



Universidad de Buenos Aires

Facultad de Ciencias Exactas y Naturales

# Maestría en Explotación de Datos y Descubrimiento del Conocimiento

Taller de Tesis I – Entrega III

Grupo 2

Miguel Kiszkurno

# Tabla de Contenidos

<b>INTRODUCCIÓN.....</b>	<b>3</b>
<b>MARCO TEÓRICO.....</b>	<b>4</b>
RELEVAMIENTO DE TRABAJOS PREVIOS Y RELEVANTES .....	4
CONCEPTOS Y TÉCNICAS DE CIENCIA DE DATOS UTILIZADOS EN EL TRABAJO .....	5
ANÁLISIS DE SERIES TEMPORALES.....	5
<b>METODOLOGÍA .....</b>	<b>7</b>
PRESENTACIÓN Y DESCRIPCIÓN DE LOS DATOS UTILIZADOS .....	7
PREPROCESAMIENTO Y LIMPIEZA DE LOS DATOS.....	11
HERRAMIENTAS.....	11
ANÁLISIS EXPLORATORIO DE DATOS (AED) .....	12
<b>RESULTADOS Y DISCUSIÓN.....</b>	<b>22</b>
PRESENTACIÓN Y ANÁLISIS DE RESULTADOS OBTENIDOS .....	22
DISCUSIÓN DE LOS RESULTADOS Y SU RELEVANCIA .....	23
LIMITACIONES Y POSIBLES MEJoras .....	23
<b>CONCLUSIÓN .....</b>	<b>24</b>
RESUMEN DE LOS HALLAZGOS PRINCIPALES .....	24
CONCLUSIONES GENERALES Y SU RELACIÓN CON LOS OBJETIVOS DEL TRABAJO .....	24
RECOMENDACIONES PARA FUTUROS TRABAJOS .....	24
<b>BIBLIOGRAFÍA Y REFERENCIAS .....</b>	<b>27</b>
<b>ANEXOS.....</b>	<b>28</b>
ANEXO 1: CÓDIGO FUENTE UTILIZADO EN EL ANÁLISIS .....	28
ANEXO 2: DESCRIPCIÓN DE ÁREAS CLAVE .....	28

## Introducción

El desarrollo socioeconómico de los países está determinado por una interacción compleja de factores. Este estudio se enfoca en identificar y comparar los factores más influyentes sobre las tendencias de crecimiento de Argentina, utilizando como referencia países con desarrollos similares tanto en Latinoamérica como en otras regiones seleccionadas.

El caso argentino es estudiado a lo largo de todo el mundo por sus particularidades. Gonzalez (2012) utiliza modelos de series temporales para estimar los efectos de las privatizaciones y la inversión directa extranjera sobre el crecimiento del país en el periodo 1971–2000; mientras que Taylor (1994) hace un estudio histórico de nuestra performance económica caracterizando las siguientes etapas: pre-1913, 1913-1930s y 1930s-1950s. Asimismo, Weisskoff (1970) intenta responder la pregunta de si el crecimiento económico en países en desarrollo llevó a inequidades en la distribución del ingreso, analizando en detalle los casos de Puerto Rico, Argentina y México.

Con una perspectiva práctica, buscamos proporcionar datos relevantes que puedan servir como referencia en la elaboración de estudios futuros. Nos centramos en comprender: ¿Qué indicadores que dan cuenta de factores socioeconómicos han influido significativamente en el desarrollo de Argentina durante las últimas seis décadas y cuál es su correlato en países con desarrollos comparables?

# Marco teórico

## Relevamiento de trabajos previos y relevantes

Los estudios en la materia de la que trata este trabajo son cuantiosos y con diversos enfoques, abordando desde la influencia del gasto público hasta la distribución del ingreso y el impacto de la inversión extranjera. A continuación, se presentan algunos de los trabajos más relevantes que sirven como base y contexto para nuestra investigación sobre el desarrollo socioeconómico de Argentina.

Estas referencias son esenciales para contextualizar y apoyar nuestro análisis del desarrollo socioeconómico de Argentina, proporcionando marcos teóricos y evidencias empíricas que nos ayudan a entender mejor los factores que influyen en el crecimiento económico y su distribución.

### Análisis del Gasto Público y el Crecimiento Económico

El estudio de Landau (1983) examina la relación entre la proporción del gasto de consumo del gobierno en el PIB y la tasa de crecimiento del PIB real per cápita en más de 100 países. Este trabajo es relevante para nuestra investigación ya que proporciona una base metodológica sólida para analizar cómo los diferentes tipos de gasto gubernamental pueden impactar el crecimiento económico. En nuestro estudio del desarrollo socioeconómico de Argentina, entender estos efectos es crucial para identificar los factores que pueden promover o inhibir el crecimiento económico en diferentes contextos regionales y temporales.

### Distribución del Ingreso y Crecimiento Económico

Weisskoff (1970) investiga si el crecimiento económico en países en desarrollo ha conducido a inequidades en la distribución del ingreso, enfocándose en Puerto Rico, Argentina y México. Este análisis es fundamental para entender las dinámicas de distribución del ingreso en el contexto del crecimiento económico, un tema central en nuestro estudio del desarrollo socioeconómico de Argentina. Weisskoff concluye que, aunque el crecimiento económico puede aumentar la riqueza nacional, a menudo exacerba las desigualdades de ingresos. Este aspecto es crítico para nuestra investigación, ya que buscamos comprender cómo las políticas económicas pueden equilibrar el crecimiento con la equidad.

### Urbanización y Desarrollo en Países en Desarrollo

Timberlake y Kentor (1984) exploran los determinantes estructurales de la urbanización periférica y sus efectos en el desarrollo nacional. Este trabajo aporta una perspectiva sobre cómo la urbanización puede influir en el desarrollo económico y social, un aspecto relevante para nuestro análisis de Argentina en comparación con otros países. El estudio resalta que la urbanización no planificada puede llevar a problemas económicos y sociales significativos, lo que subraya la importancia de políticas urbanas bien diseñadas para promover un desarrollo equilibrado. Esto es particularmente pertinente para Argentina, donde la urbanización ha tenido impactos mixtos en el desarrollo regional.

### Influencia de la Inversión Extranjera Directa

El trabajo de Newfarmer y Mueller (1975) analiza el poder económico y no económico de las corporaciones multinacionales en Brasil y México. Este estudio es relevante ya que muestra cómo la inversión extranjera directa puede afectar las economías locales, tanto positivamente como negativamente. Este análisis es útil para entender el impacto de la inversión extranjera en el crecimiento económico de Argentina y cómo puede ser gestionada para maximizar sus beneficios y minimizar sus efectos adversos. En nuestra investigación, evaluar el papel de la inversión extranjera es esencial para comprender las dinámicas de crecimiento en un contexto globalizado.

# Conceptos y técnicas de ciencia de datos utilizados en el trabajo

## Técnicas de Clustering

El clustering es una técnica de aprendizaje no supervisado que se utiliza para agrupar un conjunto de objetos de manera que los objetos dentro de un mismo grupo (o clúster) sean más similares entre sí que con los objetos de otros grupos. Las técnicas de clustering son fundamentales en el análisis exploratorio de datos, ayudando a descubrir patrones y estructuras ocultas en los datos.

### K-means

Una de las técnicas de clustering más utilizadas es K-means. Este algoritmo partitiona los datos en K clústeres, donde cada punto pertenece al clúster con el centroide más cercano. El objetivo de K-means es minimizar la variación dentro de cada clúster. Los centroides de los clústeres se actualizan iterativamente hasta que los cambios sean mínimos. Este método es especialmente útil por su simplicidad y eficiencia, aunque puede verse afectado por la elección inicial de los centroides y es sensible a los valores atípicos.

### Aplicación de K-means en el Análisis de Indicadores Socioeconómicos

En nuestro análisis, hemos utilizado K-means para clasificar y caracterizar a diferentes países basándonos en una serie de indicadores socioeconómicos disponibles en el dataset del Banco Mundial. Comenzamos con un experimento inicial utilizando el Ingreso per cápita para el año 2022, partiendo de la clasificación en cuatro categorías de ingreso según lo define el Banco Mundial. Este análisis inicial nos permitió establecer una línea de base para la clasificación de los países.

Posteriormente, realizamos varios experimentos adicionales en diferentes años y con distintos indicadores socioeconómicos para identificar diferencias significativas entre las nuevas categorizaciones y la realizada por el Banco Mundial. Esta exploración nos permitió observar cómo varían las características socioeconómicas de los países a lo largo del tiempo y cómo estas variaciones pueden influir en su clasificación.

### Evaluación de los Clústeres

Para evaluar la calidad de los clústeres resultantes, utilizamos el coeficiente de silhouette y la inercia. El coeficiente de silhouette mide la similitud entre los puntos dentro del mismo clúster en comparación con los puntos fuera del clúster, proporcionando una medida de la cohesión y la separación de los clústeres. La inercia, por otro lado, mide la suma de las distancias al cuadrado entre cada punto y el centroide de su clúster, con el objetivo de minimizar esta suma para mejorar la compacidad de los clústeres.

## Análisis de series temporales

### Modelos ARIMA

ARIMA (AutoRegressive Integrated Moving Average) es una clase de modelos utilizada para analizar y predecir series temporales. ARIMA combina tres componentes principales: autorregresivo (AR), diferenciación (I) y promedio móvil (MA). El componente AR utiliza la relación entre una observación y un número de observaciones rezagadas anteriores. El componente I se aplica para hacer que la serie sea estacionaria, eliminando tendencias y estacionalidades. Finalmente, el componente MA modela el error de la predicción como una combinación lineal de errores pasados.

### Aplicación de ARIMA en el Índice de Ingreso Nacional Bruto (GNI)

Para nuestro análisis, utilizamos el modelo ARIMA para predecir el Índice de Ingreso Nacional Bruto (GNI) de diferentes países. La serie temporal del GNI es crucial para entender las tendencias económicas y planificar políticas futuras. La previsión del GNI mediante ARIMA nos permitirá identificar patrones subyacentes y hacer predicciones basadas en datos históricos.

El proceso para aplicar ARIMA al GNI implica los siguientes pasos:

- **Visualización y análisis exploratorio de la serie temporal:** Comenzamos visualizando la serie temporal del GNI para identificar cualquier tendencia, estacionalidad o patrones.
- **Diferenciación para estacionarizar la serie:** Si la serie no es estacionaria, aplicamos diferenciación para eliminar la tendencia.
- **Selección de parámetros ARIMA ( $p$ ,  $d$ ,  $q$ ):** Utilizamos métodos como el gráfico ACF (Autocorrelation Function) y PACF (Partial Autocorrelation Function) para identificar los valores adecuados de  $p$  (orden autorregresivo),  $d$  (diferenciación) y  $q$  (orden de promedio móvil).
- **Ajuste del modelo ARIMA:** Ajustamos el modelo ARIMA a la serie temporal del GNI utilizando los parámetros seleccionados.
- **Evaluación del modelo:** Evaluamos el modelo utilizando métricas de error y validación cruzada.
- **Predicción futura:** Utilizamos el modelo ajustado para hacer predicciones futuras del GNI.

## Análisis de Componentes Principales (PCA)

El Análisis de Componentes Principales (PCA) es una técnica de reducción de dimensionalidad que se utiliza para transformar un conjunto de datos de alta dimensión en un espacio de menor dimensión. Este proceso se lleva a cabo identificando las direcciones (componentes principales) en las que varía más el conjunto de datos. PCA permite conservar la mayor cantidad posible de varianza en los datos originales mientras reduce el número de dimensiones, lo que facilita la visualización y el análisis. Los componentes principales son ortogonales entre sí y están ordenados de manera que el primer componente principal captura la mayor cantidad de varianza, el segundo componente captura la segunda mayor cantidad de varianza, y así sucesivamente.

Se utilizó PCA para visualizar la información de los clusters generados a partir de diversos subconjuntos de indicadores. Dado que los datos originales contienen múltiples indicadores como tasas de empleo, participación laboral femenina, Producto Nacional Bruto, entre otros, que pueden ser difíciles de analizar en un espacio multidimensional, PCA nos permite proyectar estos datos en un espacio bidimensional. Esta transformación facilita la visualización al condensar la información de múltiples variables en solo dos componentes principales que retienen la mayor parte de la variabilidad de los datos originales.

La reducción de dimensionalidad mediante PCA es particularmente útil en la visualización de clusters. Al reducir los datos a dos dimensiones, podemos crear gráficos de dispersión donde cada punto representa un país, coloreado según el cluster al que pertenece. Esto no solo nos permite observar la distribución y separación de los clusters de manera clara, sino que también ayuda a identificar patrones y relaciones entre los países dentro de los clusters.

