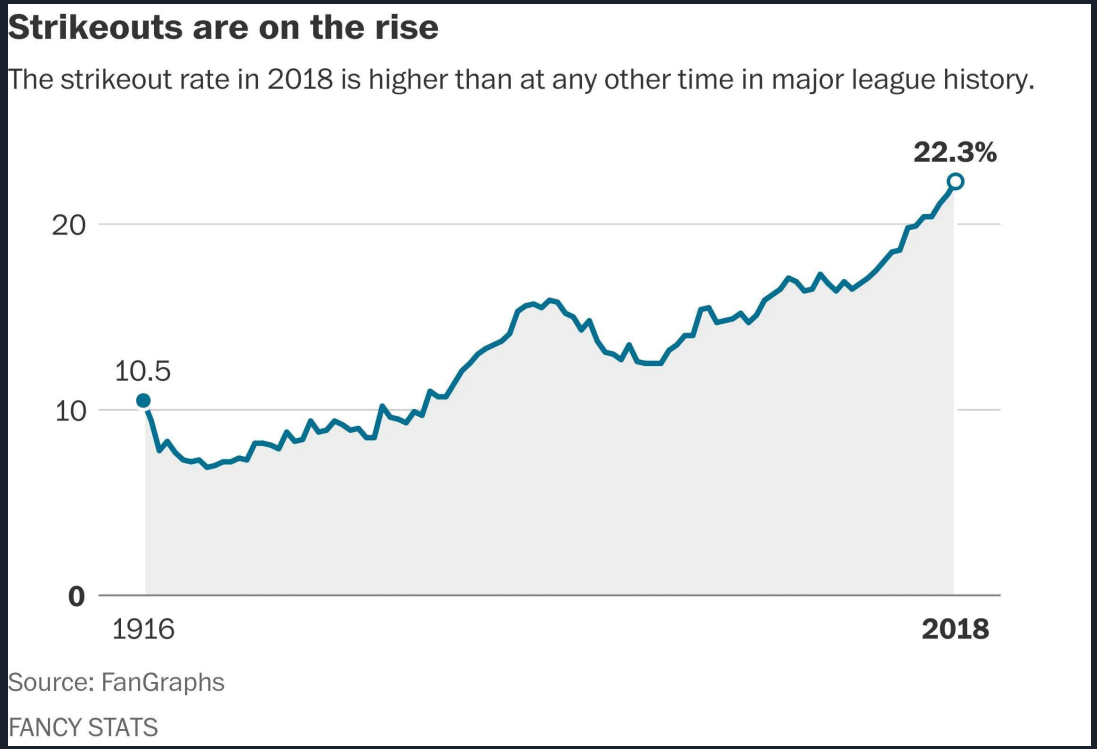
How To Strikeout the Competition in the SP Free Agency Market

Michael Scognamiglio
Joe Ramirez
nyc-mnhtn-ds-080320

What Our Project Is About

- This project analyzes strikeout numbers for a starting pitcher in a season.
- The goal of this project is to gain insights into what statistics are most important in determining a starting pitcher's season strikeout numbers.
- Through EDA, hypothesis testing, and model analysis, this project strives to find the ‘underrated’ or often ignored stats, to provide insight to MLB GMs on who to pursue in the free agency starting pitcher market.

