PROYECTO INTRO CD

“Ubicación, Clima y Calidad del Aire: Un Análisis Global”

INTERANTE DEL EQUIPO:
Alumno: José Alexander Caballero Palma
Número de lista: 4

Nombre del profesor:
Jaime Alejandro Romero Sierra

Materia:
Introducción a la Ciencia de Datos

INTRODUCCIÓN

OBJETIVO DEL PROYECTO:

El propósito de este proyecto es comprender como las condiciones climáticas afectan la calidad del aire en diferentes partes del mundo. De igual forma, estudiar los datos disponibles para identificar patrones que nos ayuden a explicar las variaciones en los niveles de contaminación dependiendo la ubicación geográfica. También me interesa explorar si aspectos astronómicos, como las fases de la luna o los horarios del amanecer y atardecer, tienen algún impacto en el clima o la calidad del aire. Con este análisis, se busca identificar irregularidades, entender tendencias a nivel global, y generar información útil para crear modelos que sirvan de apoyo en la toma de decisiones relacionadas con el medio ambiente.

JUSTIFICACIÓN Y CONTEXTO:

Estudiar todo lo relacionado con el entorno en diferentes ubicaciones geográficas es fundamental para comprender cómo las condiciones meteorológicas y ambientales influyen en la contaminación. Este análisis permite identificar patrones y tendencias que pueden ayudar a mitigar los efectos negativos de la contaminación, desarrollar estrategias más efectivas para su gestión y adaptarnos mejor a los cambios provocados por el clima. Además, explorar estas dinámicas fomenta una comprensión más completa del medio ambiente y sus interacciones.

FUENTES DE DATOS:

La base de datos fue encontrada en la página web llamada “Kaggle”, después de una busqueda de algun DataFrame, este fue por el que me decante. La base de datos cuenta con 33,417 filas **y** 40 columnas, lo que representa un total de 1,336,680 datos. Podemos observar una gran cantidad de datos e información acerca de diferentes ubicaciones geofráficas, datos meteorológicos, datos sobre la calidad del aire e incluso información astronómica.

METODOLOGÍA

PROCESO DE LIMPIEZA:

Antes de comenzar a limpiar la base de datos, realice un análisis preliminar para comprender la naturaleza y la distribución de los errores. Primero que nada comence haciendo un reusmen estadístico de los datos, después calcule el porcentaje de valores faltantes por columna, igualemente identifique si existían filas duplicadas y finalmente, analice los tipos de datos de las columnas y si eran consistentes con el contenido esperado.

Una vez realizado el análisis, procedi con las diferentes tareas de limpieza para la base de datos, comencé haciendo una eliminación o imputación de valores faltantes, donde decídi si eliminaba filas o olumnas con valores NaNs o si prefería utilizar técnicas de imputación. De igual forma, ocupe otra tarea de limpiza la cual fue, la eliminación de filas duplicadas. Otra tarea que realice, fue la de correción de tipos de datos, donde me asegure de que las columnas tengan sus tipos de datos adecuados. Finalmente, la última tarea que realice fue la de la correción de los valores “invalid”, por ejemplo los valores bbb.

ANÁLISIS EXPLORATORIO DE LOS DATOS (EDA)

1. **Descripción General de los Datos**
    - **Visión general:** El dataset tiene un total de 33,417 registros y 40 variables. Cada fila representa datos meteorológicos y de calidad del aire recopilados en diferentes ubicaciones alrededor del mundo, incluyendo información sobre las condiciones del clima, la calidad del aire y datos astronómicos.

    - **Tipos de Variables:**
        - **Categóricas**:
            - **country**: País de la ubicación.
            - **location_name**: Nombre de la ubicación o ciudad.
            - **timezone**: Zona horaria de la ubicación.
            - **condition_text**: Descripción del clima (por ejemplo, "soleado", "nublado").
            - **wind_direction**: Dirección del viento (por ejemplo, "norte", "sur").

    - **sunrise**, **sunset**, **moonset**, **moon_phase**: Información astronómica relacionada con la salida y puesta del sol, y fases de la luna.

- **Numéricas**:

    - **latitude**, **longitude**: Coordenadas geográficas.
    - **temperature_celsius**, **temperature_fahrenheit**: Temperatura en grados Celsius y Fahrenheit.
    - **wind_mph**, **wind_kph**: Velocidad del viento en millas por hora y kilómetros por hora.
    - **pressure_mb**, **pressure_in**: Presión atmosférica en milibares y pulgadas.
    - **precip_mm**, **precip_in**: Precipitación en milímetros y pulgadas.
    - **humidity**, **cloud**, **visibility_km**, **visibility_miles**: Humedad, nubosidad y visibilidad.
    - **gust_mph**, **gust_kph**: Velocidad de ráfagas de viento en millas y kilómetros por hora.
    - **air_quality_**: Variables relacionadas con la calidad del aire, como el monóxido de carbono, ozono, dióxido de nitrógeno, dióxido de azufre, PM2.5, PM10, etc.
    - **uv_index**: Índice UV (radiación ultravioleta).

- **Fechas**:

    - **last_updated**: Fecha de la última actualización de los datos.

- **Resumen estadístico:**

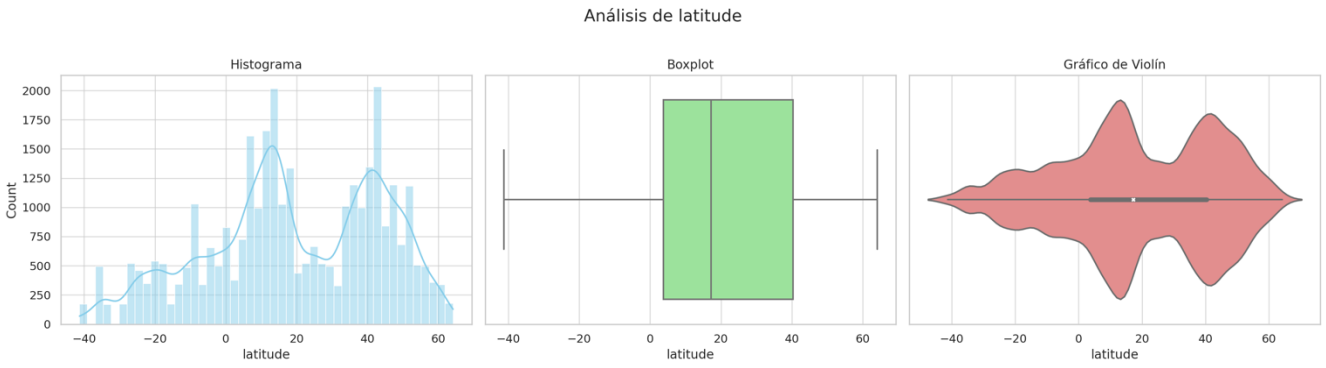
```
resumen_estadístico = df.describe(include='all')
print("Resumen estadístico:\n", resumen_estadístico)
```

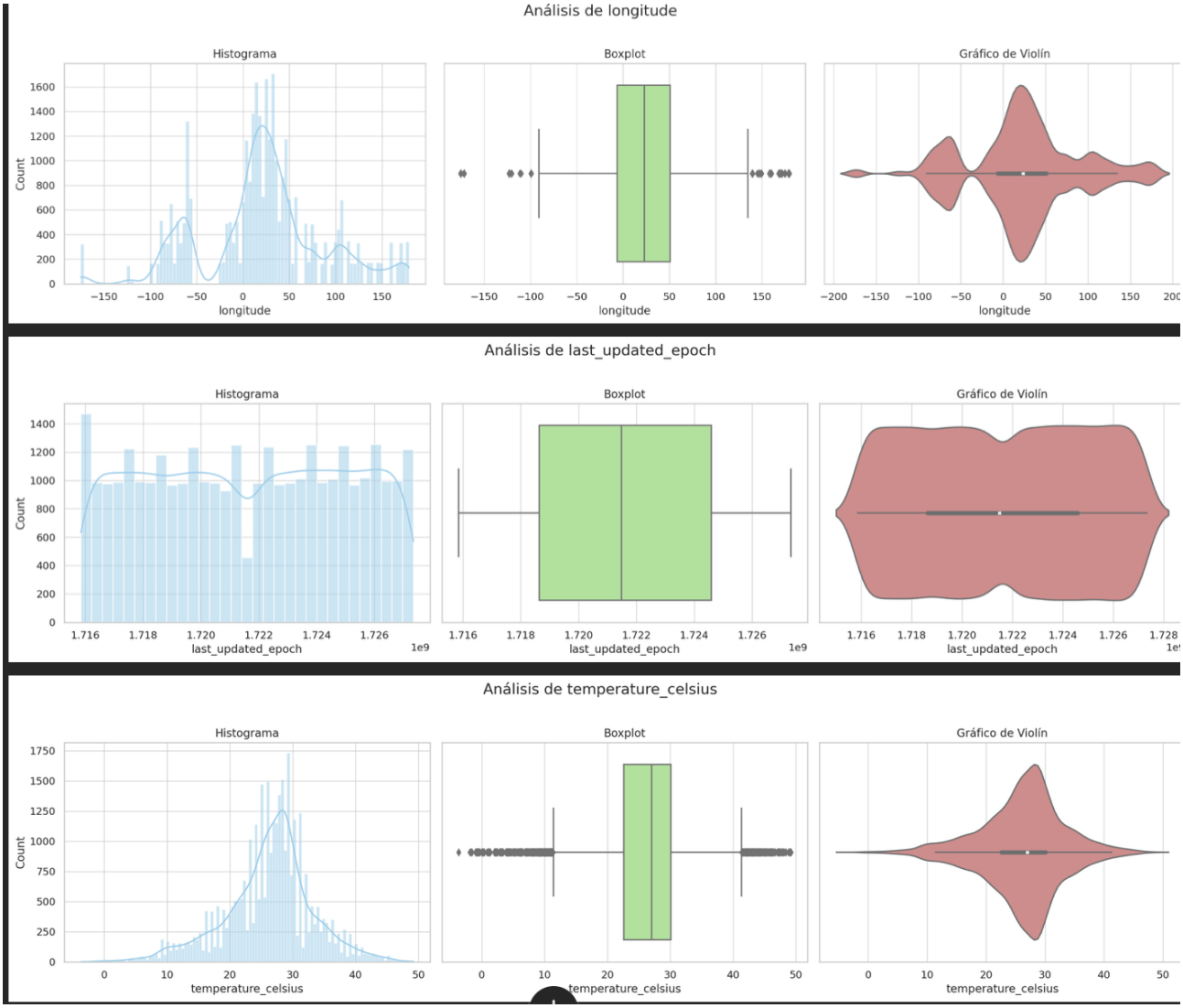
```
Resumen estadístico:
          country location_name latitude longitude      timezone  \
count       33099         33115    33083     33076         33116
unique        211           248      235       240           191
top     Bulgaria         aaaaa     40.4     aaaaa  Asia/Bangkok
freq          369           326      343       382           604
mean          NaN           NaN      NaN       NaN           NaN
std           NaN           NaN      NaN       NaN           NaN
min           NaN           NaN      NaN       NaN           NaN
25%           NaN           NaN      NaN       NaN           NaN
50%           NaN           NaN      NaN       NaN           NaN
75%           NaN           NaN      NaN       NaN           NaN
max           NaN           NaN      NaN       NaN           NaN

        last_updated_epoch last_updated temperature_celsius  \
count                33045        33080               33117
unique                 183         3776                 476
top                 aaaaa        aaaaa                28.0
freq                   352          322                 592
mean                   NaN          NaN                 NaN
std                    NaN          NaN                 NaN
min                    NaN          NaN                 NaN
25%                    NaN          NaN                 NaN
50%                    NaN          NaN                 NaN
75%                    NaN          NaN                 NaN
max                    NaN          NaN                 NaN

        temperature_fahrenheit condition_text  ... air_quality_PM2.5  \
count                33084          33084  ...              33075
unique                 794             30  ...               1818
top                   84.2  Partly cloudy  ...                0.5
freq                   535          11302  ...               1277
mean                   NaN            NaN  ...                NaN
std                    NaN            NaN  ...                NaN
min                    NaN            NaN  ...                NaN
25%                    NaN            NaN  ...                NaN
50%                    NaN            NaN  ...                NaN
75%                    NaN            NaN  ...                NaN
max                    NaN            NaN  ...                NaN
```

```
       air_quality_PM10 air_quality_us-epa-index air_quality_gb-defra-index  \
count             33100                   33078                       33093
unique             2717                       7                          11
top                 0.6                     1.0                         1.0
freq                360                   24084                       21874
mean                NaN                     NaN                         NaN
std                 NaN                     NaN                         NaN
min                 NaN                     NaN                         NaN
25%                 NaN                     NaN                         NaN
50%                 NaN                     NaN                         NaN
75%                 NaN                     NaN                         NaN
max                 NaN                     NaN                         NaN

       sunrise sunset moonrise     moonset      moon_phase moon_illumination
count    33085  33104      0.0       33105           33071             33094
unique     309    408      NaN        1442               9                78
top   05:56 AM  aaaaa      NaN  No moonset  Waxing Gibbous               0.0
freq       428    340      NaN        1005            8137              1438
mean       NaN    NaN      NaN         NaN             NaN               NaN
std        NaN    NaN      NaN         NaN             NaN               NaN
min        NaN    NaN      NaN         NaN             NaN               NaN
25%        NaN    NaN      NaN         NaN             NaN               NaN
50%        NaN    NaN      NaN         NaN             NaN               NaN
75%        NaN    NaN      NaN         NaN             NaN               NaN
max        NaN    NaN      NaN         NaN             NaN               NaN

[11 rows x 41 columns]
```

