DIPLOMATURA EN CIENCIAS DE DATOS, APRENDIZAJE AUTOMÁTICO Y SUS
APLICACIONES

Análisis y Predicción de Calidad del Aire en una Zona Urbana: Exploración de Datos Multisensor y Modelado Predictivo

Presentado por:
Alejandro Bringas
Lautaro Cabrera
Juan Carlos Recarey
Mariel Palacio

Mentora: Patricia Miranda



Descripción y Objetivo del Proyecto

La calidad del aire es un tema de vital importancia en la agenda mundial de salud ambiental.

Según la OMS, los contaminantes atmosféricos, siendo responsables de 1 de cada 9 muertes a nivel mundial.

Buscamos investigar la interacción entre factores ambientales y contaminantes, ademas de predecir estos contaminantes en función de otros contaminantes y sensores, en una zona contaminada en Italia.



Descripción del Dataset y Objetivo de Estudio



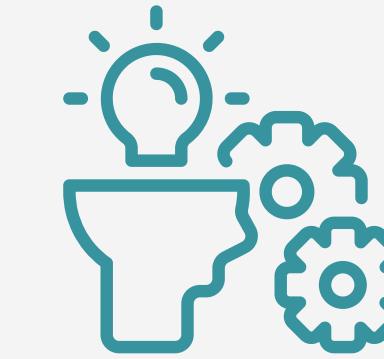
Dataset

- 9471 filas y 17 columnas, con promedios horarios de **respuestas y concentraciones de gases**, obtenidas por un dispositivo multisensor de gases instalado en una zona significativamente contaminada en Italia (Milán)
- Periodo de **1 año** a partir del mes de marzo de 2004 y abril de 2005
- Los datos faltantes se han codificado como -200.



Variables

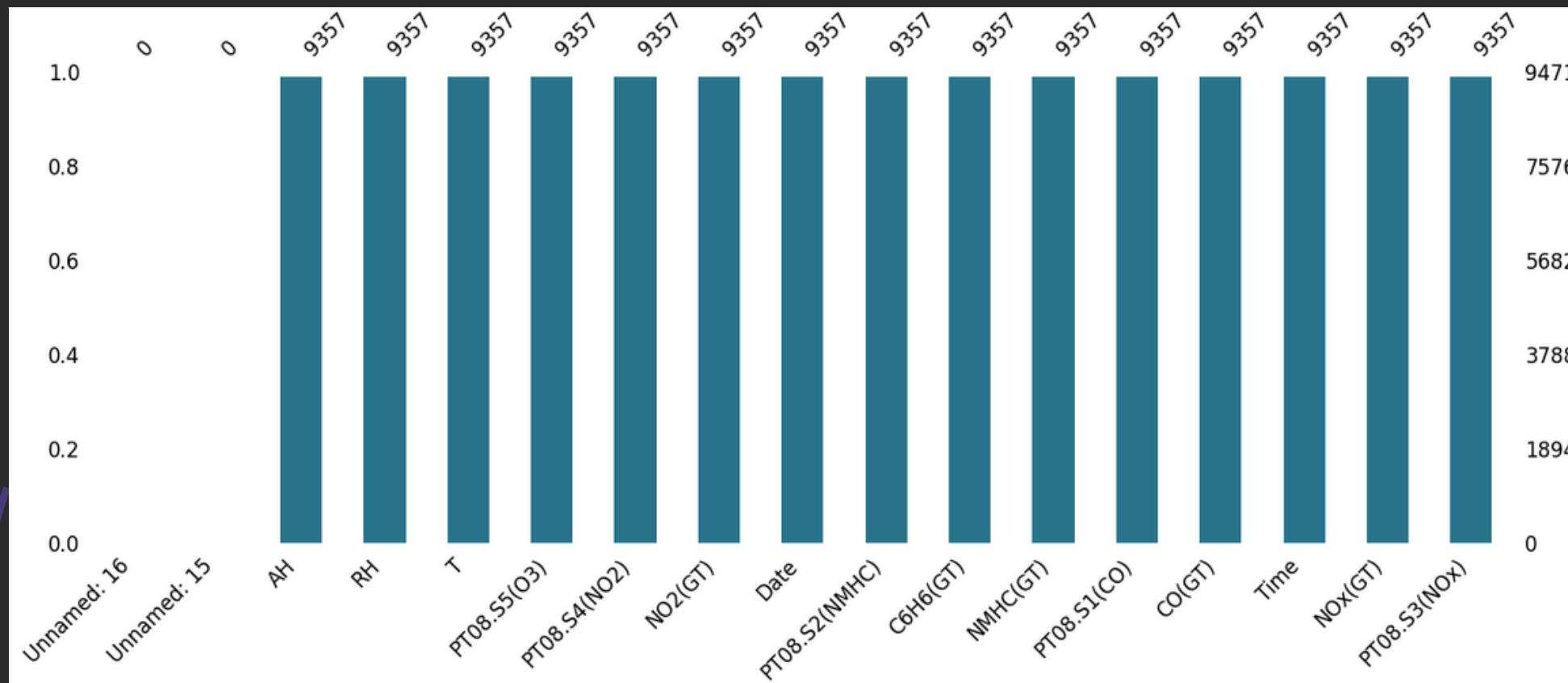
- Fecha, Hora
- Contaminantes: CO(GT), NMHC (GT), C6H6(GT), NOx(GT), NO2(GT)
- Sensores: PT08.S1, PT08.S2, PT08.S3, PT08.S35
- Climáticas: T (temperatura), HR (Humedad relativa), AH (Humedad Absoluta)



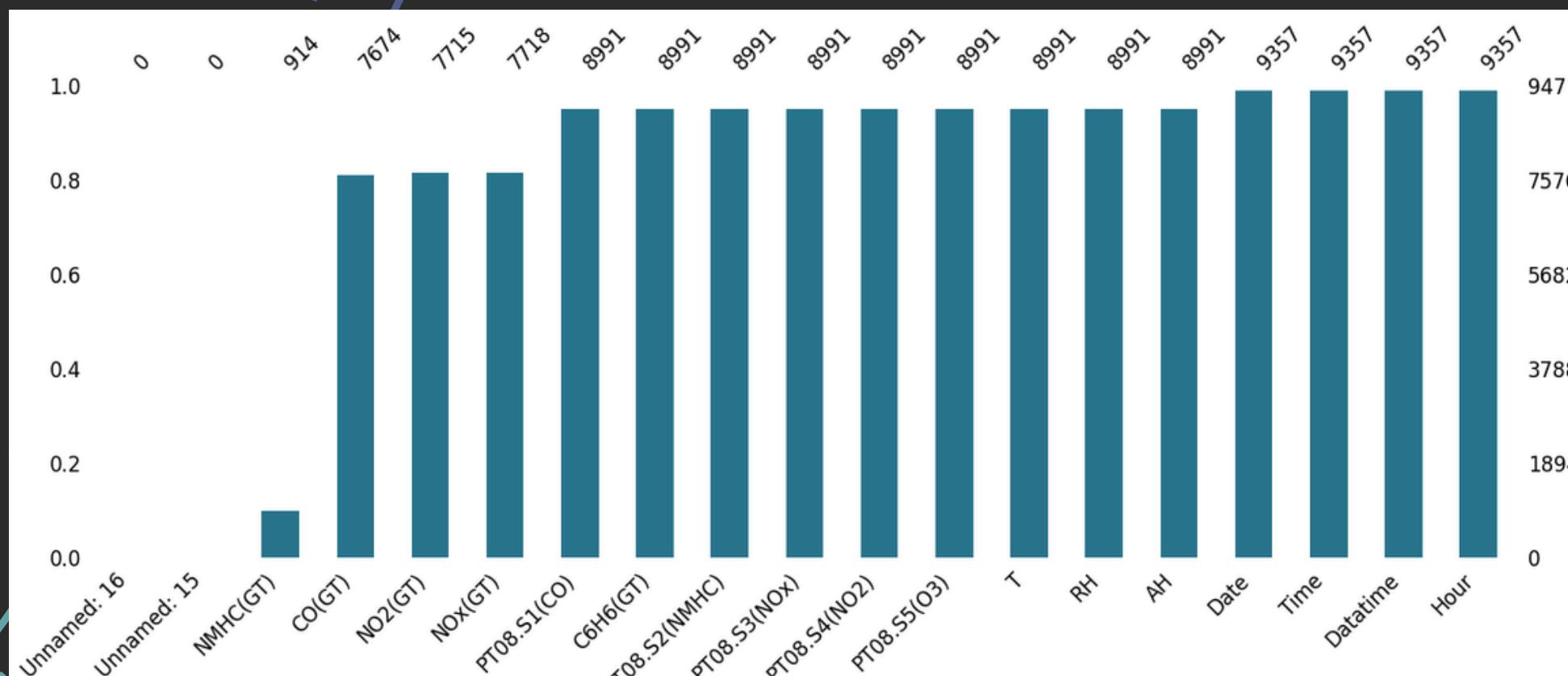
Preguntas

- Mínimos y Máximos de cada contaminante, sus variaciones durante el día.
- Distribución de los datos y sus valores anómalos.
- Asociación, relaciones y correlaciones de los sensores y contaminantes.
- Relación entre condiciones ambientales y los contaminantes

Antes Con -200



Antes Sin -200



Primeros Insights y Conclusiones

Distribución y anomalías de datos detectadas.

