What Makes a Cole-ity Start?

Emma Troast

Stanford University

Figure 1: *Gerrit Cole pitching for the New York Yankees*

Introduction

In baseball, there is certainly no shortage of metrics for evaluating players. We often use traditional counting statistics like earned runs, hits, strikeouts, or innings pitched to determine whether or not a pitcher performed well in a game, season, or career. For a single game, starting pitchers are often deemed as having recorded a "quality start" if they pitch at least 6 innings and allow no more than 3 earned runs, a somewhat arbitrary line that misses the mark in many ways.

Is there a better way to analyze and distinguish starts other than as win/loss or quality/not quality? In this project, I use Principal Component Analysis to analyze Gerrit Cole's career starts from box score data. By reducing the dimensionality of this dataset, we can better find relationships between variables, visualize data in two and three dimensions, and discover what makes a "Cole-ity" start by each principal component. This allowed me to find general trends over the course of Cole's career in Pittsburgh, Houston, and New York and discover some of Cole's best starts through the power of matrix algebra.

Methodology

Data

My data set consists of Gerrit Cole's career starts from 2013 through June 3, 2023. I will be using select box score metrics from each outing as variables for PCA.

Principal Component Analysis

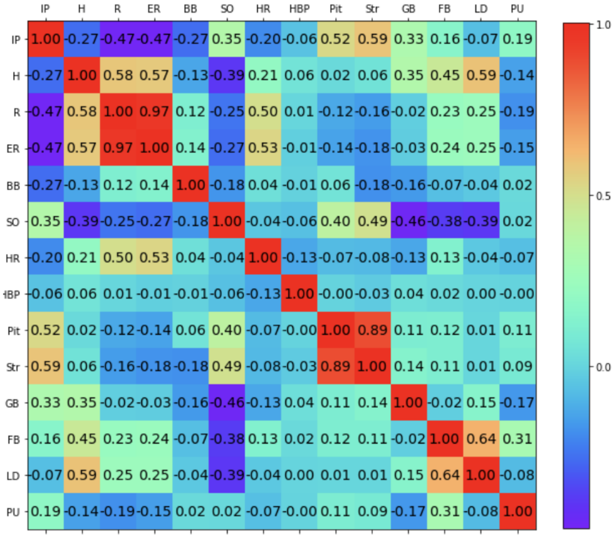
Principal Component Analysis (PCA) is a tool used in statistics to reduce the dimensionality of a data set with a large number of interrelated variables while retaining as much variance in data as possible.

PCA consists of the following steps:

1. Compute the sample covariance matrix $\frac{1}{n-1}AA^T$
2. Find SVD of $\frac{1}{\sqrt{n-1}}A = U\Sigma V^T$ and $\frac{1}{n-1}AA^T = U\Sigma\Sigma^T U^T$
3. The columns of U are the principal components in order of greatest variance. Note that $\frac{1}{n-1}AA^T U = UD$ where D contains the squares of the singular values of A on its diagonal. So, the principal components are the eigenvectors of the sample covariance matrix with corresponding eigenvalues denoting explained variance.

For the purpose of this project, I will be performing PCA with Python's scikit-learn library.

The rows of my matrix represent a start and columns represent variables such as innings pitched, hits, and runs. I standardized the features by removing the mean and scaling to variance. The covariance of standardized variables is equivalent to their correlation. Below is the covariance matrix [1].



