MLB Players Analysis

Lukas Juranek, Collin Martin

November 17, 2024

Introduction

In the MLB industry, the average cost of a team is 2.16 Billion USD with an average of 3.95 million TV streams this 2024 postseason. Profitability in the MLB comes from a straightforward yet challenging formula. It's that winning games equals greater revenue, greater ticket sales, and greater merchandise profit, every factor of the team increases as money rolls in. However, making a winning team in baseball can be much more difficult than one can expect. There are many underlying factors that go into making a winning team, usually in first place, is the amount of money a team has in their disposable. For instance, the MLB World Series champions of 2024 were the LA Dodgers whose payroll cost 339.8 million USD, the second-largest payroll in 2024 following the New York Mets.

Among the many factors that may contribute to a team's success, the players often lead the race. Whether your team can afford these players is another thing. Although there are many challenges in the realm of baseball, one is that pitching performance can make or break a team's chances of winning during all games of the season. Understanding what makes pitchers successful through pitch selection, mechanics, or physical attributes requires a deep understanding of the game and the players.

This paper examines the patterns in MLB pitching performance, utilizing unsupervised learning techniques to reveal fundamental components that drive pitching success. With this idea in mind, the following questions emerge. What statistical metrics most reliably indicate pitching success? How do metrics reveal effective pitching styles? Can unsupervised learning techniques identify distinct pitcher profiles? With the metrics given, how can we build a cheap but successful team in the MLB today?

Dataset and Preprocessing

In order to get the data we wanted we searched external sources and found a website that had previously scraped MLB.com and converted it into our desirable CSV to load into R. However, this data was still in the text format and was not downloadable. With a simple Python script unrelated to the course works we successfully scraped the site and found our dataset. Below is a section of the data set with 4233 rows and 108 columns.

primaryPositionAbbrev <chr>	winningPercentage <chr>	runsScoredPer9 <chr>	battersFaced <int>	babip <chr>	obp <dbl>	slg <dbl>	ops <dbl>	strikeoutsPer9 <chr>	baseOnBallsPer9 <chr>
P	.---	0.00	3	1.000	0.667	1.000	1.667	27.00	0.00
C	.---	0.00	4	.000	0.250	0.000	0.250	0.00	9.00
P	.---	0.00	4	.500	0.500	0.333	0.833	13.50	13.50
3B	.---	0.00	1	.---	0.000	0.000	0.000	27.00	0.00
P	.---	0.00	13	.286	0.385	0.200	0.585	10.13	10.13
P	.---	0.00	22	.231	0.182	0.143	0.325	12.00	0.00

Figure 1: Dataset Preview

We then preprocessed the data to make the data set more understandable. These steps included changing the data into numeric values for those that needed to be changed such as winPercentage, runsScoreper9, etc. The data also contained many players with very little data. For instance, if someone were to pitch only 1 inning in a game they would have been registered in the data set. To improve accuracy, we filtered out players who participated in fewer than 5 games during the 162-game season and focused only on players from the 2024 regular season. This step reduced our data set to 855 players and was a successful way of getting rid of outliers since it's hard to exclude players from our dataset as it is often for players to differ in skill and ability. We also performed feature engineering to convert some features into percentages for greater relevance in our data. This was applied to the features seen below. We then removed columns irrelevant to our analysis (a full list of which is available in the data set) and eliminated 1 feature of each pair that obtained a correlation of 1. By looking at the correlation matrix of our dataset we also made sure it contained enough correlation between features to make it suitable for the unsupervised learning methods we later applied.

