Project_2 ETL

Group_3

Dakota/ Joyce/ Sean/ Trevor

Data Source

- https://fbref.com/en/comps/12/3239/2019-2020-La-Liga-Stats

ETL process

Extract (Read the data from multiple)

player_standard_stats_df

```
In [25]: #  Import other cvs file
csv_file = "sportsref_20_21.csv"
player_standard_stats_df = pd.read_csv(csv_file, header=1, encoding='maccentraleurope')
player_standard_stats_df.head()
```

Out[25]:

	Rk	Player	Nation	Pos	Squad	Age	Born	MP	Starts	Min	...	xG	npxG	xA	npxG+xA	xG.1	xA.1	xG+xA	npxG.1	npxG+xA.1	Matches
0	1	Sabit Abdulai	gh GHA	MF	Getafe	21	1999	3	0	60	...	0.0	0.0	0.0	0.0	0.00	0.00	0.00	0.00	0.00	Matches
1	2	Marcos AcuÃa	ar ARG	DF	Sevilla	28	1991	30	26	2,330	...	1.1	1.1	2.8	4.0	0.04	0.11	0.15	0.04	0.15	Matches
2	3	Bobby Adekanye	nl NED	FW,MF	Cádiz	21	1999	3	0	36	...	0.2	0.2	0.0	0.2	0.43	0.00	0.43	0.43	0.43	Matches
3	4	Martin Agirregabiria	es ESP	DF,MF	Alavés	24	1996	26	16	1,558	...	0.2	0.2	0.9	1.1	0.01	0.05	0.06	0.01	0.06	Matches
4	5	Joseph Aidoo	gh GHA	DF	Celta Vigo	24	1995	25	14	1,289	...	0.8	0.8	0.6	1.4	0.06	0.04	0.10	0.06	0.10	Matches

5 rows × 33 columns

league_table_df

```
In [3]: tables = pd.read_html(url)
league_table_df = tables[0]
league_table_df.head()
```

Out[3]:

	Rk	Squad	MP	W	D	L	GF	GA	GD	Pts	xG	xGA	xGD	xGD/90	Attendance	Top Team Scorer	Goalkeeper	Notes
0	1	Atlético Madrid	38	26	8	4	67	25	42	86	52.4	32.7	19.6	0.52	NaN	Luis Suárez - 21	Jan Oblak	→ UEFA Champions League via league finish
1	2	Real Madrid	38	25	9	4	67	28	39	84	61.6	36.5	25.2	0.66	NaN	Karim Benzema - 23	Thibaut Courtois	→ UEFA Champions League via league finish
2	3	Barcelona	38	24	7	7	85	38	47	79	78.9	39.6	39.3	1.03	NaN	Lionel Messi - 30	Marc-André ter Stegen	→ UEFA Champions League via league finish
3	4	Sevilla	38	24	5	9	53	33	20	77	50.9	35.0	15.9	0.42	NaN	Youssef En-Nesyri - 18	Yassine Bounou	→ UEFA Champions League via league finish
4	5	Real Sociedad	38	17	11	10	59	38	21	62	60.4	36.6	23.8	0.63	NaN	Alexander Isak - 17	Álex Remiro	→ UEFA Europa League via league finish

ETL process

Transform (Clean and structure data)

Source DB	Source Table	Source column	Source Datatype	Target DB	Target Table	Target column	Target Datatype	Business Rule
Sportsref_20_21.csv	player_standard_stats	Player	string	La_liga_db	player_standard_stats	Player	string	encoding='maccentraleurope'
Sportsref_20_21.csv	player_standard_stats	Nation	string	La_liga_db	player_standard_stats	Nation	string	Split to get last 3 nation abbreviated
Sportsref_20_21.csv	player_standard_stats			La_liga_db	player_standard_stats			Dropped columns
Sportsref_20_21.csv	player_standard_stats	Min	string	La_liga_db	player_standard_stats	Min	string	Replace “,” with””
https://fbref.com/en/comps/12/10731/2020-2021-La-Liga-Stats	League_table_df			La_liga_db	League_table_table			Dropped columns

```
                                                'Pts']]
new_league_table_df.head()
```

