Taller 1 - Big Data: Machine Learning

Jorge Gallego – Juan David Martínez

February 12, 2019

- Fecha de entrega: Hasta el miércoles 20 de febrero del 2019 23:59 horas. Enviar script debidamente explicado al correo: jdavidmartinezg@gmail.com.
- Si algún concepto no se explicó en clase Google es su mejor amigo. Especialmente Stack Overflow (La biblia de todo programador).

1 Vectores

a) Genere un vector vacío de tamaño 10 y asígnelo a una variable llamada *vec1*

b) Asigne a la primera posición del vector *vec1* su nombre en mayúsculas. Luego, asigne a la última posición del mismo vector su edad.

c) Genere un vector con todos los números del 1 a 10000 y asígnelo a una variable llamada *vec2*.

d) Utilice el comando *typeof()* para determinar el tipo de objeto para *vec1* y *vec2*. Explique porqué el tipo de cada vector es diferente.

2 Factores

a) Genere un vector lógico a partir del vector *vec2* llamado *vec3*. El vector *vec3* debe tomar el valor de TRUE cuando el número correspondiente a la misma posición en *vec2* es par. En caso de que el número sea impar *vec3* debe tomar el valor FALSE.

b) Convierta el vector *vec3* a un vector de factores.

3 Listas

a) Genere una lista vacía de tamaño 10 y asígnela a una variable llamada *list1*.

b) Asigne a la primera posición de la lista *list1* su nombre en mayúsculas. Luego, asigne a la última posición de la misma lista su edad.

c) Explique la diferencia entre *list1* y *vec1* usando el comando *typeof()*.

4 Matrices

a) Genere una matriz llamada *mat1* de tamaño 100x100 a partir del vector *vec2*.

b) Multiplique la primera fila de la matriz *mat1* por su primera columna. Asigne al resultado a un vector llamado *vec4*.

c) Reemplaze la diagonal de la matriz *mat1* por los valores del vector *vec4*.

5 Data frames

a) Descargue la base de datos *Precios.csv* aquí.

b) Abra la base de datos *Precios.csv* con R Studio. Asígnele a la base *Precios.csv* el nombre *df1*.

c) Use el comando *str()* para saber la estructura de la base *df1*.

d) Usando R encuentre cual es el nombre del producto de la base *df1* con el precio más alto.

e) Usando R encuentre cual es el precio promedio para los productos de la base *df1*.

f) Usando R determine si la variable *SKU* es un identificador único de cada producto en la base *df1*.

g) Usando R determine cual es la categoría con el mayor número de productos en la base *df1*.

h) Si una persona cuenta con un presupuesto de 56 mil pesos, ¿Qué productos de la base *df1* puede comprar?

i) Si una persona quisiera comprar todos los productos de la categoría *Teléfonos y celulares*, ¿Cuánto debería gastar?

j) ¿Cuántas marcas distintas tiene la base de datos *df1*?

k) ¿Cuáles son las 2 marcas con los productos más caros (en promedio)?

6 Gráficas

a) Haga el histograma de la variable precios para la categoría *Teléfonos y celulares*. Sobre la misma gráfica haga el histograma de la categoría *Relojes*. Muestre ambos histogramas de forma similar a como se puede ver en el siguiente ejemplo: imagen.

b) Haciendo uso de *ggplot2*, realice una gráfica de barras mostrando cuántos productos tiene la base *df1* por cada categoría (agrupe las categorías con un sólo producto en un grupo llamado *otras*). Ordene las barras de mayor a menor y utilice un color distinto para cada categoría.

c) Haciendo uso de *ggplot2*, realice una gráfica tipo *boxplot* mostrando la dispersión de precios por categoría. La gráfica debería mostrar las categorías en el eje x y el precio en el eje y. Ejemplo: imagen.

** Recuerde presentar cada gráfica con sus respectivos nombres de ejes, una leyenda y un título.