Análisis de valores Nulos e Inspección de los -200

Primeros Insights y Conclusiones

Distribución y anomalías de datos detectadas.

Análisis de valores Nulos e Inspección de los -200

Análisis Con los -200

Column Name	Non-Null Count	Data Type
Date	9357	object
Time	9357	object
CO(GT)	9357	float64
PT08.S1(CO)	9357	float64
NMHC(GT)	9357	float64
C6H6(GT)	9357	float64
PT08.S2(NMHC)	9357	float64
NOx(GT)	9357	float64
PT08.S3(NOx)	9357	float64
NO2(GT)	9357	float64
PT08.S4(NO2)	9357	float64
PT08.S5(O3)	9357	float64
T	9357	float64
RH	9357	float64
AH	9357	float64
Unnamed: 15	0	float64
Unnamed: 16	0	float64

Column Name	-200 Count
Date	0
Time	0
CO(GT)	1683
PT08.S1(CO)	366
NMHC(GT)	8443
C6H6(GT)	366
PT08.S2(NMHC)	366
NOx(GT)	1639
PT08.S3(NOx)	366
NO2(GT)	1642
PT08.S4(NO2)	366
PT08.S5(O3)	366
T	366
RH	366
AH	366
Unnamed: 15	0
Unnamed: 16	0

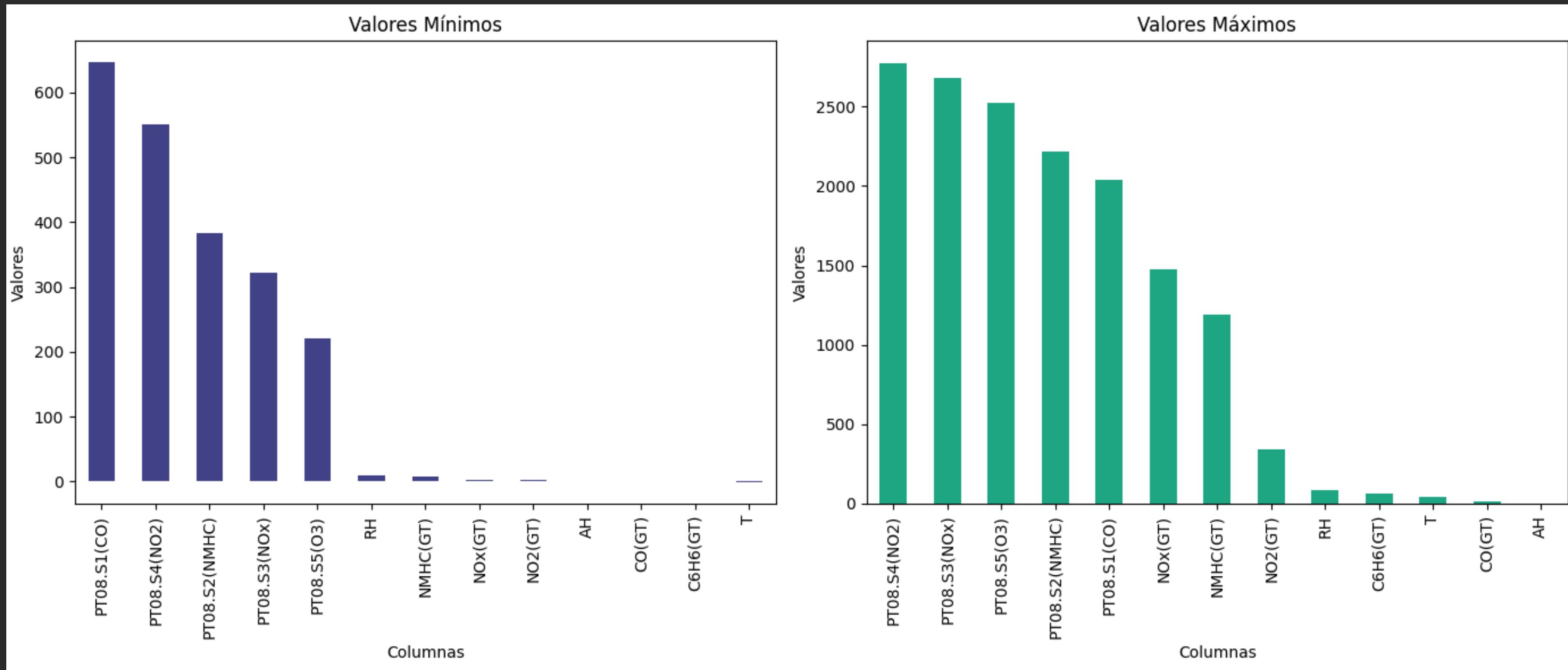
Variable	Min	Max
CO(GT)	-200	11,9
PT08.S1(CO)	-200	2040
NMHC(GT)	-200	1189
C6H6(GT)	-200	63,7
PT08.S2(NMHC)	-200	2214
NOx(GT)	-200	1479
PT08.S3(NOx)	-200	2683
NO2(GT)	-200	340
PT08.S4(NO2)	-200	2775
PT08.S5(O3)	-200	2523
T	-200	44,6
RH	-200	88,7
AH	-200	2,231
Unnamed		
Unnamed		

Variables	count	min	max
CO(GT)	7674	0,1	11,9
PT08.S1(CO)	8991	647	2040
NMHC(GT)	914	7	1189
C6H6(GT)	8991	0,1	63,7
PT08.S2(NMHC)	8991	383	2214
NOx(GT)	7718	2	1479
PT08.S3(NOx)	8991	322	2683
NO2(GT)	7715	2	340
PT08.S4(NO2)	8991	551	2775
PT08.S5(O3)	8991	221	2523
T	8991	-1,9	44,6
RH	8991	9,2	88,7
AH	8991	0,1847	2,231
Unnamed: 15	0		
Unnamed: 16	0		

