Universidad Central de Venezuela
Facultad de Ciencias
Introducción a la Ciencia de Datos

Proyecto de Ciencia de Datos
Agrupación de los jugadores de baseball de la MLB en la temporada regular 2016 de acuerdo a su rendimiento como bateadores

Autor: Jorge Flores

Introducción

El baseball es un deporte en el cual toda acción que se realiza queda registrada, esto ayuda a establecer una serie de estadísticas por jugador como el bateador, pitcher y sus respectivas posiciones. A su vez, por equipo quedan registradas las victorias, derrotas y otros tipos de logros, lo que hace al baseball muy interesante, ya que gracias a esta cualidad han surgido nuevas inquietudes para hacer este deporte más efectivo que otros.

Con un registro tan amplio de estadísticas, se trabajará sobre la de los bateadores específicamente en el cual se aplicará unas técnicas de minería de datos las cuales son conocidas como clustering, scraping web y clasificación donde el proceso de clustering es un método de aprendizaje no supervisado el cual consiste en generar grupos que comparten características similares para el conjunto de datos. El proceso de scraping web consiste en extraer todo tipo de información de la web por medio de progaramas de software como R, Python entre otros, lo cual permite hacer un estudio con mayor rigurosidad.

El motivo que inspiró a la realización de este proyecto fue la aplicación de la Ciencia de Datos en el campo del deporte. De acuerdo a lo expuesto anteriormente, en este proyecto se plantea como objetivo agrupar y clasificar a los bateadores de Major League Baseball (MLB) de la temporada regular 2016 mediante la técnica de clustering.

Planteamiento del problema

En la actualidad, los científicos de datos están adquiriendo un roll muy importante en la vida cotidiana, ya que se requiere de sus especialidades para encontrar soluciones a problemas que antes no se habían tratado. Un ejemplo de ello es aplicar las técnicas Web scraping y clustering en el ámbito deportivo, lo cual es importante ya que se puede encontrar nuevos clasificadores para los deportistas según su rendimiento, características y potencial atlético.

Con base en lo anterior puede inferirse que el método de clustering en el baseball será de gran ayuda para el análisis técnico de los jugadores y realizar una mejor estrategia para el funcionamiento del equipo.

Vásquez Fernández, Miguel (2014) en su tesis de grado **"COMBINING CLUSTERING AND TIME SERIES FOR BASEBALL FORECASTING"** a partir de la técnica clustering, profundizó en el tema de la Ciencia de Datos en el deporte. El principal aporte de esta investigación fue lograr predecir de manera favorable la complejidad de esta tarea.

Otro material que fue tomado en cuenta como antecedente para este proyecto es la película **"Moneyball"**, producida por Bennett Miller, Octubre 2011. Esta película ofrece información valiosa para el proyecto debido a que es basada en una historia de la vida real en la que se aplica análisis estadístico en el baseball con el fin de formar un buen equipo con un presupuesto reducido.

Dentro de esta perspectiva, el problema de la investigación puede ser formulado de la siguiente manera:

¿Cómo se pueden agrupar a los jugadores de la MLB de la temporada regular 2016 según su rendimiento?

De esta interrogante se desprenden las siguientes preguntas:

•¿Qué jugadores serán tomados en cuenta para la realización de proyecto?
•¿Cuáles son las características a estudiar?

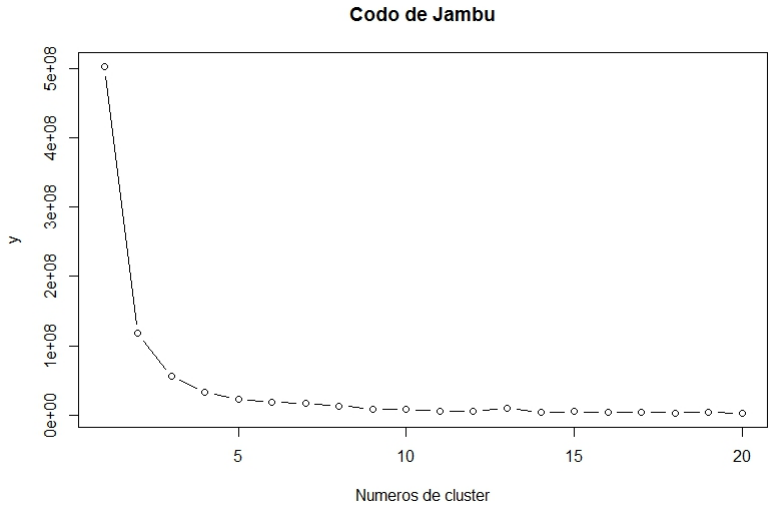
•¿Cómo será la clasificación de los jugadores luego de los métodos aplicados?

