UNSA
UNIVERSIDAD NACIONAL DE SAN AGUSTÍN DE AREQUIPA

Datos de Calidad del Aire

Albert Llica Álvarez

Junio 2025

- **Conjunto de datos:** Urban Air (Microsoft Research), mayo 2014–abril 2015.
- **Alcance:** Calidad del aire, meteorología y pronósticos en 43 ciudades chinas.
- **Volumen:** 2.89M registros de calidad del aire, 1.89M meteorológicos, 910K pronósticos.
- **Objetivo:** Analizar estructura, problemas e hipótesis para predicción de calidad del aire.

- **Componentes:**
  - 43 ciudades, 380 distritos, 437 estaciones.
  - Calidad del aire: PM2.5, PM10, NO2, CO, O3, SO2 (horario).
  - Meteorología y pronósticos (3/6/12 h).
- **Granularidad:**
  - **Espacial:** Ciudad, distrito, estación.
  - **Temporal:** Horaria (calidad, meteorología).

DESCRIPCIÓN DE LA DATASET

Archivo	Contenido	Volumen	Granularidad
city.csv	43 ciudades, datos geográficos, clústeres	43 registros	Espacial: Ciudad
district.csv	380 distritos	380 registros	Espacial: Distrito
station.csv	437 estaciones de monitoreo	437 registros	Espacial: Estación
airquality.csv	PM2.5, PM10, NO2, CO, O3, SO2	2.89M registros	Temporal: Horaria
meteorology.csv	Temperatura, presión, humedad,	1.89M registros	Temporal: Horaria
weatherforecast.csv	Pronósticos a 48 h (clima, temperatura)	910K registros	Temporal: 3/6/12 h

Muchos Dataset de Air Pollution, presenta ausencia de datos o outliers que podria sera data incorrecta debido a falla de sensores, mantenimiento, etc. Estos datos requieren ser imputados, sin embargo al ser Multivariables, las técnicas o modelos de imputación no siempre se acercan al valor real de imputación.

- **Valores nulos:** PM10 (8.16%), presión (14.74%), clima (6.57%).
- **Datos sucios:** Outliers (PM2.5 = 1463 µg/m^3), valores no físicos.
- **Duplicados:** 53,496 en pronósticos, aunque no hay duplicados al incluir todas las columnas.

OBJETIVO DEL PROYECTO

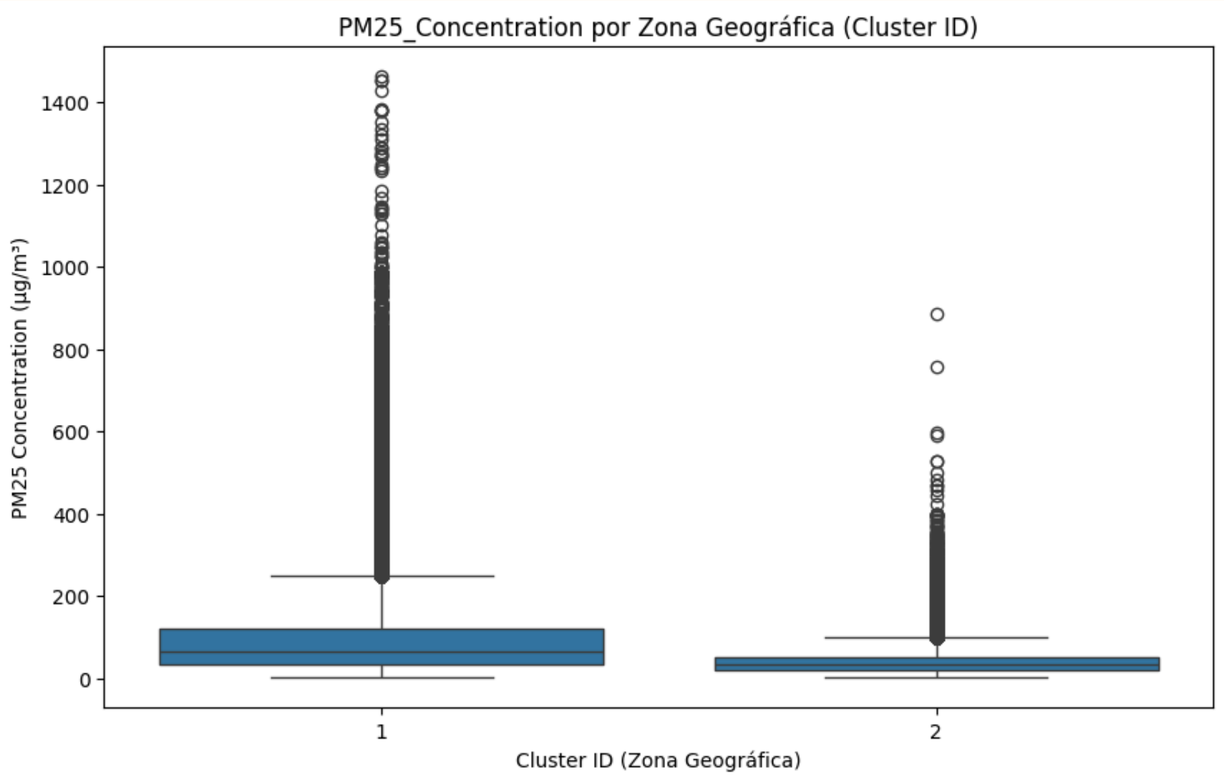
1. Explorar los datos para entender su contexto, tipos y representatividad, evaluando medidas estadísticas, calidad, correlaciones y patrones que respalden o refuten las hipótesis planteadas.

