Preparing for Your Proposal

Which client/dataset did you select and why?

I selected the Client 3: SportsStats (Olympics Dataset - 120 years of data)
SportsStats is a sports analysis firm partnering with local news and elite personal trainers to provide “interesting” insights to help their partners.
The reason behind selecting this dataset is

- I have an interest in Sports so as it is my first project I will have a better understanding of terms and the values in the data set and this will help me to work on the dataset more efficiently.
- With an analysis of the dataset, I can find patterns and hidden insights for players, hidden insights.

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

<u>Importing data :-</u>

For this I will be using Jupyter notebook using Python Language, the following steps are :-

Step 1

>> sportsstats = pd.read_csv('/Project/athlete_events.csv', index_col = 0);

>> noc=pd.read_csv('/Project/noc_regions.csv',index_col = 0)

```
In [15]: sportsstats.head()
```

Out[15]:

ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event	Medal
1	A Dijiang	M	24.0	180.0	80.0	China	CHN	1992 Summer	1992	Summer	Barcelona	Basketball	Basketball Men's Basketball	NaN
2	A Lamusi	M	23.0	170.0	60.0	China	CHN	2012 Summer	2012	Summer	London	Judo	Judo Men's Extra-Lightweight	NaN
3	Gunnar Nielsen Aaby	M	24.0	NaN	NaN	Denmark	DEN	1920 Summer	1920	Summer	Antwerpen	Football	Football Men's Football	NaN
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
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

To import data to work in sql we will import the sqlalchemy package, and will make a database in sqlite3 and then will connect to the database using the following commands :-

>> import sqlalchemy

>> engine=sqlalchemy.create_engine('sqlite:///athlete.db')

>> %load_ext sql

>>%sql sqlite:///athlete.db

```
In [16]: import sqlalchemy

In [17]: engine=sqlalchemy.create_engine('sqlite:///athlete.db')

In [18]: %load_ext sql

         The sql extension is already loaded. To reload it, use:
           %reload_ext sql

In [19]: %reload_ext sql

In [20]: %sql sqlite:///athlete.db
```

Then we will convert the “sportsstats” dataframe into sql table using :-

>> sportsstats.to_sql('athlete_data',engine)

>> noc.to_sql('noc_data',engine)

We will run the magic command to work on sql :-

>>%%sql

pragma table_info('athlete_data')

```
In [23]: sportsstats.to_sql('athlete_data',engine)
Out[23]: 271116

In [32]: %%sql
         pragma table_info('athlete_data')
          * sqlite:///athlete.db
         Done.
```

Out[32]:

cid	name	type	notnull	dflt_value	pk
0	ID	BIGINT	0	None	0
1	Name	TEXT	0	None	0
2	Sex	TEXT	0	None	0
3	Age	FLOAT	0	None	0
4	Height	FLOAT	0	None	0
5	Weight	FLOAT	0	None	0
6	Team	TEXT	0	None	0
7	NOC	TEXT	0	None	0
8	Games	TEXT	0	None	0
9	Year	BIGINT	0	None	0
10	Season	TEXT	0	None	0
11	City	TEXT	0	None	0
12	Sport	TEXT	0	None	0
13	Event	TEXT	0	None	0
14	Medal	TEXT	0	None	0

Cleaning of data

Data aggregation

1:- Removing duplicate values

To remove data from we will make a new table using the previous table and using the remove the duplicate values using following statements :-

>>%%sql

CREATE TABLE athlete_data20 AS SELECT *, ROW_NUMBER() OVER

(PARTITION BY ID ORDER BY ID) as row_num FROM athlete_data ;

DELETE FROM athlete_data20 WHERE row_num > 1;

alter table athlete_data20 drop column row_num;

SELECT * from athlete_data20 limit 10;

Drop table athlete_data;

```
In [106]: %%sql
          SELECT * from athlete_data20 limit 10
```

```
 * sqlite:///athlete.db
Done.
```

Out[106]:

ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event	Medal
1	A Dijiang	M	24.0	180.0	80.0	China	CHN	1992 Summer	1992	Summer	Barcelona	Basketball	Basketball Men's Basketball	None
2	A Lamusi	M	23.0	170.0	60.0	China	CHN	2012 Summer	2012	Summer	London	Judo	Judo Men's Extra-Lightweight	None
3	Gunnar Nielsen Aaby	M	24.0	None	None	Denmark	DEN	1920 Summer	1920	Summer	Antwerpen	Football	Football Men's Football	None
4	Edgar Lindenau Aabye	M	34.0	None	None	Denmark/Sweden	DEN	1900 Summer	1900	Summer	Paris	Tug-Of-War	Tug-Of-War Men's Tug-Of-War	Gold
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	None
6	Per Knut Aaland	M	31.0	188.0	75.0	United States	USA	1992 Winter	1992	Winter	Albertville	Cross Country Skiing	Cross Country Skiing Men's 10 kilometres	None
7	John Aalberg	M	31.0	183.0	72.0	United States	USA	1992 Winter	1992	Winter	Albertville	Cross Country Skiing	Cross Country Skiing Men's 10 kilometres	None
8	Cornelia "Cor" Aalten (-Strannood)	F	18.0	168.0	None	Netherlands	NED	1932 Summer	1932	Summer	Los Angeles	Athletics	Athletics Women's 100 metres	None
9	Antti Sami Aalto	M	26.0	186.0	96.0	Finland	FIN	2002 Winter	2002	Winter	Salt Lake City	Ice Hockey	Ice Hockey Men's Ice Hockey	None
10	Einar Ferdinand "Einari" Aalto	M	26.0	None	None	Finland	FIN	1952 Summer	1952	Summer	Helsinki	Swimming	Swimming Men's 400 metres Freestyle	None

2:- Giving Average value to required fields

Here we are showing a new column “Avg.” for Height and Weight for the players with null values for that.

>>%%sql

alter table athlete_data20 add `Avg. Weight for NULL Values` Number(5);

alter table athlete_data20 add `Avg. Height for NULL Values` number(5);

update athlete_data20 set `Avg. Weight for NULL Values`= 85.5 where Weight is null;

update athlete_data20 set `Avg. Height for NULL Values`=180.0 where Height is null;

```
 * sqlite:///athlete.db
Done.
```

Out[143]:

ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event	Medal	Avg. Weight for NULL Values	Avg. Height for NULL Values
1	A Dijiang	M	24.0	180.0	80.0	China	CHN	1992 Summer	1992	Summer	Barcelona	Basketball	Basketball Men's Basketball	None	None	None
2	A Lamusi	M	23.0	170.0	60.0	China	CHN	2012 Summer	2012	Summer	London	Judo	Judo Men's Extra-Lightweight	None	None	None
3	Gunnar Nielsen Aaby	M	24.0	None	None	Denmark	DEN	1920 Summer	1920	Summer	Antwerpen	Football	Football Men's Football	None	85.5	180
4	Edgar Lindenau Aabye	M	34.0	None	None	Denmark/Sweden	DEN	1900 Summer	1900	Summer	Paris	Tug-Of-War	Tug-Of-War Men's Tug-Of-War	Gold	85.5	180
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	None	None	None

3:- Removing NULL records from table

In [173]:
```
%%sql
delete from athlete_data20 where ID is null
```

```
 * sqlite:///athlete.db
34885 rows affected.
```

Out[173]: []

4:- Giving primary key to the table

Creating a new table with primary key

>>%%sql

CREATE TABLE athlete_data(ID number(10), Name char(30), Sex char(3), Age number(3), Height number(4), Weight number(4), Team vachar(30), NOC varchar(10), Games char(30), Year number(4), Season varchar(15), City char(30),Sport varchar(20),Event varchar(30),Medal varchar(10),`Avg. Weight for NULL Values` Number(5),`Avg. Height for NULL Values` Number(5),primary key(ID), FOREIGN KEY(NOC) REFERENCES noc_data(NOC));

INSERT INTO athlete_data SELECT * FROM athlete_data20;

Out[31]:

cid	name	type	notnull	dflt_value	pk
0	ID	number(10)	0	None	1
1	Name	char(30)	0	None	0
2	Sex	char(3)	0	None	0
3	Age	number(3)	0	None	0
4	Height	number(4)	0	None	0
5	Weight	number(4)	0	None	0
6	Team	vachar(30)	0	None	0
7	NOC	varchar(10)	0	None	0
8	Games	char(30)	0	None	0
9	Year	number(4)	0	None	0
10	Season	varchar(15)	0	None	0
11	City	char(30)	0	None	0
12	Sport	varchar(20)	0	None	0
13	Event	varchar(30)	0	None	0
14	Medal	varchar(10)	0	None	0
15	Avg. Weight for NULL Values	Number(5)	0	None	0
16	Avg. Height for NULL Values	Number(5)	0	None	0

