Project Part 2

Project Questions

1) What is the relationship between the average hit angle and average home run distance? Answering this question can provide insight on how batters can set up in the box to either raise or lower their hit angle (if there is a relationship between the two variables).
2) Does a batter's hit speed play a role in the distance they hit the ball? Here I can see whether or not a more powerful hitter tends to hit the ball with a lot of speed or not.
3) Does a player's batting average coupled with slugging percentage have a relationship with average hit angle? Answering this question will give players a better understanding of the angle they need to hit the ball to achieve the best hitting performance.

Description of Data Set

The data sets I will be using come from https://baseballsavant.mlb.com/, which is a data collection site using STATCAST powered by GCP. This site is directly maintained by the MLB, therefore making it a primary source. The data contains aggregated hitting statistics for every eligible hitter for the entire 2022 season. The data was directly collected from each player's game, with an aggregation of each player's statistics conducted after each game. Since, every eligible hitter in the MLB is accounted for in the data, the data is a population, and conclusions can only be made about MLB hitters.

The rows in the data are the names of every eligible hitter in the 2022 MLB season. The variables in the data are specific hitting statistics. The variables of interest are as follows: BBE - Batted Ball Event, LA - Launch Angle, SwSp% - sweet spot percentage (a batted-ball event with a launch angle between eight and 32 degrees), Max Exit Velocity, Average Exit Velocity, FB/LD - foul balls over line drives, GB - ground ball rate, Max HR Distance, Average Hit Distance, Average HR Distance, PA - plate appearances, BIP - balls in play, BA - batting average, and SLG - slugging percentage (total number of bases a player records per at-bat).

One potential issue of using the data set is the possibility of outliers skewing the results. For example, there may be a very poor hitter in terms of batting average but only hits home runs resulting in a high slugging percentage. This could impact the interpretation of question 3, so this is something to keep an eye on.

The data is appropriate to answer the research questions. The proposed questions want to explore certain MLB statistics which are all present in the data set.

Numerical Summaries

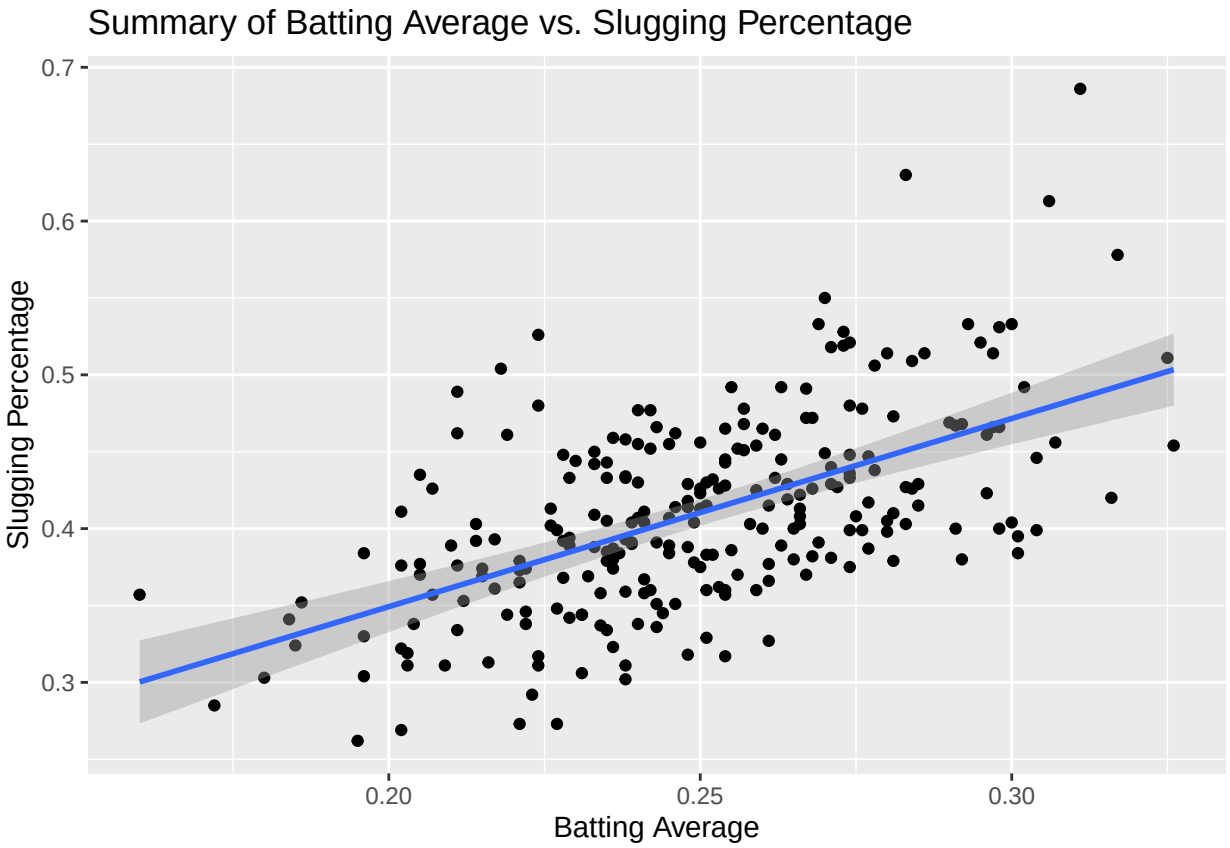
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   -1.90    9.80   12.45   12.58   15.70   24.60
```

I wanted to look at a numerical summary of the average hit angle variable. This summary is useful because one of my research questions explores the relationship between average hit angle and home run distance. Looking at exploratory statistics of average hit angle will help me see the variation of average hit angle among MLB hitters. I can see that the hit angle takes on a wide range of values. Since the range of numbers is so large (min of -1.90, max of 24.60, relatively low IQR), this gives me a good indication that different launch angles could correspond to certain home run distances.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   344.0   391.2   398.0   397.5   404.8   428.0       2
```

I also wanted to look at a summary of average home run distance. It is obvious that lower angled hit balls have a less likely chance of being a home run, but that changes with the size of the ballpark. Looking at these summary statistics, the idea of playing in a small ballpark having an impact on launch angle and home run distance can be nullified. This is because the average home run distance is comparable to the average dimensions of MLB ballparks (~400 feet).

Graphical Summary



Since my third question explores using both batting average and slugging percentage as an indicator for launch angle, I wanted to look at a graphical summary of the two to see if multicollinearity problems exist. From the graph, there seems to be a positive correlation between the two but not perfect. Since there does not seem to be evidence of multicollinearity between the two variables, I can move forward in using both as an indicator of launch angle.

Final Conclusions

Based on initial data exploration, I am confident that this data set will do a good job answering my questions of interest. The numerical and graphical summaries explain partially why. Another reason being that this data set contains 29 variables. If I run into problems during the analysis phase and cannot answer one of my questions, I can shift focus to another set of variables with ease and explore another question.

References

1) https://baseballsavant.mlb.com/
2) https://www.si.com/mlb/2021/03/24/mlb-outfield-walls-ranked-fenway-park-yankee-stadium