

13 DE JUNIO DE 2025



ANÁLISIS Y PREDICCIÓN DE LA CALIDAD DEL AIRE EN CIUDADES GLOBALES DURANTE 2024

PROYECTO INTEGRADOR

MAURO TALAMANTES VILLAGRANA

MINERÍA DE DATOS | MAESTRÍA EN CIENCIAS DEL PROCESAMIENTO DE LA INFORMACIÓN
UNIVERSIDAD AUTÓNOMA DE ZACATECAS

Contenido

Comprensión del negocio	3
Objetivo General:	3
Justificación:	3
Problema por resolver:	3
Público objetivo:	3
Comprensión de los datos	3
Adquisición y carga de los datos	4
Estadísticas descriptivas.....	4
Inspección de los valores nulos y primeras observaciones.....	5
Clasificación del índice de calidad del aire (AQI)	6
Correlación entre contaminantes y el Índice de Calidad del Aire (AQI)	6
Resumen Visual: Análisis de Contaminantes por Rango de AQI	7
Resumen visual: Matriz de Correlación entre Contaminantes y AQI	8
Resumen visual: Distribución del AQI por ciudad	9
Promedio de Contaminantes por Ciudad y Nivel AQI	10
Preparación de los datos	14
Limpieza de datos.....	14
Transformación y enriquecimiento temporal.....	14
Análisis Temporal del AQI.....	14
Modelado	22
Regresión Lineal Múltiple	22
Selección de Variables y Preprocesamiento	22
División del Conjunto de Datos.....	22
Entrenamiento del Modelo	22
Evaluación del modelo	22
Random Forest	22
Selección de Variables y Preprocesamiento	22
División del Conjunto de Datos.....	23
Entrenamiento del Modelo	23
Evaluación del modelo	23
XGBoost Regressor	23
Selección de Variables y Preprocesamiento	23

División del Conjunto de Datos.....	24
Entrenamiento del modelo.....	24
Evaluación del modelo	24
Evaluación de los resultados	24
Evaluación específica por ciudad	28
Análisis de outliers en el promedio diario del AQI	29
Despliegue o Aplicación Práctica	30
Monitoreo inteligente del aire en tiempo real.....	30
Optimización de políticas públicas ambientales	30
Modelos personalizados por ciudad	30
Identificación de eventos extremos	31
Conclusiones	31

Comprensión del negocio

Objetivo General:

El objetivo de este proyecto es analizar la calidad del aire en seis ciudades del mundo durante el año 2024, con el fin de identificar patrones de contaminación, entender la relación entre diferentes contaminantes atmosféricos y predecir el índice de calidad del aire (AQI) a lo largo del tiempo.

Justificación:

La contaminación del aire es una de las principales amenazas para la salud pública, causando aproximadamente 7 millones de muertes prematuras cada año, según la OMS. A través del análisis de datos ambientales, es posible anticipar condiciones de riesgo, apoyar decisiones gubernamentales, y fomentar acciones preventivas en la población.

Problema por resolver:

- ¿Cómo varía la calidad del aire entre distintas ciudades del mundo durante el año 2024?
- ¿Qué contaminantes influyen más en el índice AQI?
- ¿Es posible predecir la calidad del aire usando modelos de aprendizaje automático?

Público objetivo:

- Instituciones gubernamentales de salud y medio ambiente.
 - Investigadores y científicos ambientales.
 - Organizaciones internacionales como la OMS y Greenpeace.
 - Ciudadanos interesados en la salud ambiental.
-

Comprensión de los datos

El conjunto de datos Global Air Quality (2024) – 6 Cities contiene más de 52,000 registros de mediciones diarias de contaminantes atmosféricos tomadas entre enero y diciembre de 2024. Los registros incluyen datos de seis ciudades ubicadas en diferentes continentes, lo cual permite realizar un análisis geográficamente diverso y comparativo.

Las variables disponibles en el dataset principal son las siguientes:

- Date: Marca temporal de la medición (formato datetime).
- City: Ciudad donde se realizó la medición.
- CO: Monóxido de carbono.
- CO2: Dióxido de carbono.
- NO2: Dióxido de nitrógeno.
- SO2: Dióxido de azufre.

- O3: Ozono.
- PM2.5: Partículas finas.
- PM10: Partículas gruesas.
- AQI: Índice de calidad del aire (según estándar europeo).

Adquisición y carga de los datos

Para asegurar la reproducibilidad del análisis, el dataset fue descargado directamente desde **Kaggle**, utilizando credenciales autenticadas a través del archivo `kaggle.json`. Una vez descomprimido el archivo, se procedió a cargar el archivo principal **Air_Quality.csv** utilizando `pandas`, especificando que la columna `Date` debía ser interpretada como tipo `datetime` para facilitar el análisis temporal.

Una vez cargado el dataset, se procedió a examinar su estructura general mediante `df.info()` para entender su dimensionalidad, tipos de datos y presencia de valores faltantes. El resumen de la data set nos mostró que:

- **Total de registros:** 52,704.
- **Variables:** 10.
- **Columna temporal:** `Date` (tipo `datetime` con zona horaria UTC)
- **Columna categórica:** `City`.
- **8 columnas numéricas continuas:** contaminantes y el índice AQI.

Además, nos permitió identificar que la columna **CO2** contiene **solo 9,648 registros válidos**, lo que representa tan solo el **18.3% del total**. Las demás columnas no presentan valores faltantes.

Estadísticas descriptivas

Para obtener una visión general del comportamiento de las variables numéricas, se utilizó el método `describe()`, el cual permite observar medidas de tendencia central, dispersión y valores extremos.

index	CO	CO2	NO2	SO2	O3	PM2.5	PM10	AQI
count	52704	9648	52704	52704	52704	52704	52704	52704
mean	258.26	462.35	24.10	12.57	60.03	17.69	35.64	41.35
std	159.59	33.77	19.36	17.27	38.22	15.67	48.44	26.63
min	52	434	0	0	0	0.1	0.1	4.45
25%	159	445	9.7	2.3	35	6.9	10.2	22.8
50%	213	453	18.9	5.7	54	12.5	18.9	31.27
75%	306	467	33.4	16.8	78	23	37.5	57.70
max	2045	884	165.9	239.7	349	129.5	543.9	196.63

En la tabla podemos observar la presencia de valores extremos con máximos considerablemente alejados de la media, como puede ser en CO, SO2 Y PM10.

Las variables **NO₂**, **PM2.5** y **PM10** presentan asimetrías, dado que su media está más cerca del mínimo que del máximo.

CO2 tiene una menor dispersión, lo que sugiere una variabilidad reducida respecto al resto, pero esto se da que en su mayoría los datos son nulos.

El índice de calidad del aire AQI muestra una media de 41.35, lo cual en general indica niveles dentro del rango “aceptable”, aunque el valor máximo refleja casos de alta contaminación.

Inspección de los valores nulos y primeras observaciones

Date	0
City	0
CO	0
CO2	43056
NO2	0
SO2	0
O3	0
PM2.5	0
PM10	0
AQI	0

Para identificar posibles problemas de integridad en los datos, se utilizó el método **isnull().sum()** que permite contar los valores faltantes por columna, identificando que únicamente tenemos en CO2 el total de 43,056 valores faltantes (81.7%). En este caso se evaluo que lo mejor seria simplemente eliminar toda la columna.

Además, se revisaron las primeras filas del conjunto de datos usando **df.head()**, donde podemos apreciar que tenemos lecturas continuas cada hora, lo que indica que el dataset tiene una frecuencia horaria, útil para análisis temporales y detección de patrones diarios, semanales o mensuales.

index	Date	City	CO	CO2	NO2	SO2	O3	PM2.5	PM10	AQI
0	2024-01-01 00:00:00+00:00	Brasilia	323	NaN	23.8	2.8	42	12	17.1	16.8
1	2024-01-01 01:00:00+00:00	Brasilia	318	NaN	21.9	2.7	40	12.5	17.9	16
2	2024-01-01 02:00:00+00:00	Brasilia	309	NaN	19.2	2.6	39	12.1	17.3	15.599999
3	2024-01-01 03:00:00+00:00	Brasilia	295	NaN	16.3	2.4	38	11.4	16.2	15.2

2024-01-01										
4	04:00:00+00:00	Brasilia	270	NaN	13	2.1	40	10.2	14.6	16

Clasificación del índice de calidad del aire (AQI)

Se generó una nueva variable categórica que clasifica los valores del índice AQI en niveles de calidad del aire, basados en los rangos establecidos por agencias internacionales, que se clasifica en Bueno, Moderado, Insalubre (Grupos sensibles), Insalubre y Muy insalubre o peligroso.

Luego se calcularon los promedios de contaminantes asociados a cada categoría de AQI, lo cual permite comprender como se relacionan los niveles de contaminación con la calidad del aire.

	CO	NO2	SO2	O3	PM2.5	PM10
AQI_Nivel						
Bueno	205.75	19.15	7.27	51.19	10.69	16.18
Moderado	378.70	36.11	26.29	77.69	31.40	64.59
Insalubre (Grupos sensibles)	452.31	38.36	22.70	109.58	55.66	189.30
Insalubre	384.82	33.84	18.24	83.44	64.54	281.25
Muy insalubre	NaN	NaN	NaN	NaN	NaN	NaN
Peligroso	NaN	NaN	NaN	NaN	NaN	NaN

Como observamos en la tabla todos los contaminantes aumentan a medida que el AQI se vuelve más crítico.

PM2.5 y PM10 muestran una relación especialmente fuerte con los niveles “Insalubre” e “Insalubre para grupos sensibles”, lo cual respalda su papel como principales contribuyentes a un aire de mala calidad.

En niveles “Buenos”, los contaminantes se encuentran en concentraciones considerablemente más bajas.

Esta clasificación no solo permite segmentar el análisis, sino que será útil en etapas posteriores como visualizaciones y modelado.

Correlación entre contaminantes y el Índice de Calidad del Aire (AQI)

Con el objetivo de esclarecer porque en la tabla anterior algunos valores eran altos, pero no considerados graves para el AQI se identificaron cuales contaminantes influyen mas en el valor del AQI, se calcularon las correlaciones lineales entre cada uno de ellos y el AQI.

Con los resultados podemos apreciar que las partículas PM10 y PM2.5 tienen la correlación mas alta con el AQI, lo cual confirma que son los principales impulsores del índice.

El monóxido de carbono (CO) también presenta una correlación moderada indicando una contribución importante.

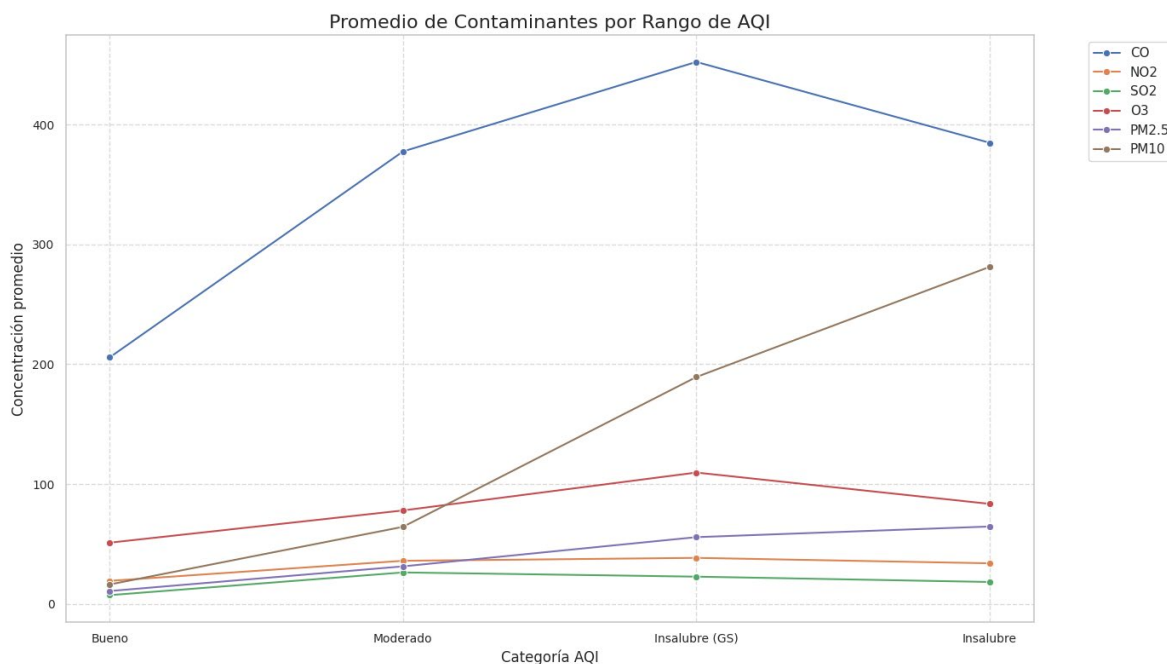
Los demás aunque relevantes, presentan correlaciones mas bajas

Correlación con AQI:	
AQI	1.000000
PM10	0.845944
PM2.5	0.822101
CO	0.537100
O3	0.452599
SO2	0.451506
NO2	0.403690

