Does it Home Run?

Looking into Home Runs from the 2025 Season
By: Garrett Johnson

rMarkdown File on Github

Research Questions

1. Which statcast variables are the best predictors of a homerun, and how much do they contribute to the model?
2. Does Pitch type change the likelihood of a batted ball resulting in a homerun?
3. How much does pitch location mean for the probability of a home run?

Data Processing Workflow

- Retrieved Statcast data via the *Baseballr* API.
- Filtered out bunts, pitchouts, and batted balls with a launch speed of NA (non-swings; used only for the regression model).
- Created a binary **is_hr** variable to model home-run outcomes using logistic regression
- Engineered an Interaction term, **launchsp_launchang**, to capture combined launch-speed and launch-angle effects.
- Created a column for **pitch_type** as a factor so it could be plugged into the regression model

Regression Variables

Model Variables

- **Launch Angle** — The vertical angle at which the ball leaves the bat.
- **Launch Speed** — The initial velocity of the ball off the bat.
- **Launch Speed × Launch Angle** — Interaction term capturing how the two variables jointly influence contact quality.
- **Pitch Type** — The type of pitch delivered.
- **Effective Speed** — The perceived velocity of the pitch, accounting for location and pitch sequencing.
- **Bat Speed** — The velocity of the bat during the swing.
- **Swing Path Tilt** — The vertical shape of the hitter's swing as it moves through the zone.

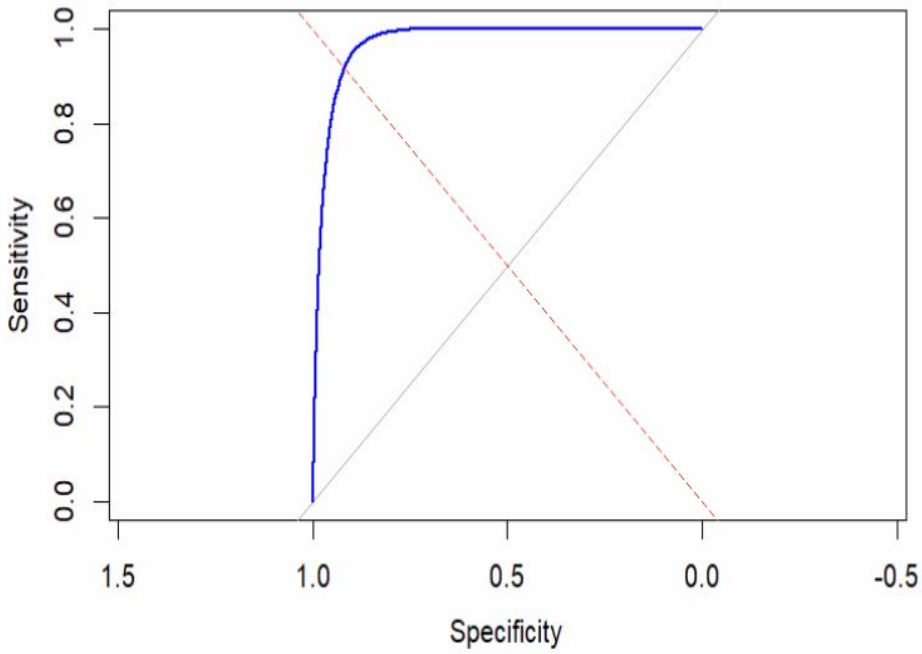
Logistic Regression Model Summary and ROC Curve

```
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)       -2.050e+01  5.640e-01 -36.353  < 2e-16 ***
launch_speed       2.128e-01  3.196e-03  66.585  < 2e-16 ***
launch_angle      -3.339e-01  5.545e-03 -60.222  < 2e-16 ***
launchsp_launchang 4.090e-03  6.061e-05  67.486  < 2e-16 ***
pitch_typefCS     -8.564e-01  1.119e+00  -0.765  0.44428
pitch_typefCU     -4.339e-02  9.487e-02  -0.457  0.64742
pitch_typefEP     -1.261e-01  3.539e-01  -0.356  0.72154
pitch_typefFA     -3.787e-01  3.736e-01  -1.014  0.31072
pitch_typefFC     -2.742e-01  8.422e-02  -3.256  0.00113 **
pitch_typefFF     -1.698e-01  7.823e-02  -2.170  0.03002 *
pitch_typefFO     -1.195e+00  1.080e+00  -1.106  0.26867
pitch_typefFS     -5.773e-02  1.166e-01  -0.495  0.62036
pitch_typefKC      2.311e-01  1.517e-01   1.523  0.12764
pitch_typefKN     -9.410e+00  1.149e+02  -0.082  0.93471
pitch_typefSC     -6.581e+00  8.827e+02  -0.007  0.99405
pitch_typefSI     -2.261e-01  8.342e-02  -2.711  0.00671 **
pitch_typefSL     -4.368e-02  7.026e-02  -0.622  0.53411
pitch_typefST     -6.351e-02  8.819e-02  -0.720  0.47143
pitch_typefSV     -1.546e-01  2.719e-01  -0.569  0.56957
pitch_typefUN     -3.817e+00  8.827e+02  -0.004  0.99655
effective_speed   -2.417e-02  5.376e-03  -4.496 6.93e-06 ***
bat_speed         -4.015e-02  4.455e-03  -9.014  < 2e-16 ***
swing_path_tilt   -1.085e-02  2.652e-03  -4.092 4.27e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 47656  on 216848  degrees of freedom
Residual deviance: 25018  on 216826  degrees of freedom
AIC: 25064

Number of Fisher Scoring iterations: 13
```



