Fundamentos para al Análisis de Datos y la Investigación

Tema 2 - Análisis Exploratorio de Datos (2/2)

José María Sarabia

Máster Universitario en Ciencia de Datos
CUNEF Universidad

Contents

1 Introducción

Contenidos:

- Análisis exploratorio de datos multidimensionales
- Medidas estadísticas multivariantes. Gráficos multidimensionales
- Casos Prácticos con R

1 Introducción

Objetivos

- Conocer los instrumentos metodológicos de un análisis exploratorio de datos multidimensionales
- Vector de medias y matrices de covarianzas y de correlaciones
- Conocer y aplicar los diferentes gráficos multidimensionales
- Realizar una análisis exploratorio multivariado con R

Contents

2 Análisis exploratorio

- *Los datos* consisten en observaciones de n individuos en los que se miden p características o variables, las misma en todos ellos. Los datos se disponen en una matriz de datos $\mathbf{X}$ de dimensiones $(n \times p)$

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

- Las filas son los individuos y las columnas las variables. El valor x_{ij} es el valor de la variable j para el individuo i.

2 Análisis exploratorio

Dada una matriz de datos multidimensionales $\{x_{ij}\}$, $i = 1, 2, \ldots, n$ (individuos) y $j = 1, 2, \ldots, p$ (variables) tenemos que distinguir dos aspectos

- **Estructura Marginal**: Se refiere a la información de las variables $x_1, \ldots, x_p$ vistas de forma aislada. De esta forma, podemos obtener el vector de medias, medianas, etc de las variables, sin hacer referencia a las posibles dependencias.

- **Estructura de Dependencia**. Se refiere al estudio de las dependencias entre variables. El tipo de *dependencia lineal* es el primero que se estudia y el más habitual. La modelización de la estructura de dependencia entre variables se realiza por media de la *teoría de cópulas*.

2 Vector de medias

- Se trata de medidas de las variables (estructura marginal). El *vector de medias muestrales* es:

$$\bar{x} = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{pmatrix}$$

donde, $\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$ es la media muestral de la variable X_j. Es un vector de dimensión $p \times 1$. El vector de medias muestrales es el centro de la nube de puntos de dimensión p.

2 Matrices de covarianzas y correlaciones

- La *matriz de covarianzas* es:

$$S_X = \begin{pmatrix} s_{11} & s_{12} & \dots & s_{1p} \\ s_{21} & s_{22} & \dots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \dots & s_{pp} \end{pmatrix}$$

- Elementos: varianzas s_{jj} y covarianzas s_{jk}

$$s_{jj} = s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2, \quad s_{jk} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$$

2 Matrices de covarianzas y correlaciones

Propiedades

- La matriz de covarianzas contiene información sobre dependencias de tipo lineal entre las variables
- La matriz S_X es simétrica
- El determinante es el producto de los autovalores $|S_X| = \prod_{j=1}^{p} \lambda_j$
- Si $|S_X| = 0$ existirá dependencia lineal entre algunas de las variables
- Traza:

$$\text{tr}(S_X) = \sum_{j=1}^{p} s_j^2 = \sum_{j=1}^{p} \lambda_j \geq 0$$

2 Matrices de covarianzas y correlaciones

- La *matriz de correlaciones* viene dada por:

$$R_X = \begin{pmatrix} 1 & r_{12} & \dots & r_{1p} \\ r_{21} & 1 & \dots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \dots & 1 \end{pmatrix}$$

- Elementos: correlaciones entre variable j y k:

$$r_{jk} = \frac{s_{jk}}{s_j s_k}$$

- El coeficiente de correlación no depende de las unidades de medida, toma valores entre -1 y 1 y el signo indica si la relación es positiva o negativa. Si $r = 0$ no existe dependencia lineal.

