Análisis de la calidad del Aire de la ciudad de Bogotá, aplicando métodos estándar de limpieza y Análisis de Datos

Juan José Garcia Sánchez

Asesora

Dayana Alejandra Barrera Buitrago

Universidad Nacional Abierta y a Distancia UNAD

Escuela de Escuela de Ciencias Básicas, Tecnología e Ingeniería ECBTI

Programa Especialización en Ciencia de Datos

2025

Resumen

La contaminación del aire constituye un problema crítico de salud pública a nivel mundial, y Colombia no es la excepción. En Bogotá, considerada la ciudad más afectada del país, los niveles de contaminantes atmosféricos superan con frecuencia los límites recomendados por la Organización Mundial de la Salud (OMS), generando impactos significativos en la salud, la calidad de vida y los costos sociales y económicos.

Este estudio se analiza la calidad del aire en Bogotá mediante un enfoque de descubrimiento de conocimiento en bases de datos (KDD), utilizando registros del Ministerio de Ambiente. Se seleccionaron ocho variables clave: PM_{10} $PM_{2.5}$,CO, NO, NO_2 , O_3 , SO_2 y la velocidad del viento; tras un proceso de limpieza y preprocesamiento, se aplicaron técnicas estadísticas descriptivas, análisis de correlación y visualización de distribuciones.

Los resultados revelan que las concentraciones promedio de PM_{10} (32.5 µg/m³), $PM_{2.5}$ (16.0 µg/m³) y NO_2 (29.3 µg/m³) superan o se aproximan a los umbrales sugeridos por la OMS (45, 15 y 25 µg/m³, respectivamente). Además, se identificaron correlaciones moderadas (r = 0.45–0.65) entre contaminantes, lo que sugiere fuentes de emisión compartidas.

Finalmente, se proponen acciones como la realización de más investigacion académica, expansión de la red de monitoreo; el uso de tecnologías más limpias ; políticas diferenciales por estrato y densidad urbana; y el fortalecimiento normativo y tecnológico en entornos residenciales, para mejorar la gestión ambiental en Bogotá.

***Palabras Clave*:** calidad del aire; análisis exploratorio de datos (EDA); validación de datos; material particulado; análisis Estadístico; Visualización de Datos, descubrimiento en bases de datos (KDD).

Abstract

Air pollution is a critical public health problem worldwide, and Colombia is no exception. In Bogotá, considered the most affected city in the country, levels of air pollutants frequently exceed the limits recommended by the World Health Organization (WHO), generating significant impacts on health, quality of life, and social and economic costs.

This study analyzes air quality in Bogotá using a knowledge discovery database (KDD) approach, utilizing records from the Ministry of Environment. Eight key variables were selected: PM_{10}, $PM_{2.5}$, CO, NO, NO_2, O_3, SO_2, and wind speed. After a cleaning and preprocessing process, descriptive statistical techniques, classification analysis, and distribution visualization were applied.

The results reveal that the average concentrations of PM_{10} (32.5 μg/m³), $PM_{2.5}$ (16.0 μg/m³), and NO_2 (29.3 μg/m³) exceed or approach the thresholds suggested by the WHO (45, 15, and 25 μg/m³, respectively). Furthermore, moderate correlations ($r = 0.45–0.65$) were identified between pollutants, suggesting shared emission sources.

Finally, actions are proposed, such as conducting more academic research, expanding the monitoring network, using cleaner technologies, implementing differentiated policies based on socioeconomic stratum and urban density, and strengthening regulations and technology in residential areas, to improve environmental management in Bogotá.

keywords: air quality; exploratory data analysis (EDA); data validation; particulate matter; statistical analysis; data visualization; database discovery (KDD).

Tabla de Contenidos

Lista de Tablas

Lista de Figuras

Lista de Apéndices

Introducción

La contaminación del aire en Colombia ha empeorado considerablemente en el siglo XXI, llegando a niveles los suficientemente altos para reducir la expectativa de vida de los colombianos: desde el año 1999 al 2022, el nivel de contaminación particulada aumento en un 52.8%, y alcanzo 3.2 veces los niveles de contaminación recomendados por la OMS (AQLI, 2022). La exposición a la contaminación atmosférica puede causar múltiples enfermedades crónicas, causando enfermedades como la neumonía, canceres en el sistema respiratorio, infecciones respiratorias, accidentes cerebrovasculares, y enfermedades del corazón(Bouza et al., 2022; Ritchie & Roser, 2021) . Estos problemas de salud causan la reducción en 2.8 años de la calidad de vida (AQLI, 2022), y la muerte de 26.4 personas por cada 100 mil habitantes en Colombia (Ritchie & Roser, 2021).

Esta mala calidad del aire no es nueva ni exclusiva de Colombia: es uno de los principales factores de riesgo de muerte, contribuyendo a 1 de cada diez muertes en el mundo (Ritchie & Roser, 2021). Casi toda la población (99%) vive en lugares donde el aire excede los límites de contaminantes establecidos por la OMS (5 μgm^{-3} concentración anual para PM2.5), y las personas que viven en países subdesarrollados están en especial riesgo (World Health Organization, 2016). Estas cifras se deben a la poca regulación existente: menos del 90% de los países poseen una regulación completa para la calidad de aire (Represa, 2020) .

En Colombia, no se tiene un programa nacional para el manejo o mejoramiento de la calidad del aire, pero se tienen políticas nacionales como la de CONPES 3943, y políticas específicas en ciertas ciudades, como en Bogotá (AQLI, 2022), ambas orientadas a incentivar la renovación del parque automotor, la implementación de mejoras técnicas y prácticas en la industria (CONPES, 2018).

La amenaza que representa la contaminación atmosférica está bien clara, pero su diagnóstico y formas de solución o manejo son problemas complejos, que dependen de la colaboración de múltiples expertos de diferentes ciencias. La ciencia de datos en este campo representa una de las herramientas básicas para recopilar, limpiar, y analizar datos para el entendimiento de la situación. En Colombia este diagnóstico es responsabilidad a nivel nacional por el ministerio de Ambiente y desarrollo Sostenible (MinAmbiente), y es ejecutado por el instituto de Hidrología, Meteorología y Estudios ambientales (IDEAM). A nivel regional el IDEAM se coordina con las corporaciones autónomas regionales (CAR) para realizar vigilancia y control en asuntos ambientales. Toda la información recopilada está disponible en la página del subsistema de información de la calidad del aire SISAIRE (http://sisaire.ideam.gov.co/ideam-sisaire-web/).

Sin embargo, esta tarea de diagnóstico se ha visto limitada en Colombia por la poca cantidad de profesionales disponibles para analizar la problemática, y por la poca discusión desde un enfoque multidisciplinario (Bouza et al., 2022). Sin embargo, previamente ya se han realizado estudios sobre la calidad del aire, como por ejemplo el estudio de (Albarracín et al., 2023), el cual recolectó datasets abiertos y usó minería de datos y aprendizaje automático para completarlos (Mura et al., 2020) describiendo los cambios de la calidad del aire en Colombia (Cadavid-Giraldo et al., 2017) estudia la calidad del aire y el potencial de la ciudad del valle de Aburra para convertirse en una ciudad inteligente. Todos estos artículos usan la ciencia de datos como herramienta para portar información y análisis como insumos para la proposición de soluciones y la toma de decisiones de las instituciones reguladoras.

En este trabajo se abordará la situación ambiental de la ciudad de Bogotá, con un enfoque particular en la contaminación del aire, un fenómeno que ha generado preocupación por sus efectos en la salud pública y la calidad de vida de los habitantes. La investigación busca indagar cuáles son los principales factores que inciden en los niveles de contaminación registrados, así como identificar tendencias y relaciones entre estos factores a lo largo del tiempo. Para ello, se trabajará con datos provenientes de estaciones de monitoreo ubicadas en diferentes sectores de la ciudad, recolectados durante un periodo de cinco años. Estos datos serán sometidos a procesos de limpieza, transformación y análisis exploratorio, empleando como guía la metodología del Descubrimiento de Conocimiento en Bases de Datos (KDD), con el fin de obtener una comprensión más profunda del fenómeno y generar hallazgos que puedan contribuir a la toma de decisiones informadas en materia ambiental.

Justificación

La calidad del aire es un problema relevante para la población debido a los impactos directos a la salud y al gasto público (Bouza et al., 2022). En ese sentido, la exposición al aire contaminado causa efectos negativos sociales y económicos en Colombia: enfermedades y muertes, restricción del desarrollo de actividades económicas, y un aumento en la carga del sistema de salud (CONPES, 2018).

Además, El gasto público en salud se ve exacerbado por el deterioro de la calidad del aire : los costos públicos debidos a muertes y enfermedades asociadas a la contaminación urbana en Colombia se estiman en los 15.4 billones de pesos, según el departamento nacional de planeación (DNP, 2017).

La investigación y el estudio de la calidad del aire brindan herramientas para comprender en profundidad los factores que afectan la calidad del aire, y permiten el desarrollo de soluciones (Bouza et al., 2022). Los estudios resultantes sirven como insumos importantes para la creación de políticas públicas que lleven a la disminución de la contaminación del aire siendo por medio del ordenamiento del territorio, y la gestión del riesgo por contaminación.

Objetivos

Objetivo General

Aplicar técnicas de análisis de datos sobre un conjunto de datos históricos de calidad del aire recolectados durante cinco años en diferentes zonas de la ciudad de Bogotá, con el propósito de identificar factores y patrones de contaminación atmosférica, que permitan comprender la dinámica del fenómeno y generar recomendaciones orientadas a mejorar la gestión ambiental en la ciudad.

Objetivos Específicos

Identificar variables relevantes para el diagnóstico, en el estudio de la calidad del aire, a partir de la literatura existente, para su análisis en este trabajo.

Realizar una limpieza y preprocesamiento de los datos sobre la calidad del aire, para facilitar el análisis y visualización entre características.

Analizar la relación entre la concentración de contaminantes, material particulado, y variables meteorológicas; y su papel en la disminución de la calidad del aire, utilizando métodos estadísticos.

Formular recomendaciones orientadas a la mitigación de la contaminación atmosférica en Bogotá, a partir de los hallazgos obtenidos mediante el análisis exploratorio y el descubrimiento de patrones en los datos históricos de calidad del aire.

Marco Teórico

Antecedentes

Existen trabajos previos sobre el análisis de datos y la calidad del aire en varias ciudades de Colombia. Varios de estos se exponen las metodológicas y prácticas para la recolección de datos abiertos sobre la calidad del aire. Otros realizan un análisis, que varía desde un análisis estadístico básico hasta un análisis exploratorio de datos completo.

En (Albarracín et al., 2023) y (Henao et al., 2021), se discute la metodología para la recolección de datos en la capital de Bogotá y Medellín. Los datasets se recolectan por medio de sensores, desde la secretaria distrital de ambiente Bogotá; y también datasets abiertos obtenidos desde otras instituciones: datos desde SIATA para la calidad del aire, y desde la colección MODIS para la actividad de incendios.

En el trabajo de (Mura et al., 2020) se presenta un análisis completo de la calidad del aire en la ciudad de Bogotá, desde el año 1998 hasta la actualidad. Se presentan la limpieza y validación completa de los datos, así como el análisis estadístico descriptivo con respectivas visualizaciones para presentar la evolución histórica. En trabajo de (Represa, 2020) no se utilizan datos de Colombia, pero se desarrolla una metodología completa de análisis de datos, priorizando la presentación de resultados y la selección de algoritmos descriptivos y predictivos para la optimización del análisis.

En los trabajos de (Henao et al., 2021) y (Gómez & Molina, 2021) se avanza más allá y se presenta un análisis Exploratorio de datos, junto con el desarrollo de modelos de aprendizaje automático, donde se presentan las distribuciones y gráficos estándares, analizando las distribuciones de variables y correlaciones. También se aplicaron algunas técnicas de aprendizaje

automático para predecir características faltantes, o predecir variables de respuestas usándose modelos lineales, regresión logística, y arboles de decisiones con el enfoque de random forests.

Conceptos Generales

Medidas de Resumen Descriptiva

La estadística descriptiva se ocupa de analizar los datos disponibles para presentar resumen de los datos conocidos, y describir sus características (Samuels et al., 2012). representa parte de las herramientas básicas usando en todas las ciencias, asegurando la posibilidad de una investigación científica reproducible (Rubio, 2019)

Tendencia Central

Hacen parte de las medidas de estadísticas básicas para describir poblaciones o muestras de estas. En su conjunto son el promedio, la moda, y los cuantiles (Samuels et al., 2012). Su propósito es explorar que valores son los esperados o más frecuentes para una población en estudio (Raykov & Marcoulides, 2013)

Dispersión

Parte de los estadísticos básicos que se usan para comprender la distribución de los datos: entender que tan comprimido o disperso esta un dataset, determinar outliers y valores extremos (Vanderplas, 2017). Dentro de estos se encuentran la desviación estándar, la varianza, covarianzas entre variables, el rango intercuartílico y la curtosis (Rubio, 2019).

Figura 1

Formula de la desviación Estándar

$$\sigma_{est} = \sqrt{\frac{\sum(Y - Y')^2}{N}}$$

Donde Y es un conjunto de datos, y N es el tamaño de la muestra.

Figura 2

Cálculo de la Varianza

$$var = \sigma_{est}^2$$

Donde sigma es la desviación estándar.

Figura 3

Cálculo de la Curtosis

$$\text{kurt} = \frac{\mu_4}{\sigma^4}$$

Donde sigma es la desviación estándar, y miu es el cuarto momento centrado (Rubio, 2019).

