Profesor: Dr. Oldemar Rodríguez Rojas
Minería de Datos 1
Programación en R
Fecha de Entrega: Jueves 16 de enero del 2023 - 8am

Tarea Número 1

- Las tareas son estrictamente de carácter individual, tareas idénticas se les asignará cero puntos.
- Todas las tareas tienen el mismo valor en la nota final del curso, es decir, el promedio de las notas obtenidas en la tareas será la nota final del curso.
- Todos los ejercicios tienen el mismo valor.

1. Basados en la conferencia que está en el Aula Virtual: "Desde la Estadística hasta la Ciencia de Datos Pasando por el concepto de Big Data" responda las siguientes preguntas:

   a) Explique las diferencias entre: Estadística, Análisis de Datos, Minería de Datos, "Machine Learning", "Big Data" y Ciencia de Datos.

   b) Explique la relación entre: La evolución de la capacidad de almacenamiento y Explosión en la cantidad de datos.

2. Dado $x = (3, -5, 31, -1, -9, 10, 0, 18)$ y dado $y = (1, 1, -3, 1, -99, -10, 10, -7)$ realice lo siguiente:

   - Introduzca x y y como vectores en R.
   - Calcule la media, la varianza, la raíz cuadrada y la desviación estándar de `y`.
   - Calcule la media, la varianza, la raíz cuadrada y la desviación estándar de `x`.
   - Calcule la correlación entre `x` y `y`.
   - Escriba un comando en R para extraer las entradas 2 a la 7 de x.
   - Escriba un comando en R para extraer las entradas de y excepto la 2 y la 7.
   - Escriba un comando en R para extraer las entradas de y menores a -3 o mayores a 10.
   - Escriba un comando en R para extraer las entradas de x mayores a 0 y que sean números pares.

3. Introduzca en R la siguiente matriz a 4×3 usando:

   ```
   A = matrix(c(1,2,3,4,5,6,7,8,9,10,11,12),nrow=4,"byrow"="true")
   ```

   Luego, obtenga algunos elementos de la matriz de la siguiente manera: `A[1,1:3]`, `A[1:4,2]`, `A[3,3]`, `A[11]`, `A[20]`, `A[5,4]`, `A[1,1,1]` y explique qué pasa en cada caso.

4. Investigue para qué sirven los comandos de R `as.matrix(...)` y `as.data.frame(...)`, explique y dé un ejemplo de cada uno.

5. Introduzca usando código R (no archivos) en un `DataFrame` la siguiente tabla de datos:

Peso	Edad	Nivel Educativo
76	25	Lic
67	23	Bach
55	19	Bach
57	18	Bach
87	57	Dr
48	13	MSc

6. En muchas ocasiones nos interesa hacer referencia a determinadas partes o componentes de un vector. Defina el vector $x = (2, -5, 4, 6, -2, 8)$, luego a partir de este vector defina instrucciones en R para generar los siguientes vectores:

   - $y = (2, 4, 6, 8)$, así definido y es el vector formado por las componentes positivas de x.
   - $z = (-5, -2)$, así definido z es el vector formado por las componentes negativas de x.
   - $v = (-5, 4, 6, -2, 8)$, así definido v es el vector x eliminada la primera componente.
   - $w = (2, 4, -2)$, así definido w es el vector x tomando las componentes con índice impares, es decir, $x[1] = 2, x[3] = 4$ y $x[5] = -2$.

7. Queremos representar gráficamente la función coseno en el intervalo $[0, 2\pi]$. Para esto creamos el vector x de la siguiente forma `x<-seq(0,2*pi,length=100)`. ¿Cuál es la diferencia entre las gráficas obtenidas por comandos `plot`?

```
x<-seq(0,2*pi,length=100)
plot(cos(x))
plot(x,cos(x),col="red")
```

8. Para tabla de Datos que viene en el archivo `DJTable.csv` el cual contiene los valores de las acciones de las principales empresas de Estados Unidos en el año 2010, usando el comando `plot` de R, grafique (en un mismo gráfico) las series de valores de las acciones de las empresas CSCO (Cisco), IBM, INTC (Intel) y MSFT (Microsoft).