2 Matrices de covarianzas y correlaciones

- La matriz R_X es simétrica,

$$R_X = D^{-1} S_X D^{-1}$$

  donde D es una matriz diagonal con elementos $(s_1, \ldots, s_p)$
- Si $|R_x| = 0$ existirá dependencia lineal entre algunas de las variables
- La traza de R_X es:

$$\text{tr}(R_X) = 1 + \cdots + 1 = \sum_{j=1}^{p} \lambda_p^{(r)} = p$$

- Se pueden calcular matrices de correlación en términos de los diferentes coeficientes de correlación:
  - Pearson: coeficiente de correlación líneal
  - Spearman: basado en rangos
  - Kendall: basado en concordancias

2 La varianza generalizada

- La *varianza generalizada* se define como el determinante de la matriz de varianzas y covarianzas (su raíz cuadrada es la desviación típica generalizada). Es una medida global de la variabilidad conjunta de las p variables.

- Propiedades
  - Está correctamente definida, puesto que $|S_X| > 0$
  - Es una medida del área (si $p = 2$), del volumen (si $p = 3$) o del hipervolumen si $p > 3$ de los datos.
- En el caso $p = 2$, si disponemos de una matriz de covarianzas muestral,

$$S = \begin{pmatrix} s_1^2 & r s_1 s_2 \\ r s_1 s_2 & s_2^2 \end{pmatrix}$$

entonces:

$$|S| = s_1^2 s_2^2 (1 - r^2), \quad |S|^{1/2} = s_1 s_2 \sqrt{1 - r^2}.$$

3 Visualización de datos multivariantes

- Como aspecto previo y lo mismo que en el caso univariante, el *análisis exploratorio gráfico multidimensional* es un *aspecto clave y complementario* al análisis exploratorio numérico.
- Algunas de las representaciones gráficas más importantes se presentan a continuación.
- *Diagramas de caja conjuntos*: Diagrama que incluye los box-plots de las diferentes variables marginales. Permite comparar las diferentes niveles de las variables
- *Gráfico de dispersión matricial (pairs)*: incluyen los diagramas de díspersión de las diferentes parejas de variables. Pueden incluir información marginal como histogramas o estimaciones kernel

3 Visualización de datos multivariantes

- *Representaciones gráficas de la matriz de correlaciones*: visualización de la significatividad y de la magnitud de la correlación.
- *Co-plots o diagramas condicionales*: Permiten obtener los diagramas de dispersión de dos variables (x_i, x_j) para los diferentes niveles de una tercera variable x_k, ya sea numérica o categórica
- *Gráficos conjuntos*: gráficos que combinan dos variables de naturaleza cuantitativa, para los niveles de una variable categórica. Por ejemplo, la altura y el peso de una muestra de individuos según género.
- *Coordenadas Paralelas*: Se representan ls variables de cada individuos unidas por líneas. Permite captar cluster de individuos y tendencias.
- *Mapas de calor y dendogramas*
- *Diagramas para bigdata mediante* `ggplot2`: diagramas exagonales y de controno

3 Visualización de datos multivariantes

- *Visulación de matrices de distancia y visualización de Cluster jerárquicos de individuos mediante dendogramas*.
- *Diagramas de caras de Chernoff* : Se trata de gráficos que representar a los individuos por medio de una cara, a partir de los datos de las p-variables. Las características de la cara incluyen: tamaño y forma de la cara; tamaño de la nariz; posición de la boca; tamaño y grosor de la sonrisa; posición, separación, tamaño e inclinación de los ojos etc.
- *Diagramas de estrella*: igualmente a los diagramas de cara, se trata de un gráfico para representar individuos. Cada individuo se representa por una estrella, con tantos rayos o ejes como variables. Cada eje representa el valor de la variable re-escalada, de modo independiente entre cada variables. En todas las estrellas se usa siempre el mismo eje para representar la misma variable.
- Paquetes de visulización: `ggplot2`, `lattice`

