Data Wrangling

Leon Felipe Davis Coropuna

Orientador: Prof Ana Maria Cuadros Valdivia

Plan de Tesis presentado la Escuela Profesional Ciencia de la Computación como paso previo a la elaboración de la Tesis Profesional.

UNSA - Universidad Nacional de San Agustín de Arequipa
Junio de 2025

Índice

1. Introducción

El objeto de estudio en este análisis es la relación entre los patrones de criminalidad reportados oficialmente (Dataset de Chicago) y la mención de incidentes delictivos en redes sociales (Dataset de tweets). La etapa de data wrangling tiene como base el trabajo de [İlgün and Dener, 2025]

1.1. Tamaño del dataset

1.1.1. Crímenes en Chicago

Este dataset contiene aproximadamente **8.32 millones de registros**, cada uno representando un crimen reportado con detalles como fecha, tipo, ubicación, arresto, entre otros. Aunque su tamaño bruto es de 3.9 GB, en memoria RAM puede superar los 5 GB debido al manejo interno de pandas.

1.1.2. Tweets 2020

Este dataset contiene **15,033 registros**, cada uno correspondiente a un tweet con metadatos asociados. El volumen no representa problemas de procesamiento, aunque su cobertura temporal es limitada. El tamaño bruto es pequeño, pero también se incrementa en memoria, posteriormente se planea seguir con el scrape lo que aumentaría considerablemente el dataset.

2. Crimes - 2001 to Present

El objeto de estudio es un incidente criminal reportado oficialmente en la ciudad de Chicago. El conjunto "Crimes – 2001 to Present" recoge incidentes delictivos reportados en la Ciudad de Chicago desde el año 2001 hasta la actualidad (con un desfase de siete días para garantizar la calidad de los datos). Esta información es provista diariamente por el sistema CLEAR (Citizen Law Enforcement Analysis and Reporting) del Departamento de Policía de Chicago. Para proteger la privacidad de las víctimas, las direcciones se muestran únicamente a nivel de manzana y no se identifican ubicaciones exactas. Aunque los datos iniciales pueden basarse en primeros reportes sin verificar (y sus clasificaciones podrían cambiar tras investigaciones posteriores), ofrecen una ventana única para estudiar la evolución y distribución espacial de la criminalidad en una de las mayores metrópolis de Estados Unidos [Chicago Police Department, 2025]. El dataset contiene la siguiente información en la tabla 1.

2.1. Uso de describe()

Como se puede apreciar en la Figura 1, la mayoría de las variables presentan valores de media y desviación estándar relativamente estables y adecuados, lo que indica una distribución más o menos homogénea. Sin embargo, para las variables espaciales `X` y `Y` se observan cambios más pronunciados en estas métricas, lo cual se debe a que son projecciones de Plane Illinois East NAD 1983.

2.2. Identificación de outliers

En la figura 10 se muestra que no existen o no hay muchos outliers mas que en X y Y las cuales pertenecen a un sistema de proyecciones propios de Illiois.

2.3. Renombramiento de Columnas

Para facilitar la comprensión y estandarizar los nombres de las columnas, se realizó el renombramiento de variables clave del conjunto de datos. Esto permite una manipulación más clara durante el análisis, evitando ambigüedades en nombres como `id`, `case_number` o `IUCR`, como se puede ver en la figura 3.

2.4. Eliminación de Duplicados y Valores Nulos

Como parte de la limpieza básica, se eliminaron las filas con valores nulos y aquellas duplicadas para evitar sesgos en los análisis estadísticos. Se calculó el porcentaje de registros eliminados para documentar el impacto de esta limpieza. La cantidad de registros totales se muestra en la figura 4.

2.5. Enriquecimiento Espacial

A partir de las coordenadas `X` e `Y`, se generaron nuevas variables espaciales. Se transformaron las coordenadas cartesianas a coordenadas polares, obteniendo el radio y el ángulo de cada punto. Además, se aplicó *KMeans* sobre las coordenadas escaladas para identificar zonas similares, y se introdujeron transformaciones geométricas mediante rotaciones de 30°, 45° y 60° para capturar relaciones espaciales que podrían no ser evidentes en la representación original.

