



INSTITUTO TECNOLÓGICO Y DE ESTUDIOS SUPERIORES DE
MONTERREY

Escuela de Ingeniería y Ciencias
Ingeniería en Ciencia de Datos y Matemáticas

Conocimiento de la naturaleza de contaminantes que influyen en la calidad del aire y sus interrelaciones

Etapas 1. Conociendo el negocio

APLICACIÓN DE MÉTODOS MULTIVARIADOS EN CIENCIA DE DATOS

Mariana Ivette Rincón Flores A01654973

Leonardo Laureles Olmedo A01659241

Carlos Mateos Perez A01654085

Renata Vargas Caballero A01025281

Andrea Bravo Avila A01028579

Alberto Lozano Cárdenas A01067141

En conjunto con: SIMA

Supervisado por

Rubí Isela Gutiérrez López
Blanca Rosa Ruiz Hernández

El trabajo realizado es para fines académicos sin fines de lucro. Queda prohibida la reproducción total o parcial de los datos (en bruto o enmascarados), resultados, modelos y conclusiones sin el previo consentimiento por escrito otorgado por SIMA.

Monterrey, Nuevo León. Fecha, 10 de septiembre de 2022

1. Introducción y Marco teórico

En un mundo cada vez más urbanizado, donde en 2018, la ONU estimó que 55 % de la población mundial vive en zonas urbanas, la calidad del aire se ha vuelto de vital importancia Nations, 2018. Conocer la calidad del aire nos ayuda a conocer la composición y concentración de los múltiples gases y partículas que se encuentran dispersas en el ambiente, éstas se deben de analizar para cumplir con un equilibrio y calidad determinado cuya finalidad es permitir que todos los seres vivos puedan disfrutar una vida saludable (no solo ciudadanos). La contaminación del aire representa un riesgo para la salud de los ciudadanos y es que conocer la calidad del aire puede representar la prevención de enfermedades/dolencias como la fatiga, dolor de cabeza, ansiedad, irritación de ojos y mucosas, entre otros de la Salud, 2018.

Para poder determinar si la contaminación está en un nivel riesgoso se pueden considerar varios factores. Los aspectos que se consideran para analizar la calidad del aire son principalmente la composición del aire que se encuentre en el momento actual, y es que los principales elementos que se encuentran en el aire son nitrógeno, oxígeno e hidrógeno. La temperatura, humedad, vientos, precipitaciones, radiación solar y presión atmosférica son factores importantes a considerar ya que pueden condicionar la dispersión y reacciones químicas de los elementos presentes en la atmósfera dependiendo de la situación actual.

La calidad del aire se mide con el índice de calidad del aire que va en una escala de 0 a 500 y establece 6 niveles de peligrosidad, entre más alto sea el índice, peor será la calidad del aire. El rango se divide en: Buena (Verde, ICA 0-50), Moderada (Amarillo, ICA 51-100), Dañina a la salud para grupos sensibles (Naranja, ICA 101-150), Dañina a la salud (Rojo, ICA 151-200), Muy dañina a la salud (Morado, ICA 201-300), Peligrosa (Marrón, ICA >300). Las variables que intervienen en la posible mejora de la calidad del aire son la calidad de las aguas, vertidos de aguas residuales, medición de emisiones atmosféricas, control de sedimentos, estudios de caracterización de suelos y residuos (Palou, 2017).

En México, la Secretaría de Salud es responsable de medir los impactos de la contaminación atmosférica en la salud, mediante el Cuadro 1 se establecen los límites permisibles de la concentración de los contaminantes en la atmósfera con su respectiva norma ("Normas Oficiales Mexicanas (NOM) de Calidad del Aire Ambiente", 2017).

Contaminante	Dato base utilizado para la evaluación	Exposición	Frecuencia tolerada	Valor límite Indicador con el que se evalúa	Norma Oficial Mexicana
Partículas PM10	Promedio 24 horas	Aguda	No se permite	75 $\mu\text{g}/\text{m}^3$ Máximo	NOM-025-SSA1-2014
		Crónica	–	40 $\mu\text{g}/\text{m}^3$ Promedio anual	
Partículas PM2.5	Promedio 24 horas	Aguda	No se permite	45 $\mu\text{g}/\text{m}^3$ Máximo	
		Crónica	–	12 $\mu\text{g}/\text{m}^3$ Promedio anual	
Ozono (O3)	Dato horario	Aguda	No se permite	0.095 ppm Máximo	NOM-020-SSA1-2014
	Promedio móvil de 8 hora		No se permite	0.070 ppm Máximo	
Dióxido de azufre (SO2)	Promedio de 8 hora	Aguda	1 vez al año	0.200 ppm Segundo máximo	NOM-022-SSA1-2010
	Promedio e 24 hora	Aguda	No se permite	0.110 ppm Máximo	
	Dato horario	Crónica	–	0.025 ppm Promedio anual	
Dióxido de nitrógeno (NO2)	Dato horario	Aguda	1 vez al año	0.210 ppm Segundo máximo	NOM-023-SSA1-1993
Monóxido de carbono (CO)	Promedio móvil de 8 hora	Aguda	1 vez al año	11 ppm Segundo máximo	NOM-021-SSA1-1993
Plomo (Pb)	Promedio aritmético de tres meses	Crónica	No se permite	1.5 g/m^3	NOM-026-SSA1-1993

Cuadro 1: Normas Oficiales Mexicanas (NOM) de Calidad del Aire Ambiente (“Normas Oficiales Mexicanas (NOM) de Calidad del Aire Ambiente”, 2017

De igual manera, la Secretaría de Desarrollo Sustentable se encarga de establecer, instrumentar y coordinar las políticas, estrategias y planes que promueven el desarrollo urbano y programas para fomentar un medio ambiente sustentable.

Sin embargo, medir la calidad del aire, a pesar de tener métricas específicas estipuladas, no es una tarea simple. La presencia de la diversidad de relieves en Nuevo León provoca un obstáculo en el posicionamiento de los agentes receptores ya que, debido a las corrientes de aire, los valores verdaderos se pueden ver modificados. También es importante tener un amplio conocimiento y capacidad de predicción de factores meteorológicos lo que va a determinar como las partículas contaminantes se mueven, expanden y reaccionan entre ellas.

Otro obstáculo es la relación costo-beneficio que resulta de implementar un sistema de medición y departamento de analítica. La Sociedad Chilena de Enfermedades Respiratorias (Matus y LUCERO CH, 2002) dice que uno de los pocos factores con los que se puede medir el beneficio de este proyecto de manera monetaria, es la disminución de casos de riesgo de enfermedades respiratorias.

En el área Metropolitana de Monterrey se cuenta con un centro que se enfoca en la recopilación y analítica de datos que rodean el estudio de la contaminación ambiental, llamado El Sistema Integral de Monitoreo Ambiental, SIMA. Fue fundada en 1992 y a lo largo de los años ha ayudado a diversos estudios para estudiar el comportamiento de la calidad del aire (“SIMA”, 2015).

1.1. Descripción del problema específico

El reto consiste en diseñar una propuesta para la detección de las relaciones que se establecen entre los contaminantes del aire y su relación con el clima en las distintas estaciones metereológicas de la Ciudad de Monterrey.

Para dirigir la solución al reto, se seguirá la metodología CRISP-DM que consta de las siguientes

etapas: comprensión del negocio (establecimiento de las expectativas del proyecto así como los criterios de aceptación del mismo), comprensión y preparación de los datos (evaluación, exploración de los datos con la ayuda de tablas y gráficos utilizando lenguajes de programación y bibliotecas de manejo de datos), modelado y evaluación (uso de los datos organizados o transformados en la etapa anterior para determinar sus características óptimas y así generar el modelo adecuado seleccionando una técnica de análisis multivariado que resuelva la problemática establecida por la competencia), despliegue (interpretación de los resultados en términos de la problemática esperada).

1.2. Solución: Idea central del proyecto

Se contempla el análisis de bases de datos con al menos diez variables interrelacionadas, mayormente variables numéricas. Asimismo se contará con un mínimo de 300 registros por base de datos. Se busca dar respuesta a varias preguntas; realizar varios análisis estadísticos con técnicas multivariadas: análisis de regresión lineal múltiple; regresión multivariada; análisis discriminante; análisis por componentes principales; análisis de factores o análisis de conglomerados, además, se brindarán recomendaciones a SIMA y la dirección de gestión de calidad del aire sobre cómo minimizar dichos problemas.

