**ETL Project (Albert Lee and Ryan Chang)**

**Extract: original data sources and how the data was formatted**

We downloaded four CSV files from Kaggle related to baseball data. Each file contains different sets of information. The first file stores the data for every at bat statistics from 2015-2018 MLB seasons. The second file stores the data for the games that took place in that same timeframe. The third file stores the data for the pitches that were thrown for every game in the same timeframe. The fourth file stores the information for each player that played in a game in that same timeframe. We imported each table into pandas dataframe. Using BeautifulSoup, we scraped www.baseball-reference.com to pull the weight of each player into the player name dataframe.

**Transform: cleaning or transformation required**

Using the pattern we saw in how the player names are transformed into a specific url link, we created the url link for each player name. Namely, we took the first five characters of the last name and the first two characters of the first name with “01” as the string to add to the end of the common url that started with “www.baseball-reference.com/”. If these url links had a successful connection (200), then we created a column with a value of 1. If it did not have a successful connection, then we inserted a value of 0 in that column. Once we ran the query for each player name, we still kept all the player names rows so that we don’t have problems with the foreign/primary key matching later on in MySQL.

**Load: final database, tables/collections, and why this was chosen**

We began the load process by first creating the 4 tables in MySQL, so that we can load the pandas dataframes into MySQL. We created the primary key for the following 3 tables: atbats, games, and player names. We did not create a primary key for pitches because the at-bats id is not unique in the data. Then we exported all of our pandas dataframes into MySQL using the pd.to\_sql script. Then within MySQL, we assigned the foreign keys so that the data are linked between the primary keys and foreign keys in the 4 different data tables (see the EER diagram on the next page for further details).

