Baseball - A Statistical Sport

Thomas Martinez, Sasha Cahill

Introduction

For our project, we wanted to see what factors affected home runs (more specifically, the percentage of plate appearances in which a player hit a home run) and batting average in the MLB. Baseball is widely known for being a game of statistics, whether due to the popularity and impact of Michael Lewis's 2002 book *Moneyball: The Art of Winning an Unfair Game* and its eponymous 2011 movie[1] or with the MLB having adopted analytical tools and methods for studying peripheral statistics such as Statcast, which was formally adopted by the league and implemented by all 30 teams in 2015[2]. Tracking of personal and team statistics goes back almost as far as organized baseball, with 1930's players such as Hack Wilson keeping personal stat sheets[3]. Statisticians and baseball fans alike would continue to develop and implement statistical approaches to baseball, with the most notable being sabermetrics by Bill James in 1977[4]. Additionally, teams such as the Oakland Athletics and Tampa Bay Rays were renowned for unique, statistical approaches to baseball in the modern era, which were needed to field cheap but competitive teams in small market environments[5]. The general idea of these methods is that they study a player's performance and stats via lesser known and acknowledged metrics such as exit velocity and hard hit percentage, which directly or indirectly impact many widely known and traditional stats such as home runs and batting average[6]. By understanding and analyzing these peripheral stats, players can improve many aspects of their play and performance whether by simply increasing certain numbers or qualitatively changing their approach to the game. These statistics can also give a trend of a player's performance and whether or not they'll improve or decline.

For example, Arizona Diamondbacks third baseman Jake Lamb discovered, following the 2015 season, that by adjusting his batting posture when up to hit he improved his exit velocity and as such improved many other batting statistics, most notably home runs. This even led to Lamb going on a streak of hitting 16 home runs in 45 games. Teams and front offices can also use statistics to decide who to play, bench, sign, trade, etc. In another example involving exit velocity, the 2014 New York Mets chose Lucas Duda over Ike Davis as their starting first baseman due to Duda's better exit velocity. Duda would then also go on to hit 57 home runs in the span of two seasons. Besides hitters, which are the main focus of this project, pitchers can also be subjected to statistical analysis and use it to their advantage, or be taken advantage of, in improving their game. In one instance, a team considered trading for Kansas City Royals closer Wade Davis but changed their minds because of his declining spin rate, a stat that is often overlooked compared to pitching stats such as pitch velocity. This decision to back out of the trade proved wise as Davis would end up injured with a flexor strain just before the trade deadline[7]. Statistics aren't everything, however, and baseball players must adjust their game to be successful, incorporating statistical analysis in varying amounts. An interview with Josh Donaldson reveals the careful observance some players have of their personal statistics and those

of other players in MLB, but also the need to customize ones game by way of feel in the absence of statistics. By incorporating simple yet effective qualitative changes such as how a player handles their bat or what part of the field they should hit the ball to, Donaldson and other stars were able to find success in the league and develop their own style of play[8]. Pitchers too can also qualitatively change their approach to the game in order to improve their results and performance. Diamondbacks pitcher Bryce Jarvis theorized that by improving his release of a pitched baseball, or extension, he could make the ball travel as little as possible and thus give the batter less time to react, helping improve his performance and prove himself in the major league[9].

Materials and Methods

While many variables were studied and considered for their effect on the aforementioned hitting outcomes of home run percentage and batting average, our project mainly looked at 4 variables in particular (a general project outline can be found in A11). These 4 variables were launch angle (the angle that a batted ball is launched at), barrel percentage (the percentage of batted balls that had the perfect combination of exit velocity and launch angle), hard hit percentage (the percentage of batted balls with an exit velocity of at least 95 miles per hour), and EV50 (the average of a player's 50 best exit velocities). These variables along with others can be referenced on the Baseball Savant website, which stores and tracks Statcast data for each player via datasets and leaderboards[10]. Additionally, home run percentage was studied rather than just home runs because home run percentage takes into account the number of games a player played in rather than the whole 162 game MLB season. Our data analysis was all done digitally on R Studio and our datasets were imported directly from Baseball Savant or modified in Google Sheets. All datasets for our project were from the 2024 MLB season. The data was analyzed using ANOVA, ETA squared, and Games-Howell, in which variables were analyzed individually or together to see if they had an effect on the hitting variables as well as any indication of statistical significance. Said analysis data is referenced in A1 and A2 of the Appendix. The data was also plotted visually via boxplots in order to see the distributions of each variable and general trends which could indicate if there were any statistically significant differences within the variables, as well as what exact parameters and numbers were needed to maximize home run percentage and batting average. These boxplots can be found in A3 through A6 of the Appendix.