Interpreting the Model Results

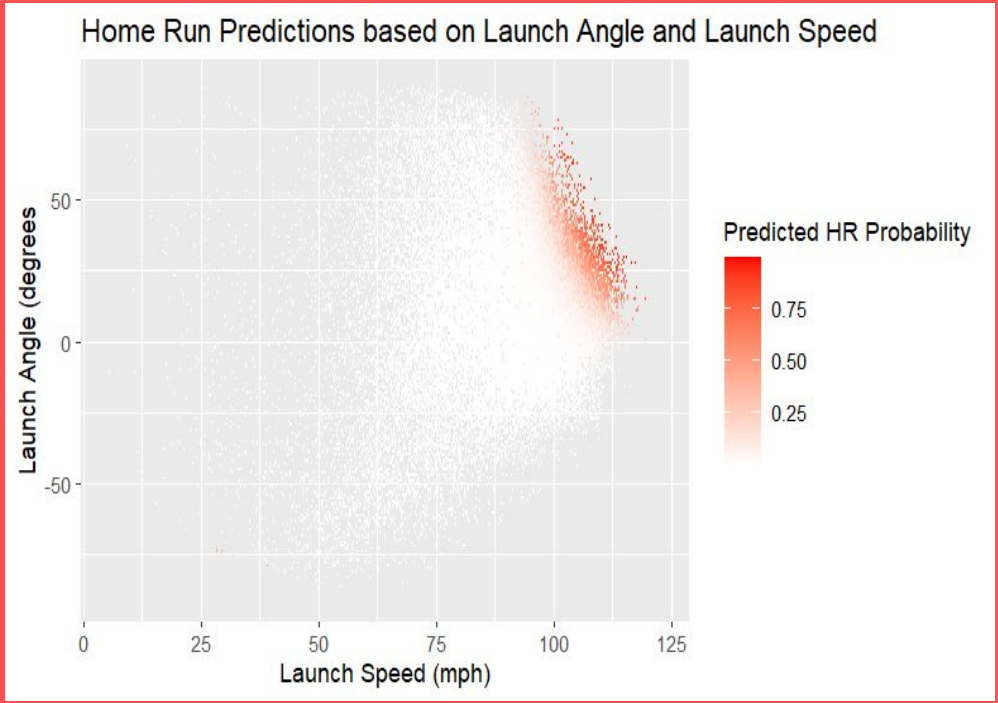
Key Takeaways

- **Launch speed, launch angle, and their interaction** are by far the strongest predictors of home-run probability with the smallest p-values ($p > .001$).
- **Effective speed, bat speed, and swing path tilt** also significantly affect HR outcomes.
- Only **three pitch types** (cutter, four-seamer, sinker) significantly differ from the baseline.
- Most pitch types contribute little once contact quality is known.

