

Análisis preliminar de calidad de datos Recopilación de datos:

Factores socioeconómicos importantes a estudiar inicialmente

- 1. Esperanza de vida al nacer (Total) | worldbank.org link: [SP.DYN.LE00.IN](#)**
- 2. Esperanza de vida al nacer (female) | worldbank.org link: [SP.DYN.LE00.FE.IN](#)**
- 3. Esperanza de vida al nacer (male) | worldbank.org link: [SP.DYN.LE00.MA.IN](#)**

El indicador sobre “**Esperanza de vida al nacer**” se refiere a la cantidad de años que viviría un recién nacido *si los patrones de mortalidad vigentes al momento de su nacimiento no cambian a lo largo de la vida del infante*.

Se considerará como el Target para el modelado, así como, es factor importante para el análisis

4. Urban population (% of total population) | worldbank.org link: [SP.URB.TOTL.IN.ZS](#)

La “**Población Urbana**” se refiere a las personas que viven en áreas urbanas según lo definen las oficinas nacionales de estadística. Los datos son recopilados y suavizados por la División de Población de las Naciones Unidas.

5. Rural population (% of total population) | worldbank.org link: [SP.RUR.TOTL.ZS](#)

La “**Población Rural**” se refiere a las personas que viven en zonas rurales según la definición de las oficinas nacionales de estadística. Se calcula como la diferencia entre la población total y la población urbana.

6. Population growth (annual %) | worldbank.org link: [SP.POP.GROW](#)

La “**Tasa de Crecimiento Poblacional**” anual para el año t es la tasa exponencial de crecimiento de la población a mitad de año desde el año $t-1$ hasta t , expresada como porcentaje. La población se basa en la definición de facto de población, que cuenta a todos los residentes independientemente de su estatus legal o ciudadanía.

7. Inflation, consumer prices (annual %) | worldbank.org link: [FP.CPI.TOTL.ZG](#)

La **Inflación**, medida por el índice de precios al consumidor, refleja el cambio porcentual anual en el costo para el consumidor promedio de adquirir una canasta de bienes y servicios que puede fijarse o cambiarse en intervalos específicos, como por ejemplo anualmente. Generalmente se utiliza la fórmula de Laspeyres.

8. Gini index | worldbank.org link: [worldbank.org link: \[SI.POV.GINI\]\(#\)](#)

El índice de **Gini** mide el grado en que la distribución del ingreso (o, en algunos casos, el gasto de consumo) entre individuos u hogares dentro de una economía se desvía de una distribución perfectamente equitativa. Una curva de Lorenz traza los porcentajes acumulados del ingreso total recibido frente al número acumulado de beneficiarios, comenzando con el individuo o el hogar más pobre. El índice de Gini mide el área entre la curva de Lorenz y una línea hipotética de igualdad absoluta, expresada como porcentaje del área máxima bajo la línea. Así, un índice de Gini de 0 representa una igualdad perfecta, mientras que un índice de 100 implica una desigualdad perfecta.

9. Inflation, GDP deflator (annual %)' | worldbank.org link: [NY.GDP.DEFL.KD.ZG](#)

La inflación medida por la tasa de crecimiento anual del deflactor implícito del PIB muestra la tasa de cambio de precios en la economía en su conjunto. El deflactor implícito del PIB es la relación entre el PIB en moneda local corriente y el PIB en moneda local constante.

10. GDP (current US\$) | worldbank.org link: [NY.GDP.MKTP.CD](#)

El **PIB** a precios de comprador es la suma del valor agregado bruto de todos los productores residentes en la economía más los impuestos sobre los productos y menos los subsidios no incluidos en el valor de los productos. Se calcula sin hacer deducciones por depreciación de activos fabricados o por agotamiento y degradación de recursos naturales. Los datos están en dólares estadounidenses actuales. Las cifras en dólares del PIB se convierten a partir de monedas nacionales utilizando tipos de cambio oficiales de un solo año. Para algunos países donde el tipo de cambio oficial no refleja el tipo efectivamente aplicado a las transacciones reales de divisas, se utiliza un factor de conversión alternativo.

11. GDP per capita (current US\$) | worldbank.org link: [NY.GDP.PCAP.CD](#)

El **PIB per cápita** es el producto interno bruto dividido por la población a mitad de año. El PIB es la suma del valor agregado bruto de todos los productores residentes en la economía más los impuestos sobre los productos y menos los subsidios no incluidos en el valor de los productos. Se calcula sin hacer deducciones por depreciación de activos fabricados o por agotamiento y degradación de recursos naturales. Los datos están en dólares estadounidenses actuales.

12. GDP per capita growth (annual %) | worldbank.org link: [NY.GDP.PCAP.KD.ZG](#)

Tasa de crecimiento porcentual anual del PIB per cápita basada en moneda local constante. El PIB per cápita es el producto interno bruto dividido por la población a mitad de año. El PIB a precios de comprador es la suma del valor agregado bruto de todos los productores residentes en la economía más los impuestos sobre los productos y menos los subsidios no incluidos en el valor de los productos. Se calcula sin hacer deducciones por depreciación de activos fabricados o por agotamiento y degradación de recursos naturales.

13. GNI (current US\$) | worldbank.org link: [NY.GNP.MKTP.CD](#)

El INB (anteriormente PNB) es la suma del valor agregado de todos los productores residentes más cualquier impuesto sobre los productos (menos subsidios) no incluido en la valoración de la producción más los ingresos netos de ingresos primarios (compensación de los empleados e ingresos de la propiedad) del exterior. Los datos están en dólares estadounidenses actuales.

14. Current health expenditure (% of GDP) | worldbank.org link: [SH.XPD.CHEX.GD.ZS](#)

Nivel de gasto corriente en salud expresado como porcentaje del PIB. Las estimaciones de los gastos corrientes en salud incluyen los bienes y servicios de atención médica consumidos durante cada año. Este indicador no incluye gastos de capital en salud, como edificios, maquinaria, TI y reservas de vacunas para emergencias o brotes.

15. Hospital beds (per 1,000 people)' | worldbank.org link: [SH.MED.BEDS.ZS](#)

Las camas hospitalarias incluyen camas para pacientes hospitalizados disponibles en hospitales y centros de rehabilitación públicos, privados, generales y especializados. En la mayoría de los casos se incluyen camas para cuidados agudos y crónicos.

16. Domestic private health expenditure per capita (current US\$)' | worldbank.org link: [SH.XPD.PVTD.PC.CD](#)

Gasto privado corriente en salud per cápita expresado en dólares estadounidenses corrientes. Las fuentes privadas nacionales incluyen fondos de hogares, corporaciones y organizaciones sin fines de lucro. Dichos gastos pueden pagarse por adelantado al seguro médico voluntario o pagarse directamente a los proveedores de atención médica.

17. Current health expenditure per capita (current US\$)' | worldbank.org link: [SH.XPD.CHEX.PC.CD](#)

Gasto privado corriente en salud per cápita expresado en dólares estadounidenses corrientes. Las fuentes privadas nacionales incluyen fondos de hogares, corporaciones y organizaciones sin fines de lucro. Dichos gastos pueden pagarse por adelantado al seguro médico voluntario o pagarse directamente a los proveedores de atención médica.

18. Mortality caused by road traffic injury (per 100,000 population)' | worldbank.org link: [SH.STA.TRAF.P5](#)

La mortalidad causada por lesiones por accidentes de tránsito se estima en muertes por lesiones mortales por accidentes de tránsito por cada 100.000 habitantes.