Results and Discussion

All the one way ANOVA tests showed that each variable had had a statistically significant effect on home run percentage and batting average at a 95% confidence interval. Multiple variables were also analyzed together in a two way or three way ANOVA test. For home run percentage, only hard hit percentage and EV50 showed statistical significance when paired together. For batting average, every two way ANOVA except for Launch Angle-EV50 and Barrel Percentage-EV50 showed statistical significance. No three way anova test for home run

percentage or batting average showed statistical significance. These values can be referenced in A1 and A2 and the Appendix. Eta squared tests were also done to see the size of the variable effects on home run percentage and batting average. Such data, along with the one way ANOVA p-values, is presented in the following table.

	Home Run Percentage				Batting Average			
	Launch Angle	Barrel %	Hard Hit %	EV50	Launch Angle	Barrel %	Hard Hit %	EV50
η^2	0.183	0.642	0.401	0.406	0.054	0.050	0.110	0.111
p	2e-16	2e-16	2e-16	0.0133	6.65e-5	2.92e-6	3.07e-5	5.6e-7

Table 1: Eta squared (η^2) values and p values (derived from the anova analysis) for each of the 4 parameters against Home Run Percentage and Batting Average. A large η^2 is > 0.14, a medium η^2 is around the interval (0.06,0.14), and a small η^2 is < 0.01. A P value of less than 0.05 indicates that the result is statistically significant.

Barrel percentage had the biggest effect on home run percentage, with an eta squared value of 0.642, followed by EV50 at 0.406, hard hit percentage at 0.401, and launch angle at 0.183. For batting average, EV50 had the biggest effect with an eta squared value of 0.111, followed by hard hit percentage at 0.110, launch angle at 0.054, and barrel percentage at 0.050. As such, every variable had a huge effect on home run percentage and a smaller but still fairly noticeable effect on batting average.

The Games Howell Test revealed which parameters were statistically significant and could therefore be used at a 95% confidence interval to make recommendations about where the four hitting statistics should land to maximize home run percentage and batting average. The Games Howell Test for each variable can be referred to in A7 through A10 in the Appendix. A launch angle of 10.5 degrees or higher would be sufficient to maximize home run percentage. Any launch angle above that may improve home run percentage, but cannot be said for certain at a 95% confidence interval. There were no statistically significant launch angle parameters that affect batting average, so although launch angle was shown in our anova analysis to be a significant factor impacting batting average, we cannot give a specific numerical recommendation. A 13.3% or above barrel percentage is favorable for home run percentage, while once again, no parameters were significant for batting average, and no recommendation can be made. A hard hit percentage at or above 49.3% is recommended for home run percentage, and a hard hit percentage of 45.9% or above is shown to be statistically impactful for batting average. An EV50 of 102.9 mph is recommended to maximize both home run percentage and batting average.

On the batting average graphs for barrel percentage, hard hit percentage, and EV50 (A4-6), a sort of "U" shaped trend can be observed, where players performing poorly in these statistics seem to still be accruing good batting averages. Although nothing can be said at a 95% confidence interval, it would be a point of further study to examine why this trend occurs. It could be that players whose batting approach is more based on bat control.