Out[5]:

	Rk	Squad	MP	W	D	L	GF	GA	GD	Pts
0	1	Atlético Madrid	38	26	8	4	67	25	42	86
1	2	Real Madrid	38	25	9	4	67	28	39	84
2	3	Barcelona	38	24	7	7	85	38	47	79
3	4	Sevilla	38	24	5	9	53	33	20	77
4	5	Real Sociedad	38	17	11	10	59	38	21	62

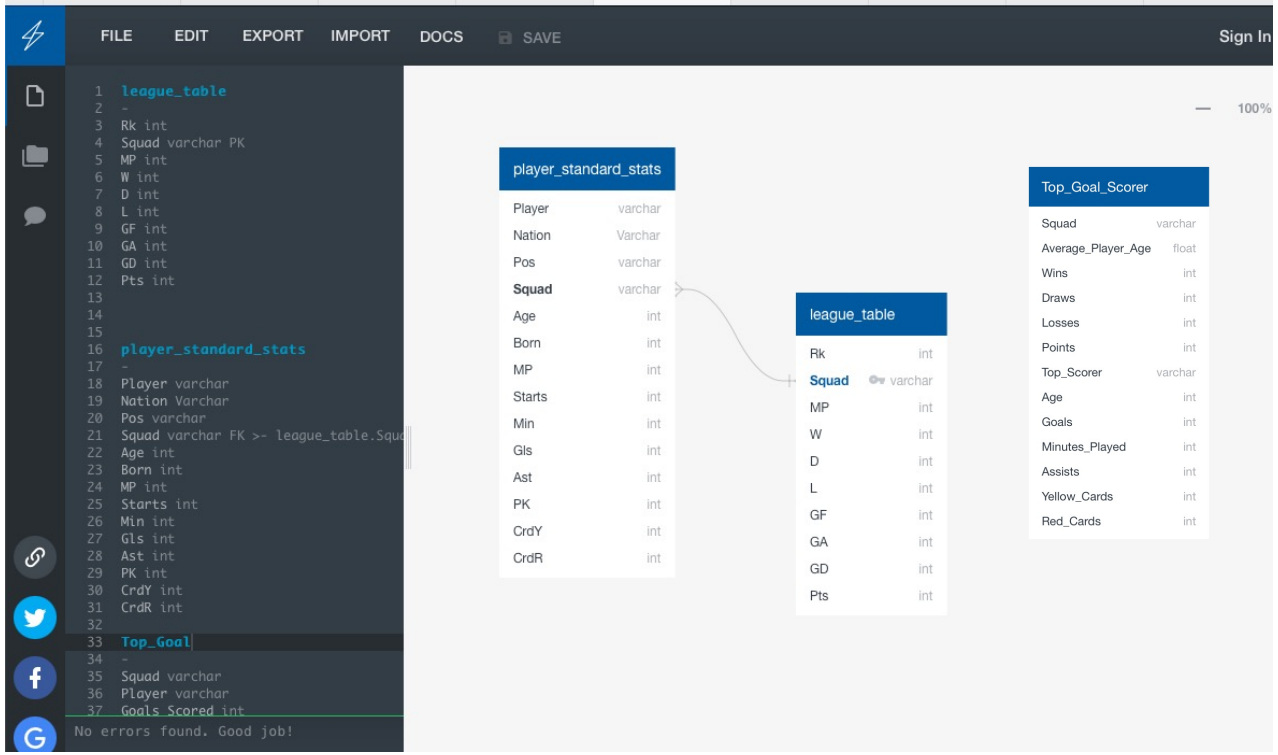
⚽ **new_league_table_df**

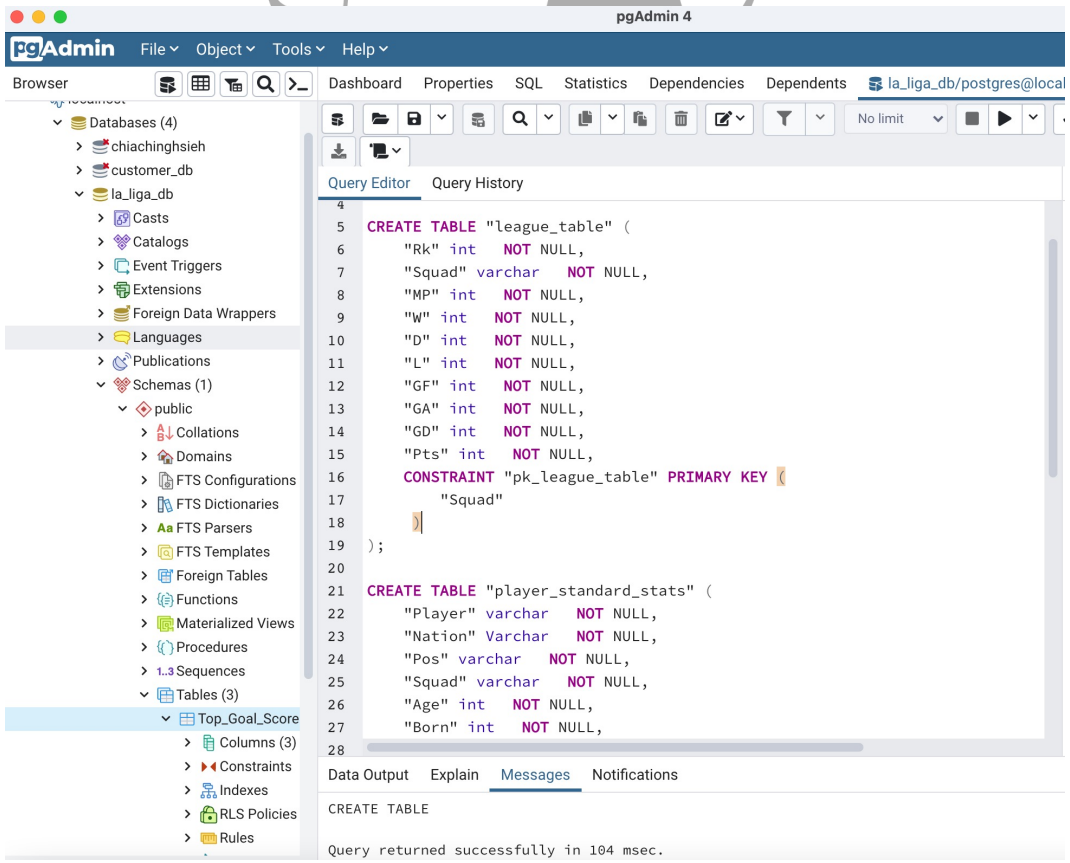
⚽ **new_player_df**

Out[9]:

	Player	Nation	Pos	Squad	Age	Born	MP	Starts	Min	Gls	Ast	PK	CrdY	CrdR
0	Sabit Abdulai	GHA	MF	Getafe	21	1999	3	0	60	0	0	0	1	0
1	Marcos AcuËa	ARG	DF	Sevilla	28	1991	30	26	2330	1	2	0	5	0
2	Bobby Adekanye	NED	FW,MF	Cádiz	21	1999	3	0	36	0	0	0	0	0
3	Martin Agirregabiria	ESP	DF,MF	Alavés	24	1996	26	16	1558	0	1	0	1	0
4	Joseph Aidoo	GHA	DF	Celta Vigo	24	1995	25	14	1289	0	0	0	3	0

