
        sex
    ''', con=engine
)
df_sex_ratio
```

	Sex	total
0	F	74522
1	M	196594



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).



Develop Your Project Proposal

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.

Questions

1. To what extent does an athlete's age impact their probability of winning a medal?
2. Is there a correlation between a country’s wealth and its likelihood of securing medals, possibly due to early and substantial investment in sports?
3. What is the distribution of medals across seasons, particularly with respect to geographic location? For instance, do countries from northern latitudes fare better in Winter Season events and the countries from south latitudes better in Summer Season?
4. Over the historical timeline, has the gender gap in athlete participation narrowed, especially in recent decades?

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.

Approach

1. **Distribution of age and Medals**

2. **Distribution of medals and countries**

3. **Distribution of men and women over the years**