Olympics Dataset - 120 years of data

- Part 1: Prepare for my project proposal
- Part 2: Develop my Project Proposal

Prepare for my proposal

- Client/Dataset: Olympics Dataset - 120 years of data
- Reason/Goal: I’m interested in Olympics, and find the pattern to have higher chance to get the golden medal
- Import data: Use Pandas library in Python to read the dataset (shown in the next page)
- Entity Relationship Diagram (ERD)

Import Data – athlete events

```
import pandas as pd
```

```
olympics_athletes = pd.read_csv('athlete_events.csv')
olympics_athletes.head()
```

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event	Medal
0	1	A Dijiang	M	24.0	180.0	80.0	China	CHN	1992 Summer	1992	Summer	Barcelona	Basketball	Basketball Men's Basketball	NaN
1	2	A Lamusi	M	23.0	170.0	60.0	China	CHN	2012 Summer	2012	Summer	London	Judo	Judo Men's Extra-Lightweight	NaN
2	3	Gunnar Nielsen Aaby	M	24.0	NaN	NaN	Denmark	DEN	1920 Summer	1920	Summer	Antwerpen	Football	Football Men's Football	NaN
3	4	Edgar Lindenau Aabye	M	34.0	NaN	NaN	Denmark/Sweden	DEN	1900 Summer	1900	Summer	Paris	Tug-Of-War	Tug-Of-War Men's Tug-Of-War	Gold
4	5	Christine Jacoba Aaftink	F	21.0	185.0	82.0	Netherlands	NED	1988 Winter	1988	Winter	Calgary	Speed Skating	Speed Skating Women's 500 metres	NaN

```
olympics_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
```

Data cleaning – athlete events

```
olympics_athletes.dropna(subset = ['Medal'], axis = 0, inplace = True)
olympics_athletes.head()
```

:

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event	Medal
3	4	Edgar Lindenau Aabye	M	34.0	NaN	NaN	Denmark/Sweden	DEN	1900 Summer	1900	Summer	Paris	Tug-Of-War	Tug-Of-War Men's Tug-Of-War	Gold
37	15	Arvo Ossian Aaltonen	M	30.0	NaN	NaN	Finland	FIN	1920 Summer	1920	Summer	Antwerpen	Swimming	Swimming Men's 200 metres Breaststroke	Bronze
38	15	Arvo Ossian Aaltonen	M	30.0	NaN	NaN	Finland	FIN	1920 Summer	1920	Summer	Antwerpen	Swimming	Swimming Men's 400 metres Breaststroke	Bronze
40	16	Juhamatti Tapio Aaltonen	M	28.0	184.0	85.0	Finland	FIN	2014 Winter	2014	Winter	Sochi	Ice Hockey	Ice Hockey Men's Ice Hockey	Bronze
41	17	Paavo Johannes Aaltonen	M	28.0	175.0	64.0	Finland	FIN	1948 Summer	1948	Summer	London	Gymnastics	Gymnastics Men's Individual All-Around	Bronze

- The “Medal” column has been cleaned since I just need the information from relation between athletes and medal

Import Data – NOC regions

```
olympics_noc = pd.read_csv('noc_regions.csv')
olympics_noc.head()
```

