Bitacora 3

Bitácora 3 Proyecto Grupal

Integrantes:

- Sebastián Calderón Segura C21517
- Sebastián Miranda Ramírez C4H274
- Josué Gustavo Rodríguez Aguilar C4J023
- Kevin David Calderon Martínez C4D511
- Santiago Paniagua Chavarría C4I249

1 Planteamiento del problema:

Posterior a la limpieza exhausiva de datos, se procede con la pregunta de investigación que funcionará como guía para el análisis del trabajo y darle finalidad a la data tratada, dicha pregunta es:

¿Cómo varía la frecuencia de los principales delitos en Costa Rica entre 2019 y 2025 según la provincia y las características de la víctima (edad y sexo)?

Para dar respuesta a dicha pregunta se procede por relaciones lineales; en este caso, al ser multivariables relacionadas con la frecuencia de delitos, se condensará la respuesta en correlaciones lineales, en específico en la técnica del coeficiente de Pearson, pues identifica qué tan fuerte y en qué dirección se asocian estas variables con los cambios en los delitos a lo largo del tiempo.El mismo se explicara con más detalle continuación.

Marco metodológico

1. Enfoque y tipo de estudio

El estudio se desarrolla bajo un enfoque cuantitativo, de tipo descriptivo–correlacional, con base en registros administrativos oficiales. Se analizan los delitos cometidos en Costa Rica

entre 2019 y 2025 (periodo disponible en la base del OIJ) con el objetivo de identificar patrones de criminalidad por género en los cantones del país y estudiar su relación con variables socioeconómicas como pobreza y desempleo.

La unidad de análisis es el evento delictivo registrado, que luego se agrupa por año, provincia/cantón, sexo y características de la víctima (principalmente edad y sexo). A nivel analítico, se trabaja con unidades agregadas del tipo cantón–año–género, que permiten vincular los delitos con indicadores socioeconómicos oficiales.

2. Fuentes de información y bases de datos

Los datos principales de criminalidad provienen de las bases de delitos del Organismo de Investigación Judicial (OIJ) y de las estadísticas policiales consolidadas entre 2019 y julio de 2025. Como apoyo contextual se consideran informes del Observatorio de Seguridad Ciudadana, el Ministerio de Seguridad Pública y los indicadores de desarrollo y género publicados por el PNUD, que incluyen información sobre condiciones socioeconómicas por cantón, como pobreza y desempleo.

3. Preparación y limpieza de los datos

Las bases originales se encontraban separadas por año y con diferencias en nombres de variables y formatos (por ejemplo, "género" vs. "sexo", o fechas en distintos formatos). Estas se unificaron en una sola base de más de 345 000 registros con 12 variables categóricas (delito, subdelito, fecha, hora, víctima, subvíctima, edad, sexo, nacionalidad, provincia, cantón y distrito).

Para mejorar la calidad de los datos se aplicaron los siguientes pasos:

- **Reclasificación de valores "NO APLICA" y "DESCONOCIDO"** como NA cuando no tenían sentido para el tipo de víctima (por ejemplo, edad o sexo asignados a "VIVIENDA" o "VEHÍCULO").
- **Construcción de tablas de frecuencias** para todas las variables categóricas (delito, víctima, edad, sexo, nacionalidad, provincia y cantón) y elaboración de gráficos de barras y mapas coropléticos para describir la distribución espacial de los delitos.
- **Detección de outliers** en las frecuencias mediante el método del rango intercuartílico (IQR) y **winsorización** de los valores extremos (por ejemplo, cantones o provincias con frecuencias muy altas), con el fin de reducir la distorsión que estos pueden generar en los análisis correlacionales sin eliminar información relevante.

A partir de esta base depurada se construyen variables cuantitativas de interés, como frecuencias anuales de delitos por provincia/cantón y sexo, y, cuando la información lo permite, tasas de delitos por población para cada cantón.

4. Etapas del análisis estadístico

Debido a la forma en que está planteada la pregunta de investigación, el análisis se organiza en tres etapas principales:

4.1. Etapa 1: Frecuencia de delitos a lo largo del tiempo

En la primera etapa se estudia cómo varía la cantidad de delitos entre 2019 y 2025, desagregada por sexo de la víctima y por provincia. Para ello:

- Se agrupan los registros por año, sexo y provincia.
- Se elaboran tablas de frecuencia y gráficos para mostrar la evolución temporal de los delitos, distinguiendo entre hombres y mujeres.
- De esta manera se identifican tendencias generales (incrementos, descensos o estabilidad en el tiempo) y se detectan años críticos para cada grupo.

En esta etapa, el año se trata como una variable cuantitativa (2019, 2020, …, 2025), lo que permite explorar si existe una tendencia lineal entre el paso del tiempo y la frecuencia de delitos.

4.2. Etapa 2: Delitos según características de la víctima (edad y sexo)

La segunda etapa se centra en la relación entre la frecuencia de delitos y las características de la víctima, especialmente edad y sexo.

