UNIVERSIDAD AMERICANA
Facultad de Ingeniería y Arquitectura

Inteligencia de Negocios

Informe de Investigación - Asignación U3T2 - Exploración y análisis de un conjunto de datos

---

Estudiante:
Gabriel David Chang Pérez

Docente:
Arlen Jeannette Lopez

Índice

Abstract

Este informe presenta un análisis exploratorio de las emisiones de dióxido de carbono (CO2) a nivel global, utilizando datos de diferentes países y variables como la densidad poblacional, el área geográfica y la población estimada en 2022. El objetivo del análisis fue identificar patrones y tendencias en las emisiones de CO2, así como su relación con otras variables relevantes. Para lograr esto, se utilizaron herramientas de análisis de datos en Python, como Pandas y Seaborn, junto con algoritmos de clustering como K-means para segmentar los países según sus características compartidas.

Los resultados encontrados en base al desarrollo del análisis del conjunto de datos con Python, mostraron que los países más industrializados, a como lo son Estados Unidos, y el Reino Unido, han sido los principales emisores de CO2 y contribuidores al cambio climático históricos en los pasados siglos. Igualmente, se destacó distintas correlaciones, algunas esperadas, y otras no tanto, en base al impacto de la dependencia y relación entre las diferentes variables del conjunto de datos, como lo es la inesperada baja correlación directa entre la densidad poblacional y las emisiones de cada país. Dichos resultados sugieren que el nivel de industrialización y las políticas energéticas tienen un impacto mayor. El análisis temporal reveló una aceleración en las emisiones desde mediados del siglo XX, reflejando el auge industrial global.

En conclusión, el análisis subraya la responsabilidad de los principales emisores en la implementación de políticas más estrictas para mitigar el cambio climático, mientras que los países con menores emisiones deben mantener políticas de sostenibilidad. Este trabajo refuerza la importancia de adoptar medidas más efectivas y coordinadas a nivel internacional para reducir las emisiones de CO2.

Introducción

El cambio climático y sus consecuencias globales han puesto de relieve la importancia de estudiar las emisiones de dióxido de carbono (CO_2), uno de los principales gases de efecto invernadero. La realización de distintos análisis y estudios en base a las emisiones de CO_2 y su relación con el cambio climático, ha tenido significativos aportes en cuanto a cómo los países y organizaciones alrededor del mundo implementan técnicas y soluciones para monitorear y reducir su huella de carbono, así concretando su contribución en diferentes aspectos y métricas.

Este trabajo tiene como objetivo principal analizar el dataset en cuestión respecto a las emisiones de CO_2 de diferentes países. Pretendo poder identificar patrones, tendencias, anomalías, y otras consideraciones relevantes en dicho conjunto de datos. Mediante la aplicación de técnicas de análisis y visualización de datos, buscaré obtener una comprensión más profunda de cómo las emisiones de carbono varían en función de las características demográficas y geográficas de los países.

Este análisis promueve su relevancia basado en la capacidad de poder ofrecer información clave a ser utilizada en la toma de decisiones a nivel internacional por gobiernos y organizaciones. Al identificar qué países son los principales emisores de CO_2 y cómo sus características influyen en estas emisiones, es posible desarrollar políticas más efectivas para mitigar el impacto del cambio climático.

Revisión Literaria

Para los objetivos planteados del trabajo en cuestión, se utilizarán diferentes técnicas, investigaciones, y modelos basados en el análisis y procesamiento de datos dentro del área de la inteligencia de negocios, específicamente mediante el uso de Python.

El análisis exploratorio de datos (EDA) es una etapa crítica en cualquier investigación basada en datos, permitiendo identificar patrones, detectar anomalías y formular hipótesis iniciales. Kashnitsky (2021) destaca la importancia de utilizar herramientas como Pandas en Python para manipular y analizar grandes conjuntos de datos de manera eficiente. Mediante técnicas aplicadas para visualizar y resumir estadísticas de los datos, podemos descubrir relaciones ocultas e implícitas entre las variables, así brindando mejores resultados para facilitar el desarrollo de modelos predictivos. En el contexto de este estudio, el EDA ha sido clave para comprender las tendencias de las emisiones de CO2 y su relación con otras variables, como la densidad poblacional y la superficie territorial.