Análisis Sin los -200

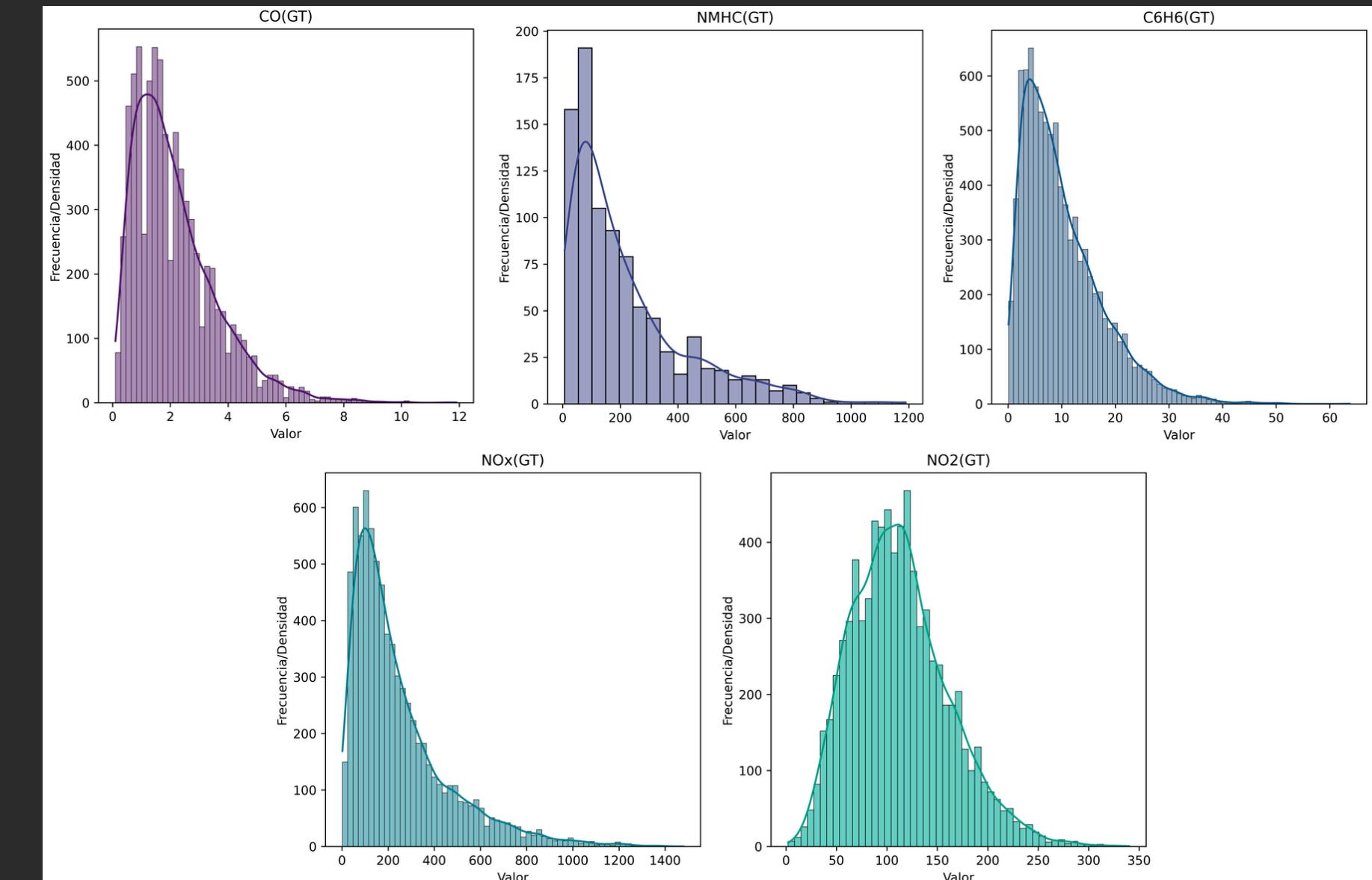
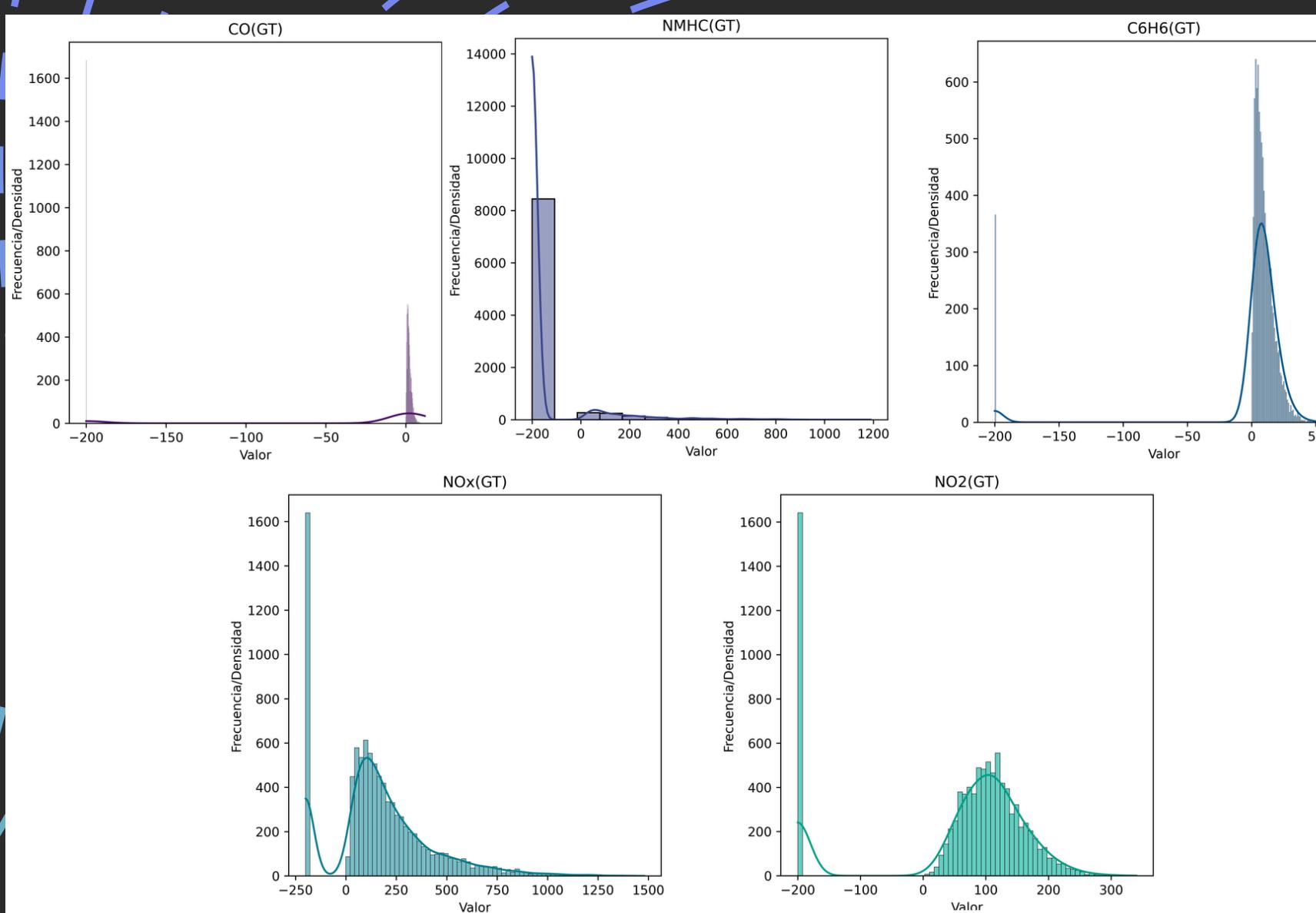
Primeros Insights y Conclusiones