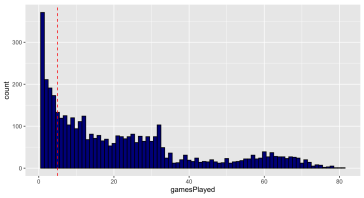
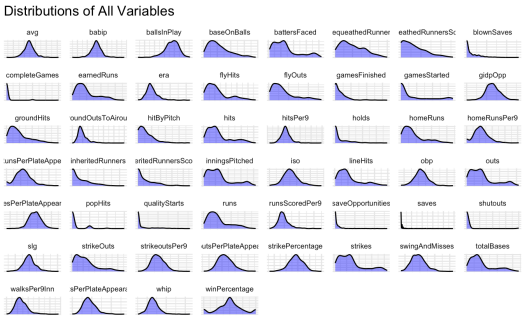


Figure 2: Data deleted from players with not enough games played

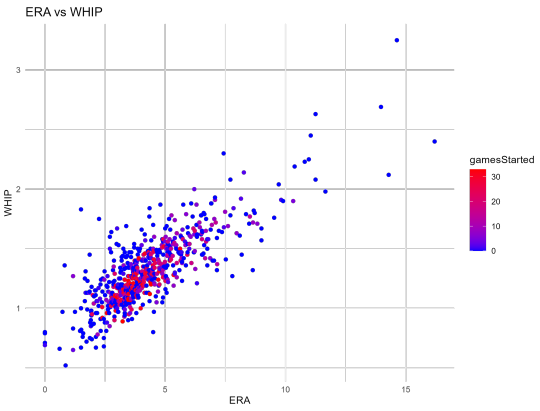
- gidpOpp = gidpOpp / battersFaced
- swingAndMisses = swingAndMisses / battersFaced
- ballsInPlay = ballsInPlay / battersFaced

Data Analysis

In our data analysis, we utilized many graphing techniques to gain insights into the features of the data presented. To start, we examined important pitching metrics like ERA(Earned Run Average, which is how many runs the pitcher gives up), Strikeouts per 9 innings, WHIP (Walks and Hits per Inning Pitched) using histograms. These graphs highlighted any skewness and outliers of our data. For example, shown below in figure 3a, the ERA histogram revealed a little bit of a right skewness to the data. Next, we used scatter plots to explore any linear relationships between features. For example, we plotted WHIP against ERA to asses if lower WHIP values correlated with Lower ERA values showing greater pitching efficiency. As you can see in the graph the two correlate well, thus the players who have fewer walks and hits earn fewer runs.



(a) Distributions of All Variables



(b) Scatter plot of ERA vs WHIP

Next, we examined the ERA between starting pitchers and relieving or closing pitchers to understand the distinct roles and performance patterns of pitchers in the league. Starting pitchers usually pitch at the beginning of the game and for longer amounts of time, around 5 to 7 innings, and are expected to set the tone for the game by facing each batter multiple times. The elongated time pitching is one reason why Starters would have a high ERA. Relievers and Closers, on the other hand, come after the starters for the last 2 to 3 innings and oftentimes in high-stress situations. These clutch-time situations would be a significant reason why closers/relievers would cause more runs. To see which causation was in place inside our data set we graphed the figure below. As you can see closers/relievers do have a more variable ERA on average due to the high-stress situations they are constantly in. However, starters tend to have a higher ERA on average due to the fact they face more batters and pitch more innings presenting the factor of fatigue in one's attributes. This comparison highlights differences in endurance, consistency, and the ability to handle the pressure of situations, which is crucial for optimizing pitching rotations and managing player performance. These features are the ones that we hope the PCA and Factor analysis can analyze.

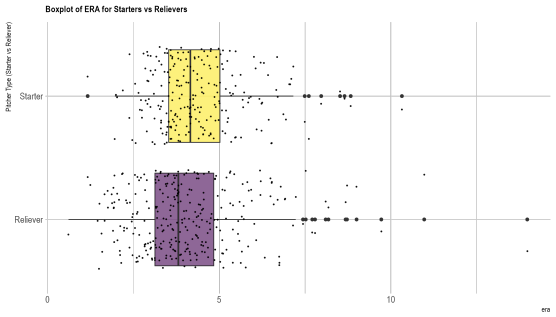


Figure 4: ERA Box plot Between Archetypes

The graph above labels two boxplots, Reliever and Starter. These are not factors of our dataset but rather engineered. Starting pitchers were selected if they had pitched more than 4 games started within the season. Which is a feat only starting pitchers could accomplish. Relieving pitchers were selected by having more than 4 games finished. Only Relieving pitchers carry out the task of ending games except in certain situations when a starting pitcher plays the whole game (which is highly unlikely thus more than 4 games are finished).

