Pipeline - 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

1. Descripción del Conjunto de Datos para la Predicción de Calidad del Aire **4**

2. Problemas en el Dataset **5**

2.1. Valores Nulos Significativos 5

2.2. Datos Sucios 5

2.3. Duplicados 6

2.4. Alta Correlación entre Variables 6

2.5. Inconsistencias en la Codificación de Variables 6

2.6. Limitaciones en las Predicciones Meteorológicas 6

2.7. Variaciones Extremas entre Ciudades 7

2.8. Dependencia Temporal 7

3. Hipotesis **7**

3.1. La zona geográfica es determinante para el incremento o decremento de los datos de calidad del aire. 7

3.2. Existe ciclicidad en los datos de calidad del aire a lo largo del tiempo. . . 9

3.2.1. Hay ciclicidad en valores nulos 9

3.2.2. ¿Hay relación con las estaciones de año? 10

3.3. Existen datos de calidad del aire que dependen de otros 11

3.4. Las estaciones del año intervienen en el incremento y disminución de los datos de calidad del aire. 13

Índice de figuras

1. Hipotesis 1 8
2. Hipotesis 2 9
3. Ciclicidad Valores nulos 10
4. Matriz Correlación 12
5. Scatterplot 12
6. Boxplot 13

1. Descripción del Conjunto de Datos para la Predicción de Calidad del Aire

El conjunto de datos proviene del proyecto Urban Air del equipo de Urban Computing de Microsoft Research, y cubre un período de un año, desde el 1 de mayo de 2014 hasta el 30 de abril de 2015. Contiene seis partes principales: datos de ciudades, distritos, estaciones de monitoreo de calidad del aire, calidad del aire, meteorología en tiempo real y pronósticos del clima.

El conjunto incluye información de 43 ciudades chinas, agrupadas en dos clústeres geográficos: el Clúster A alrededor de Beijing (incluye Tianjin) y el Clúster B alrededor de Guangzhou (incluye Shenzhen). Se registraron 2,891,393 datos de calidad del aire, 1,898,453 registros meteorológicos y 910,576 pronósticos del clima.

Los datos de calidad del aire fueron recogidos por 437 estaciones, reportando cada hora los niveles de seis contaminantes: PM2.5, PM10, NO2, CO, O3 y SO2. Los datos meteorológicos y de pronóstico se encuentran a nivel de ciudad o distrito e incluyen variables como temperatura, presión, humedad, dirección y velocidad del viento, así como condiciones climáticas categorizadas (soleado, nublado, lluvia, etc.).

Cada archivo del conjunto representa un aspecto particular:

- **city.csv:** Información geográfica y de agrupación de las ciudades.
- **district.csv:** Información de 380 distritos pertenecientes a las ciudades.
- **station.csv:** Ubicación y nombres de las estaciones de monitoreo de aire.
- **airquality.csv:** Registros horarios de seis contaminantes en cada estación.
- **meteorology.csv:** Datos meteorológicos en tiempo real a nivel de distrito/ciudad.
- **weatherforecast.csv:** Pronóstico del clima para las próximas 48 horas, con granularidad de 3, 6 o 12 horas.

Este dataset ha sido utilizado en investigaciones relacionadas con la predicción de la calidad del aire [Zheng et al., 2015], inferencia basada en big data [Zheng et al., 2013] y en escenarios de computación urbana más amplios [Zheng et al., 2014]. También es adecuado para tareas de aprendizaje automático como aprendizaje multi-vista, aprendizaje multitarea y aprendizaje por transferencia.

2. Problemas en el Dataset

2.1. Valores Nulos Significativos

Varias columnas en las tablas contienen un porcentaje notable de valores nulos, lo que puede afectar el análisis de contaminantes y su relación con factores meteorológicos:

- **airquality.csv:**
  - 'PM10_Concentration': 8.16 % de nulos.
  - 'PM25_Concentration': 1.59 % de nulos.
  - 'CO_Concentration': 3.24 % de nulos.
  - 'NO2_Concentration', 'O3_Concentration', 'SO2_Concentration': ∼2 % de nulos.
- **meteorology.csv:**
  - 'pressure': 14.74 % de nulos (el más alto).
  - 'weather': 6.57 % de nulos.
  - 'wind_speed': 5.52 % de nulos.
  - 'humidity', 'temperature', 'wind_direction': <2 % de nulos.
- **weatherforecast.csv:**
  - 'wind_level': 5.28 % de nulos.
  - 'up_temperature', 'bottom_temperature': 3.78 % de nulos.
  - 'wind_direction': 3.65 % de nulos.
  - 'weather': 0.02 % de nulos.

