

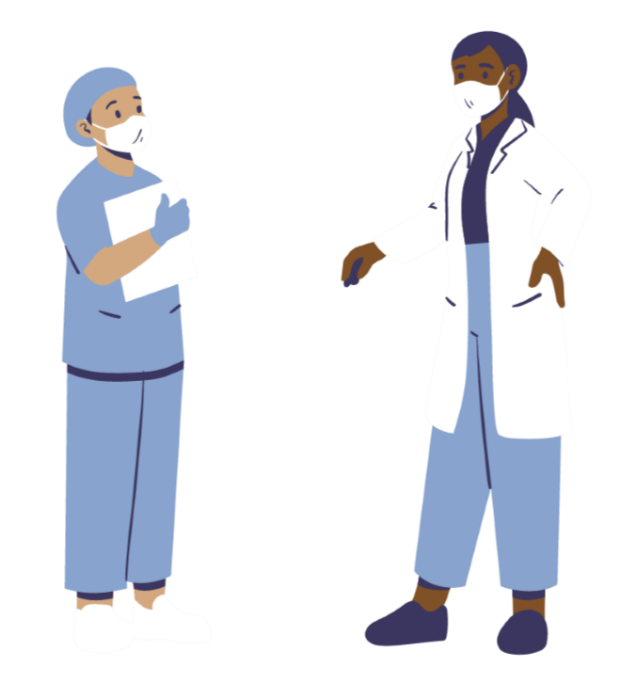
Segunda Entrega

TFM – COVID-19

INESDI

**Programa:** Máster en Business Intelligence &

Data Management (Online)



**Autores:** Grupo 6 COVID - Grupo I (Proyecto Inesdi)                                     **Tutor:** Pier Paolo Rossi

* Amaia Miranda Ulloa
* Fabián Ascheri Aguerre
* José Chavarría Montero
* Juan Carlos Valcuende Aláez
* Patricia Peña Torres

                    10 de julio de 2023

Tabla de Contenidos

[INTRODUCCIÓN 3](#_Toc139775841)

[DESCRIPCIÓN DE FUENTES DE INFORMACIÓN 4](#_Toc139775842)

[Datos de Recuento de casos COVID y Defunciones 4](#_Toc139775843)

[Fuente 1: Casos diarios y muertes por fecha modificada a la OMS 4](#_Toc139775844)

[Fuente 2: Ultimos recuentos notificados de casos y muertes 5](#_Toc139775845)

[Datos de Vacunación 5](#_Toc139775846)

[Fuente 3: Datos de Vacunación 6](#_Toc139775847)

[Fuente 4: Tipos de Vacunas 7](#_Toc139775848)

[Otras Fuentes 8](#_Toc139775849)

[Fuente 5: DATA ON TESTING FOR COVID-19 BY WEEK AND COUNTRY 8](#_Toc139775850)

[Fuente 6: Data on hospital and ICU admission rates and current occupancy for COVID-19 9](#_Toc139775851)

[Fuente 7: Datos sobre casos diarios registrados por estado en USA 11](#_Toc139775852)

[Fuente 8: Paises 11](#_Toc139775853)

[LIMPIEZA, TRANSFORMACIÓN Y ENRIQUECIMIENTO DE DATOS 13](#_Toc139775854)

[Fuente 1: OMS Daily cases and deaths by date reported to WHO 13](#_Toc139775855)

[Fuente 2: OMS Latest reported counts of cases and deaths 21](#_Toc139775856)

[Fuente 3: Datos de Vacunación 25](#_Toc139775857)

[Fuente 4: Tipos de Vacunas 31](#_Toc139775858)

[Fuente 5: Data on testing for COVID-19 by week and country 36](#_Toc139775859)

[Fuente 6: Data on hospital and ICU admission rates and current occupancy for COVID-19 40](#_Toc139775860)

[Fuente 7: Datos sobre casos diarios registrados por estado en USA 45](#_Toc139775861)

[BASE DE DATOS Y PROCEDIMIENTO DE CARGA 56](#_Toc139775862)

[Script para creación de BBDD 56](#_Toc139775863)

[Modelo de datos de la BBDD – TFM COVID-19 59](#_Toc139775864)

[Notebook Jupyter de Conexión y Carga de Ficheros a la BBDD – TFM COVID-19 60](#_Toc139775865)

[CONCLUSIONES 63](#_Toc139775866)

[ANEXOS 64](#_Toc139775867)

# INTRODUCCIÓN

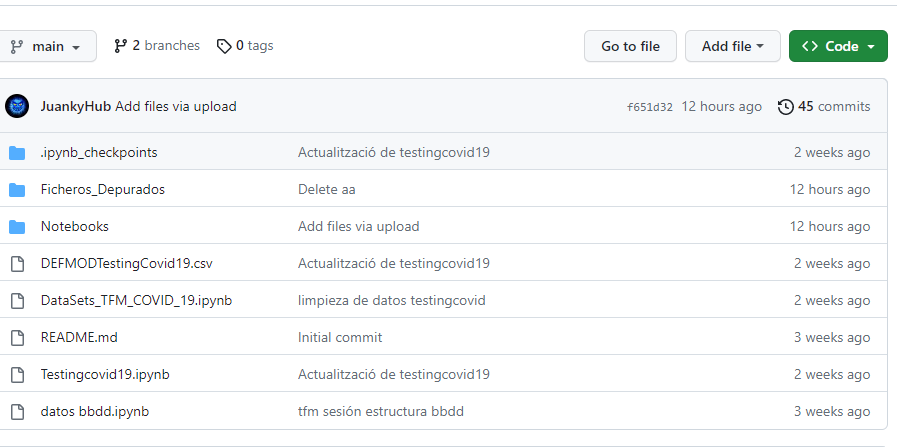
En el primer entregable de este proyecto de TFM se realizó una definición del proyecto, incluyendo una propuesta de valor luego de entender las necesidades de los potenciales usuarios. Estas necesidades se definieron mediante entrevistas, las cuales se resumieron en una sección que denominamos "insights”.

Adicionalmente, en el primer entregable se realizó una exploración y un perfilado inicial de las distintas fuentes de información, lo que permitió definir claramente aquellas que se estarían utilizando para el modelo de datos definitivo.

Este segundo avance que presentamos contiene las siguientes etapas dentro de la construcción de un modelo que permitirá brindar información relevante al potencial usuario. Dicho esto, el entregable se compone de las siguientes tres secciones:

1. Descripción de Fuentes de Información: en esta sección se explica en detalle cada una de las fuentes de información utilizadas, desde el contenido mismo de cada fuente, hasta el tipo de dato que se obtiene de cada columna.
2. Limpieza, Transformación y Enriquecimiento de Datos: en este paso se explica en detalle todo el proceso de depuración de los datos, para lo que se utilizó principalmente la librería Pandas de Python.
3. Base de Datos y Procedimiento de Carga: este capítulo considera la estructura de base de datos de SQL que se propone, así como el script de creación de la misma y el procedimiento de carga de información.

Finalmente, todo nuestro proyecto TFM se encuentra almacenado en el siguiente repositorio público de Github, para fácil acceso de cualquier persona interesada: <https://github.com/patriciaapenat/TFM.git>



# DESCRIPCIÓN DE FUENTES DE INFORMACIÓN

## Datos de Recuento de casos COVID y Defunciones

Consta de 2 ficheros que se cargan semanalmente desde la OMS [**https://covid19.who.int/data**](https://covid19.who.int/data)

Los recuentos de casos nuevos y muertes se calculan restando los recuentos totales acumulativos anteriores del recuento actual. Estos recuentos se actualizan gradualmente a lo largo del día a medida que se dispone de más información. Los recuentos diarios de casos nuevos y muertes se completan a las 23:59CET/CEST de cada día. Debido a las diferencias en los métodos de notificación, las horas límite, la consolidación de datos retrospectivos y los retrasos en la notificación, es posible que la cantidad de casos nuevos no siempre refleje los totales diarios publicados por países, territorios o áreas individuales. Debido a la tendencia reciente de los países que realizan ejercicios de conciliación de datos que eliminan un gran número de casos o muertes de sus recuentos totales, dichos datos pueden reflejarse como números negativos en los recuentos de nuevos casos/nuevas muertes, según corresponda. Esto ayudará a los usuarios a identificar cuándo se producen dichos ajustes.

### Fuente 1: Casos diarios y muertes por fecha modificada a la OMS

<https://covid19.who.int/WHO-COVID-19-global-data.csv>

|  |  |  |
| --- | --- | --- |
| **Nombre del Campo** | **Tipo de Datos** | **Descripción** |
| PAIS | VARCHAR(50) | País, territori, àrea |
| PAIS\_ISO3 | VARCHAR(3) | Código de país ISO Alpha-3 |
| PAIS\_ISO2 | VARCHAR(2) | Código de país ISO Alpha-2 |
| FECHA\_NOTIFICACION | DATE | Fecha de notificación a lOMS |
| OMS\_REGION | VARCHAR(50) | Oficinas regionales de la OMS: Los Estados miembros de la OMS se agrupan en seis regiones de la OMS: Oficina Regional para África (**AFRO**), Oficina Regional para las Américas (**AMRO**), Oficina Regional para el Sur- este Asiático (**SEARO**), Oficina Regional para Europa (**EURO**), Oficina Regional para el Mediterráneo Oriental (**EMRO**) y Oficina Regional para el Pacífico Occidental (**WPRO**). |
| CASOS\_NUEVOS | INTEGER | Nuevos casos confirmados. Se calcula restando el recuentoacumulado anterior del recuentoacumulado de casos actual.\* |
| CASOS\_ACUM | INTEGER | Casos confirmadosacumuladosnotificados en la OMS hastaahora. |
| MUERTES\_NUEVAS | INTEGER | Nuevasmuertesconfirmadas. Se calcula restando las defuncionesacumuladasanteriores de las defuncionesacumuladas actuales.\* |
| MUERTES\_ACUM | INTEGER | Las muertesconfirmadasacumuladas se han notificado a la OMS hastaahora. |

#### \*Los usuarios tienen que tener en cuenta que, además de capturar nuevos casos y muertes notificados un día cualquiera, las actualizaciones se hacen retrospectivamente para corregir los recuentos de los días anteriores según sea necesario en función de la información posterior recibida.

### Fuente 2: Ultimos recuentos notificados de casos y muertes

<https://covid19.who.int/WHO-COVID-19-global-table-data.csv>

|  |  |  |
| --- | --- | --- |
| **Nombre del Campo** | **Tipo de Datos** | **Descripción** |
| PAIS\_ISO\_3 | VARCHAR(3) | Código de país ISO Alpha-3 |
| OMS\_REGION | VARCHAR(50) | Región de l'OMS |
| CASOS\_ACUM\_TOTAL | INTEGER | Casos confirmados acumulados notificados en la OMS hasta ahora. |
| CASOS\_ACUM\_TOTAL\_POR\_100000\_HAB | DOUBLE | Casos confirmados acumulados notificados a la OMS hasta la fecha por cada 100.000 habitantes. |
| CASOS\_NUEVOS\_INFORMADOS\_ULT\_7\_DIAS | INTEGER | Se han notificado nuevos casos confirmados en los últimos 7 días. Se calcula restando el recuentoa cumulado anterior (8 días antes) del recuento acumulado de casos actual. |
| CASOS\_NUEVOS\_INFORMADOS\_ULT\_7\_DIAS\_POR\_100000\_HAB | DOUBLE | Nuevos casos confirmados notificados en los últimos 7 días por cada 100.000 habitantes. |
| CASOS\_NUEVOS\_INFORMADOS\_ULT\_24H | INTEGER | Se han notificado nuevos casos confirmados en las últimas 24 horas. Se calcula restando el recuento acumulado anterior de casos del recuento acumulado actual. |
| MUERTES\_ACUM\_INFORMADAS\_ULT\_7\_DIAS | INTEGER | Las muertes confirmadas acumuladas que se han notificado a la OMS hasta ahora. |
| MUERTES\_ACUM\_TOTAL\_POR\_100000\_HAB | DOUBLE | Muertes confirmadas acumuladas notificadas a la OMS hasta la fecha por cada 100.000 habitantes. |
| MUERTES\_NUEVAS\_INFORMADAS\_ULT\_7\_DIAS | DOUBLE | Se han notificado nuevasmuertes confirmadas en los últimos 7 días. Se calcula restando el recuento acumulado anterior de defunciones (8 días antes) del recuento acumulado actual de defunciones. |
| MUERTES\_NUEVAS\_INFORMADAS\_ULT\_7\_DIAS\_POR\_100000\_HAB | DOUBLE | Nuevas muertes confirmadas notificadas en los últimos 7 días por cada 100.000 habitantes. |
| MUERTES\_NUEVAS\_INFORMADAS\_ULT\_24H | INTEGER | Se han notificado nuevas muertes confirmadas en las últimas 24 horas. Se calcula restando el recuento acumulado anterior de defunciones del recuento acumulado actual de defunciones. |

## Datos de Vacunación

Consta de 2 ficheros que se cargan semanalmente desde la OMS [**https://covid19.who.int/data**](https://covid19.who.int/data)

Un fichero con las actualizaciones semanales sobre la introducción y administración de vacunas por países, territorios y áreas. Estos datos se recopilan de numerosas fuentes, incluidos informes directos de los Estados miembros, la revisión de la OMS de datos oficiales disponibles públicamente o datos recopilados y publicados por sitios de terceros como Our World in Data . Los datos publicados por sitios de terceros no han sido validados por la OMS, y la OMS no puede comentar sobre su precisión o integridad. Se esperan diferencias en los conteos en comparación con otras fuentes debido a los diferentes criterios de inclusión y tiempos de corte de datos.

Las dosis totales administradas, las personas vacunadas con al menos una dosis y las personas vacunadas por completo son totales acumulados desde el inicio de la vacunación en el país respectivo, hasta la última actualización de datos. Las dosis totales administradas se refieren a dosis únicas y pueden no ser iguales al número total de personas vacunadas, según el régimen de dosis específico (las personas reciben dosis múltiples). Las dosis totales administradas por 100 habitantes pueden exceder las 100, por ejemplo, cuando más de la mitad de la población recibe las dos dosis de vacuna requeridas en un régimen de dos dosis. Las tasas <0,001 por 100 habitantes pueden redondearse a 0. Cuando se utilizan múltiples vacunas en un país/territorio/área, la fecha de inicio que se muestra es equivalente a la fecha de inicio de la primera vacuna introducida. No se tienen en cuenta las suspensiones (temporales o no) del despliegue de la vacunación.

Un segundo fichero con la información de los tipos de vacunas utilizadas por los diferentes países. La mención de empresas específicas o de productos vacunales de ciertos fabricantes no implica que la OMS los apruebe o recomiende de preferencia a otros de naturaleza similar que no se mencionan. Salvo excepciones limitadas, los nombres de los productos patentados se distinguen por letras mayúsculas iniciales.

### Fuente 3: Datos de Vacunación

<https://covid19.who.int/who-data/vaccination-data.csv>

|  |  |  |
| --- | --- | --- |
| **Nombre del Campo** | **Tipo de Datos** | **Descripción** |
| PAIS | VARCHAR(50) | País, territorio, área |
| PAIS\_ISO3 | VARCHAR(3) | Código de país ISO Alpha-3 |
| CONTINENTE | VARCHAR(50) | Descripción del Continente |
| ISO\_CONTINENTE | VARCHAR(2) | Código de Continente ISO Alpha-2 |
| OMS\_REGION | VARCHAR(50) | Oficinas regionales de la OMS: Los Estados miembros de la OMS se agrupan en seis regiones de la OMS: Oficina Regional para África (**AFRO**), Oficina Regional para las Américas (**AMRO**), Oficina Regional para el Sur- este Asiático (**SEARO**), Oficina Regional para Europa (**EURO**), Oficina Regional para el Mediterráneo Oriental (**EMRO**) y Oficina Regional para el Pacífico Occidental (**WPRO**). |
| FUENTE\_DATOS | VARCHAR(50) | Indica la fuente de los datos: - REPORTING: Datos reportados por los Estados miembros, o procedentes de informes oficiales - OWID: Datos procedentes de OurWorld in Data: <https://ourworldindata.org/covid-vaccinations> |
| FECHA\_ULT\_ACTUALIZACION | DATE | Fecha de la última actualización |
| TOTAL\_VACUNACIÓN\_ACUM | DOUBLE | Total acumulado de dosis de vacunas administradas |
| NPER\_VACUNADAS\_1DOSIS | INTEGER | Número acumulado de personas vacunadas con al menos una dosis |
| TOTAL\_VACUNACION\_PER100 | DOUBLE | Total acumulado de dosis de vacunas administradas por cada 100 habitantes |
| NPER\_VACUNADAS\_1DOSIS\_PER100 | DOUBLE | Personas acumuladas vacunadas con al menos una dosis por cada 100 habitantes |
| NPER\_VACUNADAS\_DOSIS\_FULL | INTEGER | Número acumulado de personas completamente vacunadas |
| NPER\_VACUNADAS\_DOSIS\_FULL\_PER100 | DOUBLE | Número acumulado de personas completamen tevacunadas por cada 100 habitantes |
| FECHA\_PRIMERA\_VACUNA | DATE | Fecha de las primeras vacunaciones. Equivalente a la fecha  de inicio/lanzamiento de la primera vacuna administrada en un país. |
| N\_TIPOS\_VACUNAS\_USADAS | INTEGER | Número de tipos de vacunas utilizadas por país, territorio, área |
| NPER\_CON\_DOSIS\_ADIDICIONAL | DOUBLE | Las personas recibieron dosis de refuerzo o adicional |
| NPER\_CON\_DOSIS\_ADIDICIONAL\_PER100 | DOUBLE | Las personas recibieron dosis de refuerzo o adicional por cada 100 habitantes |

### Fuente 4: Tipos de Vacunas

<https://covid19.who.int/who-data/vaccination-metadata.csv>

|  |  |  |
| --- | --- | --- |
| **Nombre del Campo** | **Tipo de Datos** | **Descripción** |
| PAIS\_ISO3 | VARCHAR(3) | Código de país ISO Alpha-3 |
| CONTINENTE | VARCHAR(50) | Descripción del Continente |
| ISO\_CONTINENTE | VARCHAR(2) | Código de Continente ISO Alpha-2 |
| NOMBRE\_VACUNA | VARCHAR(100) | Nombre corto combinado de la vacuna: "Empresa - Nombre del producto |
| NOMBRE\_TIPO\_VACUNA | VARCHAR(90) | Nombre o etiqueta del producto de la vacuna, o tipos de vacuna (si no tiene nombre). |
| NOMBRE\_COMPAÑÍA | VARCHAR(90) | Autorización de comercialización del titular del producto vacunal. |
| FECHA\_PRIMERA\_VACUNA | DATE | Fecha de las primeras vacunaciones. Equivalente a la fecha de inicio/lanzamiento de la primera vacuna administrada en un país. |
| FECHA\_INICIO\_VACUNACION | DATE | Fecha de inicio/lanzamiento de la vacunación con tipos de vacuna (excluye las vacunas durante los ensayos clínicos). |
| FECHA\_FIN\_VACUNACION | DATE | Fecha de finalización del despliegue de la vacuna |
| FUENTE\_DATOS | VARCHAR(50) | Indica la fuente de datos - REPORTING: Datosreportados por los Estados miembros, o procedentes de informes oficiales - OWID: Datos procedentes de OurWorld in Data: <https://ourworldindata.org/covid-vaccinations> |

## Otras Fuentes

### Fuente 5: [DATA ON TESTING FOR COVID-19 BY WEEK AND COUNTRY](https://opendata.ecdc.europa.eu/covid19/testing/)

<https://opendata.ecdc.europa.eu/covid19/testing/>

Las cifras que se muestran para la tasa de pruebas semanales por cada 100 000 habitantes y la positividad de las pruebas semanales (%) se basan en varias fuentes de datos.

El número de casos semanales por utilizado para estimar la positividad de la prueba semanal por país o región subnacional se basa en los datos recopilados por ECDC Epidemic Intelligence. Las fuentes de información son Ministerios de Salud o Institutos Nacionales de Salud Pública (sitios web, cuentas oficiales de twitter o cuentas oficiales de Facebook), y los datos obtenidos se cotejan sistemáticamente con datos de OMS. Hay más información disponible en este enlace.

La fuente principal del total de pruebas por país o región subnacional por semana son los datos agregados presentados por los Estados miembros a TESSy. Sin embargo, cuando no estaba disponible, como solía ser el caso antes de la pandemia, el ECDC recopiló datos de fuentes públicas en línea. Estos datos se han recuperado automática o manualmente ("web-scraped") diariamente de fuentes públicas en línea nacionales/oficiales de países de la UE/EEE. Cabe señalar que existen varias limitaciones para este tipo de datos. Los datos raspados no están disponibles para todas las variables y/o países debido a la variabilidad del contenido en los sitios web nacionales.

Además, el proceso de recopilación de datos requiere una adaptación constante para evitar series temporales interrumpidas (es decir, debido a la modificación de las páginas del sitio web, tipos de datos).

La tasa de notificación de 14 días de nuevos casos de COVID-19 se basa en los datos recopilados por ECDC Epidemic Intelligence de varias fuentes y se ve afectada por la estrategia de prueba local, la capacidad del laboratorio y la eficacia de los sistemas de vigilancia. Por lo tanto, la comparación de la situación epidemiológica de la COVID-19 entre países no debe basarse únicamente en estas tasas. Sin embargo, a nivel de país individual o regional, este indicador puede ser útil para monitorear la situación nacional a lo largo del tiempo.

Las políticas de pruebas y el número de pruebas realizadas por cada 100 000 personas varían notablemente a lo largo del UE/EEE y presumiblemente aún más entre terceros países. Las pruebas más exhaustivas conducirán inevitablemente a la detección de más casos.

**Interpretation of COVID-19 data**

La tasa de notificación de 14 días de nuevos casos de COVID-19 debe usarse en combinación con otros factores, incluidas las políticas de prueba, la cantidad de pruebas realizadas, la positividad de la prueba, el exceso de mortalidad y las tasas de ingresos hospitalarios y en la Unidad de Cuidados Intensivos (UCI), al analizar la situación epidemiológica en un país. La mayoría de estos indicadores se presentan para los Estados miembros de la UE/EEE en el informe Panorama general del país.

Incluso cuando se utilizan varios indicadores en combinación, las comparaciones entre países deben hacerse con cautela y experiencia epidemiológica relevante.

|  |  |
| --- | --- |
| **Columna** | **Descripción** |
| ISO 3 | 3-letter ISO country code |
| year\_week | yyyy-Www |
| level | National (archived dataset with national subnational data to week 36, 2022 is available on ECDC’s website) |
| new\_cases | Number of new confirmed cases |
| tests\_done | Number of tests done |
| population | Numeric |
| testing\_rate | Testing rate per 100,000 population |
| positivity\_rate | Weekly test positivity (%): 100 x Number of new confirmed cases/number of tests done per week |
| testing\_data\_source | - Country API<br>- Country GitHub<br>- Country website<br>- Manual webscraping<br>- Other<br>- Survey<br>- TESSy: data provided directly by Member States to ECDC via TESSy |

### Fuente 6: Data on hospital and ICU admission rates and current occupancy for COVID-19

<https://www.ecdc.europa.eu/en/publications-data/download-data-hospital-and-icu-admission-rates-and-current-occupancy-covid-19>

Datos sobre las tasas de admisión hospitalaria y de UCI y la ocupación actual por COVID-19.

|  |  |  |
| --- | --- | --- |
| ***Nombre del Campo*** | ***Tipo de Datos*** | ***Descripción*** |
| COUNTRY | String | Pais 3-letter ISO country code |
| INDICATOR | String |  Ocupación hospitalaria diaria (número  de pacientes con COVID-19 en el hospital  en un día determinado)   Ocupación diaria de la UCI (número de  Pacientes con COVID-19 en UCI en un  dado día)   Nuevos ingresos hospitalarios semanales  por 100k (tarifa semanal de nuevo  admisiones de pacientes con COVID-19  por 100 000 habitantes)   Nuevos ingresos semanales en UCI  de pacientes con COVID-19 por cada 100k  Cuerda  (tasa semanal de nuevas admisiones  por 100 000 habitantes) |
| DATE | YYYY -MM -DD | Fecha para los indicadores de ocupación diaria |
| YEAR\_WEEK | YYYY -Www | Fecha |
| VALUE | Numeric | Número de pacientes o nuevos ingresos  por 100 000 habitantes |
| SOURCE | String | Fuente categórica de los datos:   TESSy: datos proporcionados directamente por  Los Estados miembros al ECDC a través de TESSy   Country\_API  Country\_Github  Country\_Website  External\_Github  CCI  Vigilancia  Other\_Websit |

**Descripción y descargo de responsabilidad:**

Los archivos de datos descargables contienen información sobre hospitalización y Unidad de Cuidados Intensivos (UCI) tasas de admisión y ocupación actual por COVID-19 por fecha y país. Cada fila contiene el datos correspondientes para una fecha determinada (día o semana) y por país. El archivo se actualiza semanalmente. Tú puede utilizar los datos de acuerdo con la política de derechos de autor del ECDC.

**Fuente**

Las cifras mostradas sobre las tasas de hospitalización y admisión en UCI y la ocupación actual son basado en varias fuentes de datos. La fuente principal son los datos basados en casos presentados por los Estados miembros. Sin embargo, cuando no está disponible, y especialmente para la ocupación actual, el ECDC recopila datos de Fuentes públicas en línea.

Los datos que se muestran se han recuperado automática o manualmente ("web-scraped") diariamente de Fuentes en línea públicas nacionales/oficiales de países de la UE/EEE. Cabe señalar que hay varias limitaciones a este tipo de datos. Los datos raspados no están disponibles para todas las variables y/o países debido a la variabilidad del contenido en los sitios web nacionales. Además, el proceso de recopilación de datos requiere una adaptación constante para evitar series temporales interrumpidas (es decir, debido a la modificación del sitio web) páginas, tipos de datos). Los criterios de admisión en hospitales y UCI, y las políticas para informar estos datos difiere entre países y a lo largo del tiempo, lo que puede resultar en estimaciones sesgadas derivadas de tales datos.

**Interpretación de los datos de COVID-19**

La tasa de notificación de 14 días de los nuevos casos de COVID-19 se basa en los datos recopilados por el ECDC Epidemic Intelligence de varias fuentes y se ven afectados por la estrategia de prueba local, la capacidad de los laboratorios y la eficacia de los sistemas de vigilancia. Comparando la epidemiología

Por lo tanto, la situación con respecto a COVID-19 entre países no debe basarse en estas tasas solo. Sin embargo, a nivel de país individual, este indicador puede ser útil para monitorear la situación nacional a lo largo del tiempo.

