Entrega Final

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

31 de octubre de 2023

Resumen Ejecutivo

El presente documento responde a la cuarta y última entrega del TFM – Covid-19, en el cual se sintetiza lo realizado a la fecha.

En la primera etapa de nuestra investigación realizamos un análisis de mapas de empatía, el enfoque post-Motorola y el mapeo de alineación del equipo. Durante este proceso, identificamos los intereses del usuario, que incluían temas como el COVID-19 persistente, tratamientos y factores de riesgo. Aunque estos temas resultaron muy interesantes, tuvimos que adaptarnos a los recursos disponibles, manteniendo al mismo tiempo nuestra atención en los intereses específicos de nuestro público objetivo, en particular, la relación entre la vacunación contra el COVID-19 y la mortalidad.

En una etapa posterior, ampliamos nuestro enfoque para explorar la relación entre los datos relacionados con la pandemia de COVID-19 y diversos indicadores de desarrollo del Banco Mundial. Nos centramos en variables clave, como el acceso a Internet, servicios básicos de agua potable y saneamiento, densidad de población, proporción de población rural y urbana, y el Producto Interno Bruto (PIB).

Además de estos indicadores de desarrollo, también consideramos factores geográficos y meteorológicos, como la altitud promedio sobre el nivel del mar. Incorporamos estos datos complementarios en nuestro análisis para lograr una comprensión más completa de cómo la pandemia de COVID-19 se relaciona con aspectos sociodemográficos, geográficos y económicos.

Para recopilar información esencial relacionada con la pandemia de COVID-19, utilizamos diversas fuentes de datos de acceso abierto. **La Organización Mundial de la Salud (OMS)** proporcionó detalles sobre los casos diarios y las muertes a nivel mundial [1]. Además, otra fuente de la OMS ofreció información adicional sobre los casos confirmados, las muertes y las tasas por cada 100,000 habitantes [2]. Para evaluar el progreso de la vacunación, utilizamos datos de la OMS que incluyen la administración de dosis y el número de personas vacunadas parcial o completamente, junto con información sobre los tipos de vacunas utilizadas en cada país [3] [4].

El Centro Europeo para la Prevención y el Control de Enfermedades (ECDC) proporcionó datos cruciales sobre las pruebas realizadas y la positividad por COVID-19 [5]. También recurrimos a datos sobre la ocupación hospitalaria y las unidades de cuidados intensivos (UCI) de ECDC para evaluar la capacidad de atención médica [6].

Todos estos datos se sometieron a un riguroso proceso de preprocesamiento y limpieza utilizando Python en Jupyter Notebook, específicamente diseñado para cada conjunto de datos. Este proceso de limpieza se repitió y revisó en varias ocasiones para garantizar la integridad de los datos. Además, comenzamos a trabajar con una base de datos SQL.

A medida que avanzaba nuestra investigación y el máster, nos dimos cuenta de que queríamos reorientar nuestros objetivos. En lugar de centrarnos en la modelización predictiva mediante Machine Learning, decidimos desarrollar una arquitectura en la nube y su respectiva automatización. En resumen, implementamos scripts que leen las fuentes de datos originales cada siete días, limpian los datos y los cargan en Azure. Los dashboards creados con PowerBI se conectan directamente a nuestra base de datos SQL actualizada en Azure.

Esta decisión de cambio en nuestro enfoque de investigación recibió el respaldo de nuestro tutor, y se apoyó en el estado actual de la literatura especializada en aplicaciones de Machine Learning en el ámbito de la salud pública. Los datos para modelos predictivos de COVID-19 se ven influidos por diversos factores, como mutaciones del virus, el proceso de vacunación y la inmunidad adquirida por quienes han estado previamente expuestos al virus.

En cuanto a nuestras hipótesis, debido a la falta de acceso a datos más concretos, eliminamos la hipótesis que hacía referencia a la incidencia de COVID-19 en países sin sistemas de salud pública sólidos. Los indicadores disponibles se centraban en el gasto público en salud, y no pudimos determinar a partir de qué nivel se podía hablar de una seguridad pública universal sólida propiamente dicha. Luego nos dimos cuenta de que había conflictos con los datos de temperatura y por tanto no se podía contrastar.

Esta revisión y adaptación de nuestro enfoque nos permitió avanzar en nuestra investigación de manera más efectiva y abordar preguntas relevantes en el contexto de la pandemia de COVID-19.

- Tras analizar el dashboard presentado en el primer archivo de Power BI (.pbix) que correspondería al tercer entregable, se extrajeron varias conclusiones sobre la relación entre la pandemia de COVID-19 y diversos indicadores. En el mapa de casos y muertes, se destacó que no existe una correlación directa entre los países con más casos por 100k habitantes y las muertes por 100k habitantes. Europa fue la región más afectada, con altos números de casos y muertes, pero se notó una disminución en la mortalidad en las olas de contagio más recientes debido a la progresiva vacunación.
- El análisis de casos vs. muertes reveló que no hay una correlación evidente, ya que la respuesta y preparación de cada país influyó en la mortalidad. Países como Perú mostraron un alto número de muertes por 100k habitantes, incluso sin tener uno de los mayores números de casos reportados, debido a oleadas iniciales de contagios. Además, hubo casos interesantes de países con alta tasa de contagios y baja mortalidad, como Corea del Sur, que gestionaron exitosamente la enfermedad a través de detección temprana, sistemas de salud efectivos y cooperación de la población.
- En la sección de casos/muertes vs. vacunación, se observó que no existe una correlación clara, pero se notó que a medida que aumenta el porcentaje de dosis recibidas, el número de muertes tiende a disminuir.
- El análisis de vacunación indicó que el continente asiático lidera en cantidad de dosis administradas, debido a su alta población, y que la mayoría de los países que recibieron la primera dosis también recibieron la segunda.
- En cuanto a las pruebas realizadas por país en Europa, Alemania lideró en pruebas de COVID-19 realizadas en datos recientes. Se observó una correlación moderada entre la cantidad de pruebas realizadas y los nuevos casos, pero se encontró que la secuencia de pruebas varió según si se consideraban pruebas totales o pruebas por cada 100,000 habitantes. Francia, Italia y Austria fueron los países líderes en pruebas totales, mientras que Austria, Dinamarca y Grecia lideraron en pruebas por cada 100,000 habitantes.
- En el análisis de casos en Unidades de Cuidados Intensivos (UCI) en Europa, se notó que el número de ingresos hospitalarios fue mayor que el de ingresos en UCI. Europa experimentó un pico en ingresos hospitalarios en enero de 2021, y se observó que la ocupación hospitalaria diaria superó la ocupación de la UCI en varios momentos, alcanzando su punto más alto en abril de 2021.

Teniendo en cuenta los aspectos sociodemográficos, geográficos y económicos, podemos decir que:

- En relación a la población urbana y la propagación, se observa una correlación entre la alta población urbana y el aumento de casos y muertes por COVID-19, debido a la mayor densidad de población que facilita la propagación del virus.
- En cuanto al acceso a servicios básicos e internet, se registra un incremento en casos y muertes en áreas con acceso limitado a servicios esenciales e internet. Esto podría estar relacionado con un menor desarrollo, lo que probablemente resulta en una movilidad reducida y, en última instancia, en una menor propagación del virus.
- En lo que respecta a las condiciones ambientales, como la temperatura, influyen en la propagación del virus. Las temperaturas más altas se asocian con menos contagios, sugiriendo que el virus se reproduce con mayor facilidad en climas fríos. Además, la altitud parece influir en la mortalidad del COVID-19, lo que podría estar relacionado con la disponibilidad de oxígeno y otros factores.

Este proyecto de investigación demostró ser flexible y adaptable. A medida que avanzábamos, ajustamos nuestras estrategias según los datos y resultados, optimizando nuestro trabajo. La exploración constante de nuevas posibilidades fue clave para el éxito, ya que nos permitió adaptarnos a direcciones inesperadas y aprovechar oportunidades en el camino.

Tabla de Contenidos

Introducción

El presente documento constituye la culminación de nuestro proyecto de máster sobre Covid-19. En él, consolidamos y sintetizamos las entregas previas en un único documento que ofrece una visión integral del proceso llevado a cabo a lo largo de nuestro trabajo final. A medida que avanzábamos en este proceso, nuestro equipo evolucionó de un simple grupo de trabajo a un equipo de alto rendimiento, desempeñando diferentes roles y aprovechando nuestras fortalezas individuales. Afrontamos desafíos y discrepancias, que fueron superados mediante una comunicación directa y honesta.

Los conocimientos teóricos y la experiencia adquirida durante el máster se reflejan en las entregas previas, abordando aspectos metodológicos, diseño, maquetación, estética y visualización de datos, entre otros. Nuestro trabajo comenzó con la identificación de las necesidades y problemas a abordar, en nuestro caso, la información relevante para profesionales de la salud acerca del Covid-19. Luego, diseñamos una estrategia para alcanzar nuestros objetivos, lo que implicó la búsqueda, perfilado, limpieza y transformación de los datos. Utilizamos herramientas aprendidas durante el curso, como SQL, Python y MySQL para la gestión de bases de datos, así como PowerBI para la visualización de datos. Además, consideramos la infraestructura necesaria y optamos por Microsoft Azure como plataforma de IAAS para aprovechar los beneficios del Cloud Computing.

A través de la integración de los conocimientos adquiridos en las diferentes asignaturas del curso, llegamos a la etapa final en la que extraíamos conclusiones y descubrimos nuevos conocimientos inesperados, resultado de la exploración y visualización de los datos. Este trabajo representa un esfuerzo colectivo que refleja nuestro compromiso y crecimiento a lo largo de este proyecto.

Núcleo del documento

Definición del proyecto y exploración de bases de datos

En el inicio de nuestro TFM, la primera entrega fue dedicada a la definición y planificación del proyecto. Se llevaron a cabo tanto un análisis interno, que evaluó nuestro conocimiento y familiaridad con la temática, como un análisis externo, que contextualizó la problemática del COVID-19. Estos análisis incluyeron herramientas como el mapa de empatía, el post-motorola, el team alignment map y la propuesta de valor. Además, se definieron de manera detallada los objetivos generales y específicos del proyecto.

Seguidamente, nos sumergimos en la investigación y comprensión de todas las posibles fuentes de datos, comenzando de manera lógica con la información proporcionada por la OMS para el acceso público. A medida que nuestra visión del proyecto se volvía más sólida y adquiríamos una comprensión más profunda de los datos disponibles, tomamos la decisión de priorizar las fuentes de información que el equipo de trabajo consideraba de mayor relevancia para el proyecto.

A continuación, procedimos con el perfilado y limpieza de las bases de datos preseleccionadas, así como la definición del modelo de datos. Luego, planificamos los recursos necesarios, los plazos y el alcance del proyecto. Esta etapa incluyó aspectos como recursos económicos, materiales y humanos, tiempos estimados, cronograma y el alcance esperado del proyecto. Nuestra metodología de trabajo se basó en AGILE y equipos de alto rendimiento. Finalmente, reflexionamos sobre el trabajo realizado por el equipo en esta fase inicial.

Estructura de la Base de Datos

Un paso de vital importancia en la creación de la solución definitiva que el proyecto ofrecería consistió en la meticulosa definición de la estructura de la base de datos. Cada aspecto, desde las tablas hasta los tipos de datos y las relaciones entre estas, se sometió a un análisis minucioso. El objetivo principal era asegurar la integridad, la coherencia y la inclusión de datos suficientes y pertinentes para el proyecto en curso.

A continuación, proporcionamos un desglose detallado de cada una de las tablas presentes en nuestro modelo de base de datos:

Tabla PAISES

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)	País, territorio, área
COD_CONTINENTE	VARCHAR(2)	Código de continente Alpha-2
CONTINENTE	VARCHAR(100)	Continente al que pertenece el país
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 suresteasiático **(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)	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 sureste asiático **(SEARO),** Oficina Regional para Europa **(EURO),** Oficina Regional para el Mediterráneo Oriental **(EMRO)** y Oficina Regional para el Pacífico Occidental **(WPRO).**
PAIS_NOM_2	VARCHAR (100)	País, territorio, área (Descriptivo diferente al campo País_Nom)
POBLACION	FLOAT	Número de Habitantes (2022)
LONGITUD	FLOAT	Coordenadas de Longitud
LATITUD	FLOAT	Coordenadas de Latitud
GDP_PER_CAPITA	FLOAT	El Producto Interno Bruto (PIB) Dividido Por La Población A Mitad De Año (2020)
USO_INDIVIDUAL_INTERNET	FLOAT	% de Población que ha utilizado internet en los últimos 3 meses del año 2020
USO_DE_AGUA_POTABLE	FLOAT	El porcentaje de personas que utilizan al menos servicios básicos de agua (2020)
USO_DE_SERVICIOS_ BASICOS_SANEAMIENTO	FLOAT	El porcentaje de personas que utilizan al menos servicios básicos de saneamiento, es decir, instalaciones sanitarias mejoradas que no se comparten con otros hogares (2020)
DENSIDAD_DE_POBLACION_X_KM2	FLOAT	Número de personas por km2 de superficie terrestre (2020)
POBLACION_RURAL	FLOAT	El Porcentaje de la Población rural respecto a la Población Total (2020)
POBLACION_URBANA	FLOAT	El Porcentaje de la Población urbana respecto a la Población Total (2020)
TEMPERATURA_PROMEDIO	FLOAT	Temperatura promedio por País
ALTURA_PROMEDIO	FLOAT	Altura promedio por País

Tabla COVID DAILY

Nombre del Campo	*Tipo de Datos*	*Descripción*
PAIS	VARCHAR(50)	País, territorio, área
PAIS_ISO3	VARCHAR(3)	Código de país ISO Alpha-3
PAIS_ISO2	VARCHAR(2)	Código de país ISO Alpha-2
FECHA_INFORMADA	DATE	Fecha de notificación a l'OMS (Datos Agregados por semana)
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 sureste 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 recuento acumulado anterior del recuento acumulado de casos actual. *
CASOS_ACUM	INTEGER	Casos confirmados acumulados notificados en la OMS hasta ahora.
MUERTES_NUEVAS	INTEGER	Nuevas muertes confirmadas. Se calcula restando las defunciones acumuladas anteriores de las defunciones acumuladas actuales. *
MUERTES_ACUM	INTEGER	Las muertes confirmadas acumuladas que se han notificado a la OMS hasta ahora.

