**Major League Baseball Final Project**

**Corey Schwartz**

**Project Problem and Hypothesis**

* I hope to find out more about the predictability of three baseball statistics: HR/FB (how often a hitter’s flyballs result in home runs), SBA/PA (how often a player attempts to steal a base), and BABIP (batting average on balls in play – a measure of relative luck).
* Each of these target variables are important for projecting MLB players’ statistics, and have given my team trouble at work
  + HR/FB is a proxy for a player’s power – we can control easily for a player’s ballpark and can easily see how many flyballs he tends to hit, but it is tougher to predict increases or decreases in the player’s raw power throughout his career
  + SBA/PA is a proxy for a player’s speed/baserunning ability – it is tough to predict how this will change throughout a player’s career
  + BABIP – any one player tends to have a true ability to get more hits (thus higher BABIP), but the statistic fluctuates and obscures that ability, so it is tough to know when regression will occur
* For HR/FB and SBA/PA (power and speed), I essentially want to get an “age curve”, or see how the statistics depend on a player’s age, in order to be able to project future statistics based on past ones for a given player.
  + Other indicators may exist – for example a player may have a high BABIP because he’s fast, or a player may hit a lot of doubles that were almost home runs, and those doubles may become home runs in the future with added muscle. So I will test for those indicators.
* For BABIP, I want to see how the % of groundballs, flyballs, and line drives can predict a player’s TRUE BABIP level, stripping luck out of the equation
  + I will also test for other indicators, such as speed
* I expect to ultimately find that HR/FB increases up to a certain age and then plateaus, as the player adds muscle mass, without much decline at any point. I also expect SB/PA to peak very early in one’s career and then decline, as added size/muscle means slower, older players. I expect players who hit more line drives to have higher BABIPs, as those types of hits require the fastest reactions from fielders.

**Dataset**

* I will be using data available for free from **fangraphs.com**, a baseball statistics website
* With a log-in, I was able to grab custom datatables, so I now have every relevant statistic for every MLB hitter from 2012-2016, separated by season
* I will need to merge separate CSVs (one for each season) into one dataframe
* I will need to establish a minimum number of PA (plate appearances) in each season, and remove a player if he doesn’t satisfy that for each season in the sample, because I’m using rate statistics that are easily skewed

**Project Concerns**

* Concern #1: Time
  + I am trying to learn about 3 different target variables. I may not have the time to do this. So I am prioritizing HR/FB first, then SBA/PA, then BABIP. I will do 100% of the work for the first one, before moving onto the second one, etc. If I don’t get to the third one, at least I’ll have finished the work on the others.
* Concern #2: Linearity
  + Linear regression seems to make the most sense here. But what if the relationship isn’t linear, but rather parabolic or something like that? I see this being an issue with raw power/speed statistics, and then any predictive linear relationship could be misleading
  + If I don’t use a linear regression model, I’m worried about my ability to apply the model going forward

**The Model**

* I think first I will need to look at the scatterplots of the various statistics vs. age – if it looks linear, I think linear regression (with cross-validation) would be ideal, for finding causation and also establishing a formula I can use to project statistics at my job in the future
* If it’s non-linear, I’ll need to use other methods.
  + I will probably get a list of Random Forest importance for all the potential indicators