

Ejercicio T2 UD1: Datos atípicos

Realiza un estudio de los valores atípicos de la base de datos "StudentsPerformance.csv". Para ello realiza los siguientes apartados:

1. Muestra la estructura de la base de datos y realiza un resumen descriptivo de los datos. De la base de datos, aquellas variables declaradas como enteras, cámbialas a numéricas. Visualiza mediante histogramas cada una de distribuciones de las variables "writing.score", "reading.score" y "math.score".
2. Visualiza la distribución de la variable "writing.score" mediante un gráfico de cajas. Indica si el gráfico muestra algún outlier. Si es así, indica qué valores son y en qué posición se encuentran.
3. Realiza un gráfico de dispersión de la variable "writing.score" (variable x debe ser index) en el que se muestre una línea divisoria en color rojo a 3 desviaciones típicas de la media que visualice los posibles outliers.
4. Realiza un gráfico de densidad de la variable "writing.score" en el que se muestre líneas verticales a 3 desviaciones típicas de la media.
5. Realiza el test de Grubbs y el test de Rosner para los potenciales outliers de la variable "writing.score". Interpreta los resultados obtenidos.
6. Visualiza un gráfico de dispersión con las tres variables de calificaciones, interpreta el gráfico. Visualiza también gráficos de cajas para cada una de las tres variables de calificaciones.
7. Teniendo en cuenta las tres variables de calificaciones, calcula la distancia de Mahalanobis y muestra los 5 valores más altos.
8. Representa en un gráfico bidimensional, haciendo uso de las variables "math.score" y "writing.score", la distancia mahalanobis robusta con ajuste, utilizando diferentes colores según las distancias euclidianas de las observaciones. ¿Los gráficos representados indican algún outlier?