Las políticas de pruebas y el número de pruebas realizadas por cada 100 000 personas varían notablemente a lo largo del UE/EEE y presumiblemente aún más entre terceros países. Pruebas más extensas inevitablemente llevar a que se detecten más casos.

La tasa de notificación de 14 días de nuevos casos de COVID-19 debe usarse en combinación con otros factores que incluyen políticas de prueba, número de pruebas realizadas, positividad de la prueba, exceso de mortalidad y tasas de ingresos hospitalarios y en UCI, al analizar la situación epidemiológica de un país.

La mayoría de estos indicadores se presentan para los Estados miembros de la UE/EEE en el informe Panorama general del país. Incluso cuando se usan varios indicadores en combinación, se deben hacer comparaciones entre países con precaución y experiencia epidemiológica relevante.

Fuente: [2021-01-13\_Variable\_Dictionary\_and\_Disclaimer\_hosp\_icu\_all\_data.pdf (europa.eu)](https://www.ecdc.europa.eu/sites/default/files/documents/2021-01-13_Variable_Dictionary_and_Disclaimer_hosp_icu_all_data.pdf)

### Fuente 7: Datos sobre casos diarios registrados por estado en USA

<https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_daily_reports_us>

|  |  |  |
| --- | --- | --- |
| ***Nombre del Campo*** | ***Tipo de Datos*** | ***Descripción*** |
| PROVINCE\_STATE | String | Nombre del estado en USA |
| CONFIRMED | Integer | Casos confirmados acumulados por estado |
| DEATH | Integer | Numero de individuos muertos acumulados por estado |
| RECOVERED | Float | Número de individuos recuperados acumulados por estado |
| ACTIVE | Float | Acumulado de casos confirmados que no an sido resueltos (Casos activos = número de casos totales-total de individuos recuperados -total de individuos muertos) |
| ISO3 | String | Código oficial de identificador asignado a cada país |
| DATE | DATETIME | Fecha en que se realizo el nuevo registro |

**Descripción:**

La fuente de información fue los archivos descargables que contiene datos de los casos acumulados registrados diariamente relativo a COVID-19 en cada estado referente a casos confirmados, individuos fallecidos, recuperados y casos activos

**Fuente**

La fuente de datos fue el repositorio de datos sobre COVID-19 del Centro para Sistemas de Ciencia e Ingeniería de la Universidad John Hopkings en GitHub. Los datos en este repositorio fueron recolectados por el Centro de Recursos de Coronavirus de dicha Universidad

### Fuente 8: Paises

<https://gist.github.com/wipodev/9596693c07e1152dae03f2e3e294c493>

|  |  |  |
| --- | --- | --- |
| **Nombre del Campo** | **Tipo de Datos** | **Descripción** |
| PAIS\_ISO3 | VARCHAR(3) | Código de país ISO Alpha-3 |
| PAIS\_ISO2 | VARCHAR(2) | Código de país ISO Alpha-2 |
| PAIS\_NOM | VARCHAR(100) | Decriptivo del Nombre del País |
| COD\_CONTINENTE | VARCHAR(2) | Código de Continente |
| CONTINENTE | VARCHAR(100) | Desriptivo del Nombre del Continente |
| OMS\_REGION | VARCHAR(5) | Oficinas regionales de la OMS: Los Estados miembros de la OMS se agrupan en seis regiones de la OMS: Oficina Regional para África (**AFRO**), Oficina Regional para las Américas (**AMRO**), Oficina Regional para el Sur- este Asiático (**SEARO**), Oficina Regional para Europa (**EURO**), Oficina Regional para el Mediterráneo Oriental (**EMRO**) y Oficina Regional para el Pacífico Occidental (**WPRO**). |
| DESC\_OMS\_REGION | VARCHAR(100) | Descriptivo de las Oficinas regionales de la OMS |
| PAIS\_NOM\_2 | VARCHAR(100) | Decriptivo del Nombre del País con alguna variedad en la descrpición del Nombre del Pais. |

Relación de Paises del Mundo identificados por los códigos ISO2 e ISO3 definidos por la Organización Internacional de Normalización (ISO), así como el continente al que pertenecen y la Oficina Regional de la OMS a la que son miembros.

# LIMPIEZA, TRANSFORMACIÓN Y ENRIQUECIMIENTO DE DATOS

## Fuente 1: OMS Daily cases and deaths by date reported to WHO

El fichero se puede encontrar en el siguiente enlace: <https://github.com/patriciaapenat/TFM/blob/main/Notebooks/Data_Frame_covid_daily_oms.ipynb>

**Ficheros de la organización Mundial de la Salud**

* <https://covid19.who.int/data>
* <https://ourworldindata.org/coronavirus#coronavirus-country-profiles>

**Descarga de datos**

Casos diarios y muertes por fecha notificados a la OMS: <https://covid19.who.int/WHO-COVID-19-global-data.csv>

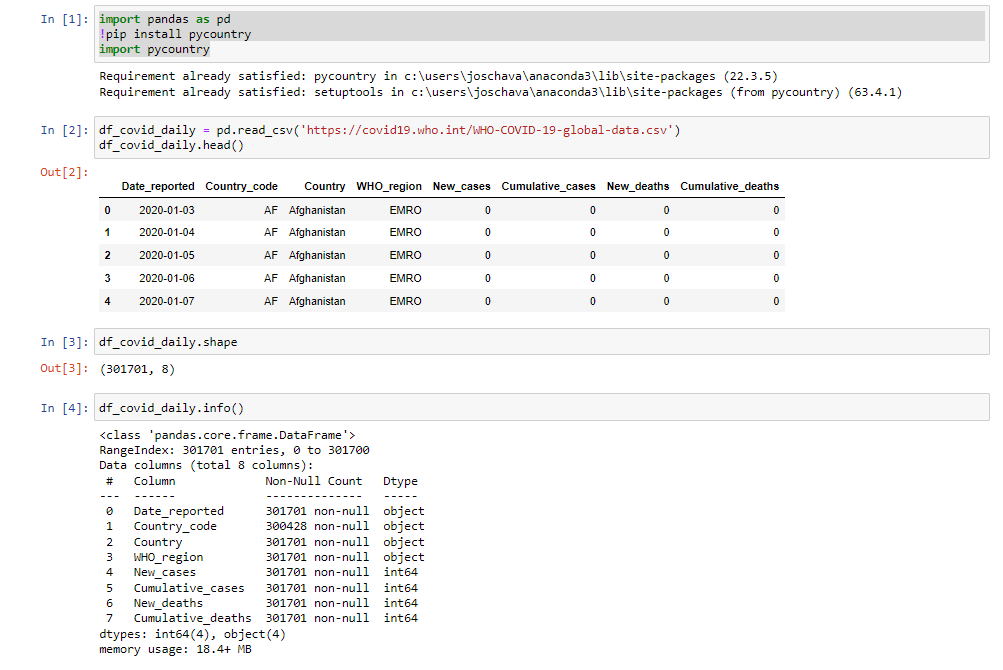
**Información del Dataset**

Los usuarios deben tener en cuenta que, además de capturar nuevos casos y muertes notificadas en un día determinado, las actualizaciones se realizan retrospectivamente para corregir los recuentos de días anteriores según sea necesario en función de la información recibida posteriormente. Consulte "Datos agregados diarios de recuento de casos y muertes" más arriba para obtener más detalles sobre el cálculo de nuevos casos/muertes.

**Exploración**

Exloración de los casos diarios y muertes por fecha notificados a la OMS.

Al dataframe lo llamaremos "df\_covid\_daily".



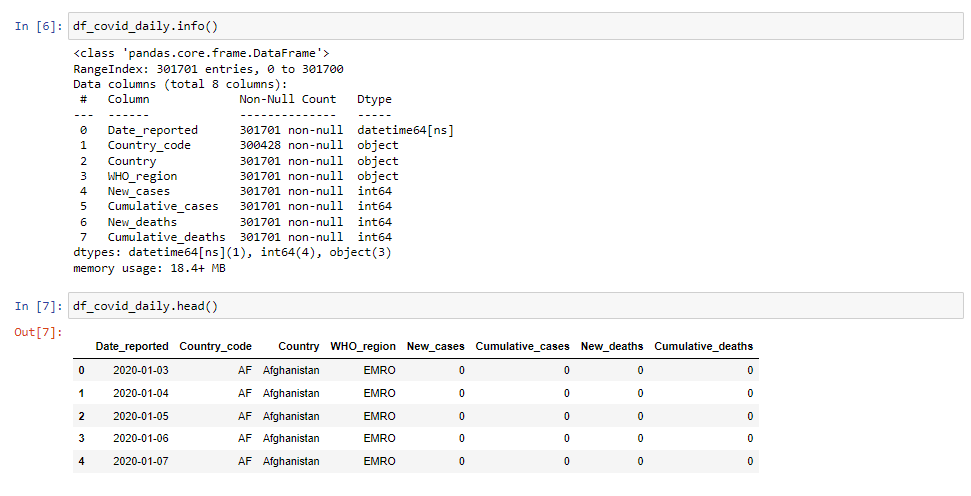
**Transformación**

Transformamos el campo de fecha (DATE\_REPORTED) que aparecen con tipo de datos cadena de caracteres



**Transformación**

Verificamos que se ha modificado el tipo de dato a fecha

****

**Exploración**

Observamos si tenemos valores duplicados

****

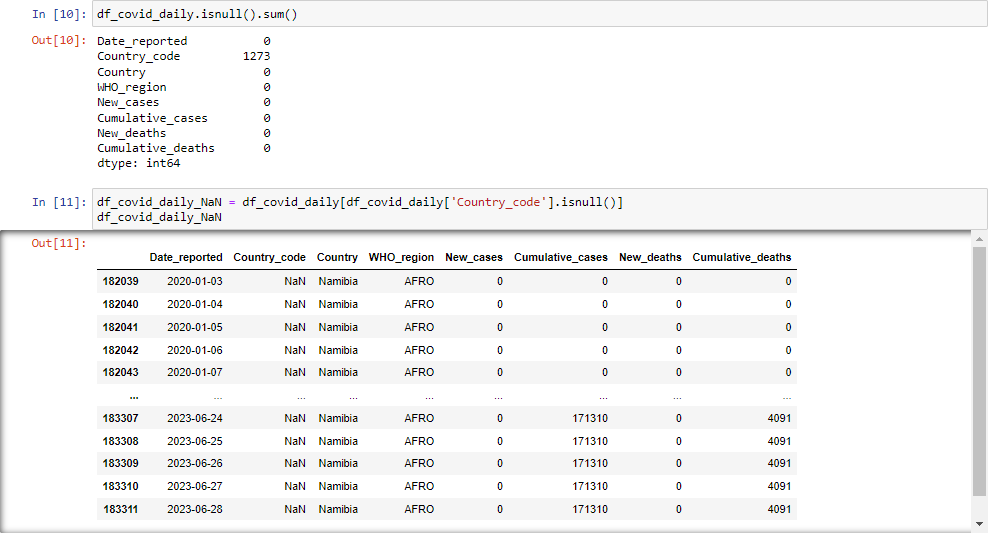
**Exploración**

Observamos si tenemos valores nulos

****

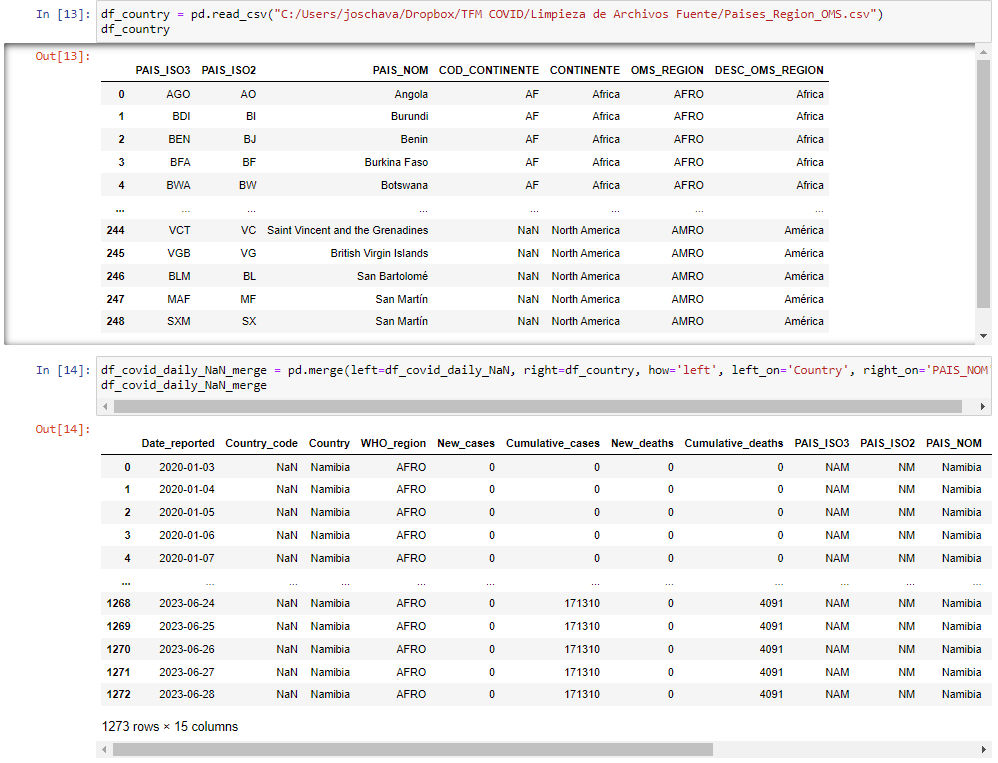
**Exploración**

Identificamos dónde están los valores nulos

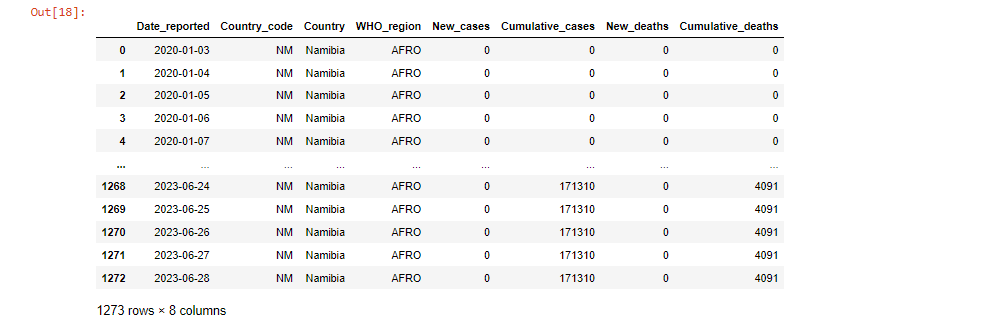
****

**Transformación**

Sustituimos los códigos de país nulos por su respectivo codigo ISO.

****

****

****

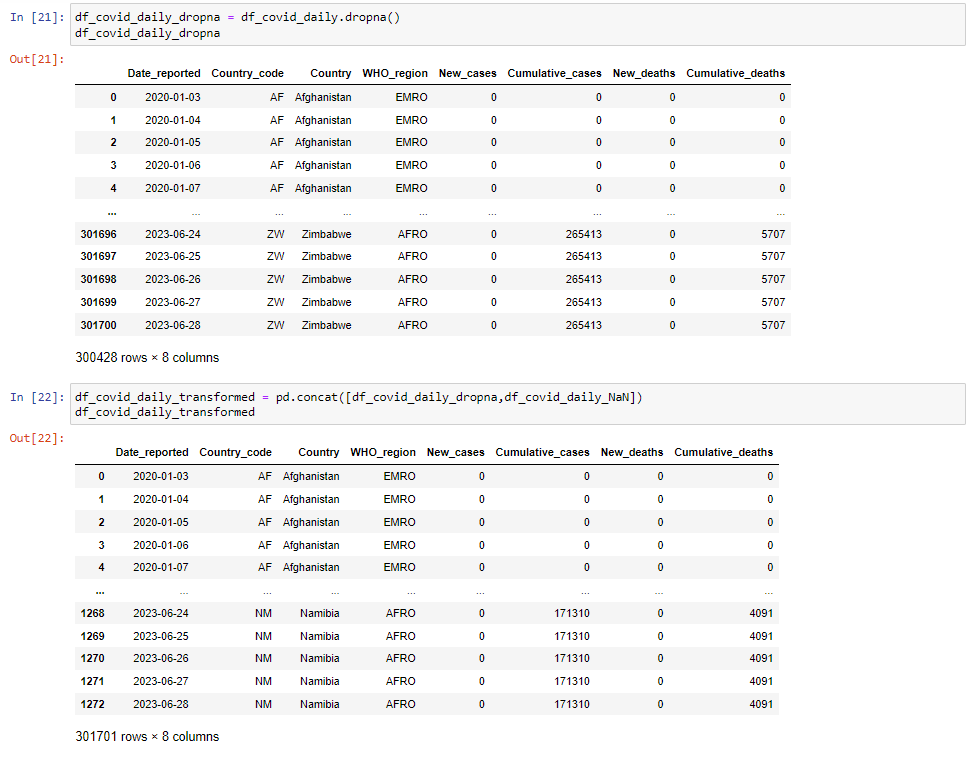
**Transformación**

Validamos que no hay nulos en el df.

****

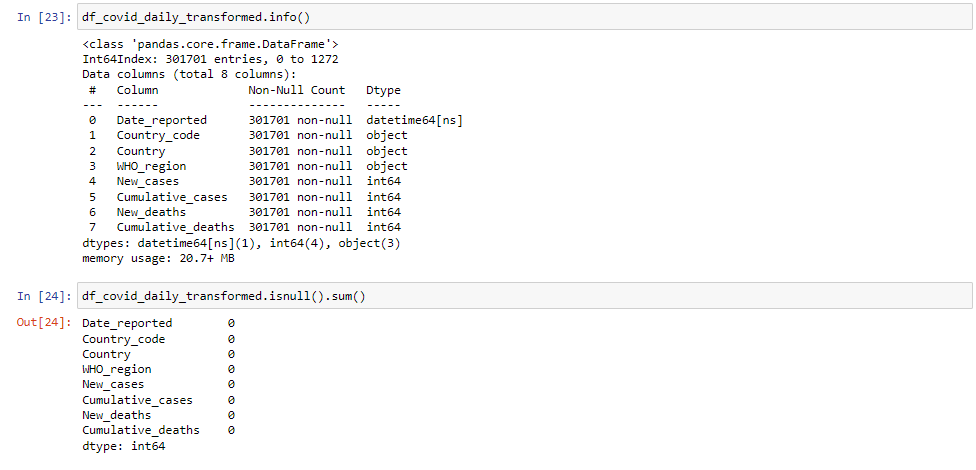
**Transformación**

Eliminamos registros con nulos en el df original (df\_covid\_daily).

****

**Transformación**

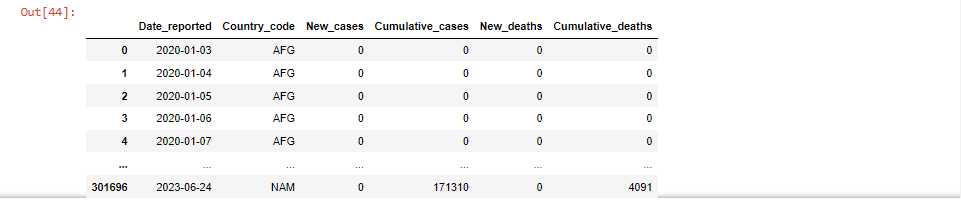
Validamos misma cantidad de registros que df inicial y comprobamos que no haya nulos.

****

**Transformación**

Nuestra base de datos incluirá códigos ISO3 para los países, por lo que debemos llamar la tabla de dimension "País" y sustituir los códigos ISO2 por sus respectivos ISO3.

****

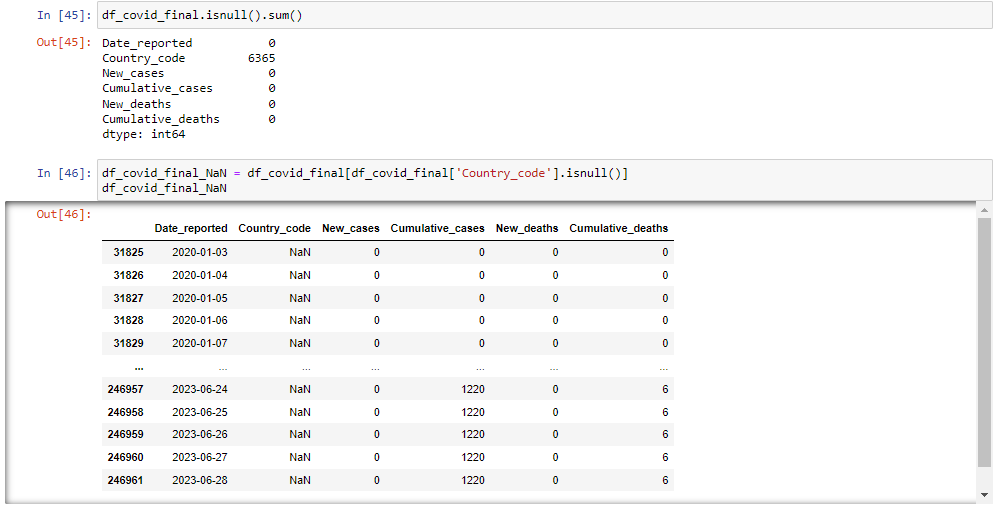
****

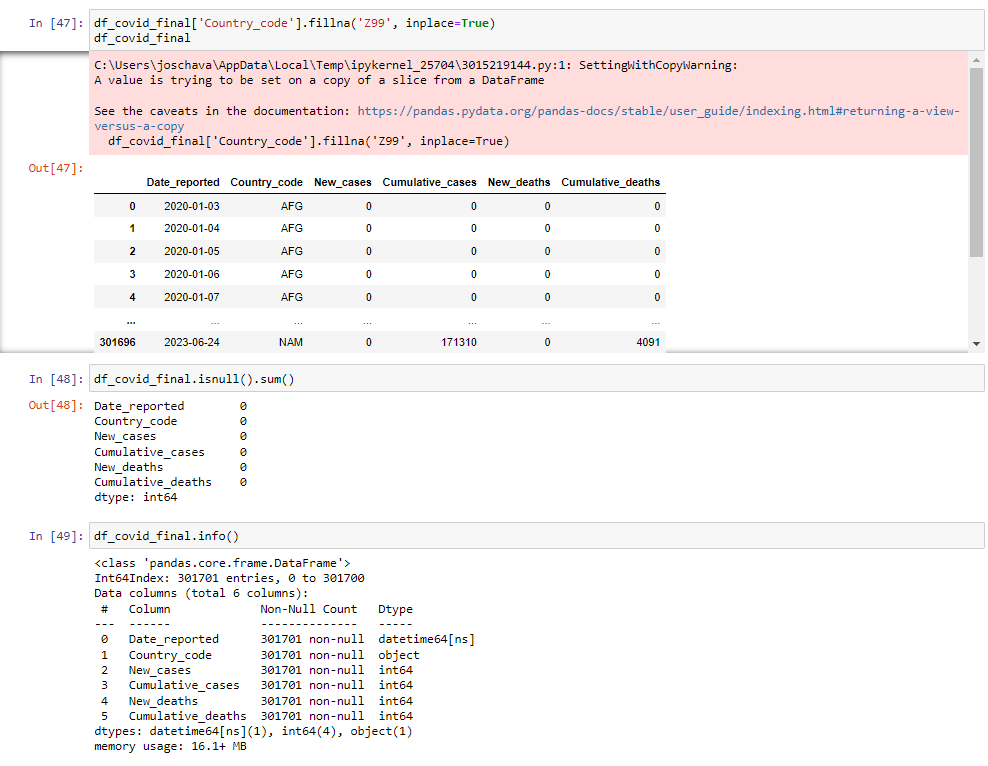
**Transformación**

Identificamos valores nulos relacionados a códigos ISO2 especiales que no aparecen en la tabla de parámetros de país.

<https://es.wikipedia.org/wiki/ISO_3166-1_alfa-2>

* El código XA está siendo utilizado por Suiza, como código de país por las Islas Canarias, a pesar de que IC ya está reservado con dicho propósito.23
* El código XI está siendo utilizado por el Gobierno del Reino Unido, como el prefijo de código de país del número EORI de Irlanda del Norte.24
* El código XK está siendo utilizado por la Comisión Europea,25​ el FMI, la SWIFT,26​ el CLDR y otras organizaciones como código de país provisional para Kosovo.27
* El código XN está siendo utilizado por la Organización Mundial de la Propiedad Intelectual como indicador para el Instituto Nórdico de Patentes, una organización internacional a la que pertenecen Dinamarca, Islandia, Noruega y Suecia.28
* El código XU está siendo utilizado por la Organización Mundial de la Propiedad Intelectual como indicador para la Unión Internacional para la Protección de las Obtenciones Vegetales
* El código XV está siendo utilizado por la Organización Mundial de la Propiedad Intelectual como indicador para el Instituto Visegrad de Patentes
* El código XX está siendo utilizado por la Organización Mundial de la Propiedad Intelectual como un indicador para estados desconocidos, otras entidades u organizaciones

****

****

**Exportamos**

Por último, procedemos con la descarga de los datos.