19. People using at least basic sanitation services (% of population)' | worldbank.org link: [SH.STA.BASS.ZS](#)

El porcentaje de personas que utilizan al menos servicios básicos de saneamiento, es decir, instalaciones sanitarias mejoradas que no se comparten con otros hogares. Este indicador abarca tanto a las personas que utilizan servicios de saneamiento básicos como a aquellas que utilizan servicios de saneamiento gestionados de forma segura. Las instalaciones de saneamiento mejoradas incluyen sistemas de descarga/vertido de agua a sistemas de alcantarillado, fosas sépticas o letrinas de pozo; letrinas de pozo mejoradas y ventiladas, sanitarios compuestos o letrinas de pozo con losas.

20. People using safely managed drinking water services (% of population)' | worldbank.org link: [SH.H2O.SMDW.ZS](https://data.worldbank.org/SH.H2O.SMDW.ZS)

El porcentaje de personas que utilizan agua potable de una fuente mejorada que sea accesible en las instalaciones, disponible cuando sea necesario y libre de contaminación fecal y química prioritaria. Las fuentes de agua mejoradas incluyen agua entubada, perforaciones o pozos entubados, pozos excavados protegidos, manantiales protegidos y agua envasada o entregada.

21. Literacy rate, adult total (% of people ages 15 and above)' | worldbank.org link: [SE.ADT.LITR.ZS](https://data.worldbank.org/SE.ADT.LITR.ZS)

La tasa de alfabetización de adultos es el porcentaje de personas de 15 años o más que pueden leer y escribir y comprender una declaración breve y sencilla sobre su vida cotidiana.

22. Control of Corruption: Estimate' worldbank.org link: [CC.ESI](https://data.worldbank.org/CC.ESI)

Control de la Corrupción captura percepciones de hasta qué punto el poder público se ejerce para beneficio privado, incluidas formas tanto pequeñas como grandes de corrupción, así como la "captura" del Estado por parte de élites e intereses privados. La estimación proporciona la puntuación del país en el indicador agregado, en unidades de una distribución normal estándar, es decir, entre aproximadamente -2,5 y 2,5.

23. Military expenditure (% of GDP)' | worldbank.org link: [MS.MIL.XPND.GD.ZS](https://data.worldbank.org/MS.MIL.XPND.GD.ZS)

Los datos sobre gastos militares del SIPRI se derivan de la definición de la OTAN, que incluye todos los gastos corrientes y de capital en las fuerzas armadas, incluidas las fuerzas de mantenimiento de la paz; ministerios de defensa y otras agencias gubernamentales involucradas en proyectos de defensa; fuerzas paramilitares, si se considera que están entrenadas y equipadas para operaciones militares; y actividades espaciales militares. Dichos gastos incluyen personal militar y civil, incluidas pensiones de jubilación del personal militar y servicios sociales para el personal; operación y mantenimiento; obtención; investigación y desarrollo militar; y ayuda militar (en los gastos militares del país donante). Se excluyen la defensa civil y los gastos corrientes de actividades militares anteriores, como los beneficios para veteranos, la desmovilización, la conversión y la destrucción de armas. Sin embargo, esta definición no se puede aplicar a todos los países, ya que requeriría información mucho más detallada que la que está disponible sobre lo que se incluye en los presupuestos militares y en las partidas de gastos militares extrapresupuestarios. (Por ejemplo, los presupuestos militares pueden cubrir o no la defensa civil, las reservas y las fuerzas auxiliares, la policía y las fuerzas paramilitares, las fuerzas de doble propósito como la policía militar y civil, las subvenciones militares en especie, las pensiones del personal militar y las contribuciones a la seguridad social pagadas de una parte del gobierno a otra.)

24. General government final consumption expenditure (current US\$)' | worldbank.org link: [NE.CON.GOV.CD](https://data.worldbank.org/NE.CON.GOV.CD)

El gasto de consumo final del gobierno general (anteriormente consumo del gobierno general) incluye todos los gastos corrientes del gobierno para compras de bienes y servicios (incluida la remuneración de los empleados). También incluye la mayoría de los gastos en defensa y seguridad nacionales, pero excluye los gastos militares del gobierno que forman parte de la formación de capital del gobierno. Los datos están en dólares estadounidenses actuales.

25. Access to electricity (% of population)' | worldbank.org link: [EG.ELC.ACCS.ZS](#)

El acceso a la electricidad es el porcentaje de la población con acceso a la electricidad. Los datos sobre electrificación se recopilan de la industria, encuestas nacionales y fuentes internacionales.

26. Food exports (% of merchandise exports)' | worldbank.org link: [TX.VAL.FOOD.ZS.UN](#)

Los alimentos comprenden los productos de las secciones 0 (alimentos y animales vivos), 1 (bebidas y tabaco) y 4 (aceites y grasas animales y vegetales) de la CUCI y la división 22 de la CUCI (semillas, nueces y granos oleaginosos).

27. Food production index (2014-2016 = 100)' | worldbank.org link: [AG.PRD.FOOD.XD](#)

El índice de producción de alimentos cubre cultivos alimentarios que se consideran comestibles y que contienen nutrientes. Se excluyen el café y el té porque, aunque son comestibles, no tienen valor nutritivo.

28. Households and NPISHs final consumption expenditure (% of GDP)' | worldbank.org link: [NE.CON.PRVT.ZS](#)

El gasto de consumo final de los hogares (anteriormente consumo privado) es el valor de mercado de todos los bienes y servicios, incluidos los productos duraderos (como automóviles, lavadoras y computadoras domésticas), adquiridos por los hogares. Excluye las compras de viviendas pero incluye el alquiler imputado de las viviendas ocupadas por sus propietarios. También incluye pagos y tarifas a los gobiernos para obtener permisos y licencias. Aquí, el gasto de consumo de los hogares incluye los gastos de las instituciones sin fines de lucro que prestan servicios a los hogares, incluso cuando el país los informa por separado. Esta partida también incluye cualquier discrepancia estadística en el uso de recursos en relación con la oferta de recursos.

29. School enrollment, tertiary (gross), gender parity index (GPI)' | worldbank.org link: [SE.ENR.TERT.FM.ZS](#)

El índice de paridad de género para la tasa bruta de matriculación en educación terciaria es la proporción de mujeres y hombres matriculados en el nivel terciario en escuelas públicas y privadas.