- Se generan tablas de contingencia que cruzan los rangos de edad (menor de edad, mayor de edad, adulto mayor) y el sexo de la víctima, con el número de delitos registrados.
- Se calculan porcentajes por grupo (por ejemplo, porcentaje de delitos contra mujeres mayores de edad vs. hombres menores de edad) y se representan mediante gráficos de barras agrupadas o apiladas.

Aunque edad y sexo son variables categóricas, al trabajar con sus frecuencias se obtienen indicadores numéricos que permiten describir qué grupos son más afectados y comparar la carga delictiva entre ellos.

4.3. Etapa 3: Distribución espacial de los delitos por provincias y cantones

En la tercera etapa se analiza la distribución espacial de los delitos.

- Se calcula la frecuencia de delitos por cantón, separando por sexo de la víctima.
- Esta información se vincula con un shapefile cantonal del INEC para generar mapas de tipo coroplético que permitan visualizar qué cantones concentran mayores niveles de criminalidad y cómo se distribuyen según género.

Esta etapa es clave para el objetivo de sectorizar la criminalidad y para preparar el terreno del análisis correlacional con variables socioeconómicas.

5. Análisis correlacional con factores socioeconómicos

Después de caracterizar los patrones temporales, demográficos y espaciales de la criminalidad, se realizó un análisis correlacional con el objetivo de explorar si las tasas de victimización nacionales presentan alguna asociación con indicadores de desarrollo humano y condiciones educativas de la población.

Dado que no fue posible obtener datos cantonales completos de pobreza ni desempleo para todos los años del periodo 2019–2025, el análisis se centró en aquellas variables socioeconómicas disponibles para todos los cantones y que cuentan con series temporales consistentes a nivel nacional. Estas variables provienen del PNUD y del INEC e incluyen:

- IDG-D: Índice de Desarrollo de Género desagregado.
- IDH: Índice de Desarrollo Humano.
- IDH-D: Índice de Desarrollo Humano desagregado.
- Escolaridad femenina: porcentaje de mujeres con al menos educación secundaria.
- Escolaridad masculina: porcentaje de hombres con al menos educación secundaria.

Para cada año, se construyeron las tasas nacionales de delitos por cada 100 000 habitantes, desagregadas por sexo, a partir de:

- La frecuencia anual de víctimas hombres y mujeres.
- La población total estimada del país.

Con estas variables se generó una matriz de correlaciones utilizando el coeficiente de Pearson, el cual permite medir la fuerza y dirección de la asociación lineal entre:

- La tasa nacional de víctimas hombres (por 100 000 habitantes).
- La tasa nacional de víctimas mujeres (por 100 000 habitantes).
- Los indicadores IDH, IDH-D, IDG-D.
- Los niveles promedio de escolaridad por sexo.

El análisis se centra en identificar si los niveles de desarrollo humano, equidad de género y escolaridad presentan alguna relación estadística con el comportamiento de las tasas de criminalidad durante el periodo 2019–2023.

6. Justificación del uso del coeficiente de correlación de Pearson

El coeficiente de correlación de Pearson (r) es una medida estadística que indica qué tan fuerte y en qué dirección se relacionan dos variables cuantitativas. Su valor va de –1 a +1:

- Valores cercanos a +1 indican una relación lineal positiva fuerte (cuando una variable aumenta, la otra también).
- Valores cercanos a –1 indican una relación lineal negativa fuerte (cuando una sube, la otra tiende a bajar).
- Valores cerca de 0 sugieren ausencia de relación lineal clara.

En este proyecto se escoge el coeficiente de Pearson por varias razones metodológicas:

1. **Las variables clave se transforman en cuantitativas.** Las tasas de victimización por cada 100 000 habitantes para hombres y mujeres son variables continuas, al igual que los indicadores socioeconómicos incluidos: IDH, IDH-D, IDG-D y porcentajes de escolaridad. Esto permite aplicar el coeficiente de Pearson.

2. **Interesa específicamente la relación lineal.** El propósito del análisis es determinar si variaciones en el nivel de desarrollo humano o en la escolaridad se asocian con aumentos o disminuciones en las tasas de criminalidad. Este tipo de relación —del tipo "a mayor desarrollo, menor tasa" o "a mayor escolaridad, menor criminalidad"— corresponde a una relación lineal, precisamente la que Pearson está diseñado para medir.

3. **Tamaño muestral grande y datos agregados.** Aunque el análisis se realiza a nivel nacional (una observación por año), las series disponibles son consistentes y comparables entre sí, lo cual permite explorar tendencias generales. Además, la estandarización de las tasas y el uso de promedios cantonales de IDH, IDH-D y escolaridad reducen la variabilidad extrema y mejoran la estabilidad de las correlaciones estimadas.

4. **Interpretación sencilla para la toma de decisiones.** Pearson ofrece un único número fácil de interpretar:

   - |r| entre 0 y 0,3 relación débil,
   - |r| entre 0,3 y 0,6 relación moderada,
   - |r| mayor a 0,6 relación fuerte. Esto facilita explicar los resultados a nivel académico y de política pública (por ejemplo, "la criminalidad tiene una correlación moderada y positiva con la pobreza cantonal").

