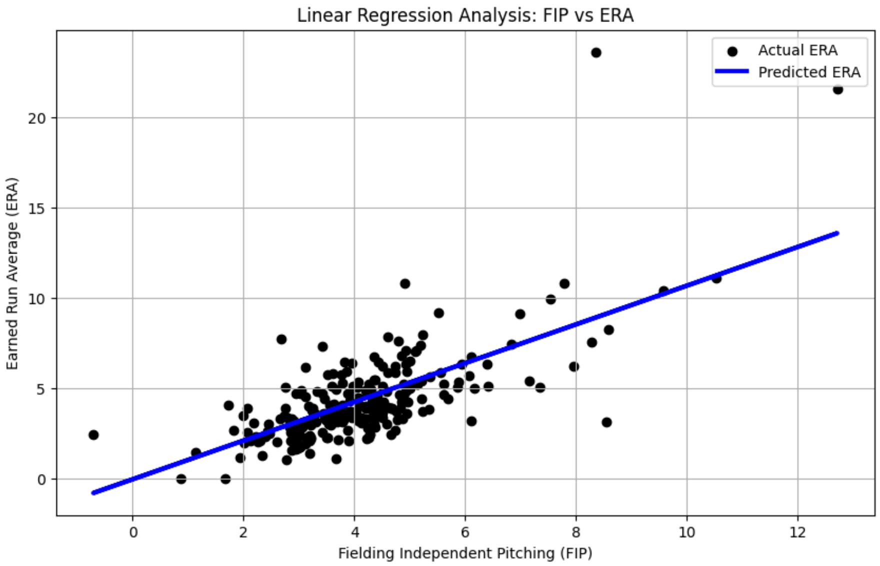
Names: Mitchell Stephens, Junhak Lee, Praneetha Popuri, Talal Alhammad, Kaya Farhat

The data set selected is 2022 MLB Player Stats (from Kaggle, taken from baseballreference.com). Our goal is to ultimately learn more about what makes up a successful pitcher. This data set includes all basic pitching stats like wins, ERA, and strikeouts along with more advanced ones like ERA+ and FIP. In order to thoroughly assess our goal, we came up with 4 questions and explored them using different machine learning models such as linear regression, k-means clustering, and logistic regression. The questions we explored discuss the significance of defense, identifying top-performing players, categorizing pitchers, and the impact of wild pitches on ERA.

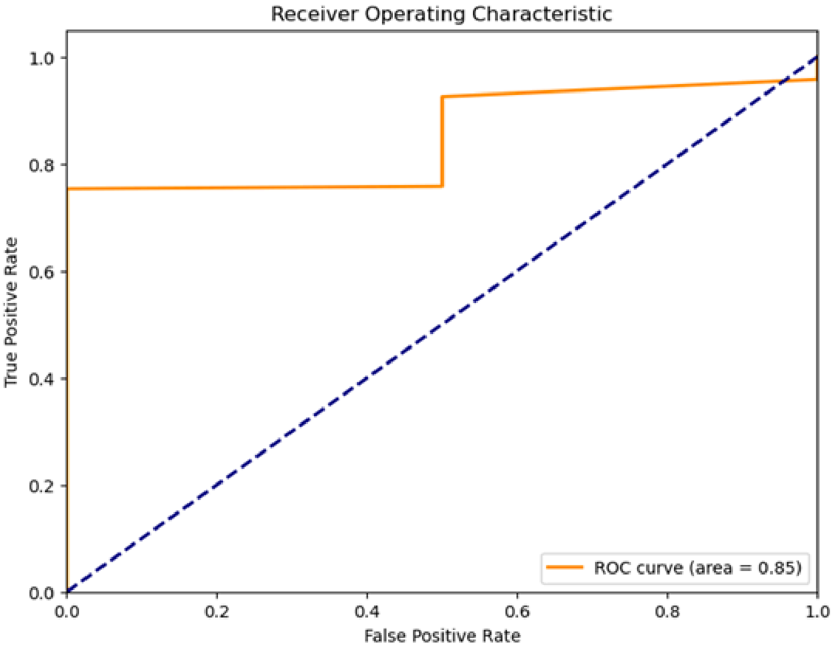
1. "How significant is non-pitcher-controlled defense when comparing ERA to FIP?"