2. Obtener patrones, características o correlaciones en la data que permitan la búsqueda en modelos de imputación de datos de calidad de aire.

3. Usar los modelos de imputación en una herramienta de visualización que permita ver la imputación realizada por cada modelo a la data faltante y permita comparar y elegir resultados.

HIPÓTESIS 1 - INFLUENCIA GEOGRÁFICA

- **Enunciado:** La zona geográfica determina los niveles de calidad del aire.
- **Justificación:** Cluster 1 (Beijing, Tianjin) con mayor PM2.5 por industrialización e inversiones térmicas. Cluster 2 (Shenzhen, Guangzhou) con mejor calidad por clima cálido.



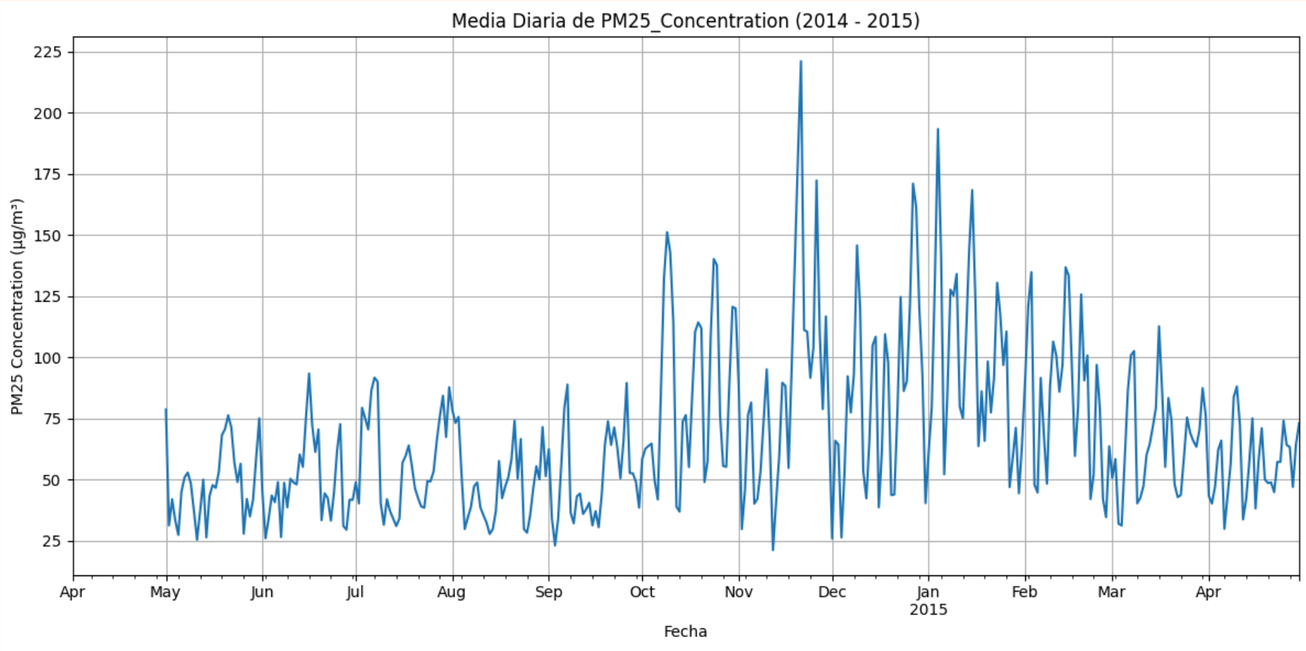
Conclusión :
El análisis muestra que las ciudades en el clúster 1 (Beijing, Tianjin) tienen niveles mucho más altos de PM2.5 que las ciudades del clúster 2 (Shenzhen, Guangzhou, Hong Kong). Esto se debe a la mayor industrialización y la quema de carbón en el norte de China, mientras que el sur disfruta de un clima más cálido que favorece la dispersión de contaminantes. Los valores atípicos no cambian la tendencia general, que resalta la influencia geográfica e industrial en la contaminación.

