

# CIENCIA DE DATOS

Comision 5  
Profesor: Lucas Paz

## Final - Grupo 10

- 1- Lourdes Y. Videla Pérez - DNI: 36.129.650 - lourdesvidelaperez@gmail.com
- 2- Rodrigo I. Reyes - DNI: 34880277 - rodrigocjsreyes666@gmail.com
- 3- Rosana M. Mauno - DNI: 28888374 - maunorosana374@gmail.com
- 4- Eduardo E. Mendiola - DNI: 28.878.552 - ezequielmz@gmail.com
- 5- Leonardo D. Lafflitto - DNI: 25152257 - neolelaff@gmail.com
- 6- Guillermo Citelli - DNI: 36099992 - guillermocitelli@gmail.com
- 7- Julieta Mena - DNI: 23521731 - Julietaayelenmena@gmail.com
- 8- Carlos Paladini - DNI: 28964898 - paladinicarlos@gmail.com
- 9- Pablo Miguel - DNI: 25433799 - pablogmiguel@gmail.com
- 10- Francisco Romero - DNI: 95890354 - romerofrancisco262@gmail.com
- 11- Maximiliano Villegas - DNI: 37661757 - maximilianov.fernandez@gmail.com
- 12- Mauro Arias - DNI: 36080832 - maurosebastianarias@gmail.com

*Optamos por explorar el conjunto de datos "World Population", el cual nos brinda valiosas ventajas para comprender las tendencias demográficas y socioeconómicas a nivel global. A través de este conjunto, podemos analizar cómo la población se distribuye entre países y cómo ha evolucionado con el tiempo. Además, nos permite comparar distintos países en términos de desarrollo y calidad de vida. Utilizando técnicas analíticas, podemos anticipar el crecimiento futuro de la población y tomar decisiones informadas en el ámbito económico. Creemos que este conjunto de datos nos proporciona una perspectiva valiosa sobre la población mundial y sus implicaciones.*

Importamos las bibliotecas que consideramos necesarias y cargamos la base de datos. En estos primeros registros, se representa la información (a grandes rasgos) sobre la población en un año específico, 2020. Las columnas principales nos muestran las siguientes variables:

- **Country:** Nombre del país. Tipo: categórica nominal.
- **Year:** Año correspondiente a los datos. Tipo: numérica discreta.
- **Population:** Población del país en el año especificado. Tipo: numérica discreta.
- **Yearly % Change:** Cambio porcentual anual en la población. Tipo: numérica continua.
- **Yearly Change:** Cambio anual en la población. Tipo: numérica discreta.
- **Migrants (net):** Número neto de migrantes (inmigrantes menos emigrantes) en el país. Tipo: numérica discreta.
- **Median Age:** Edad mediana de la población. Tipo: numérica discreta.
- **Fertility Rate:** Tasa de fertilidad. Tipo: numérica continua.
- **Density (P/Km<sup>2</sup>):** Densidad de población por kilómetro cuadrado. Tipo: numérica continua.
- **Urban Pop %:** Porcentaje de población urbana. Tipo: numérica continua.
- **Urban Population:** Población urbana total. Tipo: numérica discreta.
- **Country's Share of World Pop:** Participación del país en la población mundial. Tipo: numérica continua.
- **World Population:** Población mundial total en el año especificado. Tipo: numérica discreta.
- **Rank:** Clasificación del país en términos de población. Tipo: numérica discreta.

Los valores en estas columnas nos ofrecen información sobre la población, la demografía y la posición del país en relación con otros y la población mundial.

Luego obtenemos información sobre el índice de rango, las columnas, los tipos de datos, como así también la cantidad de valores no nulos en cada columna.

- El DataFrame contiene 4196 filas y 14 columnas.
- Podemos ver el nombre de cada columna junto con información sobre el recuento de valores no nulos y el tipo de datos.
- Observamos que algunas columnas tienen valores nulos, como "Migrants (net)", "Median Age" y "Fertility Rate".
- Los tipos de datos de las columnas incluyen float64 para columnas numéricas, int64 para enteros y object para columnas con valores de tipo de datos mixtos (como las columnas de porcentaje con símbolo %).

Realizamos operaciones de limpieza, a saber:

- **Eliminación de duplicados:** Eliminamos filas duplicadas basadas en la columna 'country'. Esto asegura que solo haya una entrada para cada país.
- **Reemplazo de valores 'N.A.' por NaN:** Reemplazamos los valores 'N.A.' con valores NaN (Not a Number).
- **Filtrado de datos por año:** Filtramos los datos para mantener solo las filas correspondientes al año 2020. Tomamos esta decisión para poder unificar y sistematizar de manera más coherente los datos.
- Verificamos la columna "Year", para asegurarnos que solo contiene información de ese año.

Procedemos a interpretar nuevamente los valores de nuestro DataSet.

- **Count (conteo):** Indica que hay 201 registros en total en la columna 'Year'.
- **Mean (promedio):** La media de los valores en la columna 'Year' es 2020, lo cual es de esperar ya que filtramos los datos.
- **Std (desviación estándar):** La desviación estándar es 0. Esto significa que todos los valores en la columna son iguales (2020), por lo que no hay variabilidad en esta columna.
- **Min (mínimo):** El valor mínimo en la columna 'Year' es 2020.

- **25% (percentil):** El percentil 25 es 2020. Esto significa que el 25% de los valores son iguales o menores a 2020.
- **50% (percentil):** El percentil 50 es 2020, lo cual es la mediana. Esto indica que la mitad de los valores son iguales o menores a 2020.
- **75% (percentil):** El percentil 75 es 2020. Esto significa que el 75% de los valores son iguales o menores a 2020.
- **Max (máximo):** El valor máximo en la columna 'Year' es 2020.

Mostramos nuevamente una vista completa del conjunto de datos, incluyendo detalles como la cantidad de filas y columnas presentes, los tipos de datos almacenados en cada columna, y la cantidad de valores que no son nulos en cada una. Además, presentamos una tupla con el recuento de filas y columnas, ofreciendo así una perspectiva clara de las dimensiones y la estructura del conjunto de datos. La ejecución de estas dos instrucciones proporcionará información detallada acerca de la estructura y el tamaño del conjunto de datos en cuestión.

- Hay un total de 201 filas en el conjunto de datos.
- El conjunto de datos tiene 14 columnas en total.
- Cada columna está etiquetada con un nombre específico, como 'country', 'Year', 'Population', etc.
- 'Non-Null Count' muestra la cantidad de valores no nulos presentes en cada columna.
- 'Dtype' indica el tipo de datos almacenados en cada columna, como 'object' para texto, 'int64' para enteros de 64 bits y 'float64' para números de punto flotante.

La última línea muestra la cantidad total de memoria utilizada por el conjunto de datos.

La tupla (201, 14) indica que hay 201 filas y 14 columnas en el conjunto de datos. Esta información es valiosa para comprender la estructura y la composición del conjunto de datos.

Rehacemos el índice para que quede desde 0 hasta 234 luego de haber eliminado las filas duplicadas. Además, hemos asignado un nuevo nombre al DataFrame con el que trabajaremos.

Contamos la cantidad de valores nulos en cada columna. Podemos ver la cantidad de valores nulos en cada columna, en 'Urban Pop %' hay 7 y en 'Urban Population' también tiene 7.

Vamos a escoger las columnas que aportan datos al análisis, suprimiendo aquellas que no se consideran importantes:

- **Year:** Al escoger solo el año 2020, se convierte en una columna con información redundante.
- **Rank:** Al considerarla meramente informativa, se decide eliminarla del set de datos.
- **Yearly change:** Se decide solamente dejar la información en % de esta variable.
- **Urban Population:** Decimos dejar la información en % de esta variable.
- **Country's share of the world:** Al igual que rank, se decide suprimirla por considerarse meramente informativa.

Después de haber eliminado ciertas columnas, nuestro DataSet quedaría de la siguiente manera:

- **Country:** Nombre del país.
- **Population:** Población del país.
- **Yearly % Change:** Cambio porcentual anual en la población.
- **Migrants (net):** Migrantes netos (migración neta) en el país.
- **Median Age:** Edad mediana de la población.
- **Fertility Rate:** Tasa de fertilidad.
- **Density (P/Km<sup>2</sup>):** Densidad de población por kilómetro cuadrado.
- **Urban Pop %:** Porcentaje de población urbana.

Este recorte de datos nos proporciona una visión más centrada en las columnas que conservamos en el conjunto de datos “poblacion\_mundial”.

Para garantizar la consistencia de los valores, realizamos algunas transformaciones:

- Utilizamos el método `.str.replace()` para eliminar los signos de porcentaje ('%') de las columnas 'Urban Pop %' y 'Yearly % Change'. Esto es importante para que los valores puedan ser tratados como números en lugar de cadenas de caracteres.
- Utilizamos el método `.astype(float)` para convertir los valores en las columnas 'Urban Pop %' y 'Yearly % Change' de tipo objeto (cadenas de caracteres) a tipo flotante. Esto convierte los valores en números decimales, lo que es necesario para realizar cálculos numéricos.
- Utilizamos `poblacion_mundial.info()` para mostrar la información actualizada sobre el conjunto de datos después de las transformaciones realizadas. Esas operaciones nos permiten transformar las columnas en formatos adecuados para el análisis numérico.

A continuación, nos encontramos en una posición para iniciar el análisis de la siguiente manera:

- 201 filas en el conjunto de datos.
- Las columnas 'Population', 'Yearly % Change', 'Migrants (net)', 'Median Age', 'Fertility Rate', 'Density (P/Km<sup>2</sup>)' y 'Urban Pop %' no tienen valores nulos y contienen datos numéricos.
- La columna 'Urban Pop %' tiene 7 valores nulos.
- La columna 'country' es de tipo objeto (cadena de caracteres).
- El conjunto de datos utiliza aproximadamente 14.1 KB de memoria.

Para resolver el problema de valores nulos, procedemos a reemplazar los valores faltantes en la columna 'Urban Pop %' por la mediana de esa misma columna. Posteriormente, verificamos nuevamente la cantidad de valores nulos en cada columna después de realizar el reemplazo. Esta corrección elimina los valores nulos de cualquier columna en el conjunto de datos.

Finalmente, realizamos cálculos de estadísticas descriptivas, redondeando los valores a dos decimales. Esto nos proporciona información sobre medidas como el valor medio (mean), la desviación estándar (std), el valor mínimo (min), los percentiles 25%, 50% (mediana) y 75%, y el valor máximo (max) para cada columna numérica.

