

Dataset: Estadísticas de Contagiados, Confirmados y Muertos por Corona Virus 2019-2020 Mundial

Luis Alberto Caro Saldivia

7 Abril de 2021

Descripción

En esta actividad se trabajó con algunos datasets sobre la contaminación del **CORONA VIRUS** en *Europa, Africa, Reino Unido, Norteamérica, Asia y Latinoamérica* desde los años 2019 y 2021. Los datos se obtuvieron mediante algoritmos de *webscraping* utilizando la librería **BeautifulSoup** recopilando las estadísticas de cantidad de personas contagiadas, confirmadas, muertas, su procedencia y totales.

Imagen Contexto

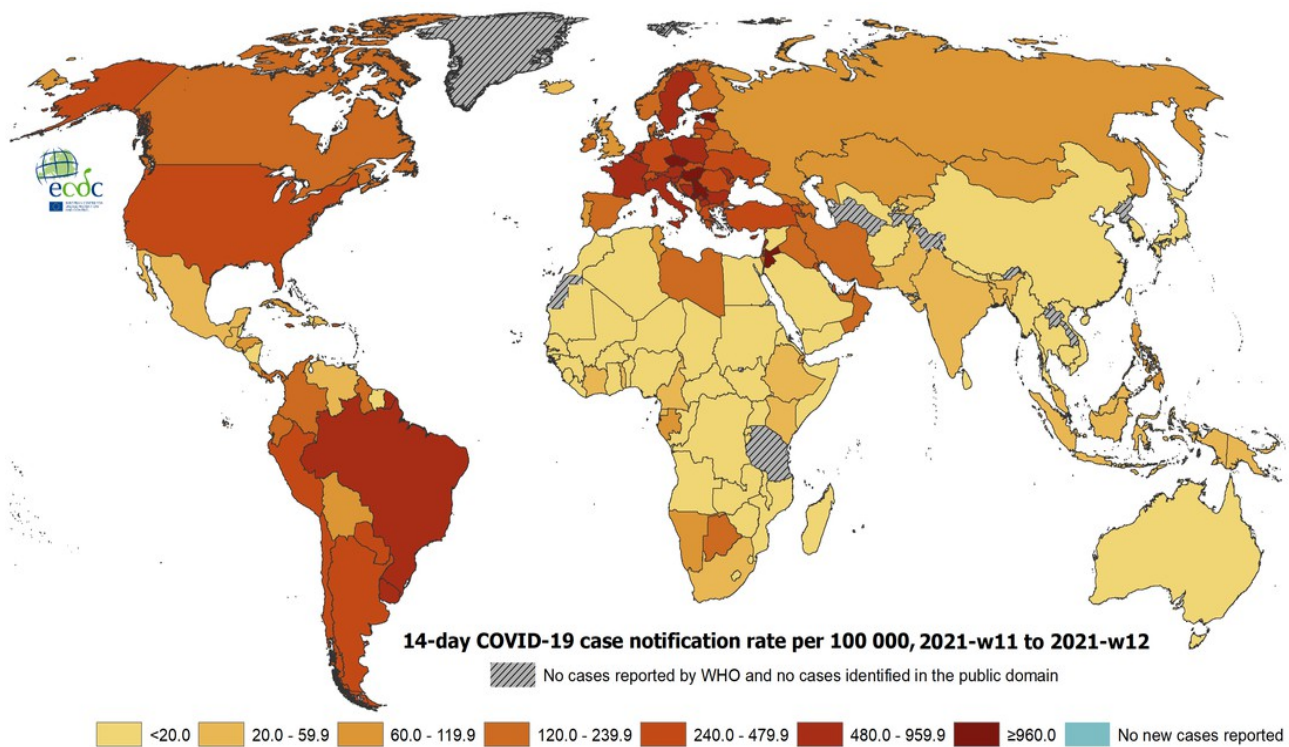


Figura 1: Mapa contaminación mundial de COVID 19

Fuente: <https://www.ecdc.europa.eu/en/geographical-distribution-2019-ncov-cases>

Contexto

La pandemia mundial que estamos viviendo desde finales del 2019 ha transformado al mundo. No solo ha tenido severas repercusiones sanitarias si no que ha transformado el curso de la historia de todos los países. La economía mundial se ha visto tremendamente afectada generando una gran recesión mundial y afectado a todas las sociedades de diferentes continentes.

Grandes pérdidas de vidas humanas con mucho sufrimiento y un intento por ganar el juego a la naturaleza para controlar este virus es fuente de mucha investigación y trabajo colaborativo a nivel mundial. Dado que esta pandemia ha puesto al mundo y sus instituciones gubernamentales en *Jaque Mate*, me interesa conocer su avance y su control sanitario a nivel mundial. Esas son las razones principales de la elección de este dataset.