## Metodología

### Presentación y descripción de los datos utilizados

El Banco Mundial categoriza sus indicadores en diversas áreas para proporcionar una visión integral del desarrollo socioeconómico de los países. Estas áreas abarcan desde la agricultura y el desarrollo rural, hasta la eficacia de la ayuda, el cambio climático y el crecimiento económico. Indicadores como la producción agrícola, las emisiones de CO<sub>2</sub> y el Producto Interno Bruto (PIB) son esenciales para evaluar el progreso y los desafíos en estos sectores. Además, incluye indicadores en sectores cruciales como la educación, la salud y la infraestructura, que miden aspectos como la tasa de matrícula en educación primaria, la esperanza de vida al nacer y el acceso a electricidad. Estos indicadores permiten un análisis detallado de las políticas y estrategias necesarias para fomentar un desarrollo sostenible y equitativo.

En términos de equidad y sostenibilidad, las áreas de género, medio ambiente y pobreza son particularmente relevantes. Los indicadores de participación laboral femenina, calidad del aire y tasa de pobreza internacional ayudan a comprender mejor las desigualdades y los impactos ambientales que afectan el desarrollo. Asimismo, las categorías de sector financiero, sector público y protección social y trabajo ofrecen datos sobre la estabilidad económica y la seguridad social, con indicadores como el crédito doméstico al sector privado, el gasto público como porcentaje del PIB y la tasa de desempleo. En conjunto, estos indicadores proporcionan un marco completo para evaluar el desarrollo económico y social de los países, permitiendo la formulación de políticas más efectivas y focalizadas.

Hay un total de 1,463 indicadores, organizados en 20 áreas clave (ver Tabla 1 en los anexos de este documento), como Educación; Ciencia y Tecnología; y Crecimiento Económico; cubriendo el período de 1960 a 2022. Los conjuntos de datos están disponibles para su descarga en el siguiente [enlace](#).

En el sitio web del Banco Mundial, los datos se almacenan y acceden de manera individual; es decir, cada indicador se encuentra en su propio archivo Excel con información correspondiente a todos los países y todos los años. Un mismo indicador puede pertenecer a varias áreas.

En la Figura 1 se muestra las áreas claves junto con las cantidades de indicadores de cada una.

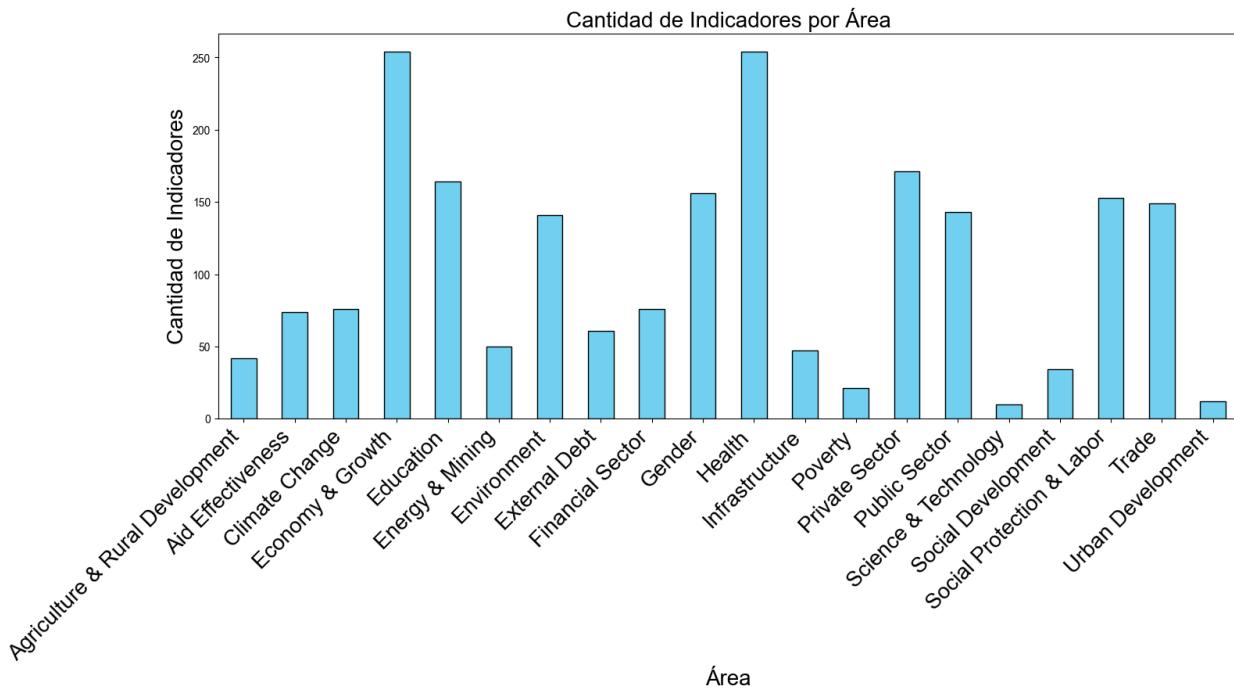


Figura 1: Cantidad de indicadores por Área agrupados por áreas claves según el Banco Mundial. Son significativas las cantidades de indicadores en el área de Economy & Growth y Health

Dentro del dataset, los países se categorizan por dos criterios: geográfico y económico. Las regiones en las que distingue el Banco Mundial a los países son las siguientes (Figura 2):

- Latin America & Caribbean
- North America
- Middle East & North Africa,
- Sub-Saharan Africa,
- Europe & Central Asia,
- East Asia & Pacific,
- South Asia

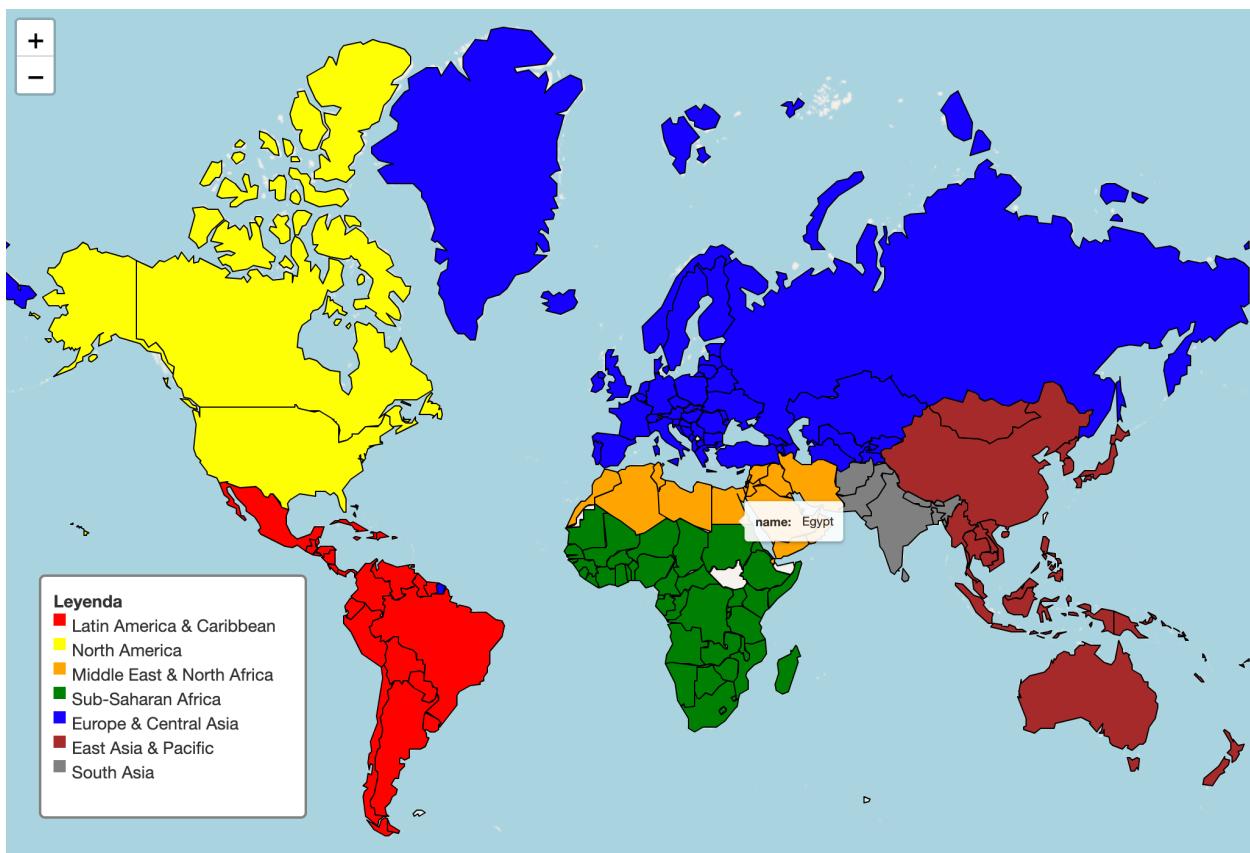


Figura 2: Mapa de los países identificados por color según su región.

El Banco Mundial categoriza a los países en función de su GNI, que es la suma del valor agregado por todos los productores residentes más cualquier impuesto sobre productos (menos subsidios) no incluido en la valoración de la producción, más los ingresos netos primarios (compensación de empleados e ingresos por propiedad) recibidos del extranjero. Los datos se expresan en dólares estadounidenses corrientes. Las categorías son las siguientes:

- **Low income:** “Low-income economies are those in which 2022 GNI per capita was \$1,135 or less”.
- **Lower middle income:** “Lower-middle-income economies are those in which 2022 GNI per capita was between \$1,136 and \$4,465”.
- **Upper middle income:** “Upper-middle-income economies are those in which 2022 GNI per capita was between \$4,466 and \$13,845”.

- **High income:** “High-income economies are those in which 2022 GNI per capita was more than \$13,845”.

En la Figura 3 puede verse el mapa mundial con los países coloreados según su categoría de ingresos. La Figura 4 muestra los países agrupados por geografía y categoría de ingresos. La Figura 5 muestra las cantidades totales de países por cada categoría de ingreso.

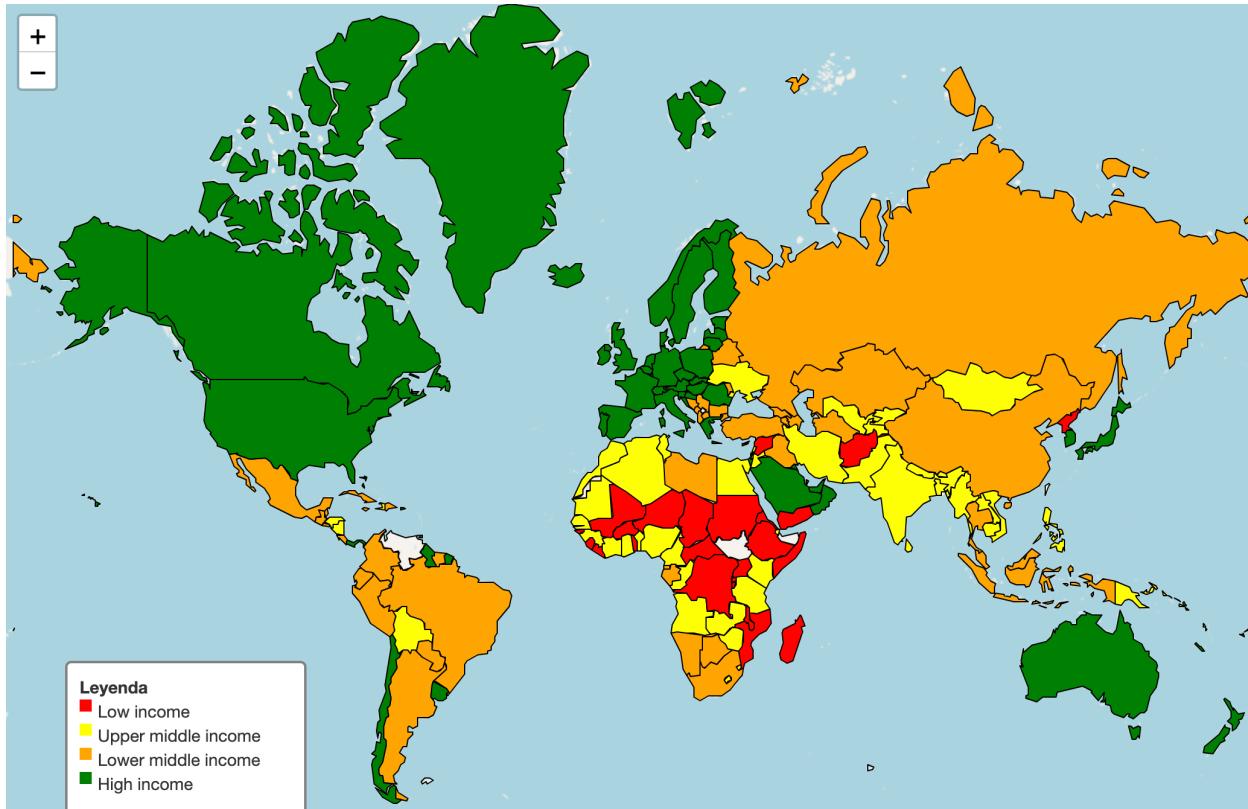


Figura 3: Los países según su nivel de ingreso dentro del mapa.

