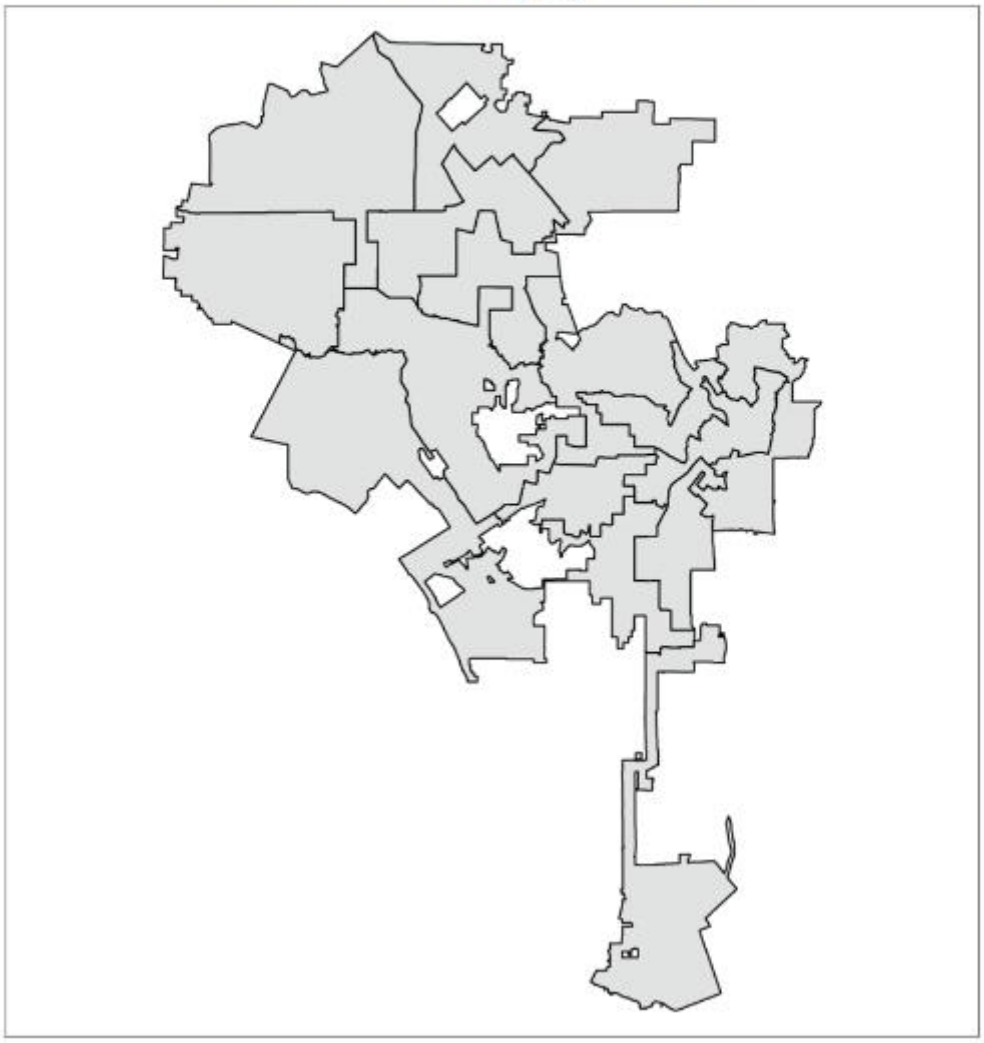
Análisis del crimen en Los Ángeles



Mario Francisco Guillem Fito

3º BIA

Datos Espaciales y Espacio-Temporales

ÍNDICE

1. Introducción

En el presente trabajo vamos a realizar una investigación sobre el crimen en la ciudad estadounidense de Los Ángeles desde 2020 hasta 2023, para ello se llevará a cabo un análisis espacio-temporal. Los datos durante el desarrollo del trabajo se irán relacionando con información social y demográfica que se comentará más adelante.

Antes de empezar a desarrollar el tema, consideramos oportunas una serie de definiciones e hipótesis para el correcto entendimiento del trabajo. En primer lugar, comenzaremos definiendo qué es un dato espacial.

Según M. Fisher y J. Wang, los datos consisten en números o símbolos que, en cierto sentido, son neutrales y, en contraste con la información, casi libres de contexto. Los datos geográficos brutos, como la temperatura en un momento y lugar específicos, son ejemplos de datos. En *Geospatial Analysis* (Longley et al., 2018) podemos ver los datos espaciales como son construidos a partir de elementos o hechos sobre el mundo geográfico. Un dato espacial vincula una ubicación geográfica (espacio geográfico), a menudo un atributo de tiempo, y un atributo (o propiedad descriptiva) de la entidad entre sí. Estos atributos pueden ser de naturaleza ambiental (por ejemplo; la temperatura), económicos (capital de una región concreta), otros pueden identificar una ubicación (direcciones postales). En particular, en el análisis de datos espaciales el tiempo es opcional, sin embargo, la ubicación geográfica es esencial ya que es la que distingue este análisis de otros con datos no espaciales. Para llevar a cabo el análisis de datos espaciales requerimos, al menos, la información de la ubicación y de los atributos, independientemente de cómo se midan los atributos. Este tipo de análisis requiere un marco espacial sobre el cual ubicaremos los fenómenos espaciales que vamos a estudiar.

Durante el estudio se contrastarán las hipótesis de autocorrelación espacial y autocorrelación temporal. La primera nos sugiere que valores temáticos tienden a ser más parecidos entre ubicaciones más próximas en el espacio que entre otros que se ubican más lejos. La autocorrelación temporal nos indica que los datos que son más próximos en el tiempo tienden a ser más parecidos que los que tienen una mayor diferencia temporal.

El tipo de dato espacial a analizar será de datos de patrones puntuales (*Spatial Point Patterns*), es decir, trataremos datos donde las localizaciones son correspondidas a sucesos en diferentes puntos de alguna región de estudio, en nuestro caso la región de estudio será el término de la ciudad de Los Ángeles. A lo largo del análisis, estos datos se agruparán en función de los distritos a los que correspondan para poder realizar análisis por regiones.

Como hemos comentado anteriormente el objetivo de nuestro estudio es relacionar los crímenes con aspectos sociodemográficos más relevantes con los que poder explicar fielmente a la realidad la situación que se está viviendo en Los Ángeles. Dichos aspectos serán:

- **Violencia de género**
- **Racismo**
- **Falta de vivienda y personas sintecho (Homeless)**
- **Desempleo**

De estos 4 aspectos se quiere profundizar especialmente en el último, “Falta de vivienda y sintecho (Homeless)”. Este es el principal problema que está teniendo la ciudad desde hace años. Actualmente y según datos de *Los Angeles Homeless Services Authority* viven en la calle 69.144 personas en todo el condado, habiendo un incremento desde el año de pandemia del 4%.

1.1 Contexto

En estos momentos Los Ángeles está sufriendo un gran aumento en el número de crímenes, alcanzando sus máximos niveles desde hace 15 años, a esta situación se le suma que se ha convertido en la ciudad con mayor número de personas viviendo en la calle de todo Estados Unidos, habiendo sobrepasado a la ciudad de Nueva York. Actualmente, en el condado de Los Ángeles vive en la calle una de cada 150 personas y se estima que el número de personas aumenta diariamente en 20 personas, según datos de *Homelessness in Los Angeles: A unique crisis demanding new solutions (*McKinsey & Company, 2023). Se espera que para el año 2028, año en el que la ciudad albergará los Juegos Olímpicos de verano tenga alrededor de 113.000 personas viviendo en la calle, tal y como se muestra en la siguiente Figura número 1.

