Dataset: COVID-19 WORLDWIDE

Fernando Alamo, Daniel Padilla Ortega

1 de abril de 2020

1. Contexto

En diciembre de 2019, un nuevo tipo de coronavirus, del SARS-CoV-2, aparece en la ciudad china de Wuhan y, tan solo tres meses más tarde, pone en jaque al mundo entero.

El día 11 de marzo de 2020, el director general de la Organización Mundial de la Salud (OMS), Tedros Adhanom Ghebreyesus, declaró que el coronavirus COVID-19 pasa de ser una epidemia a una pandemia.

En ese mismo comunicado, recordaba a todos los países la necesidad de activar y ampliar sus mecanismos de respuesta de emergencia para frenar el coronavirus COVID-19. Unos días después, las consecuencias son miles de personas contagiadas y fallecidas, hospitales desbordados, supermercados desabastecidos y economías camino del colapso. Frente a esta situación todos los países del mundo y numerosas compañías farmacéuticas comienzan una carrera para obtener una vacuna lo antes posible. Finalmente tras 10 meses de investigación se pone en producción la primera vacuna contra el Covid-19, la vacuna de Pfizer. Pronto comienzan a aparecer nuevas vacunas, sin embargo ante la cantidad masiva de demanda por parte de los países para inmunizar a su población la cadena de producción y distribución de las farmacéuticas se resiente puesto que la demanda de vacunas supera ampliamente a la oferta.

Las negociaciones de los distintos organismos con las farmacéuticas serán clave para determinar la velocidad en la vacunación en los diferentes países.

En este dataset se recogen las variables más importantes que definen la evolución de la pandemia en cada país que reporta información, durante los últimos 100 días. Estos datos permiten analizar la evolución de la pandemia y de las vacunaciones a lo largo del mundo.

Con este dataset podemos responder preguntas como las siguientes:

- ¿Qué país tiene una mayor tasa de incidencia en los últimos 14 días?
- ¿Qué país tiene una mayor tasa de vacunación en su población?
- ¿Existe una correlación entre la tasa de vacunación y la tasa de incidencia en los últimos 14 días?

Datosmacro tiene como objetivo ofrecer las principales variables económicas y sociodemográficas de cientos de países, para ofrecer una visión global de la situación económica en cada momento y en cada país. En los últimos años se ha convertido en un referente en la red para hispanohablantes de datos económicos y sociodemográficos por su fiabilidad y claridad.

Sus fuentes son los organismos oficiales de los distintos países y zonas a las que corresponden los datos: https://datosmacro.expansion.com/legal/fuentes

Esta página presenta información de manera estructurada y fácilmente accesible incluyendo los últimos meses información sobre la evolución de la pandemia por tanto consideramos que es una buena fuente de datos a la hora de referenciar nuestro proyecto.

3. Descripción del dataset

El conjunto de datos generado como parte de esta actividad práctica reúne diferentes características de la propagación del virus SARS-CoV-2, los efectos de la enfermedad COVID-19 sobre la población segmentada por países durante los últimos 100 días y la evolución de la vacunación. Algunas de las variables que se recogen en el conjunto de datos son el país, la fecha, el incremento de muertos, número de muertos totales . . .

Contiene datos de los últimos 99 días de los 188 países que reportan información oficial, generando un dataset con un total de registros de 18770, con 10 variables por registro.

