# "Análisis de la calidad del aire en España, identificando patrones para mejorar salud pública por zonas/localidades"

Ángel Ronaldo Perez Diaz, Juan José Rincón
Estudiantes Ingeniería De Sistemas Y Computación UPTC.
Colombia.

## *Sobre Los Autores.*

**Ronaldo Perez Diaz:** Egresado del colegio técnico industrial Gustavo Jiménez, en la especialidad de mecánica industrial, actualmente estudiante de octavo semestre de ingeniería de sistemas y computación en la UPTC, seccional Sogamoso.
Correo: angel.perez01@uptc.edu.co

**Juan José Rincón:** Estudiante ingeniería de sistemas y computación, UPTC, seccional Sogamoso.
Correo: juan.rincon12@uptc.edu.co

## *Resumen.*

Este estudio tiene como objetivo analizar la calidad del aire en distintas zonas y localidades de España, identificando patrones que podrían ayudar al país a prevenir y tomar decisiones para mejorar la salud pública. Se plantearon y examinaron cinco hipótesis, incluyendo las concentraciones de gases y partículas en diferentes momentos y lugares. Se hizo uso de una metodología de análisis de datos, la cual permitió evaluar las mediciones de calidad del aire proporcionadas por las autoridades. Los resultados mostraron que los mayores niveles de concentración del gas CO se presentaron en el mes de octubre, también, se encontró que las partículas PM25 presentan mayores niveles de concentración en el sector Zona Norte Polideportivo y que el gas NO2 tiene mayores concentraciones después de las seis de la tarde. También se observó que los niveles del gas O3 son más bajos en el mes de diciembre, que en el mes de octubre. En conclusión, el análisis de los resultados facilitados por las autoridades de España, podría ayudarles a tomar medidas preventivas que permitan mejorar la calidad del aire y de esta forma proteger la salud pública en las zonas más afectadas.

## *Palabras Claves:*

Calidad del aire, contaminación atmosférica, concentración de gases, salud pública, patrones de comportamiento, impacto ambiental, monitoreo ambiental.

# "Analysis of air quality in Spain, identifying patterns to improve public health by zones/localities."


## *Abstract*

*This study aims to analyze air quality in different areas and locations in Spain, identifying patterns that could help the country to prevent and make decisions to improve public health. Five hypotheses were proposed and examined, including gas and particle concentrations at different times and locations. A data analysis methodology was used to evaluate air quality measurements provided by the authorities. The results showed that the highest concentration levels of CO gas occurred in the month of October, also, it was found that PM25 particles have higher concentration levels in the North Zone Sports Center sector and that NO2 gas has higher concentrations after six o'clock in the afternoon. It was also observed that O3 gas levels are lower in December than in October. In conclusion, the analysis of the results provided by the Spanish authorities could help them to take preventive measures to improve air quality and thus protect public health in the most affected areas.*



## *Keywords:*

*Air quality, atmospheric pollution, gas concentration, public health, behavioral patterns, environmental impact, environmental monitoring.*


## *Introducción:*

La calidad del aire es un tema que ha adquirido gran relevancia en las últimas décadas debido a su impacto en la salud pública y el medio ambiente a lo largo y ancho del planeta. En España, la contaminación atmosférica es uno de los principales problemas ambientales, especialmente en áreas urbanas y metropolitanas según estudios e información recolectada por las autoridades. La exposición a partículas y gases contaminantes puede tener efectos negativos en la salud, como enfermedades respiratorias, cardiovasculares y en los peores casos, llegar a producir cáncer.

A pesar de los esfuerzos realizados por las autoridades para controlar la contaminación, aún existen áreas y localidades en España donde los niveles de contaminación son elevados. Por esta razón, resulta necesario llevar a cabo estudios que permitan conocer la calidad del aire, y establecer medidas preventivas para mejorar la salud pública, estableciendo regulaciones a las grandes fábricas para prohibir el desecho de agentes contaminantes en grandes medidas, un ejemplo claro de esto es el presentado en 2020 por la Comisión Europea, el cual se trazó un nuevo objetivo climático, proponiendo reducir las emisiones de gases de efecto invernadero en al menos un 55% para 2030, este objetivo incluye que se prohibiría la venta de nuevos vehículos a gasolina y diésel para 2035, sin embargo, este objetivo aún no se ha adoptado formalmente como política de la UE, se espera que se discuta en las próximas negociaciones del Parlamento Europeo y el Consejo de la UE sobre la Ley de Clima de la UE.

En este sentido, el objetivo del presente estudio es analizar la calidad del aire en distintas zonas y localidades de España, esto gracias al conjunto de datos proporcionado por el tutor, el cual permitió identificar patrones que podrían ayudar al país a prevenir y tomar decisiones para mejorar la salud pública de sus habitantes. El estudio se basó en la metodología de análisis de datos, la cual permitió evaluar las mediciones de calidad del aire proporcionadas por las autoridades en el conjunto de datos; los resultados obtenidos podrían ser de gran utilidad para las autoridades y la sociedad en general, ya que podrían ayudarles a tomar medidas preventivas que permitan mejorar la calidad del aire, como, por ejemplo, vivir en zonas menos contaminadas, hacer uso de tapabocas en ciertos horarios, entre otras; y de esta forma, proteger la salud pública de la población en las zonas más afectadas.

## *Metodología:*

La metodología del estudio se basó en el análisis de datos proporcionados por las autoridades en el conjunto de datos sobre mediciones de calidad del aire en distintas zonas y localidades de España. Para ello, se llevó a cabo un proceso de limpieza y transformación de los datos con el objetivo de obtener información útil, limpia y coherente. Posteriormente, se aplicaron técnicas estadísticas y de visualización de datos para analizar los patrones y tendencias en la calidad del aire en las diferentes zonas y localidades de España.

En concreto, se utilizaron técnicas de estadística descriptiva, haciendo uso de las más comunes, como: la media, la mediana, y en específico los coeficientes de correlación de Pearson (r), Spearman (rho), Kendall (tau), obteniendo de esta forma medidas de tendencia central, dispersión y forma en los datos de calidad del aire. Además, se utilizaron gráficos para visualizar los niveles de contaminación en las distintas zonas y localidades de España. Finalmente, con los datos obtenidos se pudo llevar a cabo un análisis estadístico para evaluar las posibles relaciones entre los niveles de contaminación y otros factores como los industriales, socioeconómicos, el tráfico vehicular, entre otros.

En resumen, la metodología utilizada en este estudio fue el análisis de datos mediante técnicas estadísticas, limpieza de datos, y comparación de variables según las hipótesis planteadas, para así evaluar la calidad del aire en distintas zonas y localidades de España, lo que permitió identificar patrones y tendencias que podrían ayudar a las autoridades a tomar medidas preventivas para mejorar la salud pública de la población.

## *Desarrollo.*

## *Estado del arte.*

**ARTÍCULO 5**

| | |
|---|---|
| **TÍTULO** | Inteligencia ambiental (Una Estrategia de Ciencia de Datos al Servicio del Medio Ambiente y las Personas) |
| **AUTOR** | Superintendencia del Medio Ambiente |
| **FECHA** | 2021 |
| **URL** | https://portal.sma.gob.cl/publicaciones/IA/Inteligencia_Ambiental_SMA.pdf |
| **DESCRIPCION** | El objetivo de la Superintendencia del Medio Ambiente (SMA), como principal órgano del estado encargado de asegurar el cumplimiento de la normativa ambiental, es la protección de la salud de las personas, y de la diversidad de ecosistemas que conforman el invaluable patrimonio natural nacional. Para ejercer estas funciones la SMA debe lidiar con múltiples normativas ambientales y sus especificidades, tener una cobertura en todo el territorio nacional, así como dar solución a permanentes contingencias que requiere de una reacción rápida y efectiva para mitigar efectos ambientales indeseados. En este sentido, la presente Estrategia de Inteligencia Ambiental tiene como objetivo mejorar la gestión institucional de la Superintendencia del Medio Ambiente, a través del uso intensivo de tecnologías, la ciencia de datos y la inteligencia artificial, de modo de contribuir a robustecer la fiscalización, fortalecer la potestad sancionatoria y el incentivo al cumplimiento; potenciar el seguimiento y el acceso a información pública en materia ambiental. |
| **APORTE** | La ciencia de datos tiene usos fundamentales los cuales mejoraran el medio ambiente y el diario vivir del ser humano gracias a estrategias de decisiones que van de la mano con entidades del gobierno tales como la superintendencia del medio ambiente en chile. incentiva proyectos que capturan y analizan datos, esto es de gran ayuda ya que todos estos proyectos generan una big-data de datos la cual se puede utilizar en casi cualquier modelo ambiental aplicado a la ciencia de datos. |

| **ARTÍCULO 6** | |
|---|---|
| **TÍTULO** | **Impacto de la aplicación de algoritmos de minería de datos en variables de contaminación del aire** |
| **AUTOR** | John Javier, Ortega Guamán, Orellana Cordero, Marcos Patricio |
| **FECHA** | 2018 |
| **URI** | http://dspace.uazuay.edu.ec/handle/datos/78009/11887 |

| | |
|---|---|
| **DESCRIPCIÓN** | Este informe contiene enfoques sobre el descubrimiento de lo que sus autores consideran “el mejor algoritmo de Minería de Datos para el análisis de las variables de contaminación del aire”; cuyo análisis fue posible gracias a la recopilación de datos mediante el uso de estaciones de medición de calidad del aire.<br>Después de esto, realizan el análisis de 5 algoritmos no descriptivos de minería de datos, esto, con el fin de identificar, cuál es el mejor para procesar las variables de contaminación del aire e identificar patrones de comportamiento. Se evaluaron 5 variables ambientales, las cuales son los principales generadores de contaminación: Ozono (O3), Monóxido de Carbono (CO), Dióxido de Azufre (SO2), Dióxido de Nitrógeno (NO2) y Material Particulado 2,5um (PM2.5), estos análisis se llevaron a cabo con distintos patrones de comportamiento, y en diferentes etapas del día.<br>De este modo, y aplicando la metodología CRISP-DM, ya que proporciona pautas que permiten alcanzar los resultados de manera eficiente y precisa, se logró obtener los resultados preliminares de todos los algoritmos evaluados, logrando identificar cuál era el mejor algoritmo de minería de datos que permite analizar las variables de contaminación, y concluyendo, que el agente que más influye en contaminación sobre el resto de las variables es el O3. |
| **APORTE** | Es importante a la hora de realizar análisis de datos, escoger las herramientas optimas que permitan arrojar un análisis real, coherente y que permita cumplir los objetivos de la investigación que se lleve a cabo.<br>En conclusión, se debe realizar una investigación detallada para identificar que algoritmos de minería de datos permiten procesar datos y arrojar resultados dirigidos al proyecto que se requiere, porque como se expone en este documento de investigación, hay unos más eficientes que otros.<br>Por último, y no menos importante, investigaciones como esta nos permite conocer que el análisis de datos es una herramienta optima para mostrarle al mundo la calidad de aire que esta respirando, y cuales son las repercusiones que conlleva a la salud el respirar aire de esa calidad. |