1.3. Hipótesis

Dentro de la base de datos se espera encontrar diversos indicadores para poder dar una propuesta de valor y poder generar un mejor entregable.

1.4. Objetivos

1.4.1. Objetivo general

Detectar las relaciones que se establecen entre los contaminantes del aire y su relación con el clima en las distintas estaciones meteorológicas de la Ciudad de Monterrey.

1.4.2. Objetivos específicos

- Modelación de las interrelaciones de los contaminantes que influyen en el aire.
- Modelación de factores ambientales que tienen influencia en la concentración de los contaminantes. Igualmente, detectar condiciones más desfavorables.
- Modelación de las condiciones que están fuera y dentro de las normas oficiales establecidas como normas de calidad de salud.
- Utilizar técnicas estadísticas para la modelación (Análisis multivariado) de los datos de la calidad del aire.

1.5. Justificación

La ciencia de datos es una rama muy moderna que a la vez incluye otras técnicas como Machine Learning, Deep Learning, Big Data, Artificial Intelligence, entre muchas otras, que dan oportunidad a las empresas de analizar sus datos de una manera más precisa y confiable al ser respaldadas por fundamentos matemáticos, estadísticos y de programación. La empresa socio formadora, SIMA y la dirección de gestión de calidad del aire, busca un armado de propuesta para el análisis de la relación entre los contaminantes del aire y las distintas estaciones meteorológicas de la Ciudad de Monterrey.

1.6. Descripción de las fuentes de información (datos)

El socio formador nos proporcionó una base de datos que incluye la información de los datos promedio diarios de la red de monitoreo ambiental del Estado de Nuevo León (SIMA), la información se presenta dividida en libros que corresponden a cada contaminante criterio y parámetros meteorológicos que se mide en las estaciones del SIMA, así como la descripción de cada abreviatura de las estaciones de monitoreo y las unidades de medición para contaminante criterio.

1.7. Selección del modelo

Para poder resolver esta problemática, buscaremos utilizar varios algoritmo de Machine Learning con ayuda de la regresiones, así como diferentes análisis estadísticos en los lenguajes de programación R y Python.

1.8. Descripción de la solución (alcance)

La solución mostrará las relaciones que se establecen entre los contaminantes del aire y el clima en las distintas estaciones meteorológicas de la Ciudad de Monterrey, al tener conocimiento de esta relación se podrían crear regulaciones para mantener estos contaminantes lo más bajos posibles y mejorar la calidad del clima en la región, lo que impactaría de manera positiva en la población.

1.9. Propuesta de valor

Al poder identificar cuáles son los factores que más se relacionan entre los contaminantes del aire y las distintas estaciones meteorológicas, se obtiene información de valor que puede ser utilizada para atender cuestiones relacionadas a los problemas de salud de la población ocasionadas por la mala calidad del aire.

1.10. Nombre detallado del proyecto

El proyecto busca mediante el análisis de ciertas variables encontrar patrones con el mayor número de incidencias que puedan dar información de valor para detectar las relaciones que se establecen entre los contaminantes del aire y las distintas estaciones meteorológicas de la Ciudad de Monterrey.

1.11. Nombre corto o comercial del proyecto

Cli-max

1.12. Impacto social principal

La realización de este trabajo, permitirá detectar las relaciones que se establecen entre los contaminantes del aire y su relación con el clima en las distintas estaciones meteorológicas de la Ciudad de Monterrey, por lo que se podría crear un impacto positivo ya que al encontrar las relaciones se pueden idear planes para reducir dichos contaminantes del aire y mejorar la calidad del clima lo que beneficiaría a toda la población de la Ciudad de Monterrey.

1.13. Impacto hacia los Objetivos de Desarrollo Sostenible


ODS	Justificación
	Se ha implementado la Estrategia de Descontaminación, que se estructura a través de las normas primarias de calidad ambiental que regulan la concentración de los contaminantes del aire nocivos para la salud.
	Se centra en la buena salud y el bienestar de toda la sociedad, donde la calidad del aire es un punto muy importante, ya que los diferentes contaminantes que se emiten pueden llegar a ser nocivos para la salud.
	Al ser un área urbana, contar con 176 parques industriales, implica que haya una mayor contaminación en el aire, por lo que se busca crear conciencia en las empresas para que el aire pueda

Figura 1: ODS relacionadas con el proyecto.

2. Comprensión y Preparación de los datos

2.1. Comprensión de los datos del negocio

Se nos proporcionaron 3 bases de datos y un catálogo con datos promedios diarios de la red de monitoreo ambiental del Estado de Nuevo León. El catálogo contiene información descriptiva para las variables dentro de las bases, sus definiciones, especificaciones y notas importantes a tomar en consideración. A continuación la información de las bases de datos meteorológicos y contaminantes:

2.1.1. Meteorológicos:

- Tamaño de las hojas en la base de meteorológicos: (39297, 28), (39382, 28), (39389, 28), (39391, 28), (39270, 28), (39259, 28), (39320, 28).

- Descripción de variables:

TOUT: Es la variable que marca la temperatura, es una variable numérica, sus valores se miden en grados centígrados y es una variable tipo float64.

RH: Es una variable que marca la humedad relativa por medio de porcentaje, es numérica y es de tipo int.

SR: Es una variable que nos indica la Radiación Solar, se mide por Kilo Wats por metro cuadrado, es tipo float64.

RAINF: Es una variable que nos indica la precipitación, se mide por mililitro por hora y esta variable es de tipo float64.

PRS: Es una variable que mide la presión atmosférica, su unidad de medida es el milímetro de mercurio, es de tipo float64.

WS: Es una variable que describe la velocidad del viento, se mide en kilómetros por hora y es de tipo Float 64.

WD: Es una variable que describe la dirección del viento, se mide en grados y es de tipo Int.

- Medias estadística (variables cuantitativas):

	TOUT	RH	SR	RAINF	PRS	WS	WD
count	38630.000000	38656.000000	39123.000000	38037.000000	38212.000000	32066.000000	15010.000000
mean	23.374690	59.029465	0.209101	0.011578	715.356723	7.153144	171.952965
std	7.262886	21.159404	0.254281	0.290675	5.014861	79.172452	94.538942
min	-5.670000	0.000000	0.000000	0.000000	620.300000	-9999.000000	1.000000
25%	18.690000	42.000000	0.001000	0.000000	712.100000	3.700000	112.000000
50%	24.090000	61.000000	0.045000	0.000000	714.800000	7.400000	138.000000
75%	28.330000	77.000000	0.467000	0.000000	718.200000	11.000000	251.000000
max	42.670000	98.000000	0.900000	29.300000	736.400000	34.700000	360.000000

Figura 2: Exploración de datos

Aquí podemos ver como estaban los datos antes de la limpieza, aún hay datos erróneos y repetidos, por lo que la media y desviación estándar se ven afectadas. Se observan las medidas de tendencia central y de dispersión para las variables meteorológicas.

- Medidas estadística (variables cualitativas): No se utilizaron variables cualitativas ya que no habían en la base de datos. Los únicos registros de valores no numéricos indican anotaciones sobre los registros como la validez de la hora de registro y qué tipo de registro (falla eléctrica, calibración, apagado, malas condiciones). Éstos datos son necesarios para poder enlazarlos con

los contaminantes, ya que es el único dato que sirve como parámetro para poder relacionar ambas bases de datos.

- Calidad de los datos: Para ver la calidad de los datos, utilizamos diagramas de caja y bigote.

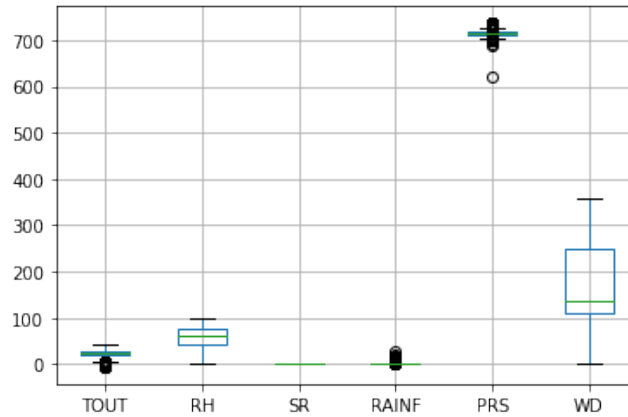


Figura 3: Box plot de variables meteorológicas

Se observa que la humedad relativa, radiación solar y el viento no tienen valores atípicos, contrario a la temperatura, precipitación y presión atmosférica que sí tienen valores fuera de los boxplots. Sin embargo al ver los valores posibles de estas variables, si es viable que en un momento dado estas mediciones aparezcan y sean reales. Por esta razón los datos normalmente considerados atípicos que se observan en la gráfica se mantendrán.