Por otra parte, este proyecto busca generar respuestas cómo encontar una agrupación de los bateadores de acuerdo a su rendimiento tomando en cuenta los datos que quedan registrados como: en cuántos juegos participó, cuántos turnos al bate tomó, cuántos hit conectó, cuantas carreras impulsó y otra serie de características que serán visualizadas mejor en la tabla de datos, esto se realizará mediante el lenguaje de programación R en el cual se aplicará la técnica web scraping para obtener los datos de la página oficial de la MLB que contiene toda la lista de los bateadores que participaron en la temporada regular 2016. Luego de dicha extracción se procede al siguiente paso en el cual se prepararán los datos para realizar la investigación.

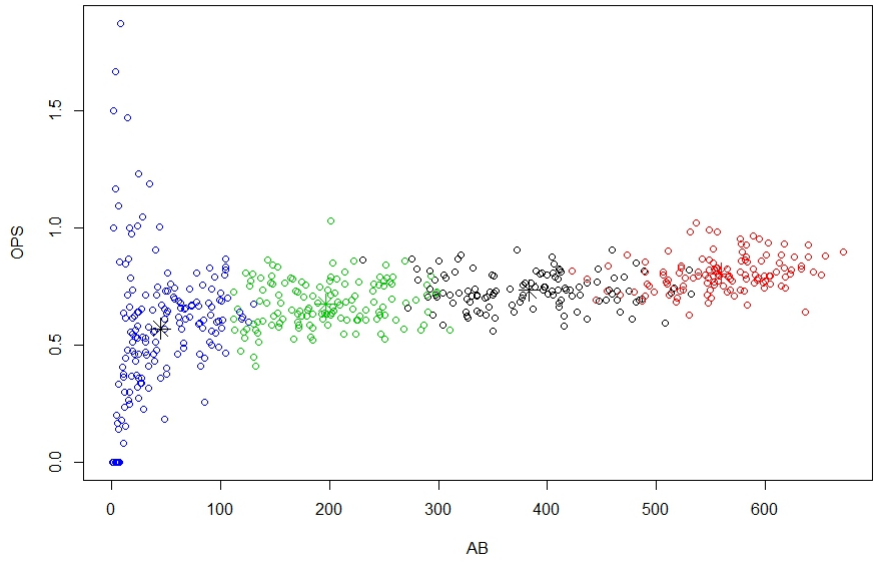
Una vez culminado el procedimiento anterior, se aplicará el algoritmo K-means y así se encontrarán las diferentes agrupaciones de los bateadores lo cual permitirá clasificarlos según su capacidad de rendimiento.

Procedimiento para la elección de los grupos

Para realizar el trabajo de investigación se aplicó el algoritmo K-means, la elección del "k" se hace mediante el codo de Jambu, donde se puede visualizar en la siguiente gráfica que el "k" puede ser 4 ó 5. En esta ocasión se eligió k=4 ya que cuando se dividió a los jugadores en 5 grupos, ese grupo adicional era de 4 individuos los cuales no eran de gran relevancia y solo disminuía el error en menos de 1%.



Esta gráfica permite observar los grupos con las observaciones de "AB" y "OPS"



Datos

Para la realización del proyecto se utilizó la siguiente base de datos (link), en donde se encuentra reflejado el rendimiento de cada bateador de la MLB temporada regular 2016.

A su vez, en este (link) se utilizan características específicas del jugador como: peso, estatura, entre otros.

A continuación se presentan las imágenes de la tabla y del perfil del jugador donde se aprecian las estadísticas que serán utilizadas para la investigación.

ESTADÍSTICAS

Jugador | Bateo | Pitcheo | Fildeo | Equipo | Novatos | Bateadores Vs. Lanzadores | Statcast

2016 | Todos los Tiempos por Año | Totales de Todos los Tiempos | Regular Season | Todos Tiempos | Activo

