Instituto Politécnico Nacional

Escuela Superior de Computo

Trabajo Terminal
2021-B106

"Análisis de crimen en el oriente del Valle de México"

Trabajo para cumplir con la opción de titulación curricular en la carrera de

Ingeniero en Sistemas Computacionales

Presentan

Mendoza Cuellar Jose Oscar	2015010680
Morales Zaranda Victor Hugo	2015010727

Director

Dr. Roberto Zagal Flores

México, CDMX a 22 de marzo de 2022

Índice

Índice de figuras

Resumen

El comportamiento en tiempo y espacio de crímenes en zonas urbanas como el Valle de México es tema vigente y que aún no se ha estudiado en extenso desde la perspectiva de la integración de datos abiertos, en este sentido es de principal utilidad para tomadores de decisiones en ámbito de política pública, para conocer las regiones donde los crímenes tienen una mayor concentración en el tiempo, caracterizando la frecuencia de ocurrencia y ubicación, con el objetivo de que estos generen nuevas estrategias en apoyo a la comunidad, así como para los habitantes del Valle de México interesados en mantenerse informados acerca del índice de delincuencia en su localidad y como este evoluciona a través del tiempo. El oriente del Valle de México es una región con una gran densidad poblacional en el país, como son las alcaldías y municipios de Nezahualcóyotl, Ecatepec, Gustavo A. Madero y Azcapotzalco. En este trabajo se propone el análisis de crimen en dicha región del valle de México desde una perspectiva espacio temporal, considerando análisis de concentración de crímenes usando minería de datos y Sistemas de Información Geográfica. El objetivo es descubrir y comprender cómo se comporta la concentración y densidad de crímenes de alto impacto en estos municipios y alcaldías, para mostrar patrones tanto geográficos como temporales.

Palabras clave: Hot Spots, Sistemas de Información Geográfica, Delitos de alto impacto, Minería de datos espaciales.

1. Introducción del trabajo terminal

1.1. Introducción y antecedentes

La violencia que sufre actualmente la población mexicana se traduce en un problema de seguridad pública, tanto por la dimensión que ha adquirido la muerte por dichas causas, como por los efectos materiales y emocionales que ocasiona, y cuyo origen se encuentra en factores históricos, demográficos, psicológicos, económicos, biológicos, sociales, entre otros. [1] Es sustancial mencionar que México es un país de altos contrastes y que nuestras raíces desde la colonización española han tenido cierta presencia de sufrimiento que después de varias luchas sociales iniciaron un proceso de transformación nacional, como lo es la lucha contra el crimen y la violencia que se presenta en nuestro día a día.

La delincuencia posee componentes geográficos inigualables ya que la mayor parte de los delitos ocurren en lugares concretos y los perpetran personas que viven o viajan a dicho lugar. [2] Si tenemos en cuenta este componente, el hecho delictivo puede ser interpretado de mejor manera y con más facilidad, partiendo del supuesto teórico de que los fenómenos sociales son dependientes del espacio donde suceden.

Cabe vislumbrar que analizar de manera precisa y oportuna la información criminal es un proceso clave para comprender los factores que impulsan las actividades delictivas, así como intuir posibles formas de operación de diversos grupos criminales o encontrar patrones que identifiquen el funcionamiento criminal en diversas zonas.

Como ya se ha mencionado, la distribución espacial de delitos puede proveer información muy valiosa, existe una gran cantidad de técnicas para el análisis de datos espaciales y geográficos, la técnica de hotspots (puntos críticos) describe áreas que tienen un alto índice de intensidad delictiva, los cuales suelen visualizarse mediante mapas, lo que facilita el análisis de áreas geográficas con relación a la delincuencia. Estas herramientas se están convirtiendo en ayuda fundamental para la vigilancia y para la comprensión y conocimiento de las diferentes zonas de densidad donde es más probable que ocurran crímenes.

De acuerdo con la Procuraduría General de Justicia de la CDMX y la Secretaría de Seguridad Pública de la CDMX, las alcaldías con mayores índices de crímenes reportados desde 2014 son: Iztapalapa, Cuauhtémoc, Gustavo A. Madero, Benito Juárez, Coyoacán y Tlalpan. [3] Debido a la gran segmentación de estas zonas y el tamaño de área que abarcan, los informes policiales suelen subestimar el crimen de manera sustancial, por lo que pueden llegar a ser engañosos. Por estos motivos la delincuencia es muy difícil de medir y los datos obtenidos deben tratarse con precaución.

Asimismo, en el Estado de México, tenemos a Ecatepec con sus casi 1.7 millones de habitantes siendo el municipio más poblado del país, por lo que no sorprende que también sea el municipio más violento para las mujeres, registrando un 17% de los feminicidios en dicha entidad. [4] Según informa la Dirección General de Seguridad Ciudadana de Nezahualcóyotl una gran parte de los homicidios registrados se vincula con el robo de

vehículo con violencia, pues el 17% de los 805 homicidios captados entre 2013 y 2018 (143) están vinculados con este delito. [5] Finalmente, la extorción aumento un 105% en la alcaldía Gustavo A. Madero entre enero y septiembre de 2019 con respecto a 2018 y el robo a negocio con violencia también creció en un 64% en relación con el mismo lapso. [6]

1.2. Planteamiento del problema

El reto es definir mecanismos de integración de datos de delitos ya que las estructuras pueden cambiar según la región, pero sobre todo identificar zonas de concentración de delitos con claridad y útiles es una tarea que depende de la densidad de los diferentes tipos delitos en ciertas zonas y periodos de tiempo esto es necesario para poder analizar el comportamiento delictivo y conocer patrones de comportamiento de delitos, la complejidad aumenta si consideramos otras fuentes de datos relacionadas como densidad de población o niveles socioeconómicos.

La principal motivación para este presente estudio es analizar el clima de violencia y crímenes en relación con las zonas de interés, las cuales consideramos que son relevantes pues son de gran afluencia no solo por los habitantes si no por gran parte de personas de la Zona Metropolitana, ya sea por cruzarlas para llegar al trabajo o centro de estudios.

Este análisis nos puede ayudar a mejorar la seguridad pública, identificar tendencias emergentes y planificar estrategias de prevención del delito. En estados unidos este análisis de zonas de concentración del delito proporciona herramientas estadísticas para ayudar a los organismos encargados de hacer cumplir la ley y a los investigadores de justicia penal en sus esfuerzos de mapeo del crimen. [7]

A lo largo de este trabajo terminal, se busca la realización de un dashboard que muestre los conocimientos adquiridos sobre la concentración de delitos en el oriente y norte del Valle de México haciendo uso de las técnicas de minería de datos y análisis espacial, generando así una fuente de información confiable, siendo esta de acceso público para informar, concientizar y funcionar como un apoyo en la medida de lo posible a los tomadores de decisiones en el ámbito político.

1.3. Objetivos

1.3.1. Objetivo particular

Desarrollar un prototipo de software para procesar datos de delitos de alto impacto del oriente del Valle de México (Nezahualcóyotl, Ecatepec, Gustavo A. Madero y Azcapotzalco) para identificar zonas de concentración de crímenes, patrones de comportamiento, así como tendencias en tiempo y en espacio, empleando técnicas data mining y de sistemas de información geográfica.

1.3.2. Objetivos Específicos

- Seleccionar y comprender las fuentes de datos de crímenes en las zonas de estudio
- Definir una metodología de data mining para caracterizar en el tiempo y el espacio los conjuntos de datos de crimen de las zonas de estudio
- Acoplar a la metodología mecanismos de análisis espacio temporal que apliquen para el procesamiento de datos de crímenes
- Estudiar y definir posibles algoritmos específicos para el análisis de datos de crímenes
- Explorar mecanismos para identificar y visualizar zonas de concentración o densidad de crímenes de alto impacto de acuerdo con la región de estudio.
- Diseñar una arquitectura de data mining que permite el descubrimiento no trivial de conocimiento en crimen, procesamientos de datos y de visualización de resultados.

1.4. Justificación

El reto es definir mecanismos de integración de datos de delitos ya que las estructuras pueden cambiar según la región, pero sobre todo identificar zonas de concentración de delitos con claridad y útiles es una tarea que depende de la densidad de los diferentes tipos delitos en ciertas zonas y periodos de tiempo esto es necesario para poder analizar el comportamiento delictivo y conocer patrones de comportamiento de delitos, la complejidad aumenta si consideramos otras fuentes de datos relacionadas como densidad de población o niveles socioeconómicos.

La principal motivación para este presente estudio es analizar el clima de violencia y crímenes en relación con las zonas de interés, las cuales consideramos que son relevantes pues son de gran afluencia no solo por los habitantes si no por gran parte de personas de la Zona Metropolitana, ya sea por cruzarlas para llegar al trabajo o centro de estudios.

Este análisis nos puede ayudar a mejorar la seguridad pública, identificar tendencias emergentes y planificar estrategias de prevención del delito. En estados unidos este análisis de zonas de concentración del delito proporciona herramientas estadísticas para ayudar a los organismos encargados de hacer cumplir la ley y a los investigadores de justicia penal en sus esfuerzos de mapeo del crimen. [7]

A lo largo de este trabajo terminal, se busca la realización de un dashboard que muestre los conocimientos adquiridos sobre la concentración de delitos en el oriente del Valle de México haciendo uso de las técnicas de minería de datos y análisis espacial, generando así una fuente de información confiable, siendo esta de acceso público para informar, concientizar y funcionar como un apoyo en la medida de lo posible a los tomadores de decisiones en el ámbito político.

2. Marco Teórico

A continuación, se describirán algunos de los procesos, conceptos, técnicas y metodologías mencionadas en el presente trabajo.

2.1. Fundamentos técnicos generales

ETL

Por su siglas en inglés (Extract, Transform and Load) es un proceso que funciona como la base de los almacenes de datos. Un sistema ETL correctamente diseñado permite extraer datos de uno o diversos sistemas de origen, hace cumplir estándares de calidad y consistencia de datos, además de ajustar los datos para que las fuentes separadas se puedan usar juntas con la finalidad de entregar los datos en un formato listo a desarrolladores para que estos puedan crear aplicaciones y los usuarios finales consigan tomar mejores decisiones basadas en conocimiento.

Aunque la construcción de un sistema ETL es una actividad que no se aprecia mucho del lado del usuario final, se estima que fácilmente consume el 70% de los recursos necesarios en la implementación y mantenimiento de una típica Data Warehouse. [8]

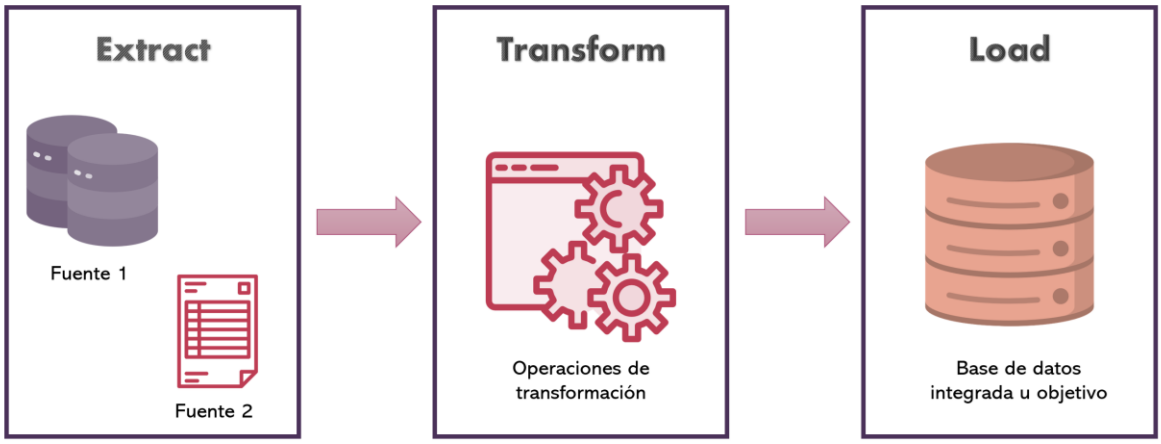

Figura 1: Fases del proceso ETL

ETL cobró popularidad en la década de 1970 cuando las organizaciones comenzaron a utilizar múltiples repositorios de datos, o bases de datos, para almacenar diferentes tipos de información de negocios. La necesidad de integrar datos que se diseminaran por estas bases de datos creció con rapidez. ETL se convirtió en el método estándar para extraer datos de diferentes fuentes y transformarlos antes de cargarlos en una fuente pretendida o destino. Con el tiempo, el número de formatos, fuentes y sistemas de datos ha aumentado enormemente.