** Es importante que los usuarios tengan en consideración que, además de registrar los nuevos casos y las defunciones notificadas en un día dado, las actualizaciones se realizan retrospectivamente para corregir los recuentos de días anteriores, según sea necesario, en función de la información posterior recibida*

Tabla CASOS HOSPITALIZADOS _UCI_EU

Nombre del Campo	***Tipo de Datos***	***Descripción***
PAIS_ISO3	VARCHAR(3)	Código de país ISO Alpha-3
PAIS_NOM	VARCHAR(100)	País, territorio, área
INDICADOR	VARCHAR(100)	Ocupación diaria del hospital, ocupación diaria de la UCI, nuevos ingresos hospitalarios semanales por 100k, nuevas admisiones semanales en UCI por 100k en paises europeos
FECHA	DATE	fecha de admisión en hospitalización o UCI
ANY_SEMANA	VARCHAR(10)	Semana del año de admisión en hospitalización o UCI
VALOR	FLOAT	Valor del Indicador correspondiente
FUENTE_ORIGEN	VARCHAR(100)	Indica la fuente de datos - REPORTING: Datos reportados por los estados miembros, o procedentes de informes oficiales

Tabla TESTING COVID_EU

Nombre del Campo	***Tipo de Datos***	***Descripción***
PAIS_ISO3	VARCHAR(3)	Código de país ISO Alpha-3
ANY_SEMANA	VARCHAR(10)	Año Semana de Admisión en Hospitalizació o UCI
NIVEL	VARCHAR(50)	Nacional (conjunto de datos archivados con datos nacionales) los datos subnacionales a la semana 36 de 2022 son disponible en ECDC sitio web)
CASOS_NUEVOS	INTEGER	Número de Nuevos casos confirmados

N_TESTS_REALIZADOS	INTEGER	Número de tests realizados
POBLACION	INTEGER	Número de habitantes
RATIO_TESTS	FLOAT	Tasa de pruebas por cada 100k habitantes
RATIO_POSITIVO	FLOAT	Positividad de la prueba semanal (%): 100 x Número de nuevos casos confirmados/número de pruebas hecho por semana
FUENTES_TESTS	VARCHAR(50)	• API del país • GitHub del país • Sitio web del país • Web Scraping manual • Otros • Encuesta • TESSy: datos proporcionados directamente por Estados miembros al ECDC a través de TESSy

Tabla VACUNACIONES

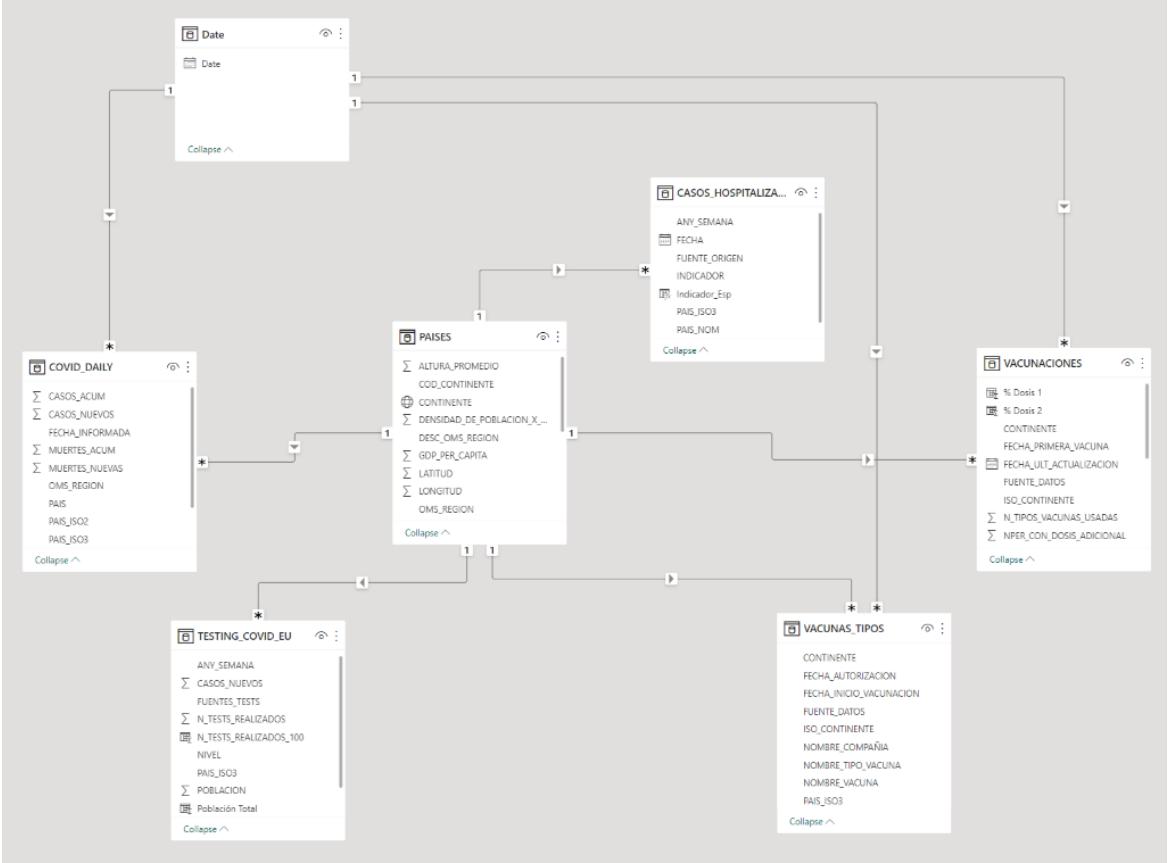
Nombre del Campo	*Tipo de Datos*	*Descripción*
PAÍS	VARCHAR(50)	País, territorio, área
PAIS_ISO3	VARCHAR(3)	Código de país ISO Alpha-3
CONTINENTE	VARCHAR(50)	Continente al que pertenece el país
ISO_CONTINENTE	VARCHAR(2)	Código de continente Alpha-2
OMS_REGION	VARCHAR(50)	Oficinas regionales de la OMS: 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 Our World in Data: https://ourworldindata.org/covid-vaccinations
FECHA_ULT_ACTUALIZACIÓN	DATE	Fecha de la última actualización
TOTAL_VACUNACIÓN_ACUM	FLOTA	Total acumulado de dosis de vacunas administradas
NPER_VACUNADAS_1DOSIS	INTEGER	Número acumulado de personas vacunadas con al menos una dosis
TOTAL_VACUNACION_PER100	FLOAT	Total acumulado de dosis de vacunas administradas por cada 100 habitantes
NPER_VACUNADAS_1DOSIS_PER100	FLOAT	Personas acumuladas vacunadas con al menos una dosis por cada 100 habitantes

NPER_VACUNADAS_DOSIS_FULL	FLOAT	Número acumulado de personas completamente vacunadas
NPER_VACUNADAS_DOSIS_FULL_PER 100	FLOAT	Número acumulado de personas completamente vacunadas 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	FLOAT	Número de tipos de vacunas utilizadas por país, territorio, área
NPER_CON_DOSIS_ADIDICIONAL	FLOAT	Las personas recibieron dosis de refuerzo o adicional
NPER_CON_DOSIS_ADIDICIONAL_PER 100	FLOAT	Las personas recibieron dosis de refuerzo o adicional por cada 100 habitantes

Tabla VACUNAS_TIPOS

Nombre del Campo	***Tipo de Datos***	***Descripción***
PAIS_ISO3	VARCHAR(3)	Código de país ISO Alpha-3
CONTINENTE	VARCHAR(50)	Continente al que pertenece el país
ISO_CONTINENTE	VARCHAR(2)	Código de continente Alpha-2
NOMBRE_VACUNA	VARCHAR(100)	Nombre corto combinado de la vacuna: "Empresa - Nombre del producto" (ver más abajo)
NOMBRE_TIPO_VACUNA	VARCHAR(90)	Nombre o etiqueta del producto de la vacuna, o tipos de vacuna (si no tiene nombre).
NOMBRE_COMPAÑÍA	VARCHAR(50)	Autorización de comercialización del titular del producto vacunal.
FECHA_AUTORIZACIÓN	DATE	Fecha en la que se autorizó el producto vacunal para su uso en el país, territorio, zona.
FECHA_INICIO_VACUNACIÓN	DATE	Fecha de inicio/lanzamiento de la vacunación con tipos de vacuna (excluye las vacunas durante los ensayos clínicos).
FUENTE_DATOS	VARCHAR(50)	Indica la fuente de datos - REPORTING: Datos reportados por los Estados miembros, o procedentes de informes oficiales - OWID: Datos procedentes de Our World in Data: https://ourworldindata.org/covid-vaccinations

La relación entre cada una de las tablas se puede observar en el siguiente diagrama, extraído de Power BI (una limitación que encontramos en el software de Azure Data Studio, es que no cuenta con un visualizador de la estructura de base de datos):



Fase ETL (Extracción, Transformación y Carga)

Extracción de Datos

Para obtener datos fiables sobre la evolución de la pandemia de Covid-19 a nivel mundial, consultamos diversas fuentes de organismos públicos, incluyendo la Organización Mundial de la Salud (OMS), el Centro Europeo para la Prevención y Control de Enfermedades (ECDC) y otros.

Transformación

Esta etapa se considera la más crítica y fundamental de todo el proceso, ya que en ella los datos obtenidos previamente se someten a procesos de tratamiento y limpieza, lo que permite que sean manipulados de manera sencilla y eficiente.

Durante esta fase, implementamos una serie de estándares y acciones mediante el uso de Python:

- **Normalización**: Se aplican reglas y normas con el objetivo de reducir la redundancia de los datos.
- **Eliminación de duplicados**: Esto conduce a la obtención de un conjunto de datos más coherente y fácil de manipular, al mismo tiempo que reduce las posibilidades de errores en el modelo de datos.
- **Verificación**: Se realizan controles en los datos para detectar posibles anomalías.
- **Clasificación**: Se agrupa y categoriza la información de manera lógica, lo que facilita su consulta y acceso en las bases de datos.

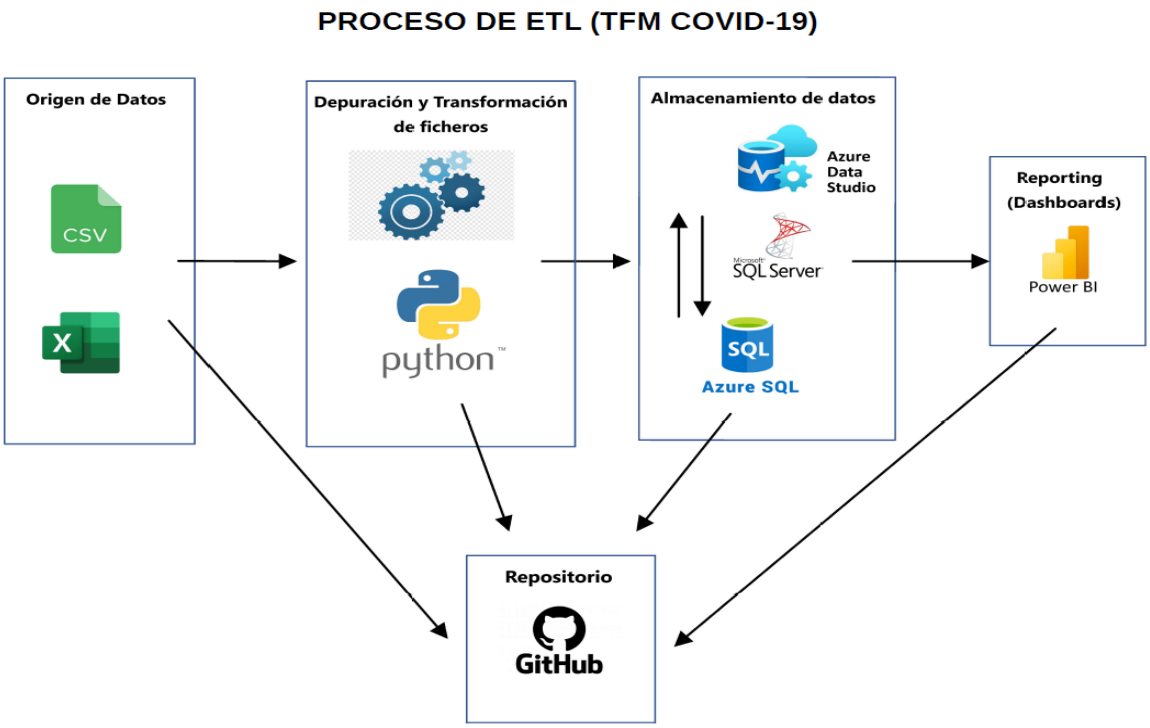
Carga de Datos

En esta etapa final del proceso ETL, nos encargamos de cargar los datos que hemos obtenido y procesado en nuestra base de datos (BBDD) de Azure Data Studio.

Inicialmente, realizamos una carga de archivos desde nuestra fuente de datos local. Sin embargo, después de superar ciertos problemas técnicos y de conexión, logramos cargar estos archivos de origen en una base de datos de Azure Data Studio. A través de una conexión con SQL Server, redirigimos nuestros datos a Azure SQL Database en la nube.

Para llevar a cabo la carga de nuestros archivos de origen, integramos nuestros Notebooks en Python que se encuentran en Github (Repositorio de Acceso público). Estos Notebooks se encargan de cargar las tablas en nuestra base de datos en Azure Cloud.

Posteriormente, utilizamos Power BI para conectarnos directamente a la base de datos en Azure. A través de esta conexión, creamos diversos paneles de control (Dashboards) que nos proporcionan información sobre la evolución de la pandemia de Covid-19, así como la evolución de la mortalidad. Además, llevamos a cabo un análisis de la correlación entre las dosis de vacunación administradas y la mortalidad, así como entre los pacientes hospitalizados y aquellos que han requerido cuidados intensivos en las unidades de cuidados intensivos (UCI).



Además, compartimos nuestro proyecto TFM en un repositorio público en Github para facilitar el acceso a cualquier persona interesada: [patriciaapenat/TFM: Covid-19, un análisis de casos y su correlación con vacunación, indicadores de desarrollo y variables ambientales (github.com)]

Notebooks

Dataframe Países

Este notebook se encarga de la carga de la "Tabla Maestra Países", que abarca códigos ISO2 e ISO3 de países a nivel mundial. Además, contiene información sobre el continente al que pertenecen, su población, el porcentaje de individuos que utilizan Internet, la proporción de personas con acceso a servicios básicos de

agua potable y saneamiento, la densidad de población (expresada en personas por kilómetro cuadrado de superficie terrestre), la relación de población rural frente al total y la proporción de población urbana en relación con el total. Asimismo, se considera el Producto Interno Bruto (PIB). Además de estos indicadores de desarrollo, se incorporan datos sobre la altura promedio sobre el nivel del mar y la temperatura promedio en regiones de interés.

