Análisis de la Incidencia Delictiva en México para la Toma de Decisiones en Seguridad Pública

Alumno: Gerardo Ulises Gonzalez Romero

Materia: Introducción a la Ciencia de Datos

Maestro: Jaime Alejandro Romero Sierra

28/11/2025

Objetivo del Proyecto

Identificar patrones y tendencias delictivas en México para generar información útil que mejore la toma de decisiones en seguridad pública. El objetivo es reducir los delitos en un 10% en las zonas más peligrosas en 24 meses, optimizando recursos, fortaleciendo la prevención y mejorando la percepción ciudadana. También se busca aumentar la eficiencia gubernamental mediante un uso más estratégico del presupuesto y acciones preventivas por parte de las fuerzas de seguridad.

¿Por qué es importante resolver o estudiar esta problemática?

La inseguridad no solo se mide en cifras: se siente en las calles, en las casas, en la forma en que vivimos. Cada delito representa una historia rota, una comunidad que pierde confianza, un gobierno que necesita actuar mejor. Estudiar esta problemática nos permite entender qué está pasando, dónde y por qué, para tomar decisiones más inteligentes y humanas.

Cuando usamos los datos para prevenir en lugar de reaccionar, cuando los recursos se asignan con estrategia, y cuando la gente empieza a sentirse más segura, no solo bajan los delitos: mejora la vida. Este proyecto busca que el cambio se note, que se sienta, y que sea sostenible.

Fuentes de datos:

• Dataset Kaggle “Crime rates - México”: Base principal con registros de delitos clasificados por tipo y entidad.

• INEGI (Instituto Nacional de Estadística y Geografía): Encuestas de victimización y percepción de seguridad.

• Secretariado Ejecutivo del Sistema Nacional de Seguridad Pública (SESNSP): Reportes oficiales de incidencia delictiva.

• Datos socioeconómicos (CONAPO, Banco de México): Indicadores de empleo, pobreza y desarrollo.

• Datos de percepción ciudadana (ENCVI, ENVIPE): Encuestas sobre confianza en instituciones y percepción de inseguridad. La combinación de estas fuentes permitirá un análisis integral que no solo mida delitos, sino también contexto social y económico.

Cantidad de datos:

- **84,233 (filas)**
- **21 variables (columnas)**

Principales características del conjunto de datos:

El conjunto de datos limpio contiene 84,233 registros y 21 variables. Cada registro representa una combinación única de entidad federativa, tipo de delito, modalidad y número de casos registrados por mes, entre los años 2015 y 2023.

Variables clave

- **Año:** Indica el periodo temporal del registro (2015–2023).
- **Entidad y Clave_Ent:** Identifican la entidad federativa donde ocurrió el delito.
- **Tipo de delito, Subtipo de delito y Modalidad:** Clasifican legalmente el hecho delictivo.
- **Bien jurídico afectado:** Describe el derecho o interés protegido por la ley que fue vulnerado.
- **Sexo/Averiguación previa y Rango de edad:** Proporcionan información demográfica asociada al registro.
- **Enero a Diciembre:** Representan el número de delitos registrados por mes.
- **Total anual:** Variable agregada que suma los delitos mensuales por registro.

Metodología

Proceso de limpieza de datos:

Esta base de datos contiene información sobre el índice de criminalidad en México, esta base tiene como finalidad ser de apoyo para los departamentos policiales de México para su buen uso, en contra de los diferentes crímenes registrados y su respectivo combate de ello.

En esta base encontraremos diversas columnas, las cuales son de gran ayuda para análisis próximos. las columnas con las cuales contamos son las siguientes: Año, Clave_Ent(Clave de Entidad), Bien jurídico afectado, Tipo de delito, Subtipo de delito, Modalidad, Meses del Año, sexo, Rango de edad.

De igual forma tenemos Renglones, de los 32 estados de México, de esta forma de se obtiene un fácil acceso a la información deseada.

No obstante, la base tiene valores repetidos, sin valor o mal catalogados etc.… aquí mismo se lleva el proceso de limpieza el cual es explicado en cada paso.

Significado de cada columna:

Año: Año del crimen

Clave_Ent: La clave de entidad es una propiedad o un conjunto de propiedades de un tipo de entidad que se utiliza para determinar la identidad de una instancia dentro de un conjunto de entidades en un modelo de datos.

Bien Jurídico: El bien jurídico se refiere a aquellos bienes, tanto materiales como inmateriales, que son considerados de sumo valor por la sociedad y que gozan de protección por parte del derecho, ya sea a través de otras ramas del derecho o específicamente del derecho penal.

Tipo de delito: Un delito se define como una conducta típica, antijurídica, imputable, culpable y sometida a una sanción penal, que constituye una infracción del derecho penal. Esta conducta puede manifestarse como una acción o una omisión, y debe estar prevista y penada por la ley.

Subtipo de delito: El término "subtipo de delito" no se encuentra directamente definido en los contextos proporcionados, pero se puede inferir su significado a partir de la clasificación detallada de los delitos según sus elementos y características.

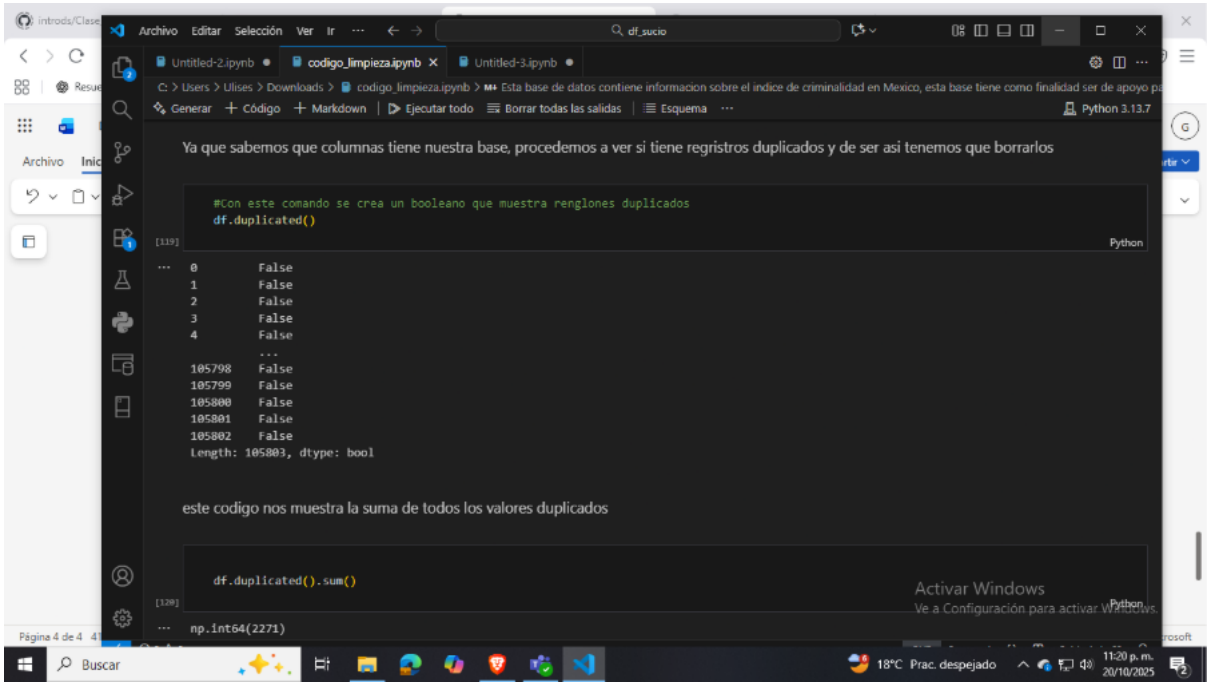
Modalidad: se refiere a las circunstancias de lugar, tiempo, modo u ocasión que califican la conducta del sujeto activo y que tienen como función determinar el quantum de la pena, afectando su medida, ya sea aumentándola o disminuyéndola.

Meses del año: Enero, febrero, marzo, abril, mayo, junio, Julio, agosto, septiembre, octubre, noviembre, diciembre

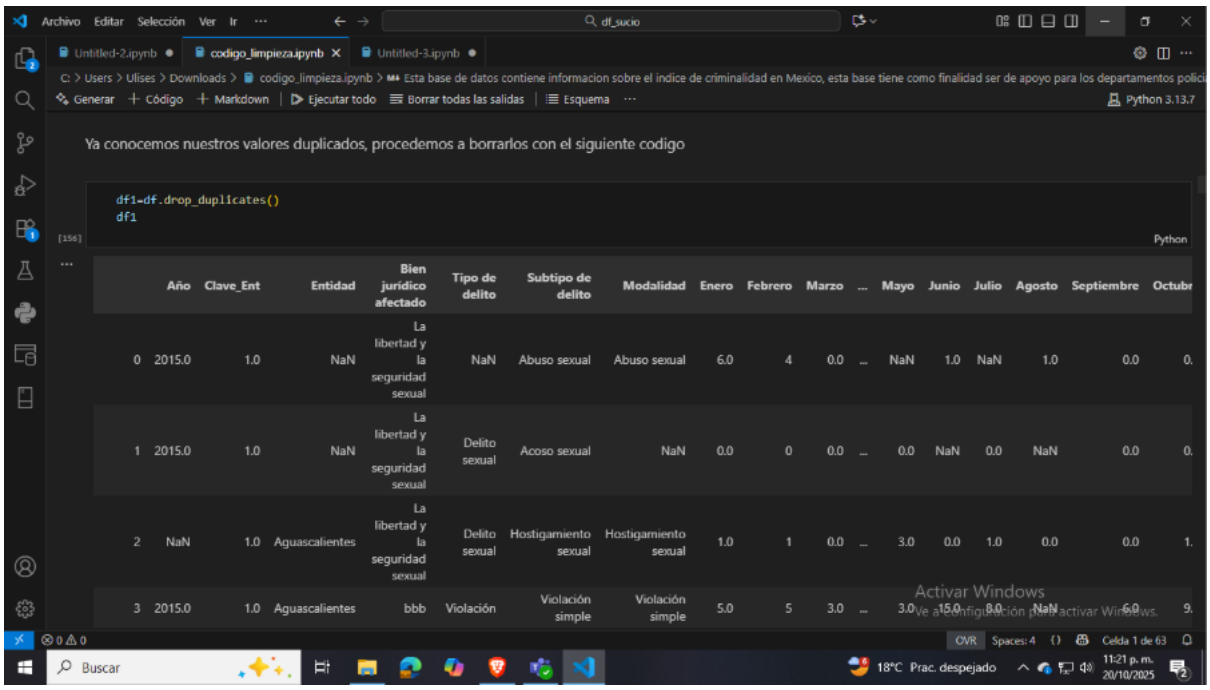
Sexo: Masculino/Femenino

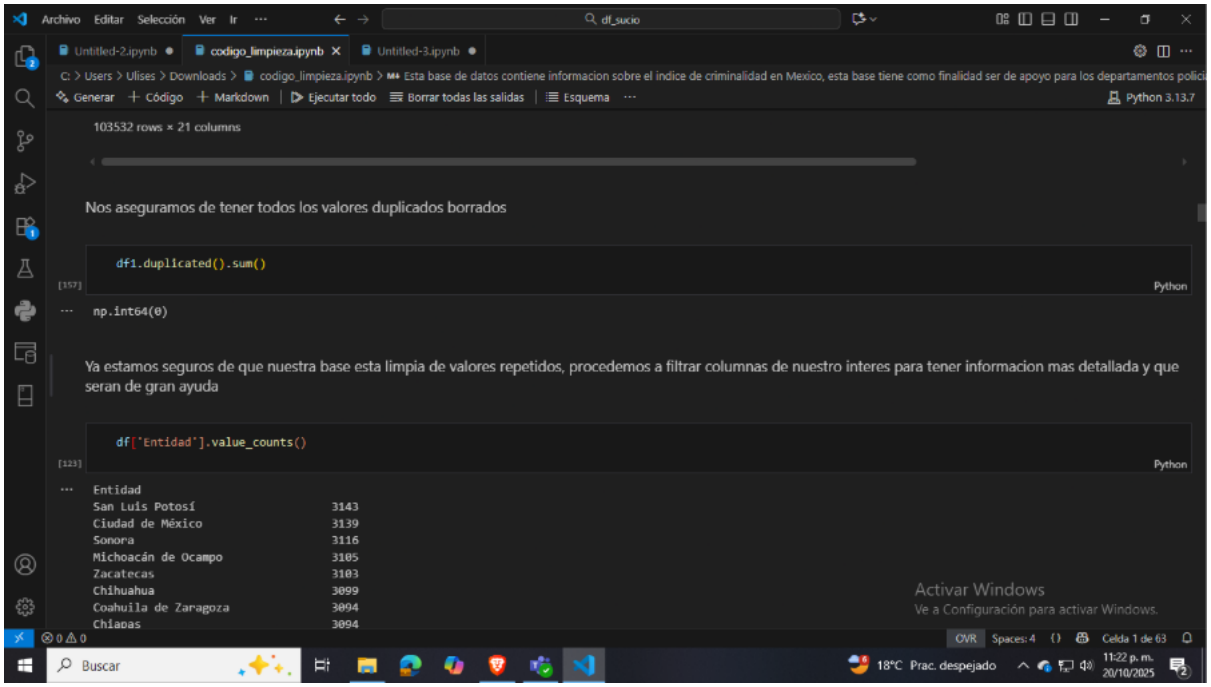
Rango de edad: Menores de edad 0-17, Mayores de edad 18-100

Primeramente, lo primero que hice al limpiar la base de datos fue borrar los duplicados:

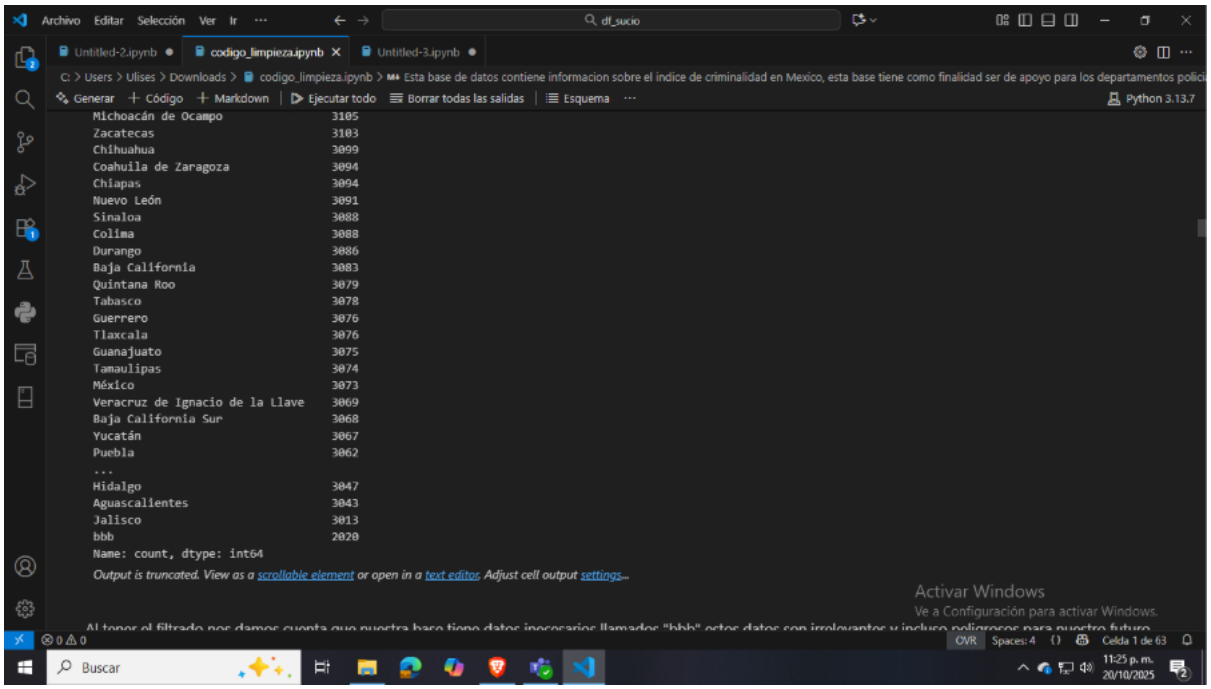


Después de ver que efectivamente contiene duplicados, pase a la tarea de borrar todos los valores duplicados

