ETL Project Summary Report

For the extraction step, we found three separate relating CSV data sources on Kaggle which contained MLB Pitch data from 2015-2018. We chose this data because it was a subject matter we were both familiar with and felt that would make the project easier working with information we had an understanding of. The file names were games.csv which contained specific data on each individual game through the four season (venue, weather, umpire info, etc.), atbats.csv which contained info for each specific at-bat (inning, outs, pitcher/batter handedness) and player\_names.csv (which gave player names that correlated with an id on the other csv files).

We transformed the three data sets by using two separate joins. First, we changed the “batter\_id” index in the atbats.csv to “id” so it would align better for upcoming join. Next, we joined player\_names.csv and atbats.csv on the “id” index using the pd.merge function in Pandas. Then, we joined this new data frame with the games.csv on the “g\_id” index using the pd.merge function in Pandas as well. After this we had all of our information aggregated into one data set.

We then loaded the data frame to a Mongo database using Pandas through the local host connection. We chose this database because simply it was the method we felt most comfortable with using and accomplishing the task successfully.