Valores mínimos y máximos de cada contaminante identificados.



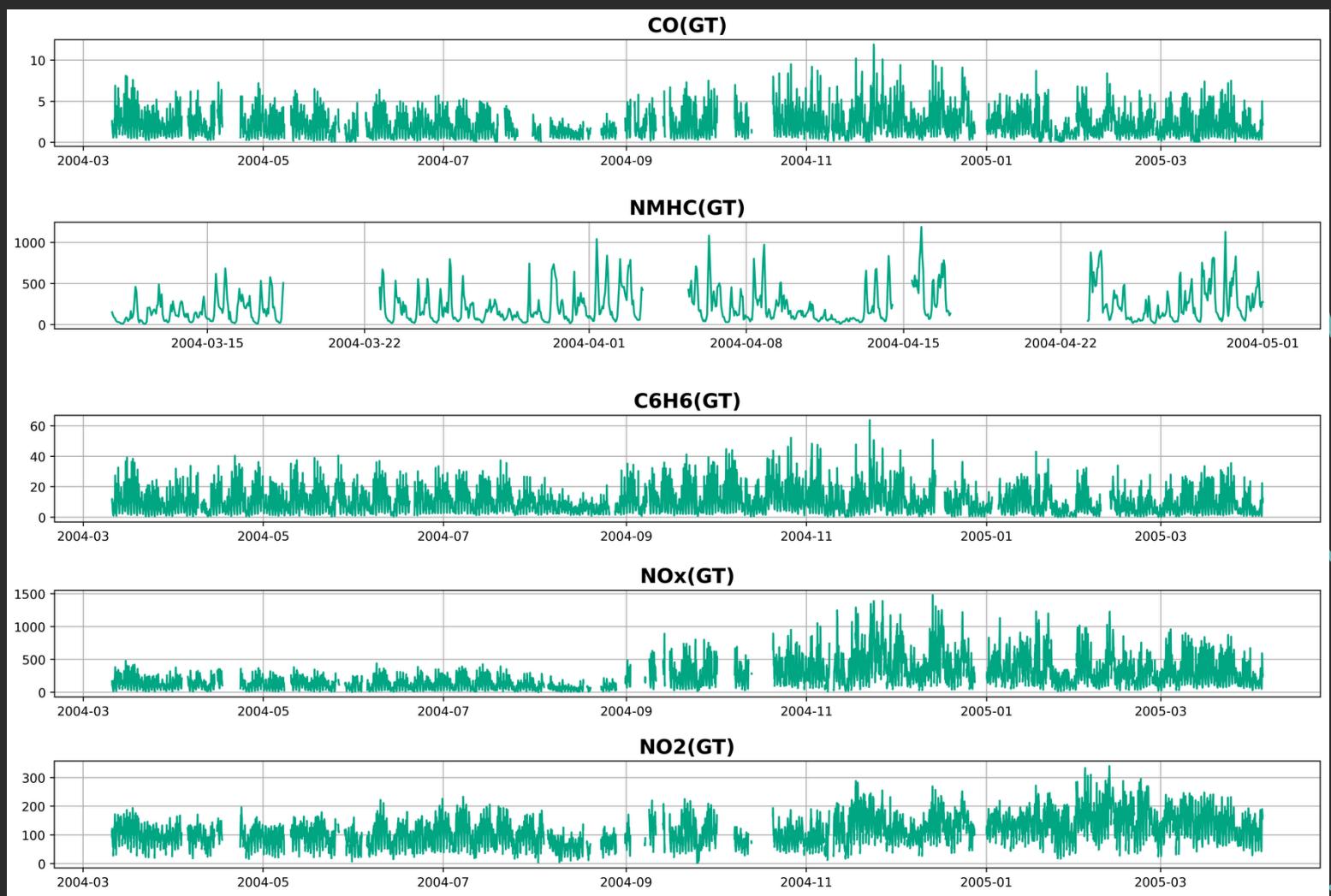
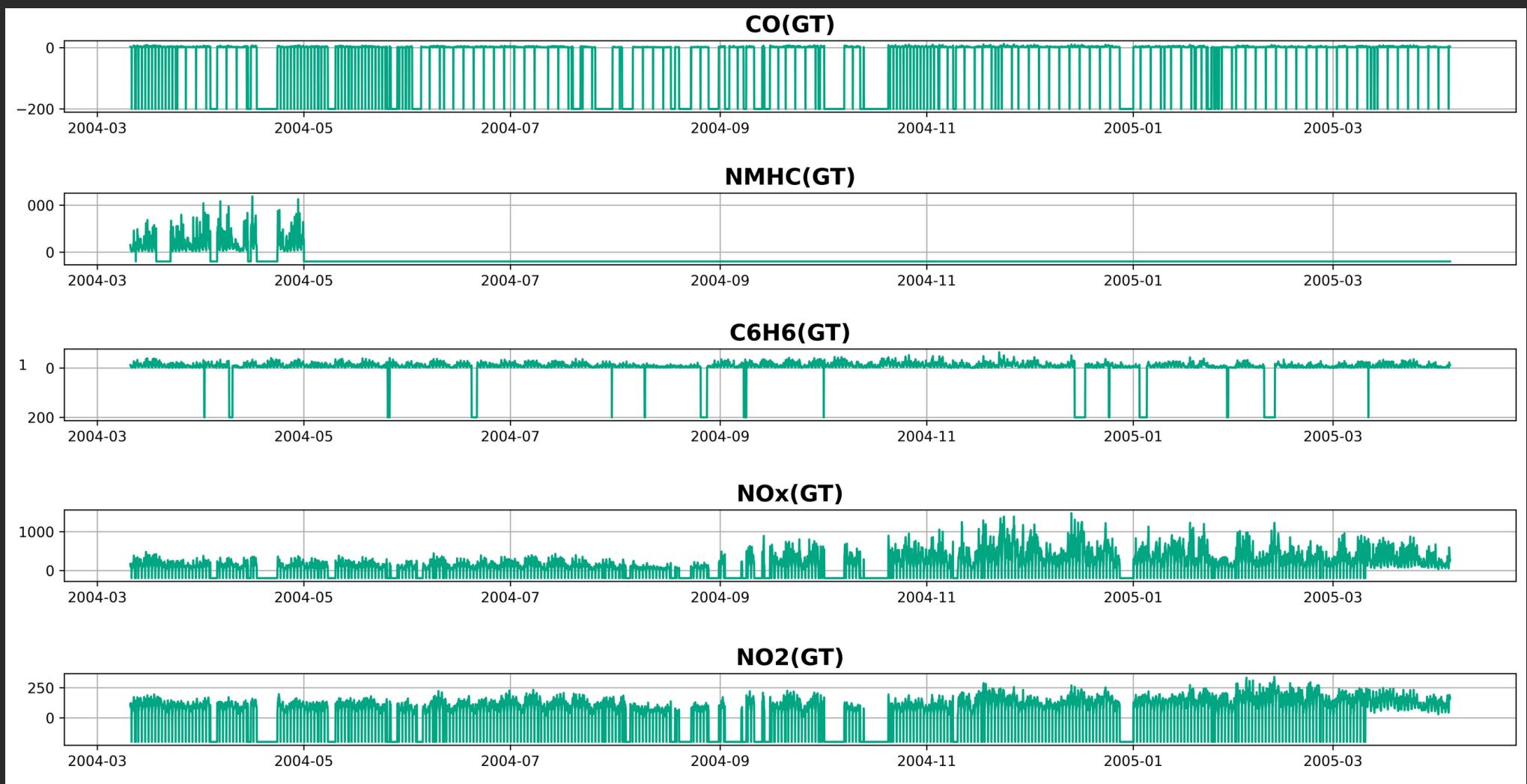
Primeros Insights y Conclusiones

