Máster Universitario en Ciencia de Datos
Fundamentos para el Análisis de Datos y la Investigación.
Prácticas y Tareas 1. Curso 2023-24. Prof. JM Sarabia

1. Análisis Exporatorio de datos

1. El fichero de datos `DatosD1_1.txt` incluye información numérica sobre la renta per capita y el nivel de educación de un conjunto de países.

   (i) Obtener medidas de posición, dispersión y forma.

   (ii) Obtener el histograma junto con el estimador kernel. Diagramas de caja.

   (iii) Comprobar el efecto de cambiar la función kernel y el ancho de banda en la estimación de la función de densidad.

   (iv) Representar conjuntamente las dos variables con los datos originales y mediante alguna transformación.

2. Realizar un análisis exploratorio con la variable `Education`, que se encuentra en la base de datos `Swiss` análogo al anterior (apartados (i) a (iii)), cambiando las opciones de las diferentes funciones.

3. A partir del fichero `survey` de la libraría `MASS` se pide:

   (i) Identificar variables cuantitativas y categóricas (factores)

   (ii) Representar dos variables cuantitativas mediante `plot(x,y)` y a continuación identificar los niveles de una variable categórica

   (iii) Representar variables cuantitativas mediante `hist(x)` y mediante `boxplot`

   (iv) Representar variables cuantitativas según niveles de algún factor

4. Los siguientes datos se refieren a la renta de una muestra de nueve individuos en dos localidades (en miles de euros)

   A: 13, 25, 15, 7, 12, 38, 42, 53, 7
   B: 4, 23, 36, 18, 39, 20, 9, 45, 12

   Representar las curvas de Lorenz. Obtener los índices de Gini mediante el paquete `ineq`. Obtener los índices de Atkinson y de Entropía para algunos valores de los parámetros.

5. Haciendo uso de paquete `simFrame`, la base de datos `eusilcP` y la variable: `eqIncome`:

   (i) Representar gráficamente la variable anterior. Obtener medidas de desigualdad basadas en cocientes de cuantiles.

   (ii) Haciendo uso del paquete `ineq`, calcular el índice de Gini y los índices de desigualdad de Atkinson y entropía generalizada según valores del parámetro.

   (iii) Considerar una nueva variable prescindiendo de valores por encima de 100.000. Calcular índice de Gini.

6. Crear un vector de caracteres de 160 votantes de cuatro partidos P1,...,P4 con distribución de voto de 40, 60, 50 y 10 respectivamente.

(i) Representaciones gráficas

(ii) Crear una tabla con los votos según porcentajes. Ordenar la tabla según porcentajes de voto.

7. La base de datos `Wage` (se encuentra en el paquete ISLR) ofrece información sobre un grupo de 3000 trabajadores hombres en una determinada región

   (i) Representar la variable `wage` según los niveles de educación, raza, salud y estado civil mediante box plot. Obtener el salario medio según los niveles de las variables categóricas.

   (ii) Estudiar de forma aislada alguna variable categórica. Construir una tabla de contingencia de dos variables categóricas

   (iii) Representar el salario `wage` en función de la edad, y a continuación obtener el mismo gráfico según los niveles de la raza, salud y estado civil.

8. Un Data Scientist se encuentra un Boxplot como el de la figura 1. ¿Qué conclusiones puede obtener?

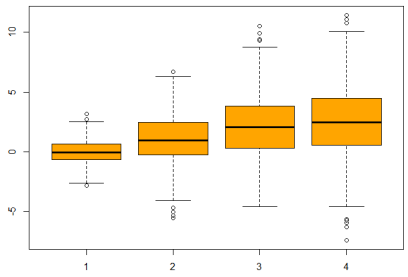

Figura 1: Box plot de cuatro variables

9. Para la base de datos `Swiss` realizar un análisis exploratorio multivariante completo, con los siguientes elementos:

   (i) Realizar un análisis exploratorio univariante, incluyendo histogramas y estimadores núcleo de las variables marginales

   (ii) Obtener medidas exploratorias multivariantes

   (iii) Visualización de los datos, mediante diversos métodos

10. La base de datos iris se refiere a un conjunto de datos que contienen información sobre 50 muestras de cada una de tres especies de Iris (setosa, virginica y versicolor), y fueron estudiados por R.A. Fisher.

    (i) Leer la base de datos, indicar las variables y los niveles de las tres especies. Obtener diagramas de caja.

    (ii) Seleccionar las 10 primeras observaciones.

    (iii) Seleccionar las variables Sepal.Length y Sepal.Width.

(iv) Seleccionar los datos correspondientes a las especiles setosa, virginica.

