Analisis Exploratorio de Datos - Datos de Calidad del Aire

Albert Daniel Llica Alvarez

Docente: Mag. Ana Maria Cuadros Valdivia

UNSA - Universidad Nacional de San Agustín de Arequipa
Junio de 2025

Índice

Índice de figuras

1. Introducción

Analizando la calidad del aire, este informe detalla un análisis exploratorio de datos recopilados entre mayo de 2014 y abril de 2015 en 43 ciudades chinas. La motivación surge de la necesidad de comprender los factores que influyen en la contaminación atmosférica, como la densidad poblacional, las condiciones climáticas y las emisiones industriales, para desarrollar herramientas de visualización interactiva que apoyen la toma de decisiones.

2. Plan de Análisis

2.1. Motivación y Contexto

La calidad del aire es un factor crítico para la salud pública, el medio ambiente y la formulación de políticas urbanas. En regiones urbanas densas, como las ciudades de China, los contaminantes como PM2.5, PM10, NO2, CO, O3 y SO2 tienen un impacto significativo en la calidad de vida [World Health Organization, 2016]. El dataset analizado, que abarca mediciones horarias de calidad del aire en 43 ciudades chinas entre mayo de 2014 y abril de 2015, presenta desafíos como valores nulos (8.16 % en PM10 y 14.74 % en presión atmosférica) y datos sucios, incluyendo valores extremos (PM25 hasta 1463 $\mu g/m^3$) y valores no físicos [Zheng et al., 2015].

El data wragling del dataset reveló patrones importantes: una ciclicidad estacional en PM2.5, con picos en invierno debido a factores como la quema de carbón y condiciones de baja dispersión [He et al., 2017], y correlaciones significativas entre contaminantes, como PM2.5 y PM10 (0.873) o SO2 con otros contaminantes (0.529–0.570). Además, variables geográficas (latitud y longitud) y meteorológicas (temperatura, velocidad del viento) influyen en las concentraciones de contaminantes [Tai et al., 2010]. Estos patrones sugieren que los modelos de imputación deben capturar dependencias temporales, espaciales y correlaciones entre variables.

Dado que los modelos de imputación, varían en precisión dependiendo de las características del dataset [Junninen et al., 2004, Gong and Ordieres-Meré, 2018], comparar sus resultados visualmente puede mejorar la calidad de los datos imputados. Este proyecto propone desarrollar un visualizador interactivo que facilite la comparación de modelos de imputación preentrenados, permitiendo a los usuarios evaluar cuál modelo preserva mejor los patrones temporales, espaciales y correlaciones del dataset, apoyando la toma de decisiones informadas.

2.2. Preguntas de Hipótesis

- ¿Influye la zona geográfica en el aumento o disminución de los niveles de calidad del aire?

- ¿Se presentan patrones cíclicos en los datos de calidad del aire a lo largo del tiempo?
- ¿Algunos registros de calidad del aire dependen de otros?

2.3. Objetivos del Análisis

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

3. Fuente de Datos

El dataset abarca datos recopilados durante un año (del 1 de mayo de 2014 al 30 de abril de 2015) en 4 ciudades principales de China —Beijing, Tianjin, Guangzhou y Shenzhen— y 39 ciudades cercanas dentro de un radio de 300 km. Estas ciudades están agrupadas en dos clústeres:

- **Clúster A**: 19 ciudades cercanas a Beijing.
- **Clúster B**: 24 ciudades cercanas a Guangzhou.