- Calidad de los datos: También se graficaron las variables contra el tiempo, para ver cómo se comportaban los datos, y poder ver la calidad de los datos

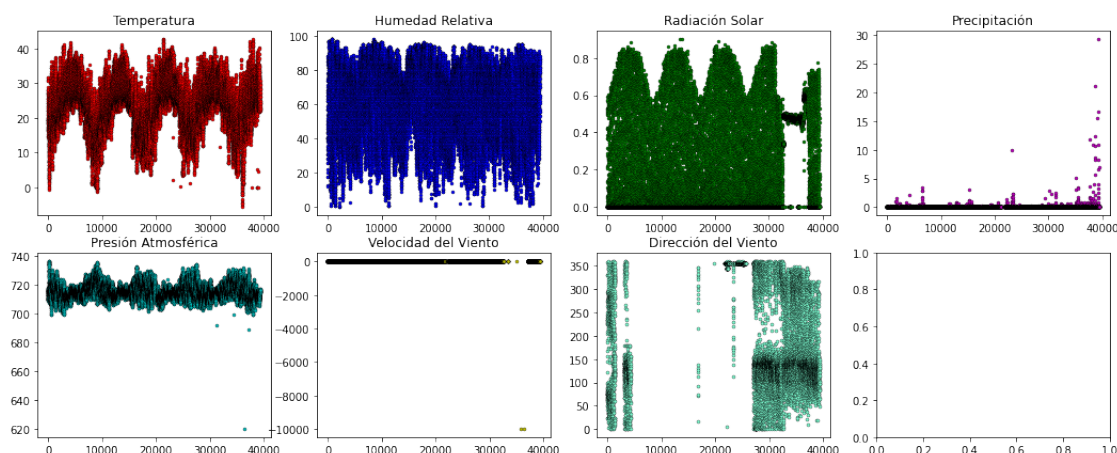


Figura 4: Gráficos de variables meteorológicas en el tiempo

Pudimos observar que en la variable de dirección del viento hay muchos datos faltantes, secciones completas que representan años donde no se registraron datos. También observamos que hay datos erróneos en la variable de velocidad del viento que no permiten ver la silueta de la distribución de los datos. Esto nos ayudará a limpiar los datos adecuadamente. En esta visualización también pudimos ver el carácter periódico de los datos, viendo como los datos meteorológicos dependen de las estaciones, especialmente en temperatura, humedad relativa y radiación solar, se ven muy claras las estaciones de los 5 años de los datos.

- Correlación entre las variables: Visualizamos la correlación entre las variables en un mapa de calor.

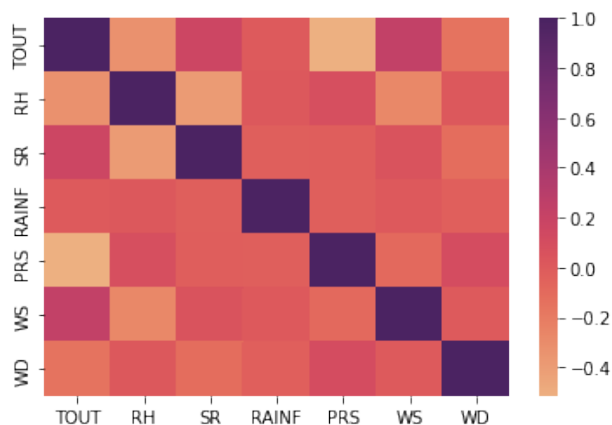


Figura 5: Mapa de calor de la matriz de correlación de los valores de datos meteorológicos

Se observa que ninguna de estas variables tiene una correlación alta, no hay ninguna correla-

ción mayor a 0.5.

- Correlación entre las variables: Cada uno de los cuadrantes del mapa de calor se puede visualizar como su propio diagrama de dispersión, o en el caso de los 1 comparando con si mismo podemos ver un hisograma, esto se muestra en la siguiente figura.

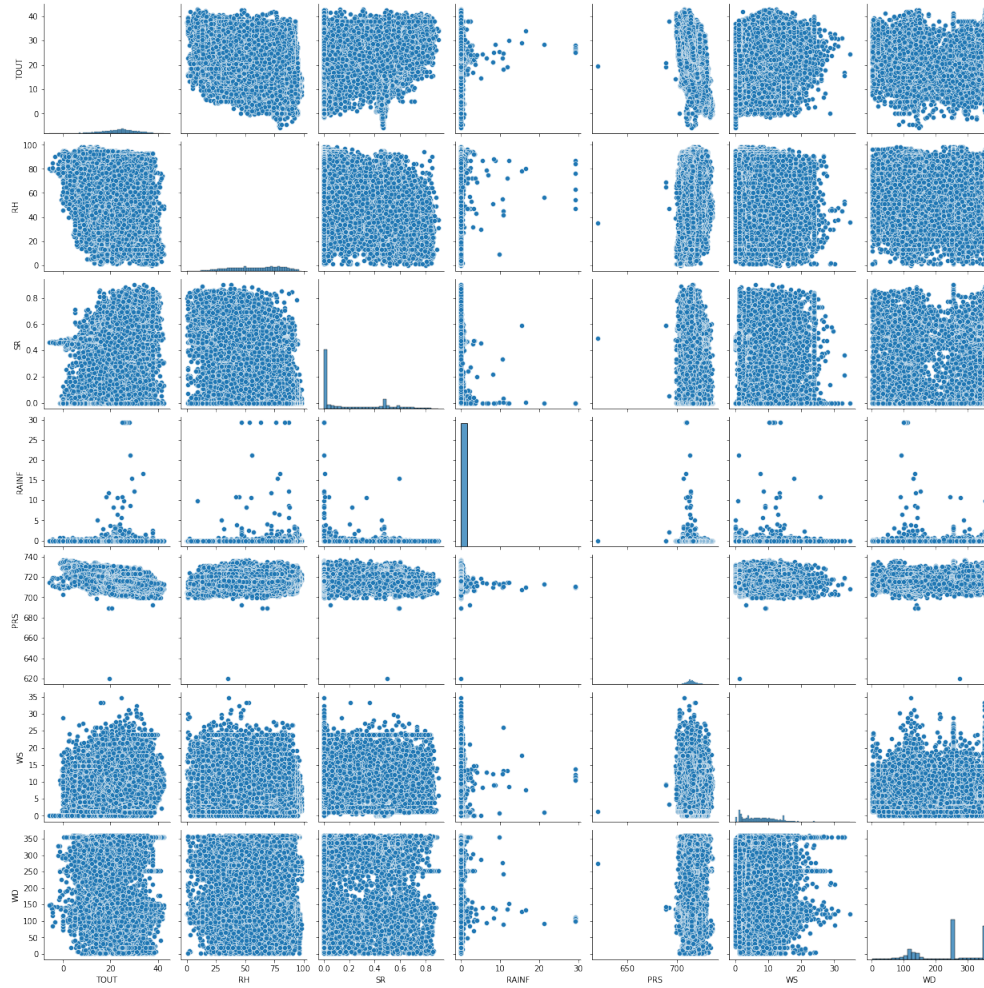


Figura 6: Pairplot de relación entre variables meteorológicas

En estas gráficas se puede ver porque no hay correlación entre las variables en la figura anterior (5), pues ninguna de las gráficas muestra formas definidas que indiquen alguna clase de relación.

