Tipología y ciclo de vida de los datos - PRAC1

Generación del dataset COVID-19_WORLD_2020 mediante los datos obtenidos a través de técnicas de *web scraping*

David Marín Sánchez
Abril de 2020

Contexto

El contexto en el que se enmarca esta práctica es la situación de pandemia de COVID-19 a la que se enfrenta el planeta y, como consecuencia, el estado de alarma decretado por el gobierno español que impone sobre la población el confinamiento, haciendo difícil mantener la mente alejada de un tema distinto. Es en este contexto donde surge el objetivo de obtener información lo más actualizada posible sobre los casos de COVID-19 alrededor del mundo.

Para obtener esta información, y poder generar el correspondiente dataset, se ha utilizado la página web https://www.worldometers.info/coronavirus/, que contiene una recopilación de datos globales de la pandemia.

Título, descripción y contenido

El título del *dataset* es COVID-19_WORLD_2020 y se almacena en el fichero CSV, publicado en zenodo, con nombre covid-19_world_2020.csv. Los datos que contiene se han obtenido a través de técnicas de web scraping sobre la web worldometers, implementadas en el notebook covid-19_world_scraper.ipynb. Cabe destacar que la estructura de la página web analizada ha ido cambiando a lo largo de los días, lo que ha hecho necesario modificar el código varias veces.

El dataset contiene datos relacionados con la evolución cuantitativa de los casos de coronavirus a lo largo de finales del mes de marzo y principios del mes de abril de 2020. Cada registro (fila) del dataset está identificado por el país (*conuntry*) y la fecha (*date*) de los datos.

Los atributos (columnas) del dataset contienen los datos recopilados desde finales de marzo a principios de abril de 2020, con periodicidad diaria:

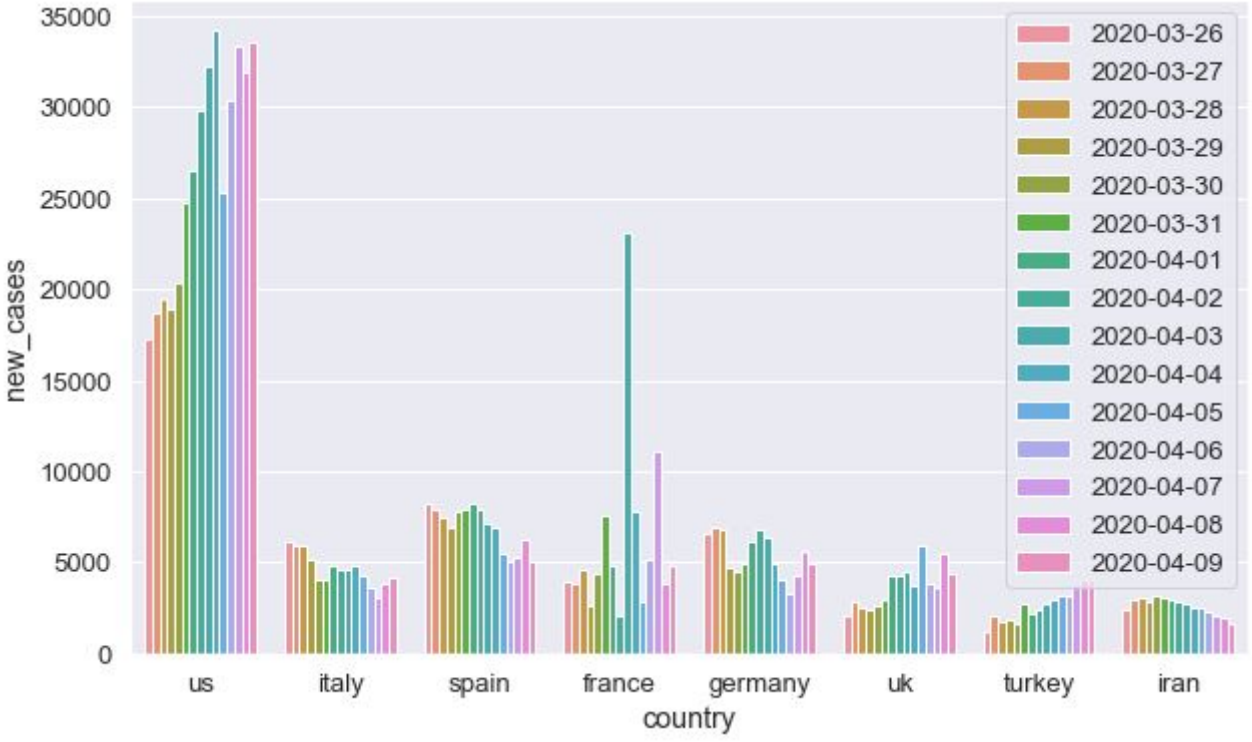
- *total_cases*: Número total de casos confirmados.
- *new_cases*: Número de nuevos casos confirmados respecto al día anterior.
- *total_deaths*: Número total de muertes confirmadas.
- *new_deaths*: Número de nuevas muertes confirmadas respecto al día anterior.
- *total_recovered*: Número de casos recuperados confirmados.
- *active_cases*: Número de casos activos confirmados.

