2024 MLB Pitching Analysis

STAT 437 Final Project

By: RJ Burjek

Fall 2024

1. Introduction
2. Dataset Discussion
3. Clusterability
4. Clustering Algorithms
5. Conclusion

Introduction / Motivation

- In Major League Baseball, the value of pitching is at an all-time high
- Three Types of Pitches: Fastballs, Breaking Balls, Offspeed Pitches
- Goal: Use 2024 MLB regular season pitching statistics to cluster pitchers and learn about pitch type usage and its effectiveness
    - Data restricted to pitchers who recorded 250 plate appearances or more

Dataset

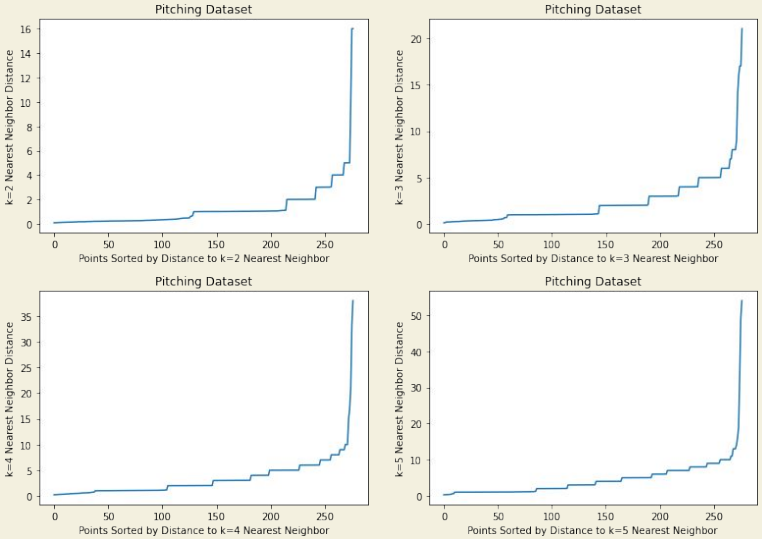
- 2024MLB_AdvancedPitchStats
  - Baseball Savant
- 277 Observations
- 12 Attributes
  - 11 Numerical
  - 1 Categorical
- Variables:
  - PA, k_percentage, bb_percentage, wOBA, barrel_batted_rate, hard_hit_percent, whiff_percent, swing_percent, pitch_hand, n_fastball, n_breaking, n_offspeed

