BUILDING BETTER HITTERS:

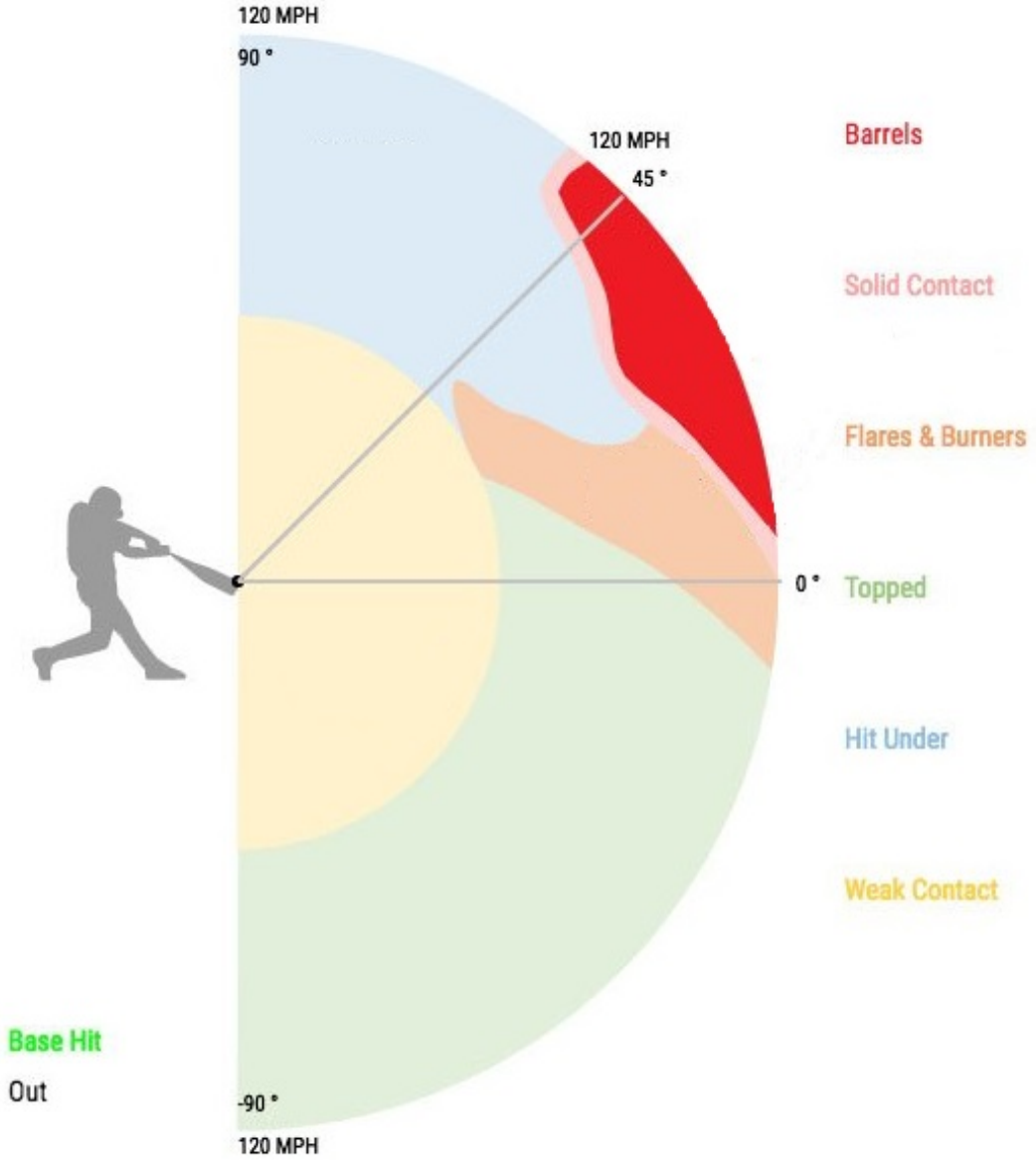
PREDICTING BASEBALL PLAYER OFFENSIVE SUCCESS

AGENDA

- Introduction to Project
- Methodology
- Findings
- Player Applications
- Future Improvements