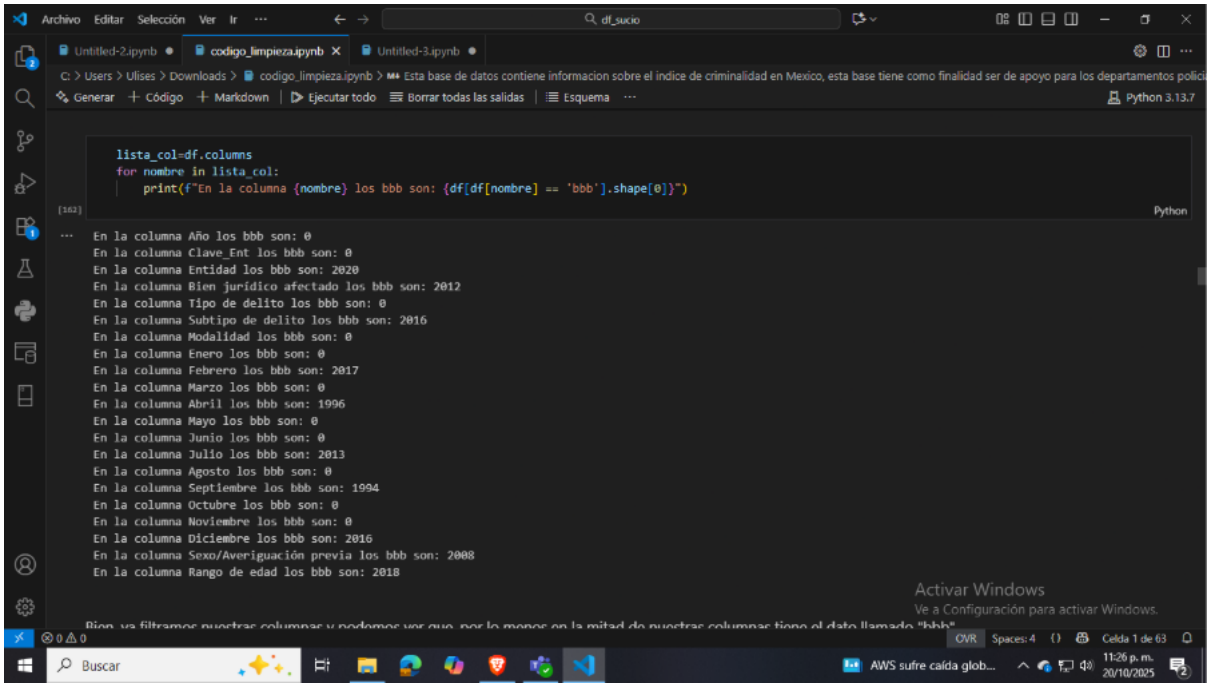


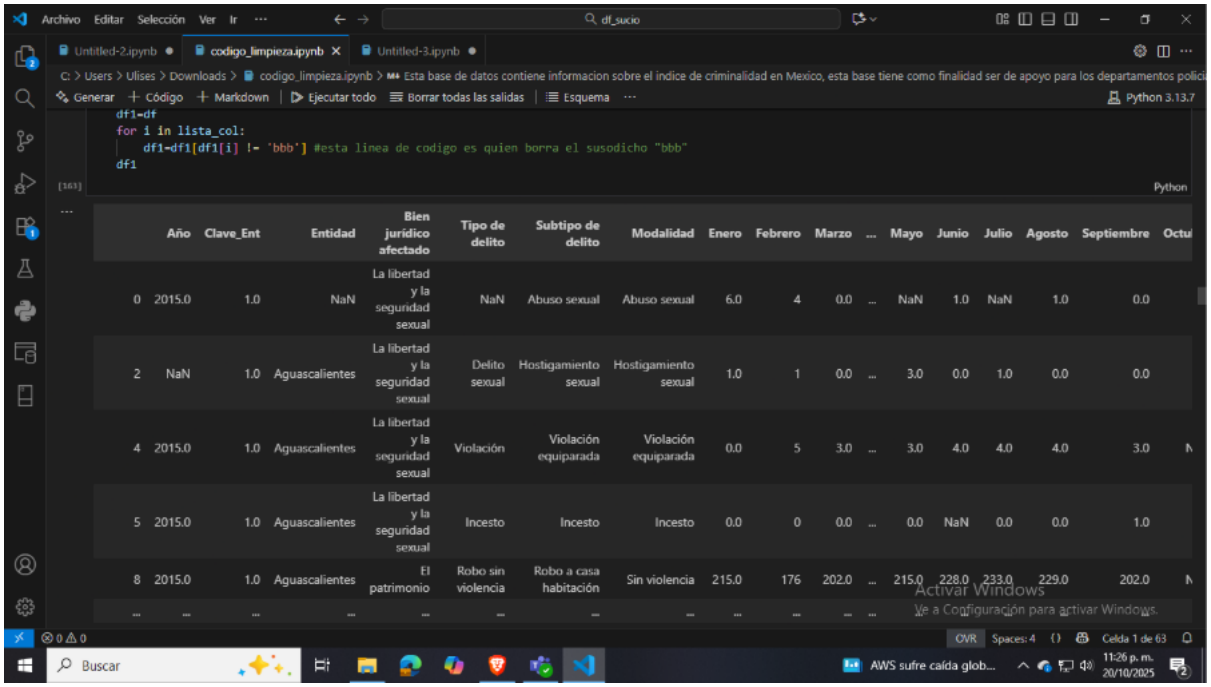


Consecutivamente filtré algunas columnas y encontré valores inválidos

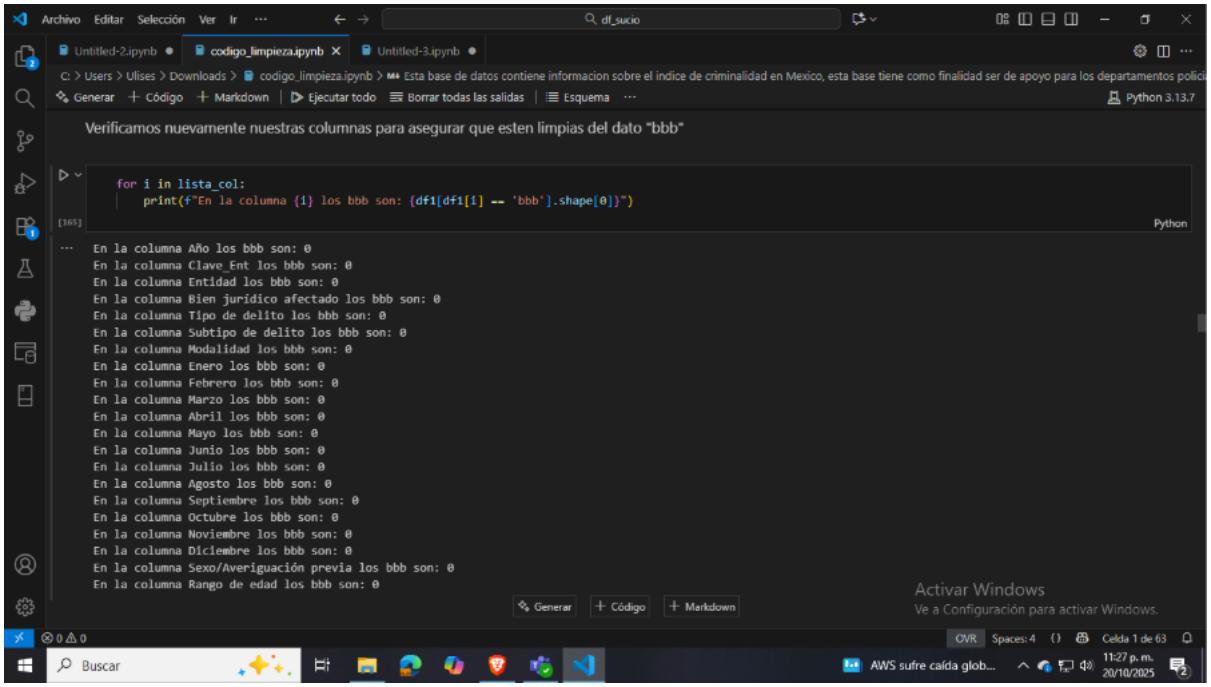


Y procedí a borrar cada uno de cada columna donde hubiera uno

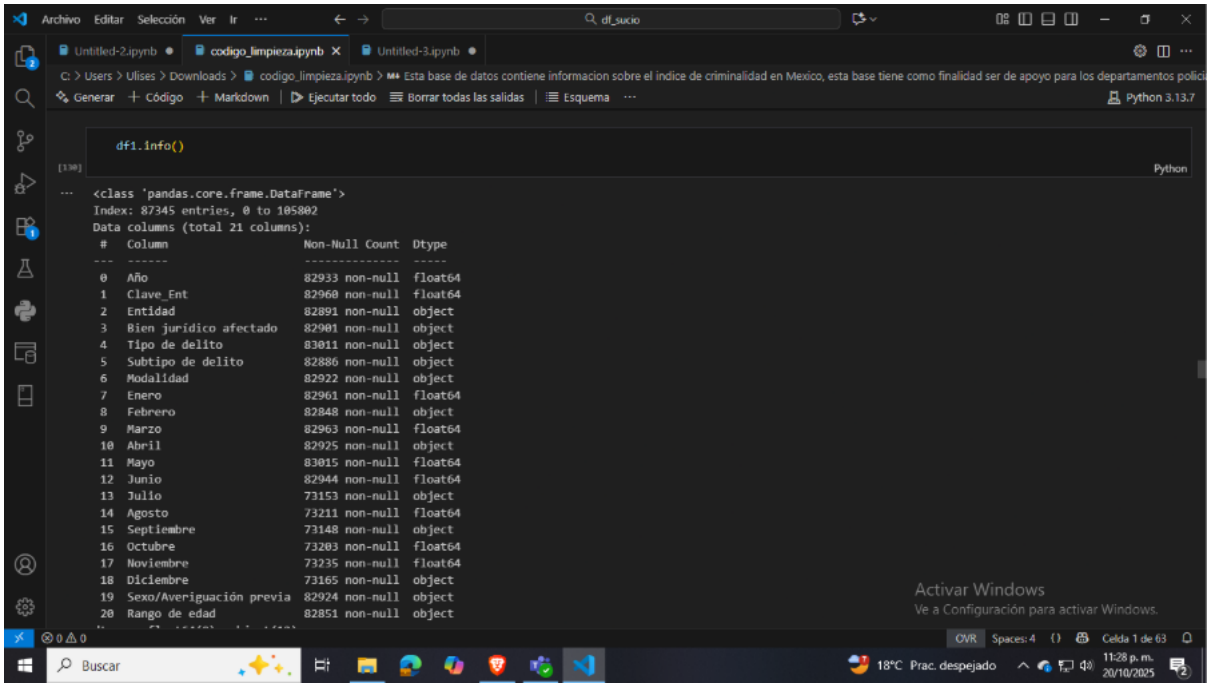


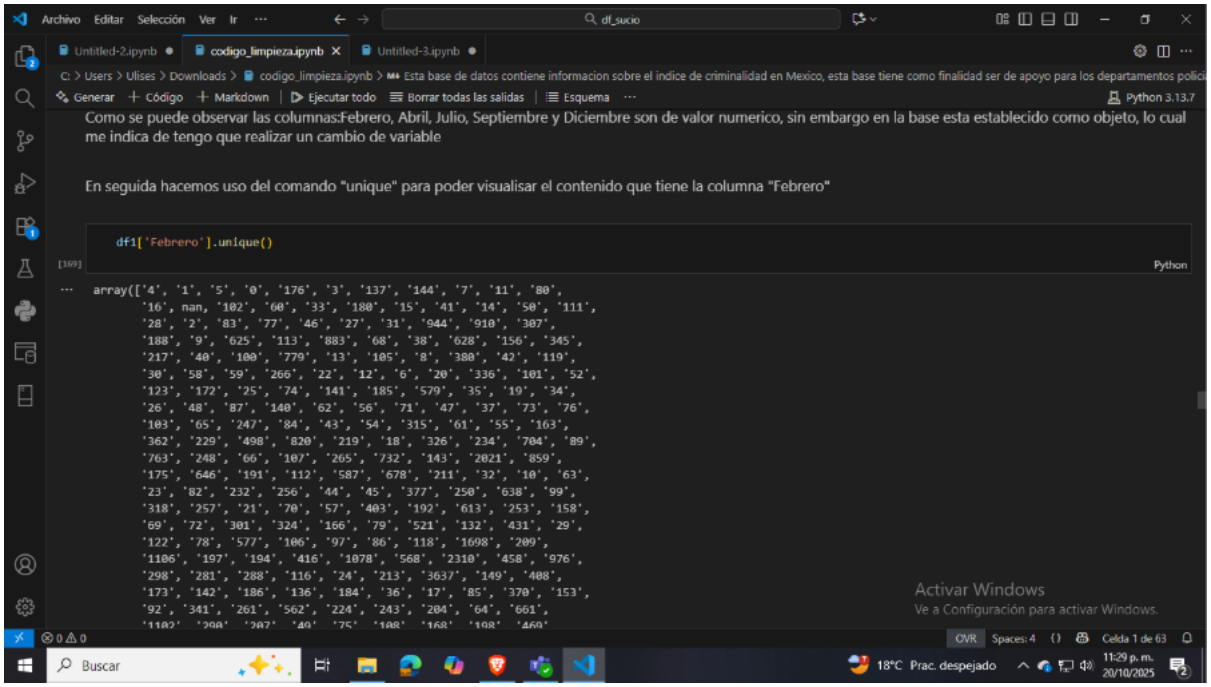


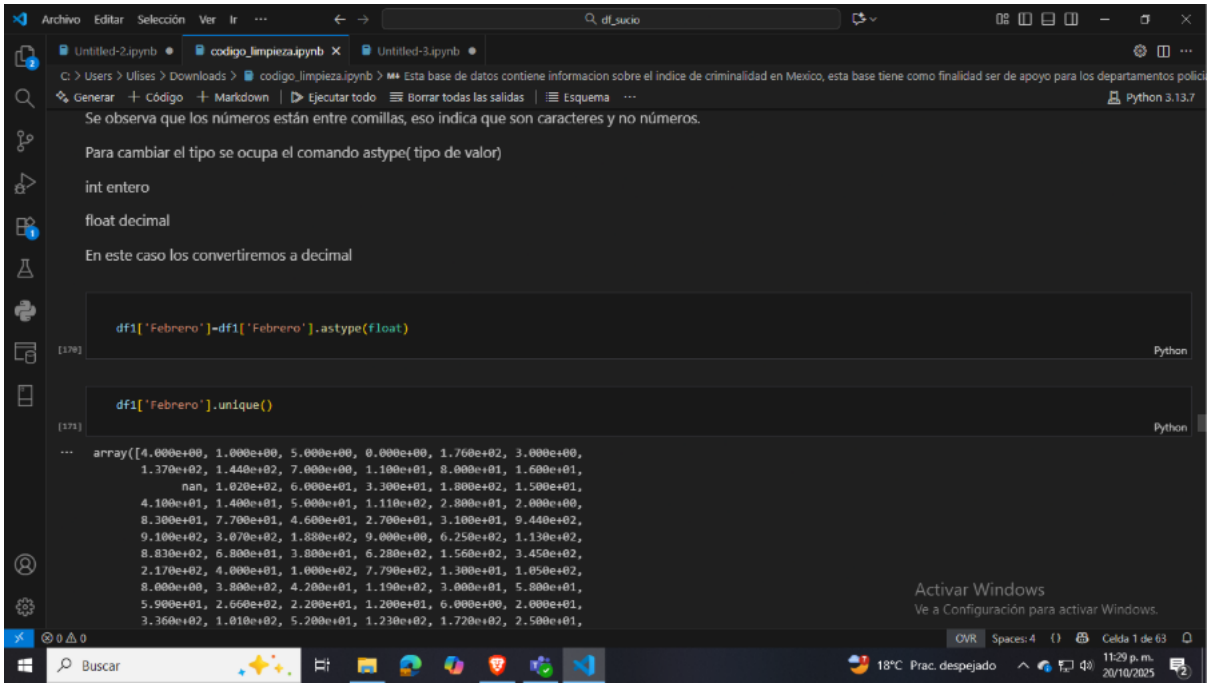
	Año	Clave_Ent	Entidad	Bien jurídico afectado	Tipo de delito	Subtipo de delito	Modalidad	Enero	Febrero	Marzo	...	Mayo	Junio	Julio	Agosto	Septiembre	Octu
0	2015.0	1.0	NaN	La libertad y la seguridad sexual	NaN	Abuso sexual	Abuso sexual	6.0	4	0.0	...	NaN	1.0	NaN	1.0	0.0	
2	NaN	1.0	Aguascalientes	La libertad y la seguridad sexual	Delito sexual	Hostigamiento sexual	Hostigamiento sexual	1.0	1	0.0	...	3.0	0.0	1.0	0.0	0.0	
4	2015.0	1.0	Aguascalientes	La libertad y la seguridad sexual	Violación	Violación equiparada	Violación equiparada	0.0	5	3.0	...	3.0	4.0	4.0	4.0	3.0	N
5	2015.0	1.0	Aguascalientes	La libertad y la seguridad sexual	Incesto	Incesto	Incesto	0.0	0	0.0	...	0.0	NaN	0.0	0.0	1.0	
8	2015.0	1.0	Aguascalientes	El patrimonio	Robo sin violencia	Robo a casa habitación	Sin violencia	215.0	176	202.0	...	215.0	228.0	233.0	229.0	202.0	N
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	

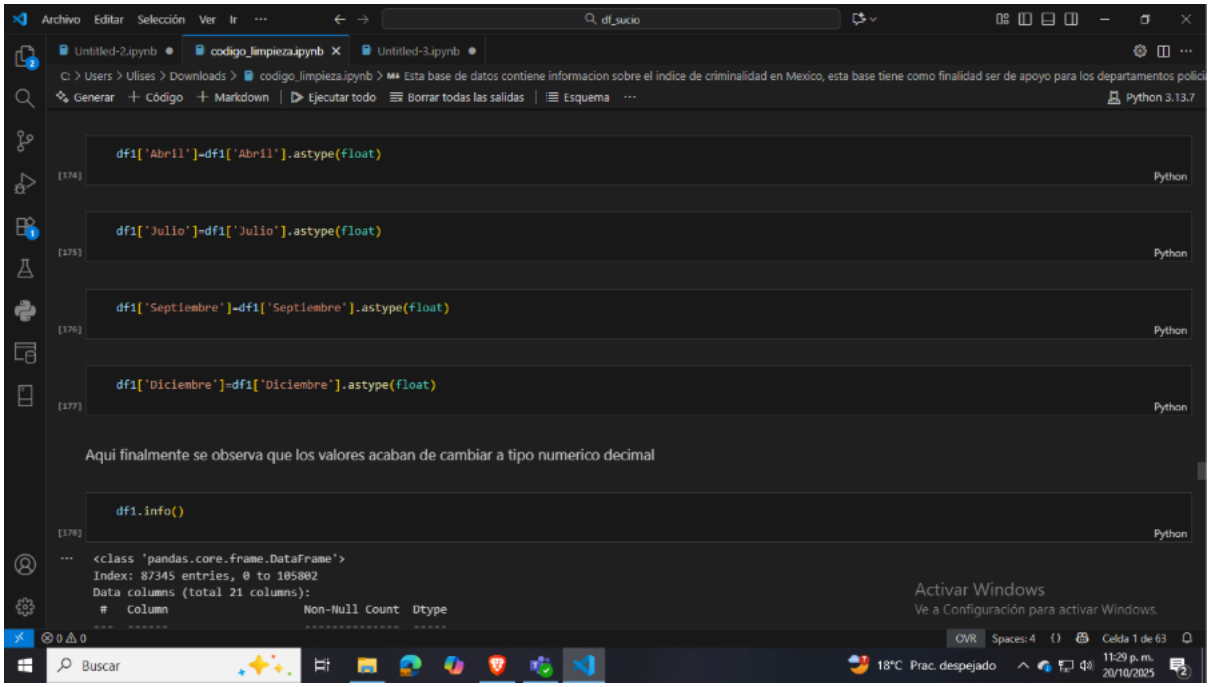


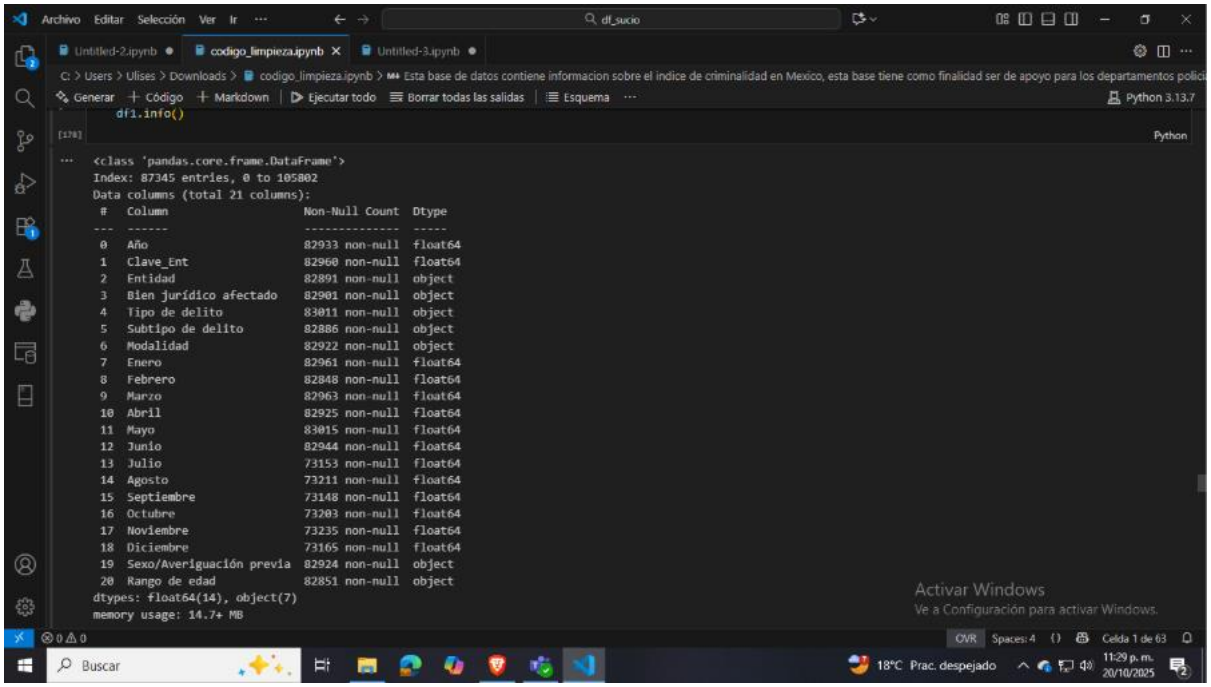
Finalmente, encontré valores clasificados de manera errónea, lo cual desato el inmediato cambio en todas las columnas donde se presenten




Como se puede observar las columnas:Febrero, Abril, Julio, Septiembre y Diciembre son de valor numerico, sin embargo en la base esta establecido como objeto, lo cual me indica de tengo que realizar un cambio de variable
En seguida hacemos uso del comando "unique" para poder visualisar el contenido que tiene la columna "Febrero"
df1['Febrero'].unique()
array(['4', '1', '5', '0', '176', '3', '137', '144', '7', '11', '80',
'16', nan, '102', '60', '33', '180', '15', '41', '14', '50', '111',
'28', '2', '83', '77', '46', '27', '31', '944', '910', '307',
'188', '9', '625', '113', '883', '68', '38', '628', '156', '345',
'217', '40', '100', '770', '13', '105', '8', '380', '42', '119',
