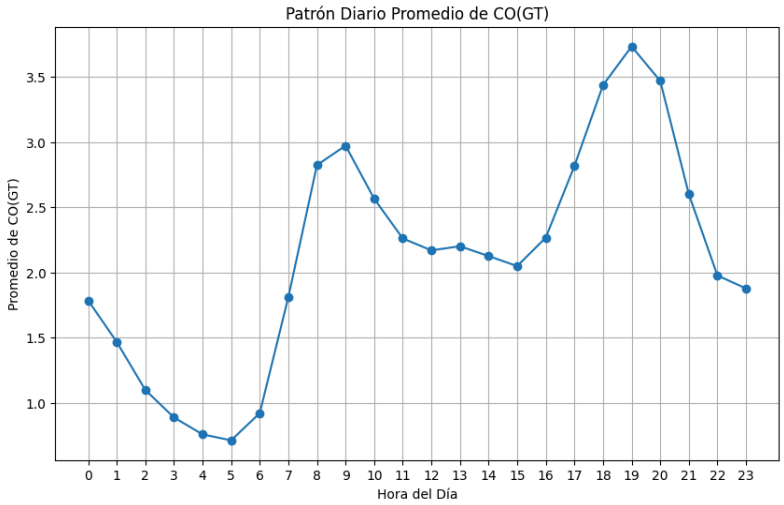
# Preguntas de Negocio Iniciales: Análisis de la Calidad del Aire




Fecha: 21 de abril de 2025

## Introducción:

En este proyecto queremos entender mejor la calidad del aire en una ciudad de Italia. Para esto, queremos analizar unos datos que se recogieron con un dispositivo especial que tiene sensores que mide diferentes gases contaminantes y el clima (como la humedad y la temperatura). Los datos nos pueden dar una idea de cómo está el aire que respiraba la gente durante un año. Con este proyecto, queremos ver qué podemos aprender de esta información para entender mejor la contaminación del aire en esa ciudad y podremos agregar información valiosa sobre estos contaminantes y sus variaciones durante el día y el año. Entender estos patrones es importante para la salud de las personas y para tomar decisiones sobre el medio ambiente.

Fuente de los datos: https://archive.ics.uci.edu/dataset/360/air+quality

## 1. Preguntas Iniciales:

**¿Cuándo se tomaron estas medidas de la calidad del aire?** Sabemos que se tomaron cada hora, durante un año entero. Esto nos da mucha información para ver cómo cambia la contaminación con el tiempo.

**¿Qué podemos aprender en general sobre la calidad del aire con estos datos?** Al mirar toda la información que tenemos, seguro podemos ver si hay días o meses con más o menos contaminación. Esto podría ser importante para la gente que vive en esa ciudad.

**¿La contaminación del aire cambia mucho a lo largo del día?** Como tenemos medidas cada hora, podemos ver si hay horas en las que la contaminación sube o baja, quizás por el tráfico o por otras actividades que pasan en la ciudad.

**¿Qué gases son los que midió este aparato?** El aparato midió varios gases contaminantes diferentes, como el monóxido de carbono, hidrocarburos no metánicos, benceno, dióxido de nitrógeno, entre otros. Queremos saber cuáles son y cómo varían sus niveles.

**¿También se midió el clima al mismo tiempo?** Sí, junto con los gases, también se tomaron datos del clima, como la temperatura y la humedad. Esto nos hace preguntarnos si el clima tiene algo que ver con la cantidad de contaminación en el aire.

**¿Hay información que nos falta en estos datos?** Vimos que hay algunos valores que están marcados como -200, que significa que no se midieron. Necesitamos saber cuántos datos faltan, en qué afectará y cómo tratarlos.

**¿Podríamos usar esta información para predecir los niveles de contaminación y tomar medidas preventivas?** Si vemos patrones en los datos del pasado, podremos saber qué gases afectan más la calidad del aire y ver los niveles de contaminación y predecir los niveles de contaminación.

**¿Podemos identificar cuáles fueron los días con la peor calidad del aire durante ese año?** Buscar los días en los que la contaminación fue más alta nos podría ayudar a entender qué pasó en esos momentos (por ejemplo, si hubo algún evento especial o si el clima fue diferente).

**¿Varía la calidad del aire de forma parecida para todos los gases que se midieron?** Quizás algunos gases contaminantes suben y bajan al mismo tiempo, mientras que otros tienen un comportamiento diferente. Entender esto nos daría una mejor idea de las fuentes de contaminación.

## 2. Preguntas Técnicas sobre el set de datos

**¿Los datos son estructurados o no estructurados?** Los datos son estructurados, organizados en filas y columnas dentro de un archivo CSV. Es tabular y facilita su procesamiento y análisis mediante herramientas como Pandas en Python.

**¿Cómo es el volumen de los datos?** El dataset contiene aproximadamente 9,358 filas y 15 columnas. Este volumen de datos se considera adecuado para la aplicación de técnicas de regresión supervisada sin necesidad de recurrir a metodologías de Big Data.

**¿Las variables objetivo son continuas o categóricas?** Estas concentraciones son valores numéricos continuos.

**¿Qué tipo de datos contiene cada columna?** representan concentraciones de gases (como CO(GT), NOx(GT), C6H6(GT)), señales de sensores electrónicos (PT08.S1(CO) hasta PT08.S5(O3)), y condiciones meteorológicas (Temperatura, Humedad Relativa, Humedad

Absoluta). También hay columnas que representan **información temporal** (Fecha y Hora), que posteriormente se combinaron en un formato datetime.

**¿Qué herramientas y librerías son apropiadas para trabajar estos datos?** Las herramientas y librerías apropiadas incluyen pandas para la manipulación y análisis de datos tabulares, seaborn y matplotlib para la visualización de datos, scikit-learn para la implementación de modelos de aprendizaje automático (regresión), statsmodels para el análisis estadístico y numpy para operaciones numéricas.

**¿Qué tipo de limpieza se debe realizar?** La limpieza principal identificada fue el tratamiento de los valores faltantes, representados por el valor -200 en las columnas numéricas. Estos valores fueron reemplazados por NaN para facilitar su manejo en el análisis. Además, las columnas 'Date' y 'Time' requirieron ser combinadas y convertidas al formato datetime en una nueva columna 'Datetime' para permitir el análisis de series temporales y la extracción de características basadas en el tiempo.

**¿Las variables objetivo son continuas o categóricas?** Estas concentraciones son valores numéricos continuos.

**¿Qué tipo de problema se resolverá?** El objetivo principal es abordar un problema de regresión. Se buscará predecir las concentraciones de gases contaminantes en el aire, siendo el Monóxido de Carbono (CO) un ejemplo clave

**¿Por qué el CO se usa como un ejemplo clave?** El Monóxido de Carbono es un gas tóxico conocido por sus efectos perjudiciales para la salud humana. Por lo tanto, predecir sus niveles es un objetivo relevante desde una perspectiva de salud pública y gestión ambiental.

Queremos predecir esta variable utilizando las otras variables como características predictoras en un modelo de regresión ya que se espera que su valor esté correlacionado o influenciado por ellas. El modelo de regresión aprenderá estas relaciones a partir de los datos históricos para poder predecir la concentración de CO en nuevas situaciones.