COVID-19 HAS CHANGED THE FINANCIAL LANDSCAPE OF BASEBALL

- Large financial losses incurred during 2020 season
- Majority of teams don't have budget to sign free agent stars
- **Result: new focus on internal player development to maximize in-house player performance**



FEATURES: FOCUS ON WHAT THE PLAYER CAN CONTROL

- Focusing on physical hitting traits, not the results – Statcast data
- Features Answer Three Main Questions:
  - How hard is the ball hit?
  - Where is it hit?
  - How much plate discipline does the batter have?
- Data Pool: 2015-2020, players with 200 plate appearances or more

WRC+ REMOVES BIAS FROM EXTERNAL FACTORS FOR EVALUATING SUCCESS

- The target: Weighted Runs Created Plus (wRC+)
- wRC+ is a weighted metric:
  - League Neutral
  - Stadium Neutral
- League Average is set at 100

wRC+ = (((wRAA/PA + League R/PA) + (League R/PA – Park Factor* League R/PA))/ (AL or NL wRC/PA excluding pitchers))*100

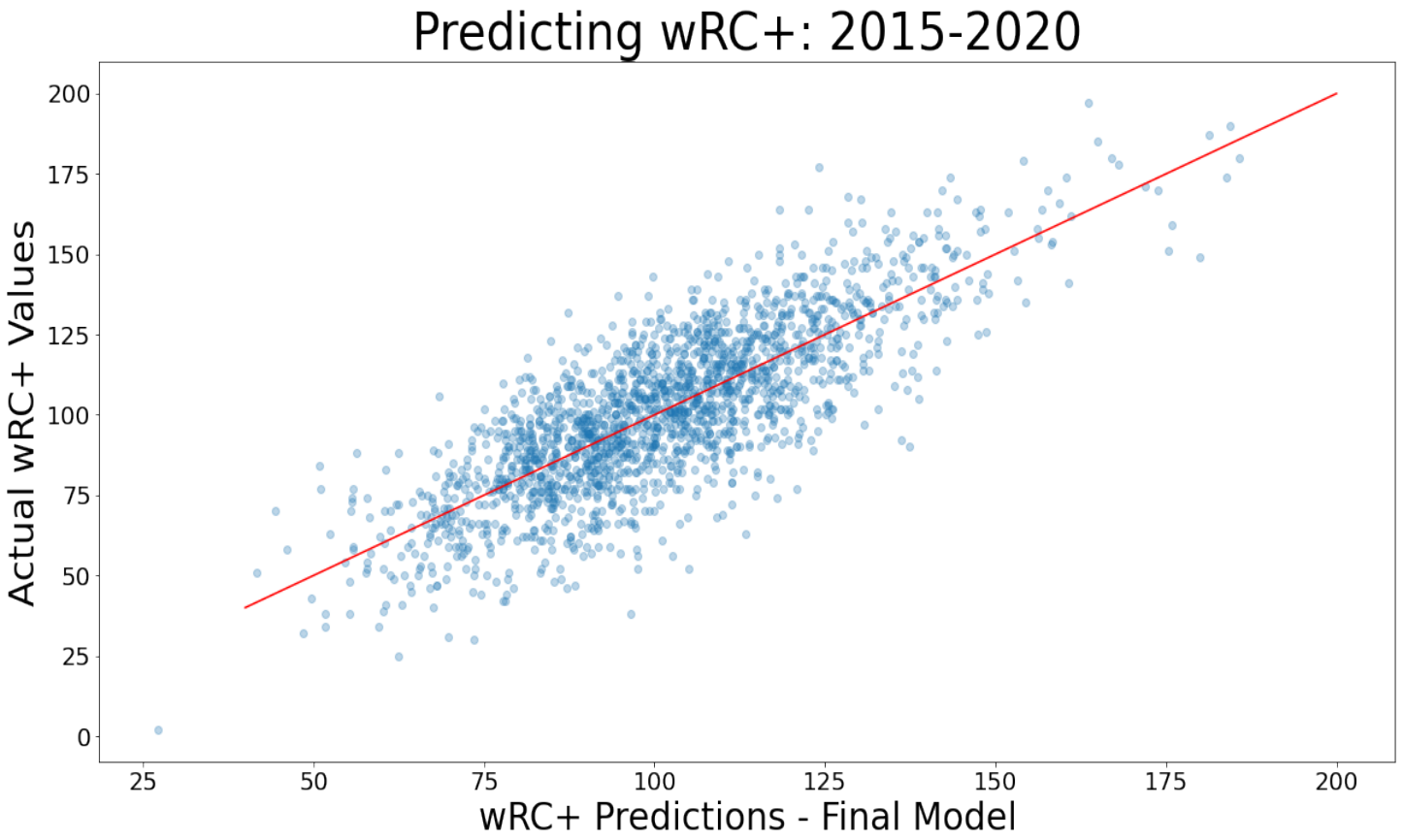
METHODOLOGY AND TOOLS

Data scraped using Selenium and BeautifulSoup:

- FanGraphs: wRC+
- Baseball Savant: Statcast

Tools used to analyze the data and build the model include:

- Python
- Pandas
- Scikit-learn



THE FINAL MODEL ACCOUNTED FOR 65% OF THE VARIANCE IN WRC+

- Test R^2: 0.653
- Test Mean Absolute Error: 12.56 wRC+ points
- Biggest Contributors to wRC+:
  - Positive:
    - Walk Rate
    - Flare/Burner Rate
    - Sprint Speed
    - In Zone Swing Rate
  - Negative:
    - Whiff Rate
    - Under Rate
    - Topped Rate
    - In Zone Contact Rate
- Final model included strong interaction terms

