1.0 Milestone 1: Project Proposal and Data Selection/Preparation.

This notebook refers to the first week of the final project of the Learn SQL Basics for Data Science Specialization.

Project guideline - week 01.

You are a data scientist working for a data analytics company. Your company has explored a multitude of data sources and is tasked with providing key insights that your customers can make actionable. Your manager has asked you to provide some data analysis guidance for one of the company's customers.

In a typical scenario, you would iteratively work with your client to understand the data wanting to be analyzed. Having a solid understanding of the data and any underlying assumptions present is crucial to the success of a data analytics project. However, in this case, you will need to do a little more of the "heavy lifting".

Project Proposal.

For this project i want to know which countries have better performance, which countries are dominant from the past and which ones are not relevant on the last years, and how were the athletes performance in Football and Basketball through the years

Questions

- 1.-Which countries have more representation, and how was their evolution.
- 2.-Which countries have the best performance in the games.
- 3.-Which countries were better performance in Futbol and Basquetball

Description

I think every team might be interest in my findigs so they could see where their weaknesses are and how they improved in the last few games and compared themselves with other teams that are better or worse and make a strategy when it comes to compete with them

Hypothesis

- 1.- USA have more medals/performance in Basketball
- 2.- Brazil have more medals/performance in Futbol
- 3.- China have had better performance through the years
- 4.- People with Age > 35 have received more medals than People with Age < 35

01. Which client/dataset did you select and why?

I selected the Client 3: SportsStats (Olympics Dataset - 120 years of data)

I chose this dataset because sports it's a really interesting topic for me and i have strong knowledge sports

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

For importing the data, I used pandas so I could read the CSV Files and then used to_sql() to store the data in MYSQL dataset I didn't clean the dataset because null values are also important data because with it we could see for example which teams have had participation on the games and didn't we at all

03. Initial exploration of data with some stats of it

In [ ]:
```
import pandas as pd
import sqlalchemy
import matplotlib.pyplot as plt
import plotly.express as px
```

In [ ]:
```
#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/Proyecto/athlete_events.csv')
noc_regions_df = pd.read_csv('C:/Users/PC/Desktop/Proyecto/noc_regions.csv')
```

In [ ]:
```
#Import the data from the csv files to Mysql

athlete_df.to_sql(
    name = 'athlete',
    con=engine,
    chunksize=1000,
    if_exists='replace'
)
noc_regions_df.to_sql(
    name = 'noc_regions',
    con=engine,
    chunksize=1000,
    index=False,
    if_exists='replace'
)
```

In [ ]:
```
athlete_data = pd.read_sql('SELECT * FROM athlete;', con=engine)
athlete_data
```

Out[ ]:

	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	None
1	2	A Lamusi	M	23.0	170.0	60.0	China	CHN	2012 Summer	2012	Summer	London	Judo	Judo Men's Extra-Lightweight	None
2	3	Gunnar Nielsen Aaby	M	24.0	NaN	NaN	Denmark	DEN	1920 Summer	1920	Summer	Antwerpen	Football	Football Men's Football	None
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	None
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
271111	135569	Andrea Ye	M	29.0	179.0	89.0	Poland-1	POL	1976 Winter	1976	Winter	Innsbruck	Luge	Luge Mixed (Men)'s Doubles	None
271112	135570	Piotr ya	M	27.0	176.0	59.0	Poland	POL	2014 Winter	2014	Winter	Sochi	Ski Jumping	Ski Jumping Men's Large Hill, Individual	None
271113	135570	Piotr ya	M	27.0	176.0	59.0	Poland	POL	2014 Winter	2014	Winter	Sochi	Ski Jumping	Ski Jumping Men's Large Hill, Team	None
271114	135571	Tomasz Ireneusz ya	M	30.0	185.0	96.0	Poland	POL	1998 Winter	1998	Winter	Nagano	Bobsleigh	Bobsleigh Men's Four	None
271115	135571	Tomasz Ireneusz ya	M	34.0	185.0	96.0	Poland	POL	2002 Winter	2002	Winter	Salt Lake City	Bobsleigh	Bobsleigh Men's Four	None

271116 rows × 15 columns

In [ ]:
```
noc_regions_data = pd.read_sql('SELECT * FROM noc_regions;', con=engine)
noc_regions_data
```

Out[ ]:

	NOC	region	notes
0	AFG	Afghanistan	None
1	ANO	Curacao	Netherlands Antilles
2	ALB	Albania	None
3	ALG	Algeria	None
4	AND	Andorra	None
...	...	...	...
225	YEM	Yemen	None
226	YMD	Yemen	South Yemen
227	YUG	Serbia	Yugoslavia
228	ZAM	Zambia	None
229	ZIM	Zimbabwe	None

