**DSC 530**

**Exploratory Data Analysis**

**Final Project**

**Milan Sherman**

The outcome of my exploratory data analysis demonstrated a few things. First, the data supports the hypothesis that first basemen are better power hitters than other field players[[1]](#footnote-1). Another outcome of this analysis is that slugging percentage is a better predictor of runs scored than batting average. Finally, we established a correlation with at-bats and slugging percentage and batting average, so that players with more at-bats generally have higher batting averages and slugging percentages. The direction of this relationship is unknown, but is likely bidirectional, i.e., players who hit more get better at it, and players who are better hitters will likely get more opportunities to hit.

As far as other variables in concerned, it would have been interesting to know which side of the plate a player batted from, for a couple of reasons.

1. My intuition suggests that there is a higher proportion of left-handed baseball players than the general population. Historically, players who bat from the left side of the plate hit better against right-handed pitchers than right-handed hitters. As most pitchers are right-handed, left-handed hitters may be more highly valued. Having this data would allow us to determine if indeed the proportion of left-handed baseball players is higher than the general population.
2. There also seems to be a higher proportion of first basemen who are left-handed, with good reason. A first baseman lines up just to the right of the foul line, meaning that most hits that they would have to field will be to their right. A left-handed player throws with their left hand and has their glove on their right hand. So if the ball is hit to their right, a left-handed first basemen is better prepared to prevent a base hit as they can reach farther to their right than a player with their glove on their left hand. If it is true that first basemen are left-handed more often than other players, it would be interesting to then explore these hitting metrics based on which side of the plate a player hits from.

In terms of assumptions, there is an implicit assumption about the value of power hitting. While we showed that power hitting is a more predictive metric than batting average with respect to runs scored, the question about whether first basemen tend to be better power hitters somewhat assumes that power hitters are valuable. Another metric that is highly correlated with runs scored is base-out percentage, which is the ratio of the number bases a player nets to the number of outs that they generate. In this analysis, we explored a hypothesis about power hitting and first basemen. A more open question is which types of players tend to have the best base-out percentage? A further assumption is about the values of runs scored. Ultimately, scoring runs is only half of the equation. No matter how many runs your team scores, you will still lose if your opponent scores more.

A few things that would improve this analysis:

1. I conducted a hypothesis test on the difference of means between first basemen and other field players (except pitchers) for each of three power hitting metrics, which showed that the difference between first basemen and all other field players *as a group* is statistically significant. An open question is whether first basemen are better power hitters compared to each type of field player. For example, right fielders were the second highest in each of the three power hitting metrics behind first basemen. But when combined with other field positions that are much worse power hitters, we fail to answer the question as to whether first basemen are better power hitters than right fielders. To answer this question we would need to conduct an ANOVA, and likely a post hoc analysis, which was not covered in this class and I ran out of time to determine how to do this in Python.
2. Given the correlation noted above between at-bats and the hitting metrics used in this analysis, it would be useful to control for number of at bats by replicating the analysis after creating separate groups based on number of at-bats to see if the results hold. For example, at the very least it would be worth splitting this data into players with fewer than 2000 at-bats and those who had more.

1. Pitchers were excluded from this analysis as they are notoriously bad hitters. Major League Baseball consists of two leagues, the American League and the National League. Pitchers do not even hit in the American League, but their at-bats are taken by designated hitter, a player who hits but does not play the field. [↑](#footnote-ref-1)