Milestone 1: Project Proposal and Data Selection/Preparation

Step 1: Preparing for Your Proposal

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

   I have chosen 'Client 3: SportsStats' with the Olympics Dataset for 120 years of data. The fact that I have chosen this client is that I spend a lot of my free time practicing sports, and I would like to get interesting insights from the dataset. in addition, the .csv files are not large and can be easily handled.

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

   First, the data was downloaded and stored locally since the volume of files is not big, and does not require Databricks or several clusters to work with. I have used my own customized VSCode text editor for coding and querying since I am used to it. I have also used Excel to check the integrity of the data and both datasets appear to be OK.

   Second, I have used pandas from Python to read the .csv files, and the built-in to_sql() function to store the data in a MySQL dataset.

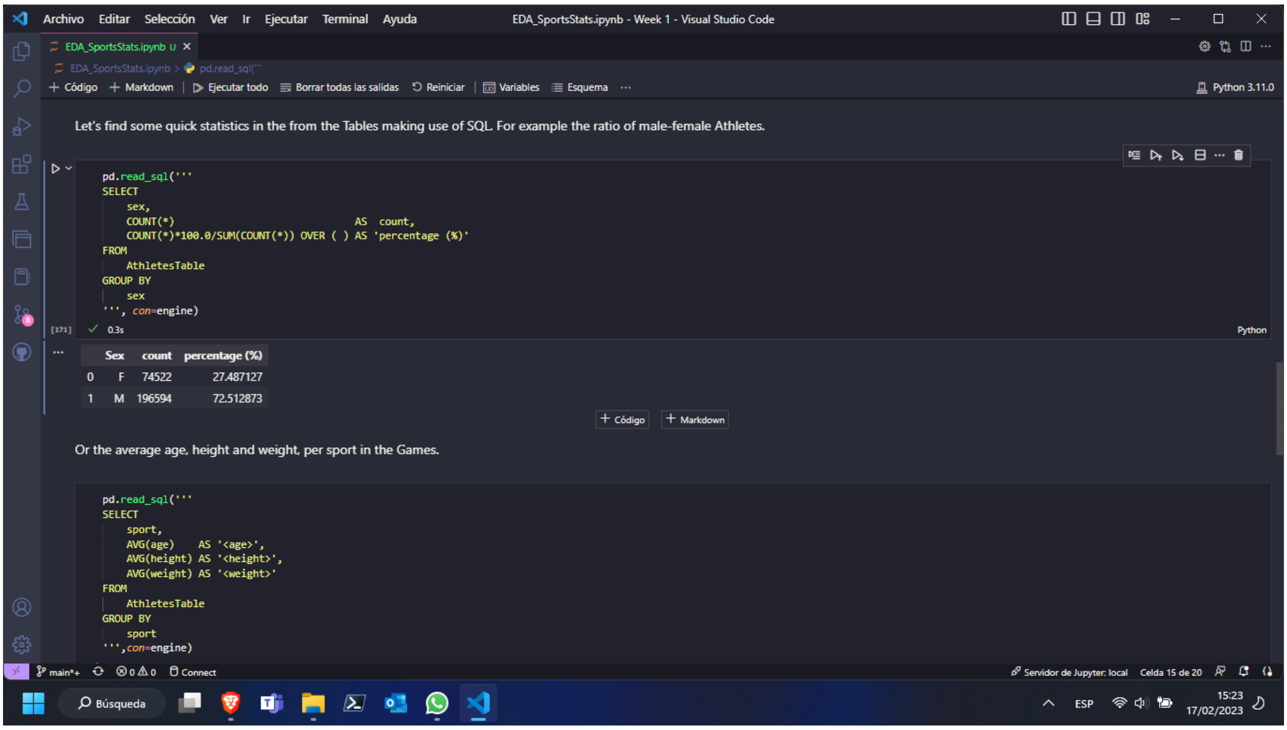
   Third, I checked the amount of NaN or NULL values to know how to deal with them, and remove them or not.