[Acceso en GitHub aquí]

In [178…
```
df_paises.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 249 entries, 0 to 248
Data columns (total 20 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   PAIS_ISO3                             249 non-null    object
 1   PAIS_ISO2                             249 non-null    object
 2   PAIS_NOM                              249 non-null    object
 3   COD_CONTINENTE                        249 non-null    object
 4   CONTINENTE                            249 non-null    object
 5   OMS_REGION                            249 non-null    object
 6   DESC_OMS_REGION                       249 non-null    object
 7   PAIS_NOM_2                            249 non-null    object
 8   POBLACION                             249 non-null    float64
 9   LONGITUD                              249 non-null    float64
 10  LATITUD                               249 non-null    float64
 11  GDP_PER_CAPITA                        249 non-null    float64
 12  USO_INDIVIDUAL_INTERNET               249 non-null    float64
 13  USO_DE_AGUA_POTABLE                   249 non-null    float64
 14  USO_DE_SERVICIOS_BASICOS_SANEAMIENTO  249 non-null    float64
 15  DENSIDAD_DE_POBLACION_X_KM2           249 non-null    float64
 16  POBLACION_RURAL                       249 non-null    float64
 17  POBLACION_URBANA                      249 non-null    float64
 18  TEMPERATURA_PROMEDIO                  249 non-null    float64
 19  ALTURA_PROMEDIO                       249 non-null    float64
dtypes: float64(12), object(8)
memory usage: 39.0+ KB
```

In [180…
```
import pyodbc
server = 'servidortfmcovid-v2.database.windows.net'
database = 'basedatostfmcovid-v2'
username = 'admintfmcovid_v2'
password = 'TFMcovid2023!_v2'
cnxn = pyodbc.connect('DRIVER={SQL Server};SERVER='+server+';DATABASE='+database+';UID='+username+';PWD='+ password)
cursor = cnxn.cursor()

# Nombre de la tabla en la base de datos
nombre_de_tabla = 'dbo.PAISES'  # Reemplaza 'MiTabla' por el nombre de tu tabla

# Eliminar registros
cursor.execute(f"DELETE FROM {nombre_de_tabla}")

# Inserta los datos del DataFrame en la tabla
for index, row in df_paises.iterrows():
    insert_query = f"INSERT INTO {nombre_de_tabla} ({', '.join(df_paises.columns)}) VALUES ({', '.join(['?'] * len(df_paises
    cursor.execute(insert_query, tuple(row))
    cnxn.commit()

# Cierra la conexión
cnxn.close()
```

Dataframe Covid Daily OMS

Contiene información de casos diarios y muertes por fecha notificados a la OMS

In [1]:
```
import pandas as pd
import numpy as np

#!pip install mysql-connector-python
#!pip install sqlalchemy
#!pip install PyMySql
#!pip install ipython-sql
```

In [2]:
```
#pip install pycountry==20.7.3
#!pip install pycountry-convert==0.7.2
```

Acceso a los datasets de la OMS:

Exloración e Información del DataFrame **df_covid_daily**

In [3]:
```
df_covid_daily = pd.read_csv('https://covid19.who.int/WHO-COVID-19-global-data.csv')
df_covid_daily.head()
```

Out[3]:

	Date_reported	Country_code	Country	WHO_region	New_cases	Cumulative_cases	New_deaths	Cumulative_deaths
0	2020-01-03	AF	Afghanistan	EMRO	0	0	0	0
1	2020-01-04	AF	Afghanistan	EMRO	0	0	0	0
2	2020-01-05	AF	Afghanistan	EMRO	0	0	0	0

In [33]:
```
df_covid_daily_new
```

Out[33]:

	PAIS	PAIS_ISO3	PAIS_ISO2	FECHA_INFORMADA	OMS_REGION	CASOS_NUEVOS	CASOS_ACUM	MUERTES_NUEVAS
0	Afghanistan	AFG	AF	03-01-2020	EMRO	0	0	0
1	Afghanistan	AFG	AF	04-01-2020	EMRO	0	0	0
2	Afghanistan	AFG	AF	05-01-2020	EMRO	0	0	0
3	Afghanistan	AFG	AF	06-01-2020	EMRO	0	0	0
4	Afghanistan	AFG	AF	07-01-2020	EMRO	0	0	0

Acceso en GitHub aquí

Dataframe Pacientes Hospitalizados UCI

Contiene Información acerca de la ocupación diaria del hospital, ocupación diaria de la UCI, nuevos ingresos hospitalarios semanales por 100k, nuevas admisiones semanales en UCI por 100k en países europeos.

Acceso al dataset de hospitalizaciones en UCI

Exloración e Información del DataFrame **df_hosp_UCI**

In [3]:
```
df_hosp_UCI = pd.read_csv ('https://opendata.ecdc.europa.eu/covid19/hospitalicuadmissionrates/csv/data.csv')
df_hosp_UCI
```

Out[3]:

	country	indicator	date	year_week	value	source	url
0	Austria	Daily hospital occupancy	2020-04-01	2020-W14	856.000000	Country_Website	NaN
1	Austria	Daily hospital occupancy	2020-04-02	2020-W14	823.000000	Country_Website	NaN
2	Austria	Daily hospital occupancy	2020-04-03	2020-W14	829.000000	Country_Website	NaN
3	Austria	Daily hospital occupancy	2020-04-04	2020-W14	826.000000	Country_Website	NaN
4	Austria	Daily hospital occupancy	2020-04-05	2020-W14	712.000000	Country_Website	NaN
...	...	...	...	...	...	...	...
28100	Sweden	Weekly new ICU admissions per 100k	2023-09-10	2023-W36	0.047836	TESSy COVID-19 combined sources	NaN
28101	Sweden	Weekly new ICU admissions per 100k	2023-09-17	2023-W37	0.038269	TESSy COVID-19 combined sources	NaN
28102	Sweden	Weekly new ICU admissions per 100k	2023-09-24	2023-W38	0.095672	TESSy COVID-19 combined sources	NaN
28103	Sweden	Weekly new ICU admissions per 100k	2023-10-01	2023-W39	0.047836	TESSy COVID-19 combined sources	NaN
28104	Sweden	Weekly new ICU admissions per 100k	2023-10-08	2023-W40	0.066971	TESSy COVID-19 combined sources	NaN

In [15]:
```
df_hosp_UCI_NEW
```

Out[15]:

	PAIS_ISO3	PAIS_NOM	indicator	date	year_week	value	source	url
0	AUT	Austria	Daily hospital occupancy	01-04-2020	2020-14	856.000000	Country_Website	NaN
1	AUT	Austria	Daily hospital occupancy	02-04-2020	2020-14	823.000000	Country_Website	NaN
2	AUT	Austria	Daily hospital occupancy	03-04-2020	2020-14	829.000000	Country_Website	NaN
3	AUT	Austria	Daily hospital occupancy	04-04-2020	2020-14	826.000000	Country_Website	NaN
4	AUT	Austria	Daily hospital occupancy	05-04-2020	2020-14	712.000000	Country_Website	NaN

Establecemos **Conexión a BBDD y Carga de Ficheros** (Azure Data Studio)

In [23]:
```
import pyodbc
server = 'servidortfmcovid-v2.database.windows.net'
database = 'basedatostfmcovid-v2'
username = 'admintfmcovid_v2'
password = 'TFMcovid2023!_v2'
cnxn = pyodbc.connect('DRIVER={SQL Server};SERVER='+server+';DATABASE='+database+';UID='+username+';PWD='+ password)
cursor = cnxn.cursor()

# Nombre de la tabla en la base de datos
nombre_de_tabla = 'dbo.CASOS_HOSPITALIZADOS_UCI_EU'  # Reemplaza 'MiTabla' por el nombre de tu tabla

# Eliminar registros
cursor.execute(f"DELETE FROM {nombre_de_tabla}")

# Inserta los datos del DataFrame en la tabla
for index, row in df_hosp_UCI_NEW.iterrows():
    insert_query = f"INSERT INTO {nombre_de_tabla} ({', '.join(df_hosp_UCI_NEW.columns)}) VALUES ({', '.join(['?'] * len
    cursor.execute(insert_query, tuple(row))
    cnxn.commit()

# Cierra la conexión
cnxn.close()
```

[Acceso en GitHub aquí]

Dataframe Testing EU

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

Cargamos el dataset y configuramos el notebook

In [3]:
```
# importar paquetes
import pandas as pd
import numpy as np
import datetime as dt
```

In [4]:
```
# Leer el archivo
df_datos4 = pd.read_csv('https://opendata.ecdc.europa.eu/covid19/testing/csv/data.csv') #cargamos los datos
df_datos4.head()
```

Out[4]:

	country	country_code	year_week	level	region	region_name	new_cases	tests_done	population	testing_rate	positivity_rate
0	Austria	AT	2020-W01	national	AT	Austria	NaN	NaN	8978929	NaN	NaN
1	Austria	AT	2020-W02	national	AT	Austria	NaN	NaN	8978929	NaN	NaN
2	Austria	AT	2020-W03	national	AT	Austria	NaN	NaN	8978929	NaN	NaN
3	Austria	AT	2020-W04	national	AT	Austria	NaN	NaN	8978929	NaN	NaN
4	Austria	AT	2020-W05	national	AT	Austria	NaN	NaN	8978929	NaN	NaN

Podemos hacer una función para obtener ISO3

In [12]:
```
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))
```

In [13]:
```
df_datos4['iso3']
```

Establecemos **Conexión a BBDD y Carga de Ficheros** (Azure Data Studio)

In [21]:
```
import pyodbc
server = 'servidortfmcovid-v2.database.windows.net'
database = 'basedatostfmcovid-v2'
username = 'admintfmcovid_v2'
password = 'TFMcovid2023!_v2'
cnxn = pyodbc.connect('DRIVER={SQL Server};SERVER='+server+';DATABASE='+database+';UID='+username+';PWD='+ password)
cursor = cnxn.cursor()

# Nombre de la tabla en la base de datos
nombre_de_tabla = 'dbo.TESTING_COVID_EU'  # Reemplaza 'MiTabla' por el nombre de tu tabla

# Eliminar registros
cursor.execute(f"DELETE FROM {nombre_de_tabla}")

# Inserta los datos del DataFrame en la tabla
for index, row in df_datos4.iterrows():
    insert_query = f"INSERT INTO {nombre_de_tabla} ({', '.join(df_datos4.columns)}) VALUES ({', '.join(['?'] * len(df_da
    cursor.execute(insert_query, tuple(row))
    cnxn.commit()

# Cierra la conexión
cnxn.close()
```

[Acceso en GitHub aquí]

Dataframe Vacunaciones OMS

El fichero contiene 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.

Acceso al dataset de la OMS que a partir de ahora se llamará **df_vacunation**

In [31]:
```
df_vacunation =pd.read_csv ('https://covid19.who.int/who-data/vaccination-data.csv')
df_vacunation.head()
```

Out[31]:

	COUNTRY	ISO3	WHO_REGION	DATA_SOURCE	DATE_UPDATED	TOTAL_VACCINATIONS	PERSONS_VACCINATED_1PLUS_DOSE	TOTA
0	Albania	ALB	EURO	REPORTING	2023-07-16	3087677.0	1349150	
1	Chile	CHL	AMRO	REPORTING	2023-06-02	66273384.0	18106832	
2	Congo	COG	AFRO	REPORTING	2022-07-31	833210.0	695760	
3	Côte d'Ivoire	CIV	AFRO	REPORTING	2023-02-19	25263932.0	13568372	
4	Denmark	DNK	EURO	REPORTING	2023-08-20	14943741.0	4754913	

Identificamos **el/los registros** donde se ecuentran los **valores nulos**

1.-TOTAL_VACCINATIONS
2.-TOTAL_VACCINATIONS_PER100
3.-FIRST_VACCINE_DATE
4.-NUMBER_VACCINES_TYPES_USED
5.-PERSONS_BOOSTER_ADD_DOSE
6.-PERSONS_BOOSTER_ADD_DOSE_PER100

In [42]:
```
df_vacunation[df_vacunation['TOTAL_VACCINATIONS'].isnull()]
```

Out[42]:

	COUNTRY	ISO3	WHO_REGION	DATA_SOURCE	DATE_UPDATED	TOTAL_VACCINATIONS	PERSONS_VACCINATED_1PLUS_DOSE	TC
155	Eritrea	ERI	AFRO	REPORTING	03-07-2022	NaN	0	

Establecemos **Conexión a BBDD y Carga de Ficheros** (Azure Data Studio)

In [64]:
```
import pyodbc
server = 'servidortfmcovid-v2.database.windows.net'
database = 'basedatostfmcovid-v2'
username = 'admintfmcovid_v2'
password = 'TFMcovid2023!_v2'
cnxn = pyodbc.connect('DRIVER={SQL Server};SERVER='+server+';DATABASE='+database+';UID='+username+';PWD='+ password)
cursor = cnxn.cursor()

# Nombre de la tabla en la base de datos
nombre_de_tabla = 'dbo.VACUNACIONES'  # Reemplaza 'MiTabla' por el nombre de tu tabla

# Eliminar registros
cursor.execute(f"DELETE FROM {nombre_de_tabla}")

# Inserta los datos del DataFrame en la tabla
for index, row in df_vacunation.iterrows():
    insert_query = f"INSERT INTO {nombre_de_tabla} ({', '.join(df_vacunation.columns)}) VALUES ({', '.join(['?'] * len(d
    cursor.execute(insert_query, tuple(row))
    cnxn.commit()

# Cierra la conexión
cnxn.close()
```

[Acceso en GitHub aquí]

Dataframe Vacunas tipo OMS

Contiene el nombre de empresas y vacunas de ciertos fabricantes y los países que las han incorporado en su proceso de vacunación.