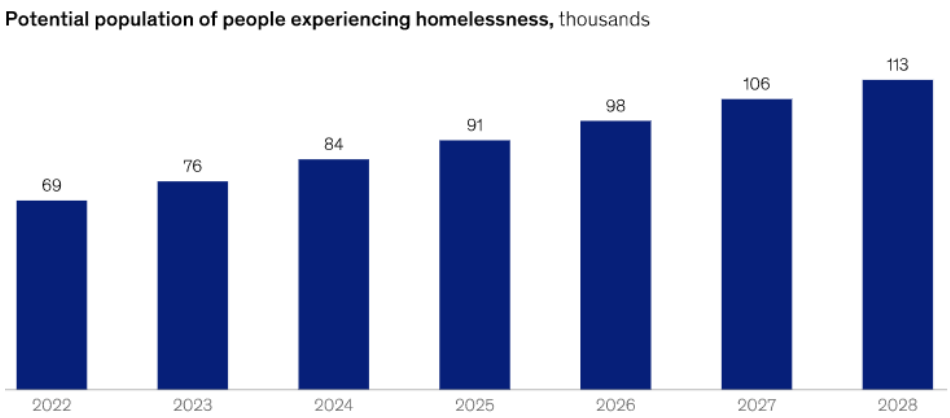


Figura 1 (McKinsey & Company, 2022)

Sin embargo, todas las medidas que se están tomando desde el gobierno parece ser que no son suficientes. Los últimos tres años han sido destinados 500 millones de dólares anuales con el objetivo de frenar dicho problema, pero las medidas no están cumpliendo su cometido.

Por otro lado, también se ha considerado oportuno estudiar el desempleo ya que California es el estado con mayor tasa de todo Estados Unidos, y comprobar si dentro de la ciudad afecta significativamente al crimen. Los otros dos factores sociodemográficos, racismo y violencia de género, han sido elegidos debido a la importancia social.

1.2 Hipótesis

Tras la descripción de los objetivos, se dispone a enumerar las hipótesis de partida del trabajo:

-El número de víctimas de sexo femenino es mayor al masculino.

-Existencia de distritos más peligrosos significativamente a otros.

-Existe de autocorrelación espacial en los crímenes.

-El factor “homeless” afecta al aumento del número de crímenes.

-Las personas de raza negra sufren más crímenes.

-Jóvenes (menores de 30 años) sufren más crímenes que personas más mayores.

-El desempleo afecta al crimen.

Queremos clarificar que los datos con los que se trabajará son de la ciudad de Los Ángeles, no del condado de Los Ángeles. En la siguiente figura se refleja visualmente el término al cual nos referimos.

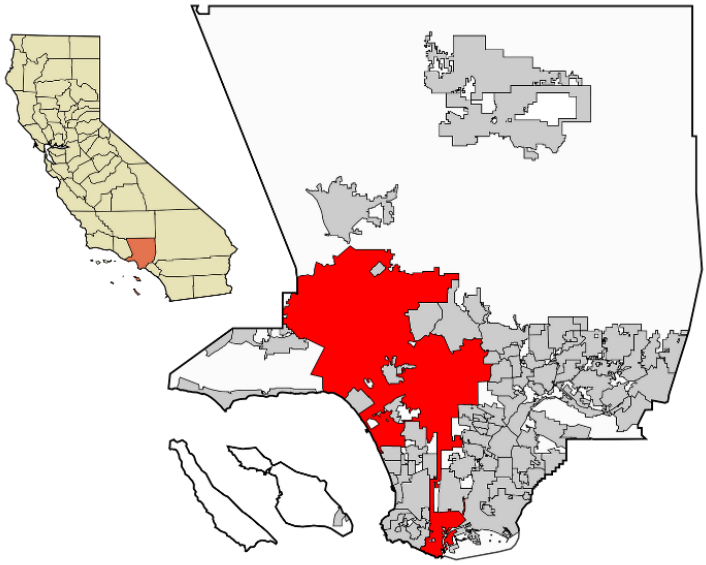


Figura 2

Como podemos ver en la Figura número 2, el color rojo representa la ciudad de Los Ángeles, estando englobada en el condado homónimo. A la izquierda de la figura, vemos el estado al que pertenece (color beige), en este caso se trata de California, ubicado en la costa oeste de los Estados Unidos de América. Dentro del condado, las zonas representadas de color blanco hacen referencia a zonas no urbanizadas como parques o reservas naturales.

2. Metodología

2.1 Datos

Los datos con los que se desarrollará el estudio han sido extraídos del repositorio de datos de la ciudad de Los Ángeles. El nombre oficial de la base de datos es "*Crime Data from 2020 to Present*", pudiéndose descargar gratuitamente desde:

https://data.lacity.org/Public-Safety/Crime-Data-from-2020-to-Present/2nrs-mtv8

Los datos son propiedad del Departamento de Policía de Los Ángeles (LAPD), conteniendo datos desde el 30 de noviembre de 2020 hasta el presente, siendo actualizada la base de datos semanalmente.

Se considera oportuno para poner en contexto describir brevemente los datos que se van a tratar durante el estudio. Como se ha comentado anteriormente los datos escogidos recogen los crímenes ocurridos en la ciudad de Los Ángeles. En dicha base de datos se cuenta a fecha de realización del trabajo 722 mil incidentes denunciados, descritos en 28 columnas.

De las 28 columnas, se procede a describir las 12 variables utilizadas en el estudio de dicha base de datos. Las restantes variables de la base de datos pueden ser consultadas en el enlace anteriormente mencionado.

DR_NO: Número identificativo del crimen formado por 2 dígitos del año, el identificativo del área y 5 dígitos (Variable Numérica)

DATE.OCC: Fecha en la que ocurrieron los hechos (Fecha)

AREA: Identificativo de las 21 áreas geográficas en las que LAPD divide la ciudad de Los Ángeles (Variable Numérica)

AREA.NAME: Nombre asociado al identificativo "AREA". (Carácter)

Rpdt.Dist.No: Número del distrito del área en la que ocurrieron los hechos. (Variable Numérica)

Crm.Cd: Código del crimen (Variable Numérica)

Crm.Cd.Desc: Descripción de "Crm.Cd" (Carácter).

Vict.Age: Edad de la víctima (Variable Numérica)

Vict.Sex: Sexo de la víctima (Carácter)

Vict.Descent: Raza de la víctima (Carácter)

LAT: Latitud

LON: Longitud

Otra base de datos consultada es el proyecto "*Neighborhood Data For Social Change*". Se trata de un repositorio de datos de la University of Southern California. En su web se pueden encontrar datos sobre demografía, educación, empleo, medioambiente, salud y transporte entre muchas otras. De la web se han obtenido datos acerca de las siguientes variables:

-Tasa de desempleo

-Población de personas sintecho

-Población inmigrante (%)

https://la.myneighborhooddata.org/data/

La Universidad agrupa los datos de las variables en barrios, sin embargo, nuestro estudio es mediante las 21 divisiones de la ciudad, es por ello que manualmente se ha realizado una media aritmética de todos los barrios correspondientes a cada distrito, asignando la media a cada distrito.

Por lo que respecta a las geometrías del perímetro de la ciudad, así como de los 21 distritos que la componen han sido extraídas de la página web oficial de la ciudad de Los Ángeles.

https://geohub.lacity.org/datasets/lahub::lapd-divisions/explore?location=34.019778%2C-118.410104%2C11.29

También se han tenido en cuenta para realizar comprobaciones una geometría diferente en la que se segmentaba la ciudad en todos los distritos que la componen (5.313). El archivo se encuentra disponible también en la web oficial de la ciudad, enlace:

https://geohub.lacity.org/datasets/39b404bd22804807ba0f0e1628e585f2/explore

2.2 Técnicas

Una vez obtenidos los datos y con el objetivo de analizarlos se han utilizado técnicas de análisis espacial y espaciotemporal mediante el lenguaje de programación con enfoque estadístico R, a través de su IDE denominado "RStudio".