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

   This preliminary or initial EDA has been carried out with Pandas and Pandas SQL libraries to query the data. Other libraries like seaborn and numpy has been used to help the EDA.

   I have performed a quick EDA with simple queries. The athlete_events.csv contains 271116 entries. Some columns can be dropped, since are not relevant to this analysis, for example, the 'Team' column contains some character in the string that should be removed (e.g.: Poland-1). This is more tedious than just using the 'NOC' since it gives us the same information. Additionally, the 'Games' column is not going to be used, and we have this same information with columns 'Year' and 'Season'. Dropping these columns will not reduce the volume of data significantly like notice a speed-up in queries but would keep the data frame cleaner and simpler.

   An initial EDA with basic queries shows that there are 271116 entries or Event_ID, while there are entries that are fully completed (Sex, years, season...) and do not contain missing values, there are some others like Age, Height, and Weight, that show missing values.


Archivo Editar Selección Ver Ir Ejecutar Terminal Ayuda
EDA_SportsStats.ipynb - Week 1 - Visual Studio Code
EDA_SportsStats.ipynb U
EDA_SportsStats.ipynb > pd.read_sql('''
+ Código + Markdown | Ejecutar todo Borrar todas las salidas Reiniciar | Variables Esquema ···
Python 3.11.0
Python
+ Código + Markdown
main*+ ⊗ 0 ⚠ 0 Connect
Servidor de Jupyter: local Celda 15 de 20
Búsqueda
ESP
15:23
17/02/2023

Let's find some quick statistics in the from the Tables making use of SQL. For example the ratio of male-female Athletes.

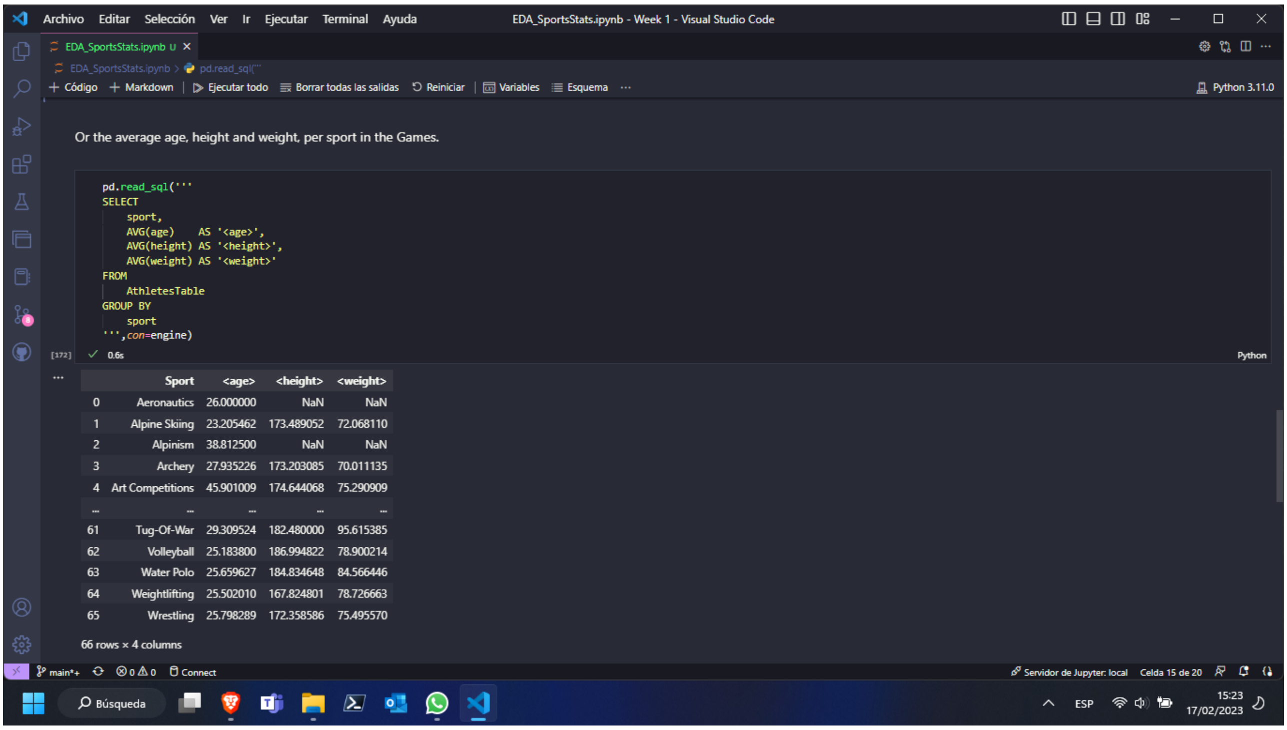
```
pd.read_sql('''
SELECT
    sex,
    COUNT(*)                               AS  count,
    COUNT(*)*100.0/SUM(COUNT(*)) OVER ( ) AS 'percentage (%)'
FROM
    AthletesTable
GROUP BY
    sex
''', con=engine)
```

