SportsStats

Capstone Project

LI RUI

Table Of Contents

Preparing Proposal

Which client did you select and why?

- I choose the SportsStats client. I selected this project since I have strong knowledge in sports and have interests in working in sports analytics potentially.

- With an analysis of the dataset, I can find patterns and hidden insights for players, hidden insights.

Steps used to import and Clean data

- For importing the data, I used pandas to read in both csv files

- Since there were many NAN values, I didn’t do much cleaning since that would falsify the data!

Initial Exploration of Data and display screenshots of work

In [11]:
```
# info of athlete df columns
athlete_events_df.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
```

Observation

- Fairly evenly distributed data types amongst the columns.
- Object type is the most ubiquitous
- Can see some missing values in data right away

In [12]:
```
# stats of athletes columns
athlete_events_df.describe()
```

Out[12]:

	ID	Age	Height	Weight	Year
count	271116.000000	261642.000000	210945.000000	208241.000000	271116.000000
mean	68248.954396	25.556898	175.338970	70.702393	1978.378480
std	39022.286345	6.393561	10.518462	14.348020	29.877632
min	1.000000	10.000000	127.000000	25.000000	1896.000000
25%	34643.000000	21.000000	168.000000	60.000000	1960.000000
50%	68205.000000	24.000000	175.000000	70.000000	1988.000000
75%	102097.250000	28.000000	183.000000	79.000000	2002.000000
max	135571.000000	97.000000	226.000000	214.000000	2016.000000

Observation

- Nothing too crazy about stats of athletes df columns, although I am curious of the 97 years old who participated

Initial Exploration of Data and display screenshots of work

In [14]:
```
# making sure only two sexes
athlete_events_df["Sex"].unique()
```

Out[14]:
```
array(['M', 'F'], dtype=object)
```

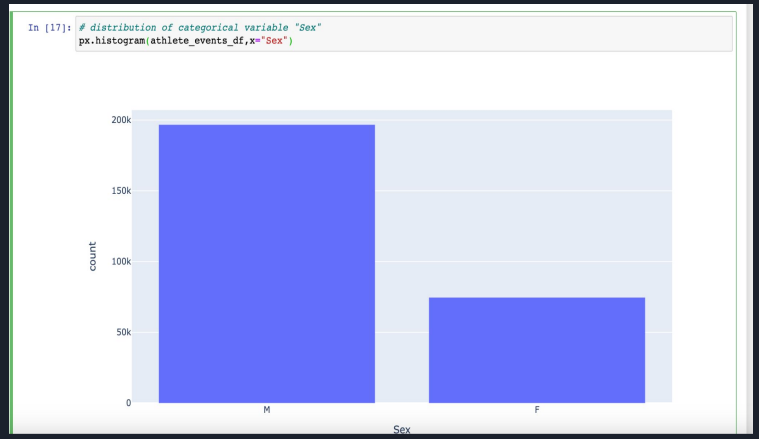
In [15]:
```
# seeing distribution of sexes
athlete_events_df["Sex"].value_counts(normalize=True)
```

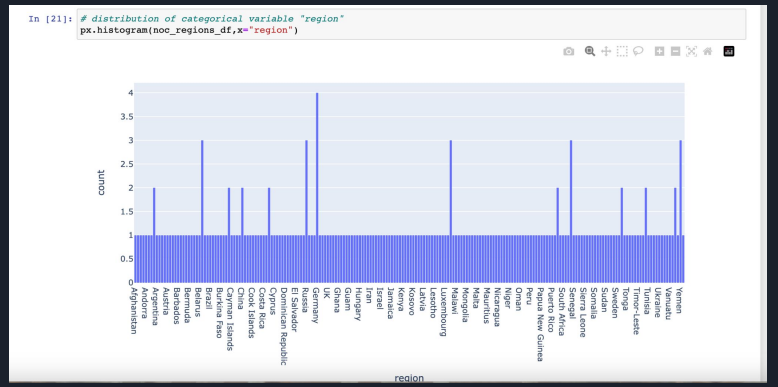
Out[15]:
```
M    0.725129
F    0.274871
Name: Sex, dtype: float64
```