Unsupervised Learning

Principal Component Analysis

To improve our MLB analysis, we used Principal Component Analysis to reduce the complexity of our dataset while retaining the essential information. With 108 original features we collected for each pitcher in the league, we aimed to eliminate the noise of the data set and simplify these metrics to fewer uncorrelated principal components. First, we needed to standardize the data to ensure each feature had a mean of 0 and a unit variance. We then applied PCA to the dataset, transforming our features into a set of principal components ranked by their contribution to the overall data variance. The number of components was chosen by Kaiser's Criterion. A criterion where if your loadings eigenvalues have a higher value than 1, they are kept as they explain more than an original variable from the data set. The number of factors received from this test was 8. With 8 components selected, we would be able to explain 84 percent of the data set variance. Although we achieved the optimal number of components which explained a high variance as well, we figured through more analysis

it would be hard to make sense of the low dimensional components as they explain very little of the variance. Thus the reason why 4 or 5 components seemed more than enough to perform analysis from this point. We felt more comfortable with the lower dimensional data as we could not make sense of the end data making it seem more reasonable to choose 4 or 5. After this process, we found the structures behind each component. For example, Component 1 primarily represented starting pitcher-like classifications and performance. This expressed positive values such as batterFaced, strikes, outs, inningsPitched, total bases, etc. while in the negative it showed walksPer9Inn, walksPerPlateAppearance, etc. providing us clear insight that these were high-performing starting pitchers within our data set. Component 2 expresses the players in the positive direction as being able to be hit off of and give up runs. These people tended to have a high SLG, AVG, and HitsPer9 while having low strikeout averages. Extracting the analysis from this model we can see that the PCA did analyze the archetypes of different pitchers which is what we were hoping for. Now with our PCA model, we dive deeper into the analysis of our data set.

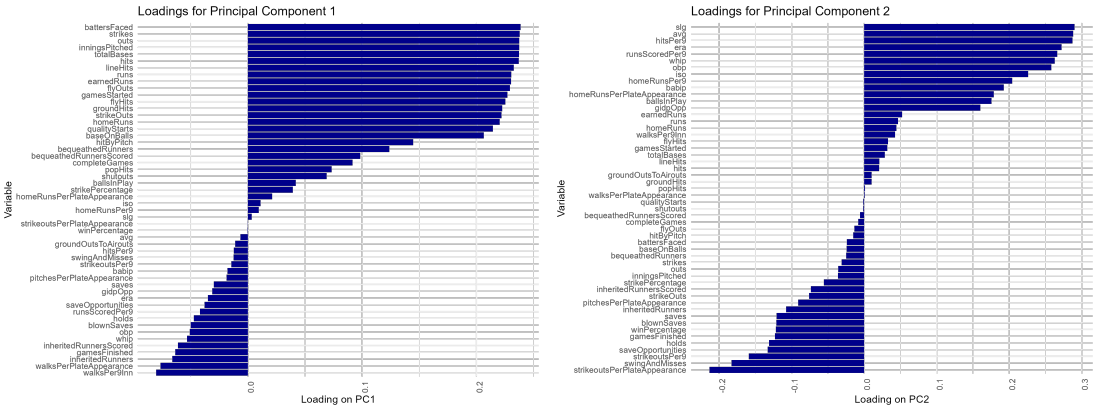
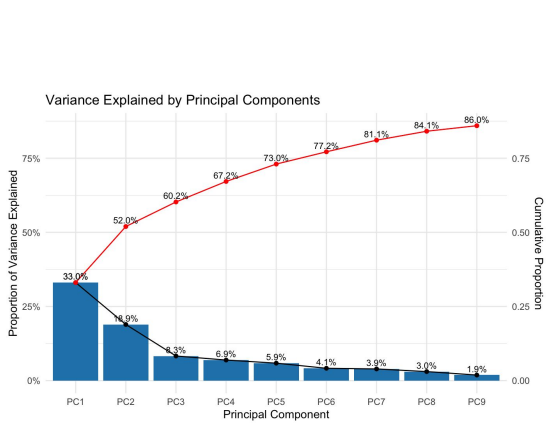
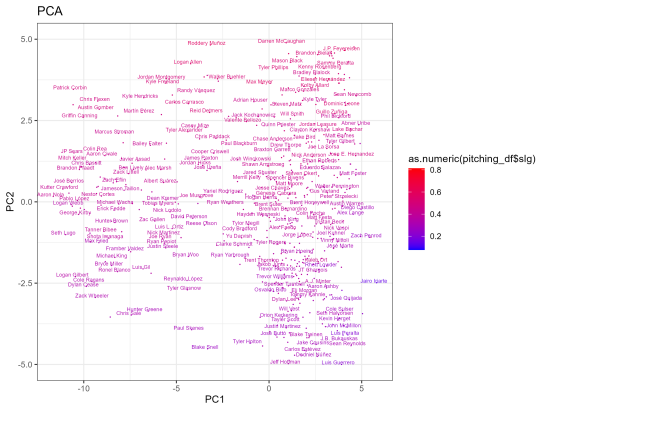