5:- Joining the two tables

Creating a new table from the result of joining the tables

>>%%sql

CREATE TABLE joined_data_athlete AS select a.*,n.notes from athlete_data as a left join noc_data as n on a.NOC=n.NOC;

In [51]:
```
%%sql
CREATE TABLE joined_data_athlete AS select a.*,n.notes from athlete_data as a left join noc_data as n on a.NOC=n.NOC
```

```
 * sqlite:///athlete.db
Done.
```

Out[51]: []

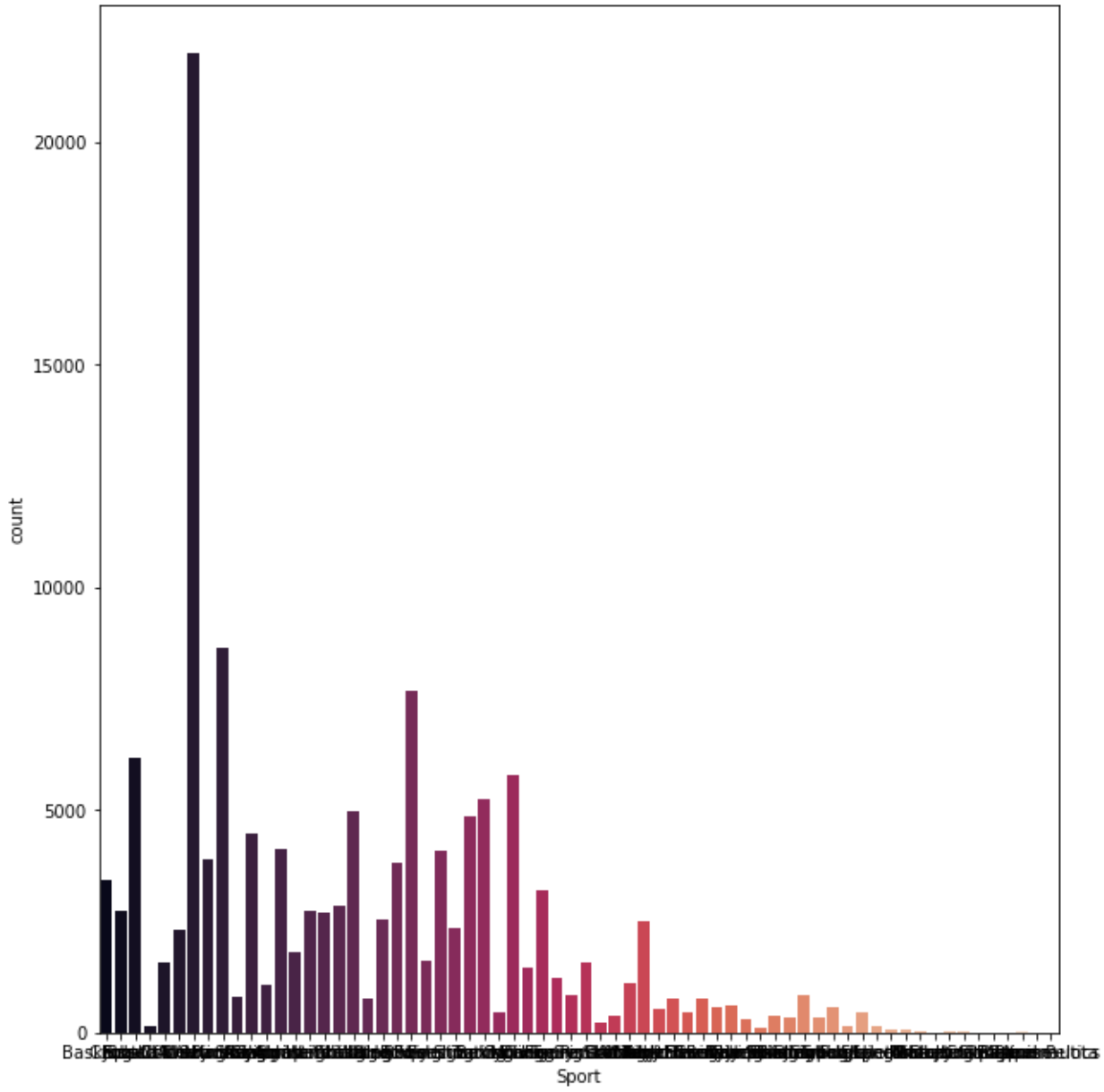
In [55]:
```
%%sql
select * from joined_data_athlete limit 5
```

```
 * sqlite:///athlete.db
Done.
```