Extraer, transformar, cargar (ETL) es ahora sólo uno de varios métodos que utilizan las organizaciones para recopilar, importar y procesar datos. [9]

Data warehouse

De acuerdo con William H. Inmon, Denominado padre del almacén de datos y un respetado arquitecto líder en la construcción de sistemas de almacenamiento de datos "Un almacen de datos es una recopilación de datos no volátil, variable en el tiempo, integrada y orientada a temas que respalda el proceso de toma de decisiones de la administración".

El almacenamiento de datos proporciona arquitecturas y herramientas para que los ejecutivos de negocios, políticos, empresarios, etc. organicen, comprendan y utilicen sistemáticamente sus datos transformados en información para tomar decisiones estratégicas.

Los sistemas de almacenamiento de datos son herramientas valiosas en el mundo competitivo y en constante evolución día a día. En los últimos años, muchas empresas han gastado millones de dólares en la construcción de almacenes de datos en toda la empresa. Mucha gente siente que, con la creciente competencia en todas las industrias, el almacenamiento de datos es la última arma de marketing imprescindible: una forma de retener a los clientes es aprendiendo más sobre sus necesidades. [10]

Análisis exploratorio de datos (EDA)

De acuerdo con IBM y su portal online [11], el análisis exploratorio de datos es ampliamente utilizado por los científicos de datos para analizar e investigar conjuntos de datos y resumir sus características principales, a menudo empleando métodos de visualización de datos y técnicas de estadística descriptiva. Esto con la intensión de determinar la mejor forma de manipular los datos disponibles para obtener las respuestas que buscamos.

El principal propósito de realizar un EDA es ayudar a ver los datos antes de hacer suposiciones. Puede servir para identificar errores obvios, así como para comprender mejor los patrones dentro de los datos, detectar valores atípicos o eventos anómalos, encontrar relaciones interesantes entre las variables, en incluso probar hipótesis o verificar suposiciones.

Dashboard

En los últimos años se han desarrollado un sin número de herramientas para la visualización de datos basadas en la web, ya que esta creciente área de estudio tiene como objetivo comunicar los datos de manera clara y efectiva a través de la representación gráfica, [12] pues podemos aprovechar las técnicas de visualización para descubrir relaciones de datos que de otro modo no serían fácilmente observables al mirar los datos sin procesar. [10]

Según Stephen Few, [13] un dashboard es “Una visualización de la información más importante y necesaria para lograr uno o más objetivos; consolidados y ordenados en una sola pantalla para que la información pueda ser monitoreada de un vistazo”.

2.2. El crimen y su estudio.

El entender el por qué las personas llegan a cometer delitos es una tarea ardua, pues al fin y al acabo no existe una verdad universal que pueda ser aplicada en todos los casos, dado que la delincuencia es el resultado de múltiples factores que se dan en usa sociedad. Entre dichos factores podemos destacar el bajo nivel educativo, una distribución de la riqueza desigual, salarios bajos o falta de empleo, usos y costumbres de cada localidad, entre otros.

Si abordamos el delito de manera más teórica y especializada, la teoría de patrones del crimen sostiene y ha demostrado que el crimen no se distribuye aleatoriamente a través del espacio o tiempo ni tampoco de una manera uniforme, sino que hay patrones de agrupamiento llamados puntos calientes (hot spot); los cuales pueden definirse como pequeños lugares con una alta concentración de delitos durante un determinado periodo de tiempo. [15]

Un factor también muy importante para tomar en cuenta es la densidad de población, pues como lo asegura el presidente de la Comisión de Seguridad de la COPARMEX, Salvador López Contreras, “la diferencia de densidad poblacional también hace que a mayor número de habitantes obvio mayor número de eventos” [16] lo que significa que con una mayor acumulación de población y por ende más delitos, se amerita también ampliar las estrategias de protección social e inteligencia en prevención del delito.

Delito

Se puede definir través del diccionario jurídico un delito como: “La acción u omisión que castigan las Leyes Penales, entendido en el sentido más general de la expresión, será una forma de comportamiento desviado que se considera grave dentro del sistema social y que es calificado como tal por los órganos legislativos con competencia para ello.” [17]

Delito de alto impacto

La coordinación nacional antisecuestro define un delito de alto impacto como “Aquellos que por el bien jurídico tutelado que dañan, la forma en que se cometen y la conmoción social que generan, además del sentimiento de inseguridad”, se han señalado como de alto impacto: El homicidio, el secuestro, la trata de personas, son algunos de los delitos de alto impacto, que es impostergable su erradicación. [18]

2.3. La información espacial y su valor (falta)

Casi todo lo que sucede a nuestro alrededor, sucede en alguna parte y como seres amantes de la información buscamos tener un registro de lo que pasa en nuestro alrededor, tanto así que de igual forma buscamos registrar el ambiente que nos rodea de la forma más específica y realista posible con diversos objetivos que van desde lo político hasta lo social.

Dichas representaciones nacen del estudio de la geografía que ha sido importante para los humanos desde tiempo ancestrales, los antiguos nómadas comprendían la ubicación de sus presas y como las manadas de animales se movían en función del clima y orografía, los colonizadores trazaban nuevos mapas en cada conquista e incluso la sociedad de hoy en día ha adquirido un sentimiento de pertenencia generado al lugar que habita.

La geografía aplicada, en forma de mapas e información espacial, ha servido para el descubrimiento, la planificación, la cooperación, y conflicto durante al menos los últimos 3.000 años. Los mapas se encuentran entre los documentos más hermosos y útiles de la civilización humana, y la información espacial tiene un gran impacto en nuestras vidas al ayudarnos a producir los alimentos que comemos, la energía que quemamos, la ropa que usamos y las diversiones que disfrutamos. [19]

Hot Spot

Un hot spot es una pequeña unidad geográfica donde el crimen se concentra, ocasionando que la posibilidad de victimizaciones en estas áreas aumente, estos pueden consistir en cualquier cosa desde un bloque que incluye las intersecciones en uno o ambos extremos, la distancia de un par de manzanas, todo un complejo de apartamentos o todo un centro comercial, es decir, no hay un tamaño predeterminado de un hot spot. [14]

//Técnicas de minería de datos y análisis espacial

//Estadísticas espaciales: CrimeStat

2.4. Zonas de Estudio

Se ha decidido expandir el área de estudio del presente Trabajo Terminal a la zona centro de la Ciudad de México, debido a que esta presenta índices delictivos similares o incluso mayores dependiendo la alcaldía, lo que aumenta la fidelidad del estudio y la congruencia de los datos obtenidos.

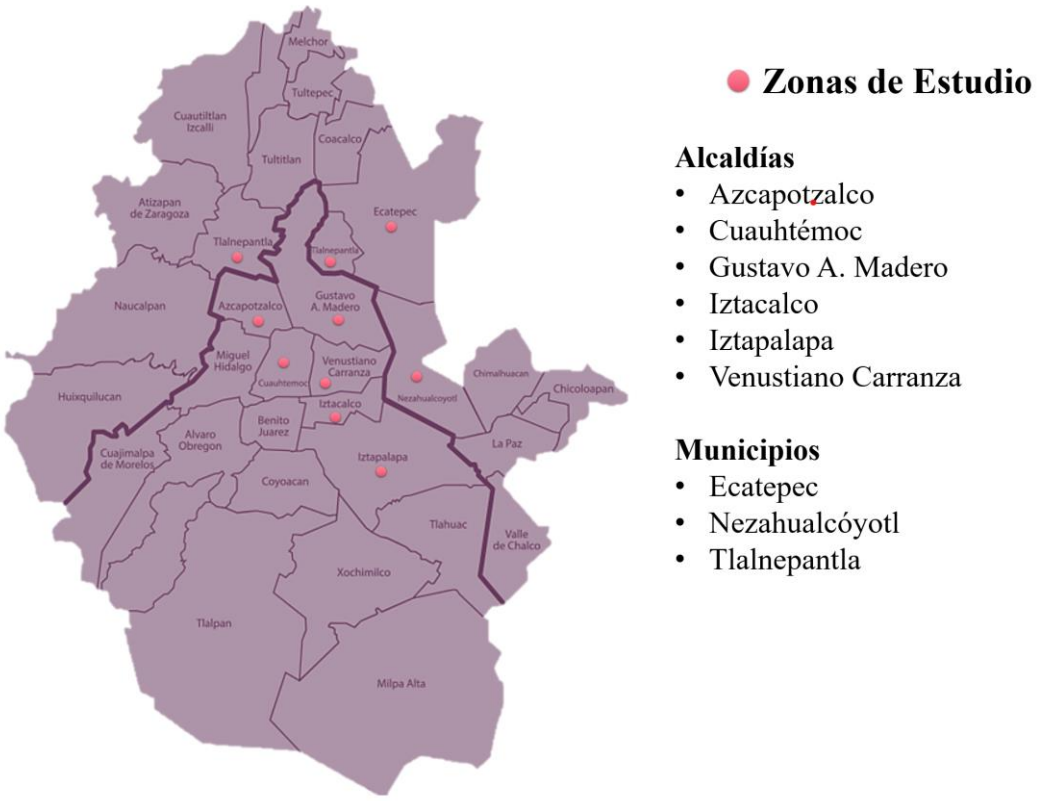

Figura 2: Mapa de las diversas localidades de estudio

3. Estado del arte

3.1. Estudios relacionados al análisis del crimen.

Se han estudiado 4 trabajos relacionados con el presente plantado, el primero es usado para nutrir el sistema de justicia de Estados Unidos, analizar datos del crimen y encontrar comportamientos que a simple vista o con un análisis superficial no se pueden encontrar, siendo necesario hacer uso de técnicas de minería de datos y de análisis espacial con el fin de visualizar la información recopilada. El otro es un trabajo desarrollado por la Universidad Autónoma del Estado de México en el que de igual forma se hizo uso de estas técnicas para representar los hallazgos encontrados en la incidencia del delito callejero en la Ciudad de México. En el siguiente tenemos una página web que es un claro ejemplo que lo que buscamos hacer, en este Dashboard se muestran datos y estadísticas de los diferentes delitos en los estados de la República Mexicana, tales como lo son robos, asesinatos, extorsión. Etc. Considerándose así una buena fuente de consulta, encajando con lo que nosotros queremos brindar, claro siempre mejorando lo ya existente pues nuestro trabajo será más específico y por ende más completo. Por último, una tesis bastante completa sobre el análisis de crímenes en las delegaciones Benito Juárez, Coyoacán y Cuauhtémoc de la

Ciudad de México relacionado con las características socioeconómicas de la zona para entender su contexto.

3.1.1. **Tabla comparativa sobre estudios** expuestos.

