# 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.

Con una perspectiva práctica, buscamos proporcionar datos relevantes que puedan servir como referencia en la elaboración de estudios futuros. Nos centraremos 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?

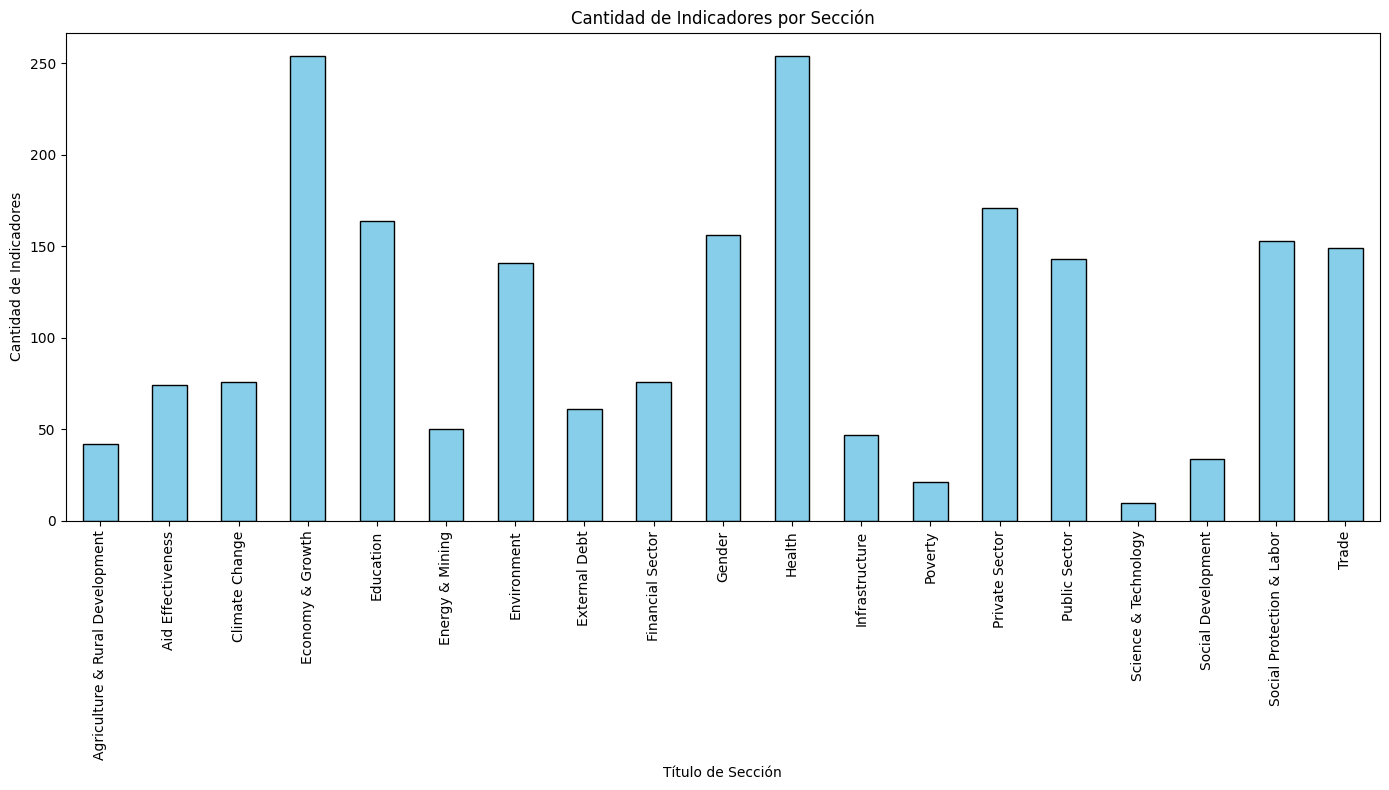
# Metodos y Materiales

## Datos

Trabajaremos con un dataset del Banco Mundial que incluye alrededor de 1462 indicadores de países en áreas clave como Educación, Ciencia y Tecnología, y Crecimiento económico, abarcando el período de 1960 a 2021. Los datasets se encuentran disponibles en este [link](https://data.worldbank.org/indicator) de descarga.

Dentro de la pagina del Banco Mundial, los datos se almacenan y acceden de manera individual. Esto es, cada indicador se encuentra en un archivo excel con la información de ese indicador para todos los paises y todos los años.