Contenido

El dataset que se ha obtenido de <https://www.ecdc.europa.eu/en/geographical-distribution-2019-ncov-cases> posee las siguientes características:

- **Índice:** Se utiliza para identificar el país o región de origen. **Tipo Numérico.**
- **País:** Nombre del país estudiado. Tipo String.
- **Contagiados:** Número de personas contagiadas por *Corona Virus*. **Tipo Numérico.**
- **Confirmados:** Número de personas confirmadas (POSITIVOS) con *Corona Virus*. **Tipo Numérico.**
- **Muertos:** Número de personas muertas por *Corona Virus*. **Tipo Numérico.**
- **Total Contagiados:** Suma total de personas contagiadas. **Tipo Numérico.**
- **Total Confirmados.** Suma total de personas confirmadas. **Tipo Numérico.**
- **Total Muertos:** Suma total de personas fallecidas. **Tipo Numérico.**

Esta información se puede consultar en el archivo **data_covid_10_19_2021.csv**. El *Script* que realiza esta operación es **ws_01.py**.

Además, se han obtenido varios archivos de estadísticas relacionadas del impacto del *Corona Virus* a nivel mundial, los cuales se pueden consultar en: <https://www.ecdc.europa.eu/en/publications-data/data-covid-19-vaccination-eu-eea>.

Toda la información de los siguientes archivos fue obtenida mediante web scraping (**ws_02.py**):

- **data.csv**
- **data.json**
- **data.xlsx**
- **data.xml**

Agradecimientos

Toda esta información es registrada periódicamente y está disponible desde el portal del **Centro Europeo para la Prevención y Control de Enfermedades** (<https://www.ecdc.europa.eu.europa>).

También se consultó la información de contagios y muertes de COVID en el **Centro de Control de Enfermedades de Nigeria (NCDC)** (<https://covid19.ncdc.gov.ng/>).

Todos los datasets son datos abiertos y públicos. Se agradece estos repositorios de abundantes datos sanitarios actualizados sobre la pandemia del Corona Virus.

Para la extracción de los datasets señalados se utilizó el lenguaje de programación **Python** (<https://python.org>) y la librería de *web scraping* **BeautifulSoup** (<https://www.crummy.com/software/BeautifulSoup/bs4/doc/>).

Inspiración

También se trabajó con el ambiente **Anaconda Navigator 1.10** mediante **Jupyter Notebook**, el cual ofrece muchas características que facilitan la programación y pruebas de los algoritmos. Además, cuenta con una abundante batería de librerías y módulos especializados. En la figura 2 se aprecia este ambiente de trabajo.

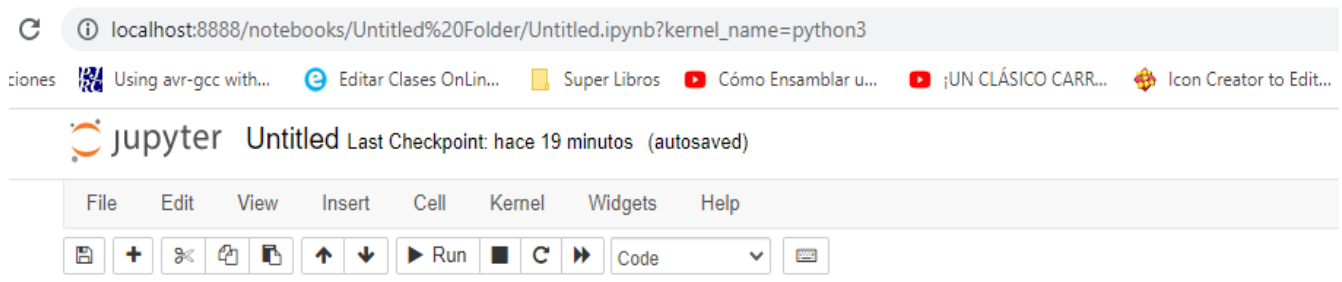


Figura 2

Por ejemplo, el *Script* **ws_05.py** fue desarrollado en esta plataforma como se aprecia en las figuras 3 y 4.