Figure 5: Loadings for Principal Component 1 and 2

Below is the PCA analysis within the first two factors. If you are a baseball fanatic you are able to see the correlation between the dimensions. For example, we have the top 10 pitchers whom have had the best innings pitched in the most bottom left in the graph, these players are (Zach Wheeler, Cole Raggans, Chirs Sale, Seth Lugo, etc.. These players in the bottom left of the graph are also known to have a very high velocity and low contact rate. Which referencing the Loadings above the best players mentioned were negative in PCA Dimmension 1 as they didn't log many hits and a low number of TotalBases and also negative in PCA Dimmension 2 which correlates to the High Strikeout per plate appearance and high swing and miss rate. This concluded our PCA model and thus the reason we went to Factor analysis to compare the two models side by side without misunderstanding the data given. Also below is the graph of the variance explained over time. Shown is the individual and the cumulative variance explained.



(a) Variance Explained by Principal Components



(b) Pitcher characteristics visualized by first two principal components

Factor Analysis

Factor analysis was used next as a method to identify structures within the dataset. A factor model using factanal with three factors was extracted from our reduced data set. The data set needed to be reduced to make use of factanal. This was done by eliminating variables with a near zero variance and also very high correlating factors with more than a .95 correlation. The data was then scaled and the output was given as follows in the figure below. Concerning the MLB.Rmd

file you can see Factor 1 very much resembles a starting pitcher with positive metrics in bequeathedRunners(the number of runners allowed after the pitcher has left the game) and gamesStarted. Factor 2 resembles the pitchers who you need to avoid while scouting. For example, this factor has a loading of .991 in the positive direction of slg which is quite bad. Finally, Factor 3 resembles closing/relieving metrics with high Inherited Runners (the runners already on base when a pitcher enters a game), gamesFinished, holds, blownSaves, and saveOpportunities. The variance portrayed by these factors covers a total of 50 percent as seen below.

	Factor1	Factor2	Factor3
SS loadings	7.807	5.501	3.843
Proportion Var	0.230	0.162	0.113
Cumulative Var	0.230	0.391	0.504