Conclusion

As stated before, every variable had at least a noticeable effect on home run percentage and batting average. Barrel percentage had the biggest effect on the former, with a 0.624 eta squared value, while EV50 had the biggest effect on the latter, with a 0.111 eta squared value. Every variable also showed statistical significance for home run percentage and batting average when passed through an anova analysis. Potential error in our study could have resulted from the way in which we set up the dataset, in that we parameterized the data in order to make the Games Howell test functional. We also had to exclude outliers to make the Games Howell test functional, something that could have also skewed the results slightly.

Further study could be conducted on any of the many peripheral statistics that have recently surfaced as of baseball statistical age. Some of these include whiff percentage (how often a platelet swings and misses), chase percentage (how often a player swings at a pitch outside the strike zone), and swing length (a measure of how long a batter's swing is). Besides home run percentage and batting average, other hitting outcomes such as on base percentage and slugging could also be analyzed in regards to peripheral statistics. As mentioned in the introduction, pitching and peripheral pitching variables could also be studied and considered, potentially resulting in an entire study on how to maximize certain pitching metrics or how to be a successful MLB pitcher in general. Other seasons could also be studied as well besides 2024, particularly seasons immediately following the MLB's implementation of Statcast in 2015 or potentially seasons prior to its implementation if such data is available.

While an understanding of advanced, peripheral statistics is necessary for a Major League player to succeed in today's game, each player's scenario may be different and they may customize their offensive approach to maximize success. While barrel percentage had the biggest effect on home runs, a player might find that improving their exit velocity could yield more home runs. Perhaps a player may focus less on hitting home runs at all and instead focus on just getting on base. These decisions could also be outside a player's control or be decided by a team as a whole. Ultimately though, how a player wants to improve their performance depends exactly on their approach to the game and what metrics they specifically want to improve. As such, an understanding of these statistics can help players improve their performance qualitatively and quantitatively.

Appendix

A-1

Table 2: Results for the three-way anova, with Home Run Percentage as a factor of each of the 4 experimental parameters (launch angle, barrel% Hard Hit %, and EV50) measured individually and against each other. Bolded rows indicate statistically significant results at a 95% confidence interval, with associated P values listed.

Launch Angle	**Anova** - Home Run Percentage		**p=2e-16**
Barrel %			**p=2e-16**
Hard Hit %			
EV50			**p=0.0133**
Launch Angle	Barrel %		
Launch Angle	Hard Hit %		
Launch Angle	EV50		
Barrel %	Hard Hit %		
Barrel %	EV50		
Hard Hit %	**EV50**		**p=0.0395**
Launch Angle	Barrel %	Hard Hit %	
Launch Angle	Barrel %	EV50	
Launch Angle	Hard Hit %	EV50	

A-2

Table 3: Results for the three-way anova, with Batting Average as a factor of each of the 4 experimental parameters (launch angle, barrel% Hard Hit %, and EV50) measured individually and against each other. Bolded rows indicate statistically significant results at a 95% confidence interval, with associated P values listed.

Launch Angle	**Anova** - Batting Average		**p=6.65e-5**
Barrel %			**p=2.92e-6**
Hard Hit %			**p=3.07e-5**
EV50			
Launch Angle	**Barrel %**		**p=0.0389**

Launch Angle	**Hard Hit %**		**p=0.0163**
Launch Angle	EV50		
Barrel %	**Hard Hit %**		**p=0.0334**
Barrel %	EV50		
Hard Hit %	**EV50**		**p=0.0493**
Launch Angle	Barrel %	Hard Hit %	
Launch Angle	Barrel %	EV50	
Launch Angle	Hard Hit %	EV50	

