



ANÁLISIS DE DATOS

TDSD

ASIGNATURA:

ANÁLISIS DE DATOS

PROFESOR:

Ing. Lorena Chulde / Ing. Juan Pablo
Zaldumbide

FECHA:

13 - 08 - 2024

PERÍODO ACADÉMICO:

2024-A

PROYECTO FINAL – BIMESTRE 2



Integrantes:

Karla Rodriguez

Isaac Quinapallo

Angel Villamil

2024-A

Objetivo:

Aplicar los conocimientos adquiridos sobre análisis de datos, extracción, limpieza, transformación, visualización de datos mediante la aplicación Power BI.

Proyecto Final Análisis de Datos

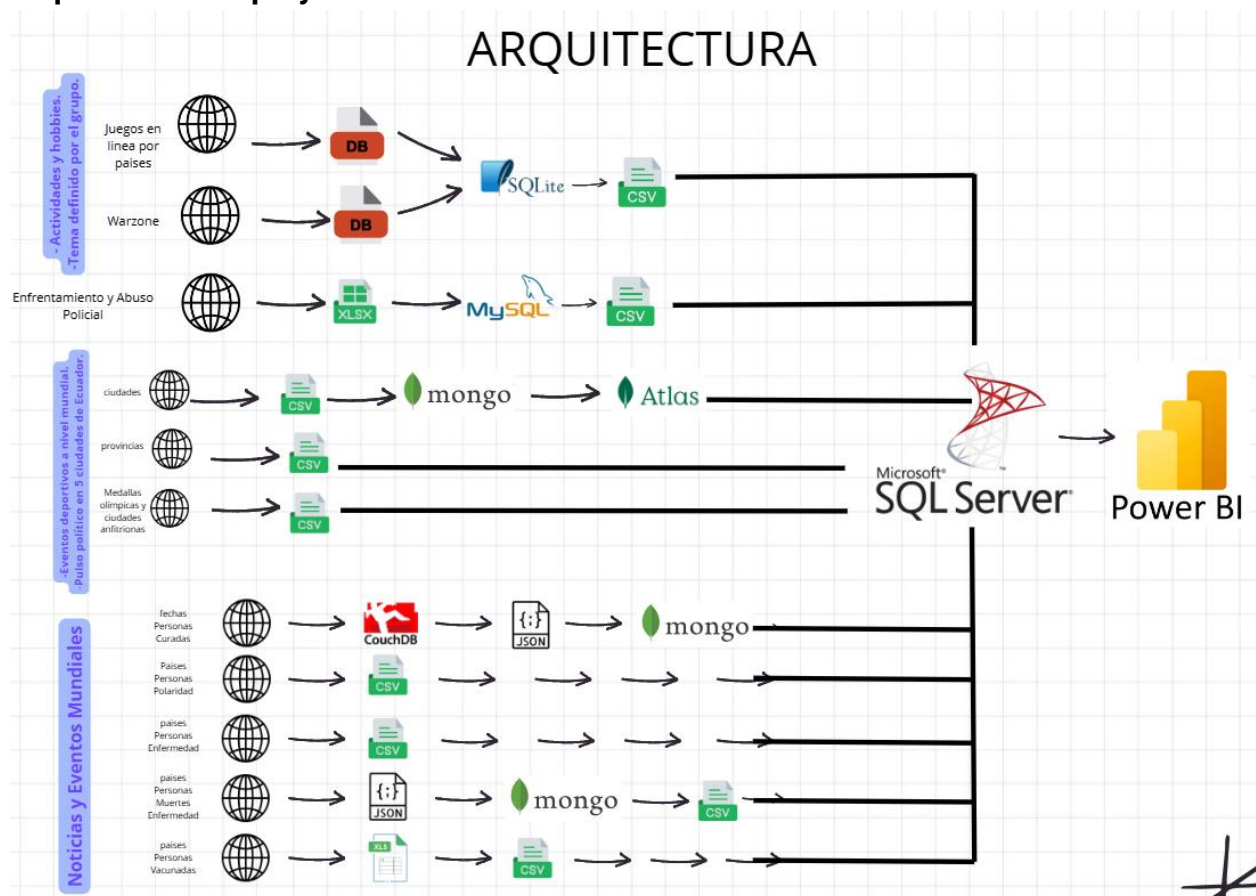
3. URL del Repositorio de GitHub (25%)

- Repositorio de GitHub con todos los script utilizados y explicando el proceso completo y las fuentes de datos.

<https://github.com/isaacquinapallo/ProyectoFinalAnalisisDeDatos.git>

Nota: Se tomará en cuenta la innovación en cualquiera de las etapas del proyecto, es un plus para la nota final.

Arquitectura del proyecto



DESARROLLO

Objetivos Generales

Recopilar datos a través de diferentes herramientas y técnicas para su análisis gráfico.

Objetivos Específicos

- Analizar los casos de estudio para el análisis
- Crear una arquitectura para recopilación de datos
- Desarrollar un cronograma de actividades.
- Dividir las actividades y herramientas a utilizar por integrante tomando como referencia la arquitectura.
- Recopilar datos de las fuentes establecidas en la arquitectura
- Realizar un análisis grafico usando las herramientas establecidas en la arquitectura.
- Documentar todo el análisis realizado por cada integrante del grupo.

Recursos y Herramientas Utilizadas

Apache CouchDB

Conocido también como CouchDB, es un gestor de bases de datos de código abierto que se clasifica como una base NoSQL diseñada para simular la web. Utiliza JSON para almacenar datos y emplea JavaScript como lenguaje para realizar consultas a través de API.

En lugar de almacenar datos y sus relaciones en tablas, CouchDB organiza cada base de datos como una colección de documentos independientes, donde cada documento contiene sus propios datos y su esquema.

El protocolo de replicación de datos de CouchDB se implementa en diversos proyectos y productos, abarcando una amplia gama de entornos informáticos, desde clústeres de servidores distribuidos a nivel global hasta dispositivos móviles y navegadores web.



Figura 1 CouchDB

Mongo DB

MongoDB es una base de datos libre y distribuida, basada en documentos y de uso general, diseñada para desarrolladores de aplicaciones modernas debido a su capacidad de escalabilidad y su tecnología adaptada a la era de la nube.

El modelo de documentos de MongoDB es intuitivo y fácil de aprender, ofreciendo a los desarrolladores todas las funciones necesarias para cumplir con requisitos complejos a cualquier escala. Además, cuenta con drivers para más de diez lenguajes de programación y está respaldada por una comunidad de desarrolladores activa.



Figura 2 MongoDB

Power BI

Power BI es una herramienta de software de Microsoft diseñada para el análisis de datos en el entorno empresarial. Destaca por su alto rendimiento en la visualización de resultados, gracias a sus gráficos dinámicos y su interfaz fácil de usar. Power BI incluye funciones de inteligencia empresarial que facilitan la comprensión de gráficos y visualizaciones, permitiendo crear informes y paneles. Además, Power BI ofrece protección integral para la transmisión de datos y se conecta a servicios en la nube y otros servicios de Microsoft.



SQL Server

SQL Server es un sistema de gestión de bases de datos relacionales (RDBMS) desarrollado por Microsoft, diseñado específicamente para el entorno empresarial. Utiliza T-SQL (Transact-SQL), un conjunto de extensiones de programación de Sybase y Microsoft que agrega varias características al SQL estándar, como control de transacciones, manejo de excepciones y errores, procesamiento de filas y uso de variables declaradas.



Mongo DB Atlas

Una base de datos global basada en la nube y totalmente gestionada por MongoDB que integra modelos de datos similares a JSON, ofrece indexación y búsqueda avanzadas, y permite escalabilidad elástica, mientras automatiza las tareas administrativas que consumen mucho tiempo.



SQLite

SQLite es un software libre que almacena datos de manera eficiente en dispositivos con recursos limitados. Soporta consultas SQL básicas y avanzadas y es compatible con dispositivos móviles y de escritorio, facilitando la portabilidad de datos sin procesos complicados.



Temas Generales de casos de estudio

Se obtendrá dashboards de los siguientes casos de estudio:

Pulso político en 20 ciudades principales de Ecuador

En abril de 2021, se celebrarán elecciones en Ecuador, lo que hace relevante explorar datos para entender la opinión pública sobre el clima político. Estos datos se recopilaban a través de Twitter, una red social frecuentemente utilizada por los candidatos presidenciales en Ecuador.

Pulso político por provincias en Ecuador

Aunque Ecuador no es conocido por ser un país donde Twitter es ampliamente utilizado, en el ámbito político, diversas redes sociales, incluida Twitter, son empleadas para expresar opiniones o apoyar a candidatos preferidos. Esto despertó el interés en conocer cuán utilizada es esta red social en distintas provincias. La recopilación de datos sobre el uso de Twitter en el contexto político en cada provincia ha arrojado resultados satisfactorios, aclarando la duda sobre el porcentaje de provincias interesadas en la política.

Juegos en línea por países y el juego Warzone

Actualmente, los juegos en línea son extremadamente populares, y con la facilidad de acceso a internet, cada país cuenta con un gran grupo de jugadores. Utilizando internet y diversas plataformas, se pueden obtener datos sobre las clasificaciones de los juegos, estadísticas de jugadores por país y los ingresos que la industria de los videojuegos genera para la economía de cada nación.

Tema definido por el estudiante

Los datos fueron recopilados mediante un archivo .xlsx exportado de data.world, enfocándose en tiroteos fatales cometidos por la policía en Estados Unidos. Este tipo de recopilación es común debido al elevado número de muertes en diferentes ciudades del país.

Eventos o noticias mundiales

Se obtuvieron datos de Kaggle relacionados con las personas vacunadas contra el Covid-19, con el fin de comprender mejor el progreso mundial un año después de haberse declarado la pandemia. También se obtuvieron datos de Statista, que proporcionan información sobre la evolución del Covid-19 desde el 3 de febrero de 2020 hasta el 12 de marzo de 2021. Finalmente, se recopilaban datos de Facebook y de algunos de los portales de noticias más populares.

Cronograma de actividades

Fechas	Actividad	Responsable	Horas
26 de julio - 29 de julio, 2024	Planificación del Proyecto		
	Selección de las herramientas a utilizar.	Karla Rodríguez	3
	Distribución de temas por miembro del equipo.	Isaac Quinapallo	3
	Selección de bases SQL y NoSQL.	Ángel Villamil	2
	Establecimiento del tiempo de investigación.	Karla Rodríguez	2

Fechas	Actividad	Responsable	Horas
30 de julio - 4 de agosto, 2024	Recolección de Datos		
	Búsqueda de datos en diferentes fuentes.	Isaac Quinapallo	10
	Creación de scripts para recopilar datos.	Ángel Villamil	2
	Ejecución de scripts para recopilar datos.	Karla Rodríguez	20
	Limpieza de datos.	Isaac Quinapallo	6

Fechas	Actividad	Responsable	Horas
5 de agosto - 7 de agosto, 2024	Concentración de Datos		
	Cargar los datos recopilados a sus respectivas bases de datos.	Ángel Villamil	6
	Creación de índices en el servidor Elasticsearch.	Karla Rodríguez	6
	Creación de scripts para migración a la herramienta Elasticsearch.	Isaac Quinapallo	5

Fechas	Actividad	Responsable	Horas
8 de agosto - 10 de agosto, 2024	Creación de Dashboards		
	Importación de los datos recopilados a las herramientas de visualización.	Karla Rodríguez	4
	Limpieza de datos.	Isaac Quinapallo	6
	Creación de visualizaciones.	Ángel Villamil	8

Fechas	Actividad	Responsable	Horas
11 de agosto - 13 de agosto, 2024	Resultados y Documentación		
	Interpretación de cada visualización.	Isaac Quinapallo	12
	Conclusiones y recomendaciones de los resultados obtenidos.	Karla Rodríguez	8
	Documentación de todo el proceso y resultados finales.	Ángel Villamil	10

Casos de Estudios

Pulso político en 20 ciudades de Ecuador

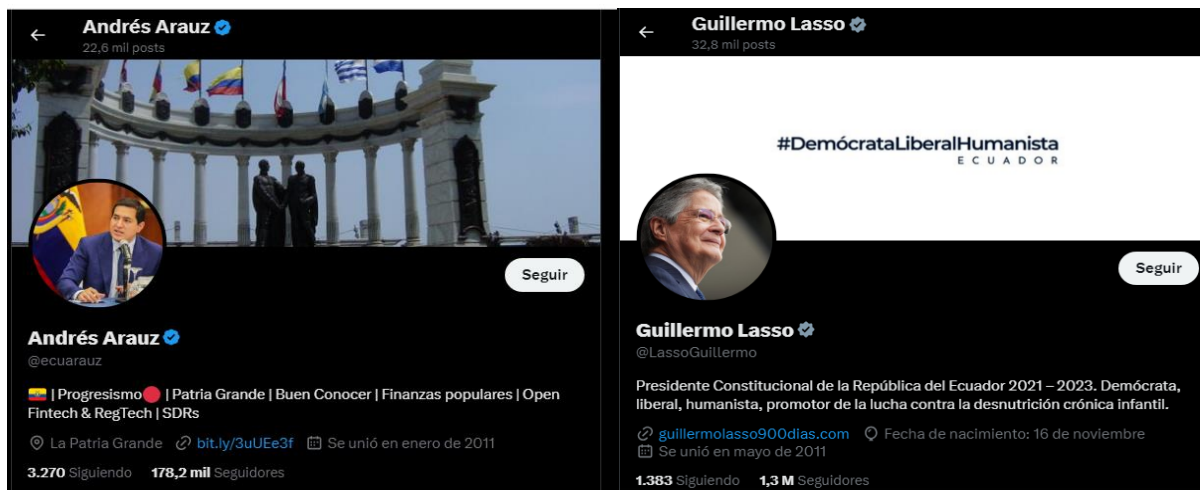
Este informe presenta el análisis del pulso político en 20 ciudades de Ecuador, basado en la recolección de datos de Twitter. La información se centró en las menciones y retweets relacionados con los candidatos de las elecciones presidenciales de Ecuador de 2021 que se realizaron el 7 de febrero de 2021 para elegir al presidente constitucional y vicepresidente constitucional de la República del Ecuador para el período 2021-2025, con los candidatos Guillermo Lasso, candidato de la alianza entre CREO y PSC, contra Andrés Arauz, candidato de la coalición UNES. Los datos se almacenaron y gestionaron utilizando MongoDB y MongoDB Atlas, y se visualizó mediante Power BI.