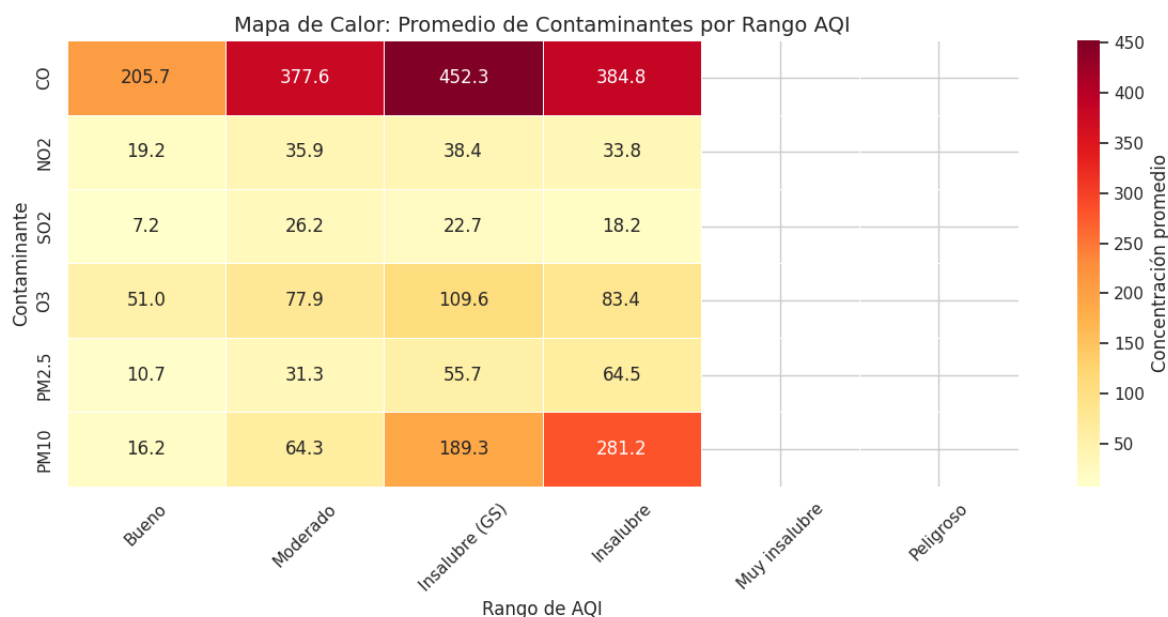
Resumen Visual: Análisis de Contaminantes por Rango de AQI

Los 2 siguientes gráficos representan la misma información que obtuvimos anteriormente sobre el rango de AQI para cada contaminante, con la intención de mostrar de forma visual su impacto y como suelen aumentar cuando empeora la contaminación del aire.

En este grafico podemos ver como CO, NO2, SO2, tienden a aumentar en los rangos Moderado e Insalubre, pero curiosamente bajan en el rango “Insalubre”, mientras que el O3 se incrementa de forma constante conforme se agrava la calidad del aire, igual que PM2.5 Y PM10 que muestran un aumento sostenido y muy marcado, especialmente en el rango de “Insalubre”, lo cual confirma la fuerte correlación con los niveles de AQI.

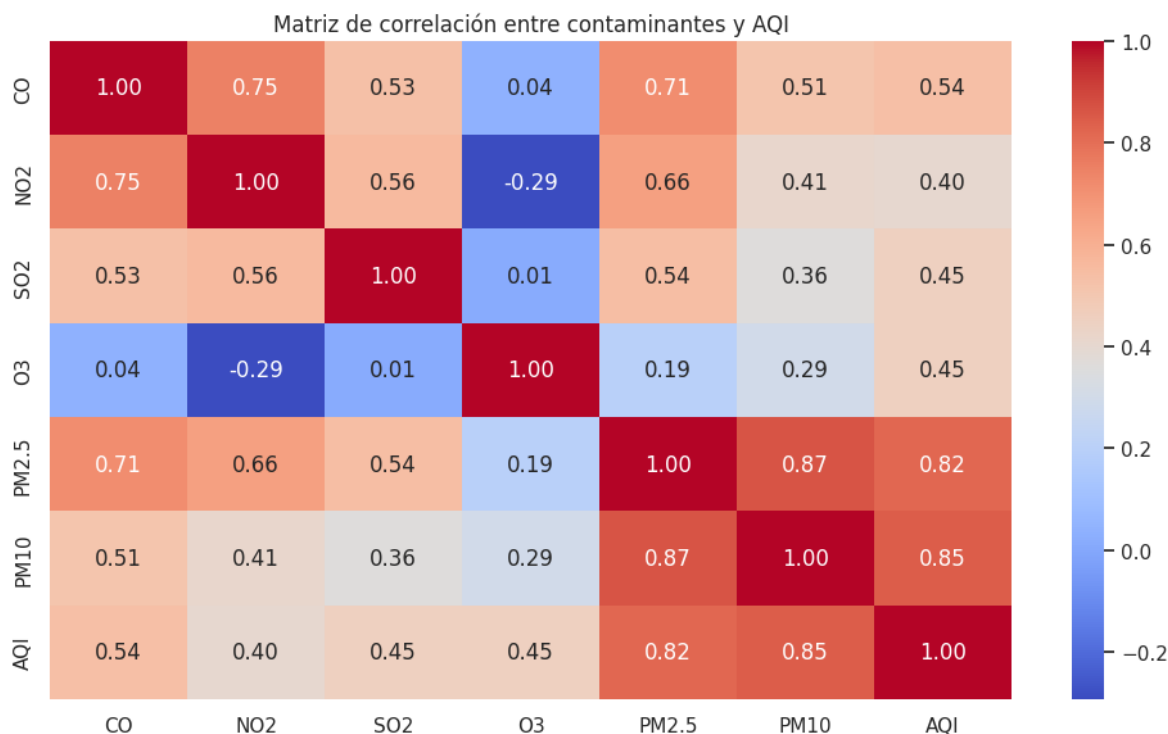


El heatmap nos ayuda a visualizar de manera más clara las diferencias de concentración entre contaminantes y rangos de AQI. Además, nos ayuda a concluir que el comportamiento no lineal de CO y SO2 podría deberse a condiciones locales o patrones de emisión específicos de cada ciudad.



Resumen visual: Matriz de Correlación entre Contaminantes y AQI

La siguiente matriz de calor muestra las correlaciones lineales (coeficiente de Pearson) entre el índice AQI y cada contaminante, así como entre los contaminantes entre sí.

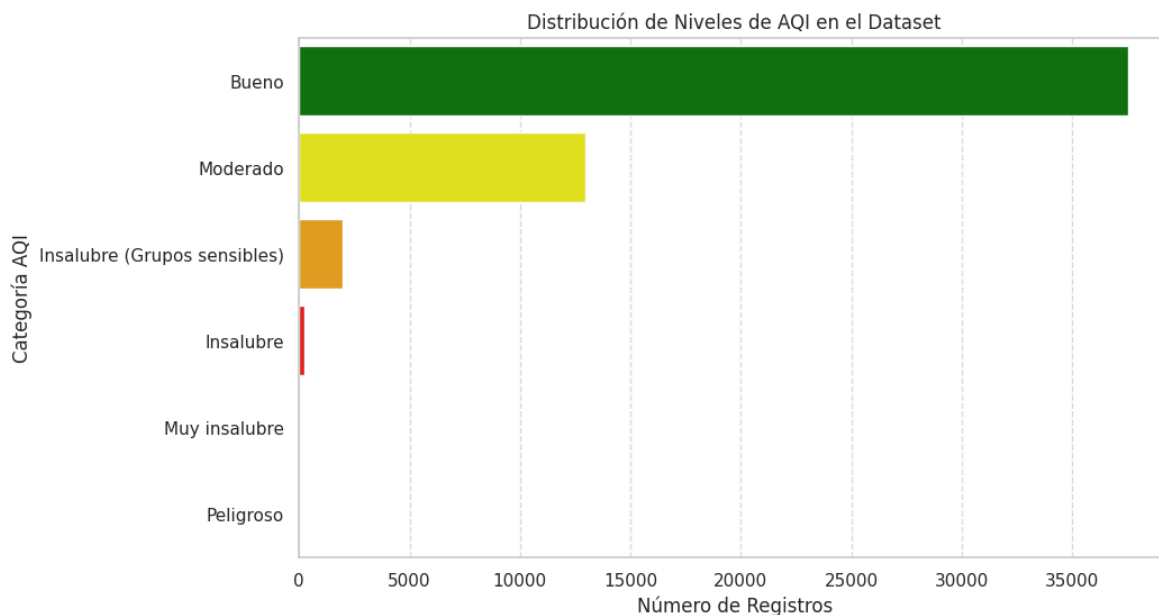


Las correlaciones contra el AQI se examinaron un poco anteriormente, aquí algo que podemos observar como extra es que hay fuerte correlación entre partículas grandes y finas, indicando que entre más se tenga de una, más habrá de la otra.

También observamos que PM2.5 tiene una correlación importante con SO2, NO2 y CO. Estas relaciones entre contaminantes sugieren agrupamientos típicos de fuentes, lo cual puede ser útil en estrategias de mitigación ambiental.

Resumen visual: Distribución del AQI por ciudad

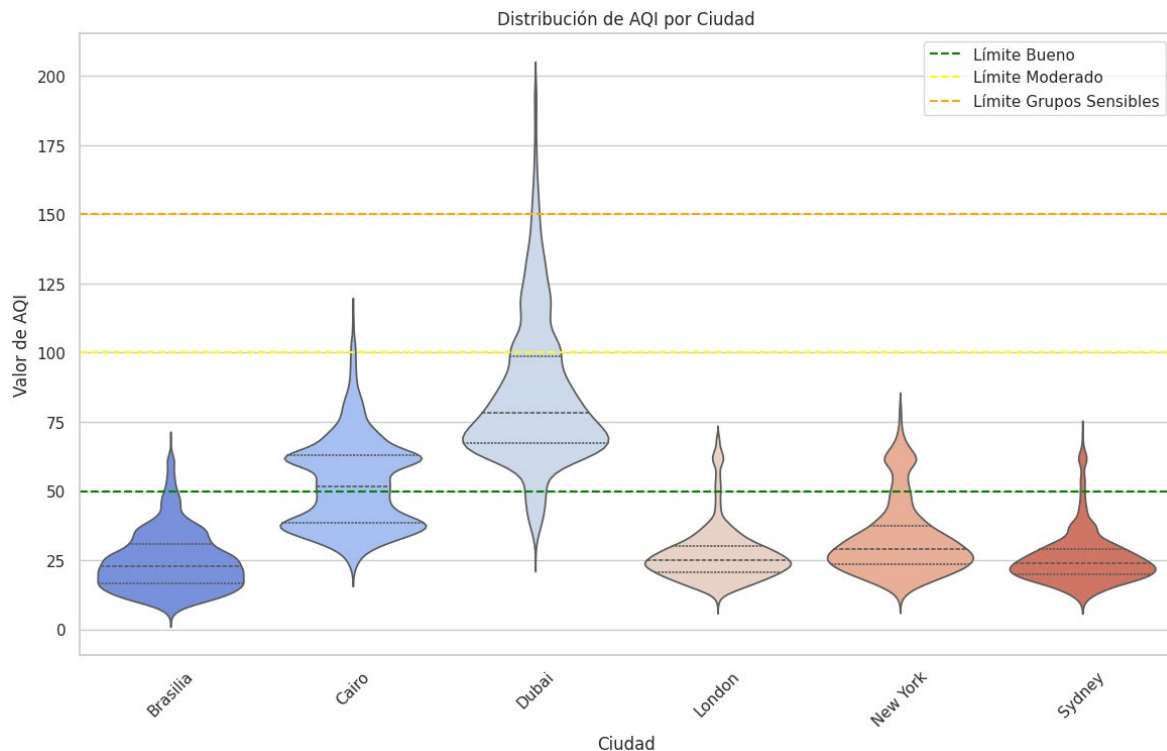
Este primer grafico nos muestra la cantidad de registros por cada nivel de calidad del aire que tenemos en el data set, como podemos observar en su mayoría los registros están en la categoría de bueno.



Este grafico nos sugiere que en general las ciudades analizadas mantienen una calidad del aire aceptable en la mayoría del tiempo, aunque existen eventos de contaminación severa.

El siguiente grafico es la densidad con distribución visual, en el podemos visualizar como:

- Brasilia, Londres, Nueva York y Sídney muestran distribuciones donde la mayoría de los valores se mantienen dentro del rango “Bueno”, con ligeros picos que sobrepasan la categoría.
- El Cairo presenta una distribución mas extendida, cruzando el rango de “Moderado” indicando mayor variabilidad y episodios de peor calidad del aire.
- Dubái, por otro lado, sobrepasa de forma mas clara el limite de Insalubre para “Grupos Sensibles”.