A-3

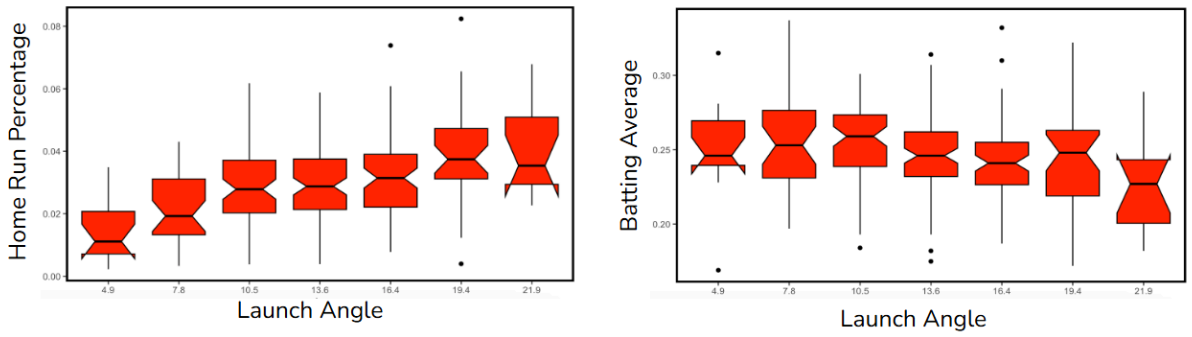


Figure 1: Boxplots showing distribution of launch angle with respect to home run percentage and batting average.

A-4

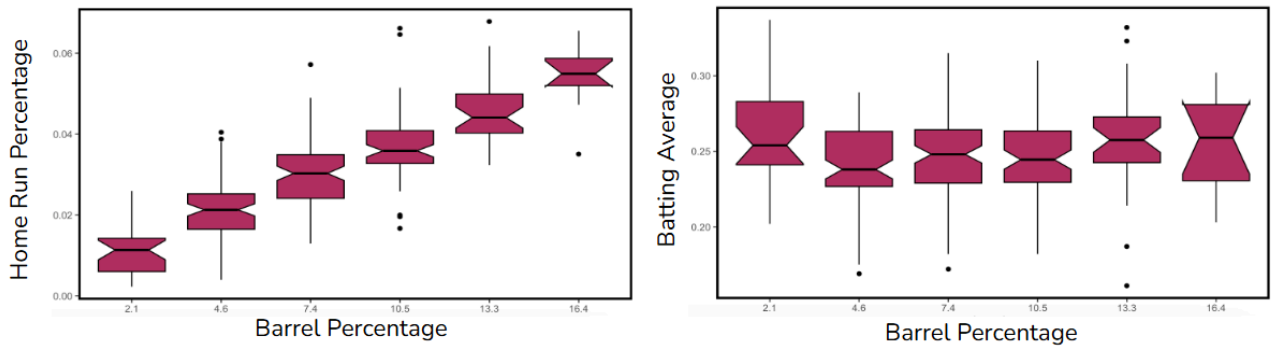


Figure 2: Boxplots showing distribution of barrel percentage with respect to home run percentage and batting average.

A-5

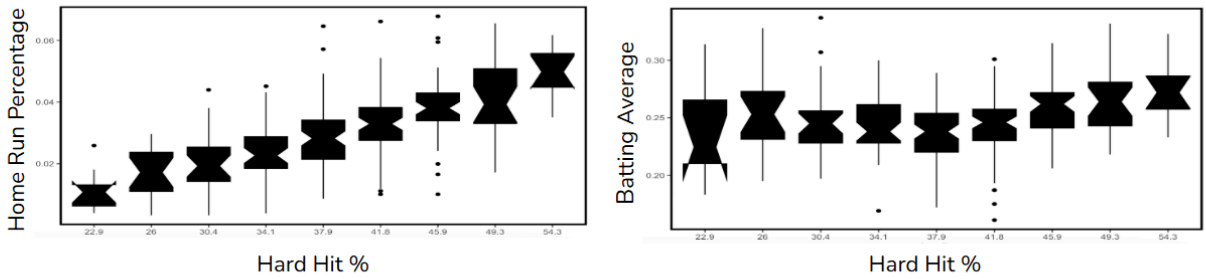


Figure 3: Boxplots showing distribution of hard hit percentage with respect to home run percentage and batting average.