2024MLB_AdvancedPitchStats

Name	PA	k_percent	bb_percent	woba	barrel_batted_rate	hard_hit_percent	whiff_percent	swing_percent	pitch_hand	n_fastball	n_breaking	n_offspeed
Verlander, Justin	396	18.7	6.8	0.337	6.9	32.5	21.3	48.7	R	48.7	41.2	10.1
Chavez, Jesse	264	20.8	7.2	0.315	6.3	35.4	17	42.3	R	78	8.6	13.4
Morton, Charlie	701	23.8	9.3	0.335	9.3	38.4	26.1	46.6	R	46.6	42.4	11
Lynn, Lance	511	21.3	8.6	0.314	10.5	39.3	24.1	48.4	R	86.2	7.5	6.3
Carrasco, Carlos	447	19.9	7.4	0.360	8.1	38.8	22.7	48.7	R	42.3	34.4	23.3
García, Luis	254	20.9	5.9	0.314	7.1	39.3	24	49.7	R	60.2	21.2	18.6
Quintana, Jose	717	18.8	8.8	0.312	6.7	38.2	22.2	43.1	L	52.9	27.8	19.3
Gibson, Kyle	722	20.9	9.4	0.323	9.2	39.1	25.9	42.9	R	60.7	29.3	9.9
Robertson, David	297	33.3	9.4	0.257	5.5	37.8	29.5	43.4	R	64.8	33.4	1.7
Anderson, Chase	255	16.5	8.2	0.329	8.5	33	23.8	46.4	R	50.4	19.2	30.4
Darvish, Yu	331	23.6	6.6	0.286	7.5	39.9	25.6	51.5	R	40.2	50.7	9
Cruz, Fernando	288	37.8	12.2	0.307	9.8	44.1	38.2	46.7	R	58.1	0	41.9
Kelly, Merrill	300	21	6.3	0.311	10.7	40.9	21.3	48.2	R	59.5	19.6	20.9
Sale, Chris	702	32.1	5.6	0.260	5.6	31.2	31	49	L	45	40.3	14.7
Strickland, Hunter	294	19.4	8.2	0.286	9	35.5	22.8	47	R	72.3	21.6	6.1
Pérez, Martín	590	18.1	8.3	0.351	8.6	41.5	21	46	L	63.2	14.1	22.7
Anderson, Tyler	765	18.6	9.5	0.308	7.9	33	27.8	49.3	L	61.5	1.4	37.1
Armstrong, Shawn	295	22.4	8.5	0.345	8.9	38.1	22.9	50.9	R	91.8	8.2	0
Cole, Gerrit	390	25.4	7.4	0.286	7.4	39.5	24.8	48.3	R	62.2	33.8	4
Eovaldi, Nathan	696	23.9	6	0.291	7.7	42.6	26.8	52.2	R	52.2	17	30.8
Gray, Sonny	671	30.3	5.8	0.292	8.9	37.9	29.8	48.7	R	59.5	34.1	6.4
Hendricks, Kyle	567	15.3	7.6	0.339	7.4	33.4	18.9	46.2	R	48.2	13.1	38.7
Hudson, Daniel	253	24.9	7.5	0.274	12.4	49.4	32.3	50.6	R	55.4	40.9	3.7
Tonkin, Michael	340	25	8.8	0.296	7.4	35	23.7	47.2	R	54.9	45.1	0
Suárez, Albert	565	19.1	7.6	0.320	7.1	38.4	24.6	50.4	R	69.5	13.3	17.2
Lorenzen, Michael	535	18.1	11	0.303	8	39.4	22.8	44.7	R	58.4	22.8	18.8
Chapman, Aroldis	264	37.1	14.4	0.289	7.1	41.3	32.3	44.5	L	61.4	26.6	12
Stripling, Ross	381	12.9	5.8	0.344	6.8	35.3	16.9	47.9	R	60.4	22	17.6
Kittredge, Andrew	287	23.3	7	0.289	8.5	39.5	27.8	53.5	R	50.6	49.4	0
Wheeler, Zack	787	28.5	6.6	0.256	8.3	33.6	27.5	51.2	R	70	22.6	7.3

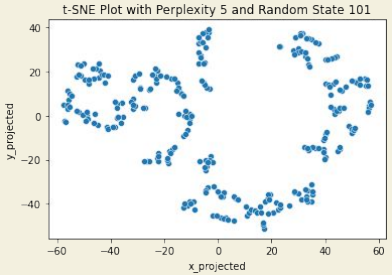
Cleaning / Exploration

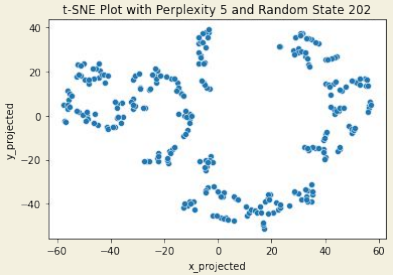
- No N/A values
- Mild evidence of outliers present in the data ->
- Little to no evidence of noise
- Summary statistics of each of the numerical variables, choose to scale “PA” only