2.1.2. Contaminantes:

- Tamaño de las hojas en la base de contaminantes: (39393, 28) (39394, 28) (39392, 28) (39393, 28) (39375, 28) (39393, 28) (39392, 28) (39392, 28)
- Descripción de variables:
 - PM10: Es una variable que mide el material particular menor a 10 micrómetros, se mide en microgramos por metro cúbico y son datos tipo Int.
 - PM2.5: Es una variable que mide el material particular menor a 2.5 micrómetros, se mide en microgramos por metro cubico y son datos tipo Int.
 - O3: Es una variable que describe el ozono, se mide en partes por billón y es un dato tipo Int.
 - SO2: Es una variable que describe el Dióxido de Azufre, se mide en partes por billón y es un dato tipo Float64.
 - NO2: Es una variable que describe el Dióxido de Nitrógeno, se mide en partes por billón y es un dato tipo Float64.
 - CO: Es la variable que describe el Monóxido de Carbono, se mide en partes por millón y es de tipo Float64.
 - NO: Es una variable que representa el Monóxido de Nitrógeno, se mide en partes por billón y es de tipo Float64.
 - NOx: Es la variable que representa el Óxido de Nitrógeno, se mide en partes por billón y es un dato de tipo Float64.
- Medidas estadística (variables cuantitativas):

	PM10	PM2.5	O3	SO2	CO	NO	NO2	NOx
count	39394.000000	39394.000000	39394.000000	39394.000000	39394.000000	39394.000000	39394.000000	39394.000000
mean	58.339006	23.217709	24.208135	9.125306	1.438258	16.066144	19.041612	31.668545
std	37.524874	14.567340	4.361973	5.305450	0.843124	9.527184	2.569077	10.533921
min	2.000000	2.000000	0.000000	0.500000	0.050000	0.100000	0.400000	0.100000
25%	35.000000	15.000000	24.208135	9.125306	0.770000	16.066144	19.041612	31.668545
50%	53.000000	23.217709	24.208135	9.125306	1.438258	16.066144	19.041612	31.668545
75%	71.000000	24.000000	24.208135	9.125306	1.770000	16.066144	19.041612	31.668545
max	687.000000	546.000000	170.000000	164.500000	12.540000	500.000000	74.800000	500.000000

Figura 7: Exploración de datos

Aquí se puede observar un primer entendimiento de las variables de contaminantes con los que estaremos trabajando. Estas son las medidas estadísticas antes de la limpieza de los datos, pero nos ayudan a saber en donde se encuentra nuestra estación (Apodaca) aproximadamente. Se observan las medidas de tendencia central y de dispersión para las variables contaminantes.

- Medidas estadística (variables cualitativas): No se utilizaron variables cualitativas porque no habían en la base de datos. Los únicos registros de valores no numéricos indican anotaciones

sobre los registros como la validez de la hora de registro y qué tipo de registro (falla eléctrica, calibración, apagado, malas condiciones). Estos datos son necesarios para poder enlazarlos con los meteorológicos, ya que es el único dato que sirve como parámetro para poder relacionar ambas bases de datos.

- Calidad de los datos: Primero se realizaron diagramas de caja y bigotes para ver la cantidad y distribución de los datos atípicos.

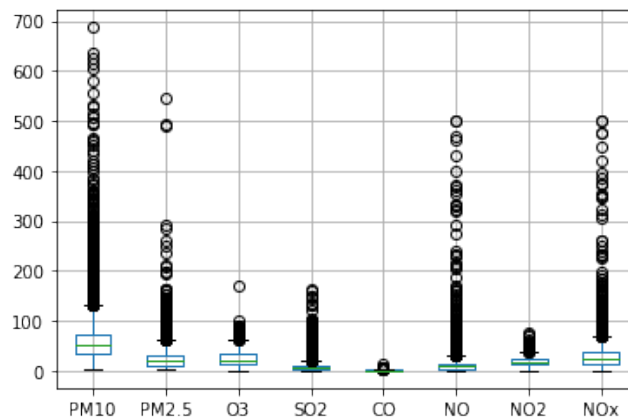


Figura 8: Box Plot de variables de Contaminantes

Se observa que existen muchos valores atípicos para todas las variables, el contaminante PM10 cuenta con el BoxPlot de mayor tamaño. Pero se puede ver que los datos atípicos no es uno solo lejos de los demás, sino que progresivamente hay menos datos mayores. También se puede ver que no hay datos atípicos en los mínimos, a diferencia de lo que se observaba en la velocidad del viento.

- Calidad de los datos: También se utilizaron las gráficas de dispersión para ver la distribución de datos en el tiempo y ver la calidad.

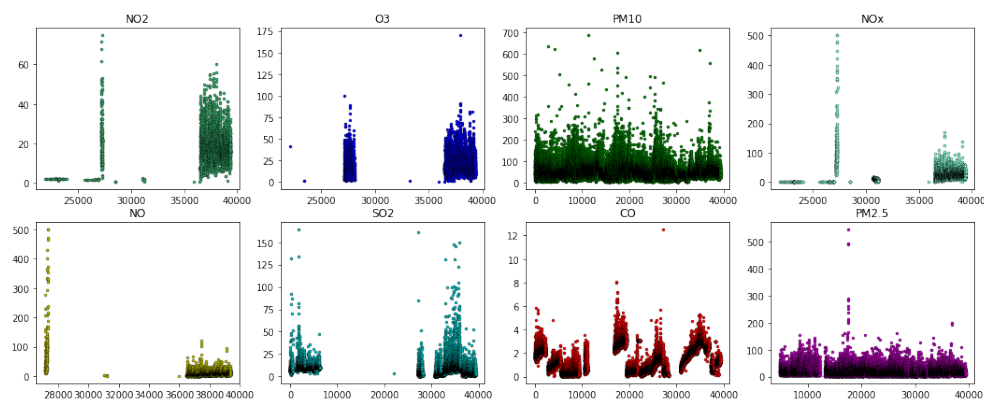


Figura 9: Subplot de Contaminantes

Se puede observar que hay mucho espacio vacio para la mayoría de los contaminantes, pero que todas las variables tienen datos que se ven completos en los años más recientes. A partir de esto vamos más adelante a utilizar únicamente datos del 2019 en adelante, donde tenemos una gran cantidad de datos significativos. También podemos ver que PM10 y PM2.5 son las variables con datos más completos, y dada su importancia en la contaminación y esto, serán de las variables de contaminantes en las que nos enfocaremos.

- Correlación entre las variables: Visualizamos la correlación entre los datos con un mapa de calor.

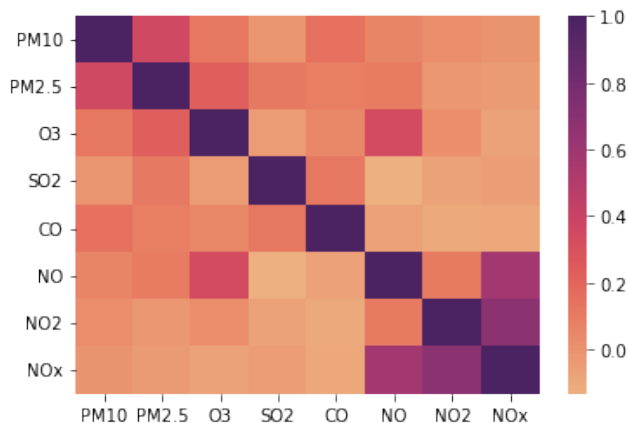


Figura 10: Mapa de calor de la matriz de correlación de los valores de contaminantes

Podemos ver que no hay una relación muy fuerte y significativa entre la mayoría de los contaminantes. Algunas de las relaciones que vemos son entre el monóxido de nitrógeno (NO) y los óxidos de nitrógeno (NOx) que tienen mucha relación, también vemos que si aumenta

el ozono disminuyen los NOx porque se consumen, NOx está relacionado con el uso de los vehículos, el ozono aumenta con el sol y el calor y esto depende tanto del día de la semana como de la hora del día, entre otros factores.

- Correlación entre las variables: Cada uno de los cuadrantes del mapa de calor se puede visualizar como su propio diagrama de dispersión, o en el caso de los 1 comparando con si mismo podemos ver un hisograma, esto se muestra en la siguiente figura.