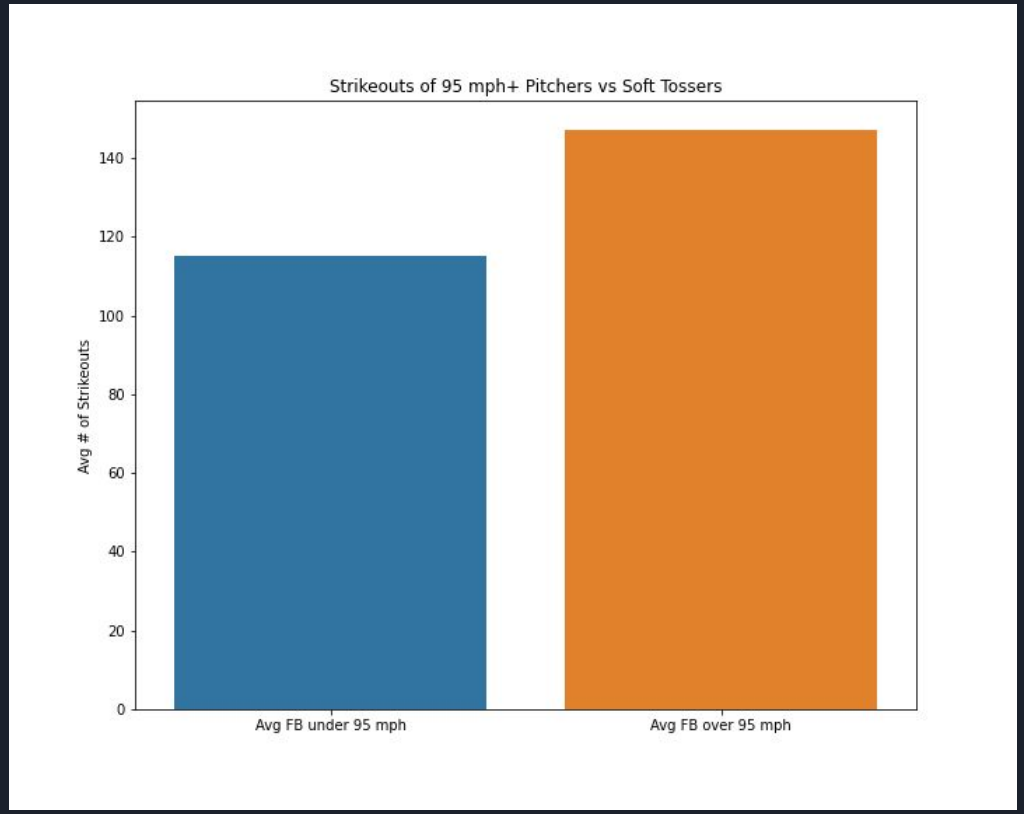
Why Strikeouts?



Dataset

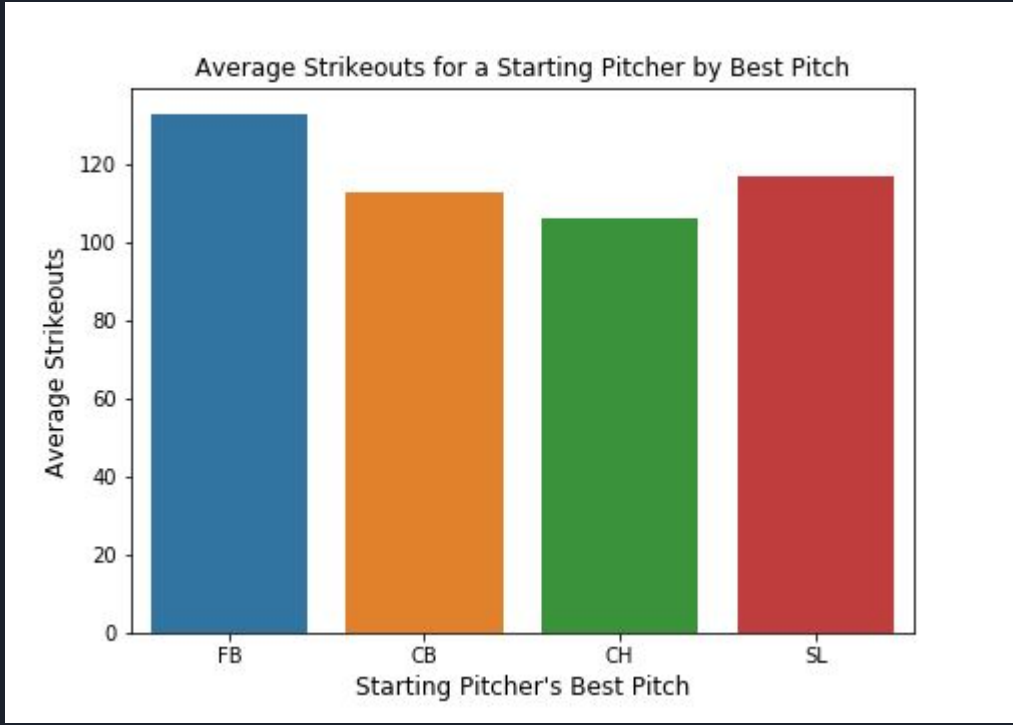
- Data Sources: Fangraphs and Baseball Reference via PyBaseball library
- Baseball Reference
  - Uses more traditional and commonly known baseball statistics
  - ERA, BB, Wins, Losses, IP, HR, Strikeouts, etc …
- FanGraphs
  - Uses Sabermetrics and Analytics
  - O-Swing%, F-Strike%, SwStr%, FB Velocity, FIP, Soft%,etc ..
- Compiled Dataset contained initially 1200 rows and 340 columns of SP season stats (2014-2019)