En primer lugar, se ha realizado un **Análisis Exploratorio de Datos** (EDA) en el que a partir de funciones de *Rbase* y diversas librerías como *lubridate* (Garrett Grolemund, Hadley Wickham 2011), o *tidyverse* (Wickham H et al., 2019) se han podido obtener este primer análisis.

Mediante la función *view*() y *summary*() se ha podido ver a simple visto los valores que tomaban las diferentes variables de la base de datos. Gracias a estas dos funciones el primer paso que se realizará será convertir los valores de una serie de variables en factores, dichas variables son: *Vict.Sex*, *Vict.Descent*, *Vict.Age*, *AREA.NAME*, *Crm.Cd*, *Crm.Cd.Desc*.

Cabe destacar que para un entendimiento más simple de los datos se ha recodificado la variable "*Vict.Descent*", cambiando así la codificación de Los Angeles Police Department a un diccionario propio. A continuación, se enuncia el diccionario que se ha seguido para remplazar la codificación inicial.

A	Other Asisan	**B**	Black	**C**	Chinese
D	Cambodian	**F**	Filipino	**G**	Guamanian
H	Hispanic/Latin/Mexican	**I**	American Indian/ Alaskan Native	**J**	Japanese
K	Korean	**L**	Laotian	**O**	Other
P	Pacific Islander	**S**	Samoan	**U**	Hawaiian
V	Vietnamese	**W**	White	**X**	Unknown
Z	Asian Indian				

Elaboración propia

Figura 3.

Una vez con los datos listos, se ha llevado a cabo una tabla mediante la librería *gt* (Iannone et al., 2022) en la que se ha establecido el número total de crímenes en Los Ángeles por años, desde 2020.

Otro aspecto que se va a tener en cuenta es el género de las víctimas del crimen. Se ha diferenciado el sexo del tipo de crimen para saber la cantidad total de crímenes sobre cada tipo de sexo. Para este proceso se ha utilizado la función *filter*() y para su posterior ilustración en un gráfico se ha utilizado la librería externa a R denominada ggplot2 (H. Wickham, 2016). El gráfico escogido es un diagrama de barras. En el análisis se han eliminado NA's y la variable "X" la cual indicaba el desconocimiento del sexo de la víctima.

Otra variable sociodemográfica que se va a analizar es la edad de las víctimas (*Vict.Age*). La edad se ha tomado como una variable discretizada, es por ello por lo que se ha realizado para su análisis otro diagrama de barras para medir como se distribuyen las edades de las víctimas. En la construcción del gráfico se ha añadido visualmente el valor medio de la edad de las víctimas. Por otro lado, se ha filtrado la población entre menores de 30 y mayores de 30 para poder estudiar la hipótesis relacionada con la edad.

La siguiente variable que se va a tener en cuenta en este análisis exploratorio de datos es la raza de la víctima (*Vict.Descent*), para poder establecer una relación con el racismo. Para la

visualización de la distribución del tipo de raza se utilizará un gráfico circular también creado con ggplot2.

El siguiente factor que se ha tenido en cuenta ha sido agrupar el número de crímenes por áreas, obteniendo así las áreas con mayor número de crímenes. Así como generar una ordenación de los crímenes más frecuentes en la ciudad de Los Ángeles tanto por un gráfico de barras como representando en una escala de colores en el mapa de Los Ángeles los crímenes.

A continuación, se dispone a comentar las técnicas utilizadas relacionadas con el **análisis espacial.**

Se han realizado estudios de la densidad de los crímenes por zonas, generando mapas de densidad y mapas de calor mediante *ggplot2* y *leaflet* (Cheng J et al., 2022).

Otra técnica utiliza ha sido la de la densidad KDE. La **estimación de densidad de Kernel** (KDE) es un método que logra estimar la función de densidad de probabilidad de una variable aleatoria, en nuestro caso, la concentración de los números de crimen a partir del área. El KDE nos permite calcular la densidad de los crímenes alrededor de cada uno de ellos. Este método está basado en la función Kernel cuártica (Silverman (1986, p. 76, ecuación 4.5)). La fórmula para calcular la densidad sería la siguiente:

$$Density = \frac{1}{(radius)^2}\sum_{i=1}^{n}\left[\frac{3}{\pi}\cdot pop_i\left(1-\left(\frac{dist_i}{radius}\right)^2\right)^2\right]$$

$$\text{For } dist_i < radius$$

i = 1,2,3,…,n son los patrones puntuales de entrada, es decir, un crimen en concreto.

Pop es el valor de campo de población del punto i, este parámetro es opcional.

Dist será la distancia entre el punto i y la ubicación (x,y).

Una vez obtenida la densidad, esta se multiplica por la cantidad de patrones puntuales (n). Por otro lado, el radio de búsqueda está fijado a la siguiente fórmula:

$$SearchRadius = 0.9 * \min\left(SD, \sqrt{\frac{1}{\ln(2)}} * D_m\right) * n^{-0.2}$$

Dm es la distancia mediana (ponderada) desde el centro medio (ponderado).

n es el número de puntos si no se usa ningún campo de población, o si se proporciona un campo de población, la suma de los valores de campo de población.

SD es la distancia estándar.

Otro método empleado ha sido la utilización de **mapas de calor**. Un mapa de estas características ayuda a identificar la distribución espacial de los datos en un simple vistazo. Cuando existe una gran concentración de datos en una zona, en el caso de estudio sería crímenes por área, se colorea en una escala de azul a rojo pasando por el color amarillo en función de la intensidad. Un problema que puede ocasionar los mapas de calor es la exactitud de los datos, es decir, no

permite reconocer a simple vista el punto exacto o diferenciar por zonas concretas los crímenes. Es por ello por lo que se ha empleado otro método de graficación distribuyendo los crímenes según las áreas en los que han tenido lugar, esta es otra opción que permite diferenciar fácilmente los distritos que más o menos crímenes han ocurrido en sus calles.

También se realizarán mapas en función del número de crímenes respecto a sus distritos.

La siguiente técnica a realizar es el test de Moran sobre diferentes variables. El **test de Moran** (Moran, P. (1950)) es un test global de autocorrelación espacial. Esta prueba permite obtener el denominado Índice de Moran (I)

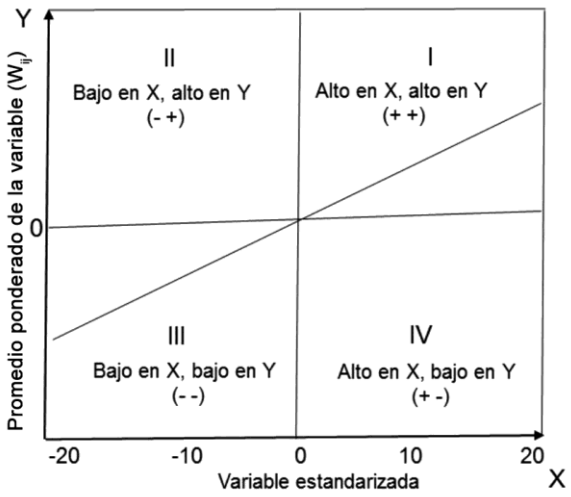
$$I = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Wij, representará el peso espacial entre la observación i y la observación j. Las variables de interés serán yi e yj.

Los valores que puede tomar el Índice serán entre -1 y 1:

- Si el Índice de Moran es 0, se dice que los datos se distribuyen al azar.
- Si es positivo sí que existe concentración
- Si es negativo, existe una excesiva dispersión, siendo aún mayor que

La hipótesis nula de partida será la nula autocorrelación espacial, es decir, se distribuyen aleatoriamente entre el espacio. Para una visualización de esta prueba se elaboró el Diagrama de dispersión de Moran:



Para obtener el Índice de moran y su posterior Diagrama utilizaremos la función *moran.test*() y *moran.plot*() respectivamente sobre las variables de desempleo, los casos totales y la población homeless. Ambas funciones pertenecen a la librería *spdep* (Bivand, R. 2022).

