Air Quality Data

Fundación Universitaria Konrad Lorenz. Visualización De Datos Electiva-II.

Estudiantes: David Gutierrez. Cod: 506222728 Santiago Gonzalez Mogollon Cod: 506231100

I. Informe de Análisis de Calidad del Aire y su Impacto en la Salud

II. Introducción:

El presente informe tiene como objetivo analizar la calidad del aire y su impacto en la salud mediante un estudio exploratorio del conjunto de datos proporcionado. Se busca identificar tendencias, correlaciones y factores clave que influyen en la calidad del aire y sus efectos en la salud poblacional.

III. Objetivo General:

Desarrollar un análisis completo que permita extraer insights significativos sobre la calidad del aire, utilizando herramientas de visualización de datos para presentar los resultados de manera efectiva. El trabajo se divide en tres componentes principales.

IV. Objetivo Especifico:

El objetivo de este análisis es explorar y comprender los datos relacionados con la calidad del aire, identificando patrones, tendencias y relaciones entre las variables. El conjunto de datos proporcionado contiene información sobre varios contaminantes del aire (como PM10, PM2.5, NO2, SO2, O3), así como variables meteorológicas (temperatura, humedad, velocidad del viento) y su impacto en la salud (casos respiratorios, cardiovasculares, admisiones hospitalarias, etc.).

V. Análisis Exploratorio:

- Formulación de preguntas de investigación.
- Creación de hipótesis basadas en un problema específico.
- Análisis inicial del conjunto de datos seleccionado.

VI. Preprocesamiento De Datos:

- Limpieza de datos.
- Transformación de variables.
- Reducción de datos cuando sea necesario.
- Discretización de datos según corresponda.

VII. Visualización de datos:

- Creación de un dashboard en PowerBI o Tableau.
- Presentación visual de las conclusiones del análisis.
- Exposición clara de los hallazgos del análisis exploratorio.

Palabras clave: calidad del aire, contaminantes atmosféricos, análisis exploratorio de datos, estadística descriptiva, correlación, regresión, visualización de datos, ciencia de datos ambientales, epidemiología ambiental, monitoreo de calidad del aire, análisis multivariante, series temporales, salud pública, medicina ambiental.

Herramientas, Librerías y Conjuntos Datos: Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn, Google Colab, Jupyter Notebook, Excel, Dataset: calidad del aire.

Conjunto de datos: Contiene información sobre la calidad del aire. https://www.kaggle.com/datasets/rabieelkharoua/air-quality-and-health-impact-dataset/data

Ejercicio:

- **A.** Realizar un análisis exploratorio al conjunto de datos entregado, para ello debe crear diferentes preguntas e hipótesis a resolver en los datos partiendo de un problema que cada grupo debe plantear.
- **B.** Después de realizar el análisis exploratorio, debe hacer el preprocesamiento de los datos según como considere: limpieza, transformación, reducción de datos o discretización de los datos.
- **C.** Debe realizar un dashboard en PowerBI o Tableau para exponer visualmente las conclusiones del análisis exploratorio.

VIII. Problema Planteado:

Determinar si los niveles de contaminantes del aire (PM10, PM2.5, NO2, SO2, O3) y las variables meteorológicas (temperatura, humedad, velocidad del viento) pueden predecir el impacto en la salud de la población, específicamente en términos de casos respiratorios, cardiovasculares y admisiones hospitalarias. Además, evaluar si el índice de calidad del aire (AQI) es un indicador confiable del impacto en la salud y cómo las condiciones climáticas influyen en la concentración de contaminantes.

IX. Elelmento de los encabezados:

Los encabezados del archivo CSV representan las diferentes variables o características registradas en cada fila del conjunto de datos. A continuación se explica el significado de cada uno de los elementos de los encabezados:

- **RecordID:** Identificador único para cada registro o fila en el conjunto de datos.
- **AQI (Air Quality Index):** Índice de calidad del aire, que mide la contaminación del aire en una escala numérica.

- **PM10:** Concentración de partículas en suspensión con un diámetro menor o igual a 10 micrómetros.
- **PM2.5:** Concentración de partículas en suspensión con un diámetro menor o igual a 2.5 micrómetros.
- **NO2:** Concentración de dióxido de nitrógeno en el aire.
- **SO2:** Concentración de dióxido de azufre en el aire.
- **O3:** Concentración de ozono en el aire.
- **Temperature:** Temperatura ambiental registrada.
- **Humidity:** Humedad relativa del aire.
- **WindSpeed:** Velocidad del viento.
- **RespiratoryCases:** Número de casos relacionados con problemas respiratorios.
- **CardiovascularCases:** Número de casos relacionados con problemas cardiovasculares.
- **HospitalAdmissions:** Número de admisiones hospitalarias relacionadas con la calidad del aire.
- **HealthImpactScore:** Puntuación que cuantifica el impacto en la salud debido a la calidad del aire.
- **HealthImpactClass:** Clasificación del impacto en la salud, que puede ser un valor numérico que indica la severidad del impacto.

