EJERCICIOS R - PRINCIPIANTES

1. Obtener la frecuencia absoluta y relativa del siguiente conjunto de datos:

a,b,c,a,b,d,e,f,a,b,c,f,g,a

2. Sobre la siguiente serie de números: 20,22,22,27,28,31,33,33,35,36,39,39,39,41,42
Obtener la media aritmética, la mediana, la moda, los cuartiles, la varianza y la desviación típica.

3. Sobre la serie de números del ejercicio 2 añadimos los valores 53 y 67. Indicar si los valores añadidos deberían tratarse como outliers en base a un diagrama de boxplot.

4. La siguiente tabla contiene datos de ventas por año de una compañía. Para cada año tenemos las ventas totales, los gastos en marketing y el número de productos diferentes vendidos ese año.

Año	Ventas totales	Gastos_mkt	Número productos
2008	1000	120	50
2009	1200	130	52
2010	1400	135	54
2011	1500	140	44
2012	1200	120	43
2013	1100	120	48
2014	1300	130	46
2015	1400	140	48

Obtener la covarianza entre ventas totales y gastos en marketing y entre ventas totales y número de productos. Obtener conclusiones acerca de la dependencia entre las variables.

5. Partiendo de los datos de la actividad 4 obtener la correlación de Pearson entre:

a. ventas totales - gastos de marketing
b. ventas totales – productos

6. Tenemos la siguiente serie de datos con edades de clientes:

22,24,26,31,34,36,37,38,39,40,43,47,51

Realizar una normalización de esta variable para que una vez normalizada la variable creada tenga media aritmética 0 y desviación típica 1.

7. En esta actividad vamos a practicar el manejo de vectores con R. Crear los siguientes vectores:

- *vec1*: vector de 5 elementos enteros: 10,20,30,15,5
- *vec2*: vector de cadenas de caracteres con 3 elementos: “curso”, “programación”, “R”
- *vec3* como resultado de multiplicar los elementos de *vec1* por 2
- *vec4* como resultado de restar *vec1* a *vec3*
- *vec5* que contenga los elementos 1 y 3 de *vec2*
- *vec6* que contenga los elementos de *vec3* que sean mayores o iguales a 20.
- Obtener longitud de *vec6*.
- *vec7* que contenga los elementos de *vec2* cuya longitud sea mayor o igual a 4.

8. En esta actividad vamos a practicar el manejo de matrices con R Crear los siguientes objetos:

- *mat1*: matriz de 3 filas y 3 columnas de números enteros
- *vec1*: vector de 5 posiciones de números enteros
- *cad1*: cadena de caracteres
- Crear una lista lista1 que contenga los tres objetos anteriores
- Obtener el valor del tercer elemento de la lista
- Obtener el tercer elemento del vector de la lista

9. En esta actividad vamos a practicar el manejo de dataframes con R. Crear un data.frame con las columnas: cliente, periodo y consumo en base a estos tres vectores:

clientes <- c("id001", "id002", "id003","id001", "id002", "id003");
periodos<- c(201506,201506,201506,201507,201507,201507)
consumo<- c(20,30,50,10,20,40)

Realizar las siguientes acciones sobre el data.frame:

a. Visualizar el data.frame

b. Obtener el número de registros del data.frame

c. Crear una nueva columna que contenga el año

d. Obtener un subconjunto con los clientes con consumo >= 20

e. Obtener un data.frame con el consumo medio a nivel de cliente

f. Sobre el data.frame del punto anterior obtener una nueva columna con el ratio entre el consumo medio de cada cliente y el consumo medio total

g. Sobre del data.frame inicial obtener histograma de la variable consumo

h. Crear un data.frame con la descripción de los periodos

datafper<- data.frame(periodos=c(201506,201507),
desc_periodo = c("Junio 2015","Julio 2015"))

i. Cruzar con el data.frame inicial para añadirle la columna desc_periodo

j. Obtener agregado sobre el data.frame resultado del paso anterior a nivel de desc_periodo

k. Partiendo del data.frame del punto anterior obtener gráfico que muestre el consumo por periodos