Otro método utilizado para comprobar la autocorrelación espacial ha sido Local Indicators of Spatial Association (**LISA**) (Anselin, Luc. 1995.). Este método permite identificar patrones locales de asociación espacial, descomponiendo e propio Índice Moran con la intención de evaluar la influencia de ubicaciones individuales exactas en los datos globales. Este índice se encarga de representar aquellos distritos con valores significativos en el número total de crímenes,

permitiendo localizar así puntos calientes o “hot spots”, cuya intensidad dependerá de la significativa presentada por los datos estudiados. Este análisis se basa en el test de moran local que representa las localizaciones. Para la obtención de este método ha sido empleada la librería *spdep* (Roger S. Bivand et al., 2013), de la cual se ha utilizado la función *localmoran*(). A partir de dicha función y una serie de ajustes se ha logrado una visualización gráfica del resultado obtenido.

2.3 Modelos

Para la generación de modelos se han generado de dos tipos **GLM** y **GAM**.

El método de GLM, **Modelos Lineales Generalizados**, engloba una gran variedad de regresiones. En nuestro caso se va a usar la denominada Regresión de Poisson.

La distribución de Poisson es una de las principales distribuciones de variables discretas. Su aplicación hace referencia a la modelización de diversas situaciones en las que interesa conocer el número de hechos que pueden llegar a suceder en un intervalo de tiempo (de 0 a infinito).

Función de cuantía: $P(x) = P(X = x) = \frac{e^{-\lambda}\lambda^{x}}{x!}$

Función de distribución: $F(x)=P(X \leq x)=\sum_{i=0}^{x} \frac{e^{-\lambda}\lambda^{i}}{i!}$

Función generatriz de momentos: $\varphi(t)=e^{\lambda(e^{t}-1)}$

Media: $\mu=E(x)=\sum_{x=0}^{\infty} x\frac{e^{-\lambda}\lambda x}{x!} = \varphi'(t=0)=\lambda$ Varianza: $\sigma^2 = E\left[(x-\mu)^2\right]=\lambda$

Moda: $\lambda-1 \leq \text{moda} \leq \lambda$

Para la generación de un modelo lineal generalizado se necesitan de 3 componentes esenciales:

-**Componente aleatorio**: la variable de respuesta Y. La distribución de Y ha de pertenecer a la Familia exponencial. En nuestro caso será el **número de crímenes en cada distrito**.

-**Componente sistemático**: las variables predictoras o covariables

-**Función link**: Función matemática que se encarga de linealizar la relación establecida entre la variable de respuesta Y y las variables predictoras mediante la transformación de la variable de respuesta.

Modelo Aditivo

En este modelo las covariables pueden tener una expresión más general que en los modelos de regresión lineal múltiple, esto es debido a la utilización de funciones que logran una mayor flexibilidad en el modelo. Sin embargo, se debe exigir:

-Normalidad en los residuos

-Homocedasticidad

-La variable Y debe tener una distribución normal.

Los **Modelos Aditivos Generalizados** como su propio nombre indica permiten generalizar los modelos aditivos. Usando los modelos aditivos generalizados se permite que la variable de

respuesta Y pueda tener una distribución diferente a la normal. Para la generación de modelos se agruparán los patrones puntuales en las 21 divisiones establecidas por el departamento de policía de Los Ángeles.

Como variables predictoras han sido utilizadas las provenientes del proyecto "Neighborhood Data for Social Change", de la University Southern California. De dicho proyecto se han obtenido datos sobre:

- Población inmigrante (%)
- Población de personas sintecho (Homeless)
- Tasa de desempleo (%)

Se han realizado **3 modelos**

Un Modelo Lineal Generalizado, como variable respuesta los casos totales y como predictores la población homeless, la población inmigrante y la tasa de desempleo.

Por otro lado, se han realizado **3 GAM's**

1. GAM: En el que analizamos el efecto espacial sin covariables
2. GAM en el que como covariables se establecen el efecto espacial y la población sintecho.
3. GAM usando como covariables los ingresos, el desempleo y el precio del alquiler.

Para comprobar la efectividad del modelo se tendrán en cuenta los valores ajustados, los residuos y el R^2, este nos explica el porcentaje de varianza de la variable dependiente explicado por la variable independiente.

Las librerías utilizadas para la generación de modelos GAM han sido:

- *gamm4* (Wood S, Scheipl F, 2020)
- *mgcv* (Wood S, 2017)

3. Resultados

Tras haber analizado los datos se procede a comentar los principales resultados obtenidos. En primer lugar, se mostrará la evolución del total de crímenes en la ciudad de Los Ángeles.

Número de crímenes en Los Ángeles

Año	Nº de casos
2020	199008
2021	208840
2022	232083
2023	37974

Elaboración propia

Figura 4.

Se puede observar como en el año 2020 se rozaban los 200 mil casos anuales, pero los años venideros han seguido una tendencia alcista por lo que respecta al número de crímenes. En el 2021 tuvieron lugar cerca de 9 mil casos más que el año anterior. Por lo que respecta al 2022, los crímenes han aumentado un 16.62% respecta al 2020. Por último, debido a que se cuentan con datos del 2023 y aún no se ha finalizado el año, no se pueden extraer grandes conclusiones a partir de esos datos.

El siguiente gráfico a comentar hace referencia a la distribución de género de las víctimas de los crímenes. Para la elaboración de este se han eliminado los valores "X", estos valores indicaban el desconocimiento del sexo de la víctima.

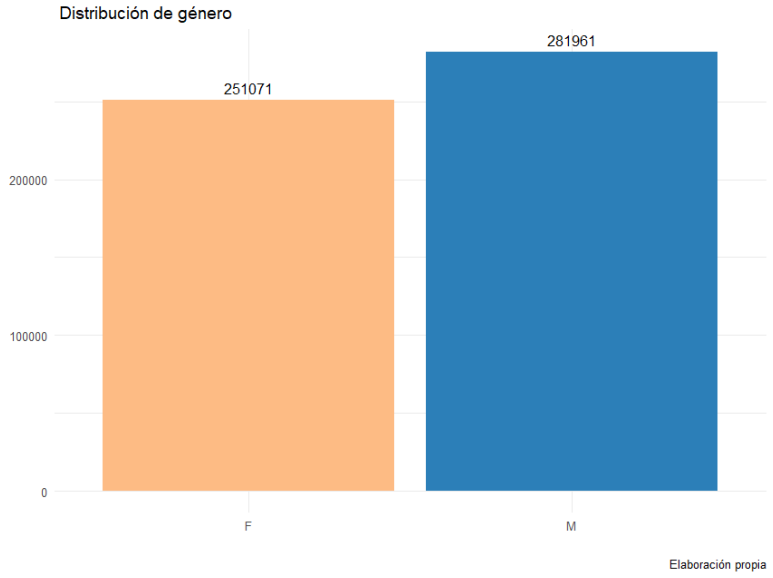


Figura 5.

Se puede observar en color crema la cantidad total de mujeres que han sufrido un crimen, siendo un total de 251.071, mientras que los hombres han sido víctimas de un total de 281.961 crímenes. Estas cantidades han sido obtenidas a partir del periodo desde 2020 hasta 2023.

Seguidamente se va a analizar la distribución de la edad. La edad se ha estudiado para conocer la aceptación o rechazo de la hipótesis de partida del trabajo: "Jóvenes (menores de 30 años) sufren más crímenes que personas mayores".