EXTREME VALUES OF BATTING AVERAGE ON BALLS IN PLAY TENDED TO LEAD TO UNDER OR OVER PREDICTIONS

Top 5 Under Predicted Players

Player - Year	Residual	BABIP
DJ LaMahieu - 2020	52.87	.370
Nelson Cruz - 2020	45.77	.360
Enrique Hernandez - 2015	44.67	.359
Chris Colabello - 2015	43.28	.411
Andres Blanco - 2015	42.32	.335

Top 5 Over Predicted Players

Player - Year	Residual	BABIP
Victor Reyes - 2018	-58.46	.277
Juan Uribe - 2016	-53.09	.227
Kendrys Morales - 2019	-47.42	.217
Kyle Schwarber - 2020	-46.66	.219
Eduardo Escobar - 2020	-45.61	.244

SO WHAT?
PLAYER APPLICATIONS

Metric	2020 Stat	Proposed Change
Walk Rate	6.5	8.0 (+1.5)
Avg. Launch Angle	14.3	13.0 (-1.3)
Flare/Burner Rate	22.4	23.4 (+1.0)
Poor/Weak Rate	4.5	3.5 (-1.0)
Out of Zone Swing Rate	30.4	25.4 (-5.0)
Whiff Rate	42.7	32.0 (-10.7)
Pull Rate	35.8	36.8 (+1.0)
Opposite Field Rate	23.9	22.9 (-1.0)
Projected 2021 wRC+ w/ Improvements:		116.45

CASE 1: IMPROVING THE BAT OF A YOUNG, STRUGGLING PLAYER

- **Keston Hiura – 2B, Milwaukee Brewers**
  - 2020 wRC+: 87 (predicted 85)
- **RECOMMENDATION:**
  - **Work with hitting coach to reduce weak contact rate**
  - **Work on pitch recognition to increase walks and reduce swings at bad pitches and swings and misses**
  - **Work on hitting pitches in front of the plate to pull**

Metric	2020 Stat	Proposed Change
Avg. Launch Angle	16.7	14 (-2.67)
Flare/Burner Rate	16.0	17.8 (+1.8)
Poor/Weak Rate	5.3	3.5 (-1.8)
Out of Zone Swing Rate	39.6	32.0 (-7.6)
Whiff Rate	41.6	30.0 (-11.6)
Pull Rate	32.8	36.8 (+4.0)
Opposite Field Rate	29	25.0 (-4.0)
First Strike Rate	67.8	62.8 (-5.0)
Projected 2021 wRC+ w/ Improvements:		132.80