Exploratory Data Analysis / Stat Testing



Test Statistic: 5.3

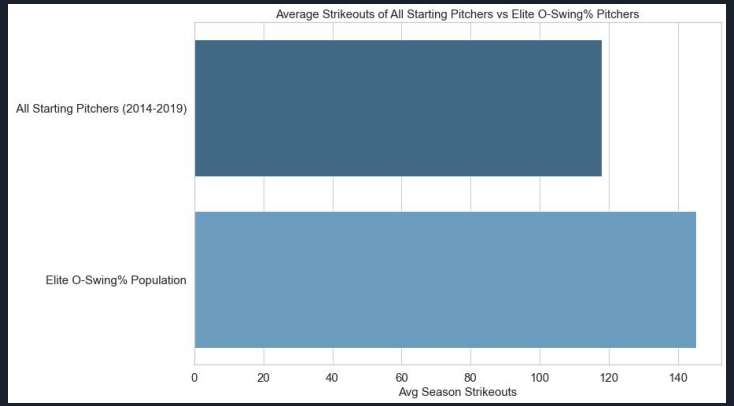
More Stats Testing & EDA



F-Statistic : 12.5

Stat Test/ EDA Summary

- GMs should look very closely at a pitcher's best pitch
- If that best pitch happens to be a fastball, GMs should then look for a pitcher with elite fastball velocity to best capitalize on a starting pitcher’s strikeout ability.
- F-Strike% is also vital in determining a starting pitcher’s ability to obtain strikeouts controlling for elite fastball velocity and high endurance. (P-Value = .002)



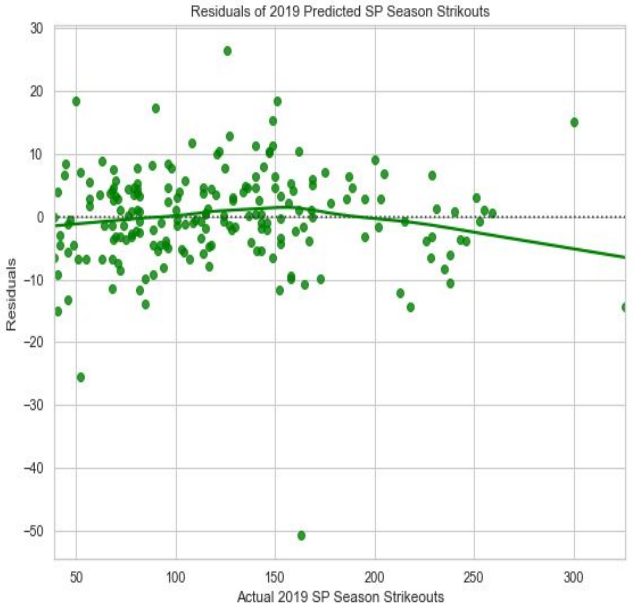
P-Value=10^-18
(essentially 0)