****

## Fuente 2: OMS Latest reported counts of cases and deaths

El fichero se puede encontrar en el siguiente enlace: <https://github.com/patriciaapenat/TFM/blob/main/Notebooks/Data_Frame_Latest%20reported%20counts%20of%20cases%20and%20deaths.ipynb>

**Ficheros de la organización Mundial de la Salud**

* <https://covid19.who.int/data>
* <https://ourworldindata.org/coronavirus#coronavirus-country-profiles>

**Descarga de datos**

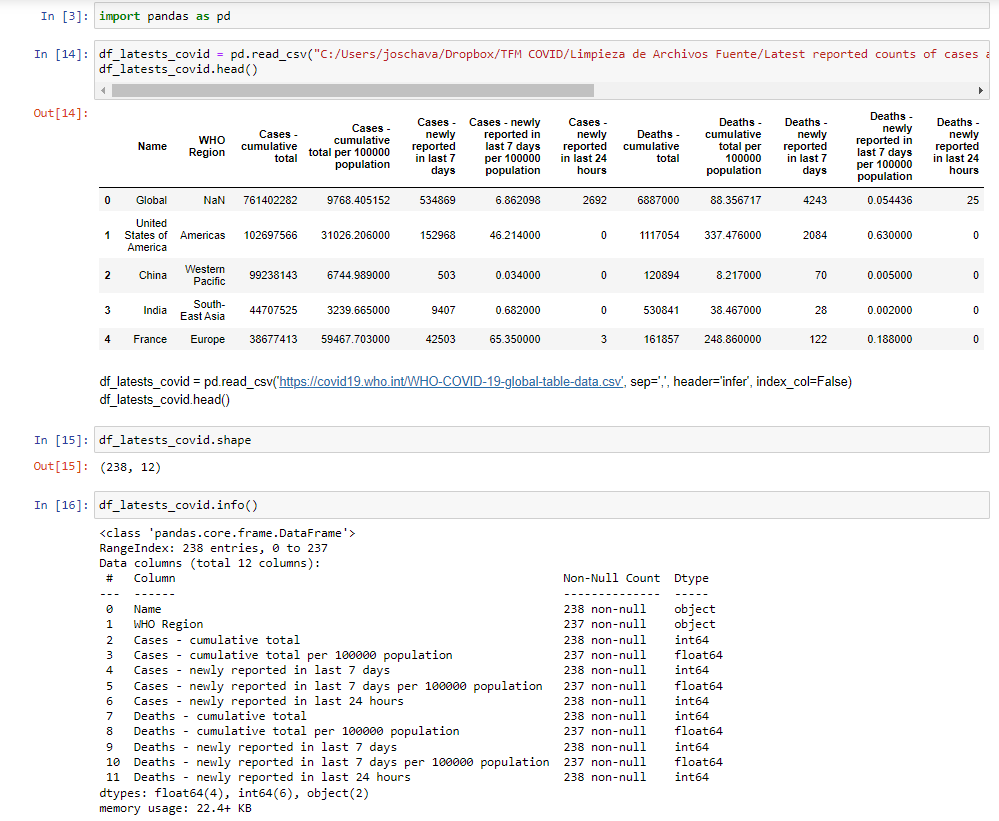
Último reporte de casos y muertes notificados a la OMS: <https://covid19.who.int/WHO-COVID-19-global-table-data.csv>

**Información del Dataset**

Los usuarios deben tener en cuenta que, además de capturar nuevos casos y muertes notificadas en un día determinado, las actualizaciones se realizan retrospectivamente para corregir los recuentos de días anteriores según sea necesario en función de la información recibida posteriormente. Consulte "Datos agregados diarios de recuento de casos y muertes" más arriba para obtener más detalles sobre el cálculo de nuevos casos/muertes.

**Exploración**

Exploración del último reporte de casos y muertes notificados a la OMS. Al dataframe lo llamaremos "df\_latest\_covid".



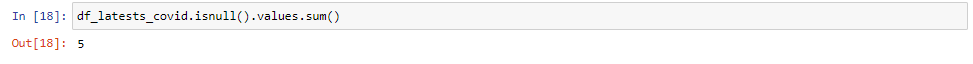
**Exploración**

Observamos si tenemos valores duplicados



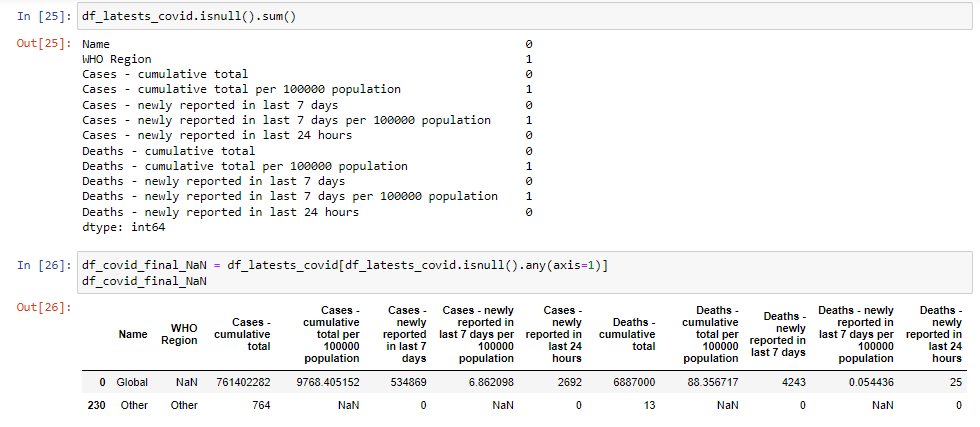
**Exploración**

Observamos si tenemos valores nulos



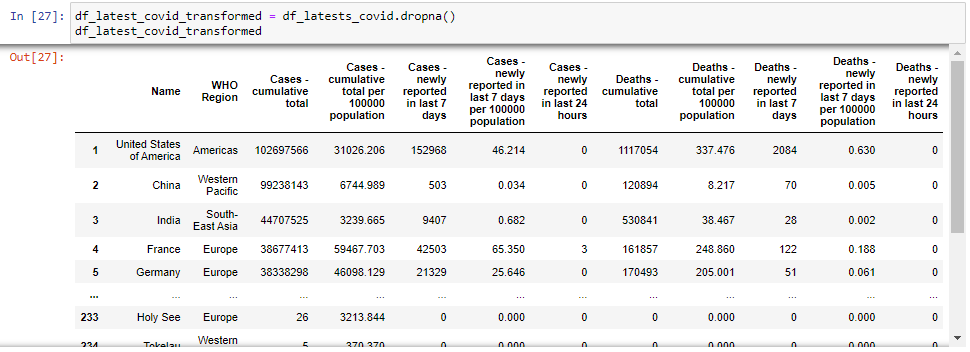
**Exploración**

Identificamos dónde están los valores nulos



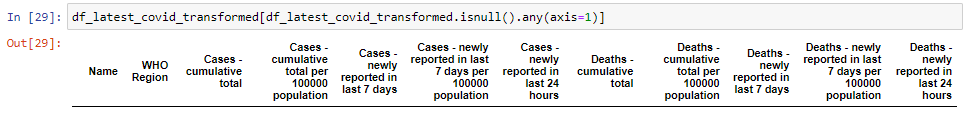
**Transformación**

Ambas líneas podemos eliminarlas de nuestro DF por ser "total" y "otros" (no podremos analizar o extraer nada de esta información).



**Transformación**

Validamos



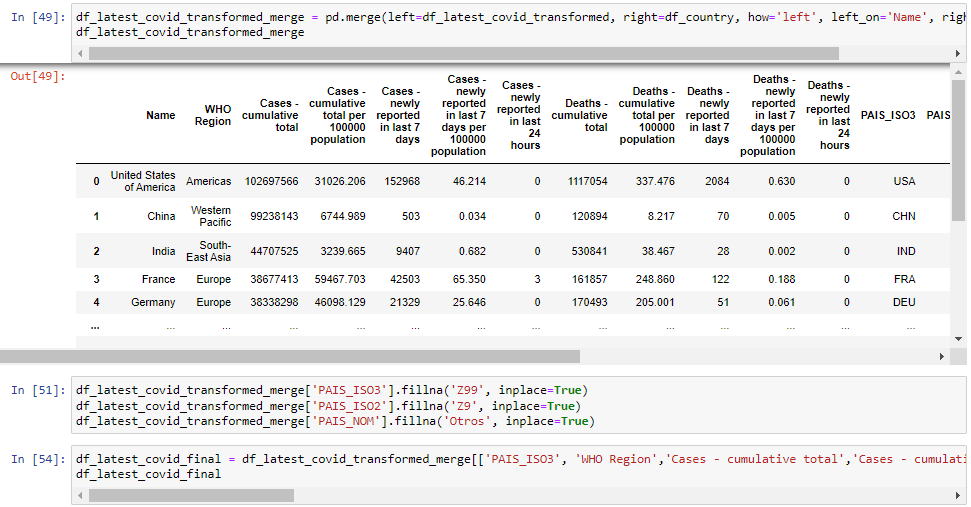
**Transformación**

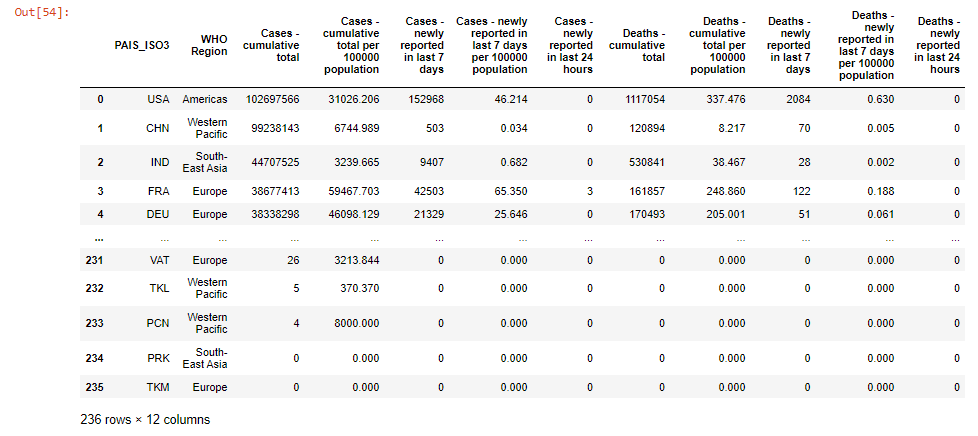
Llamamos tabla de países para obtener ISO3



**Transformación**

Cruzamos las tablas y luego limpiamos los registros





**Exportamos**

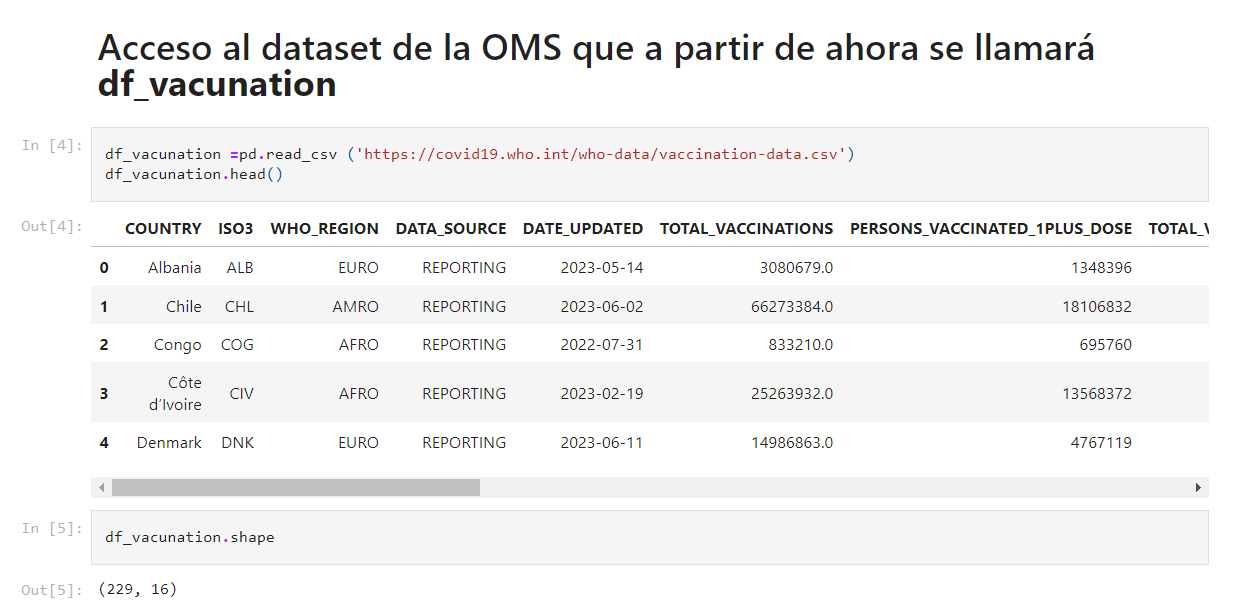
Por último, procedemos con la descarga de los datos.



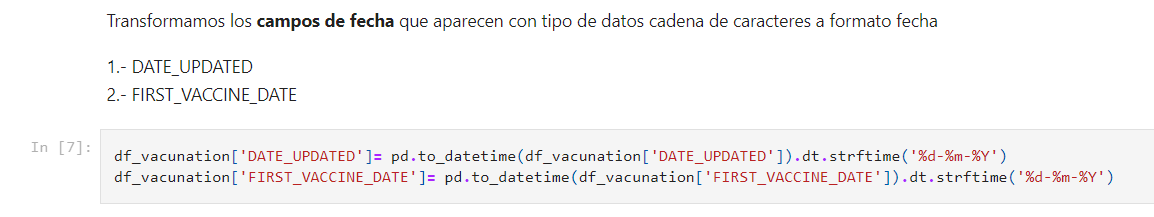
## Fuente 3: Datos de Vacunación

El fichero se puede encontrar en el siguiente enlace: <https://github.com/patriciaapenat/TFM/blob/main/Notebooks/Data_Frame_vacunaciones_oms.ipynb>

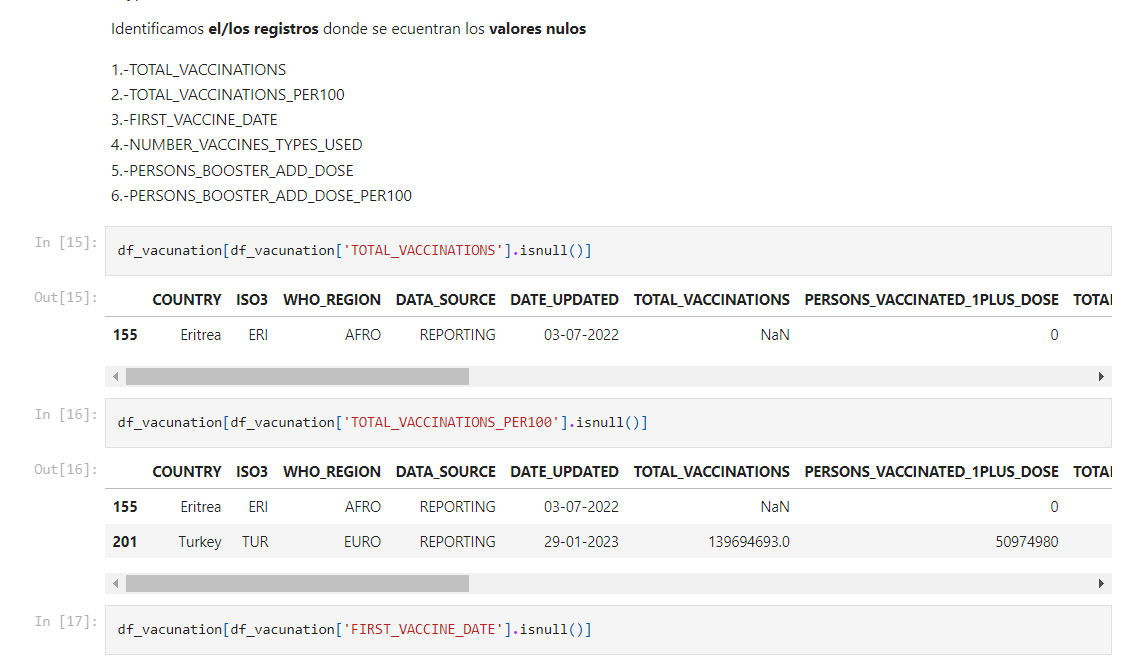


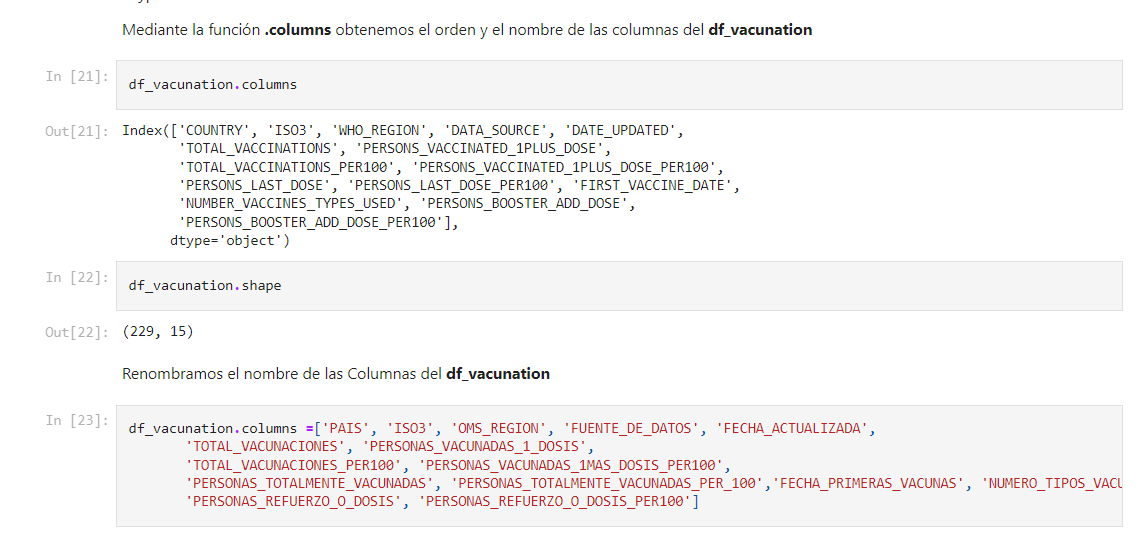
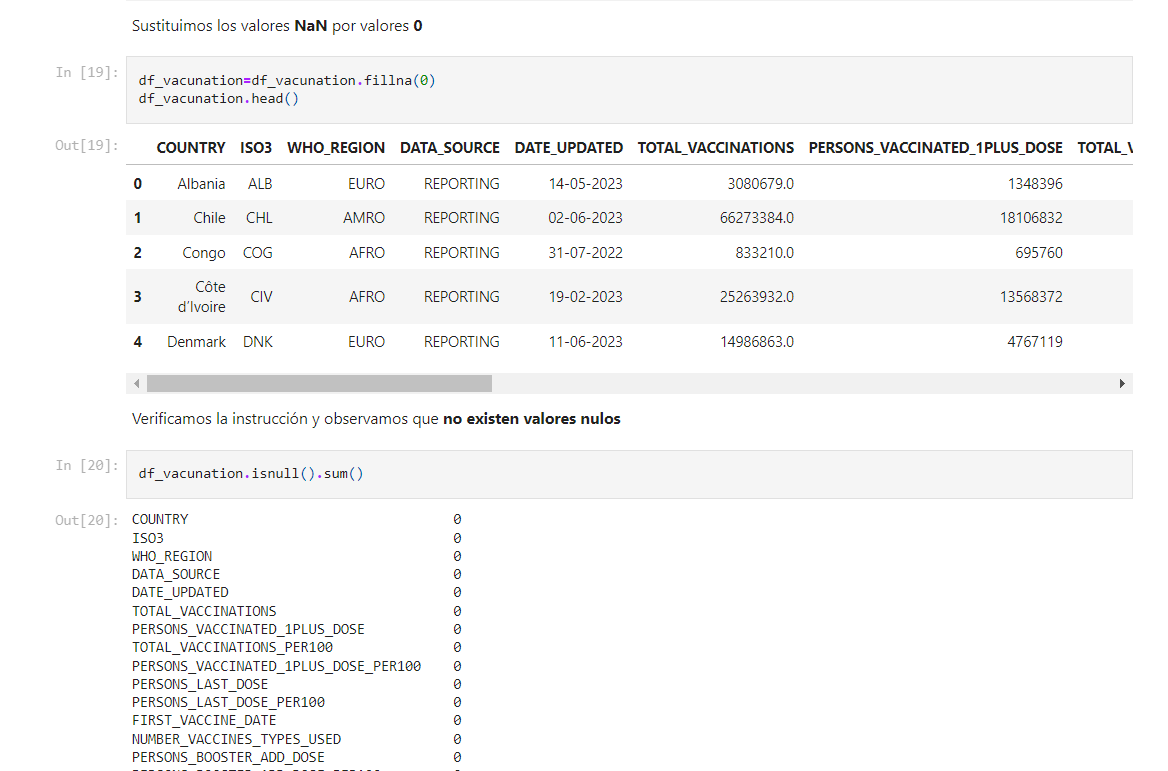