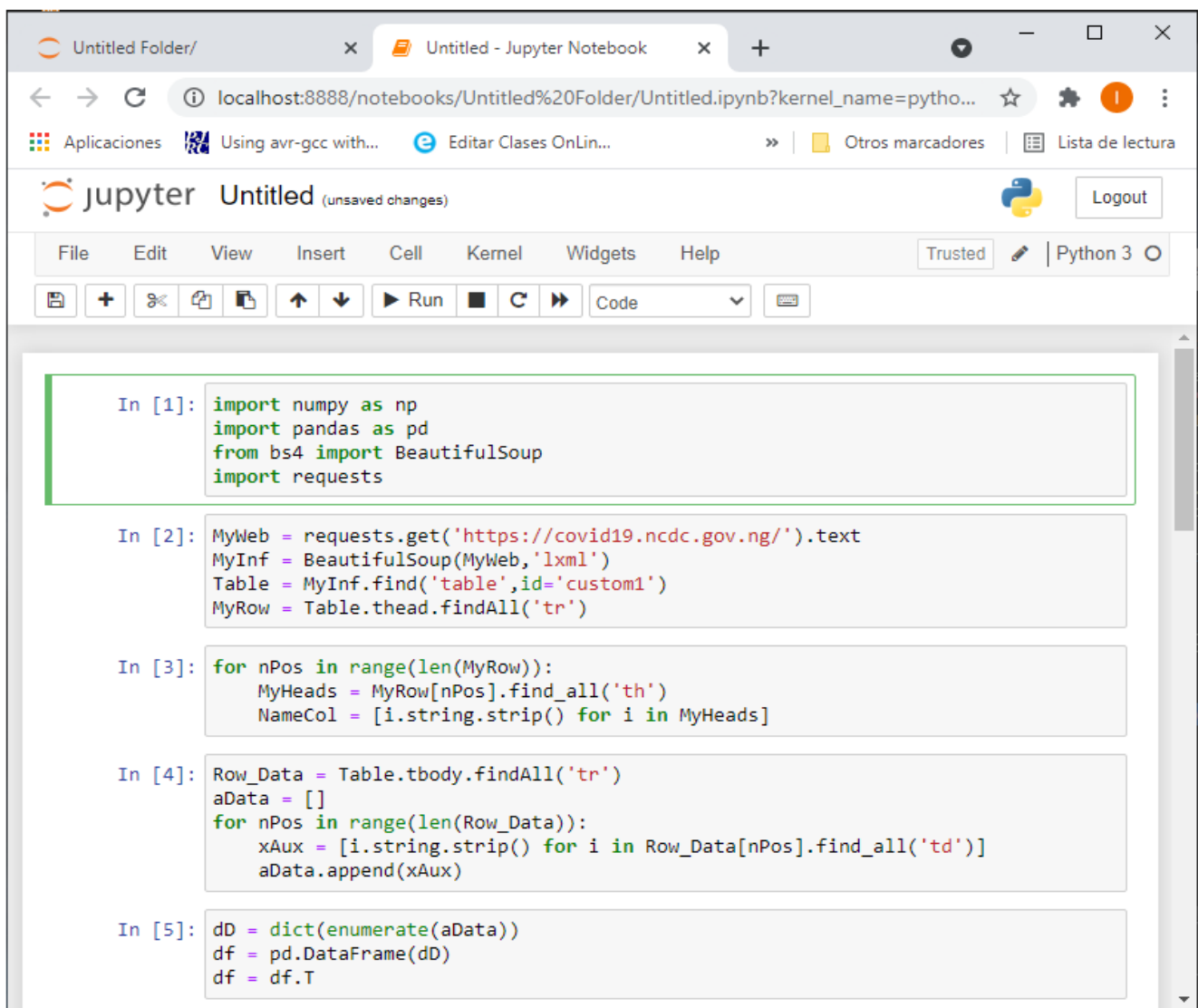


Figura 3

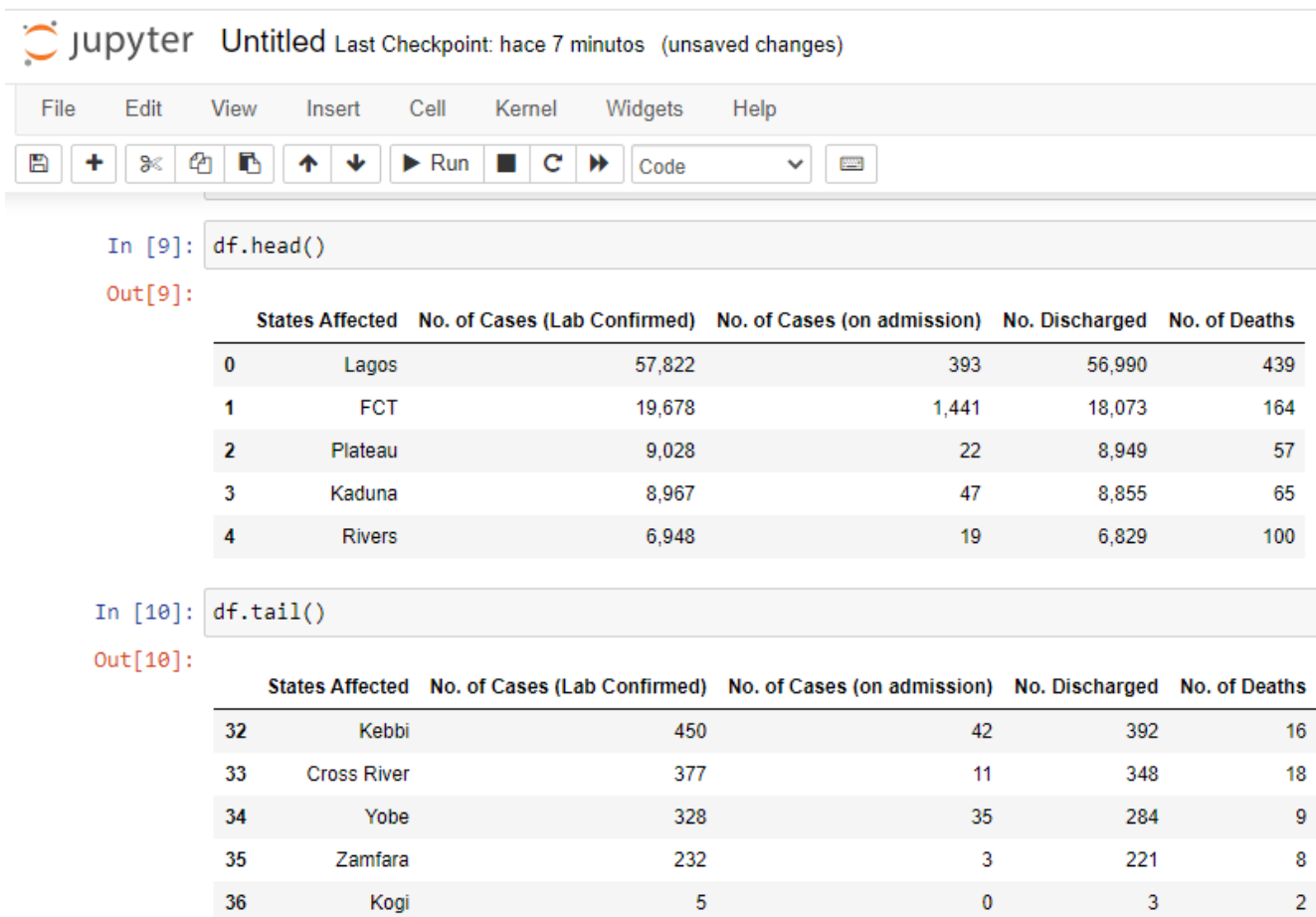


Figura 4

El esquema anterior es el **DataFrame** de **Pandas** que presenta de manera matricial la información recopilada por el *Script ws_5.py* del seguimiento de la pandemia del **Centro de Control de Enfermedades de Nigeria (NCDC)** (<https://covid19.ncdc.gov.ng/>).

Otra forma de trabajar que es bastante cómoda y fácil es mediante la utilización de algunas herramientas de web scraping que pueden ser instaladas en los navegadores (extensiones). Algunos ejemplos de extensiones:



Web Scraper - Free Web Scraping

Ofrecido por: webscraper.io



Data Scraper - Easy Web Scraping

Ofrecido por: data-miner.io



Grepsr - Web Scraping Tool

Ofrecido por: <https://www.grepsr.com>



Datap - AI-Powered Web Scraping

Ofrecido por: <https://datap.io>

Existen muchas extensiones para todos los navegadores (Chrome, Opera, FireFox, IE, etc.).

Por ejemplo, se utilizó la extensión **Data – AI- Powered Web Scraping** en **Chrome** la cual permitió exportar algunos registros de la página web <https://covid19.ncdc.gov.ng/> a formato **JSON**. Las figuras 5, 6, 7 y 8 detallan su utilización.