HIPÓTESIS 1 – INFLUENCIA GEOGRÁFICA

Conclusión :
En el Dashboard se permite evaluar la comparación de diferentes estaciones.
El Primero es del Cluster de Beijing, Mientras que el Segundo es del otro cluster, se aprecia una diferencia en los datos de calidad del aire.

HIPÓTESIS 2 - CICLICIDAD

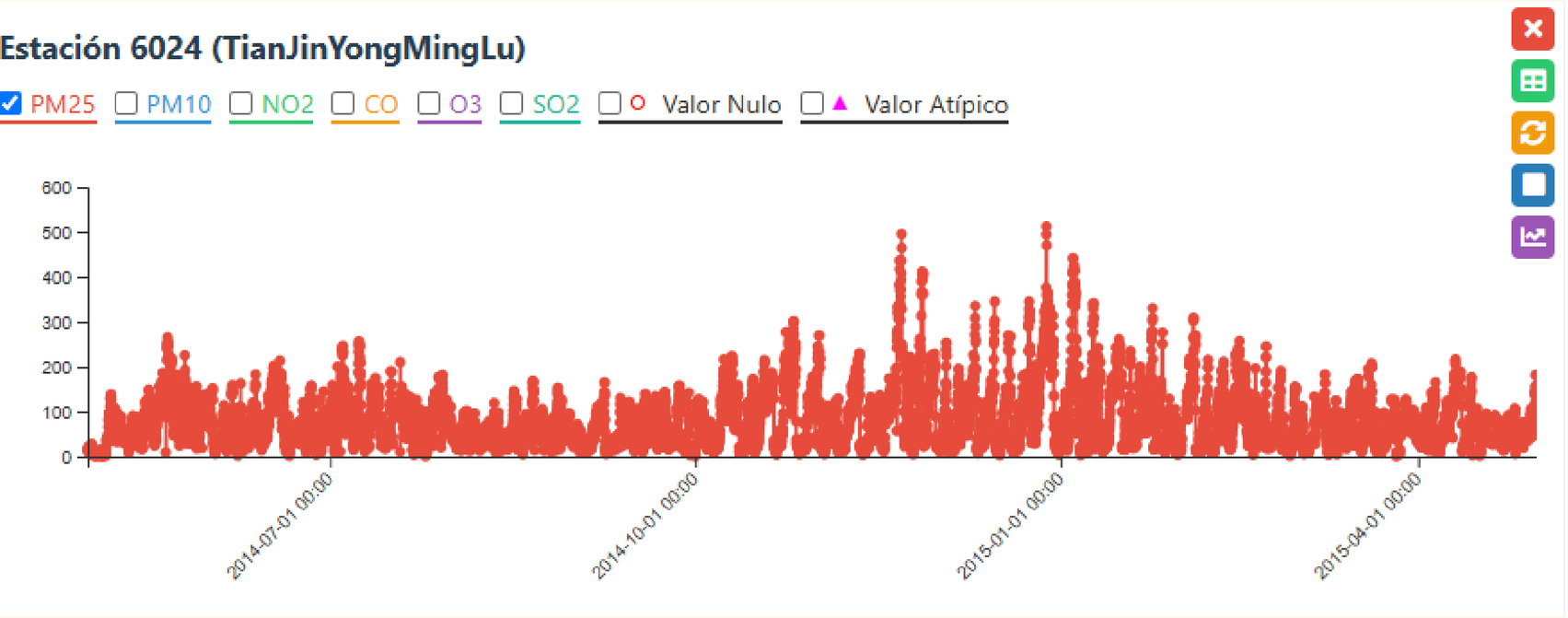
- **Enunciado:** Existe ciclicidad en los datos de calidad del aire y en los valores nulos.
- **Justificación:**
  - **Calidad del aire:** PM2.5 aumenta en invierno (80 µg/m³, quema de carbón, inversiones térmicas) y disminuye en verano (40 µg/m³, mayor dispersión). Patrones diarios por tráfico.



