Step 1: Preparing for Your Proposal

Q.1 - Which client/dataset did you select and why?

Olympic Dataset, because I have an interest in sports and the Olympics dataset will provide essential insights on athletes, nations, and different sports categories. Using this dataset I can create a great project.

Q.2 - Describe the steps you took to import and clean the data.

I used the Pandas library to read the data in the CSV file.

To clean the data

- Data Integration: Integrated data from two different files, based on a common variable
- Null values: first I found out the null values, majority of the null values were in the Medal column, based on my understanding those athletes who didn't win that year the medal column was left empty so I replaced the null value with 0.
- Duplicate data: Check for duplicate rows in the final dataframe and removed them, while keeping the first row.

Q.3 - Perform an initial exploration of data and provide some screenshots or display some stats of the data you are looking at.

```python
# data integration
df_final = pd.merge(df_athlete_events,df_noc_regions,'inner',on='NOC')
df_final.head()
```

✓ 0.2s Python

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event	Medal	region	notes
0	1	A Dijiang	M	24.0	180.0	80.0	China	CHN	1992 Summer	1992	Summer	Barcelona	Basketball	Basketball Men's Basketball	NaN	China	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	China	NaN
2	602	Abudoureheman	M	22.0	182.0	75.0	China	CHN	2000 Summer	2000	Summer	Sydney	Boxing	Boxing Men's Middleweight	NaN	China	NaN
3	1463	Ai Linuer	M	25.0	160.0	62.0	China	CHN	2004 Summer	2004	Summer	Athina	Wrestling	Wrestling Men's Lightweight, Greco-Roman	NaN	China	NaN
4	1464	Ai Yanhan	F	14.0	168.0	54.0	China	CHN	2016 Summer	2016	Summer	Rio de Janeiro	Swimming	Swimming Women's 200 metres Freestyle	NaN	China	NaN

```python
df_final.describe()
```

✓ 0.1s

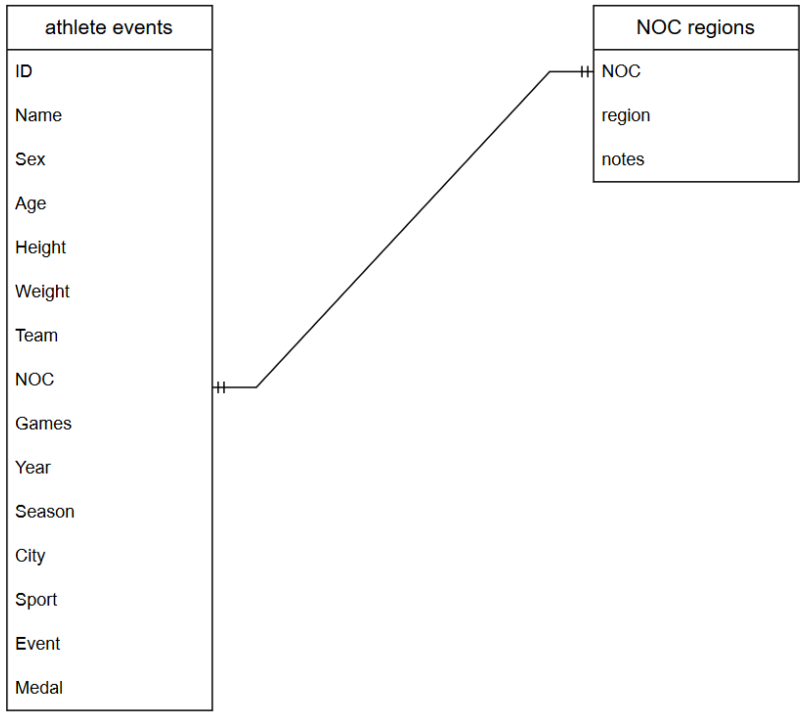
	ID	Age	Height	Weight	Year
count	270767.000000	261305.000000	210684.000000	207982.000000	270767.000000
mean	68229.276832	25.559783	175.344250	70.709523	1978.362297
std	39017.998824	6.392501	10.519556	14.350094	29.884637
min	1.000000	10.000000	127.000000	25.000000	1896.000000
25%	34630.500000	21.000000	168.000000	60.000000	1960.000000
50%	68187.000000	24.000000	175.000000	70.000000	1988.000000
75%	102065.500000	28.000000	183.000000	79.000000	2002.000000
max	135571.000000	97.000000	226.000000	214.000000	2016.000000

```python
df_final.info()
```

✓ 0.2s

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 270767 entries, 0 to 270766
Data columns (total 17 columns):
 #   Column  Non-Null Count   Dtype
---  ------  --------------   -----
 0   ID      270767 non-null  int64
 1   Name    270767 non-null  object
 2   Sex     270767 non-null  object
 3   Age     261305 non-null  float64
 4   Height  210684 non-null  float64
 5   Weight  207982 non-null  float64
 6   Team    270767 non-null  object
 7   NOC     270767 non-null  object
 8   Games   270767 non-null  object
 9   Year    270767 non-null  int64
 10  Season  270767 non-null  object
 11  City    270767 non-null  object
 12  Sport   270767 non-null  object
 13  Event   270767 non-null  object
 14  Medal   270767 non-null  object
 15  region  270746 non-null  object
```

Q.4 - Create an ERD or proposed ERD to show the relationships of the data you are exploring.



Step 2: Develop Project Proposal

Q.1 - Description | Write a 5-6 sentence paragraph describing your project; include who might be interested to learn about your findings. Who might be your audience?

This project is an analysis of the Olympics dataset, here we try to analyze the performance of athletes, which countries are dominating in the Olympics,

and which sport is the strength or weakness of different regions. The target audience is athletes, sports organizations of the nation, fans, marketing agencies,

and brands that are investing in the Olympics for their marketing

Q.2 - Questions | Create 2-3 questions that you want to answer with the data

1- Which are the best-performing athletes in their sports?

2- best-performing nations, in which sports categories they are dominating, and best athletes belong to which nations?

3- demography (age distribution) and gender participation, sports-wise gender distribution, nation-wise gender distributions?

Q.4 - Hypothesis | What are your initial hypotheses about the data?

1- Gender distribution, participation of Females should have increased over the years

2- USA has the highest gold medal count

Q.5 - Approach | Describe in 5-6 sentences what approach you are going to take in order to prove (or disprove) your hypotheses

1- for gender participation analysis over the years, I am going to take the gender column and count their number over the years then I will use a graphical representation of the data to see the variation in the data

2- to understand the best-performing nation I am going to use the nation and count of different medals they secured over the years, then plot a bar chart for comparison of different nations