Es el análisis estadístico realizado sobre dos o más variables (Raykov & Marcoulides, 2013). Se realiza para descubrir tendencias previamente desconocidas, hallando ya sea relaciones cerca de las lineales simples, relaciones monotónicas de crecimiento o decrecimiento, o por lo menos determinar si existe correlaciones entre las variables. El análisis estadístico estándar incluye el coeficiente de Pearson. Este mide la fuerza y el carácter de la relación lineal entre las variables en estudio, siendo útil para determinar la existencia de posibles correlaciones o relaciones lineales entre las variables (Rubio, 2019).

Figura 4

Coeficiente de Pearson

$$\rho = \frac{N\sum XY - (\sum X \sum Y)}{\sqrt{\{[N\sum x^2 - (\sum x)^2][N\sum y^2 - (\sum y)^2\}]}}$$

Donde N es el tamaño de la muestra; *x*, *y* son las dos variables en comparación.

Pruebas de Hipótesis

Método básico de la estadística inferencial, para comprobar si los datos son los suficientemente significantes para soportar una hipótesis estadística, que debe esta propuesta con base en la población en estudio (Samuels et al., 2012). Las pruebas T-Student y de tipo ANOVA son las más comunes, y utilizan la varianza entre dos o más variables para analizar las diferencias entre las medias (Samuels et al., 2012). El ANOVA es una generalización de la prueba T-Student, y una herramienta básica en el diseño de experimentos y el análisis de hipótesis, para comprobar significancias estadísticas de las diferentes correlaciones encontradas (Samuels et al., 2012).

Para los cálculos, se usó el lenguaje de programación de Python, debido a que su ecosistema ha evolucionado alrededor de aplicaciones científicas, desarrollándose las herramientas necesarias para el análisis de datos. El entorno de programación será el de Jupyter Notebooks. Estos se plantean usar de forma local, o por medio de la plataforma de Google Colab cuando el uso excesivo de memoria RAM se vuelva necesario.

Se usó el paquete de *pandas* (Pandas Docs, 2025) el cargue y limpieza de los datos, y se usaron los paquetes de *scipy* (The SciPy community, 2025) y *stats* (Seabold & Perktold, 2010) para aplicar técnicas de análisis sobre estos. Otros paquetes principales que incluidos: *sklearn* para la transformación y el uso de técnicas de aprendizaje automático, el uso de *matplotlib*

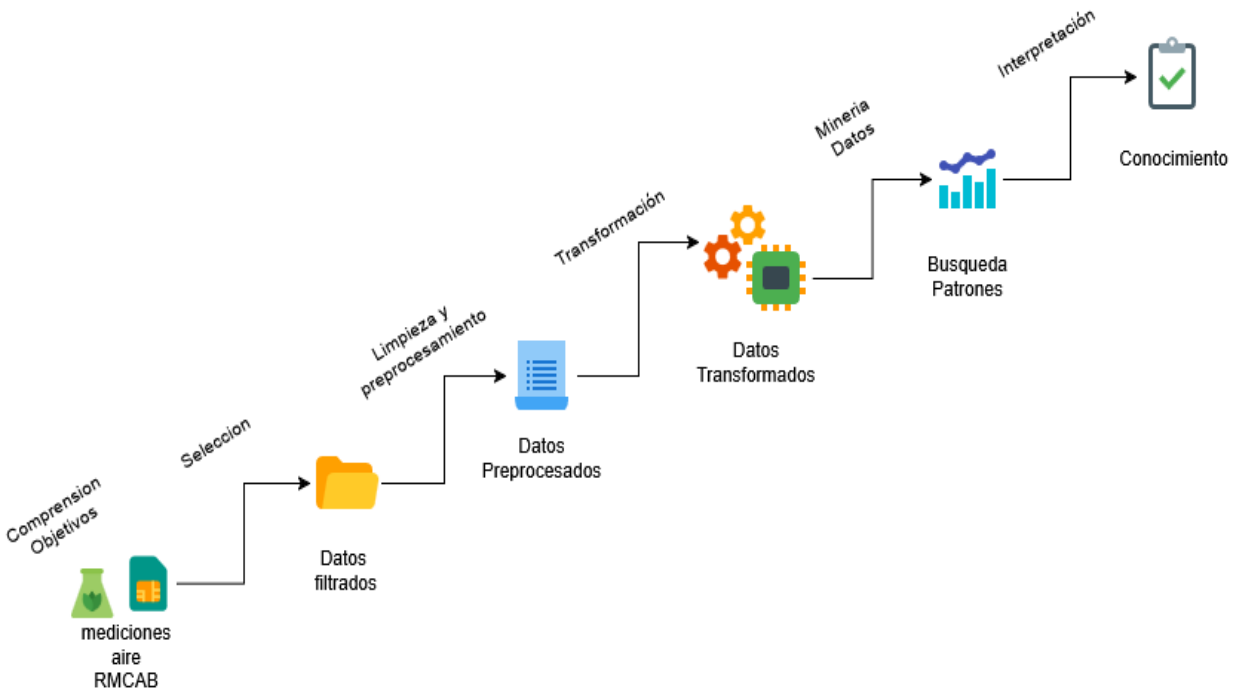
(Matplotlib development team, 2025) y de *seaborn* (Waskom, 2025) para la visualización de los resultados.

Metodología

Se eligió para este trabajo un enfoque de descubrimiento de conocimiento en bases de datos (KDD). La metodología KDD es una de las más clásicas y estructuradas, y plantea la exploración de grandes cantidades de datos que normalmente están no-estructurados (sin una forma o modelo predefinido), usando los métodos comunes de la estadística y de la ciencia de datos para encontrar relaciones o patrones desconocidos en conjuntos de datos. (Martins & Kaspars, 2023). En este trabajo se busca analizar una gran cantidad de datos, que tienen una naturaleza semiestructurada (formato CSV), y la metodología KDD es la más completa y simple para realizar un análisis exploratorio de este estilo. Las etapas principales del KDD se adaptaron en este trabajo, y se muestran en la figura 1: selección, preprocesamiento, transformación, minería de datos, e interpretación (Martins & Kaspars, 2023).

Figura 5

Etapas del Proceso KDD



Recopilación de Datos

Los datos históricos sobre la calidad del aire se obtuvieron por medio de la página del ministerio de educación, con el reporte SISAIRE (MinAmbiente, 2025). Estos datos corresponden a una recopilación que el ministerio realiza de los diferentes departamentos del país. Se seleccionaron específicamente los datos de la ciudad de Bogotá; que provienen de publicaciones de mediciones de la red de monitoreo de la calidad del aire (RMCAB) de la alcaldía. Estos datos son de acceso abierto, y fueron recolectados por medio de las estaciones ubicadas en varios sectores de la ciudad, estando disponibles desde el año 2010.

Los datos están estructurados en formato CSV, recogidos en intervalos de 24 horas, tomándolos de todas las localidades medidas de Bogotá. Los parámetros para cuantificar la contaminación del aire estandarizados: corresponden a la concentración de compuestos contaminantes; y el tamaño de su material particulado (Bouza et al., 2022). También se involucran parámetros meteorológicos, como son la temperatura, velocidades del viento, y la humedad, debido a que afectan la propagación de los contaminantes (Bouza et al., 2022). La medición en Colombia sigue esta misma línea, y el dataset está estructurado con mediciones de compuestos en el tiempo.

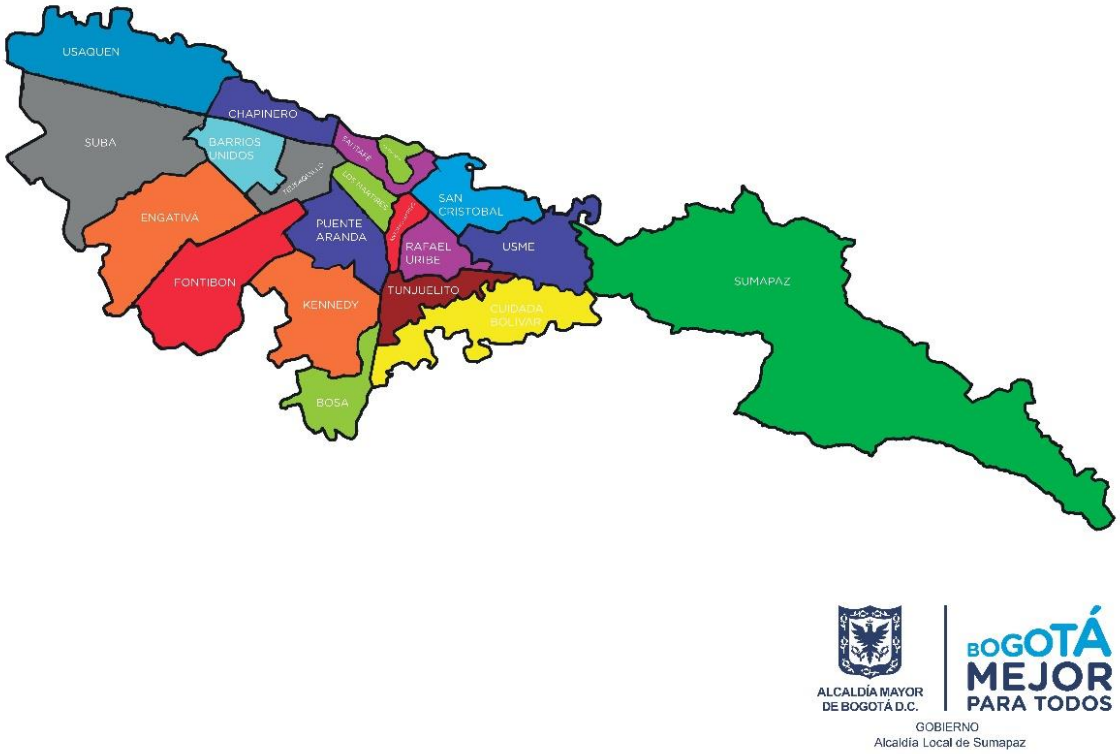
Caracterización de las Localidades

Las localidades en Bogotá son producto de la unión de varios núcleos urbanos a la ciudad de Bogotá en el año 1954, dividiéndose inicialmente en 11 zonas que evolucionaron a 20 en la actualidad (Duque-Duque & Molina Caro, 2021). Las localidades actuales son: Usaquén, Chapinero, Santa Fe, San Cristóbal, Usme, Tunjuelito, Bosa, Kennedy, Engativá, Suba, Fontibón, Ciudad Bolívar, Rafael Uribe, Uribe, Puente Aranda, La Candelaria, Los Mártires, Antonio Nariño, Teusaquillo, Barrios Unidos y Sumapaz.

Las zonas se han formado y reestructurado en pro de una organización de la ciudad que favorezca los propósitos del aumento de la participación ciudadana en la democracia, eficiencia en prestación y construcción de servicios públicos, y el mejoramiento económico y social (Duque-Duque & Molina Caro, 2021). La figura 2 muestra la ubicación de las localidades actuales.

Figura 6

Mapa de las Localidades en Bogotá



Nota. Tomado de (Alcaldia de Bogota, 2025)

La implementación de la estratificación socioeconómica en Colombia también ha tenido un efecto en el desarrollo y configuración de las localidades: Este sistema se creó en Colombia para atender su situación de desigualdad económica, subsidiando una parte de los servicios

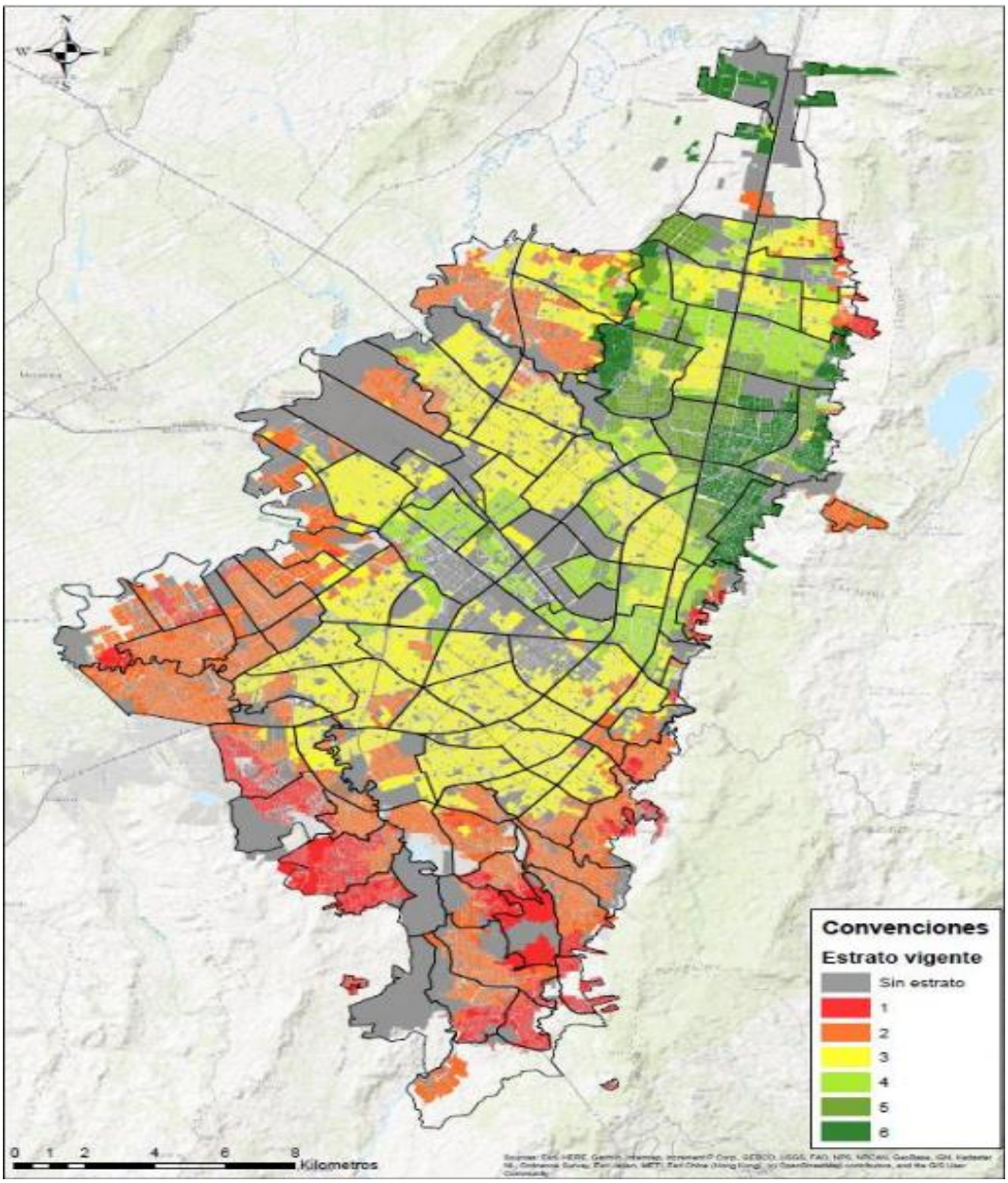
públicos de los hogares con más bajos recursos (estratos 1,2), mediante una tarifa diferencial que aumentaba el cobro a los hogares de más altos recursos económicos (estratos 4, 5, 6) (Yunda, 2019). El desarrollo urbano parece estar relacionadas y seguir el sistema de estratificación, dándose mayor una mayor densificación y transformación intensiva de casas y edificios en las manzanas de estrato 4, 5, y 6 (Yunda, 2019).

