Introducción

México ocupó el lugar 135 de 163 países en el **GLOBAL PEACE INDEX** 2025, situándose entre las naciones con mayores niveles de violencia e inseguridad. La Ciudad de México (CDMX), con una población de 9,209,944 habitantes según el Censo de Población y Vivienda 2020, constituyó la segunda entidad federativa más poblada del país, después del Estado de México. Si bien la capital no se ubicó entre las regiones más violentas, su relevancia demográfica, económica y simbólica la convirtió en un espacio clave para el análisis del fenómeno delictivo urbano.

De acuerdo con Martin A. Andresen, en CRIMINOLOGÍA AMBIENTAL (**ENVIRONMENTAL CRIMINOLOGY**), el fenómeno delictivo se sustentó en tres principios fundamentales. En primer lugar, el comportamiento delictivo se encontró condicionado por el entorno en el que ocurrió (“el lugar importa”). En segundo lugar, la distribución espacial y temporal del delito no fue aleatoria, sino que tendió a concentrarse en determinados puntos y momentos. En tercer lugar, la modificación de estas condiciones, así como la concentración estratégica de recursos (por ejemplo, presencia policial, intervenciones de diseño urbano o programas sociales) en zonas de alta incidencia (**HOT SPOTS**), pudo contribuir de manera significativa a la reducción de la inseguridad.

El objetivo de la investigación se planteó como el desarrollo de un modelo predictivo del crimen en la Ciudad de México. Para ello, en primera instancia se llevó a cabo un análisis exploratorio orientado a identificar la relación entre factores socioeconómicos y la incidencia delictiva, con énfasis en los delitos cometidos contra vehículos particulares. La elección de este tipo de delito respondió a que, de acuerdo con datos del Instituto Nacional de Estadística y Geografía (INEGI, 2024), durante 2023 se registraron 7,160 delitos de robo total o parcial de vehículo por cada 100 mil habitantes, lo que lo ubicó como el tercer delito más frecuente en la capital.

El estudio se sustentó en una base de datos oficial proporcionada por la FISCALÍA GENERAL DE JUSTICIA DE LA CIUDAD DE MÉXICO, la cual integró información detallada sobre los delitos registrados, incluyendo su localización geográfica, temporalidad y modalidad. Esta fuente constituyó el insumo principal para el análisis espacial y la construcción del modelo predictivo propuesto.

Como punto de partida, se realizó un análisis exploratorio para identificar la relación entre factores socioeconómicos y la incidencia delictiva, con énfasis en los delitos cometidos contra vehículos particulares. De este análisis se desprendió que la tasa de delitos relacionados con vehículos particulares por cada 10,000 habitantes fue significativamente mayor en alcaldías con un IDS superior al promedio de la Ciudad de México durante los días laborales, entre las 6:00 y las 18:00 horas. Este hallazgo sugirió que el nivel de desarrollo social se asoció con formas específicas de victimización, pero no permitió, por sí solo, caracterizar la diversidad de perfiles urbanos existentes. Por ello, se consideró necesario plantear una segunda hipótesis basada en técnicas de **CLUSTERING**, con el propósito de identificar patrones y relaciones ocultas en los datos que sirvieron de insumo para la fase final de predicción, incorporando características urbanas como alumbrado público y cámaras **CCTV**.

H2 Hipótesis de Clustering

Las diferencias en la carga delictiva, la intensidad de la actividad económica y la disponibilidad de alumbrado público y cámaras **CCTV** en proporción a la población generaron perfiles urbanos distinguibles, que pudieron ser segmentados mediante **CLUSTERING** en grupos de colonias funcionalmente similares.

Objetivo General

Analizar y segmentar el territorio de la Ciudad de México mediante técnicas de **CLUSTERING** aplicadas a indicadores de carga delictiva, actividad económica e infraestructura urbana de seguridad (ALUMBRADO PÚBLICO y cámaras **CCTV** por habitante), con el fin de identificar perfiles urbanos funcionalmente similares asociados a distintos niveles de riesgo delictivo.

Objetivo Particular

Depurar la base de datos integrada que combinó información socioeconómica, registros delictivos georreferenciados e indicadores de infraestructura de seguridad (ALUMBRADO PÚBLICO y cámaras **CCTV**) a nivel de alcaldía/colonia, normalizando las variables en términos per cápita para hacerlas comparables.

Aplicar y evaluar modelos de **CLUSTERING** (**K-MEANS**) sobre las variables seleccionadas para identificar grupos de UNIDADES ECONÓMICAS con patrones similares de carga delictiva.

Antecedentes

En investigaciones previas sobre análisis y predicción del delito en contextos urbanos, se emplearon técnicas de aprendizaje automático y métodos de **CLUSTERING** combinados con información espacial, socioeconómica y de infraestructura de seguridad, como se describió en los siguientes estudios.

El estudio [1] “**CRIME PREDICTION AND MONITORING IN PORTO, PORTUGAL, USING MACHINE LEARNING, SPATIAL AND TEXT ANALYTICS**” aplicó estos enfoques combinados. Utilizó datos confidenciales de la Policía de Seguridad Pública de Oporto (2016–2018), con más de 42,000 registros georreferenciados a nivel de calle. Complementariamente, integró datos censales, urbanos y de uso de suelo, así como la ubicación de estaciones de policía y cámaras de vigilancia. Además, recopiló tweets en un radio de 1 km alrededor de cada incidente delictivo, empleando más de cincuenta términos relacionados con el crimen en inglés y portugués, para realizar modelado temático y análisis de sentimientos. Esta integración de datos espaciales, socioeconómicos y textuales permitió fortalecer el análisis predictivo del crimen urbano. Primero, se empleó regresión **LASSO** para la selección de variables relevantes. Este método redujo el sobreajuste y mejoró la eficiencia del modelo al penalizar variables menos significativas. Los resultados mostraron que una mayor proporción de población con bajo nivel educativo y de jóvenes se asoció positivamente con el crimen, mientras que la educación universitaria, la presencia de

CCTV y las zonas predominantemente residenciales se relacionaron con menores tasas delictivas. Posteriormente, se desarrollaron modelos de clasificación (**LOGISTIC REGRESSION, SVM, DECISION TREE y RANDOM FOREST**) para predecir la presencia o ausencia de delitos.