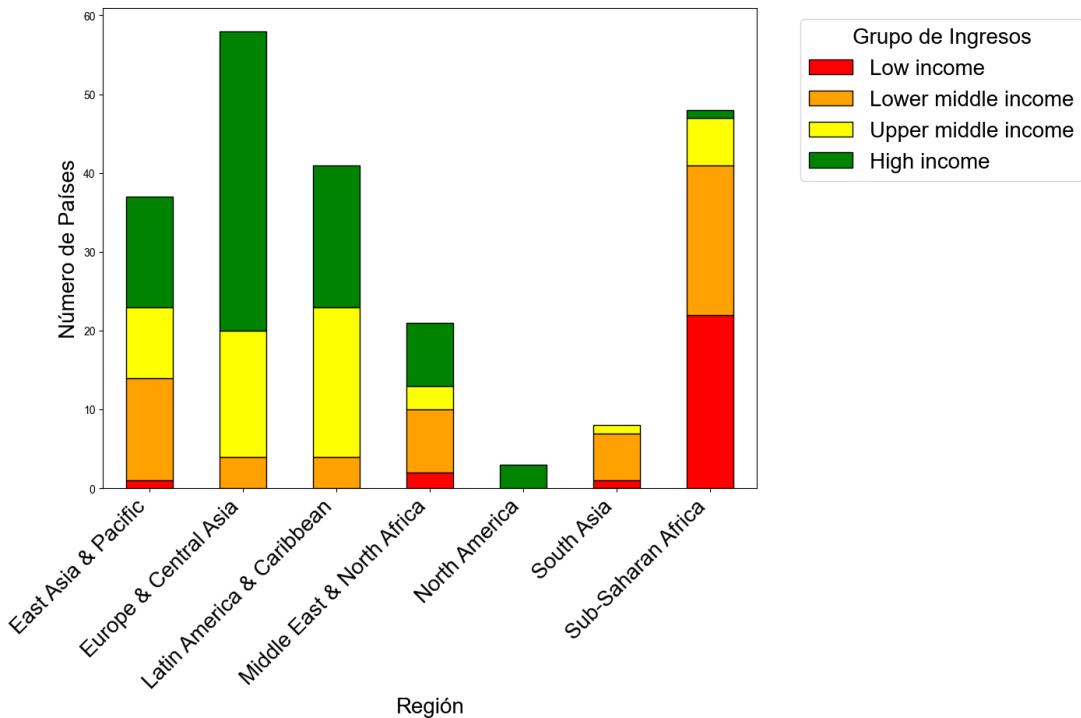


Figura 4: Muestra la cantidad de países en función de su categoría de ingreso agrupando por la región a la que pertenecen

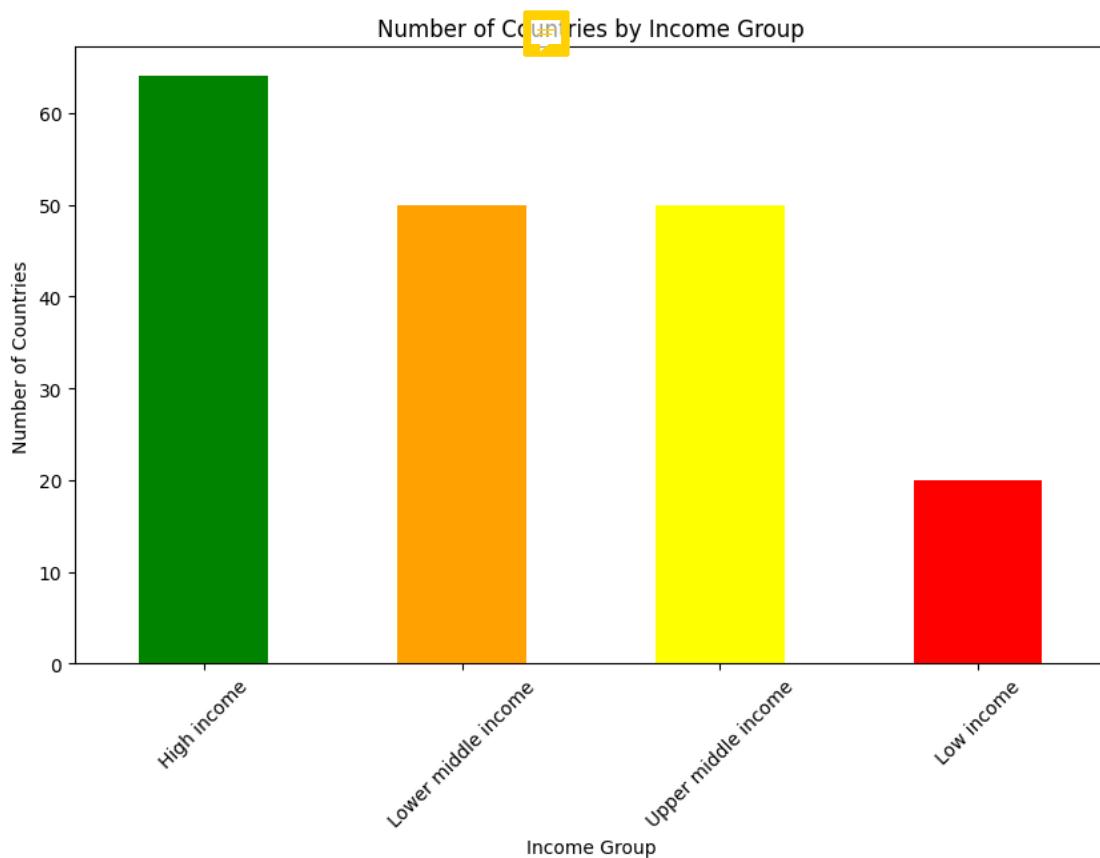


Figura 5: Cantidad total de países por categoría de ingreso

## Preprocesamiento y limpieza de los datos

Los datos se encuentran agrupados en distintas planillas. Cada indicador tiene su propio archivo, donde se encuentran los países como cabecera de las filas y los años como cabecera de la columna.

Es por esto por lo que para generar datasets fácilmente utilizables, hubo que realizar las siguientes tareas de preprocesamiento:

1. Descarga automática: Utilizando Selenium, se realizaron las descargas automáticas de todos los archivos en formato xls.
  - a. Estos archivos se convirtieron a CSV para facilitar su manipulación<sup>1</sup>
2. Generación de metadata: Luego de bajar los archivos, estos se procesan y se genera archivo maestro con la metadata necesaria para el estudio. Esta metainformación incluye:
  - a. Catálogo de indicadores: Con su código, nombre, descripción y ruta de acceso al archivo de origen dentro del repositorio de código.
  - b. Catálogo de indicadores por área clave: Cada una de las áreas clave junto con sus indicadores y su dirección url.
  - c. Catálogo de países: Los códigos, nombres, nivel de ingreso, agrupamiento geográfico y comentarios de creadores de los datos
3. Agrupamiento: Para facilitar el uso y manipulación de los datos, se procesaron los archivos individuales para generar un consolidado por área, que agrupa a todos los indicadores de esa área para todos los países.
4. Cálculo de Métricas: Dada la cantidad de archivos e indicadores, el análisis exploratorio se hace extremadamente complejo y costoso. Es por esto se automatizó el proceso de cálculo de métricas. Esto permitió analizar las cantidades de datos faltantes para los distintos experimentos.

## Herramientas

Tanto para la descarga y preprocesamiento de los datos, como para el análisis exploratorio, y el modelado, utilizaremos Python corriendo sobre la IDE Visual Studio Code.

Dentro de Python usaremos las siguientes bibliotecas y librerías:

- Selenium: Se utiliza para automatizar la descarga de datos desde la web.
- OpenPyXL: Se utiliza para manipular los archivos Excel durante el preprocesamiento de datos.
- Pandas y NumPy: Se utilizan para manipulación y análisis de datos.
- Matplotlib y Seaborn: Se utilizan para la visualización de datos.
- PDFPages de Matplotlib: Para generar reportes en PDF con gráficos de los análisis realizados.
- Holoviews y bokeh: Se utilizan para la visualización dinámica de datos
- Scikit-learn: Se utiliza para aplicar las técnicas de clustering en la etapa de modelado
  - StandardScaler
  - KMeans
  - silhouette\_score
- Statsmodels: Se utiliza para el análisis de series temporales, incluyendo modelos ARIMA.
- Jupyter Notebooks: Se utiliza para la etapa de análisis exploratorio. Permite no solo ejecutar código sino también documentar y presentar los análisis de manera interactiva.

Adicionalmente a las librerías de Python, se utilizaron los siguientes programas de oficina:

- Microsoft Excel: Para la gestión inicial de los datos descargados y la generación de la metadata necesaria.
- Microsoft Word: Para la elaboración de este documento

<sup>1</sup> El World Bank también provee los datos en formato CSV, no obstante, hacia más compleja la tarea de descarga ya que por cada indicador requería descargar más de un csv (uno para el indicador y otros dos con metadata)

## Análisis exploratorio de datos (AED)

### Datos faltantes

Dado el origen de los datos, se decidió aceptarlos como válidos. Entonces, el análisis exploratorio se basa principalmente en entender con la mayor precisión posible cuáles son los datos que faltan, y buscar estrategias que permitan imputarlos o recortarlos para poder aplicar las técnicas de clustering y de análisis de series temporales propuestas.

Como puede verse en la Figura 6, el volumen de datos hace difícil la tarea de visualizarlos todos juntos.

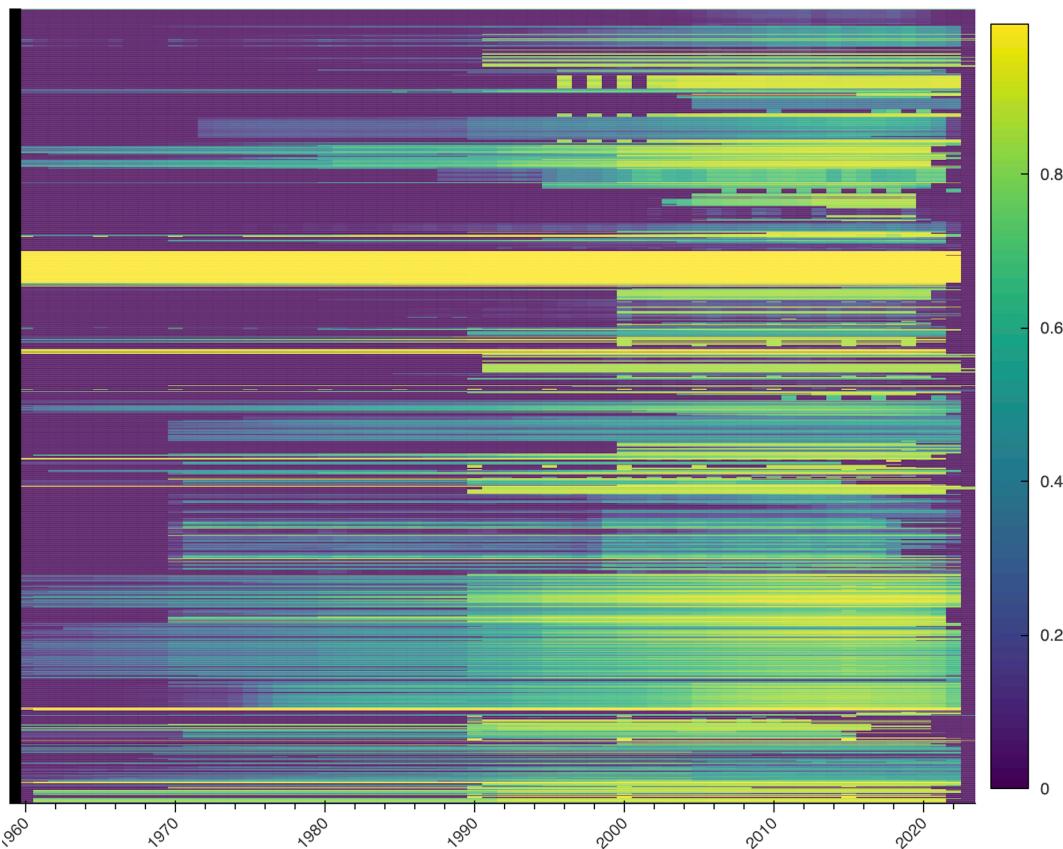


Figura 6: Heatmap de datos faltantes. En el eje de las x se visualizan los años, mientras que en el de las y se ven los indicadores. Las cajas del heatmap presentan la cantidad de datos que tienen valores distintos de nulo. Se removieron las etiquetas para facilitar la comprensión. Aun sin etiquetas puede apreciarse la alta taza de datos faltantes

En la Figura 7 puede verse un recorte de los datos a un conjunto de países Latinoamérica: Argentina, Bolivia, Chile, Colombia, Ecuador, México y Uruguay, pero esta vez agrupando los indicadores en las áreas antes descritas.