9. Repita el ejercicio anterior usando funciones del paquete `ggplot2`.

10. Cargue en un `DataFrame` el archivo `EjemploAlgoritmosRecomendacion.csv` usando el siguiente comando de `R`:

```
Datos <- read.table('EjemploAlgoritmosRecomendacion.csv',
                                    header=TRUE, sep=';',dec=',',row.names=1)
```

    y haga lo siguiente:

    - Calcule la dimensión de la Tabla de Datos.
    - Despliegue las primeras 2 columnas de la tabla de datos.
    - Ejecute un "summary" y un "str" de los datos.
    - Calcule la Media y la Desviación Estándar para todas las variables cualesquiera.

- Ahora repita los ítems anteriores pero leyendo el archivo como sigue:

```
Datos <- read.table('EjemploAlgoritmosRecomendacion.csv',
                                    header=TRUE, sep=';',dec='.',row.names=1)
```

  Explique porqué todo da mal o genera error.

11. Usando el archivo de datos `EjemploAlgoritmosRecomendacion.csv` realice lo siguiente:

   a) Grafique usando los comandos `plot` y `qplot` en el plano XY las variables `Entrega` vs `Precio`.

   b) Grafique usando comando `scatterplot3d` en 3 dimensiones las variables `Entrega`, `Precio` y `Durabilidad`.

   c) Usando el comando `cor` calcule la matriz de correlaciones de la tabla `EjemploAlgoritmos Recomendacion.csv` y grafique esta matriz de 4 formas diferentes.

   d) Usando el comando `Boxplot` encuentre los datos atípicos de la tabla de datos `Ejemplo AlgoritmosRecomendacion.csv`.

12. Cargue la tabla de datos que está en el archivo `SAheartv.csv` haga lo siguiente:

   - Calcule la dimensión de la Tabla de Datos.
   - Despliegue las primeras 3 columnas de la tabla de datos.
   - Ejecute un `summary` y un `str` de los datos.
   - Usando el comando `cor` de R calcule la correlación entre las variables tobacco y alcohol.
   - Calcule la suma de las columnas con variables cuantitativas (numéricas).
   - Calcule para todas las variables cuantitativas presentes en el archivo `SAheart.csv`: El mínimo, el máximo, la media, la mediana y para la variables `chd` determine la cantidad de `Si` y de `No`.

13. Programe en R una función que genera 200 números al azar entre 1 y 500 y luego calcula cuántos están entre el 50 y 450, ambos inclusive.

14. Desarrolle una función que calcula el costo de una llamada telefónica que ha durado t minutos sabiendo que si $t < 1$ el costo es de 0,4 dólares, mientras que para duraciones superiores el costo es de $0{,}4 + (t-1)/4$ dólares, la función debe recibir el valor de t.

15. Desarrolle una función que recibe una matriz cuadrada A de tamaño $n \times n$ y calcula su traza, es decir, la suma de los elementos de la diagonal. Por ejemplo, la traza de la siguiente matriz:

$$\begin{pmatrix} 9 & 3 & 4 \\ 1 & 3 & -1 \\ 4 & 12 & -2 \end{pmatrix}$$

   es 10.

16. Escribir una función que genere los n primeros términos de la serie de Fibonacci.

17. Escriba una función que retorne cuál es el mayor número entero cuyo cuadrado no excede de x donde x es un número real que se recibe como parámetro, utilizando `while`.

18. Crear un Data Frame con diez alumnos con su edad, año de nacimiento y número de teléfono. Deberá aparecer el nombre de la columna (edad, año de nacimiento, teléfono) y el nombre de la fila, que será el nombre del alumno al que corresponden los datos.