El dataset consta de seis partes principales:

1. **City Data (city.csv)**: Información de 43 ciudades, incluyendo ID, nombre (chino e inglés), coordenadas (latitud y longitud) y clúster (A o B).
2. **District Data (district.csv)**: Detalles de 380 distritos en las 43 ciudades, con ID, nombre y el ID de la ciudad correspondiente.
3. **Air Quality Monitoring Station Data (station.csv)**: Información de 437 estaciones de monitoreo de calidad del aire, con ID, nombre, coordenadas y el ID del distrito asociado.
4. **Air Quality Data (airquality.csv)**: 2,891,393 registros horarios de calidad del aire en las 437 estaciones. Incluye concentraciones de seis contaminantes: *PM2.5*, *PM10*, NO_2, *CO*, O_3 y SO_2. Presenta valores faltantes, especialmente en PM10 (45.1 % en Beijing).
5. **Meteorological Data (meteorology.csv)**: 1,898,453 registros horarios meteorológicos a nivel de distrito o ciudad, con variables como clima, temperatura, presión atmosférica, humedad, velocidad y dirección del viento. También contiene valores faltantes (por ejemplo, 24.2 % en la variable *clima* para Beijing).

6. **Weather Forecast Data (weatherforecast.csv)**: 910,576 registros de pronósticos meteorológicos para los próximos dos días, con granularidades temporales de 3, 6 o 12 horas. Incluye información sobre clima, temperatura, nivel de viento y dirección del viento.

3.1. Características Clave

- **Escala**: Gran volumen de datos (millones de registros) que cubren aspectos geográficos, temporales y ambientales.
- **Granularidad**: Datos disponibles a nivel de ciudad, distrito y estación, con registros horarios (para calidad del aire y meteorología) y pronósticos con diferentes granularidades temporales.
- **Aplicaciones**: El dataset ha sido utilizado para inferir la calidad del aire a nivel fino (tanto actual como futuro), así como en tareas de aprendizaje automático como aprendizaje multi-vista, multi-tarea y transferencia.

3.2. Desafíos en los Datos

- **Valores faltantes significativos**, por ejemplo, 45.1 % en los datos de PM10 en Beijing.
- **Datos sucios**, incluyendo valores atípicos o duplicados que pueden deberse a errores en la recolección o publicación de los datos.

3.3. Observaciones Relevantes

- **Distribución de la calidad del aire**: Las ciudades del norte (Beijing y Tianjin) presentan una peor calidad del aire en comparación con Guangzhou y Shenzhen. Las concentraciones de PM2.5 son especialmente altas durante los meses fríos.
- **Meteorología**: Beijing presenta una alta proporción de días soleados (47.67 %) y condiciones como niebla o polvo en aproximadamente un 10 % de los registros.

4. Análisis de Datos

4.1. Descripción de Registros

Cada archivo CSV del conjunto de datos contiene registros que representan distintos tipos de entidades u observaciones. A continuación, se detalla el significado de un registro en cada archivo:

- **airquality.csv**: Un registro representa una medición de calidad del aire en una estación de monitoreo específica en un momento determinado (con granularidad horaria). Incluye las concentraciones de seis contaminantes: PM2.5, PM10, NO_2, CO, O_3 y SO_2.
- **city.csv**: Un registro representa una ciudad. Contiene información como el identificador único de la ciudad (`city_id`), su nombre en chino e inglés, coordenadas geográficas (latitud y longitud) y el clúster al que pertenece (A o B).
- **district.csv**: Un registro representa un distrito dentro de una ciudad. Incluye su identificador (`district_id`), el nombre del distrito y el identificador de la ciudad a la que pertenece (`city_id`).
- **meteorology.csv**: Un registro representa las condiciones meteorológicas observadas en un distrito o ciudad en un momento específico (con granularidad horaria). Contiene variables como el tipo de clima, temperatura, presión atmosférica, humedad, velocidad del viento y dirección del viento.
- **station.csv**: Un registro representa una estación de monitoreo de calidad del aire. Incluye el identificador de la estación (`station_id`), su nombre en chino e inglés, coordenadas geográficas, y el identificador del distrito al que pertenece.
- **weatherforecast.csv**: Un registro representa un pronóstico meteorológico para un distrito o ciudad en un momento futuro. Tiene una granularidad temporal de 3, 6 o 12 horas e incluye información sobre el clima, dirección del viento, temperaturas máxima y mínima, y nivel del viento.

4.2. Tipo de Datos

5. Exploración y Visualización

5.1. ¿Influye la zona geográfica en el aumento o disminución de los niveles de calidad del aire?