'30', '58', '59', '266', '22', '12', '6', '20', '336', '101', '52',
'123', '172', '25', '74', '141', '185', '579', '35', '19', '34',
'26', '48', '87', '140', '62', '56', '71', '47', '37', '73', '76',
'103', '65', '247', '84', '43', '54', '315', '61', '55', '163',
'362', '229', '498', '820', '219', '18', '326', '234', '704', '89',
'763', '248', '66', '107', '265', '732', '143', '2021', '859',
'175', '646', '191', '112', '587', '678', '211', '32', '10', '63',
'23', '82', '232', '256', '44', '45', '377', '250', '638', '99',
'318', '257', '21', '70', '57', '403', '192', '613', '253', '158',
'69', '72', '301', '324', '166', '79', '521', '132', '431', '29',
'122', '78', '577', '106', '97', '86', '118', '1698', '209',
'1106', '197', '194', '416', '1078', '568', '2310', '458', '976',
'298', '281', '288', '116', '24', '213', '3637', '149', '408',
'173', '142', '186', '136', '184', '36', '17', '85', '370', '153',
'92', '341', '261', '562', '224', '243', '204', '64', '661',
Activar Windows
Ve a Configuración para activar Windows.

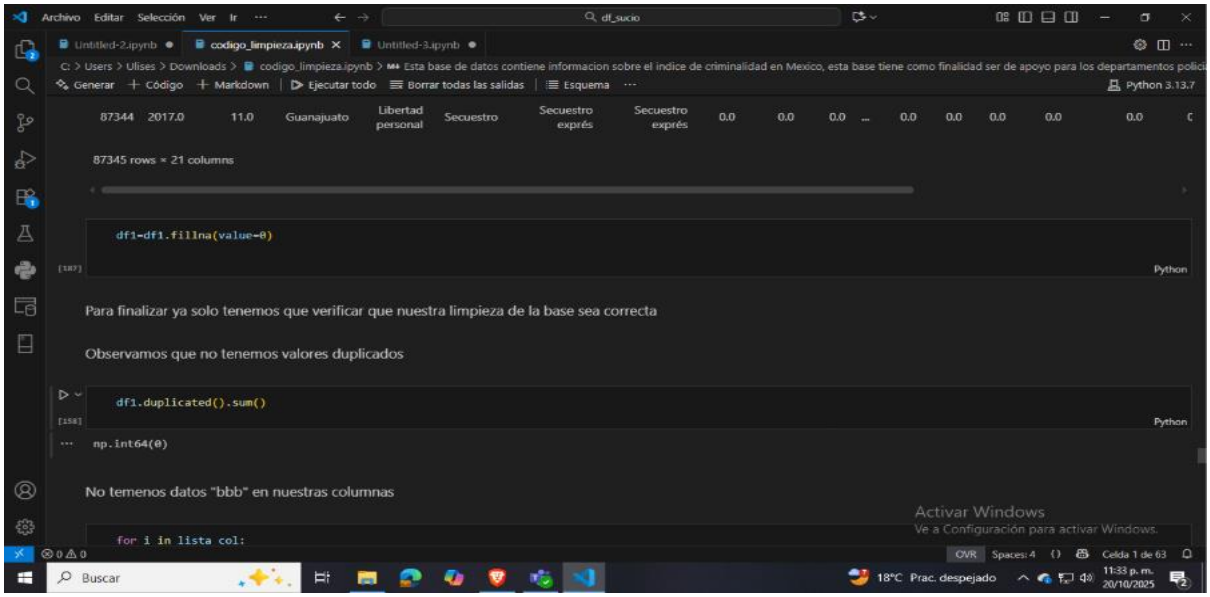

Se observa que los números están entre comillas, eso indica que son caracteres y no números.
Para cambiar el tipo se ocupa el comando astype(tipo de valor)
int entero
float decimal
En este caso los convertiremos a decimal
df1['Febrero']=df1['Febrero'].astype(float)
df1['Febrero'].unique()
array([4.000e+00, 1.000e+00, 5.000e+00, 0.000e+00, 1.760e+02, 3.000e+00,
1.370e+02, 1.440e+02, 7.000e+00, 1.100e+01, 8.000e+01, 1.600e+01,
nan, 1.020e+02, 6.000e+01, 3.300e+01, 1.800e+02, 1.500e+01,
4.100e+01, 1.400e+01, 5.000e+01, 1.110e+02, 2.800e+01, 2.000e+00,
8.300e+01, 7.700e+01, 4.600e+01, 2.700e+01, 3.100e+01, 9.440e+02,
9.100e+02, 3.070e+02, 1.880e+02, 9.000e+00, 6.250e+02, 1.130e+02,
8.830e+02, 6.800e+01, 3.800e+01, 6.280e+02, 1.560e+02, 3.450e+02,
2.170e+02, 4.000e+01, 1.000e+02, 7.790e+02, 1.300e+01, 1.050e+02,
8.000e+00, 3.800e+02, 4.200e+01, 1.190e+02, 3.000e+01, 5.800e+01,
5.900e+01, 2.660e+02, 2.200e+01, 1.200e+01, 6.000e+00, 2.000e+01,
3.360e+02, 1.010e+02, 5.200e+01, 1.230e+02, 1.720e+02, 2.500e+01,
Activar Windows
Ve a Configuración para activar Windows.