3 Visualización de datos multivariantes

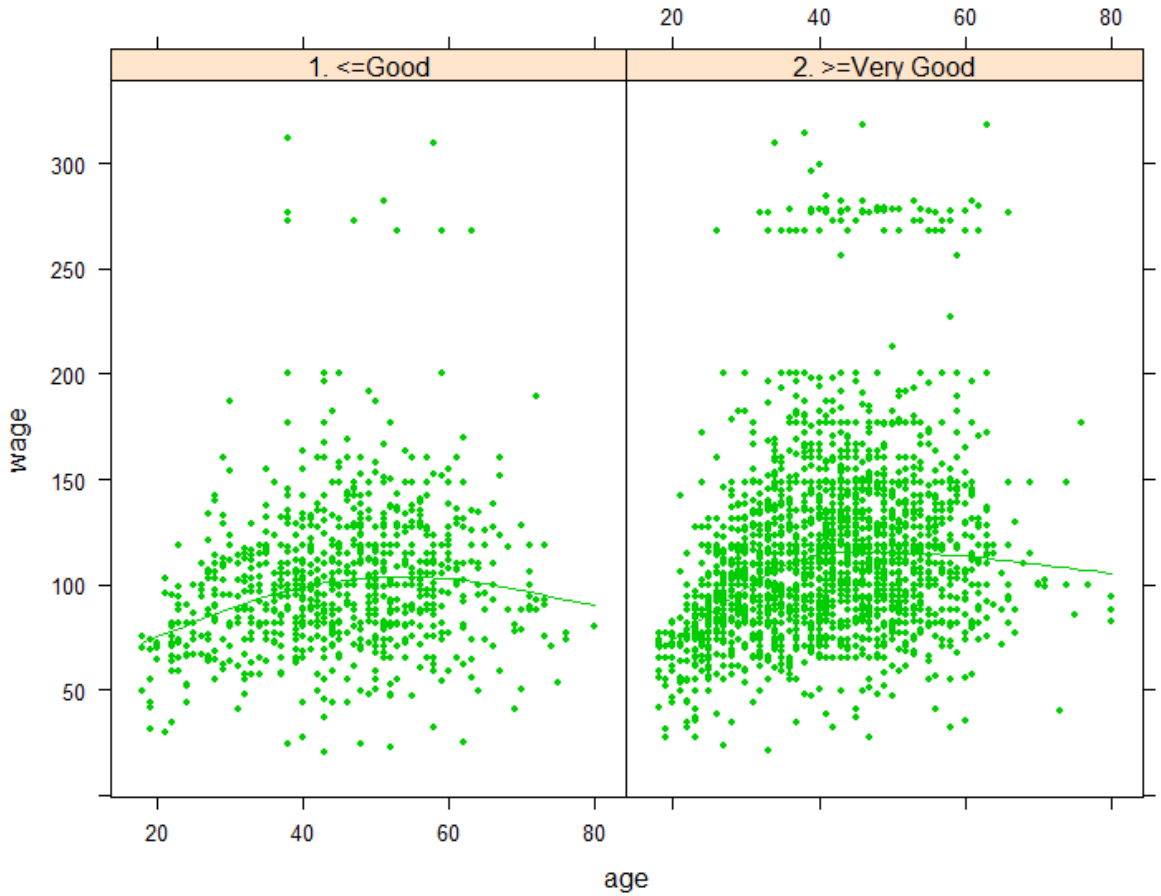

Figura: Gráficos de dos variables según niveles de una tercera variable categórica