Las tablas `city.csv` y `station.csv` contienen información geográfica, incluyendo las coordenadas de latitud y longitud, que pueden asociarse directamente con las concentraciones de contaminantes registradas en `airquality.csv`. En particular, las ciudades están agrupadas mediante un identificador de clúster (`cluster_id`) en `city.csv`, lo cual puede reflejar zonas geográficas con características similares, como el norte y el sur de China.

Esta agrupación espacial permite explorar si existen diferencias sistemáticas en la calidad del aire entre regiones. Por ejemplo, la concentración de partículas finas PM2.5 podría variar entre clústeres debido a factores geográficos como el grado de industrialización, las condiciones climáticas predominantes o la topografía de la región. Por tanto, analizar

Cuadro 1: Tipos de datos de las tablas del dataset

Tabla	Columna	Tipo Actual	Contexto y Razón para el Cambio
Air Quality	station$_i d$	int64	Identificador único de estaciones de monitoreo (437 valores únicos). Se usa para unir con `station.csv` o filtrar datos por estación. Como identificador categórico, no se realizan operaciones matemáticas. `category` reduce el uso de memoria (crucial para 2.89M registros) y refleja su naturaleza no numérica.
	time	object	Marca temporal de la medición de calidad del aire (horaria). Esencial para análisis de series temporales. `datetime64` facilita agrupaciones, filtrado por fechas y cálculos de diferencias temporales.
	PM25_Concentration, PM10_Concentration, NO2_Concentration, CO_Concentration, O3_Concentration, SO2_Concentration	float64	Concentraciones de contaminantes. Variables objetivo para predicción de calidad del aire. `float64` es adecuado, aunque se podría evaluar `float32` si se necesita reducir memoria.
City	city$_i d$	int64	Identificador único de ciudades (43 valores). Usado para unir con `district.csv`. Como categórico, ahorra memoria y refleja su uso.
	name_chinese, name_english	object	Nombres de ciudades en chino e inglés. Usados para etiquetas. `object` es adecuado.
	latitude, longitude	float64	Coordenadas geográficas. Útiles para análisis espacial. `float64` es apropiado.
	cluster$_i d$	int64	Identifica el clúster (1=Cluster A, 2=Cluster B). Se recomienda usar `category` por eficiencia.
District	district$_i d, city_i d$	int64	Identificadores para unir con `station.csv` o `meteorology.csv`. No se realizan operaciones matemáticas. `category` reduce memoria.
	name_chinese, name_english	object	Nombres de distritos para reportes o visualización. `object` es adecuado.
Meteorology	id	int64	Identificador para unir con `district.csv`. `category` es más eficiente.
	time	object	Marca temporal horaria. Requiere `datetime64` para análisis temporal.
	weather, wind_direction	float64	Códigos categóricos (17 y 10 valores). Se recomienda convertir a `category` para reflejar su naturaleza.
	temperature, pressure, humidity, wind_speed	float64	Variables meteorológicas continuas. `float64` es adecuado.
Station	station$_i d, district_i d$	int64	Identificadores para uniones. `category` mejora eficiencia.
	name_chinese, name_english	object	Nombres de estaciones. `object` es correcto.
	latitude, longitude	float64	Coordenadas para análisis espacial. `float64` es apropiado.
Weather Forecast	id	int64	Identificador para unir con `meteorology.csv`. `category` ahorra memoria.
	time_forecast, time_future	object	Marcas temporales. `datetime64` permite operaciones temporales.
	frequent, weather, wind_direction	int64, float64	Granularidad temporal y condiciones climáticas. Son categóricas, por lo que se recomienda `category`.
	up_temperature, bottom_temperature, wind_level	float64	Variables continuas. `float64` es adecuado.

la información geográfica junto con las mediciones de calidad del aire puede proporcionar evidencia sobre el impacto del entorno físico en la contaminación atmosférica.