| **ARTÍCULO 7** | |
|---|---|
| **TÍTULO** | **Descubriendo patrones de comportamiento entre contaminantes del aire: Un enfoque de minería de datos** |
| **AUTOR** | Juan Ortega (ua069259@uazuay.edu.ec, Universidad del Azuay, Ecuador)<br>Chester Sellers (csellers@uazuay.edu.ec, Universidad del Azuay, Ecuador)<br>Patricia Ortega (portega@uazuay.edu.ec, Universidad del Azuay, Ecuador)<br>Diana Arce (darce@uazuay.edu.ec, Universidad del Azuay, Ecuador)<br>Fernando Lima (flima@uazuay.edu.ec, Universidad del Azuay, Ecuador)<br>Marcos Orellana (marore@uazuay.edu.ec, Universidad del Azuay, Ecuador) |
| **FECHA** | Recepción: 09 septiembre 2018<br>Aprobación: 29 noviembre 2018 |

| | |
|---|---|
| **URL** | https://ingenieria.ute.edu.ec/enfoqueute/index.php/revista/article/view/411 |
| **DESCRIPCION** | Este artículo, contiene enfoques relacionados a los agentes contaminantes en el aire, y sus afectaciones tanto a la salud humana como al medio ambiente; es por esto, que el aire es monitoreado por estaciones, las cuales generan datos sobre los niveles de concentración de contaminantes del aire en una zona específica, dichos datos son evaluados para medir la calidad del aire, y de esta forma, centrar esfuerzos con el fin de mejorar la calidad de vida en zonas urbanas.<br>Es por esto, que es importante conocer niveles de concentración de los contaminantes, como las asociaciones entre estos; teniendo en cuenta el proceso estándar "Cross-industry para minería de datos", el presente artículo presenta un enfoque que lleva a identificar correlaciones e incidencias entre los contaminantes más nocivos en la Región Andina, estos agentes son: Ozono, Monóxido de carbono, Dióxido de azufre, Dióxido de nitrógeno y Material Particulado.<br>El presente artículo también describe un experimento usando un conjunto de datos de la estación de monitoreo de la ciudad de Cuenca, Ecuador ubicada en la Región Andina. Los resultados muestran que el enfoque propuesto es efectivo para extraer conocimiento útil de apoyo a la evaluación de la calidad del aire en zonas urbanas.<br>Además, este trabajo proporciona un punto de partida para futuras aplicaciones de minería de datos en el contexto de contaminación atmosférica en la Región Andina. |
| **APORTE** | Este articulo contiene estudios importantes que permiten guiar a las alcaldías, gobernaciones, o países para que lo pongan y practica y de esta forma mejoren la calidad de vida, un ejemplo claro donde podría ser aplicada esta investigación es la zona industrial del municipio de Sogamoso, la cual, es una de las zonas con mayores índices de contaminación, esto debido a las industrias presentes.<br>El objetivo de la investigación, es apoyar a los gestores medioambientales y urbanistas en el proceso de toma de decisiones. El enfoque propuesto permite identificar correlaciones y la incidencia entre cinco contaminantes atmosféricos nocivos en la región andina. |

| **ARTÍCULO 8** | |
|---|---|
| **TÍTULO** | **Analítica de datos: incidencia de la contaminación ambiental en la salud pública en Medellín (Colombia)** |
| **AUTOR** | Juan Sebastián Parra Sánchez, (Grupo de Investigación GIDATI-UPB).<br>Ana Isabel Oviedo Carrascal, (Ing. Sistemas. Ph. D. Ingeniería Electrónica, UPB).<br>Ferney Orlando Amaya Fernández, (Ing. Electrónico. Ph. D. Ingeniería. UPB). |
| **FECHA** | 2020-11-01 |
| **URL** | Parra Sánchez, J. S., Oviedo Carrascal, A. I., & Amaya Fernández, F. O. (2020). Analítica de datos: incidencia de la contaminación ambiental en la salud pública en |

| | |
|---|---|
| | Medellín (Colombia). Revista De Salud Pública, 22(6), 609–617. https://doi.org/10.15446/rsap.v22n6.78985 |
| **DESCRIPCION** | Analizar el impacto de la contaminación del aire por material particulado PM2,5 y su relación con el número de asistencias a entidades de salud por enfermedades respiratorias por medio de analítica de datos.<br>Se analizaron datos del Área Metropolitana de Medellín, Colombia, ciudad ubicada en un valle estrecho densamente poblado e industrializado y que ha presentado episodios críticos de contaminación en los últimos años. Se analizaron tres fuentes de datos: datos meteorológicos aportados por el SIATA (Sistema de Alerta Temprana de Medellín y el Valle de Aburrá); datos de contaminación por material particulado PM2,5 aportados por SIATA; y reportes de los RIPS (Registros Individuales de Prestación de Servicios de Salud) aportados por la Secretaría de Salud. |
| **APORTE** | Se evidenció la relación entre la concentración de PM2,5 con las asistencias médicas por los diagnósticos de IRA, EPOC y asma. En un episodio crítico de contaminación por PM2,5, se encontraron los siguientes retardos en la atención médica: entre 0 y 2 días para el IRA, 0 y 7 días para el EPOC y 0 y 5 días para el asma.<br>Se encontraron coeficientes de correlación que evidencian la asociación de la concentración de PM2,5 con las asistencias por los diagnósticos de IRA, EPOC y asma. La mayor correlación entre las tres morbilidades se presentó para el asma. La variable meteorológica de mayor correlación con la variable objetivo es la temperatura del aire para el caso de EPOC y asma. En el caso de IRA, la variable con mayor correlación es la velocidad del viento. Por otro lado, el día de la semana es una variable de gran importancia a la hora de realizar un estudio de atenciones por enfermedades.<br>Para concluir, se evidencia que este articulo demuestra como la tecnología es un gran aliado para detectar y prevenir enfermedades, sobre todo en ciudades como Medellín, y su área metropolitana, dado que, se encuentra rodeada por varias montañas, creando así, una zona en la cual el aire no circula de manera correcta, y no permite “limpiar” toda la contaminación que genera por su industria. |

| **ARTÍCULO 9** | |
|---|---|
| **TÍTULO** | **Fitorremediación: una alternativa para eliminar la contaminación** |
| **AUTOR** | María Santos Pedraza Guevara |
| **FECHA** | 2021-09-01 |
| **URL** | María Santos Pedraza Guevara. (2021). FITORREMEDIACIÓN EN CUERPOS DE AGUA CONTAMINADOS POR METALES PESADOS. Innova Biology Sciences, 1(1), 61–78. https://doi.org/10.58720/ibs.v1i1.6 |

| | |
|---|---|
| **DESCRIPCION** | El presente artículo recopila información de las técnicas de fitorremediación en cuerpos de agua por diferentes especies de plantas, su acumulación de metales pesados, ventajas y desventajas, a partir de diversos estudios de investigación.<br>Se encontró que las plantas acuáticas, para desarrollarse, son capaces de captar metales a través de diversos mecanismos, entre los cuales se encuentran: la Fito volatilización y Fito extracción, efectuadas a través de su parte aérea; la rizo filtración y Fito estabilización, realizadas desde la raíz hacia la parte inferior; la Fito filtración, que es otro mecanismo para tratamiento de aguas residuales por micrófitos y desarraigadas de cultivos hidropónicos que luego son reubicados a corrientes acuosas.<br>Para terminar, la planta "Lantana camara" puede eliminar hasta 88,93 % de Pb, Cyperus alternifolius tiene niveles de absorción de Zn (5 %), Cd (6 %), Al (13 %) y Pb (14 %). La V unguiculada absorbe 72 % de iones de Cu y 92 % de iones de Cd. |
| **APORTE** | Las plantas tienen la capacidad de eliminar iones de nutrientes de metales pesados de cuerpos de agua contaminada; es por esto, que se usa potencialmente la fitorremediación como una técnica verde amigable con el medioambiente.<br>La tasa de crecimiento rápido y la capacidad de tolerancia dependerán de la especie, la temperatura y el pH para la bioacumulación de los metales. la Fito extracción, la Fito degradación es aplicable a suelos, sedimentos de lodos y aguas superficiales, pero esto dependerá de diversos factores, como la temperatura y el pH.<br>La ventaja es el poder de limpieza a bajo coste, no utiliza reactivos químicos peligrosos ni afecta la estructura del agua. Se realiza en el lugar afectado, por ende, es una técnica ecológica amigable con el medioambiente. |

| **ARTÍCULO 10** | |
|---|---|
| **TÍTULO** | **E L IMPACTO DEL BIG DATA EN LA ECOLOGÍA Y CUIDADO DEL MEDIO AMBIENTE EN COLOMBIA** |
| **AUTOR** | Caro Caro Yeferson Norbey |
| **FECHA** | 2021-09-29 |
| **URL** | Universidad Militar Nueva Granada, Informe: https://repository.unimilitar.edu.co |
| **DESCRIPCION** | El movimiento ecológico que ha abandonado el ser humano a través de los años, ha tomado relevancia e importancia, a tal punto, de convertirse en un discurso prioritario para quienes consideran que el cuidado del medio ambiente, es necesario para la supervivencia de los grupos poblacionales, por cuanto, las comunidades ambientales, ecologistas y grandes corporaciones, visualizan su preocupación por la sostenibilidad del medio ambiente, dichas inquietudes se han convertido en puntos claves dentro de las agendas de trabajo de los gobiernos y las organizaciones |

| | |
|---|---|
| | ambientales, mediante la formulación, aplicación y ejecución de políticas públicas dirigidas a la protección del medio ambiente. |
| **APORTE** | Es de importancia recalcar en el presente ensayo el estudio de los impactos que se han generado tanto positivo, como negativamente frente a la entrada de la tecnología en todos los temas ecológicos-entorno, y cómo dichas tecnologías pueden ir entregados resultados donde los principales beneficiarios son los ecosistemas que nos rodean. En este orden de ideas y de acuerdo con el world forum economics, "Desde la llegada de los escenarios de ecosistemas de desarrollo digital, todos los sectores económicos han alcanzado su productividad, pero también, el medio ambiente se benefició en tener trabajos mucho más seguros y ecológicamente protegidos" |

## *Actividades.*

Principalmente se plantearán una serie de tareas que se irán validando en el desarrollo del artículo; principalmente se trazarán una serie de cinco (5) hipótesis:

*Planteamiento De Hipótesis.*

1) En los últimos 7 días de noviembre se presentaron los mayores niveles de concentración del gas CO.
2) El gas NO2 tiene una relación lineal con el gas CO por tal motivo son valores directamente proporcionales entre sí.
3) Las partículas PM25 presenta mayores niveles de concentración en el sector Canto Pinyo.
4) El gas NO2 presenta mayores concentraciones en las horas que van de (6 pm – 11 pm), por tal motivo se dice que su comportamiento se intensifica en las horas pico de la noche.
5) Los niveles del gas O3 son más bajos en el mes de octubre que en el mes de diciembre.

Seguido a esto se realizará un análisis exploratorio y se establecerán las tareas de preprocesamiento de los datos. Todo esto hacia el tipo de dato de cada variable, resumen estadístico, uso de las funciones select () y filter () para validar los valores de los gases contaminantes a partir del tipo de sensor, y se realizaran tareas de transformación de variables teniendo en cuenta el resultado del análisis exploratorio inicial, es importante mencionar que todas estas tareas se realizaran en R Studio.

## *Conceptos a tener en cuenta.*

**R Studio**: es un entorno de desarrollo integrado (IDE) para el lenguaje de programación R, por lo que se considera un lenguaje de programación. R es un lenguaje de programación orientado a objetos y funcional utilizado en estadística y análisis de datos, y R Studio proporciona un ambiente de desarrollo para escribir, depurar y ejecutar código en R.

**Filter** La función filter() en R Studio permite filtrar filas de un data frame que cumplan ciertas condiciones especificadas. La función se encuentra en el paquete dplyr.

