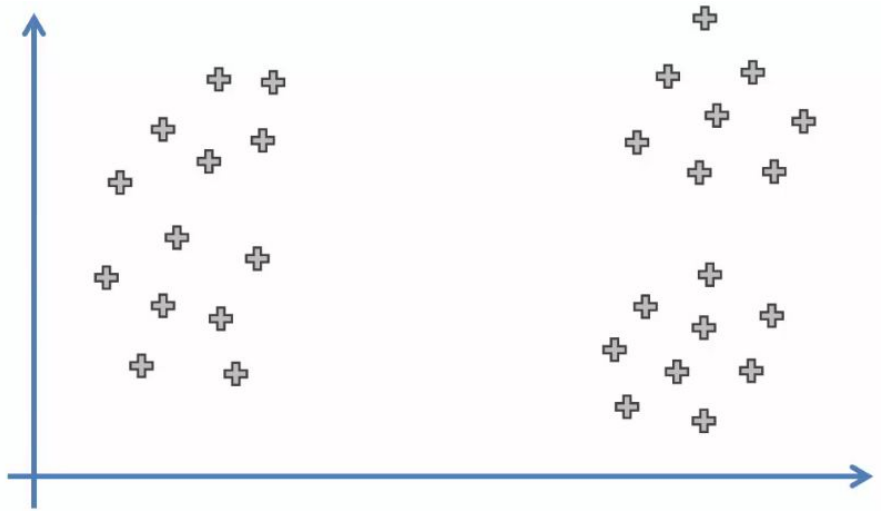
Choosing the right number of clusters

Hablamos acerca de como encontrar el número correcto de clústeres para un problema de de datos dado.

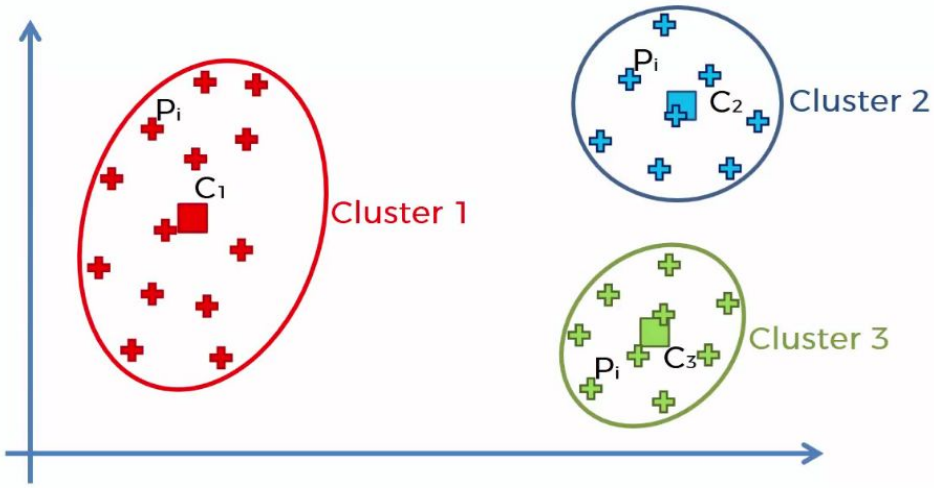
Hablaremos acerca del algoritmo que está detrás para encontrar el correcto número de clústers.

Aprenderemos como decidir qué número de clusters de entrada en mi algoritmo K-Means

Tenemos este problema de diseño nuevamente con dos variables (puede ser cualquier cantidad de columnas o variables), coordenadas X e Y



Si ejecutamos nuestro algoritmo de K-Means con tres clústers o con K predeterminada a ser tres, el resultado se verá algo como esto.



FUeron identificdos tres clústers y sabemos que se debió ejecutar Kmeans ++ solo para evitar la trampaa de inicialización aleatoria

Con el fin de entender, si por ejemplo 2 clústers podrían haber sido mejor en este escenario, o si 10 clústers hubiesen sido mejor, necesitamos una métrica en particular.

Necesitamos una forma de comprender o evaluar como funciona un cierto número de clústers comparado a diferente número de clústers y preferiblemente esa métrica debe ser cuantificable.