Selvaraj (2023) refuerza este enfoque, subrayando cómo el análisis de datos en Python, en particular para principiantes, es fundamental para procesar, limpiar y visualizar información de manera efectiva. Con los EDA, podemos generar perspectivas y predicciones valiosas y relevantes para los estudios de impactos ambientales en el mundo. Al beneficiarnos de librerías como Seabon y Matplotlib, podemos identificar correlaciones y patrones significativos que nos ayudarán a comprender mejor los datos.

Pre-procesamiento del Data Set

El Data Set proveniente de *Kaggle*, “CO2 Emissions by Country”, utilizado en este análisis incluye información detallada sobre las emisiones de CO2 por país, junto con otras variables relevantes como la densidad poblacional, el área geográfica, y la población estimada

en 2022. Los datos fueron procesados utilizando las bibliotecas **Pandas** y **NumPy** en Python, lo que facilitó la manipulación y limpieza de las columnas clave

```
Data columns (total 9 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Country               59620 non-null  object
 1   Code                  57452 non-null  object
 2   Calling Code          56097 non-null  object
 3   Year                  59620 non-null  int64
 4   CO2 emission (Tons)   59620 non-null  float64
 5   Population(2022)      53116 non-null  float64
 6   Area                  55284 non-null  float64
 7   % of World            55284 non-null  object
 8   Density(km2)          53116 non-null  object
dtypes: float64(3), int64(1), object(5)
```

```
IMPRMIENDO ESTADÍSTICAS BÁSICAS DEL DATA FRAME:
                      count          mean           std      min        25%  \
Year                59620.0  1.885000e+03  7.823108e+01   1750.0     1817.0
CO2 emission (Tons) 59620.0  1.034774e+09  1.041652e+10      0.0        0.0
Population(2022)    53116.0  3.992260e+07  1.482365e+08  11312.0  1770414.5
Area                55284.0  6.522073e+05  1.865483e+06     21.0    17704.5

                         50%         75%           max
Year                  1885.0      1953.0  2.020000e+03
CO2 emission (Tons)      0.0   8715092.0  4.170000e+11
Population(2022)   8673095.0  28629200.5  1.425887e+09
Area                110381.5    492573.0  1.709824e+07
```

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

```
Country                0.000000
Code                   3.636364
Calling Code           5.909091
Year                   0.000000
CO2 emission (Tons)    0.000000
Population(2022)      10.909091
Area                   7.272727
% of World             7.272727
Density(km2)          10.909091
dtype: float64
```

Durante el preprocesamiento, también se realizaron transformaciones en las columnas que contenían datos mixtos o valores no numéricos:

- Columna Density(km2), se eliminaron caracteres adicionales como "/km²" para convertirla correctamente a formato numérico.
- Los valores nulos utilizando filtros y se descartaron aquellos registros que contenían emisiones de CO2 igual a cero o valores indefinidos.

Metodología

Para llevar a cabo el análisis del dataset en cuestión, utilicé diferentes técnicas de análisis de datos exploratorio, visualización y algoritmos de aprendizaje no supervisado. La

primera etapa del análisis incluyó la exploración y visualización de los datos utilizando las bibliotecas Matplotlib y Seaborn. Estas herramientas fueron esenciales en el análisis para generar gráficos como diagramas de dispersión, gráficos de barras y mapas de calor, lo que permitió identificar patrones y tendencias en base a cómo se relacionan las variables del conjunto de datos. Los beneficios y la facilidad de uso de la librería Matplotlib definieron la decisión de utilizarla para la visualización de datos, y personalización completa de los gráficos a construir en base a la interpretación del data set.