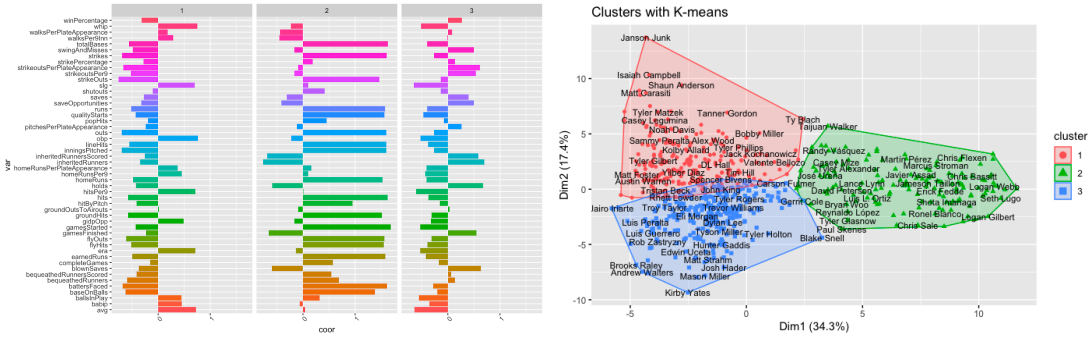
Figure 7: Factor analysis result

Unfortunately, this is the highest number of factors we could have used for our data set without receiving an error. However, the data does give us a little better representation of starters vs relievers in our dataset than the PCA did. However, the PCA explained much more data within fewer dimensions reducing the complexity of the model. For example, the PCA explained Starters vs Relievers within the 1st dimension along with hittable vs non-hittable pitchers in the 2nd dimension while the Factor analysis explained the Starters, Relievers, and bad players within 3 factors.

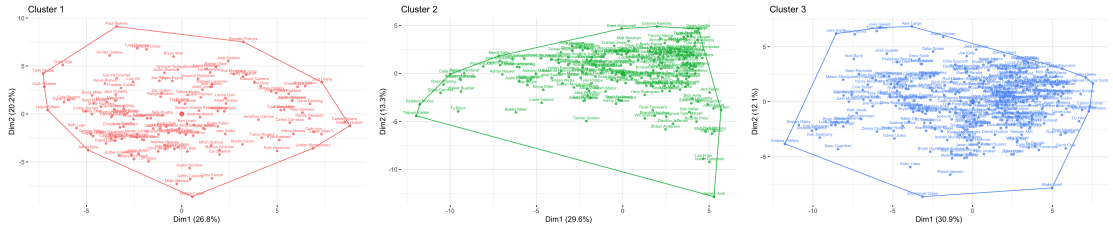
Clustering

After strong analysis and now knowing that these models are capable of correctly classifying the archetype of a current MLB pitcher we moved on to our clustering model which hopefully could do the same but better.

To uncover more patterns and groupings within the pitching dataset, we applied K-Means clustering. This method allowed us to segment the data within distinct clusters based on similarities(perfect for pitching archetypes). The cluster number was selected by using the Silhouette score that chose an optimal cluster numbering of two. However, given we are familiar with the data set and can apply our nuance, we have increased the number of clusters to 3 to add more dimensionality to it. The clustering grouped players based on metrics, as seen below, such as WalksPer9Inn, Strikeouts, runs, OBP, games started, etc. Cluster 1 seems to include high metrics such as WHIP, SLG, and OBP which correlate to pitchers who tend to give up hits and runs. We call these players the bad players. Cluster 2 looks to have high metrics in TotalBases, but also strikes, strikeouts, runs, quality starts, outs, innings pitched, and low in the era. This cluster looks to be the players who are often more efficient than anyone else they play against. Through a lot of games, they will be consistent for the team throughout the season. Cluster 3 is full of high-quality starters and relievers. Why, because metrics like Completed games, and inherited runners make them stand out as relievers/closers, and metrics like WHIP, OBP, and ERA explain that they are the good ones. Like factor analysis, the K-Means cluster was good at filtering out the bad players in Cluster 1 and leaving the high-hanging (high-performance players) fruit for the second and third clusters.



Below you see each Cluster Plot individually. Through further insights and analysis of this cluster, we know cluster 1 below is cluster 2 reflected about the x-axis for better interpretation. This cluster explicitly shows starting pitchers. Whether you are a good starting pitcher or not depends on the -x slope of the graph (The further top left you are the better of a player you are). For instance, Paul Skenes, Tyler Glasnow, and Hunter Greene are now regarded as some of the best-starting pitchers in the MLB. For cheaper players, you go further into the bottom right of the graph to get the best value for a fraction of what these top players are worth. For instance, Hunter Greene earns 8,000,000 USD per year in his contract while a person from the middle of the graph, Andrew Abbot, attains a yearly salary of 1,300,000 USD nearly 6 times less than Hunter Greene. Cluster 2 we can disregard the dataset entirely as we consider the noise of the data set. These are no good players who were either hurt the entire season or barely made the roster. Cluster 3 is mostly relieving/closing pitchers. However, the lower you go in Dim2 the better players you are seeing. For the most value for your dollar, scout from the middle of the graph. For a top-tier ace, look toward the bottom.