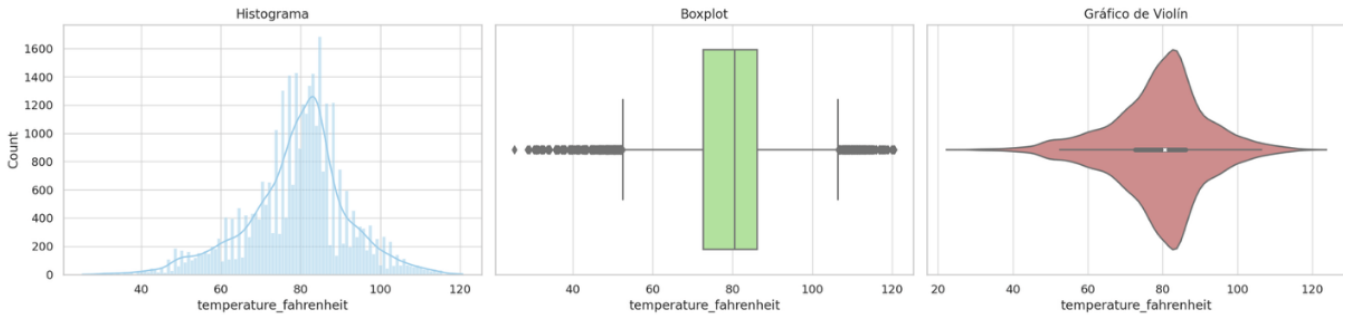
2. Visualización y Distribución de Variables Individuales:

- **Variables númericas**

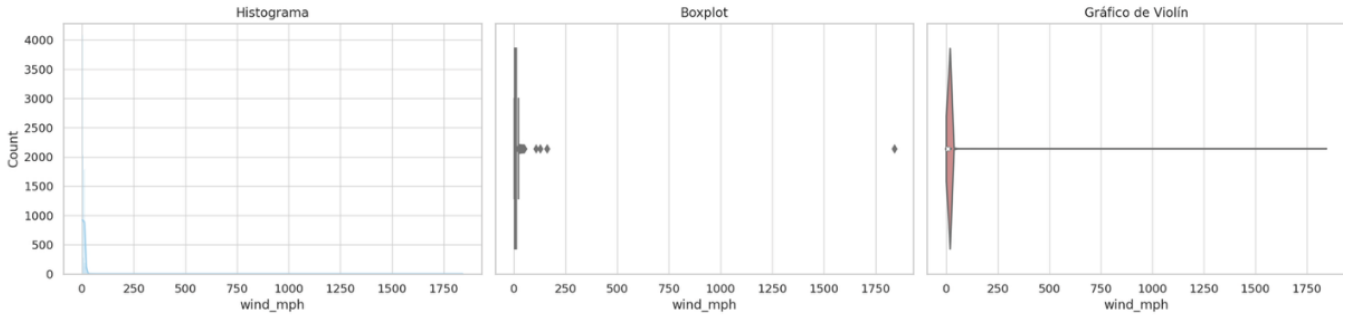



Análisis de longitude
Histograma
Boxplot
Gráfico de Violín
Count
longitude
Análisis de last_updated_epoch
Histograma
Boxplot
Gráfico de Violín
Count
last_updated_epoch
1e9
Análisis de temperature_celsius
Histograma
Boxplot
Gráfico de Violín
Count
temperature_celsius

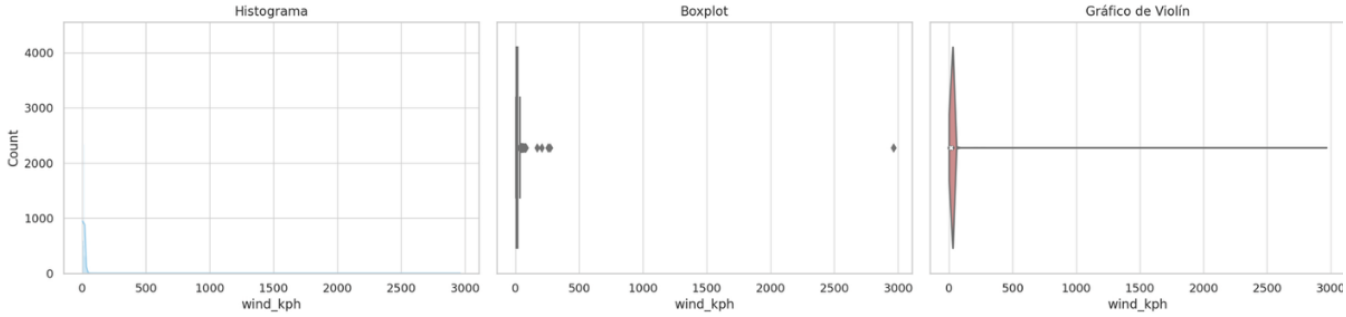
Análisis de temperature_fahrenheit



Análisis de wind_mph



Análisis de wind_kph

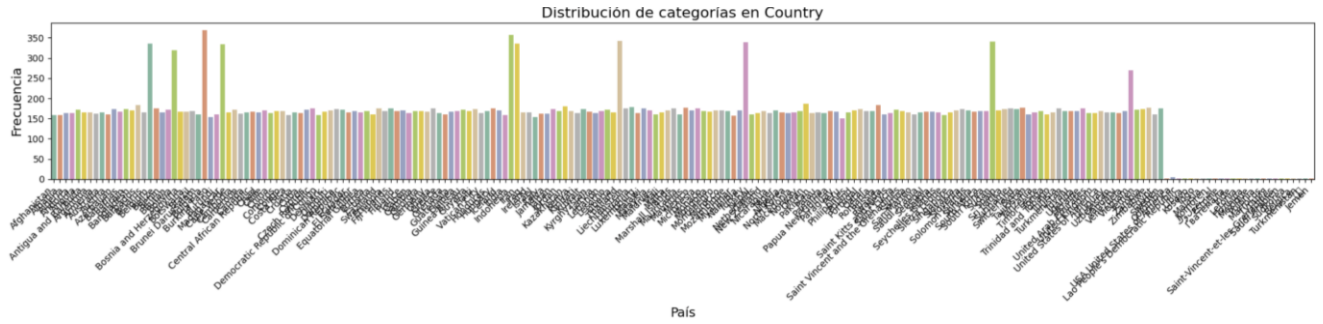


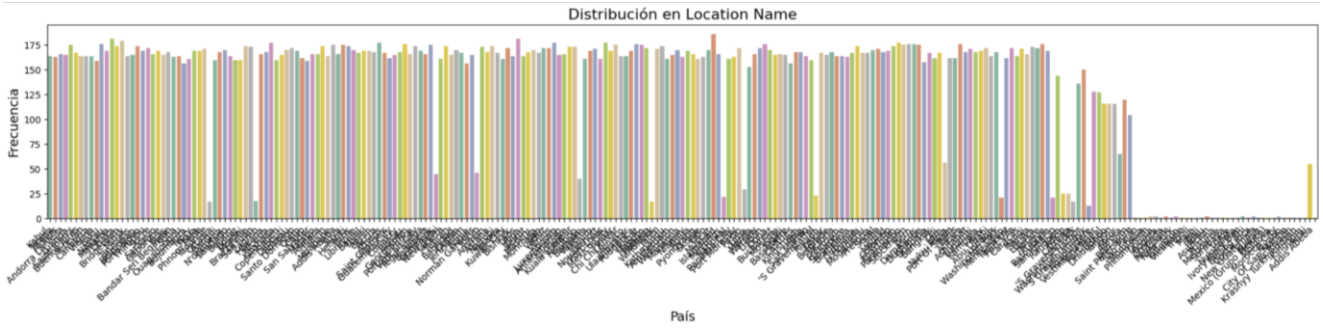
- **Distribuciones generales:**
    - Algunas variables parecen seguir una distribución normal (campana simétrica), mientras que otras presentan sesgo positivo o negativo. Esto podría indicar diferencias en cómo se distribuyen los datos o posibles transformaciones necesarias para análisis posteriores.
- **Valores atípicos:**
    - Los boxplots identifican claramente valores atípicos en ciertas variables. Estos puntos extremos podrían ser errores de medición, datos válidos o indicadores de subpoblaciones interesantes.