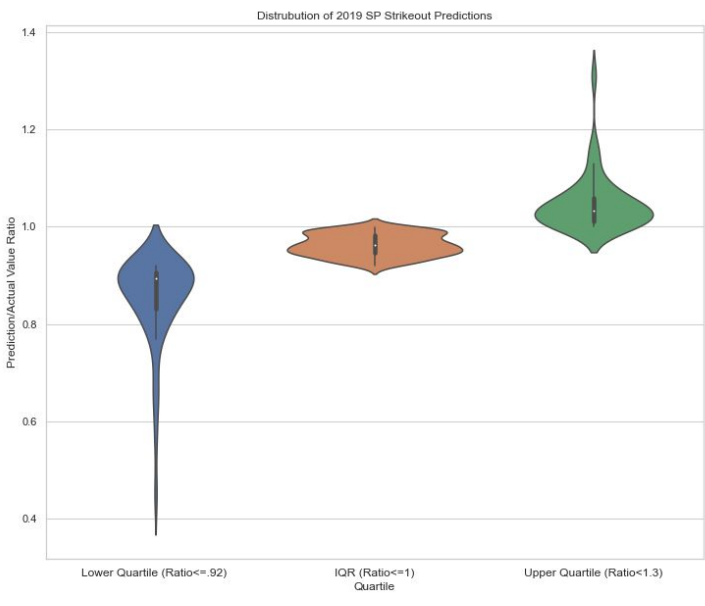
Modeling Selection

Model Version	Training RMSE	Testing RMSE
Basic Linear Regression	7.2 (strikeouts)	8.6 (strikeouts)
Linear Regression with recursive feature elimination algorithm	7.3 (strikeouts)	8.8 (strikeouts)
Linear regression with K best feature elimination method	8.5 (strikeouts)	10.3 (strikeouts)
Ridge Model	7.5 (strikeouts)	8.8 (strikeouts)

Training and Testing set consists of 2014-2018 Season SP stats, 2019 was used as holdout data.

Modeling Performance

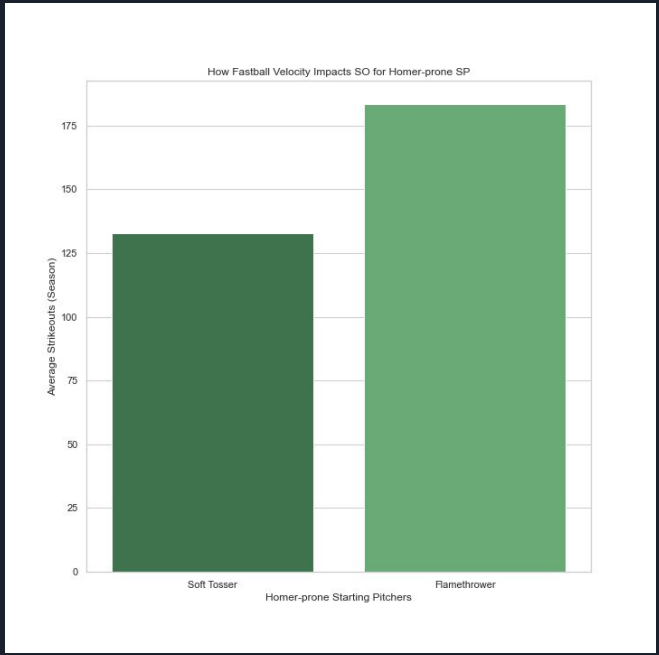


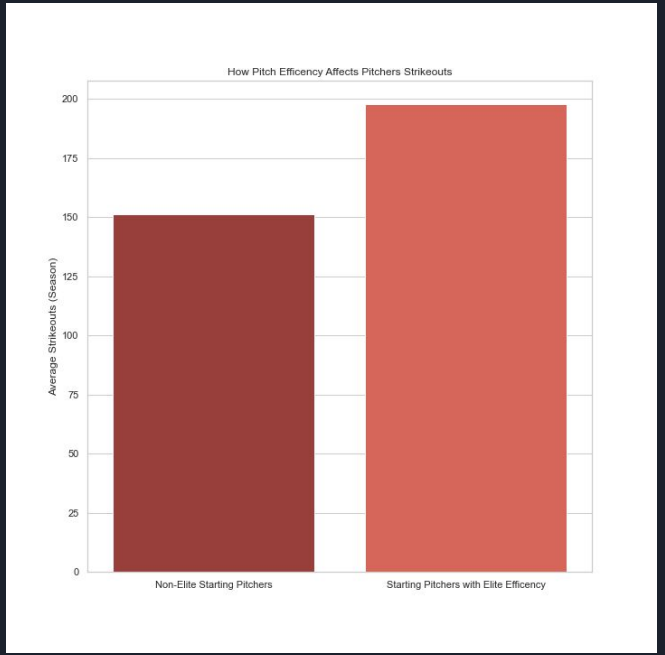


R squared = .97

Modeling Real World Takeaways

- Based on our model's performance, we feel that many important insights into our target variable can be found by model analysis.





Further Analysis / Next Steps

- Create a highly accurate interpretable models for relief pitchers.
    - This model would test to see if insights gained in our current model still hold for relief pitchers.
- Create a General Pitcher Predictive Model for ERA.
    - The goal would be to test whether the insights obtained during this modeling process would help the performance of this new predictive model.

Any Questions?