Pitch Effects

By: Cameron, Brandon, Willie, and Zach

Model Purpose

- Evaluating what to look for in pitches on an individual level
  - Can be used to evaluate pitchers arsenals
  - Measures effectiveness of different pitch types with different conditions
  - Can be used to evaluate minor league and amateur talent

Data Dictionary

- pitch_type: The type of pitch derived from Statcast.
- release_speed: Pitch velocity when leaving the hand
- release_pos_x: Horizontal Release Position of the ball measured in feet
  from the catcher's perspective.
- release_pos_z: Vertical Release Position of the ball measured in feet from
  the catcher's perspective.
- Zone: quadrant breakdown of strikezone
- batter_side: Side of the plate batter is standing.
- pitcher_throws: Hand pitcher throws with.
- horiz_pos: Horizontal movement in feet from the catcher's perspective.
- Exit_velo: speed at which the ball leaves the bat on contact
- vertical_break: Vertical movement in feet from the catcher's perspective.
- horiz_pos: Horizontal position of the ball when it crosses home plate from the catcher's perspective.
- vert_pos: Vertical position of the ball when it crosses home plate from the catcher's
  perspective.
- effective_speed: Derived speed based on the the extension of the pitcher's
  release.
- release_spin: Spin rate of pitch
- release_extension: Release extension of pitch in feet
- Exit_velo: speed at which the ball leaves the bat on contact
- spin_axis: The Spin Axis in the 2D X-Z plane in degrees from 0 to 360,
  such that 180 represents a pure backspin fastball and 0 degrees represents
  a pure topspin (12-6) curveball