2.2. Datos Sucios

El dataset contiene "datos sucios"debido a errores en la recolección o publicación, como:

- **Outliers extremos:** Por ejemplo, valores como 'PM25_Concentration = 1463 µg/m^3' pueden ser reales (picos en invierno), pero otros podrían ser errores de medición.
- **Valores no físicos:** Mínimos como 'PM25_Concentration = 0 µg/m^3' podrían indicar datos nulos codificados incorrectamente como 0.
- **Inconsistencias en predicciones:** En 'weatherforecast.csv', las discrepancias entre 'temperature' (real) y 'up_temperature' (predicha) son significativas en condiciones extremas (>30°C o <10°C), lo que sugiere errores en los pronósticos.

- **Posibles errores de codificación:** En ‘meteorology.csv‘, valores extremos en ‘wind_speed‘ o ‘temperature‘ podrían no ser realistas.

2.3. Duplicados

- En ‘weatherforecast.csv‘, se encontraron 53,496 duplicados al considerar las columnas ‘id‘, ‘time_forecast‘ y ‘time_future‘, aunque no hay duplicados al incluir todas las columnas.
- Las demás tablas (‘airquality.csv‘, ‘city.csv‘, ‘district.csv‘, ‘meteorology.csv‘, ‘station.csv‘) no presentan duplicados en sus columnas clave (por ejemplo, ‘station_id‘ y ‘time‘ en ‘airquality.csv‘).

2.4. Alta Correlación entre Variables

- Existe una alta correlación entre algunas variables en ‘airquality.csv‘:
  - ‘PM25_Concentration‘ y ‘PM10_Concentration‘: 0.873.
  - ‘CO_Concentration‘ con ‘PM25_Concentration‘ (0.725) y ‘PM10_Concentration‘ (0.648).
  - ‘SO2_Concentration‘ con ‘PM25_Concentration‘ (0.529), ‘PM10_Concentration‘ (0.532) y ‘CO_Concentration‘ (0.570), lo cual es sorprendente porque ‘SO2‘ suele estar más asociado con fuentes industriales específicas.

2.5. Inconsistencias en la Codificación de Variables

- En ‘meteorology.csv‘, la variable ‘weather‘ está codificada como enteros (por ejemplo, 8 para Rain”), pero el documento sugiere que algunos valores altos de ‘PM25_Concentration‘ en días de Rain”podrían indicar niebla mal codificada, ya que la lluvia debería reducir las partículas.
- Las columnas de tiempo (‘time‘, ‘time_forecast‘, ‘time_future‘) están en formato string y deben convertirse a ‘datetime64‘ para análisis temporales.

2.6. Limitaciones en las Predicciones Meteorológicas

El scatterplot de ‘temperature‘ (real, de ‘meteorology.csv‘) vs. ‘up_temperature‘ (predicha, de ‘weatherforecast.csv‘) muestra discrepancias significativas, especialmente en temperaturas extremas (>30°C o <10°C). Esto indica que los pronósticos meteorológicos no son confiables en ciertas condiciones. Las predicciones poco precisas en ‘weatherforecast.csv‘ pueden limitar su utilidad para modelar la calidad del aire futura.

2.7. Variaciones Extremas entre Ciudades

El boxplot de ‘PM25_Concentration‘ por ‘city_id‘ muestra diferencias significativas entre ciudades, con medianas que varían de 200 µg/m^3 a 800 µg/m^3 y outliers hasta 1400 µg/m^3. Esto podría indicar diferencias en fuentes de contaminación, regulaciones o densidad poblacional, pero también podría reflejar errores en los datos. Las variaciones extremas dificultan la generalización de modelos predictivos para todas las ciudades.