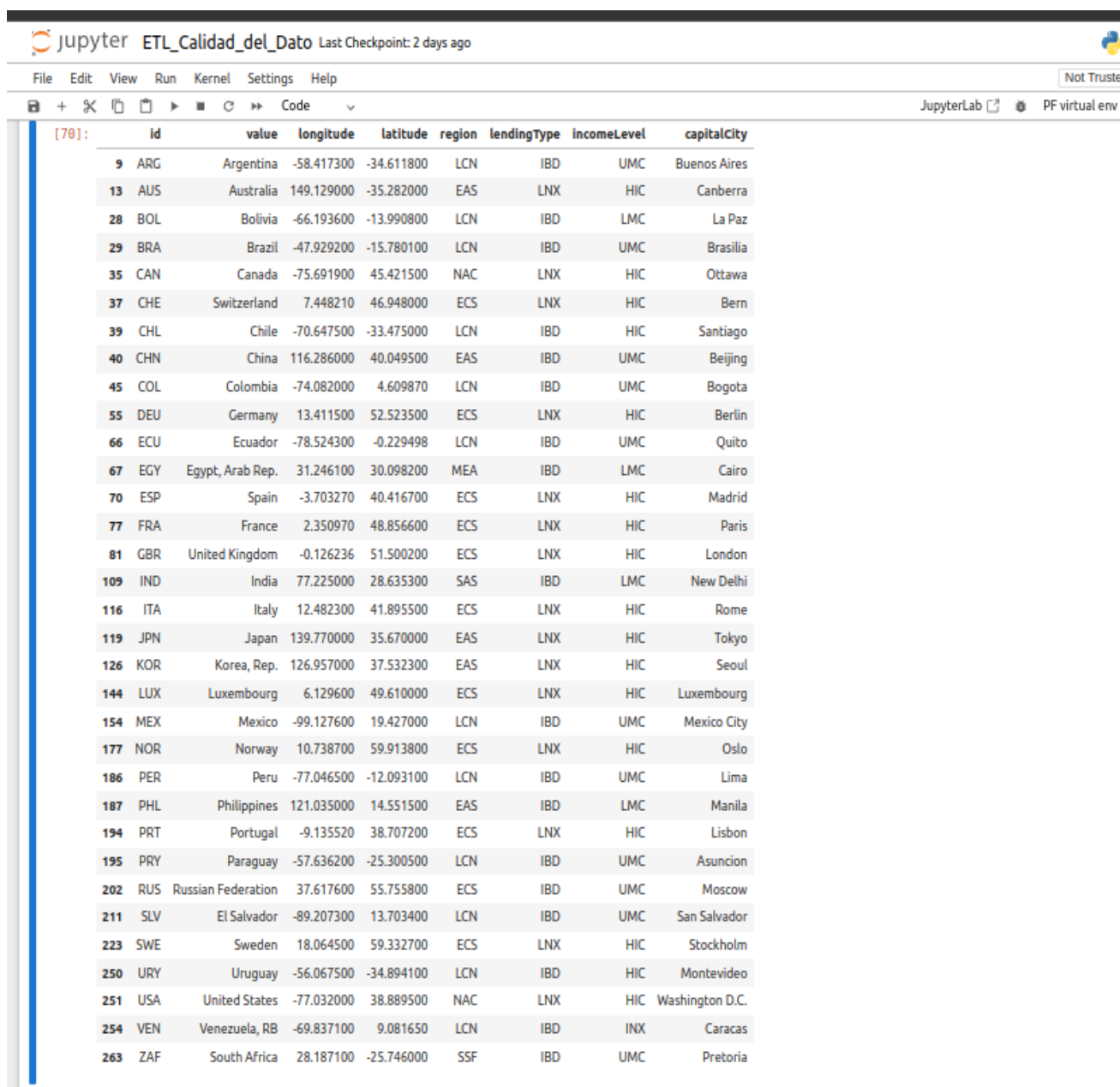
30. School enrollment, tertiary (% gross)', | worldbank.org link: [SE.ENR.TERT.FM.ZS](#)

El índice de paridad de género para la tasa bruta de matriculación en educación terciaria es la proporción de mujeres y hombres matriculados en el nivel terciario en escuelas públicas y privadas.

Estos factores socioeconómicos son importantes para comprender la situación de un país en términos de salud, desarrollo económico, acceso a servicios básicos y calidad de vida. Al analizar estos indicadores en conjunto, se puede obtener una imagen más completa de la situación socioeconómica de una nación y su posible impacto en la esperanza de vida de la población.

Se revisa la documentación proporcionada por las fuentes de datos para comprender el significado y la interpretación correcta de cada variable.

Países que formarán parte del Estudio.



JupyterLab ETL_Calidad_del_Dato Last Checkpoint: 2 days ago

File Edit View Run Kernel Settings Help

JupyterLab PF virtual env

[70]:

	id	value	longitude	latitude	region	lendingType	incomeLevel	capitalCity
9	ARG	Argentina	-58.417300	-34.611800	LCN	IBD	UMC	Buenos Aires
13	AUS	Australia	149.129000	-35.282000	EAS	LNK	HIC	Canberra
28	BOL	Bolivia	-66.193600	-13.990800	LCN	IBD	LMC	La Paz
29	BRA	Brazil	-47.929200	-15.780100	LCN	IBD	UMC	Brasilia
35	CAN	Canada	-75.691900	45.421500	NAC	LNK	HIC	Ottawa
37	CHE	Switzerland	7.448210	46.948000	ECS	LNK	HIC	Bern
39	CHL	Chile	-70.647500	-33.475000	LCN	IBD	HIC	Santiago
40	CHN	China	116.286000	40.049500	EAS	IBD	UMC	Beijing
45	COL	Colombia	-74.082000	4.609870	LCN	IBD	UMC	Bogota
55	DEU	Germany	13.411500	52.523500	ECS	LNK	HIC	Berlin
66	ECU	Ecuador	-78.524300	-0.229498	LCN	IBD	UMC	Quito
67	EGY	Egypt, Arab Rep.	31.246100	30.098200	MEA	IBD	LMC	Cairo
70	ESP	Spain	-3.703270	40.416700	ECS	LNK	HIC	Madrid
77	FRA	France	2.350970	48.856600	ECS	LNK	HIC	Paris
81	GBR	United Kingdom	-0.126236	51.500200	ECS	LNK	HIC	London
109	IND	India	77.225000	28.635300	SAS	IBD	LMC	New Delhi
116	ITA	Italy	12.482300	41.895500	ECS	LNK	HIC	Rome
119	JPN	Japan	139.770000	35.670000	EAS	LNK	HIC	Tokyo
126	KOR	Korea, Rep.	126.957000	37.532300	EAS	LNK	HIC	Seoul
144	LUX	Luxembourg	6.129600	49.610000	ECS	LNK	HIC	Luxembourg
154	MEX	Mexico	-99.127600	19.427000	LCN	IBD	UMC	Mexico City
177	NOR	Norway	10.738700	59.913800	ECS	LNK	HIC	Oslo
186	PER	Peru	-77.046500	-12.093100	LCN	IBD	UMC	Lima
187	PHL	Philippines	121.035000	14.551500	EAS	IBD	LMC	Manila
194	PRT	Portugal	-9.135520	38.707200	ECS	LNK	HIC	Lisbon
195	PRY	Paraguay	-57.636200	-25.300500	LCN	IBD	UMC	Asuncion
202	RUS	Russian Federation	37.617600	55.755800	ECS	IBD	UMC	Moscow
211	SLV	El Salvador	-89.207300	13.703400	LCN	IBD	UMC	San Salvador
223	SWE	Sweden	18.064500	59.332700	ECS	LNK	HIC	Stockholm
250	URY	Uruguay	-56.067500	-34.894100	LCN	IBD	HIC	Montevideo
251	USA	United States	-77.032000	38.889500	NAC	LNK	HIC	Washington D.C.
254	VEN	Venezuela, RB	-69.837100	9.081650	LCN	IBD	INX	Caracas
263	ZAF	South Africa	28.187100	-25.746000	SSF	IBD	UMC	Pretoria

Se preseleccionaron un total de 50 países para ser evaluados con respecto a dos aspectos primordiales:

- Los países deberán tener una economía con **nivel de ingresos (incomen) minimo del tipo medio**
- Los países deberán contener la mayor cantidad de datos para su evaluación respecto a los indicadores seleccionados.

De los resultados arrojados por el análisis fueron seleccionados treinta tres (33) países que se muestran en la tabla anterior.

