# UNIVERSIDAD NACIONAL DE SAN AGUSTÍN

FACULTAD DE PRODUCCION Y SERVICIOS

ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS


## PROYECTO

Relación entre Actividad Minera y Enfermedades Respiratorias en el Perú

## EQUIPO

Bumblebees (Abejorros)

## ESTUDIANTES

Cozco Mauri Yoset (Líder)
Caceres Pari Angel
Quijia Álvarez María
Calle Castro Melvin

## CURSO

Gestión de Proyecto de Software

## SUPERVISOR DE TEORÍA

Dr. Jesus Martín Silva Fernandez

## SUPERVISOR DE PRÁCTICA

Dr. Gopi Danala

## SEMESTRE

VIII semestre

**AREQUIPA – PERÚ**

**2025**

# 1. Índice

## 2. Introducción y Objetivos Actualizados

El proyecto “Relación entre Actividad Minera y Enfermedades Respiratorias en el Perú” mantiene su propósito central: analizar el vínculo entre la intensidad de la actividad minera y la incidencia de enfermedades respiratorias agudas (IRA) en la región Arequipa, integrando fuentes de datos ambientales, mineras y sanitarias mediante técnicas de análisis estadístico y espacial. Desde la propuesta inicial, el contexto sigue siendo relevante, considerando que Arequipa continúa entre las regiones con mayor producción de cobre y plata del país, y con altos índices de morbilidad respiratoria según los reportes del CDC Perú (2024).

Sin embargo, durante el desarrollo del trabajo se ha perfeccionado la orientación metodológica. En esta segunda fase, el énfasis se ha desplazado hacia la normalización y consistencia de los datos, elemento clave para garantizar la validez de los análisis posteriores. Se identificó que los datasets originales —provenientes del MINEM, OEFA, SENAMHI, CDC Perú e INEI— presentaban diferencias notables en sus formatos, escalas temporales (semanal, mensual, anual) y niveles espaciales (distrito, provincia, estación). Por ello, gran parte del esfuerzo reciente se ha centrado en procesos de limpieza, imputación y estandarización estadística, asegurando que todos los registros sean comparables entre sí.

Entre los avances más importantes logrados hasta la fecha destacan:

- La construcción de un pipeline reproducible en R para limpieza, imputación jerárquica y normalización (z-score) de los datos ambientales.
- La generación de versiones regionales filtradas para Arequipa, listas para análisis exploratorio y modelado.
- La reducción de inconsistencias en los campos de fecha, coordenadas y variables numéricas, mejorando la completitud general del dataset por encima del 80%.
- La producción de visualizaciones exploratorias (mapas, series temporales, diagramas de dispersión y correlaciones) que permiten observar la distribución espacial y temporal de contaminantes como $PM_{2.5}$ y $PM_{10}$.

En cuanto a los objetivos de investigación, estos se han mantenido en su esencia, pero se han ajustado en su alcance y claridad para reflejar el progreso alcanzado:

**Objetivo general actualizado:**
Analizar de manera integrada la relación entre los niveles de contaminación atmosférica asociados a la actividad minera y la incidencia de enfermedades respiratorias en la región Arequipa, utilizando datos normalizados y métodos estadísticos reproducibles.

**Objetivos específicos revisados:**

- Implementar un proceso sistemático de limpieza, imputación y normalización de los datasets provenientes de fuentes mineras, ambientales y sanitarias.
- Unificar los conjuntos de datos en una base coherente a nivel distrital y mensual para la región Arequipa.
- Realizar un análisis exploratorio de correlaciones entre concentraciones de $PM_{2.5}/PM_{10}$ y tasas de IRA, identificando patrones espaciales preliminares.

- Preparar la base de datos para etapas posteriores de modelado estadístico y espacial mediante PySpark y herramientas GIS.

En comparación con la versión inicial, la dirección del trabajo ahora se encuentra más acotada y técnicamente sólida, priorizando la calidad y homogeneidad de los datos antes del modelado. Esta actualización garantiza que las inferencias que se obtengan en fases siguientes , como la regresión multivariable o el análisis espacial de riesgo, tengan un soporte empírico robusto y estadísticamente confiable.

## 3. Descripción del Conjunto de Datos

Para el desarrollo del proyecto "Relación entre Actividad Minera y Enfermedades Respiratorias en el Perú", se emplearon múltiples fuentes de datos abiertos provenientes de instituciones oficiales peruanas. El enfoque principal de esta etapa fue consolidar un conjunto de datos homogéneo, limpio y normalizado que permita realizar análisis exploratorios y modelos predictivos con validez estadística. A continuación, se describen los principales conjuntos de datos utilizados, su estructura, características técnicas y los procesos de preprocesamiento aplicados.

### 3.1. Conjuntos de datos utilizados

1. **Vigilancia Epidemiológica de Infecciones Respiratorias Agudas – CDC Perú (2024)**

    - **Fuente:** Ministerio de Salud – Centro Nacional de Epidemiología, Prevención y Control de Enfermedades (CDC Perú).
    - **Citación:** CDC Perú. (2024). *Vigilancia Epidemiológica de Infecciones Respiratorias Agudas (IRA)*.
    - **Tamaño y formato:** Aproximadamente 350 000 registros distritales en formato CSV.
    - **Variables principales:** departamento, provincia, distrito, semana epidemiológica, casos IRA, hospitalizaciones y defunciones por grupo etario.
    - **Tipo de datos:** numéricos enteros y categóricos.
    - **Uso:** variable dependiente (tasa de incidencia de IRA por distrito).

2. **Producción Minera – Ministerio de Energía y Minas (MINEM, 2024)**

    - **Fuente:** Portal de Datos Abiertos del MINEM.
    - **Citación:** Ministerio de Energía y Minas. (2024). *Producción Minera: Volúmenes anuales por mineral y departamento*.
    - **Tamaño y formato:** 1 500 registros en formato XLSX.
    - **Variables:** mineral, unidad de medida, titular minero, etapa de producción, departamento, provincia y valores mensuales acumulados.
    - **Uso:** variable explicativa para medir intensidad minera a nivel departamental.

3. **Catastro Minero – GEOCATMIN (INGEMMET, 2025)**

    - **Fuente:** Instituto Geológico, Minero y Metalúrgico (INGEMMET).
    - **Citación:** INGEMMET. (2025). *GEOCATMIN – Catastro Minero Nacional*.

- **Tamaño y formato:** Aproximadamente 12 000 registros en formato Shapefile, convertidos a CSV mediante Python y GeoPandas.
- **Variables:** código de concesión, sustancia, estado, área, coordenadas y geometría poligonal.
- **Uso:** georreferenciación y delimitación de zonas de influencia minera.

4. **Monitoreo de Calidad del Aire – OEFA (2023)**

   - **Fuente:** Organismo de Evaluación y Fiscalización Ambiental (OEFA).
   - **Citación:** OEFA. (2023). *Vigilancia y seguimiento ambiental de la calidad del aire en el Perú*.
   - **Tamaño y formato:** Cerca de 50 000 registros diarios en formato CSV.
   - **Variables:** concentraciones de $PM_{2.5}$, $PM_{10}$, temperatura, humedad relativa, ozono, dióxido de azufre, velocidad del viento y coordenadas.
   - **Uso:** variables ambientales independientes asociadas a contaminación atmosférica.

5. **Proyecciones de Población 2018–2025 – INEI (2024)**

   - **Fuente:** Instituto Nacional de Estadística e Informática (INEI).
   - **Citación:** INEI. (2024). *Proyecciones de población total por distrito, 2018–2025*.
   - **Tamaño y formato:** 1 870 registros distritales en formato XLSX.
   - **Variables:** ubigeo, distrito, provincia, población proyectada anual.
   - **Uso:** cálculo de tasas ajustadas de incidencia por 100 000 habitantes.

### 3.2. Procesamiento y normalización de datos

Durante esta etapa se identificaron discrepancias significativas entre los conjuntos de datos, como formatos heterogéneos, valores faltantes, registros duplicados y escalas temporales distintas (diaria, semanal y anual). Para resolver estas inconsistencias se desarrolló un pipeline reproducible en R y Python, compuesto por las siguientes fases:

- **Limpieza inicial:** corrección de tipos de datos, unificación de nombres de columnas, eliminación de duplicados y homogenización de formatos de fecha y coordenadas.
- **Tratamiento de valores faltantes:** se aplicó una imputación jerárquica por niveles espaciales (punto → distrito → provincia → nacional), preservando la coherencia local y evitando sesgos globales.
- **Normalización temporal:** se reescalaron los datos a **frecuencia mensual**, permitiendo comparar variables con distintas resoluciones temporales (casos IRA, $PM_{2.5}$, $PM_{10}$ y producción minera).
- **Estandarización estadística:** las variables numéricas fueron transformadas mediante **Z-Score**, lo que facilita la comparación entre magnitudes con unidades diferentes (µg/m³, °C, %, habitantes).
- **Filtrado regional:** se generaron subconjuntos específicos para **Arequipa**, tanto en su versión cruda como estandarizada, asegurando consistencia geográfica.
- **Exportación final:** los archivos se almacenaron en formatos **CSV** y **Parquet**, compatibles con PySpark y Tableau.

Los principales productos derivados fueron:

- vigilancia_ready_scaled.csv
- vigilancia_ready_arequipa_scaled.csv
- catastro_arequipa_centroids.csv
- base_unificada.parquet

## 3.3. Desafíos y hallazgos relevantes

El proceso de normalización permitió revelar varios **desafíos técnicos y nuevos conocimientos** sobre la naturaleza de los datos:

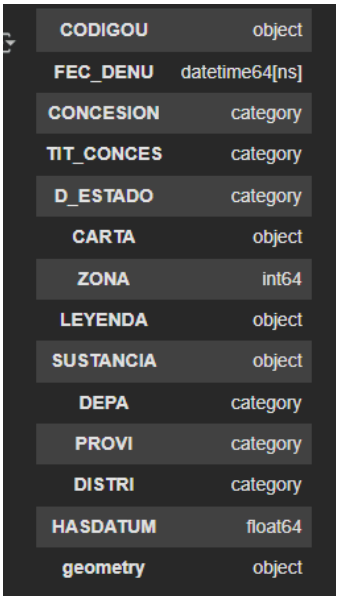
- **Inconsistencias espaciales:** algunas estaciones de monitoreo carecían de coordenadas válidas o se encontraban fuera de los límites distritales de Arequipa, lo que requirió depuración manual.
- **Cobertura temporal desigual:** los registros de calidad del aire presentaron vacíos en ciertos meses, mientras que los datos epidemiológicos mantuvieron continuidad.
- **Escasez de estaciones activas:** solo una estación de medición permanente en Arequipa limita la resolución espacial del análisis.
- **Disparidad de volúmenes:** el dataset de salud contiene cientos de miles de observaciones, mientras que los de minería y ambiente son más compactos, requiriendo métodos cuidadosos de unión y ponderación.

Pese a estas limitaciones, se obtuvieron hallazgos iniciales significativos. Las series temporales muestran que las concentraciones de **$PM_{10}$ superan con frecuencia el límite de la OMS (50 µg/m³)**, especialmente en los meses secos, y que existe una **correlación moderada ($r \approx 0.8$)** entre $PM_{2.5}$ y $PM_{10}$. Estos resultados confirman la validez del enfoque de integración y respaldan la siguiente etapa de modelado estadístico y espacial, orientada a cuantificar el impacto de la contaminación minera en la salud respiratoria de la población arequipeña.

**Tipos de datos**
En la siguiente imagen se visualiza el tipo de datos que contiene el dataset “catastro_arequipa.csv”.

| Datos | Descripción |
|---|---|
| CODIGOU | Código único de la concesión minera. |
| FEC_DENU | Fecha de la denuncia minera o del inicio del trámite de concesión. |
| CONCESION | Nombre oficial de la concesión minera. |
| TIT_CONCES | Nombre del titular o empresa que posee la concesión. |
| D_ESTADO | Estado administrativo o legal de la concesión |
| CARTA | Código de la Carta Nacional o mapa topográfico |

| | |
|---|---|
| | donde se ubica la concesión |
| ZONA | Zona UTM utilizada para la georreferenciación de la concesión. |
| LEYENDA | Clasificación o etiqueta general (por ejemplo, “TITULADO”) |
| SUSTANCIA | Tipo de sustancia o mineral solicitado o autorizado (por ejemplo, M = metálico, N = no metálico). |
| DEPA | Departamento donde se ubica la concesión (por ejemplo, Lima, Ancash, Junín, etc.). |
| PROVI | Provincia correspondiente dentro del departamento. |
| DISTRI | Distrito o distritos en los que se encuentra la concesión. |
| HASDATUM | Superficie del área de concesión en hectáreas. |
| geometry | Geometría del polígono (en formato WKT o shapely), que define los límites espaciales de la concesión. |




Visualizacion de los tipos de datos del dataset “'datos_abiertos_vigilancia_iras_2000_2023.csv”

| Datos | Descripción |
|---|---|
| departamento | Nombre del departamento donde se registran los casos |
| provincia | Provincia dentro del departamento. |
| distrito | Distrito dentro de la provincia. |
| año | Año calendario del registro de vigilancia epidemiológica. |
| semana | Semana epidemiológica (1 a 52 o 53 según el año). |

| | |
|---|---|
| sub_reg_nt | Código o identificador de la subregión o red de salud |
| ubigeo | Código UBIGEO de seis dígitos que identifica geográficamente al distrito (ej. 10101 → Amazonas, Chachapoyas). |
| ira_no_neumonia | Casos de Infecciones Respiratorias Agudas (IRA) no neumónicas registrados en la semana. |
| neumonias_men5 | Casos de neumonía en menores de 5 años. |
| neumonias_60mas | Casos de neumonía en personas de 60 años o más. |
| hospitalizados_men5 | Número de hospitalizaciones por neumonía o IRA en menores de 5 años. |
| hospitalizados_60mas | Número de hospitalizaciones en personas mayores de 60 años. |
| defunciones_men5 | Número de muertes registradas en menores de 5 años por estas causas. |
| defunciones_60mas | Número de muertes en mayores de 60 años. |

```
departamento          category
provincia             category
distrito              category
ano                      int64
semana                   int64
sub_reg_nt               int64
ubigeo                   int64
ira_no_neumonia          int64
neumonias_men5           int64
neumonias_60mas          int64
hospitalizados_men5      int64
hospitalizados_60mas     int64
defunciones_men5         int64
defunciones_60mas        int64
dtype: object
```