	IP	H	R	ER	BB	SO	HR	HBP	Pit	Str	GB	FB	LD	PU
IP	1.00	-0.27	-0.47	-0.47	-0.27	0.35	-0.20	-0.06	0.52	0.59	0.33	0.16	-0.07	0.19
H	-0.27	1.00	0.58	0.57	-0.13	-0.39	0.21	0.06	0.02	0.06	0.35	0.45	0.59	-0.14
R	-0.47	0.58	1.00	0.97	0.12	-0.25	0.50	0.01	-0.12	-0.16	-0.02	0.23	0.25	-0.19
ER	-0.47	0.57	0.97	1.00	0.14	-0.27	0.53	-0.01	-0.14	-0.18	-0.03	0.24	0.25	-0.15
BB	-0.27	-0.13	0.12	0.14	1.00	-0.18	0.04	-0.01	0.06	-0.18	-0.16	-0.07	-0.04	0.02
SO	0.35	-0.39	-0.25	-0.27	-0.18	1.00	-0.04	-0.06	0.40	0.49	-0.46	-0.38	-0.39	0.02
HR	-0.20	0.21	0.50	0.53	0.04	-0.04	1.00	-0.13	-0.07	-0.08	-0.13	0.13	-0.04	-0.07
HBP	-0.06	0.06	0.01	-0.01	-0.01	-0.06	-0.13	1.00	-0.00	-0.03	0.04	0.02	0.00	-0.00
Pit	0.52	0.02	-0.12	-0.14	0.06	0.40	-0.07	-0.00	1.00	0.89	0.11	0.12	0.01	0.11
Str	0.59	0.06	-0.16	-0.18	-0.18	0.49	-0.08	-0.03	0.89	1.00	0.14	0.11	0.01	0.09
GB	0.33	0.35	-0.02	-0.03	-0.16	-0.46	-0.13	0.04	0.11	0.14	1.00	-0.02	0.15	-0.17
FB	0.16	0.45	0.23	0.24	-0.07	-0.38	0.13	0.02	0.12	0.11	-0.02	1.00	0.64	0.31
LD	-0.07	0.59	0.25	0.25	-0.04	-0.39	-0.04	0.00	0.01	0.01	0.15	0.64	1.00	-0.08
PU	0.19	-0.14	-0.19	-0.15	0.02	0.02	-0.07	-0.00	0.11	0.09	-0.17	0.31	-0.08	1.00

Figure 2: *Covariance Matrix*

We see that runs and earned runs are highly collinear, as are pitches and strikes, which follows our intuition of the game.

Results

The first eigenvector of the covariance matrix is the principal component with the largest eigenvalue and explained variance in the data [2]. We want to reduce the data set to the fewest principal components possible while retaining a high ratio of explained variance.

The explained variance will taper off with each principal component. PC 1 accounts for 27.0% of explained variance, PC 2 for 18.0%, PC 3 for 12.8%, and so on. The first five principal components explain over 75% of variance and the first ten describe over 97%. This trend is displayed by the figure below [3].

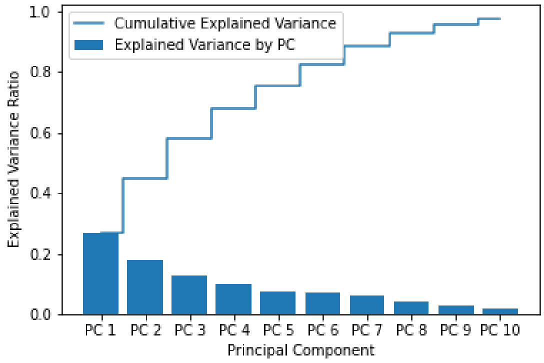


Figure 3: *Explained Variance Ratio by Principal Component*

For visualization purposes, it is often necessary to use only the first two or three principal components.

How does this provide information within the context of baseball? Principle components are linear combinations of our original variables, where the coefficient or "loading" of a variable tells us how strongly a variable affects a given component. We can examine and interpret the first three principal components below.

Table 1: *Principal Component Loadings*

	PC 1	PC 2	PC 3
IP	**-0.339**	**-0.343**	0.054
H	**0.356**	**-0.335**	0.044
R	**0.425**	-0.072	**-0.319**
ER	**0.430**	-0.063	**-0.320**
BB	0.080	0.167	-0.104
SO	**-0.325**	0.017	**-0.462**
HR	0.234	0.005	**-0.416**
HBP	0.016	-0.005	0.137
Pitches	-0.227	**-0.423**	-0.293
Strikes	-0.251	**-0.456**	-0.280
GB	0.032	-0.265	**0.391**
FB	0.190	**-0.391**	0.129
LD	0.250	**-0.340**	0.241
PU	-0.105	-0.082	0.043

The most important variables in PC 1 are IP, H, R, ER, and strikeouts. PC 1 is inversely related to IP and strikeouts and directly related to H, R, and ER. If a start has a large, negative PC 1 value, it means that Cole pitched a lot of innings and had a lot of strikeouts without giving up a lot of hits or runs. This is what a team would want from Cole as the ace of its pitching staff and a strikeout pitcher.

