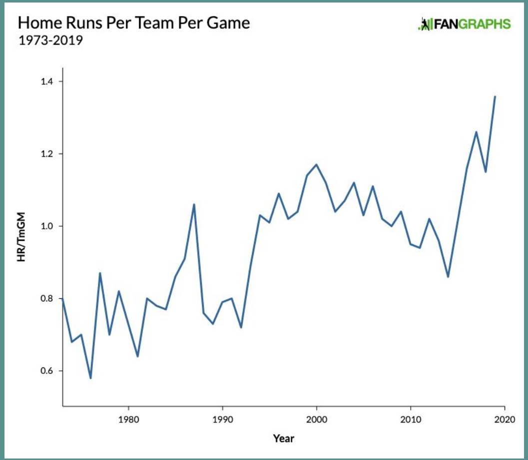
Predicting MLB Pitcher Value

Devan Bose, Jordan Denish, Malcolm Mashig, Christian Rogers

Questions We Want To Answer

- How do pitchers generally perform over time as they age and are we able to predict their performance in the future using advanced metrics?
- Are there certain types of pitchers that age better or worse?



	PLAYER	CONTRACT VALUE
1	Gerrit Cole STARTING PITCHER	$324,000,000
2	Stephen Strasburg STARTING PITCHER	$245,000,000
3	David Price STARTING PITCHER	$217,000,000
4	Max Scherzer STARTING PITCHER	$210,000,000
5	Zack Greinke STARTING PITCHER	$206,500,000

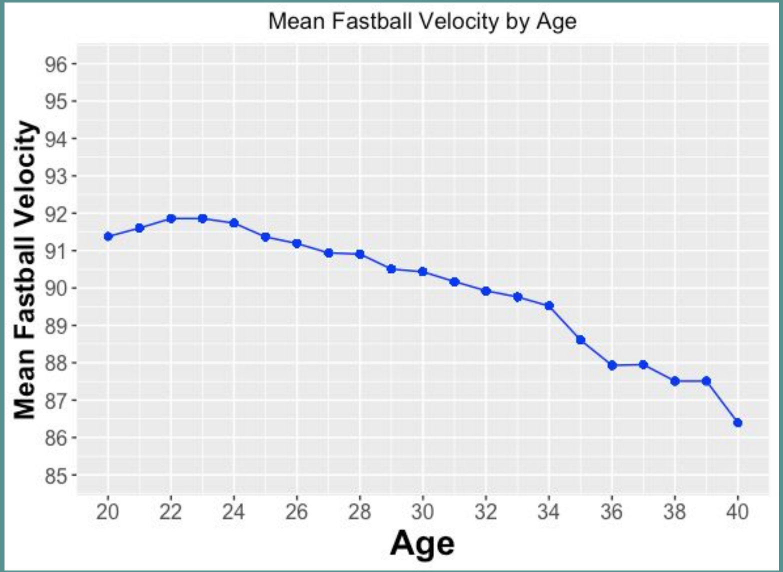
Data

- Exported initial data from fangraphs.com
  - Custom data finder (can set years, statistics, filter based on certain criteria)
- Cleaned and reorganized data in R to have one row per player, sorted by year, age, and amount of experience
- Initially included players with 100+ innings pitched in a season from 2002-2019
  - Expanded to include every season from this initial group of players in order to have the complete careers of these pitchers

	Season	Name	Team	Age	ERA	GS	IP	K/9	BB/9	K/BB	WHIP
1	2000	A.J. Burnett	Marlins	23	4.79	13	82.2	6.21	4.79	1.30	1.50
2	2001	A.J. Burnett	Marlins	24	4.05	27	173.1	6.65	4.31	1.54	1.32
3	2002	A.J. Burnett	Marlins	25	3.12	29	201.2	8.70	3.88	2.24	1.17
4	2004	A.J. Burnett	Marlins	27	3.74	19	118.0	8.31	2.90	2.87	1.18
5	2005	A.J. Burnett	Marlins	28	3.44	32	209.0	8.53	3.40	2.51	1.26
6	2006	A.J. Burnett	Blue Jays	29	3.98	21	135.2	7.83	2.59	3.03	1.30
7	2007	A.J. Burnett	Blue Jays	30	3.75	25	165.2	9.56	3.59	2.67	1.19
8	2008	A.J. Burnett	Blue Jays	31	4.00	34	220.1	9.35	3.47	2.69	1.33
9	2009	A.J. Burnett	Yankees	31	4.04	33	207.0	8.48	4.22	2.01	1.40

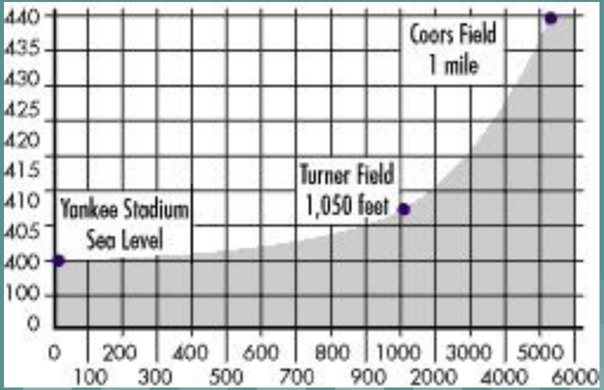
Initial Modeling

- In order to look into how well our data could answer our questions, we each worked on some preliminary modeling to look for interesting relationships between our variables.
- We found interesting relationship between age and velocity and several other pairs of our variables.
- This initial modeling helped us consider which metrics we would want to use to evaluate pitcher value.



