Springboard DSC Capstone Project 2 Milestone 1 Report:
Prediction of MLB player WAR statistic

Albert Chiu

March 2019

Problem:

MLB teams are constantly seeking to improve their rosters through free agency signings, trades and the development of minor league players in farm systems. For these teams, player salaries represent a significant financial cost and it is important to be able to effectively evaluate a player's value.

WAR (Wins Above Replacement) is commonly used as a metric to determine player value; it is the sum of a player's total offensive and defensive contributions and attempts to measure the total number of additional wins a team can expect with the player over an average replacement level player. A simplified version of the equation is:

$$WAR = {}^{(Batting\ Runs\ +\ Base\ Running\ Runs\ +\ Fielding\ Runs\ +\ Positional\ Adjustment\ +\ League\ Adjustment\ +\ Replacement\ Runs)} / {}_{(Runs\ Per\ Win)}$$

A common challenge for MLB teams is forecasting the future production of a player. Teams will acquire a star player at the height of their careers for a large contract, only to discover that they've anchored themselves and a significant portion of payroll to a player whose production is quickly diminishing. Other times, teams will overlook a young player that will eventually become a superstar or an aging veteran that continues to provide solid on-field play. If MLB teams are able to better forecast when a player's production is on the verge of decline or improvement, it will allow them to make more informed player acquisition decisions.

Data Sets:

Lahman's Baseball Database: http://www.seanlahman.com/baseball-archive/statistics/
Baseball Prospectus: https://www.baseballprospectus.com/

EDA and Data Wrangling:

For this project, I obtained two csv data sets from Baseball Prospectus (containing primarily offensive stats) and Lahman's Baseball Database (containing primarily defensive stats). The csv files were read into dataframes and then merged together after converting the Baseball Prospectus player name convention (full first and last name) into the Lahman convention (first 5 letters of last name followed by first two letters of first name). There were a lot of special cases that needed to be addressed including:

- Last names that contain multiple whitespaces, ex. Tomas de la Rosa
- Names that end with Jr. or Sr., ex. Jose Cruz Jr.
- Names that have initialed first name with whitespace separation, ex. J. T. Bruett

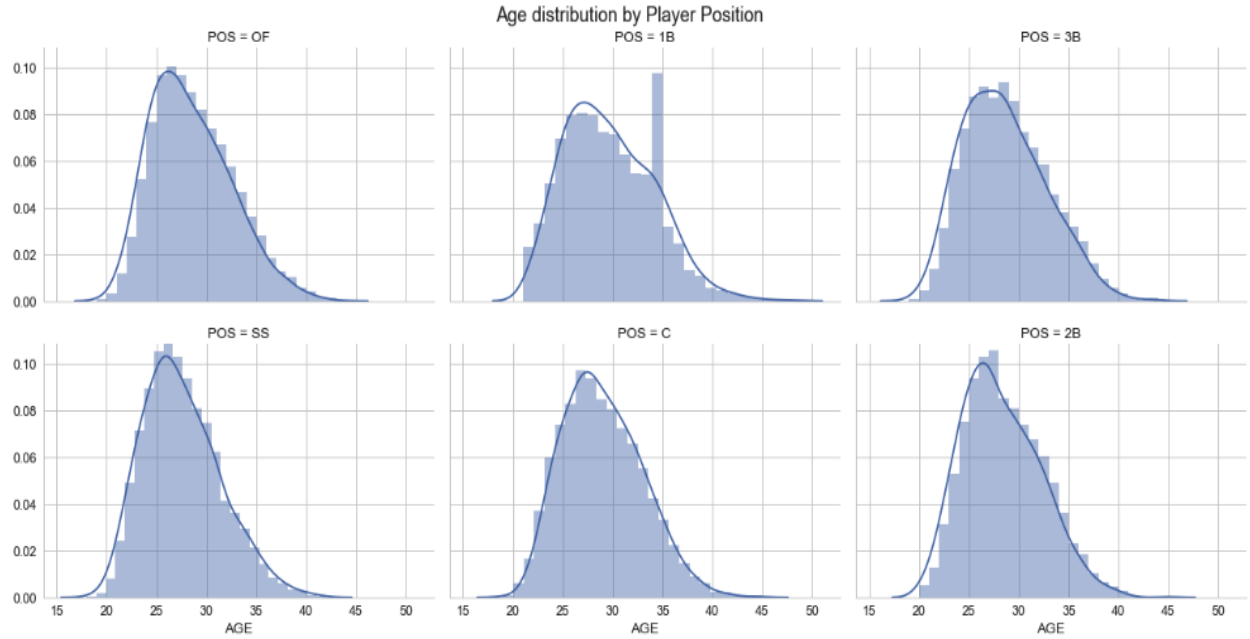
- Names that have initialed first name without whitespace separation, ex. J.T. Snow
- Names with middle initial, ex. Bobby J. Jones
- Names that contain middle names, ex. Chan Ho Park
- Names that contain whitespace in first name, ex. La Marr Hoyt