**Select ()**: es una de las funciones principales del paquete dplyr de R Studio. Permite seleccionar columnas de un data frame (conjunto de datos) de manera muy fácil y rápida. Con select(), puedes

elegir las columnas que necesitas de tu data frame y dejar de lado las que no necesitas. Esto puede ser útil si tienes un data frame grande con muchas columnas, pero solo necesitas trabajar con algunas de ellas.

A continuación, se presentan algunos conceptos básicos y causas de las partículas más comunes en el conjunto de datos:

**PM10**: son partículas finas en suspensión que tienen un diámetro menor o igual a 10 micrómetros. Estas partículas pueden ser inhaladas por las personas y llegar hasta los pulmones, lo que puede provocar problemas respiratorios y cardiovasculares. Las causas de estas partículas pueden ser variadas, como emisiones de vehículos y fuentes industriales, emisiones de calefacción, polvo de carreteras y construcción, entre otras.

**PM2.5:** son partículas finas en suspensión que tienen un diámetro menor o igual a 2.5 micrómetros. Estas partículas son aún más peligrosas que las PM10 ya que son más pequeñas y pueden penetrar aún más profundamente en los pulmones y en la corriente sanguínea, lo que puede provocar problemas de salud graves. Las causas de estas partículas son similares a las PM10, pero también pueden ser generadas por la combustión de combustibles fósiles en centrales térmicas y otros procesos industriales.

**PM1:** se refiere a las partículas finas en suspensión en el aire que tienen un diámetro aerodinámico menor o igual a 1 micrómetro (µm). Estas partículas son consideradas ultrafinas y pueden ser emitidas por fuentes naturales, como los incendios forestales y las tormentas de polvo, así como por fuentes antropogénicas, como los vehículos, la quema de combustibles fósiles y la industria. Debido a su tamaño extremadamente pequeño, las partículas PM1 pueden penetrar profundamente en los pulmones y en el torrente sanguíneo, lo que puede tener efectos negativos en la salud, como problemas respiratorios, cardiovasculares y otros problemas de salud.

**NO2:** es un gas tóxico que se forma principalmente por la combustión de combustibles fósiles en vehículos y centrales térmicas. La exposición prolongada al NO2 puede provocar problemas respiratorios y aumentar el riesgo de enfermedades cardiovasculares.

**O3:** es un gas que se forma a partir de la reacción de la luz solar con otros contaminantes como los óxidos de nitrógeno y los compuestos orgánicos volátiles. Aunque la capa de ozono es beneficiosa para la protección contra los rayos ultravioleta, el O3 en la atmósfera a nivel del suelo puede provocar problemas respiratorios, especialmente en personas con problemas respiratorios preexistentes.

**SO2:** es un gas que se forma principalmente por la combustión de combustibles fósiles en centrales térmicas y otros procesos industriales. La exposición prolongada al SO2 puede provocar problemas respiratorios y aumentar el riesgo de enfermedades cardiovasculares.

**CO**: es el monóxido de carbono, un gas inodoro e incoloro que se produce por la combustión incompleta de combustibles fósiles como el petróleo, el gas y el carbón. El CO es muy peligroso para la salud humana, ya que puede ser mortal en altas concentraciones. Puede provocar dolores de cabeza, náuseas, vómitos, mareos, debilidad y dificultades respiratorias.

**CO2**: es el dióxido de carbono, un gas inodoro e incoloro que se produce en la respiración de los seres vivos y en la combustión de combustibles fósiles. Es un gas de efecto invernadero que contribuye al calentamiento global y al cambio climático. El CO2 se ha convertido en uno de los principales contaminantes del aire en las grandes ciudades, y su concentración en la atmósfera ha aumentado significativamente en las últimas décadas debido a la actividad humana.

| Valor del coeficiente de correlación | Tipo de correlación |
|---|---|
| -1 | Correlación negativa perfecta |
| -0,9 a -0,99 | Correlación negativa muy fuerte |
| -0,7 a -0,89 | Correlación negativa fuerte |
| -0,4 a -0,69 | Correlación negativa moderada |
| -0,2 a -0,39 | Correlación negativa débil |
| -0,01 a -0,19 | Correlación negativa muy débil |
| 0 | Correlación nula |
| 0,01 a 0,19 | Correlación positiva muy débil |
| 0,2 a 0,39 | Correlación positiva débil |
| 0,4 a 0,69 | Correlación positiva moderada |
| 0,7 a 0,89 | Correlación positiva fuerte |
| 0,9 a 0,99 | Correlación positiva muy fuerte |
| 1 | Correlación positiva perfecta |

Tabla 1. Análisis del resultado del coeficiente de correlación.

## Resultados:

Al tener en cuenta el desarrollo que se planteo anteriormente, se empiezan a visualizar los resultados:

### Tipo de dato de cada variable:

Al leer el dataset proporcionado por la tutora de la asignatura, se observo que se encuentran varias columnas de datos, para las cuales se desea conocer el tipo de datos a la que pertenecen, lo que nos arroja lo que se observa en la FIG. 1, a continuación:

```
11  dataset<- read.csv("E:/es
12  View(dataset)
13
14  class(dataset$id)
15  class(dataset$name)
16  class(dataset$recvTime)
17  class(dataset$NO2)
18  class(dataset$O3)
19  class(dataset$SO2)
20  class(dataset$CO)
21  class(dataset$CO2)
22  class(dataset$PM10)
23  class(dataset$PM25)
24  class(dataset$PM1)
```

```
> class(dataset$id)
[1] "integer"
> class(dataset$name)
[1] "character"
> class(dataset$recvTime)
[1] "character"
> class(dataset$NO2)
[1] "numeric"
> class(dataset$O3)
[1] "numeric"
> class(dataset$SO2)
[1] "numeric"
> class(dataset$CO)
[1] "numeric"
> class(dataset$CO2)
[1] "integer"
> class(dataset$PM10)
[1] "integer"
> class(dataset$PM25)
[1] "integer"
> class(dataset$PM1)
[1] "integer"
```

FIG 1. Tipos de datos de cada variable en el Dataset

**Resumen Estadístico:**

**Tamaño del dataset:**

```
> dim(calidadAire)
[1] 108120     11
```

| | |
|---|---|
| **Numero de Observaciones** | **108.120** |
| **Filas** | **108.120** |
| **Columnas** | **11** |

FIG 2. Tamaño Del Dataset.

**Resumen de los datos:**

```
> summary(calidadAire)
       id             name              recvTime              NO2               O3              SO2               CO
 Min.   :251085   Length:108120      Length:108120      Min.   :   0.00   Min.   :   0.00   Min.   :  0.00   Min.   :  0.00
 1st Qu.:278115   Class :character   Class :character   1st Qu.:   0.00   1st Qu.:  27.35   1st Qu.:  0.00   1st Qu.: 63.91
 Median :305145   Mode  :character   Mode  :character   Median :   3.63   Median :  70.40   Median :  0.00   Median : 88.25
 Mean   :305145                                         Mean   :  33.71   Mean   :  70.02   Mean   : 62.76   Mean   :106.39
 3rd Qu.:332174                                         3rd Qu.:  20.39   3rd Qu.:  99.87   3rd Qu.:176.36   3rd Qu.:132.29
 Max.   :359204                                         Max.   :3895.59   Max.   :1530.53   Max.   :325.74   Max.   :885.40
                                                        NA's   :94153     NA's   :94153     NA's   :94153    NA's   :94153
      CO2             PM10             PM25              PM1
 Min.   :315.0   Min.   :   0.00   Min.   :   0.00   Min.   :   0.00
 1st Qu.:339.0   1st Qu.:   2.00   1st Qu.:   2.00   1st Qu.:   0.00
 Median :345.0   Median :   7.00   Median :   6.00   Median :   3.00
 Mean   :346.5   Mean   :  14.02   Mean   :  12.22   Mean   :   8.39
 3rd Qu.:353.0   3rd Qu.:  19.00   3rd Qu.:  17.00   3rd Qu.:  11.00
 Max.   :395.0   Max.   :3564.00   Max.   :3564.00   Max.   :3564.00
 NA's   :44840   NA's   :77247     NA's   :77247     NA's   :77247
```

FIG 3. Análisis estadístico de cada una de las variables del dataset en RStudio.

| **Análisis estadístico (variables numéricas)** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | **NO2** | **O3** | **SO2** | **CO** | **CO2** | **PM10** | **PM25** | **PM1** |
| **Media** | 33.71 | 70.02 | 62.76 | 106.39 | 346.5 | 14.02 | 12.22 | 8.39 |
| **Mediana** | 3.36 | 70.4 | 0 | 88.25 | 345.0 | 7.00 | 6.00 | 3.00 |
| **Dato mínimo** | 0.00 | 0.00 | 0.00 | 0.00 | 315.0 | 0.00 | 0.00 | 0.00 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Dato máximo** | 3895.56 | 1530.53 | 325.74 | 885.40 | 395.0 | 3564.00 | 3564.00 | 3564.00 |

Tabla 2. Análisis estadístico de cada una de las variables del dataset.

**Tareas de transformación de variables.**

Para transformar las variables utilizamos el método *separate* el cual divide los valores de una columna en 2 o más, y, divide dichos datos según el argumento que definimos para que los separe.

```
# DIVISION DE LA COLUMNA O VARIABLE:  name -> (Sensor_Zona)
calidadAire<-separate(calidadAire,name,c("Sensor","Zona"))
# DIVISION DE LA COLUMNA O VARIABLE:  rcvTime -> (Fecha_Hora)
calidadAire<-separate(calidadAire,recvTime, c("fecha", "hora"), sep = "T")
# DIVISION DE LA COLUMNA O VARIABLE:  fecha -> (anio:mes:dia)
calidadAire<-separate(calidadAire,fecha, c("anio", "mes","dia"), sep = "-")
# DIVISION DE LA COLUMNA O VARIABLE:  hora -> (hora:minutos:segundos)
calidadAire<-separate(calidadAire,hora, c("hora","minutos","segundos"), sep = ":")
View(calidadAire)
# ELIMINAMOS EL CARACTER Z EN LOS SEGUNDOS
calidadAire$segundos<-gsub("Z", "", calidadAire$segundos)
View(calidadAire)
```

```
# REVISO EL TIPO DE VARAIBLE DE LAS NUEVAS VARIABLES
class(calidadAire$Sensor)
class(calidadAire$Zona)
class(calidadAire$anio)
class(calidadAire$mes)
class(calidadAire$dia)
class(calidadAire$hora)
class(calidadAire$minutos)
class(calidadAire$segundos)
```

```
> # REVISO EL TIPO DE VARAIBLE DE LAS NUEVAS VARIABLES
> class(calidadAire$Sensor)
[1] "character"
> class(calidadAire$Zona)
[1] "character"
> class(calidadAire$anio)
[1] "character"
> class(calidadAire$mes)
[1] "character"
> class(calidadAire$dia)
[1] "character"
> class(calidadAire$hora)
[1] "character"
> class(calidadAire$minutos)
[1] "character"
> class(calidadAire$segundos)
[1] "character"
```

FIG 4. Separación de las columnas nombre, tiempo, fecha, y lectura del tipo de variable.