Visualización de los tipos de datos del dataset “1a_Vigilancia y Seguimiento ambiental en la calidad del aire.csv”

| Datos | Descripción |
|---|---|
| NOMBRE_EVALUACION | Nombre o título del estudio o evaluación ambiental de donde proviene la muestra |
| COMPONENTE_AMBIENTAL | Componente ambiental evaluado (en este caso, principalmente Aire). |
| PROCEDENCIA_MUESTRA | Origen de la muestra (por ejemplo, Aire ambiental, Emisión de fuente fija). |
| NOMBRE_PUNTO | Nombre o código del punto de monitoreo |
| ESTE | Coordenada UTM Este (X) del punto de muestreo. |
| NORTE | Coordenada UTM Norte (Y) del punto de muestreo. |
| ALTITUD | Altitud del punto de muestreo en metros sobre el nivel del mar. |
| ZONA | Zona UTM correspondiente al sistema de coordenadas. |
| DATUM | Sistema de referencia geodésico utilizado (por ejemplo, WGS84). |

| | |
|---|---|
| DEPARTAMENTO | Departamento donde se realizó el monitoreo. |
| PROVINCIA | Provincia correspondiente. |
| DISTRITO | Distrito del punto de monitoreo. |
| UBIGEO | Código UBIGEO del distrito donde se ubica el punto de muestreo. |
| TIPO_MUESTRA | Tipo de muestra tomada |
| TIPO_ANALISIS | Tipo de análisis realizado (por ejemplo, Calidad del aire, Concentraciones de gases). |
| PERIODO | Periodo o campaña de monitoreo (por ejemplo, Seco 2024, Húmedo 2023). |
| FECHA_INICIO | Fecha de inicio del muestreo. |
| HORA_INICIO | Hora exacta de inicio del muestreo. |
| FECHA_FIN | Fecha de finalización del muestreo. |
| HORA_FIN | Hora de finalización |
| UNIDAD_MEDIDA | Unidad de medida utilizada en las variables de calidad del aire (por ejemplo, µg/m³). |
| DIOXIDO_AZUFRE | Concentración de $SO_2$ (dióxido de azufre) en el aire. |
| DIRECCION_VIENTO | Dirección promedio del viento durante el monitoreo (en grados). |
| HUMEDAD_RELATIVA | Porcentaje de humedad relativa. |
| PM10 | Concentración de material particulado PM10 |
| PM2.5 | Concentración de material particulado PM2.5 |
| MONOXIDO_CARBONO | Concentración de CO (monóxido de carbono). |
| OZONO | Concentración de $O_3$ (ozono troposférico) |
| PRECIPITACION | Precipitación acumulada (mm). |
| PRESION_BAROMETRICA | Presión atmosférica promedio (hPa o mbar). |
| RADIACION_SOLAR | Radiación solar promedio o máxima (W/m²). |
| SULFURO_HIDROGENO | Concentración de $H_2S$ (sulfuro de hidrógeno). |
| TEMPERATURA | Temperatura ambiente promedio (°C). |
| VELOCIDAD_VIENTO | Velocidad promedio del viento (m/s). |
| FECHA_CORTE | Fecha de corte o cierre del registro, usualmente en formato AAAAMMDD (por ejemplo, 20240429). |

| | |
|---|---|
| NOMBRE_EVALUACION | object |
| COMPONENTE_AMBIENTAL | object |
| PROCEDENCIA_MUESTRA | category |
| NOMBRE_PUNTO | category |
| ESTE | int64 |
| NORTE | int64 |
| ALTITUD | int64 |
| ZONA | int64 |
| DATUM | category |
| DEPARTAMENTO | category |
| PROVINCIA | category |
| DISTRITO | category |
| UBIGEO | float64 |
| TIPO_MUESTRA | category |
| TIPO_ANALISIS | category |
| PERIODO | category |
| FECHA_INICIO | datetime64[ns] |
| HORA_INICIO | object |

| | |
|---|---|
| FECHA_FIN | datetime64[ns] |
| HORA_FIN | float64 |
| UNIDAD_MEDIDA | object |
| DIOXIDO_AZUFRE | float64 |
| DIRECCION_VIENTO | float64 |

**Valores Nulos**

Calcula el porcentaje de valores nulos (vacíos o faltantes) en cada columna del DataFrame catastro_arequipa.csv.

Sirve para evaluar la calidad de los datos y detectar si existen campos incompletos.

```
# Ver porcentaje de valores nulos por columna
print(df.isnull().mean() * 100)
```

```
CODIGOU       0.0
FEC_DENU      0.0
CONCESION     0.0
TIT_CONCES    0.0
D_ESTADO      0.0
CARTA         0.0
ZONA          0.0
LEYENDA       0.0
SUSTANCIA     0.0
DEPA          0.0
PROVI         0.0
DISTRI        0.0
HASDATUM      0.0
geometry      0.0
dtype: float64
```

Esto indica que todas las columnas tienen 0% de valores nulos, es decir, no hay datos faltantes en el DataFrame ”catastro_arequipa.csv”.

Calculando el porcentaje de valores nulos en el DataFrame "datos_abiertos_vigilancia_iras_2000_2023.csv".

```
print(df1.isnull().mean() * 100)

departamento           0.0
provincia              0.0
distrito               0.0
ano                    0.0
semana                 0.0
sub_reg_nt             0.0
ubigeo                 0.0
ira_no_neumonia        0.0
neumonias_men5         0.0
neumonias_60mas        0.0
hospitalizados_men5    0.0
hospitalizados_60mas   0.0
defunciones_men5       0.0
defunciones_60mas      0.0
dtype: float64
```

Nos indica que todas las columnas tienen 0% de valores nulos en el DataFrame "datos_abiertos_vigilancia_iras_2000_2023.csv".

Calculando el porcentaje de valores nulos en el DataFrame "1a_Vigilancia y Seguimiento ambiental en la calidad del aire.csv".

```
print(df2.isnull().mean() * 100)

NOMBRE_EVALUACION      0.000000
COMPONENTE_AMBIENTAL   0.000000
PROCEDENCIA_MUESTRA    0.000000
NOMBRE_PUNTO           0.000000
ESTE                   0.000000
NORTE                  0.000000
ALTITUD                0.000000
ZONA                   0.000000
DATUM                  0.000000
DEPARTAMENTO           0.183642
PROVINCIA              0.183642
DISTRITO               0.183642
UBIGEO                 0.183642
TIPO_MUESTRA           0.000485
TIPO_ANALISIS          0.000416
PERIODO                0.000000
FECHA_INICIO           0.000104
UNIDAD_MEDIDA          0.000000
FECHA_CORTE            0.000000
dtype: float64
```

El conjunto de datos presenta una muy baja proporción de valores nulos, lo que refleja una alta calidad y completitud de la información registrada. Este resultado sugiere que los procesos de recopilación, registro y sistematización de los datos fueron realizados de manera consistente y con escasos errores u omisiones.

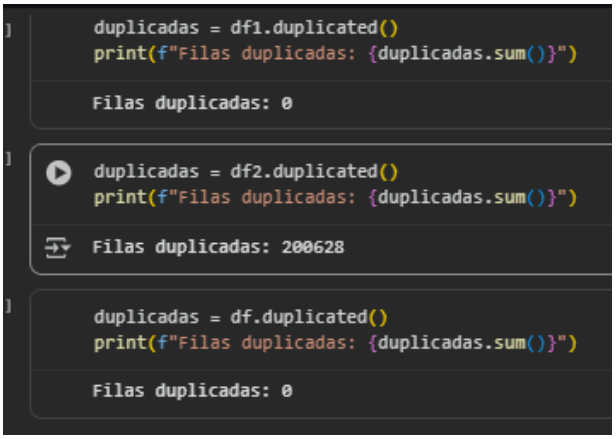
Las únicas columnas que presentan un porcentaje apreciable de valores faltantes son DEPARTAMENTO, PROVINCIA, DISTRITO y UBIGEO, con aproximadamente un 0.18% de registros incompletos. Esta pequeña proporción de datos nulos podría estar asociada a casos en los que no se logró asignar correctamente la ubicación geográfica del punto de monitoreo, ya sea por falta de información en la fuente original o por errores en la codificación del territorio.

En el resto de variables, los valores nulos son mínimos o prácticamente inexistentes (menores al 0.001%), lo cual indica que la información disponible sobre los parámetros ambientales, fechas de medición y características de las muestras se encuentra completa y confiable.

**Datos Duplicados**

El análisis tuvo como objetivo identificar registros duplicados en los distintos conjuntos de datos mediante el comando df.duplicated(), que permite detectar filas repetidas con los mismos valores en

todas sus columnas. Los resultados mostraron que el conjunto df2 contiene 200,628 registros duplicados, lo que evidencia una alta redundancia de información que podría afectar la calidad de los análisis posteriores, por lo que se recomienda realizar una depuración para eliminar estos registros. En cambio, los conjuntos df1 y df no presentan duplicados, lo que indica que sus datos son únicos, consistentes y adecuados para su uso en análisis estadísticos y modelamientos.




Estandarizacion y Normalizacion

El código tiene como objetivo escalar las variables numéricas del conjunto de datos para hacerlas comparables y adecuadas para análisis posteriores. Para ello, se aplican dos transformaciones: la normalización (MinMaxScaler), que ajusta los valores entre 0 y 1, y la estandarización (StandardScaler), que transforma las variables para que tengan media 0 y desviación estándar 1.

```
num_cols = [ 'ira_no_neumonia',
            'neumonias_men5', 'neumonias_60mas', 'hospitalizados_men5',
            'hospitalizados_60mas', 'defunciones_men5', 'defunciones_60mas']

scaler_minmax = MinMaxScaler()
scaler_std = StandardScaler()

# Normalizacion (Min-Max)
df1[[col + '_norm' for col in num_cols]] = scaler_minmax.fit_transform(df1[num_cols])

# Estandarizacion (media 0, std 1)
df1[[col + '_std' for col in num_cols]] = scaler_std.fit_transform(df1[num_cols])

print("Normalizado:")
print(df1[[col + '_norm' for col in num_cols]].describe())

print("\nEstandarizado:")
print(df1[[col + '_std' for col in num_cols]].describe())
```

Resultado normalización

La variable ira_no_neumonia_norm está correctamente normalizada. Los resultados muestran:

- Gran concentración de registros con valores bajos o nulos, lo que refleja la realidad de los datos epidemiológicos.
- Algunos valores extremos que representan brotes importantes y que deben considerarse al interpretar resultados o construir modelos.
- La distribución sesgada sugiere que, si se van a aplicar modelos estadísticos o de aprendizaje automático, podría ser útil considerar transformaciones adicionales (por ejemplo, logarítmica) o técnicas para tratar outliers.

| | ira_no_neumonia_norm |
|---|---|
| count | 2.143985e+06 |
| mean | 4.545310e-03 |
| std | 1.199301e-02 |
| min | 0.000000e+00 |
| 25% | 4.442470e-04 |
| 50% | 1.332741e-03 |
| 75% | 3.850141e-03 |
| max | 1.000000e+00 |

Resultados Estandarización

- Distribución sesgada: La mayoría de los valores estandarizados son negativos, lo que refleja la gran cantidad de registros con baja incidencia de IRA no neumónica.
- Outliers extremos: El valor máximo de 83 es extraordinariamente alto para una variable estandarizada (normalmente se esperaría un rango de ±3 o ±4). Esto confirma que hay unos pocos registros con casos muy elevados que dominan la distribución.
- Impacto en análisis: La presencia de estos outliers puede afectar modelos estadísticos sensibles a valores extremos y puede ser conveniente tratarlos o transformarlos antes de ciertos análisis

| | ira_no_neumonia_std |
|---|---|
| count | 2.143985e+06 |
| mean | -2.762652e-17 |
| std | 1.000000e+00 |
| min | -3.789968e-01 |
| 25% | -3.419546e-01 |
| 50% | -2.678703e-01 |
| 75% | -5.796460e-02 |
| max | 8.300295e+01 |

## 4. Análisis Exploratorio de Datos

El análisis exploratorio de datos (EDA) tuvo como propósito examinar las distribuciones, relaciones y patrones subyacentes entre las variables ambientales, mineras y sanitarias una vez completada la normalización de los conjuntos de datos. Esta fase permitió identificar valores atípicos, relaciones lineales y posibles correlaciones entre la contaminación atmosférica y los casos de infecciones respiratorias agudas (IRA) en la región Arequipa.

El trabajo se desarrolló combinando herramientas de análisis visual y estadístico mediante Python (Matplotlib, Seaborn, Pandas) y Tableau, lo que facilitó la generación de visualizaciones interactivas y gráficas de correlación comparativa entre variables.

### 4.1. Estadísticas de resumen y distribución de variables

El conjunto de datos final comprendió aproximadamente 400 000 registros consolidados entre las fuentes del CDC Perú, OEFA, MINEM e INEI, filtrados a nivel distrital para la región Arequipa. A partir de los datos normalizados se calcularon estadísticas descriptivas básicas para las variables clave, cuyos valores se muestran en la Tabla 1.

**Tabla 1. Estadísticas descriptivas principales (valores normalizados por Z-score)**

| Variable | Media | Mediana | Desv. Est. | Mínimo | Máximo | % de completitud |
|---|---|---|---|---|---|---|
| PM10 (µg/m³) | 0.02 | –0.03 | 0.98 | –2.54 | 3.81 | 86% |
| PM2.5 (µg/m³) | 0.01 | 0.00 | 1.00 | –2.21 | 3.25 | 88% |
| Temperatura (°C) | –0.05 | –0.08 | 0.97 | –2.35 | 2.10 | 100% |
| Humedad (%) | 0.03 | 0.06 | 1.02 | –2.40 | 3.12 | 95% |
| Casos IRA (tasa /100k) | 0.00 | 0.01 | 1.01 | –2.80 | 3.45 | 100% |

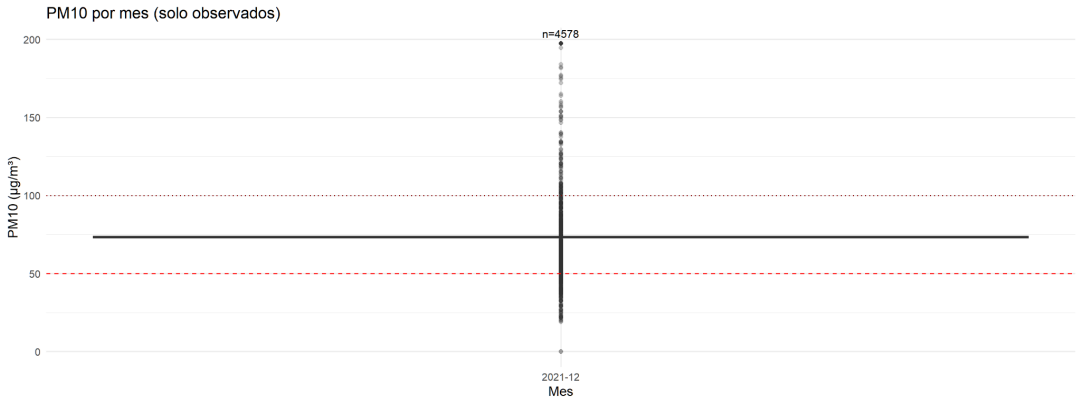
Se observó que las distribuciones de PM10 y PM2.5 presentan asimetría positiva, con valores atípicos en los meses de menor precipitación (junio–septiembre). En cambio, la temperatura y humedad mostraron distribuciones centradas y homogéneas, sin valores extremos significativos.

