Part 1: Data

Anna, Riley

2/7/2022

Overview of the Dataset

This data was collected almost entirely from baseballsavant.com. The categorical variables were collected from baseballreference.com and mlb.com. Our sample population is the 132 Major League Baseball players who had more than 502 plate appearances during the 2021 season, enough plate appearances to qualify for end-of-year awards. Batters in the sample have between 505 and 724 plate appearances. We chose qualified players out of a desire to avoid treating players with far fewer plate appearances (smaller sample sizes) with equal statistical significance. Each row/observational unit represents one batter.

The columns are the variables, 4 of which are independent categorical variables. These are handedness (what side of the plate the batter bats from), position (what field position the batter plays the most innings at), mode of acquisition (high school draftee, college draftee, or international free agent signing), and the division they play in. The other 41 variables are quantitative. Some quantitative variables are/have been made discrete, such as age or number of home runs. The quantitative variables of the most interest to us are:

b_hr_rate: the rate at which the batter hits home runs. Home runs/plate appearances.

b_k_percent: the rate at which the batter strikes out. Strikeouts/plate appearances.

b_bb_percent: the rate at which the batter walks. Walks/plate appearances.

batting_avg: the rate at which the batter gets base hits. Hits/at bats.

isolated_power: where a single is worth 1 total base, a double is worth 2, and so on, $ISO = \frac{\text{total bases - singles}}{\text{at bats}}$

exit_velocity_avg: the average velocity of the hitter's batted balls.

launch_angle_avg: the average vertical angle at which the ball leaves the hitter's bat, where 0 degrees is parallel to the ground.

sweet_spot_percent: the percentage of the hitter's batted balls that are hit between 8-32° launch angle.

z_swing_percent: the percentage of pitches in the strike zone that the batter swings at.

oz_swing_percent: the percentage of pitches outside of the strike zone that the batter swings at.

whiff_percent: the percentage of pitches at which the batter swings and misses entirely.

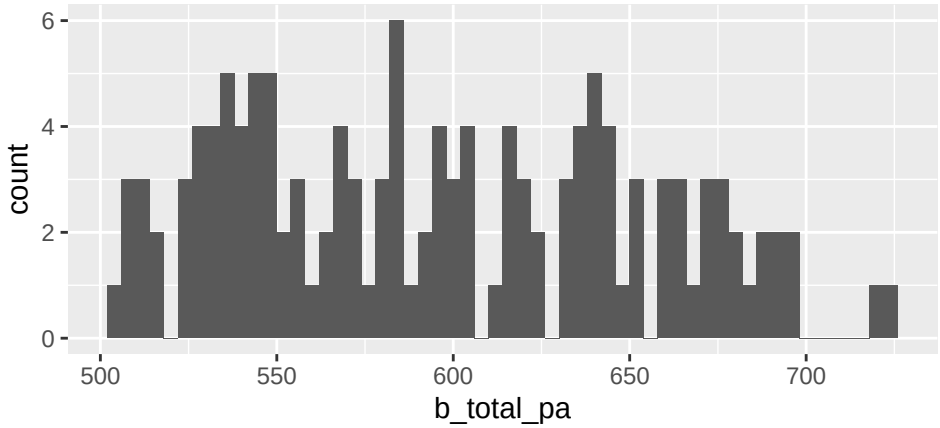
pull_percent: the percentage of a hitter's batted balls that are hit to the same side of the field as the batter's box they bat from (left/right).

groundballs_percent: the percentage of a batter's batted balls that are hit on the ground (less than 10° launch angle).

flyballs_percent: the percentage of a batter's batted balls that are hit in the air, between 25-50° launch angle.