Una vez seleccionados los países e indicadores para la realización del estudio se procede a generar un dataframe con los datos por medio de una función personalizada:

1. Listado de indicadores y dataframe resultante

jupyter ETL_Calidad_del_Dato Last Checkpoint: 2 days ago

File Edit View Run Kernel Settings Help

Not Trusted

JupyterLab PF virtual env

```
[71]: # Diccionario de indicadores
indicators = {
    'SP.DYN.LE00.FE.IN': 'Life expectancy at birth, female (years)',
    'SP.DYN.LE00.IN': 'Life expectancy at birth, total (years)',
    'SP.DYN.LE00.MA.IN': 'Life expectancy at birth, male (years)',
    'SP.URB.TOTL.IN.ZS': 'Urban population (% of total population)',
    'SP.RUR.TOTL.ZS': 'Rural population (% of total population)',
    'SP.POP.GROW': 'Population growth (annual %)',
    'FP.CPI.TOTL.ZG': 'Inflation, consumer prices (annual %)',
    'SI.POV.GINI': 'Gini index',
    'NY.GDP.DEFL.KD.ZG': 'Inflation, GDP deflator (annual %)',
    'NY.GDP.MKTP.CD': 'GDP (current US$)',
    'NY.GDP.PCAP.CD': 'GDP per capita (current US$)',
    'NY.GDP.PCAP.KD.ZG': 'GDP per capita growth (annual %)',
    'NY.GNP.MKTP.CD': 'GNI (current US$)',
    'SH.XPD.CHEX.GD.ZS': 'Current health expenditure (% of GDP)',
    'SH.MED.BEDS.ZS': 'Hospital beds (per 1,000 people)',
    'SH.XPD.PVTD.PC.CD': 'Domestic private health expenditure per capita (current US$)',
    'SH.XPD.CHEX.PC.CD': 'Current health expenditure per capita (current US$)',
    'SH.STA.TRAF.P5': 'Mortality caused by road traffic injury (per 100,000 population)',
    'SH.STA.BASS.ZS': 'People using at least basic sanitation services (% of population)',
    'SH.H2O.SMDW.ZS': 'People using safely managed drinking water services (% of population)',
    'SE.ADT.LITR.ZS': 'Literacy rate, adult total (% of people ages 15 and above)',
    'CC.EST': 'Control of Corruption: Estimate',
    'MS.MIL.XPND.GD.ZS': 'Military expenditure (% of GDP)',
    'NE.CON.GOVT.CD': 'General government final consumption expenditure (current US$)',
    'EG.ELC.ACCS.ZS': 'Access to electricity (% of population)',
    'TX.VAL.FOOD.ZS.UN': 'Food exports (% of merchandise exports)',
    'AG.PR.D.FOOD.XD': 'Food production index (2014-2016 = 100)',
    'NE.CON.PRVT.ZS': 'Households and NPISHs final consumption expenditure (% of GDP)',
    'SE.ENR.TERT.FM.ZS': 'School enrollment, tertiary (gross), gender parity index (GPI)',
    'SE.TER.ENRR': 'School enrollment, tertiary (% gross)',
}

countries = countries_df['id'].tolist()
```

jupyter ETL_Calidad_del_Dato Last Checkpoint: 2 days ago

File Edit View Run Kernel Settings Help

Not Trusted

JupyterLab PF virtual env

```
df[data[i]['name']] = data[i]['value']

return df

[73]: df = crerate_df_bm(indicators, countries)
df

[73]:
```

	country	year	Life expectancy at birth, female (years)	Life expectancy at birth, total (years)	Life expectancy at birth, male (years)	Urban population (% of total population)	Rural population (% of total population)	Population growth (annual %)	Inflation, consumer prices (annual %)	Gini index	Inflation, GDP deflator (annual %)	GDP (current US\$)	GDP per capita (current US\$)	GDP per capita growth (annual %)
0	ZAF	2022	NaN	NaN	NaN	68.335	31.665	0.841058	7.039727	NaN	5.054293	4.058697e+11	6776.480077	1.187663
1	ZAF	2021	64.999	62.341	59.458	67.847	32.153	0.998920	4.611672	NaN	6.219017	4.190156e+11	7055.055176	3.870315
2	ZAF	2020	67.964	65.252	62.178	67.354	32.646	1.223179	3.210036	NaN	5.692615	3.376196e+11	5741.641249	-7.481093
3	ZAF	2019	69.107	66.175	62.834	66.856	33.144	1.295074	4.120246	NaN	4.638081	3.885312e+11	6688.774746	-0.987175
4	ZAF	2018	68.740	65.674	62.203	66.355	33.645	1.225530	4.517165	NaN	3.745754	4.041589e+11	7048.508112	0.285736
...
2074	ARG	1964	68.261	64.363	61.037	75.844	24.156	1.599523	NaN	NaN	28.774617	2.560525e+10	1179.504098	8.382751
2075	ARG	1963	68.285	64.449	61.178	75.309	24.691	1.621661	NaN	NaN	25.591154	1.827212e+10	855.275656	-6.831393
2076	ARG	1962	68.031	64.244	61.035	74.767	25.233	1.629282	NaN	NaN	28.871842	2.445060e+10	1163.186836	-2.454333
2077	ARG	1961	68.031	64.360	61.240	74.217	25.783	1.613029	NaN	NaN	20.310698	NaN	NaN	3.740903
2078	ARG	1960	67.617	63.978	60.916	73.611	26.389	NaN	NaN	NaN	NaN	NaN	NaN	NaN

2079 rows x 32 columns

2. Visión general de la estructura del dataframe, información sobre los tipos de datos, la presencia de valores nulos y la memoria utilizada.

```
jupyter ETL_Calidad_del_Dato Last Checkpoint: 2 days ago

File Edit View Run Kernel Settings Help

+ - + Code

[74]: # Verificar tipos de datos
df.dtypes

[74]: country                object
      year                  object
      Life expectancy at birth, female (years)  float64
      Life expectancy at birth, total (years)   float64
      Life expectancy at birth, male (years)    float64
      Urban population (% of total population)  float64
      Rural population (% of total population)  float64
      Population growth (annual %)             float64
      Inflation, consumer prices (annual %)    float64
      Gini index                             float64
      Inflation, GDP deflator (annual %)        float64
      GDP (current US$)                       float64
      GDP per capita (current US$)             float64
      GDP per capita growth (annual %)         float64
      GNI (current US$)                       float64
      Current health expenditure (% of GDP)    float64
      Hospital beds (per 1,000 people)         float64
      Domestic private health expenditure per capita (current US$) float64
      Current health expenditure per capita (current US$) float64
      Mortality caused by road traffic injury (per 100,000 population) float64
      People using at least basic sanitation services (% of population) float64
      People using safely managed drinking water services (% of population) float64
      Literacy rate, adult total (% of people ages 15 and above) float64
      Control of Corruption: Estimate         float64
      Military expenditure (% of GDP)         float64
      General government final consumption expenditure (current US$) float64
      Access to electricity (% of population) float64
      Food exports (% of merchandise exports) float64
      Food production index (2014-2016 = 100) float64
      Households and NPISHs final consumption expenditure (% of GDP) float64
      School enrollment, tertiary (gross), gender parity index (GPI) float64
      School enrollment, tertiary (% gross)   float64
      dtype: object
```

```
jupyter ETL_Calidad_del_Dato Last Checkpoint: 2 days ago

File Edit View Run Kernel Settings Help

+ - + Code

[76]: # Información general del dataframe
df.info()

<class 'pandas.core.frame.DataFrame'>
Index: 1254 entries, 0 to 2053
Data columns (total 32 columns):
# Column Non-Null Count Dtype
---
0 country 1254 non-null object
1 year 1254 non-null int64
2 Life expectancy at birth, female (years) 1221 non-null float64
3 Life expectancy at birth, total (years) 1221 non-null float64
4 Life expectancy at birth, male (years) 1221 non-null float64
5 Urban population (% of total population) 1254 non-null float64
6 Rural population (% of total population) 1254 non-null float64
7 Population growth (annual %) 1254 non-null float64
8 Inflation, consumer prices (annual %) 1175 non-null float64
9 Gini index 700 non-null float64
10 Inflation, GDP deflator (annual %) 1241 non-null float64
11 GDP (current US$) 1243 non-null float64
12 GDP per capita (current US$) 1243 non-null float64
13 GDP per capita growth (annual %) 1241 non-null float64
14 GNI (current US$) 1223 non-null float64
15 Current health expenditure (% of GDP) 703 non-null float64
16 Hospital beds (per 1,000 people) 820 non-null float64
17 Domestic private health expenditure per capita (current US$) 699 non-null float64
18 Current health expenditure per capita (current US$) 704 non-null float64
19 Mortality caused by road traffic injury (per 100,000 population) 660 non-null float64
20 People using at least basic sanitation services (% of population) 753 non-null float64
21 People using safely managed drinking water services (% of population) 522 non-null float64
22 Literacy rate, adult total (% of people ages 15 and above) 227 non-null float64
23 Control of Corruption: Estimate 759 non-null float64
24 Military expenditure (% of GDP) 1243 non-null float64
25 General government final consumption expenditure (current US$) 1230 non-null float64
26 Access to electricity (% of population) 1011 non-null float64
27 Food exports (% of merchandise exports) 1156 non-null float64
28 Food production index (2014-2016 = 100) 1199 non-null float64
29 Households and NPISHs final consumption expenditure (% of GDP) 1230 non-null float64
30 School enrollment, tertiary (gross), gender parity index (GPI) 785 non-null float64
31 School enrollment, tertiary (% gross) 908 non-null float64
dtypes: float64(30), int64(1), object(1)
memory usage: 323.3+ KB
```