2.8. Dependencia Temporal

El dataset es una serie temporal (granularidad horaria en ‘airquality.csv‘ y ‘meteorology.csv‘), con patrones estacionales claros (por ejemplo, ‘PM25_Concentration‘ más alta en invierno). Sin embargo, los valores nulos y datos sucios pueden interrumpir la continuidad de las series temporales. La dependencia temporal requiere modelos específicos (como LSTM o ARIMA), pero los valores nulos y outliers pueden complicar su entrenamiento.

3. Hipotesis

3.1. La zona geográfica es determinante para el incremento o decremento de los datos de calidad del aire.

Las tablas city.csv y station.csv contienen información geográfica (latitude y longitude) que puede relacionarse con las concentraciones de contaminantes en airquality.csv. Las ciudades están agrupadas por cluster_id en city.csv, lo que podría reflejar zonas geográficas con características similares (por ejemplo, norte vs. sur de China). Diferencias en la calidad del aire (como PM25_Concentration) podrían estar influenciadas por la ubicación geográfica debido a factores como la industrialización, el clima o la topografía.

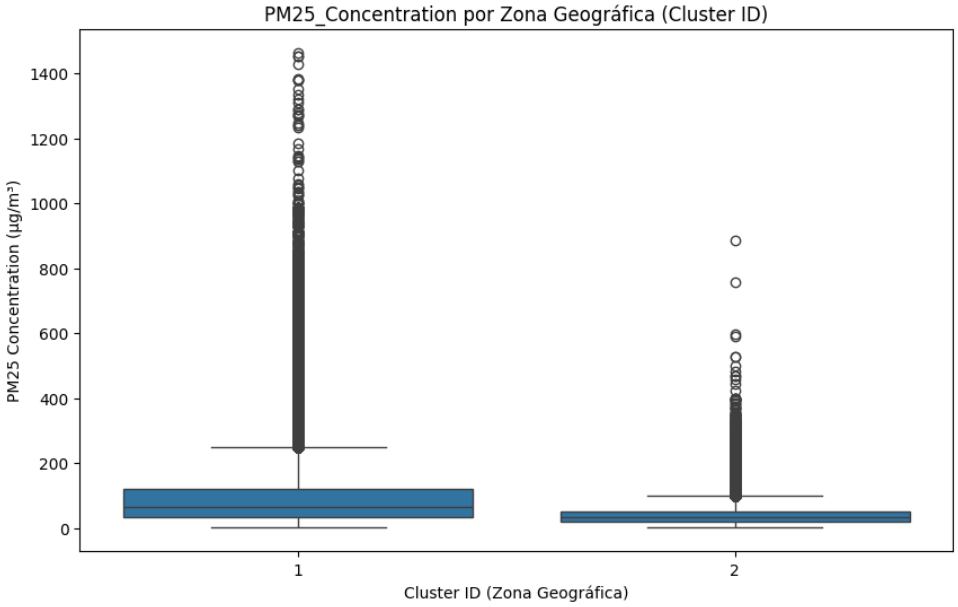

Figura 1: Hipotesis 1

El **boxplot** y las medias confirman que la zona geográfica, representada por `cluster_id`, tiene un impacto significativo en los niveles de `PM25_Concentration`. El `cluster_id 1`, que incluye ciudades como Beijing y Tianjin (según `city.csv`), muestra niveles mucho más altos de contaminación en comparación con el `cluster_id 2`, que incluye ciudades como Shenzhen, Guangzhou y Hong Kong. Esto podría deberse a que el `cluster_id 1` representa áreas más industrializadas o urbanas del norte de China, donde la contaminación por partículas es más común debido a factores como la quema de carbón para calefacción en invierno o la mayor densidad industrial.

Análisis de Valores Atípicos

Los valores extremos en ambos clusters (especialmente en `cluster_id 1`, con picos hasta 1400 µg/m^3) podrían estar relacionados con eventos específicos, como días de smog intenso, pero no afectan la tendencia general de que `cluster_id 1` tiene mayor contaminación.