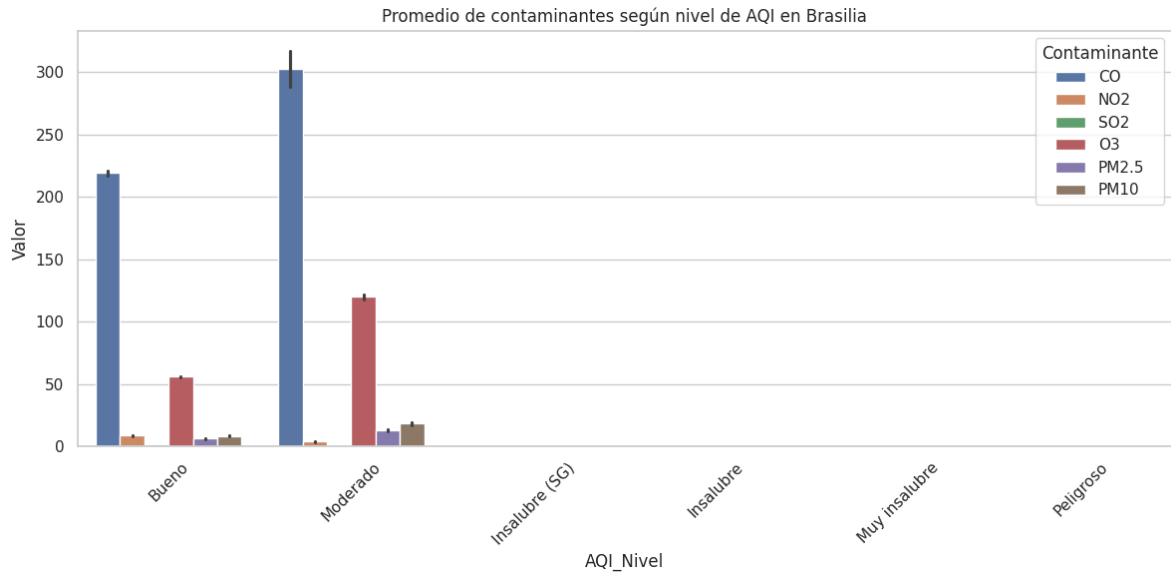


Promedio de Contaminantes por Ciudad y Nivel AQI

Este análisis permite observar cómo varía la concentración de contaminantes según el nivel de calidad del aire en cada ciudad. Se destacaron principalmente las categorías “Bueno” y “Moderado”, ya que las categorías mas críticas no presentaron suficientes registros para calcular promedios en la mayoría de las ciudades.

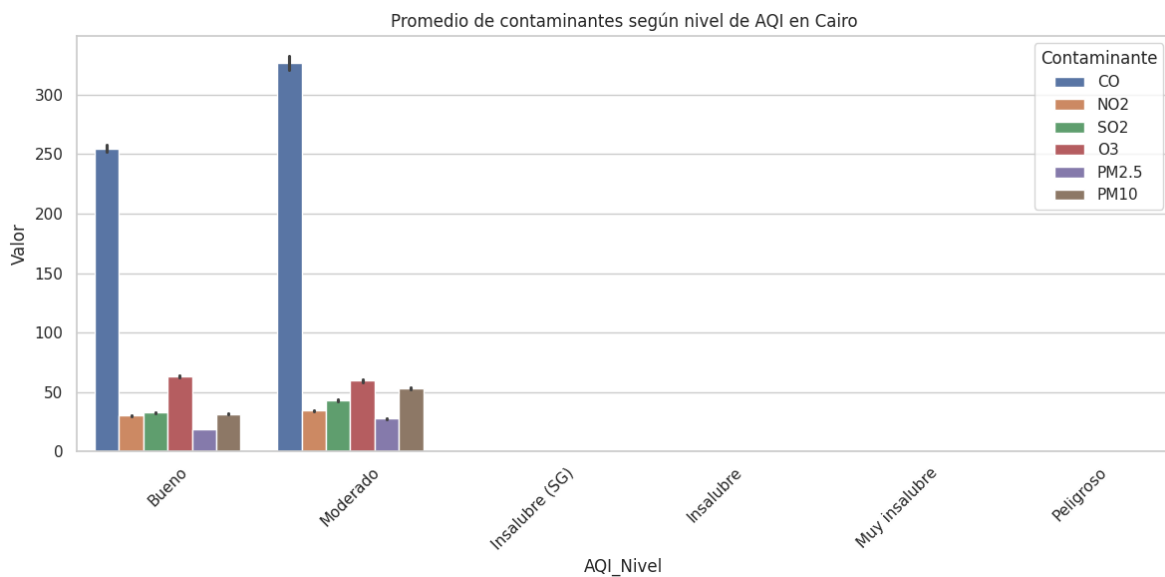
Brasilia:

- En condiciones de aire “Bueno”, Brasilia presenta bajos niveles de contaminantes, destacando un promedio de CO en 219.10 ppb, y partículas PM2.5 en solo $6.36 \mu\text{g}/\text{m}^3$.
- Al pasar a la categoría "Moderado", se observa un incremento notable de ozono (O_3) a 120.17 ppb, mientras que PM2.5 se duplica hasta $13.09 \mu\text{g}/\text{m}^3$.
- No se registraron suficientes datos en categorías más críticas para esta ciudad.



El Cairo:

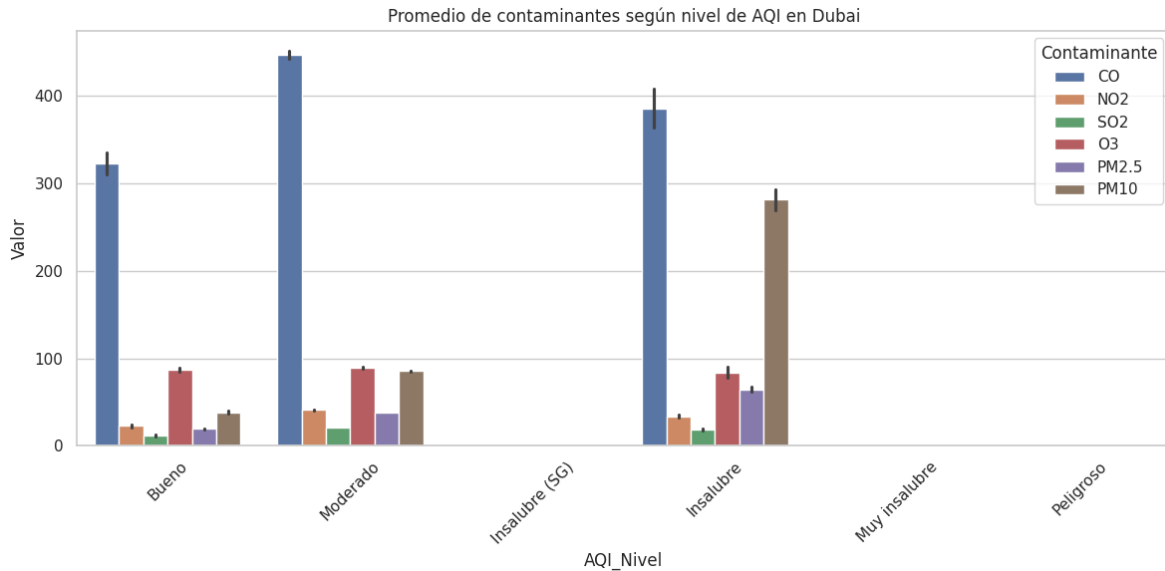
- Incluso en condiciones "Buenas", El Cairo muestra niveles elevados de contaminantes: SO₂ en 32.64 ppb, y PM10 en 31.44 µg/m³.
- En categoría "Moderado", todos los contaminantes se incrementan: el NO₂ sube a 34.43 ppb, el SO₂ a 42.94 ppb, y las PM10 alcanzan 53.20 µg/m³.
- Esta ciudad parece tener una base contaminante más alta, incluso en los mejores escenarios de AQI.



Dubái:

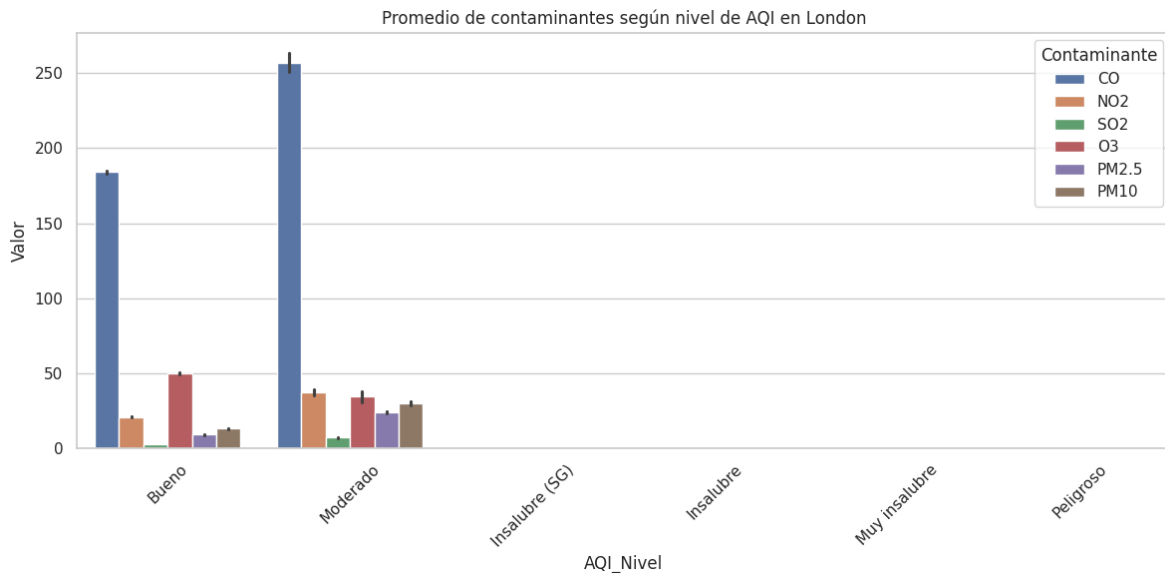
- En niveles "Buenos", Dubái ya exhibe valores elevados de CO (322.42 ppb) y PM10 (38.51 µg/m³).

- Al llegar a "Moderado", se observa un aumento generalizado, con PM2.5 en 37.86 $\mu\text{g}/\text{m}^3$, y NO₂ en 41.16 ppb.
- Interesantemente, Dubái es la única ciudad con registros en el nivel "Insalubre", donde PM10 alcanza un alarmante promedio de 281.25 $\mu\text{g}/\text{m}^3$, muy por encima del límite considerado saludable.



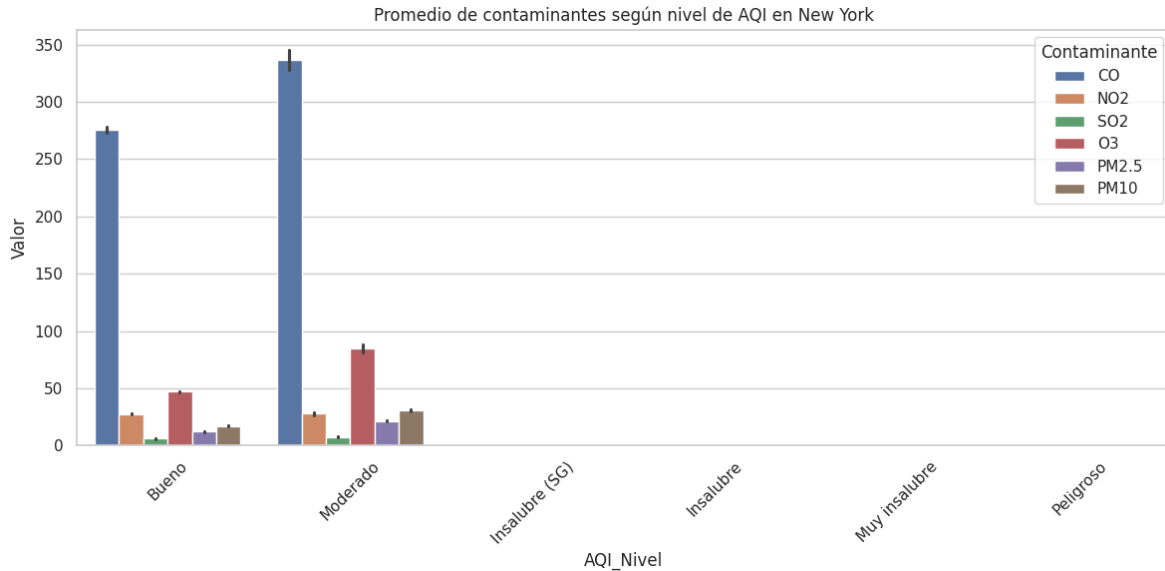
Londres:

- Durante periodos de buena calidad del aire, Londres mantiene niveles relativamente bajos, con PM2.5 en 9.25 $\mu\text{g}/\text{m}^3$ y SO₂ en solo 3.12 ppb.
- En la categoría "Moderado", los valores de NO₂ suben considerablemente a 37.40 ppb, y las partículas finas y gruesas también aumentan, aunque sin llegar a niveles críticos.



