Sports Science Analysis

Ruslan Davtian

Problem 1

Data Description & Project Goal

In this analysis, we are given two data sets from the 2019 season for Trackman and Blast Motion. The trackman data includes typical batted ball characteristics and in-game information metadata associated with typical Statcast pitch by pitch level data. The blast motion data displays metrics from practice related to swing mechanics such as bat speed, bat acceleration, swing angle, body tilt & vertical bat angle, etc. In total, there are 104 unique batters and all batters have Blast Motion data from practice. However, 42 of those hitters do not have any game data. No missing values exist within the practice data but there are missing batted ball characteristics for many of the pitches since not every pitch resulted in a batted ball. Below is an output of the first five rows of each data set. The first table corresponds with Blast Motion data and the other two tables display all variables within the trackman data set.

BatterId	Date	AttackAngle	BatSpeed	Connection	EarlyConnection	Handedness	PlanarEfficiency	RotationalAcceleration
2e612ce7	2019-01-02	0.111074	30.49020	1.428424	1.507817	5	0.727937	145.1709
2e612ce7	2019-01-02	0.222480	29.83865	1.358282	1.442910	5	0.761305	124.7508
2e612ce7	2019-01-02	0.126757	29.61909	1.339027	1.466272	5	0.713503	119.2641
2e612ce7	2019-01-02	0.248148	29.01311	1.422598	1.557318	5	0.683808	125.4251
367fb7f9	2019-01-06	0.149912	31.72581	1.501380	1.344469	5	0.771255	182.8401

Date	Inning	Top	Outs	Balls	Strikes	PitcherId	BatterId	Bats	Throws	PitchNumber	PAofInning
2019-04-30	4	Top	1	0	0	710e55d6	f70b0d82	Right	Right	123	4
2019-04-30	4	Top	1	0	1	710e55d6	f70b0d82	Right	Right	124	4
2019-04-30	4	Top	1	0	2	710e55d6	f70b0d82	Right	Right	125	4

Date	Inning	Top	Outs	Balls	Strikes	PitcherId	BatterId	Bats	Throws	PitchNumber	PAofInning
2019-05-06	5	Bottom	0	0	0	bf435272	b4417992	Right	Right	105	2
2019-05-06	5	Bottom	0	1	0	bf435272	b4417992	Right	Right	106	2

PitchofPA	PlateSide	PlateHeight	ExitSpeed	VertAngle	HorzAngle	HitSpinRate	PitchType	PitchCall	PlayResult	HitType
1	0.0856740	2.463370	NA	NA	NA	NA	Curveball	StrikeCalled	Undefined	Undefined
2	0.6468201	2.623517	NA	NA	NA	NA	Fastball	FoulBall	Undefined	Undefined
3	0.3048246	1.252048	NA	NA	NA	NA	Curveball	BallCalled	Undefined	Undefined
1	1.0006599	3.386153	NA	NA	NA	NA	Fastball	BallCalled	Undefined	Undefined
2	0.5828917	2.981364	NA	NA	NA	NA	Fastball	StrikeSwinging	Undefined	Undefined

The coaching staff is interested in building improvement plans for each hitter in order to improve their damage on contact. The main goal in this research is to aid the coaching staff by grouping similar hitters together based on their hitting attributes. This will make it easier for the player development staff to implement and tailor their training programs to groups of players with similar hitting mechanics and performance.

Data Transformations

Before any modeling or data visualizations, I needed to perform necessary data cleaning/filtering steps as well as feature engineer possibly useful variables. Since we are interested in basing group decisions from damage on contact, I decided to create a quality of contact measure by assigning a value to each play result that represents how many runs on average that result is worth. This is called wOBA value or run expectancy of each outcome and it will represent our quality of contact measure. We can use wOBA linear weights as such, (Outs = 0, Error = 0.9, Single = 0.9, Double = 1.25, Triple = 1.6, Home Run = 2). Furthermore, I created a few indicator variables that tracked whether or not a pitch resulted in a swing, a pitch landed inside the strike zone, and if a batter swung on a pitch outside the strike zone. Next, I used the hit type variable given (line drive, ground ball, popup, fly ball) and created hit type distributions for each batter like line drive percentage. There are several switch hitters in the data set so based on the pitcher throwing hand, I was able to regroup those individuals to left handed batter or right handed batter. Typically, early connection and connection blast motion metrics are reported in degrees but the data given reports in radians. Therefore, I transformed those variables into degrees using the formula $degrees = radians \times \frac{180}{\pi}$.

