

# Criminalidad en grandes ciudades: Caso Los Ángeles, California

¿Cómo varían los patrones de  
criminalidad en Los Ángeles según el  
tiempo y la ubicación?



# ¿Cuál es el contexto y el público objetivo?

**Objetivo:** Analizar la criminalidad en Los Ángeles identificando posibles mejoras a la seguridad pública.

**Datos:** Analizar patrones delictivos para asignar recursos de manera más efectiva.

**Aplicación:** Extensible a otras ciudades con problemas similares



## Público Objetivo

- Autoridades locales: Mejorar estrategias de seguridad y distribución de recursos.
  - Investigadores y analistas: Identificar patrones para políticas preventivas.
  - Comunidad: Mejorar la seguridad en colaboración con las autoridades.

## Limitaciones !

- Falta de precisión o datos incompletos.
  - No todos los delitos son denunciados.
  - Poca inclusión de factores contextuales (socioeconómicos o ambientales).

# ¿Qué preguntas queremos responder con el análisis?

## Preguntas principales:

- ¿Qué patrones se pueden identificar en los tipos de crímenes cometidos según la hora del día o el área geográfica?
- ¿Cómo pueden estos patrones ayudar a predecir futuros delitos y guiar la asignación de recursos y políticas públicas de prevención?

## Preguntas secundarias:

- ¿En qué zonas específicas de Los Ángeles se reportan más delitos violentos y no violentos?
- ¿Cuáles son los lugares y momentos con mayor incidencia de delitos en Los Ángeles?
- ¿Cuál es la tendencia de criminalidad a lo largo del año o por estaciones?
- ¿Qué tipo de delitos son más frecuentes en determinadas áreas?

# Descripción del Dataset

El dataset contiene información clave sobre el tipo, lugar y tiempo de crímenes reportados en Los Ángeles (2020-2024), junto con datos de la víctima y el estado del caso. Este análisis apoya la mejora de la seguridad pública. Fuente: [data.lacity.org](<https://data.lacity.org/Public-Safety>)

978628 registros

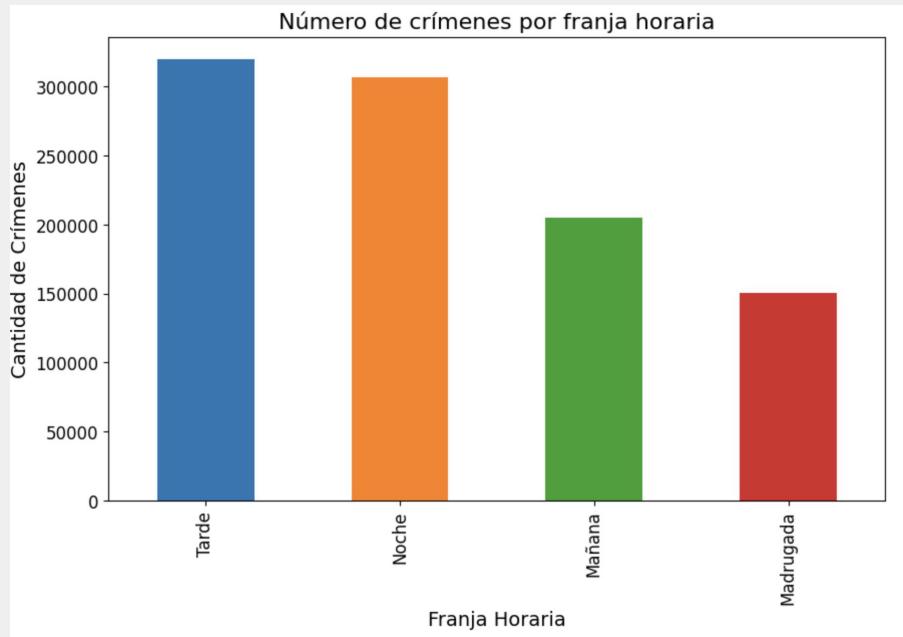
Del 01/01/2020 al 30/09/2024

28 variables

Principales variables				
 Fecha y hora	 Ubicación	 Crimen	 Víctima	 Caso
Date Reported	Area ID	Crime Code	Victim Age	Status Code
Date Occurred	Area Name	Crime Code	Victim Sex	Status Description
Time Occurred	LAT	Description	Victim Descent	
	LON	MO Codes		

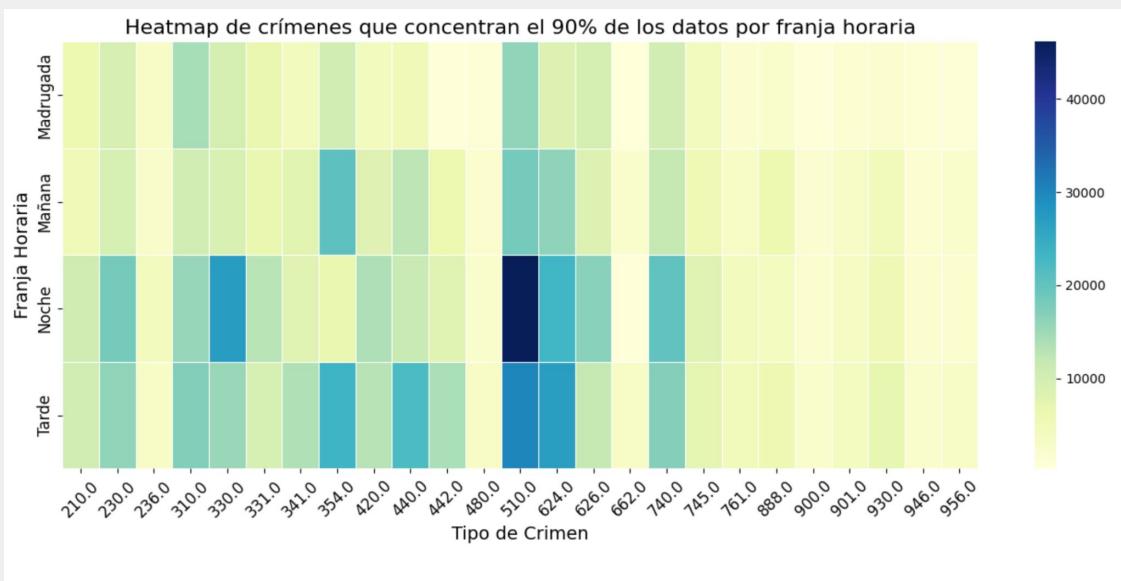
**Exploraremos  
los datos  
(Primera  
parte)**

