# Pitch Insights Report

Holly Erickson

# Executive Summary

I used a Major League Baseball (MLB) dataset from 2015 - 2019. My goal was to create actionable insights for players and coaches.

First I used a KMeans Clustering algorithm to create pitch clusters. KMeans works by finding homogeneous subgroups within the data where the data points in each cluster are as similar as possible. The pitches can be thought of in terms of speed and movement profile, consisting of the horizontal and vertical breaks. Using these three features, the data takes up three dimensions. KMeans is looking to minimize the distance between points in a cluster in the 3 Dimensional space. The data was split into 30 clusters for left-handed pitchers, and 30 clusters for right-handed pitchers. From here I performed data analysis on the clusters to determine the strike zone with the most likely chance for key outcomes: whiffs, soft contacts and ground balls.

I also predicted the exit speed of a ball coming off of a bat using a random forest regressor. The variables for the prediction were limited to those that would be available before the pitch as well as pitch characteristics. This approach combines multiple decision trees in determining the final output. A decision tree uses a tree-like graph or model of decisions to build a flow-chart. Each path in the flow-chart leads to its predictions.

# Technical Report

## Introduction and Background

Pitches can be thought of in terms of speed and movement (horizontal and vertical break), which I could get from the data using features pfxx and pfxz. Here is how Trackman defines these terms:

pfxx: The horizontal (left-right) movement of the pitch during the last 40 feet before the front of home plate, as compared to a theoretical pitch thrown at the same speed with no spin-induced movement.

pfxz: The vertical (up-down) movement of the pitch during the last 40 feet before the front of home plate, as compared to a theoretical pitch thrown at the same speed with no spin-induced movement. (Woods, 2019)

I clustered the pitches into subgroups and performed analysis on these subgroups to find patterns in their outcomes. I also created a model to predict the speed of the ball, measured in miles per hour, as it comes off the bat at the moment of contact.

## Methods

I began by splitting pitch data into pitcher handedness (right or left). The reason for this is that the release point is different for these pitches, so even if the speed and breaks are similar, the pitch will be very different from the batter's perspective.

I used the KMeans algorithm to divide pitches into clusters based on pitch speed, pfx_x and pfx_z. The data was standardized prior to clustering using scikit learn's

StandardScaler. The reason for this is that KMeans minimizes the sum of the squared euclidean distance between the data points and the cluster's centroid (arithmetic mean of all the data points that belong to that cluster). Leaving inputs at different scales is equivalent to putting more weight on variables with smaller variance.

The standard score of a sample x is calculated as:

$z = (x - u) / s$

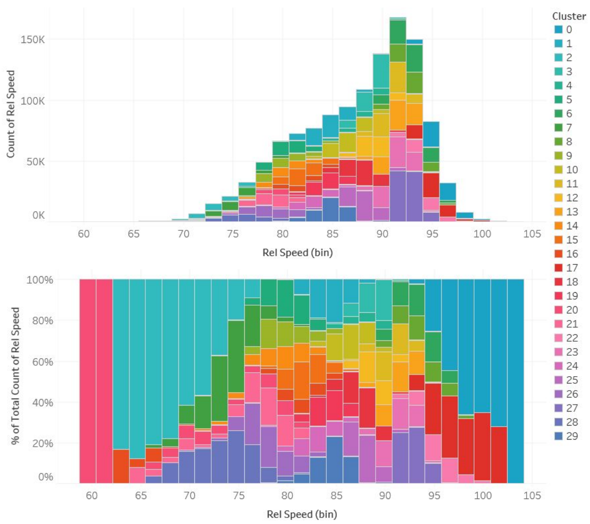
where u is the mean of the training samples and s is the standard deviation of the training samples. (scikit-learn, 2020)

In order to minimize overlap between clusters, I decided not to create separate clusters for each pitch type. The name of the pitch type (fastball, curveball, etc.) does not matter to the batter, and it's very possible for two pitches tagged as different types to have the same movement profile.