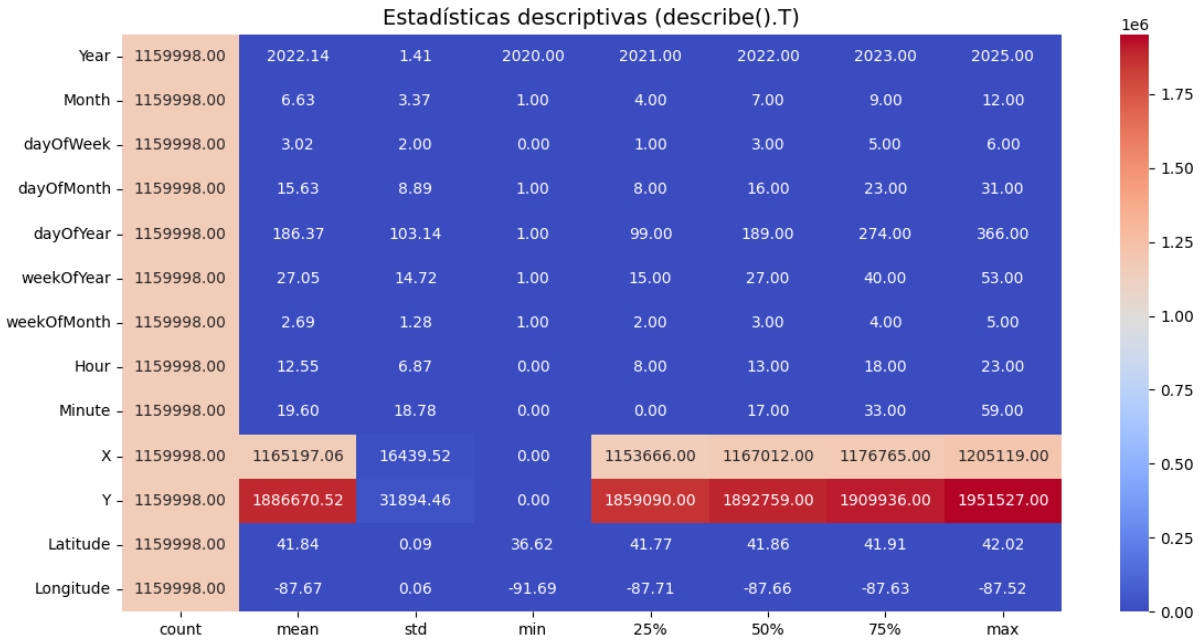


Figura 1: Estadísticas descriptivas transpuestas (`describe().T`) de las variables numéricas originales. Se destacan medias y desviaciones estándar.