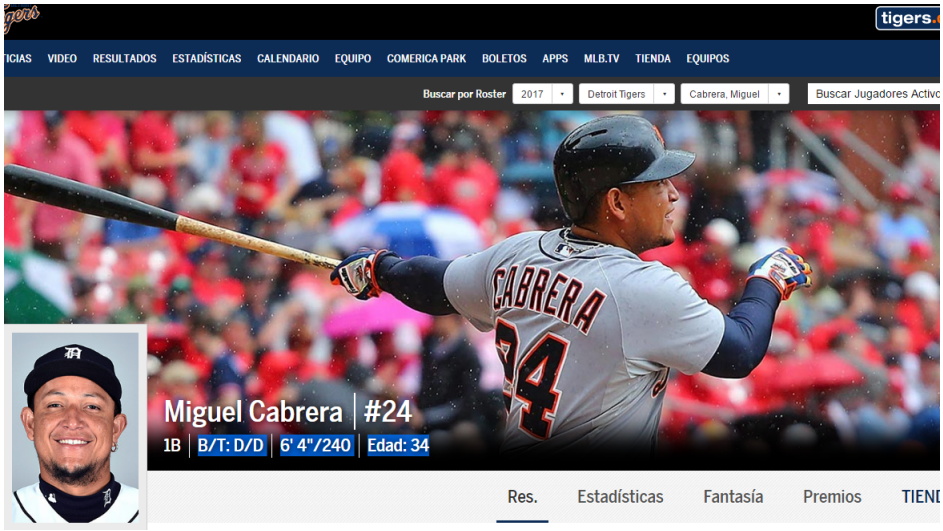
Todos los jugadores | Qualifiers

MLB | LA | LN | All Teams | Todas las posiciones | Select Split

Cronología: Al Día | Antes del All-Star | Despues del All-Star

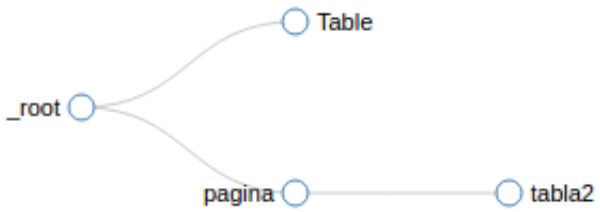
Próx Stats

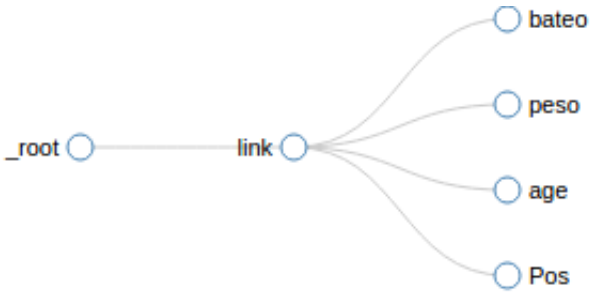
RK	Player	Team	Pos	G	AB	R	H	2B	3B	HR	RBI	BB	SO	SB	CS	AVG	OBP	SLG	OPS
1	Albers, M	CWS	P	4	1	1	1	1	0	0	0	0	0	0	0	1.000	1.000	2.000	3.000
1	Berrios, J	MIN	P	1	1	0	1	0	0	0	0	0	0	0	0	1.000	1.000	1.000	2.000
1	Ege, C	LAA	P	5	1	0	1	0	0	0	0	0	0	0	0	1.000	1.000	1.000	2.000
1	Gearrin, C	SF	P	54	1	0	1	0	0	0	0	0	0	0	0	1.000	1.000	1.000	2.000
1	Milone, T	MIN	P	1	2	0	2	0	0	0	0	0	0	0	0	1.000	1.000	1.000	2.000
6	Perez, O	WSH	P	59	3	1	2	1	0	0	0	0	0	0	0	.667	.667	1.000	1.667
6	Perez, Y	MIA	SS	12	3	5	2	1	0	0	0	0	1	4	2	.667	.667	1.000	1.667
8	Blach, T	SF	P	4	4	1	2	0	0	0	0	0	0	0	0	.500	.500	.500	1.000
8	Davidson, M	CWS	3B	1	2	1	1	0	0	0	1	0	1	0	0	.500	.500	.500	1.000
8	Gallardo, Y	BAL	P	3	4	1	2	0	0	0	0	1	1	0	0	.500	.600	.500	1.100
8	Hicks, J	DET	1B	1	2	1	1	1	0	0	0	0	0	0	0	.500	.500	1.000	1.500
8	Holland, D	TEX	P	1	2	0	1	0	0	0	0	0	1	0	0	.500	.500	.500	1.000
8	Parmelee, C	NYY	1B	6	8	4	4	1	0	2	4	0	3	0	0	.500	.500	1.375	1.875
8	Tilson, C	CWS	CF	1	2	0	1	0	0	0	0	0	0	0	0	.500	.500	.500	1.000
15	Sucre, J	SEA	C	9	25	4	12	2	0	1	5	2	5	0	0	.480	.552	.680	1.232



El método que se aplicó para la extracción de los datos fue la técnica web scraping, la cual se llevó a cabo con la extensión de Google Chrome llamada Scraper totalmente gratuita. Hay otras formas de aplicar dicha técnica, una de ellas es en el lenguaje de programación Python utilizando los paquetes (Scrapy o Beatiful Soup), sin embargo, se eligió la extensión Scraper ya que facilitó la obtención de los datos de manera exitosa.