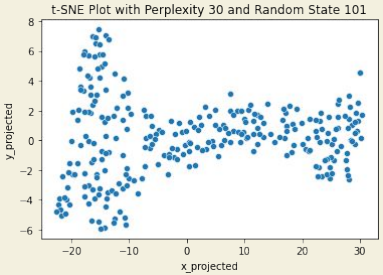


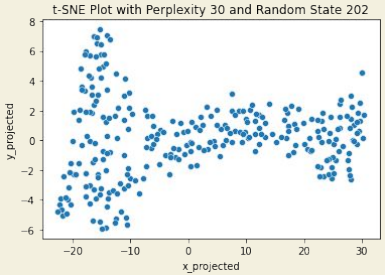
	PA	k_percent	bb_percent	woba	barrel_batted_rate	hard_hit_percent	whiff_percent	swing_percent	n_fastball	n_breaking	n_offspeed
count	277.000000	277.000000	277.000000	277.000000	277.000000	277.000000	277.000000	277.000000	277.000000	277.000000	277.000000
mean	456.086643	0.229874	0.079130	0.304040	0.077101	0.386043	0.253863	0.482603	0.566354	0.296404	0.137242
std	184.758478	0.050908	0.023606	0.036214	0.020137	0.041868	0.047537	0.027626	0.115366	0.126534	0.101878
min	250.000000	0.094000	0.021000	0.175000	0.017000	0.270000	0.128000	0.421000	0.313000	0.000000	0.000000
25%	285.000000	0.195000	0.062000	0.282000	0.065000	0.361000	0.220000	0.463000	0.484000	0.212000	0.054000
50%	390.000000	0.224000	0.077000	0.305000	0.077000	0.388000	0.248000	0.483000	0.556000	0.302000	0.126000
75%	632.000000	0.263000	0.094000	0.325000	0.089000	0.415000	0.286000	0.500000	0.633000	0.384000	0.205000
max	841.000000	0.378000	0.144000	0.413000	0.137000	0.500000	0.405000	0.563000	0.997000	0.639000	0.536000