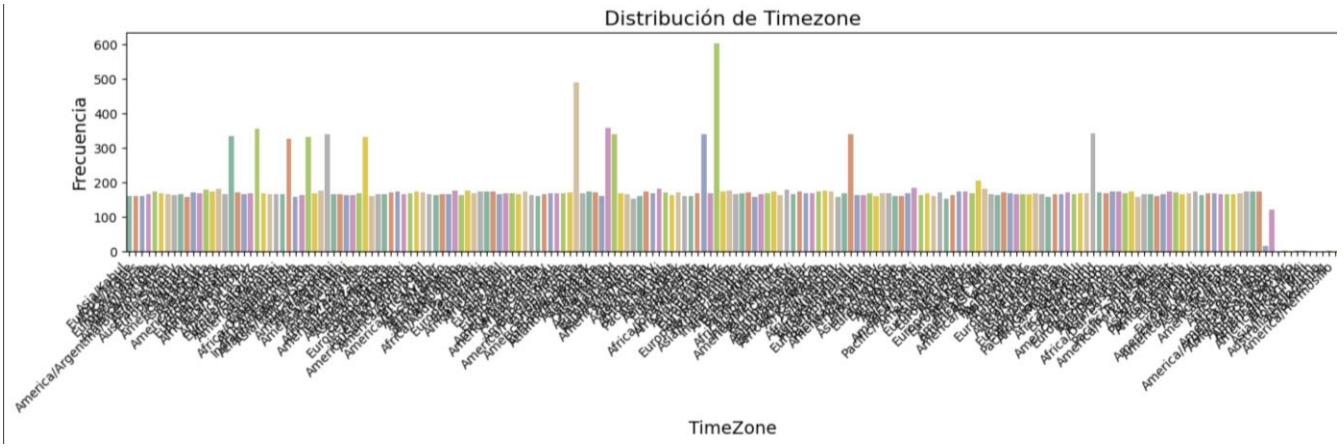
- **Simetría y dispersión:**

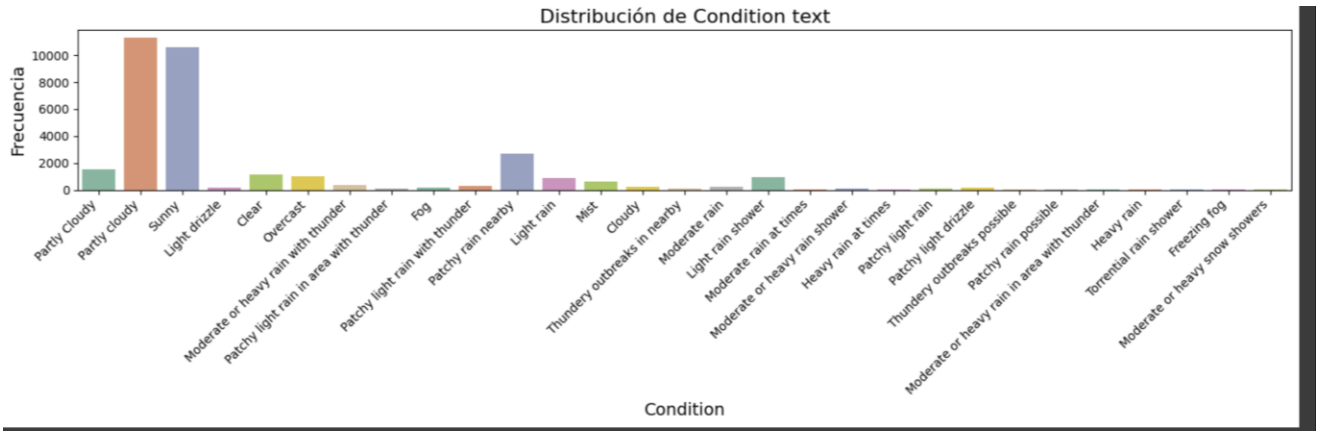
  - Los gráficos de violín destacan la simetría o asimetría y la densidad de los datos. Algunas variables muestran densidad concentrada en un rango estrecho, mientras que otras están más dispersas.

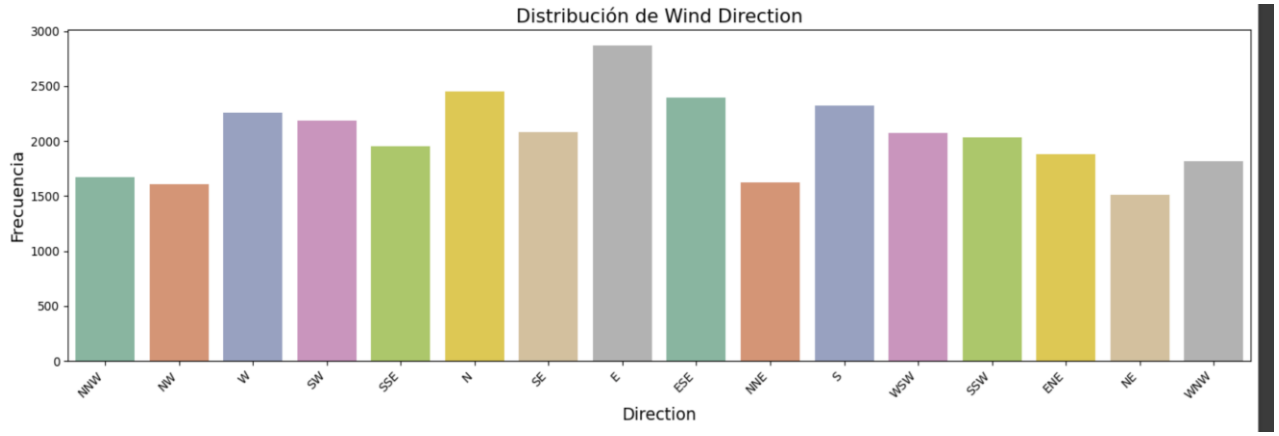
    - **Variables categotricas:**

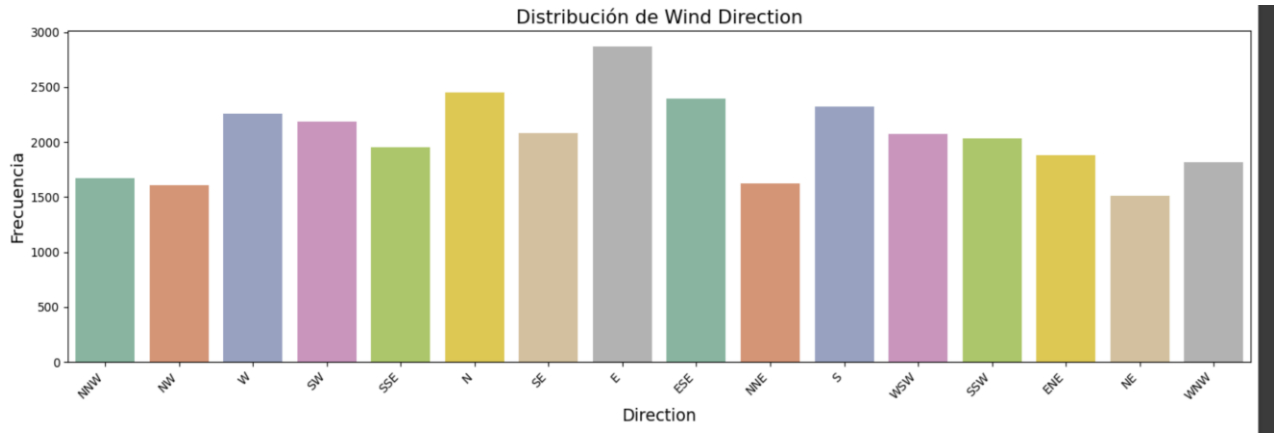












- **Países (`country`):** Si un país aparece con mucha mayor frecuencia que otros, puede indicar que los datos están sesgados hacia ciertas regiones.

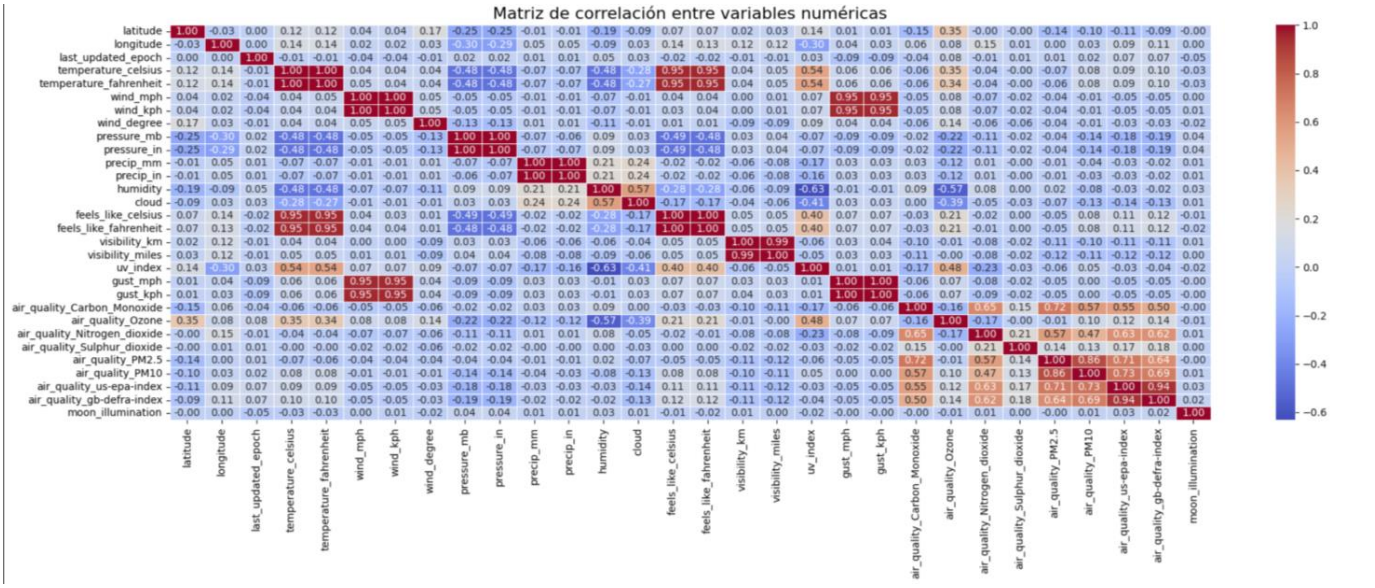
- **Nombre de la ubicación (`location_name`):** Si una ubicación tiene un número muy alto de registros, podrías identificar que el conjunto de datos está sesgado hacia esa ubicación.