Out[55]:

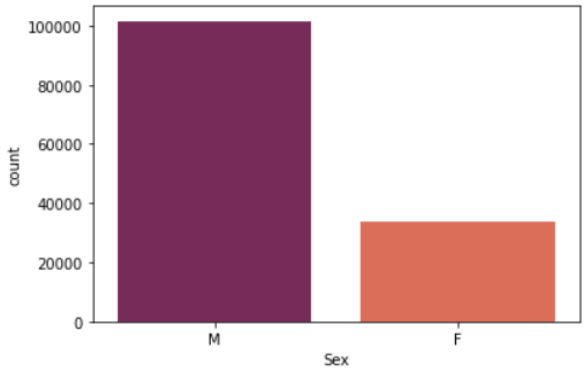
ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event	Medal	Avg. Weight for NULL Values	Avg. Height for NULL Values	notes
1	A Dijiang	M	24	180	80	China	CHN	1992 Summer	1992	Summer	Barcelona	Basketball	Basketball Men's Basketball	None	None	None	AVG.
2	A Lamusi	M	23	170	60	China	CHN	2012 Summer	2012	Summer	London	Judo	Judo Men's Extra-Lightweight	None	None	None	AVG.
3	Gunnar Nielsen Aaby	M	24	None	None	Denmark	DEN	1920 Summer	1920	Summer	Antwerpen	Football	Football Men's Football	None	85.5	180	AVG.
4	Edgar Lindenau Aabye	M	34	None	None	Denmark/Sweden	DEN	1900 Summer	1900	Summer	Paris	Tug-Of-War	Tug-Of-War Men's Tug-Of-War	Gold	85.5	180	AVG.
5	Christine Jacoba Aaftink	F	21	185	82	Netherlands	NED	1988 Winter	1988	Winter	Calgary	Speed Skating	Speed Skating Women's 500 metres	None	None	None	AVG.

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



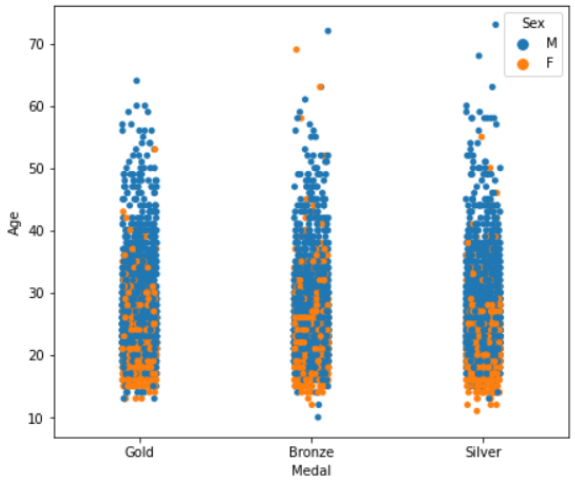
In [35]:
```
# plt.figure(figsize=(10,11))
sns.countplot(x="Sex",data=athlete_noc_combined_csv,palette="rocket")
```

Out[35]: <AxesSubplot:xlabel='Sex', ylabel='count'>



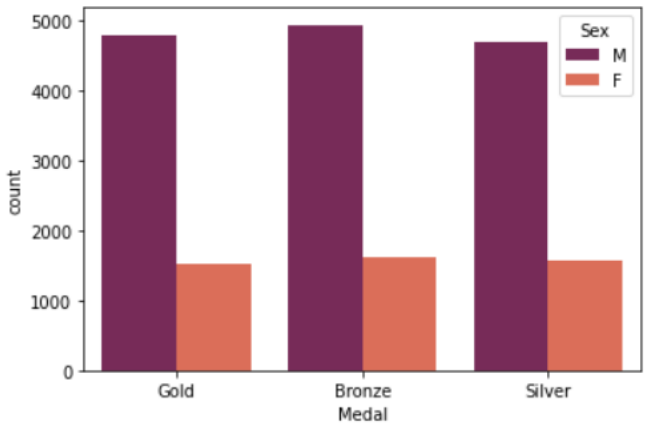
In [42]:
```
plt.figure(figsize=(7,6))
sns.stripplot(x=athlete_noc_combined_csv.Medal,y=athlete_noc_combined_csv.Age,hue=athlete_noc_combined_csv.Sex)
```

