**DS 3001 Project Proposal**

Team WHZM

**Team members**: Bao Huynh, Lindsay Maclnnis, Jessie White, Ziqian Zeng

**Needs and Benefits:**

Writing in 1919, philosopher Morris Raphael Cohen described baseball as America's national religion. In the words of sports columnist Jayson Stark, baseball has long been "a unique paragon of American culture"—a status he sees as devastated by the steroid abuse scandal. Today, as one of the most popular sports in the U.S., baseball still plays an important role in society. However, our team notices that it is difficult for normal people to have an intuitionistic way of understanding the financial movement for the teams and players, such as the salary of players, which is all public data. Therefore, our model targets users that have concerns in having a direct understanding of these huge amounts of data. Also, our model will be valuable for team or company managers that share a similar concern.

**The Data:**

The data we will be using comes from the website Baseball-Reference. This website contains an enormous amount of baseball statistics for every player and team in the MLB dating back to the late 1800s. We will be investigating whether or not some specific statistics can be used to predict the players’ salaries. We will split the players into two groups, position players and pitchers, and investigate different statistics. These statistics are very popular for evaluating a player’s performance and therefore we believe these will help us to estimate the players’ salary.

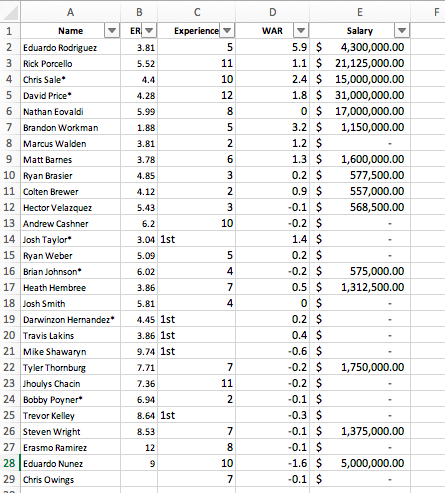
For positions Players:

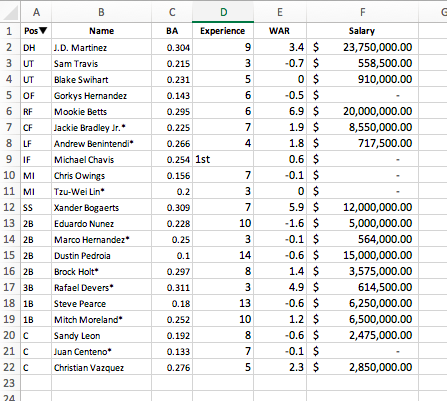
* Batting average
* Years in MLB
* WAR (Wins above replacement)

Pitchers:

* ERA (Earned Run Average)
* Years In MLB
* WAR(Wins above replacement)

Here is a sample of the data we will be pulling from the website. This dataset shows the desired statistics and salaries of the players on the Boston Red Sox. The first dataset is the pitchers and the second is position players.





**Toolkits:**

* Numpy: Data importation
* Matplotlib.pylot: Data visualization graphs
* Prettytable: Data visualization table
* Sklearn: Data regression implementation
* Functools (partial): Function pointer

**The Goal:**

We hope to build a regression model with multiple variables to predict the salary of baseball players in Major League Baseball (MLB). We plan to predict the salary in two groups based on positions: pitchers and position players. For position players, the variables we will use are batting average, years in MLB, and wins above replacement (WAR). For pitchers, the variables we will use are earned run average (ERA), years in MLB, and WAR. The ultimate goal is to find which variables influence a player’s salary, and use those variables to predict the salary of the players.