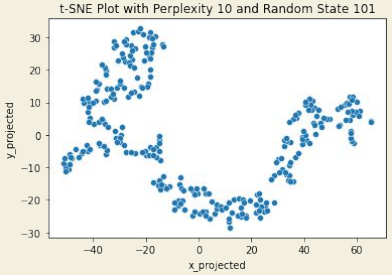
Clusterability

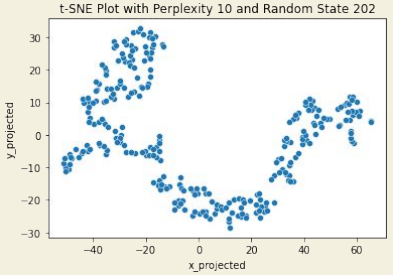


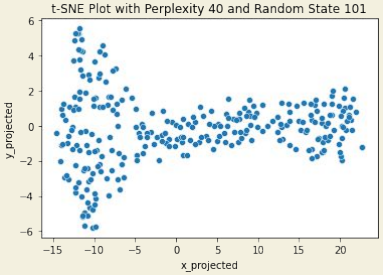


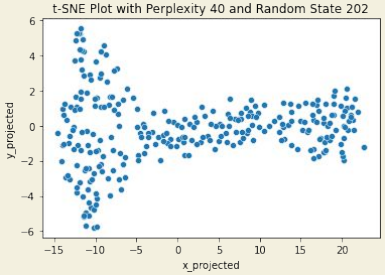


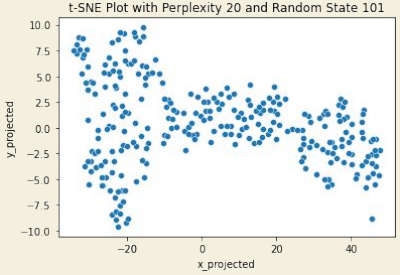


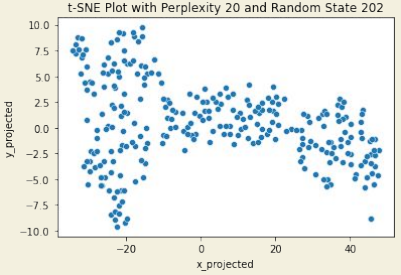


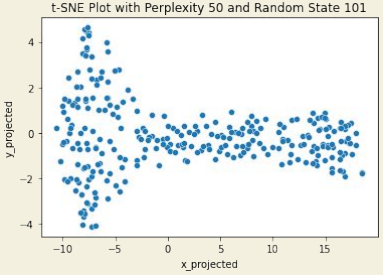


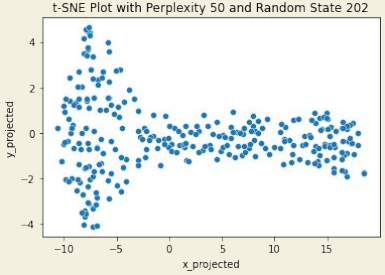












Algorithm Selection

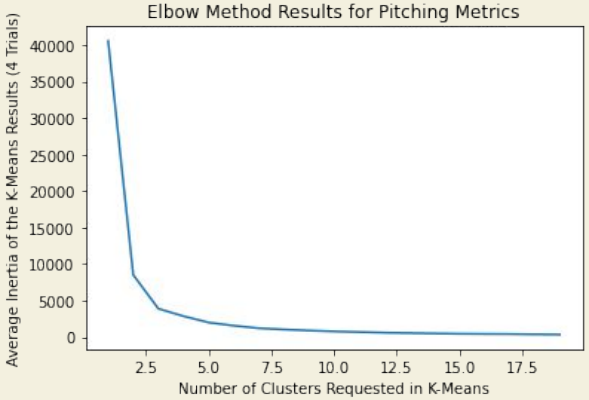
K-Means

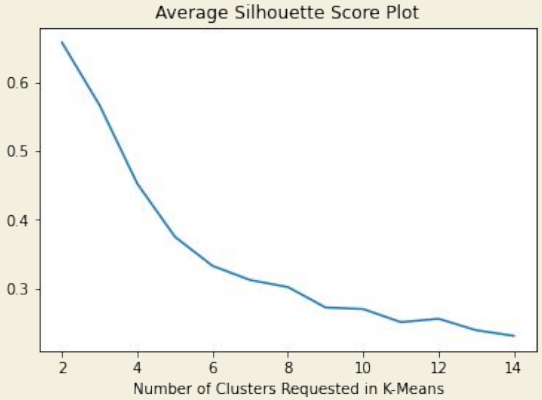
1. Ideal dataset for k-means clustering features clusters that are uniform in size, spherical in shape, non-convex, highly cohesive, and well-separated
2. No significant outliers or noise that would skew clustering results
3. Goal of this study was to find groupings of MLB pitchers from the 2024 regular season and see if there were any disparities in numerical statistics and pitching arsenals