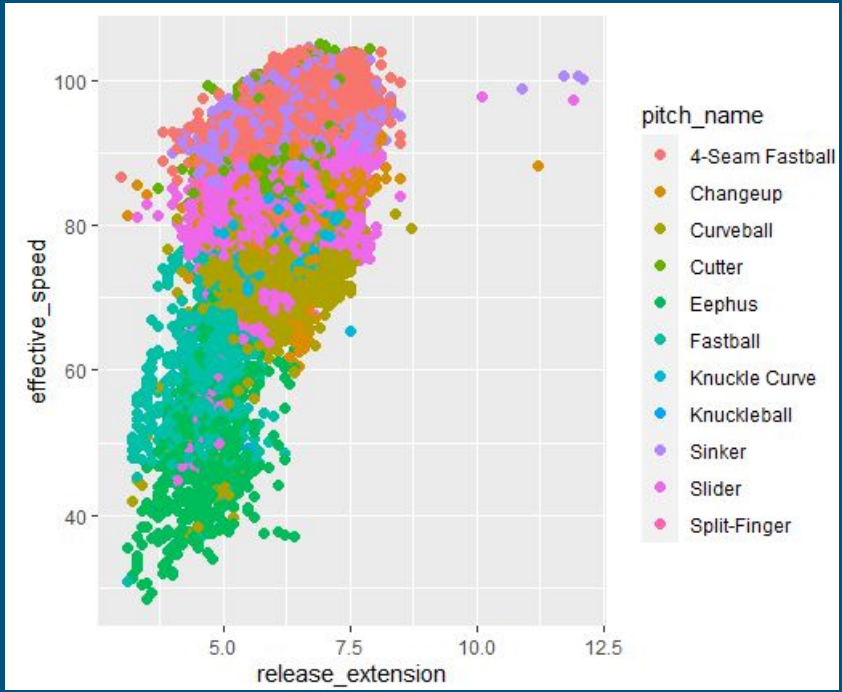
Data Acquisition and cleaning

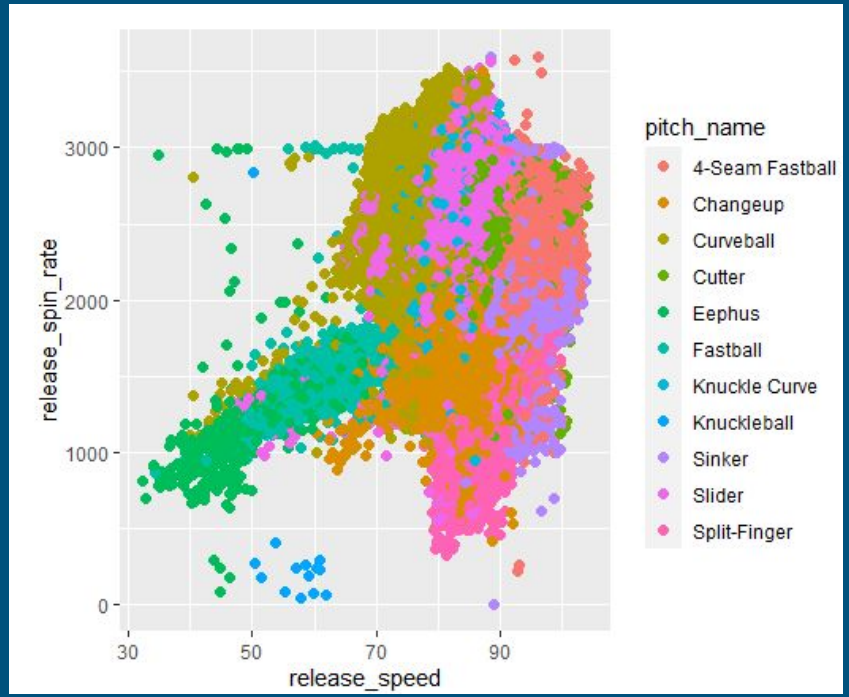
- Scrape Baseball Savant Data
  - Baseballr Package
- Select Dates by 3 day span
- 690,000+ obs and 92 variables
- Drop unnecessary variables
- Drop NA values and check for outliers
- Rename, factor and combine

```
install_github("BillPetti/baseballr")

date1 = baseballr::scrape_statcast_savant(start_date = '2022-04-07',
                                          end_date = '2022-04-10',
                                          player_type = 'pitcher')

date2 = baseballr::scrape_statcast_savant(start_date = '2022-04-11',
                                          end_date = '2022-04-14',
                                          player_type = 'pitcher')

date3 = baseballr::scrape_statcast_savant(start_date = '2022-04-15',
                                          end_date = '2022-04-18',
                                          player_type = 'pitcher')

date4 = baseballr::scrape_statcast_savant(start_date = '2022-04-19',
                                          end_date = '2022-04-21',
                                          player_type = 'pitcher')

date5 = baseballr::scrape_statcast_savant(start_date = '2022-04-22',
                                          end_date = '2022-04-25',
                                          player_type = 'pitcher')
```

Rk.	Player	Year	xBA	xSLG	xwOBA	xOBP	xISO	Avg EV (MPH)	Avg LA (°)	Barrel%
1	Rodon, Carlos	2022	.198	.309	.254	.260	.111	89	19.4	6.5
2	Verlander, Justin	2022	.207	.331	.255	.248	.124	87.8	16.9	6.3
3	Ohtani, Shohei	2022	.204	.311	.256	.260	.107	87.1	14.5	6.3
4	Cease, Dylan	2022	.184	.292	.257	.273	.109	86.8	15	6.2
5	Nola, Aaron	2022	.211	.340	.259	.248	.129	87.7	12.5	7.1

Two Summary Plots to Describe Dataset





Linear Regression-Part 1

- 3 Models: Fastball, Curveball, Changeup
  - Falls under main 3 pitching categories
  - Initial models all featured same variables