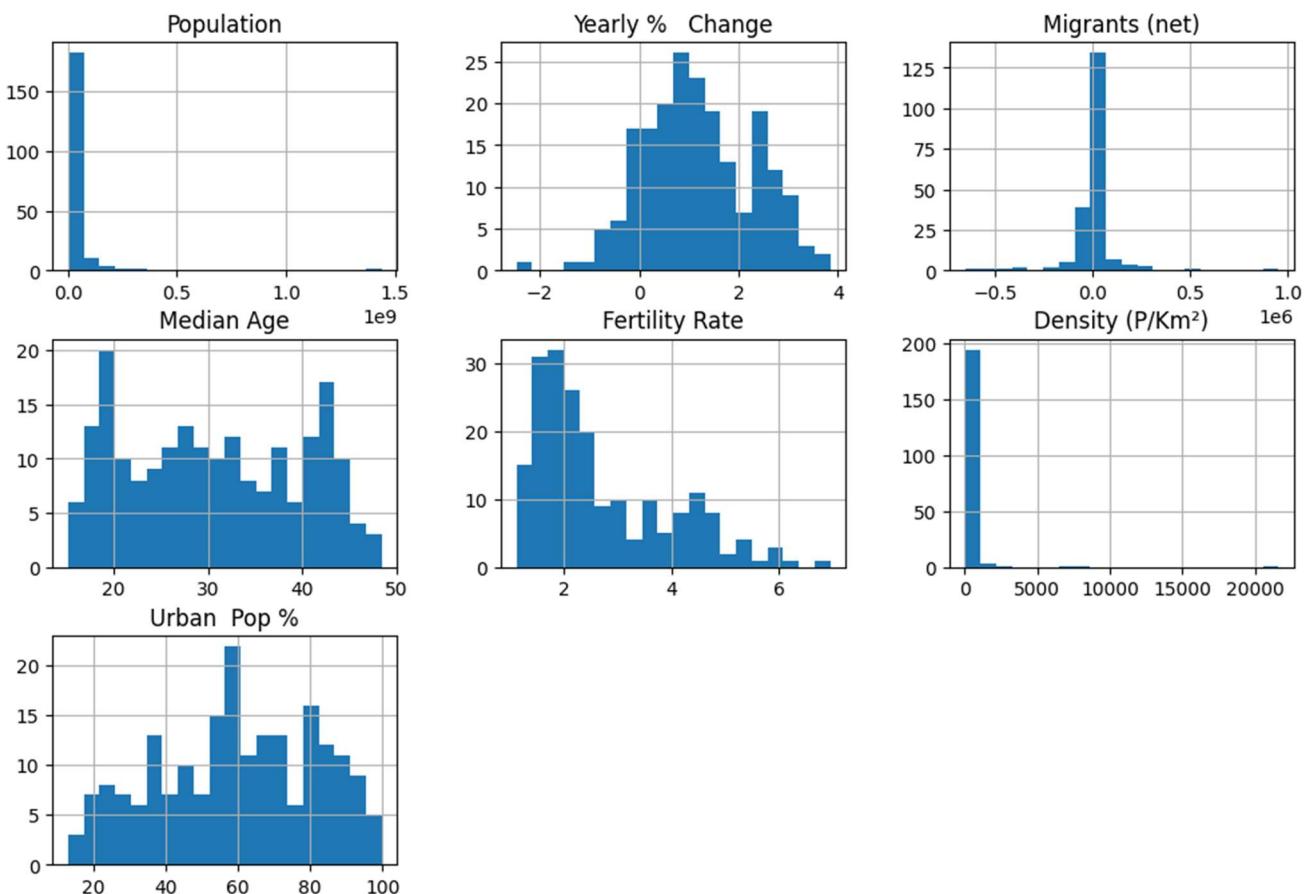
El resumen estadístico para las variables quedaría así:

- **Population:** El tamaño de la población varía significativamente, con una media de alrededor de 38.776.610 personas. La desviación estándar es bastante alta, indicando una gran variabilidad en las poblaciones de diferentes países. El rango va desde casi 98.930 a más de 1.439.324.000.
- **Yearly % Change:** La tasa de cambio anual de la población tiene una media del 1,20%, pero con una desviación estándar de 1,09%, lo que sugiere cierta variabilidad en las tasas de crecimiento entre países.
- **Migrants (net):** El promedio de migrantes netos es -5,44, lo que indica que en promedio hay más emigración que inmigración. Sin embargo, la desviación estándar es bastante alta, lo que indica que hay países con flujos migratorios significativos.
- **Median Age:** La mediana de la edad mediana es de 29,90 años, lo que sugiere una distribución relativamente joven. El rango va desde 15,20 a 48,40 años.

- **Fertility Rate:** La tasa de fertilidad tiene una media de 2,69, lo que sugiere que en promedio, las mujeres tienen alrededor de 2-3 hijos. La desviación estándar indica cierta variabilidad en la tasa de fertilidad entre países.
- **Density (P/Km<sup>2</sup>):** La densidad poblacional tiene una media de 358,67 personas por kilómetro cuadrado, pero con una desviación estándar alta de 1710,11. Esto indica variabilidad en la distribución de la población entre países.
- **Urban Pop %:** El porcentaje promedio de población urbana es del 59,68%, pero con una desviación estándar de 21,83. Esto sugiere que la proporción de población urbana varía significativamente entre países.

Población Mundial - Gráfico 1

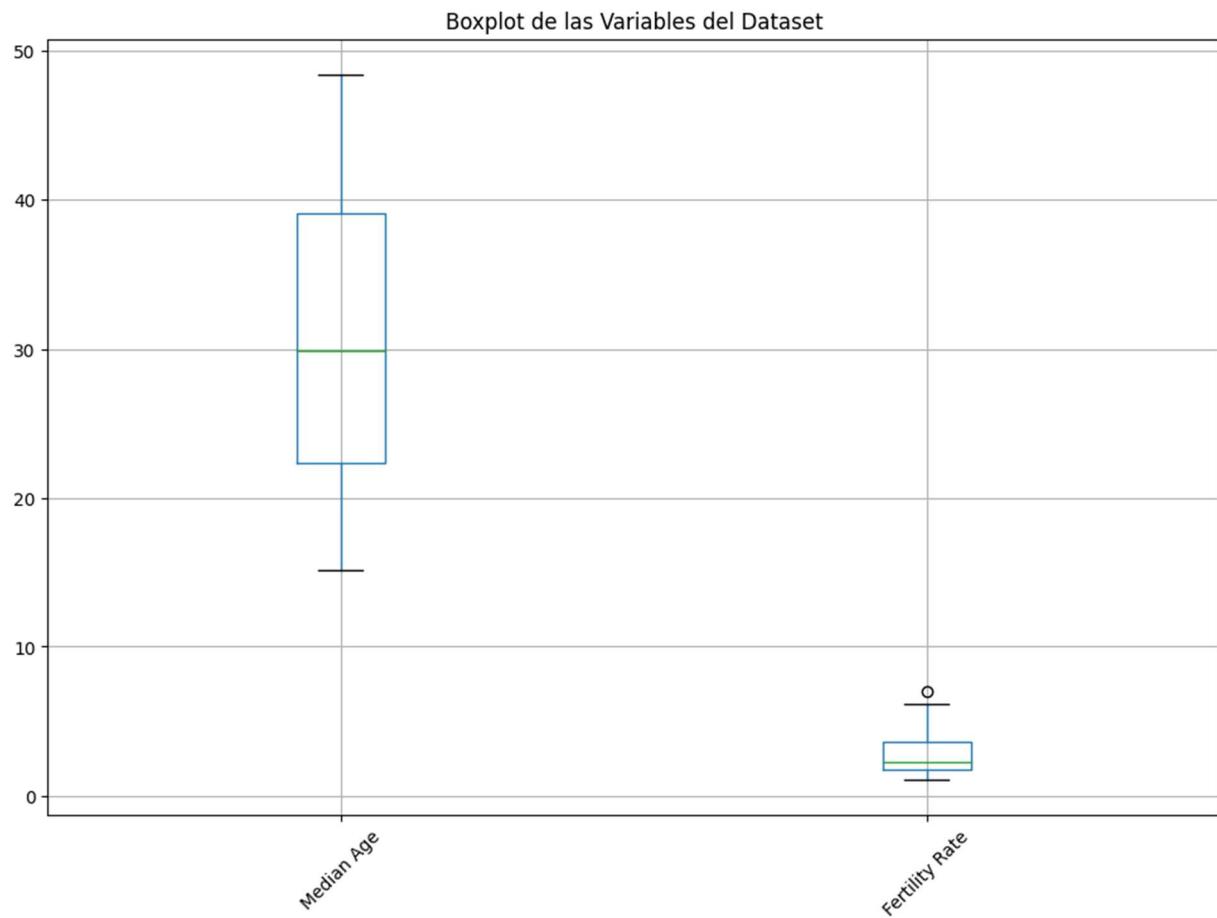
Histogramas de las Variables



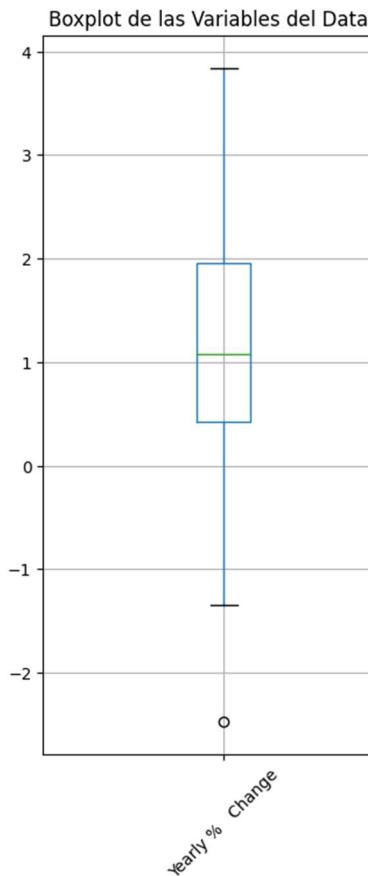
## BOXPLOT

Creamos una serie de boxplot para observar la distribución de distintas variables. Vamos a obtener de esta manera representaciones gráficas, con información valiosa sobre la mediana, los cuartiles y posibles valores atípicos. Al final expondremos algunas conclusiones.