# ¿Cómo varían los crímenes según la hora del día?



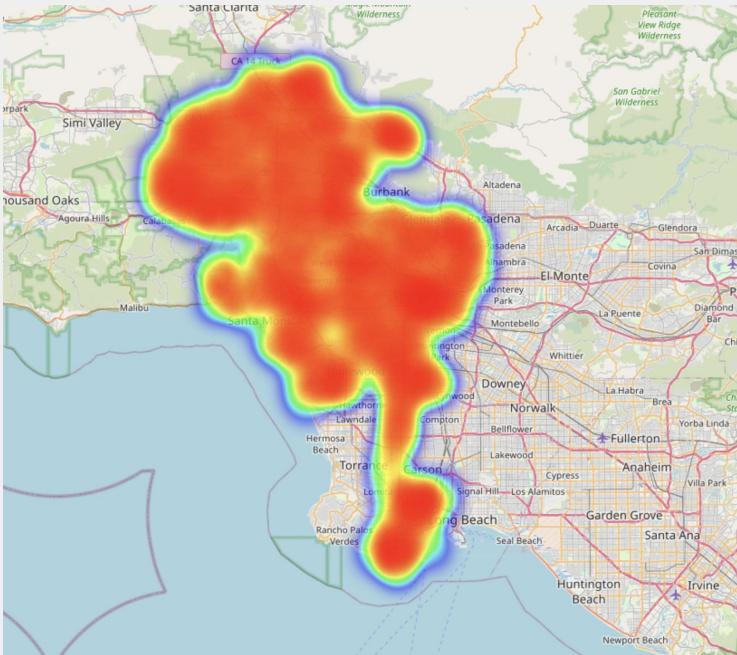
- **Tarde y Noche:** Mayor cantidad de crímenes, coincidiendo con el mayor movimiento social y comercial.
- **Madrugada:** Menor actividad delictiva, cuando las calles están más tranquilas.

# ¿Cómo varían los crímenes más comunes durante el día?



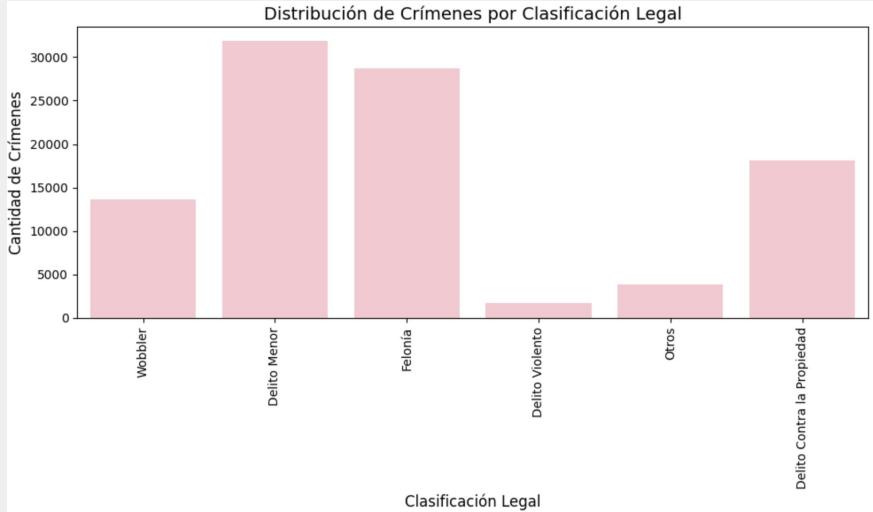
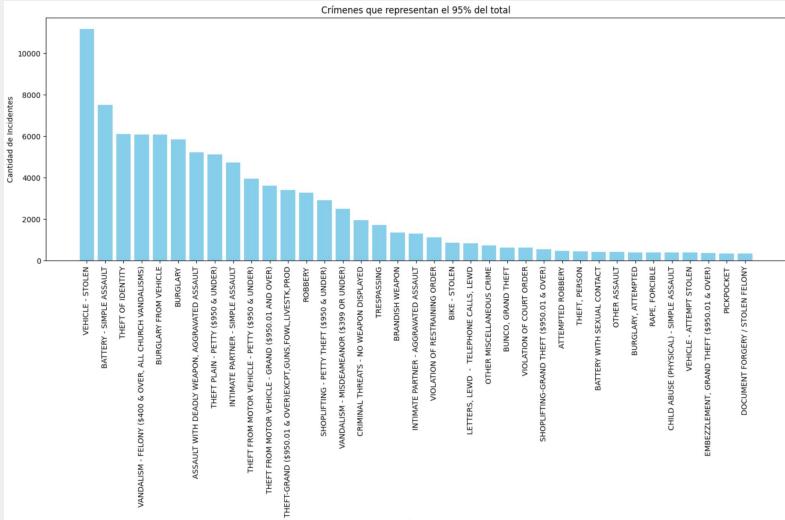
- **Hurto desde vehículo** (330) es más frecuente en la **noche**.
- **Robo de vehículos** (510) ocurre principalmente en la **mañana**.
- **Asalto simple** (624) es más común en la **tarde**.
- **Vandalismo** (740) tiene un pico en la **tarde**, relacionado con mayor actividad pública.
- La **tarde** es una franja de alta actividad delictiva.
- La **noche** concentra ciertos crímenes como **hurto desde vehículos**.