[171] ✓ 0.3s

	Sex	count	percentage (%)
0	F	74522	27.487127
1	M	196594	72.512873

Or the average age, height and weight, per sport in the Games.

```
pd.read_sql('''
SELECT
    sport,
    AVG(age)    AS '<age>',
    AVG(height) AS '<height>',
    AVG(weight) AS '<weight>'
FROM
    AthletesTable
GROUP BY
    sport
''',con=engine)
```


Archivo Editar Selección Ver Ir Ejecutar Terminal Ayuda
EDA_SportsStats.ipynb - Week 1 - Visual Studio Code
EDA_SportsStats.ipynb U
EDA_SportsStats.ipynb > pd.read_sql('''
+ Código + Markdown | Ejecutar todo Borrar todas las salidas Reiniciar | Variables Esquema ···
Python 3.11.0
Python
main*+ ⊗ 0 ⚠ 0 Connect
Servidor de Jupyter: local Celda 15 de 20
Búsqueda
ESP
15:23
17/02/2023

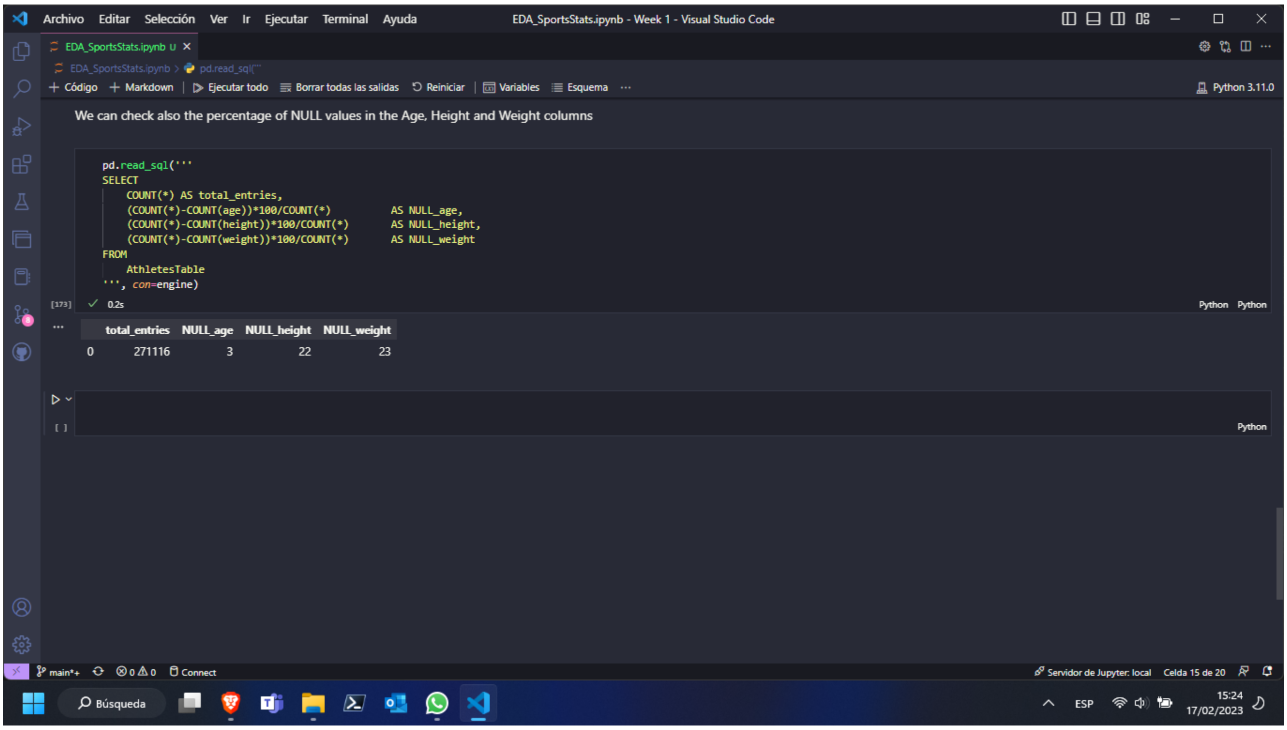
Or the average age, height and weight, per sport in the Games.

```
pd.read_sql('''
SELECT
    sport,
    AVG(age)    AS '<age>',
    AVG(height) AS '<height>',
    AVG(weight) AS '<weight>'
FROM
    AthletesTable
GROUP BY
    sport
''',con=engine)
```