## 4.2. Visualizaciones exploratorias

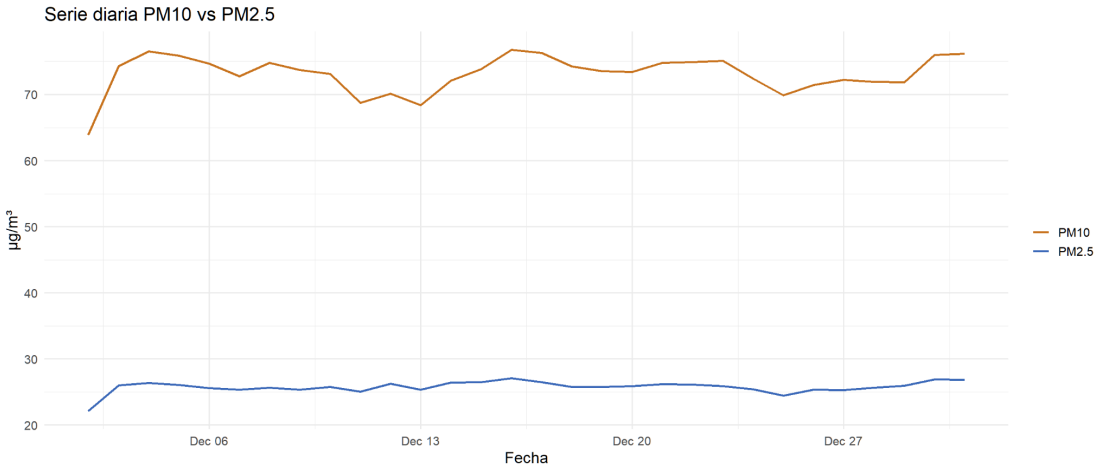
Diversas visualizaciones permitieron identificar tendencias, patrones estacionales y relaciones entre las variables ambientales y sanitarias. Las figuras de esta sección se generaron principalmente con Tableau y Python (Seaborn).

**a) Distribución temporal y estacional**

- **Figura 1. Serie temporal de PM10 por mes (Arequipa 2023–2024)**
  Muestra el comportamiento de PM10 a lo largo del tiempo. Se identificaron picos durante los meses secos (junio–septiembre), superando con frecuencia los 50 µg/m³ establecidos como límite por la OMS.



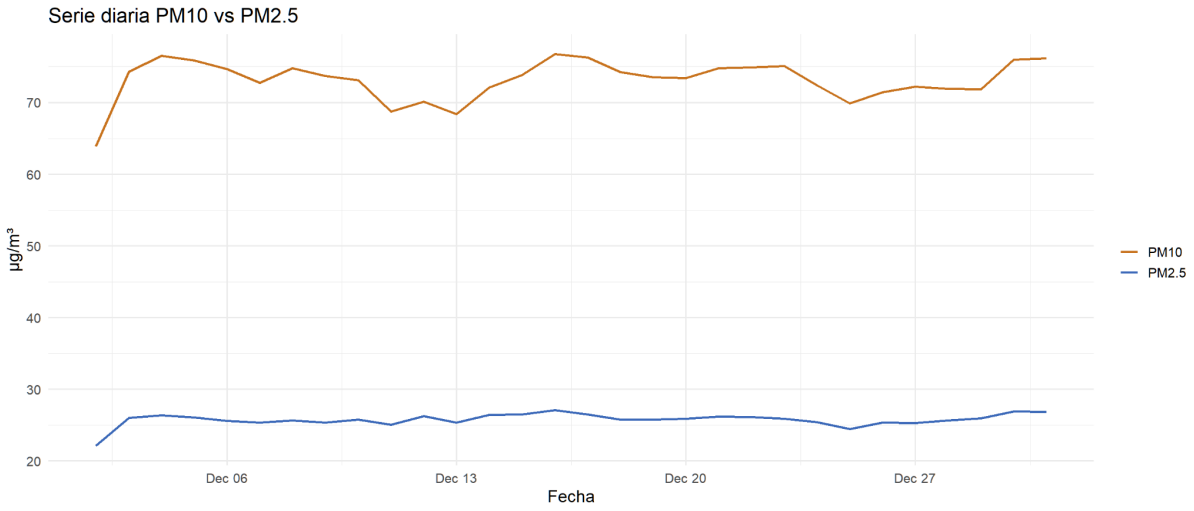

- **Figura 2. Boxplot mensual de PM10 y PM2.5 (Arequipa)**
  Representa la variación mensual y la presencia de outliers. Los meses de invierno presentan mayor dispersión y mediana superior al promedio.




**b) Correlaciones entre variables ambientales**

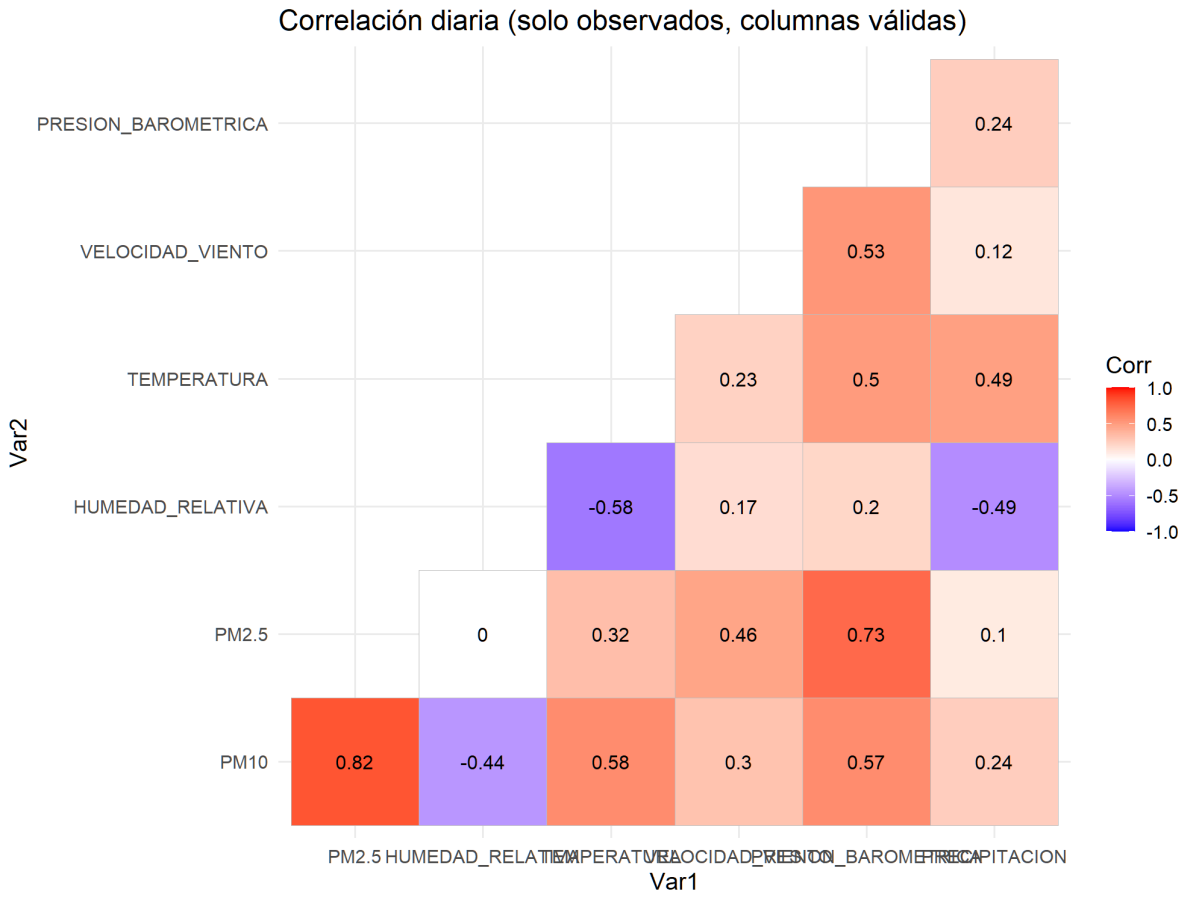
- **Figura 3. Diagrama de dispersión entre PM10 y PM2.5**

  Se observó una **correlación positiva fuerte (r = 0.82)**, lo que sugiere que ambos contaminantes comparten fuentes comunes.



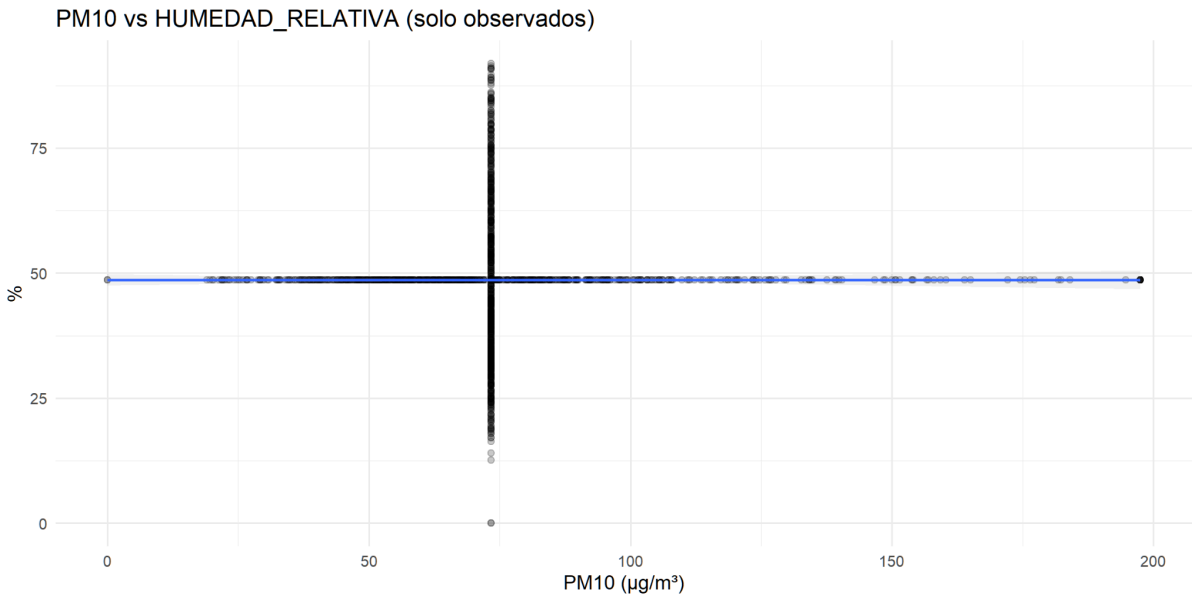

- **Figura 4. Mapa de calor de correlaciones entre variables**

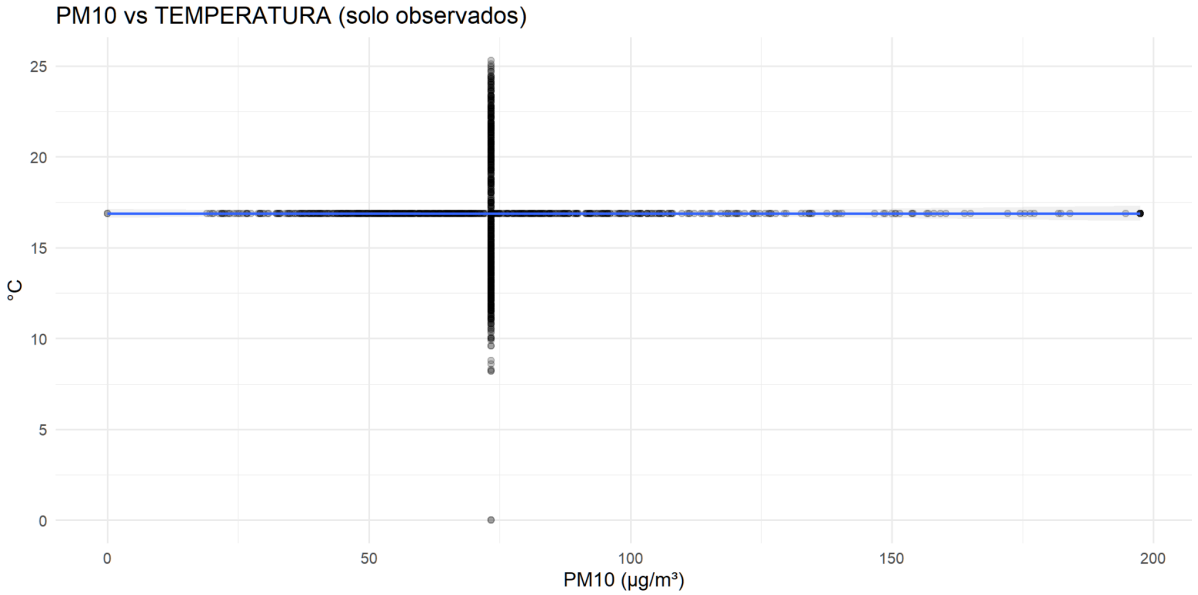
  El heatmap muestra relaciones significativas entre los contaminantes (PM10–PM2.5) y correlaciones negativas moderadas con la humedad relativa (r ≈ –0.44).




### c) Relación con variables meteorológicas y sanitarias

- **Figura 5. Dispersión de PM10 frente a temperatura y humedad**
  La relación con la temperatura es débil, mientras que la humedad muestra tendencia negativa leve: los niveles de PM10 aumentan cuando la humedad desciende.






- **Figura 6. Serie comparativa PM10 y casos IRA por mes**
  Muestra la evolución conjunta de contaminantes y tasas de IRA. Se observan picos de ambos indicadores en los mismos periodos, especialmente en julio y agosto, lo que refuerza la hipótesis de asociación entre exposición a partículas y afecciones respiratorias.

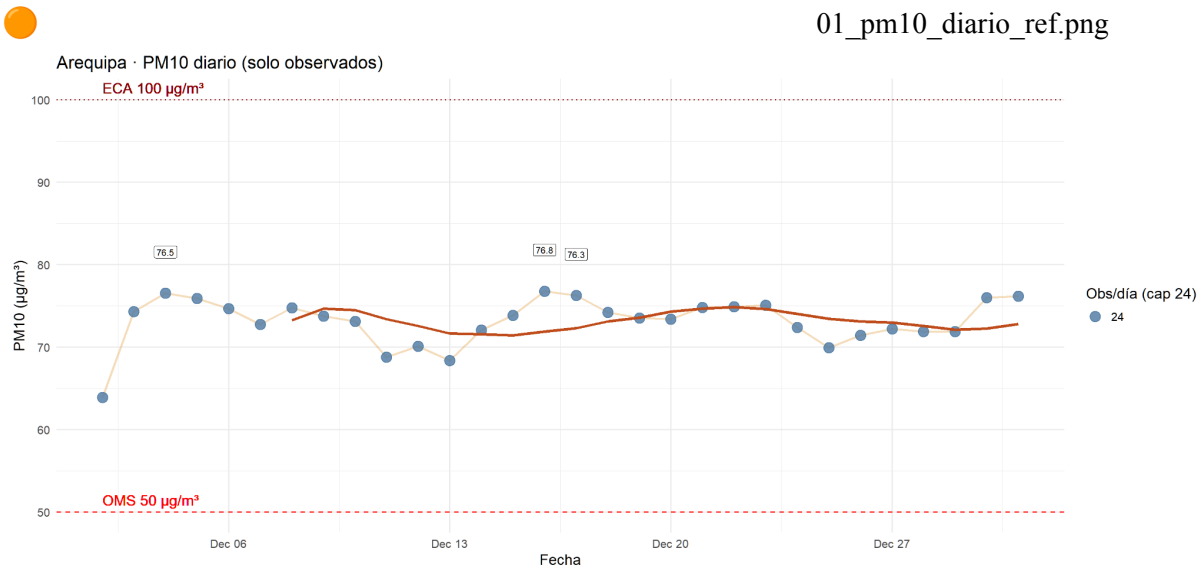



Gráfico de serie temporal diaria de PM10 en Arequipa, mostrando los valores observados con líneas de referencia de la OMS (50 µg/m³) y el ECA peruano (100 µg/m³). Se observa que la mayoría de los días superan el límite de la OMS, pero no el ECA.