CASE 2: HELPING A RISING STAR TAKE THE NEXT STEP

- **Luis Robert – OF, Chicago White Sox**
  - 2020 wRC+: 101 (predicted 102)
  - 2nd in 2020 American League Rookie of the Year Voting
- **RECOMMENDATION:**
  - **Work with hitting coach to reduce weak contact rate**
  - **Work on pitch recognition to reduce reduce swings at bad pitches, swings and misses, and taking first pitch strikes**
  - **Work on hitting pitches in front of the plate to pull**

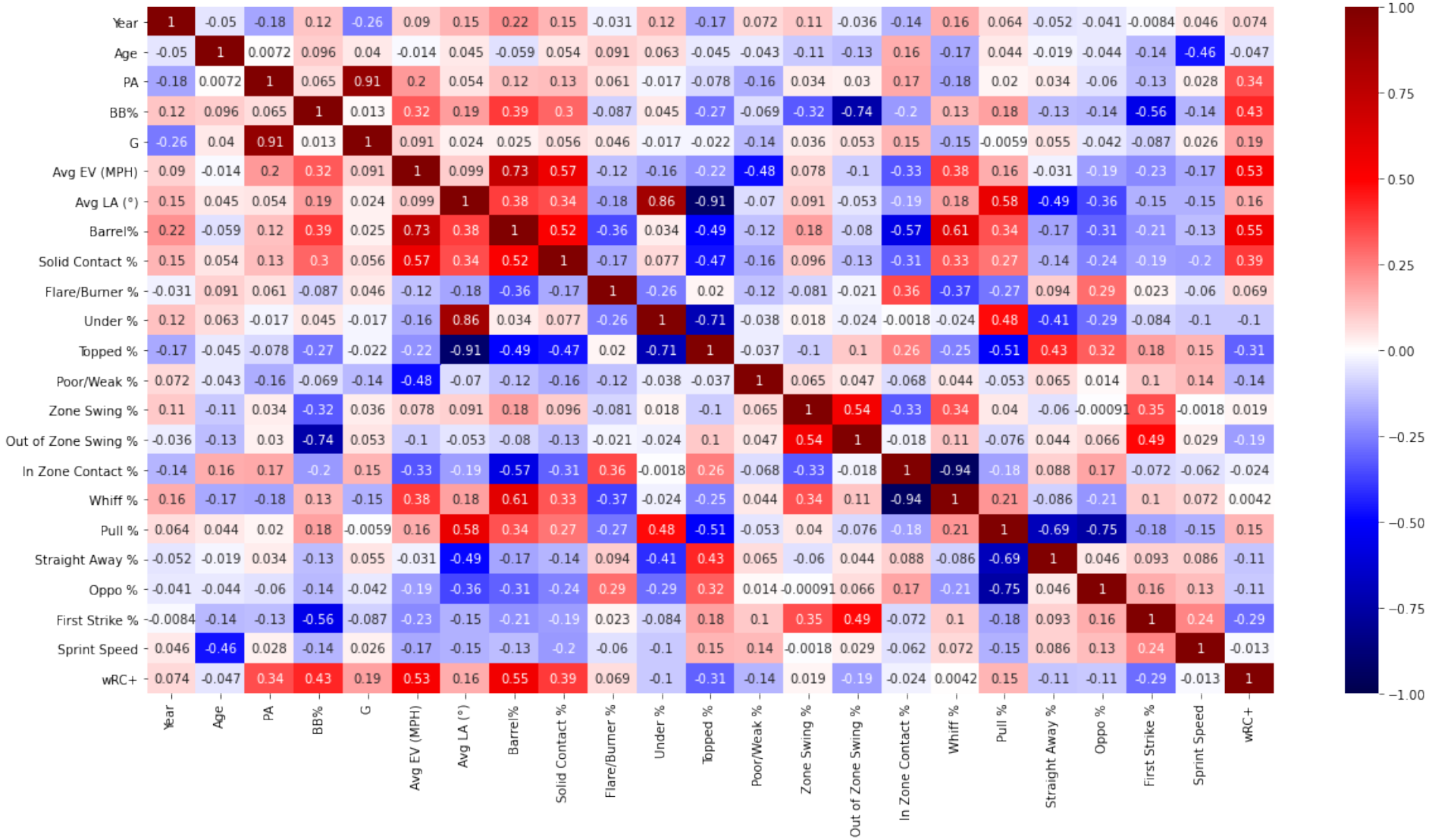
FUTURE IMPROVEMENTS

- Incorporating Batting Average on Balls in Play modeling
  - Defensive shifts
  - Spray charts
- Time Series Analysis for changes over time

THANK YOU!
ANY QUESTIONS?