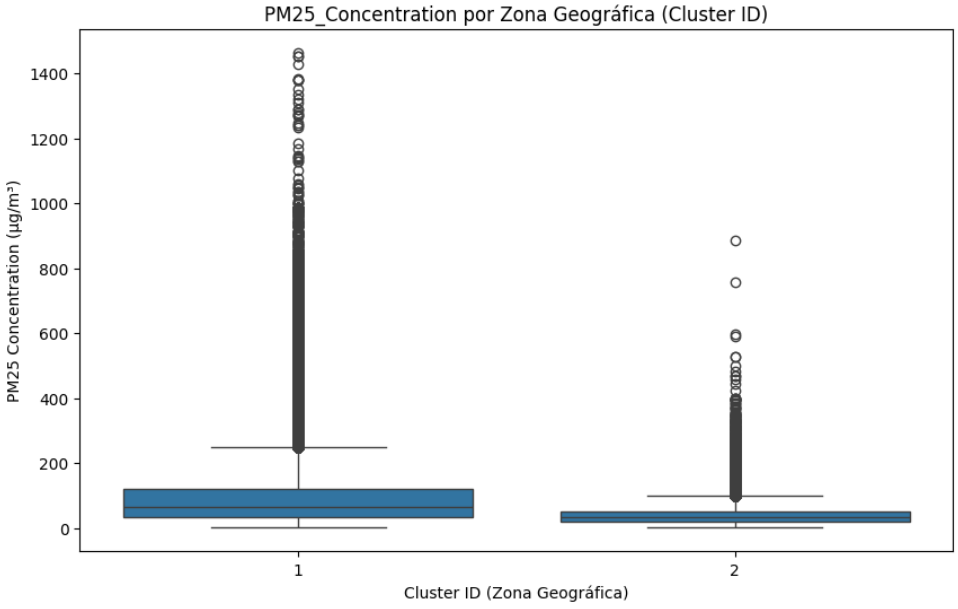

Figura 1: Hipotesis 1

Los diagramas de caja (boxplots) y los valores medios de concentración de PM2.5

confirman que la zona geográfica, representada por la variable `cluster_id`, tiene un impacto significativo en los niveles de contaminación atmosférica. En particular, el `cluster_id` 1, que agrupa ciudades como Beijing y Tianjin (según los datos de `city.csv`), muestra niveles notablemente más altos de concentración de PM2.5 en comparación con el `cluster_id` 2, que incluye ciudades como Shenzhen, Guangzhou y Hong Kong.

Esta diferencia podría atribuirse al hecho de que el `cluster_id` 1 representa regiones del norte de China, caracterizadas por una mayor densidad industrial y el uso intensivo de carbón para calefacción durante los meses fríos. Estas condiciones favorecen la acumulación de contaminantes atmosféricos, especialmente durante episodios de inversión térmica en invierno. En contraste, el `cluster_id` 2 incluye ciudades del sur del país, donde el clima más cálido, la menor dependencia de sistemas de calefacción, y la mayor dispersión de contaminantes contribuyen a una mejor calidad del aire.

Análisis de valores atípicos

Se observaron valores extremos en ambos clústeres, en particular dentro del `cluster_id` 1, con picos de concentración que alcanzan hasta los 1400 . Estos valores atípicos podrían estar relacionados con eventos puntuales, como episodios intensos de esmog. No obstante, dichos extremos no alteran la tendencia general, que indica que el `cluster_id` 1 presenta una mayor contaminación por partículas finas en comparación con el `cluster_id` 2.

Contexto geográfico

Las diferencias observadas se alinean con el contexto geográfico y socioeconómico de las ciudades incluidas en cada clúster. Las ciudades del norte de China tienden a registrar mayores niveles de contaminación debido a su grado de industrialización y a condiciones meteorológicas que dificultan la dispersión de contaminantes. En cambio, las ciudades del sur presentan características climáticas y estructurales más favorables para una mejor calidad del aire, lo que se refleja en las menores concentraciones de PM2.5 registradas.

5.2. ¿Se presentan patrones cíclicos en los datos de calidad del aire a lo largo del tiempo?

