

Fusión Semántico-Espacio-Temporal de Datos de Redes Sociales Y Gubernamentales para el Análisis de Crimen y Seguridad Pública

Leon Felipe Davis Coropuna

Problema

Actualmente, los enfoques existentes para el análisis de datos criminales enfrentan limitaciones debido a la fragmentación y heterogeneidad de la información proveniente de diversas fuentes. Esta dispersión genera registros superpuestos, incompletos o inconsistentes, lo que dificulta consolidar una visión unificada de fenómenos delictivos.

Objetivo

Desarrollar un enfoque para el análisis de datos criminales urbanos que integre de manera conjunta las dimensiones semántica, espacial y temporal, a partir de la fusión de información heterogénea proveniente de redes sociales y fuentes oficiales, con el fin de generar una visión unificada y coherente que contribuya a la toma de decisiones en seguridad pública.

Datasets

Crimes-2001 to Present

Este dataset proporciona un registro de incidentes delictivos en la ciudad de Chicago desde el año 2001 hasta la actualidad. Los datos son proporcionados por el sistema CLEAR (Citizen Law Enforcement Analysis and Reporting) del Departamento de Policía de Chicago, y son actualizados diariamente. En total, el conjunto de datos contiene más de 8.3 millones de registros y 22 columnas, incluyendo datos espaciales. Como muestra se tomó datos desde 2020.



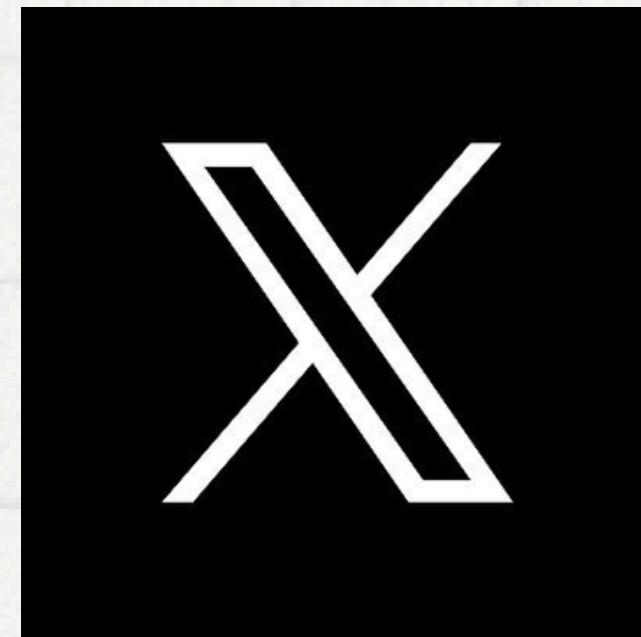
**CHICAGO
DATA PORTAL**

Link dataset

Datasets

Tweets entre enero y abril de 2020

El dataset fue obtenido mediante técnicas de scraping en la plataforma X.com (antes Twitter), utilizando palabras clave relacionadas con crímenes. Como resultado, se recolectaron aproximadamente 15,033 registros, cada uno representando un tweet individual, junto con sus respectivos metadatos. La información fue almacenada inicialmente en archivos JSON, organizados cronológicamente. Posteriormente, estos archivos fueron unificados y convertidos en un único archivo CSV con 106 atributos por registro.



x.com

Problemas de dataset

Datos faltantes y duplicados

Problema: Algunos registros contenían valores nulos o estaban duplicados.

- Registros originales: 1183866
- Registros después de limpieza: 1159998
- Registros eliminados: 23868
- Porcentaje eliminado: 2.02%



**CHICAGO
DATA PORTAL**

Link dataset

Problemas de dataset

Variables categóricas no numéricas

Problema: Las variables categóricas (como "Primary Type" o "District") no pueden ser procesadas directamente por los modelos o algoritmos.

```
#   Column           Non-Null Count  Dtype  
---  --  
0   ID              1159998 non-null   int64  
1   Case Number     1159998 non-null   object 
2   Date            1159998 non-null   object 
3   Block           1159998 non-null   object 
4   IUCR            1159998 non-null   object 
5   Primary Type    1159998 non-null   object 
6   Description     1159998 non-null   object 
7   Location Description  1159998 non-null   object 
8   Arrest           1159998 non-null   bool   
9   Domestic         1159998 non-null   bool   
10  Beat             1159998 non-null   int64  
11  District         1159998 non-null   int64  
12  Ward             1159998 non-null   float64 
13  Community Area  1159998 non-null   float64 
14  FBI Code         1159998 non-null   object 
15  X                1159998 non-null   float64 
16  Y                1159998 non-null   float64 
17  Year             1159998 non-null   int64  
18  Updated On       1159998 non-null   object 
19  Latitude          1159998 non-null   float64 
20  Longitude         1159998 non-null   float64 
21  Location          1159998 non-null   object 

dtypes: bool(2), float64(6), int64(4), object(10)  
memory usage: 188.1+ MB
```

```
#   Column           Non-Null Count  Dtype  
---  --  
0   ID              1159998 non-null   int64  
1   Case Number     1159998 non-null   object 
2   Date            1159998 non-null   datetime64[ns] 
3   Block           1159998 non-null   object 
4   IUCR            1159998 non-null   object 
5   Category        1159998 non-null   int64  
6   Description     1159998 non-null   object 
7   Location Description  1159998 non-null   object 
8   Arrest           1159998 non-null   bool   
9   Domestic         1159998 non-null   bool   
10  Beat             1159998 non-null   int64  
11  District         1159998 non-null   int64  
12  Ward             1159998 non-null   float64 
13  Community Area  1159998 non-null   float64 
14  FCode            1159998 non-null   object 
15  X                1159998 non-null   float64 
16  Y                1159998 non-null   float64 
17  Year             1159998 non-null   float64 
18  Updated On       1159998 non-null   object 
19  Latitude          1159998 non-null   float64 
20  Longitude         1159998 non-null   float64 
21  Location          1159998 non-null   object
```

Problemas de dataset

Rangos de valores diferentes

Problema: Algunas variables numéricas estaban en diferentes escalas, lo que afecta el rendimiento de los algoritmos.

```
from sklearn.preprocessing import LabelEncoder

categorical_cols = ['District', 'Category', 'Hour_Zone', 'Season']
label_encoders = {}

for col in categorical_cols:
    le = LabelEncoder()
    df[col] = le.fit_transform(df[col])
    label_encoders[col] = le

from sklearn.preprocessing import MinMaxScaler

to_scale = ['Hour', 'Minute', 'Year', 'Month', 'dayOfWeek', 'dayOfMonth',
            'dayOfYear', 'weekOfMonth', 'weekOfYear']

scaler = MinMaxScaler()
df[to_scale] = scaler.fit_transform(df[to_scale])
```

Problemas de dataset

Información temporal subutilizada

Problema: Usar solo la columna Date como una cadena o campo único limita el análisis, ya que no permite capturar patrones temporales importantes como el día de la semana, fin de semana, etc.