Entonces **¿Qué tipo de métrica podemos imponer a nuestro algoritmo de clustering que nos diga algo sobre el resultado final?**

Esta métrica es llamada suma de cuadrados dentro del clúster. Esta es la formula de como es calculada:

$$\text{WCSS} = \sum_{\text{P}_\text{i} \text{ in Cluster 1}} \text{distance}(\text{P}_\text{i}, \text{C}_1)^2 + \sum_{\text{P}_\text{i} \text{ in Cluster 2}} \text{distance}(\text{P}_\text{i}, \text{C}_2)^2 + \sum_{\text{P}_\text{i} \text{ in Cluster 3}} \text{distance}(\text{P}_\text{i}, \text{C}_3)^2$$

Este es un ejemplo para tres clústeres

$$WCSS = \sum_{Pi \text{ in Cluster } 1} distance(P_i, C_1)^2 + \sum_{Pi \text{ in Cluster } 2} distance(P_i, C_2)^2 + \sum_{Pi \text{ in Cluster } 3} distance(P_i, C_3)^2$$

Analicemos que sucede aquí.

Tenemos tres elementos y para cada uno la suma es calculada, para cada clúster, de hecho se calcula la suma dentro de ese clúster, por ello la W al comienzo **[Within Cluster Sum Squared]**

MIremos por ejemplo el clúster central:

$$\text{WCSS} = \sum_{\text{P}_\text{i} \text{ in Cluster 1}} \text{distance}(\text{P}_\text{i}, \text{C}_1)^2 + \sum_{\text{P}_\text{i} \text{ in Cluster 2}} \text{distance}(\text{P}_\text{i}, \text{C}_2)^2 + \sum_{\text{P}_\text{i} \text{ in Cluster 3}} \text{distance}(\text{P}_\text{i}, \text{C}_3)^2$$

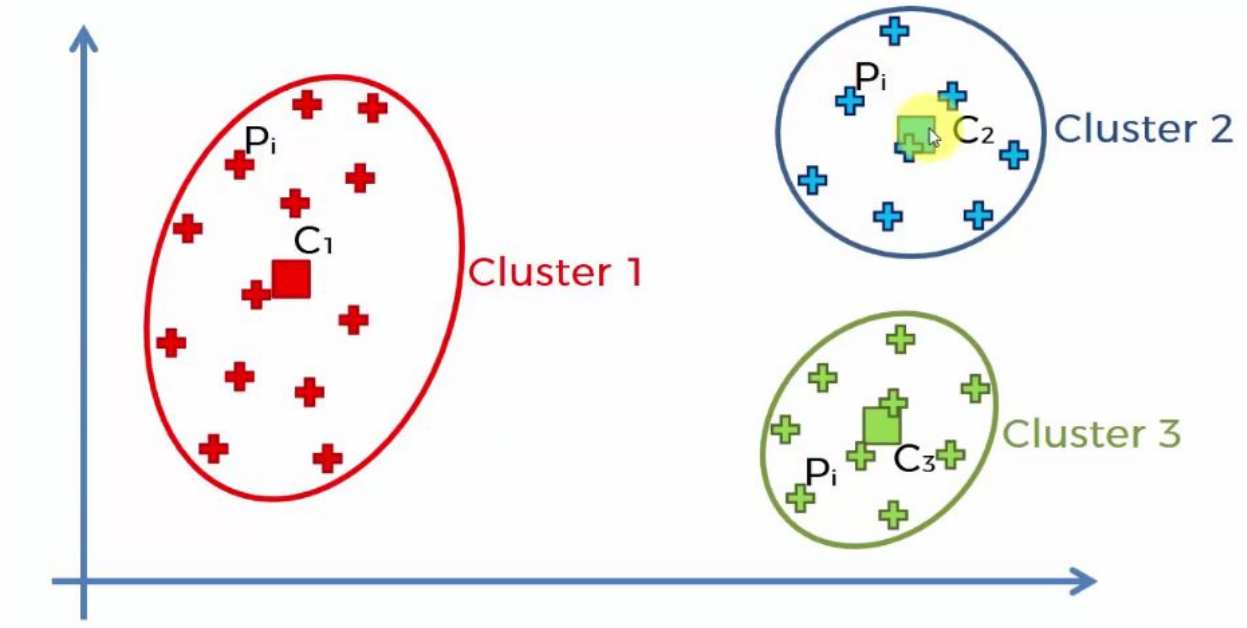
Aquí estamos tomando cada punto dentro del clúster 2