Nombre	Características	Lugar de desarrollo	Año
CrimeStat: Spatial Statistics Program for the Analysis of Crime Incident Locations **[7]**	Un sistema de software de estadísticas espaciales para el análisis de la ubicación de los incidentes delictivos. Interactúa con la mayoría de los programas GIS de escritorio. Incluye más de 100 rutinas estadísticas para el análisis espacial del crimen.	Houston Texas, National Institute of Justice, Agency of the United States Department of Justice.	2015
Spatial analysis of street crime in Mexico City **[8]**	Trabajo de investigación de patrones espaciotemporales de incidencia del delito callejero en la Ciudad de México durante 2018. Hacen uso de técnicas como Hot Spots.	México, Universidad Autónoma Del Estado De México	2018
Análisis De Los Datos Disponibles De Incidencia Delictiva	Un Dashborad web, que presenta estadísticas de diferentes delitos en los estados y algunos municipios de la República Mexicana.	México, Causa en Común.	2021
Análisis Espacial Del Delito: La Relación Entre El Delito Y Las Características Sociodemográficas En Las Delegaciones Benito Juárez, Coyoacán Y Cuauhtémoc Del D. F	Una tesis que analiza la distribución espacial de diferentes delitos como lo son el robo de vehículos, el robo a transeúnte y el homicidio, esto relacionado con el contexto socioeconómico de las delegaciones de Benito Juárez, Coyoacán y Cuauhtémoc de la ciudad de México.	Tijuana, B. C., México	2014

4. Análisis del proyecto

4.1. Necesidades detectadas

El crimen y los delitos lamentablemente ya son algo habitual en el país en que vivimos, existen zonas como las que identificamos en la figura 2, donde estos hechos ocurren con mucha más frecuencia, por este motivo se busca enriquecer, así como hacer más transparente y digerible la información de consulta que se tiene hasta el momento, esperando además que los descubrimientos exhibidos en el presente trabajo sean útiles para diversas aplicaciones.

Partiendo de este preámbulo, es necesario emplear de forma correcta y adecuada las diferentes metodologías que nos ayudan en el proceso de descubrimiento de información a partir de datos, iniciando por la búsqueda de bases de datos confiables que posteriormente puedan ser tratadas y transformadas según las necesidades del proyecto para después aplicar juiciosamente técnicas estadísticas o de minería de datos espaciales para descubrir tendencias o comportamientos, así como generar diversos mapas que identifiquen puntos de concentración.

Por lo anterior mencionado se busca desarrollar diferentes módulos, como lo son de exploración de datos, de cálculo de retención de información, de limpieza e integración o ETL, etc. haciendo uso notebooks de Jupyter con Python y sus diferentes librerías con la intención de hacerlo más demostrativo, para finalmente generar scripts en Python más reutilizables.

Así bien, las primeras preciciones encontradas para este trabajo son integrar los datos de la Cuidad de México con los del Estado de México para posteriormente realizar una adecuada limpieza tomando solo los datos de utilidad y realizando en caso de ser necesario transformaciones sobre ellos, facilitando el proceso de ETL. Después se busca el desarrollo de dos módulos.

- Módulo de análisis exploratorio del crimen que proyecte tendencias temporales por delitos
- Módulo de análisis geoespacial del crimen que genere diversos mapas relacionados a la concentración de delitos por zonas.

Cada módulo será alimentado por la base de datos histórica obtenida del ETL, esta base de datos será gestionada en Microsoft SQL Server (SQL Server). Los resultados de cada módulo serán expuestos en un dashboard web interactivo, capaz de mostrar los resultados obtenidos en cada módulo según los datos ingresados en el. Se plantea desarrollar el dashboard con un framework de Python llamado Django y usar librerías de visualización de datos tales como Dash de Plotly que están diseñadas específicamente para visualizaciones web.

4.2. Requisitos del software

En el proceso de definición de los requerimientos tendremos dos tipos de estos, los requerimientos funcionales y los requerimientos no funcionales, los cuales tendrán la siguiente nomenclatura mediante la cual los podremos identificar durante el proceso de desarrollo del trabajo terminal.

- RF-N

Los requerimientos funcionales serán identificados por la anterior nomenclatura donde RF es la abreviatura de “requerimiento funcional”, posteriormente habrá un guion medio que separa una “N” la cual indica el número del requerimiento que se está listando.

- RNF-N

Los requerimientos no funcionales serán identificados por la anterior nomenclatura donde RNF, como en el anterior caso indica una abreviatura, pero en este caso de “requerimiento no funcional”, seguida de un guion medio y el numero “N” que identifica el requerimiento no funcional evidenciado.

4.2.1. Requisitos funcionales

Los requerimientos funcionales para un sistema refieren lo que el sistema debe hacer. Son enunciados acerca de servicios que el sistema debe proveer, de cómo debería reaccionar el sistema a entradas particulares y de cómo debería comportarse el sistema en situaciones específicas. En algunos casos, los requerimientos funcionales también explican lo que no debe hacer el sistema. [20]

A continuación, en la Tabla 1 se identifican los requerimientos funcionales del presente proyecto, así como identificadores.

Tabla 1: Requerimientos funcionales

ID	Descripción
RF-1	El sistema tendrá un menú de inicio que permitirá identificar la clase de análisis que se desea hacer, dichos análisis pueden ser de tipo temporal o geoespacial.
RF-2	El sistema detectara tendencias temporales de crimen.
RF-3	El sistema realizara diversos análisis geoespaciales del crimen.
RF-4	El sistema permitirá seleccionar un rango de tiempo para el análisis temporal o geoespacial.

RF-5	El sistema permitirá seleccionar uno o varios tipos de delitos según el tipo de análisis, ya sea temporal o geoespacial.
RF-6	El sistema dejara cambiar el tipo de visualización de datos para el análisis temporal.
RF-7	El sistema implementara un módulo sobre visualización de series de tiempo.
RF-8	El sistema integrara al menos dos tipos de algoritmos para generar hot spots en mapas.
RF-9	El sistema podrá comparar al menos dos tipos de mapas con hot spots.
RF-10	El sistema integrara al menos dos tipos de algoritmos para generar mapas de calor.
RF-11	El sistema podrá generar mapas de coropletas.
RF-12	El sistema permitirá seleccionar la ubicación geográfica por alcaldía o municipio para realizar los análisis de tipo geoespacial.

4.2.2. Requisitos no funcionales

Son requerimientos que no se relacionan directamente con los servicios específicos que el sistema entrega a sus usuarios, pueden relacionarse con propiedades emergentes del sistema, como fiabilidad, tiempo de respuesta y uso de almacenamiento. De forma alternativa, pueden definir restricciones sobre la implementación del sistema. Los requerimientos no funcionales se suelen aplicar al sistema como un todo, más que a características o a servicios individuales del sistema [20]

A continuación, en la Tabla 2 se identifican los requerimientos no funcionales del presente proyecto, así como identificadores.

Tabla 2: Requerimientos no funcionales

ID	Descripción
RNF-1	El sistema funcionara como aplicación web.
RNF-2	*Desarrollo*: Los módulos de exploración, limpieza e integración de datos serán desarrollados en Jupyter Notebooks de Python.
RNF-3	*Desarrollo*: Los procesos de ETL y EDA será realizados en Python
RNF-4	*Desarrollo*: La creación de la base de datos histórica será generada en Python y gestionada con el SABD SQL Server
RNF-5	*Desarrollo*: Para el análisis geoespacial se utilizará el motor del sistema de información geografía de Qgis

RNF-6	*Desarrollo*: La interfaz del sistema será desarrollada con HTML, CSS, JavaScript.
RNF-7	El sistema debe desarrollarse en un periodo no mayor a 4 meses.
RNF-8	*Usabilidad*: la interfaz debe ser limpia, sencilla e intuitiva
RNF-9	*Usabilidad*: el sistema debe tener manuales de usuario
RNF-10	Si el sistema falla deberá de notificarlo al usuario, detenerse y reiniciarse.

4.3. Casos de uso

En su forma más sencilla, un caso de uso identifica a los actores implicados en una interacción, y nombra el tipo de interacción. En la Tabla 3 se identifica el formato a seguir para poder describir cada uno de estos casos de uso.

Tabla 3: Formato para la descripción de casos de uso

Identificador
Nombre
Descripción
Actor
Precondiciones
Postcondiciones
Escenario
Importancia
Urgencia

4.3.1. Listado de casos de uso

En la siguiente tabla se enlistas los casos de uso necesarios a tratar y atender en el desarrollo del presente proyecto

Tabla 4: Casos de uso

ID	Descripción
CU-1	Seleccionar tipo de análisis del crimen.
CU-2	Detectar tendencias del crimen.
CU-3	Realizar análisis geoespacial sobre el crimen
CU-4	Comparar dos mapas de hot spots
CU-5	Regresar al inicio.

4.3.2. Diagramas de caso de uso

Caso de uso CU-1

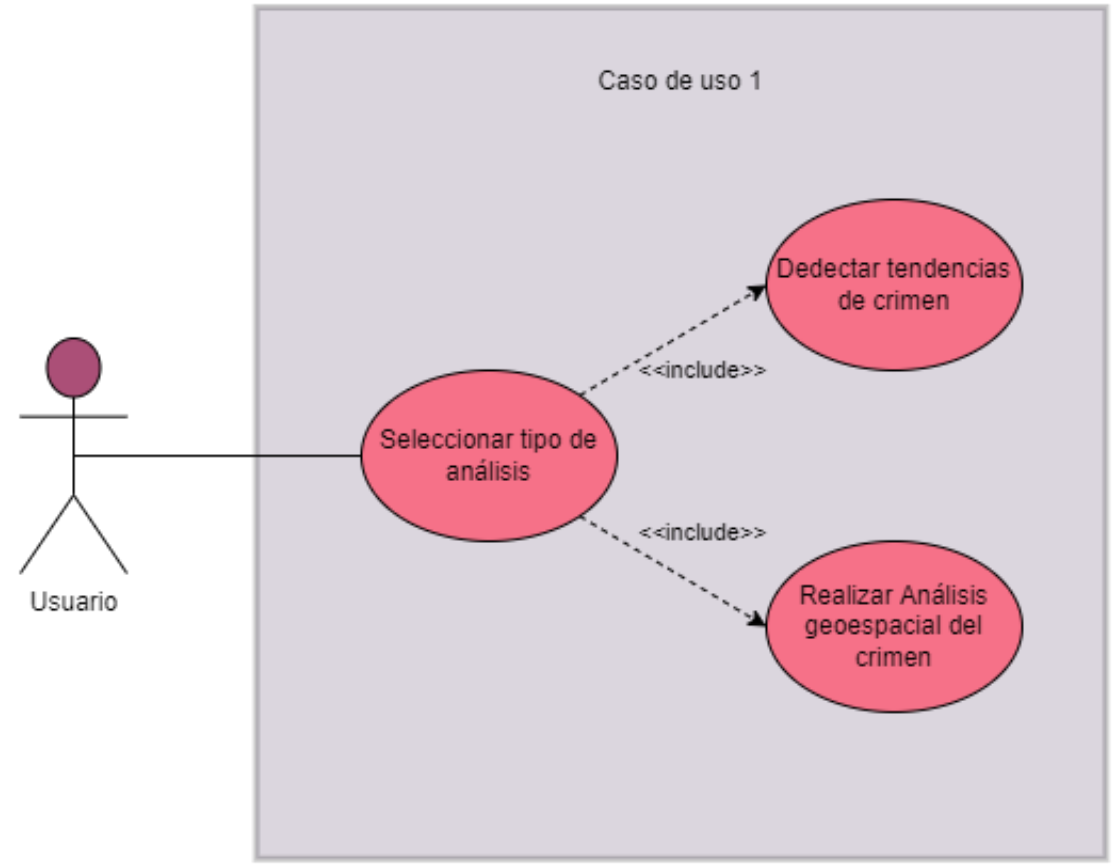

Figura 3: Diagrama del caso de uso CU-1.

A continuación, se muestra una descripción más detallada del caso de uso.

Tabla 5: Descripción caso de uso CU-1

Identificador CU-1	
Nombre	Seleccionar tipo de análisis.
Descripción	Es el inicio del sistema y aquí el usuario podrá seleccionar el tipo de análisis que desea realizar.
Actor	El usuario.
Precondiciones	Ninguna.
Postcondiciones	Ninguna.
Escenario	1. El usuario entra al inicio del sistema web. 2. El usuario selecciona el tipo de análisis que guste. 3. El sistema despliega la sección del sistema que solicito.