22	Month	1159998	non-null	int32
23	dayOfWeek	1159998	non-null	int32
24	dayOfMonth	1159998	non-null	int32
25	dayOfYear	1159998	non-null	float64
26	weekOfMonth	1159998	non-null	int64
27	weekOfYear	1159998	non-null	UInt32
28	Hour	1159998	non-null	int32
29	Minute	1159998	non-null	int32
30	Hour_Zone	1159998	non-null	category
31	BusinessHour	1159998	non-null	int64
32	Weekend	1159998	non-null	int64
33	Season	1159998	non-null	object
34	Holiday	1159998	non-null	bool
35	Rot30_X	1159998	non-null	float64
36	Rot30_Y	1159998	non-null	float64
37	Cluster	1159998	non-null	int32

¿Qué se descubrió?

Análisis de mutual information (información mutua) para cuantificar cuánta información proporciona cada variable (feature) sobre la categoría del crimen ('Category')

Feature	MI Score
FCode	2.285003
Rot30_Y	0.216821
Radius	0.216549
Rot60_Y	0.214920
Rot60_X	0.213658
Rot30_X	0.211785
Y	0.192255
Rot45_X	0.189966
Rot45_Y	0.184173
X	0.169316
ID	0.162898
Minute	0.082764
Beat	0.079316
Community Area	0.062595
Cluster	0.061754
District	0.054459
Ward	0.052541
Hour	0.044829
BusinessHour	0.036376
Hour_Zone	0.036123
Year	0.015762
Season	0.011104
Weekend	0.010702

FCode	2.445897
Angle	0.289704
Rot30_Y	0.287521
Rot45_Y	0.286572
Rot60_X	0.286064
Rot45_X	0.285376
Rot30_X	0.284664
Rot60_Y	0.284502
Radius	0.283441
Y	0.204331
ID	0.174396
X	0.170194
BusinessHour	0.126087
Minute	0.098302
Cluster	0.077721
Hour_Zone	0.074456
District	0.067808

Otro estudio

¿Qué se descubrió?

La cantidad de outliers es muy baja, pero eliminarlos directamente podría ser contraproducente

Columna - # Outliers - % Outliers
ID: 3536 (0.3048%)
Category: 0 (0.0000%)
Beat: 0 (0.0000%)
District: 0 (0.0000%)
Ward: 0 (0.0000%)
Community Area: 0 (0.0000%)
X: 7086 (0.6109%)
Y: 2 (0.0002%)
Year: 0 (0.0000%)

Year: 0 (0.0000%)
Latitude: 2 (0.0002%)
Longitude: 7089 (0.6111%)
Radius: 2 (0.0002%)
X_scaled: 7086 (0.6109%)
Y_scaled: 2 (0.0002%)
Cluster: 0 (0.0000%)
Rot30_X: 63 (0.0054%)
Rot30_Y: 2 (0.0002%)
Rot45_X: 2 (0.0002%)
Rot45_Y: 20 (0.0017%)
Rot60_X: 2 (0.0002%)
Rot60_Y: 64 (0.0055%)
Month: 0 (0.0000%)

Month: 0 (0.0000%)
dayOfWeek: 0 (0.0000%)
dayOfMonth: 0 (0.0000%)
dayOfYear: 0 (0.0000%)
weekOfMonth: 0 (0.0000%)
weekOfYear: 0 (0.0000%)
Hour: 0 (0.0000%)
Minute: 0 (0.0000%)
Hour_Zone: 0 (0.0000%)
BusinessHour: 0 (0.0000%)
Weekend: 0 (0.0000%)
Season: 0 (0.0000%)
Holiday: 0 (0.0000%)
dayOfWeek_int: 0 (0.0000%)

**Tweets entre
enero y abril de
2020**

Problemas de dataset

Columnas inservibles y valores duplicados

Problema: Algunas columnas tienen en su totalidad valores vacíos y algunos registros se repiten

Valores nulos por columna (y su porcentaje):

	Nulos	Porcentaje (%)
viewCount	15033	100.00
retweetedTweet	15033	100.00
quotedTweet	15033	100.00
place	15033	100.00
coordinates	14981	99.65
...
quotedTweet_card	14974	99.61
quotedTweet_possibly_sensitive	14501	96.46
quotedTweet_type	14263	94.88
card_options	15020	99.91
card_finished	15020	99.91

[67 rows x 2 columns]

Filas duplicadas: 97 (0.65%)

Problemas de dataset

Información temporal subutilizada

Problema: Depender solo de Date de tipo texto termina siendo una limitante

Problemas de dataset

Normalización de texto

Problema: Los textos de x.com incluyen emojis, menciones de usuarios, hashtags, etc.

Woman shot in head on Far South Side, man crit.
@JoeStreckert If you want the basics in a much.
"How to Get Away with Murder" writers created .
@dude_vol @PatriotAlways2 @prettynikkivar You .
Man Killed In Bishop Ford Freeway Shooting htt.

woman shot head far south side man critically .
want basic muchmore digestible format rick gea.
get away murder writer created gabriel maddox .
understand rare compared weekly death toll cit.
man killed bishop ford freeway shooti