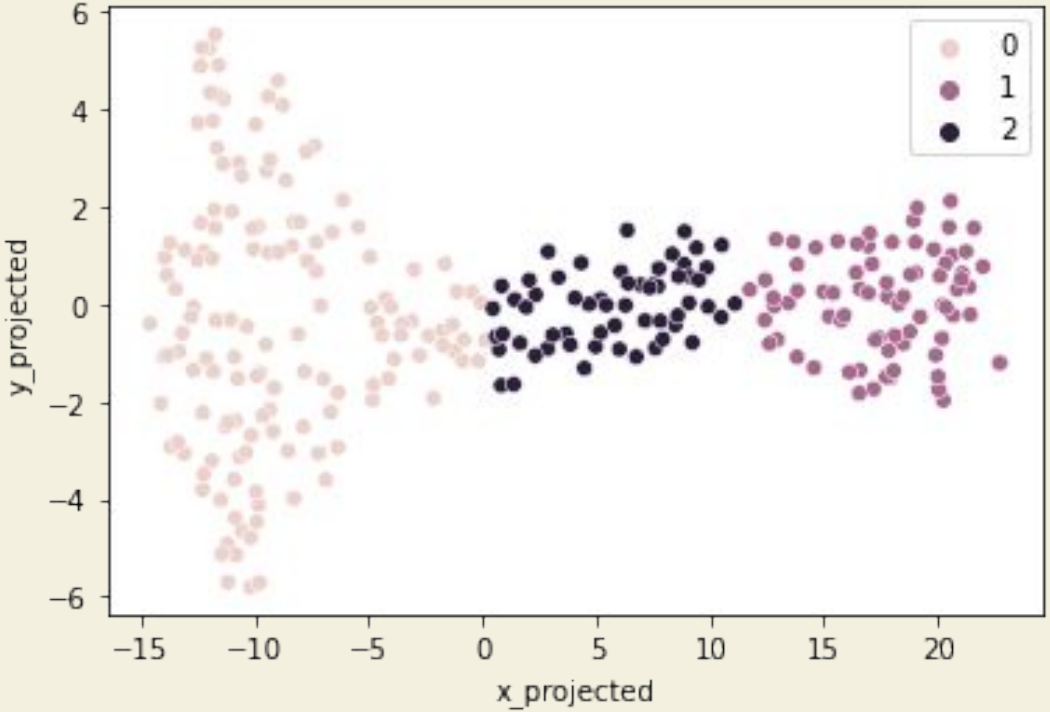
Fuzzy C-Means

1. Fuzzy c-means is effective for addressing cluster structures with a lack of separation
2. Fuzzy c-means will give each data point a cluster membership score, providing additional information about each observation
3. Analyzing the observations with the strongest membership scores will allow us to verify any conclusions drawn from the k-means clustering

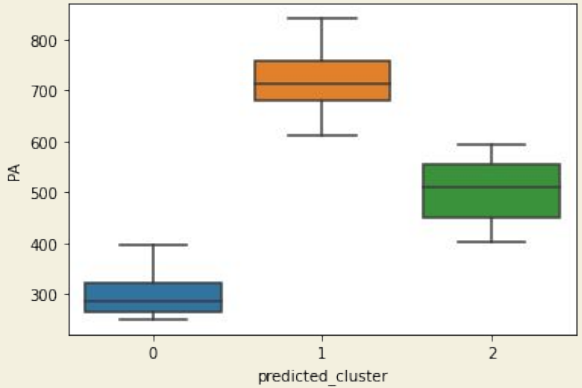
K-Means

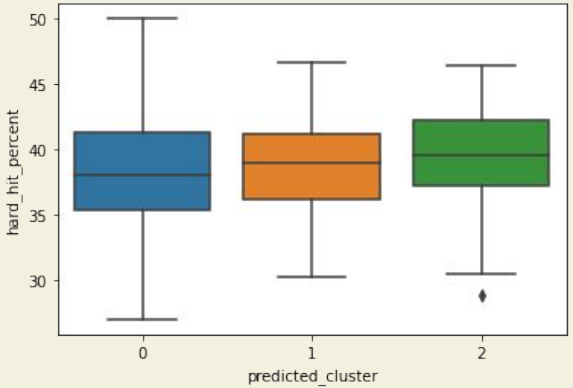

Elbow Method Results for Pitching Metrics
Average Inertia of the K-Means Results (4 Trials)
40000
35000
30000
25000
20000
15000
10000
5000
0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Number of Clusters Requested in K-Means

