Informe Final:
Análisis Exploratorio de Datos

Docente: Ana Maria CuadrosValdivia
Alumno: Leon Felipe Davis Coropuna

ANEXO

Este es el formato sugerido, puede agregar secciones pero no puede omitir las sugeridas.
Dashboard: https://dashboard-crime-tcd.vercel.app/crimes-chicago

INFORME FINAL DE ANÁLISIS EXPLORATORIO DE DATOS DEL CONJUNTO DE DATOS ……

1. **Hipótesis iniciales**:
   - **1.1. Motivación: describe cómo se originaron sus hipótesis y las razones para elegirlas**
     Estas hipótesis surgieron al observar discrepancias entre la percepción ciudadana del crimen en redes sociales y los datos oficiales reportados por las autoridades. Además, se identificaron patrones temporales y espaciales que sugieren diferencias en la forma y momento en que se reportan o comentan los delito
   - **1.2. Exprese sus hipótesis en forma de pregunta (sea claro y conciso)**
     Hipótesis 1: ¿Los usuarios de Twitter tienden a enfocarse más en crímenes impactantes como homicidios, mientras que estos delitos representan una proporción menor en los datos oficiales?
     Hipótesis 2: ¿Existen distritos que mantienen una frecuencia de crímenes constante durante el año, y otros que presentan variaciones marcadas según la estación?
     Hipótesis 3: ¿Los reportes oficiales se publican mayormente en horario laboral, mientras que los comentarios en redes sociales aumentan en horas no laborales, especialmente de madrugada?
   - **1.3. Plan de análisis:**
     Describe qué pasos siguió para investigar las hipótesis.
     1. Limpieza de datos: Se realizó el proceso de data wrangling para preparar los datos de forma estructurada y consistente.
     2. Desarrollo de aplicación: Se creó una aplicación full-stack para facilitar el análisis interactivo de los datos.
     3. Visualización: Se integraron gráficos, mapas y series de tiempo para explorar patrones y relaciones.
     4. Comprobación de hipótesis: Usando la herramienta desarrollada, se analizaron los datos para validar o refutar las hipótesis.

2. **Fuente de datos:**

2.1. **Fuente:**

Fuente 1: Crimes - 2001 to Present

- Origen y fecha: El dataset fue obtenido del portal oficial del Chicago Police Department el 4 de junio de 2025.
- Responsables de recolección: Chicago Police Department, mediante el sistema CLEAR (Citizen Law Enforcement Analysis and Reporting).
- Técnica de recolección: Registros administrativos de crímenes reportados, actualizados diariamente y anonimizados al nivel de bloque para proteger la privacidad.
- Área de conocimiento: Seguridad pública y análisis criminal.
- Objetivo computacional: Detectar patrones espacio-temporales y correlaciones entre tipo de crimen, ubicación y frecuencia.
- Variables relevantes: Fecha, tipo de crimen, ubicación (ward, distrito, comunidad), arrestos, violencia doméstica, entre otros. Estas permiten analizar tendencias, zonas críticas y la respuesta policial.
-Referencia:
https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2

Fuente 2: Datos de Twitter

- Origen y fecha: Datos recolectados durante 2025 mediante scraping con la herramienta twscrape, empleando cookies para acceder a publicaciones públicas.
Responsables de recolección: Recolección propia como parte del proyecto, siguiendo criterios éticos (sin información privada).
- Técnica de recolección: Web scraping usando un diccionario temático de palabras clave relacionadas a delitos (e.g., “asalto”, “robo”, “disparo”, “asesinato”, etc.).
- Área de conocimiento: Ciencia de datos, análisis social y procesamiento de lenguaje natural.
- Objetivo computacional: Comparar la percepción social del crimen en redes sociales con los registros oficiales, detectando picos de atención y correlaciones temporales.
- Variables relevantes: Texto del tweet, fecha, hora y hashtags. Estas ayudan a detectar qué tipos de delitos generan mayor discusión pública, y cuándo.
- Referencia: No aplica paper base específico; se utilizó twscrape: https://github.com/JustAnotherArchivist/twscrape

2.2. **Descripción:**

Describe el conjunto de datos:

a) **A nivel de atributos:**

El dataset contiene información sobre crímenes y tweets relacionados, cubriendo un periodo temporal específico (por ejemplo, un año). Tiene miles de filas y decenas de columnas que incluyen datos temporales (año, mes, día, hora), geográficos (latitud, longitud),

categóricos (tipo de crimen, distrito), y derivados (clústeres, transformaciones geométricas). La mayoría de atributos no tienen valores nulos, salvo algunos en el dataset de tweets donde varias columnas están vacías. Los atributos pueden ser categóricos (ej. tipo de crimen, estación), cuantitativos (hora, coordenadas), o nominales (día de la semana). La distribución varía según el atributo, con tendencias temporales claras y dispersión típica para datos espaciales.