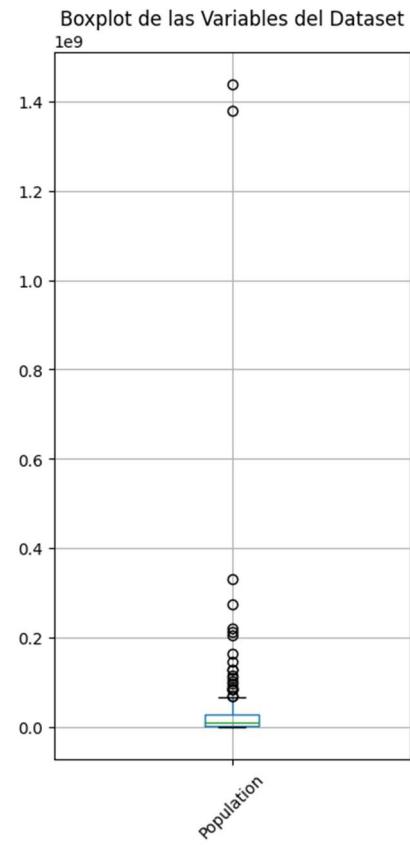
Población Mundial - Gráfico 2



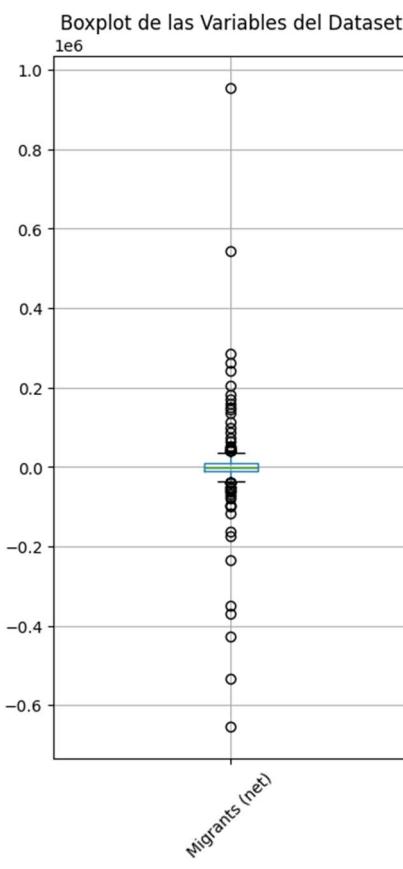
Población Mundial - Gráfico 3



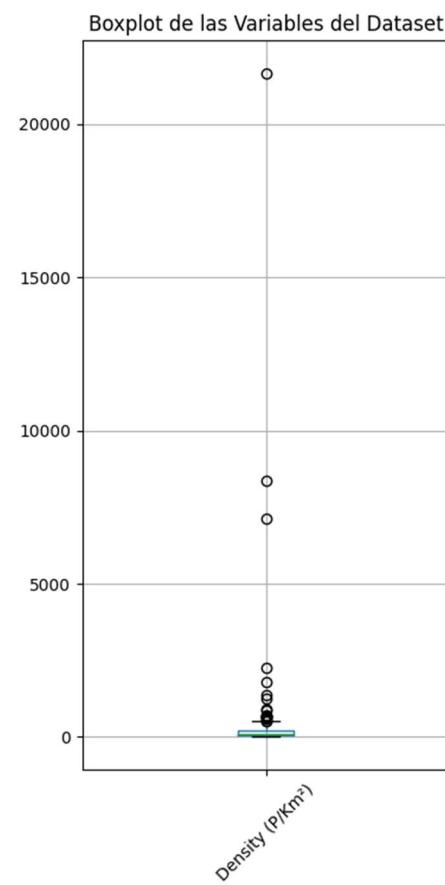
Población Mundial - Gráfico 4



Población Mundial - Gráfico 5

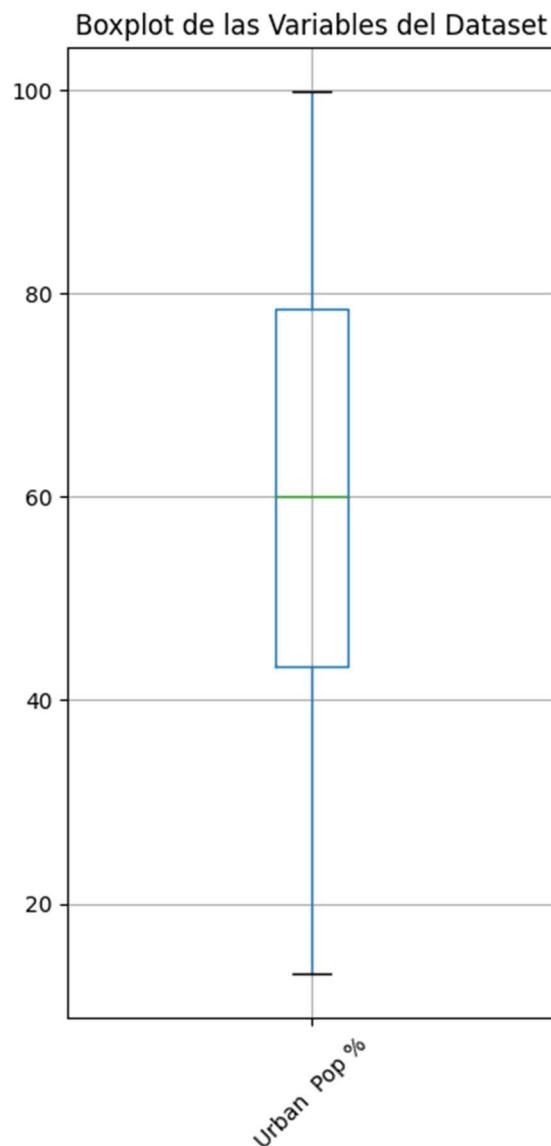


Población Mundial - Gráfico 6

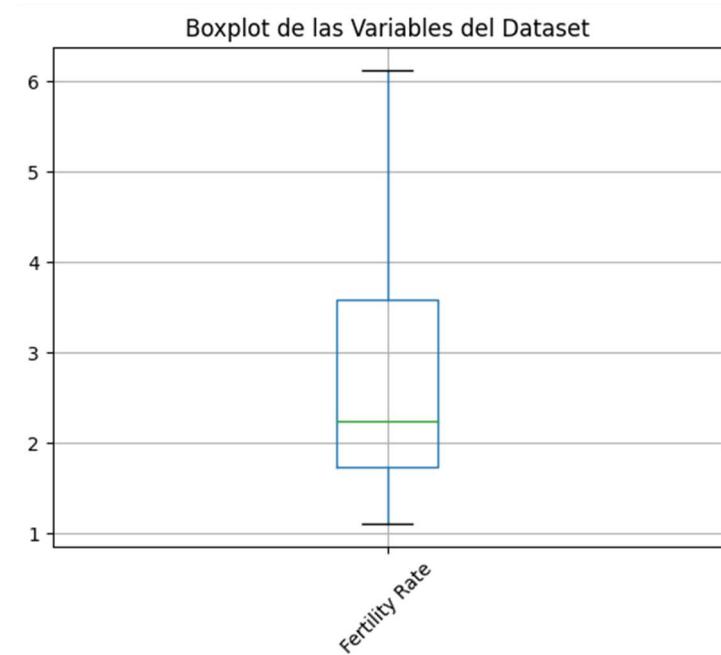


En estos gráficos podemos apreciar una marcada disparidad de datos en algunas variables clave, como la población mundial, la densidad por metro cuadrado y el número de migrantes. No obstante, cualquier dato que pudiera estar visualmente apartado de los boxplots de estas variables **no será considerado como un outliers**, dado que esta disparidad se origina de la naturaleza intrínseca de cada país, la cual puede explicar estas notables diferencias.

En contraste, en el caso de las variables como el porcentaje de población urbana, la tasa de cambio anual y la tasa de fertilidad, se observa un comportamiento más uniforme. En estas variables, la mayoría de los datos se encuentran dentro de los límites de los boxplots.



Con el objetivo de mantener la integridad del análisis, hemos decidido llevar a cabo la eliminación de los valores atípicos de la variable "Tasa de Fertilidad" utilizando el método del Rango Intercuartílico (IQR). Esto se debe a que dicha variable generalmente muestra uniformidad entre la población mundial y cualquier valor atípico en el conjunto de datos podría introducir ruido y distorsionar los resultados del análisis.



Eliminamos de los valores atípicos de la variable 'Fertility Rate' (Tasa de Fertilidad) y verificamos la distribución de la variable.

- Calculamos el primer cuartil (Q1) de la variable utilizando el percentil 25.
- Calculamos el tercer cuartil (Q3) de la variable utilizando el percentil 75.
- Calculamos el Rango Intercuartílico (IQR).
- Filtramos el DataFrame 'poblacion\_mundial' para incluir solo las filas donde la variable 'Fertility Rate' se encuentra dentro del rango definido por  $Q1 - 1.5 * IQR$  y  $Q3 + 1.5 * IQR$ .  
Esto elimina los valores atípicos de la variable.
- Creamos un nuevo boxplot solo para la variable 'Fertility Rate' después de la eliminación de valores atípicos.

## CORRELACION DE PEARSON Y COVARIANZA

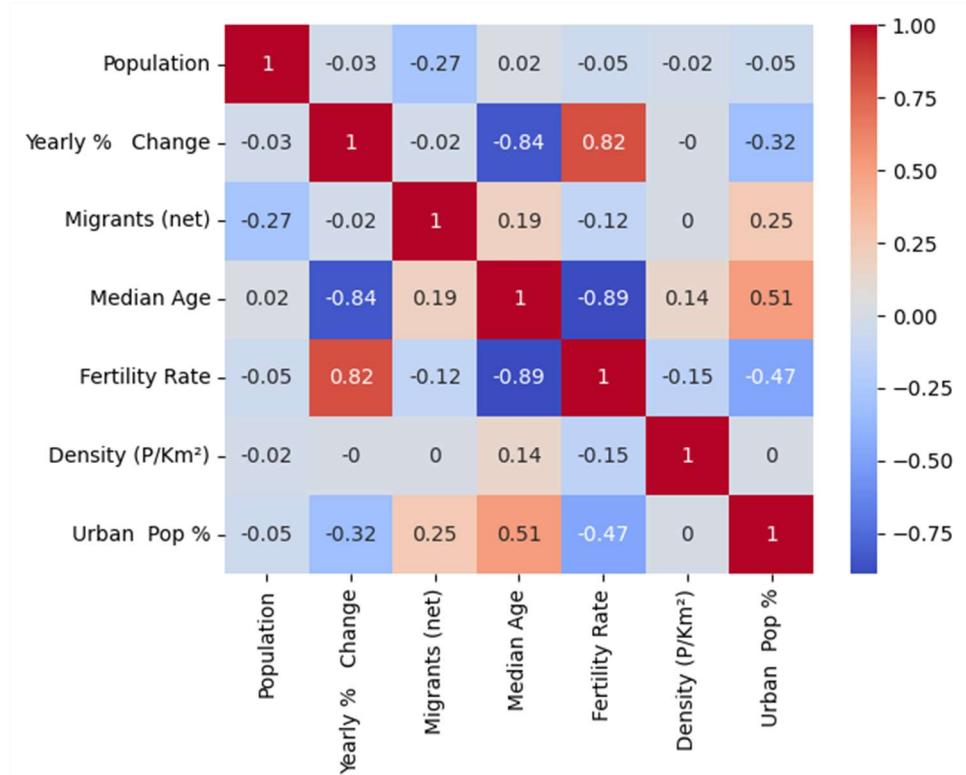
Vamos a hacer dos tareas esenciales: la generación de las matrices de correlación de Pearson y la construcción de la matriz de covarianza específicamente enfocadas en las variables numéricas:

- **Matriz de Correlación de Pearson:** En este proceso, calcularemos las correlaciones entre todas las variables presentes en el DataFrame 'poblacion\_mundial'. Para ello, empleamos la función 'corr()', que determina las relaciones de correlación entre las diversas columnas numéricas del DataFrame. Luego, aplicamos 'np.round()' para redondear los valores de correlación a dos decimales.

- **Matriz de Covarianza:** En esta fase, realizamos una selección rigurosa de las columnas numéricas del DataFrame 'poblacion\_mundial' utilizando 'select\_dtypes(include=[np.number])'. Este paso nos garantiza que solo las columnas numéricas serán consideradas en el cálculo de la matriz de covarianza. Posteriormente, empleamos 'np.cov()' para efectuar el cálculo de la matriz de covarianza. Con el parámetro 'rowvar=False', indicamos que cada columna representa una variable y que cada fila constituye una observación. Finalmente, 'np.round()' se aplica para redondear los valores de la matriz de covarianza a dos decimales.

Estas dos matrices, se convierten en herramientas sumamente valiosas para desentrañar las relaciones inherentes entre las variables numéricas dentro del conjunto de datos. La matriz de correlación proporciona una visión sobre la intensidad y dirección de las relaciones lineales entre las variables, mientras que la matriz de covarianza ofrece perspectivas sobre la variabilidad conjunta de estas variables.

Población Mundial - Gráfico 9



Generamos con Seaborn, un mapa de calor para visualizar los datos de correlación. Utilizamos los parámetros:

- **annot=True:** Nos muestra los valores de correlación en cada celda del mapa de calor.

- **cmap='coolwarm':** Define la paleta de colores a utilizar en el mapa de calor. En este caso, se utiliza una paleta que va desde tonos fríos (azules) hasta tonos cálidos (rojos).

Se pueden observar ciertas correlaciones notables en el conjunto de datos. Por un lado, se aprecia una correlación lineal positiva entre la tasa de fertilidad (Fertility Rate) y la variación anual (Yearly % Change). Asimismo, se detecta una correlación lineal negativa entre la fertilidad y la edad mediana (Median Age); esto sugiere que en poblaciones más jóvenes existe una tendencia a tener una mayor tasa de fertilidad. También se observa una correlación negativa entre la variación anual y la edad mediana, lo que indica que en países con una población más joven, no solo hay una mayor fertilidad, sino también un incremento más pronunciado en la población en general.

