# Caso Práctico para la Asignatura 'Visualización y Análisis de Datos'

Visualización Interactiva de Indicadores Ambientales en la Ciudad de Madrid

**VISUALIZACIÓN Y ANALÍTICA DE DATOS MASIVOS**

15 de mayo de 2025

Autor: Francisco Javier Agudín Álvarez

# Contenido

# Caso Práctico para la Asignatura 'Visualización y Análisis de Datos'

Visualización Interactiva de Indicadores Ambientales en la Ciudad de Madrid

## Objetivo

La presente memoria documenta el desarrollo completo del caso práctico realizado en el contexto de la asignatura "Visualización y Análisis de Datos", impartida en el Máster Universitario. El objetivo fundamental ha sido concebir, implementar y publicar un sistema de visualización interactiva orientado al análisis ambiental en el ámbito urbano de la ciudad de Madrid. Este sistema permite la exploración de datos heterogéneos, recogidos a través de redes de sensores municipales que miden calidad del aire, condiciones meteorológicas y niveles de ruido.

El desarrollo de este proyecto se ha estructurado en torno a un enfoque profesional, abordando todas las fases del proceso: adquisición de datos, depuración, transformación estructural, diseño visual, interpretación analítica y despliegue web. La tecnología central utilizada ha sido la librería Altair, que facilita la construcción de visualizaciones declarativas y reproducibles.

Como objetivos secundarios, se plantean los siguientes:

- Implementar una herramienta de visualización interactiva basada en Altair que permita explorar datos ambientales filtrados por fechas, zonas geográficas y tipo de indicador.
- Correlacionar datos ambientales con condiciones climáticas para detectar patrones en la calidad del aire y el ruido urbano.
- Facilitar la interpretación de tendencias mediante gráficos adecuados, respetando principios de visualización de datos (eficacia, claridad y percepción visual).
- Garantizar la reproducibilidad del análisis mediante código documentado y una estrategia de validación del entorno de ejecución.

En resumen, los objetivos específicos planteados en este ejercicio han sido:

- ✓ Integrar datos de calidad del aire, ruido y meteorología mediante estructuras normalizadas.
- ✓ Desarrollar un conjunto de visualizaciones interactivas claras, interpretables y accesibles vía web.
- ✓ Aplicar técnicas de limpieza, transformación y validación para garantizar la fiabilidad del análisis.
- ✓ Fomentar la reproducibilidad mediante documentación y publicación en GitHub y Binder

## Fuente de Datos

Se han usado fuentes abiertas del portal de datos de la ciudad de Madrid (https://datos.madrid.es/portal/site/egob/), concretamente:

- Datos históricos y en tiempo real sobre calidad del aire.
- Datos de ruido ambiental en diferentes zonas urbanas.
- Información meteorológica disponible para correlacionar datos ambientales con condiciones climáticas.

Las fuentes se han elegido en base a que proporcionan datos estructurados y en formato estándar, lo que facilita su tratamiento, ofrecen mediciones con la granularidad temporal suficiente para detectar tendencias, y en su conjunto permiten analizar la relación entre factores ambientales (ej. impacto de la temperatura en la calidad del aire).

Las fuentes utilizadas provienen del portal de datos abiertos del Ayuntamiento de Madrid. Se han seleccionado por su fiabilidad institucional, granularidad temporal diaria, riqueza geográfica y estandarización en los formatos de publicación. Las tres fuentes principales son:

- Datos de **calidad del aire**: incluye concentraciones de $NO_2$, PM10, $O_3$, $SO_2$ y CO. Se ofrece por estación y día, con códigos validados de magnitud y calidad.
- Datos **meteorológicos**: incorpora variables como temperatura, humedad relativa, presión atmosférica, velocidad del viento y radiación solar.
- Datos de **ruido (estaciones NMT):** recoge niveles acústicos desglosados por períodos del día y percentiles estadísticos.

# Preparación y Transformación de los Datos

## Descarga y Consolidación

Los ficheros fueron descargados desde el portal de datos abiertos del Ayuntamiento de Madrid, en formato CSV. Para facilitar la gestión, se establecieron tres grupos de ficheros: calidad del aire, meteorología y ruido. Cada grupo se consolidó en un único DataFrame, homogenizando los nombres de columnas y el rango temporal (2022–2025).