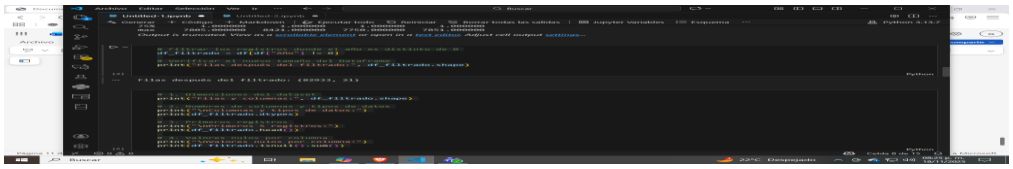


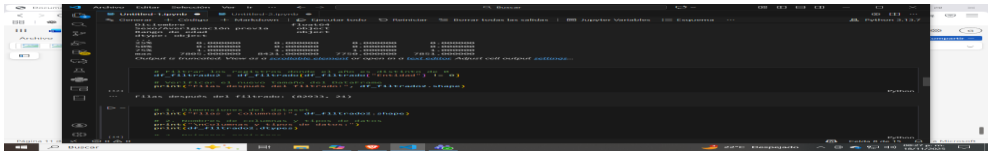


Como paso final encontré valores NaN, los cuales no aportan nada en mi base de datos y tampoco serán de ayuda en el futuro, ya que estos pueden interferir en mi análisis a futuro, los cuales pasamos a borrar.

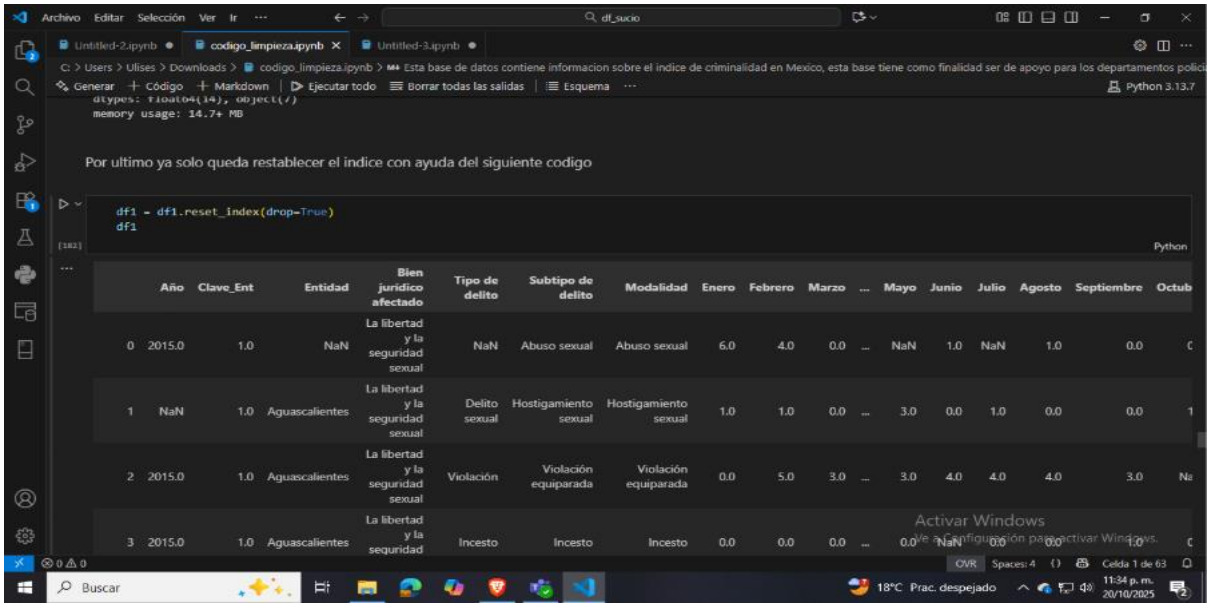
Primero cambie el nombre a de los valores a 0 para después borrarlos







Y como paso final, ya solo reinicié el índice para dar mejor presentación a la base de datos



	Año	Clave_Ent	Entidad	Bien jurídico afectado	Tipo de delito	Subtipo de delito	Modalidad	Enero	Febrero	Marzo	...	Mayo	Junio	Julio	Agosto	Septiembre	Octub
0	2015.0	1.0	NaN	La libertad y la seguridad sexual	NaN	Abuso sexual	Abuso sexual	6.0	4.0	0.0	...	NaN	1.0	NaN	1.0	0.0	0
1	NaN	1.0	Aguascalientes	La libertad y la seguridad sexual	Delito sexual	Hostigamiento sexual	Hostigamiento sexual	1.0	1.0	0.0	...	3.0	0.0	1.0	0.0	0.0	1
2	2015.0	1.0	Aguascalientes	La libertad y la seguridad sexual	Violación	Violación equiparada	Violación equiparada	0.0	5.0	3.0	...	3.0	4.0	4.0	4.0	3.0	Na
3	2015.0	1.0	Aguascalientes	La libertad y la seguridad	Incesto	Incesto	Incesto	0.0	0.0	0.0	...	0.0	NaN	0.0	0.0	1.0	0

Análisis Exploratorio de Datos (EDA)

1. **Descripción General de los Datos**

- **Visión General:**
  El Dataset contiene 84,233 filas y 21 columnas

- **Tipos de Variables:**
  La variable “Año” es numérica, “Clave_Ent” es numérica, “Entidad” es categórica, “Bien jurídico afectado” es categórica, “Tipo de delito” es categórica, “Subtipo de delito” es categórica “Modalidad” es categórica, “Enero” es numérica, “Febrero” es numérica, “Marzo” es numérica, “Abril” es numérica, “Mayo” es numérica, “Junio” es numérica, “Julio” es numérica, “Agosto” es numérica, “Septiembre” es numérica, “Octubre” es numérica, “Noviembre” es numérica, “Diciembre” es numérica, “Sexo/Averiguación previa” es categórica, “Rango de edad” es categórica.

- **Resumen Estadístico:**

  - **Para Variables Numéricas:**
    - **Año:**
      -- Media: 2019.032765
      -- Mediana**:** 2019.000000
      -- Desviación estándar: 2.588802
      -- Valor mínimo: 2015.000000
      -- Valor máximo: 2023.000000

    - **Clave_Ent:**
      -- Media: 15.625752
      -- Mediana**:** 16.000000
      -- Desviación estándar: 9.693931
      -- Valor mínimo: 0.000000
      -- Valor máximo: 32.000000
    - **Enero:**
      -- Media: 16.126253
      -- Mediana**:** 0.000000
      -- Desviación estándar: 103.883369
      -- Valor mínimo: 0.000000
      -- Valor máximo: 7565.000000
    - **Febrero:**

-- Media: 16.253405
-- Mediana**:** 0.000000
-- Desviación estándar: 107.075680
-- Valor mínimo: 0.000000
-- Valor máximo: 7888.000000

- **Marzo:**
  -- Media: 18.360526
  -- Mediana**:** 0.000000
  -- Desviación estándar: 119.630679
  -- Valor mínimo: 0.000000
  -- Valor máximo: 8418.000000
- **Abril:**
  -- Media: 16.891029
  -- Mediana**:** 0.000000
  -- Desviación estándar: 108.714001
  -- Valor mínimo: 0.000000
  -- Valor máximo: 6500.000000
- **Mayo:**
  -- Media: 18.063666
  -- Mediana**:** 0.000000
  -- Desviación estándar: 118.459830
  -- Valor mínimo: 0.000000
  -- Valor máximo: 7787.000000
- **Junio:**
  -- Media: 17.900447
  -- Mediana**:** 0.000000
  -- Desviación estándar: 113.785211
  -- Valor mínimo: 0.000000
  -- Valor máximo: 6615.000000
- **Julio:**
  -- Media: 15.237549
  -- Mediana**:** 0.000000
  -- Desviación estándar: 106.030201
  -- Valor mínimo: 0.000000
  -- Valor máximo: 7619.000000
- **Agosto:**
  -- Media: 15.730611
  -- Mediana**:** 0.000000
  -- Desviación estándar: 110.206817
  -- Valor mínimo: 0.000000
  -- Valor máximo: 7955.000000
- **Septiembre:**
  -- Media: 15.301614

-- Mediana**:** 0.000000
-- Desviación estándar: 104.300628
-- Valor mínimo: 0.000000
-- Valor máximo: 7805.000000

- **Octubre:**
  -- Media: 16.059182
  -- Mediana**:** 0.000000
  -- Desviación estándar: 110.885002
  -- Valor mínimo: 0.000000
  -- Valor máximo: 8421.000000
- **Noviembre:**
  -- Media: 15.121781
  -- Mediana**:** 0.000000
  -- Desviación estándar: 105.304424
  -- Valor mínimo: 0.000000
  -- Valor máximo: 7750.000000
- **Diciembre:**
  -- Media: 14.408478
  -- Mediana**:** 0.000000
  -- Desviación estándar: 102.138865
  -- Valor mínimo: 0.000000
  -- Valor máximo: 7851.000000

- **Para variables categóricas:**

Entidad:

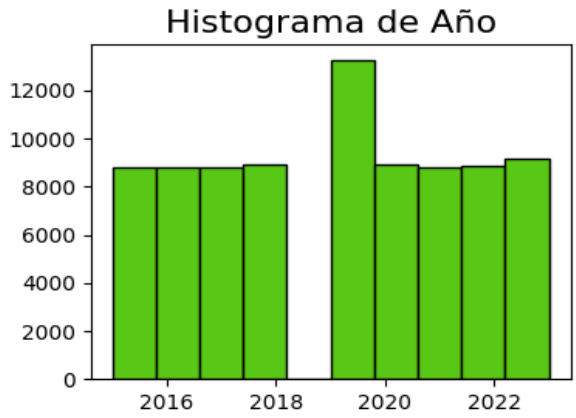
-Sonora: 2355
-San Luis Potosí: 2345
-Ciudad de México: 2344
-Chihuahua: 2335
-México: 2333
- Michoacán de Ocampo: 2331
-Quintana Roo: 2326
-Coahuila de Zaragoza: 2321
-Chiapas: 2316
-Morelos: 2316
-Durango: 2311
-Zacatecas: 2310
-Oaxaca: 2309
-Veracruz de Ignacio de la Llave: 2303
-Tamaulipas: 2297
-Nuevo León: 2297
-Querétaro: 2296

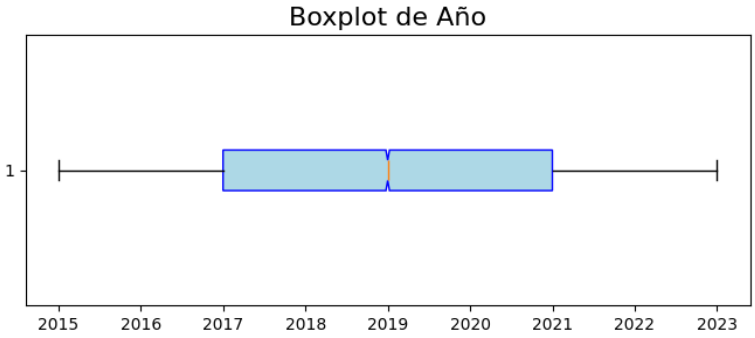
-Puebla: 2296
-Colima: 2294
-Guanajuato: 2289
-Jalisco: 2287
-Baja California: 2287
-Guerrero: 2286
-Hidalgo :2286
-Sinaloa :2285
-Nayarit: 2282
-Tlaxcala: 2277
-Tabasco: 2276
-Baja California Sur: 2269
-Aguascalientes: 2264
-Yucatán: 2239
-Campeche: 2229

Visualización y Distribución de Variables Individuales

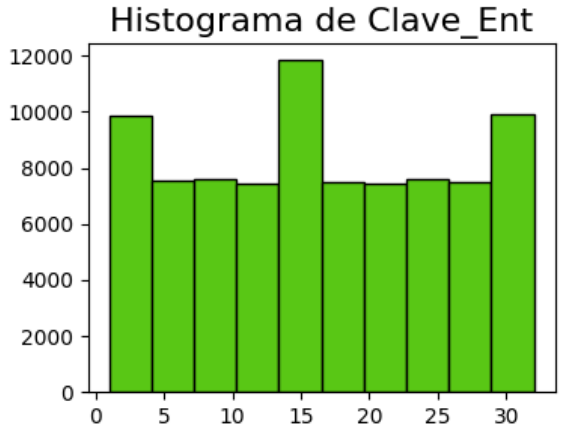
Variables numéricas:

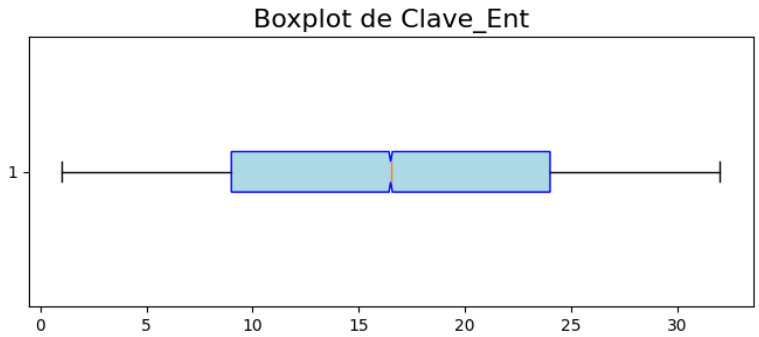
- Histograma de Año: muestra una distribución ligeramente segada a la derecha y teniendo un pico durante el 2020, lo que indica que, sucedido un aumento de crímenes después del 2018.
    - Boxplot: muestra una concentración en la media, también muestra valores atípicos en ambos lados y una forma simétricas, indicando que hay mayor concentración en los años de 2017 a 2021.