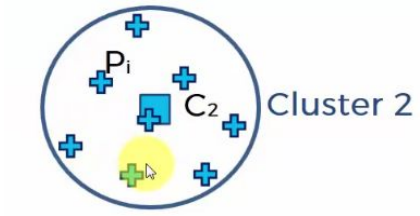
$$\sum_{P_i \text{ in Cluster 2}} \text{distance}(P_i, C_2)^2$$

Y hacemos una suma a través de esos puntos. Lo que estamos sumando es la distancia entre cada punto dentro del cluster 2 y **el centroide del Cluster 2** y elevamos al cuadrado esa distancia.

Asi que nosotros tomamos esas distancias cuadradas.

Si examinamos nuestro gráfico, por ejemplo en el clúster 2 nosotros estamos tomando el centroide y calculamos la distancia a cada punto





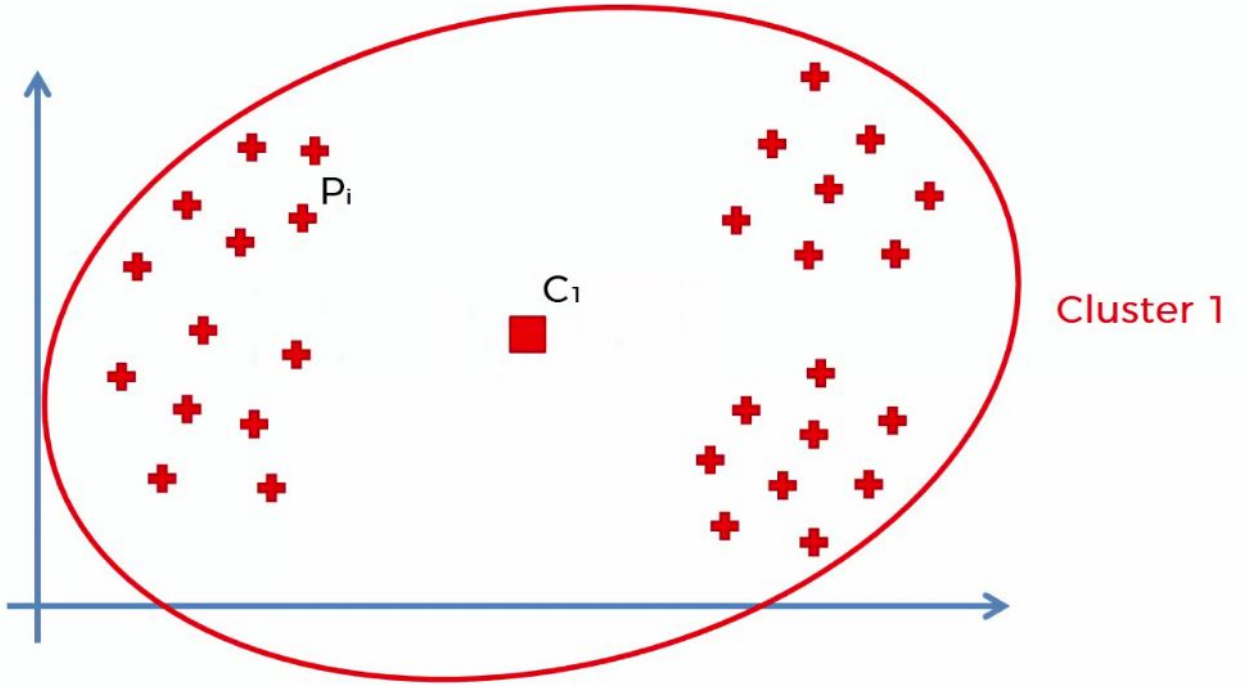
Y es ahí en donde hacemos el cuadrado de la distancia
El centroide, hasta ese punto entonces lo hacemos al cuadrado y así sucesivamente