La regresión logística con penalización L1 identificó variables como edificaciones antiguas y presencia de cámaras de seguridad como factores negativos (menor probabilidad de crimen), mientras que el bajo nivel educativo y los edificios altos se relacionaron positivamente con la ocurrencia de delitos.

El modelo de **RANDOM FOREST** mostró el mejor desempeño predictivo, con precisión = 0.79, recall = 0.99 y F1-score = 0.89, destacando como variables clave la estructura de las edificaciones, el nivel educativo básico y la proporción de población masculina.

Por otro lado, la investigación [3] "**USING BIG DATA ANALYTICS TO IDENTIFY TRENDS AND GROUP CRIMES THROUGH CLUSTERING**" mencionó que el incremento en la inseguridad de los ciudadanos en Perú se debió a factores económicos y de desigualdad social, pobreza, diseño urbano, consumo de alcohol y drogas, entre otros. En su estudio únicamente usaron la técnica de **CLUSTERING KNN** a través de la distancia euclidiana con datos históricos geoespaciales.

Finalmente, Kiani et al., en "**ANALYSIS AND PREDICTION OF CRIMES BY CLUSTERING AND CLASSIFICATION**", aplicaron el algoritmo de **K-MEANS** para agrupar delitos según patrones anuales usando datos de Inglaterra y Gales (1990-2011). Este **CLUSTERING** permitió identificar grupos con comportamientos delictivos similares y detectar estructuras recurrentes en el tiempo. Utilizaron variables basadas en la frecuencia y tipo de delitos para entrenar modelos de clasificación que predijeron tendencias futuras.

Justificación

La elección de técnicas de **CLUSTERING** y de las variables empleadas respondió al carácter contextual y no aleatorio del delito planteado por la criminología ambiental. Dado que el análisis exploratorio mostró que las alcaldías con IDS superior al promedio presentaron mayores tasas de delitos contra vehículos particulares, se consideró que la combinación de carga delictiva (delitos por 10,000 hab.), nivel socioeconómico (IDS) y disponibilidad de infraestructura urbana de seguridad (ALUMBRADO PÚBLICO y cámaras **CCTV** por habitante) pudo dar lugar a perfiles urbanos diferenciados. Bajo la hipótesis H2, estas diferencias configuraron entornos funcionalmente distintos que influyeron en las oportunidades para delinquir. El **CLUSTERING** permitió agrupar colonias o alcaldías que comparten patrones similares en estas variables, identificando grupos territoriales con estructuras de riesgo comparables y generando insumos más robustos para la etapa posterior de predicción del crimen.

Metodología

Clustering Densidad de Cámaras por Alcaldía

Para analizar la relación entre infraestructura de videovigilancia y criminalidad a nivel territorial, se llevó a cabo un ejercicio de **CLUSTERING** basado en la densidad de cámaras por alcaldía.

Se utilizaron las bases de datos de la Fiscalía de la Ciudad de México y del programa [7] MI CALLE, que contenía el registro de las cámaras instaladas en distintas colonias de la CDMX. Asimismo, se empleó una base de datos elaborada con la población total por alcaldía y su superficie en km². A partir de estas bases de datos se generaron las columnas de delitos por cada 10,000 habitantes y cámaras por cada 10,000 habitantes (PORTAL DE DATOS ABIERTOS DE LA CIUDAD DE MÉXICO, s. f.).

Las variables creadas fueron camaras_por_10k (número de cámaras de videovigilancia por cada 10,000 habitantes en cada alcaldía), IDS (Índice de Desarrollo Social de cada alcaldía) y Delitos_por_10k_hab (tasa de delitos por cada 10,000 habitantes).

En cuanto al método de **CLUSTERING**, se utilizó el algoritmo de **K-MEANS**, un método de agrupamiento no jerárquico que permitió clasificar observaciones en grupos (**CLUSTERS**) con características similares. **K-MEANS** resultó adecuado en este caso porque funcionó bien para identificar patrones generales en función de varias variables cuantitativas. Igualmente, permitió detectar grupos de alcaldías con combinaciones similares de nivel de desarrollo, densidad de cámaras y tasa de delitos, lo que facilitó la comparación entre contextos heterogéneos. Además, generó centroides que resumieron el perfil promedio de cada **CLUSTER**, útiles para interpretar posibles estrategias diferenciadas de política pública.

El procedimiento consistió, primero, en la construcción de las variables de interés: a partir del total de cámaras **CCTV** por alcaldía y su población se calculó la densidad de cámaras por cada 10,000 habitantes y, de manera análoga, se obtuvo la tasa de delitos por cada 10,000 habitantes. Posteriormente, se integraron la base de CÁMARAS y la base de DELITOS, incorporando también el IDS de cada alcaldía, de modo que cada registro concentró simultáneamente nivel de desarrollo, densidad de cámaras y carga delictiva. Dado que estas variables se encontraban en escalas distintas, se aplicó un escalamiento estándar mediante **STANDARDSCALER** para equilibrar su influencia en el algoritmo de agrupamiento. A continuación, se estimó el número óptimo de **CLUSTERS** (K) utilizando el método del codo, que evaluó la reducción en la suma de errores cuadráticos **INTRA-CLUSTER** para distintos valores de K, y se complementó este criterio con el análisis del coeficiente de silueta, que midió qué tan bien se encontró cada observación asignada a su grupo. Una vez definido el valor de K, se entrenó el modelo **K-MEANS** y se obtuvieron los centros de cada **CLUSTER** mediante el atributo cluster_centers_, lo que permitió caracterizar los grupos de alcaldías según su combinación de IDS, densidad de cámaras y tasas de delitos.

Clustering Alumbrado Público

Preparación y limpieza de datos

El análisis partió de un conjunto de aproximadamente 200,000 registros correspondientes a carpetas de investigación de la Fiscalía General de Justicia de la Ciudad de México para el periodo 2016–2024. En una primera etapa se aplicó un filtro para seleccionar únicamente los delitos vinculados con robo vehicular, incluyendo robo de vehículo particular con y sin violencia, robo de motocicleta, robo de vehículos de pedales, robo de accesorios y robo de objetos del interior de vehículos.