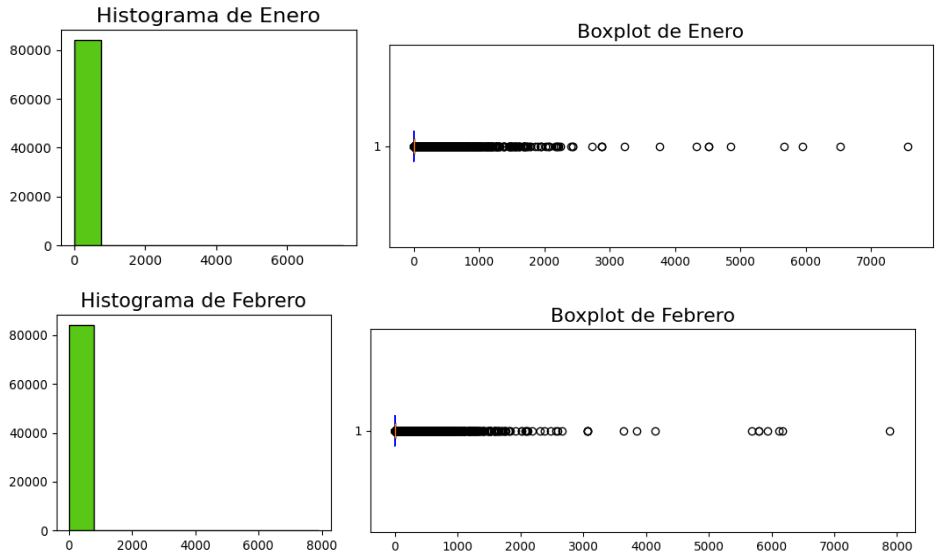


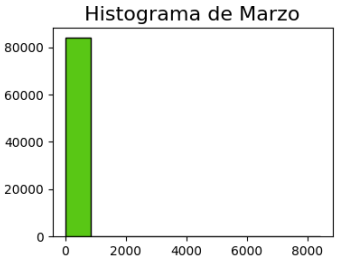
- Histograma de Clave_Ent: muestra una distribución bimodal, el primer pico se encuentra en el rango de 0 a 5, el segundo muestra pico en el rango aproximado de 14 a 16 y, por último, el tercero muestra un pico en el rango de 29 a 32, esto dice que los estados, cuya clave está en un pico, tuvieron un aumento en crímenes.
    - Boxplot: se puede observar una concentración en la media, abarcando desde 8 hasta 24 el rango de la x, de igual forma muestra valores atípicos de forma simétrica en ambos lados.

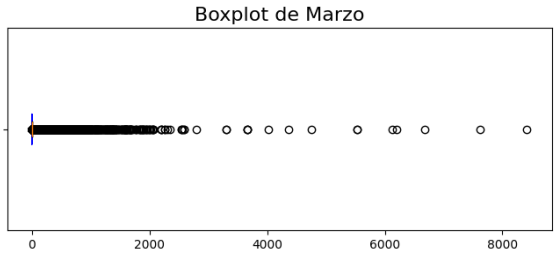


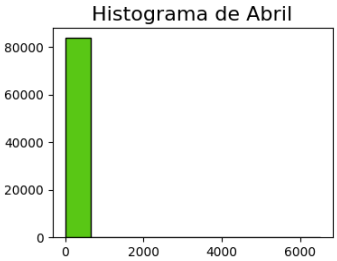


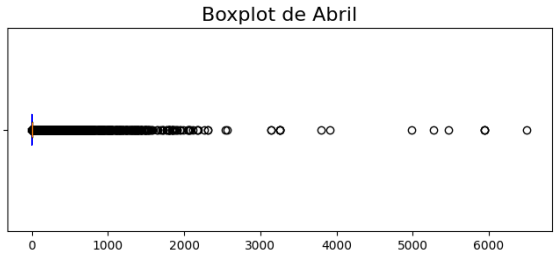
- Histograma de Enero a Diciembre: Se observa una distribución altamente segada al lado izquierdo mostrando que nuestros datos están muy cerca del 0 en el eje X, mientras que, en eje Y, sobrepasan los 8,000 datos, dando a entender que hay una gran concentración de datos en un solo periodo.
    - Boxplots de enero a diciembre: Los boxplots de enero a diciembre muestran una distribución sesgada hacia la derecha, con una alta concentración de valores en el rango bajo y múltiples outliers que se extienden hacia valores elevados. Esto indica una fuerte asimetría y alta variabilidad en los datos de estos meses.

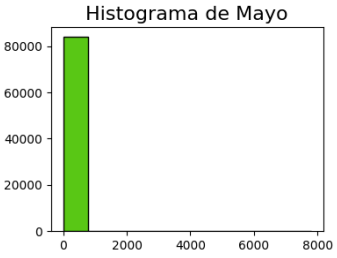


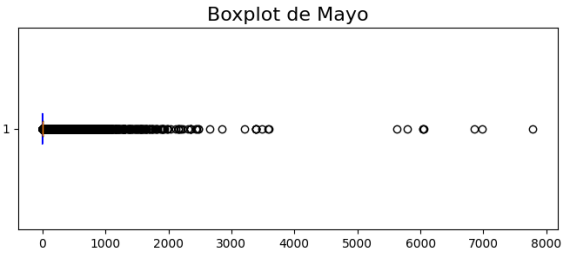

Histograma de Marzo
0
20000
40000
60000
80000
0
2000
4000
6000
8000

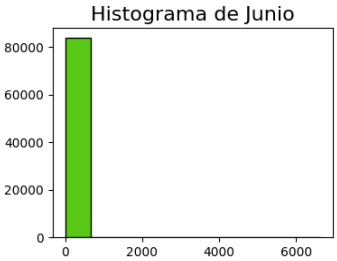

Boxplot de Marzo
1
0
2000
4000
6000
8000

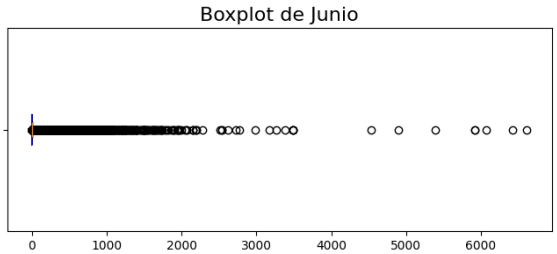

Histograma de Abril
0
20000
40000
60000
80000
0
2000
4000
6000

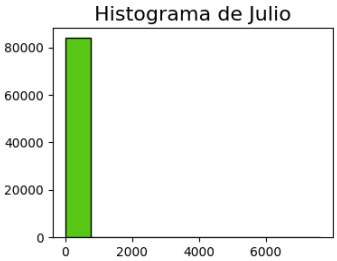

Boxplot de Abril
1
0
1000
2000
3000
4000
5000
6000

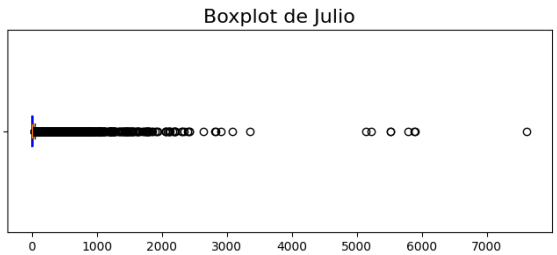

Histograma de Mayo
0
20000
40000
60000
80000
0
2000
4000
6000
8000

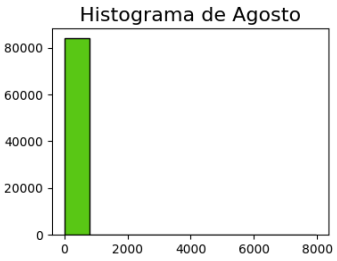

Boxplot de Mayo
1
0
1000
2000
3000
4000
5000
6000
7000
8000

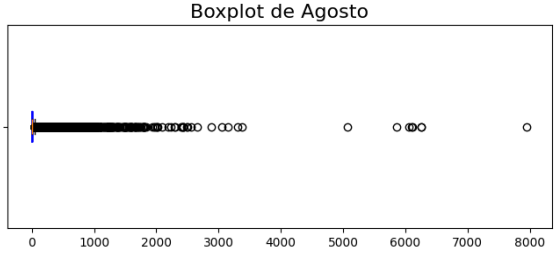

Histograma de Junio
0
20000
40000
60000
80000
0
2000
4000
6000

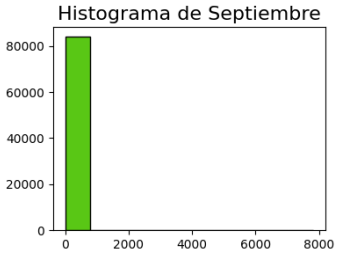

Boxplot de Junio
1
0
1000
2000
3000
4000
5000
6000

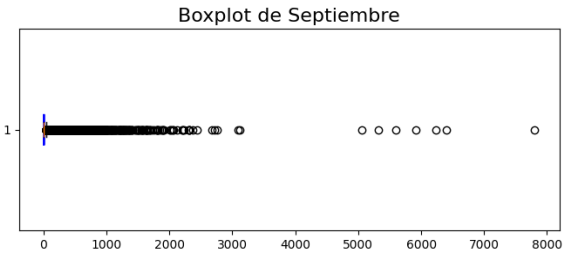

Histograma de Julio
0
20000
40000
60000
80000
0
2000
4000
6000

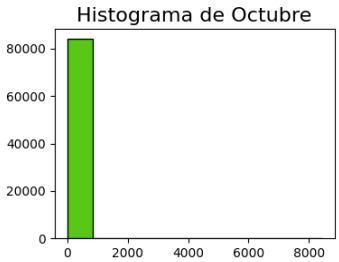

Boxplot de Julio
1
0
1000
2000
3000
4000
5000
6000
7000

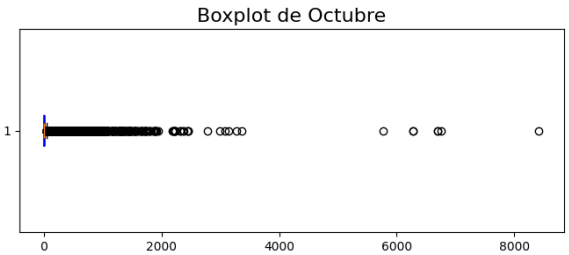

Histograma de Agosto
0
20000
40000
60000
80000
0
2000
4000
6000
8000

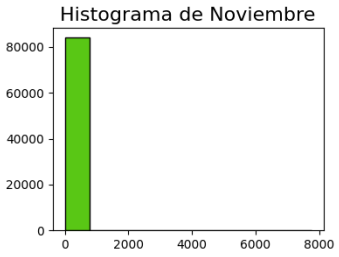

Boxplot de Agosto
1
0
1000
2000
3000
4000
5000
6000
7000
8000

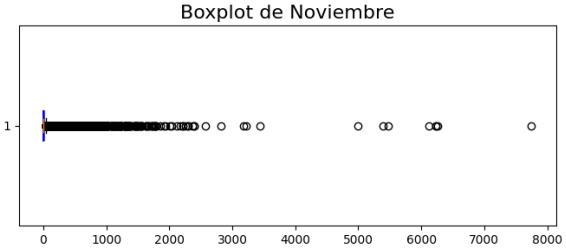

Histograma de Septiembre
80000
60000
40000
20000
0
0
2000
4000
6000
8000

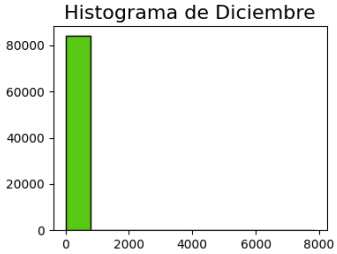

Boxplot de Septiembre
1
0
1000
2000
3000
4000
5000
6000
7000
8000

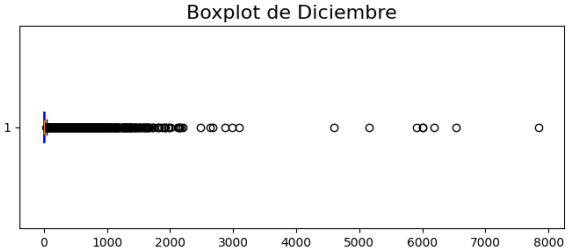

Histograma de Octubre
80000
60000
40000
20000
0
0
2000
4000
6000
8000


Boxplot de Octubre
1
0
2000
4000
6000
8000


Histograma de Noviembre
80000
60000
40000
20000
0
0
2000
4000
6000
8000


Boxplot de Noviembre
1
0
1000
2000
3000
4000
5000
6000
7000
8000

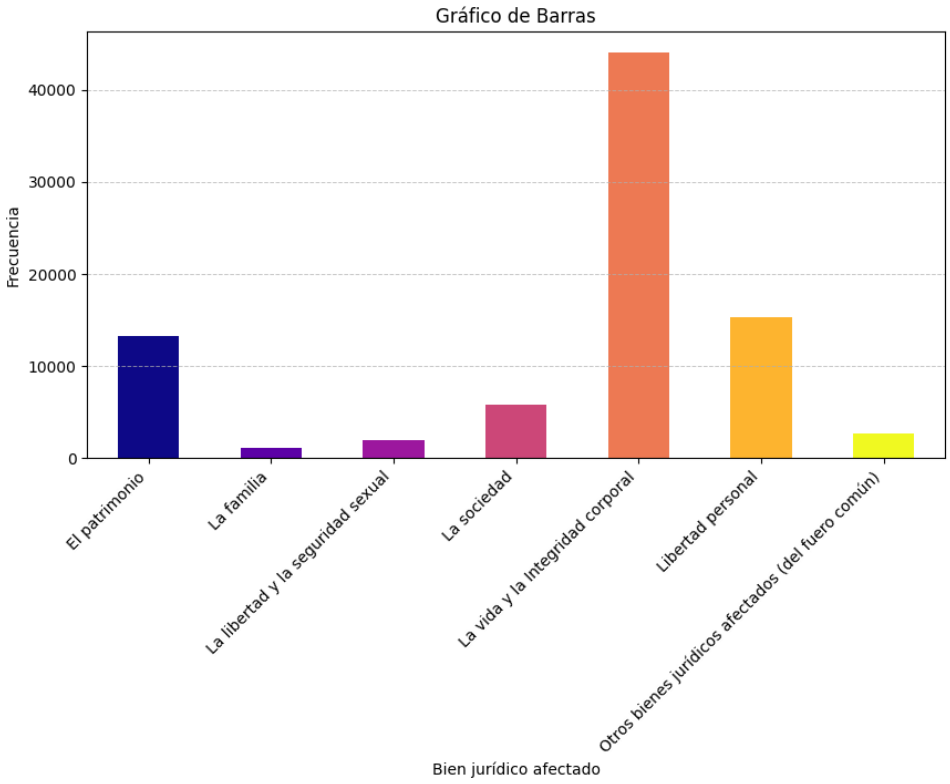

Histograma de Diciembre
80000
60000
40000
20000
0
0
2000
4000
6000
8000


Boxplot de Diciembre
1
0
1000
2000
3000
4000
5000
6000
7000
8000

Variables categóricas:

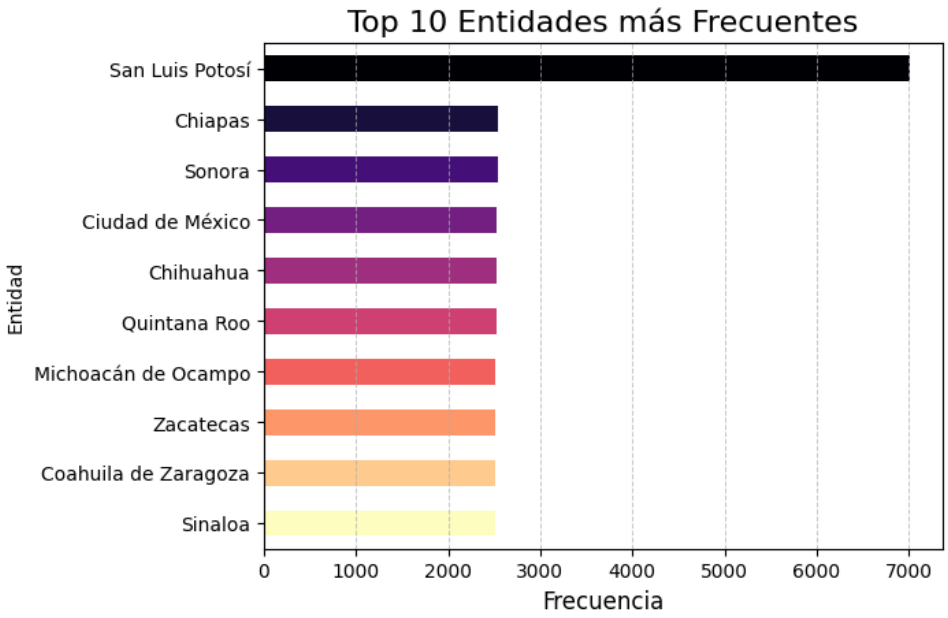
Variable: Bien Jurídico Afectado

Al analizar el bien jurídico, se observa que la categoría predominante es **'La vida y la Integridad corporal'**, abarcando el **52.35%** de los registros. Esto indica que más de la mitad de la incidencia delictiva registrada atenta directamente contra las personas, superando a los delitos patrimoniales en esta clasificación específica.



