Casa abierta al tiempo
UNIVERSIDAD AUTÓNOMA
METROPOLITANA
Unidad Cuajimalpa

Análisis de datos Liga Mexicana de baseboll con K-means

Barajas Sánchez Maximiliano
Jimenez Romero Mauricio De Jesus
Rodríguez Pérez Amanda Yael

1 de octubre de 2023

1. Introducción

Durante el desarrollo de esta práctica, se empleó el lenguaje de programación R con el propósito de manipular una base de datos más amplia, a la vez que se incorporaron datos propios a la tabla de la LMB. Mediante la inclusión de bibliotecas específicas de R, se busca la exploración y análisis de datos multivariables con el objetivo de generar gráficos mediante el algoritmo de K-means. Este enfoque nos permitirá obtener una comprensión más profunda de la información contenida en la base de datos, facilitando así la identificación de patrones y tendencias relevantes.
El método de K-means, es un algoritmo de agrupamiento que tiene como objetivo particionar un conjunto de datos en k grupos tambien llamados clusters. Se inicia eleccionando k centroides de manera aleatoria, donde k representa el número predefinido de clusters que se desean analizar, despues se asigna cada punto de datos al centroide más cercano y recalcula los centroides basándose en la media de los puntos asignados a cada grupo. Este proceso de asignación y actualización se repite iterativamente hasta que los centroides convergen y el agrupamiento se vuelve estable. Como resultado final tenemos la formación de k clusters, donde cada punto de datos pertenece al grupo cuyo centroide es el más cercano. El algoritmo es efectivo para identificar patrones y estructuras en conjuntos de datos.

2. Marco Teórico

A lo largo de esta sección explicaremos el fundamento teórico sobre el cual se basa el algoritmo K-Means. El algoritmo puede resumirse bervemente en 5 secciones:

2.1. Seleccionar Centroides

inicialmente una vez se cuenta con los datos a analizar lo que se hace es seleccionar K datos particulares de manera aleatoria, cabe recalcar que los resultados del algoritmo como su convergencia son extremadamente sensibles al como son elegidos los centroides iniciales y la cantidad de los mismos, de ahi el que existan variantes y métodos particulares de elegir estos parámetros de manera mas eficiente, en particular para elegir los centroides iniciales de manera mas óptima lo que se hace es seleccionar un centroide aleatorio, a continuación calculamos el cuadrado de la distancia de todos los puntos a dicho centroide y elegimos el punto mas lejano como nuestro proximo centroide, repetimos el proceso hasta encontrar K centroides solo que una vez poseemos mas de un centroide calculamos la suma de las distancias de cada punto a los centroides ya establecidos y tomamos el de mayor distancia como el siguiente centroide, esto lo establecen

Arthur y Vassilvitskii (2007), este método de elegir los centroides es el usado por las principales bibliotecas de aprendizaje máquina aunque a veces puede resultar computacionalmente costoso pero garantiza una convergencia mucho mas rápida así como una consistencia en los resultados replicable, una vez son elegidos los centroides. En cuanto a la cantidad de centroides existen ciertas heurísticas como el método de la curva del codo el cual aplica el algoritmo con una variedad de valores para K's de manera ascendente y lo gráfica contra la inercia (suma de los cuadrados de las distancias de los puntos al centroide mas cercano) de tal suerte que una vez comience a tener un comportamiento asintótico que se puede ver como un codo en la gráfica se establece esa instancia del método como la de la K apropiada, tal como establece Banerji(2021)

2.2. Asignación

En este paso únicamente calculamos la distancia de todos los puntos a los centroides y los categorizamos en el cluster del centroide del cual estan menos alejados.

2.3. Actualización de centroides

Una vez realizado lo anterior se cálcula un vector promedio representativo de cada etiqueta definida por el paso anterior y se repite el paso de asignación.

2.4. Condición de paro del algoritmo

En general como condición de paro se usa la diferencia entre los centroides de cada cluster consecutivos en terminos absolutos como condición para dejar de aplicar el método.

3. Metodología

Para la recopilación de datos, se llevaron a cabo tres rondas de bateo durante una sesión de clase. Posteriormente, se calculó el promedio de desempeño en cada ronda para cada estudiante, incluso para aquellos que estuvieron ausentes. Para los alumnos faltantes, se generaron números aleatorios dentro del intervalo de los promedios obtenidos por los demás estudiantes. Además, se incorporaron los datos proporcionados por la Liga Mexicana de Béisbol, así como los promedios correspondientes a los estudiantes ausentes, para completar los campos vacíos.