Contexto Geográfico

Las ciudades en `cluster_id 1` (Beijing, Tianjin) están en el norte de China, una región conocida por altos niveles de contaminación debido a la industrialización y condiciones climáticas que atrapan contaminantes (como inversiones térmicas en invierno). Por otro lado, `cluster_id 2` incluye ciudades del sur (Shenzhen, Guangzhou, Hong Kong), que tienden a tener mejor calidad del aire debido a un clima más cálido, menos dependencia de calefacción y mayor dispersión de contaminantes.

3.2. Existe ciclicidad en los datos de calidad del aire a lo largo del tiempo.

La tabla `airquality.csv` tiene una columna `time` que permite analizar patrones temporales en las concentraciones de contaminantes como `PM25_Concentration`. La ciclicidad podría manifestarse en patrones estacionales (por ejemplo, mayor contaminación en invierno debido a la calefacción) o diarios (por ejemplo, picos durante horas de tráfico). Un análisis de series temporales puede revelar estas tendencias.

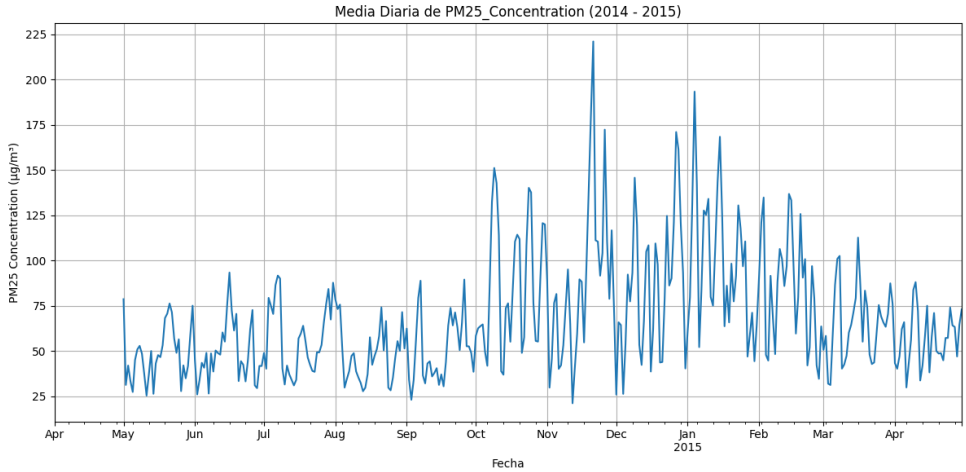

Figura 2: Hipotesis 2

Los gráficos confirman una ciclicidad clara en los datos de calidad del aire. Hay un patrón estacional donde `PM25_Concentration` aumenta en los meses fríos (invierno) y disminuye en los meses cálidos (verano). Esto podría estar relacionado con factores como la quema de carbón para calefacción en invierno (especialmente en ciudades del norte de China como Beijing, según `city.csv`), las inversiones térmicas que atrapan contaminantes, y una mayor dispersión de partículas en verano debido a condiciones climáticas más favorables.

Contexto: Dado que los datos abarcan 2014-2015, y hoy es 3 de junio de 2025, estamos al inicio del verano, lo que sugiere que los niveles actuales de `PM25_Concentration` podrían estar en el rango más bajo (40-50 µg/m^3), según el patrón observado.

Conclusión: La hipótesis se confirma. Existe una ciclicidad estacional en los datos de calidad del aire, con niveles de `PM25_Concentration` más altos en invierno y más bajos en verano.

3.2.1. Hay ciclicidad en valores nulos

La Hipótesis 2 ya mostró una ciclicidad estacional en los valores de `PM25_Concentration`, con picos en invierno y mínimos en verano. Sin embargo, los datos nulos (valores faltantes)