4. Representación gráfica

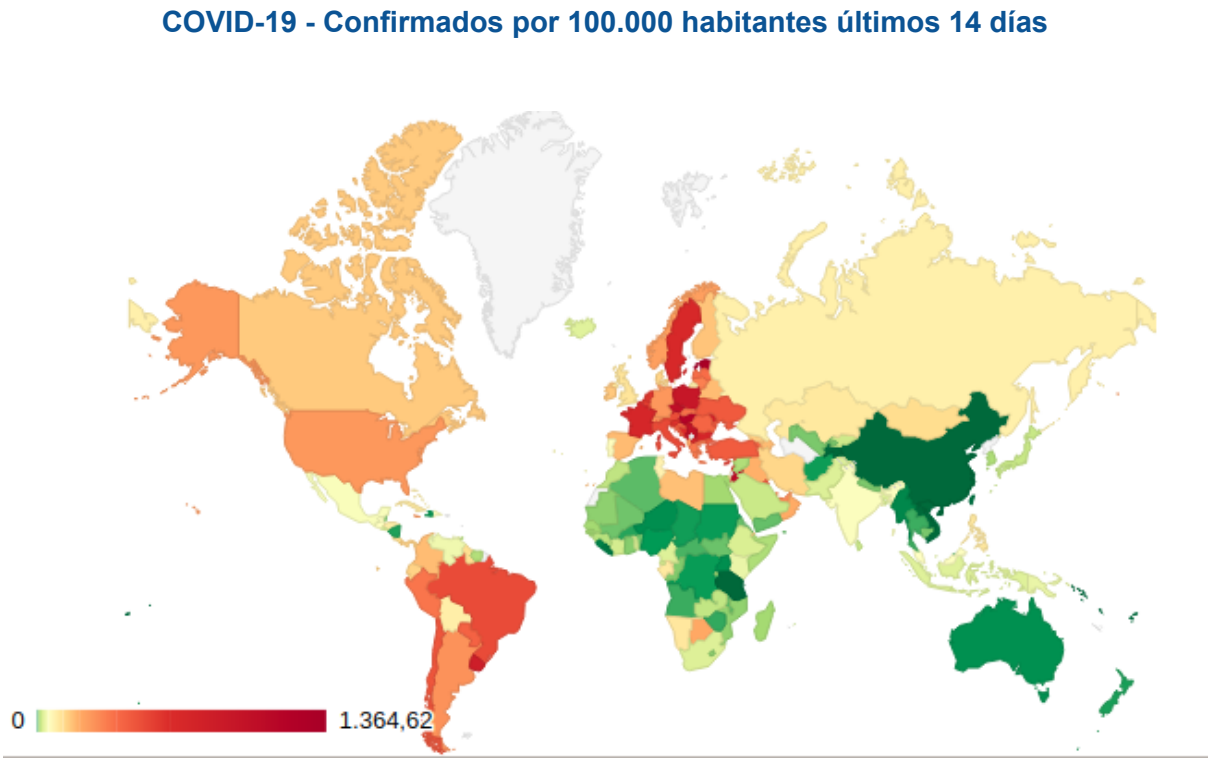

Figura 1: Estado de la pandemia provocada por el SARS-CoV-2 en el mundo a 1 de abril de 2020. La figura muestra un mapa de calor con el estado actual de la incidencia medida por la media móvil de los últimos catorce días, de los contagios registrados por 100.000 habitantes.

Esta cifra se usa para determinar el nivel de riesgo por país, clasificándolos en riesgo bajo, moderado, alto, muy alto, y extremo. Es un factor determinante para posibles medidas de restricción de movilidad entre países, pudiéndose restringir los movimientos con origen de países en riesgo muy alto o extremo.

5. Contenido

Para cada registro se recogen las siguientes características:

País: País en el que se obtienen los datos en formato string en idioma español.

Fecha: Fecha en la que se recogen los datos en formato date “dd/mm/yyyy”.

Incremento_Muertos: Defunciones registradas en el día de referencia en formato int.

Muertos: Total de defunciones registradas desde el inicio de la pandemia en formato int.

Muertos_millon: Número de muertos registrados por millón de habitantes en formato float.

Incremento_Confirmados: Número de nuevos casos confirmados en formato int.

Confirmados: Total de casos confirmados desde el inicio de la pandemia hasta la fecha en formato int.

Confirmados_100k_14d: Media móvil de los casos positivos de los últimos 14 días por cien mil habitantes en formato float.

Dosis: Número total de dosis suministradas hasta la fecha a la población. Formato int.