5. **Posibilidad de comparar por género.** El cálculo de correlaciones separadas para las tasas de víctimas hombres y de víctimas mujeres facilita identificar si los indicadores de desarrollo humano se relacionan de manera distinta con la criminalidad según el género. Esto fortalece el enfoque de género del proyecto y permite identificar patrones específicos de vulnerabilidad.

Conclusiones

El análisis de la evolución anual de las víctimas revela que el año 2019 concentra los niveles más altos de criminalidad tanto para hombres como para mujeres. En 2020 se observa un descenso significativo en todas las provincias, probablemente influenciado por las restricciones de movilidad durante la pandemia, lo cual redujo las oportunidades delictivas. A partir de 2021 se identificó un repunte progresivo que alcanza picos en 2023 o 2024 según la provincia. En todo el periodo, los hombres presentan sistemáticamente niveles más altos de victimización que las mujeres, aunque ambos grupos siguen tendencias muy similares. El aparente descenso en 2025 debe interpretarse con cautela, pues corresponde a un año con registros incompletos (enero–julio). Estos resultados muestran que la dinámica delictiva no es aleatoria, sino que sigue ciclos temporales que responden tanto a factores sociales como a condiciones extraordinarias.

Al desagregar los delitos según sexo y rangos de edad, se observa que la mayoría de víctimas corresponden a personas mayores de edad, seguidas (a una distancia considerable) por menores de edad y adultos mayores. En todas las categorías etarias, los hombres representan la mayor proporción de víctimas, aunque las mujeres mantienen una presencia significativa, especialmente en los grupos mayoritarios. Por otra parte, los mapas cantonales muestran que la criminalidad no se distribuye homogéneamente en el territorio nacional. La mayor concentración de delitos se ubica en zonas urbanas densamente pobladas, particularmente en el Gran Área Metropolitana, mientras que regiones periféricas presentan niveles notablemente menores. Estos patrones geográficos revelan desigualdades territoriales que deben considerarse al formular estrategias de prevención y políticas públicas.

Como última conclusión, la correlación entre las tasas nacionales de delitos por sexo y los indicadores encontrados (IDH-H, IDG-G y escolaridad hasta al menos la secundaria por sexo) sugiere que sí existe una relación lineal entre el nivel de desarrollo del país y la evolución general de la criminalidad entre 2019 y 2023. Sin embargo, estas asociaciones deben entenderse con su debida cautela. Pues solo se encuentra una observación general por año, y la falta de indicadores genwrales de pobreza o desempleo a nivel cantonal, en resumen, falta de información. De todas formas, los resultados sugieren que la escolaridad también podría estar relacionada con la dinámica delictiva y que estos deben seguir siendo estudiados con mayor profundidad y con información actualizada. Se concluye en general que la seguridad y el bienestar social no han de ser fenómenos aislados, sino dimensiones correlacionadas que influyen conjuntamente en la realidad delictiva de Costa Rica.

Limitaciones

Las bases de datos del OIJ no venían de forma unificada; sino separadas por año, y además presentaban diferencias en nombres de columnas y extensiones. Por lo tanto se tuvo que crear un script para leer cada año, estandarizar variables y unir todo en una sola base. Aunque esto permitió tener una base de datos unificada, cualquier error manual durante este proceso (sobre todo porque se realizó al inicio del proyecto) podría haber afectado los resultados.

Debido a que las bases de datos traía información de sexo, nacionalidad, edad o subvíctima aún cuando la víctima no era persona (por ejemplo, era un vehículo, vivienda, o empresa) se decidió cambiar todas esas variables a "NO APLICA". Esto tiene lógica, pues el mismo OIJ aclara que dichas variables describen específicamente a la víctima, y objetos no tienen ni edad ni sexo. Sin embargo, esto también implica que, si por alguna razón en algún registro se había anotado información adicional, como datos del dueño del bien afectado, esa información se pierde.

Como el archivo solo llega hasta julio de 2025, el año se presenta con menos observaciones, lo que a primera vista puede hacer parecer que hubo menos criminalidad en dicho año, cuando en realidad es simplemente que no está completo.

Se aplicó winsonorización para reducir el impacto de los valores extremadamente altos, lo que, aunque mejora la uniformidad (sobretodo en los gráficos), también modifica artificialmente las frecuencias más altas, lo que hace que los datos de cantones con delitos muy altos sean atenuados, subestimando así la concentración de los delitos en estas zonas.

Durante el análisis se usó la división cantonal correspondiente a 2024 para hacer los mapas, pero los años abarcan desde 2019. En esos años hubieron cambios a nivel cantonal, como la creación de Monteverde y Puerto Jiménez como cantones. Esto implica que en los primeros años dichos cantones no tendrán datos registrados en los mapas, pues esos datos se registraron sobre el cantón al que pertenecían antiguamente. Por tanto, en el análisis de los cantones que sufrieron este fénomeno se debe tener especial cuidado.