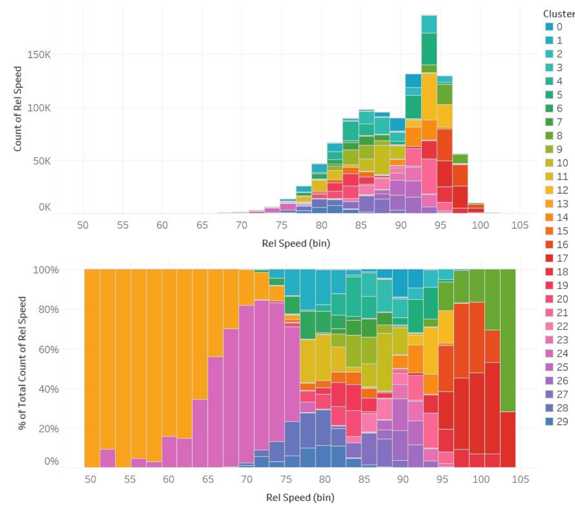
In determining the optimal number for k, I looked at using the elbow method and the silhouette score. The elbow method works by plotting the sum of squared errors (SSE) for different values of k. This value decreases toward 0 as we increase k, and the elbow usually represents where we start to have diminishing returns by increasing k. It does not work well when the data is tightly grouped though, as was the case for the MLB data. The silhouette value is a measure of how similar an object is to its own cluster compared to other clusters. The silhouette ranges from −1 to +1. Again, due to having tightly grouped data without clearly separated clusters, the silhouette score was giving me negative values for any choice k.

Instead, I plotted the data for different values of k, looking for fairly tight spread within the clusters over the three features. Here are is the spread of the data with k = 30. The lower chart in each image shows the cluster % of the total for each bin:

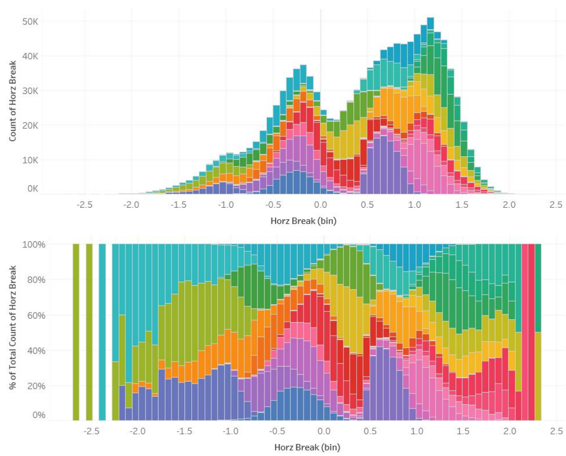
Left Speed:



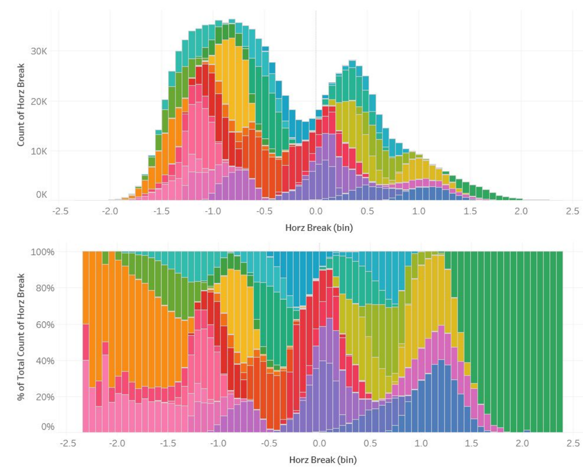

Right Speed:



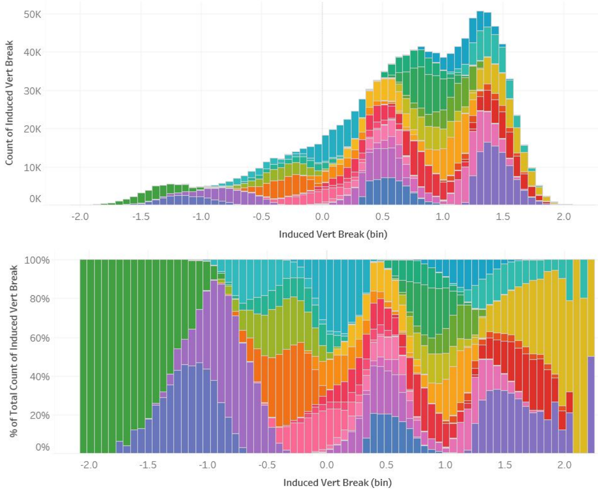

Left Horizontal Break:



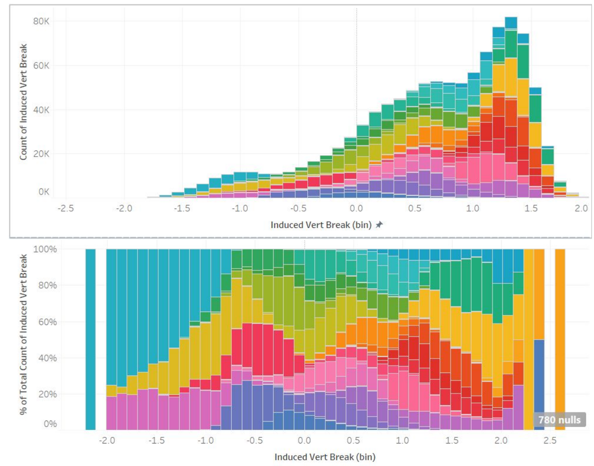

Right Horizontal Break:



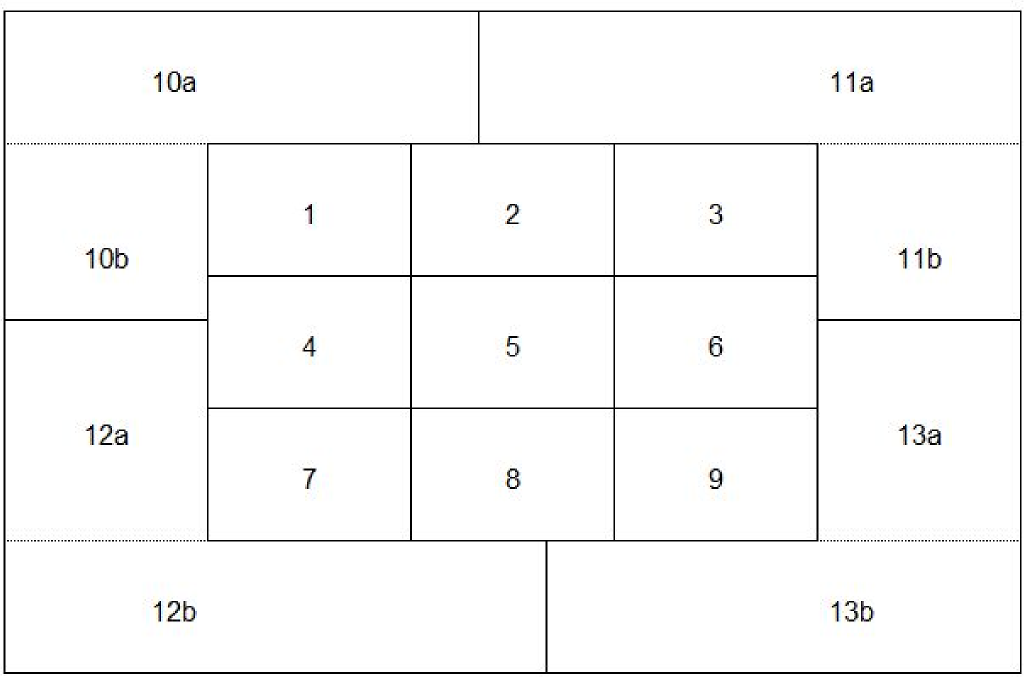
Left Vertical Break:




Right Vertical Break:




Next, I determined which zone the pitches crossed the plate in. The strike zones are labeled according to this chart, where 1 - 9 would be considered inside the strike zone and 10 - 13 are considered balls by an umpire with a good eye.




(Roegele, 2018)

I find the most common zone for key outcomes for each cluster: whiffs, soft contacts and ground balls. Whiffs are swinging strikes. Soft contact is considered a ball hit under 75 mph. Ground balls are hit with a launch angle under 10. More on these outcomes can be found in the Results section of this paper.

Next I created a model to predict the exit speed of a ball off a bat based on features found in the data sets. One of the possible input features was total spin rate (How fast the ball is spinning as it leaves the pitcher's hand, reported in the number of times the pitched ball would spin per minute ("revolutions per minute" or "rpm"). However the total spin rate is not reflective of the true spin rate as it impacts the ball, as the total spin is made up of a combination of transverse and gyrospin. The gyrospin is similar to that of a bullet, and does not impact movement. To get the transverse spin (true spin) I followed the methods prescribed by Alan Nathan. (Nathan, 2018).

