Tecnológico de Monterrey

Etapa 2. Conociendo los Datos

Aplicación de métodos multivariados en ciencia de datos Gpo(301)

Docentes:

Blanca Rosa Ruíz Hernández

Raul Gómez Muñoz

Miembros de equipo:

Santiago Tolosa Torres A00838869

David de la Cerda Naime A01571649

Francisco Moreno Alcocer A00838723

Matías Gil Abadie A01723409

Fecha: 20 de Noviembre de 2025

Introducción

Los festivales generan importantes beneficios económicos, sociales y culturales, pero también producen impactos ambientales que suelen ser menos atendidos. Este análisis busca identificar y evaluar las afectaciones ambientales ocurridas durante los días del festival Pal-Norte, utilizando métodos de análisis multivariado para obtener resultados sólidos. Para ello se seleccionaron variables clave relacionadas con la calidad del aire y factores meteorológicos que influyen en la dispersión de contaminantes, con el fin de establecer una base de referencia y permitir comparaciones con otros eventos similares.

El objetivo de este análisis es identificar y demostrar las afectaciones ambientales concentradas específicamente en los días en que se realiza el festival. Con ello buscamos establecer una base de referencia que permita no solo evaluar este evento en particular, sino también generar informes comparativos sobre otros festivales similares y sus respectivas repercusiones. Para lograrlo, emplearemos distintos métodos de análisis multivariado, ya que la complejidad del fenómeno requiere un enfoque robusto y riguroso para alcanzar resultados confiables.

La entrega anterior se centró en la limpieza de los datos con el fin de seleccionar las variables más relevantes para el análisis. Estas variables fueron elegidas para evaluar el impacto ambiental del festival Pal-Norte, ya que influyen directamente en la calidad del aire y la salud humana, especialmente durante eventos masivos.

Después de un proceso detallado de investigación y análisis, las variables seleccionadas que son necesarias para completar nuestro objetivo son las siguientes:

- **PM10:** Variable numérica que representa partículas con diámetro aerodinámico ≤ 10 μm. Su exposición se asocia con afectaciones respiratorias y cardiovasculares. Se generan principalmente por emisiones vehiculares, resuspensión de polvo, quema de combustibles y actividades de construcción
- **PM2.5:** Variable numérica correspondiente a partículas finas con diámetro ≤ 2.5 μm. Debido a su alta penetrabilidad pulmonar, están vinculadas con enfermedades respiratorias y cardiovasculares graves, así como con un mayor riesgo de cáncer pulmonar. Sus fuentes principales incluyen emisiones vehiculares, combustión de biomasa, procesos industriales y reacciones atmosféricas secundarias.
- **CO:** Variable numérica asociada al monóxido de carbono, gas tóxico producto de la combustión incompleta de combustibles fósiles o biomasa. Inhibe la capacidad de la sangre para transportar oxígeno y puede causar efectos letales a concentraciones elevadas. Las principales fuentes son el transporte motorizado y ciertas actividades industriales.
- **NOx:** Variable numérica que agrupa óxidos de nitrógeno (principalmente NO y NO2), emitidos por procesos de combustión a alta temperatura. Actúan como irritantes

respiratorios y contribuyen al empeoramiento de enfermedades pulmonares crónicas. Su origen predominante es el parque vehicular.
- **O3:** Variable numérica correspondiente al ozono troposférico, contaminante secundario formado por reacciones fotoquímicas entre NOx y COV en presencia de radiación solar. Su exposición provoca irritación respiratoria y agrava enfermedades como asma y bronquitis.

Con las variables de estudio definidas, es necesario identificar las zonas donde analizaremos su comportamiento y efectos. El festival *Pal-Norte* se lleva a cabo cada año en el mismo lugar: el Parque Fundidora, ubicado en el centro de Monterrey. Con base en ello, se seleccionaron las áreas que rodean la sede del evento, ya que son las que presentan la mayor probabilidad de afectación directa.

Las zonas consideradas son las siguientes:

1. **ZONA CENTRO:** Fundidora directamente.
2. **ZONA NORESTE**: Este de fundidora y abarca parte del municipio de Guadalupe. Ahí es donde se desplazan los vientos.
3. **ZONA NORESTE2**: Más oriental del noreste.
4. **ZONA SUR**: Cercana al corredor de Constitución-Morones Prieto que conecta con el sur.
5. **ZONA SURESTE2**: Lado más cercano a Guadalupe-Contry-Revolución.

Teniendo esto definido, es posible proceder a trabajar directamente con los datos.

Comprensión de los datos

La primera fase de esta entrega consiste en el análisis directo de nuestros datos con las variables y las zonas elegidas. Esto nos abre la oportunidad para comenzar con nuestro análisis de manera centrada.

Comenzamos realizando un *merge* de siete bases de datos distintas, cada una correspondiente a un periodo diferente entre 2020 y 2025. Es importante señalar que todas presentaban un orden distinto tanto en columnas como en filas. Unificarlas en una sola estructura nos permitió obtener una visualización más coherente y continua, iniciando en 2020 y finalizando en 2025. Además, reorganizamos las columnas para que coincidieran entre sí según las zonas, evitando así tener que desplazarse entre columnas diferentes para localizar el mismo concepto.