Se puede apreciar un grafo en el cual se muestra como se obtuvieron los datos.





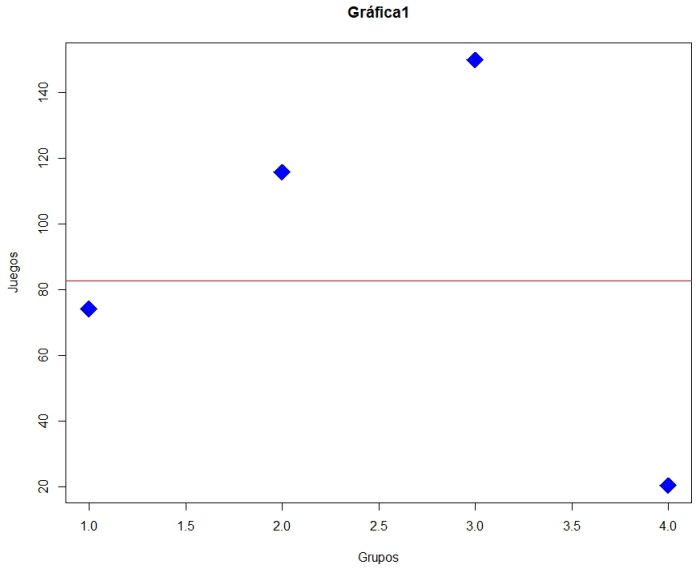
Luego de la obtención de los datos se procede a ejecutar la limpieza, la cual fue realizada en el lenguaje de programación Python utilizando las bibliotecas (Pandas y Numpy). Para dicha limpieza se toman en cuenta las siguientes observaciones. Solo se trabajó sobre los jugadores que no ocupan

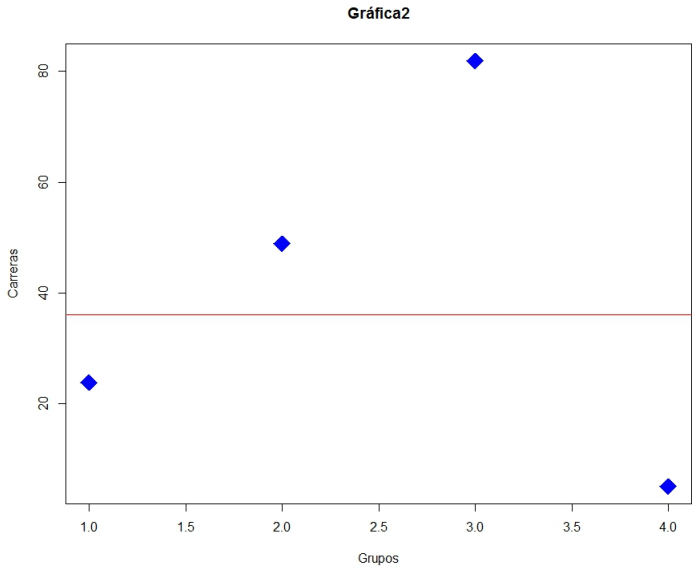
la posición de pitcher, ya que hay una gran diferencia entre la cantidad de turnos al bate del pitcher con respecto a la cantidad de turnos al bate de un jugador que ocupe otra posición. Por otra parte, se encontraron valores faltantes en la variable edad. Este problema se solucionó remplazando dichos valores por la moda, la cual era 26 años, a su vez, las variables peso y estatura estaban representadas en el sistema inglés las cuales se transformaron en medidas del sistema métrico decimal. Para la lateralidad de los bateadores se le asignó: (diestro: 1, zurdo: -1, ambidiestro: 0), además se eliminaron jugadores duplicados.