ROC Takeaways

- The Roc curve is a way to evaluate my model's performance.
- From the model we see high specificity and sensitivity meaning it has excellent predictive accuracy that isn't left up to chance.
- The AUC at 95% means the model can correctly put home runs over non-home runs 95% of the time which is statistically very high.

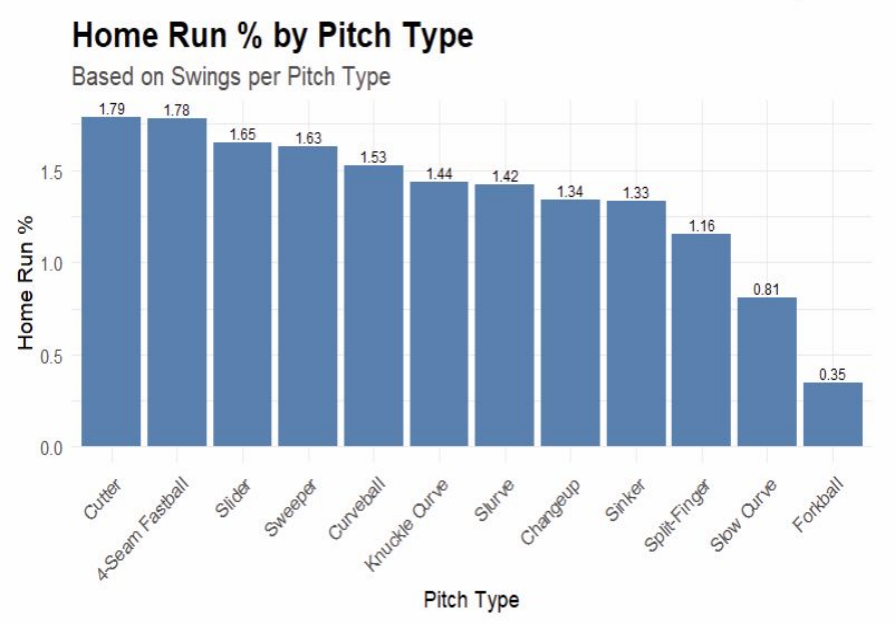
Home Run Prediction Graphic

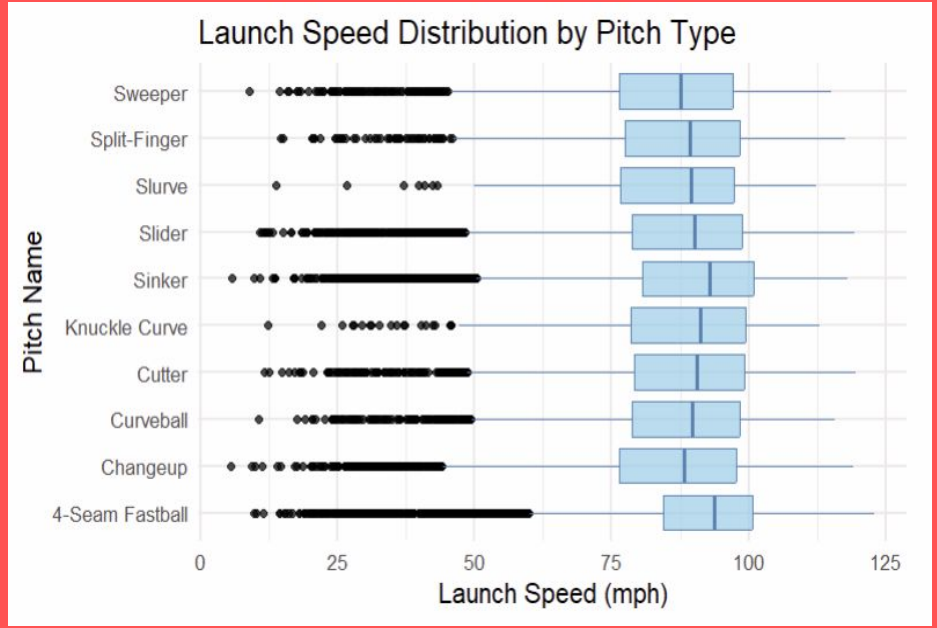


Home runs cluster at launch angles of 20–45° and exit velocities of 100–115 mph, with harder-hit balls always increasing the predicted probability of a home run.