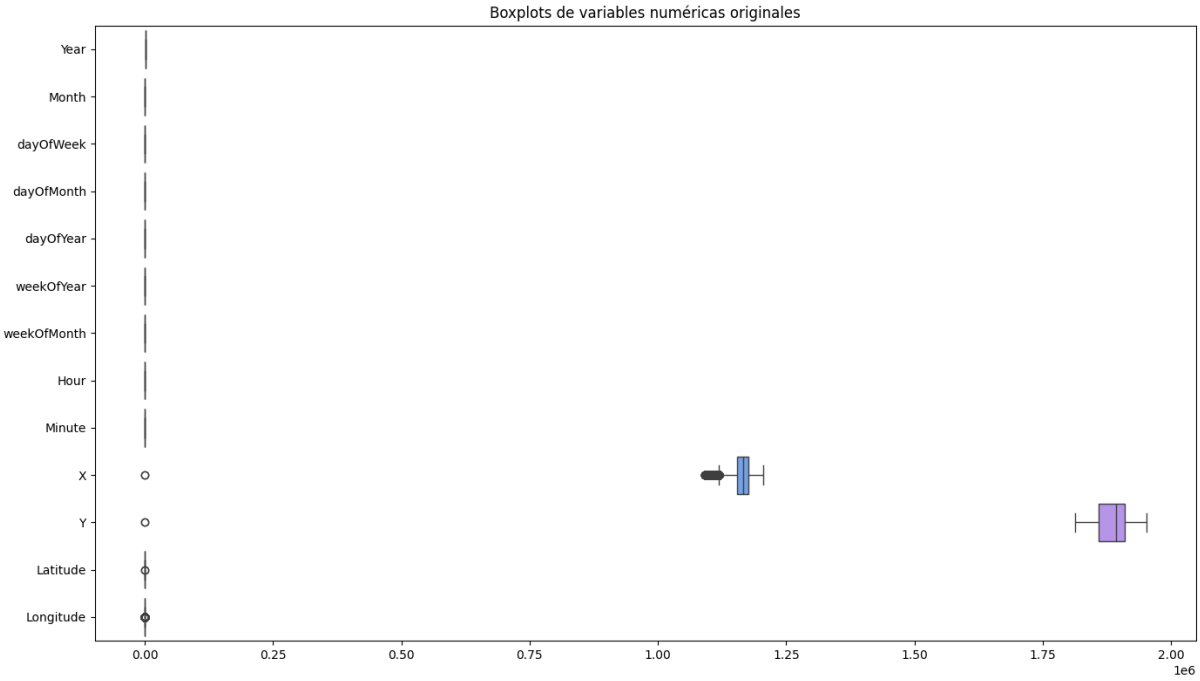


Figura 2: Outliers de cada variable no categórica

```python
df = df.rename(columns={
    'id': 'ID',
    'case_number': 'INumber',
    'IUCR': 'ICode',
    'date': 'Date',
    'district': 'District',
    'X Coordinate': 'X',
    'Y Coordinate': 'Y',
    'Primary Type': 'Category',
    'FBI Code': 'FCode'
})
```

Figura 3: Renombramiento de columnas

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

Figura 4: Total de registros eliminados y su porcentaje

2.6. Enriquecimiento Temporal

A partir de la columna de fecha (`Date`), se derivaron múltiples variables temporales como el año, mes, día de la semana, semana del año y hora. También se construyeron segmentos horarios como mañana, tarde y noche, y se distinguieron días laborables de fines de semana. Finalmente, se añadieron variables que indican la estación del año y si el día corresponde a un feriado según el calendario de EE.UU. Estas variables permiten analizar patrones delictivos en función de temporalidades específicas.

2.7. Codificación de Variables Categóricas

Se utilizó *Label Encoding* para convertir variables categóricas a formato numérico, lo cual es esencial para algoritmos que no pueden trabajar con datos no numéricos. Esta codificación se aplicó a columnas como el código IUCR, distrito, zona horaria, estación del año y marcadores booleanos derivados.

2.8. Selección de Características

Para evaluar la relevancia de las variables, se utilizó el método de **Información Mutua** (*Mutual Information*), el cual mide cuánta información proporciona una característica sobre la variable objetivo (`Category`). El proceso consistió en seleccionar las columnas

previamente identificadas como relevantes, codificar las variables categóricas aún no convertidas, y aplicar `mutual_info_classif` de `sklearn` para obtener una puntuación por característica. Finalmente, se ordenaron las variables por importancia de manera decreciente. Este paso permitió identificar qué atributos temporales y espaciales resultan más útiles para entender el tipo de crimen. Los atributos seleccionados se muestran en la figura 5 y se comparan con los de un trabajo relacionado en la figura 6.

2.9. Resumen del Preprocesamiento

Como resultado final del *data wrangling*, se obtuvo un conjunto de datos limpio y codificado, listo para tareas de análisis exploratorio o modelado predictivo. La estructura final fue verificada mediante el método `info()` para asegurar que no haya columnas faltantes o con tipos incorrectos. La figura 7 muestra las columnas y sus tipos de datos en el `DataFrame` resultante.