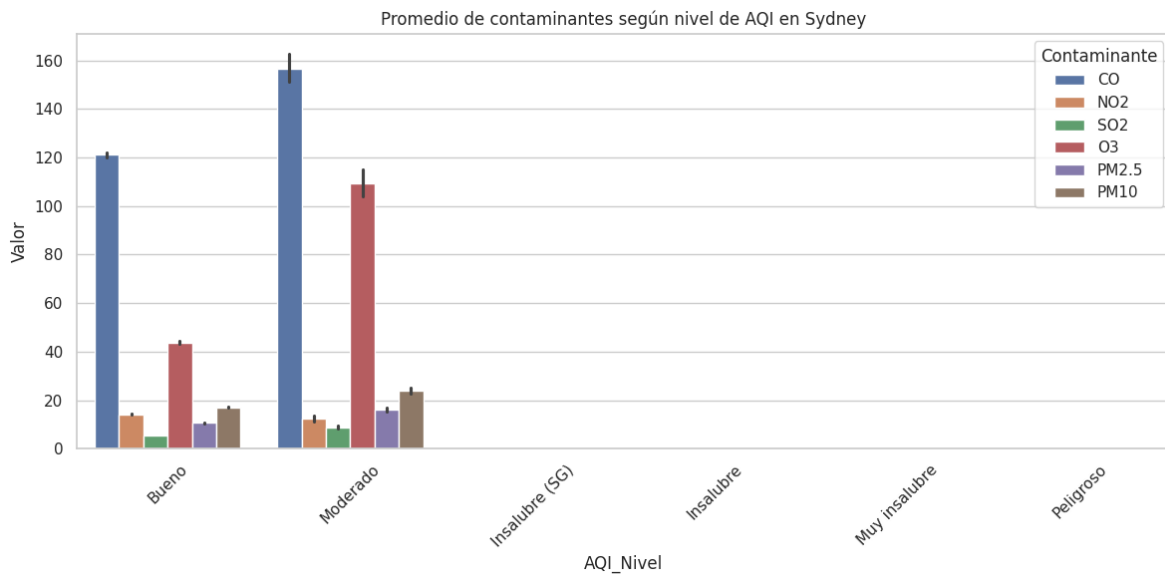
Nueva York:

- En condición "Buena", muestra niveles de ozono estables (47.13 ppb) y partículas en rangos saludables.
- En "Moderado", se observa un crecimiento del CO (336.39 ppb) y ozono (85.14 ppb), aunque sin cambios dramáticos en los niveles de azufre o dióxido de nitrógeno.



Sídney:

- Esta ciudad destaca por sus niveles particularmente bajos de CO (121.17 ppb) y PM10 en categoría "Buena".
- En categoría "Moderado", el ozono muestra un salto a 109.33 ppb, y PM2.5 se incrementa ligeramente a 16.30 $\mu\text{g}/\text{m}^3$, aunque sin alcanzar niveles preocupantes.



En este análisis podemos ver como el rango y niveles de AQI dependen mucho de la ciudad y obviamente de sus condiciones locales (datos que no tenemos aquí expuestos).

Dubái y El Cairo son las ciudades que presentan mayores niveles de contaminantes, incluso en condiciones aceptables de calidad del aire mientras que Sídney y Londres mantienen perfiles más limpios y estables.

La mayoría de las ciudades no acumulan registros suficientes en categorías críticas, lo que sugiere que los eventos extremos de contaminación son poco frecuentes (o posiblemente subreportados).

Preparación de los datos

Limpieza de datos

Para garantizar la calidad y consistencia del conjunto de datos antes de continuar con modelado, se realizó una limpieza preliminar enfocada a los valores faltantes que identificamos anteriormente.

Se tomó la decisión de eliminarla completamente para evitar sesgos o pérdidas excesivas de información al imputar dichos valores.

Posteriormente y por seguridad se eliminaron las filas con valores nulos en cualquier otro lugar (que al momento de usar el dataset ninguna otra tenía valores nulos).

Transformación y enriquecimiento temporal

Para facilitar análisis posteriores relacionados con patrones temporales se convirtió la columna Date al tipo datetime, asegurando que las fechas estuvieran correctamente interpretadas.

Basándose en esa columna, se crearon nuevas variables temporales que permiten analizar el comportamiento de los contaminantes y el AQI a distintos niveles de granularidad:

- Año (year)
- Mes (month)
- Día (day)
- Hora (hour)
- Día de la semana (weekday), con valores del 0 al 6 (lunes a domingo)
- Semana del año (week), siguiendo la norma ISO

Adicionalmente, se clasificaron las fechas en estaciones del año mediante una función que asigna la estación climática correspondiente según el mes, generando una nueva variable **seson** con categorías: Winter, Spring, Summer y Autumn.

Análisis Temporal del AQI

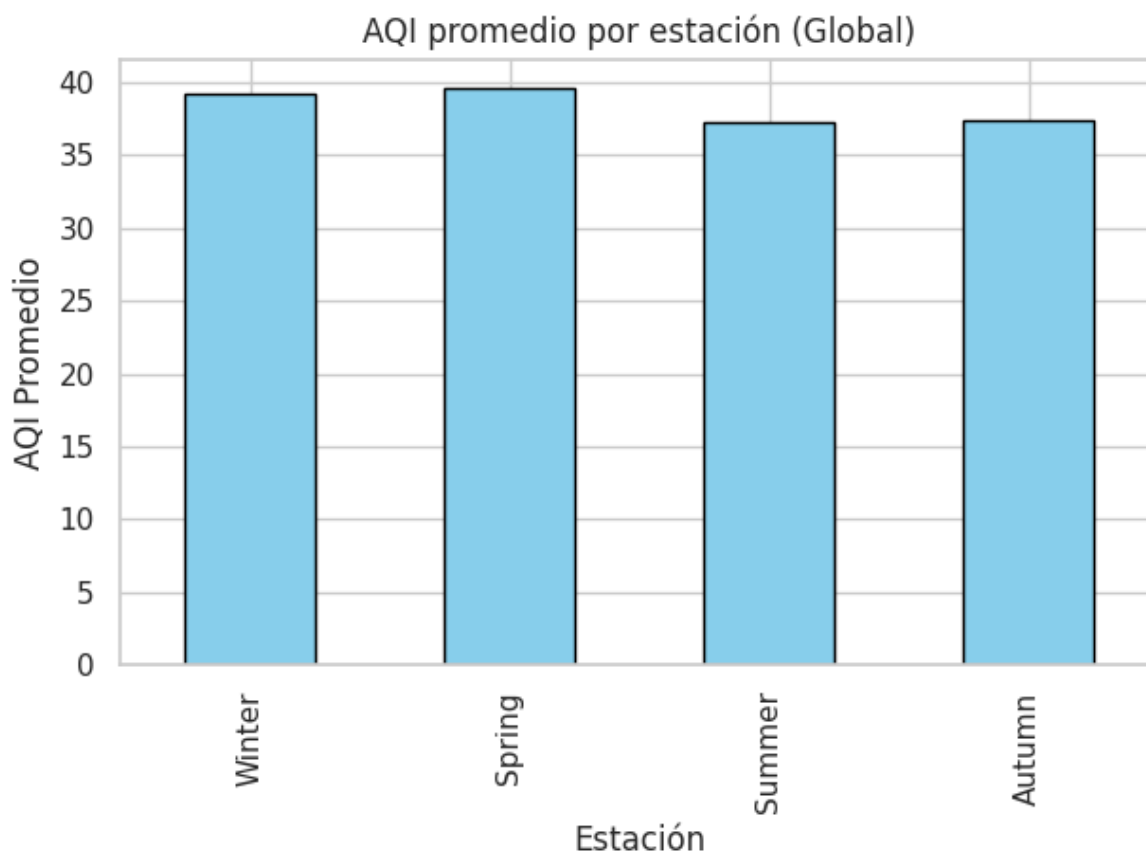
Para explorar posibles patrones estacionales y mensuales en la calidad del aire, se realizó un análisis de correlación y visualización entre el índice AQI y las variables temporales.

Correlación entre Mes y AQI

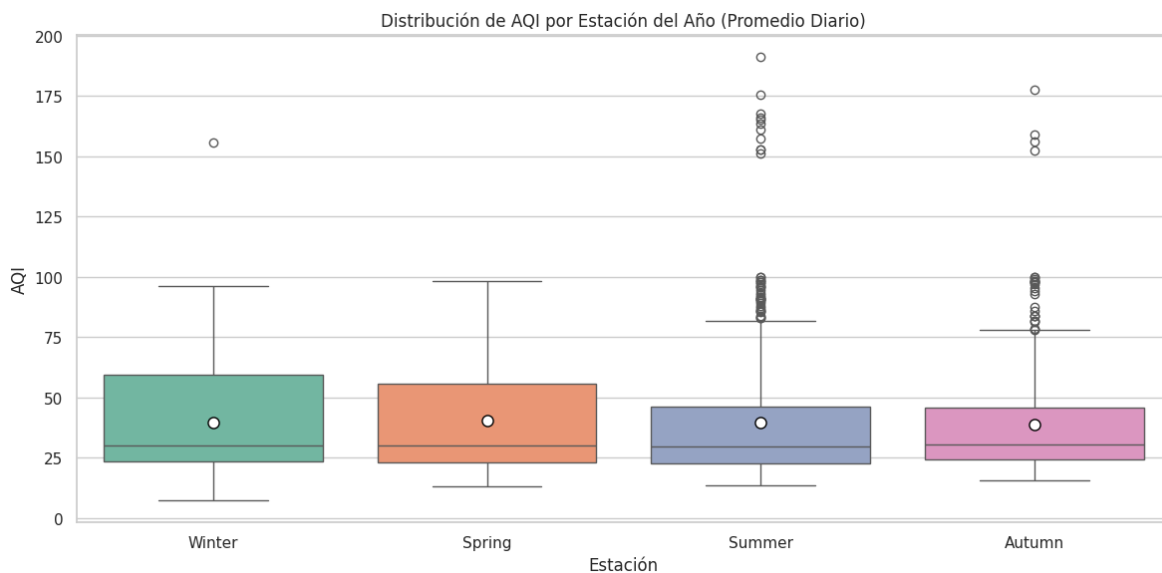
La correlación global entre el mes del año y el AQI resultó ser prácticamente nula con un coeficiente de correlación de Pearson de -0.01 y un valor p de 0.1294. Esto indica que, en términos generales, no existe una relación lineal significativa entre el mes y el AQI.

AQI Promedio por Estación

El análisis por estación mostró que el AQI promedio se mantiene relativamente constante a lo largo del año, con un leve descenso durante el verano y un ligero aumento en primavera.



Haciendo un boxplot obtuvimos que a pesar de mantener un AQI promedio en verano y otoño, también existen muchos valores atípicos que pueden indicar una pero contaminación en estos periodos.