How to Measure Value

- Metric should be **valuable** and **predictive**
- ERA (traditional metric) is very flawed
  - No information about fielding around pitcher
  - No information about park
  - Luck
- WAR (Wins Above Replacement) is valuable, but injuries make it hard to predict year-to-year
  - Injuries are mostly random and hard to predict



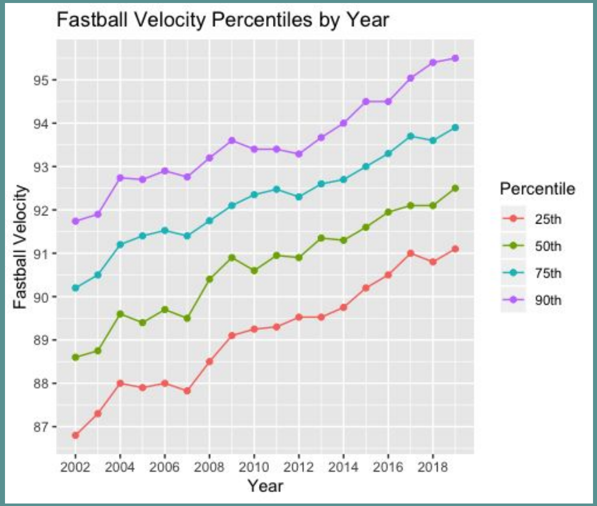
Our Value Metric: xFIP

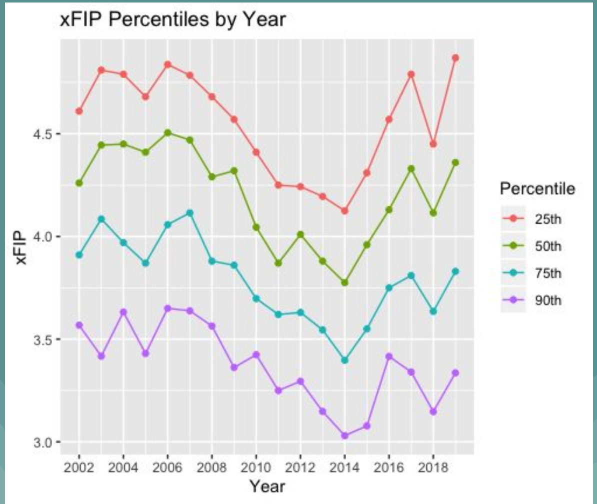
- FIP (Fielding Independent Pitching): only relies on outcomes that the pitcher can control (strikeouts, walks, home runs, hit batters)
  - Includes a constant that gives it a similar scale to ERA
- xFIP (expected FIP): use league average home run per fly ball percentage to remove further variance
- We believe that xFIP is both valuable and predictive

$$xFIP = \frac{13\times(Fly\ Balls \times LgHR/FB\%) + 3\times(BB + HBP) - 2\times K}{IP} + FIP\ constant$$

Our Methodology

- We want to build a model that can predict xFIP in future seasons.
- In order to use all of our data, to understand how pitchers will age on an equal scale, we standardized nearly all of our variables.
- In order to support the idea of standardizing, we ensured that differences between each percentile remained steady.





Lagged Linear Regression

- Use xFIP from past years to predict xFIP for the current year
- Other “lagged” variables utilized
  - Fastball velocity from past year
  - Fastball percentage from past year
- Age as a predictor
  - Yearly change is known
- Age category as a predictor
  - Young vs. Prime vs. Old

Building the Model

- "Submodels"
    1) predicted_FBvelocity ~ lag_FBvelocity + Age*Age_range
    2) predicted_FBpct ~ lag_FBpct + Age*Age_range
- xFIP Models
    1) xFIP ~ lag_xFIP + predicted_FBvelocity + predicted_FBpct + Age
    2) xFIP ~ Model1 + lag_xFIP2
    3) xFIP ~ Model2 + lag_xFIP3
- This way, we are able to directly use "real" data when we have it

Generating Predictions

- For each pitcher in each year, we generate a three-year prediction
- Steps:
    1) Calculate expected fastball velocity and percentage
    2) Plug into appropriate xFIP model
        - Model 1 if predicting 1st year with no lagxFIP2
        - Model 2 if predicting 2nd year or 1st year with lagxFIP2
        - Model 3 if predicting 3rd year
    3) Repeat steps 1 and 2 with predictions as predictors

Evaluating Beta Coefficients

- A pitcher's xFIP from their most recent season as by far the most important predictor in estimating xFIP in future years.
- Fastball velocity has a clear negative correlation with predicted xFIP value so we should expect harder throwing pitchers to have lower xFIPs that softer throwing pitchers.