230 rows × 3 columns

In [ ]:
```
#I want to know the age distribution from the athletes

age_distribution = pd.read_sql('SELECT Age FROM athlete', con = engine)

plt.figure(figsize=(10, 15))
fig = px.histogram(age_distribution, x = 'Age', nbins = 30)
fig.show()
```


In [ ]:
```
#I want to know how many athletes every team had

team = pd.read_sql('SELECT Team, Count(Team) AS Team_Count FROM athlete GROUP BY Team ORDER BY Count(Team) DESC LIMIT 20;', con = engine)

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


In [ ]:
```
#I want to know the athletes in Football and Basketball from the top 3 Teams with more Athletes

USA_Football = pd.read_sql(
"""
SELECT distinct(Name), Sex, Age, noc_regions.NOC, Games, Year
FROM athlete
INNER JOIN noc_regions ON noc_regions.NOC = athlete.NOC
WHERE athlete.NOC = 'USA'
AND Sport = 'Football'
ORDER BY Games ASC
""", con = engine
)
USA_Football
```

Out[ ]:

	Name	Sex	Age	NOC	Games	Year
0	Peter Joseph Ratican	M	17.0	USA	1904 Summer	1904
1	Joseph J. Brady	M	NaN	USA	1904 Summer	1904
2	Alexander Cudmore	M	16.0	USA	1904 Summer	1904
3	Louis John Menges	M	15.0	USA	1904 Summer	1904
4	Martin Thomas Dosling	M	17.0	USA	1904 Summer	1904
...	...	...	...	...	...	...
296	Whitney Elizabeth Engen	F	28.0	USA	2016 Summer	2016
297	Kelley Maureen O'Hara	F	27.0	USA	2016 Summer	2016
298	Hope Amelia Solo (-Stevens)	F	35.0	USA	2016 Summer	2016
299	Meghan Elizabeth Klingenberg	F	26.0	USA	2016 Summer	2016
300	Mallory Diane Pugh	F	18.0	USA	2016 Summer	2016

301 rows × 6 columns

In [ ]:
```
France_Football = pd.read_sql(
"""
SELECT distinct(Name), Sex, Age, noc_regions.NOC, Games, Year
FROM athlete
INNER JOIN noc_regions ON noc_regions.NOC = athlete.NOC
WHERE athlete.NOC = 'FRA'
AND Sport = 'Football'
ORDER BY Games ASC
""", con = engine
)
France_Football
```

Out[ ]:

	Name	Sex	Age	NOC	Games	Year
0	Georges Carnier	M	NaN	FRA	1900 Summer	1900
1	Richard Louis Pierre Allemane	M	18.0	FRA	1900 Summer	1900
2	Maurice Macaire	M	18.0	FRA	1900 Summer	1900
3	Maurice Eugne Finysse	M	20.0	FRA	1900 Summer	1900
4	Ren Paul Virgile Gaillard	M	22.0	FRA	1900 Summer	1900
...	...	...	...	...	...	...
207	Eugnie Anne Claudine Le Sommer	F	27.0	FRA	2016 Summer	2016
208	Claire Marie Anne Lavogez	F	22.0	FRA	2016 Summer	2016
209	Camille Anne Frances Abily	F	31.0	FRA	2016 Summer	2016
210	Sakina Karchaoui	F	20.0	FRA	2016 Summer	2016
211	Iodie Ginette Thomis	F	29.0	FRA	2016 Summer	2016

212 rows × 6 columns

In [ ]:
```
Italy_Football = pd.read_sql(
"""
SELECT distinct(Name), Sex, Age, noc_regions.NOC, Games, Year
FROM athlete
INNER JOIN noc_regions ON noc_regions.NOC = athlete.NOC
WHERE athlete.NOC = 'ITA'
AND Sport = 'Football'
ORDER BY Games ASC
""", con = engine
)
Italy_Football
```

Out[ ]:

	Name	Sex	Age	NOC	Games	Year
0	Enrico Zuffi	M	20.0	ITA	1912 Summer	1912
1	Giuseppe Milano	M	24.0	ITA	1912 Summer	1912
2	Franco Bontadini	M	19.0	ITA	1912 Summer	1912
3	Fabio Mario Lodovico Benedi	M	20.0	ITA	1912 Summer	1912
4	Carlo De Marchi	M	22.0	ITA	1912 Summer	1912
...	...	...	...	...	...	...
255	Flavio De Cegie	M	21.0	ITA	2008 Summer	2008
256	Domenico Criscito	M	21.0	ITA	2008 Summer	2008
257	Andrea Coda	M	23.0	ITA	2008 Summer	2008
258	Riccardo Montolivo	M	23.0	ITA	2008 Summer	2008
259	Claudio Marchisio	M	22.0	ITA	2008 Summer	2008

260 rows × 6 columns

In [ ]:
```