A-6

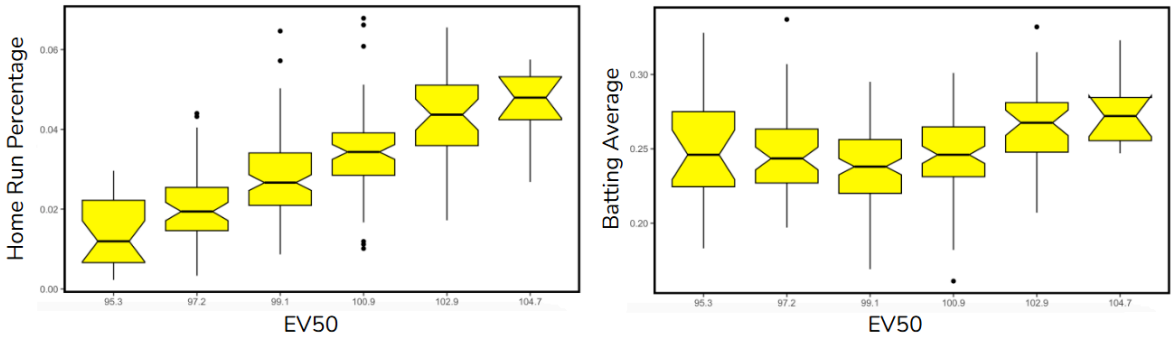


Figure 4: Boxplots showing distribution of EV50 with respect to home run percentage and batting average

A-7

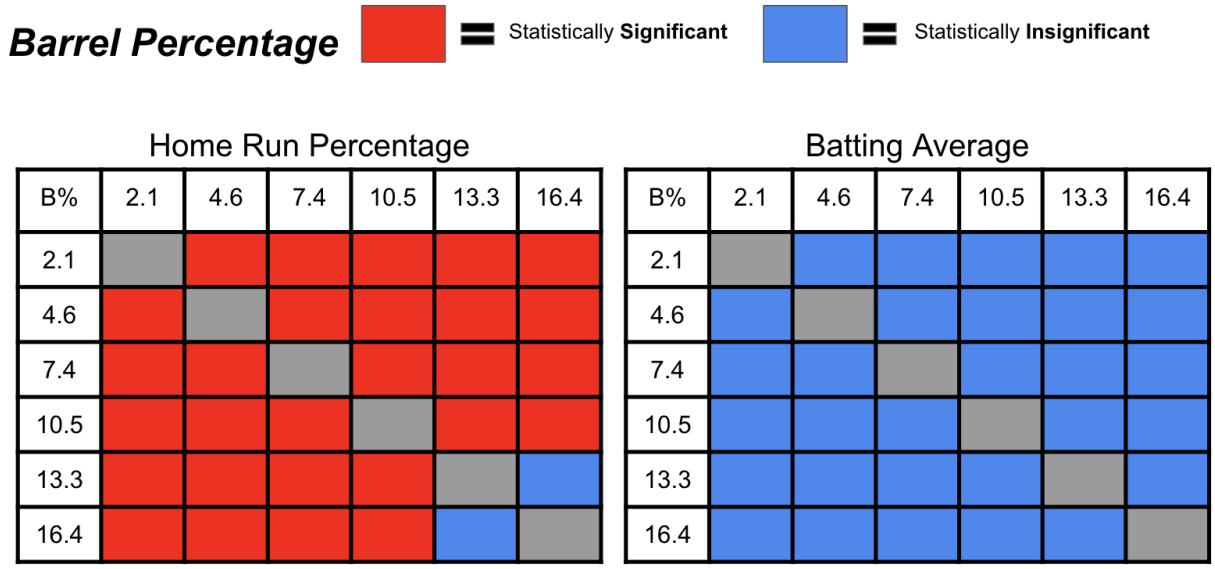


Figure 5: Games Howell test of Barrel Percentage parameters, vs Home Run Percentage and Batting Average. Red squares indicate statistically significant results, blue squares indicate statistically insignificant results.