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-0.32454	0.18149	-1.788	0.073996	.
predicted_fbp1	-0.01451	0.02679	-0.541	0.588327	
predicted_fbv1	-0.10241	0.03018	-3.394	0.000712	***
lag_age	0.01407	0.00633	2.223	0.026432	*
lag_xfip	0.48464	0.03038	15.953	< 2e-16	***
lag_xfip2	0.26639	0.02979	8.941	< 2e-16	***

A Test Case

- We tried many different combinations of predictors before ultimately, selecting our final model.
- Since our model factors in the multiple seasons, it can account for pitchers with wide variation in their xFIP values from season to season, weighing the most reason data the heaviest

	Name	Team	2018 xFIP	2019 Age	2019 FBV	2019 FBP	2019 xFIP	2020 xFIP (Predicted)	2021 xFIP (Predicted)	2022 xFIP (Predicted)
98	Lucas Giolito	White Sox	5.46	24	94.3 mph	55 %	3.66	4.13	3.93	4.01
101	Marco Gonzales	Mariners	3.59	27	88.9 mph	39.3 %	5.11	4.5	4.66	4.69

Accuracy Measures

- We performed cross validation by dividing our existing data into a training and test set 5000 times.
- Our model performs similarly on training and testing data.
- It makes sense that the predictions became worse as we extrapolated further.

	Year 1 Prediction	Year 2 Prediction	Year 3 Prediction
Train Absolute Mean Error	0.3398959	0.4423486	0.4583183
Test Absolute Mean Error	0.3516178	0.4445261	0.4783301

Displaying our Predictions

- Wanted something easy to use but also far more complex than just displaying the data frame online
- We built a Shiny app in R and exported to a website
- Seamless integration with R made it very easy to implement many features such as graphs made with ggplot

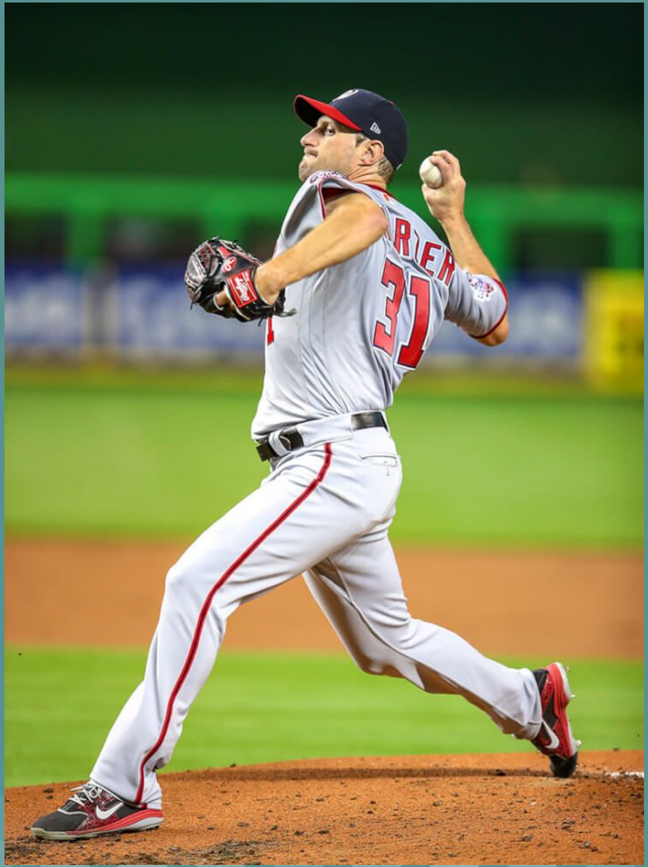
DEMO

Conclusions

- xFIP from most recent year most important predictor by large margin
  - Debunks the “down-year” theory
- Fastball velocity decreases as age increases
  - And xFIP declines alongside fastball velocity
- xFIP effectively normalizes for other more specific pitching statistics
  - Thus, these statistics did not prove to be valuable predictors
- Surprised some measures such as K/9 innings were not significant

Additional Conclusions

- The game of baseball has changed a lot over the past 20 years.
- Strikeouts and home runs have grown rapidly
- Ground ball and strikeout pitchers have become more valuable.
- We learned that predicting sports in general and especially baseball value becomes very complicated very quickly.



Back to Our Inspiration

- In December, Gerrit Cole signed a 9 year, 324 million dollar contract, worth $36 million annually.
- Gerrit Cole is 29 years and will finish out his contract at 38.
- Will he be able to perform at that contract value at 36, 37, and 38 as he ages and his fastball velocity falls?

Next Steps

- Evaluate our predictions against the 2020 season
  - Continue to improve our model
  - Model must evolve as the game of baseball evolves
- Use our model to assess fair contract value
  - Increasing size and length of contracts
  - Other factors such as past injuries, “star” value, and signing team’s championship odds need to be considered.

Questions?