Tipología y ciclo de vida de los datos
PRA 1: Web Scraping

Jonathan Zambrano
Tatiana Piccolomini

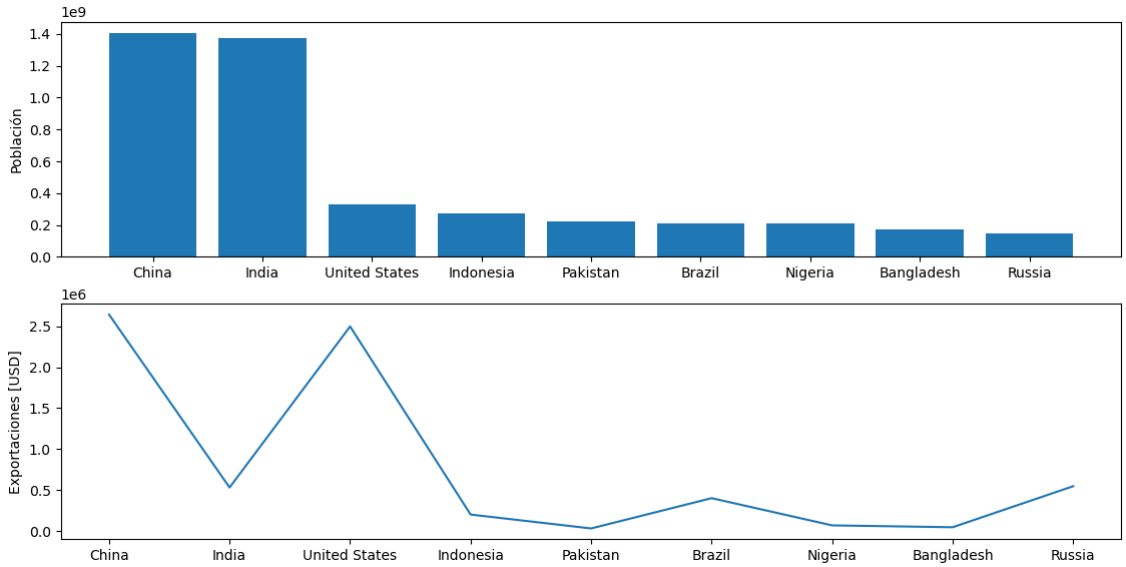
1. Contexto. Explicar en qué contexto se ha recolectado la información. Explique por qué el sitio web elegido proporciona dicha información.
Los datos presentados en el dataset han sido recabados en base a la información contenida en la web de Wikipedia, la cual, contiene un sinnúmero de páginas relativas a información y factores económicos diferenciados tanto por indicador como por país, para lo cual, el dataset reúne en una sola fuente todos estos indicadores.

La página de Wikipedia corresponde a una enciclopedia en línea que es editada a nivel mundial por más de 1 millón de usuarios, quienes sugieren cambios y añaden páginas con información de diferentes temáticas; finalmente, la información es verificada y aplicada en su versión final incluyendo información como la fuente de origen de los datos.

2. Definir un título para el dataset. Elegir un título que sea descriptivo.
El nombre que se ha definido para el dataset es: wiki_country_data. Dando referencia a que se relaciona con información económica de países obtenido de la web Wikipedia.

3. Descripción del dataset. Desarrollar una descripción breve del conjunto de datos que se ha extraído (es necesario que esta descripción tenga sentido con el título elegido).
El dataset obtenido contiene por un lado un listado de países cuya información ha sido obtenida de la web de Wikipedia (249 países), a continuación, se presenta un listado de indicadores principales en común para cada país como población, exportaciones, importaciones, entre otros; estos indicadores se han definido como claves para la interpretación del dataset. Finalmente, se complementa el dataset con un listado de información complementaría propia de cada país relacionadas a estadísticas, comercio y Finanzas Públicas. Toda esta información se encuentra basada en la forma de presentación de los datos en la página de Wikipedia.