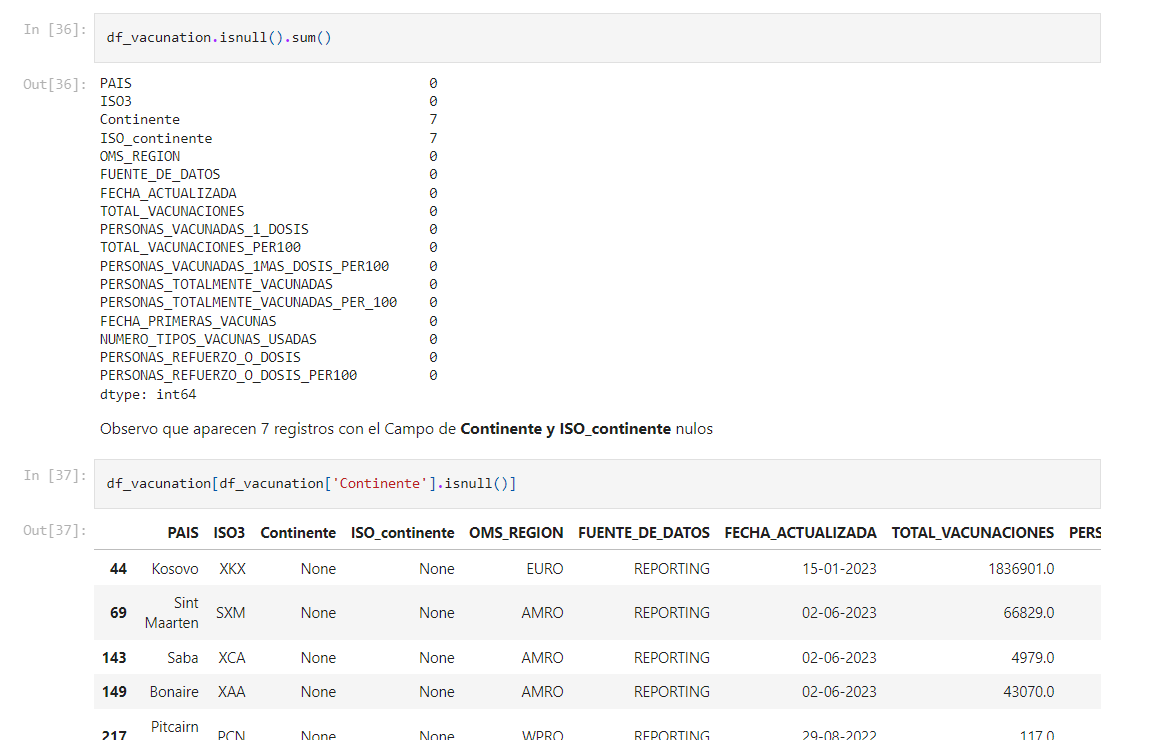


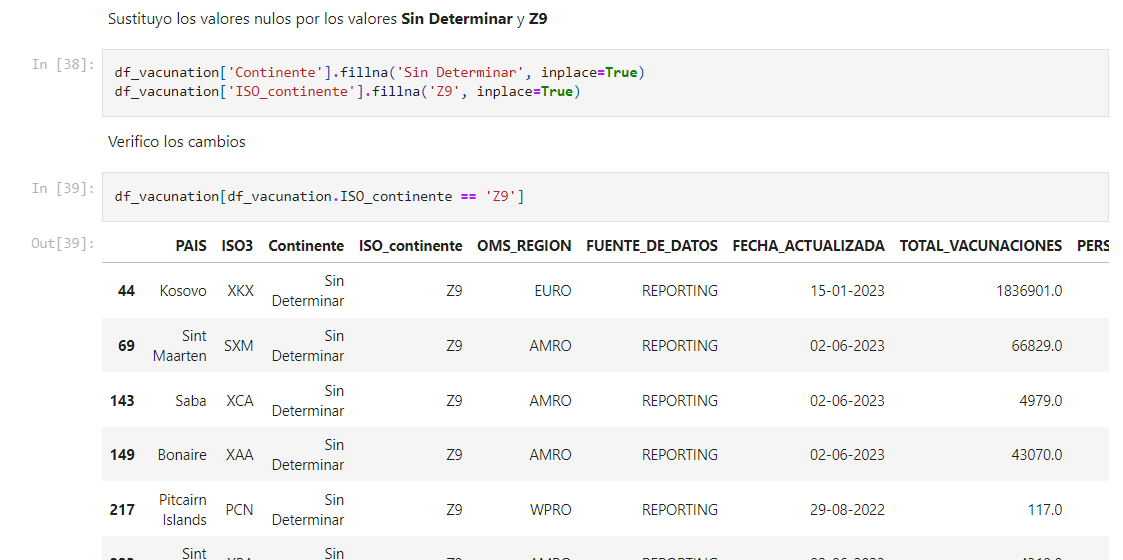


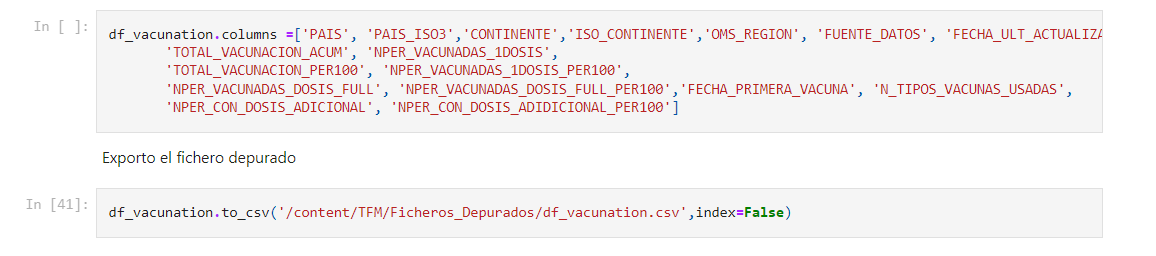






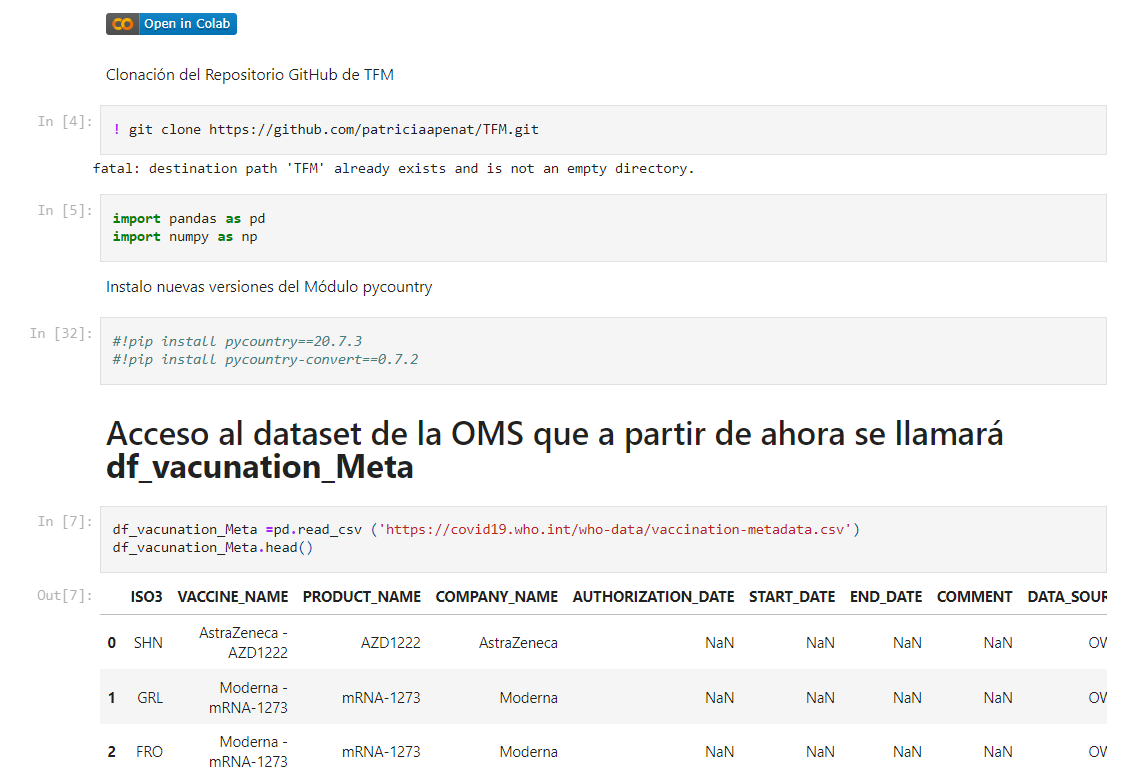


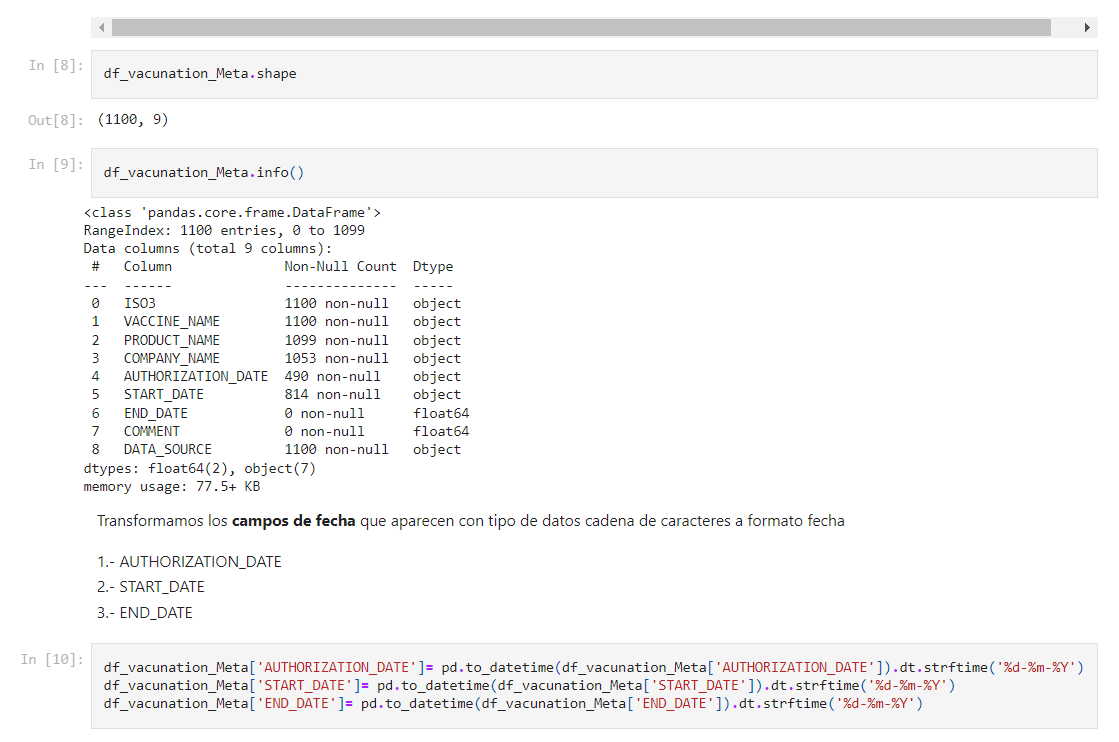




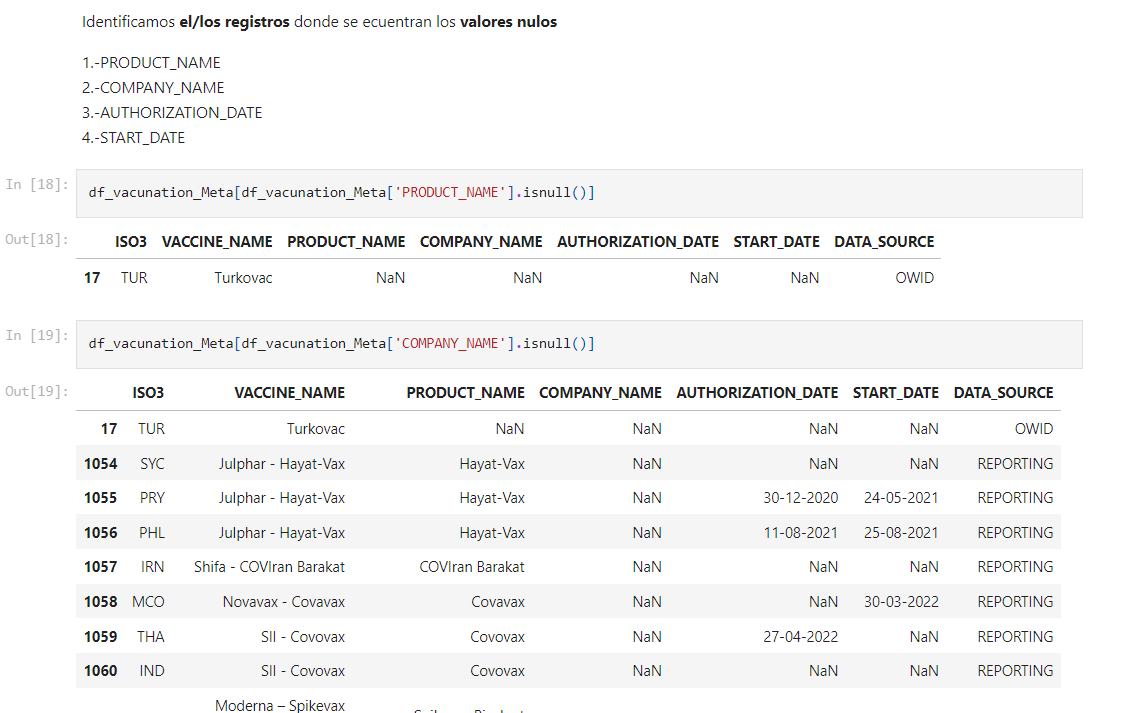
## Fuente 4: Tipos de Vacunas

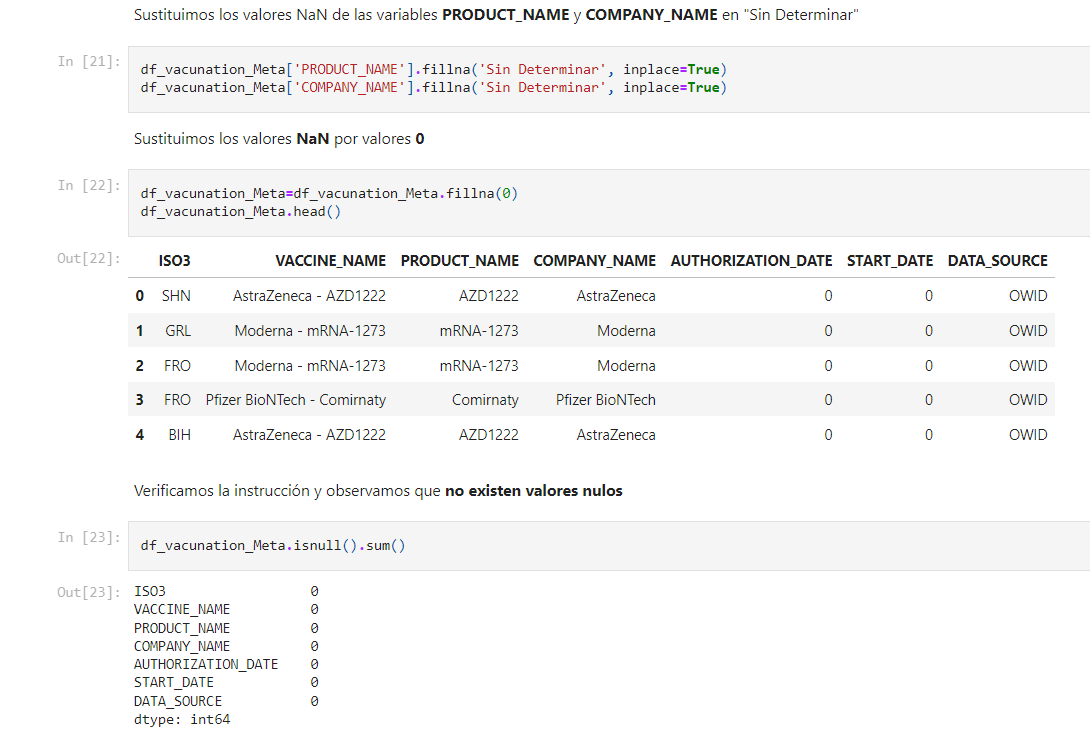
El fichero se puede encontrar en el siguiente enlace: <https://github.com/patriciaapenat/TFM/blob/main/Notebooks/Data_Frame_vacunas_tipo_oms.ipynb>

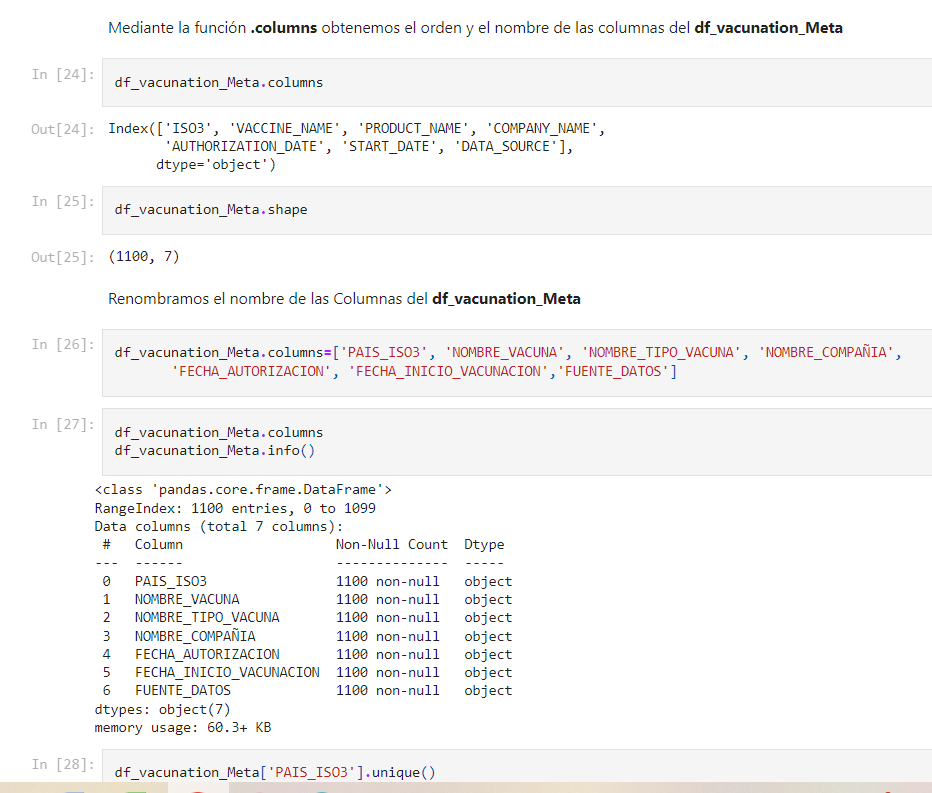


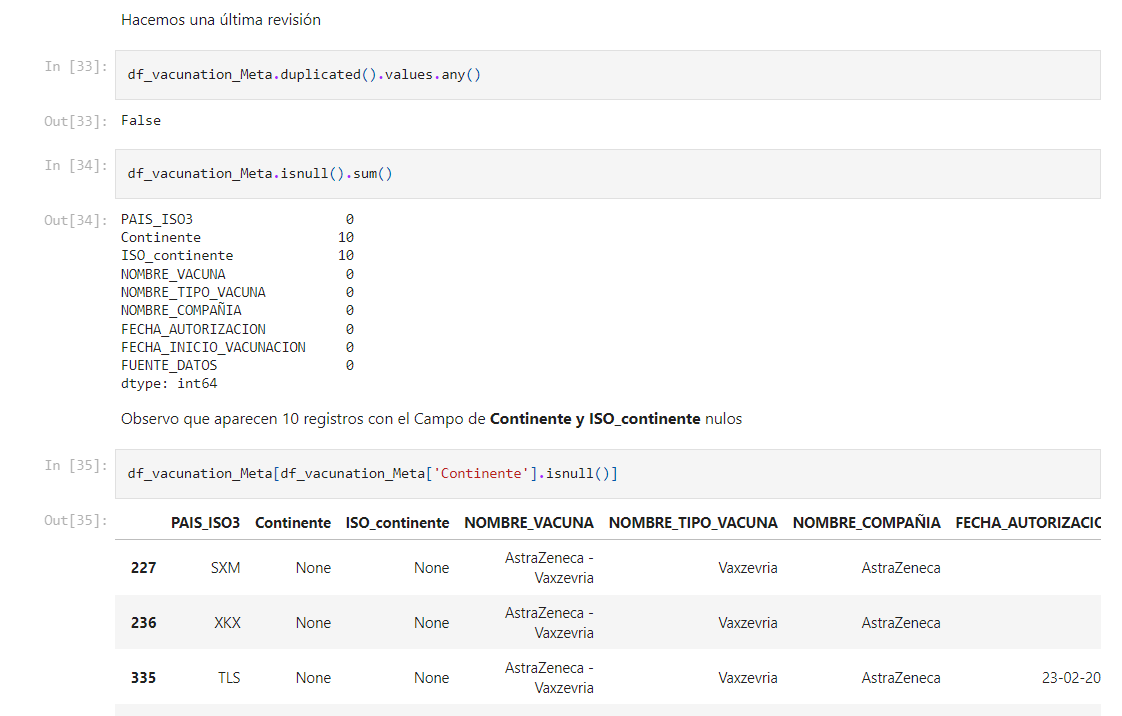


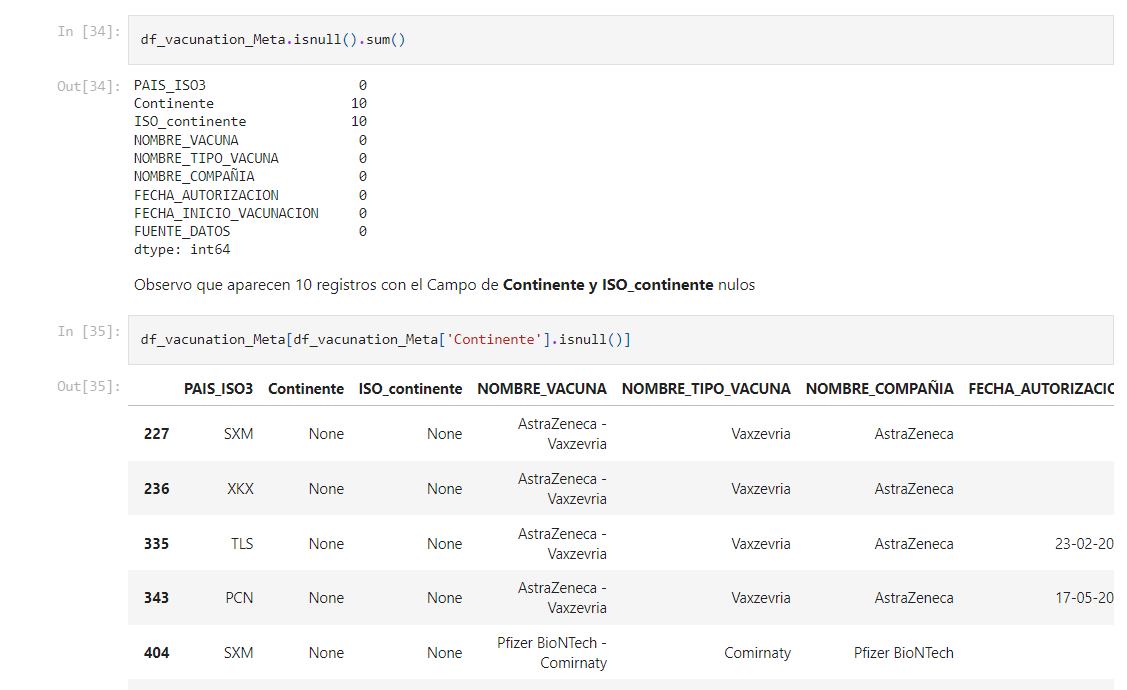


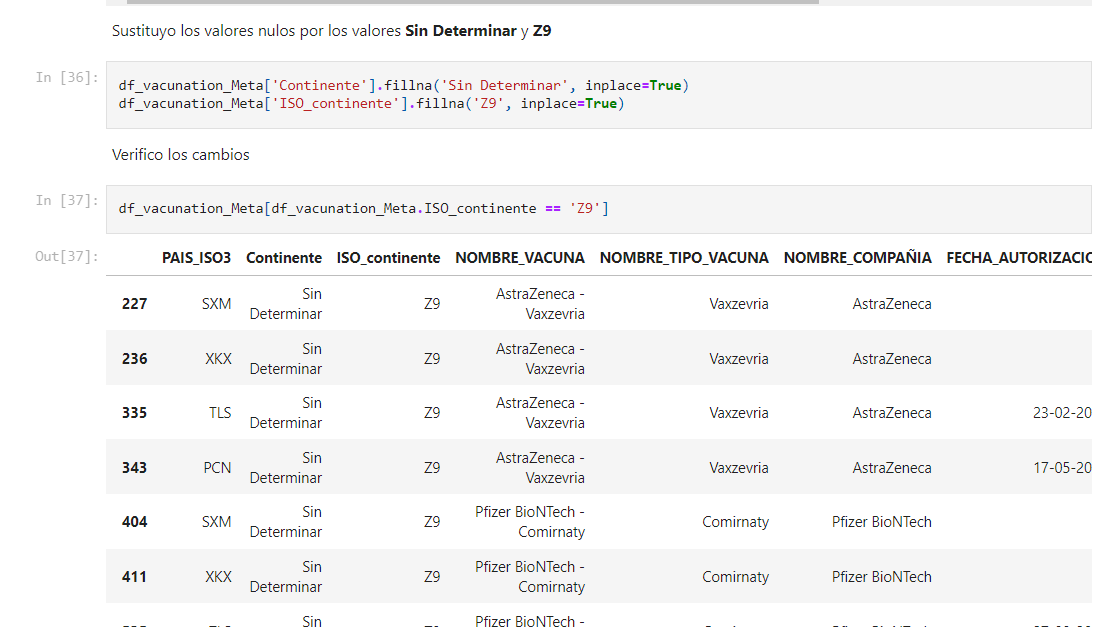


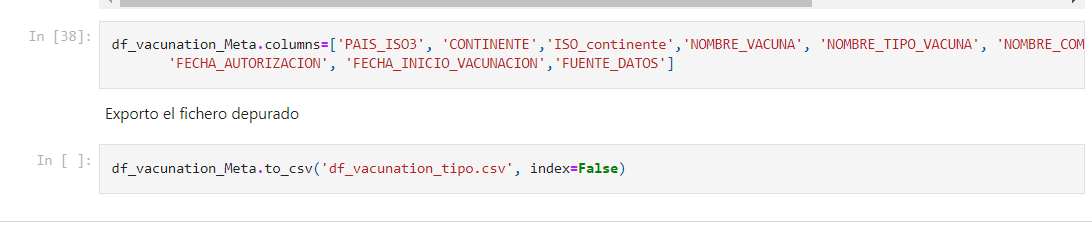












## Fuente 5: Data on testing for COVID-19 by week and country

El fichero se puede encontrar en el siguiente enlace: <https://github.com/patriciaapenat/TFM/blob/main/Notebooks/Data_Frame_Testing_eu.ipynb>

Primero se crea un diccionario con toda la información sobre las variables, hacer esto hará que sea más sencillo llevar a cabo la depuración

data\_dict = {

"country": "String",

"country\_code": "2-letter ISO country code",

"year\_week": "yyyy-Www",

"level": "National (archived dataset with national subnational data to week 36, 2022 is available on ECDC’s website)",

"region": "2-letter ISO country code where level is national",

"region\_name": "Country name where level is national",

"new\_cases": "Number of new confirmed cases",

"tests\_done": "Number of tests done",

"population": "Numeric",

"testing\_rate": "Testing rate per 100,000 population",

"positivity\_rate": "Weekly test positivity (%): 100 x Number of new confirmed cases/number of tests done per week",

"testing\_data\_source": [

"Country API",

"Country GitHub",

"Country website",

"Manual webscraping",

"Other",

"Survey",

"TESSy: data provided directly by Member States to ECDC via TESSy"

]

}

Ahora cargamos el dataset y configuramos nuestro entorno de trabajo

# importar paquetes

import pandas as pd

import numpy as np

import datetime as dt

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

# configurar los gráficos

sns.set\_style('whitegrid')

sns.set\_palette('flare')

Leemos el archivo

# Leer el archivo

df\_datos4 = pd.read\_csv('https://opendata.ecdc.europa.eu/covid19/testing/csv/data.csv') #cargamos los datos

df\_datos4.head()

Empezamos revisando la información

df\_datos4.info()

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 5460 entries, 0 to 5459

Data columns (total 12 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 country 5460 non-null object

1 country\_code 5460 non-null object

2 year\_week 5460 non-null object

3 level 5460 non-null object

4 region 5460 non-null object

5 region\_name 5460 non-null object

6 new\_cases 5265 non-null float64

7 tests\_done 4406 non-null float64

8 population 5460 non-null int64

9 testing\_rate 4406 non-null float64

10 positivity\_rate 4384 non-null float64

11 testing\_data\_source 4406 non-null object

dtypes: float64(4), int64(1), object(7)

memory usage: 512.0+ KB

# Revisamos si hay duplicados

df\_datos4.duplicated().sum().any()

False

# Revisamos si hay nulos

df\_datos4.isna().sum().any()

True

# Revisamos donde están los nulos

df\_datos4.isna().sum()

country 0  
country\_code 0  
year\_week 0  
level 0  
region 0  
region\_name 0  
new\_cases 195  
tests\_done 1054  
population 0  
testing\_rate 1054  
positivity\_rate 1076  
testing\_data\_source 1054  
dtype: int64

Una vez revisado el estatus inicial podemos empezar a hacer modificaciones

Este código convierte una columna en un DataFrame en una categoría y verifica que la conversión se realice correctamente.

# Modificar categoría

df\_datos4['testing\_data\_source'] = df\_datos4['testing\_data\_source'].astype('category'); assert df\_datos4['testing\_data\_source'].dtype == 'category'

Este código convierte la columna 'year\_week' en un DataFrame en un tipo de datos de fecha y hora utilizando la función 'pd.to\_datetime'. Luego, se extraen el número de semana y el año de la columna 'year\_week' y se almacenan en la misma columna, pero en formato de cadena de texto utilizando el método 'dt.strftime'.

# Convertir la columna 'year\_week' a tipo datetime

df\_datos4['year\_week'] = pd.to\_datetime(df\_datos4['year\_week'] + '-1', format='%Y-W%W-%w')

# Extraer el número de semana y el año

df\_datos4['year\_week'] = df\_datos4['year\_week'].dt.strftime('%Y-%W')

Ahora, tenemos que obtener el código ISO3

# Sabemos que disponemos del nombre del país en inglés

df\_datos4['country']:

# Sabemos que disponemos del nombre del país en inglés

df\_datos4['country']

0 Austria  
1 Austria  
2 Austria  
3 Austria  
4 Austria  
 ...   
5455 Sweden  
5456 Sweden  
5457 Sweden  
5458 Sweden  
5459 Sweden  
Name: country, Length: 5460, dtype: object

Podemos hacerlo utilizando una función que implemente el módulo pycountry, definimos obtener\_iso3

La función toma un parámetro country, que representa el nombre del país para el cual se desea obtener el código ISO 3. A continuación, utiliza la función pycountry.countries.get(name=country) para buscar el objeto Country correspondiente al nombre del país en la biblioteca pycountry.

Si se encuentra un objeto Country válido para el nombre del país, se devuelve su código ISO 3 utilizando el atributo alpha\_3. En caso de que no se encuentre un objeto Country válido, la función captura la excepción LookupError y no realiza ninguna acción adicional.

Finalmente, si se devuelve un código ISO 3 válido, este se asigna a la columna 'iso3' en el DataFrame 'df\_datos4' utilizando el método apply en la columna 'country'.

import pycountry

def obtener\_iso3(country):

try:

pais = pycountry.countries.get(name=country)

if pais is not None:

return pais.alpha\_3

except LookupError:

pass

return None

# Obtener el código ISO 3 correspondiente a los nombres de país en la columna 'country'

df\_datos4.insert(1, 'iso3', df\_datos4['country'].apply(obtener\_iso3))

Y ahora sólo verificamos que haya funcionado correctamente

df\_datos4['iso3']

0 AUT

1 AUT

2 AUT

3 AUT

4 AUT

...