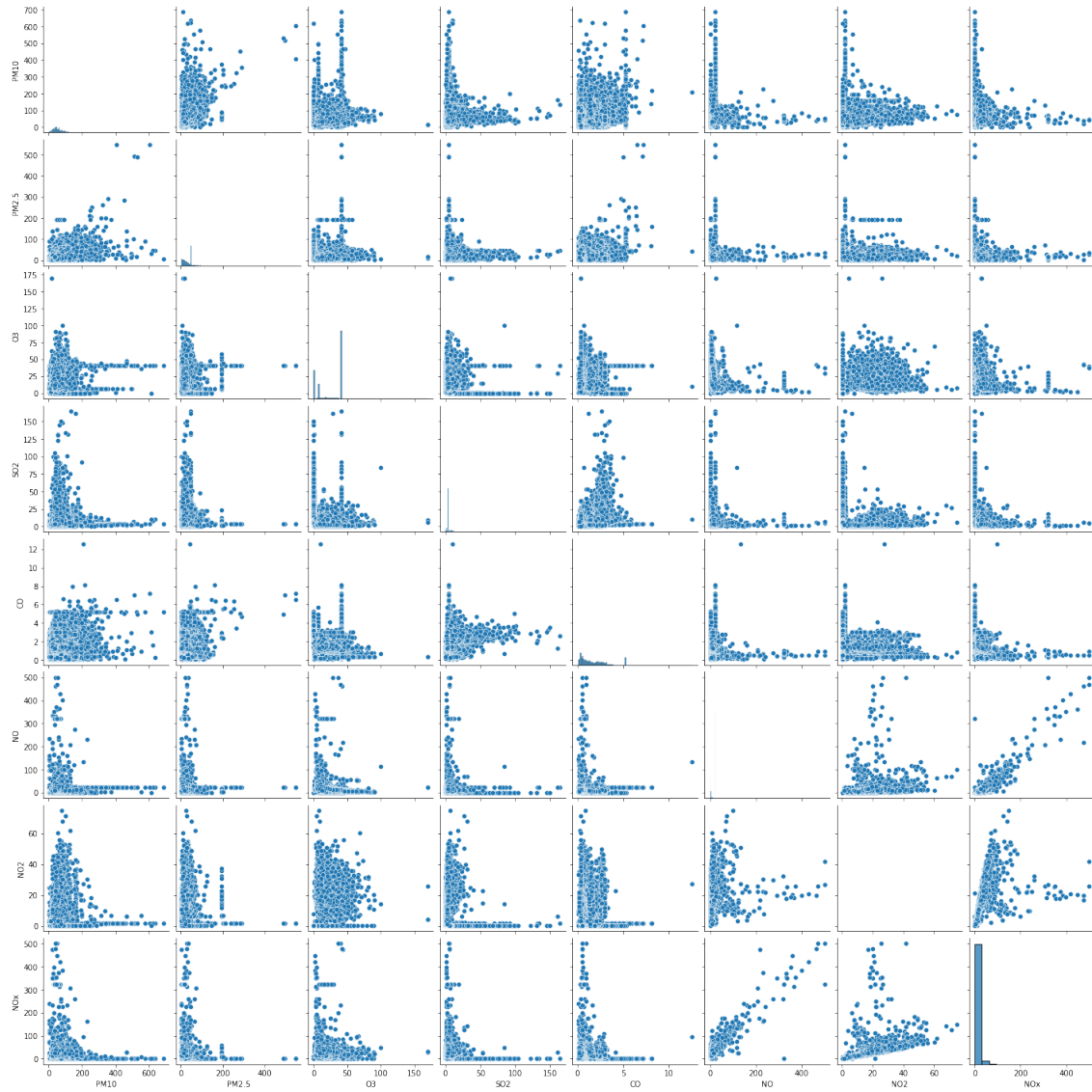


Figura 11: Pairplot de relación entre contaminantes

Se dibuja una gráfica de distribución univariada para mostrar la distribución marginal de los datos en cada columna. Nuevamente observando que no se ve ninguna forma que indique relación entre variables, excepto por NO y NOx.

2.2. Preparación de los datos

Debido a que la estación que se seleccionó fue la Noreste 2: Apodaca, Apodaca (NE2), se pudo descartar la base de datos “BD Tec Banderas 2018 2021 3Estacionesza que no contenía información de la variable elegida. Las dos bases restantes si tienen a la estación contemplada y sus registros.

2.3. Columna objetivo

Al analizar el problema de la calidad del aire, las columnas objetivo que se escogieron es la PM10, PM2.5, SO₂, las cuales son partículas relacionadas con la gran variedad de problemas de salud, además que dichos contaminantes son los que mas tienen presencia en el tiempo establecido; el objetivo es considerar las columnas objetivo con la estación Noroeste 2; este análisis se considera con la base de datos “Contaminantes”, y es que en esa base de datos solo se considera la columna NE2; por otra parte, en la base de datos “meteo”, se van a considerar todas las variables (TOUT, RH, SR, RAINF, PRS, WS, WD) para poder determinar las relaciones que existen con la estación.

2.4. Limpieza de datos

- Para eliminar los datos duplicados se utilizó la función de `drop_duplicates` de la librería de Pandas de Python, con la cual sí había instancias repetidas se eliminaban.
- Se revisaron los valores esperados y viables de cada una de las variables meteorológicas para ver si no habían datos erróneos, revisando los estadísticos del cuadro 2.
 - Sí hay temperaturas negativas, en Apodaca sí se pueden esperar variaciones de entre -5° y 43°.
 - La humedad se mide en un porcentaje por lo que 0 % y 98 % son los mínimos y máximos adecuados.
 - La radiación solar máxima es de $1kW/m^2$ por lo que valores entre 0 y 0.9 son válidos.
 - La precipitación no es un fenómeno de mucha frecuencia en Apodaca por lo que la mayoría de los estadísticos siendo de valor 0 son razonables, sin embargo, cuando han habido huracanes, sí llueve lo suficiente para generar ese máximo.
 - La presión a nivel del mar es 760 mmhg en promedio, a mayor altura menor presión, y como Monterrey no está a nivel del mar si no a mayor altitud, la presión no puede ser mayor a 760 mmhg, por lo tanto, tiene valores adecuados.
 - La velocidad del viento no puede tener valores negativos, por lo que -9999 es un valor erróneo que se corrigió durante la limpieza de datos, el máximo es un valor razonable.
 - La dirección del viento son en grados por lo que 1° a 360° es válido.

- Al haber valores faltantes, se optó por usar la técnica de backfilling, la cual le asigna el siguiente valor de la columna al dato faltante, de esta manera la periodicidad que se observa en los datos meteorológicos va a poder continuar de manera más real, ya que la media no era viable por la periodicidad de los datos meteorológicos y por la desviación estándar de los datos de los contaminantes.
- No se realizó un manejo de variables cualitativas, pues todas las variables que estaremos utilizando, como se mencionó al principio son cuantitativas. La única variable que no es manejada igual es la de Fecha, sin embargo, es necesario tener el valor exacto para poder enlazar ambas bases de datos, ya que el objetivo del reto, es poder identificar qué variables meteorológicas, se relacionan con el contaminante Pm10 en la estación NE2, y cómo esto varía con el tiempo.
- También se hizo una limpieza de acuerdo a las fechas, como se vio en la Figura 9, hay muchos datos faltantes antes del 2019, y como el metodo de backfilling no iba a proporcionar datos significantes en esos huecos, se optó por unicamente utilizar los datos a partir del 2019. Los nuevos subplots de las variables con la limpieza mencionada anteriormente se muestran a continuación: 12, 13.

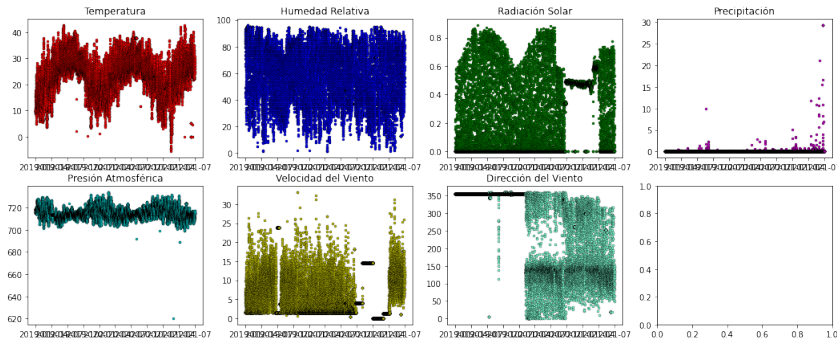


Figura 12: Subplot de datos meteorológicos limpios

- Con esta limpieza nuestro nuevo data frame tiene 10 variables más la fecha en la que se tomaron las medidas y son 21897 instancias.