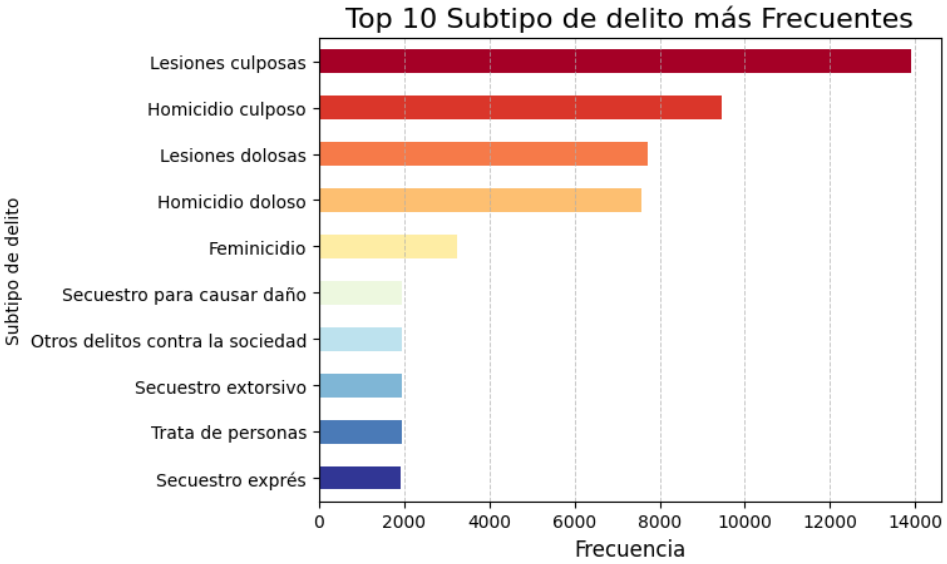
Variable: Entidad

La distribución geográfica muestra que **San Luis Potosí** es la entidad con mayor frecuencia de registros en esta base de datos procesada, representando el **8.32%** del total. Es importante notar que la distribución entre los 32 estados es más uniforme que en otras variables, lo que sugiere una cobertura nacional amplia, aunque con focos de concentración en ciertos estados.



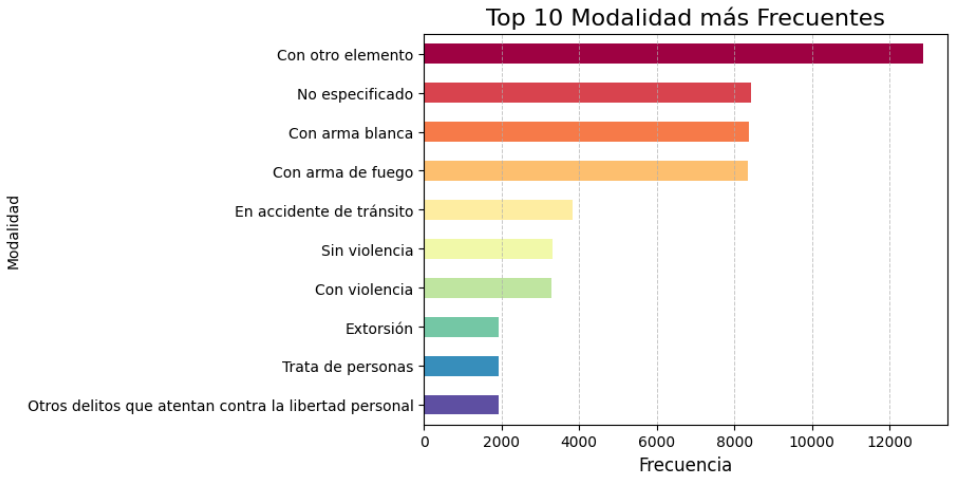
Variable: Subtipo de Delito

Dentro de los subtipos, **'Lesiones culposas'** es la categoría más frecuente (**16.51%**), seguida de cerca por otros tipos de lesiones y homicidios. La gran variedad de subtipos (61 categorías únicas) muestra la alta especificidad de los datos, aunque el volumen se concentra en los daños físicos accidentales o intencionales.



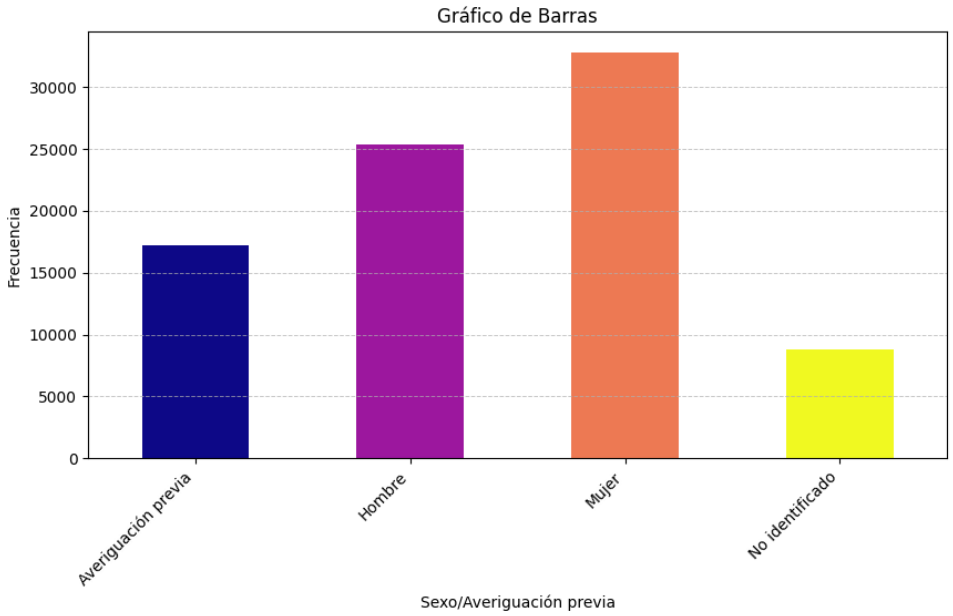
Variable: Modalidad

El análisis de modalidad revela que el **15.25%** de los delitos se cometen **'Con otro elemento'** (categoría genérica), seguido por modalidades específicas de violencia. Esta variable presenta una alta fragmentación, lo que indica que el *modus operandi* delictivo es muy variado.



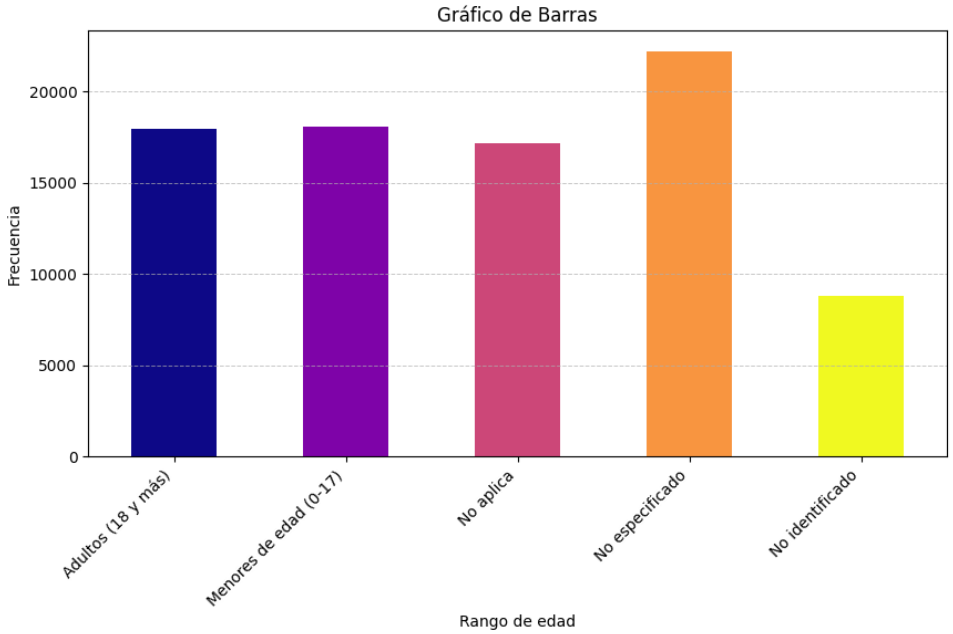
Variable: Sexo / Averiguación Previa

El **38.98%** de los registros corresponden a víctimas del sexo **'Mujer'**, siendo la categoría más alta identificada. Sin embargo, existe una proporción relevante de datos etiquetados como 'Averiguación previa' o 'No especificado', lo que señala áreas de oportunidad en la calidad del registro demográfico de las víctimas.



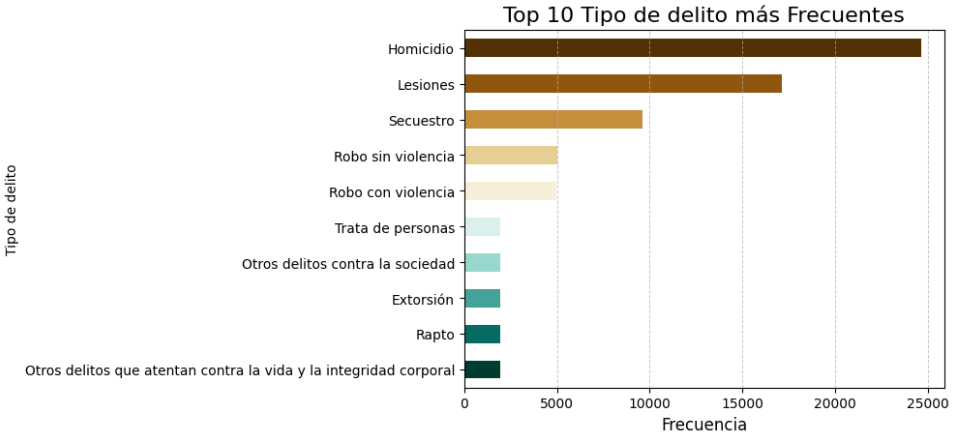
Variable: Rango de Edad

"Esta variable presenta un desafío de calidad de datos, ya que la categoría **'No especificado'** es la más frecuente con un **26.39%**. De los rangos conocidos, los adultos jóvenes suelen tener mayor representación, pero el alto porcentaje de valores no especificados limita el análisis generacional preciso."



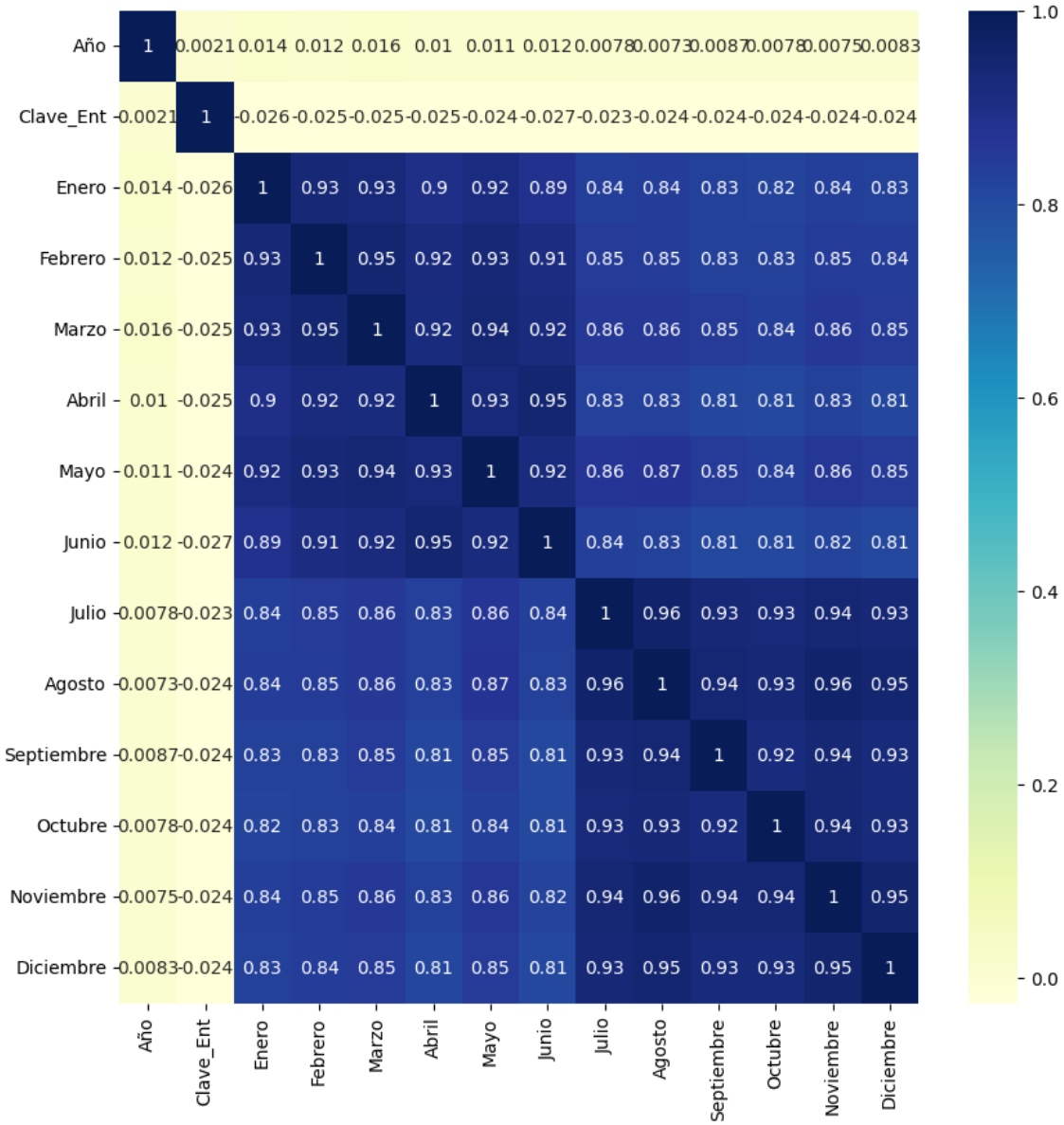
Variable: Tipo de Delito

"Esta es una de las variables más críticas del estudio. El análisis muestra que el **'Homicidio'** es la categoría con mayor número de registros procesados, representando el **29.28%** del total, seguido de cerca por delitos como **'Robo'** y **'Lesiones'**.