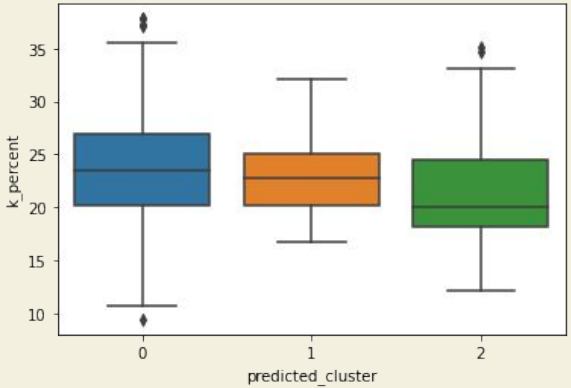

Average Silhouette Score Plot
0.6
0.5
0.4
0.3
2
4
6
8
10
12
14
Number of Clusters Requested in K-Means

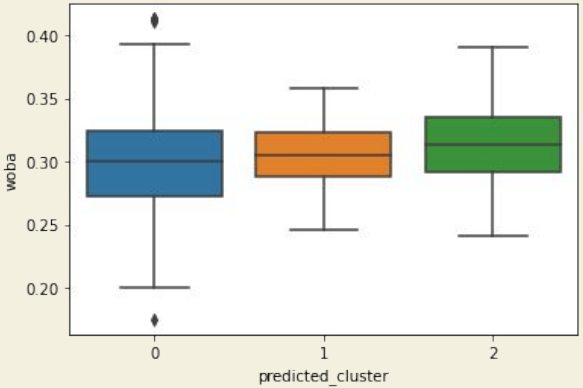

y_projected
6
4
2
0
−2
−4
−6
−15
−10
−5
0
5
10
15
20
x_projected
0
1
2

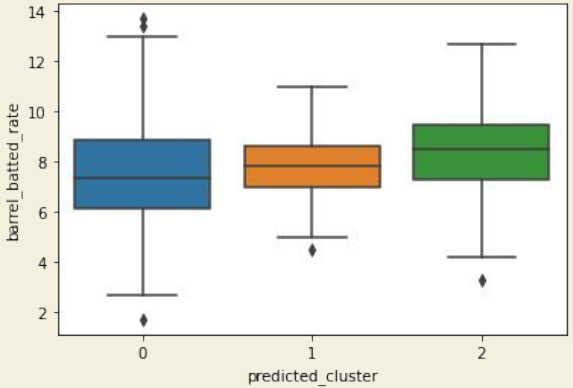
K-Means Visualization

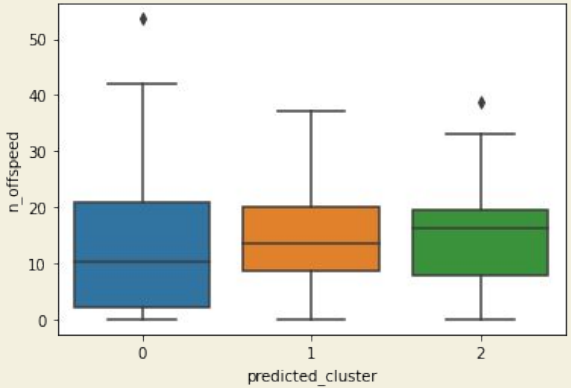






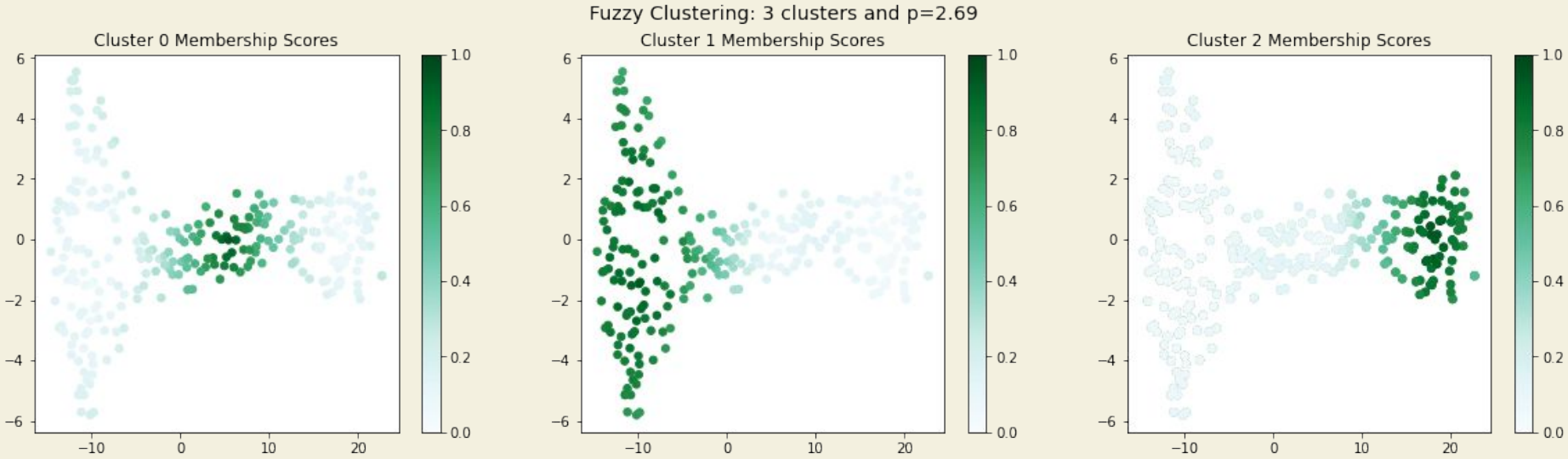