3. Crime tweets X.com Chicago

El objeto de estudio es un tweet público que menciona o podría estar relacionado con actividades delictivas. El conjunto de datos "Crime Tweets from X.com" recoge mensajes públicos relacionados con incidentes delictivos en la Ciudad de Chicago obtenidos mediante scraping en la plataforma social X.com (anteriormente conocida como Twitter). Este dataset contiene publicaciones generadas por usuarios que mencionan eventos, percepciones y reportes de crímenes, complementando la información oficial con perspectivas sociales en tiempo real. El dataset presenta la siguiente información en la tabla 2

3.1. Uso de describe()

Como se puede apreciar en la figura 8, se puede apreciar que existen datos que salen por mucho de lo normal como son los seguidores de un usuario o la cantidad de likes o retweets que recibió algún comentario, esto podría indicar un crímen muy controversial o simplemente que lo publicó una persona muy conocida. Por el resto de variables se tiene una distribución adecuada. Además en la figura 9 se observa una distribución sesgada hacia la derecha lo que indica una distribución no gaussiana.

3.2. Identificación de outliers

En la figura 10 se muestran muchos outliers provenientes de columnas relacioinadas a usuarios.

	Feature	MI_Score
0	ICode	2.396172
1	Rot30_Y	0.216835
2	Radius	0.216441
3	Rot60_Y	0.214921
4	Rot60_X	0.213040
5	Rot30_X	0.211844
6	Angle	0.211687
7	Y	0.192693
8	Rot45_X	0.189836
9	Rot45_Y	0.184477
10	X	0.169264
11	Minute	0.081840
12	Cluster	0.061339
13	District	0.053931
14	Hour	0.046175
15	BusinessHour	0.035876
16	Hour_Zone	0.034600
17	Year	0.011988
18	Weekend	0.009897
19	Season	0.009304
20	dayOfYear	0.008485
21	weekOfMonth	0.006682
22	dayOfWeek	0.006550
23	weekOfYear	0.006334
24	Month	0.005412
25	dayOfMonth	0.002081
26	Holiday	0.000425

Figura 5: Atributos seleccionados según la selección de características

ICode	2.468959
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

Figura 6: Atributos seleccionados de un paper que usa un dataset similar

```
Data columns (total 28 columns):
 #   Column         Non-Null Count    Dtype
---  ------         --------------    -----
 0   Year           1159998 non-null  int32
 1   Month          1159998 non-null  int32
 2   dayOfWeek      1159998 non-null  int32
 3   dayOfMonth     1159998 non-null  int32
 4   dayOfYear      1159998 non-null  int32
 5   weekOfYear     1159998 non-null  int64
 6   weekOfMonth    1159998 non-null  int64
 7   Hour           1159998 non-null  int32
 8   Minute         1159998 non-null  int32
 9   Hour_Zone      1159998 non-null  int64
 10  BusinessHour   1159998 non-null  int64
 11  Weekend        1159998 non-null  int64
 12  Holiday        1159998 non-null  int64
 13  Season         1159998 non-null  int64
 14  X              1159998 non-null  float64
 15  Y              1159998 non-null  float64
 16  Radius         1159998 non-null  float64
 17  Angle          1159998 non-null  float64
 18  Cluster        1159998 non-null  int32
 19  Rot30_X        1159998 non-null  float64
 20  Rot30_Y        1159998 non-null  float64
 21  Rot45_X        1159998 non-null  float64
 22  Rot45_Y        1159998 non-null  float64
 23  Rot60_X        1159998 non-null  float64
 24  Rot60_Y        1159998 non-null  float64
 25  ICode          1159998 non-null  int64
 26  District       1159998 non-null  int64
 27  Category       1159998 non-null  object
dtypes: float64(10), int32(8), int64(9), object(1)
```