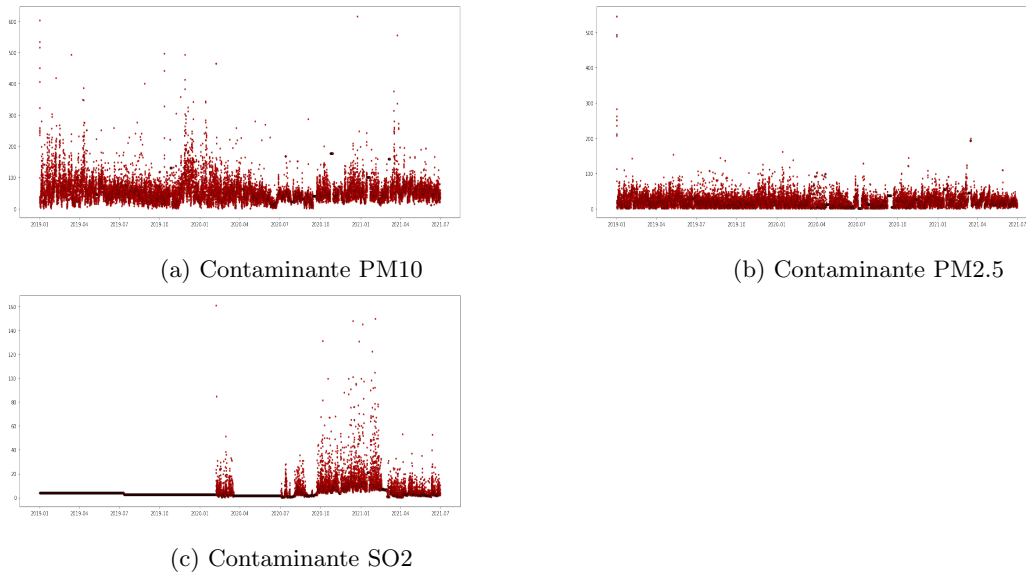


Figura 13: Subplot de variables de contaminantes limpios

En la gráfica 13c se ve bastante espacio vacío, pues este contaminante no tiene registros en 2019, y después de iniciar la pandemia en marzo 2020, parece que se dejaron de tomar registros. Sin embargo, este contaminante está muy relacionado con las fábricas y producción que hay en la región de Apodaca, por lo que aunque no hay tantos datos anteriores, nos pareció importante incluirlo en nuestros análisis.

- El código donde se llevó a cabo el proceso de limpieza se puede revisar en el siguiente enlace: [Link](#)

2.5. Exploración de los datos

Una vez que tuvimos nuestros datos limpios y seleccionados, se realizaron pruebas para ver el comportamiento de los datos y poder determinar que tipos de análisis eran viables. Se realizaron pruebas de Mardia para ver el sesgo y curtosis de los datos, y encontramos que no había normalidad en los datos, eliminando los posibles análisis que asumen normalidad en los datos, ya que preferimos enfocarnos en análisis para datos no normales que en normalizar nuestros datos, para mantener algunos componentes de los datos que se perderían con la transformación. También se realizaron pruebas de Kaiser–Meyer–Olkin para ver que tan viable era una prueba de análisis factorial. Inicialmente se hicieron las pruebas por separado, para los datos meteorológicos y para los datos de contaminantes y dio un resultado de miserable. Pero tras la selección de variables y la unión de los data sets, la prueba KMO llegó a mejores resultados, por lo que procedimos con los análisis factoriales, los cuales requieren un parámetro de rotación, al realizar las gráficas se decidió que el mejor parámetro es el *quartimax* ya que minimiza el número de factores necesarios para explicar cada variable, de igual manera requiere un parámetro de método, el cual es el del eje principal

”pa” que hará la solución del factor principal. Estos parámetros se utilizarán para ambos análisis factoriales.

3. Adecuación y/o validación del Modelo

3.1. Análisis Interdependiente

3.1.1. Análisis de conglomerados

Se realizó un análisis de conglomerados por medio del método KMeans, para realizar dicho método, primero se necesita saber con cuántos clusters se va a trabajar, esto es una de las desventajas, ya que no se conoce con certeza, por lo que se optó por realizar un gráfico de codo, la cual se observa en la Figura 12, se opta por trabajar con 2 clusters debido a que se observa el punto de codo donde la tasa de ascenso se afila en ese punto de la gráfica.

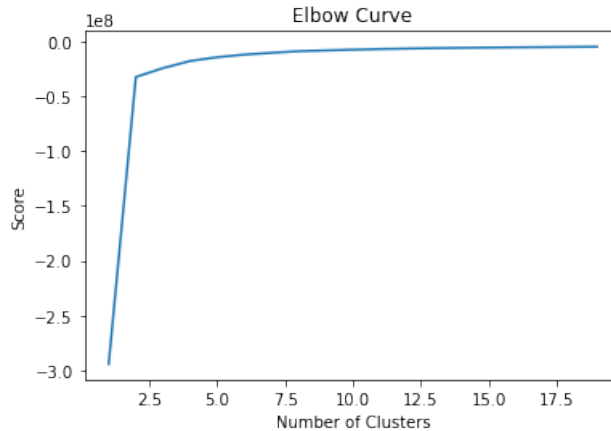


Figura 14: Curva de codo

Posteriormente, se ejecuta el algoritmo KMeans para 2 clusters donde se obtienen los centroides, que permiten conocer el punto equidistante de los objetos pertenecientes a dicho cluster y se procede a realizar la gráfica, la cuál se puede observar en la Figura 13, se observa mucho agrupamiento entre los 2 clusters, por lo que es complicado encontrar una relación entre las variables dependientes y las meteorológicas; al obtener estos resultados, se realizó la validación de la eficiencia del modelo KMeans, el cuál da como resultado 11.6 % además, se obtiene que 138,740 datos se encuentran en el primer cluster y 54,793,89 en el segundo clúster. Derivado de estos resultados, se opta por descartar este método y realizar el análisis factorial.

