Ejercicio Propuesto 1 (Resuelto)

Una empresa de comercio electrónico desea medir el volumen de gasto mensual en euros en función de las reseñas de los clientes, los cupones y el número de visitas a la web de la empresa. Vamos a utilizar el archivo <u>GASTOS1.RData</u> que contiene el gasto mensual de los clientes en euros, números de reseñas positivas realizadas por el cliente, los cupones promocionales disponibles para el cliente y el número de visitas realizadas a la página web.

Se desea clasificar a los 24 clientes encuestados por el número de visitas realizadas a la página web (Visitas), el gasto mensual de los clientes en euros (Gasto) y los cupones promocionales disponibles para el cliente (Cupones)

Se pide:

- **1. Realizar un diagrama de dispersión 3-D para mostrar la distribución de los datos y estudiar los posibles grupos que se pueden hacer.**
- 2. Utilizar un análisis clúster jerárquico. (Etiquetar los casos mediante Identificación personal, id)
 - I. Método: Vecino más lejano; Medida: Intervalo- Distancia euclídea al cuadrado; Transformar valores: Estandarizar las variables (puntuaciones Z)
 - II. Obtener la Matriz de distancia y el Dendograma
 - III. Analizar las tablas obtenidas y sacar conclusiones

Solución

Considerando las variables de interés gastos y número de visitas a la página web, repetir el análisis hecho en el ejercicio propuesto 3. Para realizar el apartado 1.

- Seleccionar en el menú principal Gráficas/Gráficos 3D/Diagrama de dispersión en 3D
- Variables explicativas: visitas y cupones; Variable explicada: gasto.
- Pestaña Opciones: Identificar Observaciones Interactivamente con el ratón.

En primer lugar:

> library(Rcmdr)

1. Realizar un diagrama de dispersión 3-D para mostrar la distribución de los datos y estudiar los posibles grupos que se pueden hacer.

En primer lugar tenemos que cargar los datos

Fichero	Editar	Datos Estadi	ticos Gráficas Modelos Di	stribucione	s Herramientas Ayud	a	
R	Conjunto	Nuevo con Cargar con	unto de datos unto de datos		ditar conjunto de datos	🔯 Visualizar conjunto de datos	Modelo: 2
R Script	R Markd	Fusionar co	njuntos de datos				
		Importar d	tos		desde archivo de text	to, portapapeles o URL	
	Conjunto de datos en paquetes		desde datos SPSS				
		Conjunto de datos activo		•	desde un archivo SAS	S exportado	
		Modificary	riables del conjunto de datos a	tivo 🕨	desde un archivo SAS	5 b7dat	
		-			desde datos Minitab.		
					desde datos STATA	(a)	
					desde un archivo de l	Excel	

R Leer archivo de texto, portapapeles o UR	L X
Introducir el nombre del conjunto de datos	: Gastos
Nombres de las variables en el archivo:	
Convertir variables de carácter en factores	\checkmark
Indicador de datos ausentes:	NA
Localización del archivo de datos	
Sistema de archivo local	
O Portapapeles	
O Dirección URL	
Separador de campos	
Espacio en blanco O Comas [,]	
○ Punto y coma [;] ○ Tabuladores	
O Otro Especificar:	
Carácter decimal	
Punto [.]	
O Coma [,]	
🔯 Ayuda 🛛 🚽 Aceptar	💥 Cancelar

Pulsamos Aceptar y elegimos el dichero: Gastos.xlsx Gastos.xlsx

Para elaborar el gráfico, tenemos que ir a la opción *Gráficas*, seleccionar Gráficos 3D, Diagrama de dispersión en 3D. Ver la siguiente imagen

Gráficas	Modelos Distribuciones Herrami	entas Ayuda	
Gama	a de colores	onjunto de datos	Modelo: Σ <no hay="" m<="" td=""></no>
Gráfic Diagr Histo Dibuj Estim Gráfic Diagr Gráfic	ca secuencial ama de puntos igrama iar una variable numérica discreta iar densidad ca de tallos y hojas rama de caja ca de comparación de cuantiles cana de caja de rimetría	EA/DataR/Jove	enes.RData")
Diagr Matri Gráfic Gráfic Gráfic Diagr	rama de dispersión iz de diagramas de dispersión ca lineal ca XY ca XY ca de las medias rama de puntos		
Gráfic	ca de barras ca de sectores	_BEA/DataR/J	ovenes.RData")
Gráfie	cos 3D	 Diagrama de dis 	spersión en 3D
Guard	dar gráfico en archivo	 Identificar obse Guardar gráfico 	rvaciones con el ratón en archivo

Se abrirá una ventana donde se muestra dos cuadros para seleccionar las *Variables explicativas* (se selecciona las variables cupones y visitas) y en *Variable explicada*, se selecciona la variable gasto.

Variables explicativa Cupones	as (elegir dos)	Variable explicad Cupones	a (elegir una)	
Basto		Gasto		
Reseñas		Reseñas		
<u>iisitas</u>	~	Visitas	~	
Gráfica por grupos.				

Por último, en *Opciones*, seleccionamos en *identificar Observaciones*, Interactivamente con el ratón. Terminamos con *Aceptar*.