# ¿Cómo varían los crímenes por zona?



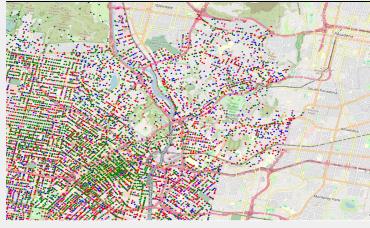
- **Zonas de menor densidad:** Áreas como *Beverly Hills* y *Santa Monica* (*centro y sur cerca de la costa*) presentan menos crímenes en comparación a sus alrededores.
- **Zonas con alta concentración criminal:** El sur de Los Ángeles y el oeste de la ciudad muestran una mayor densidad de crímenes, posiblemente vinculada a la densidad poblacional y condiciones socioeconómicas.
- **Zonas suburbanas y rurales:** Lugares como *Porter Ranch*, *Chatsworth*, y *San Fernando* (*borde del mapa*) muestran una menor actividad criminal debido a su carácter suburbano y más aislado.
- **Siguiente paso:** Combinar esta información con tipos de delito y horarios para identificar patrones más claros.

# Categorización de los códigos de crimen



- Felonías:** Crímenes graves con penas severas (prisión, grandes multas).
- Delitos Menores:** Crímenes menos graves con multas o tiempo en la cárcel del condado.
- Delitos Contra la Propiedad:** Robos y daños materiales, sin violencia directa a personas.
- Delitos Violentos:** Crímenes que implican daño físico o amenaza de daño a las personas.
- Wobbler:** Crímenes que pueden ser procesados como delito menor o felonía, según las circunstancias.

# ¿Cómo varían los crímenes por zona y categoría?

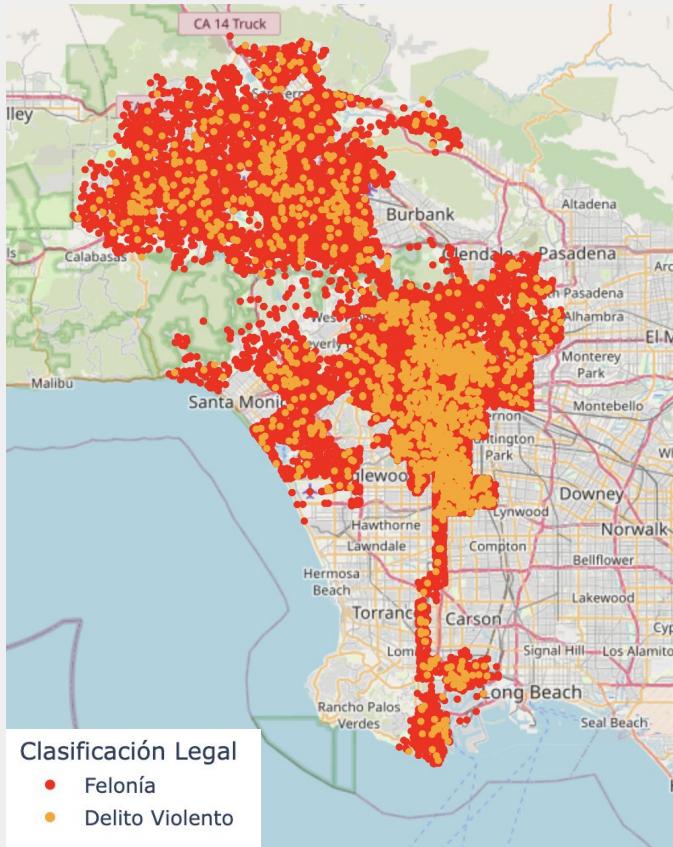


## Clasificación Legal

- Wobbler
- Delito Menor
- Felonía
- Delito Violento
- Otros
- Delito Contra la Propiedad

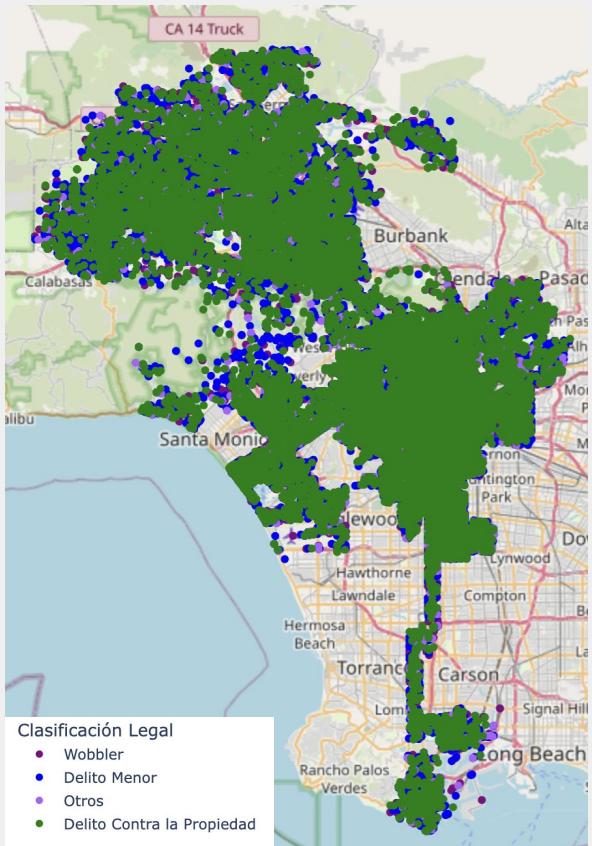
- **Alta densidad de delitos contra la propiedad** (verde): Estos delitos están concentrados en áreas densamente pobladas, como el centro de Los Ángeles y el sur, donde los robos y hurtos son frecuentes.
- **Delitos menores dispersos** (azul): Aparecen en áreas residenciales y comerciales, aunque con menos intensidad que los delitos contra la propiedad.
- **Felonías y delitos violentos** (rojo y naranja): Se ven en todo el mapa pero se concentran en zonas conflictivas del sur de Los Ángeles.
- **Zonas más seguras**: Áreas alejadas del centro y zonas costeras con menor incidencia de delitos graves o violentos.
- **Wobblers dispersos** (púrpura): Estos delitos, que pueden ser procesados como felonía o delito menor, ocurren en áreas comerciales y residenciales por igual.

# ¿Dónde predominan los crímenes violentos?



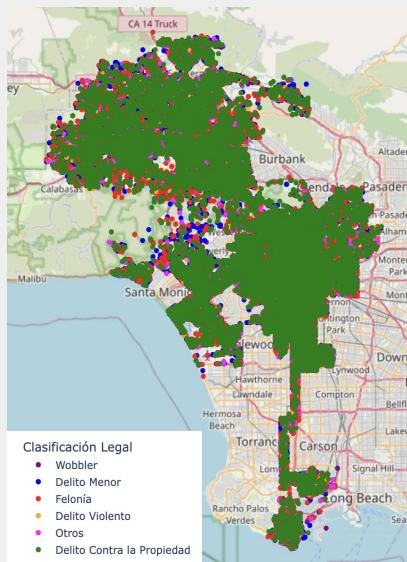
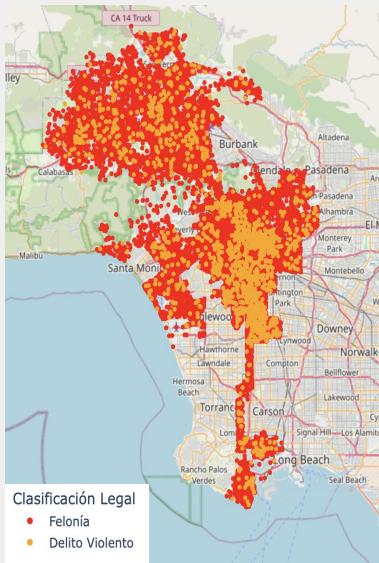
- **Alta concentración en áreas urbanas:** Los crímenes violentos predominan en zonas urbanas densamente pobladas como el centro y sur de Los Ángeles.
- **Felonías predominan:** Los delitos clasificados como felonías (rojo) son más comunes que los delitos violentos menores (naranja).
- **Zonas suburbanas más seguras:** Áreas más alejadas del centro muestran menor densidad de crímenes violentos.

# ¿Cómo se distribuyen los crímenes menos violentos?



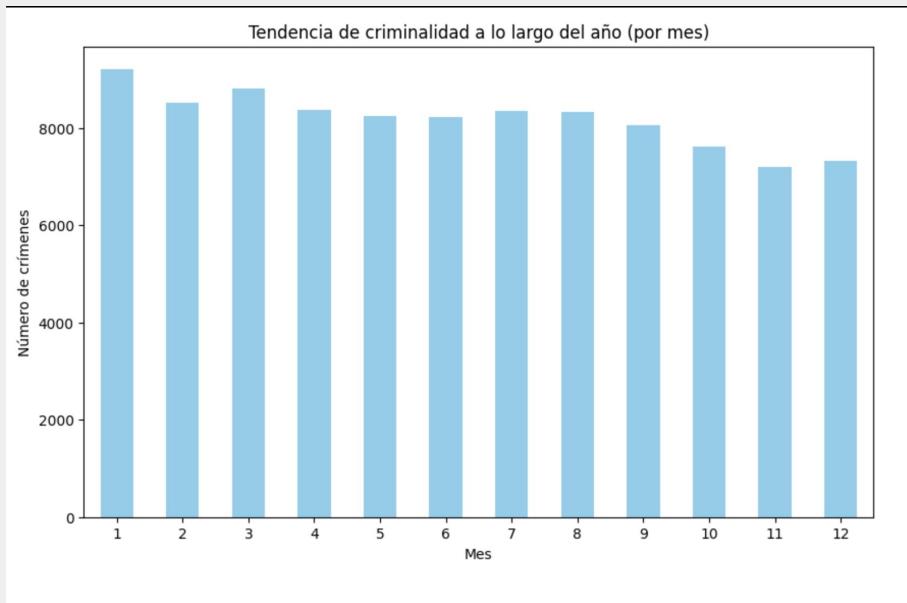
- **Alta densidad de los delitos contra la propiedad (verde):** Están concentrados en áreas urbanas densamente pobladas, como el centro y sur de Los Ángeles, indicando una mayor incidencia de robos, vandalismos y hurtos.
- **Delitos menores dispersos (azul):** Presentes en áreas residenciales y comerciales, pero más dispersos que los delitos contra la propiedad.
- **Wobblers dispersos (morado):** Distribuidos en diversas zonas, ocurren tanto en áreas residenciales como comerciales, sin una concentración específica.
- **Zonas suburbanas más seguras:** Las áreas alejadas del centro muestran una menor densidad de crímenes, especialmente en delitos menores y wobblers.

# ¿Hay zonas que pueden considerarse más violentas?



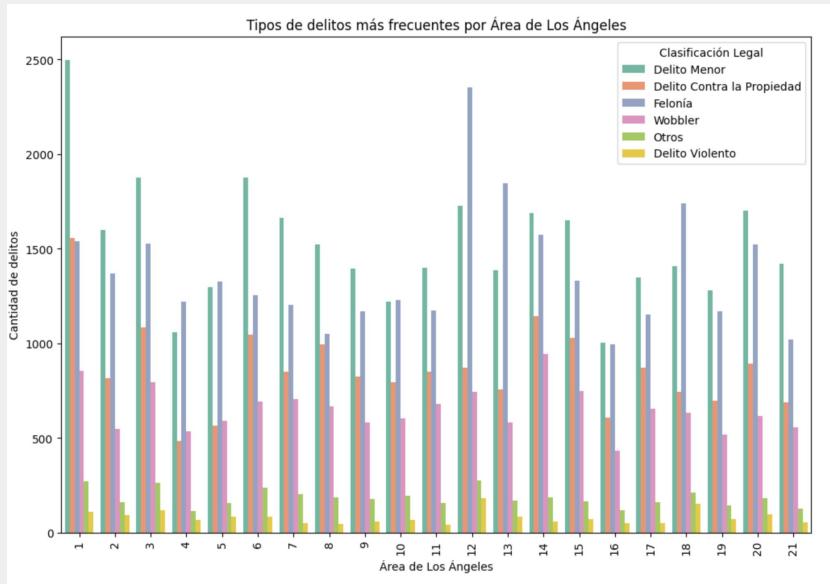
- **Concentración de delitos violentos (rojo/naranja):** Las felonías y delitos violentos predominan en el centro y sur de Los Ángeles, especialmente en zonas densamente pobladas.
- **Delitos contra la propiedad (verde):** Tienen alta densidad, especialmente en áreas urbanas y comerciales.
- **Zonas suburbanas más seguras:** Las áreas alejadas del centro muestran menos incidentes de crímenes violentos y delitos contra la propiedad.