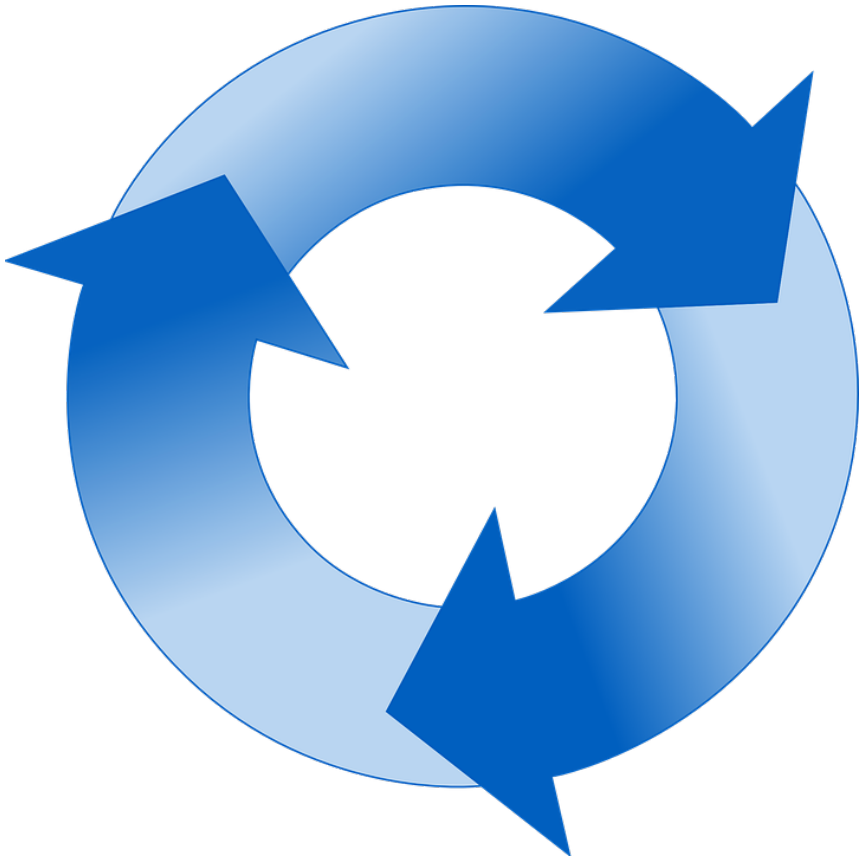
APPENDIX

FINAL FEATURES: SEABORN HEAT MAP



FINAL FEATURES: PAIR PLOT

SWING CHANGE EXAMPLE – JUSTIN TURNER



ITERATING THROUGH FEATURES INCREASED MODELING PREDICTABILITY VIA R^2 SCORES

- 0.275 — Only using Exit Velocity
- 0.438 — Original Full Features w/Position
- 0.561 — Added Additional Batted Ball Profiles, Removed Position
- 0.631 — Added Walk Rate, Whiff Rate
- 0.641 — Added strong interaction terms