1. Jugador: nobre del jugador
2. Pos: posicion del jugador
3. Equipo
4. JJ :juegos jugados
5. TB: total de bases
6. C : carreras anotadas
7. H :Hits
8. X2B: dobles
9. X3B: tripes
10. HR: home runs

11. CI: carreras impulsadas
12. BB: base por bolas
13. P: ponches
14. BR: bases robadas
15. AR: asistencias en el jardin
16. PRO: promedio del bateo
17. OBP:porcentaje de embasado
18. SLG: porcentaje de slugging
19. OPS: OPS combina el OBP y el SLG

Una vez que la tabla de datos estuvo completa, se guardó como un archivo CSV para su posterior análisis en R.

4. Discusión y resultados

Inicialmente decidimos aplicar el método de la gráfica del codo como describe Banerji (2021) en cada conjunto de datos además de hecerlo de manera estándar con u3 grupos, obtuvimos los siguientes resultados:

4.1. Ejercicio 1

Aquí se tomaron en cuenta las variables J , PRO y OBP para aplicar el método de K - Medias, al graficar la inercia o bien la suma de las distancias de todos los puntos al centroide mas cercano obtenemos la siguiente gráfica:

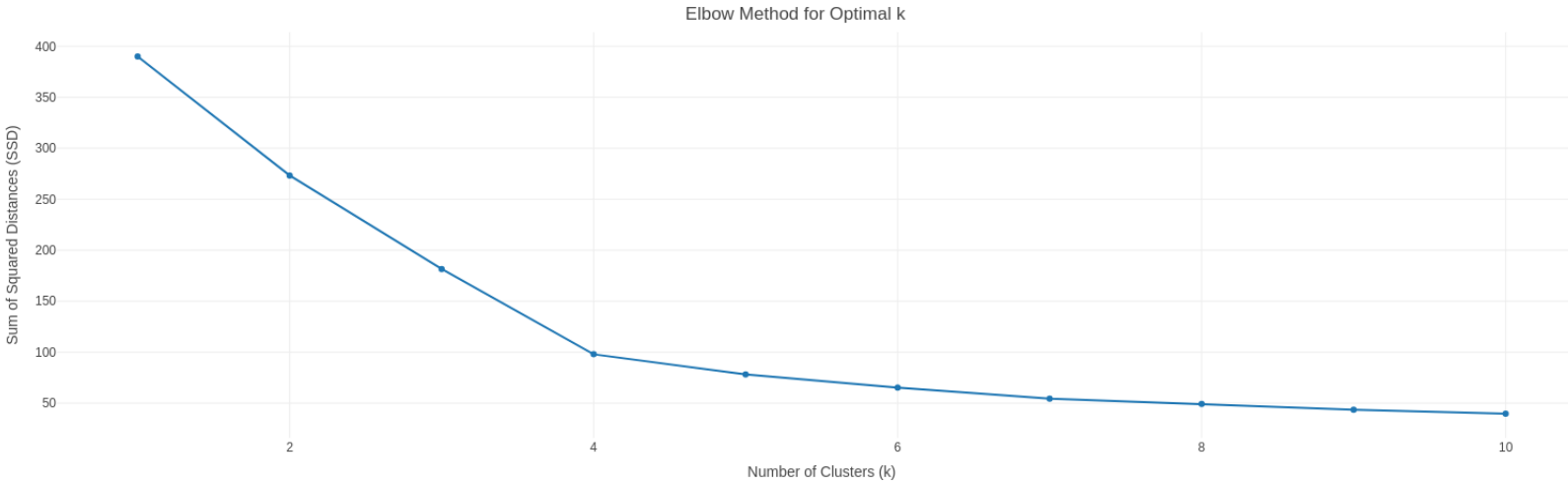

Figura 1: Gráfica del método del codo sobre el conjunto de datos del ejercicio 1, note que el çodo"se encuentra en k=4

Inicialmente como lo mencionamos anteriormente probamos con k=3 obteniendo la siguiente gráfica:

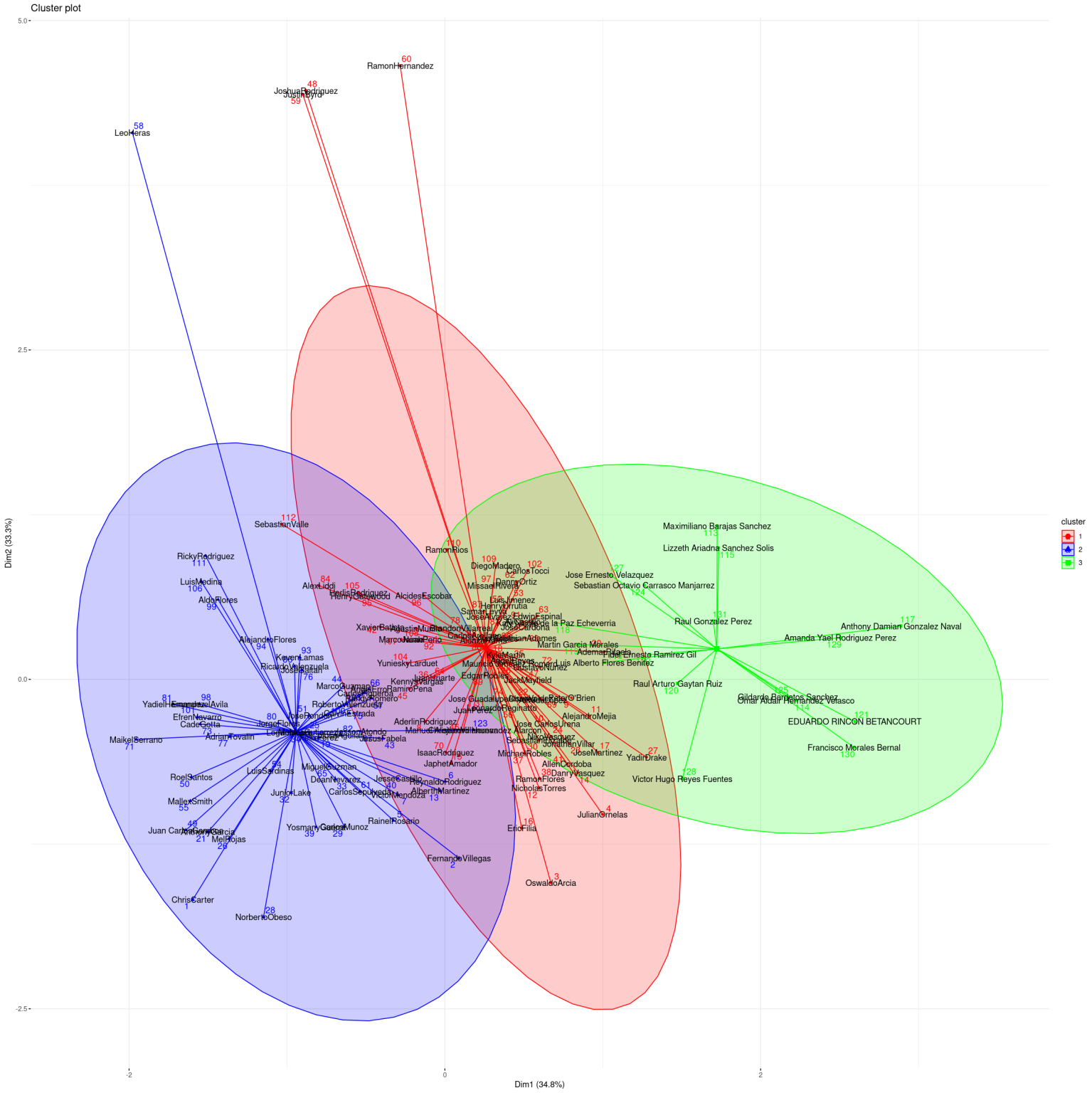

Figura 2: Gráfica de los resultados del algoritmo para el ejemplo 1 con k=3

Notamos que si aplicamos el método con k=3 una cantidad existente pero mínima con estas variables sobre las cuales aplicamos el algoritmo de K-means se mantienen extremadamente lejos de nuestros 3 centroides, esto junto con la gráfica de codo nos dan un indicio de que requerimos de un centroide mas, si aplicamos el método con k=4 obtenemos el siguiente resultado:

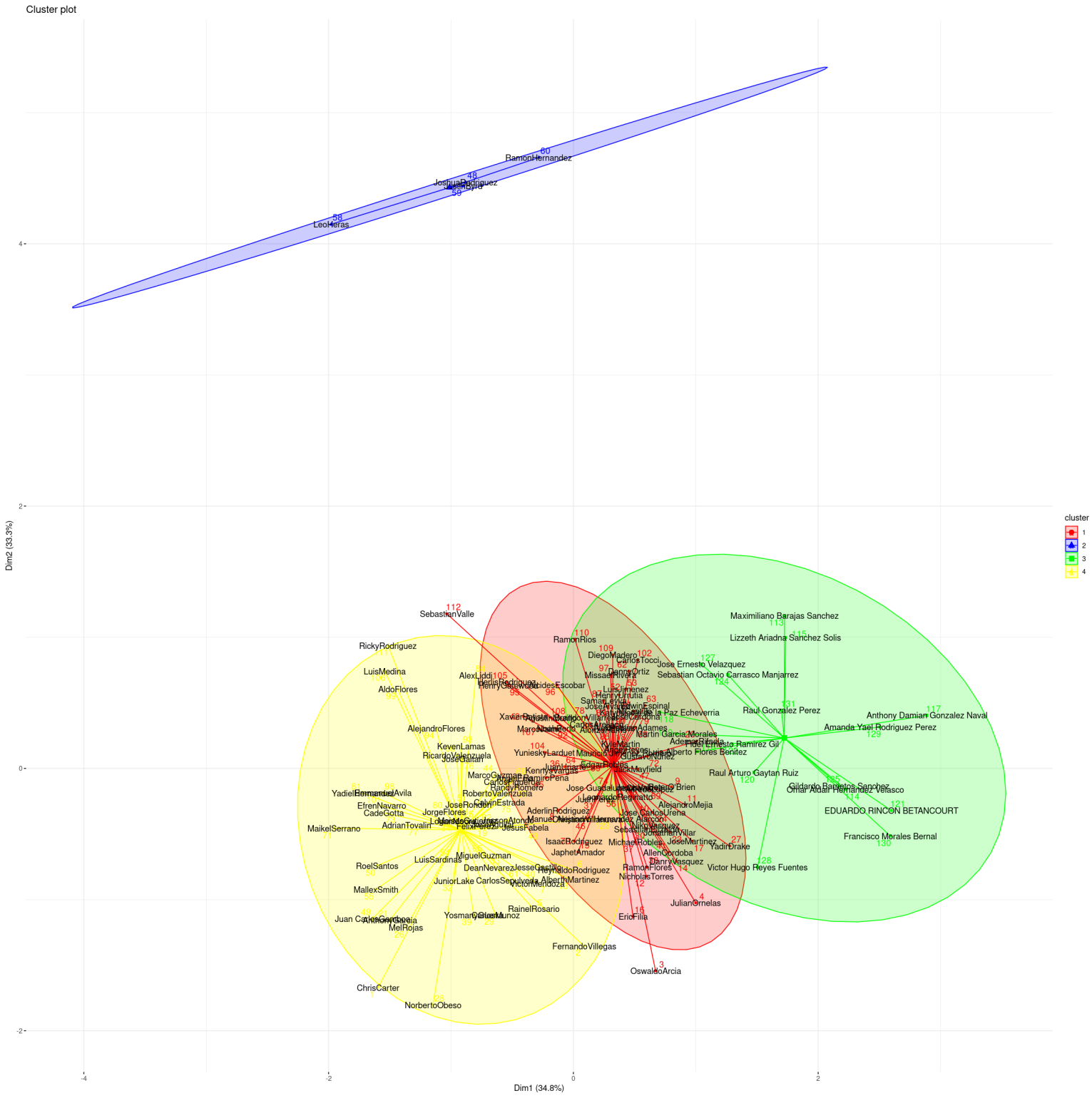

Figura 3: Gráfica de los resultados del algoritmo para el ejemplo 1 con k=4

Esta vez los datos que se encontraban a una distancia considerable de los centroides en la gráfica anterior se encuentran en su propio cluster que es mucho mas apropiado.

4.2. Ejercicio 2

Aquí se encuentra el análisis realizado con el algoritmo K-Means sobre las variables TB,CI y HR, comenzamos realizando la gráfica de codo para guiarnos a elegir una k apropiada para el método:

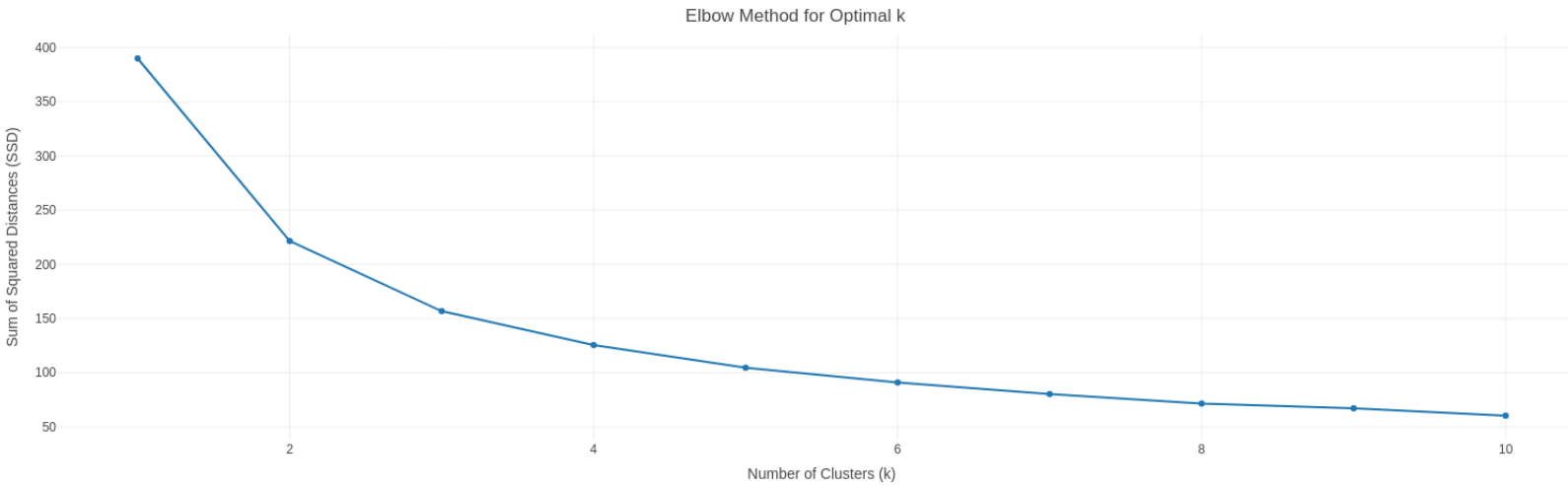

Figura 4: Gráfica de codo para los datos del ejercicio 2, note que el codo se encuentra en k=2

Inicialmente como lo mencionamos, tratamos con una k=3 y obtenemos los siguientes resultados:

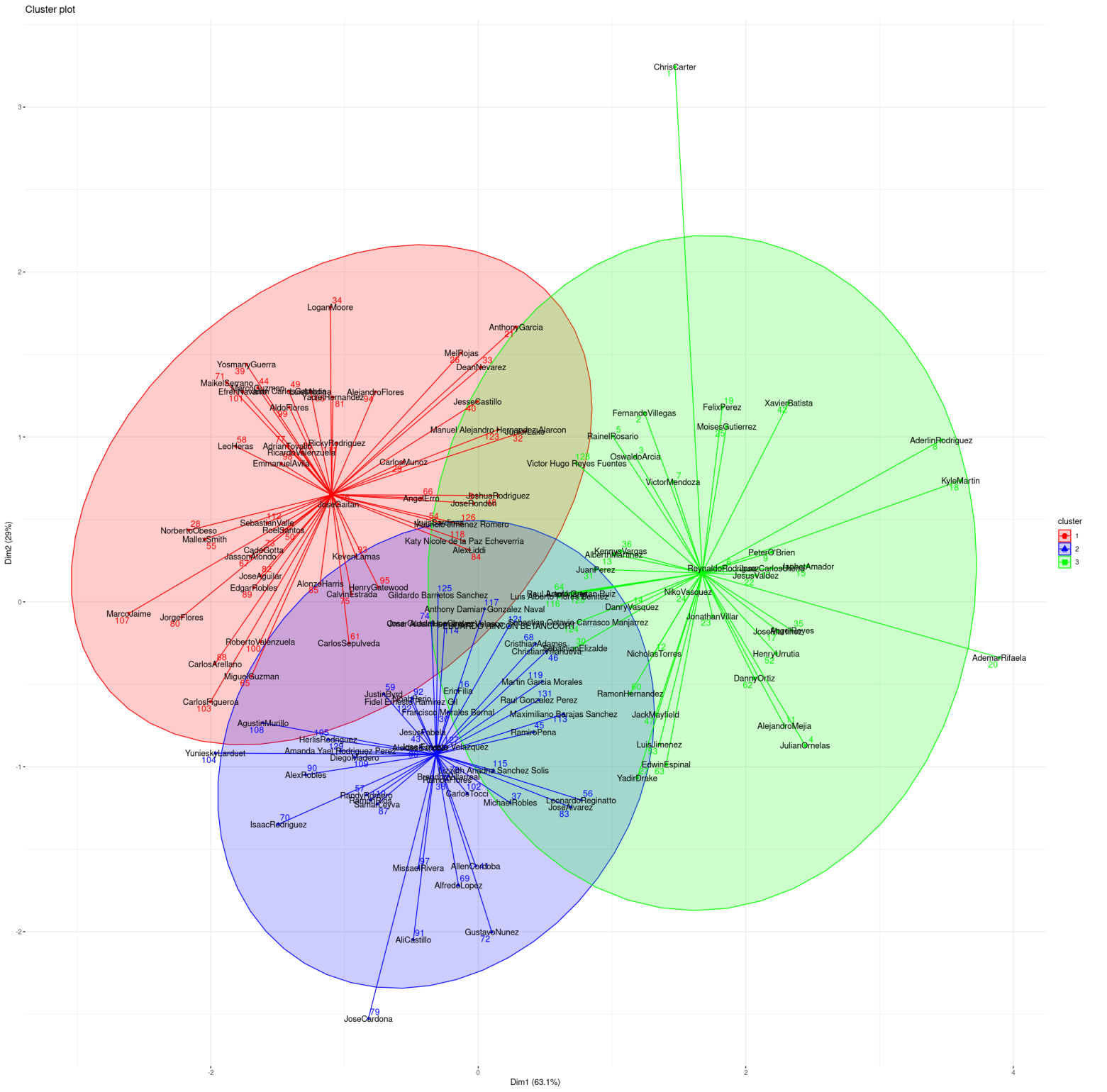

Figura 5: Gráfica del método de K-means con los datos del ejercicio 2 y k=3

Notamos que los clusters parecen estar demasiado cerca unos de otros dando lugar a regiones que podrían crear un sesgo, en especial en las intersecciones de las elipses trazadas por la gráfica, si probamos con k=2 obtenemos la siguiente gráfica:

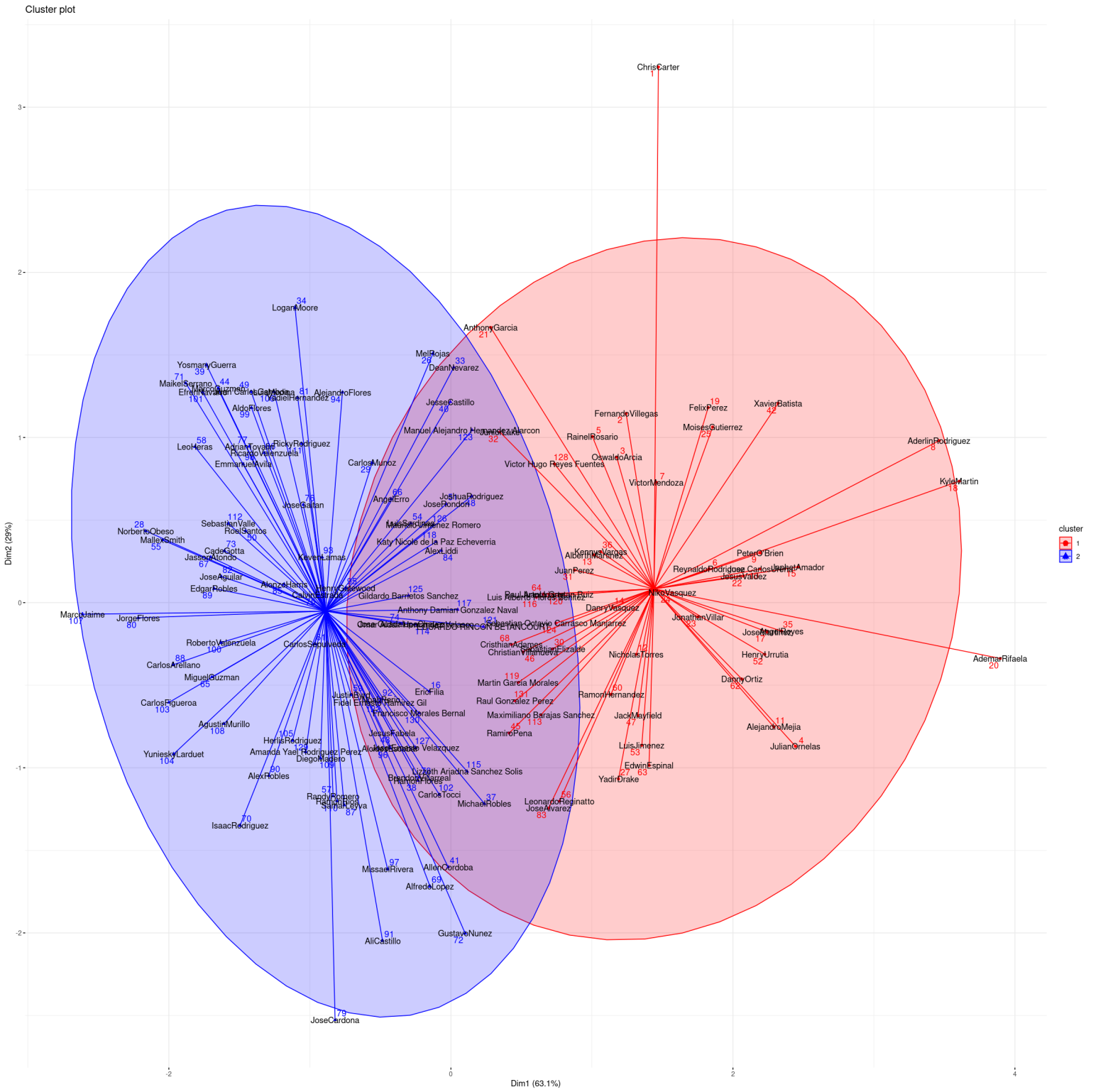

Figura 6: Gráfica del método de K-means con los datos del ejercicio 2 y k=2

Notamos que los clusters en este caso se encuentran considerablemente separados creando una dsitnición mucho mas clara entre ambos conjuntos.

4.3. Ejercicio 3

Para este caso se tomaron en cuenta 4 variables, JJ, TB, X2B y X3B, como con los incisos anteriores, graficamos el método del codo para obtener una k apropiada en primera instancia:

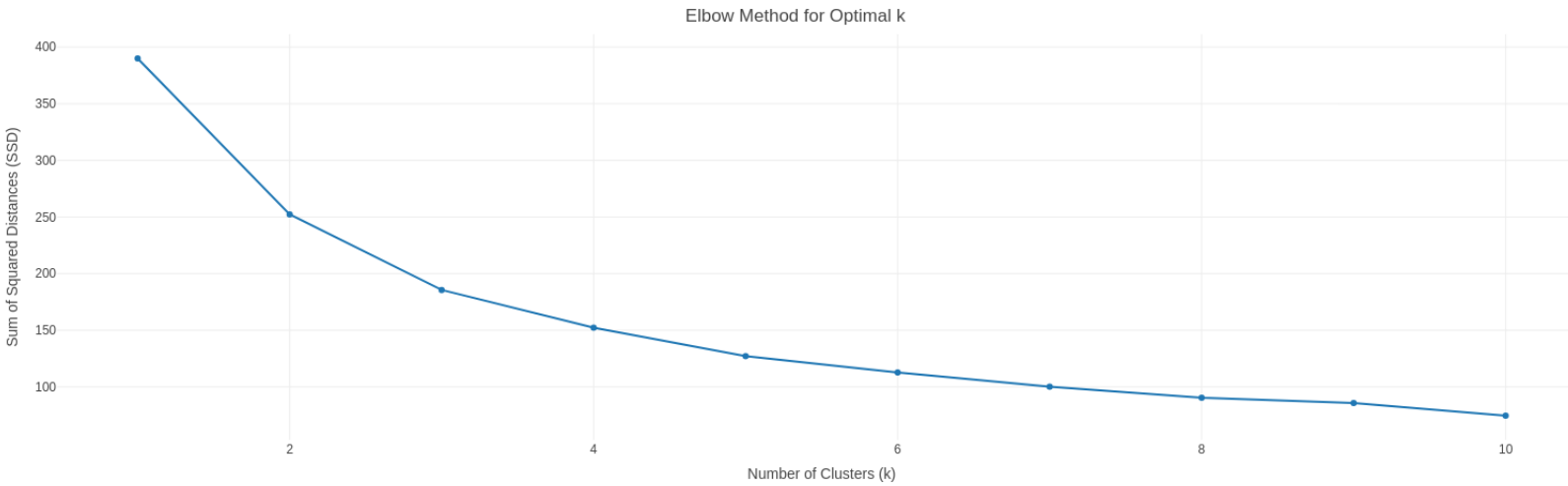

Figura 7: Gráfica del método del codo para los datos del ejercicio 3, note que el codo se encuentra en aproximadamente k=3

Procedemos a probar el algoritmo con k=3 y resulta la siguiente gráfica:

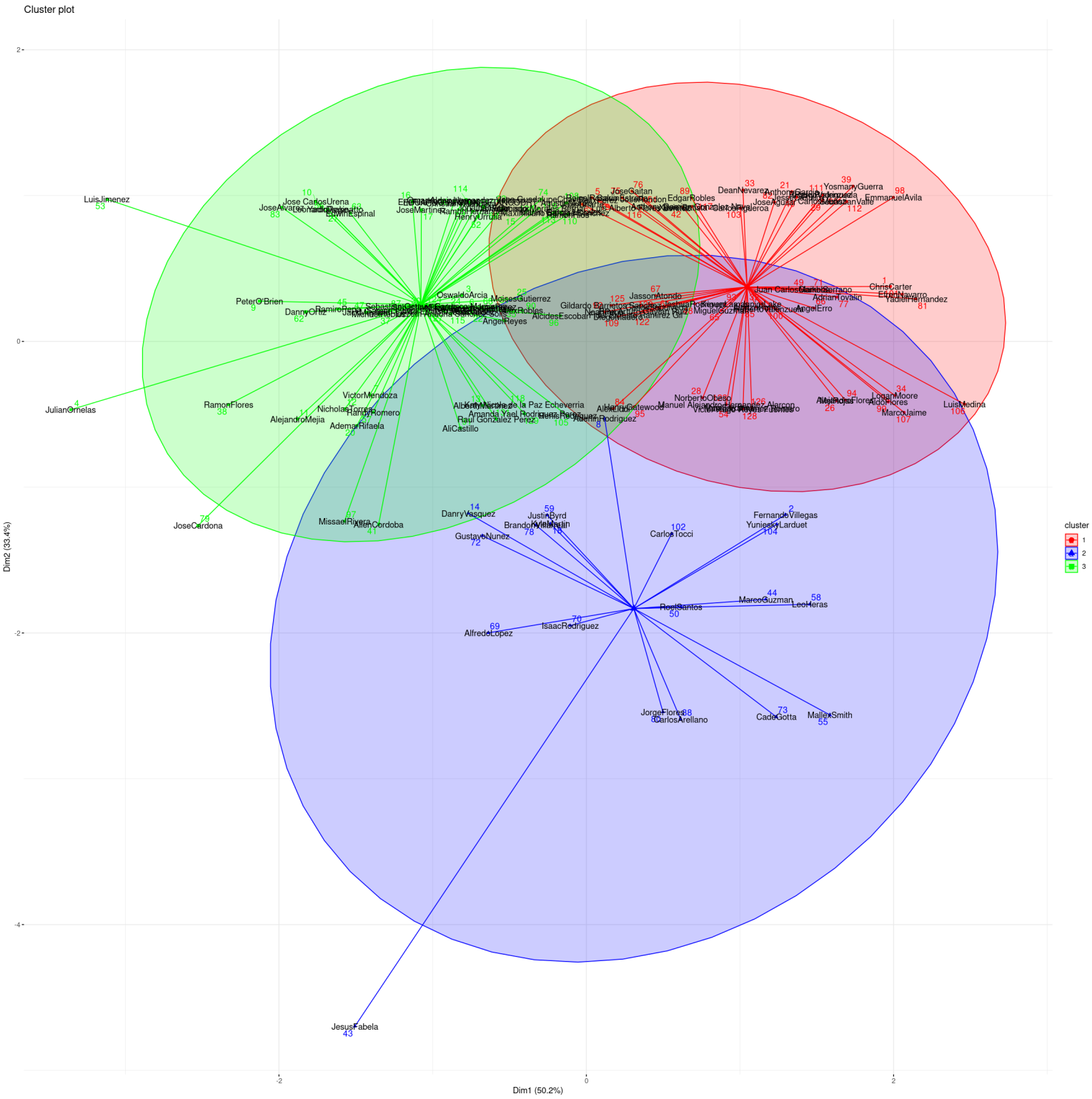

Figura 8: Gráfica del método de K-means con los datos del ejercicio 3 y k=3

Notamos que los conjuntos se encuentran apropiadamente delimitados tal como lo inferimos con el método del codo a excepción de algunos datos poco rutinarios que se encuentran lejos de los 3 centroides.

5. Conclusiones

Finalmente concluimos que el comprender el fundamento detrás del algoritmo es vital para interpretar la salida del mismo, al ser un algoritmo de aprendizaje no supervisado es crucial de igual manera que sepamos ajustar los parámetros que aun con las implementaciones de las bibliotecas de machine learning dependen aún de nosotros como usuarios de las mismas y analistas en general, de igual manera la biblioteca Cluster representa una herramienta crucial para el análisis superficial de este tipo de datos.

6. Bibliografía

1. K-Mean: Getting the Optimal Number of Clusters Banerji (2021)
2. Arthur, D., and Vassilvitskii, S. (2007). k-means++: The Advantages of Careful Seeding. In Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms (pp. 1027–1035). Society for Industrial and Applied Mathematics.
3. Barlow, H. B. (1989). Unsupervised learning. Neural computation, 1(3), 295-311.
4. Celebi, M. E., and Aydin, K. (Eds.). (2016). Unsupervised learning algorithms (Vol. 9, p. 103). Cham: Springer.