b) **A nivel de registros:**
Cada registro representa un evento de crimen o un tweet relacionado. En el dataset de crímenes, están etiquetados por categoría de delito, mientras que en tweets las etiquetas reflejan categorías predichas. Los registros tienen granularidad temporal y espacial precisa.

c) Relación entre atributos:
Se identifican correlaciones temporales y espaciales, como la concentración de ciertos crímenes en zonas específicas y horarios particulares. Además, existe relación entre categorías y ubicaciones, y patrones estacionales en la frecuencia de delitos.

Defina alguna terminología especial usada en el conjunto de datos e indique las columnas importantes tanto para el lector como para la hipótesis. Si los atributos poseen unidades de medidas diferentes, investigue cómo los artículos científicos las procesan.

El conjunto de datos contiene términos específicos como Cluster (agrupamiento de crímenes por similitud geográfica) y Category (tipo de crimen) , que son claves para el análisis. Las columnas más importantes para el lector y la hipótesis son: fecha y hora (para análisis temporal), ubicación (latitud, longitud), y categoría del crimen (para clasificar eventos).

Los atributos tienen distintas unidades de medida, por ejemplo, tiempo en horas/minutos, ubicación en grados (latitud/longitud), y distancias en unidades métricas. Los gráficos presentados — histogramas, líneas de tiempo y mapas — muestran claramente sus ejes X y Y, facilitando la interpretación básica. Para análisis más avanzados como boxplots o scatterplots, se especifica qué variables se están comparando, garantizando que el lector pueda entender fácilmente qué información se visualiza.

Realice un cuadro resumen de la descripción de los atributos.

Cuadro sobre crímenes documentos por el gobierno (Crimes - 2001 to Present)

Columna	Descripción
Year	Año del crimen.
Month	Mes del crimen.
dayOfWeek	Día de la semana (0=Lunes, 6=Domingo).
dayOfMonth	Día del mes en que ocurrió el crimen.
dayOfYear	Día del año (1–365/366).
weekOfYear	Semana del año.
weekOfMonth	Semana dentro del mes.
Información temporal	Descripción
Hour	Hora en que ocurrió el crimen.
Minute	Minuto en que ocurrió el crimen.
Hour_Zone	Zona de horas (ej. madrugada, mañana, tarde, noche).
BusinessHour	Indica si ocurrió en horario laboral (1 = sí, 0 = no).
Weekend	Indica si ocurrió fin de semana.
Holiday	Indica si fue feriado.
Season	Estación del año (0–3, ej. invierno, primavera...).
Ubicación geográfica	Descripción
Latitude	Latitud del lugar del crimen.
Longitude	Longitud del lugar del crimen.
X, Y	Coordenadas proyectadas (sistema cartesiano).
Radius	Distancia radial desde un punto de referencia.
Angle	Ángulo de dirección relativo al punto de referencia.
Transformaciones geométricas	Descripción
Rot30_X, Rot30_Y	Coordenadas tras rotación de 30 grados.
Rot45_X, Rot45_Y	Coordenadas tras rotación de 45 grados.
Rot60_X, Rot60_Y	Coordenadas tras rotación de 60 grados.
Procesamiento y clustering	Descripción
Cluster	Etiqueta del clúster (ej. resultado de KMeans).
Codificación y variables categóricas	Descripción
ICode	Código del tipo de crimen (codificado).
District	Número de distrito policial.
Category	Categoría del crimen (etiqueta objetivo).

Datos de el dataset obtenido de Twitter con scrape

Columnas principales	
ID	Identificador numérico único del tweet.
Description	Texto original del tweet.
Category	Tipo de crimen.
Confidence	Nivel de confianza del modelo en la categoría.
Confidence-skywalker	Nivel de confianza de un segundo modelo.
PredictedCrime	Predicción de crimen asociada.
Description_normalized	Versión preprocesada del texto.
Category_encoded	Representación numérica de la categoría
Datos del tweet (contenido y contexto)	
url	Enlace al tweet.
Date	Fecha y hora exacta de publicación.
lang	Idioma del tweet.
hashtags, cashtags, mentionedUsers, links	Elementos del contenido.
conversationId, conversationIdStr	ID de la conversación.
possibly_sensitive	Indica si contiene contenido sensible.
source, sourceUrl, sourceLabel	Plataforma o cliente desde donde se publicó.
Datos del usuario que publica	
user_id, user_id_str	ID del usuario.
user_username, user_displayname	Nombre de usuario y nombre mostrado.
user_url, user_rawDescription	Información del perfil.
user_created	Fecha de creación de la cuenta.
user_profileImageUrl	Imagen de perfil.
user_verified, user_blue	Verificación oficial y verificación Blue.
user_descriptionLinks, user_pinnedIds	Otros enlaces en su perfil.
user__type	Tipo de entidad (probablemente siempre üser").
Datos sobre respuestas	
inReplyToTweetId, inReplyToTweetIdStr	ID del tweet al que responde.
inReplyToUser_id, inReplyToUser_id_str	ID del usuario al que responde.
inReplyToUser_username	Nombre del usuario al que responde.
inReplyToUser__type	Tipo (probablemente üser").
Datos temporales derivados	
Year, Month, dayOfYear weekOfMonth, weekOfYear Hour, Minute	Datos temporales derivados.