Problemas de dataset

Falta de un label que indique el tipo de crimen

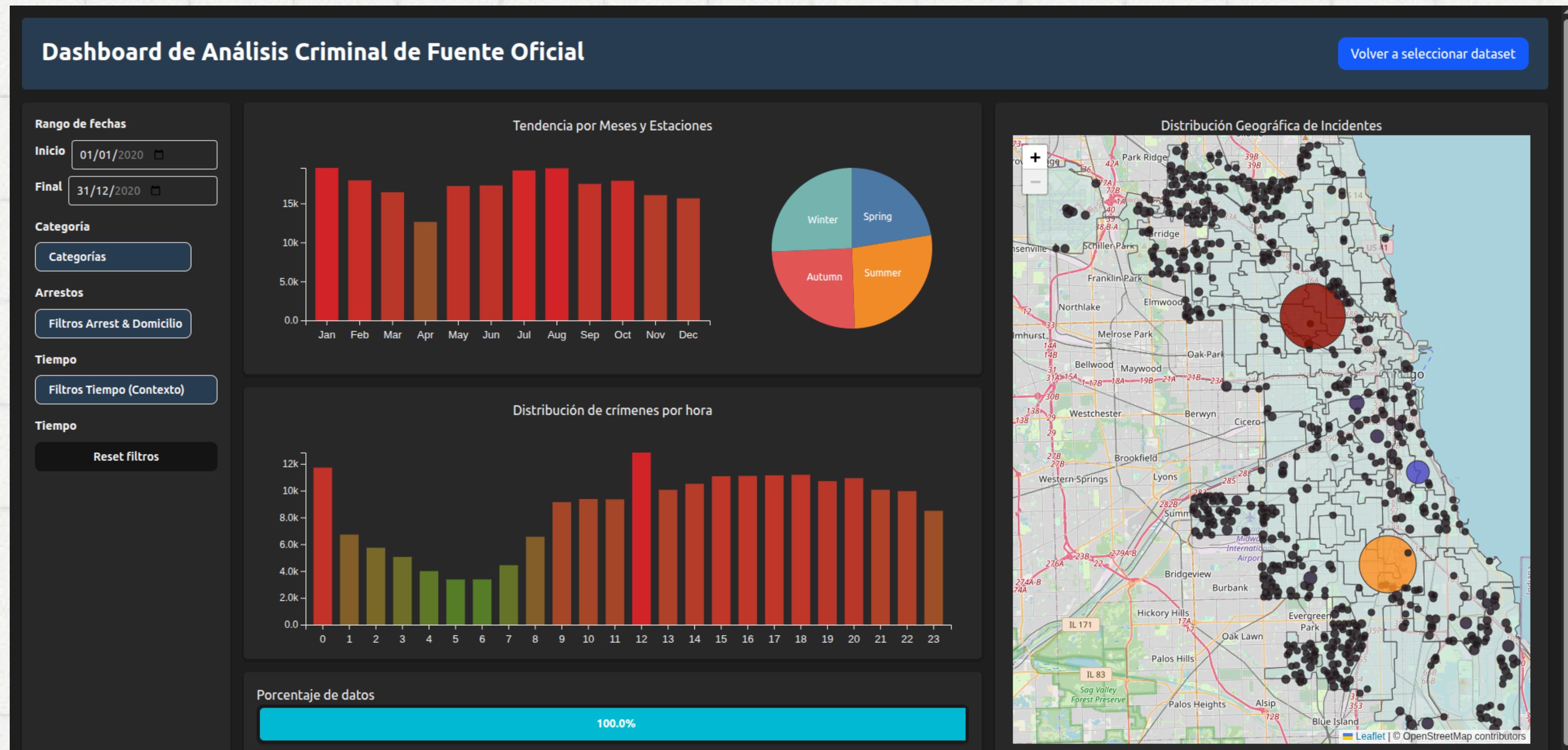
Problema: Al ser solo texto no se indica claramente que tipo de crimen es el que se ocurrió.

```
model_name = "Luna-Skywalker/BERT-crime-analysis"
tokenizer = AutoTokenizer.from_pretrained(model_name)
model = AutoModelForSequenceClassification.from_pretrained(model_name)
```

```
classifier = pipeline("zero-shot-classification", model="facebook/bart-large-mnli")
```

	PredictedCrime	Confidence
0	assault	0.529124
1	murder	0.644339
2	murder	0.606229
3	homicide	0.271090
4	homicide	0.484051

