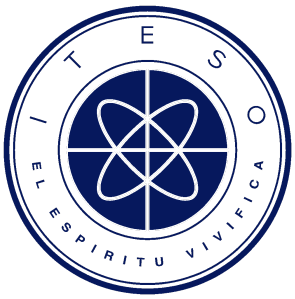
Instituto Tecnológico y de Estudios Superiores de Occidente

**PROGRAMACIÓN PARA ANÁLISIS DE DATOS**




# ¿CUÁL ES LA RELACIÓN ENTRE EL USO DE ENERGÍA NUCLEAR Y LAS EMISIONES DE CO2 DESDE 1965 A 2021?

Presentan

**Naranjo Salgado David Abraham 730697**

**Veloso Ramirez Ana Paulina 720517**

**José Jorge VIllarreal Farías**

Profesor: Mtra. Gisel Hernández Chávez

Fecha <12 05 2023>

**Revisado por**:

**Aprobado por**:

# Contenido

## Índice de Tablas

## Lista de Figuras

# Introducción

## 1.1 Enunciado del problema y preguntas de investigación

Para este proyecto se busca hacer un análisis de la relación entre la generación de energía nuclear (una de las energías con menores emisiones de CO2) y comparar la relación de las emisiones de CO2 generadas por los países que generador de energía nuclear desde 1965 hasta el 2021 utilizando como puntos de referencia las unidades de Terawhatt por hora generados por cada país anualmente y las emisiones anuales per cápita de CO2 .

Los países seleccionados para este estudio fueron los que han generado al menos un Terawhatt al año desde 1965 hasta 2021 dando un total de 36 países con 57 entradas cada uno.

Estos datos fueron obtenidos mediante los svc de

Nuclear energy generation (Ritchie, Rosado, & Roser, 2022)

Ritchie, H., Rosado, P., & Roser, M. (2022). Our World In Data. Obtenido de CO2 emissions

Las tablas utilizadas para el análisis de datos fueron mezcladas manualmente para esta entrega, en donde se desestimaron todos los países que no generaron energía nuclear dentro del rango de años, haciendo la excepción de RUSIA, ARMENIA, BIELORRUSIA, KAZAJSTÁN, KIRGUIZISTÁN los cuales se tomaron los datos faltantes desde 1965 hasta 1985 y fueron asignados los valores de la URSS, a su vez se desestimaron las sumatorias por continentes, bloques y regiones exceptuando Los valores mundiales.

También se unieron los 2 svc, mediante los países delimitando los valores de las emisiones de CO2 de 1965 a 2021 para cada uno de los 36 países que generaron energía nuclear, más la sumatoria del mundo.

Unidades de análisis

Electricity(TWh/Terawatt Hora): Razón

Valores válidos: 0 - ∞

Significado: es la cantidad de energía generada a partir de energía nuclear por país anualmente

Emissions(CO2): Razón

Valores válidos: 0 - ∞

Significado: es la cantidad de emisiones de dióxido de carbono generada anualmente por país

Preguntas de investigación

¿Cuál ha sido la tendencia por país para generación nuclear?

¿Hubo una reducción de emisiones de CO2 en los años que cada país generó mayor energía nuclear?

¿Hubo alguna relación lineal entre el punto más alto de energía nuclear con la generación de CO2?

¿Existe una tendencia significativa en la generación de energía nuclear a lo largo del tiempo por continente?

¿Existe una tendencia significativa en las emisiones anuales de CO2 a lo largo del tiempo por continente?

¿Existe una relación entre las emisiones anuales de CO2 y la población de los países?

¿Existe una relación entre las emisiones anuales de CO2 y generación de energía nuclear de los países?

## 1.2 Justificación

Este estudio es importante ya que estamos interesados en observar 2 factores relevantes en la actualidad como lo son la generación de energía y las emisiones de dióxido de carbono, y mediante estas 2 medidas queremos encontrar si existe una relación lineal entre el aumento de generacion de energia nuclear y la disminución de emisiones de dióxido de carbono.

## 1.3 Objetivos

Objetivo general:

1- Aplicar conocimientos de programación para análisis de datos en Python empleando bibliotecas como numpy, pandas, matplotlib, seaborn, statsmodels y scikit learn.
2- Recopilar datos y estructurar el dataset que permita responder a las preguntas de investigación.

3- Aplicar conocimientos de Probabilidades y Estadísticas durante la etapa de exploración y preparación de datos, fundamentalmente las relativas a estadística descriptiva, formulación y pruebas de hipótesis, así como análisis de regresión con fines explicativos en nuestro cosos con el uso de herramientas como el ANOVA RM.
4- Visualizar datos durante la exploración de los mismos
5- Documentar los resultados del proceso de análisis de datos exploratorio, con énfasis en la redacción de hallazgos y conclusiones
6- Seguir procesos y etapas de una metodología de Minería de Datos, fundamentalmente en sus fases iniciales de comprensión del negocio, comprensión de los datos, preparación de los datos para el modelado y modelado con fines explicativos

## 1.4 Vista general del documento

Capítulo 2: planeación de etapas y distribución de actividades.

Capítulo 3:Explicación de la obtención de datos así como su manipulación y características especiales.

Capítulo 4:Análisis de datos univariados con el fines exploratorios de los data frames.

Capítulo 5:Detallar la implementación del proyecto y plasmarlo mediante diagramas.

Capitulo 6:Conclusiones del equipo y de la lectura encontrada en https://www.questionpro.com/blog/exploratory-data-analysis/#:~:text=Conclusion,might%20not%20have%20found%20otherwiseLinks to an external site..

Capítulo 7: Referencias bibliográficas citadas en formato APA.

# 2 Plan de Análisis de Datos

Etapas de proyecto:

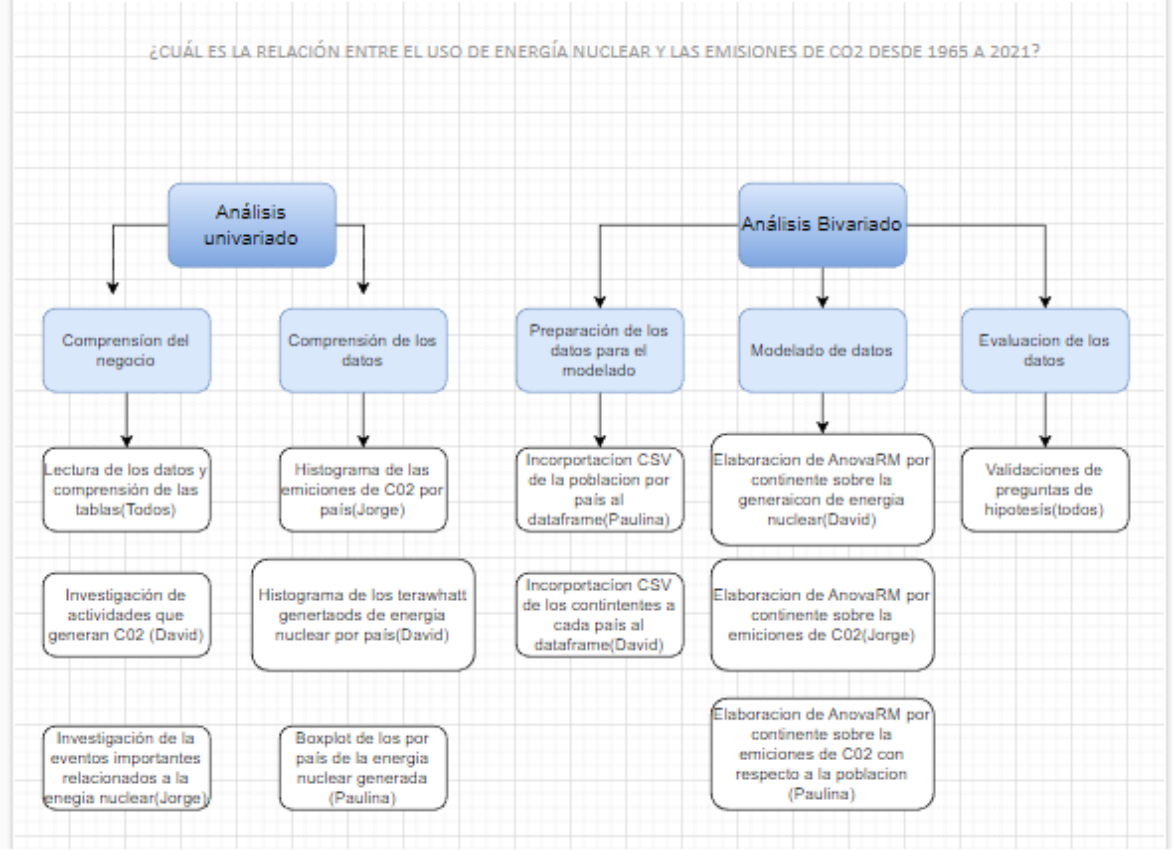
A)Análisis univariado

-Comprensión del negocio
-Comprensión de los datos

B)Análisis bivariado

-Preparación de los datos para el modelado
-Modelado de datos

-Evaluación de los datos.




*Fuente: elaboración propia*

# 3 Adquisición y comprensión de los datos

Nuclear energy generation (Ritchie, Rosado, & Roser, 2022)

| | Nuclear terawatt-hours | | | |
|---|---|---|---|---|
| Country | 1965 | 2022 | Absolute Change | Relative Change |
| Afghanistan | 2000 0.00 TWh | 2021 0.00 TWh | +0.00 TWh | |
| Africa | 0.00 TWh | 2021 12.15 TWh | +12.15 TWh | |
| Africa (BP) | 0.00 TWh | 2021 10.42 TWh | +10.42 TWh | |
| Africa (Ember) | 2000 13.01 TWh | 2021 12.15 TWh | -0.86 TWh | -7% |
| Albania | 1990 0.00 TWh | 2020 0.00 TWh | +0.00 TWh | |
| Algeria | 0.00 TWh | 2021 0.00 TWh | +0.00 TWh | |
| American Samoa | 2000 0.00 TWh | 2021 0.00 TWh | +0.00 TWh | |
| Angola | 2000 0.00 TWh | 2021 0.00 TWh | +0.00 TWh | |
| Antigua and Barbuda | 2000 0.00 TWh | 2021 0.00 TWh | +0.00 TWh | |
| Argentina | 0.00 TWh | 2021 10.17 TWh | +10.17 TWh | |

CO2 emissions (Ritchie, Rosado, & Roser, 2022)

| | Annual $CO_2$ emissions tonnes | | |
|---|---|---|---|
| Country | 1750 | 2021 | Absolute Change |
| Afghanistan | 1949 14,656.00 t | 11,874,211.00 t | +11,859,555.00 t |
| Africa | 0.00 t | 1,450,796,300.00 t | +1,450,796,300.00 t |
| Africa (GCP) | 1850 0.00 t | 1,450,782,600.00 t | +1,450,782,600.00 t |
| Albania | 1933 7,328.00 t | 4,619,109.00 t | +4,611,781.00 t |
| Algeria | 1916 3,664.00 t | 176,269,070.00 t | +176,265,406.00 t |
| Andorra | 0.00 t | 452,888.00 t | +452,888.00 t |
| Angola | 1950 186,864.00 t | 21,362,716.00 t | +21,175,852.00 t |
| Anguilla | 1990 51,296.00 t | 144,744.00 t | +93,448.00 t |
| Antarctica | 1987 3,664.00 t | 2007 10,992.00 t | +7,328.00 t |
| Antigua and Barbuda | 1957 21,984.00 t | 468,695.00 t | +446,711.00 t |
| Argentina | 1887 1,084,544.00 t | 186,448,290.00 t | +185,363,746.00 t |

Estos son los datos crudos originales, Energía Nuclear por país entre 1965 y 2021, y Emisiones de C02 por país entre 1965 y 2021, los cuales tuvimos que modificarlos debido a que algunos países desaparecieron o cambiaron a lo largo de los años, como en el caso de la unión soviética y países relacionados, por lo que nuestra solución fue, tomar los países antecesores y tomarlos como el país que son en la actualidad (unión soviética como rusia por ejemplo)

Listado de países a analizar.

Argentina ,Armenia, Netherlands , Pakistan, Romania, Russia, Slovakia , Slovenia, South Africa ,South Korea , Spain , Sweden , Switzerland , Taiwan , Ukraine , United Arab Emirates , United Kingdom , Mexico , Lithuania , Kazakhstan , Czechia , Belarus , Belgium , Brazil , Bulgaria , Canada , China , Finland , Japan , France , Germany , Hungary , India , Iran , Italy y United States.

Dando un total de 36 países con 57 entradas cada uno.

World Population Growth (Ortiz, Ritchie, Rodes & Roser, 2022)

| | Population people | | | |
|---|---|---|---|---|
| Country | 1950 | 2021 | Absolute Change | Relative Change |
| Afghanistan | 7,480,464 | 40,099,460 | +32,618,996 | +436% |
| Africa (UN) | 227,549,260 | 1,393,676,400 | +1,166,127,140 | +512% |
| Albania | 1,252,587 | 2,854,710 | +1,602,123 | +128% |
| Algeria | 9,019,866 | 44,177,964 | +35,158,098 | +390% |
| American Samoa | 19,057 | 45,056 | +25,999 | +136% |
| Andorra | 6,028 | 79,057 | +73,029 | +1,211% |
| Angola | 4,478,186 | 34,503,776 | +30,025,590 | +670% |
| Anguilla | 5,036 | 15,779 | +10,743 | +213% |
| Antigua and Barbuda | 45,456 | 93,229 | +47,773 | +105% |
| Argentina | 17,017,748 | 45,276,788 | +28,259,040 | +166% |
| Armenia | 1,385,038 | 2,790,971 | +1,405,933 | +102% |
| Aruba | 38,818 | 106,543 | +67,725 | +174% |
| Asia (UN) | 1,379,048,300 | 4,694,576,000 | +3,315,527,700 | +240% |
| Australia | 8,177,169 | 25,921,094 | +17,743,925 | +217% |
| Austria | 6,936,443 | 8,922,086 | +1,985,643 | +29% |
| Azerbaijan | 3,158,966 | 10,312,992 | +7,154,026 | +226% |
| Bahamas | 81,651 | 407,920 | +326,269 | +400% |
| Bahrain | 117,160 | 1,463,266 | +1,346,106 | +1,149% |
| Bangladesh | 39,728,540 | 169,356,240 | +129,627,700 | +326% |
| Barbados | 210,556 | 281,204 | +70,648 | +34% |

Continents according to Our World In Data (Our World In Data, 2022)

| Country | Continent 2015 |
|---|---|
| Abkhazia | Asia |
| Afghanistan | Asia |
| Akrotiri and Dhekelia | Asia |
| Albania | Europe |
| Algeria | Africa |
| American Samoa | Oceania |
| Andorra | Europe |
| Angola | Africa |
| Anguilla | North America |
| Antarctica | Antarctica |
| Antigua and Barbuda | North America |
| Argentina | South America |
| Armenia | Asia |
| Aruba | North America |
| Australia | Oceania |
| Austria | Europe |
| Austria-Hungary | Europe |
| Azerbaijan | Asia |
| Baden | Europe |
| Bahamas | North America |
| Bahrain | Asia |
| Bangladesh | Asia |

Estos siguientes 2 grupos de datos, fueron utilizados para nutrir el análisis bivariado, tomando el papel de dar una tercera columna de información para tomar conclusiones, y para poder agrupar los datos para llevar a otras conclusiones

## 3.1 Tipo de estudio

Este proyecto es un estudio de datos longitudinales, ya que se analizan los datos de 1965 hasta 2021

## 3.2 Datos en memoria externa (archivos)

**Tabla 3-1 Descripción de archivos de datos originales**

| *Nombre del archivo* | *Breve descripción* | *Formato (txt, csv, xlsx, json, mp4, jpg, etc.)* |
|---|---|---|
| *continents-according-to-our-world-in-data.csv* | *Archivo que contiene la información de los continentes* | *csv* |
| *nuclear-energy-generation.xlsx* | *Archivo con la fuente de datos central* | *xlsx* |
| *population-and-demography.csv* | *Archivo con la información de la población de los países* | *csv* |

*Fuente: elaboración propia*

**Tabla 3-2 Especificación de columnas de archivo continents-according-to-our-world-in-data.csv**

| *Nombre del campo o columna* | *Tipo csv* | *Tipo de dato según escala de medición estadística* | *Tipo sugerido en Python pandas* | *Valores válidos* |
|---|---|---|---|---|
| *Entity* | *Cadena de caracteres* | *nominal* | *category* | *Cadena con el nombre del país* |
| *Code* | *Cadena de caracteres* | *nominal* | *category* | *3 caracteres* |

| | | | | |
|---|---|---|---|---|
| *Year* | *Número* | *Intervalo* | *Int* | *4 Números* |
| *Continent* | *Cadena de caracteres* | *nominal* | *category* | *Cadena con el nombre del continente* |

*Fuente: Elaboración propia*

**Tabla 3-3 Especificación de columnas de archivo nuclear-energy-generation.xlsx**

| *Nombre del campo o columna* | *Tipo xlsx* | *Tipo de dato estadístico* | *Tipo sugerido en Python pandas* | *Valores válidos* |
|---|---|---|---|---|
| *Entity* | *Cadena* | *nominal* | *category* | *Nombre de país* |
| *Code* | *Cadena* | *nominal* | *category* | *3 Caracteres* |
| *Year* | *numérico* | *Intervalo* | *Int* | *4 números (año)* |
| *Nuclear_Electricity* | *numérico* | *Razón* | *Float32* | *0 hasta el máximo* |
| *Annual_CO2_emissions* | *numérico* | *Razón* | *Float32* | *0 hasta el máximo* |

*Fuente: Elaboración propia*

**Tabla 3-4 Especificación de columnas de archivo population-and-demography.csv**

| *Nombre del campo o columna* | *Tipo xlsx* | *Tipo de dato estadístico* | *Tipo sugerido en Python pandas* | *Valores válidos* |
|---|---|---|---|---|
| *Country* | *Cadena* | *nominal* | *category* | *Nombre de país* |
| *Year* | *numérico* | *Intervalo* | *int* | *4 números (año)* |
| *Population* | *numérico* | *Razón* | *Float32* | *0 al máximo valor* |

## 3.3 Preprocesamiento. Primera parte.

El preprocesamiento de los datos crudos iniciales, está compuesto por 2 datasets, los cuales los unimos por medio de excel (no por medio de código), generando este archivo nuclear-energy-generation.xlsx, así como eliminamos los países que tuvieron una producción nula o muy baja de energía nuclear en los años

### 3.3.1 Transformaciones de tipos de datos

```
nominales = ['Entity', 'Code']
intervalo = ['Year']
razon = ['Nuclear_Electricity', 'Annual_CO2_emissions']



df['Entity'] = df['Entity'].astype('category')
df['Code'] = df['Code'].astype('category')
df.info()
```

La única transformación significativa que hicimos en nuestros datos centrales, fue indicar las columnas de entidad y código como tipo categoría.

### 3.3.2 Datos faltantes. Imputaciones y eliminaciones

- Descartamos los países con una producción nula o muy baja de energía nuclear, centrando nuestra información en 36 países
- Hay países que cambiaron a lo largo del tiempo, por ejemplo la unión soviética, tomamos la decisión de nombrarlos como el país al que pertenece en la actualidad

### 3.3.3 Eliminación de duplicados

No tenemos Información duplicada

## 3.4 Conclusiones del capítulo

Lo que nos hace darnos cuenta este capítulo es la importancia de obtener un buen dataset ya que la confiabilidad, precisión y sesgos de estos datos va a determinar de manera muy clara la cantidad y presión de los métodos que sean usados para obtener resultados a través de dichos datasets.

Además de esto creemos que es importante poder entender los tipos de datos y el tipo de estudio con el fin de poder elaborar preguntas acordes y relevantes a los tipos de datos obtenidos.

# 4 Análisis de Datos Exploratorio

## 4.1 Análisis descriptivo univariado

### 4.1.1 Análisis descriptivo univariado de datos nominales

Para este estudio, se tuvieron columnas de tipo nominal, de intervalo y de razón, a continuación se especifican cada una de ellas.

**Nominales:** Entity y Code, que son los países y los códigos que los identifican.

**Intervalo:** Year, que son los años en los que se obtuvieron los valores buscados.

**Razón:** Nuclear electricity y Annual_CO2_emissions, qué son las cantidades de energía nuclear generada y las emisiones generadas por país y por año.

### 4.1.2 Análisis descriptivo univariado de datos ordinales

En este estudio no se tuvo ningún dato de tipo ordinal.

### 4.1.3 Análisis descriptivo univariado de datos de intervalo

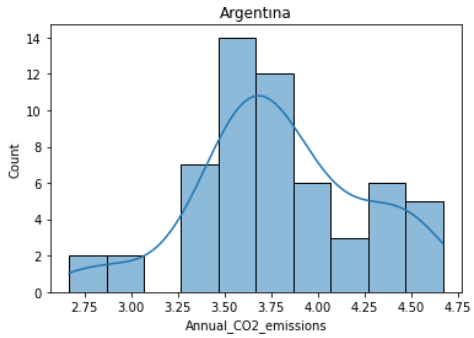
Los datos de intervalo, en este caso no fueron estudiados directamente, sino que fueron herramienta para estudiar los datos de tipo razón. Esto se debe a que nuestro estudio era de tipo longitudinal de panel, por lo que solo eran valores independientes.

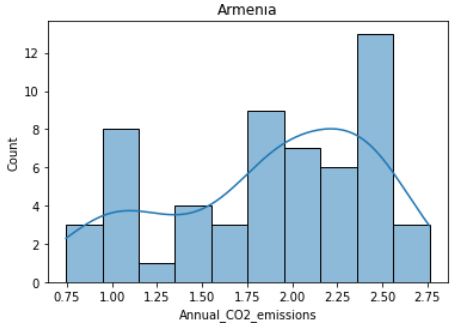
#### *4.1.3.1 Valores atípicos*

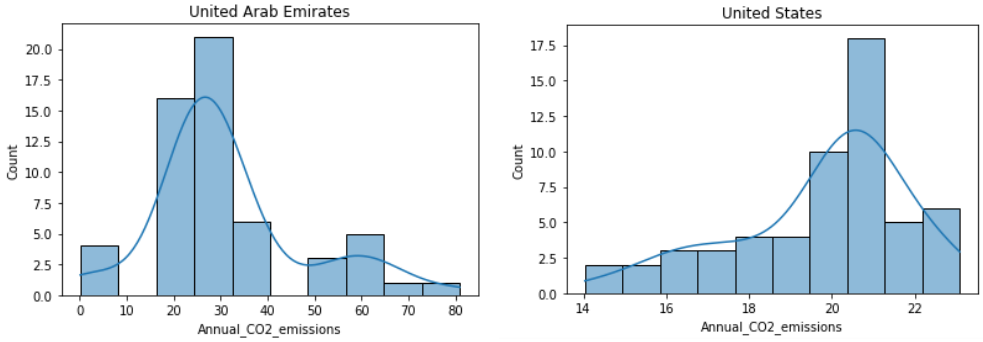
Para este caso, solo se realizaron modificaciones al momento de llevar a cabo la limpieza de datos, acotando el rango de años para que pudieran tener los mismos valores ambos datasets utilizados (Energía nuclear y Emisiones de CO2).

### 4.1.4 Análisis descriptivo univariado de datos de razón

El primer análisis que se realizó fue una serie de visualizaciones de histogramas de las emisiones anuales de CO2 para cada país, dándonos así información sobre cuál rango de valores es el que comúnmente emite cada país. Al tener 36 países distintos, se muestran algunos de estos histogramas a continuación:






A partir del análisis anterior, se obtuvo el primer indicio de que estados unidos y emiratos árabes iban a ser outliers, puesto que sus valores de emisión siempre eran mayores que los del resto. Los hallazgos principales de este estudio fueron:

- Para todos los países se observa de manera general una producción de energía nuclear intermitente, presentando alta frecuencia en valores cercanos a cero y con menor frecuencia en otros valores.
- Puesto que para una gran parte de los países presentados en el dataset las tecnologías necesarias para generar la energía nuclear fueron adoptadas posteriormente a 1965, existe una discrepancia, no permitiendo que la comparación de los histogramas sea representativa.

Posteriormente, se llevó a cabo un análisis aritmético para observar, valores de media, mínimos, máximos y total. Se muestra a continuación:

```
AritmeticOPerations.mean()
```

| Entity | Nuclear_Electricity | Annual_CO2_emissions |
|---|---|---|
| Argentina | 4.977175 | 3.799055 |
| Armenia | 0.850526 | 1.881491 |
| Belarus | 14.399708 | 7.510111 |
| Belgium | 30.501860 | 11.442265 |
| Brazil | 5.642649 | 1.676097 |
| Bulgaria | 11.684579 | 7.335813 |
| Canada | 62.099263 | 16.505319 |
| China | 54.450667 | 3.329578 |
| Czechia | 13.149123 | 13.840003 |
| Finland | 15.655494 | 10.135981 |
| France | 264.642491 | 7.277892 |
| Germany | 98.327789 | 11.754664 |
| Hungary | 9.222368 | 6.339608 |

```
AritmeticOPerations.sum()
```

| Entity | Nuclear_Electricity | Annual_CO2_emissions |
|---|---|---|
| Argentina | 283.699000 | 216.546118 |
| Armenia | 48.480000 | 107.244963 |
| Belarus | 820.783342 | 428.076306 |
| Belgium | 1738.606000 | 652.209084 |
| Brazil | 321.631000 | 95.537544 |
| Bulgaria | 666.021000 | 418.141355 |
| Canada | 3539.657995 | 940.803165 |
| China | 3103.688000 | 189.785949 |
| Czechia | 749.500000 | 788.880152 |
| Finland | 892.363157 | 577.750918 |
| France | 15084.622000 | 414.839840 |
| Germany | 5604.684000 | 670.015855 |
| Hungary | 525.675000 | 361.357681 |

```
AritmeticOPerations.min()
```

| Entity | Nuclear_Electricity | Annual_CO2_emissions |
|---|---|---|
| Argentina | 0.000000 | 2.666444 |
| Armenia | 0.000000 | 0.744803 |
| Belarus | 0.000000 | 5.293781 |
| Belgium | 0.000000 | 7.816142 |
| Brazil | 0.000000 | 0.664164 |
| Bulgaria | 0.000000 | 5.296769 |
| Canada | 0.128000 | 12.794933 |
| China | 0.000000 | 0.604760 |
| Czechia | 0.000000 | 8.722278 |
| Finland | 0.000000 | 5.542695 |
| France | 0.897000 | 4.342917 |
| Germany | 0.117000 | 7.672972 |
| Hungary | 0.000000 | 4.407595 |

```
AritmeticOPerations.max()
```

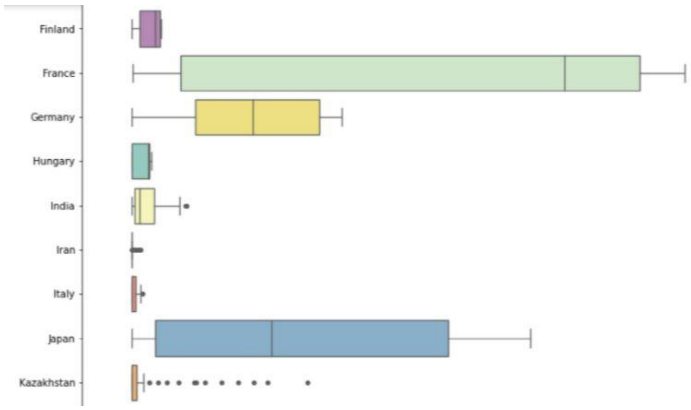
| Entity | Nuclear_Electricity | Annual_CO2_emissions |
|---|---|---|
| Argentina | 10.17000 | 4.670683 |
| Armenia | 2.57000 | 2.761147 |
| Belarus | 143.79843 | 11.900442 |
| Belgium | 50.33000 | 14.251578 |
| Brazil | 15.17000 | 2.742072 |
| Bulgaria | 20.22000 | 10.295022 |
| Canada | 107.08421 | 18.469010 |
| China | 407.50000 | 8.045740 |
| Czechia | 30.75000 | 18.386303 |
| Finland | 23.87000 | 13.937167 |
| France | 451.53000 | 10.396503 |
| Germany | 171.30000 | 14.347402 |
| Hungary | 16.29000 | 8.558931 |

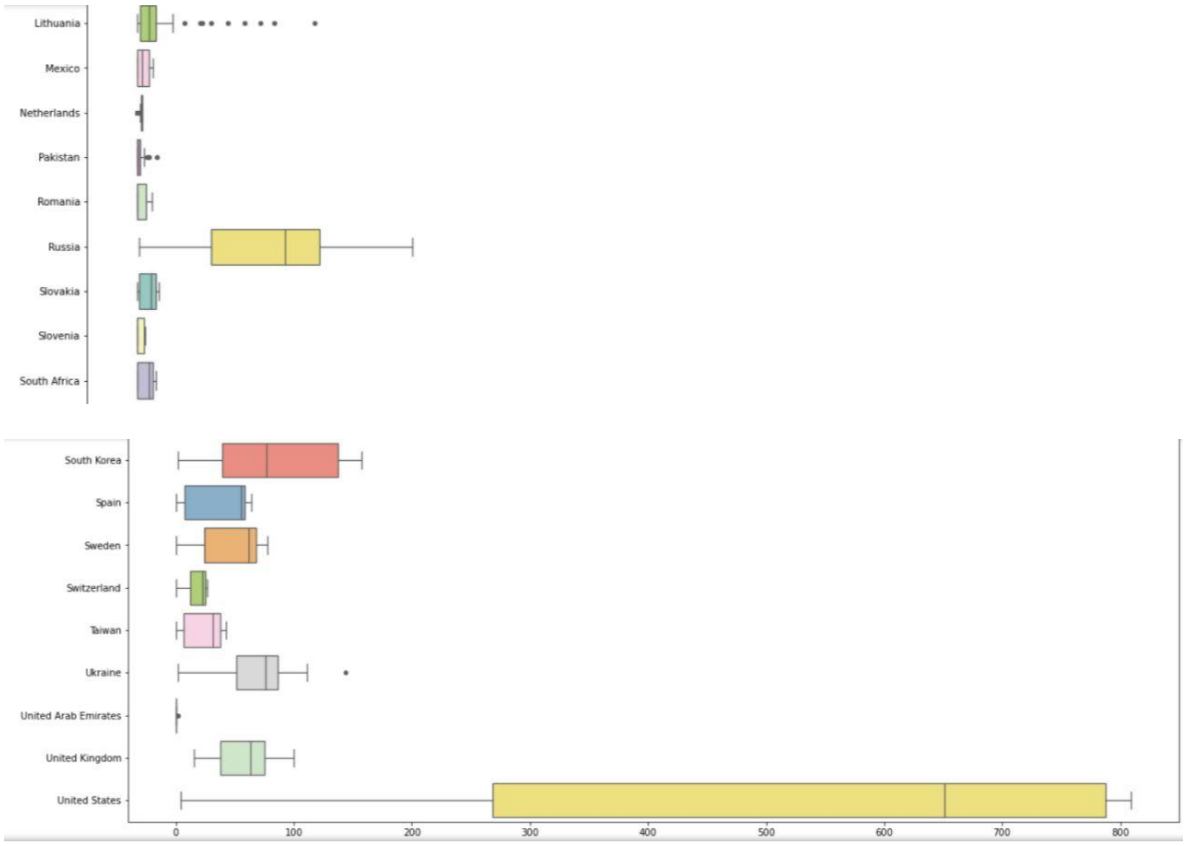
Los hallazgos, encontrados a partir de lo anterior es lo siguiente:

- Anualmente, la mayoría de los países presentan una producción media de energía nuclear menor a 20 TWh.
- De igual forma, anualmente la mayoría de los países generan al menos 10 toneladas de CO2.
- El país con menos producción de energía nuclear (United Arab Emirates) tiene la mayor tasa de CO2.
- El país con mayor producción de energía nuclear a través de los años es Japón, con un valor de 7792.162860 TWh.
- El país con menor producción de energía nuclear es United Arab Emirates.
- El país con menor cantidad de producción de CO2 es Pakistán.
- El rango de producción de energía nuclear es amplio, desde 3.35 TWh hasta 7792.162860 TWh.
- El rango de generación de emisiones de CO2 va desde 33.8875 toneladas hasta 1803.6192 toneladas.

- En la mayoría de los países la producción de energía nuclear fue nula en cierto periodo de tiempo.
- United Kingdom presenta el mínimo mayor de producción de energía nuclear.
- La producción mínima de emisiones de CO2 es de United Arab Emirates.
- United States presenta el mínimo mayor de producción de CO2.
- El país con una mayor capacidad productiva de energía nuclear es United States.
- El país con menor capacidad productiva de energía nuclear es United Arab Emirates.
- Pakistan presenta la menor cantidad de emisiones de CO2 entre los valores máximos.
- El mayor productor de emisiones de CO2 entre todos los países es United Arab Emirates.

Para observar el comportamiento general de generación de energía nuclear por país, se generaron boxplots, permitiendo así ver cómo se comportan en términos de generación. A continuación se observan dichos boxplots.




A partir de lo anterior, se pudieron observar las cantidades de energía nuclear media que producían cada país, así como los valores atípicos. En muchos casos, la generación fue muy cercana a cero, por lo que la tendencia general de varios de estos países eran hacia esa cantidad, resaltando solo países como estados unidos, francia, japón y rusia. Los hallazgos son:

- El país con una mediana mayor en términos de producción de energía nuclear es United States.
- El país con la mínima menor en términos de energía nuclear es United Arab Emirates.
- Países aparte de United States que presentan datos de generación de energía nuclear más variada son: Francia y Japón.

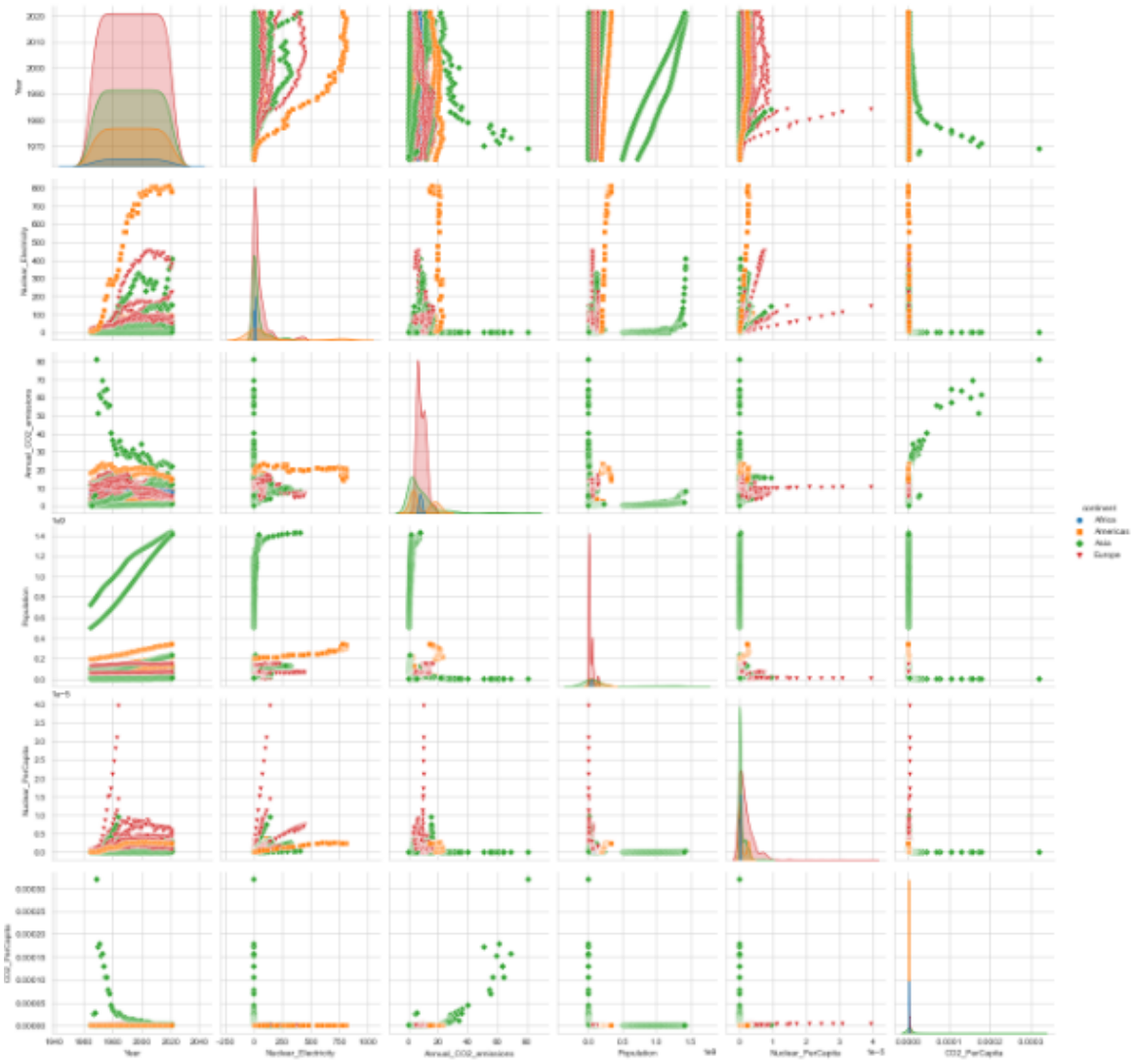
#### *4.1.4.1 Valores atípicos*

Para evitar valores atípicos, al momento de realizar la limpieza de datos, se acotaron la cantidad de países para así obtener un análisis relevante.

## 4.2 Análisis bivariado descriptivo y relacional

### 4.2.1 Entre dos variables categóricas (nominales u ordinales)

Lo primero que se realizó fue un pair plot para observar el comportamiento entre los distintos datos que se tienen en el dataset.



De lo anterior se concluyó que al tener un set de datos longitudinales de panel, un estudio longitudinal (por país) era más adecuado. No se observaron relaciones lineales claras entre los datos.

### 4.2.2 Entre una categórica (nominal u ordinal) y una numérica (de intervalo o razón)

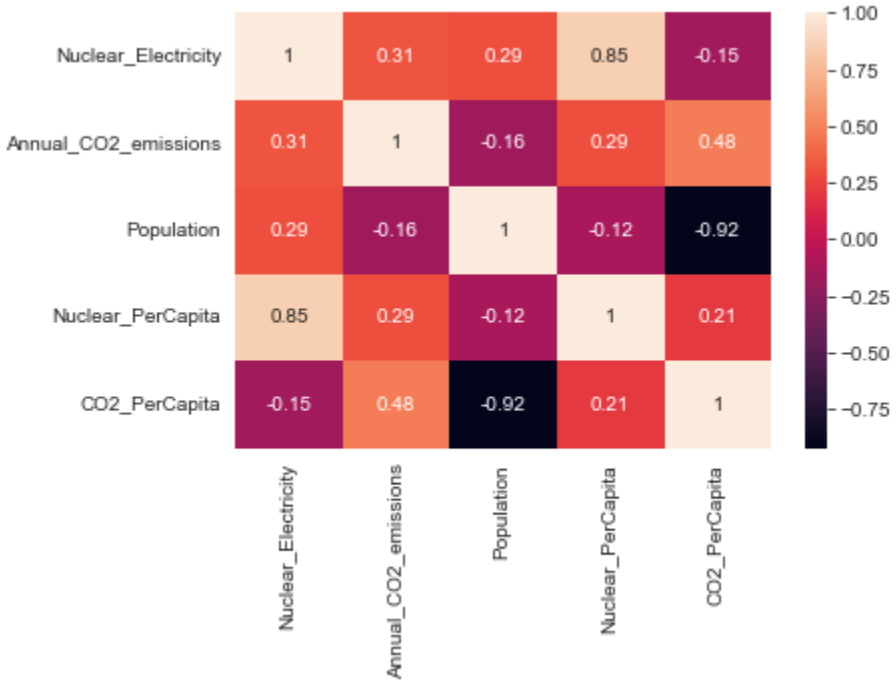
Para llevar a cabo este análisis, dos datasets fueron adicionados, uno donde se tiene la cantidad de población por países y una que divide los países por continente.

Las preguntas a responder para este análisis fueron:

- Existe una relación entre las emisiones anuales de CO2 y la población de los países.

- Existe una relación entre las emisiones anuales de CO2 y generación de energía nuclear de los países.

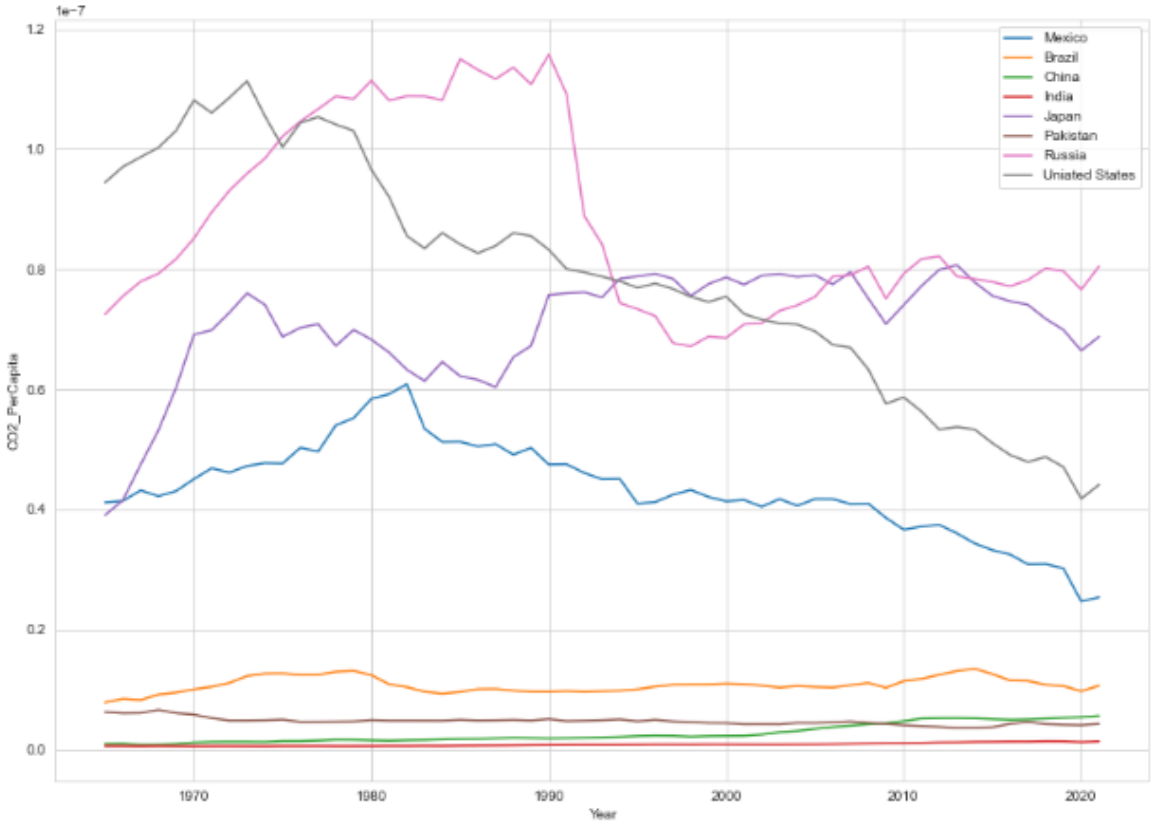
Se realizó un heatmap para observar la correlación entre los datos.




De lo anterior se encontró lo siguiente:

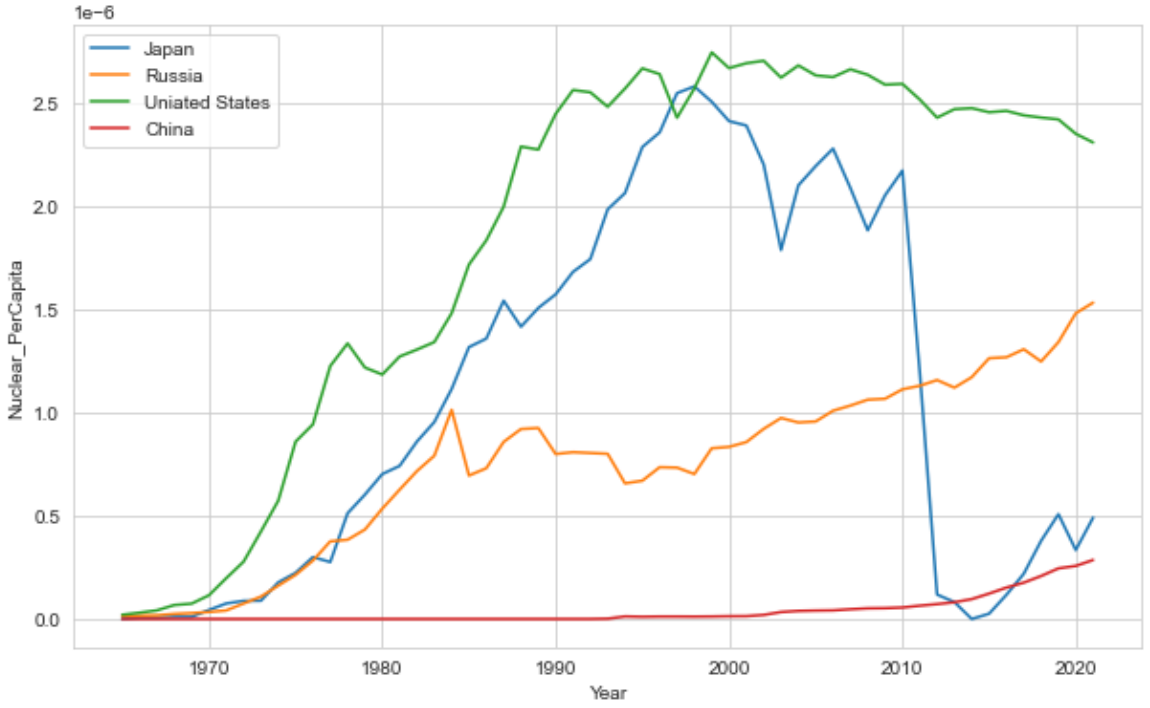
- Al realizar una correlación de los datos con el método spearman, se observa correlación negativa muy alta entre la población y el CO2 per cápita, lo que da a entender que el crecimiento poblacional y la cantidad de CO2 generado crecen en sentidos contrario, siendo lo contrario a lo esperado en la hipótesis nula 1.
- Dicha oposición a la hipótesis nula (1) se confirma con la correlación observada entre la población y las emisiones de CO2 anuales.
- De igual forma, los datos de correlación muestran que no existe una alta correlación entre las emisiones de CO2 per cápita y la energía nuclear generada per cápita.

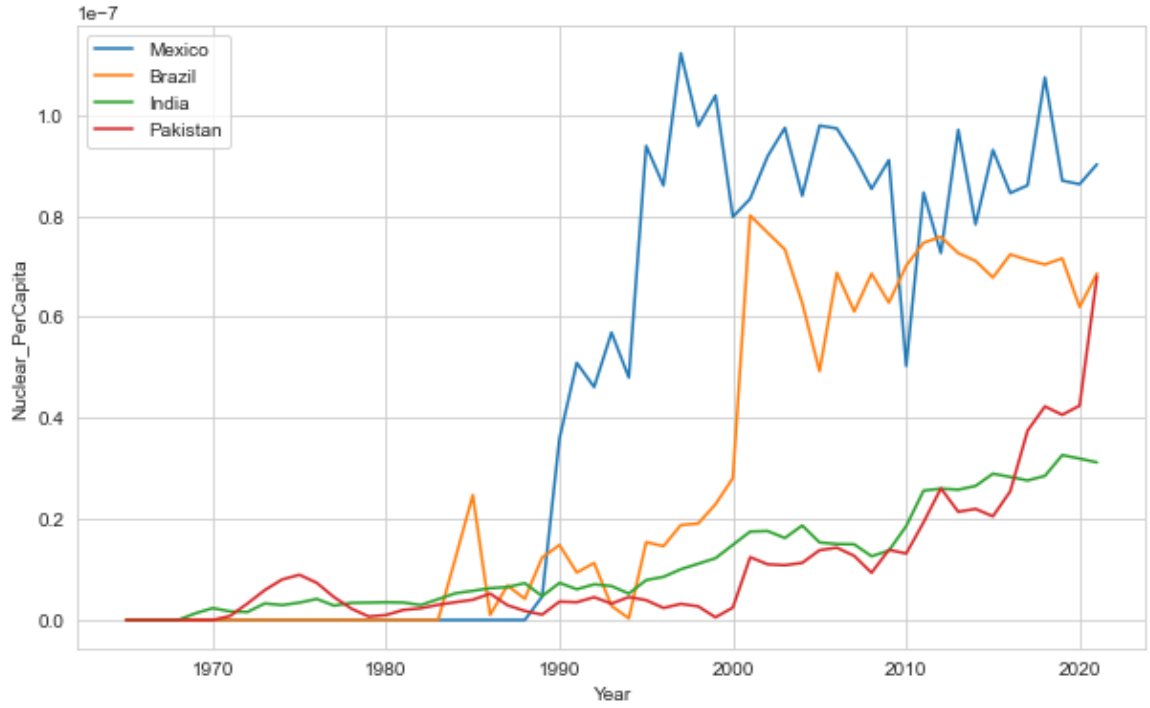
A partir de eso, se seleccionaron los países: México, Brasil, China, India, Japón, Pakistán, Rusia y Estados Unidos, esto debido a que su población total era parecida. Con los países seleccionados, se graficaron las emisiones anuales de CO2 per cápita y la generación de energía nuclear per cápita.




Hallazgos:

- Se observa que de manera general para los países observados, la tendencia de la emisión de gases per cápita se mantiene estable, confirmando que no existe una relación con el crecimiento poblacional como se establece en la hipótesis nula 1.
- La disminución de los gases emitidos per cápita para United States alrededor de 1990 se puede deber a la recesión que se sufrió durante esa época, así como se puede deber a la estanflación vivida durante 1970.
- Japón sufrió un decrecimiento de las emisiones per cápita alrededor de 1970, puede existir una relación entre la estanflación vivida por United States puesto que este es el principal aliado económico.
- La disminución de los gases per cápita generados por Rusia se puede atribuir a la disolución de la unión soviética dada en 1991, cercana al inicio de la disminución de la gráfica.




Hallazgos:

- De manera general se observa un crecimiento de la generación de energía para los países mostrados en el gráfico, siendo este continúa, mostrando pocos valles.
- Se observa una caída abrupta de la generación de energía nuclear para Japón en el 2010, esto se puede atribuir a el terremoto sufrido que a su vez ocasionó el accidente nuclear de Fukushima, lo que ocasionó un periodo de nula o poca generación de energía nuclear.

- De igual forma para los países del gráfico se observa un crecimiento en la generación de energía nuclear, siendo esta menos continua comparada con los países del gráfico anterior, esto se le puede atribuir principalmente a la falta de infraestructura para llevar a cabo dicha generación.

### 4.2.3 Entre dos numéricas (de intervalo o razón)

Para este caso, también aplica el pairplot realizado en la sección 4.2.1

## 4.3 Otros análisis exploratorios multivariados (opcional)

**Análisis ANOVA RM por continente.**

Explicación del número de continente, debido a que el mediante este método buscamos que no existan diferencias significativas en el uso de energía nuclear entre países de cada continente desde 1965 hasta 2021, debido a esto se tiene que descartar 2 continente ya que en el caso de Oceanía no existe ningún país que genere energía nuclear dentro de ese continente y el en el caso de África solo existe el caso de Sudáfrica el cual no tiene ningún otro país para poder hacer la comparación.

**ANOVA RM Europa**

Hipótesis nula:

- No hay diferencias significativas en el uso de energía nuclear entre países de Europa desde 1965 hasta 2021.

Hipótesis alternativas:

- Hay diferencias significativas en el uso de energía nuclear entre países de Europa desde 1965 hasta 2021.

Regla de aceptación si el valor de p.value es mayor al 0.05 de alpha la hipótesis nula será aceptada.

Anova

| | F Value | Num DF | Den DF | Pr > F |
|---|---|---|---|---|
| Year | 5.1356 | 56.0000 | 952.0000 | 0.0000 |

Decisión: la hipótesis nula se rechaza ya que el valor de p.value es menor al 0.05.

Conclusión:Hay diferencias significativas en el uso de energía nuclear entre países de Europa desde 1965 hasta 2021.

**ANOVA RM América**

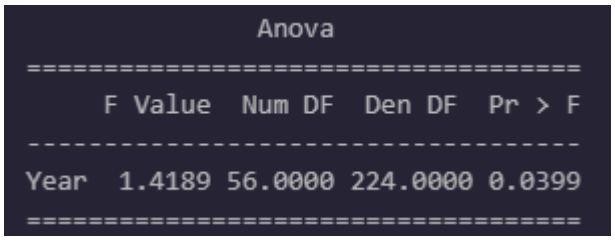
Hipótesis nula:

- No hay diferencias significativas en el uso de energía nuclear entre países de América desde 1965 hasta 2021.

Hipótesis alternativas:

- Hay diferencias significativas en el uso de energía nuclear entre países de América desde 1965 hasta 2021.

Regla de aceptación si el valor de p.value es mayor al 0.05 de alpha la hipótesis nula será aceptada.



| Anova | | | | |
|---|---|---|---|---|
| | F Value | Num DF | Den DF | Pr > F |
| Year | 1.4189 | 56.0000 | 224.0000 | 0.0399 |

Decisión: la hipótesis nula se rechaza ya que el valor de p.value es menor al 0.05.

Conclusión:Hay diferencias significativas en el uso de energía nuclear entre países de América desde 1965 hasta 2021.

**ANOVA RM Asia**

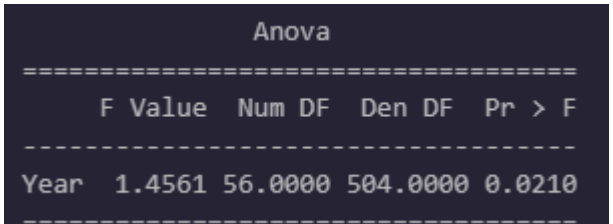
Hipótesis nula:

- No hay diferencias significativas en el uso de energía nuclear entre países de Asia desde 1965 hasta 2021.

Hipótesis alternativas:

- Hay diferencias significativas en el uso de energía nuclear entre países de Asia desde 1965 hasta 2021.

Regla de aceptación si el valor de p.value es mayor al 0.05 de alpha la hipótesis nula será aceptada.



| Anova | | | | |
|---|---|---|---|---|
| | F Value | Num DF | Den DF | Pr > F |
| Year | 1.4561 | 56.0000 | 504.0000 | 0.0210 |

Decisión: la hipótesis nula se rechaza ya que el valor de p.value es menor al 0.05.

Conclusión:Hay diferencias significativas en el uso de energía nuclear entre países de Asia desde 1965 hasta 2021.

Mixed Linear Model Regression para Energía Nuclear

```
            Mixed Linear Model Regression Results
=================================================================
Model:            MixedLM Dependent Variable: Nuclear_Electricity
No. Observations: 2052    Method:             REML
No. Groups:       5       Scale:              10706.2656
Min. group size:  57      Log-Likelihood:     inf
Max. group size:  1140    Converged:          Yes
Mean group size:  410.4
-----------------------------------------------------------------
              Coef.    Std.Err.     z      P>|z|   [0.025   0.975]
-----------------------------------------------------------------
Intercept     -0.000
Year           1.481      0.139   10.664   0.000    1.208    1.753
Group Var      0.000
=================================================================
```

*Conclusiones a partir de los datos otorgados por el modelo*

- El modelo indica que la producción de electricidad nuclear no presenta una intercepción significativa en el año cero, lo que podría indicar que la producción de electricidad nuclear en ese momento era cercana a cero.

- El coeficiente de año indica que, en promedio, la producción de electricidad nuclear aumentó en 1.481 unidades por año (esto es significativo, con un valor p < 0.05).

- El modelo también indica que la varianza de grupo es cero, lo que significa que no hay variación en la producción de electricidad nuclear entre los diferentes continentes.

Mixed Linear Model Regression para Emisiones de C02

```
            Mixed Linear Model Regression Results
=========================================================================
Model:            MixedLM Dependent Variable: Annual_CO2_emissions
No. Observations: 2052    Method:             REML
No. Groups:       5       Scale:              39.3611
Min. group size:  57      Log-Likelihood:     inf
Max. group size:  1140    Converged:          Yes
Mean group size:  410.4
-------------------------------------------------------------------------
               Coef.    Std.Err.     z      P>|z|   [0.025   0.975]
-------------------------------------------------------------------------
Intercept       0.000
Year           -0.018       0.008   -2.089   0.037   -0.034   -0.001
Group Var       0.000
=========================================================================
```

Conclusiones a partir de los datos otorgados por el modelo

- El coeficiente del año (Year) es -0.018, lo que significa que, en promedio, las emisiones de dióxido de carbono disminuyen en 0.018 toneladas por año.

- El modelo también indica que la varianza de grupo es cero, o que indica que no hay diferencias significativas en las emisiones de dióxido de carbono entre los cinco continentes

## 4.4 Conclusiones del capítulo

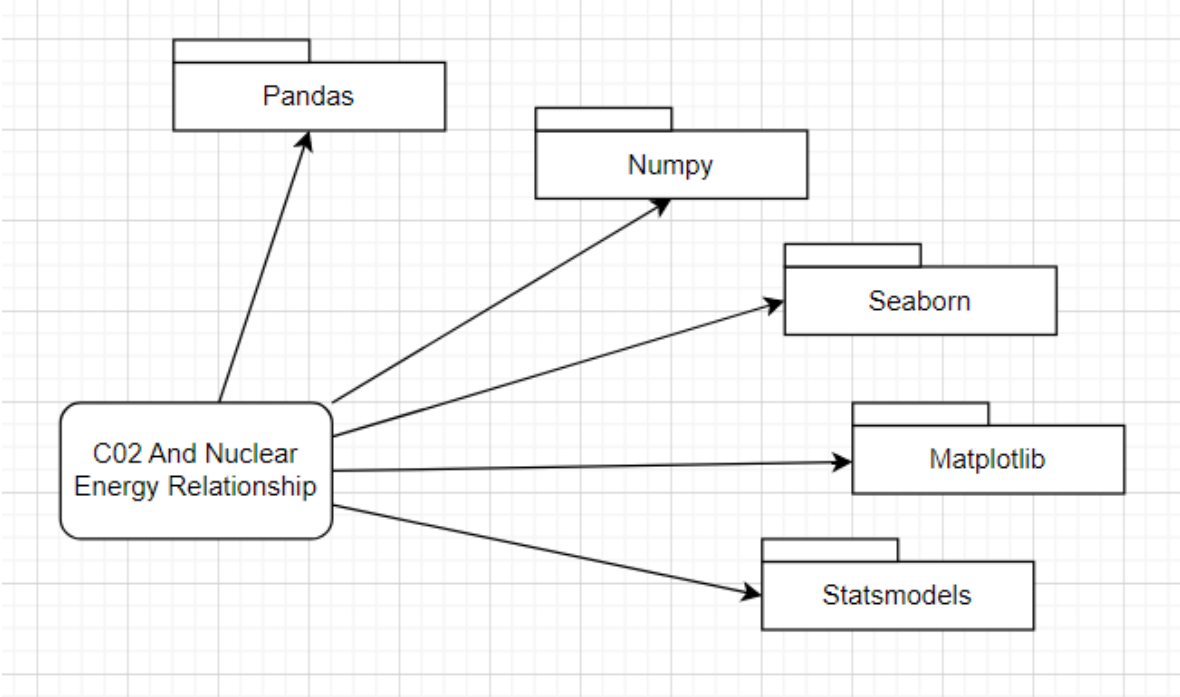
El análisis bivariado y multivariado son técnicas estadísticas que permiten explorar la relación entre dos o más variables en un conjunto de datos. Al aplicar estas técnicas en este trabajo, se pueden identificar patrones y tendencias en los datos, y tomar decisiones informadas para reducir las emisiones y aumentar la generación de energía limpia.

El análisis bivariado permite identificar la relación entre dos variables, mientras que el análisis multivariado permite identificar la relación entre tres o más variables, lo que facilita la identificación de relaciones causales y patrones complejos en los datos. En general, estas técnicas son herramientas valiosas para analizar datos complejos y tomar decisiones informadas en diversas áreas de estudio.