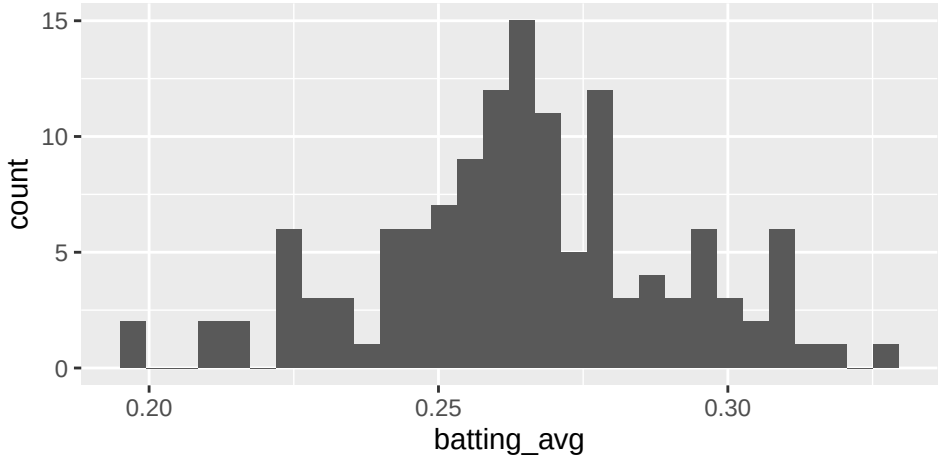
linedrives_percent: the percentage of a batter's batted balls that are hit between 10-25° launch angle.

Summary Statistics



Perhaps the most important summarizing statistic that can be found here is the standard deviation in the total number of plate appearances (b_total_pa). One standard deviation is equal to 56.2 about a mean of 597. This is significant as it pertains to counting statistics (as opposed to averages or rate statistics). For example, if we were looking to compare players on their ability to hit home runs, simply comparing home run totals would not be the best way because of the variance in number of PAs. It would be better to use home run rate, HR/PA. This variance in sample size matters less for averages/rate statistics, because they are adjusted for number of PAs and averages/rates tend to approach their "real" values in large samples (502 PAs being a pretty large sample). In our future analyses of this data set, we will confine ourselves to averages/rate statistics.

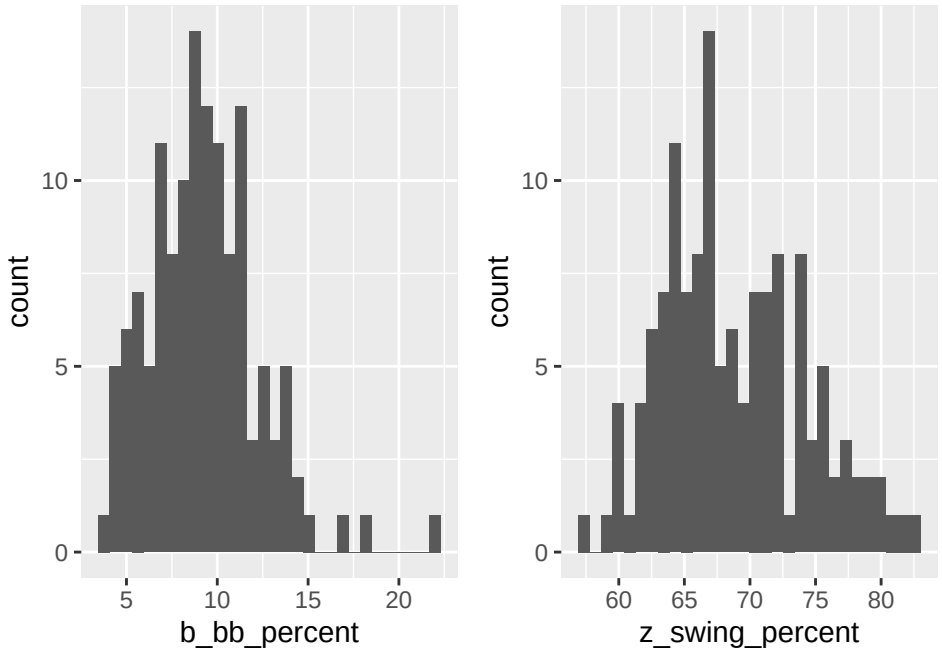
We can gain further insight by comparing some of the means here to the overall 2021 Major League Baseball (MLB) means. For example, the league-wide slash line (batting average/on base percentage/slugging percentage) in 2021 was .244/.317/.411. (see: https://www.baseball-reference.com/leagues/majors/2021.shtml). In this sample, the mean slash line is .265/.339/.457, which is better. This makes sense, as better players tend to be given more plate appearances, while many others do not receive enough to qualify. While some means in this sample may vary from the greater population of MLB players, it is a necessary sacrifice in order to conduct linear regression tests that give us a more accurate view of the relationships between these variables, each player needs to have a sufficiently high number of plate appearances. This way, we can avoid small sample noise. Lowering the plate appearance requirement for this data set would make this sample more similar to the average player, but there would still be difference in the same direction. To be able to compare the two, we have have to sample almost the entire population. We will say that this sample is taken from the population of MLB players with roughly a full season's worth of plate appearances (502+).