Después tomamos la suma de todos los cuadrados de todas esas distancias y estamos sumandolas.
Hacemos eso, para el primer cluster, para el segundo y para el tercero

$$\text{WCSS} = \sum_{P_i \text{ in Cluster 1}} \text{distance}(P_i, C_1)^2 + \sum_{P_i \text{ in Cluster 2}} \text{distance}(P_i, C_2)^2 + \sum_{P_i \text{ in Cluster 3}} \text{distance}(P_i, C_3)^2$$

Y el resultado obtenido es el total de la suma, la suma completa y ese resultado va a ser nuestra métrica y es una métrica bastante buena en términos de entendimiento o comparación de la bondad de ajuste entre dos diferentes clústers K-Means.
¿Y como sabemos eso?
Echemos un vistazo a cuando teniamos un cluster y miremos lo que la métrica WCSS dice:

Va a cambiar a medida que aumentamos el número de clusters. Aquí esta nuestro gráfico con solo un cluster y nuestro centroide ubicado:

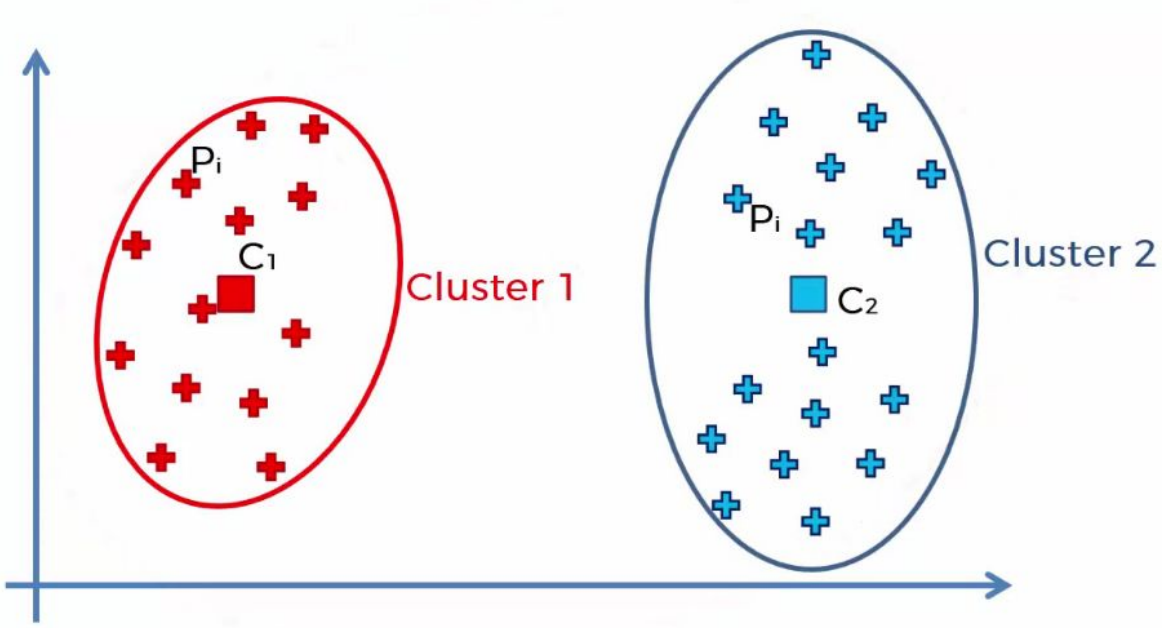


Por cada punto, tenemos que medir la distancia entre ese punto y el centroide, entonces tendremos que adicionar todas esas distancias. Podemos ver que obtendremos un valor

bastante grande, ya que este centroide está alejado de todos los puntos y esos puntos tienen que alcanzarlo.

Detallando como se vé esto, nos ayudará a recordar que distancia no es en términos de valor absoluto, pero es justo lo que se siente. Es una distancia bastante grande.