Figura 7: Información del dataframe resultante luego del preprocesamiento

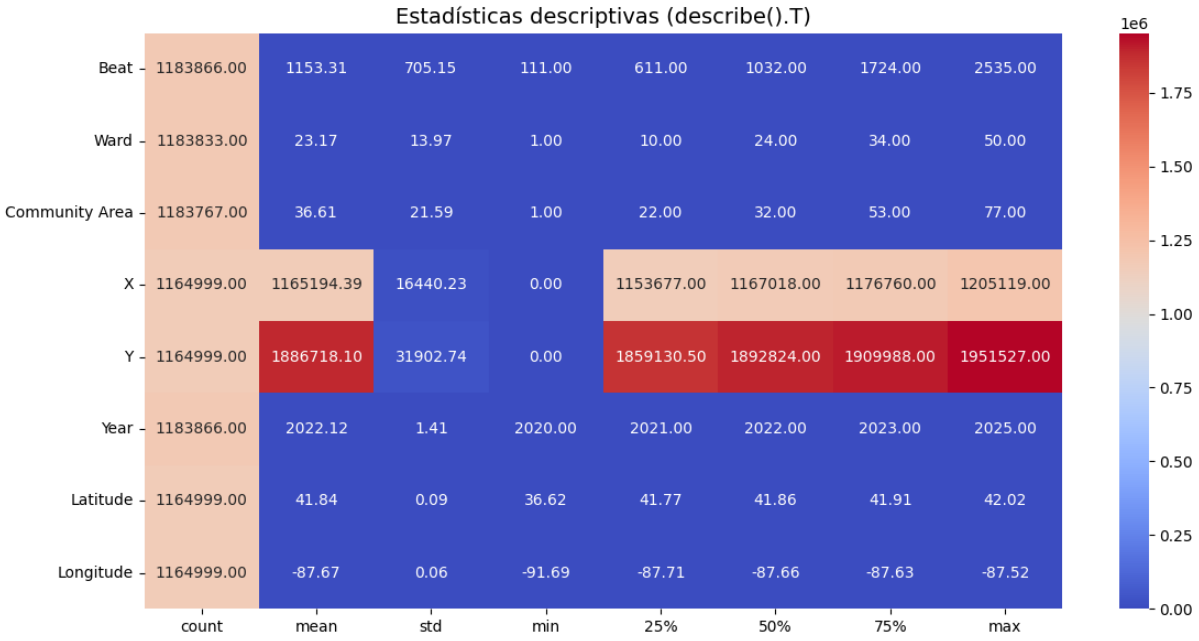


Figura 8: Histogramas de las variables numéricas originales, mostrando su distribución y variabilidad

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).