**Transformación de las nuevas variables.**

Utilizamos el método *as.nuemric* para transformar las variables de tipo *character* a numérico para aplicar correlación lineal.

```
# CONVIERTO LAS VARAIBLES QUE NECESITO
calidadAire$anio <- as.numeric(calidadAire$anio)
calidadAire$mes <- as.numeric(calidadAire$mes)
calidadAire$dia <- as.numeric(calidadAire$dia)
calidadAire$hora <- as.numeric(calidadAire$hora)
calidadAire$minutos <- as.numeric(calidadAire$minutos)
calidadAire$segundos <- as.numeric(calidadAire$segundos)
```

```
> summary(calidadAire)
       id             Sensor              Zona                anio          mes
 Min.   :251085   Length:108120      Length:108120      Min.   :2022   Min.   :10.00
 1st Qu.:278115   Class :character   Class :character   1st Qu.:2022   1st Qu.:10.00
 Median :305145   Mode  :character   Mode  :character   Median :2022   Median :11.00
 Mean   :305145                                         Mean   :2022   Mean   :11.02
 3rd Qu.:332174                                         3rd Qu.:2022   3rd Qu.:12.00
 Max.   :359204                                         Max.   :2022   Max.   :12.00

      dia             hora           minutos         segundos           NO2                O3
 Min.   : 1.00   Min.   : 0.00   Min.   : 0.00   Min.   : 0.00   Min.   :   0.00   Min.   :   0.00
 1st Qu.: 8.00   1st Qu.: 6.00   1st Qu.:12.00   1st Qu.:16.72   1st Qu.:   0.00   1st Qu.:  27.35
 Median :17.00   Median :12.00   Median :27.00   Median :35.67   Median :   3.63   Median :  70.40
 Mean   :16.16   Mean   :11.92   Mean   :26.73   Mean   :32.32   Mean   :  33.71   Mean   :  70.02
 3rd Qu.:24.00   3rd Qu.:18.00   3rd Qu.:44.00   3rd Qu.:46.45   3rd Qu.:  20.39   3rd Qu.:  99.87
 Max.   :31.00   Max.   :23.00   Max.   :59.00   Max.   :60.00   Max.   :3895.59   Max.   :1530.53
                                                                 NA's   :94153     NA's   :94153
```

FIG. 5. Trasformación de las variables, y resumen estadístico con las modificaciones implementadas en RStudio.

| **Análisis Estadístico (Nuevas Variables Numéricas)** | | | | | | |
|---|---|---|---|---|---|---|
| | **ANIO** | **MES** | **DIA** | **HORA** | **MINUTOS** | **SEGUNDOS** |
| **Media** | 2022 | 11.02 | 16.16 | 11.92 | 26.73 | 32.32 |
| **Mediana** | 2022 | 11.00 | 17.00 | 12.00 | 27.00 | 35.67 |
| **Dato mínimo** | 2022 | 10.00 | 1.00 | 0.00 | 0.00 | 0.00 |
| **Dato máximo** | 2022 | 12.00 | 31.00 | 23.00 | 59.00 | 60.00 |

Tabla 3. Análisis estadístico de las nuevas columnas.

**Aplicación De Funciones Select () Y Filter ():**

Leemos, analizamos las columnas de nuestro dataset para que no tenga valores NA (anómalos o nulos), y separamos los datos de los sensores de gases, con los sensores de partículas, creando dos nuevos dataset.

```
# PARA GASES (NO2, O3, SO2, CO)
gases<-filter(calidadAire, Sensor==("Gases")|Sensor==("gases"))
gases<-select(gases,-CO2,-PM1,-PM10,-PM25)
View(gases)
# CALCULO DE VALORES NA EN CADA COLUMNA PARA LOS GASES
(sum(is.na(gases$NO2)))
(sum(is.na(gases$O3)))
(sum(is.na(gases$SO2)))
(sum(is.na(gases$CO)))
# COMO LOS NA DAN 0 NO NESECITO IMPUTAR
```

FIG. 6. Código en R Studio para comprobar la existencia de valores anómalos.

Comprobando si existen valores nulos en el dataset (gases).

```
> # CALCULO DE VALORES NA EN CADA COLUMNA PARA LOS GASES
> (sum(is.na(gases$NO2)))
[1] 0
> (sum(is.na(gases$O3)))
[1] 0
> (sum(is.na(gases$SO2)))
[1] 0
> (sum(is.na(gases$CO)))
[1] 0
```

FIG. 7. Resultado de la consola de R Studio sobre existencia de valores nulos.

Realizamos el mismo procedimiento para el data set que contiene los sensores “sonda”

```
# PARA SONDA (GAS CO2)
sonda<-filter(calidadAire, Sensor==("SondaCO2")|Sensor==("SondaCO22")|Sensor==("sonda"))
sonda<-select(sonda,-NO2,-O3,-SO2,-CO,-PM1,-PM10,-PM25)
View(sonda)
# CALCULO DE VALORES NA EN CADA COLUMNA O VARIABLE PARA SONDA CO2
(sum(is.na(sonda$CO2)))
# COMO LOS NA DAN 0 NO NESECITO IMPUTAR
```

```
> # CALCULO DE VALORES NA EN CADA COLUMNA O VARIABLE PARA SONDA CO2
> (sum(is.na(sonda$CO2)))
[1] 0
```

FIG. 8. Análisis y resultados en R Studio para el sensor “sonda”.

Realizamos el mismo procedimiento para el data set que contiene los sensores “partícula”

```
# PARA PARTICULAS (PM1, PM10, PM25)
particulas<-filter(calidadAire, Sensor==("Particulas")|Sensor==("particula"))
particulas<-select(particulas,-NO2,-O3,-SO2,-CO,-CO2)
View(particulas)
# CALCULO DE VALORES NA EN CADA COLUMNA O VARIABLE PARA LAS PARTICULAS
(sum(is.na(particulas$PM10)))
(sum(is.na(particulas$PM25)))
(sum(is.na(particulas$PM1)))
# COMO LOS NA DAN 0 NO NESECITO IMPUTAR
```

```
> # CALCULO DE VALORES NA EN CADA COLUMNA O VARIABLE PARA LAS PARTICULAS
> (sum(is.na(particulas$PM10)))
[1] 0
> (sum(is.na(particulas$PM25)))
[1] 0
> (sum(is.na(particulas$PM1)))
[1] 0
```

FIG. 9. Análisis y resultados en R Studio para el data set del sensor “partícula”.

**Análisis bivariado y resultados de las hipótesis.**

**Hipótesis 1:**

- En los últimos 7 días de noviembre se presentaron los mayores niveles de concentración del gas CO.

*Solución:*

Inicialmente se analizaron los meses existentes en el dataset seguidamente se filtró por el mes que se plantea en la hipótesis.

```
# QUE MESES EXISTEN EN EL DATASET
unique(gases$mes)
# FILTRO LOS DATOS DE NOVIMBRE YA QUE ES EL MES A EVALUAR
gasesNoviembre<-filter(gases, mes==("11"))
View(gasesNoviembre)
```

FIG. 10. Análisis de y filtro de los meses del dataset en R Studio.

Se aplica el test de normalidad de Kolmogórov ya que la cantidad de datos del dataset es superior a 50 datos. El resultado del test es el siguiente:

```
# PASO 1 -> TEST O PRUEBA DE NORMALIDAD

# TEST KOLMOGOROV
lillie.test(gasesNoviembre$dia)          # MAS DE 50 DATOS
lillie.test(gasesNoviembre$CO)           # MAS DE 50 DATOS
```

```
> # TEST KOLMOGOROV
> lillie.test(gasesNoviembre$dia)        # MAS DE 50 DATOS

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  gasesNoviembre$dia
D = 0.072807, p-value < 2.2e-16

> lillie.test(gasesNoviembre$CO)         # MAS DE 50 DATOS

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  gasesNoviembre$CO
D = 0.14257, p-value < 2.2e-16
```

```
# PASO 2 -> ANALIZO EL RESULTADO DEL TEST

# P-VALOR -> MENOR A 0.05, LA DISTRIBUCION NO ES NORMAL
# POR TAL MOTIVO SE USAN LAS PRUEBAS NO PARAMETRICAS (SPEARMAN Y KENDALL)
# SI LA DISTRIBUCION FUESE NORMAL APLICO PRUEBAS PARAMETRICS (PEARSON)
```

FIG. 11. Implementación del test de normalidad de Kolmogórov al dataset en R Studio.

Como P-valor 0.00000000000000022 es menor a 0.05 la distribución no es normal y se debe utilizar pruebas no paramétricas (Spearman y Kendall)

```
# GENERO EL HISTOGRAMA -> PARA ANALIZAR EL COMPORTAMTIENTO DE LAS VARIABLES SELECCIONADAS

# VARAIBLE 1
hist(gasesNoviembre$dia)
# VARIABLE 2
hist(gasesNoviembre$CO)
```

FIG. 12. Generación del histograma en R Studio.

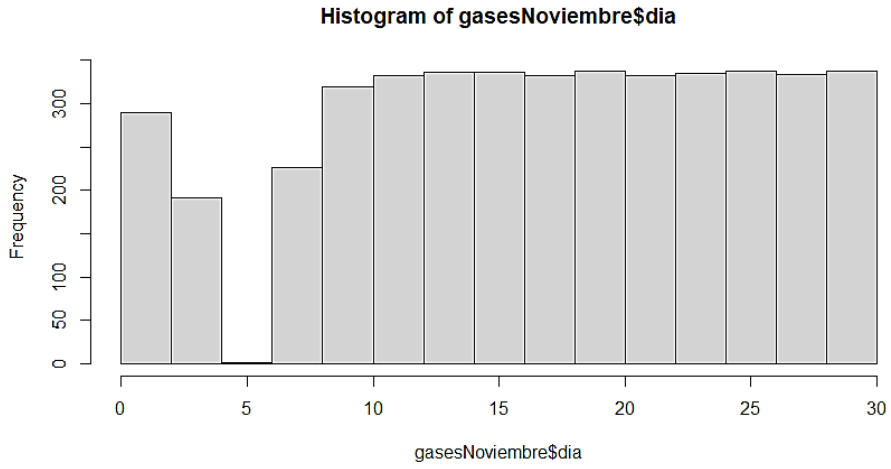


FIG. 13. Histograma del mes de noviembre vs gas en el aire en los distintos días.

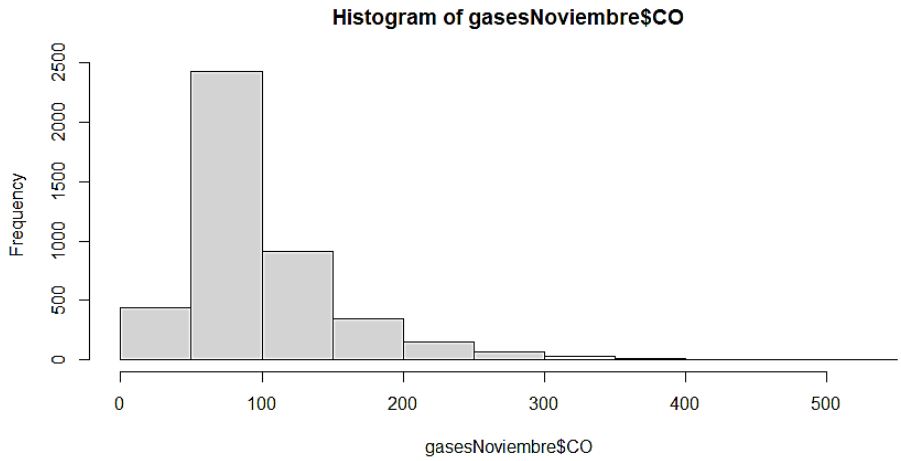


FIG. 14. Histograma del mes de noviembre vs el gas CO en el aire.