[172] ✓ 0.6s

	Sport	<age>	<height>	<weight>
0	Aeronautics	26.000000	NaN	NaN
1	Alpine Skiing	23.205462	173.489052	72.068110
2	Alpinism	38.812500	NaN	NaN
3	Archery	27.935226	173.203085	70.011135
4	Art Competitions	45.901009	174.644068	75.290909
...	...	...	...	...
61	Tug-Of-War	29.309524	182.480000	95.615385
62	Volleyball	25.183800	186.994822	78.900214
63	Water Polo	25.659627	184.834648	84.566446
64	Weightlifting	25.502010	167.824801	78.726663
65	Wrestling	25.798289	172.358586	75.495570

66 rows × 4 columns


Archivo Editar Selección Ver Ir Ejecutar Terminal Ayuda
EDA_SportsStats.ipynb - Week 1 - Visual Studio Code
EDA_SportsStats.ipynb U
EDA_SportsStats.ipynb > pd.read_sql('''
+ Código + Markdown | Ejecutar todo Borrar todas las salidas Reiniciar | Variables Esquema ...
Python 3.11.0
Python Python
Python
main*+ ⊗0 ⚠0 Connect
Servidor de Jupyter: local Celda 15 de 20
Búsqueda
ESP
15:24
17/02/2023

We can check also the percentage of NULL values in the Age, Height and Weight columns

