Using K-Means Clustering for MLB Hitters
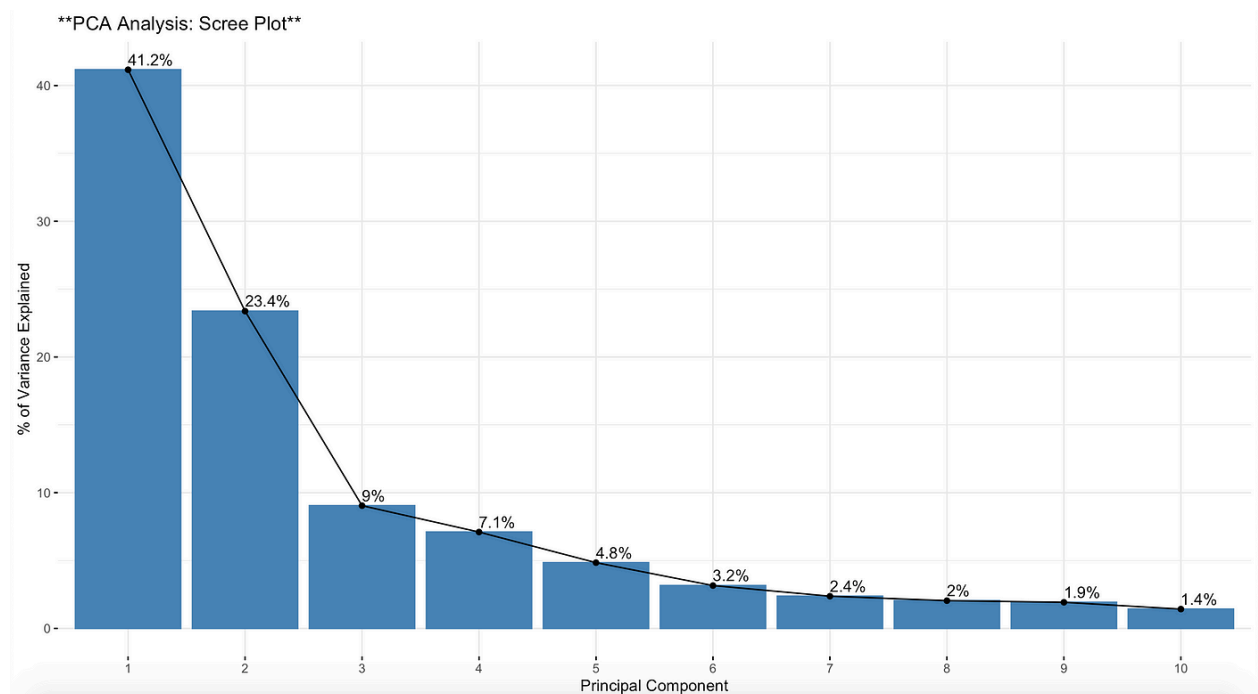