4. Representación gráfica. Presentar esquema o diagrama que identifique el dataset visualmente y el proyecto elegido
Se presentan las gráficas de los dos primeros indicadores principales del dataset, como muestra de la información que puede ser obtenida gráficamente del dataset.



5. Contenido. Explicar los campos que incluye el dataset, el periodo de tiempo de los datos y como se ha recogido.
El dataset está compuesto de los siguientes segmentos definidos a continuación:

- **País:** Corresponde al listado de países de los cuales se ha obtenido su información económica (249).
- **Indicadores principales:**
  - Población
  - Exportaciones
  - Importaciones
  - Principales Socios Comerciales
  - Exportación de Petróleo
  - Producción de Petróleo
- **Información Económica General:**
  - Moneda
  - Año Fiscal
  - Organizaciones de Comercio
  - Grupo Económico
- **Estadísticas:**
  - GDP (Producto Interno Bruto)
  - Ranking GDP
  - Crecimiento del GDP
  - GDP per Cápita
  - Ranking GDP per Cápita
  - GDP por Sector
  - GDP por componente
  - Inflación (CPI)
  - Población debajo del límite de la pobreza
  - Coeficiente de Gini
  - Human Development Index
  - Fuerza Laboral
  - Desempleo
  - Principales Industrias
- **Comercio:**
  - Productos Exportados
  - Productos Importados
- **Finanzas Públicas:**
  - Deuda Interna
  - Gasto Público
  - Reservas Internacionales

De igual manera, se debe indicar que los datos se encuentran actualizados a la fecha, aunque es posible incluir actualizaciones del dataset al ejecutar el programa para la extracción de los indicadores en fechas futuras. Adicionalmente, se indica que los indicadores se presentan en idioma inglés.

6. Agradecimientos. Presentar al propietario del conjunto de datos. Es necesario incluir citas de análisis anteriores o, en caso de no haberlas, justificar esta búsqueda con análisis similares.
Como se ha mencionado previamente, la web de donde se han obtenido los datos corresponde a Wikipedia, la cual, pertenece al grupo sin fines de lucro de Wikimedia, y el cual, es un una web de enciclopedia abierta, donde más de un millón de usuarios realizan aportes a la web, que posteriormente son aprobados y publicados.

Algunas fuentes que contienen información de indicadores económicos mundiales son:

- https://datahub.io/collections/economic-data#basic-economic-data
- https://www.google.com/publicdata/explore?ds=d5bncppjof8f9_
- https://ourworldindata.org/countries

7. Inspiración. Explique por qué es interesante este conjunto de datos y que preguntas se pretenden responder. Es necesario comparar con los análisis anteriores presentados en el apartado 6.

El dataset generado presenta indicadores globales, relevantes y diversos que sirven de base para análisis económicos basados en cada país. Estos análisis más específicos pueden considerar indicadores o países determinados; así como, al existir variables categóricas se pueden implementar algoritmos de clustering para encontrar grupos de países con economías de iguales características.

Las preguntas que se pretender responder están orientadas al crecimiento / decremento económico de los países, relaciones comerciales y principales productos importados / exportados. Así mismo, el aprovechamiento del dataset puede permitir el desarrollo de proyectos de análisis de riesgo país, descubrimiento de mercados sobrecargados o no explotados.

8. Licencia. Seleccione una de estas licencias para su dataset y explique el motivo de su selección:

El dataset generado tiene como objetivo proveer una herramienta para diversos análisis económicos, para lo cual, se espera que los posibles usuarios del dataset compartan y realicen cambios, modificaciones y mejoras, así mismo, se permitirá su uso comercial, de tal manera, que se ha seleccionado la licencia “CC BY-SA 4.0”.

9. Código. Adjuntar el código con el que se ha generado el dataset, preferiblemente en Python o, alternativamente, en R.

El código en Python se encuentra adjunto al github del proyecto (https://github.com/tatianapicc/web-scrapping-wiki-country-data) .

10. Dataset. Publicación del dataset en formato CSV en Zenodo (obtencion del DOI) con una breve descripción.

En Proceso. Una vez se cuente con la versión final del dataset, se realizará la publicación en Zenodo.