Habiendo definido nuestras variables de tipo numérico, lo que permite calcular diversas métricas matemáticas a partir de ellas, procedimos a filtrar la nueva base de datos. En esta limpieza eliminamos las zonas que no forman parte del análisis y las variables que no utilizaremos en esta primera etapa. Esto no implica que más adelante no puedan agregarse o retirarse variables;

simplemente, para esta dimensión del problema trabajaremos únicamente con las previamente seleccionadas.

Después de este proceso, se realizó un conteo de todos los valores nulos presentes en la base de datos con los que trabajaríamos durante la exploración. En total, se identificaron 233,812 valores faltantes, una cantidad considerable que podría comprometer la precisión de los resultados. Por ello, será necesario aplicar un método de imputación de datos en etapas posteriores del análisis.

Exploración de los datos

Tras seleccionar y limpiar las variables, el uso de medidas estadísticas es clave para comprender el comportamiento ambiental antes, durante y después del festival. Las medidas de tendencia central permiten establecer valores de referencia y detectar aumentos en contaminantes como PM10, PM2.5, CO, NOX y O3 durante el evento. Las medidas de dispersión ayudan a identificar variabilidad y posibles picos irregulares asociados a actividades del festival.

Los gráficos como histogramas y boxplots facilitan visualizar distribuciones y outliers, mientras que para variables cualitativas, como la dirección del viento, las tablas de frecuencia permiten detectar patrones. Finalmente, la correlación y los mapas de calor muestran cómo interactúan los contaminantes y las condiciones meteorológicas, revelando cambios relevantes durante los días del evento.

Análisis de los datos

El análisis descriptivo de las 25 variables cuantitativas correspondientes a los cinco contaminantes en cinco zonas del área metropolitana muestra patrones consistentes en todas las áreas: valores promedio moderados, medianas más bajas que las medias y rangos extremadamente amplios. Esto indica que los contaminantes presentan comportamientos altamente variables, con frecuentes picos extremos de concentración. Estos picos elevan la media pero no modifican tanto la mediana, mostrando que las concentraciones elevadas no son constantes, sino episodios puntuales.

En todas las zonas, los contaminantes PM10 y PM2.5 destacan por tener las medianas más altas y los rangos más amplios, lo que revela una presencia constante y significativa de partículas suspendidas en el aire. Su alta desviación estándar confirma que las partículas presentan fluctuaciones fuertes, posiblemente influenciadas por condiciones climáticas, movilidad urbana y eventos masivos como el festival.

Los contaminantes NOX y CO también muestran rangos muy elevados. Esto sugiere que las zonas experimentan variaciones importantes en emisiones de combustión, principalmente asociadas a tráfico vehicular y otras fuentes móviles. En particular, NOX presenta las variaciones

más extremas de todo el dataset, con rangos superiores a 300–450 unidades en casi todas las zonas, lo que evidencia episodios de contaminación altamente irregulares.

El comportamiento del O3 es distinto, aunque presenta variabilidad considerable, sus promedios y medianas son más estables. Esto concuerda con su naturaleza como contaminante secundario, dependiente de la presencia de NOX, radiación solar y condiciones atmosféricas. En varias zonas, el ozono muestra una mediana muy por debajo del rango total, indicando acumulaciones puntuales de alto impacto.

En términos generales, los resultados confirman que las zonas Centro, Noreste y Noreste2 tienen concentraciones mayores y variabilidad más pronunciada en PM10, PM2.5 y NOX, lo que coincide con su ubicación urbana, tráfico constante y cercanía al área donde se realiza el festival Pal-Norte. Las zonas Sur y Sureste2, aunque presentan valores menores en promedio, también muestran picos importantes, lo que indica que la contaminación se dispersa y puede afectar áreas más lejanas dependiendo del clima y la dirección del viento.

Visualización del análisis

A continuación se presentan cinco tablas que muestran los valores de cada variable por zona. En el apartado anterior ya se explicó el significado e interpretación de cada una de ellas.

Variable	Media	Mediana	Rango	SD	IQR
CENTRO_CO	1.93	1.79	14.55	0.99	1.31
CENTRO_NOX	21.35	14.8	362	20.47	19.2
CENTRO_PM10	59.69	53	733	33.8	36
CENTRO_PM2.5	24.47	21.05	246.93	14.85	18.25
CENTRO_O3	29.27	26	165	20.65	28