Acceso al dataset de la OMS que a partir de ahora se llamará df_vacunation_Meta

In [4]:
```
df_vacunation_Meta =pd.read_csv ('https://covid19.who.int/who-data/vaccination-metadata.csv')
df_vacunation_Meta.head()
```

Out[4]:

	ISO3	VACCINE_NAME	PRODUCT_NAME	COMPANY_NAME	AUTHORIZATION_DATE	START_DATE	END_DATE	COMMENT	DATA_S
0	SHN	AstraZeneca - AZD1222	AZD1222	AstraZeneca	NaN	NaN	NaN	NaN	
1	GRL	Moderna - mRNA-1273	mRNA-1273	Moderna	NaN	NaN	NaN	NaN	
2	FRO	Moderna - mRNA-1273	mRNA-1273	Moderna	NaN	NaN	NaN	NaN	
3	FRO	Pfizer BioNTech - Comirnaty	Comirnaty	Pfizer BioNTech	NaN	NaN	NaN	NaN	
4	JEY	Moderna - mRNA-1273	mRNA-1273	Moderna	NaN	NaN	NaN	NaN	

Con el módulo **pycountry** obtenemos datos adicionales para el dataframe y se renombran las columnas con la función **.columns**

In [28]:
```
import pycountry_convert as pc

def obtener_continente_ISO3(codigo_ISO3):
    try:
        continente_code = pc.country_alpha3_to_country_alpha2(codigo_ISO3)
        continente = pc.country_alpha2_to_continent_code(continente_code)
        continente_nombre = pc.convert_continent_code_to_continent_name(continente)
        return continente_nombre, continente
    except:
        return None, None

df_vacunation_Meta[['Continente', 'ISO_continente']] = df_vacunation_Meta['PAIS_ISO3'].apply(obtener_continente_ISO3).app

indice_ISO3 = df_vacunation_Meta.columns.get_loc("PAIS_ISO3")

df_vacunation_Meta.insert(indice_ISO3 + 1, "Continente", df_vacunation_Meta.pop("Continente"))
df_vacunation_Meta.insert(indice_ISO3 + 2, "ISO_continente", df_vacunation_Meta.pop("ISO_continente"))
```

In [29]:
```
df_vacunation_Meta
```

Out[29]:

	PAIS_ISO3	Continente	ISO_continente	NOMBRE_VACUNA	NOMBRE_TIPO_VACUNA	NOMBRE_COMPAÑIA	FECHA_AUTORIZACI
0	SHN	Africa	AF	AstraZeneca - AZD1222	AZD1222	AstraZeneca	
1	GRL	North America	NA	Moderna - mRNA-1273	mRNA-1273	Moderna	

Establecemos **Conexión a BBDD y Carga de Ficheros** (Azure Data Studio)

In [ ]:

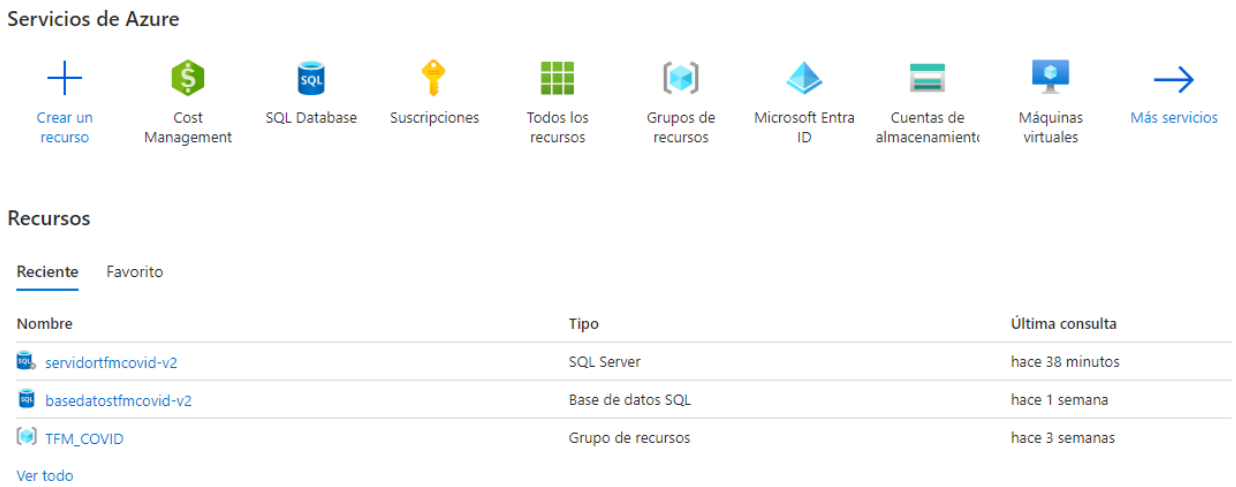
```
import pyodbc
server = 'servidortfmcovid-v2.database.windows.net'
database = 'basedatostfmcovid-v2'
username = 'admintfmcovid_v2'
password = 'TFMcovid2023!_v2'
cnxn = pyodbc.connect('DRIVER={SQL Server};SERVER='+server+';DATABASE='+database+';UID='+username+';PWD='+ password)
cursor = cnxn.cursor()

# Nombre de la tabla en la base de datos
nombre_de_tabla = 'dbo.VACUNAS_TIPOS'  # Reemplaza 'MiTabla' por el nombre de tu tabla

# Eliminar registros
cursor.execute(f"DELETE FROM {nombre_de_tabla}")

# Inserta los datos del DataFrame en la tabla
for index, row in df_vacunation_Meta.iterrows():
    insert_query = f"INSERT INTO {nombre_de_tabla} ({', '.join(df_vacunation_Meta.columns)}) VALUES ({', '.join(['?'] * 
    cursor.execute(insert_query, tuple(row))
    cnxn.commit()

# Cierra la conexión
cnxn.close()
```

[Acceso en GitHub aquí]

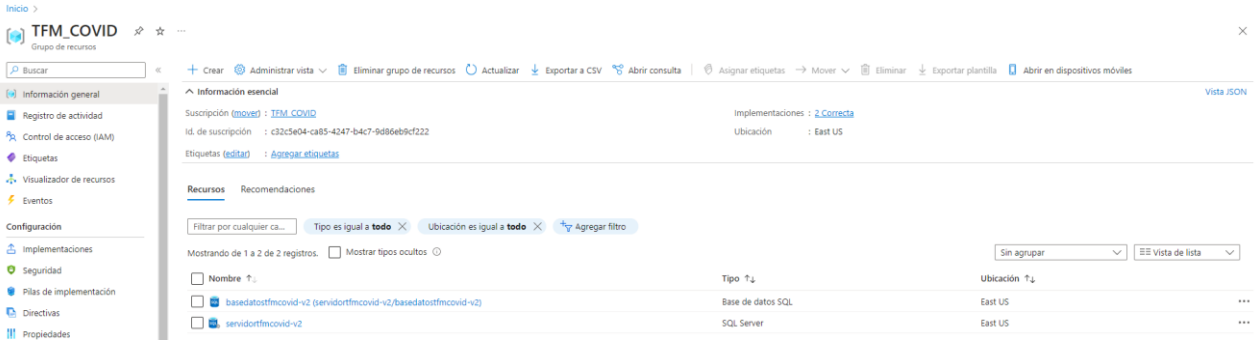
Almacenamiento en la Nube (Azure)

Una meta que el equipo de trabajo se planteó, fue la de aplicar la mayor cantidad de conocimientos adquiridos en el master para este proyecto de TFM. Dicho esto, se investigó la posibilidad de trasladar nuestra base de datos local (SQL Server) a una base de datos 100% virtual, en la nube de Microsoft Azure.

Para esto, se utilizó una cuenta personal, en la cual se crearon y configuraron los distintos elementos necesarios para el proyecto. Luego todos se agruparon bajo el grupo de recursos “TFM_COVID”.

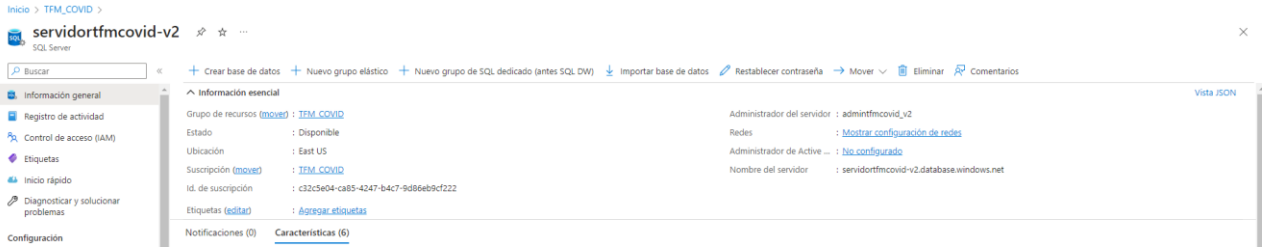


Grupo de Recursos

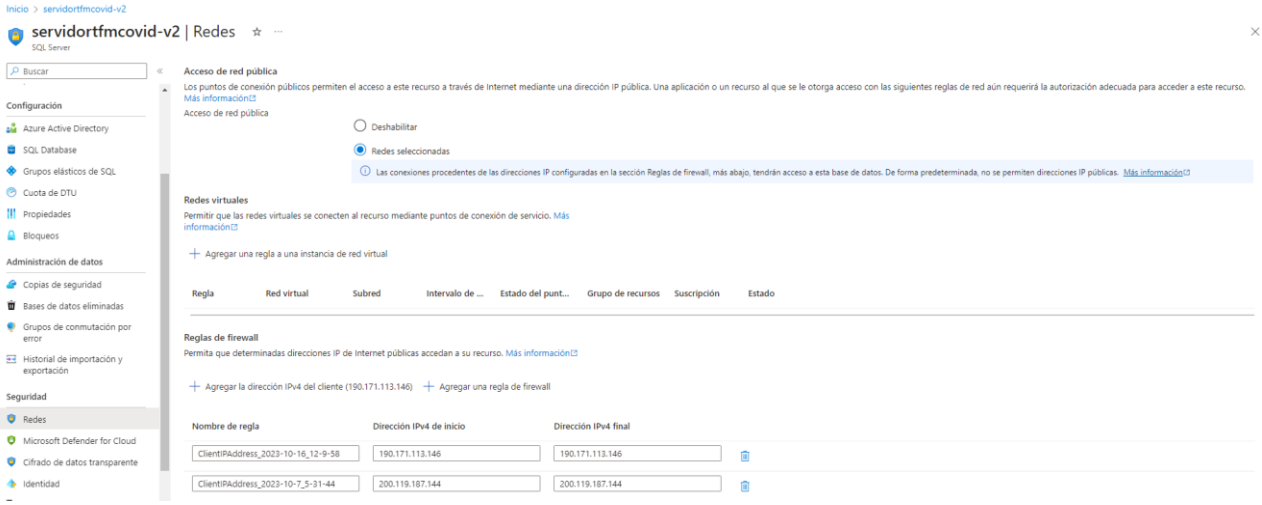


Servidor: servidortfmcovid-v2

El primer paso consistió en la creación y configuración del servidor, para el cual se realizaron todas las configuraciones necesarias, intentando mantener el coste lo más bajo posible, pero garantizando un correcto desempeño.

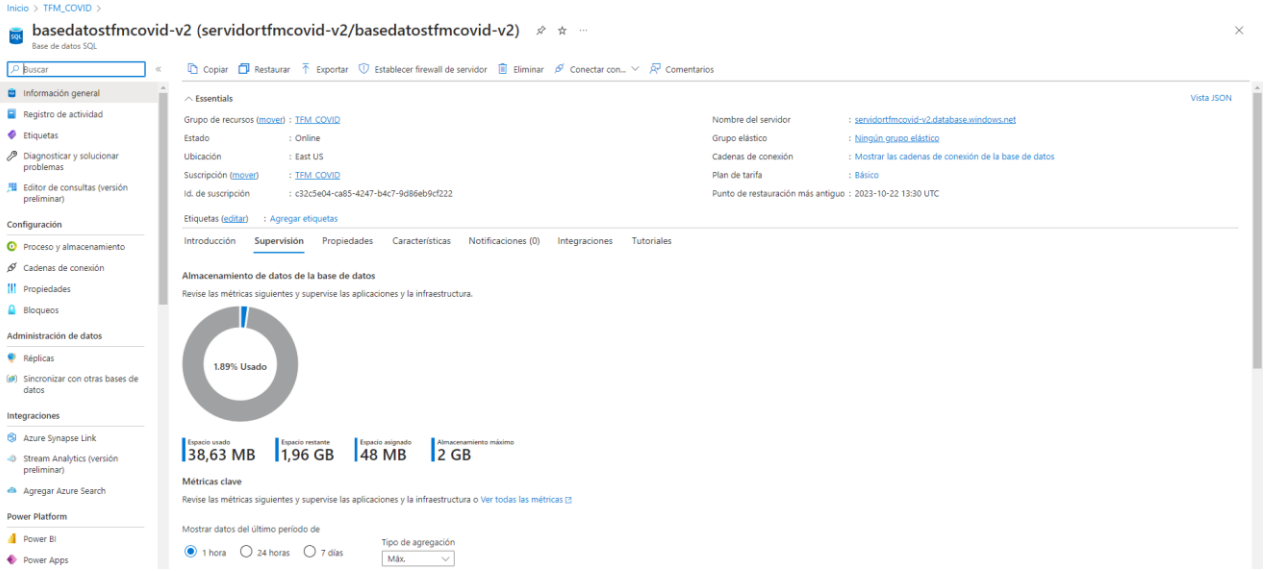


Una configuración importante que se estudió en clase ha sido la de “Reglas de Firewall”, que nos ha permitido el acceso al servidor desde distintas direcciones IP.



Base de Datos: basedatostfmcovid-v2

Ya con el servidor creado, se procedió con la creación y configuración de nuestra base de datos. En la imagen siguiente se puede apreciar la información general de la misma.

