Introduction

Sean Layman has done a fantastic job at collecting raw data collected from 1871 to 2018 of major league baseball games. All atomic data elements needed for the calculation of OPS (on base plus slug), OBP (on base percentage) and SLG (slugging) are available. This project requires OPS for the analysis which will be performed. The database has copyright 1996-2018 by Sean Lahman. I have read the license agreement which is licensed under Creative Commons Attribution and will not restrict me from using this data. The raw data needed will be from the 1954 to 2018.

As part of this Capstone Project, there are calculations that I must make based upon the Layman data. Ultimately, the OPS (on base plus slugging) is required for my analysis.

The details of the equations are as follows:

OPS = OBP + SLG

OBP = (H + BB + HBP) / (AB + BB + SF + HBP)

SLG = TB / AB

Where:

H – total number of hits of a player

BB – total number of walks (base on balls) of a player

HBP – total number of times the player was hit by a pitch

AB – total number of plate appearances (times at bat) by the player

SF – total number of sacrifice flies of a player

TB – is the total bases and is a weighted sum ( 1 for single, 2 for double, 3 for triple, 4 for HR).

So, TB is ∑ ( (nSingle \* 1) + (nDoubles \* 2) + (nTriple \* 3) + (nHomeRuns \* 4) ) where nSingles is the number of singles, nDoubles is the number of doubles, nTriple is the number of triples and nHomeRuns is the number of home runs.

OPS – on-base plus slugging

OBP – on-base percentage

SLG – slugging percentage

NOTE: all statistics are taken over a period of time (typically a year)

(source : Wikipedia)

**Data Wrangling**

The following data wrangling steps were followed to assess the data model and the quality of data.

1. Data Source and Data Model Assessment

The current data model consists of 17 entities as follows:

AllStarFull.csv – all players who played in the all-star games (starts at 1933.

Appearances.csv – statistics on player appearances with various counters such as number

of games where player batted, played 1st base etc.

AwardsManagers.csv – awards information given to managers.

AwardsPlayers.csv – awards information given to players (eg. MVP or Cy Young Award.)

AwardsShareManagers.csv = statistics on voting counts for managers.

AwardsSharePlayers.csv = statistics on voting counts for players.

Batting.csv – by year information on batting statistics for each player.

BattingPost.csv – post season (playoff) batting statistics.

CollegePlaying.csv – colleges players attended.

Fielding.csv – fielding statistics by player by year (put outs, errors etc)

FieldingOF.csv - counts where outfielder played LF, CF and RF

FieldingPost.csv – fielding statistics in the post season.

FieldingOFsplit – fielding statistics by outfield position.

HallOfFame.csv – players inducted into baseball’s hall of fame.

HomeGames.csv – home game statistics including total games played and fan attendance.

Managers.csv – information about individual team managers and teams they managed.

ManagersHalf.csv – managers appeared in partial year.

Parks.csv – name and location information on ball park stadiums.

People.csv – baseball player information such as name, DOB, and other BIO information.

Pitching.csv – pitching statistics for pitchers in baseball.

PitchingPost.csv – post season pitching statistics for pitchers in baseball.

Salaries.csv – salary history information for players.

Schools.csv – schools that players attended.

SeriesPost.csv – count statistics on how many world series each player won.

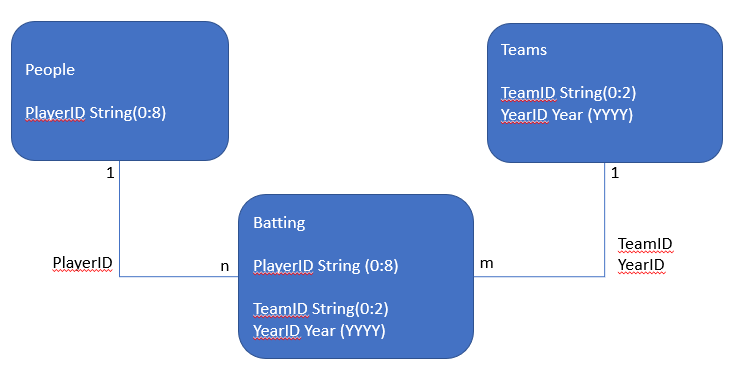
Teams.csv – yearly statistics and standings for teams.

TeamsFranchises.csv – team franchise information.

TeamsHalf.csv – split season data for teams due to 1981 player strike.