(v) A continuación, seleccionar los casos 1 a 3 y 98 a100.

(vi) Realizar gráficos en dos variables, y a continuación distinguiendo especies.

(vii) Realizar gráficos tridimensionales. Diagramas 3D y de coordenadas paralelas

11. El fichero de datos `Cotizaciones2020.txt` contiene las series de Precios de cierre ajustados diarios de IBEX 35, Banco Santander, BBVA, Repsol, Inditex y MediaSet desde julio 2019 a julio 2020.

    (i) Obtener las series de rendimientos de los precios en tiempo discreto. Estudiar las propiedades estadísticas de los rendimientos. Definir una variable categórica covid, que permita calcular los rendimientos antes y después del *covid*, eligiendo un punto de corte.

    (ii) Obtener los rendimientos en tiempo continuo y ver si existen diferencias con los antes obtenidos.

    (iii) Obtener los histogramas y los diagramas de caja sin y con la variable *covid*.

    (iv) Diagramas de dispersión del rendimiento del IBEX frente al del Santander. Diagramas de dispersión por parejas de rendimientos en un diagrama matricial.

    (v) Obtener las matrices de covarianza y de correlaciones

12. Datasets mediante Yahoo Finance

    (i) Seleccionar series de datos de cotizaciones del yahoo! finance del MCE `https://es.finance.yahoo.com/`.

    (ii) Elegimos los datasets `MEL.MC.csv`, `MAP.MC.csv` y `IBEX.csv`. Leer estos datasets de Yahoo Finance en formato csv, construir dataframe y realizar algunos análisis. Construir un modelo de regresión simple CAPM para explicar los rendimientos de MAP en términos de los rendimientos del IBEX.

13. Mediante el paquete de R `quantmod` (Quantitative Financial Modelling and Trading Framework for R):

    (i) Descargar los datos de Amazon (AMZN) durante los dos últimos años. Obtener la serie de los rendimientos, suponiendo tiempo continuo y representarla.

    (ii) Obtener el rendimiento medio, cuartiles, desviación típica y coeficientes de asimetría y curtosis

    (iii) Hallar el histograma y el estimador kernel de la serie de rendimientos.

14. Mediante el paquete de R `quantmod`, seleccionar activos nacionales e internacionales y realizar análisi exploratorio básico.

15. A partir de la base de datos `mtcars`:

    (i) Seleccionar las variables c(1,3,4,5,6,7) y obtener matrices de covarianzas y de correlaciones. Obtener autovalores y autovectores de la matriz de correlaciones. Interpretar los resultados.

    (ii) Obtener la significación estadística de la matriz de correlaciones.

(iii) Representar gráficamente la matriz de correlaciones.

(iv) Obtener una matriz con toda la información relevante del subconjunto seleccionado.

16. El paquete `NHANES` incluye información sobre una gran encuesta de salud realizada por el US National Center for Health Statistics (NCHS).

    (i) Realizar histogramas del peso y la altura según género

    (ii) Obtener gráficos bidimensionales del peso (altura) y edad según género, incluyendo una función de suavizado. Realizar gráficos del peso y altura edad según raza.

    (iii) Obtener medidas de resumen de las variables principales según niveles de algunas variables categóricas.

17. Mediante el paquete `vioplot`, realizar gráficos del tipo box plot y de violin, con las base de datos `Iris`, considerando una variable categórica.

18. Mediante la base de datos `USArrests` de tasas de criminalidad en USA

    (i) Obtener la matriz de distancias usando las variables contenidas en los datos

    (ii) Obtener la matriz de distancias entre ciudades

    (iii) Realizar un dendograma mediante cluster jerárquico y criterio de agregación `ave`.

19. Se trata de construir mapas de calor `heatmap` en `R` con diversos paquetes y el dataset `swiss`

    (i) Haciendo uso del módulo base

    (ii) Por medio del paquete `gplots`

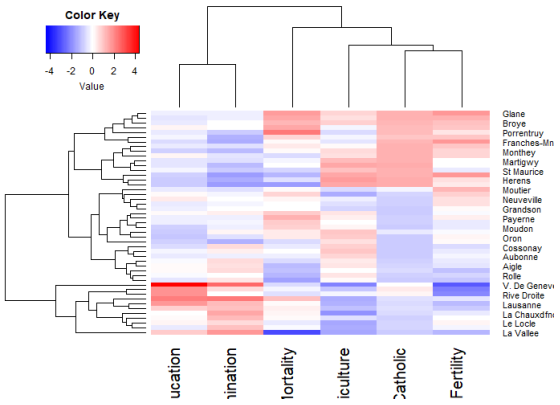

Figura 2: Heatmap de Swiss

20. Uso de gráficos para bigdata don `ggplot2`.

    (i) Obtener mediante simulación un dataset de bigdata normal con dos variables

    (ii) Dibujar los datos mediante plot y mediante diagramas exagonales y de contorno haciendo uso de `ggplot2`

