EDA (Exploratory Data Analysis)

Contenido

[Recursos útiles 1](#_Toc207035875)

[Enunciado 2](#_Toc207035876)

[1. Carga de datos y vistazo general (10 puntos) 2](#_Toc207035877)

[2. Ajuste de tipos y normalización de nombres (8 puntos) 2](#_Toc207035878)

[3. Distribución de variables numéricas (12 puntos) 3](#_Toc207035879)

[4. Análisis de valores ausentes (NA) (12 puntos) 3](#_Toc207035880)

[5. Tratamiento de valores ausentes (12 puntos) 3](#_Toc207035881)

[6. Detección de valores atípicos (outliers) en variables numéricas (15 puntos) 3](#_Toc207035882)

[7. Detección de valores atípicos / desequilibrios en variables categóricas (8 puntos) 4](#_Toc207035883)

[8. Tratamiento de outliers numéricos (12 puntos) 4](#_Toc207035884)

[9. Análisis temporal básico (series de tiempo) (8 puntos) 5](#_Toc207035885)

[10. Exportación y trazabilidad (3 puntos) 5](#_Toc207035886)

# Recursos útiles

Recuerde el directorio de recursos compartidos disponible en:

<https://nascorformacion0-my.sharepoint.com/:f:/g/personal/juan_pinuela_docente_nascorformacion_com/El63bU3PsG5CuybEDz8vH04BL1J5O88yNhhYMNYynPpZWA?e=k5slLd>

# Enunciado

Partimo de un dataset público disponible en:

<https://datosabiertos.jcyl.es/web/jcyl/risp/es/medio-ambiente/calidad_aire_historico/1284212629698.csv>

Y que también puede descargar del enlace anterior con nombre: “calidad-del-aire-datos-historicos-diarios.csv”. El objetivo es aplicar un flujo completo de análisis, limpieza y detección de patrones que permita comprender la estructura de los datos y preparar su uso en futuros modelos.

## Carga de datos y vistazo general (10 puntos)

### 1.1. Importe librerías requeridas (mínimo: `pandas`, `numpy`, `matplotlib.pyplot`).

### 1.2. Cargue el CSV en un `DataFrame` de Pandas.

### 1.3. Muestre:

* Las 5 primeras filas (`head()`).
* La estructura del dataset (`info()`).
* El número de filas y columnas.
* Un resumen estadístico de las variables numéricas (`describe()`).

### Explique brevemente: ¿qué rangos y escalas presentan las variables? ¿detecta valores extremos aparentes a simple vista?

## Ajuste de tipos y normalización de nombres (8 puntos)

### 2.1. Convierta `Fecha` a formato datetime (incluya `dayfirst=True` si corresponde).

### 2.2. Convierta `Provincia` y `Estación` a tipo categórico.

### 2.3. Normalice nombres de columnas (p. ej., quitar espacios, unificar mayúsculas/minúsculas, sustituir `µ` por `u` si procede).

### 2.4. Liste categorías únicas de `Provincia` y `Estación` e indique si observa posibles duplicidades por diferencias ortográficas o de formato.

## Distribución de variables numéricas (12 puntos)

### 3.1. Seleccione solo columnas numéricas.

### 3.2. Genere histogramas para todas las variables numéricas:

* Use una cuadrícula de subplots.
* Defina bins adecuados (justifique su elección).
* Incluya títulos y etiquetas de ejes.

### 3.3. Comente brevemente la forma de las distribuciones (simetría, colas, multimodalidad) y posibles implicaciones para pasos posteriores.

## Análisis de valores ausentes (NA) (12 puntos)

### 4.1. Calcule y muestre:

* Matriz booleana de NA (`isna()`), solo para verificar.
* Conteo total de NA por columna.
* Porcentaje de NA por columna (con 2 decimales).

### 4.2. Identifique las columnas con > 50% de NA y proponga un criterio de tratamiento (eliminación o retención con imputación).

### 4.3. Justifique (en texto) el criterio elegido para este dataset.

## Tratamiento de valores ausentes (12 puntos)

### 5.1. Haga una copia del DataFrame original (para conservarlo).

### 5.2. Elimine las columnas con > 50% de NA (si procede, según 4.2).

### 5.3. Para el resto de variables numéricas, impute NA con la media de cada columna.

### 5.4. Verifique que no queden NA en las columnas tratadas (muestre conteos post-imputación).

## Detección de valores atípicos (outliers) en variables numéricas (15 puntos)

Seleccione una variable numérica principal (p. ej., `O3 (ug/m3)`) y realice:

### 6.1. Histograma actualizado tras la imputación (comparar visualmente con 3.2 si procede).

### 6.2. Cálculo de cuartiles e IQR:

* Obtenga Q1, Q3 e IQR = Q3 − Q1.
* Defina límites de outlier:
  + Inferior = Q1 − 1.5·IQR
  + Superior = Q3 + 1.5·IQR

### 6.3. Cuantifique: número y porcentaje de registros fuera de esos límites.

### 6.4. Boxplot de la variable (indique explícitamente si usa `matplotlib` o `seaborn`).

### 6.5. Discuta: ¿los outliers son plausibles (picos de contaminación) o errores? Proponga dos estrategias posibles de tratamiento y elija una para aplicar en el punto 8.

## Detección de valores atípicos / desequilibrios en variables categóricas (8 puntos)

### 7.1. Analice la distribución de `Provincia`:

* Conteo por categoría (tabla).
* Gráfico de barras (ordenado de mayor a menor).

### 7.2. Interprete si hay desequilibrios notables o posibles errores de codificación (nombres mal escritos, espacios, codificaciones múltiples de la misma provincia).

## Tratamiento de outliers numéricos (12 puntos)

Aplique una de estas estrategias a la variable numérica principal (p. ej., `O3 (ug/m3)`), justificando su elección:

* Eliminación de registros fuera de los límites IQR (trimming).
* Winsorización: recortar valores por debajo/encima de los límites a los propios límites.
* Imputación robusta (p. ej., reemplazar outliers por la mediana).

### 8.1. Implemente la estrategia elegida.

### 8.2. Compare antes vs. después:

* Media, mediana, desviación estándar.
* Histograma o boxplot actualizado.

### 8.3. Comente el efecto del tratamiento sobre la distribución.

## Análisis temporal básico (series de tiempo) (8 puntos)

### 9.1. A partir de `Fecha`, cree columnas adicionales:

* Año y Mes (numérico o etiqueta).

### 9.2. Agregue la variable numérica principal por Mes (en todo el dataset o por `Provincia`, según prefiera).

* Muestre una tabla con medias mensuales.
* Genere un gráfico de líneas simple de la media mensual (indique ejes y título).

### 9.3. Comente si detecta estacionalidad o tendencias.

## Exportación y trazabilidad (3 puntos)

### 10.1. Guarde un CSV con el dataset limpio final (nombre sugerido: `calidad\_aire\_limpio.csv`).

### 10.2. Incluya en el notebook un resumen breve (5–8 líneas) con:

* Columnas eliminadas y por qué.
* Método de imputación usado.
* Estrategia de outliers aplicada y su justificación.