```
import pandas as pd
import sqlalchemy
import matplotlib.pyplot as plt
import plotly.express as px
```

```
#Make the connection to the dataset in mysql and read the csv files

engine = sqlalchemy.create_engine('mysql+pymysql://root:root@localhost:3306/dbcourse')

athlete_df = pd.read_csv('C:/Users/PC/Desktop/Proyects/athlete_events.csv')
noc_regions_df = pd.read_csv('C:/Users/PC/Desktop/Proyecto/noc_regions.csv')
```

First, I wanted to prove or disprove my hypothesis that USA had a better performance or more medals in Basketball.

```
USA_Basketball = pd.read_sql(
    '''
    SELECT NOC AS Team, COUNT(Medal) AS Medals
    FROM athlete
    WHERE Sport = 'Basketball'
    GROUP BY NOC
    ORDER BY Medals DESC
    LIMIT 20

    ''', con = engine
)
plt.figure(figsize=(10, 15))
fig = px.bar(USA_Basketball, y = 'Medals', x = 'Team', text = 'Medals')
fig.update_layout(uniformtext_minsize = 8, uniformtext_mode = 'hide', xaxis_tickangle=-45)
fig.show()
```

```
<Figure size 1000x1500 with 0 Axes>
```

I see that USA has more medals than other teams but that's not necessary that they had a good performance through the years, so I decided to see how many medals of each category USA and other teams had

```
Basketball_medals = pd.read_sql(
    '''
    SELECT
        NOC AS Team,
        sum(CASE WHEN Medal = 'Bronze' THEN 1 ELSE 0 END) AS Bronze,
        sum(CASE WHEN Medal = 'Silver' THEN 1 ELSE 0 END) AS Silver,
        sum(CASE WHEN Medal = 'Gold' THEN 1 ELSE 0 END) AS Gold
    FROM
        athlete
    WHERE Sport = 'Basketball'
    GROUP BY NOC
    ORDER BY Gold DESC
    LIMIT 10

    ''', con = engine
)
Basketball_medals
```

Out[ ]:

	Team	Bronze	Silver	Gold
0	USA	36.0	24.0	281.0
1	URS	48.0	50.0	48.0
2	ARG	12.0	0.0	12.0
3	YUG	24.0	48.0	12.0
4	EUN	0.0	0.0	12.0
5	EGY	0.0	0.0	0.0
6	ITA	0.0	24.0	0.0
7	JPN	0.0	0.0	0.0
8	CUB	12.0	0.0	0.0
9	RUS	36.0	0.0	0.0

Now it was time to prove or disprove my 2nd hypothesis that Brazil had better performance/more medals in Football

```
Football_medals = pd.read_sql(
    '''
    SELECT NOC AS Team, COUNT(Medal) AS Medals
    FROM athlete
    WHERE Sport = 'Football'
    GROUP BY NOC
    ORDER BY Medals DESC
    LIMIT 20

    ''', con = engine
)
plt.figure(figsize=(10, 15))
fig = px.bar(Football_medals, y = 'Medals', x = 'Team', text = 'Medals')
fig.update_layout(uniformtext_minsize = 8, uniformtext_mode = 'hide', xaxis_tickangle=-45)
fig.show()
```

```
<Figure size 1000x1500 with 0 Axes>
```

With the this graph we could see that Brazil has more medals but now we need to see how many of those are gold, silver and bronze

```
Football_Performance = pd.read_sql(
    '''
    SELECT
        NOC AS Team,
        sum(CASE WHEN Medal = 'Bronze' THEN 1 ELSE 0 END) AS Bronze,
        sum(CASE WHEN Medal = 'Silver' THEN 1 ELSE 0 END) AS Silver,
        sum(CASE WHEN Medal = 'Gold' THEN 1 ELSE 0 END) AS Gold
    FROM
        athlete
    WHERE Sport = 'Football'
    GROUP BY NOC
    ORDER BY Gold DESC
    LIMIT 10

    ''', con = engine
)
Football_Performance
```

Out[ ]:

	Team	Bronze	Silver	Gold
0	USA	12.0	24.0	66.0
1	HUN	16.0	17.0	46.0
2	GBR	1.0	8.0	36.0
3	URS	51.0	0.0	36.0
4	ARG	0.0	34.0	34.0
5	URU	0.0	0.0	31.0
6	CMR	0.0	0.0	18.0
7	GER	69.0	17.0	16.0
8	NOR	30.0	0.0	17.0
9	BRA	34.0	85.0	17.0