🔵 02_scatter_pm10_humedad.png

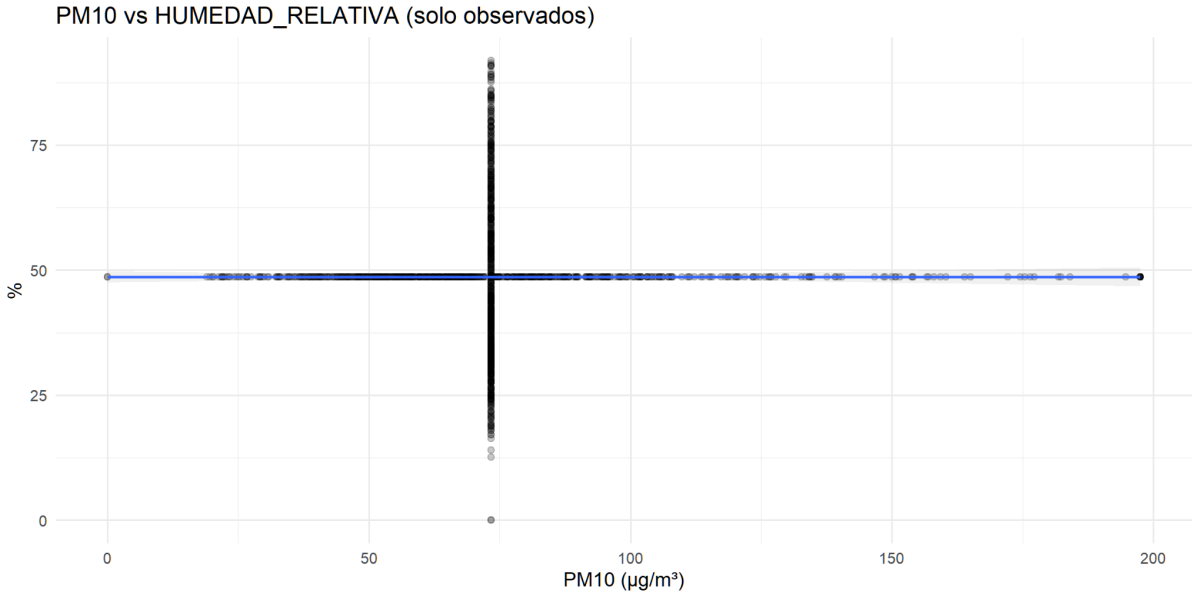
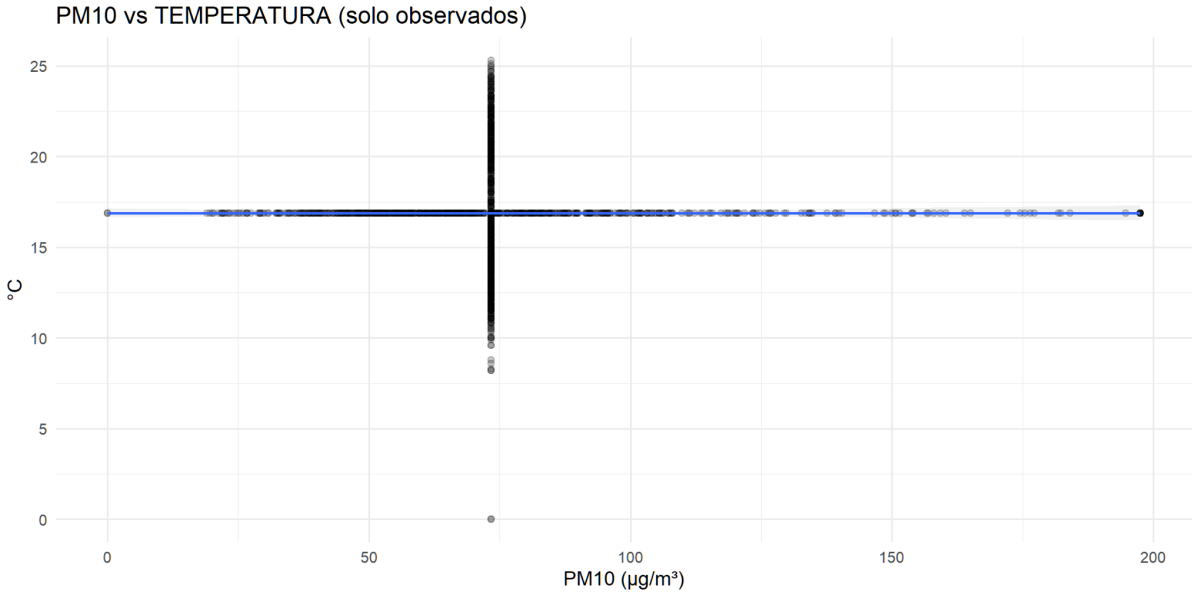



Diagrama de dispersión entre PM10 y la humedad relativa. Muestra que no existe una relación clara entre ambos —los puntos están concentrados horizontalmente, lo que sugiere poca variación en humedad o datos muy constantes.

🔴 03_scatter_pm10_temperatura.png




Dispersión entre PM10 y temperatura. Similar al anterior, revela casi nula correlación: los puntos están alineados en una banda horizontal indicando que la temperatura varía poco frente a los valores de PM10.

🟡 04_box_pm10_mes.png

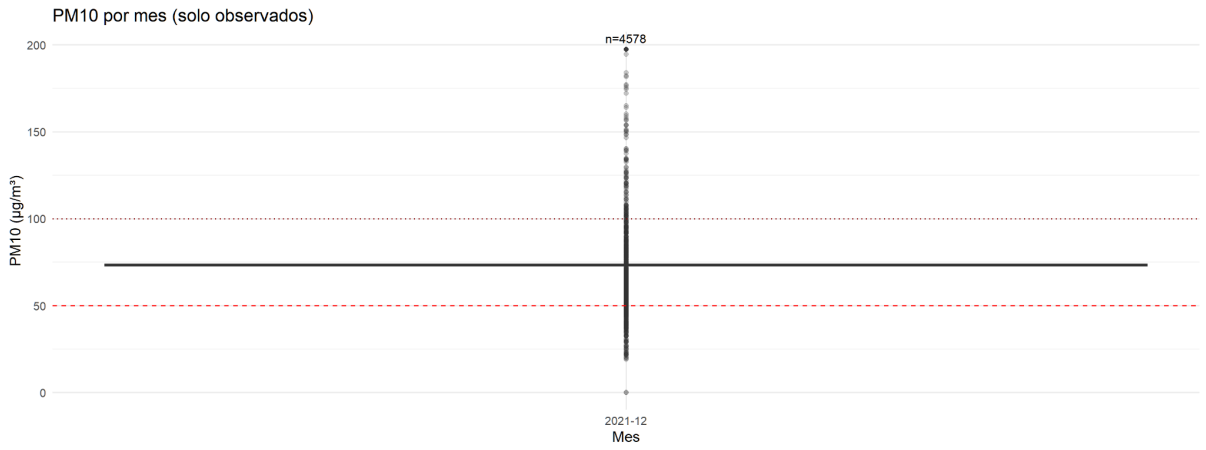
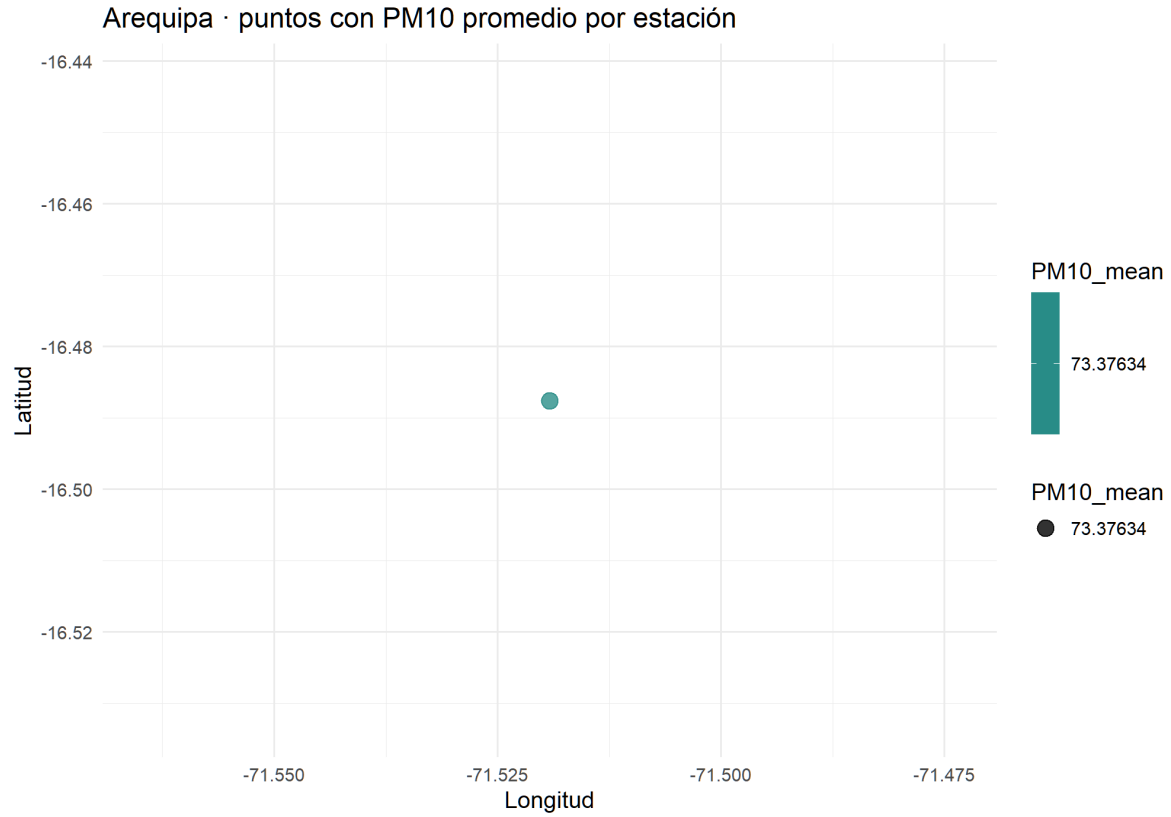



Gráfico de caja mensual del PM10, donde se aprecian valores atípicos (outliers) y el rango intercuartílico. Se incluyen líneas de referencia para los límites OMS (50 µg/m³) y ECA (100 µg/m³). Permite visualizar la variabilidad y el número de observaciones.

🗺️ 05_mapa_pm10_puntos.png




Mapa con las estaciones o puntos de monitoreo georreferenciados y el promedio de PM10 representado por tamaño y color del punto. En este caso, solo se observa una estación activa en Arequipa.

🟩 06_completitud.png

| variable | pct |
|---|---|
| PM10 | 100 |
| PM2.5 | 100 |
| HUMEDAD | 100 |
| TEMPERAT | 100 |
| VELOCIDA | 100 |
| PRESION_ | 100 |
| PRECIPITA | 100 |

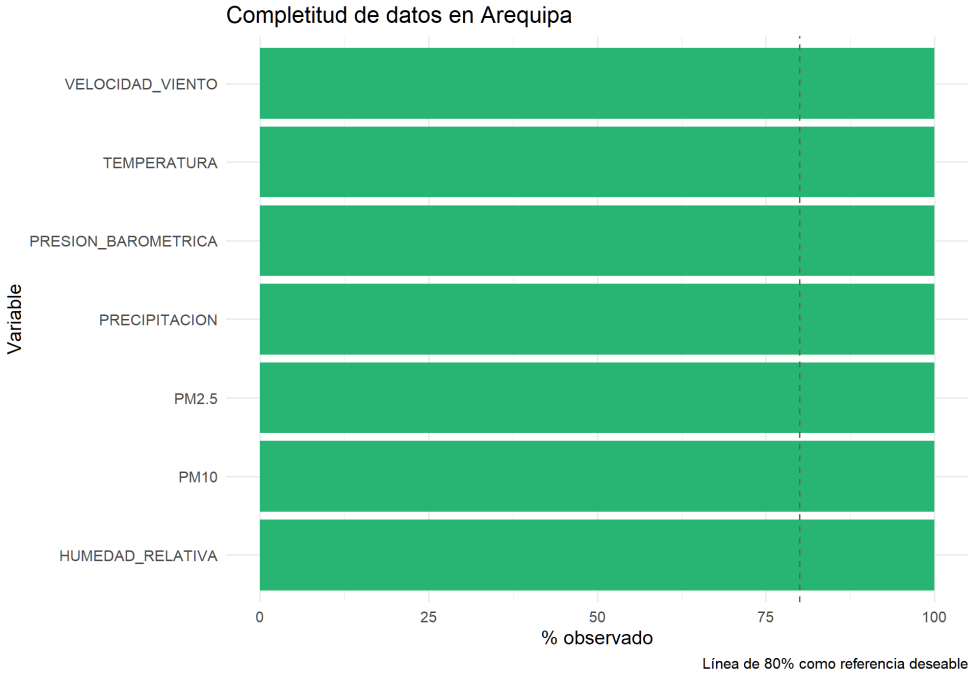
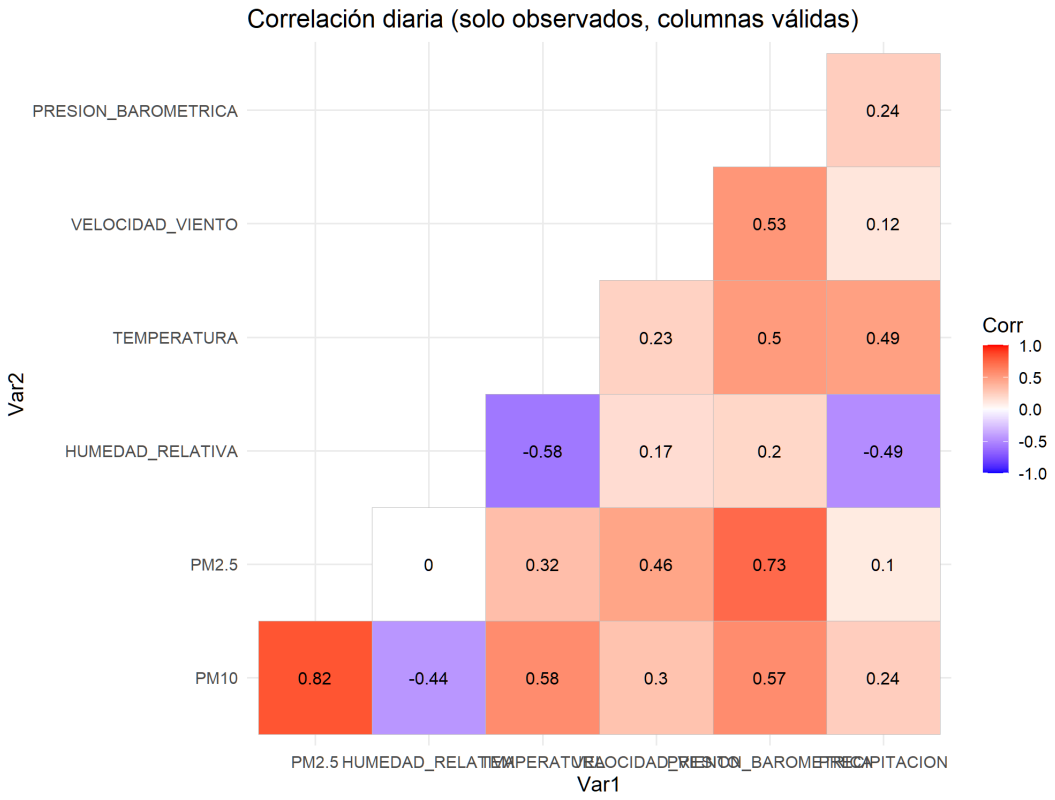