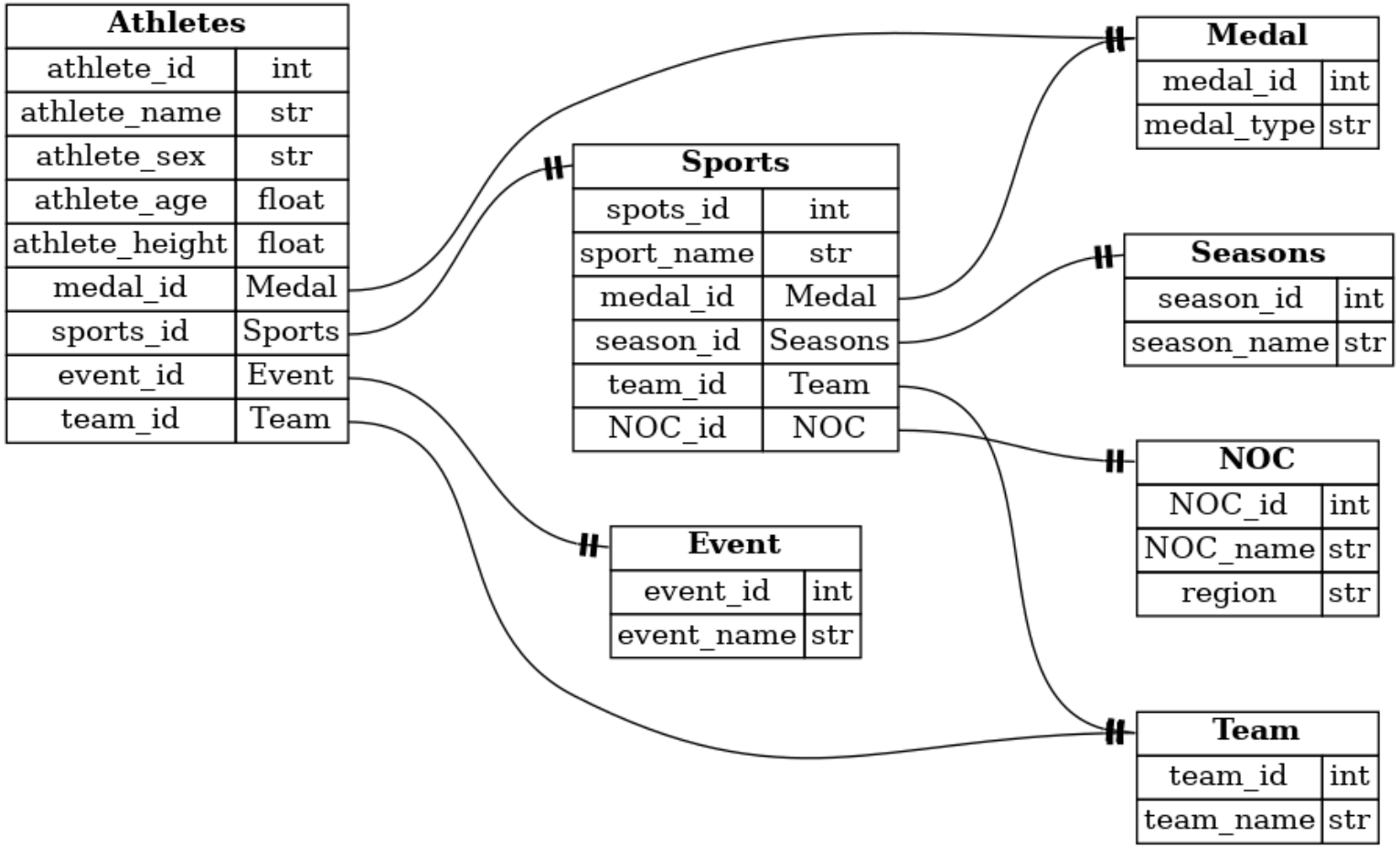
:

	NOC	region	notes
0	AFG	Afghanistan	NaN
1	AHO	Curacao	Netherlands Antilles
2	ALB	Albania	NaN
3	ALG	Algeria	NaN
4	AND	Andorra	NaN

```
olympics_noc.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
```

Entity Relationship Diagram (ERD)



Olympics Dataset - 120 years of data

Description

- Olympics is an international sports event. Thousands of athletes from different countries participate in a variety of competitions. Most of athletes who represent their country want to get the medal and bring the honor to their country. However, many factors can affect if an athlete can get the medal or not. I’m interested in finding the pattern related to the athletes who got the medal. This will be helpful for the country governments to choose good people to become Olympic athletes and win the medal. Besides, this will be useful for the chosen athletes to predict how much chance they can get the medal.

Questions

- Athletes' characteristics:

-Does an athlete get a medal related to their sex, height, age, weight? Is the medal type also related to these conditions?

-Does an athlete get a medal in specific sports related to their height and weight?

- Are there more male than female participating the Olmpics in all sport types?

- External factors:

-Does an athlete get a medal related to the NOC, their team and region?

- Does an athlete get a medal related to the season?

Hypothesis

- The ratio of female and male is the same in all sports type.
- An athlete's height, age and weight is related to if an athlete has higher chance to get the medal in specific sports.
- The competition is fair, so NOC and their team are not related to if an athlete get a medal.

Approach

- In ERD, athlete table includes the intrinsic features of athletes, and it is related to medal table and sports table. In this project, I will use aggregate function and pie chart to test my first hypothesis. Then, I will join the athlete table and medal table to test my hypothesis. Finally, I will join the NOC tables and athlete table together to test my third hypothesis.