The goal of this project is to take baseball players' batted ball data from 2015 and 2016 and to use regression analysis to determine which of these are the most predictive. Then, we will use this data to project a player's performance the next year.

Data Collection

Our group will collect data from the following sources:

Baseball Savant - Baseball Savant is a site that contains Statcast information for batters. The information that we are interested in scraping from this site includes the exit velocity of a player's batted balls and the launch angles of each of these batted balls. This site also contains a stat called "barrels" that we are interested in using. We will use this data as the independent variables to generate expected performance. The site contains comprehensive tables with information for each player's batted ball events, but there does not appear to be an API for easily collecting the data. A BeautifulSoup scraper will be necessary for this site.

Fangraphs - This is a site that contains both box score and advanced statistical information. From this site, we are interested in collecting dependent variable data such as weighted on-base average, weighted runs created plus, batting average, etc. We are also interested in some independent variable information such as the amount a player is shifted against, strikeout rate, and walk rate. This site contains tables of information but no API, so we will need to scrape this site using BeautifulSoup.

Baseball Heat Maps - From this site, we are interested in collecting data about how many days each player spent on the disabled list. This site provides the information in Excel spreadsheet format, so scraping this site will not be necessary. We will have to extract a CSV file from this data.

Data Analysis

Once we have stored the necessary data in SQL, we will use it to create two different methods for projecting a player's future performance: the Marcel the Monkey projection system (an already-established and relatively simple method for forecasting player performance, with description here: http://www.tangotiger.net/marcel/), and a proprietary method that we will synthesize ourselves using a combination of FanGraphs and Statcast data. To create Marcel projections, we need only follow the pre-specified format.

To create our own system, we will have to perform regression analyses on the data we scraped to find which of our statistics are most predictive of our dependent variables. Some examples of potential independent variables include: spray, exit velocity, launch angle, barrels, injuries, age, speed, BABIP, etc. Once we see which ones have the highest correlations, we can synthesize our own metrics for projecting expected stats such as batting average or WOBA.

Once we have computed our desired metrics for each player, we can store them back into SQL, where they will be ready for querying and presentation.

Data Presentation

We will use Django to present out data where we portray our method for analysing the data as well as our projections for future data. We will compare our analysis to Marcel the Monkey, an already established method, with ours. This comparison will provide an alternate approach to the data, as well as a meter for comparison with how we decide to represent the data with our own regression analysis.

Also, we will provide information and a ranking of batters who performed better or worse relative to how we expected them to play vs. how they actually played. We will also use this ranking list to make analytical assumptions to which players we expect to perform better or worse in the coming years.

Visitors of the site will also be able to search players individually, which will return the player's last season stats as well as our projection and the Marcel projection for the next two years.

In our presentation of the data we will display data only related to non-pitching players to avoid extraneous data. This is based off the fact that pitchers are statistically shown to be much worse hitters than batters and have a much smaller sample size. This exclusion will help avoid any extraneous answers that our projections may provide.