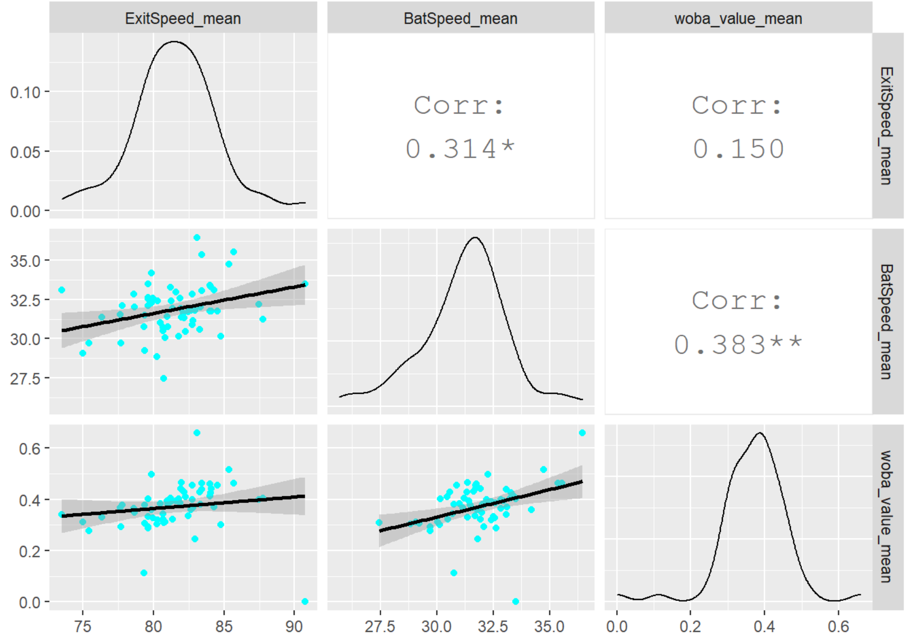
The remaining data transformation steps involve aggregating the data by player and merging the two data sets together. Since we are interested in grouping at the player level, the input features for modeling will be aggregations of the variables by player. I decided to perform mean aggregation for each of the data sets and joined both averaged data by batter id. In the following section, we will visualize our new data set to get a better understanding of the relationships between some of the variables.