La tabla `airquality.csv` contiene una columna `time` que permite realizar un análisis de series temporales sobre las concentraciones de contaminantes, en particular la variable `PM25_Concentration`. Esta dimensión temporal es fundamental para identificar patrones de ciclicidad que podrían estar relacionados con factores estacionales o de comportamiento humano.

Ciclicidad estacional y diaria

Durante los meses de invierno, se espera un incremento en los niveles de contaminación debido al uso intensivo de sistemas de calefacción, especialmente en regiones del norte de China. Este patrón estacional puede reflejarse en aumentos persistentes en la concentración de partículas finas (PM2.5). A nivel diario, es posible detectar picos de contaminación en horarios específicos, como durante las horas punta de tráfico vehicular, debido al incremento en las emisiones asociadas.

Importancia del análisis de series temporales

El análisis de series temporales permite identificar tendencias, ciclos y anomalías en los datos, lo cual es esencial para comprender la dinámica de la calidad del aire. Este tipo de análisis puede ayudar a predecir eventos de alta contaminación, evaluar el impacto de políticas ambientales, y diseñar estrategias de mitigación más eficaces.

Por lo tanto, la columna `time` no solo enriquece el análisis descriptivo, sino que también habilita técnicas avanzadas de modelado predictivo y análisis estacional, cruciales para estudios ambientales y de salud pública.