Tal como se menciona anteriormente, los indicadores se encuentran agrupados por distintas areas, cada indicador puede pertenecer a mas de un área.



Los paises se categorizan por dos criterios: Geografico y economico. Las regiones en las que distingue el banco mundial a los paises son las siguientes:

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

Mientras que las distintas categorias de ingresos en las que clasifica a los paises son las siguientes:

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

A graph with different colored bars

Description automatically generated

### Preprocesamiento

Los datos se encuentran agrupados en distintas planillas por cada indicador. El trabajo de preprocesamiento consiste en los siguientes pasos:

1. Descarga automática: Utilizamos Selenium para automatizar la descarga de datos desde la página del Banco Mundial. Selenium es una herramienta que permite controlar un navegador web a través de scripts, facilitando la interacción con elementos de las páginas web y la descarga de archivos necesarios para nuestro análisis.
   1. Navegación y selección de enlaces: Con Selenium, navegamos a la página del Banco Mundial que contiene los indicadores. Identificamos y seleccionamos todos los enlaces relevantes que contienen los indicadores de nuestro interés. Para esto, utilizamos selectores CSS y XPath para ubicar los elementos específicos en la página.
   2. Extracción de datos: Una vez identificados los enlaces, extraemos los atributos relevantes, como el texto del enlace y la URL. Estos datos se almacenan en un archivo CSV para su posterior procesamiento.
   3. Descarga de archivos: Selenium también se encarga de realizar la descarga de los archivos Excel correspondientes a cada indicador. Esto se realiza haciendo clic automáticamente en los enlaces de descarga presentes en la página web.
2. Generacion de metadata: Ademas de bajar la informacion, es importante generar un archivo maestro con la metadata necesaria para el procesamiento y entendimiento de la informacion (grupos de indicadores, significado de cada indicador, etc.)
3. Agrupamiento: Luego de descargar los archivos, y para facilitar su uso se genero un proceso que agrupa los datos en las categorias anteriormente descriptas. Esto permite, en lugar de tener 1462 archivos de datos separados, tener 19 archivos.

En la siguiente figura pueden verse las cantidades de indicadores por cada grupo.

## Analisis y Modelado

# Nuestro análisis constará de: (1) Métodos descriptivos que incluyen técnicas estadisticas y visualización de datos para caracterizar y comparar países; (2) Técnicas de clustering como K-means para identificar agrupaciones de países con patrones de desarrollo similares; y (3) Análisis de series temporales utilizando modelos como ARIMA para evaluar la evolución de los indicadores a lo largo del tiempo.

### Métodos descriptivos

El foco estara dado en entender si los datos existentes permiten la utilización de las tecnicas de clustering y de analisis de series temporales que pensamos utilizar.

En particular, existen muchos datos faltantes, que pueden dificultar la tarea. En estos casos intentaremos no imputar valores sino utilizar los indicadores que se encuentren mas completos.

Los datos son confiables pero hay muchos faltantes. Voy a tener que acotar los indicadores a aquellos que esten mas completos y tal vez a periodos de tiempo mas acotados y a regiones o categorias de paises que se encuentren mas completos para poder hacer clustering.

### Técnicas de clustering

### Análisis de series temporales

## Herramientas

Para el análisis de datos y modelado utilizaremos las siguientes herramientas:

* Python: Utilizaremos diversas bibliotecas para distintas tareas:
  + Pandas y NumPy para la manipulación y análisis de datos.
  + Matplotlib y Seaborn para la visualización de datos.
  + Scikit-learn para las técnicas de clustering, como K-means.
  + Statsmodels para el análisis de series temporales, incluyendo modelos ARIMA.
  + Selenium: Para la automatización de la descarga de datos desde la web.
  + OpenPyXL: Para manipulación de archivos Excel durante el preprocesamiento de datos.
  + PDFPages de Matplotlib: Para generar reportes en PDF con gráficos de los análisis realizados.
  + Jupyter Notebooks: Para documentar y presentar los análisis de manera interactiva.
  + Microsoft Excel: Para la gestión inicial de los datos descargados y la generación de la metadata necesaria.

# Resultado esperado

Esperamos que este enfoque analítico revele patrones subyacentes en la trayectoria de desarrollo de los países estudiados, ofreciendo así una base de conocimiento que pueda servir para futuras investigaciones y para la formulación de políticas más efectivas hacia el desarrollo sostenible.