Utilizamos la tabla de correlaciones para categorizar el valor generado por el cálculo de correlación tanto de Kendall como Spearman:

```
> # PASO 3 -> APLICO CORRELACION BASADOS EN LOS ANTERIORES PASOS
> # DIA VS GAS CO
> cor(x=gasesNoviembre$dia, y= gasesNoviembre$CO,method = "kendall")
[1] -0.03058558
> cor(x=gasesNoviembre$dia, y= gasesNoviembre$CO,method = "spearman")
[1] -0.04042567
```

FIG. 15. Resultados al aplicar las correlaciones en R Studio.

*Y el resultado nos indica que la correlación entre las dos variables es negativa muy débil.*

```
# PASO 4 -> EVALUO LAS DOS VARIABLES POR MEDIO DE UN GRAFICO PARA ENTENDER LA DISPERSION VISUALMENTE

plot(x=gasesNoviembre$dia, y= gasesNoviembre$CO,col='red')
```

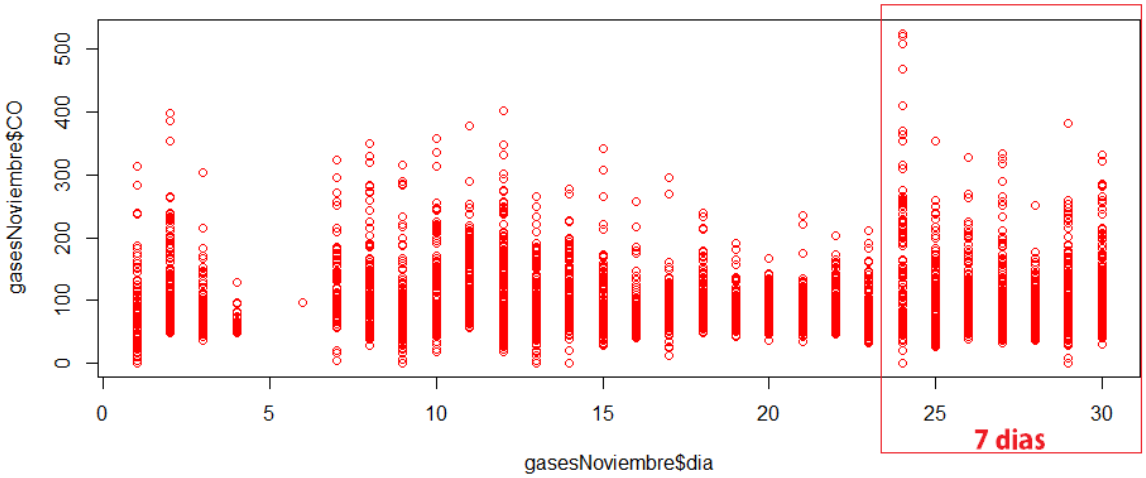


FIG. 16. Generación y observación del grafico de dispersión en R Studio.

El grafico nos indica la relación entre las dos variables

**Análisis hipótesis 1:** El 24 de noviembre fue el día en el que se presentaron mayores concentraciones del gas CO, siendo este el primer día de los últimos 7 en dicho mes, por tal motivo la hipótesis llega a ser válida. A pesar de que los datos no tienen correlación alguna

**Hipótesis 2:**

- El gas NO2 tiene una relación lineal con el gas CO por tal motivo son valores directamente proporcionales entre sí.

```
# PASO 1 -> TEST O PRUEBA DE NORMALIDAD

# TEST KOLMOGOROV
lillie.test(gases$NO2)        # MAS DE 50 DATOS
lillie.test(gases$O3)        # MAS DE 50 DATOS
```

FIG. 17. Implementación del test de normalidad de Kolmogórov al dataset en R Studio.

Se aplica el test de normalidad de Kolmogórov ya que la cantidad de datos del dataset es superior a 50 datos. El resultado del test es el siguiente:

Para el gas NO2

```
> lillie.test(gases$NO2)        # MAS DE 50 DATOS

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  gases$NO2
D = 0.44744, p-value < 2.2e-16
```

FIG. 17.1. Implementación del test de normalidad de Kolmogórov al gas NO2 en R Studio.

Para el gas O3

```
> lillie.test(gases$O3)        # MAS DE 50 DATOS

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  gases$O3
D = 0.17027, p-value < 2.2e-16
```

FIG. 17.2. Implementación del test de normalidad de Kolmogórov al gas O3 en R Studio.

Como P-valor 0.00000000000000022 es menor a 0.05 por tal motivo la distribución no es normal y se debe utilizar pruebas no paramétricas (Spearman y Kendall)

```
# PASO 2 -> ANALIZO EL RESULTADO DEL TEST

# P-VALOR -> MENOR A 0.05, LA DISTRIBUCION NO ES NORMAL
# POR TAL MOTIVO SE USAN LAS PRUEBAS NO PARAMETRICAS (SPEARMAN Y KENDALL)
# SI LA DISTRIBUCION FUESE NORMAL APLICO PRUEBAS PARAMETRICAS (PEARSON)

# GENERO EL HISTOGRAMA -> PARA ANALIZAR EL COMPORTAMTIENTO DE LAS VARIABLES SELECCIONADAS

# VARAIBLE 1
hist(gases$NO2)
# VARIABLE 2
hist(gases$O3)
```

FIG. 18. Generar los histogramas del mes de los gases en el aire.

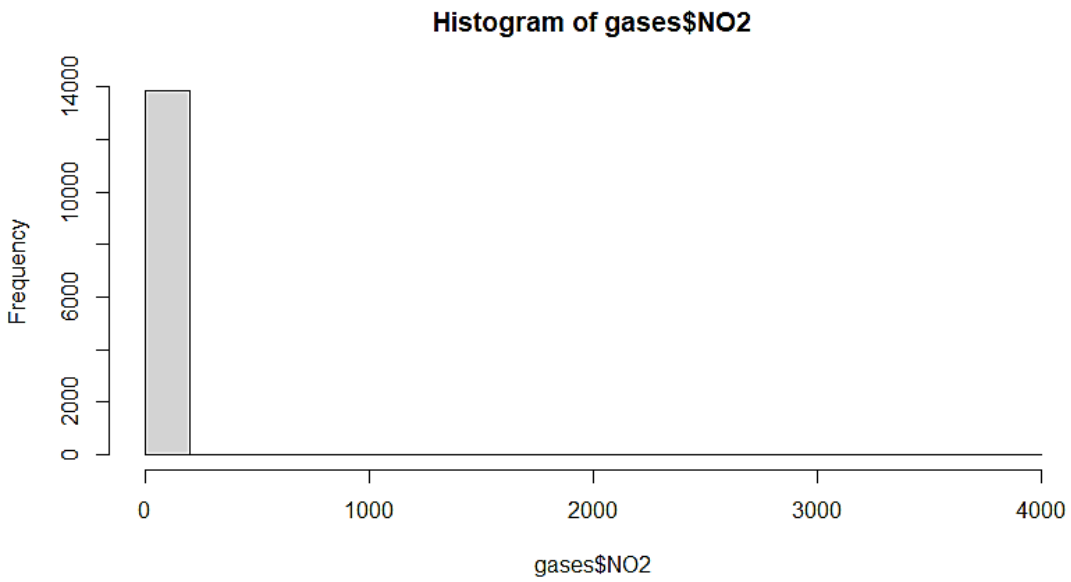


FIG. 19. Histograma del gas NO2, en el aire.

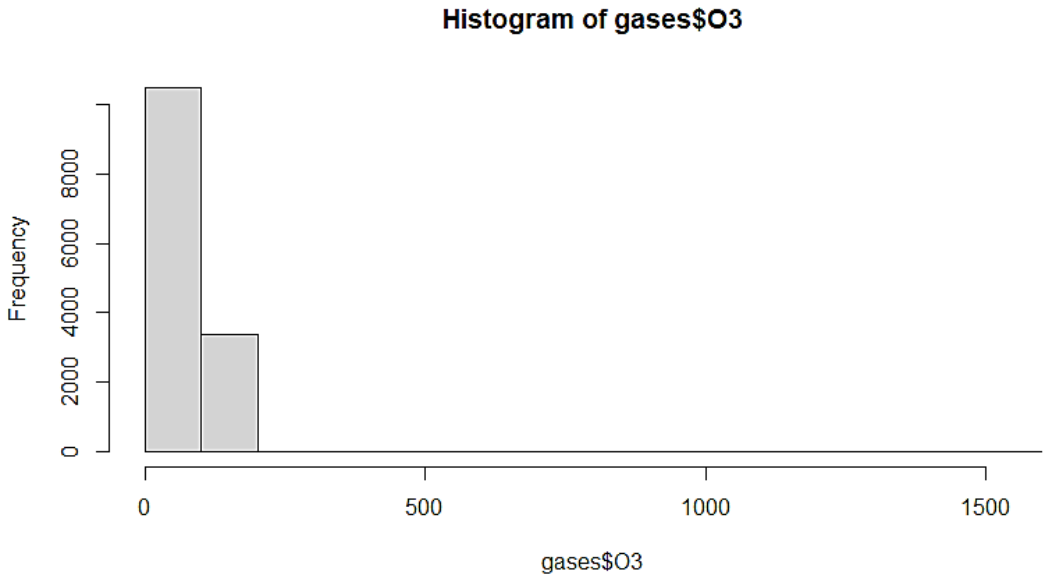


FIG. 20. Histograma del gas O3, en el aire.

Utilizamos la información de la Tabla 1. de correlaciones para categorizar el valor generado por el cálculo de correlación tanto de Kendall como Spearman.

```
# PASO 3 -> APLICO CORRELACION BASADOS EN LOS ANTERIORES PASOS
# MES VS GAS O3
cor(x=gases$NO2, y= gases$O3,method = "kendall")
cor(x=gases$NO2, y= gases$O3,method = "spearman")
```

```
> # PASO 3 -> APLICO CORRELACION BASADOS EN LOS ANTERIORES PASOS
> # NO2 VS GAS O3
> cor(x=gases$NO2, y= gases$O3,method = "kendall")
[1] -0.4593107
> cor(x=gases$NO2, y= gases$O3,method = "spearman")
[1] -0.6178678
```

FIG. 21. Resultados al aplicar las correlaciones en R Studio.

*Y el resultado nos indica que la correlación entre las dos variables es negativa moderada.*

```
# PASO 4 -> EVALUO LAS DOS VARIABLES POR MEDIO DE UN GRAFICO PARA ENTENDER LA DISPERSION VISUALMENTE

plot(x=gases$NO2, y= gases$O3,col='blue')
```

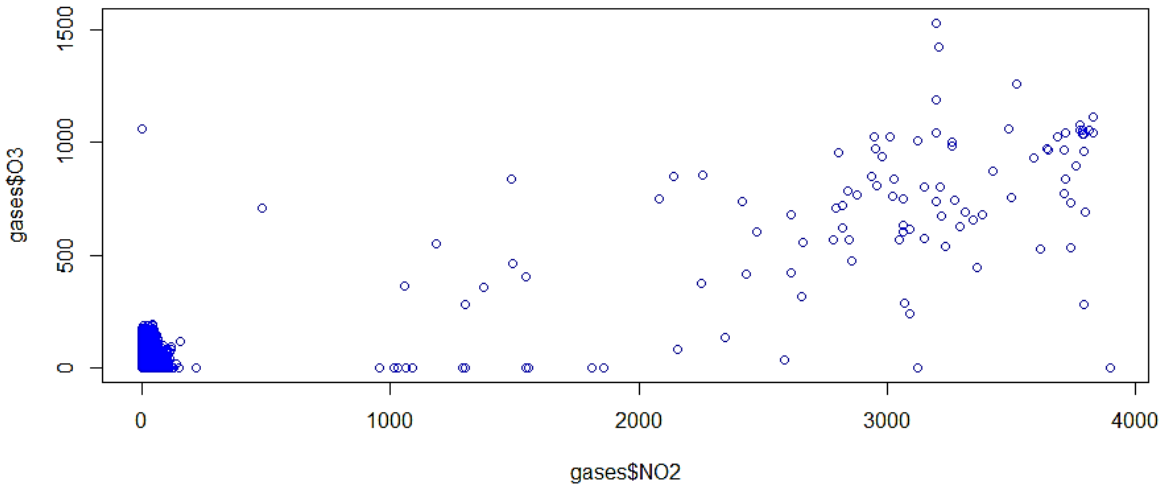


FIG. 22. Generación y observación del grafico de dispersión en R Studio.