Curveball

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     101.402936   6.173524  16.425  < 2e-16 ***
release_speed    -1.174884   0.489729  -2.399 0.016461 *
vert_break       -2.468748   0.382590  -6.453 1.16e-10 ***
release_spin_rate -0.002347   0.000596  -3.937 8.32e-05 ***
zone             -0.887721   0.064077 -13.854  < 2e-16 ***
vert_pos         -1.017755   0.418202  -2.434 0.014970 *
release_extension -2.870490   0.741239  -3.873 0.000109 ***
effective_speed   1.383096   0.477761   2.895 0.003803 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.34 on 7909 degrees of freedom
Multiple R-squared:  0.04519,    Adjusted R-squared:  0.04435
```

Fastball

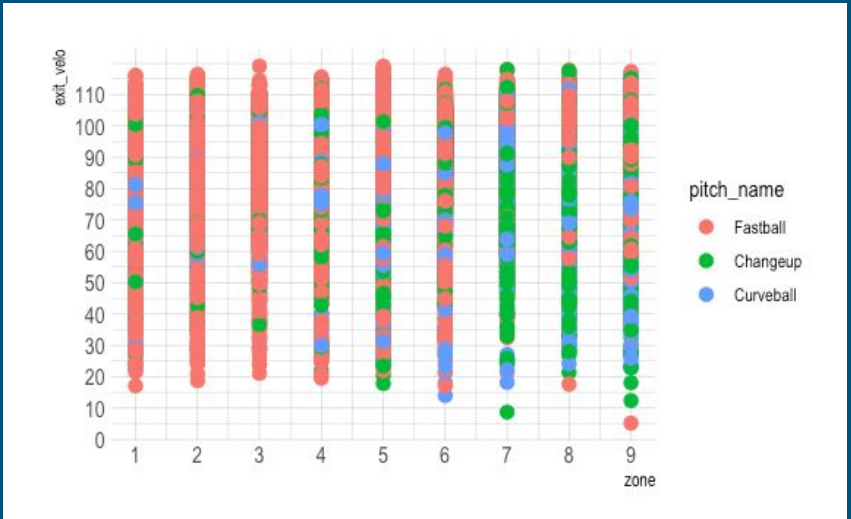
```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     101.636758   2.257532  45.021  < 2e-16 ***
vert_break        1.291985   0.349268   3.699 0.000217 ***
release_spin_rate -0.002110   0.000506  -4.170 3.06e-05 ***
zone             -0.767217   0.024000 -31.967  < 2e-16 ***
vert_pos         -4.750659   0.155441 -30.562  < 2e-16 ***
release_speed     0.059358   0.024369   2.436 0.014864 *
vert_release      0.564688   0.173262   3.259 0.001119 **
horiz_pos         1.309623   0.162735   8.048 8.77e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.6 on 28212 degrees of freedom
Multiple R-squared:  0.05435,    Adjusted R-squared:  0.05411
```

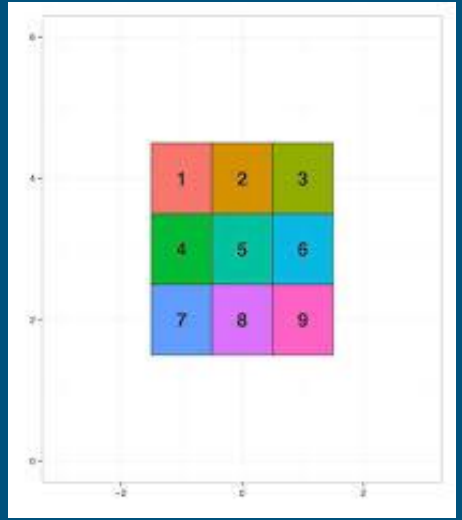
Changeup

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     1.113e+02  6.058e+00  18.376  < 2e-16 ***
release_speed  -2.353e+00  3.988e-01  -5.900 3.74e-09 ***
vert_release   -7.566e-01  3.466e-01  -2.183   0.0290 *
zone           -1.094e+00  4.372e-02 -25.013  < 2e-16 ***
vert_break     -1.086e+00  4.267e-01  -2.545   0.0109 *
horiz_pos       1.386e+00  2.223e-01   6.236 4.65e-10 ***
vert_pos       -9.077e-01  3.036e-01  -2.990   0.0028 **
effective_speed 2.545e+00  3.963e-01   6.423 1.39e-10 ***
release_spin_rate -2.131e-03  4.714e-04  -4.521 6.21e-06 ***
release_extension -3.490e+00  6.774e-01  -5.152 2.62e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.45 on 11546 degrees of freedom
Multiple R-squared:  0.07241,    Adjusted R-squared:  0.07168
```

Linear Regression cont.

Interactive Changeups





Linear Regression - Takeaways

Pros: MSE

- Fastball: Overfit
  - Train: 5.6276
  - Test:: 5.92786
- Curveball: Overfit
  - Train: 2.2641
  - Test: 2.8007
- Changeup: Underfit
  - Train: 5.757
  - Test: 1.135

Cons:

- Adjusted R-squared

- Too many outside factors to be truly accurate
  - Wind
  - Weather
  - Ballpark

