ELT Project

Soccer Teams and Player

Olivia Nicholson, Osama Ben Ismaeil, Sabrina Seaver



# Extracting the Data Sources

* We used 5 years of data from U.S Major League Soccer Salaries from Kaggle. <https://www.kaggle.com/crawford/us-major-league-soccer-salaries> The Major League Soccer Union provides the salaries of each MLS player every year. The CSV provided us with each soccer club, each player’s first and last name, position, base salary, and guaranteed compensation.
* Our second source of date was scraped from mlssoccer.com (Major League Soccer). <https://www.mlssoccer.com/stats/team> From the Team Stats page we scraped the team name, games played, goals, assists, shots, shots on goal, fouls committed, fouls suffered, offsides, corner kicks, penalty goal kicks, and penalty kick attempts.

## Transformation

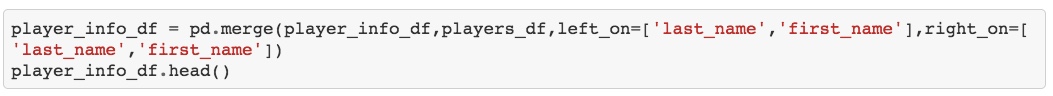
To begin work with the csvs, a dataframe was created with the columns club, last\_name, first\_name, position, base\_salary, and guaranteed compensation. Using a for-loop, the information was appended to the player\_data\_df.

A players table was then created using groupby to sort the players by last and first name to create one row for each player. The table was further cleaned by creating a column that totaled the years played by each player. The index of this table was renamed to create the player id column.



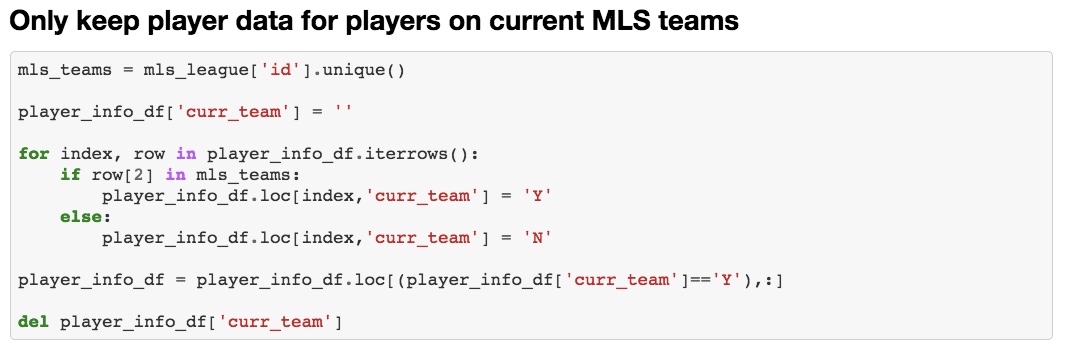
Next, a player salary table was created in order to get every player’s base salary and guaranteed compensation for each year. The salary table was then merged with the player information table so the table would now include base salary and guaranteed compensation for each year. The columns for first and last name were deleted, and each player was now listed by their player id.





The team statistics were put into a pandas dataframe, and a column with scrape date was added to the dataframe. In order to give us a column to link the player information with the team statistics it was necessary to write an Excel spreadsheet that matched the full team name to the team initials. This Excel spreadsheet was merged onto the team statistics dataframe, and the team code column was renamed and became the team id column.