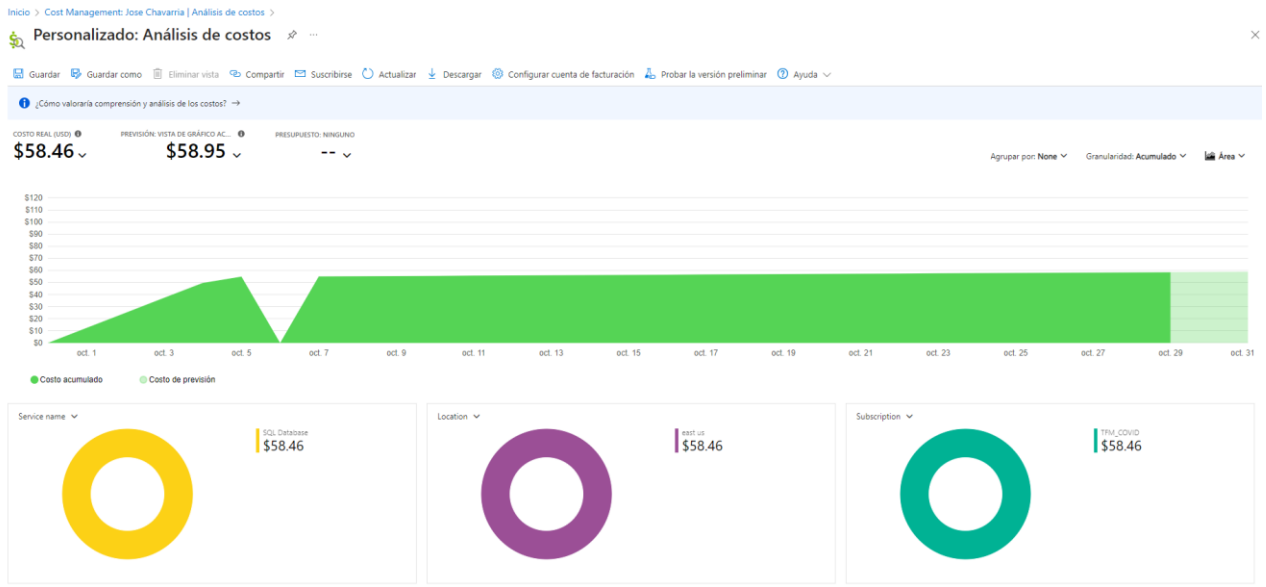


Como se puede observar, a pesar de la cantidad de datos ya ingresada (todos los datos requeridos para nuestro modelo de datos), actualmente se hace uso de poco menos del 2% de la capacidad de almacenamiento total.

Administración del Coste

Otro aspecto relevante que el equipo de trabajó consideró ha sido el de los costos. Inicialmente, durante la creación y configuración de la base de datos, se seleccionó alguna característica de manera incorrecta, aumentando el coste del servicio significativamente (casi €800 EUR al mes).

Debido a esto, el equipo de trabajo se movió rápidamente y con el apoyo del profesor Juan Luis Bermudez, se logró ajustar la configuración de manera que el costo no continuara creciendo exponencialmente. El costo estimado para final de mes es actualmente de $58.46 USD, un costo más que manejable.

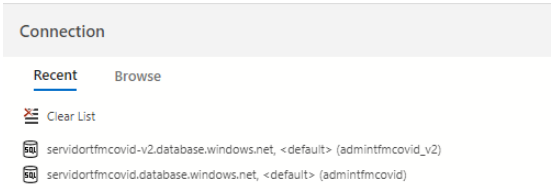


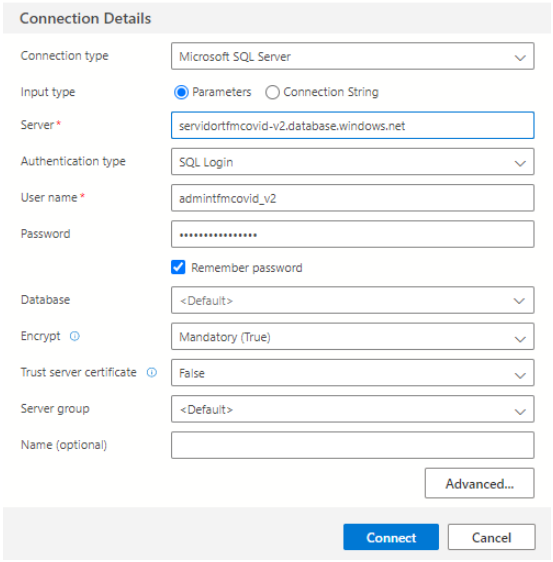
Almacenamiento en la Nube (Azure Data Studio)

Inicialmente, se utilizó el programa DBeaver para interactuar con la base de datos en Azure, sin embargo, esto nos daba varios problemas de configuración, y resultaba bastante tardado a la hora de utilizarlo.

Por recomendación del profesor Juan Luis Bermúdez, sustituimos DBeaver por Azure Data Studio, y los resultados no han podido ser mejores. La interfaz de usuario es muy sencilla de utilizar, todo se integra fácilmente y la velocidad del programa al interactuar con la nube es óptima.

Lo primero que se realizó fue configurar la conexión a nuestra base de datos en la nube:





Seguidamente, se utilizaron scripts de SQL para la creación de todas las tablas de nuestro modelo de base de datos. Algunas características se ajustaron directamente mediante el programa, como los tipos de datos de algunas columnas.

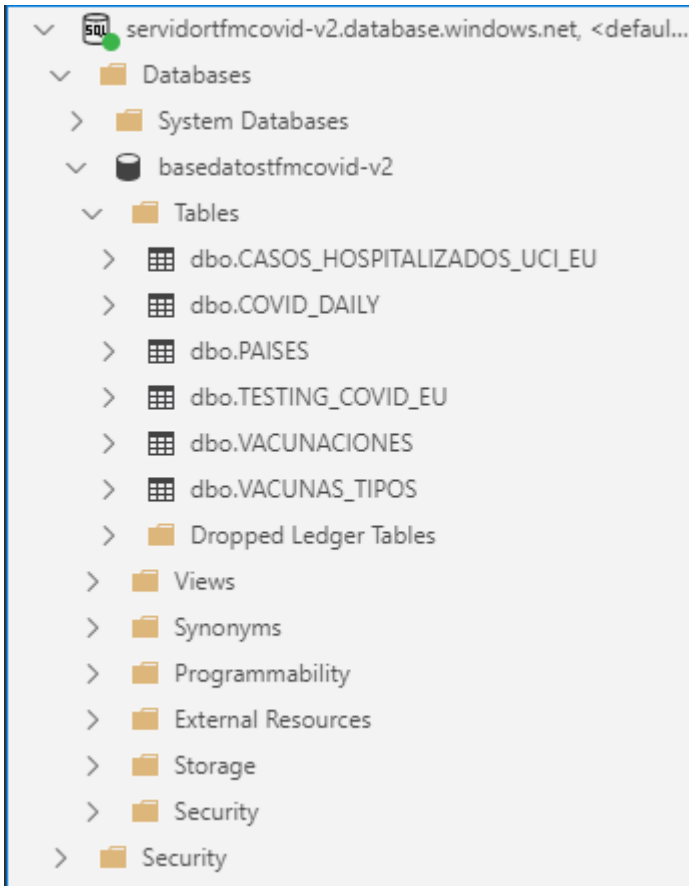


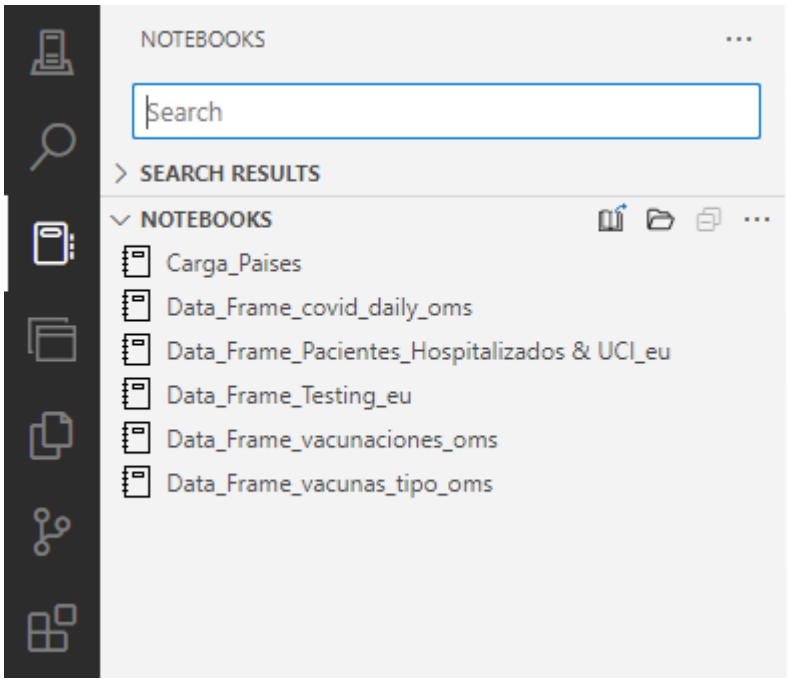
Table name: COVID_DAILY

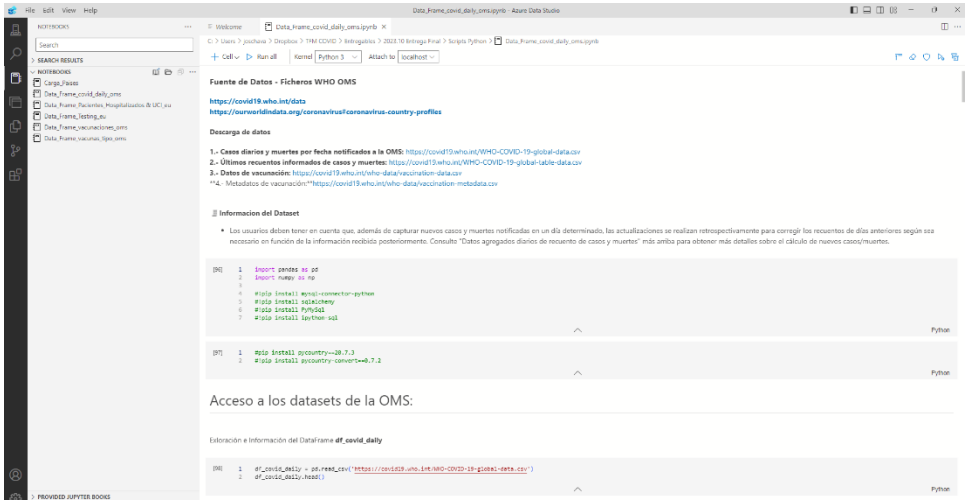
Columns | Primary Key | Foreign Keys | Check Constraints | Indexes | General

New Column | Move Up | Move Down

Move	Name	Type	Primary Key	Allow Nulls	Default Value	Remove	More Actions
=	PAIS	varchar(50)	☐	☐		🗑	...
=	PAIS_ISO3	varchar(3)	☐	☐		🗑	...
=	PAIS_ISO2	varchar(2)	☐	☐		🗑	...
=	FECHA_INFORMADA	date	☐	☐		🗑	...
=	OMS_REGION	varchar(50)	☐	☑		🗑	...
=	CASOS_NUEVOS	int	☐	☑		🗑	...
=	CASOS_ACUM	int	☐	☑		🗑	...
=	MUERTES_NUEVAS	int	☐	☑		🗑	...
=	MUERTES_ACUM	int	☐	☑		🗑	...

Otra característica interesante de este programa que ha sido de gran utilidad para el equipo de proyecto ha sido su módulo de “Notebooks”. El mismo permite abrir y actualizar los “notebooks” de python, utilizados para toda la descarga, transformación y carga de los datos en nuestra base de datos en la nube.



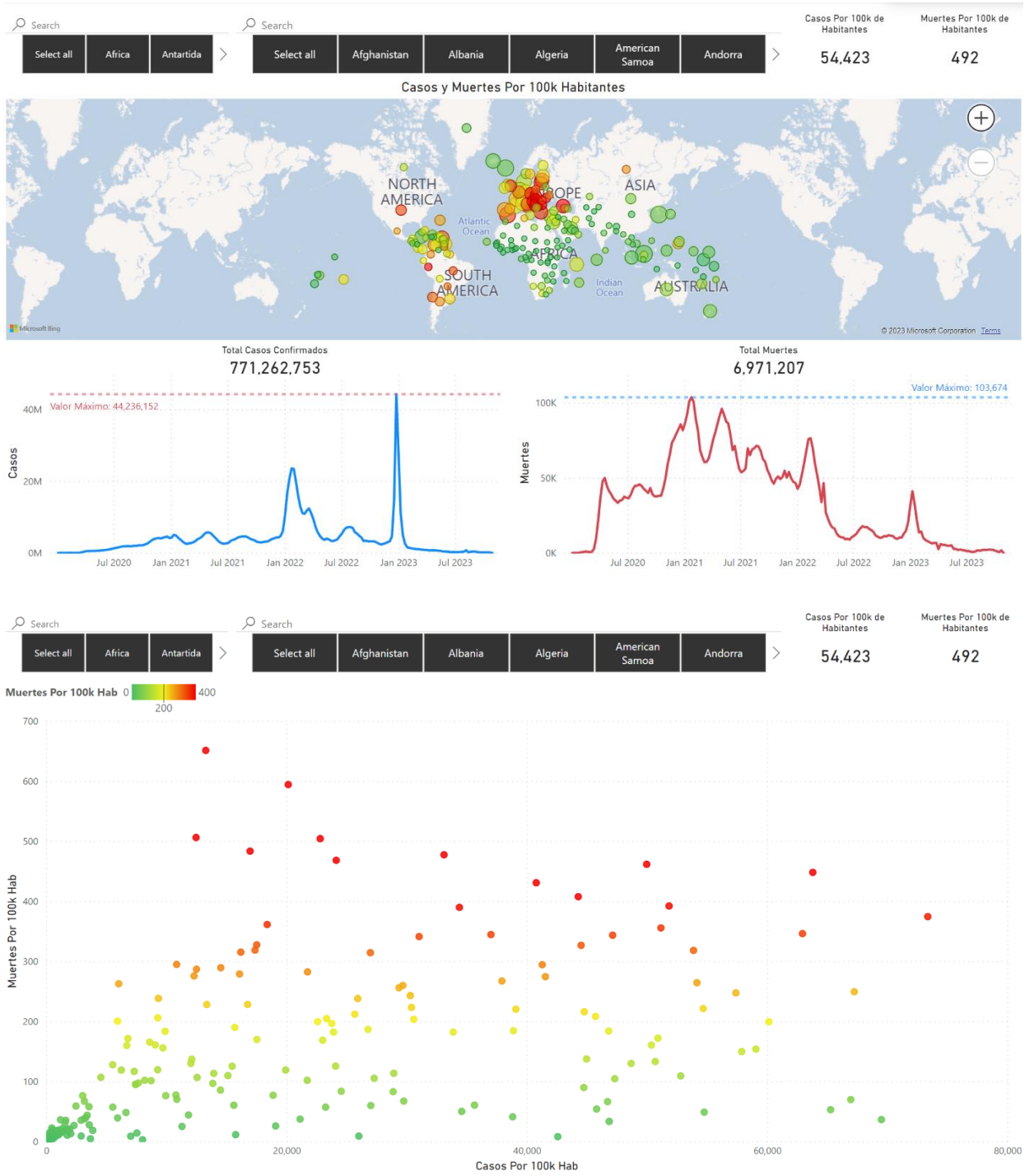


Visualización de los Datos

Enriquecimos nuestro análisis con datos relativos a diversos indicadores de desarrollo del Banco Mundial. Nos centramos en variables clave, como el acceso a Internet, servicios básicos de agua potable y saneamiento, densidad de población, proporción de población rural y urbana, y el Producto Interno Bruto (PIB). También incorporamos datos geográficos como la altitud. Todos estos datos se integraron en nuestro dataframe de países para mejorar la calidad de nuestro análisis.