También se han descargado fichero conteniendo los datos de las estaciones de medición, para cada uno de los conjuntos de datos.

## Validación Estructural

Se ha garantizado la coherencia de los campos clave (FECHA, ESTACION, MAGNITUD) mediante inspección manual y validaciones cruzadas. Los conjuntos se han organizado en tres DataFrame estructurados para su posterior explotación visual.

La preparación de los datos ha constituido una fase crítica del trabajo. Cada conjunto de datos fue procesado para unificarse estructural y semánticamente, permitiendo su integración posterior en visualizaciones conjuntas. A continuación, se describen las principales operaciones realizadas.

## Transformaciones Aplicadas

- En los datos de calidad del aire y meteorología, los valores diarios estaban codificados en columnas separadas (D01 a D31). Se transformó esta estructura de formato ancho a formato largo usando pd.melt().
- Se creó una variable FECHA combinando las columnas ANO, MES y DIA, generando un campo de tipo datetime unificado.
- Se filtraron registros marcados como no válidos (por ejemplo, códigos de validación diferentes de "V"), y se eliminaron fechas imposibles o valores nulos.

- Los valores numéricos se estandarizaron a tipo float, convirtiendo las comas decimales donde era necesario.
- Se tradujeron los códigos de magnitudes a nombres descriptivos usando diccionarios definidos manualmente (por ejemplo, 6 → NO2).
- Se realizó un cruce con tablas de metadatos para asignar a cada código de estación su nombre completo y coordenadas geográficas.

### Estructura Final de los Datos

Como resultado del procesamiento, se generaron los siguientes DataFrame principales:

- df_calidad_largo: contiene los valores diarios de contaminantes validados, con columnas normalizadas (FECHA, ESTACION_NOMBRE, MAGNITUD_NOMBRE, VALOR). Se utiliza para todos los gráficos relacionados con la calidad del aire.
- df_meteo_largo: almacena las variables meteorológicas también en formato largo, estructuradas por magnitud. Incluye temperatura, humedad relativa y presión, entre otras. Se emplea para gráficos de evolución climática y radiales.
- df_ruido_largo: contiene los datos de ruido procesados desde las estaciones NMT, con nombres normalizados de estaciones. Sirve como base para las visualizaciones de ruido y el análisis de correlación con otras variables.

Cada uno de estos conjuntos permite su integración directa con Altair, permitiendo flexibilidad en los filtros y selecciones interactivas, y la interrelación con otros conjuntos de datos.

## Visualización Interactiva con Altair

La fase de visualización ha sido concebida como el eje central del proyecto, en tanto que articula la representación interpretativa de los datos procesados. El uso de la librería Altair ha permitido generar un sistema interactivo en el que el usuario puede explorar el conjunto de variables ambientales a través de distintos niveles de agregación y enfoque.

### Diseño General

La implementación del sistema de visualización se ha concebido como un entorno modular, flexible y navegable. Altair, como herramienta declarativa de visualización, permite construir gráficos estructurados a partir de modelos de datos tabulares, lo que facilita una integración fluida con los DataFrame preparados previamente. Uno de los aspectos clave del diseño ha sido garantizar la navegabilidad del sistema, incorporando selectores interactivos vinculados a los campos relevantes (estación, variable y fecha), de modo que el usuario pueda realizar consultas visuales sin necesidad de modificar el código subyacente.

En cuanto a la lógica de diseño, se han priorizado las representaciones que permiten contrastar tanto diferencias espaciales (entre estaciones) como variaciones temporales y proporciones multivariables. Para ello, se han desplegado paneles de visualización que no solo muestran datos, sino que los contextualizan mediante el uso interactivo, codificación por color, normalización de escalas y separación de vistas en celdas.Desde el punto de vista de la programación la incorporación de elementos como alt.selection_point() y alt.binding_select() ha permitido construir filtros directos desde el navegador que afectan simultáneamente a múltiples componentes del gráfico, lo que resulta especialmente útil para el análisis.