3 Visualización de datos multivariantes

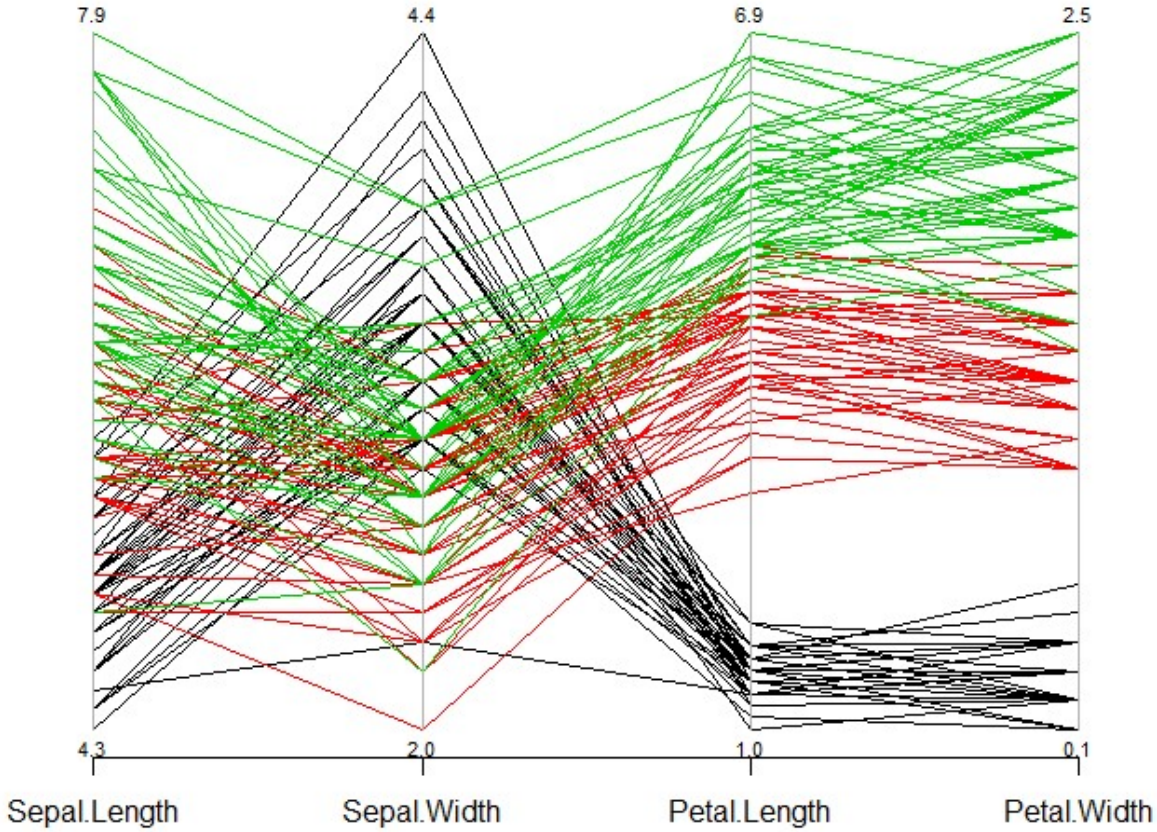

Figura: Coordenadas parallas con datos Iris

3 Visualización de datos multivariantes

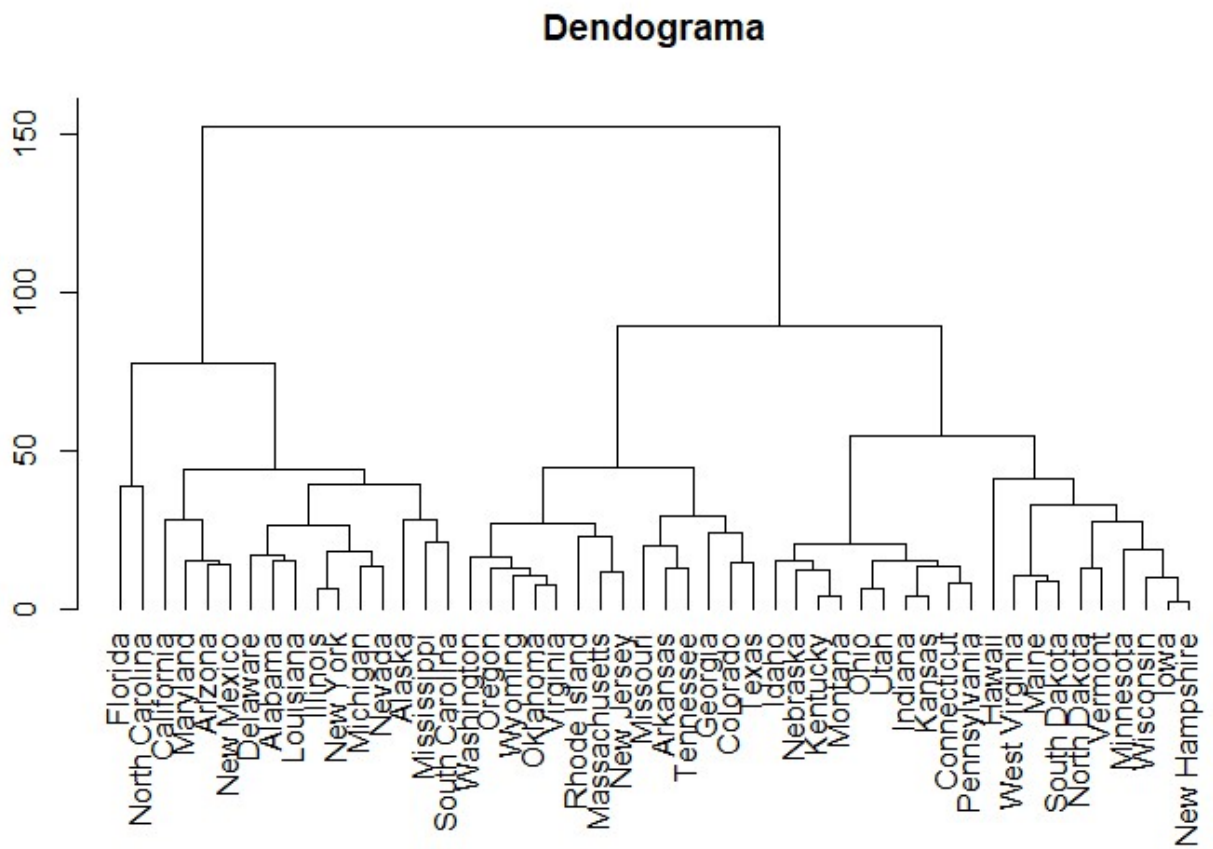

Figura: Dendograma de tasas de criminalidad en USA

Contents

4 Análisis exploratorio gráfico

- Disponemos de una serie de precios de un activo $\{P_1, P_2, \ldots, P_n\}$.
- Los rendimientos en tiempo discreto (prescindiendo de los dividendos) son:

$$R_t = \frac{P_t - P_{t-1}}{P_{t-1}} = \frac{P_t}{P_{t-1}} - 1$$

- Los rendimientos en tiempo continuo son:

$$R_t = \log\left(\frac{P_t}{P_{t-1}}\right) = \log P_t - \log P_{t-1}$$

Las dos fórmulas dan lugar a resultados parecidos (dependiendo de la frecuencia de los precios) puesto que,

$$\log\left(\frac{P_t}{P_{t-1}}\right) = \log\left(\frac{P_t - P_{t-1}}{P_{t-1}} + 1\right) \approx \frac{P_t - P_{t-1}}{P_{t-1}}$$

- Ahora, repetimos el proceso con m activos y tenemos un dataset $(\mathbf{R}^{(1)}, \ldots, \mathbf{R}^{(m)})$