Distribución y anomalías de datos detectadas.



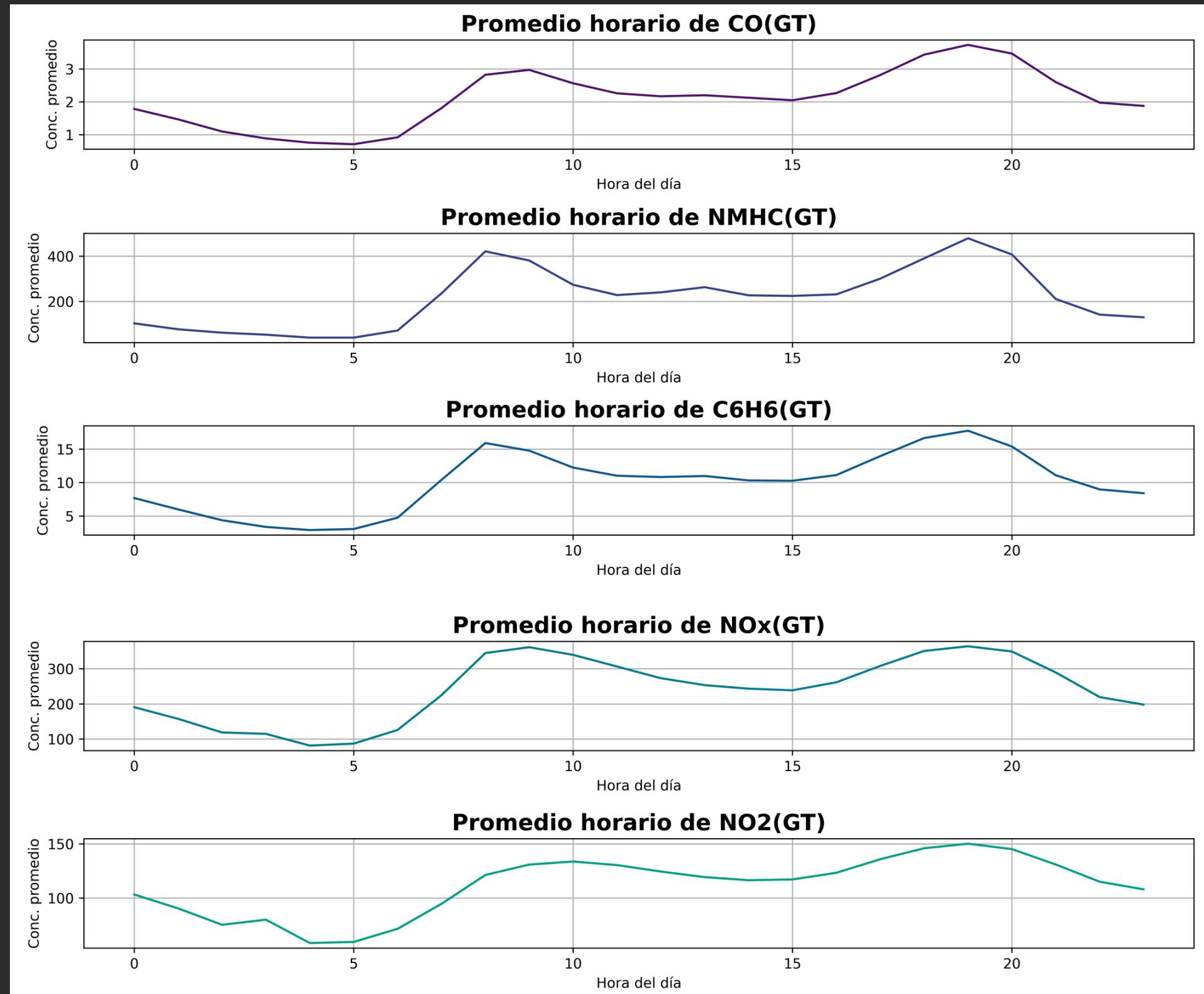
Primeros Insights y Conclusiones

Variaciones diarias de contaminantes analizadas



Primeros Insights y Conclusiones

Variaciones diarias de contaminantes analizadas



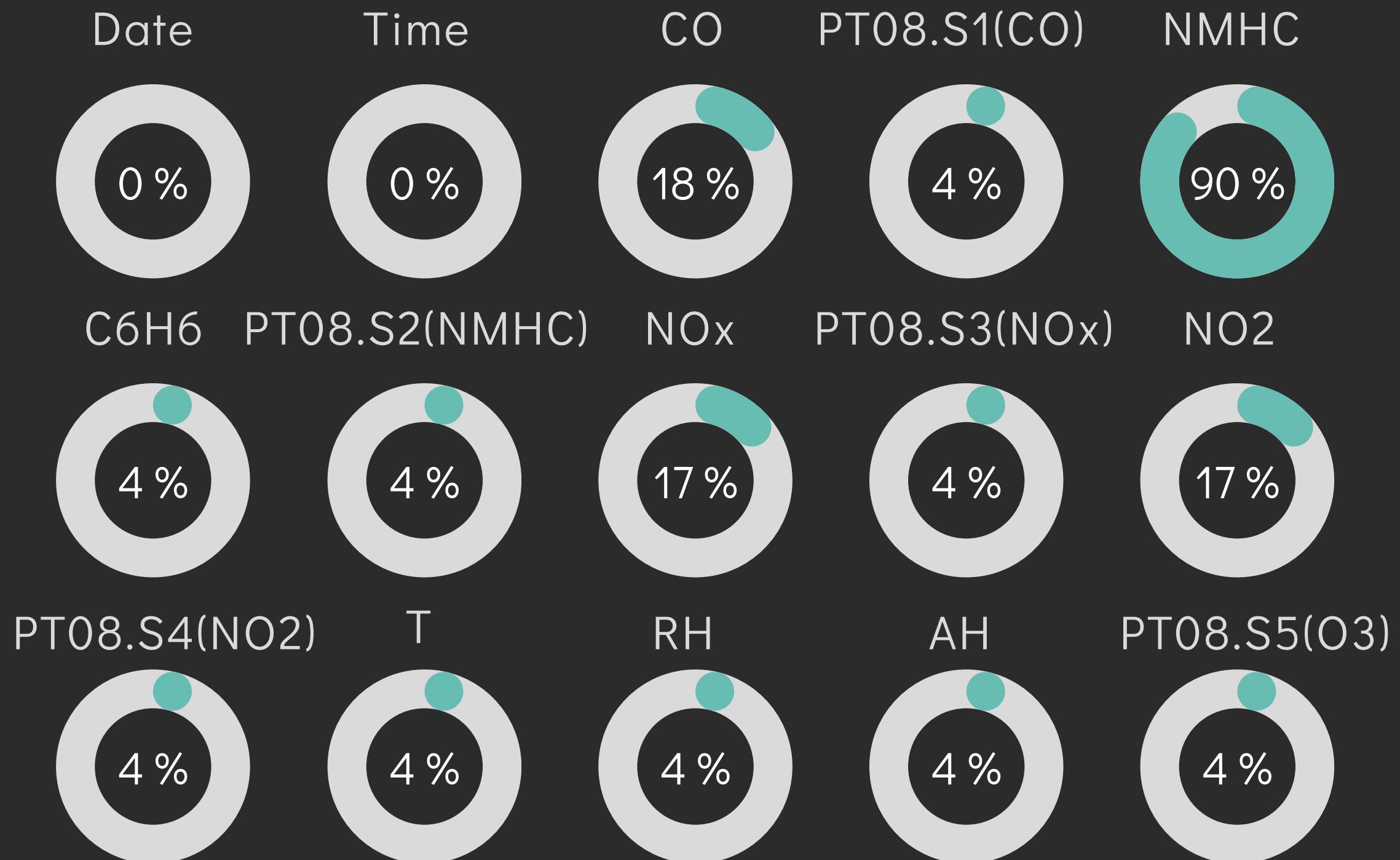
Hasta aquí

Mínimos y Máximos

- VARIABILIDAD
 - Las variables ambientales tienen poca variabilidad entre sus valores en el periodo de tiempo estudiado.
- TRAFICO: POSIBLE CAUSANTE DE PICOS
 - Existen picos de concentración de contaminantes, que parecen responder a las horas de mas tráfico a lo largo del día.