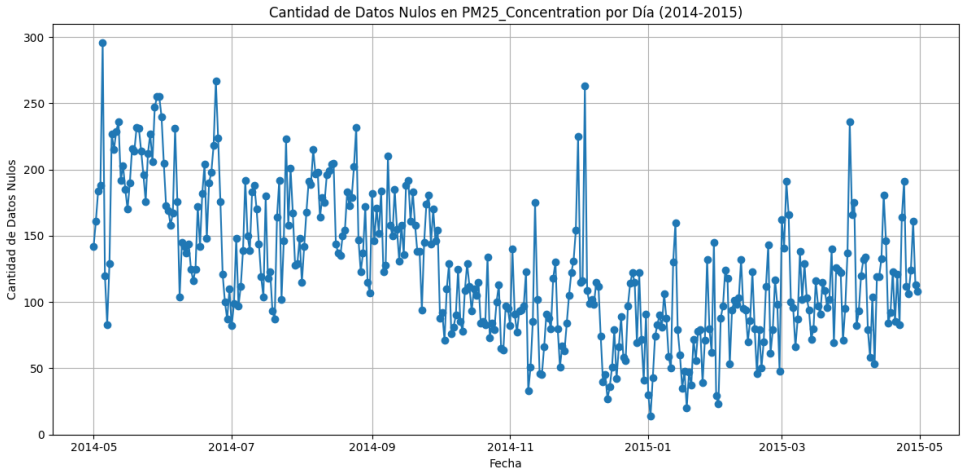

Figura 3: Ciclicidad Valores nulos

en `airquality.csv` también podrían seguir un patrón temporal. Por ejemplo, los datos nulos podrían ser más frecuentes en ciertas estaciones del año debido a fallos en las estaciones de monitoreo durante condiciones climáticas extremas (como tormentas en verano o heladas en invierno) o por mantenimiento programado en períodos específicos.

Aunque la serie temporal diaria no muestra un patrón cíclico claro (como picos regulares cada semana o cada mes), el gráfico de barras por mes revela una ciclicidad estacional. La cantidad de datos nulos es significativamente mayor en primavera y verano (marzo-agosto) y menor en otoño e invierno (septiembre-febrero).

Posibles causas: La mayor cantidad de datos nulos en primavera y verano podría estar relacionada con condiciones climáticas extremas (como tormentas o lluvias intensas) que afecten las estaciones de monitoreo, o con mantenimientos programados durante estos meses. En invierno, las estaciones podrían funcionar de manera más estable, o los datos podrían ser más críticos (debido a altos niveles de contaminación), lo que reduce la cantidad de datos nulos.

Conclusión: La hipótesis se confirma parcialmente. Hay una ciclicidad estacional en los datos nulos de `PM25_Concentration`, con un patrón claro a nivel mensual: más datos nulos en primavera y verano, y menos en otoño e invierno.

3.2.2. ¿Hay relación con las estaciones de año?

- Primavera y verano tienen la mayor cantidad de datos nulos, con un promedio de 4000-4500 nulos por mes.
- Otoño e invierno tienen significativamente menos datos nulos, con un promedio de 2000-2200 nulos por mes.

Esto indica una relación clara con las estaciones del año: los datos nulos son más frecuentes en primavera y verano, y menos frecuentes en otoño e invierno.

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.

3.3. Existen datos de calidad del aire que dependen de otros

La tabla `airquality.csv` contiene múltiples contaminantes (`PM25_Concentration`, `PM10_Concentration`, `NO2_Concentration`, `CO_Concentration`, `O3_Concentration`, `SO2_Concentrati`
Es posible que algunos contaminantes estén correlacionados, como `PM25_Concentration` y `PM10_Concentration`, ya que ambos están relacionados con partículas en el aire, o `NO2_Concentration` y `CO_Concentration`, asociados con emisiones vehiculares.