Since this is the principal component explaining the highest ratio of variance in the data, we can look at PC 1 to distinguish between Cole's good and bad starts- answering what makes a "Cole-ity" start. The start with the lowest PC 1 value came in 2018 for the Astros when Cole pitched a one-hit shutout with 16 strikeouts.

This interpretation can be applied to other principal components. Games with large, negative PC 2 values are starts in which Cole pitched deep into games, but gave up a lot of hits and hard contact. Games with large, negative PC 3 values are starts in which Cole struck out a lot of batters, but gave up a lot of home runs and very few ground balls.

The figures below are plots of the loadings in two and three dimensions [4].

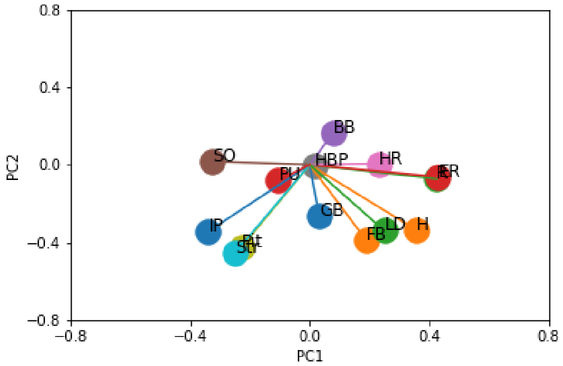


Figure 4: *2D Plot of PC Loadings*

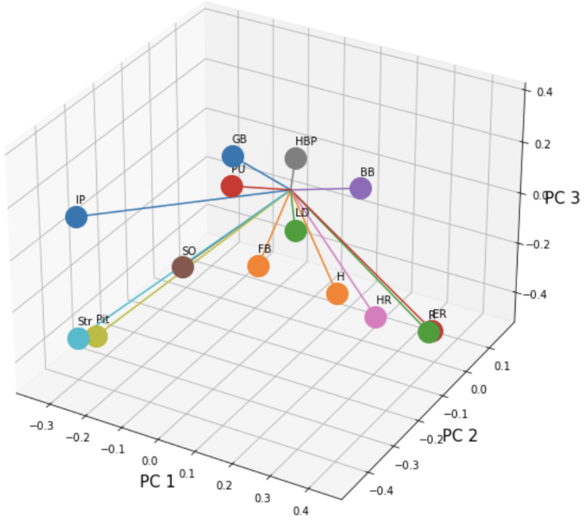


Figure 5: *3D Plot of PC Loadings*

PCA is a powerful tool for data visualization as well. It would be impossible to visualize Cole's starts with thirteen variables; however, principal components make this possible in two or three dimensions.