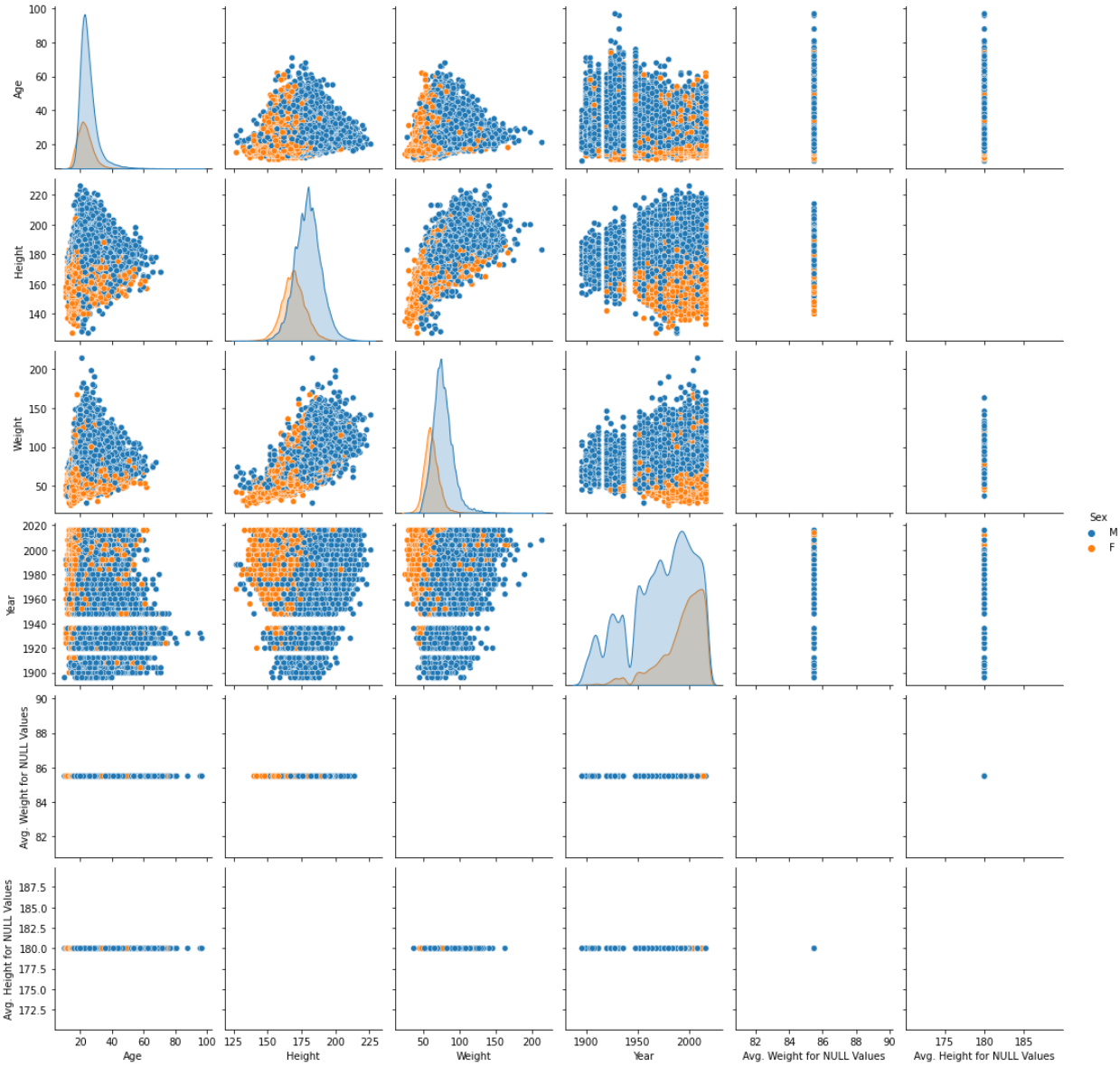
Out[42]: <AxesSubplot:xlabel='Medal', ylabel='Age'>