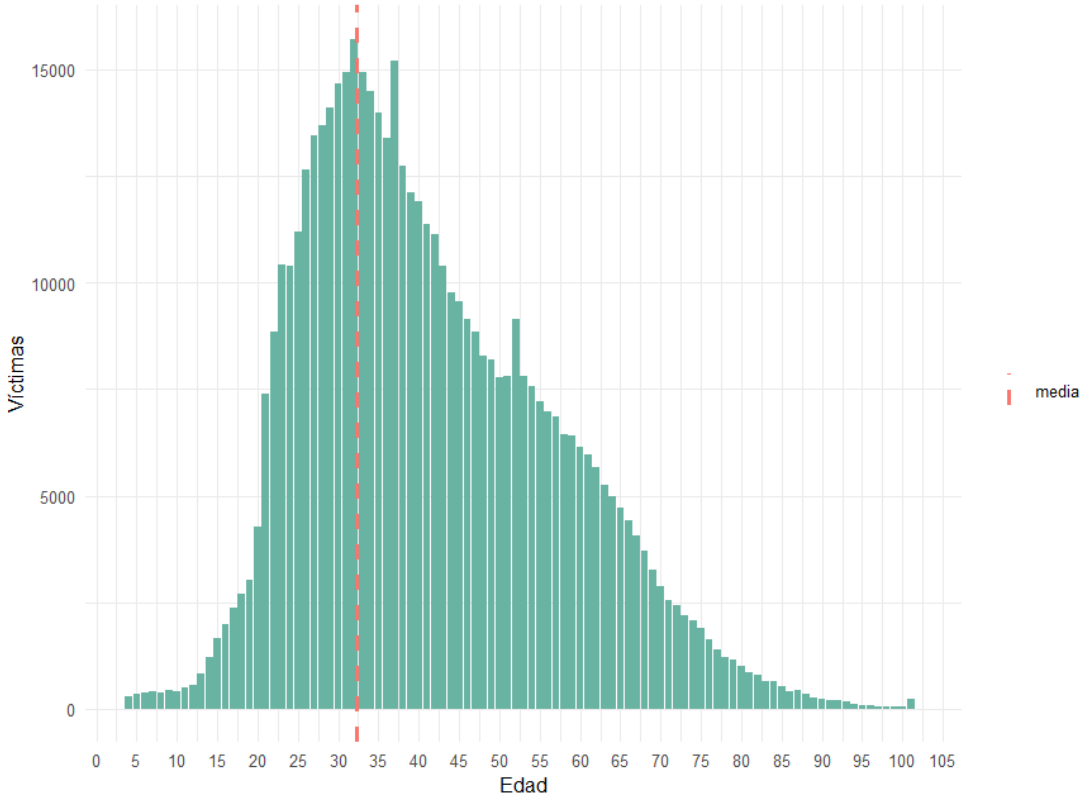


Figura 6.

Se puede observar en el gráfico superior como se distribuye la edad. Siendo la media de las víctimas de 32,23 años y la mediana de 33 años. Por otro lado, los jóvenes menores de 30 años son un total de 287.936, mientras que los mayores de 30 son 375.290 personas.

La siguiente variable de estudio es la etnia a la que pertenecían las víctimas. Su representación se ha llevado a cabo mediante un "pie chart".

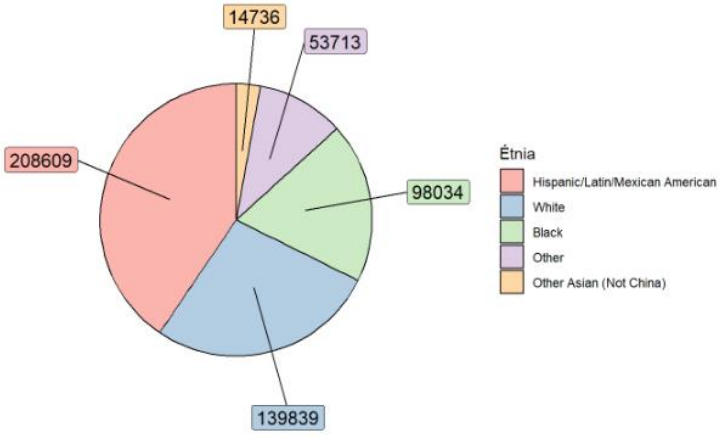


Figura 7.

Tal y como representa la imagen superior, la clase de hispanos, latinos y mejicanos americanos son los que más crímenes sufren, siendo un total de 208.609; el segundo grupo con mayor volumen serían los americanos blancos 139.839, y el tercer grupo con mayor representación serían las personas de raza negra con 98.034 personas.

A continuación, se van a comentar los resultados obtenidos sobre la autocorrelación espacial. Para determinar dichos resultados, tal y como se ha comentado anteriormente en la metodología, se ha llevado a cabo el denominado test de moran. Se considera oportuno poder visualizar los casos registrados en la ciudad de Los Ángeles.

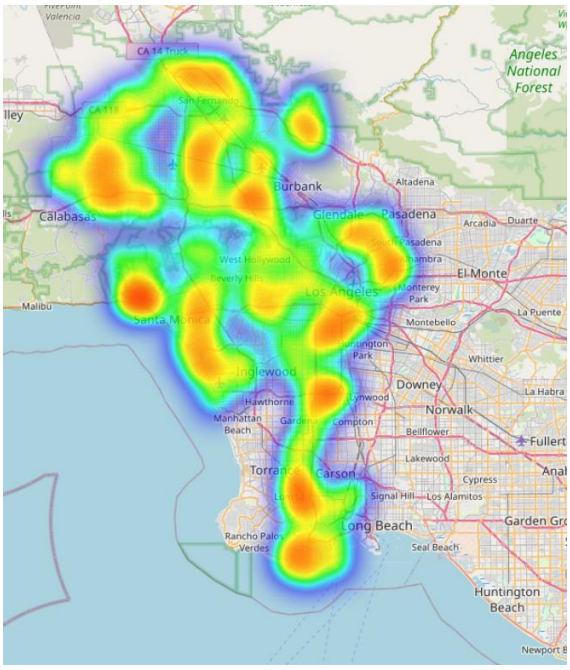


Figura 8.

Mediante la Figura número 8 podemos ver como se distribuyen los crímenes por toda la ciudad de Los Ángeles mostrando una clara concentración de casos en diferentes puntos concretos de la ciudad.

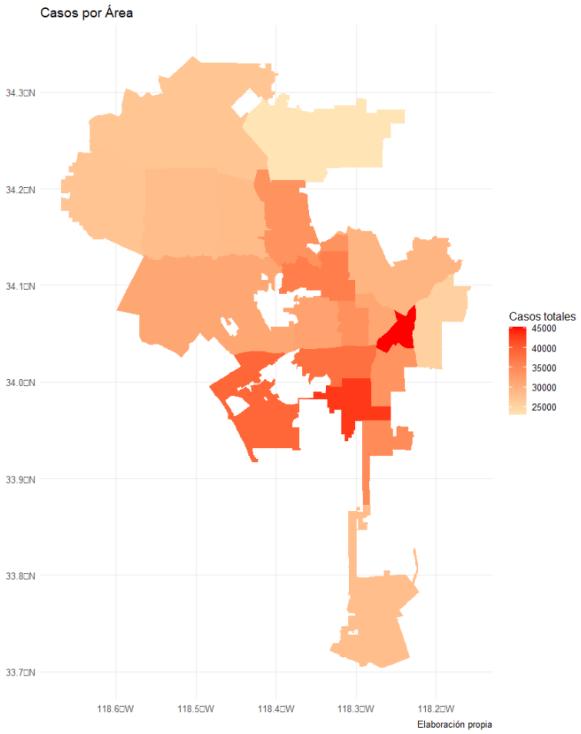


Figura 9.

Para clarificar los datos proporcionados en la figura número 8, se ha generado mediante ggplot un gráfico en el que poder mostrar el número exacto de casos por distrito. Central es el distrito con mayor número de casos acontecidos con un total de 45198 casos, mientras que el distrito que menos casos han tenido lugar entre 2020 y 2023 es la exclusiva zonal residencial de Foothill. A Central le siguen el barrio de 77th Street, y el distrito de Pacific en el cual se ubica el aeropuerto.