Aquí se visualiza el porcentaje de datos existentes sobre datos totales para cada país para cada área de interés (agrupando todos los indicadores de dicha área para todos los años). Hay escasas áreas donde este porcentaje supere el 80%. En particular, Argentina tiene un máximo de 72% en Economy & Growth.

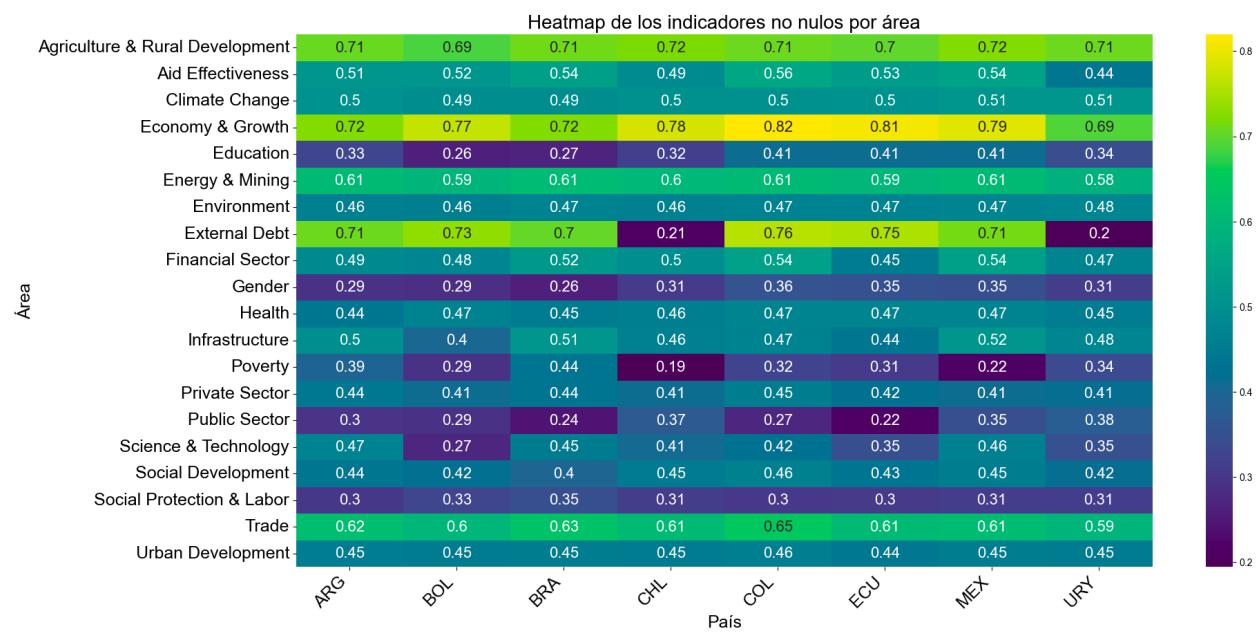


Figura 7: Heatmap con los datos faltantes para un conjunto de países seleccionados agrupando los indicadores por área.

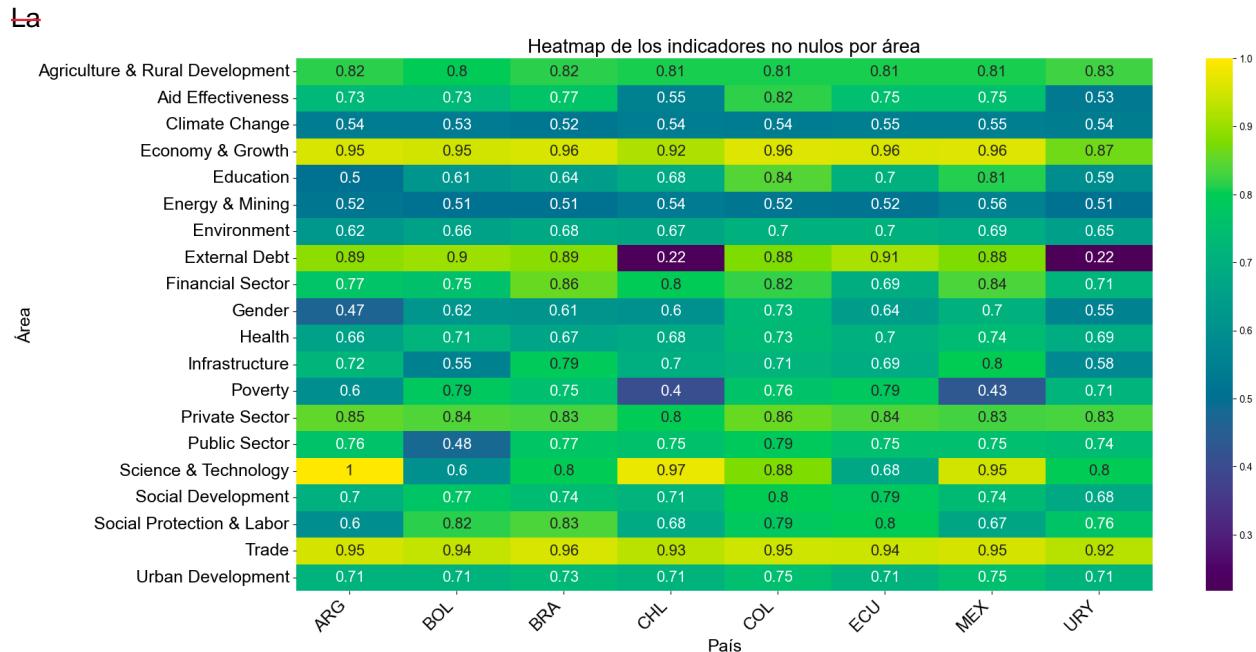


Figura 8, muestra el mismo gráfico en el periodo 2015-2018. La cantidad de datos faltantes disminuye notablemente, alcanzando picos de no nulos del 96% en algunos casos.

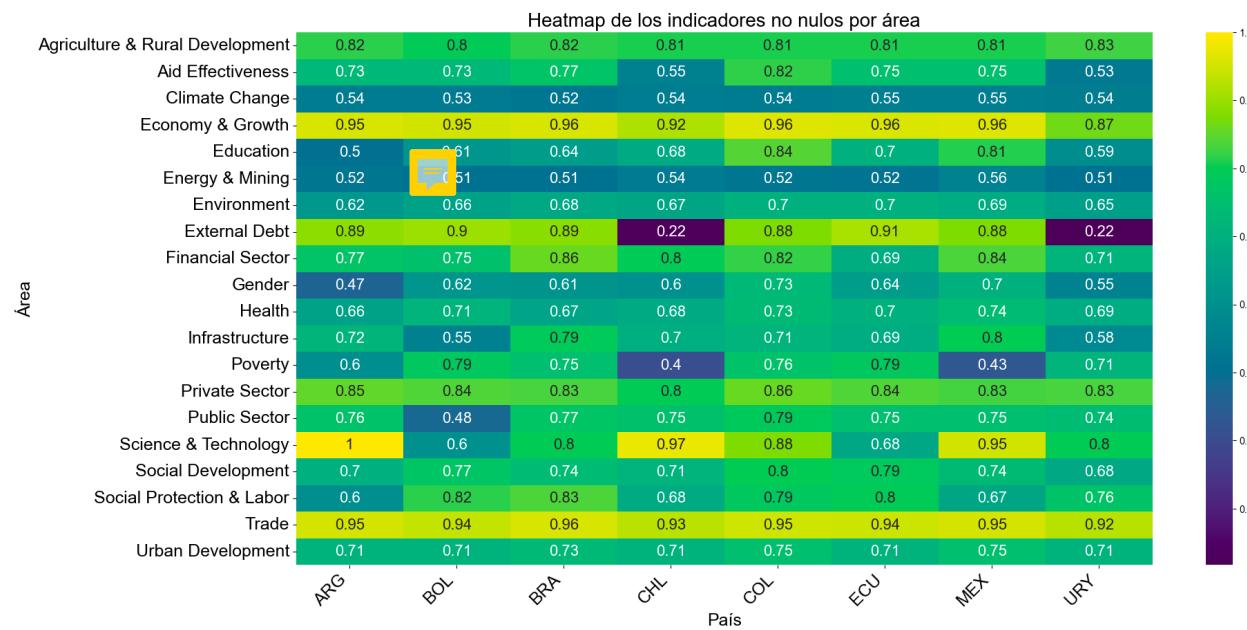


Figura 8: Datos faltantes para conjunto de países seleccionados (años 2015 a 2018)

Analizando más específicamente el área de Economy & Growth, puede verse que en el periodo 1994-2018 es cuando más datos se tienen.

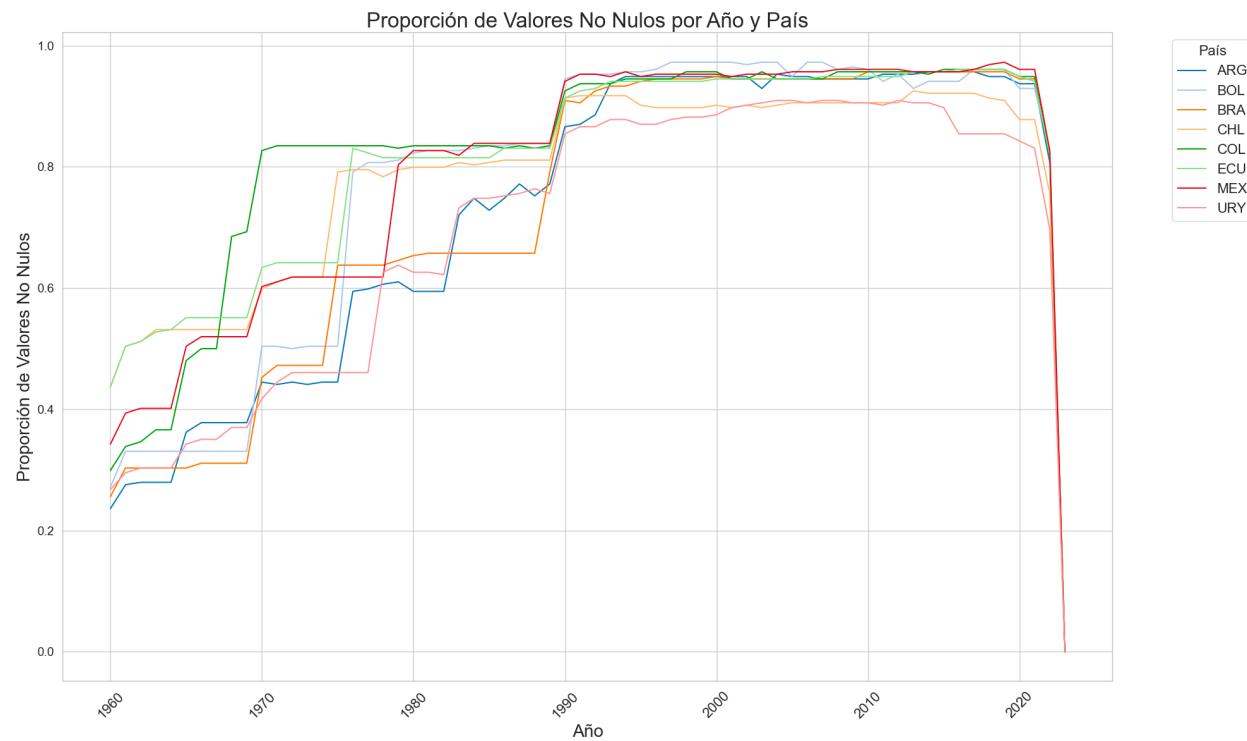


Figura 9: Cantidad de valores no nulos por año para el área de Economy &amp; Growth

Si bien hay una amplia cantidad de datos faltante, es posible plantear estrategias de mitigación que permitan llevar adelante el análisis propuesto en este trabajo.