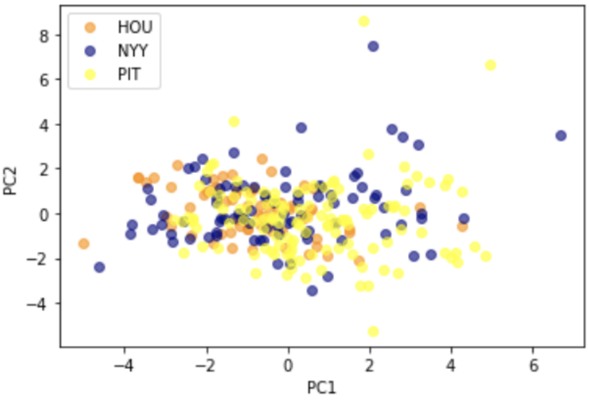


Figure 6: *2D Scatter Plot*

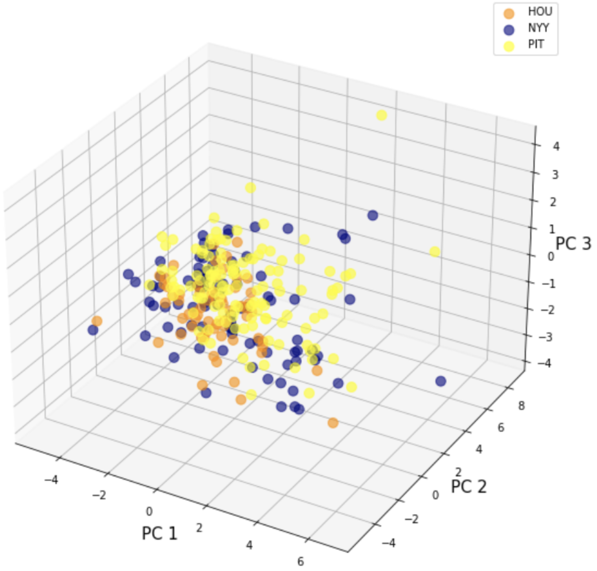


Figure 7: *3D Scatter Plot*

The scatter plot in two dimensions is more readable and interpretable. Based on our analysis of the principal components, we know a good start would have a negative PC 1 value.

What does this tell us about Gerrit Cole? To answer our original question of what makes a "Cole-ity start", we can define this as a start with a negative PC 1 value, a high volume, high strikeout, low hit, low scoring game. With this interpretation, we can analyze trends over Cole's career.

While pitching for the Pittsburgh Pirates, 54/127 or 42.5% of starts were "Cole-ity." In Houston, this rate rose to 73.8% (48/65) before falling back down to 55.7% (49/88) in New York. The median values of PC 1 in starts for these teams were 0.287, -0.880, and -0.261 respectively. This suggests that Cole's best years in terms of what PCA deems as good Gerrit Cole were in Houston.

We can also look at median PC values for each year of Gerrit Cole's career. Cole's best year by PC 1 was 2019, also the year he was the runner-up for the AL Cy Young Award.

Table 2: *Median PC Values by Year*

Year	PC 1	PC 2	PC 3
2013	0.674	0.323	1.243
2014	0.285	-0.225	0.415
2015	-0.021	**-0.687**	0.710
2016	1.218	0.319	1.163
2017	0.768	**-0.532**	0.081
2018	**-0.659**	0.044	**-0.558**
2019	**-1.042**	0.181	**-0.914**
2020	-0.211	0.187	**-0.533**
2021	-0.240	0.260	-0.430
2022	-0.115	0.347	**-0.564**
2021	-0.026	0.209	0.260

In Table 2, we can also see a change in Gerrit Cole's pitching approach and outcomes over the years. In Pittsburgh, Cole's starts had high PC 2 values, meaning he was going deep into games, but giving up a lot of hits. In Houston and New York, Cole became a pitcher who stuck out a lot of batters but also gave up a lot of home runs.

Conclusions

Principal Component Analysis is a useful tool in statistics for reducing the dimensionality of a data set. In the case of baseball, there is an abundance of statistics that capture different components of a player's performance in a game or season. Historically, arbitrary lines have been drawn to separate "good" from "bad," with the case of a quality start being one example.

Using PCA, we are able to find correlations between statistics recorded during a start and use this to explain the relationship between data points. In the case of Gerrit Cole, the first principal component can be used to identify starts characterized by a high number of innings pitched and strikeouts and a low number of hits and runs allowed. These starts can be deemed successful with our understanding of the game of baseball, specifically for Gerrit Cole, a strikeout pitcher.

It is interesting to see that even though PCA has no knowledge of what is "good" and "bad" in baseball, it can identify some of the best and worst starts of Cole's career by principal component value. Hopefully, the lowest PC 1 values are yet to come for Gerrit Cole and the New York Yankees.

References

[1] Saniya Parveez and Roberto Iriondo. *Principal component analysis (PCA) with python examples-tutorial*. Apr. 2021. URL: https://pub.towardsai.net/principal-component-analysis-pca-with-python-examples-tutorial-67a917bae9aa.

[2] Serafeim Loukas. *PCA clearly explained-how, when, why to use it and feature importance: A guide in python*. May 2023. URL: https://towardsdatascience.com/pca-clearly-explained-how-when-why-to-use-it-and-feature-importance-a-guide-in-python-7c274582c37e.

[3] Ajitesh Kumar. *PCA explained variance concepts with python example*. Apr. 2023. URL: https://vitalflux.com/pca-explained-variance-concept-python-example/.

[4] Jean-Christophe Chouinard. *PCA in Scikit-learn – Principal Component Analysis (with Python Example)*. June 2023. URL: https://www.jcchouinard.com/pca-with-python/#PCA_Examples_From_This_Tutorial.