A-8

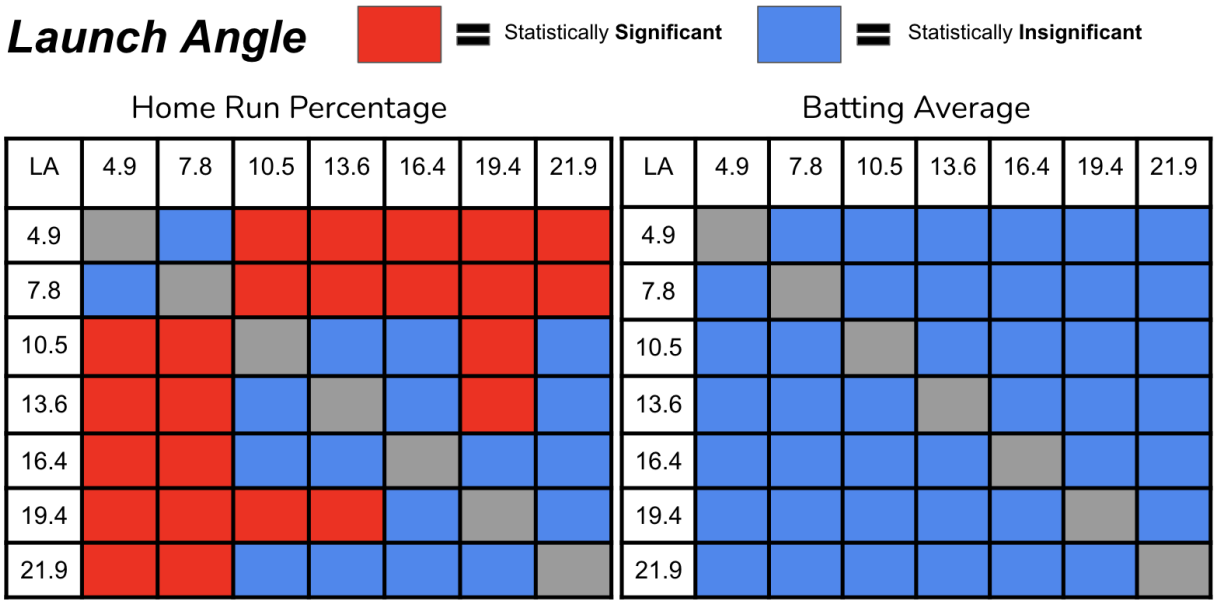


Figure 6: Games Howell test of Launch Angle parameters, vs Home Run Percentage and Batting Average. Red squares indicate statistically significant results, blue squares indicate statistically insignificant results.