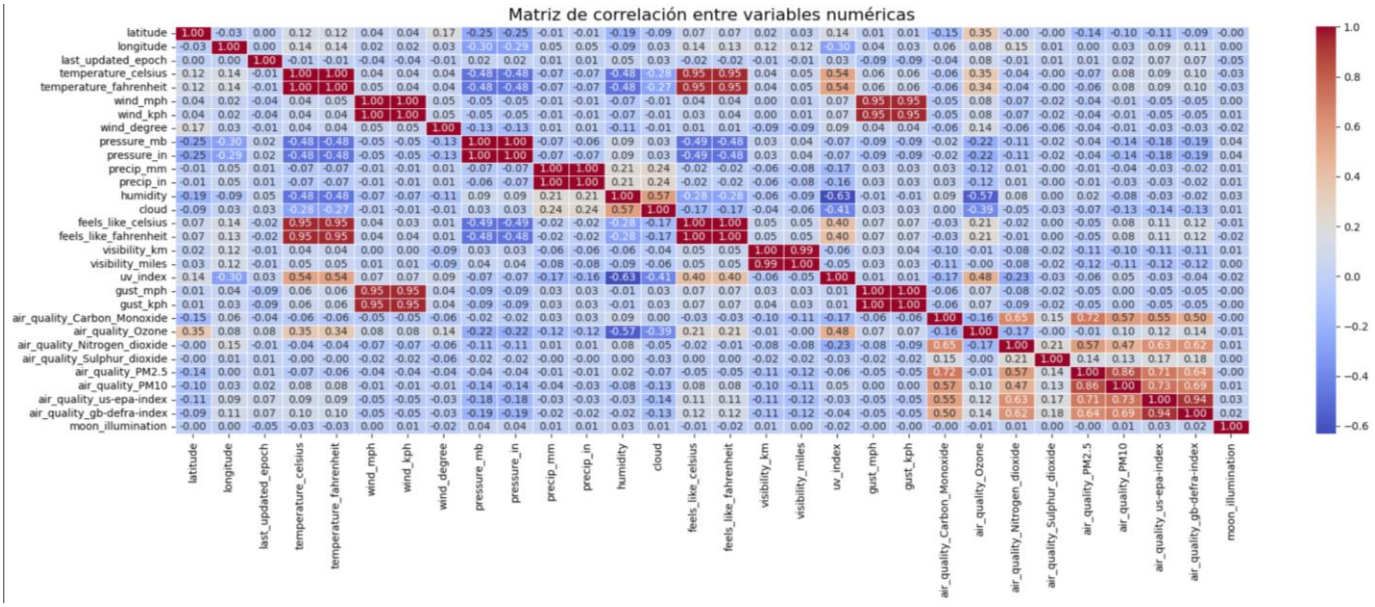
- **Descripción del clima (`condition_text`):** Si una categoría como "soleado" domina las observaciones, podría indicar que los datos provienen de días mayormente soleados.

- **Dirección del viento (`wind_direction`):** Si algunas direcciones dominan, esto puede reflejar un patrón geográfico o estacional.

3. Correlación entre Variables

• Matriz de Correlación





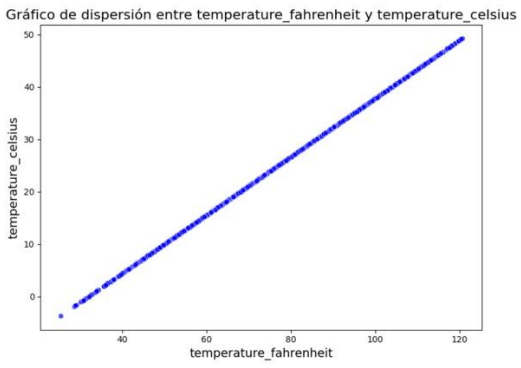
Análisis:

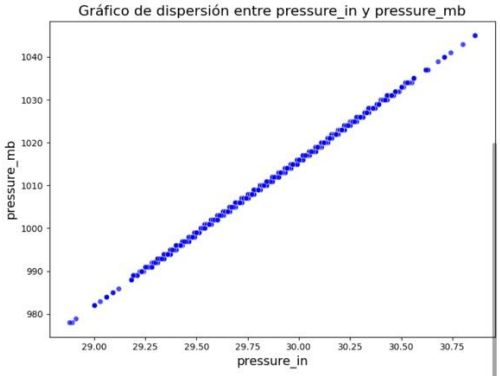
1. **Valores cercanos a +1 o -1** indican que las variables están muy relacionadas, ya sea de forma positiva (ambas aumentan o disminuyen juntas) o negativa (una aumenta mientras la otra disminuye).
2. **Valores cercanos a 0** indican que no hay una relación lineal fuerte entre las variables.

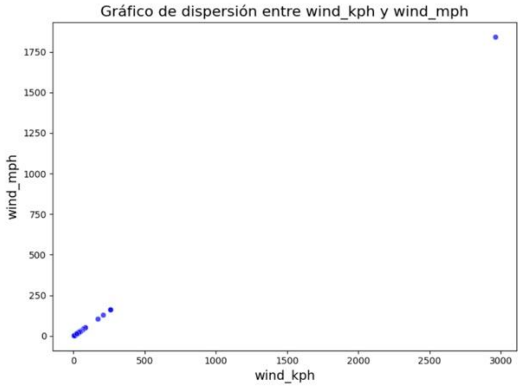
Posibles implicaciones para el modelo:

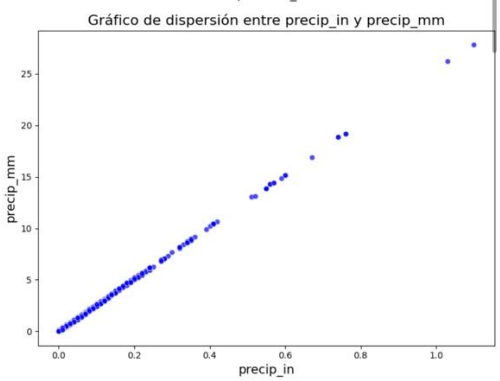
1. **Variables con alta correlación:** Si dos variables tienen una alta correlación (por ejemplo, > 0.8), se podría considerar combinar o eliminar una de ellas, ya que esto podría introducir redundancia en el modelo y causar problemas de multicolinealidad.
2. **Variables con baja correlación:** Si las variables tienen una baja correlación (cercanas a 0), puede que no estén proporcionando mucha información adicional para predecir el objetivo del modelo.

- **Parejas de Variables:**









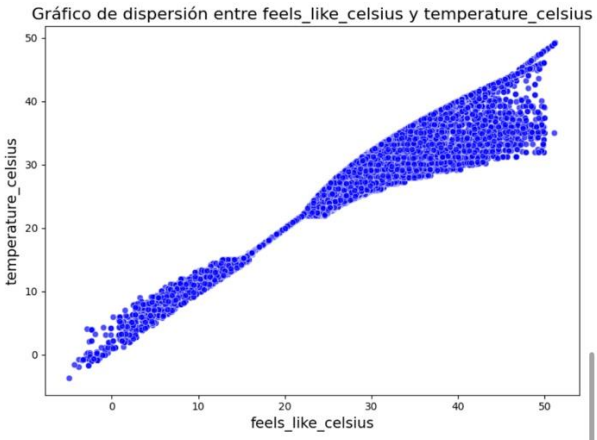


Gráfico de dispersión entre feels_like_celsius y temperature_celsius
temperature_celsius
feels_like_celsius

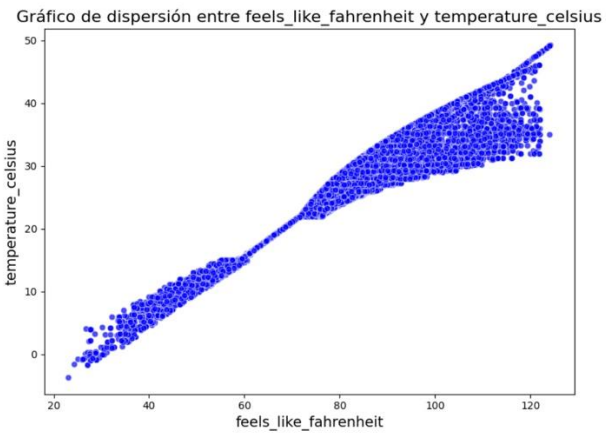


Gráfico de dispersión entre feels_like_fahrenheit y temperature_celsius
temperature_celsius
feels_like_fahrenheit

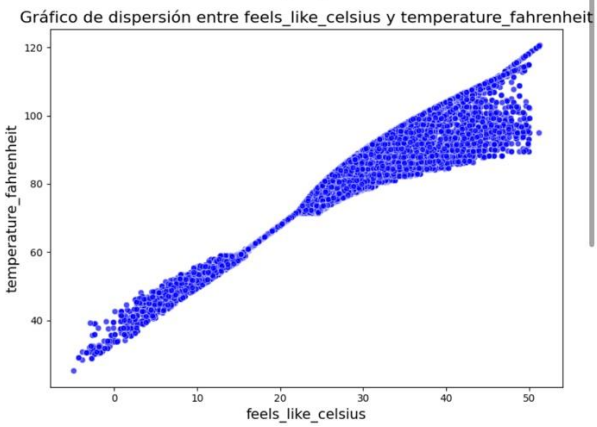


Gráfico de dispersión entre feels_like_celsius y temperature_fahrenheit
temperature_fahrenheit
feels_like_celsius

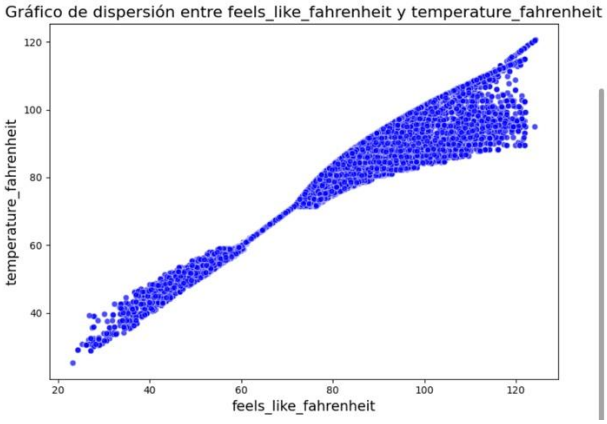


Gráfico de dispersión entre feels_like_fahrenheit y temperature_fahrenheit
temperature_fahrenheit
feels_like_fahrenheit