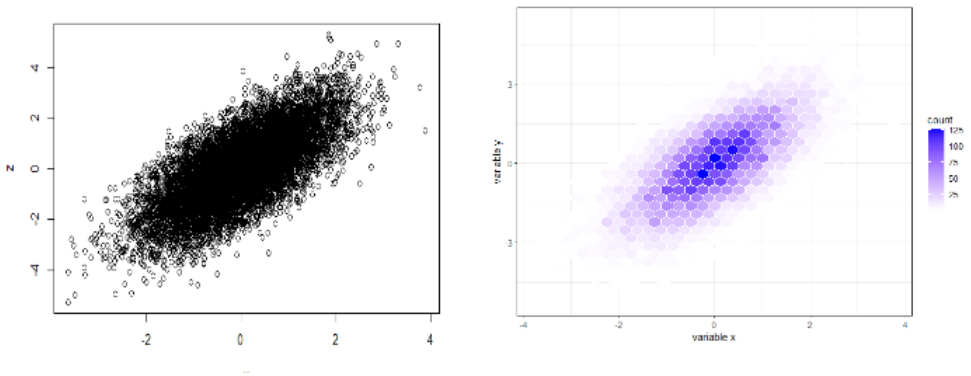

Figura 3: Plot clásico y exagonal para muestra de bigdata

2. Fundamentos de matemáticas en ciencia de datos

2.1. Fundamentos de álgebra en ciencia de datos

1. Dadas las matrices:

$$\mathbf{A} = \begin{pmatrix} 3 & 5 & 8 \\ 5 & 7 & 1 \\ -4 & 8 & -3 \end{pmatrix}, \ \mathbf{B} = \begin{pmatrix} -4 & 3 & 0 \\ 7 & 5 & 5 \\ 9 & 9 & -2 \end{pmatrix}$$

   Realizar las siguientes operaciones básicas:

   (i) $\mathbf{A} + 3\mathbf{B}$, $2\mathbf{A} - \mathbf{B}$

   (ii) $\mathbf{A}^\top\mathbf{B}$ y $\mathbf{B}^\top\mathbf{A}$. ¿Qué relación hay entre estas dos matrices?

   (iii) Hallar $\mathbf{A}^{-1}$ y $\mathbf{B}^{-1}$ y comprobar $(\mathbf{A}\cdot\mathbf{B})^{-1} = \mathbf{B}^{-1}\cdot\mathbf{A}^{-1}$.

   (iv) Medias de $\mathbf{A}$ y $\mathbf{B}$ por filas y por columnas

2. El rendimiento de un activo tiene volatilidad del 12 % y la de un segundo activo es del 18 %, con rendimientos medios del 5 y del 8 por ciento, respectivamente. La correlación entre rendimientos es 0.45. Se construye una cartera lineal donde 2 millones se invierten en el activo 1 y 3 millones en el activo 2. Hallar lel rendimiento medio y la volatilidad de la cartera.

3. Dado el sistema de ecuaciones:

$$\begin{aligned} 3x - y - 2z &= 2 \\ 8x + 4y + 4z &= 3 \\ 5x + y + z &= 7 \end{aligned}$$

   (i) Comprobar que tiene solución

   (ii) Obtener la solución mediante R

4. Dada la matriz

$$\mathbf{A} = \begin{pmatrix} 1 & 1 & 3 \\ 5 & 2 & 6 \\ -2 & -1 & -3 \end{pmatrix}$$

(i) Comprobar que $\mathbf{A}^3 = \mathbf{0}$.

(ii) Sustituir la tercera columna de $\mathbf{A}$ por la suma de las columnas dos y tres.

5. Un data scientist ha obtenido la matriz 10×10 definida como $\mathbf{A} = \{i^2 + j^2 + 1\}$, donde $i, j = 0, 1, \ldots, 9$

(i) Definir la matriz en `R`

(i) Obtener el determinante. ¿Cuál es su rango? Obtener los autovalores de la matriz. ¿Se puede tratar de una matriz definida positiva?

6. Un científico de datos construye la matriz $\mathbf{A} =$ `matriz(rnorm(30,2,1),5,5)`. A partir de la matriz A construir las matrices: $\mathbf{A}\mathbf{A}^\top$, $\mathbf{B} = \mathbf{A}\mathbf{A}^\top + \mathbf{I}$, así como la inversa de $\mathbf{B}$.