4 Análisis exploratorio gráfico

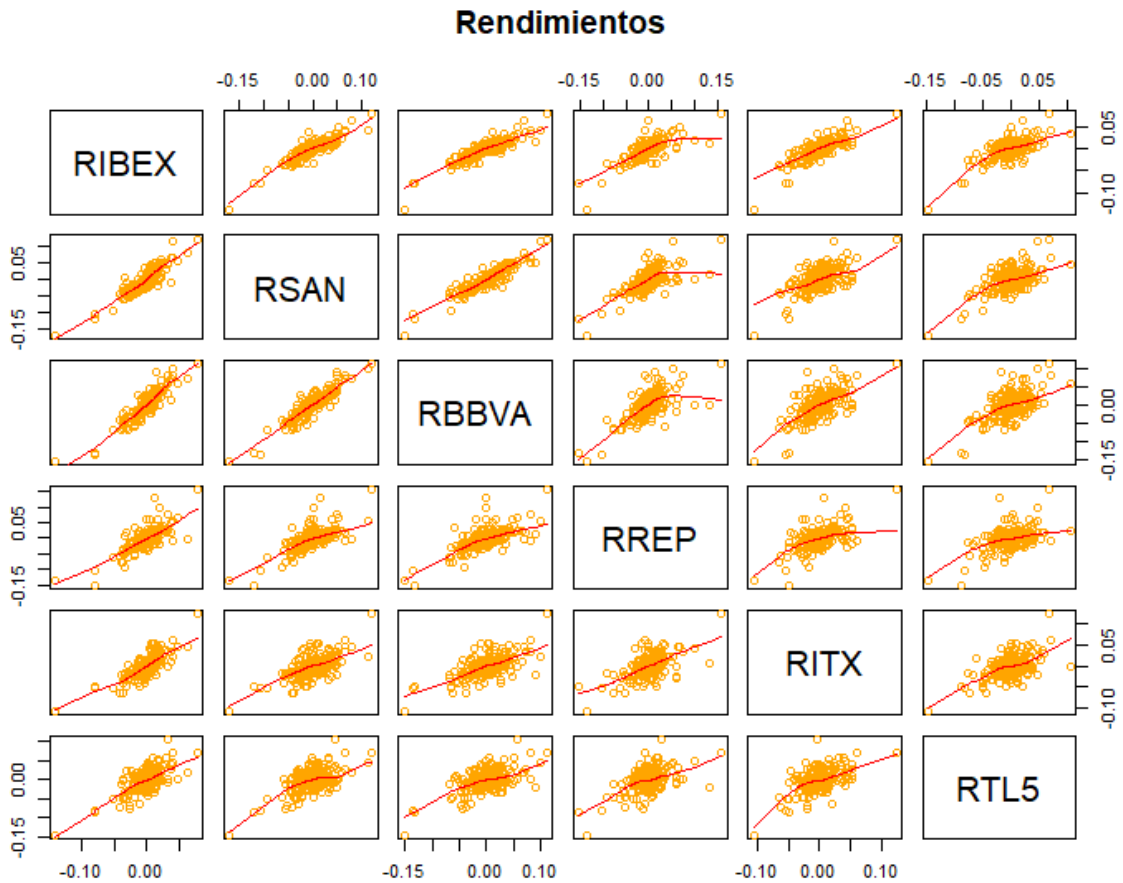

Figura: Diagrama matricial de rendimientos de activos

4 Análisis exploratorio gráfico

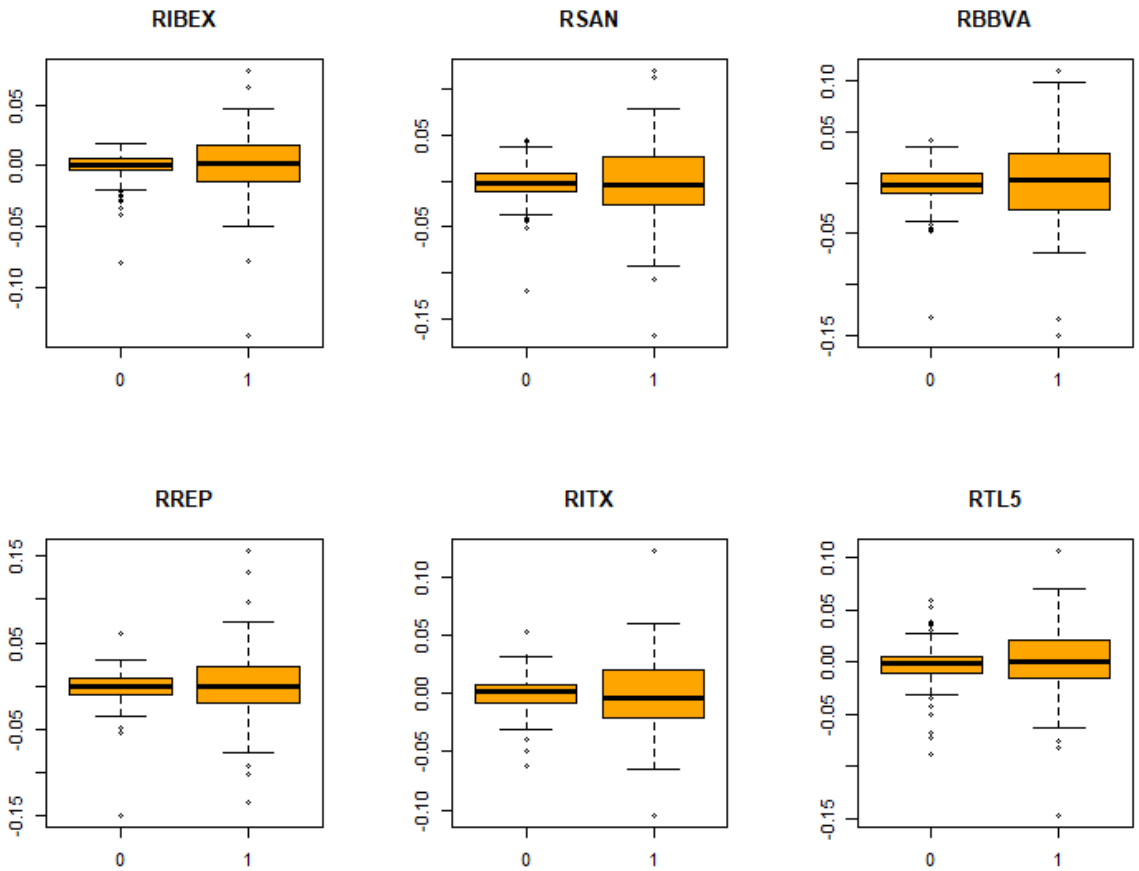

Figura: Box plot de rendimientos antes y después del covid

4 Análisis exploratorio gráfico