Importancia	Importante.
Urgencia	Inmediatamente.

Caso de uso CU-2

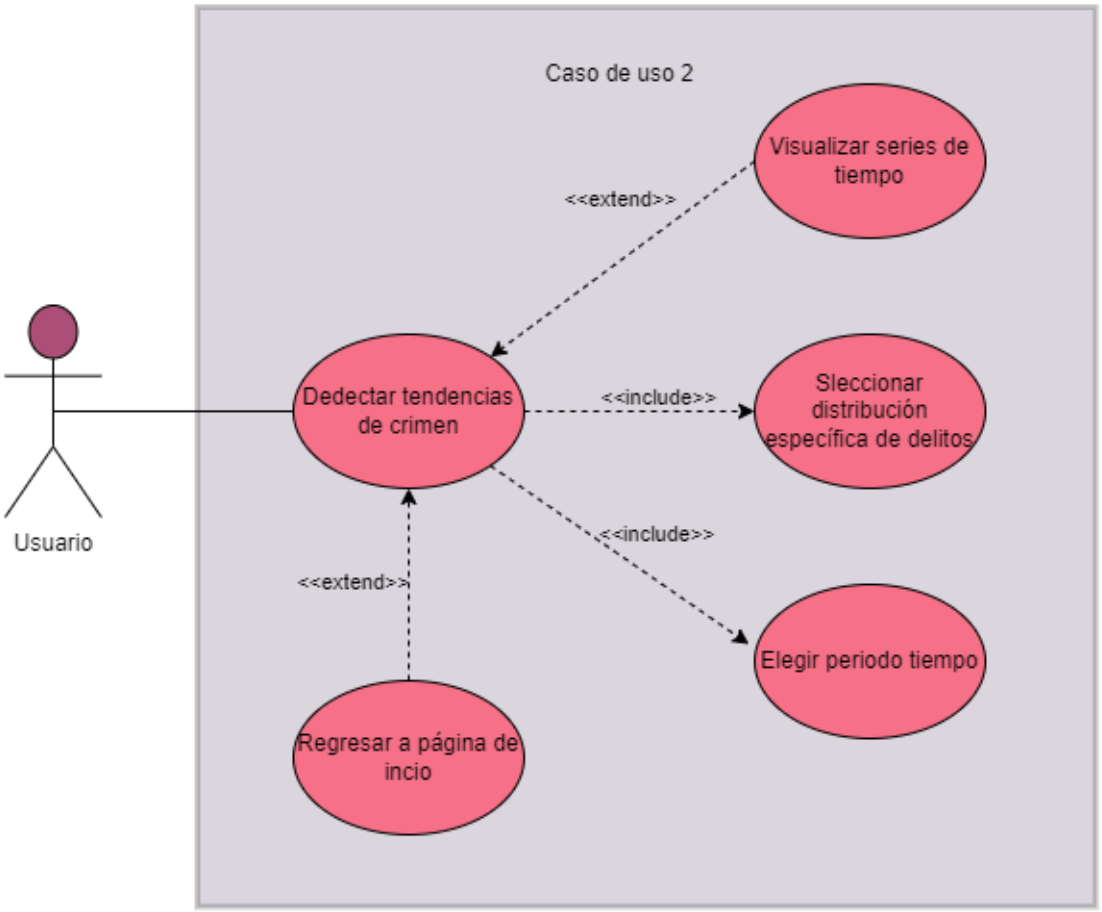

Figura 4 Diagrama de casos de uso 2.

A continuación, se muestra una descripción más detallada del caso de uso.

Tabla 6: Descripción del caso de uso CU-2

Identificador CU-2	
Nombre	Detectar tendencias del crimen.
Descripción	Sección del sistema que permite al usuario visualizar graficas referentes a tendencias respecto a delitos y periodos de tiempo seleccionados.
Actor	El usuario.

Precondiciones	El usuario debe estar en la sección para detectar tendencias del crimen.
Postcondiciones	El sistema detectara los cambios automáticamente por lo que la gráfica mostrada cambiara según sea seleccionados sus parámetros.
Escenario	1. El sistema despliega la sección para detectar tendencias del crimen. 2. La primera grafica mostrada estará en blanco. 3. El usuario deberá de seleccionar parámetros como tipo de crimen, periodo de tiempo y estilo de grafica. 4. El sistema actualizara la gráfica conforma la información cambie.
Importancia	Vital
Urgencia	Inmediatamente.

Caso de uso CU-3

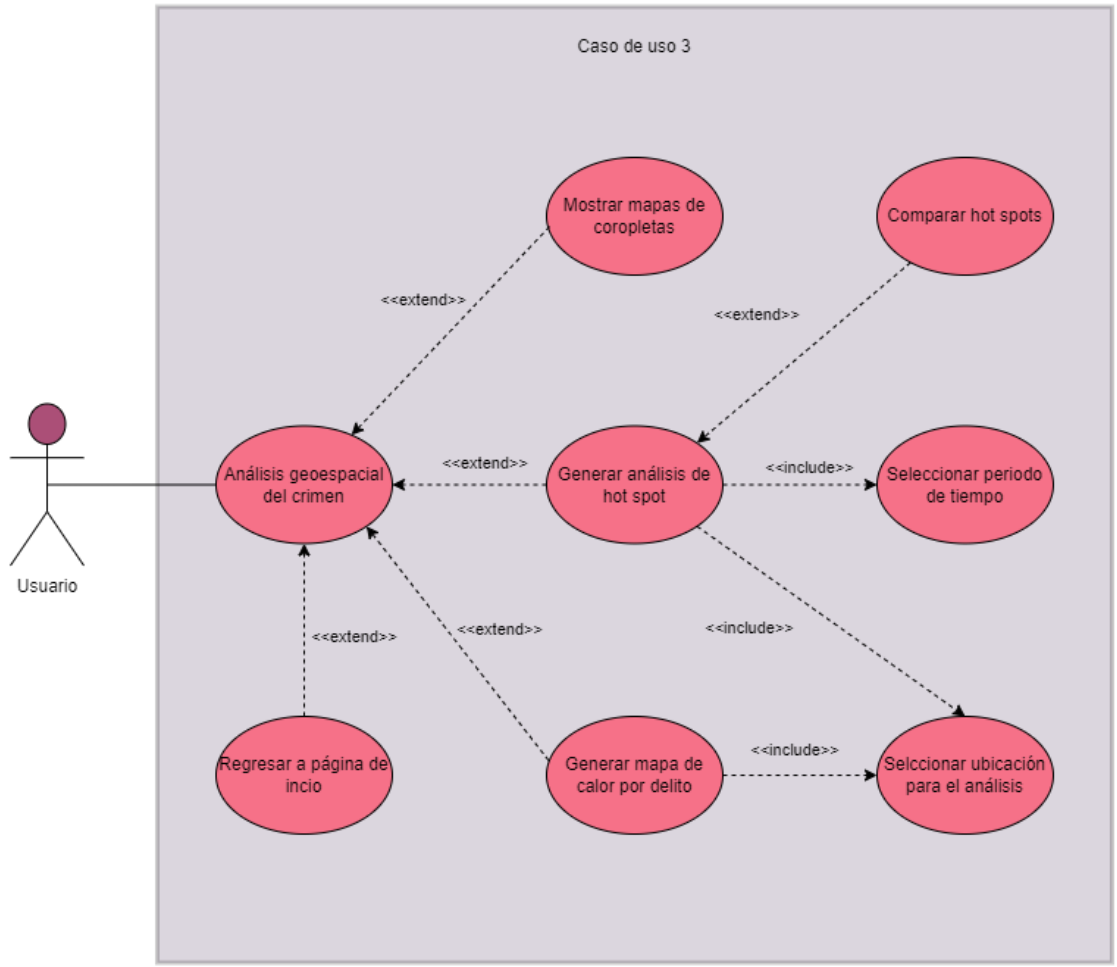

Figura 5: Diagrama de casos de uso 3.

A continuación, se muestra una descripción más detallada de los casos de uso.

Tabla 7: Descripción del caso de uso CU-3

Identificador CU-3	
Nombre	Realizar análisis geoespacial del crimen.
Descripción	Sección del sistema que permite al usuario visualizar mapas que identifiquen zonas de concentración de delitos.
Actor	El usuario.
Precondiciones	El usuario debe estar en la sección para visualizar mapas de concentración.
Postcondiciones	El sistema detectara los cambios automáticamente por lo que la gráfica mostrada cambiara según sea seleccionados sus parámetros.
Escenario	1. El sistema despliega la sección para visualizar mapas de concentración de delitos. 2. El primer mapa mostrado estará en blanco y será del valle de México. 3. El usuario deberá de seleccionar parámetros como tipo de crimen, periodo de tiempo y tipo de mapa. 4. El sistema permitirá agregar capas de visualización si así lo requiere el usuario. 5. El sistema actualizara el mapa para que este muestra la información solicitada
Importancia	Vital
Urgencia	Inmediatamente.

Tabla 8: Descripción del caso de uso CU-4

Identificador CU-4	
Nombre	Comparar dos mapas de hot spots.
Descripción	Sección del sistema que permite al usuario visualizar 2 mapas transpuestos que identifiquen zonas de concentración de delitos.
Actor	El usuario.
Precondiciones	1. El usuario debe estar en la sección para visualizar mapas de concentración. 2. Se deberá de seleccionar la opción de Comprar hot spots
Postcondiciones	El sistema detectara los cambios automáticamente por lo que el mapa mostrado cambiara según sean seleccionados sus parámetros.
Escenario	1. El sistema despliega la sección para comparar mapas de concentración de delitos por hot spots.

	2. El primer mapa mostrado estará en blanco y será del valle de México. 3. El usuario deberá de seleccionar parámetros como tipo de crimen, periodo de tiempo y tipo algoritmo de hot spot. 4. El sistema permitirá agregar capas de visualización si así lo requiere el usuario. 5. El sistema actualizara el mapa para que este muestra la información solicitada
Importancia	Vital
Urgencia	Inmediatamente.

Tabla 9: Descripción del caso de uso CU-5

Identificador CU-5	
Nombre	Regresar al inicio
Descripción	Permite al usuario regresar a la página de inicio del sistema.
Actor	El usuario.
Precondiciones	Ninguna.
Postcondiciones	El sistema enviara al usuario a la página de inicio.
Escenario	1. El sistema se encuentra en alguna sección. 2. El usuario selecciona la opción Regresar. 3. El sistema despliega la sección de inicio.
Importancia	Importante.
Urgencia	Puede esperar poco.

5. Diseño del proyecto

5.1. Arquitectura del sistema

A continuación, se muestra la arquitectura a usar para este proyecto.