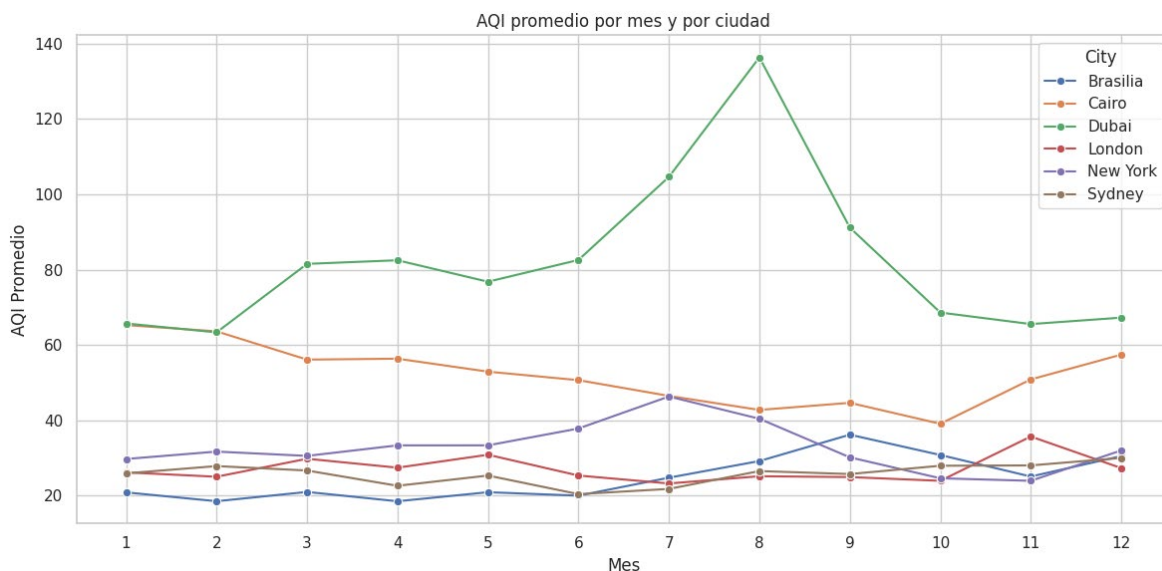


AQI promedio por mes y por ciudad

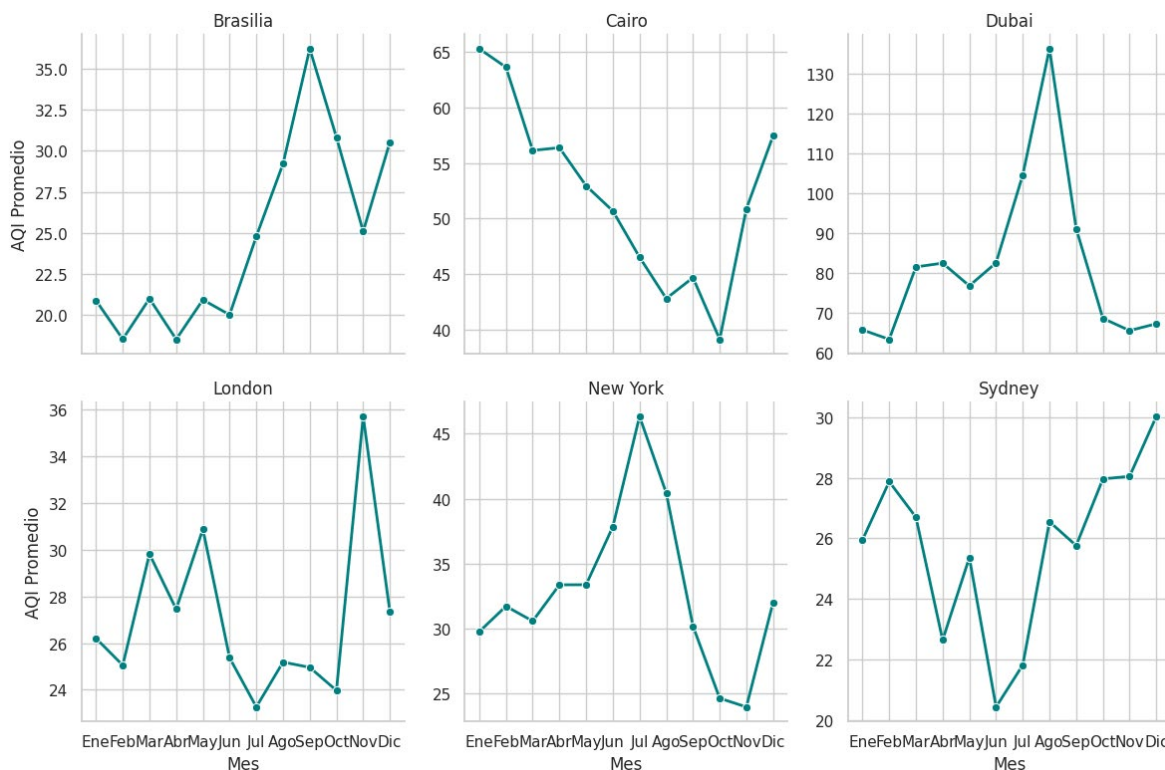
Al analizar la serie temporal mensual por ciudad, se observaron diferencias importantes. Por ejemplo, Dubái presenta niveles consistentemente elevados de AQI alcanzando casi 140 en agosto.

El Cairo por el contrario presenta niveles más altos entre noviembre y febrero.

Y los demás países se mantienen relativamente consistentes a lo largo del año.



Visto de forma individual y para poder tener una mejor visualización de cada uno tenemos el siguiente grafico:

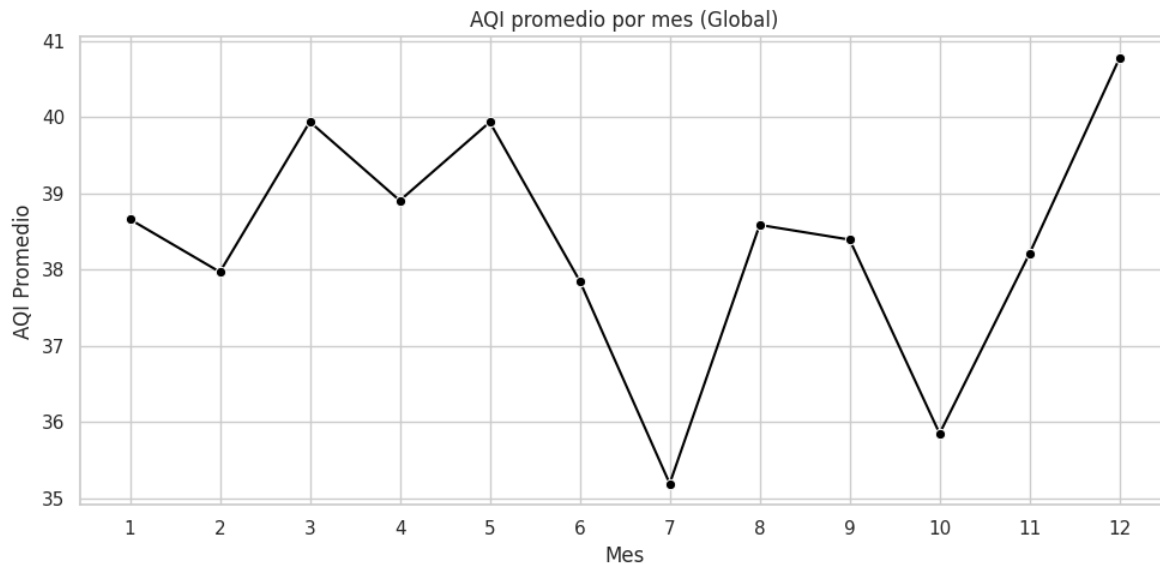


En este grafico podemos ver como cada ciudad si sigue una tendencia en ciertos meses donde su calidad de aire mejora o empeora.

En Brasilia podríamos decir que existe un deterioro paulatino de la calidad del aire, en el Cairo se podría asumir que su curva refleja los efectos estacionales del clima desértico y la actividad humana variable. Dubái muestra un patrón alarmante y marcado en términos de contaminación atmosférica. En Londres podríamos suponer que hay mayor actividad urbana en otoño, pero no tiene picos significativos. En New York cuando entra en moderate podríamos asumir que son eventos de calor urbano y tráfico vehicular.

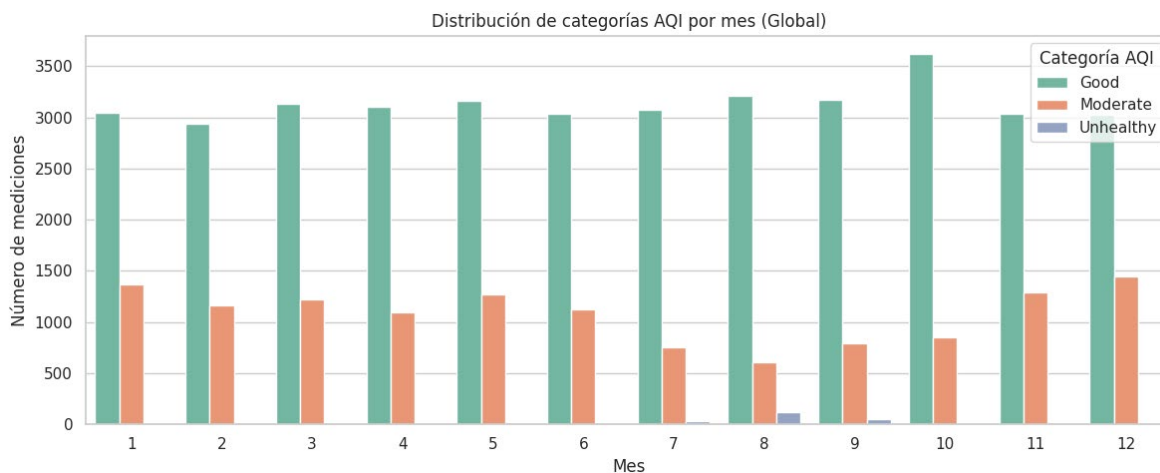
Evolución Global Mensual del AQI

La gráfica del AQI promedio mensual a nivel global evidenció que el nivel más bajo ocurre alrededor de julio, con un valor cercano a 35, mientras que diciembre presenta el valor mas alto, cercano a 41.



Distribución de Categorías AQI por mes

Finalmente, la distribución de categorías AQI por mes a nivel global mostró que los niveles “Bueno” y “Moderado” dominan las mediciones, con algunos valores de “Insalubre” apareciendo mínimamente en los meses de julio, agosto y septiembre.



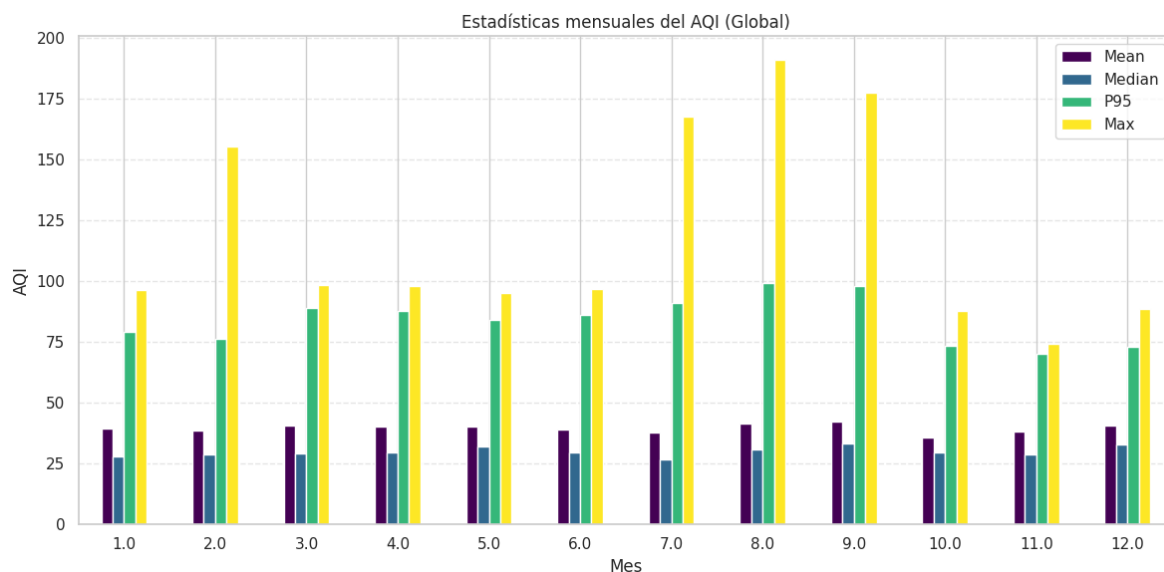
Estadísticas mensuales del AQI (Global)

La siguiente grafica muestra 4 métricas claves agrupadas por meses.

Lo primordial es que como media tenemos que conservan buenos niveles de calidad de aire estando en el rango de “Bueno”.

Como patrones observables podríamos decir que tenemos variación estacional, ya que en los meses de invierno se muestran valores mas altos en todas las métricas, especialmente en el P95 y Max. Los meses de verano presentan valores más bajos.

La brecha entre Mean y Median, y, P95 y Max es bastante significativa, indicando que, aunque la calidad del aire es generalmente moderada, hay episodios muy puntuales de contaminación severa.

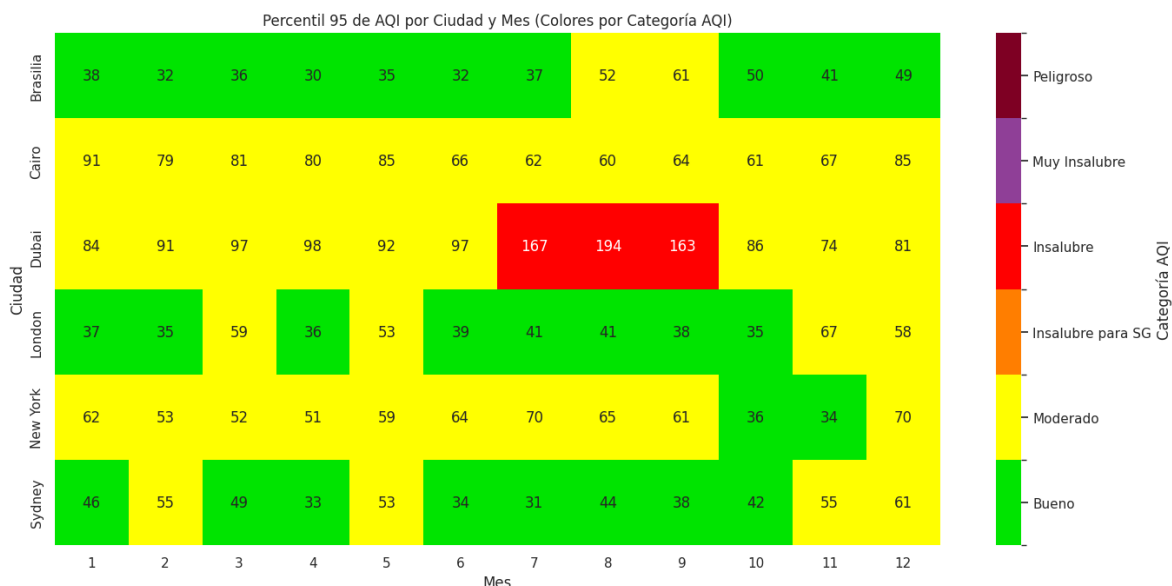


Con esta información se podría investigar a futuro las causas de picos en inversión (calefacción, inversión térmica, industria, etc), analizar si esos valores máximos son del mismo lugar y comparar contra datos meteorológicos y locales de cada ciudad para entender los patrones.

