Práctica de Laboratorio:
Análisis Exploratorio de Datos - Data Wrangling

Docente: Ana Maria CuadrosValdivia

Para realizar el Análisis Exploratorio de datos, lo primero que deberíamos hacer es intentar responder a las siguientes preguntas (data wrangling):

Paso 1: Analiza el comportamiento de tus datos.

- Un registro es una entidad, describa que representa un registro

  Cada registro representa un incidente criminal reportado en Los Ángeles (2020-presente), con detalles como ubicación, hora, tipo de crimen, y datos demográficos de la víctima.

- ¿Cuántos registros hay?
  - ¿Son demasiado pocos?
    Hay 1005109 registros, es un volumen medio.
  - ¿Son muchos y no tenemos Capacidad (CPU+RAM) suficiente para procesarlo?
    Si tenemos la capacidad computacional para manejar este volumen de datos
  - ¿Hay datos duplicados?
    No hay duplicados

- ¿Qué datos son discretos y cuáles continuos?
  - Muchas veces sirve obtener el tipo de datos: texto, int, double, float
    - ¿Cuáles son los tipos de datos de cada columna?
      - Categóricos: AREA NAME, Crm Cd Desc, Vict Sex (texto).
      - Númericos: TIME OCC (hora: 1-2359), Vict Age (edad: 0-120, con outliers).Coordenadas: LAT (33.8-34.3), LON (-118.6--118.1).
      - Fechas: DATE OCC y Date Rptd (datetime).
    - ¿Entre qué rangos están los datos de cada columna?, valores únicos, min, max
      - TIME OCC (hora: 1-2359)
      - Vict Age (edad: 0-120, con outliers)
      - LAT (33.8-34.3)
      - LON (-118.6--118.1).

- DATE OCC y Date Rptd (datetime)

- ¿Todos los datos están en su formato adecuado?
  - Premis Cd (código de premisa) es float64 pero debería ser int64 (tiene decimales por nulos).
  - Weapon Used Cd tiene nulos almacenados como float64.
- Los datos tienen diferentes unidades de medida?
  - Tiempo:
    - TIME OCC: Formato 24h sin separador (ej: 1430 = 2:30 PM).
    - Vict Age: Años (entero).
  - Geografía:
    - LAT/LON: Grados decimales (WGS84).
  - Códigos:
    - Crm Cd, Premis Cd: Enteros sin unidades.
- Cuáles son los datos categóricos, ¿hay necesidad de convertirlos en numéricos?
  - Nominales:
    - AREA NAME (ej: "Wilshire", "Central").
    - Vict Sex ("M", "F", "X").
    - Crm Cd Desc (descripción del crimen).
  - Ordinales:
    - Part 1-2 (gravedad del crimen: 1 = más grave, 2 = menos grave).

- ¿Qué representa un registro?
  - Describe qué representa cada fila.
    Cada fila del dataset Crime_Data_from_2020_to_Present.csv representa:
    - Un incidente criminal reportado en la ciudad de Los Ángeles, California, desde 2020 hasta la fecha actual.
    - Contiene detalles del crimen, información demográfica de la víctima, ubicación geográfica, y estatus del caso.
  - Si es una data etiquetada, como interpretas la información de las clases?
    - Crm Cd Desc (para clasificación multiclase: predecir tipo de crimen).
    - Status Desc (para clasificación binaria: "Adult Arrest")
      .
  - ¿Hay niveles de granularidad de los datos? Por ejemplo, datos a nivel país, región, ciudad. Años, meses, días, horas, minutos, etc.
    Estos comprenden múltiples niveles de granularidad.

- ¿Están todas las filas completas o tenemos campos con valores nulos?
  - En caso que haya demasiados nulos: ¿Queda el resto de información inútil?. Se debe agregar o combinar sus datos

        - Si hay valores nulos en algunas columnas ,pero como no supera el 80’% no se le hace imputación de datos.
    - Si se agregan datos debe comprobar que siguen el mismo comportamiento. Por ejemplo, tiene la misma media, mediana, etc.
        - Afecta en poco por ser tan pequeño el número de datos nulos.

- ¿Siguen alguna distribución?
  Usa describe() y analiza los valores.
    - Edad (Vict Age):
        - Media: 34.5 años, pero con mínimo en 0 (¿error?) y máximo en 120 (posible outlier).
        - 75% de víctimas tienen ≤42 años.
    - Hora (TIME OCC):
        - Distribución uniforme entre 100 (1:00 AM) y 2359 (11:59 PM).
        - Pico en horas pico (medianoche y tarde/noche).
    - Geografía (LAT/LON):
        - Coordenadas centradas en LA (34.05, -118.25).
        - Desviación estándar baja: crímenes concentrados en área metropolitana.

- ¿Hay correlación entre features (características)?
    - p-value < 0.001: Relación significativa entre sexo y área geográfica.
    -
    - Insight: Áreas como Central tienen mayor proporción de víctimas masculinas.