Como resultado de la evolución de la ciudad, el estrato y la densificación son más altos en las localidades que están más al norte de la ciudad: Usaquén, Suba, Chapinero y Barrios unidos concentran la mayoría de los barrios con nivel de estratos 4-6. Paralelamente, las localidades del sur concentran a los estratos más bajos, estando la mayoría en las localidades de Sumapaz, Ciudad Bolívar, Usme, San Cristóbal, Rafael Uribe y Bosa.

En la figura 3 se observa la distribución geográfica del estrato, mostrándose el aumento del estrato desde el sur hasta el norte de la ciudad. La figura 4 muestra la distribución de la densidad urbana, en forma geográfica y porcentual, indicando que el uso del área está en su mayoría en el estrato 6 (41% para comercios y oficinas, 30% para vivienda), Seguido del estrato 5 (21% para comercios y oficinas, 27% para vivienda). (Yunda, 2019). También se resalta que la mayoría de la población está en el estrato 2 (41%), pero solo tienen un uso del área para vivienda del 4%, y un 2% del uso para comercios y oficinas, siendo en general los estratos bajos los que tienen menor porcentaje del uso del área (Yunda, 2019).

Figura 7

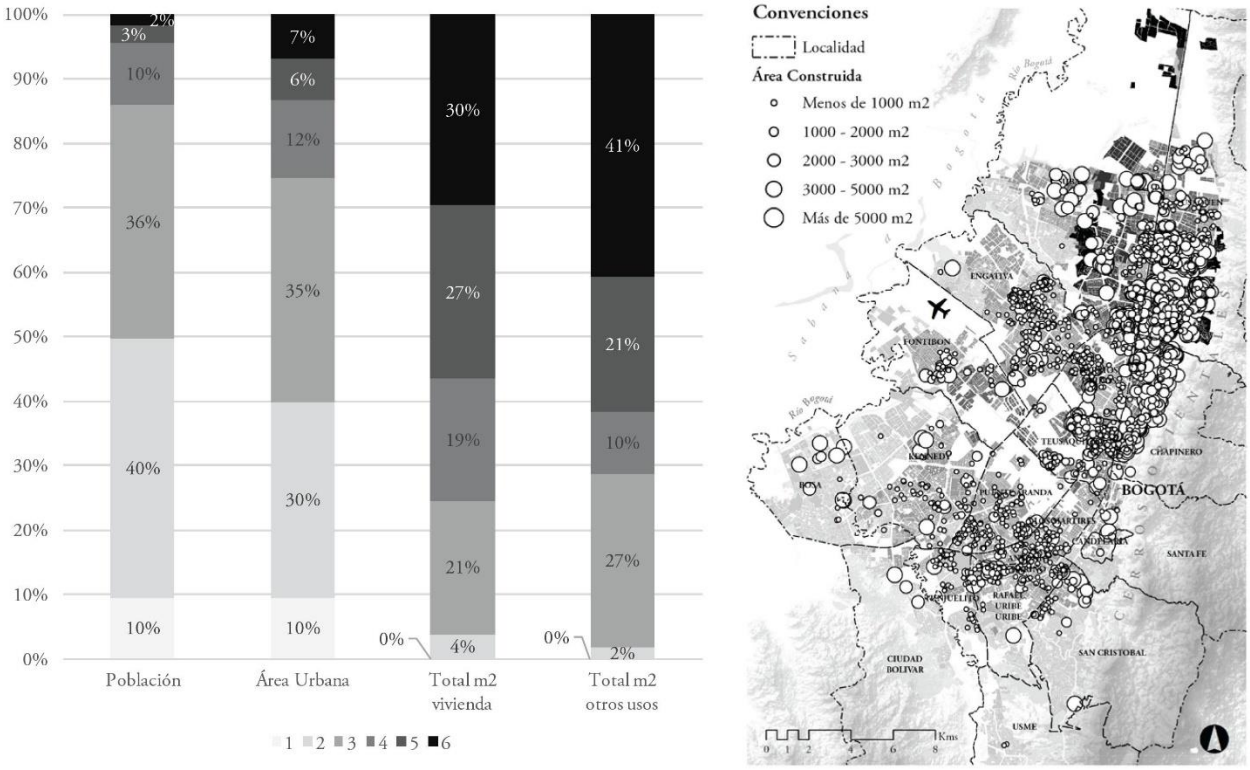
Distribución del Estrato Socioeconómico para las Localidades en 2024



Nota. Tomado de (Palma et al., 2024)

Figura 8

Distribución de la Densidad Urbana en la Ciudad de Bogotá



Nota. Tomado de (Yunda, 2019).

Análisis Exploratorio Datos

Se va a realizar un análisis estadístico unidimensional, calculando los estadísticos básicos mencionados. Después , se va a realizar un análisis bidimensional, donde se va a hacer el análisis entre pares de variables; y entre variables y la localidad donde se tomaron.

Por otra parte, se van a realizar gráficos de caja, histogramas y calcular coeficientes de Pearson (figura 4) para evaluar la fuerza lineal entre variables y existencia de correlaciones. Por otro lado, se elaborarán histogramas para entender mejor su distribución. Los posibles hallazgos encontrados serán validados posteriormente con pruebas de hipótesis sobre los coeficientes, con el propósito de verificar que las relaciones encontradas no se deban al azar.

Respecto a las localidades o estaciones, se compararán usando los resultados de variabilidad mediante gráficos de caja, para evaluar si existen diferencias significativas entre los niveles de contaminación de las localidades.

Finalmente, para poder comparar las distribuciones de las variables entre sí, se va a realizar una estandarización de las variables, y se van a realizar gráficos de caja y de cálculos de curtosis con las variables sin estandarizar.

A modo de cierre, los hallazgos obtenidos a través del análisis exploratorio de datos serán interpretados con el propósito de extraer conclusiones sobre los factores que más influyen en la calidad del aire en Bogotá. Con base en estos resultados, se identificarán áreas de oportunidad y se formularán recomendaciones orientadas a fortalecer la gestión ambiental y contribuir a la mejora de la calidad del aire en la ciudad

Resultados

En las siguientes secciones se va a discutir todo el proceso, los métodos y los estadísticos resultantes del análisis. Inicialmente, en la sección *Identificación de las Variables Determinantes* se introduce el conjunto de variables clave en el estudio, seleccionadas por su relevancia con el objetivo 1. Posteriormente, en la sección *Limpieza de los Datos*, se procede con el procesamiento, con el propósito de posibilitar la exploración de relaciones significativas y cumplir con el objetivo 2. La limpieza y el preprocesamiento se realizan antes de la selección, debido a que la selección debe realizar una evaluación sobre el conjunto limpio.

El análisis de datos se realiza a partir de la sección *Minería de Datos*; hasta la sección *Análisis Estadístico de la Interacción entre Contaminantes Atmosféricos.* Se aplican los cálculos estadísticos, de correlaciones y por localidades, cumpliendo con el objetivo 3 del proyecto. Consecuentemente, se formulan recomendaciones a partir de los hallazgos en la sección *Recomendaciones*, cumpliendo con el objetivo 4.

Estándares Recomendados por la OMS

Para poder entender y colocar en contexto los datos de la calidad del aire y los resultados de análisis, se debe tener un marco de referencia sobre sus valores esperados u óptimos para la reducción de la peligrosidad: en la tabla 1 se muestran los limites diarios recomendados por la OMS. Los datos de la OMS se convirtieron en las unidades de este trabajo, y se eligieron los promedios diarios, debido a que los puntos de datos del conjunto son promedios diarios. También hay que tener en cuenta que la OMS no ha establecido límites para el monóxido de nitrógeno (NO):

Tabla 1

Niveles Seguros de Contaminación Recomendados por la OMS

Variable	Limite recomendado por la OMS (promedio diario)	Unidad
CO	4000	$\mu g m^{-3}$
NO_2	25	$\mu g m^{-3}$
O_3	60	$\mu g m^{-3}$
PM_{10}	45	$\mu g m^{-3}$
$PM_{2.5}$	15	$\mu g m^{-3}$
SO_2	40	$\mu g m^{-3}$

Nota. Obtenido de (WHO, 2021)

Limpieza de los Datos

Para poder aprovechar los datos de las estaciones, es necesario limpiarlos y transformarlos ya que las estaciones pueden fallar o cometer errores de medición, insertando valores vacíos o numéricos inválidos. Los parámetros de medición también se encuentran separados por archivos y con rangos de tiempo dispares entre sí, dificultando su limpieza y análisis directo.

Con el objetivo de posibilitar la limpieza y el análisis, se cargaron los CSV en estructuras dataframes del paquete de *pandas*, uno de los estándares en la investigación para el análisis de datos (Pandas Docs, 2025). Sobre estos *dataframes* se eliminaron todos los datos que fueran ilógicos, y, por lo tanto, errores de medición: se removieron récords conteniendo datos negativos, y aquellos récords que cumplieran $PM_{10} < PM_{2.5}$, ya que, por definición, El PM_{10} mide la concentración de partículas con tamaño igual o menor a 10 micras, y debe contener a las partículas con tamaño menor a 2.5 micras.

Por último, se combinaron todos los *dataframes* en uno solo con la función *merge.* Esta función permite realizar combinación de bases de datos análogas a las consultas *Join* del lenguaje SQL (Pandas Docs, 2024). Para este trabajo, ser realizó una operación combinación de tipo *Outer Join* usando el nombre de las estaciones y la fecha de medición como campos para la operación. Esta permitió que se tomen las filas de ambos dataframes cuando haya colisiones, se inserte un valor nulo en los campos incompatibles, y que no se descarte ningún dato que pueda contener información importante.

Adicionalmente, para poder realizar los cálculos de correlaciones, pruebas de hipótesis y comparar variables en el análisis; es necesario realizar un proceso de estandarización de los datos; ya que las variables tienen diferentes rangos y unidades de dato. Se realizó una transformación de todas las variables con la clase *StandardScaler* del paquete de *sklearn*. La siguiente ecuación (figura 9) describe como se realiza internamente los cálculos, mostrando la normalización estadística alrededor de la media (scikit-learn, 2025):

Figura 9

Fórmula para realizar la estandarización

$$Z = \frac{x - \mu}{s}$$

Donde Z es el estadístico o variable normalizado, μ es la media de la muestra, *s* es la desviación estándar de la muestra.

Identificación de Variables Determinantes

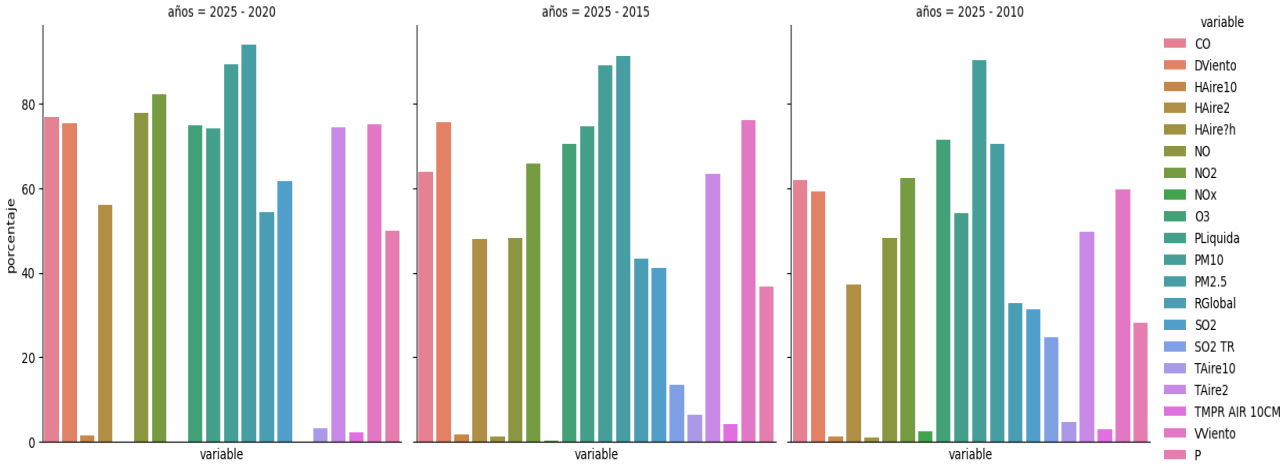
La selección de variables se basó en la búsqueda de parámetros clave en la literatura, en la disponibilidad y calidad de los datos disponibles. La calidad de los datos se midió con el porcentaje de datos no nulos que posee cada columna del conjunto de datos, donde un porcentaje

mayor implica mayor calidad. Posteriormente, se construyeron tres *dataframes* agrupando los datos por antigüedad, correspondientes a 5, 10 y 15 años antes del año 2025. Se calcularon los porcentajes de datos no nulos para cada nivel de antigüedad, con el propósito de buscar aquel nivel que tenga los porcentajes más altos, haciendo comparaciones para cada variable. En la figura 10, se pueden ver tres diagramas de barras comparando estos porcentajes. Se encontró que el porcentaje de no nulos, para la mayoría de las variables, es más alto para el nivel de 5 años atrás; donde la excepción es la variable de PM_{10} en el nivel de 15 atrás, mejorando en un 1% frente a los 5 años. En general, solo se observó una disminución de la calidad tomando los niveles de 10 o 15 años atrás.