Objetivos

- Extraer datos de Twitter utilizando palabras clave relacionadas con los candidatos.
- Utilizar MongoDB y MongoDB Atlas para almacenar y manejar los datos recopilados.
- Emplear Power BI para crear visualizaciones que muestren el pulso político en diferentes ciudades.
- Identificar patrones en la actividad política en Twitter y proporcionar insights sobre el pulso político.

Métodos y Herramientas Utilizadas

- Utilización de API de Twitter para la recolección de datos mediante palabras clave ("Arauz", "Lasso").



- Los datos se almacenaron en MongoDB y se migraron a MongoDB Atlas para análisis y visualización.

The image displays two screenshots of MongoDB management tools. The top screenshot shows the MongoDB Compass interface for a database named 'proyecto.pulso politico'. It shows a collection of documents with fields like '_id', 'created_at', 'text', 'display_text_range', 'source', 'in_reply_to_status_id', 'in_reply_to_user_id', 'in_reply_to_screen_name', 'user', 'contributors', 'is_quote_status', 'quote_count', 'reply_count', and 'retweet_count'. A notification at the bottom indicates 'Import completed. 8476 documents imported.' The bottom screenshot shows the MongoDB Atlas interface for the same database, displaying the 'proyecto.pulso politico20' collection with storage and logical data sizes, total documents, and indexes. It also shows a query filter and query results.

- Power BI fue utilizado para la creación de gráficos y mapas interactivos que presentan el número de tweets y retweets.

```
[3]: import pandas as pd
import csv

# Lista para almacenar las filas válidas
valid_rows = []

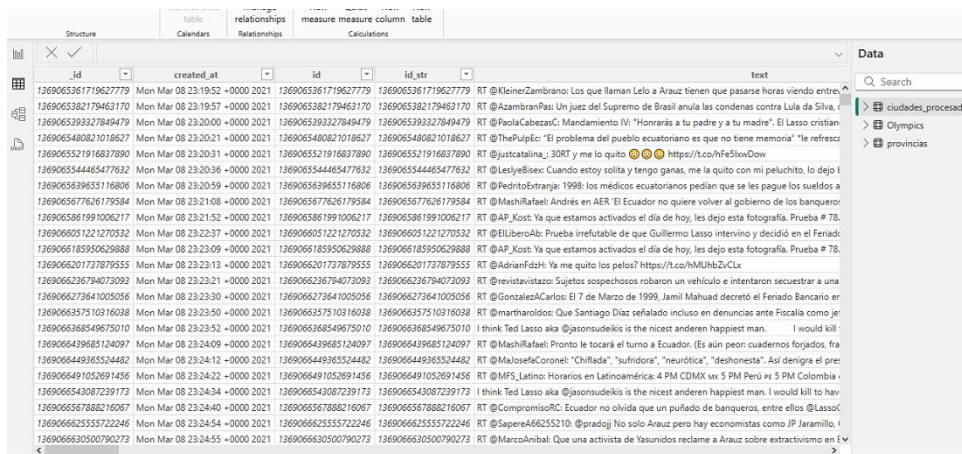
# Abrir el archivo CSV y procesarlo línea por línea
with open('ciudades.csv', newline='', encoding='utf-8') as csvfile:
    reader = csv.reader(csvfile)
    header = next(reader) # Leer la cabecera
    for row in reader:
        # Verificar si la fila tiene el número correcto de columnas
        if len(row) == len(header):
            valid_rows.append(row)

# Convertir la lista de filas procesadas en un DataFrame
df = pd.DataFrame(valid_rows, columns=header)

# Mostrar las primeras filas del DataFrame resultante
print(df.head())

# Guardar el DataFrame procesado en un nuevo archivo CSV
df.to_csv('ciudades_procesado.csv', index=False)
print("Archivo procesado guardado como 'ciudades_procesado.csv'.")
```

	_id	created_at	id
0	1369065352076931079	Mon Mar 08 23:19:50 +0000 2021	1369065352076931079
1	1369065352794148872	Mon Mar 08 23:19:50 +0000 2021	1369065352794148872
2	1369065357420294148	Mon Mar 08 23:19:51 +0000 2021	1369065357420294148
3	1369065357676339201	Mon Mar 08 23:19:51 +0000 2021	1369065357676339201
4	1369065361719627779	Mon Mar 08 23:19:52 +0000 2021	1369065361719627779



id	created_at	id_str	text
1369065361719627779	Mon Mar 08 23:19:52 +0000 2021	1369065361719627779	RT @KleinerZambrano: Los que llaman Lelo a Arauz tienen que pasarse horas viendo entre,
1369065382179463170	Mon Mar 08 23:19:57 +0000 2021	1369065382179463170	RT @AzambraPas: Un juez del Supremo de Brasil anula las condenas contra Lula da Silva,
136906539327849479	Mon Mar 08 23:20:00 +0000 2021	136906539327849479	RT @PaolaCabezasC: Mandamiento IV: "Honrarás a tu padre y a tu madre". El Lasso cristian,
1369065480821018627	Mon Mar 08 23:20:21 +0000 2021	1369065480821018627	RT @ThePulpeC: "El problema del pueblo ecuatoriano es que no tiene memoria" "le refresc
1369065521916837890	Mon Mar 08 23:20:31 +0000 2021	1369065521916837890	RT @justicialista_30RT y me lo quito 🤔🤔🤔 https://t.co/nHfellowDow
1369065544465477632	Mon Mar 08 23:20:36 +0000 2021	1369065544465477632	RT @LeijebBuen: Cuando estoy solito y tengo ganas, me la quito con mi peluchito, lo dejo l
1369065639655116806	Mon Mar 08 23:20:59 +0000 2021	1369065639655116806	RT @PedritoStrange: 1998: los médicos ecuatorianos piden que se les pague los sueldos a
1369065677626179584	Mon Mar 08 23:21:08 +0000 2021	1369065677626179584	RT @MashRafael: Andrés en AER: El Ecuador no quiere volver al gobierno de los banquero;
1369065861991006217	Mon Mar 08 23:21:52 +0000 2021	1369065861991006217	RT @AP_Kout: Ya que estamos activados el día de hoy, les dejo esta fotografía: Prueba # 78.
1369066051221270532	Mon Mar 08 23:22:37 +0000 2021	1369066051221270532	RT @EliUbersál: Prueba irrefutable de que Guillermo Lasso intervino y decidió en el Feriadi
136906618590629888	Mon Mar 08 23:23:09 +0000 2021	136906618590629888	RT @AP_Kout: Ya que estamos activados el día de hoy, les dejo esta fotografía: Prueba # 78.
1369066201737879555	Mon Mar 08 23:23:13 +0000 2021	1369066201737879555	RT @AdrianFadi: Ya me quito los pelos? https://t.co/nMjHbZcCu
1369066236794073093	Mon Mar 08 23:23:21 +0000 2021	1369066236794073093	RT @nevisativato: Sujetos sospechosos robaron un vehículo e intentaron secuestrar a una
1369066273641005056	Mon Mar 08 23:23:30 +0000 2021	1369066273641005056	RT @GonzalezCarlos: El 7 de Marzo de 1999, Jamil Mahud decretó el Feriado Bancario er
1369066357510216038	Mon Mar 08 23:23:50 +0000 2021	1369066357510216038	RT @martharoldon: Que Santiago Díaz señalado incluso en denuncias ante Fiscalía como je
1369066368549675010	Mon Mar 08 23:23:52 +0000 2021	1369066368549675010	I think Ted Lasso aka @jasonudekis is the nicest anderen happiest man. I would kill
1369066439685124097	Mon Mar 08 23:24:09 +0000 2021	1369066439685124097	RT @MashRafael: Pronto le tocará el turno a Ecuador. (Ei aún peon: cuadernos forjados, fra
1369066449365524482	Mon Mar 08 23:24:12 +0000 2021	1369066449365524482	RT @MaloselaCoronel: "Chiflada", "sufridora", "neurótica", "deshonesta". Así designa el pri
1369066491052691456	Mon Mar 08 23:24:22 +0000 2021	1369066491052691456	RT @MPS_Latino: Horarios en Latinoamérica: 4 PM CDMX vs 5 PM Perú vs 5 PM Colombia
1369066543087239173	Mon Mar 08 23:24:34 +0000 2021	1369066543087239173	I think Ted Lasso aka @jasonudekis is the nicest anderen happiest man. I would kill to hav
1369066567888216067	Mon Mar 08 23:24:44 +0000 2021	1369066567888216067	RT @CompromisoRC: Ecuador no olvida que un puñado de banqueros, entre ellos @LassoC
1369066625555722246	Mon Mar 08 23:24:54 +0000 2021	1369066625555722246	RT @SapereA66255210: @pradaji No solo Arauz pero hay economistas como JP Jaramillo, i
1369066630500790273	Mon Mar 08 23:24:55 +0000 2021	1369066630500790273	RT @MarcoAnibal: Que una activista de Yesunidos reclame a Arauz sobre extractivismo en [

Datos Recopilados

- Cantidad de Documentos Se recolectaron aproximadamente 32,000 documentos.
- Datos Incluidos: Número de tweets, retweets, fechas y contenido relevante de cada tweet.

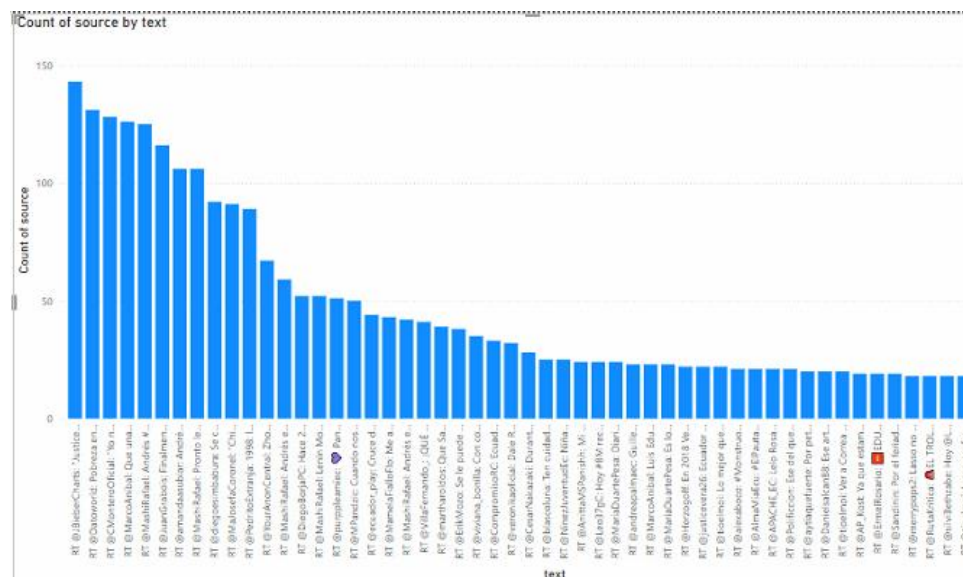
Análisis de Datos

1. La ciudad de Alausí mostró el mayor número de retweets en el ámbito político, indicando una mayor actividad en comparación con otras ciudades.
2. Conteo de Tweets: El análisis reveló que el candidato Lasso tiene un mayor número de tweets en comparación con Andrés Arauz, lo que sugiere una mayor visibilidad o apoyo en la red social.
3. Pulso Político General: Se observó un alto nivel de actividad política en Twitter, con muchas personas realizando retweets sobre la situación del país y los candidatos.