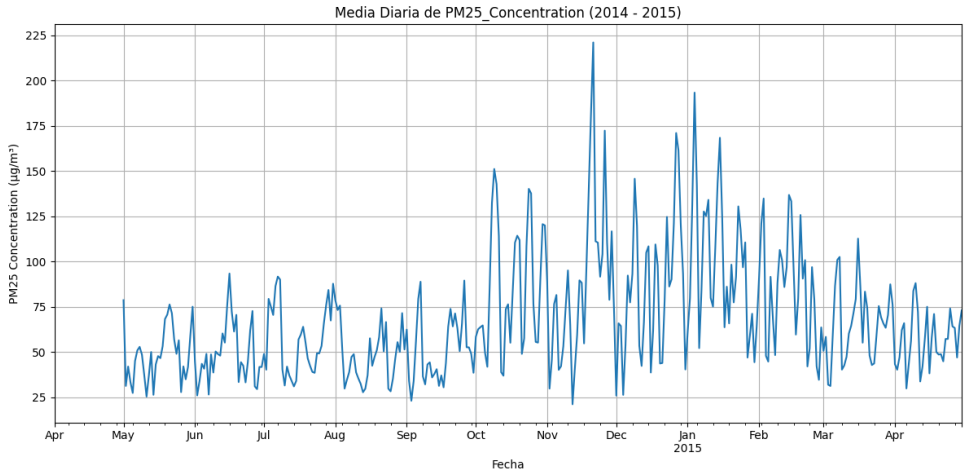

Figura 2: Hipotesis 2

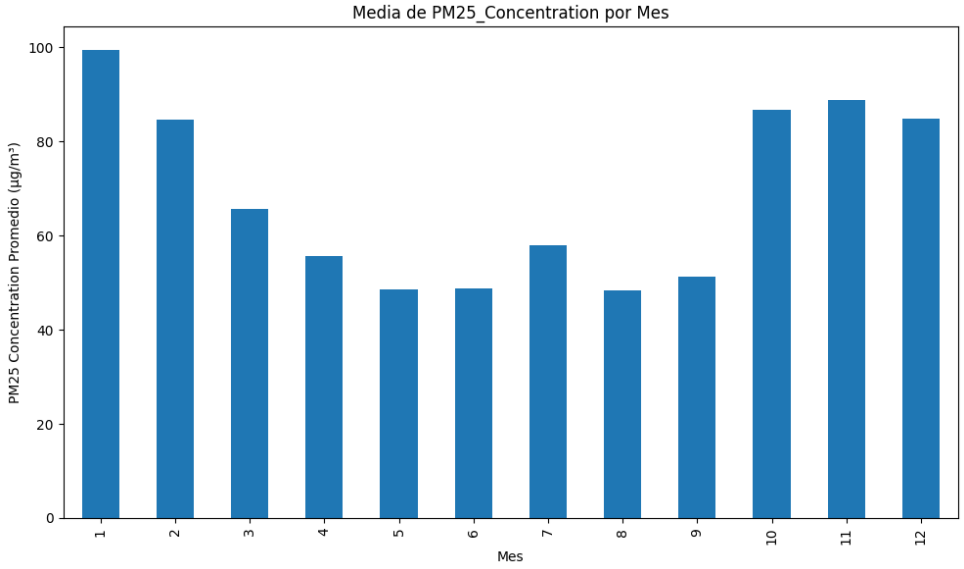

Figura 3: Hipotesis 2.1

Los gráficos obtenidos a partir del análisis de los datos de calidad del aire confirman una ciclicidad estacional clara. En particular, se observa que la concentración de PM2.5 (`PM25_Concentration`) tiende a aumentar durante los meses fríos (invierno) y a disminuir durante los meses cálidos (verano).

Este patrón puede explicarse por varios factores ambientales y socioeconómicos. Durante el invierno, especialmente en ciudades del norte de China como Beijing (según `city.csv`), es común la quema de carbón para calefacción, lo cual incrementa significativamente la emisión de partículas contaminantes. Además, las condiciones atmosféricas como las inversiones térmicas pueden atrapar estos contaminantes cerca del suelo, agravando los niveles de polución. Por el contrario, en verano, la mayor radiación solar, las precipitaciones y los vientos favorecen la dispersión de contaminantes, lo que resulta en concentraciones más bajas.

Contexto temporal

Considerando que los datos abarcan los años 2014 y 2015, y que la fecha actual es el 3 de junio de 2025 (inicio del verano), podemos estimar que los niveles actuales de `PM25_Concentration` se encontrarían en el rango más bajo del ciclo estacional, aproximadamente entre 40 y 50 μg/m^3, de acuerdo con las observaciones históricas.

Conclusión

Los datos y los gráficos confirman la hipótesis inicial: existe una marcada ciclicidad estacional en la calidad del aire, particularmente en los niveles de `PM25_Concentration`.

Este comportamiento debe ser considerado en modelos predictivos y estrategias de mitigación, especialmente en regiones con altos niveles de contaminación durante el invierno.

5.3. ¿Algunos registros de calidad del aire dependen de otros

La tabla airquality.csv contiene múltiples contaminantes (PM25_Concentration, PM10_Concentrati NO2_Concentration, CO_Concentration, O3_Concentration, SO2_Concentration). Es posible que algunos contaminantes estén correlacionados, como PM25_Concentration y PM10_Concentratio ya que ambos están relacionados con partículas en el aire, o NO2_Concentration y CO_Concentration, asociados con emisiones vehiculares.

