ETL Assignment

As we gave thought to our Project #2 we decided to look at NBA data and focus on player data comparisons and to look at metrics that would allow us to compare current NBA players to past ones.  We decided to limit past players to Hall of famers, and MVP’s(most valuable players per season), along with the 2018-2019 season player roster.

**Extract:**

For our data sources we scraped Wikipedia for Hall of Fame inductees and MVP’s to compare against the website NBA reference, which had the actual player statistics we were looking for.  The website NBA reference had CSV files split up for individual players. We then created a ‘For loop’ to scrape individual player CSV’s and joined them together, along with their stats from NBA reference.

**Transform:**

We decided on a relational database to be able to compare players. Using MySQL, we joined all our player files to to end up with a CSV, that would eventually become a pandas data frame. We cleaned up data by only using seasons defined by data(some of the data included year ranges and career). We came across some players with special characters in their names which caused Python issues, so those cases were done manually using find and replace in Excel. Some of the players in our datasets belonged to other leagues, so we dropped every league that wasn’t in the NBA. In order to clean up other areas to use not nulls to eliminate seasons that were not numbers, and eliminated cumulative seasons(that were meant to be overall stats). Lastly, we dropped all player data for the 2019-20 season since we are looking at cumulative season stats, and those games have yet to be played.

**Load:**

For our final product we will have a dataframe with MVP players, hall of famers(1988-present), and the roster of players for the 2018 season.  With our ultimate goal for Project#2 to compare each player based on their stats.