Paso 2. Análisis de outliers<

- ¿Cuáles son los Outliers? (unos pocos datos aislados que difieren drásticamente del resto y “contaminan” ó desvían las distribuciones)
- Variable Tipo de Outlier Ejemplo Posible Causa
- Vict Age Edades extremas (0, >100) 0, 120 Error de registro o datos simbólicos (ej: "0" para edad desconocida).
- TIME OCC Horas fuera de rango (1-2359) 99, 2400 Error de digitación (ej: "99" en lugar de "2359").
- LAT/LON Coordenadas fuera de LA 34.5, -118.8 Geolocalización incorrecta o crímenes en áreas limítrofes.
- Weapon Used Cd Valores extremos (>1000)
    - ➢ ¿Podemos eliminarlos? ¿Es importante conservarlos?
    - ➢ son errores de carga o son reales?

Variable	Tipo de Outlier	Ejemplo	Posible Causa
Vict Age	Edades extremas (0, >100)	0, 120	Error de registro o datos simbólicos (ej: "0" para edad desconocida).
TIME OCC	Horas fuera de rango (1-2359)	99, 2400	Error de digitación (ej: "99" en lugar de "2359").
LAT/LON	Coordenadas fuera de LA	34.5, -118.8	Geolocalización incorrecta o crímenes en áreas limítrofes.
Weapon Used Cd	Valores extremos (>1000)		

Paso 3: Visualización

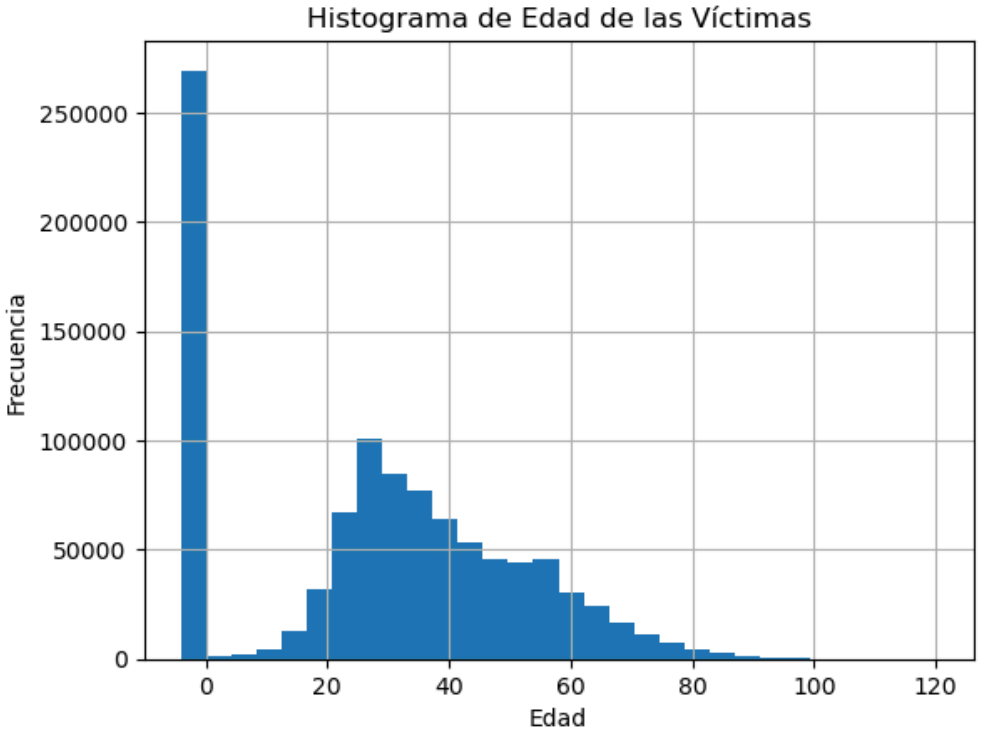
- Las variables que podemos representar son:
  - Variables categóricas: Gráfico de barras y circular
  - Variables numéricas: Una variable: histogramas, dos variables: boxplot

Gráfico de barras: comparar cantidades de una variable.

Gráfico circular: para representar porcentajes y proporciones.

Boxplot: representa los datos numéricos a través de sus cuartiles pudiendo representar los outliers.

Scatterplot: muestra el grado de relación entre dos variables.



-

Paso 4. Encuentra un problema potencial en tus datos.

- Si es un problema de tipo supervisado:
  - ¿Cuál es la columna de “salida”? ¿binaria, multiclase?
    Es multiclase la descripción del crimen
  - ¿Está balanceado el conjunto salida?
    THEFT FROM MOTOR VEHICLE (10.2%), ASSAULT (7.5%), BURGLARY (5.1%).

- ¿Cuáles parecen ser features importantes? ¿Cuáles podemos descartar?
  Todos son importantes a excepción de estos que tiene gran porcenajtaa este Cross Street (85% nulos), Crm Cd 2-4 (90% nulos), Mocodes (15% nulos y baja relevancia).
  - ¿Estamos ante un problema dependiente del tiempo? Es decir un TimeSeries.
    Si y tambien localización
  - Si fuera un problema de Visión Artificial: ¿Tenemos suficientes muestras de cada clase y variedad, para poder hacer generalizar un modelo de Machine Learning?
    Si tenemos la suficiente cantidad y variedad para un modelo de Machine Learning.

Conclusión

¿Qué podemos aprender de este análisis?

A través de este análisis hemos podido:
- Explorar la estructura y distribución de los datos de criminalidad.
- Identificar outliers y revisar valores nulos.
- Visualizar los tipos de crimen más comunes y sus ubicaciones.
- Detectar problemas de calidad y posibles mejoras en el preprocesamiento.