UNIVERSIDAD DE LAS PALMAS
DE GRAN CANARIA

Escuela de Ingeniería Informática

Métodos Estadísticos, Curso 2019-20

Ejercicios Prácticos Lab1 y Lab2

Ejercicio 1: Analizar con el comando ***search*()** los paquetes presentes en el entorno de trabajo. Con ***library***(help=package), seleccionar el paquete ***datasets***, y, dentro de los distintos conjuntos de datos, visualizar en la consola los contenidos de varios de ellos con distintas características (tipos de variables, series, etc.).

a) Analizar cómo están estructurados los datos para familiarizarse con ellos.
b) Distinguir claramente en su contenido aquellos que contengan factores y vectores.
c) Visualizar y direccionar su contenido y realizar algunos cálculos sencillos sobre el mismo.
d) Generar, utilizando **R Markdown,** un report de laboratorio que recoja la sesión y explicar en él los resultados que se han obtenido. Utilizar aquellos trozos de código R empotrados (code chunks) con sintaxis ***knitr*** que se consideren necesarios para este fin.

Alguno de estos "data sets" pueden ser utilizados como parte experimental del proyecto o trabajo de curso. Para otros paquetes puede consultarse el siguiente link https://vincentarelbundock.github.io/Rdatasets/datasets.html

Ejercicio 2: El Data Set "***MplsStops"*** de la librería ***carData*** contiene datos de incidencias de personas implicadas en actuaciones policiales por el Departamento de Policía de Minneapolis en 2017. Se pide:

a) Analizar su contenido y visualizar los factores y vectores.
b) Explicar el uso del comando ***subset***() y emplearlo para obtener un subconjunto de este data set que con tenga los vectores *race*, *gender* y *neighborhood* para el caso de actuaciones derivadas de accidentes de tráfico:

```
datos_seleccionados<-subset(datos[problem=="traffic",],
                            select = c(race, gender, neighborhood))
```

c) Utilizando el comando ***ftable***() analizar los diferentes porcentajes de accidentes de tráfico según raza y género.
d) Visualizar con el comando gráfico ***pie***() los resultados del apartado anterior.
e) Encontrar en qué zona de Minneapolis se registraron más accidentes.

Ejercicio 3: Utilizar el Data Set ***"Davis"*** de la librería ***carData***, que proporciona los datos de hombres y mujeres que realizan ejercicio regularmente de peso y altura, tanto medidos como comunicados por los/las afectados/as. El Data Set contiene datos no disponibles (NA´s). Analizar la estructura de los datos correspondientes y:

a) Estudiar y aplicar posibles soluciones para los NA´s.
b) Encontrar las variaciones de altura y peso reales en función del género. Calcular las medias, medianas y desviación estándar correspondientes.
c) Analizar las variaciones de altura y peso comunicadas en función del género. Calcular las medias, medianas y desviación estándar correspondientes.
d) Visualizar gráficamente, utilizando ***boxplot***(), una comparativa de los datos de peso medido y peso declarado por un lado y de la altura medida y la altura declarada por otro. Establecer justificadamente las conclusiones.
e) Encontrar si hay diferencias significativas entre lo medido y declarado según el género y analizar las posibles formas de corregirla

Ejercicio 4 (Opcional): Utilizar la siguiente secuencia de comandos para leer los ficheros "*empleados.txt*" y "*salarios.txt*".

```
setwd("C:/Users/Antonio/Documents/R/Scripts R")
empleados <- read.table("empleados.txt",sep = ",", header = TRUE)
salarios <- read.table("salarios.txt",sep = "\t",header = TRUE)
names(empleados)
[1] "Num_Empleado" "Fecha_nacimiento" "Nombre" "Apellido" "Genero" "Fecha_Contrato"
names(salarios)
[1] "Num_Empleado" "Salario" "Desde_Fecha" "Hasta_Fecha"
```

Estos ficheros contienen los datos de los empleados y salarios de una empresa de Ingeniería vinculados por un campo común "*Num_Empleado*".

a) Analizar el contenido de los Data Frames con los comandos ***tail***() y ***head***()
b) Razonar sobre los tipos de datos que lo integran (factores y vectores).
c) Encontrar las medias, medianas y desviaciones estándar de la variable "Salario" agrupada por la variable "*Num_empleado*" y encontrar el empleado que más cobra y el que menos en promedio.
d) Visualizar utilizando ***boxplot***() las variaciones de *"Salario"* dependiendo del empleado.
e) Utilizar el comando ***merge***() para unir los dos data frames unificados por "*Num_empleado*" y repetir los apartados c) y d) para el data frame resultante.
f) Con los comandos ***interval***() , ***now***() y ***ymd***() del paquete **lubridate**, determinar la edad de los diez empleados y añadir una nueva columna con el campo *"Edad"* al data frame resultante del apartado anterior
g) Anadir un nuevo registro al data frame del apartado e). Explicar en detalle el proceso.

