TITLE:

Correlation between the popularity sentiment of a soccer player and his influence on the rankings of his soccer club and country.

This project was completed by Rafal, Yusuf and Sri

SCOPE:

The program will deliver a master database on players with information about his affiliation with a soccer club, the country he represents, world ranking represented in the FIFA game, Instagram rankings which is a measure of # of followers and popularity.

The questions that will be answered by this program will be as follows:
1. From the list of top (FIFA game) ranked 200 players who is the most popular as measured by the Instagram popularity sentiment.
2. How many of those popular players helped their respective club hold a place in the top Ten.
3. How many of those popular players helped their respective country hold a place in the top Ten.
4. How many of those popular players helped both club and country hold a place in the top Ten.

The program ingests 4 data sources and eventually creates a normalized database with 6 tables in PostgreSQL.

The following steps were followed:

1. The database schema was defined quickdatabasediagrams.com and the entity relationship diagram is as follows:

2. The export function was used to predefine the database in PostgresSQL .
3. Once the database is ready, the ETL process can run.

EXTRACT:

The following 4 data sources are loaded. . The csv files can be found in the Resources folder

1. World_Player_Master - Acquired from the FIFA game into pandas (https://www.ea.com/games/fifa/fifa-20/ratings/fifa-20-player-ratings-top-100)
2. Top clubs - API extraction from https://www.api-football.com/
3. Top countries - API extraction from https://www.api-football.com/
4. Instagram player popularity sentiment - Web scraping from Instagram (https://www.instagram.com/)

TRANSFORM:

1. Top Clubs:
- Using pandas requests leverage API and extract information on top soccer clubs and their ranking
- Order and sort the list based on the rank.
- Create unique city_id's

2. Top Countries:
- Using pandas requests leverage API and extract information on top soccer clubs and their ranking
- Order and sort the list based on the rank.
- Create unique country_id's

3. World_Player_Master:
- create unique df_player dataframe from the fifa game
- Clean Null values
- Drop unnecessary columns
- Assign unique player_id
- Join based on player_id with player_popularity table to create player popularity ranking column

4. Player_Popularity:
- Scrape Instagram website
- Select top 200 players from Table3 and obtain equivalent popularity sentiment
- Clean extracted data and transfer to dataframe
- Merge with Table 3

5. Player_City_Results:

- Player_City_Results table will now have player_name, world_ranking, instagram ranking and club_name

6. Player_Country_Results:

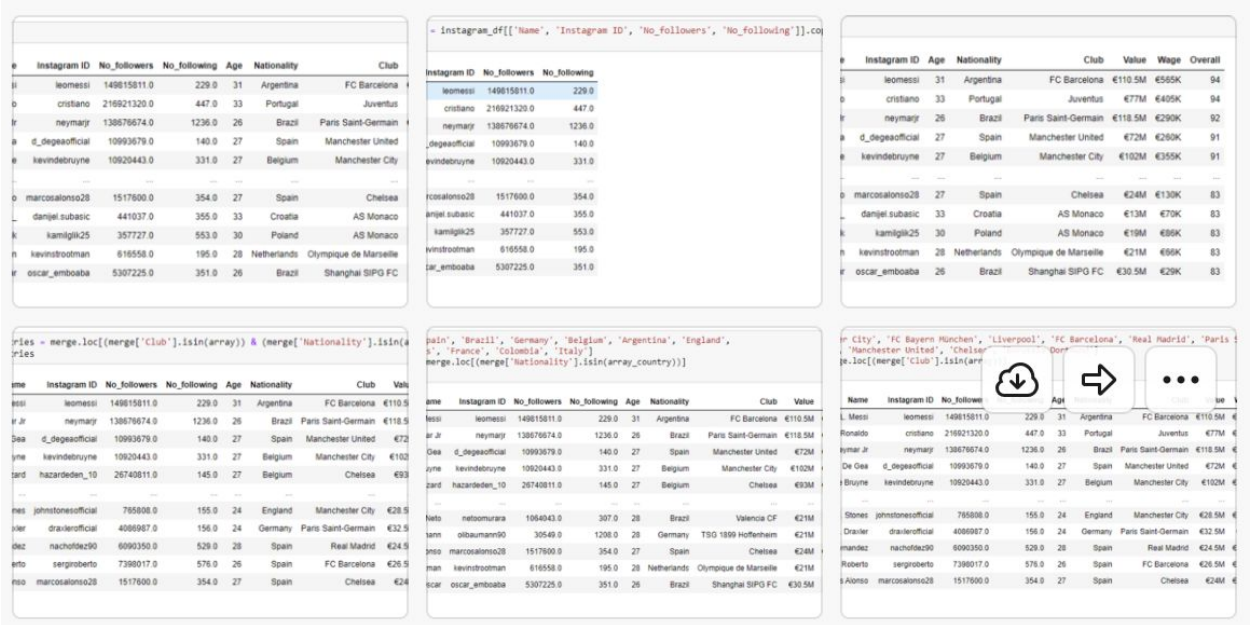
- Player_Country_Results table will now have player_name, world_ranking, instagram ranking and country_name

Finally merge players common to Table 5 and Table 6 and this will be the list of players who have represented both club and country enabling them to hold a top Ten position along with their popularity sentiment.

LOAD:

Since the tables have been pre created in postgreSQL, loading of the data can happen only once (or primary keys will be violated). A test query combining all 6 tables shows it worked correctly.

Please see the snapshot detailing the six tables below:



INFERENCES FROM DATA:

- Some of the top 200, FIFA game ranked players do not do well in the popularity ranking, or they might not be actively using Instagram. Therefore, the number of followers is less than the others.
- Although some players are very successful and very popular, they do not have much influence on the ranking of their national team or club. For example, Cristiano Ronaldo is the second most successful player all over the world, some people think that he is the best. However, his national team which is Portugal is not in top-10 in nationals ranking. This shows that football is not an individual game, but a team and system game.