## Categorización de los países (GNI per cápita)

En la siguiente figura pueden verse todos los indicadores relacionados con el GNI y su taza de nulos. En particular, nos enfocamos en el NY.GNP.MKTP.CD, que refleja el GNI total de los países. En este caso, se observa que para el 2022 hay 29 datos faltantes. Esto no invalida el análisis posterior, pero de alguna manera limita los resultados.

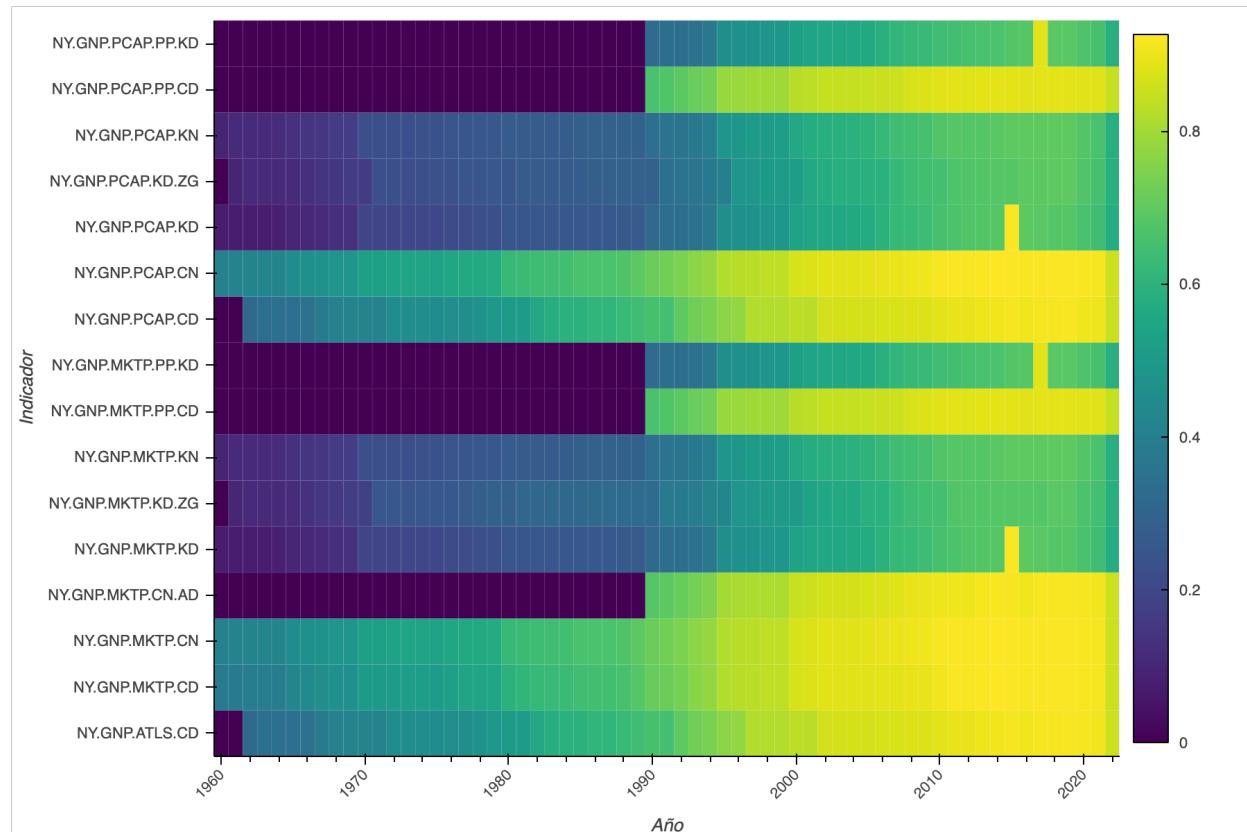


Figura 10: Heatmap con todos los indicadores relacionados con el GNI y su taza de nulos distribuidos a lo largo de los años.

La categorización propuesta por World Bank descrita anteriormente muestra una inusual cantidad de países en la categoría más alta (“High Income”) y una muy baja cantidad de países en la categoría de menor ingreso (“Low Income”). Esto es de alguna manera anti-intuitivo. Asimismo, puede verse que los países de mayores ingresos podrían considerarse Outliers según se muestra en el Boxplot de la Figura 11.

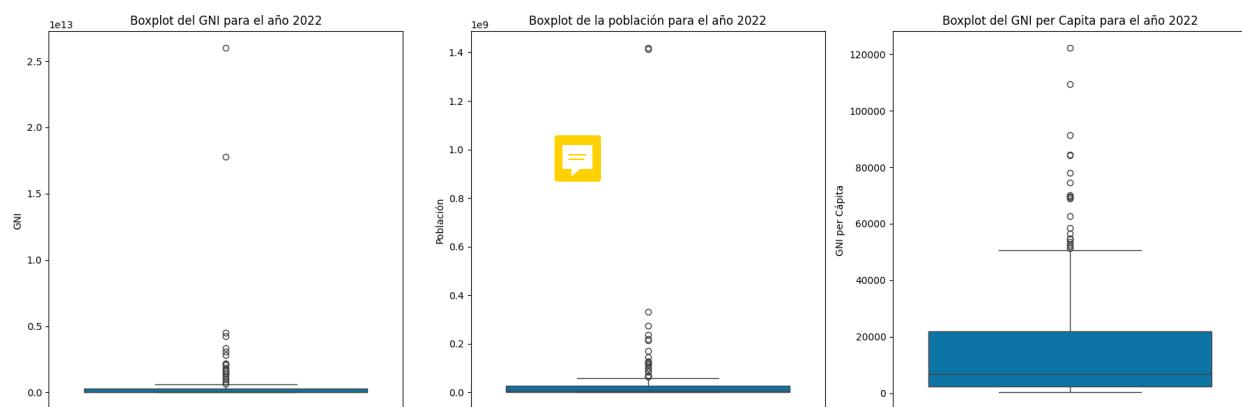


Figura 11: Boxplots de los indicadores asociados al GNI para el año 2022: (a) GNI total, (b) Población total, (c) GNI per cápita. Los potenciales outliers que se ven en el grafico son: (1) Bermuda, (2) Noruega, y (3) Suiza. Mientras que el mínimo es Burundi.

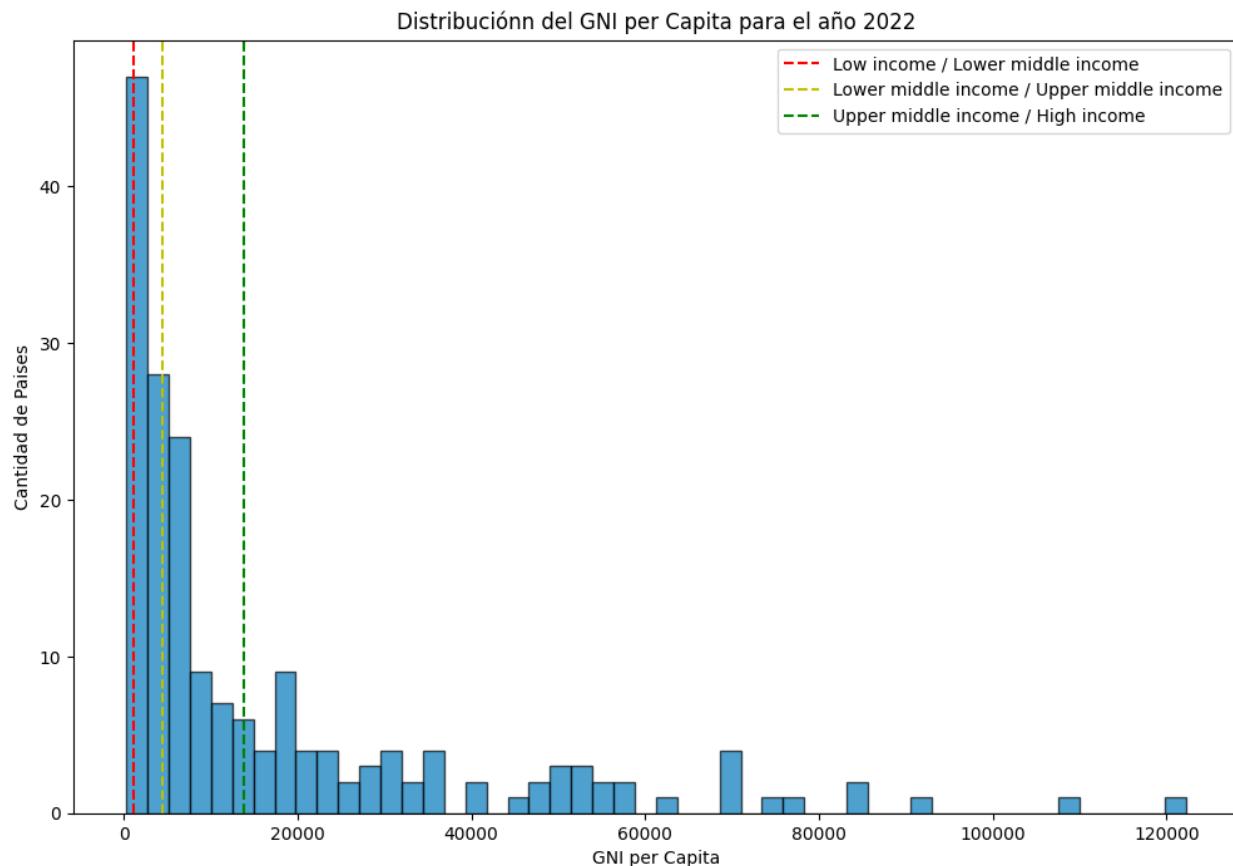


Figura 12: Distribución de los países en función de su grupo de ingreso. Las líneas amarilla, verde y roja marcan los límites a partir de los cuales los países se consideran Lower Middle Income, Upper middle income y High Income respectivamente.

En la siguiente gráfica se puede observar la distribución del GNI per cápita para el año 2022 según la categorización del Banco Mundial. Es notable la cantidad de países con ingresos altos, representados en verde, que se destacan significativamente por encima del resto. Por otro lado, los países de ingresos bajos, indicados en rojo, parecen mezclarse más entre sí, mostrando menos dispersión y situándose en la parte inferior del gráfico. Los países de ingresos medios bajos y medios altos, en amarillo y naranja respectivamente, también muestran una distribución más compacta en comparación con los de ingresos altos. Esta visualización resalta la amplia brecha entre los países de diferentes niveles de ingresos.

El grupo High income tiene una media de GNI de 41120.30, mientras que la categoría siguiente (Upper Middle income) es de 7973, lo cual representa 1.45 desviaciones estándar de diferencia. Los grupos Low income, Lower middle income y Upper middle income están más cerca entre sí en términos de GNI per capita, con distancias cercanas a cero desviaciones estándar entre sí.