Con el fin de garantizar la consistencia de los identificadores territoriales, se normalizaron los nombres de colonias mediante la eliminación de acentos, la conversión a mayúsculas y la remoción de caracteres especiales. Posteriormente, se integraron cinco bases de datos externas: (i) UNIDADES ECONÓMICAS POR COLONIA, obtenidas del Directorio Estadístico Nacional de Unidades Económicas (DENUE); (ii) NÚMERO DE HOGARES POR COLONIA; (iii) REGISTRO DE LUMINARIAS DE ALUMBRADO PÚBLICO; (iv) ÍNDICE DE MARGINACIÓN URBANA DEL CONAPO 2020; y (v) un identificador espacial por unidad territorial.

El proceso de vinculación (**MERGE**) entre bases se realizó, en primera instancia, mediante coincidencia exacta de nombres de colonia y claves territoriales. Para los casos en los que existieron discrepancias en la nomenclatura, se emplearon algoritmos de **FUZZY** **MATCHING** con un umbral de similitud del 85 %, lo que permitió resolver diferencias menores de escritura y reducir la pérdida de registros por errores tipográficos. La variable de población proveniente del Censo 2010 se descartó del análisis por considerarse desactualizada respecto al periodo de estudio.

Construcción de variables y transformaciones

Una vez integradas las distintas fuentes, los delitos individuales se agregaron a nivel de colonia, obteniéndose el conteo total de incidentes por unidad territorial. Con el objetivo de eliminar el sesgo derivado de diferencias en tamaño y permitir comparaciones válidas entre colonias de distinta extensión, se construyeron indicadores de densidad expresados por kilómetro cuadrado: delitos por km², unidades económicas por km², luminarias por km² y hogares por km².

El análisis descriptivo inicial mostró una alta asimetría en la distribución de varias de estas variables, con coeficientes de sesgo superiores a 3 y curtosis mayor a 10, lo que evidenció la presencia de valores extremos. Para mitigar este efecto y estabilizar la varianza, se aplicó una transformación logarítmica mediante la función log1p, que resultó adecuada dada la presencia de valores cercanos a cero. Posteriormente, se examinó la correlación entre las variables transformadas con el propósito de evitar problemas de multicolinealidad en la etapa de **CLUSTERING**.

Asimismo, se evaluó la pertinencia de incluir la variable relativa a cámaras de vigilancia del sistema C5. Dicho indicador se excluyó del análisis final por considerarse endógeno, en tanto su distribución espacial respondió de manera reactiva a la incidencia delictiva

preexistente, lo que habría introducido un sesgo en la identificación de patrones estructurales.

Segmentación mediante K-means

Para la segmentación territorial se implementó el algoritmo **K-MEANS** sobre las variables seleccionadas, previamente normalizadas mediante **STANDARDSCALER** con el fin de homogeneizar su escala y evitar que aquellas con valores numéricamente mayores dominaran el proceso de agrupamiento.

La selección del número óptimo de **CLÚSTERES** se realizó a través de una evaluación sistemática de diferentes valores de k entre 2 y 7. Para cada configuración se calcularon métricas de validación interna, utilizando el coeficiente de **SILHOUETTE** como indicador de cohesión y separación entre grupos, y el índice de **DAVIES–BOULDIN** como medida de la dispersión relativa de los **CLÚSTERES**. El modelo final con k = 3 obtuvo un coeficiente de **SILHOUETTE** de 0.241, valor que se consideró adecuado para describir una estructura de agrupamiento interpretable, dado el carácter heterogéneo del contexto urbano analizado.

Cada **CLÚSTER** se caracterizó mediante el cálculo de medidas resumen (promedios, medianas y desviaciones estándar) de las variables originales y transformadas, identificando además las colonias más representativas de cada grupo a partir de su proximidad euclidiana al centroide correspondiente. Finalmente, se llevó a cabo una validación cualitativa de los resultados, examinando la coherencia geográfica y urbanística de las colonias asignadas a cada segmento, con el fin de verificar que los patrones detectados por el algoritmo fueran consistentes con el conocimiento empírico del territorio y con la literatura previa sobre criminalidad urbana.

La base de datos se encontró almacenada en Drive, por lo cual primero se instaló y configuró la ruta de acceso para poder trabajar con los archivos. El primer archivo correspondió a la información de marginación, que contuvo el IDS por alcaldía, la superficie en kilómetros cuadrados y la población; todos estos datos fueron obtenidos del Gobierno de la Ciudad de México. El IDS se definió como una medida agregada que resumió ocho indicadores: vivienda, educación, bienes durables, energía, adecuación sanitaria, telecomunicaciones, salud y seguridad social de las personas.

El cálculo del IDS se basó en el método de Necesidades Básicas Insatisfechas (NBI), el cual formó parte del Método de Medición Integrada de la Pobreza (MMIP), que experimentó cambios metodológicos en años recientes. Ante esta situación, Evalúa dio a conocer las nuevas cifras del IDS 2020 con base en el Censo de Población y Vivienda 2020.

El segundo archivo correspondió a los datos de la Fiscalía de la Ciudad de México relativos a los delitos. La Dirección General de Política y Estadística Criminal, área adscrita directamente a la oficina de la Fiscal, tuvo la responsabilidad estratégica de generar la información estadística delictiva que permitió elaborar indicadores para la instrumentación de políticas criminales. El informe estadístico contempló la situación delictiva de la ciudad y detalló los delitos cometidos por lugar de los hechos y modalidad.

Por último, los datos de cámaras CCTV correspondieron al programa *Mi Calle*, el cual consistió en la instalación de tótems en domicilios de 333 colonias prioritarias de la Ciudad de México, los cuales incluyeron cámaras de videovigilancia, alertas sonoras y visuales, y botones de auxilio.

Dado que el objetivo del estudio es predecir delitos relacionados con el robo a vehículos, se acotó el universo delictivo a un conjunto específico de categorías proporcionadas por la Fiscalía. Para ello, se definió una lista de prefijos (DELITOS_PREFIJOS) que agrupa tanto robos violentos como no violentos vinculados a vehículos, incluyendo: robo de vehículo de servicio particular, robo de motocicleta, robo de vehículo de pedales, robo de accesorios de auto y robo de objetos del interior de un vehículo. A partir de estos prefijos, se filtran los registros de la base de datos para construir un subconjunto de incidentes focalizado exclusivamente en el robo a vehículos y sus modalidades asociadas, que será la base para el análisis y los modelos de predicción posteriores.