7. **Análisis de Regresión.** Una agencia inmobiliaria posee una cartera de 5 casas en una misma zona. Tenemos: y es el precio en miles de euros; x_1 metros cuadrados y x_2 número de habitaciones:

$$\mathbf{y}^\top = (122, 115, 128, 145, 108),$$
$$\mathbf{X}^\top = \begin{pmatrix} 10 & 12 & 12 & 14 & 10 \\ 4 & 4 & 5 & 5 & 3 \end{pmatrix}$$

Consideramos el modelo $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}$ estimado mediante mínimos cuadrados. Obtener:

(i) $\hat{\beta} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$. Si $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$, hallar el vector de residuos $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$

(ii) Probar que $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{P}\mathbf{y}$ donde $\mathbf{P}$ es la matriz,

$$\mathbf{P} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$$

denominada *hat matrix* (porque hace cambiar la $\mathbf{y}$ a y con $\hat{\mathbf{y}}$) y que el vector de residuos se puede escribir como

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I}_n - \mathbf{P})\mathbf{y}$$

Probar que $\mathbf{P}$ e $\mathbf{I}_n - \mathbf{P}$ son simétricas e idempotentes.

(iii) Verificar que la suma de cuadrados de los residuos se puede escribir como:

$$\mathbf{e}^\top\mathbf{e} = \mathbf{y}^\top(\mathbf{I}_n - \mathbf{P})\mathbf{y}$$

(iv) Realizar estos resultados incluyendo término independiente en el modelo

(v) Comprobar los resultados mediante el comando `lm` de R

8. Disponemos de una matriz 3×3

$$\mathbf{A} = \begin{pmatrix} 8 & -2 & -3 \\ -2 & 11 & 4 \\ -3 & 4 & 5 \end{pmatrix}$$

(i) Obtener la descomposición espectral. Comprobar sus propiedades.

(ii) Rango de $\mathbf{A}$.

(iii) Determinante y traza de $\mathbf{A}$.

(iv) Hallar los autovalores y autovectores de las matrices $\mathbf{A}^{-1}$ y $\mathbf{A}^2$.

9. Un análisis financiero dispone de la siguiente información histórica relativa al rendimiento de tres activos correlacionados.

	Activo 1	**Activo 2**	**Activo 3**
Media rend.	0.05	0.09	0.08
Desv. típ.	0.20	0.30	0.25
Matriz correlaciones	1	0.90	0.75
	0.90	1	0.5
	0.75	0.5	1

   (i) Obtener la matriz de covarianza y descomponerla según Choleski.

   (ii) Simular 30 muestras de los rendimientos de los tres activos usando la descomposición anterior. Comprobar sus propiedades.

10. Hallar la descomposición de valores singulares (SVD) de la matriz 4×3,

$$\mathbf{A} = \begin{pmatrix} 2 & 7 & 6 \\ 3 & -1 & 4 \\ 0 & 5 & 1 \\ -2 & 3 & -2 \end{pmatrix},$$

    Obtener las tres matrices de la descomposición, y comprobar que las propiedades de la descomposición.

11. Hallar la descomposición de valores singulares de la matriz 8×4,

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & 6 & -7 \\ 2 & 3 & 7 & -8 \\ -1 & -2 & 9 & -7 \\ 3 & 2 & 11 & -9 \\ 0 & 1 & 3 & 2 \\ 0 & 1 & 1 & 2 \\ 0 & 1 & 1 & 4 \\ 0 & 1 & 2 & 1 \end{pmatrix}$$

2.2. Fundamentos de cálculo en ciencia de datos

1. Representar las siguientes funciones mediante `plot`, `lines` y `curve`:

   (i) $f_1(x) = \frac{e^{-(x-a)}}{(1+e^{-(x-a)})^2}$ para $a = 0$ y a continuación $a \in \{-2, -1, 0, 1, 2\}$

   (ii) $f_2(x) = \frac{1}{1+(x-a)^2}$

   (iii) $f_3(x) = \sin(3\cos(x)e^{-x^2/10})$, en el intervalo $(-8, 5)$ y a continuación en el intervalo $(-9, 9)$

2. Hallar el mínimo de la función $y = -\exp(-(x-5)^2)$ mediante `optimize`.

3. Un análisis de datos desea estimar la función de densidad $f(x;\theta) = \theta\exp(-\theta x)$ donde $x > 0$ y $\theta > 0$.

   (i) Escribir la función logaritmo de la verosimilitud (LL).