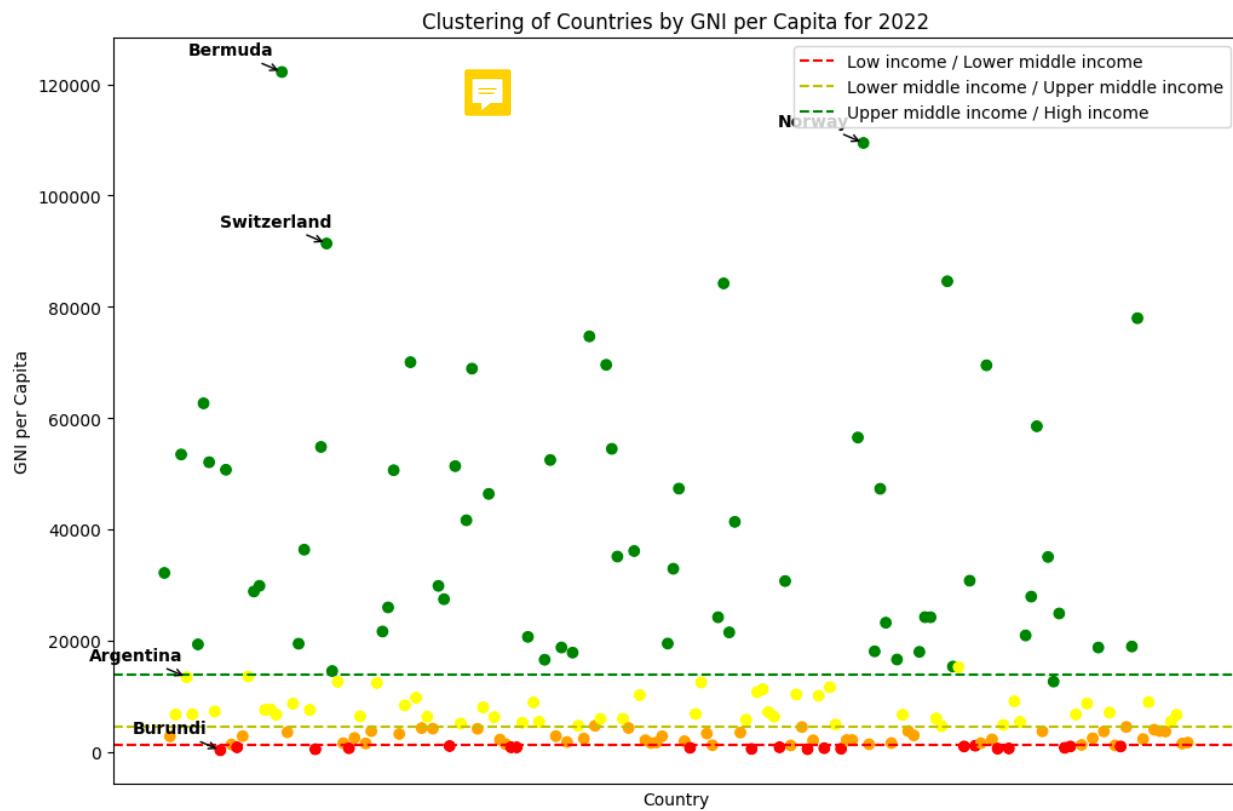


Figura 13: Scatterplot que muestra el GNI per cápita para los países según la categorización del Banco Mundial. En verde el grupo de mas altos ingresos, con sus mayores valores en Bermuda, Noruega y Suiza, mientras que del lado de mas bajos ingresos puede verse a Burundi. Argentina se sitúa en el límite entre Upper middle income y High income.

Haciendo un cálculo de la inercia y el coeficiente de silhouette para estos Clústeres, el resultado obtenido es: 796 para la inercia y 0.26 para el coeficiente de silhouette. Esto sugiere que los clústeres están dispersos y muy cercanos entre sí.

Una primera clusterización utilizando K-means con 4 grupos sobre con el GNI per cápita como único feature, logra una clasificación de mejor calidad, con un coeficiente de silhouette de  $\sim 0.7$  y una inercia de 12.72.

En esta categorización Argentina queda ubicada dentro de la categoría más baja de ingresos (Figura 14). Asimismo, aplicando el mismo algoritmo con 7 clusters, se logra una clasificación que tiene  $\sim 0.6$  de coeficiente de silhouette y  $\sim 3.5$  de inercia (Figura 15).

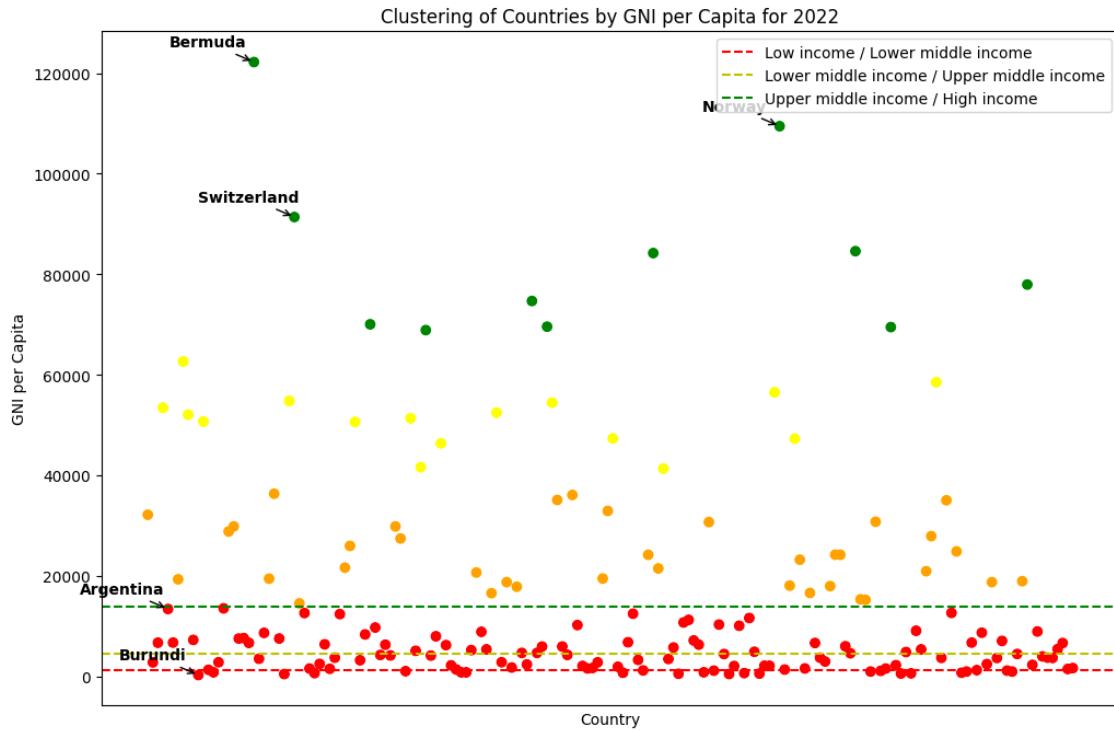


Figura 14: Scatterplot que muestra el GNI per cápita para los países categorizados utilizando K-means con 4 grupos sobre el GNI como único feature. En verde el grupo de mas altos ingresos. Argentina se sitúa en el Cluster de mas bajos ingresos.

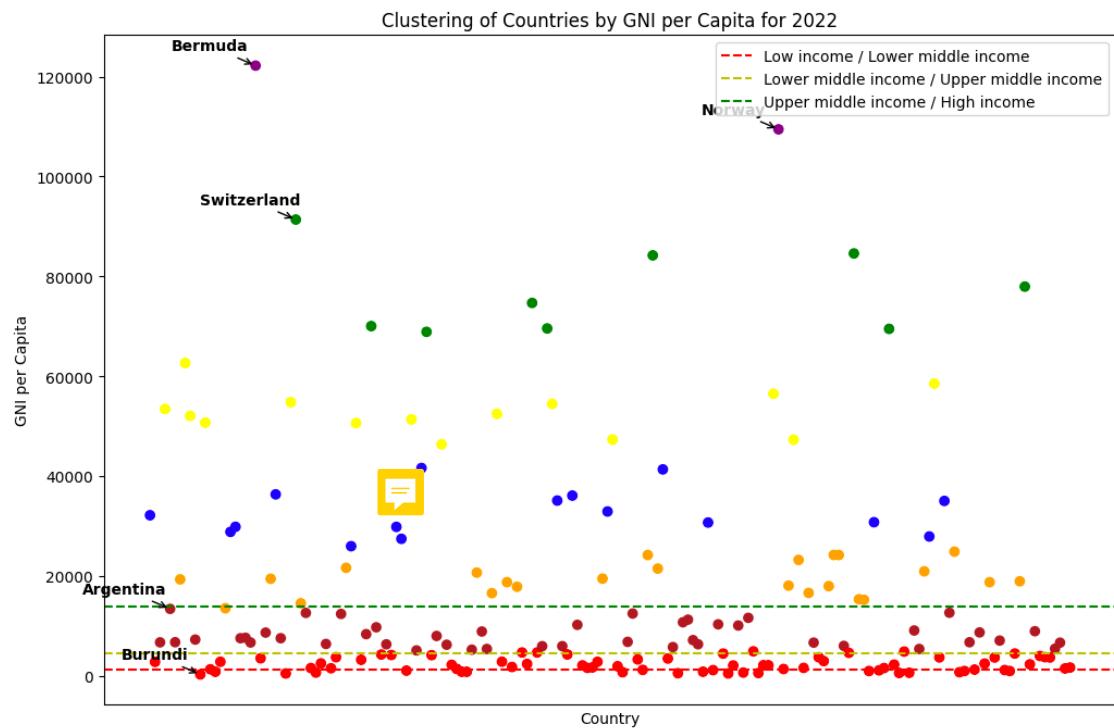


Figura 15: Scatterplot que muestra el GNI per cápita para los países categorizados utilizando K-means con 7 grupos sobre el GNI como único feature. En violeta el grupo de mas altos ingresos (que ahora cuenta únicamente con 2 elementos). Argentina se sitúa en el Cluster inmediatamente superior al de mas bajos ingresos.

La Figura 16 muestra el cambio en los criterios de clasificación utilizando el algoritmo de Clustering en lugar de los del Banco Mundial. Muchos de los catalogados como High Income pasan a las categorías de Low Income y Lower Middle Income. Aquellos coloreados en azul (indefinido) son los que no cuentan con el dato dentro del dataset.

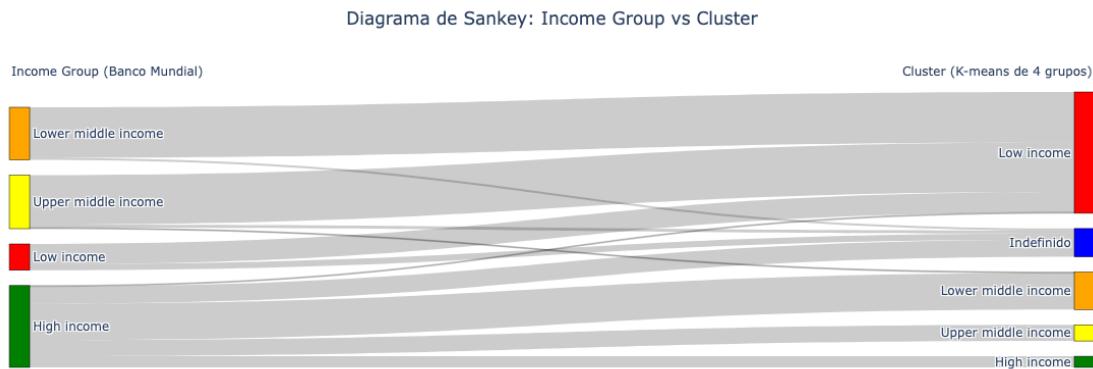


Figura 16: Scatterplot que muestra el GNI per cápita para los países categorizados utilizando K-means con 4 grupos sobre el GNI como único feature. En verde el grupo de mas altos ingresos. Argentina se situa en el Cluster de mas bajos ingresos.

## Educación

Incluye estadísticas sobre el acceso a la educación, la calidad educativa y los resultados de aprendizaje. Indicadores como la tasa de matrícula en educación primaria y el gasto público en educación son cruciales para medir el progreso en el sector educativo.

El área de Educación del Banco Mundial abarca una amplia gama de indicadores que proporcionan una visión detallada del estado y la evolución del sistema educativo en los países de todo el mundo. Estos indicadores se agrupan en varias categorías clave, tales como tasas de matrícula, deserción escolar, logro educativo y gasto en educación. Por ejemplo, las tasas ajustadas de matrícula neta en educación primaria (SE.PRM.TENR), desglosadas por género, permiten evaluar el acceso equitativo a la educación básica. Indicadores de deserción escolar en adolescentes (SE.SEC.UNER.LO.ZS) y en niños en edad primaria (SE.PRM.UNER.ZS) ofrecen datos cruciales para entender los desafíos en la retención de estudiantes. Además, los indicadores de duración de la educación obligatoria (SE.COM.DURS) y los gastos corrientes en educación primaria, secundaria y terciaria (SE.XPD.CPRM.ZS, SE.XPD.CSEC.ZS, SE.XPD.CTER.ZS) proporcionan información sobre la inversión en el sector educativo y su impacto potencial en la calidad de la educación.

Otra área importante dentro de la educación es el logro educativo, que se mide a través de indicadores como la tasa de finalización de la educación secundaria inferior (SE.SEC.CMPT.LO.ZS) y los niveles de logro educativo en la población adulta, desglosados por niveles de educación alcanzados, desde primaria hasta doctorado (SE.PRM.CUAT.ZS, SE.TER.CUAT.DO.ZS). Estos indicadores permiten evaluar no solo el acceso, sino también la permanencia y el éxito en el sistema educativo. Además, los indicadores de paridad de género en la educación (SE.PRM.GINT.FE.ZS, SE.ENR.TERT.FM.ZS) son cruciales para monitorear y promover la igualdad de oportunidades educativas entre hombres y mujeres. En conjunto, estos datos proporcionan una base sólida para analizar las políticas educativas, identificar áreas de mejora y desarrollar estrategias para aumentar la equidad y calidad en la educación a nivel global.

