Metodología del análisis de datos

Para el proyecto:
Guía preventiva ante el delito que
atenta contra la libertad personal
en Ciudad de México

Abraham Jain Jiménez
Facultad de Ciencias.
Universidad Nacional Autónoma de México.

Mayo 2025

1. Introducción

Se presenta la metodología propuesta para el análisis de datos asociado al proyecto ***Guía preventiva ante el delito hacia la libertad personal en Ciudad de México***.

El objetivo del análisis es extraer información de los datasets y proporcionar las bases para el conjunto de medidas preventivas ante el delito objetivo.

2. Elección de las bases de datos

Se trabajará principalmente con dos datasets, extraídos del Portal de Datos Abiertos del Gobierno de la Ciudad de México, y de los Datos Abiertos de Incidencia Delictiva del Gobierno de México. Además, también se usarán dos datasets más como apoyo, obtenidos de la última fuente mencionada.

Datasets principales

Nombrados a convención para el análisis correspondiente a este proyecto, tenemos:

- Dataframe 2: Carpetas de Investigación (acumulado 2016-2024)
- Dataframe 3: Cifras de Víctimas del Fuero Común, 2015 - mayo 2025

Datasets de apoyo

- Dataframe 1: Cifras de Incidencia Delictiva Municipal, 2015 - mayo 2025
- Dataframe extra: Cifras de Incidencia Delictiva Estatal, 1997 - diciembre 2017

Relevancia de cada dataset

- Dataframe 1: Nos introduce a las bases de datos de la privación a la libertad personal. Su manipulación tiene como objetivo extraer información sobre las modalidades en que se presenta el delito hacia la libertad personal, con el fin de saber cómo filtrar el Dataframe 2, el cuál es la base de datos central para el proyecto.
- Dataframe 2: Es el corazón del análisis de datos en este proyecto. Se limpia con base en la información recopilada para el Dataframe 1. Contiene información fundamental sobre las denuncias registradas: fechas, horas, modalidad del delito, colonia, alcaldía del hecho, otros delitos asociados, etc. Abarca desde 2016 hasta enero de 2025.
- Dataframe 3: Contiene información de la última década sobre registros del delito mencionado. Complementa el Dataframe 2 con información sobre las víctimas: su rango de edad y el sexo.

- Dataframe extra: Contiene registros de carpetas de investigación desde 1997 hasta 2017. Sirve de apoyo para el Dataframe 2 en cuestión de desarrollar hipótesis sobre los patrones temporales del delito de interés.

3. Metodología

A gran escala, las bases de la guía de recomendaciones para prevenir delitos hacia la libertad personal toma como base dos pasos en el análisis de datos, la limpieza de los datasets (filtrando los datos de acuerdo a los objetivos del proyecto) y el análisis exploratorio para detectar patrones, grupos, tendencias, etc, a través de distintas variables como el tiempo, categorías del delito, etc.

Todo lo descrito anteriormente se llevará a cabo usando el lenguaje Python, trabajando los datasets como dataframes de la biblioteca *Pandas*. Las visualizaciones de datos se harán con las bibliotecas *Matplotlib*, *Seaborn* y *Plotly* Express. Así mismo, otra biblioteca de utilidad será *NumPy*.

3.1. Limpieza de los datasets

Conforme a los objetivos del proyecto, se van a filtrar los dataframes de acuerdo a los delitos que afectan el bien jurídico de la libertad personal. Se aborda la imputación, eliminación y/o el mapeo de valores faltantes, atípicos, ambiguos, o cualquier otro grupo que pueda perturbar el análisis. Así mismo, se trabaja la manipulación del formato de los datos y de los dataframes.