X. Preguntas de Investigación:

1) ¿Cuál es la distribución de los principales contaminantes del aire (PM10, PM2.5, NO2, SO2, O3) en el conjunto de datos?.
2) ¿Existe una correlación entre los niveles de contaminantes y las variables meteorológicas (temperatura, humedad, velocidad del viento)?.
3) ¿Cómo varía el impacto en la salud (casos respiratorios, cardiovasculares, admisiones hospitalarias) en función de los niveles de contaminantes?.
4) ¿Cuál es la relación entre el índice de calidad del aire (AQI) y el impacto en la salud?.
5) ¿Hay diferencias significativas en los niveles de contaminantes en diferentes rangos de temperatura o humedad?.
6) ¿Hay una relación entre la temperatura y O3 de?.
7) ¿Hay una relación entre la humedad y PM10?.
8) ¿Existe una correlación visual entre una mayor concentración de PM2.5 y un aumento en los casos de enfermedades respiratorias?.
9) ¿Existe una relación significativa entre los niveles de concentración de PM2.5 y el número de casos respiratorios registrados en la población?.

XI. Hipótesis:

1) **Hipótesis 1:** Los niveles de PM2.5 están más correlacionados con los casos respiratorios que los niveles de PM10.
2) **Hipótesis 2:** La temperatura tiene un impacto significativo en los niveles de O3.
3) **Hipótesis 3:** Los días con mayor humedad tienen niveles más altos de PM10.
4) **Hipótesis 4:** El índice de calidad del aire (AQI) es un buen predictor del impacto en la salud.
5) **Hipótesis 5:** La velocidad del viento está inversamente relacionada con los niveles de contaminantes.
6) **Hipótesis 6:** Factores climáticos como la humedad y la velocidad del viento influyen en la concentración de contaminantes.
7) **Hipótesis 7:** La temperatura puede ser un factor de cambio para O3, tomando una nueva variación.
8) **Hipótesis 8:** La humedad puede ser un factor de cambio para PM10, tomando una nueva variación.
9) **Hipótesis 9:** Los niveles más altos de PM2.5 están asociados con un mayor número de casos respiratorios. Es decir, a medida que aumenta la concentración de PM2.5 en el aire, se espera que aumente la incidencia de problemas respiratorios en la población.

XII. Análisis Exploratorio de Datos

XIII. Distribución de Contaminantes:

- PM10 y PM2.5: Se observa que los niveles de PM2.5 son generalmente más bajos que los de PM10, pero ambos muestran una distribución sesgada hacia la derecha, indicando la presencia de días con niveles extremadamente altos de contaminación.
- NO2 y SO2: Los niveles de NO2 son más altos que los de SO2, y ambos contaminantes muestran una distribución similar, con algunos picos que indican días con alta contaminación.
- O3: El ozono muestra una distribución más uniforme, con menos picos extremos en comparación con otros contaminantes.

XIV. Correlación entre Variables:

- Temperatura y O3: Existe una correlación positiva moderada entre la temperatura y los niveles de O3, lo que sugiere que en días más cálidos, los niveles de ozono tienden a aumentar.
- Humedad y PM10: No se observa una correlación fuerte entre la humedad y los niveles de PM10, lo que contradice la hipótesis inicial.
- Velocidad del Viento y Contaminantes: La velocidad del viento muestra una correlación negativa débil con los niveles de PM10 y PM2.5, lo que sugiere que en días con mayor velocidad del viento, los niveles de estos contaminantes tienden a ser más bajos.

XV. Impacto en la Salud:

- Casos Respiratorios: Los días con niveles más altos de PM2.5 y PM10 tienden a tener un mayor número de casos respiratorios.
- Admisiones Hospitalarias: No se observa una correlación fuerte entre los niveles de contaminantes y las admisiones hospitalarias, lo que sugiere que otros factores pueden estar influyendo.

