MÁSTER UNIVERSITARIO EN BIOINFORMÁTICA Y BIOESTADÍSTICA

SOFTWARE PARA EL ANÁLISIS DE DATOS (SAD)

Prueba de Evaluación Continua 1 (PEC1)

CONTENIDO Y ORIENTACIONES

La PEC1 comprende los conceptos básicos de la introducción a la programación en R.
Para realizar la práctica, es necesario haber trabajo los siguientes laboratorios y recursos asociados:

- LAB 1. Introducción a R.
- LAB 2. Estadística descriptiva con R y Rcommander.
- LAB 3. Programación en R.

La PEC1 consta de diversos ejercicios organizados en secciones que corresponden a cada uno de los contenidos de los diferentes laboratorios.

Adjunto al enunciado, encontraréis los ficheros que debéis utilizar para realizar la PEC1.
Por otra parte, los ejercicios se desarrollaran en el entorno Rstudio y Rcommander.

FORMATO DE ENTREGA

Se entregará una carpeta comprimida apellido1_apellido2_nombre_SAD_PEC1.zip que contendrá:

- El documento de solución en pdf o formato RMarkdown.
- El documento de código en R.

FECHA LÍMITE DE ENTREGA

La fecha límite de entrega de la PEC1 es el **29 de Octubre de 2019 a las 23.59 h.**

PEC 1: ENUNCIADO

A lo largo de la práctica utilizaréis ficheros de datos para realizar los ejercicios. Uno de ellos, *lung_cancer_examples.csv* ha sido extraído de la plataforma Kaggle: https://www.kaggle.com/yusufdede/lung-cancer-dataset, el resto de conjunto de datos pertenecen a paquetes propios de Rstudio.

Sección 1. Importación, exportación y gestión de datos (2 puntos)

El objetivo de esta sección es importar y exportar diferentes archivos de datos y visualizarlos desde Rstudio.

Ejercicio 1.

Resuelve las siguientes cuestiones, mostrando las instrucciones de código utilizadas así como el resultado de la ejecución de dicho código:

1.1. Importad los datos del fichero *lung_cancer_examples.csv* correspondientes a la probabilidad de que un paciente tenga cáncer de pulmón a partir del estudio de algunos parámetros. Guardad estos datos en un data frame llamado *dataLungCancer* y mostrad los primeros y últimos registros de este conjunto de datos.

1.2. A partir del data frame definido, *dataLungCancer*, mostrad algunas características como:

a) Nombre de las variables que forman el conjunto de datos.

b) Estructura del conjunto de datos, es decir, tipo de de variables.

c) Tamaño de la muestra y número de variables.

d) ¿Existen valores nulos en el conjunto de datos? ¿Y en la variable *Result*?

e) ¿Existen datos perdidos (missing values) en la tabla?

1.3. Exportad el data frame *dataLungCancer* a un fichero en formato texto. Explicad cómo se realizaría con RStudio y con RCommander.

Ejercicio 2.

Resolved las siguientes cuestiones, mostrando las instrucciones de código utilizadas así como el resultado de la ejecución de dicho código:

2.1. Instalad el paquete *survival* de RStudio, cargad el conjunto de datos *lung* y guardadlo en un data frame llamado *dataLungCSurv*. Posteriormente, mostrad los primeros y últimos registros, así como el nombre de las variables, la estructura del conjunto de datos y el número de registros.

2.2. Definid un data frame, *dataLungCSurv_w*, y otro data frame,*dataLungCSurv_m*, que corresponderán a los conjuntos de datos de las mujeres y de los hombres, respectivamente. Posteriormente, exportad ambos data frame a dos ficheros (LungCSurv_w,LungCSurv_m) de tipo *csv*.

Sección 2: Estructuras de datos y análisis. (2,5 puntos)

El objetivo de esta sección es definir diferentes estructuras de datos realizar operaciones y consultas sobre un conjunto de datos.

Ejercicio 3

3.1. A partir del conjunto de datos correspondientes al data frame *dataLungCancer*, se pide:

a) ¿Cuál es la media de edad, que definiremos como *age_mean*, de los pacientes de la muestra de datos?

b) Definid una variable, *age_max_smoke*, que guarde la edad del paciente que registra el máximo valor de cigarros fumados por día.

c) Mostrad los diagnósticos resultantes (Result=1 (afectado), Result=0 (no afectado) de aquellos pacientes que superan la media de cantidad de cigarros fumados y de tomas de alcohol?

d) Definid un data frame,*paciente_result_0*, que contenga los pacientes que no estan afectados (Result=0). A partir de este data frame, definid un vector que contenga los valores màximos de las variables Age, Smokes, AreaQ y Alkhol.