The model clearly shows that optimal angles combined with high launch speeds produce nearly all predicted home run outcomes, highlighting how crucial quality of contact is.

Pitch Type Summary





Interpreting the Pitch Type Models

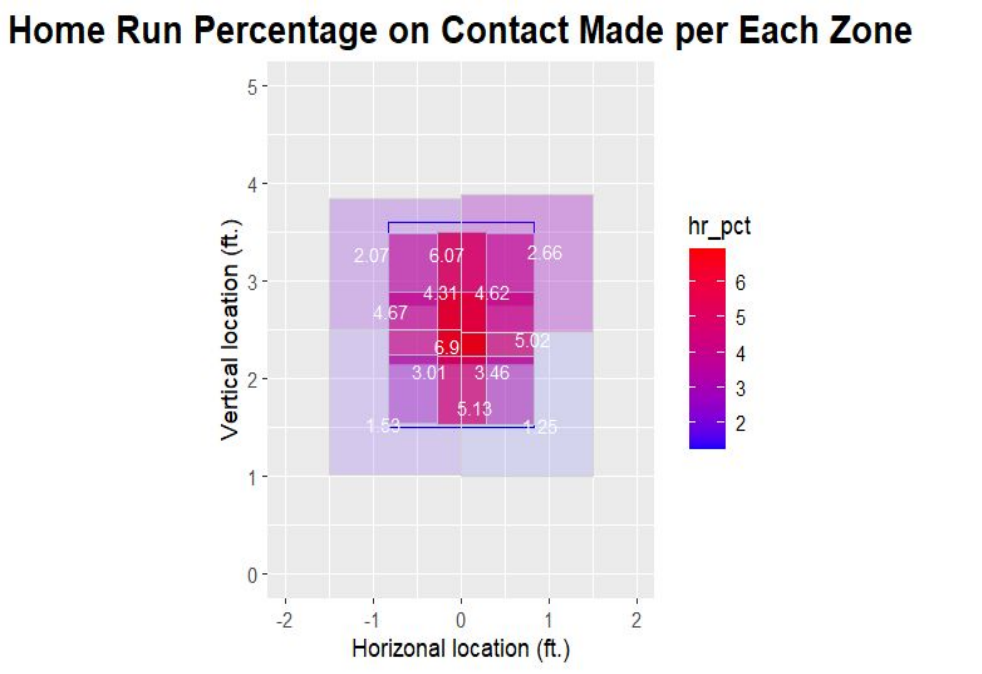
Bar Graph Takeaways

- Home Run Percentages range from .35% up to 1.79%, showing some variation but not really any drastic differences.
- Most pitches types cluster between roughly 1.1% and 1.8% home run rates, indicating relatively similar home run risk across many pitches.
- 4-Seam Fastballs and Cutters sit at the higher end, but still the gap compared to breaking and off-speed stuff isn't extreme

Box Plot Takeaways

- 4-Seam Fastballs produce the highest median launch speeds, indicating they are generally hit the hardest.
- Most other pitches have similar median launch speeds around 90-100mph, showing moderate variation in quality of contact.