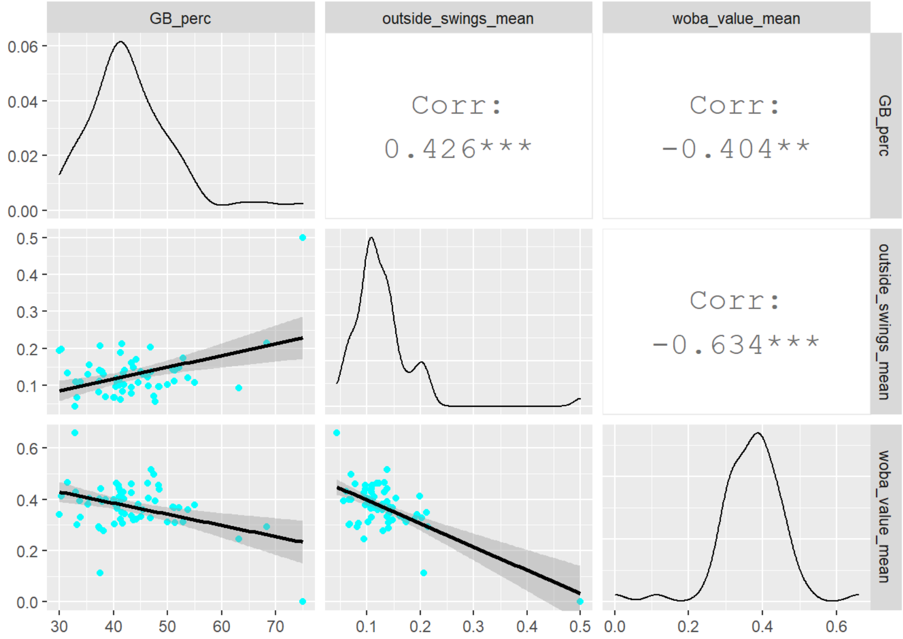
Data Visualizations

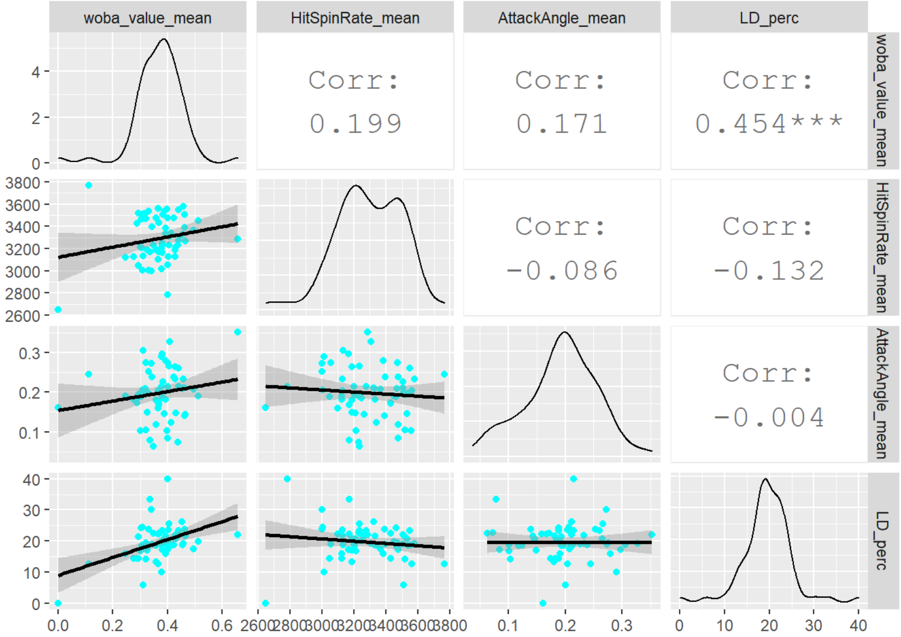
In any analysis project, it is important to visualize the data before model building. Below are advanced scatterplot matrices that show correlations, univariate density, and bivariate scatterplots. It is important to reiterate that we are now looking at averaged metrics per player and not the raw data. Also, we are interested in finding how all variables are related to damage on contact. Therefore, I included wOBA value in each plot and we can compare all other variables relative to wOBA value. There are 62 players with trackman data so we only have wOBA values for those players which means the scatterplots are plotting 62 observations.

As expected, the left plot in the first row shows positive correlation between contact damage and average speed. The harder the ball is hit and the faster the bat swing, the more damage is done on average. In contrast, the top right plot shows negative correlation between ground ball percentage and outside swings (proportion of pitches resulting in swings on pitches outside the zone) against damage on contact. Lastly, the plot on the bottom row shows slight positive correlation between wOBA value and attack angle, hit spin rate, and line drive percentage. A positive attack angle value indicates swinging up, and a negative value indicates swinging down, where zero is perfectly level. This means that we associate line drives and fly balls (more damage) with higher attack angles than ground balls (less damage).

Since we have small sample sizes, it is important to acknowledge that we may not have enough degrees of freedom to include every variable in a model and still get good results. Therefore, I chose to focus on the variables plotted above as they have the most linear associations in comparison to wOBA value. The next section will focus on the setup and implementation of the modeling algorithm for clustering hitters.


ExitSpeed_mean
BatSpeed_mean
woba_value_mean
Corr:
0.314*
Corr:
0.150
Corr:
0.383**
ExitSpeed_mean
BatSpeed_mean
woba_value_mean
0.10
0.05
0.00
35.0
32.5
30.0
27.5
0.6
0.4
0.2
0.0
75
80
85
90
27.5
30.0
32.5
35.0
0.0
0.2
0.4
0.6


GB_perc
outside_swings_mean
woba_value_mean
Corr:
0.426***
Corr:
-0.404**
Corr:
-0.634***
GB_perc
outside_swings_mean
woba_value_mean
0.06
0.04
0.02
0.00
0.5
0.4
0.3
0.2
0.1
0.6
0.4
0.2
0.0
30
40
50
60
70
0.1
0.2
0.3
0.4
0.5
0.0
0.2
0.4
0.6


woba_value_mean
HitSpinRate_mean
AttackAngle_mean
LD_perc
Corr:
0.199
Corr:
0.171
Corr:
0.454***
Corr:
-0.086
Corr:
-0.132
Corr:
-0.004
woba_value_mean
HitSpinRate_mean
AttackAngle_mean
LD_perc
4
2
0
3800
3600
3400
3200
3000
2800
2600
0.3
0.2
0.1
40
30
20
10
0
0.0
0.2
0.4
0.6
2600
2800
3000
3200
3400
3600
3800
0.1
0.2
0.3
0
10
20
30
40