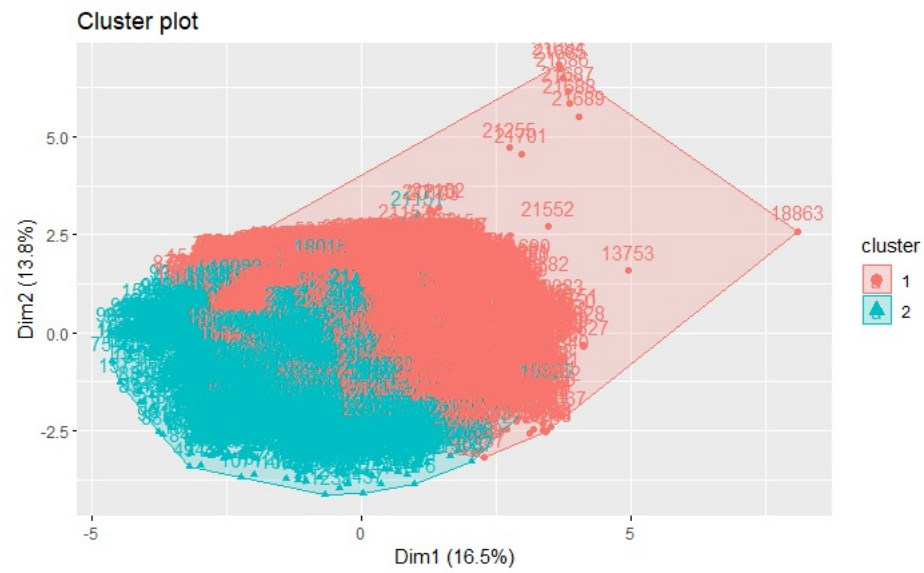


Figura 15: Cluster plot con R

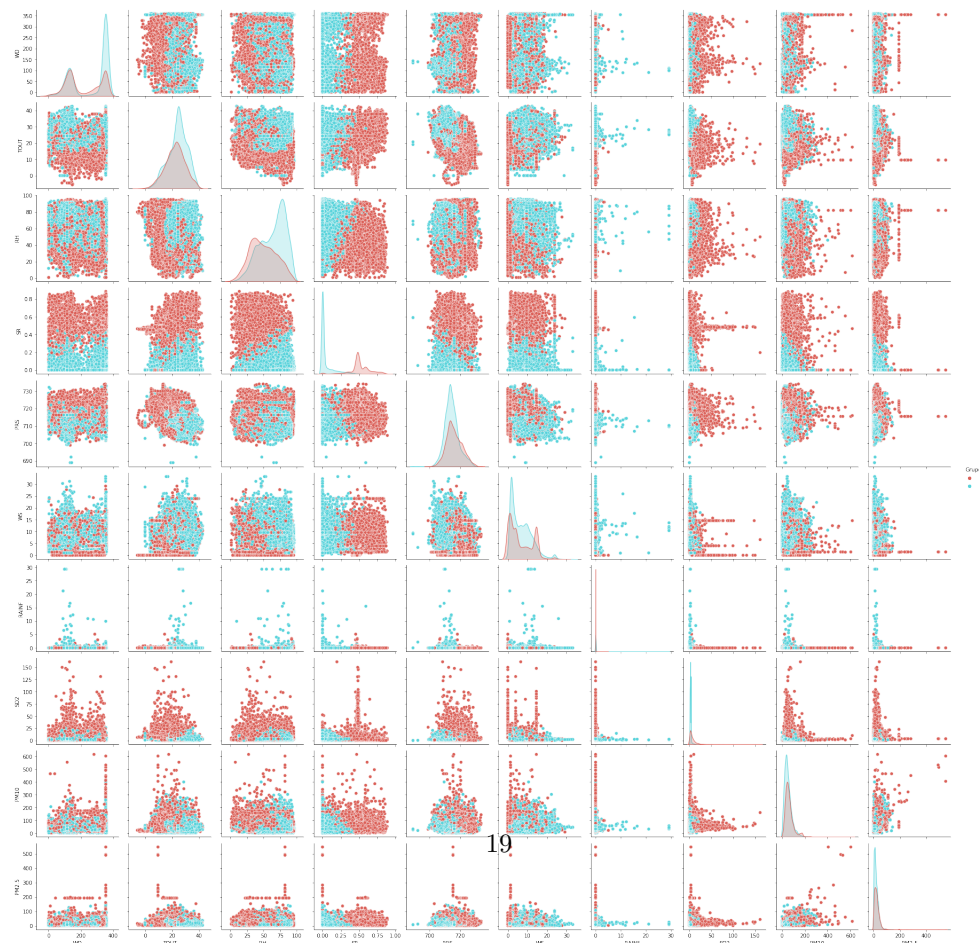


Figura 16: Pairplot con grupos por K-medias

En el pairplot podemos observar la relación por análisis conglomerados entre todas las variables. Podemos asumir relaciones por estas agrupaciones, como por ejemplo la relación entre los 3 contaminantes y la cantidad de lluvia: entre menos lluvia haya, mayor cantidad de contaminantes. Por otro lado, también podemos observar como no hay relación entre ciertas variables. Como el PM10 con las variables meteorológicas ya que no se observan cambios significativos con la cantidad de ninguna de las dos variables. Viendo las relaciones unicamente en gráficas de contaminante con variable meteorológica podemos observar que el grupo azul corresponde con un menor cantidad de contaminante, mientras que el grupo rojo indica mayor nivel de contaminante.

3.1.2. Análisis de Factores utilizando variables binarias

La base de datos cuenta con variables cuantitativas, que son numéricas continuas, pero ninguna de ellas se distribuye de manera normal, para manejar los datos de una manera diferente y buscar otro análisis, se decidió crear una variable dependiente de los contaminantes a partir de las Normas Oficiales Mexicanas (NOM) de Calidad del Aire Ambiente. Esto se logró realizando una transformación de la información numérica a binaria, comparando las concentraciones diarias de los contaminantes de los años 2019 a 2021 contra las emisiones permitidas bajo las regulaciones del gobierno. La clasificación binaria registra como 1 los valores que caen dentro del rango permitido y 0 los valores que caen fuera del mismo rango. A continuación se describen los contaminantes con sus parámetros permitidos de concentración en el aire:

$$PM_{2,5} < 45\mu g/m^3$$

$$PM_{10} < 75\mu g/m^3$$

$$SO_2 < 1,1ppm$$

Una vez teniendo las nuevas variables dependientes binarias, se realizó una matriz de correlación policórica (Figura 17) con la finalidad de poder realizar un test de Bartlett para verificar que existe suficiente correlación significativa para poder realizar un análisis de factores, el cual se puede ver confirmado en la Figura 18.

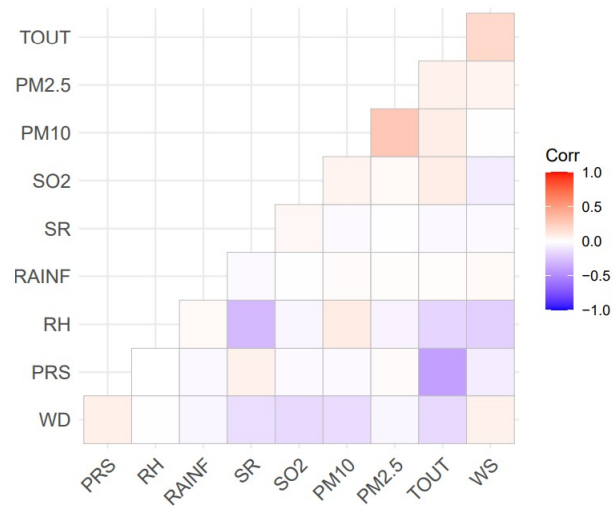


Figura 17: Matriz de correlación

```
# Test of Sphericity

Bartlett's test of sphericity suggests that there is sufficient significant correlation in the data for
factor analysis (Chisq(45) = 4261.76, p < .001).
```

Figura 18: Test de Bartlett

Una vez contando con el test de Bartlett, que indica que si hay suficiente significancia, se realizó una gráfica de scree (19) para mostrar cuántos factores se necesitan para el análisis de máxima versimilitud y análisis de factores principales, que en este caso fue 1 factor el que se plotea por arriba de la línea marcada (Figura 20).

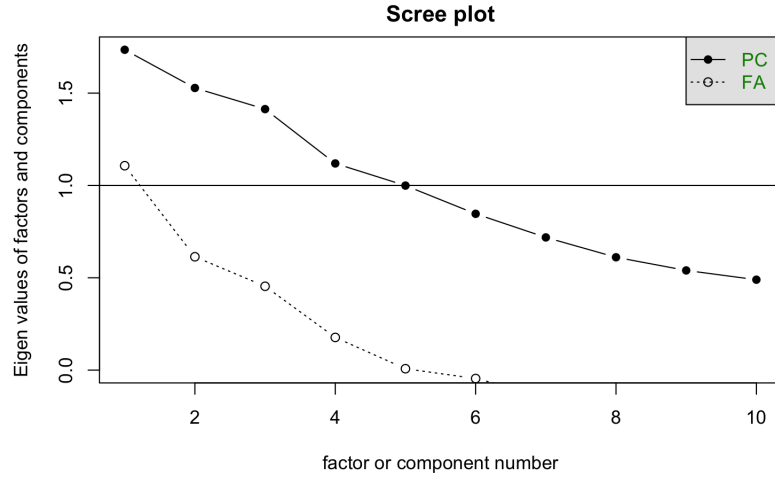


Figura 19: Gráfico de Sedimentación

	V <S3: AsIs>	PA1 <S3: AsIs>	h2 <dbl>	u2 <dbl>	com <dbl>
TOUT	2	0.87	0.7497014893	0.2502985	1
PRS	5	-0.43	0.1877200058	0.8122800	1
WD	1	-0.22	0.0505322013	0.9494678	1
WS	6	0.18	0.0330312634	0.9669687	1
PM10	9	0.18	0.0327509175	0.9672491	1
PM2.5	10	0.14	0.0197033233	0.9802967	1
SO2	8	0.13	0.0174450883	0.9825549	1
RH	3	-0.13	0.0164720862	0.9835279	1
RAINF	7	0.04	0.0014333148	0.9985667	1
SR	4	0.02	0.0004308373	0.9995692	1

Figura 20: Análisis de factores, comunales (h2) y especificidad (u2)

Al realizar el análisis, se obtuvieron cuatro componentes principales, se obtuvo que la variable de temperatura es la de mayor peso seguida de la presión atmosférica, se observa que las variables meteorológicas son las de mayor peso, la comunalidad más alta es la de la temperatura lo cual indica que es la única variable que está mayormente explicada por el factor.