There were a number of players that played multiple positions over the course of a single season so there would be a separate row for each position played. To simplify things, I calculated the position that the player played the most inning outs for the season and aggregated all their stats to that position.

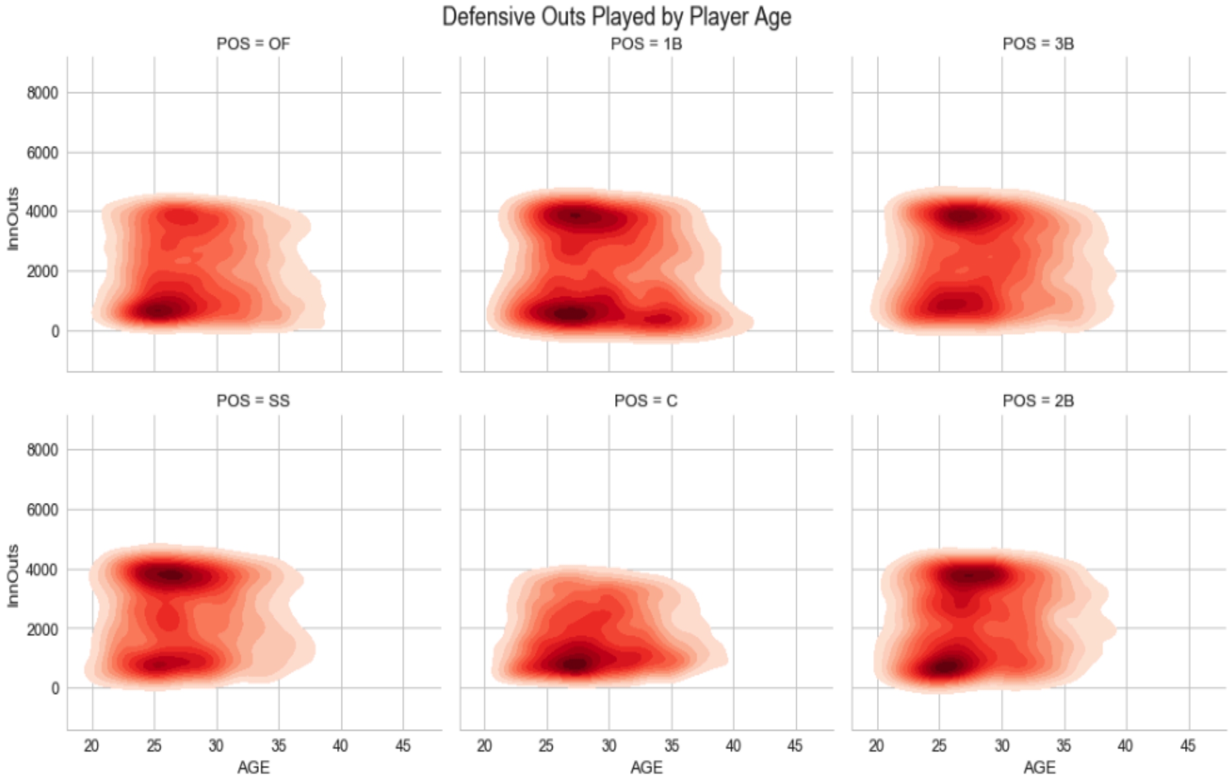
There were also numerous instances where a player would play for multiple teams over the course of one season. For these players, I merged the statistics for all teams played in a season to a single row.

The final data set after the data wrangling is a dataframe where each row represents all the statistics of a player for a single season. It contains 18,621 unique player seasons and covers the span of 40 MLB seasons (1977 through 2017).