Por último, se determinó que las variables de temperatura del aire a 10 metros (Taire10), a 2 metros (Taire2) y a 10 cm (Taire10cm), la humedad relativa (Haire) y la concentración de NOx (ver Tabla 3) presentan una calidad de datos muy baja, con niveles de completitud cercanos al 1–2 %, independientemente del nivel de antigüedad. Debido a esta deficiencia, dichas variables fueron descartadas del análisis y no serán consideradas en el estudio

Figura 10

Análisis de la Calidad de Datos del Dataset



Diccionario de Datos

Las variables de la estación donde se tomaron las mediciones del aire, y el día respectivo (Fecha inicial), son fundamentales para poder entender los datos, y están completas, por lo que se tomaron. De los artículos revisados para determinar la relevancia (Bouza et al., 2022; Mura et al., 2020; WHO, 2021), se encontró que las concentraciones de dióxido de carbono, óxido nitroso, ozono y concentración de material particulado son fundamentales para el entendimiento de la calidad del aire, ya que son los contaminantes más comunes (WHO, 2021). El dióxido de azufre y el monóxido de nitrógeno son menos comunes, pero se incluyeron, así como la velocidad del viento, por tener una gran disponibilidad de datos.

Tabla 2

Diccionario de Variables Seleccionadas

Campos	Descripción	Clasificación	Niveles
$Estacion$	Estación (nombre) donde se realizado la medición	categórica (texto)	24 valores de estación, representado localidad o una fuente de datos.
$Fecha\ inicial$	fecha del día de la medición	Categórica	texto con formato año-mes-día, desde el 2010-01-01 al 2025-03-31
CO	Concentración monóxido de carbono	numérica	número real mayor a cero, en μgm^{-3}

Campos	Descripción	Clasificación	Niveles
NO	Concentración monóxido de nitrógeno	numérica	número real mayor a cero, en μgm^{-3}
NO_2	Concentración de dióxido de nitrógeno	numérica	número real mayor a cero, en μgm^{-3}
O_3	Concentración de Ozono	numérica	número real mayor a cero, en μgm^{-3}
SO_2	Concentración de dióxido de azufre	numérica	número real positivo, en μgm^{-3}
$PM_{2.5}$	Concentración del material particulado de diámetro ≤ 2.5 micras	numérica	número real positivo, en μgm^{-3}
PM_{10}	Concentración del material particulado de diámetro a ≤ 10 micras	numérica	número real positivo, en μgm^{-3}
V_{Viento}	Velocidad del viento	numérica	número real positivo, en ms^{-1}

Variables Eliminadas

Las siguientes variables se eliminaron, ya sea por tener una calidad de datos muy baja (1-2%), como se vio en la sección *Identificación de Variables Determinantes* (Las variables de NO_x, SO2_TR, todas las variaciones de H_{aire}); o porque su tratamiento no aportaría nada útil para

el análisis de la calidad del aire, como es el caso de la radicación solar registrada, o la presión. Sin embargo, los datos de estas características todavía se encuentran en el conjunto de datos.

Tabla 3

Variables Eliminadas y No Estudiadas

Campos	Descripción	Clasificación	Niveles
NO_x	Concentración combinada de otros compuestos gaseosos de nitrógeno	numérica	número real positivo, en mgm^{-3}
SO_2TR	Concentración de azufre total reducido: compuesto de azufre gaseosos conteniendo organosulfuro	numérica	número real positivo, en μgm^{-3}
R_{global}	Radiación solar global	numérica	número real positivo, en Whm^{-2}
$Haire_{10}$	Humedad relativa del aire, medida a $10m$	numérica	porcentaje de saturación (%)
$Haire_2$	Humedad relativa del aire, medida a $2m$	numérica	porcentaje de saturación (%)
$Haire_h$	Humedad relativa del aire, medida a altura diferentes de 2 o 10 m	numérica	porcentaje de saturación (%)

Campos	Descripción	Clasificación	Niveles
$Taire_{10}$	Temperatura del aire, medida a $10m$	numérica	grados Celsius (°C)
$Taire_{2}$	Temperatura del aire, medida a $2m$	numérica	grados Celsius (°C)
$Taire_{10cm}$	Temperatura del aire, medida a $10cm$	numérica	grados Celsius (°C)
P	Presión atmosférica	numérica	número real positivo, en mmHg
$P_{liquida}$	nivel de precipitación	numérica	número real positivo, en mm
D_{Viento}	Dirección del viento	numérica	número real positivo, en grados

Minería de Datos

Se llevó a cabo un análisis detallado de la distribución y los valores representativos de las variables seleccionadas, con el fin de obtener una visión general de su comportamiento estadístico. En la tabla 4 se presentan los estadísticos descriptivos básicos, como el promedio, la desviación estándar y los cuartiles (Q1, mediana y Q3), los cuales permiten identificar tendencias centrales y niveles de dispersión. Esta información numérica se complementa con una representación gráfica mediante diagramas de caja, los cuales se encuentran en el **Gráfico 2**, y facilitan la identificación visual de la simetría, la dispersión y los rangos intercuartílicos de cada variable.

Para mejorar la legibilidad de los gráficos y evitar distorsiones visuales, se decidió eliminar los outliers de cada una de las visualizaciones. Adicionalmente, se construyeron diagramas de caja individuales para cada estación de monitoreo, con el propósito de observar las particularidades en la distribución de los datos según la localidad. Esta desagregación por zona permite detectar posibles diferencias espaciales en el comportamiento de las variables analizadas y aporta insumos relevantes para el diagnóstico a nivel territorial.

Tabla 4

Cálculo de los Estadísticos Básicos

	CO	NO	NO2	O3	PM10	PM2.5	SO2	VViento
count	25740	26299	26642	24959	30878	30878	19298	24157
mean	705.485	21.582	29.323	23.968	32.536	15.884	4.903	1.631
std	387.747	19.949	13.006	11.603	18.072	9.09	4.824	0.737
min	0.254	0.123	0.94	0	2.565	0.588	0.058	0
25%	453.19	8.854	19.149	15.377	19.288	8.875	2.248	1.142
50%	620.154	16.103	28.201	22.359	29.087	14.469	3.655	1.488
75%	863.445	27.443	38.039	30.648	41.986	21.125	6.066	1.917
max	5025.54	288.814	126.747	114.369	172.368	92.667	117.52	6.942
var	150348.109	397.96	169.162	134.624	326.608	82.623	23.271	0.543

La tabla 3 muestra que los valores de desviación estándar y varianza de todas las variables es baja, excepto la del monóxido de carbono, indicando que hay poca variabilidad para los cinco años de medición; y que para el monóxido de carbono hay mucha variación, que se produce por la existencia de mediciones grandes en ciertas partes de Bogotá, como se verá en la sección *Análisis Estadístico de la Interacción entre Contaminantes Atmosféricos*. La distribución de los datos, observada en la tabla 3 y en la figura 11, muestra que es asimétrica con sesgo hacia la izquierda, pero tendiendo a una distribución normal. En la figura 12, en la cual se muestran los

histogramas de las variables, se observa que en todos los casos se tiene una forma similar, sin presencia de colas pronunciadas, y con figuras de campana que se desvía hacia la izquierda.

Para corroborar de una forma más robusta estas observaciones de la asimetría y sesgo, se calculó la curtosis para cada una de las variables, construyéndose la tabla 5. Los valores de curtosis muestran que las variables del CO, NO Y SO_2 tienen los valores más grandes y mayores a cero, en ese sentido estos valores indican que los datos tiendes a los valores extremos, y que los histogramas tienen colas que se desvían significativamente de una distribución normal, así como picos más pronunciados. Las demás variables también tienen valores positivos, pero más cercanos a cero, indicando mayor cercanía a la distribución normal y colas menos pronunciadas.

Los picos pronunciados y colas largas de la distribución pueden indicar que posiblemente hay picos o focos ocasionales de alta contaminación para los casos del CO, NO y $SO_,$ y en algunas localidades de Bogotá. Por el contrario, para los casos del PM_{10} y $PM_{2.5}$, NO_2 O_3 y no se dan focos de contaminación, sino que se distribuye más uniformemente por localidades.

Tabla 5

Resultados del Cálculo de la Curtosis

Variable	Curtosis
CO	10.431
NO	12.602
NO_2	0.841
O_3	1.282
PM_{10}	1.553
$PM_{2.5}$	1.48
SO2	55.698
Vviento	4.299

Comparando los promedios de cada variable con los estándares de contaminación de la OMS (ver tabla 1), lo primero que se nota es que los valores promedio de las concentraciones de material particulado se acercan en al límite recomendado en el caso del PM_{10} (32.5 μgm^{-3}, se acerca a 45) y se excede en el caso del $PM_{2.5}$ (15.84 μgm^{-3} excede a 15). Por otra parte, en el caso de los compuestos contaminantes, solamente se supera el límite para del NO_2 (29.32 μgm^{-3} vs. 25), y se mantienen muy por debajo para los demás.

Por otra parte, cuando se analizan los cuartiles de la tabla y de la figura 11, se observa que más de la mitad de las mediciones supera el límite recomendado para el $PM_{2.5}$, aproximadamente el 20% de los datos supera el límite para el PM_{10} y aproximadamente el 50% supera el límite para NO_2. Para el resto de las variables, la totalidad de los datos (excluyendo outliers) se encuentran debajo de las recomendaciones. Por tanto, en Bogotá se está cumpliendo estrictamente con 3 estándares de contaminación (los estándares CO, O_3, SO_2), y estaría incumpliendo con 3 métricas de contaminación clave (PM_{10}, $PM_{2.5}$, NO_2).

Para terminar, se compararon nuevamente las variables, mediante el uso de diagramas de caja sin la presencia de valores atípicos y con la aplicación de un proceso de estandarización; que se presentan en la figura 13. Esta transformación permite que todas las variables compartan una misma escala, facilitando una comparación más objetiva de su comportamiento estadístico, independientemente de sus unidades originales. A partir de esta visualización, se aprecia que en general todas las cajas parecen tener una distribución casi simétrica alrededor de la mediana, pero desviándose hacia la izquierda; y en todos los casos teniendo tamaños similares, con la notable excepción de la variable "velocidad del viento", que evidencia una menor dispersión relativa frente al resto.

Figura 11

Diferentes Diagramas de Caja para las Concentraciones de Contaminantes, El PM, y la Velocidad del Viento, Durante todo el Tiempo