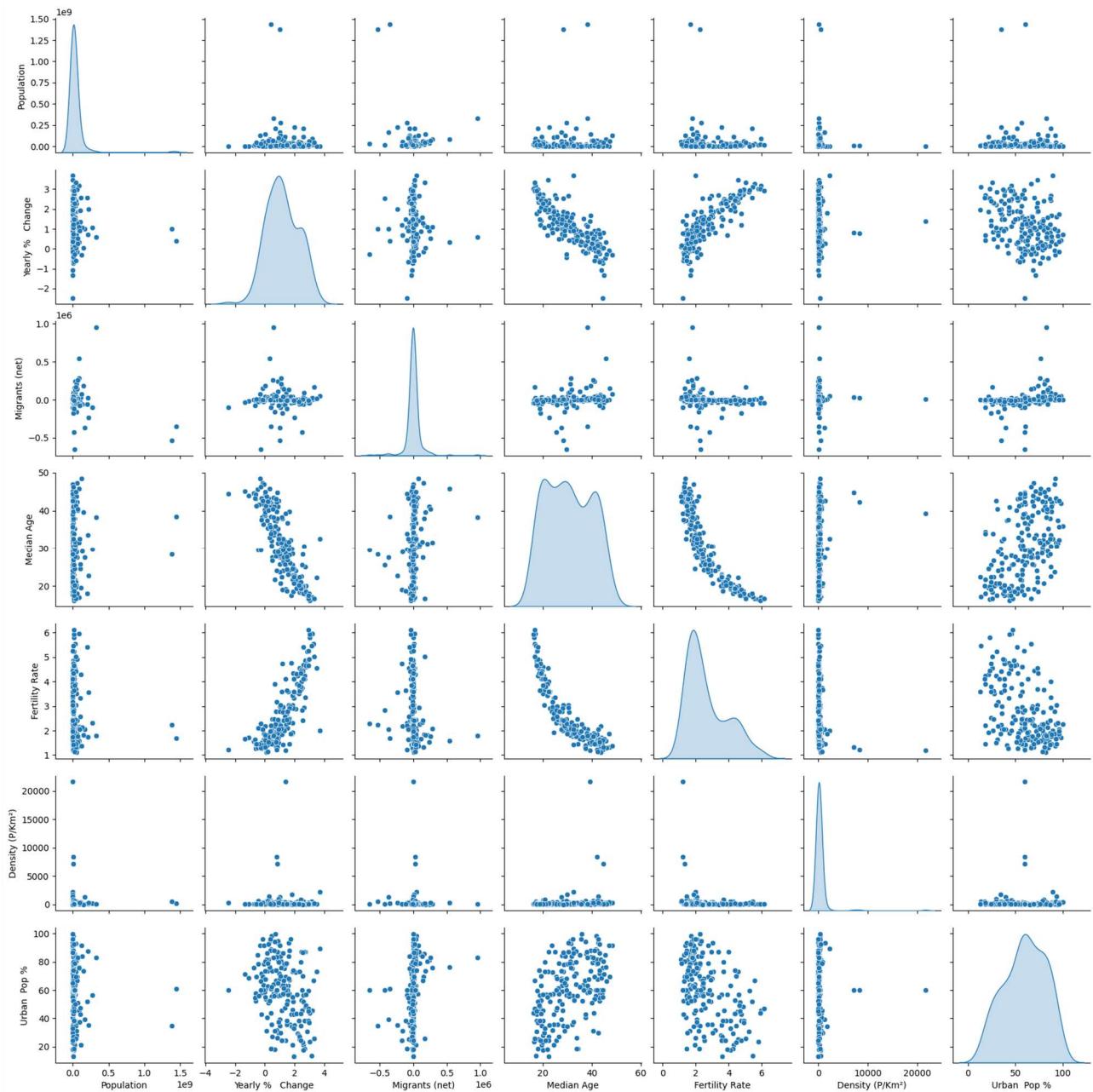
Además, se destaca una correlación mínima y positiva entre el porcentaje de población urbana (% Urban Pop) y la edad mediana. Esta asociación sugiere que en países con una población más envejecida, es más probable que haya un mayor porcentaje de población viviendo en áreas urbanas.

Sin embargo, el análisis revela que la relación entre la fertilidad y la edad de la población no sigue una tendencia lineal, sino más bien una tendencia exponencial. Esta relación es más compleja de lo que una correlación lineal podría representar.

Para visualizar estas relaciones con mayor claridad, el siguiente gráfico muestra de manera evidente cómo la relación entre fertilidad y edad de la población se aleja de la linealidad, sugiriendo una tendencia exponencial.

## SCATTERPLOTS

Generamos ahora una matriz de scatterplots utilizando la función pairplot de Seaborn para analizar las relaciones entre las características numéricas. Los gráficos de dispersión (scatterplots) muestran cómo se relacionan dos variables entre sí, y en este caso, la matriz nos permite observar múltiples relaciones de una vez. El parámetro diag\_kind='kde' indica que se mostrarán gráficos de densidad en las diagonales en lugar de histogramas.



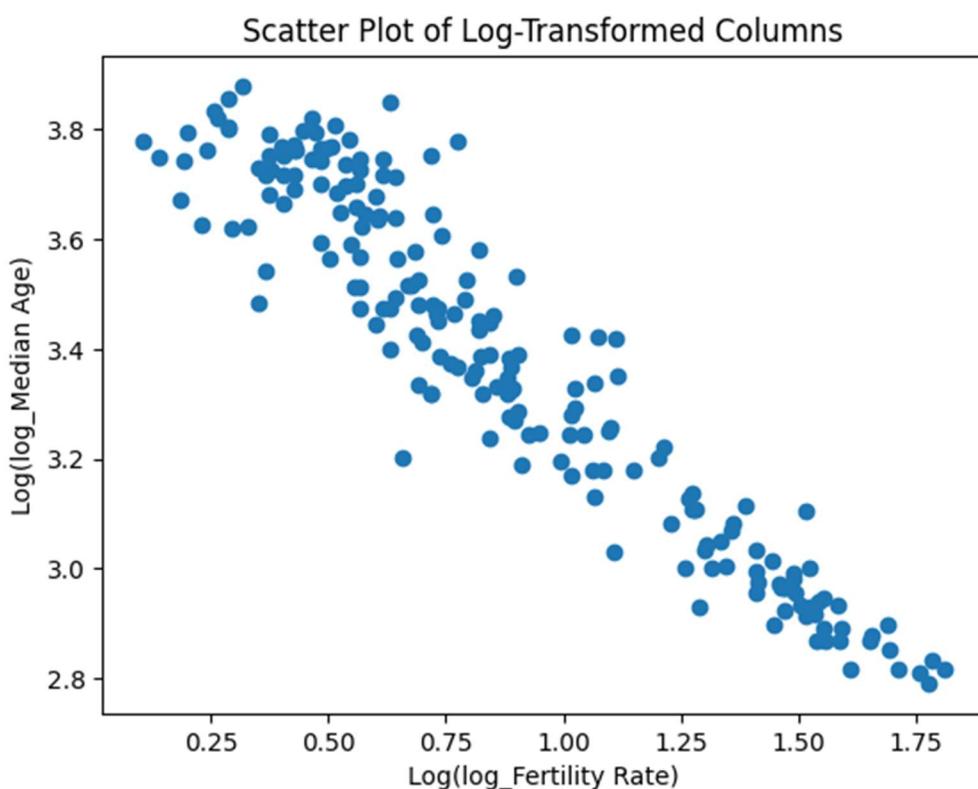
Para poder hacer un acercamiento más profundo vamos a realizar una transformación logarítmica en las columnas donde hay una correlación aparentemente exponencial, luego calculamos la correlación entre estas variables transformadas y generamos un scatter plot para visualizar. Para este procedimiento:

- Creamos una copia del DataFrame "poblacion\_mundial" llamada "log\_poblacion\_mundial".
- Realizamos las transformaciones logarítmicas en las columnas "Fertility Rate" y "Median Age" utilizando la función `np.log()`. Los resultados se almacenan en las columnas "log\_Fertility Rate" y "log\_Median Age", respectivamente.

- Calculamos la matriz de correlación para el DataFrame "log\_poblacion\_mundial" utilizando el método `.corr()`.
- Extraemos la correlación entre las variables transformadas "log\_Fertility Rate" y "log\_Median Age" de la matriz de correlación.
- Creamos un scatter plot utilizando los valores transformados. El eje x representa "Log(log\_Fertility Rate)" y el eje y representa "Log(log\_Median Age)".

El resultado es un scatter plot que visualiza la relación entre las variables transformadas utilizando una escala logarítmica. La matriz de correlación también se imprime para analizar las correlaciones entre las variables originales y transformadas.

Población Mundial - Gráfico 11



En esta matriz se pueden observar las correlaciones entre diferentes variables en el conjunto de datos.

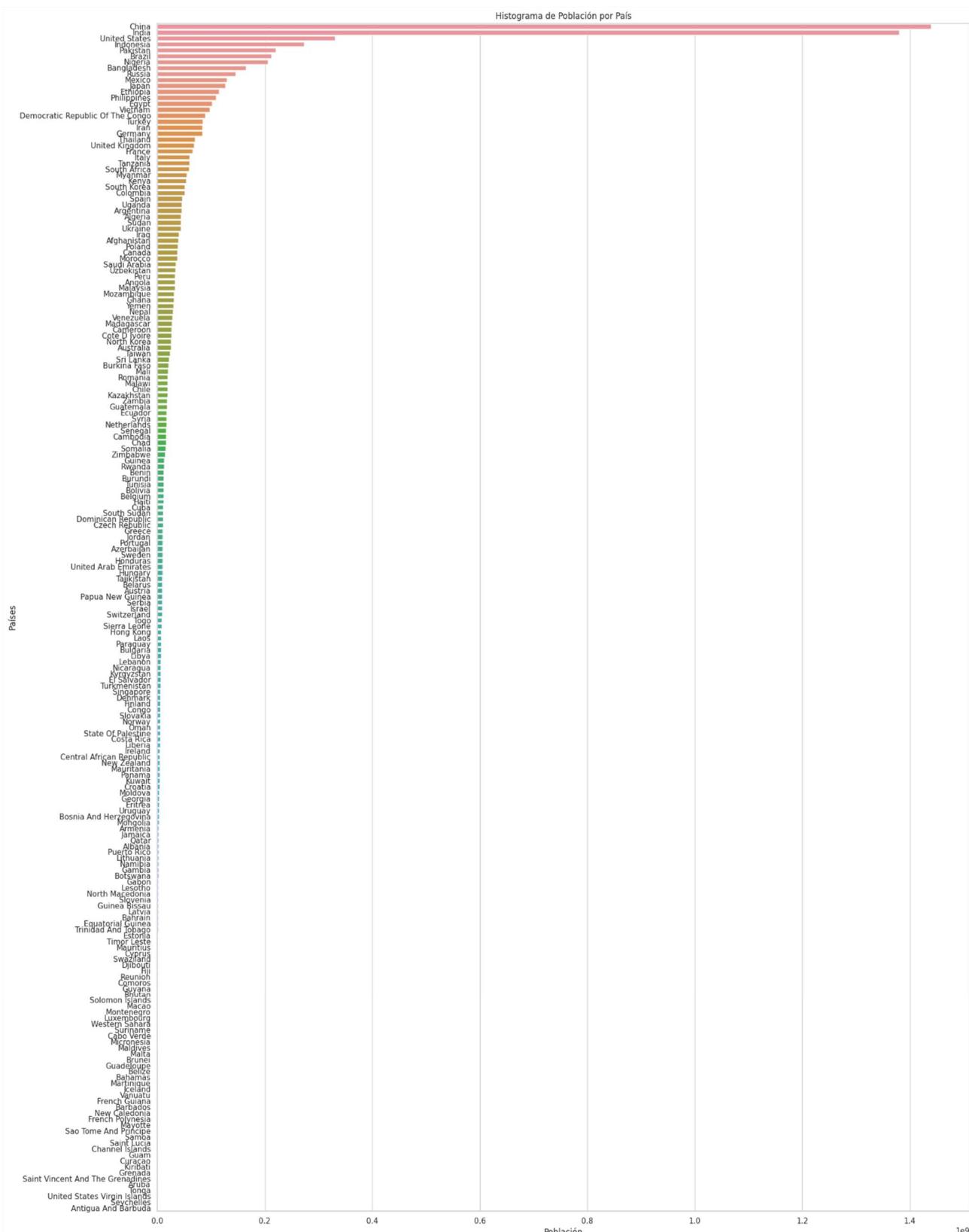
- **Population vs. Yearly % Change:** Existe una correlación negativa muy pequeña entre la población y la variación anual, lo que sugiere que en general no hay una relación fuerte entre estas dos variables.