A continuación, encontrarán nuestros dashboards: (tambien podrán encontrar el archivo .pbix aquí)

Casos y Muertes



Este tablero de información sobre la pandemia de COVID-19 presenta datos relevantes relacionados con la propagación del virus en diferentes regiones y países. Está diseñado para proporcionar una visión integral de la situación global.

En la parte superior del tablero, se encuentra un filtro por continente y datos sobre la cantidad de casos y muertes por cada 100,000 habitantes.

En el gráfico central, se representa un diagrama de burbujas del mundo. El tamaño de las burbujas representa la cantidad de casos por cada 100,000 habitantes, mientras que el color de las burbujas indica las muertes por cada 100,000 habitantes.

En la sección inferior, se muestran los totales de casos confirmados y muertes confirmadas. Todos los gráficos cuentan con la opción de filtrar el resto de los gráficos. Basándonos en esta página podemos decir que:

- No existe una correlación clara entre la cantidad de casos por 100,000 habitantes y las muertes por 100,000 habitantes. Esto se debe en parte a que el COVID-19 afectó a diferentes países en oleadas, y no todos estaban preparados para la magnitud de los contagios.
- Europa ha sido la región del mundo más afectada en términos de casos y muertes, con 185,000 casos por 100,000 habitantes y 1,500 muertes por 100,000 habitantes. Le siguen Suramérica con 32,000 casos y 629 muertes, y Norteamérica con 37,000 casos y 481 muertes. Oceanía y África han experimentado menos impacto, con Oceanía registrando 55,000 casos y 108 muertes, y África mostrando menos casos, pero una mortalidad más alta, con 6,000 casos y 119 muertes.
- Se observa que las primeras oleadas de contagio (de 2020 a mediados de 2021) fueron mucho más mortales que las oleadas más recientes, debido al progreso de la vacunación. La primera vacuna se aprobó el 11 de diciembre de 2020 (Pfizer-BioNTech).
- Cada país tiene una historia única en relación con la pandemia. Por ejemplo, Perú y Bulgaria presentan tasas de mortalidad diferentes. Perú experimentó una mortalidad elevada al inicio de la pandemia, pero luego disminuyó, a pesar de una ola de contagios en 2022. En el caso de Bulgaria, la mortalidad se mantuvo alta en 2022.

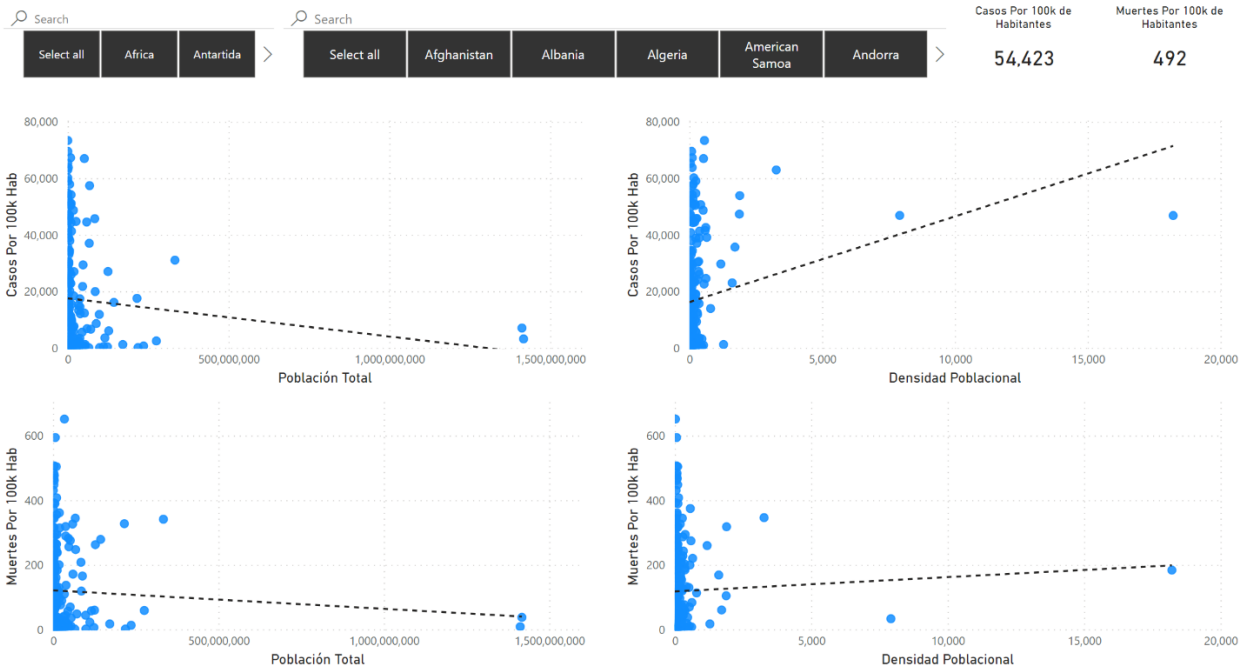
El objetivo del gráfico de dispersión es establecer una relación entre la cantidad de casos y las muertes por cada 100,000 habitantes. Los colores en el gráfico representan la cantidad de muertes por cada 100,000 habitantes. Al analizar este tablero de información, se pueden extraer las siguientes conclusiones:

1. No se observa una correlación clara entre la cantidad de casos y muertes. Esto se debe, en parte, a que el COVID-19 afectó a las naciones en olas, y no todas estaban preparadas para la magnitud de los contagios. Un ejemplo destacado es Perú, el punto rojo más alto en el gráfico, que no se encuentra entre los países con mayor cantidad de casos reportados, pero tiene la mayor tasa de muertes por 100,000 habitantes. Esto se explica por la presencia de oleadas de contagio en 2020 y 2021, cuando el país no estaba preparado, lo que resultó en un mayor número de muertes en comparación con el pico histórico de casos en enero de 2022, cuando la población ya estaba vacunada.
2. Existen casos interesantes de países con un alto número de contagios por 100,000 habitantes y una mortalidad notablemente baja, como Corea del Sur. Este país ha logrado gestionar exitosamente la enfermedad a través de:
    - Detección temprana mediante un programa agresivo de pruebas.
    - Un sistema de salud eficiente.
    - Comunicación efectiva y transparente con la población.
    - Avances tecnológicos en el control y seguimiento.
    - Cooperación de la población al seguir las indicaciones del gobierno.

Casos y muertes en función de la vacunación

Estos cuatro gráficos de dispersión pretenden identificar la relación entre la cantidad de casos y muertes reportadas, con los indicadores de vacunación (1 dosis y 2 dosis) por país. Analizando los mismo, si bien no se detecta que exista una correlación, en el caso de la cantidad de casos en función del % de dosis recibidas de la primera dosis, el no. de muertes cae, si bien no disminuyen el no. de casos por cada 100.000 habitantes. Caso similar ocurre con la 2da. Dosis. Esto se confirma al ver reflejado ese hecho en las gráficas de muerte en función del % de dosis recibidas.

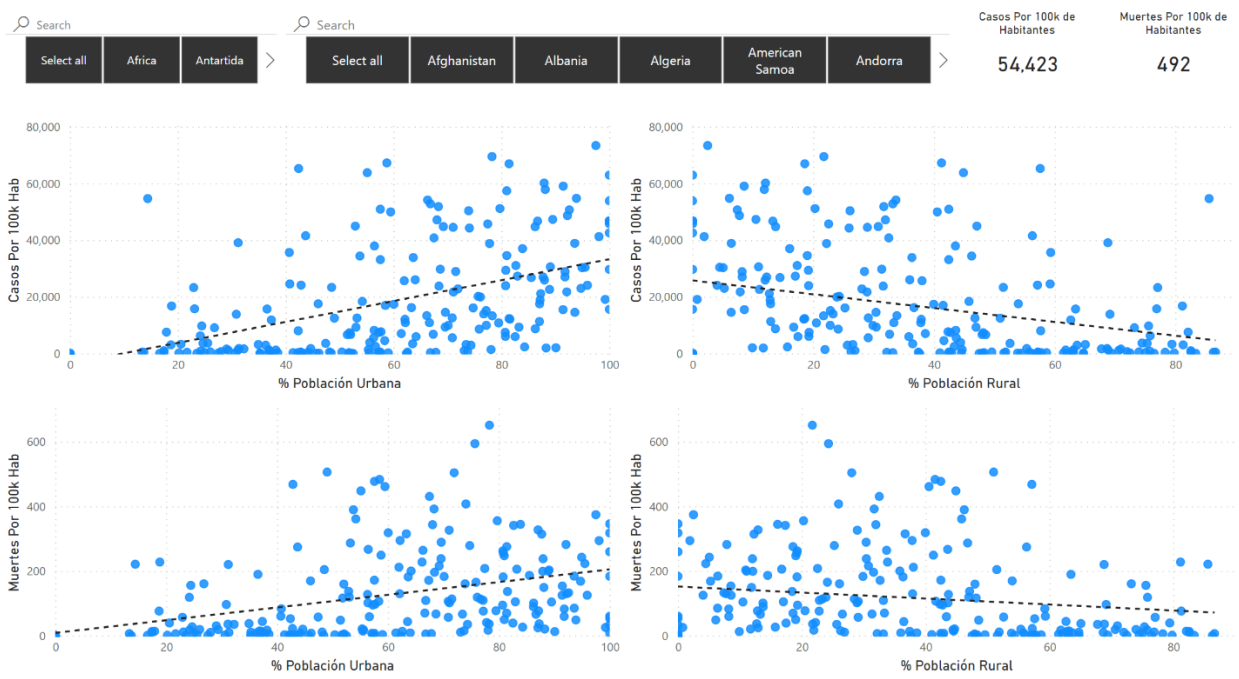
Casos y muertes en función de la población



En los gráficos de dispersión presentados, se utiliza un plano cartesiano en el que el eje x representa la población total en los gráficos de la columna izquierda y la densidad poblacional en los de la columna derecha. Por otro lado, las filas superiores muestran los casos, mientras que las filas inferiores representan las muertes, entonces, el eje y, por su parte, corresponde a la cantidad de casos o muertes, dependiendo de la ubicación.

Un patrón común en todos estos gráficos es la tendencia de los registros a agruparse en el borde derecho, con una distribución ascendente. Este patrón sugiere una relación positiva significativa entre la población y el número de casos o muertes.

Casos y muertes en función de áreas urbanas o rurales

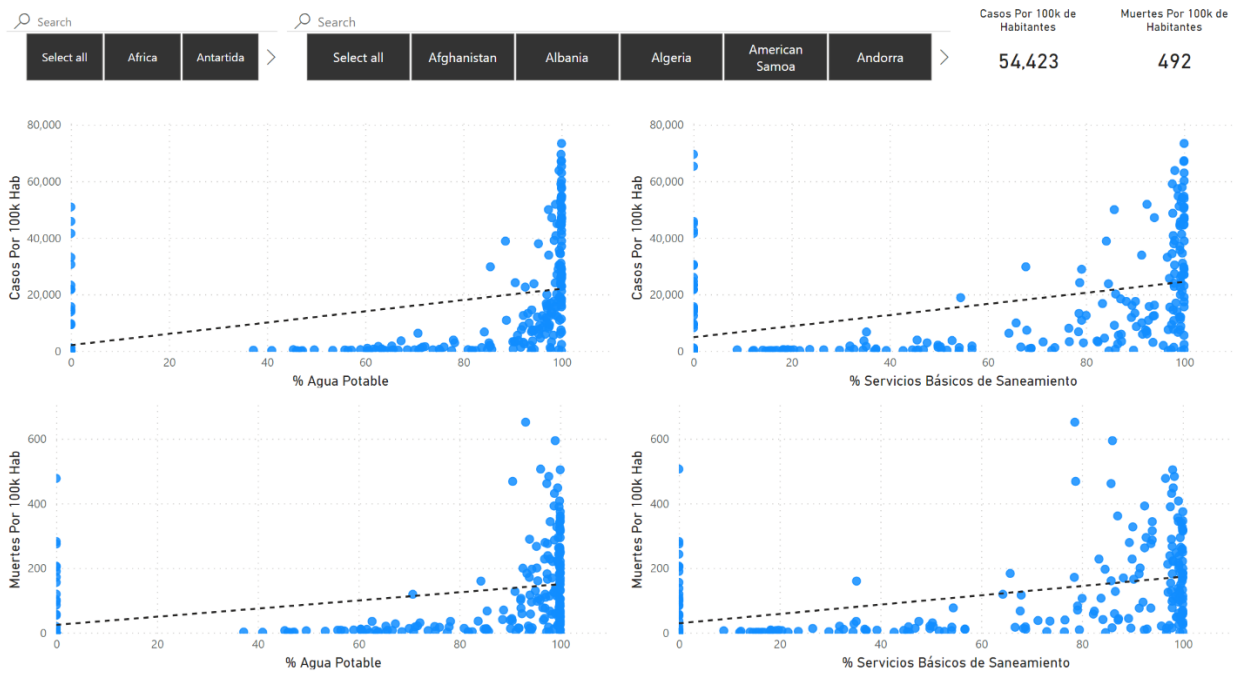


En los gráficos de dispersión presentados, se utiliza un plano cartesiano en el que el eje x representa el porcentaje de población urbana en los gráficos de la columna izquierda y el porcentaje de población rural en los de la columna derecha. Por otro lado, las filas superiores muestran los casos, mientras que las filas inferiores representan las muertes, entonces, el eje y, por su parte, corresponde a la cantidad de casos o muertes, dependiendo de la ubicación.