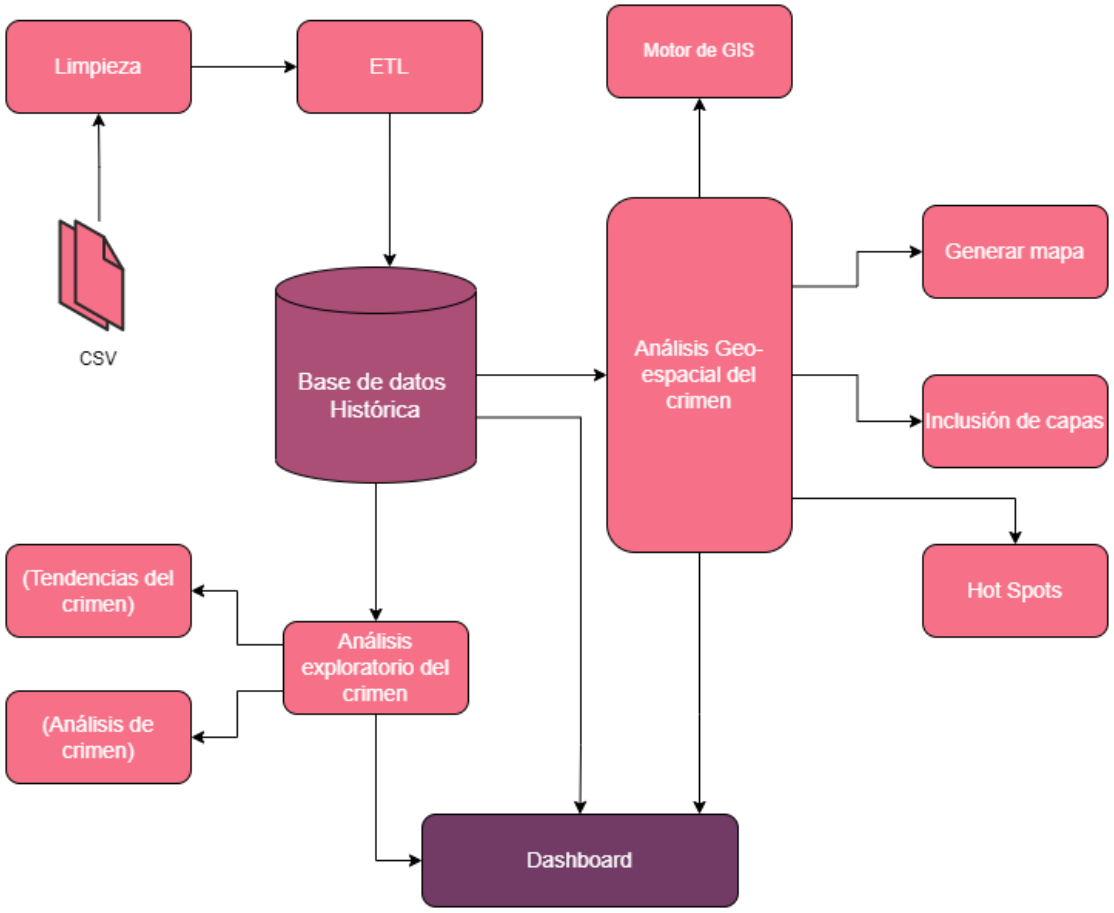

Figura 6: Diagrama de la arquitectura del sistema.

La arquitectura comienza con la base de datos de crímenes que ofrece el portal de datos abiertos de la CDMX [20], esta base se encuentra en formato CSV con varios datos faltantes y columnas que no resultan importantes, por lo que pasa a un módulo de limpieza y posteriormente a un proceso ETL en el que se realiza la transformación y ajuste de los datos como lo son la separación de información, conversión de tipo de dato, etc. Esto como ya se ha mencionado con anterioridad se realizó con notebooks de Jupyter implementados con Python y las bibliotecas Pandas y Pandas-Profiling. Posteriormente pasamos estos datos ya limpios y transformados a la base de datos histórica, para la cual usaremos SQL Server como sistema administrador de base de datos.

La base de datos histórica será utilizada en dos vertientes, una es el análisis exploratorio del crimen para obtener resultados como lo son las tendencias de delitos, donde de igual forma trabajaremos con Python y Jupyter notebooks. La otra vertiente comenzará con el análisis geoespacial del crimen, en el que tendremos módulos como la generación de mapas que identifiquen zonas de concentración bajo algoritmos de hot spots, mapas de calor y mapas de coropletas, así como la inclusión de capas con información geográfica en estos, para lo que nos apoyaremos de un motor GIS como lo es QGIS.

Para la parte del dashboard tenemos que este se nutre de la base de datos histórica, el análisis geoespacial del crimen y del análisis exploratorio del crimen, en este se representara y visualizaran toda la información en forma graficas y mapas con capas y hot spots obtenidos de los módulos ya mencionados.

5.1.1. Comportamiento de módulos o componentes

En el siguiente diagrama se darán las descripciones de los procesos y módulos que se implementarán en este proyecto de manera secuencial.

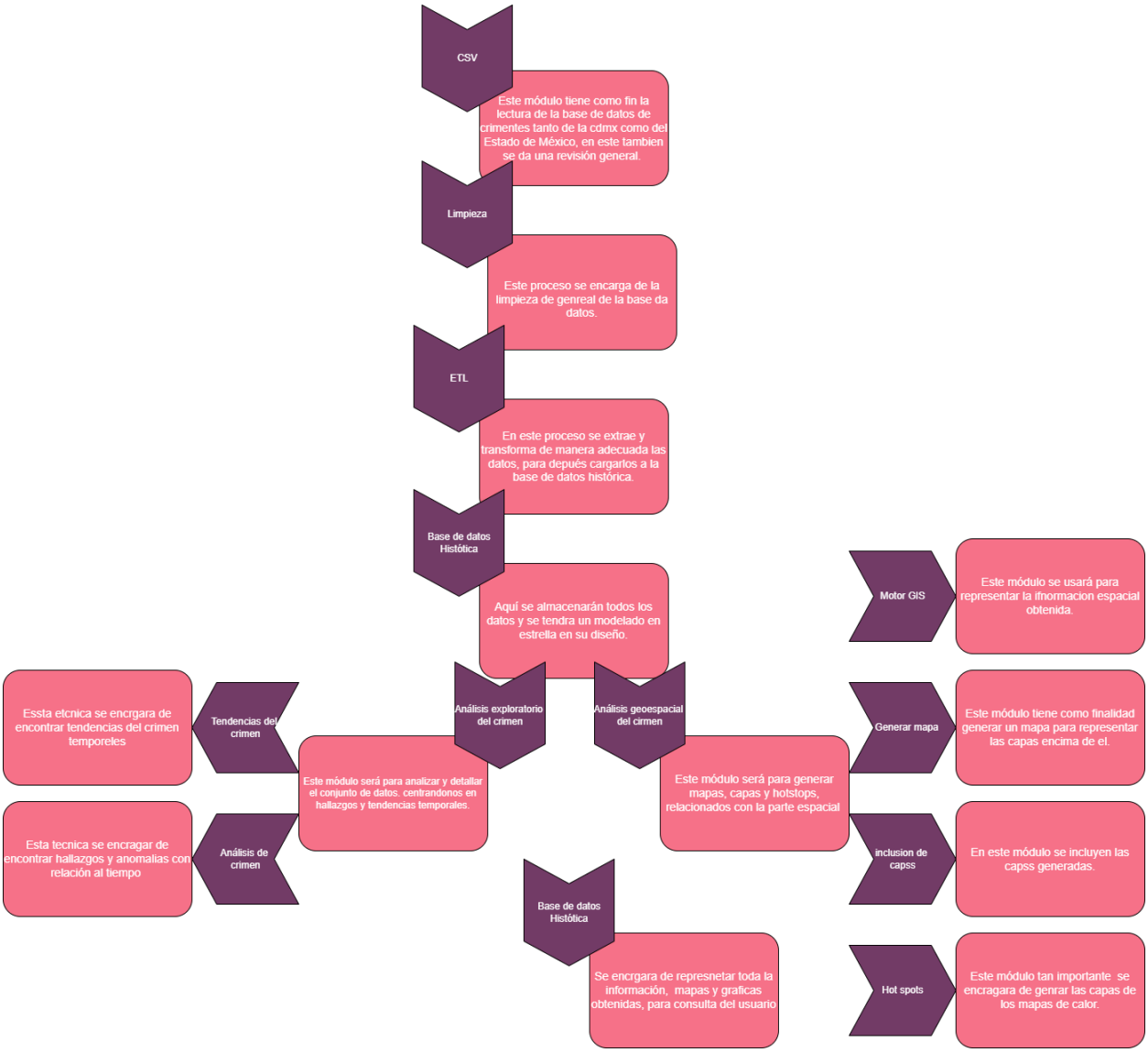

Figura 7: Diagrama del comportamiento de los módulos del sistema.

5.1.2. Diagrama general del funcionamiento del sistema

El siguiente diagrama representa el funcionamiento interno del dashboard, partiendo desde la petición que hace el usuario, hasta la visualización de los datos.

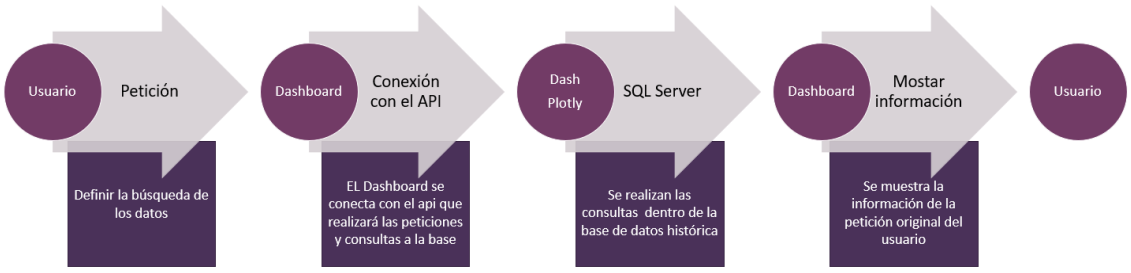

Figura 8: Diagrama del funcionamiento del sistema.

5.1.3. Diseño y análisis de la base de datos

La base de datos que ofrecen la página de datos abiertos de la CDMX en un principio contenía varias columnas y datos que no resultaban útiles, así como varios de estos estaban en formatos poco prácticos para su uso a lo largo de este proyecto, derivado de esto se tomó la decisión de realizar un preprocesamiento de la base de datos, contemplando solo hacer uso de la información que consideramos más útil, según nuestras necesidades y algunos cálculos realizados en el anexo Limpieza e integración de datos (ETL).

La base de datos original tenía la siguiente estructura en cuanto a sus columnas

Atributos del DataSet

```
print (df.columns)
```

```
Index(['ao_hechos', 'mes_hechos', 'fecha_hechos', 'ao_inicio', 'mes_inicio',
       'fecha_inicio', 'delito', 'fiscalia', 'agencia', 'unidad_investigacion',
       'categoria_delito', 'calle_hechos', 'calle_hechos2', 'colonia_hechos',
       'alcaldia_hechos', 'competencia', 'longitud', 'latitud', 'tempo'],
      dtype='object')
```

Estas columnas se describirán de mejor manera en el diccionario de datos presentado a continuación.

Numero	Atributo	Definición	Tipo de dato	Dominio
1	ao_hechos	Año en el que aconteció del delito	string	1906-2022
2	mes_hechos	Mes en el que aconteció el delito	string	Enero-Diciembre
3	fecha_hechos	Fecha completa con hora en la que aconteció el delito	string	aaaa-mm-dd hh:mm:ss
4	ao_inicio	Año en que se inició la denuncia	String	2016-2022

5	mes_inicio	Mes en que se inició la denuncia	String	Enero-Diciembre
6	fecha_inicio	Fecha completa con hora en la que se inició la denuncia	String	aaaa-mm-dd hh:mm:ss
7	delito	Especificación del delito	String	Violación Robo Allanamiento Etc…
8	fiscalía	Nombre de la fiscalía en la que se realizó la denuncia	String	
9	agencia	Clave de la agencia	String	CUH-5 TLP-3 FDS-1 IZP-8 Etc..
10	unidad_investigacion	Clave de la unidad de investigación a la que fue asignado el delito	String	UI-2SD UI-1SD UI-3SD FDS-7-03 Etc..
11	categoría_delito	Categoría que se le asigno al delito	String	Delito de bajo impacto Violación Hecho no delictivo Robo a transeúnte Robo de vehículo Etc..
12	calle_hechos	Calle en la que aconteció el delito	String	-
13	calle_hechos2	Segunda calle en la que aconteció el delito	String	-
14	colonia_hechos	Colonia en la que aconteció el delito	String	-
15	alcaldía_hechos	Alcaldía en la que aconteció el delito	String	-
16	competencia	Indica en que competencia encaja el delito	String	Fuero común Incompetencia
17	Longitud	Longitud de la ubicación del delito	String	-99.00 hasta 99.00
18	latitud	Latitud de la ubicación del delito	String	-180.00 hasta 180.00
19	Tempo	Desconocido	String	NAN

Se aplicación varios procesos de transformación de datos, pero al final del ETL se obtuvo una base de datos, limpia y completa, descrita en el siguiente diccionario de datos.