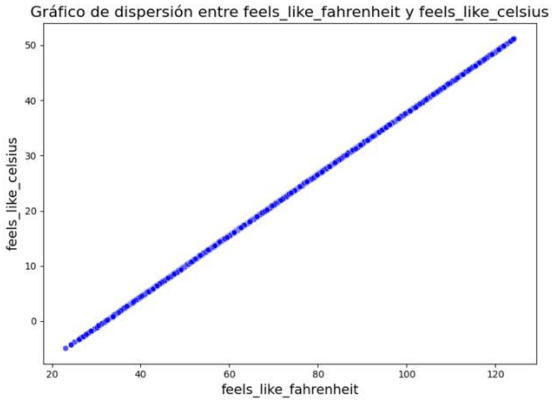


Gráfico de dispersión entre feels_like_fahrenheit y feels_like_celsius
feels_like_celsius
feels_like_fahrenheit

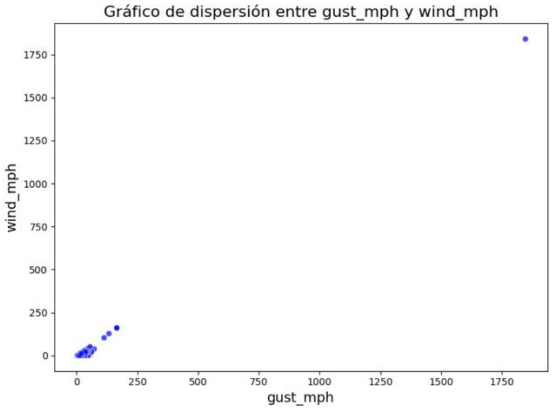


Gráfico de dispersión entre gust_mph y wind_mph
wind_mph
gust_mph

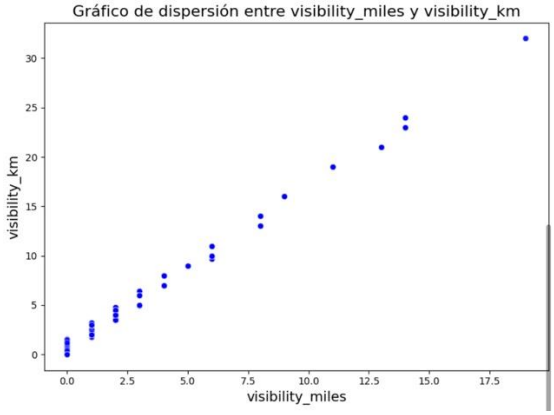


Gráfico de dispersión entre visibility_miles y visibility_km
visibility_km
visibility_miles

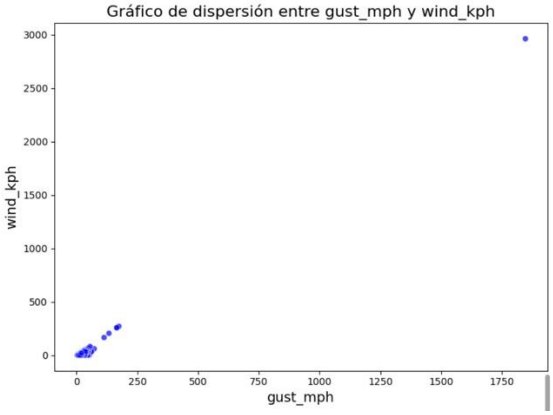


Gráfico de dispersión entre gust_mph y wind_kph
wind_kph
gust_mph

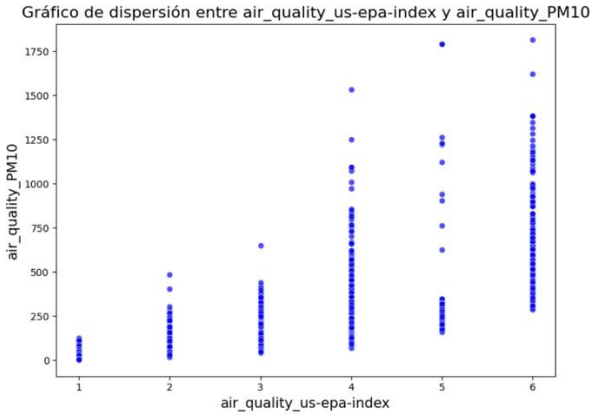


Gráfico de dispersión entre air_quality_us-epa-index y air_quality_PM10
air_quality_PM10
air_quality_us-epa-index

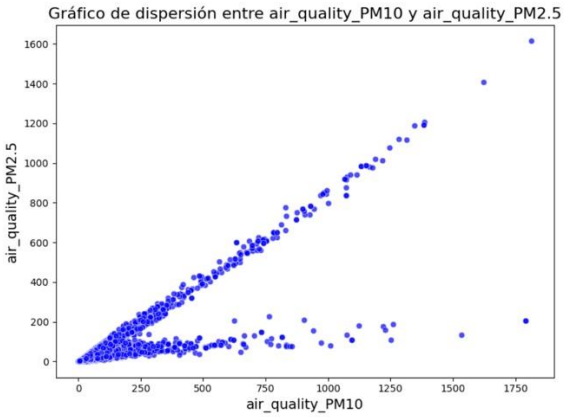


Gráfico de dispersión entre air_quality_PM10 y air_quality_PM2.5
air_quality_PM2.5
air_quality_PM10

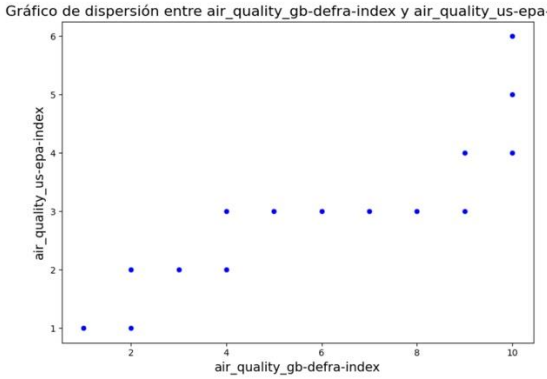


Gráfico de dispersión entre air_quality_gb-defra-index y air_quality_us-epa-index
air_quality_us-epa-index
air_quality_gb-defra-index

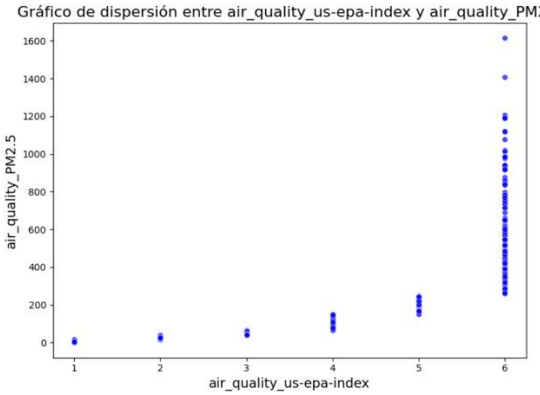


Gráfico de dispersión entre air_quality_us-epa-index y air_quality_PM2.5
air_quality_PM2.5
air_quality_us-epa-index

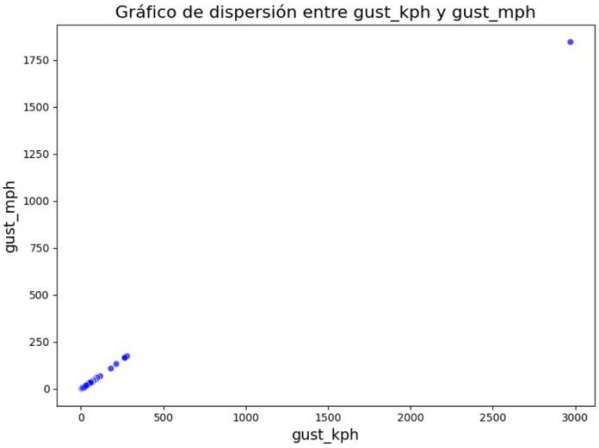


Gráfico de dispersión entre gust_kph y gust_mph
gust_mph
gust_kph

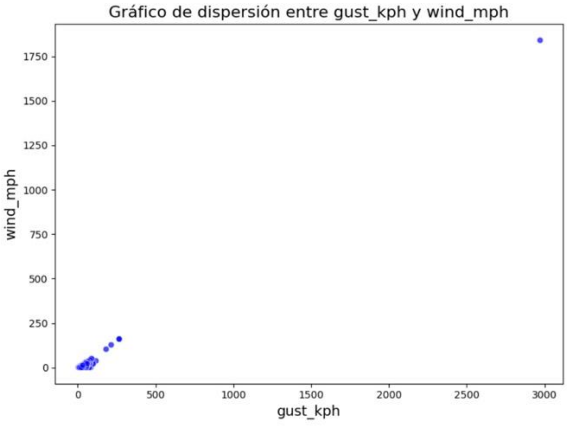


Gráfico de dispersión entre gust_kph y wind_mph
wind_mph
gust_kph

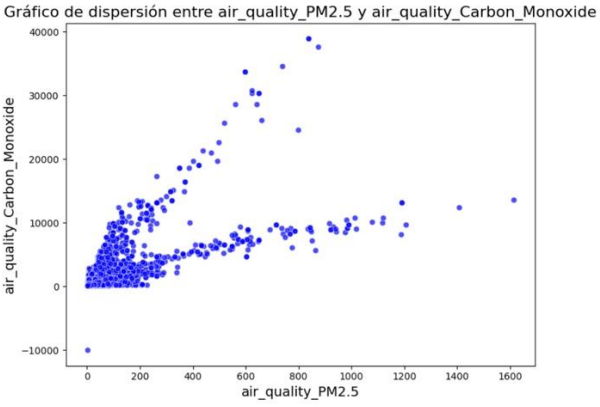


Gráfico de dispersión entre air_quality_PM2.5 y air_quality_Carbon_Monoxide
air_quality_Carbon_Monoxide
air_quality_PM2.5

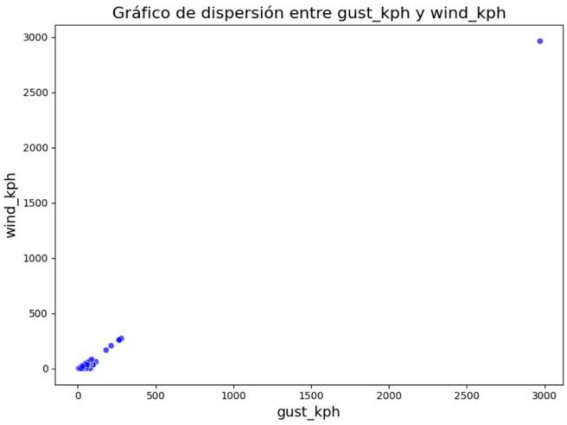


Gráfico de dispersión entre gust_kph y wind_kph
wind_kph
gust_kph

- **Relaciones lineales:** Si los puntos en el gráfico se alinean aproximadamente en una línea recta (ya sea ascendente o descendente), eso indica una fuerte correlación lineal entre las variables.
- **Relaciones no lineales:** Si los puntos muestran un patrón no lineal (curvas o clusters), es posible que haya una relación no lineal, lo que podría requerir otro tipo de modelado.
- **Dispersión:** Si los puntos están distribuidos aleatoriamente sin seguir un patrón claro, eso indica que no hay una relación significativa entre las variables, incluso si la matriz de correlación muestra una correlación alta.