Como es de esperarse, utilizando los indicadores relacionados con la fuerza laboral (fuerza laboral total, desempleo, fuerza laboral femenina, etc.) proporciona una visión completamente distinta.

La Figura 17 muestra el resultado de aplicar PCA luego de haber hecho una clasificación con K-means (utilizando 4 clusters). Como puede verse, mientras los países que antes destacaban (tanto hacia arriba como hacia abajo) ahora forman parte del mismo cluster que Argentina, ahora hay nuevos "Outliers" (en este caso China, India y Estados Unidos).

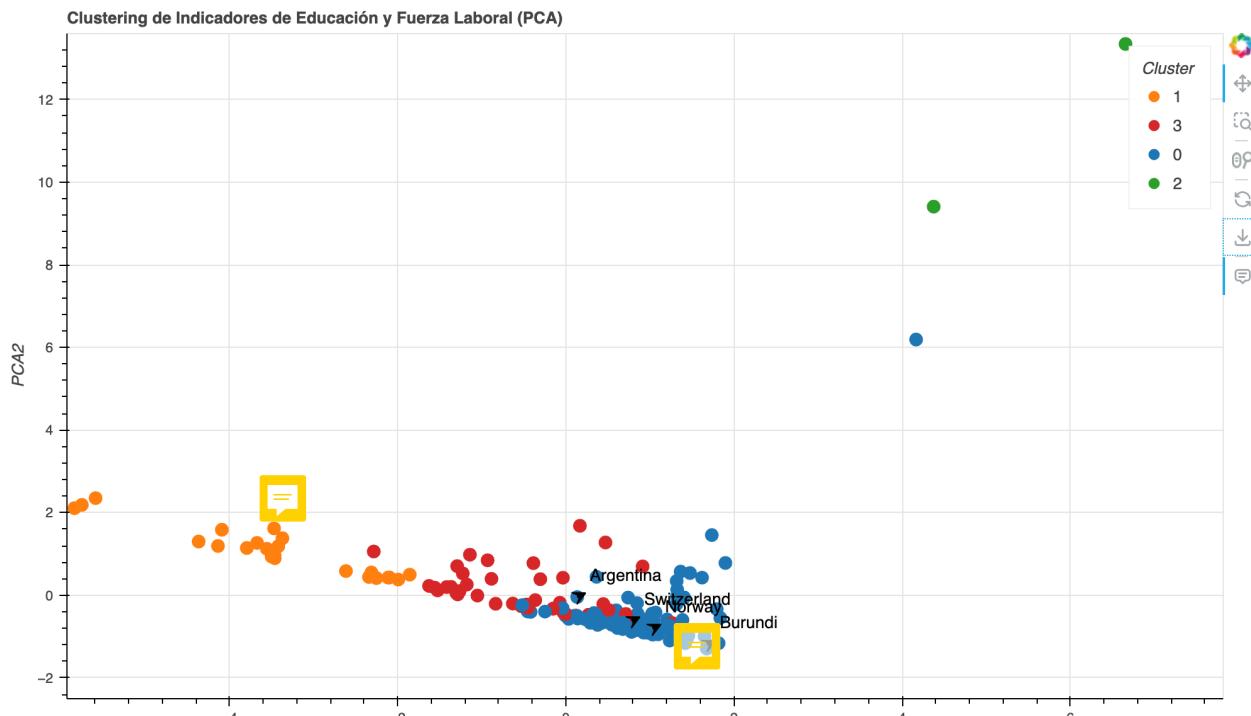


Figura 17: Scatterplot del resultado de aplicar PCA (con una varianza explicada de 77.43%) luego de hacer un Clustering de 4 clusters utilizando indicadores relacionados con la fuerza laboral y desempleo para el año 2022. Los clusters formados son completamente distintos. Se destacan China, India y Estados Unidos.

#### Clúster 0 (Azul): Economías Estables y Desarrolladas

Este clúster incluye países como Argentina, Suiza y Noruega. Las características promedio de estos países son una economía relativamente fuerte y estable, con bajas tasas de desempleo y una alta participación femenina en la fuerza laboral. Por lo tanto, el nombre "Economías Estables y Desarrolladas" refleja adecuadamente sus características.

#### Clúster 1 (Naranja): Economías en Transición con Alta Desocupación

Este clúster representa países con menores ingresos y alta tasa de desempleo. Los países en este clúster enfrentan desafíos económicos significativos, con altas tasas de desempleo, especialmente entre las mujeres. El nombre "Economías en Transición con Alta Desocupación" captura bien la situación de estos países.

#### Clúster 2 (Verde): Economías Emergentes de Gran Escala

Este clúster incluye países con grandes poblaciones y economías emergentes. Estos países tienen grandes economías con baja tasa de desempleo, pero la participación femenina en la fuerza laboral es menor comparada con otros clústeres. El nombre "Economías Emergentes de Gran Escala" es apropiado para describir este grupo.

#### Clúster 3 (Rojo): Economías Vulnerables

Este clúster incluye países como Burundi, con economías más débiles. Estos países tienen tasas de desempleo más altas y menor participación femenina en la fuerza laboral, reflejando desafíos económicos y laborales significativos. El nombre "Economías Vulnerables" es adecuado para describir las características de este clúster.

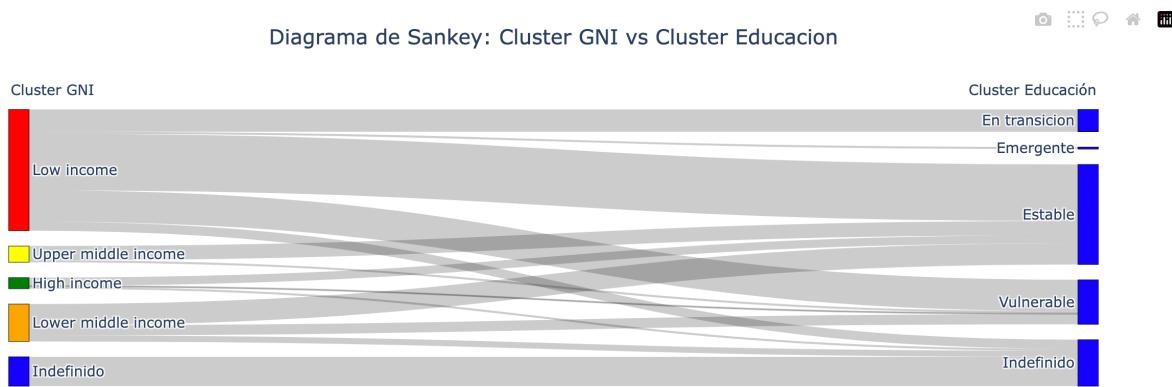


Figura 18: Diagrama de Sankey que muestra las diferencias entre los clustes conformados a partir del GNI en comparación con los clustes conformados a partir de los datos de Educación (Fuerza Laboral). Aquellos para los cuales no hay datos, se ven con la etiqueta de indefinido.

Los países considerados como Low Income se distribuyen, principalmente a las categorías de Estable y Vulnerable. Todos los países en la categoría High Income transicionan hacia la categoría Estable, con la única excepción de Qatar.

# Resultados y discusión

## Presentación y análisis de resultados obtenidos

En este trabajo se realizaron diversas técnicas de análisis y modelado para estudiar la relación entre indicadores socioeconómicos y la clasificación de países. Se aplicaron técnicas de clustering, específicamente K-means, para agrupar países según sus características económicas y laborales. Además, se utilizó el Análisis de Componentes Principales (PCA) para reducir la dimensionalidad de los datos y facilitar la visualización de los clústeres.

### Clustering y PCA

Utilizamos K-means para clasificar los países en cuatro clústeres basados en el GNI y en indicadores de la fuerza laboral y desempleo para el año 2022. Se aplicó PCA para reducir los múltiples indicadores a dos componentes principales, lo que permitió visualizar los clústeres en un gráfico bidimensional.

La primera clusterización utilizando únicamente el GNI arrojó mejores métricas que la clasificación propuesta por el Banco Mundial en cuanto a la calidad de los clusters (tanto Silhouette como la inercia mejoraron la clasificación propuesta por el organismo).

También se hicieron experimentos con distintos algoritmos y con distintas cantidades de clusters (todos disponibles en el código fuente anexo como parte de este trabajo) pero en pos de poder comparar con la clasificación mantuvimos la cantidad de clusters en 4.

Para la clusterización teniendo en cuenta los datos de fuerza laboral, los resultados del PCA mostraron que los dos primeros componentes principales explicaron el 77.43% de la varianza total, con un 43.47% explicado por el primer componente y un 33.96% por el segundo componente. Esto indica que los datos originales de alta dimensión se representaron adecuadamente en dos dimensiones, facilitando la interpretación visual de los clústeres.

Si bien los clusters logrados con los datos de fuerza laboral no son completamente distintos a los del GNI, permiten detectar otro tipo de comportamientos y permiten considerar aspectos de igualdad que el GNI enmascara.

### Análisis de Series Temporales con ARIMA

Por cuestiones de tiempo, se alcanzó a hacer un único experimento (cuyo código se encuentra disponible como adjunto en el presente trabajo) para predecir la evolución del GNI para Argentina. Se aplicó un modelo ARIMA. Se utilizó la diferenciación para hacer la serie estacionaria, y se seleccionaron los parámetros p, d y q basados en los gráficos de ACF y PACF. El modelo ajustado mostró una buena capacidad predictiva, capturando las tendencias subyacentes en los datos históricos del GNI. Aunque es necesario mayor trabajo de entrenamiento, validación y pruebas para hacer afirmaciones al respecto.

### Evolución de los factores socioeconómicos

La evolución de los factores socioeconómicos a lo largo de los años ofrece una visión general de cómo las economías y las sociedades han cambiado en respuesta a diversas políticas, eventos globales y desarrollos internos. Utilizando el clustering de 4 grupos para los años de 1980 a 2022, no se detectan grandes patrones observables ni muchos cambios significativos a lo largo del rango temporal. Esto puede deberse a que el uso de solo 4 clusters no logra captar una granularidad suficiente que facilite un análisis más detallado y revelador.

El clustering con un número reducido de grupos puede resultar en una agrupación demasiado amplia, donde países con características socioeconómicas muy diferentes terminan en el mismo clúster, diluyendo así las diferencias y similitudes significativas. Además, las transformaciones socioeconómicas suelen ser complejas y multifacéticas, requiriendo un número mayor de clústeres para capturar la diversidad de trayectorias de desarrollo de los países.

Otra consideración es la estabilidad de los clústeres a lo largo del tiempo. Los países pueden cambiar de clúster debido a cambios en sus indicadores económicos y sociales. Sin embargo, si los clústeres son demasiado generales, estos movimientos pueden no ser visibles o no proporcionar información útil. Esto sugiere que futuros estudios deberían considerar un mayor número de clústeres y quizás integrar análisis adicionales, como el uso de métodos de clustering jerárquico o análisis de trayectorias para entender mejor las dinámicas temporales.

## Datos faltantes

La falta de datos es un desafío significativo en el análisis socioeconómico longitudinal. Aunque la ausencia de datos no invalida los resultados obtenidos, ciertamente limita su aplicabilidad y robustez. En cada año, los países que tienen (o no tienen) datos varían, lo que dificulta la tarea de comparar los clústeres entre sí, ya que las muestras subyacentes cambian.

Esta inconsistencia en los datos puede introducir sesgos y afectar la precisión de los clústeres formados. Por ejemplo, un país con datos faltantes en varios indicadores puede ser asignado a un clúster incorrecto debido a la información incompleta. Además, la variabilidad en los datos disponibles entre años impide una comparación directa y clara de la evolución de los clústeres, lo que es crucial para entender las dinámicas a largo plazo.

Para mitigar estos problemas, es fundamental explorar técnicas avanzadas de imputación de datos que permitan estimar los valores faltantes de manera confiable. Métodos como la imputación múltiple, el uso de modelos predictivos basados en machine learning, o técnicas de imputación basadas en series temporales pueden ofrecer soluciones efectivas. Además, la integración de datos de múltiples fuentes y la validación cruzada de estos datos pueden ayudar a mejorar la calidad y la cobertura de los datos disponibles, proporcionando una base más sólida para el análisis.