Aaron Judge

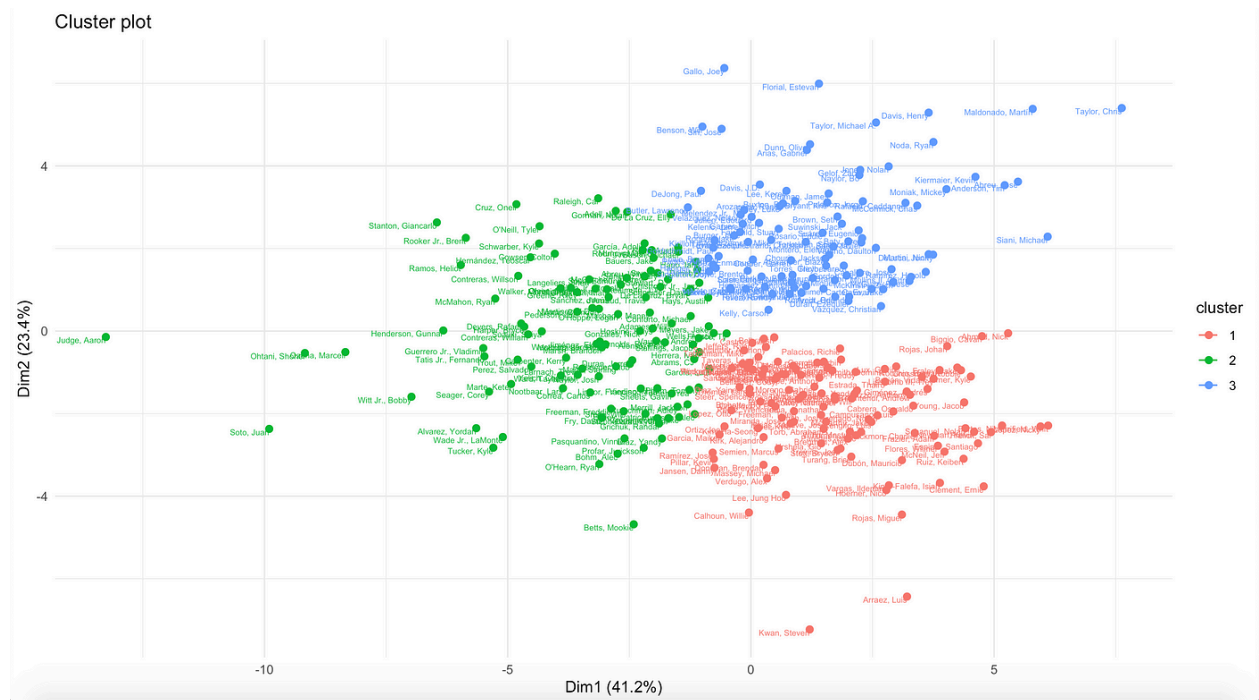**What is K-Means Clustering and Introduction?**

Clustering is a machine learning technique that divides objects into groups, or "clusters", that share similarities. It is beneficial to see how productive hitters are, and how they are grouped.

To complete the clustering I decided to use OPS as the only results-based stat, and the rest were more aligned for "performance-based" statistics to see how hitters were clustered and potentially given a preview of hitters who could turn out to have a pretty good year at the plate. Some of those statistics were:

K% and BB%, xBA, xSLG, and xISO, wOBA and xwOBA, blasts per swing rate, Squared-up Swing%, average Exit Velocity, Sweet Spot%, Chase%, In-Zone Contact %, and a few other statcast and batted ball metrics. Data is oriented from Baseball Savant as of June 13th, 2024 and hitters need at least 100 Plate Appearances to qualify.

The image above is a "Scree Plot". This image is imperative to recognize the amount of clusters you should designate. The optimal amount of clusters is at the bend or "elbow" of the plot, so for this exercise there will be three clusters.

Cluster plot

Here we can see the three clusters taking place being seperated by their three colors. We can easily see players grouped together and it passes the visual validation test. Juan Soto is sandwiched between Aaron Judge, and Shohei Otani. Joey Gallo is beside light hitters, Estevan Florial and Martin Maldonado. Contact-oriented hitter, Steven Kwan is grouped with Luis Arraez and Nico Hoerner.

Cody Bellinger

Cluster 1

Looking through the list of players in this cluster, it is clear that they are

"contact-oriented" hitters. There were 117 hitters selected into Cluster 1, and the average

K% is 16.9%. These hitters all have high In-Zone Contact% and Squared-up swing rate,

while also registering low Whiff%. This cluster has a good mix of quality and

below-average players. Some players listed within this cluster are Steven Kwan, Jose

Ramirez, Jose Altuve, Andrew Benintendi, Jeff McNeil, and Cavan Biggio. There are

some surprises in this cluster, and not the good surprises. Players can easily be valuable

hitters in this cluster due to this elite skill set, look no further than a couple of the players

that I mentioned above. However, players like Anthony Rizzo and Nolan Schanuel being

on here is not a recipe for success. Rizzo and Schanuel are first baseman, destined to be

in Cluster 2 hoping to be power hitting, successful first baseman. Rizzo and Schanuel are on opposite ends of the age spectrum, but, obviously, they share similar qualities in being contact-oriented, light hitters. I'm not sure which hitter I would buy stock into upgrading into Cluster 2. Cody Bellinger appears in this cluster, yet if I did this in 2019, he probably would show up in Cluster 2 with some of the other hard-hitters, demonstrating that you can have success from different skills within the game. Players listed in Cluster 1:

1 Kepler, Max

2 Jansen, Danny

3 Hamilton, David

4 Bruján, Vidal

5 Bellinger, Cody

6 Vargas, Ildemaro

7 Díaz, Elias

8 Estrada, Thairo

9 Albies, Ozzie

10 Frazier, Adam

11 Carroll, Corbin

12 Arenado, Nolan

13 Palacios, Richie

14 Lee, Jung Hoo

15 Trevino, Jose

16 Keith, Colt

17 Bichette, Bo

18 Steer, Spencer

19 McNeil, Jeff

20 Toro, Abraham

21 Jeffers, Ryan

22 Semien, Marcus

23 Clement, Ernie

24 Frelick, Sal

25 Rojas, Johan

26 Miranda, Jose

27 Olivares, Edward

28 Crawford, J.P.

29 Biggio, Cavan

30 Cabrera, Oswaldo

31 Lopez, Otto

32 Castro, Willi

33 Hoerner, Nico

34 Lopez, Nicky

35 Rizzo, Anthony

36 Joe, Connor

37 Ahmed, Nick

38 Mateo, Jorge

39 Altuve, Jose

40 Ortiz, Joey

41 Campusano, Luis

42 Freeman, Tyler

43 Peña, Jeremy

44 Bregman, Alex

45 Winn, Masyn

46 Farmer, Kyle

47 Santana, Carlos

48 Brennan, Will

49 Margot, Manuel

50 Merrifield, Whit

51 Hayes, Ke'Bryan

52 Young, Jacob

53 Urshela, Gio

54 Diaz, Yainer

55 Paredes, Isaac

56 Smith, Josh

57 Canha, Mark

58 Fraley, Jake

59 Heim, Jonah

60 Garcia, Maikel

61 Lux, Gavin

62 Newman, Kevin

63 Kiner-Falefa, Isiah

64 Espinal, Santiago

65 Taylor, Tyrone

66 Turang, Brice

67 Winker, Jesse

68 Verdugo, Alex

69 Donovan, Brendan

70 Thomas, Lane

71 Ruiz, Keibert

72 Kim, Ha-Seong

73 McCarthy, Jake

74 Kirk, Alejandro

75 Taveras, Leody

76 Bader, Harrison

77 Flores, Wilmer

78 Massey, Michael

79 Gurriel Jr., Lourdes

80 Pillar, Kevin

81 India, Jonathan

82 Dubón, Mauricio

83 Bleday, JJ

84 Rocchio, Brayan

85 Wong, Connor

86 Giménez, Andrés

87 Meneses, Joey

88 Lipscomb III, Trey

89 Pérez, Wenceel

90 Bell, Josh

91 Volpe, Anthony

92 Senzel, Nick

93 Schanuel, Nolan

94 Moreno, Gabriel

95 Grossman, Robbie

96 Santander, Anthony

97 Fermin, Freddy

98 Stott, Bryson

99 Blackmon, Charlie

100 Tauchman, Mike

101 Benintendi, Andrew

102 Ramírez, José

103 Rodgers, Brendan

104 Springer III, George

105 Rengifo, Luis

106 Rojas, Josh

107 Arraez, Luis

108 Lowe, Nathaniel

109 Isbel, Kyle

110 Smith, Dominic

111 Rosario, Amed

112 Fortes, Nick

113 Bogaerts, Xander

114 Calhoun, Willie

115 Turner, Justin

116 Kwan, Steven

117 Rojas, Miguel

Austin Wells

Cluster 2

Cluster 2 has more of the successful hitters, hitters who have the highest Barrel%, Hard-Hit%, bat speed, BB%, and some other quality of contact statistics. I'd argue that this is the cluster you would want to be in. The average OPS of players listed here is .797, 100 points higher than Cluster 1, and about 200 points higher than Cluster 3. Some players listed here are Kyle Tucker, Gunnar Henderson, Bobby Witt Jr., Andrew Vaughn, Gary Sanchez, and Jo Adell. There are some pleasant surprises in this cluster. I'm a fan that Austin Wells is showcased in this cluster. Although he has the lowest OPS of the 115 players listed, he has a higher Sweet Spot% than J.T Realmuto, lower Chase% than Pete Alonso, and a higher Squared-up Swing% than Vladimir Guerrero Jr. If there is a player I