Visualizaciones

- Número de Tweets por Ciudad

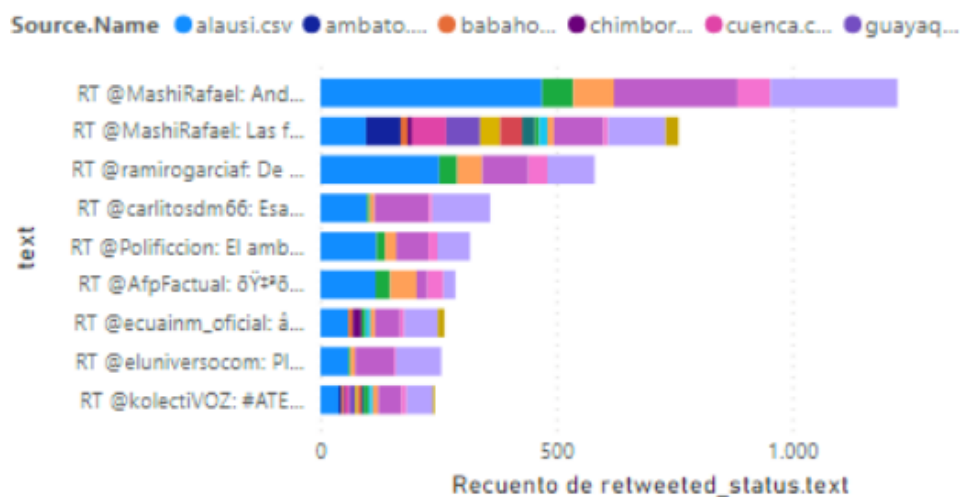
El gráfico muestra que Alausí, Quevedo y Napo son las ciudades con mayor cantidad de tweets relacionados con el pulso político.



- Número de Retweets sobre “Arauz”

La visualización destaca las ciudades con mayor número de retweets utilizando la palabra “Arauz”.

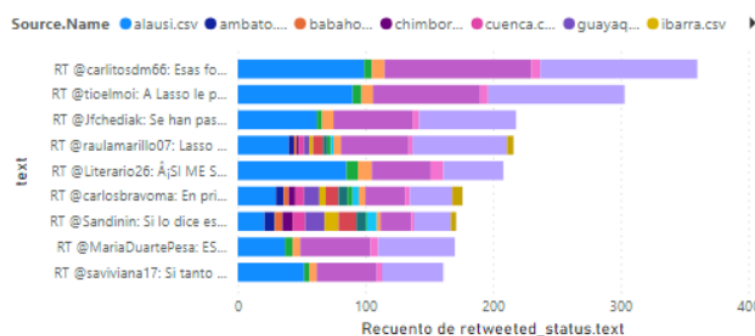
Arauz Nombrado en Tweeter



- Número de Retweets sobre “Lasso”

Se presenta la cantidad de retweets que mencionan a Daniel Lasso en cada ciudad.

Lasso Nombrado en Tweeter



Resultados Obtenidos

- Las ciudades de Alausí, Quevedo y Napo están entre las más activas en términos de número de tweets y retweets sobre el pulso político.
- Lasso tiene una mayor presencia en términos de conteo de tweets, mientras que la actividad en torno a Andrés Arauz es menor en comparación.
- Se identificaron patrones regionales en la actividad política en Twitter, proporcionando una visión sobre la distribución del apoyo a los candidatos.

Conclusiones y Recomendaciones

- Conclusiones: Este informe proporciona una visión detallada del pulso político en Ecuador a través de Twitter, utilizando herramientas avanzadas para el análisis y visualización de datos. Los resultados ofrecen una base sólida para comprender las dinámicas políticas y apoyar las estrategias de comunicación de los candidatos.
- Recomendaciones: Se recomienda un seguimiento continuo de las menciones y retweets para observar cómo evolucionan las tendencias políticas a medida que se acercan las elecciones. Además, se debe considerar la implementación de estrategias para aumentar el compromiso y la visibilidad de los candidatos con menor actividad.

Desafíos y Problemas Encontrados

- Limitaciones en la Recolección de Datos, Algunas provincias tuvieron baja actividad en Twitter o acceso limitado a internet, lo que afectó la cantidad de datos recolectados.
- Problemas Técnicos* La conversión de datos a SQL Server a veces generó problemas con los formatos de los datos, lo que dificultó el análisis preciso.

Recursos y Herramientas

- Fuentes de Datos: Twitter API.
- Herramientas de Almacenamiento y Gestión: MongoDB, MongoDB Atlas.
- Herramientas de Visualización: Power BI.

Pulso Político por Provincias en Ecuador

Este informe analiza el pulso político en las 24 provincias de Ecuador basado en datos recolectados de Twitter. Utilizando un script con credenciales de desarrollador para acceder a la API de Twitter, se recopilaban datos filtrados por términos clave relacionados con las elecciones de 2021 y temas de interés político. Se empleó la geolocalización para obtener datos precisos y se almacenaron aproximadamente 30,000 documentos. Las visualizaciones se realizaron con Power BI para representar el análisis de manera efectiva.

2. Objetivos

- **Recolectar Datos:** Utilizar la API de Twitter para recopilar información sobre el pulso político en las provincias de Ecuador.
- **Almacenar y Gestionar Datos:** Manejar los datos recopilados en una base de datos para su análisis.
- **Visualizar Información:** Crear visualizaciones que muestran el total de retweets, tweets favoritos y la actividad política por candidato en cada provincia.
- **Analizar Patrones:** Identificar patrones en la actividad política y la popularidad de los candidatos en diferentes provincias.

3. Métodos y Herramientas Utilizadas

- **Extracción de Datos:** Recolección de datos a través de la API de Twitter utilizando un script especializado. Se incluyeron palabras clave como "candidatos 2021", "votaciones 2021", "pulso político", entre otras.



- **Visualización:** Uso de Power BI para crear gráficos y dashboards interactivos que representan la actividad política.

4. Datos Recopilados

- **Cantidad de Documentos:** Aproximadamente 30,000 documentos.
- **Datos Incluidos:** Retweets, tweets favoritos, menciones de candidatos, y actividad por provincia.

5. Análisis de Datos

1. **Total de Retweets por Provincia:** Las provincias con mayor actividad política en términos de retweets son Pichincha, Azuay y Guayas. Estas provincias muestran un alto volumen de discusiones y retweets sobre temas políticos. En contraste, las

provincias con menor actividad política son Carchi, Cotopaxi y Galápagos, donde el número de retweets es significativamente bajo.

2. Total de Tweets Favoritos por Candidato:

-Andrés Arauz es el candidato presidencial con mayor número de tweets favoritos, indicando un notable nivel de apoyo y reconocimiento en la red social.

-Guillermo Lasso sigue a Arauz en popularidad, con muchos tweets agregados a favoritos.

-Yaku Pérez, Lucio Gutiérrez y Xavier Hervas también tienen menciones y tweets favoritos, aunque en menor cantidad en comparación con Arauz y Lasso.

6. Visualizaciones

• Total de Retweets por Provincia:

- Gráfico que muestra las provincias con el mayor y menor número de retweets políticos.
- **Pichincha, Azuay y Guayas** destacan como las provincias con la mayor actividad.
- **Carchi, Cotopaxi y Galápagos** presentan menor actividad política.

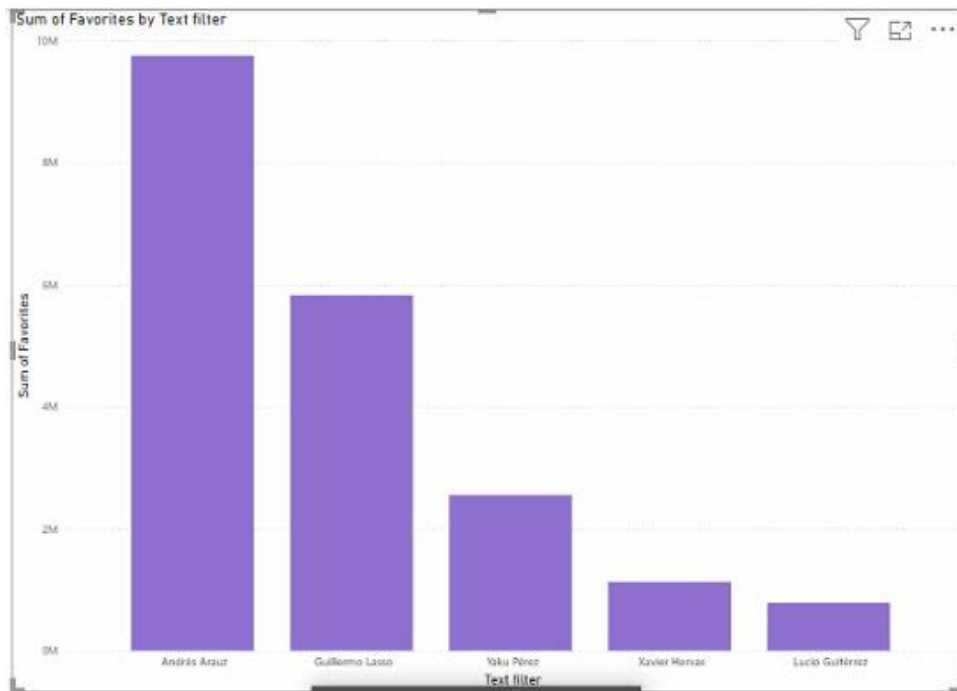
Retweet count and Place

Place ● Chimborazo ● Azuay ● Bolívar ● Bolívar ● Cañar ● Carchi

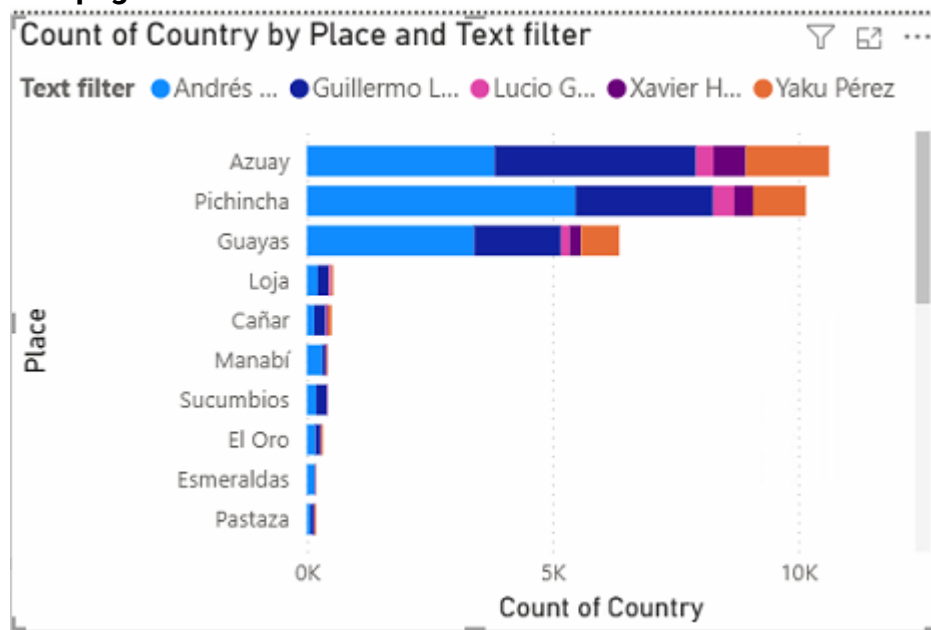


• Total de Tweets Favoritos por Candidato:

- Gráfico que ilustra el número de tweets favoritos por candidato presidencial.
- **Andrés Arauz** lidera en cantidad de tweets favoritos, seguido por **Guillermo Lasso**.
- **Yaku Pérez, Lucio Gutiérrez y Xavier Hervas** tienen una presencia menor en la sección de favoritos.



- **Total de Tweets por Provincia y Candidato:**
 - Visualización que compara el número total de tweets sobre cada candidato en diferentes provincias.
 - **Pichincha, Azuay y Guayas** muestran una mayor cantidad de tweets relacionados con los candidatos, mientras que **Carchi, Cotopaxi y Galápagos** tienen menos actividad.



7. Resultados Obtenidos

- **Pichincha, Azuay y Guayas** son las provincias más activas en términos de retweets y tweets políticos.
- **Andrés Arauz** es el candidato más destacado en tweets favoritos, seguido de cerca por **Guillermo Lasso**.
- Las provincias menos activas en política en Twitter son **Carchi, Cotopaxi y Galápagos**, lo que sugiere un menor interés o participación en temas políticos.



8. Conclusiones y Recomendaciones

- **Conclusiones:** pulso político en Ecuador, destacando las áreas con mayor y menor actividad en Twitter y el apoyo a los candidatos presidenciales. Las visualizaciones y análisis permiten una comprensión más clara de la dinámica política en las distintas provincias.
- **Recomendaciones:** Se recomienda a los candidatos enfocarse en provincias con baja actividad para aumentar su presencia y apoyo. Además, se debería considerar realizar campañas específicas para aumentar el interés en regiones con menor participación política.

9. Desafíos y Problemas Encontrados

- **Variabilidad en la Actividad:** La diferencia en la actividad política entre provincias puede deberse a diversos factores, como el acceso a Internet y el interés en temas políticos.
- **Precisión de Geolocalización:** Aunque se utilizó geolocalización para recolectar datos, la precisión de la ubicación puede afectar la representación exacta de la actividad en algunas provincias.

10. Recursos y Herramientas

- **Fuentes de Datos:** Twitter API.
- **Herramientas de Almacenamiento:** Base de datos estructurada.
- **Herramientas de Visualización:** Power BI.

Eventos deportivos a nivel mundial. (Juegos Olímpicos y Población)

el rendimiento y la participación en los Juegos Olímpicos de Verano a nivel mundial, con un enfoque en la acumulación de medallas por país y la relación entre el número de ciudades anfitrionas y el éxito olímpico. Se examina cómo los países que participan más activamente en los Juegos Olímpicos tienden a obtener más medallas, y se estudia la influencia de la inversión en infraestructura deportiva.