5455 SWE

5456 SWE

5457 SWE

5458 SWE

5459 SWE

Name: iso3, Length: 5460, dtype: object

Después de revisar los valores nulos vemos que sería mejor eliminarlos así procedemos a ello

# Eliminar nulos

df\_datos4.dropna(subset=['new\_cases', 'tests\_done', 'testing\_rate', 'positivity\_rate', 'testing\_data\_source'], inplace=True)

Y eliminamos las columnas que no utilizaremos

df\_datos4.drop(['country\_code', 'region\_name', 'region', 'country'], axis=1, inplace=True)

### Verificamos

df\_datos4.info()

<class 'pandas.core.frame.DataFrame'>

Int64Index: 4384 entries, 40 to 5459

Data columns (total 9 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 iso3 4384 non-null object

1 year\_week 4384 non-null object

2 level 4384 non-null object

3 new\_cases 4384 non-null float64

4 tests\_done 4384 non-null float64

5 population 4384 non-null int64

6 testing\_rate 4384 non-null float64

7 positivity\_rate 4384 non-null float64

8 testing\_data\_source 4384 non-null category

dtypes: category(1), float64(4), int64(1), object(3)

memory usage: 312.6+ KB

Y por último exportamos a CSV

import os

def guardar\_como\_csv(df, nombre\_archivo):

# Obtener la ruta completa del directorio actual

current\_directory = os.getcwd()

# Definir la ubicación y el nombre del archivo CSV

file\_path = os.path.join(current\_directory, nombre\_archivo)

# Guardar el DataFrame como un archivo CSV en la ubicación especificada

df.to\_csv(file\_path, index=False)

# Llamar a la función para guardar el DataFrame df\_datos4 como un archivo CSV

guardar\_como\_csv(df\_datos4, "DEFMODTestingCovid19.csv")

**Propuesta para la base de datos SQL**

* iso3: VARCHAR o CHAR (cadena de caracteres de longitud fija o variable que representa un código de país de 3 letras).
* year\_week: VARCHAR o CHAR (cadena de caracteres de longitud fija o variable que representa una semana en formato "año-semana").
* level: VARCHAR o CHAR (cadena de caracteres de longitud fija o variable que representa el nivel geográfico).
* new\_cases: FLOAT o DECIMAL (número decimal que representa la cantidad de nuevos casos).
* tests\_done: FLOAT o DECIMAL (número decimal que representa la cantidad de pruebas realizadas).
* population: INTEGER o BIGINT (número entero que representa la población).
* testing\_rate: FLOAT o DECIMAL (número decimal que representa la tasa de pruebas).
* positivity\_rate: FLOAT o DECIMAL (número decimal que representa la tasa de positividad).
* testing\_data\_source: VARCHAR o CHAR (cadena de caracteres de longitud fija o variable que representa la fuente de datos de pruebas).

## Fuente 6: Data on hospital and ICU admission rates and current occupancy for COVID-19

El fichero se puede encontrar en el siguiente enlace: <https://github.com/patriciaapenat/TFM/blob/main/Notebooks/Data_Frame_Pacientes_Hospitalizados%20%26%20UCI_eu.ipynb>

**Fuente de Datos - European Centre for Disease Prevention and Control**

[**https://www.ecdc.europa.eu/en/publications-data/download-data-hospital-and-icu-admission-rates-and-current-occupancy-covid-19**](https://www.ecdc.europa.eu/en/publications-data/download-data-hospital-and-icu-admission-rates-and-current-occupancy-covid-19)

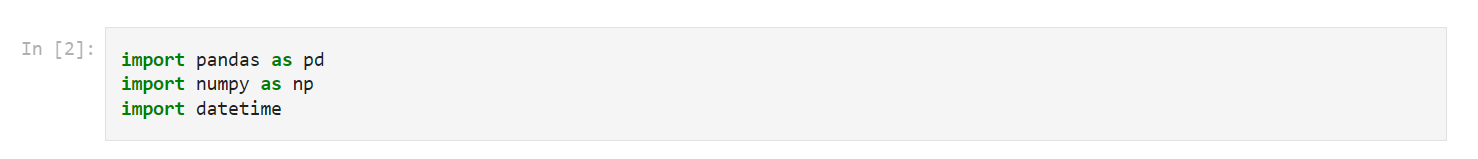
**Descarga de datos**

Ingresos en hospital UCI fueron registrados: <https://www.ecdc.europa.eu/en/publications-data/download-data-hospital-and-icu-admission-rates-and-current-occupancy-covid-19>

**Informacion del Dataset**  
Los usuarios deben tener en cuenta que los archivos de datos descargables contienen información sobre las tasas de admisión de hospitalización y Unidad de Cuidados Intensivos (UCI) y la ocupación actual para COVID-19 por fecha y país. Cada fila contiene los datos correspondientes a una determinada fecha (día o semana) y por país. El archivo se actualiza semanalmente. Puede utilizar los datos de acuerdo con la política de derechos de autor del ECDC.

El dataset que estaremos trabajando contiene Informacion acerca de Ocupación diaria del hospital, Ocupación diaria de la UCI, Nuevos ingresos hospitalarios semanales por 100k, Nuevas admisiones semanales en UCI por 100k (Daily hospital occupancy,Daily ICU occupancy, Weekly new hospital admissions per 100k, ,Weekly new ICU admissions per 100k) en paises europeos.

Columnas: country, indicator, date, year\_week, value, source, url



**Acceso al dataset de hospitalizaciones en UCI**

Exloración e información del DataFrame df\_hosp\_UCI:

****

Carga del df países para realizar el “join”:

****

****

Depuración del nuevo df:

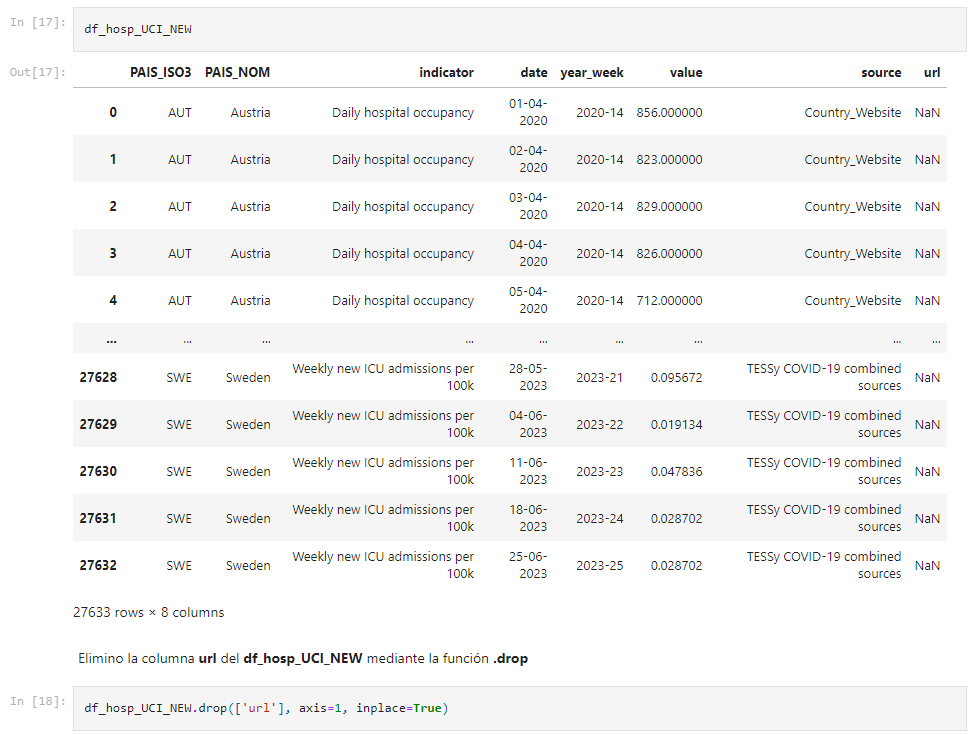
****

****

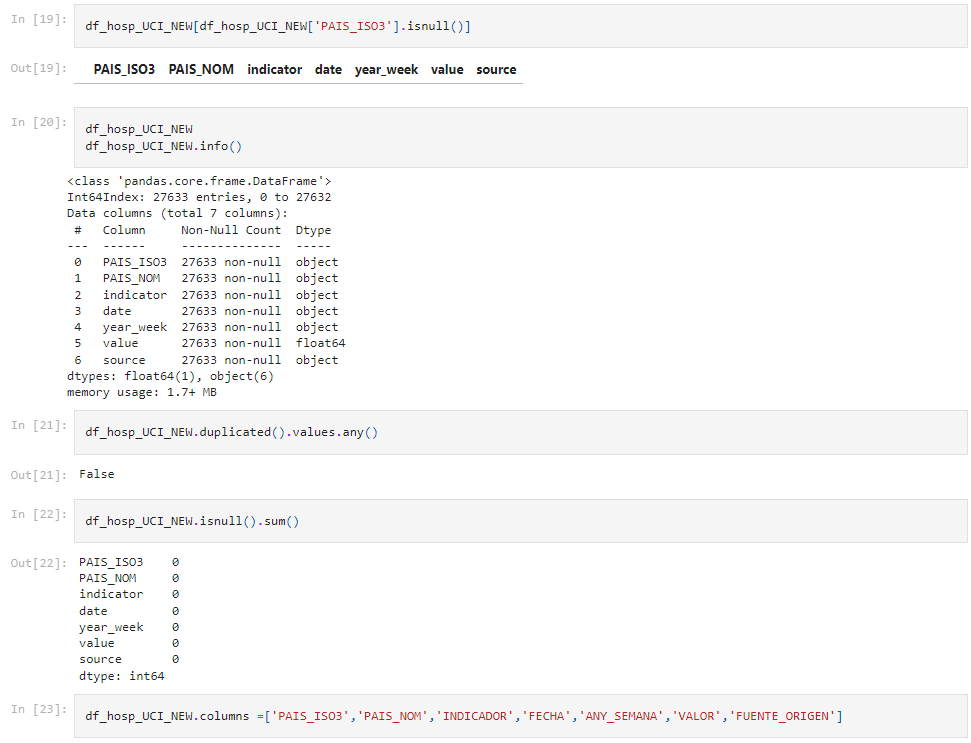
Asignación de formato fecha a columna “year\_week”:

****

Eliminación de la columna “url”:

****

Validación de nulos:

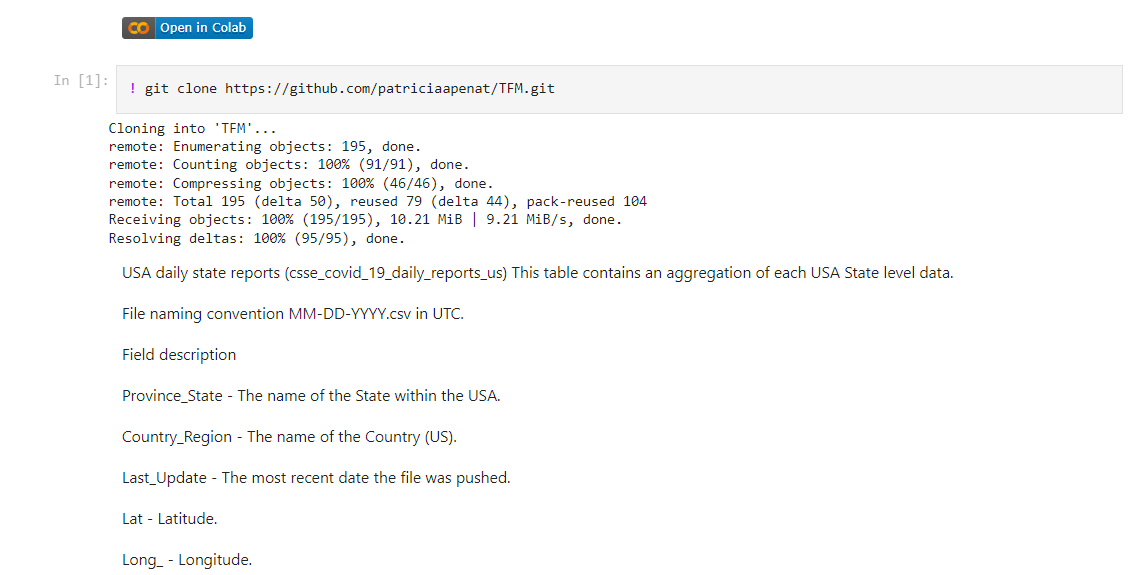
****

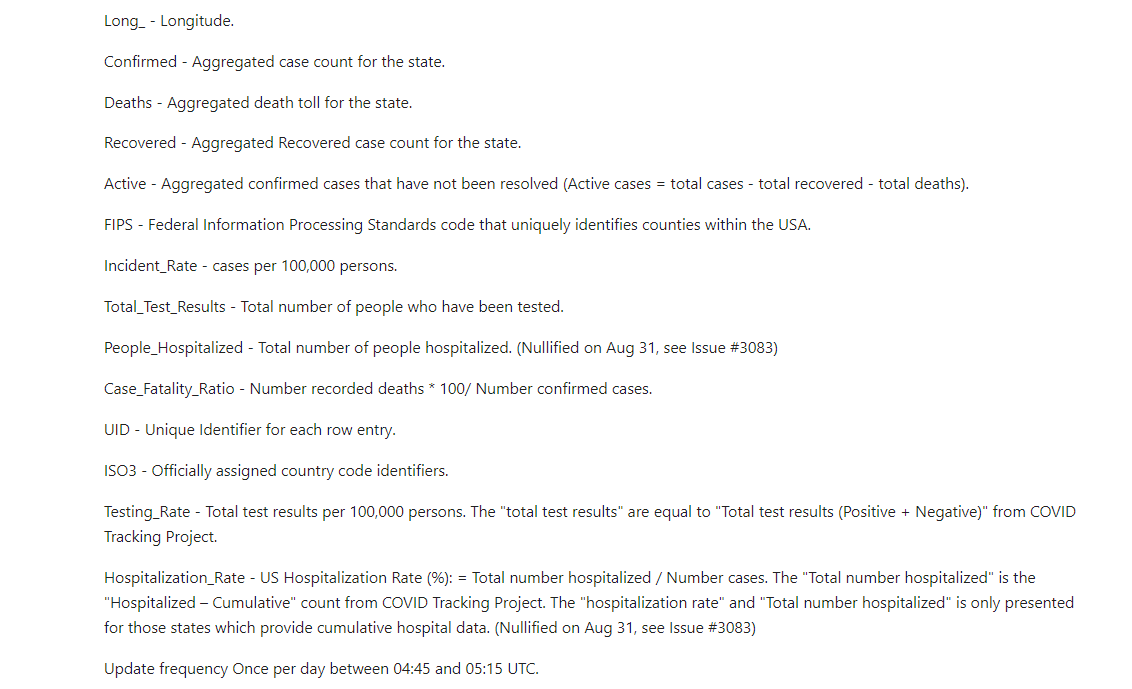
Exportación del df final a formato CSV:

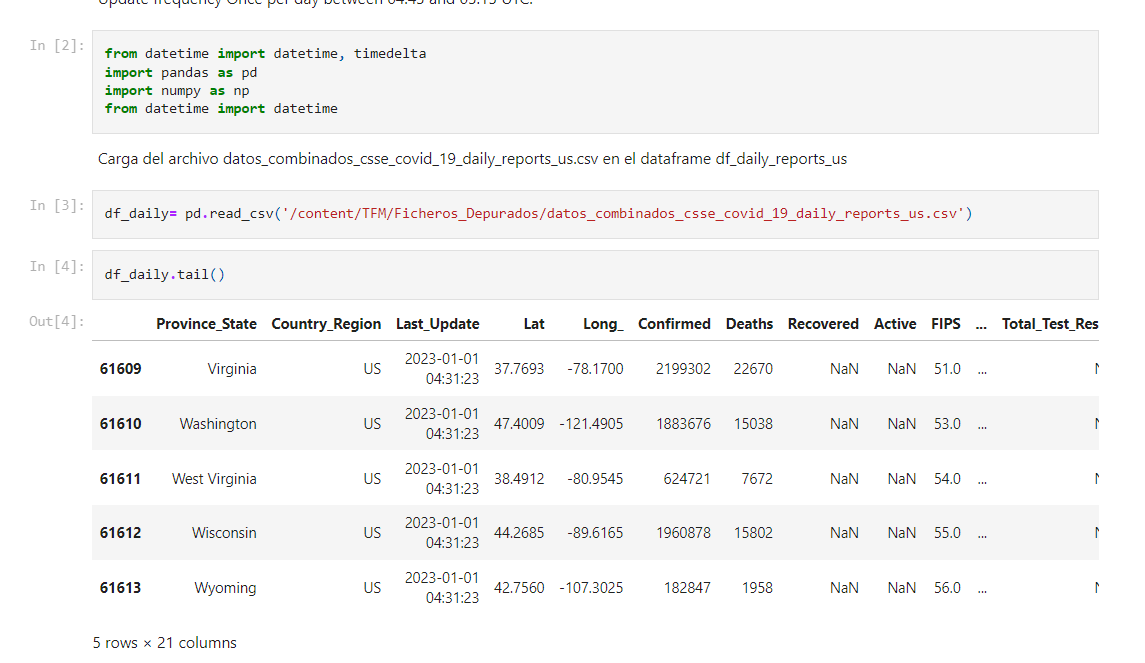
****

## Fuente 7: Datos sobre casos diarios registrados por estado en USA

El fichero se puede encontrar en el siguiente enlace: <https://github.com/patriciaapenat/TFM/blob/main/Notebooks/Data_Frame_csse_covid_daily_report_US.ipynb>