(a) Cluster 1 (b) Cluster 2 (c) Cluster 3

After these clustering models with K-Means we performed the same model but with the PCA analysis, we performed prior. With this, the first few PCA components explained a high variance of the data set and decreased rapidly after. These components provided a clear visualization(as seen in the mlb.Rmd file) of the player distribution across clusters which highlighted latent structures in our dataset. The approach offered a clear visualization of player distribution across clusters as values of the PCA while reducing noise. However, this PCA-based clustering method did not provide the level of insight we anticipated as we thought when we first created our PCA model. With more than 3 PCA dimensions the graph became messy and very hard to interpret. Although, with the number of clusters equal to those of the K-Means they closely resembled each other with most players still divided into the same group. While PCA provided us with a theoretical advantage in simplifying the data, the impact of the cluster model was deemed insignificant. For this reason, we mostly used the PCA as a tool to gain insight from the data set and how these unsupervised learning methods would perform on the dataset.

After performing K-Means clustering, we finished our cluster analysis by using Partition Around Mediods. PAM differs in that it selects actual data points as the center of each cluster rather than computing the mean. This makes PAM more tuned to outlier data which makes it a perfect model for our data as many players vary in skill and level. Using the Silhouette elbow method with PAM we achieved the same amount of centers as before with 2 but decided to choose 3 again for more interpretation. The cluster almost looks identical to the K-Means data although, the insight we gained from this method was something that validated our reasoning. As you can see below, there are two players we are interested in, Kyle Harrison and Sam Long. And if you were to recall from Figure 10 you can see that Kyle Harrison is likely a starting pitcher and Sam Long is a Reliever/Closing pitcher. If you look below you can see that they are chosen for the PAM Mediods and do follow our reasoning.

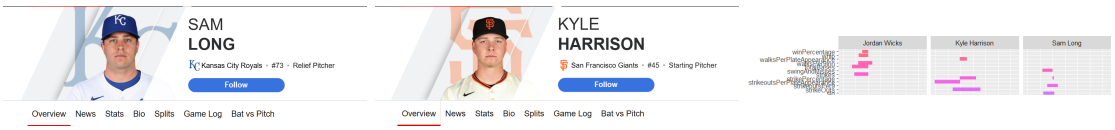


Figure 10: An example of PAM clustering distinguishing Starters from Relievers/Closers

Conclusion

Our analysis offers valuable insights into MLB pitching performance and provides a foundation for identifying cost-effective strategies to build a successful pitching line-up. Starting with data preprocessing, feature engineering, and unsupervised learning techniques, we uncovered details that helped classify the archetype of the pitcher and also the factors driving their success.

The primary questions we asked at the beginning of this paper were effectively addressed. By examining essential metrics like batters Faced, Strikes, and total bases we revealed how these metrics distinguish different pitching archetypes. For instance, our model measured starting pitchers by their ability to sustain performance over a long period, and for relieving and closing pitchers the model measured these players to be high-performing in short-term situations.

Despite the valuable insights gained, our work also highlights the complexity of baseball analytics. While pitching is a crucial factor in the realm of baseball, there are many other factors to building a successful team as this is just one piece of the broader puzzle of a championship team. Therefore in this project, we struggled to provide nuance to the game of baseball which is already one of the most statistical games ever played. However, this project underlines the potential for data-driven approaches for scouting and recruiting purposes. For example, if an MLB team can obtain 13 pitchers in their roster, it's usual that 5 starting pitchers are chosen and the rest are relievers or closers. With this knowledge, one can dive into this project with no baseball knowledge whatsoever and leave with an effective pitching lineup.