Una vez vista la distribución se procede a comentar los resultados obtenidos a partir de los diferentes test de Moran realizados.

Para realizar el test global de Moran sobre los casos por distrito. Se parte de la hipótesis nula de ausencia de autocorrelación espacial. Tras realizar dicha prueba, observamos que el p-valor obtenido es prácticamente 0, por lo que se decide rechazar la hipótesis nula y se confirma la existencia de autocorrelación espacial. Por lo que podemos afirmar que existe una correlación mayor de la que cabría esperar si las observaciones se hubieran repartido aleatoriamente por el espacio.

P-valor	**0,002**
Moran I statistic	**0,24**

Se ha realizado también el mismo test global de Moran pero la geometría escogida ha sido de barriadas, en lugar de distritos. Siendo un total de 1135 barrios, en lugar de las 21 divisiones. Para este mismo test también se ha obtenido un p-valor prácticamente 0 y un Índice de Moran también positivo.

Además, tras realizar el test de Moran, seguidamente se ha realiza una representación gráfica la autocorrelación espacial con un mapa de clusters LISA.

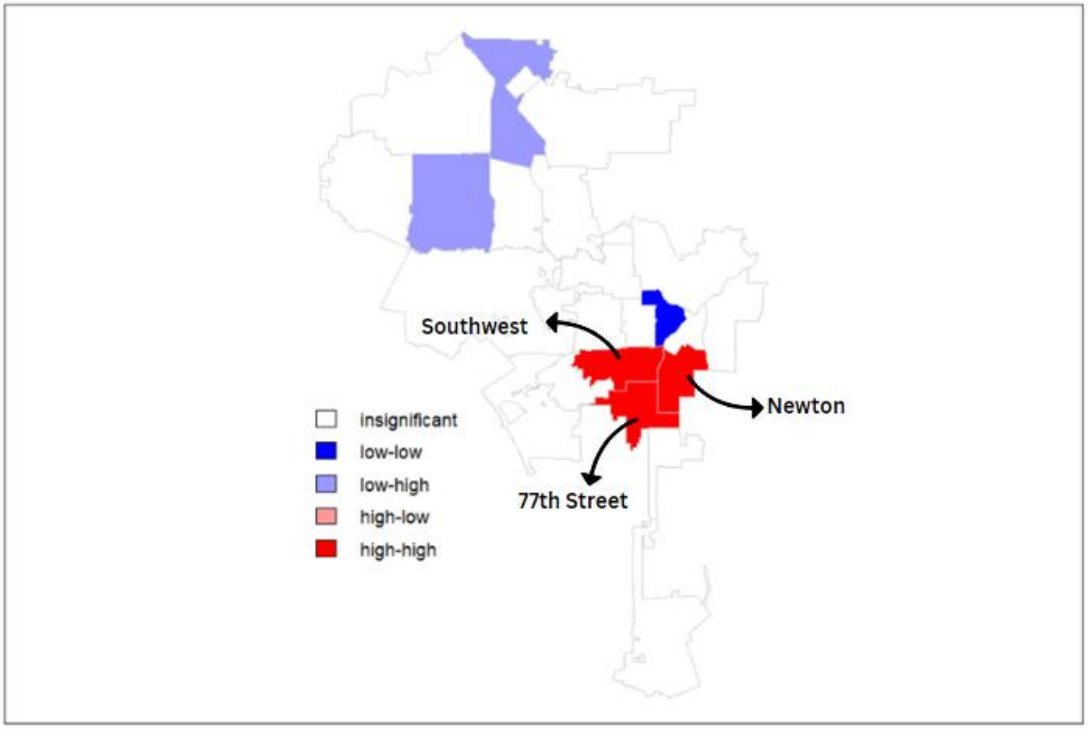


Figura 10.

Se puede observar un cluster de distritos con valores “high-high”, los cuales son las divisiones de 77th Street, Newton y Southwest. Como puntualización destaca la división de Rampart la cual es la única con “low-low” estando a a la izquierda de Central que es la división con mayor crimen de la ciudad.

Antes de comentar los modelos realizados en la metodología, se han realizado test de Moran sobre las variables de población homeless y el desempleo.

Desempleo	
P-valor	0,004858
Moran I statistic	0,3407

Homeless	
P-valor	0,59
Moran I stati	-0,0598

Por lo que respecta a la variable de desempleo, se puede decir que se rechaza la hipótesis nula y existe autocorrelación espacial. Sin embargo, por lo que respecta a la población Homeless no se puede rechazar la hipótesis nula y al ser menor que 0 se puede decir que existe una dispersión mayor de la que se tendría que esperar si los datos se distribuyeran al azar. Sin embargo, ha resultado sorprendente dicho resultado y dado que el Índice ha resultado negativo se ha considera incluir el efecto espacial en la creación del modelo.

A continuación, se procede a comentar los modelos obtenidos con los que obtener las conclusiones. Para tener en cuenta el factor espacial y las diferentes covariables para obtener el número de casos, se han realizado 3 Modelos Aditivos Generalizados.

El primero modelo GAM se ha realizado teniendo solo en cuenta el **efecto espacial** obteniendo los valores ajustados de la Figura inferior número 11. El modelo presenta un R^2 ajustado del 0.297

El siguiente modelo se ha tenido en cuenta el **desempleo y el efecto espacial**. Como resultado se obtuvo que dicha variable resultó significativa a la hora de ajustar los datos. Además, el R^2 ajustado fue de 0.756.

El último modelo ajustado fue en función de la **población sintecho**, en la que se puede ver que dicha variable es significativa además del efecto espacial. El R^2 ajustado fue de 0.896.

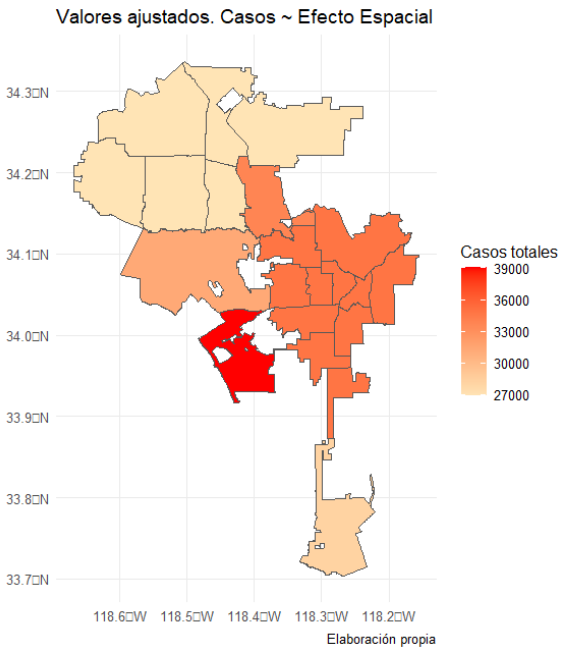


Figura 11.

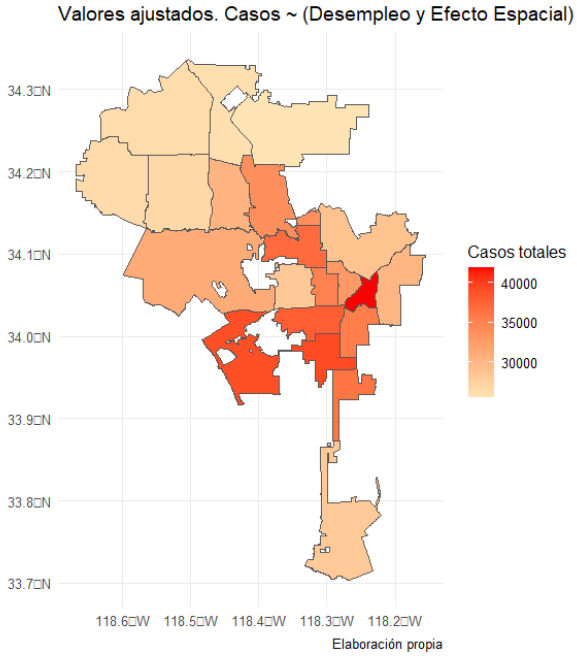


Figura 12.