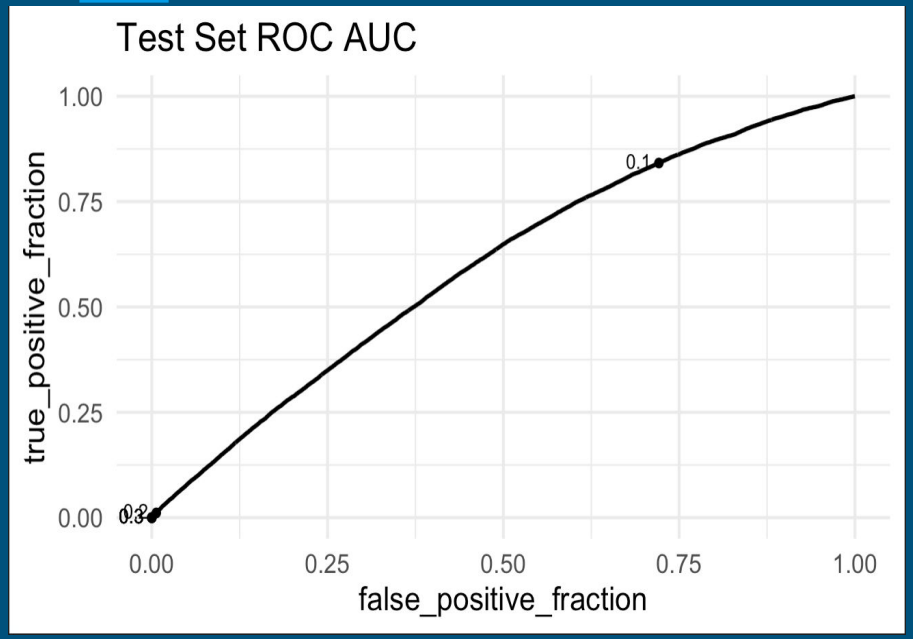
Logistic Regression

- Model fitted using:
  - Pitch_name (factored), release_speed, zone, horiz_break, vert_break, release_spin_rate, and dist_from_home
- 75/25 Train/Test split