2.3. Formato:

Describe el formato en el que se encontró el conjunto de datos original.

Fuente 1: Crimes - 2001 to Present: Se encontró en el sitio oficial de la policia de Chicago, por defecto cuenta con un formato CSV en su mayoría sin nulos ni duplicados.

2.4. Transformaciones:

Describe cualquier transformación que necesitó realizar en los datos para convertirlos en un formato utilizable.

Se calcularon ángulo y distancia radial a partir de X y Y para facilitar el análisis espacial. Se aplicaron rotaciones de 30, 45 y 60 grados para captar patrones direccionales. Se usó clustering para agrupar zonas similares. Las variables categóricas fueron codificadas con label encoding y se realizó

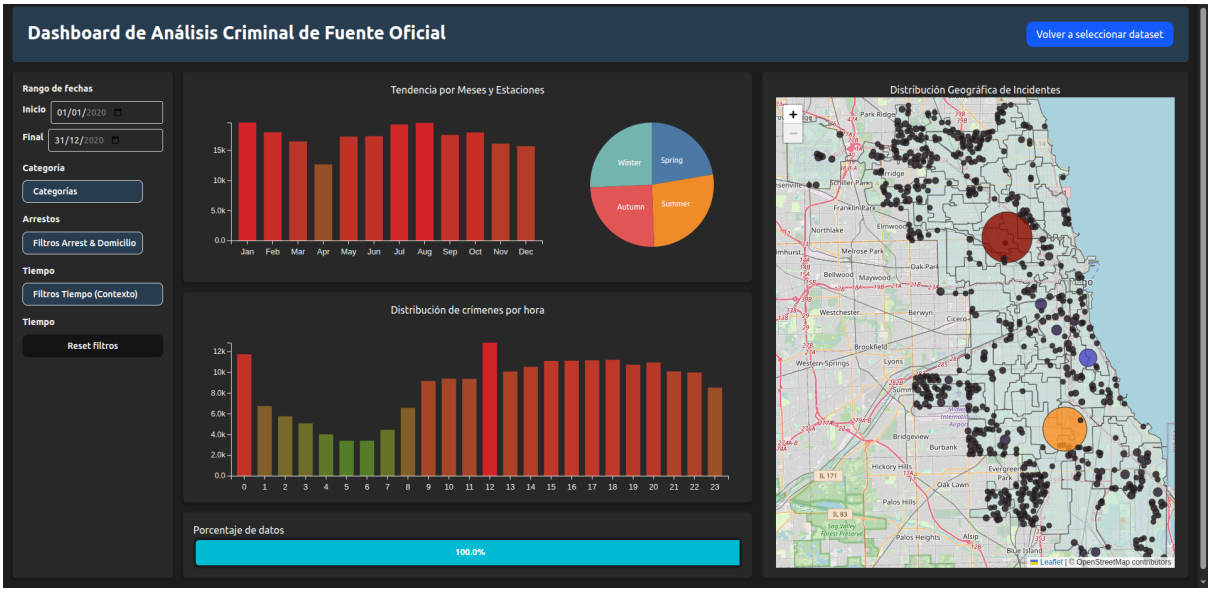
escalado para normalizar las variables numéricas. Luego de eso se tuvo que subir a una base de datos, en este caso MongoDBAtlas por lo que se tuvo que convertir en un formato .json.