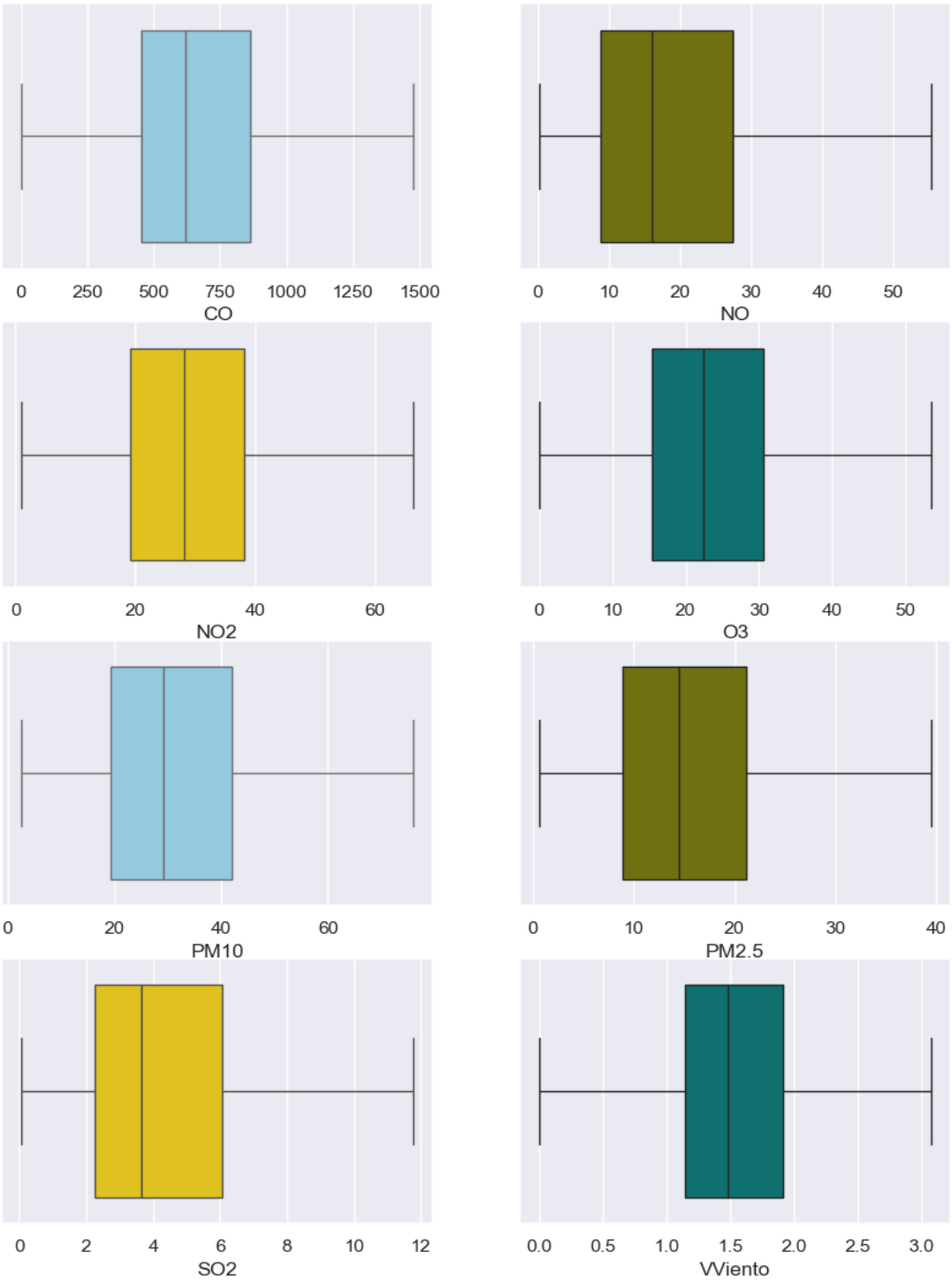


Figura 12

Histogramas para las Variables en Estudio

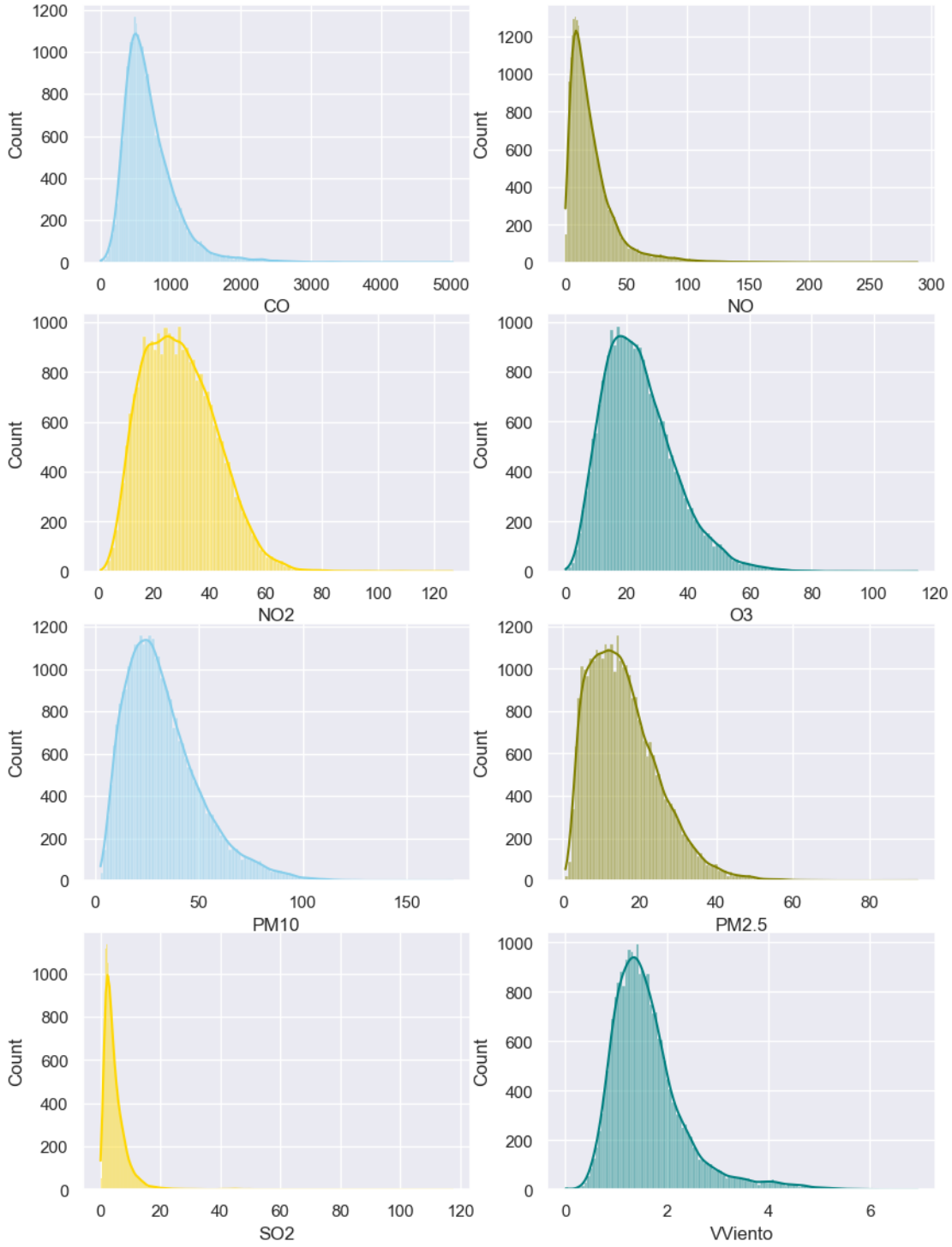
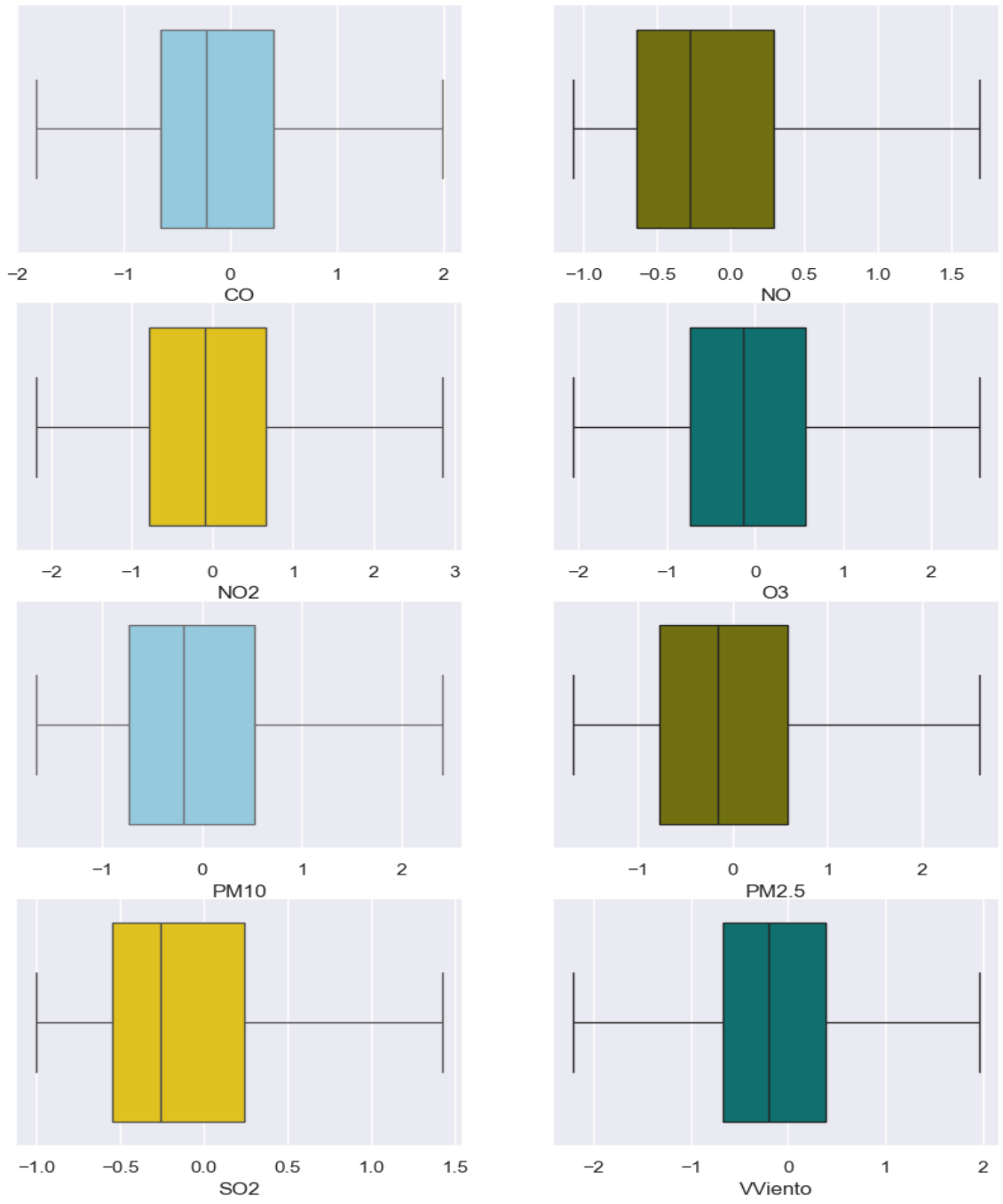


Figura 13

Diagramas de Variables Estandarizadas



Distribución de los Datos por Estación

Probando varios tipos de graficas, se determinó que la mejor forma de visualizar las diferencias entre las diferentes estaciones, y por extensión las localidades, es mediante el uso de diagramas de caja, dibujados para cada estación y para cada variable. Las figuras 14-21 muestran estas distribuciones, donde cada figura corresponde a una única variable, Asimismo, en ninguno de los diagramas se muestran los puntos outliers, para no saturar los diagramas de caja.

Realizar un análisis por localidades de esta forma resulta fundamental debido a las marcadas diferencias geográficas y demográficas que existen dentro de Bogotá. Cada localidad puede presentar condiciones particulares en cuanto a densidad poblacional, fuentes de emisión, dinámica urbana, lo que influye en la concentración y dispersión de contaminantes atmosféricos. Este proceso de desagregar la información va articulado al objetivo 3, ya que permite identificar patrones específicos, detectar posibles focos de contaminación y comprender mejor cómo varían las condiciones del aire a nivel territorial