Análisis del percentil 95 del AQI por ciudad y mes

Para poder identificar los periodos de mayor concentración de contaminantes en cada ciudad, se calculó el percentil 95 (P95) del AQI por ciudad y mes. Este enfoque permite enfocar el análisis en las condiciones mas críticas del aire excluyendo los valores extremos aislados.

Estos datos fueron usados para mostrarlos en el siguiente heatmap que permite observar de manera visual y comparativa la evolución mensual del percentil 95 del AQI.



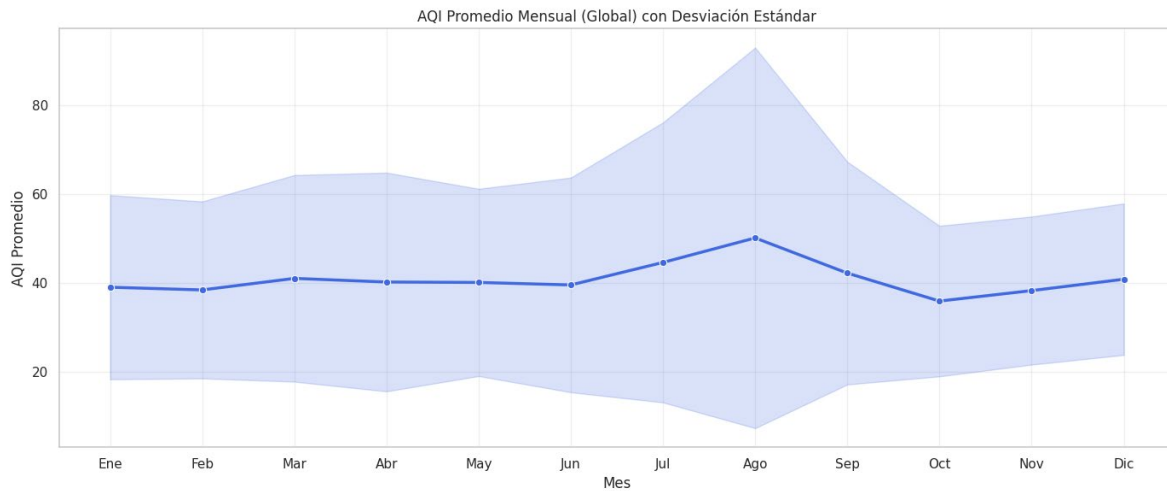
- **Brasilia** presentó una calidad del aire generalmente buena durante todo el año, con solo dos meses (agosto y septiembre) entrando en la categoría Moderate.
- **Cairo y New York** mantuvieron un perfil Moderate constante durante todo el año, lo cual sugiere niveles de contaminación más persistentes, aunque sin llegar a niveles insalubres.
- **London** se caracterizó por una mayoría de meses con calidad Good, aunque con picos Moderate entre marzo y diciembre.
- **Sydney** mostró una dinámica estacional interesante: calidad del aire Good durante gran parte del año, pero con meses Moderate intercalados, particularmente en febrero, mayo, noviembre y diciembre.
- **Dubai** fue la única ciudad que registró valores de AQI en la categoría Unhealthy: durante julio, agosto y septiembre, el percentil 95 superó los 150, indicando condiciones de aire potencialmente dañinas para la salud, incluso para la población general.

Este enfoque por percentiles confirma que, aunque muchas ciudades pueden tener promedios aceptables, los picos de contaminación son una preocupación importante, especialmente en regiones como Oriente Medio (Dubai), donde las condiciones extremas se extienden por varios meses consecutivos.

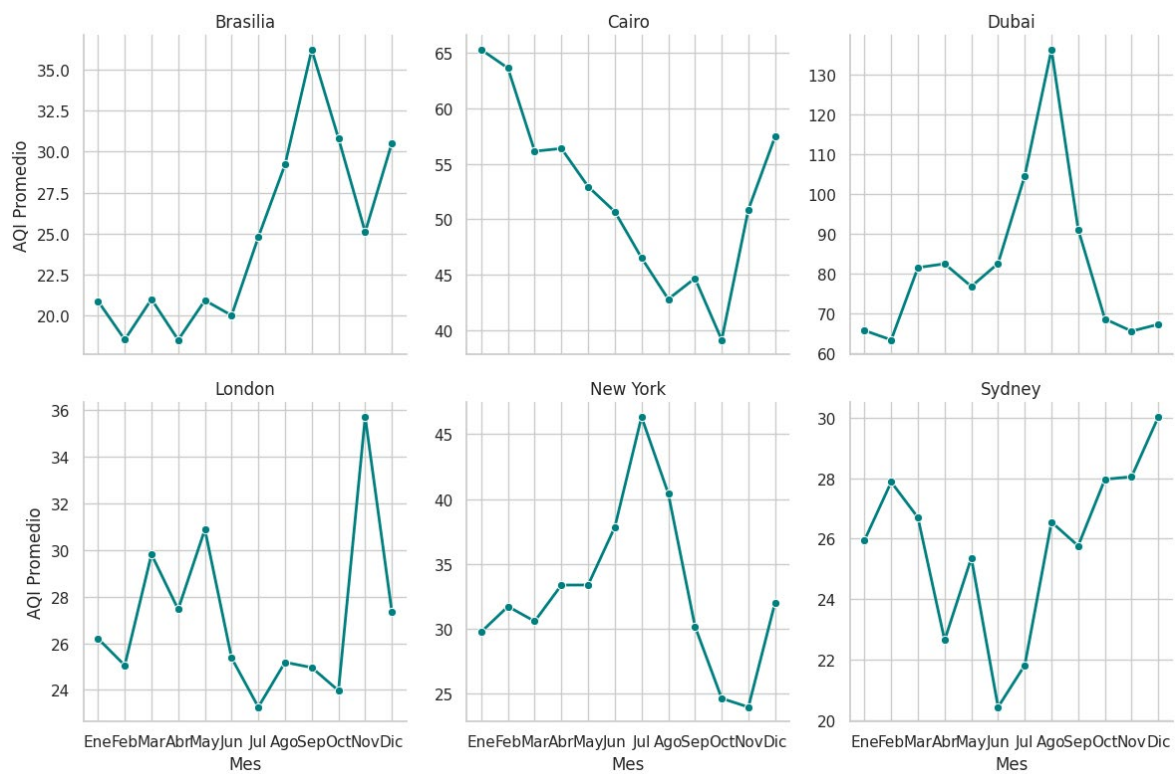
Tendencia Mensual del AQI Global Promedio

Con el propósito de entender la evolución general de la calidad del aire, se extrajeron las variables temporales clave a partir de Date. La siguiente grafica es la evolución del AQI promedio mensual global con su desviación estándar.

La desviación estándar en esta grafica nos ayuda a ver como las mediciones en promedio suelen no variar salvo en agosto, donde encontramos una gran variación, confirmando lo que se comentó anteriormente de episodios de contaminación extremos durante ese periodo.



La amplitud nos sugiere que el promedio mensual no varía drásticamente, pero que si existen diferencias marcadas especialmente en esos meses.



Modelado

Regresión Lineal Múltiple

Selección de Variables y Preprocesamiento

Para predecir el AQI se seleccionaron como variables predictoras los principales contaminantes atmosféricos: CO, NO₂, SO₂, O₃, PM_{2.5} y PM₁₀. Además, se incluyó la ciudad como variable categórica, la cual fue codificada utilizando Label Encoding, generando una nueva columna CityEncoded.

Debido a que los contaminantes tienen distintas escalas, se aplicó una estandarización mediante StandardScaler, lo cual garantiza que cada variable tenga media cero y desviación estándar uno, permitiendo una comparación más justa entre las características.

División del Conjunto de Datos

El conjunto se dividió en entrenamiento y prueba utilizando una proporción del 80/20. Con esto se aseguró que el modelo fuera evaluado sobre datos no vistos durante su entrenamiento, obteniendo una medida realista.

Entrenamiento del Modelo

Se utilizó Regresión Lineal Múltiple, que permite cuantificar la relación entre los contaminantes atmosféricos y el AQI. Este modelo, aunque sencillo, es útil como línea base para evaluar el rendimiento de técnicas más complejas posteriores.

Evaluación del modelo

El modelo fue evaluado utilizando tres métricas estándar:

- **MAE (Error Absoluto Medio):** 7.37
- **RMSE (Raíz del Error Cuadrático Medio):** 10.31
- **R² (Coeficiente de Determinación):** 0.84

El valor de **R² = 0.84** indica que el modelo logra explicar aproximadamente el **84%** de la varianza del AQI, lo cual representa un buen desempeño para un modelo lineal. Tanto el MAE como el RMSE reflejan un error promedio bajo, especialmente considerando que el AQI puede fluctuar entre valores de 0 a más de 150.

Random Forest

Selección de Variables y Preprocesamiento

Dado que algunas variables temporales como la hora, el día y el mes tienen un comportamiento cíclico (por ejemplo, la hora 23 está seguida por la 0), se transformaron en variables trigonométricas (seno y coseno) para preservar esta naturaleza circular, mejorando así el aprendizaje del modelo.

Además, se mantuvieron las variables numéricas originales relevantes, como las concentraciones de contaminantes y variables temporales ya procesadas.

División del Conjunto de Datos

El conjunto completo se dividió en datos de entrenamiento y prueba, reservando el 20% para evaluación, garantizando la representatividad y evitando sobreajuste.

Entrenamiento del Modelo

Se entrenó un **Random Forest Regressor**, un modelo basado en el ensamblaje de múltiples árboles de decisión que permite capturar relaciones no lineales y patrones complejos en los datos.

Evaluación del modelo

El modelo fue evaluado utilizando métricas de regresión:

- **MAE (Error Absoluto Medio): 2.95**, mostrando un error promedio muy bajo en las predicciones.
- **RMSE (Raíz del Error Cuadrático Medio): 5.17**, que penaliza errores grandes y confirma un alto nivel de precisión.
- **R^2 (Coeficiente de Determinación): 0.96**, indicando que el modelo explica el 96% de la variabilidad observada en el AQI.

El desempeño del Random Forest supera ampliamente al modelo lineal previo, evidenciando su capacidad para modelar las complejas interacciones entre variables meteorológicas, contaminantes y factores temporales. La incorporación de variables temporales cíclicas fue clave para capturar patrones horarios y estacionales de la calidad del aire.

Este modelo ofrece un pronóstico robusto del AQI que puede ser útil para alertas tempranas y planificación ambiental.

XGBoost Regressor

Selección de Variables y Preprocesamiento

Se reutilizaron las variables empleadas previamente, incluyendo:

- Contaminantes atmosféricos: CO, NO₂, SO₂, O₃, PM_{2.5} y PM₁₀.
- Variables temporales codificadas cíclicamente.
- Codificación de ciudad.

Para el preprocesamiento se aplicó un escalamiento estandarizado (StandardScaler) a las variables continuas, para asegurar un entrenamiento más estable y eficiente. También se confirmó que todas las variables categóricas estaban codificadas numéricamente y la variable objetivo fue nuevamente AQI.

División del Conjunto de Datos

Se mantuvo la misma participación de entrenamiento y prueba utilizada en los modelos anteriores para garantizar la comparabilidad directa entre los modelos.

Entrenamiento del modelo

Se utilizó XGBoost Regressor, un algoritmo de boosting que combina múltiples árboles de decisión en forma secuencial, optimizando los errores residuales de iteraciones anteriores. Este modelo es conocido por su alto rendimiento, manejo eficiente de valores faltantes, y capacidad para modelar relaciones no lineales y complejas.

Evaluación del modelo

El desempeño del modelo se evaluó utilizando las mismas métricas de regresión:

- **MAE (Error Absoluto Medio):** 3.68
- **RMSE (Raíz del Error Cuadrático Medio):** 5.78
- **R² (Coeficiente de Determinación):** 0.95

El modelo XGBoost mostró un rendimiento competitivo, con una capacidad explicativa del 95% de la varianza del AQI. Aunque su precisión fue ligeramente inferior al modelo de Random Forest (MAE más alto en ~0.7 puntos), sigue siendo una alternativa robusta y altamente efectiva, especialmente en contextos donde se prioriza la capacidad de generalización y manejo de relaciones no lineales.