Asimismo, se ha tenido en cuenta la cohesión estética y funcional del sistema: las visualizaciones siguen un diseño homogéneo, tanto en paleta de colores como en escalado temporal y disposición gráfica. Cada visualización fue validada individualmente y exportada en formato HTML, preservando la interactividad para su publicación en entorno web sin pérdida de funcionalidad. Se han diseñado diferentes tipos de gráficos en función del propósito analítico: distribución agregada, series temporales, perfiles multivariables, relaciones bivariadas y mapas. Se han utilizado selectores (alt.selection_point) y menús interactivos (binding_select) para ofrecer una experiencia navegable desde el navegador.

## Visualizaciones Generadas

- Barras apiladas: carga total de contaminantes por estación.
- Líneas temporales: evolución diaria de $NO_2$ y PM10 con selección por estación.
- Radiales ambientales: comparación proporcional de $NO_2$, $O_3$ y PM10 por día y estación.
- Radiales climáticos: representación normalizada de temperatura, humedad y presión.
- Gráficos de detalle: series individuales de cada variable ambiental.
- Dispersión ruido-temperatura: exploración bivariada.
- Mapa georreferenciado: distribución espacial de estaciones y valores.

## Estructura del Notebook

El notebook que sustenta este proyecto ha sido estructurado con una secuencia lógica que guía desde la carga y transformación de los datos hasta la generación e interpretación de las visualizaciones. Cada bloque de código está precedido por una celda explicativa en formato Markdown, donde se describe con claridad el propósito de la operación, las decisiones adoptadas y, en su caso, los criterios de selección y filtrado aplicados.

Estas celdas explicativas permiten una lectura fluida y comprensible del flujo de trabajo, facilitando tanto la revisión técnica como la comprensión del análisis. Se ha optado por mantener el código limpio y modular, evitando duplicidades y favoreciendo la reutilización de estructuras. En consecuencia, el propio notebook constituye una documentación viva y exhaustiva del trabajo realizado, y se remite al mismo para el detalle completo de todas las operaciones técnicas, visualizaciones y filtros incorporados.

## Interpretación de Resultados

La interpretación de los resultados conecta las visualizaciones generadas con preguntas analíticas relevantes en el ámbito ambiental urbano. A través del sistema desarrollado, se tiene la posibilidad de explorar datos en múltiples niveles de detalle, aplicando filtros por estación, variable y periodo temporal. Esta flexibilidad permite identificar patrones, relaciones y anomalías que, de otro modo, podrían haber permanecido ocultos.

Uno de los hallazgos más consistentes ha sido la diferencia de comportamiento entre estaciones de zonas conocidas por su nivel de tráfico. En los gráficos de evolución temporal de contaminantes como $NO_2$ y PM10, se observa una mayor concentración en aquellas estaciones situadas en entornos urbanos con tráfico denso. Este patrón se mantiene de forma sostenida en distintos periodos del año, lo que refuerza la hipótesis de que el tráfico rodado sigue siendo una de las principales fuentes emisoras de contaminantes en entornos urbanos como Madrid.

El análisis de los perfiles radiales diarios de calidad del aire ha permitido comparar el equilibrio relativo entre distintos contaminantes en fechas concretas. Esta representación, al normalizar los valores de cada magnitud respecto a su propio rango, facilita la identificación de días atípicos en los que un contaminante se impone de

forma desproporcionada. Por ejemplo, se detectaron episodios en los que el ozono ($O_3$) presenta picos relativamente más altos, especialmente en estaciones periféricas durante los meses de verano, lo que puede estar relacionado con fenómenos fotoquímicos favorecidos por la radiación solar.

Asimismo, el análisis del perfil radial climático ha ofrecido una herramienta para comparar condiciones meteorológicas entre estaciones y fechas. Se observó que días con alta temperatura y baja presión suelen coincidir con condiciones más inestables, lo que se refleja también en la variabilidad de los contaminantes. En este sentido, las visualizaciones integradas permiten inferencias cruzadas entre variables físicas y químicas del entorno.

En el caso de las visualizaciones detalladas por variable, se ha podido observar cómo cada magnitud presenta una estacionalidad distinta. Por ejemplo, el ozono muestra picos característicos durante los meses de verano, mientras que el $NO_2$ y el PM10 presentan máximos más frecuentes en invierno, coincidiendo con una menor dispersión atmosférica y un uso más intensivo del transporte motorizado.