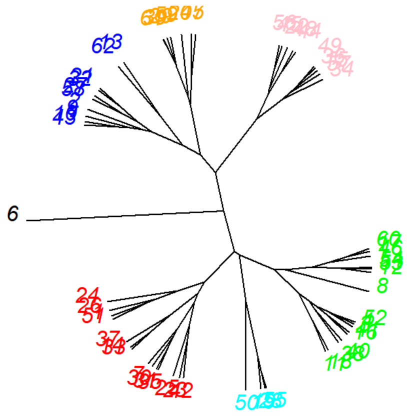
Methodology and Pre-Processing

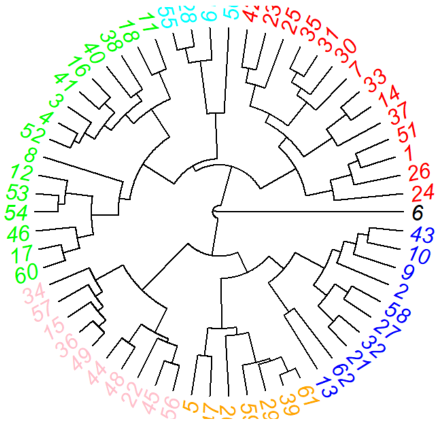
We are interested in clustering all players into groups regardless of how many game or practice reps. There may be an outlier due to batters seeing very few game or practice pitches but I felt it important to cluster every player for the coaching staff. As a result, I will run two clustering models: one for the 62 players with both practice and game data and the other 42 players with just practice data. In terms of variable selection, I used the rule-of-thumb of square root of sample size as a cut-off range for the number of variables to include in the modeling stage. I checked for multicollinearity by calculating variance inflation factors but there are no issues with variables being collinear.

Due to small samples sizes (unique number of players) and not knowing the exact number of groups in advance, I chose to use agglomerative (bottom-up) hierarchical clustering instead of kmeans. Each variable is standardized before computing dissimilarities between the variables with Euclidean as the distance metric. Lastly, I implemented the default Ward's method that minimizes the total within-cluster variance. This clustering setup is the same for both clustering problems. For visual purposes, batter id labels were removed in favor of number labels (1-62) but the order is kept the same. Batter 1 corresponds with the first batter ID in alphabetical order. We can see the results of the first cluster groupings in the next section.

Clustering Game Hitters

Below are two fancier versions of dendrograms from hierarchical clustering. A typical dendrogram such as a bracket style may be harder to interpret with 62 players but I believe viewing it as spaced out tree branches makes viewing clusters easier. Looking at the left plot, we can see five major groups (top left, top right, bottom right, bottom left, and a single node on the side). However, I chose to cut the dendrogram into these seven groups but if needed for the player development staff, we could re-adjust the number of groups. The right dendrogram is also included since the player numbers are easily viewable. These groupings are meaningless without providing numerical summaries of the variables to see the strengths and weaknesses of the groups and what makes the groupings different. However, the biggest question to address is why player 6 is by himself.





black	blue	cyan	green	orange	pink	red
1	10	4	16	7	10	14

* Number of Batters Per Group

BatterId	ExitVelo	HitSpinRate	AttackAngle	wOBAValue	BatSpeed	OutsideSwing%	LD%	GB%	RotAcceleration
1a22f94b	3.1	-3.127	-0.586	-4.125	1.016	5.942	-3.443	3.792	1.199

* Batter 6 Average Metrics Standardized

We observe that this player is a clear outlier with multiple variables at least 3 standard deviations from the mean. The main reason for his extreme values is that he only appeared in one game on July 9th, 2019 and saw a grand total of 11 pitches. In that game, he went 0-2 with a Sac Fly. With this knowledge, we cannot infer anything from his Trackman data so given more in-game at bats, we expect this player to merge into one of the other 4 major clusters. However, he has over 800 swings in practice and we can infer interesting findings from his Blast Motion data. Player 6 has a below average attack angle but above average bat speed and rotational acceleration which are critical for damage on contact. We can take a look at how this player ranks across the other groups in the next table.