Ejercicio 5: Ejercicio: Leer el fichero "***casas.txt***" que incluye el precio medio de viviendas en miles de euros por localizaciones en España. Generar un vector "Precios" a partir de los datos indicados en el fichero. Realizar a continuación las siguientes operaciones:

```
A<-rank(Precios)
B<- sort(Precios)
C<- order(Precios)
Comparativa<-data.frame(Precios,A,B,C)
Comparativa
```

Explicar la diferencia entre las diferentes columnas que resultan en cada caso y obtener las casas de precio medio superior a 190.000 €

Ejercicio 6: El fichero "***Accidentes_1969_1984_UK.txt***" contiene datos de series temporales referidas a conductores fallecidos o con lesiones graves en UK entre los años 1966 y 1984. En enero de 1983 entró en vigor la ley que obliga a la utilización del cinturón de seguridad. Entre otras variables se dispone de las siguientes:

- *DriversKilled* : conductores de automóvil muertos.
- *front*: Pasajeros asientos delanteros muertos o gravemente heridos.
- *rear*: Pasajeros asientos delanteros muertos o gravemente heridos.
- *VanKilled*: número de conductores de furgonetas
- *law*: vigencia (0/1) de obligatoriedad del cinturón

Se pide:

a) Analizar la serie temporal de fallecidos en accidentes, encontrar sus zonas de máximo valor y visualizar el efecto de entrada en vigor de la ley.
b) Analizar las relaciones existentes entre los conductores fallecidos y las víctimas según estuvieran en los asientos delanteros o traseros. Explicar y estudiar en detalle el alcance de las suposiciones establecidas en los posibles modelos.
c) Analizar y evaluar el efecto que tienen las furgonetas ligeras (tipo Van) en el conjunto de accidentes mortales antes y después de la aplicación de la ley. Justificar las respuestas

Ejercicio 7: El fichero "*Ventas_Provincia.txt*" contiene datos de ventas en euros de una empresa productora de cereales a distintas provincias españolas durante el año 2012. Se desea realizar un análisis de estos datos para valorar los procesos. Se pide:

a) Cantidades totales y las medias anuales de ventas por provincia.
b) Provincia en la que más se vende y en la que menos.
c) Estudiar la evolución de las ventas de las provincias de Cáceres, Madrid y Barcelona en el segundo semestre de 2012.
d) Utilizando los comandos gráficos de base de R, visualizar la evolución temporal de los datos del apartado c)
e) Alternativamente, utilizando ***ggplot2***() realizar una visualización de la evolución mensual de los datos del apartado c), tanto absolutos como relativos al total de ventas de la empresa. Explicar las distintas soluciones adoptadas.
f) Realizar cambios en la estética, la escala y el tema en el apartado e). Explicar las ventajas y diferencias en cada caso.

Ejercicio 8 (Opcional): Diseñar con los gráficos de base de ***R*** una función que rellene el área entre una curva dada por un vector y el eje de ordenadas, desde una posición inicial (***xlimite1***) a otra final (***xlimite2***). Se deberá incluir:

a) Parámetros de entrada: el vector de ordenadas ***x*** y de abscisas ***y***, las posiciones inicial y final de las ordenadas (*xlimite1*, *xlimite2*) y el color de relleno y el tipo. Opcionalmente puede incluirse una function como entrada que establezca la relación ***y***=f(***x***).
b) Comprobación de los parámetros de entrada que garanticen la estabilidad de la función.
c) Introducir mejoras en la llamada de la función con la posibilidad de incluir algún texto adicional en alguno de los ejes o en la propia gráfica.
d) Plantear una función alternativa utilizando ***ggplot2***().
e) Introducir mejoras en la estética, el tema y la escala en la propia llamada de la función.
f) Explicar el uso de ambas funciones en al menos tres casos de ejemplo

Ejercicio 9 (Opcional): El fichero "*Datos_hsb.txt*" contiene datos de resultados de aprendizaje de un estudio longitudinal de estudiantes USA no universitarios (High School and Beyond, 1980, National Center for Education Statistcs) relativos a sus progresos en varias materias (lectura, escritura, matemáticas, sociales y ciencias) y clasificados por categorías (sexo, raza, estatus socioeconómico, tipo de centro y de programa). Se desea realizar un análisis de estos datos utilizando ***ggplot2***() y combinado con diferentes estadísticos resumen, para ello se pide:

a) Calcular las medias de progreso por las distintas categorías y razonar posibles conclusiones.
b) Evaluar la influencia del centro en los resultados de aprendizaje de la época
c) Estudiar posibles relaciones entre la capacidad de lectura y el aprendizaje de otras disciplinas.
d) Visualizar si el sexo o la raza de origen en 1980 tenían influencias significativas en el nivel de aprendizaje.