Dashboard

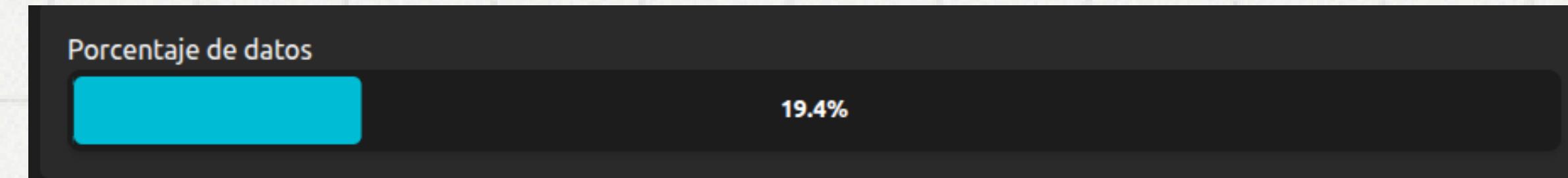
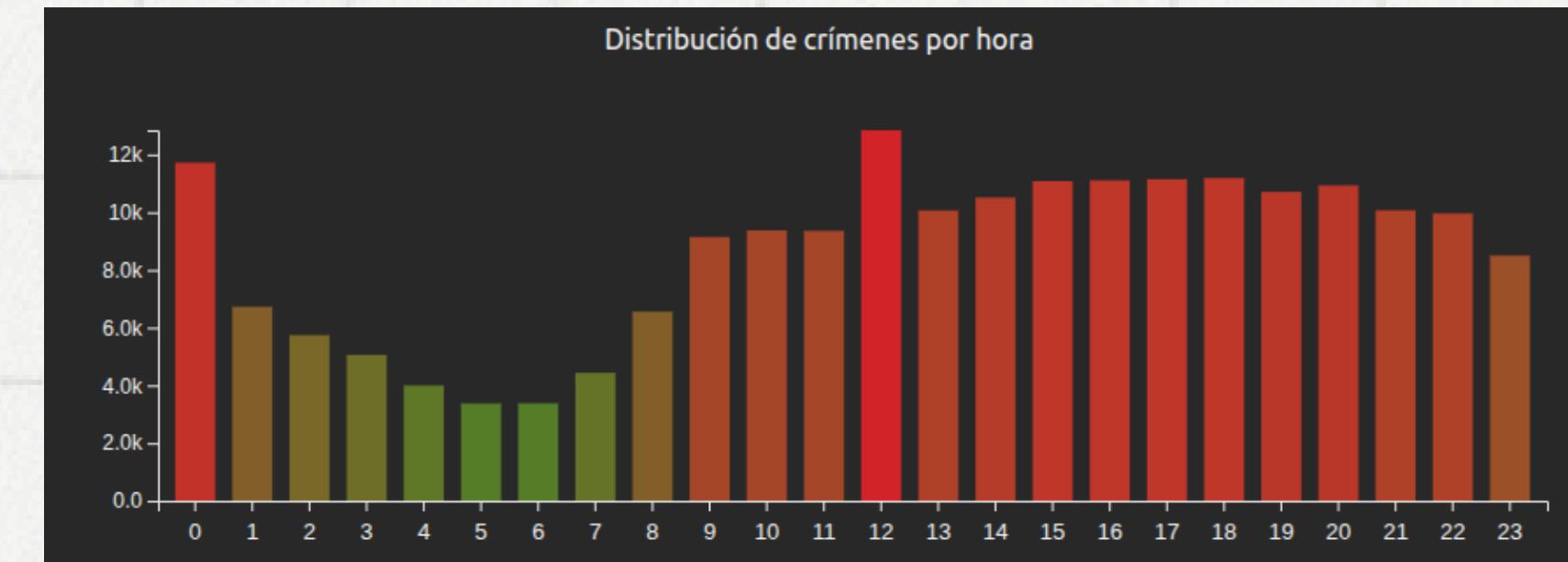
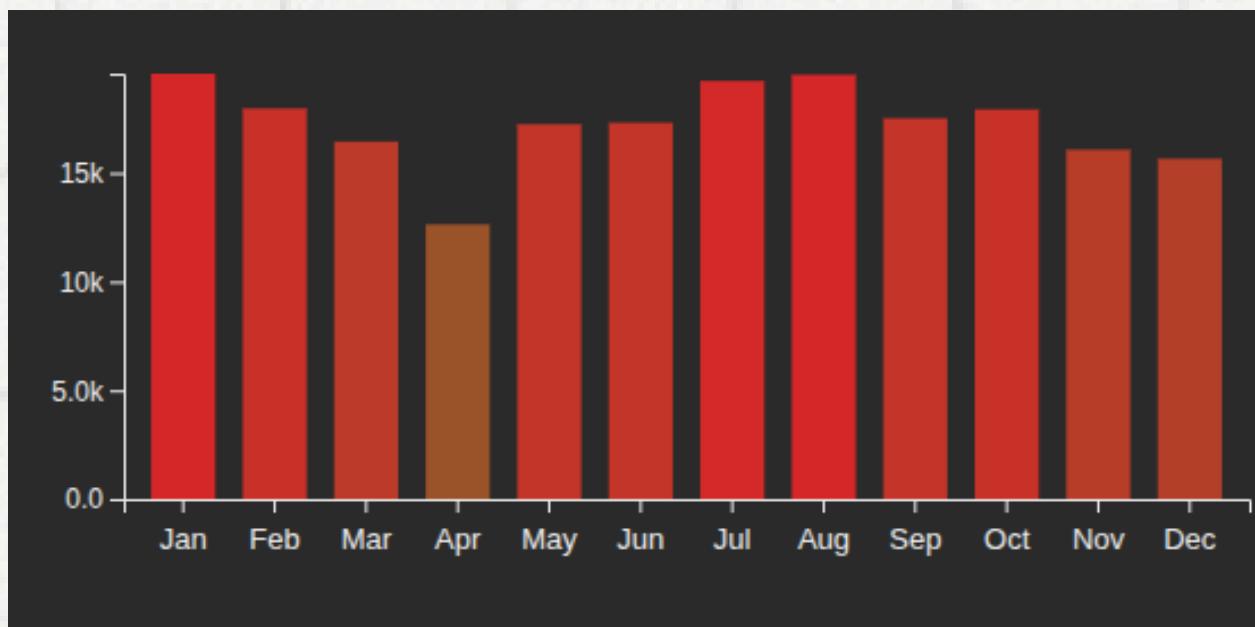


Dashboard

CrimeMonthChart: Muestra la cantidad de crímenes por mes para ver tendencias temporales.

CrimePorcentajeChart: Representa el porcentaje de cada tipo de crimen respecto al total.

CrimesByHourChart: Muestra crímenes por hora para identificar patrones diarios.