Numero	Atributo	Definición	Tipo de dato	Dominio
1	dia_hechos	Dia en el que aconteció el delito	int	0-31
2	mes_hechos	Mes en el que aconteció el delito	string	Enero-Diciembre
3	ao_hehcos	Año en el que aconteció del delito	int	1906-2022
4	delito	Especificación del delito	String	Violación Robo Allanamiento Etc…
5	fiscalía	Nombre de la fiscalía en la que se realizó la denuncia	String	
6	categoría_delito	Categoría que se le asigno al delito	String	Delito de bajo impacto Violación Hecho no delictivo Robo a transeúnte Robo de vehículo Etc..
7	calle_hechos	Calle en la que aconteció el delito	String	-
8	calle_hechos2	Segunda calle en la que aconteció el delito	String	-
9	colonia_hechos	Colonia en la que aconteció el delito	String	-
10	alcaldía_hechos	Alcaldía en la que aconteció el delito	String	-
11	longitud	Longitud de la ubicación del delito	Float	-99.00 hasta 99.00
12	latitud	Latitud de la ubicación del delito	Float	-180.00 hasta 180.00

Posteriormente ha de modelarse la

De tal forma que después de tener la anterior base de dato seleccionados los datos, se muestra el diseño de la base de datos siguiendo el modelo en estrella.

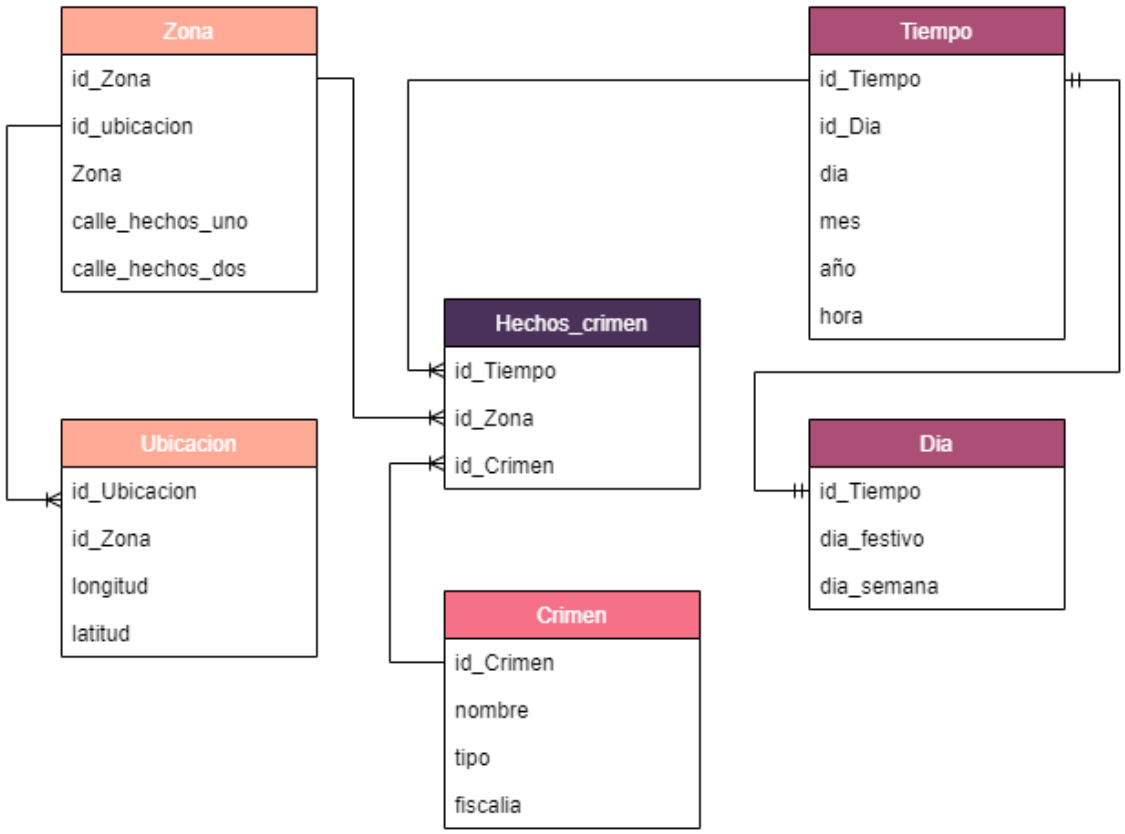

Figura 9: Diseño de la base de datos

5.2. Prototipado dashboard

Se considera que el dashboard web tenga una interfaz limpia, sencilla e intuitiva, presentara diversas gráficas y tipos de mapas, se hará uso de selectores, combobox y sliders para que se actualice en tiempo real la información al momento de seleccionarla sin necesidad de presionar botones innecesarios, pero manteniendo el control sobre la configuración que se desea hacer en la visualización de los mapas y graficas. Lo que se busca es que se tenga un tiempo de aprendizaje bastante corto y sea amigable con el usuario.

- Mockup 1 página de inicio:

Esta será la primera vista que se tenga cuando se entre a al dashboard, en ella se mostrará un pequeño resumen que hablara de en qué consiste el trabajo y permitirá seleccionar que tipo de análisis se desea realizar, si detección de tendencias del crimen o análisis geoespacial del crimen.

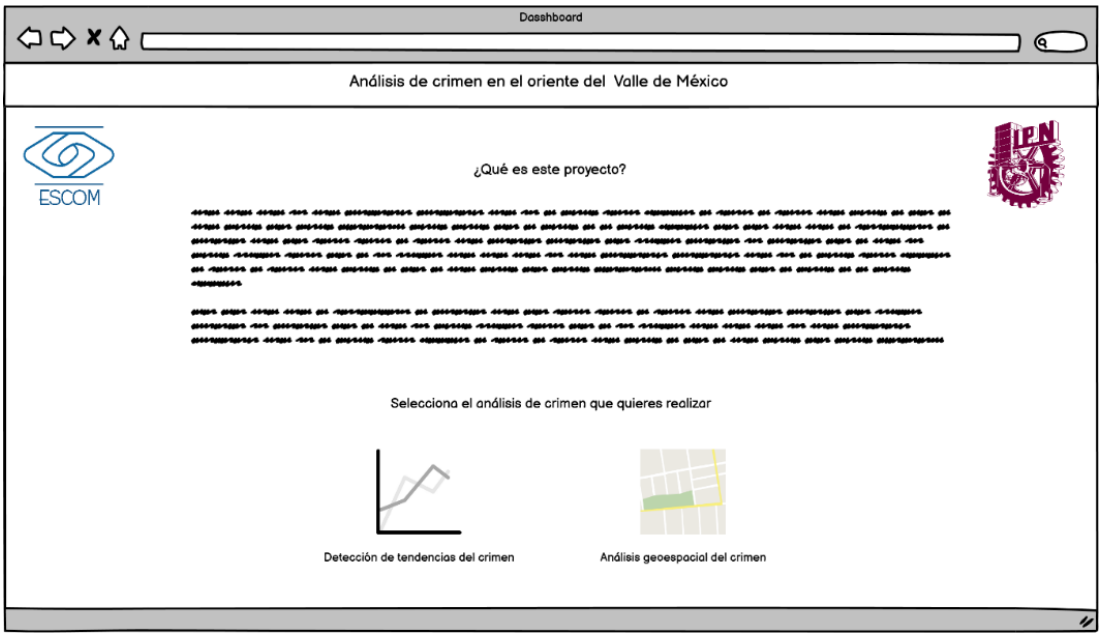

Figura 10: Mockup 1 página de inicio.

- Mockup 2 detección de tendencias del crimen:

Se mostrará una gráfica, la cual se actualizará en tiempo real según la selección del usuario, se hará uso de un combobox el cual desplegará varios tipos de gráfica, así como también otro en el que se podrá seleccionar uno o varios delitos. Para mover la gráfica a través del tiempo se contará con una slider para proporcionar un uso más fácil, por último se tendrá un botón para visualizar las series de tiempo y otro para regresar la página de inicio.

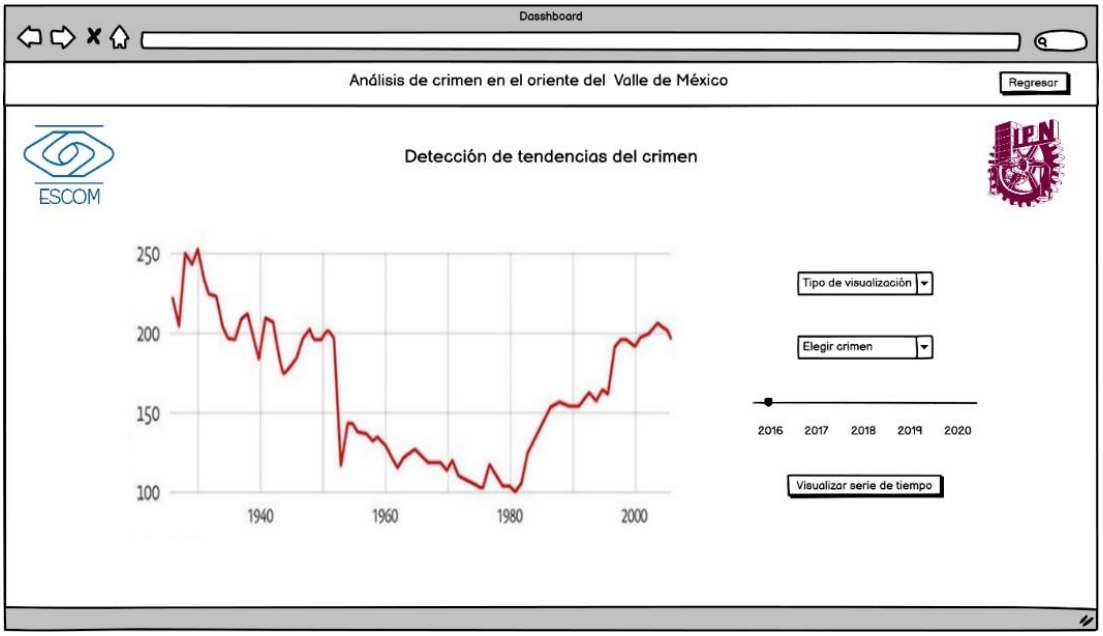

Figura 11: Mockup 2 detección de tendencias del crimen.

- Mockup 3 análisis geoespacial del crimen:

Se mostrará un mapa el cual se actualizará en tiempo real según la selección del usuario, se hará uso de un combobox que desplegará una lista para seleccionar una o mas zonas de estudio, otro combobox que de igual forma desplegará una lista para seleccionar el uno o varios crímenes y el ultimo combobox permitirá seleccionar que capas queremos que se muestran en el mapa. Para mover la gráfica a través del tiempo se contará con una slider y para seleccionar el tipo de análisis se tendrán 4 opciones, pero solo permitirá seleccionar uno a la vez. Se contará también con un botón para entrar en la función de comparar hot spots y otro para regresar a la página de inicio.

