# Primera Parte

1. Indique 4 recomendaciones para escribir código en R y muestre un ejemplo de cada uno.

* Tener la buena práctica de documentar y comentar el script, utilizando el símbolo # al inicio de cada oración. Ejemplo: #Creación la variable suma.
* Tener presente que el lenguaje R discrimina entre las mayúsculas y las minúsculas. Ejemplo: es distinto utilizar un string llamado "Cartago" y otro llamado "CARTAGO".
* Las funciones deben escribirse seguidas de paréntesis ( ). Ejemplo: sum(1,2,3).
* Los corchetes se utilizan para referir a las posiciones. Ejemplo: En una matriz de Ventas [1,3] accederá a la posición de la fila uno y la columna 3.

2. Mencione los tipos de fuentes de datos y de 2 ejemplos para cada uno.

Los tipos de fuentes de datos son los siguientes:

* Estructurados: son los datos que poseen una estructura y formato bien definido, por lo que son sencillos de acceder. Como ejemplo tenemos una hoja de cálculo de Excel y una base de datos de SQL.
* No estructurados: como su propio nombre lo indica, son un tipo de fuente de datos que no pueden ser normalizados o estructurados. Como ejemplos tenemos a los datos contenidos en un documento de texto Word o un documento tipo PDF.
* Semi estructurados: se refiere a los datos que no poseen esquema alguno y se describen así mismos (auto descriptibles), básicamente los datos y la información de la estructura se muestran en sí mismo. Como ejemplos podemos mencionar los archivos XML y los HTML.

3. ¿Cuál es la diferencia entre el clustering jerárquico y el particional? Justifique su respuesta.

La diferencia entre clustering jerárquico y un clustering particional, radica en que el primero divide un conjunto de objetos de datos en un conjunto de datos que no se traslapan (no se posicionan encima uno sobre otro), con lo que cada objeto está en cada subconjunto (están en grupos separados), mientras que el clustering jerárquico identifica los subgrupos agrupados o anidados y los organiza en una estructura similar a un árbol, en donde básicamente un clúster es la unión de dos subcluster y su raíz contiene todos los objetos.

4. Explique en que consiste la maximización de distancias inter-cluster.

En el análisis del clustering, los objetos que se incluyen dentro de un mismo grupo tienen características similares entre ellas mismas que si las tuviera con otros grupos, y el concepto de maximización de inter-cluster muestra que entre más distancia posean esos grupos distintos más homogéneos serán los objetos.

5.Explique que es clúster contiguo (nearest neighbour) y de dos ejemplos (3 pts.)

Es un conjunto de puntos que es más similar a uno o más puntos en el clúster con respecto a otros como, por ejemplo:

* 1.La Base de datos de alumnos más entre sí que los alumnos de Turismo
* 2.La Base de datos de los ciudadanos de una provincia más que los ciudadanos de Costa Rica.

6. ¿Qué es la apofenia y en que podría afectar en un análisis de clustering? (2 pts.)

Es emitir un criterio basado en experiencias, no en el análisis, afecta de una manera errona, en que podemos tener patrones o datos sin sentido, tal como por ejemplo un gráfico de agrupación de colores y vemos que varios estudiantes dicen que hay un total de 12, otros estudiantes 15 y otras personas que 11.

7. ¿Cuál es la diferencia entre un vector y una matriz (2 pts.)

La diferencia es que un vector es un arreglo lineal de valores de un mismo tipo que se definen con esta función "c()" y la matriz es un arreglo de dimensiones, es decir, fila y columna, siendo sus valores del mismo tipo.

8.¿Que es un vector de tipo factor (2 pt)

Un vector de tipo factor es el vector de una variable de tipo nominal categórica, el cual, contiene un cierto orden y por lo tanto no se puede indicar cual categoría vale más que la otra.