

TIPOLOGÍA Y CICLO DE VIDA DE LOS DATOS

Práctica 1: PRELIMINARES

PROFESORA: MIREIA CALVO GONZALEZ

ALUMNOS

ALEJANDRO PRIETO VELASCO

CORREO ELECTRONICO: alex47@uoc.edu

OSCAR RAMIREZ GONZALEZ

CORREO ELECTRONICO: oramirezgo@uoc.edu

Solución de la PEC

El objetivo de esta actividad será la creación de un dataset a partir de los datos contenidos en una web. Para su realización, se deben cumplir los siguientes puntos:

1. Contexto. Explicar en qué contexto se ha recolectado la información. Explique por qué el sitio web elegido proporciona dicha información.





Con la situación actual de guerras en el mundo y cambio climático, unido a países del primer mundo con una alto nivel de envejecimiento y baja natalidad, es más necesario que nunca hacer un estudio lo más detallado posible del estado demográfico actual de cada país, de modo que se puedan tomar decisiones respecto como que países podrían admitir migrantes provenientes de otros países, qué países pueden aportar esa inmigración además de poder estudiar la situación socio-demográfica en otros muchos aspectos como las religiones predominantes en cada país, la población reclusa, índice de desarrollo humano o de felicidad entre otros.

La información proporcionada en los ficheros de salida CSV realizados en esta práctica, nos va a ayudar a comprender la situación demográfica de cada uno de los países. Mientras que se intenta que las personas no tengan que migrar de sus países por ningún motivo forzoso, se hace necesario conocer con detalle la situación demográfica de cada país y su posibilidad de aceptar migrantes de otros países. También nos permitirá tener en cuenta otros factores como la tasa de suicidio, índice de desarrollo humano, natalidad, religiones o tasas de alfabetización que nos ayudarán a entender mejor la realidad social de estos países.

La web <https://datosmacro.expansion.com/> proporciona mucha información de distinta índole, y entre dicha información se encuentran datos demográficos de todos los países del mundo, que utilizaremos para generar unos ficheros .csv con cada uno de los 22 índices disponibles.

2. Definir un título para el dataset. Elegir un título que sea descriptivo.

Creemos que un título que describe muy bien el contenido de nuestro dataset es: **“Índices sociodemográficos para el análisis de la migración internacional”**

Al estar dividido en 22 datasets el título que vamos a darle a cada uno de ellos va a ser el número que ocupe el link seleccionado dentro de la lista de nombres que damos como opción de entrada en el programa, más el nombre que la página haya añadido al sublink dentro de la página. Por ejemplo, si vamos a descargar  contenida  en  este  enlace (<https://datosmacro.expansion.com/demografia/indice-brecha-genero-global>) el nombre del dataset que descargamos será 1_indice-brecha-genero-global.csv al ser el primero dentro de la lista que hemos sacado con todos los datos demográficos.

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).

Hemos sacado 22 índices de la página web

<https://datosmacro.expansion.com/demografia>, que se listan a continuación:

- Índice Global de la Brecha de Género
- Población
- Inmigrantes
- Remesas de migrantes
- Emigrantes totales
- Índice de Desarrollo Humano
- Índice de Progreso Social - SPI
- Índice de Paz Global
- Índice global de envejecimiento
- Natalidad
- Mortalidad
- Esperanza de vida al nacer
- Matrimonios
- Divorcios
- Suicidios
- Homicidios Intencionados
- Población reclusa
- Riesgo de pobreza
- Índice Mundial de la Felicidad
- Pirámide de población
- Tasa de alfabetización de adultos
- Religiones

La información de cada uno de estos 22 dataframes se va a volcar en 22 ficheros .csv distintos, ya que un solo fichero .csv no admite más de una carpeta dentro del mismo, como si es el caso de ficheros .xls.

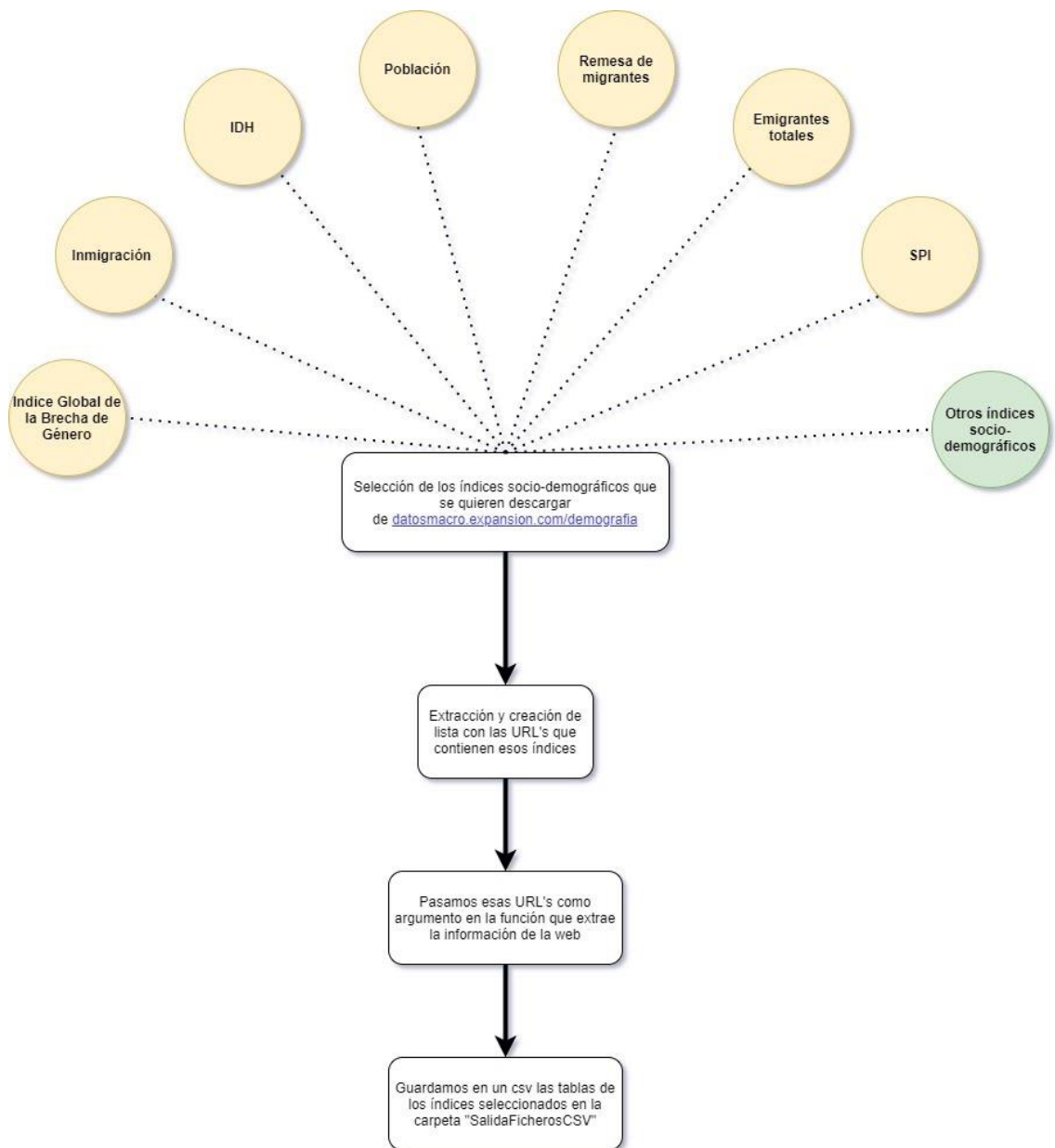
Debajo se incluye un resumen del contenido de cada fichero, y un ejemplo de la primera línea, que en todos los casos pertenece a España.

1. Índice brecha-género-global									
Países	Ranking de la Brecha de Género	Índice de la Brecha de Género	Var.						
España	8ª	0,735	6,57%						
2. Población									
Países	Fecha	Densidad	Población	Var.					
España	2019	94	47.332.614	0,46%					
3. Inmigración									
Países	Inmigrantes hombres	Inmigrantes mujeres	Inmigrantes	% Inmigrantes	Var.				
España	2.313.747	3.190.456	6.104.203	12,30%	0,15				
4. Remesas									
Países	Fecha	Saldo remesas (M.\$)	Remesas recibidas (%PIB)	Remesas recibidas (M.\$)	Remesas enviadas (%PIB)	Remesas enviadas (M.\$)	Var.		
España	2017	-1.162,00	0,81%	10.632,00	1,36%	17.874,00	1,36%		
5. Emigración									
Países	Emigrantes hombres	Emigrantes mujeres	Emigrantes	% Emigrantes	Var.				
España	666.443	778.439	1.444.342	3,05%	0,17				
6. Índice desarrollo humano									
Países	IDH	Ranking IDH	Var.						
España	0,904	25ª	0						
7. Índice progreso social									
Países	SPI	Var.							
España	87,47	-1							
8. Índice paz-global									
Países	Índice de Paz Global	Ranking Paz Global	Var.						
España	1,712	38ª	6						
9. Índice global-empoderamiento									
Países	Ranking	Índice	Externo	Competencias	Salud	Ingresos	Var.		
España	25ª	61,71	74,7	23,36	80,46	73,37	2,30%		
10. Natalidad									
Países	Fecha	Nacidos	Nacidos Hombres	Nacidos Mujeres	Tasa Natalidad	Índice de Fecund.	Var.		
España	2019	360.617	185.523	175.094	7,62%	1,24	-2,24%		
11. Mortalidad									
Países	Fecha	Muertes	Tasa mortalidad	Var.					
España	2019	416.703	6,83%	-0,26					
12. Esperanza de vida									
Países	Fecha	Esperanza de vida - Mujeres	Esperanza de vida - Hombres	Esperanza de vida	Var.				
España	2020	85,1	79,7	82,4	-1,90%				
13. Matrimonios									
Países	Fecha	Matrimonios	Edad Media primer mat. mujeres	Edad Media primer mat. hombres	Tasa de primer matrimonio - mujeres	Tasa de primer matrimonio - hombres	Tasa bruta de nupcialidad	Var.	
España	2019	161.383	33,3	36,1	0,45	0,43	3,51%	-0,04	
14. Divorcios									
Países	Fecha	Divorcios	Tasa bruta de divorcios	Var.					
España	2019	31.645	1,30%	-0,1					
15. Suicidio									
Países	Fecha	Suicidios mujeres	Suicidios hombres	Suicidios	Suicidios tasa femenina	Suicidios tasa masculina	Suicidios por 100.000	Var.	
España	2018	920	2.619	3.539	3,84	11,38	7,54	-3,46%	
16. Homicidios									
Países	Fecha	Número de Homicidios	Homicidios Mujeres	Homicidios Hombres	Homicidios Mujeres por familiares	Homicidios hombres por familiares	Homicidios por 100.000	Var.	
España	2018	290	117	173	67	24	0,62	-5,63%	
17. Población carcelaria									
Países	Fecha	Total reclusos	Reclusos por 100.000	Var.					
España	2018	58.883	126,1	0					
18. Riesgo-pobreza									
Países	Fecha	Personas en riesgo de pobreza	Umbral persona	Umbral hogar	% Riesgo Pobreza	Var.			
España	2019	3.610 m.	9.0091	18.3191	20,70%	-0,8			
19. Índice-felicidad									
Países	Ranking Felicidad	Índice Felicidad	Var.						
España	27ª	6,431	1,41%						
20. Estructura-población									
Países	Fecha	0-14 años %	15-64 años %	> 64 años %	Var.				
España	2019	14,53%	65,83%	19,50%	0,1				
21. Tasa-alfabetización									
Países	Fecha	Tasa de alfabetización mujeres	Tasa de alfabetización hombres	Tasa de alfabetización de adultos	Tasa de alfabetización jóvenes	Var.			
España	2018	97,37%	98,93%	98,44%	99,72%	0,09			
22. Religiones									
Países	Creyentes	Cristianismo	Islam	Hinduismo	Sincretismo	Budismo	Animismo	Judaísmo	Taoísmo
España	84,00%	80,56%	3,04%		0,04%	0,02%		0,03%	

Como se puede observar en la tabla resumen, tenemos parámetros tipo string, fecha, integer y float. En todos los casos aparece, como poco, el nombre del país y la variación respecto al año anterior. Este parámetro es relevante para saber la evolución de ese país para cada uno de los parámetros medidos.

4. Representación gráfica. Presentar esquema o diagrama que identifique el dataset visualmente y el proyecto elegido.



Índices sociodemográficos para el análisis de la migración internacional



5. Contenido. Explicar los campos que incluye el dataset, el periodo de tiempo de los datos y cómo se ha recogido.

Cada uno de los 22 índices tiene tipos de datos distintos. Como un ejemplo, debajo se incluye el aspecto de un par de índices:

<< 2018 Comparativa: Riesgo de pobreza						
Países	Fecha	Personas en riesgo de pobreza	Umbral persona	Umbral hogar	% Riesgo Pobreza	Var.
España [+]	2019	9.610 m.	9.009 €	18.919 €	20,7%	-0,8
Alemania [+]	2015	13.428 m.	12.401 €	26.041 €	16,7%	0

<< 2019 Comparativa: Índice de Paz Global 2020			
Países	Índice de Paz Global	Ranking Paz Global	Var.
España [+]	1,712	38° 	6
Alemania [+]	1,494	16° 	-6

En el dataset de cada índice se va a incluir la lista de parámetros que se puede encontrar en cada enlace. El principal problema que tiene este web scraping es la heterogeneidad de los campos que hay en cada uno de los índices. A parte de que cada índice tiene distinto número de campos, tenemos el problema de que, dependiendo de cada índice, un campo de una columna dada tiene un significado distinto al campo que hay en la misma posición de otro índice.

Todos los campos contienen valores ya indicados con anterioridad: fecha, string, integer y float, y no hay mayores problemas en el tratamiento de los mismos, más allá de los caracteres específicos de la lengua española, como la ñ o los acentos, y que se ha tratado con la codificación necesaria a la hora de crear los ficheros .csv, para facilitar su manejo a quién realice un análisis de dichos ficheros.

Cada índice tiene un año distinto de recogida del mismo, sin llegar a ser este homogéneo. Como evolución de este web scraping, existe la posibilidad de elegir cada índice por el año requerido, u obtener el mismo índice a lo largo de un periodo de varios años, solo introduciendo dicho año en la URL del índice en cuestión, como se puede ver a continuación:

- Índice “población” del último año disponible, en este caso 2020:

<https://datosmacro.expansion.com/demografia/poblacion>

- Índice “población” del año 2019:

<https://datosmacro.expansion.com/demografia/poblacion?anio=2019>

Como se puede observar es tan sencillo como añadir en la cadena “?anio=2019” a la URL original de ese índice.

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.

Agradecemos al periódico de la Expansión y en concreto a su repositorio Datosmacro.com por la información que proporcionan de forma totalmente gratuita. Nosotros hemos decidido centrarnos en su sección de demografía (<https://datosmacro.expansion.com/demografia>). Datosmacro es una web del periódico La Expansión que recoge índices macroeconómicos de fuentes como organismos oficiales, institutos nacionales de diferentes países, bancos centrales, etc. con el objetivo de proporcionar una visión global desde un punto de vista socio - económico.

Aunque a nivel de calidad del estudio, accesos a recursos y magnitud del proyecto no es comparable las fuentes en las que nos hemos inspirado son los informes International Migration Report de la UN (United Nations, 2021) y el World Migration Report 2020 del IOM (International Organization for Migration, 2019) en los cuales se estudian las causas y consecuencias de los movimientos migratorios y los relacionan con problemas como la actual pandemia o el cambio climático. Estos organismos recogen información directamente de los estados y hacen un estudio exhaustivo sobre que va mucho más allá del contenido de esta asignatura sobre los movimientos migratorios durante 2020.

Nuestro objetivo es poder ofrecer una herramienta que sirva para que investigadores en esta área, puedan acceder de una manera cómoda y sencilla a la información sobre los índices socio-demográficos que necesiten, sin tener que ir, por ejemplo, a las webs de los institutos nacionales de estadística de cada país para recoger la información o tener que buscar estos en algún otro recurso. De esta manera pueden hacerse una primera impresión sobre los índices que se estén estudiando y pueden complementar la información que aportan estos con otros índices que estén disponibles para descargar dentro del programa. Además, la salida del programa serán una serie de documentos CSV que permitirán que el investigador pueda explorar y trabajar esta información con herramientas como Excel.

Hemos decidido usar la web de datosmacro para obtener esta información ya que reúne información actualizada sobre varios de estos índices, además, no tiene ningún aviso legal o restricción en lo que a web scraping se refiere, incluido en el fichero robots.txt.

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

Este dataset permitirá descargar de una forma sencilla una serie de indicadores sociodemográficos de distintos países, nuestra idea inicial es que sirva para entender mejor los movimientos migratorios que se producen en el mundo y conocer mejor la realidad socio-demográfica de una serie de países para entender por qué la gente emigra o inmigra a un país en concreto.

Con indicadores como el IDH, SPI, felicidad, etc. se puede crear obtener una valiosa información que de forma sencilla refleja la realidad de un determinado país. Además, se pueden buscar correlaciones con el fin de poder predecir qué

países van a tener más inmigrantes o emigrantes, y de este modo poder hacer una previsión de movilidad de gente para facilitar la acogida en países que puedan realizar dicha acogida, y prever ayudas para aquellos países que tengan emigración de sus ciudadanos a otros países, evitando la descapitalización humana de los mismos, que es una de las peores cosas que le puede ocurrir a un país como estamos viendo, por ejemplo, en el caso de España, donde pagamos la formación a personas que terminan dando sus servicios en otros países.

Comparando nuestra extracción de datos con los datos de los que disponen en los informes International Migration Report de la UN (United Nations, 2021) y el World Migration Report 2020 del IOM (International Organization for Migration, 2019), es evidente que la magnitud y la calidad de los proyectos no tienen punto de comparación, pero nuestro objetivo no es realizar un análisis socio-demográfico exhaustivo sino más bien como se ha mencionado anteriormente, dar una idea genérica de la situación socio-demográfica de cada país de manera que se pueda entender mejor los movimientos migratorios y que sirva al investigador como punto de partida para realizar un análisis más en profundidad.

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

- Released Under CC0: Public Domain License
- Released Under CC BY-NC-SA 4.0 License
- Released Under CC BY-SA 4.0 License
- Database released under Open Database License, individual contents under Database Contents License
- Other (specified above)
- Unknown License

La página web de la que hemos recogido los datos se encuadra en el grupo de licencias Released Under CC0: Public Domain License. En el apartado de términos y condiciones no se hace una referencia explícita a limitaciones en el uso de los datos, más allá del descargo de responsabilidad por el uso de los mismos. No se pide hacer referencia a la fuente de los mismos, ni que los datos, documentos o bases de datos resultantes tengan que publicarse bajo una licencia específica.

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

Código incluido en el repositorio de Github.

<https://github.com/alex47ST3/Indices-sociodemograficos-para-el-analisis-de-la-inmigracion-internacional>

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

<https://zenodo.org/record/4657519#.YGXu-a8zZEY>

DOI: 10.5281/zenodo.4657519

Bibliography

International Organization for Migration. (2019). *WORLD MIGRATION REPORT 2020*. IOM Online Bookstore. Retrieved 04 01, 2021, from https://publications.iom.int/system/files/pdf/wmr_2020.pdf

United Nations. (2021, 01 15). *International Migration 2020 Highlights*. un.org. Retrieved 04 02, 2021, from <https://www.un.org/en/desa/international-migration-2020-highlights>

Firma de autores

Contribuciones	Firma
Investigación previa	A.P.V - O.R.G
Redacción de las respuestas	A.P.V - O.R.G
Desarrollo código	A.P.V - O.R.G