Ejercicio T2 UD1: Datos faltantes

Realiza un estudio de los valores faltantes de la base de datos “airquality”, para posteriormente realizar el tratamiento de los datos faltantes. Para ello realiza los siguientes apartados:

1. Haz un resumen de los datos faltantes de cada variable y de cada caso. ¿Qué variable presenta más valores faltantes? ¿Qué podría implicar esto sobre la calidad de los datos o la recolección de los mismos?
2. Visualiza los valores faltantes en los diferentes gráficos trabajados en clase. Comenta los resultados. ¿Notas algún patrón en los valores faltantes entre las variables? ¿Parece que los valores faltantes están correlacionados entre sí?
3. Visualiza los valores faltantes facetada por la variable "Month". ¿En qué meses se concentran los valores faltantes? ¿Podría haber alguna razón estacional para esta distribución?
4. Genera la matriz sombra de la base de datos.
5. Calcula la media de la variable Wind diferenciada por los datos completos y datos faltantes de la variable Ozone. ¿Existe una diferencia notable entre ambas medias? Si sí, ¿qué podría estar indicando esta diferencia sobre la relación entre Wind y Ozone?
6. Visualiza la distribución de la variable Wind en función de los valores faltantes de la variable Ozone mediante un gráfico de cajas. Visualiza también la distribución de la variable Temp en los mismos términos. ¿Qué diferencias encuentras entre las distribuciones de Wind y Temp cuando Ozone es faltante o completo? ¿Qué hipótesis podrías plantear sobre las posibles causas de los valores faltantes?
7. Realiza un gráfico de dispersión de la variable Wind en función de Temp, en el gráfico deben estar diferenciados los registros que sean faltantes o completos en la variable Ozone. ¿Hay patrones claros en la relación Wind-Temp que puedan estar relacionados con los valores faltantes de Ozone? ¿Qué podría significar esto?
8. Realiza el test de Little y concluye si los datos faltantes son MCAR. Si el test indica que los datos no son MCAR, ¿qué estrategias considerarías para tratar los valores faltantes? ¿Cómo cambiaría tu enfoque dependiendo del resultado?
9. Genera una base de datos a partir de la original que se hayan eliminado los datos faltantes. ¿Cuántas observaciones se pierden al eliminar las filas con valores faltantes? ¿Cómo afecta esto al tamaño de la muestra y la representatividad del análisis?
10. Imputa los valores faltantes por la media, en aquellas variables que haya valores faltantes. ¿Qué impacto tiene esta imputación sobre las distribuciones de las variables imputadas?
11. Evalúa las imputaciones realizadas en el apartado anterior mediante los gráficos pertinentes. ¿Las distribuciones de las variables imputadas son similares a las originales? ¿Qué conclusiones puedes extraer sobre la validez de la imputación por la media?
12. Imputa los valores faltantes en las variables Ozone y Solar.R, por regresión lineal y por regresión estocástica. Utiliza como variables independientes Wind y Temp. Evalúa las imputaciones realizadas mediante los gráficos pertinentes. ¿Las imputaciones realizadas por regresión preservan las relaciones entre variables? ¿Cuál de los métodos (lineal o estocástico) parece más adecuado según los gráficos obtenidos?
13. Compara la distribución de las variables Ozone y Solar.R antes y después de la imputación con los diferentes métodos. ¿Cuál de los métodos de imputación parece más realista? Justifica tu respuesta basándote en los gráficos y estadísticas descriptivas.