FINAL FEATURES: R^2 RESULTS ON TRAIN/VAL

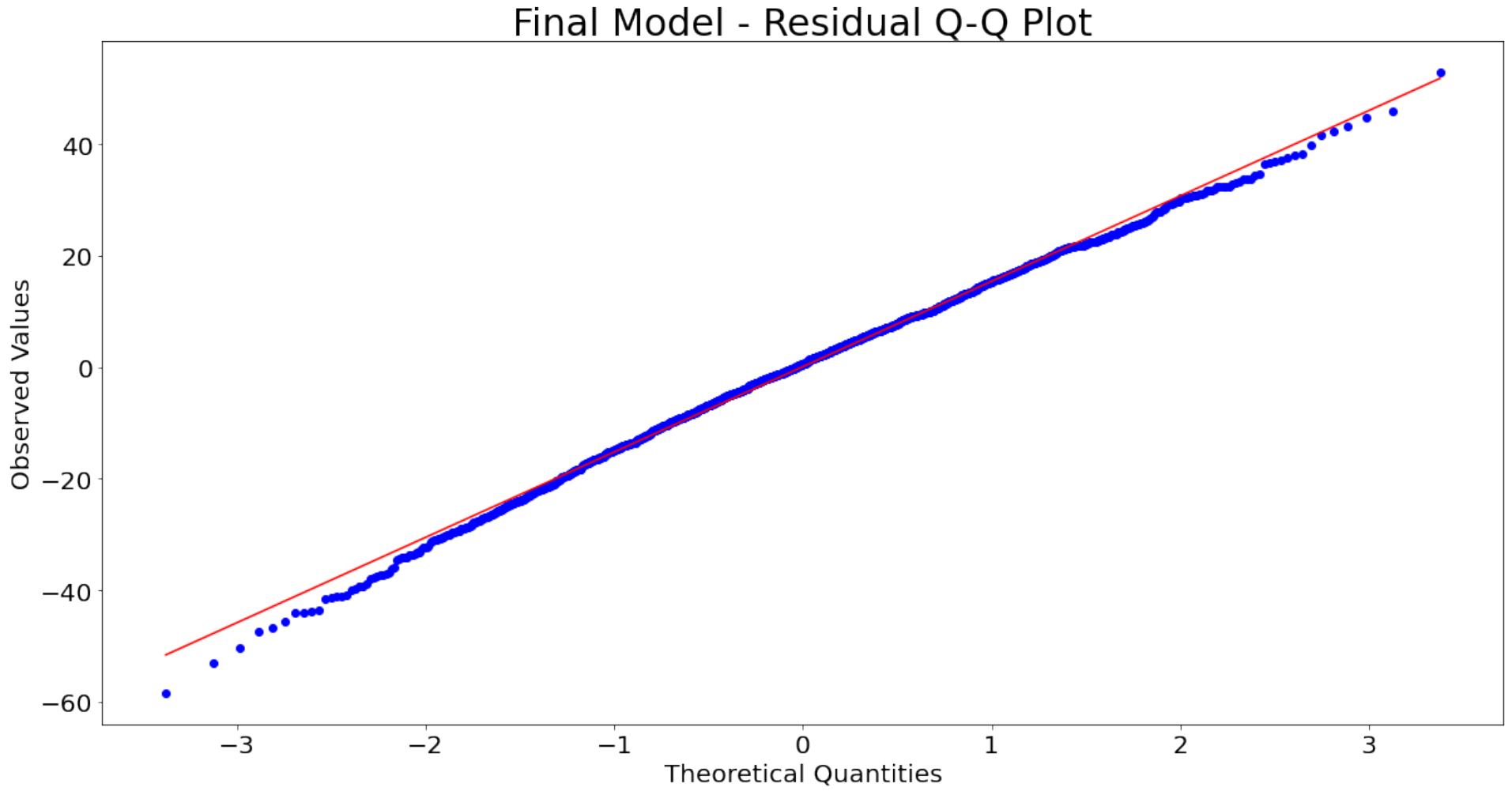
Regression Type	Full Features w/o BB% and Whiff%	Full Features w/ BB% and Whiff%	Full Features w/ BB% and Whiff%, - Oppo %	Barrel %, Topped %, Sprint Speed Only	*Full Features w/ BB%, Whiff%, & Strong Interaction Terms*
Simple Linear (no CV)	0.543	0.578	0.578	0.325	0.589
Simple Linear (w/Kfold CV)	0.543	0.617	0.617	0.297	0.626
Lasso w/ CV	0.561	0.631244	0.631244	0.304	0.6402
Ridge w/ CV	0.561	0.631267	0.631267	0.304	0.6412*
ElasticNet w/ CV	0.560	0.630	0.630	0.304	0.6406
Lasso w/ Polynomial Features & CV	n/a	0.644	n/a	n/a	n/a