# 5 Implementación

## 5.1 Diagramas de paquetes de UML

**Figura 6-1 Diagrama de paquetes UML**

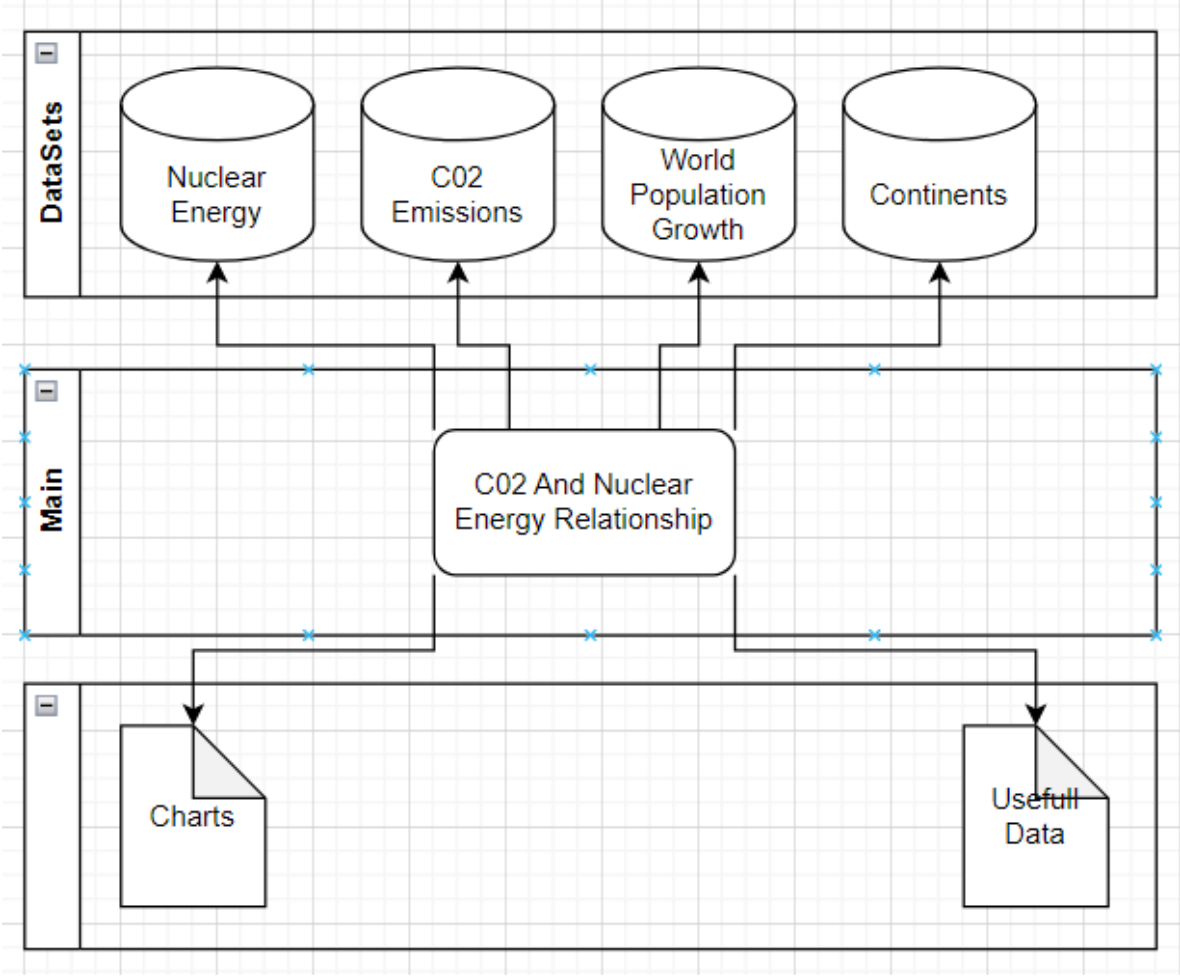



*Fuente: elaboración propia*

- Statsmodels:
  - mixedlm
  - AnovaRM
- Pandas
  - Manejo del Dataframe
- Numpy
  - Métodos para manejo de información, como media, min y max
- Seaborn
  - boxplot
  - listplot
- Matplotlib
  - Gráficas junto con seaborn

## 5.2 Diagrama de flujo de notebooks

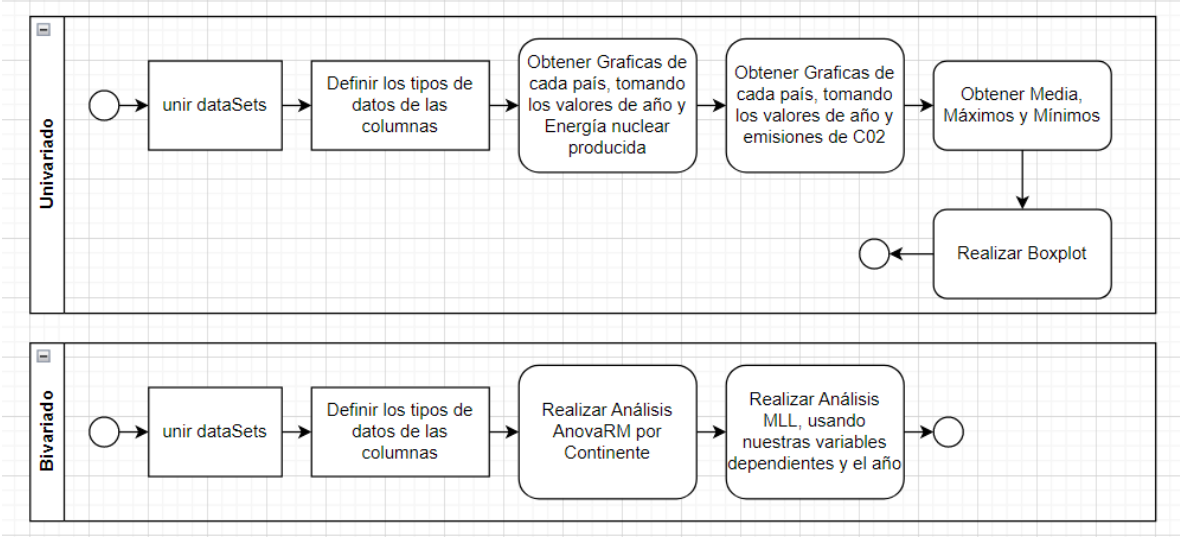
**Figura 6.2 Diagrama de entradas y salidas de un notebook**




*Fuente: elaboración propia*

**Figura *6-3***

**Figura 6.3 Diagrama de flujo de notebooks usando BPMN**



## 5.3 Conclusiones del capítulo

Mediante el uso de herramientas como las librerías Pandas y Seaborn, podemos mejorar el alcance y lograr un desarrollo de proyecto de mayor calidad y complejidad en el análisis de datos. Estas herramientas nos permiten manipular y analizar datos de manera eficiente, visualizarlos de forma clara y concisa, identificar patrones y correlaciones, y hacer predicciones precisas. Además, el uso de estas herramientas representa un reto interesante para nosotros como estudiantes, ya que es importante que comprendamos su funcionamiento y capacidades para poder aprovecharlas al máximo.

Al aprender a utilizar estas librerías, estamos adquiriendo habilidades esenciales para el análisis de datos, lo cual es crucial en una gran variedad de campos, desde la ciencia hasta los negocios. En resumen, el uso de librerías como Pandas y Seaborn es fundamental para un análisis de datos eficiente y efectivo, y es una habilidad valiosa para desarrollar como estudiantes.

# 6 Conclusiones

Como programadores, el análisis exploratorio de datos (EDA) es una técnica valiosa, especialmente en los campos de la ciencia de datos y el aprendizaje automático. Nos permite comprender mejor los datos con los que trabajamos y descubrir patrones y relaciones útiles para desarrollar modelos y algoritmos efectivos.

El EDA nos ayuda a limpiar los datos, identificando problemas como valores atípicos, datos faltantes y errores de entrada. Además, nos permite optimizar el rendimiento de nuestros modelos y algoritmos, identificando las características más importantes para ellos.

La visualización de datos es una parte clave del EDA, ya que nos permite comunicar patrones complejos de manera clara y accesible. Al ser curiosos y flexibles en nuestro enfoque, podemos descubrir patrones y relaciones que podrían ser de gran ayuda para el éxito de nuestros proyectos y nuestra carrera profesional en general.

En este trabajo se utilizó el análisis exploratorio de datos para investigar la relación entre la energía nuclear y las emisiones de dióxido de carbono en diferentes países. Gracias al EDA, pudimos descubrir patrones interesantes que podrían ayudar a informar políticas energéticas futuras. Sin duda, el análisis exploratorio de datos es una herramienta poderosa y valiosa para nuestro trabajo como programadores.

# 7 Referencias

Ritchie, H., Rosado, P., & Roser, M. (2022). *Our World In Data*. Obtenido de Nuclear Energy: https://ourworldindata.org/nuclear-energy

Ritchie, H., Rosado, P., & Roser, M. (2022). *Our World In Data*. Obtenido de CO2 emissions: https://ourworldindata.org/co2-emission

Ortiz, E., Ritchie, H., Rodes,L., & Roser, M. (2022). Our World In Data. Obtenido de World Population Growth: https://ourworldindata.org/world-population-growth

Our World In Data. (2022). Obtenido de Continents according to Our World In Data: https://ourworldindata.org/grapher/continents-according-to-our-world-in-data

Repeated Measures ANOVA - Understanding a Repeated Measures ANOVA | Laerd Statistics. (n.d.). https://statistics.laerd.com/statistical-guides/repeated-measures-anova-statistical-guide.php

Linear Mixed Effects Models - statsmodels 0.15.0 (+6). (n.d.). https://www.statsmodels.org/dev/examples/notebooks/generated/mixed_lm_example.html

GeeksforGeeks. (2022). How to Perform a Repeated Measures ANOVA in Python. GeeksforGeeks. https://www.geeksforgeeks.org/how-to-perform-a-repeated-measures-anova-in-python/

statsmodels.regression.mixed_linear_model.MixedLM - statsmodels 0.15.0 (+6). (n.d.). https://www.statsmodels.org/dev/generated/statsmodels.regression.mixed_linear_model.MixedLM.html