- GRAN IMPACTO DE LOS “-200”
 - Luego del análisis de los gráficos de distribución, vemos que los valores -200 tienen un gran impacto en todas las variables analizadas.

Curación de Datos

Pre-proceso e Eliminación de Valores Nulos con Gran Proporción



- Columnas Completamente Nulas : 'Unnamed: 15','Unnamed: 16'
- NMHC: Variable con 90% de valores -200 = Nulos.



Curación de Datos

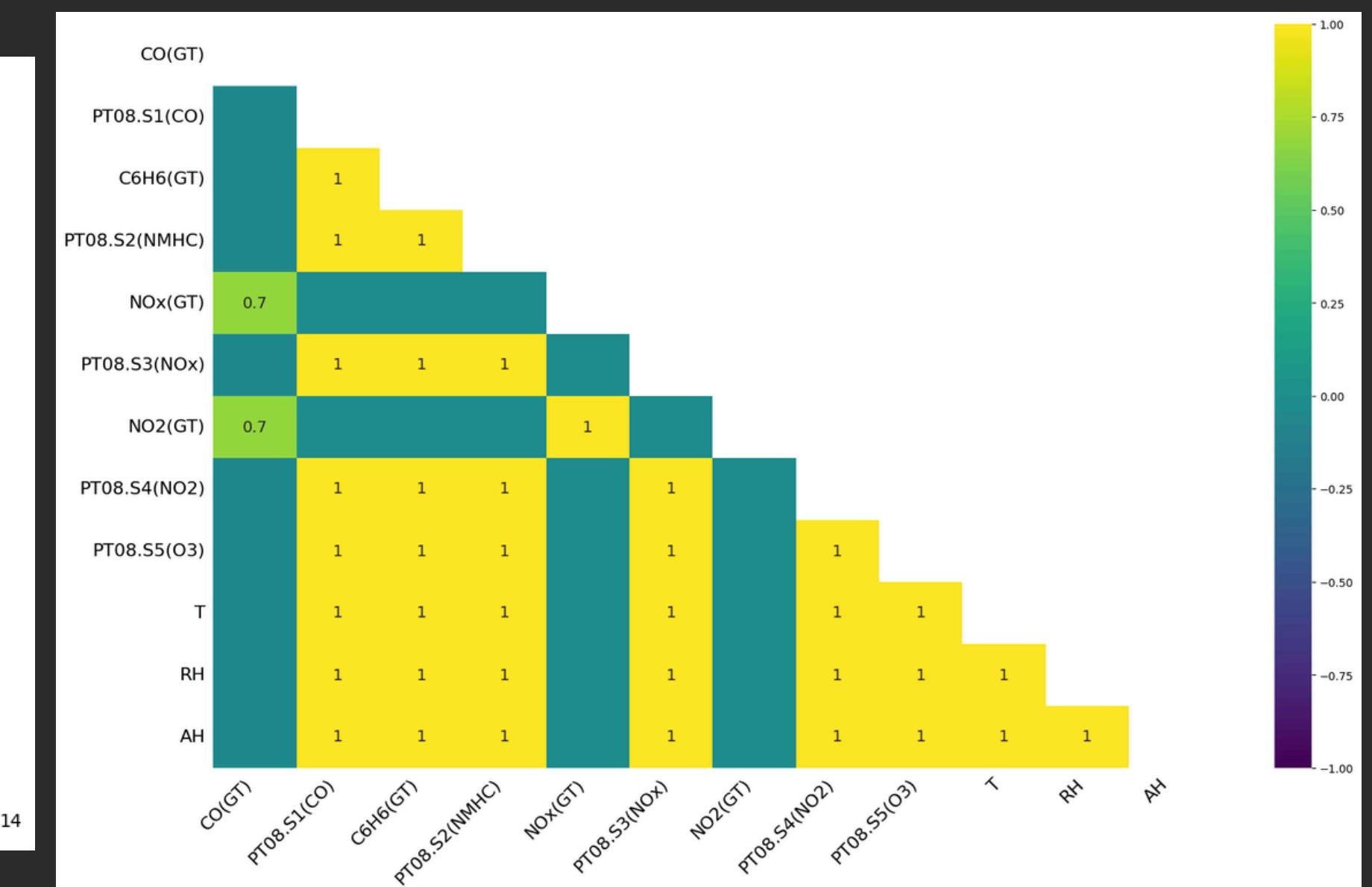
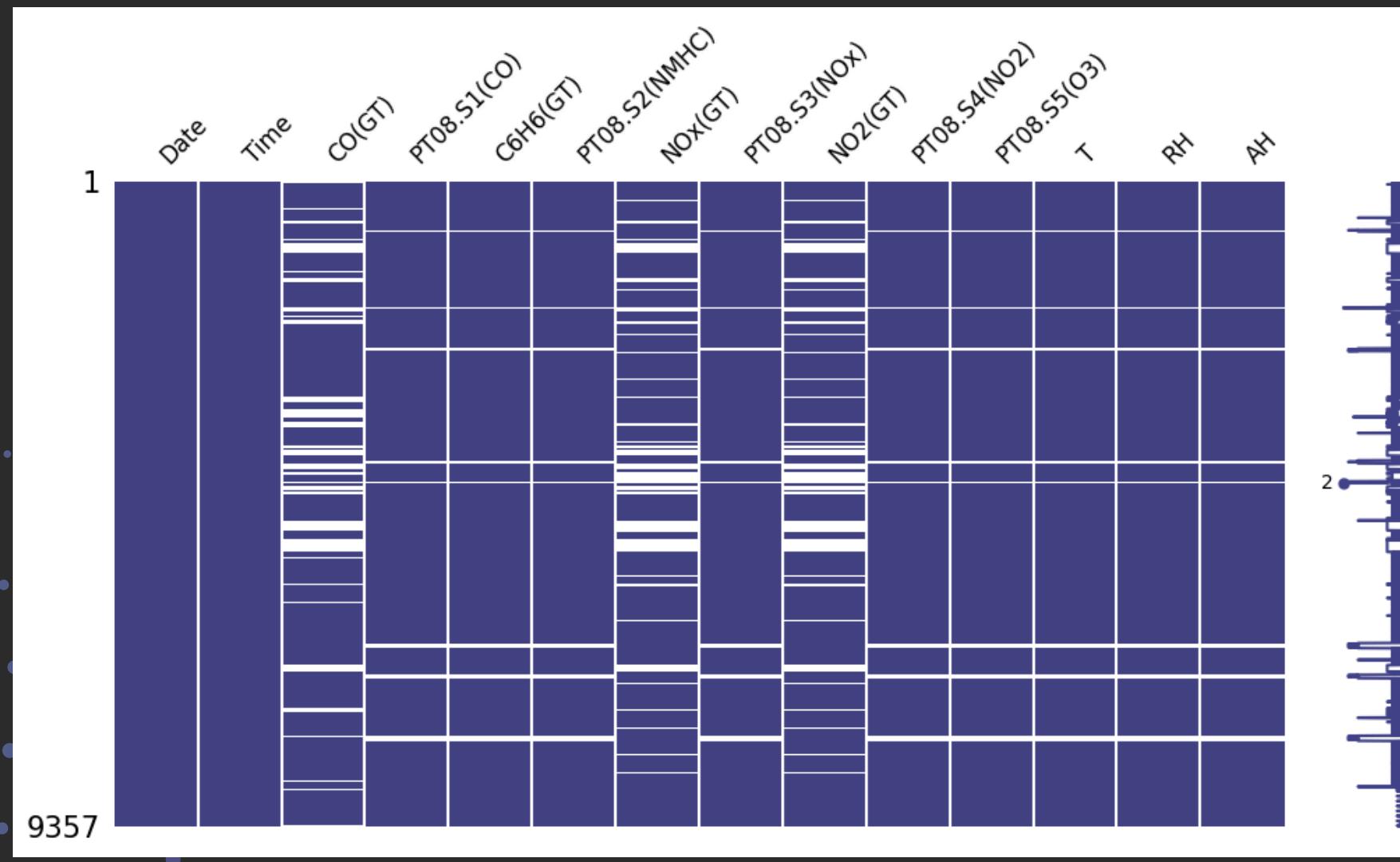
Unificación de Unidades.