Cabe destacar asimismo el valor del gráfico georreferenciado incluido en el sistema. Esta visualización permite observar de forma espacial la distribución de los valores promedio de una magnitud seleccionada, proyectados sobre la localización real de cada estación en el municipio de Madrid. A través de una codificación por color, se facilita la identificación de zonas de mayor concentración, lo que contribuye al análisis territorial del fenómeno. Por ejemplo, al representar PM10 o $NO_2$, se constata la acumulación de valores más altos en áreas de tráfico denso o zonas urbanas consolidadas, mientras que los valores más bajos se registran en estaciones periféricas o parques urbanos. Esta herramienta resulta especialmente útil para establecer comparaciones inter-estación y detectar gradientes espaciales o puntos críticos de contaminación.

Es importante señalar que cada una de estas interpretaciones ha sido acompañada de un resumen textual incluido como celda Markdown dentro del propio notebook. Estas explicaciones se sitúan junto a las visualizaciones correspondientes, actuando como guías para la lectura e interpretación de los resultados. De esta forma, se pretende favorecer la comprensión técnica del trabajo, al mismo tiempo que servir como mecanismo de comunicación.

En conjunto, el sistema desarrollado en el notebook no solo ha demostrado ser funcional desde el punto de vista visual y técnico, sino que ha permitido obtener resultados interpretables y fundamentados, alineados con hipótesis reales del comportamiento ambiental urbano en la ciudad de Madrid.

El sistema permite contrastar visualmente hipótesis relevantes desde el punto de vista ambiental:

El uso de gráficos radial y de series temporales aporta nuevas perspectivas frente a las visualizaciones temporales tradicionales. La selección interactiva potencia el análisis personalizado y la comparación contextualizada.

.

## Publicación y Acceso

Una parte importante del desarrollo es la publicación de resultados, que en este caso ha sido concebida como parte integral del trabajo, orientada a garantizar la trazabilidad de los procesos realizados, la accesibilidad de los resultados y la posibilidad de reproducir los análisis de forma autónoma desde cualquier entorno compatible. De este modo las interpretaciones pueden ser sometidas a debate y/o obtener nuevas conclusiones y análisis. Para este trabajo se ha optado por una estrategia de publicación basada en herramientas abiertas, ampliamente utilizadas en entornos profesionales y académicos.

En primer lugar, todos los elementos del proyecto se han organizado en un repositorio en GitHub. Este incluye tanto el notebook principal (notebook_altair_FJAA.ipynb) como los archivos auxiliares necesarios para la visualización, ejecución y documentación del proyecto. Se ha prestado especial atención a mantener el notebook sin las salidas generadas, de forma que el usuario pueda ejecutarlo en su entorno para obtener los resultados actualizados y verificar el funcionamiento del sistema paso a paso. Esto además se combina con la limitación de tamaño de los archivos que tiene Github en 25MB.

Para facilitar el acceso a los datos utilizados, se ha creado un directorio específico denominado data, donde se alojan los archivos CSV originales descargados del portal de datos abiertos del Ayuntamiento de Madrid. Esta estructura permite organizar de forma clara los ficheros de entrada y diferenciarlos de los scripts de procesamiento o visualización.

Los gráficos generados con Altair han sido exportados individualmente como ficheros HTML mediante el método Chart.save("nombre.html"). Este procedimiento permite conservar la interactividad en cada visualización, incluso fuera del entorno Jupyter, y facilita su consulta directa desde un navegador web estándar sin necesidad de instalar software adicional. Estas visualizaciones están ubicadas en una carpeta específica (/graficos) dentro del repositorio, lo que permite un mantenimiento ordenado y facilita la navegación.

Para la publicación web, se ha configurado GitHub Pages como sistema de entrega de contenidos. Se ha incluido un archivo index.html diseñado como portada del proyecto, en el que se enlazan de forma estructurada los gráficos exportados, el notebook y una introducción explicativa. Esta interfaz permite al usuario explorar el conjunto visual sin necesidad de conocimientos técnicos. La navegación es fluida y está organizada por categorías temáticas (calidad del aire, variables climáticas, ruido), lo que mejora la experiencia de usuario.

