1. Calidad de datos

1.1. Revisión de DataFrame

- La primera etapa consistió en revisar la dimensión del dataset: 'dataset_examen_1.csv', en donde se identificó un shape de (774952, 19). Los siguientes pasos consistieron en:
    - Imprimir el nombre de las columnas que integraban el dataset.
    - Visualizar los primeros 5 registros de la tabla, con el objetivo de tener una visión preliminar de su comportamiento y tipo de variable.
    - También se revisó el formato original en el cual se cargaron las columnas, y en esta visualización preliminar se identificaron columnas con 'Unnamed' headers que posteriormente fueron eliminadas del dataset original.

1.2. Etiquetado de variables

- c_longitud
- c_latitud
- c_geopoint
- v_ao_hechos
- v_mes_hechos
- v_agencia
- v_unidad_investigacion
- v_mes_inicio
- v_ao_inicio
- v_id
- v_categoria_delito
- v_calle_hechos
- v_alcaldia_hechos
- v_calle_hechos2
- d_fecha_hechos
- d_fecha_inicio

1.3. Regristros duplicados y nulos

- No se encontro ningun registro duplicado.
- Respecto a la completitud de las variables se encontró que la columna ‘t_calle_hechos2’ contaba con un 40.48% del total de sus valores, razón por la cual se procedió a eliminarla. El resto de las variables tuvieron un valor por arriba del 80%.

1.4. Limpieza y normalizacion de variables

A las siguientes variables se aplicó función de normalización y limpieza con el objetivo de eliminar la presencia de caracteres especiales, signos de puntuación, cadenas de texto en columnas donde únicamente estaba permitido el uso de valores numéricos enteros o flotantes, discrepancias en el idioma y escritura, así como valores no válidos (un ejemplo de ello fueron: c_longitud y c_latitud).

- d_fecha_hechos
- t_delito
- t_categoria_delito
- t_fiscalia
- t_colonia_hechos
- t_alcaldia_hechos
- d_fecha_inicio
- v_mes_inicio
- c_longitud
- c_latitud

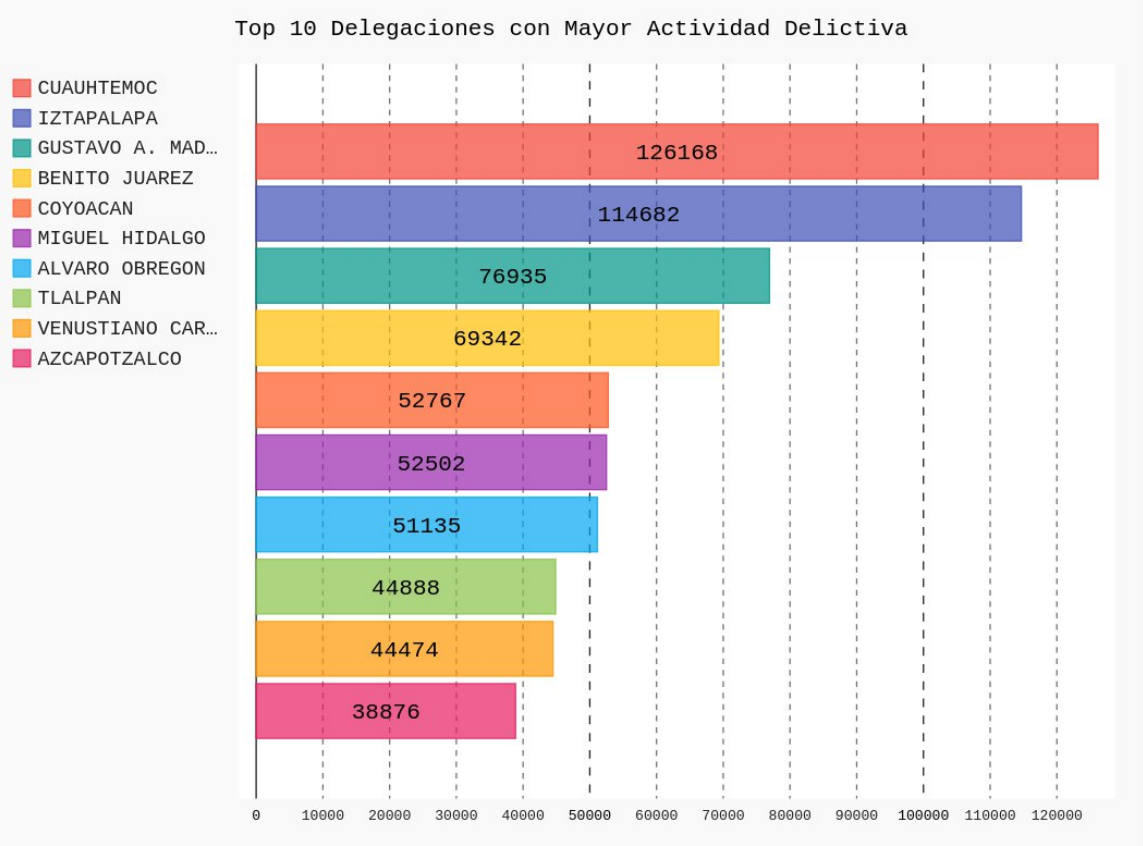
Dentro de los tratamiento aplicados a las variables, sobresalen:

- Homologación en el formato de ‘d_fecha_hechos’ pues se tenía la prescencia de fechas expresadas en formato: cadena de texto - numero, parentesis y caracteres especiales.
- En ‘v_ao_hechos’ se tenían números negativos, años superiores al actual, meses indicados con letras y años no válidos.
- A las variables de tipo fecha “d_”, después de homologarlas y ajustar valores incorrectos, se les aplicó la conversión a formato de fecha para poder manipularlas sin problemas en los pasos posteriores.

2. Análisis exploratorio de datos

2.1. Identificación de las Delegaciones con Mayor Actividad Delictiva

En el gráfico siguiente se pueden visualizar las 10 delegaciones con mayor actividad delictiva indeitficadas a través del número total de delitos cometidos durante todo el periodo comprendido en el dataset. Es importante mencionar que se tienen más de delegaciones, sin embargo para efectos de visualización de decidió mostrar únicamente aquellas con un mayor número de crímenes.

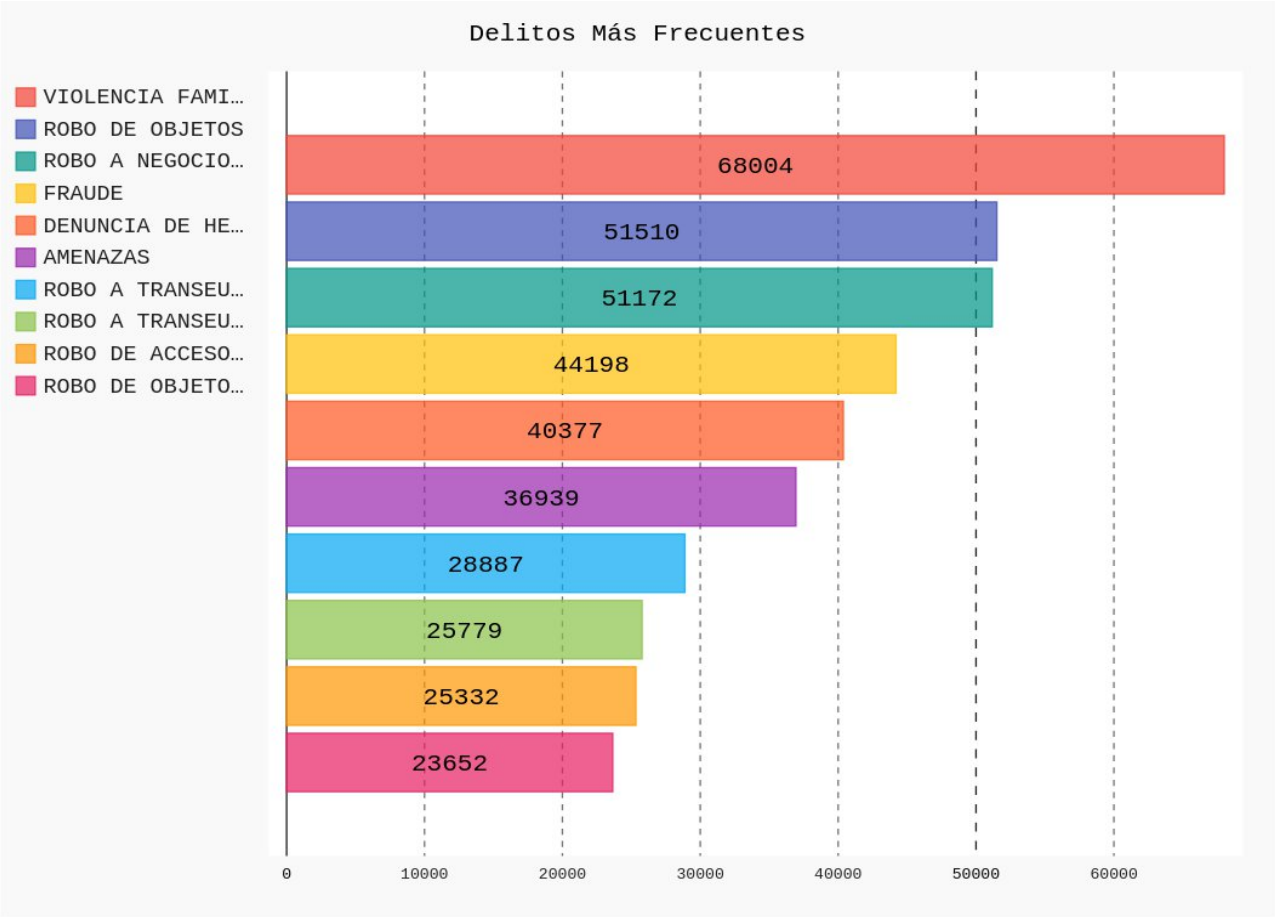


Algunas observaciones importantes de este análisis indican que el 70% de estas delegaciones fueron citadas en medios nacionales e internacionales (como Infobae - Argentina - Tier 1) como las delictivas de la Ciudad de México, en donde Iztapalapa ocupó el 1er lugar con mayor incidencia en delitos de **Violencia Familiar**.

2.2. Distribución de los delitos más frecuentes

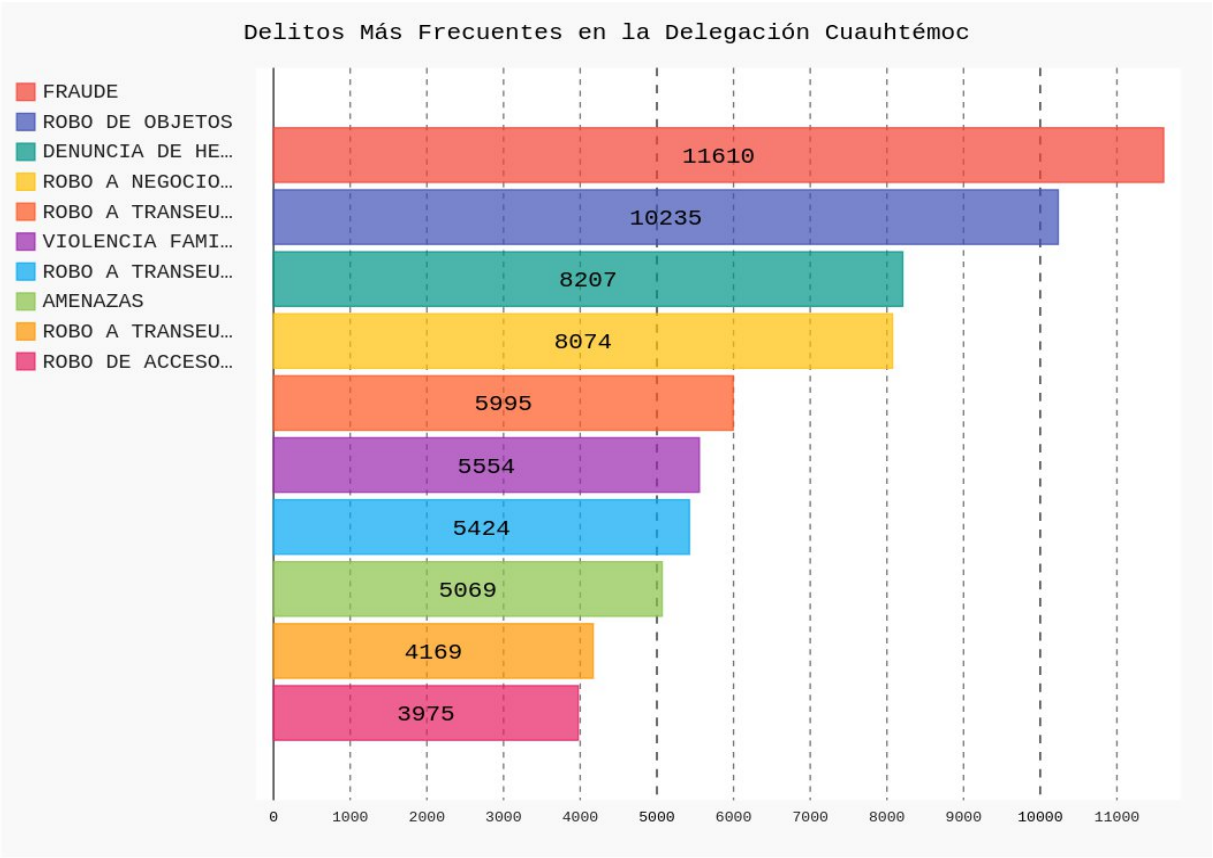
Siguiendo el mismo esquema que el análisis anterior, se identificaron los delitos con mayor frecuencia, y en el gráfico siguiente únicamente se muestra el Top 10 de éstos, ya que el resto si bien también es importante considerarlos, se encuentran por debajo de estos valores.