USA_Basquetball = pd.read_sql(
"""
SELECT distinct(Name), Sex, Age, noc_regions.NOC, Games, Year
FROM athlete
INNER JOIN noc_regions ON noc_regions.NOC = athlete.NOC
WHERE athlete.NOC = 'USA'
AND Sport = 'Basketball'
ORDER BY Games ASC
""", con = engine
)
USA_Basquetball
```

Out[ ]:

	Name	Sex	Age	NOC	Games	Year
0	Samuel J. "Sam" Balter, Jr.	M	26.0	USA	1936 Summer	1936
1	Jack Williamson Ragland	M	22.0	USA	1936 Summer	1936
2	Willard Theodore Schmidt	M	26.0	USA	1936 Summer	1936
3	Ralph English Bishop	M	20.0	USA	1936 Summer	1936
4	Frank John Lubin	M	26.0	USA	1936 Summer	1936
...	...	...	...	...	...	...
336	Tina Alexandria Charles	F	27.0	USA	2016 Summer	2016
337	Diana Lorena Taurasi	F	34.0	USA	2016 Summer	2016
338	Maya Ayni Moore	F	27.0	USA	2016 Summer	2016
339	Swinoria Delicia Augustus	F	32.0	USA	2016 Summer	2016
340	Kay Alexander Thompson	M	26.0	USA	2016 Summer	2016

341 rows × 6 columns

In [ ]:
```
France_Basketball = pd.read_sql(
"""
SELECT distinct(Name), Sex, Age, noc_regions.NOC, Games, Year
FROM athlete
INNER JOIN noc_regions ON noc_regions.NOC = athlete.NOC
WHERE athlete.NOC = 'FRA'
AND Sport = 'Basketball'
ORDER BY Games ASC
""", con = engine
)
France_Basketball
```

Out[ ]:

	Name	Sex	Age	NOC	Games	Year
0	Robert Calvy	M	24.0	FRA	1936 Summer	1936
1	Ivanne Alphonse Albert Ocorous	M	29.0	FRA	1936 Summer	1936
2	Pierre Casjus	M	26.0	FRA	1936 Summer	1936
3	Georges Carrier	M	25.0	FRA	1936 Summer	1936
4	Fernand Fruithomme	M	20.0	FRA	1936 Summer	1936
...	...	...	...	...	...	...
142	Nwal-Enih Iklyerni	F	28.0	FRA	2016 Summer	2016
143	Latifa Kamba	F	29.0	FRA	2016 Summer	2016
144	Antoine Diot	M	27.0	FRA	2016 Summer	2016
145	Sandrine Gruda	F	29.0	FRA	2016 Summer	2016
146	Isabelle Yacoubou-Dehoui	F	30.0	FRA	2016 Summer	2016

147 rows × 6 columns

In [ ]:
```
Italy_Basketball = pd.read_sql(
"""
SELECT distinct(Name), Sex, Age, noc_regions.NOC, Games, Year
FROM athlete
INNER JOIN noc_regions ON noc_regions.NOC = athlete.NOC
WHERE athlete.NOC = 'ITA'
AND Sport = 'Basketball'
ORDER BY Games ASC
""", con = engine
)
Italy_Basketball
```

Out[ ]:

	Name	Sex	Age	NOC	Games	Year
0	Livio Franceschini	M	23.0	ITA	1936 Summer	1936
1	Sergio Pagnarella	M	24.0	ITA	1936 Summer	1936
2	Ambrogio Bessi	M	21.0	ITA	1936 Summer	1936
3	Giancarlo Martinoli	M	20.0	ITA	1936 Summer	1936
4	Emilio Gavelli	M	30.0	ITA	1936 Summer	1936
...	...	...	...	...	...	...
179	Denis Marconato	M	29.0	ITA	2004 Summer	2004
180	Matteo Soragna	M	28.0	ITA	2004 Summer	2004
181	Gianluca Basile	M	29.0	ITA	2004 Summer	2004
182	Michele Mian	M	31.0	ITA	2004 Summer	2004
183	Massimo Bulleri	M	26.0	ITA	2004 Summer	2004

184 rows × 6 columns

In [ ]:
```
#I want to know how many athletes are by gender

Gender = pd.read_sql(
"""
SELECT Sex, Count(Sex) AS 'Gender'
FROM athlete
GROUP BY Sex;
""", con= engine
)

fig = px.bar(Gender, x = 'Sex', y = 'Gender', color = 'Gender')
fig.show()
```

In [ ]:
```
#Try want to know how many medals have been earned since the first games

Medal = pd.read_sql(
"""
SELECT Medal, count(Name) As Athletes
FROM athlete
WHERE Medal IS NOT NULL
GROUP BY Medal
""", con= engine
)
Medal
```

Out[ ]:

	Medal	Athletes
0	Gold	13372
1	Bronze	13295
2	Silver	13116

In [ ]:
```
from IPython.display import Image

path = 'C:/Users/PC/Desktop/Proyecto/proyecto.drawio.png'
Image(filename=path)
```

Out[ ]: