Practica-1---Web-scraping

Descripción

Ejercicio realizado como Práctica 1 de la asignatura M2.851-Tipología y ciclo de vida de los datos dentro del Master de Ciencia de Datos de la Universitat Oberta de Catalunya.

Miembros del Equipo

La actividad ha sido realizada de manera individual por Hernando Hernández Mariño.

1. Contexto

Este conjunto de datos presenta proyecciones demográficas al 1 de julio de 2021, las cuales se basan en los últimos censos, estimaciones oficiales de los países y territorios dependientes, así como de proyecciones de la ONU o el reloj de población nacional; no obstante, en estos cálculos, no se incluyen reajustes demográficos relacionados al impacto de la actual pandemia global causada por el SARS-CoV-2. La importancia de este tipo de conjunto de datos es reconocer su relevancia en investigaciones, estudios y análisis de planeación del desarrollo socioeconómico de los países en general. La información recolectada para este ejercicio se realizó con la plataforma "Wikipedia" ya que es una fuente de información abierta, de fácil acceso, colaborativa; además es posible verificar la información suministrada, pues disponen de las fuentes de donde se obtuvo toda la información plasmada en la página web.

2. Título del DataSet

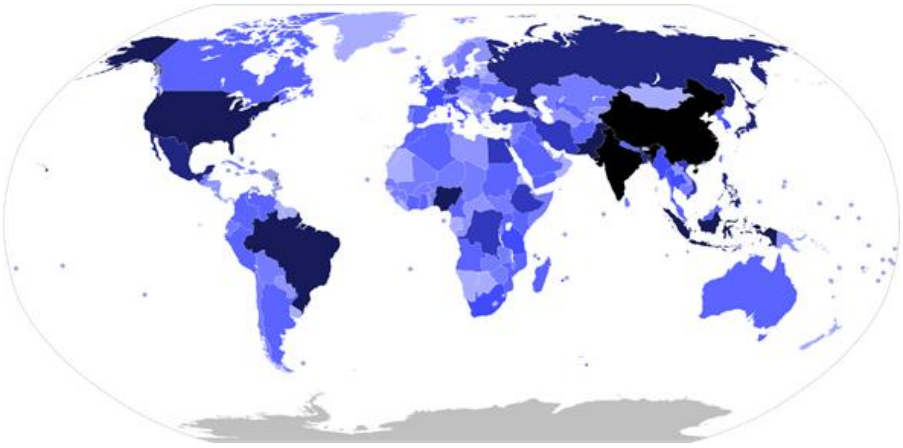
Proyecciones demográficas de países al 1 de julio de 2021.

3. Descripción del DataSet

Este dataset presenta la proyección exponencial de la cantidad de población por país al 1 de julio de 2021. También ofrece porcentajes de la población por países respecto al total mundial, las variaciones temporales (cambios medios y absolutos anuales) por países y cálculos del tiempo de duplicación de la población por país. Finalmente brinda cifras recolectadas a partir de fuentes tales como el censo más reciente, la última estimación oficial, la proyección de la ONU o el reloj de población nacional.

4. Representación Gráfica

PROYECCIÓNES DEMOGRÁFICAS POR PAÍSES-2021



Fuente: https://es.wikipedia.org/wiki/Anexo:Pa%C3%ADses_y_territorios_dependientes_por_poblaci%C3%B3n

5. Contenido

El código utilizado para realizar web-scraping en una de las páginas de Wikipedia. Inicia con algunos comandos que son ideales para hacer previo a la extracción de la información, tales como conocer la estructura de la página, el dueño y los posibles directorios que esta página puede excluir.

La página contiene información sobre la distribución poblacional del mundo por países y, esta se encuentra consolidada en una de las dos tablas que contiene la página. Por esto, se procede, en primer lugar, con la extracción del link. Luego se identifican los elementos definidos como "table" en aras de encontrar la información de interés. Una vez encontrada la tabla de interés se utiliza comandos para extraer la información de ella, la cual contiene los siguientes campos o variables:

• Países: Nombre del país a tratar.

• Proyeccion_2021: Es la cantidad poblacional proyectada al año en curso (2021) de cada país.

• Porcentaje: Es el peso porcentual que tiene la cantidad poblacional de cada país, respecto al total de la población mundial.

• CMA: Es el cambio medio anual en la cantidad poblacional de cada país.

• CAAP: Es el cambio absoluto anual promedio de la cantidad poblacional de cada país.