## Discusión de los resultados y su relevancia

Los resultados obtenidos proporcionan una visión clara de cómo los países pueden ser agrupados según sus características socioeconómicas y laborales. La técnica de PCA demostró ser efectiva para reducir la complejidad de los datos y facilitar su visualización. Los clústeres identificados revelan patrones interesantes sobre la economía y el mercado laboral de diferentes países, destacando las similitudes y diferencias significativas entre ellos.

El uso del modelo ARIMA para predecir el GNI es particularmente relevante para la planificación económica y la formulación de políticas. La capacidad del modelo para capturar las tendencias históricas y hacer predicciones razonables puede ayudar a los responsables de políticas a tomar decisiones informadas sobre intervenciones económicas y sociales, pero realmente se necesita mas trabajo de modelado y validacion para poder conseguir resultados utilizables.

Aunque el análisis actual proporciona una visión útil de los factores socioeconómicos, no arroja luz sobre los cambios a lo largo del tiempo, las limitaciones impuestas por los datos faltantes y la granularidad del clustering deben ser abordadas en investigaciones futuras para obtener conclusiones más detalladas y robustas

## Limitaciones y posibles mejoras

Una limitación de este estudio es la dependencia de los datos disponibles, que en algunos casos presentan datos faltantes. Aunque se aplicaron estrategias para mitigar este problema, la precisión y la fiabilidad de los resultados pueden verse afectadas. Futuras investigaciones podrían beneficiarse de la integración de más fuentes de datos y la aplicación de técnicas avanzadas de imputación de datos faltantes.

Además, explorar otras técnicas de clustering, como el clustering jerárquico o el clustering basado en densidad, podría proporcionar perspectivas adicionales y complementar los hallazgos obtenidos con K-means. También se podría extender el análisis a un período más largo, utilizando datos de múltiples años para observar la evolución de los países en el tiempo y revelar tendencias a largo plazo.

# Conclusión

## Resumen de los hallazgos principales

### Clustering Basado en GNI y Fuerza Laboral

- La técnica de K-means, aplicada tanto al GNI per cápita como a los indicadores de fuerza laboral, permitió identificar diferentes clústeres de países.
- La clasificación basada en GNI mostró una mejora en la calidad de los clústeres respecto a la clasificación del Banco Mundial.
- Los resultados del clustering basado en fuerza laboral revelaron patrones distintos, destacando la importancia de la participación femenina en la fuerza laboral y las tasas de desempleo.

### Análisis de Componentes Principales (PCA)

- PCA facilitó la visualización de los datos reduciendo la dimensionalidad y manteniendo el 77.43% de la varianza total en dos componentes principales.
- Esta visualización permitió identificar outliers y patrones entre los clústeres, proporcionando una herramienta visual efectiva para el análisis de los datos.

## Conclusiones generales y su relación con los objetivos del trabajo

En resumen, el uso de técnicas de clustering, y en particular de K-means, ha demostrado ser una herramienta poderosa para la clasificación y análisis de los indicadores socioeconómicos, permitiendo una mejor comprensión de las dinámicas y características de los países en estudio. Sin embargo, para mejorar la robustez y aplicabilidad de los resultados, es crucial considerar las estrategias mencionadas para manejar los datos faltantes y explorar enfoques más avanzados y diversos en futuras investigaciones.

El trabajo ha logrado clasificar y analizar los indicadores socioeconómicos de manera efectiva, utilizando técnicas de clustering y PCA para proporcionar una mejor comprensión de las dinámicas y características de los países en estudio. La utilización de K-means y PCA ha demostrado ser particularmente útil para identificar patrones y relaciones entre los indicadores seleccionados, ofreciendo una nueva perspectiva sobre la clasificación de países basada en sus características económicas y laborales.

## Recomendaciones para futuros trabajos

Considerando los resultados y las limitaciones del análisis actual, se proponen varias direcciones para investigaciones futuras que podrían mejorar la comprensión de las dinámicas socioeconómicas y ofrecer una base más sólida para la formulación de políticas. Mayor exploración de los datos existentes.

### Mayor Exploración de los Datos Existentes

En este trabajo, se utilizó un subconjunto muy pequeño de los 1463 indicadores provistos por la fuente de datos original. Aunque es de esperar que muchos de estos indicadores sean irrelevantes para nuestro análisis específico, existe un potencial significativo en explorar más a fondo la vasta cantidad de datos disponibles. A continuación, se presentan algunas recomendaciones para futuras investigaciones sobre cómo aprovechar mejor este amplio conjunto de datos.

## Identificación de Indicadores Relevantes

El primer paso hacia una mayor exploración de los datos es identificar qué indicadores adicionales pueden ser relevantes para el análisis. Esto puede implicar un proceso iterativo de selección y evaluación de indicadores basados en criterios específicos como su relación con el desarrollo económico, social y laboral. Utilizar técnicas de selección de características (feature selection) y análisis de correlación puede ayudar a identificar los indicadores más prometedores para estudios más detallados.

## Análisis Multidimensional

El uso de un mayor número de indicadores permitirá realizar análisis multidimensionales más complejos y detallados. Esto puede incluir la aplicación de técnicas de análisis factorial, análisis de componentes principales (PCA) con más dimensiones, y modelos de redes neuronales que pueden manejar grandes volúmenes de datos para detectar patrones y relaciones no lineales. Estos enfoques pueden proporcionar una visión más rica y matizada de las dinámicas socioeconómicas.

## Estudios Sectoriales

Explorar los datos existentes puede permitir estudios sectoriales específicos que aborden áreas como la educación, la salud, la infraestructura, y la tecnología, entre otros. Cada uno de estos sectores tiene indicadores específicos que, cuando se analizan en profundidad, pueden proporcionar información valiosa sobre las fortalezas y debilidades de diferentes países en esos ámbitos. Por ejemplo, indicadores relacionados con la salud pública pueden ofrecer perspectivas sobre cómo las políticas de salud afectan el desarrollo económico y social.

## Uso de Datos Temporales Completo

Como posible trabajo futuro, se propone utilizar todos los años disponibles para todos los países y realizar clustering en función de la evolución de los países a lo largo del tiempo. Esta aproximación temporal podría revelar tendencias y patrones de desarrollo que no son evidentes cuando se analiza un solo año. Al considerar la evolución de los indicadores socioeconómicos a lo largo del tiempo, se pueden identificar patrones de crecimiento y desarrollo que se pierden en un análisis estático.

## Exploración de Nuevas Técnicas de Clustering

Además de K-means, explorar otras técnicas de clustering, como el clustering jerárquico o el clustering basado en densidad, podría proporcionar perspectivas adicionales y complementar los hallazgos obtenidos con K-means. Estas técnicas alternativas pueden ofrecer una mayor flexibilidad y adaptabilidad a la estructura subyacente de los datos, permitiendo la identificación de subgrupos y patrones más sutiles en los datos.

## Problema de los Datos Faltantes

La falta de datos es un desafío significativo que debe abordarse para mejorar la robustez y aplicabilidad de los resultados. A continuación, se describen varias estrategias que podrían ser implementadas para manejar los datos faltantes:

**Búsqueda de Información Externa:** Si bien algunos de estos indicadores podrían estar disponibles en datasets de otras organizaciones, este enfoque no puede considerarse como el método por defecto. El tiempo de búsqueda y validación es una limitante, y no hay garantía de éxito. Esta técnica puede utilizarse en casos muy específicos, como obtener datos conocidos en Argentina.

**Imputar Datos Faltantes:** Para los indicadores con hasta el 5% de datos faltantes, se intenta imputarlos utilizando diversas técnicas, luego evaluar las conclusiones a las que esto conduce. Es necesario ser muy cuidadosos con la forma en que se imputan, y las conclusiones derivadas de esto deben ser tomadas con cautela. Técnicas como la imputación por la media, media móvil, último valor conocido, regresión y MICE (Multivariate Imputation by Chained Equations) pueden ser útiles.

Segmentar los Datos: Dado que hay periodos de tiempo que parecen tener mayor volumen de datos completos, lo mejor es explorar estrategias de corte que eviten la necesidad de imputar los datos faltantes de manera masiva, como por ejemplo: cortes por Región, por indicadores o esquemas mixto

## Bibliografía y Referencias

- [1] Fernando Antonio Ignacio González, Silvia London, María Emma Santos (2012). [The Journal of International Trade & Economic Development](#). The Journal of International Trade & Economic Development.
- [2] Alan M. Taylor (1994).[Three Phases of Argentine Economic Growth](#). National Bureau of economic research.
- [3] Robert J. Barro (1996). [Determinants of Economic Growth: A Cross-Country Empirical Study](#). National Bureau of economic research.
- [4] Daniel Landau (1986). [Government and Economic Growth in the Less Developed Countries: An Empirical Study for 1960-1980](#). Universidad de Connecticut
- [5] Michael Timberlake, Jeffrey Kentor (1986). [Economic Dependence, Overurbanization, and Economic Growth: A Study of Less Developed Countries](#). The sociological quarterly, official Journal of the Midwest Sociological Society.
- [6] Richard Weisskoff (1970). [INCOME DISTRIBUTION AND ECONOMIC GROWTH IN PUERTO RICO, ARGENTINA, AND MEXICO](#). The review of income and wealth.
- [7] World Bank Open Data, [Free and open access to global development data](#).

## Anexos

### Anexo 1: Código fuente utilizado en el análisis

Todo el código fuente desarrollado como parte de este trabajo se encuentra en [este](#) enlace.

### Anexo 2: Descripción de áreas clave

Área	Descripción	Ejemplos de indicadores
Agricultura y Desarrollo Rural	Incluye indicadores sobre producción agrícola, uso de la tierra, insumos agrícolas y desarrollo rural.	Producción agrícola, Uso de fertilizantes, Tierra arable
Eficacia de la Ayuda	Mide la efectividad y el impacto de la ayuda internacional.	Flujos netos de ayuda oficial al desarrollo, Asistencia bilateral
Cambio Climático	Se centra en los efectos del cambio climático y las medidas de mitigación.	Emisiones de CO2, Consumo de energía renovable
Economía y Crecimiento	Comprende datos sobre el crecimiento económico, la estructura económica y la productividad.	Producto Interno Bruto (PIB), Crecimiento del PIB per cápita
Educación	Incluye estadísticas sobre el acceso a la educación, la calidad educativa y los resultados de aprendizaje.	Tasa de matrícula en educación primaria, Gasto público en educación
Energía y Minería	Se enfoca en el suministro y el consumo de energía, así como en la explotación de recursos minerales.	Producción de energía, Consumo de electricidad per cápita
Medio Ambiente	Cubre temas como la biodiversidad, la calidad del aire y el agua, y la gestión de residuos.	Áreas protegidas, Índice de calidad del aire
Deuda Externa	Proporciona datos sobre la deuda externa de los países y su sostenibilidad.	Deuda externa total, Pagos de servicio de la deuda
Sector Financiero	Incluye estadísticas sobre la banca, los mercados financieros y el acceso al financiamiento.	Crédito doméstico al sector privado, Capitalización bursátil
Género	Examina las desigualdades de género en diversas áreas como la educación, la salud y el empleo.	Tasa de participación laboral femenina, Diferencia salarial entre géneros
Salud	Cubre aspectos de la salud pública, el acceso a servicios de salud y los resultados de salud.	Esperanza de vida al nacer, Mortalidad infantil
Infraestructura	Se centra en el desarrollo y la calidad de la infraestructura básica y avanzada.	Acceso a electricidad, Infraestructura de transporte
Pobreza	Mide la incidencia y la severidad de la pobreza.	Tasa de pobreza internacional (menos de \$1.90 al día), Índice de Gi
Sector Privado	Incluye datos sobre el desarrollo del sector privado y la actividad empresarial.	Número de empresas nuevas registradas, Facilidad para hacer negocios
Sector Público	Examina la eficiencia y la calidad del gobierno y el sector público.	Gasto público como % del PIB, Índice de gobernabilidad
Ciencia y Tecnología	Proporciona datos sobre innovación, investigación y desarrollo tecnológico.	Gasto en I+D como % del PIB, Número de patentes registradas

Desarrollo Social	Incluye estadísticas sobre bienestar social, cohesión social y servicios sociales.	Acceso a servicios de saneamiento, Índice de desarrollo humano
Protección Social y Trabajo	Se enfoca en la seguridad social, las condiciones laborales y la protección del empleo.	Cobertura de seguridad social, Tasa de desempleo
Comercio	Examina el comercio internacional y la integración económica.	Exportaciones de bienes y servicios, Balanza comercial

Tabla 1: Áreas clave en las que el World Bank agrupa los distintos indicadores socioeconómicos de las naciones y su descripción junto a algunos indicadores relevantes por área.