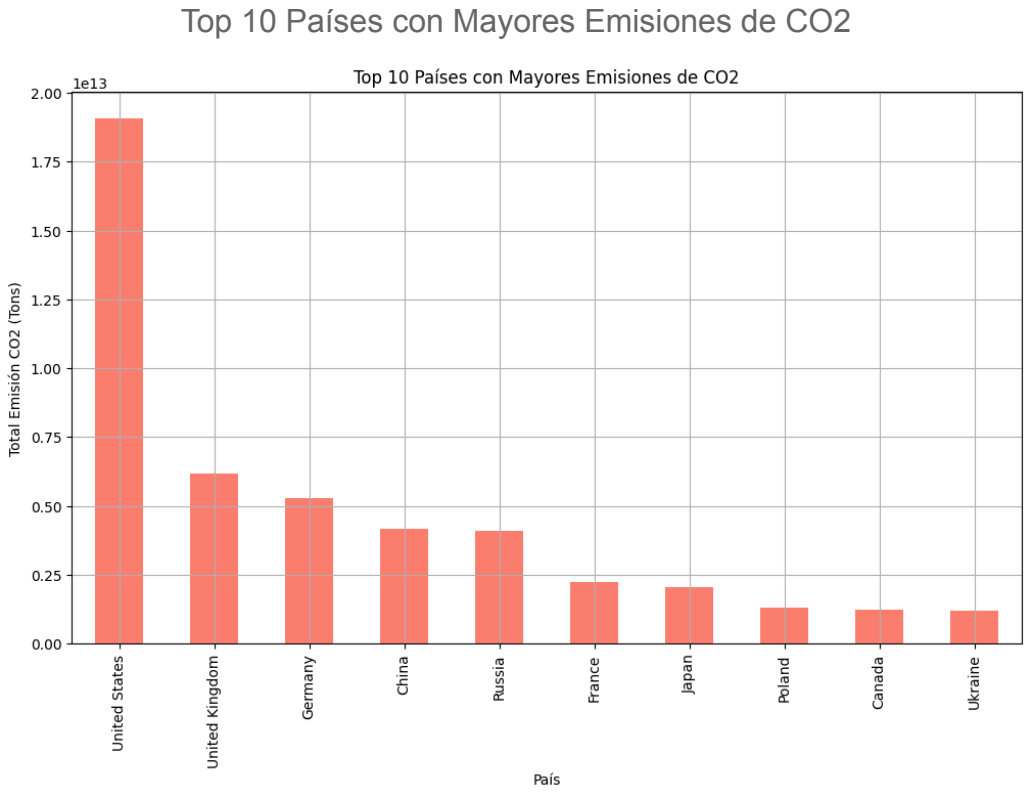
Además de la visualización, se implementó el algoritmo de K-means clustering como método de aprendizaje no supervisado para agrupar los países según sus características compartidas, específicamente la densidad poblacional y las emisiones de CO2.

K-means es ampliamente reconocido y popular por su simplicidad y eficacia en la segmentación de grandes conjuntos de datos. Mediante la aplicación de esta librería en mi código, en conjunto con la estandarización de datos, me fue posible identificar grupso de países con características similares, lo que ayudó a detectar patrones comunes en las emisiones de CO2. Opté por este algoritmo y librería por su capacidad para manejar conjuntos de datos multivariados de manera eficiente y porque ofrece resultados interpretables, lo cual es clave en un análisis orientado a la toma de decisiones.

Resultados

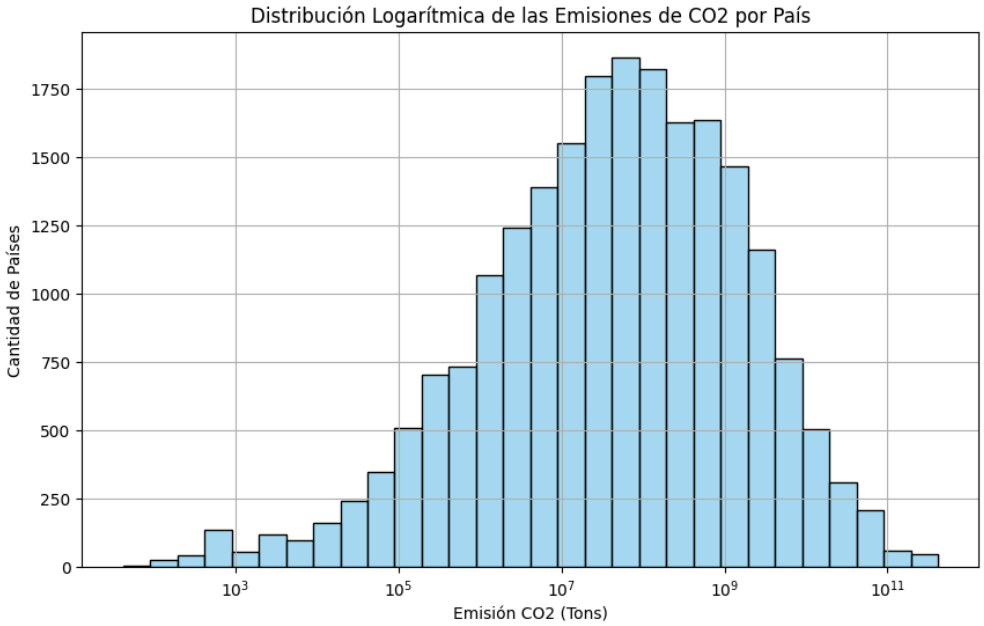
El desarrollo del análisis exploratorio y procesamiento de datos con la aplicación de las diferentes librerías y algoritmos en el código, me permitieron identificar patrones significativos

en la distribución global de emisiones y su relación con variables como la densidad poblacional y la evolución temporal. Los gráficos obtenidos permitieron visualizar tanto la magnitud de las emisiones como las diferencias entre países y grupos, así como identificar tendencias claras en la emisión acumulada a lo largo del tiempo.



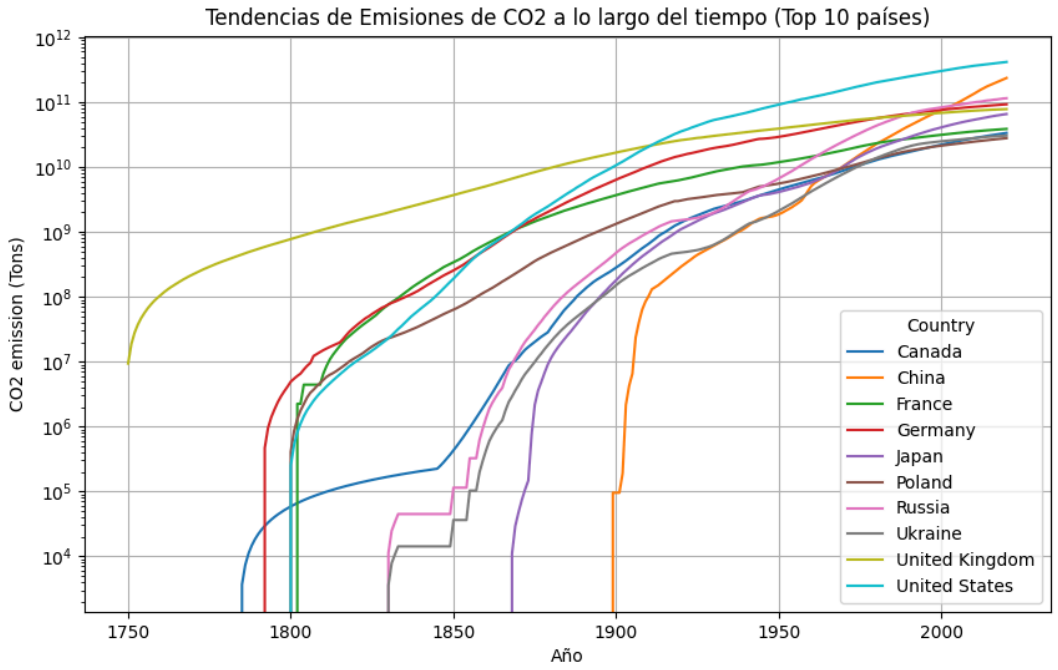
Estados Unidos lidera con un total cercano a las 20,000 millones de toneladas, seguido por Reino Unido y Alemania. Este análisis revela la significativa contribución de los países industrializados a las emisiones globales.

Distribución Logarítmica de las Emisiones de CO2 por País



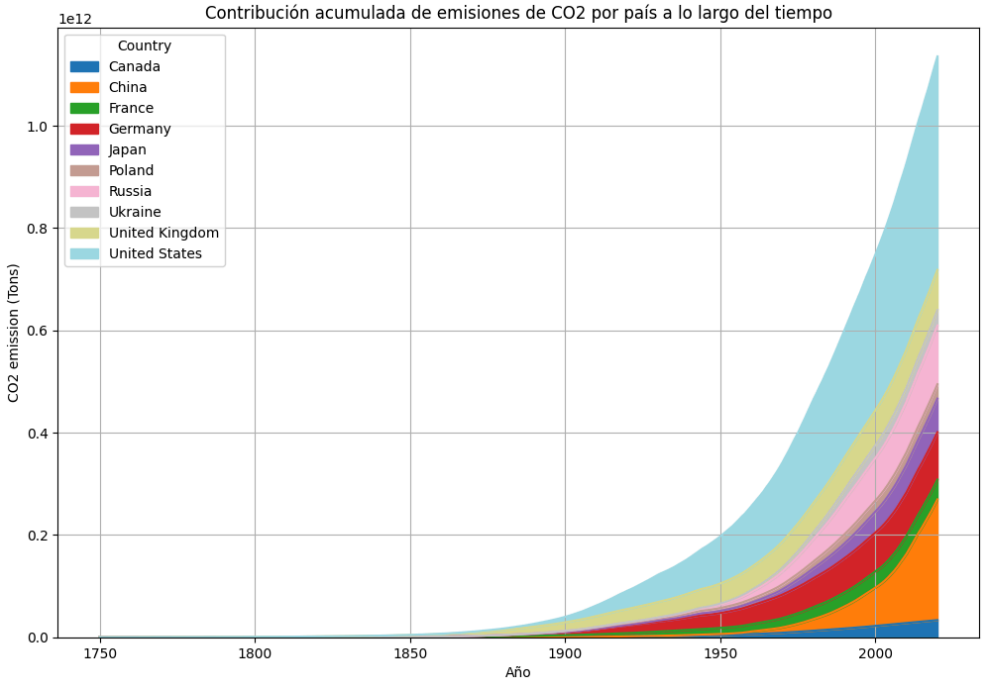
Este histograma presenta la distribución de emisiones de CO2 de todos los países, con una escala logarítmica. Se observa que la mayoría de los países se concentran en niveles más bajos de emisión, mientras que pocos países sobresalen por sus emisiones significativamente más altas.

Tendencias de Emisiones de CO2 a lo largo del Tiempo (Top 10 países)



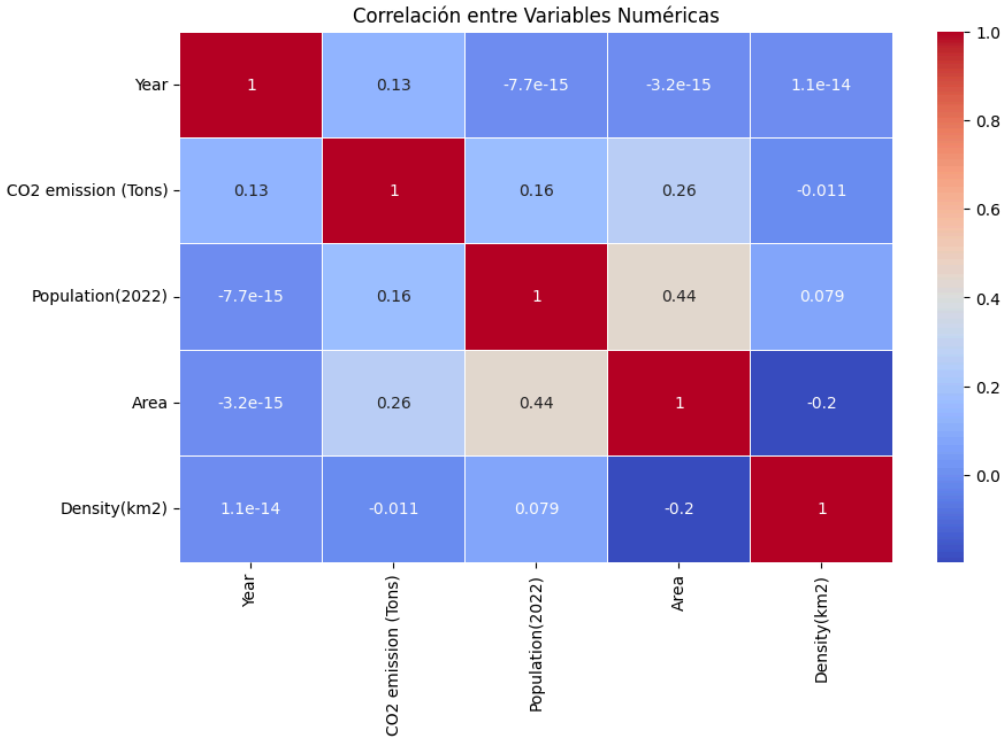
Este gráfico de líneas representa la evolución de las emisiones de CO2 en los 10 principales países emisores desde el año 1750 hasta el presente. Los países como Estados Unidos, China y Rusia muestran un aumento dramático en las emisiones a partir del siglo XX, coincidiendo con el auge industrial.

Contribución Acumulada de Emisiones de CO2 por País a lo largo del Tiempo



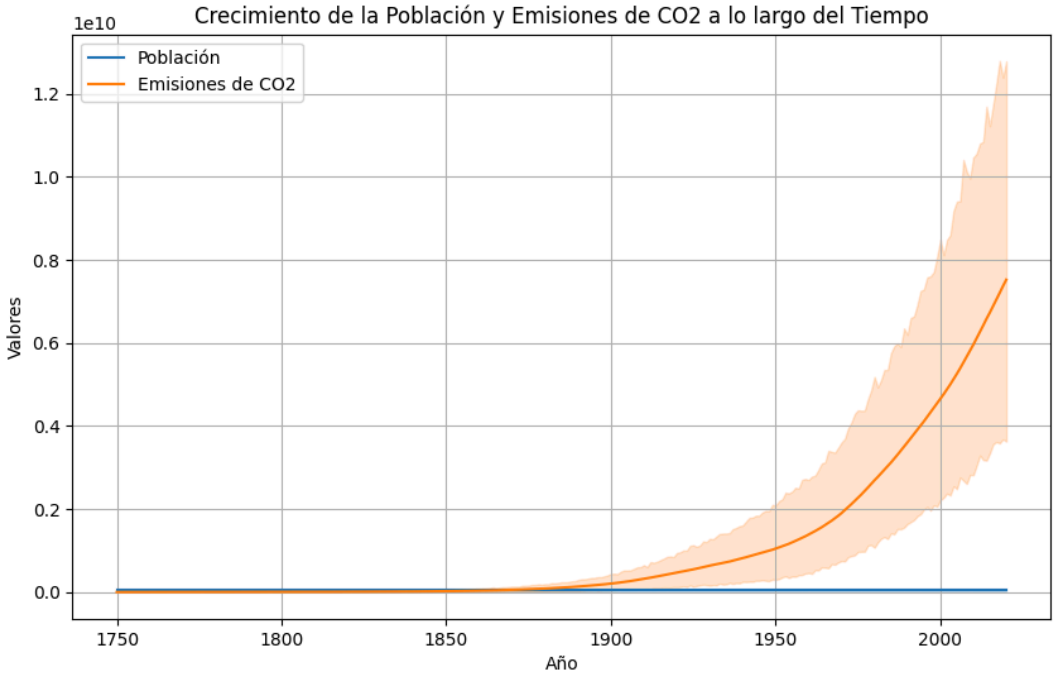
Este gráfico de áreas apiladas muestra cómo los diferentes países han contribuido a las emisiones de CO2 acumuladas a lo largo del tiempo. Los resultados destacan cómo las grandes potencias industriales han aumentado su participación proporcionalmente, sobre todo en las últimas décadas.

Correlación entre Variables Numéricas



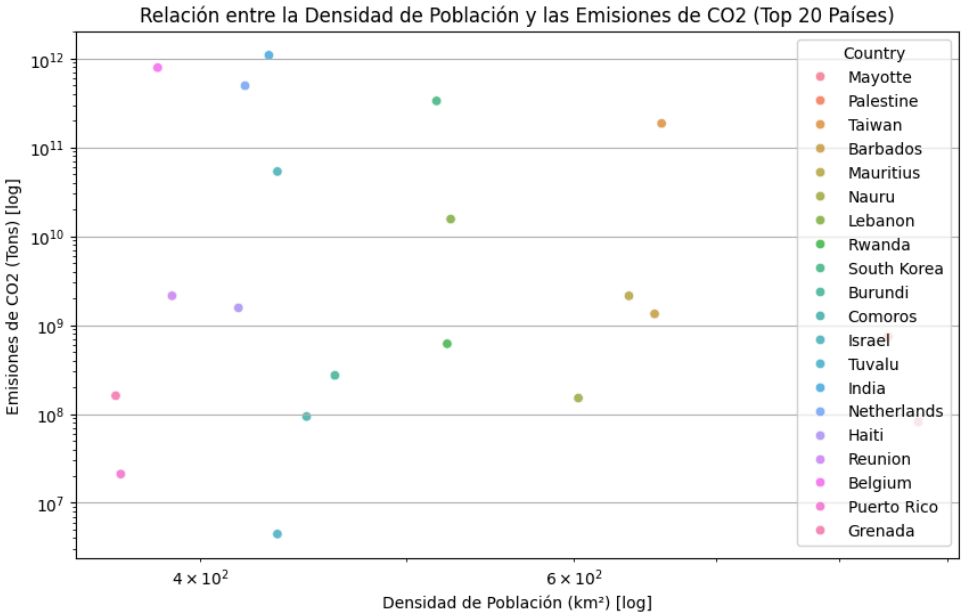
El mapa de calor de correlaciones permite observar la relación entre las variables numéricas del dataset. Aunque no hay correlaciones particularmente fuertes entre las emisiones de CO2 y otras variables como la densidad poblacional o el área, sí se destaca una ligera relación positiva entre la población y las emisiones.

Crecimiento de la Población y Emisiones de CO2 a lo largo del Tiempo



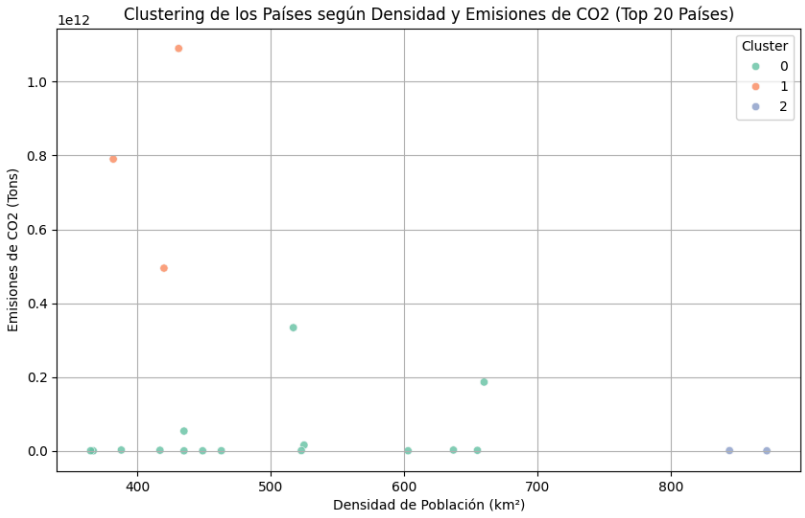
Este gráfico de líneas muestra la evolución conjunta del crecimiento poblacional y las emisiones de CO2 a lo largo del tiempo. Las emisiones de CO2 han crecido de forma mucho más acelerada que la población, lo que sugiere que otros factores, como la industrialización, han contribuido significativamente a este fenómeno.

Relación entre la Densidad de Población y las Emisiones de CO2 (Top 20 Países)



El gráfico de dispersión con escala logarítmica muestra la relación entre la densidad poblacional y las emisiones de CO2 de los 20 países más densamente poblados. Aunque algunos países con alta densidad, como Corea del Sur, también tienen altas emisiones, no parece haber una correlación directa entre ambas variables.

Clustering de los Países según Densidad y Emisiones de CO2 (Top 20 Países)



Este gráfico de dispersión presenta el resultado del algoritmo de clustering K-means, que agrupa a los 20 países según su densidad poblacional y emisiones de CO2. Los países se agrupan en tres clusters, lo que revela patrones comunes en los niveles de emisión de CO2 en función de la densidad poblacional.

Discusión de Resultados

Los resultados obtenidos e identificados mediante el desarrollo del código, y representado por las distintas visualización pintadas en el juypter notebook, destacan la gran disparidad en las emisiones de CO2 entre los países, con Estados Unidos, Reino Unido y Alemania siendo los principales contribuyentes históricos debido a su industrialización temprana. Para dichos países, estos resultados refuerzan la responsabilidad que tienen de adoptar políticas de mitigación más estrictas. A medida que las emisiones han seguido aumentando con el paso del tiempo, y visualizado los cambios en las últimas décadas, este patrón puede sugerir que las grandes economías industrializadas han tenido un impacto desproporcionado en el cambio climático.

El análisis de correlaciones reveló que no existe una relación clara entre la densidad poblacional y las emisiones de CO2, lo que indica que otros factores, como el desarrollo industrial y las políticas energéticas, juegan con un rol más relevante en las emisiones. Con el el clustering mediante los algoritmos implementados por la librería K-means, permitió identificar grupos de países con características similares, lo que podría ayudar a diseñar políticas más efectivas adaptadas a las necesidades específicas de cada grupo o segmentación de países en el mercado y las industrias.

Finalmente, el análisis temporal muestra una aceleración en el crecimiento de las emisiones de CO2 desde mediados del siglo XX, reflejando la expansión industrial global. A pesar de los esfuerzos internacionales para reducir las emisiones, los niveles actuales siguen en aumento, lo que resalta la necesidad de adoptar acciones más contundentes y coordinadas para mitigar el impacto del cambio climático.

Conclusiones

El desarrollo del análisis del conjunto de datos *“CO2 Emissions by Country”* nos ha permitido identificar tendencias importantes y patrones, entre ellos algunos esperados, otros inesperados, y aquellos desproporcionados, de ciertos países respecto a su contribución al cambio climático, y el monitoreo y medida de su huella de carbono. A partir de los resultados obtenidos, se pueden extraer las siguientes conclusiones clave:

- Los países más industrializados, como Estados Unidos y Reino Unido, han contribuido significativamente a las emisiones de CO2 globales, lo que subraya su responsabilidad en la adopción de políticas más estrictas de reducción de emisiones.

- La densidad poblacional no parece ser un predictor fuerte de las cuales dependen en su totalidad las emisiones de CO2. Esto siguiere que existen otros factores, como el nivel de industrialización y el tipo de políticas energéticas, sostienen igual o mayor relevancia respecto a las métricas y resultados de los países.

- El análisis realizado mediante clustering agrupó países con características similares en términos de densidad y emisiones, lo que podría ser útil para diseñar estrategias específicas según las necesidades de cada grupo segmentado por dichas similitudes.

Igualmente, el estudio ha revelado patrones temporales que destacan la aceleración en el crecimiento de las emisiones en las últimas décadas, especialmente desde la Revolución Industrial. Se destacan ciertos resultados por su relevancia para la toma de decisiones y acciones futuras por los países:

- La continua expansión industrial de los países sigue aumentando las emisiones, a pesar de los esfuerzos internacionales, lo que refuerza la necesidad de medidas más efectivas y globalmente coordinadas en base la naturaleza de la velocidad en la que el mundo va evolucionando y creciendo.

- Es crucial que los países con menores niveles de emisión mantengan políticas de sostenibilidad mientras los grandes emisores reduzcan drásticamente sus emisiones para mitigar el cambio climático de manera efectiva.

Referencias

CO2 Emission by countries Year wise (1750-2022). (2022, September 14). Kaggle. https://www.kaggle.com/datasets/moazzimalibhatti/co2-emission-by-countries-year-wise-17502022

Kashnitsky. (2021, May 7). Topic 1. Exploratory Data Analysis with Pandas. Kaggle. https://www.kaggle.com/code/kashnitsky/topic-1-exploratory-data-analysis-with-pandas

Selvaraj, N. (2023, April 21). A Beginner's Guide to data analysis in Python. 365 Data Science. https://365datascience.com/tutorials/python-tutorials/data-analysis-python/

TestGorilla. (2022, December 5). What is Matplotlib in Python? Top 10 advantages of Matplotlib that you should know. https://www.testgorilla.com/blog/matplotlib-in-python/

Cheng, Y., Awan, U., Ahmad, S., & Tan, Z. (2021). How do technological innovation and fiscal decentralization affect the environment? A story of the fourth industrial revolution and sustainable growth. Technological Forecasting and Social Change, 162, 120398. https://www.sciencedirect.com/science/article/abs/pii/S0040162520312245

PasanBhathiya. (2024, August 22). Customer Segmentation Using K-Means Clustering with Python. Medium. https://medium.com/@pasanbathiya246/customer-segmentation-using-k-means-clustering-with-python-86267ef14931