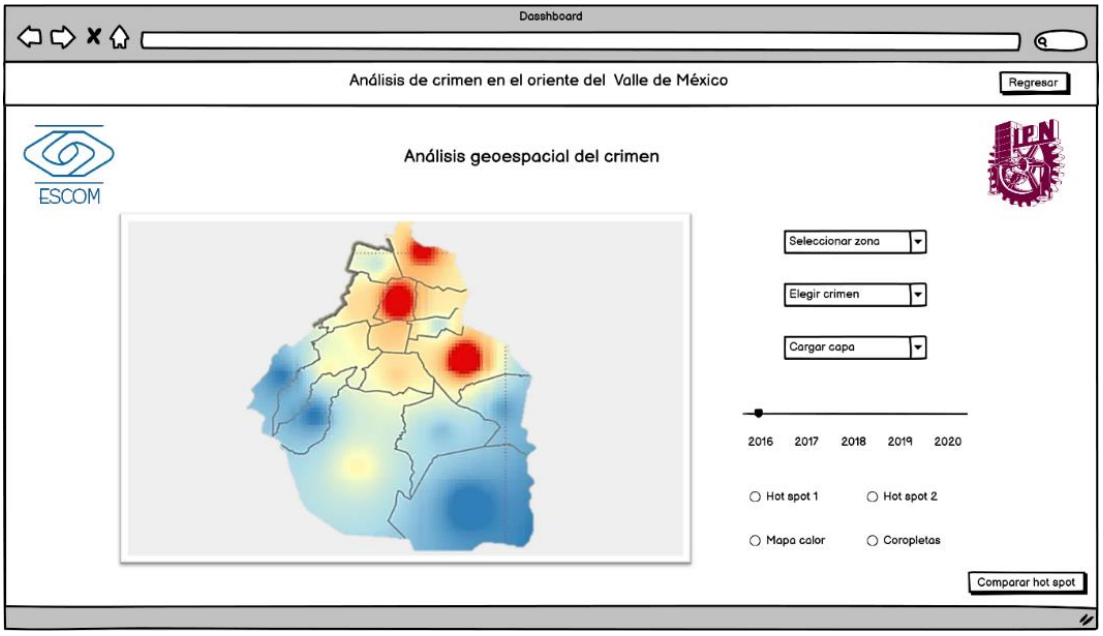

Figura 12: Mockup 3 análisis geoespacial del crimen.

- Mockup 4 Comparación de hot spots:

Para la función de comparar hot spots la funcionalidad esta basa en el mockup anterior, solo que se adecuo para poder configurar 2 hot spots, diferenciándose el uno al otro por un selector de color para representar la información de cada uno. La información de los dos se visualizará en el mismo mapa.

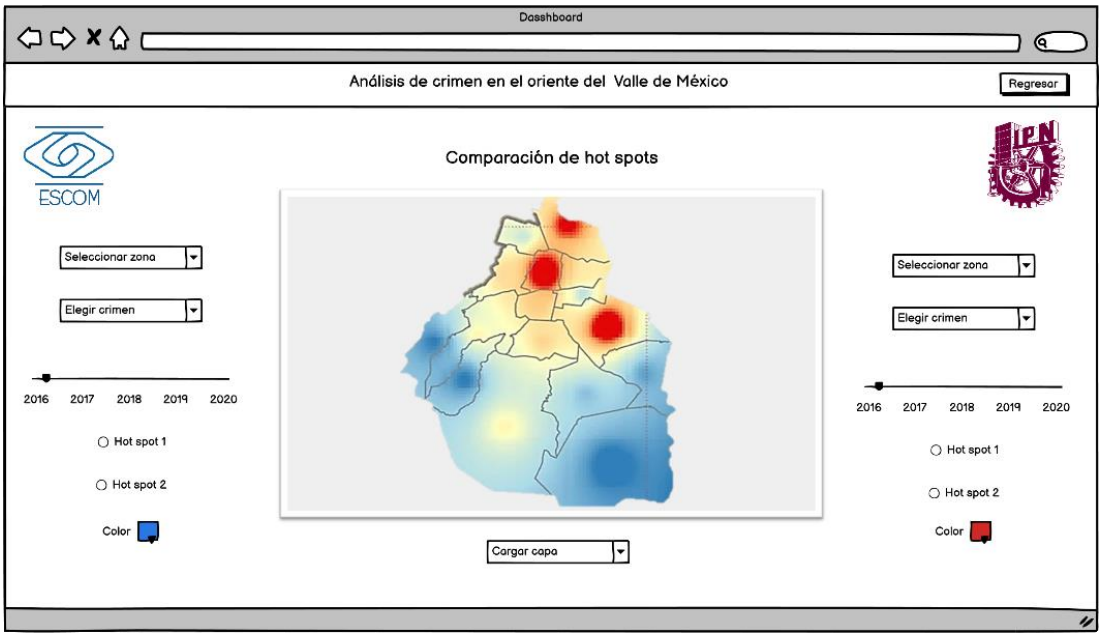

Figura 13: Mockup 4 Comparación de hot spots.

5.3. Metodologías usadas

6. Resultados preliminares

6.1. Experimentos

6.2. Ambientes y de configuraciones

6.3. Código fuente

A continuación, se proporcionarán los enlaces a los códigos empleados alojados en el repositorio del proyecto

1. Código de limpieza general de la base de datos: https://github.com/OscarMendoza-afk/TTAnalisisDeCrimen/blob/main/Codigo/LimpiezaGeneral.py
2. Notebook de Reconocimiento de la base de datos:

https://github.com/OscarMendoza-afk/TTAnalisisDeCrimen/blob/main/Notebooks/JNB_01_ReconocimientoDataSet.ipynb

3. Notebook de cálculos pata la eliminación de datos: https://github.com/OscarMendoza-afk/TTAnalisisDeCrimen/blob/main/Notebooks/JNB_02_Calculos.ipynb
4. Notebook de limpieza de la base de datos: https://github.com/OscarMendoza-afk/TTAnalisisDeCrimen/blob/main/Notebooks/JNB_03_ETL.ipynb

7. Conclusiones

8. Referencias

[1] R. A. Jiménez Ornelas, «La cifra negra de la delincuencia en México: sistema de encuestas sobre victimización.,» de *Proyectos legislativos y otros temas penales.*, Ciudad dde México, UNAM, Instituto de Investigaciones Jurídicas, 2003, pp. 167 - 190.

[2] G. P. J. M. Dávila, «La distribucion espacial de la delincuencia en el Pais Valenciano y su relacioncon algunas variables socioeconomicas,» *Investigaciones Geograficas,* vol. 6, pp. 187-205, 1988.

[3] C. A. P. G. &. L. R. Ramirez, «Exploring crime patterns in México City,» *Journal of Big Data,* vol. 6, nº 1, 2019.

[4] A. G. Rojas, «"Monstruo de Ecatepec": ¿por qué este municipio de México es el más peligroso para ser mujer?,» *BBC News,* 11 Octubre 2018.

[5] E. Jaime, Hot Spot Neza 10,000 cuadras resguardadas por vecinos, México: México Evalúa, 2020.

[6] Hallazgos Índice GLAC, «Extorsión y robo a negocio aumentaron en Gustavo A. Madero en 2019,» Animal Politico, 13 Noviembre 2019. [En línea]. Available: https://tinyurl.com/pdjnbre8. [Último acceso: 9 Noviembre 2021].

[7] N. Levine, «CrimeStat: A Spatial Statistics Program for the Analysis of Crime Incident Locations,» National Institute of Justice, Washington, D.C, 2015.

[8] K. Ralph y C. J. ., The data warehouse ETL toolkit: practical techniques for extracting, cleaning, conforming, and, Indianapolis, Estados Unidos: Wiley Publishing, Inc., 2004.

[9] SAS Institute Inc., «ETL Qué es y por qué es importante,» SAS, 2022. [En línea]. Available: https://www.sas.com/es_mx/insights/data-management/what-is-etl.html. [Último acceso: 8 Mayo 2022].

[10] J. Han, M. Kamber y J. Pei., Data mining: concepts and techniques, Waltham, Estados Unidos: Morgan Kaufmann, 2011.

[11] IBM Cloud Education, «Exploratory Data Analysis,» IBM, 25 Agosto 2020. [En línea]. Available: https://www.ibm.com/cloud/learn/exploratory-data-analysis. [Último acceso: 18 Abril 2022].

[12] A. Silberschatz, H. F. Korth y S. Sudarshan, Database system concepts, New York, Estados Unidos: McGraw-Hill, 2020.

[13] S. Few, Information Dashboard Design, Italia: O'Reilly Media, Inc., 2006.

[14] G. Farrell y W. Sousa, «Repeat Victimization and Hot Spots: The Overlap and Its,» *Repeat Victimization,*, vol. 11, pp. 221-240, 2001.

[15] M. P. Levy, «Opportunity, Environmental Characteristics, and Crime: An Analysis of Auto Theft Patterns,» U.S Departament of Justice , 2009. [En línea]. Available: https://www.ojp.gov/ncjrs/virtual-library/abstracts/opportunity-environmental-characteristics-and-crime-analysis-auto?msclkid=faef57f8d0d011ec8e19ddda8a162fd8. [Último acceso: 01 05 2022].

[16] M. F. Navarro, «Crimen organizado o delincuencia común, ¿qué afecta a la CDMX?,» Forbes, 25 06 2018. [En línea]. Available: https://www.forbes.com.mx/crimen-organizado-o-delincuencia-comun-que-afecta-a-la-cdmx/#:~:text=%E2%80%9CEn%20la%20Ciudad%20de%20M%C3%A9xico%20se%20evidencia%20la,prevenci%C3%B3n%E2%80%9D%2C%20consider%C3%B3%20el%20especialista%20en%20entrevista%20con%20Fo. [Último acceso: 01 05 22].

[17] L. R. Manzanera, «Diccionario Juridico,» 1981. [En línea]. Available: http://diccionariojuridico.mx/definicion/delito/.

[18] Cordinación Nacional Antisecuestro, «Capacitación FASP Y FORTASEG,» SEGOB, México, 2016.

[19] P. Bolstad, GIS Fundamentals: A first text on geographic information systems, 6th edition., White Bear Lake, Minnesota: XanEdu Publishing Inc, 2019.

[20] Gobierno de la Ciudad de México, «Carpetas de Investigación de la FGJ,» Portal de datos abiertos, 29 Enero 2021. [En línea]. Available: https://datos.cdmx.gob.mx/dataset/carpetas-de-investigacion-fgj-de-la-ciudad-de-

mexico/resource/48fcb848-220c-4af0-839b-4fd8ac812c0f. [Último acceso: 12 Febrero 2022].

ANEXOS

Es esta sección de detallan ciertos procesos importantes que no pueden ser completamente definidos a lo largo del desarrollo del trabajo terminal.

1) Exploración de la base de datos de crímenes de la Ciudad de México

Para la exploración de la base de datos se optó por usar el entorno virtual en Anaconda anteriormente descrito y Jupyter Notebook con la biblioteca de Pandas para la manipulación de datos.

Lo primero que se realizo fue conocer las dimensiones de la base de datos.

Dimensiones del DataSet

```
print(df.shape)
```

```
(1401331, 19)
```

Posteriormente se deben de conocer los atributos que conforman la base de datos, así como su tipo de dato y la cantidad de valores nulos que presentan.

```
print(df.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1401331 entries, 0 to 1401330
Data columns (total 19 columns):
 #   Column                Non-Null Count    Dtype
---  ------                --------------    -----
 0   ao_hechos             1400873 non-null  float64
 1   mes_hechos            1400873 non-null  object
 2   fecha_hechos          1400873 non-null  object
 3   ao_inicio             1401331 non-null  int64
 4   mes_inicio            1401331 non-null  object
 5   fecha_inicio          1401328 non-null  object
 6   delito                1401331 non-null  object
 7   fiscalia              1401329 non-null  object
 8   agencia               1401331 non-null  object
 9   unidad_investigacion  1401104 non-null  object
 10  categoria_delito      1401331 non-null  object
 11  calle_hechos          1397390 non-null  object
 12  calle_hechos2         539997 non-null   object
 13  colonia_hechos        1340993 non-null  object
 14  alcaldia_hechos       1397166 non-null  object
 15  competencia           337252 non-null   object
 16  longitud              1341941 non-null  float64
 17  latitud               1341941 non-null  float64
 18  tempo                 0 non-null        float64
```

Del paso anterior se pueden apreciar fácilmente ciertos atributos que no servirán de mucho para el análisis, pero esa tarea se realizara en la sección de limpieza de la base de datos.