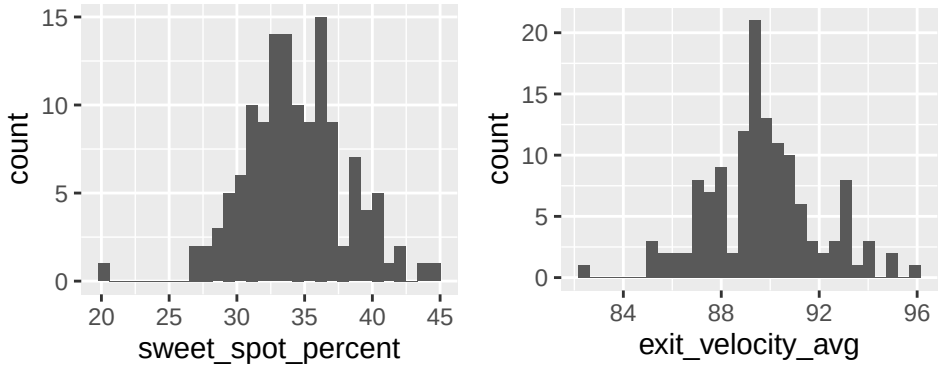
The batting average, traditionally used as the metric to determine how good a player is, has a mean of 0.265

which is also the mode, i.e. batting average of most players. Some outliers on both extremes may also explain outliers in the following two variables of interest.

Some metrics that baseball fans have begun to prioritize over the batting average include the walk rate (b_bb_percent), and the percentage of swings on pitches inside the zone (z_swing_percent), both of which have bell-shaped distributions and are skewed to the left. In particular, the walk rate has at least one outlier to the right. It will be worth investigating whether these two variables are related to the traditionally used batting average.

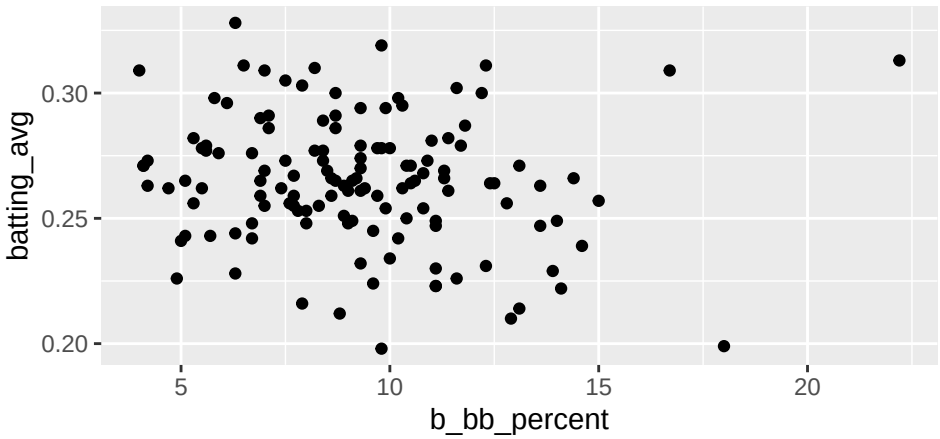


As for statistics regarding batters' swings, both the sweet spot percentage and the exit velocity average are quite symmetric bell shapes with one outlier to the left. It will be worth exploring whether these two variable are related, and whether they are explanatory variables for the determinants of a "good player" mentioned above.