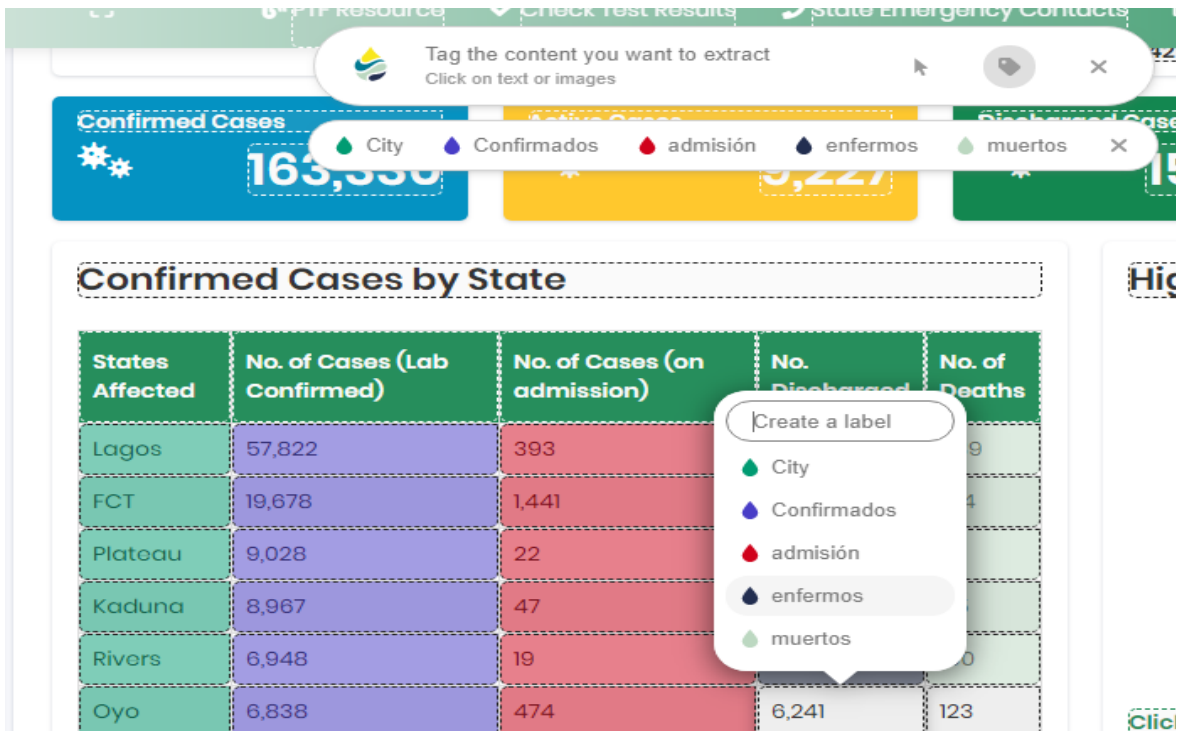


Figura 5

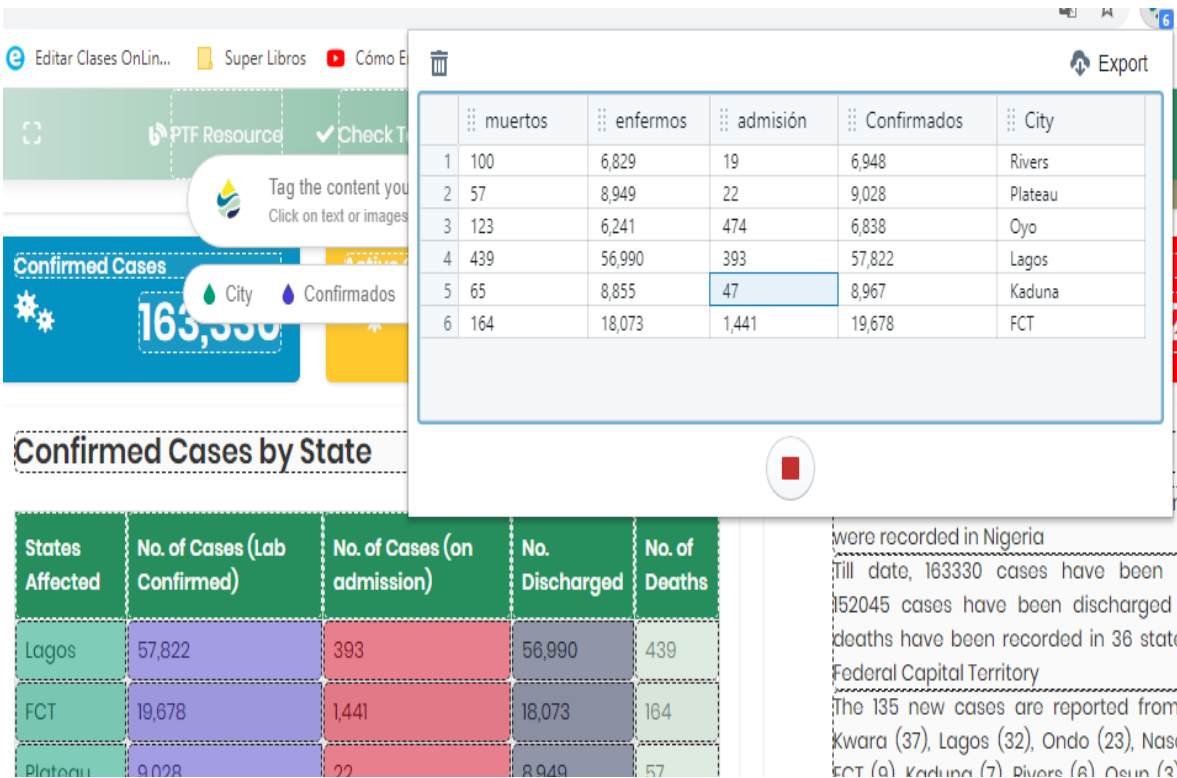