Conclusión :

Los gráficos muestran una clara ciclicidad estacional en la calidad del aire, con niveles de PM2.5 más altos en invierno debido a la quema de carbón y la inversión térmica, y más bajos en verano por una mayor dispersión de contaminantes. Esto se confirma con los datos de 2014-2015, sugiriendo que actualmente, en el inicio del verano de 2025, los niveles de PM2.5 deberían ser más bajos, entre 40-50 µg/m³.

HIPÓTESIS 2 - CICLICIDAD

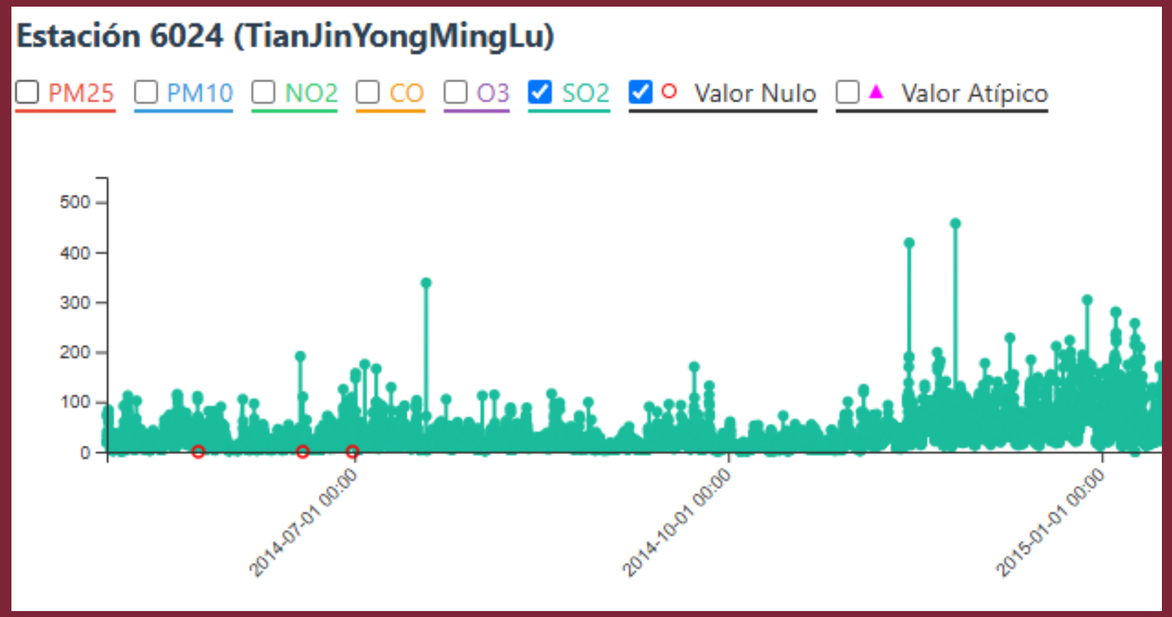


Conclusión :