10. En esta actividad vamos a practicar con estructuras de control: bucles y condicionales en R.

a. Crear dos vectores de 10 elementos: *vec1* con valores en vacío y *vec2* con valores que siguen una distribución normal con valor medio 5 y desviación estándar 1. Asignar a vec1 los valores de los elementos de *vec2* restándole a cada valor el valor medio de los elementos del vector *vec2*.

b. Crear un vector de elementos introducidos por teclado. Crear un segundo vector como resultado de restar a cada elemento del vector el elemento anterior.

c. Crear una matriz 4 x 4 con elementos que tengan una distribución normal con 10 de valor medio y 2 de desviación estándar. Sumar los elementos de la diagonal de la matriz.

d. Crear dos matrices 4 x 4 cuyos elementos que tengan una distribución normal con 10 de valor medio y 2 de desviación estándar. Crear una tercera matriz 4x4 como resultado de restar los elementos de las dos anteriores.

e. Crear una matriz 4 x 4 con dos posibles valores en los elementos:

"*Madrid*" y "*Lisboa*" generar los valores de forma aleatoria con un 60% probabilidad para "*Madrid*" y un 40% para "*Lisboa*". Una vez generada la matriz contar el número de apariciones de cada valor.

f. Obtener el traspuesto de la siguiente matriz, sin utilizar la función t() *mat1 <- matrix(c(6,3,7,32,14,9,1,5,13,5,21,2), nrow=4, ncol=3)*

11. En esta actividad vamos a practicar la creación y ejecución de funciones con R.

a. Crear una función que reciba 3 valores enteros y cree un vector con estos elementos recibidos. Validar que no haya nulos en los valores recibidos.

b. Crear una función que reciba tres vectores de 4 elementos y cree un data.frame con cada uno de los vectores como columna. Validar que todos los vectores tienen 4 elementos.

c. Crear una función que aplique la siguiente función: *y = 3*x + 5*

Aplicarla sobre todos los elementos de un vector de 20 elementos con valores entre 1 y 5.

d. Crear un data.frame partiendo de estos vectores :

ciudades<- c("Madrid","Madrid","Madrid","Lisboa","Lisboa","Lisboa")
productos<- c("P1","P1","P2","P3","P3","P4")
ventas<- c(20,22,11,10,15,20)

Empleando la función *tapply* obtener el valor medio de ventas en base a ciudades y productos. En la matriz resultado sustituir los valores NA por 0.

e. Crear una función que aplique la siguiente relación de valores:

1. 1 - 'A'
2. 2 - 'B'
3. 3 - 'C'
resto - 'D'

Aplicarla sobre una matriz 4 x4 cuyos elementos tengan valores entre 1 y 4.

12. En esta actividad vamos a realizar los pasos habituales del análisis exploratorio de datos con R.

a. Cargar el data.frame iris. iris es un data frame con 150 registros y 5 columnas: *Sepal.Length*, *Sepal.Width*, *Petal.Length*, *Petal.Width*, y *Species*. Contiene datos de 50 flores de 3 especies

b. Obtener los 20 primeros registros y una muestra aleatoria de 50.

c. Obtener tabla de frecuencia del campo *Species*.

d. Obtener el rango de valores, el valor medio y el cuartil 50 del campo *Petal.Length*

e. Obtener histograma y boxplot de *Sepal.Width*

f. Validar si hay nulos en algún campo del data.frame *Iris*

13. En esta actividad vamos a practicar el uso de gráficos. Cargar el siguiente data.frame que muestra las ventas por periodo:

periodos<- seq(201501:201509) + 201500
vper<- data.frame(periodo=periodos, ventas =
round(rnorm(9,100,20)))

a. Crear un gráfico de barras para mostrar las ventas a lo largos de los periodos, mostrar las barras en un único color (por ejemplo verde)

b. Cambiar el gráfico de barras para mostrar cada barra de un color, pasar los valores de los colores como un vector utilizando las funciones que generan escalas de

colores: *rainbow(N), heat.colors(N), gray.colors(N), topo.colors(N)*, etc…, donde N es el número de elementos del vector de colores.