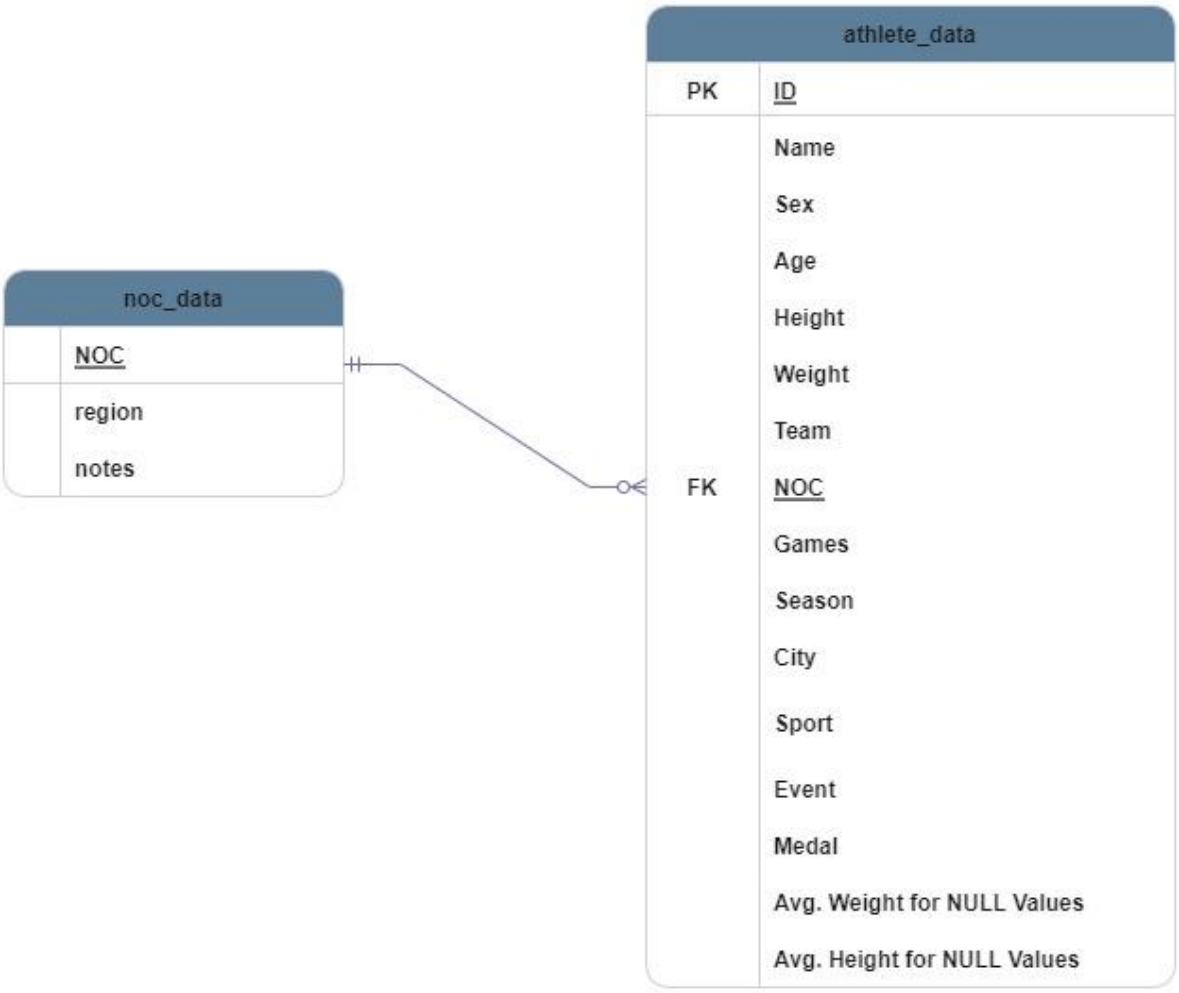
In [44]:
```
sns.countplot(x="Medal",data=athlete_noc_combined_csv,hue="Sex",palette="rocket")
```

Out[44]: <AxesSubplot:xlabel='Medal', ylabel='count'>




Age
Height
Weight
Year
Avg. Weight for NULL Values
Avg. Height for NULL Values
Sex
M
F

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



Description

My Project targets getting past Sports patterns and analysing it. Total team medals. 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.

Questions

- When was the first season ever conducted

- How many total medals were distributed
- Age Distribution
- Which sports were in the First Game
- All of the athlete events conducted
- Which country had highest number of Players

Hypothesis

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