El grafico nos indica la relación entre las dos variables

**Análisis hipótesis 2:** Se evidencia que la variable gas NO2 y gas CO tienen una relación media ya que el valor de correlación tiende a estar entre el uno negativo y el cero por tal motivo la hipótesis puede ser medianamente verdadera. La grafica nos demuestra que la mitad de los datos si llegan a correlacionarse entre sí.

**Hipótesis 3:**

- Las partículas PM25 presenta mayores niveles de concentración en el sector Canto Pinyo.

Hago uso de la función factor y unique para crear una columna la cual represente numéricamente el sector ya que como esta variable es categórica no es posible aplicar correlación lineal con la variable partícula PM25. Factor es la función que me convierte en valores numéricos los niveles o categorías, pero dicha función requiere saber cuáles son las categorías a convertir, por medio de la función unique reviso las categorías presentes en la columna que representa a los sectores.

```
# CREO UNA COLUMNA QUE ME REPRESENTA LAS ZONAS CON UN EQUIVALENTE NUMERICO
particulas$ZonaNumerica <- as.numeric(factor(particulas$Zona, levels = unique(particulas$Zona)))
```

FIG. 23. Creación de una nueva columna en el dataset.

Utilizo la función relocate para ordenar las comunas del dataset, se hace uso para mantener juntas las comunas zona (categórica) y zona numérica (numérica).

```
# CAMBIO DE POSICION EN EL DATASET DE LA COLUMNA CREADA
particulasReordenadas<-relocate(particulas, ZonaNumerica, .after = Sensor)
View(particulasReordenadas)
```

FIG. 24. Código en R Studio.

Se aplica el test de normalidad de Kolmogórov ya que la cantidad de datos del dataset es superior a 50 datos. El resultado del test es el siguiente:

```
# PASO 1 -> TEST O PRUEBA DE NORMALIDAD

# TEST KOLMOGOROV
lillie.test(particulasReordenadas$ZonaNumerica)          # MAS DE 50 DATOS
lillie.test(particulasReordenadas$PM25)                  # MAS DE 50 DATOS
```

```
> # TEST KOLMOGOROV
> lillie.test(particulasReordenadas$ZonaNumerica)          # MAS DE 50 DATOS

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  particulasReordenadas$ZonaNumerica
D = 0.21507, p-value < 2.2e-16

> lillie.test(particulasReordenadas$PM25)                  # MAS DE 50 DATOS

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  particulasReordenadas$PM25
D = 0.37866, p-value < 2.2e-16
```

FIG. 25. Implementación del test de normalidad de Kolmogórov a las columnas en R Studio.

*Como P-valor 0.00000000000000022 es menor a 0.05 por tal motivo la distribución no es normal y se debe utilizar pruebas no paramétricas (Spearman y Kendall)*

```
# GENERO EL HISTOGRAMA -> PARA ANALIZAR EL COMPORTAMTIENTO DE LAS VARIABLES SELECCIONADAS

# VARAIBLE 1
hist(particulasReordenadas$ZonaNumerica)
# VARIABLE 2
hist(particulasReordenadas$PM25)
```

FIG. 26. Generar los histogramas.

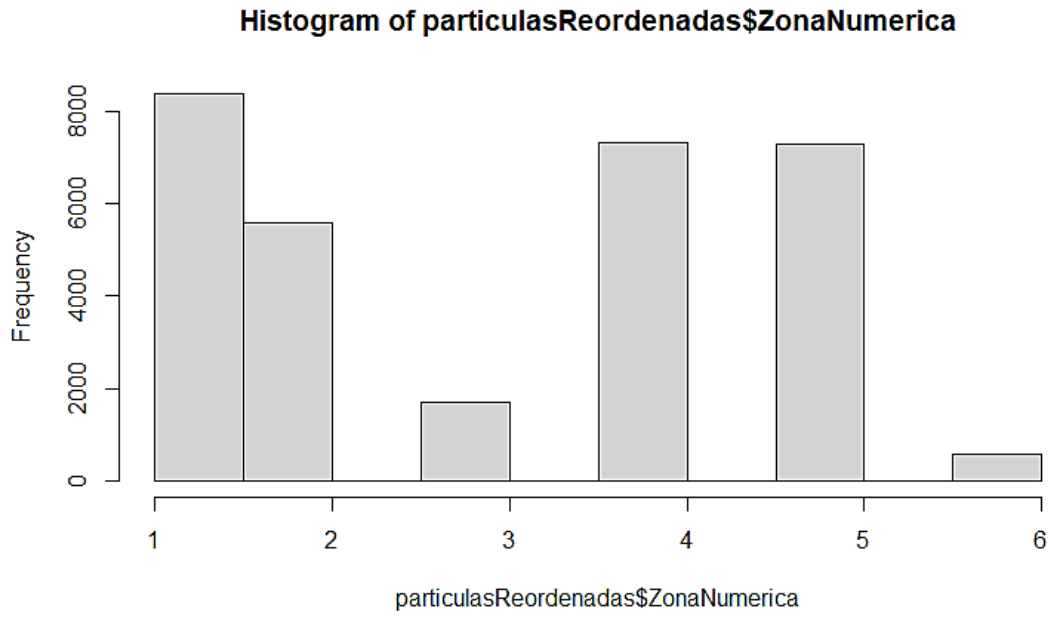


FIG. 27. Histograma de la columna zona numérica.

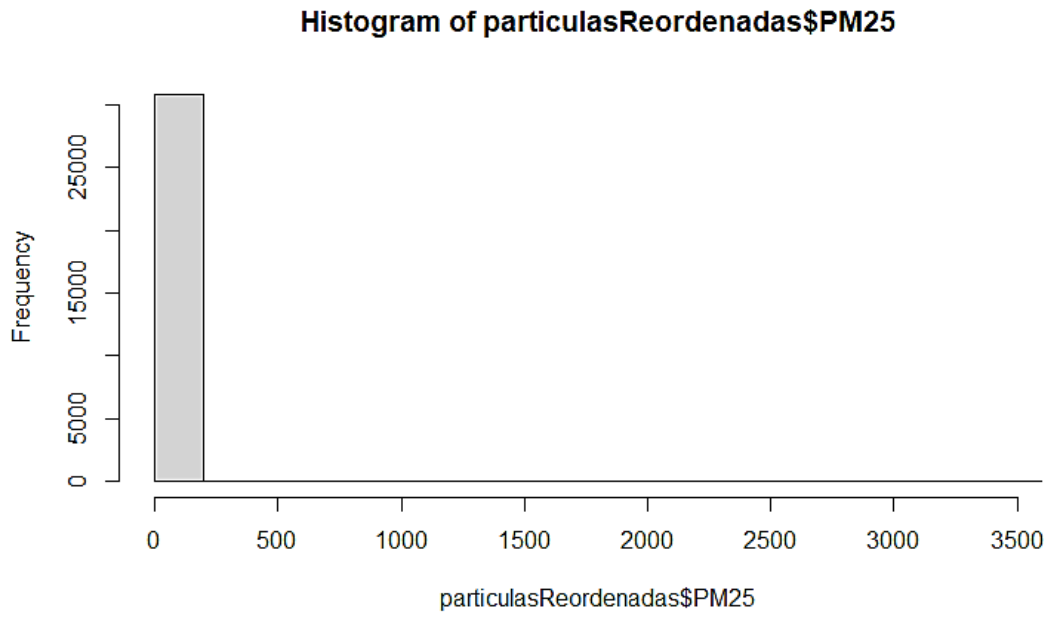


FIG. 28. Histograma de la columna PM25.

Utilizamos los valores de la tabla 1, de correlaciones para categorizar el valor generado por el cálculo de correlación tanto de Kendall como Spearman.

```
# PASO 3 -> APLICO CORRELACION BASADOS EN LOS ANTERIORES PASOS
# ZONANUMERICA VS PARTICULA PM25
cor(x=particulasReordenadas$ZonaNumerica, y= particulasReordenadas$PM25,method = "kendall")
cor(x=particulasReordenadas$ZonaNumerica, y= particulasReordenadas$PM25,method = "spearman")
```

```
> # ZONANUMERICA VS PARTICULA PM25
> cor(x=particulasReordenadas$ZonaNumerica, y= particulasReordenadas$PM25,method = "kendall")
[1] -0.06157138
> cor(x=particulasReordenadas$ZonaNumerica, y= particulasReordenadas$PM25,method = "spearman")
[1] -0.08009167
```

FIG. 29. Resultados al aplicar las correlaciones en R Studio.

*Y el resultado nos indica que la correlación entre las dos variables es negativa muy débil.*

```
# PASO 4 -> EVALUO LAS DOS VARIABLES POR MEDIO DE UN GRAFICO PARA ENTENDER LA DISPERSION VISUALMENTE

plot(x=particulasReordenadas$ZonaNumerica, y= particulasReordenadas$PM25,col='orange')
```

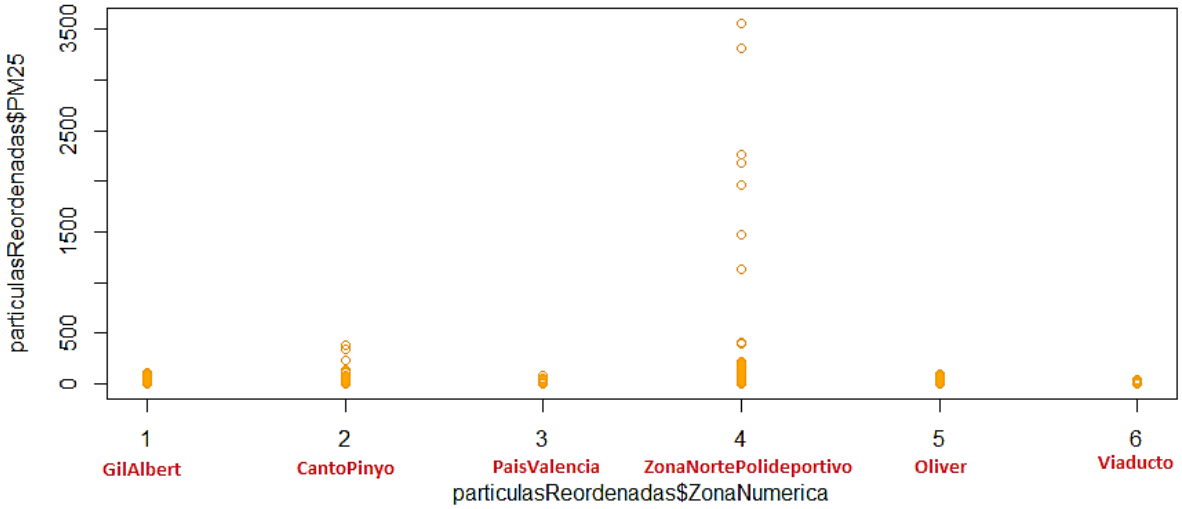


FIG. 30. Generación y observación del grafico de dispersión en R Studio.