El siguiente bloque del código se enfocaba en transformar y estandarizar los datos de la variable “total de delitos”, esto con el fin de darnos una análisis comparativo entre las diferentes alcaldías. Este paso era crucial puesto que las alcaldías de la ciudad capitalina no eran homogéneas entre ellas y sin una estandarización previa estas no podrían ser comparables. Primero, se creaba una nueva tabla de referencia, “pob_ids_ref”, a partir de los datos de “df_hipotesis”, eliminando duplicados de alcaldia_hecho_N y celdas vacías en “Población” e “IDS”. Después se conglomeraban todos los delitos sucedidos por cada alcaldía y se juntaba, con la función merge, entre el conteo de delitos y la tabla de referencia permitía asociar correctamente el total de delitos con su población e IDS correspondientes. Por último, se calculaba la tasa estandarizada de delitos por cada 10,000 habitantes, un indicador clave que normalizaba el conteo de delitos (total_delitos) por el tamaño poblacional de la alcaldía, facilitando la comparación de la incidencia delictiva entre diferentes jurisdicciones sin que afectara de sus diferentes tamaños poblacionales.

Previo a la incorporación de la variable de tasa de cámaras por 10,000 habitantes, se realizaba una inspección exploratoria del DataFrame de origen para identificar y validar las columnas de interés para el análisis. Después, se construía una tabla de referencia geoespacial al eliminar registros duplicados a nivel de alcaldía, lo que permitía generar mapeos que asociaban de forma unívoca a cada colonia con su alcaldía, latitud y longitud. Estos datos geográficos se integraban al DataFrame de cámaras (camaras_df). Para preservar la integridad y eficiencia de los datos, se creaba una copia de trabajo que contenía exclusivamente las columnas objetivo (alcaldía, totalinsta, latitud_rt, longitud_rt). Luego, esta tabla era enriquecida mediante la unión lógica (merge) con las nuevas columnas calculadas de tasa de cámaras por 10,000 habitantes y los indicadores preexistentes. Por último, las variables seleccionadas (camaras_por_10k, IDS, Delitos_por_10k_hab) eran estandarizadas utilizando el método StandardScaler para preparar el conjunto de datos para el modelado.

El análisis de correlación mostró como es que Delitos_por_10K_hab tiene una relación positiva muy fuerte con el IDS, lo que determina que en las alcaldías con un IDS alto tiene una tendencia a presentar una mayor frecuencia delictiva por habitante. Esto apoya nuestra

hipótesis inicial del proyecto, la cual indica como las concentraciones de delitos tienen una probabilidad más alta de ubicarse en zonas con una mayor actividad económica, mejores sistemas de transporte o una mayor concentración de bienes atractivos para el delito. En contraste, las demás variables no muestran una correlación lineal con el número de delitos por 10,000 habitantes. Estas variables fungen con un papel diferente en el clustering, aportando dimensionalidad adicional que ayuda a separar agrupaciones de alcaldías según su cobertura relativa de vigilancia y su contexto delictivo. Asimismo, la inclusión de estas variables no lineales ofrecen un contexto de la vigilancia en cada alcaldía, permitiendo comparar entre cada una cómo es que la incidencia delictiva cambia. Esto ayuda a identificar grupos de alcaldías con contextos de riesgo distintos y facilita la comparación entre lugares con similares niveles delictivos pero diferentes estrategias de monitoreo. Por ende, aunque no sea un predictor directo del delito por habitante, camaras_por_10k contribuye a entender la estructura del fenómeno delictivo y a detectar patrones institucionales y territoriales relevantes para la toma de estrategias basadas en datos.

#Método del codo
Se usó el método del codo para poder definir la cantidad óptima de K a usar en el clustering. De esta manera, disminuimos la suma de los errores dentro de los clústeres mientras K aumentaba. Al inicio, al agregar más clústeres, se mejoró la agrupación, pero la mejora se volvió mínima al seguir aumentando los puntos. Al llegar a ese punto donde la gráfica no cambiaba y no se vieron mejoras, decidimos dejar de modificarlo, ya que solo se complicaría el modelo.

Este bloque creó una lista llamada **INERTIA** para guardar los valores de inercia, que indicaban qué tan bien se agruparon los datos con cada número de clústeres posibles. Se definió un rango de K de 1 a 10, que representaba la cantidad de grupos que se iban a probar. Para cada K, se ajustó un modelo de **K-MEANS** con 10 intentos y se calculó la inercia, que es la suma de las distancias al cuadrado entre cada punto y su centroide.

#Slihouette
Se usó **SILHOUETTE** para confirmar la elección del número de clústeres. Con este método se vio qué tan bien estaba ubicado cada punto dentro de su grupo, midiendo la distancia promedio con otros puntos del mismo clúster y comparándolos con puntos de otros grupos. Al salir un valor alto mostró que el punto estaba bien asignado, mientras que un valor bajo o negativo indicó que podría cambiar a otro clúster. Así, con **SILHOUETTE** pudimos confirmar que el número de clústeres elegido en **K-MEANS** era el indicado, evitando errores.

En el bloque se eligió que k=3 fue el número ideal de clústeres usando el método del codo. Después se creó y entrenó el modelo **K-MEANS** con esos 3 grupos para saber a qué clúster pertenecía cada punto. Para confirmar qué tan bueno fue el agrupamiento se calcularon los valores **SILHOUETTE** para cada punto individualmente. Finalmente, se hizo un gráfico de **SILHOUETTE** donde se visualizó la cohesión y separación de los clústeres, con cada clúster representado en un bloque de colores y una línea roja que mostró el valor global, ayudando a entender la calidad del agrupamiento.

Para interpretar los resultados del agrupamiento, en primer lugar se recuperaron los centroides generados por el modelo K-Means a partir de los datos escalados con StandardScaler. Estos centroides representan el valor promedio de cada clúster en la escala estandarizada. Sin embargo, esa escala no es directamente interpretable en términos sustantivos, por lo que fue necesario desescalar dichos valores mediante la función inverse_transform del mismo escalador. De esta manera, se obtuvieron los centroides en la escala original de las variables: cámaras por cada 10 000 habitantes, IDS y delitos por cada 10 000 habitantes. Con estos valores se construyó un DataFrame (df_centroids), lo que facilitó tanto su análisis numérico como su posterior representación gráfica.