2. Objetivos

- **Recolectar Datos:** Obtener información sobre el medallero acumulado por país y los récords de ciudades anfitrionas por año.
- **Analizar el Rendimiento:** Evaluar la relación entre el número de participaciones y la cantidad de medallas obtenidas por país.
- **Visualizar Información:** Crear visualizaciones que muestren el desempeño de los países en los Juegos Olímpicos y el impacto de las ciudades anfitrionas en el medallero.

3. Métodos y Herramientas Utilizadas

- **Extracción de Datos:** Se recopiló información histórica sobre los Juegos Olímpicos de Verano, incluyendo medallas por país y datos sobre ciudades anfitrionas.
- **Almacenamiento de Datos:** Los datos se gestionaron en bases de datos estructuradas para su análisis.
- **Visualización:** Se utilizaron herramientas de visualización de datos como Power BI para representar gráficamente la información sobre medallas y ciudades anfitrionas.

4. Datos Recopilados

- **Medallero Acumulado por País:** Información sobre la cantidad total de medallas obtenidas por cada país en los Juegos Olímpicos de Verano.
- **Ciudades Anfitrionas:** Datos sobre las ciudades que han sido anfitrionas de los Juegos Olímpicos y el año en que se llevaron a cabo.

5. Análisis de Datos

3. Medallero Acumulado por País:

- **Relación con la Participación:** Los países con más participaciones en los Juegos Olímpicos tienden a acumular un mayor número de medallas. Esto puede estar relacionado con la inversión en infraestructura deportiva y la diversificación en deportes.

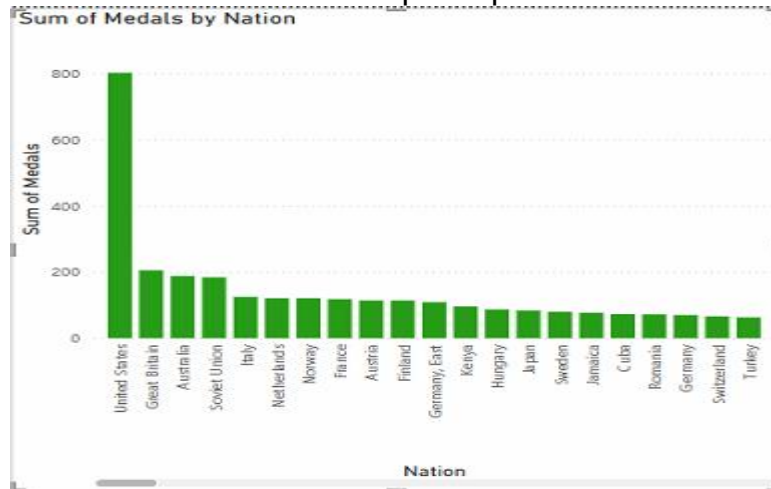
4. Récord de Ciudades Anfitrionas:

- **Impacto en el Desempeño Deportivo:** Las ciudades que han sido anfitrionas de los Juegos Olímpicos suelen recibir una mayor inversión en infraestructura deportiva, lo que puede influir en el rendimiento de los atletas y el éxito en futuras competiciones.

6. Visualizaciones

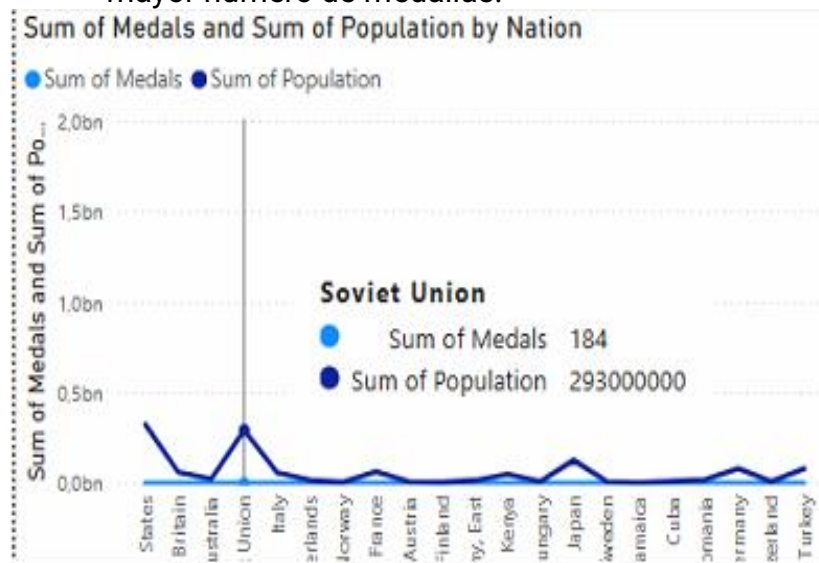
- **Medallero Acumulado por País:**

- Gráfico de barras que muestra el total de medallas acumuladas por los principales países a lo largo de los Juegos Olímpicos de Verano.
- Se destacan los países con más medallas y se compara la cantidad de medallas con el número de participaciones en los Juegos.



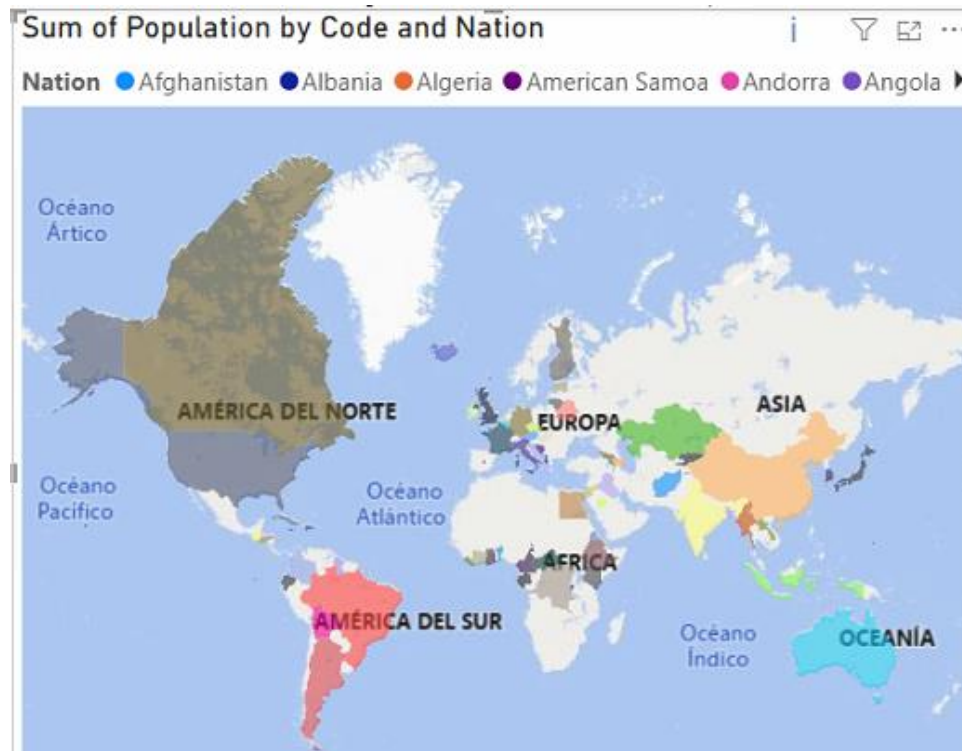
- **Participación vs. Medallas Obtenidas:**

- Gráfico de líneas que ilustra el número de participaciones en los Juegos Olímpicos y las medallas obtenidas por país.
- Se identifican patrones de inversión y participación que contribuyen a un mayor número de medallas.



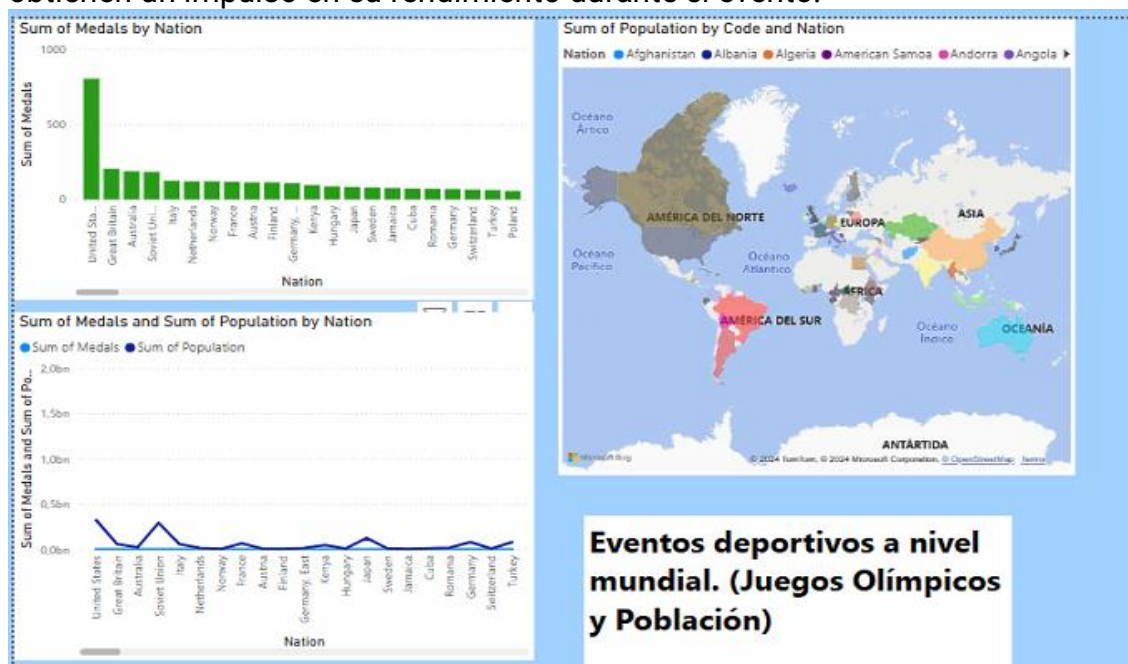
- **Ciudades Anfitrionas y Éxito Olímpico:**

- Mapa interactivo que muestra las ciudades anfitrionas de los Juegos Olímpicos y el rendimiento de los países en esos años.
- Gráfico de líneas que compara el éxito olímpico de los países anfitriones en relación con su desempeño en los Juegos.



7. Resultados Obtenidos

- **Participación y Medallas:** Los países que participan en más ediciones de los Juegos Olímpicos tienden a obtener más medallas. Esta tendencia sugiere que la experiencia y la inversión en deporte contribuyen significativamente al éxito olímpico.
- **Influencia de las Ciudades Anfitrionas:** Las ciudades anfitrionas que han invertido en infraestructura deportiva y en el desarrollo de sus atletas han mostrado un aumento en el éxito olímpico. Los países anfitriones a menudo obtienen un impulso en su rendimiento durante el evento.



8. Conclusiones y Recomendaciones

- **Conclusiones:** proporciona una visión detallada del rendimiento en los Juegos Olímpicos y cómo la participación y la inversión en infraestructura afectan la acumulación de medallas. Las visualizaciones y análisis realizados permiten una comprensión más clara de las dinámicas olímpicas a nivel mundial.
- **Recomendaciones:** Se recomienda a los países que buscan mejorar su desempeño olímpico invertir en infraestructura deportiva y fomentar una mayor participación en los Juegos Olímpicos. Para las ciudades que aspiran a ser anfitrionas, es crucial planificar y ejecutar inversiones estratégicas en instalaciones deportivas.

9. Desafíos y Problemas Encontrados

- **Variabilidad en la Inversión:** La cantidad de inversión en infraestructura varía significativamente entre las ciudades anfitrionas, lo que puede afectar los resultados y la comparación entre diferentes eventos.
- **Datos Históricos:** La recopilación de datos históricos sobre medallas y ciudades anfitrionas puede presentar desafíos debido a la variabilidad en la calidad y disponibilidad de la información.

10. Recursos y Herramientas

- **Fuentes de Datos:** Datos históricos de los Juegos Olímpicos, informes de medallas y registros de ciudades anfitrionas.
- **Herramientas de Almacenamiento:** Base de datos estructurada para gestionar la información.
- **Herramientas de Visualización:** Power BI para la creación de gráficos y dashboards interactivos.

Juegos en línea por países y el juego Warzone

Este estudio explora tres aspectos clave del mercado de videojuegos a nivel mundial:

1. **Países que juegan Warzone:** Identificamos las naciones con una base significativa de jugadores activos del popular videojuego Warzone.
2. **Países con mayores jugadores de videojuegos:** Analizamos cuáles países tienen la mayor cantidad de jugadores activos en videojuegos en general.
3. **Países que gastan más en videojuegos:** Evaluamos cuáles países destinan mayores recursos financieros a la compra y consumo de videojuegos.

Este análisis proporciona una visión integral de la distribución global de jugadores y el impacto económico del sector de videojuegos en diferentes regiones.

Extracción de Datos.

Con el primer dataset dado en .SQLITE3 se analiza el número de jugadores por millones que existen en cada país y que juegan Call of Duty WarZone.