Ahora vamos a incrementar el número de clústers a 2
K = 2 y miremos como cambia todo



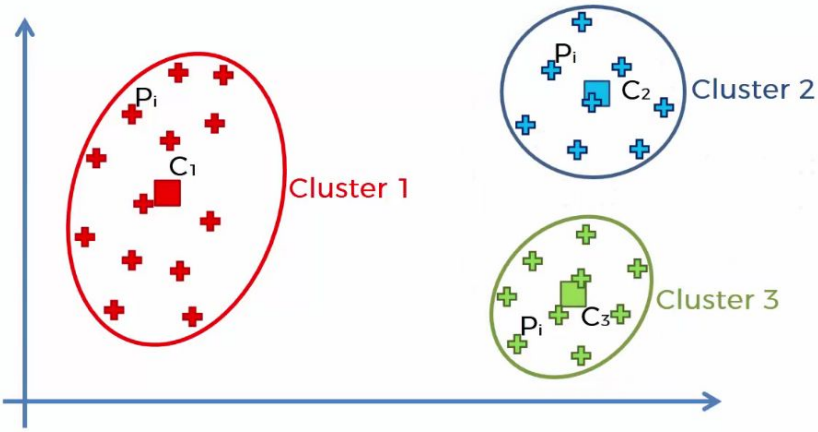
$$\text{WCSS} = \sum_{P_i \text{ in Cluster 1}} \text{distance}(P_i, C_1)^2 + \sum_{P_i \text{ in Cluster 2}} \text{distance}(P_i, C_2)^2$$

Ahora vemos que tenemos dos centroides cada uno en el centro de gravedad de sus respectivos datos.
Entonces cada uno de estos puntos ya no tiene que llegar a la mitad o la distancia no es grande, dado que cada punto tiene su centroide cerca. No es genial

Todo lo que hay que hacer es calcular la distancia de los puntos rojos al centroide rojo y de los azules al centroide azul, lo cual es cada vez menos.

Esto significa que la distancia total o el total de WCSS que miramos aquí va a ser menor que cuando teniamos un centroide.

Una vez más incrementamos el número de clústers a K = 3 y visualmente sigue asi:



$$\text{WCSS} = \sum_{P_i \text{ in Cluster 1}} \text{distance}(P_i, C_1)^2 + \sum_{P_i \text{ in Cluster 2}} \text{distance}(P_i, C_2)^2 + \sum_{P_i \text{ in Cluster 3}} \text{distance}(P_i, C_3)^2$$

Tenemos 3 clusters y en el cluster 1 nada cambió, tenemos lo mismo. Los otros clústers están separados y cada uno tiene su propio centroide y la distancia de nuevo desde cada uno de sus puntos de datos a su centroide se ha decrementado, pues ahora estos puntos están más cerca a su centroide que antes cuando habia un solo cluster para estos dos grupos de cluster 2 y 3

Entonces WCSS dice que la distancia decrementa

Pero … ¿Cuál es el límite de esto?

Qué tanto disminuirá la distancia
Bien permitamos pensar en incrementar el número de clústers a 4, 5 y 6 sucesivamente
¿Cúal es el máximo de clústers que podemos tener?

Podemos tener tantos clústers como puntos o elementos de datos en nuestro dataset. Si tenemos 50 datos o muestras, podemos tener hasta 50 clusters porque después de todo tenemos 50 o después de revisar el número de puntos no hay más puntos para agrupar cada punto tiene su propio clúster

¿ Entonces en este caso, cual seria el valor de WCSS ?

SI Tenemos un numero de clusters igual al numero de datos entonces las distancias serian reducidas a 0 y WCSS seria 0, porque cada punto de datos tiene su propio clúster y por lo tanto su propio centroide y ese centroide va a ser exactamente donde esta ese punto y por lo tanto la distancia entre el punto y el centroide va a ser 0.

$$\text{WCSS} = \sum_{P_i \text{ in Cluster 1}} \text{distance}(P_i, C_1)^2 + \sum_{P_i \text{ in Cluster 2}} \text{distance}(P_i, C_2)^2 + \sum_{P_i \text{ in Cluster 3}} \text{distance}(P_i, C_3)^2$$