A continuación, se elaboró un diagrama de dispersión donde cada punto corresponde a una alcaldía, utilizando en el eje horizontal la densidad de cámaras por 10 000 habitantes y en el eje vertical la tasa de delitos por 10 000 habitantes. Los puntos se colorearon según el clúster asignado por el modelo, lo que permite visualizar cómo se agrupan territorialmente las alcaldías en función de estas dos dimensiones. Sobre este mismo gráfico se añadieron los centroides obtenidos en la escala original, representados con un marcador distintivo y de mayor tamaño, y se etiquetaron como C0, C1, C2, etc. Esta superposición de centroides sobre los datos observados permite identificar el perfil promedio de cada clúster y comparar, de manera intuitiva, las diferencias entre grupos en términos de infraestructura de cámaras y carga delictiva. Finalmente, se incorporaron títulos, etiquetas de ejes, leyenda y rejilla para mejorar la legibilidad y facilitar la interpretación del patrón de agrupamiento resultante.

Además del análisis en dos dimensiones, se construyó una visualización tridimensional utilizando Plotly Express con el objetivo de representar simultáneamente las tres variables empleadas en el clustering: densidad de cámaras por cada 10 000 habitantes, tasa de delitos por cada 10 000 habitantes e Índice de Desarrollo Social (IDS). Para ello, se generó un gráfico de dispersión 3D donde cada punto corresponde a una alcaldía, ubicando en el eje X las cámaras por 10 000 habitantes, en el eje Y los delitos por 10 000 habitantes y en el eje Z el IDS. Los puntos se colorearon de acuerdo con el clúster asignado por el modelo K-Means, lo que permite observar cómo se distribuyen los grupos en el espacio definido por estas tres dimensiones.

Sobre esta visualización se añadieron los centroides de los clústeres, utilizando la información contenida en df_centroids. Estos centroides se incorporaron como marcadores adicionales en el gráfico 3D, representados con un símbolo en forma de “x” de color negro y etiquetados como C0, C1, C2, etc., para facilitar su identificación. Finalmente, se actualizaron los títulos de los ejes y del gráfico principal, de modo que cada dimensión quedara claramente señalada (cámaras por 10 000 habitantes, delitos por 10 000 habitantes e IDS). Esta representación tridimensional interactiva permite explorar de manera más intuitiva cómo se diferencian los clústeres cuando se consideran de forma conjunta la infraestructura de vigilancia, la carga delictiva y el nivel de desarrollo social.

Se evaluaron diferentes valores de K para el modelo K-Means, comparando en particular las configuraciones de 3 y 5 clústeres. Aunque ambos casos presentan coeficientes de silhouette moderados, la solución con K = 3 obtuvo un valor ligeramente superior (0.325 frente a 0.321 para K = 5) y generó grupos más compactos y estables. Además, con tres clústeres se obtienen perfiles territoriales claramente diferenciados: (i) alcaldías con carga delictiva baja a media y densidad intermedia de cámaras, (ii) alcaldías con alta carga

delictiva asociada a mayores niveles de desarrollo y actividad urbana, y (iii) alcaldías con mayor densidad de cámaras por habitante y niveles delictivos relativamente menores. En cambio, la solución con K = 5 fragmenta en exceso el espacio muestral, produciendo clústeres con muy pocas alcaldías y sin mejora sustantiva en la métrica de silhouette, lo que dificulta su interpretación. Por estas razones, se seleccionó K = 3 como número óptimo de clústeres para el análisis.

El análisis de K-Means con k = 3, el cual fue definido mediante el método del codo y validado con el silhouette score, revela tres patrones claros entre las alcaldías según su relación entre cámaras instaladas por cada 10,000 habitantes, delitos registrados e Índice de Desarrollo Social (IDS). La principal observación de estos patrones es la existencia de una tendencia positiva entre el aumento de IDS y cámaras por 10,000 habitantes, la frecuencia de delitos también aumenta. Analizando cada cluster, determinamos que el Clúster 1 destaca por agrupar alcaldías con alta incidencia delictiva, un IDS elevado y pocas cámaras, lo que sugiere que incluso zonas con mejores condiciones socioeconómicas mantienen una actividad delictiva intensa, posiblemente asociada a concentración de actividad económica, movilidad y una posible insuficiencia de vigilancia en áreas críticas como Cuauhtémoc o Miguel Hidalgo. En cambio, el cluster 2 tiene un control más efectivo, puesto que presenta muchas más cámaras, un IDS medio pero tiene menos delitos que el cluster 1. Finalmente, el Clúster 0 agrupa alcaldías con delincuencia baja a moderada, menos cámaras y valores de IDS también moderados, formando un bloque de condiciones equilibradas sin extremos en ninguna de las tres variables. En conjunto, en la ciudad capitalina hay alcaldías con un buen nivel de condición pero con mucha actividad delictiva, lo cual pone en evidencia la necesidad de nuevas estrategias para disminuir la recurrencia de estos delitos.

El análisis de K-Means con K = 5, demuestra que los grupos se vuelven más finos, pero menos robustos. Milpa Alta queda sola en un clúster (k=5, cluster 3)Benito Juárez y Cuauhtémoc quedan en un mismo grupo (alta delincuencia), mientras otras con delitos altos se reparten en otros clústeres. Varios clústeres tienen muy pocas observaciones (algunas 2–3 alcaldías), lo que reduce la estabilidad del análisis. El silhouette apenas mejora (de hecho baja ligeramente), lo que indica que esa fragmentación adicional no aporta mucha calidad al agrupamiento. Para 16 alcaldías, 5 clústeres implica, en promedio, 3.2 alcaldías por clúster; estadísticamente y narrativamente es débil.

#Conclusiones
La investigación permitió demostrar que el robo a vehículo en la Ciudad de México no se distribuyó de manera homogénea ni aleatoria, sino que respondió a patrones espaciales asociados tanto a características socioeconómicas como a elementos de infraestructura urbana de seguridad. A partir de la integración de bases de datos delictivas oficiales, indicadores de desarrollo social e información sobre alumbrado público y cámaras de videovigilancia, fue posible construir un marco analítico que articuló los planteamientos de la criminología ambiental con herramientas contemporáneas de análisis de datos y *machine learning*.