Adicionalmente, se ha habilitado un entorno de ejecución online mediante la plataforma Binder. A través de un botón integrado en el README.md, cualquier usuario puede lanzar una instancia del notebook en la nube, con todas las dependencias instaladas previamente a partir del archivo requirements.txt. Esta opción garantiza la reproducibilidad completa del análisis, sin necesidad de configuración local.

Se han tenido en cuenta además aspectos de buenas prácticas como la inclusión de un archivo '.gitignore', la limpieza del historial de versiones y la organización modular del código, lo que permite la escalabilidad del proyecto y su integración futura en otras plataformas o iniciativas abiertas.

El repositorio completo del proyecto se encuentra disponible en la siguiente dirección: https://github.com/pacojavi/VisualizacionDatos. Desde el archivo README.md es posible acceder tanto a la visualización en GitHub Pages como al entorno de ejecución interactivo mediante Binder, permitiendo una experiencia completa de consulta, exploración y reproducción del análisis. La interfaz web está publicada en la URL: https://pacojavi.github.io/VisualizacionDatos/

## Conclusiones

A lo largo del desarrollo de este caso práctico se han alcanzado, creo de forma satisfactoria, los objetivos planteados inicialmente. La combinación de fuentes de datos heterogéneas, la limpieza de registros, y el diseño de un sistema de visualización modular e interactivo han permitido construir una herramienta flexible y adaptada al análisis ambiental urbano.

El proceso ha resultado especialmente enriquecedor en términos de aprendizaje técnico. La necesidad de transformar estructuras de datos complejas, de construir representaciones visuales significativas, y de garantizar la visualización completa del trabajo, ha supuesto un ejercicio integral que abarca desde la ingeniería de datos hasta la comunicación analítica.

Además del valor instrumental del notebook desarrollado, el proyecto ha aportado una comprensión más profunda de los patrones ambientales que afectan a una gran ciudad como Madrid. La posibilidad de observar estos fenómenos desde distintos ejes (espacial, temporal, multivariable) confirma el potencial de la visualización interactiva como herramienta de explotación de datos del entorno utilizado..

La experiencia ha evidenciado la importancia de una documentación clara y de la publicación abierta del trabajo, principios que se han intentado aplicar para asegurar que cualquier persona que tenga acceso pueda consultar, ejecutar y extender el sistema. Considero que esta base podría ampliarse en futuras iteraciones hacia análisis predictivos o integraciones con plataformas de gestión ambiental urbana.

## Tecnologías utilizadas

El desarrollo del presente proyecto se ha realizado íntegramente en lenguaje Python, utilizando como entorno principal de trabajo Jupyter Notebook.

La limpieza, transformación y manipulación de los datos se ha llevado a cabo mediante pandas, herramienta para el tratamiento de estructuras tabulares y operaciones sobre columnas. La gestión de fechas, conversión de tipos y operaciones de filtrado se ha implementado directamente sobre los DataFrame generados a partir de los ficheros CSV descargados del portal de datos abiertos del Ayuntamiento de Madrid.

Para la visualización interactiva se ha utilizado la librería Altair, basada en el lenguaje de especificación Vega-Lite. Esta herramienta ha sido clave para definir visualizaciones estructuradas, reproducibles y exportables, compatibles con su publicación en la web mediante HTML sin pérdida de interactividad. Altair ha permitido incorporar selectores dinámicos, filtros cruzados y codificación multivariable sin necesidad de programación imperativa, favoreciendo un diseño declarativo claro.

La exportación de los gráficos a HTML y su integración en un entorno navegable se ha realizado mediante funcionalidades propias de Altair, generando archivos embebibles que pueden visualizarse directamente desde un navegador sin necesidad de entorno local.

La ejecución remota del sistema se ha resuelto con Binder, una plataforma de ejecución en la nube que permite cargar notebooks directamente desde un repositorio GitHub y reproducir el entorno especificado en un fichero requirements.txt. Este fichero contiene las dependencias necesarias (pandas, altair, vega_datasets, jupyter, entre otras), y ha sido afinado para minimizar el peso del entorno sin perder funcionalidad.

Toda la infraestructura del proyecto —incluyendo el código, los datos procesados, las visualizaciones y la documentación— se ha versionado y publicado en un repositorio GitHub, facilitando el control de cambios, la trazabilidad y la reutilización futura.