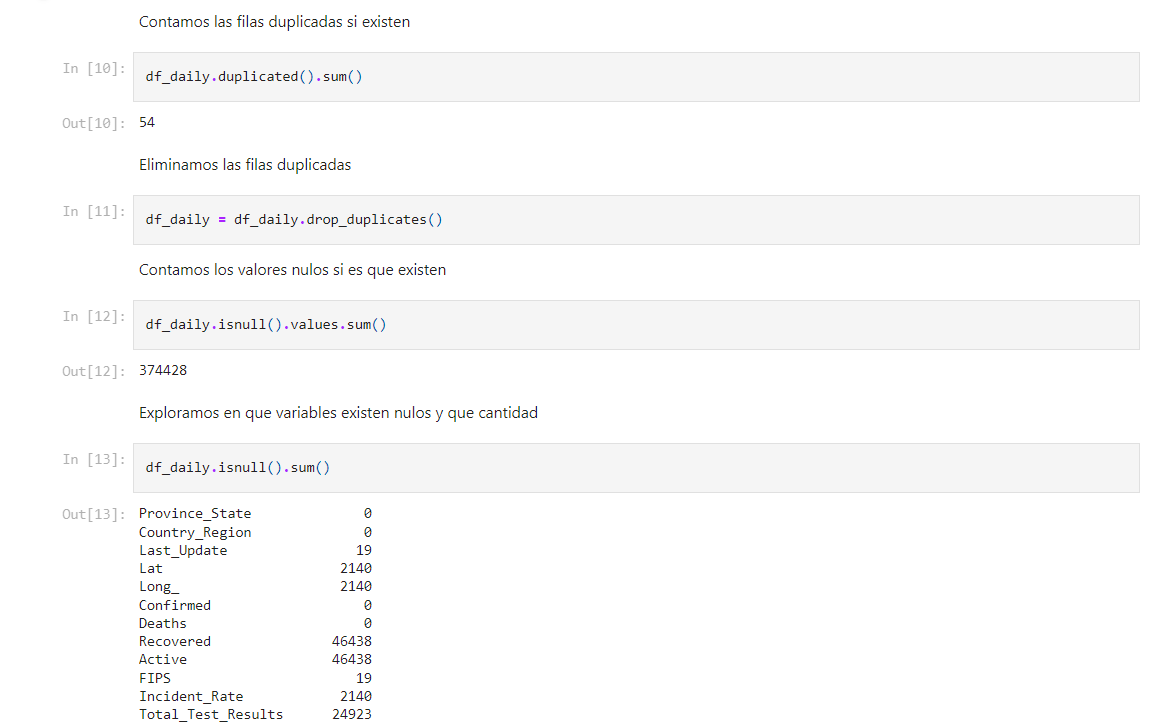


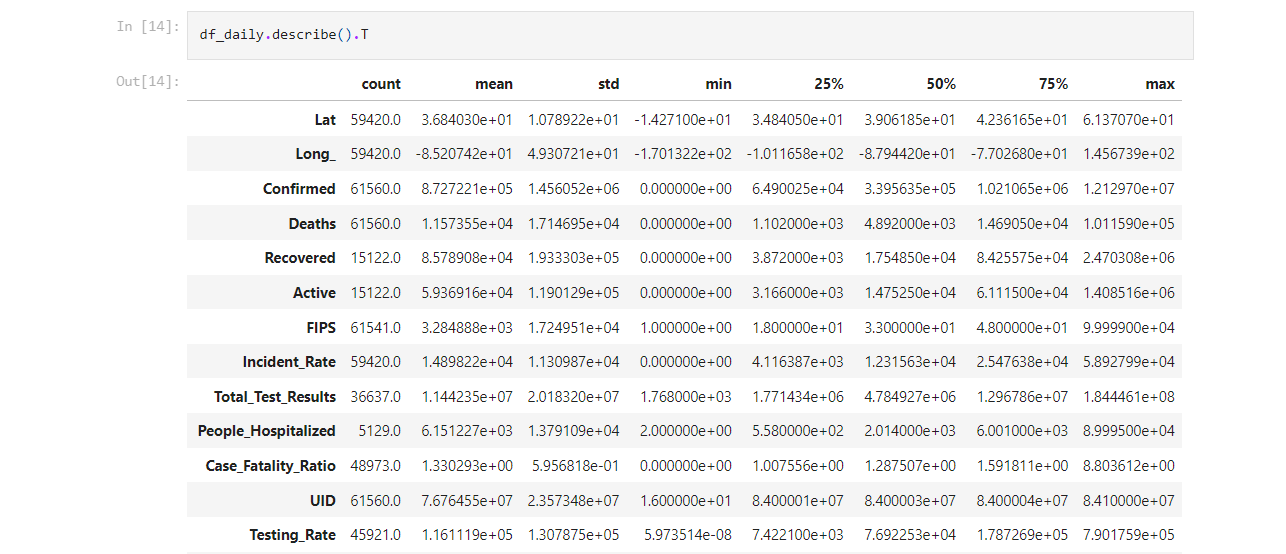


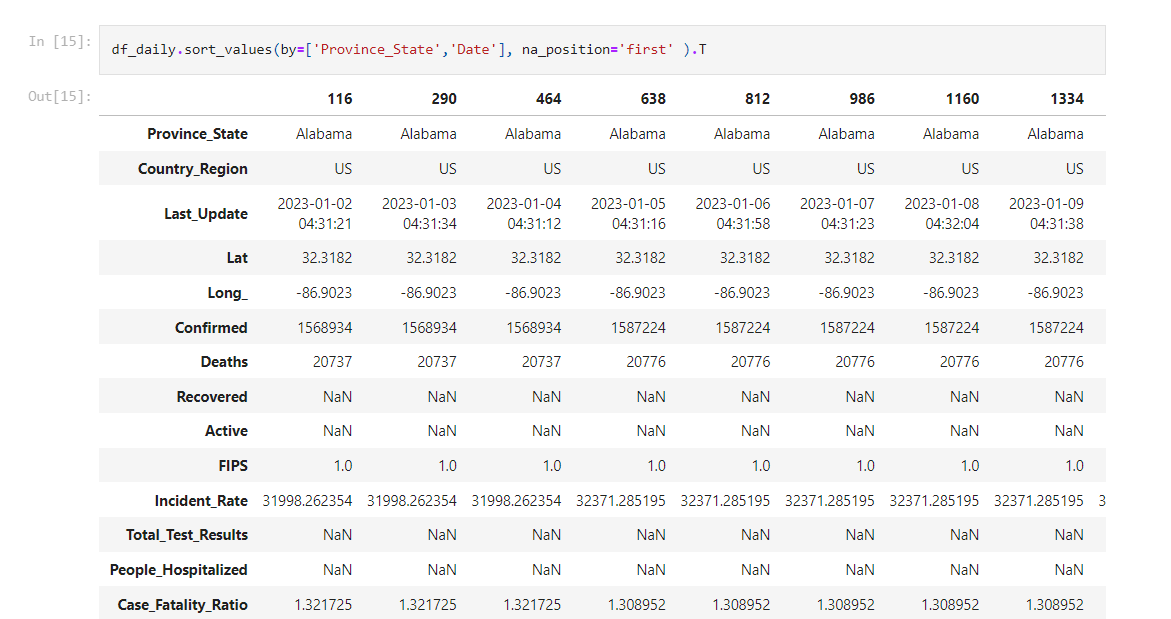


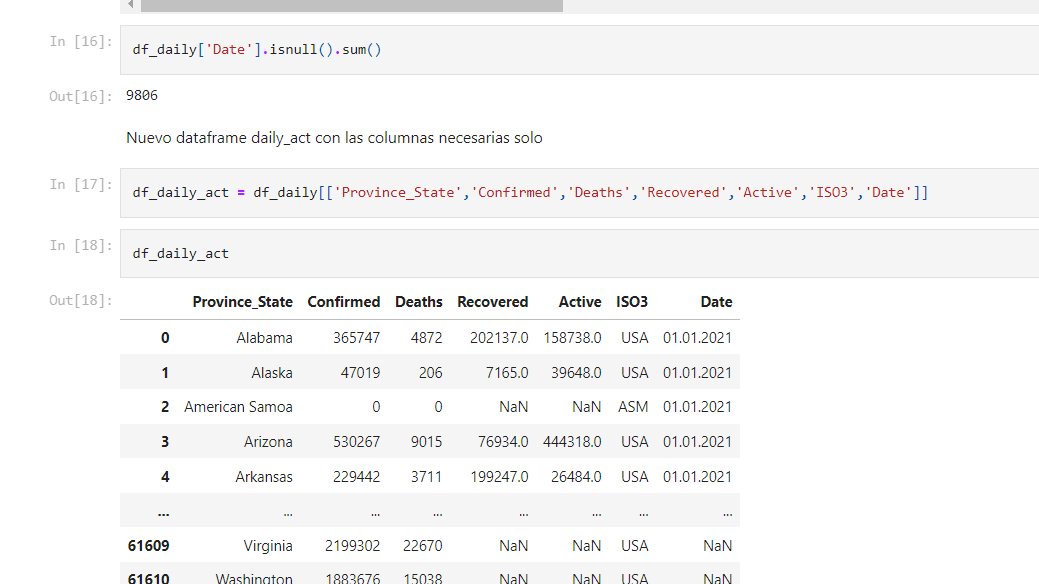


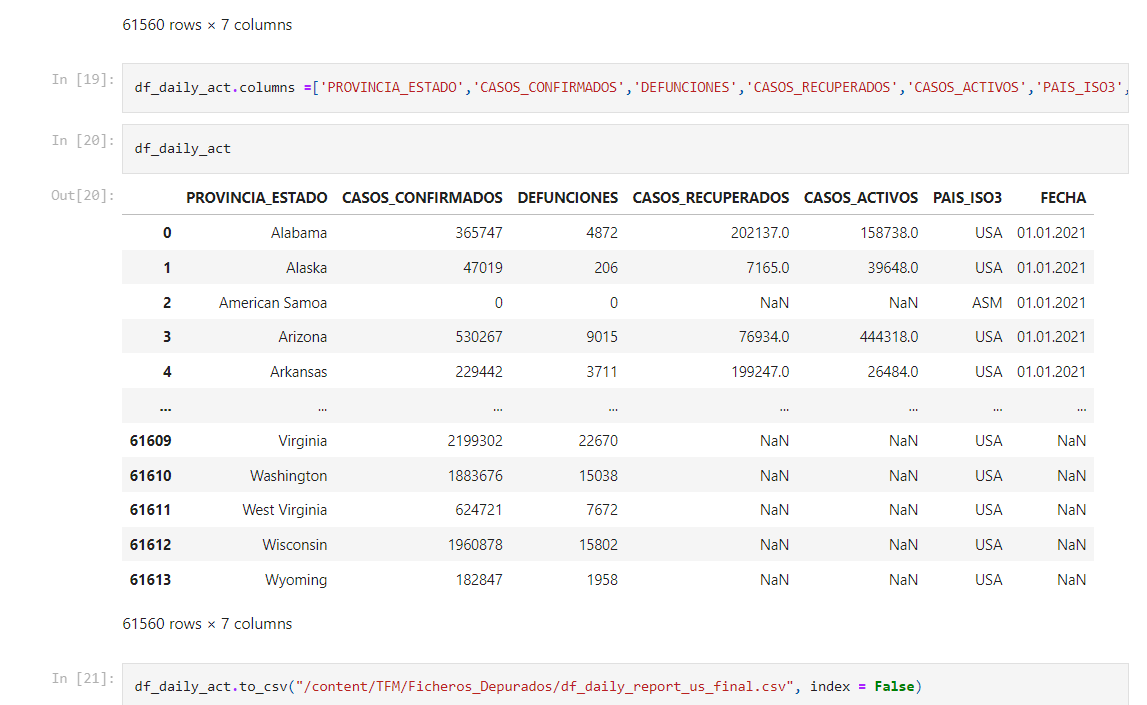


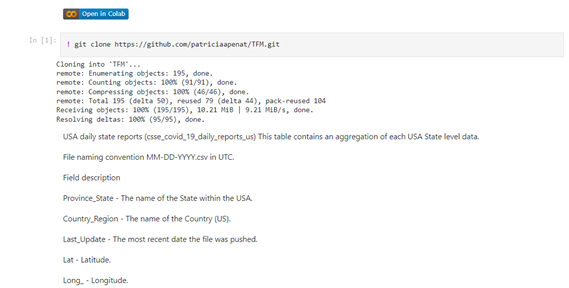


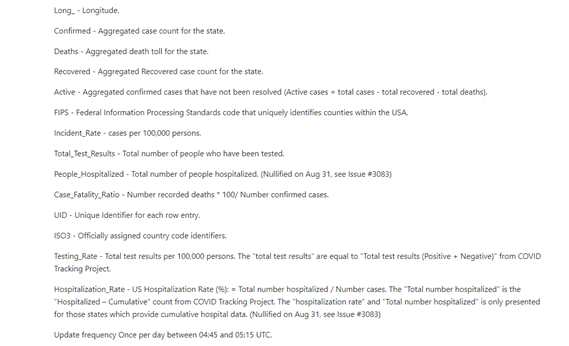


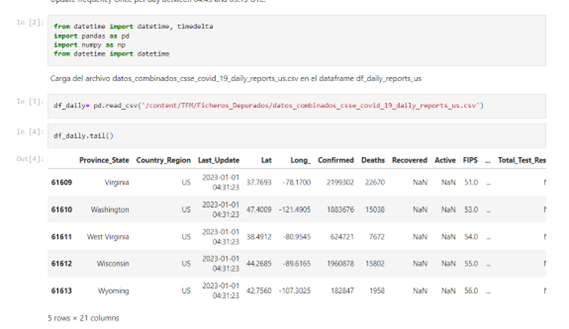






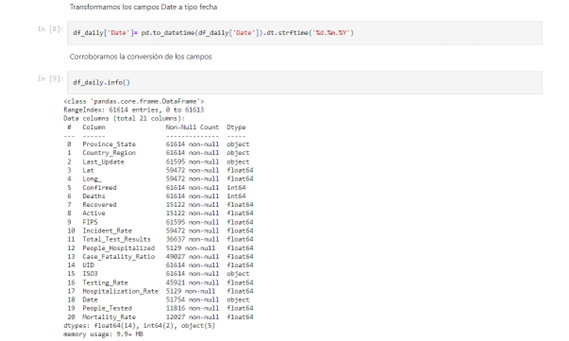


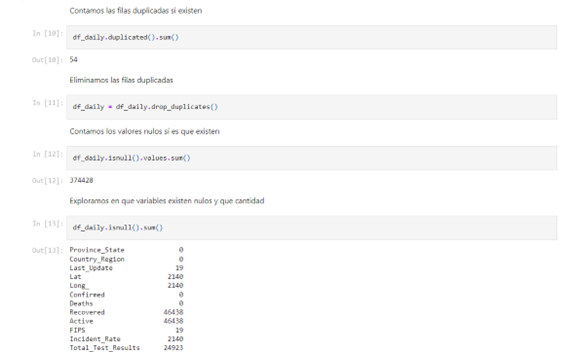






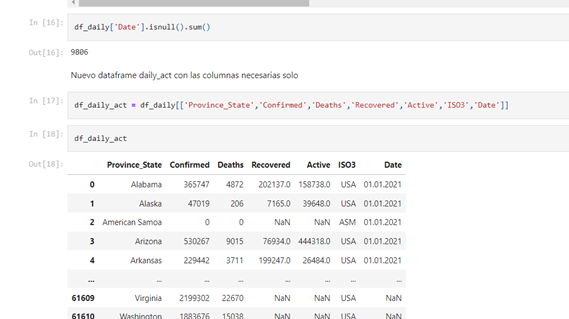


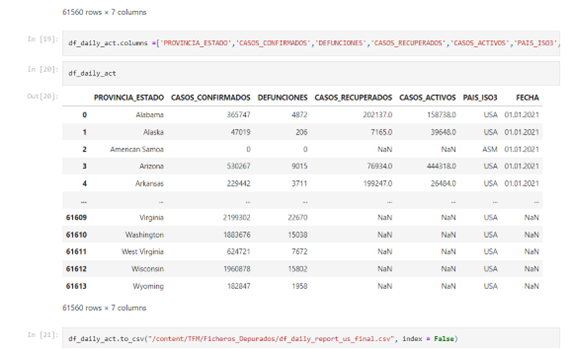












# BASE DE DATOS Y PROCEDIMIENTO DE CARGA

Creamos una BBDD MySQL donde almacenaremos todos los ficheros depurados, normalizados y transformados. El Script de la BBDD es el siguiente y podemos observar los campos que contiene cada tabla, así como la estructura y tipologia de datos de la misma.

## Script para creación de BBDD

**CREATE SCHEMA PROVA\_TFM;**

CREATE TABLE PROVA\_TFM.**PAISES (**

PAIS\_ISO3 VARCHAR(3) NOT NULL,

PAIS\_ISO2 VARCHAR(2) NOT NULL ,

PAIS\_NOM VARCHAR(100) NOT NULL,

COD\_CONTINENTE VARCHAR(2) NOT NULL,

CONTINENTE VARCHAR(100) NOT NULL,

OMS\_REGION VARCHAR(5) NOT NULL,

DESC\_OMS\_REGION VARCHAR(100) NOT NULL,

PAIS\_NOM\_2 VARCHAR(100) NOT NULL,

PRIMARY KEY (PAIS\_ISO3));

CREATE TABLE PROVA\_TFM.**VACUNACIONES (**

PAIS VARCHAR(50) NOT NULL,

PAIS\_ISO3 VARCHAR(3) NOT NULL,

CONTINENTE VARCHAR(50) NOT NULL,

ISO\_CONTINENTE VARCHAR(2) NOT NULL,

OMS\_REGION VARCHAR(50) NOT NULL ,

FUENTE\_DATOS VARCHAR(50) NULL ,

FECHA\_ULT\_ACTUALIZACION VARCHAR(10) NULL ,

TOTAL\_VACUNACION\_ACUM DOUBLE NULL,

NPER\_VACUNADAS\_1DOSIS INT NULL,

TOTAL\_VACUNACION\_PER100 DOUBLE NULL,

NPER\_VACUNADAS\_1DOSIS\_PER100 DOUBLE NULL,

NPER\_VACUNADAS\_DOSIS\_FULL INT NULL,

NPER\_VACUNADAS\_DOSIS\_FULL\_PER100 DOUBLE NULL,

FECHA\_PRIMERA\_VACUNA VARCHAR(10) NULL,

N\_TIPOS\_VACUNAS\_USADAS DOUBLE NULL,

NPER\_CON\_DOSIS\_ADICIONAL DOUBLE NULL,

NPER\_CON\_DOSIS\_ADIDICIONAL\_PER100 DOUBLE NULL,

PRIMARY KEY (PAIS\_ISO3));

CREATE TABLE PROVA\_TFM.**VACUNAS\_TIPOS (**

PAIS\_ISO3 VARCHAR(3) NOT NULL,

CONTINENTE VARCHAR(50) NOT NULL,

ISO\_CONTINENTE VARCHAR(2) NOT NULL,

NOMBRE\_VACUNA VARCHAR(100) NULL ,

NOMBRE\_TIPO\_VACUNA VARCHAR(90) NULL ,

NOMBRE\_COMPAÑIA VARCHAR(50) NULL ,

FECHA\_AUTORIZACION VARCHAR(10) NULL,

FECHA\_INICIO\_VACUNACION VARCHAR(10) NULL,

FUENTE\_DATOS VARCHAR(50) NULL);

CREATE TABLE PROVA\_TFM.**CASOS\_CONFIRMADOS\_DEFUNCIONES\_US (**

PROVINCIA\_ESTADO VARCHAR(50) NOT NULL,

CASOS\_CONFIRMADOS INT NULL,

DEFUNCIONES INT NULL,

CASOS\_RECUPERADOS INT NULL,

CASOS\_ACTIVOS INT NULL,

PAIS\_ISO3 VARCHAR(3) NOT NULL,

FECHA VARCHAR(10) NULL);

CREATE TABLE PROVA\_TFM.**CASOS\_HOSPITALIZADOS\_UCI\_EU (**

PAIS\_ISO3 VARCHAR(3) NOT NULL,

PAIS\_NOM VARCHAR(100) NOT NULL,

INDICADOR VARCHAR(100) NULL,

FECHA VARCHAR(10) NULL,

ANY\_SEMANA VARCHAR(10) NOT NULL,

VALOR DOUBLE NULL,

FUENTE\_ORIGEN VARCHAR(100) NULL);

CREATE TABLE PROVA\_TFM.**TESTING\_COVID\_EU (**

PAIS\_ISO3 VARCHAR(3) NOT NULL,

ANY\_SEMANA VARCHAR(10) NOT NULL,

NIVEL VARCHAR(50) NOT NULL,

CASOS\_NUEVOS INT NULL,

N\_TESTS\_REALIZADOS INT NULL,

POBLACION INT NULL,

RATIO\_TESTS DOUBLE NULL,

RATIO\_POSITIVO DOUBLE NULL,

FUENTES\_TESTS VARCHAR(50) NULL);

CREATE TABLE PROVA\_TFM.**COVID\_DAILY (**

PAIS VARCHAR(50) NOT NULL,

PAIS\_ISO3 VARCHAR(3) NOT NULL,

PAIS\_ISO2 VARCHAR(2) NOT NULL,

FECHA\_INFORMADA VARCHAR(10) NOT NULL,

OMS\_REGION VARCHAR(50) NOT NULL ,

CASOS\_NUEVOS INT NULL,

CASOS\_ACUM INT NULL,

MUERTES\_NUEVAS INT NULL,

MUERTES\_ACUM INT NULL );

CREATE TABLE PROVA\_TFM.**CASOS\_COVID\_ULTIMO\_RECUENTO (**

PAIS\_ISO3 VARCHAR(3) NOT NULL,

OMS\_REGION VARCHAR(50) NOT NULL ,

CASOS\_ACUM\_TOTAL INT NULL,

CASOS\_ACUM\_TOTAL\_POR\_100000\_HAB DOUBLE NULL,

CASOS\_NUEVOS\_INFORMADOS\_ULT\_7\_DIAS INT NULL,

CASOS\_NUEVOS\_INFORMADOS\_ULT\_7\_DIAS\_POR\_100000\_HAB DOUBLE NULL,

CASOS\_NUEVOS\_INFORMADOS\_ULT\_24H INT NULL,

MUERTES\_ACUM\_INFORMADAS\_ULT\_7\_DIAS INT NULL,

MUERTES\_ACUM\_TOTAL\_POR\_100000\_HAB DOUBLE NULL,

MUERTES\_NUEVAS\_INFORMADAS\_ULT\_7\_DIAS INT NULL,

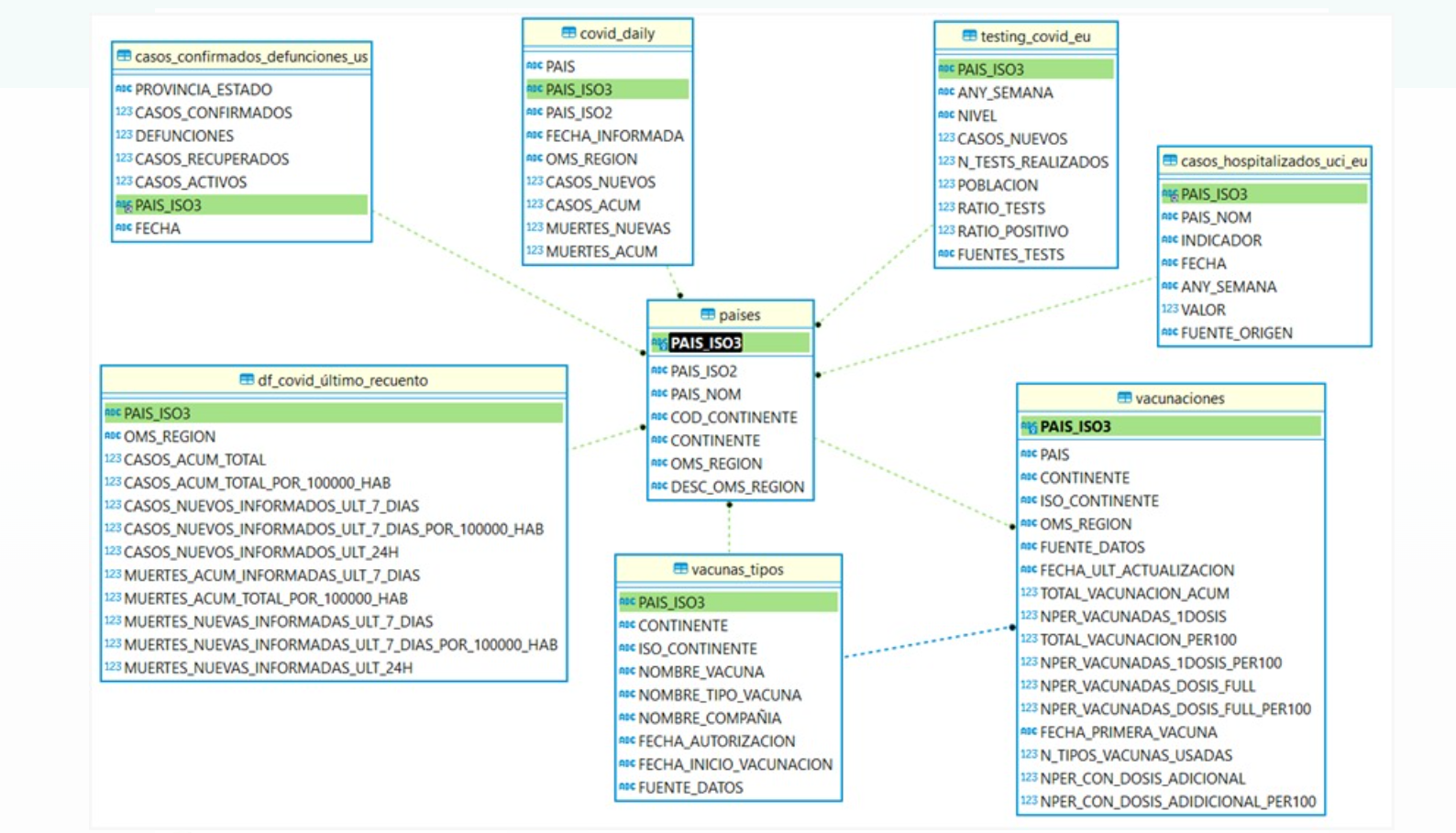
MUERTES\_NUEVAS\_INFORMADAS\_ULT\_7\_DIAS\_POR\_100000\_HAB DOUBLE NULL,

MUERTES\_NUEVAS\_INFORMADAS\_ULT\_24H INT NULL );

## Modelo de datos de la BBDD – TFM COVID-19

Hemos construido una BBDD en Modo estrella con la tabla “**Paises**” como tabla principal o maestra en la que se pueden observar las diferentes relaciones con el resto de tablas de dimensiones.

El Campo **PAIS\_ISO3** es la clave principal de la Tabla de Paises y se relaciona con el resto de tablas de dimensiones a través del mismo campo **PAIS\_ISO3** que pasa a ser la clave secundaria.



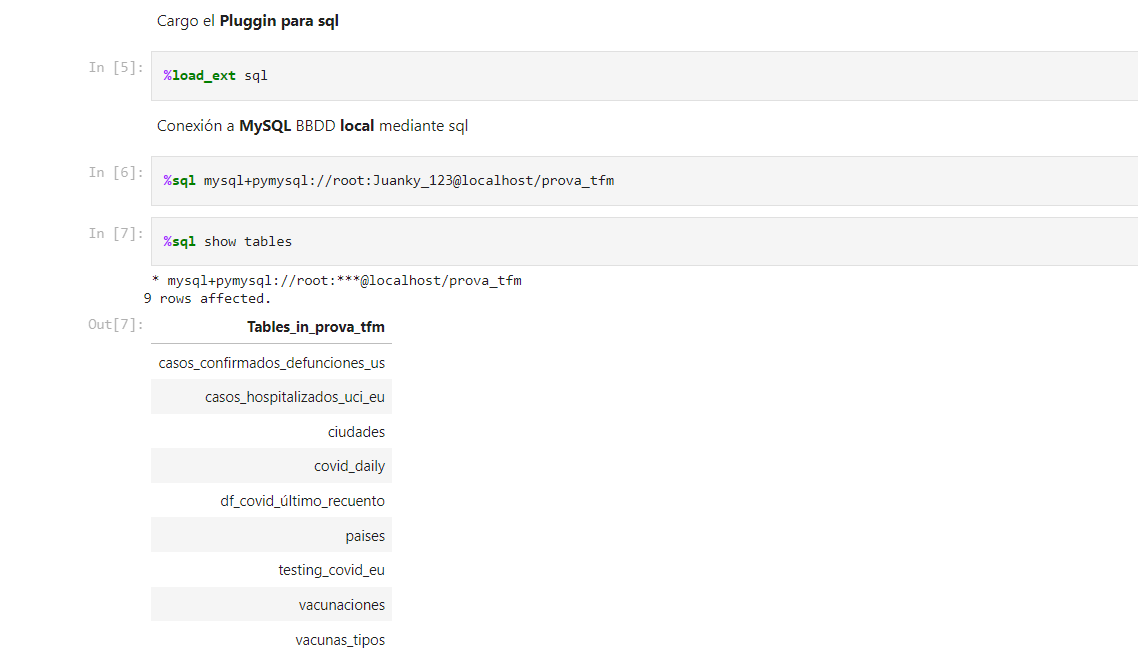
## Notebook Jupyter de Conexión y Carga de Ficheros a la BBDD – TFM COVID-19

En el Siguiente Notebook se realiza la Conexión a la BBDD, interacción con las tablas que forman la BBDD, así como la carga de ficheros depurados y actualizados.

Dicho Notebook se encuentra en el repositorio de GitHUB // TFM/Notebooks

[Carga ficheros a BBDD.ipynb](https://github.com/patriciaapenat/TFM/blob/main/Notebooks/Carga%20ficheros%20a%20BBDD.ipynb)











# CONCLUSIONES

En el presente entregable de nuestro Trabajo de Fin de Máster (TFM) se ha llevado a cabo la depuración, transformación y carga de diversas fuentes de datos relacionadas con el COVID-19, como los datos de pruebas realizadas por semana y país en Europa, informes diarios de casos en los Estados Unidos, datos de hospitalización y ocupación de unidades de cuidados intensivos, metadatos y datos de vacunación, entre otros.

La limpieza y transformación de los datos se ha realizado con el objetivo de obtener una base de datos consistente y lista para su análisis posterior. Se ha utilizado el lenguaje SQL para montar y gestionar las bases de datos, así como los dataframes generados.

En cuanto a las fuentes de información utilizadas, se ha trabajado con datos proporcionados por la Organización Mundial de la Salud (OMS), informes diarios de casos y muertes, datos de vacunación, así como datos de pruebas de COVID-19 por semana y país. También se han utilizado datos específicos de los Estados Unidos, como los informes diarios de casos por estado y datos de hospitalización, asimismo, datos sobre las pruebas de Covid-19 en el continente europeo.

Durante el proceso de depuración, transformación y enriquecimiento de los datos, se ha llevado a cabo la integración de diferentes fuentes para obtener una visión más completa de la situación de la pandemia. Además, se han desarrollado tablas adicionales, como la tabla de países con información relevante, en los anexos encontrarán también información sobre nuevas tablas constituidas como son de indicadores de desarrollo del banco mundial seleccionados por nuestro equipo y la tabla de datos climáticos globales.

En cuanto a las hipótesis formuladas para la próxima entrega, estas representan suposiciones basadas en las relaciones esperadas entre variables relevantes y la incidencia, propagación y consecuencias del COVID-19. Las hipótesis nulas, por otro lado, representan la ausencia de dichas relaciones. En el próximo proceso de análisis de datos, pruebas estadísticas, se evaluará la evidencia en apoyo o en contra de estas hipótesis, lo que contribuirá a la comprensión y conocimiento sobre la pandemia de COVID-19, así mismo estos datos se llevarán a dashboards elaborados en Power BI que serán acompañados de su respectivo storytelling.

Es importante destacar que el análisis de las hipótesis requerirá un enfoque riguroso, utilizando métodos adecuados de análisis estadístico y teniendo en cuenta otros factores que puedan influir en los resultados. Además, se respaldarán las afirmaciones con bibliografía académica especializada para garantizar la validez y la robustez de los resultados obtenidos.

En resumen, a pesar de las dificultades encontradas durante el proceso, se ha logrado cumplir con las tareas establecidas en esta segunda entrega del TFM. Se ha realizado la depuración, transformación y carga de los datos, y se ha preparado el terreno para el análisis y la respuesta a las hipótesis planteadas. El uso de Python se ha destacado como herramientas eficaces para llevar a cabo estos procesos. A medida que avancemos en la investigación, se espera obtener resultados que contribuyan al entendimiento de la pandemia de COVID-19.

# ANEXOS

En cuanto a tema de análisis de datos hemos preparado dos tablas adicionales a la tabla países para poder realizar integraciones: WDI y temp.