Group	AvgRank	ExitVelo	HitSpinRate	AttackAngle	wOBAValue	BatSpeed	OutsideSwing%	LD%	GB%	RotAcceleration
cyan	2.91	2	2	2	1	1	2	1	3	4
red	3.36	5	3	7	4	3	6	3	1	3
green	3.36	3	4	4	2	4	1	2	6	5
pink	4.27	4	5	1	3	6	4	5	5	2
blue	4.45	6	6	5	5	7	3	4	2	7
orange	4.64	7	1	3	6	5	5	6	4	6
black	5.00	1	7	6	7	2	7	7	7	1

* Group Rankings From 1 (best) to 7 (worst) Ordered by Mean Rank

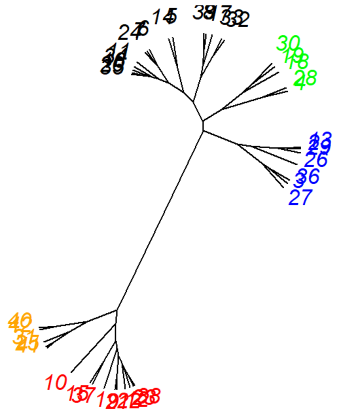
I have ranked each group for each variable according to what is considered good and bad. For exit speed, bat speed, hit spin rate, wOBA value, line drive rate, and rotational acceleration, we associate higher values as better than lower values. In contrast, we associate higher values for ground ball rate and outside swing rate as worse than lower values. Lastly, higher attack angles are preferred over lower attack angles but not always. Ground balls are associated with negative attack angles of values but popups are as a result of too high of an attack angle. However, I will rank larger attack angles as better than lower attack angles since balls in the air (fly balls, line drives) induce more damage than ground balls.

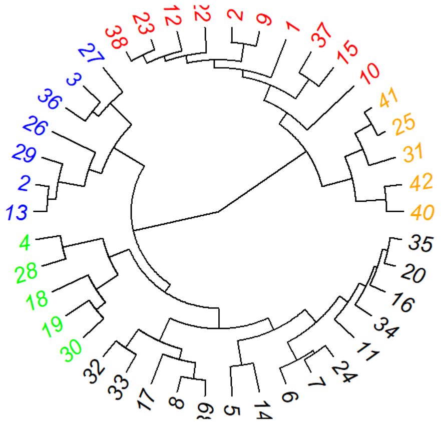
An interesting result is that our average ranking has a strong, positive correlation with wOBA value ranking which we used as a measure of damage on contact. We see that the top two groups are at least average in line drive %, attack angle, hit spin rate, outside swing %, and exit velocity. While the other groups are struggling with some aspect of hitting since they have multiple rankings above 5. Again, player 6 is an exception since he has very little Trackman but his attack angle metric from Blast motion is the lowest.

Clustering Practice Hitters

For the 42 hitters with only Blast Motion data, the clustering method is exactly the same as before except we are only using Blast Motion data to form groups. Therefore, I decided to add two more variables (connection and early connection) into the model that were not present in the previous cluster. These metrics measure the relationship between the body’s tilt and the vertical bat angle at the start of the downswing and at impact. The ideal angles are at 90 degrees which means lower or higher angles are not considered better. Hence, after running the cluster, I ranked the group averages of these variables based on the magnitude away from 90 degrees. In doing so, smaller values are preferred over larger values since smaller values indicate being closer to the ideal angle of 90 degrees.

Looking at the first tree style dendrogram on the left, there seems to be 4 clear groups, but I have broken the bottom group into two. We can also see exactly which players are in which group from the circular dendrogram. A table displaying the frequency of players per group is also shown below. Depending on the player developmental staff plans, 15 players in a group may be too much. In that case, I would only need to cut the dendrogram into the ideal number of groups.





black	blue	green	orange	red
15	7	5	5	10

* Number of Batters Per Group

From the rankings, the top two groups are generally above average in all almost all swing metrics. The black group has below average attack angle in comparison to the other groups but is at least average in all other metrics relative to the other groups. The red and orange groups really struggle in being able to apply power into their swings since they have the worst rankings in early connection, connection, and planar efficiency.