(ii) Hallar el máximo de LL a partir del vector de datos `x = rexp(10000, rate=5)` (`set.seed(2020)`). Comprobar que se trata del mismo resultado haciendo uso de la librería `MASS` y la función `fitdistr`.

4. Dada la función $f(x, y) = (x^2 + y - 11)^2 + (x + y^2 - 7)^2$,

   (i) Representarla usando los comandos `outer` y `persp`. Identificar puntos críticos

   (ii) Usando Obtener los mínimos usando `optim`

5. Estamos trabajando con la función

$$f(x, y) = -2(x^2 + y^2) + (x^4 + y^4) - 11$$

   Probar que tiene 9 puntos críticos, de los cuáles 4 son mínimos, 1 máximo y el resto puntos de silla.

6. Se trata de encontrar las dimensiones (x, y, z) de una caja rectangular abierta que contenga un volumen de 4 uc dentro de la menor superficie posible.

   (i) Probar que la función objetivo es $f(x, y) = xy + \frac{8}{x} + \frac{8}{y}$

   (ìi) Probar que f alcanza un mínimo en $(2, 2)$ empezando en $(0{,}1, 0{,}1)$

7. Obtener el mínimo de la función de dos variables $f(x, y) = (1 - x)^2 + 100(y - x^2)^2$

8. Hallar el mínimo de la función $2000x + 3000y$ sujeta a $20x + 10y \geq 200$, $25x + 50y \geq 500$ y $18x + 24y \geq 300$ con $x, y \geq 0$. Plantear el problema dual y obtener su solución.

9. Un inversor dispone de un capital de 50 millones para invertir en fondos de riesgo alto, medio y bajo con rentabilidades del 12, 6 y 2 por ciento, respectivamente. Tiene pensado invertir al menos 5 millones en fondos de riesgo medio y entre 9 y 11 millones en fondos de riesgo bajo. Tiene pensado que la inversión en fondos de riesgo alto y medio debe estar como máximo a razón de 4 a 5. Obtener la inversión óptima que maximiza la rentabilidad.

10. Un grupo inversor invertirá una cantidad C en una cartera formada por seis fondos de diferentes características, teniendo en cuenta una serie de restricciones de diversificación

Fondo	1	2	3	4	5	6
Rentabilidad esperada (%)	30	20	15	12	10	7
Riesgo	Alto	Alto	Alto	Medio	Medio	Bajo

    Se trata de obtener la proporción de capital C a invertir en cada fondo. El grupo está dispuesto a invertir entre el 50 y el 75 por ciento en fondos de alto riesgo, entre 20 y 30 por ciento en fondos de riesgo mediano y menos de un 5 por ciento en fondos de riesgo bajo. Con objeto de diversificar, la cantidad invertida en los fondos de alto riesgo 1, 2 y 3 deben estar en proporciones 1:2:3 y la inversión en los fondos de riesgo mediano 4 y 5 en proporciones 1:2. Obtener la distribución de la cartera que maximiza el rendimiento esperado (Mathur and Solow, Management Science, the art of decision making, Prentice-Hall).

11. Una forma habitual de trabajar con datos de renta es ajustarles a una curva de Lorenz poblacional, es decir, una función normalizada, creciente y convexa. Un data scientist tiene que ajustar los datos de quintiles de renta

`https://www.wider.unu.edu/project/wiid-world-income-inequality-database`

a partir de la informaicón propocionada por el Banco Mundial, incluida en la siguiente tabla.

population	RepDom-2015	Brazil-2015	Nic-2015	Peru-2014
quintiles	shares	shares	shares	shares
20 más bajo	0.05	0.03	0.05	0.04
20 segundo más bajo	0.09	0.08	0.09	0.1
20 tercero	0.14	0.13	0.14	0.15
20 segundo más alto	0.21	0.2	0.2	0.21
20 más alto	0.51	0.56	0.52	0.5

Para los siguientes cuatro modelos de curvas de Lorenz

$$\begin{aligned} y_1 &= (1-(1-x)^a)^{1/a}, \;\; 0 \le a \le 1, \\ y_2 &= 1-(1-x^a)^{1/a}, \;\; 0 \le a \le 1, \\ y_3 &= x^a, \;\; a \ge 1, \\ y_4 &= \frac{\log(\cos(x\arctan(a)))}{-\log\sqrt{1+a^2}}, \;\; a \ge 0 \end{aligned}$$

donde $0 \le x \le 1$,

(i) Comprobar que se trata de modelos genuinos, es decir, que cumplen las condiciones de curvas de Lorenz

(ii) Ajustar los modelos mediante `nls` y seleccionar el mejor de los cuatro mediante SSE.