😨 Diagrama de dispersión en 3D	×
Datos Opciones	
Número de rotaciones automáticas	0 🚖
Mostrar escala de los ejes	
Mostrar líneas de rejilla en la superficie	
Mostrar residuos cuadráticos	
Superficies a ajustar	
Mínimos cuadrados lineales	
Mínimos cuadrados cuadráticos	
Robust linear regression	
Regresión suavizada	gl = <auto></auto>
Regresión aditiva	gl(cada término) = <auto></auto>
Dibujar elipsoide que contiene el 50% de los datos	s 🗌
Color de fondo	
Negro 🔘	
Blanco 🔘	
Identificar Observaciones	
 Automáticamente 	
Interactivamente con el ratón	
○ No identificar	
Número de puntos a identificar 🛛 2 🚖	
🔞 Ayuda 🤚 Reiniciar 🚽	Aceptar 🗱 Cancelar 🥐 Aplicar

Se obtiene el siguiente gráfico



Cuando se obtiene el gráfico, presiona el lado derecho del ratón y con el cursor podrás rotar el gráfico.



Observando el gráfico podemos observar 3 grupos de sujetos considerando estas variables.

- 2. Utilizar un análisis clúster jerárquico. (Etiquetar los casos mediante Identificación personal, id)
 - IV. Método: Vecino más lejano; Medida: Intervalo- Distancia euclídea al cuadrado; Transformar valores: Estandarizar las variables (puntuaciones Z)
 - V. Obtener la Matriz de distancia y el Dendograma
 - VI. Analizar las tablas obtenidas y sacar conclusiones

Primero, vamos a estandarizar las variables con las que vamos a trabajar: Cupones, Visitas, Gasto, y trabajamos con las variables estandarizadas. Para estandarizar las variables, vamos a la opción *Datos, Modificar variables del conjunto de datos activo, Tipificar variables*.

R Commander	
Fichero Editar Datos Estadísticos Gráficas Modelos Distribucio	ones Herramientas Ayuda
Conjunto Conjunto de datos Cargar conjunto de datos	
R Script R Markd Fusionar conjuntos de datos	
scatter3d(t Importar datos ellipsoid Conjunto de datos en paquetes	<pre>SE, residuals=TRUE, bg="white", axis.scale build and build an</pre>
<pre>scatter3d(t ellipsoid scatter3d(tv~fútbol+paga, data=Jovenes, surface= ellipsoid=FALSE, id=list(method='mahal', n =2)</pre>	Recodificar variables F2 Añadir números de observaciones al conjunto de datos
<pre>scatter3d(tv~fútbol+paga, data=Jovenes, surface= ellipsoid=FALSE, id=list(method='identify')) <</pre>	FA Tipificar variables Le Convertir variable numérica en factor Convertir variables de caracteres en factores
	Reordenar niveles de factor
Salida	Descartar niveles sin uso Definir contrastes de un factor
<pre>> library(mgcv, pos=17)</pre>	Renombrar variables
<pre>> scatter3d(tv~fútbol+paga, data=Jovenes, surface)</pre>	ce: Liiminar variables dei conjunto de datos ca

Seleccionamos las variables que se va a estandarizar. Luego damos en *Aceptar* y se guardarán las variables estandarizadas en la base de datos.

R Tipificar variables		×
Variables (seleccione	una o más)	
Cupones	~	
Gasto		
ID		
Reseñas		
<u>Visitas</u>	\sim	
😳 Ayuda	🚽 Aceptar	🐺 Cancelar

En segundo lugar, para agrupar o clúster los datos, vamos a seleccionar la opción **Estadísticos, Análisis dimensional, Análisis de agrupación, Agrupación jerárquico**

Estadísticos Gráficas Modelos Distribuciones Herramientas Ayuda	
Resúmenes Tablas de contingencia Medias	s Modelo: 2 <no activo="" hay="" modelo=""></no>
Proporciones surface=FALSE, residuals=TRUE, Varianzas ify')) Test no paramétricos respective	bg="white", axis.scales=TRUE, grid=TRUE,
Análisis dimensional	g="white", axis.scales=TRUE, grid=TRUE,
Ajuste de modelos Analisis de componentes principales Análisis de componentes principales Análisis de componentes principales Análisis de componentes principales Análisis confirmatorio factorial	g="white", axis.scales=TRUE, grid=TRUE,
bol+paga, data=Jovenes, Análisis de agrupación ▶ E, id=list (method='ider	Agrupación por k-medias
	Agrupacion Jerarquica Resumir la agrupación jerárquica Agregar la agrupación jerárquica al conjunto de datos

En la pestaña **Datos**, seleccionamos las variables estandarizadas (ZCupones, ZGasto, ZVisitas).

R Agrupación jerárquica	×
Datos Opciones	
Nombre del resultado de agrupación: HClust.2	
Variables (seleccione una o más) ID Reseñas	
Z.Gasto	
Expresión de selección <todos casos="" los="" válido<br="">< ></todos>	
🔞 Ayuda 🦘 Reiniciar 🖌 Aceptar 🎇 Cancelar	Aplicar

En *Opciones*, seleccionamos en *Método de agrupación: Enlace completo* y en *Medida de la distancia: Euclídea al cuadrado*. Para terminar, damos en *Aceptar*.



Se pulsa aceptar y se muestra el siguiente Dendograma.



Cluster Dendrogram for Solution HClust.1

Observation Number in Data Set Dataset Method=complete; Distance=squared-euclidian

Teniendo en cuenta este dendograma y el diagrama de dispersión obtenido podemos pensar en una solución de tres clústeres.

- Clúster 1: (3, 12, 15, 13, 6, 10, 16, 7, 17, 22)
- Clúster 2: (2, 20, 4, 5, 11, 19,)
- Clúster 3: (1, 9, 21, 8, 23, 14, 18, 24)

Por último, tenemos la matriz de distancias:

> d <- dist(model.matrix(~-1 + Z.Cupones+Z.Gasto+Z.Visitas, Gastos))^2
> d