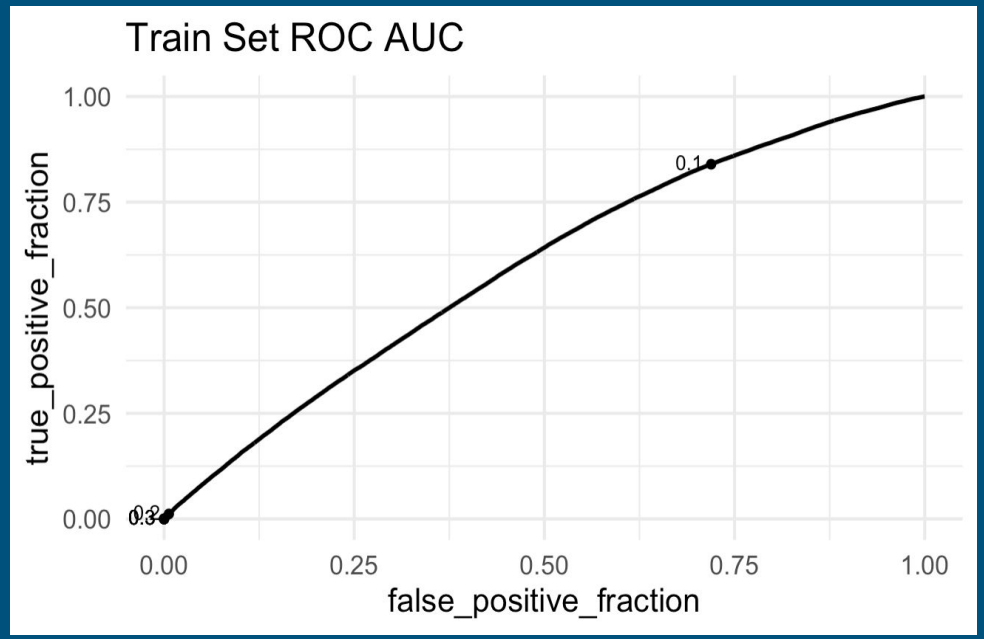
Summary Output

```
Call:
glm(formula = whiff ~ pitch_name + release_speed + zone + horiz_break +
    vert_break + release_spin_rate + dist_from_home, family = binomial,
    data = df_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.7645  -0.5582  -0.4843  -0.4010   2.7212

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)               3.277e+00  5.551e-01   5.903 3.58e-09 ***
pitch_name4-Seam Fastball -7.175e-01  1.692e-02 -42.415  < 2e-16 ***
pitch_nameSinker          -1.168e+00  1.945e-02 -60.045  < 2e-16 ***
pitch_nameChangeup         1.735e-02  1.806e-02   0.960 0.336924
pitch_nameOther           -2.880e-01  1.287e-02 -22.379  < 2e-16 ***
release_speed              3.229e-02  1.279e-03  25.255  < 2e-16 ***
zone                      -6.189e-03  1.040e-03  -5.950 2.68e-09 ***
horiz_break                2.055e-02  5.399e-03   3.807 0.000141 ***
vert_break                -7.917e-02  1.026e-02  -7.715 1.21e-14 ***
release_spin_rate          1.414e-04  1.651e-05   8.565  < 2e-16 ***
dist_from_home            -1.458e-01  9.922e-03 -14.692  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Logistic Regression ROC/AUC



Test Set AUC: 0.595



Train Set AUC: 0.596

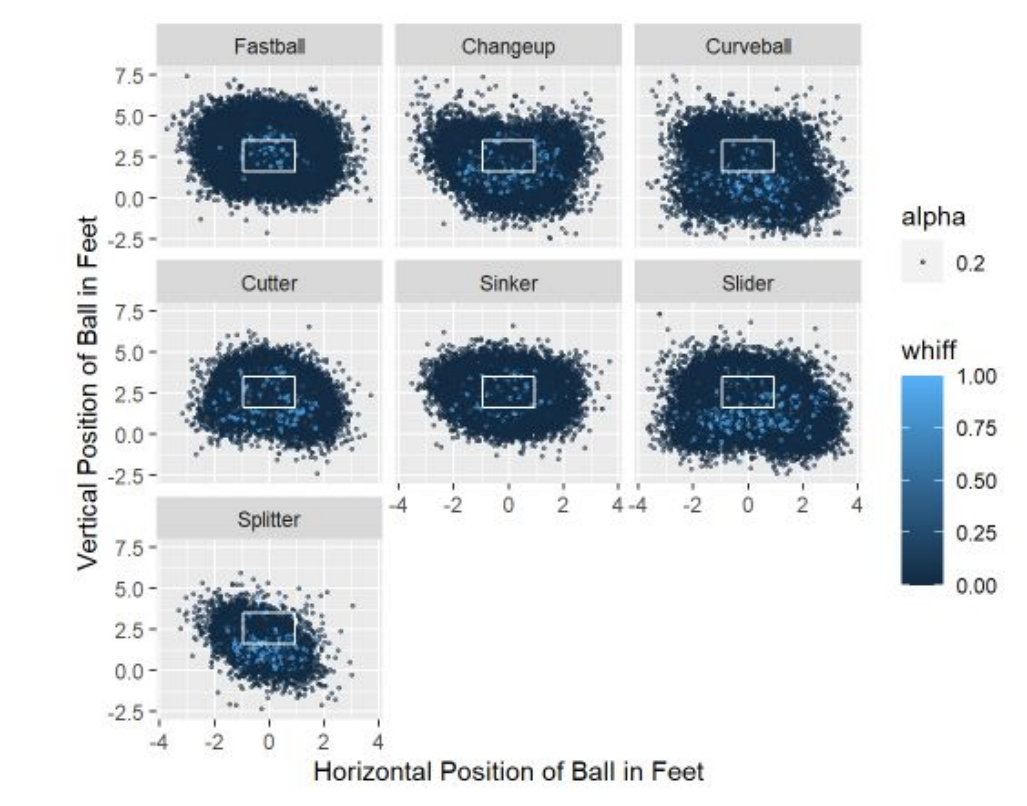
Takeaways/Findings

- Most impactful predictors:
  - Horiz_break, pitch_name (changeup) release_speed, & release_spin_rate