- **Fertility Rate vs. Yearly % Change:** Hay una correlación positiva significativa entre la tasa de fertilidad y la variación anual. Esto implica que en países con una mayor tasa de fertilidad, también tiende a haber una mayor variación anual en la población.
- **Fertility Rate vs. Median Age:** Se observa una correlación negativa alta entre la tasa de fertilidad y la edad mediana. Esto indica que en países con poblaciones más jóvenes, tiende a haber una mayor tasa de fertilidad.
- **Yearly % Change vs. Median Age:** Existe una correlación negativa alta entre la variación anual y la edad mediana. Esto sugiere que en países con poblaciones más jóvenes, la tasa de variación anual tiende a ser más alta.
- **Urban Pop % vs. Median Age:** Hay una correlación positiva alta entre el porcentaje de población urbana y la edad mediana. Esto significa que en países con una población más envejecida, tiende a haber un mayor porcentaje de población viviendo en áreas urbanas.
- **Fertility Rate vs. log\_Fertility Rate:** La correlación lineal entre los datos transformados a escala logarítmica es muy alta e inversa, con un valor de -0.948617. Una correlación de -0.948617 indica que estas dos variables tienen una fuerte relación inversa, lo que significa que a medida que una de ellas aumenta, la otra tiende a disminuir de manera proporcional.
- **Median Age vs. log\_Median Age:** Hay una correlación positiva muy alta entre la edad mediana y su versión transformada en escala logarítmica. Esto indica que la relación entre estas dos variables también se mantiene en la transformación logarítmica.

En resumen, estas correlaciones destacan patrones interesantes en tu conjunto de datos, como la relación entre la tasa de fertilidad, la edad mediana y la variación anual. También es evidente cómo las transformaciones logarítmicas preservan las relaciones originales entre las variables.

## HISTOGRAMA DE POBLACION

Para mostrar la distribución de la población de cada país en el DataSet, generamos la biblioteca Seaborn para crear un histograma.



Esta representación visual nos permite comprender las diferencias y similitudes en términos de tamaño de población entre distintos países.

El eje horizontal representa las poblaciones de los países, mientras que el eje vertical muestra los nombres de los países. Cada barra vertical en el histograma corresponde a un país y su altura está proporcionalmente relacionada con su población. Es decir, cuanta más alta sea la barra, mayor será la población del país correspondiente.

Este tipo de gráfico nos brinda una visión de cómo se distribuyen las diferentes "especies demográficas" en términos de población en todo el mundo. En este contexto, cada país puede considerarse como una "especie demográfica" con su propia población única. Al analizar el histograma, obtenemos información valiosa sobre la diversidad en cuanto al tamaño de población entre distintos países.

## DIAGRAMA DE PARETO

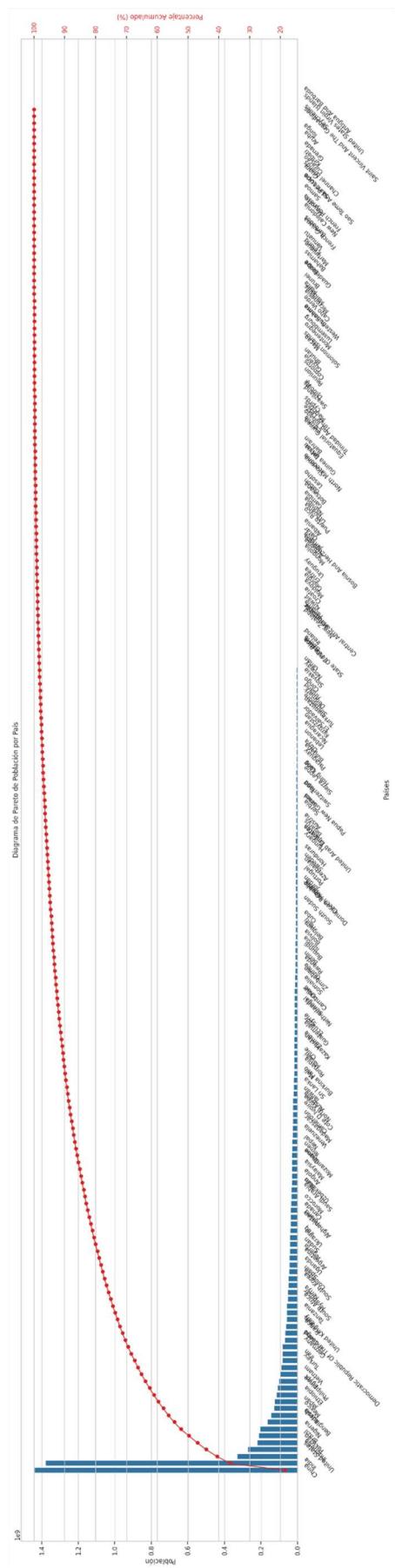
Realizaremos otro grafico para visualizar la población por país en orden descendente y como se acumula el porcentaje de la población en función de los países.

Para ello:

- Ordenamos el DataFrame "poblacion\_mundial" en orden descendente según la columna 'Population'.
- Calculamos el porcentaje acumulado de la población utilizando la función `.cumsum()` para la columna 'Population' y se almacena en una nueva columna 'Percentage'.
- Creamos un gráfico de barras utilizando Seaborn para representar la población de cada país. El eje x muestra los nombres de los países ('country') y el eje y muestra la población ('Population'). Las barras se muestran en color azul.
- Creamos una segunda serie de ejes y (`ax2`) que comparte el mismo eje x pero tiene un eje y independiente para representar el porcentaje acumulado. Se utiliza la función `plot()` para trazar la línea de porcentaje acumulado utilizando los valores de 'country' y 'Percentage'. La línea se representa en color rojo y se añaden marcadores 'o' en los puntos.

El resultado es un diagrama de Pareto que combina un gráfico de barras para representar la población por país y una línea para mostrar cómo se acumula el porcentaje de población en función de los países. Este diagrama proporciona una visión clara de cómo está distribuida la población y cómo se acumula el porcentaje de población a medida que avanzamos en la lista de países.

Población Mundial - Gráfico 13



Ahora queremos saber los países que en conjunto representan el 80% de la población mundial. Para ello utilizamos la columna de “Percentage”, que calculamos previamente.

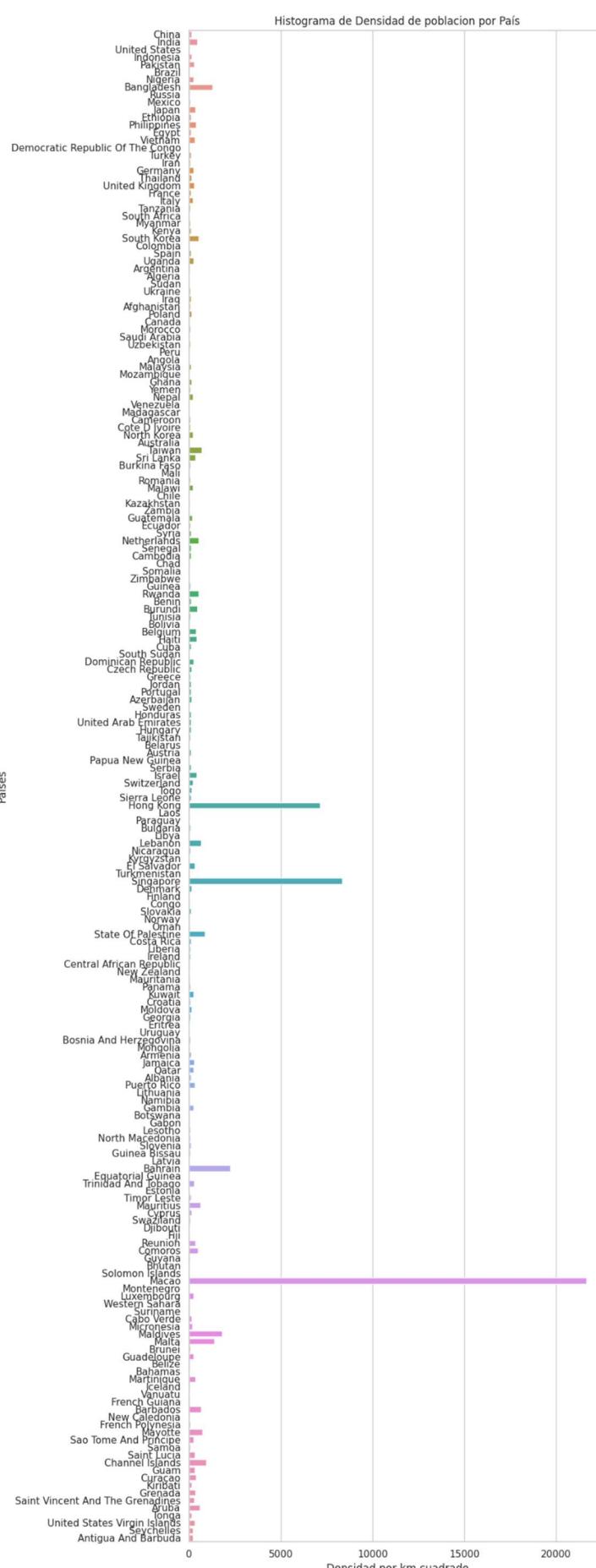
- Filtramos el DataFrame "poblacion\_mundial" para seleccionar las filas donde el valor en la columna 'Percentage' sea igual o menor al 80%. Esto significa que se están seleccionando los países que, en conjunto, representan el 80% de la población.
- De los países que cumplen el criterio anterior, se selecciona la columna 'country' (nombres de los países) y se convierte en una lista utilizando el método `.tolist()`.

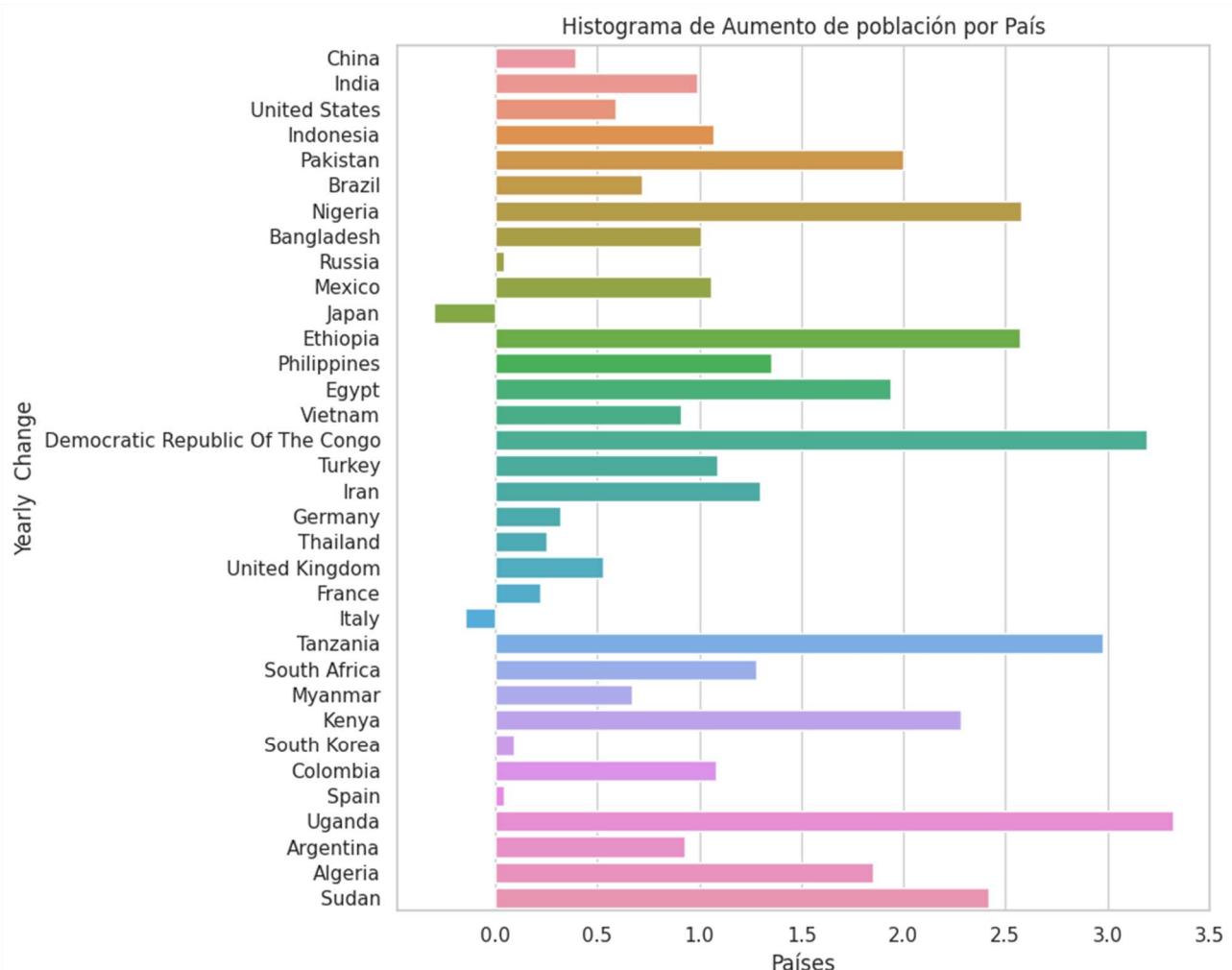
El resultado de los países que representan el 80% de la población:

'China', 'India', 'United States', 'Indonesia', 'Pakistan', 'Brazil', 'Nigeria', 'Bangladesh', 'Russia',  
'Mexico', 'Japan', 'Ethiopia', 'Philippines', 'Egypt', 'Vietnam', 'Democratic Republic Of The Congo',  
'Turkey', 'Iran', 'Germany', 'Thailand', 'United Kingdom', 'France', 'Italy', 'Tanzania', 'South Africa',  
'Myanmar', 'Kenya', 'South Korea', 'Colombia', 'Spain', 'Uganda', 'Argentina', 'Algeria', 'Sudan'.

Creamos un histograma, para representar la distribución de la densidad de población por país en orden descendente.

Población Mundial - Gráfico 14





Este histograma, representa visualmente el aumento anual de la población en los países que en conjunto representan el 80% de la población mundial.

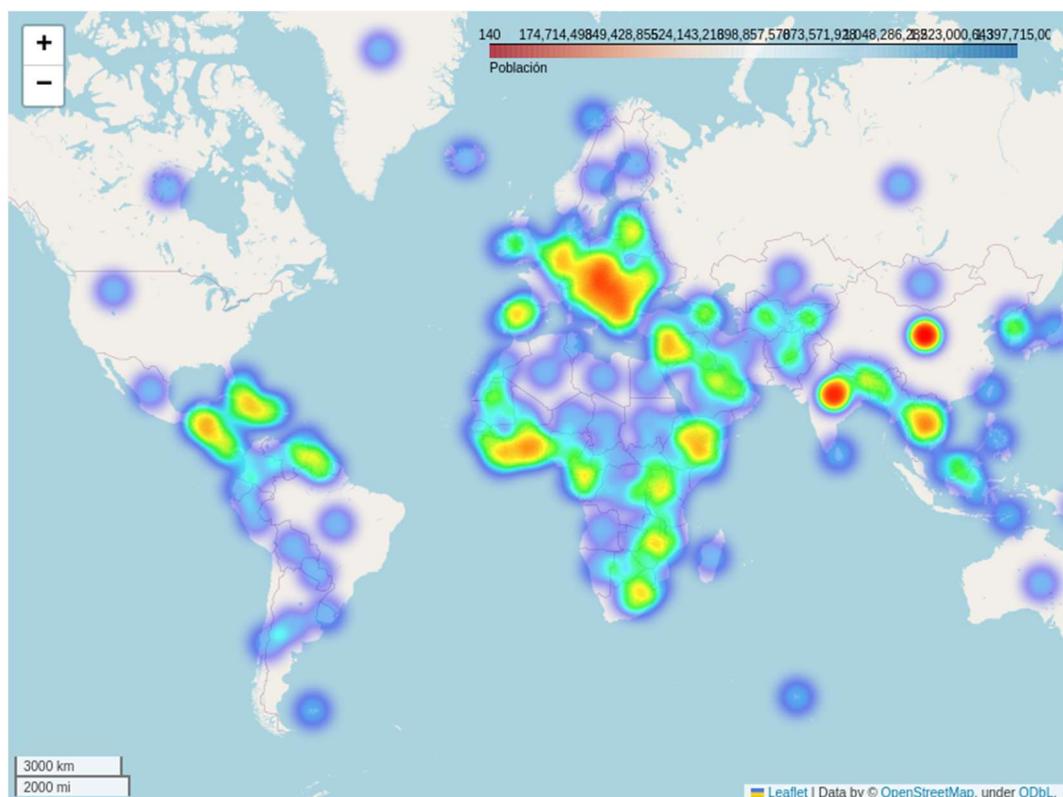
## MAPA DE CALOR

Cargamos ahora la biblioteca Folium para generar un mapa de calor (heatmap) interactivo basado en datos geoespaciales y datos de población.

- Cargamos las bibliotecas necesarias, incluidas Pandas, Geopandas, Folium, Selenium y otras.
- Cargamos un conjunto de datos geoespaciales de los países utilizando Geopandas.
- Realizamos una copia del DataFrame "poblacion\_mundial" para no modificar los datos originales.
- Renombramos las columnas en la copia del DataFrame para que coincidan con las columnas del DataFrame "world".

- Combinamos el conjunto de datos geoespaciales con los datos de población utilizando el método **merge**.
- Creamos una lista de tuplas que contiene las coordenadas geográficas y la población de cada país.
- Creamos una escala de colores para la leyenda del mapa de calor.
- Creamos el mapa de calor utilizando la biblioteca Folium y el plugin HeatMap.
- Crea un mapa de Folium y agregamos el mapa de calor y la leyenda de colores.
- Guardamos el mapa como un archivo HTML.
- Configuramos las opciones para el navegador web automatizado (Selenium) y se crea una instancia del navegador.
- Abrimos el archivo HTML en el navegador y esperamos unos segundos para que el mapa se cargue completamente.
- Capturamos una “captura de pantalla” del mapa utilizando el navegador y lo guardamos como un archivo de imagen.
- Finalmente, muestramos la imagen utilizando **IPImage**.

Población Mundial - Gráfico 16



## ANALISIS DE COMPONENTES PRINCIPALES (PCA)

El PCA es una técnica de reducción de dimensionalidad que se utiliza para encontrar las direcciones principales de variabilidad en los datos y proyectar los datos originales en un espacio de dimensiones reducidas.

- **from sklearn.decomposition import PCA:** Importamos la clase **PCA** de la biblioteca **sklearn.decomposition**.
- **X = poblacion\_mundial.drop(..., axis=1):** Crea un DataFrame **X** que contiene las características que se utilizarán para el análisis de componentes principales. Las características se seleccionan al eliminar ciertas columnas del DataFrame **poblacion\_mundial**. Las columnas eliminadas son: 'country', 'Percentage', 'Urban Pop %', 'log\_Fertility Rate' y 'log\_Median Age'.
- **Y = poblacion\_mundial.country:** Creamos una variable **Y** que contiene las etiquetas de los países. Se utilizan para identificar los países correspondientes a las muestras en el conjunto **X**.

Aplicamos el escalado estándar a las características antes de realizar el análisis de componentes principales (PCA). El escalado estándar es una técnica común en el análisis de datos que transforma las características de manera que tengan media cero y desviación estándar igual a uno. Esto es importante en el contexto de PCA porque las direcciones principales de variación pueden verse afectadas por la escala de las características originales.

Hacemos el Análisis de Componente Principal (PCA) a los datos escalados utilizando la biblioteca scikit-learn (sklearn).

- **pca = PCA():** Creamos una instancia de la clase PCA sin especificar el número de componentes principales a retener. Esto significa que inicialmente se calcularán todos los componentes principales.
- **Objeto = pca.fit(data\_scaled):** Ajustamos el modelo de PCA a los datos escalados en **data\_scaled**.

Ahora si podemos realizar el análisis más detallado del PCA y visualizamos la varianza acumulada y la proporción de varianza explicada por cada componente principal.

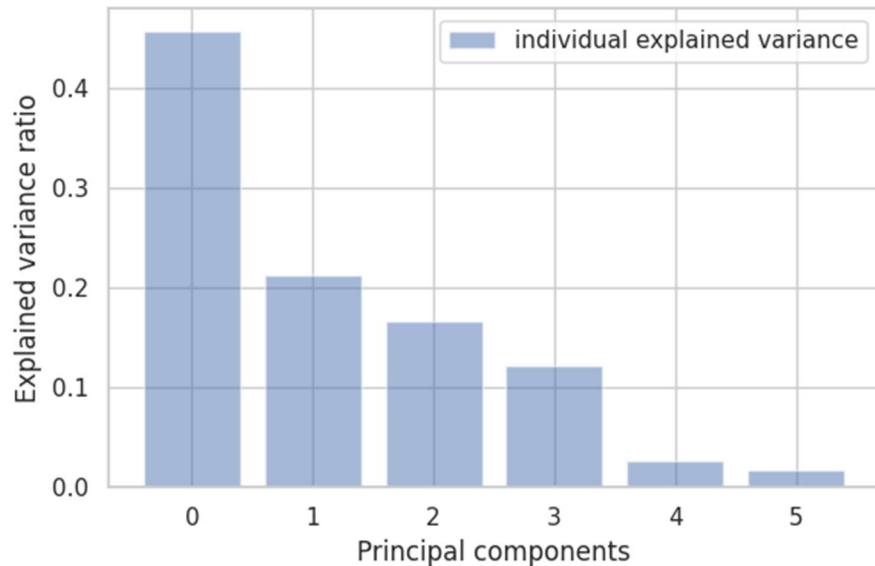
- **varianza\_acumulada = np.cumsum(objeto.explained\_variance\_ratio\_):** Calculamos la varianza acumulada de cada componente principal sumando acumulativamente la

proporción de varianza explicada (`explained_variance_ratio_`) de cada componente. Esto nos permite ver cuánta varianza total se captura a medida que aumentamos el número de componentes principales.

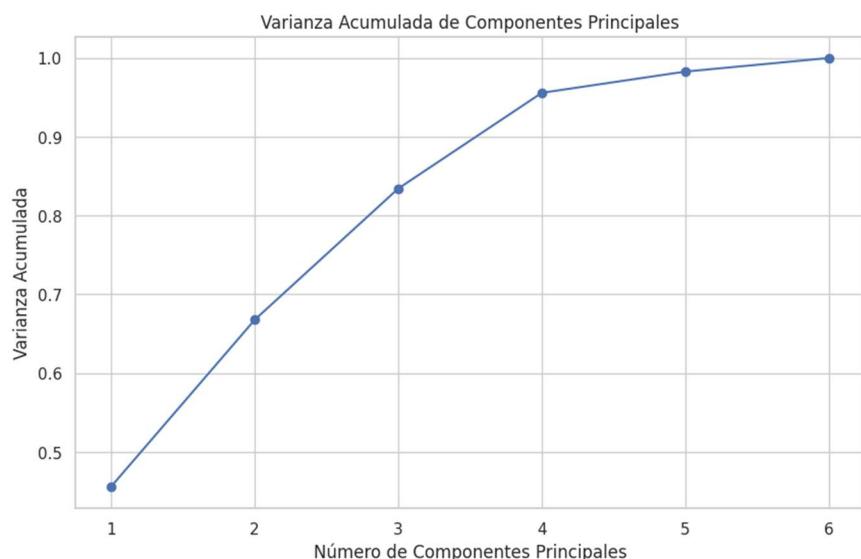
- `explained_variance = pca.explained_variance_ratio_`: Obtenemos la proporción de varianza explicada por cada componente principal y la asignamos a la variable `explained_variance`.