Observation

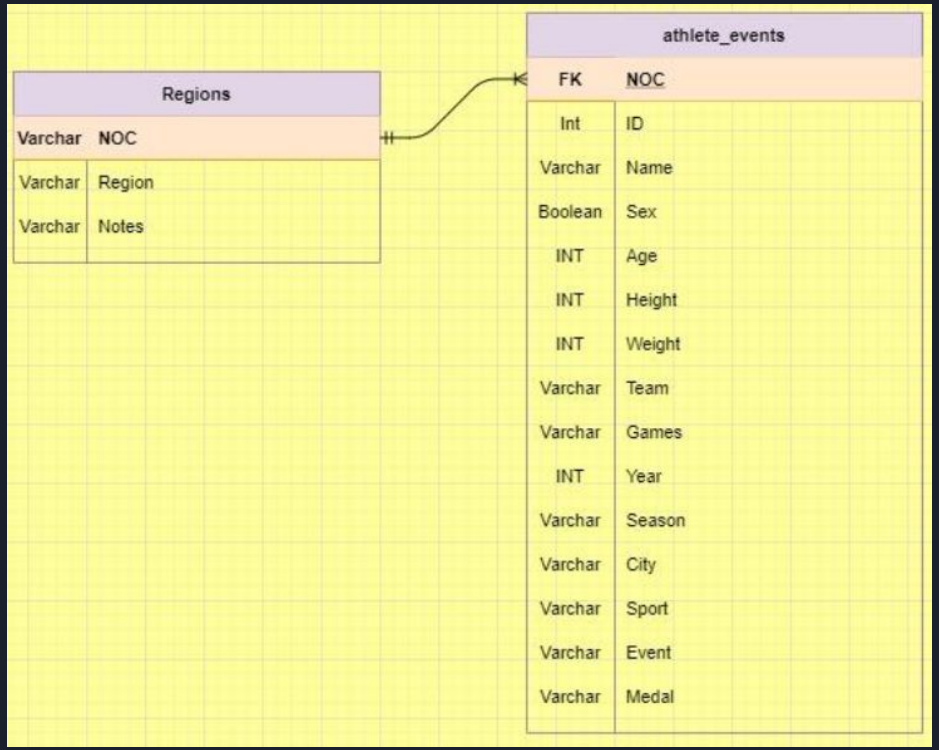
- There are more male athletes, with approximatley 75% being male and 25% being female

In [17]:
```
# distribution of categorical variable "Sex"
px.histogram(athlete_events_df,x="Sex")
```





ERD


Regions
Varchar NOC
Varchar Region
Varchar Notes
athlete_events
FK NOC
Int ID
Varchar Name
Boolean Sex
INT Age
INT Height
INT Weight
Varchar Team
Varchar Games
INT Year
Varchar Season
Varchar City
Varchar Sport
Varchar Event
Varchar Medal

Developing Project Proposal

Possible questions

- When was the first season ever conducted
- Disparity between sexes
- Which country performed the best
- Age distribution of participants
- Which country had highest number of Players

Description

- My Project targets getting past Sports patterns and analysing it.
- Get to know more insights on the data, such as when the first event was organized and in which city/country.
- This analysis will not only help Sports Coaches to identify patterns and records, but it will also help the SportsStats firm aid in their clients' decision-making.
- My audience for the projects would not be limited to Coaches/Trainers but also players who will be able to see their records/performance in past events.

Possible Hypothesis

- Women have higher number of Medals.
- Year > 1956 will have the Highest number of Events.
- More people have participated in Football.
- 3 Teams will have medals > 40.
- People with Age > 40 have received medal in any of the events.