- Less impactful predictors:
  - Dist_from_home, pitch_name (fastball, sinker & other)
- The different factors of pitch_name are highly correlated
- End Result: Omit model look to better predictive power from more complex ML models

Exponentiated Coefficients

(Intercept)	pitch_name4-Seam Fastball	pitch_nameSinker
19.2549507	0.4879312	0.3098657
pitch_nameChangeup	pitch_nameOther	release_speed
1.0167633	0.7562104	1.0329040
zone	horiz_break	vert_break
0.9945678	1.0186257	0.9214672
release_spin_rate	dist_from_home	
1.0001425	0.8691403	

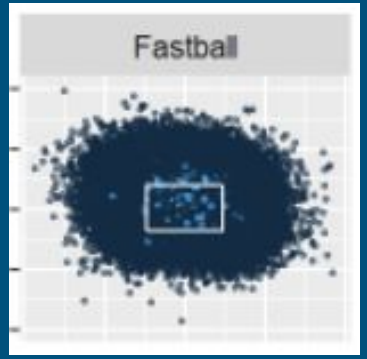
Whiffs based on pitch type and location

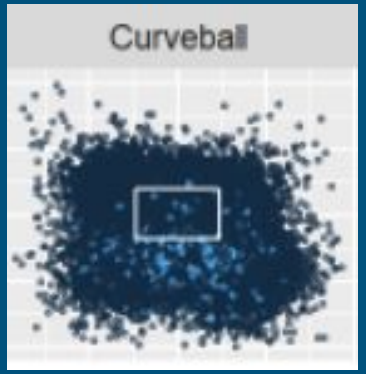


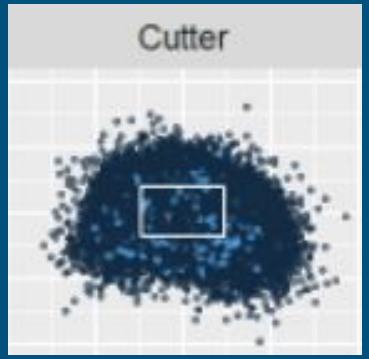
- Fastball-esque pitches have scattered whiff locations
- Breaking and offspeed pitches have a clear whiff location in lower part of zone
- Whiffs are not common inside the strike zone
- Eephus and Knuckleballs were omitted

Analysis

- Useful when you need to avoid contact
- Fastballs rely on velocity to generate whiffs
  - If a whiff is outside of the zone, it is above the zone
- Breaking and offspeed rely on movement to generate whiffs
  - Large amount of whiffs are below the zone
- Cutters and Sinkers are harder to indicate what creates whiffs
- Fastballs up and breaking balls down
- Doesn’t take batter or pitcher handedness or batter tendencies into account