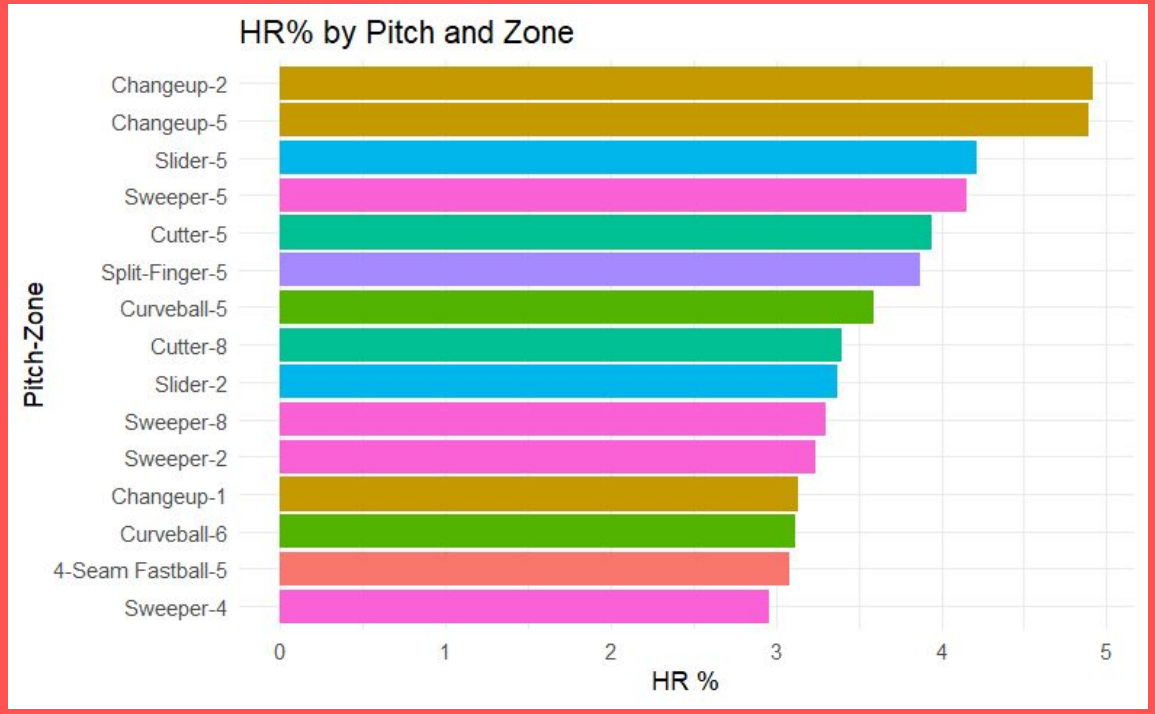
Zone Data Summary



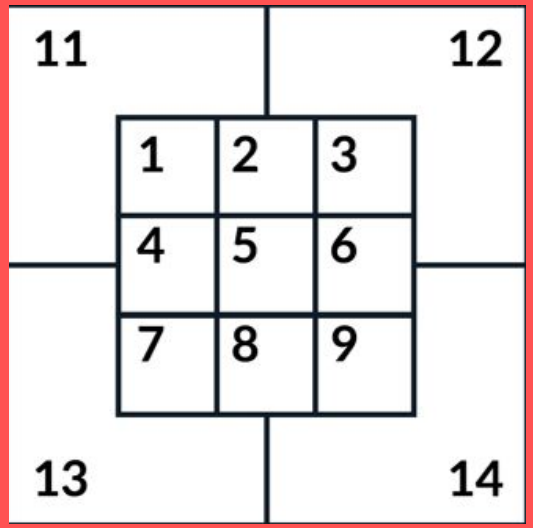
Takeaways

- Easy to see that most homeruns come from pitches when contacted near the center of the plate, which doesn't come too unexpectedly and the lower/upper middle zone comes as the second sweet spot.
- The four corners are all relatively similar with the two upper zones with slightly higher percentages.
- The bottom outside of the zone offers up the smallest home run percentages on swings.

Further Discovery on Zone and Pitch Type



For Reference of the Zones



Further Interpretation

- A combining of the two metrics shows us which pitch and placement allow for the highest home run percentage.
- From the visual it's clear that throwing changeups high in the zone or near the center is not an ideal pitch.
- In fact, breaking pitches that are hung in the middle of the zone seem to be what hitter's see best.

For more homerun data on each hitter for the 2025 Season, click the link here for further analyzation: https://7zjoqq-garrett-johnson.shinyapps.io/RegressionProject/

Summary

To Preface: Home Runs are hard to come by, but maybe we can help predict them.

1. Home runs are best predicted by launch speed, launch angle, their interaction term, effective speed, bat speed, and swing path tilt. Effective speed has a negative effect as in the faster the pitch looks, the harder it is for the batter to hit it. This further emphasizes why pitch tunneling is so important.
2. Pitch type had minimal influence on the home run probability. Both the regression results and exploratory plots showed only small differences in home run rates across pitch types.
3. Zone plots reveal that location matters more than pitch type. Pitches left middle-middle have much higher home run rates than those on the edges, regardless of the pitch thrown.
4. The most commonly hit home run pitches by percentage are hanging offspeed pitches in the middle or upper part of the zone. While less frequently thrown there, these represent the most punishable mistakes.