Group	AvgRank	AttackAngle	BatSpeed	EarlyConnection	RotAcceleration	PlanarEfficiency	Connection
green	2.33	1	1	3	5	2	2
black	2.33	4	2	2	2	3	1

Group	AvgRank	AttackAngle	BatSpeed	EarlyConnection	RotAcceleration	PlanarEfficiency	Connection
blue	3.17	5	5	1	4	1	3
red	3.50	2	4	5	1	5	4
orange	3.67	3	3	4	3	4	5

[*] Group Rankings From 1 (best) to 5 (worst) Ordered by Mean Rank

Problem 2

Introduction

Dear Player Development Staff,

I am one of the analysts on the team and have been assigned to investigate a baseball related question of interest for you.

Question & Goal

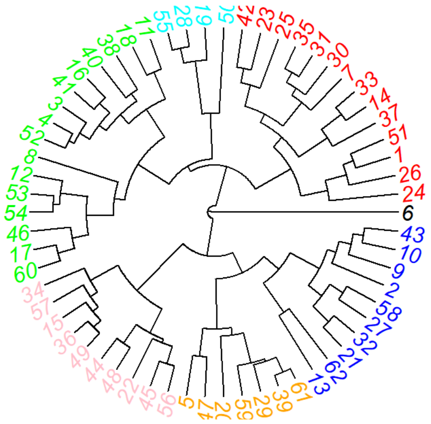
How can we group similar players together in terms of ability to do damage on contact? The goal of this research is to not only cluster our hitters, but identify which hitters are similar so we can target unique developmental hitting plans to these hitters per group. This allows us to identify different strengths and weaknesses among our hitters.

Method

We collected Blast Motion practice data as well as Trackman game data for our hitters for the 2019 season. There are a total of 104 players, but only 62 of these players have both practice and game data. The other 42 players do not have any in-game appearances. Therefore, we decided to run two separate cluster analysis for the players with game data and for the players without game data. Keep in mind that the number of groups is arbitrary and can be changed easily. There is no set in stone number so its depends on how you would like to divide up your training staff to the players. Let's look at the first group of players who have played in games.

Game & Practice

Below is a branched, circular chart that indicates how these groups are formed. Each player starts in its own group, and pairs of groups are merged as one moves up the hierarchy or to the center of the circle. The color scheme tells us which players are in which group. We decided that there are seven groups that are differential in their swing abilities so we went up the hierarchy in the chart until seven groups exist. The next question is what are the strengths and weakness of each group. The ranking chart relative to other groups below answers that question. It's important to note that player 6 only has played in one game so his rankings for Trackman specific data are not accurate. Given more playing time, he will eventually merge into one of the groups. But based on his Blast Motion rankings, we might have an educated guess as to what group he should go to.



Group	AvgRank	ExitVelo	HitSpinRate	AttackAngle	wOBAValue	BatSpeed	OutsideSwing%	LD%	GB%	RotAcceleration
cyan	2.91	2	2	2	1	1	2	1	3	4
red	3.36	5	3	7	4	3	6	3	1	3
green	3.36	3	4	4	2	4	1	2	6	5
pink	4.27	4	5	1	3	6	4	5	5	2
blue	4.45	6	6	5	5	7	3	4	2	7
orange	4.64	7	1	3	6	5	5	6	4	6

Group	AvgRank	ExitVelo	HitSpinRate	AttackAngle	wOBAValue	BatSpeed	OutsideSwing%	LD%	GB%	RotAcceleration
black	5.00	1	7	6	7	2	7	7	7	1

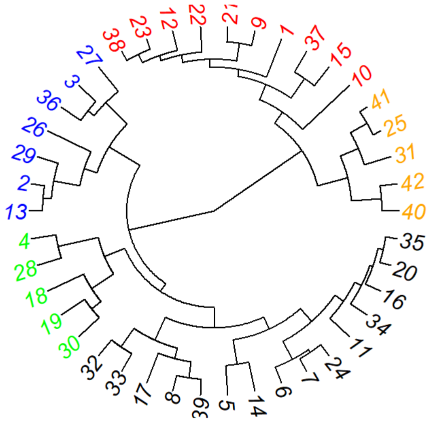
* Group Rankings From 1 (best) to 7 (worst) Ordered by Average Rank

