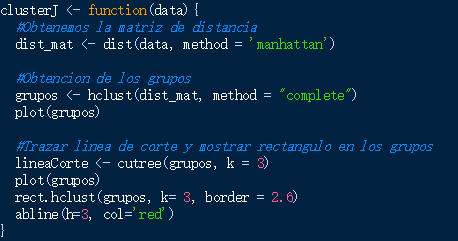
Consideraciones:

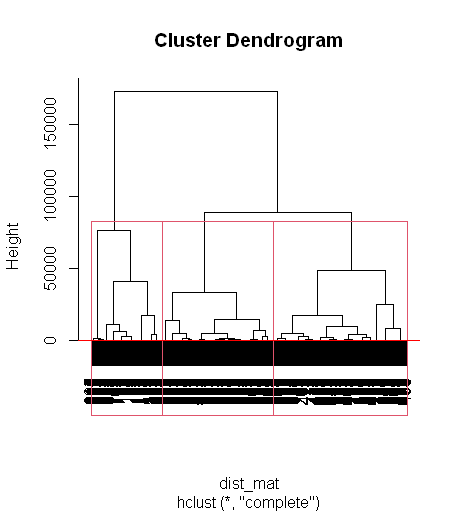
Para esta practica vamos a trabajar con el siguiente data set de abstinencia al trabajo obtenido de la siguiente dirección: <https://archive.ics.uci.edu/ml/machine-learning-databases/00445/> se nos indico que por contener muchos valores con 0 se podían limpiar o se podían trabajar con ellos, solo que nos iba a dar diferentes resultados, en este caso se opto por dejar esos datos para no modificar tanto al dataset.

Método Clúster:



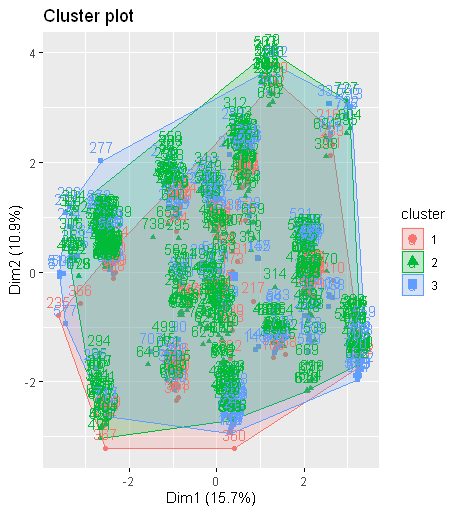
Para la distancia entre variables vamos a utilizar el método “manhattan” dado que este método no se ve tan afectado por outliers (Un outlier es una observación anormal y extrema en una muestra estadística o serie temporal de datos que puede afectar potencialmente a la estimación de los parámetros del mismo.) la elección de calcular las distancias de esta manera es por lo antes mencionado de conservas las filas que contiene uno o más ceros.

Se define la línea de corte a 3 grupos, ya que con diferente cantidad de grupos no siempre tenían una buena distribución o había grupos demasiado pequeños, la salida es la siguiente:



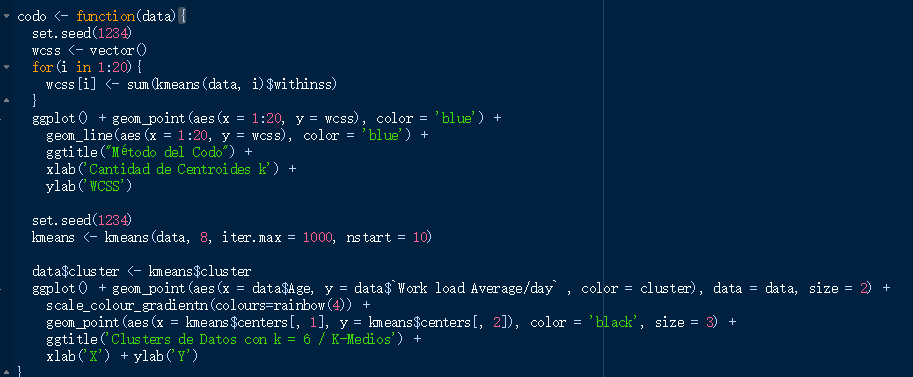
Algoritmo K-Means:

Para este algoritmo vamos igual a obtener las distancias con el método “manhattan” por las mismas razones antes mencionadas, vamos a dejar la cantidad de grupos en 3 con lo que obtenemos la siguiente gráfica:

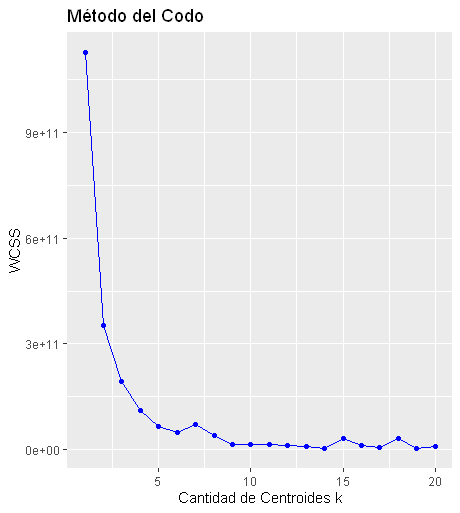


Después de obtener la gráfica podríamos decir que se podría incluir todo en uno solo, ya que no se nota muy claro la división entre cada grupo ya que están encimado uno de otro, pero lo dejamos de esta manera, ahora pasaremos al análisis para ver cuantos grupos son los recomendados para trabajar con el algoritmo de Clúster.