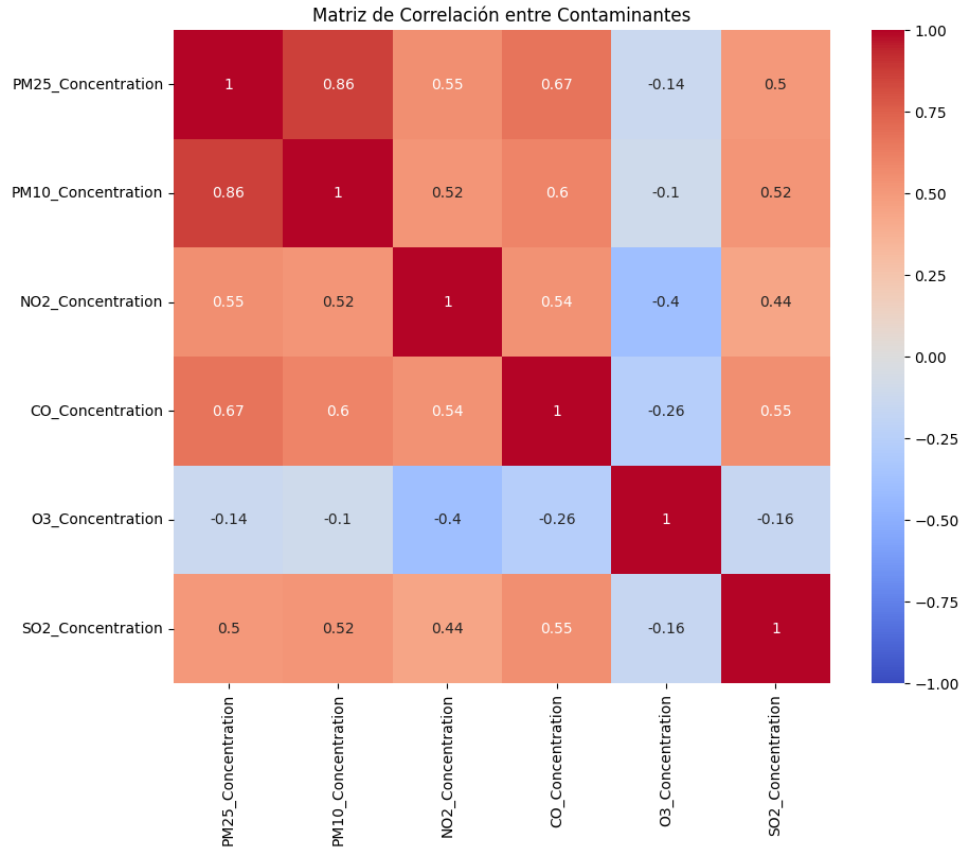

Figura 4: Hipotesis 3

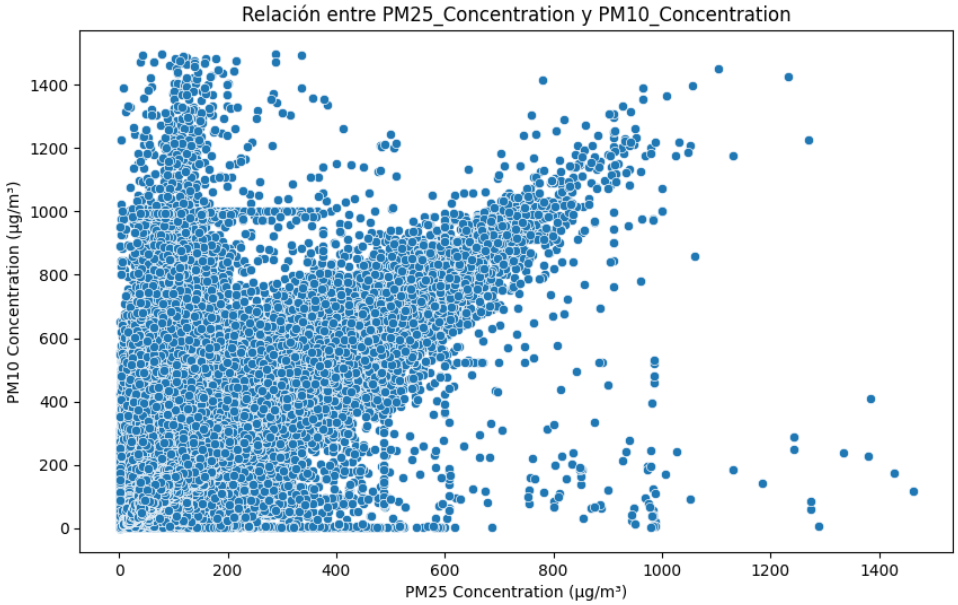

Figura 5: Hipotesis 3.1

La matriz de correlación y el scatterplot confirman que existen datos de calidad del aire que son proporcionales entre sí. La relación más fuerte se observa entre `PM25_Concentration` y `PM10_Concentration`, con una correlación de 0.86, lo que indica que estas dos variables están estrechamente relacionadas, probablemente debido a que provienen de fuentes similares, como la quema de combustibles fósiles o el tráfico.

Otras relaciones significativas, como la correlación de 0.67 entre `PM25_Concentration` y `CO_Concentration`, también apoyan la idea de proporcionalidad entre contaminantes asociados a emisiones vehiculares y procesos industriales.