El primer gráfico nos muestra la proporción de varianza explicada por cada componente, lo que nos ayuda a seleccionar el número adecuado de componentes para el análisis. El segundo gráfico de varianza acumulada nos da una idea de cuánta varianza total se captura a medida que aumentamos el número de componentes.

*Población Mundial - Gráfico 17*



*Población Mundial - Gráfico 18*

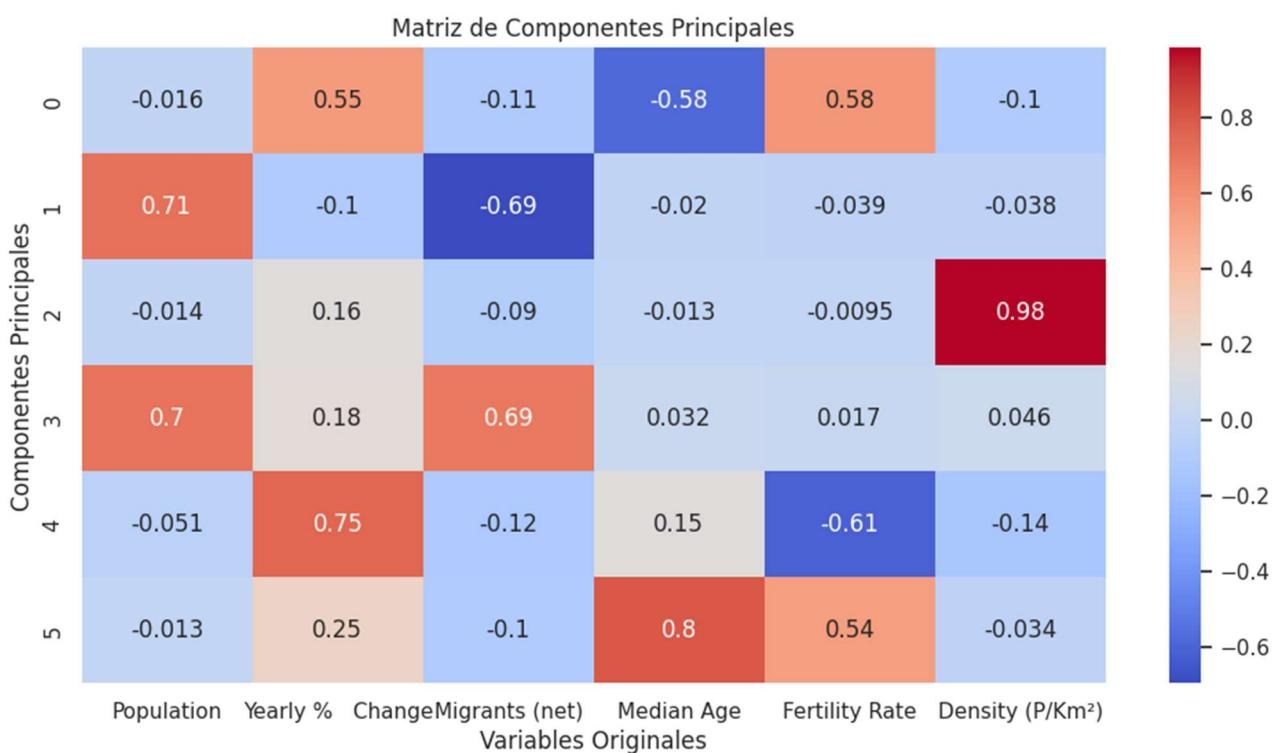


La matriz de componentes principales nos muestra cómo las variables originales contribuyen a cada uno de los componentes principales calculados mediante el análisis de componentes principales (PCA). Cada fila de la matriz representa un componente principal, y cada columna representa una variable original del conjunto de datos. Los valores en la matriz indican la magnitud y dirección de la contribución de cada variable a cada componente principal.

- Cada componente principal es una combinación lineal de las variables originales.
- Los valores positivos o negativos indican la dirección de la contribución de cada variable al componente principal.
- Cuanto mayor sea el valor absoluto de un número en la matriz, mayor será la importancia de la variable correspondiente en la creación del componente principal.
- Las variables que tienen valores más cercanos a cero en un componente principal tienen una contribución menor a ese componente.

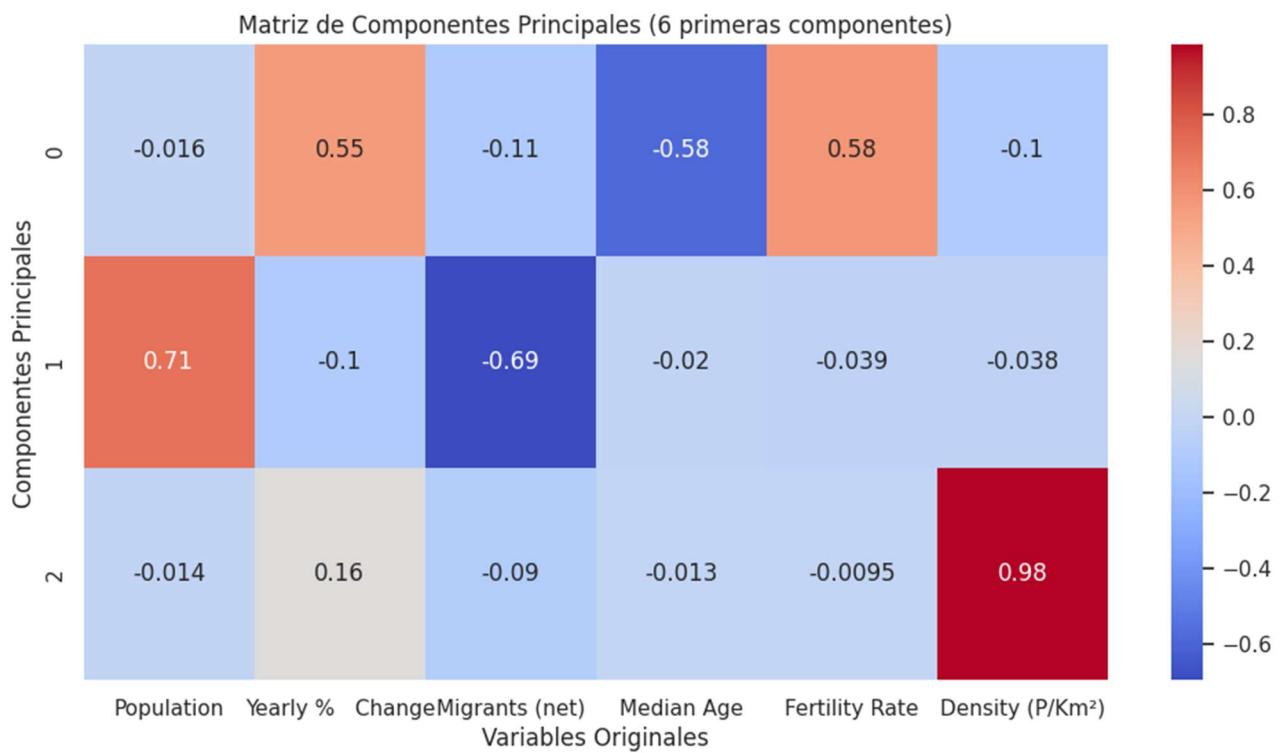
Esta matriz nos permite entender cómo las variables originales están relacionadas y cómo contribuyen a la formación de los componentes principales. Dado que la matriz está normalizada, es posible interpretar las relaciones de magnitud entre las variables y los componentes sin preocuparse por las escalas originales de las variables.

Población Mundial - Gráfico 19



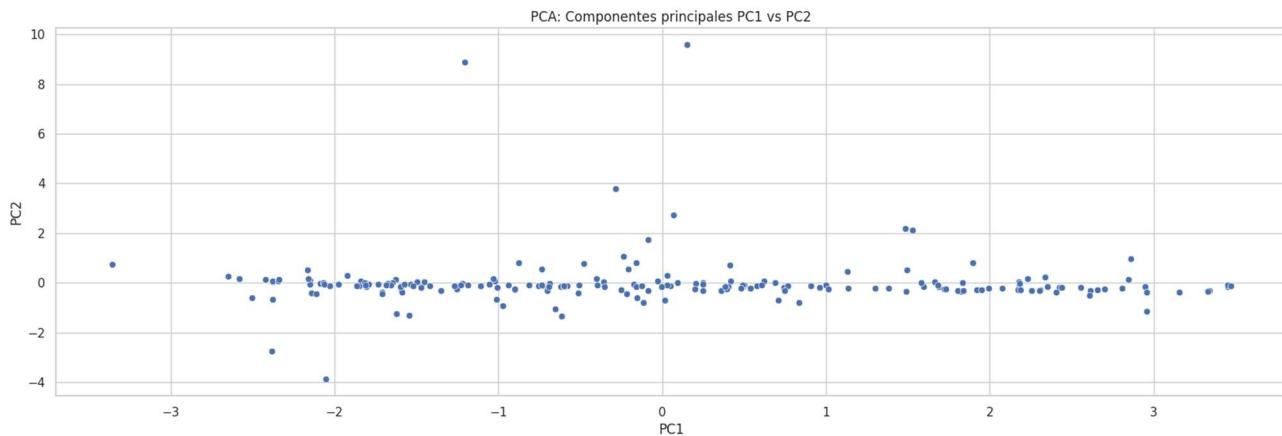
Hacemos otra muestra de la matriz de componentes principales, limitando la visualización a los primeros tres componentes principales.

Población Mundial - Gráfico 20



Transformamos los datos escalados utilizando el modelo de PCA previamente ajustado. La transformación se realiza seleccionando los primeros tres componentes principales. Esto crea una nueva matriz de datos `data_transformed` donde cada fila representa una observación y cada columna corresponde a una de las tres primeras componentes principales y posteriormente creamos un gráfico de dispersión (scatter plot) que muestra las observaciones transformadas en el espacio de las dos primeras componentes principales (PC1 y PC2).

Este gráfico de dispersión nos permite visualizar cómo las observaciones se distribuyen en el espacio de los dos primeros componentes principales. Podemos observar patrones de agrupamiento, dispersión y relaciones entre las observaciones en función de estos dos componentes principales.



## CLUSTERING

El clustering, también conocido como agrupamiento o segmentación, es una técnica en la minería de datos y el aprendizaje automático que busca dividir un conjunto de datos en grupos o clústeres, donde los elementos en cada grupo son más similares entre sí que con elementos de otros grupos. Su objetivo es descubrir patrones subyacentes, estructuras ocultas y relaciones naturales en los datos sin necesidad de etiquetas predefinidas. Algunos propósitos clave del clustering son:

- **Descubrimiento de estructura:** Revela patrones y agrupaciones en los datos, especialmente útiles en grandes conjuntos de datos.
- **Segmentación de clientes:** Agrupa clientes con características similares para personalizar estrategias de marketing.
- **Análisis de mercado:** Identifica grupos de productos comprados juntos para decisiones de colocación.
- **Búsqueda de anomalías:** Detecta puntos atípicos en los datos, como valores atípicos.
- **Comprendión de datos científicos:** Ayuda a encontrar grupos en datos complejos como secuencias genéticas.
- **Segmentación de imágenes y señales:** Agrupa elementos similares en imágenes y señales para reconocimiento de patrones.
- **Agrupamiento de documentos:** Organiza documentos similares para explorar información textual.
- **Reducción de dimensionalidad:** Puede ser usada como técnica de reducción de dimensiones al considerar representantes de clústeres en lugar de puntos individuales.