Original Nox (ppb)	Convertido Nox(µg/m3)
166	257,99591
103	160,0817996
131	203,599182
172	267,3210634
131	203,599182
89	138,3231084
62	96,3599182
62	96,3599182
45	69,93865031
.....

Original CO(mg/m3)	Convertido(µg/m3)
2,6	2600
2	2000
2,2	2200
2,2	2200
1,6	1600
1,2	1200
1,2	1200
1	1000
0,9	900
0,6	600
....

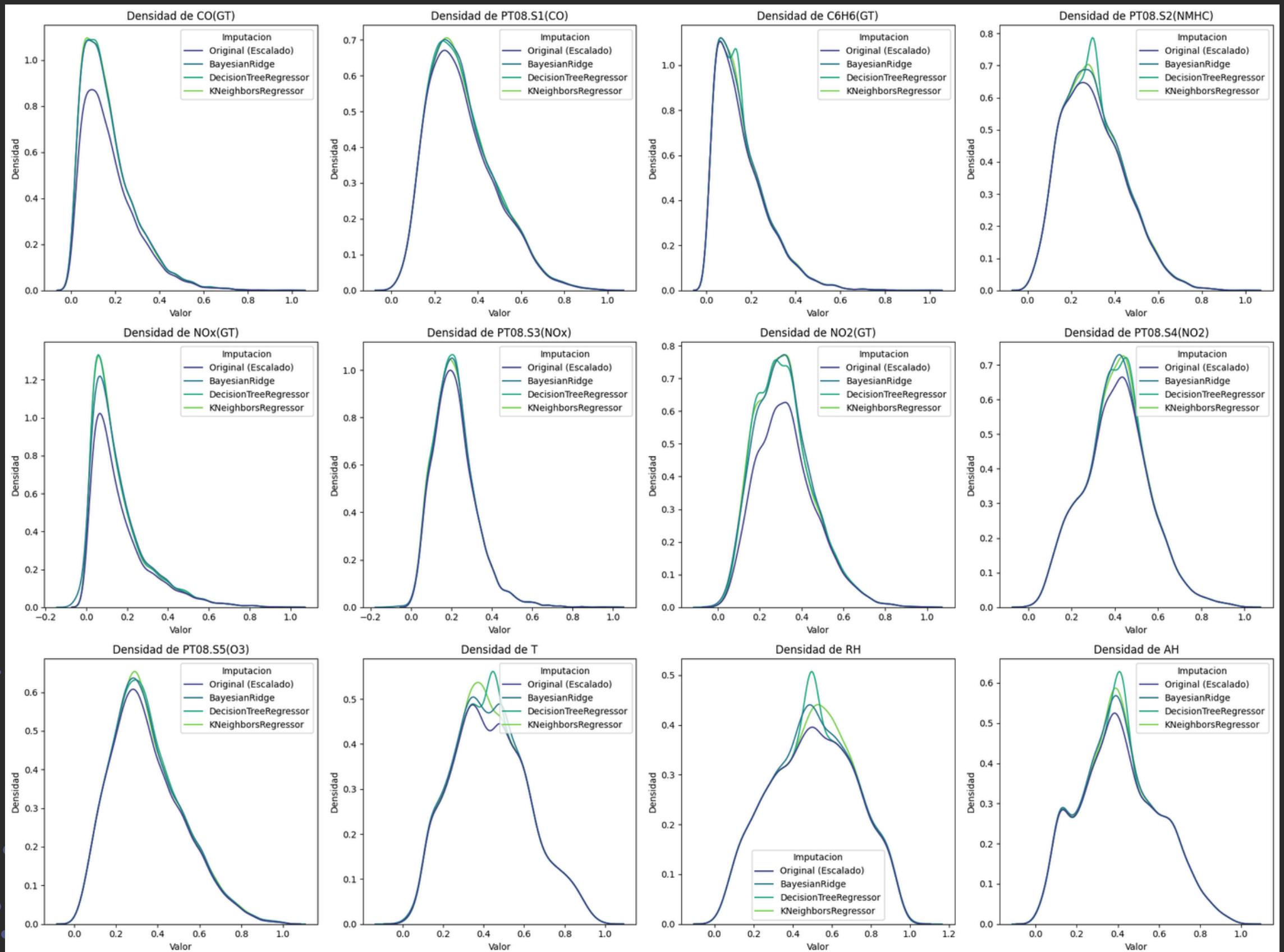
Curación de Datos

Análisis de los Valores Nulos para su Imputación



Curación de Datos

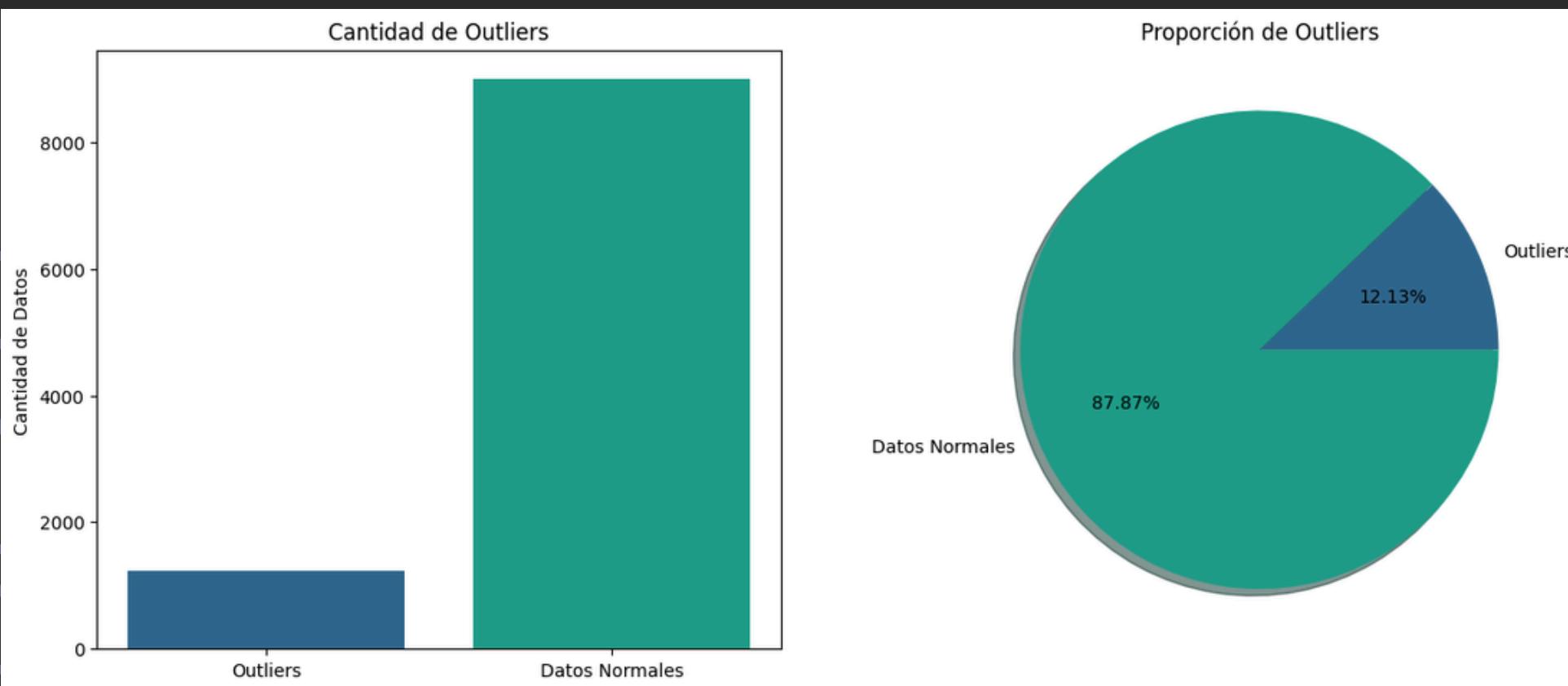
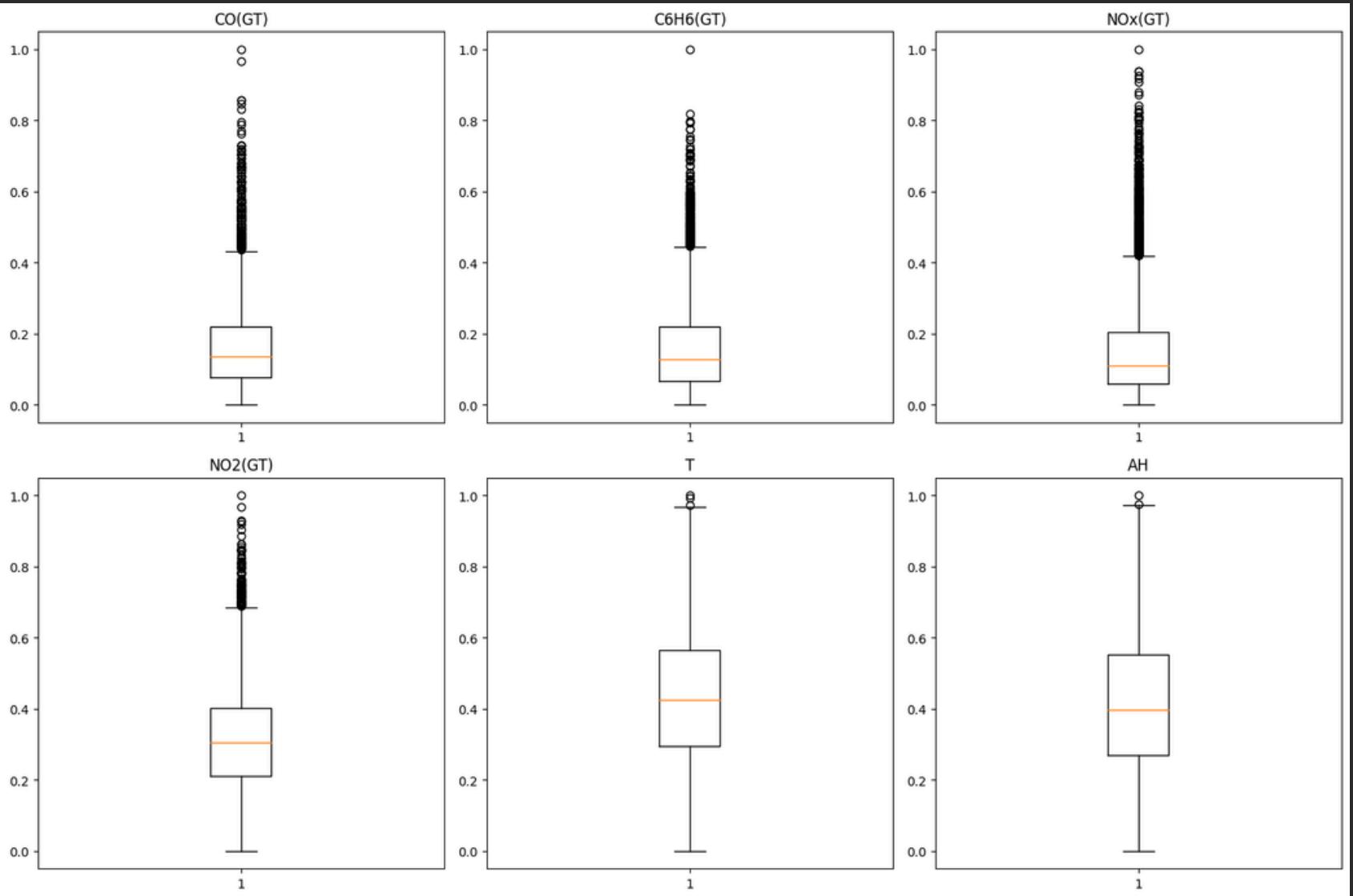
Escalado e Imputación de los Valores Nulos



