**Clustering Pitchers to Predict Batter-Pitcher Outcomes**

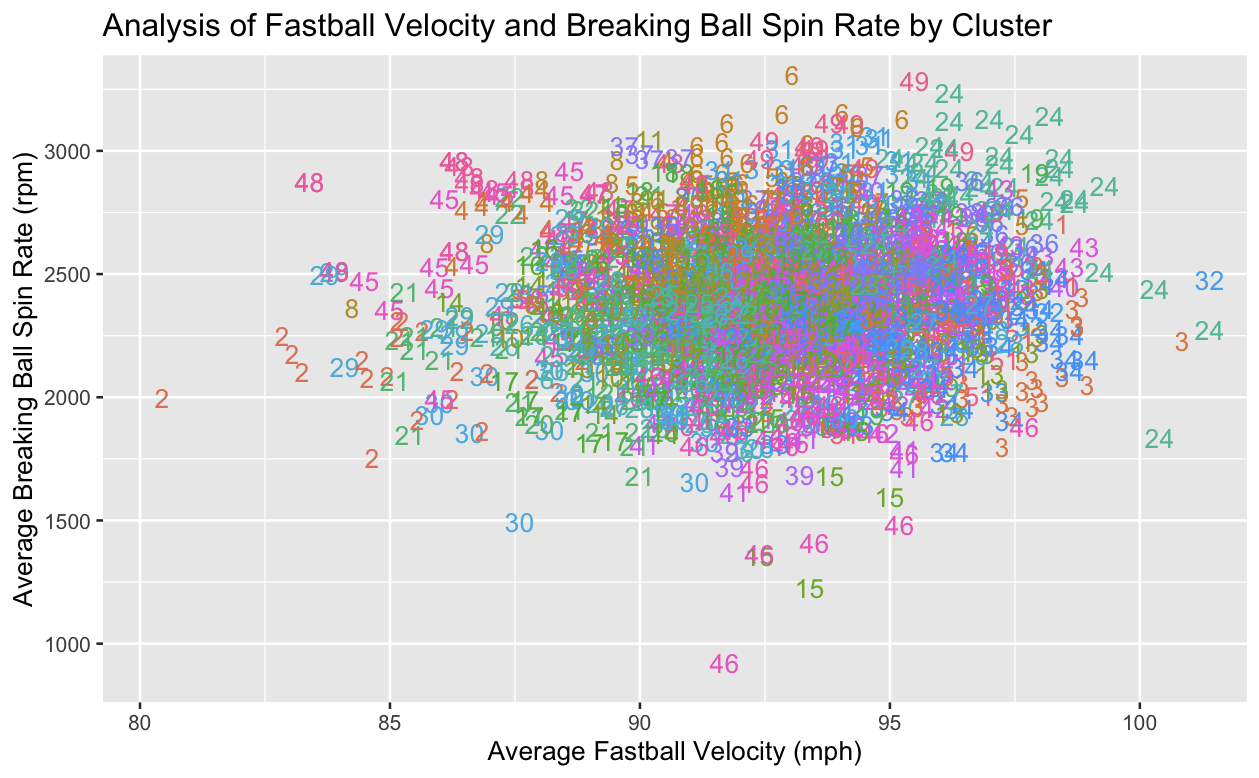
Baseball

**Introduction:**

While watching Major League Baseball, an oft-heard refrain is that this batter “hits this pitcher well.” As research by Tom Tango [1], Dave Cameron [2], and others has shown, this data carries little predictive value, as the sample size of batter-pitcher matchups is too small to draw any meaningful conclusion. But what if the question is if this batter “hits*this type of pitcher* well?” While every pitcher has unique characteristics, pitchers can be grouped via the type, speed, and frequency of pitches they throw. For example, while almost no starter can match the traits of Jacob deGrom, high-powered relievers such as Edwin Diaz bear close resemblance in the characteristics of their pitches. Therefore, is it possible that if a batter hits Diaz (and other similar pitchers) well, that the batter would be a good matchup against deGrom? This paper tests this notion by clustering pitchers based on pitch repertoires and aggregating batters’ records against each cluster to model batting average in a given at-bat. This model is compared to a similar model, but with the batters’ records versus the pitcher in a given at-bat used as inputs rather than the batters’ record against the entire cluster.