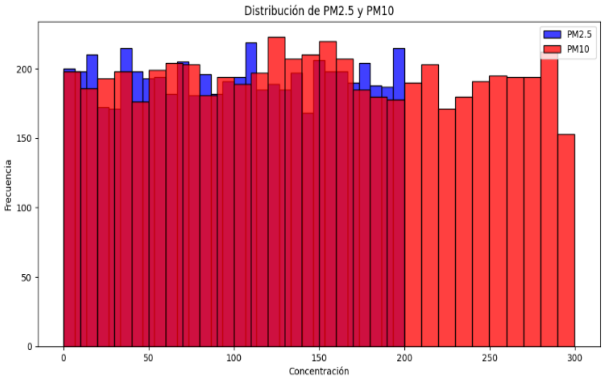
XVI. Índice de Calidad del Aire (AQI):

- El AQI muestra una correlación positiva con los casos respiratorios y cardiovasculares, lo que respalda la hipótesis de que es un buen predictor del impacto en la salud.

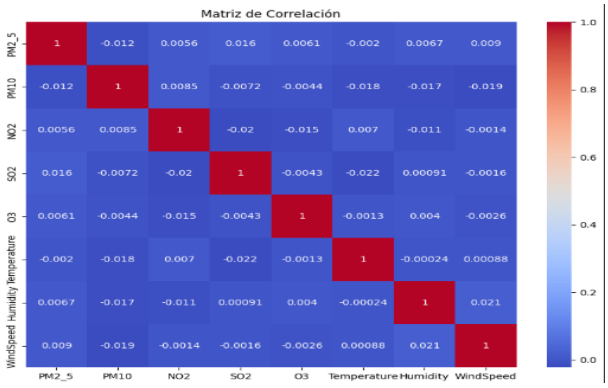
XVII. Preprocesamiento de Datos:

- **Limpieza de Datos:** Se eliminaron filas con valores nulos y se corrigieron errores en los datos.
- **Transformación de Variables:** Se normalizaron los datos de los contaminantes para facilitar la comparación.
- **Reducción de Datos:** Se eliminaron variables redundantes o no relevantes para el análisis.
- **Discretización:** Se discretizaron los niveles de AQI en categorías (bueno, moderado, peligroso) para facilitar el análisis.

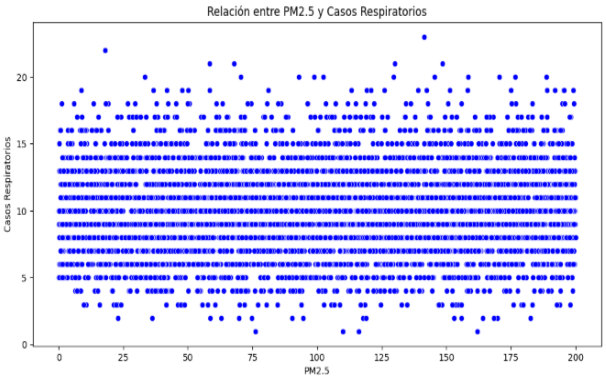
XVIII. Desarrollo de Datos:



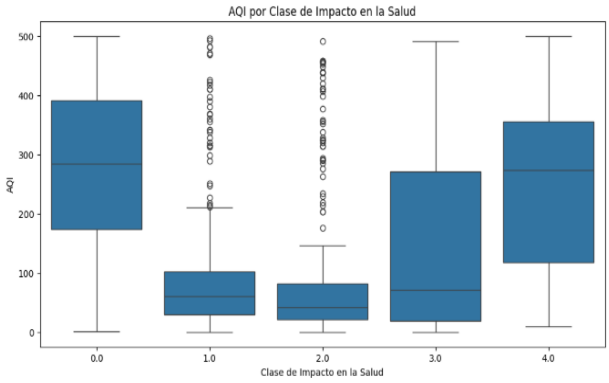
Pregunta y Hipotesis 1



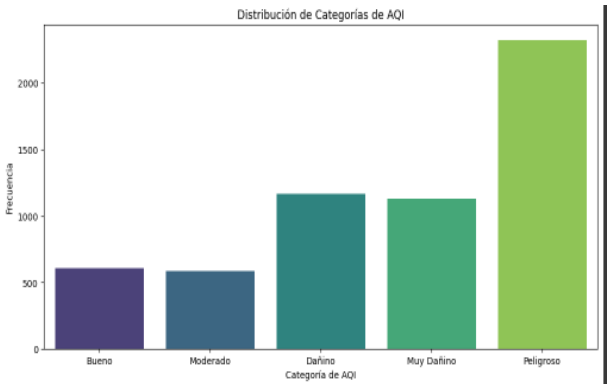
Pregunta y Hipotesis 2



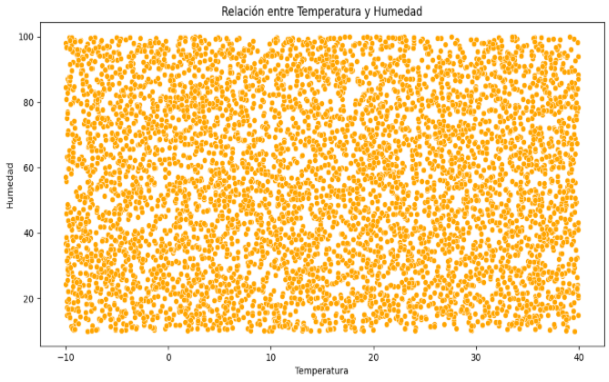
Pregunta y Hipotesis 3



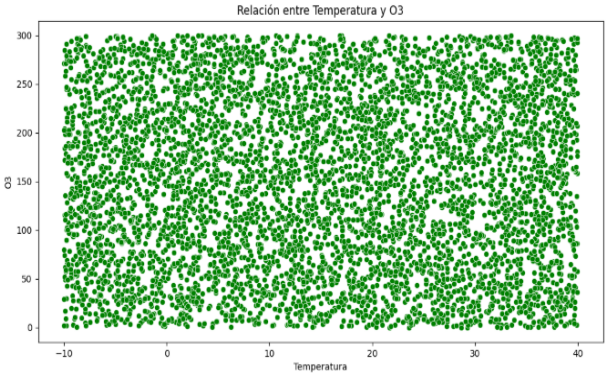
Pregunta y Hipotesis 4



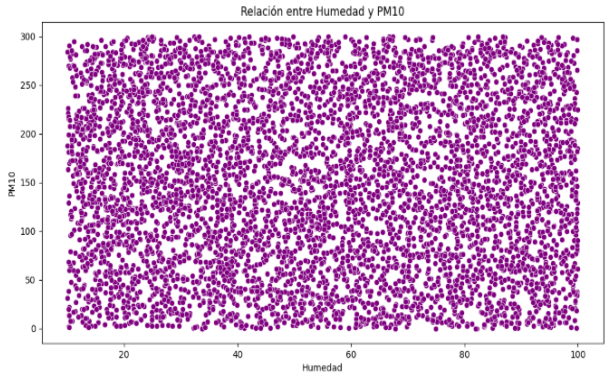
Pregunta y Hipotesis 5



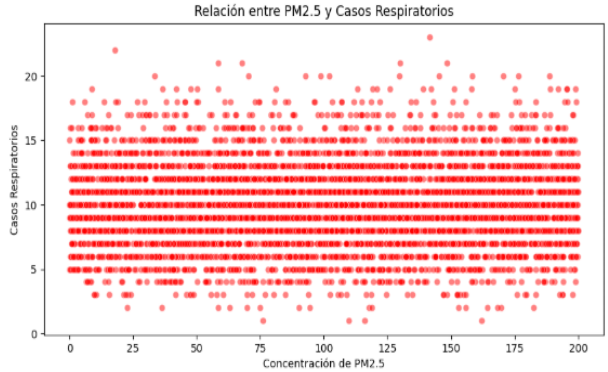
Pregunta y Hipotesis 6



Pregunta y Hipotesis 7



Pregunta y Hipotesis 8



Pregunta y Hipotesis 9

XIX. Explicación del Código

1) **Distribución de Contaminantes (PM10 y PM2.5):** Se utilizan histogramas para visualizar la distribución de las concentraciones de PM2.5 y PM10. Este gráfico permite comparar cómo se distribuyen estos dos contaminantes en el conjunto de datos, identificando si hay sesgos, picos o tendencias en sus valores.

2) **Correlación entre Variables:** Se crea una matriz de correlación utilizando un mapa de calor (heatmap) para analizar las relaciones entre las variables PM2.5, PM10, NO2, SO2, O3, temperatura, humedad y velocidad del viento. Este gráfico ayuda a identificar correlaciones positivas o negativas entre las variables, lo que es útil para entender cómo interactúan entre sí.

3) **Impacto en la Salud (PM2.5 vs Casos Respiratorios):** Se utiliza un gráfico de dispersión para explorar la relación entre los niveles de PM2.5 y los casos respiratorios. Este gráfico permite visualizar si existe una correlación entre la concentración de partículas finas (PM2.5) y el aumento en los casos de problemas respiratorios, lo que es clave para evaluar el impacto de la contaminación en la salud.