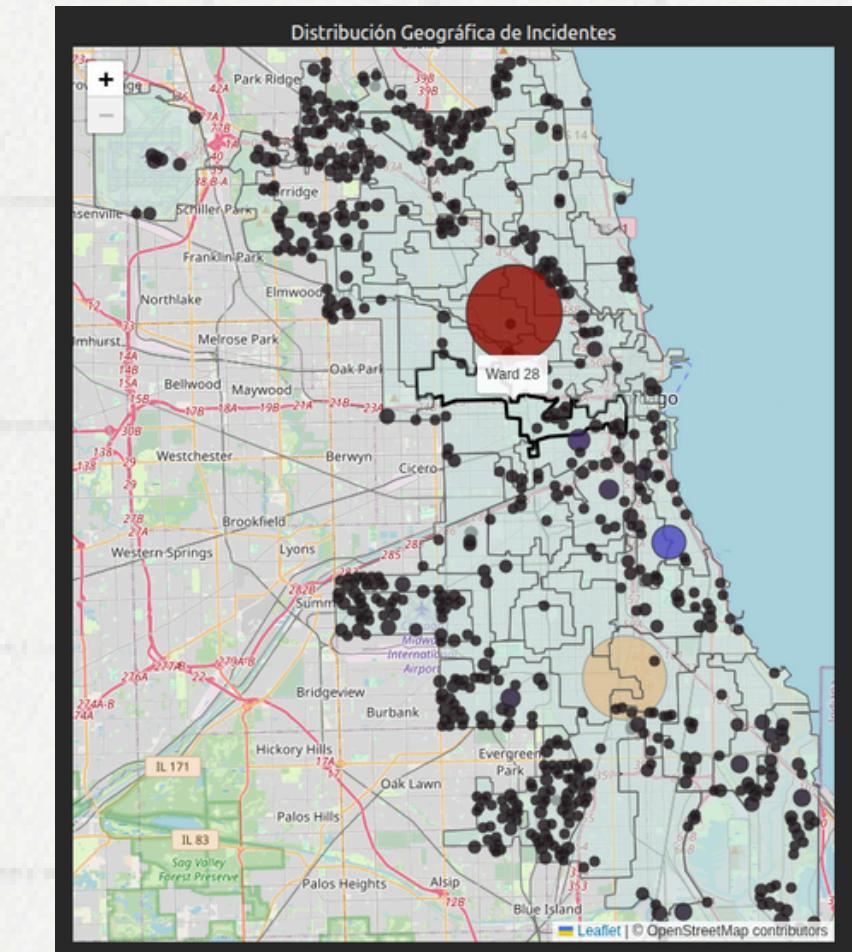
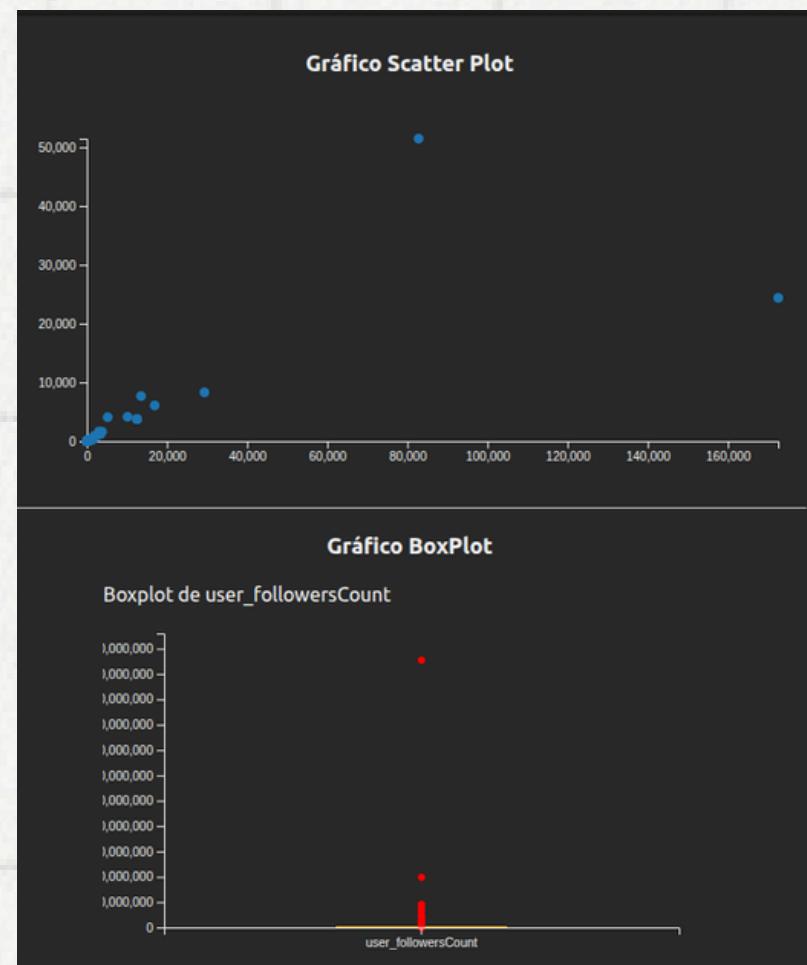


Dashboard

CrimeStationChart: Distribución de crímenes por distrito o estación geográfica.

CrimesMapChart: Ubica crímenes en el mapa para ver zonas con alta incidencia.

Tweets Boxplot y ScatterPlot: Visualizan dispersión y relaciones en datos de tweets relacionados con crímenes.