Personas_vacunadas: Número total de personas hasta la fecha a las que se les a suministrado una dosis de la vacuna. Formato int.

Completamente_vacunadas: Número total de personas hasta la fecha que han recibido las dos dosis de la vacuna. Formato int.

%_completamente_vacunadas: Porcentaje de personas completamente vacunadas hasta la fecha sobre la población total. Formato float.

Los autores de la web https:// datosmacro.expansion.com/ recopilan información sobre la evolución del COVID-19 desde el inicio de la pandemia. Lamentablemente esta página sólo permite el acceso a los datos por país de los últimos 99 días, por lo que podría usarse este scraper para generar una BBDD acumulativa desde el momento que se comenzara a usar de forma recurrente.

Las fuentes utilizadas por datos macro.expansión son el Ministerio de Sanidad de España y The Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU).

6. Agradecimientos

Los datos han sido recolectados desde la base de datos online DatosMacro. Para ello, se ha hecho uso del lenguaje de programación Python y de técnicas de Web Scraping para extraer la información alojada en las páginas HTML.

Tal y como se ha comentado con anterioridad Datosmacro tiene como objetivo ofrecer las principales variables económicas y sociodemográficas de cientos de países, para ofrecer una visión global de la situación económica en cada momento y en cada país. Se puede encontrar análisis de todo tipo, desde deuda de los municipios de España (https://datosmacro.expansion.com/deuda/espana/municipios) hasta el ranking por paises del Indice de Desarrollo Humano (https://datosmacro.expansion.com/idh).

7. Inspiración

El presente conjunto de datos puede ser utilizado desde diferentes ámbitos. Nuestra principal inspiración procede de los trabajos realizados por la OMS (https://covid19.who.int/) y la Johns Hopkins University (https://coronavirus.jhu.edu/map.html) en referencia sus análisis realizados acerca de las principales variables sobre la pandemia del Covid-19. Estas dos entidades han conseguido integrar y proponer trabajos de análisis y síntesis que son presentados en un dashboard lo que supone que sean fácilmente accesibles, navegables y explotables por cualquier tipo de usuario sin necesidad de disponer de conocimientos técnicos acerca de la materia.

Con este dataset se pretende extraer la información necesaria para poder diseñar y mantener herramientas como los dashboard presentados por la OMS y la JHU y que sirvan para un uso general.

Destacamos los principales agentes que pueden beneficiarse de este conjunto de datos y como:

Empresas y organizaciones privadas: Para analizar la evolución de la pandemia y las vacunaciones y así optimizar su toma de decisiones en cuanto a su actividad.

Particulares: Para disponer de información que les permita mantener actividades en entornos seguros y prever el riesgo en cada momento. Para analizar y comparar los efectos de las medidas sanitarias implementadas en los diferentes países.

Medios de comunicación: Para analizar y comparar los efectos de las medidas sanitarias implementadas en los diferentes países. Para informar sobre la evolución de la pandemia.

8. Licencia

La licencia escogida para la publicación de este conjunto de datos ha sido GNU GENERAL **PUBLIC LICENSE v3.0.** Los motivos que han llevado a la elección de esta licencia tienen que ver con la idoneidad de las cláusulas que esta presenta en relación con el trabajo realizado:

Se debe proveer el nombre del creador del conjunto de datos generado, indicando los cambios que se han realizado. De esta manera, se reconoce el trabajo ajeno y en qué medida se han realizado aportaciones en relación con el trabajo original.

Se permite un uso comercial. Esto haría que incrementen las probabilidades de que una empresa utilice los datos generados y realicen trabajos de calidad que reporten cierto reconocimiento al autor original.

Las contribuciones realizadas a posteriori sobre el trabajo publicado bajo esta licencia deberán distribuirse bajo la misma. Esto hace que el trabajo del autor original continúe distribuyendo bajo los términos que él mismo planteó.

9. Código fuente y dataset

Tanto el código fuente escrito para la extracción de datos como el dataset generado pueden ser accedidos a través de este enlace.