El grafico nos indica la relación entre las dos variables

La tabla 4, que se observa a continuación, muestra el valor numérico que representa a cada sector:

| **ZONA NUMÉRICA** | **ZONA** |
|---|---|
| 1 | GilAlbert |
| 2 | CantoPinyo |
| 3 | PaisValencia |
| 4 | ZonaNortePolideportivo |
| 5 | Oliver |
| 6 | Viaducto |

Tabla 4. Número de zona y sector.

**Análisis hipótesis 3:** La partícula *PM 25* presenta los mayores niveles en el sector 4 que según la tabla es el sector **Zona Norte Polideportivo** por tal motivo la hipótesis es falsa. El sector **Canto Pinyo** es el segundo que representa los mayores niveles del gas PM 25 pero como los datos se alejan bastante entre estos dos sectores, su influencia en la hipótesis es nula. Hay que tener en cuenta que la correlación entre los datos es prácticamente nula.

**Hipótesis 4:**

- El gas NO2 presenta mayores concentraciones en las horas que van de (6 pm – 11 pm), por tal motivo se dice que su comportamiento se intensifica en las horas pico de la noche.

Reviso las horas presentes en el dataset para analizar esta variable, como la hora es militar va de 0 a 23 horas, entendemos que 6 pm se ve representado por 18 y 11 pm se representa por 23

```
# ANALIZO QUE HORAS SE TRABAJAN EN EL DATA SET
unique(gases$hora)
```

Se aplica el test de normalidad de Kolmogórov ya que la cantidad de datos del dataset es superior a 50 datos. El resultado del test es el siguiente:

```
# PASO 1 -> TEST O PRUEBA DE NORMALIDAD

# TEST KOLMOGOROV
lillie.test(gases$hora)        # MAS DE 50 DATOS
lillie.test(gases$NO2)        # MAS DE 50 DATOS
```

```
> # TEST KOLMOGOROV
> lillie.test(gases$hora)        # MAS DE 50 DATOS

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  gases$hora
D = 0.079301, p-value < 2.2e-16

> lillie.test(gases$NO2)        # MAS DE 50 DATOS

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  gases$NO2
D = 0.44744, p-value < 2.2e-16
```

FIG. 31. Implementación del test de normalidad de Kolmogórov a las columnas en R Studio.

*Como P-valor 0.00000000000000022 es menor a 0.05 por tal motivo la distribución no es normal y se debe utilizar pruebas no paramétricas (Spearman y Kendall)*

```
# PASO 2 -> ANALIZO EL RESULTADO DEL TEST

# P-VALOR -> MENOR A 0.05, LA DISTRIBUCION NO ES NORMAL
# POR TAL MOTIVO SE USAN LAS PRUEBAS NO PARAMETRICAS (SPEARMAN Y KENDALL)
# SI LA DISTRIBUCION FUESE NORMAL APLICO PRUEBAS PARAMETRICAS (PEARSON)

# GENERO EL HISTOGRAMA -> PARA ANALIZAR EL COMPORTAMTIENTO DE LAS VARIABLES SELECCIONADAS
```

FIG. 32. Analizar los resultados para aplicar las correlaciones en R Studio.


FIG. 33. Histograma de la columna hora.

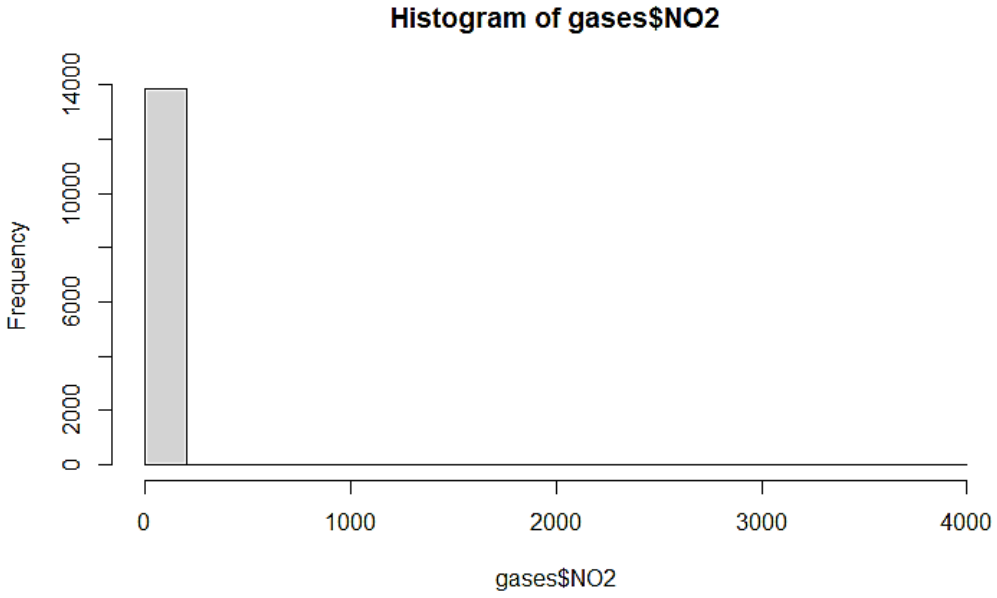


FIG. 34. Histograma de la columna NO2.

Utilizamos los resultados de la tabla 1 de correlaciones para categorizar el valor generado por el cálculo de correlación tanto de Kendall como Spearman.

```
# PASO 3 -> APLICO CORRELACION BASADOS EN LOS ANTERIORES PASOS
# HORA VS GAS NO2
cor(x=gases$hora, y= gases$NO2,method = "kendall")
cor(x=gases$hora, y= gases$NO2,method = "spearman")
```

```
> # HORA VS GAS NO2
> cor(x=gases$hora, y= gases$NO2,method = "kendall")
[1] 0.1559399
> cor(x=gases$hora, y= gases$NO2,method = "spearman")
[1] 0.2143267
```

FIG. 35. Resultados al aplicar las correlaciones en R Studio.

El resultado nos indica que la correlación entre las dos variables según el test de Kendall es positiva muy débil, según el test de Spearman la correlación es positiva débil.

```
# PASO 4 -> EVALUO LAS DOS VARIABLES POR MEDIO DE UN GRAFICO PARA ENTENDER LA DISPERSION VISUALMENTE

plot(x=gases$hora, y= gases$NO2,col='purple')
```

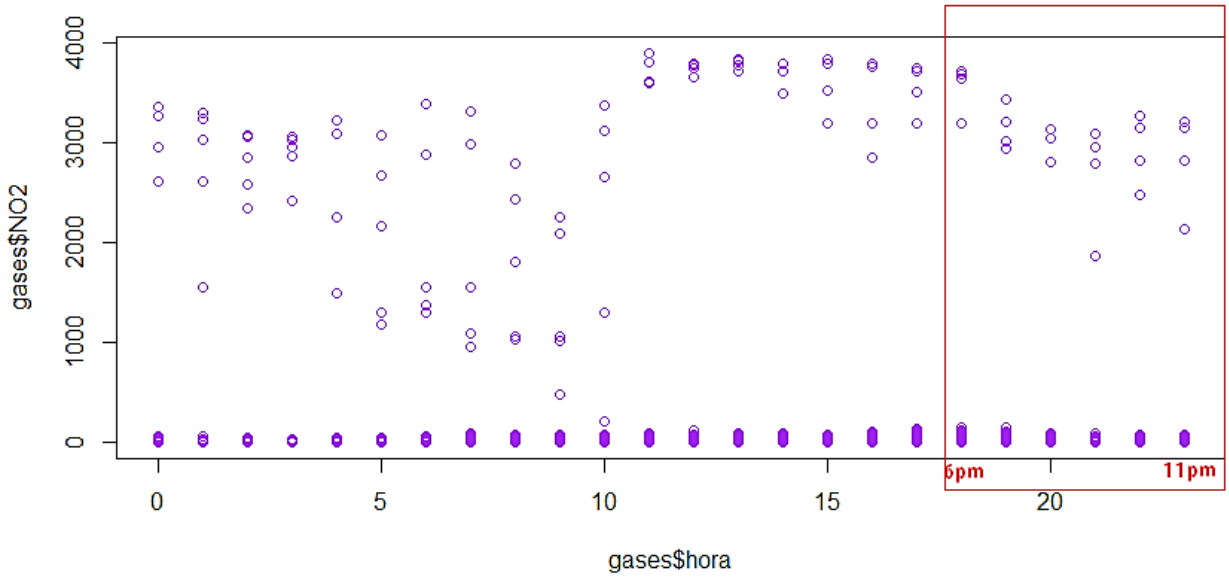


FIG. 36. Generación y observación del grafico de dispersión en R Studio.

El grafico nos indica la relación entre las dos variables.

**Análisis hipótesis 4:** La hipótesis llega a ser falsa ya que el gas NO2 presenta los mayores niveles de concentración a las 11 am, dichos niveles se mantienen de las 6pm en adelante estos niveles se reducen. Hay que tener en cuenta que la correlación entre los datos es no es nula, pero es muy débil.

**Hipótesis 5:**

- Los niveles del gas O3 son más bajos en el mes de octubre que en el mes de diciembre.

Se aplica el test de normalidad de Kolmogórov ya que la cantidad de datos del dataset es superior a 50 datos. El resultado del test es el siguiente:

```
# PASO 1 -> TEST O PRUEBA DE NORMALIDAD

# TEST KOLMOGOROV
lillie.test(gases$mes)          # MAS DE 50 DATOS
lillie.test(gases$O3)           # MAS DE 50 DATOS
```

```
> # TEST KOLMOGOROV
> lillie.test(gases$mes)          # MAS DE 50 DATOS

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  gases$mes
D = 0.23608, p-value < 2.2e-16

> lillie.test(gases$O3)           # MAS DE 50 DATOS

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  gases$O3
D = 0.17027, p-value < 2.2e-16
```

FIG. 36. Implementación del test de normalidad de Kolmogórov a las columnas en R Studio

```
# PASO 2 -> ANALIZO EL RESULTADO DEL TEST

# P-VALOR -> MENOR A 0.05, LA DISTRIBUCION NO ES NORMAL
# POR TAL MOTIVO SE USAN LAS PRUEBAS NO PARAMETRICAS (SPEARMAN Y KENDALL)
# SI LA DISTRIBUCION FUESE NORMAL APLICO PRUEBAS PARAMETRICAS (PEARSON)

# GENERO EL HISTOGRAMA -> PARA ANALIZAR EL COMPORTAMTIENTO DE LAS VARIABLES SELECCIONADAS
```

FIG. 37. Analizar los resultados para aplicar las correlaciones en R Studio.