c. Crear un gráfico de tarta que muestre la distribución de ventas por periodo. Mostrar cada porción en un color.

d. Crear un data.frame como la siguente:

	Cliente1	**Cliente2**	**Cliente3**	**Cliente4**	**Cliente5**
201501	97	104	115	83	80
201502	80	95	84	90	118
201503	132	167	130	94	104
201504	92	94	108	86	72
201505	106	92	124	129	90
201506	128	96	115	89	103
201507	107	58	113	91	105
201508	119	89	87	105	78
201509	119	109	91	85	121

El valor numérico se refiere a las ventas de un periodo para un determinado cliente. Tanto las columnas, como las filas del data.frame deben estar etiquetadas. Partiendo de esta tabla mostrar un mapa de calor. Mostrar el gráfico con el color por defecto y después cambiarlo a escala de grises.

Crear un gráfico de mosaico partiendo de esta tabla.

e. Gráfico que represente las ventas de una compañía por países.

1. Crear una tabla que tenga ventas por países, para los siguientes países: España, China, Alemania y Japón. Para obtener los códigos de los países incorporar la tabla countryExData que pertenece a la librería rworldmap.

2. Basándonos en la librería rworldmap, pintar un mapa del mundo con los países con ventas resaltados en colores

f. Utilizando la librería ggmap obtener un mapa de una determinada zona geográfica y localizar 3 puntos sobre el mapa. Para obtener la longitud y la latitud se puede utilizar la web: http://www.latlong.net/

14. En esta actividad vamos a practicar la regresión lineal en R. Queremos conocer si en una empresa existe una relación entre el gasto en publicidad y sus ingresos. La variable G tiene el histórico de gasto anual en publicidad (2005-2016) y la variable I tiene los ingresos de la empresa cada uno de esos años

G<-c(20,25,21,30,22,23,19,24,21,23,28,27)

I<-c(229,235,230,242,231,233,226,232,230,232,238,236)

Pintar el gasto en el eje de coordenadas y el ingreso en abscisas. Obtener el grado de correlación entre las variable y realizar un modelo de regresión lineal simple que permita estimar los ingresos para los próximos 5 años tiendo en cuenta que se va a realizar la siguiente inversión en publicidad anual:

GP <- c(30,32,34,36,38)

15. En esta actividad vamos a practicar diferentes algoritmos de machine learning para realizar predicciones con R. Partiendo del dataset *SAheart* de la librería *ElemStatLearn* realizar predicciones utilizando dos tipos de algoritmos: árbol decisión y regresión logística.

El dataset *SAheart* tiene datos de un estudio sobre enfermedades del corazón. Tiene 462 registros y 10 variables: *edad, consumo de alcohol,tabaco, historial familiar,obesidad, ldl(colesterol)* y *typea. chd* es la variable respuesta que indica si se produce enfermedad del corazón o no. Crear 2 modelos para predecir la variable chd uno en base a árbol de decisión y otro en base a regresión logística.

16. En esta actividad vamos a practicar el uso de series temporales en R. Partiendo del dataset *co2* (serie histórica de valores de concentraciones Co2 que incluye de 1959 a 1997 tomados en la estación rusa Vostok) de la librería *forecast*. Vamos a realizar el siguiente tratamiento con la serie temporal

a. Descomponer la serie en: tendencia, estacionalidad y ruido

b. Pintar en una gráfica las diferentes partes de la serie temporal

c. Validar si hay autocorrelación en la serie.

d. Realizar una predicción empleando el algoritmo arima. Tomamos como intervalo para entrenamientos los datos de 1959 a 1994 y predecimos de 1995 a 1997. Pintar una gráfica con la estimación y calcular la precisión de la estimación.