19. Programe funciones en **R** para calcular:

    - El número de permutaciones con repetición de r objetos tomados de n.
    - El número de permutaciones sin repetición, o arreglo, de r objetos tomados de n.
    - El número de permutaciones con repetición de n objetos, de los cuales solo k son distintos.
    - El número de combinaciones es un subconjunto desordenado de r objetos seleccionados en un conjunto que contiene n.
    - El número de combinaciones con repetición es un subconjunto desordenado de r objetos seleccionados de un conjunto que contiene n en los cuales se pueden repetir.
    - El número de particiones de n objetos en r clases, el cual se conoce como el **número de Stirling de segundo tipo**.

20. Desarrolle una función **R** que recibe un **DataFrame** que retorna la cantidad de entradas de este DataFrame que son divisibles entre 3.

21. Desarrolle una función **R** que recibe un **DataFrame** y dos números de columna y que retorna en una lista el nombre de las variables correspondientes a las columnas, la covarianza y la correlación entre esas dos variables.

22. Importe directamente desde Excel en **R** el archivo `EjemploAlgoritmosRecomendación.xlsx` el cual contiene los promedios de evaluación de 100 personas que adquirieron los mismos productos o muy similares en la tienda AMAZON. Luego ejecute un `str(...)` y un `summary(...)` con esta tabla de datos.

23. Programe la siguiente función recursiva:

$$U(n) = \begin{cases} 5 & si \quad n = 0 \\ -5 & si \quad n = 1 \\ 2 & si \quad n = 2 \\ 4U_{n-1} - 15U_{n-2} + U_{n-3} & si \quad n \geq 3 \end{cases}$$

24. Programe la siguiente función $f(x) = x^n$ de forma recursiva.

25. El archivo `bosques_energia.csv` el cual contiene el resultado de dos estudios, uno en el que se mide la superficie boscosa en proporción al terreno total de cada país y el otro corresponde al consumo de energía renovable en proporción al consumo de energía total. Estos estudios se hicieron en varios países y en varios años distintos. Con estos datos realize lo siguiente:

    a) Convierta los datos a datos `tidy`.

    b) Elimine las variables innecesarias y explique el motivo.

c) Con la ayuda del paquete `dplyr`, realice lo siguiente:

- Con un gráfico muestre la evolución del consumo de energía renovable para los países Canada, Paraguay, Perú y China. Debe mostrar tanto el gráfico como la tabla de datos con la que se realizó el gráfico.
- Con un gráfico muestre los 10 países con mayor superficie bosocosa promedio para los años analizados. Debe mostrar tanto el gráfico como la tabla de datos con la que se realizó el gráfico.

26. El archivo `DatosEducacion.csv` contiene información de las escuales primarias de varios países durante los años 2013 a 2019. Las variables están por filas, los valores de dichas variables están en forma columna por año.

    Nota: No olvide revisar el archivo con un bloc de notas.

    Cargue la tabla de datos y luego realice lo siguiente:

    a) Convierta el dataset a uno `tidy`. Elimine las variables innecesarias y los valores con `NA`.

    b) Agrupe el dataset por país y promedie los resultados, no incluya la variable fecha. Además, cambie los nombres de las variables a unos más 'cortos'.

    c) Construya una variable con el porcentaje de estudiantes que repiteron el año y otra con el gasto público bruto si se repartiera en partes iguales a cada estudiante. Luego elimine las variables que se utilizaron para crear las 2 variables anteriores. Por último, repita los 2 ejercicios anteriores con estos nuevos datos.

    d) Con base a los resultados obtenidos, responda a las siguientes preguntas:

    - De los 2 resultados obtenidos anteriormente, ¿Cuál le hace más sentido, según los países agrupados y sus características?
    - ¿A que se debe que sean tan diferentes los 2 resultados obtenidos?

Entregables: Debe entregar un documento autreproducible `HTML` con todos los códigos y salidas.