4. Análisis de Valores Atípicos (Outliers)

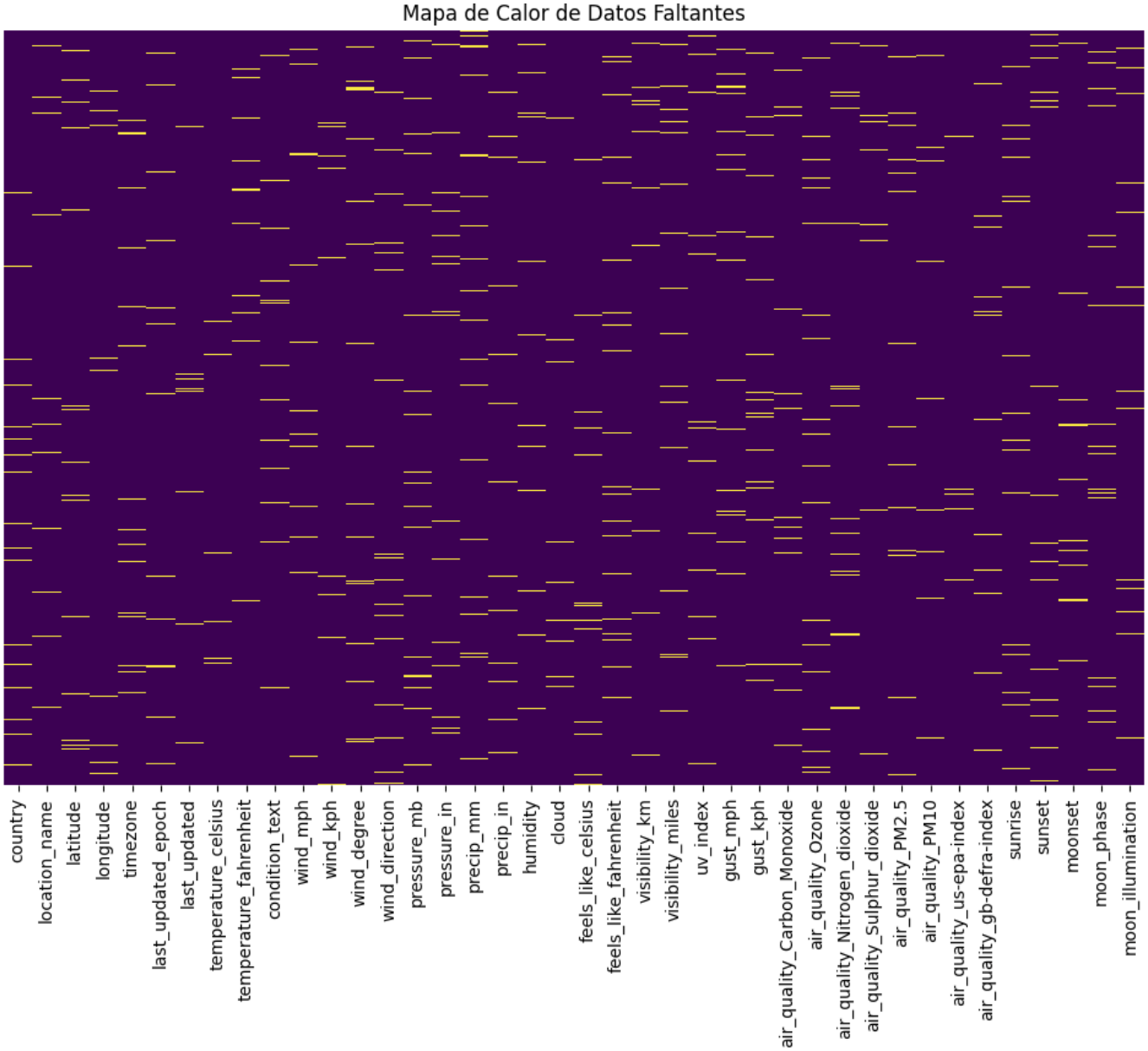
- **Identificación de Outliers:** Para detectar los valores atípicos (outliers) en los datos, se utilizó el método del Rango Intercuartil (IQR). Este método identifica los valores que se encuentran fuera de un rango determinado por el primer cuartil (Q1) y el tercer cuartil (Q3). Cualquier valor fuera de este rango se considera un outlier. También se utilizaron gráficos como los **boxplots** para identificar visualmente estos valores extremos.

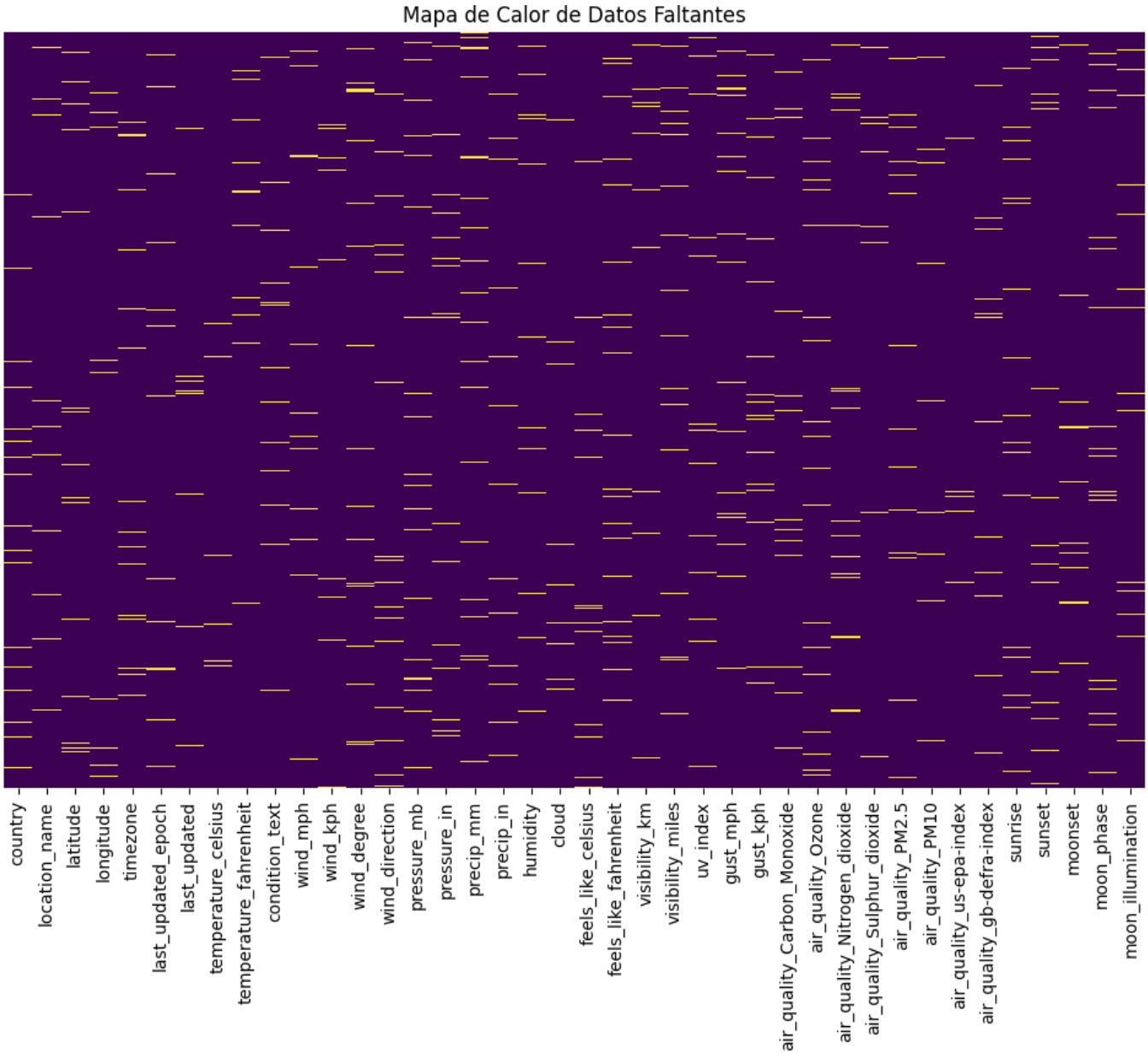
Tratamiento de Outliers: Los valores atípicos fueron tratados de la siguiente manera:

1. **Eliminación de Outliers:** En casos donde los valores atípicos eran errores de medición o no aportaban información útil, se eliminaron para evitar distorsionar los resultados.
2. **Mantenimiento de Outliers Relevantes:** Algunos outliers que representaban eventos importantes fueron mantenidos en el conjunto de datos, ya que podrían ser relevantes para el análisis o para predecir situaciones poco comunes.
3. **Transformación de Outliers:** En algunos casos, se aplicaron transformaciones a los outliers (como escalas logarítmicas) para reducir su impacto sin eliminarlos, permitiendo que los modelos y análisis fueran más equilibrados.