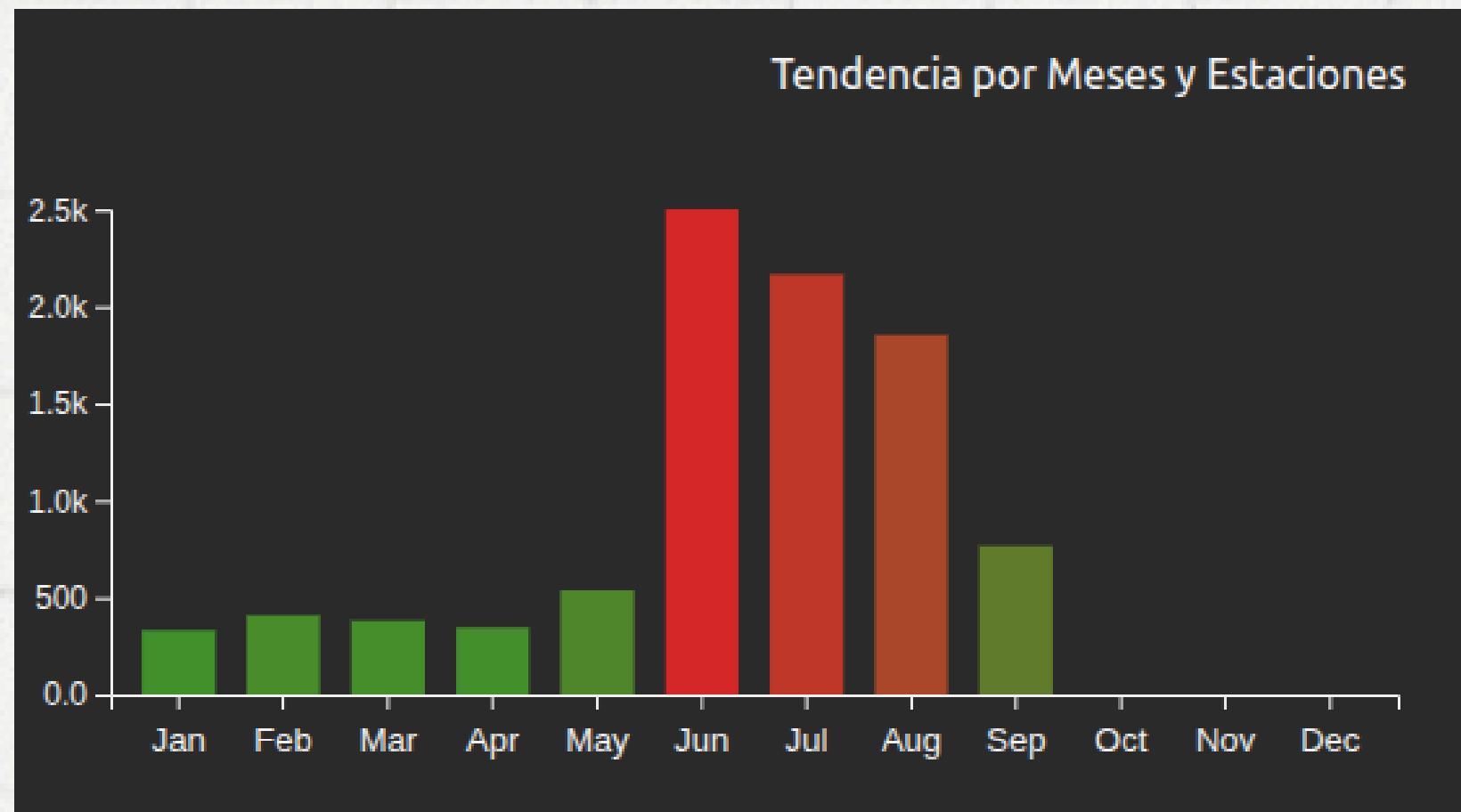


Hipótesis

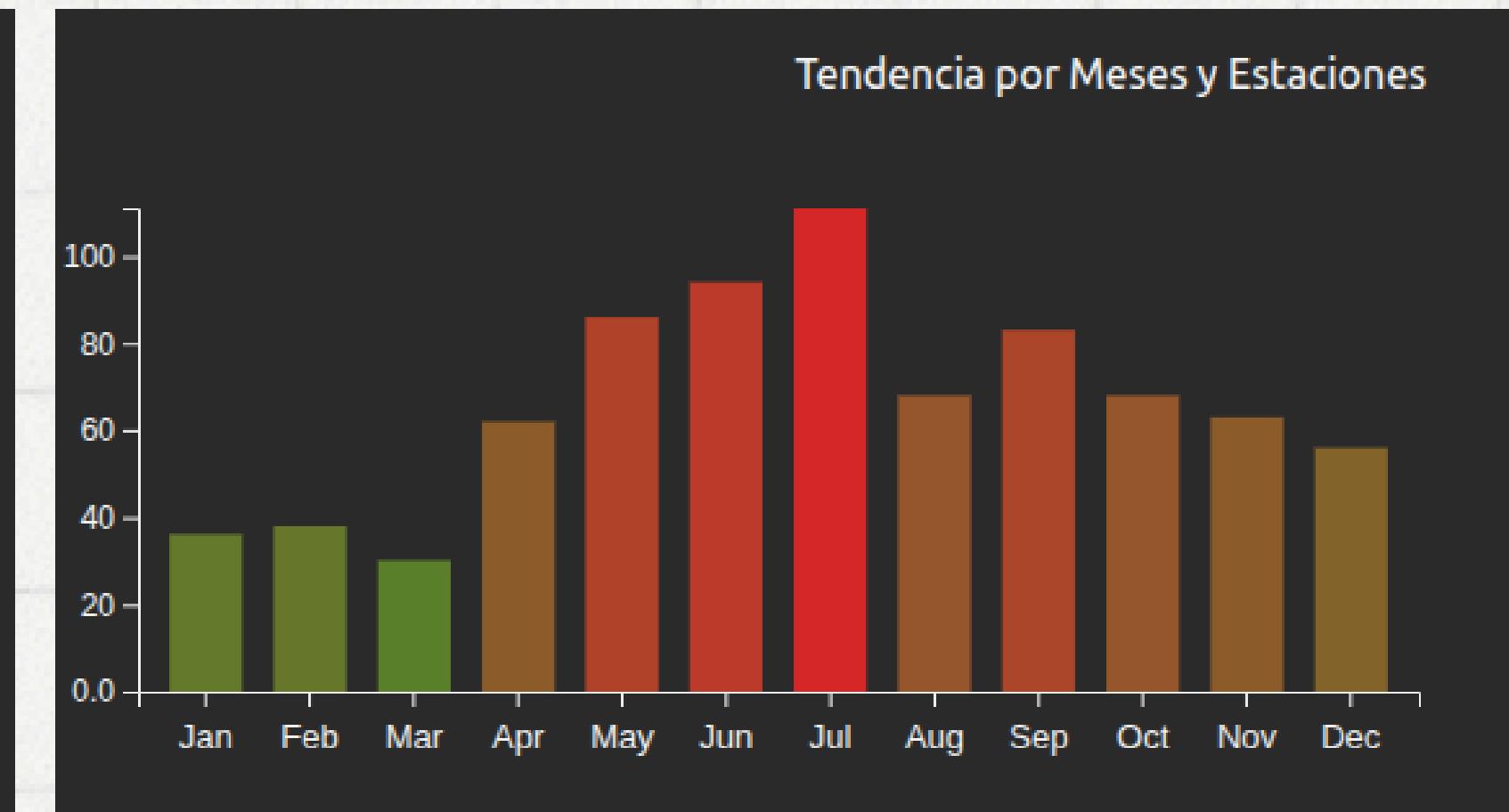
Hipótesis

Los usuarios de Twitter se enfocan más en crímenes impactantes como homicidios, mientras que en los datos oficiales la proporción de estos delitos es menor.