Cuadro 1: Descripción de columnas del dataset procesado para análisis de crimen

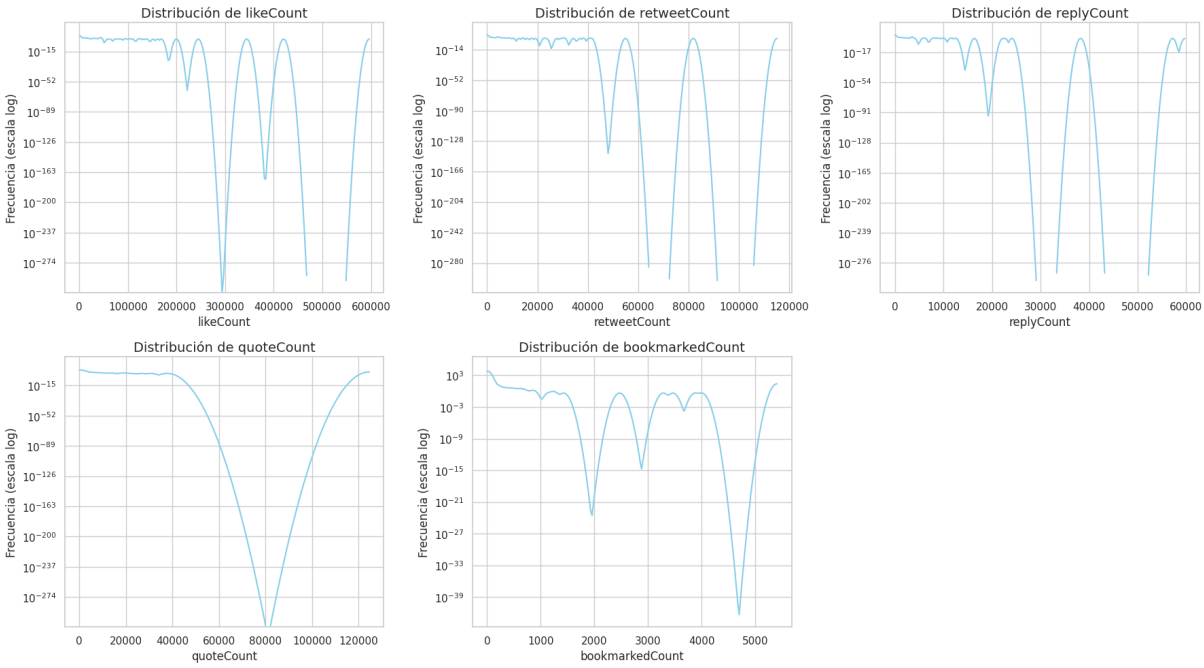


Figura 9: Distribución en escala logarítmica de variables numérticas

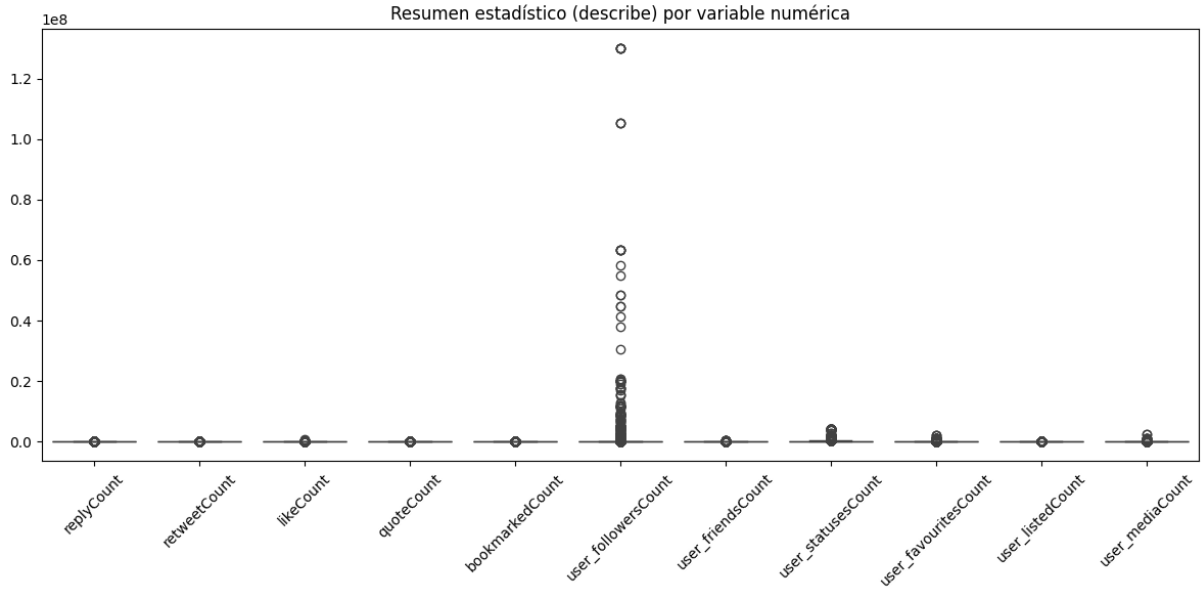
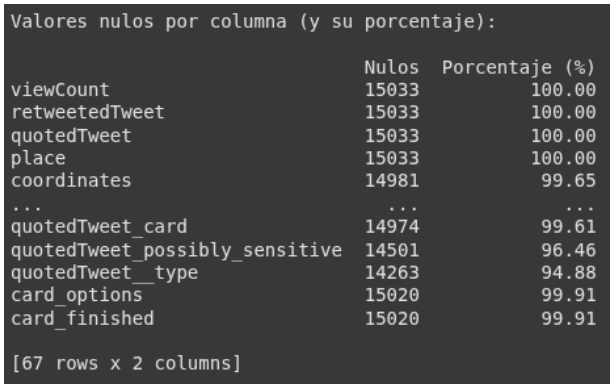


Figura 10: Outliers de cada variable no categórica de tweets



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]

Figura 11: Total de registros eliminados y su porcentaje en tweets

3.3. Renombramiento de Columnas

Para facilitar la comprensión y estandarizar los nombres de las columnas, con el fin de posteriormente aplicar un buen match entre el dataset de tweets y el dataset de crimenes oficiales.