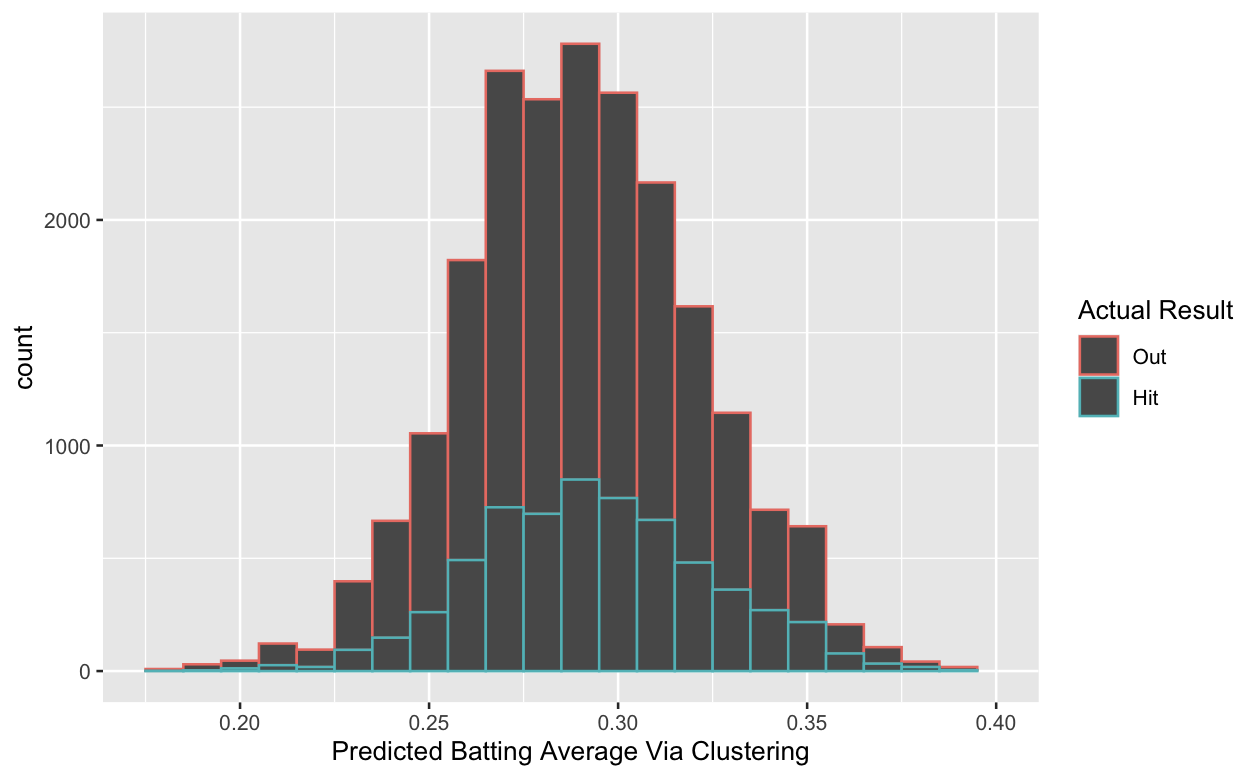
**Methods:**

For this paper, data is used from *Baseball Savant* [3] to find plate appearance records and pitch repertoire characteristics from 2017-2019. For each pitch that a pitcher throws, the average velocity (mph), average spin rate (rpm), and frequency are used in the clustering. The clusters are produced using Ward’s Agglomerative Hierarchical Clustering with grouping designed to optimize the tradeoff between pitcher similarity and sample size of innings pitched. After clustering, three logistic regression models to predict a hit are created: a model with just pitcher batting average and batting average, a model that also adds and batter-pitcher record (at-bats and hits), and a model that also adds batter-cluster record (at-bats and hits). To test the models, ANOVA, Goodness of Fit, and Hosmer-Lemeshow tests are used.



**Results:**

The results from the tests all show statistically significant (P<.01) evidence that using data from clustered pitchers produces more predictive projections of at-bat outcomes. When applying the models to testing data, the cluster-based model most closely matches the actual outcomes.



**Conclusion:**

This study helps to show that batters doing consistently well against certain “types” of pitchers is more meaningful than an individual pitcher. The utilities of this result are wide-ranging, such as a manager debating which batter to start when the first batter has good numbers against the pitcher and the second batter has good numbers against the entire pitcher’s “type.” This study would advocate for starting the second batter. Similarly, in many gambling or daily-fantasy settings, knowing a batter has strong results against the type of pitcher they are facing is a strong indicator that the batter would a good play. Finally, this study reinforces larger sample sizes’ value when conducting statistical research.

**References**

[1] Tango, T. M., Lichtman, M. G., & Dolphin, A. E. (2014). *The book: Playing the percentages in baseball*. TMA Press.

[2] Cameron, D. (2013, May 14). *The absurdities of batter/pitcher match-up numbers*. FanGraphs Baseball. Retrieved September 20, 2021, from https://blogs.fangraphs.com/the-absurdities-of-batterpitcher-match-up-numbers/.

[3] *Baseball savant: Trending MLB Players, Statcast and visualizations*. baseballsavant.com. (2019). Retrieved September 20, 2021, from https://baseballsavant.mlb.com/.