3. Cálculo de porcentajes de valores nulos por variable

```
jupyter ETL_Calidad_del_Dato Last Checkpoint: 2 days ago

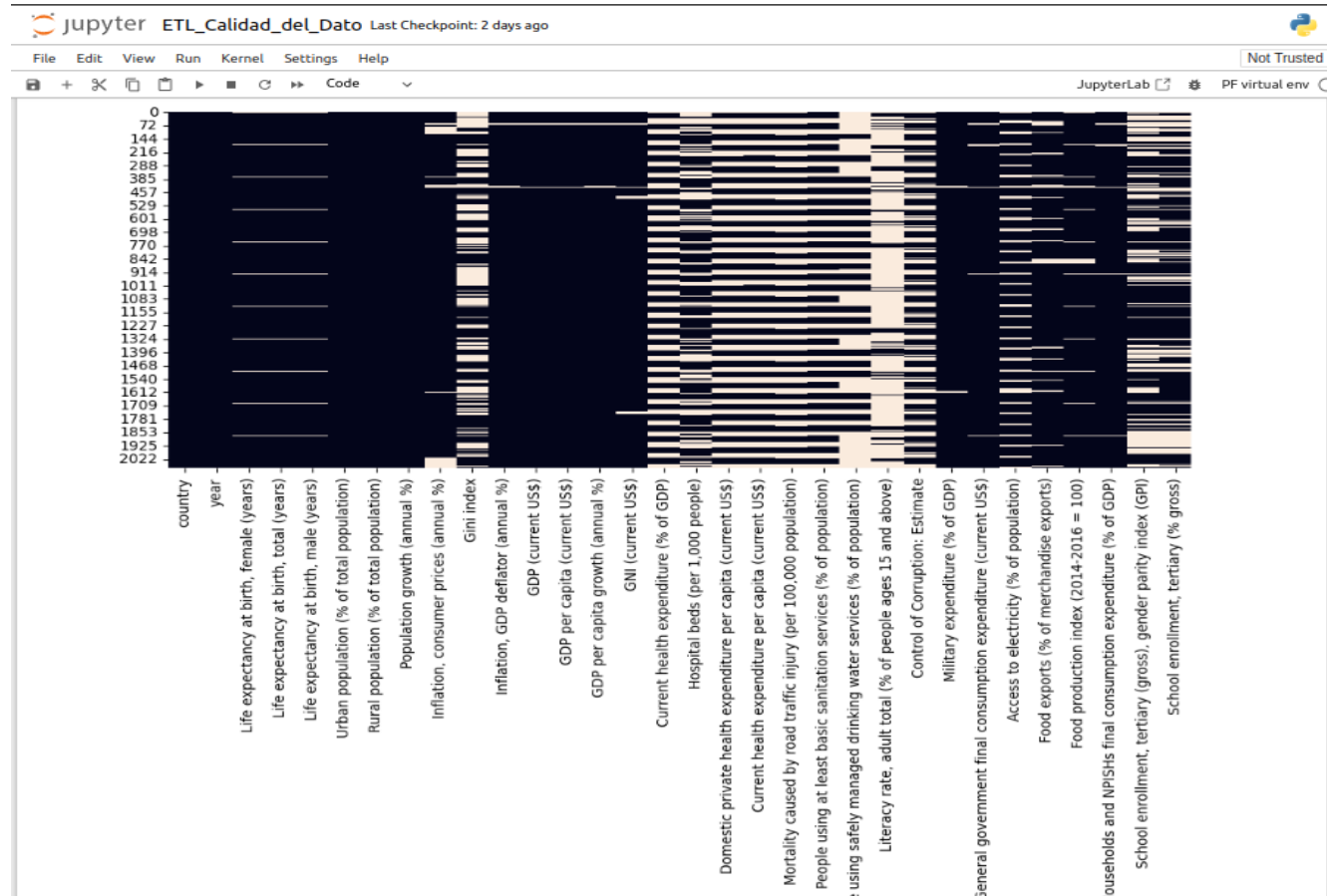
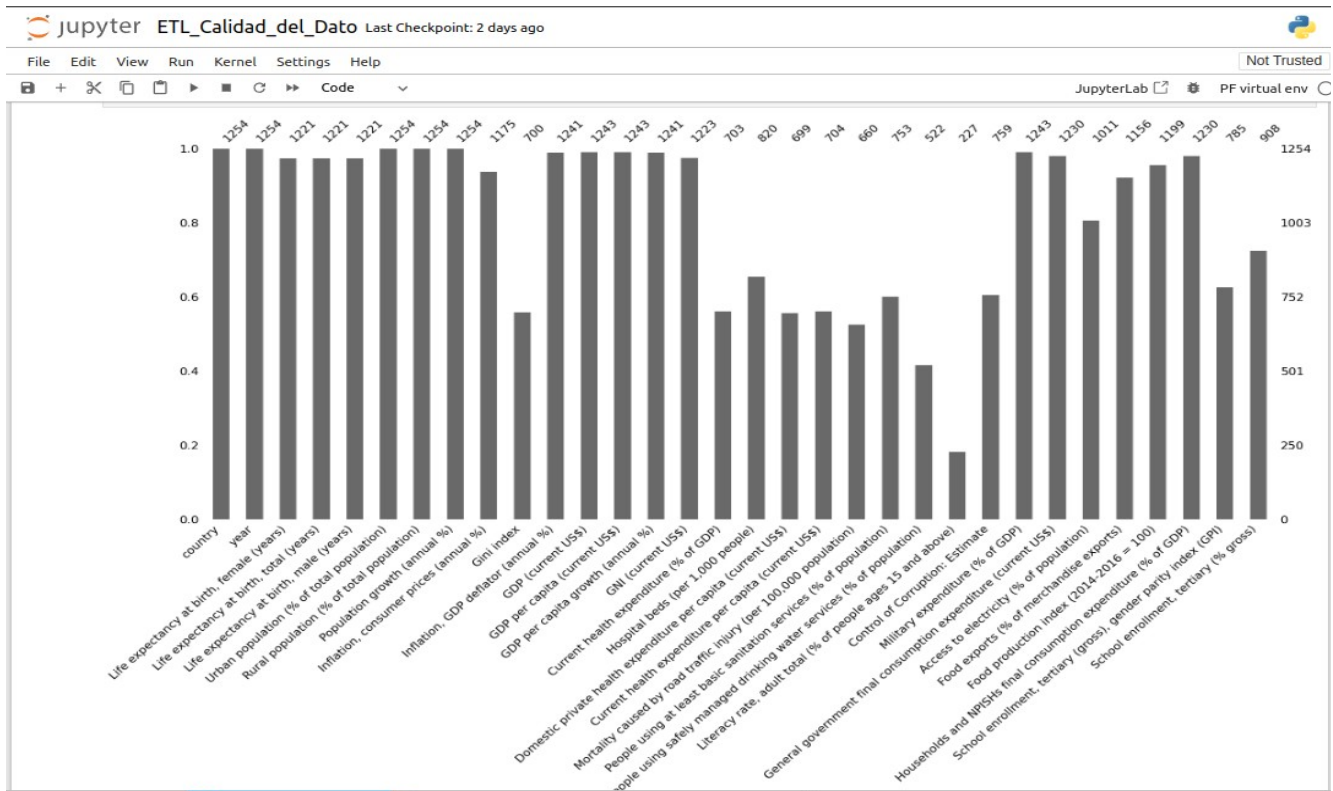
File Edit View Run Kernel Settings Help

+ - + Code

[80]: print("Valores nulos en porcentaje")
      print('*'*100)
      # Valores Nulos
      df.isnull().sum()*100/df.shape[0]

Valores nulos en porcentaje
*****
[80]: country                0.000000
      year                  0.000000
      Life expectancy at birth, female (years) 2.631579
      Life expectancy at birth, total (years) 2.631579
      Life expectancy at birth, male (years) 2.631579
      Urban population (% of total population) 0.000000
      Rural population (% of total population) 0.000000
      Population growth (annual %) 0.000000
      Inflation, consumer prices (annual %) 44.178628
      Gini index 1.036683
      Inflation, GDP deflator (annual %) 0.877193
      GDP (current US$) 0.877193
      GDP per capita (current US$) 1.036683
      GDP per capita growth (annual %) 2.472089
      GNI (current US$) 43.939394
      Current health expenditure (% of GDP) 34.609250
      Hospital beds (per 1,000 people) 44.258373
      Domestic private health expenditure per capita (current US$) 43.859649
      Current health expenditure per capita (current US$) 47.368421
      Mortality caused by road traffic injury (per 100,000 population) 39.952153
      People using at least basic sanitation services (% of population) 50.373206
      People using safely managed drinking water services (% of population) 81.897927
      Literacy rate, adult total (% of people ages 15 and above) 39.473684
      Control of Corruption: Estimate 0.877193
      Military expenditure (% of GDP) 1.913876
      General government final consumption expenditure (current US$) 19.377990
      Access to electricity (% of population) 7.814992
      Food exports (% of merchandise exports) 4.385965
      Food production index (2014-2016 = 100) 1.913876
      Households and NPISHs final consumption expenditure (% of GDP) 37.400319
      School enrollment, tertiary (gross), gender parity index (GPI) 27.591707
      School enrollment, tertiary (% gross)
      dtype: float64
```