• PCMATA: Es el cambio medio absoluto total anual en cifras porcentuales.

• AED: La cantidad de años que se tardaría el país en lograr duplicar su población actual.

• CMR: Es el censo más reciente, fuente utilizada para las proyecciones poblacionales y demás cálculos estadísticos.

Estas variables son las que contienen la información de cada país y, es la que conforman el DataSet resultante del código propuesto.

6. Agradecimientos

El propietario o autor de los datos son los colaboradores de Wikipedia y la edición es de Wikipedia, La enciclopedia libre. Dentro de las proyecciones poblacionales, por lo general, es pertinente aclarar algunas datos o cifras referentes a las poblaciones de algunos países o territorios dependientes. Es el caso de la República Popular China. Por ejemplo, Hong Kong y Macao son regiones administrativas especiales chinas, las cuales mantienen sus propios institutos de estadística separados de China e incluso continúan realizando sus respectivos censos en fechas diferentes respecto de ella, son consideradas como dependencias aparte. Por su lado, la isla de Taiwán (la autodenominada República de China) — la cual ha sido independiente de facto desde 1949— es considerada como una entidad política aparte y por lo tanto su población no es incluida dentro de la de China continental, a diferencia de lo que sucede en el caso de las estimaciones anuales elaboradas por la División de Población de las Naciones Unidas.

Otro caso es la Federación de Rusia. Debido a la anexión de la república de Crimea y de la ciudad de Sebastopol a Rusia, proceso no reconocido por las autoridades ucranianas y que finalizó formalmente el 1 de enero de 2015 como consecuencia directa del referéndum al respecto realizado el 16 de marzo de 2014, en este anexo no aparecen incorporadas dentro de Ucrania las poblaciones de la actual Crimea y de Sebastopol.

También la población de Francia solo se refiere a su región metropolitana. Por otra parte, el resultado preliminar del censo de Marruecos realizado entre el 1 y el 20 de septiembre de 2014, así como también la proyección exponencial basada en el mismo para el 1 de julio de 2019 excluye la población del disputado territorio del Sahara Occidental (la autodenominada República Árabe Saharaui Democrática). Así mismo, la proyección para Serbia no incluye la población del estado parcialmente reconocido de la República de Kosovo, cuyo parlamento declaró unilateralmente su independencia el 17 de febrero de 2008.

Otro tanto ocurre con: la población del Estado de Palestina, la cual comprende la de Cisjordania (o la de la Ribera Occidental) y la Franja de Gaza, las cuales están separadas por un corredor de territorio israelí; la proyección de población georgiana para mediados de 2019 no incluye los estados con reconocimiento limitado de Abjasia ni de Osetia del Sur; la proyección de Moldavia para inicios de 2019 excluye a la región separatista de Transnistria; y tanto la proyección como la estimación oficial de Chipre se refieren al área de origen griego de dicha isla mediterránea, por lo que no incluyen los habitantes de la autodenominada República Turca del Norte de Chipre, la cual solamente es reconocida por Turquía.

Finalmente destacar que varios países —pertenecientes a los continentes africano o asiático— no han tenido un censo nacional de población en varios años, lo que en cierta medida tiende a afectar la precisión de sus estimaciones demográficas. Entre aquellos se encuentran (incluyendo entre paréntesis el año de sus últimos censos respectivos): Líbano (1970), Afganistán (1979), Congo Oriental (1984, entonces Zaire y desde 1997 conocido como República Democrática del Congo), Somalia (1987), Uzbekistán (1989), Eritrea (1994) e Irak (1997).

7. Inspiración

Las proyecciones de población son importantes dado que permiten estudiar los efectos de las variaciones de los principales componentes demográficos, así como la manera en que se reflejan en el volumen y la estructura por sexo y edad dentro de los países. Ahora bien, particularmente, las proyecciones demográficas suponen una simulación estadística que ayudan a conocer cuál será la evolución futura de la población en un país bajo determinadas hipótesis de fecundidad, mortalidad y migración. Desde la década de 1950 se ha reunido información relevante para elaborar estimaciones de la mortalidad, la fecundidad y la migración, puesto que son los componentes demográficos fundamentales para conocer la dinámica de la población en cuanto a su estructura por sexo y edad. Desde esa misma fecha se ha impulsado también el desarrollo de estadísticas nacionales, principalmente en lo que se refiere a los censos de población y estadísticas vitales.