ERD & DataBase

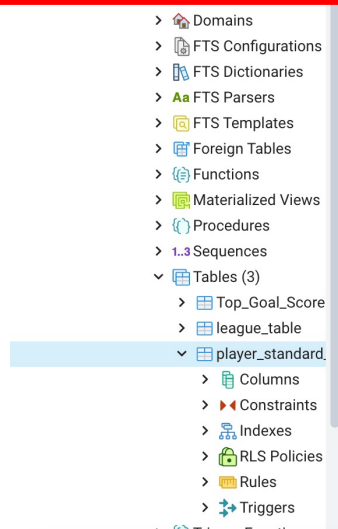


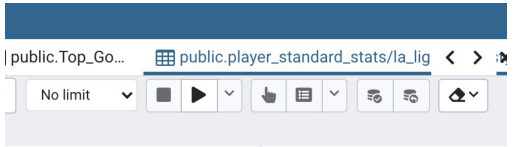


ETL process

Load (Write the data into database for storage)

```
In [12]: # Connect to local database

         from sqlalchemy import create_engine
         rds_connection_string = "postgres:postgres@localhost:5432/la_liga_db"
         engine = create_engine(f'postgresql://{rds_connection_string}')

In [13]: #Check for tables
         engine.table_names()

Out[13]: ['league_table', 'player_standard_stats', 'Top_Goal_Scorer']

In [14]: # Use pandas to load csv into database
         new_league_table_df.to_sql(name='league_table', con=engine, if_exists='append', index=False)
         new_player_df.to_sql(name='player_standard_stats', con=engine, if_exists='append', index=False)
```





Scratch Pad

	Player character varying	Nation character varying	Pos character varying	Squad character varying	Age integer	Born integer	MP integer	Starts integer	Min integer	Gls integer
1	Sabit Abdulai	GHA	MF	Getafe	21	1999	3	0	60	
2	Marcos AcuÑa	ARG	DF	Sevilla	28	1991	30	26	2330	
3	Bobby Adekanye	NED	FW,MF	Cádiz	21	1999	3	0	36	
4	Martin Agirregabiria	ESP	DF,MF	Alavés	24	1996	26	16	1558	
5	Joseph Aidoo	GHA	DF	Celta Vigo	24	1995	25	14	1289	
6	Carlos Akapo	EQG	DF	Cádiz	27	1993	13	10	865	
7	Paul Akouokou	CIV	MF	Betis	22	1997	10	4	378	
8	Jony Álamo	ESP	MF	Elche	18	2001	1	1	73	
9	Jordi Alba	ESP	DF	Barcelona	31	1989	35	34	3025	
10	Raúl Albiol	ESP	DF	Villarreal	34	1985	35	35	3115	
11	Paco Alcácer	ESP	FW	Villarreal	26	1993	27	19	1544	
12	Pedro Alcalá	ESP	DF	Cádiz	31	1989	13	8	909	
13	Rubén Alcaraz	ESP	MF	Valladolid	29	1991	30	25	2165	
14	Iván Alejo	ESP	MF,FW	Cádiz	25	1995	22	6	827	
15	Carles AleÑá	ESP	MF,FW	Getafe	22	1998	22	15	1366	
16	Carles AleÑá	ESP	MF	Barcelona	22	1998	2	0	48	

```
In [15]: # #querying  Join the two table on Squad col
         Join_df=pd.read_sql_query('select * from league_table as lt join player_standard_stats as ps on lt."Squad"=ps."Squad
         Join_df.head()
```

Out[15]:

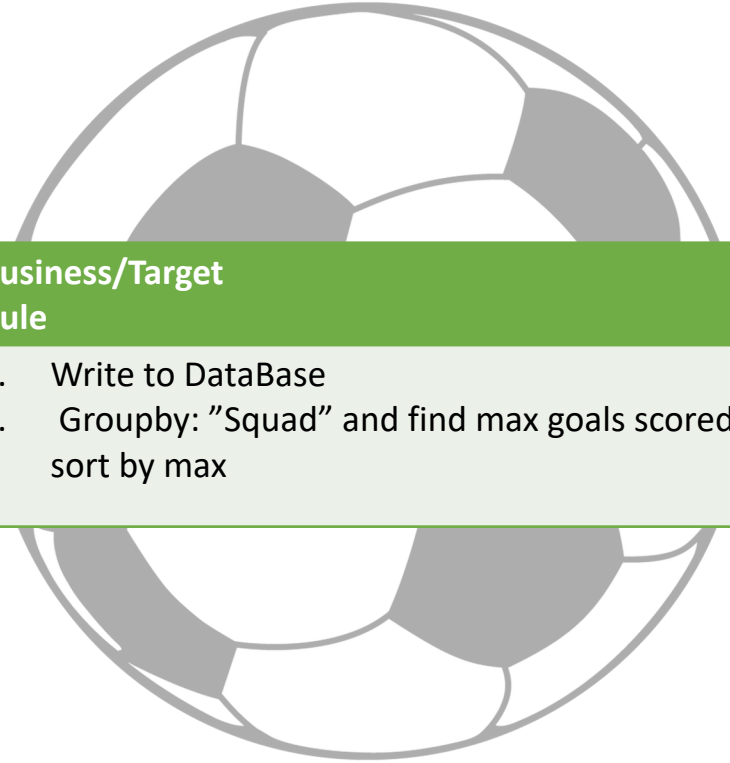
	Rk	Squad	MP	W	D	L	GF	GA	GD	Pts	...	Age	Born	MP	Starts	Min	Gls	Ast	PK	CrdY	CrdR
0	15	Getafe	38	9	11	18	28	43	-15	38	...	21	1999	3	0	60	0	0	0	1	0
1	4	Sevilla	38	24	5	9	53	33	20	77	...	28	1991	30	26	2330	1	2	0	5	0
2	12	Cádiz	38	11	11	16	36	58	-22	44	...	21	1999	3	0	36	0	0	0	0	0
3	16	Alavés	38	9	11	18	36	57	-21	38	...	24	1996	26	16	1558	0	1	0	1	0
4	8	Celta Vigo	38	14	11	13	55	57	-2	53	...	24	1995	25	14	1289	0	0	0	3	0

5 rows × 24 columns

Join Two Table

Join_df=pd.read_sql_query('select * from league_table as lt join player_standard_stats as ps on lt."Squad"=ps."Squad"', con=engine)

Target Table



Source DB	Source Table	Target DB	Target Table	Business/Target Rule
La_liga_db	player_standard_stats &League_table	La_liga_db	Top_Goal_Scorer	1. Write to DataBase 2. Groupby: ”Squad” and find max goals scored then sort by max

```
df4 = Top_player.groupby('Squad')['Age'].agg(['mean’])
df4 = df4.rename(columns={"mean": "Average_Player_Age",})
df4 = df4.reset_index()

df3 = Top_player.sort_values(['Gls'],ascending=False).groupby(['Squad']).head(1)
df3 = df3.set_index('Squad').join(df4.set_index('Squad’))
df3 = df3.round({'Average_Player_Age': 1})
#insert new datafrmae into database new table
df3.to_sql(name='Top_Goal_Scorer', con=engine, if_exists='append', index=True)
```

Table Sample

Squad	Average_Player_Age	Wins	Draws	Losses	Points	Top_Scorer	Age	Goals	Minutes_Played	Assists	Yellow_Cards	Red_Cards
Barcelona	25.1	24	7	7	79	Lionel Messi	33	30	3023	9	4	0
Real Madrid	25.2	25	9	4	84	Karim Benzema	32	23	2894	9	2	0
Villarreal	26.3	15	13	10	58	Gerard Moreno	28	23	2673	7	3	0
Sevilla	26.9	24	5	9	77	Youssef En-Nesyri	23	18	2311	0	2	0
Real Sociedad	23.7	17	11	10	62	Alexander Isak	20	17	2340	2	4	0

Thank you?