4. Visualización de variables con valores faltantes y distribución de los mismos



5. Selección de indicadores con menos del 30% de datos nulos.

```
[84]: # Columnas a eliminar
drop_columns = [
    'Gini index',
    'Current health expenditure (% of GDP)',
    'Hospital beds (per 1,000 people)',
    'Domestic private health expenditure per capita (current US$)',
    'Current health expenditure per capita (current US$)',
    'Mortality caused by road traffic injury (per 100,000 population)',
    'People using at least basic sanitation services (% of population)',
    'People using safely managed drinking water services (% of population)',
    'Literacy rate, adult total (% of people ages 15 and above)',
    'Control of Corruption: Estimate',
    'School enrollment, tertiary (gross), gender parity index (GPI)'
]
```

6. Visualizar nueva estructura del dataframe.

JupyterLab ETL_Calidad_del_Dato Last Checkpoint: 2 days ago

File Edit View Run Kernel Settings Help

Not Trusted

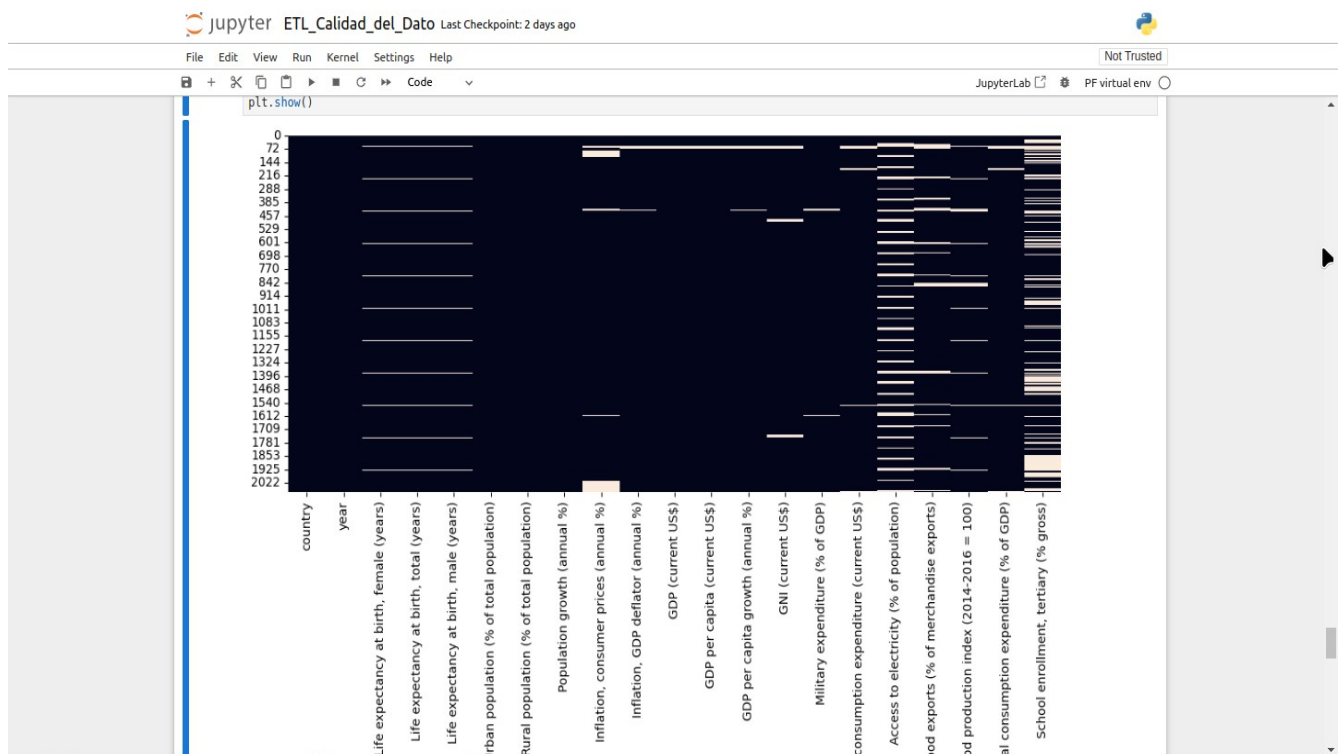
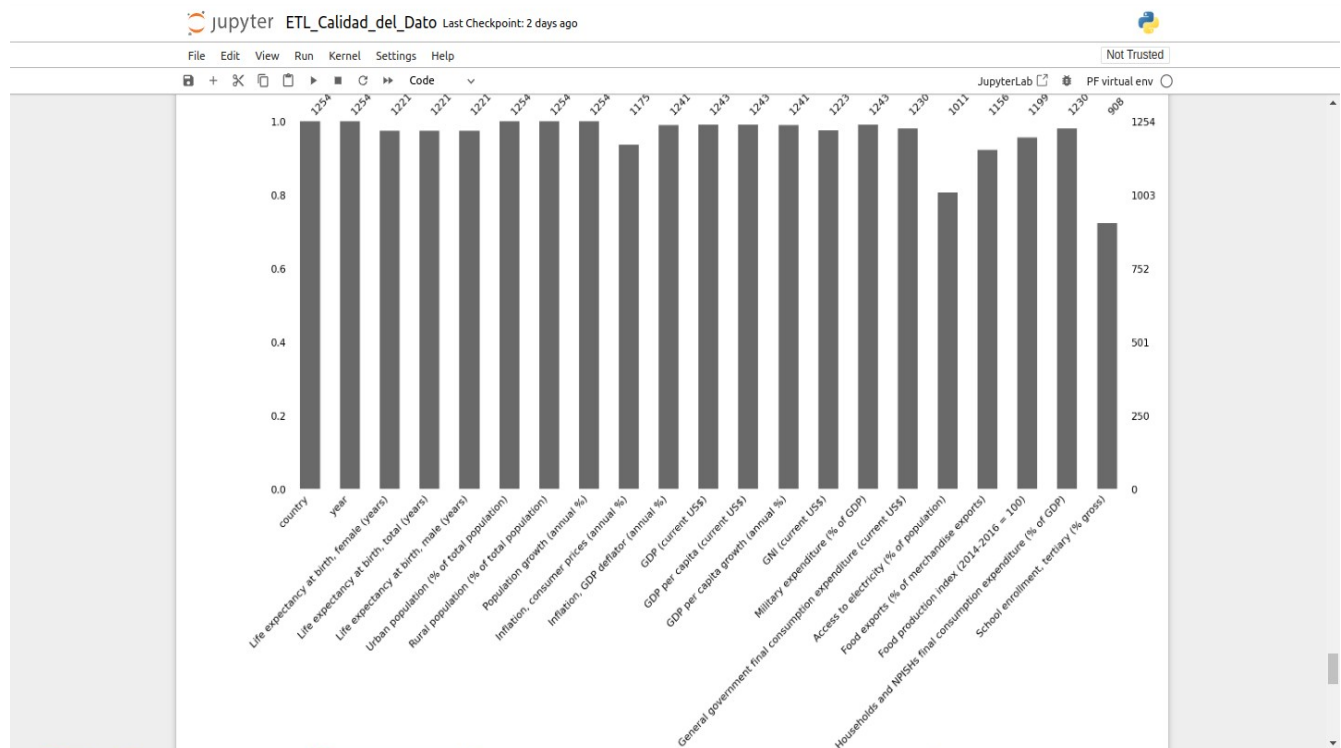
JupyterLab PF virtual env

[85]:

	country	year	Life expectancy at birth, female (years)	Life expectancy at birth, total (years)	Life expectancy at birth, male (years)	Urban population (% of total population)	Rural population (% of total population)	Population growth (annual %)	Inflation, consumer prices (annual %)	Inflation, GDP deflator (annual %)	GDP (current US\$)	GDP per capita (current US\$)	GDP per capita growth (annual %)	GN