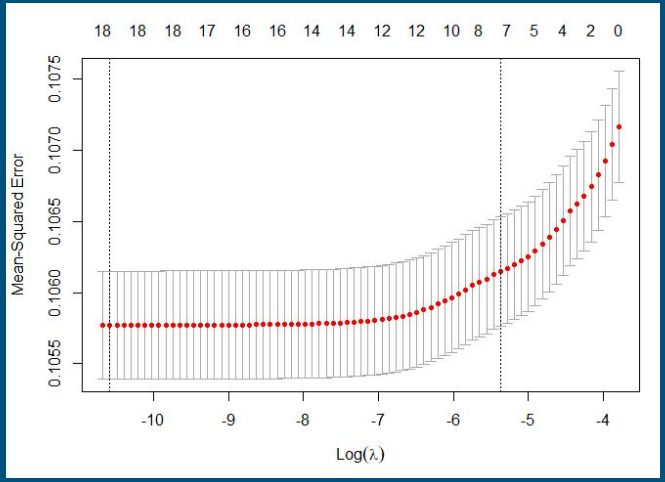
Lasso Regression

Important factors:

- 4 - Seam Fastball, Sinker, Eephus, Split Finger

Least important factors:

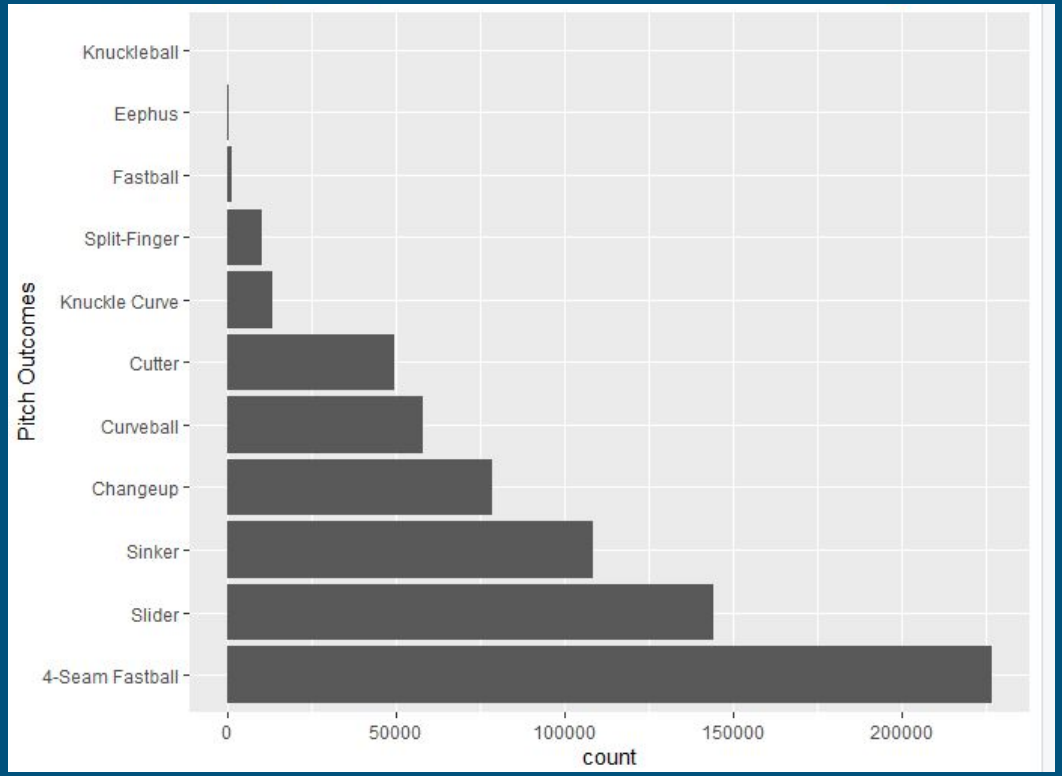
- Release speed, Release Spin Rate, Right handed batters



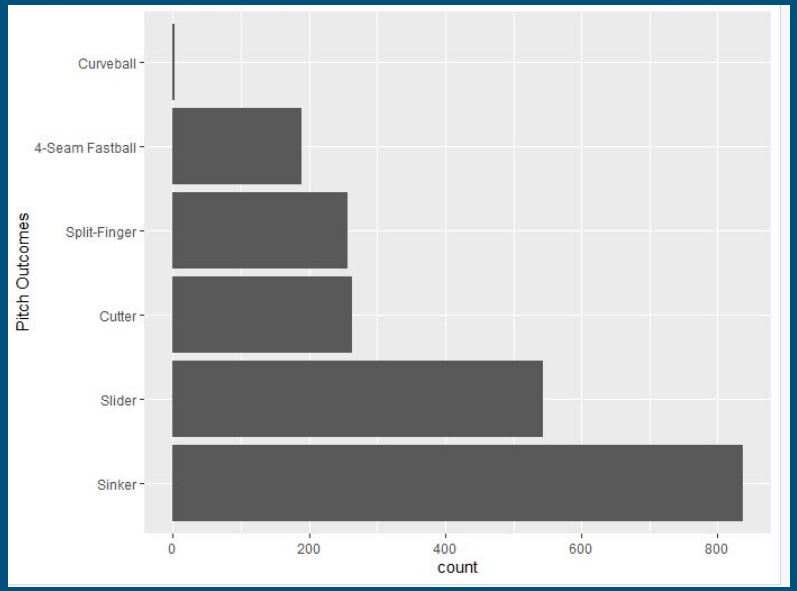
	s1
(Intercept)	-0.267
vert_break	0.001
horiz_break	0.005
zone	-0.001
pitch_name4-Seam Fastball	-0.070
pitch_nameChangeup	0.032
pitch_nameCurveball	0.006
pitch_nameCutter	-0.042
pitch_nameEephus	0.065
pitch_nameFastball	-0.017
pitch_nameKnuckle Curve	0.005
pitch_nameKnuckleball	0.027
pitch_nameSinker	-0.099
pitch_nameSlider	0.022
pitch_nameSplit-Finger	0.076
release_speed	0.004
release_spin_rate	0.000
spin_axis	0.000
batter_sideL	-0.012
batter_sideR	.

Takeaways and External Variables

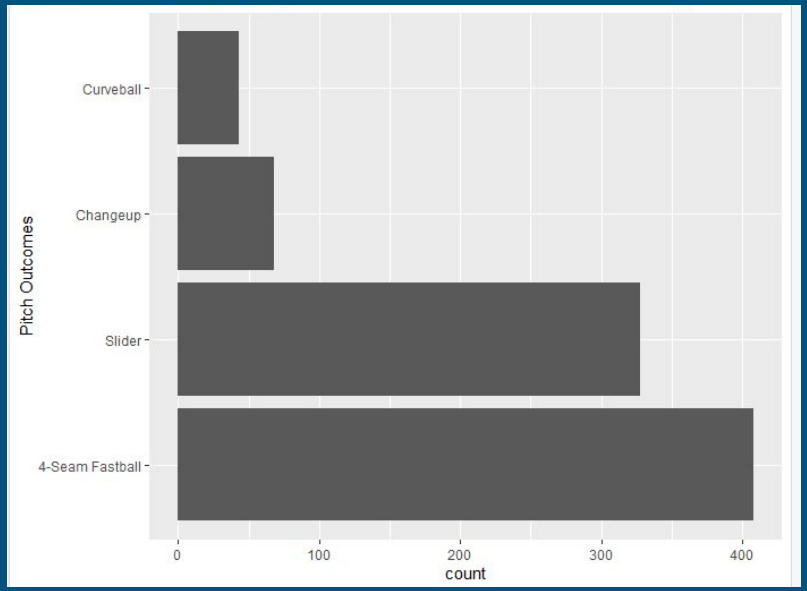
- Small sample size
  - Eephus
  - Split - Finger
- Majority of the batters are right handed
  - “Platoon splits”



Marcus Stroman

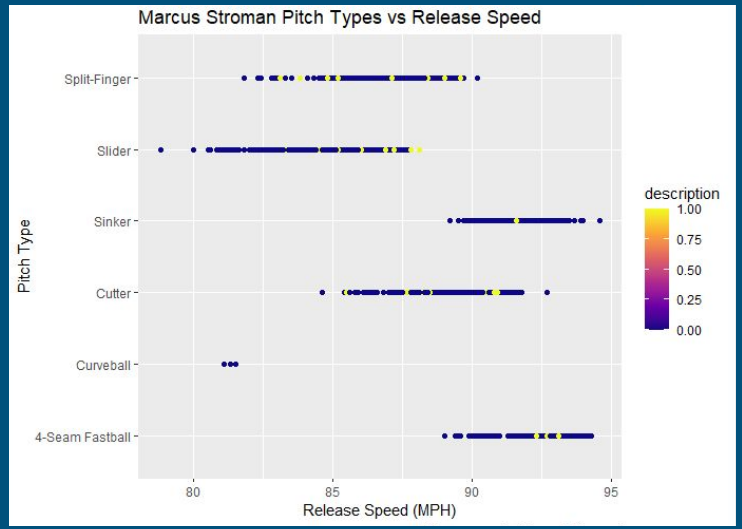


Jacob deGrom

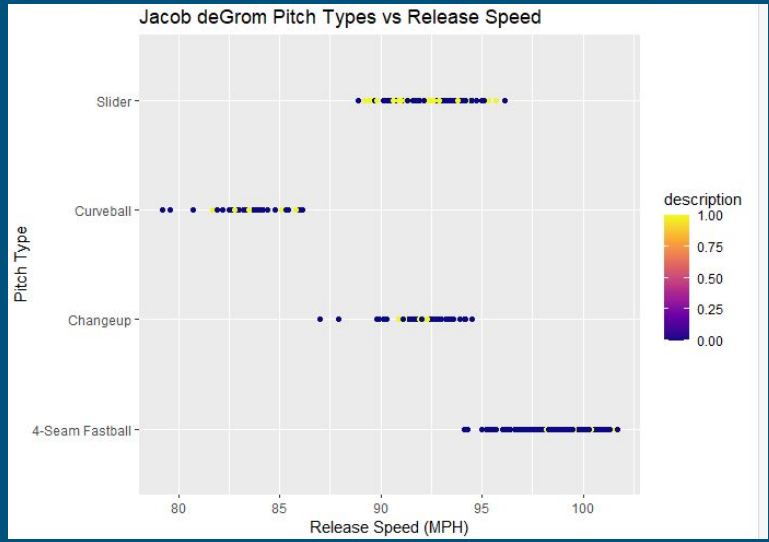


Marcus Stroman and Jacob deGrom

	s1
(Intercept)	0.089
vert_break	.
horiz_break	.
zone	0.004
pitch_name4-Seam Fastball	.
pitch_nameCurveball	.
pitch_nameCutter	0.040
pitch_nameSinker	-0.023
pitch_nameSlider	.
pitch_nameSplit-Finger	.
release_speed	.
release_spin_rate	.
spin_axis	0.000
batter_sideL	.
batter_sideR	.



	s1
(Intercept)	-1.893
vert_break	-0.037