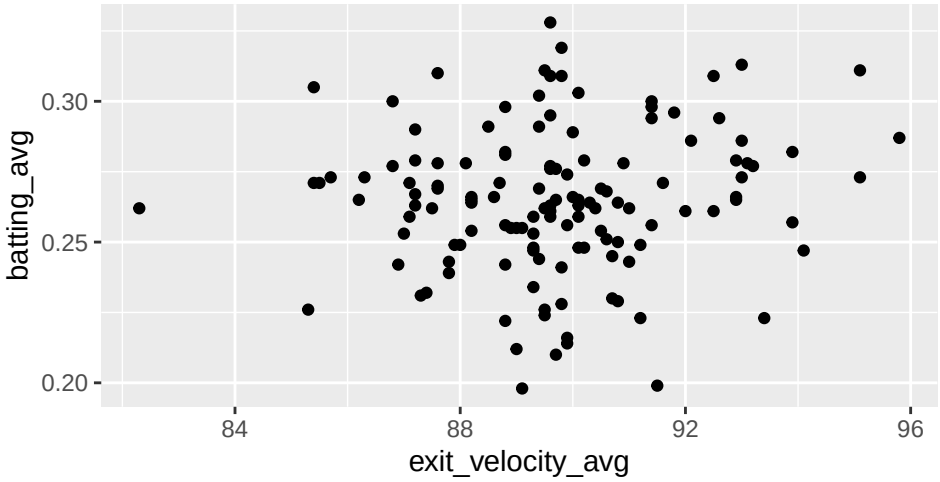


Graphical Analysis

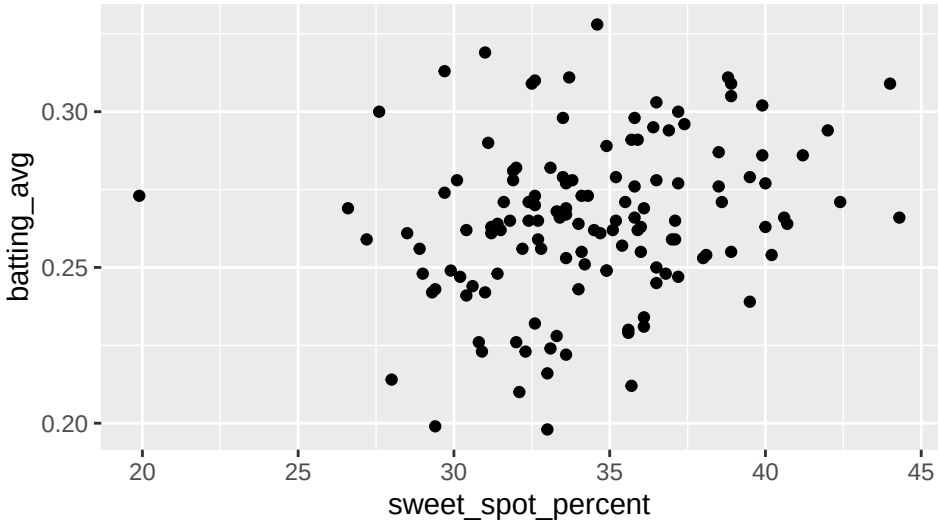
Following analysis above regarding traditional and newer conceptions of a "good player," initial plots don't seem to show a strong relationship between the walk rate and the batting average, and may even show some negative correlation even though



The following scatter plot also seems to indicate that the exit velocity of a hit is not strongly related to the batting average, implying that the former variable may not be as significant in predicting whether a player gets base hits.