To settle on the features I used a combination of Pearson's Correlation, Kbest selector with f regression, random forest feature importance, recursive feature importance with linear regression, and a meta-transformer for selecting features based on importance weights. I found which features were weighted the highest out of these five methods. My final list of features is ['BatterSide', 'HorzBreak', 'InducedVertBreak', 'PlateLocHeight', 'PlateLocSide','RelSpeed', 'Spin Efficiency','True Spin (rpm)','ay', 'release_pos_x', 'release_pos_y', 'release_pos_z',  'balls', 'strikes', 'Elevation']

I tried several regression-based methods for my model, and determined that Random Forest Regressor performed the best of these. This could be due to the correlation between some of the input features causing models like lasso and ridge regression to become unstable. I used a grid search cv for hyperparameter tuning for each of these models.

## Results

For each cluster, I determined the highest success percentage for the three key outcomes (whiff %, soft contact %, ground ball%) at 4 levels:

1. Top or bottom of the zone and inside or outside side of the zone
2. Top or bottom of the zone and inside or outside side of the zone for left-handed and for right-handed batters
3. Zone (1 - 13)
4. Zone (1 - 13) for left-handed and for right-handed batters

I put the results for each cluster into two spreadsheets. insights.csv shows the top performer at each of the levels. It includes the sample size, percent that it occurred in that region, the variable it is calculating, and a string that details the output in a way that a pitcher or coach could easily understand. Note that ground ball and soft contact percents are out of balls in that region that are hit into play, not out of all pitches. This matches how GBP and SCP are calculated by the MLB. Whiff percent is out of all pitches. The sample size is the size within that region / zone.

| Cluster | level | sample | percent | var | sample desc | string |
|---|---|---|---|---|---|---|
| 5 | Best Zone | 12 | 50 | SCP | zone [1] | Your highest SCP is in zone [1] at 50.0%. |
| 5 | Best Zone | 4842 | 26.8 | Whiff | zone [9] | Your highest Whiff is in zone [9] at 26.8%. |
| 6 | Best Zone | 846 | 81.6 | GBP | zone [9] | Your highest GBP is in zone [9] at 81.6%. |
| 6 | Best Zone | 348 | 56.9 | SCP | zone [11] | Your highest SCP is in zone [11] at 56.9%. |
| 6 | Best Zone | 3984 | 13.6 | Whiff | zone [3] | Your highest Whiff is in zone [3] at 13.6%. |
| 7 | Best Zone | 60 | 100 | GBP | zone [3] | Your highest GBP is in zone [3] at 100.0%. |
| 7 | Best Zone | 60 | 60 | SCP | zone [1] | Your highest SCP is in zone [1] at 60.0%. |
| 7 | Best Zone | 1704 | 16.9 | Whiff | zone [7] | Your highest Whiff is in zone [7] at 16.9%. |
| 8 | Best Zone | 120 | 90 | GBP | zone [12] | Your highest GBP is in zone [12] at 90.0%. |
| 8 | Best Zone | 120 | 45 | SCP | zone [12] | Your highest SCP is in zone [12] at 45.0%. |
| 8 | Best Zone | 2406 | 14.5 | Whiff | zone [3] | Your highest Whiff is in zone [3] at 14.5%. |
| 9 | Best Zone | 210 | 85.7 | GBP | zone [13] | Your highest GBP is in zone [13] at 85.7%. |
| 9 | Best Zone | 60 | 60 | SCP | zone [1, 11] | Your highest SCP is in zone [1, 11] at 60.0%. |
| 9 | Best Zone | 2688 | 22.3 | Whiff | zone [7] | Your highest Whiff is in zone [7] at 22.3%. |
| 10 | Best Zone | 48 | 75 | GBP | zone [1] | Your highest GBP is in zone [1] at 75.0%. |
| 10 | Best Zone | 144 | 45.8 | SCP | zone [12] | Your highest SCP is in zone [12] at 45.8%. |
| 10 | Best Zone | 4998 | 16 | Whiff | zone [9] | Your highest Whiff is in zone [9] at 16.0%. |
| 11 | Best Zone | 84 | 64.3 | GBP | zone [12] | Your highest GBP is in zone [12] at 64.3%. |
| 11 | Best Zone | 360 | 36.7 | SCP | zone [10] | Your highest SCP is in zone [10] at 36.7%. |
| 11 | Best Zone | 5226 | 15.6 | Whiff | zone [2] | Your highest Whiff is in zone [2] at 15.6%. |
| 12 | Best Zone | 774 | 89.9 | GBP | zone [13] | Your highest GBP is in zone [13] at 89.9%. |
| 12 | Best Zone | 174 | 51.7 | SCP | zone [11] | Your highest SCP is in zone [11] at 51.7%. |
| 12 | Best Zone | 10224 | 13.7 | Whiff | zone [13] | Your highest Whiff is in zone [13] at 13.7%. |
| 13 | Best Zone | 192 | 81.2 | GBP | zone [13] | Your highest GBP is in zone [13] at 81.2%. |
| 13 | Best Zone | 156 | 38.5 | SCP | zone [10] | Your highest SCP is in zone [10] at 38.5%. |