After assessing the data model, it appears that three entities are required for my analysis: Batting entity, People entity and Teams entity. All of the columns of the entities were not required. From Batting, 'playerID', 'yearID', 'teamID', 'G', 'AB', 'H', '2B', '3B', 'HR', 'SF', 'BB', 'HBP' were used. From People (players), 'playerID', 'birthYear','birthMonth', 'birthDay', 'nameFirst', 'nameLast', 'debut', 'finalGame' were used. At this time, no team level data was incorporated or is required.

As the project unfolds, additional entities may be added to the model. The current entity relationship diagram is as follows:

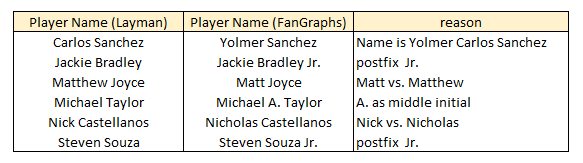


1. Data Quality Assessment.
2. Missing information
   1. Created a DataFrame where at least one necessary column had a NaN value.
   2. The decision was made to not try to fill the NaN cell with a value, rather, it was decided to do analysis and see how many years of data could be used by avoiding years of data with NaN values.
   3. There were two main areas of missing data in the critical Batting entity: SF (sacrifice flies) and HBP (hit by pitch) which are required for OPS calculation. According to wikipedia, “Batters have not been charged with a time at-bat for a sacrifice hit since 1893, but baseball has changed the sacrifice fly rule multiple times. The sacrifice fly as a statistical category was instituted in 1908, only to be discontinued in 1931. The rule was again adopted in 1939, only to be eliminated again in 1940, before being adopted for the last time in 1954.” Not being counted as an at bat prior to 1954 and not being tracked would skew the OBP denominator portion of the calculation.
   4. Based upon NaN data, the analysis showed from 1954 onward the Layman data had all required columns in order to do the proper analysis. So, data prior to 1954 will be ignored.
   5. Followed same process with People (players) entity and it also showed that after 1954 all data was present that was required for project.
3. Referential Integrity between entities
   1. Data was one-to-many without loss of information when joining People to Batting. Data was merged as inner joins between the Batting entity and the People entity without loss of data. See diagram above for relationships. At this point, no need to join Teams to Batting.
   2. People and Teams entities needed to have no redundancy, and analysis on primary keys (identifying attributes) was performed. The analysis showed the playerID for People and yearID and teamID for Teams were primary keys and had foreign keys in the Batting entity.
4. Additional Required Columns
   1. There were additional data columns required for the analysis as follows:

Batting Entity: ‘1B’ number of base hits, ‘TB’ total bases, ‘SLG’ slugging, ‘OBP’ on base percentage and OPS on base plus slug were added to the DataFrame. Added average yearly at bats (avg\_yrly\_AB) and years played (years\_played) columns to Batting entity which will be used later during statistical analysis.

People Entity: added ‘birthdate’ which will be used to calculate players age. Model had three integers: year, month and day (2001 11 30). Once birthdate was created from three columns, dropped year, month and day. Added playername. Model had first name and last name in separate columns. Created playername as first name + last name (eg. Hank Aaron).

1. Data Quality of existing data
   1. Datatype Conversions – there were various data type conversions that were done. Most were floating point which were converted to int64 (SF and HBP). Two columns were converted from strings to datetime values in the People entity (debut and finalGame).
   2. nameFirst (first name of player) had space issues. For example, J.D. Martinez was J. D. Martinez. Got rid of spaces in first name. Made comparisons to FanGraphs (which did not have spaces) much easier because I had to join on player name between the two data sets of information.
   3. Statistics collected by player by year had some issues. Players could be traded mid-year and therefore have multiple records for player by year. Since we don’t care which team the player was on, performed a group by to sum up data to a single player by year.
   4. Filtered batting data based upon the average yearly at bats of a player. We want to look at starters or players who played a significant amount of time. Initially the chosen number is 300 at bats on average per year for a player over their entire career in MLB (Major League Baseball). This average is arbitrary and can easily be changed.
   5. Did an independent automated validation of the data by accessing one year of FanGraph baseball batting statistics and comparing all required batting statistics with the Batting entity. There were 261 players compared between Layman and FanGraphs. Other than the anomalies below, all player’s statistics were exact comparisons. The results of the anomalies were as follows:



The player names were important because that was the only way to join the two entities (Layman Batting and FanGraph data) together using merge on DataFrame. The statistical data for the players in the chart above is correct. Names were inconsistent which there is no need to fix.

In addition, one player Manual Margot had a difference of .001 for OBP when comparing Layman to FanGraphs which is not enough of a difference to be concerned with. Also, Ender Inciarte played one more game according to FanGraphs than the Layman data in 2017. Upon research online at Baseball Reference, Layman data appears to be correct. Again, the difference is not significant.