Algoritmo del codo:

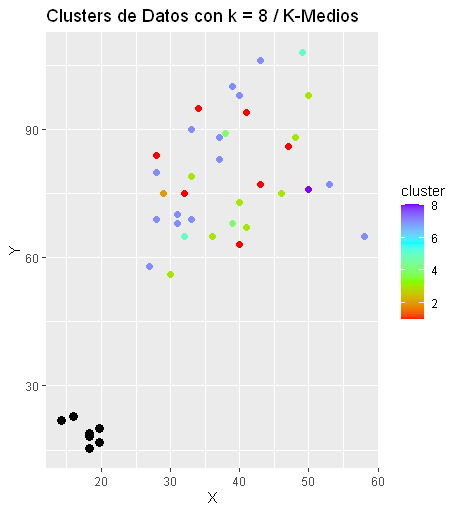


Para esta parte del código es necesario fijar una semilla y aplicaremos la función kmeans al conjunto de datos, variando en cada caso el valor de k, y acumulando los valores de WCSS para posterior graficar estos resultados, lo que nos da la siguiente gráfica:



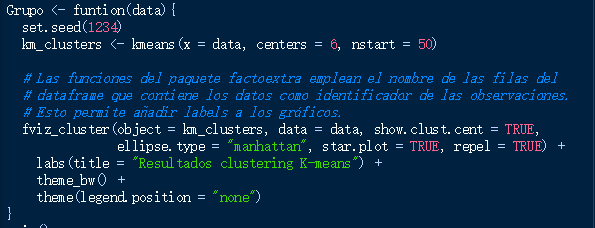
Para saber cuantos grupos son óptimos para trabajar, tenemos que escoger un punto en donde ya no se dejan de producir variaciones importantes del valor de WCSS al aumentar k. En este caso, vemos que esto se produce a partir de k >= 8

A partir de eso volvemos a graficar el data set, utilizando los valores de x = Age y los valores de y = Weight, con el valor de k = 8, lo que nos resulta en la siguiente gráfica:

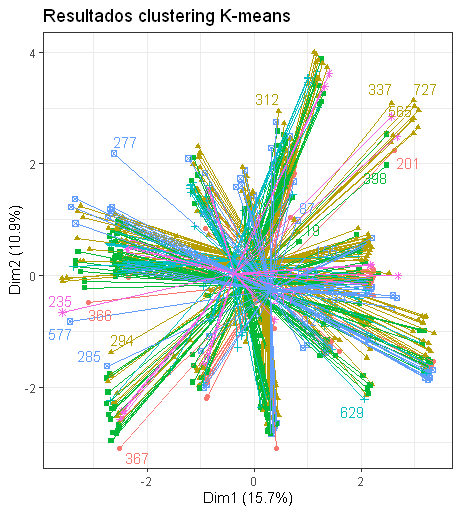


Ponemos la misma semilla que en la parte anterior y procedemos a graficar con cada color los atributos que corresponden a cada grupo.

Se puede sintetizar con el siguiente código:



Con ese código obtenemos la siguiente gráfica:



En teoría tendrían que verse cada grupo con los respectivos valores que pertenecen a dicho grupo, pero como se menciono anteriormente en la parte de k-means, todos los datos están muy juntos, al menos se diferencian por colores.

Conclusiones:

En esta práctica pusimos en practica el algoritmo de Clúster y el algoritmo de k-means, el cual es un algoritmo de agrupamiento donde se busca que cada grupo describa de mejor manera a cada uno de los integrantes, este se puede obtener de varias maneras, pero lo más usual es usarlo mediante la distancia entre sus componentes, lo cual se puede realizar por varios métodos, como:

* Distancia euclídea
* Distancia de Manhattan
* Jackknife correlation (Pearson)
* Indice Jaccard
* Distanica coseno

Después de obtener la matriz de distancia con alguno de los métodos anteriores se tienen que agrupar los valores, igual hay varios métodos para hacerlo, como:

* Método de agrupación de enlace simple
* Método de agrupación de enlace completo
* Método de agrupación de vinculación promedio
* Método de Ward

Los métodos de conglomerados jerárquicos proporcionan un dendograma que es la representación gráfica que permite visualizar las relaciones entre las distintas OTU. Obviamente, como en todo análisis multifactorial, algo de información se pierde durante el agrupamiento, pero es de suma utilidad para resumir una gran cantidad de información.

En está practica al parecer por algunos de los valores que tienen cero, seria lo que afecta en la visualización de los grupos, igual podría ser por los métodos seleccionado de distancia, en este caso yo use el método manhattan por que calcula la diferencia absoluta entre cada punto y es muy blanda con los parámetros que están o muy fuera de rango o variables que son demasiado volátiles.

Igual en la parte de averiguar cuales son los datos que pertenecen a cada grupo no se logra visualizar a que grupo pertenece cada valor, esto podría ser por los métodos seleccionado o por la forma de graficarlos.

Pero con más practica y con quizá los datos ya filtrados o eliminando esos datos de 0 se podrían obtener mejores resultados que los aquí presentados.