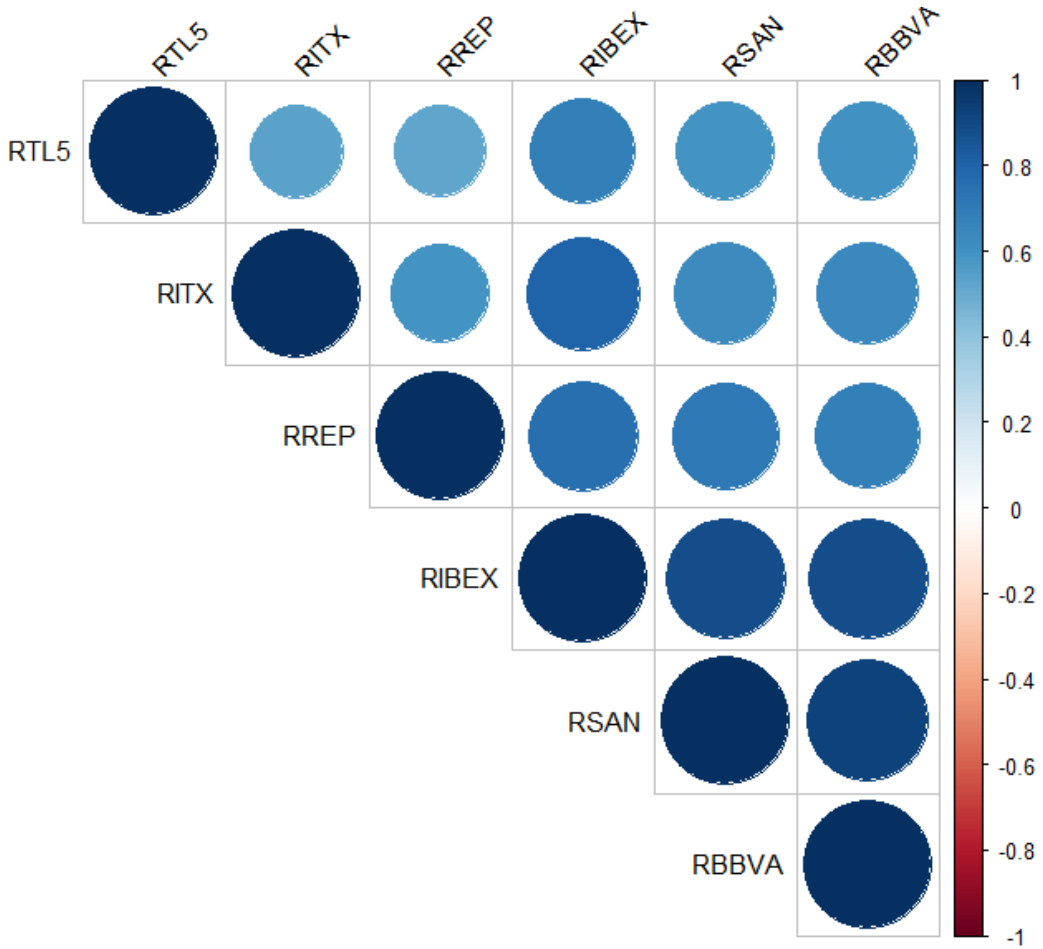

Figura: Visualización matriz de correlaciones

Fundamentos para al Análisis de Datos y la Investigación

Tema 2 - Análisis Exploratorio de Datos (2/2)

José María Sarabia

Máster Universitario en Ciencia de Datos
CUNEF Universidad