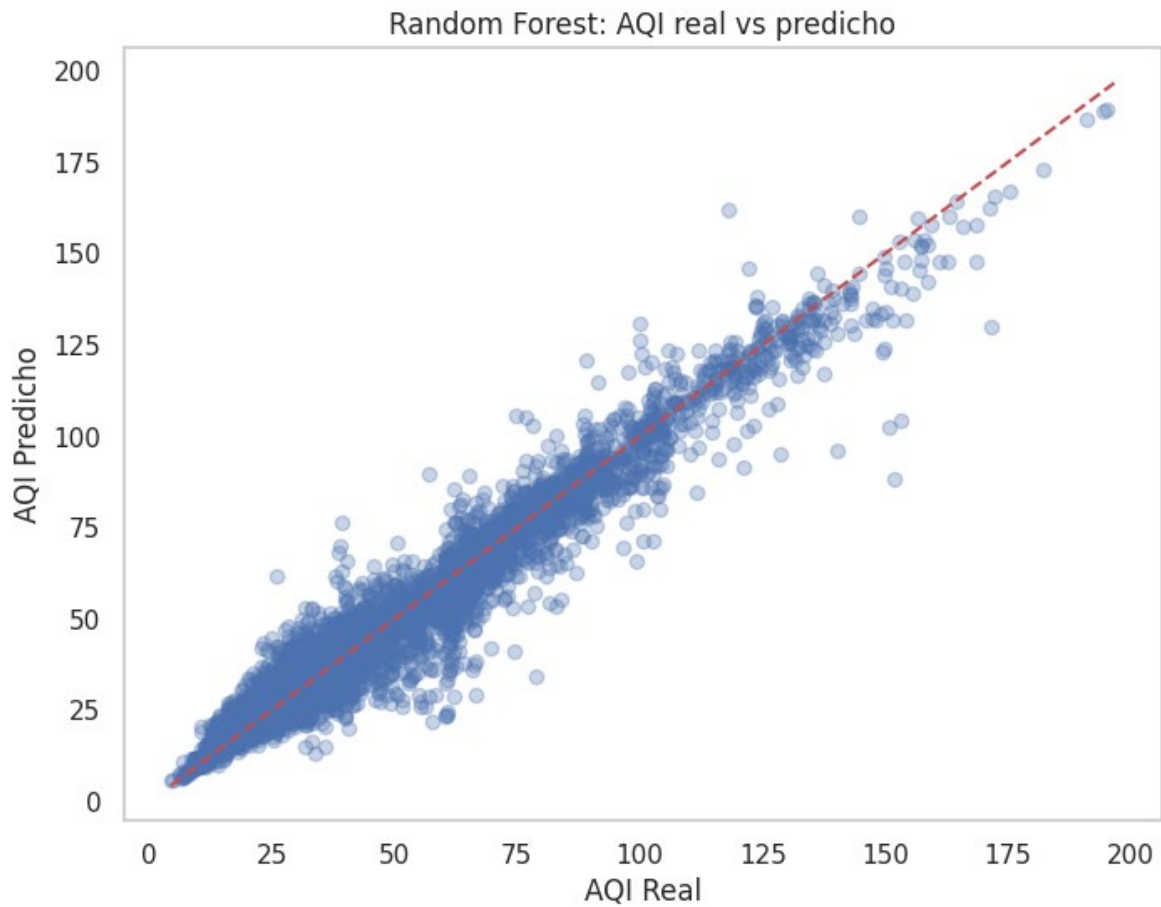
Evaluación de los resultados

Una vez entrenados y evaluados los 3 modelos de regresión, se procedió a comparar sus desempeños utilizando métricas estándar para problemas de regresión: MAE, RMSE y R2. A continuación un resumen de los resultados obtenidos:

Modelo	MAE	RMSE	R2
Regresión Lineal	7.37	10.31	0.84
Random Forest	2.95	5.17	0.96
XGBoost	3.68	5.78	0.95

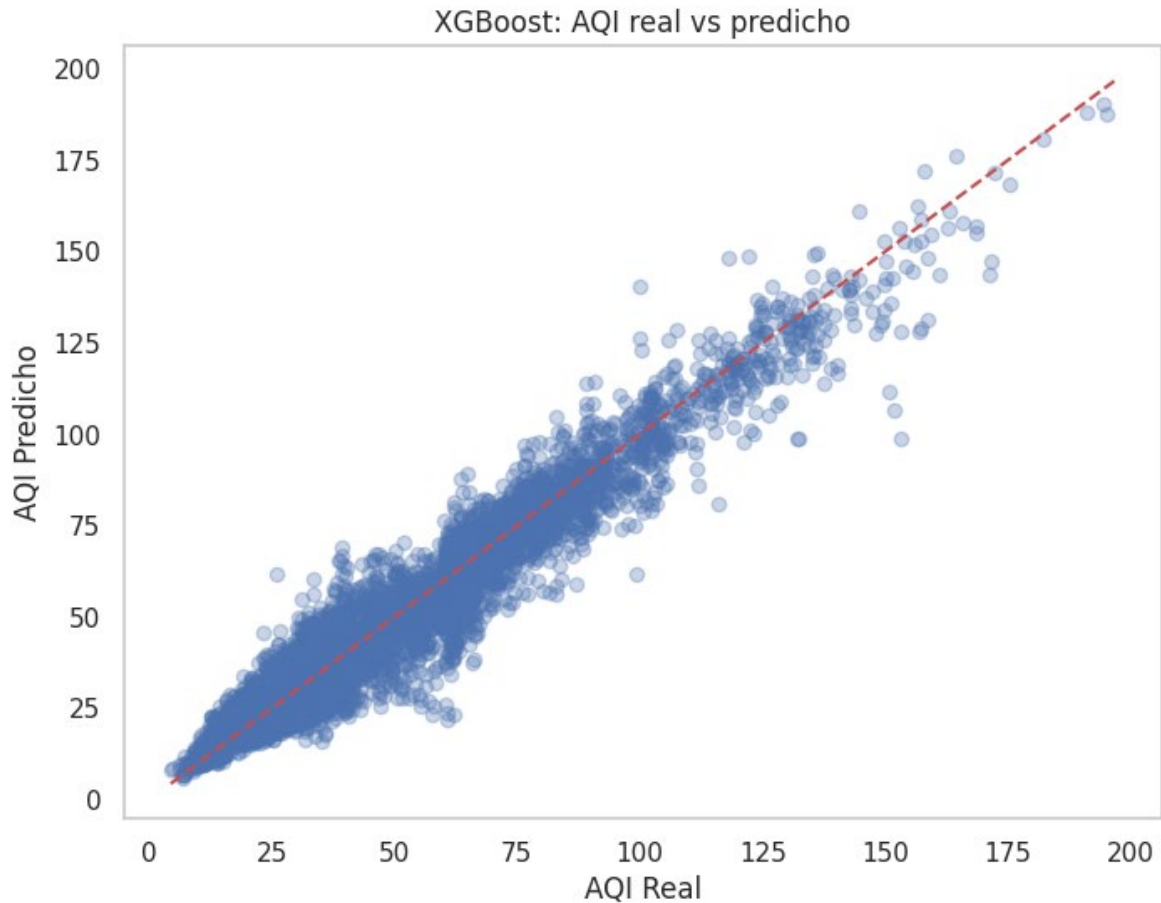
Como podemos apreciar, **Random Forest** fue el modelo con mejor rendimiento general. Obtuvo el menor error absoluto y cuadrático, así como el mayor coeficiente R2, lo cual indica una excelente capacidad para modelar la relación entre los contaminantes y el índice AQI. También mostró buena robustez sin requerir ajustes complejos.

Se incorporó un gráfico de dispersión para comparar los valores reales contra los predichos. En el gráfico de random forest podemos apreciar como hay un buen ajuste y concentración de puntos alrededor de la línea roja de referencia. Random Forest destaca en sus predicciones con valores AQI reales a comparación de los otros 2.



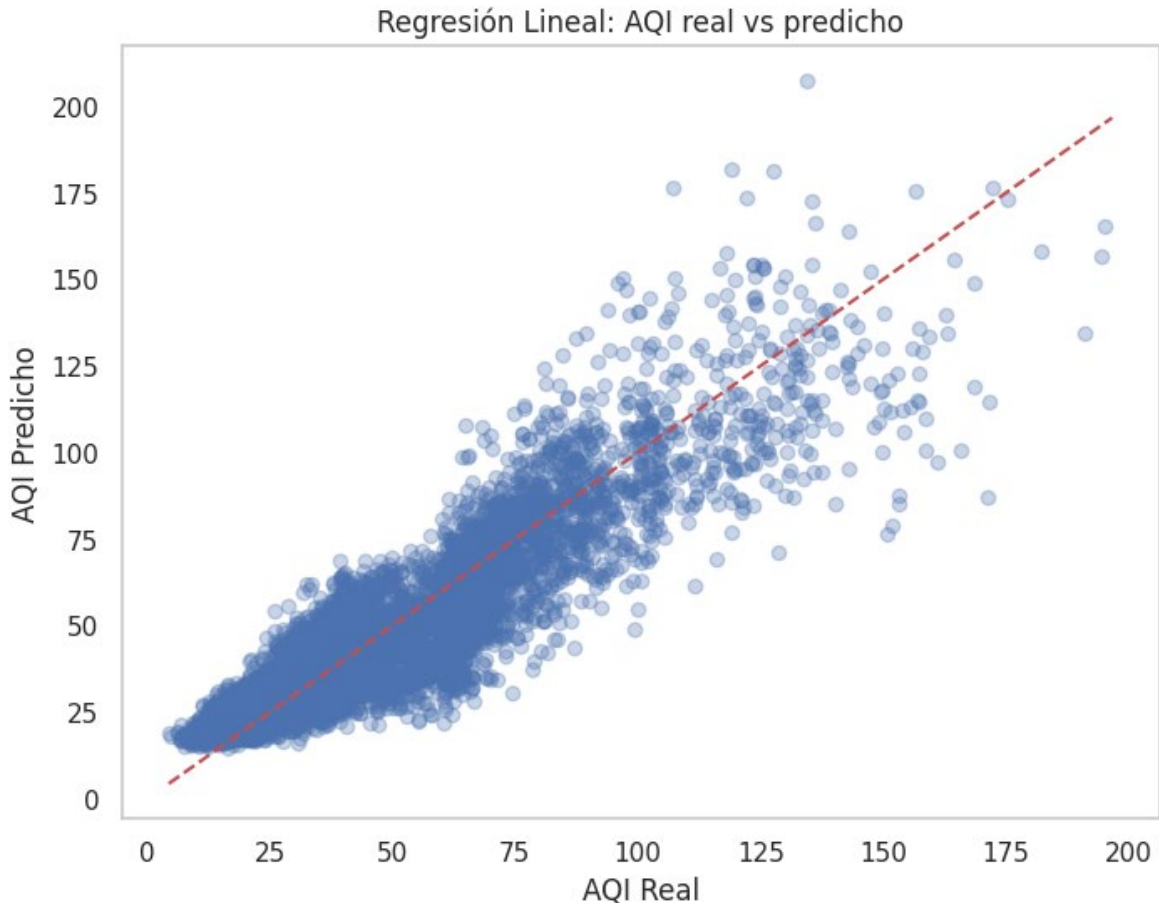
El **XGBoost** aunque fue ligeramente menos preciso que el RF, también ofreció resultados muy sólidos. Su rendimiento sugiere que es capaz de capturar relaciones no lineales complejas, lo es una opción útil si se desea escalar el modelo a otros contextos o integrar más variables en el futuro.

En el grafico de dispersión podemos apreciar igual un ajuste bueno, ligeramente mas disperso que Random Forest.



Por último, la **Regresión Lineal** presentó un desempeño considerablemente inferior. Aunque su implementación es sencilla y su interpretación directa, los errores obtenidos muestran que **no logra capturar adecuadamente la complejidad del fenómeno de la contaminación del aire**, lo cual es esperable en presencia de relaciones no lineales y efectos temporales cíclicos.

En el caso de la Regresión Lineal, se observa una dispersión significativa de los puntos, especialmente en valores altos de AQI. Esto es coherente con su bajo valor de R^2 y su mayor error absoluto.



Estos resultados nos responden directamente una de las preguntas centrales del proyecto:

¿Es posible predecir la calidad del aire usando modelos de aprendizaje automático?

Donde Si, modelos como random forest y XGBoost logran predecir el AQI con un alto grado de precisión, lo que demuestra que el aprendizaje automático es una herramienta poderosa para anticipar condiciones de riesgo ambiental.

Además, al observar las métricas de error, se identifica que estos modelos podrían utilizarse como base para sistemas de alerta temprana, informes ciudadanos o incluso apoyar políticas públicas, lo que se alinea con el público objetivo del proyecto.