Como se puede visualizar, **Violencia Familiar,** ocupa el 1er lugar y de la información recolectada en el análisis anterior, su principal foco se ubica en **Iztapalapa**. Seguida por **Robo de Objetos** y **Robo a Negocio Sin Violencia**.



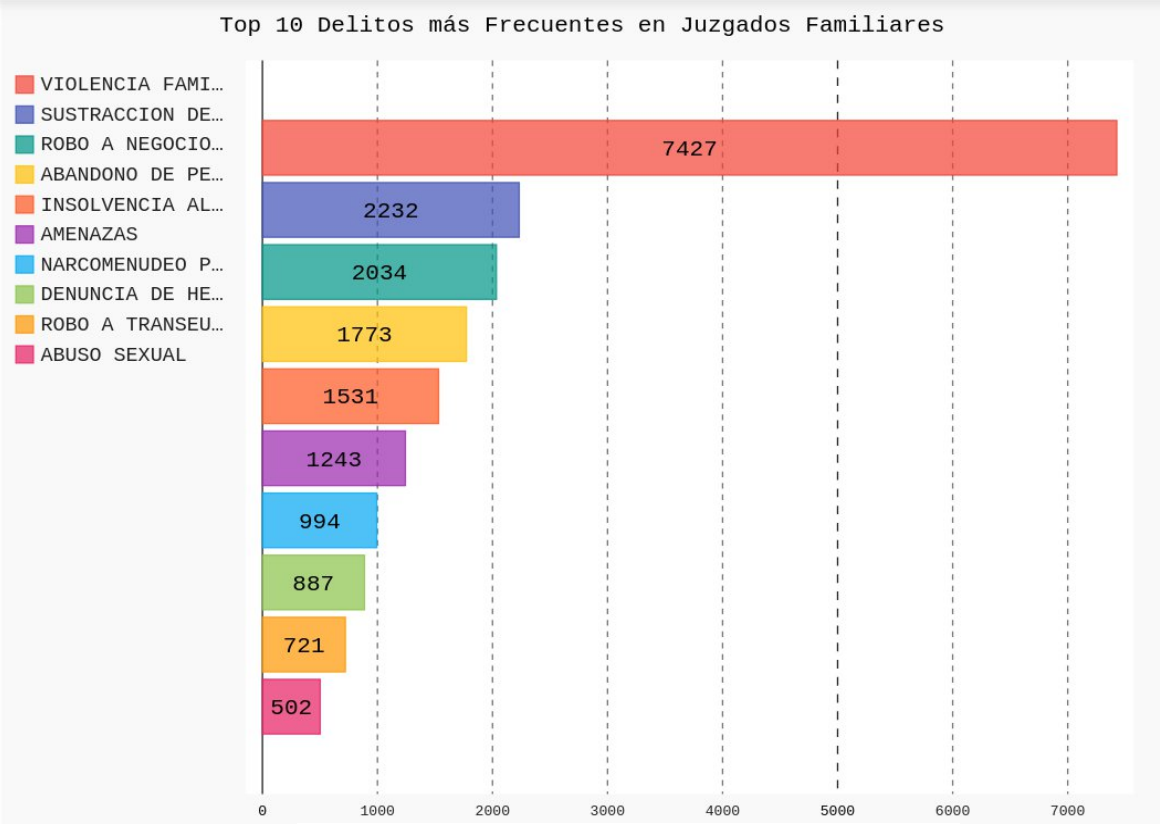
2.3. Tipo de delitos más frecuentes en la delegación con mayor incidencia delictiva.