Con el segundo dataset se puede analizar:

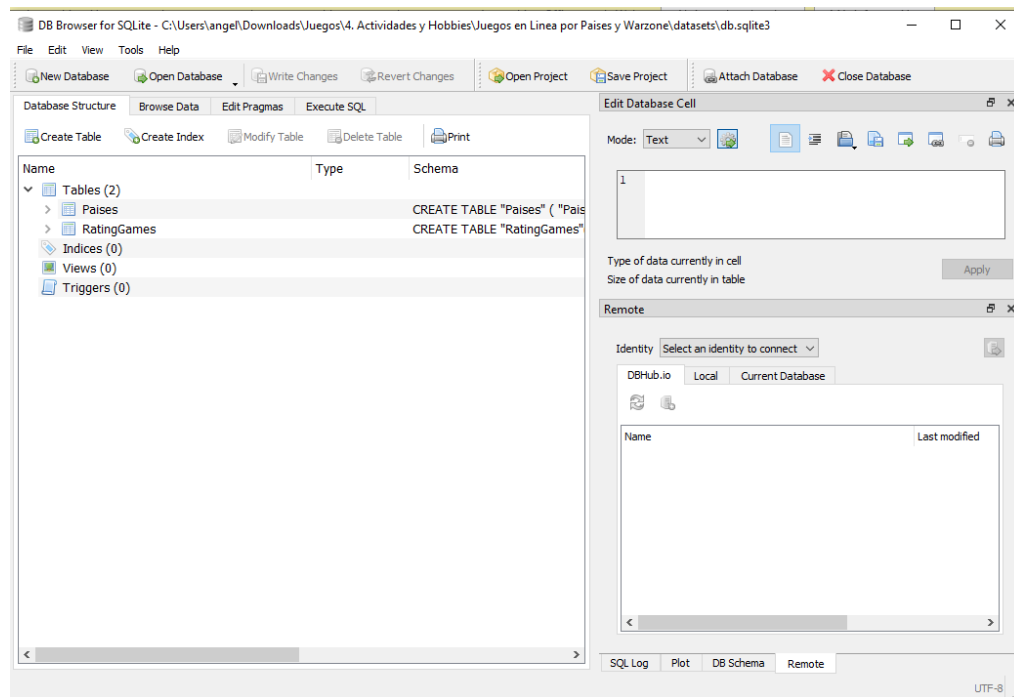
- Country: país de origen - Population: población por millones en el país en el campo Country. - Languages: el idioma que hablan los pobladores del país. - Gamers: la población por millones de jugadores en línea. - USD: dinero generado por la población jugadora

Análisis de Datos.

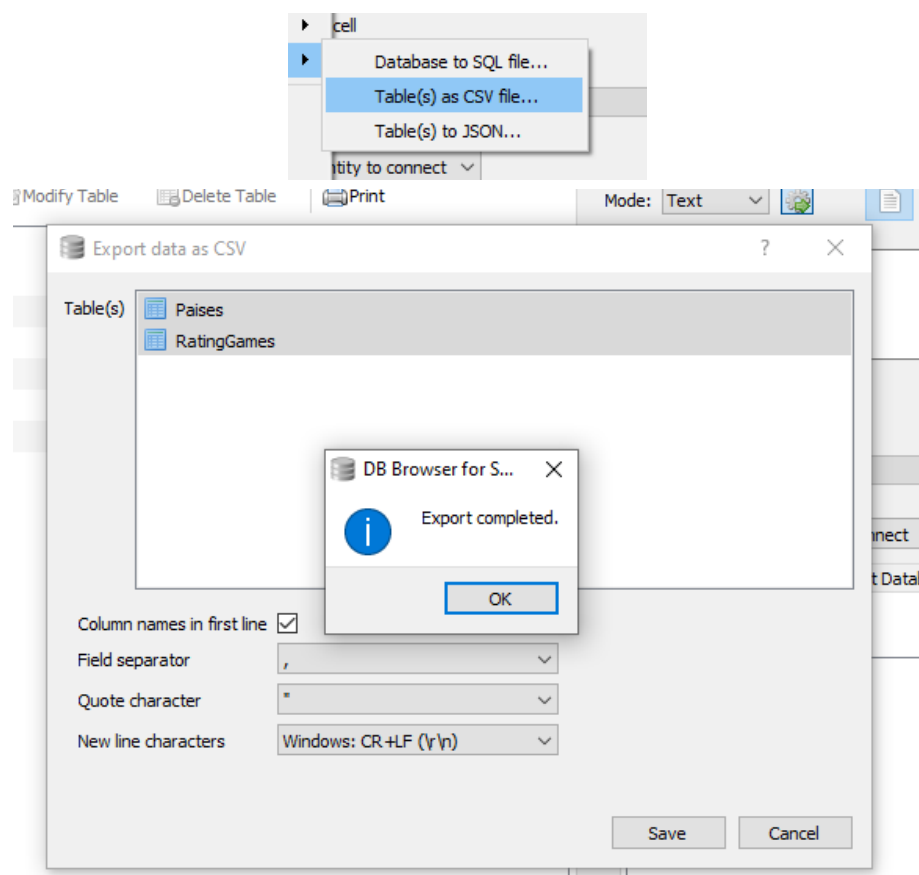
Descargamos desde Gamalytic el archivo .sqlite3

The screenshot shows the Gamalytic website interface. At the top, there's a search bar containing 'Call of Duty®: Warzone™' and buttons for 'Sign Up' and 'Log in'. The left sidebar contains a navigation menu with links: Home, Steam Analytics, Games List, Publishers List, Genres and Tags, Years, Blog, About, and Pricing. The main content area is titled 'Call of Duty®: Warzone™ - Steam Stats' and includes a subtitle 'Revenue, player data and other stats on Call of Duty®: Warzone™'. Below this is a 'Steam Header' section with a large image of Call of Duty: Warzone. Underneath the header is an 'Overview' section with the following details: Developers: Infinity Ward, Raven Software, Beenox, Treyarch, High Moon Studios, Sledgehammer Games, Activision Shanghai, Demonware, Toys for Bob; Publishers: Activision; Release date: Wed Nov 16 2022; Price: \$0; Genres: Action, Free To Play; Languages: English, French, Italian, German, Spanish - Spain, Arabic, Japanese, Korean, and more. A 'Download as SQLITE3' button is visible in the bottom right corner of the overview section.

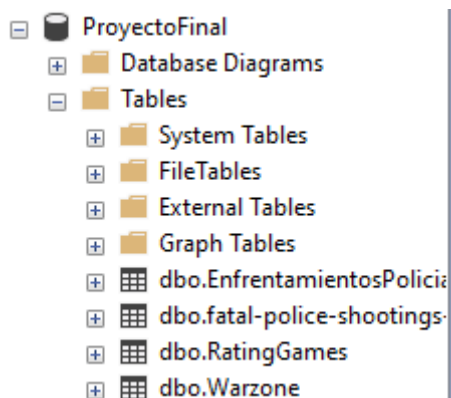
Abrimos la database desde SQLite DB Browser para abrir el archivo db.sqlite3



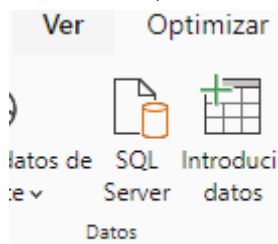
Una vez abierto exportamos ambas tablas en formato CSV para estas Subirlas a SQL Server



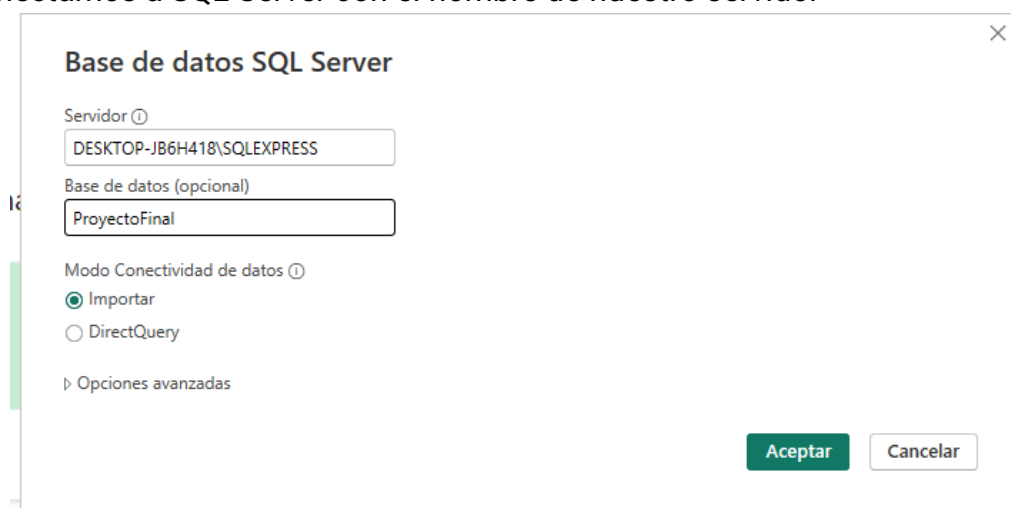
Una vez exportadas en CSV, subimos a SQL Server para realizar su conexión a Power BI de una manera más directa y rápida.