En primer lugar, el análisis exploratorio confirmó que las alcaldías con un Índice de Desarrollo Social (IDS) superior al promedio concentraron tasas significativamente más altas de delitos relacionados con vehículos particulares durante los días hábiles y dentro de

horarios de intensa actividad urbana. Este hallazgo se alineó con la literatura que señala que determinadas formas de criminalidad, como el robo a vehículo, tienden a concentrarse en zonas con mayor densidad económica, flujo de personas y bienes, y oportunidades situacionales para delinquir. Así, se corroboró el principio de que “el lugar importa”: no sólo importó la presencia de desventajas estructurales, sino también la configuración de contextos urbanos que generan una mayor exposición de posibles víctimas y objetivos.

Un aporte central del estudio fue la construcción y selección de features específicamente diseñadas para capturar la estructura urbana y el riesgo delictivo más allá de los conteos brutos de incidentes. La transformación de los delitos, unidades económicas, luminarias y hogares a indicadores de densidad por kilómetro cuadrado, así como la creación de tasas por habitante (delitos por 10 000 hab., cámaras por 10 000 hab.), permitió eliminar el sesgo asociado al tamaño de las unidades territoriales y hacer comparables colonias y alcaldías con superficies y poblaciones muy distintas. De igual forma, la incorporación del IDS como indicador sintético de desarrollo social, junto con las variables de infraestructura de seguridad, generó un conjunto de variables que resumió de manera más fiel la interacción entre contexto socioeconómico, diseño urbano y condiciones de vigilancia. El análisis estadístico previo (asimetrías, curtosis, correlaciones) y las transformaciones logarítmicas contribuyeron a estabilizar estas variables, lo que se reflejó en clústeres más compactos y mejor definidos.

Sobre esta base, la aplicación de técnicas de clustering mediante K-Means permitió segmentar el territorio en grupos de alcaldías con perfiles funcionales diferenciados, tomando en cuenta simultáneamente la carga delictiva, el nivel de desarrollo social y la disponibilidad de infraestructura de seguridad por habitante. La solución con K = 3 clústeres se mostró metodológicamente más robusta y sustantivamente más interpretable que configuraciones con mayor número de grupos, al combinar un mejor desempeño en métricas de validación interna con una estructura de perfiles territoriales coherente. Se identificaron así tres grandes patrones: (i) alcaldías con baja a moderada incidencia delictiva y niveles intermedios de desarrollo e infraestructura; (ii) alcaldías con alta carga delictiva asociada a mayores niveles de desarrollo y actividad urbana, pero con densidades de cámaras relativamente menores; y (iii) alcaldías con mayor densidad de cámaras por habitante y niveles delictivos relativamente inferiores. La forma en que las variables derivadas en particular, delitos por 10 000 habitantes, cámaras por 10 000 habitantes y densidades por km²— estructuraron estos clústeres mostró que el diseño cuidadoso de las features fue determinante para revelar patrones que no serían visibles con datos crudos.

Este resultado tuvo implicaciones relevantes para la política pública. Por un lado, evidenció que una mayor dotación de infraestructura de videovigilancia no se distribuyó necesariamente en los territorios de mayor riesgo relativo, lo que sugiere posibles desajustes entre la localización de recursos y los patrones de victimización. Por otro lado, mostró que la alta incidencia delictiva no pudo explicarse únicamente por condiciones socioeconómicas desfavorables: también se observó una importante concentración de delitos en zonas con IDS alto, posiblemente asociada a la intensidad de la actividad económica, la movilidad diaria y la abundancia de objetivos atractivos (vehículos, mercancías, flujos de tránsito). Ello reforzó la idea de que las políticas de prevención deben considerar tanto la vulnerabilidad estructural como la dinámica de uso del espacio urbano.

Metodológicamente, el trabajo aportó un esquema reproducible de integración, limpieza, transformación y estandarización de datos provenientes de múltiples fuentes institucionales, así como una reflexión explícita sobre la pertinencia de determinadas variables (por ejemplo, la exclusión de cámaras C5 por su carácter endógeno). El uso combinado de indicadores por kilómetro cuadrado, tasas por habitante, transformaciones logarítmicas, escalamiento estandarizado y métricas de validación interna permitió construir clústeres con mayor consistencia estadística y relevancia sustantiva. Este andamiaje ofrece una base sólida para la posterior incorporación de modelos predictivos supervisados que utilicen los clústeres y las *features* derivadas como insumos para estimar probabilidades de ocurrencia de delitos en distintos escenarios territoriales.

Interpretación de la matriz de correlación entre variables

Se calculó la correlación entre las variables clave del dataset:

Análisis:

La variable `delitos_anuales` tiene correlaciones moderadas con `hogares` (0.34) y `ue` (0.31), y correlaciones bajas con `AlumP_sum` (0.18) y `sup_km2` (0.09). Esto indica que la incidencia delictiva absoluta está parcialmente relacionada con la densidad poblacional y la actividad económica, pero apenas relacionada con la superficie de la colonia o el alumbrado público.

Se observan fuertes correlaciones entre las variables socioeconómicas:
`hogares` y `ue` tienen correlación alta (0.73), mostrando que colonias con más hogares tienden a concentrar más unidades económicas.
`hogares` y `AlumP_sum` tienen correlación alta (0.79), indicando que la infraestructura de alumbrado público está asociada con la densidad de hogares.
`ue` y `AlumP_sum` correlacionan moderadamente (0.59), sugiriendo que las áreas comerciales suelen estar mejor iluminadas.

`sup_km2` tiene correlaciones moderadas con `AlumP_sum` (0.44) y `hogares` (0.33), reflejando que colonias más grandes pueden tener más hogares e infraestructura, pero la relación no es lineal.
Uso en clustering K-Means:

Las variables `delitos_anuales`, `hogares`, `ue`, `AlumP_sum` y `sup_km2` aportan diferentes dimensiones de información:
`delitos_anuales` distingue colonias según la incidencia delictiva.
`hogares` y `ue` representan características socioeconómicas.
`AlumP_sum` y `sup_km2` reflejan infraestructura y tamaño de la colonia.
Este conjunto de variables permitirá identificar perfiles de colonias combinando densidad poblacional, actividad económica, infraestructura y seguridad, facilitando la segmentación en clusters significativos.