# ¿Cuál es la tendencia de la criminalidad en el año?



- **Alta actividad a inicios de año (enero):** Se observa un mayor número de crímenes reportados durante el mes de enero.
- **Tendencia estable:** A lo largo de los meses siguientes, la criminalidad muestra una tendencia relativamente estable, con pequeñas variaciones en la cantidad de crímenes.
- **Descenso gradual hacia el final del año:** Hay una ligera disminución en la criminalidad en los últimos meses (noviembre y diciembre), lo que podría indicar una menor actividad criminal en el invierno.
- **Estacionalidad baja:** No se observan cambios estacionales drásticos en la criminalidad a lo largo del año, lo que sugiere que los crímenes ocurren de manera constante durante todo el año.

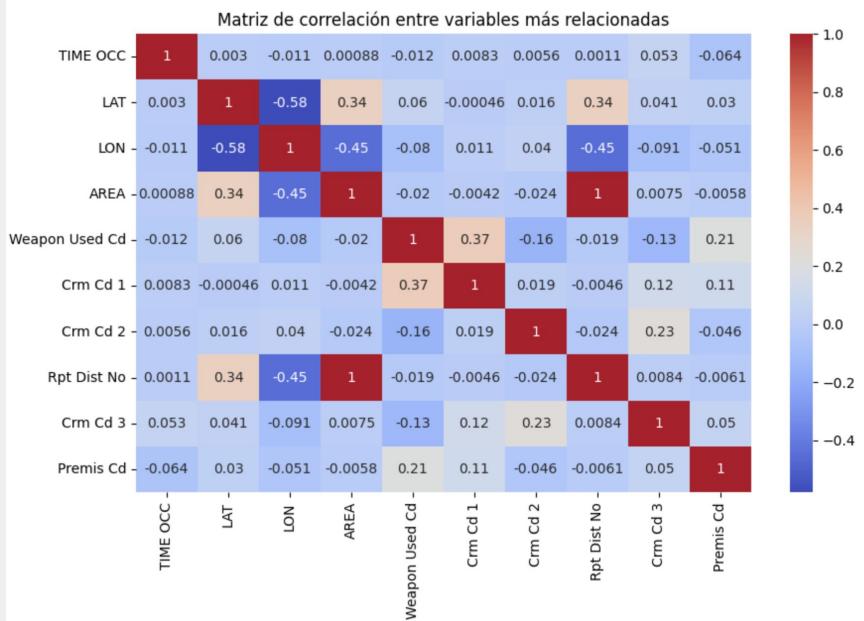
# ¿Qué tipos de delitos son más comunes en determinadas áreas?



- **Felonías (morado claro):** Áreas 12 (77th Street), 13 (Newton), 18 (Southeast) tienen altos niveles de crímenes graves como robos con violencia y homicidios.
- **Delitos Violentos (amarillo):** Áreas 12 (77th Street), 18 (Southeast) y 3 (Southwest) con la mayor incidencia de violencia física.
- **Delitos Menores (verde):** Área 1 (Central), seguida por 6 (Hollywood) y 3 (Southwest), con infracciones como vandalismo y pequeños robos.
- **Delitos Contra la Propiedad (naranja):** Áreas 1 (Central), 14 (Pacific), y 6 (Hollywood) con altos niveles de robos y allanamientos.

El área 12 (77th Street) es la más peligrosa en términos de crímenes graves y violencia, mientras que el área 1 (Central) destaca por delitos menores y contra la propiedad.

# ¿Qué nos revela la correlación entre variables?



## Coordenadas y tiempo:

- **LON** y **TIME OCC** muestran una correlación negativa (-0.45), lo que indica que ciertos crímenes en ciertas longitudes ocurren a horas específicas.
- **LAT** y **TIME OCC** tienen una correlación baja (0.003), mostrando poco impacto de la latitud en el horario de los crímenes.

## Armas y tipo de crimen:

- **Weapon Used** y **Crm Cd 1** muestran una correlación moderada (0.37), lo que sugiere que el uso de armas está relacionado con el tipo de crimen.

## Relación entre ubicación y tipo de crimen:

- **LAT** y **LON** no presentan correlaciones fuertes con el tipo de crimen, lo que sugiere una distribución homogénea de los crímenes en varias áreas.

La baja correlación indica que no podemos basarnos en estas variables para obtener un modelo preciso, pero podemos identificar otros factores que pueden influir en los patrones criminales.

# Conclusiones (Primera Parte)

**1. ¿Qué patrones se pueden identificar en los tipos de crímenes cometidos según la hora del día o el área geográfica?**

**Patrones temporales:** Los crímenes contra la propiedad, como robos y hurtos, son más frecuentes por la tarde y noche en áreas densamente pobladas.

**Delitos violentos:** Suelen ocurrir de noche y madrugada, especialmente en áreas como 77th Street (Área 12) y Newton (Área 13).

**Distribución geográfica:** Áreas como Central (Área 1) y Hollywood (Área 6) tienen más delitos menores, mientras que las zonas suburbanas son más seguras.

**2. ¿En qué zonas específicas de Los Ángeles se reportan más delitos violentos y no violentos?**

**Delitos violentos:** Las áreas 12 (77th Street), 13 (Newton) y 18 (Southeast) destacan por su alta incidencia de delitos graves y violentos.

**Delitos no violentos:** Las áreas 1 (Central) y 6 (Hollywood) tienen altos niveles de delitos menores y contra la propiedad, aunque menos violentos.

# Conclusiones (Primera Parte)

**3. ¿Cuáles son los lugares y momentos con mayor incidencia de delitos en Los Ángeles?**

**Lugares:** Las áreas con más crímenes son 12 (77th Street), 1 (Central), 6 (Hollywood) y 13 (Newton), con altos niveles de criminalidad.

**Momentos:** Los delitos contra la propiedad y menores ocurren más en la tarde, mientras que los delitos violentos son más frecuentes durante la noche y madrugada.

**4. ¿Cuál es la tendencia de criminalidad a lo largo del año o por estaciones?**

**Tendencia anual:** La criminalidad es constante durante el año, con una ligera baja en los últimos meses y un pico en enero y febrero.

**Tendencia estacional:** El invierno y la primavera tienen más crímenes, mientras que el verano y el otoño son más estables.

# Conclusiones (Primera Parte)

**5. ¿Cuáles delitos son más comunes en determinadas áreas?**

**Felonías:** Las áreas 12 (77th Street), 13 (Newton) y 18 (Southeast) destacan por la alta incidencia de crímenes graves.

**Delitos violentos:** Predominan en las áreas 12 (77th Street), 18 (Southeast) y 3 (Southwest), indicando mayor riesgo de violencia física.

**Delitos menores:** Más comunes en las áreas 1 (Central), 6 (Hollywood) y 12 (77th Street).

**Delitos contra la propiedad:** Frecuentes en las áreas 1 (Central), 14 (Pacific) y 6 (Hollywood).

**6. ¿Es posible predecir futuros delitos a partir de un modelo con estos datos ?**

**Baja correlación entre variables:** Las relaciones entre las variables del dataset son débiles, lo que limita su poder predictivo.

**Relación moderada entre uso de armas y delitos:** El uso de armas tiene cierta relación (0.37) con delitos violentos.

**Geolocalización como factor clave:** Aunque la correlación de ubicación es baja, sigue siendo un factor importante para identificar patrones de criminalidad.

Para generar un modelo predictivo **los datos actuales parecen no ser suficientes**, por lo que se sugiere integrar variables adicionales y utilizar modelos más avanzados.

# Algoritmos de Machine Learning (Segunda parte)

# **Eligiendo una nueva pregunta**

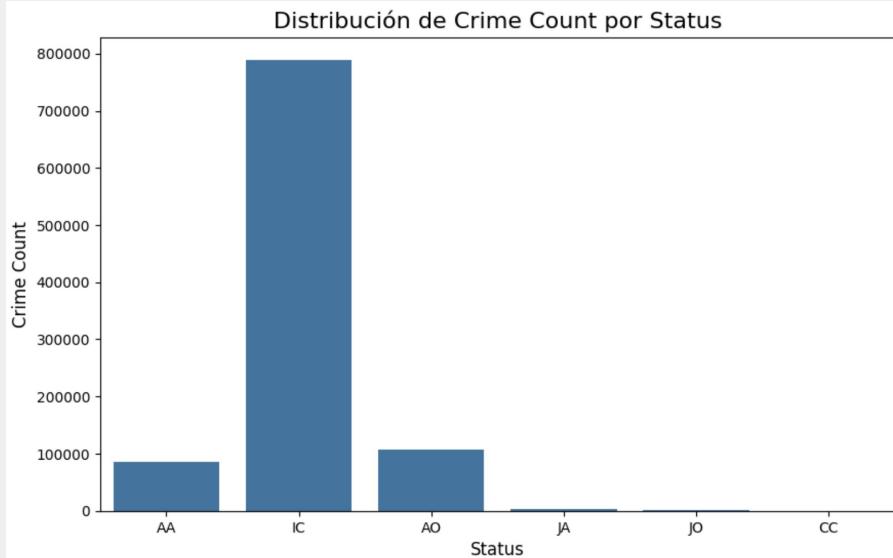
## **Del análisis anterior:**

- Nos dimos cuenta que había poca relación entre las variables de estudio como para poder construir un modelo de calidad.
- Sin embargo tenemos muchas variables que pueden arrojar información de valor para la comunidad.

## **Ahora nos preguntamos:**

- ¿Es posible predecir si un crimen va a ser resuelto o no basándonos en algunas características?
- ¿Cuáles son las características que influyen más en el estado de resolución de un crimen?
- Hipótesis
  - “Es posible clasificar el estado de resolución de un crimen basándonos en sus características”

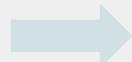
# ¿Cuáles son los status posibles para los crímenes reportados en LA?



Status	Descripción	Cantidad de Crímenes
AA	Adulto Arrestado	85.346
IC	Investigación Continúa	789.360
AD	Otro Adulto	107.087
JA	Arresto Juvenil	3.252
JO	Otros Juveniles	1.821
CC	Unknown	6
NaN	Unknown	-

# Transformemos los datos

Status	Descripción	Cantidad de Crímenes
AA	Adulto Arrestado	85.346
IC	Investigación Continúa	789.360
AD	Otro Adulto	107.087
JA	Arresto Juvenil	3.252
JO	Otros Juveniles	1.821
CC	Unknown	6
NaN	Unknown	-

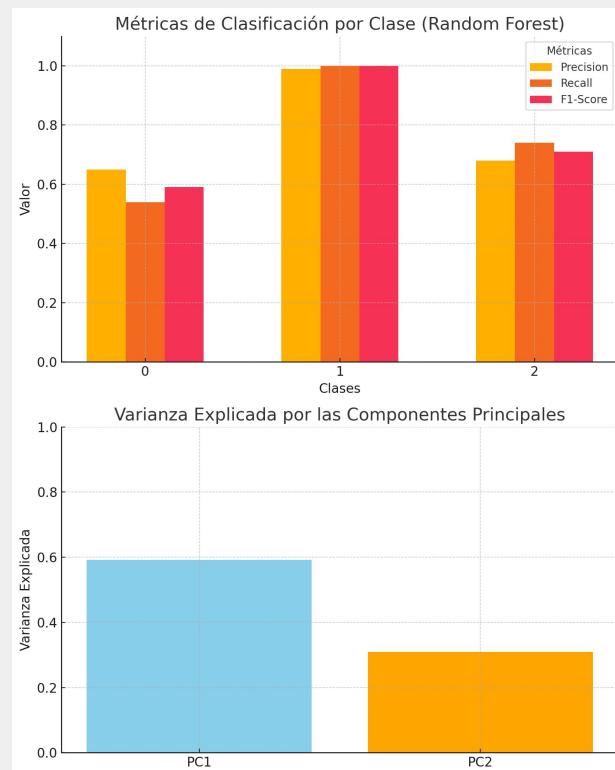


Status Modified	Descripción	Cantidad de Crímenes
AR	Arrestado	85.598
IC	Investigación Continúa	789.360
NA	No Arrestado	108.980

Además, se definieron las variables  
Adult\_crime y Young\_Crime

# Modelo 1: Random Forest con dos componentes principales

Variable	Significado
Part 1-2	Tipo de reporte según el sistema UCR (crímenes menores o graves)
Crm_Cd_Merged	Código del crimen fusionado que prioriza el crimen principal (Crm Cd 1) pero toma Crm Cd si es necesario
Adult_crime	Indica si el crimen está relacionado con adultos (1) o no (0).
Status_modified_encoded	La variable objetivo que indica el estado modificado (arrestado, no arrestado, investigación continua).



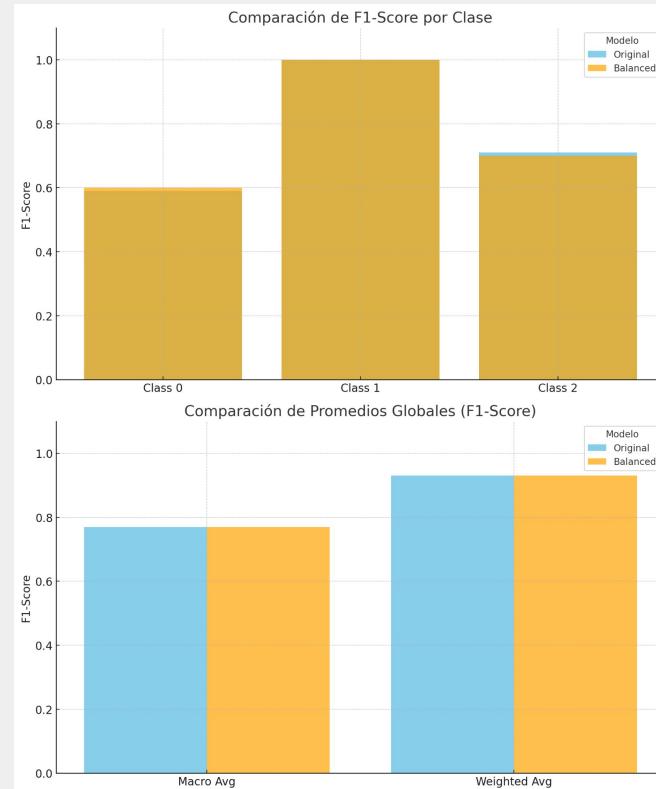
# Modelo 1: Observaciones

## Del análisis anterior:

- La clase 1 domina el dataset (789360 registros), lo que explica su excelente rendimiento pero afecta las predicciones para las clases 0 y 2. Podríamos agregar balanceo para mejorar este rendimiento
- Actualmente, se capturan el 90% de la variabilidad con dos componentes. Agregar una tercera componente podría mejorar el rendimiento sin agregar demasiada complejidad.
- Los F1 score muestran el desempeño moderado del modelo para la clase 0 (0.59) y la clase 2 (0.71) lo que refleja dificultades para clasificar correctamente casos de esa clase.
- El modelo tiene un accuracy de 0.93, lo que significa que predice correctamente el resultado en el 93% de los casos.
- El modelo tiene un 0.78 de macro average, lo que se traduce en que tiene un desempeño del 78% en todas las clases, si se considera cada clase por igual.

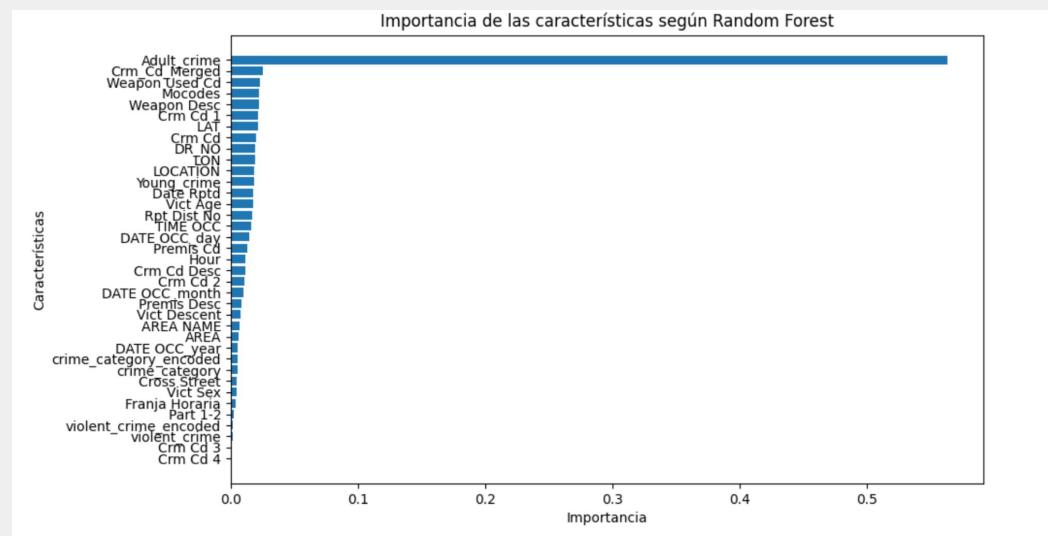
# Modelo 2: Random Forest con balanceo de clases

- Considerando la gran diferencia que hay entre los datos disponibles para cada una de las clases del estudio, introducimos el balanceo de clases para obtener un modelo que considerara la representatividad de cada clase.
- En el gráfico se compara el performance de cada modelo. Se observa poca variación respecto al modelo anterior.
- En este caso, el incluir balanceo al análisis de componentes principales ayuda ligeramente a mejorar el rendimiento pero estas mejoras no son dramáticas, probablemente porque la clase 1 sigue dominando en el dataset.