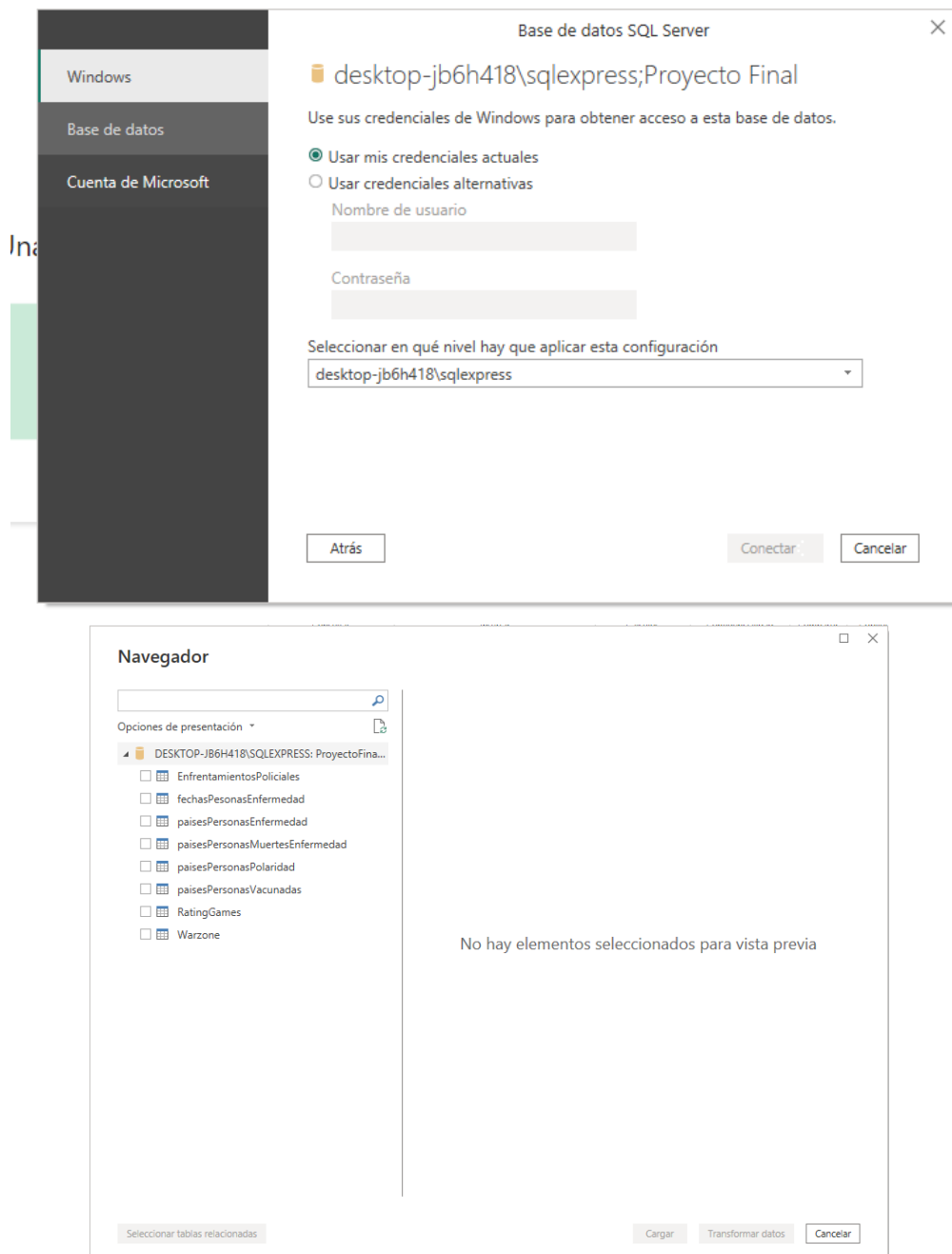


Subimos la información a Power BI desde SQL Server



Nos conectamos a SQL Server con el nombre de nuestro servidor





Visualización de información

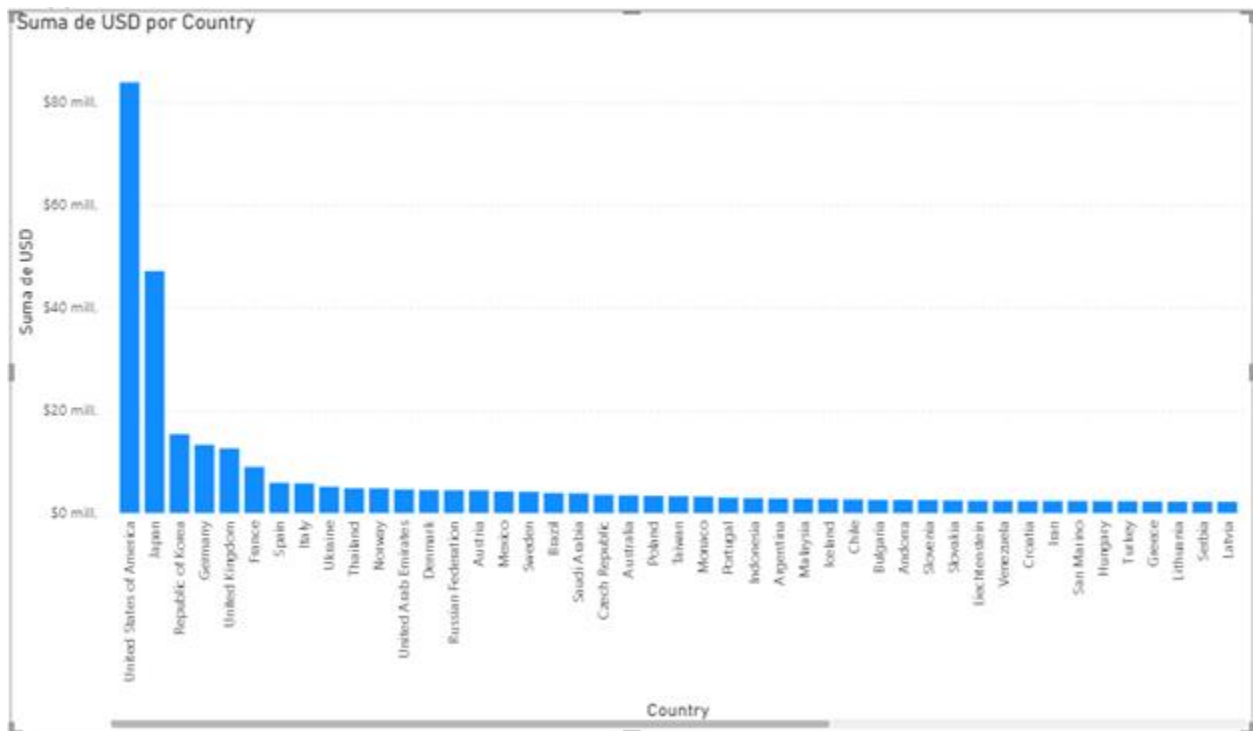
En el primer gráfico se visualiza en que países del mundo hay mayores cantidad de jugadores especialmente en el juego Warzone.



En el segundo gráfico se muestra que país es el que tiene la mayor cantidad de jugadores a nivel general, mostrando que China es dominante.

Country	Suma de Population
China	3761070
United States of America	869166
Indonesia	709686
Brazil	560838
Nigeria	531600
Bangladesh	438570
Russian Federation	382752
Thailand	366804
Mexico	348198
Japan	337566
Ethiopia	297696
Egypt	265800
Total	12856746

La tercera gráfica muestra cuales son los países que más dinero gastan en videojuegos, la sorpresa es no ver a China, pero lo que no se conocía es que China cuenta con políticas muy estrictas para evitar a su población que gaste en videojuegos y evitando adicciones en videojuegos.



Enfrentamientos Policiales en EEUU

Para obtener los datos, se descargó un archivo .XLSX del sitio web data.world. Este archivo se lo modificó para guardarlo en .CSV que contiene información sobre los enfrentamientos de la policía en los Estados Unidos, proporcionando detalles sobre los delincuentes, como el tipo de arma, género, edad, entre otros.

Este estudio desea conocer estos 3 casos:

1. Que armas han usado los delincuentes por edad.
2. Muertes por sexo y en que ciudad han ocurrido más.
3. Que sexo huyo y como de los policías.

Análisis de datos

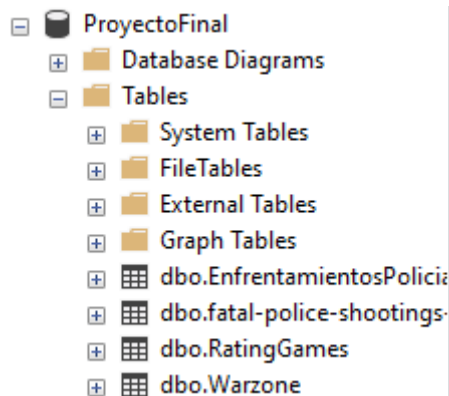
Descargamos la información de esta database desde .dataworld que fue descargada en xlsx, y debemos transformarla a CSV.

The screenshot shows the Data World interface for the 'Fatal Police Shootings' dataset. The page includes a search bar, navigation tabs (Overview, Discussion, Activity), and a detailed description of the dataset. The description mentions that the data is sourced from the Washington Post and includes details about the victims, the circumstances of the shootings, and the methodology used for data collection. The dataset is shared with everyone and is updated daily. The file 'fatal-police-shootings-data.csv' is listed as the primary data source. The page also features a 'Launch workspace' button and a 'Learn more' link for enterprise users.

Guardamos el archivo en CSV

The screenshot shows a file download dialog box. The 'File name' field contains 'fatal-police-shootings-data.csv'. The 'File type' dropdown is set to 'CSV (MS-DOS) (*.csv)'. The 'Save in' field shows the path 'ANGEL Villamil'. The 'Etiquetas' (Tags) field is empty.

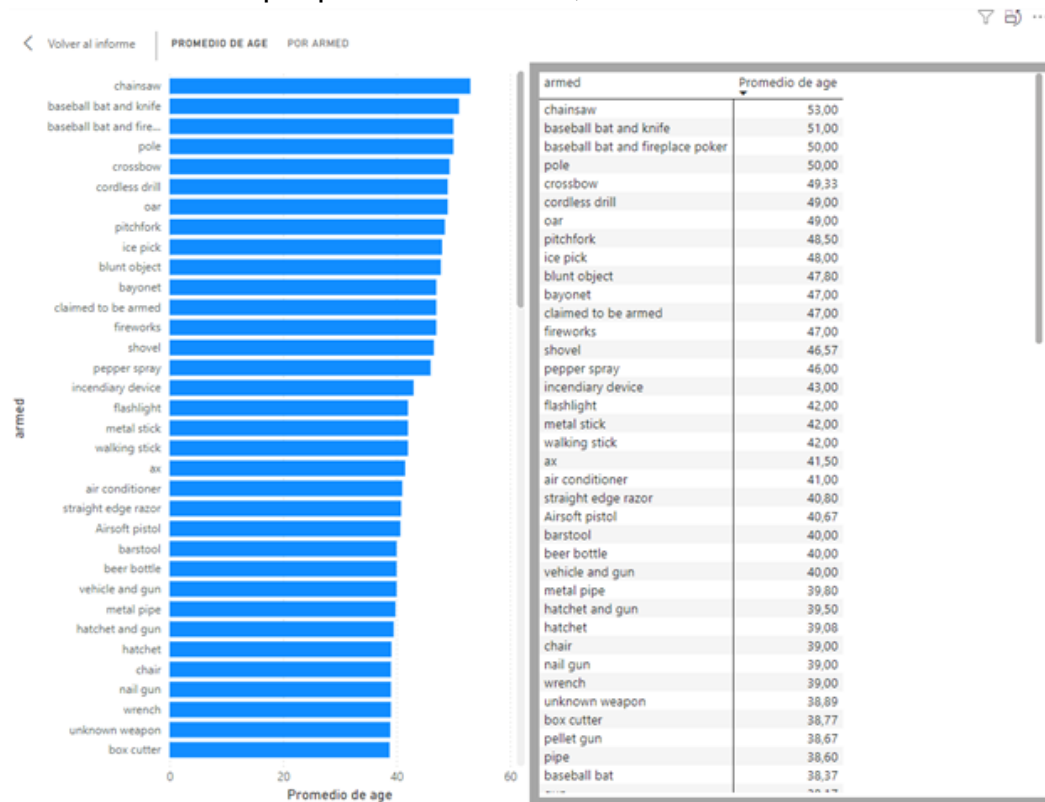
Y conectamos con SQL Server



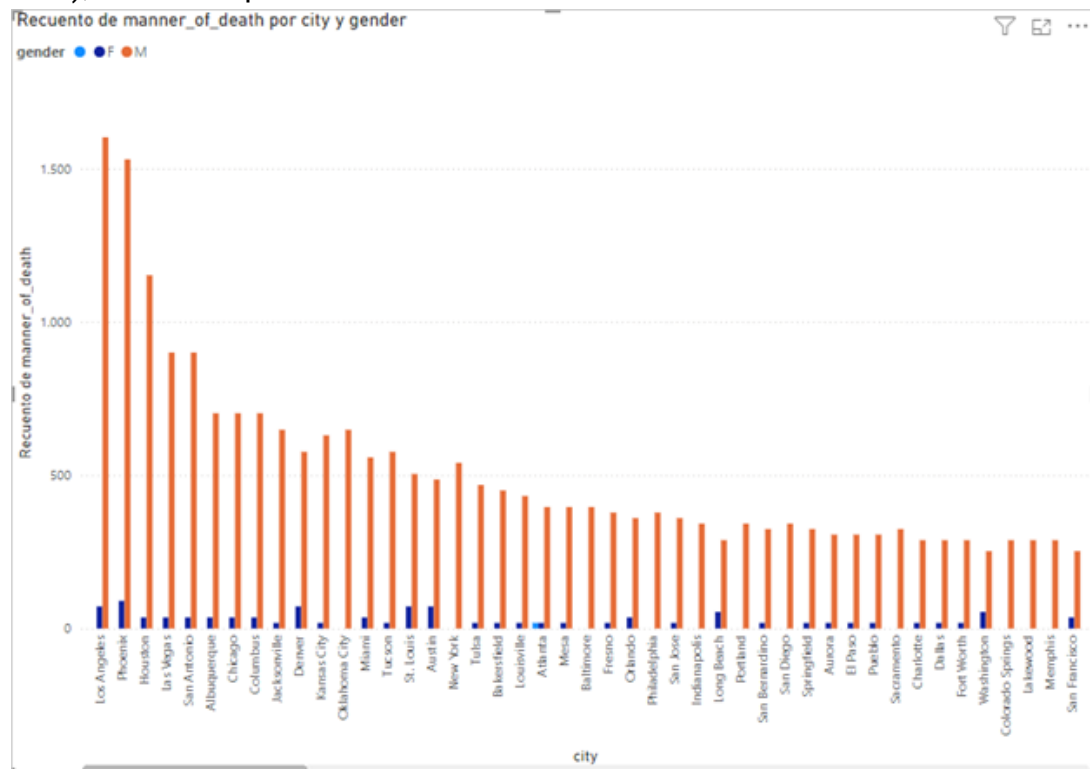
Y se vuelve a repetir el mismo paso de exportación para POWER BI.

Visualización de graficas

En el primer gráfico respondiendo el primer caso de estudio podemos ver en que tipo de armas fueron usadas por promedio de edad, indicandonos en forma de barra y tabla.

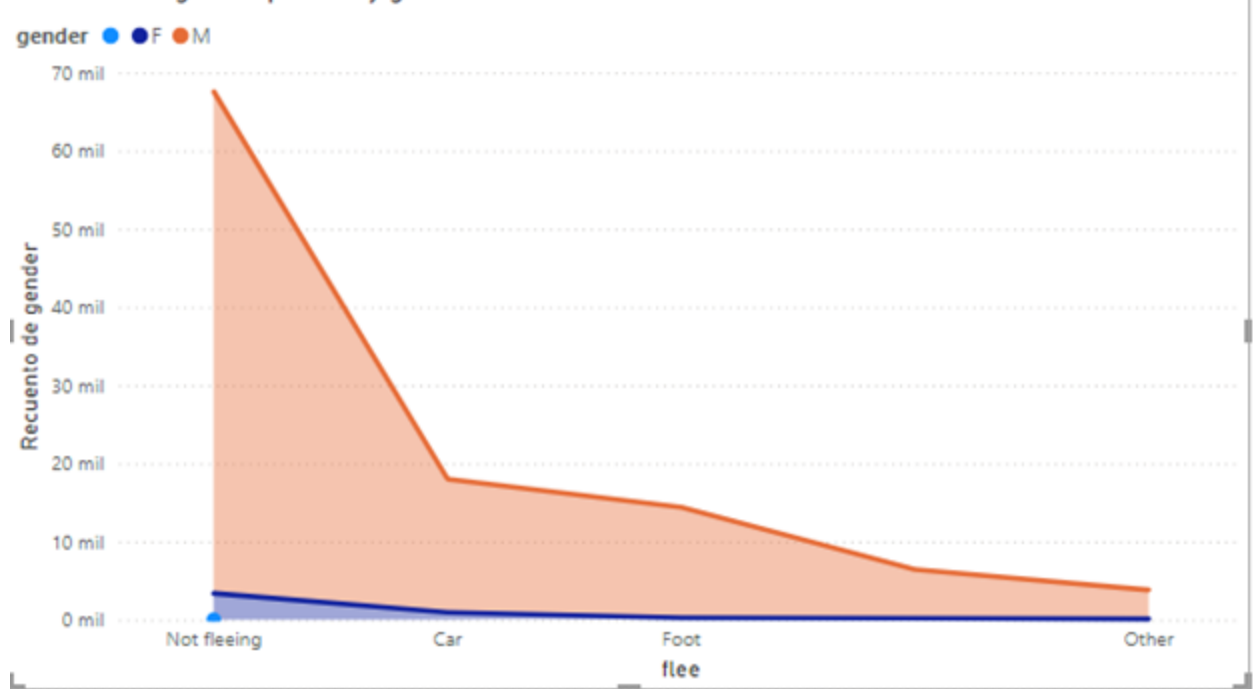


En el segundo gráfico según el segundo caso de estudio, se visualiza el recuento de muertes que hay por ciudad en EEUU, que a su vez muestra que genero (Masculino o Femenino), son los implicados en esta estadística.