En general, se observa que los países con mayor población urbana también tienen mayor incidencia de casos y muertes por COVID-19. Esto se debe a que las zonas urbanas suelen tener una mayor densidad de población, lo que facilita la propagación del virus.

Sin embargo, hay algunos países que presentan una incidencia menor de casos y muertes por COVID-19. Estos países suelen ser de menor tamaño y con una economía menos desarrollada, lo que puede dificultar la propagación del virus.

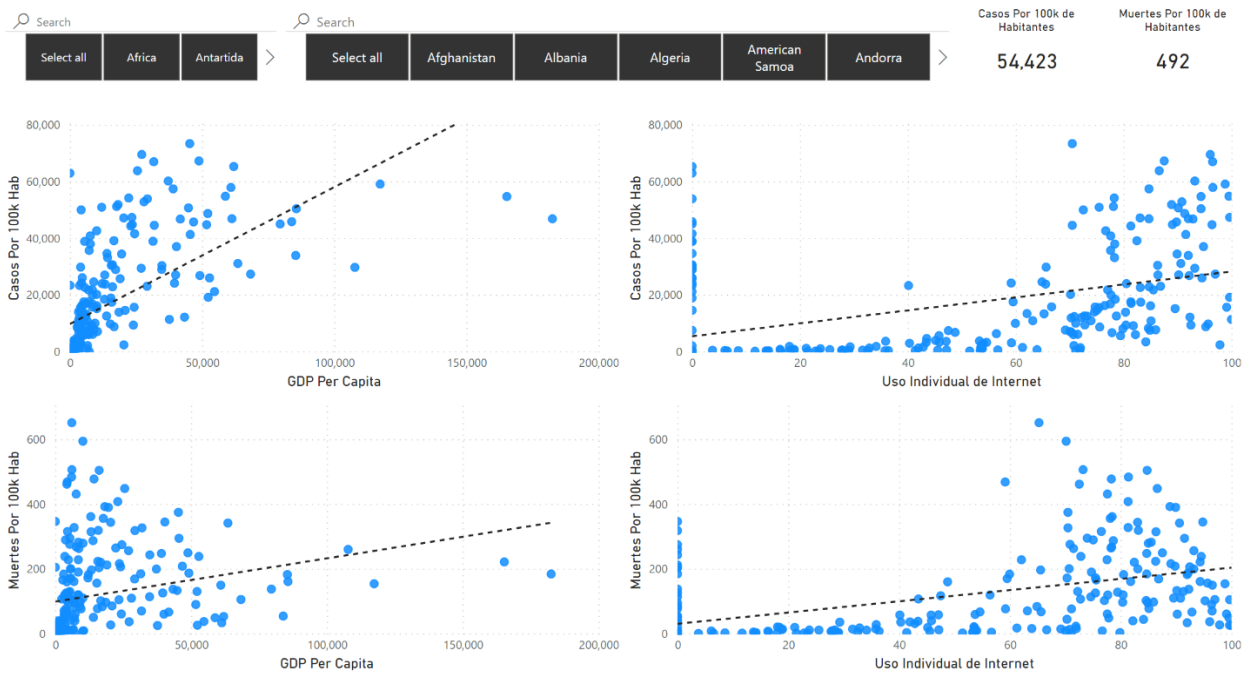
Casos y muertes en función de acceso a servicios básicos



En los gráficos de dispersión presentados, se utiliza un plano cartesiano en el que el eje x representa el porcentaje de agua potable en los gráficos de la columna izquierda y el porcentaje de servicios básicos de saneamiento en los de la columna derecha. Por otro lado, las filas superiores muestran los casos, mientras que las filas inferiores representan las muertes, entonces, el eje y, por su parte, corresponde a la cantidad de casos o muertes, dependiendo de la ubicación.

Estos gráficos de dispersión pretenden identificar la relación entre la cantidad de casos y muertes reportadas, con los indicadores de agua y servicios básicos de saneamiento. Se puede identificar que a mayor porcentaje de agua potable y servicio de saneamiento existe un mayor porcentaje de casos y muertes reportadas.

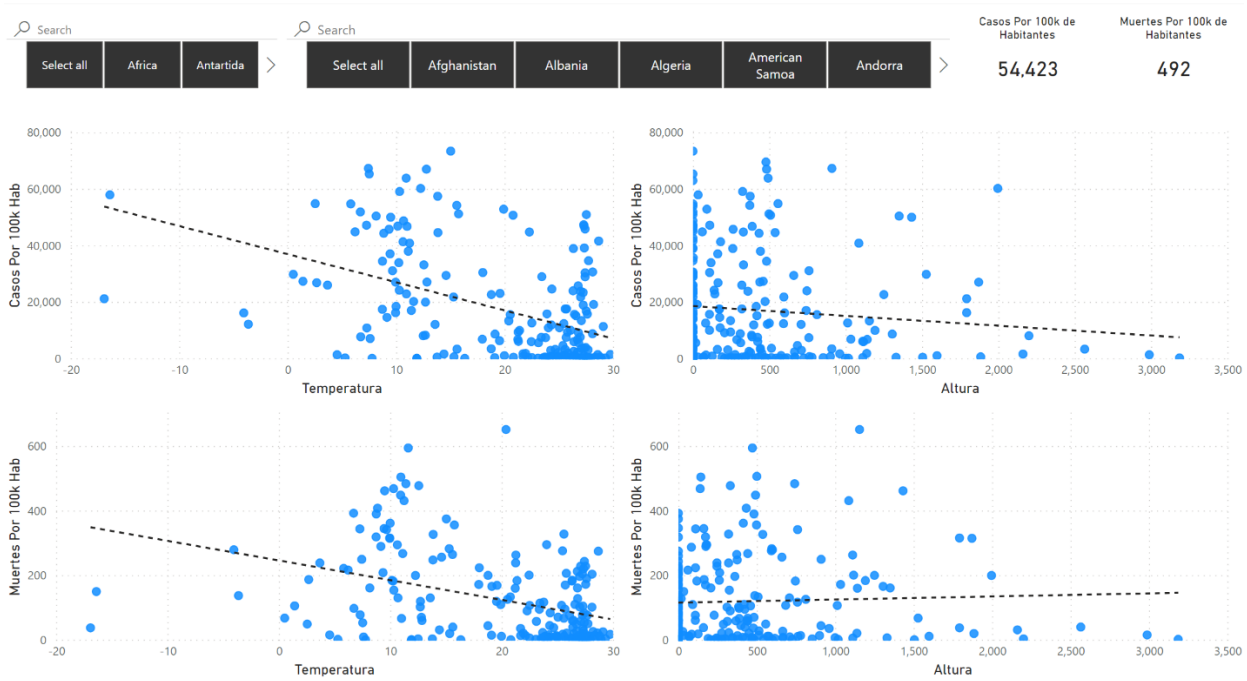
Casos y muertes en función del PIB



En los gráficos de dispersión presentados, se utiliza un plano cartesiano en el que el eje x representa el GDP per cápita en los gráficos de la columna izquierda y uso individual de internet en los de la columna derecha. Por otro lado, las filas superiores muestran los casos, mientras que las filas inferiores representan las muertes, entonces, el eje y, por su parte, corresponde a la cantidad de casos o muertes, dependiendo de la ubicación.

Estos gráficos de dispersión pretenden identificar la relación entre la cantidad de casos y muertes reportadas, con los indicadores de GDP por persona y uso individual de internet. En cuanto a los gráficos de dispersión de casos y muertes reportadas y GDP per cápita, se puede identificar que, a mayor porcentaje de casos y muertes reportadas, se presenta cuando el GDP per cápita es menor. Por otro lado, en cuanto a los gráficos de dispersión de casos y muertes reportadas y uso individual de internet, se puede identificar que, a mayores casos de casos y muertes esta correlacionada con mayor número de uso individual de internet.

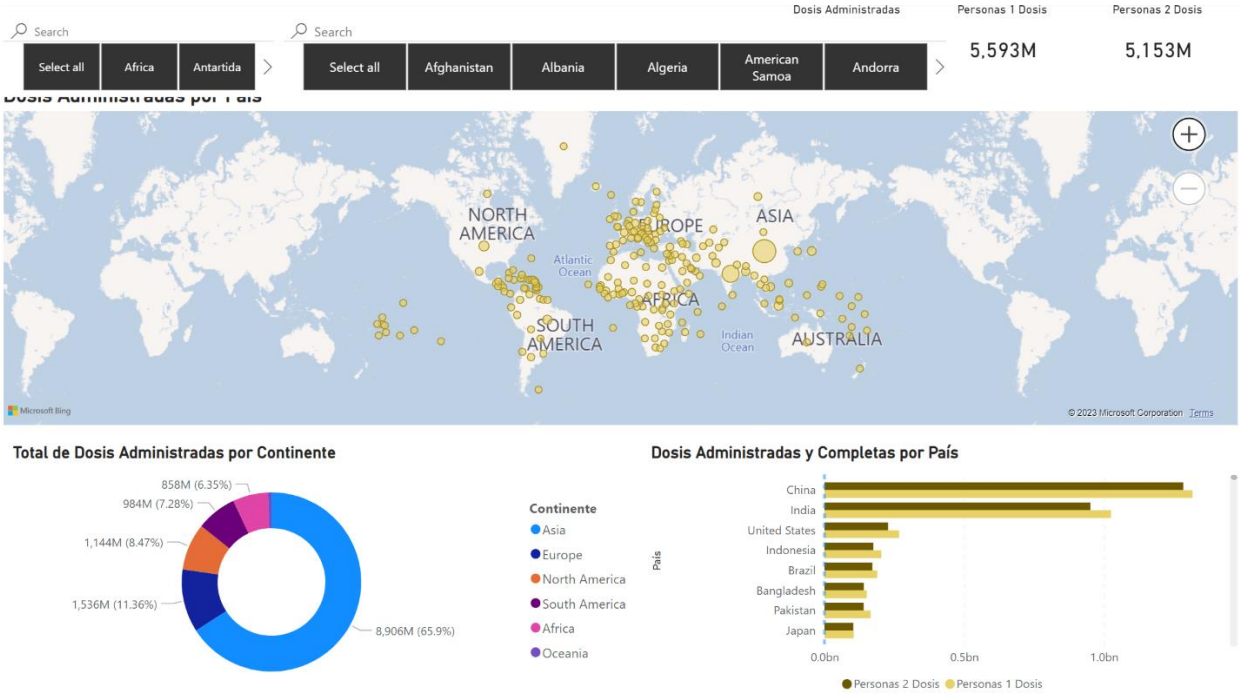
Casos y muertes en función de condiciones ambientales



En los gráficos de dispersión presentados, se utiliza un plano cartesiano en el que el eje x representa la temperatura en los gráficos de la columna izquierda y la altura en los de la columna derecha. Por otro lado, las filas superiores muestran los casos, mientras que las filas inferiores representan las muertes, entonces, el eje y, por su parte, corresponde a la cantidad de casos o muertes, dependiendo de la ubicación.

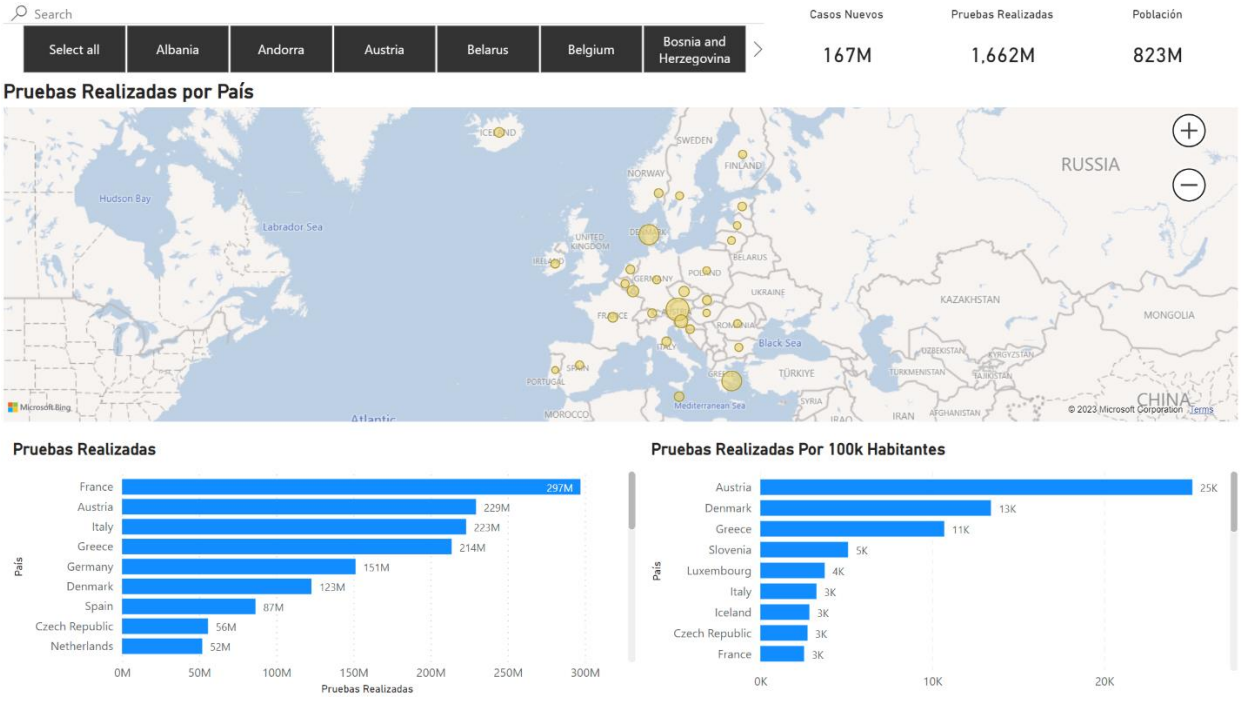
Estos gráficos de dispersión pretenden identificar la relación entre la cantidad de casos y muertes reportadas, con los indicadores de temperatura y altura. En cuanto a los gráficos de dispersión de casos y muertes reportadas y temperatura, se puede identificar que, existe un menor porcentaje de casos y muertes reportadas, se presenta cuando la temperatura es mayor. Por otro lado, en cuanto a los gráficos de dispersión de casos y altura, se puede identificar que, su línea de tendencia es casi recta, sin embargo, cuando se compara con el grafico de dispersión de muertes y alturas, su línea de tendencia tiende a inclinarse, es decir, que parece indicar la altura influyo en la mortalidad.