3.4. Eliminación de Duplicados y Valores Nulos

En este caso existen muchísimas columnas vacías donde muchas de ellas directamente tienen 100 % de valores nulos por lo que es imposible de imputar, además de ello estas columnas directamente no representan información importante para el análisis ya que lo más importante son el comentario, la fecha y metadata del tweet más no del usuario, el resto es presindible 11.

3.5. Enriquecimiento Temporal

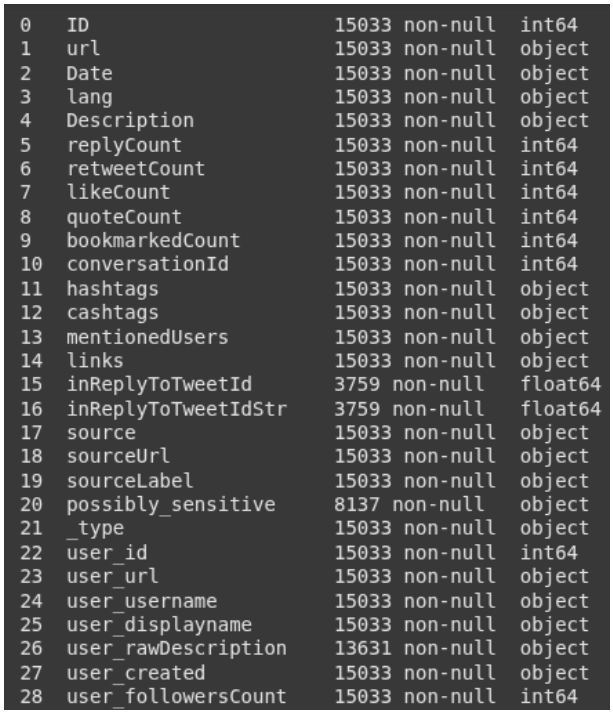
De manera similar al dataset de crimenes del gobierno, aquí también se generan nuevos datos para tener al alcance información directa sobre la fecha.

3.6. Resumen del Preprocesamiento

Como resultado final del *data wrangling*, se obtuvo un conjunto de datos listo para tareas como etiquetado para predecir el tipo de crimen comentado en el tweet, identificación de entidades (NER), etc. La estructura final se muestra en la figura 12.

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.

Cuadro 2: Descripción de columnas del dataset de tweets



0	ID	15033 non-null	int64
1	url	15033 non-null	object
2	Date	15033 non-null	object
3	lang	15033 non-null	object
4	Description	15033 non-null	object
5	replyCount	15033 non-null	int64
6	retweetCount	15033 non-null	int64
7	likeCount	15033 non-null	int64
8	quoteCount	15033 non-null	int64
9	bookmarkedCount	15033 non-null	int64
10	conversationId	15033 non-null	int64
11	hashtags	15033 non-null	object
12	cashtags	15033 non-null	object
13	mentionedUsers	15033 non-null	object
14	links	15033 non-null	object
15	inReplyToTweetId	3759 non-null	float64
16	inReplyToTweetIdStr	3759 non-null	float64
17	source	15033 non-null	object
18	sourceUrl	15033 non-null	object
19	sourceLabel	15033 non-null	object
20	possibly_sensitive	8137 non-null	object
21	_type	15033 non-null	object
22	user_id	15033 non-null	int64
23	user_url	15033 non-null	object
24	user_username	15033 non-null	object
25	user_displayname	15033 non-null	object
26	user_rawDescription	13631 non-null	object
27	user_created	15033 non-null	object
28	user_followersCount	15033 non-null	int64

Figura 12: Información del dataframe resultante luego del preprocesamiento del df tweets

Referencias

[Chicago Police Department, 2025] Chicago Police Department (2025). Crimes - 2001 to present. https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2/about$_data. Datasetactualizadoal29demayode2025. ProporcionadoporChicagoPoliceDepartmen$

[İlgün and Dener, 2025] İlgün, E. G. and Dener, M. (2025). Exploratory data analysis, time series analysis, crime type prediction, and trend forecasting in crime data using machine learning, deep learning, and statistical methods. *Neural Computing and Applications*.