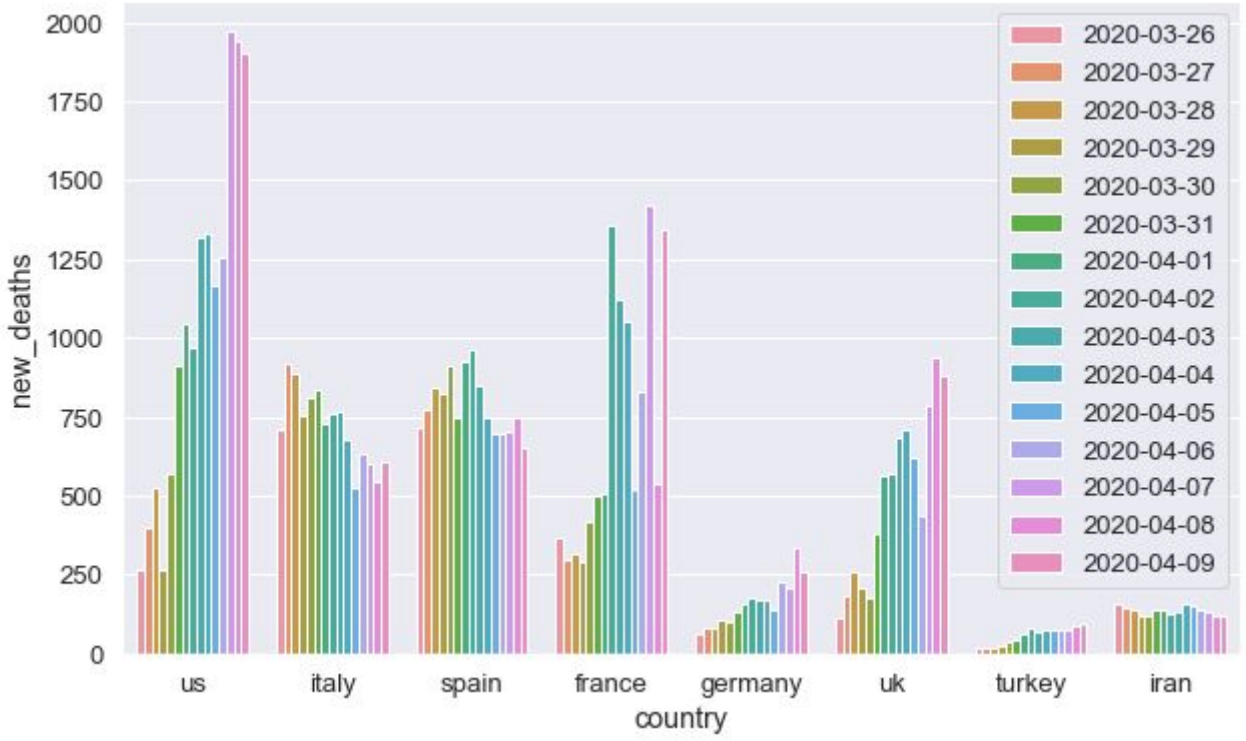
- *servious_critical*: Número de casos serios o críticos confirmados.
- *total_cases_1M_pop*: Número total de casos confirmados por cada millón de habitantes.
- *total_deaths_1M_pop*: Número total de muertes confirmadas por cada millón de habitantes.
- *total_tests*: Número de tests realizados.
- *tests_1M_pop*: Número de tests realizados por cada millón de habitantes.