Figura 14

Diagrama de Caja por Estación para CO

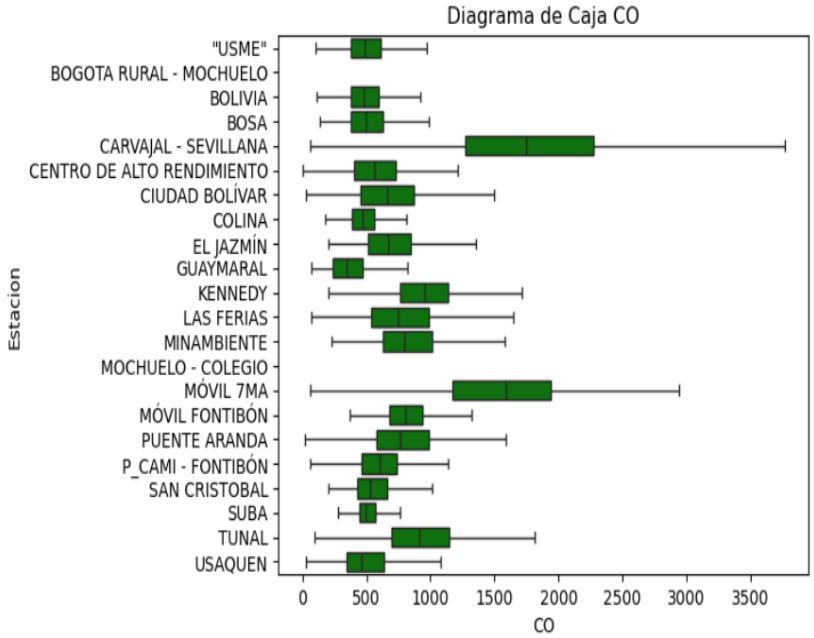


Figura 15

Diagrama de Caja por Estación para PM10

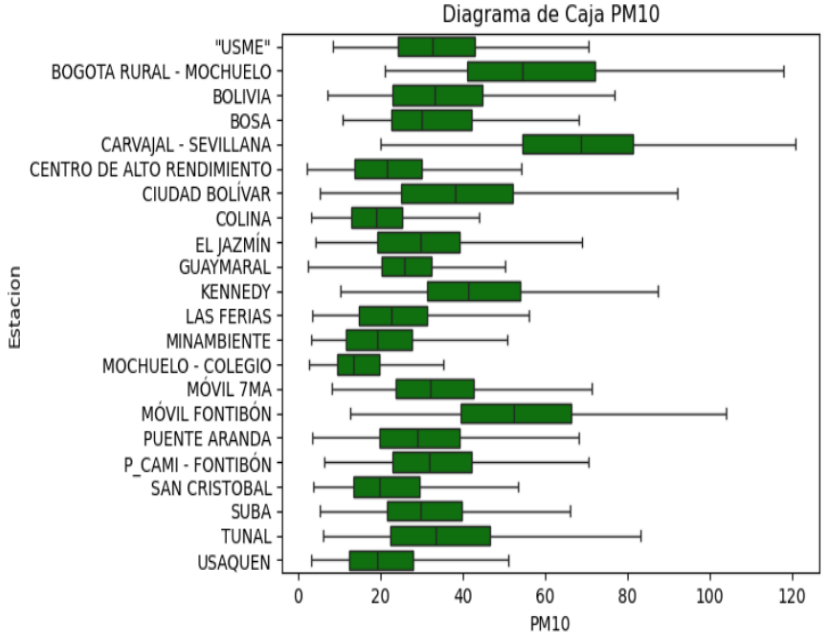


Figura 16

Diagrama de Caja por Estación para PM 2.5

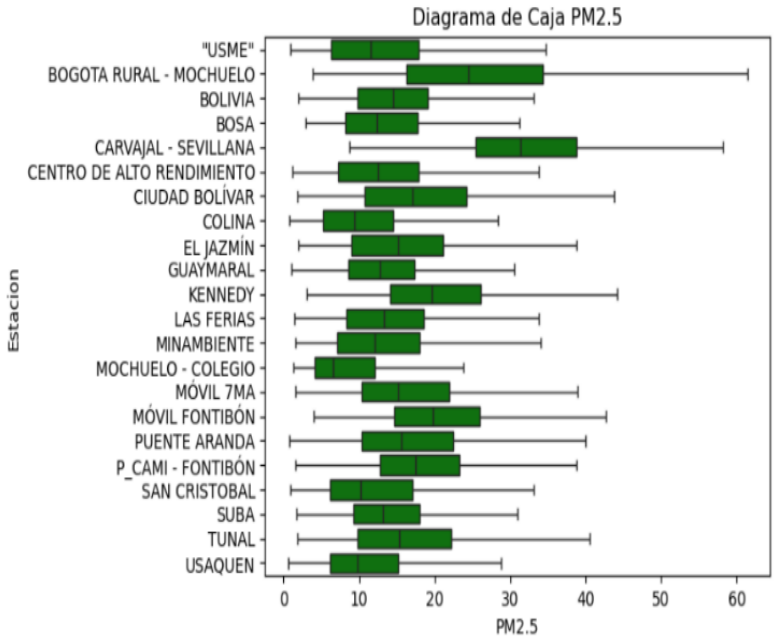


Figura 17

Diagrama de Caja por Estación para NO2

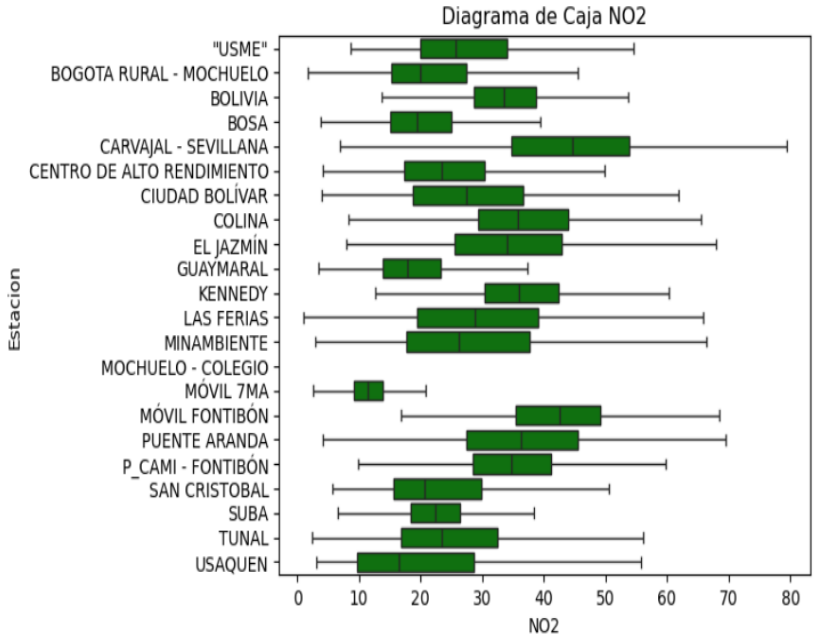


Figura 18

Diagrama de Caja por Estación para SO2

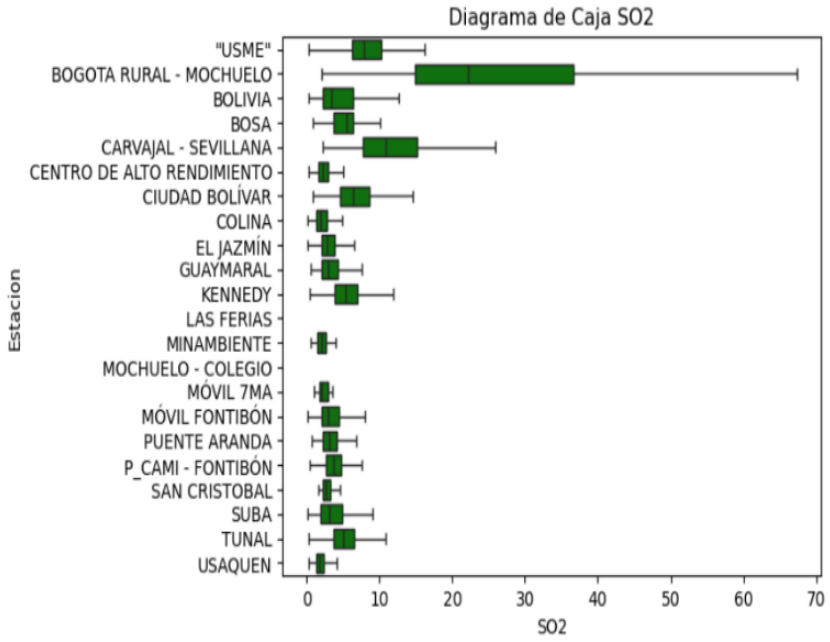


Figura 19

Diagrama de Caja por Estación para O3

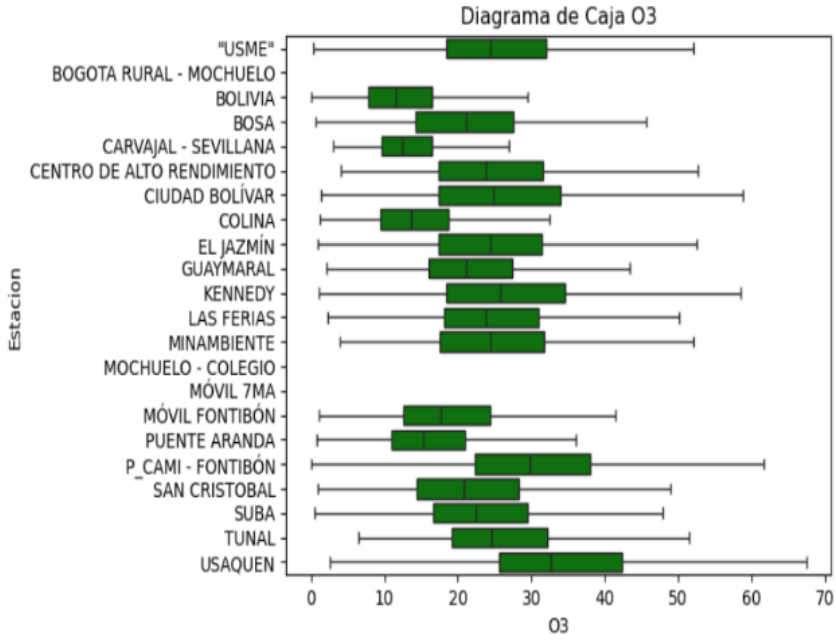


Figura 20

Diagrama de Caja por Estación para NO

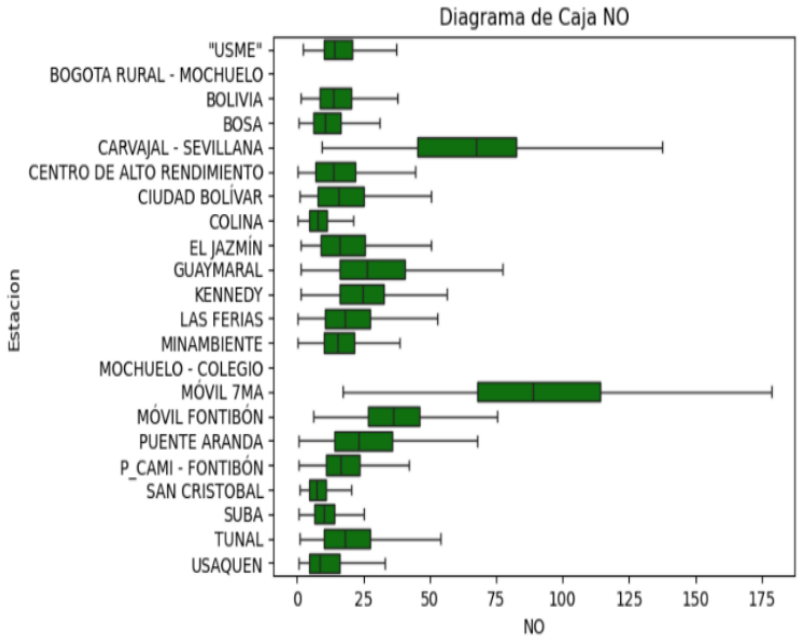
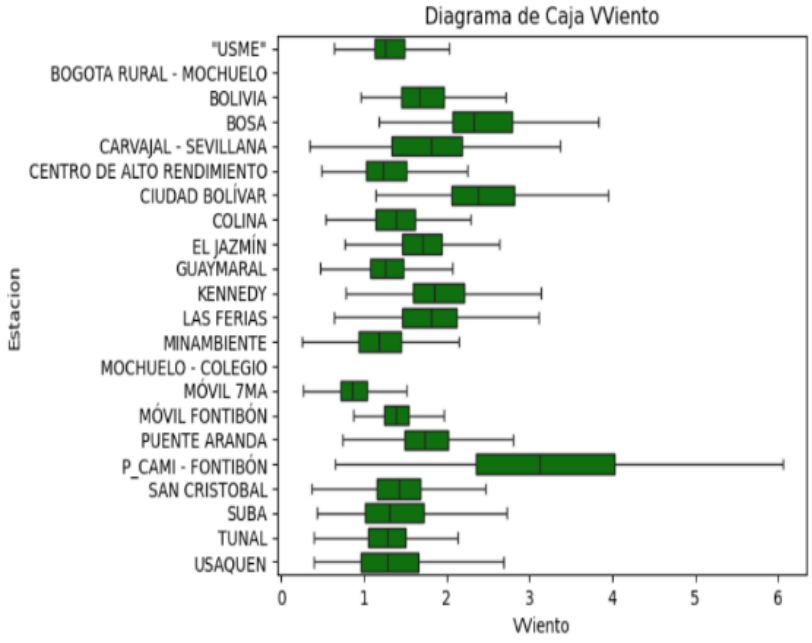


Figura 21

Diagrama de Caja por Estación para Velocidad del Aire



Análisis por Localidades

Separando y visualizando los datos por estación, se nota que no en todas las localidades se mantienen a cabal los niveles recomendados, sino que hay excepciones. Para empezar, se observa en las figuras 15 y 16 que para las estaciones de la zona rural de Bogotá (Mochuelo), y para la estación ubicada en el centro de alto rendimiento (Teusaquillo) la distribución de ambos PM_{10} y $PM_{2.5}$ supera por mucho los limites recomendados, con niveles entre 45 y 60 μgm^{-3} para PM_{10} (estándar de 45 μgm^{-3}) y con niveles entre 20 y 35 μgm^{-3} para $PM_{2.5}$ (estándar 15 μgm^{-3}). También en las estaciones del Tunal (Tunjuelito), Móvil Fontibón (Fontibón), Kennedy, Ciudad Bolívar, Carvajal – Sevillana (Kennedy), y Bolivia se presentan valores superiores para el límite de $PM_{2.5}$ (entre 15 y 30 μgm^{-3}).

Adicionalmente, de la figura 17 se concluye que la mayoría de las localidades se presentan niveles superiores al límite para el caso del NO_2 : Bolivia (Engativá), Carvajal – Sevillana (Kennedy), Colina (Suba), Ciudad Bolívar, El Jazmín (Puente Aranda), Kennedy, Las Ferias (Engativá), Móvil Fontibón (Fontibón), Puente Aranda, P_CAMI – Fontibón (Fontibón) ; todas mostrando medianas en el rango de 29 – 45 μgm^{-3}. Concretamente, todas las localidades están ubicadas en el área noroeste de Bogotá, mostrando que es un problema generalizado en esa área.

Por otra parte, otras diferencias notables se encuentran en la figura 14, donde los niveles de concentración del CO son bastante similares en todas las localidades (alrededor de las 500 a 1000 μgm^{-3}), excepto en las de Carvajal-Sevillana (Kennedy) y Móvil 7ma (Chapinero) donde son significativamente mayores (entre 1500 y 1200 μgm^{-3}). Sin embargo, en todos los casos se está bastante por debajo de la concentración limite recomendado (4000 μgm^{-3}).

En la figura 18, se observa que para la estación de Bogotá Rural – Mochuelo (Usme) hay concentraciones de SO_2 significativamente mayores, con la caja alcanzando niveles cercanos a limite recomendado.

De los niveles de O_3 vistos en la figura 19 no se pueden observar alguna diferencia significativa, todas las localidades parecen tener niveles similares (entre 20 y 35 $\mu g m^{-3}$), solamente los valores extremos superando o acercándose al nivel limite (60 $\mu g m^{-3}$).

De la figura 20, las estaciones de Carvajal – Sevillana (Kennedy) y Móvil 7ma (Chapinero) presentan niveles altos del contaminante NO, en el rango de 75 – 125 $\mu g m^{-3}$. No se está violando ningún límite especifico en este caso, pero se debe tener en cuenta que el NO es un posible precursor de para la generación del NO_2 , por lo que se deberían buscar opciones para disminuir sus niveles en el aire.

La velocidad del aire (figura 21) es más grande en las estaciones de P_CAMI Fontibón, Ciudad Bolívar y Bosa (mayor a 3 m/s). En estas localidades, y en las aledañas se esperaría que los contaminantes se difundieran más fácilmente. Estas locaciones se encuentran en la parte noroeste de la ciudad de Bogotá, indicando que en esas áreas hay una mayor circulación del aire en comparación con las demás localidades.