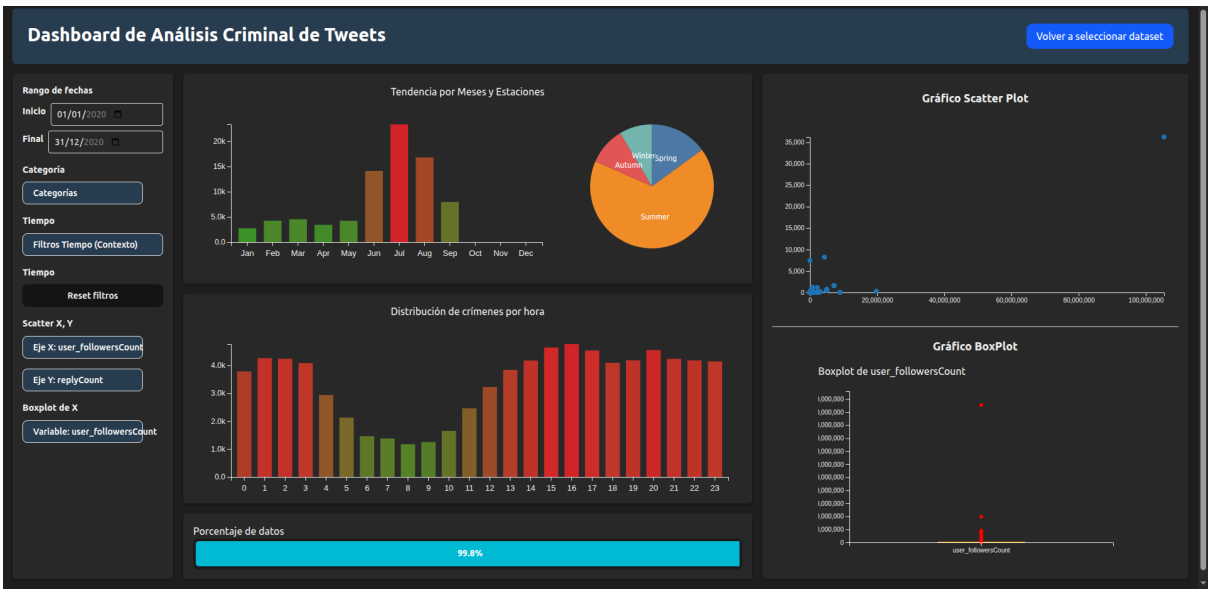
2.5. Limpieza de datos:

En el dataset de crímenes, solo el 3% de valores estaban vacíos y las columnas contenían la información esperada. En cambio, el dataset de tweets tenía cerca de 70 columnas 100% nulas, imposibilitando la imputación. Sin embargo, las columnas clave, como la fecha y el texto del tweet, estaban completas y disponibles.

3. Exploración:

Describe los pasos que realizó para investigar las hipótesis. Coloque las visualizaciones, con título y leyendas, explique lo que aprendió de ellas. Debe tener al menos 10 gráficos exploratorios.





4. **Conclusión:**
   Qué conocimiento obtuvo del análisis exploratorio de datos.
   Para cada hipótesis incluya sus conclusiones intermedias y finales.

Hipótesis 1: ¿Los usuarios de Twitter tienden a enfocarse más en crímenes impactantes como homicidios, mientras que estos delitos representan una proporción menor en los datos oficiales?

- Conclusión intermedia:

Al analizar la distribución de categorías en los tweets, se observa una sobrerrepresentación de crímenes graves (homicidios, asaltos violentos) en comparación con la distribución oficial de crímenes, donde estos representan un porcentaje menor relativo a delitos menores.

- Conclusión final:

Los usuarios de Twitter efectivamente enfatizan crímenes más impactantes, lo que puede generar una percepción pública sesgada respecto a la frecuencia real de dichos delitos en los datos oficiales.

Hipótesis 2: ¿Existen distritos que mantienen una frecuencia de crímenes constante durante el año, y otros que presentan variaciones marcadas según la estación?

- Conclusión intermedia:

El análisis temporal por distrito muestra que algunos distritos tienen una frecuencia estable de crímenes mes a mes, mientras que otros presentan picos y caídas relacionadas con estaciones específicas (por ejemplo, aumento en verano o en meses festivos).

- Conclusión final:

Se confirma la hipótesis: la dinámica del crimen varía según la zona, con distritos más estables y otros con variabilidad estacional, lo cual es relevante para diseñar políticas focalizadas.

Hipótesis 3: ¿Los reportes oficiales se publican mayormente en horario laboral, mientras que los comentarios en redes sociales aumentan en horas no laborales, especialmente de madrugada?

- Conclusión intermedia:

Los registros oficiales muestran mayor concentración durante horarios laborales típicos (9 a 18 horas), mientras que la actividad en Twitter sobre crímenes aumenta notablemente en la noche y madrugada.

- Conclusión final:

Se confirma que los comentarios en redes sociales tienen un patrón temporal diferente, con más actividad fuera del horario laboral, reflejando probablemente un comportamiento social distinto a la hora de reportar o comentar eventos delictivos.