3.2. A partir del conjunto de datos correspondientes al data frame *dataLungCSurv*, se pide:

a) Definid una matriz que muestre las columnas *sex* y *age*.

b) Mostrad los datos de los hombres (*sex=1*) del conjunto de datos que fallecieron (*status=2*).

c) En base a los resultados del apartado anterior, de los hombres que fallecieron, comprobad si el paciente que registraba mayor pérdida de peso poseía también un valor de *ph.ecog* superior a 3. Tened en cuenta que, en el caso de la variable *wt.loss* aparecen *missing values* que se aconseja eliminar para realizar cálculos.

Sección 3: Estadística descriptiva y gráficos. (3 puntos)

El objetivo de esta sección es estudiar los conceptos relacionados con estadística descriptiva en R y realizar diferentes tipos de gráficos en RStudio y, según el caso, en RCommander.

Ejercicio 4.

A partir del data frame *dataLungCancer* se pide resolver las siguientes cuestiones que deberéis resolver en RStudio y Rcommander:

4.1 Realizad un resumen estadístico de *dataLungCancer* que muestre los parámetros básicos más importantes.

4.2. Definid un vector con las edades (*Age*) de los pacientes y otro vector con el resultado (*Result*) del diagnóstico. Etiquetad la variable *Result* con “Afectado” si el valor es 1 y “No afectado” si el valor es 0.

4.3. A partir del vector de las edades, ordenad el vector, calcular la media y la varianza y desviación estándar.

4.4. Mostrad las tablas de frecuencias relativas y absolutas de los vectores definidos al apartado 4.2. Posteriormente representad una tabla de frecuencias relativas cruzadas, de manera que podamos visualitzar cuántos pacientes según edad, estan o no afectados.

4.5. Realizad los siguientes gráficos:

a) Diagrama de tallo y hojas de la variable *age*.
b) Diagrama de cajas y bigotes de la variable *smoke*.
c) Histograma de la variable *Alkhol*.
d) Diagrama de puntos de la variable *Alkhol*.
e) Combinad los anteriores gráficos en una representación gráfica común. Podéis utilizar la función layout() para ajustar la distribución de los gráficos.

Ejercicio 5.

A partir del data frame *dataLungCSurvr* y utilizando el paquete de gráficos ggplot2, se pide resolver las siguientes cuestiones:

a) Realizad un gráfico de tipo qplot de la variable *age* i la variable *wt.loss*.
b) Realizad un gráfico de barras para las dos variables *age* y *status*.

Ejercicio 6.

A partir del conjunto de datos *dataLungCancer*, realizad un breve estudio de regresión y correlación lineal, resolviendo las siguientes cuestiones:

a) Realizad un diagrama de cajas sobre las variables *Alkhol* y *Age*. ¿Qué se intuye de este gráfico?
b) Realizad un modelo de regresión para las dos variables anteriores y un diagrama de puntos que ajuste al modelo anterior. ¿Qué conclusiones obtenéis?
c) Realizad la matriz de correlación del conjunto de datos *dataLungCancer*. ¿Qué podéis afirmar?
d) Calculad los residuos del modelo ajustado anterior y realizad un gráfico de normalidad.
e) De los apartados anteriores, ¿qué podéis concluir?

Sección 4: Programación en R: Instrucciones condicionales y repetitivas (2,5 puntos)

Los contenidos a evaluar en esta sección corresponden a las instrucciones condicionales y repetitivas en R. No es necesario definir funciones para resolver los ejercicios.

Ejercicio 7.

A partir del paquete *MASS* de Rstudio, seleccionamos el conjunto de datos *Melanoma*. Desde la pestaña Package de RStudio, se puede acceder al contenido del paquete *MASS* y, por tanto, a la descripción de cada conjunto de datos.
Se pide:

a) ¿Cuántos hombres y cuántas mujeres han realizado las pruebas?
b) ¿Cuál es la media de edades de aquellos pacientes que murieron por melanoma?