3.1.3. Análisis de Factores utilizando variables continuas

Se hizo un segundo análisis de factores, pero esta vez sin usar variables binarias para hacer un análisis interdependiente. Dado a la elección de variables se generó un data frame con el fin de poder encontrar ciertos patrones de comportamiento dentro de nuestras variables elegidas. Como primera instancia se generó una matriz de correlación para poder observar de una manera gráfica la existencia de correlación en las variables.

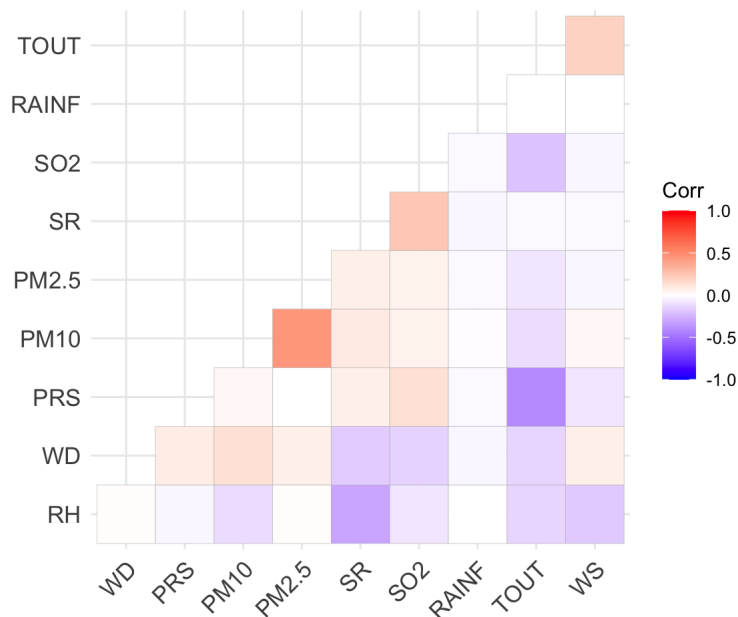


Figura 21: Matriz de correlación

Una vez con la gráfica de correlación y ver que había cierta correlación en las variables pero no de una manera gráfica fuerte se decidió hacer un test de Bartlett con el fin de saber si la correlación entre las variables era significativa y suficiente para que los resultados del análisis den información relevante. Una vez con el resultado se puede corroborar que existe suficiente correlación en las variables.

```
# Test of Sphericity

Bartlett's test of sphericity suggests that there is sufficient significant correlation in the data for factor analysis (Chisq(45) = 5827.30, p < .001).
```

Figura 22: Test de Bartlett

Después de tener el test de Bartlett, que determinó que sí hay suficiente relación entre las variables (22), se realizó una gráfica de scree (23) para mostrar cuántos factores se necesitan para el análisis de máxima verosimilitud y análisis de factores principales, que en este caso fue 1 los factores que se plantean por arriba de la línea marcada en uno.

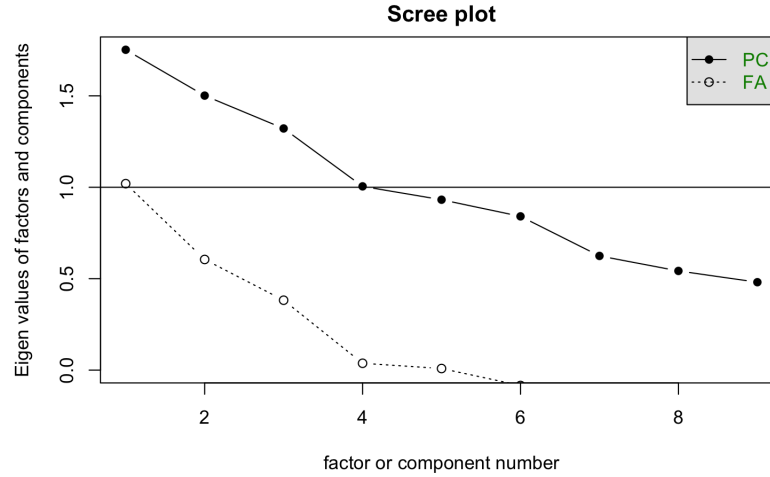


Figura 23: Gráfico de Sedimentación

	V <S3: AsIs>	PA1 <S3: AsIs>	h2 <dbl>	u2 <dbl>	com <dbl>
TOUT	1	-0.72	0.511716087	0.4882839	1
PRS	4	0.47	0.219136324	0.7808637	1
SO2	7	0.30	0.091490097	0.9085099	1
PM10	8	0.25	0.062051826	0.9379482	1
PM2.5	9	0.25	0.060339260	0.9396607	1
WS	5	-0.20	0.041127956	0.9588720	1
SR	3	0.15	0.023832531	0.9761675	1
RAINF	6	-0.04	0.001566625	0.9984334	1
RH	2	0.04	0.001400813	0.9985992	1

Figura 24: Análisis de factores, componentes y pesos

Los resultados del análisis indican que se tienen cuatro componentes principales, se puede observar que la variable principal y con mayor peso es la de temperatura, seguida por la de radiación solar, después la de dirección del viento y por último la variable de velocidad del viento. Lo interesante es que los componentes principales del análisis factorial son únicamente variables meteorológicas. Cabe mencionar que la única variable que tiene una comunalidad arriba del 0.70 es la de temperatura, lo cual indica que es la única variable que está mayormente explicada por el factor. De igual manera, la especificidad al ser contraria a la comunalidad y explicar la parte de la varianza relacionada a los factores únicos o específicos, tiene valores muy altos en todas las variables a excepción de la variable de temperatura.

4. Conclusiones

Recapitulando el proyecto, primero se exploraron las bases de datos, se planteó la pregunta objetivo que busca encontrar la relación entre los contaminantes y los datos meteorológicos de la zona de Apodaca, la cual fue la base para escoger las variables a analizar.

Una vez contando con esas variables limpias, y sabiendo qué relación se busca hallar, fue más sencillo plantear la etapa del modelaje, en la que se realizó el análisis de conglomerados y el análisis factorial para ambos casos, continuo y binario. Y con los resultados arrojados para cada análisis se concluyó que en la matriz de correlación del análisis utilizando variables binarias y en la matriz de correlación del análisis utilizando variables continuas se presentan 2 pares de relaciones fuertes en común entre los contaminantes PM10 y PM2.5, y las variables meteorológicas de temperatura y velocidad del viento; con la diferencia de que además en el análisis de variables continuas se encontró una tercera fuerte relación entre el dióxido de azufre y la radiación solar. Además del análisis de conglomerados donde se encontró una fuerte relación entre la concentración de contaminación con los días en los que existe una precipitación. Cabe mencionar que a pesar de no contar con la mitad de los datos del año 2021 y los datos del 2022, con el análisis realizado es posible predecir que los meses con más concentración de contaminación serán los meses de abril a julio, debido a las altas temperaturas, sequías, falta de humedad y obviamente el gran problema de la falta de lluvias y agua en todo el estado de Nuevo León.

Como conclusión, con estas relaciones, se pueden definir las bases causales de las concentraciones de contaminación en la región de Apodaca, y ésta información a su vez queda abierta a poder ser utilizada para analizar más a fondo los principales factores de contaminación causados por la urbanización de la zona, hablese del tráfico, zona industrial, cambio climático, entre otros.

Referencias

- de la Salus, O. M. (2018). *Organización Panamericana de la Salud*. <https://www.paho.org/es/temas/calidad-aire-salud/contaminacion-aire-ambiental-exterior-vivienda-preguntas-frecuentes>
- Matus, P., & LUCERO CH, R. (2002). Norma primaria de calidad del aire. *Revista chilena de enfermedades respiratorias*, 18(2), 112-122.
- Nations, U. (2018). *Department of Economic and Social Affairs Population Dynamics*. <https://population.un.org/wup/>
- Normas Oficiales Mexicanas (NOM) de Calidad del Aire Ambiente. (2017). *Comisión Federal para la Protección contra Riesgos Sanitarios*.
- Palou, S. (2017). ¿Cómo se mide la calidad del aire. *SP*.
- SIMA. (2015). *Sistema Integral de Monitoreo Ambiental*.