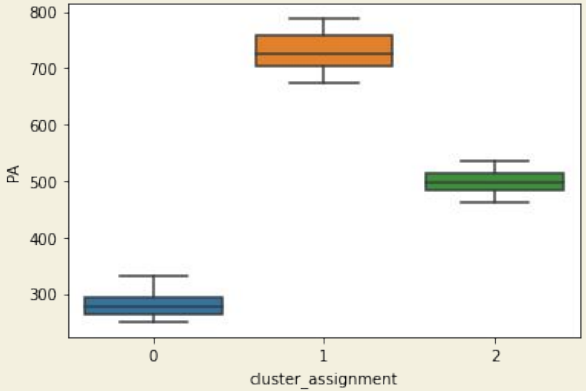


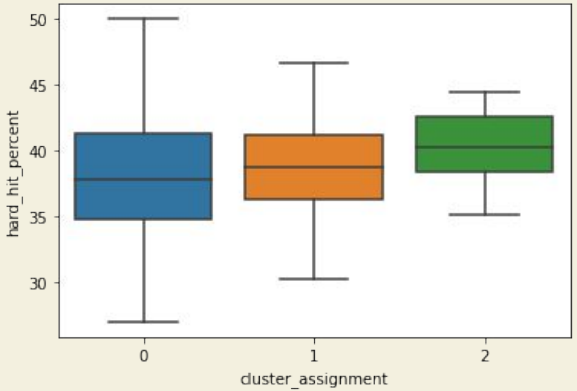
Fuzzy C-Means

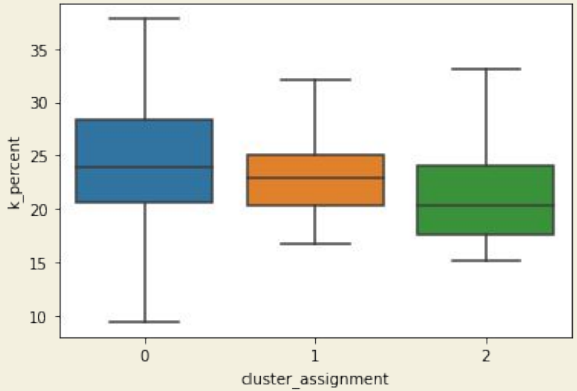
	0	1	2	3	4	5	6	7	8	9	10	11	12	13
p	1.010000	1.220000	1.430000	1.640000	1.850000	2.060000	2.270000	2.480000	2.69000	2.900000	3.110000	3.320000	3.530000	3.740000
ndpc	0.999994	0.977728	0.945585	0.902287	0.845576	0.782184	0.719668	0.662728	0.61327	0.571441	0.536545	0.507605	0.483633	0.463747