could buy stock into, it would be the lefty-hitting platoon catcher from the Yankees.

Players listed in Cluster 2:

1 d'Arnaud, Travis

2 Abreu, Wilyer

3 Rooker Jr., Brent

4 Acuña Jr., Ronald

5 Fry, David

6 Stephenson, Tyler

7 Chisholm Jr., Jazz

8 Alvarez, Yordan

9 McMahon, Ryan

10 Contreras, William

11 Pham, Tommy

12 Freeman, Freddie

13 García, Luis

14 Happ, Ian

15 Harper, Bryce

16 McCutchen, Andrew

17 Martinez, J.D.

18 Ohtani, Shohei

19 Soto, Juan

20 Stallings, Jacob

21 Guerrero Jr., Vladimir

22 Riley, Austin

23 Swanson, Dansby

24 France, Ty

25 Judge, Aaron

26 Bailey, Patrick

27 Gonzales, Nick

28 Seager, Corey

29 Cronenworth, Jake

30 Carpenter, Kerry

31 Contreras, Willson

32 Larnach, Trevor

33 Perez, Salvador

34 Sheets, Gavin

35 Rodríguez, Julio

36 Turner, Trea

37 Westburg, Jordan

38 Rutschman, Adley

39 Correa, Carlos

40 Nimmo, Brandon

41 Sánchez, Gary

42 Schneider, Davis

43 Marsh, Brandon

44 Raleigh, Cal

45 Sánchez, Jesús

46 Ramos, Heliot

47 Conforto, Michael

48 O'Neill, Tyler

49 Tatis Jr., Fernando

50 Herrera, Iván

51 Morel, Christopher

52 Schwarber, Kyle

53 Betts, Mookie

54 Caratini, Victor

55 Walker, Christian

56 Bohm, Alec

57 Vierling, Matt

58 Henderson, Gunnar

59 García, Adolis

60 Chapman, Matt

61 Merrill, Jackson

62 Pederson, Joc

63 Hays, Austin

64 Marte, Starling

65 Pasquantino, Vinnie

66 Jiménez, Eloy

67 Machado, Manny

68 Meyers, Jake

69 Cowser, Colton

70 Hoskins, Rhys

71 Witt Jr., Bobby

72 Wells, Austin

73 Muncy, Max

74 Devers, Rafael

75 Olson, Matt

76 Bauers, Jake

77 Ozuna, Marcell

78 Trout, Mike

79 Refsnyder, Rob

80 Yelich, Christian

81 Profar, Jurickson

82 De La Cruz, Elly

83 Cruz, Oneil

84 Suzuki, Seiya

85 Adell, Jo

86 Realmuto, J.T.

87 Gorman, Nolan

88 Smith, Will

89 Duran, Jarren

90 Greene, Riley

91 Reynolds, Bryan

92 Burleson, Alec

93 Ward, Taylor

94 Marte, Ketel

95 O'Hearn, Ryan

96 O'Hoppe, Logan

97 Wade Jr., LaMonte

98 Langeliers, Shea

99 Díaz, Yandy

100 Mountcastle, Ryan

101 Vaughn, Andrew

102 De La Cruz, Bryan

103 Nootbaar, Lars

104 Hernández, Teoscar

105 Sosa, Edmundo

106 Stewart, DJ

107 Busch, Michael

108 Adames, Willy

109 Lindor, Francisco

110 Grichuk, Randal

111 Abrams, CJ

112 Alonso, Pete

113 Tucker, Kyle

114 Naylor, Josh

115 Stanton, Giancarlo

Cluster 3