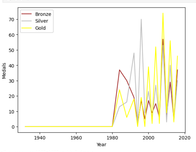
Now it's time for mi 3rd hypothesis, let's see if Chine had a better performance through the years

```
China_Performance = pd.read_sql(
    '''
        SELECT Year,
            sum(CASE WHEN Medal = 'Bronze' THEN 1 ELSE 0 END) AS Bronze,
            sum(CASE WHEN Medal = 'Silver' THEN 1 ELSE 0 END) AS Silver,
            sum(CASE WHEN Medal = 'Gold' THEN 1 ELSE 0 END) AS Gold,
            sum(CASE WHEN Medal IN ('Bronze', 'Silver', 'Gold') THEN 1 ELSE 0 END) AS Total_Medals
        FROM athlete
        WHERE NOC = 'CHN'
        group by Year
        ORDER BY Year DESC
        ''', conn= engine
)

plt.plot(China_Performance.Year, China_Performance.Bronze, color = "brown", label = "Bronze")
plt.plot(China_Performance.Year, China_Performance.Silver, color = "silver", label = "Silver")
plt.plot(China_Performance.Year, China_Performance.Gold, color = "yellow", label = "Gold")

plt.xlabel("Year")
plt.ylabel("Medals")
plt.legend()

plt.figure(figsize=(10, 15))
fig = px.line(China_Performance, x = 'Year', y= 'Total_Medals', text='Total_Medals')
fig.show()
```



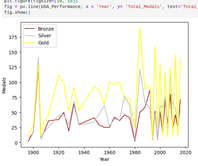

```
<Figure size 1000x1500 with 0 Axes>
```

Now let's compare China with the other 2 teams that had more medals through the years to see which team had a better performance

```
USA_Performance = pd.read_sql(
    '''
        SELECT Year,
            sum(CASE WHEN Medal = 'Bronze' THEN 1 ELSE 0 END) AS Bronze,
            sum(CASE WHEN Medal = 'Silver' THEN 1 ELSE 0 END) AS Silver,
            sum(CASE WHEN Medal = 'Gold' THEN 1 ELSE 0 END) AS Gold,
            sum(CASE WHEN Medal IN ('Bronze', 'Silver', 'Gold') THEN 1 ELSE 0 END) AS Total_Medals
        FROM athlete
        WHERE NOC = 'USA'
        group by Year
        ORDER BY Year DESC
        ''', conn= engine
)
plt.plot(USA_Performance.Year, USA_Performance.Bronze, color = "brown", label = "Bronze")
plt.plot(USA_Performance.Year, USA_Performance.Silver, color = "silver", label = "Silver")
plt.plot(USA_Performance.Year, USA_Performance.Gold, color = "yellow", label = "Gold")

plt.xlabel("Year")
plt.ylabel("Medals")
plt.legend()

plt.figure(figsize=(10, 15))
fig = px.line(USA_Performance, x = 'Year', y= 'Total_Medals', text='Total_Medals')
fig.show()
```

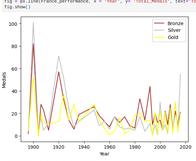



```
<Figure size 1000x1500 with 0 Axes>
```

```
France_performance = pd.read_sql(
    '''
        SELECT Year,
            sum(CASE WHEN Medal = 'Bronze' THEN 1 ELSE 0 END) AS Bronze,
            sum(CASE WHEN Medal = 'Silver' THEN 1 ELSE 0 END) AS Silver,
            sum(CASE WHEN Medal = 'Gold' THEN 1 ELSE 0 END) AS Gold,
            sum(CASE WHEN Medal IN ('Bronze', 'Silver', 'Gold') THEN 1 ELSE 0 END) AS Total_Medals
        FROM athlete
        WHERE NOC = 'FRA'
        group by Year
        ORDER BY Year DESC
        ''', conn= engine
)
plt.plot(France_performance.Year, France_performance.Bronze, color = "brown", label = "Bronze")
plt.plot(France_performance.Year, France_performance.Silver, color = "silver", label = "Silver")
plt.plot(France_performance.Year, France_performance.Gold, color = "yellow", label = "Gold")

plt.xlabel("Year")
plt.ylabel("Medals")
plt.legend()

plt.figure(figsize=(10, 15))
fig = px.line(France_performance, x = 'Year', y= 'Total_Medals', text='Total_Medals')
fig.show()
```




```
<Figure size 1000x1500 with 0 Axes>
```