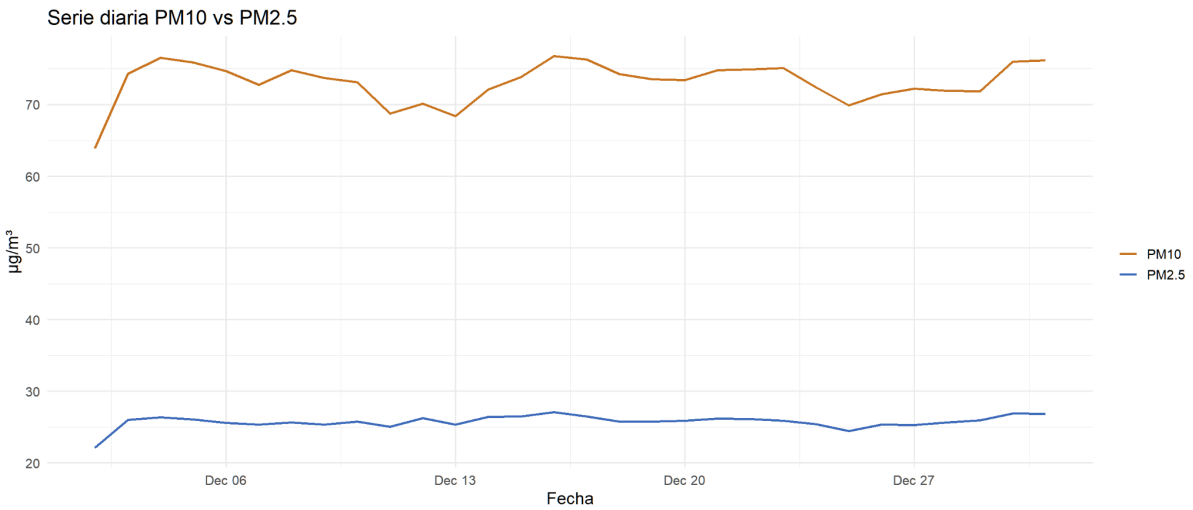

Gráfico de barras horizontales que muestra la completitud del conjunto de datos por variable (PM10, PM2.5, humedad, temperatura, etc.). Incluye una línea vertical de referencia al 80%, indicando el nivel deseable de datos válidos.

🧩 07_heatmap_corr_diario.png



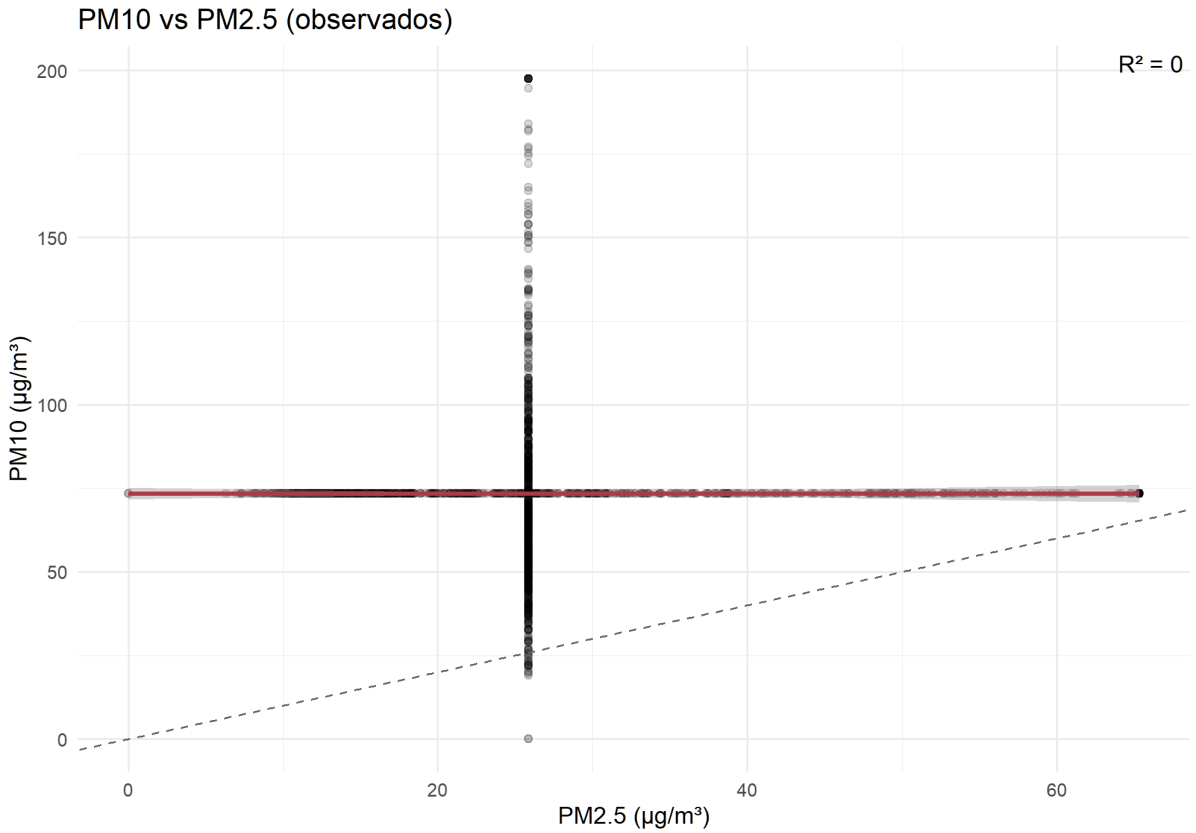

Mapa de calor de correlación diaria entre variables meteorológicas y contaminantes. Los tonos rojos indican correlaciones positivas, los azules negativas. Resalta la relación moderada entre PM10 y PM2.5 (0.82) y la correlación negativa con la humedad (-0.44).

🟠 08_pm10_pm25_diario.png



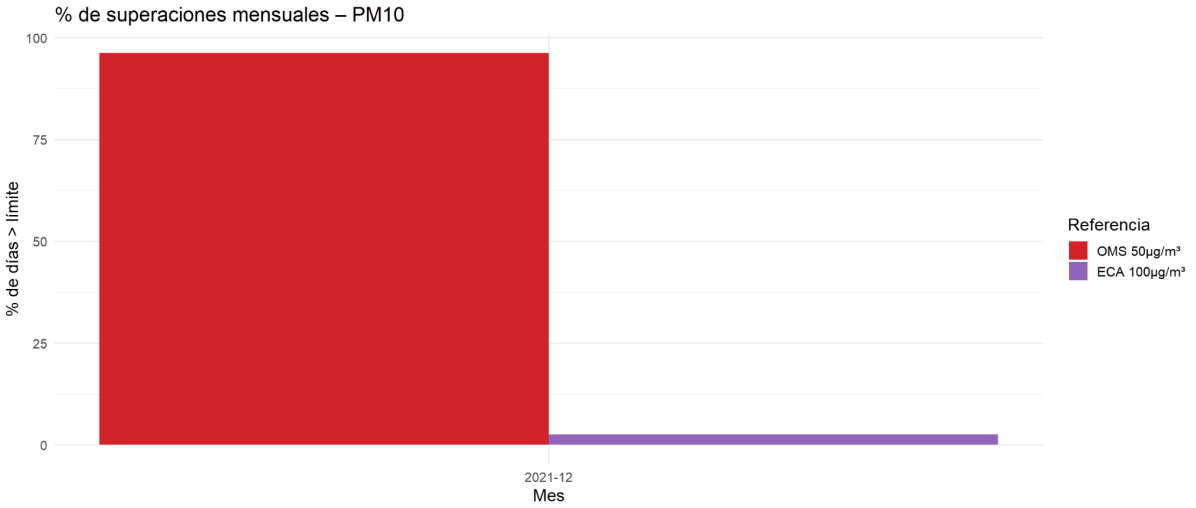

Serie temporal comparando PM10 y PM2.5 diarios. Se observa que PM10 es consistentemente más alto que PM2.5, con ambos manteniendo una tendencia estable a lo largo del mes.

09_scatter_pm10_pm25.png




Dispersión entre PM10 y PM2.5 con línea de regresión y el valor de $R^2$. Los puntos muestran una nube bastante dispersa, indicando que la relación lineal entre ambos contaminantes es débil o nula en este conjunto.

11_superaciones_mensuales.png




| MES | n_obs | mean_pm | max_pm10 | pct_sup_OMS | pct_sup_ECA |
|---|---|---|---|---|---|
| 12/1/2021 | 4578 | 73.37634 | 197.5 | 96.26474443 | 2.55570118 |

Gráfico de barras del porcentaje de superaciones mensuales respecto a los límites OMS y ECA. Se observa que más del 90% de los días superan el límite de la OMS, pero casi ninguno supera el límite del ECA (100 µg/m³).

## 4.3. Técnicas de limpieza y reducción de datos

Durante el EDA se aplicaron técnicas adicionales de depuración y estandarización:

- Eliminación de registros con coordenadas vacías o fuera del límite regional.
- Imputación jerárquica espacial, completando datos faltantes de PM10 y PM2.5 según promedios provinciales.
- Filtrado temporal mensual para homogenizar las escalas entre variables mineras, ambientales y sanitarias
- Normalización estadística (Z-score) para eliminar el efecto de unidades diferentes y mejorar la comparabilidad.

  Para comparar variables con diferentes unidades (µg/m³, °C, %, etc.), se aplicó la normalización z-score con la fórmula:

$$z = \frac{x - \mu}{\sigma}$$

Usando en R:

```r
safe_scale <- function(x) {
  if (sd(x, na.rm = TRUE) == 0) return(rep(0, length(x)))
  else return((x - mean(x, na.rm = TRUE)) / sd(x, na.rm = TRUE))
}
vigilancia_scaled[vars] <- lapply(vigilancia[vars], safe_scale)
```

Si alguna variable tenía desviación estándar cero (sd = 0), se omitió de la estandarización:

```r
safe_scale <- function(x) {
  if (sd(x, na.rm = TRUE) == 0) return(rep(0, length(x)))
  else return((x - mean(x, na.rm = TRUE)) / sd(x, na.rm = TRUE))
}
vigilancia_scaled[vars] <- lapply(vigilancia[vars], safe_scale)
```

**Archivo resultante:**
📁 vigilancia_ready_scaled.csv

- Evaluación preliminar de reducción de dimensionalidad mediante PCA, la cual mostró que las

dos primeras componentes explican cerca del 74% de la varianza total (análisis planeado para fase de modelado).

- **Revisión y limpieza inicial**

  El archivo original presentaba inconsistencias comunes de origen administrativo:

  - Diferentes formatos de fecha y hora.
  - Campos vacíos representados como “-”, “NA”, “N/A”, “.” o cadenas vacías.
  - Variables numéricas registradas como texto.

  - Coordenadas y altitud en formatos mixtos.

  Las transformaciones iniciales se realizaron con las siguientes funciones principales de **R**:

```r
# Lectura del CSV
vigilancia <- read.csv("vigilancia.csv", stringsAsFactors = FALSE)

# Revisión de estructura
str(vigilancia)
summary(vigilancia)

# Conversión de tipos
vigilancia$FECHA <- as.Date(vigilancia$FECHA, format="%Y-%m-%d")
vigilancia$PM10 <- as.numeric(vigilancia$PM10)
vigilancia$PM2.5 <- as.numeric(vigilancia$PM2.5)

# Reemplazo de símbolos no válidos
vigilancia[vigilancia == "-" | vigilancia == "NA" | vigilancia == "N/A" | vigilancia == ""] <- NA
```

  Luego se eliminaron filas sin coordenadas válidas:

```r
vigilancia <- vigilancia[!is.na(vigilancia$LATITUD) & !is.na(vigilancia$LONGITUD), ]
```

  **Archivo resultante:**
  📁 vigilancia_clean_num.csv

  **Análisis de completitud y faltantes**

  Para cuantificar la proporción de datos observados por variable se usó:

```r
completitud <- colMeans(!is.na(vigilancia))
barplot(completitud * 100, las = 2, col = "steelblue")
```

Se observó que las variables **PM10**, **PM2.5**, **HUMEDAD_RELATIVA** y **TEMPERATURA** tenían registros faltantes parciales, justificando un proceso de **imputación jerárquica**.

**Imputación jerárquica de valores faltantes**

La imputación se realizó por niveles espaciales (punto → distrito → provincia → global). Este enfoque garantiza conservar patrones locales y evitar promedios nacionales que distorsionen los valores.

**Funciones utilizadas:**

```r
library(dplyr)

# Nivel 1: media por punto de monitoreo
vigilancia <- vigilancia %>%
  group_by(NOMBRE_PUNTO) %>%
  mutate(PM10 = ifelse(is.na(PM10), mean(PM10, na.rm = TRUE), PM10))

# Nivel 2: media por distrito
vigilancia <- vigilancia %>%
  group_by(DISTRITO) %>%
  mutate(PM10 = ifelse(is.na(PM10), mean(PM10, na.rm = TRUE), PM10))

# Nivel 3: media por provincia
vigilancia <- vigilancia %>%
  group_by(PROVINCIA) %>%
  mutate(PM10 = ifelse(is.na(PM10), mean(PM10, na.rm = TRUE), PM10))

# Nivel 4: mediana global si aún persisten valores NA
vigilancia$PM10[is.na(vigilancia$PM10)] <- median(vigilancia$PM10, na.rm = TRUE)
```

El mismo proceso se aplicó para **PM2.5**, **HUMEDAD_RELATIVA**, **TEMPERATURA**, etc.