La tercera gráfica en resupuesta al tercer caso de estudio de este tema, muestra que género huyo y de que manera, si fue a pie, en auto, de otra manera o incluso no logró escapar, que demuestra la gran efectividad policiaca en EEUU.

Recuento de gender por flee y gender



Noticias y Eventos Mundiales

1. Definición del Caso de Estudio

El caso de estudio se centra en el análisis global de eventos y noticias relacionados con el impacto del COVID-19 en varios aspectos, como la tasa de infección, la mortalidad, la vacunación y la polarización de opiniones. La finalidad es comprender mejor cómo se ha desarrollado la pandemia a nivel mundial y cómo ha afectado a diferentes países y poblaciones.

2. Enfoque

El objetivo principal de este proyecto es analizar y visualizar datos relacionados con el COVID-19 para identificar patrones y tendencias globales que permitan una mejor comprensión del impacto de la pandemia. Para ello, se recolectaron datos relevantes sobre el COVID-19 desde diferentes fuentes fiables como Kaggle y Statista, se transformaron y limpiaron para asegurar su compatibilidad y consistencia para su análisis, y se utilizaron herramientas como Power BI para visualizar la información mediante gráficos y mapas interactivos. Finalmente, se interpretaron los resultados obtenidos para proporcionar insights significativos sobre la evolución de la pandemia y su impacto global.

4. Actividades Realizadas

- *Extracción de Datos:* Isaac Quinapallo se encargó de la extracción de datos relevantes sobre el COVID-19 de plataformas como Kaggle y Statista.

- **Transformación de Datos:** Se desarrollaron scripts para transformar y limpiar los datos obtenidos, asegurando su coherencia y formato adecuado.
- **Visualización:** Se integraron los datos en Power BI, creando gráficos y mapas interactivos que representan la evolución y el impacto del COVID-19.
- **Análisis de Resultados:** Interpretación de los datos visualizados para identificar patrones y proporcionar recomendaciones basadas en los resultados.

6. Recursos y Herramientas Utilizadas

- **Fuentes de Datos:** Kaggle, Statista.
- **Herramientas de Transformación y Limpieza:** Scripts personalizados en Python para conversión de formatos y limpieza de datos.
- **Herramientas de Visualización:** Power BI para la creación de gráficos y análisis visuales.
- **Bases de Datos NoSQL:** MongoDB para el almacenamiento y gestión de datos estructurados y no estructurados.

7. Arquitectura de la Solución

La arquitectura de la solución incluye la recolección de datos en formatos JSON, CSV y XLS, su transformación y limpieza, almacenamiento en bases de datos MongoDB, y posterior visualización en Power BI. La estructura permite un flujo de trabajo eficiente desde la obtención de datos hasta la presentación de resultados visuales.

8. Extracción de Datos

Los datos fueron extraídos de Kaggle y Statista, centrados en cinco casos de estudio clave:

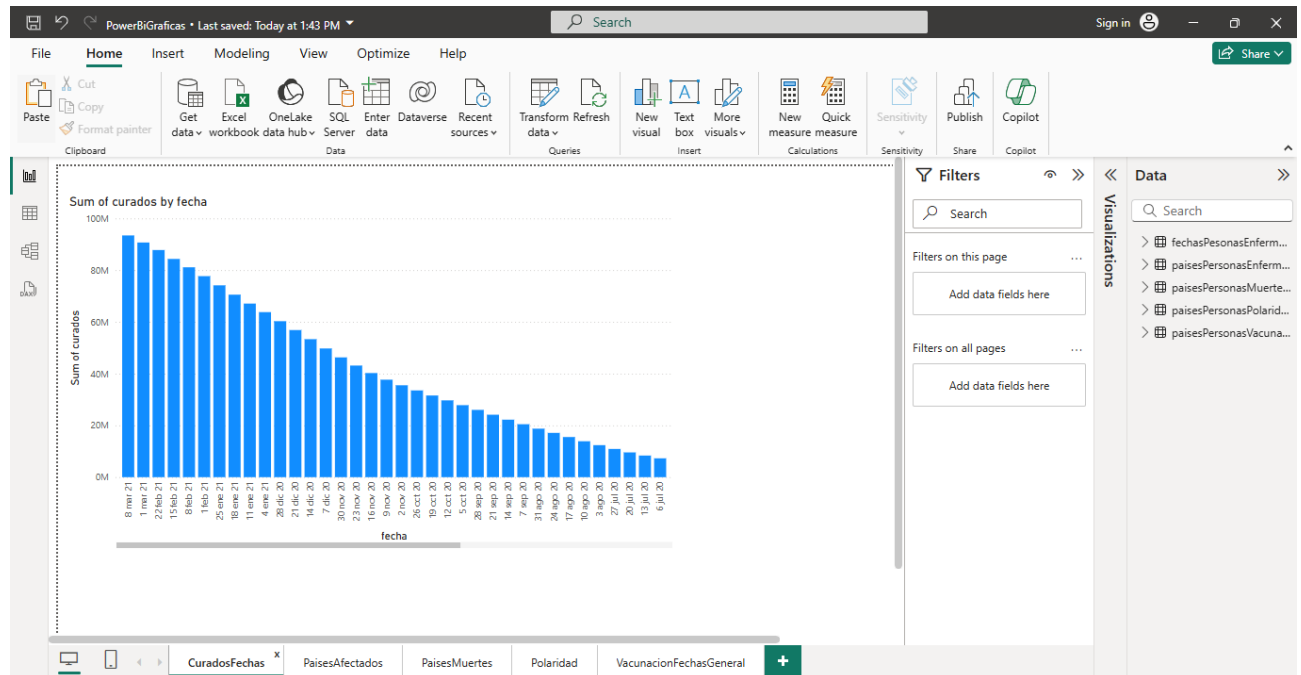
- **FechasPersonasCuradas:** Incluye información sobre las personas curadas de COVID-19 desde 2020 hasta 2021.
- **PaísesPersonasPolaridad:** Mide la polaridad global en términos de opiniones neutras, positivas y negativas.
- **PaísesPersonasEnfermas:** Proporciona datos sobre el número de personas afectadas por COVID-19 en diferentes países.
- **PaísesPersonasMuertesEnfermedad:** Muestra el número de personas fallecidas por la enfermedad en distintos países.
- **PaísesPersonasVacunadas:** Contiene información sobre el número de personas vacunadas, el país, y las fechas de vacunación de 2020 a 2021.

9. Análisis y Visualización de Información

FechasPersonasCuradas: Se creó un gráfico que analiza el índice de recuperación de personas desde 2020 hasta 2021, destacando las tendencias en la curación a lo largo del tiempo.

Tasa de Curación Por Fechas

Figura

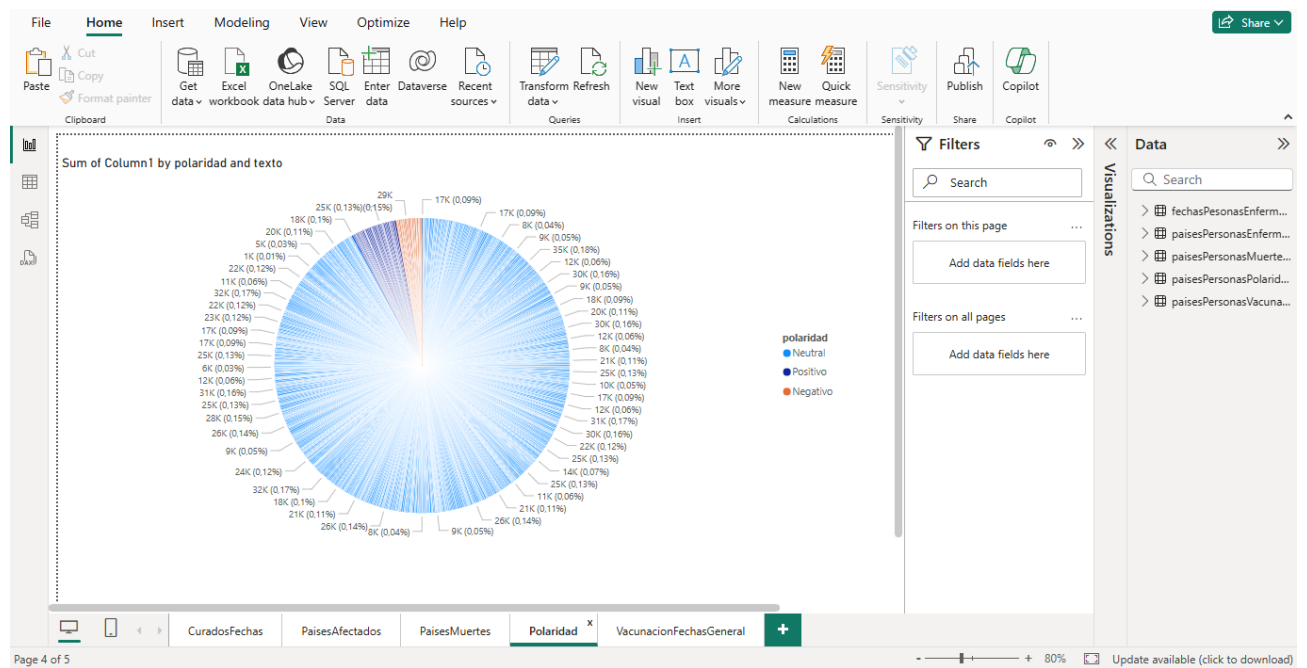


Nota: En este grafico se muestra la tasa de vacunación por País de acuerdo a los índices generados por Who (World Health Organization)

PaísesPersonasPolaridad: Se generó un gráfico que muestra la polaridad global respecto al COVID-19, categorizando las opiniones en neutras, positivas y negativas.

Tasa de Porcentaje de Polaridad

Figura

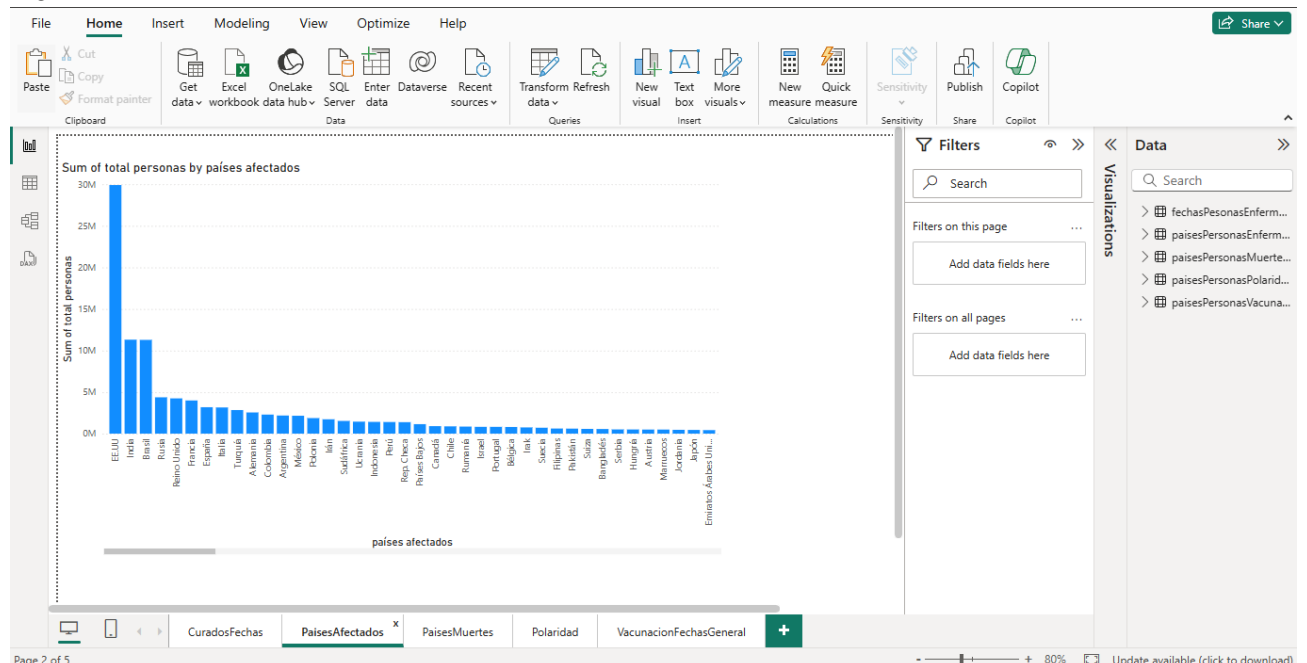


Nota: En este grafico se muestra la tasa de Polaridad Global tanto de valores Neutrales como Negativos y Positivos.

PaísesPersonasEnfermas: Se desarrolló un gráfico que muestra el índice de países afectados y el número de personas enfermas por cada país.