Contexto La correlación negativa de `O3_Concentration` con otros contaminantes refleja un comportamiento opuesto, consistente con su formación fotoquímica. Esto explica que los niveles de ozono tienden a ser más altos en verano, cuando otros contaminantes pueden estar en niveles más bajos debido a una mayor dispersión y reacciones químicas atmosféricas.

Conclusión La hipótesis se confirma: existe proporcionalidad entre varios contaminantes, especialmente entre `PM25_Concentration` y `PM10_Concentration`, lo que ayuda a comprender mejor las fuentes y dinámicas de la contaminación atmosférica en las ciudades estudiadas.

6. Conclusiones

- El análisis exploratorio del dataset de calidad del aire en 43 ciudades chinas entre mayo de 2014 y abril de 2015 permitió identificar patrones importantes y desafíos en los datos que impactan el desarrollo de modelos predictivos y herramientas de visualización.
- Se confirmó que la calidad del aire varía significativamente según la zona geográfica, siendo las ciudades del norte, como Beijing y Tianjin, las que presentan mayores concentraciones de contaminantes, especialmente en invierno debido a factores climáticos y actividades humanas específicas. Además, se observaron patrones cíclicos claros en la contaminación, con picos estacionales relacionados con condiciones meteorológicas y emisiones industriales. Tambien se encontro una correlación entre los datos de calidad del aire.
- Los datos presentaron problemas relevantes, como un alto porcentaje de valores faltantes y la presencia de datos atípicos o no físicos, lo que subraya la necesidad de técnicas robustas de imputación. Las correlaciones entre contaminantes y variables meteorológicas sugieren que los modelos deben integrar dependencias temporales, espaciales y multivariadas para obtener resultados precisos.
- Finalmente, el desarrollo de un visualizador interactivo para comparar modelos de imputación surge como una herramienta valiosa para mejorar la calidad de los datos y facilitar la toma de decisiones informadas en la gestión ambiental. Este enfoque permite preservar y evaluar los patrones complejos del dataset, contribuyendo a una mejor comprensión y manejo de la contaminación atmosférica en contextos urbanos.

Anexo

- Codigo fuente: https://github.com/AlbertLlica/Dashboard-TCD-AirQuality
- Data Wragling y Pipeline: https://colab.research.google.com/drive/1Yz6ipkvvNIyYcovGWTCuaE
  ?authuser=2scrollTo=eOfzm3gX9Szx

Referencias

[Gong and Ordieres-Meré, 2018] Gong, B. and Ordieres-Meré, J. (2018). Prediction of daily maximum ozone threshold exceedances by preprocessing and ensemble artificial neural networks in the beijing-tianjin-hebei region. *Atmospheric Environment*, 181:158–167.

[He et al., 2017] He, J., Gong, S., Yu, Y., Yu, L., Wu, L., Mao, H., Song, C., Zhao, S., Liu, H., Li, X., and Li, R. (2017). Air pollution characteristics and their relation to meteorological conditions during 2014–2015 in major chinese cities. *Environmental Pollution*, 223:484–496.

[Junninen et al., 2004] Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J., and Kolehmainen, M. (2004). Methods for imputation of missing values in air quality data sets. *Atmospheric Environment*, 38(18):2895–2907.

[Tai et al., 2010] Tai, A. P. K., Mickley, L. J., and Jacob, D. J. (2010). Correlations between fine particulate matter (pm2.5) and meteorological variables in the united states: Implications for the sensitivity of pm2.5 to climate change. *Atmospheric Environment*, 44(32):3976–3984.

[World Health Organization, 2016] World Health Organization (2016). Ambient air pollution: A global assessment of exposure and burden of disease. *World Health Organization Report*. Accessed: June 2025.

[Zheng et al., 2015] Zheng, Y., Yi, X., Li, M., Li, R., Shan, Z., Chang, E., and Li, T. (2015). Forecasting fine-grained air quality based on big data. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2267–2276. ACM.