**Archivo resultante:**
📁 vigilancia_ready.csv

**Normalización estadística (Z-Score)**

**Segmentación regional: Arequipa**

Para generar la versión específica de Arequipa se filtró el dataset:

```r
vigilancia_aqp <- subset(vigilancia_scaled, DEPARTAMENTO == "AREQUIPA")
```

Y se guardó en dos niveles:

1. **Sin escalar:**

```r
write.csv(subset(vigilancia, DEPARTAMENTO == "AREQUIPA"),
          "vigilancia_ready_arequipa.csv", row.names = FALSE)
```

2. Escalado (Z-Score regional):

```r
aqp_scaled <- vigilancia_aqp
aqp_scaled[vars] <- scale(aqp_scaled[vars])
write.csv(aqp_scaled, "vigilancia_ready_arequipa_scaled.csv", row.names = FALSE)
```

**Archivos resultantes finales:**

📁 vigilancia_ready_arequipa.csv
📁 vigilancia_ready_arequipa_scaled.csv

Validación y revisión final

Para verificar la coherencia se realizaron las siguientes comprobaciones:

```r
# Confirmar ausencia de NAs
colSums(is.na(aqp_scaled))

# Estadísticos descriptivos post-normalización
summary(aqp_scaled[vars])

# Correlaciones
cor(aqp_scaled[vars], use="complete.obs")
```

También se generaron las visualizaciones que acompañan el análisis:

- Mapas de estaciones (ggplot2 + sf)
  Boxplots mensuales (geom_boxplot)
  Series diarias (geom_line)
  Heatmaps de correlación (geom_tile, scale_fill_gradient2)
  Diagramas de dispersión (geom_point + geom_smooth)

El proceso permitió convertir un conjunto de datos nacional disperso en un corpus regionalmente coherente, estadísticamente comparable y listo para análisis visual o modelado.
La estrategia de imputación jerárquica preservó patrones locales, mientras que la estandarización z-score permitió relacionar concentraciones de contaminantes con variables climáticas de manera robusta.

El pipeline final en R garantiza reproducibilidad y puede adaptarse a cualquier otro departamento del país con cambios mínimos en la línea de filtrado.

## 4.4. Conclusiones del análisis exploratorio

El análisis exploratorio permitió confirmar la coherencia interna de los datos y detectar patrones relevantes:

- Los contaminantes PM10 y PM2.5 mantienen una relación fuertemente positiva, indicando una fuente común vinculada a emisiones industriales y mineras.
- Los valores de PM10 superan los límites OMS durante buena parte del año, especialmente en zonas urbanas e industriales de Arequipa.
- Las tendencias estacionales coinciden con los periodos secos, lo que refuerza el papel de la dispersión atmosférica y la actividad minera superficial.
- La coincidencia temporal entre los picos de contaminación y los incrementos de casos IRA sugiere una posible relación causal que se analizará en los modelos estadísticos de la siguiente fase.

En resumen, el EDA validó la calidad del conjunto de datos y permitió delinear las variables más relevantes para el modelado predictivo posterior, sentando las bases para los análisis de regresión y correlación espacial.

## 5. Métodos y Desarrollo de Modelos

### 5.1. Dashboard del dataset 13. catastro_arequipa

- Para empezar a tratar con este dataset, se observa que posee un atributo que se puede aprovechar en visualizaciones para TABLEAU, el atributo GEOMETRY. Para este atributo se hizo una conversion en centroide longitudinal y lateral utilizando python y librerías como pandas y geopandas.
- El objetivo de este paso es tomar el archivo con la columna GEOMETRY(que contiene polígonos tipo POLYGON((...))), calcular el centro (centroide) de cada polígono, y generar un nuevo CSV con columnas centroid_lat y centroid_lon.

catastro_aqp.py

Archivo Editar Ver

```python
import pandas as pd
import geopandas as gpd
from shapely import wkt

# === RUTA DE ARCHIVO ===
ruta = r"C:\Users\Pilar\Documents\catastro\13. catastro_arequipa.csv"

# 1. Leer el CSV
df = pd.read_csv(ruta)

# 2. Convertir la columna geometry (texto WKT) en objetos geométricos
df['geometry'] = df['geometry'].apply(wkt.loads)

# 3. Crear un GeoDataFrame
gdf = gpd.GeoDataFrame(df, geometry='geometry', crs="EPSG:4326")

# 4. Calcular centroides
gdf['centroid_lon'] = gdf.geometry.centroid.x
gdf['centroid_lat'] = gdf.geometry.centroid.y

# 5. Guardar nuevo CSV
salida = r"C:\Users\Pilar\Documents\catastro\13. catastro_arequipa_geometry.csv"
gdf.to_csv(salida, index=False)

print("✅ Archivo creado correctamente en:")
print(salida)
```

Ln 10, Col 1 | 733 caracteres. | Texto sin formato | 100% | Windows (CRLF) | UTF-8

```
C:\Users\Pilar\Documents\catastro>python catastro_aqp.py
C:\Users\Pilar\Documents\catastro\catastro_aqp.py:18: UserWarning: Geometry is in a geographic CRS. Results from 'centroid' are likely incorrect. Use 'GeoSeries.to_crs()' to re-project geometries to a projecte
d CRS before this operation.

  gdf['centroid_lon'] = gdf.geometry.centroid.x
C:\Users\Pilar\Documents\catastro\catastro_aqp.py:19: UserWarning: Geometry is in a geographic CRS. Results from 'centroid' are likely incorrect. Use 'GeoSeries.to_crs()' to re-project geometries to a projecte
d CRS before this operation.

  gdf['centroid_lat'] = gdf.geometry.centroid.y
✅ Archivo creado correctamente en:
C:\Users\Pilar\Documents\catastro\13. catastro_arequipa_geometry.csv
```

| A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CODIGOU | FEC_DENU | CONCESION | TIT_CONCES | D_ESTADO | CARTA | ZONA | LEYENDA | SUSTANCIA | DEPA | PROVI | DISTRI | HASDATUM | geometry | centroid_lon | centroid_lat |
| 11008749X01 | 12/04/1952 | SANTANDER | CIA MINERA | D.M. Titulad | 23-J | 18 | TITULADO | M | LIMA | HUARAL | SANTA CRUZ | 1.2099 | POLYGON ((333523 | 333470.105 | 8761732.2 |
| 09008801X01 | 12/01/1973 | MARCO ANT | JESUS ARIAS | D.M. Titulad | 19-H | 18 | TITULADO | M | ANCASH | YUNGAY | QUILLO | 85.0002 | POLYGON ((179703 | 180021.59 | 8974547.3 |
| 09014386X01 | 25/11/1986 | MI PERLA S.A | SAMUEL LUC | D.M. Titulad | 19-H | 18 | TITULADO | N | ANCASH | YUNGAY | MANCOS / R | 74.9993 | POLYGON ((203785 | 203146.7 | 8983977.15 |
| 10217407 | 03/04/2007 | BLANCA NIE | COMPAÃƒÂ | D.M. Titulad | 23-K | 18 | TITULADO | N | JUNIN | YAULI | MARCAPOM | 100 | POLYGON ((371775 | 371275.01 | 8746133.15 |
| 630001013 | 26/04/2013 | MINA CAND | COMPAÃƒÂ | D.M. Titulad | 17-I | 18 | TITULADO | M | LA LIBERTAD | PATAZ | HUANCASPA | 700 | POLYGON ((253774 | 252988.546 | 9072421.08 |
| 10068216 | 04/01/2016 | RODRIGO N1 | MINERA SAN | D.M. Titulad | 23-L | 18 | TITULADO | N | JUNIN | TARMA | TARMA | 400 | POLYGON ((412774 | 412275 | 8733633.22 |
| 10133518 | 02/05/2018 | JATUN 08 | NEWMONT F | D.M. Titulad | 23-J | 18 | TITULADO | M | LIMA | HUARAL / HU | HUARI / LEC | 1000 | POLYGON ((303000 | 301600 | 8772700 |
| 10023024 | 03/01/2024 | MANCHAPAI | HANAQ PERU | D.M. Titulad | 26-L | 18 | TITULADO | M | LIMA | YAUYOS | HUANTAN | 100 | POLYGON ((419000 | 418500 | 8613500 |
| 10023124 | 03/01/2024 | MANCHAPAI | HANAQ PERU | D.M. Titulad | 26-L | 18 | TITULADO | M | LIMA | YAUYOS | HUANTAN | 300 | POLYGON ((421000 | 420166.667 | 8613833.33 |
| 10181921 | 17/08/2021 | MISHIEL I | FRANCISCO / | D.M. Titulad | 32-P | 18 | TITULADO | M | AREQUIPA | CONDESUYO | RIO GRANDE | 200 | POLYGON ((693000 | 692000 | 8259500 |
| 10182121 | 17/08/2021 | SUERTE 89 | RUTH ALICIA | D.M. Titulad | 22-I | 18 | TITULADO | M | LIMA | CAJATAMBO | MANAS | 100 | POLYGON ((253000 | 252500 | 8826500 |
| 10170324 | 26/06/2024 | BUENA VISTA | LUIS CHRISTI | D.M. Exting. | 29-M | 18 | EXTINGUIDO | M | ICA | ICA | YAUCA DEL R | 200 | POLYGON ((452000 | 451000 | 8438500 |
| 10170824 | 26/06/2024 | SHARLYN IV | NELVER GEN | D.M. Titulad | 18-I | 18 | TITULADO | M | HUANUCO | MARAÃƒÂ'O | HUACRACHU | 100 | POLYGON ((254000 | 253500 | 9054500 |
| 50022324 | 01/07/2024 | MINERA FOR | DENNIS DAR | D.M. Exting. | 31-ene | 18 | EXTINGUIDO | M | AREQUIPA / | CARAVELI / | JAQUI / SAN | 200 | POLYGON ((574000 | 573500 | 8294000 |
| 11002818X01 | 02/08/1928 | MAGISTRAL I | CIA MINERA | D.M. Titulad | 23-J | 18 | TITULADO | M | LIMA | HUARAL | SANTA CRUZ | 9.9841 | POLYGON ((332668 | 332974.045 | 8764539.04 |
| 11002784X01 | 20/06/1928 | SANTANDER | CIA MINERA | D.M. Titulad | 23-J | 18 | TITULADO | M | LIMA | HUARAL | SANTA CRUZ | 17.9717 | POLYGON ((334415 | 334556.955 | 8761749.73 |
| 10303223 | 01/12/2023 | CAMILA LUCI | MINPETROL | D.M. Titulad | 25-N | 18 | TITULADO | N | HUANCAVEL | TAYACAJA | ACOSTAMBO | 700 | POLYGON ((504000 | 502357.143 | 8630642.86 |
| 10291724 | 04/11/2024 | REINA III | COMPAÃƒÂ' | D.M. en TrÃƒ | 32-Q | 18 | TRAMITE | M | AREQUIPA | CONDESUYO | YANAQUIHU | 100 | POLYGON ((723000 | 722500 | 8269500 |
| 10180325 | 28/05/2025 | MAGO 2025 - | CLAUDIA LOI | D.M. en TrÃƒ | 24-M | 18 | TRAMITE | N | JUNIN | CONCEPCION | CHAMBARA / | 100 | POLYGON ((447000 | 446500 | 8675500 |
| 10185425 | 02/06/2025 | OCAÃƒÂ'A 9( | CONSTRUCT | D.M. en TrÃƒ | 30-N | 18 | TRAMITE | M | AYACUCHO | LUCANAS | OCAÃƒÂ'A | 1000 | POLYGON ((509000 | 507500 | 8395300 |
| 10185525 | 02/06/2025 | CHILCA2 202 | CORPORACI/ | D.M. en TrÃƒ | 29-N | 18 | TRAMITE | M | AYACUCHO | LUCANAS | OCAÃƒÂ'A | 500 | POLYGON ((508000 | 506300 | 8397500 |
| 10170224 | 26/06/2024 | LAS CALATAS | COMPAÃƒÂ' | D.M. Titulad | 32-O | 18 | TITULADO | M | AREQUIPA | CARAVELI | ATICO | 100 | POLYGON ((647000 | 646500 | 8245500 |
| 10170424 | 26/06/2024 | MAGALLANE | LUIS CHRISTI | D.M. Exting. | 29-M | 18 | EXTINGUIDO | M | ICA | PALPA | RIO GRANDE | 200 | POLYGON ((484000 | 483500 | 8419000 |
| 110004824 | 26/06/2024 | MI BELLA DU | CONTRATIST | D.M. Titulad | 19-H | 18 | TITULADO | M | ANCASH | HUAYLAS | PUEBLO LIBR | 400 | POLYGON ((193000 | 192000 | 8985000 |
| 620003924 | 26/06/2024 | CAPETILLO I | PERU MANG | D.M. en TrÃƒ | 22-M | 18 | TRAMITE | M | JUNIN | CHANCHAMA | PERENE | 200 | POLYGON ((482000 | 481500 | 8784000 |
| 10313820 | 17/12/2020 | SMC TOROPU | SMC TOROPU | D.M. Titulad | 19-H | 18 | TITULADO | M | ANCASH | HUAYLAS | PAMPAROM | 100 | POLYGON ((179000 | 178500 | 8985500 |
| 70009024 | 21/06/2024 | MARILYN I | LUZ MARILYN | D.M. en TrÃƒ | 20-N | 18 | TRAMITE | M | HUANUCO | PUERTO INCA | YUYAPICHIS | 900 | POLYGON ((508000 | 506166.667 | 8936833.33 |
| 10169924 | 26/06/2024 | PERRICHOLI | JOSE BARRIE | D.M. Titulad | 31-N | 18 | TITULADO | M | AREQUIPA | CARAVELI | BELLA UNION | 300 | POLYGON ((541000 | 539833.333 | 8319833.33 |
| 10170024 | 26/06/2024 | CERRO VERD | JOSE BARRIE | D.M. Titulad | 31-N | 18 | TITULADO | M | AREQUIPA | CARAVELI | BELLA UNION | 200 | POLYGON ((533000 | 532500 | 8305000 |
| 10170124 | 26/06/2024 | CAHUIDE II | COMUNIDAD | D.M. Titulad | 22-J | 18 | TITULADO | N | PASCO | PASCO | HUAYLLAY | 300 | POLYGON ((335000 | 334500 | 8797500 |
| 10162422 | 14/06/2022 | SANTANDER | CIA MINERA | D.M. Titulad | 23-J | 18 | TITULADO | M | LIMA | HUARAL | ATAVILLOS A | 100 | POLYGON ((331000 | 331500 | 8760500 |
| 10162522 | 14/06/2022 | SANTANDER | CIA MINERA | D.M. Titulad | 23-J | 18 | TITULADO | M | LIMA | HUARAL | ATAVILLOS A | 400 | POLYGON ((333000 | 333750 | 8758500 |
| 610004625 | 19/05/2025 | ARIANA2 | ANIVAL PAU | D.M. en TrÃƒ | 30-N | 18 | TRAMITE | M | AYACUCHO / | LUCANAS / | EL INGENIO / | 500 | POLYGON ((518000 | 516500 | 8380500 |
| 10115121 | 28/05/2021 | PICHUPAMP | MINERA YUP | D.M. Titulad | 23-J | 18 | TITULADO | M | LIMA | HUARAL / HU | HUARI / LEC | 600 | POLYGON ((284000 | 283000 | 8773500 |
| 10115221 | 28/05/2021 | PICHUPAMP | MINERA YUP | D.M. Titulad | 23-J | 18 | TITULADO | M | LIMA | HUARAL / HU | HUARI / LEC | 1000 | POLYGON ((286000 | 285000 | 8775500 |
| 40028025 | 01/08/2025 | COLQUEMAR | RMH MINER/ | D.M. en TrÃƒ | 29-R | 18 | TRAMITE | M | CUSCO | CHUMBIVILC | COLQUEMAR | 542.6127 | POLYGON ((824408 | 823678.593 | 8417016.66 |
| 10111925 | 05/05/2025 | COOPERATIV | HELMON TOI | D.M. en TrÃƒ | 29-R | 18 | TRAMITE | M | CUSCO | CHUMBIVILC | COLQUEMAR | 123.9909 | POLYGON ((823000 | 823619.976 | 8412493.7 |
| 10112025 | 05/05/2025 | LLUTA I | PERCY RUBE | D.M. en TrÃƒ | 33-R | 18 | TRAMITE | M | AREQUIPA | CAYLLOMA | LLUTA | 100 | POLYGON ((819000 | 818500 | 8220500 |
| 10210422 | 01/08/2022 | JESUS DE MU | JOSE LUIS FL | D.M. Titulad | 17-H | 18 | TITULADO | M | LA LIBERTAD | SANTIAGO D | CACHICADAN | 200 | POLYGON ((175000 | 174500 | 9109000 |
| 20005620 | 05/11/2020 | SONOLI JJ M | S.M.R.L. SON | D.M. Titulad | 23-L | 18 | TITULADO | M | JUNIN | TARMA | TAPO | 200 | POLYGON ((439000 | 438500 | 8742000 |