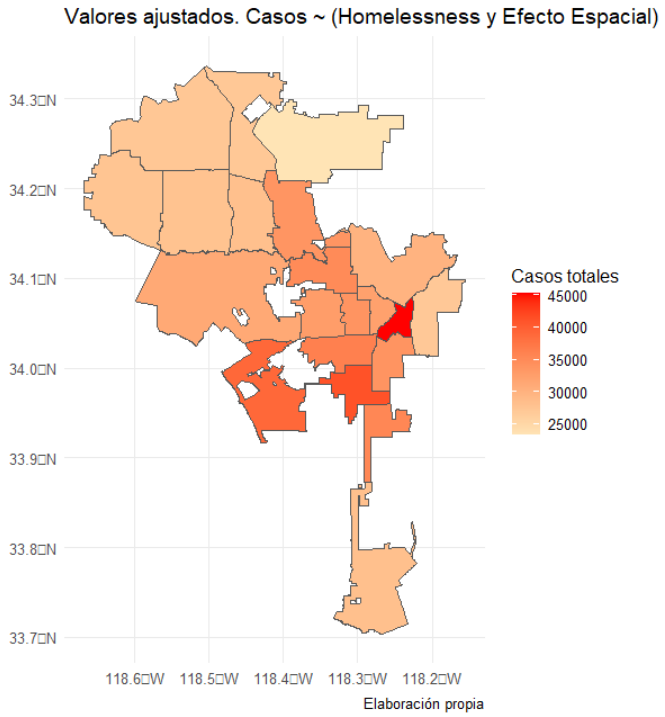


Figura 13.

Aunque a simple vista pueda parecer que el modelo que ajuste mejor es el segundo modelo, (ajustando los casos en función del desempleo y del efecto espacial), hay que fijarse en la escala para ver que es menor y proporciona valores inferiores a los deseados. Por consiguiente, el modelo que mejor ajusta los datos es el modelo aditivo generalizado que modela los casos en función de las personas sintecho en las diferentes divisiones, teniendo en cuenta el efecto espacial, proporcionando de los 3 modelos GAM el mayor R^2. Cabe destacar que los residuos de los modelos del desempleo y de homeless son significativos por lo que existen factores que se desconocen que están explicando gran parte del modelo.

4. Conclusiones / Discusiones.

Los primeros resultados a comentar hacen referencia a la primera hipótesis de partida del trabajo, es decir, "**Hay más víctimas mujeres que hombres**". Sin embargo, se tiene que **rechazar** la hipótesis nula puesto que los hombres representan un mayor porcentaje de las víctimas totales que las mujeres.

Cabe destacar que el **21.83%** (54.828) de las denuncias de las mujeres han sido por **violencia de género**, englobando desde agresiones de pareja, violaciones, insultos, secuestros e incumplimiento de la orden de alejamiento de sus parejas. A modo de comparación, la ciudad de Madrid contando con una población de 3.23 millones frente a las 3.85 millones en Los Ángeles, se han registrado 77.230 denuncias; siendo un total de más de 22 mil denuncias que en los Ángeles. Los datos de Madrid engloban el mismo período (2020-2023).

La siguiente hipótesis de partida que ha tenido que ser rechazada ha sido la de "**Jóvenes (menores de 30 años) sufren más crímenes que personas más mayores**". Tras analizar la distribución de las edades de las víctimas y su posterior comentario donde se citaban el total de

víctimas jóvenes **se rechaza la hipótesis**, puesto que el número de personas con más de 30 años supera al de menos de 30 años. 375.290 victimas frente a 287.936 respectivamente. Esta hipótesis de partida se planteó a partir de un estudio de la ONU (Global Homicide Report, 2013) en el que se citaba que casi la mitad de las víctimas de crímenes su edad estaba comprendida entre 15 y 29 años, pero en el caso de Los Ángeles se puede observar que no se da el caso.

La tercera hipótesis de partida "**Personas de raza negra sufren más crímenes**" también ha de ser **rechazada**, la etnia que sufre un mayor número de crímenes son los latinos, americanos de origen mejicano o hispanos; representando un total del 40.51% de las denuncias recibidas por el departamento de policía de Los Ángeles. Respecto a las personas de raza negra, han sido victimas del 19.03% de las denuncias. La segunda etnia que más víctimas recoge son los blancos americanos con un total del 27.15%, por lo que no se puede establecer.

Por otro lado, y tal y como se anuncia en un estudio del "Pew Research Center" (Noe-Bustamante, 2022), el 32% de las personas negras de Los Ángeles sienten un mayor miedo a ser discriminados por delitos de odio debido a su raza, sin embargo, la raza más discriminada diariamente son los asiáticos, sintiéndose ofendidos por algún acto de la sociedad el 61% de los orientales. Utilizando este estudio y los datos disponibles del Departamento de Policía una conclusión que se puede extraer es que las personas negras y asiáticas se sienten más odiados, pero denuncian menos los hechos ante la policía.

Una vez analizada la Figura número 9 podemos decir que sí que **existen distritos más peligrosos significativamente que otros**. Un ejemplo claro es la diferencia entre el distrito con mayor número de crímenes, siendo el distrito de Central con un total de 45.198 casos y el distrito con menor número de denuncias es Foothill con un total de 22.991. Central es el denominado "Downtown" de Los Ángeles, en esta ubicación tiene lugar el distrito financiero y los principales estadios y museos y centros comerciales de la ciudad, mientras que Foothill es una zona residencial. En Central se ubica el denominado Skid Row que comentaremos en las siguientes páginas.

Una vez realizado el test de Moran y rechazada la hipótesis nula de ausencia de autocorrelación espacial, ya que el p-valor es prácticamente 0, también se acepta la hipótesis de partida del trabajo. Por consiguiente, **existe una correlación espacial** mayor de la que se tendría que esperar si las observaciones se repartieran aleatoriamente por la ciudad de Los Ángeles.

Sin embargo, tras el modelo generado para ajustar lo número de casos por distrito en función del efecto espacial, está covariable ha sido significativa solo al 10%, por consiguiente, los residuos han sido significativos, por lo que se ha necesitado generar más modelos que tengan en cuenta más covariables.

La siguiente variable por comentar es la influencia del desempleo en el crimen, para ello tras realizar un modelo aditivo generalizado que ajustara los casos en función del desempleo, vemos que dicho modelo nos ha proporcionado como significativa la covariable de desempleo y el coeficiente de determinación es bastante elevado. Es por ello que tras los resultados obtenidos el modelo ajusta notablemente bien los datos, además el índice de la covariable de desempleo es significativa, por lo que podemos afirmar que **el desempleo afecta significativamente al crimen.**

La última hipótesis por comentar es si el factor 'homeless' afecta al aumento del crimen. Tras generar un Modelo Aditivo Generalizado y mostrar significatividad en el componente sistemático y un R^2 bastante elevado, siendo del 86.79% podemos decir que ajusta muy bien el modelo y

que **la cantidad de personas que viven en la calle influye significativamente al número de crímenes en el área.**

Tras investigaciones sociodemográficas, se ha encontrado una explicación a la gran cantidad de crímenes y de personas 'homeless' en el distrito de Central. Se debe al denominado por los locales "Skid Row", una zona de entorno a 50 manzanas en la que se encuentran principalmente almacenes, viviendas sociales, fabricas, moteles y las propias tiendas de campaña.