Interpretación de la correlación de variables con delitos anuales

Se calculó la correlación de delitos anuales con otras variables numéricas del dataset:

Análisis:

La variable `delitos_anuales` se correlaciona moderadamente con `hogares` (0.34) y `ue` (0.31), lo que indica que colonias con más hogares y más unidades económicas tienden a registrar un mayor número absoluto de delitos.
La correlación con `AlumP_sum` (0.18) y `sup_km2` (0.09) es baja, lo que sugiere que la presencia de alumbrado público o la superficie de la colonia no explica fuertemente la variación en el número de delitos.
Todas las variables seleccionadas son útiles para un **clustering K-Means** porque:
`delitos_anuales` permite diferenciar colonias con diferentes niveles de incidencia delictiva.
`hogares` y `ue` capturan información socioeconómica y de actividad económica que puede influir en patrones de delito.
`AlumP_sum` y `sup_km2` aportan información contextual sobre infraestructura y densidad espacial.
Es recomendable normalizar o transformar las variables** antes del clustering (por ejemplo, mediante logaritmo o estandarización) para asegurar que todas tengan igual peso en la creación de clusters.

Este análisis servirá como base para definir los perfiles de las colonias mediante K-Means, combinando aspectos de densidad poblacional, actividad económica e infraestructura urbana con la incidencia delictiva.
Relación con delitos:

Se observa una correlación POSITIVA entre cámaras y delitos. Las colonias con más cámaras (verdes) también tienen más delitos. Esto NO significa que las cámaras causen delitos, sino que las cámaras se instalan reactivamente en zonas de alta criminalidad.
Además tenemos el problema de endogeneidad, esta es una variable endógena. Las cámaras no se distribuyen aleatoriamente sino que se colocan donde ya hay problemas de seguridad. Incluirla en el clustering puede distorsionar los resultados porque está correlacionada con la variable objetivo (delitos).

Razones para excluirla:
Primera, la endogeneidad mencionada hace que la variable sea consecuencia del fenómeno que intentas segmentar, no una causa independiente. Incluirla crearía clusters circulares donde las zonas de alto riesgo se identifican porque tienen muchas cámaras, que a su vez se instalaron porque eran zonas de alto riesgo.
Segunda, la distribución extremadamente sesgada (skewness alto, muchos ceros, outliers extremos) dificulta que K-means la procese adecuadamente. Incluso con transformación logarítmica, los valores de 0 son problemáticos.
==========
CLUSTER 0
==========
Tamaño: 649 colonias (50.1%)

CARACTERÍSTICAS PROMEDIO:

Delitos por km2: 73.85 (↑ vs promedio: 44.33)
Unidades económicas por km2: 866.31 (↑ vs promedio: 681.99)
Alumbrado público por km2: 335.01 (↑ vs promedio: 233.94)
Hogares por km2: 6919.68 (↑ vs promedio: 4953.89)

ESTADÍSTICAS DETALLADAS:
Delitos por 1k - Mín: 2.42, Máx: 1860.83
Delitos por 1k - Mediana: 43.67
Delitos por 1k - Desv. Est.: 120.84

COLONIAS MÁS REPRESENTATIVAS (cercanas al centroide):
  - EL MOLINO: 51.4 delitos/km2, 697.8 UE/1k
  - MEDIA LUNA: 42.7 delitos/km2, 673.2 UE/1k
  - INDUSTRIAL I: 46.5 delitos/km2, 674.3 UE/1k
  - 20 DE NOVIEMBRE: 42.6 delitos/km2, 823.2 UE/1k
  - AMERICA: 42.1 delitos/km2, 758.5 UE/1k

==========
...
 - SAN PEDRO APOSTOL BARR: 7.3 delitos/km2, 446.1 UE/1k
  - EX EJIDO SAN FRANCISCO CULHUACAN I: 6.0 delitos/km2, 650.4 UE/1k
  - SAN PEDRO XALPA AMPL I: 9.2 delitos/km2, 483.1 UE/1k
  - VICTORIA DE LAS DEMOCRACIAS: 11.3 delitos/km2, 613.4 UE/1k

Interpretación del Clustering de Robo Vehicular en CDMX

Resumen

El análisis de clustering con K-means identificó tres perfiles distintos de colonias en la Ciudad de México basándose en densidades espaciales de delitos vehiculares, unidades económicas, alumbrado público y hogares por kilómetro cuadrado. Los resultados revelan patrones claros que permiten orientar estrategias de prevención diferenciadas.

Cluster 0: Corredores Urbanos de Alto Riesgo

Tamaño: 651 colonias (50.2%)

Características Distintivas

Este cluster representa la mitad de todas las colonias analizadas y concentra las zonas con mayor problemática de robo vehicular. La densidad de delitos promedio es de 73.68 por km², lo cual supera en un 66% al promedio general de la ciudad. Estas colonias presentan simultáneamente la mayor densidad comercial (865 UE/km²), la mayor concentración de hogares (6,914 hogares/km²) y la mejor infraestructura de iluminación (334 luminarias/km²).

Interpretación

La paradoja de este cluster es reveladora: a pesar de tener el mejor alumbrado público, presenta la mayor criminalidad. Esto indica que la iluminación por sí sola no es suficiente cuando existe una alta concentración de actividad económica y residencial. Las colonias representativas como El Molino, Media Luna e Industrial I ejemplifican zonas mixtas con intensa actividad comercial donde la alta afluencia vehicular genera más oportunidades para los delincuentes.

Heterogeneidad Interna

La desviación estándar de 120.70 delitos/km² y el rango que va desde 2.42 hasta 1,860.83 delitos/km² indican que este cluster es internamente heterogéneo. Algunas colonias dentro de este grupo requieren intervención urgente mientras otras están en niveles manejables. Se recomienda un análisis secundario para identificar las colonias más críticas dentro de este cluster.

Recomendaciones de Política Pública

La intervención en este cluster debe ser multifactorial. El alumbrado existente debe complementarse con cámaras de vigilancia C5, patrullaje focalizado en horarios de alta afluencia vehicular, y regulación de estacionamientos en vía pública. Dado que concentra aproximadamente el 70-80% de los delitos totales, este cluster debe ser la prioridad máxima de inversión en seguridad.