Los gráficos muestran una clara ciclicidad estacional en la calidad del aire, con niveles de PM2.5 más altos en invierno debido a la quema de carbón y la inversión térmica, y más bajos en verano por una mayor dispersión de contaminantes. Esto se confirma con los datos de 2014-2015, sugiriendo que actualmente, en el inicio del verano de 2025, los niveles de PM2.5 deberían ser más bajos, entre 40-50 $\mu g/m^3$.

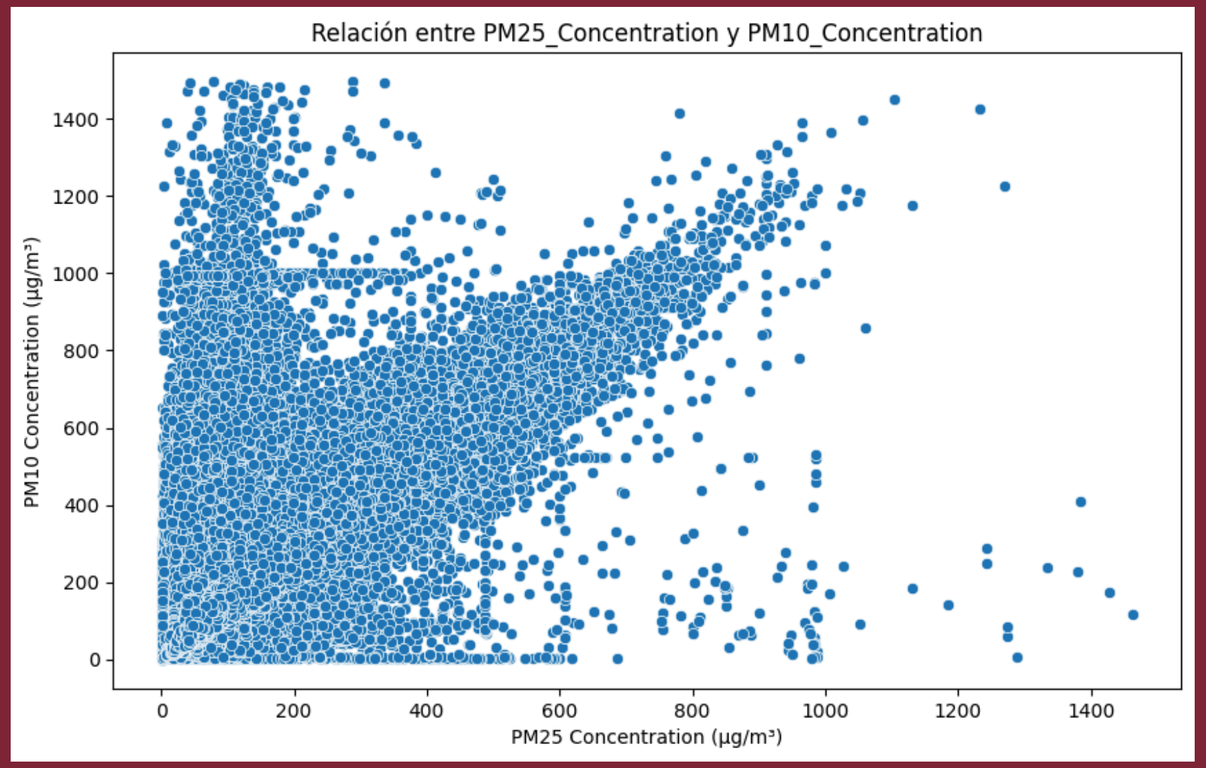
HIPÓTESIS 2.1 - CICLICIDAD RELACIONADA A ESTACIONES