Al respecto, la comunidad internacional ha elaborado recomendaciones para los levantamientos censales y los sistemas de estadísticas vitales que apuntaron a mejorar las formas de conocer los niveles, las tendencias y la estructura de la fecundidad, la mortalidad y la migración, y consecuentemente la dinámica poblacional y su composición por sexo y edad. Estas recomendaciones -que están en continua revisión -abarcan desde aspectos conceptuales y metodológicos- para la recolección de la información, incluyendo el diseño de los cuestionarios, hasta los tabulados básicos necesarios para la aplicación de las técnicas y metodologías de estimación de los componentes demográficos.

No obstante, se presentan limitaciones o precisiones como las anotadas en el apartado anterior (6). Particularmente, en este conjunto de datos, es importante advertir que mientras el total mundial de la columna de las proyecciones corresponde a la sumatoria de las poblaciones de los distintos países y territorios, no sucede lo mismo respecto del total de la columna con los estimados oficiales, censos nacionales, etc. Éste último corresponde a un cálculo diario aproximado basado en el reloj de población global mantenido por parte del sitio

web WorldOdometers.info, el cual a su vez se basa en tal sentido en la publicación World Population Prospects ("Perspectivas de la población mundial") elaborada por la División de Población de la Organización de las Naciones Unidas (ONU).

Pese a todo lo anterior, la importancia de estas proyecciones demográficas radica en su utilización posterior en otros ejercicios, como por ejemplo en la proyección del gasto en pensiones, en la estimación del crecimiento del PIB o los niveles de empleo, la planeación de servicios sociales, civiles o educativos, entre otros temas. Y es que con el aumento de la capacidad de los computadores, su uso prácticamente universal y la mayor disponibilidad de información, surgen nuevas necesidades como la de elaborar proyecciones a largo plazo; en esta dirección las estimaciones y proyecciones poblacionales permiten situar a los países en etapas de la transición demográfica, analizar las tendencias demográficas y los impactos futuros de estos cambios demográficos en la estructura de la población.

Ahora bien, específicamente las preguntas que podrían responder el conjunto de datos descrito -en esta práctica- son por ejemplo: ¿Cuál es la distribución de la población por continentes, subcontinentes o regiones especificas?; ¿Cuál es la variación anual de la población por países?; ¿Cuál es promedio de años para que eventualmente, la población se duplique por países en términos comparativos? Si bien, es posible que las proyecciones demográficas utilizadas para responder este tipo de preguntas les falte certeza (por las razones anotadas en párrafos anteriores de los apartados 6 y 7), pueden impulsar planes estratégicos y ser usadas por gobiernos para orientar políticas familiares y de asistencia médica. También las corporaciones las pueden emplear para decidir dónde invertir. Y en general pueden incidir en la toma de decisiones de organizaciones privadas y gubernamentales en relación al tema de las pensiones, el crecimiento económico.

8. Licencia

Al revisar el código fuente de la página web seleccionadas se identifica que la licencia esta bajo CC BY-SA 3.0, la cual permite compartir, copiar y redistribuir el material en cualquier medio o formato. Así mismo, adaptar, remezclar, transformar y construir sobre el material para cualquier propósito, incluso comercialmente. Lo anterior, siempre se conceda el crédito adecuado y proporcionando un enlace a la licencia e indicar si se realizaron modificaciones.

Dentro del listado proporcionado se selecciona CC BY-SA 4.0 License, dado que se debe proveer el nombre del creador del conjunto de datos generado, indicando los cambios que se han realizado; lo cual permite reconocimientos a los autores. Por otra parte, esta licencia consiente el uso comercial; lo cual favorece el uso de los datos generados por los empresarios y/o emprendedores. Y finalmente, las contribuciones que se realicen, bajo esta licencia, garantizan que estas mismas se lleven a cabo también por esta misma licencia; lo cual es clave para los autores, pues se les admite que continúen distribuyendo bajo los términos que el autor o autores plantearon.

9. Código utilizado

El código utilizado para la extracción de la información ya se encuentra cargado dentro del repositorio para consulta y revisión.

10. Dataset

El Dataset creado como resultado del ejercicio elaborado se encuentra cargado dentro del repositorio para consulta y revisión en formato CSV.

Contribuciones	Firma
Investigación previa	HHM
Redacción de las respuestas	HHM
Desarrollo código	HHM