Fielding Independent Pitching (FIP) was utilized to predict Earned Run Average (ERA), focusing on pitching independent of defense and external conditions. Our linear regression model, demonstrating an R^2 of 0.50, indicates that FIP explains about half of ERA's variability. The regression coefficient of 1.07 suggests a direct correlation, where each increase in FIP similarly raises ERA. However, a mean squared error of 3.05 points out the model's limitations, underscoring that defense and other factors also significantly affect ERA. A scatter plot supports these findings, highlighting FIP's significant yet incomplete influence on ERA, with defense playing a crucial role.



2. “Can we identify top baseball players based on wins, strikeouts, walks, and an ERA under 4.0?”

We evaluated MLB pitcher performance using logistic regression, predicting high performance with an ERA under 4.0 based on wins, strikeouts, and walks. We removed any players with less than 5 games as their results are irrelevant for our purpose. The optimal regularization strength, determined via cross-validation, was approximately 0.06, with precision for high performers at 99.1%.



We got an AUC of 0.85, meaning that there's an 85% chance that the model will correctly classify a randomly chosen positive instance as more likely to be positive than negative instance.

Feature	Coefficient
W	+0.16
SO	+0.04
BB	-0.11

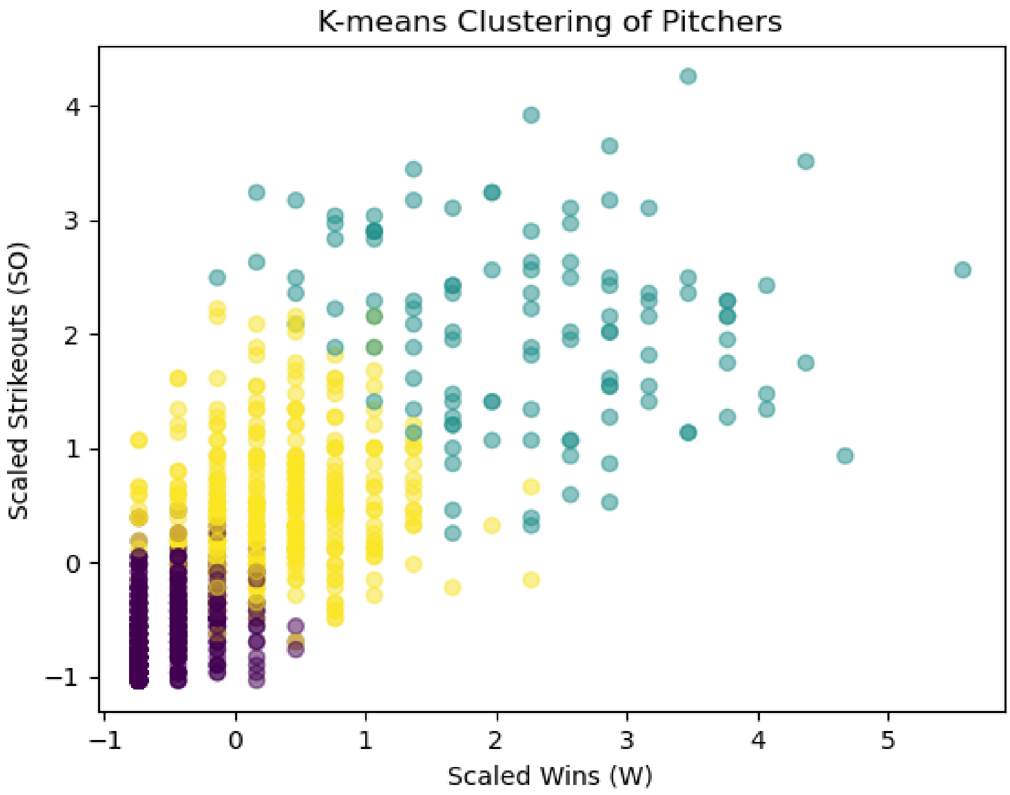
After feature analysis, we get these coefficients because the higher the number of wins, the better; likewise, the better a pitcher throws the ball, the higher the number of strikeouts, and the fewer walks allowed, the better.

3. “Can we group pitchers by their game results and pitching style?”