Homicidios Tweets

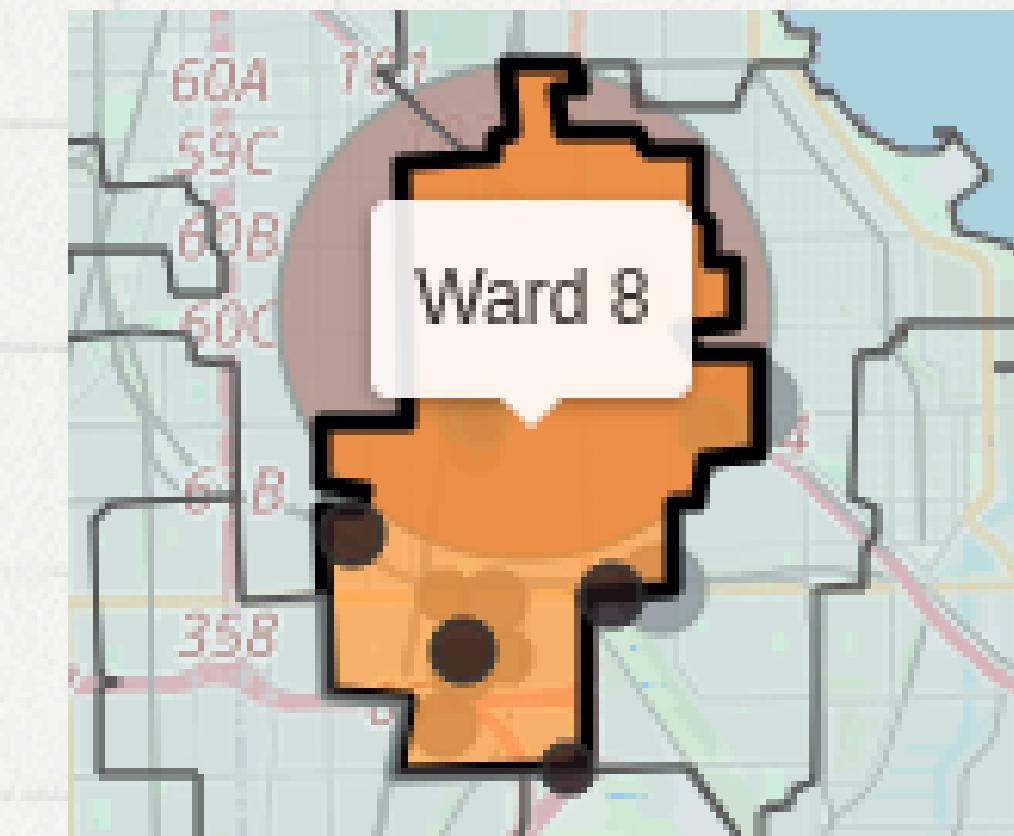
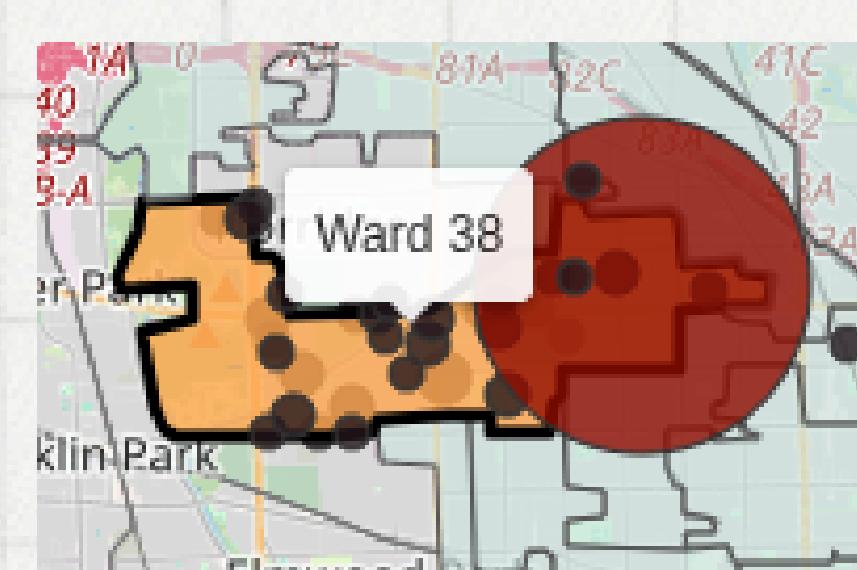
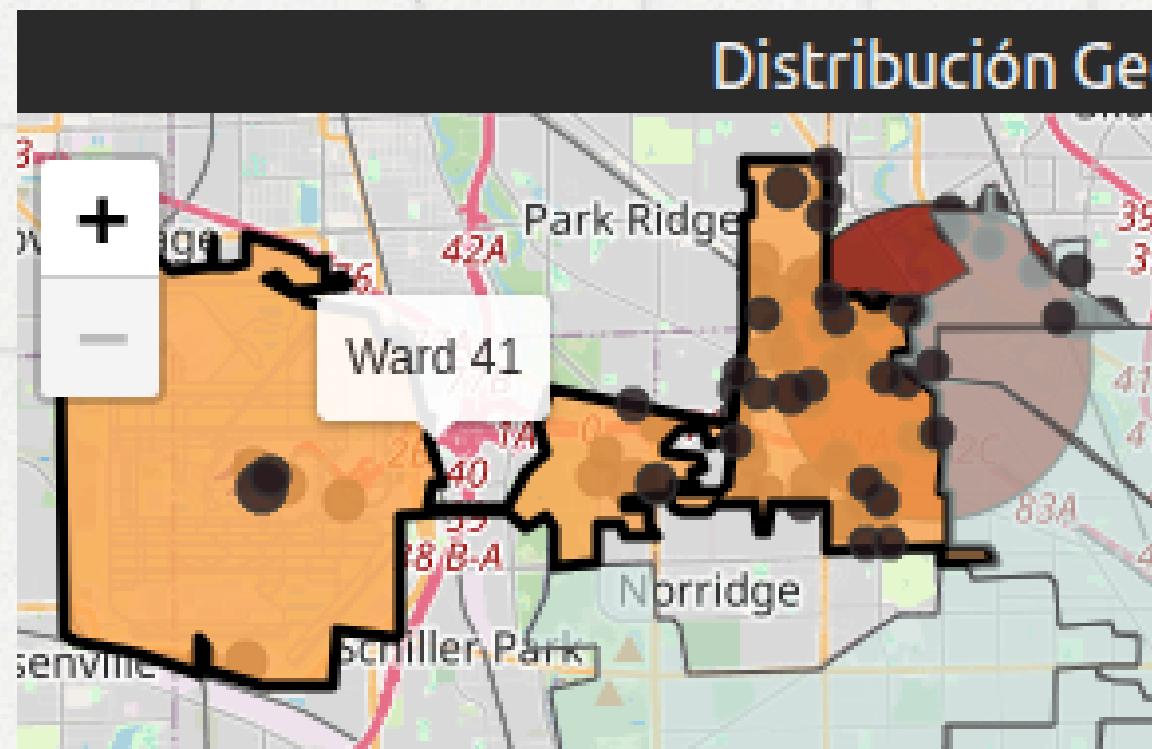
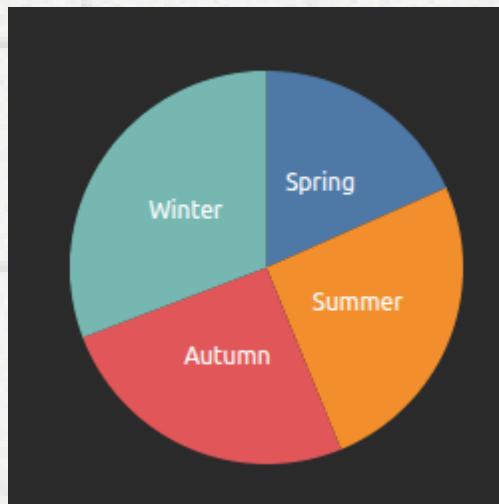