0	ZAF	2022	NaN	NaN	NaN	68.335	31.665	0.841058	7.039727	5.054293	4.058697e+11	6776.480077	1.187663	3.97
1	ZAF	2021	64.999	62.341	59.458	67.847	32.153	0.998920	4.611672	6.219017	4.190156e+11	7055.055176	3.870315	4.10
2	ZAF	2020	67.964	65.252	62.178	67.354	32.646	1.223179	3.210036	5.692615	3.376196e+11	5741.641249	-7.481093	3.31
3	ZAF	2019	69.107	66.175	62.834	66.856	33.144	1.295074	4.120246	4.638081	3.885312e+11	6688.774746	-0.987175	3.78
4	ZAF	2018	68.740	65.674	62.203	66.355	33.645	1.225530	4.517165	3.745754	4.041589e+11	7048.508112	0.285736	3.92
...
2049	ARG	1989	74.735	71.425	68.090	86.613	13.387	1.487655	NaN	3046.091152	7.662973e+10	2382.338066	-8.527911	7.02
2050	ARG	1988	74.381	71.052	67.711	86.233	13.767	1.508087	NaN	381.246344	1.268902e+11	4004.009589	-2.570392	1.21
2051	ARG	1987	74.004	70.564	67.148	85.843	14.157	1.510974	NaN	127.539918	1.088109e+11	3485.690232	1.164690	1.03
2052	ARG	1986	73.638	70.119	66.655	85.445	14.555	1.511328	NaN	77.292237	1.058724e+11	3443.191431	4.561114	1.01
2053	ARG	1985	73.253	69.651	66.141	85.038	14.962	1.513403	NaN	607.447498	8.815089e+10	2910.508329	-6.613093	8.34

1254 rows × 15 columns

7. Se verifican las variables con datos faltantes una vez realizado la eliminación de las columnas con problemas y la nueva distribución



8. Evaluación de integridad de los datos

Se verifica si los datos recopilados tienen todos los campos requeridos, es decir si existe algún campo faltante o si existen registros duplicados. Se examina la consistencia de los datos en términos de formatos, unidades de medida