- **Posible explicación:**
  - En primavera y verano (marzo-agosto), las condiciones climáticas como tormentas, lluvias intensas o altas temperaturas podrían afectar las estaciones de monitoreo, causando más datos nulos. Por ejemplo, mayo y junio son meses típicamente lluviosos en muchas regiones de China, lo que podría interrumpir las mediciones.
  - En otoño e invierno (septiembre-febrero), las condiciones podrían ser más estables para las estaciones de monitoreo, o los datos podrían ser más prioritarios debido a los altos niveles de contaminación (como se vio en la Hipótesis 4), lo que reduce la cantidad de datos nulos.
  - También podría haber factores técnicos, como mantenimientos programados en primavera/verano, cuando los niveles de contaminación son más bajos (según la Hipótesis 4), y menos interrupciones en invierno, cuando los datos son más críticos.



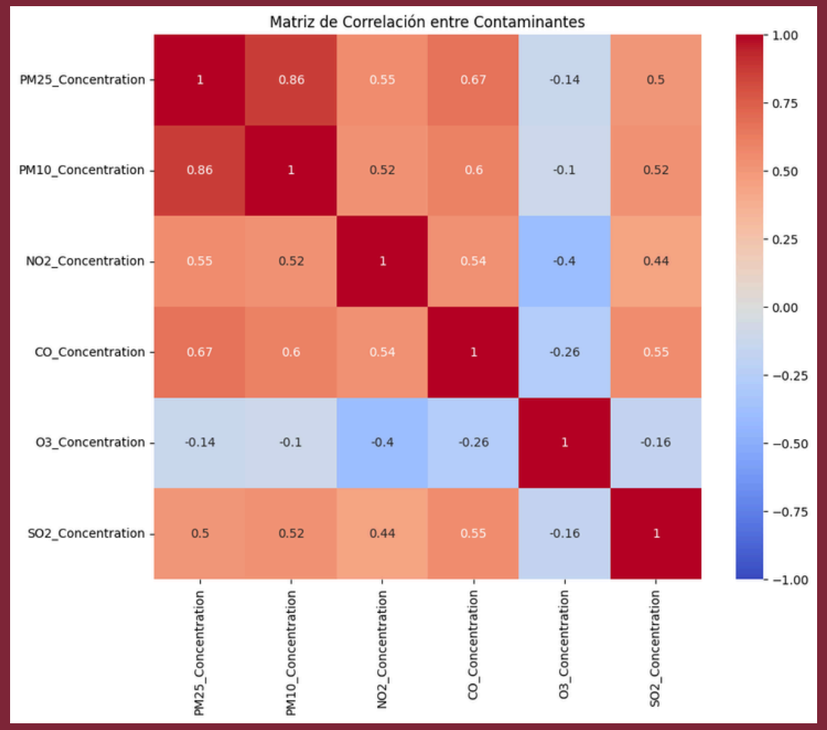
HIPÓTESIS 3 - DEPENDENCIA ENTRE CONTAMINANTES

- **Enunciado:** Algunos contaminantes están correlacionados.
- **Justificación:** PM2.5 y PM10 (0.86) por fuentes comunes (combustión, tráfico). O3 con correlación negativa (formación fotoquímica).