Limpieza del Dataframe 1:

Este proceso es importante, pues el Dataframe 2 de carpetas de investigación sobrepasa el millón de registros y es complicado filtrar para tipos de delitos hacia la libertad personal sin conocer bien sus clasificaciones. El proceso es el siguiente:

1. Se filtra el dataframe para la entidad Ciudad de México.

2. Se filtra para el bien jurídico afectado correspondiente a la libertad personal.

3. Se extrae la información sobre los tipos (o clasificaciones) legales del delito hacia la libertad personal.

Limpieza del Dataframe 2:

Corresponde a los registros de carpetas de investigación abiertas. Su limpieza es fundamental pues es el dataset más completo que se encontró para cumplir con los objetivos del proyecto. El proceso es el siguiente:

1. Con base a la información extraída del Dataframe 1, se filtra para delitos que contengan esta misma información.

2. Se eliminan columnas que no aporten información para el objetivo con el fin de hacer el archivo CSV más ligero.

3. Se aborda el manejo de valores faltantes. Los valores NaN más relevantes en cuanto a repercusiones en el análisis, se asocian a la columna de la alcaldía donde sucedió el delito, la cual es una variable fundamental, por ello es necesario imputar los valores faltantes. Esta imputación se realizará mediante la información de las columnas de agencia y fiscalía que atendieron el caso, las cuales tienen información próxima a la alcaldía del hecho.

4. Se convierten fechas y horas a formato *datetime64* y se ordenan los datos con base a estas.

Limpieza del Dataframe 3:

Este proceso es análogo a la limpieza del Dataframe 1, pero se necesita abarcar más detalles, expuestos a continuación:

- Se cambia el formato del dataframe de ancho a largo sobre los meses y el número de registros del delito, para una correcta visualización con la biblioteca Seaborn.

- Se aborda el manejo de valores faltantes. Estos se presentan en la columna de número de registros. Su interpretación es la ausencia de registros y se imputan como cero.

- Se filtra para registros distintos de cero.

- Se ordena por año de reporte y se filtra por este mismo para coincidir con el Dataframe 2.

Limpieza del Dataframe extra:
Este proceso es completamente similar a la limpieza del Dataframe 1.

3.2. Análisis de datos para la guía preventiva

Un análisis de datos se llevará a cabo profundamente en los Dataframes 2 y 3, mientras que en el Dataframe extra solo se analizará el aspecto en cuanto a patrones temporales. En esta parte consideramos también los outliers.

Para cada Dataframe, el procedimiento consiste en abordar los siguientes ocho puntos:

Análisis del Dataframe 1:

1. Distribución de frecuencia de cada clasificación del delito hacia la libertad personal de acuerdo a la fecha de apertura de las carpetas de investigación entre enero 2016 y enero 2025.
2. Análisis del número de carpetas por fechas del hecho.
3. Distribución de la frecuencia del delito hacia la libertad personal a lo largo del tiempo para detectar patrones mediante un algoritmo de clustering (K-means).
4. Análisis del delito hacia la libertad personal y sus clasificaciones por alcaldía.
5. Análisis de ocurrencias del delito en CDMX y sus alcaldías en intervalos temporales de una hora.
6. Análisis del delito hacia la libertad personal por colonia.
7. Análisis sobre involucrados en el delito hacia la libertad personal por alcaldía. Esta información puede servir para detectar patrones en el *modus operandi*,

Análisis del Dataframe 3:

8. Distribución del delito hacia la libertad personal por sexo y edad de las víctimas.

Análisis del Dataframe extra:

2.1. Implementar el mismo algoritmo de clustering del punto 2. para comprobar o rechazar hipótesis de los patrones a lo largo del tiempo.

3.3. Información adicional

Se incluirá, como apéndice, información que no es necesariamente esencial para la parte preventiva al delito.

A. Análisis de los tipos de delito hacia la libertad personal y su asociación explícita con otros delitos de acuerdo a este dataset.

B. Análisis sobre detenciones por alcaldía.

C. Análisis de la relación entre la fecha del hecho y la fecha de inicio de la carpeta de investigación.

D. Análisis de los días semanales que corresponden al día del acontecimiento del delito y al día de la apertura de la carpeta para ver algunos otros patrones temporales.