**final model. Final Model test R^2 was 0.653*

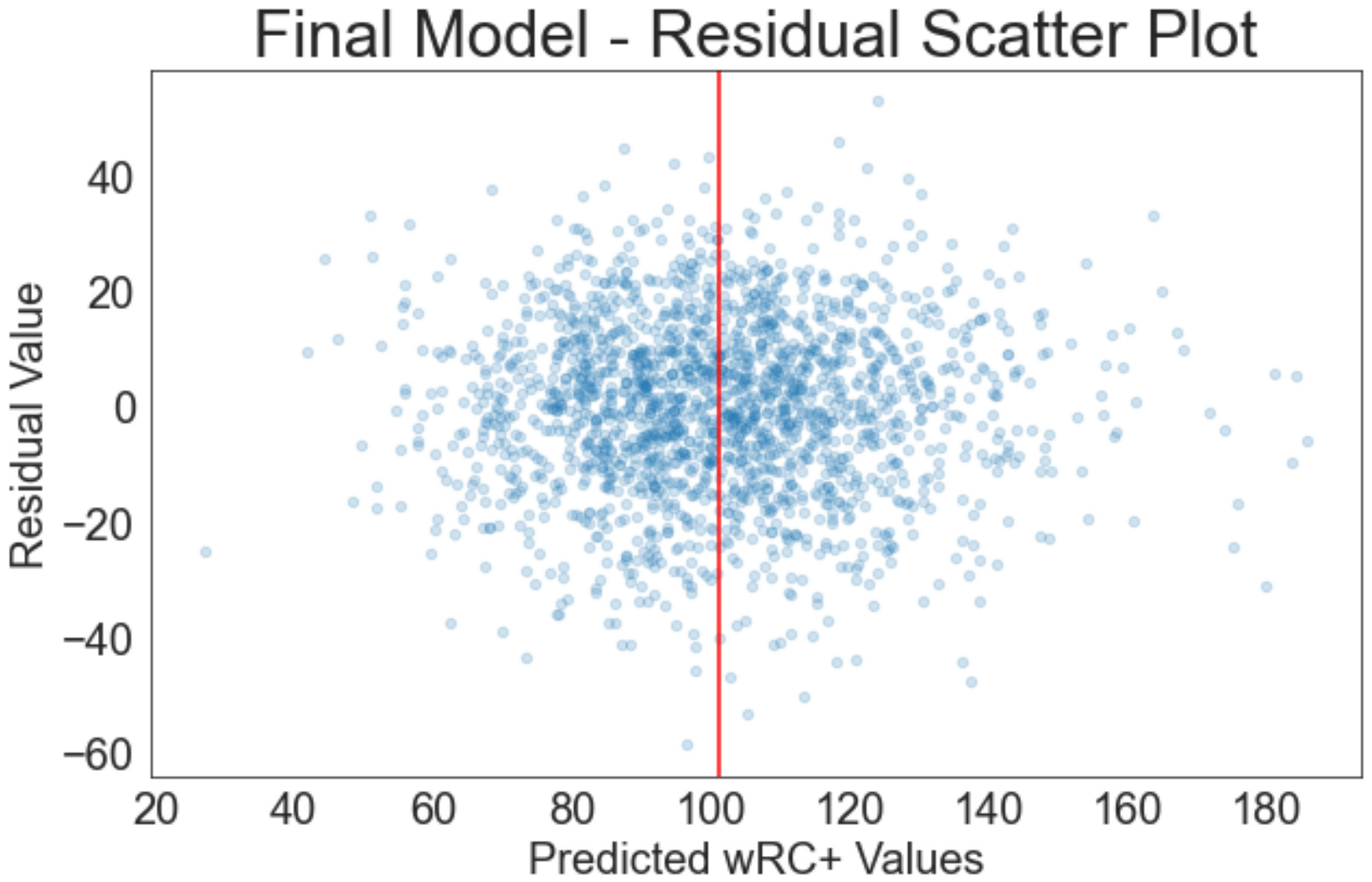
BABIP: LINEAR REGRESSION R^2 RESULTS

Regression Type	BABIP (Only Feature)
Simple Linear (no CV)	0.339
Simple Linear (w/Kfold CV)	0.257
Lasso w/ CV	0.260
Ridge w/ CV	0.260
ElasticNet w/ CV	0.259