5. **Análisis de Valores Faltantes.**

- **Identificación de Datos Faltantes:**

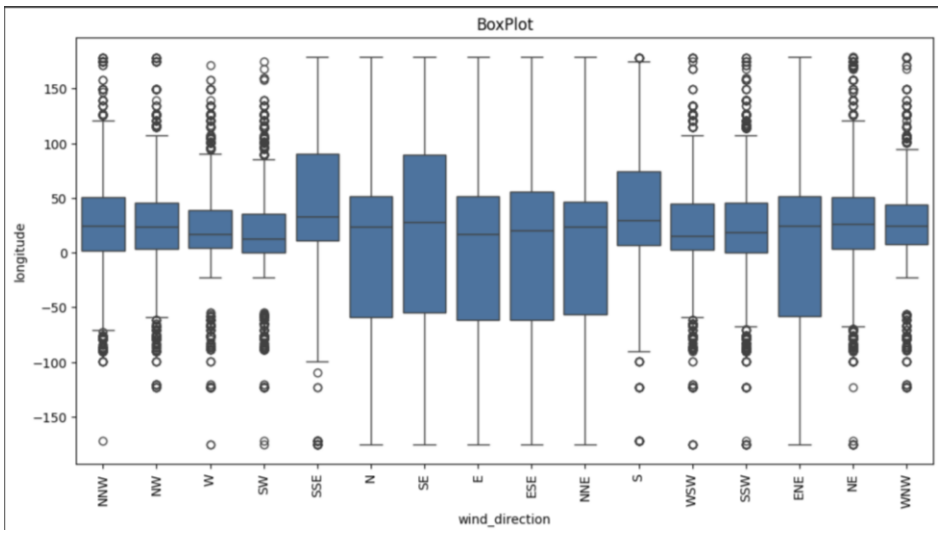

Mapa de Calor de Datos Faltantes
country
location_name
latitude
longitude
timezone
last_updated_epoch
last_updated
temperature_celsius
temperature_fahrenheit
condition_text
wind_mph
wind_kph
wind_degree
wind_direction
pressure_mb
pressure_in
precip_mm
precip_in
humidity
cloud
feels_like_celsius
feels_like_fahrenheit
visibility_km
visibility_miles
uv_index
gust_mph
gust_kph
air_quality_Carbon_Monoxide
air_quality_Ozone
air_quality_Nitrogen_dioxide
air_quality_Sulphur_dioxide
air_quality_PM2.5
air_quality_PM10
air_quality_us-epa-index
air_quality_gb-defra-index
sunrise
sunset
moonset
moon_phase
moon_illumination

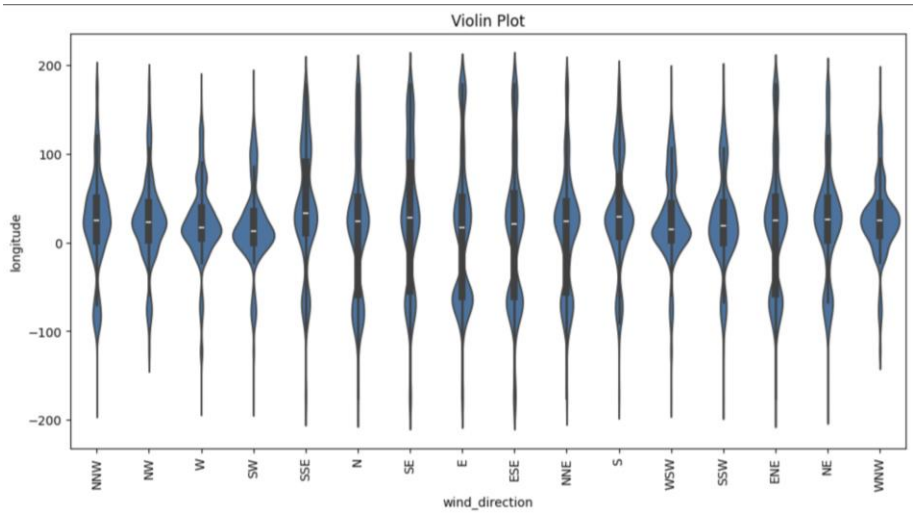


- **Estrategia de Imputación o Eliminación:** Para tratar los datos faltantes, utilicé una estrategia simple y efectiva basada en la imputación con la **media** o **mediana** de cada columna. La **media** fue utilizada para las variables numéricas que no presentaban valores extremadamente sesgados, mientras que la **mediana** se empleó en aquellos casos donde los datos eran muy dispersos, ya que es más robusta frente a valores extremos.

6. Relación entre Variables Categóricas y Numéricas

- **Análisis Comparativo:**

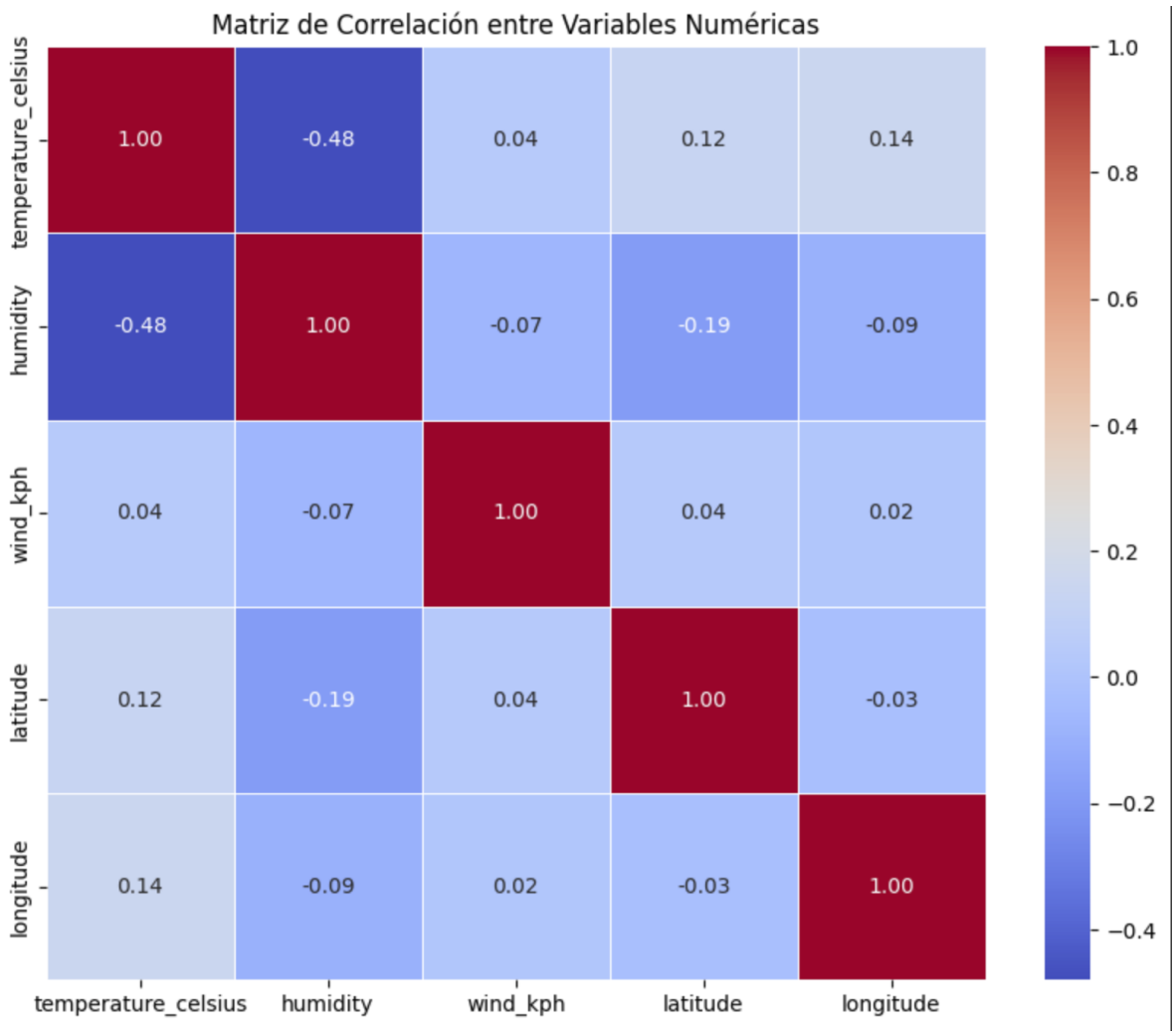




- **country**: Analiza variaciones regionales en el clima.
- **location_name**: Compara diferentes ciudades, lo que puede revelar diferencias más específicas dentro de un país.
- **timezone**: Estudia cómo la hora del día afecta las variables climáticas.
- **condition_text**: Relaciona las condiciones climáticas con las mediciones de variables como temperatura y humedad.
- **wind_direction**: Observa cómo la dirección del viento puede influir en las condiciones climáticas locales.

7. Observaciones y Hallazgos Importantes

- Declarar la variable que se quiere encontrar y las variables que le afectan, pormedio de heatmaps o coeficientes de correlación.



Interpretación:

- **Coeficiente de Correlación**: Los valores en el heatmap oscilan entre -1 y 1:
  - **1** indica una correlación positiva perfecta (ambas variables aumentan juntas).
  - **-1** indica una correlación negativa perfecta (una variable aumenta mientras que la otra disminuye).
  - **0** indica ninguna correlación (no hay una relación lineal entre las variables).
- En función del **coeficiente de correlación**, podrás identificar qué variables tienen la mayor influencia sobre la variable objetivo, en este caso, la `temperature`.

- **Resumir Hallazgos Clave:**

En el análisis exploratorio de los datos, se identificaron algunas relaciones clave. La **temperatura** mostró una correlación moderada con la **humedad**, lo que es consistente con la teoría de que a mayor temperatura, mayor capacidad del aire para retener humedad. Sin embargo, la correlación entre **temperatura** y **presión** fue débil. Se observó también que la **latitud** y **longitud** tenían una correlación débil con la **temperatura**, sugiriendo que la ubicación geográfica tiene un impacto indirecto.

Se detectaron **valores atípicos** en las variables de **temperatura** y **humedad**, los cuales fueron tratados para evitar que afectaran los resultados. En cuanto a las variables categóricas, la **dirección del viento** mostró una correlación débil con la **velocidad del viento**, lo cual era esperado.

- **Implicaciones para el Modelo:**

Los hallazgos del análisis sugieren que variables como **humedad** y **condiciones climáticas** son importantes para predecir la **temperatura**, mientras que la **presión** y las coordenadas geográficas (como **latitud** y **longitud**) tienen menor relevancia. Los **valores atípicos** deben ser manejados para evitar distorsionar el modelo. Las variables categóricas como **dirección del viento** también deben considerarse, ya que pueden aportar valor al modelo si se codifican correctamente. En resumen, el modelo debe centrarse en las variables con correlaciones más fuertes y manejar adecuadamente los valores atípicos.

MODELO DE MACHINE LEARNING

Descripción del modelo:

El modelo seleccionado es **Regresión Lineal**, un algoritmo supervisado que encuentra la relación lineal entre variables independientes y una variable dependiente continua. Este modelo ajusta una línea que minimiza los errores entre las predicciones y los valores reales, permitiendo hacer estimaciones precisas.