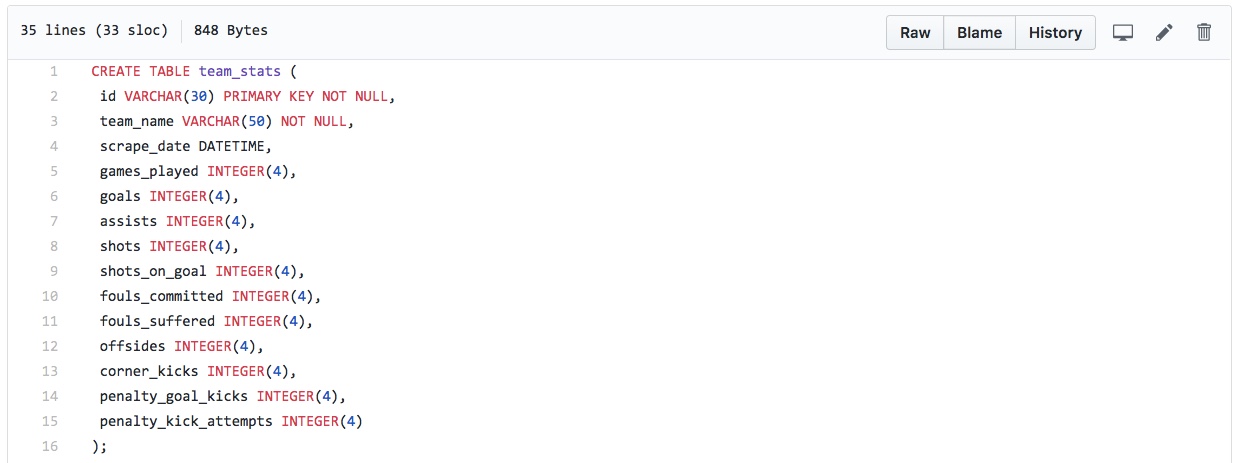
After trying to link the two tables in the database, it became apparent when creating a foreign key that some teams were listed on the player information table that were not on the team statistics table. We identified the players on the player information table that were not on the current team statistics table and dropped those players. 

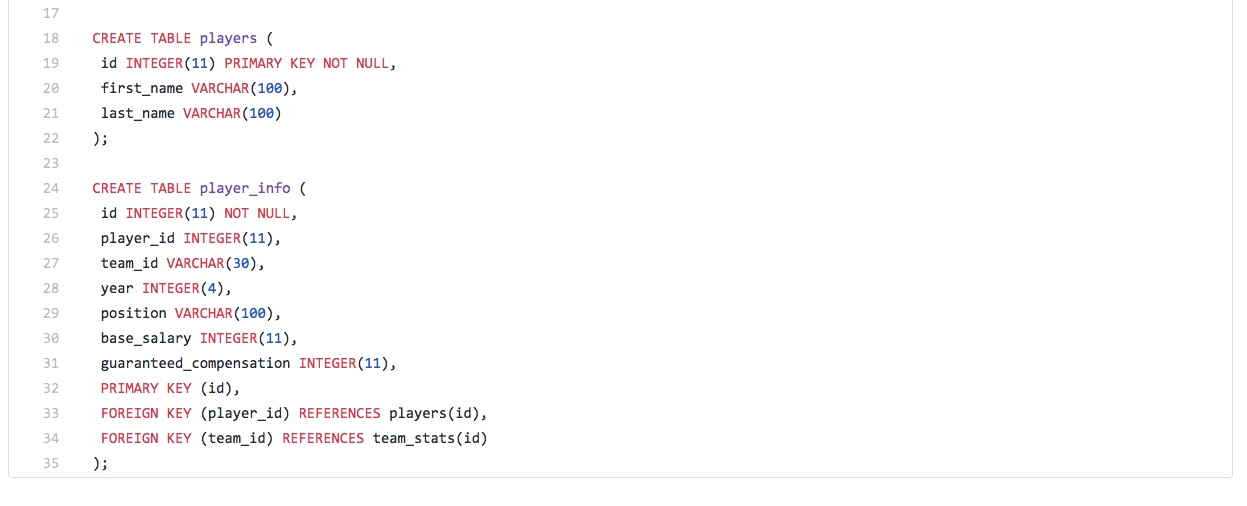
## Loading the Data

The three dateframes were then loaded into MySQL, a relational database.



The team stats table has a primary id key, which references the player info table team id column. Both the players table and player info table reference each other through player id columns, which are primary keys for each table.





## Final Production Database

The three tables in our production are team\_stats, player\_info, and players.