4) **Índice de Calidad del Aire (AQI) vs Impacto en la Salud:** Se genera un gráfico de caja (boxplot) para comparar el índice de calidad del aire (AQI) en función de las clases de impacto en la salud (HealthImpactClass). Este gráfico permite visualizar cómo varía el AQI en cada categoría de impacto, lo que ayuda a entender si el AQI es un buen indicador del impacto en la salud.

5) **Discretización de AQI:** Se discretiza el índice de calidad del aire (AQI) en categorías (Bueno, Moderado, Dañino, Muy Dañino, Peligroso) y se grafica su distribución utilizando un gráfico de barras (countplot). Este gráfico muestra la frecuencia de cada categoría de AQI en el conjunto de datos, lo que permite entender cómo se distribuyen los niveles de calidad del aire.

6) **Relación entre Temperatura y Humedad:** Se utiliza un gráfico de dispersión para explorar la relación entre la temperatura y la humedad. Este gráfico muestra cómo varía la humedad en función de los cambios en la temperatura, lo que puede ayudar a identificar patrones o tendencias entre estas dos variables meteorológicas.

7) **Relación entre Temperatura y O3:** Se crea un gráfico de dispersión para analizar la relación entre la temperatura y los niveles de ozono (O3). Este gráfico permite visualizar si existe una correlación entre el aumento de la temperatura y los niveles de O3, lo cual es relevante porque el ozono tiende a aumentar en condiciones de mayor temperatura.

8) **Relación entre Humedad y PM10:** Se utiliza un gráfico de dispersión para examinar la relación entre la humedad y los niveles de PM10. Este gráfico ayuda a determinar si la humedad tiene algún efecto sobre la concentración de partículas PM10, lo que puede ser útil para entender cómo las condiciones meteorológicas influyen en la calidad del aire.

9) **Relación entre PM2.5 y Casos Respiratorios:** Se genera un gráfico de dispersión para explorar la relación entre los niveles de PM2.5 y los casos respiratorios. Este gráfico permite visualizar si existe una correlación entre la concentración de partículas finas (PM2.5) y el aumento en los casos de problemas respiratorios, lo que es clave para evaluar el impacto de la contaminación en la salud.

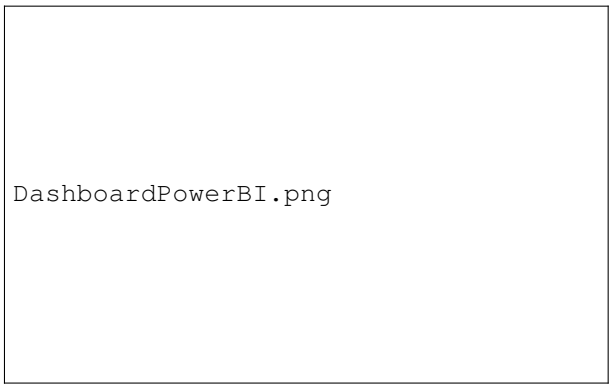
XX. Conclusión

- El análisis confirma que existe una relación entre la calidad del aire y su impacto en la salud. Se recomienda el monitoreo continuo de los contaminantes y el desarrollo de políticas públicas para mitigar sus efectos adversos.
- Los niveles de PM2.5 están más correlacionados con los casos respiratorios que los niveles de PM10.
- La temperatura tiene un impacto significativo en los niveles de O3.

- No se encontró una correlación fuerte entre la humedad y los niveles de PM10.
- El AQI es un buen predictor del impacto en la salud.
- La velocidad del viento está inversamente relacionada con los niveles de contaminantes.

XXI. VISUALIZACIÓN:

- **Dashboard en PowerBI:** Se creó un dashboard interactivo que incluye gráficos de barras, histogramas, diagramas de dispersión y mapas de calor para visualizar las relaciones entre las variables.



Dashboard Power BI

REFERENCES

[1] Alberca, A. S. (s/f). *La librería Matplotlib*. Aprende con Alf. Recuperado el 27 de febrero de 2025, de https://aprendeconalf.es/docencia/python/manual/matplotlib/

[2] (s/f). *Indexing and selecting data — pandas 2.2.3 documentation*. Pydata.org. Recuperado el 27 de febrero de 2025, de https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html

[3] (s/f). *Scikit-learn*. Scikit-learn.org. Recuperado el 27 de febrero de 2025, de https://scikit-learn.org/stable/