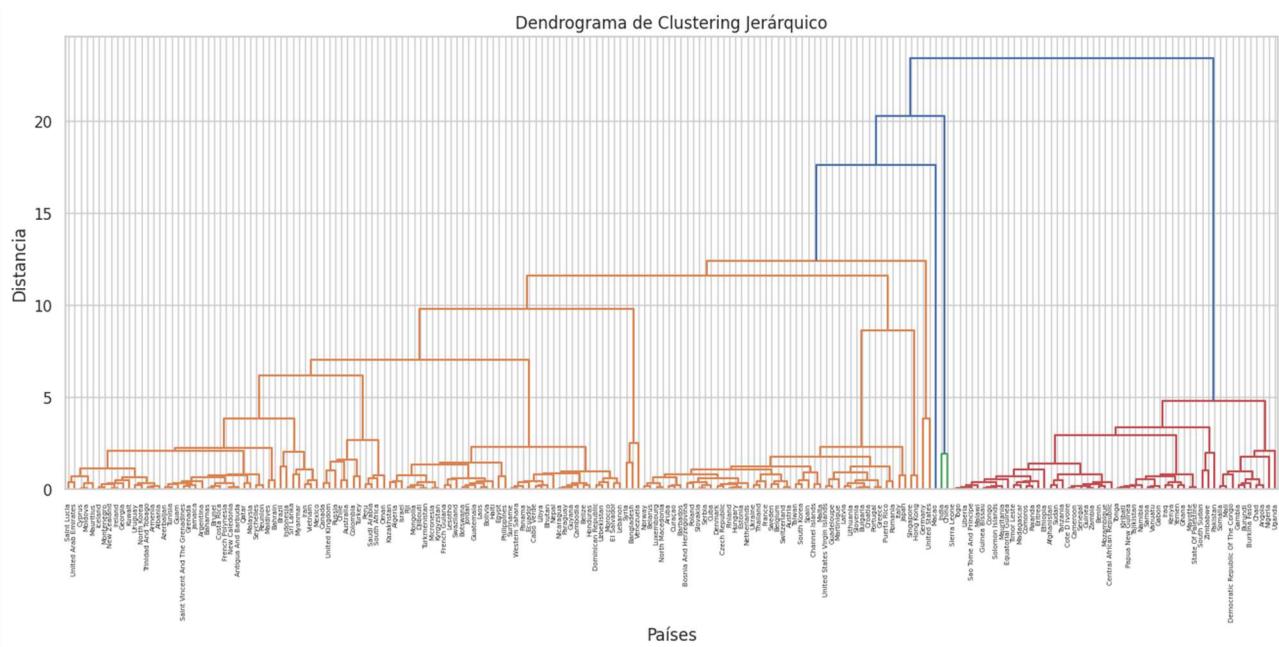
En resumen, el clustering busca hallar estructuras y relaciones en los datos sin etiquetas predefinidas. Esto ofrece comprensión de datos y genera información valiosa para decisiones y estrategias en varias áreas.

# DENDROGRAMA DE CLUSTERING JERÁRQUICO

Vamos a generar un dendrograma de clustering jerárquico, el cual puede ser visualizado como un "árbol genealógico" que muestra cómo las poblaciones de países se agrupan y se relacionan en función de sus atributos demográficos.

En este contexto, el código realiza un proceso de clustering jerárquico utilizando la biblioteca SciPy, con el objetivo de identificar similitudes entre países basadas en sus características demográficas. Tras la conversión y estandarización de los datos, se procede a calcular una matriz de enlace jerárquico. Este proceso se asemeja a construir un árbol que gradualmente conecta a los países, agrupándolos de acuerdo a su similitud en términos de atributos demográficos. El método 'ward' se emplea para determinar cómo los grupos se fusionan y forman las "ramas" del dendrograma.

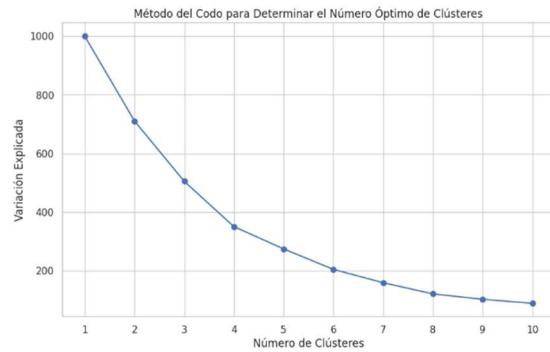
Población Mundial - Gráfico 22



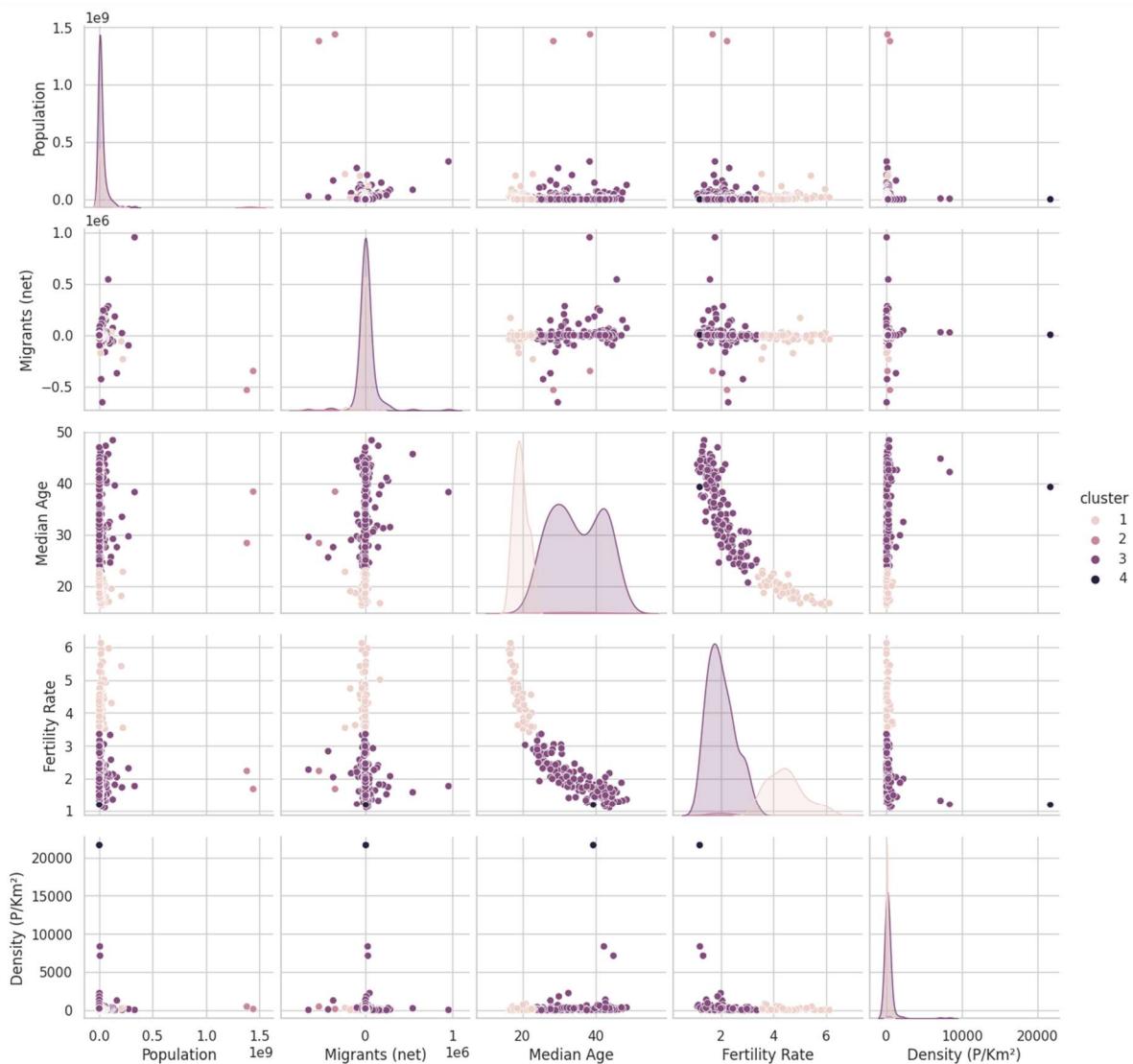
Realizamos un análisis llamado "Método del Codo" para encontrar el mejor número de clústeres en K-Means. Calculamos la variación explicada para diferentes valores de k (número de clústeres) y observamos cómo cambia a medida que k aumenta. Queremos encontrar el punto en el que agregar más clústeres ya no reduce significativamente la variación explicada, formando un "codo" en el gráfico. El código itera a través de valores de k del 1 al 10, realiza K-Means y guarda la

variación explicada en la lista "inertia". Luego, traza un gráfico que muestra cómo la variación explicada disminuye con k. Este análisis ayuda a determinar el número óptimo de clústeres para el conjunto de datos. El punto del "codo" en el gráfico sugiere cuántos clústeres son adecuados sin complicar el modelo. El gráfico del "Método del Codo" guía la elección del número óptimo de clústeres en K-Means.

Población Mundial - Gráfico 23



Población Mundial - Gráfico 24



En este contexto, llevamos a cabo un análisis de agrupamiento utilizando el algoritmo K-Means. Luego, creamos gráficos de dispersión para visualizar los grupos generados en relación con las variables numéricas presentes en el conjunto de datos. Las variables involucradas son: 'Population' (Población), 'Migrants (net)' (Migrantes netos), 'Median Age' (Edad media), 'Fertility Rate' (Tasa de fertilidad) y 'Density (P/Km<sup>2</sup>)' (Densidad de población).

En cada gráfico de dispersión, cada país se representa como un punto coloreado según su clúster de pertenencia. Los ejes del gráfico combinan diferentes pares de características numéricas mencionadas previamente. Esto permite una observación detallada de cómo los países se distribuyen en función de estas características y cómo se forman los clústeres. Se generan un total de 25 gráficos, organizados en una matriz de 5 filas y 5 columnas. Los gráficos en la diagonal principal son histogramas que muestran la distribución de cada variable según los clústeres. Los gráficos fuera de la diagonal principal son gráficos de dispersión que visualizan las relaciones entre pares de variables en relación con los clústeres.

Este análisis de agrupamiento y los gráficos de dispersión resultantes ofrecen una visión holística de cómo las poblaciones se agrupan en función de sus características compartidas.

Esta representación gráfica nos ayuda a comprender mejor cómo se relacionan las características demográficas y cómo se forman los grupos en el contexto de estos datos.

Por ultimo vamos a visualizar los centros de los clústeres en relación a diferentes características.

- **num\_clusters = 4:** Definimos el número de clústeres que se utilizarán en el algoritmo de K-Means.
- **kmeans = KMeans(n\_clusters=num\_clusters, random\_state=42):** Creamos una instancia del algoritmo K-Means con el número de clústeres especificado.
- **poblacion\_mundial['cluster'] = kmeans.fit\_predict(scaled\_features):** Ajustamos el modelo K-Means a los datos escalados y asignamos a cada muestra el número de clúster al que pertenece.
- **def plot\_cluster\_centers(...):** Definimos una función para graficar los centros de los clústeres en un espacio bidimensional.
- **sns.scatterplot(...):** Creamos un gráfico de dispersión para visualizar las muestras coloreadas por el número de clúster al que pertenecen.

- `centers = scaler.inverse_transform(kmeans.cluster_centers_)`: Transformamos los centros de los clústeres de nuevo a la escala original.
- `sns.scatterplot(...)`: Agregamos marcadores negros 'X' en el gráfico para representar los centros de los clústeres.

Con este código podemos visualizar la distribución de las muestras en relación con los centros de los clústeres en un espacio bidimensional. Podemos ver como las muestras se agrupan en función de diferentes características y cómo los centros de los clústeres están ubicados en ese espacio. Esto proporciona una idea de cómo el algoritmo K-Means ha agrupado los datos en función de las características seleccionadas.

Población Mundial - Gráfico 25