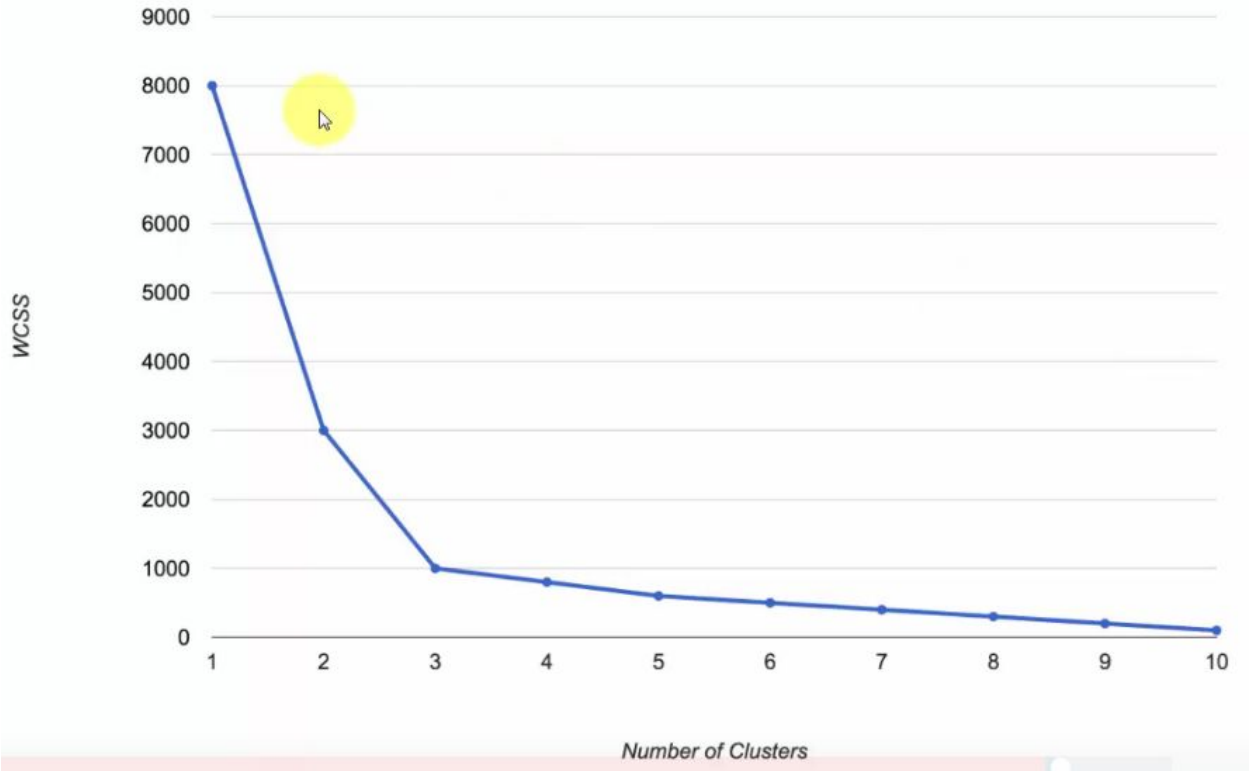
Si elevamos esa distancia 0 al cuadrado va a ser 0 y si le sumamos todos las distancias 0 al cuadrado, va a ser 0
Como podemos ver, WCSS decrementará gradualmente desde un valor sustancial cuando tenemos un clúster a 0 cuando incrementamos el número de clústers

Así que esta es una buena métrica, pero al mismo tiempo, está constantemente decreciendo la distancia asi que constantemente esta mejorando Y eso es porque, como puede ver, cuanto menos dice el WCSS o cuanto mayor es el número de clústeres, mejor es la bondad de ajuste, estábamos ajustando nuestros datos mejor y mejor y, por lo tanto, el menor WCSS es el mejor.

¿Pero como encontramos el valor óptimo de ajuste si este sigue mejorando?