Análisis Estadístico de la Interacción entre Contaminantes Atmosféricos

Se realizo una prueba de correlación de Pearson entre cada par de variables en el dataset, usando el paquete de scipy. De los resultados de la sección *Minería de Datos* se puede observar que las variables no tienen una distribución estrictamente normal; por lo que se transformaron previamente mediante una estandarización para los cálculos. Por otra parte, se usó un mapa de calor para la mejor visualización de los coeficientes resultantes, que se puede ver en la **figura 22**.

También se probó la fuerza estadística de la existencia de correlación mediante una prueba de hipótesis T-Student, que también se presentan en un mapa de calor en la **figura 23**.

El coeficiente de correlación entre variables solo muestra una correlación lineal fuerte para la relación PM_{10} y $PM_{2.5}$. El resto de los pares muestra coeficientes en el rango 0.45 – 0.65, indicando correlaciones lineales con fuerza media. Todos los pares con ozono (O_3) son débiles y cercanos a cero, indicando que casi no hay correlación.

Por otra parte, se calculó el valor P para cada uno de los coeficientes, con el propósito de medir la significancia estadística de los resultados, mediante pruebas de hipótesis de tipo T-Student. Los valores P se pueden observar en el mapa de calor de la figura 23; todos siendo prácticamente cero y menor al umbral estándar de 0.05; lo que demuestra que los valores hallados probablemente no se deben al azar y son significativos.

Figura 22

Mapa de Calor con los Valores del Coeficiente de Pearson

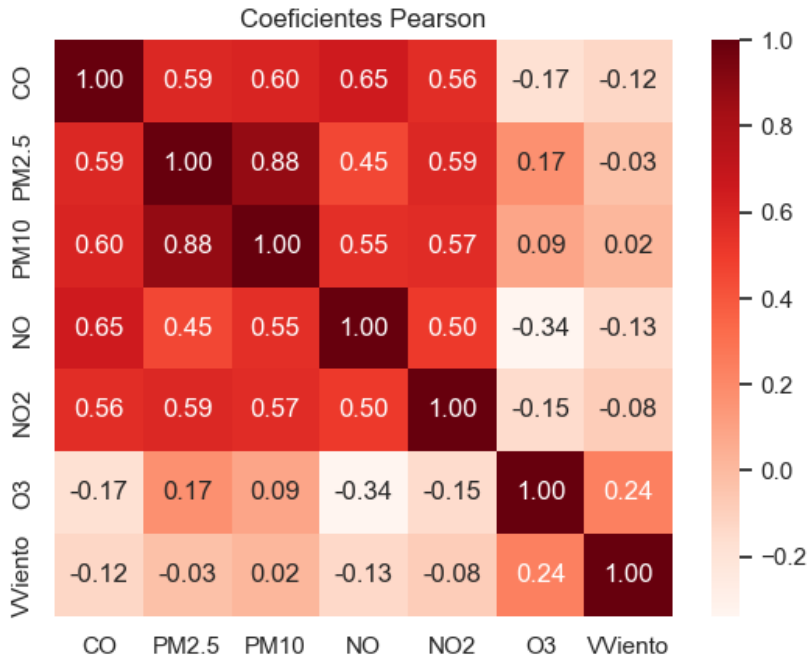
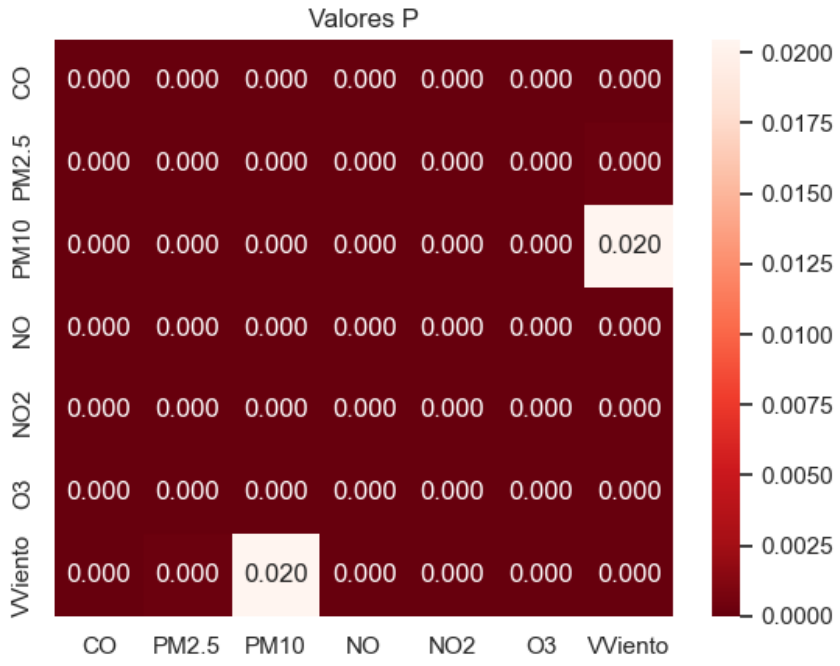


Figura 23

Valores P para las Pruebas T-student entre Variables



Los resultados de las pruebas de hipótesis entre variables muestran que las correlaciones más significativas se dan entre las variables de PM_{10} y $PM_{2.5}$ (r = 0.88). Esto se debe a la naturaleza física de las partículas, ya que la concentración PM_{10}, por definición, de contener la concentración $PM_{2.5}$; por lo que la correlación no sería del todo inesperada o ilógica. Por el contrario, el O_3 no muestra una relación fuerte con ninguna de las variables (entre -.17 y 0.24), y se encontraron relaciones de fuerza media entre pares de las demás variables (entre 0.45 y 0.65).

Posiblemente, las correlaciones de fuerza media significan que ambos contaminantes provienen de fuentes similares, o las concentraciones están relacionadas por medio de alguna reacción química, como en el caso del NO y NO_2 , ya que el monóxido es recursos del dióxido, y el dióxido puede degradarse a otras formas del óxido de nitrógeno. La falta de correlación del

ozono con alguna otra variable podría deberse a su falta de reactividad, o a que su fuente no produce algún otro de los contaminantes observados.

Conclusiones

Se llevó a cabo una revisión de la literatura con el propósito de identificar las variables pertinentes para el análisis exploratorio de datos. Como resultado de este proceso, Se seleccionaron 8 variables para su estudio: CO, PM_{10}, $PM_{2.5}$, NO, NO_2 O_3, SO_2 , velocidad del viento (Vviento). Asimismo, se descartaron once variables que fueron consideradas no significativas para los objetivos del estudio o que presentaban una calidad de datos insuficiente, lo cual impedía su utilización.

Los resultados del análisis muestran que, en términos generales, la ciudad de Bogotá cumple con la mayoría de los estándares recomendados por la OMS. No obstante, se identificaron concentraciones que superan o se aproximan a los límites establecidos para contaminantes críticos como el dióxido de nitrógeno (29.32 NO_2 sobre 25 μgm^{-3}) y el material particulado (32.5 PM_{10} se acerca a 45 μgm^{-3}y 15.9 $PM_{2.5}$ sobre 15 μgm^{-3}); especialmente en determinadas localidades del norte de la ciudad, como Usaquén, Teusaquillo, Kennedy, y Fontibón. Estas zonas coinciden con áreas de alta densidad urbana y mayor estrato, lo que sugiere la influencia de factores como el tráfico vehicular, el uso intensivo de fuentes de energía doméstica y posibles dinámicas propias del desarrollo urbano. Este hallazgo destaca la importancia de considerar el contexto territorial y la importancia de formular políticas públicas diferenciadas según las condiciones de cada sector.

Mediante la aplicación de pruebas de hipótesis, se identificó una correlación positiva fuerte (r = 0.88) entre las concentraciones de material particulado PM_{10} y $PM_{2.5}$, Por otro lado, las asociaciones observadas entre el resto de las variables mostraron magnitudes débiles a moderadas, con coeficientes de correlación que oscilaron entre 0.45 y 0.65. Estos resultados sugieren la necesidad de complementar el análisis estadístico con datos más detallados y

específicos sobre las fuentes de emisión de contaminantes, para identificar posibles correlaciones adicionales significativas. La integración de esta información permitiría un entendimiento más profundo de las dinámicas contaminantes, y fortalecería la precisión en la formulación de estrategias de mitigación ambiental.

Se elaboraron tres recomendaciones derivadas directamente del análisis de los datos de calidad del aire en Bogotá. A partir de los hallazgos, se establecieron propuestas concretas orientadas a la expansión y modernización de la red de monitoreo; la elaboración de políticas diferenciales por densidad y estrato; y la necesidad de investigar y ajustar regulaciones relacionadas con fuentes domésticas y agrícolas; Así, las recomendaciones presentadas no solo responden a las situaciones detectadas en el análisis, sino que también buscan fortalecer las capacidades de gestión ambiental para promover un desarrollo urbano más sostenible.

Recomendaciones

Los datos analizados evidencian que, si bien Bogotá ha mostrado una tendencia a la baja en algunos indicadores clave de contaminación (CO, O_3, SO_2), todavía persisten niveles preocupantes en las concentraciones de material particulado (45 μgm^{-3} PM_{10} y 15 μgm^{-3} $PM_{2.5}$); como se evidencia en las localidades de Kennedy, Fontibón, Ciudad Bolívar, Teusaquillo, Chapinero y Mochuelo; donde se ven las concentraciones de 45 a 60 μgm^{-3} para PM_{10} ; y entre 15 a 30 μgm^{-3} $PM_{2.5}$; que continua siendo iguales o superiores a los límites. La contaminación en estas zonas parece estar relacionada su alta densidad urbana y mayor nivel socioeconómico.

Ante este panorama, y como parte de las acciones recomendadas en el marco del objetivo 4, se recomienda el fortalecimiento y modernización de la red de monitoreo en estas localidades, con el propósito de comprender en mayor detalle las causas de las variaciones locales. Adicionalmente, se propone que se desarrollen e implementen políticas diferenciales respecto al estrato y densidad urbana, con el propósito de mitigar las emisiones en las zonas donde se requiere: se pueden construir zonas verdes; o implementar zonas de bajas emisiones y crear o fortalecer controles vehiculares. según el contexto territorial. Se esta recomendado tener en cuenta el estrato debido a la historia de Bogotá de usar el estrato como herramienta de desarrollo urbano, como se mostró en la sección *Caracterización de las Localidades*.

De igual forma, se destacan como problemáticos los niveles significativos de dióxido de nitrógeno (NO_2), que en estaciones como Carvajal–Sevillana (Kennedy) superan en el 75 % de los registros el umbral de 25 µg/m³ establecido por la OMS. Las fuentes de este contaminante se asocian principalmente a procesos de combustión en estufas de gas domésticas, actividades agrícolas y emisiones del parque automotor. Se recomienda promover la transición hacia

sistemas de cocción más limpios en zonas residenciales, como las estufas eléctricas, y desarrollar campañas de educación ambiental sobre prácticas domésticas sostenibles y revisar las rutas de tráfico vehicular cercanas a estaciones críticas. Por otra parte, dado el papel del NO como precursor del NO_2; y los niveles elevados (50 a 125 $\mu g m^{-3}$) que se encontraron en estaciones localidades como Kennedy y Chapinero; se sugiere investigar la dinámica de propagación del NO (como las fuentes principales, rutas químicas concretas en el aire), con el fin de proponer acciones preventivas frente al NO_2.

Respecto a las diferencias en la velocidad del aire que se encontraron, principalmente en las zonas de Fontibón, Bosa y Ciudad Bolívar donde son significativamente más altas (2 a 4 ms^{-1}, cuando el promedio global es 1.6 ms^{-1}); este conocimiento puede ser estudiado, y aprovechado con el propósito de construir zonas industriales donde los contaminantes no tengan riesgo de acumularse en niveles peligrosos. También puede usarse la información para la búsqueda de terrenos adecuados, y la construcción de parques, zonas verdes o de corredores ambientales; como una estrategia de mitigación del problema de contaminación.

Referencias

Albarracín, K., Consuegra, A., & Aguilar-Arias, J. (2023). Particulate matter 10 µm (PM10), 2.5 µm (PM2.5) datasets gathered by direct measurement, low-cost sensor and by public air quality stations in Fontibón, Bogotá D.C., Colombia. *Data in Brief*, *49*, 109323. https://doi.org/10.1016/j.dib.2023.109323

Alcaldia de Bogota. (2025). *Mapa de Bogota por Localidades*. http://www.sumapaz.gov.co/mi-localidad/mapas

Bouza, E., Vargas, F., Alcázar, B., Álvarez, T., Asensio, Á., Cruceta, G., Gracia, D., Guinea, J., Angel, M., Linares, C., Muñoz, P., Olier, E., Pastor, P., Luisa, M., Querol, X., Tovar, J., Urrutia, I., Villar, F., & Palomo, E. (2022). Air pollution and health prevention: A document of reflection. *Revista Española de Quimioterapia*. https://doi.org/10.37201/req/171.2021

Cadavid-Giraldo, N., Zuelima, A., Guarín, O., & Jimenez, H. (2017). Air Quality in Aburrá Valley: What Can We Expect from the Smart City. *Ingeniería Y Ciencia*, *17*(33), 185–222. https://doi.org/10.17230/ingciencia.17.33.8

CONPES. (2018). *Política para el Mejoramiento De la Calidad del Aire*. https://www.minambiente.gov.co/wp-content/uploads/2021/08/conpes-3943-de-2018.pdf

Duque-Duque, N., & Molina Caro, R. M. (2021). Características de las localidades de Bogotá en contraste con las unidades administrativas territoriales de otras cinco grandes ciudades. *Cuadernos Latinoamericanos de Administración*, *17*(33). https://doi.org/10.18270/cuaderlam.v17i33.3613

Gómez, N., & Molina, I. (2021). *Incidencia de la calidad el aire en el desarrollo urbano sostenible. Metodología de pronóstico basado en herramientas de aprendizaje*

automático [Universitat Politècnica de València]. https://doi.org/10.4995/Thesis/10251/168398

Henao, J., Rendón, A., Hernández, S., Giraldo-Ramirez, P., Robledo, V., Posada-Marín, J., Bernal, N., Salazar, J., & Mejía, J. (2021). Differential Effects of the COVID-19 Lockdown and Regional Fire on the Air Quality of Medellín, Colombia. *Atmosphere*. https://doi.org/10.3390/atmos12091137

Martins, J., & Kaspars, O. (2023). Knowledge Discovery Frameworks and Characteristics. *Baltic J. Modern Computing*, *11*(4), 686–702. https://www.bjmc.lu.lv/fileadmin/user_upload/lu_portal/projekti/bjmc/Contents/11_4_08_Jansevskis.pdf

Matplotlib development team. (2025). *Matplotlib API Reference*. https://matplotlib.org/stable/api/index

MinAmbiente. (2025). *SISAIRE-IDEAM*. http://sisaire.ideam.gov.co/ideam-sisaire-web/consultas.xhtml

Mura, I., Franco, J., Bernal, L., Melo, N., Díaz, J., & Akhavan-Tabatabaei, R. (2020). A Decade of Air Quality in Bogotá: A Descriptive Analysis. *Frontiers in Environmental Science*, *8*(65). https://doi.org/https://doi.org/10.3389/fenvs.2020.00065

Pandas Docs. (2024). *pandas.DataFrame.merge*. https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.merge.html

Pandas Docs. (2025). *Pandas API reference*. https://pandas.pydata.org/docs/reference/index.html

Raykov, T., & Marcoulides, G. (2013). *Basic Statistics an Introduction with R*. Roman and Littlefield.