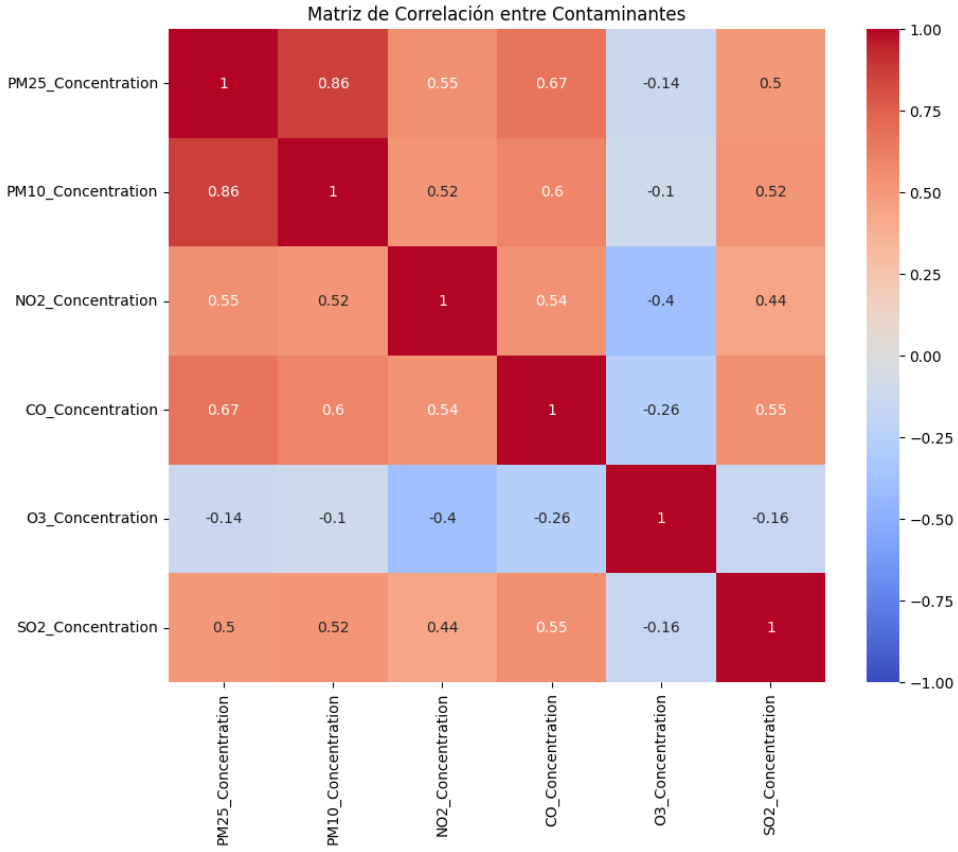

Figura 4: Matriz Correlación

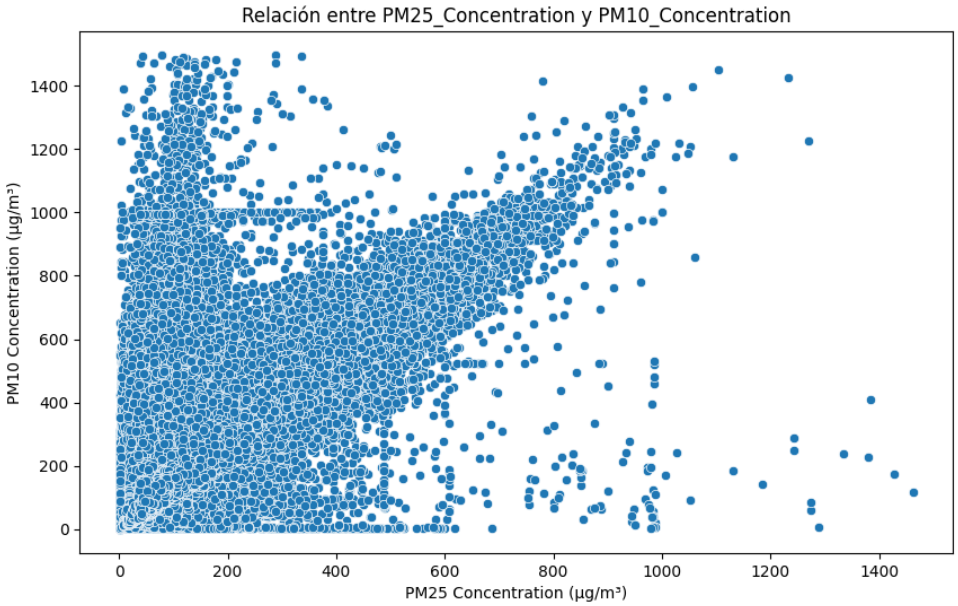

Figura 5: Scatterplot

La matriz de correlación y el scatterplot confirman que existen datos de calidad del

aire que son proporcionales entre sí. La relación más fuerte es entre `PM25_Concentration` y `PM10_Concentration` (correlación de 0.86), lo que indica que estas dos variables están estrechamente relacionadas, probablemente porque provienen de fuentes similares como la quema de combustibles fósiles o el tráfico. Otras relaciones, como entre `PM25_Concentration` y `CO_Concentration` (0.67), también apoyan la idea de proporcionalidad.