Homicidios Gobierno



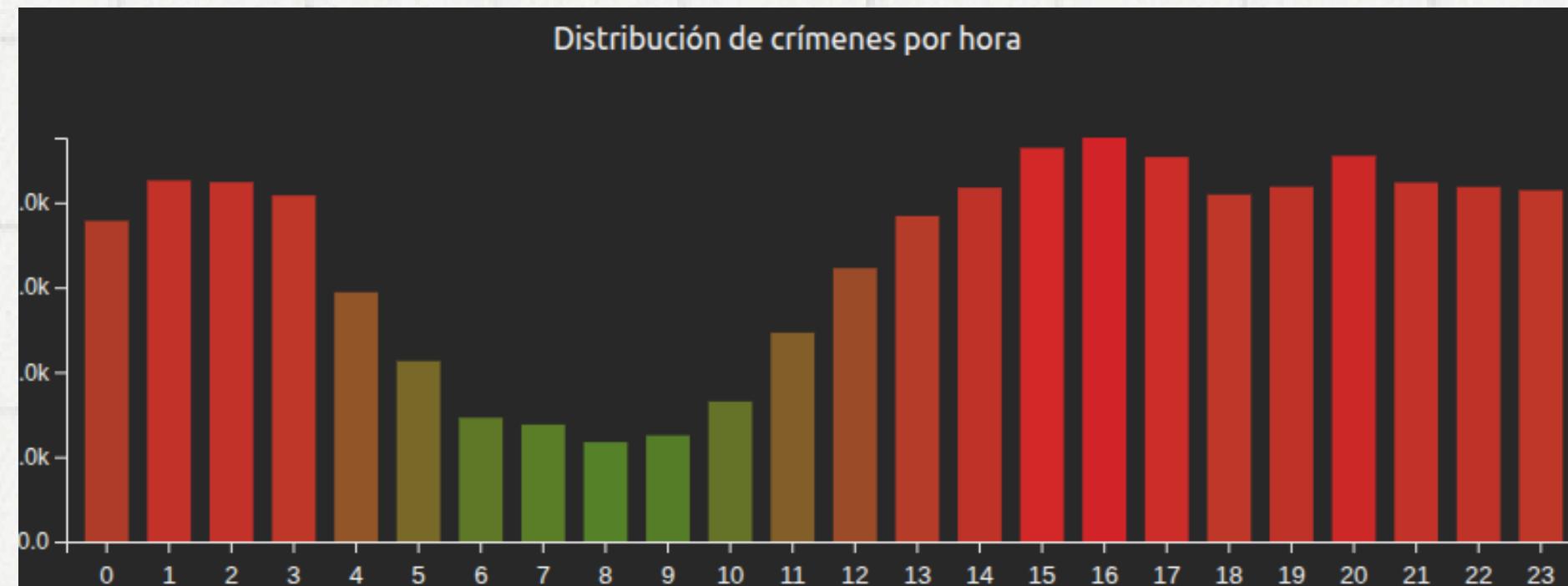
Hipótesis

Algunos distritos mantienen una frecuencia de crímenes constante durante el año, mientras que otros presentan variaciones marcadas según la estación.



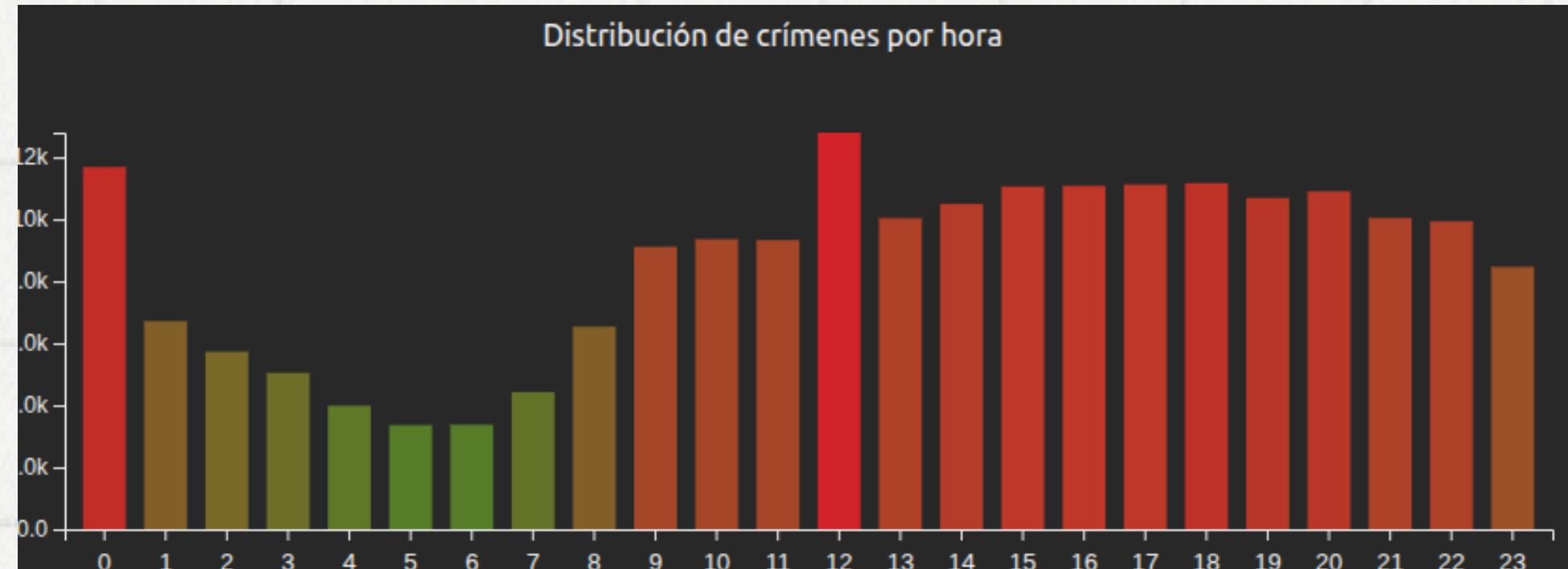
Hipótesis

Los reportes oficiales se publican mayormente en horario laboral, mientras que los comentarios en redes sociales aumentan en horas no laborales, especialmente de madrugada.



Tweets

Gobierno



Conclusiones

El análisis mostró patrones claros en tiempo y lugar de los crímenes, con diferencias entre distritos y estaciones. Los tweets resaltan crímenes graves, mientras los reportes oficiales son más variados. Además, reportes oficiales ocurren en horario laboral y los comentarios en redes, en horarios nocturnos y fines de semana.

Link

<https://dashboard-crime-tcd.vercel.app>

Gracias