Análisis Exploratorio de Calidad del Aire en EE.UU.

1. Introducción

Este proyecto tiene como objetivo realizar un análisis exploratorio de un conjunto de datos sobre la calidad del aire en Phoenix, Arizona (EE.UU.). Se utilizan mediciones de contaminantes como NO_2, O_3, SO_2 y CO, junto con sus respectivos índices de calidad del aire (AQI). Este análisis nos permite comprender el comportamiento de estos contaminantes, identificar relaciones y proponer recomendaciones para futuros estudios.

2. Conjunto de datos

El conjunto de datos fue obtenido de Kaggle y contiene 1000 registros. Incluye columnas que representan:
- Concentraciones promedio y máximas de NO_2, O_3, SO_2 y CO.
- Índices de calidad del aire (AQI) para cada contaminante.
- Información temporal y geográfica (fecha, ciudad, estado).

Todas las mediciones provienen de Phoenix, Arizona.

3. Limpieza de datos

Durante la limpieza se eliminaron las columnas 'State', 'City' y 'Address' al presentar valores constantes en todas las filas, lo cual no aporta variabilidad para el análisis.
También se eliminaron las filas que contenían valores nulos en las columnas 'SO2 AQI' y 'CO AQI'.
Finalmente, se convirtió la columna 'Date Local' al tipo de dato fecha (`datetime`) para facilitar el análisis temporal.

Se encontraron columnas con valores cero frecuentes, como:
- SO2 AQI: 11.6%
- SO2 1st Max Hour: 16.5%
- CO 1st Max Hour: 11.2%

Estos ceros no fueron tratados como valores faltantes, ya que probablemente indican días sin contaminación medible.

4. Análisis Exploratorio de Datos (EDA)

Se utilizó la librería ydata-profiling para generar un informe completo del conjunto de datos. Entre los hallazgos más relevantes se encuentran:
- CO AQI tiene una distribución sesgada a la derecha, con valores concentrados entre 10 y

20.
- SO2 AQI también presenta distribución sesgada, con un pico en valores bajos (entre 0 y 10).
- Las variables AQI están altamente correlacionadas con sus valores máximos y promedios respectivos.
- Algunas variables como 'Date Local' y 'Unnamed: 0' tienen valores únicos por fila, y pueden usarse como identificadores.

5. Explicación de alertas del EDA

Las siguientes alertas fueron detectadas automáticamente por ydata-profiling:

◈ Variables constantes: 'State Code', 'County Code', etc., que no fueron eliminadas por no haber sido mencionadas explícitamente.
◈ Altas correlaciones: por ejemplo, 'CO AQI', 'CO 1st Max Value', y 'CO Mean' están fuertemente correlacionadas. Esto es esperable, ya que el AQI se deriva directamente de las mediciones del contaminante.
◈ Columnas con ceros: variables como 'SO2 AQI' y 'SO2 1st Max Hour' tienen una alta proporción de ceros, lo cual debe interpretarse cuidadosamente.
◈ Columnas únicas o uniformes: 'Date Local' y 'Unnamed: 0' tienen valores únicos, lo que indica que cada fila corresponde a un día diferente de medición.

6. Visualizaciones

Se utilizaron gráficos de histograma y pairplot para explorar la distribución y relaciones entre variables. Algunos hallazgos visuales:

- CO AQI: valores distribuidos mayormente entre 10 y 20, con muy pocos días superando 40.
- SO2 AQI: gran cantidad de días con valor 0, lo cual es visible en el histograma.
- Pairplot (Seaborn): evidencia visual de correlaciones altas entre promedios, máximos y AQI.

Nota: Las visualizaciones interactivas completas se encuentran en el archivo 'EDA_Contaminacion_USA.html'.

7. Conclusiones y próximos pasos

Este análisis permitió identificar patrones de contaminación en Phoenix y comprender la estructura del dataset. Se evidenció alta redundancia entre variables, lo cual puede aprovecharse para reducir dimensionalidad en modelos predictivos. Aunque no hubo valores nulos luego de la limpieza, la presencia de ceros debe considerarse al hacer inferencias.

Próximos pasos:
- Analizar múltiples ciudades para obtener resultados más generalizables.
- Clasificar días según niveles AQI (bueno, moderado, malo).
- Explorar modelado predictivo con aprendizaje automático.