Entonces, además de las tablas de covid tenemos

* paises, que contiene información sobre cada país, región OMS, continente, ISO2, ISO3, etc.
* WID, que contiene información de 2018 a 2022 de algunos indicadores de desarrollo seleccionados por el equipo
* temp, que contiene datos sobre el clima global en 2013, si bien sería mejor tener datos más cercanos ha sido imposible encontrar un dataset similar con estas características puesto que muchos suelen tener registros por segundos (lo que nos resulta imposible de manejar por cuestiones informáticas) o por estaciones de control climático sin divisiones territoriales

**Preparación del entorno de trabajo**

import pandas as pd

import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt

import IPython as display

import datetime as dt

# configurar los gráficos

sns.set\_style('whitegrid')

sns.set\_palette('mako')

**Tabla países**

Debido a que hay código como NN y NA siempre hay que revisar los NaN o Z9

print(paises.loc[paises['COD\_CONTINENTE'] == 'Z9','CONTINENTE'])

Se agrega el código faltante

paises.loc[(paises['COD\_CONTINENTE'] == 'Z9') & (paises['CONTINENTE'] == 'North America'), 'COD\_CONTINENTE'] = 'NA'

paises.loc[(paises['COD\_CONTINENTE'] == 'Z9') & (paises['CONTINENTE'] == 'Europe')

Y agregamos las coordenadas

coord = pd.read\_csv("C:/Users/Patricia/Desktop/Github/TFM/DF Covid/world\_country\_and\_usa\_states\_latitu

# Realiza la fusión utilizando el código ISO2 como clave

paises = paises.merge(coord[['country\_code', 'longitude', 'latitude']],

left\_on='PAIS\_ISO2', right\_on='country\_code', how='left')

paises.loc[pd.isna(paises['longitude'])]

# Define las coordenadas de longitud y latitud correspondientes

longitudes = [17.0, -177.0, -68.98, np.nan, -62.83, -63.06, np.nan]

latitudes = [-22.0, 0.0, 12.17, np.nan, 17.9, 18.07, np.nan]

# Itera sobre los países faltantes y actualiza los valores correspondientes en la tabla

for i, pais in enumerate(['Namibia', 'United States Minor Outlying Islands', 'Curaçao', 'Sudán del Sur', 'San Bartolomé', 'San Martín', 'San Martín']):

index = paises[paises['PAIS\_NOM'] == pais].index[0]

paises.at[index, 'longitude'] = longitudes[i]

paises.at[index, 'latitude'] = latitudes[i]

paises.sort\_values('PAIS\_NOM')

A table with black text

Description automatically generated

**Indicadores de Desarrollo**

WDI = pd.read\_csv("C:/Users/Patricia/Desktop/Github/TFM/DF Covid/WDIData.csv")

Nos quedamos únicamente con los años que nos interesa revisar

WDI = WDI[['Country Name', 'Country Code', 'Indicator Name', 'Indicator Code', '2018', '2019',

Como vemos que hay regiones también procedemos a eliminarlas, para esto revisamos en la tabla paises y todos aquellos registro que el WDI['Country Code'] que no estén en paises['PAIS\_ISO3'] los eliminamos

WDI = WDI[WDI['Country Code'].isin(paises['PAIS\_ISO3'].unique())]

WDI['Indicator Code'].nunique()

indicators = WDI['Indicator Name'].unique()

A screenshot of a computer

Description automatically generated

np.savetxt(f"C:/Users/Patricia/Desktop/Github/TFM/DF Covid/WDI\_indicators.txt", indicators

Después de exportarlos a txt y realizar una lectura cuidadosa declaramos los que queremos mantener, a continuación, una explicación de cada indicador cortesía de ChatGPT pero revisado manualmente

* Access to clean fuels and technologies for cooking (% of population) : Este indicador muestra el porcentaje de la población que tiene acceso a combustibles y tecnologías limpias para cocinar, como gas natural, electricidad o cocinas eficientes. Proporciona información sobre el nivel de acceso a fuentes de energía seguras y menos contaminantes para las actividades culinarias.
* Access to electricity (% of population) : Este indicador muestra el porcentaje de la población que tiene acceso a la electricidad. Sirve para evaluar el nivel de electrificación de un país y su capacidad para brindar servicios básicos a la población.
* Adequacy of social insurance programs (% of total welfare of beneficiary households) Este indicador representa el porcentaje del bienestar total de los hogares beneficiarios que es cubierto por los programas de seguro social. Mide la efectividad y la cobertura de los programas de seguro social en la protección del bienestar de los hogares.
* Adequacy of social safety net programs (% of total welfare of beneficiary households): Este indicador muestra el porcentaje del bienestar total de los hogares beneficiarios que es cubierto por los programas de redes de seguridad social. Evalúa la efectividad y la cobertura de los programas de protección social en la prevención y mitigación de la pobreza.
* Adjusted net national income (annual % growth) : Este indicador representa el crecimiento anual porcentual del ingreso nacional neto ajustado. El ingreso nacional neto ajustado tiene en cuenta los factores externos, como el agotamiento de recursos naturales y las emisiones de carbono, para proporcionar una medida más precisa del crecimiento económico sostenible.
* Adjusted net national income per capita (annual % growth) : Este indicador muestra el crecimiento anual porcentual del ingreso nacional neto ajustado per cápita. Proporciona información sobre el crecimiento económico ajustado a la población y puede indicar cambios en el nivel de vida de los ciudadanos.
* Air transport, passengers carried : Este indicador muestra la cantidad de pasajeros transportados por vía aérea en un determinado período de tiempo. Sirve como una medida del volumen y la importancia del transporte aéreo en un país o región.
* Annual freshwater withdrawals, domestic (% of total freshwater withdrawal) : Este indicador muestra el porcentaje del total de extracciones de agua dulce que se utiliza para uso doméstico en un año determinado. Ayuda a evaluar la disponibilidad y la gestión del agua potable para uso residencial en relación con el uso total de agua.
* Bank capital to assets ratio (%) : Este indicador muestra el porcentaje de los activos totales de un banco que está respaldado por capital. Proporciona una medida de la solidez financiera de un banco y su capacidad para absorber pérdidas.
* Bank liquid reserves to bank assets ratio (%) : Este indicador muestra el porcentaje de los activos totales de un banco que se mantiene como reservas líquidas, como efectivo o depósitos en bancos centrales. Sirve como indicador de la capacidad del banco para hacer frente a retiros de fondos y afrontar situaciones de estrés financiero.
* Birth rate, crude (per 1,000 people) : Este indicador muestra el número promedio de nacimientos por cada 1,000 personas en un año determinado. Proporciona información sobre la tasa de natalidad en una población y puede indicar tendencias demográficas y cambios en la estructura poblacional.
* Births attended by skilled health staff (% of total) : Este indicador muestra el porcentaje de nacimientos en los que el parto es atendido por personal de salud capacitado. Sirve como indicador de la calidad de la atención médica durante el parto y puede reflejar el acceso a servicios de salud materna adecuados.
* Cause of death, by communicable diseases and maternal, prenatal and nutrition conditions (% of total) : Este indicador muestra el porcentaje de muertes causadas por enfermedades transmisibles y condiciones relacionadas con la salud materna, prenatal y nutrición. Ayuda a evaluar la carga de enfermedades prevenibles y las condiciones de salud en una población.
* Cause of death, by injury (% of total) : Este indicador muestra el porcentaje de muertes causadas por lesiones. Proporciona información sobre la incidencia y la gravedad de diferentes tipos de lesiones y puede ayudar a orientar los esfuerzos de prevención y respuesta.
* Cause of death, by non-communicable diseases (% of total) : Este indicador muestra el porcentaje de muertes causadas por enfermedades no transmisibles, como enfermedades cardiovasculares, cáncer, diabetes y enfermedades respiratorias crónicas. Ayuda a evaluar la carga de enfermedades crónicas y la efectividad de las estrategias de prevención y control.
* Community health workers (per 1,000 people) : Este indicador muestra la cantidad de trabajadores de salud comunitarios por cada 1,000 personas. Los trabajadores de salud comunitarios son profesionales de la salud capacitados que brindan atención básica y promueven la salud en comunidades locales.
* Control of Corruption: Estimate : Este indicador representa una estimación del control de la corrupción en un país. Evalúa la percepción y la efectividad de las medidas anticorrupción, así como la transparencia y la integridad en la administración pública.
* Control of Corruption: Percentile Rank : Este indicador muestra el rango percentil en el que se encuentra un país en términos de control de la corrupción. Proporciona una comparación relativa de la situación de un país en relación con otros países en cuanto a la lucha contra la corrupción.
* Coverage of social insurance programs (% of population) : Este indicador muestra el porcentaje de la población que está cubierta por programas de seguro social. Ayuda a evaluar la efectividad y la amplitud de los programas de protección social en un país.
* Coverage of social protection and labor programs (% of population) : Este indicador muestra el porcentaje de la población que está cubierta por programas de protección social y laboral. Incluye medidas como la seguridad social, el desempleo, la capacitación laboral y otros programas destinados a proteger y apoyar a los trabajadores y sus familias.
* Coverage of social safety net programs (% of population) : Este indicador muestra el porcentaje de la población que está cubierta por programas de redes de seguridad social. Los programas de redes de seguridad social proporcionan apoyo y asistencia a las personas y familias en situaciones de vulnerabilidad económica o social.
* Current education expenditure, total (% of total expenditure in public institutions) : Este indicador muestra el porcentaje del gasto total en educación en relación con el gasto total en instituciones públicas. Proporciona información sobre la inversión en educación en relación con otros sectores y puede indicar el compromiso y la prioridad de un país con la educación.
* Current health expenditure (% of GDP) : Este indicador muestra el porcentaje del Producto Interno Bruto (PIB) de un país que se destina al gasto en salud. Ayuda a evaluar la inversión en salud en relación con el tamaño de la economía y puede reflejar el nivel de compromiso de un país con la salud de su población.
* Current health expenditure per capita, PPP (current international dls) : Este indicador muestra el gasto en salud por persona en dólares internacionales de paridad de poder adquisitivo (PPP). Proporciona una medida del gasto promedio en salud ajustado a las diferencias de poder adquisitivo entre países.
* Death rate, crude (per 1,000 people) : Este indicador muestra la tasa de mortalidad general por cada 1,000 personas en un año determinado. Proporciona información sobre el nivel de mortalidad en una población y puede indicar cambios en la salud y la calidad de vida.
* Cause of death, by communicable diseases and maternal, prenatal and nutrition conditions (% of total) : Este indicador muestra el porcentaje de muertes causadas por enfermedades transmisibles y condiciones relacionadas con la salud materna, prenatal y nutrición. Ayuda a evaluar la carga de enfermedades prevenibles y las condiciones de salud en una población.
* Cause of death, by injury (% of total) : Este indicador muestra el porcentaje de muertes causadas por lesiones. Proporciona información sobre la incidencia y la gravedad de diferentes tipos de lesiones y puede ayudar a orientar los esfuerzos de prevención y respuesta.
* Cause of death, by non-communicable diseases (% of total) : Este indicador muestra el porcentaje de muertes causadas por enfermedades no transmisibles, como enfermedades cardiovasculares, cáncer, diabetes y enfermedades respiratorias crónicas. Ayuda a evaluar la carga de enfermedades crónicas y la efectividad de las estrategias de prevención y control.
* Community health workers (per 1,000 people) : Este indicador muestra la cantidad de trabajadores de salud comunitarios por cada 1,000 personas. Los trabajadores de salud comunitarios son profesionales de la salud capacitados que brindan atención básica y promueven la salud en comunidades locales.
* Control of Corruption: Estimate : Este indicador representa una estimación del control de la corrupción en un país. Evalúa la percepción y la efectividad de las medidas anticorrupción, así como la transparencia y la integridad en la administración pública.
* Control of Corruption: Percentile Rank: Este indicador muestra el rango percentil en el que se encuentra un país en términos de control de la corrupción. Proporciona una comparación relativa de la situación de un país en relación con otros países en cuanto a la lucha contra la corrupción.
* Coverage of social insurance programs (% of population) : Este indicador muestra el porcentaje de la población que está cubierta por programas de seguro social. Ayuda a evaluar la efectividad y la amplitud de los programas de protección social en un país.
* Coverage of social protection and labor programs (% of population) : Este indicador muestra el porcentaje de la población que está cubierta por programas de protección social y laboral. Incluye medidas como la seguridad social, el desempleo, la capacitación laboral y otros programas destinados a proteger y apoyar a los trabajadores y sus familias.
* Coverage of social safety net programs (% of population) : Este indicador muestra el porcentaje de la población que está cubierta por programas de redes de seguridad social. Los programas de redes de seguridad social proporcionan apoyo y asistencia a las personas y familias en situaciones de vulnerabilidad económica o social.
* Current education expenditure, total (% of total expenditure in public institutions) : Este indicador muestra el porcentaje del gasto total en educación en relación con el gasto total en instituciones públicas. Proporciona información sobre la inversión en educación en relación con otros sectores y puede indicar el compromiso y la prioridad de un país con la educación.
* Current health expenditure (% of GDP) : Este indicador muestra el porcentaje del Producto Interno Bruto (PIB) de un país que se destina al gasto en salud. Ayuda a evaluar la inversión en salud en relación con el tamaño de la economía y puede reflejar el nivel de compromiso de un país con la salud de su población.
* Current health expenditure per capita, PPP (current international dls) : Este indicador muestra el gasto en salud por persona en dólares internacionales de paridad de poder adquisitivo (PPP). Proporciona una medida del gasto promedio en salud ajustado a las diferencias de poder adquisitivo entre países.
* Death rate, crude (per 1,000 people) : Este indicador muestra la tasa de mortalidad general por cada 1,000 personas en un año determinado. Proporciona información sobre el nivel de mortalidad en una población y puede indicar cambios en la salud y la calidad de vida.
* GNI growth (annual %) : Este indicador muestra el crecimiento anual del Ingreso Nacional Bruto (INB) de un país. El INB es la suma del PIB y los ingresos netos del exterior. El crecimiento del INB refleja la evolución de la economía y puede indicar cambios en la capacidad productiva y el nivel de ingresos de un país.
* GNI per capita growth (annual %) : Este indicador muestra el crecimiento anual del Ingreso Nacional Bruto (INB) per cápita de un país. El INB per cápita es el INB dividido por la población. El crecimiento del INB per cápita refleja el cambio porcentual en el ingreso promedio de los habitantes de un país.
* Government Effectiveness: Estimate : Este indicador representa una estimación de la eficacia del gobierno de un país. Evalúa la calidad de los servicios y las políticas gubernamentales, incluyendo la capacidad para implementar y hacer cumplir las leyes, la transparencia en la gestión y la eficiencia en la prestación de servicios públicos.
* Gross national expenditure (% of GDP) : Este indicador muestra el porcentaje del gasto nacional bruto en relación con el Producto Interno Bruto (PIB) de un país. El gasto nacional bruto incluye el consumo final, la inversión y las exportaciones netas. Proporciona información sobre la demanda interna y el nivel de actividad económica.
* Human capital index (HCI) (scale 0-1) : Este indicador es un índice que mide el capital humano en un país en una escala de 0 a 1, donde 1 representa el máximo nivel de capital humano. El capital humano se refiere al conocimiento, las habilidades y la salud de la población, y el HCI evalúa el nivel de desarrollo humano en términos de capital humano.
* Immunization, DPT (% of children ages 12-23 months) : Este indicador muestra el porcentaje de niños de 12 a 23 meses que han sido inmunizados contra la difteria, el tétanos y la tos ferina. Proporciona información sobre la cobertura de vacunación y la protección de los niños contra estas enfermedades.
* Immunization, HepB3 (% of one-year-old children) : Este indicador muestra el porcentaje de niños de un año de edad que han sido inmunizados contra la hepatitis B. Proporciona información sobre la cobertura de vacunación y la protección de los niños contra esta enfermedad.
* Immunization, measles (% of children ages 12-23 months) : Este indicador muestra el porcentaje de niños de 12 a 23 meses que han sido inmunizados contra el sarampión. Proporciona información sobre la cobertura de vacunación y la protección de los niños contra esta enfermedad.
* Incidence of HIV, all (per 1,000 uninfected population) : Este indicador muestra la incidencia del VIH en una población, medida como el número de nuevos casos de infección por VIH por cada 1,000 personas no infectadas en un año determinado. Proporciona información sobre la propagación de la infección por VIH y el impacto de la epidemia en la salud de la población.
* Incidence of malaria (per 1,000 population at risk) : Este indicador muestra la incidencia de la malaria en una población, medida como el número de nuevos casos de malaria por cada 1,000 personas en riesgo de contraer la enfermedad en un año determinado. Proporciona información sobre la carga de la malaria y la efectividad de las medidas de prevención y control.
* Incidence of tuberculosis (per 100,000 people) : Este indicador muestra la incidencia de la tuberculosis en una población, medida como el número de nuevos casos de tuberculosis por cada 100,000 personas en un año determinado. Proporciona información sobre la carga de la tuberculosis y la efectividad de los programas de control de la enfermedad.
* Increase in poverty gap at dls 1.90 ( dls 2011 PPP) poverty line due to out-of-pocket health care expenditure (% of poverty line) : Este indicador muestra el incremento en la brecha de pobreza en la línea de pobreza de dls 1.90 (en paridad de poder adquisitivo de 2011) debido a los gastos de atención médica pagados directamente por las personas. Indica el impacto de los gastos de salud en la situación de pobreza de la población.
* Increase in poverty gap at 3.20 dls (dls 2011 PPP) poverty line due to out-of-pocket health care expenditure (% of poverty line) : Este indicador muestra el incremento en la brecha de pobreza en la línea de pobreza de 3.20 dls (en paridad de poder adquisitivo de 2011) debido a los gastos de atención médica pagados directamente por las personas. Indica el impacto de los gastos de salud en la situación de pobreza de la población.
* Individuals using the Internet (% of population) : Este indicador muestra el porcentaje de la población que utiliza Internet. Proporciona información sobre la penetración de Internet en una población y refleja el acceso a las tecnologías de la información y la comunicación.
* Inflation, consumer prices (annual %) : Este indicador muestra la tasa de inflación anual, medida como el cambio porcentual en los precios de los bienes y servicios de consumo en un año determinado. Indica la variación de los precios en la economía y puede tener impacto en el poder adquisitivo de la población.
* Intentional homicides (per 100,000 people) : Este indicador muestra la tasa de homicidios intencionales por cada 100,000 personas en una población. Proporciona información sobre la seguridad y la violencia en un país y puede reflejar la situación de la delincuencia y el orden público.
* Land area (sq. km): Este indicador muestra el área total de tierra de un país medida en kilómetros cuadrados. Proporciona información sobre el tamaño geográfico de un país y puede ser relevante para evaluar su capacidad de recursos naturales y desarrollo territorial.
* Level of water stress: freshwater withdrawal as a proportion of available freshwater resources: Este indicador muestra el nivel de estrés hídrico en relación con la disponibilidad de recursos de agua dulce. Mide la proporción de agua dulce que se retira con respecto a la cantidad total disponible en una región o país. Proporciona información sobre el uso sostenible del agua y los desafíos relacionados con su escasez.
* Life expectancy at birth, female (years): Este indicador muestra la esperanza de vida al nacer para las mujeres, es decir, la cantidad promedio de años que se espera que viva una mujer al nacer en un determinado país. Proporciona una medida del estado de salud y calidad de vida de las mujeres en una población.
* Life expectancy at birth, male (years): Este indicador muestra la esperanza de vida al nacer para los hombres, es decir, la cantidad promedio de años que se espera que viva un hombre al nacer en un determinado país. Proporciona una medida del estado de salud y calidad de vida de los hombres en una población.
* Life expectancy at birth, total (years): Este indicador muestra la esperanza de vida al nacer para la población en general, es decir, la cantidad promedio de años que se espera que viva una persona al nacer en un determinado país. Proporciona una medida general del estado de salud y calidad de vida en una población.
* Literacy rate, adult total (% of people ages 15 and above): Este indicador muestra el porcentaje de la población adulta de 15 años en adelante que puede leer y escribir. Proporciona información sobre el nivel de alfabetización en una población y puede ser un indicador del nivel educativo y el acceso a oportunidades.
* Mobile cellular subscriptions (per 100 people): Este indicador muestra el número de suscripciones de telefonía móvil por cada 100 personas en un país. Proporciona información sobre la penetración de los servicios de telefonía móvil y la disponibilidad de comunicación en una población.
* Mortality rate attributed to unsafe water, unsafe sanitation and lack of hygiene (per 100,000 population): Este indicador muestra la tasa de mortalidad atribuida al agua insalubre, saneamiento deficiente y falta de higiene. Mide el número de muertes relacionadas con la falta de acceso a agua potable segura, saneamiento adecuado e higiene. Proporciona información sobre los riesgos para la salud asociados a estas condiciones y la necesidad de mejorar la infraestructura y las prácticas sanitarias.
* Multidimensional poverty headcount ratio (% of total population): Este indicador muestra el porcentaje de la población que se encuentra en situación de pobreza multidimensional. La pobreza multidimensional considera varios aspectos, como el acceso a servicios básicos, la educación, la salud y el nivel de vida. Proporciona una medida más completa de la pobreza que solo considerar el ingreso.
* Multidimensional poverty index (scale 0-1): Este indicador muestra el índice de pobreza multidimensional en una escala del 0 al 1, donde 0 representa la ausencia total de pobreza multidimensional y 1 representa la pobreza multidimensional más extrema. El índice combina diferentes dimensiones de la pobreza para proporcionar una medida compuesta de la pobreza en un país.
* Multidimensional poverty intensity (average share of deprivations experienced by the poor): Este indicador muestra la intensidad de la pobreza multidimensional, es decir, la cantidad promedio de privaciones experimentadas por las personas en situación de pobreza. Proporciona información sobre la gravedad y la profundidad de la pobreza multidimensional en un país.
* Number of people pushed below the 3.65 dls (dls 2017 PPP) poverty line by out-of-pocket health care expenditure: Este indicador muestra el número de personas que han caído por debajo de la línea de pobreza de dls 3.65 (en paridad de poder adquisitivo de 2017) debido a los gastos de atención médica pagados directamente por ellas. Indica el impacto de los gastos de salud en la situación económica de la población.
* Number of people spending more than 10% of household consumption or income on out-of-pocket health care expenditure: Este indicador muestra el número de personas que gastan más del 10% del consumo o ingreso familiar en gastos de atención médica pagados directamente por ellas. Proporciona información sobre la carga económica de los gastos de salud en los hogares y la capacidad de afrontar dichos gastos.
* Number of people spending more than 25% of household consumption or income on out-of-pocket health care expenditure: Este indicador muestra el número de personas que gastan más del 25% del consumo o ingreso familiar en gastos de atención médica pagados directamente por ellas. Proporciona información sobre la carga económica significativa de los gastos de salud en los hogares y su impacto en el bienestar financiero de las familias.
* Number of surgical procedures (per 100,000 population): Este indicador muestra el número de procedimientos quirúrgicos realizados por cada 100,000 personas en una población. Proporciona información sobre la disponibilidad y acceso a servicios quirúrgicos en un país, lo cual puede ser indicativo de la infraestructura de salud y la capacidad de brindar atención médica.
* Nurses and midwives (per 1,000 people): Este indicador muestra el número de enfermeras y parteras por cada 1,000 personas en una población. Proporciona información sobre la disponibilidad de profesionales de enfermería y partería, que desempeñan un papel crucial en la prestación de servicios de salud y atención materna e infantil.
* People practicing open defecation (% of population): Este indicador muestra el porcentaje de la población que practica la defecación al aire libre, es decir, la falta de acceso a servicios de saneamiento adecuados. Proporciona información sobre las condiciones de higiene y la falta de infraestructura sanitaria en una población.
* People using at least basic drinking water services (% of population): Este indicador muestra el porcentaje de la población que utiliza al menos servicios básicos de agua potable. Mide la proporción de personas que tienen acceso a una fuente mejorada de agua potable, lo cual es esencial para la salud y el bienestar de la población.
* People using at least basic sanitation services (% of population): Este indicador muestra el porcentaje de la población que utiliza al menos servicios básicos de saneamiento. Mide la proporción de personas que tienen acceso a instalaciones de saneamiento mejoradas, lo cual es fundamental para la salud pública y la prevención de enfermedades.
* People using safely managed drinking water services (% of population): Este indicador muestra el porcentaje de la población que utiliza servicios de agua potable gestionados de manera segura. Mide la proporción de personas que tienen acceso a una fuente de agua potable segura, lo cual implica que la fuente está protegida contra la contaminación y garantiza una calidad adecuada del agua.
* People using safely managed sanitation services (% of population): Este indicador muestra el porcentaje de la población que utiliza servicios de saneamiento gestionados de manera segura. Mide la proporción de personas que tienen acceso a instalaciones de saneamiento que están diseñadas de manera segura, protegidas contra la contaminación y aseguran una disposición adecuada de los desechos.
* People with basic handwashing facilities including soap and water (% of population): Este indicador muestra el porcentaje de la población que cuenta con instalaciones básicas para el lavado de manos, incluyendo jabón y agua. Proporciona información sobre la disponibilidad de recursos para la higiene personal, lo cual es fundamental para prevenir enfermedades y promover la salud.
* Political Stability and Absence of Violence/Terrorism: Estimate: Este indicador muestra una estimación de la estabilidad política y la ausencia de violencia/terrorismo en un país. Evalúa la percepción y el nivel de seguridad en términos de estabilidad política y la presencia o ausencia de actos violentos o terroristas.
* Political Stability and Absence of Violence/Terrorism: Percentile Rank: Este indicador muestra el ranking de un país en términos de estabilidad política y la ausencia de violencia/terrorismo en comparación con otros países. Proporciona una medida relativa de la estabilidad política y la seguridad en un país.
* Population ages 65 and above, total: Este indicador muestra el número total de personas en una población que tienen 65 años o más. Proporciona información sobre la estructura demográfica y la distribución de edades en una población.
* Population density (people per sq. km of land area): Este indicador muestra la densidad de población, es decir, el número de personas por kilómetro cuadrado de área terrestre. Proporciona información sobre la concentración y distribución de la población en un área geográfica determinada.
* Population growth (annual %): Este indicador muestra la tasa de crecimiento anual de la población. Indica el cambio porcentual en el tamaño de la población durante un año específico y refleja la dinámica de crecimiento demográfico de un país.
* Population in largest city: Este indicador muestra el número de personas que viven en la ciudad más grande de un país. Proporciona información sobre la concentración urbana y la importancia relativa de una ciudad en términos de población.
* Population in the largest city (% of urban population): Este indicador muestra el porcentaje de la población urbana que vive en la ciudad más grande de un país. Proporciona información sobre la proporción de la población urbana que se concentra en la ciudad más grande en comparación con otras áreas urbanas.
* Population in urban agglomerations of more than 1 million (% of total population): Este indicador muestra el porcentaje de la población total que vive en aglomeraciones urbanas con más de 1 millón de habitantes. Proporciona información sobre la urbanización y la concentración de la población en áreas urbanas densamente pobladas.
* Population living in slums (% of urban population): Este indicador muestra el porcentaje de la población urbana que vive en asentamientos informales (slums). Proporciona información sobre las condiciones de vivienda precarias y la falta de servicios básicos en los asentamientos urbanos más pobres.
* Population, female: Este indicador muestra el número total de mujeres en una población. Proporciona información sobre la estructura demográfica y la distribución de género en una población.
* Population, male: Este indicador muestra el número total de hombres en una población. Proporciona información sobre la estructura demográfica y la distribución de género en una población.
* Population, total: Este indicador muestra el número total de personas en una población. Es una medida fundamental para comprender el tamaño de una población y su evolución a lo largo del tiempo.
* Poverty headcount ratio at national poverty lines (% of population): Este indicador muestra el porcentaje de la población que vive por debajo de la línea de pobreza nacional. Proporciona información sobre la incidencia de la pobreza en un país y la proporción de personas que viven en condiciones de pobreza.
* Prevalence of current tobacco use (% of adults): Este indicador muestra el porcentaje de adultos que son consumidores actuales de tabaco. Proporciona información sobre la prevalencia del consumo de tabaco en una población y su impacto en la salud pública.
* Prevalence of moderate or severe food insecurity in the population (%): Este indicador muestra el porcentaje de la población que experimenta inseguridad alimentaria moderada o grave. Mide la falta de acceso regular y adecuado a alimentos suficientes y nutricionalmente adecuados en una población.
* Prevalence of severe food insecurity in the population (%): Este indicador muestra el porcentaje de la población que experimenta inseguridad alimentaria grave. Mide la falta de acceso regular y adecuado a alimentos suficientes y nutricionalmente adecuados en una población, en su forma más severa.
* Prevalence of undernourishment (% of population): Este indicador muestra el porcentaje de la población que padece desnutrición. Mide la proporción de personas que no reciben suficientes nutrientes para satisfacer sus necesidades alimentarias diarias.
* Proportion of population pushed below the dls 3.65 (dls 2017 PPP) poverty line by out-of-pocket health care expenditure (%): Este indicador muestra el porcentaje de la población que cae por debajo de la línea de pobreza de dls 3.65 debido a los gastos de atención médica pagados de su propio bolsillo. Proporciona información sobre el impacto económico de los gastos de salud en la población.
* Proportion of population spending more than 10% of household consumption or income on out-of-pocket health care expenditure (%): Este indicador muestra el porcentaje de la población que gasta más del 10% de su consumo o ingreso familiar en gastos de atención médica pagados de su propio bolsillo. Mide el impacto económico de los gastos de salud en la población.
* Railways, passengers carried (million passenger-km): Este indicador muestra el número total de pasajeros transportados por ferrocarril en millones de pasajeros-kilómetro. Proporciona información sobre la utilización y la importancia relativa del transporte ferroviario en un país.
* Real interest rate (%): Este indicador muestra la tasa de interés real, es decir, la tasa de interés ajustada por inflación. Refleja el rendimiento real de las inversiones y el costo real del endeudamiento, teniendo en cuenta el efecto de la inflación.
* Risk of catastrophic expenditure for surgical care (% of people at risk): Este indicador muestra el riesgo de gastos catastróficos debido a la atención quirúrgica, como porcentaje de las personas en riesgo. Mide la proporción de personas que enfrentan un alto nivel de gastos en atención quirúrgica en relación con sus ingresos o capacidad de pago.
* Risk of impoverishing expenditure for surgical care (% of people at risk): Este indicador muestra el riesgo de empobrecimiento debido a los gastos en atención quirúrgica, como porcentaje de las personas en riesgo. Mide la proporción de personas cuyos gastos en atención quirúrgica los empujan por debajo de la línea de pobreza o reducen su capacidad adquisitiva.
* Rural population (% of total population): Este indicador muestra el porcentaje de la población que vive en áreas rurales en relación con la población total. Proporciona información sobre la distribución de la población entre áreas rurales y urbanas en un país.
* Survival to age 65, female (% of cohort): Este indicador muestra el porcentaje de mujeres de una cohorte que sobreviven hasta los 65 años de edad. Proporciona información sobre la esperanza de vida y la salud de las mujeres en un país.
* Survival to age 65, male (% of cohort): Este indicador muestra el porcentaje de hombres de una cohorte que sobreviven hasta los 65 años de edad. Proporciona información sobre la esperanza de vida y la salud de los hombres en un país.
* Total alcohol consumption per capita (liters of pure alcohol, projected estimates, 15+ years of age): Este indicador muestra el consumo total de alcohol per cápita en litros de alcohol puro, proyectado para personas de 15 años en adelante. Proporciona información sobre los niveles de consumo de alcohol en una población.
* Tuberculosis case detection rate (% of all forms): Este indicador muestra el porcentaje de casos de tuberculosis detectados en relación con todos los casos de tuberculosis, incluyendo todas las formas de la enfermedad. Mide la efectividad de los sistemas de detección y diagnóstico de la tuberculosis en un país.
* Tuberculosis treatment success rate (% of new cases): Este indicador muestra el porcentaje de casos nuevos de tuberculosis que se tratan exitosamente. Mide la eficacia del tratamiento de la tuberculosis en un país y su capacidad para curar la enfermedad.
* UHC service coverage index: Este indicador muestra el índice de cobertura de servicios de salud universal (UHC, por sus siglas en inglés). Mide el grado de cobertura de servicios de salud esenciales y accesibles para toda la población, sin importar su capacidad de pago.
* Urban population (% of total population): Este indicador muestra el porcentaje de la población que vive en áreas urbanas en relación con la población total. Proporciona información sobre la distribución de la población entre áreas urbanas y rurales en un país.