Of the three clusters, this is probably the cluster you do not want to be in. Cluster 3 has the highest K%, Whiff%, and Chase%, while also ranking below average in most quality of contact metrics. The average OPS of the 97 players listed in Cluster 3 is .600. So… not that good. There are quality hitters here, they just might be overperforming. Some players listed here are Ezequiel Tovar, Daulton Varsho, Andy Pages, Chris Taylor, Tim Anderson and Javier Baez. Unfortunately, there are some names that are in this cluster that are a displeasing revelation. After a 36 home run season, rewarded with a 42 million dollar contract, the San Francisco Giants might be disappointed to see Jorge Soler here. Paul Goldschmidt and Gleyber Torres are both set to become free agents at the end of this year, so we will see how much money they potentially have costed them from this season. There are a lot of rookies in this list, Brett Baty, Jackson Chourio, Evan Carter, Andy Pages, Wyatt Langford to just name a few, so I may want to update this at the end of the season to see how, or if they have progressed. Players listed in Cluster 3:

1 Caballero, José

2 Varsho, Daulton

3 Ramírez, Harold

4 Davis, J.D.

5 Taylor, Michael A.

6 Tovar, Ezequiel

7 Bryant, Kris

8 DeJong, Paul

9 Torkelson, Spencer

10 Raley, Luke

11 Valdez, Enmanuel

12 Drury, Brandon

13 Kiermaier, Kevin

14 Duran, Ezequiel

15 Noda, Ryan

16 Dunn, Oliver

17 Cooper, Garrett

18 McGuire, Reese

19 Kelenic, Jarred

20 Arozarena, Randy

21 Pages, Andy

22 Moore, Dylan

23 Perkins, Blake

24 Baty, Brett

25 Mullins II, Cedric

26 Candelario, Jeimer

27 Rogers, Jake

28 Gordon, Nick

29 Garver, Mitch

30 Kelly, Carson

31 Rortvedt, Ben

32 Gallo, Joey

33 Moniak, Mickey

34 Polanco, Jorge

35 Vázquez, Christian

36 Singleton, Jon

37 Velázquez, Nelson

38 Fairchild, Stuart

39 Torres, Gleyber

40 Montero, Elehuris

41 Maldonado, Martín

42 Burger, Jake

43 Naylor, Bo

44 Benson, Will

45 Arcia, Orlando

46 Anderson, Tim

47 Brown, Seth

48 Schuemann, Max

49 Siri, Jose

50 Harris II, Michael

51 McCann, James

52 Neto, Zach

53 Tellez, Rowdy

54 Renfroe, Hunter

55 Buxton, Byron

56 Arias, Gabriel

57 Florial, Estevan

58 Carter, Evan

59 Abreu, José

60 Yastrzemski, Mike

61 Triolo, Jared

62 Duvall, Adam

63 Encarnacion-Strand, Christian

64 Kirilloff, Alex

65 Nevin, Tyler

66 Haniger, Mitch

67 Báez, Javier

68 Butler, Lawrence

69 Rosario, Eddie

70 McKinstry, Zach

71 Lowe, Brandon

72 Alexander, Blaze

73 Suárez, Eugenio

74 Hernández, Enrique

75 DeLuca, Jonny

76 Langford, Wyatt

77 Rivera, Emmanuel

78 Lee, Korey

79 Doyle, Brenton

80 Siani, Michael

81 Chourio, Jackson

82 Soler, Jorge

83 Suwinski, Jack

84 Castellanos, Nick

85 Jones, Nolan

86 Goldschmidt, Paul

87 Gelof, Zack

88 Melendez Jr., MJ

89 Cave, Jake

90 Rafaela, Ceddanne

91 Martini, Nick

92 Amaya, Miguel

93 McCormick, Chas

94 Julien, Edouard

95 Outman, James

96 Davis, Henry

97 Taylor, Chris

Data from Baseball Savant

The link to the code and writing pdf can be found on my Github: *[Here](#)*

You can also find this on Twitter/X: @EthanMann02