Hay un sacrificio que viene con esa mejora y ese es exactamente el caso

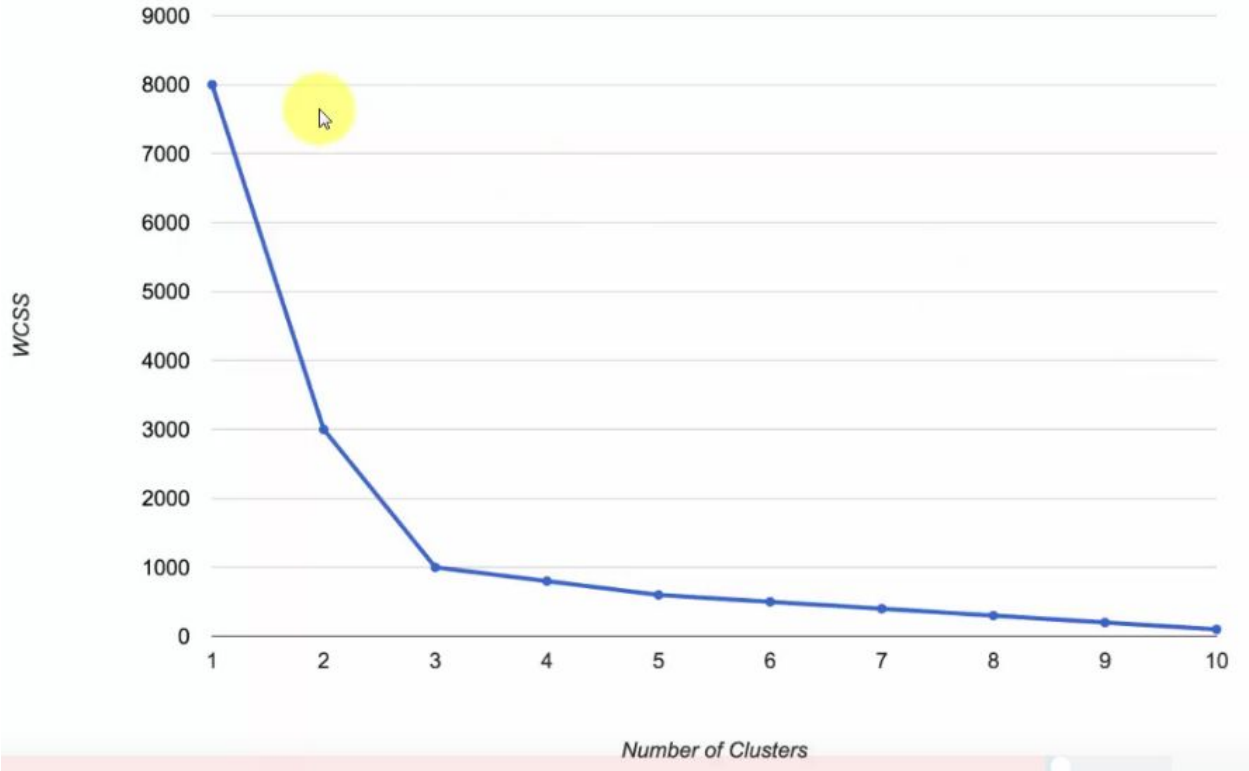
Miremos esto en un gráfico



Este gráfico representa como la distancia WCSS cambia a medida que incrementamos el número de clúster.
Como podemos ver, al comienzo el WCSS dice que comienza en un numero bastante grande y no importa cual es, su valor se mide en valor absoluto

Lo que importa es como cambia la relativa comparación entre diferentes métodos de k-means con diferentes números de clustering

WCSS de 8000 con 1 cluster
WCSS de 3000 con 2 clusters
WCSS de 1000 con 3 clusters