*Como P-valor 0.00000000000000022 es menor a 0.05 por tal motivo la distribución no es normal y se debe utilizar pruebas no paramétricas (Spearman y Kendall)*

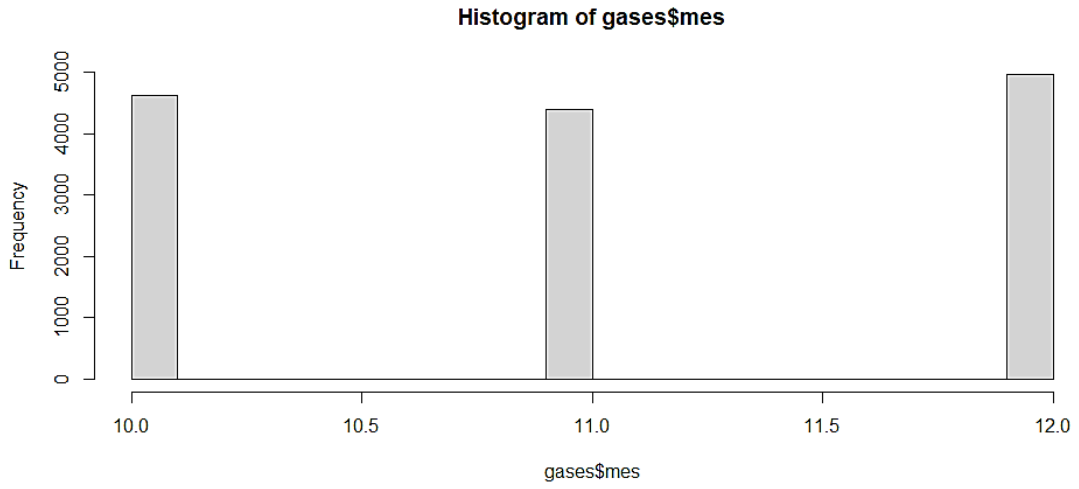


FIG. 38. Histograma de la columna mes en el D.S. gases.

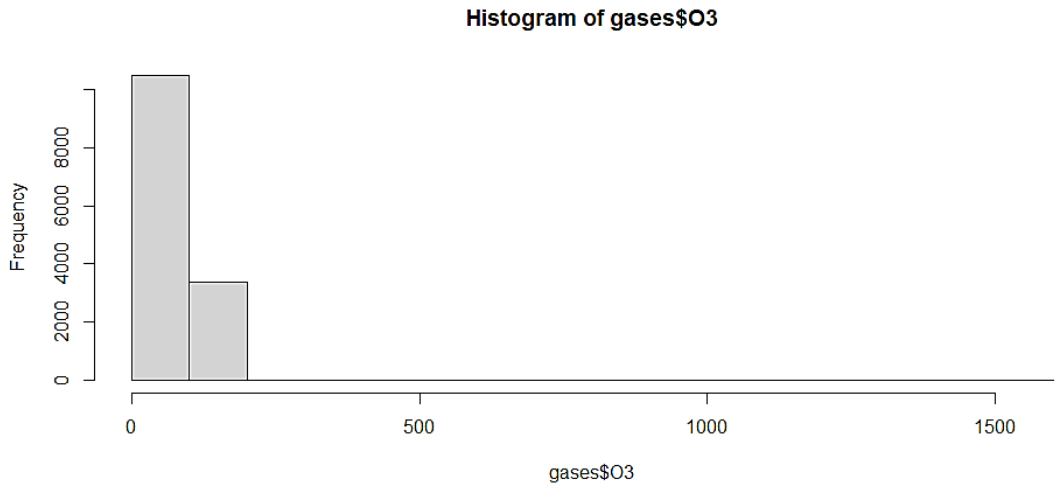


FIG. 39. Histograma de la columna O3 en el D.S. gases.

Utilizamos los resultados de la tabla 1 de correlaciones para categorizar el valor generado por el cálculo de correlación tanto de Kendall como Spearman.

```
# PASO 3 -> APLICO CORRELACION BASADOS EN LOS ANTERIORES PASOS
# MES VS GAS O3
cor(x=gases$mes, y= gases$O3,method = "kendall")
cor(x=gases$mes, y= gases$O3,method = "spearman")
```

```
> # MES VS GAS O3
> cor(x=gases$mes, y= gases$O3,method = "kendall")
[1] -0.1090703
> cor(x=gases$mes, y= gases$O3,method = "spearman")
[1] -0.139956
```

FIG. 40. Resultados al aplicar las correlaciones en R Studio.

*El resultado nos indica que la correlación entre las dos variables es negativa muy débil.*

```
# PASO 4 -> EVALUO LAS DOS VARIABLES POR MEDIO DE UN GRAFICO PARA ENTENDER LA DISPERSION VISUALMENTE
plot(x=gases$mes, y= gases$O3,col='black')
```

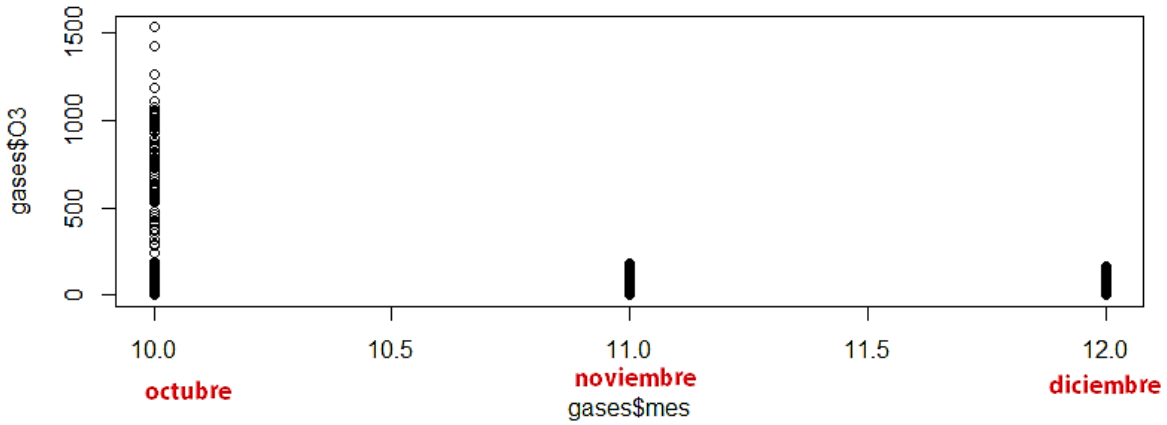


FIG. 41. Generación y observación del grafico de dispersión en R Studio.

*El grafico nos indica la relación entre las dos variables*

La siguiente tabla #5 muestra el valor numérico que representa a cada sector:

| MES NUMÉRICO | MES |
|---|---|
| 10 | Octubre |
| 11 | Noviembre |
| 12 | Diciembre |

Tabla 5. Valores del mes y mes

*Análisis hipótesis 5:* La hipótesis es falsa ya que el gas O3 presenta sus mayores niveles en el mes de octubre, y según la hipótesis este sería el mes con los niveles más bajos, según el grafico el mes con los niveles más bajos del gas O3 es diciembre. Hay que tener en cuenta que la correlación entre los datos es prácticamente nula.

## *Discusión de resultados.*

En primer lugar, se observa que la hipótesis en cuanto al gas O3 es falsa, ya que sus mayores niveles se presentan en el mes de octubre, en lugar del mes de diciembre como se planteó inicialmente. Además, se destaca que la correlación entre los datos es prácticamente nula.

En segundo lugar, la hipótesis en cuanto al gas NO2 también se ve afectada, ya que se observa que los mayores niveles de concentración de este gas se presentan a las 11 am y se mantienen desde las 6 pm en adelante, lo cual difiere de la hipótesis inicial. Sin embargo, se debe tener en cuenta que la correlación entre los datos es débil y no nula.

Por último, en cuanto a la partícula PM25, se puede concluir que la hipótesis es falsa, ya que los mayores niveles se presentan en el sector 4, que corresponde a la Zona Norte Polideportivo. Aunque el sector Canto Pinyo también presenta altos niveles de PM25, se debe tener en cuenta que la distancia entre estos dos sectores es grande, lo que indica que la influencia de Canto Pinyo en la hipótesis es nula. Es importante resaltar que la correlación entre los datos es prácticamente nula en este caso.

En conclusión, a pesar de que la hipótesis inicial no se cumple completamente, los resultados obtenidos son valiosos y permiten tener una mejor comprensión de la calidad del aire en la zona estudiada. Además, se pueden establecer nuevas preguntas de investigación y líneas de trabajo a partir de estos resultados.

## *Conclusiones.*

Después de analizar los resultados obtenidos en este estudio, se pueden hacer las siguientes conclusiones:

La hipótesis planteada sobre los niveles de concentración del gas O3 en los diferentes meses del año ha resultado ser falsa. Según el análisis realizado, los mayores niveles de este gas se presentan en el mes de octubre, mientras que se esperaba que fueran más bajos. Además, se encontró una correlación prácticamente nula entre los datos.

Respecto a la concentración del gas NO2, se puede concluir que la hipótesis también resulta ser falsa. Los mayores niveles de concentración de este gas se presentan a las 11 am, en lugar de en la tarde, como se esperaba. Además, se encontró una correlación débil entre los datos.

En cuanto a la partícula PM 25, se pudo determinar que la hipótesis también resulta ser falsa. Los mayores niveles de esta partícula se presentan en el sector 4, que corresponde a la Zona Norte Polideportivo, y no en el sector Canto Pinyo como se esperaba. Aunque la correlación entre los datos es prácticamente nula, se puede

concluir que el sector 4 tiene una mayor influencia en los niveles de concentración de PM 25.

En conclusión, los resultados obtenidos en este estudio no apoyan las hipótesis planteadas inicialmente. Es importante tener en cuenta que los datos presentan diferentes niveles de correlación y que hay factores adicionales que pueden afectar los niveles de concentración de los gases y partículas en el aire. Estos resultados pueden ser útiles para futuras investigaciones y para la implementación de políticas públicas orientadas a mejorar la calidad del aire en la zona estudiada.


### *Agradecimientos.*

Agradecemos principalmente a nuestra tutora, *ing. Mónica Duran*, por su orientación, conocimientos y experiencia, transmitidos a lo largo del semestre, y que han sido fundamentales para el éxito de este estudio. A la plataforma de datos abiertos del Gobierno de España por proporcionar públicamente el conjunto de datos utilizado en este estudio, lo que nos permitió llevar a cabo un análisis detallado sobre la calidad del aire en distintas zonas y localidades de España. Finalmente, agradecemos a nuestras familias y seres queridos por su apoyo incondicional durante todo este proyecto, sin su motivación y paciencia, no hubiéramos tenido la capacidad de lograr este resultado.


### *Referencias:*


Universidad Militar Nueva Granada, Informe: https://repository.unimilitar.edu.co

María Santos Pedraza Guevara. (2021). FITORREMEDIACIÓN EN CUERPOS DE AGUA CONTAMINADOS POR METALES PESADOS. Innova Biology Sciences, 1(1), 61–78. https://doi.org/10.58720/ibs.v1i1.6

Parra Sánchez, J. S., Oviedo Carrascal, A. I., & Amaya Fernández, F. O. (2020). Analítica de datos: incidencia de la contaminación ambiental en la salud pública en Medellín (Colombia). Revista De Salud Pública, 22(6), 609–617. https://doi.org/10.15446/rsap.v22n6.78985

Calidad del aire en España: https://datos.gob.es/es/catalogo/l01030092-mediciones-de-calidad-del-aire

RStudio Team. (2021). RStudio: Integrated Development Environment for R. RStudio. https://www.rstudio.com/

Lozano, L., Hernández, J., & Chirino, E. (2020). Datos abiertos y análisis de calidad del aire en la ciudad de Caracas, Venezuela. Acta Científica Venezolana, 71(4), 78-85. Recuperado de http://sedici.unlp.edu.ar/handle/10915/108542

Lozano, C., Benavides, J. A., & León, L. (2019). Evaluación de la calidad del aire en la ciudad de Quito mediante el análisis de indicadores ambientales. EPN. Escuela Politécnica Nacional. https://bibdigital.epn.edu.ec/handle/15000/19825

Forbes Argentina. Recuperado de https://www.forbesargentina.com/innovacion/como-inteligencia-artificial-puede-ser-clave-lucha-cambio-climatico-n18566