We investigated whether we could group pitchers based on their game results and pitching style. To answer this, we used K-means clustering in order to segment pitchers into distinct groups, which would reveal performance patterns.

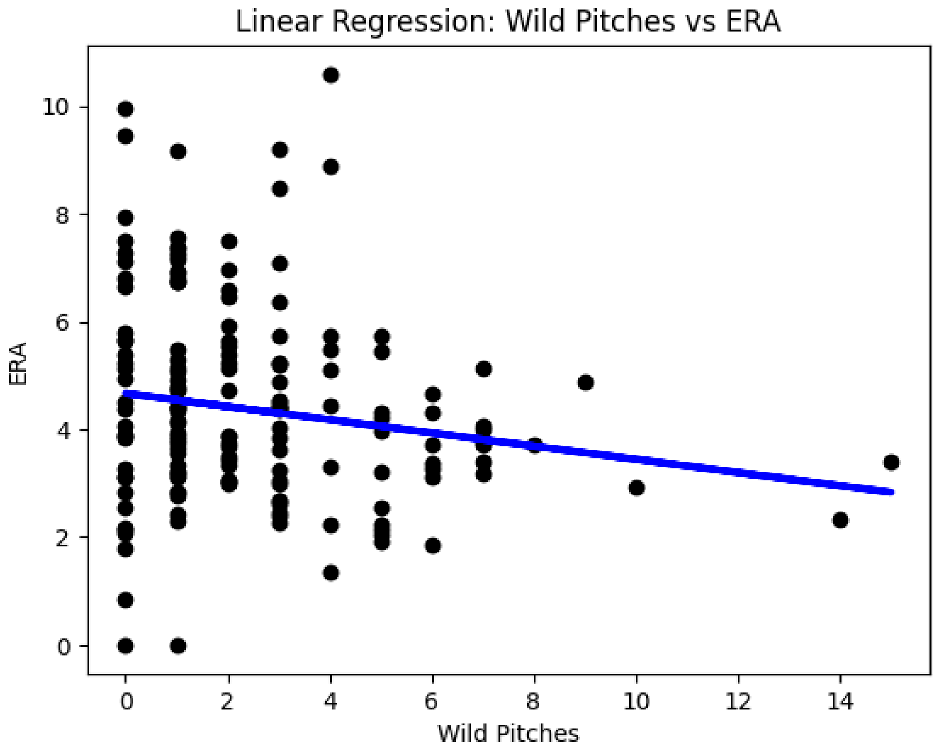
In this analysis, we used K-means clustering on the following pitching statistics: wins, losses, strikeouts, ERA, walks, and hits per 9 innings. After preprocessing the data with standard scaling, there were three distinct clusters: purple, yellow, and cyan.

The purple cluster likely represents less experienced or relief pitchers, while the yellow cluster consists of pitchers with moderate wins and strikeouts. The cyan cluster exhibits higher strikeouts, indicating power pitchers. The overall results of this model are satisfactory, with the exception of a couple points that seem to be an ambiguous color, likely due to the overlapping points with low opacity.



4. “Do wild pitches have an effect on ERA?”

We assumed that wild pitches would lead to a higher ERA, since those pitches are the least predictable and are unable to be caught by the catcher. Using a linear regression model, however, we yielded an extremely low accuracy score of 0.031. Therefore, we concluded that wild pitches do not have a significant effect on ERA.



After thorough analysis, there are clear overall trends that emerge. It is clear that striking out batters while also limiting walks is a clear indicator of a pitcher's success. Throwing wild pitches does not seem to matter much. Pitchers with lots of movements on pitches will be successful in increasing strikeouts even if that is at the expense of wild pitches. Increasing strikeouts also allows for less balls put in play as fielding still has a large impact on a pitchers success as some have a difference of 1 or 2 runs in their FIP and ERA and in a game when the average team only scores around 5 runs a game this is a huge difference in a pitchers success. These trends of increasing strikeouts while throwing harder with more spin is one that has taken over baseball in the last decade even if it is at the expense of the pitchers health with more and more pitchers suffering injuries each season.

Name	Proposal	Coding	Presentation	Report
Talal Alhammad	1	1	1	1
Kaya Farhat	1	1	1	1
Mitchell Stephens	1	1	1	1
Praneetha Popuri	1	1	1	1
Junhak Lee	1	1	1	1