Práctica No3: Compañías en EEUU
Tarea 3

Vamos a trabajar con un conjunto de datos consistente en 79 compañías estadounidenses. Para ello tendrás que descargar el archivo de datos çompanies79.csv"que se encuentra en nuestro repositorio.

Tenemos información de 8 variables:

1. Nombre de la compañía
2. Bienes (en millones de dólares)
3. Ventas (en millones de dólares)
4. Valor de mercado (en millones de dólares)
5. Beneficios (en millones de dólares)
6. Flujo de fondos (en millones de dólares)
7. Empleados (en millones de personas)
8. Sector

Nota que las variables V2 a V7 son numéricas.

1. Dibuja un scatterplot múltiple de las 6 variables numéricas, pero coloreando respecto del sector de cada compañía (el sector es la variable categórica V8). Nota que la visualización no es clara, pues los datos se amontonan.

2. Para solucionar esto, crea una nueva tabla con las mismas columnas que la original, pero que las columnas numéricas sean los logaritmos de cada una de las columnas respectivas. Llama Xlog a esta nueva tabla y realiza un scatterplot nuevamente de las columnas numéricas coloreadas por sector. ¿Se ve mejor? ¿A qué crees que se deba esto?

3. Del scatterplot original, observas relaciones lineales entre algunos pares de variables? ¿Cuáles? De esas mismas variables, observa cómo se ven los respectivos scatterplots de Xlog.

4. Obtén el vector de medias muestrales del conjunto de datos original. ¿Te parecen parecidos o muy diferentes los valores para cada una de las medias de cada variable?

5. Obtén la matriz de varianzas y covarianzas muestrales del conjunto de datos. ¿Las varianzas para cada variable dónde se encuentran? ¿Es una matriz simétrica? ¿Por qué?

6. Obtén la matriz de correlaciones muestrales del conjunto de datos. ¿Las correlaciones para cada variable dónde se encuentran? ¿Es una matriz simétrica? ¿Por qué?

7. Para cuantificar la fuerza de las relaciones lineales entre las variables, ¿es mejor usar la matriz de covarianzas o la de correlaciones?

Práctica No4: Compañías en EEUU
Tarea 4

Realizar el mismo análisis (y obtener los mismos resultados) que se hizo para trabajar atípicos con RStudio y distancia de Mahalanobis pero en Python. Ojo: en Python, Mahalanobis se calcula con *scipy.spatial.distance.mahalanobis*.