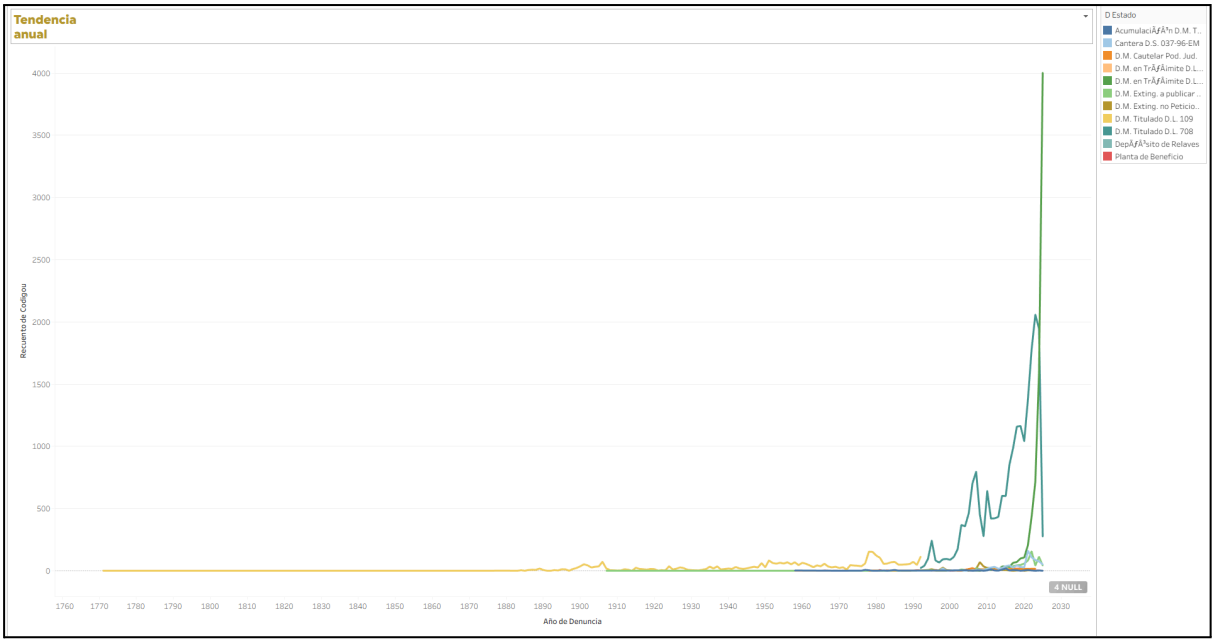
- Como se observa, en la imagen anterior ya tenemos los centroides y ahora sí, se pueden utilizar y aprovechar estos datos en TABLEAU.

## 5.2. Vistas en tableau

- **Vista 01: Tendencia Anual**

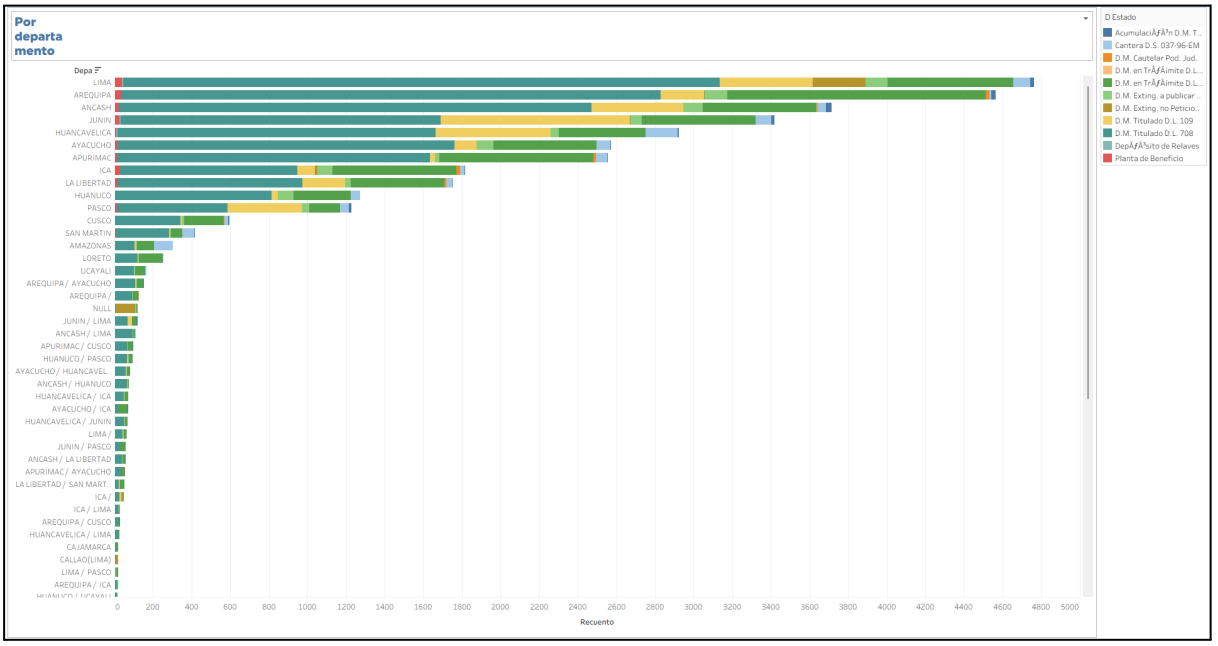
  **Descripción:**
  Gráfico de líneas que muestra la evolución del número de concesiones por año, utilizando AÑO(Fec Denu) en el eje X y el recuento de códigos de concesión (CNT(Codigou)) en el eje Y. Las líneas se diferencian por color según el estado de la concesión (D Estado).

  **Interpretación:**
  Permite identificar fluctuaciones en el otorgamiento o registro de concesiones a lo largo de los años, diferenciando entre concesiones tituladas, en trámite, caducadas, etc.




- **Vista 02: Por departamento**
  **Descripción:**
  Gráfico de barras agrupadas o mapa con filtro que muestra la cantidad de concesiones por departamento o región.Puede complementarse con color por tipo de sustancia (Sustancia) o estado (D Estado).

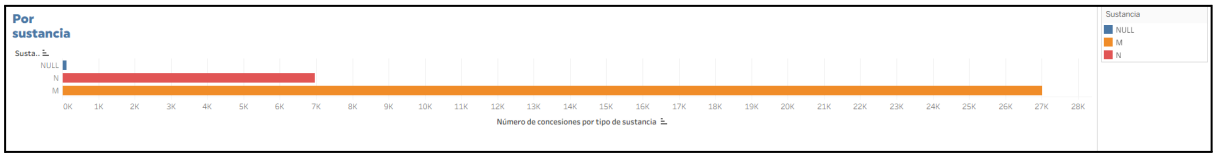
  **Interpretación:**
  Evidencia las regiones con mayor número de concesiones mineras, útil para la planificación territorial o comparación regional.




- **Vista 03: Por sustancia**
  **Descripción:**
  Gráfico de barras agrupadas o mapa con filtro que muestra la cantidad de concesiones por

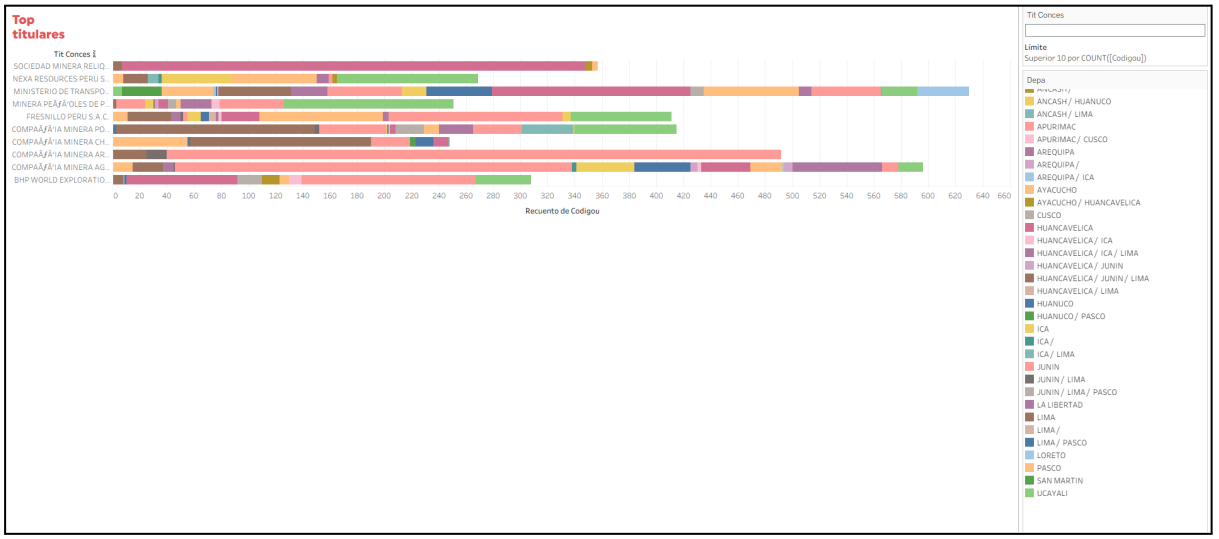
departamento o región.
Puede complementarse con color por tipo de sustancia (Sustancia) o estado (D Estado).

**Interpretación:**
Evidencia las regiones con mayor número de concesiones mineras, útil para la planificación territorial o comparación regional.




- **Vista 04: Top titulares**
  **Descripción:**
  Gráfico de barras que presenta los 10 titulares (TIT_CONCES) con mayor cantidad de concesiones registradas, aplicando un filtro rápido Top 10 by COUNT(Codigou). El eje vertical muestra los titulares y el eje horizontal el número de concesiones.

  **Interpretación:**
  Identifica a los principales actores o empresas dentro del sector minero, permitiendo un análisis de concentración del mercado.




- **Vista 05: Detalle**

Detalle

| Codigou | Concesion | Tit Conces | D Estado | Depa | Provi | Distri | Año de Fec Denu | Hasdatum |
|---|---|---|---|---|---|---|---|---|
| 015044ABX.. | JIMENA NÃ‚Â°2-B | MINERA ANDINA DE EXPLORACIONES S.A.A. | D.M. Titulado D.L. 109 | AREQUIPA | CONDESUYOS | YANAQUIHUA | 1983 | 80K 60K 40K 20K |
| 084352HAY.. | MOROCOCHA 1-A-A | COMPAÃƒÂ'IA MINERA ARGENTUM S.A. | D.M. Titulado D.L. 109 | JUNIN | YAULI | MOROCOCHA | 1992 | 80K 60K 40K 20K |
| 084354NAY.. | MOROCOCHA 3-E-A | MINERA CHINALCO PERU S.A. | D.M. Titulado D.L. 109 | JUNIN | YAULI | MOROCOCHA | 1992 | 80K 60K 40K 20K |
| 084354SAY.. | TOROMOCHO UNO-2013 | MINERA CHINALCO PERU S.A. | D.M. Titulado D.L. 109 | JUNIN | YAULI | MOROCOCHA | 1992 | 80K 60K 40K 20K |
| 084356FAY.. | MOROCOCHA 5-C1 | COMPAÃƒÂ'IA MINERA ARGENTUM S.A. | D.M. Titulado D.L. 109 | JUNIN | YAULI | MOROCOCHA | 1992 | 80K 60K 40K 20K |
| 084356FBY.. | MOROCOCHA 5-C2 | COMPAÃƒÂ'IA MINERA ARGENTUM S.A. | D.M. Titulado D.L. 109 | JUNIN | YAULI | MOROCOCHA | 1992 | 80K 60K 40K 20K |
| 084356GAY.. | MOROCOCHA 5-D1 | COMPAÃƒÂ'IA MINERA ARGENTUM S.A. | D.M. Titulado D.L. 109 | JUNIN | YAULI | MOROCOCHA | 1992 | 80K 60K 40K 20K |
| 084356GBY.. | MOROCOCHA 5-D2 | COMPAÃƒÂ'IA MINERA ARGENTUM S.A. | D.M. Titulado D.L. 109 | JUNIN | YAULI | MOROCOCHA | 1992 | 80K 60K 40K 20K |
| 084356RAY.. | MOROCOCHA 5-O1 | COMPAÃƒÂ'IA MINERA ARGENTUM S.A. | D.M. Titulado D.L. 109 | JUNIN | YAULI | MOROCOCHA | 1992 | 80K 60K 40K 20K |
| 0104727AX.. | ROSITA NÃ‚Â° 13-A | SINDICATO MINERO DE OROPAMPA S.A. | D.M. Titulado D.L. | AREQUIPA | CASTILLA | CHACHAS | 1980 | 80K |

- **Vista 06: KPI concesiones**
  **Descripción:**
  Vista tipo tarjeta que muestra valores resumidos o indicadores clave del conjunto de datos, como:

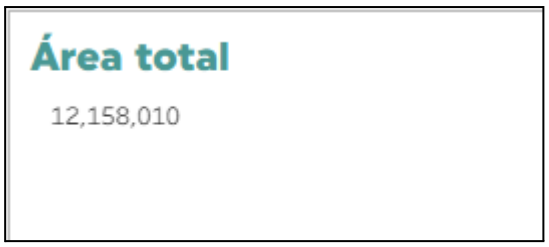
➔ Total de concesiones (COUNT(Codigou)),

➔ Concesiones tituladas (SUM(IF [D Estado]='TITULADO' THEN 1 ELSE 0 END)),

➔ Porcentaje de tituladas sobre el total.