Vacunación



Este gráfico muestra un comparativo de vacunación por país, en términos de cantidad de dosis administradas. Claramente el continente asiático es el que cuenta con la mayor cantidad de habitantes, y por consiguiente es por mucho el continente con mayor cantidad de dosis administradas. También se puede ver que, en la gran mayoría de los países, los que recibieron la 1ª. dosis también recibieron la 2da.

Pruebas realizadas por país (Europa)



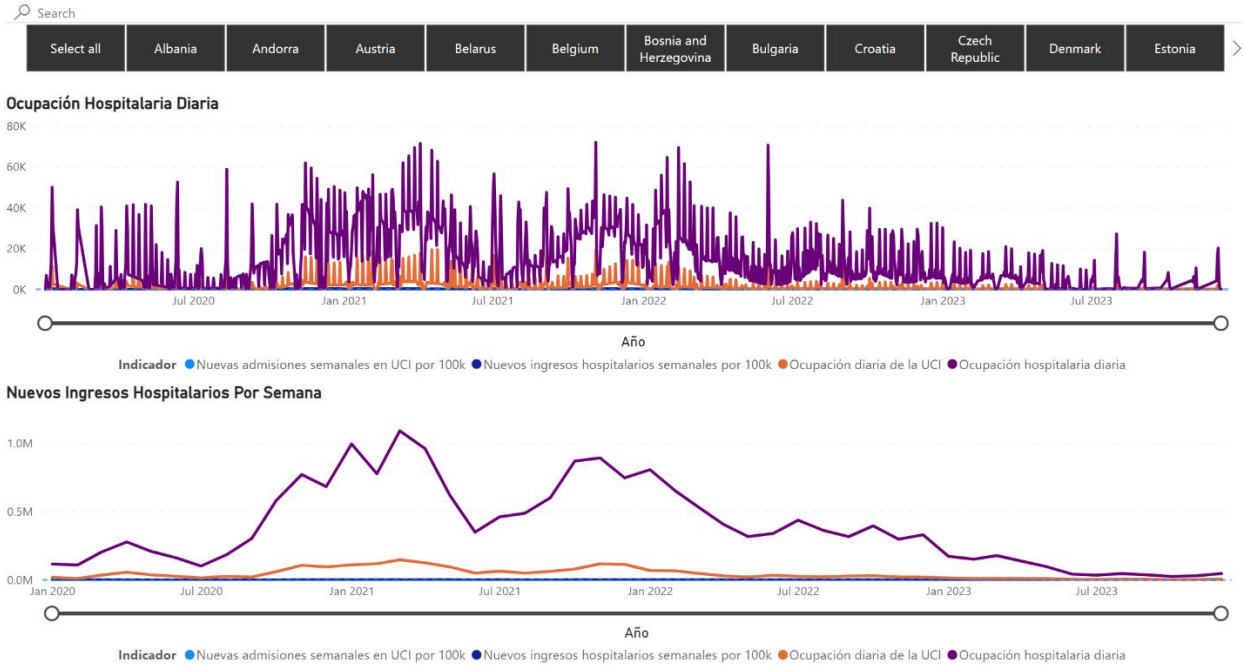
El gráfico presente muestra datos relacionados con la tasa de pruebas semanales por cada 100,000 habitantes y la positividad de las pruebas semanales (%), basados en diversas fuentes de datos.

El número de casos semanales utilizado para calcular la positividad de la prueba semanal por país o región subnacional se deriva de los datos recopilados por ECDC Epidemic Intelligence.

A partir de las visualizaciones en el panel de control y otras realizadas (no incluidas), se pueden destacar los siguientes hallazgos:

- Los datos más recientes indican que Alemania lidera en términos de la cantidad de pruebas de COVID-19 realizadas.
- Se observa una correlación moderada de 0.56 entre el número de pruebas realizadas (N_TEST_REALIZADOS) y los nuevos casos (CASOS_NUEVOS), lo cual es esperable, aunque no tan alta como se podría haber anticipado.
- La secuencia de pruebas realizadas varía dependiendo de si se consideran el total de pruebas realizadas o las pruebas por cada 100,000 habitantes.
- En términos absolutos de pruebas realizadas, los países líderes son Francia, Italia y Austria. Sin embargo, al considerar las pruebas por cada 100,000 habitantes, los principales países son Austria, Dinamarca y Grecia

Ingresos en UCI (Europa)



En el gráfico presente, se presentan datos relativos a la hospitalización y la tasa de admisión en Unidades de Cuidados Intensivos (UCI) debido a COVID-19, desglosados por fecha y país.

A nivel regional, se observa una tendencia en Europa durante el período de 2020 a 2023. El número de nuevos ingresos hospitalarios semanales por cada 100,000 habitantes ha sido consistentemente mayor que la tasa de nuevos ingresos semanales en UCI por cada 100,000 habitantes. Además, se destaca que, en enero de 2021, Europa experimentó su punto más alto en términos de nuevos ingresos hospitalarios semanales por cada 100,000 habitantes.

En cuanto a la ocupación hospitalaria, se nota que el número de pacientes con COVID-19 en los hospitales en un día dado ha sido más alto que la ocupación diaria de las UCI. También se aprecia que, en abril de 2021, la ocupación hospitalaria diaria alcanzó su nivel más elevado desde enero de 2020.

Insights

Los datos presentados en los dashboards sobre la pandemia de COVID-19 proporcionan una visión integral de la situación a nivel mundial, lo que nos permite extraer valiosas conclusiones y orientar futuras acciones.

- Impacto de la pandemia: La pandemia ha tenido un impacto sustancial tanto en la salud como en la economía a nivel global, subrayando la necesidad de respuestas efectivas a nivel mundial.
- Papel crucial de la vacunación: La vacunación se revela como un factor clave en la reducción de casos y muertes por COVID-19, reforzando la importancia de programas de inmunización amplios y efectivos.
- Población urbana y propagación: Se observa una correlación entre la alta población urbana y el aumento de casos y muertes por COVID-19, debido a la mayor densidad de población que facilita la propagación del virus.
- Acceso a servicios básicos e internet: Existe un incremento en casos y muertes en áreas con limitado acceso a servicios básicos e internet. Esto podría vincularse a un menor desarrollo, lo que probablemente conduce a una movilidad reducida y, en última instancia, a una menor propagación del virus.
- Condiciones ambientales: Las condiciones climáticas, como la temperatura, influyen en la propagación del virus. Las temperaturas más altas se asocian con menos contagios, sugiriendo que el virus se reproduce con mayor facilidad en climas fríos. Además, la altitud parece afectar la mortalidad del COVID-19, lo que podría estar relacionado con la disponibilidad de oxígeno y otros factores.

Estas conclusiones generales apuntan a una clara dirección para futuras investigaciones y acciones preventivas. Además, se pueden destacar observaciones específicas:

- Ausencia de correlación directa: No se ha establecido una relación clara entre la cantidad de casos y muertes por COVID-19, ya que la pandemia ha afectado a diferentes países en oleadas, y no todos estaban preparados para enfrentar la magnitud de los contagios.
- Densidad de población en áreas urbanas: La mayor densidad de población en áreas urbanas contribuye a la propagación del virus, explicando por qué los países con poblaciones urbanas más grandes suelen tener una mayor incidencia de casos y muertes por COVID-19.
- Investigación sobre condiciones ambientales: Las implicaciones de las condiciones ambientales en la propagación del COVID-19 requieren investigaciones más profundas para desarrollar estrategias de prevención más efectivas.

Basándonos en estas conclusiones y observaciones, se proponen las siguientes recomendaciones:

- Campañas de vacunación continuas: Los gobiernos deben mantener y ampliar las campañas de vacunación para aumentar la cobertura en todos los grupos de edad.
- Investigación ambiental: Los gobiernos y organismos pertinentes deben llevar a cabo investigaciones para comprender mejor las implicaciones de las condiciones ambientales en la propagación del COVID-19, lo que permitirá el desarrollo de estrategias de prevención más efectivas. Estas acciones son esenciales para proteger la salud y la economía a nivel global.

Conclusiones

Este proyecto de investigación demostró ser dinámico y adaptable en su ejecución. Aunque inicialmente seguimos una planificación sólida, a medida que avanzamos en el máster, abrazamos la flexibilidad de ajustar nuestras ideas y enfoques en función de la dirección que nos marcaban los datos y los resultados. Este enfoque exploratorio nos llevó a tomar decisiones informadas para alinear nuestro trabajo con los objetivos finales.

En el transcurso de la investigación, exploramos diversas posibilidades, a menudo descartando ideas iniciales en favor de opciones más alineadas con nuestros objetivos. En cierto momento, consideramos la inclusión de tecnologías como BigML, pero con el tiempo, y a medida que profundizamos en el entorno de Cloud Computing y sus capacidades de automatización, nos sentimos más cómodos centrándonos en la nube y ajustando nuestras estrategias de acuerdo con ello.

Este enfoque evolutivo y adaptable no solo resultó esencial para optimizar nuestro proyecto, sino que también nos permitió extraer valiosos insights que arrojaron luz sobre la pandemia de COVID-19 y sus relaciones con una serie de factores. Algunos de los hallazgos más significativos incluyen:

- La falta de correlación directa entre la cantidad de casos y muertes por COVID-19, ya que la pandemia afectó a diferentes países en oleadas y no todos estaban preparados para la magnitud de los contagios.
- La importancia crítica de la vacunación en la reducción de casos y muertes por COVID-19, resaltando la necesidad de programas de vacunación efectivos a nivel mundial.
- La influencia de la alta población urbana en la propagación del virus debido a la mayor densidad de población en áreas urbanas.
- La relación entre el acceso limitado a servicios básicos e Internet y un aumento en los casos y muertes por COVID-19 en áreas menos desarrolladas.
- Las implicaciones de las condiciones ambientales, como la temperatura y la altitud, en la propagación y mortalidad del COVID-19, que sugieren la necesidad de investigaciones adicionales.

En última instancia, este proyecto de investigación reafirma la importancia de mantener una mente abierta y explorar constantemente nuevas posibilidades. La investigación nos llevó en direcciones inesperadas, lo que nos permitió adaptarnos y aprovechar al máximo las oportunidades que se presentaron en el camino. Este enfoque no solo optimizó nuestro proyecto, sino que también contribuyó a la comprensión y abordaje de la pandemia de COVID-19 en el contexto global.

Bibliografía

1. **Organización Mundial de la Salud**. (s. f.). *Datos globales de COVID-19 de la OMS* [Datos en formato CSV]. acceso aquí.
2. **Organización Mundial de la Salud**. (s. f.). *Tabla global de datos de COVID-19 de la OMS* [Datos en formato CSV]. acceso aquí.
3. **Organización Mundial de la Salud**. (s. f.). *Datos de vacunación de la OMS* [Datos en formato CSV]. acceso aquí.
4. **Organización Mundial de la Salud**. (s. f.). *Metadatos de vacunación de la OMS* [Datos en formato CSV]. acceso aquí.
5. **European Centre for Disease Prevention and Control**. (s. f.). *Datos de pruebas ECDC* [Página web]. acceso aquí.
6. **European Centre for Disease Prevention and Control**. (s. f.). *Datos de hospitalización y ocupación de UCI de ECDC* [Datos en formato CSV]. acceso aquí.
7. **World Development Indicators (WDI)** (s. f.). *Datos de Tabla paises socioeconómicos y demográficos [*Datos en formato CSV]. acceso aquí.

Anexos

Entregable 1

En la primera entrega de nuestro proyecto de máster sobre el Covid-19, se abordó la definición del proyecto y su viabilidad, que incluyó análisis interno y externo. Luego, se procedió con la definición detallada del proyecto y el perfilado de las bases de datos seleccionadas. Posteriormente, se presentó la planificación que abarcó recursos, tiempos estimados y alcance. Finalmente, se concluyó con una reflexión sobre el trabajo realizado por el equipo en este entregable. La gestión se basó en metodologías AGILE y equipos de alto rendimiento con relaciones horizontales entre los miembros del equipo.

Acceso al documento completo en nuestro GitHub

Entregable 2

En esta segunda entrega de nuestro Trabajo de Fin de Máster (TFM), se llevó a cabo la depuración, transformación y carga de diversas fuentes de datos relacionadas con el COVID-19. Estos datos incluyeron 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. Se integraron diferentes fuentes para obtener una visión más completa de la situación de la pandemia. Además, se formularon hipótesis para el próximo análisis de datos, respaldadas con bibliografía académica especializada. A pesar de las dificultades encontradas durante el proceso, se lograron cumplir con las tareas establecidas en esta segunda entrega del TFM. La investigación contribuyó al entendimiento de la pandemia de COVID-19, y se destacó el uso de Python como una herramienta eficaz en estos procesos.

Acceso al documento completo en nuestro GitHub

Entregable 3

Esta entrega implica la elaboración de los dashboards iniciales con Power BI, que son de carácter descriptivo de los datos y se realizará un análisis de las diferencias observadas. El archivo Power BI (.pbix) desarrollado por el equipo de trabajo se presenta en esta entrega e incluye varias visualizaciones. En el "Mapa de Casos y Muertes," se permite el filtrado por continente y se muestran casos y muertes por cada 100,000 habitantes. La visualización "Casos vs. Muertes" relaciona la cantidad de casos con muertes por 100,000 habitantes, destacando la falta de correlación debido a las distintas olas de contagio. También, se presentan gráficos que exploran la relación entre casos, muertes y la vacunación.

Además, se ofrecen datos relacionados con las pruebas realizadas por país en Europa, indicando que Alemania lidera en términos de pruebas de COVID-19 realizadas. Se señala una correlación moderada entre el número de pruebas y nuevos casos. En el caso de las UCI (Unidades de Cuidados Intensivos), se observa que los ingresos hospitalarios superaron a las admisiones en UCI, y se destaca un aumento en la ocupación hospitalaria en enero de 2021.

En las conclusiones generales, se reconoce la importancia de diseñar, validar y probar los tableros antes de lanzar un producto. Se identifican desafíos en algunos tableros, como la dificultad de uso y la falta de filtros

clave. Además, se planea complementar los datos con información adicional de cada país, como cantidad de turistas, temperatura promedio y otros indicadores de desarrollo y salud.

Acceso al documento completo en nuestro GitHub