FINAL MODEL RESIDUAL Q-Q PLOT

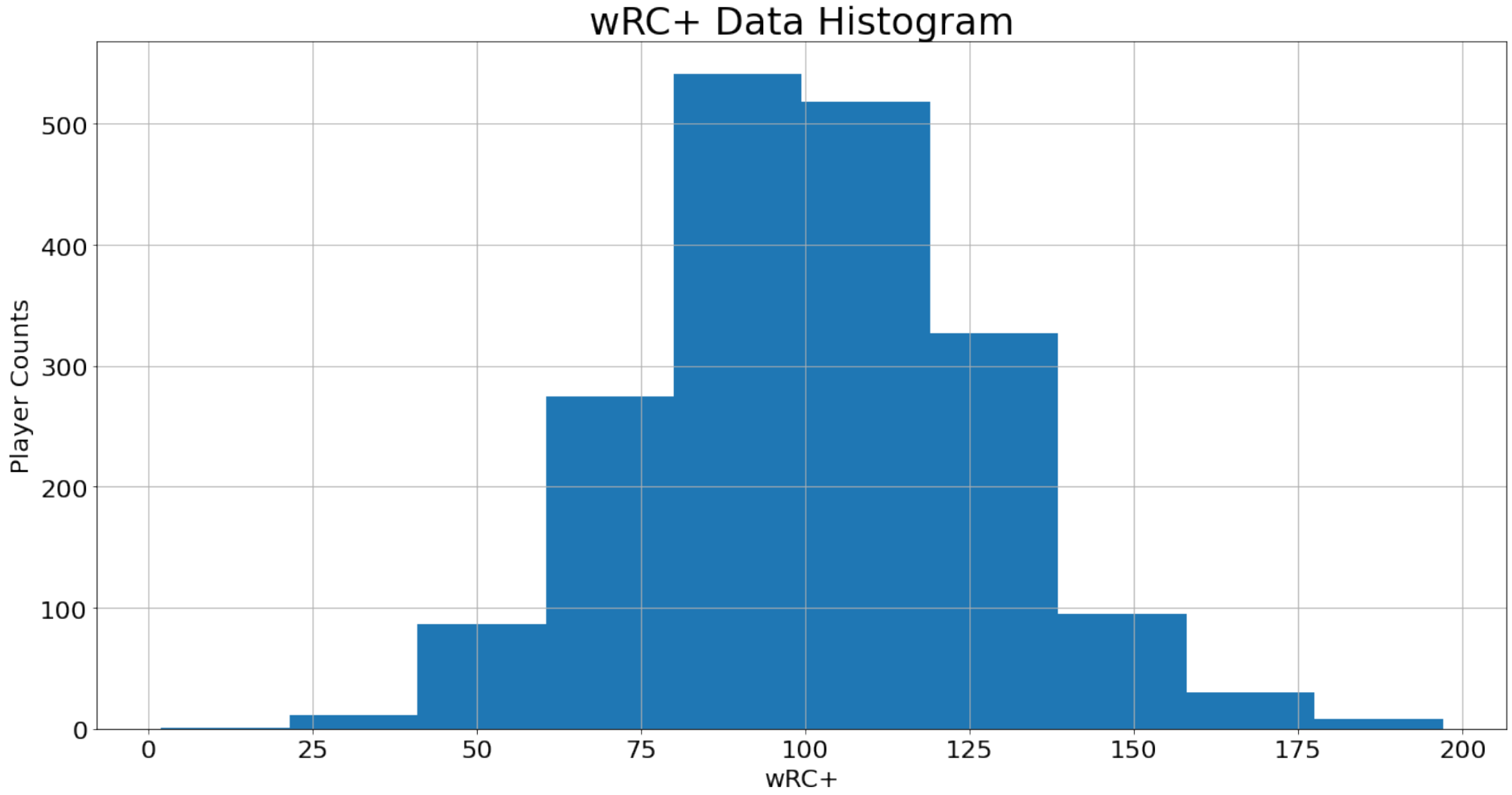


FINAL MODEL RESIDUAL PLOT



- Mean Value of predicted wRC+ in Red Line (101.20)
- No obvious trend or pattern, seems to be centered around mean

WRC+ FEATURE HISTOGRAM



TEST DATA: REGRESSION MODEL