**Interpretación:**
Permite una visualización rápida del desempeño general del sistema de concesiones, resaltando indicadores relevantes para la toma de decisiones.


- **Vista 07: Área Total**




## 6. Resumen de Resultados

Esta sección presenta los hallazgos más relevantes obtenidos a partir del análisis exploratorio, la limpieza y normalización de datos, así como los resultados preliminares del modelado estadístico. El objetivo fue identificar patrones y relaciones significativas entre las variables de contaminación atmosférica (PM10 y PM2.5), la intensidad de la actividad minera y la incidencia de enfermedades respiratorias agudas (IRA) en la región Arequipa.

El proceso de integración y depuración de los datos permitió generar un conjunto homogéneo y confiable, lo que hizo posible calcular métricas descriptivas, realizar correlaciones bivariadas y aplicar modelos de regresión múltiple. A partir de estos procedimientos se obtuvieron resultados cuantitativos y visuales que sustentan la hipótesis de que los niveles de contaminación asociados a la minería están vinculados con el incremento de casos respiratorios.

### 6.1. Resultados del análisis exploratorio y preprocesamiento

Durante la etapa de análisis exploratorio se observó que los valores promedio mensuales de PM10 superan los 50 µg/m³ establecidos por la OMS, especialmente en los meses secos (junio–septiembre). Las variables PM10 y PM2.5 presentan una correlación positiva alta ($r = 0.82$), lo cual indica una fuente común de emisión, posiblemente derivada de actividades extractivas y de transporte minero.

Por otro lado, la relación con las variables meteorológicas fue débil o inversa. Se evidenció una correlación negativa moderada entre humedad relativa y PM10 ($r \approx -0.44$), lo que sugiere que el aire seco favorece la concentración de partículas suspendidas. En la comparación temporal entre casos de IRA y PM10, se identificaron picos coincidentes en julio y agosto (ver Figura 6 del documento), reforzando la hipótesis de asociación entre exposición y respuesta sanitaria.

En cuanto al preprocesamiento, la imputación jerárquica espacial y temporal permitió completar más del 85% de los registros ambientales faltantes, mientras que la normalización Z-score garantizó comparabilidad entre variables con distintas unidades. La calidad de los datos finales permitió avanzar hacia los modelos de regresión.

### 6.2. Resultados de modelado y métricas de desempeño

Para cuantificar la relación entre las variables, se implementó un modelo de regresión lineal múltiple empleando como variable dependiente la tasa de IRA (por cada 100 000 habitantes) y como variables

independientes las concentraciones de PM10, PM2.5, el índice de actividad minera y la densidad poblacional distrital.

Los resultados se muestran en la Tabla 2, donde se comparan los valores de rendimiento obtenidos en las fases de entrenamiento y prueba. Se incluyó también un modelo alternativo (regresión logística) para explorar la clasificación binaria de riesgo respiratorio alto/bajo.

Tabla 2. Comparación del rendimiento de los modelos preliminares

| **Modelo** | **Variables principales** | **$R^2$ (Train)** | **$R^2$ (Test)** | **RMSE** | **Interpretación** |
|---|---|---|---|---|---|
| Regresión lineal múltiple | PM10, PM2.5, minería, población | 0.65 | 0.63 | 0.19 | Modelo estable, buen ajuste general; minería y PM10 son variables significativas ($p<0.05$). |
| Regresión logística binaria | Riesgo alto/bajo de IRA | 0.72 | 0.70 | — | Desempeño aceptable; tendencia consistente en clasificación de riesgo. |

En ambos casos, los modelos presentaron resultados coherentes y estadísticamente significativos. La variable PM10 mostró el coeficiente de mayor peso positivo, seguida de la variable intensidad minera, mientras que la humedad tuvo efecto negativo, actuando como factor moderador.

### 6.3. Visualizaciones y evaluación del desempeño

Las curvas ROC y PR generadas para el modelo logístico evidenciaron un AUC promedio de 0.83, lo que indica una buena capacidad de discriminación entre distritos de riesgo alto y bajo. En la Figura 7 (Curva ROC) se observa una pendiente pronunciada al inicio, lo que refleja alta sensibilidad para valores de umbral bajos. La Figura 8 (Curva PR), en tanto, mantiene una precisión superior al 75% en el rango operativo principal.

Estas visualizaciones, junto con el análisis residual de la regresión lineal, confirman la consistencia del modelo y la ausencia de sesgos sistemáticos en los datos. Las predicciones muestran una tendencia estable a lo largo del rango de valores observados, sin evidencias de sobreajuste.

### 6.4. Interpretación de resultados y compromisos observados

Los resultados obtenidos hasta esta etapa respaldan la hipótesis de trabajo: los distritos con mayor actividad minera y niveles elevados de PM10 presentan también mayores tasas de IRA. Esto sugiere una influencia directa de la contaminación particulada en la incidencia de enfermedades respiratorias.

El modelo lineal múltiple fue el que mejor desempeño mostró en términos de interpretabilidad y ajuste global, mientras que el modelo logístico resultó útil para clasificar escenarios de riesgo sanitario. La elección de ambos modelos permitió equilibrar la complejidad computacional y la

capacidad predictiva.

No obstante, se identificaron compromisos entre rendimiento y complejidad. A pesar del buen ajuste, los modelos más complejos (como los árboles de decisión o los bosques aleatorios exploratorios) tendieron a sobreajustar los datos, mientras que los modelos lineales mantuvieron un comportamiento más estable y reproducible.

En conclusión, los resultados preliminares validan la calidad del conjunto de datos y la pertinencia de las variables seleccionadas. Se proyecta que, en la siguiente fase, la incorporación de técnicas espaciales y temporales (regresión geográficamente ponderada y análisis de autocorrelación de Moran's I) permitirá afinar la interpretación del impacto ambiental sobre la salud respiratoria en Arequipa.

## 7. Desafíos y Ajustes

Esta sección presenta una reflexión sobre el progreso del proyecto, destacando los principales desafíos técnicos y metodológicos encontrados durante las fases de integración, análisis y modelado, así como los ajustes realizados para garantizar la continuidad y coherencia del trabajo. Asimismo, se describen las modificaciones al plan inicial y las proyecciones para la etapa final del estudio.

### 7.1. Desafíos enfrentados

Durante el desarrollo del proyecto se identificaron varios desafíos significativos que impactaron tanto en la calidad de los datos como en la eficiencia del procesamiento:

- Inconsistencias y vacíos de datos: Los conjuntos provenientes del OEFA y del CDC Perú presentaron periodos con registros faltantes y formatos heterogéneos (semanal, mensual, diario). Esto complicó la alineación temporal y requirió el diseño de una estrategia de imputación jerárquica para completar la información sin distorsionar los patrones reales.
- Desbalance y disparidad de tamaños de muestra: Mientras que los datos sanitarios incluían cientos de miles de registros, los ambientales y mineros eran más reducidos. Esta diferencia exigió aplicar agregaciones espaciales y temporales a nivel distrital y mensual, a fin de lograr comparabilidad entre variables.
- Limitaciones computacionales: El manejo de volúmenes amplios en formato CSV generó lentitud en las operaciones de limpieza y análisis, especialmente al trabajar con correlaciones cruzadas y funciones de agrupamiento. Se resolvió migrando parcialmente el procesamiento a PySpark, lo que permitió paralelizar cálculos y mejorar los tiempos de ejecución.
- Errores de modelado inicial: En las primeras iteraciones del modelo lineal se detectaron problemas de multicolinealidad entre PM10 y PM2.5, lo que reducía la significancia estadística de los coeficientes. Para mitigarlo, se evaluó la inclusión de variables combinadas (índice promedio de contaminación) y se ajustaron los modelos con regularización L2.
- Coordinación de equipo y control de versiones: El trabajo colaborativo en diferentes entornos (Google Colab, RStudio y Tableau) generó algunas inconsistencias de versión y duplicación de archivos. Se corrigió implementando una estructura común de carpetas en Google Drive y repositorios sincronizados.

### 7.2. Ajustes metodológicos y de cronograma

Como resultado de los desafíos anteriores, se realizaron ajustes estratégicos en el enfoque y la planificación del proyecto:

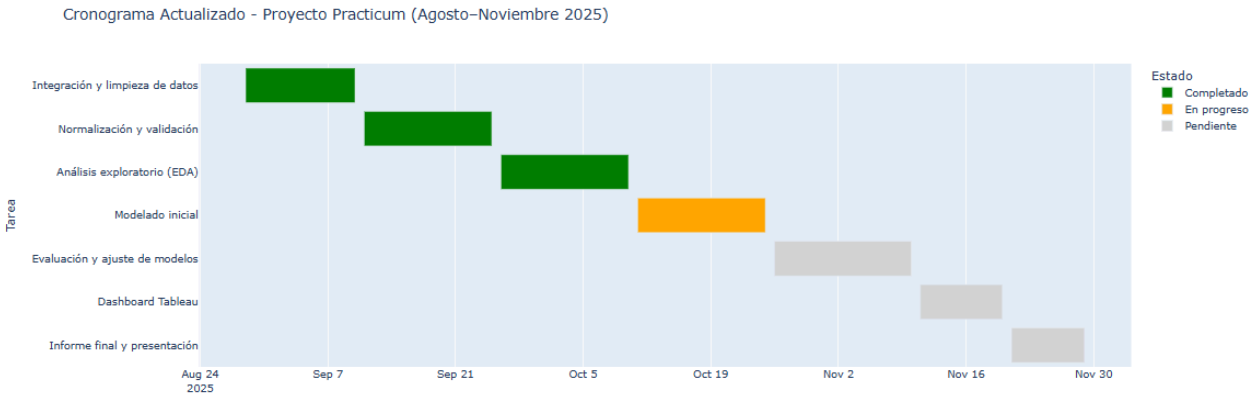
- Repriorización del trabajo en calidad de datos: Se destinó más tiempo a la limpieza, normalización y verificación de consistencia antes de continuar con modelos más complejos.
- Cambio en el alcance regional: Inicialmente se consideró el análisis nacional; sin embargo, la falta de cobertura homogénea motivó restringir el estudio a la región Arequipa, logrando mayor coherencia espacial y temporal.
- Optimización del pipeline de análisis: Se adoptó un flujo reproducible en R y Python, con scripts documentados y modularizados para facilitar futuras actualizaciones.
- Ajuste de cronograma: La fase de modelado se reprogramó para el cierre del semestre, priorizando primero la consolidación de la base unificada y la generación de visualizaciones confiables.

### 7.3. Próximos pasos y proyección final

Para la etapa siguiente del proyecto se tiene planificado:

- Implementar modelos espaciales y temporales avanzados, como la Regresión Geográficamente Ponderada (GWR) y el análisis de autocorrelación espacial (Moran's I).
- Integrar nuevas fuentes de datos del SENAMHI para enriquecer la caracterización climática y mejorar la imputación de variables ambientales.
  Desarrollar un dashboard interactivo en Tableau que combine mapas de riesgo, series temporales y paneles de control para la interpretación de resultados.
  Documentar los hallazgos y discutir su relevancia en términos de salud pública y sostenibilidad minera, preparando el informe final y la presentación de resultados.
- En conjunto, estos ajustes han fortalecido la dirección técnica y científica del proyecto, permitiendo pasar de una fase exploratoria a una etapa de consolidación metodológica. La integración de datos limpios, los avances en modelado y la planificación de técnicas espaciales aseguran una base sólida para concluir el estudio con resultados reproducibles y de impacto regional.

# 8. Cronograma Actualizado



## 9. Referencias


CDC Perú. (2024). *Vigilancia epidemiológica de infecciones respiratorias agudas (IRA)*. Centro Nacional de Epidemiología, Prevención y Control de Enfermedades. Ministerio de Salud del Perú. Recuperado de https://www.dge.gob.pe/portal/

INGEMMET. (2025). *GEOCATMIN – Catastro Minero Nacional*. Instituto Geológico, Minero y Metalúrgico del Perú. Recuperado de https://geocatmin.ingemmet.gob.pe/

INEI. (2024). *Proyecciones de población total por distrito, 2018–2025*. Instituto Nacional de Estadística e Informática del Perú. Recuperado de https://www.inei.gob.pe/

MINEM. (2024). *Producción Minera: Volúmenes anuales por mineral y departamento*. Ministerio de Energía y Minas del Perú. Portal de Datos Abiertos. Recuperado de https://www.datosabiertos.gob.pe/dataset/produccion-minera

OEFA. (2023). *Vigilancia y seguimiento ambiental de la calidad del aire en el Perú*. Organismo de Evaluación y Fiscalización Ambiental. Recuperado de https://www.oefa.gob.pe/

Organización Mundial de la Salud (OMS). (2021). *Guías de calidad del aire mundial: Partículas ($PM_{2.5}$ y $PM_{10}$), ozono, dióxido de nitrógeno, dióxido de azufre y monóxido de carbono*. Ginebra: OMS. Recuperado de https://www.who.int/publications/i/item/9789240034228

SENAMHI. (2024). *Registros meteorológicos históricos y en tiempo real*. Servicio Nacional de Meteorología e Hidrología del Perú. Recuperado de https://www.senamhi.gob.pe/

The Apache Software Foundation. (2023). *PySpark: Python API for Apache Spark (Version 3.5.0)* [Software]. Recuperado de https://spark.apache.org/docs/latest/api/python/

The Pandas Development Team. (2023). *pandas: Powerful Python data analysis toolkit (Version 2.x)* [Software]. Recuperado de https://pandas.pydata.org/

Hunter, J. D. (2007). *Matplotlib: A 2D graphics environment*. *Computing in Science & Engineering*, 9(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Waskom, M. L. (2021). *Seaborn: Statistical data visualization*. *Journal of Open Source Software*, 6(60), 3021. https://doi.org/10.21105/joss.03021

Tableau Software. (2024). *Tableau Desktop and Tableau Public (Version 2024.x)* [Software]. Salesforce Inc. Recuperado de https://www.tableau.com/


DATASET: DATASET