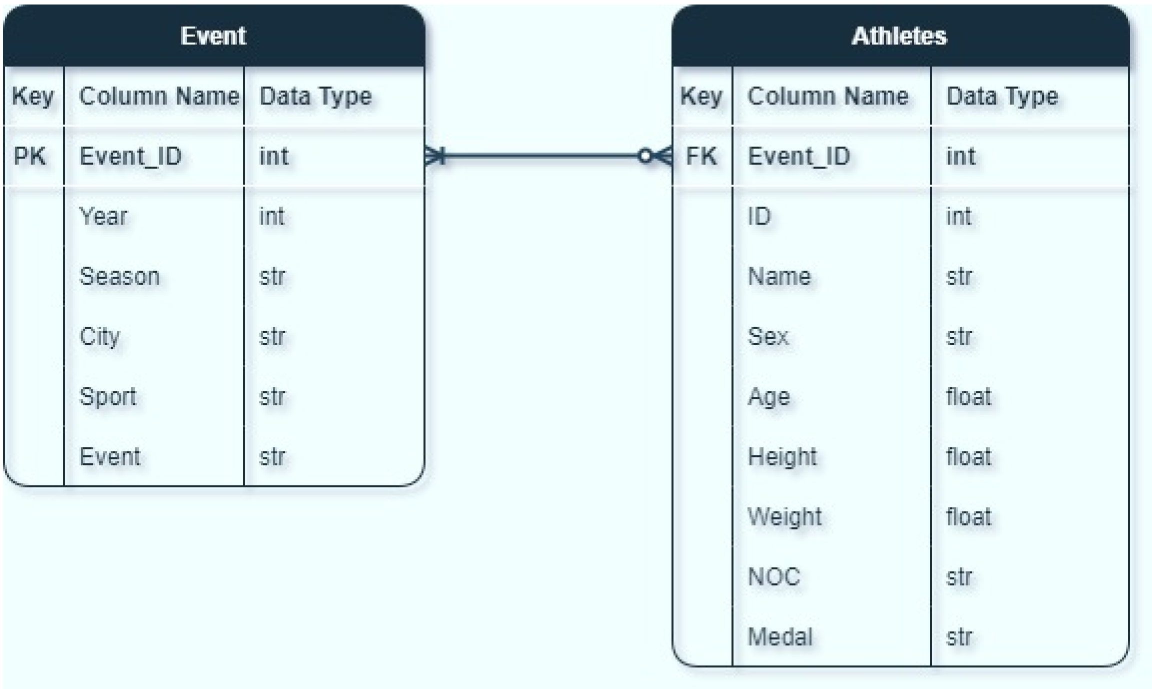
```
pd.read_sql('''
SELECT
    COUNT(*) AS total_entries,
    (COUNT(*)-COUNT(age))*100/COUNT(*)         AS NULL_age,
    (COUNT(*)-COUNT(height))*100/COUNT(*)      AS NULL_height,
    (COUNT(*)-COUNT(weight))*100/COUNT(*)      AS NULL_weight
FROM
    AthletesTable
''', con=engine)
```

[173] ✓ 0.2s

```
   total_entries  NULL_age  NULL_height  NULL_weight
0         271116         3           22           23
```

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

   The ERD shown below was intended for a small relational database, splitting them into two tables, the athletes and the event. Some modifications have been needed, for example, the column 'ID' had no unique values, so it could not be used as a primary key (PK), so a new column "Event_ID" in the 'Event' table has been added as a PK, and as a FK in the 'Atheletes' Table.



Event

Key	Column Name	Data Type
PK	Event_ID	int
	Year	int
	Season	str
	City	str
	Sport	str
	Event	str

Athletes

Key	Column Name	Data Type
FK	Event_ID	int
	ID	int
	Name	str
	Sex	str
	Age	float
	Height	float
	Weight	float
	NOC	str
	Medal	str

Step 2: Develop a Project Proposal

Description

The purpose of this project is to get some insight from the data to obtain several statistics based on athletes during different Olympic events in the last 120 years. The audience could be directed to sports enthusiasts and *followers, or even coaches/trainers might find them useful. This data might be also relevant for Sports media and curiosity channels of communication.*

Questions

- *How relevant is the athletes' age to affect the chance to obtain a medal in the event?*
- *What countries have more chances to get medals, those with more or fewer resources to invest in sports since early years?*
- *How is the Season-countries distribution? Are northern countries more likely to get medals in the Winter Seasons?*
- *Over the years, has the participation of men and women athletes reached equality? the participation of both are more equal in the last decades?*

Hypothesis

- *Countries at higher latitudes have better performance (medals) in Winter Sports*.
- *Female and Male participants tend to be equilibrated over the years.*
- *Developed countries have more medals on their records.*
- *It has to be an age of around 25 years, for the best winning medals.*

Approach

- *Distribution of age and Medals*
- *Distribution of medals and countries*
- *Distribution of men and women over the years*