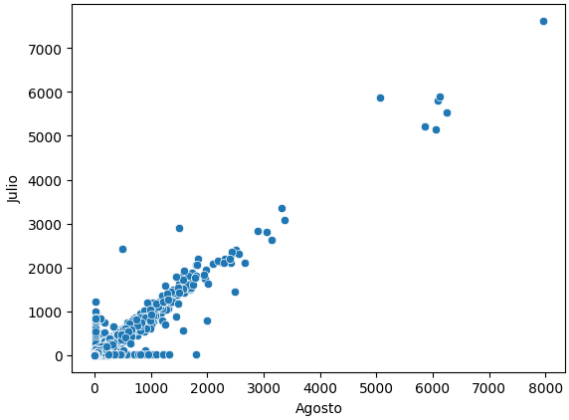
Correlación entre Variables

Matriz de correlación:

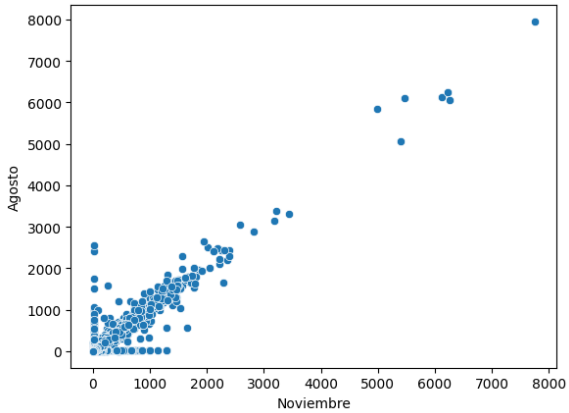


Parejas de Variables:

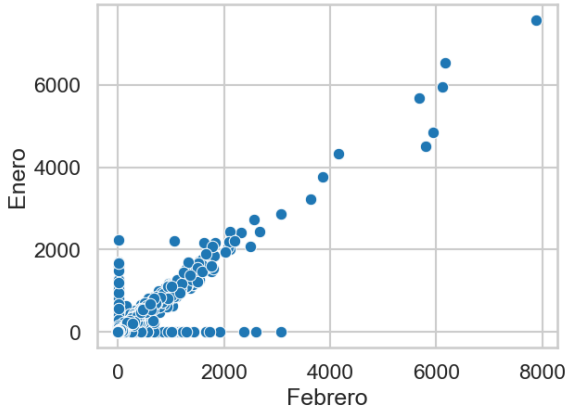
Agosto vs. Julio: La correlación que presenta es de:0.96 la más alta se da entre estos dos meses consecutivos, sugiriendo que el verano presenta un patrón delictivo casi idéntico.



Noviembre vs. Agosto: La correlación que tienen es de:0.95 lo que implica que existe una fuerte conexión entre el cierre del año y el verano, lo que implica que los focos de delincuencia no se desplazan significativamente a lo largo del segundo semestre.



Febrero vs. Enero: La correlación que presentan es la siguiente:0.93 lo que indica que el inicio del año también muestra consistencia alta, aunque ligeramente menor que el cierre del año.

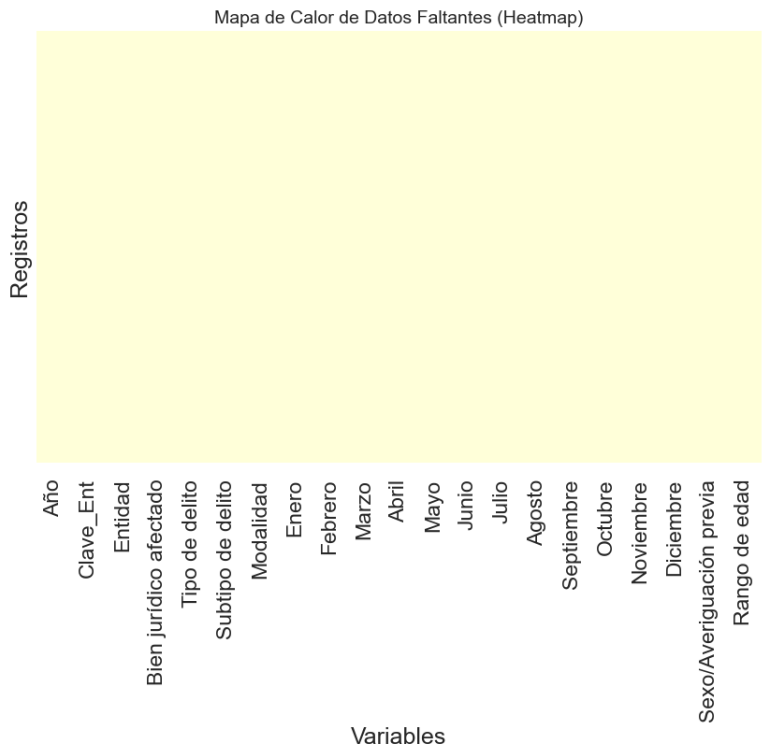


Análisis de Valores Atípicos (Outliers):

Tratamiento y Justificación: Se decidió **CONSERVAR** la totalidad de los valores atípicos.

Justificación: En el contexto de este proyecto de **Seguridad Pública**, los valores extremos no corresponden a errores de medición, sino que representan precisamente los **focos rojos** de criminalidad que buscamos analizar (zonas de alta densidad como CDMX, Estado de México y Jalisco). Eliminar estos registros sería contraproducente, ya que "cegaría" al modelo ante las problemáticas más graves del país. Mantenerlos permite que el análisis refleje la realidad de la distribución delictiva, donde unos pocos estados concentran la mayor parte de la incidencia.

Análisis de Valores Faltantes



Identificación: Se realizó un análisis de integridad de datos utilizando la función isnull().

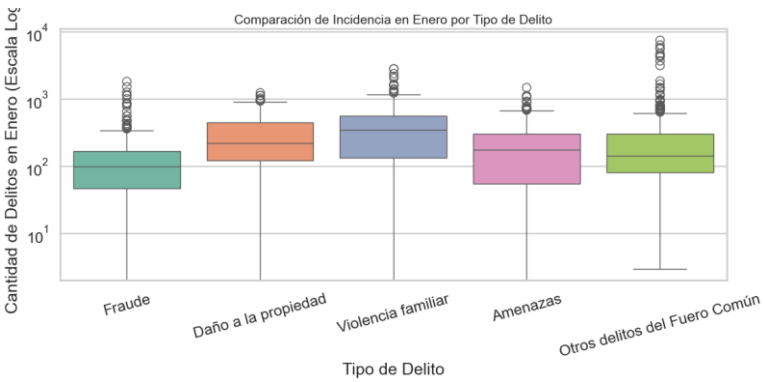
- **Porcentaje de Faltantes:** El análisis final arroja un **0% de valores faltantes** en el dataset procesado
- **Visualización:** El mapa de calor presentado anteriormente se muestra completamente uniforme, confirmando visualmente la ausencia de datos nulos o vacíos en la estructura del dataframe.

Estrategia de Imputación: Para llegar a este estado de calidad, se aplicó una estrategia de **Imputación por Lógica de Negocio** y **Moda**:

1. **Variables Numéricas (Meses):** Los valores nulos encontrados en las columnas de los meses (Enero-Diciembre) fueron imputados con la constante **0**.
    - Justificación*:* En el contexto de incidencia delictiva, la ausencia de un dato numérico en un mes específico se interpreta como la "no ocurrencia" de delitos en ese periodo, no como un error de captura desconocido.
2. **Variables Categóricas:** Para las variables cualitativas, se verificó la integridad y, en caso de inconsistencias, se optó por la eliminación de filas corruptas para no introducir sesgos artificiales mediante la moda.

Relación entre Variables Categóricas y Numéricas

Análisis Comparativo (Boxplot):



Se analizó la relación entre la variable categórica **Tipo de delito** y la variable numérica **Enero** (Solo como ejemplo)

Interpretación:

- **Tendencia Central:** El delito de **"Violencia familiar"** presenta la mediana más alta en el mes de enero (348 delitos), confirmándose como la categoría con mayor volumen base constante. Le siguen **"Daño a la propiedad"** (222) y **"Amenazas"** (177).
- **Dispersión:** Se observa que **"Otros delitos del Fuero Común"** tiene una dispersión significativa lo que indica que en enero este delito puede tener cifras muy bajas en algunos estados y muy altas en otros.

Observaciones y Hallazgos Importantes

Identificar variable objetivo y variables influyentes:

- **Variable Objetivo (Target):** Target_Total_Anual. Se construyó sumando la incidencia de los 12 meses para obtener el volumen delictivo total por registro.
- **Variables Influyentes:**
    - **Tipo de Delito:** Es la variable categórica más determinante. Delitos como **"Violencia familiar"** y **"Robo"** presentan medianas significativamente más altas que el resto.
    - **Meses (Predictores):** Variables como **'Agosto'** y **'Julio'** mostraron una correlación superior a 0.95 con el total anual, indicando que son fuertes predictores del comportamiento anual.

Hallazgos clave:

- **Patrones:** Se observa una distribución de **"Cola Larga"**. La mayoría de los registros tienen pocos delitos, pero un pequeño grupo (focos rojos) concentra una cantidad masiva de incidencia.
- **Outliers Relevantes:** Se detectaron **8,288 registros atípicos** (superiores a 256 delitos). El valor máximo es de **80,883 delitos**, lo cual refleja la disparidad extrema de seguridad entre distintos municipios/estados.
- **Variables Desbalanceadas:** Existe un desbalance notable en el tipo de delito; categorías como **"Homicidio"** representan cerca del 29% de los registros, mientras que otros delitos tienen presencia mínima. Además, el **22% de los registros son ceros** (no hubo delitos).
- **Correlaciones Fuertes:** Se identificó alta multicolinealidad entre los meses (ej. Agosto vs Julio, correlación: 0.96), lo que confirma una fuerte estacionalidad.
- **Calidad de Datos:** El dataset procesado presenta un **0% de valores nulos**, confirmando una limpieza exitosa.

Implicaciones para el modelo:

- **Multicolinealidad:** Dado que los meses individuales están correlacionados por encima de 0.90, no se usarán todos como predictores independientes para evitar redundancia; se priorizará el uso de variables categóricas (`Tipo de delito`, `Entidad`) y la variable temporal `Año`.
- **Manejo de Outliers:** A pesar de su magnitud extrema, se decidió **conservar los outliers** ya que representan los puntos críticos de seguridad pública que el modelo debe ser capaz de detectar.
- **Estrategia de Modelado:** Debido a la naturaleza no lineal y la presencia de muchos ceros, se optará por un modelo basado en árboles llamado "**Random Forest**" en lugar de una regresión lineal simple, ya que maneja mejor estas irregularidades.

Modelo de Machine Learning

Descripción del Modelo:

- **Nombre del Modelo:** Random Forest
- **Tipo de Aprendizaje:** Aprendizaje Supervisado.
- **Justificación**: El modelo es "supervisado" porque lo entrenamos utilizando datos etiquetados; es decir, le enseñamos al algoritmo tanto las características de entrada (Entidad, Tipo de delito) como la respuesta correcta esperada (el Total de Delitos Anuales) para que aprenda la relación entre ellas.
- **Tipo de Problema:** Regresión.
- **Justificación***:* El problema se clasifica como de regresión porque la variable objetivo, Target_Total_Anual**,** es numérica y continua (representa una cantidad de delitos, por ejemplo: 150, 5000, 80000). El objetivo del modelo es predecir un número específico, no clasificar en una categoría (como "Alto/Bajo").

Justificación:

Se seleccionó el modelo **R**andom Forest Regressor, basándonos en los siguientes criterios:

- **Tipo de Variable Objetivo:** La variable objetivo “Target_Total_Anual” es numérica y continu**a** (representa la cantidad de delitos acumulados). Por lo tanto, el problema corresponde a una tarea de Regresión y no de Clasificación.
- **Complejidad de los Datos y No-Linealidad:** El análisis exploratorio “EDA” reveló que la incidencia delictiva no sigue un comportamiento lineal simple. A diferencia de una Regresión Lineal tradicional, el Random Forest tiene la capacidad de capturar relaciones no lineales complejas entre las variables categóricas y el volumen de incidencia.
- **Manejo de Outliers:** Dado que conservamos los valores atípicos (registros con hasta 80,000 delitos) por ser críticos para el análisis de seguridad, necesitamos un modelo robusto. Random Forest es menos sensible a los outliers que otros algoritmos, evitando que unos pocos casos extremos distorsionen por completo las predicciones generales.
- **Interpretabilidad para la Toma de Decisiones:** Aunque es un modelo complejo permite extraer la Importancia de las Variables. Esto es crucial para el objetivo del proyecto, ya que nos permite explicar a las autoridades no solo cuántos delitos ocurrirán, sino qué factores tienen mayor peso en esa predicción.

Implementación y Entrenamiento

- División de datos: Utilizamos la función train_test_split para separar nuestros datos en dos grupos:

- **Entrenamiento (80%):** Datos que el modelo usará para aprender.
- **Prueba (20%):** Datos que guardamos para hacerle un examen final al modelo.

-Entrenamiento del modelo: Seleccionamos el algoritmo Random Forest Regressor por su capacidad para manejar datos complejos. Lo configuramos con 100 árboles de decisión para asegurar un buen equilibrio entre precisión y velocidad. Durante este paso, el modelo analizó las relaciones entre el "Tipo de Delito", la "Entidad" y la cantidad de delitos ocurridos.