- Possible existencia de eventos especiales al solo registrar CO,NOx y NO2.
- Escalado de la totalidad de los datos - MinMaxScaler
- Imputación de la totalidad de los datos nulos - KKN

Curación de Datos

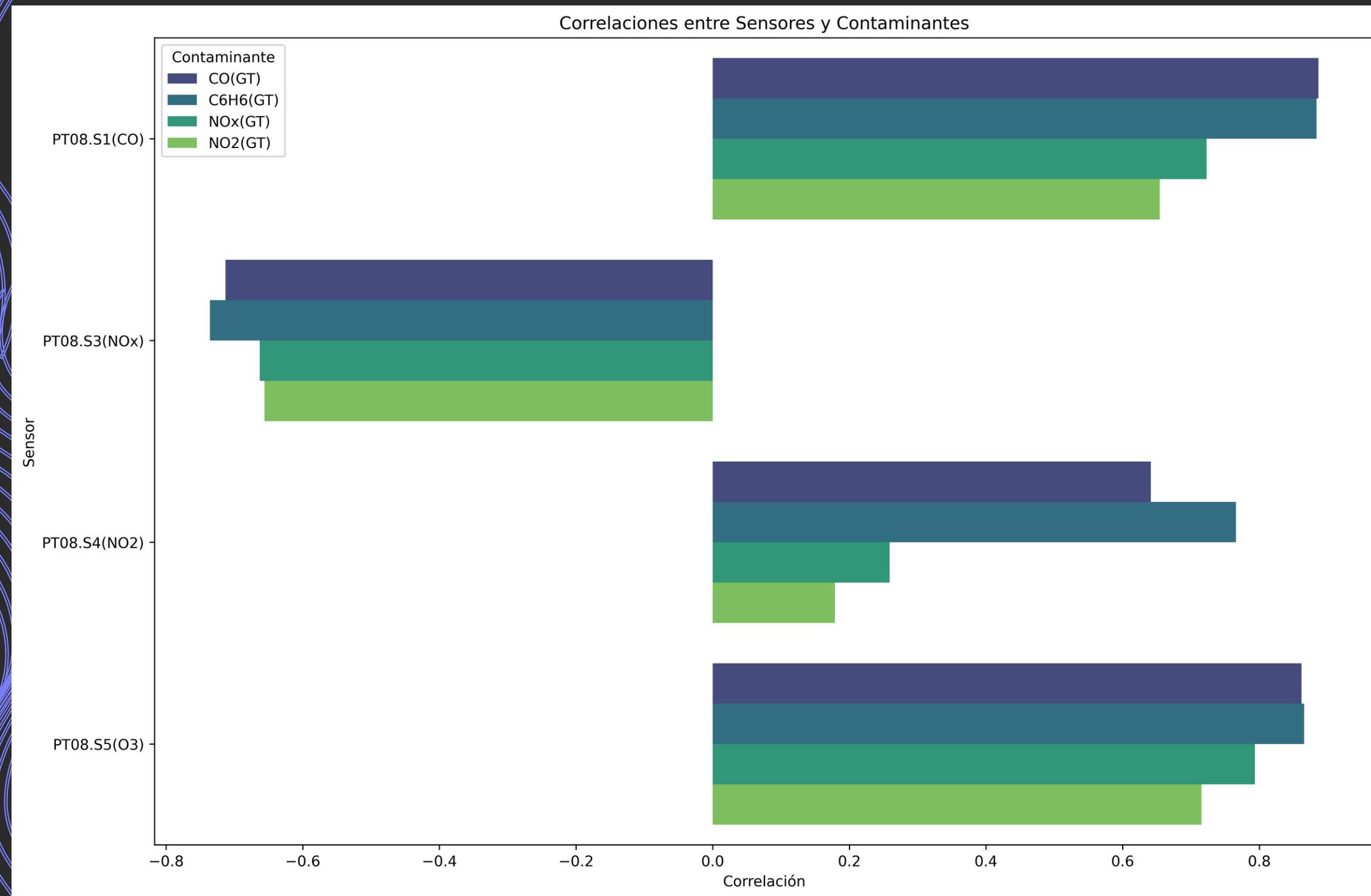
Análisis de los Outliers



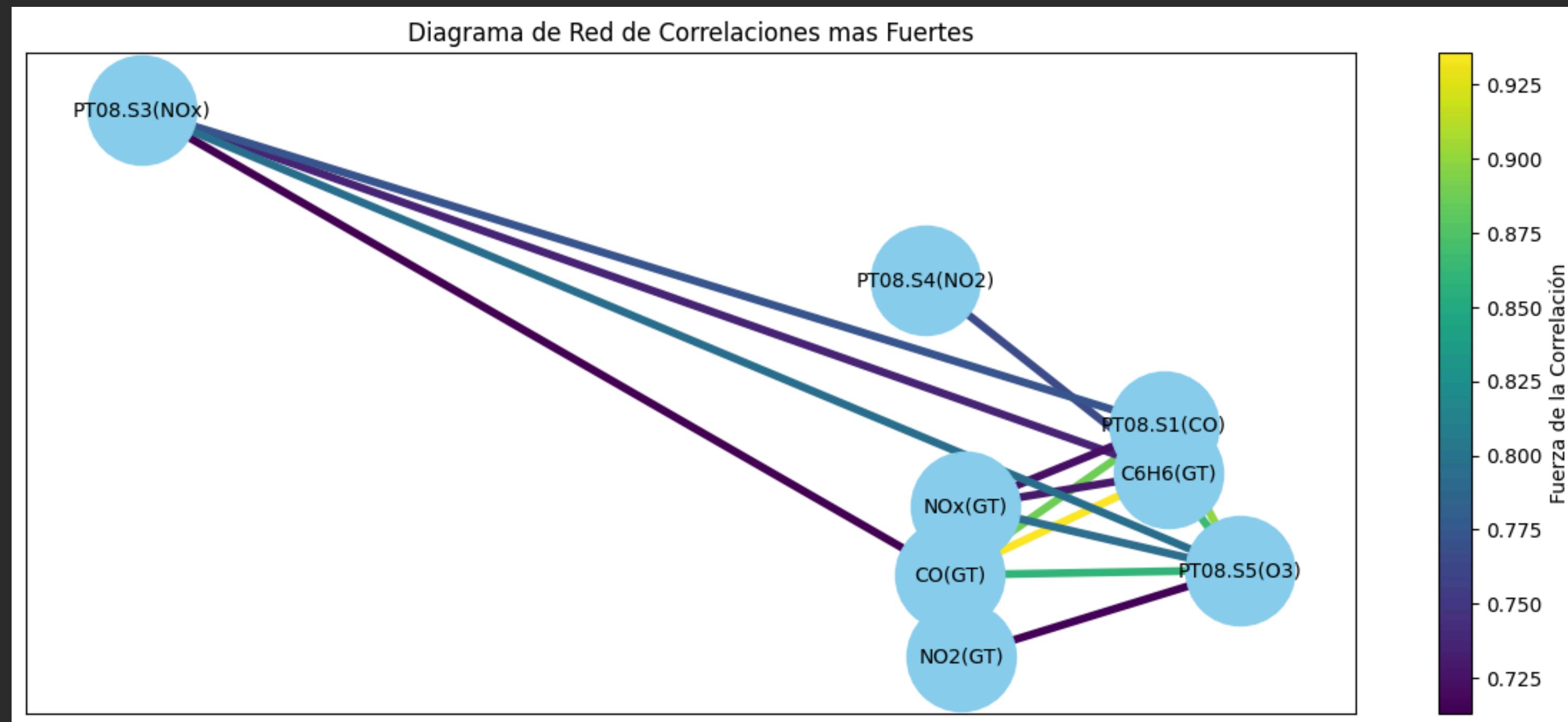
- Proporción no elevada de outliers.
- Análisis de los outliers con cuidado por ser posiblemente claves para el comportamiento del nivel de contaminación del aire .
- No se retiraron los outliers.

Análisis Exploratorio de Datos

Correlaciones entre Sensores y contaminantes.

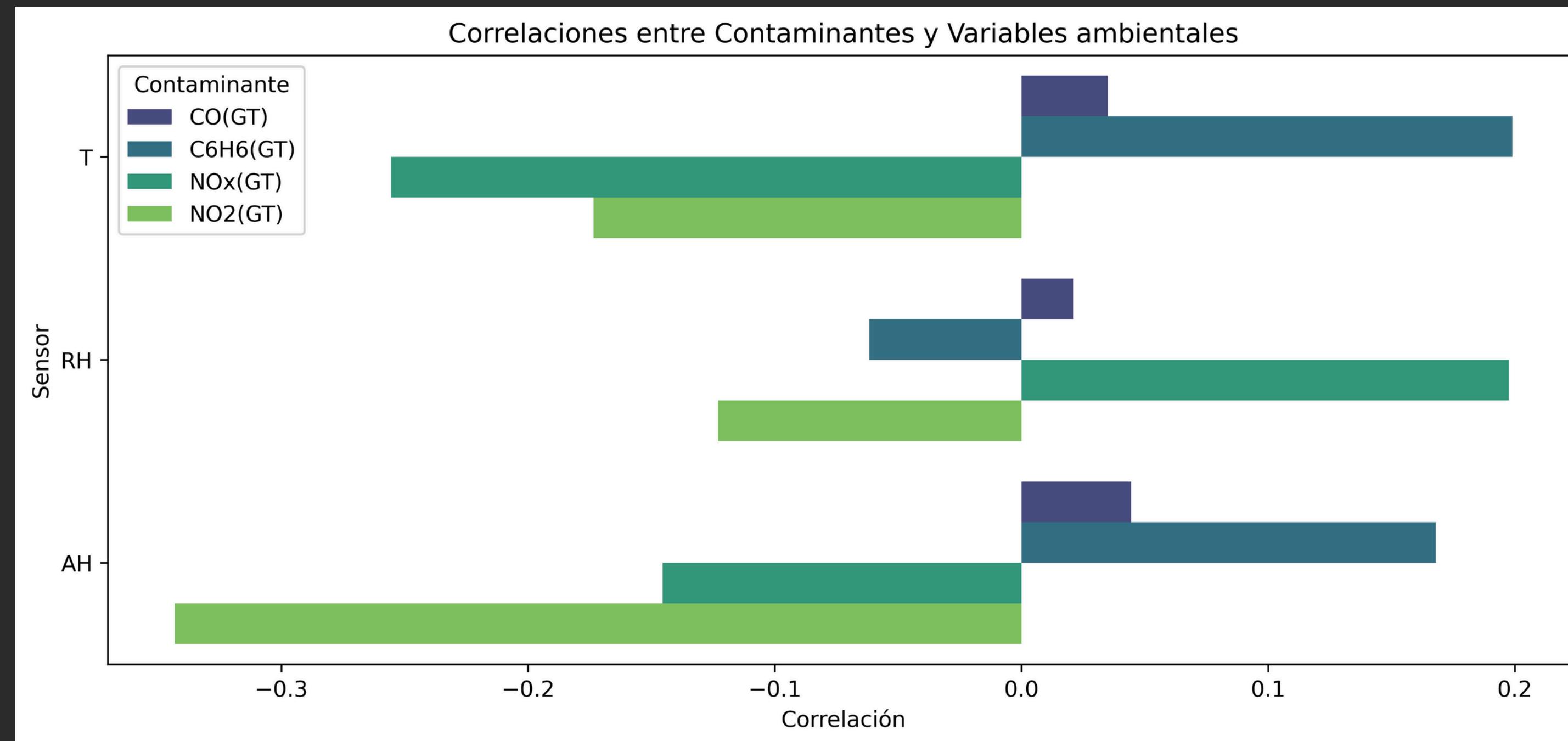


Análisis Exploratorio de Datos



Análisis Exploratorio de Datos

Impacto de las condiciones ambientales en la contaminación



Conclusiones

Relaciones entre Sensores y Contaminantes

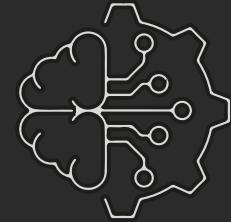
- LOS SENSORES NO ESTÁN EXCLUSIVAMENTE RELACIONADOS CON SU CONTAMINANTE ESPECÍFICO
 - hay correlaciones de distintas intensidades entre sensores y contaminantes, y entre contaminantes entre sí.
- CORRELACIONES NEGATIVAS
 - No se observaron correlaciones negativas significativas entre contaminantes ni con variables climáticas, lo cual es relevante para predicciones climáticas.
 - El sensor PT08.S3 (NOx) muestra correlaciones negativas con NO y NOx, lo que podría indicar un fallo en su funcionamiento.
- POSIBLE FALTA DE CALIBRACION
 - La falta de calibración podría explicar por qué cada sensor se correlaciona con varios contaminantes

Conclusiones

Relaciones entre Contaminantes y Variables Ambientales

- TEMPERATURA (T)
 - Correlación baja y positiva con CO (0.035) y C6H6 (0.20).
 - Correlación negativa con NO₂ (-0.17) y NO_x (-0.26).
- HUMEDAD ABSOLUTA (AH)
 - Correlaciones bajas y positivas con CO (0.044) y C6H6 (0.17).
 - Correlaciones bajas y negativas con NO₂ (-0.34) y NO_x (-0.15)
- HUMEDAD RELATIVA (RH)
 - Correlaciones bajas y positivas con CO (0.021) y NO_x (0.20).
 - Correlaciones bajas y negativas con C6H6 (-0.062) y NO₂ (-0.12)

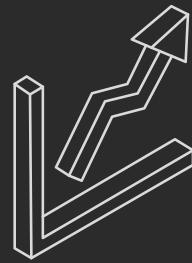
Próximos Pasos



Aplicación de Modelos de Machine Learning para Insights Relevantes.



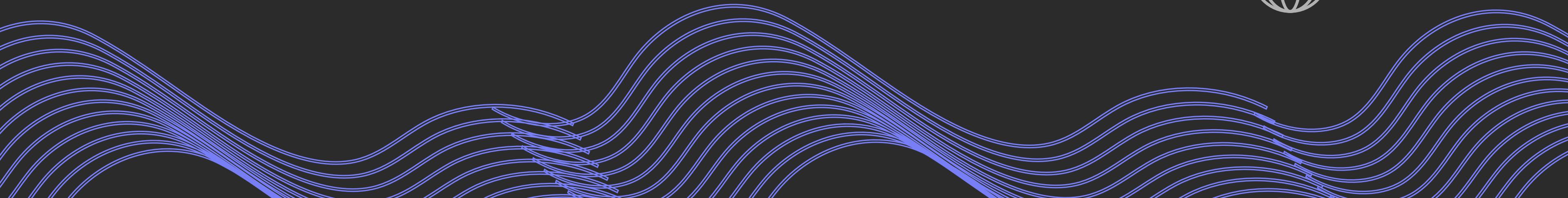
Evaluación de Relación entre Contaminantes y Factores Ambientales.



Optimización de Modelos Predictivos.



Integración de Nuevos Datos para Mayor Precisión.



¡Gracias por ver el video!



¿PREGUNTAS? EN LOS COMENTARIOS

CONTACTO

mariel.palacio@mi.unc.edu.ar