Represa, N. (2020). *Elaboración e implementación de una propuesta metodológica para la evaluación y gestión de la calidad del aire mediante el enfoque de la ciencia de datos* [Universitat Politècnica de València]. https://doi.org/10.4995/Thesis/10251/144645

Ritchie, H., & Roser, M. (2021). Air Pollution. *Our World in Data*. https://ourworldindata.org/air-pollution

Rubio, M. (2019). *Estadisticas con Aplicaciones en R*. Utadeo.

Samuels, M., Witmer, J., & Shaffner, A. (2012). *Fundamentos de Estadística Para Las Ciencias de la Vida* (4th ed.). Pearson.

scikit-learn. (2025). *StandardScaler*. https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

Seabold, S., & Perktold, J. (2010). statsmodels: Econometric and statistical modeling with python. *9th Python in Science Conference*.

The SciPy community. (2025). *Scipy API Reference*. https://docs.scipy.org/doc/scipy/

Vanderplas, J. (2017). *Python Data Science Handbook*. O'Really.

Waskom, M. (2025). *Seaborn API reference*. https://seaborn.pydata.org/api.html#objects-api

WHO. (2021). *WHO global air quality guidelines: particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide*. https://www.who.int/publications/i/item/9789240034228

World Health Organization. (2016). *Health as the pulse of the new urban agenda: United Nations conference on housing and sustainable urban development*. https://iris.who.int/bitstream/handle/10665/250367/9789241511445-eng.pdf

Yunda, J. G. (2019). Densificación y estratificación social en Bogotá: distribución sesgada de la inversión privada. *EURE (Santiago), 45*, 237–257

Apéndices

Apéndice A

Código para Importar Paquetes Necesarios

```
from enum import StrEnum
from pathlib import Path
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from scipy.stats import pearsonr, chisquare, f_oneway
from data_science_tools.stats import medidas_resumen
from data_science_tools.graficos import heatmap
from sklearn.preprocessing import StandardScaler
from sklearn.compose import ColumnTransformer
formato = "{:.3f}"
pd.options.display.float_format = formato.format
```

Apéndice B

Código para la Limpieza de los Datos

Cargar datos a un dataframe, realizar una combinación y eliminar los posibles datos nulos.

```
r_reporte = Path("datasets/reporte_sisaire.csv")

df_reporte = pd.read_csv(r_reporte)


#limpiado de datos, el archivo csv debe estar en el mismo directorio

df_reporte.drop_duplicates(inplace=True)

#Transformar fechas a datetime objects
df_reporte["Fecha inicial"] = pd.to_datetime(df_reporte["Fecha inicial"], for
mat="%Y-%m-%d")

#limpieza hecha para los tres modelos: modelo lineal, logistico, y knn. se el
imina la columna "wip"
#df_reporte = df_crudo.dropna(axis=1)
```

Apéndice C

Código para Creación de Variables Enum

Un mecanismo para definir objetos que contienen los nombres de las variables, y usarlos como aliases en el Código

```
class At(StrEnum):
    est = 'Estacion'
    f = 'Fecha inicial'
    años = "años"

class Otras(StrEnum):
    h10 = 'HAire10'
    hh = 'HAire?h'
    nox = 'NOx'
    so2_tr = 'SO2 TR'
    tempcm = 'TMPR AIR 10CM'
    dirv = 'DViento'
    h2 = 'HAire2'
    prec = 'PLiquida'
    temp10 = 'TAire10'
    temp2 = 'TAire2'
    rad = 'RGlobal'
    p = 'P'

class Var(StrEnum):
    co = 'CO'
    no = 'NO'
    no2 = 'NO2'
    o3 = 'O3'
    pm10 = 'PM10'
    pm2= 'PM2.5'
    so2 = 'SO2'
    velv = 'VViento'
```

Dividir Datos por rango de años, y obtener conteo de datos no nulos

```
fecha_max = pd.to_datetime("2025-03-31",  format="%Y-%m-%d")
fecha_antes5 = fecha_max - pd.DateOffset(years=5)
fecha_antes10 = fecha_max - pd.DateOffset(years=10)
fecha_antes15 = fecha_max - pd.DateOffset(years=15)

df_rep5 = df_reporte[df_reporte[At.f] >= fecha_antes5]
df_rep10 = df_reporte[df_reporte[At.f] >= fecha_antes10]
df_rep15 = df_reporte[df_reporte[At.f] >= fecha_antes15]

def obtener_prop_validos(df, año):
    conteo_no_nulos = df.notna().sum()

    conteo_fracc = conteo_no_nulos[2:]/conteo_no_nulos[At.est] * 100

    conteo_fracc = pd.concat([conteo_fracc], axis=1)

    conteo_fracc.reset_index(inplace=True)

    conteo_fracc.rename(columns = {0:"porcentaje", "index":"variable"}, inpla
ce=True)
```

```
    conteo_fracc["años"] = f"2025 - {2025 - año}"

    return conteo_fracc


df_por_intervalo = (df_rep5, df_rep10, df_rep15)
años = (5, 10, 15)

array_conteo = [obtener_prop_validos(df, año) for df, año in zip(df_por_inter
valo, años)]

conteo_fracc = pd.concat(array_conteo)

ejes_comp_años = sns.catplot(data=conteo_fracc, x="variable", y="porcentaje",
col="años", kind="bar", hue="variable", legend="full")
ejes_comp_años.set(xticks=[])
plt.show()
```

Versión Final de los datos a 5 años, el dataframe datos contiene los datos sin index por años, el dataframe datos_año contiene indices por años para poder acceder a los datos como si fueran agrupados por años

```
df_rep5[At.años] =  df_rep5[At.f].apply(lambda x: x.year)
variables_eliminar = [var for var in Otras]

datos = df_rep5.drop(variables_eliminar, axis=1)
datos_año = datos.set_index([At.años])
```

Eliminar los registros Negativos

```
variables = [var for var in Var]
datos = datos[(datos[variables] >= 0).all(axis=1)]
```

Eliminar los registros que tienen valores de PM inconsistentes

```
datos = datos[datos[Var.pm10] >= datos[Var.pm2]]
```

análisis Unidimensional

Medidas de Resumen del dataset años 2020 - 2025

```
resum = [medidas_resumen(datos, col) for col in Var]
resumen = pd.concat(resum, axis=1)
resumen
```

Medidas individuales para cada año

```
resum = [medidas_resumen(datos_año.loc[2020], col) for col in Var]
resumen = pd.concat(resum, axis=1)
resumen
```

año 2021

```
resum = [medidas_resumen(datos_año.loc[2021], col) for col in Var]
resumen = pd.concat(resum, axis=1)
resumen
```

año 2022

```
resum = [medidas_resumen(datos_año.loc[2022], col) for col in Var]
resumen = pd.concat(resum, axis=1)
resumen
```

año 2023

```
resum = [medidas_resumen(datos_año.loc[2023], col) for col in Var]
resumen = pd.concat(resum, axis=1)
resumen
```

año 2024

```
resum = [medidas_resumen(datos_año.loc[2024], col) for col in Var]
resumen = pd.concat(resum, axis=1)
resumen
```

año 2025

```
resum = [medidas_resumen(datos_año.loc[2025], col) for col in Var]
resumen = pd.concat(resum, axis=1)
resumen
```

Apéndice D

Cálculo de la Curtosis para las Variables

```
datos[[var for var in Var]].kurt()
```

Apéndice E

Código para los Gráficos de Caja

Gráficos de caja para cada Estación

```
for var in Var:
    ax = sns.boxplot(data=datos, x=var, y=At.est,  orient="h", color="g",  sh
owfliers=False )
    ax.set_title(f"Diagrama de Caja {var}")
    plt.show()
```

Graficos de Caja para las variables en general (Sin Estandarizar)

```
sns.set_theme(style="darkgrid")

fig1, axs1 = plt.subplots(4, 2, figsize=(10, 14))
colores = ("skyblue", "olive", "gold", "teal", "skyblue", "olive", "gold", "t
eal")

def crear_ejes(dim1, dim2):
    for i in range(dim1):
        for j in range(dim2):
            yield (i, j)

for var, color,  coord in zip (Var, colores, crear_ejes(4, 2)):
    eje = axs1[*coord]
    sns.boxplot(data=datos, x=var,  orient="h", color=color, showfliers=False
, ax=eje )


fig1.suptitle("Diagramas de Caja de Contaminantes", y=0.9)
plt.show()
```

Apéndice F

Código para los Histogramas

Histogramas para las variables en general, Sin Estandarizar

```
sns.set_theme(style="darkgrid")

fig2, axs2 = plt.subplots(4, 2, figsize=(10, 14))

colores = ("skyblue", "olive", "gold", "teal", "skyblue", "olive", "gold", "t
eal")

def crear_ejes(dim1, dim2):
    for i in range(dim1):
        for j in range(dim2):
            yield (i, j)

for var, color, coord in zip (Var, colores, crear_ejes(4, 2)):
    eje = axs2[*coord]
    sns.histplot(data=datos, x=var, kde=True, color=color, ax=eje)

fig2.suptitle("Histogramas de los Contaminantes", y=0.9)
plt.show()
```

Apéndice G

Código para el Análisis Bidimensional e Inferencial.

Diagramas de Dispersion por Pares

```
sns.pairplot(datos[[var for var in Var]])
plt.show()
```

Estandarización de las variables: Para la comparación de la variabilidad de las variables, se realizó una estandarización de las siguientes variables: *cantidad de personas con acceso internet*, *población*, *índice*: Esto por medio de la clase `StandardScaler`. Después se realizó los diagramas de caja respectivos para comparar la distribución de las variables numéricas.

```
# usar columnstransformer para trasnformar las columnas necesarias unicamente
datos.columns = datos.columns.astype(str)

variables = [var.value for var in Var]

prep = ColumnTransformer([
    ("escalamiento", StandardScaler(), variables ),
    ],
    remainder='drop' )

display(prep)

#construir el dataframe necesario
datos_std = pd.DataFrame(prep.fit_transform(datos[variables]), columns=variab
les )

ColumnTransformer(transformers=[('escalamiento', StandardScaler(),
                                 ['CO', 'NO', 'NO2', 'O3', 'PM10', 'PM2.5',
                                  'SO2', 'VViento'])])
```

Graficos de Caja para las variables en general (Estandarizado)

```
fig3, axs3 = plt.subplots(4, 2, figsize=(10, 14))

colores = ("skyblue", "olive", "gold", "teal", "skyblue", "olive", "gold", "t
eal")

for var, color, coord in zip (Var, colores, crear_ejes(4, 2)):
    eje = axs3[*coord]

    sns.boxplot(data=datos_std, x=var.value, color=color, showfliers=False, a
x=eje)


fig3.suptitle("Comparación Variables Estandarizadas", y=0.9)
plt.show()
```

Apéndice H

Código para los Coeficientes de Pearson, Pruebas de Hipótesis, Mapas de Calor

El siguiente fragmento de Código calcula los coeficientes de correlación de pearson, y los valores *P* de la prueba de correlación asociada al mismo tiempo. Se hacen pares de las variables numéricas en estudio secuencialmente, y toda la información se consigna en dos DataFrames: uno llamado `corr` que contiene los coeficientes, y otro llamado `valores_p` que contiene los valores p de las pruebas estadísticas.

```
variables = datos_std[[Var.co, Var.pm2, Var.pm10, Var.no, Var.no2, Var.o3, Va
r.velv]].dropna()


corr = pd.DataFrame(index=variables.columns, columns=variables.columns)

valores_p = corr.copy()

for variable_1 in variables.columns:
    coefs = []
    p_valores = []

    for variable_2 in variables.columns:
        coef, p_valor = pearsonr(variables[variable_1], variables[variable_2]
)
        coefs.append(coef)
        p_valores.append(p_valor)

    corr[variable_1] = coefs
    valores_p[variable_1] = p_valores
```

Mapa de Calor Coeficientes de Correlación

```
heatmap(corr, "Coeficientes Pearson")
plt.show()
```

Mapa de Calor Valores P de las pruebas estadísticas de correlación

```
heatmap(valores_p, "Valores P", formato=".3f", mapa="Reds_r")
plt.show()
```