A-9

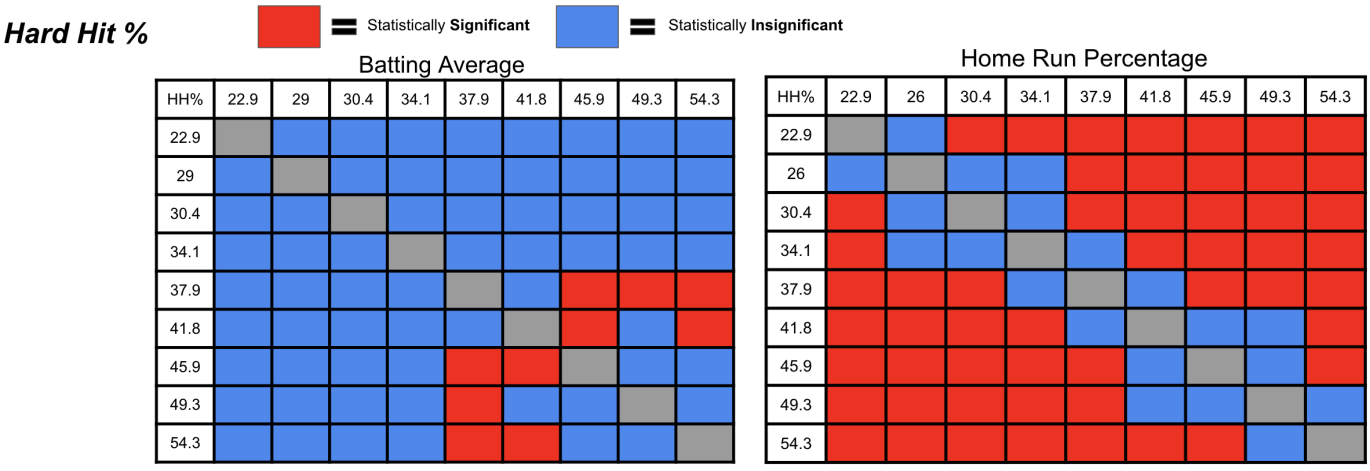
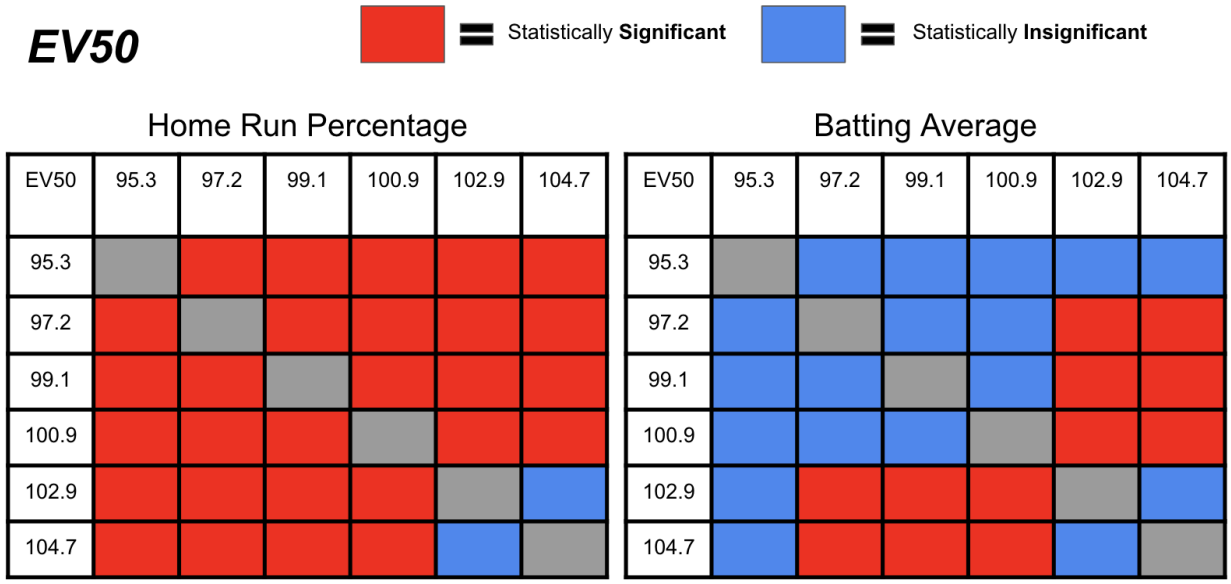


Figure 7: Games Howell test of Hard Hit percentage parameters, vs Home Run Percentage and Batting Average. Red squares indicate statistically significant results, blue squares indicate statistically insignificant results.

A-10



EV50	95.3	97.2	99.1	100.9	102.9	104.7
95.3						
97.2						
99.1						
100.9						
102.9						
104.7						

EV50	95.3	97.2	99.1	100.9	102.9	104.7
95.3						
97.2						
99.1						
100.9						
102.9						
104.7						

Figure 8: Games Howell test of EV50 parameters, vs Home Run Percentage and Batting Average. Red squares indicate statistically significant results, blue squares indicate statistically insignificant results.

