**Data Summary**

In 1995, Sean Lahman, an author and journalist, made a database of extensive baseball data freely available to the public. This collection of data has been ongoing and will be used in this analysis.

"[Lahman's] database contains pitching, hitting, and fielding statistics for Major League Baseball from 1871 through 2019. It includes data from the two current leagues (American and National), the four other "major" leagues (American Association, Union Association, Players League, and Federal League), and the National Association of 1871-1875."

This database contains the following main tables: People, Batting, Pitching, and Fielding. But also includes a collection of secondary tables: Teams, Post-Season play, Awards, Hall of Fame, AllstarFull, Managers, FieldingOF, ManagersHalf, Salaries, Appearances, Schools, and CollegePlaying. A more thorough reading of this database can be found at [Sean Lahman's website](http://www.seanlahman.com/files/database/readme2017.txt).

All data concerning forecasted (variables preceded by m), or ex ante results were obtained from BaseballGuru.com. A website where sabermetricians, baseball historians, and other experts can submit an analysis concerning baseball for the public to read. The website provides a yearly batting and pitching statistics for the years 2005 through 2021. The target audience for these forecasts are fantasy baseball players. Even though, this forecasted data set is not of the same caliber as predictions made by professional teams, it relies on the same forecasting methodology where AB is the most difficult to predict and most influential factor in all other batting metric forecasts. Baseball Guru’s forecasts are based on data from the last three years with recent years carrying more weight. Forecasts are not performed for players without MLB experience leading to substantial differences in the number of player observations per year. Requiring the normalization of ex ante and ex post data in order for accurate comparisons to be made.

The data sets contain player level data with the variables of interest mAB (forecasted at bats), mAVG (forecasted batting average), mH (forecasted hits), mOBP (forecasted on base percentage), and mSLG (forecasted slugging percentage). The data sets from baseballguru.com also contains the same player identification as in Sean Lahman’s datasets allowing for an easy merge of data. These variables will be compared to the player level data obtained from the Sean Lahman data set. The Lahman data set contains the variables of interest AB (at bats), H (hits), 1B (singles), 2B (doubles), 3B (triples), HR (homeruns), HBP (hit by pitch), SF (sacrifice fly) and BB (walks). These extra variables are required from the Lahman data set because slugging percentage and on base percentage are not explicitly supplied but have been calculated within the Lahman data set.

In order to adeptly analyze ex post verse ex ante baseball offensive statistics, special attention will be placed on analyzing the at bats (AB) statistics. The metric AB is considered one of the most difficult of metric to predict in baseball due to its multitude of confounding factors. This metric is also a key component of all forecasted batting statistics making its impact on the overall batting predictions extensive. By observing the accuracy of these ex ante verse ex post AB statistics, an estimation of the overall accuracy of forecasted batting statistics can be obtained.