**ETL Project**

**Team Members:** Marcel Walker, David Romaine, Kevin Cha, Rebekah Vinson, Donald Stegman

The original data sources came from https://www.kaggle.com/justinas/nba-players-data ,and https://www.kaggle.com/drgilermo/nba-players-stats. The data came in the form of CSV files.

The data was transformed by removing an unnecessary asterisk. We created data tables to facilitate speed and access of various nba player information using key value pairs. One of the tables that was created contained player name and a created, unique player id. We set this unique id as the index for the dataframes that would populate these tables. We made another table that contained college name and college id. Then a table that was country of origin name and country id.

The split tables from the all-seasons dataframe was divided into demographic which contained; player\_name, team\_abbreviation, age, player\_height, player\_weight, college, country, draft\_year, draft\_round, draft\_number.

Another table that contained statistics; player\_name, gp, pts, reb, ast, net\_rating, oreb\_pct, dreb\_pct, usg\_pct, ts\_pct, ast\_pct, season.

The reason behind this split is because of the logical choice when referring to a player. Analyst would most likely be referring to one or the other at a given time.

After this we loaded in another csv called “season 78” which contained player name and win share data by season. We preserved this dataframe because it only contained information relevant to this data. However, we did change the column names to lower case versions of themselves so that the ORM could read the column names into the database without errors. Next, we proceeded to divide “all season” into smaller dataframes. We called these dataframes; demographic, team, stat, season78, country, college, and player.

We checked all of our table variables to make sure they were of the correct data type(int, float, or varchar(object)).

Lastly, the data was loaded to pgadmin by using a database connection driver called psycopg2. Because player id was a created column, and our data reached back all the way to 1978 we returned NaN values. To avoid issues with this we removed all “NOT NULL” attributes from the sql schema. In addition, our original ERD did not end up reflecting the changes we made during development, therefore while it has been included it should be noted that it is no longer accurate. The ERD still served as a valuable outline to begin production. Finally, we ran several queries to ensure our data was uploaded correctly.