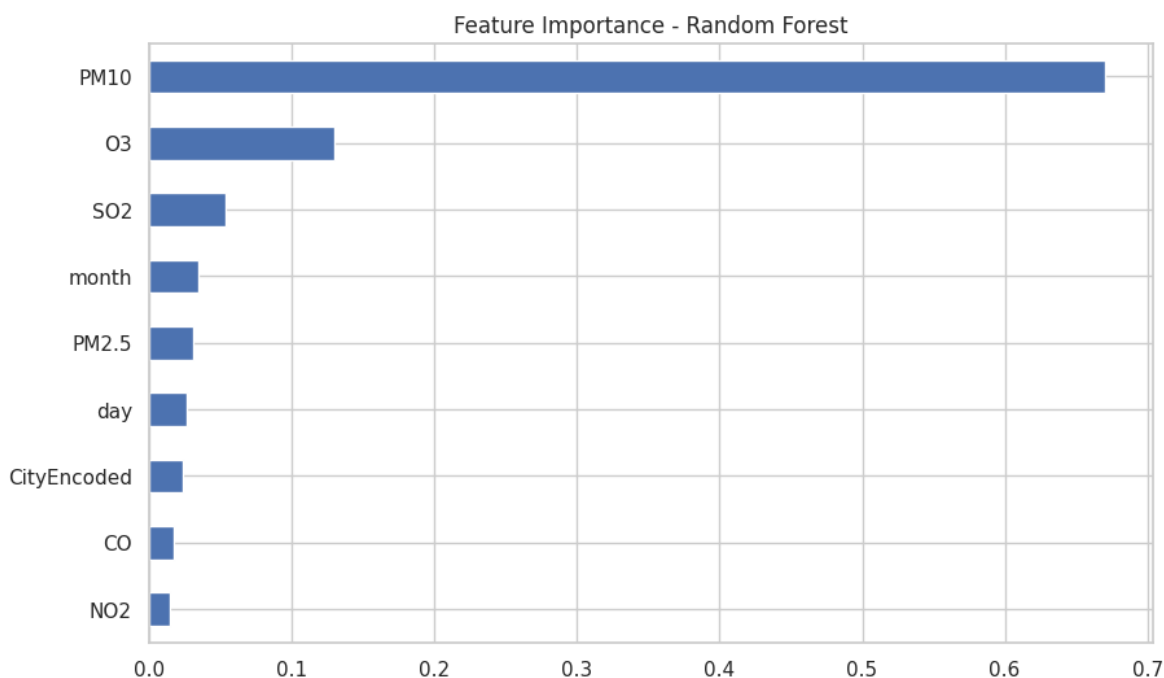
También considerando los gráficos de dispersión, podemos apreciar que tienen a fallar en sus predicciones mientras mayor sea el AQI, pero esto es algo completamente esperado dado que en el análisis inicial vimos que no tenemos casi valores grandes de AQI y que la mayoría se encuentran entre el rango de “Bueno” y “Moderado”.

Interpretación del Modelo: Importancia de Características

Para reforzar la interpretación del modelo de Random Forest, se evaluó la importancia relativa de cada variable en el proceso de predicción. En este caso, se generó un gráfico de barras para ver el peso de cada característica y se imprimieron los porcentajes correspondientes.

Los resultados obtenidos muestran que la variable PM10 es, por un amplio margen, la mas influyente en la predicción del índice de calidad del aire, aportando un 67.06% del peso total del modelo. Le siguen, aunque con menor contribución O3 con 13.01%, y SO2 con 5.31%. Las variables temporales como el mes y el día, así como la ciudad codificada también tienen cierta relevancia, aunque marginal.

Estos resultados son consistentes desde la perspectiva ambiental y la formación del valor AQI, pero podemos comprender de mejor manera que variables están impulsando las predicciones del modelo entrenado.



- PM10: 67.06%
- O3: 13.01%
- SO2: 5.31%
- Month: 3.42%
- PM2.5 : 3.03%
- Day: 2.60%
- CityEncoded: 2.38%
- CO: 1.72%
- NO2: 1.46%

Evaluación específica por ciudad

Para enriquecer aún más el análisis y evaluar el desempeño del modelo en contextos urbanos individuales, se realizó un entrenamiento independiente por ciudad. Esta estrategia permite observar si ciertos entornos urbanos presentan mayor predictibilidad que otros, y si los modelos generalistas pueden verse superados por enfoques especializados.

Se utilizó un modelo **XGBoost** para cada ciudad de forma independiente, entrenando y evaluando sobre subconjuntos específicos. La siguiente tabla resume las métricas obtenidas:

Ciudad	MAE	RMSE	R ²
Brasilia	0.52	1.54	0.99
Dubai	3.94	40.17	0.90
Sydney	2.14	11.39	0.88
New York	3.17	23.84	0.87
London	2.00	14.03	0.86
Cairo	4.24	35.42	0.85

Los resultados revelan un rendimiento particularmente alto en **Brasilia**, con un R² de **0.99**, seguido por Dubai y Sydney. Esta alta precisión en Brasilia puede estar relacionada con una menor variabilidad de datos o una correlación más clara entre los contaminantes y el índice AQI.

En contraste, ciudades como **Cairo** y **Dubai** presentan errores absolutos más elevados (MAE superiores a 3.9), lo que sugiere condiciones ambientales más complejas o factores externos que afectan la predicción del AQI y que podrían requerir modelos más sofisticados o fuentes adicionales de información.

Este análisis no solo valida la aplicabilidad del modelo general, sino que también abre la posibilidad de diseñar **modelos locales optimizados** para entornos urbanos específicos, lo cual resulta especialmente valioso para políticas públicas diferenciadas por región.

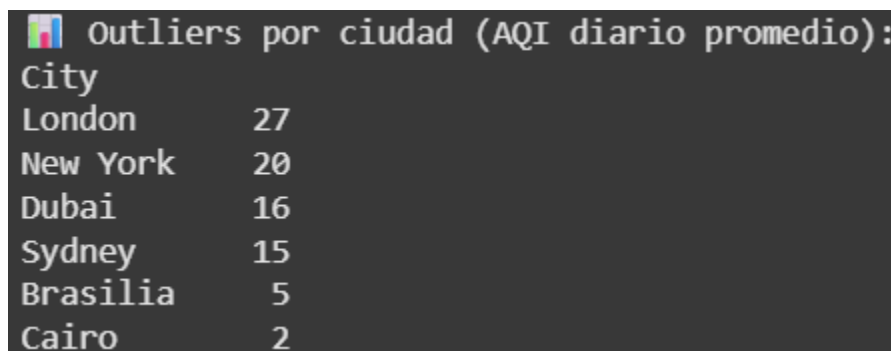
Análisis de outliers en el promedio diario del AQI

Con el fin de explorar la variabilidad extrema del índice de calidad del aire (AQI), se realizó un análisis de outliers sobre el promedio diario del AQI por ciudad. Para ello:

- Se eliminó la información horaria para trabajar con fechas puras.
- Se calculó el promedio diario del AQI por ciudad.
- Se identificaron valores atípicos utilizando el método del rango intercuartílico (IQR), considerando como outliers aquellos días cuyo AQI promedio supera el tercer cuartil en más de 1.5 veces el IQR.

Este análisis pone en evidencia que Londres y Nueva York presentan la mayor cantidad de días con valores anómalos de AQI, lo cual puede deberse a eventos puntuales como incendios, tráfico elevado o condiciones meteorológicas particulares. Estos outliers pueden afectar el rendimiento del modelo, especialmente si no son bien representados en el conjunto de entrenamiento.

En contraste, Cairo y Brasilia muestran una distribución mucho más estable, con muy pocos valores atípicos. Esto concuerda con los buenos resultados obtenidos previamente en la predicción del AQI en dichas ciudades, particularmente en el caso de Brasilia.



City	
London	27
New York	20
Dubai	16
Sydney	15
Brasilia	5
Cairo	2

Despliegue o Aplicación Práctica

Tras entrenar y evaluar múltiples modelos de regresión para predecir el Índice de Calidad del Aire (AQI), los resultados obtenidos permiten visualizar diversos escenarios de aplicación práctica, tanto desde una perspectiva técnica como social. A continuación, se presentan las principales rutas de uso e implementación del modelo.

Monitoreo inteligente del aire en tiempo real

El modelo predictivo puede integrarse en sistemas de monitoreo ambiental para ofrecer una estimación confiable del AQI incluso cuando faltan sensores o existen fallos temporales en la recolección de datos. Esto permitiría a las ciudades:

- Mantener alertas tempranas en caso de niveles críticos.
- Estimar el impacto de actividades humanas (tráfico, eventos) en la calidad del aire.
- Sugerir recomendaciones automatizadas a la población en función de las predicciones.

Optimización de políticas públicas ambientales

Gracias al análisis de importancia de características, se identifica que contaminantes como el **PM10** (con una contribución del 67.06%) y el **O₃** (13.01%) tienen un fuerte impacto en el AQI. Esta información resulta estratégica para:

- Enfocar regulaciones más estrictas en fuentes emisoras de partículas PM10.
- Diseñar campañas específicas para reducir los niveles de ozono en ciudades críticas.

Modelos personalizados por ciudad

Se entrenaron modelos independientes para cada ciudad, lo que permitió capturar mejor las particularidades locales. Los resultados mostraron que ciudades como **Brasilia** alcanzan una precisión altísima ($R^2 = 0.98$), lo que demuestra la viabilidad de modelos locales. Esta segmentación permite:

- Desplegar sistemas personalizados de pronóstico por ciudad.
- Incorporar variables adicionales de interés local (e.g., humedad, tráfico, altitud).
- Mejorar la aceptación ciudadana al percibir recomendaciones más ajustadas a su contexto.

Identificación de eventos extremos

El análisis de outliers en el AQI diario reveló que ciudades como Londres y Nueva York experimentan más días con calidad del aire fuera de lo normal. Esto sugiere la posibilidad de integrar mecanismos de vigilancia automática para detectar posibles eventos críticos o anomalías, como incendios o contaminaciones puntuales.

Conclusiones

A lo largo de este proyecto, se abordó el desafío de predecir el Índice de Calidad del Aire (AQI) mediante técnicas de minería de datos, utilizando un conjunto de datos multicidad con mediciones de contaminantes atmosféricos durante el año 2024. El proceso completo, guiado por la metodología CRISP-DM, permitió no solo entrenar modelos con alto desempeño, sino también obtener conocimientos clave sobre los factores que afectan la calidad del aire en distintas regiones del mundo.

Entre los principales hallazgos se destacan los siguientes:

- **Modelos con alta precisión:** Algoritmos como XGBoost y Random Forest lograron desempeños notables, con valores de R^2 cercanos o superiores al 0.90, lo que demuestra una capacidad robusta para predecir el AQI a partir de contaminantes medidos diariamente.
- **Influencia dominante del PM10:** El análisis de importancia de características mostró que el **PM10** es el principal predictor del AQI, representando más del 67% de la contribución del modelo. Este hallazgo refuerza la necesidad de enfocar esfuerzos de mitigación en este contaminante.
- **Variabilidad por ciudad:** Los modelos entrenados de manera específica para cada ciudad reflejaron que el comportamiento del aire y su predicción están profundamente ligados al contexto local. Por ejemplo, Brasilia presentó un desempeño prácticamente perfecto ($R^2 = 0.98$), mientras que otras ciudades como El Cairo mostraron más desafíos.
- **Identificación de eventos extremos:** El análisis de outliers reveló que ciudades como Londres y Nueva York tienen mayor frecuencia de días con condiciones fuera de lo normal, lo que puede ser crucial para diseñar estrategias de respuesta ante emergencias ambientales.

Aprendizajes obtenidos

A nivel técnico, este proyecto permitió consolidar habilidades en todo el ciclo de ciencia de datos: desde la preparación y limpieza del dataset, hasta el modelado, validación y evaluación. Se profundizó en el manejo de algoritmos de regresión, interpretación de resultados, y sobre todo, en cómo traducir los hallazgos en conocimiento útil. También se aprendió a tomar

decisiones basadas en evidencia y a adaptarse a las particularidades de los datos reales, como el desbalance o la variabilidad por región.

Limitaciones del análisis

Pese a los buenos resultados, es importante reconocer ciertas limitaciones:

- **Datos estáticos de un solo año (2024):** Al contar únicamente con datos de un solo año, el modelo no puede generalizar sobre tendencias a largo plazo o estacionales a nivel multianual.
- **Ausencia de variables contextuales:** Factores como temperatura, humedad, velocidad del viento, políticas ambientales locales, o eventos excepcionales (como incendios o festividades) no fueron considerados, lo cual podría afectar la precisión en ciertos casos.
- **Modelo entrenado sobre datos históricos:** Aunque los modelos predicen bien con los datos disponibles, no se probó su desempeño en un entorno completamente nuevo o en tiempo real.

Posibles mejoras futuras

Para futuras versiones del proyecto, se podrían considerar las siguientes mejoras:

- **Ampliar el periodo de análisis** para incluir varios años, lo cual permitiría detectar patrones estacionales, cambios en políticas ambientales o eventos atípicos.
- **Incluir variables climáticas y socioeconómicas**, que pueden tener un impacto directo sobre la calidad del aire y enriquecer los modelos.
- **Implementar modelos en tiempo real** que permitan predicciones automatizadas y alertas para la población en caso de niveles críticos de contaminación.
- **Desarrollar dashboards interactivos** o herramientas visuales que faciliten la toma de decisiones por parte de autoridades ambientales o usuarios comunes.