Figura 6

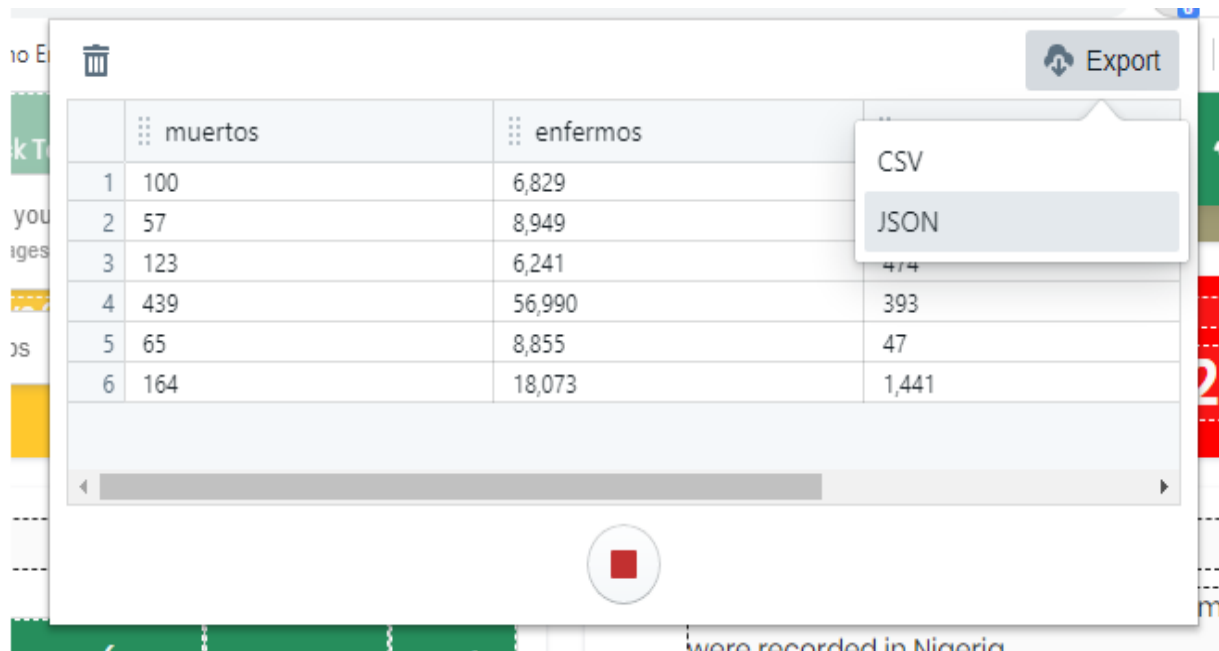


Figure 7