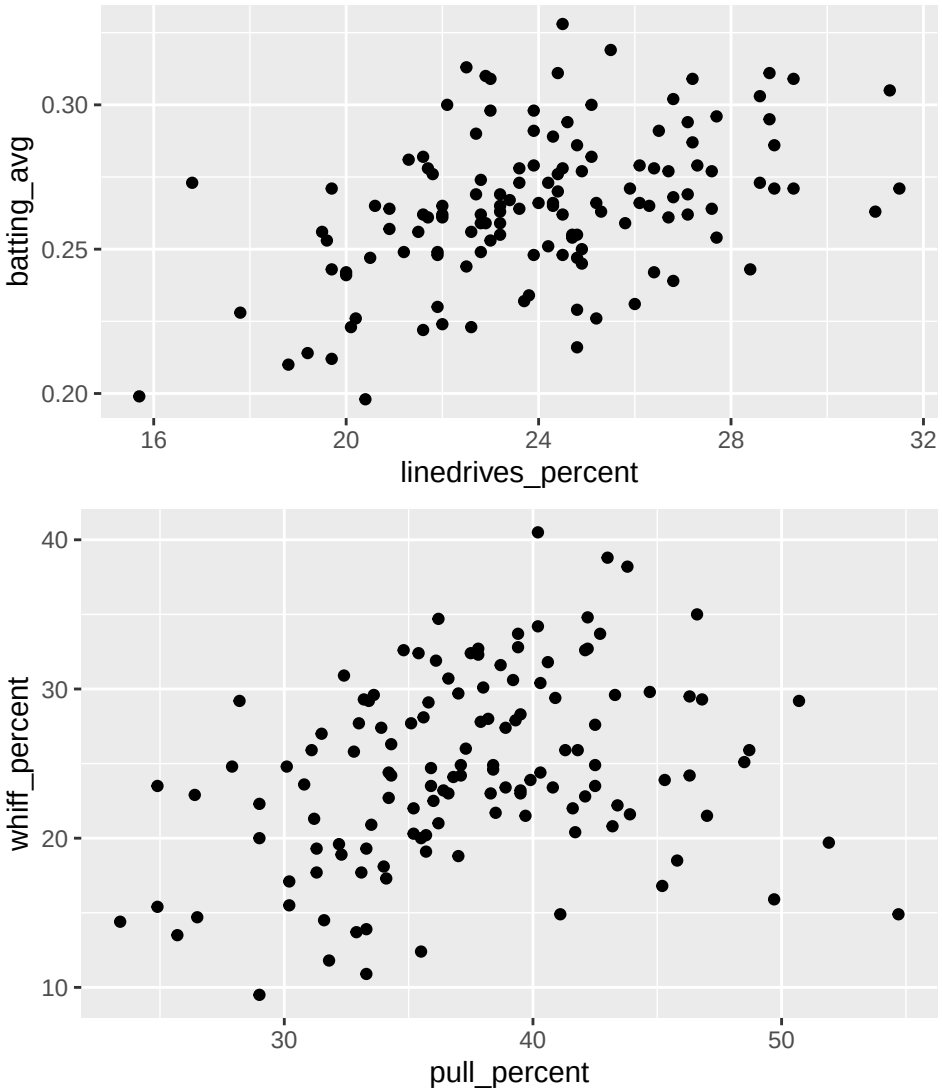


However, there does seem to be an upward trend relating a player's sweet spot percentage and his batting average.



Another initial scatter plot worth noting shows that a player's batting average increases steadily with the

percentage of line drives (balls batted at 10-25° launch angle), which makes sense as the latter is considered an optimal hit.



Discussion

This preliminary analysis of our dataset shows most variables are have bell-shaped distributions, but skews and asymmetry may warrant some investigation. We have also observed some outliers in multiple variables; determining whether these outliers originate from the same players, and consideration of the outliers' effects on future inference will be required. The apparent lack of relationship between certain variables is surprising and may be so, or other relationships may become apparent with transformations.

As discussed, our sample of players who qualified for end-of-year awards is quite representative of our population of MLB players. However, a broader population of, for example, all competitive baseball players, may not be suitable since our sample is limited to male American players.