Cluster 1: Zonas Periféricas de Baja Densidad

Tamaño: 167 colonias (12.9%)

Características Distintivas

Este es el cluster más pequeño y presenta los indicadores más bajos en todas las dimensiones. La densidad de delitos es de apenas 18.24 por km², representando una reducción del 59% respecto al promedio. La densidad comercial es mínima (117 UE/km²), el alumbrado es escaso (68 luminarias/km²) y la densidad de hogares es la más baja (1,066 hogares/km²).

Interpretación

Estas colonias corresponden principalmente a zonas periféricas, rurales o de conservación ecológica dentro de la Ciudad de México. Colonias como La Magdalena Petlacalco, San Pedro de los Pinos y Santa María Insurgentes representan áreas con baja actividad urbana general. La criminalidad es baja no por la efectividad de medidas de seguridad, sino porque existe menor actividad económica y vehicular, lo que reduce las oportunidades de robo.

Advertencia Estadística

A pesar del bajo promedio, el valor máximo de 1,422.51 delitos/km² indica la presencia de al menos un outlier extremo que merece investigación individual. La mediana de apenas 2.36 delitos/km² confirma que la gran mayoría de estas colonias son efectivamente seguras.

Recomendaciones de Política Pública

No se requiere inversión urgente en seguridad para este cluster. Sin embargo, es importante monitorear el crecimiento urbano en estas zonas, ya que la expansión de la mancha urbana podría incrementar gradualmente los niveles de criminalidad. La inversión debe enfocarse en infraestructura básica conforme estas zonas se desarrollen.

Cluster 2: Zonas Residenciales de Riesgo Moderado

Tamaño: 478 colonias (36.9%)

Características Distintivas

Este cluster presenta el mejor balance entre seguridad y desarrollo urbano. La densidad de delitos es de 13.41 por km², un 70% menor que el promedio general. Mantiene una actividad comercial cercana al promedio (629 UE/km²), iluminación moderada (154 luminarias/km²) y densidad residencial intermedia (3,638 hogares/km²).

Interpretación

Estas colonias representan zonas residenciales consolidadas con actividad comercial equilibrada. Colonias como Torres de Potrero, San Pedro Apóstol y Ex Ejido San Francisco Culhuacán ejemplifican barrios donde existe vida urbana activa pero con niveles de criminalidad controlados. La combinación de densidad comercial moderada con menor concentración vehicular genera un entorno más seguro.

Consistencia Interna

La desviación estándar de apenas 15.47 delitos/km² es significativamente menor que en los otros clusters, indicando que este grupo es internamente homogéneo. El rango de 0.07 a 113.73 delitos/km² es el más compacto de los tres clusters, lo que facilita la implementación de políticas estandarizadas.

Recomendaciones de Política Pública

Este cluster representa el modelo a mantener y potencialmente replicar. Las intervenciones deben enfocarse en prevención y mantenimiento de la infraestructura existente. Incrementar la iluminación en un 30-50% podría consolidar estos niveles de seguridad. Estas colonias son candidatas ideales para programas de vigilancia comunitaria y prevención social del delito.

Hallazgos Clave

La Iluminación No Es Suficiente Por Sí Sola

El Cluster 0 demuestra que tener la mejor iluminación (334 luminarias/km²) no garantiza seguridad cuando existe alta densidad comercial. El ratio de iluminación por unidad económica es un mejor predictor que la iluminación absoluta. El Cluster 2 logra menor criminalidad con menos iluminación porque tiene menor densidad comercial.

Concentración del Problema

El 50.2% de las colonias (Cluster 0) concentra la mayoría de los delitos vehiculares. Esto confirma el principio de Pareto aplicado a la criminalidad urbana y permite focalizar recursos de manera eficiente.

Importancia de la Densidad Comercial

La variable más predictiva de criminalidad vehicular es la densidad de unidades económicas. Las zonas con alta concentración comercial generan mayor circulación y estacionamiento de vehículos, incrementando las oportunidades delictivas independientemente de otros factores.

El análisis de clustering revela una realidad incómoda para la política de seguridad pública en la Ciudad de México: la inversión en infraestructura de iluminación, tradicionalmente considerada como solución universal contra la delincuencia, resulta insuficiente e incluso irrelevante cuando no se considera el contexto de densidad comercial y actividad económica de cada zona. El Cluster 0, que agrupa más de la mitad de las colonias analizadas y concentra la mayor problemática de robo vehicular, cuenta paradójicamente con la mejor infraestructura de alumbrado público de toda la ciudad. Este hallazgo desafía directamente las estrategias convencionales de prevención situacional y sugiere que los recursos invertidos en iluminación en zonas de alta densidad comercial tienen un retorno mínimo si no se acompañan de medidas complementarias como vigilancia tecnológica, patrullaje focalizado y regulación del estacionamiento en vía pública. La concentración del 50% de las colonias problemáticas en un solo cluster representa tanto un desafío como una oportunidad: los recursos pueden focalizarse de manera eficiente en lugar de dispersarse en intervenciones generalizadas de bajo impacto.

Desde una perspectiva metodológica, este ejercicio demuestra las limitaciones de trabajar con datos urbanos desactualizados y la necesidad de desarrollar indicadores alternativos cuando las fuentes tradicionales no están disponibles. La decisión de utilizar densidades espaciales por kilómetro cuadrado en lugar de tasas poblacionales per cápita resultó ser no solo una solución práctica ante la obsolescencia del censo 2010, sino posiblemente un

enfoque más apropiado para el análisis de delitos vehiculares, donde la variable relevante no es cuántas personas habitan una zona sino cuánta actividad económica y circulación vehicular genera. Sin embargo, la alta heterogeneidad interna del Cluster 0, evidenciada por una desviación estándar que supera a la propia media, indica que este agrupamiento requiere refinamiento adicional mediante subclustering para identificar con precisión las colonias que demandan intervención urgente versus aquellas que, aunque clasificadas como alto riesgo, operan en niveles manejables. El verdadero valor de este análisis no radica en las clasificaciones finales sino en la identificación de las variables que realmente predicen la criminalidad vehicular: la densidad de unidades económicas emerge como el factor dominante, relegando a la iluminación y la densidad residencial a roles secundarios que solo cobran relevancia cuando se analizan en proporción a la actividad comercial de cada zona.