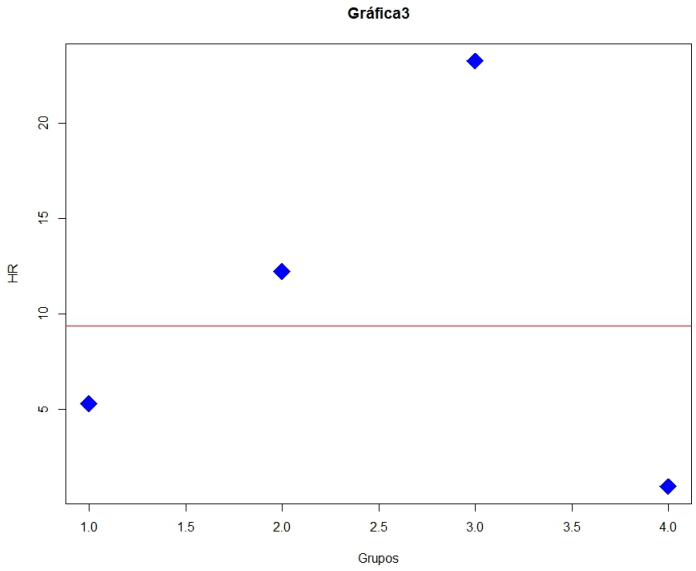
Nota: El código de limpieza se encuentra en el siguiente (link).

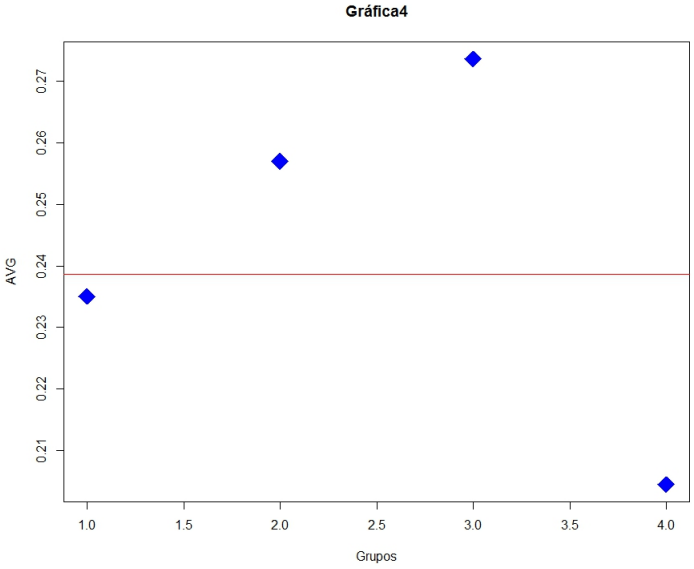
Análisis Exploratorio

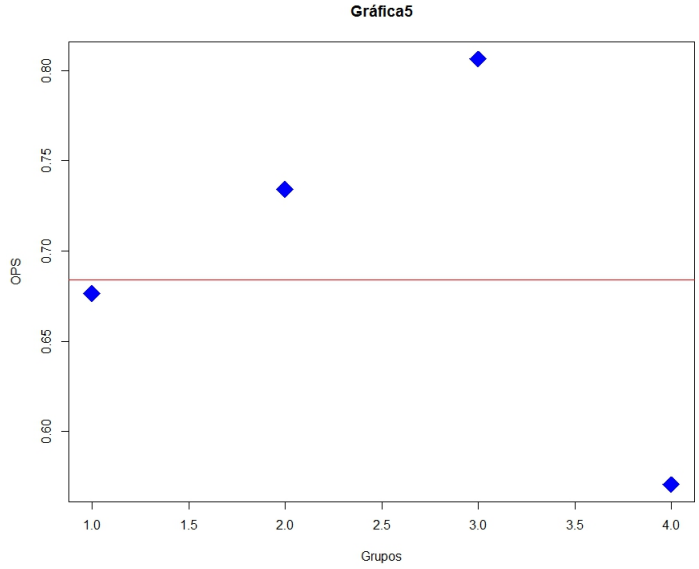
Luego de emplear la técnica de clustering se obtuvo un total de 4 grupos. En las siguientes gráficas se puede observar el comportamiento que presenta cada grupo considerando las siguientes variables: (Juegos, Carreras, HR, AVG, OPS), identificando la recta como el promedio de todo el conjunto de datos en su respectiva variable.




Gráfica2
Carreras
20
40
60
80
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Grupos


Gráfica3
HR
5
10
15
20
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Grupos


Gráfica4
AVG
0.21
0.22
0.23
0.24
0.25
0.26
0.27
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Grupos


Gráfica5
OPS
0.60
0.65
0.70
0.75
0.80
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Grupos

Ahora bien, los métodos estadísticos para llevar a cabo la investigación fueron los siguientes:

Análisis de correlación el cual permite observar la relación que existe entre las variables. A su vez, la moda que es el valor que tiene mayor frecuencia absoluta en el conjunto de datos, en esta ocasión como se mencionó anteriormente fue empleada para sustituir los valores faltantes en la variable edad.