A continuación, se identifican los valores mínimos, máximos, el promedio, la desviación estándar y un desglose por cuartiles de los atributos numéricos. Donde se pueden apreciar datos interesantes como que el primer delito registrado data de 1906, o el promedio de la latitud y la longitud, definiendo talvez una de las zonas más peligrosas o con mayor concentración de delitos de la CDMX.

Descripcion de los datos numericos de la base de datos

```
print(df.describe())
```

```
          ao_hechos     ao_inicio      longitud       latitud  tempo
count  1.400873e+06  1.401331e+06  1.341941e+06  1.341941e+06    0.0
mean   2.018462e+03  2.018617e+03 -9.913714e+01  1.938701e+01    NaN
std    2.022233e+00  1.728106e+00  6.015875e-02  7.029166e-02    NaN
min    1.906000e+03  2.016000e+03 -1.002319e+02  1.909535e+01    NaN
25%    2.017000e+03  2.017000e+03 -9.917560e+01  1.933889e+01    NaN
50%    2.018000e+03  2.019000e+03 -9.914198e+01  1.938953e+01    NaN
75%    2.020000e+03  2.020000e+03 -9.909932e+01  1.943780e+01    NaN
max    2.022000e+03  2.022000e+03 -9.894686e+01  1.958333e+01    NaN
```

Después se analizará el mejor rango de años para trabajar, esto considerando el porcentaje de recuperación de información, el rango debe de comprender un alto porcentaje de datos, pero no debe ser tan grande para facilitar su estudio.

Conteo ordenado de delitos registrados por año

```
print(df['ao_hechos'].value_counts(dropna=False).head(10))
print("")
for i in range (1, 10):
    numDelitos = sum(list(df['ao_hechos'].value_counts(dropna=False).head(i)))
    porcentajeRec = numDelitos * 100 / df.shape[0]
    print(str(i) + " años: "+ str(numDelitos) + " delitos, porcentaje de recuperacion: " + str(porcentajeRec))
```

```
2018.0    254383
2019.0    244563
2017.0    227453
2021.0    218659
2020.0    203033
2016.0    195100
2022.0     28029
2015.0     16088
2014.0      4374
2013.0      2244
Name: ao_hechos, dtype: int64

1 años: 254383 delitos, porcentaje de recuperacion: 18.152956011106582
2 años: 498946 delitos, porcentaje de recuperacion: 35.60514967555845
3 años: 726399 delitos, porcentaje de recuperacion: 51.83636128794696
4 años: 945058 delitos, porcentaje de recuperacion: 67.44002666036789
5 años: 1148091 delitos, porcentaje de recuperacion: 81.9286093007291
6 años: 1343191 delitos, porcentaje de recuperacion: 95.85108728772859
7 años: 1371220 delitos, porcentaje de recuperacion: 97.85125712626068
8 años: 1387308 delitos, porcentaje de recuperacion: 98.99930851454795
9 años: 1391682 delitos, porcentaje de recuperacion: 99.31144033779314
```

De la imagen anterior se aprecian los 10 años con mayor número de delitos cometidos, así como una cuenta acumulativa de delitos por año y su respectivo porcentaje de recuperación de datos, definiendo un posible rango de estudio del año 2016 al 2021.

A continuación, se muestran varias clasificaciones de los atributos mas representativos en la base de datos.

- Clasificación de los principales delitos registrados

```
print(df['delito'].value_counts(dropna=False).head(10))
```

```
VIOLENCIA FAMILIAR                                  151656
FRAUDE                                               90375
ROBO DE OBJETOS                                      81389
AMENAZAS                                             78163
ROBO A NEGOCIO SIN VIOLENCIA                         68808
ROBO A TRANSEUNTE EN VIA PUBLICA CON VIOLENCIA       67860
ROBO DE ACCESORIOS DE AUTO                           44976
DENUNCIA DE HECHOS                                   40159
ROBO DE OBJETOS DEL INTERIOR DE UN VEHICULO          38538
ROBO DE VEHICULO DE SERVICIO PARTICULAR SIN VIOLENCIA    34423
```

- Clasificación de las principales categorías de los delitos cometidos

```
print(df['categoria_delito'].value_counts(dropna=False).head(10))
```

```
DELITO DE BAJO IMPACTO                              1106875
ROBO A TRANSEUNTE EN VÍA PÚBLICA CON Y SIN VIOLENCIA      77984
ROBO DE VEHÍCULO CON Y SIN VIOLENCIA                70313
HECHO NO DELICTIVO                                  59507
ROBO A NEGOCIO CON VIOLENCIA                        22214
ROBO A REPARTIDOR CON Y SIN VIOLENCIA               14032
ROBO A PASAJERO A BORDO DEL METRO CON Y SIN VIOLENCIA     11895
LESIONES DOLOSAS POR DISPARO DE ARMA DE FUEGO        9657
VIOLACIÓN                                            8295
HOMICIDIO DOLOSO                                     7764
```

- Clasificación de las fiscalías con mayor número de delitos registrados

```
print(df['fiscalia'].value_counts(dropna=False).head(10))
```

```
INVESTIGACIÓN EN IZTAPALAPA                         114947
INVESTIGACIÓN EN CUAUHTEMOC                         114405
INVESTIGACIÓN EN GUSTAVO A. MADERO                   78355
INVESTIGACIÓN EN BENITO JUÁREZ                       74528
INVESTIGACIÓN EN ÁLVARO OBREGÓN                      54129
INVESTIGACIÓN EN COYOACÁN                            53685
INVESTIGACIÓN EN MIGUEL HIDALGO                      53210
FISCALÍA DE INVESTIGACIÓN TERRITORIAL EN IZTAPALAPA    50114
INVESTIGACIÓN EN TLALPAN                             48244
INVESTIGACIÓN EN VENUSTIANO CARRANZA                 45444
```

- Clasificación de las colonias con mayor número de delitos registrados

```
print(df['colonia_hechos'].value_counts(dropna=False).head(10))
```

```
NaN                 60338
CENTRO              44872
DOCTORES            26923
DEL VALLE CENTRO    19729
ROMA NORTE          16523
NARVARTE            14434
BUENAVISTA          13271
MORELOS             12407
JUÁREZ              11494
AGRÍCOLA ORIENTAL   11411
```

- Clasificación de las calles con mayor número de delitos registrados

```
print(df['calle_hechos'].value_counts(dropna=False).head(10))
```

```
CALZADA DE TLALPAN             5995
EJE CENTRAL LAZARO CARDENAS    5112
SIN CALLES DEL SAP             5046
CALZADA IGNACIO ZARAGOZA       4769
AVENIDA TLAHUAC                4053
NaN                            3941
AV. INSURGENTES SUR            3403
PASEO DE LA REFORMA            3374
CALZADA DE GUADALUPE           3374
INSURGENTES SUR                3164
```

Limpieza e integración de datos (ETL)

Antes de realizar la limpieza e "integración" de la base de datos histórica hay que definir ciertas necesidades iniciales y realizar cálculos para saber cuánta información podemos perder en el proceso, esto con la intensión de conservar la mayor cantidad de datos.

De igual forma se habla sobre una integración de datos, ya que al momento de realizar el proceso ETL en código se secciono la base de datos original y después se concateno con las sub bases que obtuvimos por alcaldía.

Para el proceso ETL se realizó los siguiente:

- Atributos de la base de datos

```
print(df.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1401331 entries, 0 to 1401330
Data columns (total 19 columns):
 #   Column                Non-Null Count    Dtype  
---  ------                --------------    -----  
 0   ao_hechos             1400873 non-null  float64
 1   mes_hechos            1400873 non-null  object 
 2   fecha_hechos          1400873 non-null  object 
 3   ao_inicio             1401331 non-null  int64  
 4   mes_inicio            1401331 non-null  object 
 5   fecha_inicio          1401328 non-null  object 
 6   delito                1401331 non-null  object 
 7   fiscalia              1401329 non-null  object 
 8   agencia               1401331 non-null  object 
 9   unidad_investigacion  1401104 non-null  object 
 10  categoria_delito      1401331 non-null  object 
 11  calle_hechos          1397390 non-null  object 
 12  calle_hechos2         539997 non-null   object 
 13  colonia_hechos        1340993 non-null  object 
 14  alcaldia_hechos       1397166 non-null  object 
 15  competencia           337252 non-null   object 
 16  longitud              1341941 non-null  float64
 17  latitud               1341941 non-null  float64
 18  tempo                 0 non-null        float64
dtypes: float64(4), int64(1), object(14)
memory usage: 203.1+ MB
None
```

Una vez que se realizó la exploración de la base de datos, consultando las cifras y sabiendo que información aporta cada a tributo se ha decidido eliminar algunos de estos, para solo dejar los que nos resultaran útiles, pues por ejemplo "tempo" es una columna completamente vacía, que no aporta nada, de esta forma se decidió eliminar las siguientes columnas de la base de datos.

- Eliminación de columnas

```
dfL = df.drop([
            'tempo',
            'ao_inicio',
            'mes_inicio',
            'fecha_inicio',
            'agencia',
            'unidad_investigacion',
            'competencia'] , axis=1)
```

De tal forma que nuestra base de datos se quedaría con los siguientes atributos

- Columnas de la base de datos

```
print(df.columns)
```

```
Index(['ao_hechos', 'mes_hechos', 'fecha_hechos', 'ao_inicio', 'mes_inicio',
       'fecha_inicio', 'delito', 'fiscalia', 'agencia', 'unidad_investigacion',
       'categoria_delito', 'calle_hechos', 'calle_hechos2', 'colonia_hechos',
       'alcaldia_hechos', 'competencia', 'longitud', 'latitud', 'tempo'],
      dtype='object')
```

Nos interesa saber que día fue el delito, así como la hora en la que se suscitó, tenemos estos datos en la columna "fecha_hechos" pero debemos separarlos

- Vista e la base con las nuevas columnas

```
dfS = dfS.drop(['fecha_hechos'] , axis=1)

dfS = dfS[['dia_hechos', 'mes_hechos', 'ao_hechos', 'hora_hechos', 'delito', 'fiscalia', 'categoria_delito',
           'calle_hechos', 'calle_hechos2', 'colonia_hechos', 'alcaldia_hechos', 'longitud', 'latitud']]
print(dfS)
```

```
        dia_hechos mes_hechos  ao_hechos hora_hechos  \
0               01      Enero     2016.0    22:16:00
1               01      Enero     2016.0    20:50:00
2               02    Febrero     2016.0    00:30:00
3               01      Enero     2016.0    22:00:00
4               12  Diciembre     2015.0    12:00:00
...            ...        ...        ...         ...
1401326         02    Febrero     2022.0    06:26:00
1401327         02    Febrero     2022.0    12:00:00
1401328         05       Mayo     2021.0    09:00:00
1401329         02    Febrero     2022.0    15:50:00
1401330         02    Febrero     2022.0    12:40:00
```

En esta limpieza también se eliminarán los registros de las zonas que no son de estudio, así como los años que no nos son de utilidad, para solo quedarnos con nlos registros de 2016 a 2021.

- Información de la base de datos limpia

```
print(dfFinal.info())
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 714383 entries, 1 to 1400430
Data columns (total 13 columns):
 #   Column            Non-Null Count   Dtype  
---  ------            --------------   -----  
 0   dia_hechos        714383 non-null  object 
 1   mes_hechos        714383 non-null  object 
 2   ao_hechos         714383 non-null  float64
 3   hora_hechos       714383 non-null  object 
 4   delito            714383 non-null  object 
 5   fiscalia          714383 non-null  object 
 6   categoria_delito  714383 non-null  object 
 7   calle_hechos      713254 non-null  object 
 8   calle_hechos2     293996 non-null  object 
 9   colonia_hechos    713800 non-null  object 
 10  alcaldia_hechos   714383 non-null  object 
 11  longitud          714383 non-null  float64
 12  latitud           714383 non-null  float64
dtypes: float64(3), object(10)
memory usage: 76.3+ MB
None
```

Análisis exploratorio de datos (EDA)