A-11

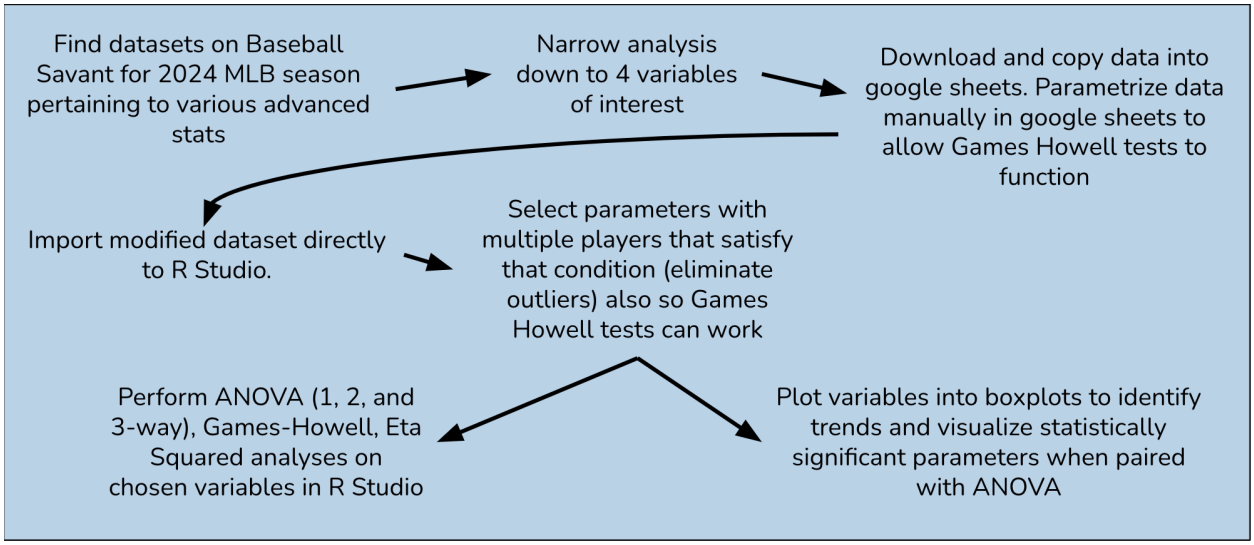


Figure 9: General flowchart of the project

Work Cited

1. Slusser, S. Michael Lewis on A's 'Moneyball' legacy. SFGate (2011). https://www.sfgate.com/athletics/article/Michael-Lewis-on-A-s-Moneyball-legacy-2309126.php
2. Nesbitt, S. J., Dodd, R., Sarris, E., Statcast at 10: From MLB's secret project to inescapable part of modern baseball. The Athletic (2024). https://www.nytimes.com/athletic/5627303/2024/07/10/mlb-statcast-10-year-anniversary/
3. Smith, Dave. A Number of Changes. National Baseball Hall of Fame https://baseballhall.org/discover-more/stories/baseball-history/changing-nature-of-statistics
4. Silverman, J., How Sabermetrics Works. HowStuffWorks (2024). https://entertainment.howstuffworks.com/sabermetrics.htm
5. Posnanski, J., Why the Tampa Bay Rays Are Baseball's True Moneyball Franchise. Esquire (2024). https://www.esquire.com/sports/a60441972/tampa-bay-rays-mlb-2024/
6. Clegg, C., Statcast 101: Barrels, Launch Angle, and Sweet Spot Percentage. The Dynasty Dugout (2023). https://www.thedynastydugout.com/p/statcast-101-barrels-launch-angle-sweet-spot
7. Chen, A. The Metrics System: How MLB's Statcast is creating baseball's new arms race. Sports Illustrated (2016). https://www.si.com/mlb/2016/08/26/statcast-era-data-technology-statistics
8. Laurila, David. Josh Donaldson Talks Hitting. Fangraphs. (September 25, 2021) https://blogs.fangraphs.com/josh-donaldson-talks-hitting/

9. Piecoro, N. Former top pick Bryce Jarvis hoping adjustments will unlock his potential. AZCentral (2023). https://www.azcentral.com/story/sports/mlb/diamondbacks/2023/03/03/former-top-pick-bryce-jarvis-hoping-adjustments-will-unlock-his-potential/69964895007/
10. Baseball Savant. MLB. https://baseballsavant.mlb.com/leaderboard/statcast