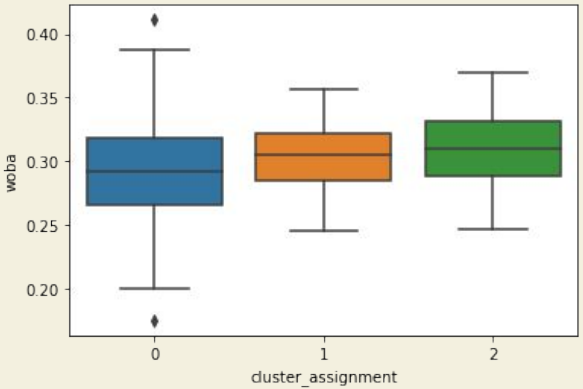


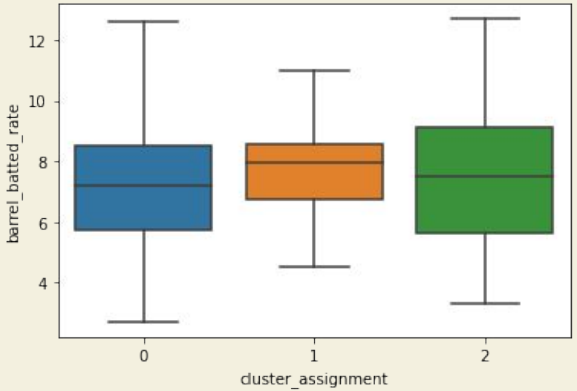
Fuzzy C-Means Visualization

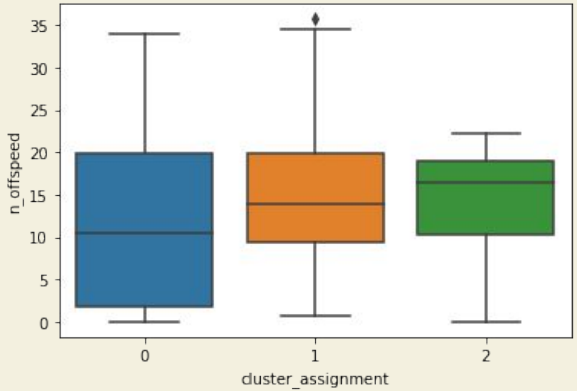












Conclusion

Takeaways:

- In conclusion, there is evidence to suggest a lack of effectiveness of offspeed pitches in Major League Baseball in 2024.
- Evidence to suggest batters were able to make better and harder contact against pitchers that utilized offspeed pitches in their arsenal more

Shortcomings:

- The size of the dataset along with the nature of the observations
- 277 observations is still relatively small, even smaller when straddle points were not accounted for in fuzzy c-means
- Not a predictive model, further supervised learning methods would need to be used to implement this idea in a real-life pitching scenario

References

- Baseball Savant. (n.d.). MLB Advanced Media, LP. https://baseballsavant.mlb.com/.

- Lebovitch, J. (2023). *Baseball Pitches - A Comprehensive Guide*. RPP Baseball. https://rocklandpeakperformance.com/baseball-pitches-a-comprehensive-guide/.

- Madison, B. (2024). *In Baseball, Just How Important is Pitching?*. The Shepherdstown Chronicle. https://www.shepherdstownchronicle.com/sports/2024/02/02/in-baseball-just-how-important-is-pitching/#:~:text=Many%20baseball%20experts%20from%20Major,the%20ball%20toward%20the%20batter.

- Mahlke, A. (2023). "Economics, Baseball, and the Value of Pitching". Twins Daily News. https://twinsdaily.com/news-rumors/minnesota-twins/economics-baseball-and-the-value-of-pitching-r13581/.