(Relative to own team only)

- cyan
  - strengths: Above average or great in almost all metrics
  - weaknesses: Average at best rotational acceleration, room for improvement here to generate more power at contact
- red
  - strengths: Lowest GB %, good bat speed
  - weaknesses: Lowest group average attack angle, might not be necessarily bad, check to make sure these players don’t chop down on the ball too often (negative attack angle)
- green
  - strengths: High LD %, Good discipline taking outside pitches
  - weaknesses: High GB%, could improve damage on contact by increasing rotational acceleration and bat speed
- pink
  - strengths: Highest group average attack angle, great rotational acceleration,
  - weaknesses: Highest group average attack angle, may not be a weakness, check how often they swing too much up on the ball, check FB% and Popup%. Room for improvement in bat speed which may increase exit velocity and LD%
- blue
  - strengths: Low GB%
  - weaknesses: Lots of room for improvement, lacks power, low exit velo, bad speed, rotational acceleration, hit spin rate
- orange
  - strengths: Highest group average hit spin rate
  - weaknesses: Below average in most metrics, room for improvement for contact damage
- black (Trackman Data Not Accurate, only 1 game played)
  - strengths: Highest average rotational acceleration and good bad speed
  - weaknesses: Given more at bats, he may be one of the best hitters among all players

Practice Only

We took a similar approach to group these players with only practice data and no game reps. Only difference is we could only use Blast Motion metrics so we included all of them to group these hitters. Below is the same type of group chart along with the rankings.



Group	AvgRank	AttackAngle	BatSpeed	EarlyConnection	RotAcceleration	PlanarEfficiency	Connection
green	2.33	1	1	3	5	2	2
black	2.33	4	2	2	2	3	1
blue	3.17	5	5	1	4	1	3
red	3.50	2	4	5	1	5	4
orange	3.67	3	3	4	3	4	5

* Group Rankings From 1 (best) to 5 (worst) Ordered by Average Rank

(relative to own team only)

- green
    - strengths: Highest group average attack angle and bat speed

    - weaknesses: Lowest group average rotational acceleration
- black
    - strengths: Good body positioning with respect to the bat angle throughout swing
    - weaknesses: No clear weakness, second lowest average attack angle but may not be considered a weakness
- blue
    - strengths: Highest group average % of swing on swing plane, great early connection near 90 degrees
    - weaknesses: Low bad speed and rotational acceleration, could increase damage on contact by increasing those metrics
- red
    - strengths: Highest group average rotational acceleration
    - weaknesses: Lowest group average % of swing on swing plane, connection away from 90 degrees throughout swing, opposite from blue group
- orange
    - strengths: No clear strengths, much work needed
    - weaknesses: Average or below average in all metrics

Problem 3

I managed to feature engineer variables such as outside zone swing percentage, batted ball distribution (FB%, GB%, LD%), and wOBA value as a quantitative measure for contact damage. However, I would have liked to compute expected wOBA value based on more Statcast hitting metrics not included such as landing distance, direction, launch angle, hang time. Then, I would use xwOBA value that is created from process-based stats instead of the wOBA value that I defined which came from the actual results. I would also need a lot more batted balls as observations for each each player. Expected wOBA would be a better representation of a player’s true damage on contact measure than wOBA value.

I believe there are several areas where this clustering algorithm setup is least trustworthy. First, I lost information about player variability from simply taking the mean of each metric by player. I could have included standard deviation as well but since I made up my mind in clustering players using aggregations at the player level, the sample size decreased significantly. I was left with small number of observations and could not include too many features into the model (curse of dimensionality). This is the reason why I did not include Early Connection and Connection metrics from Blast Motion into the first clustering problem as well as Vertical & Horizontal Angle. Given more time, I would definitely investigate which combination of variables if not all to include. I did some research and came across a paper that performed hierarchical clustering with repeated measures but due to time and difficulty, I did not try to run clustering with repeated measures. This is certainly something than can be done given every player has multiple practice or game plate appearances. In my analysis, I ran two clusters but I would have liked to run just one cluster on all 104 players but that would require every player to have game and practice reps. Batter ID 1a22f94b certainly influenced results with only having data on 1 game played. Aggregating by mean is not the best approach if players have small sample size. His values were outliers but I decided to not remove any player from the data, but explain the reasoning of why he was in a group of size 1.