Esta zona tiene sus inicios en 1976 cuando las autoridades permitieron permanecer a las personas sin hogar en esta zona para que los turistas de Los Ángeles no les vieran en las principales calles. La recesión de principios de los 80 derivó en un aumento de desempleo, sumándose así al comienzo de la epidemia de crack en la ciudad. El número de personas sin hogar aumentó y muchas personas se trasladaron a otras zonas, pero fuera de Skid Row la policía aplicaba enérgicamente la ley. En aquella época era delito sentarse, tumbarse o dormir en un espacio público, es por ello que al final todas las personas sintecho acaban en Skid Row, donde la policía le permitía acampar y drogarse a la vista de todo el mundo.

Tras posteriores recesiones como la del 2001 o la Crisis hipotecaria en 2008, aumentó a grandes niveles las personas vivienda en Skid Row, pero con la llegada del coronavirus aumentó la crispación social en la zona. Aumentó el número de personas que vivían en esta zona, pero las enfermedades y la drogadicción sumadas a la pandemia de la COVID-19 propició en una gran cantidad de muertes. Se estima que en la actualidad viven en este barrio informal cerca de 12 mil personas, entre las que según informan las autoridades la mayoría tienen problemas con el alcohol, problemas psiquiátricos, veteranos de guerra, expresidiarios o personas que han perdido su hogar. Es razonable que en este ambiente de crispación y sumado al alcohol y las drogas y la necesidad de alimentarse y tener un sitio donde dormir; sea el lugar donde más denuncias se reciben.

La ciudad de Los Ángeles en enero de este año activó el estado de emergencia sobre la crisis "homeless". El Gobierno de California y el Ayuntamiento de la ciudad de Los Ángeles han tomado durante años políticas sociales y de reinserción con el objetivo de eliminar dicho barrio informal, pero tan solo no han sido suficientes las medidas, sino que ha aumentado año tras año el número de personas que viven en estas calles. Durante los últimos 2 años, la iniciativa "*Homeless Initiative*" ha gastado más de 500 millones al año sin obtener grandes resultados, pero con la activación del estado de emergencia permitirá agilizar todo tipo de trámite relacionado con las personas sintecho.

A modo de conclusión, se puede decir que la ciudad de Los Ángeles se está enfrentando a un problema social de gran magnitud. El crimen, el desempleo, las personas sin hogar… son muchos de los rompecabezas que están teniendo las autoridades de la ciudad durante los últimos años, sin embargo, todas las medidas que están tomando parecen ser ineficaces ante la situación. Con los Juegos Olímpicos a la vuelta de la esquina y la continua evolución ascendente del crimen en Los Ángeles es un problema que se debe atajar lo más pronto posible, sino la reputación y la imagen que tiene la ciudad puede ir decayendo rápidamente.

5. Bibliografía

-Fischer, M.M. y Wang, J. (2011). Spatial Data Analysis. Models, Methods and Techniques. Springer.

-Rastegar, M (2021, 5 marzo). *Domestic Violence Stats in California. Esfandi Law Firm.* Esfandi Law Group.

https://esfandilawfirm.com/domestic-violence-stats-in-california/#:~:text=Domestic%20Violence%20Statistics%3A%20Los%20Angeles&text=The%20Los%20Angeles%20Police%20respond,for%20domestic%20violence%20each%20year

-Los Angeles County Domestic (2020, marzo) *Intimate Partner Violence: A Data Snapshot*. Los Angeles County Domestic Violence Council

http://publichealth.lacounty.gov/dvcouncil/resources/docs/snapshot_0320.pdf

-*Madrid – (Datos y estadísticas sobre violencia de género por comunidad autónoma*, s. f.) Disponible en:

https://www.epdata.es/datos/datos-graficos-violencia-genero/49/madrid/304

-ONU (2013) Estudio mundial sobre el homicidio, disponible en:

https://www.unodc.org/documents/gsh/pdfs/GLOBAL_HOMICIDE_Report_ExSum_spanish.pdf

-Ramos, A. (2022). Los delitos de odio en el condado Los Ángeles alcanzan su nivel más alto en 19 años, según reporte. CNN. Disponible en:

https://cnnespanol.cnn.com/2022/12/08/los-crimenes-de-odio-en-el-condado-los-angeles-alcanzan-su-nivel-mas-alto-en-19-anos-segun-reporte/

-Tidyverse: Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019). "Welcome to the tidyverse." *Journal of Open Source Software*, **4**(43), 1686. doi:10.21105/joss.01686.

-Lubridate: Grolemund G, Wickham H (2011). "Dates and Times Made Easy with lubridate." Journal of Statistical Software, 40(3), 1–25. https://www.jstatsoft.org/v40/i03/.

-GT: Iannone R, Cheng J, Schloerke B, Hughes E, Lauer A, Seo J (2023). gt: Easily Create Presentation-Ready Display Tables. https://gt.rstudio.com/, https://github.com/rstudio/gt.

-Ggplot2: Wickham H (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org.

-Dplyr: Wickham H, François R, Henry L, Müller K, Vaughan D (2023). dplyr: A Grammar of Data Manipulation. https://dplyr.tidyverse.org, https://github.com/tidyverse/dplyr.

-Sf: Pebesma E, Bivand R (2023). Spatial Data Science: With applications in R. Chapman and Hall/CRC. https://r-spatial.org/book/.

-RColorBrewer: Neuwirth E (2022). _RColorBrewer: ColorBrewer Palettes_. R package version 1.1-3, https://CRAN.R-project.org/package=RColorBrewer.

-GGmap: D. Kahle and H. Wickham. ggmap: Spatial Visualization with ggplot2. The R Journal, 5(1), 144-161. http://journal.r-project.org/archive/2013-1/kahle-wickham.pdf

-Leaflet: Cheng J, Karambelkar B, Xie Y (2022). _leaflet: Create Interactive Web Maps with the JavaScript 'Leaflet' Library_. R package version 2.1.1, https://CRAN.R-project.org/package=leaflet.

-Spdep: Roger S. Bivand, Edzer Pebesma, Virgilio Gomez-Rubio, 2013. Applied spatial data analysis with R, Second edition. Springer, NY. https://asdar-book.org/

-Tmap: Tennekes M (2018). "tmap: Thematic Maps in R." _Journal of Statistical Software_, *84*(6), 1-39. doi:10.18637/jss.v084.i06 <https://doi.org/10.18637/jss.v084.i06>.

-Homelessness in Los Angeles: A unique crisis demanding new solutions. (2023, 24 marzo). McKinsey & Company. Disponible en:

https://www.mckinsey.com/industries/public-and-social-sector/our-insights/homelessness-in-los-angeles-a-unique-crisis-demanding-new-solutions

-Holmes, N. S. (2021, 21 noviembre). Places in LA, Explained - Nathan S. Holmes - Medium. Medium. Disponible en:

https://nsholmes21.medium.com/the-puzzle-of-places-in-la-18e8a537e3ac

-Statista. (2022, 14 noviembre). Tasa de paro en los estados de EE. UU. en 2021. Disponible en:

https://es.statista.com/estadisticas/634616/tasa-de-desempleo-estatal-en-ee-uu/

-ArcGIS Pro | Cómo funciona la densidad kernel— Documentación. (s. f.). Disponible en:

https://pro.arcgis.com/es/pro-app/latest/tool-reference/spatial-analyst/how-kernel-density-works.htm

-Silverman, B. W. Estimación de densidad para las estadísticas y el análisis de datos. New York: Chapman and Hall, 1986.

-Anselin, Luc. 1995. "Local Indicators of Spatial Association — LISA." Geographical Analysis 27: 93–115.

-Moran, P. (1950) A Test for the Serial Independence of Residuals. Biometrika, 37, 178-181. http://dx.doi.org/10.1093/biomet/37.1-2.178

6. Apéndice

Todo el código creado, los datos utilizados y las geometrías pueden ser consultadas y descargarse en:

https://github.com/marioguillem/Crimen-En-Los-Angeles