After seeing all the Teams medals in the categories I mentioned, I wanted to see the medal ratio through the years to see if something could be correlated to teams winning more or less medals

```
Medal_ratio = pd.read_sql(
    '''
    SELECT
    Year,
        CAST(medal_count AS FLOAT) / total_count AS medal_ratio
    FROM
    (
        SELECT
            Year,
            COUNT(*) AS total_count,
            SUM(CASE
                WHEN Medal IS NOT NULL THEN 1 ELSE 0
                END) AS medal_count
            FROM
            athlete
            GROUP BY
            Year
    )newtable
    ORDER BY year DESC
''', conn= engine
)
plt.plot(Medal_ratio.Year, Medal_ratio.medal_ratio, color = "red", label = "ratio")

plt.xlabel("Year")
plt.ylabel("Medals")
plt.legend()
```

Out[ ]:

```
<matplotlib.legend.Legend at 0x13481d77a70>
```

let's take a look at my last hypothesis, that People with age > 35 have more medals thant age < 35

```
AgeAthletes = pd.read_sql(
    '''
    SELECT
    *
    FROM
    (
        SELECT
                COUNT(Distinct Name) as Athletes
            FROM athlete
                WHERE Age > 35
        ) AS Older,
        COUNT(distinct Name) as Younger
FROM athlete
WHERE Age < 35
''', conn= engine
)
AgeAthletes
```

Out[ ]:

	Older	Younger
0	88854	122148

```
YoungerMedals = pd.read_sql(
    '''
    SELECT
        COUNT(Medal) AS Medals_total,
        SUM(Case WHEN Medal = 'Gold' THEN 1 ELSE 0 END) AS Gold,
        sum(CASE WHEN Medal = 'Silver' THEN 1 ELSE 0 END) AS Silver,
        sum(CASE WHEN Medal = 'Bronze' THEN 1 ELSE 0 END) AS Bronze
FROM athlete
WHERE Age < 35
''', conn= engine
)
YoungerMedals
```

Out[ ]:

	Medals_total	Gold	Silver	Bronze
0	36245	12270.0	11877.0	12098.0

```
OlderMedals = pd.read_sql(
    '''
    SELECT
        COUNT(Medal) AS Medals_total,
        SUM(Case WHEN Medal = 'Gold' THEN 1 ELSE 0 END) AS Gold,
        sum(CASE WHEN Medal = 'Silver' THEN 1 ELSE 0 END) AS Silver,
        sum(CASE WHEN Medal = 'Bronze' THEN 1 ELSE 0 END) AS Bronze
FROM athlete
WHERE Age > 35
''', conn= engine
)
OlderMedals
```

Out[ ]:

	Medals_total	Gold	Silver	Bronze
0	2321	778.0	788.0	755.0

Submit 2-3 key points you may have discovered about the data, e.g. new relationships? Aha's! Did you come up with additional ideas for other things to review?

- 1.- I found out that besides Brazil had more medals in Football, they're in 9th place in Teams with more gold medals
- 2.- I found out that China actually won medals after 1984
- 3.- I found that that the medal ratio fluctuated at first and then it stabilized

Did you prove or disprove any of your initial hypotheses? If so, which one and what do you plan to do next?

I prove mi first hypothesis that USA has better performance in Basketball that other teams. My other 3 hypothesis were disproved. In the second hypothesis, brazil had more medals in Football but there are 9 teams that have more gold medals, so we could say that they dont have better performance through the years that other teams. Mi 3rd hypothesis was also disprove because despite China in recent years has become a competitive team, before 1980 they had a bad performance therefore less medals than other teams. In my last hypothesis, the younger athletes that won a medal are the 29% of the total of younger athletes and in the other hand, the older athletes that won a medal are 26% of the total of older athletes, so my hypothesis were disproved

What additional questions are you seeking to answer?

1.- The annual performance of different countries varies in the winter olympics