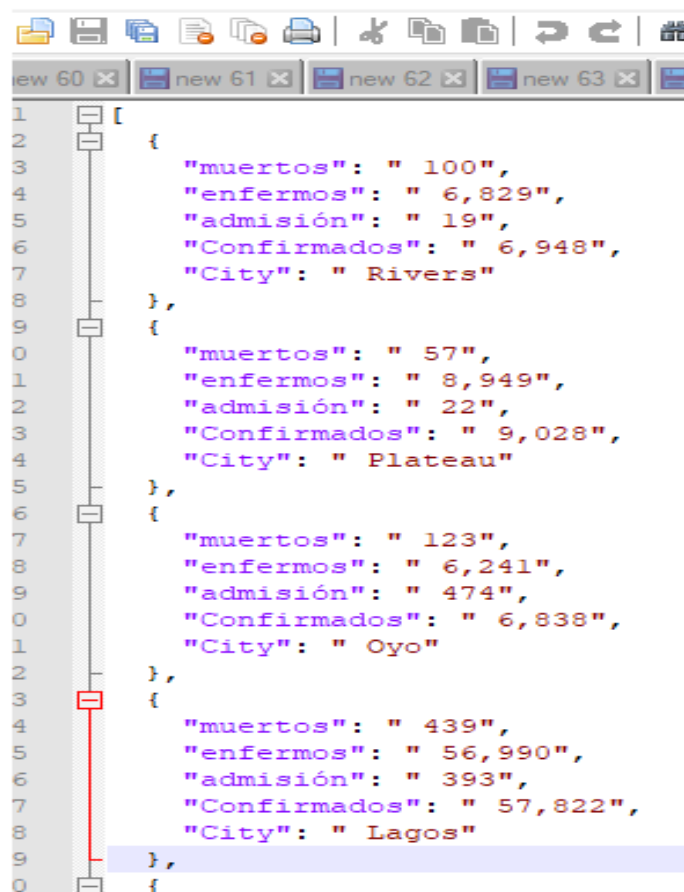
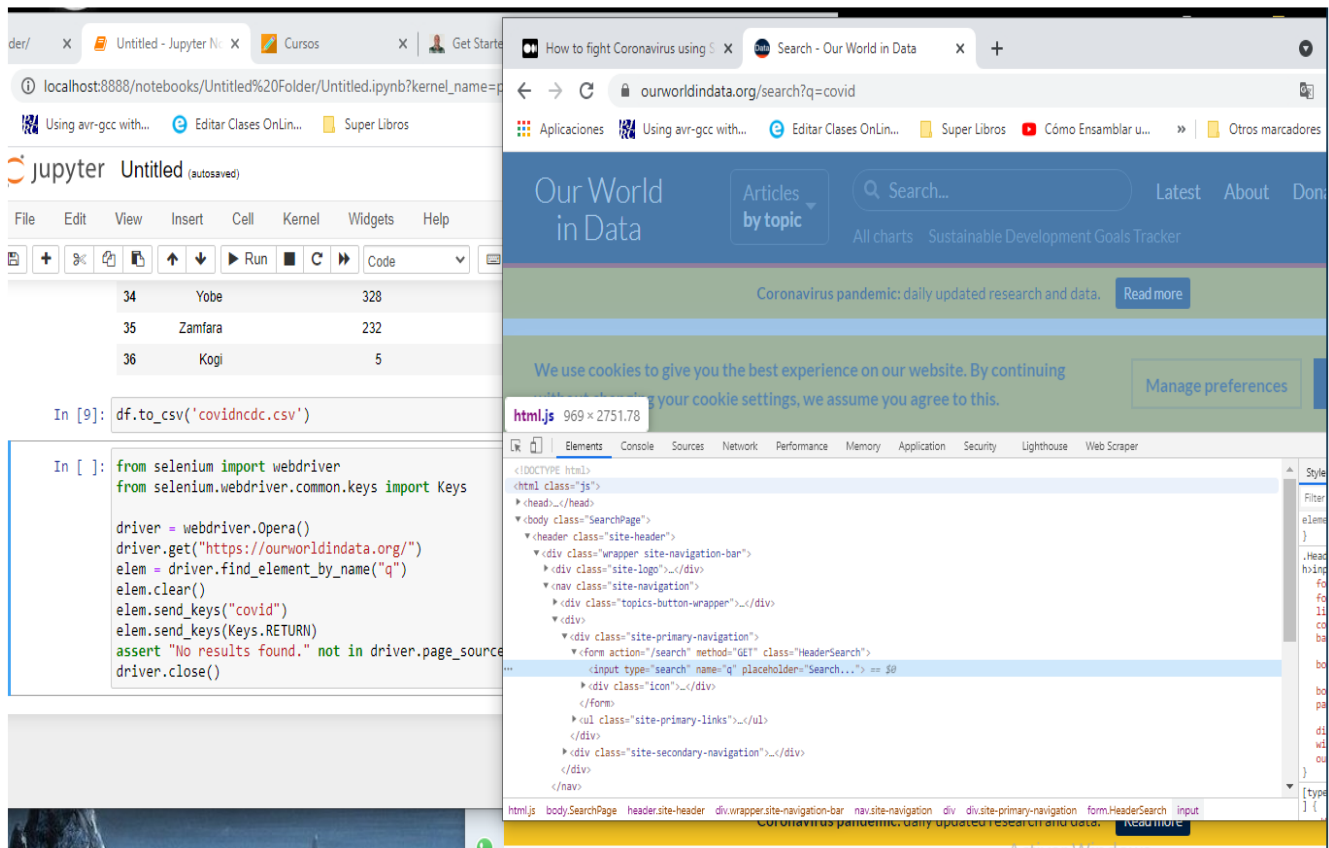