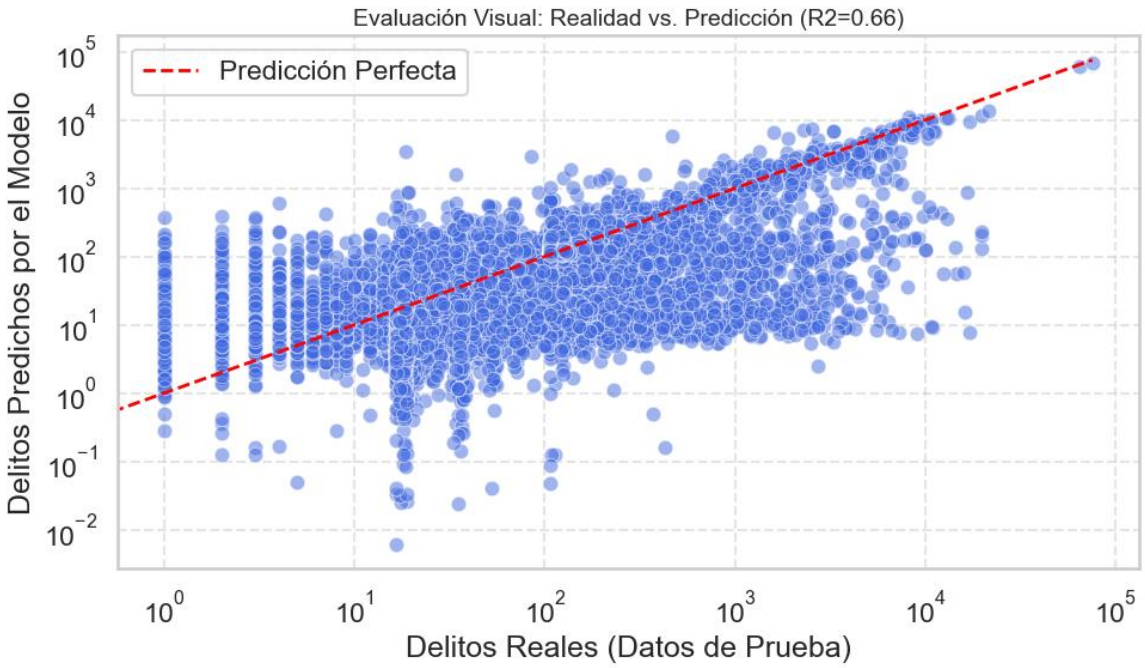
- Predicción: Una vez entrenado, le pedimos al modelo que predijera la cantidad de delitos para el grupo de prueba (los datos que nunca había visto). Finalmente, convertimos esas predicciones (que estaban en escala logarítmica) nuevamente a números reales para poder compararlas con la realidad.

Resultados y Evaluación

Para evaluar la eficacia del modelo Random Forest, se calcularon las siguientes métricas sobre el conjunto de prueba (el 20% de datos desconocidos):

1. **Coeficiente de Determinación (R^2): 0.65**
    a. **Interpretación:** El modelo logra explicar el 65% de la variabilidad de la incidencia delictiva. Esto significa que es capaz de identificar correctamente las tendencias generales de criminalidad basándose en el tipo de delito y la entidad, lo cual es un resultado sólido considerando la complejidad social del problema.
2. **Error Absoluto Medio (MAE): 131 delitos**
    a. **Interpretación:** En promedio, las predicciones del modelo se desvían por 131 delitos respecto al valor real.
    b. **Análisis***:* Este margen de error es aceptable para estados con alta incidencia (donde ocurren miles de delitos), pero indica que el modelo puede ser menos preciso en municipios pequeños con muy baja actividad delictiva.

Análisis Gráfico (Realidad vs. Predicción):

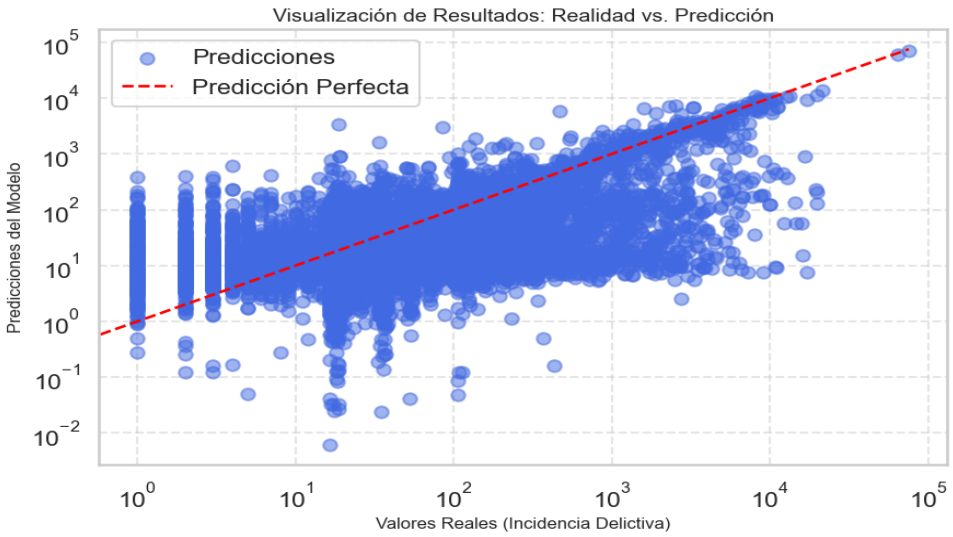


La gráfica de dispersión muestra que los puntos se alinean con la diagonal roja (predicción perfecta) en los rangos medios y altos. Esto confirma que el modelo es efectivo para detectar focos rojos de inseguridad, cumpliendo con el objetivo principal del proyecto de apoyar la toma de decisiones en seguridad pública.

Visualización de resultados

Visualización de Resultados (Regresión)

Gráfico de Dispersión (Realidad vs. Predicción):



Interpretación:

- **Alineación:** Se observa que los puntos tienden a alinearse sobre la diagonal roja (línea de predicción perfecta), especialmente en los rangos medios y altos de incidencia delictiva. Esto confirma que el modelo ha aprendido correctamente la tendencia general.
- **Dispersión:** Existe una dispersión visible en los valores más bajos (parte inferior izquierda), lo cual es consistente con el Error Absoluto Medio (MAE) reportado anteriormente. Esto indica que el modelo tiene mayor dificultad para predecir con exactitud en municipios con muy pocos delitos, pero es robusto para identificar las zonas de mayor criminalidad.

Conclusión del modelo

¿El modelo predice con buena precisión?

El modelo Random Forest Regressor alcanzó un coeficiente de determinación de 0.65, lo que significa que logra explicar el 65% del comportamiento delictivo. Si bien el Error Absoluto Medio de 131 delitos indica un margen de error aceptable para los estados con alta incidencia, el modelo presenta limitaciones para predecir con exactitud en municipios pequeños con muy baja actividad delictiva. En general, se considera una herramienta útil y robusta para identificar tendencias y focos rojos, cumpliendo su propósito de inteligencia para la seguridad pública.

¿Qué variables fueron más influyentes?

El análisis de importancia de características reveló que el Tipo de Delito es el factor más determinante para la predicción, seguido por la Entidad Federativa. Específicamente,

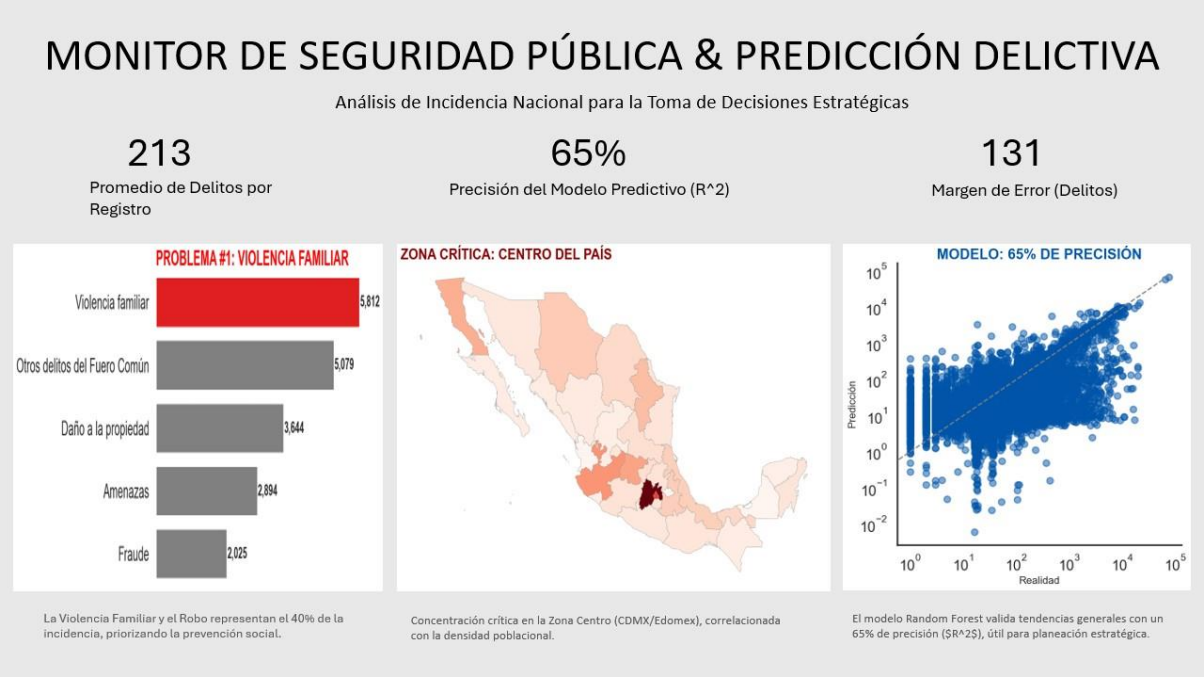
categorías como "Violencia Familiar" y "Robo", así como ubicaciones geográficas como "Estado de México", tienen el mayor peso en el modelo.

¿Qué mejoras podrían aplicarse?

Para futuras iteraciones y aumentar la precisión por encima del 80%, se recomienda:

1. **Segmentación:** Crear dos modelos separados; uno especializado en delitos de alto volumen y otro para incidencias bajas, para reducir el sesgo causado por la gran cantidad de ceros.
2. **Enriquecimiento de Datos:** Incorporar variables externas como densidad poblacional, presupuesto de seguridad o tasas de desempleo por estado.

Dashboard



Uso y Beneficios del Dashboard

Público Objetivo: Esta herramienta visual está dirigida principalmente a **Directivos de Seguridad Pública, Mandos Policiales y Tomadores de Decisiones Gubernamentales** encargados de la asignación de presupuesto y estrategias de prevención del delito.

Apoyo a la Toma de Decisiones: El dashboard transforma millones de registros de datos crudos en inteligencia accionable, permitiendo:

1. **Asignación Inteligente de Recursos:** Al visualizar claramente los "Focos Rojos" geográficos (Estado de México y CDMX), los mandos pueden decidir enviar mayor número de patrullas y presupuesto a la zona centro del país, en lugar de dispersar recursos ineficientemente en estados con baja incidencia.
2. **Estrategias de Prevención Focalizada:** Al identificar que la **"Violencia Familiar"** es el delito número uno, las autoridades pueden dejar de tratar la inseguridad como un problema genérico y comenzar a implementar programas sociales y unidades especializadas en conflictos domésticos.

Simplificación y Democratización de Datos: El tablero permite obtener insights inmediatos sin necesidad de conocimientos técnicos avanzados:

- **Lectura Rápida:** En menos de 5 segundos, el usuario comprende la magnitud del problema (213 delitos promedio) y la confiabilidad de la herramienta (65% de precisión).
- **Validación del Modelo:** El gráfico de dispersión traduce la complejidad matemática del algoritmo Random Forest en una visualización sencilla: si los puntos siguen la línea, la predicción es confiable. Esto fomenta una cultura de decisiones basadas en evidencia en lugar de la intuición política.

Conclusiones y Futuras Líneas de Trabajo

Resumen de Hallazgos y Cumplimiento de Objetivos

El objetivo principal de este proyecto fue analizar la incidencia delictiva en México para generar herramientas de inteligencia que apoyen la toma de decisiones en seguridad pública. Tras el procesamiento de datos y la implementación del modelo de Machine Learning, se concluye lo siguiente:

1. **Concentración del Delito (Ley de Pareto):** Se identificó que la inseguridad no es un fenómeno homogéneo. Un pequeño grupo de categorías (**Violencia Familiar** y **Robo**) y entidades federativas (**Estado de México** y **CDMX**) concentran la gran mayoría de la incidencia. Esto cumple con el objetivo de identificar "focos rojos" para la asignación prioritaria de recursos.
2. **Validación del Modelo Predictivo:** El modelo **Random Forest Regressor** logró explicar el **65% ($R^2 = 0.65$)** de la variabilidad delictiva. Si bien existe un margen de error (MAE: 131 delitos), el modelo demostró ser robusto para detectar tendencias generales y zonas de alto riesgo, validando su utilidad como herramienta de planeación estratégica.

3. **Estacionalidad Rígida:** El análisis de correlación reveló una consistencia temporal extremadamente alta ($r > 0.95$) entre los meses. Esto implica que los patrones delictivos son estructurales y no aleatorios; un municipio violento en enero tiende a permanecer así todo el año.

Posibles Mejoras y Futuras Líneas de Investigación

Reconociendo las limitaciones actuales del modelo (especialmente la dificultad para predecir en municipios con cero o muy baja incidencia), se proponen las siguientes líneas de trabajo:

- Enriquecimiento de Datos (Variables Externas):

Actualmente, el modelo predice basándose solo en el historial delictivo. Para futuras iteraciones, es crucial incorporar variables sociodemográficas como densidad poblacional, tasas de desempleo, presupuesto policial y niveles de educación. Esto permitiría entender no solo dónde ocurre el delito, sino por qué ocurre.

- Modelado Segmentado (Manejo de Ceros):

Dado que el 22% de los registros son ceros, se recomienda implementar un enfoque de "Modelos de Dos Etapas" (Hurdle Models) o Regresión Inflada de Ceros. Primero, un modelo de clasificación determinaría la probabilidad de que ocurra un delito (Sí/No), y posteriormente, el modelo de regresión predeciría la cantidad.

- Análisis de Series de Tiempo:

Dada la fuerte correlación mensual, se sugiere explorar modelos específicos de series temporales (como ARIMA o Prophet) para realizar pronósticos a futuro (ej. predecir la incidencia de 2026) en lugar de solo explicar el comportamiento actual.

Referencias y Bibliografía

Fuentes de Datos:

- **Kaggle (Fuente del Dataset):**
  - Nombre del Dataset: *Incidencia Delictiva en México (Datos del SESNSP).*
  - URL: [https://www.kaggle.com/datasets/beelzabi/crimen-mx]
- **Fuente Oficial (Origen de los datos):**
  - Secretariado Ejecutivo del Sistema Nacional de Seguridad Pública (SESNSP). (2023). *Incidencia Delictiva del Fuero Común*. Gobierno de México. Recuperado de: https://www.gob.mx/sesnsp/acciones-y-programas/datos-abiertos-de-incidencia-delictiva

Recursos Técnicos y Librerías:

- **Scikit-learn Developers.** (2023). *Scikit-learn: Machine Learning in Python*. Documentación técnica sobre Random Forest y métricas de evaluación. Recuperado de: https://scikit-learn.org/stable/
- **Pandas Development Team.** (2023). *Pandas: Powerful Python Data Analysis Toolkit*. Recuperado de: https://pandas.pydata.org/
- **Waskom, M.** (2021). *Seaborn: Statistical Data Visualization*. Recuperado de: https://seaborn.pydata.org/

Bibliografía Académica y Metodológica:

- **Breiman, L.** (2001). Random Forests. *Machine Learning*, 45(1), 5-32. (Referencia teórica del algoritmo utilizado).
- **Tukey, J. W.** (1977). *Exploratory Data Analysis*. Addison-Wesley. (Referencia metodológica para el uso de Boxplots y Rango Intercuartílico - IQR).