Por otra parte, la media aritmética que es el valor promedio de las muestras y fue la principal medida que se utilizó, gracias a esta se logró clasificar a los grupos y atribuirle la etiqueta de (Bajo, Regular, Alto, Muy Alto). Por último, la varianza fue considerada ya que se aplicó un análisis de los principales componentes el cual permite reducir la dimensión de las observaciones perdiendo la menor cantidad de información posible.

Análisis de los resultados

En la temporada regular del 2016 de la MLB se efectuaron un total de 162 juegos. Mediante el presente trabajo de investigación se obtuvieron los siguientes resultados:

El promedio de participación de los bateadores fue de 83 juegos, es decir, el 51% de toda la temporada regular. Observando los 4 grupos por separado existe una gran diferencia de los jugadores que pertenecen a cada uno de estos, la cual se puede percibir de la siguiente forma:

- El grupo 1 participó en el 46% de los juegos.
- El grupo 2 participo en el 71% de los juegos.
- El grupo 3 participo en el 93% de los juegos.
- El grupo 4 participo en el 21% de los juegos.

Esto indica que hay dos grupos que están por encima del promedio y los otros dos están por debajo.

Otra característica que se observó fue en la variable "carreras", en donde el máximo de carreras anotadas por un jugador en la temporada fue de 123, y el promedio general del grupo completo fue de 36 carreras, ahora bien, el promedio por grupo se refleja de la siguiente manera: el grupo 1 anotó 24 carreras, el grupo 2 anotó 49 carreras, el grupo 3 anotó 82 carreras y por

último el grupo 4 anotó 5 carreras.

En la variable home run (HR) se observó que la cantidad máxima de HR dentro de la temporada alcanzada por un jugador fue de 47 y el promedio general de HR por jugador es de 9.

Luego examinando los grupos se obtuvo que el promedio de HR del grupo 1 fue de 5 HR, el grupo 2 de 12 HR, el grupo 3 de 23 HR y grupo 4 de 1 HR. Hasta ahora se puede apreciar que el grupo 4 está muy por debajo del promedio y el grupo 3 muy por encima del promedio, con ayuda de esto y las gráficas anteriores se logró etiquetar a los jugadores según su rendimiento de la siguiente forma:

- Grupo 1 Regular con una cantidad de 136 jugadores.
- Grupo 2 Alto con una cantidad de 117 jugadores.
- Grupo 3 Muy Alto con una cantidad de 121 jugadores.
- Grupo 4 Bajo con una cantidad con una cantidad de 173 jugadores.

Esto denota que el 56% de los bateadores están por debajo del promedio en las principales características.

Luego de finalizado el proceso de la investigación surgió una nueva inquietud, la cual arrojó la siguiente interrogante: ¿Qué equipo al terminar la temporada regular tenía mejor grupo de bateadores?

Para ello se presenta la siguiente imagen que contiene la tabla:

	ARI	ATL	BAL	BOS	CHC	CIN	CLE	COL	CWS	DET	HOU	KC	LAA	LAD	MIA	MIL	MIN	NYM
Alto	3	3	4	2	1	3	3	3	3	6	5	5	4	4	6	2	7	4
Bajo	6	2	6	9	3	7	5	6	6	6	6	4	7	8	8	7	2	6
Muy Alto	2	4	5	6	6	4	5	4	4	4	3	4	3	5	4	4	3	3
Regular	4	7	4	3	7	4	4	6	6	4	5	6	7	2	2	4	5	7

	NYY	OAK	PHI	PIT	SD	SEA	SF	STL	TB	TEX	TOR	WSH
Alto	3	3	3	8	5	6	2	5	4	4	3	3
Bajo	6	9	5	6	9	8	5	4	3	4	5	5
Muy Alto	5	3	4	3	1	3	5	3	4	6	7	4
Regular	5	3	4	3	5	2	5	4	6	4	4	4

De acuerdo con esta tabla, el equipo de Texas tenía el mejor grupo de bateadores al terminar la temporada regular 2016, sin embargo, el campeón fue el equipo de Chicago Cubs, lo que demuestra que un equipo para ser ganador no sólo requiere excelencia en el bateo sino también en la defensiva.