Figura 8

Generación del archivo JSON.

Y por último se utilizó la librería **Selenium** (<https://selenium-python.readthedocs.io>) que permite automatizar mediante códigos muchos procesos de búsqueda e interactuar directamente dentro del navegador que estemos utilizando. Aquí un ejemplo de su uso con navegador **Opera** desde **Anaconda Python**, en **Figura 9**.



Selenium es una librería muy poderosa que permite trabajar con todos los navegadores clásicos para realizar operaciones de *webscraping* y automatizar muchas tareas de búsqueda de información en Internet.

Licencia

He seleccionado para estos datasets la licencia **CC0: Public Domain Licence**, pues en el portal <https://creativecommons.org/share-your-work/public-domain/cc0/> se declara lo siguiente:

CC0 : Sin derechos reservados

CC0 permite a los científicos, educadores, artistas y otros creadores y propietarios de contenido protegido por derechos de autor o base de datos renunciar a esos intereses en sus obras y, por lo tanto, colocarlas de la manera más completa posible en el dominio público, de modo que otros puedan construir libremente, mejorar y reutilizar las obras para cualquier propósito sin restricciones bajo derechos de autor o leyes de bases de datos.

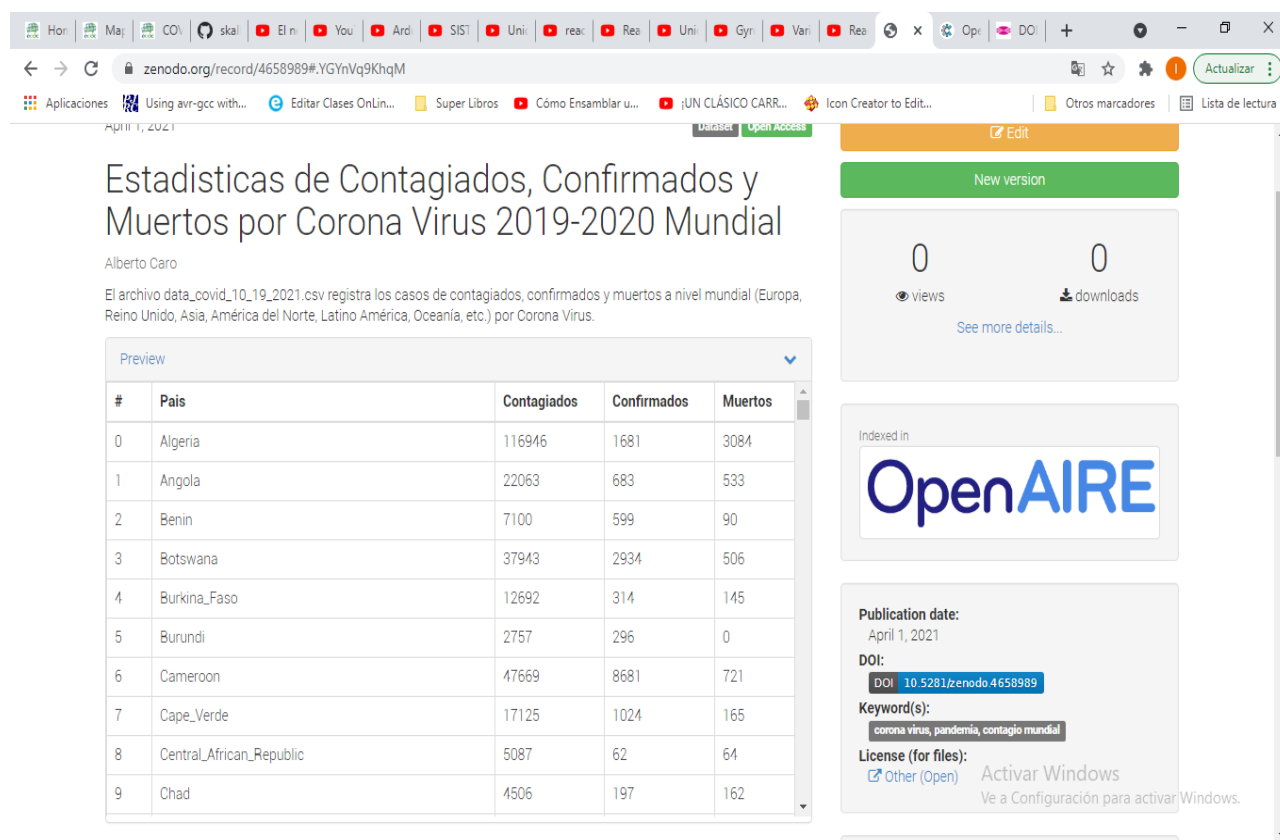
Creo que estos **datasets** cumplen con estas características pues su utilización busca un bien mayor el cual permite a todos los países y centros de salud la utilización de estos datos para sus investigaciones que permitan erradicar este mortal virus.

Todo el material referenciado en este documento y los *Scripts* en **Python** se pueden acceder en:

<https://github.com/skalberto/mitaller>

DataSet y DOI

Aquí se presenta la inscripción del dataset CSV en Zenodo.



The screenshot shows a Zenodo record page for a dataset titled "Estadísticas de Contagiados, Confirmados y Muertos por Corona Virus 2019-2020 Mundial" by Alberto Caro. The page includes a preview of a CSV file with columns for country, infected, confirmed, and dead. The right sidebar shows 0 views and 0 downloads, and the dataset is indexed in OpenAIRE. The DOI is 10.5281/zenodo.4658989.

Estadísticas de Contagiados, Confirmados y Muertos por Corona Virus 2019-2020 Mundial

Alberto Caro

El archivo data_covid_10_19_2021.csv registra los casos de contagiados, confirmados y muertos a nivel mundial (Europa, Reino Unido, Asia, América del Norte, Latino América, Oceanía, etc.) por Corona Virus.

#	Pais	Contagiados	Confirmados	Muertos
0	Algeria	116946	1681	3084
1	Angola	22063	683	533
2	Benin	7100	599	90
3	Botswana	37943	2934	506
4	Burkina_Faso	12692	314	145
5	Burundi	2757	296	0
6	Cameroon	47669	8681	721
7	Cape_Verde	17125	1024	165
8	Central_African_Republic	5087	62	64
9	Chad	4506	197	162

Indexed in

OpenAIRE

Publication date: April 1, 2021

DOI: 10.5281/zenodo.4658989

Keyword(s): corona virus, pandemia, contagio mundial

License (for files): Other (Open)

DOI: 10.5281/zenodo.4658989

Link: <https://zenodo.org/record/4658989#.YGYnVq9KhqM>

Recursos

1. Lawson, R. (2015). Web Scraping with Python. Packt Publishing Ltd. Chapter 2. Scraping the Data
2. Mitchel, R. (2015). Web Scraping with Python: Collecting Data from the Modern Web. O'Reilly Media, Inc. Chapter 1. Your First Web Scrap
3. <https://www.anaconda.com/products/individual>
4. <https://selenium-python.readthedocs.io/>