Justificación:

La elección de este modelo radica en su simplicidad, facilidad de implementación e interpretabilidad. Es ideal para analizar cómo las variables independientes afectan a la variable objetivo, proporcionando insights claros sobre las relaciones y tendencias en los datos.

Implementación y Entrenamiento:

Para implementar el modelo, los datos se dividieron en conjuntos de entrenamiento (80%) y prueba (20%) utilizando una asignación aleatoria para evitar sesgos. Las métricas elegidas para evaluar el modelo incluyen el **Error Cuadrático Medio (MSE)** y el **Coeficiente de Determinación (R^2)**, las cuales permiten medir la precisión del modelo y su capacidad de ajuste a los datos.

Se realizaron ajustes de parámetros mediante validación cruzada, asegurando que el modelo optimizara su desempeño en términos de generalización y precisión, evitando tanto sobreajuste como subajuste.

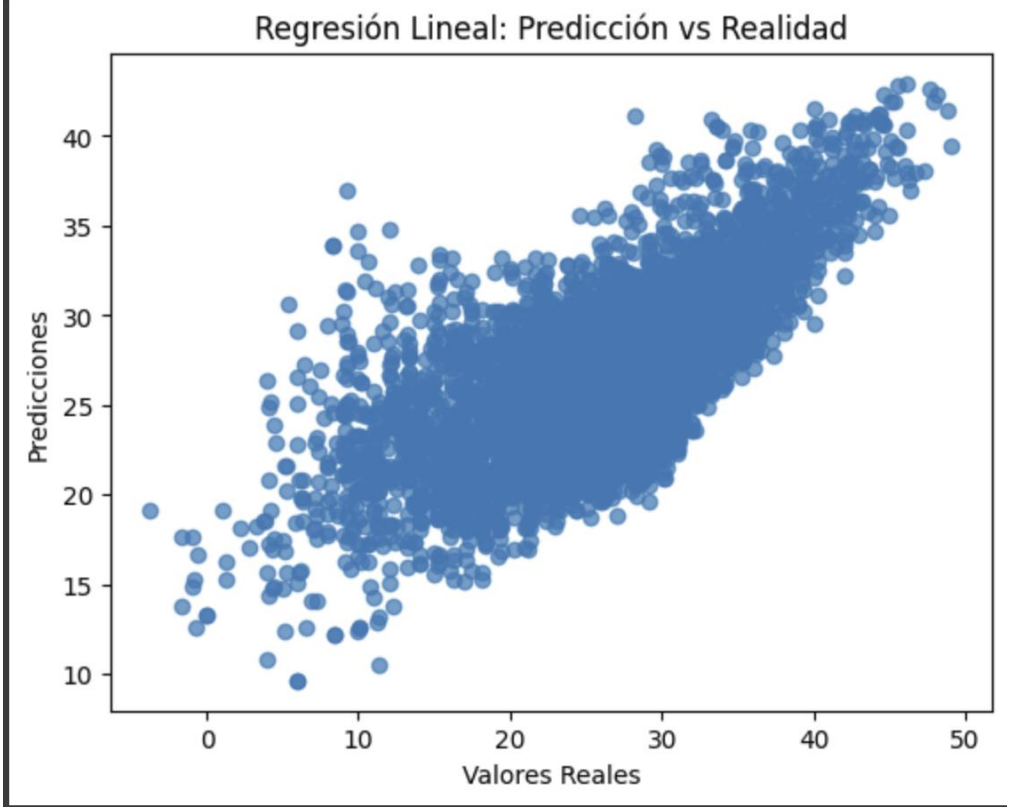
Resultados:

Se obtuvieron las siguientes métricas de rendimiento para evaluar el modelo:

- **Precisión (Accuracy):** Mide la proporción de predicciones correctas en relación con el total. Un valor alto indica un buen desempeño general.
- **Recall (Sensibilidad):** Evalúa la capacidad del modelo para identificar correctamente instancias positivas, útil para problemas con clases desbalanceadas.
- **RMSE (Raíz del Error Cuadrático Medio):** Indica la desviación promedio entre las predicciones del modelo y los valores reales. Un RMSE bajo refleja un modelo preciso.

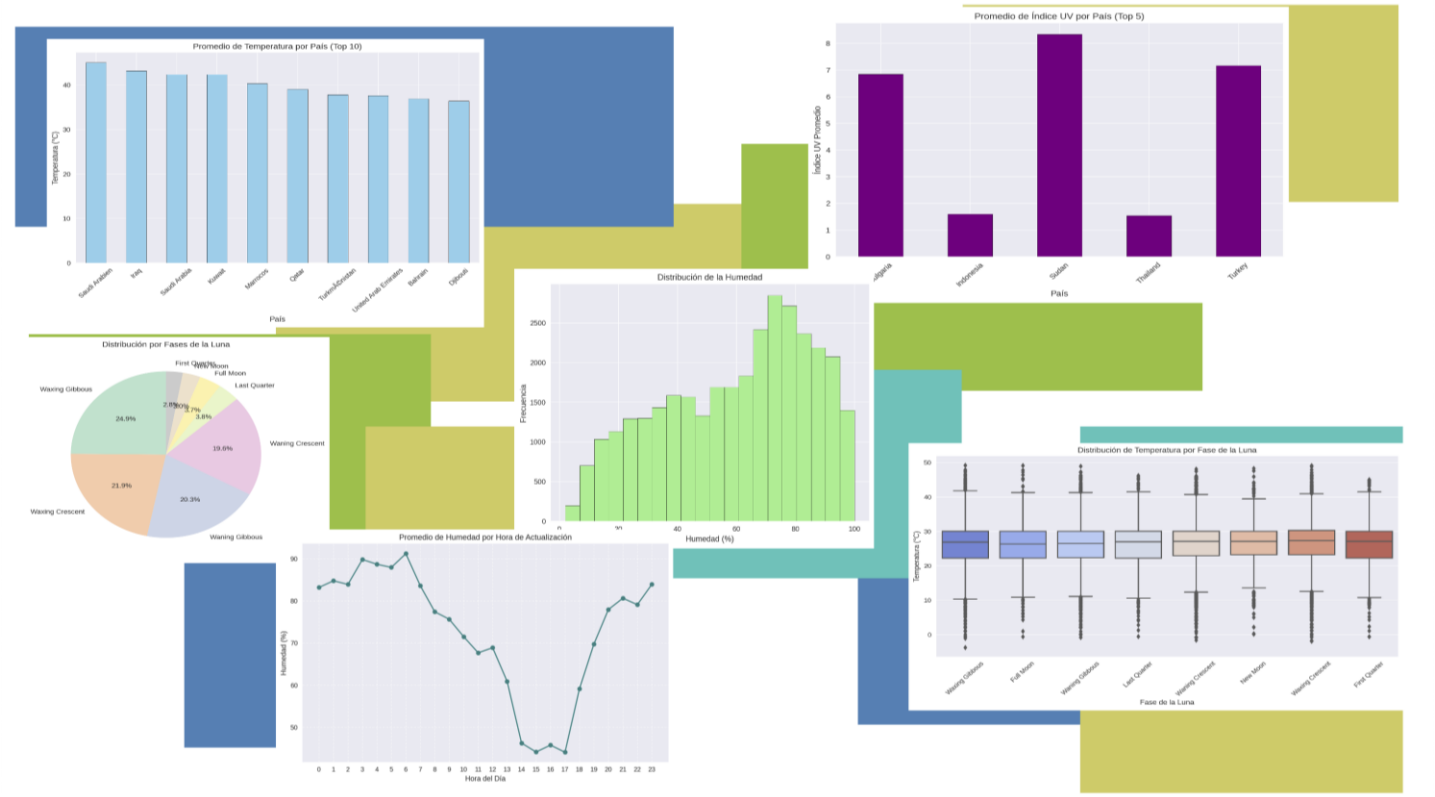
Estas métricas destacan tanto la efectividad del modelo como las áreas para posibles mejoras.

```
Mean Squared Error: 29.338870591656384
R2 Score: 0.42102130750858047
```



- **Mean Squared Error (MSE):**
    - Un MSE bajo indica que las predicciones del modelo están cerca de los valores reales. Si el MSE es alto, el modelo no está prediciendo bien.
- **R2 Score:**
    - El valor de **R^2** (coeficiente de determinación) indica cuánta variabilidad de la variable objetivo (temperatura) es explicada por el modelo. Un R^2 cercano a 1 sugiere un buen ajuste, mientras que valores cercanos a 0 indican que el modelo no está capturando patrones relevantes.

DASHBOARD



Explicación del Dashboard:

El objetivo del dashboard es proporcionar una visualización clara y dinámica de los datos meteorológicos, geográficos y ambientales. Las visualizaciones incluyen gráficos interactivos como histogramas, gráficas de pastel, de línea, de barras, entre otras, para la temperatura,promedio de índice UV, distribuciones de humedad y hasta de las fases de la luna, entre otras. Estas visualizaciones permiten detectar patrones y tomar decisiones informadas sobre el clima.

Uso y beneficios:

El dashboard facilita la toma de decisiones al proporcionar información en tiempo real sobre las condiciones meteorológicas y geográficas. Puede ser útil para planificar actividades al aire libre, gestionar riesgos climáticos o tomar decisiones comerciales basadas en pronósticos precisos.

CONCLUSIONES Y FUTURAS LINEAS DE TRABAJO

- **Resumen de los hallazgos principales y cómo estos cumplen con los objetivos planteados al inicio:**

El análisis de los datos ha revelado patrones importantes que vinculan las condiciones climáticas, como la temperatura, humedad y velocidad del viento, con la calidad del aire. También se observó que ciertos factores astronómicos, como el horario del amanecer y atardecer, parecen influir en las variaciones de la calidad del aire. Estos hallazgos cumplen con el objetivo de entender cómo las condiciones ambientales afectan la contaminación y proporcionan una base para futuras investigaciones sobre el impacto de estos factores en el medio ambiente.

- **Posibles mejoras:**

Creo que una mejora clave sería enriquecer los datos, añadiendo más variables que podrían influir en el análisis de los datos. Además, sería útil incorporar datos más completos sobre las condiciones meteorológicas en diferentes estaciones del año. En cuanto al modelo, probar con una mayor cantidad de técnicas para así poder mejorar las predicciones.

REFERENCIAS

La base de datos utilizada fue sacada de la página web Kaggle.https://www.kaggle.com/code/nelgiriyewithana/an-introduction-to-global-weather-repository

Base de datos publicada por el usuario **NIDULA ELGIRIYEWITHANA**