Tasa de Enfermedad por País

Figura

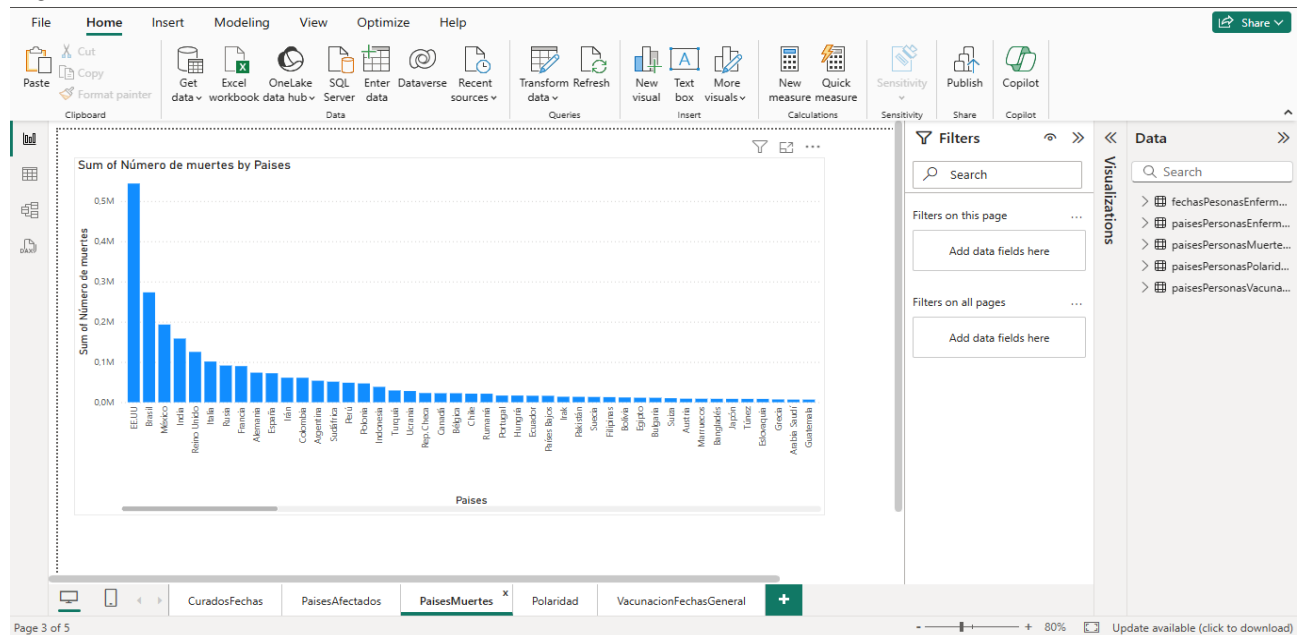


Nota: En este grafico se muestra la tasa de Enfermedad Global de las Personas en diferentes Países.

PaísesPersonasMuertesEnfermedad: Se realizó un gráfico de barras que presenta el número de personas fallecidas por COVID-19 en cada país.

Tasa de Muertes de Covid-19 por País

Figura

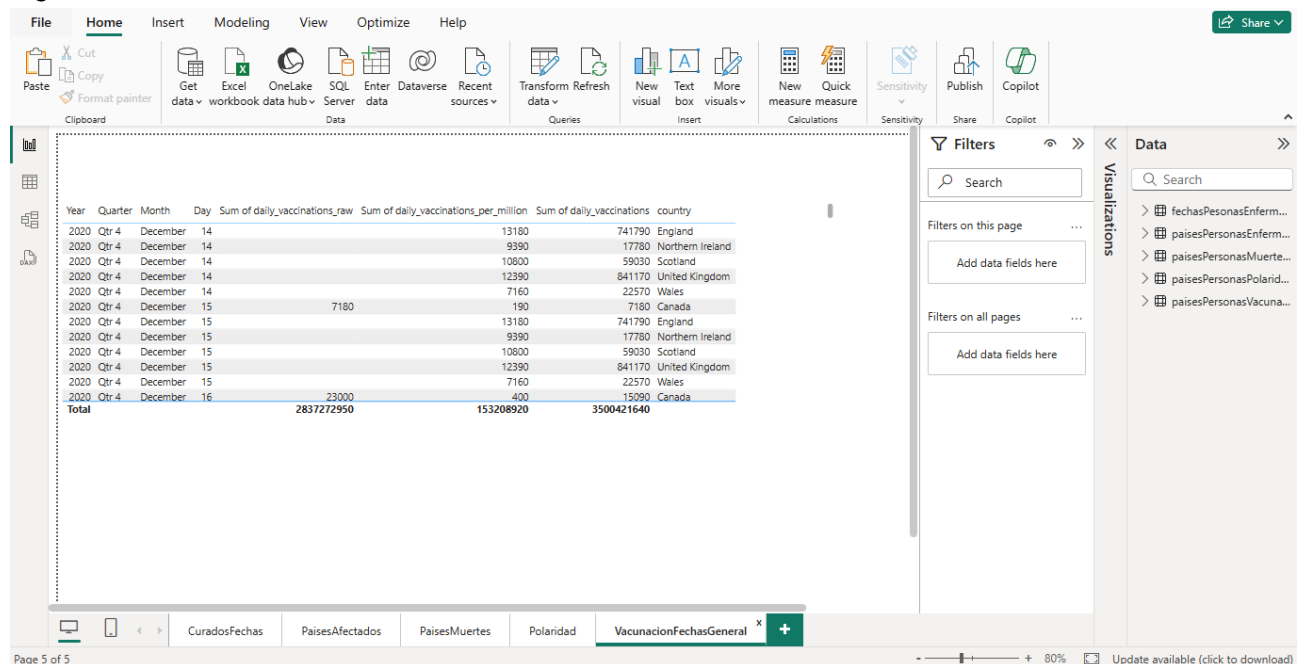


Nota: En este grafico se muestra la tasa de Muerte por Covid-19 Global de las Personas en diferentes Países.

PaísesPersonasVacunadas: Se creó una tabla que indica el número total de personas vacunadas por país, incluyendo detalles como el número de vacunaciones y las fechas de vacunación, desglosadas por mes desde 2020 hasta 2021.

Tasa de Vacunación por País y Fecha

Figura



Nota: En este grafico se muestra la tasa de Vacunación Global y sus Fechas con su respectivo País.

11. Resultados Obtenidos

Los resultados revelaron patrones clave en la propagación del COVID-19, con diferencias notables entre países en términos de respuesta y manejo de la pandemia. Los gráficos mostraron correlaciones entre las tasas de vacunación y la reducción de casos y muertes.

12. Conclusiones

- Conclusiones: La pandemia del COVID-19 ha tenido un impacto diverso en diferentes regiones, influenciado por factores como políticas de salud pública, acceso a vacunas y respuesta gubernamental.
- Recomendaciones: Se recomienda un enfoque coordinado a nivel global para futuras pandemias, con un énfasis en la equidad en el acceso a recursos médicos y vacunas.

Suma total de los datasets

Se realizó la suma de los datasets mediante código en SQL Server

```
USE ProyectoFinal;

SELECT
    t.name AS NombreTabla,
    SUM(p.rows) AS TotalFilas
FROM
    sys.tables AS t
INNER JOIN
    sys.partitions AS p ON t.object_id = p.object_id
WHERE
    p.index_id IN (0, 1) -- 0 = HEAP, 1 = CLUSTERED
    AND t.name IN (
        'Olympics',
        'EnfrentamientosPoliciales',
        'fechasPersonasEnfermedad',
        'paísesPersonasMuertesEnfermedad',
        'paísesPersonasEnfermedad',
        'paísesPersonasPolaridad',
        'paísesPersonasVacunadas',
        'provincias',
        'RatingGames',
        'Warzone'
    )
GROUP BY
    t.name
ORDER BY
    TotalFilas DESC;
```

	NombreTabla	TotalFilas
1	RatingGames	265800
2	paísesPersonasEnfermedad	150246
3	paísesPersonasMuertesEnfermedad	147132
4	Olympics	141024
5	provincias	123584
6	Warzone	122642
7	fechasPersonasEnfermedad	120320
8	EnfrentamientosPoliciales	110214
9	paísesPersonasPolaridad	0
10	paísesPersonasVacunadas	0

El total de esta suma de los datasets es **1,088,962**.

Y falta sumar los últimos dos de la tabla ya que son de 100 mil datos cada uno, sumando un nuevo total de **1,288,962**.

Conclusiones y recomendaciones del Proyecto

La recopilación de datos es un proceso muy útil cuando se necesita obtener información relevante sobre un tema de interés tanto a nivel nacional como internacional. Esto permite presentar datos estadísticos de manera efectiva.

El análisis de las gráficas muestra que la población en países orientales tiende a tener un menor índice de jugadores en plataformas en línea, mientras que los países latinoamericanos son los que más ingresos generan en este sector.

Herramientas como Power BI y Tableau mejoran la visualización de los datos recolectados, facilitando su comprensión y aprendizaje.

Desafíos y problemas encontrados

Uno de los desafíos que enfrentamos es la recopilación de datos por geolocalización en Twitter, ya que en algunas provincias no hay actividad en la plataforma o no tienen acceso a internet, lo que dificulta obtener la cantidad de datos deseada.

Además, al importar datos directamente a SQL Server, a veces los conjuntos de datos se crean con caracteres de tipo String, lo que limita la capacidad de generar estadísticas precisas durante el análisis.

La falta de herramientas adecuadas para procesar grandes volúmenes de datos puede ralentizar el análisis y generar cuellos de botella en el flujo de trabajo.

El equipo que maneja la importación y el análisis de datos no está del todo capacitado y en las mejores prácticas.

Link de GitHub del Proyecto

El repositorio del proyecto, que incluye todo el código, datasets, y documentaciones, está disponible en:

- <https://github.com/isaacquinapallo/ProyectoFinalAnalisisDeDatos.git>

ProyectoFinalAnalisisDeDatos Public Watch 1

main 1 Branch 0 Tags Go to file Add file Code

VillamilA Create readme.md a2b5988 · 7 hours ago 33 Commits

4. Actividades y Hobbies/Juegos en Linea por ...	Update readme.md	7 hours ago
5.- Noticias y Eventos Mundiales	Create readme.md	7 hours ago
Enfrentamientos Policiales y Abuso	Update readme.md	7 hours ago
HistorialDeActividades.pdf	Add files via upload	2 days ago
InformeProyectoFinal_IQuinapallo_AVillamil_KR...	Add files via upload	yesterday
README.md	Update README.md	7 hours ago

README

Proyecto Final – Bimestre 2

Arquitectura

- La arquitectura general del proyecto se muestra en la siguiente imagen:

El diagrama de arquitectura, titulado 'ARQUITECTURA', muestra el flujo de datos del proyecto. A la izquierda, tres fuentes de datos: 'Juegos en linea por paises', 'Warzone' y 'Enfrentamiento y Abuso Policial', cada una con un icono de globo. 'Juegos en linea por paises' y 'Warzone' fluyen a través de bases de datos ('DB') a 'SQLite'. 'Enfrentamiento y Abuso Policial' fluye a través de un archivo 'xlsx' a 'MySQL'. Tanto 'SQLite' como 'MySQL' exportan datos a archivos 'CSV'. Estos archivos 'CSV' se conectan a 'Microsoft SQL Server' (representado por una antena) y finalmente a 'Power BI' (representado por un gráfico de barras).

Link de los Videos

Youtube

Juegos Online Warzone y Enfrentamientos Policiales en EEUU

<https://youtu.be/cAY8350KNp0>

<https://youtu.be/bLGqD7c-uIU>

Referencias

1. Vidmar. [Online]. Available: <http://www.halcyon.com/pub/journals/21ps03-vidmar>
2. Gamalytic. (2023). Gamalytic.com. <https://gamalytic.com/game/1962663>
3. Apache. (2020). Logstash [Online]. Available: <https://www.elastic.co/es/logstash>
4. A. Robledano. (2020). ¿Qué es MySQL? [Online]. Available: <https://openwebinars.net/blog/que-es-mysql/>
5. Apache. (2020). Kibana [Online]. Available: <https://www.elastic.co/es/what-is/kibana>
6. M. Alvarez. (2003). ¿Qué es Python? [Online]. Available: <https://desarrolloweb.com/articulos/1325.php>
7. IONOS. (2020). CouchDB [Online]. Available: <https://www.ionos.es/digitalguide/hosting/cuestiones-tecnicas/presentacion-de-couchdb/>
8. Elastic. (2021). Elasticsearch [Online]. Available: <https://www.elastic.co/es/what-is/elasticsearch>
9. Davinci. (2020). ¿Qué es Logstash? [Online]. Available: <https://www.davincigroup.es/que-es-logstash-ejemplo-practico-de-uso/>
10. Tableau. (2021). ¿Qué es Tableau? [Online]. Available: <https://www.tableau.com/es-es/why-tableau/what-is-tableau>
11. M. Parada. (2019). ¿Qué es SQL Server? [Online]. Available: <https://openwebinars.net/blog/que-es-sql-server>
12. Korda, J. (2019, January 11). *Summer & Winter Olympic Games*. Data.world; data.world. <https://data.world/johayes13/summer-winter-olympic-games>
13. Data Society. (2016, November 29). *Fatal Police Shootings*. Data.world; data.world. <https://data.world/data-society/fatal-police-shootings>