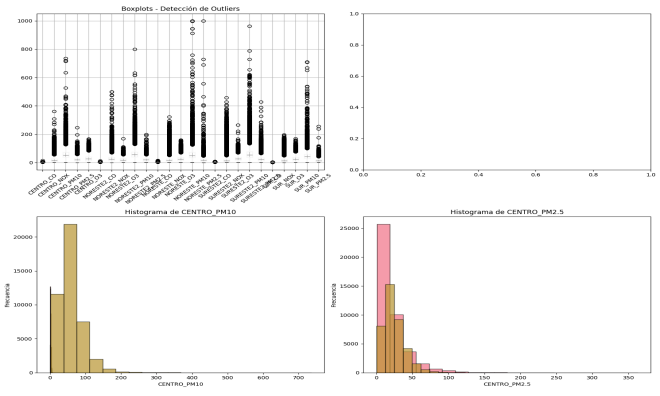
Variable	Media	Mediana	Rango	SD	IQR
NORESTE2_CO	1.74	1.71	12.54	0.71	0.88
NORESTE2_NOX	31.96	25.3	499.5	27.95	22
NORESTE2_O3	21.46	18	169	13.88	18
NORESTE2_PM10	63.06	56	797	38.06	36
NORESTE2_PM2.5	21.84	19	197	14.67	18

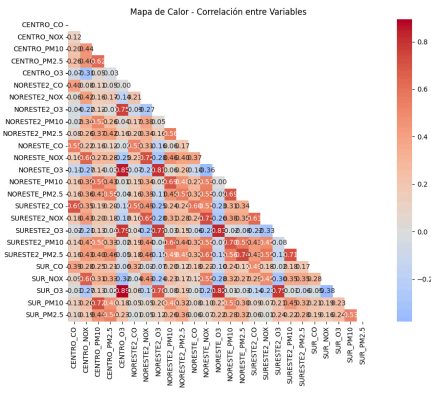
Variable	Media	Mediana	Rango	SD	IQR
NORESTE_CO	1.56	1.32	15.25	0.94	0.99
NORESTE_NOX	25.45	17.8	320.9	26.88	15.8
NORESTE_O3	27.69	25	160	17.41	22
NORESTE_PM10	59.48	51	998	40.84	36
NORESTE_PM2.5	21.43	18	999	16.81	14.56

Variable	Media	Mediana	Rango	SD	IQR
SURESTE2_CO	1.5	1.47	8.25	0.82	1.02
SURESTE2_NOX	25.63	12.8	457.4	36.34	17.3
SURESTE2_O3	29.02	26	264	16.8	21
SURESTE2_PM10	65.12	54	961	45.06	38
SURESTE2_PM2.5	27.32	22	426	20.08	20

Variable	Media	Mediana	Rango	SD	IQR
SUR_CO	1.02	0.85	3.54	0.56	0.87
SUR_NOX	20.94	14.3	196.6	20.69	16.5
SUR_O3	28.7	25	168	19.38	25
SUR_PM10	47.93	42	708	28.43	28
SUR_PM2.5	16.53	14	256	11.83	14

Visualizamos de igual manera las gráficas realizadas con histogramas, boxplots y la matriz de correlación:





Los boxplots muestran que todas las variables presentan una gran cantidad de valores atípicos, lo cual evidencia episodios frecuentes de contaminación elevada en todas las zonas. Las partículas PM10 y PM2.5 son las más variables y las que registran los picos más altos, indicando condiciones críticas de calidad del aire. Los histogramas del área Centro complementan este hallazgo ya que tanto PM10 como PM2.5 tienen distribuciones sesgadas hacia la derecha, con la mayoría de valores en rangos bajos o moderados, pero con colas largas que representan picos extremos de contaminación. Estos resultados confirman que la contaminación atmosférica presenta alta variabilidad y episodios agudos, lo cual es fundamental para comparar los días del festival con días normales y determinar su impacto ambiental.

La matriz de correlación revela que la contaminación atmosférica en Monterrey tiene un comportamiento altamente interdependiente entre zonas. PM10, PM2.5 y NOX muestran correlaciones moderadas y fuertes tanto dentro como fuera de cada zona, lo cual indica que las emisiones locales incluyendo actividades masivas como un festival se dispersan y afectan áreas más amplias. El ozono presenta las correlaciones más altas entre zonas, confirmando su naturaleza como contaminante regional. En conjunto, estas relaciones evidencian que las emisiones generadas durante el festival Pal-Norte podrían tener efectos medibles no solo en Fundidora, sino también en zonas como Noreste, Noreste2, Sureste2 y Sur debido a patrones de viento y reacciones fotoquímicas.

Optimización en la calidad de los datos. El camino a tomar para la mejora del análisis

Hasta ahora hemos integrado todas las bases de datos, realizado la limpieza inicial y ejecutado un primer análisis, lo que confirma que vamos por buen camino. Sin embargo, detectamos una cantidad considerable de valores nulos, por lo que fue necesario aplicar un método de imputación para mejorar la precisión del análisis y evitar pérdida de información.

Para ello utilizamos KNN Imputer, un enfoque multivariante que estima los valores faltantes a partir de los registros más similares, aprovechando las relaciones entre todas las variables. Esto nos permitió obtener una serie continua, coherente y adecuada para análisis estadísticos y modelos predictivos. El hecho de haber realizado esta imputación, nos otorgó un total de 44925 filas, 26 columnas y 1,123,100 valores para realizar futuros cálculos.

Con esto establecido, el siguiente paso es seguir ajustando las variables, refinar los métodos utilizados y avanzar hacia un modelo final más robusto que capture con mayor precisión el comportamiento real de los datos.

Repositorio de github: GitHub