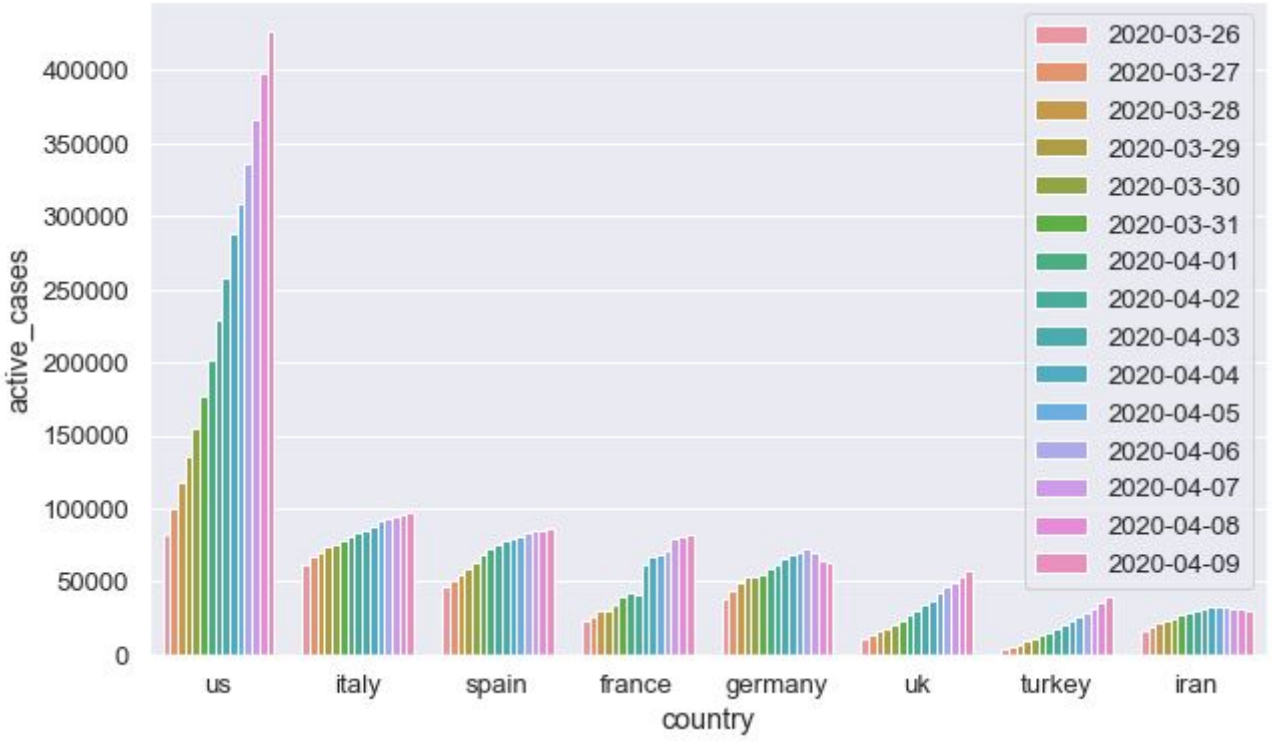
Es importante tener en cuenta que no existe un único criterio mundial de como contabilizar casos y defunciones y, además, existe una infradetección de casos por la limitada cantidad de pruebas para confirmar la infección que se están realizando. Por lo tanto, los valores del dataset muy probablemente sean distintos a los valores reales, pero de todas formas pueden ser útiles para observar tendencias y desarrollar modelos predictivos.

Representación gráfica

En las siguientes imágenes se pueden observar algunos de los datos contenidos en el dataset. En concreto se puede ver la evolución temporal del número de nuevos casos confirmados, de nuevas defunciones confirmadas y de casos activos confirmados, en los países con mayor número de casos activos confirmados, de los últimos días del mes de marzo y primeros días del mes de abril de 2020.




new_deaths
2000
1750
1500
1250
1000
750
500
250
0
us
italy
spain
france
germany
uk
turkey
iran
country
2020-03-26
2020-03-27
2020-03-28
2020-03-29
2020-03-30
2020-03-31
2020-04-01
2020-04-02
2020-04-03
2020-04-04
2020-04-05
2020-04-06
2020-04-07
2020-04-08
2020-04-09


active_cases
400000
350000
300000
250000
200000
150000
100000
50000
0
us
italy
spain
france
germany
uk
turkey
iran
country
2020-03-26
2020-03-27
2020-03-28
2020-03-29
2020-03-30
2020-03-31
2020-04-01
2020-04-02
2020-04-03
2020-04-04
2020-04-05
2020-04-06
2020-04-07
2020-04-08
2020-04-09

La disparidad en la tasa de letalidad observada para distintos países puede ser debida, al menos en parte, a los distintos niveles de detección y la disparidad de criterios que siguen cada uno de los países para contabilizar casos y defunciones.

Inspiración y agradecimientos

El conjunto de datos es de interés por la grave crisis sanitaria y socioeconómica mundial que está suponiendo la pandemia de COVID-19 del 2020. Con estos datos se pretenden responder preguntas relacionadas con la distribución y evolución temporal de la pandemia. Además, con la cantidad suficiente de datos, se podría llegar a desarrollar un modelo epidemiológico eficaz para predecir la evolución futura.

Agradecimiento especial a la web worldometers que se encarga de recopilar los datos mundiales sobre la pandemia de COVID-19 de 2020, facilitando en gran medida la tarea objetivo de este trabajo.

Agradecimiento también a otros sitios web que están recopilando datos y realizando análisis sobre este tema y que sirven como motivación de esta práctica. A continuación se muestran dos ejemplos destacados:
https://analisi.transparenciacatalunya.cat/Salut/Registre-de-test-de-COVID-19-realitzats-a-Cataluny/xuwf-dxjd
https://biocomsc.upc.edu/en/covid-19/informativedocument
http://covid19.webs.upv.es/
https://jmico.blogs.upv.es/
https://lab.montera34.com/covid19/

Licencia

La licencia bajo la que se publica el conjunto de datos es CC0: Public Domain License
https://creativecommons.org/publicdomain/zero/1.0/deed.es