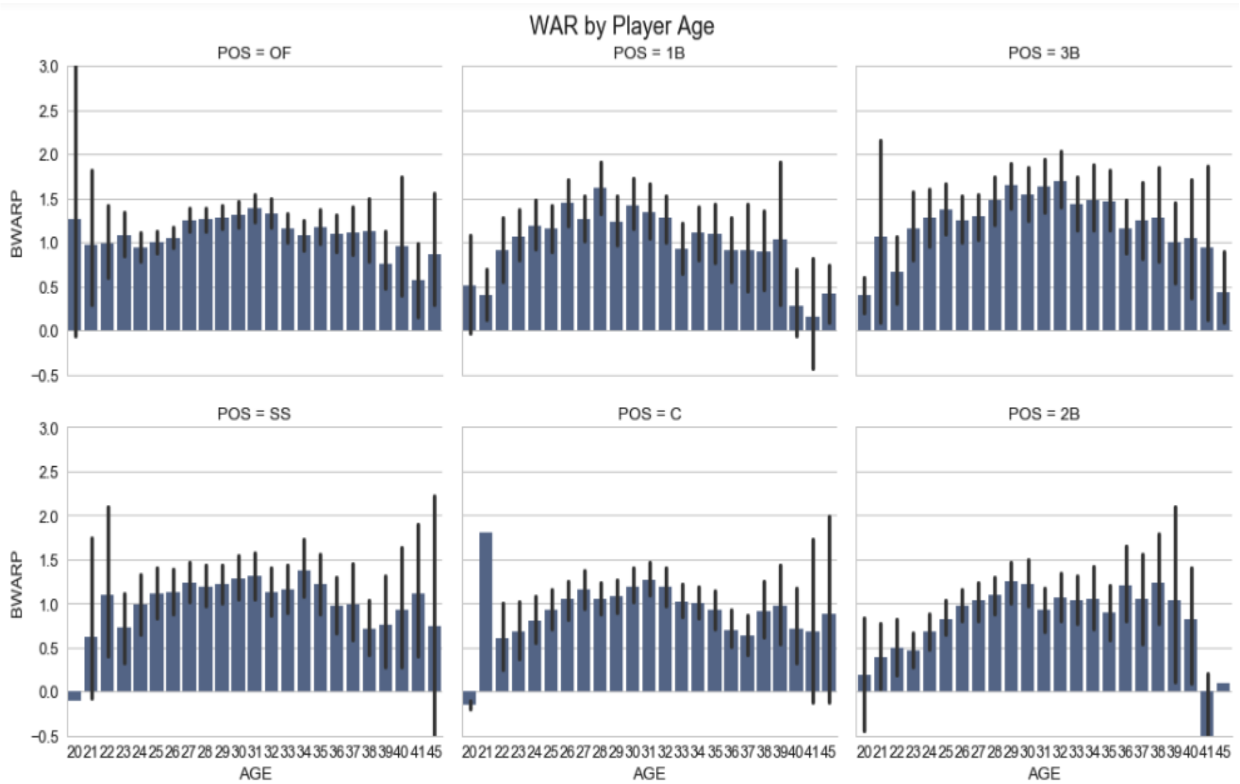
Data Visualization:



When we examine the age distribution of the players by position, it appears that most position are filled by players in their late 20's and early 30s. First basemen seem to have slight longer longevity than other positions, while shortstop tend to skew younger.



When we create contour plots of the defensive outs played by players by position, we can see that different positions have different tendencies. Catchers tend to player fewer defensive outs than other positions; for 1B, 2B, 3B and SS, we see two peaks for players in their mid to late 20's that play approx. 4000 outs and those that play several hundred outs.



When we plot the WAR by age for each position, we observe that for most positions, WAR peaks in late 20's and early 30's. Additionally, we see that there is more variance in WAR statistic for younger players (less than 25 yo) and older players (more than 35 yo) than players in their late 20's and early 30's.

Inferential Statistics:

Performing some statistics on the data set, we gain some insights:

1. When plotting a correlation table for the features in the data set, we observe that WAR is most highly correlated with TB, DRC+, RBI, R, HR, BB and DRAA features.
2. For most positions, players in the top 5% of BWARP have a statistically different OBP than players in the bottom 95% of BWARP.
3. For shortstops, it appears that there is a larger difference in DRAA between players below the 50th BWARP percentile and those above the 50th BWARP percentile.
4. For most positions, players in the top 5% of BWARP have a statistically different age than players in the bottom 95% of BWARP.