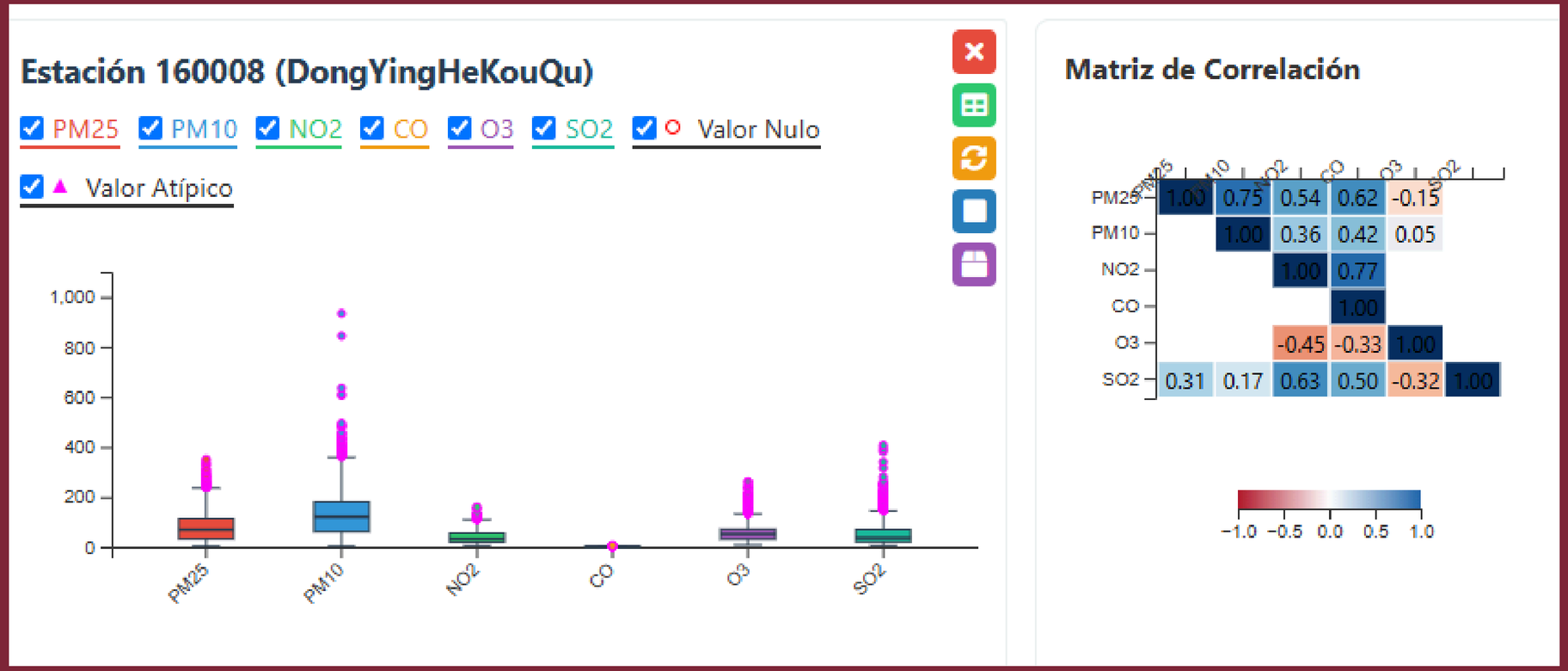
HIPÓTESIS 3 - DEPENDENCIA ENTRE CONTAMINANTES

- **Enunciado:** Algunos contaminantes están correlacionados.
- **Justificación:** PM2.5 y PM10 (0.86) por fuentes comunes (combustión, tráfico). O3 con correlación negativa (formación fotoquímica).



Conclusión :
Existe una fuerte correlación entre varios contaminantes, especialmente entre PM2.5 y PM10, con una correlación de 0.86, lo que sugiere que provienen de fuentes similares. La correlación negativa de O3 con otros contaminantes refleja su comportamiento opuesto, consistente con su formación en condiciones fotoquímicas.

HIPÓTESIS 3 - DEPENDENCIA ENTRE CONTAMINANTES



- **Cluster 1:** Mayor contaminación (200–800 µg/m^3).
- **Ciclicidad:** PM2.5 alto en invierno, bajo en verano.
- **Nulos:** Más frecuentes en primavera/verano.
- Factores geográficos y estacionales son clave.
- Puede haber correlación en diferentes datos de calidad del Aire

GRACIAS