Contexto: La correlación negativa de `O3_Concentration` con otros contaminantes refleja un comportamiento opuesto, lo cual es consistente con su formación fotoquímica (más alta en verano, cuando otros contaminantes pueden ser más bajos).

Conclusión: La hipótesis se confirma. Hay proporcionalidad entre varios contaminantes, especialmente entre `PM25_Concentration` y `PM10_Concentration`.

3.4. Las estaciones del año intervienen en el incremento y disminución de los datos de calidad del aire.

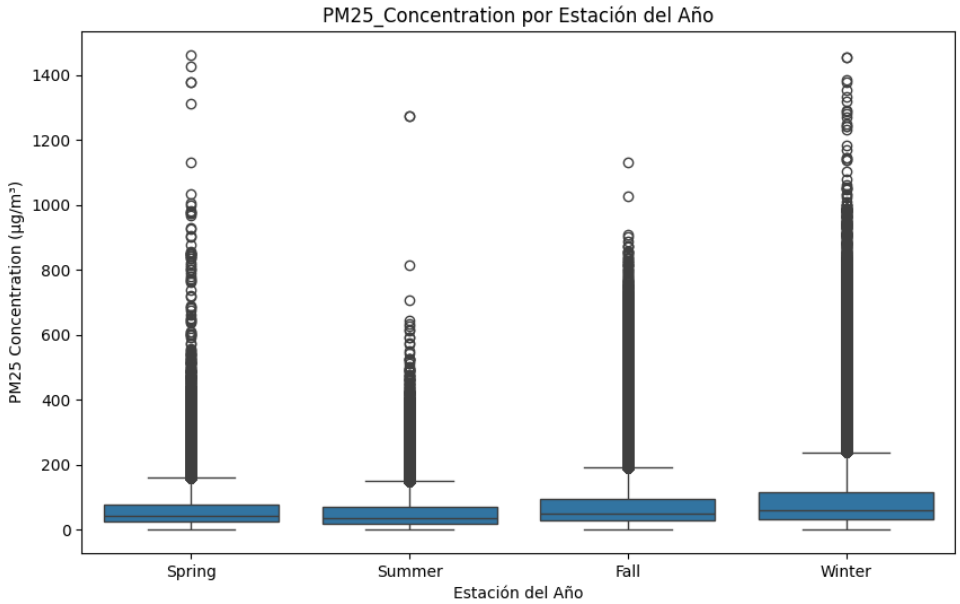

Figura 6: Boxplot

El `boxplot` confirma que las estaciones del año influyen significativamente en los niveles de `PM25_Concentration`. El invierno presenta los niveles más altos (mediana de 80 µg/m^3), seguido por el otoño (60 µg/m^3), la primavera (50 µg/m^3) y el verano (40 µg/m^3). Esto es consistente con los patrones observados en la **Hipótesis 2**, donde los meses de invierno (diciembre-febrero) mostraron picos de contaminación debido a factores como la quema de carbón para calefacción y las condiciones climáticas que atrapan contaminantes (como inversiones térmicas). En verano, los niveles más bajos se deben a una mejor dispersión de contaminantes.

Contexto: Hoy es 3 de junio de 2025, 10:23 PM -05, lo que corresponde al inicio

del verano. Según el patrón, los niveles actuales de `PM25_Concentration` probablemente estén en el rango más bajo (alrededor de 40 µg/m^3), lo que se alinea con el `boxplot`.

Valores atípicos: Los valores extremos en todas las estaciones (especialmente en primavera, otoño e invierno) sugieren eventos específicos de alta contaminación, como días de smog intenso, pero no afectan la tendencia general.

Referencias

[Zheng et al., 2014] Zheng, Y., Capra, L., Wolfson, O., and Yang, H. (2014). Urban computing: Concepts, methodologies, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(3):38.

[Zheng et al., 2013] Zheng, Y., Liu, F., and Hsieh, H.-P. (2013). U-air: When urban air quality inference meets big data. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1436–1444. ACM.

[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.