## Modelo 3: Random forest con 4 PCs y variables seleccionadas.

- Elegimos un nuevo conjunto de variables a través de un análisis exploratorio aplicando random forest, seleccionando las características más importantes y haciendo PCA.
- En el gráfico se puede ver la importancia de las características según random forest.
- Seleccionamos como variables principales y en el modelo elegimos 4 PCs:
  - Adult\_Crime : indicador si es un crimen de adulto)
  - Crm\_Cd\_merged: indicador del código del crimen)
  - Weapon\_used\_cd: indicador del arma utilizada
  - Mocodes Descripción del arma



## Modelo 3: Observaciones

Modelo	Clase 0 (F1-Score)	Clase 1 (F1-Score)	Clase 2 (F1-Score)	Macro Avg (F1-Score)	Weighted Avg (F1-Score)	Accuracy
Random Forest (4PCs)	0.59	1.00	0.71	0.77	0.93	0.93
Random Forest (balanced)	0.60	1.00	0.70	0.77	0.93	0.93
Random Forest (2 PCs)	0.59	1.00	0.71	0.77	0.93	0.93

- Los tres modelos tienen un comportamiento similar, aunque el de balanceo mejora ligeramente el F1-Score respecto a los otros dos.
- Los tres modelos logran una precisión global de 0.93 por lo que el funcionamiento es relativamente bueno.
- El modelo con 4 PCs captura más información del dataset pero no muestra mayor impacto, por lo que si queremos simplicidad el modelo de 2 PCs logra resultados similares a los otros pero más compacto

## Modelo 4: XGBoost

- Este algoritmo basado en árboles de decisión combina modelos simples para generar un modelo final robusto y preciso.
- En este caso lo entrenamos con los datos del modelo basado en 2 PCs. Estos fueron los resultados:

Modelo	Clase 0 (F1-Score)	Clase 1 (F1-Score)	Clase 2 (F1-Score)	Macro Avg (F1-Score)	Weighted Avg (F1-Score)	Accuracy
XGBoost	0.59	1.00	0.71	0.76	0.93	0.93
Random Forest (4PCs)	0.59	1.00	0.71	0.77	0.93	0.93
Random Forest (balanced)	0.60	1.00	0.70	0.77	0.93	0.93
Random Forest (2 PCs)	0.59	1.00	0.71	0.77	0.93	0.93

- El desempeño del modelo basado en XGBoost es muy parecido respecto a los otros dos modelos, solo disminuyendo ligeramente el Macro F1 score.

## Modelo 5: Regresión Logística

- Para seguir con nuestro análisis, utilizamos un modelo de regresión logística, que se caracteriza por su facilidad para interpretar y su balanceo, además de su eficiencia para entrenar y evaluar. En este caso estos fueron los resultados:

Modelo	Clase 0 (F1-Score)	Clase 1 (F1-Score)	Clase 2 (F1-Score)	Macro Avg (F1-Score)	Weighted Avg (F1-Score)	Accuracy
Regresión Logística	0.54	1.00	0.66	0.73	0.92	0.92
XGBoost	0.59	1.00	0.71	0.76	0.93	0.93
Random Forest (4PCs)	0.59	1.00	0.71	0.77	0.93	0.93
Random Forest (balanced)	0.60	1.00	0.70	0.77	0.93	0.93
Random Forest (2 PCs)	0.59	1.00	0.71	0.77	0.93	0.93

- La Regresión Logística muestra valores más bajos en precisión, recall y F1-Score promedio en comparación con los otros modelos. Esto indica que tiene más dificultad para manejar clases menos representadas.

# Conclusiones (Segunda Parte)

- Para nuestros datos, el modelo que resulta preferible para la variable status es el random Forest. Tanto con o sin balanceo arroja desempeños similares y mejores que otros modelos.
- Para que el modelo pueda predecir correctamente, vamos a necesitar los siguientes datos:
  - Part 1-2: Que corresponde al tipo de crimen según el sistema UCR.
  - Cr,\_Cd\_Merged: un código que identifica el tipo de crimen y que se selecciona de una lista pre establecida.
  - Adult\_crime: Un indicador sobre si el crimen involucra adultos o no.
- El modelo internamente divide los datos en múltiples árboles de decisión. Cada árbol evalúa los valores de cada variable y ahí decide como clasificarlas. Al final los árboles votan y se combina el resultado para obtener una predicción final.
- Este modelo entrega al usuario predicciones confiables. Al ser entrando con datos históricos de alta precisión, las variables que se utilizan para generar la clasificación tienen sentido práctico para los usuarios y puede ajustarse a diferentes escenarios según los datos proporcionados.
- En general nuestro modelo obtuvo una precisión del 93%, es decir, acierta 93 de cada 100 casos, lo cual demuestra que es muy bueno clasificando la mayoría de las situaciones

# Conclusiones (Segunda Parte)

- El 7% restante vienen siendo errores, esto puede ser crítico si estas predicciones fallidas ocurren en casos importantes. Es por eso que a pesar de la confianza del modelo, es importante analizar los errores que comete y su causa.
- Este modelo utiliza las variables relacionadas con el tipo de crimen, la gravedad, y si involucra adultos para hacer una predicción. Basándose en patrones históricos, nos indica si es probable que ocurra un arresto, no haya arresto, o el caso quede en investigación continua. Mientras más precisos sean los datos ingresados, más confiable será la predicción.
- Es altamente probable, que una de las causas por la que las clases 0 y 2 hayan sido clasificadas erróneamente en un porcentaje tan alto sea porque hay muchísimos más datos pertenecientes a la clase 1 en los datos del estudio, por ende, una estrategia que podríamos evaluar para mejorar la precisión del modelo es aplicar sobremuestreo, de este modo estaríamos generando de manera consistente datos que equilibren el modelo y puedan mejorar su capacidad de clasificación