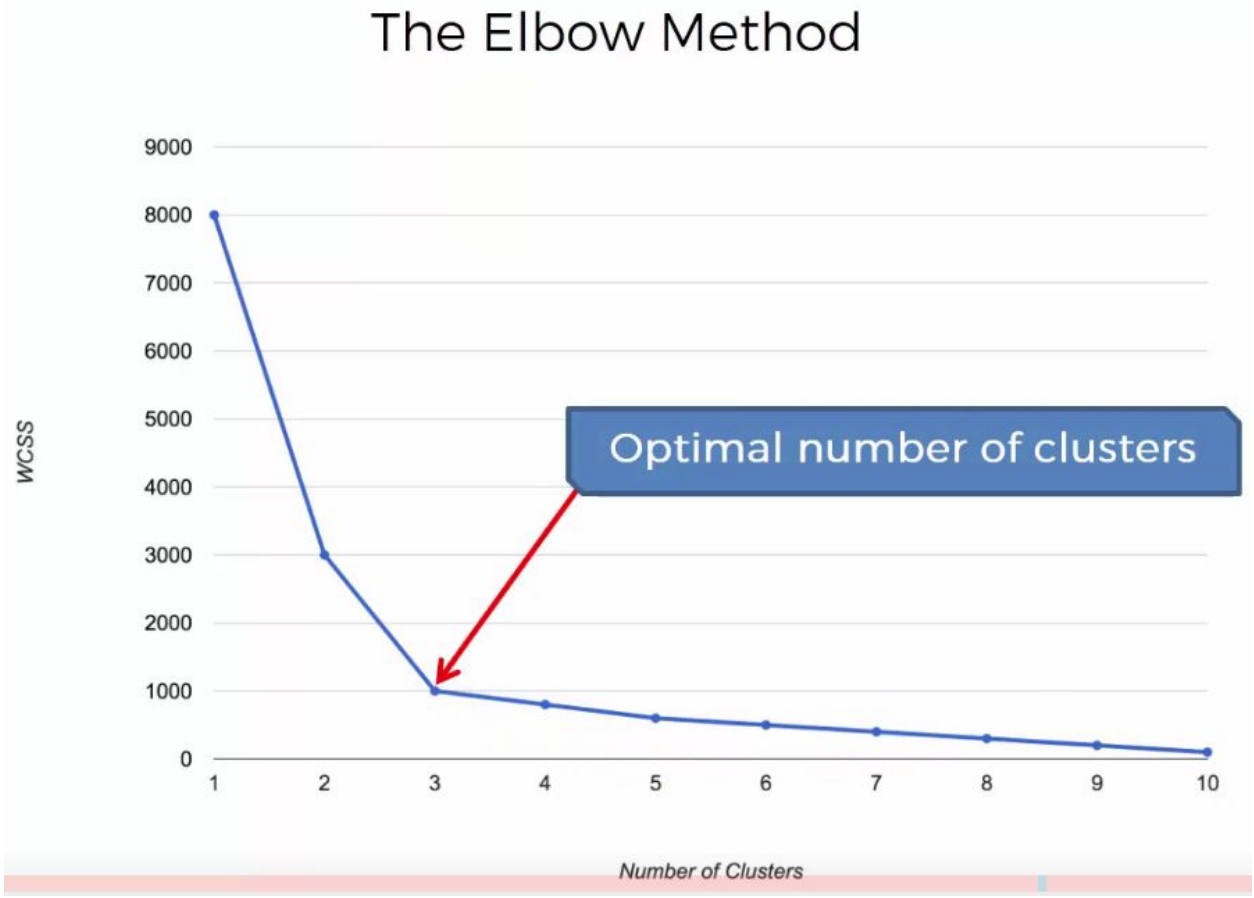


De tres a cuatro lo que sucede es que pasa de 1000 a tal vez 800 y de ahí a 600, 500 y así sucesivamente
Si miramos las primeras dos mejoras o los primeros dos cambios los saltos fueron enormes
De ahí en adelante, el WCSS no se reduce sustancialmente

Este es nuestro consejo para seleccionar el óptimo número de clústeres y el método para hacerlo es el método de Elbow que es muy visual. La gráfica se parece a un codo.

Es en la forma del codo en la gráfica, en donde la caida pasa, de ser bastante grande o sustancial a ser normal o no sustancial o no tan buena la reducción, por lo tanto, ese punto en la tabla va a ser la cantidad óptima de clusters a tener.

En este caso tres clústers



Este método es bastante arbitrario, por lo tanto a veces las situaciones no son tan obvias, a veces el codo podría no ser tan evidente como en este caso y por lo tanto alguien podría elegir un número de clústers para que alguien más pueda venir y elegir un número, otro numero de clústers.

Pero esa es una decisión que se debe tomar como científico de datos, a veces necesito decidir como el algoritmo debe ser estructurado.
Y este es uno de los casos, porque estamos decidiendo que tipo de algoritmo K-means clustering ejecutar para esa entrada de valor de K clústeres.

Y si, esto puede ser arbitrario, pero si usted no esta realmente seguro, entonces usted solo necesita ejecutar K-Means por ejemplo con tres clusters, mirar como estaba antes y ver cual es la diferencia y hacer esa decisión sobre cual es el K optimo para mi analisis, porque al final del dia eres la persona que crea este análisis y debes decidir que cantidad de clústers es óptima, por lo que el método de Elbow es solo un enfoque que puede ayudarme con esa decisión, pero al final del día es mi decisión.

Ahora tengo bastantes elementos sobre K-Means y puedo implementarlo