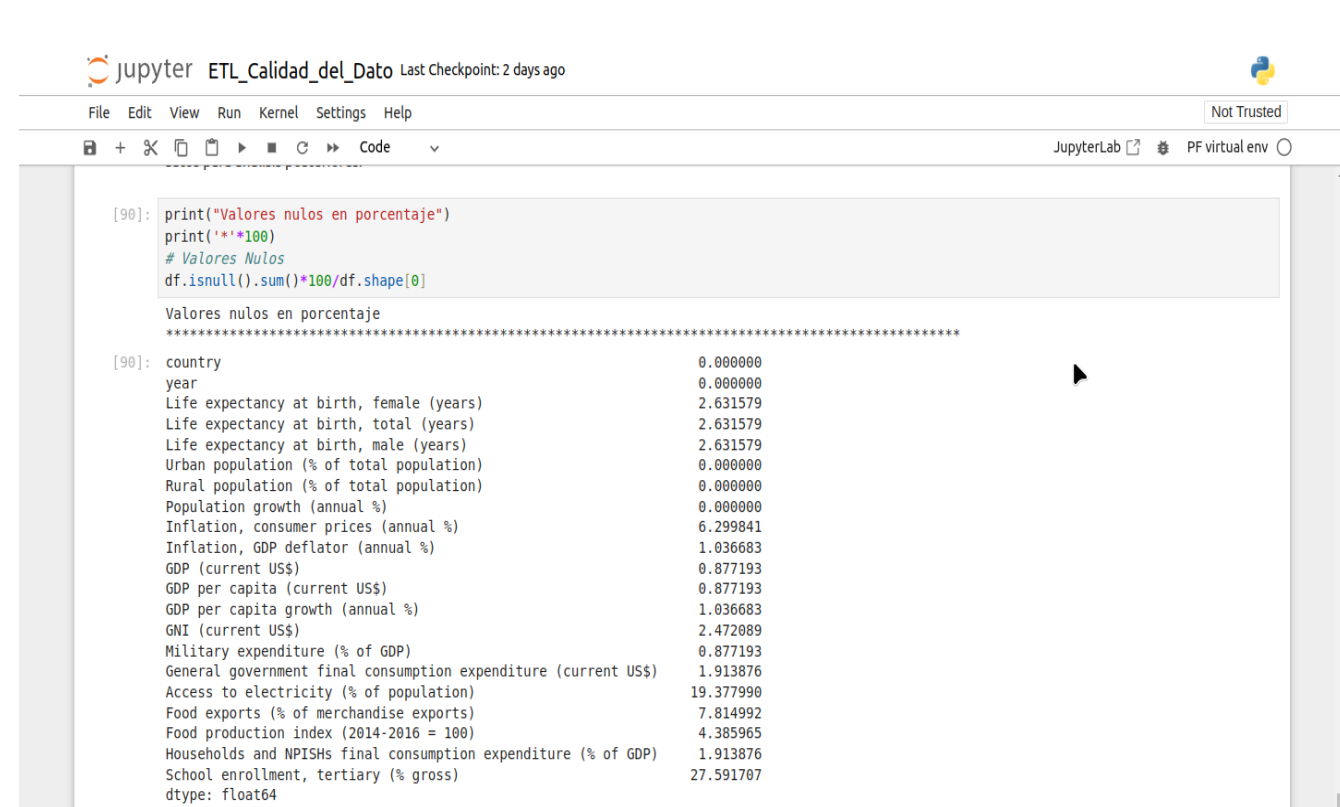
8.1. Verificación de registros repetidos y eliminación de los mismos de existir



```
[88]: # Verificar duplicados y eliminarlos si existen.
print(f'Tamaño del set antes de eliminar las filas repetidas: {df.shape}')
df.drop_duplicates(inplace=True)
print(f'Tamaño del set después de eliminar las filas repetidas: {df.shape}')
```

Tamaño del set antes de eliminar las filas repetidas: (1254, 21)
Tamaño del set después de eliminar las filas repetidas: (1254, 21)

8.2. Determinación de porcentaje de valores nulos por variable.



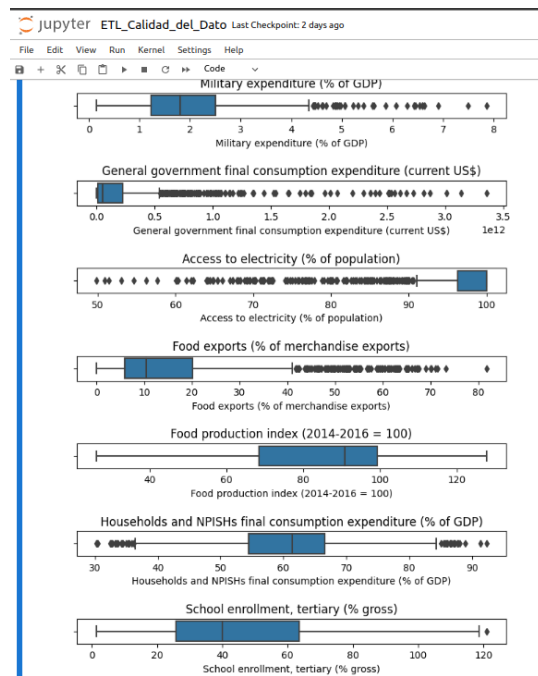
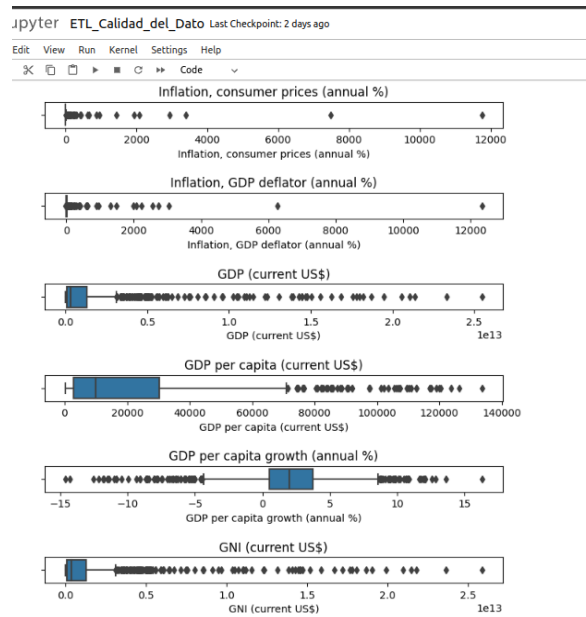
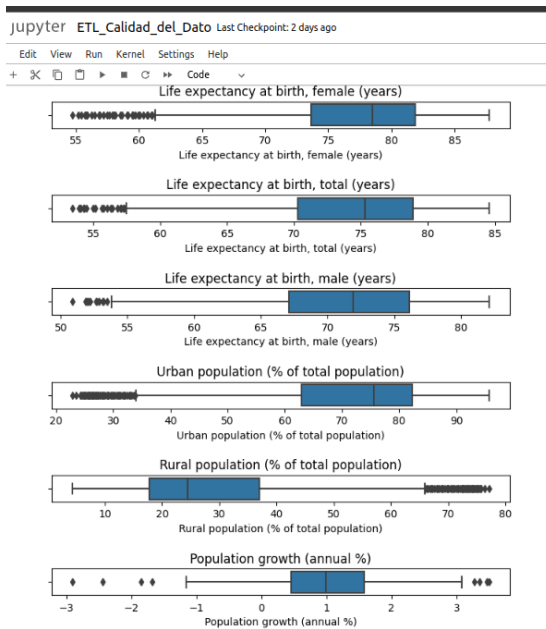
```
[90]: print("Valores nulos en porcentaje")
print('*'*100)
# Valores Nulos
df.isnull().sum()*100/df.shape[0]
```

Valores nulos en porcentaje

country	0.000000
year	0.000000
Life expectancy at birth, female (years)	2.631579
Life expectancy at birth, total (years)	2.631579
Life expectancy at birth, male (years)	2.631579
Urban population (% of total population)	0.000000
Rural population (% of total population)	0.000000
Population growth (annual %)	0.000000
Inflation, consumer prices (annual %)	6.299841
Inflation, GDP deflator (annual %)	1.036683
GDP (current US\$)	0.877193
GDP per capita (current US\$)	0.877193
GDP per capita growth (annual %)	1.036683
GNI (current US\$)	2.472089
Military expenditure (% of GDP)	0.877193
General government final consumption expenditure (current US\$)	1.913876
Access to electricity (% of population)	19.377990
Food exports (% of merchandise exports)	7.814992
Food production index (2014-2016 = 100)	4.385965
Households and NPISHs final consumption expenditure (% of GDP)	1.913876
School enrollment, tertiary (% gross)	27.591707

dtype: float64

8.3. Evaluación de valores extremos (outliers).



Observaciones:

En los gráficos anteriores se evidencian valores que se encuentran fuera de los parámetros normales los cuales se pueden tratar como outliers, sin embargo tomando en cuenta que la data **representa economías de distintos tamaños y fortaleza** la existencia de esos valores es factible, por lo que se decide mantener todo el conjunto de datos para análisis posteriores.