horiz_break	0.014
zone	-0.009
pitch_name4-Seam Fastball	-0.249
pitch_nameChangeup	0.157
pitch_nameCurveball	-0.102
pitch_nameSlider	.
release_speed	0.021
release_spin_rate	0.000
spin_axis	-0.001
batter_sideL	-0.011
batter_sideR	0.000



Comparison of performance of each of the models

Linear Regression:

- Was overfit for fastballs and curveballs; underfit for changeups
- Had a less than ideal R squared value
- Had too many outside factors to be a viable option

Logistic Regression:

- Not overfit or underfit
- Predictive power barely better than coinflip
- Omit and look for more complex ML model

Comparison of performance of each of the models

Whiffs based on pitch and location:

- Useful as a general tool for pitchers and catchers
- Doesn’t take into account advanced insights
- Unbiased

Lasso Regression:

- Fastballs and sinkers were most important factors in generating whiffs
- However it differs based on the player
- Mixed results as 2 of the most important pitches are rarely used

Conclusion regarding whether or not the model should be implemented

- Model should not be fully implemented
- Too many models returned inconclusive results
- There were some valuable takeaways (pitch type and whiff correlation)

Works Cited

- https://baseballsavant.mlb.com/
- http://www.sthda.com/english/wiki/ggplot2-title-main-axis-and-legend-titles
- https://billpetti.github.io/

GitHub Repo