indicators\_to\_keep = [

'EG.CFT.ACCS.ZS', 'EG.ELC.ACCS.ZS', 'per\_si\_allsi.adq\_pop\_tot',

'per\_sa\_allsa.adq\_pop\_tot', 'NY.ADJ.NNTY.KD.ZG', 'NY.ADJ.NNTY.PC.KD.ZG',

'IS.AIR.PSGR', 'ER.H2O.FWDM.ZS', 'FB.BNK.CAPA.ZS',

'FD.RES.LIQU.AS.ZS', 'SP.DYN.CBRT.IN', 'SH.STA.BRTC.ZS',

'SH.DTH.COMM.ZS', 'SH.DTH.INJR.ZS', 'SH.DTH.NCOM.ZS',

'SH.MED.CMHW.P3', 'CC.EST', 'CC.PER.RNK',

'per\_si\_allsi.cov\_pop\_tot', 'per\_allsp.cov\_pop\_tot', 'per\_sa\_allsa.cov\_pop\_tot',

'SE.XPD.CTOT.ZS', 'SH.XPD.CHEX.GD.ZS', 'SH.XPD.CHEX.PP.CD',

'SP.DYN.CDRT.IN', 'SH.STA.DIAB.ZS', 'EN.CLC.DRSK.XQ',

'SH.XPD.GHED.CH.ZS', 'SH.XPD.GHED.GD.ZS', 'SH.XPD.GHED.GE.ZS',

'SH.XPD.PVTD.CH.ZS', 'SE.PRM.CUAT.ZS', 'SH.XPD.EHEX.CH.ZS',

'IT.NET.BBND.P2', 'IT.MLT.MAIN.P2', 'AG.LND.FRST.ZS',

'NY.GDP.MKTP.CD', 'NY.GDP.DEFL.ZS', 'NY.GDP.DEFL.ZS.AD',

'NY.GDP.MKTP.KD.ZG', 'NY.GDP.PCAP.CD', 'NE.CON.GOVT.ZS',

'SI.POV.GINI', 'NY.GNP.MKTP.KD.ZG', 'NY.GNP.PCAP.KD.ZG',

'NE.DAB.TOTL.ZS', 'HD.HCI.OVRL', 'SH.IMM.IDPT',

'SH.IMM.HEPB', 'SH.IMM.MEAS', 'SH.HIV.INCD.TL.P3',

'SH.MLR.INCD.P3', 'SH.TBS.INCD', 'SH.UHC.NOP1.ZG',

'SH.UHC.NOP2.ZG', 'IT.NET.USER.ZS', 'FP.CPI.TOTL.ZG',

'VC.IHR.PSRC.P5', 'AG.LND.TOTL.K2', 'ER.H2O.FWST.ZS',

'SP.DYN.LE00.MA.IN', 'SP.DYN.LE00.FE.IN', 'SP.DYN.LE00.IN',

'SE.ADT.LITR.ZS', 'IT.CEL.SETS.P2', 'SH.STA.WASH.P5',

'I.POV.MDIM', 'SI.POV.MDIM.XQ', 'SI.POV.MDIM.IT',

'SH.UHC.NOP2.TO', 'SH.UHC.OOPC.10.TO', 'SH.UHC.OOPC.25.TO',

'SH.SGR.PROC.P5', 'SH.MED.NUMW.P3', 'SH.STA.ODFC.ZS',

'SH.H2O.BASW.ZS', 'SH.STA.BASS.ZS', 'SH.H2O.SMDW.ZS',

'SH.STA.SMSS.ZS', 'SH.STA.HYGN.ZS', 'PV.EST',

'PV.PER.RNK', 'SP.POP.65UP.TO', 'EN.POP.DNST',

'SP.POP.GROW', 'EN.URB.LCTY', 'EN.URB.LCTY.UR.ZS',

'EN.URB.MCTY.TL.ZS', 'EN.POP.SLUM.UR.ZS', 'SP.POP.TOTL.FE.IN',

'SP.POP.TOTL.MA.IN', 'SP.POP.TOTL', 'SI.POV.NAHCSH.PRV.SMOK',

'SN.ITK.MSFI.ZS', 'SN.ITK.SVFI.ZS', 'SN.ITK.DEFC.ZS',

'SH.UHC.NOP2.ZS', 'SH.UHC.OOPC.10.ZS', 'IS.RRS.PASG.KM',

'FR.INR.RINR', 'SH.SGR.CRSK.ZS', 'SH.SGR.IRSK.ZS',

'SP.RUR.TOTL.ZS', 'SP.DYN.TO65.FE.ZS', 'SP.DYN.TO65.MA.ZS',

'SH.ALC.PCAP.LI', 'SH.TBS.DTEC.ZS', 'SH.TBS.CURE.ZS',

'SH.UHC.SRVS.CV.XD', 'SP.URB.TOTL.IN.ZS'

]

WDI = WDI[['Country Name', 'Country Code', 'Indicator Name', 'Indicator Code', '2018', '2019', '2020', '2021', '2022']]

WDI = WDI[WDI['Indicator Code'].isin(indicators\_to\_keep)]

WDI['Indicator Code'].nunique()

Por tanto, habríamos podido pasar de 1478 a 108, es importante tener en cuenta que estos indicadores son una preselección de cuestiones que podrían relacionarse con nuestro foco de investigación

Al final habremos obtenido una tabla con los datos de 2018 a 2022 de Los Indicadores de Desarrollo Mundial (WDI) del Worl Bank por países según su ISO3.

Obtenido de: World Bank. (Revisado en julio de 2023). World Development Indicators. Recuperado de: <https://datacatalog.worldbank.org/search/dataset/0037712/World-Development-Indicators>

Sólo nos queda obtener el ISO2

# Agregar la columna 'PAIS\_ISO2' a 'WDI' utilizando el mapeo

WDI['PAIS\_ISO2'] = WDI['Country Code'].map(paises.set\_index('PAIS\_ISO3')['PAIS\_ISO2'].to\_dict())

WDI

A screenshot of a computer

Description automatically generated

Dataset obtenido de: <https://www.kaggle.com/datasets/berkeleyearth/climate-change-earth-surface-temperature-data>

nRowsRead = None

temp = pd.read\_csv('C:/Users/Patricia/Desktop/Github/TFM/DF Covid/Temperatura/GlobalLandTemperaturesByCountry.csv', delimiter=',', nrows = nRowsRead)

temp.dataframeName = 'GlobalLandTemperaturesByCountry.csv'

nRow, nCol = temp.shape

print(f'There are {nRow} rows and {nCol} columns')



Debido a que es un dataset enorme nos quedamos sólo con el 2000 en adelante

temp = temp[temp['dt'].str.startswith('20')]

temp.loc[:, 'dt'] = pd.to\_datetime(temp['dt'], format='%Y-%m-%d')

temp.loc[:, 'DIA\_DEL\_AÑO'] = temp['dt'].dt.dayofyear

temp.loc[:, 'MES'] = temp['dt'].dt.month

temp.loc[:, 'AÑO'] = temp['dt'].dt.year

temp.sort\_values(['AÑO', 'DIA\_DEL\_AÑO'], inplace=True)

temp.drop('dt', axis=1, inplace=True)

Sólo nos faltaría obtener el ISO2

# Agregar la columna 'PAIS\_ISO2' a 'WDI' utilizando el mapeo

temp['PAIS\_ISO2'] = temp['Country'].map(paises.set\_index('PAIS\_NOM')['PAIS\_ISO2'].to\_dict())

Cómo vemos hay distintos registros para un mismo país y mes, tenemos que hacer la media entre ellos para que tengamos el registro medio de cada país en cada mes

temperatura = temp.groupby(['PAIS\_ISO2', 'MES'])['AverageTemperature'].mean().reset\_index()

y habremos obtenido un dataset con la temperatura media mensual de los países según su ISO2

Es importante tener en cuenta que estos datos no entran lo suficiente en detalle para tener una idea real del clima en un país pues nos da una media para todo el territorio sin tomar en cuenta fenómenos o variaciones geográficas, pero para el propósito de saber si en invierno o verano nos funcionarán

ANÁLISIS INDIVIDUAL POR DATASET

DATA\_FRAME\_COVID\_DAILY\_OMS GLOBAL

coviddaily = coviddaily.drop("Unnamed: 0", axis=1) # Se elimina la fila unnamed, esta fila correspond

coviddaily['FECHA\_INFORMADA'] = pd.to\_datetime(coviddaily['FECHA\_INFORMADA'], format='%d-%m-%Y')

coviddaily['DIA\_DEL\_AÑO'] = coviddaily['FECHA\_INFORMADA'].dt.dayofyear

coviddaily['MES'] = coviddaily['FECHA\_INFORMADA'].dt.month

coviddaily['AÑO'] = coviddaily['FECHA\_INFORMADA'].dt.year

coviddaily.sort\_values(['AÑO', 'DIA\_DEL\_AÑO'], inplace=True)

Graficamos los casos y muertes nuevas

plt.figure(figsize=(10, 6)) # Tamaño de la figura

# Gráfico de casos nuevos

sns.lineplot(x='FECHA\_INFORMADA', y='CASOS\_ACUM', data=coviddaily)

plt.xlabel('Fecha') # Etiqueta del eje x

plt.ylabel('Casos Nuevos') # Etiqueta del eje y

plt.title('Evolución de Casos Nuevos') # Título del gráfico

plt.xticks(rotation=45) # Rota las etiquetas del eje x para una mejor legibilidad

plt.grid(True) # Muestra la cuadrícula

plt.show() # Muestra el gráfico

A graph with a line

Description automatically generated

plt.figure(figsize=(10, 6)) # Tamaño de la figura

# Gráfico de muertes nuevas

sns.lineplot(x='FECHA\_INFORMADA', y='CASOS\_ACUM', data=coviddaily)

plt.xlabel('Fecha') # Etiqueta del eje x

plt.ylabel('Muertes Nuevas') # Etiqueta del eje y

plt.title('Evolución de Muertes Nuevas') # Título del gráfico

plt.xticks(rotation=45) # Rota las etiquetas del eje x para una mejor legibilidad

plt.grid(True) # Muestra la cuadrícula

plt.show() # Muestra el gráfico

A graph with a line

Description automatically generated

# Agrupar los datos por país y calcular la suma de casos y muertes acumulados

grouped = coviddaily.groupby('PAIS\_ISO3').agg({'CASOS\_ACUM': 'sum', 'MUERTES\_ACUM': 'sum'})

# Obtener los nombres de los países y los valores de casos y muertes acumulados

paises = grouped.index

cases = grouped['CASOS\_ACUM']

deaths = grouped['MUERTES\_ACUM']

# Configurar el tamaño de la figura y los ejes

fig, ax = plt.subplots(figsize=(10, 30))

# Crear el gráfico de barras horizontales para casos acumulados

ax.barh(paises, cases, label='Cases - Cumulative Total')

# Crear el gráfico de barras horizontales para muertes acumuladas

ax.barh(paises, deaths, label='Deaths - Cumulative Total')

# Configurar las etiquetas de los ejes y el título del gráfico

ax.set\_xlabel('Count')

ax.set\_ylabel('PAIS\_ISO3')

ax.set\_title('Cases and Deaths - Cumulative Total by Country')

# Mostrar una leyenda

ax.legend()

# Ajustar el espacio entre las barras

plt.tight\_layout()

# Mostrar el gráfico de barras

plt.show()

A graph with lines on it

Description automatically generated

A graph with lines on it

Description automatically generated

Para calcular la correlación entre país y casos acumulados usaríamos el coeficiente de correlación de Spearman

# Calcular el coeficiente de correlación de Spearman

correlacion\_spearman, p\_valor = stats.spearmanr(coviddaily['PAIS'], coviddaily['CASOS\_ACUM'])

# Imprimir el coeficiente de correlación y el valor p

print('Coeficiente de correlación de Spearman:', correlacion\_spearman)

print('Valor p:', p\_valor)

Coeficiente de correlación de Spearman: -0.05202680413632658

Valor p: 7.49643620172126e-180

La correlación de Spearman mide la relación monotónica entre dos variables. En este caso, el coeficiente de correlación de -0.052 indica una correlación débil y negativa entre la variable categórica (PAIS) y la variable numérica (CASOS\_ACUM). Esto sugiere que no hay una relación lineal clara entre el país y el número acumulado de casos de COVID-19.

El valor p extremadamente bajo (7.496e-180) indica una fuerte evidencia en contra de la hipótesis nula de no correlación. En otras palabras, es altamente improbable que la correlación observada se deba al azar. Sin embargo, debido a que la correlación es muy cercana a cero, su relevancia práctica puede ser limitada

COVID DAILY US

coviddaily\_us = pd.read\_csv("C:/Users/Patricia/Desktop/Github/TFM/DF Covid/df\_daily\_report\_us\_final.csv")

coviddaily\_us['Date'] = pd.to\_datetime(coviddaily\_us['Date'], format='%d.%m.%Y')

coviddaily\_us['DIA\_DEL\_AÑO'] = coviddaily\_us['Date'].dt.dayofyear

coviddaily\_us['MES'] = coviddaily\_us['Date'].dt.month

coviddaily\_us['AÑO'] = coviddaily\_us['Date'].dt.year

coviddaily\_us.sort\_values(['AÑO', 'DIA\_DEL\_AÑO'], inplace=True)

plt.figure(figsize=(10, 6)) # Tamaño de la figura

# Gráfico de casos nuevos

sns.lineplot(x='Date', y='Confirmed', data=coviddaily\_us)

plt.xlabel('Fecha') # Etiqueta del eje x

plt.ylabel('Casos confirmados') # Etiqueta del eje y

plt.title('Evolución de Casos') # Título del gráfico

plt.xticks(rotation=45) # Rota las etiquetas del eje x para una mejor legibilidad

plt.grid(True) # Muestra la cuadrícula

plt.show() # Muestra el gráfico

A graph with a line

Description automatically generated

plt.figure(figsize=(10, 6)) # Tamaño de la figura

# Gráfico de muertes nuevas

sns.lineplot(x='Date', y='Deaths', data=coviddaily\_us)

plt.xlabel('Fecha') # Etiqueta del eje x

plt.ylabel('Muertes') # Etiqueta del eje y

plt.title('Evolución de Muertes') # Título del gráfico

plt.xticks(rotation=45) # Rota las etiquetas del eje x para una mejor legibilidad

plt.grid(True) # Muestra la cuadrícula

plt.show() # Muestra el gráfico

A graph with a line

Description automatically generated

**DATOS UCI**

uci\_hosp = pd.read\_csv("C:/Users/Patricia/Desktop/Github/TFM/DFCovid/df\_daily\_report\_us\_final.csv")

uci\_hosp.isna().sum()

Unnamed: 0 0

Province\_State 0

Confirmed 0

Deaths 0

Recovered 46492

Active 46492

ISO3 0

Date 9860

dtype: int64

No contamos con la columna Recovered pero deberían ser aquellos casos confirmados que no han fallecido

# Reemplazar los NaN en "Recovered" con los valores de "Confirmed" que no figuran como "Deaths"

uci\_hosp['Recovered'] = uci\_hosp['Recovered'].fillna(uci\_hosp['Confirmed'] - uci\_hosp['Deaths'])

# Convertir la columna "Date" a tipo de datos datetime

uci\_hosp['Date'] = pd.to\_datetime(uci\_hosp['Date'], format='%d.%m.%Y')

Eliminamos aquellos sin fecha pues no son registros que podamos situar

uci\_hosp.dropna(subset=['Date'], inplace=True)

Organizamos las fechas

uci\_hosp['DIA\_DEL\_AÑO'] = uci\_hosp['Date'].dt.dayofyear

uci\_hosp['MES'] = uci\_hosp['Date'].dt.month

uci\_hosp['AÑO'] = uci\_hosp['Date'].dt.year

uci\_hosp.sort\_values(['AÑO', 'DIA\_DEL\_AÑO'], inplace=True)

Graficamos los casos y muertes confirmadas

plt.figure(figsize=(10, 6)) # Tamaño de la figura

# Gráfico de casos

sns.lineplot(x='Date', y='Confirmed', data=uci\_hosp, color='blue', label='Confirmed')

sns.lineplot(x='Date', y='Deaths', data=uci\_hosp, color='red', label='Deaths')

plt.xlabel('Fecha') # Etiqueta del eje x

plt.ylabel('Casos') # Etiqueta del eje y

plt.title('Evolución de Casos') # Título del gráfico

plt.xticks(rotation=45) # Rota las etiquetas del eje x para una mejor legibilidad

plt.grid(True) # Muestra la cuadrícula

plt.legend() # Muestra la leyenda con los nombres de las líneas

plt.show() # Muestra el gráfico

A graph with a line

Description automatically generated

COVID TESTING AT US

testing = pd.read\_csv("C:/Users/Patricia/Desktop/Github/TFM/DF Covid/df\_testing\_covid\_eu.csv")

# Corr por defecto será Pearson que es el más utilizado

print("Coeficiente de correlación de Pearson:", testing['RATIO\_TESTS'].corr(testing['RATIO\_POSITIVO']))

Coeficiente de correlación de Pearson: -0.18767002912753783

Por tanto hablaríamos de una correlación negativa debil entre testing\_rate y positivity\_rate

# Jointgrid con histograma marginal Testing\_rate x positivity\_rate

g = sns.jointplot(data=testing, x='RATIO\_TESTS', y='RATIO\_POSITIVO', kind="kde", cmap="flare", fill=True)

g.plot\_joint(sns.kdeplot, cmap="mako", fill=True)

g.plot\_marginals(sns.histplot, color='#3D3164')

# Configurar etiquetes i títol

g.set\_axis\_labels("Testing rate", "Positivity rate")

plt.show()

A graph of a test

Description automatically generated

Conteo de representación de regiones

# Calcular el conteo de registros por región

region\_counts = testing['PAIS\_ISO3'].value\_counts()

# Crear la visualización del conteo de registros por región

plt.figure(figsize=(12, 6)) # Ajustar el tamaño de la figura

sns.countplot(data=testing, x='PAIS\_ISO3', palette='mako', order=region\_counts.index)

plt.xlabel('Región')

plt.ylabel('Contador')

plt.title('Representación de cada región en el dataset')

plt.xticks(rotation=90) # Rotar las etiquetas de las regiones

plt.tight\_layout() # Ajustar el espaciado

plt.show()

A graph of a number of people

Description automatically generated

Cluster maps de correlaciones

data = testing[['CASOS\_NUEVOS','N\_TESTS\_REALIZADOS', 'POBLACION', 'RATIO\_TESTS', 'RATIO\_POSITIVO']]

sns.clustermap(data.corr(), annot=True, cmap='mako')

plt.title('Clustermap - Correlación entre variables')

plt.show()

A screenshot of a computer screen

Description automatically generated

Cases and deaths

casesanddeath = pd.read\_csv("C:/Users/Patricia/Desktop/Github/TFM/DF Covid/df\_Latest reported counts of cases and deaths.csv")

En este no hemos realizado gráficos pues creemos que su potencial estaría en poder realizar un mapa coroplético.

Vaccination global

vacc = pd.read\_csv("C:/Users/Patricia/Desktop/Github/TFM/DF Covid/df\_vacunation.csv")

En este también lo idóneo sería hacer cruce con cases and deaths pero a realizar en la próxima entrega

Vaccines type

vacctype = pd.read\_csv("C:/Users/Patricia/Desktop/Github/TFM/DF Covid/df\_vacunation\_tipo.csv")

vacctype['NOMBRE\_VACUNA'].nunique()

38

Vemos que el dataset recoge 38 vacunas diferentes

# Contar el número de vacunas autorizadas y reordenar el dataset

vacunas\_por\_pais = vacctype.groupby('PAIS\_ISO3')['NOMBRE\_VACUNA'].nunique().reset\_index().sort\_values(by='NOMBRE\_VACUNA', ascending=False)

# Realizar merge con el dataframe de países para obtener el nombre completo del país

vacunas\_por\_pais = vacunas\_por\_pais.merge(paises[['PAIS\_ISO3', 'PAIS\_NOM']], left\_on='PAIS\_ISO3', right\_on='PAIS\_ISO3', how='left')

# Configurar estilo de Seaborn

sns.set(style="whitegrid")

# Crear el gráfico de barras horizontales con paleta 'mako'

plt.figure(figsize=(12, 45))

sns.barplot(data=vacunas\_por\_pais, x='NOMBRE\_VACUNA', y='PAIS\_NOM', palette='mako')

plt.xlabel('Número de vacunas aprobadas')

plt.ylabel('País')

plt.title('Número de vacunas aprobadas por país')

plt.xticks(rotation=0)

# Mostrar el gráfico

plt.show()

A graph of a bar chart

Description automatically generated

A graph of a bar chart

Description automatically generated

Podemos ver que los países que figuran con más vacunas aprobadas son Iran, Portugal, Bahrain, Somalia y Sudán

Los países con menor cantidad de vacunas aprobadas son Burundi, Rusia, Islas Pitcairn, Mali y Tuvalu que habrían aprobado sólo una

**ANÁLISIS COMBINADO**

Hipótesis

* Hipótesis de investigación (H1): Los países en los que hace más frío tienen una mayor incidencia de Covid-19.
  + Hipótesis nula (H01): No hay relación entre la temperatura promedio de un país y la incidencia de Covid-19.
* Hipótesis de investigación (H2): En la época más fría del año aumentan los casos de Covid-19.
  + Hipótesis nula (H02): No hay diferencia en la cantidad de casos de Covid-19 entre la época más fría y otras estaciones del año.
* Hipótesis de investigación (H3): Hay mayor mortalidad por Covid-19 en países sin salud universal.
  + Hipótesis nula (H03): No hay diferencia en la tasa de mortalidad por Covid-19 entre países con y sin salud universal.
* Hipótesis de investigación (H4): En las zonas urbanas hay mayor incidencia de Covid-19.
  + Hipótesis nula (H04): No hay diferencia en la incidencia de Covid-19 entre zonas urbanas y zonas no urbanas.
* Hipótesis de investigación (H5): El acceso a las medidas higiénicas afecta directamente en la transmisión de Covid-19.
  + Hipótesis nula (H05): No hay relación entre el acceso a las medidas higiénicas y la transmisión de Covid-19.
* Hipótesis de investigación (H6): La vacunación ha reducido la mortalidad del Covid-19.
  + Hipótesis nula (H06): La vacunación no ha reducido la mortalidad del Covid-19.

Estas hipótesis nos permiten formular suposiciones basadas en las relaciones esperadas entre variables relevantes y la incidencia, propagación y consecuencias del Covid-19. Las hipótesis nulas representan la ausencia de dichas relaciones. A través de análisis de datos y pruebas estadísticas, podremos evaluar la evidencia en apoyo o en contra de estas hipótesis, lo que contribuirá a nuestra comprensión y conocimiento sobre la pandemia de Covid-19.

Es importante destacar que las hipótesis deben ser probadas y evaluadas de manera rigurosa, utilizando métodos adecuados de análisis estadístico y teniendo en cuenta otros factores que puedan influir en los resultados. Asimismo, se respaldarán las afirmaciones con bibliografía académica especializada.