| Cluster | level | sample | percent | var | sample desc | string |
|---|---|---|---|---|---|---|
| 13 | Best Zone LR Split | 138 | 39.1 | SCP | zone [10] | Against R handed batters: Your highest SCP is in zone [10] at 39.1%. |
| 13 | Best Zone LR Split | 96 | 56.2 | SCP | zone [11] | Against L handed batters: Your highest SCP is in zone [11] at 56.2%. |
| 13 | Best Zone LR Split | 1356 | 14.2 | Whiff | zone [2] | Against L handed batters: Your highest Whiff is in zone [2] at 14.2%. |
| 13 | Best Zone LR Split | 2826 | 17.8 | Whiff | zone [3] | Against R handed batters: Your highest Whiff is in zone [3] at 17.8%. |

The second spreadsheet, insights_full_list.csv shows the success percentage for each region/zone for the three key outputs, not just the top performer. This way a player could see what the second-best, or perhaps least successful place to aim would be.

For my random forest regressor predicting the exit speed of the ball off the bat using features in the dataset that would be available before the pitch as well as pitch characteristics, the best score I was able to achieve was .28. This is the R^2 score for the model. R^2 (coefficient of determination) has a best possible score of 1.0 and it can be negative. A constant model that always predicts the expected value of y, disregarding the input features, would get a R^2 score of 0.0. “It represents the proportion of variance (of y) that has been explained by the independent variables in the model. It provides an indication of goodness of fit and therefore a measure of how well unseen samples are likely to be predicted by the model, through the proportion of explained variance.” (Scikit-Learn, 2020)

Baseball data is notoriously noisy, so I was happy to have an $R^2$ value that did indicate that my model explained some of the variance. The hyperparameters for this model were a max-depth of 10, with n_estimators set to 500.

## Discussion

There is lots of potential to expand on these ideas. I looked at three key outputs from the clusters, but you could easily repeat these methods to find other metrics such as hard hit percentage, foul balls, etc.

The same could be said for the model to predict exit speed. A similar model could be constructed to predict launch angle, etc. The challenge with baseball data is that there are many factors to consider. Two identical pitches will not always perform the same, even against the same batter. However, we have shown that some of the variables can be explained with the data on-hand.

## Acknowledgments

I would like to acknowledge Alan Nathan for his help in confirming I was calculating the True Spin correctly. I would like to acknowledge my classmates for their thought-provoking discussion throughout the course. I would like to thank my instructor, Catherine Williams, for her guidance and positive feedback throughout the process.

## References

Dabbura, A. (2018, Sept 17). *“K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks.”* Retrieved from https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a.

Nathan, A. (2018). *“Determining the 3D Spin Axis from Statcast Data.”* Retrieved from http://baseball.physics.illinois.edu/trackman/SpinAxis.pdf

Roegele, J. (2018, March 28). *“The 2017 Strike Zone.”* Retrieved from https://tht.fangraphs.com/the-2017-strike-zone/

Scikit-Learn. (2020). *“Metrics and scoring: quantifying the quality of predictions¶.”* Retrieved from https://scikit-learn.org/stable/modules/model_evaluation.html#r2-score

Scikit-Learn. (2020). *“sklearn.preprocessing.StandardScaler.”* Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

Wood, J. (2019) *“Radar Measurement Glossary of Terms.”* Retrieved from https://trackman.zendesk.com/hc/en-us/articles/115002776647-Radar-Measurement-Glossary-of-Terms