El planteamiento para trabajar en este análisis se deriva de investigar los tipos de delitos más frecuentes en la delegación con mayor incidencia delictiva. Después de trabajar en identificar el comportamiento del volumen total de delitos suscitados dentro de cada delegación, se identificó que "**Cuauhtémoc**" fue la que tuvo una mayor cantidad de crímenes cometidos, con la distribución que se muestra en el siguiente gráfico:



2.4. En la fiscalia de "Juzgados Familiares", ¿Cuáles son los delitos más frecuentes?

Tal y como podemos ver en el siguiente gráfico, el Top 10 de delitos con mayor frecuencia dentro de esta categoría corresponde a:



t_delito	Sum
VIOLENCIA FAMILIAR	7427
SUSTRACCION DE MENORES	2232
ROBO A NEGOCIO SIN VIOLENCIA	2034
ABANDONO DE PERSONA	1773
INSOLVENCIA ALIMENTARIA	1531
AMENAZAS	1243
NARCOMENUDEO POSESION SIMPLE	994
DENUNCIA DE HECHOS	887
ROBO A TRANSEUNTE EN VIA PUBLICA CON VIOLENCIA	721
ABUSO SEXUAL	502
LESIONES INTENCIONALES POR GOLPES	316
ROBO A TRANSEUNTE DE CELULAR CON VIOLENCIA	272
CORRUPCION DE MENORES	265
PORNOGRAFIA INFANTIL	210
PORTACION ARMA/PROHIB.	149

3. Ingeniería de variables

A continuación se enlistan las variables agregadas a partir de las variables existentes:

- difference_in_datetime: esta variable fua agregada para identificar el comportamiento del tiempo entre el momento en que un delito sucedia y era reportado respectivamente (computado a partir de los días transcurridos entre las variables: fecha_hecho y fecha_inicio.
- categoria_delito_norm:esta variable corresponde a la normalización de v_categoria_delito para identificar a los delitos como: Alto, Medio y Bajo Impacto.
- delegacion_norm. al igual que la variable anterior, ésta se creo para normalizar la alcadía de la ciudad de acuerdo a la zona de pertenencia.
- weekday: describe el nombre de la semana que se desprende de la fecha_hecho.
- Quarter: es utilizada para identificar el trimestre del año en el que se ubica el delito de acuerdo a la variable fecha_hecho.
- IsWeekend: Se utilizó para identificar si el delito tuvo lugar en fin de semana o no.
- Monthday: Fue utilizada para identificar el día de acuerdo al mes calendario en el ocurrió el delito (día del mes).

Cuestionario

¿Qué es la ciencia de datos?

La ciencia de datos es una combinación multidisciplinaria de inferencia de datos, desarrollo de algoritmos y tecnología para resolver problemas analíticamente complejos con ayuda de herramientas estadísticas.

Además, contribuye a revelar tendencias y genera información que las empresas pueden utilizar para tomar decisiones comerciales de manera inteligente, basada en datos y resultados.

¿Qué habilidades debe dominar un científico de datos?

Un Científico de Datos debe dominar las siguientes habilidades:
Conocimientos en matemáticas y estadística.

Habilidades sólidas y robustas en programación y manejo de bases de datos.

Habilidades para comunicar de manera efectiva sus hallazgos de forma sencilla y en términos que sean fáciles de entender e interpretar por el negocio.

Establecer canales de comunicación efectivos con expertos en materia empresarial y liderazgo.

Conseguir elaborar gráficos atractivos, explicables y fáciles de interpretar por el negocio.

Del lado de Soft Skills, un Científico de Datos debe ser: estratégico, proactivo y cooperativo, además de innovador y apasionado por su trabajo en la manipulación/tratamiento de la información y su respectiva comunicación con stakeholders.

¿Qué es una tabla analítica?

Una tabla analítica es el resultado de procesar los datos desde su estado en crudo, aplicando las herramientas estadísticas, matemáticas y de programación necesarias para que la información sea utilizada para la toma de decisiones dentro de un ambiente empresarial, así como para servir de datos de entrada para modelos de machine learning.

¿Qué es una tabla analítica?

Una tabla analítica es el resultado de procesar los datos desde su estado en crudo, aplicando las herramientas estadísticas, matemáticas y de programación necesarias para que la información sea utilizada para la toma de decisiones dentro de un ambiente empresarial, así como para servir de datos de entrada para modelos de machine learning.

¿Qué es la ingeniería de variables?

Es el proceso de tranformar datos en características/features que permitan representar de mejor forma el problema, así como brindar un mejor entendimiento del mismo. A través de este proceso se manipulan los datos para corregir errores, ajustar variables y crear nuevas (si así se requiere) y obtener como resultado un mejor rendimiento durante su uso en modelos de machine learning.

Describe la ingeniería de variables posible por cada tipo de variable

Codificación a Nivel Ordinal

Codificación a Nivel Nominal => Variables Categóricas

* One-Hot Encoding
* Count Encoding
* Target Encoding

Variables Continuas

* Min-Max Standard Scaler
* Standard Scaler

Texto

* Count Vectorizer
* TF-IDF Vectorizer